Estimating conditional hazard functions and densities with the highly-adaptive lasso

Anders Munch[1], Thomas A. Gerds[1], Mark J. van der Laan[2], and Helene C. W. Rytgaard[1]

[1]Section of Biostatistics, University of Copenhagen
[2]Devision of Biostatistics, University of California, Berkeley

April 18, 2024

Abstract

We consider estimation of conditional hazard functions and densities over the class of multivariate càdlàg functions with uniformly bounded sectional variation norm when data are either fully observed or subject to right-censoring. We demonstrate that the empirical risk minimizer is either not well-defined or not consistent for estimation of conditional hazard functions and densities. Under a smoothness assumption about the data-generating distribution, a highly-adaptive lasso estimator based on a particular data-adaptive sieve achieves the same convergence rate as has been shown to hold for the empirical risk minimizer in settings where the latter is well-defined. We use this result to study a highly-adaptive lasso estimator of a conditional hazard function based on right-censored data. We also propose a new conditional density estimator and derive its convergence rate. Finally, we show that the result is of interest also for settings where the empirical risk minimizer is well-defined, because the highly-adaptive lasso depends on a much smaller number of basis function than the empirical risk minimizer.

1 Introduction

Let $\mathcal{D}_M^d$ be the space of multivariate càdlàg functions $f\colon [0,1]^d \to \mathbb{R}$ with sectional variation norm bounded by $M < \infty$. For a suitable loss function $L\colon \mathcal{D}_M^d \times [0,1] \to \mathbb{R}$ and a probability measure P on $[0,1]^d$, we consider the parameter

$$f^* = \operatorname*{argmin}_{f \in \mathcal{D}_M^d} P[L(f,\cdot)], \quad \text{where} \quad P[L(f,\cdot)] = \int_{[0,1]^d} L(f,x)\,\mathrm{d}P(x). \tag{1}$$

The empirical risk minimizer estimates f^* by minimizing $\mathbb{P}_n[L(f,\cdot)]$ over $\mathcal{D}_M^d$, where $\mathbb{P}_n$ is the empirical measure of a dataset $\{X_i\}_{i=1}^n$ of i.i.d. observations $X_i \sim P$. The highly-adaptive lasso (HAL) estimator proposed by van der Laan [2017] estimates f^* by minimizing $\mathbb{P}_n[L(f,\cdot)]$ over a sieve, i.e., a growing subset of the parameter space [Grenander, 1981, Geman, 1981, Geman and Hwang, 1982, Walter and Blum, 1984]. For particular choices of loss functions and sieves, the HAL estimator and the empirical risk minimizer will be identical, but they might also be different. We demonstrate that the use of a sieve is necessary for conditional hazard and density estimation, as empirical risk minimizers over the class of cadlag functions are not formally well-defined in these settings. A recent result by van der Laan [2023] formally established that a particular data-adaptive choice of sieve is sufficient to achieve the asymptotic convergence rate of $n^{-1/3}\log(n)^{2(d-1)/3}$ when a smoothness assumption is imposed on the measure P. We use this result to theoretically study a conditional hazard function estimator and a novel conditional density estimator based on a HAL.

Estimation of function-valued parameters over the function class $\mathcal{D}_M^d$ is of interest because the bound on the sectional variation norm works like a non-parametric sparsity constraint that to some extend allows us to avoid the curse of dimensionality. A particularly important application is in targeted or debiased machine learning [van der Laan and Rose, 2011, Chernozhukov et al., 2018], for which non-parametric estimators of regression functions, conditional densities, and conditional hazard functions are needed. A targeted or debiased estimator relies on the ability to estimate such nuisance parameters faster than rate $n^{-1/4}$. It has been shown that this rate can be achieved independently of the dimension of the covariate space when the nuisance parameter is assumed to belong to $\mathcal{D}_M^d$ [van der Laan, 2017, Bibaut and van der Laan, 2019]. In this paper we take a closer look at this important result for conditional densities and hazard functions. In addition, our work is relevant for estimation of regression functions: The HAL estimator can be constructed using a number of basis functions that scales linearly in the sample size n while the empirical risk minimizer needs a number of basis function that is of order n^d [Fang et al., 2021].

The challenge with estimation of densities and hazard functions over $\mathcal{D}_M^d$ is illustrated in Figure 1. Informally, the sectional variation norm measures how much a function fluctuates without taking into account where in the domain of the function the fluctuations happen. A consequence is that we can redistribute the probability mass assigned by a given càdlàg density function without changing its sectional variation norm in such a way that the log-likelihood loss is decreased. A similar issue occurs with right-censored data in survival analysis, and our Proposition 10 formally shows that the empirical risk minimizer is in general either not well-defined or not consistent for the conditional hazard function. On the other hand, a consistent HAL estimator of a conditional hazard function does exist.

Earlier related work on non-parametric functional estimation used Sobolev spaces [Goldstein and Messer, 1992, Bickel and Ritov, 1988, Stone, 1980, Goldstein and Khasminskii, 1996]. Estimation over the class of multivariate càdlàg functions with uniformly bounded sectional variation norm was introduced in [van der Laan, 2017]. Estimation of conditional hazard functions in the presence of censoring has traditionally been done using kernel smoothing or local linear polynomials [e.g., Ramlau-Hansen, 1983, McKeague and Utikal, 1990, van Keilegom and Veraverbeke, 2001, Spierdijk, 2008], while more recent

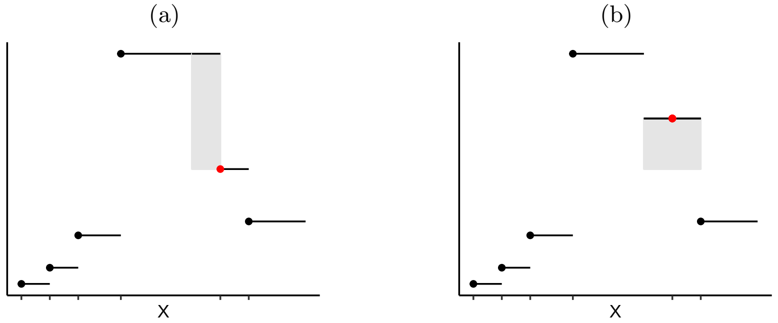

Figure 1: Illustration of two càdlàg densities, where the ticks at the x-axis denote observed data points, and the y-coordinates of the black dots denote the likelihood given to these points by the densities. Panel (a) shows a given càdlàg density, while panel (b) shows an adjusted density that has the same variation norm but assigns a higher likelihood to the observed data. Note that the function in panel (b) is a density, because the gray boxes in panels (a) and (b) have the same area.

approaches use boosting [Schmid and Hothorn, 2008, Hothorn, 2020, Lee et al., 2021]. Conditional hazard function estimation based on HAL was proposed by Rytgaard et al. [2022, 2023]. Fang et al. [2021] considered estimation of regression functions over the class of functions with uniformly bounded Hardy-Krause variation [Krause, 1903, Hardy, 1906, Owen, 2005, Aistleitner and Dick, 2015], which is closely related to the class of functions considered here.

The remainder of the article is organized as follows. In Section 2 we introduce our notation and review the properties of multivariate càdlàg functions with bounded sectional variation norm. Section 3 contains a formal definition of the general loss based estimation problem and two (potentially different) estimators; the empirical risk minimizer and a HAL estimator. In Section 4 we define a projection of f^* onto a data-adaptive sieve, which allows us to derive the asymptotic convergence rate directly for the HAL estimator without assuming it to be identical to the empirical risk minimizer. In Sections 5-7 we apply our general results to special cases. In Section 5 we consider the setting of censored survival data observed in continuous time, and show that while the HAL estimator is well-defined, the empirical risk minimizer is in general either ill-defined or inconsistent. In Section 6 we consider conditional density estimation and propose a new estimator. Section 7 considers an example from the regression setting, where the empirical risk minimizer is well-defined, and we illustrate the dramatic reduction in the number of basis functions needed to calculate the HAL estimator compared to the empirical risk minimizer. Section 8 contains a discussion of our results. Appendices A-C contain proofs.

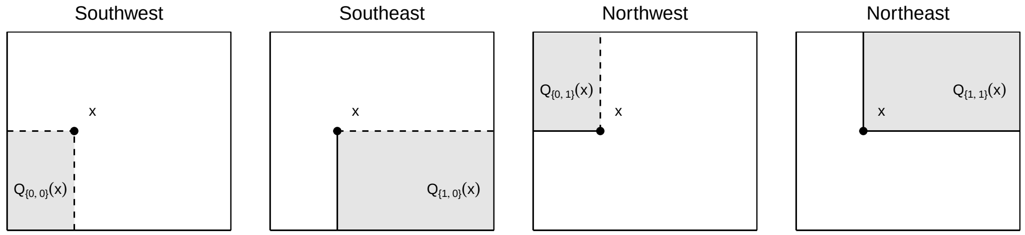

Figure 2: The four quadrants $Q_{\{0,0\}}(\mathbf{x})$, $Q_{\{1,0\}}(\mathbf{x})$, $Q_{\{0,1\}}(\mathbf{x})$, and $Q_{\{1,1\}}(\mathbf{x})$ spanned by the point $\mathbf{x} \in [0,1]^2$ and each of the vertices in the unit square. A sequence which is contained in one of the quadrants and converges to u, converges from 'southwest', 'southeast', 'northwest', or 'northeast'. That the function f is càdlàg means that the limit of the function f should exist when we approach it from any of these four directions, and the limit should agree with the function value at u when we approach it from 'northeast'.

2 Multivariate càdlàg functions with bounded sectional variation norm

For $d = 1$ the definition of a càdlàg function is given by its name – it is a function that is continuous from the right with left-hand limits. When $d > 1$ we can approach a point from an infinite number of directions, and thus the concepts 'right' and 'left' are not defined. In dimension $d \in \mathbb{N}$ we define càdlàg functions as follows. For any $u \in [0,1]$ and $a \in \{0,1\}$ we define the interval

$$I_a(u) = \begin{cases} [u,1] & \text{if } a = 1, \\ [0,u) & \text{if } a = 0. \end{cases}$$

For any $\mathbf{u} = (u_1, \ldots, u_d) \in [0,1]^d$ and $\mathbf{a} = (a_1, \ldots, a_d) \in \{0,1\}^d$ define the quadrant $Q_{\mathbf{a}}(\mathbf{u}) = I_{a_1}(u_1) \times \cdots \times I_{a_d}(u_d)$.

Definition 1 (Multivariate càdlàg function)**.** A function $f\colon [0,1]^d \to \mathbb{R}$ is *càdlàg* if for all $\mathbf{u} \in [0,1]^d$, $\mathbf{a} \in \{0,1\}^d$, and any sequence $\{\mathbf{u}_n\} \subset Q_{\mathbf{a}}(\mathbf{u})$ which converges to $\mathbf{u}$ as $n \to \infty$, the limit $\lim_{n\to\infty} f(\mathbf{u}_n)$ exists, and $\lim_{n\to\infty} f(\mathbf{u}_n) = f(\mathbf{u})$ for $\mathbf{a} = \mathbf{1}$.

Neuhaus [1971] first generalized the concept of a càdlàg function to the multivariate setting. Our Definition 1 is an equivalent definition used by, e.g., Czerebak-Morozowicz et al. [2008] and Ferger [2015]. We use $\mathcal{D}^d$ to denote the collection of all càdlàg functions with domain $[0,1]^d$. The content of Defintion 1 is illustrated in Figure 2 for $d = 2$.

A bit of notation is needed to formally define the section of a càdlàg function and the sectional variation norm. For any non-empty subset $s \subset [d] = \{1, \ldots, d\}$ let $\pi_s\colon \{1, \ldots, |s|\} \to [d]$ be the unique increasing function such that $\operatorname{Im}(\pi_s) = s$, i.e., π_s provides the ordered indices of $1, \ldots, d$ included in s. For any $\mathbf{x} \in [0,1]^d$ we define the s-section of the vector $\mathbf{x}$ as

$$\mathbf{x}_s = (x_{\pi_s(1)}, \ldots, x_{\pi_s(|s|)}) \in [0,1]^{|s|},$$

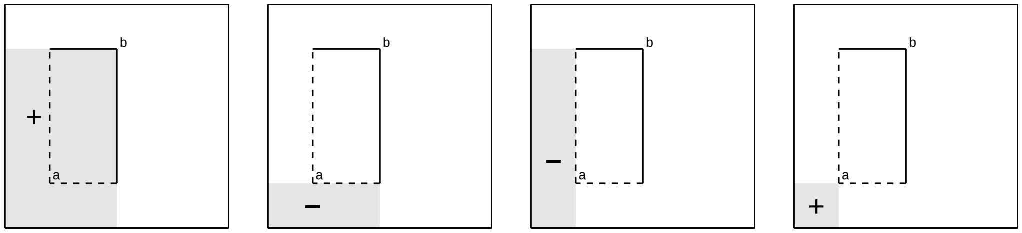

Figure 3: The area of the box $(\mathbf{a}, \mathbf{b}]$ can be calculated by first calculating the gray area in the leftmost figure, subtracting the gray areas in the two middle figures, and then adding the gray area in the rightmost figure.

i.e., $\mathbf{x}_s$ is the ordered tuple in $[0,1]^{|s|}$ consisting of all components of $\mathbf{x}$ with index in s. Note that for a singleton $s = \{i\}$, we have $\mathbf{x}_{\{i\}} = x_i$. Defining

$$\overline{\mathbf{x}}_s = (\mathbb{1}\{1 \in s\}x_1, \mathbb{1}\{2 \in s\}x_2, \ldots, \mathbb{1}\{d \in s\}x_d) \in [0,1]^d.$$

the s-section of f is the function

$$f_s \colon [0,1]^{|s|} \longrightarrow \mathbb{R} \quad \text{such that} \quad f_s(\mathbf{x}_s) = f(\overline{\mathbf{x}}_s), \quad \forall \mathbf{x} \in [0,1]^d.$$

In words, f_s is the function that appears when all arguments of f that are not in s are fixed at zero. For vectors $\mathbf{a}, \mathbf{b} \in [0,1]^d$ we write

$$\begin{aligned} \mathbf{a} \preceq \mathbf{b} \quad &\text{if} \quad a_k \leq b_k, \quad \text{for} \quad k = 1, \ldots, d, \\ \mathbf{a} \prec \mathbf{b} \quad &\text{if} \quad a_k < b_k, \quad \text{for} \quad k = 1, \ldots, d, \end{aligned}$$

and we define closed and half-open boxes by $[\mathbf{a}, \mathbf{b}] = \{\mathbf{x} \in [0,1]^d : \mathbf{a} \preceq \mathbf{x} \preceq \mathbf{b}\}$ and $(\mathbf{a}, \mathbf{b}] = \{\mathbf{x} \in [0,1]^d : \mathbf{a} \prec \mathbf{x} \preceq \mathbf{b}\}$, respectively. For a box $A = (\mathbf{a}, \mathbf{b}] \subset [0,1]^d$ with $\mathbf{a} \prec \mathbf{b}$, let

$$\mathcal{V}(A) = \{\mathbf{v} = (v_1, \ldots, v_d) : v_i = a_i \text{ or } v_i = b_i\}$$

denote the set of vertices of the box A. The *quasi-volume* assigned to the box $A = (\mathbf{a}, \mathbf{b}]$ by the function f is

$$\Delta(f; A) = \sum_{\mathbf{v} \in \mathcal{V}(A)} (-1)^{H(\mathbf{v})} f(\mathbf{v}), \quad \text{with} \quad H(\mathbf{v}) = \sum_{k=1}^{d} \mathbb{1}\{v_k = a_k\}.$$

The idea is that the volume of the box A as measured by f can be computed by calculating volumes of boxes with corners at $\mathbf{0}$, as illustrated in Figure 3 for $d = 2$. Let ρ denote a finite partition of $(0,1]$ given by

$$\rho = \{(x_{l-1}, x_l] : l = 1, \ldots, L\}, \quad \text{with} \quad 0 = x_0 < x_1 < \cdots < x_L = 1.$$

For any collection $\rho_1, \ldots, \rho_d$ of univariate partitions, we define a partition $\mathcal{P}$ of $(0,1]^d$ by

$$\mathcal{P} = \{I_1 \times I_2 \times \cdots \times I_d : I_k \in \rho_k, k = 1, \ldots, d\}. \tag{2}$$

For a function $f\colon [0,1]^d \to \mathbb{R}$ the *Vitali variation* is defined as

$$V(f) = \sup_{\mathcal{P}} \sum_{A \in \mathcal{P}} |\Delta(f; A)|,$$

where the supremum is taken over all partitions given by equation (2). The *sectional variation norm* of a function $f\colon [0,1]^d \to \mathbb{R}$ is the sum of the Vitali variation of all its sections plus the absolute value of the function at $\mathbf{0}$, i.e.,

$$\|f\|_v = |f(\mathbf{0})| + \sum_{s \in \mathcal{S}} V(f_s), \quad \text{with} \quad \mathcal{S} = \{s \subset [d] : s \neq \emptyset\}.$$

For $M \in (0, \infty)$ we use $\mathcal{D}_M^d$ to denote the space of càdlàg functions $f\colon [0,1]^d \to \mathbb{R}$ with $\|f\|_v \leq M$.

