UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

PhD thesis

Targeted learning with right-censored data

Anders Munch
Section of Biostatistics, University of Copenhagen

Academic advisors
Thomas Alexander Gerds
Claus Thorn Ekstrøm

Phd thesis

Targeted learning with right-censored data

Anders Munch

Section of Biostatistics
University of Copenhagen

Academic advisors

Thomas Alexander Gerds

Claus Thorn Ekstrøm

Assessment committee

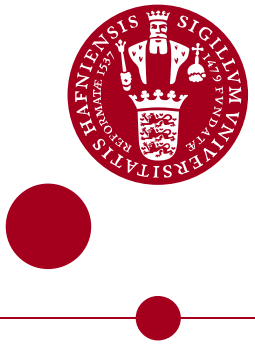
Thomas Scheike

Morten Overgaard

Jan Beyersmann

This thesis has been submitted to the Graduate School of the Faculty of Health and Medical Sciences, University of Copenhagen on August 31, 2023

UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES



PhD thesis

Targeted learning with right-censored data

Anders Munch
Section of Biostatistics, University of Copenhagen

Academic advisors
Thomas Alexander Gerds
Claus Thorn Ekstrøm

This thesis has been submitted to the Graduate School of the Faculty of Health and Medical Sciences, University of Copenhagen on August 31, 2023

Acknowledgments

This thesis was written at the Section of Biostatistics at the University of Copenhagen. I would like to thank everyone at the section, in particular my main supervisor Thomas Gerds. Had it not been for Thomas' lessons and the friendly atmosphere of the section, I would probably not have become a biostatistician. I would also like to thank Mark van der Laan for many discussions and for hosting me for four months in Berkeley.

Contents

Summary

The aim of this thesis is to contribute to the advancement of statistically rigorous methods that enable the utilization of data-adaptive estimators based on continuous-time observations that may be right-censored. Right-censoring often occurs when subjects are observed over a period of time, which is a typical situation in biostatistics. Conventional statistical methods for handling this type of data are based on (semi-)parametric models or simple non-parametric models. Importantly, for these approaches to provide valid statistical inference, the models have to be pre-specified. An appealing alternative is to use data-adaptive methods like machine learning, which provide more flexible models and tools for adapting the models to the observed data. For instance, cross-validation or super learning uses the observed data to select a model from a collection of candidate models.

The challenge with machine learning-based estimation strategies is to conduct valid statistical inference. Targeted learning addresses this challenge using semi-parametric efficiency theory. Although extensively studied for causal inference, the adaptation of targeted learning to right-censored problems in continuous-time data is less mature.

The thesis is comprised of a synopsis with eight chapters and three manuscripts. The synopsis gives an introduction to the central theoretical concepts that underpin the topics covered by the manuscripts. We first give an overview of the steps or 'road map' of targeted learning and introduce some concepts from semi-parametric efficiency theory. We then discuss identifiability conditions for right-censored data. We also provide some background on super learning and the highly-adaptive lasso, which are two data-adaptive estimation techniques commonly used in targeted learning.

The three manuscripts extend the framework and tools of targeted learning to settings with right-censored data in three different directions. The first manuscript applies the general framework of targeted learning to the illness-death model. We construct a class of estimators of the state occupation probabilities that can leverage data-adaptive estimators of the state transitions in the model. The second manuscript discusses the challenges facing the statistician who wants to construct a super learner from right-censored data. In this manuscript we also propose a new super learner and compare it to existing methods. The third manuscript formally extends the highly-adaptive lasso to settings that include conditional density and hazard function estimation.

Resumé

Målet med denne afhandling er at bidrage til udviklingen af valide statistiske metoder, der muliggør anvendelsen af data-adaptive estimatorer baseret på højrecensureret data observeret i kontinuert tid. Højrecensurering er et normalt fænomen, når data observeres over tid, hvilket typisk er situationen i biostatistik. Konventionelle statistiske metoder designet til denne type data bygger på (semi-)parametriske modeller eller simple ikke-parametriske modeller. For at opnå valid statistisk inferens, er det nødvendigt at disse modeller prespecificeres. Et attraktivt alternativ er at anvende data-adaptive metoder som *machine learning*, der giver mere fleksible modeller og værktøjer til at tilpasse modellerne til det observerede data. For eksempel bruger krydsvalidering eller *super learning* det observerede data til at vælge en model fra en samling af kandidatmodeller.

Udfordringen ved *machine learning*-baserede estimeringsstrategier er at opnå valid statistisk inferens. *Targeted learning* håndterer denne udfordring ved hjælp af semi-parametrisk efficiensteori. Disse metoder er omfattende studeret for problemer i kausal inferens, men er mindre udviklede for højrecensureringsproblemer i kontinuert tid.

Afhandlingen består af en oversigt med otte kapitler og tre manuskripter. Oversigten giver en introduktion til de centrale teoretiske koncepter, der ligger til grund for de emner, manuskripterne omhandler. Vi giver først et overblik over trinene i *targeted learning* og introducerer nogle begreber fra semi-parametrisk efficiensteori. Derefter drøfter vi identifikationsbetingelser for højrecensureret data. Vi giver også en baggrund for manuskripterne, der omhandler *super learning* og *the highly-adaptive lasso*, som er to data-adaptive estimeringsteknikker, der ofte anvendes i *targeted learning*.

De tre manuskripter udvider metoder fra *targeted learning* til situationer med højrecensureret data i tre forskellige retninger. Det første manuskript anvender den generelle *targeted learning*-metodik på *illness-death*-modellen. Vi konstruerer en klasse af estimatorer for tilstandssandsynlighederne, der tillader brugen af data-adaptive metoder til at estimere overgangssandsynlighederne i modellen. Det andet manuskript drøfter de udfordringer, som statistikeren står over for, når han eller hun ønsker at konstruere en *super learner* baseret på højrecensureret data. I dette manuskript foreslår vi også en ny *super learner* og sammenligner den med eksisterende metoder. Det tredje manuskript udvider formelt *the highly-adaptive lasso* til situationer, der inkluderer betinget tætheds- og hazardfunktionsestimering.

Chapter 1

Objectives and overview

The overall objective of the thesis presented here is the development of statistically sound methods that allow us to use data-adaptive estimators when the available data are observed in continuous time and subject to right-censoring.

Many examples in biostatistics concern data recorded on subjects over time. Right-censoring means that information about a subject is available only up to some (random) time point, and is typically ubiquitous for time-dynamic data due to dropout or end of the study period. Under suitable assumptions, it is possible to infer the underlying dynamics of the population of interest, even though we only have access to a 'corrupted' sample, where some subjects are right-censored. To identify features of the uncensored population of interest we need to model the dynamics of the system and the censoring mechanism. Traditional statistical approaches use either (semi-)parametric models or non-parametric models where some components of the available data are ignored. An attractive alternative is to use flexible, data-adaptive methods such as machine learning. A challenge with this approach is that 'naive' plug-in estimates obtained using machine learning typically do not come with valid confidence intervals.

The development of targeted learning within the last decade or two has demonstrated how we can use semi-parametric efficiency theory to obtain valid statistical inference for estimators that use machine learning. Targeted learning is also referred to as 'debiased machine learning' but we use the term 'targeted learning' in this synopsis. While targeted learning has been studied and developed extensively for causal inference problems, the theory is less developed for right-censored problems when data are observed in continuous time.

The thesis consists of three manuscripts and a synopsis. The synopsis is organized as follows. In Chapter 2 we give an overview of the components of targeted learning. Chapter 3 provides a brief introduction to some of the central theoretical concepts underlying targeted learning. These concepts are central for Manuscript I. Chapter 4 discusses coarsened data in general and right-censored survival data in particular. Right-censoring plays a role in all three manuscripts. In Chapter 5 we discuss super learning based on right-censored data, which is the topic of Manuscript II. Chapter 6 introduces the highly-adaptive lasso, which is the object of study in Manuscript III. Chapter 7 provides summaries of the three manuscripts, and Chapter 8 discusses limitations and topics for future research. Appendix A contains some technical results.

Throughout this synopsis, we give a high-level introduction without paying too much attention to technicalities. We refer to Bickel et al. [1993], van der Vaart [2000], and van der Vaart and Wellner [1996] for precise definitions and details.

Chapter 2

The modules of targeted learning

Targeted learning can be decomposed into different steps. A visual summary is given in Figure 2.1, and in this section we give a high-level description of each step.

Parameter of interest The philosophy of targeted learning is that a statistical analysis should start with the definition of a scientifically meaningful parameter of interest [van der Laan and Rose, 2011, Petersen and van der Laan, 2014]. Mathematically, this means that we should find a map $\theta\colon \mathcal{Q} \to \Theta$ defined on a collection of probability measures $\mathcal{Q}$. Each element $Q \in \mathcal{Q}$ denotes a distribution of some population of interest. The map θ is called the target parameter. We illustrate the general idea with an example.

Example 2.1 (Illness-death model)
The PROVA trial investigated the effect of propanolol and sclerotherapy on variceal bleeding and death among cirrhotic patients [PROVA Study Group, 1991]. A set of baseline variables was measured once at randomization. A relevant question could be the following. Among patients treated with sclerotherapy, what is the average risk of being alive and having experienced variceal bleeding a year after receiving treatment? We can model this as an illness-death model without recovery [Fix and Neyman, 1951, Sverdrup, 1965, Andersen et al., 2012]. Let $W \in \mathbb{R}^d$ be a vector of baseline variables and $X(t) \in \{0, 1, 2\}$ a non-decreasing stochastic process for $t \in [0, 1]$ with $X(0) = 0$. The states X can occupy is interpreted as 'healthy', 'ill', and 'dead', respectively. In our example, 'ill' means that a patient has experienced variceal bleeding. Let $\mathcal{Q}$ denote a collection of probability measures where each element $Q \in \mathcal{Q}$ determines a distribution for $(W, \{X(t) : t \in [0, 1]\})$. Our target parameter is the map $\theta\colon \mathcal{Q} \to [0, 1]$ defined as $\theta(Q) = Q(X(1) = 1)$. •

The choice of $\mathcal{Q}$ should ideally reflect the assumptions we are willing to make about the population of interest. For instance, if the hazard of death in Example 2.1 depends on whether or not a patient has experienced bleeding but not on the time at which bleeding occurred, the population can be modeled with a (semi-)Markov model [Andersen et al., 2012]. A common approach is to impose parametric assumptions such that $\mathcal{Q}$ can be indexed by a Euclidean parameter set. Our main focus is on the non-parametric case where essentially no assumptions are imposed on $\mathcal{Q}$; we discuss the precise definition of a 'non-parametric model' in Section 3.3.

Identifiability In many biostatistical examples, observations from the population of interest are not available. For instance, in the PROVA trial (Example 2.1), not all patients were observed for a whole year. Some dropped out of the study, and some were included

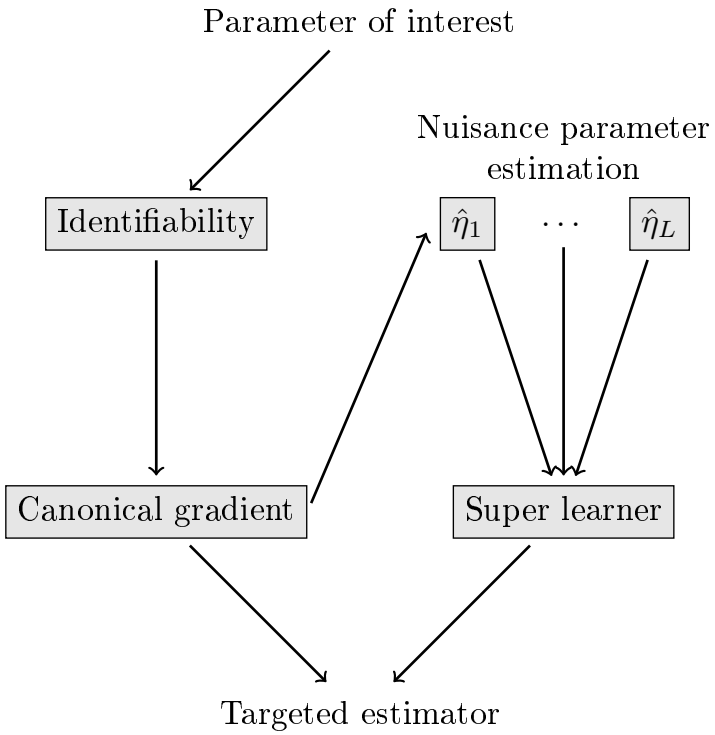


Figure 2.1: The modules and dependencies of targeted learning. We discuss identifiability in Chapter 4, the canonical gradient in Chapter 3, an example of a nuisance parameter estimator in Chapter 6, and super learning in Chapter 5.

later than a year before the study ended. These patients are right-censored, because we only see what happens up to some time point but not what happens after that. Similarly, if we want to identify a causal effect from observational data, the population of interest is a randomized group of patients that we do not have data from [van der Laan and Robins, 2003, Hernán and Robins, 2020]. In Chapter 4 we discuss how a distribution Q of interest can be identified from the observed data distribution P. When the distribution Q can be identified from P we can write $Q(P)$, and the target parameter $\theta\colon \mathcal{Q} \to \Theta$ can then be identified as $\Psi(P) = \theta(Q(P))$ for some map $\Psi\colon \mathcal{P} \to \mathbb{R}$, where $\mathcal{P}$ is a collection of probability measures for the population we observe.

Canonical gradient The parameter $\Psi\colon \mathcal{P} \to \Theta$ can be estimated using an observed data set $\{O_i\}_{i=1}^n$ of i.i.d. observations $O_i \sim P$ for some $P \in \mathcal{P}$. When the parameter Ψ is smooth enough it admits a (canonical) gradient $\varphi_P \in \mathcal{L}_P^2$ at $P \in \mathcal{P}$; the definitions of a 'gradient' and 'smooth enough' are given in Section 3.1. Let $\hat{P}_n$ be an estimator of P and define the one-step estimator [Pfanzagl and Wefelmeyer, 1982, Bickel et al., 1993, Kennedy, 2022]

$$\hat{\Psi}_n^* = \Psi(\hat{P}_n) + \mathbb{P}_n[\varphi_{\hat{P}_n}]. \tag{2.1}$$

In Section 3.2 we provide some intuition for why the one-step estimator is a good idea. The main attraction is that for many examples it holds that if $\hat{P}_n = P + o_P(n^{-1/4})^\dagger$ then $\sqrt{n}(\hat{\Psi}_n^* - \Psi(P)) \rightsquigarrow \mathcal{N}(0, P[\varphi_P^2])$.

†As we have not defined a norm on the space that contains P and $\hat{P}_n$, the o_P notation is used informally here and in the remainder of the chapter. The intended meaning is that P is estimated by $\hat{P}_n$ with an estimation error that decreases with n at a rate faster than $n^{-1/4}$.

Nuisance parameter estimation To construct the one-step estimator we need an estimator of P. We can often characterize P by a vector of nuisance parameters which we know how to estimate [van der Laan and Rose, 2011, Chernozhukov et al., 2018a]. For instance, a one-step estimator of the average treatment effect [Hernán and Robins, 2020] can be constructed using any estimator of a regression function [e.g., Kennedy, 2016], and in Manuscript I we show that a one-step estimator for the state occupation probability considered in Example 2.1 can be constructed using estimators of the hazard functions for all possible state transitions. We can then ensure that $\hat{P}_n = P + o_P(n^{-1/4})$ by using estimators of each nuisance parameter that are consistent at rate $n^{-1/4}$. The $n^{-1/4}$-rate of convergence is significantly weaker than the parametric $n^{-1/2}$-rate of convergence and allows nuisance parameters to be estimated with data-adaptive methods. In addition, the high-level convergence rate condition gives flexibility in the choice of estimator.

One possible choice of estimator for a nuisance parameter is the highly-adaptive lasso (HAL) estimator which we describe in Chapter 6. Under the assumption that the nuisance parameter belongs to the space of multivariate càdlàg functions with uniformly bounded sectional variation norm, the HAL estimator of the nuisance parameter will fulfill the needed convergence rate condition for any dimension of the covariate space. This observation was used by van der Laan [2017b] to construct a general targeted estimator that is valid in a broad range of settings.

Super learner While nuisance parameters can be estimated at rate $n^{-1/4}$ without imposing parametric assumptions, the minimax framework teaches us that *some* assumptions are generally needed to achieve this rate of convergence [Ibragimov and Has'Minskii, 1981, Wainwright, 2019]. As stated above, the assumption of a bounded sectional variation norm is one possibility, but smoothness, sparsity, or monotonicity assumptions could also be imposed. It is difficult to know in advance which (if any) of these assumptions are true about the unknown data-generating distribution. In addition, data-adaptive estimators typically rely on one or more hyperparameters that are difficult to pre-specify.

The super learner is one possible strategy for addressing this challenge. The super learner is a meta-algorithm that combines a collection of candidate estimators into a new estimator with performance guaranteed to be almost as good as the best performing estimator [van der Laan and Dudoit, 2003, van der Vaart et al., 2006, van der Laan et al., 2007]. The super learner fits nicely into the targeted learning framework, as it provides a flexible and general method for constructing estimators of nuisance parameters that converge at a sufficiently fast rate. To see this, imagine that we have a collection of candidate estimators $\{\hat{\eta}_l\}_{l=1}^L$ for estimating the nuisance parameter η. For each $l = 1, \ldots, L$, let $\mathcal{C}_l$ be a statement such that if $\mathcal{C}_l$ is true then $\eta_l = \eta + o_P(n^{-1/4})$. For instance, if $\hat{\eta}_1$ is the HAL estimator, $\mathcal{C}_l$ is the statement that η is a càdlàg function with sectional variation norm bounded by some fixed constant. Letting $\hat{\eta}_{\text{sl}}$ denote the super learner, it holds that $\hat{\eta}_{\text{sl}} = \eta + o_P(n^{-1/4})$ if just one of $\mathcal{C}_l$, $l = 1, \ldots, L$ is true. A simple example is when $\{\hat{\eta}_l\}_{l=1}^L$ is a collection of parametric models. In this case, $\hat{\eta}_{\text{sl}} = \eta + o_P(n^{-1/4})$ if just one of the models is correctly specified.

