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 a (possibly unbounded) empirical process. We deduce non-asymptotic risk bounds for an Hellinger-type loss on possibly random models. When the models are convex, 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. Besides, the robustness properties of ρ-estimators are not, in general, shared by maximum likelihood estimators. Subsequently, we present an alternative procedure to ρ-estimation, more numerically friendly, that yields a piecewise polynomial estimator. We prove theoretical results and carry out some numerical simulations that show the benefits of our approach compared with a more classical one based on maximum likelihood.

1. INTRODUCTION

1.1. Statistical settings. In the present paper, we are interesting in estimating a unknown function s that appears in one of the following frameworks.

Framework 1 (Density Estimation)**.** *Let X be a real-valued random variable with density function s with respect to the Lebesgue measure μ. Our aim is to estimate the density s 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. We suppose that T is independent of C and that T admits a density f with respect to the Lebesgue measure μ. The target function is the hazard rate s defined for all $t \in \mathbb{R}$ by*

$$s(t) = \frac{f(t)}{\mathbb{P}(T \geq t)}.$$

The observations are $(X_i, D_i)_{1 \leq i \leq n}$ where $X_i = \min\{T_i, C_i\}$ and $D_i = 1_{T_i \leq C_i}$.

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: July, 2017.

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

Key words and phrases. ρ-estimator, T-estimator, robust tests, 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* $\mathbb{R}$*,*

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

for a suitable non-negative measurable function f*.*

Our aim is to estimate the transition rate s *from state* 1 *to* 0 *defined for* $t > 0$ *by*

$$s(t) = \frac{f(t)}{\mathbb{P}\left(X_{t-} = 1\right)},$$

from the observation of n *independent copies* $\{X_t^{(i)}, t \geq 0\}$ *of* $\{X_t, t \geq 0\}$*.*

In all these frameworks, we shall always suppose that $n \geq 3$. Numerous estimation strategies are possible, and we propose here to broaden the approach developed in [BBS17] and named "ρ-estimation".

1.2. On ρ-estimation in framework 1. The procedure described in [BBS17] fits into the scheme of some ideas relating estimation and test as it appeared as early as the 70's in the works of Lucien Le Cam. Given two densities f and g, a statistical test is a decision rule that tends to decide which one is the closer of s. In order to give meaning to this notion of closeness, we consider the Hellinger distance h, which is defined for two non-negative integrable functions f and g by

$$h^2(f,g) = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(t)} - \sqrt{g(t)}\right)^2 \mathrm{d}t.$$

Several tests appeared in the literature, see [Bir06, Bir12, Bar11] and the references therein. In ρ-estimation, the idea is to try to estimate $h^2(s,f) - h^2(s,g)$ as explained in Section 1.4 of [BBS17]. The smaller this difference, the better f. Conversely, the larger this difference, the better g. Unfortunately, $h^2(s,f) - h^2(s,g)$ is difficult to directly estimate. A solution, which actually follows from [Bar11], is to estimate an approximation $T_E(f,g)$ of $h^2(s,f) - h^2(s,g)$. This estimator, we shall name $T(f,g)$, can also be interpreted as a test between f and g.

We then consider models S, that is collections of densities, which translate, in mathematical terms, the knowledge we have on the target s. A model may correspond to different assumptions on s, such as parametric, regularity, or qualitative ones. There exist several ways of deducing an estimator on S from testing, see [Bir06] for a recent reference. We may also cite the procedures of [Sar14, Sar16], which are based on $T(f,g)$, and which give a glimpse of what can be expected for these estimators in numerical simulations. In [BBS17], $T(f,g)$ is not only used as a test, but more precisely as an estimate of $h^2(s,f) - h^2(s,g)$. We may then form the criterion $\gamma(f) = \sup_{g\in S} T(f,g)$ and interpret it as an estimator of $h^2(s,f) - \inf_{g\in S} h^2(s,g)$. It then remains to minimize this criterion to define the ρ-estimator (if such a minimizer does not exist, take an approximate minimizer).

Before going any further, we need to mention that we may construct several variants of the ρ-estimation procedure that lead to similar theoretical results, at least in density estimation. We may for instance, change the estimator $T(f,g)$, see [BB17] and our Section 2.3. In particular, in density estimation, it will be convenient to deal with (non-negative) functions f and g that may not be densities. When f and g are densities, in framework 1, our definition of $T(f,g)$ coincides with the one of [BB17]. Besides, as it will be explained later, some results remain true when the criterion γ is replaced by a T-procedure [Bir06] or by the one of Section 4.1 of [Sar14]. It turns out that ρ-estimators (and some variants) satisfy interesting statistical properties. We briefly present below four of them: generality, optimality, robustness, and superminimaxity.

First of all, the theoretical performances of ρ-estimators rely on the behaviour of the random process $T(f,g)-T_E(f,g)$. The deviations of this process can be controlled according to different notions that aim at measuring the "complexity", or "massiveness" of S (entropy with bracketing, universal entropy, metric dimension, ...). Thereby, we may control the risks of ρ-estimators in various models of interest, including in particular the ones for which other procedures may not work, such as the maximum likelihood one (see [BBS17, BB16] for examples).

We may moreover generally compute an upper-bound $R_S(n)$ of the maximal risk $\sup_{s\in S} h^2(s,\hat{s})$ of a ρ-estimator $\hat{s}$. The quality of the estimator $\hat{s}$ can then be assessed by comparing $R_S(n)$ to the minimax bound $\inf_{\tilde{s}}\sup_{s\in S} h^2(s,\tilde{s})$, where the infimum is taken over all estimators $\tilde{s}$ with values in S. The rate of convergence of the minimax bound $\inf_{\tilde{s}}\sup_{s\in S} h^2(s,\tilde{s})$ to 0 is usually called the optimal minimax rate of convergence. Yet, $R_S(n)$ achieves this rate, up to a possible logarithmic factor, in all cases we know.

This minimax point of view supposes that s 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 s lies in S but more generally when s 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

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

where C is a universal constant (that is a numerical number). This inequality asserts that a small error in the choice of the model S induces a small error in the estimation of s. This is a robustness property. We recall that such a property is not shared by all estimators, and in particular by the maximum likelihood one, which may perform very poorly when $s \notin S$ but is close to S (when this closeness is measured by the Hellinger distance, see Section 2.3 of [Bir06] for an example).

The rate given by $R_S(n)$ stands for the worst-case rate over all densities s of S. This rate may therefore be very pessimistic in the sense that the estimation may be much faster for some densities $s \in S$. One may actually refine the preceding risk bound to take into account this phenomenon (named superminimaxity in [BB16]). More precisely, it is shown in [BB16] a non-asymptotic risk bound of the form

$$\mathbb{E}[h^2(s,\hat{s})] \le C' \inf_{f\in\bar{S}} \left\{h^2(s,f) + R_S(f,n)\right\} \quad \text{whatever the density } s, \tag{1}$$

where C' is a universal constant, $\bar{S}$ a subset of S, and where $R_S(f,n)$ tends to 0 faster than the optimal rate on S. For illustration purposes, consider the model

$$S = \{f1_I,\, I \text{ is an interval of } \mathbb{R} \text{ and } f \text{ a non-increasing density on } I\}\,, \tag{2}$$

where 1_I denotes the indicator function of I. Then, $\bar{S}$ consists of piecewise constant densities belonging to S,

$$R_S(f,n) = \frac{d(f)}{n}\log_+^3\left(\frac{n}{d(f)}\right),$$

where $d(f)$ is the number of pieces of f, and $\log_+ x = \max\{\log x, 1\}$. In particular, when s is piecewise constant, the rate of convergence is parametric (up to a logarithmic term). If we now only suppose that s belongs to S, we may bound the infimum (1) from above by $C_s(\log^3 n/n)^{1/3}$, where C_s only depends on s, which corresponds, up to a logarithmic factor, to the expected rate of convergence. The previous reasoning is not only valid for this particular model S but is more general and holds true for numerous models S of interest corresponding to different qualitative assumptions on the density (s may be piecewise monotone, piecewise log-concave, $\sqrt{s}$ may be piecewise convex-concave...).

There is moreover two additional properties of ρ-estimators we now briefly mention. First, ρ-estimators can be related to maximum likelihood ones. This will be further detailed in Section 1.5. Second, it is possible to introduce penalties into the criterion γ, leading to penalized ρ-estimators and allowing to cope with model selection.

1.3. On hazard rate and transition intensity estimation. In this paper, we propose to extend the scope of ρ-estimation to these two statistical settings. The first one, namely hazard rate estimation, frequently appears in different domains such as reliability or survival analysis. Typically, in medical studies, T may represent the lifetime of a patient, and the hazard rate s at time t,

$$\begin{aligned} s(t) &= \frac{f(t)}{\mathbb{P}\left(T \geq t\right)}, \\ &= \lim_{h\to 0}\frac{\mathbb{P}\left(t \leq T \leq t+h \mid T \geq t\right)}{h}, \end{aligned}$$

stands for the propensity of succumbing just after t, given survival to time t. In practice, some patients may leave the study before dying, which makes the data censored. The random variable C then gives the time of leaving and $D = 1_{T\leq C}$ indicates whether the patient dies ($D = 1$) or leaves the study ($D = 0$).

Note that this problem differs from the one of density estimation, even when the data are not censored, that is when one observes $T_1, \dots, T_n$. Indeed, although that the survival function $t \mapsto \mathbb{P}\left(T \geq t\right)$ may be estimated at a parametric rate, it is different to put an assumption on s and on f. The problem of hazard rate estimation is moreover more delicate as it cannot be uniformly estimated on $[0, +\infty)$. It is actually better estimated in regions of high value of $\mathbb{P}\left(T \geq t\right)$ than in regions of low value, see for instance [Efr16].

The problem of transition intensity estimation of a Markov process may also be encountered in various domains. For example, 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 s at time t,

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

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 s coincides with the hazard rate of T when the Markov process is defined by $X_t = 1_{T\geq t}$.

In the literature, numerous estimators are 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, and a piecewise constant estimator based on a device inspired from [Bir06]. Non-asymptotic studies seem, however, to be more scarce. We may cite [BC05, BC08, Pla09, AD10, RB06] for procedures based on (penalized) $\mathbb{L}^2$ contrasts, [vdG95, DR02] for maximum likelihood estimators, and [BB09] for a piecewise constant estimator whose partition is suitably selected from the data.

1.4. A generalized procedure, a new probabilistic tool, and new risk bounds. We propose in this paper to consider a general setup that will make it possible the simultaneous study of the three frameworks.

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

$$h^2(f,g) = \frac{1}{2}\int_0^\infty \left(\sqrt{f(t)}-\sqrt{g(t)}\right)^2 \left(\frac{1}{n}\sum_{i=1}^n 1_{X_i\geq t}\right) \mathrm{d}t,$$

and in framework 3,

$$h^2(f,g) = \frac{1}{2}\int_0^\infty \left(\sqrt{f(t)}-\sqrt{g(t)}\right)^2 \left(\frac{1}{n}\sum_{i=1}^n 1_{X_{t-}^{(i)}=1}\right) \mathrm{d}t.$$

The quality of an estimator $\hat{s}$ is therefore assessed by $h^2(s,\hat{s})$: the smaller $h^2(s,\hat{s})$, the better the estimator. Note that this Hellinger-type distance puts more weight on regions of $\mathbb{R}$ where estimation seems easier.

As explained in Section 1.2, we shall estimate an approximation $T_E(f,g)$ of $h^2(s,f)-h^2(s,g)$. The resulting estimator will then be defined as a minimizer, or more precisely as an approximate minimizer of $\gamma(f) = \sup_{g\in S} T(f,g)$ where $T(f,g)$ estimates $T_E(f,g)$. Similarly to framework 1 controlling the Hellinger-type risk of a ρ-estimator requires to control the deviations of a centered empirical process $T(f,g)-T_E(f,g)$. This process becomes however unbounded in frameworks 2 and 3. We shall carry out a result to control these deviations and use it to study the theoretical performances of ρ-estimators. More precisely, we shall establish a risk bound akin to the one (1) obtained in density estimation by [BB16]. Actually, in density estimation, our estimation term $R_S(f,n)$ is slightly smaller. Moreover, this bound is valid in frameworks 2 and 3 without

any additional assumption on these statistical settings. This new risk bound being very similar to (1), all the rates obtained by [BB16] for ρ-estimators in density estimation under qualitative assumptions can be transferred in frameworks 2 and 3 (up to a minor improvement).

1.5. On maximum likelihood estimation. A ρ-estimator $\hat{s}$ is defined as a solution of a minimax problem, which seems, unfortunately, be numerically difficult to solve in general. It turns out that this optimization problem may sometimes reduce to that of log likelihood maximization (which is thus more easier to solve).

This phenomenon can be explained by the local behaviour of the estimator $T(f,g)$. Indeed, when f and g are very close densities in framework 1, $T(f,g)$ roughly behaves as the difference of $L(g)-L(f)$ where $L(\cdot)$ denotes the log-likelihood. This behaviour was exploited in density estimation in Section 5 of [BBS17], and Sections 2.6, and 3.4 of [Sar16], to make a connection between ρ- and maximum likelihood estimation (see also the numerical simulations in [Sar16]).

When T is moreover a convex-concave function, as it is the case in the present paper in our three different statistical settings, ρ-estimators mainly reduce to that of maximum likelihood ones when the models are convex (up to a minor modification in the definition of the log likelihood if S does not only consist of densities in framework 1). We recover, in particular, a result of Su Weijie included very recently in [BB17] in the context of framework 1.

It is worthwhile to notice that these two estimation methods differ in general. In particular, a maximum likelihood estimator may be very sensitive to a small model error measured by the Hellinger distance, as it was already mentioned in Section 1.2. Moreover, maximum likelihood estimation may fail in models where ρ-estimation works (as it is the case for example for the model S defined in (2)).

1.6. Practical estimation. In practice, the model S is often chosen according to the data, and should therefore theoretically considered as random. In this paper, we will also be able to control the risks of ρ-estimators for random models. As a corollary, we may consider models S consisting of estimators, in which case the procedure amounts to performing estimator selection. The numerical complexity of this selection rule is roughly of the order of the square of the number of estimators. It may therefore be implemented in practice when this number is not too large.

Estimator selection may be an alternative way to build, in practice, estimators with nice theoretical properties on models S where the computation of ρ-estimators seems be numerically intractable. The idea is to proceed in two steps. We decompose the model S as a union $S=\bigcup_{m\in\mathcal{M}} S_m$ of convex models S_m. We maximize the log likelihood on S_m to get a ρ-estimator $\hat{s}_m$ and then select among them.

Here, we shall consider the model $S_{\ell,r}$ consisting of non-negative piecewise polynomial functions of degree at most r based on at most ℓ pieces. Although maximum likelihood estimators do not exist on $S_{\ell,r}$, ρ-estimators do exist, and we may even control their Hellinger-type risks. Unfortunately, we do not know how to build these ρ-estimators in practice. As explained above, a solution is to consider for each (finite) partition m of $\mathbb{R}$ into intervals, the model $S_{\ell,r,m}\subset S_{\ell,r}$ consisting of functions which are polynomial on each interval I of m. Then, $S_{\ell,r}=\bigcup_{m\in\mathcal{M}} S_{\ell,r,m}$ where the union is taken over all partitions m of $\mathbb{R}$ into at most ℓ intervals. We may then build in practice for each $m\in\mathcal{M}$ the maximum likelihood estimator $\hat{s}_m$ on $S_{\ell,r,m}$. Selecting among

all the estimators $\hat{s}_m$ is theoretically feasible but does not yield a practical procedure as $\mathcal{M}$ is infinite. This is the reason why, we shall rather consider a finite (but very large) collection $\widehat{\mathcal{M}} \subset \mathcal{M}$ of partitions depending on the data, and carry out a new procedure to select among the estimators $\hat{s}_m$, $m \in \widehat{\mathcal{M}}$. Although the large cardinality of $\widehat{\mathcal{M}}$, dynamic programming makes it possible the computation of the selected estimator in practice, at least when n is not too large (the numerical complexity significantly increases with n). We study the selected estimator from a theoretical and practical point of view. We first prove an oracle inequality, and then carry out a small numerical study in which we compare this procedure with a selection rule based on maximum likelihood.

We finally explain how we can adapt our procedure to build an estimator with nice statistical properties when s belongs, or is close to, the more general model $S_r = \cup_{\ell=1}^{\infty} S_{\ell,r}$. The risk bound we get 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 built on $S_{\ell,r}$ (up to slightly modifications). This best estimator, is of course, unknown in practice as the best choice of ℓ depends on the unknown function s. This procedure explains therefore how to choose ℓ from the data, or in other terms, how to build an estimator adaptive with respect to ℓ. Interestingly, this estimator can be built in practice, when n is small enough (even if the computation of a single estimator $\hat{s}_{\ell,r}$ seems be numerically intractable).

1.7. Organization of the paper. We carry out in Section 2 the general statistical setting that encompasses the three different frameworks. We then explain our estimation procedure for deterministic models 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, as well as the required assumptions on models S. We then present our main theorem on the theoretical performances of ρ-estimators. In Section 4, we deal with random models and estimator selection as explained in Section 1.6. Section 5is devoted to numerical simulations. The proofs are deferred to Section 6.

2. The ρ-estimation method

2.1. Statistical setting and notations. 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 $M(A)$. More precisely, we set in density estimation,

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

and in hazard rate estimation,

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

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

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

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

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

Our aim is to estimate s from the observation of the random measures N and M.

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

$$h^2(f,g) = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(t)} - \sqrt{g(t)}\right)^2 \mathrm{d}M(t),$$

for all non-negative measurable functions f, and g which are integrable 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 $x, y \in \mathbb{R}$, $x \wedge y = \min(x, y)$. 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 finite set A by $|A|$. We set $\log_+ x = \max\{\log x, 1\}$ for all $x > 0$. The notations $c, c', C, C', \ldots$ are for the constants. These constants may change from line to line.

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

Consider two two arbitrary functions f, g of $\mathcal{S}$ and the model $S = \{f, g\}$. As explained in the introduction, we begin by defining an approximation $T_E(f, g)$ of $h^2(s, f) - h^2(s, g)$.

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 $f, g \in \mathcal{S}$, we set

$$T_E(f,g) = \int_{\mathbb{R}} \psi\left(\frac{g(x)}{f(x)}\right) s(x)\,\mathrm{d}M(x) - \frac{1}{4}\int_{\mathbb{R}} (g(x) - f(x))\,\mathrm{d}M(x).$$

In this definition, and throughout the paper, we use the conventions $0/0 = 1$ and $x/0 = \infty$ for all $x > 0$. The quantity $T_E(f, g)$ is unknown in practice as it involves s, but can easily be estimated by

$$T(f,g) = \int_{\mathbb{R}} \psi\left(\frac{g(x)}{f(x)}\right) \mathrm{d}N(x) - \frac{1}{4}\int_{\mathbb{R}} (g(x) - f(x))\,\mathrm{d}M(x).$$

Some computations show:

Lemma 1. *For all $f, g \in \mathcal{S}$,*

$$\frac{1}{3}h^2(s,f) - 3h^2(s,g) \leq T_E(f,g) \leq 3h^2(s,f) - \frac{1}{3}h^2(s,g). \tag{3}$$

In particular, if $T_E(f,g)$ is non-negative, then $h^2(s,g) \leq 9h^2(s,f)$. Conversely, if $T_E(f,g)$ is non-positive, then $h^2(s,f) \leq 9h^2(s,g)$. In other words, the sign of $T_E(f,g)$ allows us to know which function among f, g is the closest of s (up to a multiplicative constant).

For more general sets S, and $f \in S$, we are interested in evaluating $h^2(s,f) - h^2(s,S)$. The smaller this number, the better f. As $T_E(f,g)$ is roughly of the order of $h^2(s,f) - h^2(s,g)$, it is natural to approximate $h^2(s,f) - h^2(s,S)$ by $\gamma_E(f) = \sup_{g\in S} T_E(f,g)$ and to study the properties of the minimizers of γ_E.

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

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

Minimizing γ_E over S yields a function $\bar{f} \in S$ (assuming such a function exists) such that,

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

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

The main interest of $\gamma_E(f)$ with respect to $h^2(s,f) - h^2(s,S)$ lies in the fact that it can be estimated in practice by $\gamma(f) = \sup_{g\in S} T(f,g)$. It remains to minimize this criterion to define our estimator as described below.

2.3. The procedure. Let for $f, g \in \mathcal{S}$,

$$T(f,g) = \int_{\mathbb{R}} \psi\left(\frac{g(x)}{f(x)}\right) \mathrm{d}N(x) - \frac{1}{4}\int_{\mathbb{R}} (g(x) - f(x)) \, \mathrm{d}M(x). \tag{4}$$

Let S be a model and for $f \in S$, $\gamma(f) = \sup_{g\in S} T(f,g)$. Any function $\hat{s} \in S$ satisfying

$$\gamma(\hat{s}) \leq \inf_{f\in S} \gamma(f) + 1/n \tag{5}$$

is called ρ-estimator.

Remark 1. Since S is not necessarily countable, $\hat{s}$ may not be measurable in general, which means that it is slightly abusive to say that $\hat{s}$ is an estimator. Moreover, bounding the risk $h^2(s,\hat{s})$ of the ρ-estimator requires the use of the outer probability measure $\mathbb{P}^\star$ and the outer expectation $\mathbb{E}^\star$. We refer for instance to Chapter 1.2 of [VDVW96] for the definitions and the properties of $\mathbb{P}^\star$, $\mathbb{E}^\star$. However, in the examples of models S described in Section 3.2, we shall see that there exists a deterministic, countable and dense subset $\bar{S}_{count} \subset S$ in the metric space $(\mathbb{L}^1_+(\mathbb{R}, M), h)$ such that $\gamma(f) = \sup_{g\in \bar{S}_{count}} T(f,g)$ and $\inf_{f\in S} \gamma(f) = \inf_{f\in \bar{S}_{count}} \gamma(f)$. For such models S, $\hat{s}$ can be chosen in a measurable way, the outer probability measure $\mathbb{P}^\star$ can be replaced by $\mathbb{P}$, and the outer expectation $\mathbb{E}^\star$ by $\mathbb{E}$.

Remark 2. As explained in the above heuristics, the present procedure depends on two key ingredients. First, we need to approximate $h^2(s,f) - h^2(s,g)$ by a quantity $T_E(f,g)$ that satisfies an inequality akin to (3). Second, we need a random variable $T(f,g)$ that can be computed

in practice and that is close enough to $T_E(f,g)$ (for more details about the meaning of "close enough", we refer to Section 3.1). There may exist other choices of $T_E(f,g)$ and $T(f,g)$ that met these criteria and for which the estimator $\hat{s}$ defined by (5) would perform similarly. For the sake of simplicity, we stick throughout the paper to our particular definition of $T_E(f,g)$ and $T(f,g)$. When f and g are supposed to be densities in framework 1, $T_E(f,g)$ becomes $\int_{\mathbb{R}} \psi(g/f)s \,\mathrm{d}\mu$. We then recover an approximation of $h^2(s,f) - h^2(s,g)$ that appears in [BB17].

Remark 3. Minimizing γ to derive a good estimator based on $T(f,g)$ is not the only solution. We could, for instance, mimic the definition of T-estimators of [Bir06], or draw inspiration from [Sar14]. This would lead to the criteria $\wp_1$ and $\wp_2$ defined for $f \in S$ by

$$\wp_1(f) = \sup_{\substack{g\in S \\ T(f,g)\geq 0}} h^2(f,g), \quad \text{and} \quad \wp_2(f) = \sup_{g\in S} \left\{\alpha h^2(f,g) + T(f,g)\right\},$$

where α is a fixed number $\alpha \in (0,1/2)$. The estimator $\tilde{s}_i$ of the i^{th} procedure would be defined as a minimizer of $\wp_i$ over S, or more precisely, as any element $\tilde{s}_i \in S$ such that

$$\wp_i(\tilde{s}_i) \leq \inf_{f\in S} \wp_i(f) + 1/n.$$

It actually turns out that our main result in Section 3 (Theorem 4) remains true for the two estimators $\tilde{s}_1$ and $\tilde{s}_2$ (this requires, however, to modify the proofs, and to adapt the constants in the risk bounds. These constants may depend on α for the second procedure). We do not develop these points to keep this paper a reasonable size.

