HAL
open science

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

Mathieu Sart

## To cite this version:

Mathieu Sart. Estimating a density, a hazard rate, and a transition intensity via the $\rho$-estimation method. 2018. hal-01557973v3

**HAL Id: hal-01557973**

**https://hal.science/hal-01557973v3**

Preprint submitted on 31 Jan 2018 (v3), last revised 22 Jul 2020 (v5)

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

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

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

MATHIEU SART

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

## 1. Introduction

1.1. **Statistical settings.** In the present paper, we are interesting in estimating a unknown function $s$ 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 $\mu$. 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. The variable $C$ may take the value $+\infty$. We suppose that $T$ is independent of $C$ and that $T$ admits a density $f$ with respect to the Lebesgue measure $\mu$. The target function is the hazard rate $s$ defined for $t \geq 0$ 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 $D_i = \begin{cases} 1 & \text{if } T_i \leq C_i, \\ 0 & \text{otherwise.} \end{cases}$*

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

*Date*: January, 2018.

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

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

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

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

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

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

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

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

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

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

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

1.2. **On $\rho$-estimation in framework 1.** We begin by outlining some of the underlying ideas of this estimation procedure. First, the method fits into the scheme of a series of papers using tests to construct estimators. Given two densities $f$ and $g$, a test is, intuitively, a decision rule that decides which one is the best for estimating $s$. In order to measure the quality of estimation, we consider the Hellinger distance $h$, which is defined for two non-negative integrable functions $f$ and $g$ by

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

We try, in $\rho$-estimation, to estimate $h^2(s, f) - h^2(s, g)$. The smaller this difference, the better $f$. Conversely, the larger this difference, the better $g$. Unfortunately, it seems difficult to build consistent estimators of $h^2(s, f) - h^2(s, g)$ with good properties (without additional assumptions on $s$). A way to circumvent the issue, and which follows from [Bar11], is to estimate a good approximation $T_E(f, g)$ of $h^2(s, f) - h^2(s, g)$. We will denote such estimators by $T(f, g)$.

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. When $S$ only consists of two densities, $S = \{f, g\}$, an estimator $\hat{s}$ on $S$ may be deduced from the test: we may set $\hat{s} = g$ when $T(f, g) > 0$ and $\hat{s} = f$ when $T(f, g) < 0$. When the model $S$ is more general, several methods exist in the literature to define an estimator. A key theoretical reference is [Bir06]. In $\rho$-estimation, $T(f, g)$ is viewed as an approximate value 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 approximation of $h^2(s, f) - \inf_{g \in S} h^2(s, g)$. It then remains to minimize $\gamma$ to define the $\rho$-estimator $\hat{s}$ (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 $\rho$-estimation procedure that may lead to similar theoretical results, at least in density estimation. For instance, there exist several ways to define the estimator $T(f,g)$ as explained in [BB17]. It turns out that $\rho$-estimators (including some variants) satisfy interesting statistical properties. We briefly present below four of them: generality, optimality, robustness, and superminimaxity.

First of all, $\rho$-estimators exist on some models $S$ for which the maximum likelihood method does not work. Several examples are known in the literature. A very simple one is

$$S = \{f\mathbb{1}_I,\ I \text{ is a closed interval of } \mathbb{R} \text{ and } f \text{ a non-increasing density on } I\}\,, \tag{1}$$

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

The quality of a $\rho$-estimator $\hat{s}$ relies on the behaviour of the error $T(f,g) - T_E(f,g)$. This error can be controlled according to different notions that aim at measuring the "complexity", or "massiveness" of the model $S$ (entropy with bracketing, metric dimension, covering numbers...). We may deduce an upper-bound $R_S(n)$ of the maximal risk $\sup_{s\in S}\mathbb{E}[h^2(s,\hat{s})]$. This result can then be compared to the minimax bound $\inf_{\tilde{s}}\sup_{s\in S}\mathbb{E}[h^2(s,\tilde{s})]$, where the infimum is taken over all estimators $\tilde{s}$ with values in $S$. The rate of convergence of $\inf_{\tilde{s}}\sup_{s\in S}\mathbb{E}[h^2(s,\tilde{s})]$ to 0 is usually called the optimal minimax rate of convergence. Yet, $R_S(n)$ achieves this rate, up to possible logarithmic factors, 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 $\rho$-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 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. Such a property is not shared in general by the maximum likelihood estimator. It may indeed 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

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

where $C'$ is a universal constant, $\bar{S}$ a suitable subset of $S$, and where $R_S(f,n)$ tends to 0 at a (usually) faster rate than $R_S(n)$. For illustration purposes, consider the model $S$ defined by (1). Then, $\bar{S}$ consists of piecewise constant densities belonging to $S$,

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

$d(f)$ is the number of pieces of $f$, and $\log_+ x = \max\{\log x, 1\}$. In particular, when the support of $s$ is an interval on which $s$ is non-increasing and piecewise constant, that is when $s \in \bar{S}$, the rate of estimation is parametric (up to some power of $\log_+(n/d(f))$). If we now suppose that $s$ belongs to $S \setminus \bar{S}$, some computations allows to bound (2) from above by $C_s(\log^2 n)n^{-2/3}$, where $C_s$ only depends on $s$. This 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 other models $S$ corresponding to qualitative assumptions on the density (for instance, $s$ may be piecewise monotone or $\sqrt{s}$ may be piecewise convex-concave).

There are moreover two additional properties of $\rho$-estimators we now briefly mention. First, $\rho$-estimators can be related to maximum likelihood ones. Second, it is possible to introduce penalties into the criterion $\gamma$, leading to penalized $\rho$-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 $\rho$-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 $t$,

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

measures the tendency of dying 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 $D = \mathbb{1}_{T \leq C}$ indicates whether the patient dies ($D = 1$) or leaves the study ($D = 0$).

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 $t$,

$$\begin{aligned} s(t) &= \frac{f(t)}{\mathbb{P}\left(X_{t-} = 1\right)}, \\ &= \lim_{h \to 0} \frac{\mathbb{P}\left(X_{t+h} = 0 \mid X_{t-} = 1\right)}{h}, \end{aligned}$$

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 $X_t = \mathbb{1}_{T \geq t}$.

In the literature, numerous estimators have been proposed to deal with (at least) one of these two frameworks. We may cite wavelet estimators, Kernel estimators, maximum likelihood estimators, procedures based on $\mathbb{L}^2$ contrasts... However, non-asymptotic studies seem be rather scarce. We refer to [BC05, RB06, BC08, Pla09, AD10] for results concerning procedures based on (penalized) $\mathbb{L}^2$ contrasts. We may cite [vdG95, DR02] for a study of non-asymptotic properties of maximum likelihood estimators. We refer to [BB09] for results concerning a selection rule based on pairwise comparisons of histogram type estimators.

1.4. **A generalized procedure.** As in [BB09, AD10, Bar11], we consider a more general statistical setting, namely the problem of estimating the intensity of a random measure, in order to define the estimators. This statistical setting will be very convenient to present our results in a concise way.

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

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

and in framework 3,

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

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

We define an approximation $T_E(f,g)$ of $h^2(s,f) - h^2(s,g)$ and an estimator $T(f,g)$ of $T_E(f,g)$. Given a model $S$, that is a collection of possible candidates for estimating $s$, we define our estimator $\hat{s}$ on $S$ as a minimizer, or more precisely as an approximate minimizer of $\gamma(f) = \sup_{g\in S} T(f,g)$. Similarly to framework 1, bounding above the Hellinger-type risk $h(s,\hat{s})$ of a $\rho$-estimator $\hat{s}$ requires to control the deviations of the error $T(f,g) - T_E(f,g)$. We carry out a uniform exponential inequality to control these deviations. We deduce risk bounds for $\rho$-estimators over a particular class of models $S$. This class includes, for instance, models consisting of piecewise polynomial functions, multimodal functions, piecewise convex-concave functions or piecewise log-concave functions. We establish in the three frameworks a risk bound akin to the one obtained in density estimation by [BB16]. Actually, in density estimation, the control of the empirical process is a bit more accurate than in this last paper, leading to a slightly sharper estimation term $R_S(f,n)$ (although probably not optimal). Moreover, the closeness of the results in the three frameworks allows to transfer the rates of convergence obtained in [BB16] in framework 1 when $s$ is multimodal or when $\sqrt{s}$ is piecewise convex-concave to frameworks 2 and 3 (with a minor improvement).

1.5. **On maximum likelihood estimation.** The $\rho$-estimation procedure differs from that of maximum likelihood. Nevertheless, the two approaches are very close in some situations.

This phenomenon may be understood by looking at the local behaviour of the estimator $T(f,g)$. Indeed, under some assumptions, $T(f,g)$ roughly behaves as the difference of log likelihoods $L(g) - L(f)$ when $f$ and $g$ are very close densities in framework 1. If we could replace $T(f,g)$ by $L(g)-L(f)$, $\gamma(f) = \sup_{g\in S} T(f,g)$ would become $\sup_{g\in S} L(g) - L(f)$, and a minimizer of $\gamma(\cdot)$ would be a maximum likelihood estimator.

We show that $\rho$-estimation may coincide with maximum likelihood estimation when the model is either convex or of the form $S = \left\{f^2,\, f \in \mathcal{F}\right\}$ where $\mathcal{F}$ is convex and consists of non-negative functions. In particular, a maximum likelihood estimator on such a model $S$ is a $\rho$-estimator. A similar result for convex sets of densities was obtained by Su Weijie in the context of framework 1 and was recently included in [BB17]. Consequently, our theoretical risk bounds, which apply to $\rho$-estimators, may also apply to maximum likelihood ones for these models (in the three frameworks).

1.6. **Estimator selection.** The practical computation of $\rho$-estimators seems unfortunately to be numerically out of reach in numerous models. In some cases, a model $S$ may be written as a union $S = \bigcup_{m\in\mathcal{M}} S_m$ of models $S_m$ satisfying the convex-type assumptions described in the preceding section. Thereby, one may maximize the likelihood to define a $\rho$-estimator $\hat{s}_m$ on $S_m$. All that remains then is to select an estimator among $\{\hat{s}_m,\, m \in \mathcal{M}\}$. This solution may be more numerically friendly (when $\mathcal{M}$ is not too large), but the quality of selected estimator needs to be shown.

In the present paper, we will apply this idea to cope with a particular model $S = S_{\ell,r}$. This model consists of non-negative piecewise polynomial functions of degree at most $r$ and based on at most $\ell$ consecutive intervals. Although maximum likelihood estimators do not exist on $S_{\ell,r}$, $\rho$-estimators do exist, and we may even control their Hellinger-type risks. Unfortunately, we do not know how to build these $\rho$-estimators in practice. Alternatively, we may consider collections $m$ of intervals, and define models $S_{\ell,r,m} \subset S_{\ell,r}$ consisting of functions which are polynomial on each interval $I$ of $m$. Then, $S_{\ell,r} = \bigcup_{m\in\mathcal{M}_\ell} S_{\ell,r,m}$ where the union is taken over a suitable (infinite) family $\mathcal{M}_\ell$ of collections $m$. For each $m \in \mathcal{M}_\ell$, we may define a $\rho$-estimator $\hat{s}_m$ on the convex model $S_{\ell,r,m}$ by maximizing the likelihood. Selecting among all the estimators $\hat{s}_m$ is theoretically feasible but does not yield a practical procedure as $\mathcal{M}_\ell$ is infinite. This is the reason why we will replace $\mathcal{M}_\ell$ by a finite, but usually very large, family $\widehat{\mathcal{M}}_\ell \subset \mathcal{M}_\ell$. We will carry out a new procedure, inspired from [Sar14], to select among the estimators $\hat{s}_m$, $m \in \widehat{\mathcal{M}}_\ell$. Although the large cardinal of $\widehat{\mathcal{M}}_\ell$, dynamic programming allows to reduce the numerical complexity, and makes it possible the practical implementation of the procedure in some favourable situations. We prove an oracle inequality for the selected estimator from which, we deduce, when $r = 0$, a risk bound very similar to the one we would obtain for the $\rho$-estimator. Besides, we carry out a numerical study in which we compare, in the context of density estimation, our procedure with a selection rule based on maximum likelihood only.

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

1.7. **Organization of the paper.** We carry out in Section 2 the general statistical setting that encompasses the three frameworks. We then explain the estimation procedure and relate it to the maximum likelihood one. In Section 3, we present the probabilistic tool that enables us to control the risk of $\rho$-estimators. We then present the required assumptions on the models as well as our main result on the theoretical performances of $\rho$-estimators. In Section 4, we deal with estimator selection to define a piecewise polynomial estimator as explained in Section 1.6. Section 5 is devoted to numerical simulations. The proofs are deferred to Section 6.

## 2. The $\rho$-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,

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

and in hazard rate estimation,

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

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

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

and consider

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

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

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

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

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

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

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 $x, y \in \mathbb{R}$, $x \wedge y = \min(x, y)$, $x \vee y = \max(x, y)$. The positive part of a real valued function $f$ is denoted by $f_+$ and its negative part by $f_-$. The distance between a point $x$ and a set $A$ in a metric space $(E, d)$ is denoted by $d(x, A) = \inf_{y \in A} d(x, y)$. We denote the cardinal of a set $A$ by $|A|$, and its complement by $A^c$. We set $\log_+ x = \max\{\log x, 1\}$ for all $x > 0$. The notations $c, c', C, C', \ldots$ 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 $\mu$ in framework 2. Let now $S$ be a subset of $\mathcal{S}$. Such set will be named model. Our aim is to define an estimator $\hat{s}$ with values in $S$ such that $h(s, \hat{s})$ is as small as possible.

Consider two arbitrary functions $f, g$ of $\mathcal{S}$. 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 $\psi$ be the real-valued function defined for $x \geq 0$ by $\psi(x) = \frac{\sqrt{x}-1}{\sqrt{x}+1}$, and $\psi(+\infty) = 1$. For $f, g \in \mathcal{S}$, we set

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

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

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

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

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).

Let $S$ be a model 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 $\gamma_E$.

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 $\gamma_E$ 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 $\bar{f}$ is, up to a multiplicative constant, the closest function of $s$ among the ones of $S$.

The approximation $T_E(f,g)$ is certainly unknown in practice as it involves $s$. It can however be suitably estimated by

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

Similarly, $\gamma_E(f)$ is unknown but can be estimated by $\gamma(f) = \sup_{g\in S} T(f,g)$. It then remains to minimize this criterion to define the estimator as described below.

2.3. **The procedure.** Let for $f, g \in \mathcal{S}$, $T(f,g)$ be given by (4). Let $S$ be a model and $\gamma(f) = \sup_{g\in S} T(f,g)$. Any estimator $\hat{s} \in S$ satisfying

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

is called $\rho$-estimator.

