Journal of Machine Learning Research 23 (2022) 1-59 Submitted 6/19; Revised 10/21; Published 12/21

Empirical Risk Minimization under Random Censorship

Guillaume Ausset GUILLAUME.AUSSET@TELECOM-PARIS.FR
LTCI, Télécom Paris, Institut Polytechnique de Paris
BNP Paribas

Stephan Clémençon STEPHAN.CLEMENCON@TELECOM-PARIS.FR
LTCI, Télécom Paris, Institut Polytechnique de Paris

François Portier FRANCOIS.PORTIER@ENSAI.FR
CREST UMR 9194, ENSAI, Univ Rennes

Editor: Ingo Steinwart

Abstract

We consider the classic supervised learning problem where a continuous non-negative random label Y (e.g. a random duration) is to be predicted based upon observing a random vector X valued in $\mathbb{R}^d$ with $d \geq 1$ by means of a regression rule with minimum least square error. In various applications, ranging from industrial quality control to public health through credit risk analysis for instance, training observations can be *right censored*, meaning that, rather than on independent copies of (X, Y), statistical learning relies on a collection of $n \geq 1$ independent realizations of the triplet $(X,\ \min\{Y,\ C\},\ \delta)$, where C is a nonnegative random variable with unknown distribution, modelling censoring and $\delta = \mathbb{I}\{Y \leq C\}$ indicates whether the duration is right censored or not. As ignoring censoring in the risk computation may clearly lead to a severe underestimation of the target duration and jeopardize prediction, we consider a *plug-in* estimate of the true risk based on a Kaplan-Meier estimator of the conditional survival function of the censoring C given X, referred to as *Beran risk*, in order to perform empirical risk minimization. It is established, under mild conditions, that the learning rate of minimizers of this biased/weighted empirical risk functional is of order $O_{\mathbb{P}}(\sqrt{\log(n)/n})$ when ignoring model bias issues inherent to plug-in estimation, as can be attained in absence of censoring. Beyond theoretical results, numerical experiments are presented in order to illustrate the relevance of the approach developed.

Keywords: Censored data, empirical risk minimization, U-processes, statistical learning theory, survival data analysis.

1. Introduction

Covering a wide variety of practical applications, distribution-free regression can be considered one of the flagship problems in statistical learning. In the most standard setup, (X, Y) is a random pair defined on a certain probability space with (unknown) joint probability distribution P, where the output Y is a real-valued square integrable random variable (r.v.) and X models some input information, valued in $\mathbb{R}^d$, supposedly useful to predict Y. In this context, one is interested in building a (measurable) function $f : \mathbb{R}^d \to \mathbb{R}$ minimizing the (expected quadratic) risk

$$R_P(f) = \mathbb{E}\left[(Y - f(X))^2\right], \tag{1}$$

which is finite as soon as the r.v. $f(X)$ is square integrable. Obviously, the minimizer of (1) is the *regression function* $f^\star(X) = \mathbb{E}[Y \mid X]$. As the distribution of (X, Y) is unknown in practice, the Empirical Risk Minimization paradigm (ERM in abbreviated form, see e.g. Györfi et al. (2002)) suggests considering solutions $\widehat{f}_n$ of the minimization problem, also referred to as *least squares regression*, $\min_{f\in\mathcal{F}} \widehat{R}_n(f)$, where $\widehat{R}_n(f)$ is a statistical estimate of the risk $R_P(f)$ computed from a training sample $\mathcal{D}_n = \{(X_1, Y_1),\ \dots,\ (X_n, Y_n)\}$ of independent copies of (X, Y). In general the empirical version

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2, \tag{2}$$

is considered. This boils down to replacing P in the risk functional $R_P(\cdot)$ with the empirical distribution of the (X_i, Y_i)'s. The class $\mathcal{F}$ of predictive functions is supposed to be of controlled complexity (e.g. of finite VC dimension), while being rich enough to contain a reasonable approximant of the minimizer of R_P, $f^\star(x)$. In a framework stipulating in addition that the random variables Y and $f(X)$, $f \in \mathcal{F}$, are sub-Gaussian, ERM is proved to yield rules with good generalization properties, see e.g. Györfi et al. (2002); Bartlett et al. (2005); Lecué and Mendelson (2016) (notice, however, that, in heavy-tail situations, alternative strategies are preferred, refer to Lugosi and Mendelson (2016) for instance).

In many applications such as industrial reliability, see Mann et al. (1974), or clinical trials, the r.v. Y to be predicted represents a duration, e.g. the lifespan of a manufactured component or the time to recovery of sick patients, and it is far from uncommon in survival analysis that the data at disposal to learn a predictive rule are not composed of independent realizations $(X_1, Y_1),\ \dots,\ (X_n, Y_n)$ of distribution P but of observations $(X_1,\ \tilde{Y}_1,\ \delta_1),\ \dots,\ (X_n,\ \tilde{Y}_n,\ \delta_n)$, where the observed durations are of the form

$$\tilde{Y}_i = \min\{C_i,\ Y_i\} \text{ with } i \in \{1,\ \dots,\ n\},$$

the random variables C_i modelling a possible right censoring, and the δ_i are binary variables indicating whether censoring has occurred for each duration. Of course, other types of censoring (e.g. left/interval/progressive censoring) can be encountered in practice and result in partially observed durations. Since the results established in this paper can be straightforwardly extended to a more general framework, focus is here on the right censoring case. Whereas the asymptotic theory of statistical estimation based on censored data is very well documented in the literature (see e.g. Fleming and Harrington (1991); Andersen et al. (1993) and the references therein), the issues raised by censoring in statistical learning has received much less attention and it is the major purpose of this article to investigate how ERM can be extended to this setup with sound generalization guarantees.

As the empirical risk (2) cannot be computed directly from the data available, we rely on the *inverse of the probability of censoring weighted* (IPCW) approach, see Gerds et al. (2017) for a recent review. In that, we build first a plug-in (biased) estimator of the risk (1) by means of the Beran estimator of the conditional survival function of the censoring (Beran, 1981; Dabrowska, 1989; van Keilegom and Veraverbeke, 1996) and minimize next the resulting risk estimate, referred to as *Beran risk* and that can be interpreted as a weighted version of the empirical risk process based on the observations. The asymptotic behaviour of such weighted empirical risk has been first considered in the seminal contributions of Stute (1993, 1996) and refined recently in Lopez (2011); Lopez et al. (2013).

In this paper, more in the spirit of the popular statistical learning theory of empirical risk minimization, nonasymptotic maximal deviation bounds for this risk functional, much more complex than a basic empirical process due to the strong dependency exhibited by the terms averaged to compute it, are established by means of linearization techniques combined with concentration results pertaining to the theory of U-processes. We prove that, under appropriate conditions, minimizers of the Beran risk proposed have good generalization properties, achieving learning rate bounds of order $O_{\mathbb{P}}(\sqrt{\log(n)/n})$ when ignoring the model bias impact on the plug-in estimation step, the same as ERM in absence of any censoring. Beyond this theoretical analysis, illustrative numerical results are also displayed, providing strong empirical evidence of the relevance of the approach promoted. They reveal in particular that, even if the estimator of the conditional survival function plugged is only moderately accurate, Beran risk minimizers significantly outperform approaches ignoring censoring. Eventually, we point out that some of the results established in this paper have been preliminarily presented in an elementary form at the 2018 NeurIPS Machine Learning for Healthcare workshop (ML4HEALTH), see Ausset et al. (2018).

The rest of the paper is organized as follows. The framework we consider for statistical learning based on censored training data is detailed in section 2, where notions pertaining to survival data analysis involved in the subsequent study are also briefly recalled and a nonasymptotic uniform bound for the Beran estimator of the conditional survival function of the censoring is also stated. In section 3, the statistical version of the expected quadratic risk we propose, based on the Beran estimator previously studied, is introduced and the performance of its minimizers is analysed. Illustrative numerical results are displayed in section 4, while several concluding remarks are collected in section 5. Technical proofs are postponed to the appendices.

2. Background - Preliminaries

In this section, we first describe at length the probabilistic setup considered in this paper and recall basic concepts of *censored data analysis*, which the subsequent analysis relies on. Next, we establish a nonasymptotic bound for the deviation between the conditional survival function of the random censoring and its Beran estimator under adequate smoothness assumptions. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point x by δ_x. When well-defined, the convolution product between two real-valued Borelian functions on $\mathbb{R}^d$ $g(x)$ and $w(x)$ is denoted by $(g * w)(x) = \int_{x' \in \mathbb{R}^d} g(x - x')w(x')dx'$. The left-limit at $s > 0$ of any càdlag̀ function S on $\mathbb{R}_+$ is denoted by $S(s-) = \lim_{t \uparrow s} S(t)$.

2.1 The Statistical Framework

In this paper, we consider a pair (X, Y) of random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with unknown joint distribution P and where Y, representing a duration, takes positive values only and X models some information valued in $\mathbb{R}^d$, $d \geq 1$, a priori useful to predict Y. We assume that X's marginal distribution has a density $g(x)$ w.r.t. Lebesgue measure on $\mathbb{R}^d$. We are concerned with building a prediction rule $f : \mathbb{R}^d \to \mathbb{R}_+$ with minimum expected quadratic risk $R_P(f)$, see Eq. (1), based on a training dataset $\mathcal{D}_n = \{(X_1,\ \tilde{Y}_1,\ \delta_1),\ \ldots,\ (X_n,\ \tilde{Y}_n,\ \delta_n)\}$ composed of $n \geq 1$ independent realizations of the random triplet $(X,\ \tilde{Y},\ \delta)$, where $\tilde{Y} = \min\{Y,\ C\}$, C is a positive r.v. defined on $(\Omega, \mathcal{A}, \mathbb{P})$ and $\delta = \mathbb{I}\{Y \leq C\}$ indicates whether the duration is (right) censored ($\delta = 0$) or not ($\delta = 1$). The following hypothesis is required in the present study.

Assumption 1 (Conditional independence) *The random variables Y and C are conditionally independent given the input X and we have $Y \neq C$ with probability one.*

Naturally, many other types of censoring can be encountered in practice. However, since the goal of the present paper is to explain the main ideas to apply the ERM principle to censored data rather than dealing with the problem at the highest level of generality, we restrict our attention to the type of right random censoring introduced above. Though simple, it covers many situations. Addressing the problem in a more complex probabilistic framework, where Y and C are not conditionally independent given X anymore for instance, will be the subject of future research. The assumption stipulating that $\{Y = C\}$ is a zero-probability event is quite general, insofar as it allows considering situations where Y and/or C are discrete variables. Under conditional independence, it is obviously satisfied when the r.v. Y is continuous.

Easy to state but difficult to solve, the statistical learning problem we consider here is of considerable importance. In a wide variety of applications, the input information is of increasing granularity and described by a random vector of very large dimension d, while (censored) data are progressively becoming massively available. Machine-learning techniques are thus expected to complement traditional approaches, based on statistical modelling, in order to produce more flexible/accurate predictive models based on censored data. Incidentally, we point out that the problem under study can be viewed as a very specific type of *transfer learning* problem, see e.g. Pan and Yang (2010) insofar as, due to the censoring, the distribution of the training/source data is not that of the test/target data. However, the source domain coincides here with the target one and the predictive task (regression) remains the same.

Weighted empirical risk. Discarding censored observations to evaluate the risk of a candidate function $f(x)$ would lead to the quantity

$$\bar{R}_n(f) = \sum_{i=1}^{n} \delta_i \left(\tilde{Y}_i - f(X_i)\right)^2 / \sum_{i=1}^{n} \delta_i, \tag{3}$$

with $0/0 = 0$ by convention, which is clearly a biased estimate of $R_P(f)$ in general, since, by virtue of the strong law of large numbers, it converges to $\mathbb{E}[(Y - f(X))^2 \mid Y \leq C]$ with probability one. One may easily check that the minimizer of this functional is given by

$$\bar{f}^*(X) = \mathbb{E}[Y\mathbb{I}\{Y \leq C\} \mid X]/\mathbb{P}\{Y \leq C \mid X\},$$

which significantly differs from $f^*(X)$ in general. Observing that, by means of a straightforward conditioning argument, one can write the risk as

$$R_P(f) = \mathbb{E}\left[\frac{\delta(\tilde{Y} - f(X))^2}{S_C(\tilde{Y}- \mid X)}\right], \tag{4}$$

where $S_C(u \mid x) = \mathbb{P}\{C > u \mid X = x\}$ denotes the conditional survival function of the random right censoring given X, we propose to estimate the risk (1) by computing first a nonparametric estimator $\hat{S}_C(u \mid x)$ of $S_C(u \mid x)$ and by plugging it next into (4), so as to obtain

$$\widetilde{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i(\tilde{Y}_i - f(X_i))^2}{\hat{S}_C(\tilde{Y}_i- \mid X_i)}, \tag{5}$$

which approximates the unknown quantity whose expectation is equal to (4)

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i(\tilde{Y}_i - f(X_i))^2}{S_C(\tilde{Y}_i- \mid X_i)}, \tag{6}$$

the conditional survival function of C given X being itself unknown. Observe that the risk estimate (5) can be viewed as a *weighted version* of the sum of the observed squared errors $(\tilde{Y}_i - f(X_i))^2$, just like (3) except that the i-th weight is not $\delta_i / \sum_{j \le n} \delta_j$ anymore but $\delta_i / \hat{S}_C(\tilde{Y}_i- \mid X_i)$. In the terminology of survival analysis, the weighted empirical risk (5) is usually referred to as an IPCW (Gerds et al., 2017) estimate and a natural strategy to learn a predictive function in the censored framework described above then consists in solving the minimization problem

$$\inf_{f \in \mathcal{F}} \widetilde{R}_n(f), \tag{7}$$

over an appropriate class $\mathcal{F}$. In Rotnitzky and Robins (1992) or Section 3.3 in van der Laan and Robins (2003), a parametric estimator (Cox, 1972) of $S_C(Y \mid X)$ is used to infer the risk. However, such an approach is naturally limited by well-known misspecification issues, inherent in the choice of the parametric model. In van der Laan and Robins (2003); Rubin and van der Laan (2007), the misspecification problem is alleviated by the use of a *doubly robust loss* which allows for misspecification of one of the models (either S_C or else S_T) and improves in addition the efficiency of the procedure. This has been further investigated in Molinaro et al. (2004); Steingrimsson et al. (2016, 2019) , where different methodologies are proposed to build classification trees. The use of the Kaplan-Meier estimator (Kaplan and Meier, 1958) for $S_C(u \mid x)$ has been considered in several papers (Stute, 1993, 1996; Bang and Tsiatis, 2002; Kohler et al., 2002). Even if the censoring model is free from any parametric modelling, the assumptions required to ensure consistency are quite strong as the distribution of C is supposed to be independent from X, see Stute (1996) for more details. In particular, the weights used are independent from X. To overcome the previous restrictions, the Beran estimator (Beran, 1981), which is a kernel smoothing version of the Kaplan-Meier estimate (see the next section), can be employed instead of the Cox or the Kaplan-Meier estimators. Such an approach is promoted and studied in Lopez (2011); Lopez et al. (2013). When using the Beran approach to estimate $S_C(u \mid x)$, as detailed in the next subsection, the risk functional (5) is referred to as the *Beran risk* throughout the article. Based on accuracy results for kernel-based Beran estimators of the conditional survival function $S_C(\cdot \mid x)$ such as those subsequently presented, the performance of solutions of (7) is investigated in the next section. We point out that, as highlighted in section 4, alternative inference strategies for conditional survival function estimation can be considered. For simplicity, here we restrict our attention to kernel-smoothing techniques, although the analysis carried out can be extended to other nonparametric methods (e.g. partition-based techniques, nearest neighbours).

Our results are therefore comparable to those of Lopez (2011), as both studies are concerned with the Beran risk, relaxing in particular the restrictive assumption on the dependence between C and X used in Stute (1993, 1996). In Lopez (2011), an asymptotic representation of the estimation error is established when the input variable is univariate ($d = 1$). An extension with a single index model is considered in Lopez et al. (2013). The proof technique is based on the asymptotic equicontinuity of the empirical process and imposes strong conditions on the bandwidth choice, e.g. $nh^3 \to \infty$ (see Theorem 3.3 in Lopez (2011) and Theorem 3.1 in Lopez et al. (2013)). The (nonasymptotic) analysis carried out in the present paper is quite different and based on two steps: 1) the risk estimate is linearized and 2) concentration results for generalized U-processes are used to describe its fluctuations uniformly over the class of predictive functions considered (see e.g. Clémençon and Portier (2018)). Notice additionally that the approach we adopt to establish nonasymptotic rate bounds requires weaker conditions, i.e. namely $nh^{2d}/|\log(h^d)| \to \infty$ in the d-dimensional case.

Integration domain. As any (conditional) survival function, $S_C(y \mid x)$ vanishes as y tends to infinity. In order to avoid dealing with the asymptotic behaviour of the conditional survival function of the censoring and stipulating decay rate assumptions for its tail behaviour, in the analysis carried out in section 3 we restrict the study of the prediction problem to a (borelian) domain $\mathcal{K} \subset \mathbb{R}_+ \times \mathbb{R}^d$ such that $S_C(y \mid x)$ stays bounded away from 0 on it and consider the risk

$$R_{P,\mathcal{K}}(f) = \mathbb{E}\left[\frac{\delta\left(\tilde{Y} - f(X)\right)^2}{S_C(\tilde{Y}- \mid X)}\mathbb{I}\{(\tilde{Y}, X) \in \mathcal{K}\}\right], \tag{8}$$

as well as its empirical counterpart

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i(\tilde{Y}_i - f(X_i))^2}{S_C(\tilde{Y}_i- \mid X_i)}\mathbb{I}\{(\tilde{Y}_i, X_i) \in \mathcal{K}\}. \tag{9}$$

As discussed at length below, the present analysis distinguishes itself from previous works, relying on the IPCW approach as well, in several respects. First, the problem of regression (in presence of censoring) is tackled here from the angle of prediction, not as the problem of estimating the conditional expectation f^* with minimum $L_2(\mu_X)$-error, denoting X's marginal distribution by μ_X. Although an estimator of Beran's type of C's conditional survival function given X is involved in the empirical risk construction as explained above, the goal pursued here is to ensure that the predictor $\tilde{f}_n$ obtained by solving (7) has a small *excess risk* $R_P(\tilde{f}_n) - R_P(f^*)$ with large probability. As will be discussed in detail in the next section, establishing such nonasymptotic guarantees for statistical learning in the censored context, in the form of generalization bounds, yields technical difficulties, which are far from straightforward to overcome, when avoiding the simplifying assumption, hardly met in practice, that the output variable Y is independent from the r.v. C modelling the censoring mechanism (see Assumption 1, offering a much more realistic framework for regression based on censored training data).

In contrast, we derive in this article sound theoretical results, providing nonasymptotic guarantees for the risk minimizers by jointly estimating the intrinsic loss and the censoring mechanism.

2.2 Preliminary Results

In this subsection, we briefly recall the Beran approach to estimate a (conditional) survival function by means of a kernel smoothing procedure and state a uniform bound for the deviations between the conditional survival function of C given X and its Beran estimator, involved in statistical learning framework developed in the next section for distribution-free censored regression. As shall be discussed below, this result refines those obtained in Dabrowska (1989) and Du and Akritas (2002), which are of similar nature, except that they are related to the estimation of the conditional survival function of the duration Y given X, denoted by $S_Y(u \mid x) = \mathbb{P}\{Y > u \mid X = x\}$, rather than that of the conditional survival function of the censoring C given X. Define the conditional integrated hazard function of the right censoring C given X

$$\Lambda_C(u \mid x) = -\int_0^u \frac{S_C(ds \mid x)}{S_C(s- \mid x)}, \tag{10}$$

and the conditional subsurvival functions $H(u \mid x) = \mathbb{P}\{\tilde{Y} > u \mid X = x\}$ and $H_0(u \mid x) = \mathbb{P}\{\tilde{Y} > u,\ \delta = 0 \mid X = x\}$ for $u \geq 0$ and $x \in \mathbb{R}^d$. As we have (under Assumption 1), $H_0(du \mid x) = S_Y(u- \mid x)S_C(du \mid x)$ and $H(u- \mid x) = S_Y(u- \mid x)S_C(u- \mid x)$, we obtain

$$\Lambda_C(u \mid x) = -\int_0^u \frac{H_0(ds \mid x)}{H(s- \mid x)}.$$

Here, we propose to build an estimate of $\Lambda_C(u \mid x)$ by plugging into formula (10) Nadaraya-Watson type kernel estimates of the conditional subsurvival functions and derive from it an estimator of $S_C(u \mid x)$. Of course, alternative estimation techniques can be considered for this purpose. Throughout the paper, $K : \mathbb{R}^d \to \mathbb{R}^+$ is a symmetric bounded *kernel function*, i.e. a bounded nonnegative Borelian function, integrable w.r.t. Lebesgue measure such that $\int K(x)dx = 1$, $K(x) = K(-x)$ for all $x \in \mathbb{R}^d$, see Wand and Jones (1994). We assume it lies in the linear span of functions w, whose subgraphs $\{(s, u) : w(s) \geq u\}$, can be represented as a finite number of Boolean operations among sets of the form $\{(s, u) : p(s, u) \geq \zeta(u)\}$, where p is a polynomial on $\mathbb{R}^d \times \mathbb{R}$ and ζ an arbitrary real-valued function. This assumption guarantees that the collection of functions

$$\{K((x - \cdot)/h) :\ x \in \mathbb{R}^d,\ h > 0\},$$

is a bounded VC type class, see Giné et al. (2004), a property that will be useful to establish our results. Although very technical at first glance, this hypothesis is very general and is satisfied by kernels of the form $K(x) = \zeta(p(x))$, p being any polynomial and ζ any bounded real function of bounded variation (see Nolan and Pollard (1987)) or when the graph of K is a pyramid (truncated or not). For any bandwidth $h > 0$ and $x \in \mathbb{R}^d$, we set $K_h(x) = K(h^{-1}x)/h^d$. Based on the kernel estimators given by

$$\hat{H}_{0,n}(u,x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\tilde{Y}_i > u,\ \delta_i = 0\} K_h(x - X_i), \tag{11}$$

$$\hat{H}_n(u,x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\tilde{Y}_i > u\} K_h(x - X_i), \tag{12}$$

$$\hat{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \tag{13}$$

define the conditional subsurvival function estimates

$$\hat{H}_{0,n}(u \mid x) = \frac{\hat{H}_{0,n}(u,x)}{\hat{g}_n(x)} \quad \text{and} \quad \hat{H}_n(u \mid x) = \frac{\hat{H}_n(u,x)}{\hat{g}_n(x)},$$

as well as the (biased) estimators of $\Lambda_C(u \mid x)$ and $S_C(u \mid x)$

$$\hat{\Lambda}_{C,n}(u \mid x) = -\int_0^u \frac{\hat{H}_{0,n}(ds \mid x)}{\hat{H}_n(s- \mid x)}, \tag{14}$$

$$\hat{S}_{C,n}(u \mid x) = \prod_{s \le u} \left(1 - \Delta\hat{\Lambda}_{C,n}(s \mid x)\right), \tag{15}$$

with $\Delta\Lambda(t) = \Lambda(t) - \Lambda(t-)$, which are classically referred to as the conditional Nelson-Aalen and Beran estimators (Dabrowska, 1989). Let $b > 0$ and define the set

$$\Gamma_b = \left\{(y,x) \in \mathbb{R}_+ \times \mathbb{R}^d \,:\, S_Y(y|x) \wedge S_C(y|x) \wedge g(x) > b\right\},$$

which is supposed to be non-empty. On this set, one may guarantee that $\hat{H}_{0,n}(y,x)$ and $\hat{H}_{0,n}(y,x)$ are both away from 0 with high probability, which permits the study of the fluctuations of (15). In addition, the mild and standard smoothness assumption below is required in the analysis to control the estimation bias.

Assumption 2 *For all $u \in \mathbb{R}_+$, the functions $x \mapsto H(u \mid x)$, $x \mapsto H_0(u \mid x)$ and $x \mapsto g(x)$ are twice continuously differentiable on $\mathbb{R}^d$ with all partial derivatives bounded by L.*

The result stated below provides a uniform bound for the deviation between $S_C(u \mid x)$ and its estimator (15).

Proposition 1 *Suppose that Assumptions 1 and 2 are fulfilled. Then, there exist constants $M_1 > 0$, $M_2 > 0$ and $h_0 > 0$ depending on b, L and K only such that, for all $\epsilon \in (0,1)$, we have with probability greater than $1 - \epsilon$:*

$$\sup_{(t,x)\in\Gamma_b} |\hat{S}_{C,n}(t \mid x) - S_C(t \mid x)| \le M_1 \times \left\{\sqrt{\frac{|\log(h^{d/2}\epsilon)|}{nh^d}} + h^2\right\},$$

as soon as $h \le h_0$ and $nh^d \ge M_2|\log(h^{d/2}\epsilon)|$.