We now give two alternative descriptions of $\mathcal{D}_M^d$. The first characterizes $\mathcal{D}_M^d$ as the closure of the collection of rectangular piece-wise constant functions (Proposition 2), and the second puts $\mathcal{D}_M^d$ into a one-to-one correspondence with finite signed measures (Proposition 3).

Define the function space $\mathcal{F}^d = \{\mathbb{1}_{[\mathbf{x},\mathbf{1}]} : \mathbf{x} \in [0,1]^d\}$, let $\operatorname{Span}(\mathcal{F}^d)$ denote all linear combinations of elements from $\mathcal{F}^d$, and define $\mathcal{R}_M^d = \{f \in \operatorname{Span}(\mathcal{F}^d) : \|f\|_v \leq M\}$. An example of an element in $\mathcal{F}^d$ is shown in Figure 4 (a).

Proposition 2. *Consider $\mathcal{R}_M^d$ and $\mathcal{D}_M^d$ as subspaces of the Banach space of all bounded functions $f\colon [0,1]^d \to \mathbb{R}$ equipped with the supremum norm. Then $\mathcal{D}_M^d = \overline{\mathcal{R}_M^d}$, that is, $\mathcal{R}_M^d \subset \mathcal{D}_M^d$ and for any function $f \in \mathcal{D}_M^d$ there exists a sequence of functions $\{f_n\} \subset \mathcal{R}_M^d$ such that $\|f - f_n\|_\infty \to 0$.*

Proof. See Appendix A. □

In the following, let $\|\mu\|_{\mathrm{TV}} = \mu_+([0,1]^d) + \mu_-([0,1]^d)$ denote the total variation norm of the measure μ.

Proposition 3. *For any $f \in \mathcal{D}_M^d$ there exists a unique signed measure μ_f on $[0,1]^d$ such that*

$$f(\mathbf{x}) = \mu_f([\mathbf{0}, \mathbf{x}]), \quad \forall \mathbf{x} \in [0,1]^d,$$

and $\|\mu_f\|_{\mathrm{TV}} = \|f\|_v$. For any signed measure μ on $[0,1]^d$ with $\|\mu_f\|_{\mathrm{TV}} \leq M$ there exists a unique function $f_\mu \in \mathcal{D}_M^d$ such that

$$f_\mu(\mathbf{x}) = \mu([\mathbf{0}, \mathbf{x}]), \quad \forall \mathbf{x} \in [0,1]^d.$$

Proof. A similar result is proved by Aistleitner and Dick [2015]. We use their result in our proof in Appendix A which is for càdlàg functions. □

Proposition 3 shows that the class of functions considered by Fang et al. [2021] is identical to the class $\mathcal{D}_M^d$ up to a constant.

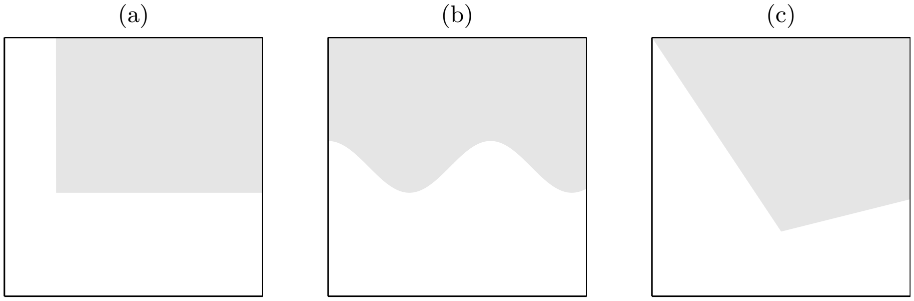

Figure 4: Illustration of functions $f\colon [0,1]^2 \to \mathbb{R}$ that are 0 on the white area and 1 on the shaded area. The function in panel (a) is càdlàg. The functions in panels (b) and (c) are not càdlàg.

Based on Proposition 3 we can define the integral with respect to a function $f \in \mathcal{D}_M^d$ as the integral with respect to the measure μ_f. We use the notation $\mathrm{d}f = \mathrm{d}\mu_f$ and $|\,\mathrm{d}f| = \mathrm{d}|\mu_f|$, where $|\mu| = \mu_+ + \mu_-$. The connection between càdlàg functions and measures is the key component underlying the HAL estimator. The HAL estimator is motivated by the following representation of functions in $\mathcal{D}_M^d$ which is due to Gill et al. [1995] and van der Laan [2017].

Proposition 4. *For any $f \in \mathcal{D}_M^d$ we can write*

$$f(\mathbf{x}) = f(\mathbf{0}) + \sum_{s\in\mathcal{S}} \int_{[0,1]^{|s|}} \mathbb{1}_{(\mathbf{0}_s,\mathbf{x}_s]}(\mathbf{u})\,\mathrm{d}f_s(\mathbf{u}),$$

and

$$\|f\|_v = |f(\mathbf{0})| + \sum_{s\in\mathcal{S}} \int_{[0,1]^{|s|}} \mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}(\mathbf{u})|\,\mathrm{d}f_s|(\mathbf{u}),$$

where $\mathcal{S} = \{s \subset [d] : s \neq \emptyset\}$.

Proof. See Appendix A. □

Proposition 2 showed that $\mathcal{D}_M^d$ is the closure of the piece-wise constant functions $\mathcal{R}_M^d$. Proposition 5 implies that piece-wise constant functions that are not in $\mathcal{R}_M^d$, like the ones in Figures 4 (b) and (c), are not càdlàg.

Proposition 5. *Let $f : [0,1]^d \to \mathcal{K} \subset \mathbb{R}$ for some finite set $\mathcal{K}$. If $f \notin \mathcal{R}_M^d$ then f is not càdlàg.*

Proof. See Appendix A. □

3 Empirical risk minimization and the HAL estimator

We now consider a general setup for loss-based estimation. We assume given an i.i.d. dataset $O_i \sim P$, $i = 1, \ldots, n$, with data on the form

$$O = (X, Y) \in \mathcal{O} = [0,1]^d \times \mathcal{Y}, \quad \text{for} \quad \mathcal{Y} \subset \mathbb{R}. \tag{3}$$

We use $\mathbb{P}_n$ to denote the empirical measure corresponding to a data set $\{O_i\}_{i=1}^n$. Let L be a loss function $L\colon \mathcal{D}_M^d \times \mathcal{O} \to \mathbb{R}$. We define the target parameter

$$f^* = \operatorname*{argmin}_{f \in \mathcal{D}_M^d} P[L(f, \cdot)], \tag{4}$$

which formally depends on M but we suppress that in the notation. A natural estimator of f^* is the substitution estimator, also known as the *the empirical risk minimizer*,

$$\operatorname*{argmin}_{f \in \mathcal{D}_M^d} \mathbb{P}_n[L(f, \cdot)]. \tag{5}$$

The optimization problem in equation (5) reduces to a finite but high-dimensional optimization problem for the squared error loss [Fang et al., 2021]. We conjecture that this can be generalized to loss functions for which we can write [Bibaut and van der Laan, 2019, Assumption 2]

$$L(f, (\mathbf{x}, y)) = \tilde{L}(f(\mathbf{x}), y), \quad \forall f \in \mathcal{D}_M^d, \quad \text{for some function} \quad \tilde{L}\colon [0,1]^d \times \mathcal{Y} \longrightarrow \mathbb{R}_+. \tag{6}$$

Note that equation (6) does not hold in general for the negative log-likelihood as we demonstrate in Section 5.

We now turn to define the HAL estimator [van der Laan, 2017]. The HAL estimator is motivated from the representation given by Proposition 4, which shows that we can estimate $f \in \mathcal{D}_M^d$ by estimating the signed measures generated by its sections. Let $\delta_{X_{s,i}}$ be the Dirac measure at the s-section of X_i and define the estimator of the signed measure of the s-section of f,

$$\mathrm{d}f_{\beta^s,n} = \sum_{i=1}^n \beta_i^s \delta_{X_{s,i}}, \quad \text{with unknown parameter vector} \quad \beta^s = (\beta_1^s, \ldots, \beta_n^s) \in \mathbb{R}^n.$$

This gives the following data-dependent model for estimation of $f \in \mathcal{D}_M^d$,

$$f_{\beta,n}(\mathbf{x}) = \beta_0 + \sum_{s \in \mathcal{S}} \sum_{i=1}^n \beta_i^s \mathbb{1}\{X_{s,i} \preceq \mathbf{x}_s\}, \quad \text{with} \quad \beta = \{\beta^s : s \in \mathcal{S}\} \cup \{\beta_0\}, \tag{7}$$

where $\mathcal{S} = \{s \subset [d] : s \neq \emptyset\}$. As $|\mathcal{S}| = \sum_{j=1}^d \binom{d}{j} = 2^d - 1$ we have that $\beta \in \mathbb{R}^{m(d,n)}$ with $m(d,n) = n(2^d - 1) + 1$. We refer to the indicator functions in equation (7) as basis function. Some examples of basis functions are given in Figure 5 for $d = 2$. By Proposition 2, any $f_{\beta,n}$ is an element of $\mathcal{D}^d$ and we have

$$\|f_{\beta,n}\|_v = \|\beta\|_1 = |\beta_0| + \sum_{s \in \mathcal{S}} \sum_{i=1}^n |\beta_{i,s}|. \tag{8}$$

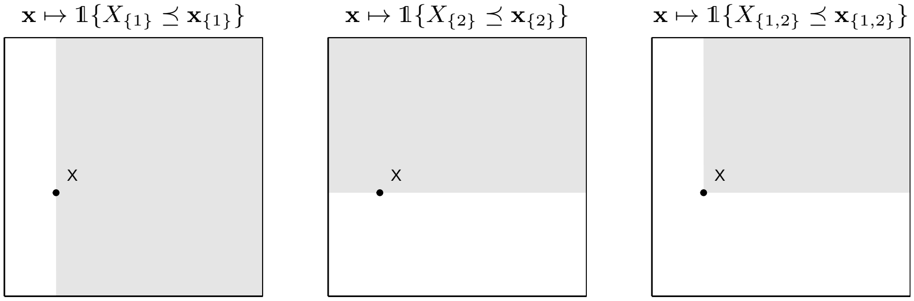
$\mathbf{x} \mapsto \mathbb{1}\{X_{\{1\}} \preceq \mathbf{x}_{\{1\}}\}$
$\mathbf{x} \mapsto \mathbb{1}\{X_{\{2\}} \preceq \mathbf{x}_{\{2\}}\}$
$\mathbf{x} \mapsto \mathbb{1}\{X_{\{1,2\}} \preceq \mathbf{x}_{\{1,2\}}\}$
X
X
X

Figure 5: Examples of the basis functions that are used to construct the HAL estimator for $d = 2$.

Denote the space of all functions of this form by

$$\mathcal{D}_n^d := \{f_{\beta,n} : \beta \in \mathbb{R}^{m(d,n)}\} \subset \mathcal{D}^d,$$

and denote similarly the subspace of these function with a sectional variation norm bounded by a fixed constant $M < \infty$ by

$$\mathcal{D}_{M,n}^d := \{f_{\beta,n} : \beta \in \mathbb{R}^{m(d,n)}, \|\beta\|_1 \leq M\} \subset \mathcal{D}_M^d.$$

A *highly-adaptive lasso (HAL) estimator* is then defined as

$$\hat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{D}_{M,n}^d} \mathbb{P}_n[L(f, \cdot)]. \tag{9}$$

We refer to any minimizer as a HAL estimator.

4 Convergence rates using a projection

In this section we show that a HAL estimator $\hat{f}_n$ enjoys the same convergence rate as has been shown to hold for the empirical risk minimizer, when this is well-defined, under an additional smoothness assumption (see Assumption 7 and the following discussion). In addition, we derive asymptotic convergence rates for a HAL estimator in a setting where the empirical risk minimizer is not well-defined. We denote by $N_{[]}(\varepsilon, \mathcal{H}, \|\cdot\|)$ the bracketing number for a function space $\mathcal{H}$ with respect to a norm $\|\cdot\|$. The bracketing number is the minimum number of brackets with a norm smaller than ε needed to cover $\mathcal{H}$ [van der Vaart and Wellner, 1996]. We use $\|\cdot\|_\infty$ to denote the supremum norm and $\|\cdot\|_v$ to denote the sectional variation norm, while for a measure μ we use $\|\cdot\|_\mu$ to denote the $\mathcal{L}^2(\mu)$-norm. We use λ to denote Lebesgue measure. Recall that the data is of the form $O = (X, Y)$ with $X \in [0, 1]^d$. For all non-empty subsets $s \subset \{1, \dots, d\}$ we let P_s denote the marginal distribution of X_s. We let $\mu_{f_s^*}$ denote the measures generated by the sections f_s^*. Note that the measures $\mu_{f_s^*}$ and P_s operate on the same measure space $[0, 1]^{|s|}$. We assume that $\mathbb{1}_{(\mathbf{0}_s, \mathbf{1}_s]} \cdot \mu_{f_s^*} \ll \mathbb{1}_{(\mathbf{0}_s, \mathbf{1}_s]} \cdot P_s$ and write the Radon-Nikodym derivatives as

$$\mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}\frac{\mathrm{d}f_s^*}{\mathrm{d}P_s} = \frac{\mathrm{d}\{\mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}\cdot\mu_{f_s^*}\}}{\mathrm{d}\{\mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}\cdot P_s\}}, \quad \text{for} \quad s\in\mathcal{S}.$$

Assumption 6 (Smoothness of the loss function)**.** For a loss function L define the function space $\mathcal{L}_M = \{L(f,\cdot) : f\in\mathcal{D}_M^d\}$. There exist constants $C<\infty$, $\eta>0$, and $\kappa\in\mathbb{N}$ such that the following conditions hold.

(i) $\|L(f,\cdot)\|_\infty \le C$ for all $f\in\mathcal{D}_M^d$.

(ii) $C^{-1}\|f-f^*\|_\lambda^2 \le P[L(f,\cdot)-L(f^*,\cdot)] \le C\|f-f^*\|_\lambda^2$ for all $f\in\mathcal{D}_M^d$.

(iii) $N_{[\,]}(\varepsilon,\mathcal{L}_M,\|\cdot\|_P) \le CN_{[\,]}(\varepsilon/C,\mathcal{D}_M^d,\|\cdot\|_\lambda)^\kappa$ for all $\varepsilon\in(0,\eta)$.

Assumption 6 (ii) is a standard assumption [e.g., van der Vaart and Wellner, 1996]. Some general conditions on the loss functions can be given to ensure that Assumption 6 (iii) holds, see for instance Lemma 4 in Appendix B in [Bibaut and van der Laan, 2019].

Assumption 7 (Data-generating distribution)**.** There is a constant $C<\infty$ such that the following conditions hold.

(i) The target parameter f^* is an inner point of $\mathcal{D}_M^d$ with respect to the sectional variation norm, i.e., $\|f^*\|_v < M$.

