Electronic Journal of Statistics
Vol. 11 (2017) 532–569
ISSN: 1935-7524
DOI: 10.1214/17-EJS1231

Support vector regression for right censored data[*]

Yair Goldberg

Department of Statistics
The University of Haifa
Mount Carmel, Haifa 31905, Israel
e-mail: ygoldberg@stat.haifa.ac.il

and

Michael R. Kosorok

Department of Biostatistics
The University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
e-mail: kosorok@unc.edu

Abstract: We develop a unified approach for classification and regression support vector machines for when the responses are subject to right censoring. We provide finite sample bounds on the generalization error of the algorithm, prove risk consistency for a wide class of probability measures, and study the associated learning rates. We apply the general methodology to estimation of the (truncated) mean, median, quantiles, and for classification problems. We present a simulation study that demonstrates the performance of the proposed approach.

Keywords and phrases: Support vector regression, right censored data, generalization error, universal consistency, misspecification models.

Received February 2016.

1. Introduction

In many medical studies, estimating the failure time distribution function, or quantities that depend on this distribution, as a function of patient demographic and prognostic variables, is of central importance for risk assessment and health planing. Frequently, such data is subject to right censoring.

The goal of this paper is to develop tools for analyzing such data using a machine learning approach. Machine learning techniques have proved themselves useful in many real-world data-analysis problems, which are often of high dimension and require nonlinear methods (Breiman, 2001; Hofmann et al., 2008). Despite the success of machine learning techniques for i.i.d. data, there has been

[*] The authors thank Danyu Lin for many helpful discussions and suggestions. The first author was funded in part by a Gillings Innovation Laboratory (GIL) award at the UNC Gillings School of Global Public Health and by NSF grant DMS-1407732. The second author was funded in part by NCI grant CA142538 and by NSF grant DMS-1407732.

little attempt to rigourously adapt these methods to right censored data. In this paper we propose a support vector machine (SVM) learning method for right censored data. The choice of SVM is motivated by the fact that SVM learning methods are easy-to-compute techniques that enable estimation under weak or no assumptions on the distribution (Steinwart and Chirstmann, 2008). Moreover, when using SVM, only the function of interest is needed to be estimated directly and there is no need to estimate the whole (possibly high-dimensional) distribution of the failure time given the covariates.

Support vector machine learning methods are a collection of algorithms that attempt to minimize the risk with respect to some loss function. An SVM learning method typically minimizes a regularized version of the empirical risk over some reproducing kernel Hilbert space (RKHS). The resulting minimizer is referred to as the SVM decision function. The SVM learning method is the mapping that assigns to each data set its corresponding SVM decision function.

We adapt the SVM framework to right censored data as follows. First, we represent the distribution's quantity of interest as a Bayes decision function, i.e., a function that minimizes the risk with respect to a loss function. We then construct a novel data-dependent version of this loss function using inverse-probability-of-censoring weighting (Robins et al., 1994). We then minimize a regularized empirical risk with respect to this data-dependent loss function to obtain what we refer to as a censored SVM decision function. Finally, we define the censored SVM learning method as the mapping that assigns for every censored data set its corresponding censored SVM decision function.

Note that unlike the standard SVM decision function, the proposed censored SVM decision function is obtained as the minimizer of a data-dependent loss function. In other words, for each data set, a different minimization loss function is defined. Moreover, minimizing the empirical risk no longer consists of minimizing a sum of i.i.d. observations. Consequently, we were required to develop novel theoretical techniques for the study of the generalization properties of the censored SVM learning method.

We prove a number of theoretical results for the proposed censored SVM learning method. We first prove that the censored SVM decision function is measurable and unique. We then show that the censored SVM learning method is a measurable learning method. We provide a probabilistic finite-sample bound on the difference in risk between the learned censored SVM decision function and the Bayes risk. We further show that the SVM learning method is consistent for every probability measure for which the censoring is independent of the failure time given the covariates, and the probability that no censoring occurs is positive given the covariates. We compute learning rates for the censored SVM learning method. Finally, we provide a simulation study that demonstrates the performance of the proposed censored SVM learning method. Our results are obtained under some conditions on the approximation RKHS and on the loss function, which can be easily verified. We also assume that the estimation of censoring probability at the observed points is consistent.

We note that a number of other learning algorithms have been suggested for survival data. Biganzoli et al. (1998) and Ripley and Ripley (2001) used neural networks. Segal (1988), Hothorn et al. (2004), Ishwaran et al. (2008), and Zhu and Kosorok (2011), among others, suggested versions of splitting trees and random forests for survival data. Johnson et al. (2004), Shivaswamy et al. (2007), Shim and Hwang (2009), and Zhao et al. (2011), among others, suggested versions of SVM different from the proposed censored SVM. The theoretical properties of most of these algorithms have never been studied. Exceptions include the consistency proof of Ishwaran and Kogalur (2010) for random survival trees, which requires the assumption that the feature space is discrete and finite. Consistency result in the context of support vector regression were given by Eleuteri and Taktak (2011). In the context of multistage decision problems, Goldberg and Kosorok (2012b) proposed a Q-learning algorithm for right censored data for which a theoretical justification is given, under the assumption that the censoring is independent of both failure time and covariates. In the context of individualized treatment regimes, Zhao et al. (2015) presented a weighted outcome learning algorithm that can handle right censored data. However, we believe that the proposed censored SVM and the accompanying theoretical evaluation given in this paper represent a significant innovation in developing methodology for learning in survival data.

The proposed censored SVM approach uses inverse-probability-of-censoring weighting to construct the data-dependent loss function, which in turn requires estimation of the censoring probability at observed failure times. This potential drawback is offset by the benefit of not having to estimate the entire failure time distribution. We remark that in many applications it is reasonable to assume that the censoring mechanism is simpler than the failure-time distribution. For example, the censoring distribution may be known in advance by the researcher. Also, in many studies the main reason for censoring is administrative and hence it does not depend on the complicated structure of the covariates. In the latter case, efficient estimators, such as the Kaplan-Meier estimator, are readily available for the censoring distribution. Even when the censoring distribution does depend on the explanatory variables, it can be simpler than that of the failure time, and hence beneficial to use the proposed method. As an example, consider the case in which the explanatory variables include high-dimensional gene expression data. While it is reasonable to assume that the patient drop-out time distribution is independent of the genetic data, such an assumption for the failure time distribution is unlikely. We present results for both correctly specified and misspecified censoring models. We also discuss in detail the special cases of the Kaplan-Meier and the Cox model estimators (Fleming and Harrington, 1991).

While the main contribution of this paper is the proposed censored SVM learning method and the study of its properties, an additional contribution is the development of a general machine learning framework for right censored data. The principles and definitions that we discuss in the context of right censored data, such as learning methods, measurability, consistency, and learning rates, are independent of the proposed SVM learning method. This framework can be adapted to other learning methods for right censored data, as well as for learning methods for other missing data mechanisms.

The paper is organized as follows. In Section 2 we review right-censored data and SVM learning methods. In Section 3 we briefly discuss the use of SVM for right-censored data when no censoring is present. Section 4 discusses the difficulties that arise when applying SVM to right censored data and presents the proposed censored SVM learning method. Section 5 contains the main theoretical results, including finite sample bounds and consistency. Simulations appear in Section 6. Concluding remarks appear in Section 7. The lengthier key proofs are provided in the Appendix. Finally, a link to the Matlab code for both the algorithm and the simulations can be found in Supplementary Material (Goldberg and Kosorok, 2017).

2. Preliminaries

In this section, we establish the notation used throughout the paper. We begin by describing the data setup (Section 2.1). We then discuss loss functions (Section 2.2). Finally we discuss SVM learning methods (Section 2.3). The notation for right censored data generally follows Fleming and Harrington (1991) (hereafter abbreviated FH91). For the loss function and the SVM definitions, we follow Steinwart and Chirstmann (2008) (hereafter abbreviated SC08).

2.1. Data setup

We assume the data consist of n independent and identically-distributed random triplets $D = \{(Z_1, U_1, \delta_1), \dots, (Z_n, U_n, \delta_n)\}$. The random vector Z is a covariate vector that takes its values in a set $\mathcal{Z} \subset \mathbb{R}^d$. The random variable U is the observed time defined by $U = T \wedge C$, where T is a failure time that takes its values in $\mathcal{T} = [0, \tau]$, for some positive constant τ, C is the censoring time, and where $a \wedge b = \min(a, b)$. The indicator $\delta = \mathbf{1}\{T \leq C\}$ is the failure indicator, where $\mathbf{1}\{A\}$ is 1 if A is true and 0 otherwise, i.e., $\delta = 1$ whenever a failure time is observed. We note that in general failure times can take values larger than τ. In such cases we replace the failure times with their clipped-at-τ value (see, for example, Karrison, 1997; Zucker, 1998).

Let $S(t|Z) = P(T > t|Z)$ be the survival functions of T, and let $G(t|Z) = P(C \geq t|Z)$ be the left-hand limit of the survival function of C given the covariate vector Z.

We assume that the censoring mechanism can be described by some simple model. Below, we consider two examples. More details regarding these examples are given in Appendix A.1.

Example 1. ***Independent censoring:*** *Assume that C is independent of both T and Z. Then the Kaplan-Meier estimator is a consistent and efficient estimator for the survival function G (FH91).*

Example 2. ***The proportional hazards model:*** *Assume that the hazard of C given Z is of the form $e^{Z'\boldsymbol{\beta}}d\Lambda$ for some unknown vector $\boldsymbol{\beta} \in \mathbb{R}^d$ and some continuous unknown nondecreasing function Λ with $\Lambda(0) = 0$ and $0 < \Lambda(\tau) <$*

∞. Then a consistent estimator for survival function G is obtained by combining the Cox's estimator for the vector β and Breslow's estimator for Λ (FH91).

We need the following two assumptions:

(A1) There is a constant $K > 0$, such that $\inf_{z\in\mathcal{Z}} G(\tau|z) \geq 2K > 0$.
(A2) C is independent of T, given Z.

The first assumption assures that there is a positive probability of no censoring over the observation time range ($\mathcal{T} = [0, \tau]$). The second assumption is standard in survival analysis and ensures that the joint nonparametric distribution of the survival and censoring times, given the covariates, is identifiable. For a comprehensive discussion about these assumptions, we refer the reader to (Tsiatis, 2006, Chapters 6-7).

Remark 1. *For Example 1 and for Example 2 when $\mathcal{Z}$ is compact, Assumption (A1) holds whenever*

$$P(C \geq \tau) > 0\,.$$

Moreover, if τ is not chosen to be the largest failure time, then this assumption can be validated: if there is at least one (possibly clipped-at-τ) failure time $T_i = \tau$ then (A1) holds almost surely.

Remark 2. *By Assumption (A1), $\inf_{z\in\mathcal{Z}} G(\tau|z) \geq 2K > 0$, and thus if the estimator $\hat{G}_n$ is consistent for G, then, for all n large enough, $\inf_{z\in\mathcal{Z}} \hat{G}_n(\tau|z) > K > 0$. Let $P_{\hat{G}_n,n} = P(\inf_Z \hat{G}_n(\tau|Z) < K)$.*

2.2. Loss functions

Let the input space $(\mathcal{Z}, \mathcal{A})$ be a measurable space. Let the response space $\mathcal{Y}$ be a closed subset of $\mathbb{R}$. Let P be a measure on $\mathcal{Z} \times \mathcal{Y}$.

A function $L : \mathcal{Z} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$ is a *loss function* if it is measurable. We say that a loss function L is *convex* if $L(z, y, \cdot)$ is convex for every $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$. We say that a loss function L is *locally Lipschitz continuous* with Lipschitz local constant function $c_L(\cdot)$ if for every $a > 0$

$$\sup_{\substack{z\in\mathcal{Z}\\ y\in\mathcal{Y}}} |L(z, y, s) - L(z, y, s')| < c_L(a)|s - s'|\,, \quad s, s' \in [-a, a]\,.$$

We say that L is *Lipschitz continuous* if there is a constant c_L such that the above holds for any a with $c_L(a) = c_L$.

For any measurable function $f : \mathcal{Z} \mapsto \mathbb{R}$ we define the *L-risk* of f with respect to the measure P as $\mathcal{R}_{L,P}(f) = E_P[L(Z, Y, f(Z))]$. We define the *Bayes risk* $\mathcal{R}^*_{L,P}$ with respect to loss function L and measure P as $\inf_f \mathcal{R}_{L,P}(f)$, where the infimum is taken over all measurable functions $f : \mathcal{Z} \mapsto \mathbb{R}$. A function $f^*_{L,P}$ that achieves this infimum is called a Bayes decision function.

We now present a few examples of loss functions and their respective Bayes decision functions. In the next section we discuss the use of these loss functions for right censored data.

Example 3. ***Binary classification:*** *Assume that* $\mathcal{Y} = \{-1, 1\}$*. We would like to find a function* $f : \mathcal{Z} \mapsto \{-1, 1\}$ *such that for almost every* z*,* $P(f(z) = Y|Z = z) \geq 1/2$*. One can think of* f *as a function that predicts the label* y *of a pair* (z, y) *when only* z *is observed. In this case, the desired function is the Bayes decision function* $f^*_{L,P}$ *with respect to the loss function* $L_{\mathrm{BC}}(z, y, s) = \mathbf{1}\{y \cdot \mathrm{sign}(s) \neq 1\}$*. In practice, since the loss function* L_{BC} *is not convex, it is usually replaced by the hinge loss function* $L_{\mathrm{HL}}(z, y, s) = \max\{0, 1 - ys\}$*.*

Example 4. ***Expectation:*** *Assume that* $\mathcal{Y} = \mathbb{R}$*. We would like to estimate the expectation of the response* Y *given the covariates* Z*. The conditional expectation is the Bayes decision function* $f^*_{L,P}$ *with respect to the squared error loss function* $L_{\mathrm{LS}}(z, y, s) = (y - s)^2$*.*

Example 5. ***Median and quantiles:*** *Assume that* $\mathcal{Y} = \mathbb{R}$*. We would like to estimate the median of* $Y|Z$*. The conditional median is the Bayes decision function* $f^*_{L,P}$ *for the absolute deviation loss function* $L_{\mathrm{AD}}(z, y, s) = |y - s|$*. Similarly, the* α*-quantile of* Y *given* Z *is obtained as the Bayes decision function for the loss function*

$$L_\alpha(z, y, s) = \begin{cases} -(1-\alpha)(y - s) & \text{if } s \geq y \\ \alpha(y - s) & \text{if } s < y \end{cases} \quad , \ \alpha \in (0, 1)\,.$$

Note that the functions L_{HL}, L_{LS}, L_{AD}, and L_α for $\alpha \in (0, 1)$ are all convex. Moreover, all these functions except L_{LS} are Lipschitz continuous, and L_{LS} is locally Lipschitz continuous when $\mathcal{Y}$ is compact.