The technical proof is given in the Appendix section (refer to the latter for a description of the constants M_1, M_2 and h_0 involved in the result stated above). A similar result, for the conditional survival function of Y given X, is proved in Dabrowska (1989), see Theorem 2.1 therein. Observe also that choosing $h = h_n \sim n^{-1/(d+4)}$ yields a rate bound of order $O_{\mathbb{P}}(\sqrt{\log(n)/n^{4/(d+4)}})$.

Prediction *vs* Estimation. Finally, we emphasize that conditional density/expectation estimation is not the goal we pursue here, the regression framework considered in the next section having to do with *prediction*, i.e. the construction of a predictive rule $\tilde{f}_n(X)$ from censored training data with 'good' predictive capacity. The learning procedure we investigate in this article consists in minimizing a plug-in estimation of the risk (4) and consequently involves the nonparametric estimator (15), the accuracy of the prediction is measured by the excess risk, not by the estimation error $\mathbb{E}[(\tilde{f}_n(X) - f^*(X))^2]$. This contrasts with estimation techniques, which consist in forming directly an estimator of the conditional expectation $f^*(X)$ under specific (smoothness) assumptions. In addition, we point out that alternative flexible (local averaging) methods could be naturally used to compute estimators of $H_0(u, x)$, $H(u, x)$ and $g(x)$ and consequently estimators of $S_C(u \mid x)$ and $\Lambda_C(u \mid x)$, including *k-nearest neighbours*, *decision trees* or *random forest*. However, whereas the accuracy of nonparametric estimators based on kernel smoothing under appropriate smoothness hypotheses can be rather easily studied, it is much less convenient to establish rates for estimators produced by tree-based techniques for example (one generally prefers to investigate estimators built by means of variants, involving 'random splitting' for instance, quite different from the algorithms used in practice). For this reason, the predictive performance of extensions of the statistical learning approach under study, based on estimators of $S_C(t \mid x)$ built by means of tree-based techniques, are studied from an empirical angle only in this article, see section 4 for further details.

3. Generalization Bounds for Kaplan-Meier Risk Minimizers

It is the purpose of this section to investigate the excess of risk (8) related to a domain $\mathcal{K} \subset \mathbb{R}_+ \times \mathbb{R}^d$ of minimizers $\widetilde{f}_n(x)$ of the Kaplan-Meier risk (9) over a class $\mathcal{F}$ of predictive functions that is of controlled complexity (see the technical assumptions below), while being rich enough to yield a small bias $R(f^*) - R(\bar{f}^*)$, denoting $R_{P,\mathcal{K}}(\cdot)$ by $R(\cdot)$ for simplicity throughout the present section. We consider here the situation where, for all $i \in \{1, \ldots, n\}$, the estimate of the quantity $S_C(\tilde{Y}_i \mid X_i)$ plugged into (6) is obtained by evaluating the kernel smoothing estimator of $S_C(y \mid x)$ investigated in subsection 2.2 and based on the subsample $\{(X_j, \tilde{Y}_j, \delta_j) : 1 \le j \le n, j \ne i\}$ at $(y, x) = (\tilde{Y}_i, X_i)$. The corresponding versions of the kernel estimators (11), (12), (13) and those of (14) and (15) are respectively denoted by $\hat{H}_{0,n}^{(i)}(y \mid x)$, $\hat{H}_n^{(i)}(y \mid x)$, $\hat{g}_n^{(i)}(x)$, $\hat{\Lambda}_{C,n}^{(i)}(y \mid x)$ and $\hat{S}_{C,n}^{(i)}(y \mid x)$. This yields the *leave-one-out* estimator of the risk of any candidate f

$$\widetilde{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i (\tilde{Y}_i - f(X_i))^2}{\hat{S}_{C,n}^{(i)}(\tilde{Y}_i - \mid X_i)} \mathbb{I}\{(\tilde{Y}_i, X_i) \in \mathcal{K}\},$$

that is well-defined on the event $\bigcap_{i=1}^{n}\{\hat{S}^{(i)}_{C,n}(\tilde{Y}_i- \mid X_i)>0\}$. As we clearly have

$$R(\widetilde{f}_n)-\inf_{f\in\mathcal{F}}R(f)\leq 2\sup_{f\in\mathcal{F}}\left|\widetilde{R}_n(f)-R(f)\right|,$$

the key of the analysis is the control of the fluctuations of the process $\{\widetilde{R}_n(f)-R(f):\ f\in\mathcal{F}\}$. Slightly more generally, we establish below a uniform deviation bound for processes of type

$$Z_n(\varphi)=\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\varphi(\tilde{Y}_i,X_i)}{\hat{S}^{(i)}_{C,n}(\tilde{Y}_i- \mid X_i)}\right)-\mathbb{E}\left[\varphi(Y,X)\right],\ \ \varphi\in\Phi,$$

where the indexing class Φ fulfills the following property allowing us to control the fluctuations of the pseudo-variables $\hat{S}^{(i)}_{C,n}(\tilde{Y}_i- \mid X_i)$, as in Proposition 1.

Assumption 3 *There exists a domain $\mathcal{K}\subset\Gamma_b$ such that $\varphi(y,x)=0$ as soon as $(y,x)\notin\mathcal{K}$ for all $\varphi\in\Phi$.*

Equipped with these notations, observe that $\widetilde{R}_n(f)-R(f)=Z_n(\varphi)$ when $\varphi(Y,X)=(Y-f(X))^2\mathbb{I}\{(\tilde{Y},X)\in\mathcal{K}\}$.

Linearization. Whereas in the standard regression framework or in classification ERM can be straightforwardly studied by means of maximal deviation inequalities for empirical processes, the form of the process $\{Z_n(\varphi):\ \varphi\in\Phi\}$ of interest is very complex since the terms averaged in (5) are obviously far from being independent due to the presence of the plugged leave-one-out estimators of the quantities $S_C(\tilde{Y}_i- \mid X_i)$. The subsequent analysis is all the more technically difficult that, in contrast to most works devoted to statistical censored data analysis (see subsection 2.1 for more details), the simplifying assumption, unrealistic in many situations in practice, that Y and C are independent is avoided here, cf Assumption 1. Our approach to the study of the fluctuations of the process Z_n consists in linearizing the statistic $Z_n(\varphi)$, i.e. approximating $Z_n(\varphi)$ by a standard i.i.d. average in the L_2-sense, as stated in the next proposition. In order to make this decomposition explicit, further notations are needed. We set, for all $i\in\{1,\ \ldots,\ n\}$,

$$\begin{aligned}\hat{a}^{(i)}_n(t\mid x)=&-\int_0^t\frac{c(u\mid x)}{H(u,x)}\left(\hat{H}^{(i)}_{0,n}(du,x)-H_0(du,x)\right)\\&+\int_0^t\frac{c(u\mid x)}{H(u,x)^2}\left(\hat{H}^{(i)}_n(u,x)-H(u,x)\right)\hat{H}^{(i)}_{0,n}(du,x),\\\hat{b}^{(i)}_n(t\mid x)=&-\int_0^t\frac{c(u\mid x)}{H(u,x)^2\hat{H}^{(i)}_n(u,x)}\left(\hat{H}^{(i)}_n(u,x)-H(u,x)\right)^2\hat{H}^{(i)}_{0,n}(du,x)\\&-\int_0^t\frac{\left(\hat{S}^{(i)}_{C,n}(u- \mid x)-S_C(u- \mid x)\right)}{S_C(u\mid x)}\hat{\Delta}^{(i)}_n(du\mid x),\end{aligned}$$

where

$$\begin{aligned}\hat{\Delta}^{(i)}_n(du\mid x)&=\hat{\Lambda}^{(i)}_{C,n}(du\mid x)-\Lambda_C(du\mid x),\\c(u\mid x)&=\frac{S_C(u- \mid x)}{S_C(u\mid x)}.\end{aligned}$$

Equipped with these notations, we can now state the following result.

Proposition 2 (KM risk decomposition) *Suppose that Assumptions 1, 2 and 3 are fulfilled. There exist constants $h_0 > 0$ and $M_1 > 0$ that depends on b, L and K only such that*

(i) for any $n \geq 2$ and $\epsilon \in (0,1)$, provided that $h \leq h_0$ and $nh^d \geq M_1 |\log(h^{d/2}\epsilon)|$, the event

$$\mathcal{E}_n \overset{def}{=} \bigcap_{i \leq n} \left\{ \forall (t,x) \in \mathcal{K},\ \ \hat{S}_{C,n}^{(i)}(t,x) \geq b/2 \text{ and } \hat{H}_n^{(i)}(t,x) \geq b^3/2 \right\},$$

occurs with probability greater than $1-\epsilon$;

(ii) for all $\varphi \in \Phi$ and $n \geq 2$, we have on the event $\mathcal{E}_n$:

$$Z_n(\varphi) = L_n(\varphi) + M_n(\varphi) + R_n(\varphi),$$

where

$$L_n(\varphi) = \frac{1}{n}\sum_{i=1}^n \left(\delta_i \frac{\varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(\tilde{Y}, X)}{S_C(\tilde{Y} \mid X)} \right] \right),$$

$$M_n(\varphi) = -\frac{1}{n}\sum_{i=1}^n \delta_i \varphi(\tilde{Y}_i, X_i) \frac{\hat{a}_n^{(i)}(\tilde{Y}_i \mid X_i)}{S_C(\tilde{Y}_i \mid X_i)},$$

$$R_n(\varphi) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \left\{ -\hat{b}_n^{(i)}(\tilde{Y}_i \mid X_i) + \frac{\left(S_C(\tilde{Y}_i \mid X_i) - \hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i) \right)^2}{S_C(\tilde{Y}_i \mid X_i)\hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i)} \right\}.$$

The proof is given in the Appendix section. Observe that the term $L_n(\varphi)$ is a basic centred i.i.d. sample mean statistic and its uniform rate of convergence $1/\sqrt{n}$ can be recovered by applying maximal deviation bounds for empirical processes under classic complexity assumptions such as those stipulated below, whereas the term $M_n(\varphi)$ is more complicated, since it involves multiple sums. It is dealt with by means of results pertaining to the theory of U-processes (de la Peña and Giné, 1999), by showing that it can be decomposed as $M_n(\varphi) = L_n'(\varphi) + R_n'(\varphi)$, the sum of a linear term and a second-order term. The term $R_n(\varphi) + R_n'(\varphi)$ is a remainder term (second order) and shall be proved to be negligible with respect to $L_n(\varphi) + L_n'(\varphi)$.

The theory of U-processes is used next to describe the uniform behaviour of $M_n + R_n$. Such concentration results are also used in Clémençon et al. (2008) and Papa et al. (2016) in simpler situations, where the residuals take the form of a degenerate U-statistic. In our case, due to the presence of a leave-one-out estimate of the survival function, the U-processes that arise do not have all their diagonal terms (e.g., the sum indexes $1 \leq i, j \leq n$ are restrained to $i \neq j$). This is of particular interest because results dealing with U-processes are in most cases stated for such sums (see Lemma 7 and Corollary 8 in the Appendix section) and, more importantly, removing diagonal terms improves the estimation accuracy by reducing the bias (see also Delyon and Portier (2016), remark 4).

To obtain uniform concentration inequalities over the function class Φ, it is standard (Nolan and Pollard, 1987; Giné and Guillou, 2001) to assume the following type of control on the complexity of the class.

Assumption 4 *The set Φ of real-valued functions on $\mathbb{R}_+ \times \mathbb{R}^d$ is a bounded* VC *type of class with parameter (A, v) and constant envelope M_Φ.*

The formal definition of VC classes is given in the Appendix section. By means of these assumptions, the following result, proved in the Appendix section, describes the order of magnitude of the fluctuations of the process Z_n.

Proposition 3 *Suppose that Assumptions 1-4 are fulfilled. There exist constants h_0, M_1, M_2 and M_3 that depend on (A, v), M_Φ, L, K and b only, such that, for all $n \geq 2$ and $\epsilon \in (0,1)$, the event*

$$\sup_{\varphi \in \Phi} |Z_n(\varphi)| \leq M_1 \left(\sqrt{\frac{\log(M_2/\epsilon)}{n}} + \frac{|\log(\epsilon h^{d/2})|}{nh^d} + h^2 \right),$$

occurs with probability greater than $1 - \epsilon$ provided that $h \leq h_0$, $nh^{2d} \geq M_3 |\log(\epsilon h^d)|$.

The risk excess probability bound stated in the following theorem shows that, remarkably, minimizers of the Kaplan-Meier risk attain the same learning rate as that achieved by classic empirical risk minimizers in absence of censoring, when ignoring the model bias effect induced by the plug-in estimation step (*cf* choice of the bandwidth h).

Theorem 4 *Suppose that Assumptions 1-4 are fulfilled. There exist constants h_0, M_1, M_2 and M_3 that depend on (A, v), M_Φ, L, K and b only, such that, for all $n \geq 2$ and $\epsilon \in (0,1)$, the event*

$$|R(\tilde{f}_n) - R(f^\star)| \leq M_1 \left(\sqrt{\frac{\log(M_2/\epsilon)}{n}} + \frac{|\log(\epsilon h^{d/2})|}{nh^d} + h^2 \right),$$

occurs with probability greater than $1 - \epsilon$ provided that $h \leq h_0$, $nh^{2d} \geq M_3 |\log(\epsilon h^d)|$.

The proof is a direct application of Proposition 3. A similar bound for the expectation of the risk excess of minimizers of the empirical Kaplan-Meier risk can be classically derived with quite similar arguments, details are left to the reader. We finally point out that, given that Proposition 3 holds true for a fairly general class of functions Φ, the guarantees provided by Theorem 4 can be naturally extended to more general risk measures than that defined by the quadratic loss.

4. Numerical Experiments

Beyond the theoretical generalization guarantees established in the previous section, we now examine at length the performance of the predictive approach proposed in the context of regression based on censored data from an empirical perspective. We present various experiments using both synthetic and real data, and compare it to alternative methods documented in the survival analysis literature standing as natural competitors. As shall be seen below, the experimental results obtained provide strong empirical evidence of the relevance of the Kaplan-Meier empirical risk minimization approach described in section 2 and analysed theoretically in section 3. All the experiments and figures displayed in this article can be reproduced using the code available at `https://github.com/aussetg/ipcw`.

4.1 Experimental Setup

Before presenting and discussing the numerical results obtained, we first describe the experimental schemes used here to investigate the predictive capacity of the learning procedure under random censoring previously studied.

4.1.1 Data Generative Models

In all the synthetic experiments we have carried out, the generation of the data is based either on the proportional hazard model of Cox and Oakes (1984) or else on the accelerated time failure model of Buckley and James (1979); both commonly used for parametric modelling and statistical estimation of conditional survival functions in the censored setup. Samples of the triplet $(\tilde{Y}, \delta, X)$ are obtained by specifying the marginal distribution of X, as well as the conditional distribution of (Y, C) given X. For simplicity, the input r.v. X is here uniformly distributed on the unit square $[0,1]^d$, for $d \in \{2,4,8\}$. Only the results for $d = 4$ are presented below, while those obtained for $d \in \{2,8\}$ are available through the link mentioned above.

Cox Model. The first survival model we use to simulate synthetic data stipulates that

$$S_Y(y \mid x) = \exp\left(-\exp\left(\beta^T x\right) y\right) \text{ and } S_C(y \mid x) = \exp\left(-\exp\left(\beta_C^T x\right) y\right), \tag{16}$$

where β and β_C are parameters in $\mathbb{R}^d$. Given X, the conditional distribution of Y is thus exponential with parameter $\exp(\beta^T X)$, while that of C is exponential with parameter $\exp(\beta_C^T X)$.

Accelerated Failure Time Model (AFT). The second generative model we considered assumes that

$$\log(Y) = -\beta^T X + \varepsilon_0 \text{ and } \log(C) = -\beta_C^T X + \varepsilon_1, \tag{17}$$

where the r.v. ε_0 (respectively ε_1) is independent from X. Different accelerated failure time models can thus be generated, depending on the distributions D_0 and D_1 chosen for ε_0 and ε_1. Three distributions have been used: Normal (N) with mean and variance $(3/2, 1)$, Laplace (L) with location and scale $(1,1)$ and Gamma (G) with shape and scale $(0,1)$. Denoting by $\text{AFT}(D_0, D_1)$ the model such that $(\varepsilon_0, \varepsilon_1) \sim D_0 \otimes D_1$, the variants $\text{AFT}(N,N)$, $\text{AFT}(N,L)$ and $\text{AFT}(N,G)$ have been simulated. Since the results obtained for these AFT models are quite similar to those based on the Cox model, only the latter are presented below. We refer to the link aforementioned for a description of the results based on the data generated through the AFT models.

Parameters β and β_C. In the Cox and AFT models, the level of censoring can be tuned by carefully choosing the parameters β and β_C. In order to guarantee that the censoring is informative, we use the following parametrization:

$$\begin{aligned} \beta^T &= \big[\overbrace{1 \quad \cdots \quad 1}^{\lceil d/2 \rceil} \quad 0 \quad \cdots \quad 0\big], \\ \beta_C^T &= \lambda \left[1 \quad 0 \quad 1 \quad 0 \quad 1 \quad \cdots\right], \end{aligned}$$

where the tuning parameter $\lambda > 0$ controls the level of censoring $1 - p$ with $p = \mathbb{P}\{Y \leq C\}$ and $u \in \mathbb{R} \mapsto \lceil u \rceil$ is the ceiling function. For a targeted censoring level p, the parameter λ can be empirically determined so that $\sum_{i=1}^n \delta_i \simeq np$.

4.1.2 Plugged estimator of the conditional survival function $S_C(\cdot \mid x)$

The estimate of the risk (9) one seeks to minimize is partly determined by the choice of the estimator $\hat{S}_C(\cdot \mid x)$ of $S_C(\cdot \mid x)$ plugged into it. We consider for $\hat{S}_C$ the kernelized Kaplan-Meier estimator (15), in its standard version denoted by $\hat{S}_C^{\texttt{Kern}}(\cdot \mid x)$ and in its leave-one-out version as well, denoted by $\hat{S}_C^{(i)\,\texttt{Kern}}(\cdot \mid x)$. We denote by $\hat{S}_C^{\texttt{KM}}(\cdot)$ the Kaplan-Meier estimator of C's (unconditional) survival function, which can be seen as the limit of $\hat{S}_C^{\texttt{Kern}}$ when $h \to \infty$ and yields the Kaplan-Meier risk considered in Stute (1995). In addition, we used $\hat{S}_C^{(i)\,\texttt{KNN}}(\cdot \mid x)$, the estimator obtained by replacing the kernel smoothing involved in (9) by a nearest neighbour averaging, in a leave-one-out fashion. Finally, we also considered $\hat{S}_C^{\texttt{RF}}$, the survival random forest estimator proposed in Ishwaran et al. (2008). From each estimator of S_C, one computes a plug-in estimation of the risk:

$$\begin{array}{rlrl}
\text{IPCW} & \displaystyle\sum_{i=1}^{n} \delta_i \frac{(\tilde{Y}_i - f(X_i))^2}{\hat{S}_C^{\texttt{Kern}}(\tilde{Y}_i|X_i)} & \text{IPCW LoO} & \displaystyle\sum_{i=1}^{n} \delta_i \frac{(\tilde{Y}_i - f(X_i))^2}{\hat{S}_C^{(i)\,\texttt{Kern}}(\tilde{Y}_i|X_i)} \\
\text{IPCW Forest} & \displaystyle\sum_{i=1}^{n} \delta_i \frac{(\tilde{Y}_i - f(X_i))^2}{\hat{S}_C^{\texttt{RF}}(\tilde{Y}_i|X_i)} & \text{IPCW Stute} & \displaystyle\sum_{i=1}^{n} \delta_i \frac{(\tilde{Y}_i - f(X_i))^2}{\hat{S}_C^{\texttt{KM}}(\tilde{Y}_i)} \\
\text{IPCW KNN} & \displaystyle\sum_{i=1}^{n} \delta_i \frac{(\tilde{Y}_i - f(X_i))^2}{\hat{S}_C^{(i)\,\texttt{KNN}}(\tilde{Y}_i|X_i)} & \text{IPCW Oracle} & \displaystyle\sum_{i=1}^{n} \delta_i \frac{(\tilde{Y}_i - f(X_i))^2}{S_C(\tilde{Y}_i|X_i)} \\
\text{Naive} & \displaystyle\sum_{i=1}^{n} (\tilde{Y}_i - f(X_i))^2 & \text{Observed} & \displaystyle\sum_{i=1}^{n} \delta_i (\tilde{Y}_i - f(X_i))^2.
\end{array} \tag{18}$$

The naive and observed empirical risks introduced above correspond to strongly biased estimators of the true risk (1) of course. Note that the normalizing constant is ignored here, insofar as it has no impact on the empirical risk minimizer. However, when estimating the true risk itself, it is necessary to correctly normalize the previous quantities in order to obtain complete case estimators. A point of comparison is the oracle risk

$$\text{Oracle} \quad \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2,$$

i.e. the empirical risk in absence of any censoring (i.e. when all the Y_i's are observed). For each risk, a prediction rule $\tilde{f}_n^{\star}$ is built by (approximately) minimizing it over a certain class $\mathcal{F}$. The results are depicted in Fig.1 for various sizes of the (censored) training sample and different censoring levels, the prediction error being evaluated by means of a test (uncensored) sample of size 5000: learning a predictive function by minimizing an IPCW estimator (here IPCW LoO) of the risk always outperforms naive alternatives, the gain in predictive performance naturally becoming more pronounced as the level of censoring $1 - p$ increases. Unsurprisingly, when most of the points are observed (i.e. $p \to 1$), all methods reach roughly the same error, all the losses in (18) being equal for $p = 1$, as depicted by Fig.2. Notice also in Fig.2 that the IPCW estimator performs best compared to the naive methods for a moderate level of censoring. This can be explained by the fact that in absence of censoring the methods are equivalent and when censoring reaches a very high level, there is not enough data to estimate reliably the IPCW weights. Of course, the phenomenon is exaggerated in the present example, as the training set is of size 1000: with a censoring level of 90%, only 100 observations are then available for the conditional estimation of the weights.

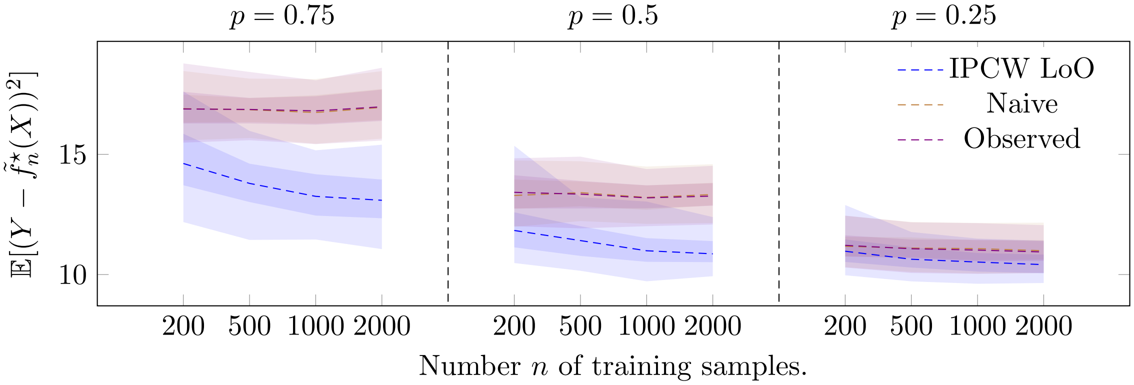

Figure 1: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for data generated by the Cox model (16), when minimization is performed over the class of affine predictive rules.

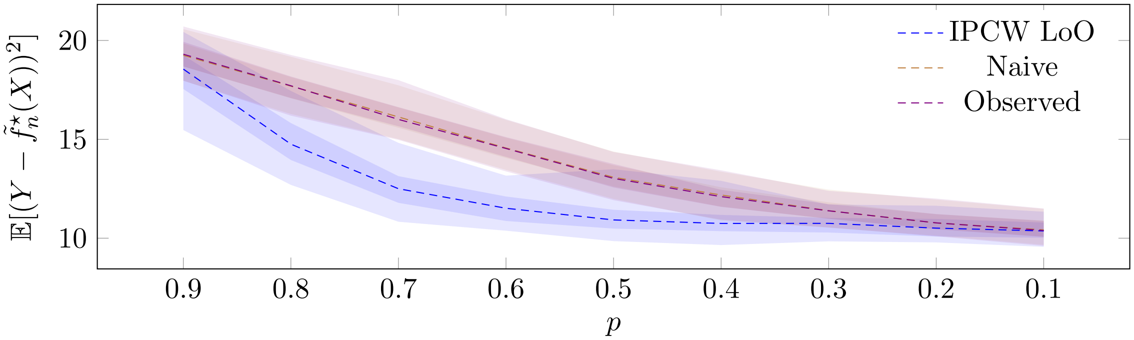

Figure 2: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for data generated by the Cox model (16) with $n = 1000$, when minimization is performed over the class of affine predictive rules.