(ii) $\mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}\cdot\mu_{f_s^*} \ll \mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}\cdot P_s$ and $\|\mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}\,\mathrm{d}f_s^*/\,\mathrm{d}P_s\|_\infty \le C$ for all $s\in\mathcal{S}$.

Assumption 7 (ii) is substantial, as it imposes an additional smoothness condition on f^*. For instance, if P is dominated by Lebesgue measure, Assumption 7 (ii) implies that the measures generated by the sections of f^* must also be dominated by Lebesgue measure, hence f^* must be continuous. We discuss the necessity of this assumption further in Section 7.

Our main result (Theorem 9) relies on Lemma 8 which is based on a construction given in Appendix B of [van der Laan, 2023]. A proof of the lemma is given at the end of this section.

Lemma 8. *For any $f\in\mathcal{L}^2(\lambda)$ and $n\in\mathbb{N}$ there exists a (random) function $\hat\pi_n(f)\in\mathcal{D}_{M,n}^d$ such that*

$$\hat\pi_n(f) = \operatorname*{argmin}_{h\in\mathcal{D}_{M,n}^d}\|h-f\|_\lambda.$$

For any f^* fulfilling Assumption 7, it holds that

$$\|\hat\pi_n(f^*)-f^*\|_\lambda = O_P(n^{-1/2}).$$

Theorem 9. *If Assumptions 6 and 7 hold, and $\hat f_n$ is a HAL estimator as defined in equation (9), then*

$$\|\hat f_n - f^*\|_\lambda = O_P(n^{-1/3}\log(n)^{2(d-1)/3}).$$

Proof. We can write

$$\|\hat{f}_n - f^*\|_\lambda \leq \|\hat{f}_n - \hat{\pi}_n(f^*)\|_\lambda + \|\hat{\pi}_n(f^*) - f^*\|_\lambda,$$

where $\hat{\pi}_n$ is the projection defined in Lemma 8. As Assumption 7 is assumed to hold, the second term on the right hand side is of order $O_P(n^{-1/2})$. The first term of the right hand side can be analyzed using classical results from empirical process theory. A detailed proof showing that $\|\hat{f}_n - \hat{\pi}_n(f^*)\|_\lambda = O_P(n^{-1/3}\log(n)^{2(d-1)/3})$ is given in Appendix B. □

Proof of Lemma 8. By definition of $\mathcal{D}^d_{M,n}$, any element $h \in \mathcal{D}^d_{M,n}$ can be written as $h = \sum_{k=1}^{m(d,n)} \beta_k h_k$, for some coefficients $\beta = (\beta_1, \ldots, \beta_{m(d,n)})$ and indicator functions h_k. Minimizing $h \mapsto \|h - f\|_\lambda$ over $\mathcal{D}^d_{M,n}$ is thus equivalent to minimizing

$$\mathcal{G}(\beta) = \int_{[0,1]^d} \left\{ \left(\sum_{k=1}^{m(d,n)} \beta_k h_k \right)^2 - 2 \sum_{k=1}^{m(d,n)} \beta_k h_k f^* \right\} \mathrm{d}\lambda$$

over the set $\mathcal{B}_M = \{\beta \in \mathbb{R}^{m(d,n)} : \|\beta\|_1 \leq M\}$. Writing

$$\mathcal{G}(\beta) = \sum_{k=1}^{m(d,n)} \sum_{l=1}^{m(d,n)} \beta_k \beta_l \int_{[0,1]^d} h_k h_l \,\mathrm{d}\lambda - 2 \sum_{k=1}^{m(d,n)} \beta_k \int_{[0,1]^d} h_k f^* \,\mathrm{d}\lambda,$$

shows that $\mathcal{G}$ is continuous. As $\mathcal{B}_M$ is compact, it follows that a minimum is attained.

To show the second statement of the lemma, we follow the proof of Lemma 23 in [van der Laan, 2023] and define the random function

$$f^*_n(\mathbf{x}) = f^*(\mathbf{0}) + \sum_{s \in \mathcal{S}} \int_{(\mathbf{0}_s, \mathbf{x}_s]} \frac{\mathrm{d}f^*_s}{\mathrm{d}P_s} \,\mathrm{d}\mathbb{P}_{s,n}, \tag{10}$$

where $\mathbb{P}_{s,n}$ is the empirical measure of the s-section of the data $\{X_i\}_{i=1}^n$, i.e., the empirical measure obtained from $\{X_{s,i}\}_{i=1}^n$. This function is well-defined by Assumption 7 (ii). We next show that

$$\|f^*_n - f^*\|_\lambda = O_P(n^{-1/2}), \tag{11}$$

and

$$P\big(f^*_n \in \mathcal{D}^d_{M,n}\big) \longrightarrow 1. \tag{12}$$

To see this, we use the representation given by Proposition 4 and Assumption 7 (ii) to write

$$f^*(\mathbf{x}) = f^*(\mathbf{0}) + \sum_{s \in \mathcal{S}} \int_{(\mathbf{0}_s, \mathbf{x}_s]} \mathrm{d}f^*_s = f^*(\mathbf{0}) + \sum_{s \in \mathcal{S}} \int_{(\mathbf{0}_s, \mathbf{x}_s]} \frac{\mathrm{d}f^*_s}{\mathrm{d}P_s} \,\mathrm{d}P_s,$$

from which we obtain

$$f^*_n(\mathbf{x}) - f^*(\mathbf{x}) = \sum_{s \in \mathcal{S}} \int_{(\mathbf{0}_s, \mathbf{x}_s]} \frac{\mathrm{d}f^*_s}{\mathrm{d}P_s} \,\mathrm{d}[\mathbb{P}_{s,n} - P_s] = n^{-1/2} \sum_{s \in \mathcal{S}} \mathbb{G}_{s,n}\left[\mathbb{1}_{(\mathbf{0}_s, \mathbf{x}_s]} \frac{\mathrm{d}f^*_s}{\mathrm{d}P_s} \right],$$

where $\mathbb{G}_{s,n}$ denotes the empirical process of the s-section of the data. As $\{\mathbb{1}_{(\mathbf{0}_s,\mathbf{x}_s]} : \mathbf{x}_s \in (0,1]^{|s|}\}$ is a Donsker class [van der Vaart and Wellner, 1996], it follows from the preservation properties of Donsker classes and the assumption that $\mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}\,\mathrm{d}f_s^*/\,\mathrm{d}P_s$ is uniformly bounded, that also

$$\mathcal{F}_s^* = \left\{ \mathbb{1}_{(\mathbf{0}_s,\mathbf{x}_s]} \frac{\mathrm{d}f_s^*}{\mathrm{d}P_s} : \mathbf{x}_s \in (0,1]^{|s|} \right\}$$

is a Donsker class. As this holds for any section $s \in \mathcal{S}$, we have

$$\|f_n^* - f^*\|_\infty \leq n^{-1/2} \sum_{s\in\mathcal{S}} \sup_{f\in\mathcal{F}_s^*} |\mathbb{G}_{s,n}[f]| = n^{-1/2} \sum_{s\in\mathcal{S}} O_P(1) = O_P(n^{-1/2}),$$

which in particular shows equation (11). To show equation (12), note that

$$f_n^*(\mathbf{x}) = f^*(\mathbf{0}) + \sum_{s\in\mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}(X_{s,i}) \frac{\mathrm{d}f_s^*}{\mathrm{d}P_s}(X_{s,i}) \mathbb{1}\{X_{s,i} \preceq \mathbf{x}_s\}. \tag{13}$$

Equation (13) shows that $f_n^* \in \mathcal{D}_n^d$, and by equation (8)

$$\begin{aligned} \|f_n^*\|_v &= f^*(\mathbf{0}) + \sum_{s\in\mathcal{S}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]}(X_{s,i}) \left| \frac{\mathrm{d}f_s^*}{\mathrm{d}P_s} \right| (X_{s,i}) \\ &= f^*(\mathbf{0}) + \sum_{s\in\mathcal{S}} \mathbb{P}_{s,n} \left[\mathbb{1}_{(\mathbf{0}_s,\mathbf{1}_s]} \frac{|\,\mathrm{d}f_s^*|}{\mathrm{d}P_s} \right], \end{aligned}$$

where we use $|\,\mathrm{d}f_s^*|/\,\mathrm{d}P_s$ to denote the Radon-Nikodym derivative of $|\mu_{f_s^*}|$ with respect to P_s on $(\mathbf{0}_s, \mathbf{1}_s]$. The last equality follows from the properties of the Jordan-Hahn decomposition and the fact that P_s is a positive measure. By Assumption 7 (ii) and the law of large numbers this implies that $\|f_n^*\|_v \xrightarrow{P} \|f^*\|_v$. As $\|f^*\|_v < M$ by Assumption 7 (i) it follows that $P(\|f_n^*\|_v < M) \to 1$, which shows equation (12).

Define the indicator variable $\eta_n = \mathbb{1}\{f_n^* \in \mathcal{D}_{M,n}^d\}$. Note that equation (12) implies that $P(\eta_n = 1) \to 1$ and thus [e.g., Schuler et al., 2023, Lemma 2] yields

$$(1 - \eta_n) = o_P(a_n^{-1}) \quad \text{for any sequence } a_n \longrightarrow \infty. \tag{14}$$

When $\eta_n = 1$ we have $f_n^* \in \mathcal{D}_{M,n}^d$ an thus by definition of $\pi_n(f^*)$ we have $\eta_n\|\pi_n(f^*) - f^*\|_\lambda \leq \eta_n\|f_n^* - f^*\|_\lambda$. From this it follows that

$$\begin{aligned} \|\pi_n(f^*) - f^*\|_\lambda &\leq \eta_n\|f_n^* - f^*\|_\lambda + (1 - \eta_n)\|\pi_n(f^*) - f^*\|_\lambda \\ &\leq \|f_n^* - f^*\|_\lambda + (1 - \eta_n)2M = O_P(n^{-1/2}), \end{aligned}$$

where the last equality follows from equations (11) and (14). □

5 Right-censored data

Let $T \in \mathbb{R}_+$ be a time to event variable and $W \in [0,1]^{d-1}$ a covariate vector. In this section we discuss estimation of the hazard function $\alpha(t, \mathbf{w})$, for $t \in [0,1]$ and $\mathbf{w} \in$

$[0,1]^{d-1}$, where

$$\alpha(t,\mathbf{w}) = \lim_{\varepsilon \searrow 0} \frac{P(T \in [t, t+\varepsilon] \mid T \geq t, W = \mathbf{w})}{\varepsilon}.$$

We parameterize the log-hazard function as a multivariate càdlàg function with bounded sectional variation norm,

$$\log \alpha(t,\mathbf{w}) = f(t,\mathbf{w}), \quad \text{with} \quad f \in \mathcal{D}^d_M. \tag{15}$$

Let $C \in \mathbb{R}_+$ be a right-censoring time. We assume conditional independent censoring, i.e., $C \perp\!\!\!\perp T \mid W$. As we are only interested in the conditional hazard function for $t \in [0,1]$, we can focus on the truncated event time $T \wedge 1$. We observe $O = (W, \tilde{T}, \Delta)$, where $\tilde{T} = T \wedge 1 \wedge C$ and $\Delta = \mathbb{1}\{T \leq (C \wedge 1)\}$. The right-censored data fits into the setup described in Section 3 by setting $X = (W, \tilde{T})$, $Y = \Delta$, and $\mathcal{Y} = \{0,1\}$. We denote by n' the number of unique time points, and by $\tilde{T}_{(1)} < \tilde{T}_{(2)} < \tilde{T}_{(n')}$ the ordered sequence of observed unique time points. We define $\tilde{T}_{(0)} = 0$.

As loss function we use the negative log of the partial likelihood for f [Cox, 1975, Andersen et al., 2012],

$$L^{\mathrm{pl}}(f, O) = \int_0^{\tilde{T}} e^{f(u,W)}\,\mathrm{d}u - \Delta f(\tilde{T}, W). \tag{16}$$

The remainder of this section is organized as follows. We start by showing that the empirical risk minimizer according to the partial likelihood loss is either not defined or not consistent. We then show that the HAL estimator is well-defined and derive its asymptotic convergence rate.

Proposition 10 gives a formal statement of the problem described in Figure 1 in Section 1. To demonstrate the problem it is sufficient to consider the univariate case without covariates.

Proposition 10. *Let $f^\circ \in \mathcal{D}^1_M$ be given. If there exists a $j \in \{1, \dots, n'-1\}$ such that $f^\circ(\tilde{T}_{(j)}) > f^\circ(\tilde{T}_{(j+1)})$, then*

$$f^\circ \notin \operatorname*{argmin}_{f \in \mathcal{D}^1_M} \mathbb{P}_n[L^{\mathrm{pl}}(f, \cdot)].$$

Proof. See Appendix C.1. □

Proposition 10 implies that any estimator $\hat{f}_n \in \mathcal{D}^1_M$ of the log-hazard function which decreases between two time points is not an empirical risk minimizer. Thus, unless the hazard function that generated the data is non-decreasing, an empirical risk minimizer either does not exist or is inconsistent. Proposition 11 on the other hand shows that a HAL estimator can be found as the solution to a convex optimization problem.

Proposition 11. *Let $f_{\beta,n}$ be the data-dependent model defined in equation (7). The problem*

$$\min_{\|\beta\|_1 \leq M} \mathbb{P}_n[L^{\mathrm{pl}}(f_{\beta,n}, \cdot)], \tag{17}$$

is convex and has a solution. For any solution $\hat{\beta}$, $f_{\hat{\beta},n}$ is a HAL estimator, i.e.,

$$f_{\hat{\beta},n} \in \operatorname*{argmin}_{f \in \mathcal{D}^d_{M,n}} \mathbb{P}_n[L^{\mathrm{pl}}(f,\cdot)].$$

Proof. See Appendix C.1. □

We assume that the conditional hazard function for the right-censoring time exists on $[0,1)$ for all $\mathbf{w} \in [0,1]^{d-1}$ and denote it by $\gamma(t,\mathbf{w})$. We assume that γ is uniformly bounded for all $(t,\mathbf{w}) \in [0,1) \times [0,1]^{d-1}$. Without loss of generality we can take

$$P(\tilde{T} = 1 \mid W = \mathbf{w}) = P(\tilde{T} = 1, \Delta = 0 \mid W = \mathbf{w}) = \exp\left\{ -\int_{[0,1)} \gamma_0(s,\mathbf{w})\,\mathrm{d}s \right\}. \tag{18}$$

As T and C are assumed conditionally independent given W, any two uniformly bounded conditional hazard functions α and γ together with a marginal distribution for the covariate vector W uniquely determine a distribution P for the observed data O through equation (18). We write α_P and γ_P for the two conditional hazard functions corresponding to a distribution P, and let $f_P = \log \alpha_P$. We assume that W has a Lebesgue density and denote this with ω_P.

Lemma 12. *Let P be a distribution such that $\|\gamma_P\|_\infty < \infty$, $\varepsilon < \omega_P < 1/\varepsilon$, for some $\varepsilon > 0$, and $f_P \in \mathcal{D}^d_M$. Then for all $f \in \mathcal{D}^d_M$,*

$$P[L^{\mathrm{pl}}(f,\cdot) - L^{\mathrm{pl}}(f_P,\cdot)] \asymp \|f - f_P\|^2_\lambda.$$

Proof. The lemma essentially follow from general properties of the Kullback-Leibler divergence. However, due to the point-mass at $t = 1$, a few additional arguments are needed which we present in Appendix C.1. □

Corollary 13. *Let P be a distribution such that $\|\gamma_P\|_\infty < \infty$, $\varepsilon < \omega_P < 1/\varepsilon$, for some $\varepsilon > 0$, and let $\hat{f}_n$ be a HAL estimator based on the negative partial log-likelihood loss defined in equation (16). If Assumption 7 holds, then*

$$\|\hat{f}_n - f_P\|_\lambda = o_P(n^{-1/3}\log(n)^{2(d-1)/3}).$$

Proof. Corollary 13 follows from Theorem 9 and Lemma 12. Details are given in Appendix C.1. □

We illustrate the HAL estimator of a conditional hazard function and the effect of the sectional variation with the following example. Consider a study that enrolls patients between the age of 20 and 60 to study the effect of a treatment on death within one year after treatment. We simulate an artificial dataset such that the hazard of death does not depend on age in the untreated group, while the hazard of death among treated patients is lowered for patients younger than 40, but increased for older patients. Censoring is generated independently of covariates and event times. As noted by Rytgaard et al.