2.3. Support vector machine (SVM) learning methods

Let L be a convex locally Lipschitz continuous loss function. Let H be a separable reproducing kernel Hilbert space (RKHS) of a bounded measurable kernel on $\mathcal{Z}$ (for details regarding RKHS, the reader is referred to SC08, Chapter 4).

Let $D_0 = \{(Z_1, Y_1), \ldots, (Z_n, Y_n)\}$ be a set of n i.i.d. observations drawn according to the probability measure P. Fix λ and let H be as above. Define the *empirical SVM decision function*

$$f_{D_0,\lambda} = \operatorname*{argmin}_{f \in H} \lambda \|f\|^2_H + \mathcal{R}_{L,D_0}(f)\,, \tag{1}$$

where

$$\mathcal{R}_{L,D_0}(f) \equiv \mathbb{P}_n L(Z, Y, f(Z)) \equiv \frac{1}{n}\sum_{i=1}^{n} L(Z_i, Y_i, f(Z_i))$$

is the empirical risk, and where $\mathbb{P}_n$ is the empirical measure, i.e., $\mathbb{P}_n f(X) = n^{-1}\sum_{i=1}^n f(X_i)$. Define Pf to be the expectation of f with respect to P.

For some sequence $\{\lambda_n\}$, define the *SVM learning method* $\mathfrak{L}$, as the map

$$\begin{aligned} (\mathcal{Z} \times \mathcal{Y})^n \times \mathcal{Z} &\mapsto \mathbb{R} \\ (D_0, z) &\mapsto f_{D_0,\lambda_n}(z) \end{aligned} \tag{2}$$

for all $n \geq 1$. We say that $\mathfrak{L}$ is *measurable* if it is measurable for all n with respect to the minimal completion of the product σ-field on $(\mathcal{Z} \times \mathcal{Y})^n \times \mathcal{Z}$. We say that $\mathfrak{L}$ is (L-risk) P-consistent if for all $\varepsilon > 0$

$$\lim_{n\to\infty} P(D_0 \in (Z \times \mathcal{Y})^n \,:\, \mathcal{R}_{L,P}(f_{D_0,\lambda_n}) \leq \mathcal{R}^*_{L,P} + \varepsilon) = 1\,. \tag{3}$$

We say that $\mathfrak{L}$ is *universally consistent* if for all distributions P on $\mathcal{Z} \times \mathcal{Y}$, $\mathfrak{L}$ is P-consistent.

We now briefly summarize some known results regarding SVM learning methods needed for our exposition. More advanced results can be obtained using conditions on the functional spaces and clipping. We will discuss these ideas in the context of censoring in Section 5.

Theorem 1. *Let $L : \mathcal{Z} \times \mathcal{Y} \times \mathbb{R} \mapsto [0,\infty)$ be a convex Lipschitz continuous loss function such that $L(z,y,0)$ is uniformly bounded. Let H be a separable RKHS of a bounded measurable kernel on the set $\mathcal{Z} \subset \mathbb{R}^d$. Choose $0 < \lambda_n < 1$ such that $\lambda_n \to 0$, and $\lambda_n^2 n \to \infty$. Then*

(a) The empirical SVM decision function f_{D_0,λ_n} exists and is unique.
(b) The SVM learning method $\mathfrak{L}$ defined in (2) is measurable.
(c) The L-risk $\mathcal{R}_{L,P}(f_{D_0,\lambda_n}) \xrightarrow{\mathrm{P}} \inf_{f\in H} \mathcal{R}_{L,P}(f)$.
(d) If the RKHS H is dense in the set of integrable functions on $\mathcal{Z}$, then the SVM learning method $\mathfrak{L}$ is universally consistent.

The proof of (a) follows from SC08, Lemma 5.1 and Theorem 5.2. For the proof of (b), see SC08, Lemma 6.23. The proof of (c) follows from SC08 Theorem 6.24. The proof of (d) follows from SC08, Theorem 5.31, together with Theorem 6.24.

3. SVM for survival data without censoring

In this section we present a few examples of the use of SVM for survival data but without censoring. We show how different quantities obtained from the conditional distribution of T given Z can be represented as Bayes decision functions. We then show how SVM learning methods can be applied to these estimation problems and briefly review theoretical properties of such SVM learning methods. In the next section we will explain why these standard SVM techniques cannot be employed directly when censoring is present.

Let (Z,T) be a random vector where Z is a covariate vector that takes its values in a set $\mathcal{Z} \subset \mathbb{R}^d$, T is survival time that takes it values in $\mathcal{T} = [0,\tau]$ for some positive constant τ, and where (Z,T) is distributed according to a probability measure P on $\mathcal{Z} \times \mathcal{T}$.

Note that the conditional expectation $P[T|Z]$ is the Bayes decision function for the least squares loss function L_{LS}. In other words $P[T|Z = z]$ minimizes $P[L_{\mathrm{LS}}(Z,T,\cdot)|Z = z]$ P-almost surely, (see Example 4). Similarly, the conditional median and the α-quantile of $T|Z$ can be shown to be the Bayes decision functions for the absolute deviation function L_{AD} and L_α, respectively (see

Example 5). In the same manner, one can represent other quantities of the conditional distribution $T|Z$ using Bayes decision functions.

Defining quantities computed from the survival function as Bayes decision functions is not limited to regression (i.e., to a continuous response). Classification problems can also arise in the analysis of survival data (see, for example, Ripley and Ripley, 2001; Johnson et al., 2004). For example, let ρ, $0 < \rho < \tau$, be a cutoff constant. Assume that survival to a time greater than ρ is considered as death unrelated to the disease (i.e., remission) and a survival time less than or equal to ρ is considered as death resulting from the disease. Denote

$$Y(T) = \begin{cases} 1 & T > \rho\,, \\ -1 & T \leq \rho\,. \end{cases} \tag{4}$$

In this case, the decision function that predicts remission when the probability of $Y = 1$ given the covariates is greater than $1/2$ and failure otherwise is a Bayes decision function for the binary classification loss L_{BC} of Example 3.

Let $D_0 = \{(Z_1, T_1), \ldots, (Z_n, T_n)\}$ be a data set of n i.i.d. observations distributed according to P. Let $Y_i = Y(T_i)$ where $Y(\cdot) : \mathcal{T} \mapsto \mathcal{Y}$ is some deterministic measurable function. For regression problems, Y is typically the identity function and for classification Y can be defined, for example, as in (4). Let L be a convex locally Lipschitz continuous loss function, $L : \mathcal{Z} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$. Note that this includes the loss functions $L_{\mathrm{LS}}, L_{\mathrm{AD}}$, L_α, and L_{HL}. Define the empirical decision function as in (1) and the SVM learning method $\mathfrak{L}$ as in (2). Then it follows from Theorem 1 that for an appropriate RKHS H and regularization sequence $\{\lambda_n\}$, $\mathfrak{L}$ is measurable and universally consistent.

4. Censored SVM

In the previous section, we presented a few examples of the use of SVM for survival data without censoring. In this section we explain why standard SVM techniques cannot be applied directly when censoring is present. We then explain how to use inverse probability of censoring weighting (Robins et al., 1994) to obtain a censored SVM learning method. Finally, we show that the obtained censored SVM learning method is well defined.

Let $D = \{(Z_1, U_1, \delta_1), \ldots, (Z_n, U_n, \delta_n)\}$ be a set of n i.i.d. random triplets of right censored data (as described in Section 2.1). Let $L : Z \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$ be a convex locally Lipschitz loss function. Let H be a separable RKHS of a bounded measurable kernel on $\mathcal{Z}$. We would like to find an empirical SVM decision function. In other words, we would like to find the minimizer of

$$\lambda \|f\|_H^2 + \mathcal{R}_{L,D}(f) \equiv \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^{n} L(Z_i, Y(T_i), f(Z_i)) \tag{5}$$

where $\lambda > 0$ is a fixed constant, and $Y : \mathcal{T} \mapsto \mathcal{Y}$ is a known function. The problem is that the failure times T_i may be censored, and thus unknown. While a simple solution is to ignore the censored observations, it is well known that this can lead to severe bias (Tsiatis, 2006).

In order to avoid this bias, one can reweight the uncensored observations. Note that at time T_i, the i-th observation has probability $G(T_i|Z_i) \equiv P(C_i \geq T_i|Z_i)$ not to be censored, and thus, one can use the inverse of the censoring probability for reweighting in (5) (Robins et al., 1994).

More specifically, define the random loss function $L^n : (\mathcal{Z} \times \mathcal{T} \times \{0,1\})^n \times (\mathcal{Z} \times \mathcal{T} \times \{0,1\} \times \mathbb{R}) \mapsto \mathbb{R}$ by

$$L^n(D,(z,u,\delta,s)) = \begin{cases} \frac{L(z,Y(u),s)}{\hat{G}_n(u|z)}, & \delta = 1, \\ 0, & \delta = 0, \end{cases}$$

where $\hat{G}_n$ is the estimator of the survival function of the censoring variable based on the set of n random triplets D (see Section 2.1). When D is given, we denote $L^n_D(\cdot) \equiv L^n(D,\cdot)$. Note that in this case the function L^n_D is no longer random. In order to show that L^n_D is a loss function, we need to show that L^n_D is a measurable function.

Lemma 2. *Let L be a convex locally Lipschitz loss function. Assume that the estimation procedure $D \mapsto \hat{G}_n(\cdot|\cdot)$ is measurable. Then for every $D \in (\mathcal{Z} \times \mathcal{T} \times \{0,1\})^n$ the function $L^n_D : (\mathcal{Z} \times \mathcal{T} \times \{0,1\}) \times \mathbb{R} \mapsto \mathbb{R}$ is measurable.*

Proof. By Remark 2, the function $\hat{G}_n(u|z) \mapsto 1/\hat{G}_n(u|z)$ is well defined. Since by definition, both Y and L are measurable, we obtain that $(u,z,\delta,s) \mapsto \delta L(z,Y(u),s)/\hat{G}_n(u|z)$ is measurable. □

We define the *empirical censored SVM decision function* to be

$$\begin{aligned} f^c_{D,\lambda} &= \operatorname*{argmin}_{f \in H} \lambda \|f\|^2_H + \mathcal{R}_{L^n_D,D}(f) \\ &\equiv \operatorname*{argmin}_{f \in H} \lambda \|f\|^2_H + \frac{1}{n} \sum L^n_D\big(Z_i,U_i,\delta_i,f(Z_i)\big) . \end{aligned} \tag{6}$$

The existence and uniqueness of the empirical censored SVM decision function is ensured by the following lemma:

Lemma 3. *Let L be a convex locally Lipschitz loss function. Let H be a separable RKHS of a bounded measurable kernel on $\mathcal{Z}$. Then there exists a unique empirical censored SVM decision function.*

Proof. Note that given D, the loss function $L^n_D(z,u,\delta,\cdot)$ is convex for every fixed z, u, and δ. Hence, the result follows from Lemma 5.1 together with Theorem 5.2 of SC08. □

Note that the empirical censored SVM decision function is just the empirical SVM decision function of (1), after replacing the loss function L with the loss function L^n_D. However, there are two important implications to this replacement. Firstly, empirical censored SVM decision functions are obtained by minimizing a different loss function for each given data set. Secondly, the second expression in the minimization problem (6), namely,

$$\mathcal{R}_{L_D^n,D}(f) \equiv \frac{1}{n}\sum_{i=1}^{n} L_D^n\big(Z_i, U_i, \delta_i, f(Z_i)\big),$$

is no longer constructed from a sum of i.i.d. random variables.

We would like to show that the learning method defined by the empirical censored SVM decision functions is indeed a learning method. We first define the term learning method for right censored data or *censored learning method* for short.

Definition 1. *A censored learning method $\mathfrak{L}^c$ on $\mathcal{Z} \times \mathcal{T}$ maps every data set $D \in (\mathcal{Z} \times \mathcal{T} \times \{0,1\})^n$, $n \geq 1$, to a function $f_D : \mathcal{Z} \mapsto \mathbb{R}$.*

Choose $0 < \lambda_n < 1$ such that $\lambda_n \to 0$. Define the *censored SVM learning method* $\mathfrak{L}^c$, as $\mathfrak{L}^c(D) = f^c_{D,\lambda_n}$ for all $n \geq 1$. The measurability of the censored SVM learning method $\mathfrak{L}^c$ is ensured by the following lemma, which is an adaptation of Lemma 6.23 of SC08 to the censored case.

Lemma 4. *Let L be a convex locally Lipschitz loss function. Let H be a separable RKHS of a bounded measurable kernel on $\mathcal{Z}$. Assume that the estimation procedure $D \mapsto \hat{G}_n(\cdot|\cdot)$ is measurable. Then the censored SVM learning method $\mathfrak{L}^c$ is measurable, and the map $D \mapsto f^c_{D,\lambda_n}$ is measurable.*

Proof. First, by Lemma 2.11 of SC08, for any $f \in H$, the map $(z, u, f) \mapsto L(z, Y(u), f(z))$ is measurable. The survival function $\hat{G}_n$ is measurable on $(\mathcal{Z} \times \mathbb{R} \times \{0,1\})^n \times (\mathcal{Z} \times \mathbb{R})$ and by Remark 2, the function $D \mapsto \delta_i/\hat{G}_n(u_i|z_i)$ is well defined and measurable. Hence $D \mapsto n^{-1}\sum_{i=1}^{n} \frac{\delta_i L(z_i, Y(u_i), f(z_i))}{\hat{G}_n(u_i|z_i)}$ is measurable. Note that the map $f \mapsto \lambda_n \|f\|_H^2$ where $f \in H$ is also measurable. Hence we obtain that the map $\phi : (\mathcal{Z} \times \mathcal{T} \times \{0,1\})^n \times H \mapsto \mathbb{R}$, defined by

$$\phi(D, f) = \lambda \|f\|_H^2 + \mathcal{R}_{L_D^n,D}(f)\,,$$

is measurable. By Lemma 3, f^c_{D,λ_n} is the only element of H satisfying

$$\phi(D, f^c_{D,\lambda_n}) = \inf_{f \in H} \phi(D, f)\,.$$

By Aumann's measurable selection principle (SC08, Lemma A.3.18), the map $D \mapsto f^c_{D,\lambda_n}$ is measurable with respect to the minimal completion of the product σ-field on $(\mathcal{Z}\times\mathcal{T}\times\{0,1\})^n$. Since the evaluation map $(f, z) \mapsto f(z)$ is measurable (SC08, Lemma 2.11), the map $(D, z) \mapsto f^c_{D,\lambda_n}(z)$ is also measurable. □

5. Theoretical results

In the following, we discuss some theoretical results regarding the censored SVM learning method proposed in Section 4. In Section 5.1 we discuss function clipping which will serve as a tool in our analysis. In Section 5.2 we discuss finite sample bounds. In Section 5.3 we discuss consistency. Learning rates are discussed in Section 5.4. Finally, censoring model misspecification is discussed in Section 5.5.