2.4. Connection with the maximum likelihood estimator. In the context of framework 1 (density estimation), if the functions of S are densities, we may maximize the log likelihood on S to define the celebrated maximum likelihood estimator. When S is a subset of $\mathcal{S} = \mathbb{L}^1_+(\mathbb{R},\mu)$, the functions f of S may not be densities, and we need to extend this definition. We propose to define $L(f)$ by

$$L(f) = \int_{\mathbb{R}} \log f \,\mathrm{d}N - \int_{\mathbb{R}} f \,\mathrm{d}M \quad \text{for all } f \in \mathcal{S}, \tag{6}$$

and to call maximum likelihood estimator any estimator maximizing L on S. In the above formula, and throughout the paper, the convention $\log 0 = -\infty$ is used. This definition coincides with the usual one in framework 1 when S consists of densities. Moreover, L is the log likelihood in frameworks 2 and 3, see, for instance, equation (3.2) of [Ant89] (where it is used that s is the Aalen's multiplicative intensity of a counting process).

It turns out that ρ-estimators and maximizers of L coincide with probability 1 for numerous models S of interest. To explain this phenomenon, remark that $T(f,g)$ also writes

$$T(f,g) = \int_{\mathbb{R}} \tanh\left(\log g - \log f\right) \,\mathrm{d}N - \frac{1}{4}\int_{\mathbb{R}} (g-f) \,\mathrm{d}M \quad \text{for all } f,g \in \mathcal{S}.$$

As $\tanh(x) \simeq x/4$ when $x \simeq 0$, we deduce that if $\tilde{s}$ maximizes L and $g \simeq \tilde{s}$,

$$\begin{aligned} T(\tilde{s},g) &\simeq \frac{1}{4}\left(\int_{\mathbb{R}} \log g \,\mathrm{d}N - \int_{\mathbb{R}} \log \tilde{s} \,\mathrm{d}N\right) - \frac{1}{4}\left(\int_{\mathbb{R}} g \,\mathrm{d}M - \int_{\mathbb{R}} \tilde{s} \,\mathrm{d}M\right) \\ &\simeq \frac{1}{4}\left(L(g) - L(\tilde{s})\right). \end{aligned}$$

Thereby, $T(\tilde{s}, g)$ is likely non-positive. Under suitable properties of S, this result does not only occur when $g \simeq \tilde{s}$, but also for all $g \in S$, 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 $\mathcal{S}$. Let $\mathcal{X}$ be a subset of $\mathbb{R}$ such that $\{x \in \mathbb{R},\, f(x) \neq 0\} \subset \mathcal{X}$ for all $f \in S$. Define*

$$L_{\mathcal{X}}(f) = \int_{\mathcal{X}} \log f \,\mathrm{d}N - \int_{\mathcal{X}} f \,\mathrm{d}M \quad \text{for all } f \in S,$$

and suppose that $\sup_{g \in S} L_{\mathcal{X}}(g) \notin \{-\infty, +\infty\}$.

If there exists $\tilde{s} \in S$ such that $L_{\mathcal{X}}(g) \leq L_{\mathcal{X}}(\tilde{s})$ for all $g \in S$, 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 $g \in S$, $L_{\mathcal{X}}(g) \leq L_{\mathcal{X}}(\hat{s})$, and $\hat{s}$ maximizes $L_{\mathcal{X}}$ over S.

When $\mathcal{X} = \mathbb{R}$, $L_{\mathcal{X}} = L$, which means that results on maximum likelihood estimators may be derived from that of ρ-estimators and vice versa. We recover a result of Su Weijie included very recently in [BB17] in framework 1. Using sets $\mathcal{X}$ not equal to $\mathbb{R}$ may be of interest to maintain a connection between these two approaches in models where the maximum likelihood estimator does not exist, as explained below.

We consider the convex model S in framework 1 defined by

$$S = \left\{ f 1_{(0,+\infty)},\ f \text{ is a non-increasing function of } \mathcal{S} \text{ on } \mathbb{R} \right\}. \tag{7}$$

When the random variables X_i are positive, which in particular holds true μ a.s. if s 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 $\mathcal{X} = \mathbb{R}$ that this estimator is, in this case, a ρ-estimator. When some of the random variables X_i are non-positive, $L(g) = -\infty$ for all $g \in S$, and we cannot maximize L over S to design an estimator. However, the ρ-estimation approach works and still coincides with the maximum likelihood one, up to some minor modifications. Indeed, in this case, the preceding theorem can be used with $\mathcal{X} = (0, +\infty)$. Then, $L_{\mathcal{X}}(f)$ takes the form

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

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

$$f \mapsto \frac{1}{n_0} \sum_{\substack{i \in \{1,\dots,n\} \\ X_i > 0}} \log f(X_i) - \int_0^\infty f(t) \,\mathrm{d}t$$

over $f \in S$, where n_0 is the number of positive random variables among $X_1, \dots, X_n$. One can verify that the estimator that maximizes $L_{\mathcal{X}}$ over S, and which is thus the 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 the ρ-estimator is not a density unless that the observations X_i are all positive.

One may wonder whether it is relevant to estimate s by an estimator that is not a density. The following well-known lemma claims that it is always possible to transform an estimator into a density and relate the risks of these two estimators.

Lemma 2. *Let, in density estimation, f be a non-zero function of $\mathcal{S}$ and*

$$g(x) = \frac{f(x)}{\int_{\mathbb{R}} f(t)\,\mathrm{d}t} \quad \text{for all } x \in \mathbb{R}.$$

Then,

$$h^2(s,g) \leq 2h^2(s,f).$$

However, the constant 2 cannot be improved in general, and $h^2(s,g)$ may be larger than $h^2(s,f)$. Moreover, as we shall see in Section 3.3, the ρ-estimator $\hat{s} = (n_0/n)\tilde{s}$ satisfies the following additional property: $h^2(s1_{(0,+\infty)}, \hat{s})$ tends a.s. to 0 when n goes to $+\infty$ whenever s is non-increasing on $(0,+\infty)$. By the way, we may specify the rate of convergence, see Section 3.5. This result means that $\hat{s}$ will consistently estimate s on $(0,+\infty)$. This phenomenon cannot occur for density estimators with values in S when the probability $\mathbb{P}(X < 0)$ is positive.

3. Risk bounds of ρ-estimators

3.1. An exponential inequality. We recall that the definition of ρ-estimators is based on the minimization of a criterion γ on S. This criterion γ uses the approximation $T(f,g) \simeq T_E(f,g)$ where $f, g \in S$ as explained in Section 2.2. 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 $\varphi \in \mathcal{S}$, the random variable

$$Z(\varphi) = \int_{\mathbb{R}} \varphi(x)\,\mathrm{d}N(x) - \int_{\mathbb{R}} \varphi(x)s(x)\,\mathrm{d}M(x).$$

This variable is centered in each of the different statistical settings. Note that $Z(\varphi)$ measures the approximation error of $T_E(f,g)$ by $T(f,g)$ when $\varphi = \psi(g/f)$. The aim of the theorem below is to control the fluctuations of $Z(\varphi)$.

Theorem 2. *Let $\mathcal{F} \subset \mathcal{S}$ be a set of functions φ such that $|\varphi(x)| \leq 1$ for all $\varphi \in \mathcal{F}$, $x \in \mathbb{R}$. Assume that for all $t \in (0,1)$, the sets $\{x \in \mathbb{R}, \varphi_+(x) > t\}$ and $\{x \in \mathbb{R}, \varphi_-(x) > t\}$ are unions of at most d intervals ($d \geq 1$). Let, for $\varphi \in \mathcal{F}$,*

$$\upsilon(\varphi) = \int_{\mathbb{R}} \varphi^2(x)s(x)\,\mathrm{d}M(x).$$

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

$$|Z(\varphi)| \leq C\left\{\sqrt{\upsilon(\varphi)\log_+(1/\upsilon(\varphi))\left(\frac{d\log_+(n/d)}{n} + \xi\right)} + \frac{d\log_+(n/d)}{n} + \xi\right\}. \tag{8}$$

Moreover, for all $\varepsilon \in (0,1]$,

$$|Z(\varphi)| \leq \varepsilon\upsilon(\varphi) + C'_\varepsilon\left\{\frac{d\log_+^2(n/d)}{n} + \xi\log_+(1/\xi)\right\}. \tag{9}$$

In the above inequalities, C is a universal constant while C'_ε only depends on ε.

In this theorem, and throughout the paper, we use the convention that the empty set $\emptyset$ is an interval. In framework 1, $Z(\varphi)$ and $v(\varphi)$ merely writes

$$Z(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i) - \mathbb{E}[\varphi(X_i)]\right) \quad \text{and} \quad v(\varphi) = \mathbb{E}[\varphi^2(X_1)].$$

In framework 2, $Z(\varphi)$ and $v(\varphi)$ take the form

$$Z(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left(\varphi(X_i)1_{D_i=1} - \int_0^{X_i}\varphi(t)s(t)\,\mathrm{d}t\right),\ v(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{X_i}\varphi^2(t)s(t)\,\mathrm{d}t.$$

In framework 3,

$$Z(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left(\varphi(T_{1,0}^{(i)})1_{T_{1,0}^{(i)}<\infty} - \int_0^{+\infty}\varphi(t)1_{X_{t-=1}^{(i)}}s(t)\,\mathrm{d}t\right),\ v(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{+\infty}\varphi^2(t)1_{X_{t-=1}^{(i)}}s(t)\,\mathrm{d}t.$$

In each of these two latter frameworks, $v(\varphi)$ plays the role of a moment of order 2, as in density estimation (with the noticeable difference that $v(\varphi)$ is now random). Note that we do not require that $\int_0^{X_i}\varphi(t)s(t)\,\mathrm{d}t$ and $\int_0^{+\infty}\varphi(t)1_{X_{t-=1}^{(i)}}s(t)\,\mathrm{d}t$ are bounded almost surely.

In this theorem, we measure the complexity of the collection $\mathcal{F}$ by using the notion of "union of intervals". This notion is general enough to control the risks of ρ-estimators in numerous models S of interest. It is however possible to weaken this assumption in framework 1. Indeed, the proof of the theorem is based on a uniform control of $Z(1_A)$ where A spans a collection of the form

$$\mathcal{A}_t = \left\{\left\{x\in\mathbb{R},\, \varphi_+(x) > t\right\}, \varphi\in\mathcal{F}\right\} \cup \left\{\left\{x\in\mathbb{R},\, \varphi_-(x) > t\right\}, \varphi\in\mathcal{F}\right\}. \tag{10}$$

In density estimation, this uniform control follows from Vapnik-Chervonenkis inequalities for relative deviation. We may therefore deal with any class $\mathcal{F}$ for which these inequalities apply. Actually, we may prove:

Corollary 1. *Consider framework 1 and an at most countable set $\mathcal{F}\subset\mathcal{S}$ of functions φ such that $|\varphi(x)|\le 1$ for all $x\in\mathbb{R}$, $\varphi\in\mathcal{F}$. Let for $t\in(0,1)$, $\mathcal{A}_t$ be the collection defined by (10), and $S_{\mathcal{A}_t}(n)$ be the Vapnik-Chervonenkis shatter coefficient defined by*

$$S_{\mathcal{A}_t}(n) = \max_{x_1,\dots,x_n\in\mathbb{R}}\left|\left\{\{x_1,\dots,x_n\}\cap A, A\in\mathcal{A}_t\right\}\right|.$$

Suppose that there exists $\sigma^2\in(0,1]$ such that $\sup_{\varphi\in\mathcal{F}}\mathbb{E}[\varphi^2(X_1)]\le\sigma^2$. Then, there exists a universal constant C such that

$$\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}}|Z(\varphi)|\right] \le C\left[\sigma\sqrt{\frac{\sup_{t\in(0,1)}\log_+|S_{\mathcal{A}_t}(2n)|\log_+(1/\sigma)}{n}} + \frac{\sup_{t\in(0,1)}\log_+|S_{\mathcal{A}_t}(2n)|}{n}\right].$$

In the literature, several papers study this expectation under different assumptions on $\mathcal{F}$, see [GK06], Chapter 13 of [BLM13] and [Bar16]. Our result can be compared to Inequality (2.8) of [Bar16]. Indeed, suppose that $\mathcal{F}$ is weak VC-major with dimension d in the sense of Definition 2.3 of [Bar16]. Then, $\mathcal{A}_t$ is Vapnik-Chervonenkis and Sauer's lemma implies

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

where C' is a numerical number. If we put aside the constant C', the main difference between this inequality and the one 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 2 is well tailored for bounding the risk of a ρ-estimator. Indeed, when $\varphi = \psi(g/f)$, the random variable $\upsilon(\varphi)$ can be related to the Hellinger distances between s, f and g:

Lemma 3. *For all $f, g \in \mathcal{S}$,*

$$\int_{\mathbb{R}} \psi^2(g/f) s \, \mathrm{d}M \leq 3\left(h^2(s,f) + h^2(s,g)\right).$$

Under suitable assumptions on the collection $\mathcal{F} = \{\psi(f/g), f, g \in S\}$, Inequality (9) roughly says that with high probability (and $\varepsilon = 1/18$):

$$|T(f,g) - T_E(f,g)| \leq \frac{1}{6}\left(h^2(s,f) + h^2(s,g)\right) + D_S(n) \quad \text{for all } f, g \in S. \tag{11}$$

The term $D_S(n)$ depends on the probability of the event on which (11) is true and the complexity of S. The approximation $T(f,g) \simeq T_E(f,g)$ is then accurate enough to control the risk of the ρ-estimator $\hat{s}$. It indeed suffices to mimic the computations done in Section 2.2: we deduce from (3), that for all $f, g \in S$,

$$\frac{1}{6}h^2(s,f) - \frac{19}{6}h^2(s,g) - D_S(n) \leq T(f,g) \leq \frac{19}{6}h^2(s,f) - \frac{1}{6}h^2(s,g) + D_S(n).$$

Therefore,

$$\frac{1}{6}h^2(s,f) - \frac{19}{6}h^2(s,S) - D_S(n) \leq \gamma(f) \leq \frac{1}{6}h^2(s,f) - \frac{19}{6}h^2(s,S) + D_S(n),$$

and hence,

$$\begin{aligned} \frac{1}{6}h^2(s,\hat{s}) - \frac{19}{6}h^2(s,S) - D_S(n) &\leq \gamma(\hat{s}) \\ &\leq \inf_{f \in S} \gamma(f) + 1/n \\ &\leq \frac{19}{6}\inf_{f \in S} h^2(s,f) - \frac{1}{6}h^2(s,S) + D_S(n) + 1/n. \end{aligned}$$

Finally, the risk of a ρ-estimator $\hat{s}$ is bounded above by

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

It remains to explain the assumptions to put on the model S to make inequality (11) more precise and rigorous.

3.2. Assumptions on models. We shall be able to control the risks of ρ-estimators for models S satisfying the assumption below.

Assumption 1. *There exists $\bar{S} \subset \mathcal{S}$ such that for all $t > 0$, $f \in \bar{S}$, $g \in S$ the set*

$$\{x \in \mathbb{R},\ g(x) > t f(x)\}$$

is a union of at most $d_S(f) \geq 1$ intervals.

This assumption is satisfied for numerous models of interest. We carry out below three examples. The first model is the collection $\mathcal{P}_{\ell,0}$ of piecewise constant functions defined by

$$\mathcal{P}_{\ell,0} = \left\{ \sum_{j=1}^{\ell} a_j 1_{K_j},\ (a_j)_{1\le j\le \ell} \in \mathbb{R}_+^{\ell},\ K_j \text{ is an interval of } \mathbb{R} \text{ of finite length} \right\}. \tag{12}$$

More generally, we may consider the set of (non-negative) piecewise polynomial functions defined for $r \ge 1$ by

$$\mathcal{P}_{\ell,r} = \left\{ \sum_{j=1}^{\ell} P_j 1_{K_j},\ P_j \text{ is a polynomial function of degree at most } r, \text{ which is non-negative on an interval } K_j \text{ of } \mathbb{R} \text{ and is such that } P_j 1_{K_j} \in \mathcal{S} \right\}. \tag{13}$$

We may also consider the slight variant:

$$\mathcal{P}_{\ell,r,+} = \left\{ \sum_{j=1}^{\ell} (P_j)_+ 1_{K_j},\ P_j \text{ is a polynomial function of degree at most } r, \text{ such that } (P_j)_+ 1_{K_j} \in \mathcal{S}, \text{ and } K_j \text{ is an interval of } \mathbb{R} \right\}. \tag{14}$$

We may also deal with the collection $\mathcal{F}_k$ of piecewise monotone functions defined for $k \ge 1$ by

$$\mathcal{F}_k = \left\{ \sum_{j=1}^{k} f_j 1_{K_j}, \text{where } K_j \text{ is an interval of } \mathbb{R},\ f_j \in \mathcal{S} \text{ is monotone on } K_j \right\}. \tag{15}$$

Note that unimodal functions belong to $\mathcal{F}_k$ as soon as $k \ge 2$.

Proposition 3. *Assumption 1 is fulfilled with:*

- $S = \mathcal{P}_{\ell,r}$, $\bar{S} \subset \mathcal{P}_{\ell,r}$ *and for all* $f \in \bar{S}$, $d_S(f) = (r+2)(2\ell+1)$.
- $S = \mathcal{P}_{\ell,r,+}$, $\bar{S} \subset \mathcal{P}_{\ell,r,+}$ *and for all* $f \in \bar{S}$, $d_S(f) = 4(3r+1)(2\ell+1)$.
- $S = \mathcal{F}_k$, $\bar{S} \subset \cup_{\ell=1}^{\infty} \mathcal{P}_{\ell,0}$ *and for all* $f \in \mathcal{P}_{\ell,0}$, $d_S(f) = 2(k+\ell+1)$.

Two additional models S satisfying this assumption may be found in [BB16]: the collection of non-negative piecewise concave-convex functions, and the collection of non-negative log-concave functions.

When Assumption 1 is met for a model S, it also holds for any model $S' \subset S$ with $\bar{S}' \subset \bar{S}$. As an illustration, consider a (non-empty) collection m of disjoint intervals of $\mathbb{R}$ that are right-closed and not-reduced to a singleton. Let $\mathcal{P}_{cont}(m)$ be the model of non-negative continuous piecewise affine functions defined by

$$\mathcal{P}_{cont}(m) = \left\{ \sum_{K\in m} P_K 1_K,\ P_K \in \mathcal{S} \text{ is a non-negative affine function such that } x \mapsto \sum_{K\in m} P_K(x) 1_K(x) \text{ is continuous on } \mathbb{R} \right\}.$$

As $\mathcal{P}_{cont}(m) \subset \mathcal{P}_{|m|,1}$, Assumption 1 is fulfilled with $S = \mathcal{P}_{cont}(m)$, $\bar{S} \subset \mathcal{P}_{|m|,1}$, and $d_S = 3(2|m|+1)$. Note that Theorem 1 applies as $\mathcal{P}_{cont}(m)$ is convex and shows that ρ-estimators

on $\mathcal{P}_{cont}(m)$ coincide with maximizers of $L_{\mathcal{X}}$ with $\mathcal{X} = \bigcup_{K\in m} K$ (to see that $\sup_{g\in S} L_{\mathcal{X}}(g)$ is finite, we refer to Lemma 4 in Section 4.2).

Remark: when S corresponds to one of the models defined above, we may always choose the ρ-estimator in a measurable way. For instance, if $S = \mathcal{P}_{\ell,0}$, we may apply the end of Section 2.3 with the countable set $\bar{S}_{count} = \overline{\mathcal{P}_{\ell,0}}_{count}$ defined by

$$\overline{\mathcal{P}_{\ell,0}}_{count} = \left\{ \sum_{j=1}^{\ell} a_j 1_{K_j},\ (a_j)_{1\leq j\leq \ell} \in \mathbb{Q}_+^{\ell}, K_j \text{ is an interval of finite length with endpoints in } \mathbb{Q} \right\}.$$

If S is defined by (15), we may use $\bar{S}_{count} = \bigcup_{\ell=1}^{\infty} \left(\mathcal{F}_k \cap \overline{\mathcal{P}_{\ell,0}}_{count}\right)$.

3.3. Main theorem. We now state our main result.

Theorem 4. *Let S be a model such that Assumption 1 is satisfied with $\bar{S} \subset S$. Then, any ρ-estimator $\hat{s}$ built on S satisfies for all $\xi > 0$,*

$$\mathbb{P}^{\star}\left[h^2(s,\hat{s}) \geq \inf_{f\in\bar{S}} \left\{ c_1 h^2(s,f) + c_2 \frac{d_S(f)}{n} \log_+^2\left(\frac{n}{d_S(f)}\right) + c_3 \xi \log_+(1/\xi) \right\}\right] \leq e^{-n\xi}, \tag{16}$$

and thus

$$\mathbb{E}^{\star}\left[h^2(s,\hat{s})\right] \leq \inf_{f\in\bar{S}} \left\{ c_1 \mathbb{E}\left[h^2(s,f)\right] + c_2' \frac{d_S(f)}{n} \log_+^2\left(\frac{n}{d_S(f)}\right) \right\}. \tag{17}$$

In the above inequalities, c_1, c_2, c_2', c_3 are universal positive constants.

Remark 1. The only risk bounds we are aware of that are written in terms of Hellinger distance h in frameworks 2 and 3 are those of [BB09] for a piecewise constant estimator whose partition is selected from the data. We refer to their Theorem 5, Proposition 8 and 9. They estimate s on a (deterministic) interval I and assume that $\int_I s(t)\,\mathrm{d}t$ is finite. This integrate deteriorates their risk bound and thus the theoretical performances of their estimator. Moreover, in hazard rate estimation, $\int_0^{\infty} s(t)\,\mathrm{d}t = \infty$, which implies that their estimation interval I must be of finite length, while our theorem makes it possible the estimation of s on the whole line $\mathbb{R}$.

It is very common in the literature to restrict the estimation of a hazard rate to a deterministic interval I of finite length on which the survival function $t \mapsto \mathbb{P}(X \geq t)$ is lower bounded by a positive constant (see [BC05, BC08, Pla09, AD10]). Note that the interval I may depend on the data in practice. Moreover, the lower bound on the survival function may influence the rates of convergence of the estimators (when it decreases with n) as soon as the loss does not favour the regions of I where the estimation is easier.

Remark 2. When $s \in \bar{S}$, the risk of the ρ-estimator built on S is bounded by

$$\mathbb{E}^{\star}\left[h^2(s,\hat{s})\right] \leq C \frac{d_S(s)}{n} \log_+^2\left(\frac{n}{d_S(s)}\right).$$

The rate of estimation of s becomes then parametric (up to a logarithmic term). When $s \notin \bar{S}$, the estimator automatically achieves the best trade-off between the bias (approximation) term

$h^2(s,f)$ and the variance (complexity) term $(d_S(f)/n)\log_+^2(n/d_S(f))$:

$$\mathbb{E}^\star\left[h^2(s,\hat{s})\right] \le CR(s) \quad \text{with } R(s) = \inf_{f\in\bar{S}}\left\{\mathbb{E}\left[h^2(s,f)\right] + \frac{d_S(f)}{n}\log_+^2\left(\frac{n}{d_S(f)}\right)\right\}. \tag{18}$$

It remains to compute $R(s)$ to deduce (an upper bound of) the rate of convergence of the ρ-estimator when $s \in S$. This rate is often non-parametric, and may therefore be much slower than the rate we would obtain if s does belong to $\bar{S}$ (see Section 3.5 for an example).

This phenomenon (faster rate of convergence when $s \in \bar{S}$) has already been put forward in [BB16] for ρ-estimators in density estimation (see their Theorem 2) and has been named superminimaxity. Our theorem improves their result in the sense that our approximation term involves a smaller exponent in the logarithmic term. Moreover, we show that superminimaxity is not specific to density estimation but also occurs in frameworks 2 and 3.