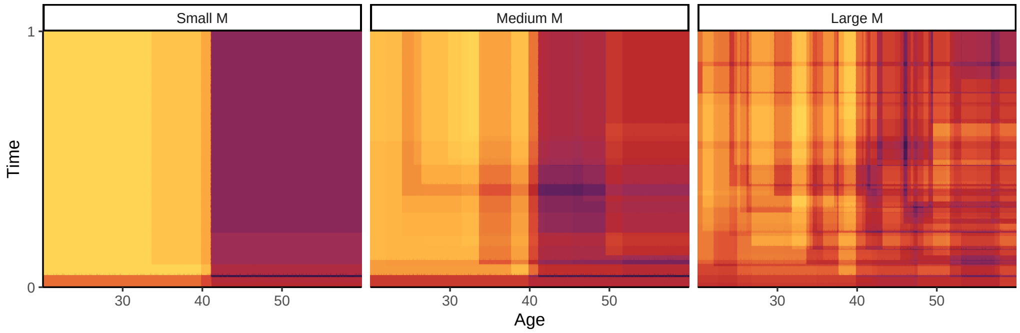

Figure 6: The HAL estimator of the hazard function for the treated group based on a sample size of 200 from the simulated study with darker values corresponding to higher values of the hazard function. Estimates are shown for three different values of the sectional variation norm (M).

[2023], the loss in equation (39) can be recognized as the negative log-likelihood of a Poisson model. This implies that we can use existing software from the R-packages `glmnet` [Friedman et al., 2010, Tay et al., 2023] and `hal9001` [Hejazi et al., 2020, Coyle et al., 2022] to construct a HAL estimator. The HAL estimator, computed on a simulated dataset of 200 patients, is displayed for the treated group in Figure 6 across various values of the sectional variation norm M. We illustrate the corresponding estimate of the conditional survival function for both treatment groups in Figure 7.

6 Density estimation

Let $U \in [0,1]$ and $W \in [0,1]^{d-1}$ and consider estimation of the conditional density of U given W. In this section the available data are $O = (U, W)$, i.e., in the notation of the general setup of Section 3, $X = (U, W)$ and no additional variable Y is observed. We parameterize the conditional density as an element of

$$\mathcal{P}_M^d = \left\{ p\colon [0,1]^d \to \mathbb{R}_+ \;\middle|\; \log p(u, \mathbf{w}) = f(u, \mathbf{w}) - \log\left(\int_0^1 e^{f(z,\mathbf{w})}\, \mathrm{d}z \right), f \in \mathcal{D}_M^d \right\}. \quad (19)$$

This parametrization is a natural one and has been used before for (univariate) density estimation [e.g., Leonard, 1978, Silverman, 1982, Gu and Qiu, 1993]. Note that any element of $\mathcal{P}_M^d$ is a conditional density, and that $\mathcal{P}_M^d$ includes all conditional densities p such that $\log p \in \mathcal{D}_M^d$. Define the data-adaptive model

$$\mathcal{P}_{M,n}^d = \left\{ p \in \mathcal{P}_M^d \;\middle|\; \log p(u, \mathbf{w}) = f(u, \mathbf{w}) - \log\left(\int_0^1 e^{f(z,\mathbf{w})}\, \mathrm{d}z \right), f \in \mathcal{D}_{M,n}^d \right\},$$

and a HAL estimator as

$$\hat{p}_n \in \operatorname*{argmin}_{p \in \mathcal{P}_{M,n}^d} \mathbb{P}_n[-\log p]. \quad (20)$$

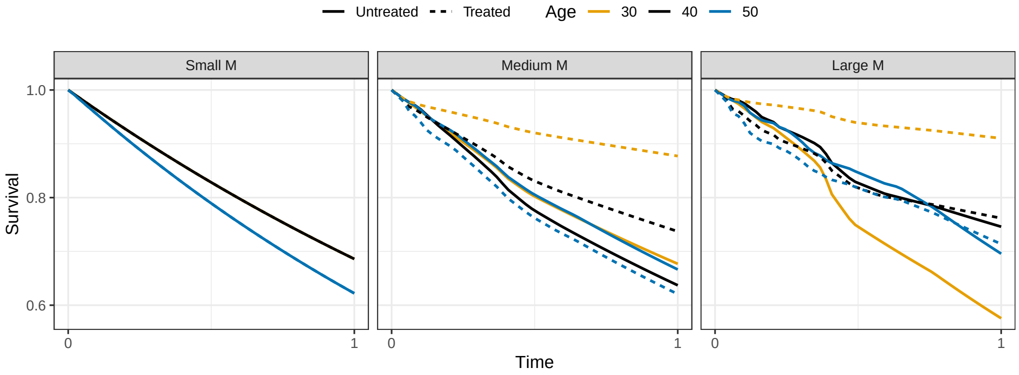

Figure 7: Estimates of the survival function derived from the HAL estimator stratified on treatment and three different age values based on a sample of 200 patients from the simulated study. Estimates are shown for three different values of the sectional variation norm (M).

Proposition 14 shows that a HAL estimator is well-defined and can be found as the solution to a convex optimization problem.

Proposition 14. *Define the set of indices $\mathcal{I} = \{\{1\} \cup s : s \subset \{2, \ldots, d\}\}$ and let*

$$g_{\beta,n}(\mathbf{x}) = \sum_{i=1}^{n} \sum_{r \in \mathcal{I}} \beta_i^r \mathbb{1}\{X_{r,i} \preceq \mathbf{x}_r\}, \quad \text{with} \quad \beta = \{\beta^r = (\beta_1^r, \ldots, \beta_n^r) : r \in \mathcal{I}\}.$$

The problem

$$\min_{\|\beta\|_1 \leq M} \mathbb{P}_n\left[\bar{L}(g_{\beta,n}, \cdot)\right], \quad \text{with} \quad \bar{L}(g, O) = \log\left(\int_0^1 e^{g(z,W)}\, \mathrm{d}z\right) - g(U, W), \tag{21}$$

is convex and has a solution. For any solution $\hat{\beta}$,

$$p_{\hat{\beta},n} \in \operatorname*{argmin}_{p \in \mathcal{P}^d_{M,n}} \mathbb{P}_n[-\log p],$$

where

$$\log p_{\hat{\beta},n}(u, \mathbf{w}) = g_{\hat{\beta},n}(u, \mathbf{w}) - \log\left(\int_0^1 e^{g_{\hat{\beta},n}(z,\mathbf{w})}\, \mathrm{d}z\right).$$

Proof. See Appendix C.2 □

Proposition 14 shows that the HAL estimator defined in equation (20) does not need to include basis functions that are only functions of w, so the number of basis functions is reduced to $|\mathcal{I}| = n2^{d-1}$.

We assume that $(U, W) \sim P$ for some distribution $P \ll \lambda$. For a distribution P, let p_P denote the conditional density of U given W and ω_P the marginal density of W with respect to λ.

Corollary 15. *Let P be a distribution such that $\varepsilon < \omega_P < 1/\varepsilon$, for some $\varepsilon > 0$, and $p_P \in \mathcal{P}_M^d$, and let $\hat{p}_n$ be a HAL estimator as defined in equation (20). If Assumption 7 holds when f^* is the minimizer of $f \mapsto P[\bar{L}(f, \cdot)]$ over $\mathcal{D}_M^d$, then*

$$\|\hat{p}_n - p_P\|_\lambda = o_P(n^{-1/3} \log(n)^{2(d-1)/3}).$$

Proof. Corollary 15 follows from Theorem 9. See Appendix C.2 for a detailed proof. □

A density can be obtained from a hazard function. This implies that an alternative density estimator can be constructed by first using the HAL estimator defined in Section 5 to estimate the corresponding log-hazard function and then transforming this into a density. We refer to this estimator as a 'HAL hazard parametrization' and to the estimator defined in equation (20) as a 'HAL density parametrization'. We compare these two estimators in Figure 8, where we have fitted both estimators to a simulated univariate dataset. The estimators are implemented using the convex optimization package `CVXR` in `R` [Fu et al., 2020] and the bounds M on the sectional variation norms are selected using cross-validation. We see that the estimator based on the hazard parametrization can exhibit erratic behavior at the end of the interval. The reason is that assuming a log-hazard function belongs to $\mathcal{D}_M^d$ implies that the corresponding density will not integrate to one. To see this, observe that the conditional survival function associated with a log-hazard function $f \in \mathcal{D}_M^d$ evaluated at $t = 1$ is

$$\exp\left\{-\int_0^1 e^{f(z,\mathbf{w})} \,\mathrm{d}z\right\} \geq \exp\left\{-e^M\right\} > 0.$$

Thus when the support of U is $[0, 1]$, the assumption that the log-hazard belongs to $\mathcal{D}_M^d$ will be wrong by definition for any $M < \infty$. We argue that the parametrization in equation (19) is better suited when U is known to have support in $[0, 1]$.

7 Least-squares regression

Let $O = (X, Y)$ for $X \in [0, 1]^d$ and $Y \in [-B, B]$ for some $B < \infty$, and define

$$f_P(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}], \quad \text{when} \quad (X, Y) \sim P.$$

In this section we consider estimation of f_P using the squared error loss

$$L^{\text{se}}(f, O) = (f(X) - Y)^2. \tag{22}$$

We here use ω_P to denote the Lebesgue density of X which we assume to exist.

Corollary 16. *Let P be a distribution such that $\varepsilon < \omega_P < 1/\varepsilon$, for some $\varepsilon > 0$, and $f_P \in \mathcal{D}_M^d$, and let $\hat{f}_n$ be a HAL estimator based on the squared error loss defined in equation (22). If Assumption 7 holds, then*

$$\|\hat{f}_n - f_P\|_\lambda = o_P(n^{-1/3} \log(n)^{2(d-1)/3}).$$

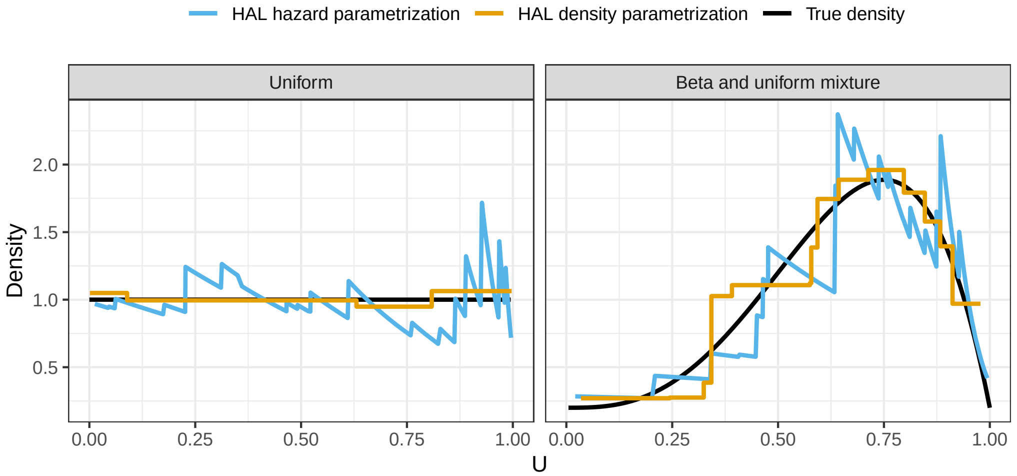

Figure 8: Two different density estimators under two different data-generating distributions. We generated 200 samples from a uniform distribution (left panel) and a mixture of a beta distribution and a uniform distribution (right panel). The 'HAL hazard parametrization' refers to a density estimator obtained from a HAL estimator of a hazard function, which was defined in Section 5. The 'HAL density parametrization' refers to the HAL estimator defined in equation (20).

Proof. We show that Assumption 6 holds for the squared error loss, and so Corollary 16 follows from Theorem 9. First note that because the squared error loss is a strictly proper scoring rule [Gneiting and Raftery, 2007], the assumption that $f_P \in \mathcal{D}_M^d$ implies that $f^* = f_P$ a.e. Conditions 6 (i)-(ii) hold by the definition of the squared error loss and the assumption that Y and ω_P are bounded. Condition 6 (iii) holds by Proposition 3 and Lemma 4 in Appendix B of [Bibaut and van der Laan, 2019]. □

For the squared error loss an empirical risk minimizer as defined in equation (5) exists. This was formally shown by Fang et al. [2021]. The authors also derive an algorithm for finding a collection of basis functions that is sufficient to construct an empirical risk minimizer. We illustrate the difference between the HAL estimator and the empirical risk minimizer by comparing the number of basis functions needed to calculate the two estimators for different sample sizes and dimensions. The results are shown in Figure 9. We see that a HAL estimator can be constructed using much fewer basis functions.

8 Discussion

Our main result relies on the smoothness assumption 7 (ii). However, another HAL estimator could be defined using a finer sieve than the one we have chosen, and we expect that Assumption 7 (ii) can be relaxed. An interesting question is how much we can reduce

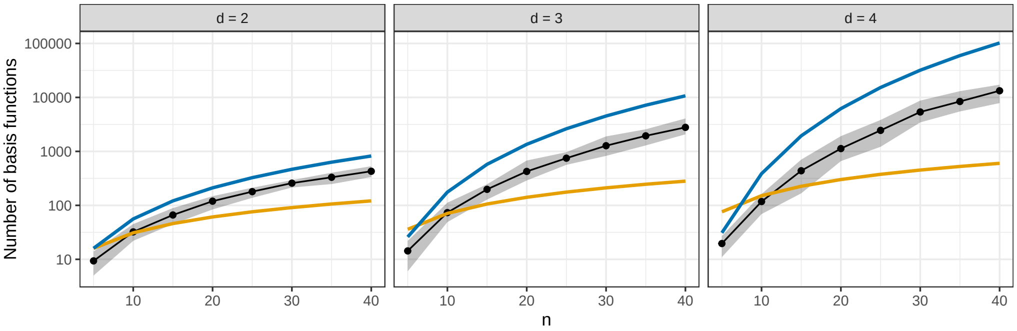

Figure 9: The black line is the average number of basis functions needed to calculate the empirical risk minimizer with ribbons denoting the 2.5%- and 97.5%-quantiles based on 200 simulations of uniformly distributed covariates. The number of observations is denoted by n and the dimension by d. The blue line is a deterministic upper bound on this number (see Lemma 3.5 of Fang et al. [2021]). The orange line is the number of basis functions needed to calculate a HAL estimator.

the number of basis functions and still achieve the same rate of convergence, and whether we need to impose additional smoothness assumptions for this to hold.

Throughout this paper we have stated that an empirical risk minimizer does not exist or is inconsistent when a density or a hazard function is estimated. Formally, our Proposition 10 does not rule out, however, that a consistent empirical risk minimizer can exist in the special case that the data-generating hazard function is non-decreasing. If the data-generating hazard function is believed to be monotone, it is natural to use shape-constrained estimators [Groeneboom and Jongbloed, 2014]. An interesting direction for future research is to investigate HAL-like estimators when biologically motivated monotonicity constraints are imposed.

A Càdlàg functions and measures

To prove the results from Section 2, we start by proving the following two lemmas.