5.1. Clipped censored SVM learning method

In the following section we bound the risk of the censored SVM decision function. Since this function belongs to an RKHS, it need not be bounded. Nevertheless, for some loss functions, we may be able to improve the censored SVM decision function by clipping it to get values in a bounded set. Clipping was introduced by Bartlett (1998) in the context of neural networks. See also Bousquet and Elisseeff (2002), Wu et al. (2007), and Steinwart et al. (2007), among others, in the context of SVM.

Before defining clipping formally, we consider the following example. Let the loss L in the minimization problem (1) be the hinge-loss $L_{\mathrm{HL}}(z, y, s) = \max\{0, 1 - ys\}$, and let the response space be $\mathcal{Y} = \{-1, 1\}$. In this case, we can improve the risk of any function f that gets values outside the segment $[-1, 1]$. Indeed, let $\Omega_1 \equiv \{z : f(z) > 1\}$, and $\Omega_{-1} \equiv \{z : f(z) < -1\}$. Define the clipped function of f by

$$\widehat{f}(z) \equiv \begin{cases} 1 & z \in \Omega_1\,, \\ -1 & z \in \Omega_{-1}\,, \\ f(z) & \text{otherwise}\,. \end{cases}$$

Note that

$$\begin{aligned}\mathcal{R}_{L_{\mathrm{HL}},P}(\widehat{f}) \equiv E_P[L_{\mathrm{HL}}(Z, Y, \widehat{f}(Z))]\\ =&E_P[\max\{0, 1 - Yf(Z)\}|(\Omega_1 \cup \Omega_{-1})^c]P((\Omega_1 \cup \Omega_{-1})^c)\\ &+ E_P[\max\{0, 1 - Y\}|\Omega_1]P(\Omega_1) + E_P[\max\{0, 1 + Y\}|\Omega_{-1}]P(\Omega_{-1})\\ \leq&\mathcal{R}_{L_{\mathrm{HL}},P}(f)\,.\end{aligned}$$

Moreover, $\widehat{f}$ has strictly smaller risk than f whenever there is a subset of positive probability of $\Omega_1 \cup \Omega_{-1}$ in which $\mathrm{sign}\{f(Z)\} \neq Y$. We say that the loss function L_{HL} *can be clipped* at 1 since $L_{\mathrm{HL}}(z, y, 1) \leq L_{\mathrm{HL}}(z, y, s)$ for any $s > 1$ and $L_{\mathrm{HL}}(z, y, -1) \leq L_{\mathrm{HL}}(z, y, s)$ for any $s < -1$, and for all $(z, y) \in \mathcal{Z} \times \mathcal{Y}$. Since L_{HL} can be clipped, replacing any function f by its clipped version $\widehat{f}$ reduces the risk.

This example demonstrates that one may control the boundlessness of the censored SVM decision function by replacing the obtained function with its clipped version. In the following proofs, we consider the analysis of loss functions that can be clipped, which, in our setting, include many of the standard loss functions. Results for loss functions that cannot be clipped are beyond the scope of this paper, but they can be proved under more stringent assumptions. We refer those interested to `http://arxiv.org/pdf/1202.5130v1.pdf`.

More formally, we say that a loss function L can be clipped at $M > 0$, if, for all $(z, y, s) \in \mathcal{Z} \times \mathcal{Y} \times \mathbb{R}$,

$$L(z, y, \widehat{s}) \leq L(z, y, s)$$

where

$$\widehat{s} = \begin{cases} -M & \text{if } s \leq -M \\ s & \text{if } -M < s < M \\ M & \text{if } s \geq M \end{cases}$$

(see SC08, Definition 2.22). In other words, a loss function L can be clipped if, when clipping its last argument, the values of the loss function are lower (or equal) to the values when the last argument is not clipped. The loss functions L_{HL}, L_{LS}, L_{AD}, and L_α can be clipped at some M when $\mathcal{Y} = \mathcal{T}$ or $\mathcal{Y} = \{-1, 1\}$. The constant M can be computed explicitly (SC08, Chapter 2).

In our context the response variable Y usually takes it values in a bounded set (see Section 3). When the response space is bounded, we have the following criterion for clipping. Let L be a distance-based loss function, i.e., $L(z, y, s) = \phi(s - y)$ for some function ϕ. Assume that $\lim_{r \to \pm\infty} \phi(r) = \infty$. Then L can be clipped at some M (SC08, Chapter 2).

For a function f, we define $\widehat{f}$ to be the clipped version of f, i.e., $\widehat{f} = \max\{-M, \min\{M, f\}\}$. Finally, we note that the clipped censored SVM learning method, that maps every data set $D \in (\mathcal{Z} \times \mathcal{T} \times \{0, 1\})^n$, $n \geq 1$, to the function $\widehat{f}^c_{D,\lambda}$, is measurable, where $\widehat{f}^c_{D,\lambda}$ is the clipped version of $f^c_{D,\lambda}$ defined in (6). This follows from Lemma 4, together with the measurability of the clipping operator.

5.2. Finite sample bounds

We would like to establish a finite-sample bound for the generalization of clipped censored SVM learning methods. We first need some notation. Define the censoring estimation error

$$Err_n(t, z) = \hat{G}_n(t|z) - G(t|z)\,, \qquad (t, z) \in \mathcal{T} \times \mathcal{Z}$$

to be the difference between the estimated and true survival functions of the censoring variable.

Let H be an RKHS over the covariates space $\mathcal{Z} \subset \mathbb{R}^d$. Define the n-th dyadic entropy number $e_n(H, \|\cdot\|_H)$ as the infimum over ε, such that H can be covered with no more than 2^{n-1} balls of radius ε with respect to the metric induced by the norm. For a bounded linear transformation $S : H \mapsto F$ where F is a normed space, we define the dyadic entropy number $e_n(S)$ as $e_n(SB_H, \|\cdot\|_F)$ where B_H is the unit ball of H. For details, the reader is referred to Appendix 5.6 of SC08.

We need the following assumptions:

(B1) The loss function $L : \mathcal{Z} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$ is a locally Lipschitz continuous loss function that can be clipped at $M > 0$ such that the supremum bound

$$L(z, y, s) \leq B \tag{7}$$

holds for all $z, y, s \in \mathcal{Z} \times \mathcal{Y} \times [-M, M]$ and for some $B > 0$. Moreover, there is a constant $q > 0$ such that

$$|L(z, y, s) - L(z, y, 0)| \leq c|s|^q$$

for all $z, t, s \in \mathcal{Z} \times \mathcal{Y} \times \mathbb{R}$ and for some $c > 0$.

(B2) H is a separable RKHS of a measurable kernel over $\mathcal{Z}$ and P is a distribution over $\mathcal{Z} \times \mathcal{T}$ for which there exist constants $\vartheta \in [0,1]$ and $V > B^{2-\vartheta}$ such that

$$P\left[\left(L \circ \widehat{f} - L \circ f^*_{L,P}\right)^2\right] \leq V\left(P\left[L \circ \widehat{f} - L \circ f^*_{L,P}\right]\right)^{\vartheta} \tag{8}$$

for all $z, y, s \in \mathcal{Z} \times \mathcal{Y} \times [-M, M]$ and $f \in H$; and where $L \circ f$ is shorthand for the function $(z, y) \mapsto L(z, y, f(z))$.

(B3) There are constants $a > 1$ and $0 < p < 1$, such that for all $i \geq 1$ the following entropy bound holds:

$$P[e_i(\mathrm{id} : H \mapsto L_2(\mathbb{P}_n))] \leq a i^{-\frac{1}{2p}}\,, \tag{9}$$

where $\mathrm{id} : H \mapsto L_2(\mathbb{P}_n)$ is the embedding of H into the space of square integrable functions with respect to the empirical measure $\mathbb{P}_n$.

Before we state the main result of this section, we present some examples for which the assumptions above hold:

Remark 3. *When $\mathcal{Y}$ is contained in a compact set, Assumption (B1) holds with $q = 1$ for L_{HL}, L_{AD} and L_α and with $q = 2$ for L_{LS} (recall the definitions of the loss functions from Section 2.2).*

Remark 4. *Assumption (B2) holds trivially for $\vartheta = 0$ with $V = B^2$. It holds for L_{LS} with $\vartheta = 1$ for compact $\mathcal{Y}$ (SC08, Example 7.3). Under some conditions on the distribution, it also holds for L_{AD} and L_α (SC08, Eq. 9.29).*

Remark 5. *When $\mathcal{Z} \subset \mathbb{R}^d$ is compact, the entropy bound (9) of Assumption (B3) is satisfied for smooth kernels such as the polynomial and Gaussian kernels for all $p > 0$ (see SC08, Section 6.4). The assumption also holds for Gaussian kernels over $\mathbb{R}^d$ for distributions P_Z with positive tail exponent (see SC08, Section 7.5).*

We are now ready to establish a finite sample bound for the clipped censored SVM learning methods:

Theorem 5. *Let L be a loss function and H be an RKHS such that assumptions (B1)–(B3) hold. Let $f_0 \in H$ satisfies $\|L \circ f_0\|_\infty \leq B_0$ for some $B_0 \geq B$. Let $\hat{G}_n(t|Z)$ be an estimator of the survival function of the censoring variable and assume (A1)–(A2). Then, for any fixed regularization constant $\lambda > 0$, $n \geq 1$, and $\eta > 0$, with probability not less than $1 - 3e^{-\eta} - P_{\hat{G}_n,n}$,*

$$\begin{aligned}
&\lambda \|f^c_{D,\lambda}\|^2_H + \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P} \\
&\qquad \leq 5\lambda\|f_0\|^2_H + 8(\mathcal{R}_{L,P}(f_0) - \mathcal{R}^*_{L,P}) + \frac{3B_0}{K^2}\mathbb{P}_n|Err_n(T,Z)| + \frac{2B_0\eta}{Kn} \\
&\qquad\quad + 4\left(\frac{72\tilde{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + 3W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}\,.
\end{aligned}$$

where W is a constant that depends only p, M, B, ϑ, V and K, where $\tilde{V} \equiv \max\{V/2K, (B/(2K))^{2-\vartheta}\}$, and where $P_{\hat{G}_n,n}$ is defined in Remark 2.

The proof appears in Appendix A.3.

For the Kaplan-Meier estimator (see Example 1), bounds of the random error $\|Err_n\|_\infty$ were established (Bitouzé et al., 1999). In this case, we can replace the bound of Theorem 5 with a more explicit one.

Specifically, let $\hat{G}_n$ be the Kaplan-Meier estimator. Let $0 < K_S = P(T \geq \tau)$ be a lower bound on the survival function at τ. Then, for every $n \geq 1$ and $\varepsilon > 0$ the following Dvoretzky-Kiefer-Wolfowitz-type inequality holds (Bitouzé et al., 1999, Theorem 2):

$$P(\|\hat{G}_n - G\|_\infty > \varepsilon) < \frac{5}{2}\exp\{-2nK_S^2\varepsilon^2 + D_o\sqrt{n}K_S\varepsilon\}\,,$$

where D_o is some universal constant (see Wellner, 2007, for a bound on D_o). Some algebraic manipulations then yield that for every $\eta > 0$ and $n \geq 1$

$$P\left(\|\hat{G}_n - G\|_\infty > \frac{\sqrt{2\eta} + D_o}{K_S\sqrt{n}}\right) < \frac{5}{2}e^{-\eta}\,. \tag{10}$$

We also have

$$\begin{aligned} P_{\hat{G}_n,n} \equiv P(\inf_Z \hat{G}_n(\tau|Z) < K) &\leq P\left(\|\hat{G}_n - G\|_\infty > K\right) \\ &< \frac{5}{2}\exp\{-2nK_S^2K^2 + D_o\sqrt{n}K_SK\}\,. \end{aligned}$$

As a result, we obtain the following corollary:

Corollary 6. *Consider the setup of Theorem 5. Assume that the censoring variable C is independent of both T and Z. Let $\hat{G}_n$ be the Kaplan-Meier estimator of G. Then for any fixed regularization constant λ, $n \geq 1$, and $\eta > 0$, with probability not less than $1 - \frac{11}{2}e^{-\eta} - \frac{5}{2}\exp\{-2nK_S^2K^2 + D_o\sqrt{n}K_SK\}$,*

$$\begin{aligned} &\lambda\|f_{D,\lambda}^c\|_H^2 + \mathcal{R}_{L,P}(\widehat{f_{D,\lambda}^c}) - \mathcal{R}_{L,P}^* \\ &\quad\leq 5\lambda\|f_0\|_H^2 + 8(\mathcal{R}_{L,P}(f_0) - \mathcal{R}_{L,P}^*) + \frac{3B_0(\sqrt{2\eta} + D_o)}{K^2K_S\sqrt{n}} + \frac{2B_0\eta}{Kn} \\ &\qquad + 4\left(\frac{72\tilde{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + 3W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}\,. \end{aligned}$$

where W is a constant that depends only on p, M, B, ϑ, V and K.

5.3. $\mathcal{P}$-universal consistency

In this section we discuss consistency of the clipped version of the censored SVM learning method $\mathfrak{L}^c$ proposed in Section 4. In general, P-consistency means that (3) holds for all $\varepsilon > 0$. Universal consistency means that the learning method is P-consistent for every probability measure P on $\mathcal{Z} \times \mathcal{T} \times \{0, 1\}$. In the following we discuss a more restrictive notion than universal consistency, namely $\mathcal{P}$-universal consistency. Here, $\mathcal{P}$ is the set of all probability distributions for which conditions (A1)–(A2) hold for some constant K. We say that a censored learning method is $\mathcal{P}$-universally consistent if (3) holds for all $P \in \mathcal{P}$. We note that when the first assumption is violated for a set of covariates $\mathcal{Z}_0$ with positive probability, there is no hope of learning the optimal function for all $z \in \mathcal{Z}$, unless some strong assumptions on the model are enforced. The second assumption is required for proving consistency of the learning method $\mathfrak{L}^c$ proposed in Section 4. However, it is possible that other censored learning techniques will be able to achieve consistency for a larger set of probability measures.

In order to show $\mathcal{P}$-universal consistency, we utilize the bound given in Theorem 5. We need the following additional assumptions:

(B4) For all distributions P on $\mathcal{Z}$, $\inf_{f \in H} \mathcal{R}_{L,P}(f) = \mathcal{R}^*_{L,P}$.

(B5) $\hat{G}_n$ is consistent for G and there is a finite constant $s > 0$ such that $P(\|Err_n\|_\infty \geq bn^{-1/s}) \to 0$ for any $b > 0$.