Targeted estimator The one-step estimator defined in equation (2.1) is one example of a targeted estimator. There are other ways of constructing a targeted estimator based on the canonical gradient and estimators of relevant nuisance parameters. Targeted minimum-loss estimation (TMLE) [van der Laan and Rubin, 2006, van der Laan and Rose, 2011] uses a plug-in estimator $\Psi(\hat{P}_n^*)$ where $\hat{P}_n^*$ is constructed such that $\mathbb{P}_n[\varphi_{\hat{P}_n^*}] = o_P(n^{-1/2})$. Debiased machine learning (DML) [Chernozhukov et al., 2018a] uses a Neyman-orthogonal score function $\psi(\cdot; \theta, \eta)$ indexed by a nuisance parameter η, such that the target parameter

solves $P[\psi(\cdot;\theta,\eta(P))] = 0$ in θ. When the score function is linear in θ, the one-step estimator is a DML estimator.[‡]

We emphasize the modularity of the targeted learning approach. Each gray box in Figure 2.1 is a separate module that can be analyzed independently of the other modules. For instance, the formula for the asymptotic variance of a targeted estimator is the same for any choice of estimators of the nuisance parameters, as long as they converge fast enough. Similarly, the theoretical properties of the super learner holds for any collection of learners. In this way, targeted learning is notably different from 'traditional' statistical approaches that rely on the delta method and asymptotic linearity of $\hat{P}_n$ to establish asymptotic linearity of $\Psi(\hat{P}_n)$. When we rely on the delta method, the choice of estimator for the nuisance parameter determines the asymptotic distribution of $\Psi(\hat{P}_n)$. As $\hat{P}_n$ is typically not asymptotically linear when $\hat{P}_n$ is estimated using data-adaptive methods, it is difficult to obtain valid statistical inference using traditional methods.

We illustrate the importance of employing a targeting step when data-adaptive estimators are used in Figure 2.2. The figure is constructed from simulated data generated as described in Appendix E of Manuscript I. To estimate the target parameter introduced in Example 2.1 we use a flexible penalized Poisson regression to estimate all transition hazard functions in the illness-death model and cross-validation to select the penalty parameter. This gives an estimator $\hat{P}_n$ of P. From Figure 2.2 we see that the targeted estimator $\hat{\Psi}_n^*$ has a lower bias than the 'naive' plug-in estimator $\Psi(\hat{P}_n)$.

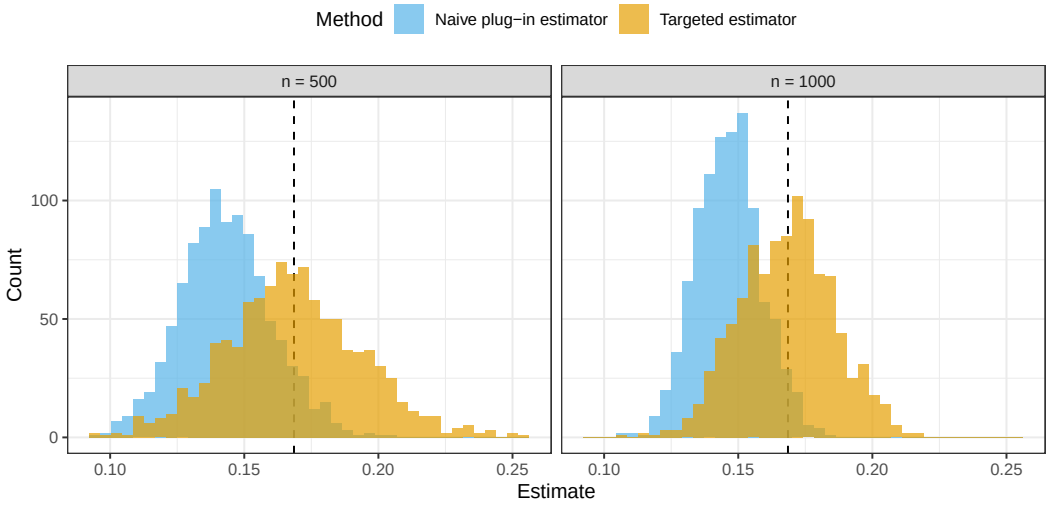


Figure 2.2: Results of 1000 simulations of a naive plug-in estimator and a targeted estimator of the state occupation probability in an illness-death model stratified on number of samples (n). The naive plug-in estimator refers to the estimator $\Psi(\hat{P}_n)$, and the targeted estimator refers to the one-step estimator $\hat{\Psi}_n^*$, see equation (2.1). The dashed vertical line is the state occupation probability according to the data-generating distribution.

[‡]Technically, a DML estimator uses cross-fitting, which is not used in our definition of the one-step estimator. Cross-fitting involves fitting some nuisance parameter estimators in separate parts of the data, while the one-step, as defined in equation (2.1), relies on an additional Donsker class regularity condition.

Chapter 3

Efficiency and targeting

In this chapter, let $\Psi\colon \mathcal{P} \to \mathbb{R}$ be a real-valued target parameter. When we estimate the unknown distribution $P \in \mathcal{P}$ from data, we expect to get closer to P as the sample size increases. We might therefore hope to understand the asymptotic behavior of estimators of Ψ by studying the local behavior of Ψ around $P \in \mathcal{P}$ – at least if Ψ is smooth enough. The formal development of this idea has a long history [e.g., Stein, 1956, Koshevnik and Levit, 1977, Beran, 1977, Levit, 1978, Begun et al., 1983, Pfanzagl and Wefelmeyer, 1982, Newey, 1990, van der Vaart, 1991, Robins and Rotnitzky, 1992, Lu and Tsiatis, 2008] and led to the development of a general theory of semi-parametric efficiency, which is described in several monographs [Bickel et al., 1993, van der Vaart, 2000, van der Laan and Robins, 2003, Tsiatis, 2007, Kosorok, 2008]. We discuss some of the central concepts in Section 3.1. Several authors have used tools from semi-parametric efficiency theory to construct estimators of low-dimensional target parameters when an infinite-dimensional nuisance parameter has to be estimated [e.g., Bickel and Ritov, 1988, Andrews, 1994, Newey, 1994, Birgé and Massart, 1995, Laurent et al., 1996, Newey et al., 1998]. More recently, focus has been on how semi-parametric efficiency theory allows the use of data-adaptive estimation methods [e.g., van der Laan and Rubin, 2006, van der Laan and Rose, 2011, Kandasamy et al., 2014, van der Laan and Rose, 2018, Chernozhukov et al., 2018a, Kennedy, 2022]. We provide some intuition for this in Section 3.2. Section 3.3 briefly discusses the meaning of 'non-parametric models'.

3.1 Pathwise differentiability

We assume for simplicity that $\mathcal{P} \ll \mu$ for some σ-finite measure μ, and for $P \in \mathcal{P}$ we write $p = \mathrm{d}P/\,\mathrm{d}\mu$. To study the local behavior of Ψ around P, we study how $\Psi(P_t)$ behaves along paths $t \mapsto P_t \in \mathcal{P}$, where $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$ and $P_0 = P$. We use the notation $\{P_t\}$ for any such path. Note that the path $\{P_t\}$ is a one-dimensional parametric submodel of $\mathcal{P}$ with parameter space $t \in (-\varepsilon, \varepsilon)$. Thus we can define the score function for $\{P_t\}$ at P in the usual manner as the derivative of $\log p_t(o)$ at $t = 0$, for all $o \in \mathcal{O}$.[†] The tangent space for $\mathcal{P}$ at P is the closure in $\mathcal{L}^2_P$ of the linear span of the collection of all score functions. A useful result is that the tangent space can always be represented as a subspace of

$$\mathcal{H}_P = \{f \in \mathcal{L}^2_P : P[f] = 0\}. \tag{3.1}$$

We use $\dot{\mathcal{P}}_P \subset \mathcal{H}_P$ to denote (this representation of) the tangent space. We say that $\Psi\colon \mathcal{P} \to \mathbb{R}$ is pathwise differentiable at P if there exists an element $g_P \in \mathcal{L}^2_P$ such that for all paths

[†]We here restrict attention to submodels for which a score function exists. Technically, a score function is defined as a limit in $\mathcal{L}^2_P$ [van der Vaart, 2000, chapter 25.3].

$\{P_t\}$,

$$\frac{\partial}{\partial t}\Big|_{t=0} \Psi(P_t) = \langle g_P, \dot{\ell} \rangle_P, \quad \text{where } \dot{\ell} \text{ is the score function for } \{P_t\} \text{ at } P. \tag{3.2}$$

Any element $g_P \in \mathcal{L}_P^2$ fulfilling equation (3.2) is called a gradient. There exists a unique gradient φ_P such that $\varphi_P \in \dot{\mathcal{P}}_P$, which can be found as the projection of any gradient onto $\dot{\mathcal{P}}_P$. We refer to φ_P as the canonical gradient.

An equivalent definition of pathwise differentiability is that Ψ is Hadamard (or compactly) differentiable at P tangentially to $\dot{\mathcal{P}}_P$ when $\mathcal{P}$ is equipped with the Hellinger metric [Bickel et al., 1993, Remark 2 in Appendix 5]. In general, there are more than one interesting ways to define differentiability of a map Ψ defined on an infinite-dimensional space. That a map is differentiable essentially means that it can be approximated locally by a linear map. As described by Reeds [1976], different concepts of differentiability varies in their requirements to the validity of the linear approximation. For instance, Fréchet differentiability requires the linear approximation to be valid uniformly, while Gâteaux differentiability only requires the approximation to be valid along straight lines [Reeds, 1976, Serfling, 1980, Shapiro, 1990]. Hadamard differentiability is somewhere in between as it requires the linear approximation to be valid along any path $\{P_t\}$. In addition to the various types of differentiability, an infinite-dimensional space admits several non-equivalent norms. Picking the right norm and the right concept of differentiability is important for establishing a useful theory [Reeds, 1976, Gill et al., 1989]. Equipping $\mathcal{P}$ with the Hellinger metric is useful because the tangent space inherits a Hilbert space structure. This allows us to use geometric arguments from general Hilbert space theory.

The canonical gradient can be used to derive facts about estimation of Ψ under the model $\mathcal{P}$. One example is that the information bound for estimation of Ψ under $\mathcal{P}$ can be read off from the canonical gradient. For the finite-dimensional model $\{P_t\}$, the information bound for estimation of Ψ under $\{P_t\}$ is defined as

$$\mathcal{I}_P(\Psi; \{P_t\}) = \frac{P[\dot{\ell}^2]}{\left(\frac{\partial}{\partial t}\big|_{t=0} \Psi(P_t)\right)^2}, \quad \text{with } \dot{\ell} \text{ the score function for } \{P_t\} \text{ at } P.$$

The information bound is motivated by the Cramér-Rao bound, which tells us that any asymptotically unbiased estimator of Ψ under $\{P_t\}$ will have asymptotic variance bounded below by the inverse of $\mathcal{I}_P(\Psi; \{P_t\})$. As estimation of Ψ under the submodel $\{P_t\}$ should be easier than estimation of Ψ under the whole model $\mathcal{P}$, the information bound for Ψ under $\mathcal{P}$ is defined as

$$\mathcal{I}_P(\Psi; \mathcal{P}) := \inf_{\{P_t\}} \mathcal{I}_P(\Psi; \{P_t\}).$$

By using the definition of the canonical gradient and the Cauchy-Schwarz inequality one can show that $\mathcal{I}_P(\Psi; \mathcal{P}) = (P[\varphi_P^2])^{-1}$.

A related result is the following. An estimator $\hat{\Psi}_n$ is asymptotically linear for Ψ under $\mathcal{P}$ when $\hat{\Psi}_n - \Psi(P) = \mathbb{P}_n[I_P] + o_P(n^{-1/2})$, for some function $I_P \in \mathcal{L}_P^2$ with $P[I_P] = 0$ for all $P \in \mathcal{P}$. The function I_P is called the estimators influence function. The estimator is called regular at P if the weak limit of $\sqrt{n}(\hat{\Psi}_n - \Psi(P))$ is stable under small perturbations to the data-generating distribution; see, e.g., [van der Vaart, 1991] for a precise definition. A regular asymptotically linear estimator is called a RAL estimator. By Proposition 3.3.1 in [Bickel et al., 1993], an asymptotically linear estimator is regular if and only if its influence function is a gradient. As φ_P is the projection of any gradient onto $\dot{\mathcal{P}}_P$, it immediately follows that if I_P is the influence function for a RAL estimator of Ψ under $\mathcal{P}$, then $P[I_P^2] \geq P[\varphi_P^2]$. Hence, if a RAL estimator has φ_P as its influence function, it has lowest possible asymptotic variance among all RAL estimators. Such an estimator is called asymptotically efficient. For this reason, the canonical gradient is also referred to as the efficient influence function. Pathwise

differentiability is necessary for a parameter to be estimable by a RAL estimator [van der Vaart, 1991]. In particular, if Ψ can be estimated by a RAL estimator, $\mathcal{I}_P(\Psi; \mathcal{P})$ is finite. The reverse statement does not hold, see for instance [Ritov and Bickel, 1990].

Bickel et al. [1993], Gerds [2002], Ichimura and Newey [2022], and Kennedy [2022] discuss a practically useful way of finding a candidate for the canonical gradient. The idea is to find the Gâteaux derivative of Ψ at P in the direction of the Dirac measure δ_O with point-mass at $O \in \mathcal{O}$. The Gâteaux derivative is the ordinary derivative of the map $t \mapsto \Psi(P + t(\delta_O - P))$ at $t = 0$. Note that this derivative can be calculated using basic mathematical tools, unlike solving equation (3.2) which is an integral equation. To see why this works, let K_h be a distribution indexed by $h > 0$ such that $K_h \ll \mu$ and $K_h \rightsquigarrow \delta_O$ for $h \to 0$. Let k_h be the μ-density of K_h and note that the score function of the model $\{P + t(K_h - P)\}$ at P is $(k_h - p)/p$. Hence, by definition of the canonical gradient φ_P,

$$\frac{\partial}{\partial t}\Psi(P + t(K_h - P)) = \langle \varphi_P, (k_h - p)/p \rangle_P = K_h[\varphi_P] - P[\varphi_P] = K_h[\varphi_P].$$

Letting $h \downarrow 0$ and assuming we can exchange limits, we obtain

$$\frac{\partial}{\partial t}\Psi(P + t(\delta_O - P)) = \varphi_P(O).$$

For this approach to be formally valid, the steps above would have to be established rigorously. In addition, if the model $\mathcal{P}$ is restricted, we need to construct the paths such that $\{P + t(K_h - P)\} \subset \mathcal{P}$ which is not automatically guaranteed. In any case, the approach outlined above can be used heuristically to motivate an estimators, which can then be analyzed formally. We use this strategy in Manuscript I.

3.2 A targeted estimator

A 'naive' plug-in estimator of a low-dimensional target parameter based on data-adaptive estimators of nuisance parameters can perform poorly. This was illustrated in Figure 2.2 in Chapter 2. The poor performance is due to a bias term that the plug-in estimator inherits from the nuisance parameter estimation. The one-step estimator defined in equation (2.1) uses the canonical gradient to remove the first order asymptotic bias of the plug-in estimator. To see how this works, consider the expansion,

$$\begin{aligned}
\hat{\Psi}^*_n - \Psi(P) &= \Psi(\hat{P}_n) + \mathbb{P}_n[\varphi_{\hat{P}_n}] - \Psi(P) \\
&= \Psi(\hat{P}_n) + \mathbb{P}_n[\varphi_{\hat{P}_n}] - \Psi(P) \pm n^{-1/2}\mathbb{G}_n[\varphi_{\hat{P}_n} - \varphi_P] \\
&= \mathbb{P}_n[\varphi_P] + \underbrace{\Psi(\hat{P}_n) + P[\varphi_{\hat{P}_n}] - \Psi(P)}_{(*)} + n^{-1/2}\mathbb{G}_n[\varphi_{\hat{P}_n} - \varphi_P],
\end{aligned} \tag{3.3}$$

where we use $P[\varphi_P] = 0$ for the last equality. For a fixed function $f \in \mathcal{L}^2_P$, $\mathbb{G}_n[f] \rightsquigarrow \mathcal{N}(0, P[f^2])$, and so we might hope that if $\|f_n\|_P \xrightarrow{P} 0$ then $\mathbb{G}_n[f_n] \xrightarrow{P} 0$. This is not true in general, but if the sequence of random functions belongs to a Donsker class, it is [van der Vaart, 2000, Lemma 19.24]. The Donsker class assumption can be relaxed by using sample splitting, but we do not discuss sample splitting in this synopsis. Letting $f_n = \varphi_{\hat{P}_n} - \varphi_P$ it follows from equation (3.3) that if $\varphi_{\hat{P}_n} - \varphi_P$ belongs to a Donsker class and converges to zero in probability, then $\hat{\Psi}^*_n - \Psi_P = \mathbb{P}_n[\varphi_P] + (*) + o_P(n^{-1/2})$. Hence if we can argue that $(*) = o_P(n^{-1/2})$, it follows that the one-step estimator is an efficient RAL estimator.