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

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

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

Remark 2. Two ingredients are required to define the procedure. 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). 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].

2.4. **Connection with maximum likelihood estimation.** The $\rho$-estimation method may be related to that of maximum likelihood. Since a model $S$ may not only consist of densities in framework 1, a traditional solution to define the maximum likelihood estimator is to add a Lagrange term in the log likelihood. More precisely, we define in the three frameworks

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

and call maximum likelihood estimator any estimator maximizing $L(\cdot)$ on $S$. In the above formula, and throughout the paper, the convention $\log 0 = -\infty$ is used. In framework 2, $L(f)$ may also be interpreted as a log likelihood (it is equal to the usual log likelihood when $f$ is a hazard rate, up to some terms constant in $f$). The same is true for framework 3, see the literature of counting processes *e.g.* equation (3.2) of [Ant89] (using that $s$ is an Aalen's multiplicative intensity).

We may write

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

As $\tanh(x) \simeq x$ 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 $\rho$-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*

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

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

*If there exists an estimator $\tilde{s} \in S$ 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 $\rho$-estimator. Conversely, assume that there exists a $\rho$-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 $L_{\mathcal{X}}(\cdot)$ 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 $\rho$-estimators and vice versa. We recover the result of Su Weijie when $S$ consists of densities in framework 1. Using sets $\mathcal{X}$ not equal to $\mathbb{R}$ may be of interest to remove some observations that would make the log likelihood identically equal to $-\infty$. In that case, we rather estimate the restriction of $s$ to $\mathcal{X}$ as illustrated in the example below.

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

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

When the random variables $X_i$ are positive, which in particular holds true almost surely 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 $\rho$-estimator. When some of the random variables $X_i$ are non-positive, $L(g) = -\infty$ for all $g \in S$, and we cannot maximize $L(\cdot)$ over $S$ to design an estimator. However, the $\rho$-estimation approach works and still coincides with the maximum likelihood one, up to minor modifications. Indeed, in this case, the preceding theorem can be used with $\mathcal{X} = (0, +\infty)$. 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 is a density and maximizes the map

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

over the densities $f$ of $S$, where $n_0$ is the number of positive random variables among $X_1, \dots, X_n$. One can verify that the estimator that maximizes $L_{\mathcal{X}}(\cdot)$ over $S$, and which is thus the $\rho$-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 $\rho$-estimator is not a density unless all the observations $X_i$ are positive. This is due to the fact that $\hat{s}$ estimates in this case the restriction of $s$ to $(0, +\infty)$ (which is not a density when some observations $X_i$ are negative).

It is sometimes convenient to consider models $S$ of the form $S = \left\{ f^2,\ f \in \mathcal{F} \right\}$ where $\mathcal{F}$ consists of non-negative functions. The set $\mathcal{F}$ can then be interpreted as a translation of the knowledge one has on $\sqrt{s}$. For instance, if $\mathcal{F}$ denotes the set of non-negative concave functions on $[0, +\infty)$ vanishing on $(-\infty, 0)$, the assumption $s \in S$ means that $\sqrt{s}$ is concave on $[0, +\infty)$ with support in $[0, +\infty)$.

It turns out that the connection between $\rho$- and maximum likelihood estimators established by Theorem 1 remains valid when the convexity assumption is put on $\mathcal{F}$ instead of $S$.

**Theorem 2.** *Let $\mathcal{F}$ be a convex set of non-negative functions such that $S = \left\{f^2,\, f \in \mathcal{F}\right\}$ is included in $\mathcal{S}$. Let $\mathcal{X}$ be a subset of $\mathbb{R}$ such that $\{x \in \mathbb{R},\, f(x) \neq 0\} \subset \mathcal{X}$ for all $f \in \mathcal{F}$. Then, if $\sup_{g\in S} L_{\mathcal{X}}(g) \notin \{-\infty, +\infty\}$, the conclusions of Theorem 1 apply to $S$: any maximizer $\tilde{s} \in S$ of $L_{\mathcal{X}}(\cdot)$ on $S$ vanishes $\gamma(\cdot)$, and any $\hat{s} \in S$ vanishing $\gamma(\cdot)$ maximizes $L_{\mathcal{X}}(\cdot)$ over $S$.*

## 3. Risk bounds of $\rho$-estimators

3.1. **A uniform exponential inequality.** We recall that the definition of $\rho$-estimators is based on the minimization of a criterion $\gamma$ on $S$. This criterion $\gamma$ 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 $\rho$-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

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

This variable is centered in each of the 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 theorem below allows to control the deviations of $Z(\varphi)$ under suitable assumptions on $\mathcal{F}$.

**Theorem 3.** *Let $\mathcal{F} \subset \mathcal{S}$ be a set of functions $\varphi$ such that $|\varphi(x)| \leq 1$ for all $\varphi \in \mathcal{F}$, $x \in \mathbb{R}$. Let $\mathcal{A}$ be the collection of sets defined by*

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

*Suppose:*

- *in frameworks 1 and 2 that $\mathcal{A}$ is a Vapnik-Chervonenkis class of dimension at most $2d$. Moreover, there exists an at most countable set $\mathcal{A}' \subset \mathcal{A}$ satisfying the following assertion: for all $A \in \mathcal{A}$, there exists a sequence $(A_m)_{m\geq 0} \in \mathcal{A}'^{\mathbb{N}}$ such that $\lim_{m\to+\infty} \mathbb{1}_{A_m}(x) = \mathbb{1}_A(x)$ for every $x \in \mathbb{R}$.*
- *in framework 3, that each set $A \in \mathcal{A}$ is a union of at most $d$ intervals.*

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

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

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

$$|Z(\varphi)| \leq \varepsilon \upsilon(\varphi) + C_\varepsilon \left\{ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right\}. \tag{8}$$

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

A class consisting of unions of at most $d$ intervals is Vapnik-Chervonenkis with dimension at most $2d$. Therefore, the condition in framework 3 is stronger than in the two first frameworks. It remains however general enough to control the risks of $\rho$-estimators in several models $S$ of interest

(see the next section). As a by-product of the proof of the theorem, we get the following proposition which may be of independent interest:

**Proposition 4.** *Consider framework 1 and an at most countable set $\mathcal{F} \subset \mathcal{S}$ of functions $\varphi$ such that $|\varphi(x)| \leq 1$ for all $x \in \mathbb{R}$, $\varphi \in \mathcal{F}$. Let for $t \in (0,1)$, $\mathcal{A}_t$ be the collection of sets defined by*

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

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

$$S_{\mathcal{A}_t}(2n) = \max_{x_1,\dots,x_{2n} \in \mathbb{R}} |\{\{x_1, \dots, x_{2n}\} \cap A, A \in \mathcal{A}_t\}|.$$

*Let $\sigma^2 = \sup_{\varphi \in \mathcal{F}} \mathbb{E}[\varphi^2(X)]$, $p_t = \sup_{\varphi \in \mathcal{F}} \mathbb{P}(|\varphi(X)| > t)$, and $r_n = \sup_{t \in (0,1)} \log_+ |S_{\mathcal{A}_t}(2n)|$. Then, there exist universal constants $C, C'$ such that*

$$\begin{aligned}\mathbb{E}\left[\sup_{\varphi \in \mathcal{F}} |Z(\varphi)|\right] &\leq C \inf_{\eta \in (0,1)} \left\{\sigma\sqrt{\log(1/\eta)} + \int_0^\eta \sqrt{p_t}\ \mathrm{d}t\right\} \sqrt{\frac{r_n}{n}} + C\frac{r_n}{n} \\ &\leq C'\left[\sigma\sqrt{\frac{r_n \log_+(1/\sigma)}{n}} + \frac{r_n}{n}\right]. \end{aligned} \tag{9}$$

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

Our proposition is similar to a result of [Bar16] when the random variables $X_i$ are identically distributed. Actually, our inequality (9) is a bit sharper when $\mathcal{A}_t$ is Vapnik-Chervonenkis with dimension $d$ (apart from constants). Such an assumption corresponds to a notion of (weak) VC-major class (see [Bar16]). In that case, Sauer's lemma [Sau72] implies

$$\mathbb{E}\left[\sup_{\varphi \in \mathcal{F}} |Z(\varphi)|\right] \leq C''\left[\sigma\sqrt{\frac{d\log_+(n/d)\log_+(1/\sigma)}{n}} + \frac{d\log_+(n/d)}{n}\right], \tag{10}$$

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

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

**Lemma 2.** *For all $f, g \in \mathcal{S}$,*

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

Now, under suitable assumptions on the collection $\mathcal{F} = \{\psi(g/f), f, g \in S\}$, Inequality (8) roughly says that with high probability (and $\varepsilon = 1/24$):

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

The term $D_S(n)$ depends on the probability of the event on which (11) holds 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 a $\rho$-estimator $\hat{s}$. We may do as 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,

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

and hence,

$$\begin{aligned} \frac{1}{6}h^2(s,\hat{s}) - \frac{19}{6}h^2(s,S) - D_S(n) &\leq \gamma(\hat{s}) \\ &\leq \inf_{f\in S} \gamma(f) + 1/n \\ &\leq \frac{18}{6}h^2(s,S) + D_S(n) + 1/n. \end{aligned}$$

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

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

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

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

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

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

  *for every $x \in \mathbb{R}$.*
- *in framework 3, that the sets $I \in \mathcal{I}_d$ are unions of at most $d$ intervals.*

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

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

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

Let $\mathcal{P}_{\ell,r}$ be the model defined for $\ell \geq 1$, $r \geq 0$ by

$$\mathcal{P}_{\ell,r} = \left\{ \sum_{j=1}^{\ell} P_j \mathbb{1}_{K_j},\ P_j \text{ is a polynomial function of degree at most } r, \text{ which is non-negative on an interval } K_j \text{ of } \mathbb{R} \text{ of finite length} \right\}. \tag{12}$$

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

We may also consider the model consisting of piecewise monotone functions defined for $k \geq 1$ by

$$\mathcal{F}_k = \mathcal{S} \cap \left\{ \sum_{j=1}^{k} f_j \mathbb{1}_{K_j}, K_j \text{ is an interval, and } f_j \text{ is monotone on } K_j \right\}. \tag{13}$$

Note that multimodal functions with $k-1$ modes belong to $\mathcal{F}_k$.

We introduce for all interval $K$ the set

$$\mathcal{G}(K) = \{f, f \text{ is either convex or concave on the interior of } K\},$$

and define

$$\mathcal{G}_k = \mathcal{S} \cap \left\{ \sum_{j=1}^{k} f_j \mathbb{1}_{K_j}, K_j \text{ is an interval, and } f_j \in \mathcal{G}(K_j) \right\},$$

$$\mathcal{G}'_k = \mathcal{S} \cap \left\{ \sum_{j=1}^{k} e^{f_j} \mathbb{1}_{K_j}, K_j \text{ is an interval, and } f_j \in \mathcal{G}(K_j) \right\}.$$

The proposition below ensues from [BB16]:

**Proposition 6.** *Let for $d \geq 1$, $\mathcal{I}_d$ be the collection of unions of at most $d$ intervals. Let $\mathcal{P}'_{\ell,1} = \left\{f,\ f \in \mathcal{S},\ e^f \in \mathcal{P}_{\ell,1}\right\}$. Assumption 2 is fulfilled with:*

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

### 3.3. **A uniform risk bound.**

**Theorem 7.** *Let $(\mathcal{I}_d)_{d\geq 1}$ be a non-decreasing collection of Borel sets that fulfils Assumption 1. For all $\xi > 0$, there exists an event which holds true with probability larger than $1 - e^{-n\xi}$ and on which: for all model $S$ satisfying Assumption 2 and all $\rho$-estimator $\hat{s}$ on $S$,*

$$h^2(s,\hat{s}) \leq \inf_{f \in \bar{S}} \left\{ c_1 h^2(s,f) + c_2 \frac{d_S(f)}{n} \log_+^2\left(\frac{n}{d_S(f)}\right) + c_3 \xi \log_+(1/\xi) \right\}. \tag{14}$$

*In particular,*

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq \inf_{f \in \bar{S}} \left\{ c_1 \mathbb{E}\left[h^2(s,f)\right] + c'_2 \frac{d_S(f)}{n} \log_+^2\left(\frac{n}{d_S(f)}\right) \right\}. \tag{15}$$

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

Remark 1. When $s \in \bar{S}$, the risk of a $\rho$-estimator on $S$ is bounded by

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

The rate of estimation of $s$ then becomes parametric (up to the logarithmic term $\log_+^2(n/d_S(s))$). 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 $(d_S(f)/n)\log_+^2(n/d_S(f))$:

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq CR(s) \quad \text{with } R(s) = \inf_{f\in\bar{S}}\left\{\mathbb{E}\left[h^2(s,f)\right] + \frac{d_S(f)}{n}\log_+^2\left(\frac{n}{d_S(f)}\right)\right\}. \tag{16}$$

It remains to compute $R(s)$ to deduce (an upper bound of) the rate of convergence of the $\rho$-estimator when $s \in S$. This rate may 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 been put forward in [BB16] for $\rho$-estimators in density estimation and has been named superminimaxity. We show here that this phenomenon also occurs for $\rho$-estimators in frameworks 2 and 3. If we now restrict to density estimation, we observe that our risk bound slightly improves the one of [BB16] in the sense that our approximation term involves a smaller exponent on the logarithm. Moreover, the risk of a $\rho$-estimator is controlled on an event that does not depend on the model but rather on the class of models for which Assumption 2 is met. This result is useful to perform (piecewise) polynomial estimator selection, see Section 4. However, our Assumption 2 is more stringent than the assumptions on models in [BB16].

Remark 2. When the framework varies, the Hellinger loss $h$ in (15) changes, but the variance term remains the same. Moreover, the bias term in (15) in frameworks 2 and 3 is not larger than the bias term in density estimation (when one extends the definition of $s$ in frameworks 2 and 3 to $(-\infty,0)$ by setting $s(x)=0$ when $x<0$). We may therefore only focus on density estimation to bound above $R(s)$ in (16). The risk of the $\rho$-estimator is then bounded by this upper bound in the three frameworks (up to a multiplicative constant).

Remark 3. It follows from a crude application of the triangular inequality and from (15) that

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

If we know how to bound $R(g)$ for $g \in S$, this inequality says that the risk of the $\rho$-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 $\hat{s}$ 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.