Truncation of the estimator $\hat{S}_C(\cdot \mid x)$. In the theoretical analysis carried out in the previous section, we placed ourselves on a restricted set Γ_b. However, in practice, we employ a truncation approach by simply removing the last jump of the estimated survival function. For instance, $\hat{S}_C^{\texttt{Kern}}(y|x)$ is taken as

$$\prod_{\substack{\tilde{Y}_i \leq y \\ \tilde{Y}_i < \max_{j:\, \delta_j = 0} Y_j}} \Big(1 - \Delta\hat{\Lambda}_{C,n}(\tilde{Y}_i \mid x)\Big)\,. \tag{19}$$

Observe that, although it is not a survival function anymore, it is still a relevant estimator. This alleviates possible difficulties caused by the frequent edge case where the last individual is observed ($\delta = 1$), since, in the case where (15) is used, we have then $\delta_n / \hat{S}_C(\tilde{Y}_n \mid X_n) = \infty$. Of course, it would have been possible to decide to force a restriction on Γ_b by considering the flooring $\max(b, \hat{S}_{C,n}(y \mid x))$ rather than $\hat{S}_{C,n}(y \mid x)$. However, b then becomes an hyperparameter of the procedure that has to be tuned by the practitioner. In the survival analysis literature, it is common to consider the restricted mean survival time, the conditional distribution of the restricted variable $\min(Y, \tau)$ being easier to learn than that of Y; see e.g. Royston and Parmar (2011) or Steingrimsson et al. (2016, 2019); Steingrimsson and Morrison (2020). In this case the conditional survival function of the restricted variable is equal to the truncated conditional survival function of Y. Our approach therefore consists in using the restricted mean censoring time as the target for the weights in order to reduce the noise. We evaluated truncation of the survival function and flooring by comparing the predictive performance attained by the rule learnt from data generated by means of the Cox model, when choosing successively $\hat{S}_C^{\texttt{Kern}}, \hat{S}_C^{(i)\,\texttt{Kern}}$ and $\hat{S}_C^{(i)\,\texttt{KNN}}$ for $\hat{S}_C$. As depicted by Fig. 3 for the specific case where $\hat{S}_C = \hat{S}_C^{(i)\,\texttt{Kern}}$ (and this remains true in the other cases), a wrong choice for b may have serious consequences, while the truncation approach of Eq. (19) consistently produces good results. Consequently, the truncated version is always used in the following experiments.

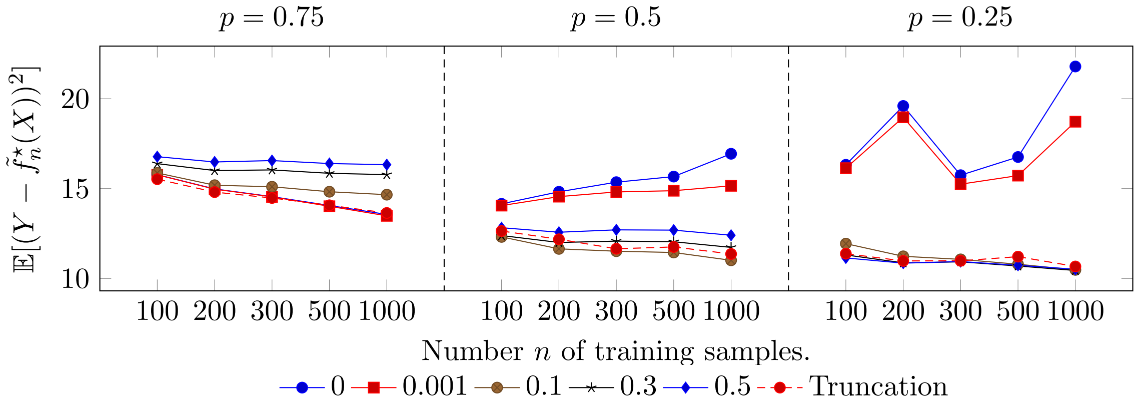

Figure 3: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ when $\mathcal{F}$ is the class of affine functions, choosing the IPCW LoO risk estimator and the Cox model (16) for generating the data. The curves correspond to different floors b and to the truncation approach of (19).

Calibration of $\hat{S}_C(\cdot \mid x)$. In order to fully specify the estimator $\hat{S}_C(\cdot \mid x)$, it may be necessary to choose specific hyperparameters. Without censoring, and therefore without having to resort to the IPCW approach, one would select the various hyperparameters by way of a cross-validation; this approach is, however, impossible in our case as the loss itself is unknown and only estimated, worse any modification of the parameters of $\hat{S}_C$ results in a modification of the estimator of the loss we wish to minimize. One possible solution is to rely on a *surrogate loss* i.e. an auxiliary loss that we are able to compute exactly and on which a cross-validation is therefore possible. For $\hat{S}_C^{\texttt{Kern}}$ and $\hat{S}_C^{(i)\,\texttt{Kern}}$, we consider $\hat{m}_h(x)$ the nonparametric kernel regression of $\tilde{Y}$ w.r.t. X, known as the Nadaraya-Watson estimator, and the surrogate loss $\mathbb{E}[|\tilde{Y} - \hat{m}_h(X)|^2]$ which is then minimized using cross-validation with respect to h. In this way, a value for the bandwidth parameter h_{cv}^* is obtained and might be used in $\hat{S}_C^{\texttt{Kern}}$ and $\hat{S}_C^{(i)\,\texttt{Kern}}$. This approach is also easily applied to set the number of neighbours involved in $\hat{S}_C^{(i)\,\texttt{KNN}}$. As a close relative of the task of estimating S_C, the previous regression loss is a good candidate for the surrogate cross-validation.

In the specific case of $\hat{S}_C^{(i)\,\texttt{Kern}}$ and $\hat{S}_C^{(i)\,\texttt{KNN}}$, we experimentally studied the impact of the choice of h and k on the prediction performance $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$. The results for the specific case of $\hat{S}_C^{(i)\,\texttt{Kern}}$ are given in Fig. 4 and demonstrate the need to be over-conservative rather than under-conservative in the choice of these hyperparameters. In our experiments, choosing h at least equal to the value h^*_{cv} obtained by minimizing the surrogate, and up to 5 times this value, is a safe choice. Consequently, we use $h = 5h^*_{cv}$ in the following experiments. For $\hat{S}_C^{\texttt{RF}}$, given the large number of hyperparameters, the default parameters selected by the package's authors have been used.

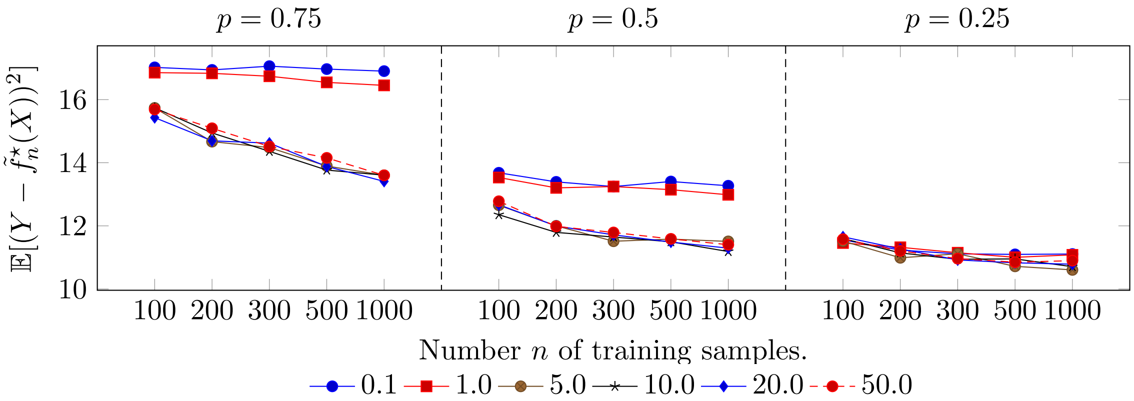

Figure 4: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for varying bandwidth h, with $\mathcal{F}$ a random forest model, IPCW LoO estimator of S_C, and data following the Cox model (16).

4.2 Experimental Results based on Synthetic Data

We now present the results obtained from the data generated by means of the model previously described.

Risk estimation. While not the focus of the predictive approach studied in this paper, it is of interest to evaluate the quality of the estimation of $\mathbb{E}[\varphi(Y, X)]$, related to a certain function φ, attained by the IPCW method. In order to make computations easier, we choose to study functionals of the form $\varphi(Y, X) = Y \exp\left(-X^T\beta\right)$, where Y follows the Cox model described in Equation (16). In this case $\mathbb{E}[\varphi(Y, X)] = 1$. For a single random dataset $\mathcal{D}_n = \{(X_i, Y_i, \delta_i) : i = 1, \ldots, n\}$ of size n, the estimation error is given by

$$\mathrm{err}_n = \left| 1 - \frac{1}{n} \sum_{(X_i, Y_i, \delta_i) \in \mathcal{D}_n} \frac{\delta_i}{\hat{S}_C(\tilde{Y}_i \mid X_i)} \tilde{Y}_i \exp\left(-X_i^\intercal \beta\right) \right|.$$

Based on $M = 100$ simulated datasets, we study the distribution of err_n for varying sizes n. We represent the median, 5% and 95% quantiles of err_n in Fig. 5 for each survival estimator. As can be seen in Fig. 5, while the naive uncorrected method results in a poor approximation of the considered expectation (as expected, since it is strongly biased), the

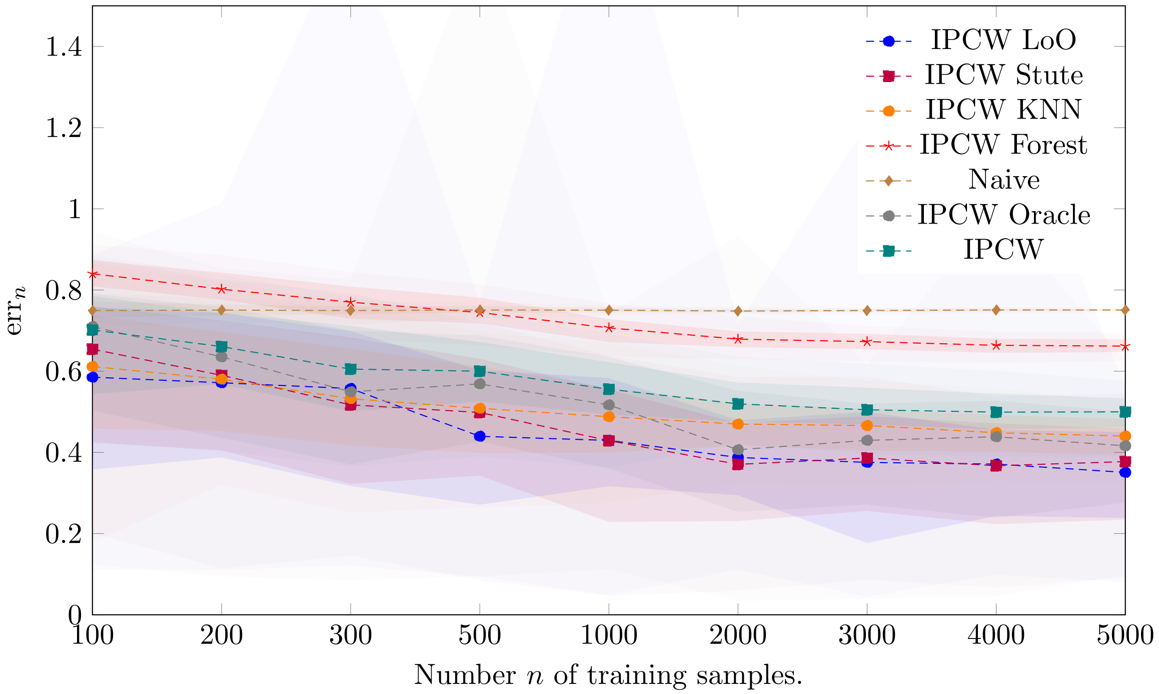

Figure 5: Estimation error for the IPCW Risks of (18) compared to the naive method, for $p = 1/4$ and data following the Cox model (16).

IPCW reweighting errors converge towards 0. One should pay specific attention to the particularly good performance of the leave-one-out version of the IPCW estimators. We also point out that low-bias estimators of S_C, such as the Random Forest estimator, can underperform significantly compared to their high bias counterparts such as the unconditional Stute estimator. This behaviour is consistent with the observations made in the previous discussion about calibration. It is illustrated by Fig. 4. We empirically observe that the IPCW estimator of the risk with oracle weights (i.e. computed from the true conditional survival function $S_C(. \mid x)$) may be less accurate than plug-in versions (i.e. computed from an estimator of the conditional $\hat{S}_C(. \mid x)$) and exhibits a much higher variance. Intuitively, this phenomenon can be explained by the fact that the empirical weights governed by the value $1/S_C(Y_i|X_i)$ can grow arbitrarily large for observations in the tail. This phenomenon is reduced for the estimated version LoO (resp. KNN) because of the truncation (see the implementation details above) and the over-conservative choice of the bandwidth (resp. of the number of neighbours). A similar phenomenon occurs for estimated of importance sampling type, for which the weights appearing in the denominator need to be tuned finely, see Delyon and Portier (2020).

Predictive performance. In this paper, we are concerned with the *predictive* task, rather than risk estimation. Hence, we now focus on the problem of minimizing (1). We study the prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for the following types of predictive model $\mathcal{F}$: Support vector Regression (SVR), Random Forests and Linear Regression. Although the choices we made are far from being exhaustive, they correspond to tools commonly used by practitioners.

Following the experimental scheme presented in subsection 4.1 we first generate train sets of varying size n and test sets of fixed size 5000 according to the data generative models described in 4.1.1. We then estimate on the train set the weights corresponding to each risk described in Eq. (18) with hyperparameters chosen using the procedure given in 4.1.2 before computing $\tilde{f}_n^\star$, the minimizer of the resulting empirical risk over the class $\mathcal{F}$ considered. We finally estimate the prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ using the test dataset. Each experiment is entirely replicated (including the sampling of the train and test sets) 100 times in order to obtain reliable statistics for the distribution of the true risk of the learning procedure.

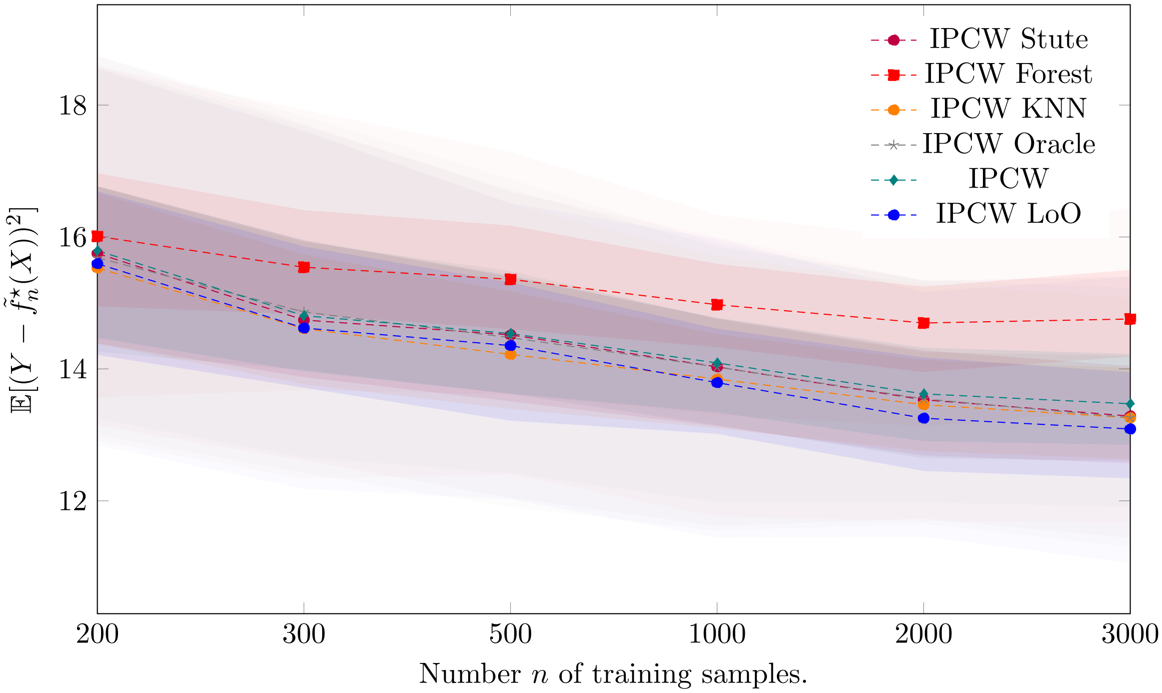

Figure 6: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for different estimators of S_C using the linear regression model for data generated by the Cox model (16).

As already observed earlier in Fig. 5, Fig. 6 and Fig. 7 show that the IPCW LoO predictor systematically outperforms the other predictors in our experiments, no matter the level of censoring p and across different distributions as can be seen here for the specific case of the Cox model and the AFT(N, L). Consequently, any further mention of IPCW implicitly refers to the IPCW LoO version from now on and all subsequent experiments involve the use of $\hat{S}_C^{(i)\,\texttt{Kern}}$. We also underline that these results hold true, no matter the predictive model $\mathcal{F}$ considered, as can be seen by examining Fig. 8 and no matter the underlying distribution, see Fig. 9. It is noteworthy that the methods are all the most different as the number of observations in the training set increases. As the conditional estimators become more and

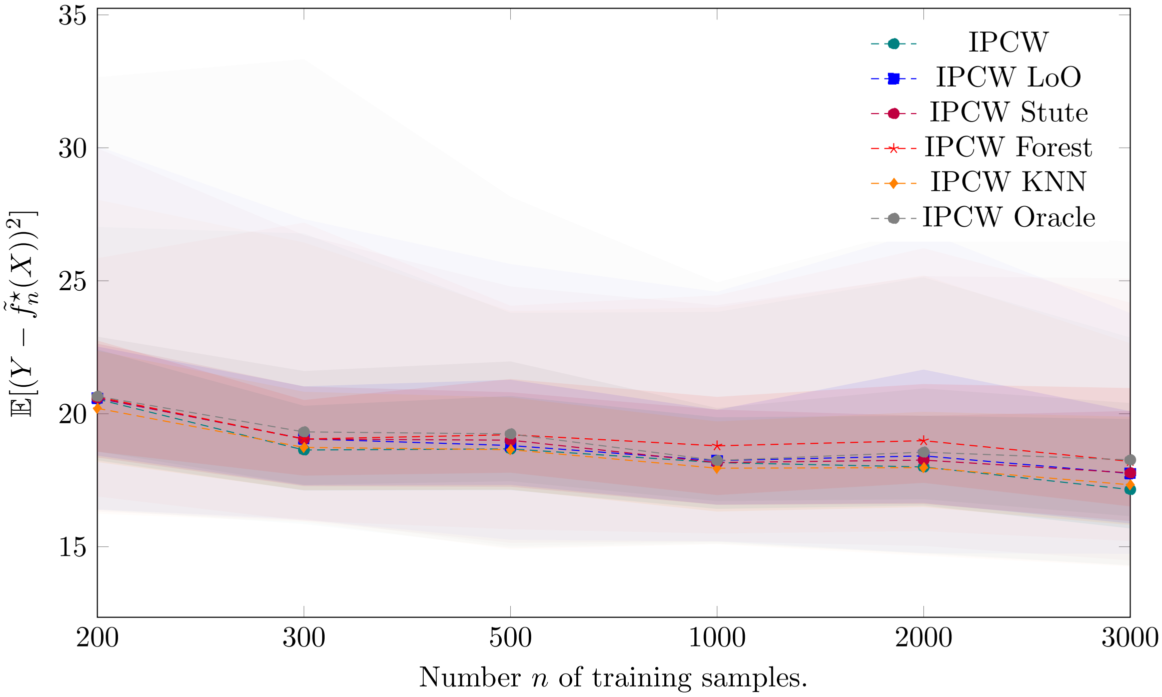

Figure 7: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for different estimators of S_C using the linear regression model for data generated by the AFT(N, L) model (17).

more complex, more and more data are required to differentiate them from the unconditional versions.

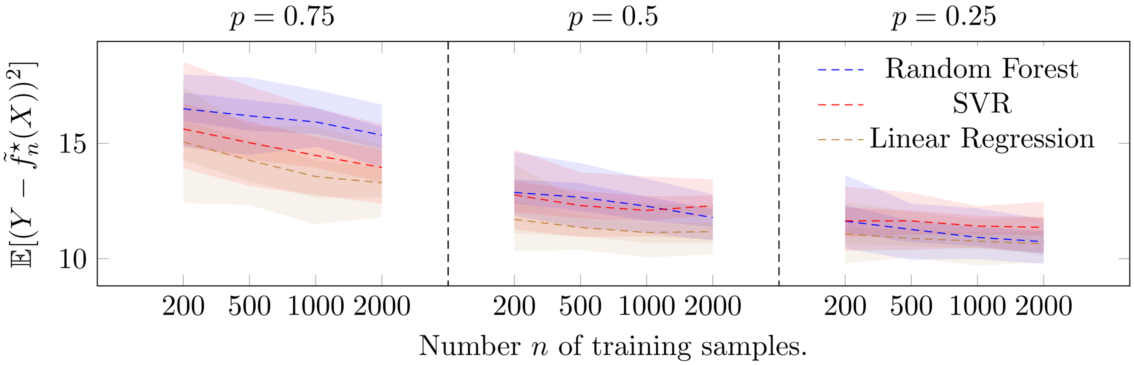

Figure 8: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for the three predictive models (SVR, random forest and linear regression) using the IPCW LoO for data generated by the Cox model (16).

Finally, we compare variants of popular machine learning methods implementing the IPCW technique promoted in this paper to standard state-of-the-art procedures from the survival analysis literature that do not rely on (re-weighted) risk minimization. Such

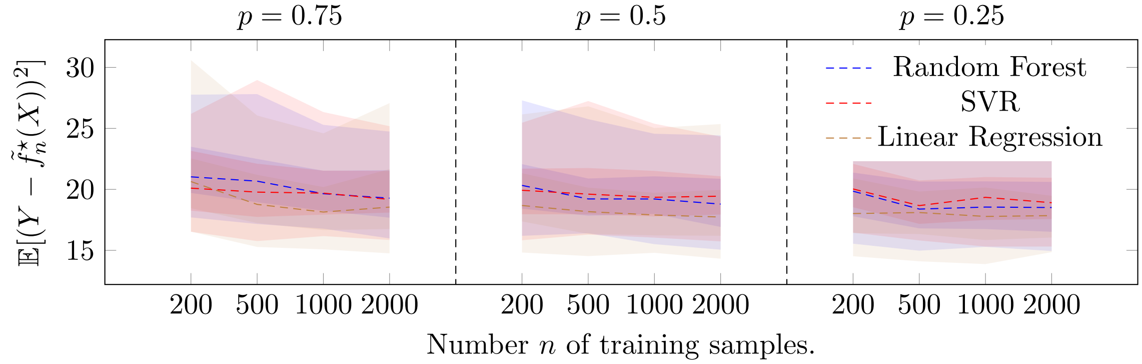

Figure 9: Prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$ for the three predictive models (SVR, random forest and linear regression) using the IPCW LoO for data generated by the $\text{AFT}(N, L)$ model (17).

techniques include classic statistical methods based on the preliminary estimation of the survival function, as already mentioned in Section 1 (see e.g. van der Laan and Robins (2003)), the survival function estimator being next used to estimate the downstream quantity of interest in a plug-in fashion, provided that the latter can be expressed as an integral w.r.t. the survival function, just like the conditional mean. An alternative approach, in the spirit of machine learning methods, consists in designing losses tailored to the censored regression problem, either through transformation models in Van Belle et al. (2011), or else by adapting the SVM methodology, as done in e.g. Van Belle et al. (2007); Pölsterl et al. (2015, 2016). We also include the method of Hothorn et al. (2006) which shares similarities with the approach investigated in this paper and uses a boosting technique to optimize a loss reweighted by unconditional Kaplan-Meier weights, as well as the technique proposed in Ishwaran et al. (2008) that builds a recursive splitting of the feature space $\mathcal{X}$ by maximizing a measure of inter-cluster dissimilarity of the survival functions, the resulting clusters being then used for downstream tasks such as classification, regression, or quantile estimation. We compare the predictive performance of ten estimators of the regression function based on statistical models documented in the survival literature with that of five predictive functions learned using the IPCW risk minimization approach. The IPCW versions of the machine learning techniques for regression considered in these experiments, corresponding to the approach studied in the present article, have been implemented with `Scikit-Learn` (Pedregosa et al., 2011), combined with our own implementation of the LoO IPCW predictor we propose. For the survival machine learning methods mentioned above, we use the reference implementations of the `Scikit Survival` package (Pölsterl, 2020). The canonical implementation of Ishwaran and Kogalur (2007) is used for Random Survival Forest. The default values for the hyperparameters are used in every case. All experiments are based on $n = 200$ training observations. Results for all methods can be found in Table 1. While the undeniable superiority of IPCW methods compared to the standard survival techniques may appear surprising at first glance. However, keeping in mind that the performance measure is here the prediction error $\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$, it is expected that directly minimizing an estimator of the prediction error yields better results than a two-stage procedure, which consists in estimating first the underlying distribution and forming next an estimator of its mean.