Remark 3. The right-hand side of (17) is of the same form in each framework. The only difference lies in the Hellinger loss h. Note that in frameworks 2 and 3, $\mathbb{E}\left[h^2(s,f)\right]$ is smaller than the (square of the) Hellinger distance in density estimation:

$$\mathbb{E}\left[h^2(s,f)\right] \le \frac{1}{2}\int_{\mathbb{R}}\left(\sqrt{s}-\sqrt{f}\right)^2 \mathrm{d}\mu.$$

This means that it suffices to bound the right-hand side of (17) in density estimation to automatically derive a bound on $\mathbb{E}^\star\left[h^2(s,\hat{s})\right]$ in frameworks 2 and 3.

Remark 4. It follows from a crude application of the triangular inequality and from (17) that

$$\mathbb{E}^\star\left[h^2(s,\hat{s})\right] \le C\inf_{g\in S}\left\{\mathbb{E}\left[h^2(s,g)\right] + R(g)\right\}.$$

If we know how to bound $R(g)$ for $g \in S$, this inequality says that the risk of the ρ-estimator is not only controlled when s does belong to S but also when there exists $g \in S$ such that $s \simeq g$, that is when s is close to S. In other words, the Hellinger risk of the ρ-estimator cannot substantially increase when s does not belong to S but is close to S. Such a result may be interpreted as a robustness property.

Remark 5. When the assumptions of Theorem 1 are met with $\mathcal{X} = \mathbb{R}$, a maximum likelihood estimator is a ρ-estimator. Its Hellinger risk is therefore bounded by (16) and (17). In particular, the maximum likelihood estimator inherits the robustness property described in the preceding remark. It is worth mentioning that such a result is not true for more general models S, as shown by the following toy example in density estimation borrowed from Section 2.3 of [Bir06]. Let $S = \{\theta^{-1}1_{[0,\theta]}, \theta > 0\}$. Then, Assumption 1 holds with $d_S(f) = 1$, $\bar{S} = S$, and the ρ-estimator $\hat{s}$ built on S is measurable and satisfies

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

We believe that the $\log^2 n/n$ term is suboptimal. However, if

$$s = (1-2n^{-1})1_{[0,1/10]} + 2n^{-1}1_{[9/10,1]},$$

then $h^2(s,S) \le 5/(4n)$ for $n \ge 4$ and thus $\mathbb{E}\left[h^2(s,\hat{s})\right] \le C'\log^2 n/n$ for some constant C'. Now, if $\tilde{s}$ designs the maximum likelihood estimator, $\mathbb{E}\left[h^2(s,\tilde{s})\right] > 0.38$, which does not tend to 0 when n goes to infinity (see [Bir06] for the computations leading to the upper bound of $h^2(s,S)$ and the lower bound on $\mathbb{E}\left[h^2(s,\tilde{s})\right]$). We may also refer to [Sar16] for numerical simulations

highlighting the interest of such a robustness property for an estimation procedure closely related to the present one.

Remark 6. When the criterion γ vanishes at a point $\hat{s}$, which typically happens when $\hat{s}$ maximizes $L_{\mathcal{X}}$, the constant c_1 appearing in front of the bias term $h^2(s,f)$ in Theorem 4 can be improved:

Proposition 5. *Let S be a model such that Assumption 1 is satisfied with $\bar{S} \subset S$. Suppose that there exists $\hat{s} \in S$ such that $\gamma(\hat{s}) = 0$.*

Then, (16) and (17) hold for all $\varepsilon > 0$ with $c_1 = c_{1,\varepsilon}$, $c_2 = c_{2,\varepsilon}$, $c_2' = c_{2,\varepsilon}'$ such that $c_{1,\varepsilon} \geq 9$ and $\lim_{\varepsilon \to 0} c_{1,\varepsilon} = 9$.

The constant $c_1 = c_{1,\varepsilon}$ may therefore be made as close as 9 as wished. We do not know to what extent this result can be improved. We only know that $c_1 = c_{1,\varepsilon}$ cannot be smaller than 2 as shown by the following elementary example.

Proposition 6. *Consider framework 1, $p \in (0,1)$, $\varepsilon \in (0,1)$ and observations $X_1, \dots, X_n$ with density*

$$s := s_{p,\varepsilon} = p\varepsilon^{-1} 1_{[0,\varepsilon]} + (1-p) 1_{[1,2]}.$$

Let $m = \{[0,1],[1,2]\}$, and S be the model of piecewise constant densities defined by

$$S = \left\{ a 1_{[0,1]} + (1-a) 1_{[1,2]},\, a \in [0,1] \right\}. \tag{19}$$

Since S is convex, the ρ-estimator $\hat{s}$ coincides with the maximum likelihood estimator defined by

$$\hat{s} = \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[0,1]}(X_i) \right) 1_{[0,1]} + \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[1,2]}(X_i) \right) 1_{[1,2]},$$

and vanishes γ: $\gamma(\hat{s}) = 0$. Moreover, for all $\eta \in (1,2)$, there exist p, ε such that

$$\lim_{n \to +\infty} h^2(s_{p,\varepsilon}, \hat{s}) \geq \eta h^2(s_{p,\varepsilon}, S) \quad \textit{almost surely.}$$

Note that the constant 2 is optimal for the particular model (19) as it follows from (2.8) of [BR06] that

$$\mathbb{E}\left[h^2(s_{p,\varepsilon}, \hat{s})\right] \leq h^2(s_{p,\varepsilon}, \bar{s}) + \frac{1}{2n} \quad \text{with } \bar{s} = \left(\int_{[0,1]} s_{p,\varepsilon} \, \mathrm{d}\mu \right) 1_{[0,1]} + \left(\int_{[1,2]} s_{p,\varepsilon} \, \mathrm{d}\mu \right) 1_{[1,2]},$$

and Lemma 2 of [BB09] asserts that $h^2(s_{p,\varepsilon}, \bar{s}) \leq 2h^2(s_{p,\varepsilon}, S)$. However, we do not know whether this constant is optimal for more general convex models S.

Remark 7. In framework 1, we do not suppose that S consists of densities. This gives to ρ-estimators an additional property as we briefly mentioned in Section 2.4. Indeed, consider a class $\mathcal{F}$ of densities, an interval I of $\mathbb{R}$, and the model $S = \{f 1_I,\, f \in \mathcal{F}\}$. We define the random measures N' and M' by $N'(A) = N(A \cap I)$, $M'(A) = M(A \cap I)$ for all $A \in \mathcal{B}(\mathbb{R})$. Note that $\mathbb{E}[N'(A)] = \int_A s 1_I \, \mathrm{d}M'$. Since the functions of S vanish outside I, we may replace N, M in the procedure of Section 2.3 by N', M'. This would not change the estimator, or the empirical process to control. In particular, if Assumption 1 is fulfilled with $\bar{S} \subset S$, the ρ-estimator satisfies

for all $\xi > 0$,

$$\mathbb{P}^\star\left[h^2(s1_I, \hat{s}1_I) \geq \inf_{f\in\bar{S}}\left\{c_1 h^2(s1_I, f1_I) + c_2\frac{d_S(f)}{n}\log_+^2\left(\frac{n}{d_S(f)}\right) + c_3\xi\log_+(1/\xi)\right\}\right] \leq e^{-n\xi}.$$

Thereby, when all functions of S vanish outside a common interval I, the ρ-estimator actually estimates the restriction of s to I (that may not be a density).

3.4. A first illustration of Theorem 4. Let $\mathcal{P}_{\ell,0}$ be the collection of step functions defined by (12). Then, the ρ-estimator $\hat{s}$ built on $\mathcal{P}_{\ell,0}$ satisfies in the three statistical settings

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C\left\{\mathbb{E}\left[h^2(s,\mathcal{P}_{\ell,0})\right] + \frac{\ell}{n}\log_+^2\left(\frac{n}{\ell}\right)\right\}. \tag{20}$$

The first term $\mathbb{E}\left[h^2(s,\mathcal{P}_{\ell,0})\right]$ can be interpreted as an approximation term that is small if s is close to a step function of $\mathcal{P}_{\ell,0}$. When s does belong to $\mathcal{P}_{\ell,0}$, the bound becomes

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

This result is, in general, slightly suboptimal as it is possible to do better in density estimation. Indeed, $\mathcal{P}_{\ell,0}$ being VC subgraph (of dimension proportional to ℓ), we derive from Theorem 12 of [BBS17],

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C'\left\{h^2(s,\mathcal{P}_{\ell,0}) + \frac{\ell}{n}\log_+\left(\frac{n}{\ell}\right)\right\},$$

where C' is universal. The logarithmic term in this inequality is mandatory, in view of results on minimax lower bounds (Proposition 2 of [BM98]). We conjecture that it is possible to improve the exponent in the logarithmic term in (20) in frameworks 2 and 3.

3.5. Risks of ρ-estimators for piecewise monotone functions. Let $(\mathbb{L}^2(\mathbb{R}, M_E), d_2)$ be the metric space of square integrable functions on $\mathbb{R}$ with respect to the measure M_E defined by $M_E(A) = \mathbb{E}[M(A)]$ for all $A \in \mathcal{B}(\mathbb{R})$. Let S be a model satisfying the assumptions of Theorem 4. We know from (18) that the ρ-estimator $\hat{s}$ satisfies

$$\mathbb{E}^\star\left[h^2(s,\hat{s})\right] \leq CR(s) \quad \text{with } R(s) = \inf_{f\in\bar{S}}\left\{\frac{1}{2}d_2^2(\sqrt{s},f) + \frac{d_S(f)}{n}\log_+^2\left(\frac{n}{d_S(f)}\right)\right\}, \tag{21}$$

where C is universal. It then remains to bound above $R(s)$ to control the risk of $\hat{s}$.

As pointed out by Remark 2 of Section 3.3, a similar result to our Theorem 4 has already been established by [BB16] in density estimation. In this framework, the only difference between our inequality (21) and their risk bound lies in the exponent of the logarithmic term. Thereby, many bounds on $R(s)$ can be deduced from their results. Dealing with the other statistical settings requires now little supplementary work. To avoid redundancy and to keep this paper a reasonable size, we restrict ourselves to one example of model S.

Let $S = \mathcal{F}_k$ be the collection of piecewise monotone functions on at most k pieces defined by (15). It follows from Proposition 3 that

$$R(s) \leq C' \inf_{\ell\geq 1}\left[\inf_{f\in\mathcal{F}_k\cap\mathcal{P}_{\ell,0}}\left\{\frac{1}{2}d_2^2(\sqrt{s},f) + \frac{k+\ell}{n}\log_+^2\left(\frac{n}{k+\ell}\right)\right\}\right],$$

where C' is universal.

We now need to introduce some notations. We define for any interval K and function f,

$$V_K(f) = \sup_{x\in K} f(x) - \inf_{x\in K} f(x).$$

Let $\mathcal{M}_k$ be the family gathering all the collections m of at most k disjoint intervals of $\mathbb{R}$. Define for $m \in \mathcal{M}_k$,

$$\mathcal{F}(m) = \left\{ \sum_{K\in m} f_K 1_K, \text{ where } f_K \in \mathcal{S} \text{ is monotone on } K \right\}.$$

Note that $\mathcal{F}_k = \bigcup_{m\in\mathcal{M}_k} \mathcal{F}(m)$. We define for $m \in \mathcal{M}_k$ and $f \in \mathcal{F}(m)$ of the form $f = \sum_{K\in m} f_K 1_K$,

$$L_m(f) = \sum_{K\in m} \left[M_E(K) V_K^2 \left(f_K\right) \right]^{1/3}.$$

In this equality, we use the convention $+\infty \times 0 = 0$ when $M_E(K) = +\infty$. For $f \in \mathcal{F}_k$, we set

$$L(f) = \inf_{m\in\mathcal{M}_k} L_m(f).$$

The result is the following.

Corollary 2. *There exists a measurable ρ-estimator on $\mathcal{F}_k$, ($k \geq 1$) and any measurable ρ-estimator $\hat{s}$ on $\mathcal{F}_k$ satisfies*

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C \inf_{f\in\mathcal{F}_k} \left\{ d_2^2\left(\sqrt{s}, f\right) + L(f) \left(\frac{\log^2 n}{n}\right)^{2/3} + \frac{k\log^2 n}{n} \right\}. \tag{22}$$

In particular, if s does belong to $\mathcal{F}_k$, then $f = \sqrt{s}$ also belongs to $\mathcal{F}_k$ and hence,

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C \left[L\left(\sqrt{s}\right) \left(\frac{\log^2 n}{n}\right)^{2/3} + \frac{k\log^2 n}{n} \right]. \tag{23}$$

In the preceding inequalities, C is a universal constant.

We now make more explicit the above inequality (23) when $k = 2$ and s is unimodal. We distinguish the cases according to the different frameworks.

Consider framework 1 or 3. Then, $M_E(K)$ is equal or smaller than the Lebesgue measure $\mu(K)$ of K. Therefore, if s has support included into an interval of length L_{supp},

$$L(\sqrt{s}) \leq 2 L_{supp}^{1/3} \|s\|_\infty^{1/3},$$

where $\|s\|_\infty = \sup_{x\in\mathbb{R}} s(x)$, and the bound (23) becomes

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C' \left[L_{supp}^{1/3} \|s\|_\infty^{1/3} \left(\frac{\log^2 n}{n}\right)^{2/3} + \frac{\log^2 n}{n} \right].$$

In framework 2, suppose that X has finite expectation: $\mathbb{E}[X] < \infty$. Then, for all interval $K \subset [0, +\infty)$,

$$\begin{aligned} M_E(K) &\leq \int_0^\infty \mathbb{P}\left(X \geq t\right) \mathrm{d}t \\ &\leq \mathbb{E}[X]. \end{aligned}$$

Therefore,

$$L(\sqrt{s}) \leq 2\left(\mathbb{E}(X)\right)^{1/3} \|s\|_\infty^{1/3}.$$

The risk of the ρ-estimator $\hat{s}$ is then bounded above by

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C'\left[\left(\mathbb{E}(X)\right)^{1/3} \|s\|_\infty^{1/3} \left(\frac{\log^2 n}{n}\right)^{2/3} + \frac{\log^2 n}{n}\right].$$

Note that we do not require that the support of s be of finite length.

4. Selecting among estimators

It is often difficult in practice to find a global minimum of γ and thus to build ρ-estimators on non-convex models S. In this section, we propose new criteria based both on estimator selection and $T(f,g)$, that are more numerically friendly.

4.1. Random models. Let $\hat{S}$ be a possibly random model, that is a model that may depend on the data. Mathematically, this means that $\hat{S}$ maps Ω to the set of subsets of $\mathcal{S}$. We may build a ρ-estimator on $\hat{S}$ in the same way as we would do if the model were deterministic. More precisely, we define $\gamma_2(f)$ for $f \in \hat{S}$ by

$$\gamma_2(f) = \sup_{g \in \hat{S}} T(f,g),$$

where T is given by (4). We say that $\hat{s}$ is a ρ-estimator built on the random model $\hat{S}$ if it satisfies

$$\gamma_2(\hat{s}) \leq \inf_{g \in \hat{S}} \gamma_2(g) + 1/n.$$

The following theorem generalizes Theorem 4 to random models $\hat{S}$.

Theorem 7. *Suppose that Assumption 1 is satisfied with $S = \hat{S}$ and a possibly random subset $\bar{S} = \bar{\hat{S}} \subset \mathcal{S}$. For all $\xi > 0$, the ρ-estimator $\hat{s}$ built on the random model $\hat{S}$ satisfies*

$$\mathbb{P}^\star\left[h^2(s,\hat{s}) \geq c \inf_{f \in \hat{S} \cap \bar{\hat{S}}} \left\{h^2(s,f) + \frac{d_{\hat{S}}(f)}{n} \log_+^2\left(\frac{n}{d_{\hat{S}}(f)}\right) + \xi \log_+\left(1/\xi\right)\right\}\right] \leq e^{-n\xi}.$$

In particular,

$$\mathbb{E}^\star\left[h^2(s,\hat{s})\right] \leq C\mathbb{E}^\star\left[\inf_{f \in \hat{S} \cap \bar{\hat{S}}} \left\{h^2(s,f) + \frac{d_{\hat{S}}(f)}{n} \log_+^2\left(\frac{n}{d_{\hat{S}}(f)}\right)\right\}\right].$$

In the above inequalities, c and C are universal constants.

We may use random models to address the problem of estimator selection as explained below. Let Λ be an at most countable set, and let for each $\lambda \in \Lambda$, $\hat{s}_\lambda$ be an estimator of s. Building a ρ-estimator on the model $\hat{S} = \{\hat{s}_\lambda, \lambda \in \Lambda\}$ yields an estimator of the form $\hat{s} = \hat{s}_{\hat{\lambda}}$, that is a particular estimator of the family. The risk of this selected estimator $\hat{s}_{\hat{\lambda}}$ is then bounded above by the following corollary, which immediately ensues from Theorem 7.

Corollary 3. *Let Λ be an at most countable set and let $\{\hat{s}_\lambda, \lambda \in \Lambda\}$ be a collection of estimators. Suppose that there exists a deterministic model S satisfying Assumption 1 such that $\bar{S} \subset S$, and such that each estimator $\hat{s}_\lambda$ has values in $\bar{S}$. Building a ρ-estimator on the random model $\hat{S} = \{\hat{s}_\lambda, \lambda \in \Lambda\}$ amounts to selecting an estimator among $\{\hat{s}_\lambda, \lambda \in \Lambda\}$: the ρ-estimator is of the form $\hat{s} = \hat{s}_{\hat{\lambda}}$ and satisfies for all $\xi > 0$,*

$$\mathbb{P}\left[h^2(s, \hat{s}_{\hat{\lambda}}) \geq c \inf_{\lambda \in \Lambda}\left\{h^2(s, \hat{s}_\lambda) + \frac{d_S(\hat{s}_\lambda)}{n}\log_+^2\left(\frac{n}{d_S(\hat{s}_\lambda)}\right) + \xi \log_+(1/\xi)\right\}\right] \leq e^{-n\xi}.$$

In particular,

$$\begin{aligned}\mathbb{E}\left[h^2(s, \hat{s}_{\hat{\lambda}})\right] &\leq C\mathbb{E}\left[\inf_{\lambda \in \Lambda}\left\{h^2(s, \hat{s}_\lambda) + \frac{d_S(\hat{s}_\lambda)}{n}\log_+^2\left(\frac{n}{d_S(\hat{s}_\lambda)}\right)\right\}\right] \\ &\leq C \inf_{\lambda \in \Lambda}\mathbb{E}\left[h^2(s, \hat{s}_\lambda) + \frac{d_S(\hat{s}_\lambda)}{n}\log_+^2\left(\frac{n}{d_S(\hat{s}_\lambda)}\right)\right].\end{aligned} \tag{24}$$

In the above inequalities, c and C are universal constants.

We now give an illustrative example. We consider $\ell \geq 1$, $r \geq 0$ and an at most countable collection $\{\hat{s}_\lambda,\ \lambda \in \Lambda\}$ of non-negative piecewise polynomial estimators of degree at most r based on at most ℓ pieces. In other words, each $\hat{s}_\lambda$ has values into $\mathcal{P}_{\ell,r}$. The model $S = \mathcal{P}_{\ell,r}$ satisfies Assumption 1 with $\bar{S} = S$ and $d_S = (r+2)(2\ell+1)$ (see the first point of Proposition 3). Therefore, the selected estimator $\hat{s}_{\hat{\lambda}}$ satisfies

$$\mathbb{E}\left[h^2(s, \hat{s}_{\hat{\lambda}})\right] \leq C'\left\{\inf_{\lambda \in \Lambda}\mathbb{E}\left[h^2(s, \hat{s}_\lambda)\right] + \frac{(r+1)\ell \log_+^2\left(n/(\ell(r+1))\right)}{n}\right\}, \tag{25}$$

for some universal constant C'. The risk of the selected estimator $\hat{s}_{\hat{\lambda}}$ is therefore bounded above, up to the multiplicative constant C' and an estimation term of the order of $(r+1)\ell\log_+^2\left(n/(\ell(r+1))\right)/n$, by the risk of the best estimator of the family.

This risk bound is always worse than the one we would obtain for a ρ-estimator $\hat{s}$ built on the model $S = \mathcal{P}_{\ell,r}$:

$$\mathbb{E}\left[h^2(s, \hat{s})\right] \leq C''\left\{\mathbb{E}\left[h^2(s, \mathcal{P}_{\ell,r})\right] + \frac{(r+1)\ell \log_+^2\left(n/(\ell(r+1))\right)}{n}\right\}, \tag{26}$$

where C'' is universal (see Theorem 4). The interest of $\hat{s}_{\hat{\lambda}}$ is practical: the construction of $\hat{s}$ seems to be numerically difficult whereas the selected estimator $\hat{s}_{\hat{\lambda}}$ can be computed in a reasonable amount of time as soon as Λ is finite and not too large.

4.2. Selecting among a special collection of piecewise polynomial estimators. As we see in (25), we should take Λ as large as possible to improve on the theoretical performances of the selected estimator. In this section, we propose to define a very large collection of piecewise polynomial ρ-estimators $\{\hat{s}_\lambda,\ \lambda \in \Lambda\}$. Despite the large cardinality of Λ, we shall explain that it is possible to tackle the problem of estimator selection in practice.

We need to introduce the following notations. We consider a (possibly random) subset $\hat{I} \subset \{1,\dots,n\}$ and real-valued random variables $(Y_i)_{i\in\hat{I}}$ so that the measure N writes

$$N(A) = \frac{1}{n}\sum_{i\in\hat{I}} 1_A(Y_i) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

More precisely, we define 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, $\hat{I} = \{i \in \{1,\dots,n\},\, T_{1,0}^{(i)} < \infty\}$, $Y_i = T_{1,0}^{(i)}$. Note that the random variables $(Y_i)_{i\in\hat{I}}$ are distinct almost surely, which enables us to order them: $Y_{(1)} < Y_{(2)} < \cdots < Y_{(\hat{n})}$ where $\hat{n} = |\hat{I}| \le n$.

Let now $\mathcal{M}$ be the class of finite (non-empty) collections m of disjoint intervals K of $\mathbb{R}$ that are right-closed and not reduced to a singleton. Let $r \ge 0$, $m \in \mathcal{M}$ and

$$\mathcal{P}_r(m) = \Bigg\{ \sum_{K\in m} f_K 1_K, \text{ for all } K \in m,\ f_K 1_K \in \mathcal{S}, \\ \text{and } f_K \text{ is a polynomial function of degree at most } r \Bigg\}.$$

We may compute a piecewise polynomial ρ-estimator $\hat{s}_m$ on the convex model $\mathcal{P}_r(m)$:

Lemma 4. *Let $m \in \mathcal{M}$ and for $K \in m$,*

$$\mathcal{P}_r(K) = \{f 1_K,\ f \text{ is a polynomial function of degree at most } r \text{ such that } f 1_K \in \mathcal{S}\}.$$

Then, $\sup_{f\in\mathcal{P}_r(K)} L_K(f)$ is finite and achieved at a point $\hat{s}_K$. Moreover, $\hat{s}_m = \sum_{K\in m} \hat{s}_K$ maximizes $L_{\mathcal{X}}$ over $\mathcal{P}_r(m)$ where $\mathcal{X} = \bigcup_{K\in m} K$. It is a ρ-estimator on the model $S = \mathcal{P}_r(m)$ that vanishes γ.

When $\bigcup_{K\in m} K = \mathbb{R}$, or more precisely when all the $(Y_i)_{i\in\hat{I}}$ lie in $\bigcup_{K\in m} K$, $\hat{s}_m$ is also the maximum likelihood estimator on the model $\mathcal{P}_r(m)$.