We now claim that the condition $(*) = o_P(n^{-1/2})$ is something we can in general expect when $\|P - P_n\| = o_P(n^{-1/4})$, for some suitable norm $\|\cdot\|$. To see this, define the path $\{\hat{P}_{n,t}\}$

as $\hat{P}_{n,t} = \hat{P}_n + t(P - \hat{P}_n)$ and consider the function $t \mapsto \Psi(\hat{P}_{n,t})$. Assuming this function is suitably smooth we can do a Taylor expansion at $t = 0$ to obtain

$$\Psi(\hat{P}_{n,1}) = \Psi(\hat{P}_{n,0}) + \frac{\partial}{\partial t}\Big|_{t=0} \Psi(\hat{P}_{n,t}) + \text{Rem}(\hat{P}_{n,1}, \hat{P}_{n,0}), \tag{3.4}$$

where $\text{Rem}(\hat{P}_{n,1}, \hat{P}_{n,0})$ is some lower order remainder term. Note that the score function of $\{\hat{P}_{n,t}\}$ at $\hat{P}_n$ is $(p - \hat{p}_n)\hat{p}_n^{-1}$, so by the definition of the canonical gradient and the path $\{\hat{P}_{n,t}\}$, equation (3.4) is equivalent to

$$\begin{aligned}
\Psi(P) &= \Psi(\hat{P}_n) + \left\langle \varphi_{\hat{P}_n}, \frac{p - \hat{p}_n}{\hat{p}_n} \right\rangle_{\hat{P}_n} + \text{Rem}(P, \hat{P}_n) \\
&= \Psi(\hat{P}_n) + \langle \varphi_{\hat{P}_n}, p - \hat{p}_n \rangle_\mu + \text{Rem}(P, \hat{P}_n) \\
&= \Psi(\hat{P}_n) + P[\varphi_{\hat{P}_n}] - \hat{P}_n[\varphi_{\hat{P}_n}] + \text{Rem}(P, \hat{P}_n) \\
&= \Psi(\hat{P}_n) + P[\varphi_{\hat{P}_n}] + \text{Rem}(P, \hat{P}_n),
\end{aligned} \tag{3.5}$$

where we use that $\varphi_{\hat{P}_n}$ is a zero-mean function under $\hat{P}_n$ for the last equality. Rearranging equation (3.5) we see that

$$(*) = -\text{Rem}(P, \hat{P}_n), \tag{3.6}$$

where $(*)$ was defined in equation (3.3). Thus $P[\varphi_{\hat{P}_n}]$ can be interpreted as a first order approximation of $\Psi(P) - \Psi(\hat{P}_n)$, and $(*)$ as a second order error term when Ψ is suitably smooth [Robins et al., 2008, 2009, Fisher and Kennedy, 2021, Kennedy, 2022]. More formally, it can in many cases be established that

$$\text{Rem}(P, \hat{P}_n) = \sum_{j,l} O_P(\|\nu_j - \hat{\nu}_{j,n}\|_P \|\nu_l - \hat{\nu}_{l,n}\|_P), \tag{3.7}$$

where $\{\nu_l : l = 1, \ldots, L\}$ is a set of nuisance parameters taking values in $\mathcal{L}_P^2$ [e.g., van der Laan and Robins, 2003, Bang and Robins, 2005, Chernozhukov et al., 2018b, Rytgaard et al., 2021, Rotnitzky et al., 2021, Kennedy, 2022]. For the illness-death model introduced in Example 2.1 we show in Manuscript I that equation (3.7) holds (Proposition 1 in Appendix C). Importantly, equation (3.7) implies that if $\|\nu_l - \hat{\nu}_{l,n}\|_P = o_P(n^{-1/4})$ for all $l = 1, \ldots, L$, then $\text{Rem}(P, \hat{P}_n) = o_P(n^{-1/2})$. By equations (3.3) and (3.6), this in turn implies that the one-step estimator $\hat{\Psi}_n^*$ is asymptotically linear when $\varphi_{\hat{P}_n}$ belongs to a Donsker class. That φ_P can be interpreted as a derivative of the map Ψ at P justifies the Taylor expansion in equation (3.5), which is the key to understand why the one-step estimator allows the use of data-adaptive estimation of nuisance parameters.

While equation (3.5) provides a general heuristic argument for why $\text{Rem}(P, \hat{P}_n) = o_P(n^{-1/2})$ when $\|\nu_l - \hat{\nu}_{l,n}\|_P = o_P(n^{-1/4})$ for all $l = 1, \ldots, L$, the term needs to be analyzed for each parameter Ψ to formally establish equation (3.7). Some general results for the form of the remainder term exist [Robins et al., 2008, Chernozhukov et al., 2018b, Rotnitzky et al., 2021]. In Section 4.3 we argue that formally bounding the remainder term posses some additional challenges for right-censored survival data observed in continuous time.

3.3 Non-parametric models

It is common to consider estimation under a so-called 'non-parametric model' [e.g., Kennedy, 2016, Benkeser et al., 2017, Carone et al., 2018, Colangelo and Lee, 2020, Rotnitzky et al., 2021, Fisher and Kennedy, 2021]. In Manuscript I we use the term 'fully non-parametric model'. This term is not particularly precise because a model can be 'not parametric' in

many ways. What we in fact mean by 'fully non-parametric' is that the tangent space is saturated, which means that $\dot{\mathcal{P}}_P = \mathcal{H}_P$ for all $P \in \mathcal{P}$, where $\mathcal{H}_P$ is the space of all zero-mean functions under P, as defined in equation (3.1). Data-adaptive methods are typically used because we are not willing to make (strict) assumptions about the model $\mathcal{P}$. Models with a saturated tangent space are an important special case. For instance, when the tangent space is saturated we see immediately from the definition of a gradient in equation (3.2) that there can exist only one gradient which will automatically be the canonical gradient. Thus, when $\mathcal{P}$ has a saturated tangent space all RAL estimators are asymptotically equivalent and efficient.

Many standard regularity assumptions lead to models with a saturated tangent space, as formally shown in Proposition 3.1. A proof is given in Appendix A.1.

Proposition 3.1. *Let $\mathcal{O} = [0,1]^d$ and $\mathcal{P}$ be the collection of all probability measures P on $\mathcal{O}$ such that $P = p \cdot \lambda$ for some density p where λ is Lebesgue measure. Define the following submodels of $\mathcal{P}$,*

- $\mathcal{P}^\varepsilon = \{P \in \mathcal{P} : \varepsilon < \|p\|_\infty < 1/\varepsilon\}$ *for some* $\varepsilon \in (0,1)$*;*
- $\mathcal{S}^k = \{P \in \mathcal{P}^\varepsilon : p \in C^k\}$*, where C^k denotes the space of k times continuously differentiable functions;*
- $\mathcal{V}^M = \{P \in \mathcal{P}^\varepsilon : p$ *is càdlàg and* $\|p\|_v < M\}$*, where $\|\cdot\|_v$ is the sectional variation norm [van der Laan, 2017b, Manuscript III];*
- $\mathcal{M} = \{P \in \mathcal{P}^\varepsilon : p$ *is non-decreasing*$\}$*.*

The models $\mathcal{P}^\varepsilon$, $\mathcal{S}^k$, $\mathcal{V}^M$, and $\mathcal{M}$ all have saturated tangent spaces.

Proposition 3.1 demonstrates that boundedness constraints, smoothness constraints, and even some shape constraints do not change the tangent space. This implies that the information bounds for estimation of a target parameter Ψ under any of the models defined in Proposition 3.1 are the same. It can also happen that the information bound is the same for $\mathcal{P}' \subset \mathcal{P}$ and $\mathcal{P}$, even though $\dot{\mathcal{P}}'_P$ is a proper subset of $\dot{\mathcal{P}}_P$. This is the case for models based on coarsened data [e.g., van der Vaart, 2000, Chapter 25.5.3]. The examples in this section serve to emphasize the asymptotic nature of the efficiency theory outlined in Section 3.1, as also discussed by Robins and Ritov [1997]. We return to this point in Section 8.

Chapter 4

Censored data

When data are censored it means that some information is missing. Heitjan and Rubin [1991] defined a general framework for missing and incomplete data using the concept of 'coarsened data'. To ensure that the distribution of the original data of interest can be identified from the distribution of the coarsened version of the data, we have to impose assumptions on the how the data were coarsened. The right assumption is that data should be coarsened at random (CAR). The original formulation of CAR given by Heitjan and Rubin [1991] was made for discrete sample spaces, and was generalized by Jacobsen and Keiding [1995] and Gill et al. [1997]. We discuss the general framework of coarsened data and CAR in Section 4.1. Identifiability conditions for right-censored data are well-studied [e.g., Lagakos and Williams, 1978, Kalbfleisch and MacKay, 1979, Heitjan, 1993, Kalbfleisch and Prentice, 2011, Andersen et al., 2012, Overgaard and Hansen, 2021, Røysland et al., 2022], and in Section 4.2 we briefly discuss right-censoring as a special case of coarsening. In Section 4.3 we consider targeted learning under CAR with a focus on right-censored survival data.

4.1 Coarsening at random

Let $Z \in \mathcal{Z}$ be a random variable of interest which is only partly observed. Heitjan and Rubin [1991] formalize 'partly observed' by saying that the data we observe no longer takes values in the sample space $\mathcal{Z}$ but instead in the power set of the sample space. The observed variable is a coarsened version of the original variable Z of interest, because the observed subset includes Z, but potentially also many other points. In this way, only partial information about Z is conserved.

Example 4.1 (Illness-death model)
Consider the illness-death model introduced in Example 2.1 but assume for simplicity that no baseline variables are measured. Define T_0 as the time at which a patient leaves state 0, and T as the time at which a patient enters state 2. If a patient drops out of the study at a random time point, we only have partial information about T_0 and T. Each observation corresponds to knowing that (T_0, T) belongs to a subset of $\mathbb{R}^2$. We illustrate this in Figure 4.1. Here $Z = (T_0, T)$ is the full data while the gray subsets represent possible coarsened observations of Z. •

A coarsened data model can be specified by a model $\mathcal{Q}$ for the full data, together with a model $\mathcal{G}$ for the conditional coarsening mechanism [Gill et al., 1997]. We follow Nielsen [2000] and van der Laan and Robins [2003] and explicitly define a coarsening variable C.

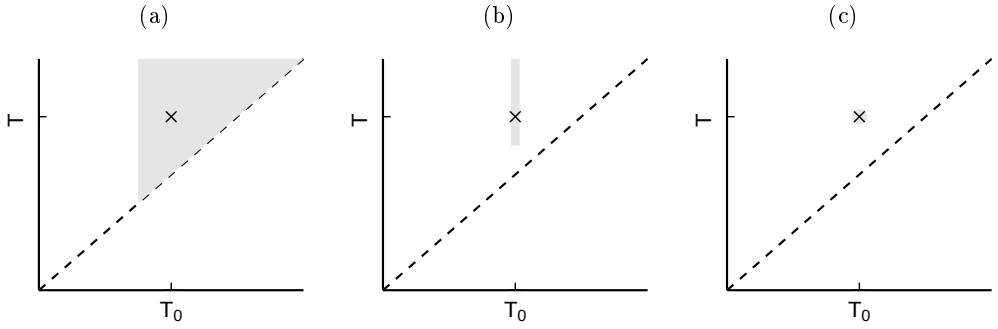


Figure 4.1: Visualization of coarsened data for the illness-death model. The cross represents the full data $Z = (T_0, T)$ and the gray areas the coarsened data determined by when a patient leaves the study. Note that the dashed line corresponds to observations where a patient dies without falling sick. In panel (a) the patient leaves the study before leaving the healthy state. In panel (b) the patient is observed to fall sick but leaves the study before dying. In panel (c) we observe the full data as the singleton $\{Z\}$.

What we observe is

$$\mathcal{X} = \Phi(Z, C) \quad \text{such that almost surely} \quad Z \in \mathcal{X}, \tag{4.1}$$

for some known function Φ, with $Z \sim Q$ for some $Q \in \mathcal{Q}$ and $C \mid Z = z \sim G(\cdot \mid z)$ for some $G \in \mathcal{G}$. Given Φ, the observed data are completely determined by Q and G, and so we can define the observed data distribution as $\mathcal{P} = \{P_{Q,G} : Q \in \mathcal{Q}, G \in \mathcal{G}\}$, where $P_{Q,G}$ denotes the distribution of the random variable $\mathcal{X}$ defined in equation (4.1).

If we assume nothing about the coarsening mechanism we cannot expect to identify Q from an observed data distribution P. First of all, we need to assume that the conditional probability of observing fully informative, uncoarsened data is positive – this is referred to as positivity. When working with coarsened data, however, positivity is not enough. To ensure identifiability, we can make the additional assumption of CAR. Roughly speaking, CAR states that for any $A \subset \mathcal{Z}$ and $z \in A$, knowing that $Z = z$ provides no more information about $\mathcal{X}$ than knowing that $Z \in A$. Intuitively, this means that the conditional censoring mechanism is only a function of the observed data. Under CAR we can thus learn the censoring mechanism from the data we get to observe – additional information about the full data is unnecessary.

When CAR holds, the likelihood for the observed data factorizes into components indexed by $Q \in \mathcal{Q}$ and $G \in \mathcal{G}$, and when positivity holds, Q is identifiable from the observed data distribution P [Robins and Rotnitzky, 1992, van de Laan, 1995, Gill et al., 1997]. Importantly, CAR can be said to be the minimal requirement for identifiability, in the sense that while CAR restricts the joint model for Z and C, the model $\mathcal{P}$ remains unrestricted when $\mathcal{Q}$ is unrestricted. More precisely, when $\mathcal{Q}$ is unrestricted the tangent space for the observed data model is saturated [Gill et al., 1997, van der Vaart, 2000].

4.2 Right-censored data

While the framework of set-valued random variables elegantly captures the idea that only partial information is available in the observed data, it is often more natural to think of the observed data in another manner. For instance, when a stochastic process is subject to right-censoring, van der Vaart [2004] represents the coarsened data as a stopped process

where the stopping time is the censoring time, and defines CAR directly with reference to the conditional density for the stopped process. In Manuscript I we followed this approach and exploited that all information about an illness-death process can be captured by the random times at which the process changes state. In this way, CAR reduces to a requirement concerning ordinary conditional probabilities for Euclidean random variables. We illustrate how this works in a simple setting in the following.

A common formulation of a right-censored survival problem is to say that we observe $O = (\tilde{T}, \Delta, W)$, where $W \in \mathbb{R}^d$ is a vector of baseline covariates, $\tilde{T} = T \wedge C$, and $\Delta = \mathbb{1}\{T \leq C\}$ for some event time $T \in \mathbb{R}_+$ and censoring time $C \in \mathbb{R}_+$. The observations $(\tilde{T}, \Delta)$ can be bijectively mapped to subsets of $\mathbb{R}$ as follows,

$$\begin{aligned} \mathbb{R} \times \{0,1\} \ni (t,0) &\iff (t, \infty) \subset \mathbb{R}, \\ \mathbb{R} \times \{0,1\} \ni (t,1) &\iff \{t\} \subset \mathbb{R}. \end{aligned} \tag{4.2}$$

Formally, the coarsened data $\mathcal{X}$ defined in equation (4.1) are the subsets on the right-hand side of equation (4.2), but the bijective correspondence means that we can express CAR as an assumption about the conditional distribution of O given (T, W). Writing $P_{(\tilde{T},\Delta)|(T,W)}$ for the conditional distribution of $(\tilde{T}, \Delta)$ given (T, W), CAR is the statement that for all $t' \leq t$, $P_{(\tilde{T},\Delta)|(T,W)}(\mathrm{d}t', \delta \mid t, w)$ does not depend on t. As $\tilde{T} = T$ when $\Delta = 1$ we can write

$$\begin{aligned} &P_{(\tilde{T},\Delta)|(T,W)}(\mathrm{d}t', \delta \mid t, w) \\ &= \mathbb{1}_{\{1\}}(\delta) P_{(\tilde{T},\Delta)|(T,W)}(\mathrm{d}t', 1 \mid t', w) + \mathbb{1}_{\{0\}}(\delta) P_{(\tilde{T},\Delta)|(T,W)}(\mathrm{d}t', 0 \mid t, w) \\ &= \mathbb{1}_{\{1\}}(\delta) P_{C|(T,W)}(C \geq t' \mid t', w) + \mathbb{1}_{\{0\}}(\delta) P_{C|(T,W)}(C < t' \mid t, w). \end{aligned} \tag{4.3}$$

It is common to assume that $T \perp\!\!\!\perp C \mid W$, in which case $P_{C|(T,W)}(C < t' \mid t, w) = P_{C|W}(C < t' \mid w)$. Thus $T \perp\!\!\!\perp C \mid W$ implies that the right-hand side in equation (4.3) does not depend on t, which shows that conditional independence of the event and censoring time implies CAR. On the other hand, conditional independence is stronger than CAR. For instance, if $C = \Delta(T + 5) + (1 - \Delta)C^\circ$, for $C^\circ \perp\!\!\!\perp T \mid W$, in general $C \not\perp\!\!\!\perp T \mid W$, but $P_{C|(T,W)}(C < t' \mid t, w) = P_{C^\circ|(T,W)}(C^\circ < t' \mid t, w) = P_{C^\circ|W}(C^\circ < t' \mid w)$ for $t' \leq t$, so CAR still holds.

For right-censored data, the observed data distribution $P_{Q,G}$ is indexed by a model $Q \in \mathcal{Q}$ for the full data $Z = (T, W)$ and a conditional censoring mechanism $G \in \mathcal{G}$. When we assume that the observed data is generated by conditionally independent event and censoring times, $\mathcal{G}$ can be taken to be a family of conditional survival functions governing the conditional distribution of C given W.

In many biostatistical applications it is sufficient to identify the distribution of T only on some interval $[0, \tau]$, for some $\tau < \infty$. For instance, we are often interested in survival probabilities up to some fixed time point τ. When this is the case, we could argue that an observation for which $\Delta = 0$ and $\tilde{T} > \tau$ should not in fact be treated as a censored observation because it contains full information about what happened on the interval $[0, \tau]$. We can accommodate this by, for example, defining the full data as $Z_\tau = (T_\tau, W)$, where

$$T_\tau = \begin{cases} T & \text{if } T \leq \tau \\ \infty & \text{if } T > \tau \end{cases},$$

and the observed, coarsened data as $O_\tau = (\tilde{T}_\tau, \Delta_\tau, W)$, where

$$\tilde{T}_\tau = \mathbb{1}\{C > \tau\} T_\tau + \mathbb{1}\{C \leq \tau\} \tilde{T} \quad \text{and} \quad \Delta_\tau = \Delta \mathbb{1}\{\tilde{T} \leq \tau\} + \mathbb{1}\{\tilde{T} > \tau\}$$

Assuming CAR for O_τ (given Z_τ) is weaker than assuming CAR for O (given Z). Another way to accommodate a time-horizon is to truncate all observed event times with values above

τ to τ and treat them as censored, which again implies a weaker CAR assumption than what is needed for CAR to hold for O. The latter strategy reflects the common situation where patients in a study will be subject to administrative censoring at the end of the study period. In either case, we introduce (artificial) point-mass for the observed data at ∞ or τ, which can be a bit annoying. We dealt explicitly with this in Manuscript III.