Before we state the main result of this section, we present some examples for which the assumptions above hold:

Remark 6. *Assumption (B4) holds when the loss function L is locally Lipschitz continuous, $R_{L,P}(0) < \infty$, and the RKHS H is dense in $L_1(\mu)$ for all distribution μ on $\mathcal{Z}$, where $L_1(\mu)$ is the space of equivalence classes of integrable functions. (SC08, see Theorem 5.31).*

Remark 7. *Assume that $\mathcal{Z}$ is compact. A continuous kernel k whose corresponding RKHS H is dense in the class of continuous functions over the compact set $\mathcal{Z}$ is called universal. Examples of universal kernels include the Gaussian kernels, and other Taylor kernels. For more details, the reader is referred to SC08, Chapter 4.6. For universal kernels, Assumption (B4) holds for L_{LS}, L_{HL}, L_{AD}, and L_α. (SC08, Corollary 5.29).*

Remark 8. *Assume that $\hat{G}_n$ is consistent for G. When $\hat{G}_n$ is the Kaplan-Meier estimator, Assumption (B5) holds for all $s > 2$ (Bitouzé et al., 1999, Theorem 3). Similarly, when $\hat{G}_n$ is the proportional hazards estimator (see Example 2), under some conditions, Assumption (B5) holds for all $s > 2$ (see Goldberg and Kosorok, 2012a, Theorem 3.2 and its conditions).*

Now we are ready for the main result.

Theorem 7. *Let L be a loss function and H be an RKHS of a bounded kernel over $\mathcal{Z}$. Assume (A1)–(A2) and (B1)–(B5). Let $\lambda_n \to 0$, where $0 < \lambda_n < 1$, and $\lambda_n^{\max\{q/2,p\}} n \to \infty$, where q is defined in Assumption (B1). Then the clipped censored learning method $\mathfrak{L}^c$ is $\mathcal{P}$-universally consistent.*

Proof. Define the approximation error

$$A_2(\lambda) = \lambda \|f_{P,\lambda}\|^2_H + \mathcal{R}_{L,P}(f_{P,\lambda}) - \mathcal{R}^*_{L,P}, , \tag{11}$$

where $f_{P,\lambda} \equiv \operatorname{argmin}_{f\in H} \lambda\|f\|^2_H + \mathcal{R}_{L,P}(f)$. By Theorem 5, for $f_0 = f_{P,\lambda}$ we obtain

$$\begin{aligned}\lambda\|f^c_{D,\lambda}\|^2_H &+ \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P}\\ &\le 8A_2(\lambda) + \frac{3B_0}{K^2}\mathbb{P}_n|Err_n(T,Z)| + \frac{2B_0\eta}{Kn}\\ &\quad + 4\left(\frac{72\tilde{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + 3W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} .\end{aligned}\tag{12}$$

for any fixed regularization constant $\lambda > 0$, $n \ge 1$, and $\eta > 0$, with probability not less than $1 - 3e^{-\eta} - P_{\hat{G}_n,n}$.

By the definition of B_0, choosing $f_0 = f_{P,\lambda}$, it is required that $\|L \circ f_{P,\lambda}\|_\infty \le B_0$. Define $B_0 = B + c_o(A_2(\lambda)/\lambda)^{q/2}$ for $c_o = c(\sup_{z\in\mathcal{Z}}\sqrt{k(z,z)})^{q/2}$ and where c and q are defined in Assumption (B1). We now show that indeed $\|L\circ f_{P,\lambda}\|_\infty \le B_0$. Since the kernel k is bounded, it follows from Lemma 4.23 of SC08 that $\|f_{P,\lambda}\|_\infty \le \sup_{z\in\mathcal{Z}}\sqrt{k(z,z)}\|f_{P,\lambda}\|_H$. By the definition of $A_2(\lambda)$, $\|f_{P,\lambda}\|_H \le (A_2(\lambda)/\lambda)^{1/2}$. Note that for all $(z, y) \in \mathcal{Z}\times\mathcal{Y}$

$$L(z, y, f_{P,\lambda}(z)) \le L(x, y, 0) + |L(z, y, f_{P,\lambda}(z)) - L(x, y, 0)| \le B + c|f_{P,\lambda}(z)|^q .$$

Thus

$$\begin{aligned}\|L\circ f_{P,\lambda}\|_\infty \le B + c\|f_{P,\lambda}\|^q_\infty &\le B + c(\sup_{z\in\mathcal{Z}}\sqrt{k(z,z)}\|f_{P,\lambda}\|_H)^q\\ &\le B + c_o\left(\frac{A_2(\lambda)}{\lambda}\right)^{\frac{q}{2}} = B_0 .\end{aligned}\tag{13}$$

Taking $\lambda = \lambda_n$, it follows from Assumption (B4), together with Lemma 5.15 of SC08 that $A_2(\lambda_n)$ converges to zero as n converges to infinity. Clearly $4\left(\frac{72\tilde{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}}$ converges to zero. We also have that $\frac{2\eta}{Kn}\left\{B + c_o\left(\frac{A_2(\lambda_n)}{\lambda_n}\right)^{q/2}\right\}$ converges to zero since $\lambda_n^{q/2} n \to \infty$. By Assumption (B5), both $\mathbb{P}_n Err_n$ and $P_{\hat{G}_n,n}$ converge to zero. Finally, $W\left(\frac{a^{2p}}{\lambda_n^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}$ converges to zero since $\lambda_n^p n \to \infty$. Hence, for every fixed η, the right hand side of (12) converges to zero, which implies (3). Since (3) holds for every $P \in \mathcal{P}$, we obtain $\mathcal{P}$-universal consistency. □

5.4. Learning rates

In the previous section we discussed $\mathcal{P}$-universal consistency which ensures that for every probability $P \in \mathcal{P}$, the clipped learning method $\mathfrak{L}^c$ asymptotically learns the optimal function. In this section we would like to study learning rates.

We define learning rates for censored learning methods similarly to the definition for regular learning methods (see SC08, Definition 6.5):

Definition 2. *Let $L : \mathcal{Z} \times \mathcal{Y} \times \mathbb{R} \mapsto [0, \infty)$ be a loss function. Let $P \in \mathcal{P}$ be a distribution. We say that a censored learning method $\mathfrak{L}^c$ learns with a rate $\{\varepsilon_n\}_n$, where $\{\varepsilon_n\} \subset (0, 1]$ is a sequence decreasing to 0, if for some constant $c_P > 0$, all $n \geq 1$, and all $\eta \in [0, \infty)$, there exists a constant $c_\eta \in [1, \infty)$ that depends on η, such that*

$$P(D \in (\mathcal{Z} \times \mathcal{T} \times \{0,1\})^n \, : \, \mathcal{R}_{L,P}(f^c_{D,\lambda}) \leq \mathcal{R}^*_{L,P} + c_P c_\eta \varepsilon_n) \geq 1 - e^{-\eta} \, .$$

In order to study the learning rates, we need an additional assumption:

(B6) There exist constants c_1 and $\beta \in (0, 1]$ such that $A_2(\lambda) \leq c_1 \lambda^\beta$ for all $\lambda \geq 0$, where A_2 is the approximation error function defined in (11).

Let $\mathcal{P}_{K_0}$ be the set of all probability distributions for which (A2) holds, and for which condition (A1) holds for some constant $K \geq K_0$.

Lemma 8. *Let L be a loss function and H be an RKHS of a bounded kernel over $\mathcal{Z}$. Assume (A1)–(A2) and (B1)–(B6). Then for every $K > 0$, the learning rate of the clipped $\mathfrak{L}^c$ for all $P \in \mathcal{P}_K$ is given by*

$$n^{-\min\left\{\frac{\beta}{(2-p-\vartheta+\vartheta p)\beta+p}, \frac{2\beta\tilde{s}}{q+(2-q)\beta}\right\}} \, ,$$

where q, ϑ, p, s, and β, are as defined in Assumptions (B1), (B2), (B3), (B5), and (B6), respectively, and where $\tilde{s} = \min\{1, 1/s\}$.

Before we provide the proof, we derive learning rates for two specific examples.

Example 6. ***Fast Rate:*** *Assume that the censoring mechanism is known, the loss function is the square loss, the kernel is Gaussian, $\mathcal{Z}$ is compact, $\mathcal{Y}$ is bounded. It follows that (B1) holds for $q = 2$, (B2) holds for $\vartheta = 1$ (SC08, Example 7.3), (B3) holds for all $0 < p < 1$ (SC08, Theorem 6.27), and (B5) holds for all $s > 0$. Thus the obtained rate is $n^{-\beta+\varepsilon}$, where $\varepsilon > 0$ is an arbitrarily small number.*

Example 7. ***Standard Rate:*** *Assume that the censoring mechanism follows the proportional hazards assumption, the loss function is either L_{HL}, L_{AD} or L_α, the kernel is Gaussian, $\mathcal{Z}$ is compact. It follows that (B1) holds for $q = 1$, (B2) holds trivially for $\vartheta = 0$, (B3) holds for all $0 < p < 1$, and (B5) holds for all $s > 2$. Thus the obtained rate is $n^{-\beta/(1+\beta)+\varepsilon}$, where $\varepsilon > 0$ is an arbitrarily small number. To see that note that*

$$\beta \leq 1 \Rightarrow (1-p)\beta \leq 1 - p \Rightarrow 2\beta - p\beta + p < 1 + \beta \Rightarrow \frac{1}{2\beta - p\beta + p} \geq \frac{1}{\beta + 1}$$
$$\Rightarrow \frac{\beta}{(2 - p - \vartheta + \vartheta p)\beta + p} \equiv \frac{\beta}{2\beta - p\beta + p} \geq \frac{\beta}{\beta + 1} - \varepsilon \equiv \frac{2\beta(\frac{1}{2} - \tilde{\varepsilon})}{q + (2-q)\beta} \, .$$

Proof of Lemma 8. Using the assumptions above, we replace the bound on $\lambda \|f^c_{D,\lambda}\|^2_H + \mathcal{R}_{L,P}(\wideparen{f^c_{D,\lambda}}) - \mathcal{R}^*_{L,P}$ that appears in (12) with quantities that depend on n, λ, η and some constants $c_1, \dots, c_4$ that can depend on p, M, ϑ, c_1, V, and K but not on P or η and constants $\tilde{c}_1, \tilde{c}_2, \tilde{c}_3$ that depend only on η and K.

Note that by Assumption (B6), $A_2(\lambda) \leq c_1 \lambda^{\beta}$ for some constant c_1. Using this fact, and the definition of B_0 in (13),

$$\begin{aligned}\left(\frac{2\eta}{Kn} + \frac{3\mathbb{P}_n Err_n(T,Z)}{K^2}\right) B_0 &= \left(\frac{2\eta}{Kn} + \frac{3\mathbb{P}_n Err_n(T,Z)}{K^2}\right)\left(B + c_o\left(\frac{A_2(\lambda)}{\lambda}\right)^{q/2}\right) \\ &\leq \left(\frac{2\eta}{Kn} + \frac{3\mathbb{P}_n Err_n(T,Z)}{K^2}\right)\left(B + c_o c_1 \lambda^{(\beta-1)q/2}\right).\end{aligned}$$

Note that $3W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} = c_2 n^{-\frac{1}{2-p-\vartheta+\vartheta p}} \lambda^{-\frac{p}{2-p-\vartheta+\vartheta p}}$ for some constant c_2.

By Assumption (B5) and the fact that $\|Err_n\|_\infty < 1$, there exists a constant $\tilde{c}_1 = c(\eta)$ that depends only on η, such that for all $n \geq 1$,

$$P(\|Err_n\|_\infty > \tilde{c}_1 n^{-1/s}) < e^{-\eta}. \tag{14}$$

Hence, when $\|Err_n\|_\infty > \tilde{c}_1 n^{-1/s}$ we have

$$\begin{aligned}\left(\frac{2\eta}{Kn} + \frac{3\mathbb{P}_n Err_n(T,Z)}{K^2}\right) B_0 &< \frac{c_4}{2}(\tilde{c}_2 n^{-1} + \tilde{c}_1 n^{-1/s})\left(1 + \lambda^{(\beta-1)q/2}\right) \\ &\leq c_4 \tilde{c}_3 n^{-\tilde{s}}\left(1 + \lambda^{(\beta-1)q/2}\right).\end{aligned}$$

where $\tilde{s} = \min\{1, 1/s\}$, for some constant $\tilde{c}_3 \geq 1$. Hence, with probability not less than $1 - 4e^{-\eta} - P_{\hat{G}_n,n}$, we have

$$\begin{aligned}\mathcal{R}_{L,P}(f^c_{D,\lambda}) - \mathcal{R}^*_{L,P} \leq c_5 \tilde{c}_3 &\left(\lambda^{\beta} + n^{-\frac{1}{2-p-\vartheta+\vartheta p}} \lambda^{-\frac{p}{2-p-\vartheta+\vartheta p}} + n^{-\tilde{s}} \lambda^{q(\beta-1)/2}\right) \\ &+ c_5 \tilde{c}_3 n^{-\tilde{s}} + 4\left(\frac{72\tilde{V}\eta}{n}\right)^{1/(2-\vartheta)}.\end{aligned}$$

Write $\lambda = n^{-\rho/\beta}$ for some $\rho > 0$ and note that

$$\begin{aligned}&\lambda^{\beta} + n^{-\frac{1}{2-p-\vartheta+\vartheta p}} \lambda^{-\frac{p}{2-p-\vartheta+\vartheta p}} + n^{-\tilde{s}} \lambda^{q(\beta-1)/2} \\ &\quad = n^{-\rho} + n^{-\frac{1}{2-p-\vartheta+\vartheta p}} (n^{-\rho/\beta})^{-\frac{p}{2-p-\vartheta+\vartheta p}} + n^{-\tilde{s}} (n^{-\rho/\beta})^{q(\beta-1)/2} \\ &\quad = n^{-\rho} + n^{-\frac{\beta - p\rho}{\beta(2-p-\vartheta+\vartheta p)}} + n^{-\frac{\beta\tilde{s} + \rho q(\beta-1)/2}{\beta}} \leq 3n^{-\min\{\rho, \frac{\beta-p\rho}{\beta(2-p-\vartheta+\vartheta p)}, \frac{2\beta\tilde{s}+\rho q(\beta-1)}{2\beta}\}}.\end{aligned}$$

Choosing

$$\rho = \min\left\{\frac{\beta}{(2-p-\vartheta+\vartheta p)\beta + p}, \frac{2\beta\tilde{s}}{q + (2-q)\beta}\right\},$$

we obtain

$$\lambda^{\beta} + n^{-\frac{1}{2-p-\vartheta+\vartheta p}} \lambda^{-\frac{p}{2-p-\vartheta+\vartheta p}} + n^{-\tilde{s}} \lambda^{q(\beta-1)/2} \leq 3n^{-\rho}$$

Using the fact that $0 < \beta \leq 1$ and $q > 0$, one can show that $\tilde{s} \geq \frac{2\beta\tilde{s}}{q+(2-q)\beta}$. Similarly, using the fact that $\vartheta \in [0,1]$ and $0 < p < 1$, one can show that $\frac{1}{2-\vartheta} \geq \frac{\beta}{(2-p-\vartheta+\vartheta p)\beta+p}$. Hence, $n^{-\min\{\tilde{s},\frac{1}{2-\vartheta}\}} \leq n^{-\rho}$. By the bound in (14), we also have that for $n > (\frac{\tilde{c}_1}{K})^s$,

$$P_{\hat{G}_n,n} \equiv P(\inf_Z \hat{G}_n(\tau|Z) < K) \leq P(\|Err_n\|_\infty > K)$$
$$\leq P(\|Err_n\|_\infty > \tilde{c}_1 n^{-1/s}) < e^{-\eta}\,.$$

It then follows that for all $n > (\frac{\tilde{c}_1}{K})^s$,

$$P\left(\mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \inf_{f\in H}\mathcal{R}_{L,P}(f) \leq c_P c_\eta n^{-\rho}\right) \geq 1 - 5e^{-\eta}\,,$$

for some constants $c_P \geq 1$ that depends on p, M, ϑ, c, B, V, and K but is independent of η, and c_η that depends only on η and K but not on P.