Lemma 17. *For a function $f\colon [0,1]^d \to \mathbb{R}$ and a sequence of functions $f_n\colon [0,1]^d \to \mathbb{R}$, $n \in \mathbb{N}$, assume that $\|f_n - f\|_\infty \to 0$ when $n \to \infty$. If $f_n \in \mathcal{D}_M^d$ for all $n \in \mathbb{N}$ then $f \in \mathcal{D}_M^d$.*

Proof. Neuhaus [1971] shows that the uniform limit of a sequence of càdlàg functions is also càdlàg. It thus only remains so be shown that $\|f\|_v \leq M$. Assume for contradiction that this is not the case. We thus assume that $\|f\|_v > M + \varepsilon$ for some $\varepsilon > 0$, which by definition means that there must exist finite partitions $\mathcal{P}_s$ of all faces $(\mathbf{0}_s, \mathbf{1}_s]$, $\emptyset \neq s \subset [d]$ such that

$$\sum_{s \in \mathcal{S}} \sum_{A \in \mathcal{P}_s} |\Delta(f; A)| > M + \varepsilon, \quad \text{where} \quad \mathcal{S} = \{s \subset [d] : s \neq \emptyset\}$$

The sum above is made up of $\kappa = \sum_s |\mathcal{P}_s| 2^{|s|} < \infty$ number of terms on the form $\pm f(\mathbf{x})$ for some $\mathbf{x} \in [0,1]^d$. By assumption we can find $n_0 \in \mathbb{N}$ such that $\|f_n - f\|_\infty < \varepsilon/\kappa$ for all $n \geq n_0$, and thus

$$M < \sum_{s \in \mathcal{S}} \sum_{A \in \mathcal{P}_s} |\Delta(f_n; A)| \leq \|f_n\|_v, \quad \forall n > n_0.$$

This contradicts the fact that $f_n \in \mathcal{D}^d_M$ for all $n \in \mathbb{N}$, so we must have $\|f\|_v \leq M$. □

Lemma 18. *Let f be a function that is right-continuous in each of its coordinates with $\|f\|_v \leq M$. There exists a sequence $\{f_n\} \subset \mathcal{R}^d_M$ such that $\|f - f_n\|_\infty \to 0$ for $n \to \infty$.*

Proof. By Theorem 3 (a) in [Aistleitner and Dick, 2015] there exists a unique, finite signed measure μ_f such that $f(\mathbf{x}) = \mu([\mathbf{0}, \mathbf{x}])$. By the Jordan-Hahn decomposition theorem we may write $\mu_f = \alpha P^+ - \beta P^-$, where P^+ and P^- are uniquely determined probability measures with $P^+ \perp P^-$, and $\alpha, \beta \in [0, \infty)$. Letting F^+ and F^- denote the associated cumulative distribution functions, we have that $f = \alpha F^+ - \beta F^-$. By Theorem 3 (a) in [Aistleitner and Dick, 2015] and because $P^+ \perp P^-$ we have

$$M \geq \|f\|_v = \|\mu_f\|_{\mathrm{TV}} = \alpha \|P^+\|_{\mathrm{TV}} + \beta \|P^-\|_{\mathrm{TV}} = \alpha + \beta. \tag{23}$$

Let P^+_n and P^-_n denote the empirical measures obtained from i.i.d. samples from P^+ and P^-, respectively. Let F^+_n and F^-_n denote the associated empirical distribution functions, and define $F_n = \alpha F^+_n - \beta F^-_n$. As $P^+_n \perp P^-_n$ almost surely we have

$$\|F_n\|_v = \alpha \|P^+_n\|_{\mathrm{TV}} + \beta \|P^-_n\|_{\mathrm{TV}} = \alpha + \beta \quad \text{a.s.}$$

The multivariate version of the Dvoretzky-Kiefer-Wolfowitz theorem [Dvoretzky et al., 1956, Naaman, 2021] and the Borel-Cantelli lemma imply that $\|F^+_n - F^+\|_\infty \to 0$ and $\|F^-_n - F^-\|_\infty \to 0$ almost surely. Hence there must exist deterministic sequences of discrete measures p^+_n and p^-_n with associated cumulative distribution functions f^+_n and f^-_n such that

$$p^+_n \perp p^-_n, \quad \forall n \in \mathbb{N}, \tag{24}$$

and

$$\|f^+_n - F^+\|_\infty \longrightarrow 0 \quad \text{and} \quad \|f^-_n - F^-\|_\infty \longrightarrow 0. \tag{25}$$

Note that f^+_n is a linear combination of the indicator functions $\{\mathbb{1}_{[\mathbf{x}_i, \mathbf{1}]}\}^n_{i=1}$, where $\{\mathbf{x}_i\}^n_{i=1}$ are the support points of the discrete measure p^+_n, and similarly for f^-_n. Hence, with $f_n = \alpha f^+_n - \beta f^-_n$, we have that $f_n \in \mathrm{Span}(\mathcal{F}^d)$. By equations (23) and (24),

$$\|f_n\|_v = \alpha + \beta \leq M,$$

so $f_n \in \mathcal{R}^d_M$ for all $n \in \mathbb{N}$. Equation (25) gives that $\|f_n - f\|_\infty \to 0$ which concludes the proof. □

Proof of Proposition 2. For $f_1, f_2 \in \mathcal{D}^d$ and $\alpha, \beta \in \mathbb{R}$ the function $f = \alpha f_1 + \beta f_2$ is càdlàg, so $\mathcal{R}^d_M \subset \mathcal{D}^d_M$. It thus follows from Lemma 17 that $\overline{\mathcal{R}^d_M} \subset \mathcal{D}^d_M$. As any $f \in \mathcal{D}^d_M$ is right-continuous in each of its coordinates, the reverse inclusion follows from Lemma 18. □

Proof of Proposition 3. Any function $f \in \mathcal{D}_M^d$ is by definition right-continuous in each of its arguments so the first statement follows immediately from Theorem 3 (a) in [Aistleitner and Dick, 2015]. For the second statement, we know by Theorem 3 (b) in [Aistleitner and Dick, 2015] that there exists a right-continuous function f_μ with $\|f_\mu\|_v = M < \infty$ such that $f_\mu(\mathbf{x}) = \mu([\mathbf{0}, \mathbf{x}])$. By Lemma 18, f_μ can be approximated uniformly by a sequence of functions $f_n \in \mathcal{D}_M^d$. Lemma 17 then implies that $f_\mu \in \mathcal{D}_M^d$. □

Proof of Proposition 4. For the first statement we use that we can partition a box $[\mathbf{0}, \mathbf{x}]$ into 'half-closed' lower dimensional faces with corners at $\mathbf{0}$, the point $\mathbf{0}$, and the remaining 'half-closed interior' of the box, i.e.,

$$[\mathbf{0}, \mathbf{x}] = \{\mathbf{0}\} \cup \Big(\bigcup_{s \in \mathcal{S}} A(\mathbf{x}; s), \Big), \quad \text{for} \quad A(\mathbf{x}; s) = A_1(\mathbf{x}; s) \times \cdots \times A_d(\mathbf{x}; s),$$

where

$$\mathcal{S} = \{ s \subset \{1, \ldots, d\} : s \neq \emptyset \}, \quad \text{and} \quad A_i(\mathbf{x}; s) = \begin{cases} (0, x_i] & \text{if } i \in s \\ \{0\} & \text{if } i \notin s \end{cases},$$

and we define $(0, 0] = \emptyset$. Using this and Proposition 3 we can write

$$f(\mathbf{x}) = \mu_f([\mathbf{0}, \mathbf{x}]) = \mu_f(\{\mathbf{0}\}) + \sum_{s \in \mathcal{S}} \mu_f(A(\mathbf{x}; s)) \tag{26}$$

Any section f_s of f is also a càdlàg function with bounded sectional variation norm and hence generates a measure on the cube $[0,1]^{|s|}$ through the relation

$$f_s(\mathbf{x}) = \mu_{f_s}([\mathbf{0}_s, \mathbf{x}]), \quad \text{for all} \quad \mathbf{x} \in [0,1]^{|s|}. \tag{27}$$

By definition of the section f_s it follows that the measure assigned to a box in $[0,1]^{|s|}$ by μ_{f_s} is the same as the measure assigned by μ_f when this space is considered as a subspace of $[0,1]^d$, i.e.,

$$\mu_{f_s}([\mathbf{0}_s, \mathbf{x}_s]) = \mu_f([\mathbf{0}, \overline{\mathbf{x}}_s]), \quad \text{for} \quad \mathbf{x} \in [0,1]^d.$$

By the uniqueness of the measures generated by f and each f_s it follows that

$$\mu_f(A(\mathbf{x}; s)) = \mu_{f_s}((\mathbf{0}_s, \mathbf{x}_s]). \tag{28}$$

By equations (26) and (28) we then have

$$f(\mathbf{x}) = f(\mathbf{0}) + \sum_{s \in \mathcal{S}} \mu_{f_s}((\mathbf{0}_s, \mathbf{x}_s]) = f(\mathbf{0}) + \sum_{s \in \mathcal{S}} \int_{(\mathbf{0}_s, \mathbf{x}_s]} \mathrm{d} f_s.$$

The second statement follows because $A(\mathbf{1}; s)$, for $s \in \mathcal{S}$, are disjoint sets, so the measures

$\mathbb{1}_{A(\mathbf{1};s)} \cdot \mu_f$ are mutually singular. Hence,

$$\begin{aligned}
\|\mu_f\|_{\mathrm{TV}} &= \left\| \mathbb{1}_{\{\mathbf{0}\}} \cdot \mu_f + \sum_{s\in\mathcal{S}} \mathbb{1}_{A(\mathbf{1};s)} \cdot \mu_f \right\|_{\mathrm{TV}} \\
&= \left\|\mathbb{1}_{\{\mathbf{0}\}} \cdot \mu_f\right\|_{\mathrm{TV}} + \sum_{s\in\mathcal{S}} \left\|\mathbb{1}_{A(\mathbf{1};s)} \cdot \mu_f\right\|_{\mathrm{TV}} \\
&= \int_{\{\mathbf{0}\}} \mathrm{d}|\mu_f| + \sum_{s\in\mathcal{S}} \int_{A(\mathbf{1};s)} \mathrm{d}|\mu_f| \\
&= |f(\mathbf{0})| + \int_{(\mathbf{0}_s,\mathbf{1}_s]} |\,\mathrm{d}f_s|.
\end{aligned}$$

□

Proof of Proposition 5. Let $B_r(\mathbf{x})$ be the ball around the point $\mathbf{x} \in [0,1]^d$ with radius $r > 0$. For a function $f : [0,1]^d \to \mathcal{K} \subset \mathbb{R}$ with $\mathcal{K}$ finite, we now claim that

$$\forall \mathbf{x} \in [0,1]^d, \forall \mathbf{a} \in \{0,1\}^d, \exists r > 0, \forall z, y \in B_r(\mathbf{x}) \cap Q_{\mathbf{a}}(\mathbf{x}) : f(z) = f(y), \qquad (*)$$

implies $f \in \mathcal{R}^d_M$. To see this, assume that $(*)$ holds. Define the covering

$$\mathcal{B} = \{B(\mathbf{x}) : \mathbf{x} \in [0,1]^d\},$$

where $B(\mathbf{x})$ is an open ball around $\mathbf{x}$ such that for any $\mathbf{a} \in \{0,1\}^d$, f is constant on $B_{r_{\mathbf{x}}}(\mathbf{x}) \cap Q_{\mathbf{a}}(\mathbf{x})$ for some $r_{\mathbf{x}} > 0$. Such an open ball exists around any $\mathbf{x}$ by $(*)$. As $[0,1]^d$ is compact there exists a finite subset $\{B(\mathbf{x}^1), \ldots, B(\mathbf{x}^J)\} \subset \mathcal{B}$ that covers $[0,1]^d$. Consider now any box of the form

$$I(\mathbf{j}) = I_1(j_1) \times \cdots I_d(j_d), \quad \text{where} \quad I_i(j) = \begin{cases} [0, x_i^1) & \text{if } j = 1, \\ [x_i^j, x_i^{j+1}) & \text{if } 0 < j < J, \\ [x_i^J 1] & \text{if } j = J, \end{cases}$$

for all unique sequences $\mathbf{j} = (j_1, \ldots, j_d) \in \{1, \ldots, J\}^d$. These boxes partition $[0,1]^d$, and by construction of the covering $\{B(\mathbf{x}^1), \ldots, B(\mathbf{x}^J)\}$, f is constant on $B(\mathbf{x}^j) \cap I(\mathbf{j})$ for all j and $\mathbf{j}$. As any $I(\mathbf{j})$ is connected and $\{B(\mathbf{x}^1), \ldots, B(\mathbf{x}^J)\}$ is an open cover, it follows that f is constant on each $I(\mathbf{j})$. Hence $f \in \mathcal{R}^d_M$, and thus we have proved the initial claim. The proposition now follows by noting that this implies that if $f \notin \mathcal{R}^d_M$, then $(*)$ is false, i.e.,

$$\exists \mathbf{x} \in [0,1]^d, \exists \mathbf{a} \in \{0,1\}^d, \forall r > 0, \exists z, y \in B_r(\mathbf{x}) \cap Q_{\mathbf{a}}(\mathbf{x}) : f(z) \neq f(y).$$

Thus, if $f \notin \mathcal{R}^d_M$, we can find a point $\mathbf{x} \in [0,1]^d$, a vertex $\mathbf{a} \in \{0,1\}^d$ and a sequence $r_n \searrow 0$ such that for all $n \in \mathbb{N}$, f is not constant on $B_{r_n}(\mathbf{x}) \cap Q_{\mathbf{a}}(\mathbf{x})$. This in turn implies that we can find a sequence $\{\mathbf{x}_n\} \in B_{r_n}(\mathbf{x}) \cap Q_{\mathbf{a}}(\mathbf{x})$ such that $f(\mathbf{x}_n) \neq f(\mathbf{x}_{n-1})$. Clearly, $\mathbf{x}_n \in Q_{\mathbf{a}}(\mathbf{x})$ and $\mathbf{x}_n \to \mathbf{x}$, but as $f(\mathbf{x}) \in \mathcal{K}$ for all $\mathbf{x} \in [0,1]^d$, $f(\mathbf{x}_n)$ cannot converge. Hence f is not càdlàg. □

B Results from empirircal process theory

Recall the notation $\mathcal{L}_M = \left\{L(f,\cdot) : f \in \mathcal{D}_M^d\right\}$ for a loss function $L\colon \mathcal{D}_M^d \times \mathcal{O} \to \mathbb{R}_+$, and the Assumption 6 (iii), which we restate her for convenience:

$$\exists C < \infty, \eta > 0, \kappa \in \mathbb{N} : N_{[\,]}(\varepsilon, \mathcal{L}_M, \|\cdot\|_P) \leq C N_{[\,]}(\varepsilon/C, \mathcal{D}_M^d, \|\cdot\|_\lambda)^\kappa, \quad \forall \varepsilon \in (0,\eta). \tag{B}$$

Define

$$\Gamma_n(\delta) = \sup_{\|f-f^*\|_\lambda<\delta} |\mathbb{G}_n[L(f,\cdot) - L(f^*,\cdot)]| \quad \text{with} \quad f \in \mathcal{D}_M^d, \tag{29}$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ is the empirical process.