4.3 Targeted learning under coarsening at random

CAR fits well into the general setting of semi-parametric efficiency theory outlined in Section 3.1 as demonstrated by Robins and Rotnitzky [1992], van de Laan [1995], Gill et al. [1997], and van der Vaart [1991, 2000, 2004]. In particular, for the special case of censored longitudinal data, the tangent space can be represented as a collection of martingale integrals, and projections onto the tangent space can be calculated as martingale integrals of conditional expectations. This representation can be used to derive candidate influence functions for an observed data model $\mathcal{P}$ when an influence function for the full data model $\mathcal{Q}$ is known, and the projection formula provides a general recipe for finding the canonical gradient in such models. These results are expanded in detail in [van der Laan and Robins, 2003] and [Tsiatis, 2007] and specific attention to the case when observations are made in continuous-time are given in [van der Vaart, 2004] and [Rytgaard et al., 2022].

Under CAR and positivity, we know that Q is identifiable from P which means that there is a map $\upsilon\colon \mathcal{P} \to \mathcal{Q}$ such that $\upsilon(P_{Q,G}) = Q$ for all $Q \in \mathcal{Q}$ and $G \in \mathcal{G}$. If θ is a target parameter defined on $\mathcal{Q}$, an alternative strategy for finding the canonical gradient of $\Psi = \theta \circ \upsilon$ is to start directly with this functional. When the parameter is $\theta\colon \mathcal{Q} \to \mathbb{R}$ is linear, the derivation of the canonical gradient under a model with a saturated tangent space is straightforward. However, as $\upsilon\colon \mathcal{P} \to \mathcal{Q}$ is typically not linear the derivation of the canonical gradient of $\Psi\colon \mathcal{P} \to \mathbb{R}$ is more complicated.

We now highlight a particular challenge facing targeted learning based on right-censored survival data observed in continuous time. To simplify the exposition we consider the setting where no baseline covariates are measured. Recall from Section 3.2 that a central ingredient of targeted learning was that the remainder term $(*)$ defined in equation (3.3) could be interpreted as a second order error term. For right-censored data, the remainder is typically composed of terms on the form

$$\hat{R}_n = \int_0^\tau \hat{H}_n(s)[\hat{\Lambda}_n - \Lambda](\mathrm{d}s), \tag{4.4}$$

where Λ is a cumulative hazard function and $\hat{\Lambda}_n$ an estimator of Λ [e.g., van der Laan and Robins [2003], Rytgaard et al. [2021, 2022], Manuscript I]. Here $\hat{H}_n$ is an estimation error, for instance $\hat{H}_n = \hat{S}_n - S$, for a survival function S and an estimator $\hat{S}_n$ of S. If there exists a fixed σ-finite measure μ such that we can write

$$[\hat{\Lambda}_n - \Lambda] = [\hat{\lambda}_n - \lambda] \cdot \mu, \quad \text{for all} \quad n \in \mathbb{N}, \tag{4.5}$$

for some functions λ and $\hat{\lambda}_n$, we have

$$\hat{R}_n = \int_0^\tau \hat{H}_n(s)[\hat{\lambda}_n(s) - \lambda(s)]\mu(\mathrm{d}s).$$

When $\mu \ll P$ it follows from the Cauchy-Schwarz inequality that

$$|\hat{R}_n| \leq \|\hat{H}_n\|_P \|\hat{\lambda}_n - \lambda\|_P.$$

In this case, $\hat{R}_n$ can be interpreted as a second order remainder term because if $\|\hat{H}_n\|_P = o_P(r_n^{-1})$ and $\|\hat{\lambda}_n - \lambda\|_P = o_P(r_n^{-1})$ for some rate $r_n \to \infty$ then $\hat{R}_n = o_P(r_n^{-2})$.

Unfortunately, for many estimators used in survival analysis, equation (4.5) does not hold. For instance, if Λ is estimated with the Nelson-Aalen estimator [Nelson, 1969, 1972] and observations are made in continuous time, there is no fixed measure μ such that equation (4.5) holds. It appears challenging to provide useful general bounds on terms on the form given in equation (4.4). Gill et al. [1995] provide bounds on the form

$$\hat{R}_n \lesssim \|\hat{H}_n\|_\infty \|\hat{\Lambda}_n - \Lambda\|_v, \tag{4.6}$$

where $\|\cdot\|_v$ is the total variation norm and $\|\cdot\|_\infty$ is the supremum norm for functions with domain $[0, \tau]$. The total variation is however too strong a norm to be useful in this case. Assume again that $\hat{\Lambda}_n$ is the Nelson-Aalen estimator of Λ, and that Λ is absolutely continuous. Considered as measures, $\hat{\Lambda}_n$ and Λ are mutually singular, and hence it follows from, e.g., Theorem 3 in [Aistleitner and Dick, 2014] and the Jordan-Hahn decomposition, that $\|\hat{\Lambda}_n - \Lambda\|_v = \hat{\Lambda}_n(\tau) + \Lambda(\tau)$. Thus, while $\|\hat{\Lambda}_n - \Lambda\|_\infty = O_P(n^{-1/2})$ [Andersen et al., 2012], we have $\|\hat{\Lambda}_n - \Lambda\|_v \to 2\Lambda(\tau) > 0$. This demonstrates that even when we can establish the fast rates $\|\hat{\Lambda}_n - \Lambda\|_\infty = O_P(n^{-1/2})$ and $\|\hat{H}_n\|_\infty = O_P(n^{-1/2})$, equation (4.6) might not provide a better bound than $\hat{R}_n = O_P(n^{-1/2})$.

Without imposing some assumptions on the structure of $\hat{\Lambda}_n$ we cannot in general improve the bound in equation (4.6); for instance, in Appendix A.2 we exhibit functions $\hat{H}_n$ and $\hat{\Lambda}_n - \Lambda$ such that $\|\hat{H}_n\|_\infty \to 0$ and $\|\hat{\Lambda}_n - \Lambda\|_\infty \to 0$, but $\hat{R}_n \to \infty$. In Appendix C.1 of Manuscript I we use empirical process theory to bound terms of the form given in equation (4.4) for the special case when Λ is estimated with the Nelson-Aalen estimator. We briefly discuss possible extensions of this approach in Section 8.

Chapter 5

Super learning with right-censored data

Most modern data-adaptive estimators depend on tuning hyperparameters. For example, the lasso depends on the choice of L_1 penalty [Tibshirani, 1996], a spline model on the number of knot points [Wahba, 1990, Wood, 2017], and a neural network on the number of hidden layers [Rosenblatt, 1958, Rumelhart et al., 1986]. The tuning of hyperparameters can be phrased as a model selection problem, because each choice of hyperparameter provides a model. In practice, it is difficult to pre-specify a suitable model and the values for hyperparameters.

A common approach to this problem is cross-validation, where data are split in training and test samples such that models are fitted and their performance evaluated in independent data sets [Stone, 1974, Geisser, 1975]. More generally, the super learner [van der Laan et al., 2007] is a meta-algorithm for combining a given set of algorithms or learners into a new learner. In the language of super learning, models or algorithms are referred to as 'learners', and a family of models is referred to as a 'library' of learners. A library can be a family of models indexed by a hyperparameter, but it can also be a collection of unrelated parametric, semi-, and non-parametric models. Another word for super learning is 'stacked regression' [Wolpert, 1992, Breiman, 1996], and ordinary cross-validation is equivalent to the so-called discrete super learner which combines learners by simply picking the one that performs best.

Importantly, the super learner can be shown to behave almost as well as the best learner in the library. This essentially means that the price we pay to use the data to select a learner from the library is very small, at least in terms of convergence rates. For instance, if a library contains a learner that is consistent at some rate, then the super learner will be consistent at the same rate up to a potential factor of $\log(M_n)$, where M_n is the number of learners in the library [van der Laan and Dudoit, 2003, van der Vaart et al., 2006].

Formally, a super learner for some parameter $\Psi\colon \mathcal{P} \to \Theta$ is defined using a loss function $L\colon \Theta \times \mathcal{O} \to \mathbb{R}_+$ and a library of learners $\mathcal{A}$. Each element $a \in \mathcal{A}$ is a deterministic[†] map $\mathbb{P}_n \mapsto a(\mathbb{P}_n) \in \Theta$, for all $n \in \mathbb{N}$. For some $K_n \in \{1, \ldots, n-1\}$, consider a (random) partition of the data set $\{O_i\}_{i=1}^n$ into K_n subsets, which are referred to as folds. Define $\mathbb{P}_n^k$ as the empirical measure of the k'th fold and $\mathbb{P}_n^{-k}$ as the empirical measure of all the data excluding the k'th fold. For all $a \in \mathcal{A}$ and $i \in \{1, \ldots, n\}$, define

$$l_i(a) = L(a(\mathbb{P}_n^{-k}), O_i), \quad \text{for } O_i \in \mathbb{P}_n^k, \text{ and } k \in \{1, \ldots, K_n\}.$$

A super learner can be defined as the learner $\sum_{a \in \mathcal{A}} \hat{u}_n(a)a$, where the weights $\hat{u}_n(a)$ are constructed using the holds-out losses $\{(l_i(a) : a \in \mathcal{A})\}_{i=1}^n$.

[†]For many algorithms, a might in fact not be deterministic for fixed $\mathbb{P}_n$. For instance, random forests use random splitting and can thus return different estimates when applied twice to the same data set. For simplicity, we ignore this in the following.

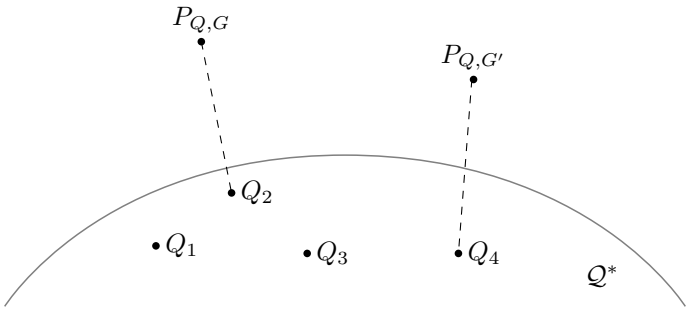


Figure 5.1: The least false model in the family $\mathcal{Q}^*$ according to the negative partial log-likelihood depends on the whole data-generating distribution $P_{Q,G}$.

We see that the super learner itself depends on several hyperparameters: The loss function L, the library $\mathcal{A}$, the number of folds K_n, the method used to split the data into folds, and the method used to construct the weights $\hat{u}_n(a)$. We consider the case where K is fixed and the random partition yields approximately equally sized folds. We focus on the discrete super learner which uses the weights $\hat{u}_n(a) = \mathbb{1}\{a = \min_{a' \in \mathcal{A}} \sum_{i=1}^n l_i(a')\}$, assuming the minimizer is unique. In this chapter we discuss the choice of loss function when data are right-censored. Sections 5.1 and 5.2 discuss two of the most common choices, and Section 5.3 briefly introduce our proposal for a new super learner given in Manuscript II. Throughout this chapter we consider the data setting described in Section 4.2, i.e., we assume that the available data $O \sim P$ are on the form $O = (\tilde{T}, \Delta, W)$ where $\tilde{T} = T \wedge C$ and $\Delta = \mathbb{1}\{T \leq C\}$ for some event and censoring times T and C such that $T \perp\!\!\!\perp C \mid W$. We recall that we use $Q \in \mathcal{Q}$ to denote the distribution of $Z = (T, W)$ and $G \in \mathcal{G}$ to denote the conditional survival function for the censoring distribution.

5.1 The negative partial log-likelihood loss

A commonly used loss function in survival analysis is the negative log of the partial likelihood [Cox, 1975, Andersen et al., 2012, Liestbl et al., 1994, Li et al., 2016, Bender et al., 2020, Kvamme and Borgan, 2021, Lee et al., 2021]. As discussed in Section 4.1, the likelihood for the observed data factorizes into components indexed by $Q \in \mathcal{Q}$ and $G \in \mathcal{G}$. This means that the partial log-likelihood loss can be optimized for the parameter Q without specifying or modeling the censoring distribution. The average loss, however, can still depend on the censoring distribution, because the average is taken with respect to the observed data distribution.

In is well-known that parameters estimated in survival analysis can depend on the censoring distribution and the dependence has been studied in detail for the Cox model [Struthers and Kalbfleisch, 1986, Hjort, 1992, Fine, 2002, Whitney et al., 2019]. The dependence on the censoring distribution can be understood by relating the negative partial log-likelihood to the Kullback-Leibler divergence. Maximum likelihood estimation is equivalent to minimizing the Kullback-Leibler divergence, which is defined as

$$D_{\mathrm{KL}}(P_1 \,||\, P_2) = P_1\left[\log \frac{p_1}{p_2}\right], \quad \text{where} \quad P_1 = p_1 \cdot \nu, \quad \text{and} \quad P_2 = p_2 \cdot \nu,$$

for some σ-finite measure ν such that $\{P_1, P_2\} \ll \nu$. For a sub-family $\mathcal{P}^* \subset \mathcal{P}$ it holds under regularity conditions that the limit of the maximum likelihood estimator based on independent samples from P_0 is the minimizer of $P \mapsto D_{\mathrm{KL}}(P_0 \,||\, P)$ over $\mathcal{P}^*$ [e.g., van der

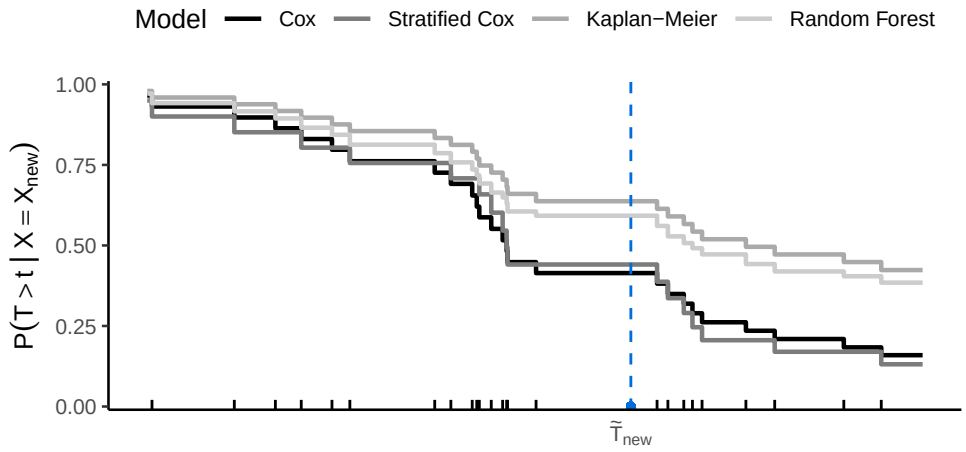


Figure 5.2: Estimated conditional survival curves for two Cox models [Cox, 1972, Therneau, 2022], the Kaplan-Meier estimator [Kaplan and Meier, 1958, Gerds, 2019], and a random survival forest [Ishwaran et al., 2008, Ishwaran and Kogalur, 2023] fitted in R [R Core Team, 2020]. All estimators assign probability only to the ticks at the x-axis, which denote observed time points in the training sample where an event occurred. A new sample with event time $\tilde{T}_{\text{new}}$ occurs with probability zero according to any of these four models.

Vaart, 2000, Wainwright, 2019]. This also holds when $P_0 \notin \mathcal{P}^*$ and thus maximum likelihood estimation under a misspecified family can be understood as estimating the model that minimizes the Kullback-Leibler divergence to the data-generating distribution. The minimizing model is referred to as the least false model.

Consider now a situation were the likelihood for $P \in \mathcal{P}$ factorizes into two terms, say $\ell_1(\theta, O)$ and $\ell_2(\nu, O)$ for two variational independent parameters θ and ν determining the distribution $P = P_{\theta,\nu}$. Using the definition of D_{KL} we can write the expected loss under P_{θ_0,ν_0} according to the loss function $-\log \ell_1$ as

$$\begin{aligned}
&P_{\theta_0,\nu_0}[-\log \ell_1(\theta, \cdot)] \\
&= P_{\theta_0,\nu_0}[-\log \ell_1(\theta, \cdot)] \pm P_{\theta_0,\nu_0}[\log \ell_2(\nu_0, \cdot) + \log \ell_1(\theta_0, \cdot)] \\
&= P_{\theta_0,\nu_0}[\log \ell_2(\nu_0, \cdot) + \log \ell_1(\theta_0, \cdot) - \{\log \ell_2(\nu_0, \cdot) + \log \ell_1(\theta, \cdot)\}] \\
&\quad - P_{\theta_0,\nu_0}[\log \ell_1(\theta_0, \cdot)] \\
&= D_{\text{KL}}(P_{\theta_0,\nu_0} \,||\, P_{\theta,\nu_0}) - P_{\theta_0,\nu_0}[\log \ell_1(\theta_0, \cdot)].
\end{aligned} \tag{5.1}$$

As $P_{\theta_0,\nu_0}[\log \ell_1(\theta_0, \cdot)]$ does not depend on θ, we see from equation (5.1) that minimizing the risk $P_{\theta_0,\nu_0}[-\log \ell_1(\theta, \cdot)]$ with respect to θ is equivalent to minimizing the Kullback-Leibler divergence $D_{\text{KL}}(P_{\theta_0,\nu_0} \,||\, P_{\theta,\nu_0})$ with respect to θ. Thus, loss-based estimation with respect to a partial log-likelihood loss can still be interpreted as estimating a least false model, but the distance with respect to which the least false model is defined depends on the whole distribution P_{θ_0,ν_0} and not only on θ_0.

The argument above applies to any factorizing likelihood, and hence to any CAR model. For the special case of right-censored data, we can use equation (5.1) with the parametrization $\theta = Q$ and $\nu = G$. It follows from properties of the Kullback-Leibler divergence that Q will always have smallest risk according to the partial log-likelihood under $P_{Q,G}$ for all $G \in \mathcal{G}$. However, for a misspecified model $\mathcal{Q}^* \subset \mathcal{Q}$ that does not contain Q, the least false model in $\mathcal{Q}^*$ depends on G. This point is illustrated in Figure 5.1.