We define $\widehat{\mathcal{M}}_0 = \{\emptyset\}$ and $\hat{s}_\emptyset = 0$. When $\hat{n} \ge 2$, we define the collection $\widehat{\mathcal{M}}$ that gathers all the partitions m of $[Y_{(1)}, Y_{(\hat{n})}]$ of the form

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

where $k \ge 0$ 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 $\ell \in \{1,\dots,\hat{n}-1\}$,

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

We consider a random variable $\hat{\ell}$ with values in $\{0,\dots,\max\{\hat{n}-1, 0\}\}$. The aim of this section is to explain how we can select, in practice, an estimator among the family $\{\hat{s}_m,\ m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$. Note that we must suppose in general that $\hat{\ell}$ is random since $\hat{n}$ is also random in frameworks 2 and 3.

The simplest solution would be to use the selection rule described in Section 4.1. Unfortunately, the numerical complexity of this procedure depends heavily on $|\widehat{\mathcal{M}}_{\hat{\ell}}|$, which is usually very large and makes the procedure numerically intractable in practice. We propose in this section an alternative way that improves on its numerical cost.

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

$$m_K \vee K = \left\{K' \cap K,\, K' \in m_K,\, K' \cap K \neq \emptyset\right\}. \tag{27}$$

We now consider some positive real number L and define the criterion γ_3 for $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ by

$$\gamma_3(\hat{s}_m) = \sum_{K \in m} \sup_{m' \in \widehat{\mathcal{M}}_{\hat{\ell}}} \left\{ T\left(\hat{s}_m 1_K, \hat{s}_{m'} 1_K\right) - L(r+1)\frac{|m' \vee K| \log_+^2\left(n/(r+1)\right)}{n} \right\}. \tag{28}$$

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 γ_3:

$$\gamma_3(\hat{s}_{\hat{m}}) = \min_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \gamma_3(\hat{s}_m). \tag{29}$$

Note that the above minimum is achieved since $\widehat{\mathcal{M}}_{\hat{\ell}}$ is finite.

Theorem 8. *There exists a universal constant L_0 such that if $L \geq L_0$, any estimator $\hat{s}_{\hat{m}}$ minimizing (29) satisfies for all $\xi > 0$, and probability larger than $1 - e^{-n\xi}$,*

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

In particular,

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

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

This inequality (31) is akin to the one (25) obtained for the first selection rule (we only slightly loose on the variance term). The procedure described in this section is definitely more complex, but has a decisive advantage: it can be implemented in practice thanks to dynamic programming (at least when $\hat{n}$ is not too large, since the numerical complexity on the algorithm significantly increases with $\hat{n}$). For more informations on this algorithm, we refer to [Kan92] and Section 4.2 of [CR04] (see also the Appendix of [Sar14] for a quite similar optimization problem).

When $|\hat{I}| \leq 1$, which may occur in frameworks 2 and 3, then $\hat{\ell} = 0$, $\hat{s}_{\hat{m}} = \hat{s}_\emptyset = 0$, and $\widehat{\mathcal{M}}_0 = \{\emptyset\}$. The probability of this event is $\alpha^{n-1}(n - n\alpha + \alpha)$ where $\alpha = \mathbb{P}(T \geq C)$ in framework 2 and $\alpha = \mathbb{P}(T_{1,0} = +\infty)$ in framework 3. It is likely small (unless that T, C, or $T_{1,0}$ depend on n). However, if this event realizes, inequality (30) becomes straightforward and useless. Actually, in that case, $T(0, f) \leq 1/n$ for all $f \in \mathcal{S}$. Thereby, although that $\hat{s}_{\hat{m}} = 0$ looks very poor, it is the ρ-estimator on any model S. In particular, a quick glance at the proof of Theorem 4 shows that for all $\xi > 0$, there exists an event which holds true with probability larger than $1 - e^{-n\xi}$ and on which: for all $\ell \geq 1$,

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

where C is universal.

When $\hat{n}$ is too large, the procedure becomes unfortunately numerically intractable and we do not know how to select an estimator among $\{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$ in a reasonable amount of

time. However, we may find a way to cope if we reduce the collection $\widehat{\mathcal{M}}_{\hat{\ell}}$ as follows. Let $\mathcal{G} \subset \{2, \dots, \hat{n}-1\}$ such that $|\mathcal{G}| \geq \hat{\ell} - 1$ and

$$\mathcal{N}_{\mathcal{G},\hat{\ell}} = \left\{ (1, n_1, \dots, n_{\hat{\ell}-1}, \hat{n}),\, (n_1, \dots, n_{\hat{\ell}-1}) \in \mathcal{G}^{\hat{\ell}-1}, n_1 < n_2 < \cdots < n_{\hat{\ell}-1} \right\}.$$

We then define $\widehat{\mathcal{M}}_{\mathcal{G},0} = \{\emptyset\}$ and the collection $\widehat{\mathcal{M}}_{\mathcal{G},\hat{\ell}}$ when $\hat{\ell} \in \{1, \dots, \hat{n}\}$ of partitions m of the form

$$m = \left\{ [Y_{(1)}, Y_{(n_1)}], (Y_{(n_1)}, Y_{(n_2)}], (Y_{(n_2)}, Y_{(n_3)}], \dots, (Y_{(n_{\hat{\ell}-1})}, Y_{(\hat{n})}] \right\},$$

where $(1, n_1, n_2, \dots, n_{\hat{\ell}-1}, \hat{n})$ belongs to $\mathcal{N}_{\mathcal{G},\hat{\ell}}$ with the convention that $m = [Y_{(1)}, Y_{(\hat{n})}]$ when $\hat{\ell} = 1$. We may then select an estimator among $\{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\mathcal{G},\hat{\ell}}\}$ by adapting the preceding procedure. More precisely, the selection rule is defined by (28) and (29) where $\widehat{\mathcal{M}}_{\hat{\ell}}$ is replaced by $\widehat{\mathcal{M}}_{\mathcal{G},\hat{\ell}}$. The performance of the resulting estimator is then given by (30) and (31), where $\widehat{\mathcal{M}}_{\hat{\ell}}$ is once more replaced by $\widehat{\mathcal{M}}_{\mathcal{G},\hat{\ell}}$. From a theoretical point of view, we should take $\mathcal{G}$ large, typically $\mathcal{G} = \{2, \dots, \hat{n}-1\}$. However, from a practical point of view, reducing the cardinality of the set $\mathcal{G}$ makes faster the construction of the estimator. The choice of $\mathcal{G}$ is therefore a compromise between the theoretical properties of the estimator and its time construction.

Since $\hat{s}_m$ is a maximum likelihood estimator, it is natural to compare our estimator $\hat{s}_{\hat{m}}$ to the one $\hat{s}_{\tilde{m}}$ that maximizes the log likelihood $L(\hat{s}_m)$ over $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$. Numerical simulations carried out in Section 5 suggest that our estimator may perform better (provided that L is suitably chosen, see the next section). We do not know theoretical results for the maximum likelihood estimator $\hat{s}_{\tilde{m}}$. However, if $\mathcal{M}_\ell$ is a deterministic collection of partitions m such that $|m| = \ell$, then results concerning the maximum likelihood estimator defined as a maximizer of $m \mapsto L(\hat{s}_m)$ over $m \in \mathcal{M}_\ell$ 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 of this estimator in density estimation. Note that they put some restriction either on s, or on the minimal length of the intervals K of the partitions $m \in \mathcal{M}_\ell$. Besides, contrary to ours, their upper-bounds involve the Kullback Leibler divergence.

4.3. Selecting among a special collection of piecewise polynomial estimators: a calibration free approach. Although the procedure described in the preceding section can be implemented in practice (if $\hat{n}$ is not too large), and that the estimator possesses nice statistical properties, it remains an important practical issue: the choice of L. This parameter, is, indeed, involved in the construction of the estimator and a bad choice of L may deteriorate its performances.

A simple solution to avoid this pitfall, is to proceed as follows. We consider a (non-empty, but at most countable) collection $\mathcal{L}$ of positive numbers. For each $L \in \mathcal{L}$, we may use the procedure described in the preceding section with the parameter L in (28) to select an estimator among the collection $\{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$. The selected estimator is written $\hat{s}_{\hat{m}_L}$ to emphasize that it depends on L. We now use the procedure of Section 4.1 to select an estimator among the collection $\{\hat{s}_{\hat{m}_L},\, L \in \mathcal{L}\}$.

We apply Theorem 7 with $\hat{S} = \bar{\hat{S}} = \{\hat{s}_{\hat{m}_L},\, L \in \mathcal{L}\}$. The resulting estimator $\hat{s} = \hat{s}_{\hat{m}_{\hat{L}}}$ satisfies for all $\xi > 0$ and probability larger than $1 - e^{-n\xi}$,

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

where C is a universal constant. If $\mathcal{L}$ contains at least one number L larger than L_0, we derive from (30),

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

where C' is a universal constant.

Interestingly, this estimator $\hat{s}$ does not depend on the particular choice of a calibration parameter L but rather on a collection $\mathcal{L}$. The larger $\mathcal{L}$, the better the risk bound. However, the numerical complexity of the whole procedure increases with the size of $\mathcal{L}$.

4.4. Building an adaptive piecewise polynomial estimator. In Section 4.2, our aim was to find a good partition m of size $\hat{\ell}$, that is a partition for which the piecewise polynomial estimator $\hat{s}_m$ of degree at most r based on m is close to the target unknown function s. The idea was to select an estimator among the collection $\{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$. Note that this estimator depends on the preliminary choice of $\hat{\ell}$. We explain in this section how to avoid this dependency, that is, how to choose $\hat{\ell}$ from the data.

We set $\widehat{\mathcal{M}}_{0,lower} = \emptyset$ and $\hat{s}_\emptyset = 0$. When $\hat{n} \geq 2$, we define for $k \in \{1,\dots,\hat{n}-1\}$ the collection $\widehat{\mathcal{M}}_{k,lower}$ of partitions $m \in \widehat{\mathcal{M}}$ whose cardinal is at most k:

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

We consider a random variable $\hat{k}$ with values in $\{0,\dots,\max\{\hat{n}-1,0\}\}$ and adapt the procedure of Section 4.2 to select an estimator among $\{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{k},lower}\}$.

We consider some $L > 0$ and set for $m \in \widehat{\mathcal{M}}_{\hat{k},lower}$,

$$\gamma_4(\hat{s}_m) = \sum_{K\in m}\sup_{m'\in\widehat{\mathcal{M}}_{\hat{k},lower}}\left\{T\left(\hat{s}_m 1_K, \hat{s}_{m'}1_K\right) - L(r+1)\frac{|m'\vee K|\log_+^2\left(n/(r+1)\right)}{n}\right\}.$$

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

$$\gamma_4(\hat{s}_{\hat{m}}) + 2L(r+1)\frac{|\hat{m}|\log_+^2\left(n/(r+1)\right)}{n} = \inf_{m\in\widehat{\mathcal{M}}_{\hat{k},lower}}\left\{\gamma_4(\hat{s}_m) + 2L(r+1)\frac{|m|\log_+^2\left(n/(r+1)\right)}{n}\right\}. \tag{32}$$

Theorem 9. *There exists a universal constant L_0 such that if $L \geq L_0$, any estimator $\hat{s}_{\hat{m}}$ satisfying (32) satisfies for all $\xi > 0$, and probability larger than $1 - e^{-n\xi}$,*

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

In particular,

$$\mathbb{E}\left[h^2(s, \hat{s}_{\hat{m}})\right] \leq C'\mathbb{E}\left[\inf_{m \in \widehat{\mathcal{M}}_{\hat{k},lower}} \left\{ h^2(s, \hat{s}_m) + L\frac{(r+1)|m| \log_+^2(n/(r+1))}{n} \right\} \right]. \tag{33}$$

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

Note that this risk bound improves when $\hat{k}$ grows up. We may even set $\hat{k} = \hat{n}$, in which case $\widehat{\mathcal{M}}_{\hat{k},lower} = \widehat{\mathcal{M}}$. This selection rule provides better theoretical results than the preceding one since (33) implies

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

The right-hand side of this inequality corresponds to the bound (31) achieved by the estimator of Section 4.2 when the choice of $\hat{\ell}$ is the best possible among $\{1, \dots, \hat{k}\}$. The main difference between the two procedures is that the present one does not depend on $\hat{\ell}$ any more (but only on $\hat{k}$). Moreover, it turns out that the right-hand side of (33) can be put in a more convenient form when $\hat{k} = \hat{n}$:

Lemma 5. *Let $\xi > 0$ and $L \geq 1$. There exists a universal constant C such that with probability larger than $1 - e^{-n\xi}$,*

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

In particular,

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

where C' is universal.

Therefore, when $\hat{k} = \hat{n}$, and $L \geq \max\{1, L_0\}$, the estimator $\hat{s}_{\hat{m}}$ satisfies

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

where C'' is a universal constant and where

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

This term $\mathcal{R}(\ell)$ can be interpreted as an upper-bound of the risk of the ρ-estimator built on $\mathcal{P}_{\ell,r}$, barely worse than the one given by Theorem 4 and that is written in (26).

Remark. As in Section 4.2, we could compare this estimator with the one that maximizes a penalized log-likelihood criterion of the form $m \mapsto L(\hat{s}_m) - \mathrm{pen}(m)$ over $m \in \widehat{\mathcal{M}}_{\hat{k},lower}$. We do not know theoretical results for this estimator, but refer to [RMG10] for a numerical study. It is, however, rather difficult to numerically compare this estimator with ours due to calibration issues in the penalties.

5. Numerical simulations

We consider framework 1, $r = 0$, $\ell \in \{1, \dots, n\}$, and the (random) collection $\widehat{\mathcal{M}}_\ell$ consisting in partitions of $[X_{(1)}, X_{(n)}]$ of size ℓ. For each $m \in \widehat{\mathcal{M}}_\ell$, we consider the usual histogram estimator defined by

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

Note that this estimator is the ρ-estimator and the maximum likelihood estimator on the random model $\mathcal{P}_0(m)$. We carry out in this section a numerical study to compare two selection rules described in Section 4.2.

- 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

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

- 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

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

  For each $L \in \mathcal{L}$, we use the procedure of Section 4.2 specified in (28) and (29) to get a partition $\hat{m}_L \in \widehat{\mathcal{M}}_\ell$. We then use the procedure of Section 4.1 to select an estimator among $\{s_{\hat{m}_L},\, L \in \mathcal{L}\}$ as explained in Section 4.3. 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 s:

Example 1. *s is the density of a Normal distribution*

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

Example 2. *s is the density of a log Normal distribution*

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

Example 3. *s is the density of an exponential distribution*

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

Example 4. *s is the density of a mixture of uniform distributions*

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

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

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

We estimate the probability that the two procedures yield the same estimator by

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

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. Risks 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, s does

	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. Risks for simulated data with $n = 100$, $N_{\rm rep} = 1000$.

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, s is not piecewise constant, and the superior 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 $\widehat{R}(2,\ell)$.

6. Proofs

6.1. Proof of Lemma 1. Let $\sqrt{q} = (\sqrt{f} + \sqrt{g})/2$. Then,

$$\frac{1}{2}\int_{\mathbb{R}} \frac{\sqrt{g}-\sqrt{f}}{\sqrt{q}} \left(\sqrt{s}-\sqrt{q}\right)^2 \mathrm{d}M = \frac{1}{2}\int_{\mathbb{R}} \frac{\sqrt{g}-\sqrt{f}}{\sqrt{q}} s \,\mathrm{d}M + \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{g}-\sqrt{f}\right)\sqrt{q}\,\mathrm{d}M - \int_{\mathbb{R}} \left(\sqrt{g}-\sqrt{f}\right)\sqrt{s}\,\mathrm{d}M.$$

Note that

$$h^2(s,g) - h^2(s,f) = \frac{1}{2}\int_{\mathbb{R}} (g-f)\,\mathrm{d}M + \int_{\mathbb{R}} \sqrt{s}\left(\sqrt{f}-\sqrt{g}\right)\mathrm{d}M.$$

Therefore,

$$\begin{aligned}\frac{1}{2}\int_{\mathbb{R}} \frac{\sqrt{g}-\sqrt{f}}{\sqrt{q}} \left(\sqrt{s}-\sqrt{q}\right)^2 \mathrm{d}M &= \frac{1}{2}\int_{\mathbb{R}} \frac{\sqrt{g}-\sqrt{f}}{\sqrt{q}} s \,\mathrm{d}M + \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{g}-\sqrt{f}\right)\sqrt{q}\,\mathrm{d}M \\ &\quad -\frac{1}{2}\int_{\mathbb{R}} (g-f)\,\mathrm{d}M + h^2(s,g) - h^2(s,f) \\ &= T_E(f,g) + h^2(s,g) - h^2(s,f). \qquad (34)\end{aligned}$$

Now,

$$\begin{aligned}\frac{1}{2}\int_{\mathbb{R}} \frac{\sqrt{g}-\sqrt{f}}{\sqrt{q}}\left(\sqrt{s}-\sqrt{q}\right)^2 \mathrm{d}M &= \int_{\mathbb{R}} \frac{\sqrt{g}-\sqrt{f}}{\sqrt{f}+\sqrt{g}}\left(\sqrt{s}-\frac{\sqrt{f}+\sqrt{g}}{2}\right)^2 \mathrm{d}M \\ &\leq \int_{\mathbb{R}} \left(\sqrt{s}-\frac{\sqrt{f}+\sqrt{g}}{2}\right)^2 \mathrm{d}M \\ &\leq \frac{1}{4}\int_{\mathbb{R}} \left(\left(\sqrt{s}-\sqrt{f}\right)+\left(\sqrt{s}-\sqrt{g}\right)\right)^2 \mathrm{d}M.\end{aligned}$$

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

$$\begin{aligned}\frac{1}{2}\int_{\mathbb{R}} \frac{\sqrt{g}-\sqrt{f}}{\sqrt{q}}\left(\sqrt{s}-\sqrt{q}\right)^2 \mathrm{d}M &\leq \frac{1+\alpha}{4}\int_{\mathbb{R}} \left(\sqrt{s}-\sqrt{f}\right)^2 \mathrm{d}M + \frac{1+\alpha^{-1}}{4}\int_{\mathbb{R}} \left(\sqrt{s}-\sqrt{g}\right)^2 \mathrm{d}M \\ &\leq \frac{1+\alpha}{2}h^2(s,f) + \frac{1+\alpha^{-1}}{2}h^2(s,g).\end{aligned}$$

We now plug this inequality into (34) to get

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

The right-hand side of (3) 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 f and g),

$$T_E(g,f) \leq 3h^2(s,g) - \frac{1}{3}h^2(s,f).$$

Yet, $T_E(f,g) = -T_E(g,f)$ and hence $T_E(f,g) \geq \frac{1}{3}h^2(s,f) - 3h^2(s,g)$ as wished. □.

6.2. Proof of Theorem 1. In each of the frameworks described in Section 1, the measure N can be put of the form $N(A) = n^{-1}\sum_{i\in\hat{I}} 1_A(Y_i)$ where $\hat{I} \subset \{1,\ldots,n\}$, and where Y_i are suitable real-valued random variables. For instance, in framework 1, $\hat{I} = \{1,\ldots,n\}$, $Y_i = X_i$, in framework 2, $\hat{I} = \{i \in \{1,\ldots,n\}, D_i = 1\}$, $Y_i = X_i$, and in framework 3, $\hat{I} = \{i \in \{1,\ldots,n\}, T_{1,0}^{(i)} < \infty\}$, $Y_i = T_{1,0}^{(i)}$.

Set $\hat{J} = \{i \in \hat{I},\, Y_i \in \mathcal{X}\}$. Then, for $f, g \in S$, $T(f,g)$ and $L_{\mathcal{X}}(f)$ take the form

$$\begin{aligned}T(f,g) &= \frac{1}{n}\sum_{j\in\hat{J}} \psi\left(\frac{g(Y_j)}{f(Y_j)}\right) - \frac{1}{4}\int_{\mathcal{X}} \left(g(x)-f(x)\right)\mathrm{d}M(x) \\ L_{\mathcal{X}}(f) &= \frac{1}{n}\sum_{j\in\hat{J}} \log f(Y_j) - \int_{\mathcal{X}} f(x)\,\mathrm{d}M(x).\end{aligned}$$

The proof is straightforward if $\hat{J} = \emptyset$ since then $4T(f,g) = L_{\mathcal{X}}(g) - L_{\mathcal{X}}(f)$ and $4\gamma(f) = \sup_{g\in S} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(f)$. We suppose from now on that $\hat{J} \neq \emptyset$.

Claim 1. *Let $\bar{S} = \{f \in S,\, L_{\mathcal{X}}(f) \neq -\infty\}$ and $\bar{f} \in \bar{S}$. Then, the signs of $\sup_{g\in\bar{S}} T(\bar{f},g)$ and $\sup_{g\in\bar{S}} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f})$ are equal.*

Proof. Let $\bar{S}_1 = \left\{ g \in \bar{S},\, g = \bar{f},\, N \text{ a.s} \right\}$ and $\bar{S}_2 = \bar{S} \setminus \bar{S}_1$.

When $g \in \bar{S}_1$,

$$\begin{aligned} T(\bar{f}, g) &= -\frac{1}{4} \int_{\mathcal{X}} \left(g(x) - \bar{f}(x) \right) \mathrm{d}M(x) \\ &= \frac{1}{4} \left(L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f}) \right). \end{aligned}$$

Therefore, $\sup_{g \in \bar{S}_1} T(\bar{f}, g)$ and $\sup_{g \in \bar{S}_1} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f})$ have same sign.

Suppose now that $\sup_{g \in \bar{S}_2} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f})$ is non-positive. Let $g \in \bar{S}_2$, $u \in [0,1]$ and $\zeta = g - \bar{f}$. Note that $\bar{f} + u\zeta = (1-u)\bar{f} + ug \in S$ and thus $L_{\mathcal{X}}(\bar{f} + u\zeta) - L_{\mathcal{X}}(\bar{f}) \leq 0$. Moreover, for all $j \in \hat{J}$, $\bar{f}(Y_j) + u\zeta(Y_j) \geq \min\{\bar{f}(Y_j), g(Y_j)\} > 0$. Thereby, $L_{\mathcal{X}}(\bar{f} + u\zeta) \neq -\infty$ and $\bar{f} + u\zeta \in \bar{S}$. Actually, $\bar{f} + u\zeta$ belongs to $\bar{S}_2$ as soon as $u \neq 0$.

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

$$\begin{aligned} \wp_1(u) &= L_{\mathcal{X}}(\bar{f} + u\zeta) - L_{\mathcal{X}}(\bar{f}) \\ &= \frac{1}{n} \sum_{j \in \hat{J}} \log \left(\frac{\bar{f}(Y_j) + u\zeta(Y_j)}{\bar{f}(Y_j)} \right) - u \int_{\mathcal{X}} \zeta(x) \, \mathrm{d}M(x). \end{aligned}$$

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

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

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

$$\begin{aligned} &\wp_1(0) = \wp_2(0) = 0 \\ &\wp_1'(0) = \wp_2'(0) = \frac{1}{n} \sum_{j \in \hat{J}} \frac{\zeta(Y_j)}{\bar{f}(Y_j)} - \int_{\mathcal{X}} \zeta(x) \, \mathrm{d}M(x) \\ &\wp_1''(0) = \wp_2''(0) = -\frac{1}{n} \sum_{j \in \hat{J}} \left(\frac{\zeta(Y_j)}{\bar{f}(Y_j)} \right)^2. \end{aligned}$$

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

We recall that $\wp_1(u)$ is non-positive for all $u \in [0,1]$ as previously explained. In particular, $\wp_1'(0) \leq 0$. The above computations show the existence of $u_1 \in (0,1]$ such that $\wp_2(u) \leq 0$ for all $u \in [0, u_1]$. Note that the function $u \mapsto \psi\left(1 + u\zeta(Y_j)/\bar{f}(Y_j)\right)$ for $j \in \hat{J}$ is concave, whatever $\bar{f}(Y_j)$ and $\zeta(Y_j)$. Therefore $\wp_2$ is concave, which implies that $\wp_2$ is non-positive on $[0,1]$. In particular, $\wp_2(1) = T(\bar{f}, g) \leq 0$. As $g \in \bar{S}_2$ is arbitrary, $\sup_{g \in \bar{S}_2} T(\bar{f}, g)$ is non-positive.