In particular, this robustness property applies to maximum likelihood estimators when assumptions of Theorem 1 or 2 are met with $\mathcal{X} = \mathbb{R}$. It is worth mentioning that this result is not true in general for maximum likelihood estimators. We refer to Section 2.3 of [Bir06] for an example in density estimation. We carry out below another example in framework 2. Let $\alpha > 0$ and $s_\alpha(t) = t^{-1}\mathbb{1}_{t\geq\alpha}$ be the hazard rate of a Pareto distribution. Suppose that the density $f$ of $T$ is a mixture of two Pareto distributions,

$$f(t) = \varepsilon\frac{\eta}{t^2}\mathbb{1}_{t\geq\eta} + (1-\varepsilon)\frac{1}{t^2}\mathbb{1}_{t\geq 1},$$

where $\varepsilon = 1/n$ and where $\eta \in (0, 1/2)$. Suppose moreover that the data are not censored: $C = +\infty$ almost surely. The survival function $G$ and the hazard rate $s$ of $T$ are given by the formulas

$$G(t) = \left(1 - \varepsilon + \varepsilon\frac{\eta}{t}\right)\mathbb{1}_{t\in[\eta,1)} + \frac{1-\varepsilon+\varepsilon\eta}{t}\mathbb{1}_{t\geq 1}$$
$$s(t) = \frac{\varepsilon\eta}{(1-\varepsilon)t+\varepsilon\eta}\frac{\mathbb{1}_{t\in[\eta,1)}}{t} + \frac{\mathbb{1}_{t\geq 1}}{t}.$$

The maximum likelihood estimator on $S = \{s_\alpha, \alpha > 0\}$ is $\tilde{s} = s_{\tilde{\alpha}}$ where $\tilde{\alpha} = \min_{1\leq i\leq n} X_i$. We are now interested in its risk. Let $\mathcal{A}_n$ be the event on which there exists at least one observation $X_i$ smaller than $2\eta$ and $\mathcal{B}_n$ be the event on which $n/2$ observations are larger than $t_0$ where $G(t_0) = 3/4$. Then, $\mathbb{P}(\mathcal{A}_n) = 1 - (1-\varepsilon/2)^n = 1 - (1 - 1/(2n))^n > 1 - e^{-1/2}$ and $\mathbb{P}(\mathcal{B}_n) = \mathbb{P}(Y \geq n/2)$ where $Y$ is binomially distributed with parameters $(n, 3/4)$. Therefore, for all $c \in (0,1)$, and $n$ large enough, $\mathbb{P}(\mathcal{B}_n) \geq c$. Since $G(1) \geq 1 - 1/n \geq 3/4$ when $n \geq 4$, $t_0 \geq 1$ and we deduce that on $\mathcal{A}_n \cap \mathcal{B}_n$,

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

Some elementary computations show (using that $\varepsilon \leq 1/2$),

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

Therefore, $h^2(s,\tilde{s}) > 0.04\log(1/(2\eta))$ and $\mathbb{E}[h^2(s,\tilde{s})] > 0.04\log(1/(2\eta))\mathbb{P}(\mathcal{A}_n \cap \mathcal{B}_n)$. Finally, there exist universal constants $c'$, $n_0$ such that for all $n \geq n_0$:

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

In particular, the risk of a maximum likelihood estimator $\tilde{s}$ can be made arbitrarily large by playing with $\eta$. We now turn to $\rho$-estimation. The model $S$ fulfils Assumption 2 with $d_S(f) = 1$, $\bar{S} = S$ and where $\mathcal{I}_1$ is the collection of intervals. A $\rho$-estimator $\hat{s}$ on $S$ satisfies therefore $\mathbb{E}[h^2(s,\hat{s})] \leq C(\mathbb{E}[h^2(s,S)] + \log^2 n/n)$. The variance term $\log^2 n/n$ is likely not optimal. However, the bias term can be bounded above by $\mathbb{E}[h^2(s,S)] \leq \mathbb{E}[h^2(s,s_1)] \leq \mathbb{P}(T \in [\eta,1]) = \varepsilon(1-\eta) \leq 1/n$. Consequently,

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

where $C'$ is universal. This example illustrates the lack of robustness of the maximum likelihood estimator in this model.

When a $\rho$-estimator $\hat{s}$ vanishes the criterion $\gamma$, which typically happens when $\hat{s}$ maximizes $L_{\mathcal{X}}$ on some models, the constant $c_1$ appearing in front of the bias term $h^2(s,f)$ in Theorem 7 can be improved:

**Proposition 8.** *Let $(\mathcal{I}_d)_{d\geq 1}$ be a non-decreasing collection of Borel sets that fulfils Assumption 1. For all $\xi > 0$, there exists an event which holds true with probability larger than $1 - e^{-n\xi}$ and on which: for all model $S$ satisfying Assumption 2 and all $\rho$-estimator $\hat{s}$ on $S$ such that $\gamma(\hat{s}) = 0$, the risk bound (14) holds true for all $\varepsilon > 0$ with $c_1 = c_{1,\varepsilon}$, $c_2 = c_{2,\varepsilon}$, $c_3 = c_{3,\varepsilon}$, 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 for estimators $\hat{s}$ maximizing $L_{\mathcal{X}}$ under convex type assumptions

on the models. However, $c_1 = c_{1,\varepsilon}$ cannot be smaller than 2 in general as shown by the following elementary example.

Consider framework 1, a finite collection of disjoint intervals $m$, and the model $S$ consisting of piecewise constant densities based on $m$. Since $S$ is convex, the usual histogram estimator $\hat{s}$ is a $\rho$-estimator that vanishes $\gamma$. Moreover, it is known that this estimator may not converge to the closest function of $s$ in the model (for the Hellinger distance), see [BR06]. In particular, let $p \in (0,1)$ to be specified later, $s = \mathbb{1}_{[0,p]} + (1-p)\mathbb{1}_{(1,2]}$ and $S$ be the model of piecewise constant densities based on the partition $\{[0,1],(1,2]\}$. Then, $\mathbb{E}[h^2(s,\hat{s})]$ converges to $h^2(s,\bar{s}_1)$ where $\bar{s}_1 = p\mathbb{1}_{[0,1]} + (1-p)\mathbb{1}_{(1,2]}$. Define

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

The fraction

$$\frac{h^2(s,\bar{s}_1)}{h^2(s,\bar{s}_2)} = \frac{p(1-\sqrt{p})}{1-\sqrt{1-p(1-p)}}$$

can be made arbitrarily close to 2 by choosing $p$ small enough. Therefore, for all $\eta \in (1,2)$, and $p$ small enough,

$$\lim_{n\to+\infty} \mathbb{E}[h^2(s,\hat{s})] \geq \eta h^2(s,\bar{s}_2) \geq \eta h^2(s,S),$$

and thus $c_1 \geq \eta$.

3.4. **Risks of $\rho$-estimators for models consisting of step functions.** For illustration purposes, consider the collection $\mathcal{P}_{\ell,0}$ of step functions defined by (12). Then, any $\rho$-estimator $\hat{s}$ on $\mathcal{P}_{\ell,0}$ satisfies

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C\left\{\mathbb{E}\left[h^2(s,\mathcal{P}_{\ell,0})\right] + \frac{\ell \log_+^2(n/\ell)}{n}\right\}. \tag{17}$$

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. When $s$ does belong to $\mathcal{P}_{\ell,0}$, the bound becomes

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

This result is, in general, slightly suboptimal as it is possible to do better in density estimation. Indeed, we may relate in framework 1 the performance of a $\rho$-estimator to the Vapnik-Chervonenkis dimension of the subgraphs of the model. Yet, this dimension is of the order of $\ell$ for $\mathcal{P}_{\ell,0}$ (see [BB17]). We then derive from the proof of Theorem 12 of [BBS17],

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

where $C'$ is universal. This last inequality is also shown and discussed in Section 7.4 of [BB17]. In particular, the logarithm in this inequality cannot be avoided since the variance term corresponds to the optimal minimax rate of convergence on $\mathcal{P}_{\ell,0}$ (Proposition 2 of [BM98]).

3.5. **Risks of $\rho$-estimators for models consisting of 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 7. We know from (16) that a $\rho$-estimator $\hat{s}$ satisfies

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq CR(s) \quad \text{with } R(s) = \inf_{f\in\bar{S}} \left\{ \frac{1}{2} d_2^2(\sqrt{s}, f) + \frac{d_S(f)}{n} \log_+^2\left(\frac{n}{d_S(f)}\right) \right\}, \tag{18}$$

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

Two bounds on $R(s)$ in density estimation can be deduced from the results of [BB16]: when $s$ is piecewise monotone and when $\sqrt{s}$ is piecewise convex-concave. Dealing with the two other statistical settings requires little supplementary work (see Remark 2 in Section 3.3). We obtain bounds that are roughly of the order of $(\log^2 n/n)^{2/3}$ when $s$ is piecewise monotone, and $(\log^2 n/n)^{4/5}$ when $\sqrt{s}$ is piecewise convex-concave. In the sequel, we propose to make explicit the first bound only.

We define for $K \subset \mathbb{R}$ and $f \in \mathcal{S}$, $V_K(f) = \sup_{x\in K} f(x) - \inf_{x\in K} f(x)$. Let $\mathcal{M}_k$ be the family gathering all the collections $m$ of at most $k$ disjoint intervals of $\mathbb{R}$, and for $m \in \mathcal{M}_k$,

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

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

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

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 1.** *Any $\rho$-estimator $\hat{s}$ on $\mathcal{F}_k$ $(k \geq 1)$ satisfies*

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C \inf_{f\in\mathcal{F}_k} \left\{ d_2^2\left(\sqrt{s}, f\right) + L(f)\left(\frac{\log^2 n}{n}\right)^{2/3} + \frac{k\log^2 n}{n} \right\}. \tag{19}$$

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

$$\mathbb{E}\left[h^2(s,\hat{s})\right] \leq C \left[ L\left(\sqrt{s}\right)\left(\frac{\log^2 n}{n}\right)^{2/3} + \frac{k\log^2 n}{n} \right]. \tag{20}$$

*In the preceding inequalities, $C$ is a universal constant.*

We may also make inequality (20) more explicit when $k = 2$ and $s$ is unimodal by bounding $L(\sqrt{s})$ from above. We distinguish the cases according to the different frameworks.

Consider framework 1. Then, $M_E(K) = \mu(K)$. Therefore, if the support of $s$ is of finite length $L_{supp}$,

$$L(\sqrt{s}) \leq 2L_{supp}^{1/3}(\sup_{x\in\mathbb{R}} s(x))^{1/3}.$$

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

$$L(\sqrt{s}) \leq 2\,(\mathbb{E}(X))^{1/3}\,\big(\sup_{x\geq 0} s(x)\big)^{1/3}.$$

As to framework 3, define the time $T_1 = \int_{I_{\text{obs}}} \mathbb{1}_{X_{t-}=1}\,\mathrm{d}t$ during which the (left limit) of the observed Markov process is in state 1 and suppose that $\mathbb{E}(T_1) < \infty$. Then, $M_E(K) \leq \mathbb{E}(T_1)$ and thus

$$L(\sqrt{s}) \leq 2\,(\mathbb{E}(T_1))^{1/3}\,\big(\sup_{x\in I_{\text{obs}}} s(x)\big)^{1/3}.$$

## 4. Selecting among estimators

It is often difficult in practice to find a global minimum of $\gamma$ and thus to build $\rho$-estimators. In particular, we do not know how to construct a $\rho$-estimator on the model $S = \mathcal{P}_{\ell,r}$. In this section, we propose an alternative way, more numerically friendly, to define a piecewise polynomial estimator and study its properties.

4.1. **Piecewise polynomial estimator selection.** The quality of a $\rho$-estimator is, in the present paper, assessed by means of Theorem 7. The event on which (14) is valid depends on the collections $(\mathcal{I}_d)_{d\geq 1}$ but not on the model $S$. In particular, the risk bound remains true when the model $S$ vary randomly among the class of models for which Assumption 2 is fulfilled.

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

$$\mathbb{E}\left[h^2(s, \hat{s}_{\hat{\lambda}})\right] \leq C \left\{ \inf_{\lambda\in\Lambda} \mathbb{E}\left[h^2(s, \hat{s}_\lambda)\right] + \frac{(r+1)\ell \log_+^2\left(n/(\ell(r+1))\right)}{n} \right\}, \tag{21}$$

where $C$ is a universal constant.

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

$$\mathbb{E}\left[h^2(s, \hat{s})\right] \leq C' \left\{ \mathbb{E}\left[h^2(s, \mathcal{P}_{\ell,r})\right] + \frac{(r+1)\ell \log_+^2\left(n/(\ell(r+1))\right)}{n} \right\}, \tag{22}$$

where $C'$ is universal. 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 $\Lambda$ is finite and not too large (and when the computation of the estimators $\hat{s}_\lambda$ is fast enough).

4.2. **Selecting among a special collection of piecewise polynomial estimators.** As we see in (21), we should take $\Lambda$ as large as possible to improve on the theoretical performances of the selected estimator. However, the numerical complexity of the procedure depends heavily on the number of estimators. In this section, we propose another way, inspired from [Sar14], to deal with a special, but possibly very large, collection of piecewise polynomial $\rho$-estimators.

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

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

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

**Lemma 3.** *Let $m \in \mathcal{M}$ and for $K \in m$,*

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

*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 $\rho$-estimator on the model $S = \mathcal{P}_r(m)$ that vanishes $\gamma$.*

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

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

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

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

We consider a random variable $\hat{\ell}$ with values in $\{1, \ldots, \hat{n} - 1\}$. For each $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$, we define the $\rho$-estimator $\hat{s}_m$ on the model $\mathcal{P}_r(m)$ as in Lemma 3. The aim of this section is to explain how we can select an estimator among the family $\{\hat{s}_m,\ m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$.