From a practical perspective, a perhaps more pressing concern is that the partial likelihood

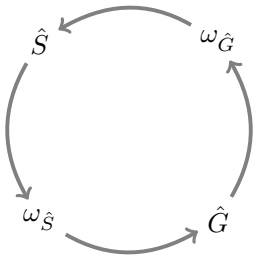


Figure 5.3: To use an inverse probability of censoring weighted loss function for estimation S we need the inverse probability of censoring weight $\omega_{\hat{G}}$, which depends on an estimator of G. To estimate G we need the inverse probability of event weight $\omega_{\hat{S}}$, which in turn depends on an estimator of S.

is useless for evaluating performance of the most commonly used survival estimators. Many common estimators of the conditional survival function are piece-wise constant with jumps at the event times observed in the training sample; some examples are given in Figure 5.2. The likelihood according to a model with a piece-wise constant survival function is zero for all time points at which the survival function does not jump. When data are recorded in continuous time, we never observe the exact same event times in an independent test sample. Thus any such estimator almost surely assigns zero likelihood to any independent test sample, and so any such estimator will have infinite loss in any test sample.

5.2 Inverse probability of censoring weighted loss functions

A theoretically elegant way of selecting a loss function for super learning is to first decide on a loss function that is defined for the full data $Z = (T, W) \sim Q$. In this section we assume that the parameter of interest is the conditional survival function $S(t \mid w) = Q(T > t \mid W = w)$, which we sometimes write as S_Q to emphasize that S is associated with a particular measure Q. A loss function for S with respect to the full data is a functional $L^F \colon \mathcal{S} \times \mathcal{Z} \to \mathbb{R}$, where $\mathcal{S}$ is the class of all conditional survival functions. Under CAR, the average loss under Q according to L^F can be identified from censored data using inverse probability of censoring weights [Graf et al., 1999, van der Laan and Dudoit, 2003, Hothorn et al., 2006, Gerds and Schumacher, 2006]. The inverse probability of censoring weighted version of L^F is the functional $L(\cdot, \cdot, G) \colon \mathcal{S} \times \mathcal{O} \to \mathbb{R}$ indexed by $G \in \mathcal{G}$, defined as

$$L(S, O; G) = L^F(S, (\tilde{T}, W))\Delta\omega_G(\tilde{T}, W), \quad \text{where} \quad \omega_G(t, w) = G(t- \mid w)^{-1}. \tag{5.2}$$

Under regularity conditions and CAR it holds that

$$P_{Q,G}[L(S, \cdot; G)] = Q[L^F(S, \cdot)], \quad \text{for all } S \in \mathcal{S}, Q \in \mathcal{Q}, \text{ and } G \in \mathcal{G}.$$

In most applications we are typically only able to estimate $S(\cdot \mid w)$ on some bounded interval $[0, \tau]$, and we discuss this in Appendix A.3. For simplicity of exposition we consider the unbounded case in this section.

In practice, G is unknown and has to be estimated to construct the weights ω_G. When little is known about G, one strategy is to build a super learner for G. This can be done in much the same way as above by noting that when G is the parameter of interest, observations with $\Delta = 0$ are now fully informative, while $\Delta = 1$ can be interpreted as meaning that information about the exact censoring time is missing; see Figures 5.4 (a) and (b) for an illustration of this point. To construct a super learner for G we can reverse the roles of S

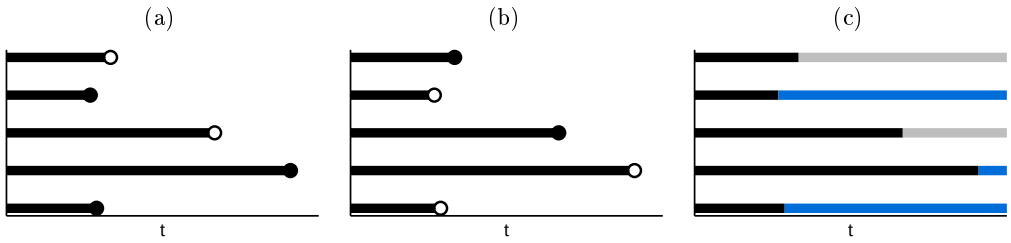


Figure 5.4: Three different interpretations of the same observed data. Each line represents an observation. Panel (a) gives the standard interpretation of right-censored data, where a black dot means that an event was observed, while a white dot means that the subject was censored. When the censoring distribution is the parameter of interest, the censored observations are interpreted as fully informative, which is illustrated in panel (b). In panel (c) each observation is considered to be in one of three mutually exclusive states at each time point.

and G and define the S-indexed loss function

$$L_c(G, O; S) = L^F(G, (\tilde{T}, W))(1 - \Delta)\omega_S(\tilde{T}, W),$$

as then (under regularity conditions),

$$P_{Q,G}[L_c(G^*, \cdot; S_Q)] = (\bar{G} \otimes H_Q)[L^F(G^*, \cdot)], \quad \text{for all } G, G^* \in \mathcal{G} \text{ and } Q \in \mathcal{Q},$$

where H_Q denotes the marginal distribution of W and $\bar{G} = 1 - G$. The problem with this approach is that we make estimation of G dependent on an estimate of S, which was the estimation problem we originally wanted to solve. If we are not willing to pre-specify an estimator of the censoring distribution G, we enter a circular estimation problem as illustrated in Figure 5.3.

To avoid the need to pre-specify an estimator of the censoring distribution, Han et al. [2021] and Westling et al. [2021] recently suggested to enter the circle in Figure 5.3 with some initial estimator $\hat{G}^{(0)}$, follow the arrows, and wait for convergence. More formally, the idea is to use two libraries $\mathcal{A}$ and $\mathcal{B}$ of learners for estimating S and G, respectively. We start by picking an estimator of the censoring distribution, $\hat{G}_n^{(0)}$, and then find the estimator $\hat{S}_n^{(1)}$ as the learner in $\mathcal{A}$ with best performance according to the loss $L(\cdot, \cdot; \hat{G}_n^{(0)})$; we then find the estimator $\hat{G}_n^{(1)}$ as the learner in $\mathcal{B}$ with best performance according to the loss $L_c(\cdot, \cdot; \hat{S}_n^{(1)})$; and so on. It is an open question whether this algorithm will always converge, and whether it will converge to the pair (S_Q, G) corresponding to the data-generating distribution $P_{Q,G}$ when the libraries $\mathcal{A}$ and $\mathcal{B}$ contain learners that are consistent for S_Q and G, respectively. We conjecture that convergence will in general be difficult to guarantee. In Appendix A.4 we investigate a population version of this algorithm and provide an example where there exist an infinite number of points toward which the algorithm can converge.

5.3 The state learner

Except for the suggestion made by Han et al. [2021] and Westling et al. [2021] that we discussed at the end of Section 5.2, most existing approaches for super learning with survival data do not address the fundamental problems we have illustrated in Figures 5.2 and 5.3. Polley and van der Laan [2011] considered super learning for right-censored data but assumed that observations were made in discrete time. Verweij and van Houwelingen [1993] and Golmakani and Polley [2020] constructed a super learner based on the partial likelihood for

Cox models [Cox, 1972], which restricts the library to contain only learners on a special form. Other approaches employ inverse propbability of censoring weights but (tacitly) assume that we know how to model the censoring distribution [Molinaro et al., 2004, Keles et al., 2004, Hothorn et al., 2006, Gonzalez Ginestet et al., 2021]. When an estimator of the censoring distribution can be pre-specified, it is also possible to use censoring unbiased transformations [Fan and Gijbels, 1996, Steingrimsson et al., 2019] or pseudo-values [Andersen et al., 2003, Mogensen and Gerds, 2013, Sachs et al., 2019] to construct a super learner. Finally, it has been suggested to use Harrell's c-index [Brown et al., 1974, Harrell et al., 1982, 1996] to evaluate performance in hold-out samples [Simon et al., 2011, Zhao and Feng, 2020] as this can be calculated without estimating the censoring mechanism. Harrell's c-index, however, is determined by the censoring distribution, so it would seem a better choice to use the c-index that is defined for the full data and identified through inverse probability of censoring weights [Gerds et al., 2013]. This again requires estimation of the censoring mechanism. Furthermore, the c-index is not necessarily maximized at the data-generating distribution [Blanche et al., 2019], which means that the c-index is not a promising performance measure. In particular, a super learner that includes a consistent learner in its library might not itself be consistent when the c-index is used for evaluating performance.

In Manuscript II we propose an alternative approach that is based on viewing the observed data as a simple multi-state system with two absorbing states. We refer to this super learner as the state learner. The observed multi-state system is illustrated in Figure 5.4 (c). As all states except the initial state are absorbing, the system can be described completely through the state occupation probabilities. Letting $\eta(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 1\} + 2\mathbb{1}\{\tilde{T} \leq t, \Delta = 0\}$, the conditional state occupation probability is

$$F(t, k, x) = P(\eta(t) = k \mid X = x), \quad \text{for} \quad t \in [0, \tau], k \in \{0, 1, 2\}, x \in \mathbb{R}^d. \tag{5.3}$$

We suggest to build a super learner for the function F. As described in Manuscript II, a library for learning F can be constructed from libraries for learning S and G, and the state learner can easily be extended to settings where a competing risk is present. To evaluate the performance of learners of F, we suggest to use the integrated Brier score, but other loss functions could be used. As F is a feature of the observed data distribution, a loss function for F will not depend on unknown nuisance parameters. We provide a summary of the results we have established for the state learner in Section 7.

Chapter 6

The highly-adaptive lasso

The highly-adaptive lasso (HAL) is an example of a non-parametric function-valued estimator than can be used to estimate parameters $\Psi\colon \mathcal{P} \to \mathcal{F}$, defined as

$$\Psi(P) = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, P[L(f, \cdot)], \tag{6.1}$$

for some suitable loss function L and function space $\mathcal{F}$. Examples include densities, regression functions, and hazard functions. Replacing P with $\mathbb{P}_n$ leads to estimators known as M-estimators, minimal contrast estimator, or empirical risk minimizers [Huber et al., 1967, Pfanzagl, 1969, Reiss, 1978, Vapnik, 1991, van der Vaart and Wellner, 1996, van der Vaart, 2000]. Sieve estimators use models $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of increasing complexity to estimate Ψ [Grenander, 1981, Geman, 1981, Geman and Hwang, 1982, Walter and Blum, 1984, Shen, 1997]. In Manuscript III we argue that the HAL estimator should be defined and interpreted as a data-adaptive sieve estimator.

Several loss functions have been used to define a HAL estimator, but all HAL estimators use $\mathcal{F} = \mathcal{D}_M^d$ for some $M \in (0, \infty)$, with $\mathcal{D}_M^d$ denoting the class of càdlàg functions $f\colon [0,1]^d \to \mathbb{R}$ with sectional variation norm bounded by M [van der Laan, 2017b, Benkeser and van der Laan, 2016, Bibaut and van der Laan, 2019, Hejazi et al., 2020, Malenica et al., 2023, Manuscript III]. Any element of $\mathcal{D}_M^d$ generates a signed measure with total variation bounded by M. Perhaps surprisingly, this guarantees that Ψ can be estimated at a rate faster than $n^{-1/4}$ for any dimension $d \in \mathbb{N}$. This essentially 'dimension-free' rate of convergence is the main reason for studying the HAL estimator.

Manuscript III provides a detailed description of the space of càdlàg functions and the sectional variation norm. In this chapter we give a brief overview of the proof techniques used to establish the rate of convergence for the HAL estimator, and we explain why some care is needed when densities and hazard functions are the parameters of interest. In Section 6.1 we discuss metric entropies, which provide a classical way to derive convergence rates of M-estimators [van de Geer, 1990, 1993, Birgé and Massart, 1993, Shen and Wong, 1994, van der Vaart and Wellner, 1996]. In Section 6.2 we discuss the assumption of a bounded sectional variation norm in the context of targeted learning and right-censored data.

6.1 Bracketing entropy

We consider data with values in $\mathcal{O} = [0,1]^d$ and let $\mathcal{F}$ denote a class of measurable real-valued functions with domain $[0,1]^d$. In this section we use bold font to denote a (non-random) point $\mathbf{x} \in [0,1]^d$. For a loss function $L\colon \mathcal{F} \times [0,1]^d \to \mathbb{R}$, define the function class

$\mathcal{L} = \{L(f, \cdot) : f \in \mathcal{F}\}$. For all $L \in \mathcal{L}$, $\mathbb{G}_n[L]$ is a random variable, which means that we can think of $\mathbb{G}_n$ as a stochastic process indexed by $L \in \mathcal{L}$. Letting $L_P = L(\Psi(P), \cdot)$, the modulus of continuity of this stochastic process is defined for $\delta > 0$ and $n \in \mathbb{N}$ as[†]

$$\Gamma_n(\delta) = \sup_{L \in \mathcal{L}_\delta(P)} |\mathbb{G}_n[L - L_P]|, \quad \text{where} \quad \mathcal{L}_\delta(P) = \{L \in \mathcal{L} : \|L - L_P\|_P < \delta\}.$$

By studying the behavior of the modulus of continuity for $\delta \downarrow 0$ and $n \to \infty$, we can derive upper bounds for how fast an empirical risk minimizer converges to $\Psi(P)$, see for instance [van der Vaart and Wellner, 1996, Theorem 3.4.1] or [Kosorok, 2008, Theorem 14.4].

The modulus of continuity can be bounded by bounding the complexity of the function class $\mathcal{F}$ [van der Vaart and Wellner, 1996, 2011]. One way to measure the complexity of a function class is through the bracketing entropy integral. For the $\mathcal{L}_P^2$ norm, the bracketing entropy integral is defined for $\delta > 0$ as

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

where $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_P)$ denotes the number of brackets with $\|\cdot\|_P$-norm smaller than ε needed to cover $\mathcal{F}$. A bracket is a pair of functions l and u such that $l(\mathbf{x}) \leq u(\mathbf{x})$ for all $\mathbf{x} \in [0,1]^d$, and the norm of the bracket is $\|u - l\|_P$. That a collection $\mathcal{B}$ of brackets covers $\mathcal{F}$ means that for every $f \in \mathcal{F}$ there exists a bracket $(l, u) \in \mathcal{B}$ such that that $l(\mathbf{x}) \leq f(\mathbf{x}) \leq u(\mathbf{x})$ for all $\mathbf{x} \in [0,1]^d$.

For the sole purpose of illustration, consider a setting where $\mathcal{F}$ consists of univariate functions that take values in $[-1, 1]$ and are piece-wise constant on the equally spaced grid with mesh size $1/K$ for some $K < \infty$. Brackets of size ε for this function class can be constructed by chopping the co-domain $[-1, 1]$ into a grid with mesh size ε and then constructing piece-wise constant functions that take values only on this grid. We illustrate this construction in Figure 6.1. If we impose no restrictions on the function class $\mathcal{F}$, we need all possible brackets to cover $\mathcal{F}$. This means that the bracketing number is ε^{-K}. By imposing 'local smoothness' conditions on elements of $\mathcal{F}$, we can limit the number of brackets needed to cover $\mathcal{F}$; for instance, we can limit $\mathcal{F}$ to consist of functions that make jumps of restricted size. Figure 6.1 (b) gives an example of the type of brackets we can restrict attention to under this assumption. We can also control the bracketing number by imposing a 'global smoothness' condition by, for instance, restricting the total sum of all jumps. Figure 6.1 (c) gives an example of the type of brackets we need to consider under this assumption.

The original proof demonstrating that, for a suitable loss function, the empirical risk minimizer over $\mathcal{D}_M^d$ converge to $\Psi(P)$ at some rate $r_n = o(n^{-1/4})$, only relied on the fact that $\mathcal{D}_M^d$ is a Donsker class [van der Laan, 2017b]. Later, Bibaut and van der Laan [2019] used that any element $f \in \mathcal{D}_M^d$ can be represented as a linear combination of multivariate cumulative distribution functions [Bibaut and van der Laan, 2019, Proposition 1; Manuscript III, Propositions 2.3 and 2.4] to derive a tighter bound on the rate of convergence by using a bound on the bracketing entropy for the class of multivariate cumulative distribution functions given by Gao [2013].

As we discuss in Section 6.2, for some loss functions the empirical risk minimizer over the whole space $\mathcal{D}_M^d$ is not well-defined [Manuscript III, Proposition 5.1]. Given a data set $\{X_i\}_{i=1}^n \subset [0,1]^d$, we can alleviate this by restricting the minimization problem to the data-

[†]The modulus of continuity is not exclusively defined for the $\|\cdot\|_P$, but can be defined for any distance measure. We are only interested in the special case where the distance is $\|\cdot\|_P$, and so we use this distance in the definition for notational convenience.

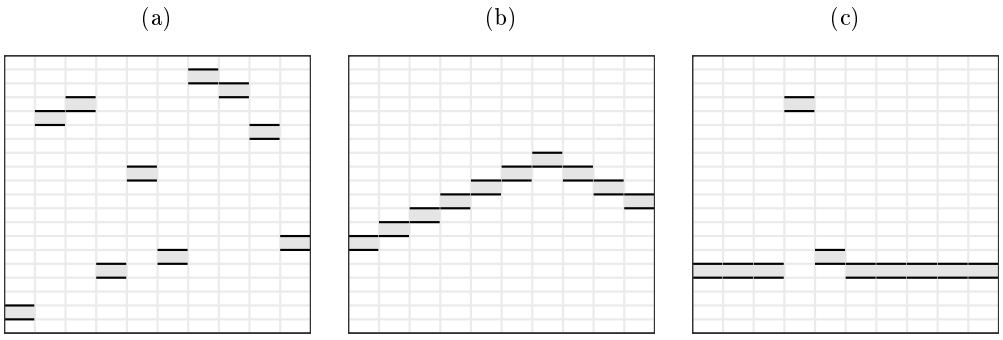


Figure 6.1: Examples of brackets for a class $\mathcal{F}$ of piece-wise constant functions $f\colon [0,1] \to [-1,1]$. The upper and lower boundaries of the gray area correspond to a bracket (l,u). Panels (a)-(c) provide examples of brackets. Brackets like the one in panel (b) can be used to cover $\mathcal{F}$ when the functions in $\mathcal{F}$ can be assumed to not fluctuate much locally. Brackets like the one in panel (c) can be used when the functions can be assumed to not fluctuate much globally.

dependent function class

$$\mathcal{F}_n = \left\{ f_{\beta,n}\colon [0,1]^d \to \mathbb{R} \;\middle|\; f_{\beta,n}(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \sum_{s \in \mathcal{S}} \beta_i^s \mathbb{1}\{X_{r,i} \leq x_r, r \in s\} \right\},$$

where we use the notation $X_i = (X_{1,i}, \ldots, X_{d,i})$, and $\mathcal{S}$ denotes all non-empty subsets of $\{1, \ldots, d\}$. We refer to the random points

$$X_i^s = (\mathbb{1}\{1 \in s\} X_{1,i}, \ldots, \mathbb{1}\{d \in s\} X_{d,i}) \in [0,1]^d, \quad \text{for all} \quad s \in \mathcal{S}, i \in \{1, \ldots, n\}, \tag{6.2}$$

as knot-points. In words, any $f_{\beta,n}$ is a linear combination of basis functions, where the set of basis functions consists of all indicator functions of boxes spanned by a knot-point and $\mathbf{1}$. In Manuscript III we define the HAL estimator as the empirical risk minimizer over $\mathcal{F}_n$. From this perspective, the HAL estimator is a sieve estimator where the sieve $\mathcal{F}_n$ is determined completely by the observed data.