		$\mathbb{E}[(Y - \tilde{f}_n^\star(X))^2]$		
	Method	$p = 0.75$	$p = 0.5$	$p = 0.25$
`Scikit Survival`	Survival Gradient Boosting	3.19	3.55	3.61
	Component-wise Survival Gradient Boosting	3.19	3.87	4.23
	Cox Proportional Hazards	7.86	7.61	7.03
	Coxnet	7.62	7.39	6.85
	Kernel Survival SVM	4.02	3.92	4.13
	Survival SVM	4.04	4.09	3.94
	Hinge Loss Survival SVM	8.10	8.28	8.09
	Minlip Survival SVM	3.27	3.96	4.22
	Random Survival Forest	2.01	2.94	2.78
`Scikit Learn`	Ridge + IPCW	**1.75**	**1.49**	**1.24**
	Kernel Ridge + IPCW	2.07	1.60	1.35
	Linear Regression + IPCW	1.81	**1.49**	**1.24**
	Random Forest + IPCW	1.85	1.57	1.36
	SVR + IPCW	1.87	1.66	1.42

Table 1: Performance on the Cox dataset

4.3 Experimental Results based on Real Data

The performance of the IPCW risk minimization approach is now investigated on the TCGA Cancer data (Grossman et al., 2016) using solely the RNA transcriptomes as informative variables. All models are trained on $n = 8080$ patients with a censoring rate of 18%, we measure on the remaining 1449 observed patients the prediction error, as well as the Harrel concordance index defined by

$$\frac{\sum_{Y_j \leq Y_i} \delta_j \mathbb{I}\{f(X_j) > f(X_i)\}}{\sum_{Y_j \leq Y_i} \delta_j}, \tag{20}$$

which can be seen as an extension of the classical AUC metric for the standard classification problem to censored data, measuring how *well ordered* the predicted death times are. Note that, as a complete case statistic, the same methodology as presented in this paper could be applied to estimate the AUC type statistic $\mathbb{P}\{f(X_1) > f(X_2) \mid Y_1 > Y_2\}$, using IPCW weights. However, because it has been used for evaluating alternative methods, we use the criterion (20) here in order to facilitate comparisons and avoid to discuss the impact of specific IPCW approaches to compute the weights. For all the results, we use the IPCW methodology presented earlier. The Cox proportional hazards model was, however, learned after variable selection via a Lasso regression so as to augment performance.

We observe from the results in Table 2 that, as expected, the predictors built through IPCW risk minimization significantly outperform their competitors, including the standard

Cox model, for the prediction task. The large improvement compared to the Cox approach may surprise at first but is not unexpected actually, insofar as the seemingly less sophisticated IPCW approach is specifically designed for the purpose of *prediction*. By directly minimizing an estimate of the loss of interest, it naturally achieves a lower test prediction error than that reached by the traditional two-stage approach in statistics, which consists in estimating first the distribution and deducing next an estimator of the minimizer of the loss of interest (here, such an approach would consist in building first an estimator of Y's conditional density given X based on the censored sample and taking next its expectation to form an estimator of the theoretical minimizer f^*). The Cox estimator only controls the likelihood of the model without any concern for the predictive performance. In particular, extreme errors are not penalized in any way while those are hurtful to the overall L^2 error. More interestingly, we see that, while the difference is not as pronounced, the IPCW predictors also outperform the Cox estimator with respect to the concordance index. Notice incidentally that the concordance index, as an extension of the Wilcoxon-Mann-Whitney or AUC statistic, is an empirical criterion that can hardly be optimized directly in practice (using gradient ascent techniques for instance) because of the nonsmooth character of the pairwise loss function it involves, see e.g. section 7 in Clémençon et al. (2008), but it is often used for evaluating performance *a posteriori* by practitioners. Remarkably, as the IPCW risk minimization approach can be combined with highly sophisticated learners (such as random forests), without any modification or increase in complexity, it is possible to significantly increase its predictive capacity, while edging the standard survival techniques on auxiliary metrics as well.

	IPCW		Naive		Observed	
Method	L^2 Error (*years*)	Concordance	L^2 Error	C	L^2 Error	C
Cox	18.78	0.6095	–	–	–	–
SVR	2.768	0.563	2.796	0.575	2.795	0.543
Linear Regression	3.193	0.594	4.971	0.557	3.898	0.508
Ridge	3.193	0.594	4.962	0.5573	3.896	0.5077
Kernel Ridge	2.683	0.597	2.704	0.592	2.956	0.513
Random Forest	**2.577**	**0.630**	2.636	0.603	2.878	0.542

Table 2: Results of the IPCW approach on the TCGA Cancer data.

5. Conclusion

In the present article, we have presented both theoretical and experimental work on statistical learning based on censored data. Precisely, we considered the problem of learning a predictive/regression function when the output variables related to the training observations are subject to random right censoring under mild assumptions. Following in the footsteps of the approach introduced in Stute (1995), we studied from a nonasymptotic perspective the performance of predictive functions built by minimizing a weighted version of the empirical (quadratic) risk, constructed by means of the Kaplan-Meier methodology. Learning rate bounds describing the generalization ability of such predictive rules have been proved, through the study of the fluctuations of the Kaplan-Meier risk functional, relying on linearization techniques combined with concentration results for U-processes. These theoretical results have also been confirmed by various numerical experiments, supporting the approach promoted. A difficult question, that will be the subject of further research, is the design of model selection methods (structural risk minimization) to pick automatically the optimal hyperparameters for the plugged estimator $\hat{S}_{C,n}$. Indeed, this is far from straightforward, insofar as changing the hyperparameters or the model modifies the loss that is being optimized, which makes standard methods such as cross-validation unsuitable.

Appendix A. Concentration inequalities for VC classes and permanence properties

For completeness, concentration results as well as preservation properties of VC classes, extensively used in the subsequent proofs, are recalled. For the sake of generality, this section is independent from the rest of the paper. For a function $f : S \to \mathbb{R}$, we define $\|f\|_\infty = \sup_{x\in S} |f(x)|$ and $\|f\|_A = \sup_{x\in A} |f(x)|$.

Concentration inequalities over VC classes. The following concentration inequalities provide uniform bound on empirical sums over VC classes of functions. We start by recalling the definition of a VC class.

Definition 5 *Let $(S, \mathcal{S})$ be a measurable space. A class $\mathcal{F}$ of real-valued functions defined on S is called* VC *of parameter $(A, v) \in (0, +\infty)^2$ and constant envelope $U_{\mathcal{F}} > 0$ if for any probability measure Q on $(S, \mathcal{S})$ and any $\epsilon \in (0, 1)$:*

$$\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon U_{\mathcal{F}}) \leq (A/\epsilon)^v,$$

where $\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon)$ denotes the smallest number of $L_2(Q)$-balls of radius less than ϵ required to cover class Φ (covering number), see e.g. Nolan and Pollard (1987) and Giné and Guillou (2001).

The following inequality for empirical processes over VC classes is stated in Einmahl and Mason (2000); Giné and Guillou (2001) under various forms. The present version is taken from Giné and Sang (2010).

Lemma 6 *Let $\xi_1, \xi_2, \ldots$ be i.i.d. r.v.'s valued in a measurable space $(S, \mathcal{S})$ and $\mathcal{U}$ be a class of functions on S, uniformly bounded and of VC-type with constant (A, v) and envelope $U : S \to \mathbb{R}$. Set $\sigma^2(u) = var(u(\xi_1))$ for all $u \in \mathcal{U}$. There exist constants $C_1 > 0$, $C_2 \geq 1$, $C_3 > 0$ (depending on v and A) such that $\forall t > 0$ satisfying*

$$C_1 \sigma \sqrt{n \log\left(\frac{2\|U\|_\infty}{\sigma}\right)} \leq t \leq \frac{n\sigma^2}{\|U\|_\infty},$$

then

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n \{u(\xi_i) - \mathbb{E}[u(\xi_i)]\}\right\|_{\mathcal{U}} > t\right\} \leq C_2 \exp\left(-C_3 \frac{t^2}{n\sigma^2}\right).$$

The previous result is extended to the case of degenerated U-processes over VC classes (Major, 2006, Theorem 2).

Lemma 7 *Let $\xi_1, \xi_2, \ldots$ be an i.i.d. sequence of random variables taking their values in a measurable space $(S, \mathcal{S})$ and distributed according to a probability measure P. Let $\mathcal{H}$ be a class of functions on S^k uniformly bounded such that $\mathcal{H}$ is of* VC *type with constants (A, v) and envelope G. For any $H \in \mathcal{H}$, set $\sigma^2(H) = var(H(\xi_1, \ \ldots, \ \xi_k))$ and assume that*

$$\forall j \in \{1, \ldots, k\}, \ \mathbb{E}[H(\xi_1, \ldots, \xi_k) \mid \xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ \ldots, \xi_k] = 0 \textit{ with probability one.}$$

Then, there exist constants $C_1 > 0$, $C_2 \geq 1$, $C_3 > 0$ (depending on v and A) such that for all $t > 0$ satisfying

$$C_1 \sigma \left(n \log \left(\frac{2\|G\|_\infty}{\sigma} \right) \right)^{k/2} \leq t \leq \sigma \left(\frac{n\sigma}{\|G\|_\infty} \right)^k,$$

then

$$\mathbb{P}\left\{ \left\| \sum_{(i_1, \ldots, i_k)} H(\xi_{i_1}, \ldots, \xi_{i_k}) \right\|_{\mathcal{H}} > t \right\} \leq C_2 \exp\left(-C_3 \frac{1}{n} \left(\frac{t}{\sigma} \right)^{2/k} \right),$$

where

$$\|G\|_\infty^2 \geq \sigma^2 \geq \|\mathrm{Var}(H)\|_{\mathcal{H}}^2.$$

The following result is directly derived from that stated above by specifying an appropriate value of t.

Corollary 8 *Let $\xi_1, \xi_2, \ldots$ be an i.i.d. sequence of random variables taking their values in a measurable space $(S, \mathcal{S})$ and distributed according to a probability measure P. Let $\mathcal{H}$ be a class of functions on S^k uniformly bounded such that $\mathcal{H}$ is of* VC *type with constants (A, v) and envelope G. For any $H \in \mathcal{H}$, set $\sigma^2(H) = var(H(\xi_1, \ \ldots, \ \xi_k))$ and assume that*

$$\forall j \in \{1, \ \ldots, \ k\}, \ \ \mathbb{E}[H(\xi_1, \ \ldots, \ \xi_k) \mid \xi_1, \ \ldots, \ \xi_{j-1}, \ \xi_{j+1}, \ \ldots, \ \xi_k] = 0 \textit{ with probability one.}$$

Then, there exist constants $C_1 > 0\,, C_2 \geq 1\,, C_3 > 0$ (depending on v and A) such that

$$\mathbb{P}\left\{ \left\| \sum_{(i_1, \ldots, i_k)} H(\xi_{i_1}, \ldots, \xi_{i_k}) \right\|_{\mathcal{H}} \leq t(n, \sigma, \epsilon) \right\} > 1 - \epsilon,$$

with

$$t(n, \sigma, \epsilon) = \sigma n^{k/2} \left(C_1 \left(\log \left(\frac{2\|G\|_\infty}{\sigma} \right) \right)^{k/2} + \left(\frac{\log(C_2/\epsilon)}{C_3} \right)^{k/2} \right),$$

provided that

$$\|G\|_\infty^2 \left(C_1^{2/k} \log \left(\frac{2\|G\|_\infty}{\sigma} \right) + \frac{\log(C_2/\epsilon)}{C_3} \right) \leq n\sigma^2,$$
$$\sup_{H \in \mathcal{H}} \sigma^2(H) \leq \sigma^2 \leq \|G\|_\infty^2.$$

VC **type classes of functions - Permanence properties.** In the subsequent sections, several results are obtained by applying the concentration bounds recalled above to specific classes of functions built up from the elements of the class Φ and other functions such as $K_h(x)$, $S_C(u \mid x)$ or $g(x)$. To show that these specific classes are VC, we rely on the following lemmas which exhibits situations where the VC type property is preserved, while controlling the constants (A, v) involved. In what follows the kernel K is assumed to satisfy the hypotheses introduced in section 2.2.

Lemma 9 *(see Nolan and Pollard (1987), Lemma 22, Assertion (ii)) The class* $\{z \mapsto K(h^{-1}(x-z)) \ : \ x \in \mathbb{R}^d, \ h > 0\}$ *is a bounded* VC *class of functions.*

The following result is an extension of a result established in the proof of Proposition 8 in Portier and Segers (2018).

Lemma 10 *Let* (V, W) *be a pair of random variables taking their values in* $\mathbb{R}^q$ *and in* $\mathbb{R}^d$ *respectively, denote by* $f_0(v \mid W)$ *the density of the conditional distribution of the random variable* V *given* W*, supposed to be absolutely continuous w.r.t. Lebesgue measure on* $\mathbb{R}^q$*. Let* $\mathcal{F}$ *be a bounded VC class of functions defined on* $\mathbb{R}^q \times \mathbb{R}^d$ *with parameter* (A, v) *and constant envelop* $U_{\mathcal{F}}$*. The class* $\mathcal{G} = \{w \in \mathbb{R}^d \mapsto \mathbb{E}[f(V, W) \mid W = w] : f \in \mathcal{F}\}$ *is a bounded* VC *class of functions with parameter* (A, v) *and constant envelop* $U_{\mathcal{F}}$*.*

Proof Let Q be a probability measure on $\mathbb{R}^d$. Consider $\tilde{Q}$ the probability measure defined through

$$d\tilde{Q}(v) = \int f_0(v|w)Q(\mathrm{d}w)\mathrm{d}v.$$

Let $\epsilon \in (0, 1)$ and consider the centres $f_1, \ \ldots, \ f_N$ of an $\epsilon U_{\mathcal{F}}$-covering of the VC class $\mathcal{F}$ with respect to the metric $L_2(\tilde{Q})$. Let $g \in \mathcal{G}$, i.e., $g : w \in \mathbb{R}^d \mapsto \mathbb{E}[f(V, W) \mid W = w]$ with f in $\mathcal{F}$. Define $g_k = \mathbb{E}[f_k(V, W) \mid W = w]$, for $k = 1, \ldots, N$. There exists $k \in \{1, \ldots, N\}$ such that

$$\begin{aligned}
\int (g(w) - g_k(w))^2 \, Q(\mathrm{d}w) &\leq \int \mathbb{E}\big[(f(V, W) - f_k(V, W))^2 \mid W = w\big] \, Q(\mathrm{d}w) \\
&= \iint (f(v, w) - f_k(v, w))^2 \, f_0(v|w) \, \mathrm{d}v \, Q(\mathrm{d}w) \\
&= \int (f(v, w) - f_k(v, w))^2 \, \tilde{Q}(\mathrm{d}v) \leq \epsilon^2 U_{\mathcal{F}}^2,
\end{aligned}$$

using Jensen's inequality and Fubini's theorem. Consequently, we have:

$$\mathcal{N}\big(\mathcal{G}, L_2(Q), \epsilon U_{\mathcal{F}}\big) \leq \mathcal{N}\big(\mathcal{F}, L_2(\tilde{Q}), \epsilon U_{\mathcal{F}}\big) \leq \left(\frac{A}{\epsilon}\right)^v.$$

Since the constant $U_{\mathcal{F}}$ is an envelop for the class $\mathcal{G}$, the result is established.

■

Lemma 11 *Let Ψ be a VC class of functions defined on $\mathbb{R}^q \times \mathbb{R}^d$ with constant envelop $U > 0$ that satisfies the following Lipschitz property: for all $\psi \in \Psi$, $z \in \mathbb{R}^q$, $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,*

$$|\psi(z, x) - \psi(z, y)| \leq \kappa \|x - y\|.$$

with $\kappa > 0$. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a positive function such that $\int K(u)du = 1$ and $v_K = \int \|u\|^2 K(u)du < \infty$. The class $\mathcal{F} = \{(z, x) \mapsto \int \psi(z, x - hu)K(u)du \ : \ \psi \in \Psi, \, 0 < h \leq \tilde{h}\}$ is a bounded measurable VC class of functions with constant envelope $(\kappa \tilde{h} \sqrt{v}_K + U)$.

Proof Let $0 < \epsilon \leq 1$ and $h_k = k\epsilon\tilde{h}$, $k = 1, \ldots, \lfloor 1/\epsilon \rfloor$, an $(\epsilon\tilde{h})$-subdivision of the interval $(0, \tilde{h}]$. Let Q be a probability measure on $\mathbb{R}^q \times \mathbb{R}^d$. For each k, define μ_k as the probability measure of the random variable $(Z, X - h_k U)$ when $(Z, X, U) \sim Q \times K$. Let $\Psi_{k,j}$, $j = 1, \ldots, N$ be an ϵU-cover of the function class Ψ with respect to $L^2(\mu_k)$. Let $h \in (0, \tilde{h}]$ and $\psi \in \Psi$. For any measurable function f and any k, we have

$$\left\| \int f(z, x - h_k u)K(u)\mathrm{d}u \right\|_{L_2(Q)} \leq \|f\|_{L_2(\mu_k)}.$$

As a consequence, for each k there exists j such that

$$\left\| \int (\psi(z, x - h_k u) - \psi_{k,j}(z, x - h_k u))K(u)\mathrm{d}u \right\|_{L_2(Q)} \leq \epsilon U.$$

Besides, from the Lipschitz property, there exists k such that

$$\left\| \int (\psi(z, x - hu) - \psi(z, x - h_k u))K(u)\mathrm{d}u \right\|_{L_2(Q)} \leq \epsilon\kappa\tilde{h}\sqrt{v}_K.$$

The triangle inequality allows to claim that there exists j and k such that

$$\left\| \int (\psi(z, x - hu) - \psi_{k,j}(z, x - h_k u))K(u)\mathrm{d}u \right\|_{L_2(Q)} = \epsilon(\kappa\tilde{h}\sqrt{v}_K + U).$$

There are $1/\epsilon \times A\epsilon^{-v}$ such functions $\Psi_{k,j}$ meaning that

$$\mathcal{N}\big(\mathcal{F}, \| \cdot \|_{L_2(Q)}, \epsilon(\kappa\tilde{h}\sqrt{v}_K + U)\big) \leq A\epsilon^{-(v+1)},$$

where $(\kappa\tilde{h}\sqrt{v}_K + U)$ is indeed an envelop for the class $\mathcal{F}$. ■

We conclude the section by a preservation result for the product and the inverse.

Lemma 12 *Suppose that $\mathcal{F}$ and $\mathcal{G}$ are two VC classes defined on S with parameters $(A_\mathcal{F}, v_\mathcal{F})$ and $(A_\mathcal{G}, v_\mathcal{G})$ and constant envelops $U_\mathcal{F}$ and $U_\mathcal{G}$, respectively. Then it holds:*

(i) The class $\mathcal{F}\mathcal{G} = \{fg \, : \, f \in \mathcal{F}, \, g \in \mathcal{G}\}$ is VC with parameter $(2(A_\mathcal{F} \vee A_\mathcal{G}), v_\mathcal{F} + v_\mathcal{G})$ and envelop $U_\mathcal{F} U_\mathcal{G}$.

(ii) In addition, if for all $f \in \mathcal{F}$ and $x \in S$, $f(x) \geq b_\mathcal{F}$, then the class $\mathcal{F}^{-1} = \{1/f \, : \, f \in \mathcal{F}, \, g \in \mathcal{G}\}$ is VC with parameter $(A_\mathcal{F} U_\mathcal{F}/b_\mathcal{F}, v_\mathcal{F})$ and envelop $1/b_\mathcal{F}$.

Proof Let $0 < \epsilon \leq 1$ and f_k, $k = 1, \ldots, N_{\mathcal{F}}$ the centers of an $(\epsilon U_{\mathcal{F}})$-covering of $\mathcal{F}$. Similarly, denote by g_k, $k = 1, \ldots, N_{\mathcal{G}}$ the centers of an $(\epsilon U_{\mathcal{G}})$-covering of $\mathcal{G}$. By applying the operation $(f_k \wedge U_{\mathcal{F}}) \vee -U_{\mathcal{F}}$, we can assume without loss of generality that the f_k (resp. g_k) are bounded by $U_{\mathcal{F}}$ (resp. $U_{\mathcal{G}}$). Then for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, there are $k \in \{1, \ldots, N_{\mathcal{F}}\}$ and $j \in \{1, \ldots, N_{\mathcal{G}}\}$ such that

$$\|fg - f_k g_j\| \leq U_{\mathcal{G}} \|f - f_k\| + U_{\mathcal{F}} \|g - g_j\| \leq 2\epsilon U_{\mathcal{G}} U_{\mathcal{F}},$$

which implies that $\mathcal{N}(\mathcal{F}\mathcal{G}, L_2(Q), 2\epsilon U_{\mathcal{G}} U_{\mathcal{F}}) \leq (A_{\mathcal{F}}/\epsilon)^{v_{\mathcal{F}}} (A_{\mathcal{G}}/\epsilon)^{v_{\mathcal{G}}}$. Taking $\epsilon' = 2\epsilon$ gives the result. For the second point, taking $f_k \geq b_{\mathcal{F}}$, we have

$$\|f^{-1} - f_k^{-1}\| \leq \frac{1}{b_{\mathcal{F}}^2} \|f - f_k\| \leq \frac{U_{\mathcal{F}}}{b_{\mathcal{F}}^2} \epsilon,$$

and the result follows taking $\epsilon' = (U_{\mathcal{F}}/b_{\mathcal{F}})\epsilon$.

■

Appendix B. Integration results

In this section we establish useful bounds related to these quantities: kernel smoothers, integrals with respect to signed measures, survival functions and hazard functions namely. This corresponds to Lemmas 13, 14, 15 and 16, respectively. As the previous section, this section is independent from the rest of the paper.