Let n_η be the smallest integer greater than $(\frac{\tilde{c}_1}{K})^s$. Let $\tilde{c}_\eta = \max\{Bc_\eta n^\rho_\eta, c_\eta\}$. Hence, for every $n \leq n_\eta$

$$c_P\tilde{c}_\eta n^{-\rho} \geq \tilde{c}_\eta n^{-\rho} \geq Bc_\eta n^\rho_\eta n^{-\rho} \geq B\,,$$

where the first inequality follows since $c_P \geq 1$ and the second from the definition of $\tilde{c}_\eta$. Since $\mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) \leq B$, and $\tilde{c}_\eta \geq c_\eta$, we have that for all $n \geq 1$,

$$P\left(\mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \inf_{f\in H}\mathcal{R}_{L,P}(f) \leq c_P\tilde{c}_\eta n^{-\rho}\right) \geq 1 - 5e^{-\eta}\,,$$

which concludes the proof. □

5.5. Misspecified censoring model

In Section 5.3 we showed that under conditions (B1)–(B5) the clipped censored SVM learning method $\mathfrak{L}^c$ is $\mathcal{P}$-universally consistent. While one can choose the Hilbert space H and the loss function L in advance such that conditions (B1)–(B4) hold, condition (B5) need not hold when the censoring mechanism is misspecified. In the following, we consider this case.

Let $\hat{G}_n(t|z)$ be the estimator of the survival function for the censoring variable. The deviation of $\hat{G}_n(t|z)$ from the true survival function $G(t|z)$ can be divided into two terms. The first term is the deviation of the estimator $\hat{G}_n(t|z)$ from its limit, while the second term is the difference between the estimator limit and the true survival function. More formally, let $G_P(t|z)$ be the limit of the estimator under the probability measure P, and assume it exists. Define the errors $Err_n(t,z)$ as

$$Err_{n1}(t,z) + Err_2(t,z) \equiv \left(\hat{G}_n(t|z) - G_P(t|z)\right) + \left(G_P(t|z) - G(t|z)\right).$$

Note that Err_{n1} is a random function that depends on the data, the estimation procedure, and the probability measure P, while Err_2 is a fixed function that depends only on the estimation procedure and the probability measure P. When the model is correctly specified, and the estimator is consistent, the second term vanishes.

Theorem 9. *Let L be a loss function and H be an RKHS of a bounded kernel over $\mathcal{Z}$. Assume (A1)–(A2) and (B1)–(B4). Let $\lambda_n \to 0$, where $0 < \lambda_n < 1$ and $\lambda_n^{\max\{q/2,p\}} n \to \infty$. Assume that*

$$P(\|\hat{G}_n - G_P\|_\infty \geq bn^{-1/s}) \to 0 \tag{15}$$

Then, for every fixed $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(D : \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) \leq \mathcal{R}^*_{L,P} + \frac{3B}{K^2}\,|P(G_P - G)| + \varepsilon\right) = 1\,,$$

where $D \in (\mathcal{Z} \times \mathcal{T} \times \{0,1\})^n$

Proof. By (12), for every fixed $\eta > 0$ and $n \geq 1$,

$$\begin{aligned} &\lambda\|f^c_{D,\lambda}\|^2_H + \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P} \\ &\qquad\leq 8A_2(\lambda_n) + \frac{3B_0}{K^2}\|\hat{G}_n - G_P\|_\infty + \frac{2B_0\eta}{Kn} + 4\left(\frac{72\tilde{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} \\ &\qquad\quad + 3W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{3B_0}{K^2}\mathbb{P}_n Err_2\,, \end{aligned} \tag{16}$$

for any fixed regularization constant $\lambda > 0$, $n \geq 1$, and $\eta > 0$, with probability not less than $1 - 3e^{-\eta} - P_{\hat{G}_n,n}$. Since $P(\|\hat{G}_n - G_P\|_\infty \geq bn^{-1/s}) \to 0$, it follows from the same arguments as in the proof of Theorem 7, that the first expression on the RHS of (16) converges in probability to zero. By the law of large numbers, $\mathbb{P}_n Err_2 \overset{\text{a.s.}}{\to} P(G_P - G)$. Note that by (13),

$$\frac{3B_0}{K^2}\mathbb{P}_n Err_2 \equiv \frac{3B + 3c_o\left(\frac{A_2(\lambda)}{\lambda}\right)^{\frac{q}{2}}}{K^2}\mathbb{P}_n Err_2 \overset{\text{a.s.}}{\to} \frac{3B}{K^2}P(G_P - G)\,,$$

since $A_2(\lambda_n)$ converges to zero as n converges to infinity (see the proof of Theorem 7), and the result follows. □

Theorem 9 proves that even under misspecification of the censored data model, the clipped censored learning method $\mathfrak{L}^c$ achieves the optimal risk up to a constant that depends on $P(G_P - G)$, which is the expected distance of the limit of the estimator from the true distribution. If the estimator estimates reasonably well, one can hope that this term is small, even under misspecification.

We now show that the additional condition (15) of Theorem 9 holds for both the Kaplan-Meier estimator and the Cox model estimator.

Example 8. ***Kaplan-Meier estimator:*** *Let $\hat{G}_n$ be the Kaplan-Meier estimator of G. Let G_P be the limit of $\hat{G}_n$. Note that G_P is the marginal distribution of the censoring variable. It follows from* (10) *that condition* (15) *holds for all $s > 2$.*

Example 9. ***Cox model estimator:*** *Let $\hat{G}_n$ be the estimator of G when the Cox model is assumed (see Example 2). Let G_P be the limit of $\hat{G}_n$. It has been shown that the limit G_P exists, regardless of the correctness of the proportional hazards model (Goldberg and Kosorok, 2012a). Moreover, for all $\varepsilon > 0$, and all n large enough,*

$$P(\|\hat{G}_n - G_P\|_\infty > \varepsilon) \leq \exp\{-W_1^2 n\varepsilon^2 + W_1 W_2 \sqrt{n}\varepsilon\},$$

where W_1, W_2 are universal constants that depend on the set $\mathcal{Z}$, the variance of Z, the constants K and K_S, but otherwise do not depend on the distribution P (see Goldberg and Kosorok, 2012a, Theorem 3.2, and conditions therein). Fix $\eta > 0$ and write

$$\varepsilon = \frac{\sqrt{4\eta + W_2^2} + W_2}{2W_1\sqrt{n}}.$$

Then for all n large enough we have

$$\begin{aligned}
&\varepsilon = \frac{\sqrt{4\eta + W_2^2} + W_2}{2W_1\sqrt{n}} \Leftrightarrow 2\varepsilon W_1\sqrt{n} - W_2 = \sqrt{4\eta + W_2^2} \\
\Leftrightarrow& 4\varepsilon^2 W_1^2 n - 4\varepsilon W_1 W_2 \sqrt{n} + W_2^2 = 4\eta + W_2^2 \\
\Leftrightarrow& 4\varepsilon^2 W_1^2 n - 4W_1 W_2 \varepsilon\sqrt{n} = 4\eta \Leftrightarrow \eta = W_1^2 n\varepsilon^2 - W_1 W_2 \sqrt{n}\varepsilon.
\end{aligned}$$

Using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ we have that

$$\varepsilon = \frac{\sqrt{4\eta + W_2^2} + W_2}{2W_1\sqrt{n}} \leq \frac{\sqrt{4\eta} + \sqrt{W_2^2} + W_2}{2W_1\sqrt{n}} = \frac{\sqrt{\eta} + W_2}{W_1\sqrt{n}}.$$

Hence

$$\limsup_{n\to\infty} P\left(\|\hat{G}_n - G_P\|_\infty > \frac{\sqrt{\eta} + W_2}{W_1\sqrt{n}}\right) < e^{-\eta}.$$

Consequently, condition (15) *holds for all $s > 2$.*

6. Simulation study

In this section we illustrate the use of the censored SVM learning method proposed in Section 4 via a simulation study. We consider five different data-generating mechanisms, including one-dimensional and multidimensional settings, and different types of censoring mechanisms. We compute the censored SVM decision function with respect to the absolute deviation loss function L_{AD}.

For this loss function, the Bayes risk is given by the conditional median (see Example 5). We choose to compute the conditional median and not the conditional mean, since censoring prevents reliable estimation of the unrestricted mean survival time when no further assumptions on the tail of the distribution are made (see discussions in Karrison, 1997; Zucker, 1998; Chen and Tsiatis, 2001). We compare the results of the SVM approach to the results obtained by the Cox model and to the Bayes risk. We test the effects of ignoring the censored observations. Finally, for multidimensional examples, we also check the benefit of variable selection.

The algorithm presented in Section 4 was implemented in the Matlab environment. For the implementation we used the Spider library for Matlab[1]. A link to the Matlab code for both the algorithm and the simulations can be found in Supplementary Material. The distribution of the censoring variable was estimated using the Kaplan-Meier estimator (see Example 1). We used the Gaussian RBF kernel $k_\sigma(x_1, x_2) = \exp(\sigma^{-2}\|x_1 - x_2\|_2^2)$, where the width of the kernel σ was chosen using cross-validation. Instead of minimizing the regularized problem (6), we solve the equivalent problem (see SC08, Chapter 5):

$$\text{Minimize } \mathcal{R}_{L_D^n, D}(f) \text{ under the constraint } \|f\|_H^2 < \lambda^{-1} ,$$

where H is the RKHS with respect to the kernel k_σ, and λ is some constant chosen using cross-validation. Note that there is no need to compute the norm of the function f in the RKHS space H explicitly. The norm can be obtained using the kernel matrix K with coefficients $k_{ij} = k(Z_i, Z_j)$ (see SC08, Chapter 11). The risk of the estimated functions was computed numerically, using a randomly generated data set of size 10000.

In some simulations the failure time is distributed according to the Weibull distribution (Lawless, 2003). The density of the Weibull distribution is given by

$$f(t) = \frac{\kappa}{\rho}\left(\frac{t}{\rho}\right)^{\kappa-1} e^{-(t/\rho)^\kappa}\mathbf{1}\{t \geq 0\} ,$$

where $\kappa > 0$ is the shape parameter and $\rho > 0$ is the scale parameter. Assume that κ is fixed and that $\rho = \exp(\beta_0 + \beta' Z)$, where β_0 is a constant, β is the coefficient vector, and Z is the covariate vector. In this case, the failure time distribution follows the proportional hazards assumption, i.e., the hazard rate is given by $h(t|Z) = \exp(\beta_0 + \beta' Z) d\Lambda(t)$, where $\Lambda(t) = t^\kappa$. When the proportional hazards assumption holds, estimation based on Cox regression is consistent and efficient (see Example 2; note that the distribution discussed there is of the censoring variable and not of the failure time, nevertheless, the estimation procedure is similar). Thus, when the failure time distribution follows the proportional hazards assumption, we use the Cox regression as a benchmark.

In the first setting, the covariates Z are generated uniformly on the segment $[-1, 1]$. The failure time follows the Weibull distribution with shape parameter

[1] The Spider library for Matlab can be downloaded form `http://www.kyb.tuebingen.mpg.de/bs/people/spider/`.

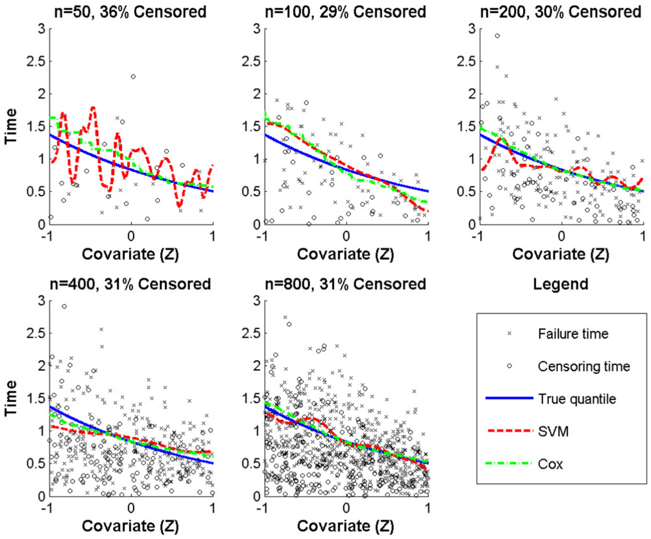

Fig 1. *Weibull failure time, proportional hazards (Setting 1): The true conditional median (solid blue), the SVM decision function (dashed red), and the Cox regression median (dot-dashed green) are plotted for samples of size $n = 50, 100, 200, 400$ and 800. The censoring percentage is given for each sample size. An observed failure times is represented by an* ×, *and an observed censoring time is represented by an* $\circ$.

2 and scale parameter $-0.5Z$. Note that the proportional hazards assumption holds. The censoring variable C is distributed uniformly on the segment $[0, c_0]$ where the constant c_0 is chosen such that the mean censoring percentage is 30%. We used 5-fold-cross-validation to choose the kernel width and the regularization constant among the set of pairs

$$(\lambda^{-1}, \sigma) = (0.1 \cdot 10^i, 0.05 \cdot 2^j), \qquad i, j \in \{0, 1, 2, 3\}.$$

In practice, choosing a grid of values for the cross-validation procedure can be done by first choosing a coarse grid and then choosing a finer grid at the vicinity of points that are of interest (see also Chapelle et al., 2002). We repeated the simulation 100 times for each of the sample sizes $50, 100, 200, 400$, and 800.