Lemma 19. *If (B) holds and $\|L(f,\cdot)\|_\infty < C$, then there exists an $\eta > 0$ such that for for all $n \in \mathbb{N}$ and $\delta \in (0,\eta)$,*

$$\mathbb{E}_P^*\left[\Gamma_n(\delta)\right] \lesssim \delta^{1/2}|\log(\delta)|^{d-1} + \frac{|\log(\delta)|^{2(d-1)}}{\delta\sqrt{n}}.$$

In particular, when $r_n = n^{1/3}\log(n)^{-2(d-1)/3}$ we have

$$n^{-1/2}\,\mathbb{E}_P^*\left[\Gamma_n(r_n)\right] = O(r_n^{-2}).$$

Proof. Define the entropy integral

$$J_{[\,]}(\delta, \mathcal{H}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[\,]}(\varepsilon, \mathcal{H}, \|\cdot\|)}\,\mathrm{d}\varepsilon.$$

Lemma 3.4.2 in van der Vaart and Wellner [1996] provides the bound

$$\mathbb{E}_P^*\left[\Gamma_n(\delta)\right] \lesssim J_{[\,]}(\delta, \mathcal{L}_M, \|\cdot\|_P)\left(1 + \frac{J_{[\,]}(\delta, \mathcal{L}_M, \|\cdot\|_P)}{\delta^2\sqrt{n}}C\right). \tag{30}$$

Bibaut and van der Laan [2019] established that

$$\log N_{[\,]}(\varepsilon, \mathcal{D}_M^d, \|\cdot\|_\lambda) \lesssim \varepsilon^{-1}|\log(\varepsilon/M)|^{2(d-1)},$$

for $\varepsilon \in (0,1)$, and so we have by assumption

$$\log N_{[\,]}(\varepsilon, \mathcal{L}_M, \|\cdot\|_P) \leq \log C + \kappa \log N_{[\,]}(\varepsilon/C, \mathcal{D}_M^d, \|\cdot\|_\lambda) \lesssim \varepsilon^{-1}|\log(\varepsilon)|^{2(d-1)}, \tag{31}$$

for small enough ε. Using integration by parts we have

$$\begin{aligned}\int_0^\delta \sqrt{\varepsilon^{-1}|\log\varepsilon|^{2(d-1)}}\,\mathrm{d}\varepsilon &= (-1)^{d-1}\int_0^\delta \varepsilon^{-1/2}(\log\varepsilon)^{d-1}\,\mathrm{d}\varepsilon \\ &= (-1)^{d-1}\left(\delta^{1/2}(\log\delta)^{d-1} - (d-1)\int_0^\delta \varepsilon^{1/2}(\log\varepsilon)^{d-2}\varepsilon^{-1}\,\mathrm{d}\varepsilon\right) \\ &= \delta^{1/2}|\log\delta|^{d-1} + (d-1)\int_0^\delta \varepsilon^{-1/2}|\log\varepsilon|^{d-2}\,\mathrm{d}\varepsilon.\end{aligned}$$

As the second term on the right vanishes for $\delta \to 0$, we can use this and equation (31) to obtain

$$J_{[\,]}(\delta, \mathcal{L}_M, \|\cdot\|_P) \lesssim \delta^{1/2} |\log \delta|^{d-1},$$

and so equation (30) gives

$$\mathbb{E}_P^* \left[\Gamma_n(\delta)\right] \lesssim \delta^{1/2} |\log \delta|^{d-1} \left(1 + \frac{\delta^{1/2}|\log \delta|^{d-1}}{\delta^2 \sqrt{n}} M\right) \lesssim \delta^{1/2}|\log \delta|^{d-1} + \frac{|\log \delta|^{2(d-1)}}{\delta \sqrt{n}},$$

which was the first statement of the lemma. For the second statement, set $\delta = r_n^{-1}$ and obtain for all $n \geq 3$,

$$\begin{aligned}
n^{-1/2} r_n^2 \, \mathbb{E}\left[\Gamma_n(r_n^{-1})\right] &\lesssim n^{-1/2} r_n^2 \left(r_n^{-1/2} |\log(r_n)|^{d-1} + \frac{r_n |\log(r_n)|^{2(d-1)}}{\sqrt{n}}\right) \\
&\leq n^{-1/2} r_n^2 \left(r_n^{-1/2} |\log(n)|^{d-1} + \frac{r_n |\log(n)|^{2(d-1)}}{\sqrt{n}}\right) \\
&= n^{-1/2} r_n^2 \left(n^{-1/6} |\log(n)|^{4(d-1)/3} + \frac{n^{1/3} |\log(n)|^{4(d-1)/3}}{\sqrt{n}}\right) \\
&= n^{-1/2} r_n^2 \left(n^{-1/6} |\log(n)|^{4(d-1)/3} + n^{-1/6} |\log(n)|^{4(d-1)/3}\right) \\
&= n^{1/6} |\log(n)|^{-4(d-1)/3} 2 n^{-1/6} |\log(n)|^{4(d-1)/3} \\
&= 2
\end{aligned}$$

□

Proof of Theorem 9. We apply theorem 3.4.1 from van der Vaart and Wellner [1996] to a HAL estimator $\hat{f}_n$. This yields that $\hat{f}_n$ converges to $\hat{\pi}_n(f^*)$ at rate r_n if there exist numbers $0 \leq \delta_n < \eta$ such that the following conditions hold for all $n \in \mathbb{N}$ and $\delta \in (\delta_n, \eta)$.

(C1) Define for $0 < a < b$ the hollow sphere $B_{(a,b)}(f_0) = \{f \in \mathcal{D}_M^d : a < \|f - f_0\| < b\}$. It holds that

$$\inf_{f \in B_{(\delta/2,\delta)}(\hat{\pi}_n(f^*))} P[L(f, \cdot) - L(\hat{\pi}_n(f), \cdot)] \gtrsim \delta^2.$$

(C2) There exists a function $\varphi_n \colon (\delta_n, \eta) \to \mathbb{R}$ such that $\delta \mapsto \varphi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and

$$\mathbb{E}^* \left[\Gamma_n(\delta)\right] \lesssim \varphi_n(\delta) \quad \text{and} \quad n^{-1/2} \varphi_n(r_n^{-1}) \leq r_n^{-2},$$

where $\mathbb{E}^*$ denotes outer expectation.

(C3) $\mathbb{P}_n[L(\hat{f}_n, \cdot)] \leq \mathbb{P}_n[L(\hat{\pi}_n(f^*), \cdot)] + O_P(r_n^{-2})$.

(C4) $\|\hat{f}_n - \hat{\pi}_n(f^*)\|_\lambda \xrightarrow{P^*} 0$, where P^* denotes outer probability.

By Assumption 6 (ii) there exists $C \in (1, \infty)$ such that

$$\frac{1}{C}\|f - f^*\|^2 \leq P[L(f, \cdot) - L(f^*, \cdot)] \leq C\|f - f^*\|^2.$$

Set $\delta_n = 4C^2\|\hat{\pi}_n(f^*) - f^*\|$. If $f \in B_{(\delta/2,\delta)}(\hat{\pi}_n(f^*))$ we have that

$$\delta/2 < \|f - \hat{\pi}_n(f^*)\| \leq \|f - f^*\| + \|f^* - \hat{\pi}_n(f^*)\| = \|f - f^*\| + \frac{\delta_n}{4C^2} < \|f - f^*\| + \delta/4,$$

so

$$\|f - f^*\| > \delta/4. \tag{32}$$

Equation (32) shows that $B_{(\delta/2,\delta)}(\hat{\pi}_n(f^*)) \subset B_{(\delta/4,\infty)}(f^*)$, so

$$\inf_{f \in B_{(\delta/2,\delta)}(\hat{\pi}_n(f^*))} P[L(f,\cdot) - L(\hat{\pi}_n(f),\cdot)] \geq \inf_{f \in B_{(\delta/4,\infty)}(f^*)} P[L(f,\cdot) - L(\hat{\pi}_n(f),\cdot)]. \tag{33}$$

As

$$P[L(\hat{\pi}_n(f),\cdot) - L(f^*,\cdot)] \leq C\|f^* - \hat{\pi}_n(f)\|^2 = \frac{C\delta_n^2}{16C^4} = \frac{\delta_n^2}{16C^3} < \frac{\delta^2}{16C^3},$$

we have

$$\begin{aligned} P[L(f,\cdot) - L(\hat{\pi}_n(f),\cdot)] &= P[L(f,\cdot) - L(f^*,\cdot)] - P[L(\hat{\pi}_n(f),\cdot) - L(f^*,\cdot)]. \\ &> P[L(f^*,\cdot) - L(f,\cdot)] - \frac{\delta^2}{16C^3} \\ &> \frac{1}{C}\|f^* - f\|^2 - \frac{\delta^2}{16C^3}, \end{aligned}$$

so when $f \in B_{(\delta/4,\infty)}(f^*)$, we have

$$P[L(f,\cdot) - L(\hat{\pi}_n(f),\cdot)] > \frac{1}{C}\frac{\delta^2}{16} - \frac{\delta^2}{16C^3} > \delta^2 \frac{1}{16C}\left(1 - \frac{1}{C^3}\right) \gtrsim \delta^2.$$

Together with equation (33) this shows that condition (C1) holds. Because of Assumptions 6 (i) and Assumptions 6 (iii), condition (C2) holds by Lemma 19. By definition of $\hat{f}_n$ and $\hat{\pi}_n(f^*)$, we have

$$\mathbb{P}_n[L(\hat{f}_n,\cdot)] \leq \mathbb{P}_n[L(\hat{\pi}_n(f^*),\cdot)], \tag{34}$$

so condition (C3) is trivially true. We now show that condition (C4) holds. By Assumption 6 (ii) and equation (34) we can write

$$\begin{aligned} \frac{1}{C}\|f^* - \hat{f}_n\|_\lambda^2 &\leq P[L(\hat{f}_n,\cdot) - L(f^*,\cdot)] \\ &= (P - \mathbb{P}_n)[L(\hat{f}_n,\cdot) - L(f^*,\cdot)] + \mathbb{P}_n[L(\hat{f}_n,\cdot) - L(f^*,\cdot)] \\ &\leq (P - \mathbb{P}_n)[L(\hat{f}_n,\cdot) - L(f^*,\cdot)] + \mathbb{P}_n[L(\hat{\pi}_n(f^*),\cdot) - L(f^*,\cdot)] \\ &= (P - \mathbb{P}_n)[L(\hat{f}_n,\cdot) - L(f^*,\cdot)] + P[L(\hat{\pi}_n(f^*),\cdot) - L(f^*,\cdot)] \\ &\quad - (P - \mathbb{P}_n)[L(\hat{\pi}_n(f^*),\cdot) - L(f^*,\cdot)] \\ &\leq 4 \sup_{L \in \mathcal{L}_M} |(\mathbb{P}_n - P)[L]| + P[L(\hat{\pi}_n(f^*),\cdot) - L(f^*,\cdot)]. \end{aligned}$$

By Assumption 6 (ii) and Lemma 8, the second term on the right hand side converges to zero in probability. Proposition 1 in [Bibaut and van der Laan, 2019] and Theorem 2.4.1 in [van der Vaart and Wellner, 1996] together with Assumption 6 (iii) imply that $\mathcal{L}_M$ is a Glivenko-Cantelli class of functions. This implies that also the first term on the right converges to zero in probability, so $\|f^* - \hat{f}_n\|_\lambda \xrightarrow{P^*} 0$. By Lemma 8, we also have $\|f^* - \hat{\pi}_n(f^*)\|_\lambda \xrightarrow{P^*} 0$, so this implies condition (C4). □

C Additional proofs

C.1 Right-censored data

Proof of Proposition 10. Let $f^\circ \in \mathcal{D}_M^1$ be a function and $j \in \{1, \dots, n-1\}$ an index such that $f^\circ(\tilde{T}_{(j)}) > f^\circ(\tilde{T}_{(j+1)})$. We shall construct a function $\check{f} \in \mathcal{D}_M^1$ such that $\mathbb{P}_n[L^{\text{pl}}(\check{f}, \cdot)] < \mathbb{P}_n[L^{\text{pl}}(f^\circ, \cdot)]$ when L^{pl} is the negative log-likelihood defined in equation (16). This implies that f° cannot be the minimizer of the empirical risk over $\mathcal{D}_M^1$. To find $\check{f}$ we first define

$$V = \inf_{u \in [\tilde{T}_{(j)}, \tilde{T}_{(j+1)}]} f^\circ(u),$$

and

$$f_\varepsilon(t) = \mathbb{1}\{t \in [\tilde{T}_{(j)} + \varepsilon, \tilde{T}_{(j+1)})\} V + \mathbb{1}\{t \notin [\tilde{T}_{(j)} + \varepsilon, \tilde{T}_{(j+1)})\} f^\circ(t),$$

for $\varepsilon \in (0, [\tilde{T}_{(j+1)} - \tilde{T}_{(j)}]/2)$. In words, f_ε is identical to f°, except on the interval $[\tilde{T}_{(j)} + \varepsilon, \tilde{T}_{(j+1)})$ where it is constant and equals V. Note that by the assumption that $f^\circ(\tilde{T}_{(j)}) > f^\circ(\tilde{T}_{(j+1)})$ we must have

$$f^\circ(\tilde{T}_{(j)}) > V. \tag{35}$$

As f° is càdlàg, so is f_ε, and as f_ε does not fluctuate more than f°, we must have $\|f_\varepsilon\|_v \leq \|f^\circ\|_v$. Thus $f_\varepsilon \in \mathcal{D}_M^1$. Now, by equation (35) and because f° is continuous from the right, we can find a $\delta > 0$ and an $\varepsilon_0 > 0$ such that $f^\circ(t) > V + \delta$ for all $t \in [\tilde{T}_{(j)}, \tilde{T}_{(j)} + \varepsilon_0]$. Thus, if we define $\check{f} = f_{\varepsilon_0/2}$ and $\mathcal{I} = (\tilde{T}_{(j)} + \varepsilon_0/2, \tilde{T}_{(j)} + \varepsilon_0)$, this implies that $\check{f}(t) \leq f^\circ(t)$ for all $t \in [0,1]$ and $\check{f}(t) < f^\circ(t) - \delta$ for $t \in \mathcal{I}$. This in turn implies that

$$\int_0^{\tilde{T}_i} e^{\check{f}(u)} \, \mathrm{d}u = \int_0^{\tilde{T}_i} e^{f^\circ(u)} \, \mathrm{d}u, \quad \text{for all} \quad i \leq j, \tag{36}$$

and

$$\int_0^{\tilde{T}_i} e^{\check{f}(u)} \, \mathrm{d}u < \int_0^{\tilde{T}_i} e^{f^\circ(u)} \, \mathrm{d}u, \quad \text{for all} \quad i > j. \tag{37}$$

Finally, we have by construction that

$$\check{f}(\tilde{T}_i) = f^\circ(\tilde{T}_i) \quad \text{for all} \quad i \in \{1, \dots, n\}. \tag{38}$$

Equations (36)-(38) together imply that $\mathbb{P}_n[L^{\text{pl}}(\check{f}, \cdot)] < \mathbb{P}_n[L^{\text{pl}}(f^\circ, \cdot)]$. □

Proof of Proposition 11. Let $\mathcal{B}_M = \{\beta \in \mathbb{R}^{m(d,n)} : \|\beta\|_1 \leq M\}$. By construction, for any $\mathbf{w}$ and $\beta \in \mathcal{B}_M$ the map $s \mapsto f_{\beta,n}(s, \mathbf{w})$ is constant on $[\tilde{T}_{(j-1)}, \tilde{T}_{(j)})$ for all $j = 1, \dots, n'$, and thus we can write

$$\int_0^{\tilde{T}_i} e^{f_{\beta,n}(s, W_i)} \, \mathrm{d}s = \sum_{j=1}^{n'} \mathbb{1}\{\tilde{T}_i \geq \tilde{T}_{(j-1)}\}(\tilde{T}_{(j)} \wedge \tilde{T}_i - \tilde{T}_{(j-1)}) e^{f_{\beta,n}(\tilde{T}_{(j-1)}, W_i)}. \tag{39}$$

For any t and $\mathbf{w}$, the map $\beta \mapsto f_{\beta,n}(t, \mathbf{w})$ is linear and as $z \mapsto e^z$ is convex and non-decreasing it follows that $\beta \mapsto e^{f_{\beta,n}(t,\mathbf{w})}$ is convex [Boyd and Vandenberghe, 2004, Section 3.2.4]. Thus equation (39) implies that the map $\beta \mapsto \mathbb{P}_n[L^{\text{pl}}(f_{\beta,n}, \cdot)]$ is convex, and as $\mathcal{B}_M$ is convex it follows that the problem in (17) is convex. The minimum is attained because the map $\beta \mapsto \mathbb{P}_n[L^{\mathrm{pl}}(f_{\beta,n}, \cdot)]$ is continuous and $\mathcal{B}_M$ is compact. □

Lemma 20. *Let $T \in [0, 1+\varepsilon]$ for some $\varepsilon > 0$, $W \in [0,1]^{d-1}$, and assume that the conditional distribution of $T \mid W = w$ has a Lebesgue density for all $w \in [0,1]^{d-1}$. For q a conditional density function for T given W, let h_q the associated hazard function and S_q the associated survival function. Let $\mathcal{Q}_M = \{q : \sup_{t\in[0,1]} \|h_q(t,\cdot)\|_\infty \leq M\}$ for some $M < \infty$. Let λ denote Lebesgue measure on $[0,1]^d$. The following holds for all $q, p \in \mathcal{Q}_M$.*

(i) $\|S_q - S_p\|_\lambda \leq \|q-p\|_\lambda$ *and* $\|S_q - S_p\|_\lambda \leq \|h_q - h_p\|_\lambda$.