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

$$m_K \vee K = \left\{ K' \cap K,\ K' \in m_K,\ K' \cap K \neq \emptyset \right\}. \tag{23}$$

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

$$\gamma_2(\hat{s}_m) = \sum_{K \in m} \sup_{m' \in \widehat{\mathcal{M}}_{\hat{\ell}}} \left\{ T\left(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K\right) - L(r+1) \frac{|m' \vee K| \log_+^2\left(n/(r+1)\right)}{n} \right\}. \tag{24}$$

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

$$\gamma_2(\hat{s}_{\hat{m}}) = \min_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} \gamma_2(\hat{s}_m). \tag{25}$$

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

**Theorem 9.** *There exists a universal constant $L_0$ such that if $L \geq L_0$, any estimator $\hat{s}_{\hat{m}}$ minimizing (25) satisfies for all $\xi > 0$, and probability larger than $1 - e^{-n\xi}$,*

$$h^2(s, \hat{s}_{\hat{m}}) \leq C \left\{ \inf_{m \in \widehat{\mathcal{M}}_{\hat{\ell}}} h^2(s, \mathcal{P}_r(m)) + L \frac{(r+1)\hat{\ell} \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right\}. \tag{26}$$

*In particular,*

$$\mathbb{E}\left[h^2(s,\hat{s}_{\hat{m}})\right] \leq C'\mathbb{E}\left[\inf_{m\in\widehat{\mathcal{M}}_{\hat{\ell}}} h^2(s,\mathcal{P}_r(m)) + L\frac{(r+1)\hat{\ell}\log_+^2(n/(r+1))}{n}\right]. \tag{27}$$

*In the above inequalities, $C$ and $C'$ are universal constants. Moreover, the event of probability larger than $1-e^{-n\xi}$ on which (26) holds may be chosen independently of $(Y_i)_{i\in\hat{I}}$, $\hat{\ell}$, $\hat{n}$, $r$ and the value of $L$ (when $L \geq L_0$).*

This last inequality also says

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

This risk bound is very similar to the one (21) obtained for the first selection rule (we only slightly loose on the variance term).

The numerical complexity of this procedure may be significantly reduced by using an algorithm of dynamic programming. The estimator $\hat{s}_{\hat{m}}$ may then be computed in practice when the $\hat{s}_K$ can be constructed sufficiently fast and when $n$, $\hat{n}$, $\hat{\ell}$ are small enough (we need to compute $\mathcal{O}(\hat{n}^2)$ $T(\hat{s}_K\mathbb{1}_{K\cap K'},\hat{s}_{K'}\mathbb{1}_{K\cap K'})$, and then use an algorithm whose numerical complexity is of the order of $\mathcal{O}(\hat{n}^4\hat{\ell}^2)$). Unfortunately, the practical computation of the estimator may take too long in other situations. For more informations on this algorithm, we refer to [Kan92] and Section 4.2 of [CR04].

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

The bias term in (26) depends on the collection $\widehat{\mathcal{M}}_{\hat{\ell}}$ and thus on the choice of $\{Y_i,\, i\in\hat{I}\}$. In general, this bias term may be larger than the one we would obtain for a $\rho$-estimator on $\mathcal{P}_{\hat{\ell},r}$. Nevertheless, it may be controlled in favourable situations as explained below.

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

$$N(A) = N(\{Y_i,\, i\in\hat{I}\}\cap A) \quad \text{for all } A\in\mathcal{B}(\mathbb{R}). \tag{28}$$

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

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

The lemma below allows to bound the bias term when $r = 0$.

**Lemma 4.** *Suppose that condition (28) is met. Let $r = 0$, and for $\ell \geq 1$, $\mathcal{M}'_\ell$ be the family that gathers all the collections $m$ of disjoint intervals of the form*

$$m = \{[x_1, x_2], (x_2, x_3], (x_3, x_4], \dots, (x_\ell, x_{\ell+1}]\}, \tag{29}$$

*where $x_1 < x_2 < \cdots < x_{\ell+1}$ (with the convention that $m = \{[x_1, x_2]\}$ when $\ell = 1$). Then, for all $\xi > 0$, the following holds with probability larger than $1 - e^{-n\xi}$: for all $\ell \in \{1, \dots, \hat{n} - 1\}$, $m \in \mathcal{M}'_\ell$ written as in (29) and such that $Y_{(1)}$ and $Y_{(\hat{n})}$ belong to $[x_1, x_{\ell+1}]$, there exists $m' \in \widehat{\mathcal{M}}_\ell$ such that the $\rho$-estimator $\hat{s}_{m'}$ satisfies*

$$h^2(s, \hat{s}_{m'}) \leq C \left\{ h^2(s, \mathcal{P}_0(m)) + \frac{\ell \log_+^2 (n/\ell)}{n} + \xi \log_+(1/\xi) \right\}. \tag{30}$$

In particular, suppose that $s$ vanishes outside $[0, 1]$, and $r = 0$. Let $\mathcal{P}'_{\hat{\ell},0}$ be the collection of step functions based on partitions $m$ of $[0, 1]$ belonging to $\mathcal{M}'_{\hat{\ell}}$. Then, we deduce from (27) and the preceding lemma when (28) holds, $\{Y_i,\, i \in \hat{I}\} \subset [0, 1]$ and $L \geq \max\{L_0, 1\}$,

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

where $C''$ is a universal constant. This risk bound, corresponds, up to slight modifications (in the constant and the variance term), to the one we would obtain for the $\rho$-estimator on $\mathcal{P}'_{\hat{\ell},0}$.

4.3. **About $L$.** The preceding procedure suffers from a major drawback: the choice of $L$. This parameter, is, indeed, involved in the construction of the estimator $\hat{s}_{\hat{m}}$ and must be chosen by the statistician. Unfortunately, Theorem 9 only applies when $L \geq L_0$, and its proof gives a value of $L_0$ that is too large to be used in practice.

A simple solution to overcome this difficulty 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 (24) 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 then select an estimator among $\{\hat{s}_{\hat{m}_L},\, L \in \mathcal{L}\}$ as explained in Section 4.1.

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

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

where $C$ is a universal constant. If $\mathcal{L}$ contains at least one number $L$ larger than $L_0$, we derive from (26),

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

where $C'$ is a universal constant.

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 $\mathcal{L}$, and the constant $C'$ above may be larger than in (26).

4.4. **Adaptive piecewise polynomial estimation.** In this section, we modify the previous procedure in order to choose the number $\hat{\ell}$ of pieces from the data.

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 $\{1, \dots, \hat{n}-1\}$ and aim at selecting an estimator among $\{\hat{s}_m,\ m \in \widehat{\mathcal{M}}_{\hat{k},lower}\}$.

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

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

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

$$\gamma_3(\hat{s}_{\hat{m}}) + 2L(r+1)\frac{|\hat{m}| \log_+^2\left(n/(r+1)\right)}{n} = \inf_{m \in \widehat{\mathcal{M}}_{\hat{k},lower}} \left\{ \gamma_3(\hat{s}_m) + 2L(r+1)\frac{|m| \log_+^2\left(n/(r+1)\right)}{n} \right\}. \tag{31}$$

**Theorem 10.** *There exists a universal constant $L_0$ such that if $L \geq L_0$, any estimator $\hat{s}_{\hat{m}}$ satisfying (31) satisfies for all $\xi > 0$, and probability larger than $1 - e^{-n\xi}$,*

$$h^2(s, \hat{s}_{\hat{m}}) \leq C \inf_{m \in \widehat{\mathcal{M}}_{\hat{k},lower}} \left\{ h^2(s, \mathcal{P}_r(m)) + L\frac{(r+1)|m| \log_+^2(n/(r+1))}{n} + \xi \log_+(1/\xi) \right\}. \tag{32}$$

*In particular,*

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

*In the above inequalities, $C$ and $C'$ are universal constants. Moreover, the event of probability larger than $1 - e^{-n\xi}$ on which (32) holds may be chosen independently of $(Y_i)_{i \in \hat{I}}$, $\hat{k}$, $\hat{n}$, $r$ and the value of $L$ (when $L \geq L_0$).*

This risk bound improves when $\hat{k}$ grows up. Moreover, (33) implies

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

The right-hand side of this inequality corresponds to the bound (27) achieved by the estimator of Section 4.2 when the choice of $\hat{\ell}$ is the best possible among $\{1, \dots, \hat{k}\}$.

The quality and the construction of the estimator $\hat{s}_{\hat{m}}$ still depends on $\{Y_i,\, i \in \hat{I}\}$. However, when $\hat{k} = \hat{n}$, and when $\{Y_i,\, i \in \hat{I}\}$ is rich enough, the infimum in (33) can be taken over the infinite collection $\mathcal{M}$ (up to a modification of $C'$), as shown below.

**Lemma 5.** *Suppose that $\{Y_i,\, i \in \hat{I}\}$ is chosen in such a way that $N$ satisfies (28). There exists a universal constant $C$ such that for all $\xi > 0$ and probability larger than $1 - e^{-n\xi}$: for all $m \in \mathcal{M}$, there exists $m' \in \widehat{\mathcal{M}}$ such that*

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

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

Therefore, when $\hat{k} = \hat{n}$, when (28) is satisfied, and when $L \geq \max\{1, L_0\}$, we deduce from (33) that $\hat{s}_{\hat{m}}$ satisfies

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

where $C''$ is a universal constant. Now, $\cup_{\substack{m \in \mathcal{M}, \\ |m| = \ell}} \mathcal{P}_r(m)$ is dense in $\mathcal{P}_{\ell,r}$ in the metric space $(\mathcal{S}, h)$, and hence

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

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 a $\rho$-estimator on $\mathcal{P}_{\ell,r}$ (up to constants), barely worse than the one given by Theorem 7 and that is written in (22).

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 when $\{Y_i,\, i \in \hat{I}\}$ is random, but refer to [RMG10] for a numerical study in framework 1.

## 5. Numerical simulations

We consider framework 1, $r = 0$, $\ell \in \{1, \dots, n\}$, $\{Y_i,\, i \in \hat{I}\} = \{X_1, \dots, X_n\}$ and the (random) collection $\widehat{\mathcal{M}}_\ell$ consisting of partitions of $[X_{(1)}, X_{(n)}]$ of size $\ell$ defined in Section 4.2. For each $m \in \widehat{\mathcal{M}}_\ell$, we consider the $\rho$- and maximum likelihood estimator $\hat{s}_m$ on $\mathcal{P}_0(m)$ defined by

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

We carry out in this section a numerical study to compare two selection rules described in Sections 4.2 and 4.3.

- 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 $\rho$-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 (24) and (25) to get a partition $\hat{m}_L \in \widehat{\mathcal{M}}_\ell$. We then use the procedure of Section 4.1 to pick out an estimator among $\{\hat{s}_{\hat{m}_L},\, L \in \mathcal{L}\}$ 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*
$$s(x) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{1}{2}\log^2 x} \mathbb{1}_{(0,+\infty)}(x) \quad \textit{for all } x \in \mathbb{R}.$$

**Example 3.** *$s$ is the density of an exponential distribution*
$$s(x) = e^{-x} \mathbb{1}_{[0,+\infty)}(x) \quad \textit{for all } x \in \mathbb{R}.$$

**Example 4.** *$s$ is the density of a mixture of uniform distributions*
$$s(x) = \frac{1}{2} \times 3\mathbb{1}_{[0,1/3]}(x) + \frac{1}{8} \times 3\mathbb{1}_{[1/3,2/3]}(x) + \frac{3}{8} \times 3\mathbb{1}_{[2/3,1]}(x) \quad \textit{for all } x \in \mathbb{R}.$$

We simulate $N_{\text{rep}}$ samples $(X_1, \ldots, X_n)$ 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 $i \in \{1, \ldots, N_{\text{rep}}\}$, $\hat{s}_{\hat{m}^{(k,\ell,i)}}$ be the value of the estimator corresponding to the $k^{\text{th}}$ procedure and the $i^{\text{th}}$ sample. We evaluate the quality of the estimators by
$$\widehat{R}(k,\ell) = \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} h^2\big(s, \hat{s}_{\hat{m}^{(k,\ell,i)}}\big).$$
We estimate the probability that the two estimators coincide by
$$\widehat{P}_{equal}(\ell) = \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} \mathbb{1}_{\hat{m}^{(2,\ell,i)} = \hat{m}^{(1,\ell,i)}}$$
Results are summarized in Figures 1 (when $n = 50$) and 2 (when $n = 100$).

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

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

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

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

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 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 robustness properties of the second procedure may be useful. The fractions $\widehat{R}(2,\ell)/\widehat{R}(1,\ell)$ suggest indeed that the second procedure improves the risk of the first one by a few percent, at least when the size $\ell$ of the partitions is well adapted to the underlying density, that is when $\ell$ corresponds to the smallest values of $\widehat{R}(1,\ell)$ and $\widehat{R}(2,\ell)$.

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

## 6. Proofs

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

$$\mathcal{X} = \{x \in \mathbb{R},\ g(x) \neq 0 \ \text{ or } \ f(x) \neq 0\}.$$

Then,

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

Note that

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

Therefore,

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

Now,

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

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

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

We now put 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. $\square$.

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

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

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

$$L_{\mathcal{X}}(f) = \frac{1}{n}\sum_{j\in\hat{J}} \log f(Y_j) - \int_{\mathcal{X}} f(x)\, \mathrm{d}M(x).$$

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, $\sup_{g\in\bar{S}} T(\bar{f}, g) = 0$ if and only if $\sup_{g\in\bar{S}} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f}) = 0$.*

*Proof.* Suppose that $\sup_{g\in\bar{S}} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f}) = 0$.

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, $T(\bar{f}, g) \leq 0$.

Let now $g \in \bar{S}_2$, $u \in [0, 1]$ and $\zeta = g - \bar{f}$. Note that $\bar{f} + u\zeta = (1 - u)\bar{f} + ug \in \bar{S}$ and thus $L_{\mathcal{X}}(\bar{f} + u\zeta) - L_{\mathcal{X}}(\bar{f}) \leq 0$. 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.

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

Likewise, $\sup_{g\in\bar{S}} T(\bar{f},g) = 0$ implies $\sup_{g\in\bar{S}} L_{\mathcal{X}}(g) - L_{\mathcal{X}}(\bar{f}) = 0$.

□

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 $u \in [0,1)$, $L_{\mathcal{X}}(f_u) \neq -\infty$ 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 $\rho$-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 $\wp_3(u) \geq |\hat{J}'|/(2n) > 0$, 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 $L_{\mathcal{X}}(g) \neq -\infty$, $L_{\mathcal{X}}(g) \leq L_{\mathcal{X}}(\hat{s})$. This inequality being true if $L_{\mathcal{X}}(g) = -\infty$, the proof is complete. □

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

$$T_2(f,g) = T(f^2, g^2) = \frac{1}{n} \sum_{j \in \hat{J}} \psi_2 \left( \frac{g(Y_j)}{f(Y_j)} \right) - \frac{1}{4} \int_{\mathcal{X}} \left( g^2(x) - f^2(x) \right) \mathrm{d}M(x),$$

$$L_{\mathcal{X},2}(f) = L_{\mathcal{X}}(f^2) = \frac{2}{n} \sum_{j \in \hat{J}} \log f(Y_j) - \int_{\mathcal{X}} f^2(x) \, \mathrm{d}M(x).$$

The proof is very similar to the one of Theorem 1. The main change lies in the replacement of the symbols $S$, $T$, $L_{\mathcal{X}}$ by $\mathcal{F}$, $T_2$, $L_{\mathcal{X},2}$. We will only give some insight into why Claim 1 remain valid under these modifications.