In Figure 1, the conditional median obtained by the censored SVM learning method and by Cox regression are plotted. The true median is plotted as a reference. In Figure 2, we compare the risk of the SVM method to the median of the survival function obtained by Cox regression (to which we refer as the Cox regression median). We also examined the effect of ignoring the censored observations by computing the standard SVM decision function for the data set in which all the censored observations were deleted. Finally, we examined the effect of model misspecification of the censoring mechanism. For that we draw C from a Weibull distribution with shape parameter 2 and scale parameter $-0.5Z + \log(1.5)$ to ensure 30% censoring. Both figures show that even though the SVM does not use the proportional hazards assumption for estimation,

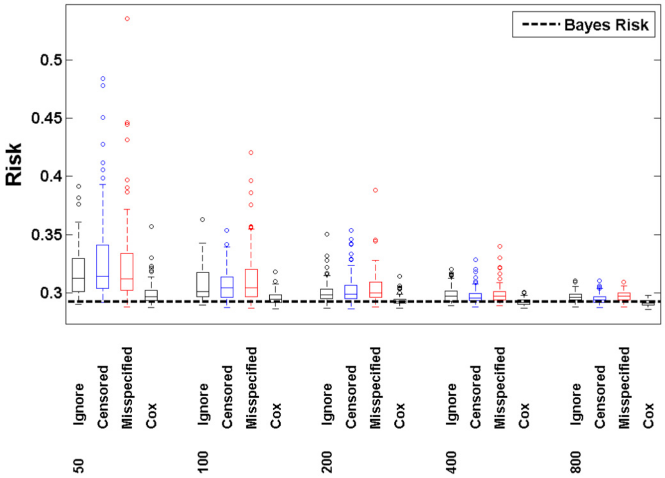

Fig 2. *Weibull failure time, proportional hazards (Setting 1): Distribution of the risk for different sizes of data set, for standard SVM that ignores the censored observations (Ignore), for censored SVM (Censored), for misspecified censoring mechanism (Misspecified), and for the Cox regression median (Cox). Bayes risk is denoted by a black dashed line. Each box plot is based on 100 repetitions of the simulation for each size of data set.*

the results are comparable to those of Cox regression for larger sample sizes. Figure 2 also shows that there is a non-negligible price for ignoring the censored observations and for misspecification.

The second setting differs from the first setting only in the failure time distribution. In the second setting the failure time distribution follows the Weibull distribution with scale parameter $-0.5Z^2$. Note that the proportional hazards assumption holds for Z^2, but not for the original covariate Z. In Figure 3, the true, the SVM median, and the Cox regression median are plotted. In Figure 4, we compare the risk of SVM to that of Cox regression. Both figures show that in this case SVM does better than Cox regression. Figure 4 also shows the price of ignoring censored observations and of misspecifying the censoring model.

The third and forth settings are generalizations of the first two, respectively, to 10-dimensional covariates. The covariates Z are generated uniformly on $[-1,1]^{10}$. The failure time follows the Weibull distribution with shape parameter 2. The scale parameter of the third and forth settings are $-0.5Z_1+2Z_2-Z_3$ and $-0.5(Z_1)^2+2(Z_2)^2-(Z_3)^2$, respectively. Note that these models are sparse, namely, they depend only on the first three variables. The censoring variable C is distributed uniformly on the segment $[0,c_0]$, where the constant c_0 is chosen such that the mean censoring percentage is 40%. We used 5-fold-cross-validation to choose the kernel width and the regularization constant among the set of pairs

$$(\lambda^{-1},\sigma) = (0.1\cdot 10^i, 0.2\cdot 2^j)\,, \qquad i,j\in\{0,1,2,3\}\,.$$

The results for the third and the forth settings appears in Figure 5 and Figure 6, respectively. We compare the risk of standard SVM that ignores cen-

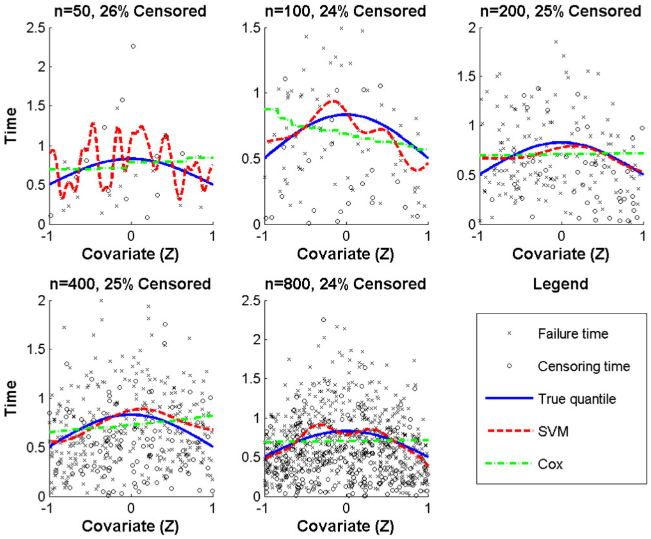

Fig 3. *Weibull failure time, non-linear proportional hazards (Setting 2): The true conditional median (solid blue), the SVM decision function (dashed red), and the Cox regression median (dot-dashed green) are plotted for samples of size* $n = 50, 100, 200, 400$ *and* 800. *The censoring percentage is given for each sample size. An observed failure times is represented by an* ×, *and an observed censoring time is represented by an* ○.

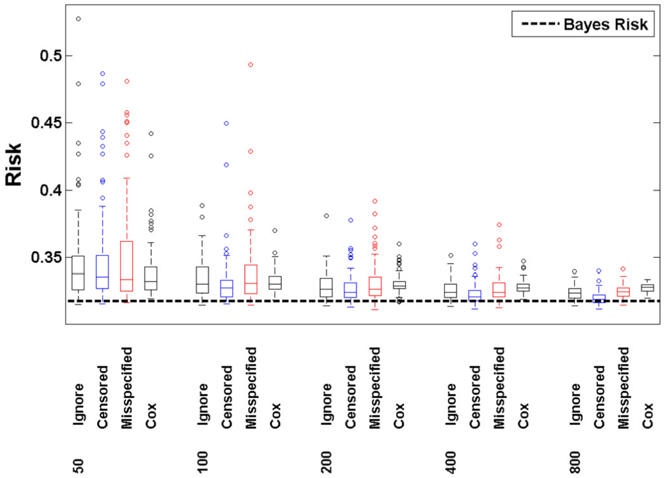

Fig 4. *Weibull failure time, non-linear proportional hazards (Setting 2): Distribution of the risk for different sizes of data set, for standard SVM that ignores the censored observations (Ignore), for censored SVM (Censored), for misspecified censoring mechanism (Misspecified), and for the Cox regression median (Cox). Bayes risk is denoted by a black dashed line. Each box plot is based on 100 repetitions of the simulation for each size of data set.*

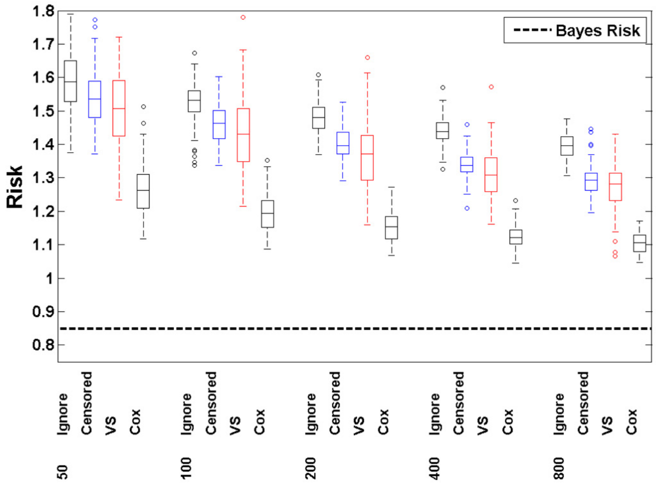

Fig 5. *Multidimensional Weibull failure time (Setting 3): Distribution of the risk for different data set sizes, for standard SVM that ignores the censored observations (Ignore), for censored SVM (Censored), for censored SVM with variable selection (VS), and for the Cox regression median with Lasso (Cox). Bayes risk is denoted by a black dashed line. Each box plot is based on 100 repetitions of the simulation for each size of data set.*

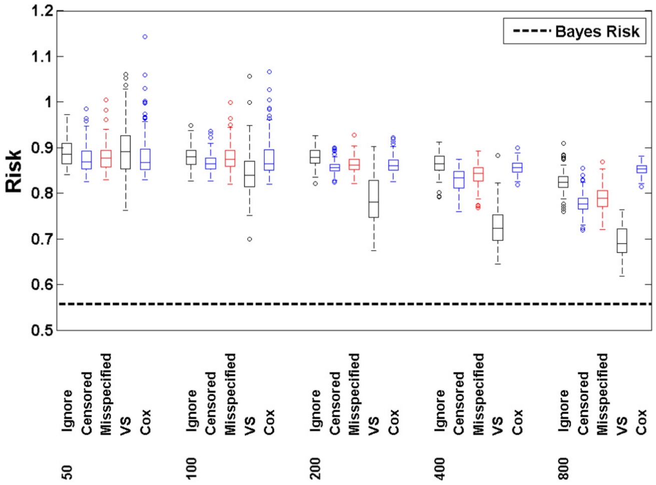

Fig 6. *Multidimensional Weibull failure time, non-linear proportional hazards (Setting 4): Distribution of the risk for different data set sizes, for standard SVM that ignores the censored observations (Ignore), for censored SVM (Censored), for censored SVM with variable selection (VS), and for the Cox regression median with Lasso (Cox). Bayes risk is denoted by a black dashed line. Each box plot is based on 100 repetitions of the simulation for each given data set size.*

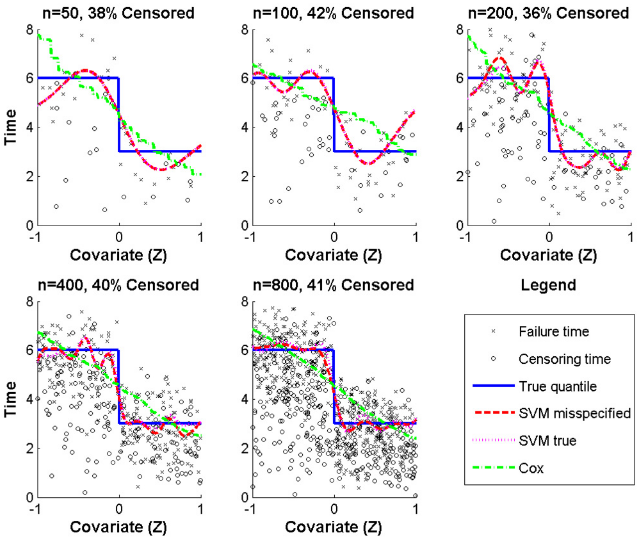

Fig 7. *Step function median, Weibull censoring time (Setting 5): The true conditional median (solid blue), the SVM decision function using the Kaplan-Meier estimator for the censoring (dashed red), the SVM decision function using the Cox estimator for censoring (doted magenta), and the Cox regression median (dot-dashed green) are plotted for samples of size* $n = 50, 100, 200, 400$ *and* 800. *The censoring percentage is given for each sample size. An observed failure times is represented by an* ×, *and an observed censoring time is represented by an* ◦.

sored observations, censored SVM, censored SVM with variable selection, and Cox regression with variable selection. We performed variable selection for censored SVM based on recursive feature elimination as in Guyon et al. (2002, Section 2.6). We performed variable selection for the Cox model using the Lasso (Tibshirani, 1997)[2]. Similarly to Settings 1 and 2, we examined the effect of model misspecification of the censoring mechanism by drawing C from a Weibull distribution with shape parameter 2 and scale parameter $-0.5Z_1 + \log(1.5)$ to ensure 40% censoring. When the proportional hazards assumption holds (Setting 3), SVM performs reasonably well, although the Cox model performs better as expected. When the proportional hazard assumption fails to hold (Setting 4), SVM performs better and it seems that the risk of Cox regression converges, but not to the Bayes risk (see Example 9 for discussion). Both figures show that variable selection achieves a slightly smaller median risk with the price of higher variance and that ignoring the censored observations and misspecifying the censoring model may lead to higher risk.

In the fifth setting, we consider a non-smooth conditional median. We also investigate the influence of using a misspecified model for the censoring mechanism. The covariates Z are generated uniformly on the segment $[-1, 1]$. The

[2] For the implementation of Lasso for Cox we used the Glmnet library for Matlab that can be found at `http://web.stanford.edu/~hastie/glmnet_matlab/`.

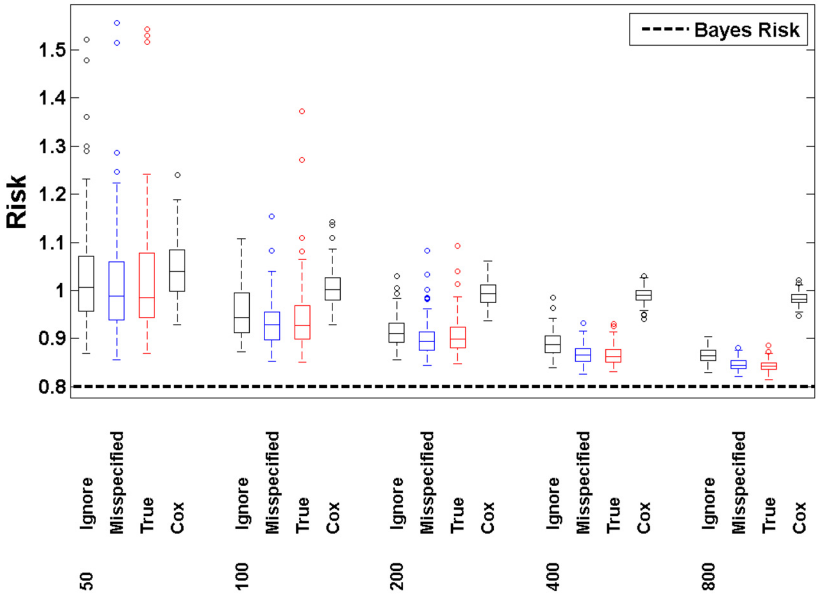

FIG 8. *Step function median, Weibull censoring time (Setting 5): Distribution of the risk for different sizes of data set, for standard SVM that ignores the censored observations (Ignore), for censored SVM with the Kaplan-Meier estimator for censoring (Misspecified), for censored SVM with the Cox estimator for censoring (True), and for the Cox regression median (Cox). The Bayes risk is denoted by a black dashed line. Each box plot is based on 100 repetitions of the simulation for each size of data set.*

failure time is normally distributed with expectation $3 + 3\mathbf{1}\{Z < 0\}$ and variance 1. Note that the proportional hazards assumption does not hold for the failure time. The censoring variable C follows the Weibull distribution with shape parameter 2, and scale parameter $-0.5Z + \log(6)$ which results in mean censoring percentage of 40%. Note that for this model, the censoring is independent of the failure time only given the covariate Z (see Assumption (A2)). Estimation of the censoring distribution using the Kaplan-Meier corresponds to estimation under a misspecified model. Since the censoring follows the proportional hazards assumption, estimation using the Cox estimator corresponds to estimation under the true model. We use 5-fold-cross-validation to choose the regularization constant and the width of the kernel, as in setting 1.

In Figure 7, the conditional median obtained by the censored SVM learning method using both the misspecified and true model for the censoring, and by Cox regression, are plotted. The true median is plotted as a reference. In Figure 8, we compare the risk of the SVM method using both the misspecified and true model for the censoring. We also examined the effect of ignoring the censored observations. Both figures show that in general SVM does better than the Cox model, regardless of the censoring estimation. The difference between the misspecified and true model for the censoring is small and the corresponding curves in Figure 7 almost coincide. Figure 8 shows again that there is a non-negligible price for ignoring the censored observations.