To derive convergence rates for the HAL estimator defined in this way, we construct an auxiliary oracle function $f_n^* \in \mathcal{F}_n$, whose coefficients $\{\beta_i^s\}$ depend on the function $\Psi(P)$ to be estimated. We then exploit that the knot-points $\{X_i^s\}$ are distributed as random samples from P and from the marginals of P, which means that the difference between f_n^* and $\Psi(P)$ can be analyzed as a sum of empirical processes. The difference between the oracle function f_n^* and the HAL estimator is then analyzed separately using the bracketing number in [Bibaut and van der Laan, 2019]. The construction of the auxiliary oracle function depends on an additional smoothness assumption, which we believe is a necessary condition. Whether this additional smoothness assumption is necessary essentially boils down to whether, for any fixed $\mathbf{z} \in [0,1]^d$, the function $\mathbb{1}_{[\mathbf{z},\mathbf{1}]}$ can be approximated sufficiently fast in $\mathcal{L}_P^2$-norm using functions from $\mathcal{F}_n$, when P is dominated by Lebesgue measure.

6.2 A global smoothness condition

In Section 6.1 we discussed how the complexity of a function class can be controlled by imposing local or global smoothness conditions. The HAL estimator works when the function to be estimated has sectional variation norm bounded by a fixed constant. The sectional variation norm measures the total fluctuation of a function and can be seen as a global smoothness condition.

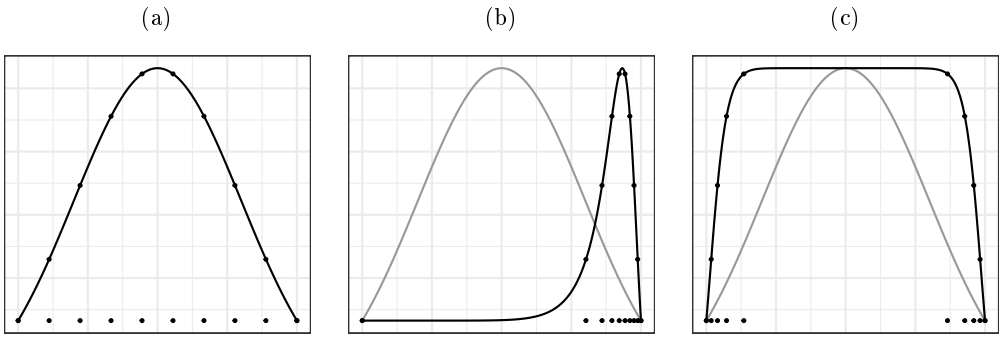


Figure 6.2: Panel (b) and panel (c) are two warped versions of the function in panel (a) obtained by transforming the domain $[0, 1]$ with two different monotone, boundary-preserving functions. The three functions have different $\mathcal{L}^1_P$- and Sobolev-norm, but the same total variation norm. This follows because the total variation norm is the supremum over the total sum of jumps a function makes between a finite set of points. The black dots illustrate that any finite collection of points can be moved so that the total sum of jumps agree for the original and the warped functions.

One can argue that the global smoothness assumption underlying the HAL estimator makes it particularly attractive as an estimator of nuisance parameters in the setting of targeted learning. It is well-known that estimators of nuisance parameters have to be undersmoothed to give good performance for estimation of low-dimensional target parameters if a targeting step is not employed [Laurent et al., 1996, Goldstein and Khasminskii, 1996, Newey et al., 1998, Bickel et al., 2003, Paninski and Yajima, 2008]. Point-wise smoothness is not particularly useful for an estimator of a nuisance parameter that is plugged into a smooth functional. Intuitively, a smooth functional works much like an integral operator that smears out fluctuations. Informally, one could say that the HAL estimator spends the complexity of the function space in the right way, namely by ensuring enough structure such that $n^{-1/4}$-rate estimation is possible, but without imposing irrelevant point-wise smoothness conditions. A theoretical understanding of the performance of the HAL estimator in the context of targeted learning is ongoing research [van der Laan, 2017a, Ertefaie et al., 2020, Qiu et al., 2021, van der Laan et al., 2022, van der Laan, 2023].

In some cases, a global smoothness condition can clash with a loss function that measures performance in terms of an $\mathcal{L}^r_P$-norm. Warping a function will typically change its $\mathcal{L}^1_P$- and Sobolev-norm, but will always leave the sectional variation norm unchanged. We illustrate this in Figure 6.2. This becomes a problem for right-censored survival data observed in continuous time. In this setting, the partial log-likelihood loss for the conditional Lebesgue hazard function h is

$$\int_0^{\tilde{T}} h(s \mid W) \, \mathrm{d}s - \Delta \log h(\tilde{T} \mid W)$$

where the data $(\tilde{T}, \Delta, W)$ are as described in Section 4.2. The presence of the $\mathcal{L}^1_P$-norm in this loss implies that empirical risk minimization over the whole space $\mathcal{D}^d_M$ is an ill-defined optimization problem. As discussed in Section 6.1, we solve this problem in Manuscript III by defining the HAL estimator as a data-adaptive sieve estimator.

Chapter 7

Summary of manuscripts and contributions

We here give summaries of the three manuscripts that make up the main part of this thesis.

Manuscript I *Targeted estimation of state occupation probabilities for the non-Markov illness-death model*

In this manuscript we derive a class of targeted estimators for the state occupation probabilities in the irreversible illness-death model when observations are made in continuous time. We allow all state transitions to depend non-parametrically on a vector of baseline covariates as well on the history of the process. We make the minimal assumptions of coarsening at random and positivity to ensure identifiability of the state occupation probabilities. This allows for complex and realistic censoring dependencies. Our proposed class of estimators allow data-adaptive estimation of the state transition probabilities, as long as the conditional hazard functions (with respect to Lebesgue measure) are estimated. We derive a computationally efficient algorithm for calculating a general class of integrals constructed from piece-wise constant functions. We expect that this can be of independent use for other types of targeted estimators based on right-censored data. While general targeted estimation methods exist for longitudinal data [van der Laan and Rose, 2011, Rytgaard et al., 2022], our proposed class of estimators are, to the best of our knowledge, the first concrete example of a data-adaptive method for estimation of state occupation probabilities.

We also pay some attention to the special case where no baseline variables are measured. It is then natural to use the Nelson-Aalen estimator to estimate some of the cumulative transition hazard functions. Our general result does not apply in this setting, but we use empirical process theory to establish a similar result via a different route. As far a we are aware, this provides the first method that formally handles remainder terms for targeted learning based on right-censored data, when estimators such as the Nelson-Aalen estimator are used.

Manuscript II *The state learner – a super learner for right-censored data*

In this manuscript we propose a new super learner that can be used with right-censored data and in the presence of competing events. We refer to this super learner as the state learner. We establish theoretical guarantees in the form of a consistency result and a finite sample oracle inequality. We discuss the advantages of the state learner compared to existing super learners for right-censored data, and conduct numerical studies to investigate the performance of the state learner. Our proposal is a theoretically justified super learner that can include any type of learner in its library, and does not rely on a pre-specified estimator

of the censoring distribution. Such a super learner did not exist before, so the state learner is an important contribution to the development of robust data-adaptive estimation methods for right-censored data.

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

This manuscript provides a detailed characterization of the space of multivariate càdlàg functions, the sectional variation norm, and the highly-adaptive lasso (HAL) estimator. We provide the first formal study of the HAL estimator in the context of conditional hazard and density estimation. We demonstrate that the HAL estimator should be defined as a data-adaptive sieve estimator, because empirical risk minimization over the space of càdlàg functions with bounded sectional variation norm is not in general well-defined. We derive a general result for the convergence rate of HAL estimators under some smoothness assumptions on the loss function and the data-generating distribution. As an example we demonstrate that this provides a formal justification for using the HAL estimator to estimate conditional hazard functions from right-censored data. We also demonstrate the usefulness of our general result for conditional density estimation and non-parametric least squares regression.

Chapter 8

Perspectives and topics for further research

Each of the three manuscripts included in this thesis attempted to solve a problem. In the process, new problems appeared. We highlight some of the limitation of our proposals in this chapter, and point to some interesting questions for future research which have not been addressed in the thesis.

Finite sample inference The semi-parametric efficiency theory outlined in Chapter 3 provides an elegant way of characterizing the asymptotic distribution of RAL estimators. However, as we tried to highlight in Section 3.3, some aspects are lost from an asymptotic viewpoint. Indeed, we might imagine that while many non-parametric models have the same information bound, some non-parametric models can lead to estimators with dramatically improved finite sample performance. Similarly, in a non-parametric model, all RAL estimators are asymptotically equivalent, but some specific RAL estimators might outperform others in finite samples. These concerns have not been addressed in this thesis. In Manuscript I we examined the finite sample performance of our proposed estimator through a small numerical study. Larger simulation studies in more general settings should be conducted to investigate finite sample performance systematically.

Except for a few recent results [van der Laan, 2017a, Chernozhukov et al., 2023, Singh, 2021], finite sample inference for targeted estimators has not been studied theoretically. The expansion in equation (3.3) in Section 3.2 suggests that finite sample performance of the one-step estimator could be investigated by studying the empirical process term $\mathbb{G}_n[\varphi_{\hat{P}_n} - \varphi_P]$ and the remainder term $(*)$. The empirical process term can be studied by using, for instance, the bracketing entropy we discussed in Section 6.1. The remainder term might be studied by considering higher-order von Mises expansions [von Mises, 1947, Serfling, 1980, Robins et al., 2008, van der Laan et al., 2021]. We hope to investigate whether this is a viable strategy for establishing finite sample inference in future work.

In Manuscript III we derived asymptotic convergence rates for the HAL estimator. We did not consider whether these rates could be improved to finite sample bounds. In the context of least squares regression with fixed design, Fang et al. [2021] provide finite sample bounds for the empirical risk minimizer over the class of càdlàg functions with bounded sectional variation norm. It would be interesting to see if their method could be extended to provide finite sample bounds for the HAL estimator in the context of density and hazard function estimation.

Bounding integral operator differences In Section 4.3 we argued that the remainder term $(*)$ defined in equation (3.3) is in general more difficult to handle when data is right-

censored and observed in continuous time. This is due to the difficulty of bounding terms that involve an integral operator difference as in equation (4.4). The technique we use in Appendix C.1 of Manuscript I is essentially based on expressing the difference between a cumulative hazard function and its Nelson-Aalen estimator as an empirical process term. We expect that the same strategy can be applied when cumulative hazard functions are estimated with Cox models.

In Appendix C of Manuscript II we show that we can obtain an estimator of the conditional survival function S given an estimator of conditional state occupation probability function F defined in equation (5.3). However, convergence rates for an estimator $\hat{F}_n$ do not directly translate into the same convergence rate for the derived estimator $\hat{S}_n$. This is an important drawback of the state learner. The problem here is essentially the same problem of bounding terms that involve an integral operator difference. Devising methods that allows us to bound such terms in decent generality is an interesting topic for future research. Overgaard et al. [2017] used results for p-variation norms [Dudley et al., 2011] to establish a general theory for estimation equations based on pseudo-observations. These norms could perhaps prove useful in bounding remainder terms appearing in targeted learning.

Computationally feasible HAL estimator Our investigation in Manuscript III of the HAL estimator did not consider any aspects of practical implementation. With the current implementations of the HAL estimator it is difficult to investigate its performance in numerical studies that reflect realistic data-generating mechanisms. Schuler et al. [2022] provide a boosting framework to approximate the HAL estimator. Designing methods for approximation the HAL estimator with computationally feasible algorithms is important if the estimator is to be used in practice.

CAR and local independence In Chapter 4 we discussed identifiability from the high-level perspective of coarsened data. We expect that CAR can be expressed in a more concise manner by focusing on multi-state models. In particular, the concept of local independence provides an anti-symmetric way of talking about dependencies between stochastic process [Schweder, 1970, Aalen, 1987, Didelez, 2008, Mogensen et al., 2020]. This might provide a useful tool for expressing and visualizing various identifiability assumptions such as CAR. Røysland et al. [2022] provide some results in this direction.

Appendix A

Some technical arguments

We collect here some technical arguments that did not fit into the main text. Section A.1 contains a proof of Proposition 3.1. Section A.2 defines a fluctuating function and provides an example where the supremum norm is not sufficient to bound terms on the form in equation (4.4) that we discussed in Section 4.3. Section A.3 considers inverse probability of censoring weighted loss functions when attention is restricted to a bounded interval. Section A.4 discusses the iterative algorithm proposed by Westling et al. [2021] and Han et al. [2021] that we mentioned in Section 5.2.

A.1 Saturated tangent spaces

Let in this subsection $\mathcal{O} = [0,1]^d$ and $\mathcal{P}$ be the collection of all probability measures on $[0,1]^d$ with a Lebesgue density. For $P \in \mathcal{P}$ we write p for the density of P with respect to Lebesgue measure. Recall that $\mathcal{H}_P$ is the collection of all functions $h \in \mathcal{L}_P^2$ with $P[h] = 0$. For any $P \in \mathcal{P}$ and $h \in \mathcal{H}_P$ with $\|h\|_\infty < \infty$, define

$$P_t^h = \exp\left\{th - \log\left(P[e^{th}]\right)\right\} \cdot P. \tag{A.1}$$

Lemma A.1. *(i) For any $h \in \mathcal{H}_P$ with $\|h\|_\infty < \infty$, the path $\{P_t^h : t \in \mathbb{R}\}$, with P_t^h defined in equation (A.1), is contained in $\mathcal{P}$ and has score function h at P.*

(ii) The space $\{h \in \mathcal{H}_P : \|h\|_\infty < \infty\}$ is dense in $\mathcal{H}_P$.

(iii) Let C_P^∞ denote the space of infinitely continuously differentiable functions on $[0,1]^d$ with zero P-integral. For any P such that $\|p\|_\infty < \infty$, C_P^∞ is dense in $\mathcal{H}_P$.

Proof. Statements (i) and (ii) are restatements of Example 3.2.1 in [Bickel et al., 1993]. Statement (iii) follows from, e.g., Theorem 2.15 in [Grubb, 2008]. □

Proof of Proposition 3.1. We consider the four models in turn.

$\mathcal{P}^\varepsilon$: When h is uniformly bounded,

$$\left\|\exp\left\{th - \log\left(P[e^{th}]\right)\right\}\right\|_\infty \longrightarrow 1, \quad \text{for} \quad t \longrightarrow 0, \tag{A.2}$$

so for small enough t, $P_t^h \in \mathcal{P}^\varepsilon$ for any $P \in \mathcal{P}^\varepsilon$. Hence, by Lemma A.1 (i) and (ii), $\dot{\mathcal{P}}_P^\varepsilon = \mathcal{H}_P$ for any $P \in \mathcal{P}^\varepsilon$.

$\mathcal{S}^k$: By definition, $p_t^h \in C^k$ when $h \in C_P^\infty$ and $P \in \mathcal{S}^k$, and as any $h \in C_P^\infty$ is uniformly bounded it follows from equation (A.2) that $P_t^h \in \mathcal{S}^k$ for small enough t. Hence, by Lemma A.1 (i) and (iii), $\dot{\mathcal{S}}_P^k = \mathcal{H}_P$ at any $P \in \mathcal{S}^k$.

$\mathcal{V}^M$: When $h \in C_P^\infty$ and $P \in \mathcal{V}^M$, p_t^h is càdlàg. By definition of the sectional variation norm (see for instance Section 2 of Manuscript III), $\|hf\|_v \leq \|h\|_\infty \|f\|_v$, and so it follows from equation (A.2) that $P_t^h \in \mathcal{V}^M$ for t small enough. As this holds for any $h \in C_P^\infty$, $\dot{\mathcal{V}}_P^M = \mathcal{H}_P$ by Lemma A.1 (i) and (iii).

$\mathcal{M}$: For a bounded non-decreasing function $f \in \mathcal{H}_P$, the density of P_t^f must also be non-decreasing for $P \in \mathcal{M}$. Hence $\{P_t^f\} \subset \mathcal{M}$ and so $f \in \dot{\mathcal{M}}_P$ by Lemma A.1 (i). Take now some $h \in C_P^\infty$. It follows from Theorem 2 in [Aistleitner and Dick, 2014] that we can write $h = h_+ - h_-$ for some non-decreasing, bounded functions $h_+, h_- \in \mathcal{H}_P$. As both h_+ and h_- are elements of the tangent space by the argument above, and the tangent space is the closed linear span of all score functions, it follows that h is in the tangent space. As $h \in C_P^\infty$ was arbitrary it follows from Lemma A.1 (iii) that $\dot{\mathcal{M}}_P = \mathcal{H}_P$. □

A.2 A wildly fluctuating function

For all $n \in \mathbb{N}$, define the function $f_n \colon [0,1] \to \mathbb{R}$ as

$$f_n(x) = \sum_{i=1}^{n} \frac{(-1)^i}{n^{1/4}} \mathbb{1}_{\left[\frac{i-1}{n}, \frac{i}{n}\right)}(x).$$

Note that for any $n \in \mathbb{N}$, f_n is càdlàg and has bounded variation norm, so the (Lebesgue-Stieltjes) integral $\int_0^1 g \, \mathrm{d}f_n$ is well-defined for any Borel-measurable and bounded function g.

Proposition A.2. *It holds that* $\|f_n\|_\infty \to 0$ *and* $\int_0^1 f_n \, \mathrm{d}f_n \to \infty$ *when* $n \to \infty$.