Lemma 13 *Let Ω an open convex subset of $\mathbb{R}^d$. Suppose that f is twice differentiable on Ω such that the greatest eigenvalue of the Hessian matrix is uniformly bounded by $M > 0$, then, if the kernel K is symmetric, i.e., $K(u) = K(-u)$, we have: for all $h > 0$,*

$$\sup_{x \in \Omega} |(K_h * f)(x) - f(x)| \leq \frac{M}{2} h^2 \int \|z\|^2 K(z) \mathrm{d}z.$$

Proof The proof follows the same lines as the proof of Lemma 11 given in Delyon and Portier (2020). ■

Lemma 14 *Let $\theta \in (0, 1)$, $h : \mathbb{R}_+ \to [1, \infty[$ be Borelian, increasing, with limit $1/\theta$ at $+\infty$ and ν be any signed measure on $\mathbb{R}_+$. Then, we have: $\forall T > 0$, $\forall t \in [0, T]$,*

$$\left| \int_0^t h \mathrm{d}\nu \right| \leq \frac{2}{\theta} \sup_{s \in [0,T]} \left| \int_0^s \mathrm{d}\nu \right|.$$

Proof Recall first the identity

$$\sup_{t \geq 0} \left| \int_0^t \mathrm{d}\nu \right| = \sup_{f \in DE} \left| \int f \mathrm{d}\nu \right|, \tag{21}$$

where DE is the space of non-increasing functions valued in $[0,1]$ and vanishing at infinity (see e.g. Dudley (1992)). Since h is increasing from 1 to $1/\theta$, we have for any signed measure ν (whose restriction to $[0,T]$ is denoted by $\nu_{[0,T]}$),

$$\left|\int_0^t h \mathrm{d}\nu\right| = \theta^{-1}\left|\int_0^t \mathrm{d}\nu + \theta\int_0^t \left(h-\theta^{-1}\right)\mathrm{d}\nu\right| \leq 2\theta^{-1}\sup_{f\in DE}\left|\int f\mathrm{d}\nu_{[0,T]}\right|.$$

Then applying (21) we obtain that

$$\left|\int_0^t h\mathrm{d}\nu\right| \leq \frac{2}{\theta}\sup_{s\geq 0}\left|\int_0^s \mathrm{d}\nu_{[0,T]}\right| = \frac{2}{\theta}\sup_{s\in[0,T]}\left|\int_0^s \mathrm{d}\nu\right|.$$

■

Lemma 15 *Let $\tau > 0$. Let $S^{(1)}$ and $S^{(2)}$ be càd-làg non-increasing functions on $\mathbb{R}_+$ such that $S^{(1)}(0) = S^{(2)}(0) = 1$ and $S^{(2)}(\tau) \geq \theta > 0$. For $k \in \{1,\ 2\}$, $\Lambda^{(k)}(t) = -\int_0^t S^{(k)}(\mathrm{d}u)/S^{(k)}(u-)$ is the corresponding cumulative hazard function. We have:*

$$\|S^{(1)} - S^{(2)}\|_{[0,\tau]} \leq 2\theta^{-1}\|\Lambda^{(1)} - \Lambda^{(2)}\|_{[0,\tau]}.$$

Proof Let $t \in [0,\tau]$. As $S^{(2)}(t) > 0$, the integration by part argument of Theorem 3.2.3 in Fleming and Harrington (1991) yields

$$\frac{S^{(1)}(t) - S^{(2)}(t)}{S^{(2)}(t)} = -\int_0^t \frac{S^{(1)}(u-)}{S^{(2)}(u)}(\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u)). \tag{22}$$

Set $\Delta_1(\mathrm{d}u) = (\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u))/S^{(2)}(u)$ and apply the integration by parts formula (refer to page 305 in Shorack and Wellner (2009) for instance) to get

$$\frac{S^{(1)}(t) - S^{(2)}(t)}{S^{(2)}(t)} = -\int_0^t S^{(1)}(u-)\Delta_1(\mathrm{d}u) = -S^{(1)}(t)\Delta_1(t) + \int_0^t \Delta_1(u)S^{(1)}(\mathrm{d}u).$$

Then, as $S^{(2)}(t) \leq 1$, we obtain that

$$|S^{(1)}(t) - S^{(2)}(t)| \leq \left(S^{(1)}(t)|\Delta_1(t)| + (1 - S^{(1)}(t))\sup_{u\in[0,\tau]}|\Delta_1(u)|\right) \leq \sup_{u\in[0,\tau]}|\Delta_1(u)|.$$

We conclude by using Lemma 14 with $d\nu = d(\Lambda^{(1)} - \Lambda^{(2)})$ and $h = 1/S^{(2)}$. ■

Lemma 16 *Let $0 < \theta_1,\ \theta_2 < 1$ and $\tau > 0$. For $k \in \{1,\ 2\}$, define $\Lambda^{(k)}(t) = \int_0^t G^{(k)}(\mathrm{d}u)/H^{(k)}(u)$, where $G^{(k)} : [0,\tau] \to [0,\beta]$ is càd-làg non-decreasing and $H^{(k)} : [0,\tau] \to [\theta_k, 1]$ is Borelian non-increasing. Then, we have:*

$$\|\Lambda^{(1)} - \Lambda^{(2)}\|_{[0,\tau]} \leq \frac{2}{\theta_1}\|G^{(1)} - G^{(2)}\|_{[0,\tau]} + \frac{\beta}{\theta_1\theta_2}\|H^{(1)} - H^{(2)}\|_{[0,\tau]}.$$

Proof Let $t \in [0, \tau]$. Observe that, by triangular inequality,

$$\left|\Lambda^{(1)}(t) - \Lambda^{(2)}(t)\right| = \left|\int_0^t \frac{d(G^{(1)} - G^{(2)})}{H^{(1)}} + \int_0^t \frac{(H^{(2)} - H^{(1)})}{H^{(1)}H^{(2)}} dG^{(2)}\right|$$
$$\leq \frac{2}{\theta_1}\|G^{(1)} - G^{(2)}\|_{[0,\tau]} + \frac{\beta}{\theta_1\theta_2}\|H^{(2)} - H^{(1)}\|_{[0,\tau]},$$

where the bound for the second term on the right hand side is straightforward and that for the first term can be deduced from the application of Lemma 14 with the measure ν equal to $A \mapsto \int_A d(G^{(1)} - G^{(2)})$ and the function h equal to $1/H^{(1)}$. ■

Lemma 17 *Let $\tau > 0$. Let $S^{(1)}$ and $S^{(2)}$ be càd-làg non-increasing functions on $\mathbb{R}_+$ such that $S^{(1)}(0) = S^{(2)}(0) = 1$ and $S^{(2)}(\tau) \geq \theta > 0$. For $k \in \{1,\ 2\}$, define $\Lambda^{(k)}(t) = -\int_0^t S^{(k)}(u-)S^{(k)}(\mathrm{d}u)$ and suppose that $\Lambda^{(k)}(t) = \int_0^t G^{(k)}(\mathrm{d}u)/H^{(k)}(u)$, where $G^{(k)} : [0, \tau] \to [0, \beta]$ and $H^{(k)} : [0, \tau] \to [\theta, 1]$ are respectively non-decreasing and non-increasing borelian functions. Then, there exists a constant $C_{\theta,\beta} > 0$, depending only on θ and β, such that*

$$\sup_{t\in[0,\tau]}\left|\int_0^t \frac{\left(S^{(1)}(u-) - S^{(2)}(u-)\right)}{S^{(2)}(u)}\left(\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u)\right)\right| \leq$$
$$C_{\theta,\beta}\left(\|H^{(1)} - H^{(2)}\|^2_{[0,\tau]} + \|G^{(1)} - G^{(2)}\|^2_{[0,\tau]} + \|W\|_{[0,\tau]}\right),$$

where

$$W(t) = \int_{u=0}^t \int_{s=0}^u \frac{S^{(2)}(s-)\left(G^{(1)}(ds) - G^{(2)}(ds)\right)}{S^{(2)}(s)H^{(2)}(s)} \frac{d\left(G^{(1)}(\mathrm{d}u) - G^{(2)}(\mathrm{d}u)\right)}{S^{(2)}(u)H^{(2)}(u)}.$$

Proof

The proof consists in showing first that there exist constants $C_{1,\theta,\beta}$ and $C_{2,\theta,\beta}$ such that

$$\sup_{t\in[0,\tau]}\left|\int_0^t \frac{(\hat{S}^{(1)}(u-) - S^{(2)}(u-))}{S^{(2)}(u)}(\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u))\right| \leq$$
$$C_{1,\theta,\beta}(\|G^{(1)} - G^{(2)}\|^2_{[0,\tau]} + \|H^{(1)} - H^{(2)}\|^2_{[0,\tau]}) + \|\Pi\|_{[0,\tau]}, \quad (23)$$

where

$$\Pi(t) = \int_0^t \Delta_2(u)\Delta_1(\mathrm{d}u), \quad \Delta_2(t) = \int_0^t S^{(2)}(u-)\Delta_1(\mathrm{d}u), \quad \Delta_1(t) = \int_0^t S^{(2)}(u)^{-1}\Delta(\mathrm{d}u),$$

and $\Delta = \Lambda^{(1)} - \Lambda^{(2)}$, and next that

$$\|\Pi - W\|_{[0,\tau]} \leq C_{2,\theta,\beta}\left(\|H^{(1)} - H^{(2)}\|^2_{[0,\tau]} + \|G^{(1)} - G^{(2)}\|^2_{[0,\tau]}\right). \quad (24)$$

In order to establish (23), we successively apply (22), Fubini's theorem and the integration by part formula:

$$\begin{aligned}
&\int_{u=0}^{t}(S^{(1)}(u-)-S^{(2)}(u-))\Delta_1(\mathrm{d}u)\\
&\quad=-\int_{u=0}^{t}\int_{v=0}^{u-}S^{(1)}(v-)\Delta_1(dv)S^{(2)}(u-)\Delta_1(\mathrm{d}u)\\
&\quad=-\int_{v=0}^{t}\left(\int_{u=v}^{t}S^{(2)}(u-)\Delta_1(\mathrm{d}u)\right)S^{(1)}(v-)\Delta_1(dv)\\
&\quad=-\Delta_2(t)\int_0^t S^{(1)}(v-)\Delta_1(dv)+\int_0^t S^{(1)}(v-)\Pi(dv)\\
&\quad=-\Delta_2(t)\left(S^{(1)}(t)\Delta_1(t)-\int_0^t\Delta_1(u)S^{(1)}(\mathrm{d}u)\right)+S^{(1)}(t)\Pi(t)-\int_0^t\Pi(u)S^{(1)}(\mathrm{d}u)\\
&\quad\leq 2\|\Delta_2\|_{[0,\tau]}\|\Delta_1\|_{[0,\tau]}+2\|\Pi\|_{[0,\tau]}. \qquad (25)
\end{aligned}$$

From (21), we deduce that $\|\Delta_2\|_{[0,\tau]}\leq\|\Delta_1\|_{[0,\tau]}$ (because $S^{(2)}\mathbb{I}_{[0,\tau]}$ belongs to the space DE) and, from Lemma 14, it follows that $\|\Delta_1\|_{[0,\tau]}\leq 2\theta^{-1}\|\Delta\|_{[0,\tau]}$. Apply next Lemma 16 to obtain

$$\|\Delta_2\|_{[0,\tau]}\|\Delta_1\|_{[0,\tau]}\leq\frac{8}{\theta^2}\left(\frac{4}{\theta^2}\|G^{(1)}-G^{(2)}\|^2_{[0,\tau]}+\frac{\beta^2}{\theta^4}\|H^{(1)}-H^{(2)}\|^2_{[0,\tau]}\right).$$

Combined with (25), this proves (23). For (24), the application of the Taylor expansion

$$\frac{1}{x}=\frac{1}{a}-\frac{(x-a)}{a^2}+\frac{(x-a)^2}{xa^2}, \qquad (26)$$

yields

$$d\Delta=\frac{d(G^{(1)}-G^{(2)})}{H^{(2)}}-\frac{(H^{(1)}-H^{(2)})dG^{(1)}}{(H^{(2)})^2}+\frac{(H^{(1)}-H^{(2)})^2dG^{(1)}}{(H^{(2)})^2H^{(1)}}.$$

Set $c(s)=S^{(2)}(s-)/S^{(2)}(s)$. It follows that

$$\begin{aligned}
\Pi(t)=\int_{u=0}^{t}\int_{s=0}^{u}c(s)\Bigg(&\frac{\left(G^{(1)}(ds)-G^{(2)}(ds)\right)}{H^{(2)}(s)}-\frac{\left(H^{(1)}(s)-H^{(2)}(s)\right)G^{(1)}(ds)}{H^{(2)}(s)^2}\\
&+\frac{\left(H^{(1)}(s)-H^{(2)}(s)\right)^2G^{(1)}(ds)}{H^{(2)}(s)^2H^{(1)}(s)}\Bigg)\Delta_1(\mathrm{d}u).
\end{aligned}$$

Observe that

$$\begin{aligned}
&\Pi(t) - W(t) = \\
&-\int_{u=0}^{t}\int_{s=0}^{u} c(s)\frac{\left(G^{(1)}(ds) - G^{(2)}(ds)\right)}{H^{(2)}(s)}\frac{\left(H^{(1)}(u) - H^{(2)}(u)\right)G^{(1)}(\mathrm{d}u)}{S^{(2)}(u)H^{(1)}(u)H^{(2)}(u)} \\
&+\int_{u=0}^{t}\int_{s=0}^{u} c(s)\frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)G^{(1)}(ds)}{H^{(2)}(s)^2}\frac{\left(H^{(1)}(u) - H^{(2)}(u)\right)G^{(1)}(\mathrm{d}u)}{S^{(2)}(u)H^{(1)}(u)H^{(2)}(u)} \\
&-\int_{u=0}^{t}\int_{s=0}^{u} c(s)\frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)G^{(1)}(ds)}{H^{(2)}(s)^2}\frac{\left(G^{(1)}(\mathrm{d}u) - G^{(2)}(\mathrm{d}u)\right)}{S^{(2)}(u)H^{(2)}(u)} \\
&+\int_{u=0}^{t}\int_{s=0}^{u} \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)^2 G^{(1)}(ds)}{H^{(2)}(s)^2 H^{(1)}(s)}\Delta_1(\mathrm{d}u) = A + B + C + D.
\end{aligned}$$

We next bound each term on the right hand side of the equation above. Successively apply Lemma 14 and (21) to get

$$\begin{aligned}
\left|\int_0^u c(s)\frac{\left(G^{(1)}(ds) - G^{(2)}(ds)\right)}{H^{(2)}(s)}\right| &\le \frac{2}{\theta^2}\sup_u\left|\int_0^u S^{(2)}(s-)\Big(G^{(1)}(ds) - G^{(2)}(ds)\Big)\right| \\
&= \frac{2}{\theta^2}\sup_u\left|\int S^{(2)}(s-)\mathbb{I}\left\{s \le u\right\}\Big(G^{(1)}(ds) - G^{(2)}(ds)\Big)\right| \\
&\le \frac{2}{\theta^2}\|G^{(1)} - G^{(2)}\|_{[0,\tau]}.
\end{aligned}$$

Because, for any $u \in [0,\tau]$, $1/\{S^{(2)}(u)H^{(1)}(u)H^{(2)}(u)\} \le 1/\theta^3$, we can write

$$\begin{aligned}
|A| &\le \frac{1}{\theta^3}\int_{u=0}^{t}\left|\int_{s=0}^{u} c(s)\frac{\left(G^{(1)}(ds) - G^{(2)}(ds)\right)}{H^{(2)}(s)}\right|\left|H^{(1)}(u) - H^{(2)}(u)\right|G^{(1)}(\mathrm{d}u) \\
&\le \frac{1}{2\theta^3}\int_{u=0}^{t}\Bigg\{\left(\int_0^u c(s)\frac{\left(G^{(1)}(ds) - G^{(2)}(ds)\right)}{H^{(2)}(s)}\right)^2 \\
&\qquad + \Big(H^{(1)}(u) - H^{(2)}(u)\Big)^2\Bigg\}G^{(1)}(\mathrm{d}u) \\
&\le \beta\left(\frac{2}{\theta^7}\|G^{(1)} - G^{(2)}\|^2_{[0,\tau]} + \|H^{(1)} - H^{(2)}\|^2_{[0,\tau]}\right).
\end{aligned}$$

In addition, because for any $u \in [0,\tau]$, $c(u)/(H^{(2)}(u))^2 \le 1/\theta^3$ we have: $\forall t \in [0,\tau]$,

$$\begin{aligned}
|B| &\le \left(\frac{1}{\theta^3}\right)^2\int_{u=0}^{t}\int_{s=0}^{t}\left|H^{(1)}(s) - H^{(2)}(s)\right|G^{(1)}(ds)\left|H^{(1)}(u) - H^{(2)}(u)\right|G^{(1)}(\mathrm{d}u) \\
&= \frac{1}{\theta^6}\left(\int_{s=0}^{t}\left|H^{(1)}(s) - H^{(2)}(s)\right|G^{(1)}(ds)\right)^2 \\
&\le \frac{\beta^2}{\theta^6}\left\|H^{(1)} - H^{(2)}\right\|^2_{[0,\tau]}.
\end{aligned}$$

Define $\Gamma_2(t) = \int_0^t (G^{(1)}(\mathrm{d}u) - G^{(2)}(\mathrm{d}u))/(S^{(2)}(u)H^{(2)}(u))$. Applying Fubini's theorem, we get

$$\begin{aligned}
|C| &= \left| \int_{u=0}^{t} \int_{s=0}^{u} c(s) \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right) G^{(1)}(ds)}{H^{(2)}(s)^2} \frac{\left(G^{(1)}(\mathrm{d}u) - G^{(2)}(\mathrm{d}u)\right)}{S^{(2)}(u)H^{(2)}(u)} \right| \\
&= \left| \int_{s=0}^{t} \int_{u=s}^{t} \frac{\left(G^{(1)}(\mathrm{d}u) - G^{(2)}(\mathrm{d}u)\right)}{S^{(2)}(u)H^{(2)}(u)} c(s) \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right) G^{(1)}(ds)}{(H^{(2)}(s))^2} \right| \\
&\leq \frac{1}{\theta^3} \int_{s=0}^{t} \left\{ |\Gamma_2(t) - \Gamma_2(s)| \times \left| H^{(1)}(s) - H^{(2)}(s) \right| \right\} G^{(1)}(ds) \\
&\leq \frac{2\beta}{\theta^3} \left\| \Gamma_2 \right\|_{[0,\tau]} \left\| H^{(1)} - H^{(2)} \right\|_{[0,\tau]}.
\end{aligned}$$

Then, using Lemma 14, it follows that

$$\begin{aligned}
|C| &\leq 2(1/\theta^3)\beta(2/\theta^3)\|G^{(1)} - G^{(2)}\|_{[0,\tau]}\|H^{(1)} - H^{(2)}\|_{[0,\tau]} \\
&\leq (1/\theta^3)\beta(2/\theta^3)(\|G^{(1)} - G^{(2)}\|_{[0,\tau]}^2 + \|H^{(1)} - H^{(2)}\|_{[0,\tau]}^2).
\end{aligned}$$

The last term can be treated by means of Fubini's theorem. Indeed, because $\|\Delta_1\|_{[0,\tau]} \leq 2(\beta/\theta)$ and for any $u \in [0,\tau]$, $1/\{H^{(2)}(u)^2 H^{(1)}(u)\} \leq 1/\theta^3$, we have

$$\begin{aligned}
|D| &= \left| \int_{u=0}^{t} \int_{s=0}^{u} \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)^2 G^{(1)}(ds)}{(H^{(2)}(s))^2 H^{(1)}(s)} \Delta_1(\mathrm{d}u) \right| \\
&\leq \int_{s=0}^{t} \left| \left(\int_{u=s}^{t} \Delta_1(\mathrm{d}u) \right) \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)^2 G^{(1)}(ds)}{H^{(2)}(s)^2 H^{(1)}(s)} \right| \\
&\leq \frac{2}{\theta^3}\beta\|\Delta_1\|_{[0,\tau]}\|H^{(1)} - H^{(2)}\|_{[0,\tau]}^2 \\
&\leq \frac{4\beta^2}{\theta^4}\|H^{(1)} - H^{(2)}\|_{[0,\tau]}^2.
\end{aligned}$$

Putting all this together, the triangular inequality leads to (24) .

■

Appendix C. Proof of Proposition 1

We start by establishing 3 useful lemmas, namely Lemma 18, 19 and 20. Then the proof will follow easily. Define

$$\begin{aligned}
H_{0,h}(y,x) &= \mathbb{E}\left[\hat{H}_{0,n}(y,x)\right], \\
H_h(y,x) &= \mathbb{E}\left[\hat{H}_n(y,x)\right].
\end{aligned}$$

and

$$\begin{aligned}
H_0(y,x) &= H_0(y \mid x)g(x), \\
H(y,x) &= H(y \mid x)g(x).
\end{aligned}$$

Lemma 18 *Under Assumption 2, there exists $C_0 > 0$ depending only on K and L such that for all $h > 0$,*

$$\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d} |H_{0,h}(t,x) - H_0(t,x)| \leq C_0 h^2,$$
$$\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d} |H_h(t,x) - H(t,x)| \leq C_0 h^2.$$

Proof The proof results from the application of Lemma 13 combined with the smoothness assumptions stipulated. ■

Lemma 19 *Under Assumption 2, There exist constants $M_1 > 0$ and $h_0 > 0$ depending only on K and L such that:*

$$\mathbb{P}\left\{\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d} |\hat{H}_{0,n}(t,x) - H_{0,h}(t,x)| \leq \sqrt{\frac{M_1|\log(\epsilon h^{d/2})|}{nh^d}}\right\} \geq 1-\epsilon,$$
$$\mathbb{P}\left\{\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d} |\hat{H}_n(t,x) - H_h(t,x)| \leq \sqrt{\frac{M_1|\log(\epsilon h^{d/2})|}{nh^d}}\right\} \geq 1-\epsilon,$$

provided that $h \leq h_0$ and $M_1|\log(\epsilon h^{d/2})| \leq nh^d$.

Proof The exponential inequalities stated above directly result from the application of Corollary 8 to the uniformly bounded VC-type classes (see Lemma 9 and 12) $\{(y,x') \in \mathbb{R}_+\times\mathbb{R}^d \mapsto \mathbb{I}\{y > u\}K((x-x')/h) : (x,u,h) \in \mathbb{R}^d\times\mathbb{R}_+\times\mathbb{R}_+^*\}$ and $\{(y,\delta,x') \in \mathbb{R}_+\times\{0,1\}\times\mathbb{R}^d \mapsto \mathbb{I}\{y > u,\ \delta = 0\}K((x-x')/h) : (x,u,h) \in \mathbb{R}^d\times\mathbb{R}_+\times\mathbb{R}_+^*\}$ whose VC constants are independent from h, with constant envelope $||K||_\infty$, with $k = 1$ and $\sigma^2 = c_{K,L}^2 h^d$ with $c_{K,L} = \sqrt{L\int K^2(x)dx}$. This gives that

$$\mathbb{P}\left\{\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d} |\hat{H}_{0,n}(t,x) - H_{0,h}(t,x)| \leq t\right\} \geq 1-\epsilon,$$
$$\mathbb{P}\left\{\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d} |\hat{H}_n(t,x) - H_h(t,x)| \leq t\right\} \geq 1-\epsilon,$$

with

$$t = \frac{c_{K,L}}{\sqrt{nh^d}}\left(\left(\frac{1}{C_3}\log\left(\frac{C_2}{\epsilon}\right)\right)^{1/2} + C_1\left(\log\left(\frac{2\|K\|_\infty}{c_{K,L}h^{d/2}}\right)\right)^{1/2}\right),$$

provided that $h^{d/2}c_{K,L} \leq ||K||_\infty$ and

$$\frac{\|K\|_\infty^2}{c_{K,L}^2}\left(\frac{1}{C_3}\log\left(\frac{C_2}{\epsilon}\right) + C_1^2\log\left(\frac{2\|K\|_\infty}{c_{K,L}h^{d/2}}\right)\right) \leq nh^d.$$

Since, for any positive numbers a, b, γ, it holds that $a^\gamma + b^\gamma \leq 2^\gamma(a+b)^\gamma$, we find, taking h_0 sufficiently small, that $t^2 \leq \tilde{M}_1|\log(\epsilon h^{d/2})|/nh^d$ for some constant $\tilde{M}_1 > 0$. Finally, taking h_0 sufficiently small ensures that $\log(C_2)/C_3 + C_1^2\log(2\|K\|_\infty/c_{K,L}) \leq C_1^2\log(1/h^{d/2})$, for any $h \leq h_0$, which permits to ensure that the previous condition is satisfied whenever $\tilde{M}_2 |\log(\epsilon h^{d/2})| \leq nh^d$, for some $\tilde{M}_2 > 0$. Take $M_1 = \tilde{M}_1 + \tilde{M}_2$ to obtain the desired result. ■

Lemma 20 *Suppose that Assumptions 1 and 2 are fulfilled. There exist constants $M_1 > 0$ and $h_0 > 0$ depending only on b, L and K such that:*

$$\mathbb{P}\left\{\inf_{(t,x)\in\Gamma_b} \hat{H}_n(t,x) \geq \frac{3b^3}{4}\right\} \geq 1-\epsilon,$$

provided that $h \leq h_0$ and $M_1|\log(\epsilon h^{d/2})| \leq nh^d$.