Similar arguments show that if $\sup_{g \in \bar{S}_2} T(\bar{f}, g)$ is non-positive, then $\sup_{g \in \bar{S}_2} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f}) \leq 0$.

□

Let $\tilde{s} \in S$ such that $L_{\mathcal{X}}(\tilde{s}) \geq L_{\mathcal{X}}(g)$ for all $g \in S$ and $L_{\mathcal{X}}(\tilde{s}) \neq -\infty$. The above claim then shows that $T(\tilde{s}, g) \leq 0$ for all $g \in S$ such that $L_{\mathcal{X}}(g) \neq -\infty$. Choose now $g \in S$ such that

$L_{\mathcal{X}}(g) = -\infty$. Define for $u \in [0,1]$, $f_u = (1-u)\tilde{s} + ug \in S$ and note that $f_1 = g$. If $u \in [0,1)$, $L(f_u) \neq -\infty$ and thus $T(\tilde{s}, f_u) \leq 0$. The continuity of the map $u \in [0,1] \mapsto T(\tilde{s}, f_u)$ ensures that $T(\tilde{s}, g) \leq 0$. Finally, $\gamma(\tilde{s}) = 0$.

Conversely, let $\hat{s}$ be a ρ-estimator satisfying $\gamma(\hat{s}) = 0$. We begin by proving that $L_{\mathcal{X}}(\hat{s}) \neq -\infty$. Consider $g \in S$ such that $L_{\mathcal{X}}(g) \neq -\infty$ and define for $u \in [0,1]$, $f_u = (1-u)\hat{s} + ug \in S$,

$$\begin{aligned}\wp_3(u) &= T(\hat{s}, f_u) \\ &= \frac{1}{n}\sum_{j\in\hat{J}} \psi\left(\frac{(1-u)\hat{s}(Y_j) + ug(Y_j)}{\hat{s}(Y_j)}\right) - \frac{1}{4}\int_{\mathcal{X}} (f_u(x) - \hat{s}(x))\, \mathrm{d}M(x).\end{aligned}$$

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

$$\wp_3(u) = \frac{|\hat{J}'|}{n} + \frac{1}{n}\sum_{j\in\hat{J}\setminus\hat{J}'} \psi\left(\frac{(1-u)\hat{s}(Y_j) + ug(Y_j)}{\hat{s}(Y_j)}\right) - \frac{1}{4}\int_{\mathcal{X}} (f_u(x) - \hat{s}(x))\, \mathrm{d}M(x).$$

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

6.3. Proof of Theorem 2. We introduce the random measure M_s defined by

$$M_s(A) = \int_A s(t)\, \mathrm{d}M(t) \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

Note that $\mathbb{E}[N(A)] = \mathbb{E}[M_s(A)]$ for all $A \in \mathcal{B}(\mathbb{R})$.

The lemma below shows that a bound on $Z(\varphi)$ can be derived from results on deviations of random variables $N(A) - \mathbb{E}[N(A)]$ and $M_s(A) - \mathbb{E}[M_s(A)]$.

Lemma 6. *Let $\mathcal{F}$ be a collection of functions of $\mathcal{S}$ such that $|\varphi| \leq 1$ for all $\varphi \in \mathcal{F}$. Consider a collection $\mathcal{A} \subset \mathcal{B}(\mathbb{R})$ of measurable sets such that*

$$\mathcal{A} \supset \bigcup_{t\in(0,1)} \{\{x \in \mathbb{R},\, \varphi_+(x) \geq t\}, \varphi \in \mathcal{F}\} \cup \{\{x \in \mathbb{R},\, \varphi_-(x) \geq t\}, \varphi \in \mathcal{F}\}.$$

Suppose that there exist α, β and an event on which: for all $A \in \mathcal{A}$,

$$|N(A) - \mathbb{E}[N(A)]| \leq \sqrt{\frac{\alpha}{n}}\left(\sqrt{N(A)} + \sqrt{\mathbb{E}[N(A)]}\right) + \frac{\beta}{n}. \tag{35}$$

$$|M_s(A) - \mathbb{E}[M_s(A)]| \leq \sqrt{\frac{\alpha}{n}}\left(\sqrt{M_s(A)} + \sqrt{\mathbb{E}[M_s(A)]}\right) + \frac{\beta}{n}. \tag{36}$$

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

$$|Z(\varphi)| \leq C\left\{\sqrt{\frac{\alpha}{n}\upsilon(\varphi)\log_+(1/\upsilon(\varphi))} + \frac{\alpha+\beta}{n}\right\}.$$

The constant C appearing in the preceding inequality is universal.

The proof of this result is delayed to Section 6.4 below. It remains to verify that inequalities (35) and (36) hold true in our different statistical settings. This is the purpose of the following lemma whose proof is deferred to Section 6.5.

Lemma 7. *Let $\mathcal{A}$ be an at most countable collection of measurable subsets of $\mathbb{R}$ and $S_{\mathcal{A}}(n)$ be the Vapnik-Chervonenkis shatter coefficient defined by*

$$S_{\mathcal{A}}(n) = \max_{x_1,\dots,x_n\in\mathbb{R}} |\{\{x_1,\dots,x_n\}\cap A, A\in\mathcal{A}\}|.$$

Let $\xi>0$. 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 (35) holds for all $A\in\mathcal{A}$ with $\alpha = c[\log_+|S_{\mathcal{A}}(2n)|+n\xi]$ and $\beta=0$.

Let for $d\geq 1$, $\overline{\mathcal{I}}_d$ be the class of unions of at most d intervals with endpoints in $\mathbb{Q}\cup\{-\infty,+\infty\}$. Then, $\overline{\mathcal{I}}_d$ is at most countable,

$$\log_+\left|S_{\overline{\mathcal{I}}_d}(2n)\right| \leq 4d\log_+(n/d), \tag{37}$$

and (36) holds true on Ω_ξ for all $A\in\overline{\mathcal{I}}_d$ with $\alpha=\beta=c'[d\log_+(n/d)+n\xi]$ where c' is a universal constant.

To prove Theorem 2, note that Lemma 7 implies that (35) and (36) hold true for the collection $\overline{\mathcal{I}}_d$ with $\alpha=\beta=c''[d\log_+(n/d)+n\xi]$. Let now $\mathcal{I}_d$ be the class of unions of at most d intervals, that is

$$\mathcal{I}_d = \left\{ \bigcup_{j=1}^d I_j,\ I_j \text{ is a (possibly empty) interval of } \mathbb{R} \right\}.$$

Then, for all $\epsilon>0$, $A\in\mathcal{I}_d$, there exists $\bar{A}\in\overline{\mathcal{I}}_d$ such that

$$|N(A)-N(\bar{A})|\leq\epsilon \quad\text{and}\quad |M_s(A)-M_s(\bar{A})|\leq\epsilon.$$

Thereby, (35) and (36) hold with $\mathcal{A}=\mathcal{I}_d$ (up to a modification of β). Lemma 6 finally implies (8).

We then deduce (9) from some elementary computations. As $x\mapsto x\log_+(1/x)$ is non-decreasing, we have for all $x\leq y$, $x\log_+(1/x)y\leq y^2\log_+(1/y)$. Moreover, when $x\geq y$, $\log_+(1/x)\leq\log_+(1/y)$ and hence $x\log_+(1/x)y\leq x\log_+(1/y)y$. Thereby, for all $x,y>0$,

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

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

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

where C_ε depends on ε. By using this result with $x=\upsilon(\varphi)$ and $y=(d/n)\log_+(n/d)$.

$$\begin{aligned} 2\sqrt{\upsilon(\varphi)\log_+(1/\upsilon(\varphi))\left(\frac{d\log_+(n/d)}{n}\right)} &\leq \varepsilon\upsilon(\varphi)+C_\varepsilon\left(\frac{d\log_+(n/d)}{n}\right)\log_+\left(\frac{1}{\frac{d\log_+(n/d)}{n}}\right)\\ &\leq \varepsilon\upsilon(\varphi)+C_\varepsilon\frac{d\log_+^2(n/d)}{n}. \end{aligned}$$

Similarly,

$$2\sqrt{v(\varphi)\log_+(1/v(\varphi))\xi} \le \varepsilon v(\varphi) + C_\varepsilon \xi \log_+(1/\xi).$$

Therefore,

$$2\sqrt{v(\varphi)\log_+(1/v(\varphi))\left(\frac{d\log_+(n/d)}{n}+\xi\right)} \le 2\varepsilon v(\varphi) + C_\varepsilon\left[\frac{d\log_+^2(n/d)}{n} + \xi\log_+(1/\xi)\right],$$

and (9) follows from (8). □

6.4. Proof of Lemma 6. Without lost of generality, we may assume that the functions φ of $\mathcal{F}$ are non-negative. We suppose moreover that we are on an event on which (35) and (36) hold true. Let for $t \in (0,1)$, $A_t = \{x \in \mathbb{R},\, \varphi(x) \ge t\}$. Then, for all $x \in \mathbb{R}$,

$$\varphi(x) = \int_0^1 1_{A_t}(x)\,\mathrm{d}t,$$
$$\varphi^2(x) = 2\int_0^1 t 1_{A_t}(x)\,\mathrm{d}t.$$

Let $\varepsilon > 0$. We have,

$$\begin{aligned} |Z(\varphi)| - \varepsilon v(\varphi) &= \left|\int_0^1 (N(A_t) - M_s(A_t))\,\mathrm{d}t\right| - 2\varepsilon\int_0^1 tM_s(A_t)\,\mathrm{d}t \\ &\le \int_0^1 \{|N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t)\}\,\mathrm{d}t. \end{aligned} \tag{38}$$

We now bound above $|N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t)$ and this requires some elementary but tedious computations. The result is summarized in the claim below whose proof is postponed after the present one.

Claim 2. *For all $t \in (0,1)$ and $\varepsilon > 0$,*

$$|N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t) \le \frac{329\alpha + 70\beta}{16n} + \frac{19\alpha}{2n\varepsilon t}.$$

We derive from (38) that for all $\eta \in (0,1]$,

$$\begin{aligned} |Z(\varphi)| - \varepsilon v(\varphi) &\le \int_0^\eta |N(A_t) - M_s(A_t)|\,\mathrm{d}t + \int_\eta^1 \left(\frac{329\alpha+70\beta}{16n} + \frac{19\alpha}{2n\varepsilon t}\right)\mathrm{d}t \\ &\le \int_0^\eta |N(A_t) - M_s(A_t)|\,\mathrm{d}t + \frac{329\alpha+70\beta}{16n} + \frac{19\alpha}{2n\varepsilon}\log(1/\eta). \end{aligned} \tag{39}$$

We need to bound above the integral appearing in the right-hand side of this inequality. We have,

$$\int_0^\eta |N(A_t) - M_s(A_t)|\,\mathrm{d}t \le \int_0^\eta |N(A_t) - \mathbb{E}[N(A_t)]|\,\mathrm{d}t + \int_0^\eta |M_s(A_t) - \mathbb{E}[M_s(A_t)]|\,\mathrm{d}t. \tag{40}$$

By using (35) and the inequalities $xy \le x^2 + 1/4y^2$, $(x+y)^2 \le 2x^2 + 2y^2$,

$$N(A_t) - \mathbb{E}[N(A_t)] \le \frac{1}{2}(N(A_t) + \mathbb{E}[N(A_t)]) + \frac{\alpha+\beta}{n},$$

and thus

$$N(A_t) \leq 3\mathbb{E}[N(A_t)] + 2\frac{\alpha+\beta}{n}.$$

By plugging this inequality into (35),

$$\begin{aligned} |N(A_t) - \mathbb{E}[N(A_t)]| &\leq \sqrt{\frac{\alpha}{n}}\sqrt{\mathbb{E}[N(A_t)]} + \sqrt{\frac{\alpha}{n}}\sqrt{3\mathbb{E}[N(A_t)] + 2\frac{\alpha+\beta}{n}} + \frac{\beta}{n} \\ &\leq (1+\sqrt{3})\sqrt{\frac{\alpha}{n}}\sqrt{\mathbb{E}[N(A_t)]} + \frac{\sqrt{2}\alpha + \sqrt{2\alpha\beta} + \beta}{n} \\ &\leq (1+\sqrt{3})\sqrt{\frac{\alpha}{n}}\sqrt{\mathbb{E}[N(A_t)]} + 2.2\frac{\alpha+\beta}{n}. \end{aligned} \tag{41}$$

The constant 2.2 above comes from the inequality $2\sqrt{xy} \leq x + y$ and the bound $1 + 1/\sqrt{2} \leq \sqrt{2} + 1/\sqrt{2} \leq 2.2$. Similarly,

$$\begin{aligned} |M_s(A_t) - \mathbb{E}[M_s(A_t)]| &\leq (1+\sqrt{3})\sqrt{\frac{\alpha}{n}}\sqrt{\mathbb{E}[M_s(A_t)]} + 2.2\frac{\alpha+\beta}{n} \\ &\leq (1+\sqrt{3})\sqrt{\frac{\alpha}{n}}\sqrt{\mathbb{E}[N(A_t)]} + 2.2\frac{\alpha+\beta}{n}. \end{aligned}$$

We now derive from (40) that,

$$\begin{aligned} \int_0^\eta |N(A_t) - M_s(A_t)| \,\mathrm{d}t &\leq 2(1+\sqrt{3})\sqrt{\frac{\alpha}{n}}\int_0^\eta \sqrt{\mathbb{E}[N(A_t)]}\,\mathrm{d}t + 4.4\frac{\alpha+\beta}{n} \\ &\leq 2(1+\sqrt{3})\sqrt{\frac{\alpha}{n}}\sqrt{\eta}\sqrt{\int_0^\eta \mathbb{E}[N(A_t)]\,\mathrm{d}t} + 4.4\frac{\alpha+\beta}{n}. \end{aligned}$$

Now, $\mathbb{E}[N(A_t)] \leq \mathbb{E}[N(\mathbb{R})] \leq 1$ in our frameworks and thus,

$$\int_0^\eta |N(A_t) - M_s(A_t)| \,\mathrm{d}t \leq 2(1+\sqrt{3})\sqrt{\frac{\alpha}{n}}\eta + 4.4\frac{\alpha+\beta}{n}.$$

Plugging this last inequality into (39) finally shows that there exists a universal constant C such that for all $\varepsilon > 0$, $\eta \in (0,1]$,

$$|Z(\varphi)| \leq \varepsilon \upsilon(\varphi) + C\left\{\frac{\alpha+\beta}{n} + \sqrt{\frac{\alpha}{n}}\eta + \frac{\alpha}{n\varepsilon}\log(1/\eta)\right\}.$$

By choosing suitably ε, we deduce,

$$|Z(\varphi)| \leq 2\sqrt{C}\sqrt{\frac{\alpha}{n}}\left(\sqrt{\upsilon(\varphi)\log(1/\eta)} + \frac{\sqrt{C}}{2}\eta\right) + C\frac{\alpha+\beta}{n}.$$

It remains to set $\eta^2 = \min\{\upsilon(\varphi), 1\}$ to prove the lemma. □

Proof of Claim 2. As $\mathbb{E}[N(A_t)] = \mathbb{E}[M_s(A_t)]$,

$$|N(A_t) - M_s(A_t)| - \varepsilon t M_s(A_t) \leq |N(A_t) - \mathbb{E}[N(A_t)]| + |\mathbb{E}[M_s(A_t)] - M_s(A_t)| - \varepsilon t M_s(A_t).$$

By using (35) and (36),

$$\begin{aligned}|N(A_t)-M_s(A_t)|-\varepsilon t M_s(A_t) &\leq \sqrt{\frac{\alpha}{n}}\left(\sqrt{N(A_t)}+\sqrt{\mathbb{E}[N(A_t)]}\right)+\sqrt{\frac{\alpha}{n}}\left(\sqrt{M_s(A_t)}+\sqrt{\mathbb{E}[M_s(A_t)]}\right)\\ &\quad -\varepsilon t M_s(A_t)+\frac{2\beta}{n}\\ &\leq \sqrt{\frac{\alpha}{n}}\sqrt{N(A_t)}+\sqrt{\frac{\alpha}{n}}\sqrt{M_s(A_t)}+2\sqrt{\frac{\alpha}{n}}\sqrt{\mathbb{E}[M_s(A_t)]}\\ &\quad -\varepsilon t M_s(A_t)+\frac{2\beta}{n}.\end{aligned}$$

By using the relation $2\sqrt{xy}\leq ax+a^{-1}y$ for all $a>0$,

$$\begin{aligned}\sqrt{\frac{\alpha}{n}}\sqrt{M_s(A_t)} &\leq \frac{\varepsilon t}{2}M_s(A_t)+\frac{\alpha}{2\varepsilon n t}\\ 2\sqrt{\frac{\alpha}{n}}\sqrt{\mathbb{E}[M_s(A_t)]} &\leq \frac{\varepsilon t}{4}\mathbb{E}[M_s(A_t)]+\frac{4\alpha}{\varepsilon n t}.\end{aligned}$$

Therefore,

$$|N(A_t)-M_s(A_t)|-\varepsilon t M_s(A_t)\leq \sqrt{\frac{\alpha}{n}}\sqrt{N(A_t)}+\frac{\varepsilon t}{4}\left[\mathbb{E}[M_s(A_t)]-2M_s(A_t)\right]+\frac{2\beta}{n}+\frac{9\alpha}{2n\varepsilon t}.$$

By noticing that for all $u>0$,

$$\sqrt{\frac{\alpha}{n}}\sqrt{N(A_t)}\leq \frac{u}{2}N(A_t)+\frac{\alpha}{2nu},$$

we deduce,

$$|N(A_t)-M_s(A_t)|-\varepsilon t M_s(A_t)\leq \frac{u}{2}N(A_t)+\frac{\varepsilon t}{4}\left[\mathbb{E}[M_s(A_t)]-2M_s(A_t)\right]+\frac{2\beta}{n}+\frac{9\alpha}{2n\varepsilon t}+\frac{\alpha}{2nu}.$$

We suppose from now on that $\varepsilon t<1/2$. Then,

$$\begin{aligned}|N(A_t)-M_s(A_t)|-\varepsilon t M_s(A_t)\leq \frac{u}{2}N(A_t)+\frac{1}{8}\left(\mathbb{E}[M_s(A_t)]-2M_s(A_t)\right)_+ &+\frac{2\beta}{n}\\ &+\frac{9\alpha}{2n\varepsilon t}+\frac{\alpha}{2nu}.\end{aligned}\tag{42}$$

We choose u according to the sign of $N(A_t)-M_s(A_t)$:

- If $N(A_t)\geq M_s(A_t)$, we set $u=2\varepsilon t/(1+2\varepsilon t)$. In this case,

$$(1+\varepsilon t)/(1-u/2)=1+2\varepsilon t.\tag{43}$$

- If $N(A_t)<M_s(A_t)$, we set $u=2\varepsilon t/(1-2\varepsilon t)$. In this case,

$$(1-\varepsilon t)/(1+u/2)=1-2\varepsilon t.\tag{44}$$

When $N(A_t)\geq M_s(A_t)$, we deduce from (42),

$$(1-u/2)N(A_t)-(1+\varepsilon t)M_s(A_t)\leq \frac{1}{8}\left(\mathbb{E}[M_s(A_t)]-2M_s(A_t)\right)_+ +\frac{2\beta}{n}+\frac{9\alpha}{2n\varepsilon t}+\frac{\alpha}{2nu}.$$

Therefore, using (43) and $1/(1-u/2) \leq 2$,

$$\begin{aligned} N(A_t) - (1+2\varepsilon t)M_s(A_t) &\leq \frac{1}{1-u/2}\left[\frac{1}{8}\left(\mathbb{E}[M_s(A_t)] - 2M_s(A_t)\right)_+ + \frac{2\beta}{n} + \frac{9\alpha}{2n\varepsilon t} + \frac{\alpha}{2nu}\right] \\ &\leq \frac{1}{4}\left(\mathbb{E}[M_s(A_t)] - 2M_s(A_t)\right)_+ + \frac{4\beta}{n} + \frac{9\alpha}{n\varepsilon t} + \frac{\alpha}{nu}. \end{aligned} \tag{45}$$

When $N(A_t) < M_s(A_t)$, (42) yields

$$(1-\varepsilon t)M_s(A_t) - (1+u/2)N(A_t) \leq \frac{1}{8}\left(\mathbb{E}[M_s(A_t)] - 2M_s(A_t)\right)_+ + \frac{2\beta}{n} + \frac{9\alpha}{2n\varepsilon t} + \frac{\alpha}{2nu}.$$

Therefore, by using (44) and $1/(1+u/2) \leq 1$,

$$\begin{aligned} (1-2\varepsilon t)M_s(A_t) - N(A_t) &\leq \frac{1}{1+u/2}\left[\frac{1}{8}\left(\mathbb{E}[M_s(A_t)] - 2M_s(A_t)\right)_+ + \frac{2\beta}{n} + \frac{9\alpha}{2n\varepsilon t} + \frac{\alpha}{2nu}\right] \\ &\leq \frac{1}{8}\left(\mathbb{E}[M_s(A_t)] - 2M_s(A_t)\right)_+ + \frac{2\beta}{n} + \frac{9\alpha}{2n\varepsilon t} + \frac{\alpha}{2nu}. \end{aligned} \tag{46}$$

In both cases, we have

$$|N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t) \leq \frac{1}{4}\left(\mathbb{E}[M_s(A_t)] - 2M_s(A_t)\right)_+ + \frac{4\beta}{n} + \frac{9\alpha}{n\varepsilon t} + \frac{\alpha}{nu}.$$

Since $1/u \leq 1 + 1/(2\varepsilon t)$,

$$|N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t) \leq \frac{1}{4}\left(\mathbb{E}[M_s(A_t)] - 2M_s(A_t)\right)_+ + \frac{\alpha+4\beta}{n} + \frac{19\alpha}{2n\varepsilon t}. \tag{47}$$

It remains to bound above $\mathbb{E}[M_s(A_t)] - 2M_s(A_t)$. Yet, (36) claims that

$$\mathbb{E}\left[M_s(A_t)\right] - M_s(A_t) \leq \sqrt{\frac{\alpha}{n}}\left(\sqrt{M_s(A_t)} + \sqrt{\mathbb{E}[M_s(A_t)]}\right) + \frac{\beta}{n}.$$

By using

$$\sqrt{\frac{\alpha}{n}}\left(\sqrt{M_s(A_t)} + \sqrt{\mathbb{E}[M_s(A_t)]}\right) \leq \frac{3\alpha}{2n} + \frac{1}{3}\left(M_s(A_t) + \mathbb{E}[M_s(A_t)]\right),$$

we deduce

$$\mathbb{E}\left[M_s(A_t)\right] - 2M_s(A_t) \leq \frac{9\alpha+6\beta}{4n}.$$

By putting this inequality into (47), we finally obtain when $\varepsilon t < 1/2$,

$$|N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t) \leq \frac{25\alpha+70\beta}{16n} + \frac{19\alpha}{2n\varepsilon t}.$$

Since $\varepsilon \mapsto |N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t)$ is non-increasing, we deduce when $\varepsilon t \geq 1/2$ that

$$\begin{aligned} |N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t) &\leq \frac{25\alpha+70\beta}{16n} + \frac{19\alpha}{n} \\ &\leq \frac{329\alpha+70\beta}{16n}. \end{aligned}$$

As a straightforward consequence, we get for all $\varepsilon t > 0$,

$$|N(A_t) - M_s(A_t)| - 2\varepsilon t M_s(A_t) \leq \frac{329\alpha+70\beta}{16n} + \frac{19\alpha}{2n\varepsilon t},$$

which proves the claim $\square$

6.5. Proof of Lemma 7.

6.5.1. Proof of (35). The proof of (35) is based on Vapnik-Chervonenkis inequalities for relative deviation (see, for instance page 24 of [DL12]). We recall them below:

Theorem 10. *Let $Z_1, \dots, Z_n$ be n independent and identically distributed variables with values in a space $\mathcal{X}$. Let $\mathcal{A}'$ be an at most countable collection of measurable sets. Define the empirical measure $\nu_n(A') = n^{-1}\sum_{i=1}^n 1_{A'}(Z_i)$, $\nu(A') = \mathbb{E}[\mu_n(A')]$ and the Vapnik-Chervonenkis shatter coefficient*

$$S_{\mathcal{A}'}(n) = \max_{z_1,\dots,z_n\in\mathcal{X}} \left|\left\{\{z_1,\dots,z_n\}\cap A', A'\in\mathcal{A}'\right\}\right|.$$

Then, for all $t>0$,

$$\mathbb{P}\left(\sup_{A'\in\mathcal{A}'} \frac{\nu(A')-\nu_n(A')}{\sqrt{\nu(A')}} \geq t\right) \leq 4S_{\mathcal{A}'}(2n)e^{-nt^2/4}$$
$$\mathbb{P}\left(\sup_{A'\in\mathcal{A}'} \frac{\nu_n(A')-\nu(A')}{\sqrt{\nu_n(A')}} \geq t\right) \leq 4S_{\mathcal{A}'}(2n)e^{-nt^2/4}.$$

This implies in particular:

$$\mathbb{P}\left(\sup_{A'\in\mathcal{A}'} \left|\sqrt{\nu_n(A')}-\sqrt{\nu(A')}\right| \geq t\right) \leq 8S_{\mathcal{A}'}(2n)e^{-nt^2/4}. \tag{48}$$

Assume that we are within framework 1. Then, the random measure N is the empirical measure of $X_1,\dots,X_n$. Now (48) with $\mathcal{A}'=\mathcal{A}$,

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

shows that (35) holds true with probability larger than $1-e^{-n\xi}$, $\alpha = nt^2$, $\beta = 0$.