Proof. The first statement follows because

$$\sup_{x \in [0,1]} \left| \frac{(-1)^i}{n^{1/4}} \mathbb{1}_{\left[\frac{i-1}{n}, \frac{i}{n}\right)}(x) \right| = n^{-1/4},$$

for all $i = 1, \ldots, n$. The second statement follows because

$$\int_0^1 f_n \, \mathrm{d}f_n = \sum_{i=1}^{n} \frac{(-1)^i}{n^{1/4}} f_n\left(\frac{i-1}{n}\right) = \sum_{i=1}^{n} \frac{(-1)^i}{n^{1/4}} \frac{(-1)^i}{n^{1/4}} = \sum_{i=1}^{n} \frac{1}{n^{1/2}} = n^{1/2}.$$ □

A.3 Inverse probability of censoring weighted loss functions on bounded intervals

Let Q denote the distribution of $Z = (T, W) \in \mathcal{Z} = \mathbb{R}_+ \times \mathcal{W}$, for some $\mathcal{W} \subset \mathbb{R}^d$, C a censoring time such that $C \perp\!\!\!\perp T \mid W$, and $G(\cdot \mid w)$ the conditional survival function for C given $W = w$. Let $P_{Q,G}^\tau$ denote the distribution of $O_\tau = (\tilde{T}, \Delta_\tau, W)$, where

$$\Delta_\tau = \Delta \mathbb{1}\{\tilde{T} \leq \tau\} + \mathbb{1}\{\tilde{T} > \tau\}.$$

Let Θ denote the parameter space for a parameter θ defined on $\mathcal{Q}$. Let $L^F \colon \Theta \times \mathcal{Z} \to \mathbb{R}$ be a loss function such that

$$L^F(\theta, (s, w)) = L^F(\theta, (\tau+, w)) \quad \text{for all } s > \tau, \theta \in \Theta, w \in \mathcal{W}. \tag{A.3}$$

Define the inverse probability of censoring weights $\omega_G^\tau(t,w) = G(t - \wedge\tau \mid w)^{-1}$, where $G(t - \wedge\tau \mid w)$ can be explicitly written as

$$\lim_{h\downarrow 0} G((t-h)\wedge\tau \mid w) = \begin{cases} G(\tau) & \text{if } t > \tau \\ G(t-) & \text{if } t \leq \tau \end{cases}. \tag{A.4}$$

Define the inverse probability of censoring weighted loss function

$$L_\tau(\theta, O_\tau; G) = L^F(\theta, (\tilde{T}, W))\Delta_\tau \omega_G^\tau(\tilde{T}, W). \tag{A.5}$$

Note that the definition in equation (A.5) is identical to the definition given in equation (5.2) when $\tau = \infty$ and $\tilde{T} < \infty$ a.s.

Proposition A.3. *For some $\tau > 0$, assume that equation* (A.3) *holds and that $G(\tau \mid w) > 0$ for all $w \in \mathcal{W}$. Then $P_{Q,G}^\tau[L_\tau(\theta, \cdot; G)] = Q[L^F(\theta, \cdot)]$.*

Proof. We have

$$\begin{aligned} &P_{Q,G}^\tau[L_\tau(\theta, \cdot; G)] \\ &= \mathbb{E}[L^F(\theta, (\tilde{T}, W))\Delta \mathbb{1}\{\tilde{T} \leq \tau\}\omega_G^\tau(\tilde{T}, W)] \\ &\quad + \mathbb{E}[L^F(\theta, (\tilde{T}, W))\mathbb{1}\{\tilde{T} > \tau\}\omega_G^\tau(\tilde{T}, W)]. \end{aligned} \tag{A.6}$$

Note that equations (A.4) and (A.3) imply

$$\begin{aligned} &\mathbb{E}[L^F(\theta, (\tilde{T}, W))\mathbb{1}\{\tilde{T} > \tau\}\omega_G^\tau(\tilde{T}, W)] \\ &= \mathbb{E}[L^F(\theta, (\tilde{T}, W))\mathbb{1}\{\tilde{T} > \tau\}G(\tau, W)^{-1}] \\ &= \mathbb{E}[L^F(\theta, (\tau+, W))\mathbb{1}\{\tilde{T} > \tau\}G(\tau, W)^{-1}] \\ &= \int_{\mathcal{W}} L^F(\theta, (\tau+, w))G(\tau, w)^{-1} S_Q(\tau \mid w)G(\tau \mid w)H_Q(\mathrm{d}w) \\ &= \int_{\mathcal{W}} L^F(\theta, (\tau+, w))S_Q(\tau \mid w)H_Q(\mathrm{d}w), \end{aligned} \tag{A.7}$$

where we let H_Q denote the marginal distribution of W. Letting $\bar{S}_Q = 1 - S_Q$ we have by equation (A.4),

$$\begin{aligned} &\mathbb{E}[L^F(\theta, (\tilde{T}, W))\Delta \mathbb{1}\{\tilde{T} \leq \tau\}\omega_G^\tau(\tilde{T}, W)] \\ &\mathbb{E}[L^F(\theta, (\tilde{T}, W))\Delta \mathbb{1}\{\tilde{T} \leq \tau\}G^{-1}(\tilde{T}-, W)] \\ &= \int_{\mathcal{W}} \int \mathbb{1}_{[0,\tau]}(s) L^F(\theta, (s, w))G^{-1}(s-, w)G(s- \mid w)\bar{S}_Q(\mathrm{d}s \mid w)H_Q(\mathrm{d}w) \\ &= \int_{\mathcal{W}} \int \mathbb{1}_{[0,\tau]}(s) L^F(\theta, (s, w))\bar{S}_Q(\mathrm{d}s \mid w)H_Q(\mathrm{d}w) \end{aligned} \tag{A.8}$$

Hence equation (A.6)-(A.8) imply

$$\begin{aligned}
P^{\tau}_{Q,G}[L_\tau(\theta,\cdot;G)] &= \int_{\mathcal{W}} \int \mathbb{1}_{[0,\tau]}(s) L^F(\theta,(s,w)) \bar{S}_Q(\mathrm{d}s \mid w) H_Q(\mathrm{d}w) \\
&\quad + \int_{\mathcal{W}} L^F(\theta,(\tau+,w)) S_Q(\tau \mid w) H_Q(\mathrm{d}w) \\
&= \int_{\mathcal{W}} \int \mathbb{1}_{[0,\tau]}(s) L^F(\theta,(s,w)) \bar{S}_Q(\mathrm{d}s \mid w) H_Q(\mathrm{d}w) \\
&\quad + \int_{\mathcal{W}} L^F(\theta,(\tau+,w)) \int \mathbb{1}_{(\tau,\infty)}(s) \bar{S}_Q(\mathrm{d}s \mid w) H_Q(\mathrm{d}w) \\
&= \int_{\mathcal{W}} \int \mathbb{1}_{[0,\tau]}(s) L^F(\theta,(s,w)) \bar{S}_Q(\mathrm{d}s \mid w) H_Q(\mathrm{d}w) \\
&\quad + \int_{\mathcal{W}} \int \mathbb{1}_{(\tau,\infty)}(s) L^F(\theta,(s,w)) \bar{S}_Q(\mathrm{d}s \mid w) H_Q(\mathrm{d}w) \\
&= Q[L^F(\theta,\cdot)],
\end{aligned}$$

where we used equation (A.3) for the second to last equality. □

A.4 Iterative inverse probability of censoring weighted super learner

We consider the algorithm proposed by Han et al. [2021] and Westling et al. [2021] described at the end of Section 5.2. We assume in the following that $\mathcal{A} = \mathcal{B} = \mathcal{S}$, with $\mathcal{S}$ being the collection of all conditional survival functions. We also assume that performance is assessed using an infinite hold-out data set. Given an initial value $G^{(0)}$ we iteratively define

$$\begin{aligned}
S^{(k)} &= \operatorname*{argmin}_{S^* \in \mathcal{S}} P_{Q,G}[L(S^*,\cdot;G^{(k-1)})], \quad \text{and} \\
G^{(k)} &= \operatorname*{argmin}_{G^* \in \mathcal{G}} P_{Q,G}[L_c(G^*,\cdot;S^{(k)})],
\end{aligned} \tag{A.9}$$

for $n \in \mathbb{N}$. Consider now some initial value $G^{(0)}$ such that $G^{(0)} \neq G$. We may then write

$$\begin{aligned}
&P_{Q,G}[L(S^*,\cdot;G^{(0)})] \\
&= \mathbb{E}_{Q,G}[L^F(S^*,(\tilde{T},W))\Delta\omega_{G^{(0)}}(\tilde{T},W)] \\
&= \mathbb{E}_{Q,G}\left[L^F(S^*,(\tilde{T},W))\Delta\frac{\omega_{G^{(0)}}(\tilde{T},W)}{\omega_G(\tilde{T},W)}\omega_G(\tilde{T},W)\right] \\
&= H_Q\left[\int_0^\infty L^F(S^*,(u,\cdot))\frac{\omega_{G^{(0)}}(u,\cdot)}{\omega_G(u,\cdot)}\omega_G(u,\cdot)G(u- \mid \cdot)\bar{S}_Q(\mathrm{d}u \mid \cdot)\right] \\
&= H_Q\left[\int_0^\infty L^F(S^*,(u,\cdot))\frac{\omega_{G^{(0)}}(u,\cdot)}{\omega_G(u,\cdot)}\bar{S}_Q(\mathrm{d}u \mid \cdot)\right] \\
&= Q\left[L^F(S^*,\cdot)\frac{\omega_{G^{(0)}}}{\omega_G}\right],
\end{aligned}$$

where $\bar{S}_Q = 1 - S_Q$. If $Q[\omega_{G^{(0)}}/\omega_G]$ is finite, we can define

$$\check{Q} = q^{(1)} \cdot Q \quad \text{with} \quad q^{(1)} = \frac{\omega_{G^{(0)}}}{\omega_G}\left(Q\left[\frac{\omega_{G^{(0)}}}{\omega_G}\right]\right)^{-1}. \tag{A.10}$$

Then the measure $\check{Q}$ is a probability measure and we have

$$P_{Q,G}[L(S^*,\cdot;G^{(0)})] \propto \check{Q}[L^F(S^*,\cdot)]. \tag{A.11}$$

Now, when the loss function L^F is a strictly proper scoring rule [Gneiting and Raftery, 2007], $S^* \mapsto \check{Q}[L^F(S^*, \cdot)]$ is uniquely minimized at $S_{\check{Q}}$. Thus $S^* \mapsto P_{Q,G}[L(S^*, \cdot; G^{(0)})]$ is also minimized at $S_{\check{Q}}$ and so $S^{(1)} = S_{\check{Q}}$. As we assumed that $G^{(0)} \neq G$, we have $\omega_{G^{(0)}} \neq \omega_G$ and so $S^{(1)} \neq S_Q$. Now starting from $S^{(1)}$ we can obtain a learner $G^{(1)}$ in a similar way by minimizing $G^* \mapsto P_{Q,G}[L_c(G^*, \cdot, S^{(1)})]$. By the same line of arguments, $G^{(1)} \neq G$. In this way, we obtain a sequence of learners $\{G^{(k)}\}_{k=0}^{\infty}$ and $\{S^{(k)}\}_{k=1}^{\infty}$ such that $G^{(k)} \neq G$ and $S^{(k)} \neq S_Q$ for all k.

Example A.4 describes a simple situation where an infinite number of pairs (G^*, S^*) are fixed points for the algorithm defined in equation (A.9).

Example A.4
For $\alpha > 0$, let $\mathrm{Exp}(\alpha)$ be the survival function of the exponential distribution with rate parameter α. Let $Q_\alpha = S_\alpha = \mathrm{Exp}(\alpha)$ and $G_\beta = \mathrm{Exp}(\beta)$, and define $P_{\alpha,\beta}$ as the distribution of $(\tilde{T}, \Delta)$ with $\tilde{T} = T \wedge C$ and $\Delta = \mathbb{1}\{T \leq C\}$ for $(T, C) \sim Q_\alpha \otimes G_\beta$. Fix $\alpha, \beta \in \mathbb{R}_+$. For any $\beta^* \in \mathbb{R}_+$,

$$\frac{\omega_{G_{\beta^*}}(t)}{\omega_{G_\beta}(t)} = \frac{G_\beta(t-)}{G_{\beta^*}(t-)} = e^{-(\beta-\beta^*)t},$$

and so

$$Q_\alpha\left[\frac{\omega_{G_{\beta^*}}}{\omega_{G_\beta}}\right] = \int_0^\infty \alpha e^{-(\beta+\alpha-\beta^*)t}\,\mathrm{d}s.$$

For any $\beta^* < \beta + \alpha$ the integral above is finite, so we may define $\check{Q}$ as in equation (A.10), and we see that $\check{Q} = Q_{\beta+\alpha-\beta^*}$. Hence $S^{(1)} = S_{\beta+\alpha-\beta^*}$ by equation (A.11) for any strictly proper scoring rule L^F. Similarly, we have

$$\frac{\omega_{S_{\beta+\alpha-\beta^*}}(t)}{\omega_{S_\alpha}(t)} = \frac{S_\alpha(t-)}{S_{\beta+\alpha-\beta^*}(t-)} = e^{-(\alpha-(\beta+\alpha-\beta^*))t} = e^{(\beta-\beta^*)t},$$

and so

$$G_\beta\left[\frac{\omega_{Q_{\beta+\alpha-\beta^*}}}{\omega_{Q_0}}\right] = \int_0^\infty \beta e^{-\beta^* t}\,\mathrm{d}s,$$

which implies $G^{(1)} = G_{\beta^*}$. This shows that for any $\beta^* < \beta + \alpha$, the pair $(G_{\beta^*}, S_{\beta+\alpha-\beta^*})$ is a fixed point for the iterative algorithm defined in equation (A.9). •

Bibliography

O. O. Aalen. Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987(3-4): 177–190, 1987.

C. Aistleitner and J. Dick. Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality. *arXiv preprint arXiv:1406.0230*, 2014.

P. K. Andersen, J. P. Klein, and S. Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

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

D. W. Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72, 1994.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

J. M. Begun, W. J. Hall, W.-M. Huang, and J. A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, 11(2):432–452, 1983.

A. Bender, D. Rügamer, F. Scheipl, and B. Bischl. A general machine learning framework for survival analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 158–173. Springer, 2020.

D. Benkeser and M. J. van der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016.

D. Benkeser, M. Carone, M. J. van der Laan, and P. B. Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

R. Beran. Robust location estimates. *The Annals of Statistics*, pages 431–444, 1977.

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 derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

P. J. Bickel, Y. Ritov, et al. Nonparametric estimators which can be "plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003.

L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

L. Birgé and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, 23(1):11–29, 1995.

P. Blanche, M. W. Kattan, and T. A. Gerds. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics*, 20(2):347–357, 2019.

L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

J. Brown, M. Hollander, and R. M. Korwar. Nonparametric tests of independence for censored data, with applications to heart transplant studies. *Reliability and Biometry:*

Statistical Analysis of Lifelength, F. Proschan and RJ Serfling, eds. Philadelphia, pages 327–354, 1974.

M. Carone, A. R. Luedtke, and M. J. van der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, 2018.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018a. doi: 10.1111/ectj.12097. URL `https://doi.org/10.1111/ectj.12097`.

V. Chernozhukov, W. K. Newey, and R. Singh. Learning l2-continuous regression functionals via regularized Riesz representers. *arXiv preprint arXiv:1809.05224*, 8, 2018b.

V. Chernozhukov, W. K. Newey, and R. Singh. A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika*, 110(1):257–264, 2023.

K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

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

V. Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

R. M. Dudley, R. Norvaiša, and R. Norvaiša. *Concrete functional calculus*. Springer, 2011.

A. Ertefaie, N. S. Hejazi, and M. J. van der Laan. Nonparametric inverse probability weighted estimators based on the highly adaptive lasso. *arXiv preprint arXiv:2005.11303*, 2020.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Routledge, 1996.

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.

J. Fine. Comparing nonnested Cox models. *Biometrika*, 89(3):635–648, 2002.

A. Fisher and E. H. Kennedy. Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2):162–172, 2021.

E. Fix and J. Neyman. A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, 23(3):205–241, 1951.

F. Gao. Bracketing entropy of high dimensional distributions. In *High Dimensional Probability VI: The Banff Volume*, pages 3–17. Springer, 2013.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

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.

T. Gerds. *Nonparametric efficient estimation of prediction error for incomplete data models*. PhD thesis, Freiburg (Breisgau), Univ., Diss., 2003, 2002.

T. A. Gerds. *prodlim: Product-Limit Estimation for Censored Event History Analysis*, 2019. URL `https://CRAN.R-project.org/package=prodlim`. R package version 2019.11.13.

T. A. Gerds and M. Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–2184, 2013.

R. D. Gill, J. A. Wellner, and J. Præstgaard. Non-and semi-parametric maximum likelihood estimators and the von Mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics*, pages 97–128, 1989.

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.

R. D. Gill, M. J. Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

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.

M. K. Golmakani and E. C. Polley. Super learner for survival data prediction. *The International Journal of Biostatistics*, 16(2):20190065, 2020.

P. Gonzalez Ginestet, A. Kotalik, D. M. Vock, J. Wolfson, and E. E. Gabriel. Stacked inverse probability of censoring weighted bagging: A case study in the infcarehiv register. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(1):51–65, 2021.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

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

G. Grubb. *Distributions and operators*, volume 252. Springer Science & Business Media, 2008.

X. Han, M. Goldstein, A. Puli, T. Wies, A. Perotte, and R. Ranganath. Inverse-weighted survival games. *Advances in Neural Information Processing Systems*, 34, 2021.

F. Harrell, K. Lee, and D. Mark. Tutorial in biostatistics: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 15: 361–387, 1996.

F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

D. F. Heitjan. Ignorability and coarse data: Some biomedical examples. *Biometrics*, pages 1099–1109, 1993.

D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *The annals of statistics*, pages 2244–2253, 1991.

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.

M. Hernán and J. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

N. L. Hjort. On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique*, pages 355–387, 1992.

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

P. J. Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA: University of California Press, 1967.

I. A. Ibragimov and R. Z. Has'Minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 1981.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

H. Ishwaran and U. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2023. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 3.2.2.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *The Annals of Statistics*, 23(3):774–786, 1995.