Proof Define

$$\mathcal{A}_n = \left\{\sup_{(t,x)\in\Gamma_b} |H(t,x) - \hat{H}_n(t,x)| \leq \frac{b^3}{4}\right\}.$$

By virtue of Assumption 1, for any $(t,x) \in \Gamma_b$, we have: $H(t|x) = S_C(t|x)S_Y(t|x) \geq b^2$. As a consequence of $\hat{H}_n(t,x) \geq H(t,x) - |H(t,x) - \hat{H}_n(t,x)|$, $\mathcal{A}_n \subset \{\inf_{(t,x)\in\Gamma_b} \hat{H}_n(t,x) \geq 3b^3/4\}$. Hence we only have to prove that event $\mathcal{A}_n$ occurs with probability $1-\epsilon$ at least. By virtue of Lemma 18, as soon as $h \leq \sqrt{3b^3/(8C_0)}$, we have

$$\sup_{(t,x)\in\Gamma_b} |H_h(t,x) - H(t,x)| \leq \frac{3b^3}{8},$$

and thus

$$\left\{\sup_{(t,x)\in\Gamma_b} |\hat{H}_n(t,x) - H_h(t,x)| \leq \frac{3b^3}{8}\right\} \subset \mathcal{A}_n.$$

Simply use Lemma 19 to ensure that the event in the left-hand side holds with probability $1-\epsilon$ whenever $M_1|\log(\epsilon h^{d/2})| \leq nh^d$ (where M_1 now depends on b, L and K) and $h \leq h_0$. ■

Now we conclude the proof. We start by using Lemma 20 to get that $\inf_{(t,x)\in\Gamma_b} \hat{H}_n(t,x) \geq 3b^3/4$ happens with probability $1-\epsilon/3$. We suppose that this event is realized in the following. Let $(t,x) \in \Gamma_b$ and define

$$\tau_x = \inf\{t \geq 0 : \ \min\{S_C(t|x),\ S_Y(t|x)\} > b\}.$$

Observing that the choice of kernel K guarantees that $\hat{S}_{C,n}(\cdot|x)$ is a (random) survival function, we first apply Lemma 15 with $S^{(1)} = \hat{S}_{C,n}(\cdot|x)$, $S^{(2)} = S_C(\cdot|x)$ and $\theta = b$ to get:

$$\|\hat{S}_{C,n}(\cdot|x) - S_C(\cdot|x)\|_{[0,\tau_x]} \leq (2/b)\|\hat{\Lambda}_{C,n}(\cdot|x) - \Lambda_C(\cdot|x)\|_{[0,\tau_x]}. \tag{27}$$

Applying Lemma 16 with $\Lambda^{(1)}(u) = \Lambda_C(u \mid x) = -\int_0^u H_0(ds,x)/H(s-,x)$, $\Lambda^{(2)}(u) = \hat{\Lambda}_{C,n}(u \mid x) = -\int_0^u \hat{H}_{0,n}(ds,x)/\hat{H}_n(s-,x)$, $\beta = 1$, $\theta_1 = b^3 \leq H(s,x)$, $\theta_2 = 3b^3/4$ (because

$\inf_{(t,x)\in\Gamma_b} \hat{H}_n(t,x) \geq b^3/4)$, next yields

$$\left|\hat{\Lambda}_{C,n}(\cdot|x) - \Lambda_C(\cdot|x)\right|_{[0,\tau_x]} \leq \frac{2}{b^3}\|\hat{H}_{0,n}(\cdot,x) - H_0(\cdot \mid x)g(x)\|_{[0,\tau_x]} + \frac{4}{3b^6}\|\hat{H}_n(\cdot,x) - H(\cdot \mid x)g(x)\|_{[0,\tau_x]}. \tag{28}$$

Combining (27) and (28), using Lemma 18 and taking the supremum over x such that $g(x) > b$, we obtain that, the following bound holds true:

$$\begin{aligned}
&\sup_{(t,x)\in\Gamma_b} |\hat{S}_{C,n}(t \mid x) - S_C(t \mid x)| \\
&\leq \frac{4}{b^4}\sup_{(t,x)\in\Gamma_b} |\hat{H}_{0,n}(t,x) - H_0(t,x)| + \frac{8}{3b^7}\sup_{(t,x)\in\Gamma_b} |\hat{H}_n(t,x) - H(t,x)| \\
&\leq \frac{4}{b^4}\sup_{(t,x)\in\Gamma_b} |\hat{H}_{0,n}(t,x) - H_{0,h}(t,x)| + \frac{4}{b^4}C_0h^2 + \frac{8}{3b^7}\sup_{(t,x)\in\Gamma_b} |\hat{H}_n(t,x) - H_h(t,x)| + \frac{8}{3b^7}C_0h^2.
\end{aligned} \tag{29}$$

Lemma 19 with the probability level $\epsilon/3$ allows us to bound the 2 previous random terms. Combined with the union bound (with 3 events having probability smaller than $\epsilon/3$), permits claiming that with probability greater than $1-\epsilon$:

$$\sup_{(t,x)\in\Gamma_b} \left|\hat{S}_{C,n}(t \mid x) - S_C(t \mid x)\right| \leq \frac{4}{b^4}\left(1 + \frac{2}{3b^3}\right)\left\{C_0h^2 + \sqrt{\frac{M_1\left|\log(\epsilon h^{d/2})\right|}{nh^d}}\right\},$$

provided that (to apply Lemma 19) $h \leq h_0$ and $nh^d \geq M_1|\log(3\epsilon h^{d/2})|$. Examining the different terms and taking h_0 small enough lead to the stated result.

Appendix D. Proof of Proposition 2

Proof of (i): Observe that: $\forall i \in \{1, \ldots, n\}$,

$$\sup_{(t,x)\in\mathcal{K}} |\hat{H}_{0,n}^{(i)}(t,x) - \hat{H}_{0,n}(t,x)| \leq 2||K||_\infty/((n-1)h^d), \tag{30}$$

$$\sup_{(t,x)\in\mathcal{K}} |\hat{H}_n^{(i)}(t,x) - \hat{H}_n(t,x)| \leq 2||K||_\infty/((n-1)h^d). \tag{31}$$

The result follows from the union bound and that each of these events

$$\mathcal{B}_n^{(1)} := \bigcap_{i\leq n}\left\{\forall(t,x)\in\mathcal{K},\ \ \hat{H}_n^{(i)}(t,x) \geq b^3/2\right\},$$

$$\mathcal{B}_n^{(2)} := \bigcap_{i\leq n}\left\{\forall(t,x)\in\mathcal{K},\ \ \hat{S}_{C,n}^{(i)}(t,x) \geq b/2\right\},$$

has probability $1-\epsilon/2$ under the mentioned condition on (n,h). Apply Lemma 20 to choose (n,h) such that with probability $1-\epsilon/2$,

$$\inf_{(t,x)\in\mathcal{K}} \hat{H}_n(t,x) \geq 3b^3/4.$$

Using (31) and the triangle inequality, we get that $\mathcal{B}_n^{(1)}$ has probability $1-\epsilon/2$ provided that $2\|K\|_\infty/((n-1)h^d) \leq b^3/4$.

Suppose that event $\mathcal{B}_n^{(1)}$ is realized. The same reasoning as that used in the proof of Proposition 1 (see (27),(28),(29)), with $S^{(1)}(\cdot) = S_C(\cdot|x)$, $S^{(2)}(\cdot) = S_{C,n}^{(i)}(\cdot|x)$, $\beta = 1$, $\theta_1 = b^3$ and $\theta_2 = b^3/2$ (as $\mathcal{B}_n^{(1)}$ is realized), combined with the triangular inequality, yields: $\forall i \in \{1, \ldots, n\}$,

$$\sup_{(t,x)\in\mathcal{K}} |\hat{S}_{C,n}^{(i)}(t|x) - S_C(t|x)| \leq$$
$$\frac{4}{b^4}\left(\sup_{(t,x)\in\mathcal{K}} |\hat{H}_{0,n}^{(i)}(t,x) - \hat{H}_{0,n}(t,x)| + \sup_{(t,x)\in\mathcal{K}} |\hat{H}_{0,n}(t,x) - H_0(t,x)|\right)$$
$$+ \frac{4}{b^7}\left(\sup_{(t,x)\in\mathcal{K}} |\hat{H}_n^{(i)}(t,x) - \hat{H}_n(t,x)| + \sup_{(t,x)\in\mathcal{K}} |\hat{H}_n(t,x) - H(t,x)|\right).$$

We further assume that

$$\left(\frac{4}{b^4} + \frac{4}{b^7}\right) 2\|K\|_\infty/((n-1)h^d) \leq b/4,$$

which is realized whenever h_0 is small enough and M_1, appearing in the condition $nh^d \geq M_1|\log(h^{d/2}\epsilon)|$, is large enough. From $\mathcal{K} \subset \Gamma_b$ and (30)-(31), it results that

$$\left\{\sup_{(t,x)\in\mathcal{K}} \left|\hat{H}_{0,n}(t,x) - H_0(t,x)\right| \leq \frac{b^5}{32}\right\} \bigcap \left\{\sup_{(t,x)\in\mathcal{K}} \left|\hat{H}_n(t,x) - H(t,x)\right| \leq \frac{b^8}{32}\right\}$$

is included in the set $\mathcal{B}_n^{(2)}$. Following the treatment of (29), it is easy to see that the latter event occurs with probability $1-\epsilon/2$ whenever $h \geq h_0$ is small enough (for the bias) and $nh^d \geq M_1|\log(h^{d/2}\epsilon)|$.

Proof of (ii). We suppose that the event $\mathcal{E}_n$ is realized. For all $i \in \{1, \ldots, n\}$, recall that

$$\hat{\Lambda}_{C,n}^{(i)}(du \mid x) = -\frac{\hat{H}_{0,n}^{(i)}(du, x)}{\hat{H}_n^{(i)}(u-, x)}, \qquad \hat{\Delta}_n^{(i)} = (\hat{\Lambda}_{C,n}^{(i)} - \Lambda_C),$$

and that $c(s \mid x) = S_C(s- \mid x)/S_C(s \mid x)$. It results from Theorem 3.2.3 in (Fleming and Harrington, 1991, page 97) that

$$\frac{\hat{S}_{C,n}^{(i)}(t \mid x) - S_C(t \mid x)}{S_C(t \mid x)} =$$
$$-\int_0^t c(u \mid x)\hat{\Delta}_n^{(i)}(du \mid x) - \int_0^t \frac{(\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x))}{S_C(u \mid x)} \hat{\Delta}_n^{(i)}(du \mid x).$$

The Taylor expansion (26) gives that

$$\begin{aligned}\hat{\Delta}_n^{(i)}(du \mid x) = & \frac{(\hat{H}_{0,n}^{(i)}(du,x) - H_0(du,x))}{H(u,x)} - \frac{(\hat{H}_n^{(i)}(u,x) - H(u,x))\hat{H}_{0,n}^{(i)}(du,x)}{H(u,x)^2} \\ & + \frac{(\hat{H}_n^{(i)}(u,x) - H(u,x))^2 \hat{H}_{0,n}^{(i)}(du,x)}{H(u,x)^2 \hat{H}_n^{(i)}(u,x)},\end{aligned}$$

which implies that

$$\frac{\hat{S}_{C,n}^{(i)}(t \mid x) - S_C(t \mid x)}{S_C(t \mid x)} = \hat{a}_n^{(i)}(t \mid x) + \hat{b}_n^{(i)}(t \mid x), \tag{32}$$

where

$$\begin{aligned}\hat{a}_n^{(i)}(t \mid x) = & -\int_0^t \frac{c(u \mid x)}{H(u,x)}(\hat{H}_{0,n}^{(i)}(du,x) - H_0(du,x)) \\ & + \int_0^t \frac{c(u \mid x)}{H(u,x)^2}(\hat{H}_n^{(i)}(u,x) - H(u,x))\hat{H}_{0,n}^{(i)}(du,x), \\ \hat{b}_n^{(i)}(t \mid x) = & -\int_0^t \frac{c(u \mid x)}{H(u,x)^2 \hat{H}_n^{(i)}(u,x)}(\hat{H}_n^{(i)}(u,x) - H(u,x))^2 \hat{H}_{0,n}^{(i)}(du,x) \\ & - \int_0^t \frac{(\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x))}{S_C(u \mid x)} \hat{\Delta}_n^{(i)}(du \mid x).\end{aligned}$$

Now, using (26), we obtain that: $\forall \varphi \in \Phi$,

$$\begin{aligned}Z_n(\varphi) &= \frac{1}{n}\sum_{i=1}^n \left\{ \delta_i \frac{\varphi(\tilde{Y}_i, X_i)}{\hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(\tilde{Y}, X)}{S_C(\tilde{Y} \mid X)}\right]\right\} \\ &= \frac{1}{n}\sum_{i=1}^n \left\{ \delta_i \frac{\varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y} \mid X)}\right]\right\} \\ &\quad - \frac{1}{n}\sum_{i=1}^n \delta_i \varphi(\tilde{Y}_i, X_i) \left(\frac{\hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i) - S_C(\tilde{Y}_i \mid X_i)}{S_C^2(\tilde{Y}_i \mid X_i)} \right) \\ &\quad + \frac{1}{n}\sum_{i=1}^n \delta_i \varphi(\tilde{Y}_i, X_i) \frac{(S_C(\tilde{Y}_i \mid X_i) - \hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i))^2}{S_C^2(\tilde{Y}_i \mid X_i)\hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i)}.\end{aligned}$$

Then, using (32), we retrieve the expected terms

$$Z_n(\varphi) = L_n(\varphi) + M_n(\varphi) + R_n(\varphi),$$

which proves (ii).

Appendix E. Proof of Proposition 3

The proof is based on the decomposition stated in Proposition 2, combined with the lemmas below that permit to control each term involved in it. Their proofs are given in the next section of the Appendix.

The term $L_n(\varphi)$ is a basic i.i.d. (centred) average. As shown in the lemma stated below, its uniform fluctuations can be controlled by standard results in empirical process theory.

Lemma 21 *Suppose that the hypotheses of Proposition 3 are fulfilled. Then, for any $\epsilon \in (0,1)$, we have with probability at least $1-\epsilon$:*

$$\sup_{\varphi\in\Phi} |L_n(\varphi)| \leq \sqrt{\frac{M_1 \log(M_2/\epsilon)}{n}},$$

provided that $n \geq M_1 \log(M_2/\epsilon)$, where $M_1 > 0$ and $M_2 > 1$ are constants depending on (A, v), K, M_Φ, and b only.

We now turn to the term $M_n(\varphi)$. Observe it can be decomposed as

$$M_n(\varphi) = V_{n,1}(\varphi) + B_{n,1}(\varphi) + V_{n,2}(\varphi) + B_{n,2}(\varphi)$$

where

$$\begin{aligned}
V_{n,1}(\varphi) &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \int_0^{\tilde{Y}_i} \frac{c(u \mid X_i)}{H(u, X_i)^2} \left(\hat{H}_n^{(i)}(u, X_i) - H_h(u, X_i)\right) \hat{H}_{0,n}^{(i)}(du, X_i),\\
B_{n,1}(\varphi) &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \int_0^{\tilde{Y}_i} \frac{c(u \mid X_i)}{H(u, X_i)^2} \left(H_h(u, X_i) - H(u, X_i)\right) \hat{H}_{0,n}^{(i)}(du, X_i),\\
V_{n,2}(\varphi) &= -\frac{1}{n}\sum_{i=1}^n \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \int_0^{\tilde{Y}_i} \frac{c(u \mid X_i)}{H(u, X_i)} \left(\hat{H}_{0,n}^{(i)}(du, X_i) - H_{0,h}(du, X_i)\right),\\
B_{n,2}(\varphi) &= -\frac{1}{n}\sum_{i=1}^n \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \int_0^{\tilde{Y}_i} \frac{c(u \mid X_i)}{H(u, X_i)} \left(H_{0,h}(du, X_i) - H_0(du, X_i)\right).
\end{aligned}$$

Next we treat the bias terms $B_{n,1}$ and $B_{n,2}$.

Lemma 22 *Under the assumptions of Proposition 3, for any $\epsilon \in (0,1)$, we have, with probability $1-\epsilon$:*

$$\sup_{\varphi\in\Phi} |B_{n,1}(\varphi)| \leq M_1 h^2,$$
$$\sup_{\varphi\in\Phi} |B_{n,2}(\varphi)| \leq M_1 h^2,$$

provided that $n \geq M_2 |\log(h^{d/2}\epsilon)|$, where $M_1 > 0$, $M_2 > 0$ depend only on M_Φ, K, L and b.

Now we consider $V_{n,1}(\varphi)$. For simplicity, we set $K_{ij} = K_h(X_i - X_j)$ for $1 \le i,\ j \le n$. We have:

$$\begin{aligned}
V_{n,1}(\varphi) &= \frac{1}{n(n-1)} \sum_{\substack{(i,j)\\ i\neq j}} \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \\
&\quad \times \nVdash_{\tilde{Y}_j \le \tilde{Y}_i} \frac{(1-\delta_j) K_{ij} c(\tilde{Y}_j \mid X_i)}{H(\tilde{Y}_j, X_i)^2} \left(\hat{H}_n^{(i)}(\tilde{Y}_j, X_i) - H_h(\tilde{Y}_j, X_i) \right) \\
&= \frac{1}{n(n-1)^2} \sum_{\substack{(i,j,k)\\ i\neq j, i\neq k}} v_{i,j,k}(\varphi) \\
&= V'_{n,1}(\varphi) + V''_{n,1}(\varphi),
\end{aligned}$$

where, for all $1 \le i, j, k \le n$,

$$v_{i,j,k}(\varphi) \quad = \quad \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \nVdash_{\tilde{Y}_j \le \tilde{Y}_i} \frac{(1-\delta_j) K_{ij} c(\tilde{Y}_j \mid X_i)}{H(\tilde{Y}_j, X_i)^2} \left(\nVdash_{\tilde{Y}_k > \tilde{Y}_j} K_{ik} - H_h(\tilde{Y}_j, X_i) \right)$$

and

$$\begin{aligned}
V'_{n,1}(\varphi) &= \frac{1}{n(n-1)^2} \sum_{\substack{(i,j,k)\\ i\neq j, i\neq k, j\neq k}} v_{i,j,k}(\varphi), \\
V''_{n,1}(\varphi) &= \frac{1}{n(n-1)^2} \sum_{\substack{(i,j)\\ i\neq j}} v_{i,j,j}(\varphi).
\end{aligned}$$

The lemma stated below provides a uniform bound for $V''_{n,1}(\varphi)$.

Lemma 23 *Under the assumptions of Proposition 3, we have, with probability* 1*:*

$$\sup_{\varphi \in \Phi} \left| V''_{n,1}(\varphi) \right| \le \frac{M_1}{nh^d}$$

where $M_1 > 0$ *depends only on* M_Φ, K, L *and* b.

We now consider $V'_{n,1}(\varphi)$. Set $Z_k = (X_k, \tilde{Y}_k, \delta_k)$ for $k \in \{1,\ \ldots,\ n\}$. It can be decomposed as follows:

$$V'_{n,1}(\varphi) = \frac{n-2}{n-1} \left\{ U^{(1)}_{n,1}(\varphi) + U^{(2)}_{n,1}(\varphi) + U^{(3)}_{n,1}(\varphi) + L'_n(\varphi) \right\},$$

with

$$\begin{aligned}
U^{(1)}_{n,1}(\varphi) &= \frac{1}{n(n-1)(n-2)} \sum_{\substack{(i,j,k)\\ i\neq j, i\neq k, j\neq k}} \big\{ v_{i,j,k}(\varphi) - \mathbb{E}[v_{i,j,k}(\varphi) | Z_j,\ Z_k] - \\
&\qquad \mathbb{E}[v_{i,j,k}(\varphi) | Z_i,\ Z_k] + \mathbb{E}[v_{i,j,k}(\varphi) | Z_k] \big\},\\
U^{(2)}_{n,1}(\varphi) &= \frac{1}{n(n-1)} \sum_{\substack{(j,k)\\ j\neq k}} \left\{ \mathbb{E}[v_{i,j,k}(\varphi) | Z_j,\ Z_k] - \mathbb{E}[v_{i,j,k}(\varphi) | Z_k] \right\},\\
U^{(3)}_{n,1}(\varphi) &= \frac{1}{n(n-1)} \sum_{\substack{(i,k)\\ i\neq k}} \left\{ \mathbb{E}[v_{i,j,k}(\varphi) | Z_i,\ Z_k] - \mathbb{E}[v_{i,j,k}(\varphi) | Z_k] \right\},\\
L'_n(\varphi) &= \frac{1}{n} \sum_k \mathbb{E}[v_{i,j,k}(\varphi) | Z_k],
\end{aligned}$$

where i, j and k always denote pairwise distinct indexes, the varying amounts of indexes in the summations being the results of the successive marginalizations necessary to obtain degenerate U-processes. Observe that, for all $\varphi \in \Phi$ and pairwise distinct indexes $i,\ j$ and k in $\{1,\ \ldots,\ n\}$, we have with probability one:

$$\mathbb{E}[v_{i,j,k}(\varphi) \mid Z_i,\ Z_j] = \mathbb{E}[v_{i,j,k}(\varphi) \mid Z_i] = \mathbb{E}[v_{i,j,k}(\varphi) \mid Z_j] = 0.$$

The quantities $U^{(k)}_{n,1}(\varphi)$, $k \in \{1,\ 2,\ 3\}$ are thus degenerate U-statistics of degree 3, 2 and 2 respectively, whereas $L'_n(\varphi)$ is a basic (centred) i.i.d. average. The following result is essentially proved by applying Corollary 8, once the complexity assumptions related to the classes of kernels involved in the definition of these degenerate U-processes have been established. It shows that the terms $U^{(k)}_{n,1}(\varphi)$'s are uniformly negligible.

Lemma 24 *Suppose that the hypotheses of Proposition 3 are fulfilled. There exist constants M_1, M_2 and h_0 depending on (A, v), M_Φ, L, K and b only, such that for any $\epsilon \in (0, 1)$, each of the following events holds true with probability at least $1 - \epsilon$:*

$$\begin{aligned}
\sup_{\varphi\in\Phi} |U^{(1)}_{n,1}(\varphi)| &\le \left(\frac{M_1 |\log(\epsilon h^{d/2})|}{nh^d} \right)^{3/2},\\
\sup_{\varphi\in\Phi} |U^{(2)}_{n,1}(\varphi)| &\le \frac{M_1 |\log(\epsilon h^{d/2})|}{nh^d},\\
\sup_{\varphi\in\Phi} |U^{(3)}_{n,1}(\varphi)| &\le \frac{M_1 |\log(\epsilon h^{d/2})|}{nh^d},
\end{aligned} \tag{33}$$

as soon as $h \le h_0$ and $M_2 |\log(\epsilon h^d)| \le nh^{2d}$.

Maximal deviation inequalities for the $L'_n(\varphi)$ can be obtained by means of classical results in empirical process theory, like for $L_n(\varphi)$.

Lemma 25 *Suppose that the hypotheses of Proposition 3 are fulfilled. Then, for any* $\epsilon \in (0,1)$*, we have with probability at least* $1-\epsilon$*:*

$$\sup_{\varphi\in\Phi}|L_n'(\varphi)| \leq \sqrt{\frac{M_1\log(M_2/\epsilon)}{n}},$$

as soon as $M_2|\log(\epsilon h^d)| \leq nh^{2d}$ *and* $h \leq h_0$ *where* h_0*,* $M_1 > 0$ *and* $M_2 > 1$ *are constants depending on* (A,v)*,* K*,* M_Φ*,* L *and* b *only.*

The two preceding lemmas combined with the union bound directly yield the following result.

Corollary 26 *Suppose that the hypotheses of Proposition 3 are fulfilled. There exist constants* M_1*,* M_2*,* M_3 *and* h_0 *depending on* (A,v)*,* M_Φ*,* L*,* K *and* b *only such that for any* $\epsilon \in (0,1)$*, we have with probability greater than* $1-\epsilon$*:*

$$\sup_{\varphi\in\Phi}\left|V_{n,1}'(\varphi)\right| \leq M_1\left(\sqrt{\frac{\log(M_2/\epsilon)}{n}} + \frac{|\log(\epsilon h^{d/2})|}{nh^d} + \left(\frac{|\log(\epsilon h^{d/2})|}{nh^d}\right)^{3/2}\right),$$

as soon as $h \leq h_0$*,* $M_3|\log(\epsilon h^d)| \leq nh^{2d}$*.*

We next deal with the term $V_{n,2}(\varphi)$.

Lemma 27 *Suppose that the hypotheses of Proposition 3 are fulfilled. There exist constants* M_1*,* M_2*,* M_3 *and* h_0 *depending on* (A,v)*,* M_Φ*,* L*,* K *and* b *only such that for any* $\epsilon \in (0,1)$*, we have with probability greater than* $1-\epsilon$*:*

$$\sup_{\varphi\in\Phi}|V_{n,2}(\varphi)| \leq M_1\left(\sqrt{\frac{\log(M_2/\epsilon)}{n}} + \frac{|\log(\epsilon h^{d/2})|}{nh^{d/2}}\right),$$

as soon as $h \leq h_0$*,* $M_3|\log(\epsilon h^{d/2})| \leq nh^d$*.*

Finally, we consider the residual $R_n(\varphi)$. Recall first that, for all $\varphi \in \Phi$, we have $R_n(\varphi) = R_n'(\varphi) + R_n''(\varphi)$, where

$$R_n'(\varphi) = -\frac{1}{n}\sum_{i=1}^n \frac{\delta_i\varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)}\hat{b}_n^{(i)}(\tilde{Y}_i \mid X_i),$$

$$R_n''(\varphi) = \frac{1}{n}\sum_{i=1}^n \delta_i\varphi(\tilde{Y}_i, X_i)\frac{\left(S_C(\tilde{Y}_i \mid X_i) - \hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i)\right)^2}{S_C^2(\tilde{Y}_i \mid X_i)\hat{S}_{C,n}^{(i)}(\tilde{Y}_i \mid X_i)}.$$

Each of the quantities, $R_n'(\varphi)$ and $R_n''(\varphi)$, is treated separately. We start with $R_n''(\varphi)$.