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

**Claim 2.** *Let $\bar{\mathcal{F}} = \{g \in \mathcal{F},\, L_{\mathcal{X},2}(g) \neq -\infty\}$ and $\bar{f} \in \bar{\mathcal{F}}$. Then, $\sup_{g \in \bar{\mathcal{F}}} T_2(\bar{f}, g) = 0$ if and only if $\sup_{g \in \bar{\mathcal{F}}} L_{\mathcal{X},2}(g) - L_{\mathcal{X},2}(\bar{f}) = 0$.*

*Sketch of the proof.* We prove that $\sup_{g \in \bar{\mathcal{F}}} L_{\mathcal{X},2}(g) - L_{\mathcal{X},2}(\bar{f}) = 0$ implies $\sup_{g \in \bar{\mathcal{F}}} T_2(\bar{f}, g) = 0$. The proof of the converse is similar. Let $\bar{\mathcal{F}}_1 = \left\{ g \in \bar{\mathcal{F}},\, g = \bar{f},\, N \text{ a.s} \right\}$ and $\bar{\mathcal{F}}_2 = \bar{\mathcal{F}} \setminus \bar{\mathcal{F}}_1$. As in the proof of Claim 1, $T_2(\bar{f}, g) = (L_{\mathcal{X},2}(g) - L_{\mathcal{X},2}(\bar{f}))/4$ when $g \in \bar{\mathcal{F}}_1$ and is thus non-positive. Let now $g \in \bar{\mathcal{F}}_2$, $u \in [0,1]$ and $\zeta = g - \bar{f}$. Note that $\bar{f} + u\zeta = (1-u)\bar{f} + ug \in \bar{\mathcal{F}}$ and thus $L_{\mathcal{X},2}(\bar{f} + u\zeta) - L_{\mathcal{X},2}(\bar{f}) \leq 0$.

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

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

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

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

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

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

□

6.4. **Proof of Theorem 3.** Let $M_s$ be the (random) measure defined by

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

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

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

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

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

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

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

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

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

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

$$|Z_Q(\varphi)| \leq C\left\{\sqrt{\alpha \upsilon_Q(\varphi) \log_+(1/\upsilon_Q(\varphi))} + \alpha + \beta\right\}, \tag{36}$$

*where*

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

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

$$|Z_Q(\varphi)| \leq \varepsilon \upsilon_Q(\varphi) + C_\varepsilon \alpha \log_+(1/\alpha) \; + C\beta. \tag{37}$$

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

The proof of this result is delayed to Section 6.5 below. As $N$ is an empirical measure, we can prove that (35) is valid with $Q = N$ on an event of high probability by using, for example, the Vapnik-Chervonenkis inequalities for relative deviations. More precisely:

**Lemma 7.** *Let $\mathcal{A} \subset \mathcal{B}(\mathbb{R})$ be a collection of Borel sets and $S_{\mathcal{A}}(2n)$ be the Vapnik-Chervonenkis shatter coefficient of $\mathcal{A}$ defined by*

$$S_{\mathcal{A}}(2n) = \max_{x_1, \ldots, x_{2n} \in \mathbb{R}} |\{\{x_1, \ldots, x_{2n}\} \cap A, A \in \mathcal{A}\}|.$$

*Suppose that there exists an at most countable set $\mathcal{A}' \subset \mathcal{A}$ such that: for all $A \in \mathcal{A}$, there exists a sequence $(A_m)_{m \geq 0} \in \mathcal{A}'^{\mathbb{N}}$ satisfying $\lim_{m \to +\infty} \mathbb{1}_{A_m}(x) = \mathbb{1}_A(x)$ for every $x \in \mathbb{R}$.*

*Let now $\xi > 0$. Then, there exist a universal constant $c_1$ and an event $\Omega_{\xi,1}$ such that $\mathbb{P}[\Omega_{\xi,1}] \geq 1 - e^{-n\xi}$ and on which (35) holds for $Q = N$ and for all $A \in \mathcal{A}$ with $\alpha = c_1[\log_+ |S_{\mathcal{A}}(2n)| + n\xi]/n$ and $\beta = 0$.*

We now need to study the deviations of $Z_{M_s}$. There is nothing to prove in framework 1 since then $Z_{M_s} = 0$. The lemma below allows to control the deviations of $Z_{M_s}(\varphi)$ for all bounded function $\varphi$. It is proved in Section 6.7.

**Lemma 8.** *Define $G(t) = \mathbb{P}(X \geq t)$ in framework 2 and $G(t) = \mathbb{P}(X_{t-} = 1)\mathbb{1}_{I_{obs}}(t)$ in framework 3. For all $\xi > 0$, there exists an event $\Omega_{\xi,2}$ such that $\mathbb{P}[\Omega_{\xi,2}] \geq 1 - e^{-n\xi}$ and on which: for all $\varepsilon > 0$, and measurable function $\varphi$ such that $|\varphi(x)| \leq 1$ for all $x \in \mathbb{R}$,*

$$|Z_{M_s}(\varphi)| \leq \varepsilon \upsilon_{M_s}(\varphi) + C'_\varepsilon \inf_{\alpha \geq 0} \left\{ [\xi + 1/n] \int_0^\alpha s \, \mathrm{d}\mu + \int_\alpha^\infty sG \, \mathrm{d}\mu + \xi + 1/n \right\}. \tag{38}$$

*Moreover, in framework 2,*

$$|Z_{M_s}(\varphi)| \leq \varepsilon \upsilon_{M_s}(\varphi) + C''_\varepsilon (\xi + 1/n) \log_+ [1/(\xi + 1/n)]. \tag{39}$$

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

The computation of the infimum in (38) may be avoided in framework 3 when one restricts the collection among which the functions $\varphi$ may vary. More precisely, a bound on $Z_{M_s}(\varphi)$ may be deduced from the following result to be proved in Section 6.8 and from Lemma 7.

**Lemma 9.** *Consider framework 3. Let for $d \geq 1$, $\mathcal{I}_{2d}$ be the class of unions of at most $2d$ intervals. Then, there exist a universal constant $c_2$ and an event $\Omega_{\xi,2}$ such that $\mathbb{P}[\Omega_{\xi,2}] \geq 1 - e^{-n\xi}$ and on which (35) holds true for $Q = M_s$ for all $A \in \mathcal{I}_{2d}$ with $\alpha = \beta = c_2[d \log_+(n/d) + n\xi]/n$.*

We are now in position to prove (8). In framework 1, we use Lemmas 6 and 7 to deduce on $\Omega_{\xi,1}$:

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

Sauer's lemma shows that there exists $c'_1$ such that $\log_+ |S_{\mathcal{A}}(2n)| \leq c'_1 d \log_+(n/d)$. Inequality (8) then follows from elementary computations.

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

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

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

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

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

$$|Z(\varphi)| \leq \varepsilon(\upsilon_N(\varphi) + \upsilon_{M_s}(\varphi)) + (C_\varepsilon + C'_\varepsilon) \left[ \frac{d \log_+^2(n/d)}{n} + \xi \log_+(1/\xi) \right]. \tag{40}$$

Now,

$$\upsilon_N(\varphi) \leq \mathbb{E}\left[ \int_{\mathbb{R}} \varphi^2 \, \mathrm{d}N \right] = \mathbb{E}\left[ \int_{\mathbb{R}} \varphi^2 \, \mathrm{d}M_s \right]. \tag{41}$$

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

$$\begin{aligned}\left|\int_{\mathbb{R}} \varphi^2 \,\mathrm{d}M_s - \mathbb{E}\left[\int_{\mathbb{R}} \varphi^2 \,\mathrm{d}M_s\right]\right| &\leq \frac{1}{2}\mathbb{E}\left[\int_{\mathbb{R}} \varphi^4 \,\mathrm{d}M_s\right] + C(\xi + 1/n)\log_+\left[1/(\xi+1/n)\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[\int_{\mathbb{R}} \varphi^4 \,\mathrm{d}M_s\right] + C\left[\frac{d\log_+^2(n/d)}{n} + \xi\log_+(1/\xi)\right], \end{aligned} \tag{42}$$

where $C$ is universal. In framework 3,

$$\left\{x \in \mathbb{R},\, \varphi^2(x) > t\right\} = \left\{x \in \mathbb{R},\, \varphi_+(x) > \sqrt{t}\right\} \bigcup \left\{x \in \mathbb{R},\, \varphi_-(x) > \sqrt{t}\right\}$$

is a union of at most $2d$ intervals. Lemma 9 with Lemma 7 show that (42) remains valid on $\Omega_{\xi+(\log 2)/n,2}$. Moreover, as $|\varphi| \leq 1$,

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

By using (41),

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

We put this inequality in (40) and use $\upsilon_{M_s}(\varphi) \leq \upsilon(\varphi)$ to show (8) on $\Omega_{\xi+(\log 2)/n,1} \cap \Omega_{\xi+(\log 2)/n,2}$. Moreover,

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

□

6.5. **Proof of Lemma 6.** For convenience, and to make the proof more readable, we introduce a new notation. Given $x, y \in \mathbb{R}$, the assertion: there exists a universal constant $C$ such that $x \leq Cy$ is written in the sequel as $x \preccurlyeq y$. The claim below follows from elementary computations.

**Claim 3.** *When (35) holds true,*

$$|Q(A) - \mathbb{E}[Q(A)]| \preccurlyeq \sqrt{\alpha \min\{Q(A), \mathbb{E}[Q(A)]\}} + \alpha + \beta. \tag{43}$$

*Proof of Claim 3.* For reasons of symmetry, we may suppose that $Q(A) \geq \mathbb{E}[Q(A)]$ to prove (43). We derive from (35),

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

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

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

where $C$ is universal. This shows (43). □

Without loss of generality, we prove the lemma when the functions $\varphi$ of $\mathcal{F}$ are non-negative. We suppose moreover that we are on an event on which (35) holds true. Let for $t \in (0,1)$, $A_{\varphi,t} = \{x \in \mathbb{R},\, \varphi(x) > t\}$. As in [Bar16], the notion of integral is a great help: for all $x \in \mathbb{R}$,

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

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

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

we get

$$\begin{aligned}
|Z(\varphi)| - \varepsilon\upsilon(\varphi) &= \left|\int_0^1 \left(Q(A_{\varphi,t}) - \mathbb{E}[Q(A_{\varphi,t})]\right)\,\mathrm{d}t\right| - 2\varepsilon \min\left\{\int_0^1 tQ(A_{\varphi,t})\,\mathrm{d}t, \int_0^1 t\mathbb{E}[Q(A_{\varphi,t})]\,\mathrm{d}t\right\} \\
&\leq \int_0^1 \left\{|Q(A_{\varphi,t}) - \mathbb{E}[Q(A_{\varphi,t})]| - 2\varepsilon t \min\left\{Q(A_{\varphi,t}), \mathbb{E}[Q(A_{\varphi,t})]\right\}\right\}\,\mathrm{d}t \\
&\leq \int_0^\eta |Q(A_{\varphi,t}) - \mathbb{E}[Q(A_{\varphi,t})]|\,\mathrm{d}t \\
&\qquad + \int_\eta^1 \left\{|Q(A_{\varphi,t}) - \mathbb{E}[Q(A_{\varphi,t})]| - 2\varepsilon t \min\left\{Q(A_{\varphi,t}), \mathbb{E}[Q(A_{\varphi,t})]\right\}\right\}\,\mathrm{d}t. \qquad (44)
\end{aligned}$$

It follows from (43) and the inequality $2\sqrt{xy} \leq \varepsilon x + \varepsilon^{-1}y$,

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

We deduce,

$$|Z(\varphi)| - \varepsilon\upsilon(\varphi) \preccurlyeq \sqrt{\alpha}\int_0^\eta \sqrt{\mathbb{E}[Q(A_{\varphi,t})]}\,\mathrm{d}t + (\alpha/\varepsilon)\log(1/\eta) + \alpha + \beta. \qquad (45)$$

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

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

It remains to use $\mathbb{E}[Q(A_{\varphi,t})] \leq 1$ to prove (36).

Elementary computations then show (37). As $x \mapsto x\log_+(1/x)$ is non-decreasing, for all $x \leq y$, $xy\log_+(1/x) \leq y^2\log_+(1/y)$. Moreover, when $x \geq y$, $\log_+(1/x) \leq \log_+(1/y)$ and hence $xy\log_+(1/x) \leq xy\log_+(1/y)$. Therefore, for all $x, y > 0$,

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

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

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

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

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

This proves (37). □

6.6. **Proof of Lemma 7.** Let $A \in \mathcal{A}$ and $(A_m)_{m\geq 1}$ be the sequence given by the theorem. Then, $N(A_m)$ converges to $N(A)$. Moreover, it follows from the dominated convergence theorem that $\mathbb{E}\left[\int_{\mathbb{R}} |\mathbb{1}_{A_m} - \mathbb{1}_A| \, \mathrm{d}N\right]$ converges to 0. In particular, $\mathbb{E}[N(A_m)]$ converges to $\mathbb{E}[N(A)]$. We deduce that

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

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

**Theorem 11** (Vapnik-Chervonenkis inequalities for relative deviation)**.** *Let $Z_1, \dots, Z_n$ be $n$ independent and identically distributed random variables with values in a space $\mathcal{X}$. Let $\mathcal{B}$ be an at most countable collection of measurable sets. Define the empirical measure $\nu_n(B) = n^{-1}\sum_{i=1}^n \mathbb{1}_B(Z_i)$, $\nu(B) = \mathbb{E}[\mu_n(B)]$ and the Vapnik-Chervonenkis shatter coefficient*

$$S_{\mathcal{B}}(2n) = \max_{z_1,\dots,z_{2n}\in\mathcal{X}} |\{\{z_1,\dots,z_{2n}\}\cap B, B\in\mathcal{B}\}|.$$

*Then, for all $t > 0$,*

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

*In particular,*

$$\mathbb{P}\left(\sup_{B\in\mathcal{B}} \left|\sqrt{\nu_n(B)} - \sqrt{\nu(B)}\right| \geq t\right) \leq 8S_{\mathcal{B}}(2n)e^{-nt^2/4}. \tag{46}$$

Assume that we are within framework 1. Then, the random measure $N$ is the empirical measure of $X_1, \dots, X_n$. Now (46) with $\mathcal{B} = \mathcal{A}$,

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

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

The proof in frameworks 2 and 3 is very similar since $N$ is an empirical measure for suitable random variables with values in $\mathcal{X} = \mathbb{R}\times\{0,1\}$: $Z_i = (X_i, \mathbb{1}_{D_i=1})$ in framework 2 and $Z_i = (T_{1,0}^{(i)}\mathbb{1}_{T_{1,0}^{(i)}\in I_{\text{obs}}}, \mathbb{1}_{T_{1,0}^{(i)}\in I_{\text{obs}}})$ in framework 3. We apply (46) with $\mathcal{B} = \{A\times\{1\}, A\in\mathcal{A}\}$. Moreover,

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

We end the proof as in framework 1. □

6.7. **Proof of Lemma 8.** We define for $t \geq 0$, $V(t) = \mathbb{1}_{X \geq t}$ in framework 2 and $V(t) = \mathbb{1}_{X_{t-}=1}\mathbb{1}_{I_{\text{obs}}}(t)$ in framework 3.