7. Concluding remarks

We studied an SVM framework for right censored data. We proposed a general censored SVM learning method and showed that it is well defined and measurable. We derived finite sample bounds on the deviation from the optimal risk. We proved risk consistency and computed learning rates. We discussed misspecification of the censoring model. Finally, we performed a simulation study to demonstrate the practical performances of censored SVM method.

We believe that this work illustrates an important approach for applying support vector machines to right censored data, and to missing data in general. However, many open questions remain and many possible generalizations exist. First, we assumed that censoring is independent of failure time given the covariates, and the probability that no censoring occurs is positive given the covariates. It should be interesting to study the consequences of violation of one or both assumptions. Second, we have used the inverse-probability-of-censoring weighting to correct the bias induced by censoring. This can be improved, for example, by using augmented inverse-probability-of-censoring weighting estimators (Tsiatis, 2006, Chapter 9). Such estimators were developed for outcome weighted learning of individual treatment rules using an RKHS framework (Zhao et al., 2015). It would be worthwhile to investigate how to develop such methods for SVMs. Third, we discussed only right-censored data and not general missing mechanisms. We believe that further development of SVM techniques that are able to better utilize the data and to perform under weaker assumptions and in more general settings is of great interest.

Appendix A

A.1. Survival function estimators

Since we are interested in estimating the censoring survival function and not the failure time survival function, some precaution is needed in order to apply the standard theory that appears in (FH91). For $t \in [0,\tau]$, define $\mathbf{N}(t) = \mathbf{1}\{U \leq t, \delta = 0\}$ and $\mathbf{Y}(t) = \mathbf{1}\{U > t\} + \mathbf{1}\{U = t, \delta = 0\}$. Note that $\mathbf{N}(t)$ is the counting process for the censoring, and not for the failure events, and $\mathbf{Y}(t)$ is the at-risk process for observing a censoring time. For a cadlag function A on $(0,\tau]$, define the product integral $\phi(A)(t) = \prod_{0<s\leq t}(1 + dA(s))$ (van der Vaart and Wellner, 1996). For a real-valued function f, we define $f(t-) = \lim_{s \nearrow t} f(s)$ when the limit exists.

Example 10. ***Independent censoring:*** *Assume that C is independent of both T and Z. Define*

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n d\mathbf{N}(s)}{\mathbb{P}_n \mathbf{Y}(s)}\,.$$

Then $\hat{G}_n(t) = \phi(-\hat{\Lambda})(t-)$ is the Kaplan-Meier estimator for G. It can be shown, similarly to Kosorok (2008, Chapter 12), that $\hat{G}_n$ is a consistent estimator for the survival function G.

Example 11. ***The proportional hazards model:*** *Consider the case that the hazard of C given Z is of the form $e^{Z'\boldsymbol{\beta}}d\Lambda$ for some unknown vector $\boldsymbol{\beta} \in \mathbb{R}^d$ and some continuous unknown nondecreasing function Λ with $\Lambda(0) = 0$ and $0 < \Lambda(\tau) < \infty$. Let $\hat{\boldsymbol{\beta}}$ be the zero of the estimating equation*

$$\Phi_n(\boldsymbol{\beta}) = \mathbb{P}_n \int_0^\tau \left(Z - \frac{\mathbb{P}_n Z\mathbf{Y}(s)e^{\boldsymbol{\beta}'Z}}{\mathbb{P}_n \mathbf{Y}(s)e^{\boldsymbol{\beta}'Z}} \right) d\mathbf{N}(s)\,.$$

Define

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n d\mathbf{N}(s)}{\mathbb{P}_n \mathbf{Y}(s)e^{\hat{\boldsymbol{\beta}}'Z}}\,.$$

Then it can be shown, similarly to Kosorok (2008, Chapters 4 and 12), that $\hat{G}_n(t|z) = \phi(-e^{\hat{\boldsymbol{\beta}}'z}\hat{\Lambda})(t-)$ is a consistent estimator for survival function G.

A.2. Auxiliary results

The following result is used to prove Theorem 5 and is based on results from SC08. Since it is not stated as a result there, we state the result and sketch the proof.

Theorem 10. *Let L be a loss function and H be an RKHS that satisfies assumptions (B1)–(B3). Fix $\lambda > 0$ and $\eta > 0$, and let $f \in H$. Then for all $n \geq 72\eta$, with probability not less than $1 - e^{-\eta}$,*

$$\begin{aligned}
&(P - \mathbb{P}_n)(L \circ \widehat{f} - L \circ f^*_{L,P}) \\
&\quad < \frac{17}{27}\left(\lambda\|f\|_H^2 + P(L \circ \widehat{f} - L \circ f^*_{L,P}) + \left(\frac{72V\eta}{n}\right)^{\frac{1}{2-\vartheta}} + r^* \right) \\
&\qquad + W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}\,,
\end{aligned}$$

where $W > 1$ is a constant that depends only on p, M, ϑ, and V, but not on f, and where

$$r^* = \inf_{f \in H} \lambda\|f\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}) - \mathcal{R}^*_{L,P}\,. \tag{17}$$

Proof. The proof is based on the proofs of Theorems 17.16, 17.20, and 17.23 of SC08. We now present a sketch of the proof for completeness.

We first note that if $a^{2p} > \lambda^p n$, it follows from (7) that the bound holds for $W \geq 4B$. Thus, we consider the case in which $a^{2p} \leq \lambda^p n$.

For every function $f \in H$, define the functions $h_f : \mathcal{Z} \times \mathcal{T} \mapsto \mathbb{R}$ as

$$h_f(z, t) = L(z, t, f(z)) - L(z, t, f^*_{L,P}(z))\,,$$

for all $z, t \in \mathcal{Z} \times \mathcal{T}$. Define

$$g_{f,r} = \frac{Ph_{\widehat{f}} - h_{\widehat{f}}}{\lambda\|f\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}) - \mathcal{R}^*_{L,P} + r}\,, \qquad f \in H,\ r > r^*\,.$$

Note that for every $f \in H$, $\|g_{f,r}\|_\infty \le 2Br^{-1}$. It can be shown (SC08, Eq. 7.43 and the discussion there) that $Pg_{f,r}^2 \le Vr^{\vartheta-2}$. Using Talagrand's inequality (SC08, Theorem 7.5) we obtain

$$P\left(\sup_{f\in H} \mathbb{P}_n g_{f,r} \le (1+\gamma)P[\sup_{f\in H}|\mathbb{P}g_{f,r}|] + \sqrt{\frac{2\eta V r^{\vartheta-2}}{n}} + \left(\frac{2}{3}+\frac{1}{\gamma}\right)\frac{2\eta B}{nr}\right) \tag{18}$$

is greater than or equals to $1-e^{-\eta}$ for every fixed $\gamma > 0$. Using Assumption (B3), it can be shown that there is a constant $\tilde{W}$ that depends only on p, M, ϑ, and V, such that for every $r > \max\{\tilde{W}\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}, r^*\}$

$$P[\sup_{f\in H}|\mathbb{P}g_{f,r}|] \le \frac{8}{30} \tag{19}$$

(see proofs of Theorems 7.20 and 7.23 of SC08, for details). Substituting $\gamma = 1/4$ in (18), and using the bound (19), we obtain that with probability of not less than $1-e^{-\eta}$,

$$\sup_{f\in H} \mathbb{P}_n g_{f,r} \le \frac{1}{3} + \sqrt{\frac{2\eta V r^{\vartheta-2}}{n}} + \frac{28\eta B}{3nr} \tag{20}$$

for all $r > \max\{\tilde{W}\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}, r^*\}$.

Using the fact that $n \ge 72\eta$, some algebraic manipulations (see SC08, proof of Theorem 7.23 for details) yield that for all $r \ge \left(\frac{72V\eta}{n}\right)^{1/(2-\vartheta)}$

$$\sqrt{\frac{2\eta V r^{\vartheta-2}}{n}} \le \frac{1}{6} \qquad , \qquad \frac{28\eta B}{3nr} \le \frac{7}{54}\,. \tag{21}$$

Fix $f \in H$. Using the definition of $g_{f,r}$, together with the estimates in (21) for the probability bound (20), we obtain that for

$$r = \tilde{W}\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \left(\frac{72V\eta}{n}\right)^{1/(2-\vartheta)} + r^*\,,$$

the inequality

$$(P-\mathbb{P}_n)(L\circ\widehat{f} - L\circ f^*_{L,P}) < \frac{17}{27}\left(\lambda\|f\|_H^2 + P(L\circ\widehat{f} - L\circ f^*_{L,P}) + r\right)$$

holds with probability not less than $1-e^{-\eta}$, and the desired result follows. □

Lemma 11. *Let L be a loss function and H be an RKHS that satisfies assumptions (B1)–(B3). Let $f \in H$ be such that $\|L(z,y,f(z))\|_\infty \le B$. Fix an $\eta > 0$, Then for all $n \ge 8\eta$, with probability not less than $1 - e^{-\eta}$,*

$$(\mathbb{P}_n - P)(L \circ f - L \circ f^*_{L,P}) < P(L \circ f - L \circ f^*_{L,P}) + \left(\frac{2V\eta}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{4B\eta}{3n}$$

For proof, see SC08, proof of Theorem 7.2.

A.3. Proof of Theorem 5

Proof. Let $\Omega_n \equiv \{\inf_Z \hat{G}_n(\tau|Z) > K\}$. Note that for any event A,

$$P(A) = 1 - P(A^c|\Omega_n)P(\Omega_n) - P(A^c|\Omega_n^c)P(\Omega_n^c) \ge 1 - P(A^c|\Omega_n) - P(\Omega_n^c)\,.$$

In this proof we show a bound for the expression $\lambda\|f^c_{D,\lambda}\|^2_H + \mathcal{R}_{L,P}(\widehat{f^c_{D,\lambda}}) - \mathcal{R}^*_{L,P}$ that given Ω_n does not hold with probability not greater than $3e^{-\eta}$ holds. By definition, $P(\Omega_n^c) \equiv P_{\hat{G}_n,n}$. Hence the bound that we find holds with probability not less that $1 - 3e^{-\eta} - P_{\hat{G}_n,n}$.

Note that by the definition of $f^c_{D,\lambda}$, for all $f_0 \in H$,

$$\lambda\|f^c_{D,\lambda}\|^2_H + \mathcal{R}_{L^n_D,D}(\widehat{f^c_{D,\lambda}}) \le \lambda\|f_0\|^2_H + \mathcal{R}_{L^n_D,D}(f_0),$$

where $\mathcal{R}_{L^n_D,D}(f) = \mathbb{P}_n \delta L(Z,Y(U),f(Z))/\hat{G}_n(U|Z)$. Hence,

$$\begin{aligned}
\lambda\|f^c_{D,\lambda}\|^2_H &+ \mathcal{R}_{L,P}(\widehat{f^c_{D,\lambda}}) - \mathcal{R}^*_{L,P} \\
&\le \lambda\|f_0\|^2_H + \mathcal{R}_{L^n_D,D}(f_0) - \mathcal{R}_{L^n_D,D}(\widehat{f^c_{D,\lambda}}) + \mathcal{R}_{L,P}(\widehat{f^c_{D,\lambda}}) - \mathcal{R}^*_{L,P} \\
&= \big(\lambda\|f_0\|^2_H + \mathcal{R}_{L,P}(f_0) - \mathcal{R}^*_{L,P}\big) + \big(\mathcal{R}_{L^n_D,D}(f_0) - \mathcal{R}_{L_G,D}(f_0)\big) \\
&\quad + \Big(\mathcal{R}_{L_G,D}(f_0) - \mathcal{R}_{L,P}(f_0) + \mathcal{R}_{L,P}(\widehat{f^c_{D,\lambda}}) - \mathcal{R}_{L_G,D}(\widehat{f^c_{D,\lambda}})\Big) \qquad (22) \\
&\quad + \Big(\mathcal{R}_{L_G,D}(\widehat{f^c_{D,\lambda}}) - \mathcal{R}_{L^n_D,D}(\widehat{f^c_{D,\lambda}})\Big) \\
&\equiv A_n + B_n + C_n + D_n\,,
\end{aligned}$$

where

$$\mathcal{R}_{L_G,D}(f) \equiv \mathbb{P}_n L_G(Z,U,\delta,f(Z)) \equiv \mathbb{P}_n \delta L(Z,Y(U),f(Z))/G(T|Z)\,,$$

i.e., $\mathcal{R}_{L_G,D}$ is the empirical loss function with the true censoring distribution function, and $L_G(Z,U,\delta,f(Z)) \equiv \delta L(Z,Y(U),f(Z))/G(T|Z)$.

In the following we will bound the expressions A_n,B_n, C_n, and D_n.

Bounding $\boldsymbol{A_n}$: For every function $f \in H$, define the functions $h_f : \mathcal{Z} \times \mathcal{T} \mapsto \mathbb{R}$ as

$$h_f(z,t) = L_G(z,t,f(z)) - L_G(z,t,f^*_{L,P}(z))\,,$$

for all $z,t \in \mathcal{Z} \times \mathcal{T}$. Using this notation, we can rewrite $A_n \equiv \lambda\|f_0\|^2_H + Ph_{f_0}$ since $P[L_G(Z,U,\delta,f(Z))] = P[L(Z,U,f(Z))]$.

Bounding $\boldsymbol{B_n}$: Note that by the definition of f_0, $\|L \circ f_0\|_\infty \leq B_0$. Hence,

$$\begin{aligned}|\mathcal{R}_{L_G,D}(f_0) - \mathcal{R}_{L_D^n,D}(f_0)| &\equiv \left|\mathbb{P}_n \frac{\delta L(Z,Y,f_0(Z))}{G(T|Z)} - \mathbb{P}_n \frac{\delta L(Z,Y,f_0(Z))}{\hat{G}_n(T|Z)}\right| \\ &= \left|\mathbb{P}_n \frac{\delta L(Z,Y,f_0(Z))}{G(T|Z)\hat{G}_n(T|Z)} \left(\hat{G}_n(T|Z) - G(T|Z)\right)\right| \\ &\leq \frac{B_0}{2K^2} \mathbb{P}_n |(\hat{G}_n - G)(T|Z)| \,,\end{aligned} \tag{23}$$

where the last inequality follows from condition (A1) and by assuming that Ω_n holds.

Bounding $\boldsymbol{C_n}$: First, using conditional expectation, we obtain that for every $f \in H$,

$$\begin{aligned}\mathcal{R}_{L,P}(f) \equiv P[L(Z,Y,f(Z))] &= P\left[P\left[\frac{\delta}{G(T|Z)} L(Z,Y,f(Z))\middle| Z,T\right]\right] \\ &= P[L_G(Z,U,\delta,f(Z)] = \mathcal{R}_{L_G,P}(f)\,.\end{aligned} \tag{24}$$

Therefore, we can rewrite the term C_n as

$$\begin{aligned}C_n &\equiv \mathcal{R}_{L_G,D}(f_0) - \mathcal{R}_{L,P}(f_0) + \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}_{L_G,D}(\widehat{f}^c_{D,\lambda}) \\ &= \left(\mathcal{R}_{L_G,D}(f_0) - \mathcal{R}_{L_G,D}(f^*_{L,P})\right) - \left(\mathcal{R}_{L_G,P}(f_0) - \mathcal{R}_{L_G,P}(f^*_{L,P})\right) \\ &\quad + \left(\mathcal{R}_{L_G,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}_{L_G,P}(f^*_{L,P})\right) - \left(\mathcal{R}_{L_G,D}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}_{L_G,D}(f^*_{L,P})\right)\,,\end{aligned} \tag{25}$$

where $f^*_{L,P}$ is the Bayes decision function.