Lemma 28 *Suppose that the assumptions of Proposition 3 are satisfied. Then, for all* $\epsilon \in (0,1)$*, we have with probability greater than* $1-\epsilon$

$$\sup_{\varphi\in\Phi}\left|R_n''(\varphi)\right| \leq M_1\left(\frac{|\log(\epsilon h^{d/2})|}{nh^d} + \frac{1}{(nh^d)^2} + h^4\right),$$

as soon as $h \leq h_0$ *and* $M_2|\log(\epsilon h^{d/2})| \leq nh^d$*, where* M_1 *and* M_2 *are nonnegative constants depending on* K*,* L*,* M_Φ *and* b *only.*

We now state a uniform bound for $R'_n(\varphi)$.

Lemma 29 *Suppose that the assumptions of Proposition 3 are satisfied. Then, for all $\epsilon \in (0,1)$, we have with probability greater than $1-\epsilon$*

$$\sup_{\varphi\in\Phi} \left|R'_n(\varphi)\right| \leq M_1 \left(\frac{|\log(\epsilon h^{d/2})|}{nh^d} + \frac{\sqrt{|\log(\epsilon h^{d/2})|}}{(nh^d)^{3/2}} + \frac{1}{nh^d} + \frac{1}{(nh^d)^2} + h^2 \right),$$

as soon as $h \leq h_0$ and $M_2|\log(\epsilon h^{d/2})| \leq nh^d$, where M_1 and M_2 are nonnegative constants depending on K, L, M_Φ and b only.

Now we can conclude the proof of Proposition 3 by gathering each of the previous results. First note that they all are valid under the condition that $h \leq h_0$ and $M_1|\log(\epsilon h^{d/2})| \leq nh^d$ and $n \geq M_2 \log(M_3/\epsilon)$. By taking h_0 small enough, the last requirement is no longer necessary. In addition, if $nh^d > 1$ and $|\log(\epsilon h^{d/2})| > 1$ we guarantee that $|\log(\epsilon h^{d/2})|/nh^d \geq 1/nh^d \geq 1/(nh^d)^2$ and $|\log(\epsilon h^{d/2})|^{1/2} \leq |\log(\epsilon h^{d/2})|^{3/2}$, leading to

$$\frac{\sqrt{|\log(\epsilon h^{d/2})|}}{(nh^d)^{3/2}} \leq \left(\frac{|\log(\epsilon h^{d/2})|}{nh^d} \right)^{3/2} \leq \frac{|\log(\epsilon h^{d/2})|}{nh^d}.$$

Using this manipulation, we obtain the stated result.

Appendix F. Intermediary Results

Here we prove lemmas involved in the argument of Proposition 3's proof. Recall that, under the assumptions stipulated: $\forall (t,x) \in \mathcal{K}$,

$$H(t,x) \geq b^3, \quad S_C(t \mid x) \geq b, \quad c(t \mid x) \leq 1/b, \quad H_h(t \mid x) \leq L. \tag{34}$$

F.1 Proof of Lemma 21

The proof is a direct application of Corollary 8 to the i.i.d. sequence $\{(X_n, \tilde{Y}_n, \delta_n) :\ n \geq 1\}$ and the class of functions

$$(x,u,\delta) \in \mathcal{K} \times \{0,1\} \mapsto \frac{\delta\varphi(u,x)}{S_C(u \mid x)},$$

indexed by $(\varphi,\ h) \in \Phi\times]0,h_0]$. The previous class is of VC type in virtue of Lemma 12. We choose $\sigma = \|G\|_\infty = 2M_\Phi/b$, the bound obtained for $L_n(\varphi)$ is simply

$$\begin{aligned}
&(2M_\Phi/b)n^{-1/2} \left(\left(C_1^2 \log(2)\right)^{1/2} + \left(\frac{\log(C_2/\epsilon)}{C_3} \right)^{1/2} \right) \\
&\leq (4M_\Phi/b)n^{-1/2} \left(C_1^2 \log(2) + \frac{\log(C_2/\epsilon)}{C_3} \right)^{1/2},
\end{aligned}$$

where the constants C_1, C_2, C_3 are the ones of Corollary 8. Easy manipulations give the result.

F.2 Proof of Lemma 22

Taking the supremum of each element in the sum we find that

$$|B_{n,1}(\varphi)| \leq \frac{M_\Phi}{b^8} \sup_{(u,x)\in\mathcal{K}} |H_h(u,x) - H(u,x)| \sup_{(u,x)\in\mathcal{K}} |\hat{H}_{0,n}^{(i)}(u,x)|.$$

An appeal to Lemma 18, Lemma 19 combined with (30) gives the first result. Concering $B_{n,2}$, we write

$$|B_{n,2}(\varphi)| \leq \frac{M_\Phi}{b} \sup_{(t,x)\in\mathcal{K}} \left| \int_0^t \frac{c(u \mid x)}{H(u,x)} \left(H_{0,h}(du,x) - H_0(du,x)\right)\right|.$$

Because for any signed measure ν on $\mathbb{R}_+$ and any measurable function f with total variation at most 1 vanishing at infinity, we have (Dudley, 1992),

$$|\int f(u)\,\nu(\mathrm{d}u)| \leq \sup_{t\in\mathbb{R}} |\int_0^t \nu(\mathrm{d}u)|,$$

we conclude that

$$|B_{n,2}(\varphi)| \leq M \sup_{(u,x)\in\mathcal{K}} |H_{0,h}(u,x) - H_0(u,x)|.$$

where $M > 0$ depends only on L, b and M_Φ. Conclude by using the bound given in Lemma 18.

F.3 Proof of Lemma 23

Observe that, for $i \neq j$, we have

$$v_{i,j,j}(\varphi) = -\frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \mathbb{I}\{\tilde{Y}_j \leq \tilde{Y}_i\} \frac{(1-\delta_j) K_{ij} c(\tilde{Y}_j \mid X_i)}{H(\tilde{Y}_j, X_i)^2} H_h(\tilde{Y}_j, X_i).$$

It follows from (34) that

$$|v_{i,j,j}(\varphi)| \leq \frac{M_\Phi}{b^8} \|K\|_\infty h^{-d} L,$$

and since $V''_{n,1}(\varphi)$ is a sum over $n(n-1)$ such terms divided by $n(n-1)^2$ we get the stated bound.

F.4 Proof of Lemma 24

We will use the expression

$$v_{i,j,k}(\varphi) = w_{i,j,k}(\varphi) K_{ik} K_{ij} - \mathbb{E}[w_{i,j,k}(\varphi) K_{ik} K_{ij} \mid Z_i, Z_j]$$

with

$$w_{i,j,k}(\varphi) = \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \mathbb{I}\{\tilde{Y}_j \leq \tilde{Y}_i\} \frac{(1-\delta_j) c(\tilde{Y}_j \mid X_i)}{H(\tilde{Y}_j, X_i)^2} \mathbb{I}\{\tilde{Y}_k > \tilde{Y}_j\}.$$

Using (34), we have that

$$|w_{i,j,k}| \leq \frac{M_\Phi}{b^8}. \tag{35}$$

Recall that $\mathbb{E}[v_{i,j,k}(\varphi) \mid Z_i,\ Z_j] = \mathbb{E}[v_{i,j,k}(\varphi) \mid Z_i] = \mathbb{E}[v_{i,j,k}(\varphi) \mid Z_j] = 0$. As a result, the quantities $U_{n,1}^{(k)}(\varphi)$, $k \in \{1,\ 2,\ 3\}$ are degenerate U-statistics of degree 3, 2 and 2 respectively. For this reason we can apply Corollary 8 to each of them as soon as their respective kernels are shown to form VC classes. The kernel of $h^{2d}n(n-1)^2U_{n,1}^{(1)}$ is

$$h^{2d}\{v_{i,j,k}(\varphi) \quad - \quad \mathbb{E}\left[v_{i,j,k}(\varphi)|Z_j, Z_k\right] \quad - \quad \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_i, Z_k\right] \quad + \quad \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_k\right]\}.$$

Lemma 10 and Corollary 17 in Nolan and Pollard (1987) implies that it is of VC type with constant envelop $8M_\Phi/b^8\|K\|_\infty^2$ as soon as $\{v_{i,j,k}(\varphi)\}$ is of VC type with envelop $M_\Phi/b^8\|K\|_\infty^2$. The later is true in virtue of Lemma 9 and Lemma 12. The same arguments implies that the kernels of $\{h^{2d}n(n-1)U_{n,1}^{(2)}(\varphi)\}$ and $\{h^{2d}n(n-1)U_{n,1}^{(3)}(\varphi)\}$ are of VC type with the constant envelop $4M_\Phi/b^8\|K\|_\infty^2$. In what follows we specify, for each $U_{n,1}^{(k)}(\varphi)$, the value of σ to use in the application of Corollary 8.

The bound for $U_{n,1}^{(1)}(\varphi)$. Observe that

$$\begin{aligned}\mathbb{E}\left[\left(h^{2d}w_{i,j,k}(\varphi)K_{ik}K_{ij}\right)^2\right] &\leq \left(\frac{M_\Phi}{b^8}\right)^2 \mathbb{E}\left[K\left(\frac{X_1 - X_2}{h}\right)^2 K\left(\frac{X_1 - X_3}{h}\right)^2\right] \\ &\leq \left(\frac{M_\Phi}{b^8}\right)^2 L^2 c_K^4 h^{2d}.\end{aligned}$$

where $c_K^2 = \int K^2(x)\mathrm{d}x$. Since we have a sum of 8 terms in the U-statistics of interest, $h^{2d}n(n-1)^2U_{n,1}^{(1)}(\varphi)$, each having an L_2-norm smaller that $\mathbb{E}[h^{4d}v_{i,j,k}(\varphi)^2]$ (by Jensen's inequality), we obtain a bound for the resulting variance (using Minkowski's inequality), in $8^2(M_\Phi/b^8)^2L^2c_K^4h^{2d}$. We apply Corollary 8 with $k=3$ and a value for σ larger than the previous bound. We take $\sigma^2 = 8^2(M_\Phi/b^8)^2L^2c_K^4h^dh_0^d$ (note that $h \leq h_0$) and $\|G\|_\infty = 8M_\Phi\|K\|_\infty^2/b^8$. The conditions are

$$\frac{\|K\|_\infty^4}{L^2c_K^4h_0^d}\left(C_1^{2/3}\log\left(\frac{2\|K\|_\infty^2}{h^{d/2}h_0^{d/2}Lc_K^2}\right) + \frac{\log(C_2/\epsilon)}{C_3}\right) \leq nh^d$$

and

$$L^2c_K^4h^dh_0^d \leq \|K\|_\infty^4,$$

where C_1, C_2 and C_3 are the constants in Corollary 8. The latter conditions are indeed of the type $h \leq h_0$ and $nh^d \geq M_2|\log(\epsilon h^{d/2})|$. This gives

$$\sup_{\varphi\in\Phi}|h^{2d}n(n-1)^2U_{n,1}^{(1)}(\varphi)| \leq \tilde{M}_1h^{d/2}n^{3/2}\left(C_1\left(\log\left(\frac{\tilde{M}_2}{h^{d/2}}\right)\right)^{3/2} + \left(\frac{\log(C_2/\epsilon)}{C_3}\right)^{3/2}\right),$$

where $\tilde{M}_1$ and $\tilde{M}_2$ are constants depending on M_Φ, L, K, b, and h_0. To recover the stated result, one just needs to multiply the previous bound by $1/(n(n-1)(n-2)h^{2d})$ and to use similar manipulations as the ones presented at the end of the proof of Lemma 19.

The bound for $U^{(2)}_{n,1}(\varphi)$. In what follows, we use the shortcut $\mathbb{E}[\cdot|Z_i, Z_j] = \mathbb{E}[\cdot|i, j]$. The kernel of $h^{2d}n(n-1)U^{(2)}_{n,1}(\varphi)$ is given by

$$\begin{aligned}
&h^{2d}\{\mathbb{E}[v_{i,j,k}(\varphi)|j,k] - \mathbb{E}[v_{i,j,k}(\varphi)|k]\}\\
&= h^{2d}\{\mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}|j,k] - \mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}|j]\\
&\quad - \mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}|k] + \mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}]\}.
\end{aligned}$$

By Jensen's inequality and Minkowski's inequality, the variance is then smaller than

$$4^2h^{4d}\mathbb{E}[\mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}|j,k]^2] \le 4^2h^{4d}(M_\Phi/b^8)^2\mathbb{E}\left[K_{ij}K_{ik} \mid j,k\right]^2.$$

But we have

$$\begin{aligned}
\mathbb{E}\left[K_{ij}K_{ik} \mid j,k\right] &= \int K_h(x-X_j)K_h(x-X_k)g(x)\mathrm{d}x\\
&\le L\int K(u)K_h(X_j-X_k+hu)\mathrm{d}u\\
&\le LK^*_h(X_k-X_j),
\end{aligned}$$

where $K^* = K * K$ and $K^*_h(u) = K^*(u/h)/h^d$ (note that $\int K^*(u)\,du = 1$ and $\|K^*\|_\infty \le \|K\|_\infty$). This implies that

$$\begin{aligned}
4^2h^{4d}\mathbb{E}\left[\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij}|j,k\right]^2\right] &\le 4^2h^{4d}\left(\frac{M_\Phi L}{b^8}\right)^2\mathbb{E}\left[K^{*2}_{jk}\right]\\
&\le 4^2h^{3d}\left(\frac{M_\Phi L}{b^8}\right)^2 Lc^2_{K^*}
\end{aligned}$$

where $c^2_K = \int K^2(x)\mathrm{d}x$. The bound (33) is thus obtained by applying Corollary 8 to $h^{2d}n(n-1)U^{(2)}_{n,1}(\varphi)$ with $k = 2$ and

$$\begin{aligned}
\sigma^2 &= 4^2h^{2d}h_0^d\left(\frac{M_\Phi L}{b^8}\right)^2 Lc^2_{K^*}\\
\|G\|_\infty &= 4\frac{M_\Phi\|K\|_\infty}{b^8}.
\end{aligned}$$

The bound for $U^{(3)}_{n,1}(\varphi)$. Based on conditioning arguments, we have

$$\mathbb{E}[v_{i,j,k}(\varphi)|Z_k] = A_k(\varphi) - \mathbb{E}[A_k(\varphi)],$$

where $A_k(\varphi) = \mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}|k]$. We now show that the class of functions

$$\left\{Z_k \mapsto h^dA_k(\varphi) : \varphi\in\Phi\right\}$$

is a VC class with constant envelop. Define

$$\begin{aligned}
\beta_1(Z_i, Z_k) &= K_{ik}\frac{\delta_i\varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)}\\
\beta_2(Z_i, Z_k) &= \int M(X_i + hu, Z_i, Z_k)K(u)du\\
M(X_j, Z_i, Z_k) &= \int \frac{\mathbb{I}\{u \le \tilde{Y}_i,\ u < \tilde{Y}_k\}c(u \mid X_i)}{H(u, X_i)^2}\,H_0(du \mid X_j),
\end{aligned}$$

and observe that

$$\begin{aligned}
&\mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}|i,k] \\
&= \beta_1(Z_i,Z_k)\mathbb{E}\left[\frac{\mathbb{I}\{\tilde{Y}_j \leq \tilde{Y}_i\}(1-\delta_j)c(\tilde{Y}_j \mid X_i)}{H(\tilde{Y}_j,X_i)^2}\mathbb{I}\{\tilde{Y}_k > \tilde{Y}_j\}K_{ij} \mid i,k\right] \\
&= \beta_1(Z_i,Z_k)\mathbb{E}[M(X_j,Z_i,Z_k)K_{ij} \mid i,k] \\
&= \beta_1(Z_i,Z_k)\int M(X_i+hu,Z_i,Z_k)g(X_i+hu)K(u)du,
\end{aligned}$$

where we have used Assumption 1 and the fact that $H_0(du|x) = S_Y(u-|x)S_C(du|x)$. Because for any f with total variation at most 1 vanishing at infinity, we have (Dudley, 1992),

$$\begin{aligned}
\left|\int f(u)\left\{H_0(du \mid x) - H_0(du \mid x')\right\}\right| &\leq \sup_{u\in\mathbb{R}}|H_0(u \mid x) - H_0(u \mid x')| \\
&\leq L\|x-x'\|,
\end{aligned}$$

where the last inequality is a consequence of Assumption 2. The same holds true for g in virtue of Assumption 2. Hence, the map M is uniformly Lipschitz with respect to X_j. Appealing to Lemma 11, we obtain that the kernel $h^d\mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}|i,k]$ is VC with constant envelop $\|K\|_\infty M_\Phi L/b^8$. The same holds true for $h^dA_k(\varphi)$ by Lemma 10. Moreover, observe that for all $(\varphi,h)\in\Phi\times]0,h_0]$, using (35), we have almost surely,

$$|A_k(\varphi)| \leq \frac{M_\Phi}{b^8}\mathbb{E}[K_{ij}K_{ik}|Z_k].$$

Because

$$\begin{aligned}
\mathbb{E}[K_{ij}K_{ik}|Z_k] &= \iint K_h(x-y)K_h(x-X_k)g(x)g(y)dxdy \\
&= \iint K(z)K_h(x-X_k)g(x)g(x-hz)dxdz \\
&\leq L\int K_h(x-X_k)g(x)\underbrace{\left(\int K(z)dz\right)}_{1}dx \\
&\leq L^2,
\end{aligned}$$

it follows that

$$\mathbb{E}[(h^dA_k(\varphi))^2] \leq h^{2d}\left(\frac{M_\Phi L^2}{b^8}\right)^2.$$

Applying Corollary 8 to the kernel $h^d\{A_k(\varphi)-\mathbb{E}[A_k(\varphi)]\}$ with $k=1$ and $\|G\| = 2\|K\|_\infty M_\Phi L/b^8$ and $\sigma^2 = h^{2d}(M_\Phi L^2/b^8)^2$ yields the bound

$$\sup_{\varphi\in\Phi}|nh^dL_n'(\varphi)| \leq \frac{M_\Phi L^2\sqrt{n}h^d}{b^8}\left(C_1\sqrt{\log(2)} + \sqrt{\log(C_2/\epsilon)/C_3}\right),$$

with probability $1-\epsilon$, provided a condition of the type

$$\begin{aligned} nh^{2d} &\geq M_2 |\log(h^d \epsilon)| \\ h &\leq h_0. \end{aligned}$$

Straightforward calculations then give the desired result.

F.5 Proof of Lemma 27

For all $\varphi \in \Phi$, we first set

$$w_{ij}(\varphi) = \frac{\delta_i \varphi(\tilde{Y}_i, X_i) \mathbb{I}\{\tilde{Y}_j \leq \tilde{Y}_i\}(1-\delta_j) K_{ij} c(\tilde{Y}_j \mid X_i)}{S_C(\tilde{Y}_i \mid X_i) H(\tilde{Y}_j \mid X_i)}$$

and observe next that

$$\begin{aligned} V_{n,2}(\varphi) &= -\frac{1}{n} \sum_{i=1}^n \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \int_0^{\tilde{Y}_i} \frac{c(u \mid X_i)}{H(u, X_i)} d\left(\hat{H}_{0,n}^{(i)}(u, X_i) - H_{0,h}(u, X_i)\right) \\ &= -\frac{1}{n(n-1)} \sum_{i \neq j} \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \\ &\times \left(\frac{\mathbb{I}\{\tilde{Y}_j \leq \tilde{Y}_i\}(1-\delta_j) K_{ij} c(\tilde{Y}_j \mid X_i)}{H(\tilde{Y}_j \mid X_i)} - \int_0^{\tilde{Y}_i} \frac{c(u \mid X_i)}{H(u, X_i)} dH_{0,h}(u, X_i) \right) \\ &= -\frac{1}{n(n-1)} \sum_{i \neq j} \{ w_{ij}(\varphi) - \mathbb{E}[w_{ij}(\varphi) \mid Z_j] \} \\ &= U_{n,2}^{(1)}(\varphi) + U_{n,2}^{(2)}(\varphi), \end{aligned}$$

where

$$U_{n,2}^{(1)}(\varphi) = -\frac{1}{n(n-1)} \sum_{i \neq j} \big\{ w_{ij}(\varphi) - \mathbb{E}[w_{ij}(\varphi) \mid Z_j] - \mathbb{E}[w_{ij}(\varphi) \mid Z_i] + \mathbb{E}[w_{12}(\varphi)] \big\}, \tag{36}$$

$$U_{n,2}^{(2)}(\varphi) = \frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}[w_{ij}(\varphi) \mid Z_i] - \mathbb{E}[w_{12}(\varphi)] \right\}. \tag{37}$$

Hence, $V_{n,2}(\varphi)$ can be decomposed as the sum of a degenerate U-statistic (36) and an i.i.d. average (37). Note also that, by (34), we have

$$|w_{ij}(\varphi)| \leq \frac{M_\Phi}{b^5} K_{ij}.$$

The bound for $U_{n,2}^{(1)}(\varphi)$. In virtue of Lemma 9 and Lemma 12, the collection of kernels of the degenerate U-statistics

$$\left\{ h^d n(n-1) U_{n,2}^{(1)}(\varphi) : (\varphi, h) \in \Phi \times]0, h_0] \right\}$$

forms a class of VC type with constants depending only on (v, A), K and h_0. In addition, these terms are all bounded by $4 \times M_\Phi \|K\|_\infty / b^5$ and we have:

$$\operatorname{Var}\left(h^d w_{ij}(\varphi)\right) \leq \left(\frac{4M_\Phi}{b^5}\right)^2 h^d L c_K^2.$$

It thus results from the application of Corollary 8 with $k = 2$ and $\sigma^2 = 4^2 \times \left(M_\Phi / b^5\right)^2 h^d L c_K^2$ that, with probability greater than $1 - \epsilon$

$$\sup_{\varphi \in \Phi} \left| h^d n(n-1) U_{n,1}^{(2)}(\varphi) \right| \leq n h^{d/2} \tilde{M}_1 \left(C_1 \log\left(\frac{\tilde{M}_2}{h^{d/2}}\right) + \frac{\log\left(C_2/\epsilon\right)}{C_3} \right), \tag{38}$$

where $\tilde{M}_1$ and $\tilde{M}_2$ depends on M_Φ, K, L, b, provided a condition of the type $h \leq h_0$ and $nh^d \geq M_2 |\log(\epsilon h^{d/2})|$.

The bound for $U_{n,2}^{(2)}(\varphi)$. Following the proof of Lemma 25, the collection of kernels of $U_{n,2}^{(2)}(\varphi)$ is of VC type with constant envelop. Besides, we have, with probability one:

$$\mathbb{E}[w_{ij}(\varphi) \mid Z_i] \leq \frac{M_\Phi L}{b^5},$$

and therefore $\operatorname{Var}\left(\mathbb{E}[w_{ij}(\varphi) \mid Z_i]\right) \leq \left(\frac{M_\Phi L}{b^5}\right)^2$. Applying thus Corollary 8 with $k = 1$, $\sigma^2 = 4 \times \left(M_\Phi L / b^5\right)^2$ and $\|G\|_\infty = 2 \times M_\Phi L / b^5$, we obtain that, with probability $1 - \epsilon$,

$$\sup_{\varphi \in \Phi} \left| n U_{n,2}^{(2)}(\varphi) \right| \leq C\sqrt{n} \left(C_1 \sqrt{\log(2)} + \sqrt{\frac{\log(C_2/\epsilon)}{C_3}} \right), \tag{39}$$

where C depends on M_Φ, K, L, b, provided a condition of the type $n \geq M_3 |\log(M_4/\epsilon)|$ holds true but this is already implied by $h \leq h_0$ and $nh^d \geq M_2 |\log(\epsilon h^{d/2})|$ whenever h_0 is small. The bound stated in the lemma results from rearranging the bounds (38) and (39).