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

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

The proof of this lemma is deferred to Sections 6.7.1 and 6.7.2 below. We now set for $i \in \{1, \ldots, n\}$, $t \geq 0$, $V_i(t) = \mathbb{1}_{X_i \geq t}$ in framework 2 and $V_i(t) = \mathbb{1}_{X^{(i)}_{t-}=1}\mathbb{1}_{I_{\text{obs}}}(t)$ in framework 3. Let then

$$G_n(t) = \frac{1}{n}\sum_{i=1}^n V_i(t).$$

Note that $G(t) = \mathbb{E}[G_n(t)]$. We define for $\alpha > 0$,

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

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

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

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

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

We deduce,

$$\mathbb{E}\left[\exp\left(\lambda R_\alpha\right)\right] \leq \frac{1}{\int_0^\alpha s\,\mathrm{d}\mu}\int_0^\alpha s(t)\mathbb{E}\left[\exp\left(u(\sqrt{G_n(t)} - \sqrt{G(t)})^2\right)\right]\mathrm{d}t. \tag{47}$$

Now,

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

The random variable $nG_n(t)$ is binomially distributed with parameters $(n, G(t))$. Therefore, we derive from Theorems 3 and 4 of [Oka59] that

$$\mathbb{P}\left[\left|\sqrt{G_n(t)} - \sqrt{G(t)}\right| \geq \sqrt{\frac{\log x}{u}}\right] \leq \frac{2}{x^{n/u}}.$$

Hence,

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

By (47), $\mathbb{E}\left[\exp\left(\lambda R_\alpha\right)\right] \leq 2$ and by Markov's inequality,

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

□

We now prove Lemma 8. We have,

$$\begin{aligned}|Z_{M_s}(\varphi)| &\leq \int_0^\infty s\left|\varphi\left(G_n-G\right)\right|\,\mathrm{d}\mu\\ &\leq \int_{G_n\leq G} s\left|\varphi\left(\sqrt{G_n}-\sqrt{G}\right)\left(\sqrt{G}-\sqrt{G_n}+2\sqrt{G_n}\right)\right|\,\mathrm{d}\mu\\ &\quad+\int_{G_n>G} s\left|\varphi\left(\sqrt{G_n}-\sqrt{G}\right)\left(\sqrt{G_n}-\sqrt{G}+2\sqrt{G}\right)\right|\,\mathrm{d}\mu\\ &\leq \int_0^\infty s|\varphi|\left(\sqrt{G_n}-\sqrt{G}\right)^2\,\mathrm{d}\mu+2\int_0^\infty s|\varphi|\left|\sqrt{G_n}-\sqrt{G}\right|\sqrt{\min\{G_n,G\}}\,\mathrm{d}\mu.\end{aligned}$$

Define

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

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

$$\begin{aligned}|Z_{M_s}(\varphi)| &\leq R+2\sqrt{R}\sqrt{\int_0^\infty s\varphi^2\min\{G_n,G\}\,\mathrm{d}\mu}\\ &\leq R+2\sqrt{R}\sqrt{\upsilon_{M_s}(\varphi)}.\end{aligned}$$

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

$$|Z_{M_s}(\varphi)| \leq \varepsilon\upsilon_{M_s}(\varphi)+(1+\varepsilon^{-1})R. \tag{48}$$

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

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

Now,

$$R\leq R_\alpha+\int_\alpha^\infty s\left(G_n+G\right)\,\mathrm{d}\mu. \tag{49}$$

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

$$R_\alpha \leq 3\left(\xi+\frac{\log 4}{n}\right)\int_0^\alpha s\,\mathrm{d}\mu. \tag{50}$$

Now,

$$\int_\alpha^\infty sG_n \,\mathrm{d}\mu = \frac{1}{n}\sum_{i=1}^n \int_\alpha^\infty s(t)V_i(t)\,\mathrm{d}t.$$

Claim 4 shows that the assumptions of Bernstein's deviation inequality are satisfied (Proposition 2.9 of [Mas07]). Therefore, with probability larger than $1-(1/2)e^{-n\xi}$,

$$\int_\alpha^\infty s\left[G_n - G\right]\,\mathrm{d}\mu \le 2\sqrt{\left(\xi+\frac{\log 2}{n}\right)\int_\alpha^\infty sG\,\mathrm{d}\mu} + \xi + \frac{\log 2}{n}.$$

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

$$\int_\alpha^\infty sG_n\,\mathrm{d}\mu \le 2\int_\alpha^\infty sG\,\mathrm{d}\mu + 2\left(\xi+\frac{\log 2}{n}\right). \tag{51}$$

By putting (50) and (51) into (49), we get with probability larger than $1-e^{-n\xi}$,

$$R \le 3\left[\frac{\log 4}{n}+\xi\right]\int_0^\alpha s\,\mathrm{d}\mu + 3\int_\alpha^\infty sG\,\mathrm{d}\mu + 2\xi + \frac{2\log 2}{n}.$$

It then remains to use the definition of $\alpha$ and (48) to prove (38). Moreover, in framework 2,

$$\int_\alpha^\infty sG\,\mathrm{d}\mu = \int_\alpha^\infty \frac{f(t)}{\mathbb{P}\left(T\ \ge t\right)}\mathbb{P}\left(T\ge t\right)\mathbb{P}\left(C\ge t\right)\mathrm{d}t \le \int_\alpha^\infty f\,\mathrm{d}\mu,$$

and

$$\int_0^\alpha s\,\mathrm{d}\mu = \int_0^\alpha \frac{f(t)}{\mathbb{P}\left(T\ \ge t\right)}\,\mathrm{d}t = -\log\left(\int_\alpha^\infty f\,\mathrm{d}\mu\right),$$

which gives (39). □

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

$$J_k = \int_{\substack{t_1,\dots,t_k\in A\\ t_1<t_2<\dots<t_k}} \left(\prod_{j=1}^k s(t_j)\right)\mathbb{P}\left(X\ge t_k\right)\,\mathrm{d}t_1\,\mathrm{d}t_2\dots\,\mathrm{d}t_k.$$

We have,

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

Now,

$$J_k \le \int_{\substack{t_1,\dots,t_{k-1}\in A\\ t_1<t_2<\dots<t_{k-1}}} \left(\prod_{j=1}^{k-1} s(t_j)\right)\left(\int_{t_{k-1}}^\infty s(t_k)\mathbb{P}\left(X\ge t_k\right)\,\mathrm{d}t_k\right)\,\mathrm{d}t_1\,\mathrm{d}t_2\dots\,\mathrm{d}t_{k-1},$$

and

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

Therefore,

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

By induction, $J_k \leq J_1$. □

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

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

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

*Then, for $\mu$-almost all $u > t$,*

$$\frac{\mathrm{d}\mu_B}{du}(u) = \mathbb{P}\left(B \text{ and } X_{u-} = 1\right) s(u). \tag{52}$$

*Proof.* First of all, $\mu_B$ is absolutely continuous with respect to the Lebesgue measure $\mu$ and admits therefore a Radon-Nikodym derivative. We now aim to show that this derivative is given by (52) for almost all $u > t$.

Let $Z_h(u)$ be the random variable giving the number of jumps of the Markov process in $[u-h, u+h]$. Then, $\mathbb{P}\left(Z_h(u) \geq 2\right) = o(h)$ when $h \to 0$. We deduce,

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

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

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

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

$$\begin{aligned}\mu_B([u,u+h]) &= \mathbb{P}\left(B, X_{u-} = 1\right)\mathbb{P}\left(X_{u+h} = 0 \mid X_{u-} = 1\right) + o(h) \\ &= \mathbb{P}\left(B, X_{u-} = 1\right)\frac{\mathbb{P}\left(X_{u-} = 1, X_{u+h} = 0\right)}{\mathbb{P}\left(X_{u-} = 1\right)} + o(h).\end{aligned} \tag{53}$$

Now,

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

Finally, by putting this inequality into (53),

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

which proves (52). □

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

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

We have,

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

Yet,

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

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

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

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

6.8. **Proof of Lemma 9.** Let $\overline{\mathcal{I}}_{2d}$ be the class of unions of at most $2d$ intervals with endpoints in $(\mathbb{Q} \cap [0, +\infty)) \cup \{+\infty\}$. Then, $\overline{\mathcal{I}}_{2d}$ is at most countable, and we only need to prove (35) when $A \in \overline{\mathcal{I}}_{2d}$.

The random measure $M_s$ is of the form

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

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

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

We measure the complexity of the family $\{f_A,\, A \in \overline{\mathcal{I}}_{2d}\}$ by means of the notion of entropy with bracketing:

**Claim 7.** *For all $\delta > 0$, there exists a collection $\mathcal{C}_\delta$ of functions of the form $f_A$ with $A \in \overline{\mathcal{I}}_{2d}$. The cardinal of this set can be bounded by $\log |\mathcal{C}_\delta| \leq cd \log_+(1/\delta^2)$, where $c$ is a universal constant. Moreover, for all $A \in \overline{\mathcal{I}}_{2d}$, there exist $f_{A_1}, f_{A_2} \in \mathcal{C}_\delta$ such that $f_{A_1} \leq f_A \leq f_{A_2}$ and such that for all $k \geq 1$,*

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

*Proof.* First of all, we only need to prove the claim when $\delta$ is smaller than 1, what we will do in the sequel.

We endow $\overline{\mathcal{I}}_{2d}$ with the distance *dist* defined for $A_1, A_2 \in \overline{\mathcal{I}}_{2d}$ by

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

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

$$dist(A_1, A_2) = \int_0^\infty \left| \mathbb{1}_{A_1}(t) - \mathbb{1}_{A_2}(t) \right| f(t) \, \mathrm{d}t,$$

where $f(t) = s(t)\mathbb{E}[V_1(t)]$ is a non-negative function satisfying $\int_0^\infty f(t) \, \mathrm{d}t \leq 1$.

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

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

Since $F$ is a continuous non-decreasing function such that $F([0, +\infty)) \subset [0, 1]$, there exist an even integer $\ell \in [2, 8d/\delta^2 + 2]$, and $\ell$ non-negative numbers $(x_1, x_2, \ldots,\ x_{\ell-1}, x_\ell) \in \{0\} \times \mathbb{Q}^{\ell-2} \times \{+\infty\}$ such that

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

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

When $k \leq \ell/2$, choosing $k$ disjoint closed intervals whose endpoints belong to $\mathcal{X}$ amounts to choosing $2k$ numbers among $\mathcal{X}$. When $k > \ell/2$, we cannot find $k$ disjoint closed intervals with

endpoints in $\mathcal{X}$. The cardinality of $\overline{\mathcal{I}}_{dis}$ is therefore bounded by

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

Standard arguments (see, for instance, exercise 2.14 of [BLM13]) show that $|\overline{\mathcal{I}}_{dis}| \leq (\ell e/(2d))^{2d}$. Using now that $\ell \leq 8d/\delta^2 + 2$, we derive that

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

for a suitable universal constant $c$.

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

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

and

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

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

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

and hence,

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

Now,

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

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

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

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

$$\log|\mathcal{B}_\delta| \leq \log 2 + \log|\mathcal{C}_\delta| \ \leq c_1 d \log_+(1/\delta^2),$$

where $c_1$ is a universal constant. We set $H(\delta) = c_1 d \log_+(1/\delta^2)$ and for $\sigma \in (0,1]$,

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

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

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

Consequently,

$$E \preccurlyeq \sigma\sqrt{nd\log_+(1/\sigma^2)} + d\log_+(1/\sigma^2). \tag{54}$$

Consider $\xi > 0$ and define $J$ as the (possibly empty) set of non-negative integers $j$ such that $2^{-j} \geq d/(2n)$. Let, for $j \in J$, $x_j = 2\log(j+1)+1+n\xi$, $\overline{\mathcal{A}}_j = \left\{A \in \overline{\mathcal{I}}_{2d},\, 2^{-j-1} \leq \mathbb{E}[M_s(A)] \leq 2^{-j}\right\}$. Claims 4 and 7 show that assumptions of Corollary 6.9 of [Mas07] are satisfied with $\mathcal{F} = \left\{f_A, -f_A,\, A \in \overline{\mathcal{A}}_j\right\}$, $\sigma^2 = 2^{-j+1}$, $b = 1$, and $H(\delta) = c_1 d\log_+(1/\delta^2)$. There exists therefore an event $\Omega_j$ such that $\mathbb{P}(\Omega_j) \geq 1 - e^{-x_j}$ and on which: for all $A \in \overline{\mathcal{A}}_j$,

$$n\,|M_s(A) - \mathbb{E}[M_s(A)]| \preccurlyeq E + \sigma\sqrt{nx_j} + x_j.$$

Hence,

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

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

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

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

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

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

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

We deduce from these computations that (35) holds true with $Q = M_s$ on the event $\Omega' \bigcap \left(\cap_{j\in J}\Omega_j\right)$ for all $A \in \bigcup_{j\in J}\overline{\mathcal{A}}_j \cup \overline{\mathcal{A}}$ with $\alpha$, $\beta$ of the form $c_2(d\log_+(n/d) + n\xi)/n$. Now, $\overline{\mathcal{I}}_{2d} = \bigcup_{j\in J}\overline{\mathcal{A}}_j \cup \overline{\mathcal{A}}$, and

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

□

6.9. **Proof of Proposition 4.** Suppose without loss of generality that the functions $\varphi$ are non-negative. Consider $\varepsilon > 0$ and $\eta \in (0,1)$. We derive from the beginning of the proof of Theorem 3 that for all $\varphi \in \mathcal{F}$, and $t \in (0,1)$, there exists $A_t \in \mathcal{A}_t$ (we omit the subscript $\varphi$) such that

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

Therefore,

$$\begin{aligned}\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}}|Z(\varphi)|\right] \leq \varepsilon\sigma^2 &+ \int_0^{\eta}\mathbb{E}\left[\sup_{A_t\in\mathcal{A}_t}|N(A_t) - \mathbb{E}[N(A_t)]|\right]\mathrm{d}t\\ &+ \int_{\eta}^1 \mathbb{E}\left[\sup_{A_t\in\mathcal{A}_t}\{|N(A_t) - \mathbb{E}[N(A_t)]| - 2\varepsilon t\mathbb{E}[N(A_t)]\}\right]\mathrm{d}t.\end{aligned} \tag{55}$$

Let now $\xi > 0$. As (35) holds true for all $A \in \mathcal{A}_t$, on an event $\Omega_{\xi,t}$ such that $\mathbb{P}\left[\Omega_{\xi,t}\right] \geq 1 - e^{-n\xi}$, with $\alpha = c[\log_+|S_{\mathcal{A}_t}(2n)| + n\xi]/n$, $\beta = 0$, we deduce from Claim 3 that on this event: for all $A_t \in \mathcal{A}_t$,

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

and using that $\sqrt{xy}\ \leq C/(8\varepsilon t)x + (2\varepsilon t/C)y$,

$$|N(A_t) - \mathbb{E}[N(A_t)]| - 2\varepsilon t\mathbb{E}[N(A_t)] \leq C'\left\{\frac{\log_+|S_{\mathcal{A}_t}(2n)| + n\xi}{n} + \frac{\log_+|S_{\mathcal{A}_t}(2n)| + n\xi}{n\varepsilon t}\right\}.$$

In these two inequalities, $C$ and $C'$ are universal constants.