(ii) $\|h_p - h_q\|_\lambda \leq (e^M + Me^{2M})\|p-q\|_\lambda$ *and* $\|p-q\|_\lambda \leq (e^M + M)\|h_p - h_q\|_\lambda$.

Proof of Lemma 20. The first inequality in statement (i) follows from Jensen's inequality:

$$\begin{aligned}
\|S_q - S_p\|_\lambda^2 &= \int_{[0,1]^{d-1}} \int_{[0,1]} \{[S_q - S_p](s,\mathbf{w})\}^2 \,\mathrm{d}s\,\mathrm{d}\mathbf{w} \\
&= \int_{[0,1]^{d-1}} \int_{[0,1]} \left\{\int_0^s [p-q](u,\mathbf{w})\,\mathrm{d}u\right\}^2 \mathrm{d}s\,\mathrm{d}\mathbf{w} \\
&= \int_{[0,1]^{d-1}} \int_{[0,1]} s^2 \left\{\int_0^s [p-q](u,\mathbf{w})\frac{\mathrm{d}u}{s}\right\}^2 \mathrm{d}s\,\mathrm{d}\mathbf{w} \\
&\leq \int_{[0,1]^{d-1}} \int_{[0,1]} s^2 \int_0^s \{[p-q](u,\mathbf{w})\}^2 \frac{\mathrm{d}u}{s}\,\mathrm{d}s\,\mathrm{d}\mathbf{w} \qquad (40)\\
&\leq \int_{[0,1]^{d-1}} \int_{[0,1]} \int_0^s \{[p-q](u,\mathbf{w})\}^2 \,\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}\mathbf{w} \\
&\leq \int_{[0,1]^{d-1}} \int_{[0,1]} \int_0^1 \{[p-q](u,\mathbf{w})\}^2 \,\mathrm{d}u\,\mathrm{d}s\,\mathrm{d}\mathbf{w} \\
&= \|q-p\|_\lambda^2.
\end{aligned}$$

For the second inequality in statement (i), let H_q denote the conditional cumulative hazard function associated with the density q. By the mean value theorem we may write

$$e^{-H_q} - e^{-H_p} = e^{-H_{q,p}}(H_p - H_q),$$

for some positive function $H_{q,p}$. Hence by Hölder's inequality

$$\|S_q - S_p\|_\lambda^2 \leq \|H_p - H_q\|_\lambda^2 = \int_{[0,1]^{d-1}} \int_{[0,1]} \left\{\int_0^s [h_p - h_q](u,\mathbf{w})\,\mathrm{d}u\right\}^2 \mathrm{d}s\,\mathrm{d}\mathbf{w},$$

and so Jensen's inequality (in the same way as in equation (40)) gives that

$$\|S_q - S_p\|_\lambda^2 \leq \|h_q - h_p\|_\lambda^2 .$$

This shows statement (i). To obtain the first inequality in statement (ii) we write

$$\begin{aligned}
\|h_q - h_p\|_\lambda = \left\|\frac{q}{S_q} - \frac{p}{S_p}\right\|_\lambda &\leq \left\|\frac{1}{S_q}(q-p)\right\|_\lambda + \left\|p\left(\frac{1}{S_q} - \frac{1}{S_p}\right)\right\|_\lambda \\
&\leq \|S_q^{-1}\|_\infty \|q-p\|_\lambda + \|p\|_\infty \left\|\left(\frac{1}{S_q} - \frac{1}{S_p}\right)\right\|_\lambda ,
\end{aligned}$$

where we use $\|\cdot\|_\infty$ to denote the supremum norm on $[0,1]^d$. By the mean value theorem we may write

$$\frac{1}{S_q} - \frac{1}{S_p} = \frac{1}{S_{q,p}^2}(S_q - S_p),$$

for a function $S_{q,p}$ such that $S_q \wedge S_p < S_{q,p} < S_q \vee S_p$. It follows that

$$\begin{aligned}\|h_q - h_p\|_\lambda &\leq \|S_q^{-1}\|_\infty \, \|q-p\|_\lambda + \|p\|_\infty \|(S_q \wedge S_p)^{-2}\|_\infty \, \|(S_q - S_p)\|_\lambda \\ &\leq \left(\|S_q^{-1}\|_\infty + \|p\|_\infty \|(S_q \wedge S_p)^{-2}\|_\infty\right) \|q-p\|_\lambda \,,\end{aligned} \tag{41}$$

where the last inequality follows from statement (i). As $p \in \mathcal{Q}_M$ and $p = h_p S_p$ we have $\|p\|_\infty \leq M$. As $\|S_p^{-1}\|_\infty \leq \|\exp\{\int_0^1 h_p(s,\cdot)\,\mathrm{d}s\}\|_\infty \leq e^M$ and $p,q \in \mathcal{Q}_M$, S_q^{-1}, we have

$$\left(\|S_q^{-1}\|_\infty + \|p\|_\infty \|(S_q \wedge S_p)^{-2}\|_\infty\right) \leq e^M + Me^{2M}.$$

This shows the first inequality in statement (ii). For the second inequality we write

$$\begin{aligned}\|q-p\|_\lambda &= \|(h_q - h_p)S_q\|_\lambda + \|h_p(S_q - S_p)\|_\lambda \\ &\leq \|S_q\|_\infty \|h_q - h_p\|_\lambda + \|h_p\|_\infty \|S_q - S_p\|_\lambda \\ &\leq \left(\|S_q\|_\infty + \|h_p\|_\infty\right) \|h_q - h_p\|_\lambda, \\ &\leq \left(e^M + M\right) \|h_q - h_p\|_\lambda,\end{aligned}$$

where the second to last inequality follows from statement (i). □

Proof of Lemma 12. To show Lemma 12 we need to find constants $0 < c_M < C_M < \infty$ such that

$$P[L^{\mathrm{pl}}(f,\cdot) - L^{\mathrm{pl}}(f_P,\cdot)] \geq c_M \|f - f_P\|_\lambda^2 \tag{42}$$

and

$$P[L^{\mathrm{pl}}(f,\cdot) - L^{\mathrm{pl}}(f_P,\cdot)] \leq C_M \|f - f_P\|_\lambda^2. \tag{43}$$

Let P_f denote the distribution of the observed data induced by the marginal density P_W, the conditional hazard for censoring γ_P, and the conditional log-hazard for the event time of interest f. Let $\nu = P_W \otimes (\lambda \otimes \tau + \delta_{\{1\}\times\{0\}})$ denote a measure on the sample space $\mathcal{O} = [0,1]^{d-1} \times [0,1] \times \{0,1\}$ where λ denotes Lebesgue measure, τ the counting measure, δ Dirac measure, and $\lambda \otimes \tau$ and $\delta_{\{1\}\times\{0\}}$ are considered as measures on $[0,1]\times\{0,1\}$. Then for every $f \in \mathcal{D}_M^d$, $P_f \ll \nu$ and if we let p_f denote the Radon-Nikodym derivative of P_f with respect to ν we have a.s.,

$$\begin{aligned}p_f(\mathbf{w},t,\delta) &= \left(e^{f(t,\mathbf{w})} \exp\left\{-\int_0^t [e^{f(s,\mathbf{w})} + \gamma_P(s,\mathbf{w})]\,\mathrm{d}s\right\}\right)^\delta \\ &\quad \times \left(\gamma_P(t,\mathbf{w})^{\mathbb{1}_{[0,1)}(t)} \exp\left\{-\int_0^t [e^{f(s,\mathbf{w})} + \gamma_P(s,\mathbf{w})]\,\mathrm{d}s\right\}\right)^{1-\delta} \\ &= \left(e^{f(t,\mathbf{w})}\right)^\delta \exp\left\{-\int_0^t e^{f(s,\mathbf{w})}\,\mathrm{d}s\right\} \\ &\quad \times \left(\gamma_P(t,\mathbf{w})^{\mathbb{1}_{[0,1)}(t)}\right)^{1-\delta} \exp\left\{-\int_0^t \gamma_P(s,\mathbf{w})\,\mathrm{d}s\right\}, \\ &=: q_f(\mathbf{w},t,\delta) g(\mathbf{w},t,\delta),\end{aligned} \tag{44}$$

where q_f denotes a component of the likelihood that depends only on f, and g denotes a component that depends only on γ_P. From this it follows that

$$
\begin{aligned}
D_{\mathrm{KL}}(P_{f_0} \,||\, P_f) &= \int \log \frac{p_{f_0}}{p_f} p_{f_0} \,\mathrm{d}\nu \\
&= \int_{[0,1]^d \times \{0,1\}} \Bigg[\int_0^t e^{f(s,\mathbf{w})} \,\mathrm{d}s - \delta f(t, \mathbf{w}) \\
&\qquad - \left(\int_0^t e^{f_0(s,\mathbf{w})} \,\mathrm{d}s - \delta f_0(t, \mathbf{w}) \right) \Bigg] p_{f_0}(\mathbf{w}, t, \delta) \,\mathrm{d}\nu(\mathbf{w}, t, \delta) \\
&= P_{f_0}[L^{\mathrm{pl}}(f, \cdot)] - P_{f_0}[L^{\mathrm{pl}}(f_0, \cdot)],
\end{aligned}
\tag{45}
$$

where D_{KL} is the Kullback-Leiber divergence. Following [van der Vaart, 2000, p. 62] we have

$$
\begin{aligned}
D_{\mathrm{KL}}(P_{f_0} \,||\, P_f) &\geq \int \left(\sqrt{p_{f_0}} - \sqrt{p_f} \right)^2 \mathrm{d}\nu \\
&\geq \left(\|(\sqrt{p_{f_0}} + \sqrt{p_f})^2\|_\infty \right)^{-1} \int \left(p_{f_0} - p_f \right)^2 \mathrm{d}\nu \\
&\geq \left(4e^M (\|\gamma_P\|_\infty \vee 1) \right)^{-1} \int \left(p_{f_0} - p_f \right)^2 \mathrm{d}\nu \\
&= \left(4e^M (\|\gamma_P\|_\infty \vee 1) \right)^{-1} \int \left(p_{f_0} - p_f \right)^2 \mathrm{d}\nu.
\end{aligned}
\tag{46}
$$

Let $S_f(t, \mathbf{w}) = \exp(-\int_0^t e^{f(s,\mathbf{w})} \,\mathrm{d}s)$ denote the conditional survival function associated with the conditional hazard function e^f, and $q_f^* = e^f S_f$ the conditional density associated with the conditional hazard function e^f. We have

$$
\begin{aligned}
\int \left(p_{f_0} - p_f \right)^2 \mathrm{d}\nu &= \int g^2 \left(q_{f_0} - q_f \right)^2 \mathrm{d}\nu \\
&\geq \int_{[0,1]^{d-1} \times [0,1) \times \{1\}} g^2 \left(q_{f_0} - q_f \right)^2 \mathrm{d}\nu \\
&\geq e^{\|\gamma_P\|_\infty} \int_{[0,1]^{d-1}} \int_0^1 (q_{f_0}^* - q_f^*)^2 \,\mathrm{d}(\lambda \otimes P_W) \\
&\geq e^{-\|\gamma_P\|_\infty} \|\omega_P^{-1}\|_\infty \|q_{f_0}^* - q_f^*\|_\lambda^2.
\end{aligned}
\tag{47}
$$

By assumption, $e^{-\|\gamma_P\|_\infty} \|\omega_P^{-1}\|_\infty > 0$ and so Lemma 20 (ii) and the mean value theorem imply that

$$
\int \left(p_{f_0} - p_f \right)^2 \mathrm{d}\nu \geq \tilde{c}_M \|f_0 - f\|_\lambda^2,
\tag{48}
$$

for some $\tilde{c}_M \in (0, \infty)$. Combining equations (45), (46) and (48) gives inequality (42).

To show inequality (43) we use, e.g., [Gibbs and Su, 2002, Theorem 5] to argue that

$$
D_{\mathrm{KL}}(P_{f_0} \,||\, P_f) \leq \int \frac{(p_{f_0} - p_f)^2}{p_f} \,\mathrm{d}\nu.
$$

Using the decomposition in equation (44) we obtain

$$\begin{aligned} D_{\mathrm{KL}}(P_{f_0} \,||\, P_f) &\leq \int \frac{g^2(q_{f_0} - q_f)^2}{q_f g} \,\mathrm{d}\nu \\ &= \int \frac{g(q_{f_0} - q_f)^2}{q_f} \,\mathrm{d}\nu \\ &\leq \exp\{M + e^{-M}\}(\|\gamma_P\|_\infty \vee 1) \int (q_{f_0} - q_f)^2 \,\mathrm{d}\nu, \end{aligned} \tag{49}$$

where we used that $1/q_f$ is bounded by $\exp\{M + e^{-M}\}$ for all $f \in \mathcal{D}_M^d$, and that g is bounded by $\|\gamma_P\|_\infty \vee 1$. Using that $[0,1]^{d-1} \times \{1\} \times \{1\}$ is a null set under ν, we can write

$$\begin{aligned} &\int (q_{f_0} - q_f)^2 \,\mathrm{d}\nu \\ &= \int_{[0,1]^{d-1}\times[0,1)\times\{1\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu + \int_{[0,1]^{d-1}\times[0,1]\times\{0\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu \\ &= \int_{[0,1]^{d-1}\times[0,1)} (q_{f_0}^* - q_f^*)^2 \,\mathrm{d}(\lambda \otimes P_w) + \int_{[0,1]^{d-1}\times[0,1]} (S_{f_0} - S_f)^2 \,\mathrm{d}(\lambda \otimes P_w) \\ &\leq \int_{[0,1]^{d-1}\times[0,1]} \left\{(q_{f_0}^* - q_f^*)^2 + (S_{f_0} - S_f)^2\right\} \mathrm{d}(\lambda \otimes P_w) \\ &\leq \|\omega_P\|_\infty \left(\int_{[0,1]^d} (q_{f_0}^* - q_f^*)^2 + (S_{f_0} - S_f)^2 \,\mathrm{d}\lambda \right) \\ &\leq \|\omega_P\|_\infty \left(\|q_{f_0}^* - q_f^*\|_\lambda^2 + \|S_{f_0}^* - S_f^*\|_\lambda^2 \right), \end{aligned} \tag{50}$$

and the inequality (43) then follows from Lemma 20 combined with equations (45), (49) and (50). □

Proof of Corollary 13. First note that because condition 7 (i) is assumed to hold, $f_P = f^*$ a.e. by Lemma 12. Thus Corollary 13 follows from Theorem 9 if we can show that Assumption 6 is true. Assumption 6 (i) follows by definition of the loss function and 6 (ii) follow from Lemma 12 as γ_P and ω_P are assumed uniformly bounded. It thus only remains to show 6 (iii). To do so, let $\varepsilon > 0$ be given and let $[l_1, u_1], \ldots, [l_K, u_K]$ denote a collection of ε-brackets with respects to $\|\cdot\|_\lambda$ covering $\mathcal{D}_M^d$. By definition of the bracketing number we can take $K = N_{[\,]}(\varepsilon, \mathcal{D}_M^d, \|\cdot\|_\lambda)$. Define for all $k = 1, \ldots, K$,

$$\tilde{l}_k(t, \delta, \mathbf{w}) = \delta l_k(t, \mathbf{w}) - \int_0^t e^{u_k(s,\mathbf{w})} \,\mathrm{d}s, \quad \text{and} \quad \tilde{u}_k(t, \delta, \mathbf{w}) = \delta u_k(t, \mathbf{w}) - \int_0^t e^{l_k(s,\mathbf{w})} \,\mathrm{d}s.$$