F.6 Proof of Lemma 28

Use the triangle inequality, (30) and (31) with Lemmas 19 and 18 to get that, with probability $1 - \epsilon$,

$$\max_{i=1,\dots n} \sup_{(t,x) \in \mathcal{K}} |\hat{H}_{0,n}^{(i)}(t,x) - H_0(t,x)| \leq \tilde{M}_1 \left(\frac{1}{nh^d} + \sqrt{\frac{|\log(\epsilon h^{d/2})|}{nh^d}} + h^2 \right),$$

$$\max_{i=1,\dots n} \sup_{(t,x) \in \mathcal{K}} |\hat{H}_{n}^{(i)}(t,x) - H(t,x)| \leq \tilde{M}_2 \left(\frac{1}{nh^d} + \sqrt{\frac{|\log(\epsilon h^{d/2})|}{nh^d}} + h^2 \right).$$

We suppose further that both previous inequalities are realized. Note that under the mentioned condition on (n, h), it holds that $\forall (t,x) \in \mathcal{K}$

$$\inf_{i=1,\dots,n} \hat{H}_n^{(i)}(t,x) \geq \frac{b^3}{2}.$$

In a similar fashion as in the proof of Proposition 1 (see (27),(28) and (29)), we apply Lemma 15 to get that

$$\sup_{(t,x)\in\mathcal{K}} |\hat{S}^{(i)}_{C,n}(t|x) - S_C(t|x)| \leq (2/b) \sup_{(t,x)\in\mathcal{K}} |\hat{\Lambda}^{(i)}_{C,n}(t|x) - \Lambda_C(t|x)|.$$

Then, we apply Lemma 16, with $\theta_1 = b^3$, $\theta_2 = b^3/2$, $\beta = 1$, to finally obtain that: $\forall i \in \{1, \ldots, n\}$,

$$\begin{aligned}
&\sup_{(t,x)\in\mathcal{K}} |\hat{S}^{(i)}_{C,n}(t|x) - S_C(t|x)| \\
&\leq \frac{2}{b}\left(\frac{2}{b^3}\sup_{(t,x)\in\mathcal{K}} |\hat{H}^{(i)}_{0,n}(t,x) - H_0(t,x)| + \frac{2}{b^6}\sup_{(t,x)\in\mathcal{K}} |\hat{H}^{(i)}_{n}(t,x) - H(t,x)|\right) \\
&\leq M_1\left(\frac{1}{nh^d} + \sqrt{\frac{|\log(\epsilon h^{d/2})|}{nh^d}} + h^2\right),
\end{aligned}$$

Hence provided that $h \leq h_0$ and $nh^d \geq M_2|\log(\epsilon h^{d/2})|$, we have

$$\left|R''_n(\varphi)\right| \leq \frac{2M_\Phi}{b^3} \sup_{(t,x)\in\mathcal{K}} \left|S_C(t \mid x) - \hat{S}^{(i)}_{C,n}(t \mid x)\right|^2.$$

F.7 Proof of Lemma 29

Recall first that

$$R'_n(\varphi) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \varphi(\tilde{Y}_i, X_i)}{S_C(\tilde{Y}_i \mid X_i)} \hat{b}^{(i)}_n(\tilde{Y}_i \mid X_i),$$

where

$$\begin{aligned}
\hat{b}^{(i)}_n(t \mid x) = &-\int_0^t \frac{c(u \mid x)}{H(u,x)^2 \hat{H}^{(i)}_n(u,x)} (\hat{H}^{(i)}_n(u,x) - H(u,x))^2 \hat{H}^{(i)}_{0,n}(du,x) \\
&- \int_0^t \frac{(\hat{S}^{(i)}_{C,n}(u- \mid x) - S_C(u- \mid x))}{S_C(u \mid x)} \hat{\Delta}^{(i)}_n(du \mid x).
\end{aligned}$$

and

$$\hat{\Delta}^{(i)}_n(du \mid x) = \hat{\Lambda}^{(i)}_{C,n}(du \mid x) - \Lambda_C(du \mid x).$$

The following argument is based on Lemma 17, stated in section B. Note that, on the event $\mathcal{E}_n$, we have:

$$\begin{aligned}
\left|\hat{b}^{(i)}_n(t \mid x)\right| &\leq \frac{2}{b^{10}} \int \left(\hat{H}^{(i)}_n(u,x) - H_h(u,x)\right)^2 \hat{H}^{(i)}_n(du,x), \\
&\qquad + \left|\int_0^t \frac{(\hat{S}^{(i)}_{C,n}(u- \mid x) - S_C(u- \mid x))}{S_C(u \mid x)} \hat{\Delta}^{(i)}_n(du \mid x)\right| \\
&\leq \frac{2}{b^{10}} \sup_{(u,x)\in\Gamma_b} \left|\hat{H}^{(i)}_n(u,x) - H_h(u,x)\right|^2 \\
&\qquad + \left|\int_0^t \frac{(\hat{S}^{(i)}_{C,n}(u- \mid x) - S_C(u- \mid x))}{S_C(u \mid x)} \hat{\Delta}^{(i)}_n(du \mid x)\right|.
\end{aligned}$$

The application of the Lemma 17, with $S^{(2)}(u) = S_C(u \mid x)$, $S^{(1)}(u) = \hat{S}^{(i)}_{C,n}(u \mid x)$, $\beta = 1$, $\theta = b$ and

$$\begin{aligned}
\Lambda^{(1)}(u) = \hat{\Lambda}^{(i)}_{C,n}(u \mid x) = -\int_{s=0}^{u} \frac{\hat{H}^{(i)}_{0,n}(ds, x)}{\hat{H}^{(i)}_{n}(s-, x)}, \\
\Lambda^{(2)}(u) = \Lambda_C(u \mid x) = -\int_{s=0}^{u} \frac{H_0(ds, x)}{H(s-, x)},
\end{aligned}$$

yields,

$$\begin{aligned}
&\frac{1}{C}\left| \int_0^t \frac{\left(\hat{S}^{(i)}_{C,n}(u- \mid x) - S_C(u- \mid x)\right)}{S_C(u \mid x)} \hat{\Delta}^{(i)}_n(du \mid x) \right| \\
&\qquad \le \sup_{(u,x)\in\Gamma_b} \left| \hat{H}^{(i)}_n(u,x) - H(u,x) \right|^2 \\
&\qquad + \sup_{(u,x)\in\Gamma_b} \left| \hat{H}^{(i)}_{0,n}(u,x) - H_0(u,x) \right|^2 \\
&\qquad + \sup_{(u,x)\in\Gamma_b} \left| \hat{W}^{(i)}_n(u,x) \right|,
\end{aligned}$$

where $C > 0$ depends on b and $\hat{W}^{(i)}_n(t, x)$ is defined as

$$\int_0^t \int_0^u c(s \mid x) \frac{\left(\hat{H}^{(i)}_{0,n}(ds, x) - H_0(ds, x)\right)}{H(s, x)} \frac{\left(\hat{H}^{(i)}_{0,n}(du, x) - H_0(du, x)\right)}{S_C(u \mid x) H(u, x)}.$$

Using (30) and (31) combined with Lemma 18 and Lemma 19, we obtain that with probability at least $1 - \epsilon$:

$$\begin{aligned}
&\sup_{(u,x)\in\Gamma_b} \left| \hat{H}^{(i)}_n(u,x) - H(u,x) \right|^2 + \sup_{(u,x)\in\Gamma_b} \left| \hat{H}^{(i)}_{0,n}(u,x) - H_0(u,x) \right|^2 \\
&\qquad\qquad \le M_1 \left(\frac{1}{(nh^d)^2} + \frac{|\log(\epsilon h^{d/2})|}{nh^d} + h^4 \right),
\end{aligned}$$

as soon as $h \le h_0$ and $M_2 |\log(\epsilon h^{d/2})| \le nh^d$. It remains to show that, with probability at least $1 - \epsilon$:

$$\max_{i\in\{1,\,\ldots,\,n\}} \sup_{(u,x)\in\Gamma_b} \left| \hat{W}^{(i)}_n(u,x) \right| \le M_1 \left(\frac{|\log(h^{d/2}\epsilon)|}{nh^d} + h^2 \right),$$

as soon as $h \le h_0$ and $M_2 |\log(\epsilon h^{d/2})| \le nh^d$. We first define $\hat{W}_{n,1}(t, x)$, for all $(t, x) \in \mathcal{K}$,

$$\int_0^t \int_0^u c(s \mid x) \frac{\left(\hat{H}_{0,n}(ds, x) - H_0(ds, x)\right)}{H(s, x)} \frac{d\left(\hat{H}_{0,n}(du, x) - H_0(du, x)\right)}{S_C(u \mid x) H(u, x)}$$

and notice that, since $c(s \mid x)/(H(s,x)S_C(u \mid x)H(u,x)) \leq 1/b^8$ (using (34)), we have by virtue of (30)

$$\max_{i\in\{1,\, \dots,\, n\}} \sup_{(t,x)\in\mathcal{K}} \left| \hat{W}_{n,1}(t,x) - \hat{W}_n^{(i)}(t,x) \right| \leq C/(nh^d),$$

where C is a constant depending on b and K only. Let

$$\alpha_1(u,x) = (S_C(u \mid x)H(u,x))^{-1} \int_0^u c(s \mid x) \frac{(H_{0,h}(ds,x) - H_0(ds,x))}{H(s,x)}$$

and note that

$$\sup_{(u,x)\in\mathcal{K}} |\alpha_1(u,x)| \leq M_1 \sup_{(u,x)\in\mathcal{K}} |H_{0,h}(u,x) - H_0(u,x)| \leq M_2 h^2.$$

We define $\hat{W}_{n,2}(t,x)$ as

$$\int_0^t \int_0^u c(s \mid x) \frac{\left(\hat{H}_{0,n}(ds,x) - H_{0,h}(ds,x)\right)}{H(s,x)} \frac{\left(\hat{H}_{0,n}(du,x) - H_{0,h}(du,x)\right)}{S_C(u \mid x)H(u,x)},$$

we have

$$\begin{aligned}
\hat{W}_n(t,x) &= \hat{W}_{n,2}(t,x) \\
&+ \int_0^t \int_0^u c(s \mid x) \frac{\left(\hat{H}_{0,n}(ds,x) - H_{0,h}(ds,x)\right)}{H(s,x)} \frac{(H_{0,h}(du,x) - H_0(du,x))}{S_C(u \mid x)H(u,x)} \\
&+ \int_0^t \alpha_1(u,x) \left(\hat{H}_{0,n}(du,x) - H_0(du,x)\right) \\
&+ \int_0^t \alpha_1(u,x) \left(H_{0,h}(du,x) - H_0(du,x)\right).
\end{aligned}$$

Applying Fubini's theorem in the second term, we see that the last three terms are similar. We give the details only for the second one. We have

$$\begin{aligned}
&\left| \int_0^t \alpha_1(u,x)(\hat{H}_{0,n}(du,x) - H_0(du,x)) \right| \\
&\quad \leq \int_0^t |\alpha_1(u,x)|(\hat{H}_{0,n}(du,x) + H_0(du,x)) \\
&\quad \leq \sup_{(u,x)\in\mathcal{K}} |\alpha_1(u,x)| \sup_{(u,x)\in\mathcal{K}} |\hat{H}_{0,n}(u,x) + H_0(u,x)| \\
&\quad \leq M_2 h^2 \sup_{(u,x)\in\mathcal{K}} |\hat{H}_{0,n}(u,x) + H_0(u,x)|
\end{aligned}$$

In addition, observe that

$$\begin{aligned}
\hat{W}_{n,2}(t,x) &= n^{-2} \sum_{i=1}^n \sum_{j=1}^n v_{ij}(t,x) \\
&= n^{-2} \sum_{i\neq j} v_{ij}(t,x) + n^{-2} \sum_{i=1}^n v_{ii}(t,x) \\
&:= U_n(t,x) + M_n(t,x),
\end{aligned}$$

where, for all $1 \leq i,\ j \leq n$, we set

$$\begin{aligned}
v_{ij}(t,x) &= u_{ij}(t,x) - \mathbb{E}[u_{ij}(t,x)|Z_i] - \mathbb{E}[u_{ij}(t,x)|Z_j] + \mathbb{E}[u_{1,2}(t,x)],\\
u_{ij}(t,x) &= \xi_{i,j}(x)\mathbb{I}\{\tilde{Y}_i \leq t\}K_h(X_i - x)K_h(X_j - x),\\
\xi_{i,j}(x) &= \frac{\delta_i\delta_j c(\tilde{Y}_j \mid x)}{S_C(\tilde{Y}_i,x)H(\tilde{Y}_i,x)H(\tilde{Y}_j,x)}\mathbb{I}\{\tilde{Y}_j \leq \tilde{Y}_i\}.
\end{aligned}$$

Because we have, for all $(t,x) \in \mathcal{K}$,

$$\mathbb{E}[v_{12}(t,x)|Z_1] = \mathbb{E}[v_{12}(t,x)|Z_2] = 0,$$

the collection of random variables

$$\{n^2h^{2d}U_n(t,x) : (t,x,h) \in \mathcal{K}\times]0,h_0]\}$$

is a degenerate U-process of order 2. The related class of kernels is uniformly bounded by $4||K||_\infty^2/b^8$ and of VC type, by virtue of classic permanence properties recalled in Appendix A. Observe in addition that

$$\begin{aligned}
\operatorname{Var}\left(h^{2d}v_{12}(t,x)\right) &\leq h^{4d}4^2\mathbb{E}[u_{12}^2(t,x)]\\
&\leq h^{4d}\left(\frac{4}{b^8}\right)^2 E[K_{1x}^2K_{2x}^2]\\
&\leq h^{2d}\left(\frac{4}{b^8}\right)^2 L^2c_K^4.
\end{aligned}$$

Applying Corollary 8 with $k = 2$ and

$$\begin{aligned}
\sigma^2 &= h^d h_0^d\frac{4}{b^8}L^2c_K^4,\\
\|G\|_\infty &= 4\frac{\|K\|_\infty^2}{9b^8},
\end{aligned}$$

we obtain that, with probability greater than $1 - \epsilon$,

$$\sup_{(t,x)\in\mathcal{K}} |U_n(t,x)| \leq M_1\frac{|\log(\epsilon h^{d/2})|}{nh^d},$$

as soon as $M_2|\log(\epsilon h^{d/2})| \leq nh^d$ and $h \leq h_0$. Notice now that, for all $(t,x) \in \mathcal{K}$, $M_n(t,x) = L_n(t,x) + R_n(t,x)$, where

$$\begin{aligned}
L_n(t,x) &= n^{-2}\sum_{i=1}^{n}\{v_{ii}(t,x) - \mathbb{E}[v_{11}(t,x)]\},\\
R_n(t,x) &= n^{-1}\mathbb{E}[v_{11}(t,x)].
\end{aligned}$$

Observing that, for all $(t, x) \in \mathcal{K}$, $|h^{2d} v_{11}(t, x)| \leq 4||K||_\infty^2 / b^8$ and

$$\begin{aligned}
\mathrm{Var}\left(h^{2d} v_{11}(t, x)\right) &\leq h^{4d} 4^2 \mathbb{E}[u_{11}^2(t, x)] \\
&\leq h^{4d} \left(\frac{4}{b^8}\right)^2 \mathbb{E}[K_{1x}^4] \\
&\leq h^d \left(\frac{4}{b^8}\right)^2 L \int K^4(x) dx.
\end{aligned}$$

Hence, the application of Corollary 8 with $k = 1$ to the empirical sums $\{n^2 h^{2d} L_n(t, x) : (t, x, h) \in \mathcal{K} \times]0, h_0]\}$ permits to get that, with probability at least $1 - \epsilon$,

$$\sup_{(t,x) \in \mathcal{K}} |L_n(t, x)| \leq \tilde{M}_1 \frac{\sqrt{|\log(\tilde{M}_2 / h^{d/2})|} + \sqrt{\log(C_2/\epsilon)/C_3}}{(nh^d)^{3/2}},$$

where $\tilde{M}_1$ and $\tilde{M}_2$ are constants depending on K, b and L. The previous bound is valid whenever $M_2 |\log(\epsilon h^{d/2})| \leq nh^d$ and $h \leq h_0$. We also have

$$n^{-1} \mathbb{E}[|v_{11}(t, x)|] \leq 4n^{-1} \mathbb{E}[|u_{11}(t, x)|] \leq (4/b^8) L c_K^2 / (nh^d).$$

This leads to the stated results.

References

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer US, New York, NY, 1993. ISBN 978-0-387-94519-4. doi: 10.1007/978-1-4612-4348-9.

G. Ausset, F. Portier, and S. Clémençon. Machine Learning for Survival Analysis: Empirical Risk Minimization for Censored Distribution Free Regression with Applications. In *NeurIPS ML4H Workshop*, Montreal, Canada, 2018.

H. Bang and A. A. Tsiatis. Median Regression with Censored Cost Data. *Biometrics*, 58(3): 643–649, 2002. ISSN 0006-341X.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, Aug. 2005. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053605000000282.

R. Beran. Nonparametric regression with randomly censored survival data. Technical report, 1981.

J. Buckley and I. James. Linear Regression with Censored Data. *Biometrika*, 66(3):429–436, 1979. ISSN 0006-3444. doi: 10.2307/2335161.

S. Clémençon and F. Portier. Beating Monte Carlo Integration: A Nonasymptotic Study of Kernel Smoothing Methods. In *International Conference on Artificial Intelligence and Statistics*, pages 548–556. PMLR, Mar. 2018.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and Empirical Minimization of U-statistics. *Annals of Statistics*, 36(2):844–874, Apr. 2008. ISSN 0090-5364, 2168-8966. doi: 10.1214/009052607000000910.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 0035-9246.

D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman and Hall/CRC, Boca Raton, 1984. ISBN 978-1-315-13743-8. doi: 10.1201/9781315137438.

D. M. Dabrowska. Uniform Consistency of the Kernel Conditional Kaplan-Meier Estimate. *Annals of Statistics*, 17(3):1157–1167, Sept. 1989. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176347261.

V. de la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Probability and Its Applications. Springer-Verlag, New York, 1999. ISBN 978-0-387-98616-6. doi: 10.1007/978-1-4612-0537-1.

B. Delyon and F. Portier. Integral approximation by kernel smoothing. *Bernoulli*, 22(4): 2177–2208, Nov. 2016. ISSN 1350-7265. doi: 10.3150/15-BEJ725.

B. Delyon and F. Portier. Safe and adaptive importance sampling: A mixture approach. *Annals of Statistics*, Mar. 2020.

Y. Du and M. G. Akritas. Uniform strong representation of the conditional Kaplan-Meier process. *Mathematical Methods of Statistics*, 11(2):152–182, 2002.

R. M. Dudley. Frechet Differentiability, p-Variation and Uniform Donsker Classes. *Annals of Probability*, 20(4):1968–1982, Oct. 1992. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176989537.

U. Einmahl and D. M. Mason. An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators. *Journal of Theoretical Probability*, 13(1):1–37, Jan. 2000. ISSN 1572-9230. doi: 10.1023/A:1007769924157.

T. Fleming and D. Harrington. Counting Processes and Survival Analysis. 1991. doi: 10.2307/2290673.

T. A. Gerds, J. Beyersmann, L. Starkopf, S. Frank, M. J. van der Laan, and M. Schumacher. The Kaplan-Meier Integral in the Presence of Covariates: A Review. *From Statistics to Mathematical Finance*, pages 25–42, 2017. doi: 10.1007/978-3-319-50986-0.

E. Giné and A. Guillou. On consistency of kernel density estimators for randomly censored data: Rates holding uniformly over adaptive intervals. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 37(4):503–522, July 2001. ISSN 0246-0203. doi: 10.1016/S0246-0203(01)01081-0.

E. Giné and H. Sang. Uniform asymptotics for kernel density estimators with variable bandwidths. *Journal of Nonparametric Statistics*, 22(6):773–795, Aug. 2010. ISSN 1048-5252. doi: 10.1080/10485250903483331.

E. Giné, V. Koltchinskii, and J. Zinn. Weighted uniform consistency of kernel density estimators. *Annals of Probability*, 32(3B):2570–2605, July 2004. ISSN 0091-1798, 2168-894X. doi: 10.1214/009117904000000063.

R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12):1109–1112, Sept. 2016. ISSN 0028-4793. doi: 10.1056/NEJMp1607591.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002. ISBN 978-0-387-95441-7. doi: 10.1007/b97848.

T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006. ISSN 14654644. doi: 10.1093/biostatistics/kxj011.

H. Ishwaran and U. B. Kogalur. Random survival forests for R. *R News*, 7(2):25–31, 2007.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008. ISSN 19326157. doi: 10.1214/08-AOAS169.

E. L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. ISSN 0162-1459. doi: 10.2307/2281868.

M. Kohler, K. Máthé, and M. Pintér. Prediction from Randomly Right Censored Data. *Journal of Multivariate Analysis*, 80(1):73–100, Jan. 2002. ISSN 0047259X. doi: 10.1006/jmva.2000.1973.

G. Lecué and S. Mendelson. Learning subgaussian classes : Upper and minimax bounds. *arXiv:1305.4825 [math, stat]*, Sept. 2016.

O. Lopez. Nonparametric Estimation of the Multivariate Distribution Function in a Censored Regression Model with Applications. *Communications in Statistics - Theory and Methods*, 40(15):2639–2660, Aug. 2011. ISSN 0361-0926, 1532-415X. doi: 10.1080/03610926.2010.489175.

O. Lopez, V. Patilea, and I. van Keilegom. Single index regression models in the presence of censoring depending on the covariates. *Bernoulli*, 19(3):721–747, Aug. 2013. ISSN 1350-7265. doi: 10.3150/12-BEJ464.

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22(3), 2016.

P. Major. An estimate on the supremum of a nice class of stochastic integrals and U-statistics. *Probability Theory and Related Fields*, 134(3):489–537, Mar. 2006. ISSN 0178-8051, 1432-2064. doi: 10.1007/s00440-005-0440-9.

N. R. Mann, R. E. Schafer, and N. D. Singpurwalla. *Methods for Statistical Analysis of Reliability and Life Data*. Wiley, 1974. ISBN 978-0-471-56737-0.

A. M. Molinaro, S. Dudoit, and M. J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1 SPEC. ISS.):154–177, 2004. ISSN 10957243. doi: 10.1016/j.jmva.2004.02.003.

D. Nolan and D. Pollard. U-Processes: Rates of Convergence. *Annals of Statistics*, 15(2): 780–799, June 1987. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176350374.

S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct. 2010. ISSN 1558-2191. doi: 10.1109/TKDE.2009.191.

G. Papa, A. Bellet, and S. Clémençon. On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 694–702. Curran Associates, Inc., 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

S. Pölsterl. Scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. ISSN 1533-7928.

S. Pölsterl, N. Navab, and A. Katouzian. Fast Training of Support Vector Machines for Survival Analysis. In A. Appice, P. P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, and C. Soares, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 243–259. Springer International Publishing, 2015. ISBN 978-3-319-23525-7.

S. Pölsterl, N. Navab, and A. Katouzian. An Efficient Training Algorithm for Kernel Survival Support Vector Machines. In *CML PKDD MLLS 2016*, 2016.

F. Portier and J. Segers. On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis*, 166(C):160–181, 2018.

A. Rotnitzky and J. M. Robins. Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. *AIDS Epidemiology*, 88(424):1473, 1992. ISSN 01621459. doi: 10.1007/978-1-4757-1229-2_14.

P. Royston and M. K. B. Parmar. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30(19):2409–2421, 2011. ISSN 1097-0258. doi: 10.1002/sim.4274.

D. Rubin and M. J. van der Laan. A Doubly Robust Censoring Unbiased Transformation. *The International Journal of Biostatistics*, 3(1), Jan. 2007. ISSN 1557-4679. doi: 10.2202/1557-4679.1052.

G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Jan. 2009. ISBN 978-0-89871-684-9. doi: 10.1137/1.9780898719017.

J. A. Steingrimsson and S. Morrison. Deep learning for survival outcomes. *Statistics in Medicine*, 39(17):2339–2349, 2020. ISSN 1097-0258. doi: 10.1002/sim.8542.

J. A. Steingrimsson, L. Diao, A. M. Molinaro, and R. L. Strawderman. Doubly robust survival trees: Doubly Robust Survival Trees. *Statistics in Medicine*, 35(20):3595–3612, Sept. 2016. ISSN 02776715. doi: 10.1002/sim.6949.

J. A. Steingrimsson, L. Diao, and R. L. Strawderman. Censoring Unbiased Regression Trees and Ensembles. *Journal of the American Statistical Association*, 114(525):370–383, Jan. 2019. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1407775.

W. Stute. Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45(1), 1993. doi: 10.1006/jmva.1993.1028.

W. Stute. The Central Limit Theorem Under Random Censorship. *Annals of Statistics*, 23 (2):422–439, Apr. 1995. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176324528.

W. Stute. Distributional Convergence under Random Censorship when Covariables are Present. *Scandinavian Journal of Statistics*, 23(4):461–471, 1996.

V. Van Belle, K. Pelckmans, J. A. K. Suykens, and S. Van Huffel. Support Vector Machines for Survival Analysis. *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare*, 2007.

V. Van Belle, K. Pelckmans, J. A. K. Suykens, and S. V. Huffel. Learning Transformation Models for Ranking and Survival Analysis. *Journal of Machine Learning Research*, 12: 819–862, 2011. ISSN 15324435.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer New York, New York, NY, first edition, 2003. ISBN 978-0-387-21700-0.

I. van Keilegom and N. Veraverbeke. Uniform strong convergence results for the conditional kaplan-meier estimator and its quantiles. *Communications in Statistics - Theory and Methods*, 25(10):2251–2265, Jan. 1996. ISSN 0361-0926. doi: 10.1080/03610929608831836.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Number 60. Chapman & Hall, Boca Raton, FL, U.S., Dec. 1994. ISBN 978-0-412-55270-0.