We integrate these inequalities with respect to $\xi$ to get

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

and

$$\mathbb{E}\left[\sup_{A_t\in\mathcal{A}_t}\{|N(A_t) - \mathbb{E}[N(A_t)]| - 2\varepsilon t\mathbb{E}[N(A_t)]\}\right] \leq C''\left[\frac{\log_+|S_{\mathcal{A}_t}(2n)|}{n} + \frac{\log_+|S_{\mathcal{A}_t}(2n)|}{n\varepsilon t}\right]$$

where $C''$ is universal. We now deduce from (55),

$$\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}}|Z(\varphi)|\right] \leq \varepsilon\sigma^2 + \frac{C''}{n\varepsilon}\int_\eta^1 \frac{\log_+|S_{\mathcal{A}_t}(2n)|}{t}\,\mathrm{d}t$$
$$+\frac{C''}{\sqrt{n}}\int_0^\eta \sqrt{\left(\sup_{A_t\in\mathcal{A}_t}\mathbb{E}[N(A_t)]\right)\left(\log_+|S_{\mathcal{A}_t}(2n)|\right)}\,\mathrm{d}t + \frac{C''}{n}\int_0^1 \log_+|S_{\mathcal{A}_t}(2n)|\,\mathrm{d}t.$$

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

$$\mathbb{E}\left[\sup_{\varphi\in\mathcal{F}}|Z(\varphi)|\right] \leq \frac{C'''}{\sqrt{n}}\inf_{\eta\in(0,1)}\left\{\sigma\sqrt{\int_\eta^1 \frac{\log_+|S_{\mathcal{A}_t}(2n)|}{t}\,\mathrm{d}t} + \int_0^\eta \sqrt{\left(\sup_{A_t\in\mathcal{A}_t}\mathbb{E}[N(A_t)]\right)\left(\log_+|S_{\mathcal{A}_t}(2n)|\right)}\,\mathrm{d}t\right\}$$
$$+\frac{C'''}{n}\int_0^1 \log_+|S_{\mathcal{A}_t}(2n)|\,\mathrm{d}t,$$

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

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

$$\mathcal{X} = \{x\in\mathbb{R},\ g(x)\neq 0 \text{ or } f(x)\neq 0\}.$$

Then,

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

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

6.11. **Proof of Proposition 5 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

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

where $P_K$ and $Q_K$ are non-negative polynomial functions on $K$ 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 $\mathcal{X} = \{b_1, \ldots, b_k\}$ with $b_1 < b_2 < \cdots < b_k$ and $k \leq r$. We set $b_0 = -\infty$ and $b_{k+1} = +\infty$. For all $j \in \{0, \ldots, k\}$, $R_K$ 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. □

6.12. **Proof of Theorem 7.** 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 $\Omega_\xi$ such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$ and on which any $\rho$-estimator $\hat{s}$ on $S$ satisfies

$$h^2(s, \hat{s}) \leq \inf_{f \in \bar{S}} \left\{c_1 h^2(s, f) + c_2 \vartheta(d_S(f)) + c_3 \xi \log_+(1/\xi)\right\}. \tag{56}$$

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

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

Let $f, g \in \mathcal{S}$. 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 all odd integer $d \geq 3$,

$$\mathscr{G}_d = \{\psi(g/f),\ f, g \in \mathcal{S}, d_g(f) = (d-1)/2\}.$$

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

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

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

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

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

$$\{x \in \mathbb{R},\, \varphi_+(x) > t\} = \{x \in \mathbb{R},\, g(x) > uf(x)\}\,,$$

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

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

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

Now, for all $r \geq 1$,

$$\{x \in \mathbb{R},\, g(x) > (1 - 1/(2r))(1/u)f(x)\} \in \mathcal{I}_{(d-1)/2}.$$

Therefore,

$$\{x \in \mathbb{R},\, g(x) \geq (1/u)f(x)\} = \bigcap_{r=1}^{\infty} \{x \in \mathbb{R},\, g(x) > (1 - 1/(2r))(1/u)f(x)\}$$

belongs to $\mathcal{J}_d$ and $\{x \in \mathbb{R},\, \varphi_-(x) > t\} = \mathbb{R} \setminus \{x \in \mathbb{R},\, g(x) \geq (1/u)f(x)\}$ belongs to $\bar{\mathcal{J}}_d$. □

**Claim 9.** *The collection $\bar{\mathcal{J}}_d$ is Vapnik-Chervonenkis with dimension at most $2d-1 \leq 2d$. Moreover, in framework 3, each set $A \in \bar{\mathcal{J}}_d$ is a union of at most $(d+1)/2 \leq d$ intervals.*

*Proof.* Let $x_1, \ldots, x_{2n} \in \mathbb{R}$ and a non-increasing sequence $(A_r)_{r \geq 1}$. Then,

$$\left| \bigcap_{r=1}^{\infty} (\{x_1, \ldots, x_{2n}\} \cap A_r) \right| = \lim_{r \to +\infty} |\{x_1, \ldots, x_{2n}\} \cap A_r|\,.$$

The non-increasing sequence $(|\{x_1, \ldots, x_{2n}\} \cap A_r|)_{r \geq 1}$ consists of integers. Therefore, there exists $r_0$ such that $|\{x_1, \ldots, x_{2n}\} \cap A_r| = |\{x_1, \ldots, x_{2n}\} \cap A_{r_0}|$ for all $r \geq r_0$. Hence,

$$\bigcap_{r=1}^{\infty} (\{x_1, \ldots, x_{2n}\} \cap A_r) = \{x_1, \ldots, x_{2n}\} \cap A_{r_0}.$$

This implies that

$$\{\{x_1, \ldots, x_{2n}\} \cap A,\, A \in \mathcal{J}_d\} = \left\{\{x_1, \ldots, x_{2n}\} \cap A,\, A \in \mathcal{I}_{(d-1)/2}\right\}$$

and $S_{\mathcal{J}_d}(2n) = S_{\mathcal{I}_{(d-1)/2}}(2n)$. Therefore, $\mathcal{J}_d$ is Vapnik-Chervonenkis with dimension at most $d-1$. We deduce that $\bar{\mathcal{J}}_d$ is Vapnik-Chervonenkis with dimension at most $2(d-1)+1 \leq 2d$ (see, for instance, exercise 4.1 of [DL12]).

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

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

□

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

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

$$T(f,g) \leq (3+4\varepsilon)h^2(s,f) - \frac{4-3\varepsilon}{12}h^2(s,g) + c_1 \min\{\vartheta(d_g(f)), \vartheta(d_f(g))\} + c_2\xi \log_+(1/\xi). \tag{58}$$

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

*Proof.* Let $d \geq 3$ be an odd integer. Claims 8 and 9 show that the assumptions of Theorem 3 are satisfied with $\mathcal{F} = \mathcal{G}_d$. Therefore, there exists for all $\xi > 0$ an event $\Omega_\xi(d)$ such that $\mathbb{P}[\Omega_\xi(d)] \geq 1 - e^{-n\xi}$ and on which: for all $\varepsilon > 0$, $\varphi \in \mathscr{G}_d$ of the form $\varphi = \psi(g/f)$, with $f, g \in \mathcal{S}$,

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

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

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

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

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

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

$$|Z(\varphi)| \leq \varepsilon \upsilon(\varphi) + c' \min\{\vartheta(d_g(f)), \vartheta(d_f(g))\} + c\xi\log_+(1/\xi). \tag{59}$$

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

$$T(f,g) \leq 3h^2(s,f) - \frac{1}{3}h^2(s,g) + Z(\psi(g/f)). \tag{60}$$

Therefore, we deduce from Lemma 2 and from (59) that on $\Omega_\xi$: for all $f, g \in \mathcal{S}$,

$$T(f,g) \le (3+4\varepsilon)h^2(s,f) - \frac{4-3\varepsilon}{12}h^2(s,g) + c' \min\{\vartheta(d_g(f)), \vartheta(d_f(g))\} + c\xi \log_+(1/\xi),$$

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

We now finish the proof of Theorem 7. Lemma 10 implies that on $\Omega_\xi$: for all $f, g \in S$,

$$T(f,g) \le (3+4\varepsilon)h^2(s,f) - \frac{4-3\varepsilon}{12}h^2(s,g) + c_1\vartheta(d_g(f)) + c_2\xi\log_+(1/\xi). \tag{61}$$

Thus, for all $f \in S$,

$$\gamma(f) \le (3+4\varepsilon)h^2(s,f) - \frac{4-3\varepsilon}{12}h^2(s,S) + c_1 \sup_{g\in S}\vartheta(d_g(f)) + c_2\xi\log_+(1/\xi). \tag{62}$$

By using $T(f,g) = -T(g,f)$, we deduce from (61) that for all $f, g \in S$,

$$\frac{4-3\varepsilon}{12}h^2(s,g) - (3+4\varepsilon)h^2(s,f) - c_1\vartheta(d_g(f)) - c_2\xi\log_+(1/\xi) \le T(g,f).$$

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

$$\begin{aligned}\frac{4-3\varepsilon}{12}h^2(s,\hat{s}) - (3+4\varepsilon)h^2(s,f) - c_1\sup_{g\in S}\vartheta(d_g(f)) - c_2\xi\log_+(1/\xi) &\le T(\hat{s},f)\\ &\le \gamma(\hat{s})\\ &\le \gamma(f) + 1/n.\end{aligned} \tag{63}$$

Using now (62) and $1/n \le \vartheta(d_g(f))$, we deduce when $\varepsilon \in (0, 1/12)$,

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

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

When $f \in \bar{S}$ and $g \in S$, Assumption 2 says that $d_g(f)$ may be defined by $d_g(f) = d_S(f)$. Therefore, (64) becomes

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

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

6.13. **Proof of Proposition 8.** We keep the notations introduced in the proof of Theorem 7. It follows from (63) that on $\Omega_\xi$: for all $f \in S$,

$$\frac{4-3\varepsilon}{12}h^2(s,\hat{s}) - (3+4\varepsilon)h^2(s,f) - c_1\sup_{g\in S}\vartheta(d_g(f)) - c_2\xi\log_+(1/\xi) \le T(\hat{s},f),$$

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

$$h^2(s,\hat{s}) \le \inf_{f\in S}\left\{c_{1,\varepsilon}h^2(s,f) + c_{2,\varepsilon}\sup_{g\in S}\vartheta(d_g(f)) + c_{2,\varepsilon}\xi\log_+(1/\xi)\right\} \quad \text{with } c_{1,\varepsilon} = 12\frac{3+4\varepsilon}{4-3\varepsilon}.$$

It then remains to use Assumption 2 to define $d_g(f)$ by $d_g(f) = d_S(f)$ when $f \in \bar{S}$, $g \in S$. □

6.14. **Proofs of Theorems 9 and 10.** It is convenient for ease of demonstration to encompass the two procedures into a more general selection rule we now describe. Theorems 9 and 10 follow from Theorem 12 below. Their proofs are given in Sections 6.14.2 and 6.14.3.

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 $\gamma_4$ by

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

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

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

The risk of this estimator is bounded above as follows.

**Theorem 12.** *Let $(\mathcal{I}_d)_{d \geq 1}$ be a non-decreasing collection of Borel sets that fulfils Assumption 1, and $d_g(f)$ be the notation introduced in Section 6.12. Let for $\xi > 0$, $\Omega_\xi$ be the event given by Lemma 10. This event, which depends on $(\mathcal{I}_d)_{d \geq 1}$, is such that $\mathbb{P}(\Omega_\xi) \geq 1 - e^{-n\xi}$. We suppose that there exist two real valued maps, $\Delta \geq 0$ on $\hat{\Lambda}$, and $d \geq 1$ on $\mathcal{S}$ such that*

$$d_{\hat{s}_\lambda}(\hat{g}) \leq d(\hat{g}) + \Delta(\lambda) \quad \textit{for all } \lambda \in \hat{\Lambda},\ \hat{g} \in \hat{S}_\lambda. \tag{65}$$

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

$$\mathrm{pen}_{1,\lambda}(\hat{g}) \leq \mathrm{pen}_1(\hat{g}) + \mathrm{pen}_2(\lambda) \quad \textit{for all } \hat{g} \in \hat{S},\ \lambda \in \hat{\Lambda}. \tag{66}$$

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

$$\begin{aligned} \mathrm{pen}_{1,\lambda}(\hat{g}) &\geq L_1 \frac{d(\hat{g})}{n} \log_+^2\left(\frac{n}{d(\hat{g})}\right) \\ \mathrm{pen}_1(\hat{f}) &\geq L_1 \frac{d(\hat{f})}{n} \log_+^2\left(\frac{n}{d(\hat{f})}\right) \\ \mathrm{pen}_2(\lambda) &\geq L_1 \frac{\Delta(\lambda)}{n} \log_+^2\left(\frac{n}{\Delta(\lambda)}\right) \end{aligned} \tag{67}$$

*then, for all $\xi > 0$, on $\Omega_\xi$:*

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

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

6.14.1. *Proof of Theorem 12.* Let $\varepsilon \in (0, 1/12)$. The following assertion holds true on $\Omega_\xi$: for all $f, g \in \mathcal{S}$

$$T(f,g) \leq (3+4\varepsilon)h^2(s,f) - \frac{4-3\varepsilon}{12}h^2(s,g) + c_1 \min\{\vartheta(d_g(f)), \vartheta(d_f(g))\} + c_2\xi\log_+(1/\xi) \tag{68}$$

where $c_1, c_2$ only depend on $\varepsilon$.