The proof in frameworks 2 and 3 is very similar since N is an empirical measure for suitable random variables with values in $\mathcal{X} = \mathbb{R}\times\{0,1\}$: $Z_i = (X_i, 1_{D_i=1})$ in framework 2 and $Z_i = (T^{(i)}_{1,0}1_{T^{(i)}_{1,0}<\infty}, 1_{T^{(i)}_{1,0}<\infty})$ in framework 3. We apply (48) with $\mathcal{A}' = \{A\times\{1\},\ A\in\mathcal{A}\}$. Note that the Vapnik-Chervonenkis shatter coefficient $S_{\mathcal{A}'}(2n)$ can be upper bounded by

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

We end the proof as in framework 1. □

6.5.2. Proof of (37). It follows from Lemma 1 of [BB16] that the Vapnik-Chervonenkis dimension of $\overline{\mathcal{I}}_d$ is at most $2d$. By using Sauer's lemma (see [Sau72]), we deduce

$$\left|S_{\overline{\mathcal{I}}_d}(2n)\right| \leq \sum_{j=0}^{2d} C^j_{2n}.$$

By using a classical inequality (see, for instance, exercise 2.14 of [BLM13]), we deduce when $d \leq n$,

$$\left|S_{\overline{\mathcal{I}}_d}(2n)\right| \leq (en/d)^{2d},$$

and when $d \geq n$,

$$\left|S_{\overline{\mathcal{I}}_d}(2n)\right| \leq e^{2d}.$$

This implies (37). □

6.5.3. Proof of (36). Note that there is nothing to prove in framework 1 as M_s is deterministic. We define $V_i(t) = 1_{X_i \geq t} 1_{[0,+\infty)}(t)$ in framework 2 and $V_i(t) = 1_{X_{t^-}^{(i)}=1} 1_{(0,+\infty)}(t)$ in framework 3. Then, M_s is of the form

$$M_s(A) = \frac{1}{n}\sum_{i=1}^{n} \int_A s(t)V_i(t)\,\mathrm{d}t \quad \text{for all } A \in \mathcal{B}(\mathbb{R}).$$

We need the proposition below whose proof is delayed to Sections 6.5.5 and 6.5.6.

Proposition 11. *For all $k \geq 1$, $i \in \{1, \dots, n\}$ and $A \in \mathcal{B}(\mathbb{R})$,*

$$\mathbb{E}\left[\left(\int_A s(t)V_i(t)\,\mathrm{d}t\right)^k\right] \leq k!\mathbb{E}\left[M_s(A)\right].$$

There exist independent random variables $Z_1, \dots, Z_n$ such that M_s is of the form

$$M_s(A) = \frac{1}{n}\sum_{i=1}^{n} f_A(Z_i) \quad \text{with } f_A(Z_i) = \int_A s(t)V_i(t)\,\mathrm{d}t.$$

The core of the proof is based on the exponential inequality given by Corollary 6.9 of [Mas07]. The aim of the lemma below is to show that its assumptions are satisfied. Its proof is postponed after the present one.

Lemma 8. *For all $\delta > 0$, there exists a collection $\mathcal{C}_\delta$ of functions of the form f_A with $A \in \overline{\mathcal{I}}_d$. The cardinality of this set can be bounded by $\log|\mathcal{C}_\delta| \leq c_1 d \log_+(1/\delta^2)$, where c_1 is a universal constant. Moreover, for all $A \in \overline{\mathcal{I}}_d$, 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 $k \geq 1$,*

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

Set for $\delta > 0$, $\mathcal{B}_\delta = \mathcal{C}_\delta \cup \{-f,\, f \in \mathcal{C}_\delta\}$. Note that

$$\log|\mathcal{B}_\delta| \leq \log 2 + \log|\mathcal{C}_\delta| \leq c_2 d \log_+(1/\delta^2),$$

where c_2 is a universal constant. We set $H(\delta) = c_2 d \log_+(1/\delta^2)$ and for $\sigma \in (0, 1]$,

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

By differentiation,

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

There exists therefore a universal constant c_3 such that

$$E \leq c_3\left[\sigma\sqrt{nd\log_+(1/\sigma^2)} + d\log_+(1/\sigma^2)\right]. \tag{49}$$

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 $j \in J$, $x_j = 2\log(j+1) + 1 + n\xi$, $\overline{\mathcal{A}}_j = \left\{A \in \overline{\mathcal{I}}_d,\, 2^{-j-1} \leq \mathbb{E}[M_s(A)] \leq 2^{-j}\right\}$. The 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 $H(\delta) = c_2 d\log_+(1/\delta^2)$. Consequently, there exists an event Ω_j such that $\mathbb{P}(\Omega_j) \geq 1 - e^{-x_j}$ and on which: for all $A \in \overline{\mathcal{A}}_j$,

$$n\left|M_s(A) - \mathbb{E}[M_s(A)]\right| \leq c_4\left[E + \sigma\sqrt{nx_j} + x_j\right],$$

where c_4 is universal. Therefore,

$$|M_s(A) - \mathbb{E}[M_s(A)]| \leq c_5\left[\sigma\sqrt{\frac{d\log_+(1/\sigma^2) + x_j}{n}} + \frac{d\log_+(1/\sigma^2) + x_j}{n}\right].$$

As $\sigma^2 \leq 4\mathbb{E}[M_s(A)]$, and $\sigma^2 \geq d/n$, we get

$$|M_s(A) - \mathbb{E}[M_s(A)]| \leq c_6\left[\sqrt{\mathbb{E}[M_s(A)]}\sqrt{\frac{d\log_+(n/d) + x_j}{n}} + \frac{d\log_+(n/d) + x_j}{n}\right].$$

Note that $x_j \leq c_7\left(\log_+(n/d) + n\xi\right)$ and hence,

$$|M_s(A) - \mathbb{E}[M_s(A)]| \leq c_8\left[\sqrt{\mathbb{E}[M_s(A)]}\sqrt{\frac{d\log_+(n/d) + n\xi}{n}} + \frac{d\log_+(n/d) + n\xi}{n}\right].$$

Let now $\overline{\mathcal{A}} = \left\{A \in \overline{\mathcal{I}}_d,\, \mathbb{E}[M_s(A)] \leq d/(2n)\right\}$. We apply Corollary 6.9 of Massart 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 $A \in \overline{\mathcal{A}}$,

$$\begin{aligned}|M_s(A) - \mathbb{E}[M_s(A)]| &\leq c_5\left[\sigma\sqrt{\frac{d\log_+(1/\sigma^2) + n\xi + \log 2}{n}} + \frac{d\log_+(1/\sigma^2) + n\xi + \log 2}{n}\right], \\ &\leq c_9\left[\sigma\sqrt{\frac{d\log_+(n/d) + n\xi}{n}} + \frac{d\log_+(n/d) + n\xi}{n}\right].\end{aligned}$$

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

$$|M_s(A) - \mathbb{E}[M_s(A)]| \leq c_{10}\left[\frac{d\log_+(n/d) + n\xi}{n}\right].$$

As a straightforward consequence, the following inequality also holds for $A \in \overline{\mathcal{A}}$:

$$|M_s(A) - \mathbb{E}[M_s(A)]| \leq c_{10} \left[\sqrt{\mathbb{E}[M_s(A)]} \sqrt{\frac{d \log_+(n/d) + n\xi}{n}} + \frac{d \log_+(n/d) + n\xi}{n} \right].$$

Now, note that $\overline{\mathcal{I}}_d = \bigcup_{j \in J} \overline{\mathcal{A}}_j \cup \overline{\mathcal{A}}$, which shows that inequality (36) holds on the event $\Omega' \bigcap \left(\cap_{j \in J} \Omega_j \right)$ with $\alpha = c_{11} \left(d \log_+(n/d) + n\xi \right)$. Moreover,

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

□

6.5.4. Proof of Lemma 8. First of all, we only need to prove the lemma when δ is smaller than 1, what we shall do in the sequel.

We endow $\overline{\mathcal{I}}_d$ with the distance *dist* defined for $A_1, A_2 \in \overline{\mathcal{I}}_d$ by

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

Note that $dist(A_1, A_2)$ can also be written as

$$dist(A_1, A_2) = \int_{\mathbb{R}} |1_{A_1}(t) - 1_{A_2}(t)| \; f(t) \, \mathrm{d}t,$$

where $f(t) = s(t)\mathbb{E}[V_1(t)]$ is a non-negative function satisfying $\int_{\mathbb{R}} f(t) \, \mathrm{d}t \leq 1$.

We introduce the real valued function F defined by

$$F(x) = \int_{-\infty}^{x} f(t) \, \mathrm{d}t \quad \text{for all } x \in \mathbb{R}.$$

Since F is a continuous non-decreasing function such that $F(\mathbb{R}) \subset [0, 1]$, there exist an even integer $\ell \in [2, 4d/\delta^2 + 2]$, and ℓ numbers $(x_1, x_2, \ldots, \; x_{\ell-1}, x_\ell) \in \{-\infty\} \times \mathbb{Q}^{\ell-2} \times \{+\infty\}$ such that

$$\max_{1 \leq i \leq \ell - 1} \{F(x_{i+1}) - F(x_i)\} \leq \delta^2/(4d).$$

Note that we may suppose that $\ell \geq d$. Let $\mathcal{X} = \{x_1, x_2, \ldots, x_\ell\}$, and $\overline{\mathcal{I}}_{dis}$ be the collection of union of at most d closed intervals whose endpoints belong to $\mathcal{X}$. To avoid any ambiguity, we recall that $[x_k, +\infty)$ and $(-\infty, x_k]$ are closed intervals.

When $k \leq \ell/2$, choosing k disjoint closed intervals whose endpoints belong to $\mathcal{X}$ amounts to choosing $2k$ numbers among $\mathcal{X}$. When $k > \ell/2$, we cannot find k disjoint closed intervals with endpoints in $\mathcal{X}$. The cardinality of $\overline{\mathcal{I}}_{dis}$ is therefore bounded by

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

Standard arguments (see, for instance, exercise 2.14 of [BLM13]) show that $|\overline{\mathcal{I}}_{dis}| \leq (\ell e/d)^d$. Using now that $\ell \leq 4d/\delta^2 + 2$, we derive that

$$\log |\overline{\mathcal{I}}_{dis}| \leq c_1 d \log_+(1/\delta^2)$$

for a suitable universal constant c_1.

For each set $A \in \overline{\mathcal{I}}_d$, we now show that there exist $A_1, A_2 \in \overline{\mathcal{I}}_{dis}$ such that $f_{A_1} \leq f_A \leq f_{A_2}$ and $dist(A_1, A_2) \leq \delta^2/2$. Let $A \in \overline{\mathcal{I}}_d$ be written as $A = \bigcup_{k=1}^d A_k$ where A_k is an interval whose endpoints are $a_k \leq b_k$. For each $k \in \{1, \ldots, d\}$, there exist $a_k^{(1)}, a_k^{(2)}, b_k^{(1)}, b_k^{(2)} \in \mathcal{X}$ such that

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

and

$$F\big(a_k^{(2)}\big) - F\big(a_k^{(1)}\big) \leq \delta^2/(4d), \quad F\big(b_k^{(2)}\big) - F\big(b_k^{(1)}\big) \leq \delta^2/(4d).$$

Define the closed intervals

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

Then, $A_1 = \bigcup_{k=1}^d A_k^{(1)}$ and $A_2 = \bigcup_{k=1}^d A_k^{(2)}$ belong to $\overline{\mathcal{I}}_{dis}$ and satisfy $f_{A_1} \leq f_A \leq f_{A_2}$. Moreover,

$$A_2 \Delta A_1 \subset \bigcup_{k=1}^d [a_k^{(1)}, a_k^{(2)}) \cup (b_k^{(1)}, b_k^{(2)}],$$

and hence,

$$\begin{aligned}
dist(A_1, A_2) &\leq \sum_{k=1}^d \int_{[a_k^{(1)}, a_k^{(2)}) \cup (b_k^{(1)}, b_k^{(2)}]} f(t)\, \mathrm{d}t \\
&\leq \sum_{k=1}^d \Big(F\big(a_k^{(2)}\big) - F\big(a_k^{(1)}\big) + F\big(b_k^{(2)}\big) - F\big(b_k^{(1)}\big)\Big) \\
&\leq \sum_{k=1}^d \big(\delta^2/(4d) + \delta^2/(4d)\big) \\
&\leq \delta^2/2.
\end{aligned}$$

Now,

$$\mathbb{E}\left[\left(f_{A_2}(Z_1) - f_{A_1}(Z_1)\right)^k\right] = \mathbb{E}\left[\left(\int_{A_2 \setminus A_1} V_1(t) s(t)\, \mathrm{d}t\right)^k\right].$$

We deduce from Proposition 11,

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

Yet, $\mathbb{E}\left[M_s(A_2 \setminus A_1)\right] = dist(A_1, A_2) \leq \delta^2/2$, which completes the proof with $\mathcal{C}_\delta = \{f_A,\, A \in \overline{\mathcal{I}}_{dis}\}$. □

6.5.5. Proof of Proposition 11 in framework 2. Remark that for all $a \in \mathbb{R}$,

$$\begin{aligned}\int_a^\infty s(t)\mathbb{P}\left(X \geq t\right)\mathrm{d}t &= \int_a^\infty f(t)\mathbb{P}\left(C \geq t\right)\mathrm{d}t \\ &\leq \left(\int_a^\infty f(u)\,\mathrm{d}u\right)\mathbb{P}\left(C \geq a\right) \\ &\leq \mathbb{P}\left(T \geq a\right)\mathbb{P}\left(C \geq a\right) \\ &\leq \mathbb{P}\left(X \geq a\right). \qquad (50)\end{aligned}$$

We define for $k \geq 1$,

$$J_k = \int_{\substack{u_1,\dots,u_k \in A \\ u_1<u_2<\dots<u_k}} \left(\prod_{j=1}^k s(u_j)\right)\mathbb{P}\left(X \geq u_k\right)\,\mathrm{d}u_1\,\mathrm{d}u_2\dots\,\mathrm{d}u_k.$$

We have,

$$\begin{aligned}\mathbb{E}\left[\left(\int_A s(u)1_{X\geq u}\,\mathrm{d}u\right)^k\right] &= \mathbb{E}\left[\int_{A^k}\prod_{j=1}^k s(u_j)1_{X\geq u_j}\,\mathrm{d}u_1\,\mathrm{d}u_2\dots\,\mathrm{d}u_k\right] \\ &= \int_{A^k}\left(\prod_{j=1}^k s(u_j)\right)\mathbb{P}\left(X \geq \max\{u_1,\dots,u_k\}\right)\,\mathrm{d}u_1\,\mathrm{d}u_2\dots\,\mathrm{d}u_k \\ &= k!J_k.\end{aligned}$$

Now,

$$J_k \leq \int_{\substack{u_1,\dots,u_{k-1} \in A \\ u_1<u_2<\dots<u_{k-1}}} \left(\prod_{j=1}^{k-1} s(u_j)\right)\left(\int_{u_{k-1}}^\infty s(u_k)\mathbb{P}\left(X \geq u_k\right)\,\mathrm{d}u_k\right)\,\mathrm{d}u_1\,\mathrm{d}u_2\dots\,\mathrm{d}u_{k-1}.$$

By using (50) with $a = u_{k-1}$,

$$\begin{aligned}J_k &\leq \int_{\substack{u_1,\dots,u_{k-1} \in A \\ u_1<u_2<\dots<u_{k-1}}} \left(\prod_{j=1}^{k-1} s(u_j)\right)\mathbb{P}\left(X \geq u_{k-1}\right)\,\mathrm{d}u_1\,\mathrm{d}u_2\dots\,\mathrm{d}u_{k-1} \\ &\leq J_{k-1}.\end{aligned}$$

By induction, $J_k \leq J_1 = \mathbb{E}[M_s(A)]$. □

6.5.6. Proof of Proposition 11 in framework 3.

Claim 3. *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 measurable set $A \in \mathcal{B}(\mathbb{R})$ by*

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

Then, for μ-almost all $u > t$,

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

Proof. First of all, μ_B is absolutely continuous with respect to the Lebesgue measure 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, $\mathbb{P}(Z_h(u) \geq 2) = o(h)$ when $h \to 0$. We deduce,

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

When $Z_h(u) = 1$, $T_{1,0} \in [u, u+h]$ is equivalent to $X_{u-} = 1$ and $X_{u+h} = 0$. This yields

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

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

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

Now,

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

Finally, by plugging this inequality into (52),

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

which proves (51). □

We now return to the proof of Proposition 11. Without loss of generality, we may suppose that $A \subset [0, +\infty)$. Define for $k \geq 1$,

$$J_k = \int_{\substack{u_1,\ldots,u_k \in A \\ u_1 < u_2 < \cdots < u_k}} \left(\prod_{j=1}^{k} s(u_j) \right) \mathbb{P}\left(X_{u_1-} = 1, \ldots, X_{u_k-} = 1\right) \mathrm{d}u_1\, \mathrm{d}u_2 \ldots \mathrm{d}u_k.$$

We have,

$$\begin{aligned}\mathbb{E}\left[\left(\int_A s(u)1_{X_{u-}=1}\,\mathrm{d}u\right)^k\right] &= \mathbb{E}\left[\int_{A^k}\prod_{j=1}^k s(u_j)1_{X_{u_j-}=1}\,\mathrm{d}u_1\,\mathrm{d}u_2\ldots\,\mathrm{d}u_k\right]\\ &= \int_{A^k}\left(\prod_{j=1}^k s(u_j)\right)\mathbb{P}\left(X_{u_1-}=1,\ldots,X_{u_k-}=1\right)\,\mathrm{d}u_1\,\mathrm{d}u_2\ldots\,\mathrm{d}u_k\\ &= k!J_k.\end{aligned}$$

Yet,

$$J_k \le \int_{\substack{u_1,\ldots,u_{k-1}\in A\\ u_1<u_2<\cdots<u_{k-1}}}\left(\prod_{j=1}^{k-1}s(u_j)\right)\left(\int_{u_{k-1}}^\infty s(u_k)\mathbb{P}\left(X_{u_1-}=1,\ldots,X_{u_k-}=1\right)\,\mathrm{d}u_k\right)\mathrm{d}u_1\,\mathrm{d}u_2\ldots\,\mathrm{d}u_{k-1}.$$

Let $B=\left[X_{u_1-}=1,\ldots,X_{u_{k-1}-}=1\right]\in\mathcal{F}_{u_{k-1}}$. Then,

$$\begin{aligned}\int_{u_{k-1}}^\infty s(u_k)\mathbb{P}\left(X_{u_1-}=1,\ldots,X_{u_k-}=1\right)\,\mathrm{d}u_k &= \int_{u_{k-1}}^\infty \frac{\mathrm{d}\mu_B}{\mathrm{d}u}(u)\,\mathrm{d}u\\ &= \mu_B([u_{k-1},+\infty))\\ &= \mathbb{P}\left(X_{u_1-}=1,\ldots,X_{u_{k-1}-}=1 \text{ and } T_{1,0}\ge u_{k-1}\right)\\ &\le \mathbb{P}\left(X_{u_1-}=1,\ldots,X_{u_{k-1}-}=1\right).\end{aligned}$$

Therefore, $J_k\le J_{k-1}$ and by induction $J_k\le J_1=\mathbb{E}[M_s(A)]$. □

6.6. Proof of Corollary 1. The proof follows closely the one of Theorem 2. Suppose without loss of generality that the functions φ are non negative. Consider $\varepsilon>0$ and $\eta\in(0,1)$. According to (38), for all $\varphi\in\mathcal{F}$, and $t\in(0,1)$, there exists $A_t\in\mathcal{A}_t$ such that

$$\begin{aligned}|Z(\varphi)| &\le \varepsilon\sigma^2+\int_0^1\left\{|N(A_t)-\mathbb{E}[N(A_t)]|-2\varepsilon t\mathbb{E}[N(A_t)]\right\}\,\mathrm{d}t\\ &\le \varepsilon\sigma^2+\int_0^\eta |N(A_t)-\mathbb{E}[N(A_t)]|\,\mathrm{d}t+\int_\eta^1\left\{|N(A_t)-\mathbb{E}[N(A_t)]|-2\varepsilon t\mathbb{E}[N(A_t)]\right)\}\,\mathrm{d}t.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}}|Z(\varphi)|\right] \le \varepsilon\sigma^2+\int_0^\eta\mathbb{E}\left[\sup_{A_t\in\mathcal{A}_t}|N(A_t)-\mathbb{E}[N(A_t)]|\right]\mathrm{d}t\\ +\int_\eta^1\mathbb{E}\left[\sup_{A_t\in\mathcal{A}_t}\left\{|N(A_t)-\mathbb{E}[N(A_t)]|-2\varepsilon t\mathbb{E}[N(A_t)]\right\}\right]\mathrm{d}t.\end{aligned}\tag{53}$$

Let now $\xi>0$. As (35) holds true for all $A\in\mathcal{A}_t$, on an event $\Omega_{\xi,t}$ such that $\mathbb{P}\left[\Omega_{\xi,t}\right]\ge 1-e^{-n\xi}$, with $\alpha=c[\log_+|S_{\mathcal{A}_t}(2n)|+n\xi]$, $\beta=0$, we deduce from Claim 2 that for all $A_t\in\mathcal{A}_t$.

$$|N(A_t)-\mathbb{E}[N(A_t)]|-2\varepsilon t\mathbb{E}[N(A_t)]\le C\left\{\frac{\log_+|S_{\mathcal{A}_t}(2n)|+n\xi}{n}+\frac{\log_+|S_{\mathcal{A}_t}(2n)|+n\xi}{n\varepsilon t}\right\}\quad\text{on }\Omega_{\xi,t}.$$

Moreover, we derive from (41), that for all $A_t \in \mathcal{A}_t$,

$$|N(A_t) - \mathbb{E}[N(A_t)]| \le C \left[\sqrt{\frac{\log_+ |S_{\mathcal{A}_t}(2n)| + n\xi}{n}} \sqrt{\mathbb{E}[N(A_t)]} + \frac{\log_+ |S_{\mathcal{A}_t}(2n)| + n\xi}{n} \right] \quad \text{on } \Omega_{\xi,t}.$$

In these two inequalities, C is a universal constant.