Any element in $\mathcal{L}_M$ is on the form $L^{\mathrm{pl}}(f, \cdot)$ for some $f \in \mathcal{D}_M^d$. If $[l_k, u_k]$ is a bracket containing f then it follows that $[\tilde{l}_k, \tilde{u}_k]$ contains $L^{\mathrm{pl}}(f, \cdot)$. Thus $[\tilde{l}_1, \tilde{u}_1], \ldots, [\tilde{l}_K, \tilde{u}_K]$ is a collection of brackets covering $\mathcal{L}_M$. If we let $\mathbb{E}$ denote expectation under P we have by the triangle inequality

$$\|\tilde{l}_k - \tilde{u}_k\|_P \leq \mathbb{E}\left[\Delta\left\{l_k(\tilde{T},W) - u_k(\tilde{T},W)\right\}^2\right]^{1/2} + \mathbb{E}\left[\left\{\int_0^{\tilde{T}} e^{u_k(s,W)} - e^{l_k(s,W)}\,\mathrm{d}s\right\}^2\right]^{1/2}.$$

By equation (18), $\Delta = \Delta\mathbb{1}\{\tilde{T} < 1\}$ a.s., which implies

$$\Delta\left\{l_k(\tilde{T},W) - u_k(\tilde{T},W)\right\}^2 \leq \mathbb{1}\{\tilde{T} < 1\}\left\{l_k(\tilde{T},W) - u_k(\tilde{T},W)\right\}^2 \quad \text{a.s.},$$

and so

$$\begin{aligned}
&\mathbb{E}\left[\Delta\left\{l_k(\tilde{T},W) - u_k(\tilde{T},W)\right\}^2\right] \\
&\leq \mathbb{E}\left[\mathbb{1}\{\tilde{T} < 1\}\left\{l_k(\tilde{T},W) - u_k(\tilde{T},W)\right\}^2\right] \\
&= \int_{[0,1]^{d-1}}\int_0^1 \{l_k(s,\mathbf{w}) - u_k(s,\mathbf{w})\}^2\, h(s,\mathbf{w})e^{-\int_0^s h(u,\mathbf{w})\,\mathrm{d}u}\omega_P(\mathbf{w})\,\mathrm{d}s\,\mathrm{d}\mathbf{w},
\end{aligned}$$

where we use $h(\cdot,\mathbf{w})$ to denote the conditional hazard for $\tilde{T}$ on $[0,1)$ given $W = \mathbf{w}$. By assumption, $\|h\omega_P\|_\infty \leq B$ for some finite constant B, and so we obtain

$$\mathbb{E}\left[\Delta\left\{l_k(\tilde{T},W) - u_k(\tilde{T},W)\right\}^2\right]^{1/2} \leq B\|l_k - u_k\|_\lambda.$$

By Jensen's inequality and the mean value theorem we similarly obtain

$$\begin{aligned}
\mathbb{E}\left[\left\{\int_0^{\tilde{T}} e^{u_k(s,W)} - e^{l_k(s,W)}\,\mathrm{d}s\right\}^2\right]^{1/2} &\leq \mathbb{E}\left[\tilde{T}\int_0^{\tilde{T}}\left(e^{u_k(s,W)} - e^{l_k(s,W)}\right)^2\mathrm{d}s\right]^{1/2} \\
&\leq \mathbb{E}\left[\int_0^1\left(e^{u_k(s,W)} - e^{l_k(s,W)}\right)^2\mathrm{d}s\right]^{1/2} \\
&\leq e^M\,\mathbb{E}\left[\int_0^1\{u_k(s,W) - l_k(s,W)\}^2\,\mathrm{d}s\right]^{1/2} \\
&\leq e^M\|\omega_P\|_\infty\|u_k - l_k\|_\lambda,
\end{aligned}$$

and so we have

$$\|\tilde{l}_k - \tilde{u}_k\|_P \leq \left(B + e^M\|\omega_P\|_\infty\right)\|u_k - l_k\|_\lambda.$$

Thus $[\tilde{l}_1,\tilde{u}_1],\ldots,[\tilde{l}_K,\tilde{u}_K]$ is a collection of $(B+e^M\|\omega_P\|_\infty)\varepsilon$-brackets covering $\mathcal{L}_M$, which shows that $N_{[\,]}(\varepsilon,\mathcal{L}_M,\|\cdot\|_P) \leq N_{[\,]}(\varepsilon/(B+e^M\|\omega_P\|_\infty),\mathcal{D}_M^d,\|\cdot\|_\lambda)$. □

C.2 Density estimation

Proof of Proposition 14. Define $\tilde{\mathcal{B}}_M = \{\beta \in \mathbb{R}^{\tilde{m}(d,n)} : \|\beta\|_1 \leq M\}$ where $\tilde{m}(d,n) = n2^{d-1}$. To show that $\beta \mapsto \mathbb{P}_n[\bar{L}(g_{\beta,n}, \cdot)]$ is convex, take $\beta_1, \beta_0 \in \tilde{\mathcal{B}}_M$. Note that for any $u \in [0,1]$, $\mathbf{w} \in [0,1]^{d-1}$, and $\alpha \in [0,1]$,

$$\exp\left\{g_{\alpha\beta_1+(1-\alpha)\beta_0}(z,\mathbf{w})\right\} = \left(\exp\left\{g_{\beta_1}(z,\mathbf{w})\right\}\right)^{\alpha}\left(\exp\left\{g_{\beta_0}(z,\mathbf{w})\right\}\right)^{1-\alpha}.$$

By Hölder's inequality,

$$\begin{aligned}\int_0^1 e^{g_{\alpha\beta_1+(1-\alpha)\beta_0}(z,\mathbf{w})}\,\mathrm{d}z &= \int_0^1 \left(\exp\left\{g_{\beta_1}(z,\mathbf{w})\right\}\right)^{\alpha}\left(\exp\left\{g_{\beta_0}(z,\mathbf{w})\right\}\right)^{1-\alpha}\mathrm{d}z\\ &\leq \left(\int_0^1 \exp\left\{g_{\beta_1}(z,\mathbf{w})\right\}\mathrm{d}z\right)^{\alpha}\left(\int_0^1 \exp\left\{g_{\beta_0}(z,\mathbf{w})\right\}\mathrm{d}z\right)^{1-\alpha},\end{aligned}$$

which implies

$$\begin{aligned}\log\left(\int_0^1 e^{g_{\alpha\beta_1+(1-\alpha)\beta_0}(s,\mathbf{w})}\,\mathrm{d}s\right) &\leq \alpha\log\left(\int_0^1 e^{g_{\beta_1}(s,\mathbf{w})}\,\mathrm{d}s\right)\\ &\quad + (1-\alpha)\log\left(\int_0^1 e^{g_{\beta_0}(s,\mathbf{w})}\,\mathrm{d}s\right).\end{aligned}$$

From this it follows that

$$\mathbb{P}_n[\bar{L}(g_{\alpha\beta_1+(1-\alpha)\beta_0}, \cdot)] \leq \alpha\mathbb{P}_n[\bar{L}(g_{\beta_1}, \cdot)] + (1-\alpha)\mathbb{P}_n[\bar{L}(g_{\beta_0}, \cdot)],$$

so $\beta \mapsto \mathbb{P}_n[\bar{L}(g_{\beta,n}, \cdot)]$ is convex. Because $\tilde{\mathcal{B}}_M$ is convex the problem in (21) is convex, and because $\beta \mapsto \mathbb{P}_n[\bar{L}(g_{\beta,n}, \cdot)]$ is continuous, the minimum is attained. To show the second statement in the proposition, note that

$$\mathbb{P}_n[-\log p] = \mathbb{P}_n[\bar{L}(\log p, \cdot)] \quad \text{for any} \quad p \in \mathcal{P}^d_{M,n}. \tag{51}$$

Observe that if $a \colon [0,1]^d \to \mathbb{R}$ is a function such that $a(u,\mathbf{w}) = a(0,\mathbf{w})$ for all $u \in [0,1]$ and $\mathbf{w} \in [0,1]^{d-1}$, then for any $f \in \mathcal{D}^d_M$ and $O \in [0,1]^d$,

$$\begin{aligned}\bar{L}(f+a, O) &= \log\left(\int_0^1 e^{f(s,W)+a(s,W)}\,\mathrm{d}s\right) - (f(U,W) - a(U,W))\\ &= \log\left(e^{a(0,W)}\int_0^1 e^{f(s,W)}\,\mathrm{d}s\right) - (f(U,W) - a(0,W))\\ &= \bar{L}(f,O).\end{aligned} \tag{52}$$

In particular, this holds when $a(\mathbf{x}) = b\mathbb{1}\{X_{s,i} \preceq \mathbf{x}_s\}$ for some $b \in \mathbb{R}$, $i \in \{1,\ldots,n\}$, and $s \notin \mathcal{I}$. Hence by definition of $\mathcal{P}^d_{M,n}$ we have for any $p \in \mathcal{P}^d_{M,n}$ that $\mathbb{P}_n[\bar{L}(\log p, \cdot)] = \mathbb{P}_n[\bar{L}(g_{\beta,n}, \cdot)]$ for some $\beta \in \tilde{\mathcal{B}}_M$. By equation (51), we thus have that for any $p \in \mathcal{P}^d_{M,n}$, $\mathbb{P}_n[-\log p] = \mathbb{P}_n[\bar{L}(g_{\beta,n}, \cdot)]$ for some $\beta \in \tilde{\mathcal{B}}_M$. The result then follows from the definition of $g_{\hat{\beta},n}$. □

Proof of Corollary 15. Define the log-density

$$f_1^*(u,\mathbf{w}) = f^*(u,\mathbf{w}) - \log\big(\int_0^1 e^{f^*(z,\mathbf{w})}\,\mathrm{d}z\big),$$

and note that $f_1^* \in \mathcal{P}_M^d$. Equations (51) and (52) imply that $P[\bar{L}(f^*,\cdot)] = P[-\log f_1^*]$ and thus $p_P = f_1^*$ a.e., because the log-likelihood is a strictly proper scoring rule [Gneiting and Raftery, 2007] and $p_P \in \mathcal{P}_M^d$ by assumption. For any HAL estimator $\hat{p}_n$ we can write $\log \hat{p}_n(u,\mathbf{w}) = g_{\hat{\beta},n}(u,\mathbf{w}) - \log\big(\int_0^1 e^{g_{\hat{\beta},n}(z,\mathbf{w})}\,\mathrm{d}z\big)$, for some solution $\hat{\beta}$ to the problem (21). By equation (52), $g_{\hat{\beta},n}$ is a HAL estimator for the loss $\bar{L}$ as defined in equation (9). To prove Corollary 15 it suffices to show that

$$\|g_{\hat{\beta},n} - f^*\|_\lambda = o_P(n^{-1/3}\log(n)^{2(d-1)/3}). \tag{53}$$

We show that Assumption 6 holds for $\bar{L}$, which imply that equation (53) is true by Theorem 9. Assumption 6 (i) holds because all $f \in \mathcal{D}_{M,n}^d$ are uniformly bounded, and Assumption 6 (ii) holds by properties of the Kullback-Leibler divergence because we assume that ω_P is uniformly bounded away from zero and infinity [Gibbs and Su, 2002]. Assumption 6 (iii) is established by the same arguments used in the proof of Corollary 13. □

References

C. Aistleitner and J. Dick. Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality. *Acta Arithmetica*, 167(2):143–171, 2015. URL http://eudml.org/doc/279219.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes.* Springer Science & Business Media, 2012.

A. F. Bibaut and M. J. van der Laan. Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*, 2019.

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivates. *Sankhyā A*, 50:381–393, 1988.

S. P. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

J. R. Coyle, N. S. Hejazi, R. V. Phillips, L. W. van der Laan, and M. J. van der Laan. *hal9001: The scalable highly adaptive lasso*, 2022. URL https://github.com/tlverse/hal9001. R package version 0.4.3.

E. Czerebak-Morozowicz, Z. Rychlik, and M. Urbanek. Almost sure functional central limit theorems for multiparameter stochastic processes. *Condensed Matter Physics*, 2008.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *The Annals of Statistics*, 49(2):769–792, 2021.

D. Ferger. Arginf-sets of multivariate cadlag processes and their convergence in hyperspace topologies. *Theory of Stochastic Processes*, 20(2):13–41, 2015.

J. Friedman, R. Tibshirani, and T. Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

A. Fu, B. Narasimhan, and S. Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020. doi: 10.18637/jss.v094.i14.

S. Geman. Sieves for nonparametric estimation of densities and regressions. *Reports in Pattern Analysis*, 99, 1981.

S. Geman and C.-R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The annals of Statistics*, pages 401–414, 1982.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

R. D. Gill, M. J. Laan, and J. A. Wellner. Inefficient estimators of the bivariate survival function for three models. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 545–597, 1995.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

L. Goldstein and R. Khasminskii. On efficient estimation of smooth functionals. *Theory of Probability & Its Applications*, 40(1):151–156, 1996.

L. Goldstein and K. Messer. Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.*, 20:1306–1328, 1992.

U. Grenander. *Abstract inference*. Wiley, 1981.

P. Groeneboom and G. Jongbloed. *Nonparametric estimation under shape constraints*. Cambridge University Press, 2014.

C. Gu and C. Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, 21(1):217–234, 1993.

G. H. Hardy. On double Fourier series and especially those which represent the double zeta-function with real and incommensurable parameters. *Quart. J. Math*, 37(1):53–79, 1906.

N. S. Hejazi, J. R. Coyle, and M. J. van der Laan. hal9001: Scalable highly adaptive lasso regression inr. *Journal of Open Source Software*, 5(53):2526, 2020.

T. Hothorn. Transformation boosting machines. *Statistics and Computing*, 30(1):141–152, 2020.

M. Krause. Über Mittelwertsätze im Gebiete der Doppelsummen and Doppelintegrale. *Leipziger Ber*, 55:239–263, 1903.

D. K. Lee, N. Chen, and H. Ishwaran. Boosted nonparametric hazards with time-dependent covariates. *Annals of statistics*, 49(4):2101, 2021.

T. Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):113–132, 1978.

I. W. McKeague and K. J. Utikal. Inference for a nonlinear counting process regression model. *The Annals of Statistics*, 18(3):1172–1187, 1990.

M. Naaman. On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statistics & Probability Letters*, 173:109088, 2021.

G. Neuhaus. On weak convergence of stochastic processes with multidimensional time parameter. *The Annals of Mathematical Statistics*, 42(4):1285–1295, 1971.

A. B. Owen. Multidimensional variation for quasi-monte carlo. In *Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday*, pages 49–74. World Scientific, 2005.

H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pages 453–466, 1983.

H. C. Rytgaard, T. A. Gerds, and M. J. van der Laan. Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. *The Annals of Statistics*, 50(5):2469–2491, 2022.

H. C. Rytgaard, F. Eriksson, and M. J. van der Laan. Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*, 79(4):3038–3049, 2023.

M. Schmid and T. Hothorn. Flexible boosting of accelerated failure time models. *BMC bioinformatics*, 9:1–13, 2008.

A. Schuler, Y. Li, and M. van der Laan. The selectively adaptive lasso. *arXiv preprint arXiv:2205.10697*, 2023.

B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982.

L. Spierdijk. Nonparametric conditional hazard rate estimation: a local linear approach. *Computational Statistics & Data Analysis*, 52(5):2419–2434, 2008.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8:1348–1360, 1980.

J. K. Tay, B. Narasimhan, and T. Hastie. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31, 2023. doi: 10.18637/jss.v106.i01.

M. van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The international journal of biostatistics*, 13(2), 2017.

M. van der Laan. Higher order spline highly adaptive lasso estimators of functional parameters: Pointwise asymptotic normality and uniform convergence rates. *arXiv preprint arXiv:2301.13354*, 2023.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

I. van Keilegom and N. Veraverbeke. Hazard rate estimation in nonparametric regression with censored data. *Annals of the Institute of Statistical Mathematics*, 53:730–745, 2001.

G. G. Walter and J. R. Blum. A simple solution to a nonparametric maximum likelihood estimation problem. *The Annals of Statistics*, pages 372–379, 1984.