Let $\lambda \in \hat{\Lambda}$ and $\hat{g} \in \hat{S}_\lambda$. We use (68) with $f = \hat{s}_\lambda$, $g = \hat{g}$:

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

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

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

If $L_1$ is large enough,

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

and hence

$$\gamma_4(\hat{s}_\lambda) \leq (3+4\varepsilon)h^2(s,\hat{s}_\lambda) + \mathrm{pen}_2(\lambda) + c_2\xi\log_+(1/\xi). \tag{69}$$

We now derive from (68) that for all $\hat{g} \in \hat{S}_{\hat{\lambda}}$,

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

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

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

If $L_1$ is large enough,

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

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

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

We deduce from (66),

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

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

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

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

6.14.2. *Proof of Theorem 9.* We will apply the selection rule developed in Section 6.14 to pick out an estimator among $\{\hat{s}_\lambda,\, \lambda \in \widehat{\Lambda}\} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$. For this purpose, we need to explain the values of the different parameters involved in the procedure. We define $\mathcal{I}_d$ as the collection of unions of at most $d$ intervals. We set $\hat{S} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$, and for $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$,

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

Note that the assumption $\hat{S} \subset \bigcap_{m\in\widehat{\mathcal{M}}_{\hat{\ell}}}\hat{S}_m$ of Theorem 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 (23). A function $\hat{g} \in \hat{S}_m$ of the form $\hat{g} = \sum_{K\in m}\hat{s}_{m_K}\mathbb{1}_K$ is piecewise polynomial. 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

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

Let $\bar{S} = \bigcup_{k=1}^{\infty}\mathcal{P}_{k,r}$ and note that $\hat{S}_m \subset \bar{S}$ for all $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$. Let $f \in \bar{S}$ and $k \geq 1$ be the smallest integer for which $f \in \mathcal{P}_{k,r}$. It follows from Proposition 5 that Assumption 2 is satisfied with $S = \mathcal{P}_{\hat{\ell}\vee k,r}$ and $d_{\mathcal{P}_{\hat{\ell}\vee k,r}}(f) = (r+2)(2(\hat{\ell}\vee k)+1)$. In particular, for all $m \in \widehat{\mathcal{M}}_{\hat{\ell}}$ and $f \in \bar{S}$, we may set since $\hat{s}_m \in \mathcal{P}_{\hat{\ell},r}$,

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

We now define $d$ for $f \in \bar{S}$ and $\Delta$ 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 (65) 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 (67) 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 $\hat{g} \in \hat{S} = \{\hat{s}_m,\, m \in \widehat{\mathcal{M}}_{\hat{\ell}}\}$.

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

*Proof of Claim 10.* We have,

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

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

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 (66).

Note now that the selection rules described in Sections 6.14 and 4.2 coincide. Theorem 12 controls the risk of the selected estimator: for all $\xi > 0$, there exists an event $\Omega_\xi$ of probability larger than $1 - e^{-n\xi}$, and on which:

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

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

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

where $C'$ is a universal constant. It then remains to use the fact that $\hat{s}_m$ is a $\rho$-estimator on $\mathcal{P}_r(m)$ to get a bound on $h^2(s, \hat{s}_m)$ on the same event $\Omega_\xi$ (the event that appears in Theorem 7 to control the risk of a $\rho$-estimator is the same that the one that appears in Theorem 12. It is, in each case, defined by Lemma 10). □

6.14.3. *Proof of Theorem 10.* The proof is almost the same than the one of Theorem 9. 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 6.14:

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

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

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

□

6.15. **Proof of Lemma 3.** As in the proof of Theorem 1, the measure $N$ can be put of the form $N(A) = n^{-1}\sum_{i\in\hat{I}} \mathbb{1}_A(Y_i)$ where $\hat{I} \subset \{1,\dots,n\}$, and where the $Y_i$ are suitable real-valued random variables.

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

6.16. **Proof of Lemmas 4 and 5.** The following claim will be useful in the sequel.

**Claim 11.** *Let $\xi > 0$ and $\Omega_\xi$ be the event given by Lemma 10 when $\mathcal{I}_d$ is the collection of unions of at most $d$ intervals. Then, $\mathbb{P}(\Omega_\xi) \ge 1 - e^{-n\xi}$. Let $\eta \ge 0$, $r \ge 0$, and $m, m' \in \mathcal{M}$. The following holds on $\Omega_\xi$: for all piecewise polynomial estimators $\hat{s}_m \in \mathcal{P}_r(m)$, $\hat{s}_{m'} \in \mathcal{P}_r(m')$ such that $T(\hat{s}_m, \hat{s}_{m'}) \ge -\eta$,*

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

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

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

*In the above inequalities, $C$ is universal.*

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

$$T(\hat{s}_m, \hat{s}_{m'}) \le (3+4\varepsilon)h^2(s, \hat{s}_m) - \frac{4-3\varepsilon}{12}h^2(s, \hat{s}_{m'}) + c_1\vartheta(d_{\hat{s}_{m'}}(\hat{s}_m)) + c_2\xi\log_+(1/\xi), \tag{70}$$

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

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

Yet, $|m''| \le |m| + |m'|$. Thereby, $\hat{s}_m$ and $\hat{s}_{m'}$ belong to $\mathcal{P}_{|m|+|m'|,r}$ and it follows from Proposition 5 that we may set

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

We now bound above $\vartheta(d_{\hat{s}'_m}(\hat{s}_m))$ in (70), and then use $T(\hat{s}_m, \hat{s}_{m'}) \ge -\eta$ to prove the first inequality of the claim. The second one follows from Theorem 7 and Proposition 5. □

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

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

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

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

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

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

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

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

We now prove that $f \leq \hat{s}_{\bar{m}}$. When $x \notin [Y_{(1)}, Y_{(\hat{n})}]$, $f(x) = \hat{s}_{\bar{m}}(x) = 0$. When $x \in [Y_{(1)}, Y_{(\hat{n})}]$, there exist $j \in \{1, \dots, \ell\}$ such that $x \in I_j$ and $j' \leq j$ such that $x \in J_{j'}$. Therefore, $f(x) = \hat{s}_{I_{j'}}$. By using that $\hat{s}_{I_{j'}} \leq \hat{s}_{I_j}$, we finally deduce that $f(x) \leq \hat{s}_{\bar{m}}(x)$.

Consider an interval $I_j \in \bar{m}$ and let us denote the cardinal of $\left\{Y_{(i)},\, Y_{(i)} \in I_j,\, i \in \{1, \dots, \hat{n}\}\right\}$ by $k_j$. When $k_j \geq 3$, there exists at least $k_j - 2$ random variables $Y_{(i)}$ that belong to $I_j$ but not to $\cup_{\substack{j' \in \{1,\dots,\hat{n}\} \\ j' \neq j}} K_{j'}$. Such $Y_{(i)}$ belong therefore to $J_j$ and satisfy $f(Y_{(i)}) = \hat{s}_{\bar{m}}(Y_{(i)})$. Therefore,

$$\begin{aligned} \left|\left\{Y_{(i)},\, f(Y_{(i)}) \neq \hat{s}_{\bar{m}}(Y_{(i)}),\, i \in \{1, \dots, \hat{n}\}\right\}\right| &= \sum_{j=1}^{\ell} \left|\left\{Y_{(i)},\, f(Y_{(i)}) \neq \hat{s}_{\bar{m}}(Y_{(i)}),\, Y_{(i)} \in I_j,\, i \in \{1, \dots, \hat{n}\}\right\}\right| \\ &\leq 2\ell. \end{aligned} \tag{71}$$

It follows from $f \leq \hat{s}_{\bar{m}}$, (71) and (28) that $T(f, \hat{s}_{\bar{m}}) \leq 2\ell/n$. We now use Claim 11 to get on $\Omega_\xi$

$$h^2(s, f) \leq C \left\{ h^2(s, \hat{s}_{\bar{m}}) + \frac{\ell}{n} \log_+^2 (n/\ell) + \xi \log_+(1/\xi) \right\}, \tag{72}$$

where $C$ is universal.

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

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

By using that $f \in \mathcal{P}_0(\bar{m}') \subset \mathcal{P}_0(m')$ and (72),

$$\begin{aligned} h^2(s, \hat{s}_{m'}) &\leq C' \left\{ h^2(s, f) + \frac{\ell}{n} \log_+^2 (n/\ell) + \xi \log_+(1/\xi) \right\}, \\ &\leq C'' \left\{ h^2(s, \hat{s}_{\bar{m}}) + \frac{\ell}{n} \log_+^2 (n/\ell) + \xi \log_+(1/\xi) \right\}, \end{aligned} \tag{73}$$

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

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

Therefore, Claim 11 asserts that

$$h^2(s, \hat{s}_{\bar{m}}) \leq C'''' \left\{ h^2(s, \hat{s}_m) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\}, \tag{74}$$

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

$$h^2(s, \hat{s}_m) \leq C''''' \left\{ h^2(s, \mathcal{P}_0(m)) + \frac{\ell}{n}\log_+^2(n/\ell) + \xi \log_+(1/\xi) \right\}. \tag{75}$$

It remains to put inequalities (73), (74) and (75) together to finish the proof. □

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

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

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

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

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

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

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

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

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

For each $j \in \{1, \ldots, \ell\}$, 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 $j \in \{2, \ldots, \ell - 1\}$, $J_{2j} = (\beta_j, \alpha_{j+1}]$ and for $j \in \{2, \ldots, \ell\}$, $J_{2j-1} = (\alpha_j, \beta_j]$. 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 $j \in \{1, \ldots, \ell\}$. The collection $m' = \{J_j, j \in \{1, \ldots, 2\ell - 1\}\}$ defines a partition of $\widehat{\mathcal{M}}$ such that $|m'| \leq 2\ell - 1$. We define the $\rho$-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 11 that on $\Omega_\xi$:

$$h^2(s, \hat{s}_{m'}) \leq C \left\{ h^2(s, \tilde{s}_{m'}) + \frac{(r+1)|m|}{n} \log_+^2 \left( \frac{n}{(r+1)|m|} \right) + \xi \log_+(1/\xi) \right\}, \tag{76}$$

where $C$ is universal.

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

$$T\left(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} \mathbb{1}_{J_{2j-1}}\right) \leq 0, \tag{77}$$

since $\hat{s}_{J_{2j-1}}$ maximizes $L_{J_{2j-1}}$ over $\mathcal{P}_r(J_{2j-1})$ and that $\hat{s}_{K_j} \mathbb{1}_{J_{2j-1}} \in \mathcal{P}_r(J_{2j-1})$. When $J_{2j-1} = \emptyset$, $T\left(\hat{s}_{J_{2j-1}}, \hat{s}_{K_j} \mathbb{1}_{J_{2j-1}}\right) = 0$, and thus (77) also holds.

We define

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

We deduce from (77) that $T(\tilde{s}_{m'} \mathbb{1}_A, \hat{s}_m \mathbb{1}_A) \leq 0$. Therefore,

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

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

$$T(\tilde{s}_{m'}, \hat{s}_m) \leq \int_{B \cap A^c} \psi\left(\hat{s}_m/0\right) \, \mathrm{d}N \leq N(B \cap A^c). \tag{78}$$

Now,

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

Since $\alpha_j, \beta_j \in \{Y_i, i \in \hat{I}\}$, we deduce from (28) that $N(K_j) - N(J_{2j-1}) = N(\{\alpha_j\})$. In each of the frameworks, $N(\{\alpha_j\}) \leq 1/n$ and thus $N(B \cap A^c) \leq \ell/n$. By using (78), we get $T(\tilde{s}_{m'}, \hat{s}_m) \leq \ell/n$. Claim 11 with $\eta = \ell/n \leq |m|/n$ ensures that on $\Omega_\xi$:

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

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

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

where $C''$ is universal. It then remains to combine the two last inequalities with (76) to finish the proof. □

## References


[AD10] Nathalie Akakpo and Cécile Durot. Histogram selection for possibly censored data. *Mathematical Methods of Statistics*, 19(3):189–218, 2010.

[Ant89] Anestis Antoniadis. A penalty method for nonparametric estimation of the intensity function of a counting process. *Annals of the Institute of Statistical Mathematics*, 41(4):781–807, 1989.

[Bar11] Yannick Baraud. Estimator selection with respect to Hellinger-type risks. *Probability Theory and Related Fields*, 151(1-2):353–401, 2011.

[Bar16] Yannick Baraud. Bounding the expectation of the supremum of an empirical process over a (weak) vc-major class. *Electronic journal of statistics*, 10(2):1709–1728, 2016.

[BB09] Yannick Baraud and Lucien Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probability Theory and Related Fields*, 143:239–284, 2009.

[BB16] Yannick Baraud and Lucien Birgé. $\rho$-estimators for shape restricted density estimation. *Stochastic Processes and their Applications*, 126(12):3888–3912, 2016.

[BB17] Yannick Baraud and Lucien Birgé. $\rho$-estimators revisited: general theory and applications. *arXiv preprint*, 2017.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[BBS17] Yannick Baraud, Lucien Birgé, and Mathieu Sart. A new method for estimation and model selection: $\rho$-estimation. *Inventiones mathematicae*, 207(2):425–517, 2017.

[BC05] Elodie Brunel and Fabienne Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā: The Indian Journal of Statistics*, pages 441–475, 2005.

[BC08] Elodie Brunel and Fabienne Comte. Adaptive estimation of hazard rate with censored data. *Communications in Statistics—Theory and Methods*, 37(8):1284–1305, 2008.

[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.

[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.

[GK06] Evarist Giné and Vladimir Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

[GN15] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

[Gre56] Ulf Grenander. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(1):70–96, 1956.

[Kan92] Yuichiro Kanazawa. An optimal variable cell histogram based on the sample spacings. *The Annals of Statistics*, 20(1):291–304, 1992.

[Mas07] Pascal Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2007. École d'été de Probabilités de Saint-Flour.

[Oka59] Masashi Okamoto. Some inequalities relating to the partial sum of binomial probabilities. *Annals of the institute of Statistical Mathematics*, 10(1):29–35, 1959.

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

[RB06] Patricia Reynaud-Bouret. Penalized projection estimators of the aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006.

[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.

[Sau72] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[vdG95] Sara van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, 23(5):1779–1801, 1995.

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

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