Using this notation of h_f defined above, we can rewrite (25) as

$$\begin{aligned}C_n &\equiv (\mathbb{P}_n - P)h_{f_0} + (P - \mathbb{P}_n)h_{\widehat{f}^c_{D,\lambda}} \\ &= (\mathbb{P}_n - P)(h_{f_0} - h_{\widehat{f_0}}) + (\mathbb{P}_n - P)h_{\widehat{f_0}} + (P - \mathbb{P}_n)h_{\widehat{f}^c_{D,\lambda}} \\ &\equiv C_{n,1} + C_{n,2} + C_{n,3}\,.\end{aligned} \tag{26}$$

Bounding $\boldsymbol{C_{n,1}}$: In order to bound $C_{n,1}$, we use Bernstein's inequality (see, for example, SC08, Theorem 6.12), and hence we first show that it is bounded and bound its variance. Since $L_G(z,t,f_0(z)) - L_G(z,t,\widehat{f_0}(z)) \geq 0$, we obtain from the definition of L_G, (7) and the bound on f_0 that $h_{f_0} - h_{\widehat{f_0}} \in [0, B_0/2K]$. It thus follows that

$$\begin{aligned}\mathrm{Var}(h_{f_0} - h_{\widehat{f_0}}) &\equiv P\left[\left((h_{f_0} - h_{\widehat{f_0}}) - P(h_{f_0} - h_{\widehat{f_0}})\right)^2\right] \\ &\leq P(h_{f_0} - h_{\widehat{f_0}})^2 \leq \frac{B_0}{2K} P(h_{f_0} - h_{\widehat{f_0}})\,.\end{aligned}$$

Using Bernstein's inequality for the function $h_{f_0} - h_{\widehat{f_0}} - P(h_{f_0} - h_{\widehat{f_0}})$, we obtain that with probability not less than $1 - e^{-\eta}$,

$$(\mathbb{P}_n - P)(h_{f_0} - h_{\widehat{f_0}}) \le \sqrt{\frac{\eta B_0 P(h_{f_0} - h_{\widehat{f_0}})}{Kn}} + \frac{B_0\eta}{3Kn}\,.$$

Using $\sqrt{ab} \le \frac{a}{2} + \frac{b}{2}$, we obtain

$$\sqrt{\frac{\eta B_0 P(h_{f_0} - h_{\widehat{f_0}})}{Kn}} \le P(h_{f_0} - h_{\widehat{f_0}}) + \frac{B_0\eta}{4Kn}\,,$$

which leads to the bound

$$C_{n,1} \le P(h_{f_0} - h_{\widehat{f_0}}) + \frac{7B_0\eta}{12Kn}\,, \tag{27}$$

which holds with probability not less than $1 - e^{-\eta}$.

Bounding $C_{n,2}$: In order to bound $C_{n,2}$ we first show that Assumption (B2) holds also for L_G with the constant $\bar{V} = V/2K$. Write

$$\begin{aligned}
P(h^2_{\widehat{f_0}}) &= P\left[\left(\frac{\delta}{G(T|Z)}(L(Z,Y,\widehat{f_0}(Z)) - L(Z,Y,f^*_{L,P}(Z)))\right)^2\right] \\
&= P\left(P\left[\left(\frac{\delta}{G(T|Z)}(L(Z,Y,\widehat{f_0}(Z)) - L(Z,Y,f^*_{L,P}(Z)))\right)^2 |T,Z\right]\right) \\
&\le \frac{1}{2K}P\left[\left(L(Z,Y,\widehat{f_0}(Z)) - L(Z,Y,f^*_{L,P}(Z))\right)^2\right] \\
&\le \frac{V}{2K}\left(P\left[L(Z,Y,\widehat{f}(Z)) - L(Z,Y,f^*_{L,P}(Z))\right]\right)^{\vartheta}\,,
\end{aligned} \tag{28}$$

where for the first inequality we use $E(\delta^2|T,Z) = G(T|Z)$, and that $G(T) > 2K$; and the last inequality follows from (8). Noting that $\|L_G(z,y,\widehat{f_0}(z))\|_\infty \le B/2K$, we can apply Lemma 11, and obtain that with probability not less than $1 - e^{-\eta}$, for all $n \ge 8\eta$,

$$C_{n,2} \equiv (\mathbb{P}_n - P)h_{\widehat{f_0}} < Ph_{\widehat{f_0}} + \left(\frac{2\bar{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{2B\eta}{3Kn}\,.$$

Bounding $C_{n,3}$: By Theorem 10, with probability not less than $1 - e^{-\eta}$, for all $n \ge 72\eta$,

$$\begin{aligned}
(P - \mathbb{P}_n)h_{\widehat{f^c_{D,\lambda}}} < \frac{17}{27}&\left(\lambda\|f^c_{D,\lambda}\|^2_H + Ph_{\widehat{f^c_{D,\lambda}}} + \left(\frac{72\bar{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + (\lambda\|f_0\|^2_H + Ph_{f_0})\right) \\
&+ W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}\,,
\end{aligned}$$

where $W > 1$ is a constant that depends only on p, M, ϑ, and $\bar{V}$, and where we used the fact that $r^* \le \lambda\|f_0\|^2_H + Ph_{f_0}$ where r^* is defined in (17).

Bounding $\boldsymbol{D_n}$: Note that by assumption (B1), $\|L \circ \widehat{f}^c_{D,\lambda}\|_\infty \le B$. Hence, similarly to (23),

$$|\mathcal{R}_{L_G,D}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}_{L^n_D,D}(\widehat{f}^c_{D,\lambda})| \le \frac{B}{2K^2}\mathbb{P}_n|(\hat{G}_n - G)(T|Z)|\,,$$

where the last inequality follows from condition (A1).

Summarizing, we obtain that with probability not less than $1-3e^{-\eta}$, for all $n \ge 72\eta$,

$$\begin{aligned}
&\lambda\|f^c_{D,\lambda}\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P} \\
&\quad\le (\lambda\|f_0\|_H^2 + Ph_{f_0}) + \frac{B_0}{2K^2}\mathbb{P}_n|(\hat{G}_n - G)(T|Z)| + \left(P(h_{f_0} - h_{\widehat{f_0}}) + \frac{7B_0\eta}{12Kn}\right) \\
&\qquad + \left(Ph_{\widehat{f_0}} + \left(\frac{2\bar{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + \frac{2B\eta}{3Kn}\right) \\
&\qquad + \frac{17}{27}\left(\lambda\|f^c_{D,\lambda}\|_H^2 + Ph_{\widehat{f^c_{D,\lambda}}} + \left(\frac{72\bar{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + \lambda\|f_0\|_H^2 + Ph_{f_0}\right) \\
&\qquad + W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}} + \frac{B}{2K^2}\mathbb{P}_n|(\hat{G}_n - G)(T|Z)|\,.
\end{aligned}$$

Let $\tilde{V} = \max\{\bar{V}, (B/(2K))^{2-\vartheta}\}$ and $n \ge 72\eta$, then

$$\frac{2B\eta}{3Kn} = \frac{4}{3}\cdot\frac{B}{2K}\cdot\frac{1}{72}\cdot\frac{72\eta}{n} \le \frac{1}{54}\tilde{V}^{\frac{1}{2-\vartheta}}\left(\frac{72\eta}{n}\right)^{\frac{1}{2-\vartheta}}\,.$$

Hence, using the facts that $6 \le 36^{1/(2-\vartheta)}$, $B_0 > B$, and that by conditional expectation (24), $\mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P} \equiv Ph_{\widehat{f^c_{D,\lambda}}}$, we obtain

$$\begin{aligned}
&(1-\frac{17}{27})\left(\lambda\|f^c_{D,\lambda}\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P}\right) \\
&\qquad\le \left(1+\frac{17}{27}\right)\lambda\|f_0\|_H^2 + \left(2+\frac{17}{27}\right)Ph_{f_0} + \frac{B_0}{K^2}\mathbb{P}_n|(\hat{G}_n - G)(T|Z)| \\
&\qquad\quad + \frac{7B_0\eta}{12Kn} + \frac{36^{\frac{1}{2-\vartheta}}}{6}\left(\frac{2\bar{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} \qquad (29) \\
&\qquad\quad + \left(\frac{1}{54}+\frac{17}{27}\right)\left(\frac{72\bar{V}\eta}{n}\right)^{\frac{1}{2-\vartheta}} + W\left(\frac{a^{2p}}{\lambda^p n}\right)^{\frac{1}{2-p-\vartheta+\vartheta p}}\,.
\end{aligned}$$

Since

$$\begin{aligned}
&\lambda\|f^c_{D,\lambda}\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P} \\
&\qquad < 3\left(1-\frac{17}{27}\right)(\lambda\|f^c_{D,\lambda}\|_H^2 + \mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P})\,,
\end{aligned}$$

multiplying both sides of (29) and rounding up the constants on the right hand side we obtain the result.

Until now we assumed that $n \geq 72\eta$. Assume now that $n < 72\eta$. Since we used the fact that $n \geq 72\eta$ only to bound $C_{n,2}$ and $C_{n,3}$, we obtain that for all n,

$$\begin{aligned}\lambda\|f^c_{D,\lambda}\|^2_H + &\mathcal{R}_{L,P}(\widehat{f}^c_{D,\lambda}) - \mathcal{R}^*_{L,P}\\ &\leq \ (\lambda\|f_0\|^2_H + Ph_{f_0}) + \frac{B_0}{K^2}\mathbb{P}_n|(\hat{G}_n - G)(T|Z)|\\ &\quad + \left(P(h_{f_0} - h_{\widehat{f_0}}) + \frac{7B_0\eta}{12Kn}\right) + C_{n,2} + C_{n,3}\,.\end{aligned}$$

Recall that $C_{n,2} + C_{n,3} \equiv (P - \mathbb{P}_n)(h_{\widehat{f_0}} + h_{\widehat{f}^c_{D,\lambda}})$. For every clipped function $\widehat{f}$, $\|L_G \circ \widehat{f} - L_G \circ f^*_{L,P}\| \leq B/(2K)$. Hence, by the definition of $h_{\widehat{f_0}} + h_{\widehat{f}^c_{D,\lambda}}$, we obtain that $(P - \mathbb{P}_n)(h_{\widehat{f_0}} + h_{\widehat{f}^c_{D,\lambda}}) \leq 4 \cdot (B/2K)$. Using the fact that $B/(2K) \leq \tilde{V}^{1/(2-\vartheta)}$, we obtain that

$$(P - \mathbb{P}_n)h_{\widehat{f_0}} + h_{\widehat{f}^c_{D,\lambda}} \leq 4\left(\frac{72\tilde{V}\eta}{n}\right)^{1/(2-\vartheta)}$$

and thus the result follows also for the case $n < 72\eta$. □

Supplementary Material

Matlab code

(doi: 10.1214/17-EJS1231SUPP; .zip). Please read the file README.pdf for details on the files in this folder.

References

P. L. Bartlett. The sample complexity of pattern classification with neural networks. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998. MR1607706

E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statist. Med.*, 17(10):1169–1186, 1998.

D. Bitouzé, B. Laurent, and P. Massart. A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.*, 35(6):735–763, 1999.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. MR1929416

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001. MR1874152

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1–3):131–159, 2002.

P. Chen and A. A. Tsiatis. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics*, 57(4):1030–1038, 2001. MR1950418

A. Eleuteri and A. F. G. Taktak. Support Vector Machines for Survival Regression. In E. Biganzoli, A. Vellido, F. Ambrogi, and R. Tagliaferri, editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, number 7548, pages 176–189. Springer, 2011.

T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. Wiley, 1991.

Y. Goldberg and M. R. Kosorok. An exponential bound for Cox regression. *Statistics & Probability Letters*, 82(7):1267–1272, 2012a.

Y. Goldberg and M. R. Kosorok. Q-learning with censored data. *The Annals of Statistics*, 40(1):529–560, 2012b.

Y. Goldberg and M. R. Kosorok. Supplement to "Support vector regression for right censored data". 2017. DOI: 10.1214/17-EJS1231SUPP.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008. MR2418654

T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Tröger. Bagging survival trees. *Statistics in Medicine*, 23(1):77–91, 2004.

H. Ishwaran and U. B. Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80(13–14):1056–1064, 2010.

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.

B. A. Johnson, D. Y. Lin, J. S. Marron, J. Ahn, J. Parker, and C. M. Perou. Threshhold analyses for inference in high dimension low sample size datasets with censored outcomes. Unpublished manuscript, 2004.

T. G. Karrison. Use of Irwin's restricted mean as an index for comparing survival in different treatment groups–Interpretation and power considerations. *Controlled Clinical Trials*, 18(2):151–167, 1997.

M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

J. F. Lawless. *Statistical Models and Methods for Lifetime Data*. Wiley, 2003.

B. D. Ripley and R. M. Ripley. Neural networks as statistical methods in survival analysis. In Ri. Dybowski and V. Gant, editors, *Clinical Applications of Artificial Neural Networks*, pages 237–255. Cambridge University Press, 2001.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. MR1294730

M. R. Segal. Regression Trees for Censored Data. *Biometrics*, 44(1), 1988.

J. Shim and C. Hwang. Support vector censored quantile regression under random censoring. *Computational Statistics & Data Analysis*, 53(4):912–919, 2009.

P. K. Shivaswamy, W. Chu, and M. Jansche. A support vector approach to censored targets. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), Omaha, Nebraska, USA*, pages 655–660. IEEE Computer Society, 2007.

I. Steinwart and A. Chirstmann. *Support Vector Machines*. Springer, 2008.

I. Steinwart, D. Hush, and C. Scovel. An oracle inequality for clipped regularized risk minimizers. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1321–1328. MIT Press, Cambridge, MA, 2007.

R. Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996. MR1385671

J. Wellner. On an exponential bound for the Kaplan-Meier estimator. *Lifetime Data Analysis*, 13(4):481–496, 2007.

Q. Wu, Y. Ying, and D. Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007. MR2297018

Y. Zhao, D. Zeng, M. A. Socinski, and M. R. Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.

Y. Zhao, D. Zeng, E. B Laber, R. Song, M. Yuan, and M. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168, 2015. MR3335102

R. Zhu and M. R. Kosorok. Recursively Imputed Survival Trees. *Journal of the American Statistical Association*, 107(497):331–340, 2011.

D. M. Zucker. Restricted mean life with covariates: Modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442):702–709, 1998.