We may integrate theses inequalities with respect to ξ to get

$$\mathbb{E}\left[\sup_{A_t \in \mathcal{A}_t} \{|N(A_t) - \mathbb{E}[N(A_t)]| - 2\varepsilon t \mathbb{E}[N(A_t)]\} \right] \le 2C \left[\frac{\log_+ |S_{\mathcal{A}_t}(2n)|}{n} + \frac{\log_+ |S_{\mathcal{A}_t}(2n)|}{n\varepsilon t} \right]$$

and

$$|N(A_t) - \mathbb{E}[N(A_t)]| \le 2C \left[\sqrt{\frac{\log_+ |S_{\mathcal{A}_t}(2n)|}{n}} \sqrt{\mathbb{E}[N(A_t)]} + \frac{\log_+ |S_{\mathcal{A}_t}(2n)|}{n} \right].$$

Set $\Gamma = \sup_{t\in(0,1)} \log_+ |S_{\mathcal{A}_t}(2n)|$. We may plug the two latter inequalities into (53) to get

$$\begin{aligned} \mathbb{E}\left[\sup_{\varphi \in \mathcal{F}} |Z(\varphi)| \right] &\le \varepsilon\sigma^2 + 2C\sqrt{\frac{\Gamma}{n}} \int_0^\eta \sqrt{\mathbb{E}[N(A_t)]}\, \mathrm{d}t + 2C\frac{\Gamma}{n} + 2C\frac{\Gamma \log(1/\eta)}{n\varepsilon} \\ &\le \varepsilon\sigma^2 + 2C\sqrt{\frac{\Gamma}{n}}\eta + 2C\frac{\Gamma}{n} + 2C\frac{\Gamma \log(1/\eta)}{n\varepsilon}. \end{aligned}$$

It remains to choose ε and η to prove the corollary, as it was done at the end of the proof of Theorem 2. □

6.7. Proof of Lemma 3. Let $\sqrt{q} = (\sqrt{f} + \sqrt{g})/2$. Then,

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

We complete the proof by using $h^2(f,g) \le 2h^2(s,f) + 2h^2(s,g)$. □

6.8. Proof of Proposition 3

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

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

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

$$f = \sum_{K \in m} P_K 1_K \quad \text{and} \quad g = \sum_{K \in m} Q_K 1_K,$$

where P_K and Q_K are non-negative polynomial functions of degree at most r. Let $R_K = P_K - tQ_K$. Now,

$$\{x \in \mathbb{R},\, g(x) > tf(x)\} = \bigcup_{K \in m} \{x \in K,\, R_K(x) > 0\}.$$

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

Finally, for all $K \in m$, $\{x \in K,\, R_K(x) > 0\}$ is always a union of at most $r/2 + 1$ intervals, which implies that $\{x \in \mathbb{R},\, g(x) > tf(x)\}$ is a union of at most $(r/2 + 1)(4\ell + 2)$ intervals. □

Proof of Assumption 1 for $S = \mathcal{P}_{\ell,r,+}$. Let $f, g \in \mathcal{P}_{\ell,r,+}$. As in the preceding proof, there exists a partition m such that $|m| \leq 4\ell + 2$ and on which

$$f = \sum_{K \in m} (P_K)_+ 1_K \quad \text{and} \quad g = \sum_{K \in m} (Q_K)_+ 1_K,$$

where P_K and Q_K are polynomial functions on K of degree at most r. Now,

$$\{x \in \mathbb{R},\, g(x) > tf(x)\} = \bigcup_{K \in m} \{x \in K,\, (P_K(x))_+ > t(Q_K(x))_+\}.$$

Let $A_K = \{x \in K, P_K(x) > 0\}$, $B_K = \{x \in K, Q_K(x) > 0\}$, $R_K = P_K - tQ_K$. Then,

$$\begin{aligned}\{x \in \mathbb{R},\, g(x) > tf(x)\} &= \bigcup_{K \in m} \{x \in A_K,\, P_K(x) > t(Q_K(x))_+\} \\ &= \bigcup_{K \in m} J_K,\end{aligned}$$

where

$$J_K = \{x \in A_K \cap B_K,\, R_K(x) > 0\} \cup (A_K \setminus B_K).$$

If $P_K = 0$, then $J_K = \emptyset$. If $R_K = 0$, then $P_K = tQ_K$, $A_K = B_K$ and thus $J_K = \emptyset$. If now $Q_K = 0$, then $R_K = P_K$ and $J_K = A_K$. In this case, we deduce from the same arguments as the ones developed in the preceding proof that J_K is a union of at most $r/2 + 1$ intervals.

We now assume that P_K, Q_K and R_K are not equal to 0. Let $\mathcal{X}$ be the set gathering the zeros of the three polynomial functions P_K, Q_K and R_K. If $\mathcal{X} = \emptyset$, the signs of P_K, Q_K and R_K do not vary on K and J_K may thus be either empty or the interval K. Suppose now that $\mathcal{X} \neq \emptyset$. Since P_K, Q_K and R_K are of degree at most r, and are not equal to 0, $|\mathcal{X}| \leq 3r$. We may write as $\mathcal{X} = \{b_1, \dots, b_k\}$ with $b_1 < b_2 < \cdots < b_k$. We define $b_0 = -\infty$ and $b_{k+1} = +\infty$. For all $j \in \{0, \dots, k\}$, the signs of P_K, Q_K and R_K are constant on (b_j, b_{j+1}). Therefore, the sets $\{x \in A_K \cap B_K,\, R_K(x) > 0\}$ and $\{x \in A_K \setminus B_K,\, P_K(x) > 0\}$ are unions of at most $3r+1$ intervals. The set J_K is hence a union of at most $6r+2$ intervals.

Finally, $\{x \in \mathbb{R},\, g(x) > tf(x)\} = \bigcup_{K \in m} J_K$ is a union of at most $(4\ell+2)(6r+2)$ intervals. □

Proof of Assumption 1 for $S = \mathcal{F}_k$. Let $f \in \mathcal{F}_k$ and $g \in \mathcal{F}_k \cap \mathcal{P}_{\ell,0}$. Let m_1 be a partition of $\mathbb{R}$ into intervals such that $|m_1| \leq 2k+1$ and such that f is monotone on each interval of m_1. Similarly, let m_2 be a partition of $\mathbb{R}$ into intervals such that $|m_2| \leq 2\ell+1$ such that g is constant on each interval of m_2. Set $m = \{K_1 \cap K_2,\, (K_1, K_2) \in m_1 \times m_2,\, K_1 \cap K_2 \neq \emptyset\}$. Then, m is a partition of $\mathbb{R}$ into intervals such that $|m| \leq |m_1| + |m_2| \leq 2k + 2\ell + 2$. Moreover, for all $K \in m$, f is monotone on K and g is constant on K. Now,

$$\{x \in \mathbb{R},\, g(x) > tf(x)\} = \bigcup_{K \in m} \{x \in K,\, g(x) > tf(x)\}.$$

As f is monotone on K and g is constant, the set $\{x \in K,\, g(x) > tf(x)\}$ is a (possibly empty) interval. Therefore, $\{x \in \mathbb{R},\, g(x) > tf(x)\}$ is a union of at most $|m|$ intervals. □

6.9. Proof of Theorem 4. Let for $d \geq 1$,

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

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

$$h^2(s, \hat{s}) \leq \inf_{f \in \bar{S}} \left\{ c_1 h^2(s, f) + c_2 \vartheta(d_S(f)) + c_3 \xi \log_+(1/\xi) \right\}. \tag{54}$$

We introduce the following notations. Let for $d \geq 1$, $\mathcal{I}_d$ be the class of unions of at most d intervals. Let $f, g \in \mathcal{S}$. Suppose that there exists $d \geq 1$ such that for all $t > 0$, the set $\{x \in \mathbb{R},\, g(x) > tf(x)\}$ belongs to $\mathcal{I}_d$. Then, $d_g(f)$ stands for any number d such that $\{x \in \mathbb{R},\, g(x) > tf(x)\}$ belongs to $\mathcal{I}_d$ (for all $t > 0$). If the preceding assumption does not hold, we set $d_g(f) = +\infty$.

We define for $d \geq 2$,

$$\mathscr{G}_d = \{\psi(g/f),\, g \in \mathcal{S}, f \in \mathcal{S},\, d_g(f) = d-1\}.$$

We shall apply Theorem 2 to the class $\mathcal{F} = \mathscr{G}_d$.

We begin with the following elementary claim:

Claim 4. *We have,*

- *For all $J \in \mathcal{I}_d$, $\mathbb{R} \setminus J \in \mathcal{I}_{d+1}$.*
- *For all non-increasing sequence $(J_n)_{n \geq 1}$ of $\mathcal{I}_d$, $\bigcap_{n \geq 1} J_n$ belongs to $\mathcal{I}_d$.*

The set $\mathscr{G}_d$ enjoys the following properties:

Claim 5. *The functions $\varphi \in \mathscr{G}_d$ satisfy $|\varphi| \leq 1$. Moreover, for all $t \in (0,1)$, $\varphi \in \mathscr{G}_d$, $\{x \in \mathbb{R},\ \varphi_+(x) > t\} \in \mathcal{I}_{d-1}$ and $\{x \in \mathbb{R},\ \varphi_-(x) > t\} \in \mathcal{I}_d$.*

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

$$\begin{aligned}\{x \in \mathbb{R},\ \varphi_+(x) > t\} &= \{x \in \mathbb{R},\ \psi_+(g(x)/f(x)) > t\}\\ &= \{x \in \mathbb{R},\ f(x) \neq 0,\ \psi_+(g(x)/f(x)) > t\} \cup \{x \in \mathbb{R},\ f(x) = 0,\ g(x) > 0\}\\ &= \{x \in \mathbb{R},\ f(x) \neq 0,\ g(x) > uf(x)\} \cup \{x \in \mathbb{R},\ f(x) = 0,\ g(x) > 0\},\end{aligned}$$

where $u = \psi^{-1}(t)$. Therefore,

$$\{x \in \mathbb{R},\ \varphi_+(x) > t\} = \{x \in \mathbb{R},\ g(x) > uf(x)\} \in \mathcal{I}_{d-1},$$

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

$$\{x \in \mathbb{R},\ \varphi_-(x) > t\} = \{x \in \mathbb{R},\ \psi_+(f(x)/g(x)) > t\}.$$

By exchanging the role of f and g in the above computations, we derive

$$\begin{aligned}\{x \in \mathbb{R},\ \varphi_-(x) > t\} &= \{x \in \mathbb{R},\ f(x) > ug(x)\}\\ &= \{x \in \mathbb{R},\ g(x) < (1/u)f(x)\}.\end{aligned}$$

By using the first point of Claim 4, for all $n \geq 1$,

$$\{x \in \mathbb{R},\ g(x) \leq (1/u + 1/n)f(x)\} \in \mathcal{I}_d.$$

Yet,

$$\{x \in \mathbb{R},\ g(x) < (1/u)f(x)\} = \bigcap_{n=1}^{\infty} \{x \in \mathbb{R},\ g(x) \leq (1/u + 1/n)f(x)\}.$$

The second point of Claim 4 ensures that $\{x \in \mathbb{R},\ g(x) < (1/u)f(x)\}$ belongs to $\mathcal{I}_d$, which completes the proof. □

The lemma below is at the core of the proof of Theorem 4.

Lemma 9. *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 $\varepsilon \in (0, 1/9)$, $f, g \in \mathcal{S}$,*

$$T(f,g) \leq 3(1+\varepsilon)h^2(s,f) - \frac{1-9\varepsilon}{3}h^2(s,g) + c_1\vartheta(d_g(f)) + c_2\xi\log_+(1/\xi). \tag{55}$$

In the above inequality, c_1 and c_2 only depend on ε. Besides, we use the conventions $\vartheta(+\infty) = +\infty$ when $d_g(f) = \infty$.

Proof. Let $d \geq 2$. Theorem 2 shows the existence of an event $\Omega_\xi(d)$ such that $\mathbb{P}[\Omega_\xi(d)] \geq 1 - e^{-n\xi}$ and on which: for all $\varepsilon > 0$, $\varphi \in \mathscr{G}_d$ of the form $\varphi = \psi(g/f)$, with $f, g \in \mathcal{S}$,

$$|Z(\varphi)| \leq \varepsilon\upsilon(\varphi) + c\left[\vartheta\left(d_g(f)+1\right) + \xi\log_+(1/\xi)\right].$$

In this inequality, c only depends on ε. Since

$$\vartheta_1(d_g(f)+1) \leq 2\vartheta_1(d_g(f)),$$

we get

$$|Z(\varphi)| \leq \varepsilon\upsilon(\varphi) + 2c\vartheta\left(d_g(f)\right) + c\xi\log_+(1/\xi).$$

Let $\Omega_\xi = \bigcap_{d=2}^{\infty} \Omega_{\xi+(2\log(1+d))/n}(d)$. Then,

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

Moreover, on Ω_ξ: for all $f, g \in \mathcal{S}$, $\varphi = \psi(g/f)$ such that $d_g(f) < \infty$,

$$\begin{aligned}
|Z(\varphi)| &\le \varepsilon \upsilon(\varphi) + 2c\vartheta(d_g(f)) + c\left[\left(\xi + \frac{2\log(1+d_g(f))}{n}\right)\log_+\left(\frac{1}{\xi + \frac{2\log(1+d_g(f))}{n}}\right)\right] \\
&\le \varepsilon \upsilon(\varphi) + 2c\vartheta(d_g(f)) + \frac{2c\log(1+d_g(f))}{n}\log_+\left(\frac{n}{2\log(1+d_g(f))}\right) + c\xi\log_+(1/\xi) \\
&\le \varepsilon \upsilon(\varphi) + c'\vartheta(d_g(f)) + c\xi\log_+(1/\xi), \qquad (56)
\end{aligned}$$

where c' only depends on ε. This last inequality remains true when $d_g(f) = \infty$ using the convention $\vartheta(+\infty) = +\infty$.

Now, it follows from (3) that for all $f, g \in \mathcal{S}$,

$$T(f,g) \le 3h^2(s,f) - \frac{1}{3}h^2(s,g) + Z(\psi(g/f)). \qquad (57)$$

Therefore, we deduce from Lemma 3 and from (56) that on Ω_ξ: for all $f, g \in \mathcal{S}$,

$$T(f,g) \le 3(1+\varepsilon)h^2(s,f) - \frac{1-9\varepsilon}{3}h^2(s,g) + c'\vartheta_1(d_g(f)) + c\xi\log_+(1/\xi),$$

which proves (55) with $c_1 = c'$ and $c_2 = c$. □

We now finish the proof of Theorem 4. Assumption 1 says that we can define $d_g(f)$ by $d_g(f) = d_S(f)$ for all $f \in \bar{S}$, $g \in S$. Lemma 9 implies that on Ω_ξ: for all $f \in \bar{S}$, $g \in S$,

$$T(f,g) \le 3(1+\varepsilon)h^2(s,f) - \frac{1-9\varepsilon}{3}h^2(s,g) + c_1\vartheta(d_S(f)) + c_2\xi\log_+(1/\xi). \qquad (58)$$

Thus, for all $f \in \bar{S}$,

$$\gamma(f) \le 3(1+\varepsilon)h^2(s,f) - \frac{1-9\varepsilon}{3}h^2(s,S) + c_1\vartheta(d_S(f)) + c_2\xi\log_+(1/\xi). \qquad (59)$$

By using $T(f,g) = -T(g,f)$, we deduce from (58) that for all $f \in \bar{S}$, $g \in S$,

$$\frac{1-9\varepsilon}{3}h^2(s,g) - 3(1+\varepsilon)h^2(s,f) - c_1\vartheta(d_S(f)) - c_2\xi\log_+(1/\xi) \le T(g,f).$$

Any ρ-estimator $\hat{s}$ satisfies on Ω_ξ: for all $f \in \bar{S}$,

$$\begin{aligned}
\frac{1-9\varepsilon}{3}h^2(s,\hat{s}) - 3(1+\varepsilon)h^2(s,f) - c_1\vartheta(d_S(f)) - c_2\xi\log_+(1/\xi) &\le T(\hat{s},f) \\
&\le \gamma(\hat{s}) \\
&\le \gamma(f) + 1/n. \qquad (60)
\end{aligned}$$

Using now (59) and $1/n \le \vartheta(d_S(f))$, we deduce when $\varepsilon \in (0, 1/9)$,

$$h^2(s,\hat{s}) \le \inf_{f\in\bar{S}}\left\{c_{1,\varepsilon}h^2(s,f) - h^2(s,S) + c_{2,\varepsilon}\vartheta(d_S(f)) + c_{2,\varepsilon}\xi\log_+(1/\xi)\right\} \quad \text{with } c_{1,\varepsilon} = 18\frac{1+\varepsilon}{1-9\varepsilon},$$

and with $c_{2,\varepsilon}$ depending only on ε. We now choose ε arbitrarily among $(0,1/9)$ to prove the theorem. □

6.10. Proof of Proposition 5. It follows from Lemma 9 that there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which: for all $f \in \bar{S}$,

$$\frac{1-9\varepsilon}{3} h^2(s,\hat{s}) - 3(1+\varepsilon)h^2(s,f) - c_1\vartheta(d_S(f)) - c_2\xi\log_+(1/\xi) \leq T(\hat{s},f),$$

where c_1, c_2 depend only on ε. Now $\gamma(\hat{s}) = 0$ and hence $T(\hat{s},f) \leq 0$. This leads to

$$h^2(s,\hat{s}) \leq c_{1,\varepsilon} h^2(s,f) + c_{2,\varepsilon}\left[\vartheta(d_S(f)) + \xi\log_+(1/\xi)\right], \quad \text{with } c_{1,\varepsilon} = 9\frac{1+\varepsilon}{1-9\varepsilon}.$$

□

6.11. Proof of Proposition 6. The maximum likelihood estimator $\hat{s}$ converges almost surely to

$$\bar{s}_1 = p1_{[0,1]} + (1-p)1_{[1,2]}.$$

Therefore $h^2(s_{p,\varepsilon},\hat{s})$ converges a.s. to $h^2(s_{p,\varepsilon},\bar{s}_1)$. Define now

$$\bar{s}_2 = \frac{p\varepsilon}{p\varepsilon+1-p}1_{[0,1]} + \frac{1-p}{p\varepsilon+1-p}1_{[1,2]} \in S.$$

We may verify that

$$h^2(s_{p,\varepsilon},\bar{s}_1) = p(1-\sqrt{\varepsilon}) \quad \text{and} \quad h^2(s_{p,\varepsilon},\bar{s}_2) = 1 - \sqrt{1-p(1-\varepsilon)}.$$

For all $\eta \in (1,2)$, there exist p,ε such that $h^2(s_{p,\varepsilon},\bar{s}_1)/h^2(s_{p,\varepsilon},\bar{s}_2) > \eta$, and thus,

$$\lim_{n\to+\infty} h^2(s_{p,\varepsilon},\hat{s}) \geq \eta h^2(s_{p,\varepsilon},\bar{s}_2) \geq \eta h^2(s_{p,\varepsilon},S) \quad \text{almost surely.}$$

□

6.12. Proof of Theorem 7. Let $\varepsilon \in (0,1/9)$. Lemma 9 asserts the existence of an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which: for all $f \in \bar{\hat{S}}$, $g \in \hat{S}$,

$$T(f,g) \leq 3(1+\varepsilon)h^2(s,f) - \frac{1-9\varepsilon}{3}h^2(s,g) + c_1\vartheta(d_{\hat{S}}(f)) + c_2\log_+(1/\xi)\xi. \tag{61}$$

As $T(g,f) = -T(f,g)$, we also have,

$$T(g,f) \geq \frac{1-9\varepsilon}{3}h^2(s,g) - 3(1+\varepsilon)h^2(s,f) - c_1\vartheta(d_{\hat{S}}(f)) - c_2\log_+(1/\xi)\xi. \tag{62}$$

In the above inequalities, c_1, c_2 only depend on ε. Therefore, we deduce from (62) that on Ω_ξ: for all $f \in \hat{S} \cap \bar{\hat{S}}$,

$$\begin{aligned} \frac{1-9\varepsilon}{3}h^2(s,\hat{s}) - 3(1+\varepsilon)h^2(s,f) - c_1\vartheta(d_{\hat{S}}(f)) - c_2\log_+(1/\xi)\xi &\leq T(\hat{s},f) \\ &\leq \gamma_2(\hat{s}) \\ &\leq \gamma_2(f) + 1/n. \end{aligned} \tag{63}$$

By using (61), and $1/n \leq \vartheta(d_{\hat{S}}(f))$,

$$\begin{aligned}\gamma_2(f) &\leq 3(1+\varepsilon)h^2(s,f) - \frac{1-9\varepsilon}{3}\inf_{g\in\hat{S}}\{h^2(s,g)\} + 2c_1\vartheta(d_{\hat{S}}(f)) + c_2\log_+(1/\xi)\xi \\ &\leq 3(1+\varepsilon)h^2(s,f) + 2c_1\vartheta(d_{\hat{S}}(f)) + c_2\log_+(1/\xi)\xi.\end{aligned}$$

Plugging this last inequality into (63) leads to the result. □

6.13. Proofs of Theorems 8 and 9. The two procedures carried out in Sections 4.2and 4.4 are particular cases of a more general selection rule we now describe. Theorems 8 and 9 follow from Theorem 12 below. Their proofs are given in Sections 6.13.2 and 6.13.3.

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

We consider for each $\lambda \in \hat{\Lambda}$ a (possibly random) model $\hat{S}_\lambda \subset \mathcal{S}$. We associate to each $\lambda \in \hat{\Lambda}$, $\hat{g} \in \hat{S}_\lambda$, two penalty terms $\mathrm{pen}_{1,\lambda}(\hat{g})$ and $\mathrm{pen}_2(\lambda)$. We finally define the criterion γ_5 by

$$\gamma_5(\hat{s}_\lambda) = \sup_{\hat{g}\in\hat{S}_\lambda}\left[T(\hat{s}_\lambda,\hat{g}) - \mathrm{pen}_{1,\lambda}(\hat{g})\right].$$

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

$$\gamma_5(\hat{s}_{\hat{\lambda}}) + 2\mathrm{pen}_2(\hat{\lambda}) \leq \inf_{\lambda\in\hat{\Lambda}}\{\gamma_5(\hat{s}_\lambda) + 2\mathrm{pen}_2(\lambda)\} + 1/n.$$

The risk of this estimator is bounded above as follows.

Theorem 12. *We assume that there exist two real valued maps, $\Delta \geq 0$ on $\hat{\Lambda}$, and $d \geq 1$ on $\mathcal{S}$ such that*

$$d_{\hat{s}_\lambda}(\hat{g}) \leq d(\hat{g}) + \Delta(\lambda) \quad \textit{for all } \lambda \in \hat{\Lambda},\ \hat{g} \in \hat{S}_\lambda. \tag{64}$$

We suppose that there exist a (possibly random) model $\hat{S} \subset \bigcap_{\lambda\in\hat{\Lambda}} \hat{S}_\lambda$ and a map pen_1 on $\hat{S}$ such that

$$\mathrm{pen}_{1,\lambda}(\hat{g}) \leq \mathrm{pen}_1(\hat{g}) + \mathrm{pen}_2(\lambda) \quad \textit{for all } \hat{g} \in \hat{S},\ \lambda \in \hat{\Lambda}. \tag{65}$$

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

$$\begin{aligned}\mathrm{pen}_{1,\lambda}(\hat{g}) &\geq L_1\frac{d(\hat{g})}{n}\log_+^2\left(\frac{n}{d(\hat{g})}\right) \\ \mathrm{pen}_1(\hat{f}) &\geq L_1\frac{d(\hat{f})}{n}\log_+^2\left(\frac{n}{d(\hat{f})}\right) \\ \mathrm{pen}_2(\lambda) &\geq L_1\frac{\Delta(\lambda)}{n}\log_+^2\left(\frac{n}{\Delta(\lambda)}\right),\end{aligned} \tag{66}$$

then, for all $\xi > 0$,

$$\mathbb{P}^\star\left[h^2(s,\hat{s}_{\hat{\lambda}}) \geq c\left(\inf_{\lambda\in\hat{\Lambda}}\left\{h^2(s,\hat{s}_\lambda) + \mathrm{pen}_2(\lambda)\right\} + \inf_{\hat{g}\in\hat{S}}\left\{h^2(s,\hat{g}) + \mathrm{pen}_1(\hat{g})\right\} + \xi\log_+(1/\xi)\right)\right] \leq e^{-n\xi}.$$

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

Remark: we recall that the notation $d_g(f)$ appearing in (64) is defined at the beginning of the proof of Theorem 4.

6.13.1. Proof of Theorem 12. Let $\varepsilon \in (0, 1/9)$. We deduce from Lemma 9 and from the equality $T(g,f) = -T(f,g)$ that there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which: for all $f, g \in \mathcal{S}$, such that $d_g(f) < +\infty$,

$$T(f,g) \leq 3(1+\varepsilon)h^2(s,f) - \frac{1-9\varepsilon}{3}h^2(s,g) + c_1\vartheta(d_g(f)) + c_2\xi\log_+(1/\xi) \tag{67}$$

$$T(g,f) \geq \frac{1-9\varepsilon}{3}h^2(s,g) - 3(1+\varepsilon)h^2(s,f) - c_1\vartheta(d_g(f)) - c_2\xi\log_+(1/\xi), \tag{68}$$

where c_1, c_2 only depend on ε.

Let $\lambda \in \hat{\Lambda}$ and $\hat{g} \in \hat{S}_\lambda$. Note that we may use (67) with $f = \hat{s}_\lambda$, $g = \hat{g}$ and $d_{\hat{s}_\lambda}(\hat{g}) = d(\hat{g}) + \Delta(\lambda)$ (we may always increase $d_{\hat{s}_\lambda}(\hat{g})$). We get for all $\lambda \in \hat{\Lambda}$, $\hat{g} \in \hat{S}_\lambda$,

$$\begin{aligned} T(\hat{s}_\lambda, \hat{g}) &\leq 3(1+\varepsilon)h^2(s,\hat{s}_\lambda) - \frac{1-9\varepsilon}{3}h^2(s,\hat{g}) + c_1\vartheta\left(d(\hat{g}) + \Delta(\lambda)\right) + c_2\xi\log_+(1/\xi) \\ &\leq 3(1+\varepsilon)h^2(s,\hat{s}_\lambda) + c_1\vartheta(d(\hat{g}) + \Delta(\lambda)) + c_2\xi\log_+(1/\xi) \\ &\leq 3(1+\varepsilon)h^2(s,\hat{s}_\lambda) + c_1\vartheta(d(\hat{g})) + c_1\vartheta(\Delta(\lambda)) + c_2\xi\log_+(1/\xi). \end{aligned}$$

In this inequality, we use the convention explained in the theorem when $\Delta(\lambda) = 0$. If L_1 is large enough,

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

and hence

$$\gamma_5(\hat{s}_\lambda) \leq 3(1+\varepsilon)h^2(s,\hat{s}_\lambda) + \mathrm{pen}_2(\lambda) + c_2\xi\log_+(1/\xi). \tag{69}$$

We derive from (68) with $g = \hat{s}_\lambda$, $f = \hat{g}$,

$$T(\hat{s}_\lambda, \hat{g}) \geq \frac{1-9\varepsilon}{3}h^2(s,\hat{s}_\lambda) - 3(1+\varepsilon)h^2(s,\hat{g}) - c_1\vartheta\left(d(\hat{g}) + \Delta(\lambda)\right) - c_2\xi\log_+(1/\xi).$$

Therefore, using this inequality with $\lambda = \hat{\lambda}$, we get for all $\hat{g} \in \hat{S}_{\hat{\lambda}}$,

$$\begin{aligned} \frac{1-9\varepsilon}{3}h^2(s,\hat{s}_{\hat{\lambda}}) &\leq T(\hat{s}_{\hat{\lambda}}, \hat{g}) + 3(1+\varepsilon)h^2(s,\hat{g}) + c_1\vartheta(d(\hat{g}) + \Delta(\hat{\lambda})) + c_2\xi\log_+(1/\xi) \\ &\leq T(\hat{s}_{\hat{\lambda}}, \hat{g}) - \mathrm{pen}_{1,\hat{\lambda}}(\hat{g}) + 3(1+\varepsilon)h^2(s,\hat{g}) + c_1\vartheta(d(\hat{g})) + \mathrm{pen}_{1,\hat{\lambda}}(\hat{g}) \\ &\quad + c_1\vartheta(\Delta(\hat{\lambda})) + c_2\xi\log_+(1/\xi). \end{aligned}$$

If L_1 is large enough, $\mathrm{pen}_{1,\hat{\lambda}}(\hat{g}) \geq 2\vartheta(d(\hat{g}))$, $\mathrm{pen}_2(\hat{\lambda}) \geq 4c_1\vartheta(\Delta(\hat{\lambda}))$, $\mathrm{pen}_{1,\hat{\lambda}}(\hat{g}) \geq 4/n$, and hence,

$$\begin{aligned} \frac{1-9\varepsilon}{3}h^2(s,\hat{s}_{\hat{\lambda}}) \leq T(\hat{s}_{\hat{\lambda}}, \hat{g}) - \mathrm{pen}_{1,\hat{\lambda}}(\hat{g}) + 3(1+\varepsilon)h^2(s,\hat{g}) + \frac{3}{2}\mathrm{pen}_{1,\hat{\lambda}}(\hat{g}) + \frac{1}{2}\mathrm{pen}_2(\hat{\lambda}) \\ + c_2\xi\log_+(1/\xi) - 1/n. \end{aligned}$$

This inequality being true for all $\hat{g} \in \hat{S}_{\hat{\lambda}}$, we get

$$\begin{aligned} \frac{1-9\varepsilon}{3}h^2(s,\hat{s}_{\hat{\lambda}}) \leq \gamma_5(\hat{s}_{\hat{\lambda}}) + \frac{1}{2}\mathrm{pen}_2(\hat{\lambda}) + \inf_{\hat{g}\in\hat{S}_{\hat{\lambda}}}\left\{3(1+\varepsilon)h^2(s,\hat{g}) + \frac{3}{2}\mathrm{pen}_{1,\hat{\lambda}}(\hat{g})\right\} \\ + c_2\xi\log_+(1/\xi) - 1/n. \end{aligned}$$

In particular, as $\hat{S} \subset \hat{S}_{\hat{\lambda}}$,

$$\frac{1-9\varepsilon}{3}h^2(s,\hat{s}_{\hat{\lambda}}) \le \gamma_5(\hat{s}_{\hat{\lambda}}) + \frac{1}{2}\mathrm{pen}_2(\hat{\lambda}) + \inf_{\hat{g}\in\hat{S}}\left\{3(1+\varepsilon)h^2(s,\hat{g}) + \frac{3}{2}\mathrm{pen}_{1,\hat{\lambda}}(\hat{g})\right\} + c_2\xi\log_+(1/\xi) - 1/n.$$

We deduce from (65),

$$\frac{1-9\varepsilon}{3}h^2(s,\hat{s}_{\hat{\lambda}}) \le \gamma_5(\hat{s}_{\hat{\lambda}}) + 2\mathrm{pen}_2(\hat{\lambda}) + \inf_{\hat{g}\in\hat{S}}\left\{3(1+\varepsilon)h^2(s,\hat{g}) + \frac{3}{2}\mathrm{pen}_1(\hat{g})\right\} + c_2\xi\log_+(1/\xi) - 1/n.$$

By using the definition of $\hat{\lambda}$ and (69), we deduce that for all $\lambda \in \hat{\Lambda}$,

$$\begin{aligned}\frac{1-9\varepsilon}{3}h^2(s,\hat{s}_{\hat{\lambda}}) &\le \gamma_5(\hat{s}_\lambda) + 2\mathrm{pen}_2(\lambda) + \inf_{\hat{g}\in\hat{S}}\left\{3(1+\varepsilon)h^2(s,\hat{g}) + \frac{3}{2}\mathrm{pen}_1(\hat{g})\right\} + 2c_2\xi\log_+(1/\xi)\\ &\le 3(1+\varepsilon)h^2(s,\hat{s}_\lambda) + 3\mathrm{pen}_2(\lambda) + \inf_{\hat{g}\in\hat{S}}\left\{3(1+\varepsilon)h^2(s,\hat{g}) + \frac{3}{2}\mathrm{pen}_1(\hat{g})\right\}\\ &\quad + 2c_2\xi\log_+(1/\xi).\end{aligned}$$

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

6.13.2. Proof of Theorem 8. We shall apply the selection rule developed in Section 6.13 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 set $\hat{S} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$, and for $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$,

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

Note that the assumption $\hat{S} \subset \bigcap_{m\in\widehat{\mathcal{M}}_{\hat{\ell}}}\hat{S}_m$ of Theorem 12 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 (27). A function $\hat{g} \in \hat{S}_m$ of the form $\hat{g} = \sum_{K\in m}\hat{s}_{m_K}1_K$ is piecewise constant. In the sequel, $m(\hat{g})$ designs a partition of $\widehat{\mathcal{M}}$ of the form

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

with minimal length that is such that

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

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 $f \in \bar{S}$ and $k \ge 1$ be the smallest integer for which $f \in \mathcal{P}_{k,r}$. It follows from Proposition 3 that one may define $d_{\mathcal{P}_{\hat{\ell}\vee k,r}}(f) = (r+2)(2(\hat{\ell}\vee k)+1)$. In particular, for all $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ and $f \in \bar{S}$, we may set since $\hat{s}_m \in \mathcal{P}_{\hat{\ell},r}$,

$$d_{\hat{s}_m}(f) = (r+2)(2\inf_{\substack{k\ge1\\ \mathcal{P}_{k,r}\ni f}}(\hat{\ell}\vee k)+1).$$

We now define d for $f \in \bar{S}$ and Δ for $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ by

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

We define d arbitrarily when $f \notin \bar{S}$. Note that (64) 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 $\hat{g} \in \hat{S}_m$ by

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

The first penalty satisfies the lower bound (66) since

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

It remains to define $\mathrm{pen}_1(\hat{g})$ for $\hat{g} \in \hat{S} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$. We need the claim below whose proof is deferred after the present proof.

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

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 (65).

Note now that the selection rules described in Sections 6.13 and 4.2 coincide. Theorem 12 controls the risk of the selected estimator: for all $\xi > 0$, with probability larger than $1 - e^{-n\xi}$,

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

where C is a universal constant.

It remains to use the definition of the penalty terms to finish the proof. □

Proof of Claim 6. We have,

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

Since m and m' are partitions into intervals, we deduce that $|m(\hat{s}_{m'})| \leq |m| + |m'|$. □

6.13.3. Proof of Theorem 9. The proof is almost the same than the one of Theorem 8. 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 6.13:

$$\hat{S} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{k},lower}\}$$

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

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

□

6.14. Proof of Lemma 4. Note that if $K \cap \{Y_{(1)}, \ldots, Y_{(\hat{n})}\} = \emptyset$ then,

$$L_K(f) = -\int_K f(t)\,\mathrm{d}M(t),$$

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

Let V be the Radon–Nikodym derivative of M with respect to the Lebesgue measure μ. Then, $V = 1$ in framework 1, $V(t) = n^{-1}\sum_{i=1}^n 1_{X_i\geq t} 1_{[0,+\infty)}(t)$ in framework 2 and $V(t) = n^{-1}\sum_{i=1}^n 1_{X_{t^-}^{(i)}=1} 1_{(0,+\infty)}(t)$ in framework 3. Let k be the largest integer of $\{1,\ldots,\hat{n}\}$ 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 $V(t) \geq 1/n$ for all $t \in (Y_{(k)} - \alpha, Y_{(k)})$.

Let now $f \in \mathcal{P}_r(K)$. Then, $L_K(f)$ takes the form

$$L_K(f) = \frac{1}{n}\sum_{i\in\hat{I}} (\log f(Y_i))\, 1_K(Y_i) - \int_K f(t)V(t)\,\mathrm{d}t,$$

and is bounded above by

$$L_K(f) \leq \log_+\left(\sup_{t\in K'} f(t)\right) - \frac{1}{n}\int_{Y_{(k)}-\alpha}^{Y_{(k)}} f(t)\,\mathrm{d}t.$$

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

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

There exists C such that $\|f\|_\infty \leq C\|f\|_1$ for all $f \in \mathcal{P}_r(K)$. Now,

$$L_K(f) \leq \log_+(C\|f\|_1) - \frac{\|f\|_1}{n}.$$

The continuous map L_K tends therefore to $-\infty$ when $\|f\|_1 \to +\infty$, which proves the existence of $\hat{s}_K$.

For the second part of the lemma, we use Theorem 1 to deduce that $T(\hat{s}_K, f_K) \leq 0$ for all $f_K \in \mathcal{P}_r(K)$. If $f \in \mathcal{P}_r(m)$ is of the form $f = \sum_{K\in m} f_K$,

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

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

6.15. Proof of Lemma 5. The following claim will be useful in the sequel.

Claim 7. *Let $\xi > 0$, $\eta \geq 0$, $r \geq 0$, and m, m' be two finite (non-empty) collections of disjoint intervals of $\mathbb{R}$. There exists an event Ω_ξ that only depends on ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$, and on which the following holds: for all piecewise polynomial estimators $\hat{s}_m \in \mathcal{P}_r(m)$, $\hat{s}_{m'} \in \mathcal{P}_r(m')$ such that $|m'| \leq 2|m| + 1$ and such that $T(\hat{s}_m, \hat{s}_{m'}) \geq -\eta$,*

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

where C is universal.

Proof. Let $\varepsilon = 1/18$. We deduce from Lemma 9 that there exists an event Ω_ξ that only depends on ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which:

$$T(\hat{s}_m, \hat{s}_{m'}) \leq 3(1+\varepsilon)h^2(s, \hat{s}_m) - \frac{1-9\varepsilon}{3}h^2(s, \hat{s}_{m'}) + c_1\vartheta(d_{\hat{s}_{m'}}(\hat{s}_m)) + c_2\xi\log_+(1/\xi), \tag{70}$$

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, $|m''| \leq |m| + |m'| \leq 3|m| + 1$. Thereby, $\hat{s}_m$ and $\hat{s}_{m'}$ belong to $\mathcal{P}_{3|m|+1,r}$ and it follows from Proposition 3 that we may set

$$d_{\hat{s}'_m}(\hat{s}_m) = (r+2)(2(3|m|+1)+1).$$

We now bound above $\vartheta(d_{\hat{s}'_m}(\hat{s}_m))$ in (70), and then use $T(\hat{s}_m, \hat{s}_{m'}) \geq -\eta$ to get the result. □

We recall that $\mathcal{M}$ stands for the class of finite (non-empty) collections m of disjoint intervals K of $\mathbb{R}$ that are right-closed and not reduced to a singleton. Let $m \in \mathcal{M}$ and $\hat{s}_m$ be the ρ-estimator defined in Lemma 4. This estimator is of the form

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

The claim below shows the existence of a partition $m' \in \widehat{\mathcal{M}}$ and a ρ-estimator $\hat{s}_{m'}$ which performs as well as $\hat{s}_m$.

Claim 8. *Let $\xi > 0$, $m \in \mathcal{M}$ and $\hat{s}_m$ be the ρ-estimator defined above. There exist an event Ω_ξ that only depends on ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$, a partition $m' \in \widehat{\mathcal{M}}$ such that $|m'| \leq 4|m| - 3$, and a piecewise polynomial ρ-estimator $\hat{s}_{m'}$ defined as in Lemma 4 such that on Ω_ξ:*

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

where C is universal.

Proof. We first suppose that

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

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 $m_1 = \{K_j,\, j \in \{1, \dots, \ell\}\}$ where $1 \leq \ell \leq |m|$ and where K_j is an interval with endpoints a_j, b_j satisfying $a_1 < b_1 \leq a_2 < b_2 < \dots$. When $K \in m$ does not belong to m_1, $\hat{s}_K = 0$ and hence

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

For each $j \in \{1, \dots, \ell\}$, 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 $j \in \{2, \dots, \ell - 1\}$, $J_{2j} = (\beta_j, \alpha_{j+1}]$ and for $j \in \{2, \dots, \ell\}$, $J_{2j-1} = (\alpha_j, \beta_j]$. 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 $j \in \{1, \dots, \ell\}$. The collection $m' = \{J_j, j \in \{1, \dots, 2\ell - 1\}\}$ defines a partition belonging to $\widehat{\mathcal{M}}$ such that $|m'| \leq 2\ell - 1$. We define the ρ-estimator

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

where $\hat{s}_A$ maximizes L_A 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

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

Note that $\tilde{s}_{m'}$ also belongs to the random model $\mathcal{P}_r(m')$ and hence $T(\hat{s}_{m'}, \tilde{s}_{m'}) \leq 0$. We deduce from Claim 7 that there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which:

$$h^2(s, \hat{s}_{m'}) \leq C\left\{h^2(s, \tilde{s}_{m'}) + \frac{(r+1)|m|}{n} \log_+^2\left(\frac{n}{(r+1)|m|}\right) + \xi \log_+(1/\xi)\right\}, \tag{71}$$

where C is universal.

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

$$T\left(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} 1_{J_{2j-1}}\right) \leq 0, \tag{72}$$

since $\hat{s}_{J_{2j-1}}$ maximizes $L_{J_{2j-1}}$ over $\mathcal{P}_r(J_{2j-1})$ and that $\hat{s}_{K_j} 1_{J_{2j-1}} \in \mathcal{P}_r(J_{2j-1})$. When $J_{2j-1} = \emptyset$, $T\left(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} 1_{J_{2j-1}}\right) = 0$, and thus (72) also holds. We define

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

We deduce from (72) that $T(\tilde{s}_{m'} 1_A, \hat{s}_m 1_A) \leq 0$. Therefore,

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

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

$$T(\tilde{s}_{m'}, \hat{s}_m) \leq \int_{B \cap A^c} \psi\left(\hat{s}_m/0\right) \, \mathrm{d}N \leq N(B \cap A^c). \tag{73}$$

Now,

$$\begin{aligned} N(B \cap A^c) &= N(B) - \sum_{j=1}^{\ell} N((\alpha_j, \beta_j]) \\ &= N(B) - \sum_{j=1}^{\ell} N([\alpha_j, \beta_j]) + \sum_{j=1}^{\ell} N(\{\alpha_j\}) \\ &= N(B) - \sum_{j=1}^{\ell} N(K_j) + \sum_{j=1}^{\ell} N(\{\alpha_j\}) \\ &= \sum_{j=1}^{\ell} N(\{\alpha_j\}). \end{aligned}$$

In each of the frameworks, $N(\{\alpha_j\}) \leq 1/n$. We then deduce from (73) that

$$T(\tilde{s}_{m'}, \hat{s}_m) \leq \ell/n.$$

We deduce from Claim 7 with $\eta = \ell/n \leq |m|/n$ that on Ω_ξ:

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

where C' is universal. By plugging this inequality into (71), we derive that

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

where C'' is universal, which proves the claim when

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

If this equality does not hold, we define the set $\mathcal{X}$ gathering the $Y_{(i)}$ that do not belong to $\bigcup_{K \in m} K$. Then, there exist $k \in \{1, \dots, |m| + 1\}$ and k disjoint intervals $I_1, \dots, I_k$ that are right-closed and not reduced to a singleton such that

$$\mathcal{X} = \bigcup_{j=1}^{k} I_j \cap \left\{ Y_{(1)}, \dots, Y_{(\hat{n})} \right\},$$

and such that $K \cap I_j = \emptyset$ for all $K \in m$, $j \in \{1, \dots, k\}$. If $m_{new} = m \bigcup \cup_{j=1}^{k} K_j$, and if $\hat{s}_{m_{new}}$ is a ρ-estimator on $\mathcal{P}_r(m_{new})$ defined as in Lemma 4, $T(\hat{s}_m, \hat{s}_{m_{new}}) \geq 0$ and hence

$$h^2(s, \hat{s}_{m_{new}}) \leq C''' \left\{ h^2(s, \hat{s}_m) + \frac{(r+1)|m|}{n} \log_+^2 \left(\frac{n}{(r+1)|m|} \right) + \xi \log_+(1/\xi) \right\}. \tag{74}$$

Moreover,

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

and we may thus apply the same arguments as before by replacing m by m_{new} to get $m' \in \widehat{\mathcal{M}}$ such that $|m'| \leq 2|m_{new}| - 1 \leq 4|m| - 3$ and such that

$$h^2(s, \hat{s}_{m'}) \leq C'''' \left\{ h^2(s, \hat{s}_{m_{new}}) + \frac{(r+1)|m|}{n} \log_+^2 \left(\frac{n}{(r+1)|m|} \right) + \xi \log_+(1/\xi) \right\}. \tag{75}$$

Putting (74) and (75) together ends the proof. □

We return to the proof of Lemma 5. Note that the event Ω_ξ appearing in Claim 8 is defined in Lemma 9. Since $\hat{s}_m$ is a ρ-estimator on the model $\mathcal{P}_r(m)$, it follows from Theorem 4 that on the same event Ω_ξ:

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

where C' is universal. We deduce from the last claim that there exist $m' \in \widehat{\mathcal{M}}$, and a piecewise polynomial ρ-estimator $\hat{s}_{m'}$based on m' such that on Ω_ξ,

$$h^2(s, \hat{s}_{m'}) \leq C'' \left\{ h^2(s, \mathcal{P}_r(m)) + \frac{(r+1)|m|}{n} \log_+^2 \left(\frac{n}{(r+1)|m|} \right) + \xi \log_+(1/\xi) \right\}, \tag{76}$$

where C'' is universal. As $|m'| \leq 4|m| - 3 \leq 4|m|$, and $L \geq 1$,

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

where C''' is universal. Define $\mathcal{P}'_{r,\ell} = \bigcup_{\substack{m \in \mathcal{M} \\ |m| \leq \ell}} \mathcal{P}_r(m)$. As m is arbitrary among $\mathcal{M}$ and $m' \in \widehat{\mathcal{M}}$,

$$\inf_{m' \in \widehat{\mathcal{M}}} \left\{ h^2(s, \hat{s}_{m'}) + L\frac{(r+1)|m'|}{n} \log_+^2 \left(\frac{n}{r+1} \right) \right\}$$
$$\leq C'''' \inf_{\ell \geq 1} \left\{ h^2(s, \mathcal{P}'_{r,\ell}) + L\frac{(r+1)\ell}{n} \log_+^2 \left(\frac{n}{r+1} \right) + \xi \log_+(1/\xi) \right\}, \tag{77}$$

where C''' is universal. We may verify that $\mathcal{P}'_{r,\ell}$ is dense in $\mathcal{P}_{r,\ell}$ in the metric space $(\mathcal{S}, h)$. This means that we may replace $\mathcal{P}'_{r,\ell}$ in (77) by $\mathcal{P}_{r,\ell}$, which ends 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. Rho-estimators for shape restricted density estimation. *Stochastic Processes and their Applications*, 126(12):3888–3912, 2016.

[BB17] Yannick Baraud and Lucien Birgé. Rho-estimators revisited: general theory and applications. *arXiv preprint arXiv:1605.05051*, 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.

[Bir12] Lucien Birgé. Robust tests for model selection. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, volume 9, pages 47–64. IMS Collections, 2012.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BM98] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[BR06] Lucien Birgé and Yves Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10:24–45, 2006.

[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.

[Efr16] Sam Efromovich. Minimax theory of nonparametric hazard rate estimation: efficiency and adaptation. *Annals of the Institute of Statistical Mathematics*, 68(1):25–75, 2016.

[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.

[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*, pages 291–304, 1992.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[Pla09] Sandra Plancade. Non parametric estimation of hazard rate in presence of censoring. *hal preprint hal-00410799*, 2009.

[RB06] Patricia Reynaud-Bouret. Penalized projection estimators of the aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006.

[RMG10] Yves Rozenholc, Thoralf Mildenberger, and Ursula Gather. Combining regular and irregular histograms by penalized likelihood. *Computational Statistics & Data Analysis*, 54(12):3313–3323, 2010.

[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.

[Sar16] Mathieu Sart. Robust estimation on a parametric model via testing. *Bernoulli*, 22(3):1617–1670, 2016.

[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*, pages 1779–1801, 1995.

[VDVW96] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

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

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