J. Kalbfleisch and R. MacKay. On constant-sum models for censored survival data. *Biometrika*, 66(1):87–90, 1979.

J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. M. Robins. Influence functions for machine learning: Nonparametric estimators for entropies, divergences and mutual informations. *arXiv preprint arXiv:1411.4342*, 2014.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

S. Keles, M. J. van der Laan, and S. Dudoit. Asymptotically optimal model selection method with right censored outcomes. *Bernoulli*, 10(6):1011–1037, 2004.

E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Y. A. Koshevnik and B. Y. Levit. On a non-parametric analogue of the information matrix. *Theory of Probability & Its Applications*, 21(4):738–753, 1977.

M. R. Kosorok. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

S. W. Lagakos and J. Williams. Models for censored survival analysis: A cone class of variable-sum models. *Biometrika*, 65(1):181–189, 1978.

B. Laurent et al. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2):659–681, 1996.

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

B. Y. Levit. Infinite-dimensional informational lower bounds. *Theor. Prob. Appl*, 23:388–394, 1978.

Y. Li, K. S. Xu, and C. K. Reddy. Regularized parametric regression for high-dimensional survival analysis. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 765–773. SIAM, 2016.

K. Liestbl, P. K. Andersen, and U. Andersen. Survival analysis and neural nets. *Statistics in medicine*, 13(12):1189–1200, 1994.

X. Lu and A. A. Tsiatis. Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika*, 95(3):679–694, 2008.

I. Malenica, R. V. Phillips, D. Lazzareschi, J. R. Coyle, R. Pirracchio, and M. J. van der Laan. Multi-task highly adaptive lasso. *arXiv preprint arXiv:2301.12029*, 2023.

S. W. Mogensen, N. R. Hansen, et al. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.

U. B. Mogensen and T. A. Gerds. A random forest approach for competing risks based on pseudo-values. *Statistics in medicine*, 32(18):3102–3114, 2013.

A. M. Molinaro, S. Dudoit, and M. J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1): 154–177, 2004.

W. Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1 (1):27–52, 1969.

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

W. K. Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2): 99–135, 1990.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

W. K. Newey, F. Hsieh, and J. Robins. Undersmoothing and bias corrected functional estimation. *SSRN*, 1998.

S. F. Nielsen. Relative coarsening at random. *Statistica Neerlandica*, 54(1):79–99, 2000.

M. Overgaard and S. N. Hansen. On the assumption of independent right censoring. *Scandinavian Journal of Statistics*, 48(4):1234–1255, 2021.

M. Overgaard, E. T. Parner, and J. Pedersen. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, 45:1988 – 2015, 2017. doi: 10.1214/16-AOS1516. URL `https://doi.org/10.1214/16-AOS1516`.

L. Paninski and M. Yajima. Undersmoothed kernel entropy estimators. *IEEE Transactions on Information Theory*, 54(9):4384–4388, 2008.

M. L. Petersen and M. J. van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418, 2014.

J. Pfanzagl. On the measurability and consistency of minimum contrast estimates. *Metrika*, 14(1):249–272, 1969.

J. Pfanzagl and W. Wefelmeyer. *Contributions to a general asymptotic statistical theory*. Springer, 1982.

E. C. Polley and M. J. van der Laan. Super learning for right-censored data. In M. J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 249–258. Springer, 2011.

T. PROVA Study Group. Prophylaxis of first hemorrhage from esophageal varices by sclerotherapy, propranolol or both in cirrhotic patients: a randomized multicenter trial. *Hepatology*, 14(6):1016–1024, 1991.

H. Qiu, A. Luedtke, and M. Carone. Universal sieve-based strategies for efficient estimation using machine learning tools. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 27(4):2300, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

J. Reeds. On the definition of von Mises functionals. *Ph. D. dissertation, Harvard Univ.*, 1976.

R. D. Reiss. Consistency of minimum contrast estimators in non-standard cases. *Metrika*, 25:129–142, 1978.

Y. Ritov and P. J. Bickel. Achieving information bounds in non and semiparametric models. *The Annals of statistics*, 18(2):925–938, 1990.

J. Robins, L. Li, E. Tchetgen, A. van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pages 335–422. Institute of Mathematical Statistics, 2008.

J. Robins, L. Li, E. Tchetgen, and A. W. van der Vaart. Quadratic semiparametric von Mises calculus. *Metrika*, 69:227–247, 2009.

J. M. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319, 1997.

J. M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS epidemiology: methodological issues*, pages 297–331, 1992.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 2021.

K. Røysland, P. Ryalen, M. Nygård, and V. Didelez. Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses. *arXiv preprint arXiv:2202.02311*, 2022.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

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*, 2021.

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.

M. C. Sachs, A. Discacciati, Å. H. Everhov, O. Olén, and E. E. Gabriel. Ensemble prediction of time-to-event outcomes with competing risks: A case-study of surgical complications in Crohn's disease. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 68(5):1431–1446, 2019.

A. Schuler, Y. Li, and M. J. van der Laan. The selectively adaptive lasso. *arXiv preprint arXiv:2205.10697*, 2022.

T. Schweder. Composable markov processes. *Journal of applied probability*, 7(2):400–410, 1970.

R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

A. Shapiro. On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487, 1990.

X. Shen. On methods of sieves and penalization. *The Annals of Statistics*, 25(6):2555–2591, 1997.

X. Shen and W. H. Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615, 1994.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1, 2011.

R. Singh. A finite sample theorem for longitudinal causal inference with machine learning: Long term, dynamic, and mediated effects. *arXiv preprint arXiv:2112.14249*, 2021.

C. Stein. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, pages 187–196. University of California Press, 1956.

J. A. Steingrimsson, L. Diao, and R. L. Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(525):370–383, 2019.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

C. A. Struthers and J. D. Kalbfleisch. Misspecified proportional hazard models. *Biometrika*, 73(2):363–369, 1986.

E. Sverdrup. Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. *Scandinavian Actuarial Journal*, 1965(3-4):184–211, 1965.

T. M. Therneau. *A Package for Survival Analysis in R*, 2022. URL `https://CRAN.R-project.org/package=survival`. R package version 3.4-0.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

S. van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990.

S. van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *The Annals of Statistics*, 21(1):14–44, 1993.

M. J. van de Laan. Efficient and inefficient estimation in semiparametric models. *CWI Tracts*, 1995.

M. J. van der Laan. Finite sample inference for targeted learning. *arXiv preprint arXiv:1708.09502*, 2017a.

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

M. J. 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. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, 2003.

M. J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

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

M. J. van der Laan and S. Rose. *Targeted learning in data science*. Springer, 2018.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

M. J. van der Laan, Z. Wang, and L. van der Laan. Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290*, 2021.

M. J. van der Laan, D. Benkeser, and W. Cai. Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. *The International Journal of Biostatistics*, 2022.

A. van der Vaart. On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991.

A. van der Vaart. On Robins' formula. *Statistics & Decisions*, 22(3):171–200, 2004.

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

A. van der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192, 2011.

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

A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.

V. Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

P. J. Verweij and H. C. van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–2314, 1993.

R. von Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

M. Wainwright. *High-Dimensional Statistics – A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

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.

T. Westling, A. Luedtke, P. Gilbert, and M. Carone. Inference for treatment-specific survival curves using machine learning. *arXiv preprint arXiv:2106.06602*, 2021.

D. Whitney, A. Shojaie, and M. Carone. Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 34(4):591, 2019.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

S. N. Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

L. Zhao and D. Feng. Deep neural networks for survival analysis using pseudo values. *IEEE journal of biomedical and health informatics*, 24(11):3308–3314, 2020.

Manuscripts

Manuscript I

Title	Targeted estimation of state occupation probabilities for the non-Markov illness-death model
Authors	Anders Munch, Marie Skov Breum, Torben Martinussen, and Thomas A. Gerds
Status	Published in *Scandinavian Journal of Statistics* (2023)

Targeted estimation of state occupation probabilities for the non-Markov illness-death model

Anders Munch, Marie Skov Breum, Torben Martinussen, and Thomas A. Gerds

Abstract

We use semi-parametric efficiency theory to derive a class of estimators for the state occupation probabilities of the continuous-time irreversible illness-death model. We consider both the setting with and without additional baseline information available, where we impose no specific functional form on the intensity functions of the model. We show that any estimator in the class is asymptotically linear under suitable assumptions about the estimators of the intensity functions. In particular, the assumptions are weak enough to allow the use of data-adaptive methods, which is important for making the identifying assumption of coarsening at random plausible in realistic settings.

We suggest a flexible method for estimating the transition intensity functions of the illness-death model based on penalized Poisson regression. We apply this method to estimate the nuisance parameters of an illness-death model in a simulation study and a real world application.

1 Introduction

In epidemiology and medical research, illness-death models are used to analyze the time to an absorbing state (death) when the subjects can change from an initial state (healthy) to an intermediate state (illness) at some intermediate time point [Andersen et al., 2012]. We consider the irreversible illness-death model shown in Figure 1 where transitions from the illness state back to the healthy state are not possible. The transition intensity from illness to death can depend on the onset time of illness, which is a challenge for a non-parametric approach. We consider right-censored observations of the illness-death process under the coarsening at random assumption [Gill et al., 1997, van der Vaart, 2004] where the current rate of censoring among subjects in the illness state can depend on the onset time of illness. We also allow the transition intensities to depend on a set of covariates measured at baseline. We use semi-parametric efficiency theory to derive estimators for the (marginal) state occupation probabilities based on the efficient influence function.

Under the stronger assumption of independent censoring it has been shown that the Aalen-Johansen estimator is consistent for the state occupation probabilities [Datta and Satten, 2001, Glidden, 2002], and non-parametric estimators have also been derived for the transition probabilities [Meira-Machado et al., 2006, Allignol et al., 2014, de Uña-Álvarez and Meira-Machado, 2015, Titman, 2015]. More recently landmark approaches have been discussed [Putter and Spitoni, 2018, Maltzahn et al., 2021]. For the case where censoring among subjects in the illness state is allowed to depend on the onset time of illness, several inverse probability of censoring weighted (IPCW) estimators have been suggested [Datta et al., 2000, Datta and Satten, 2002, Gunnes et al., 2007].

Our approach builds on the tools from semi- and non-parametric efficiency theory [Bickel et al., 1993, van der Vaart, 2000, 1991, van der Laan and Robins, 2003, Tsiatis, 2006]. This allows us to combine data-adaptive methods with asymptotic inference [van der Laan and Rubin, 2006, van der Laan and Rose, 2011, 2018, Chernozhukov et al., 2018]. We focus on the estimation of the probability of being ill and alive at some fixed time horizon after baseline.

However, our general approach can in principle be applied to other path-wise differentiable functionals of the data-generating distribution. We describe a class of estimators motivated by the efficient influence function for the parameter of interest under a fully non-parametric model. The estimators are asymptotically linear and normally distributed as long as the estimators of the transition hazard functions converge to the hazard functions of the data-generating distribution at a rate faster than $n^{-1/4}$. Following van der Laan and Rose [2011] and Chernozhukov et al. [2018] we refer to our proposed estimator as a 'targeted' or 'debiased' estimator. The targeting/debiasing step is essential when using data-adaptive methods, as demonstrated by our empirical study. Furthermore, an important feature of our main result (Theorem 4.2) is that we do not need to know the exact asymptotic distribution of the estimators of the nuisance parameters but only their rate of convergence. This is particularly attractive when using data-adaptive model selection such as cross-validation.

We derive the details of the estimators for the state occupation probabilities also in the setting without baseline covariates. The classical illness-death model without baseline covariates has a long history and has been studied intensively [Fix and Neyman, 1951, Sverdrup, 1965, Andersen et al., 2012, Xu et al., 2010, de Uña-Álvarez and Meira-Machado, 2015, Allignol et al., 2014, Meira-Machado et al., 2006, Datta et al., 2000]. In this setting our estimator simplifies (Lemma 4.3) and can be seen as the result of applying the general estimation strategy referred to as 'targeted learning' [Petersen and van der Laan, 2014, van der Laan and Rose, 2011] to this classical problem. The setting with baseline covariates is important for applications, because covariates can help make the coarsening at random assumption more plausible. As the functional relationship between the baseline covariates and the transition intensities are unknown in real applications, the ability to use flexible data-adaptive estimators is important in this context.

In Section 2 we introduce the setting and our notation, and in Section 3 we define our target parameter and its Gâteaux derivative. Section 4 contains our main result which establishes the asymptotic distribution of a class of estimators under a set of conditions on the estimators of the transition hazard functions of the illness-death model. In Section 5 we suggest a data-adaptive method to estimate these functions, and Section 6 contains an empirical study demonstrating the importance of the targeting step. In Section 7 we apply our method to estimate the effect of sclerotherapy on variceal bleeding and death among cirrhotic patients. Section 8 contains a discussion of our results. More proofs, simulations, and technical derivations are provided in the supplementary material (Appendices A-E).

2 Setting and notation

We consider an illness-death process $\{X(t)\}_{t\geq 0}$ with state space $\{0,1,2\}$ as shown in Figure 1 and assume that all subjects start in state 0, i.e., $X(0)=0$. We define respectively the time at which the subject leaves state 0 and the time at which the subject enters the absorbing state,

$$T_0 = \inf\{t>0 \,:\, X(t)\neq 0\} \quad \text{and} \quad T = \inf\{t>0 \,:\, X(t)=2\}.$$

Let variable η indicate the state of the process X at time T_0, i.e., $\eta = X(T_0) = 1 + \mathbb{1}\{T_0 = T\}$, and introduce the counting processes $N_{0k}(t) = \mathbb{1}\{T_0 \leq t, \eta = k\}$ for $k \in \{1,2\}$ and $N_{12}(t) = \mathbb{1}\{T \leq t, \eta = 1\}$. Further, denote by $W \in \mathcal{W} \subseteq \mathbb{R}^p$ a set of covariates measured at baseline ($t=0$). Note that while the process X can change over time we assume that the covariates W are only measured once at baseline (see Section 8 for a discussion of this assumption).

We assume that the intensity processes of the counting processes with respect to the filtra-

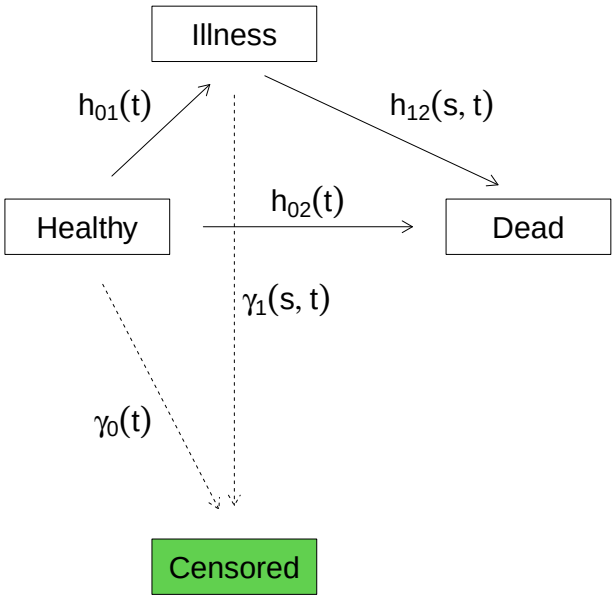


Figure 1: The illness-death model without recovery. At any time the subjects are in one of the three states 'Healthy', 'Illness', 'Dead'. 'Censored' means end of follow-up which can happen when the subject is in state '0' or in state '1'.

tion $\mathcal{F}_t$ generated by $(N_{01}(t), N_{02}(t), N_{12}(t), W)$ can be described by deterministic functions $h_{0k}\colon \mathbb{R}_+ \times \mathcal{W} \to \mathbb{R}_+$ for $k \in \{1,2\}$ and $h_{12}\colon \mathcal{U} \times \mathcal{W} \to \mathbb{R}_+$, where $\mathcal{U} = \{(t,s) \in \mathbb{R}^2 \mid s \leq t\}$:

$$\mathbb{E}(N_{0k}(\mathrm{d}t)|\mathcal{F}_{t-}) = h_{0k}(t,W)Y_0(t)\mathrm{d}t, \quad \text{where} \quad Y_0(t) = \mathbb{1}\{t \leq T_0\},$$
$$\mathbb{E}(N_{12}(\mathrm{d}t)|\mathcal{F}_{t-}) = h_{12}(t,T_0,W)Y_1(t)\mathrm{d}t, \quad \text{where} \quad Y_1(t) = \mathbb{1}\{T_0 < t \leq T\}.$$

We introduce a right-censoring time C at which the observation of the process X stops. Instead of T_0 and T, we only observe $\tilde{T}_0 = T_0 \wedge C$ and $\tilde{T} = T \wedge C$, and the censoring indicators $\Delta_0 = \mathbb{1}\{T_0 < C\}$ and $\Delta = \mathbb{1}\{T < C\}$. The right-censored counting processes are $\tilde{Y}_0(t) = \mathbb{1}\{t \leq \tilde{T}_0\}$, $\tilde{Y}_1(t) = \mathbb{1}\{\tilde{T}_0 < t \leq \tilde{T}\}$, $\tilde{N}_{0k}(t) = \mathbb{1}\{\tilde{T}_0 \leq t, \eta = k, \Delta_0 = 1\}$, for $k \in \{1,2\}$ and $\tilde{N}_{12}(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 1, \eta = 1\}$ and the censoring process is given by $\tilde{N}_C(t) = \mathbb{1}\{\tilde{T} \leq t, \Delta = 0\}$. The following definition of the coarsening at random (CAR) assumption is based on van der Vaart [2004, section 2.1].

Assumption 2.1 (CAR)**.** There is a measurable function $\tilde{r}$ which characterizes the density of the conditional distribution r of C given (T_0, T, W) such that almost everywhere:

$$r(u|T_0,T,W) = \tilde{r}(u|T_0 \wedge u, T \wedge u, W).$$

We assume that the distribution Q of the uncensored random variables (T_0, T, W) belongs to a family of probability measures $\mathcal{Q}$ and that the distribution P of the censored random variables $(\tilde{T}_0, \tilde{T}, \Delta_0, \Delta_1, W)$ belongs to a family of probability measures $\mathcal{P}$. Under CAR, without additional assumptions about $\mathcal{Q}$ the set $\mathcal{P}$ is not restricted and $Q \in \mathcal{Q}$ is identifiable from $P \in \mathcal{P}$ under a positivity assumption [Gill et al., 1997, van der Vaart, 2004]. Positivity means that the probability of observing the variables (T_0, T) is strictly positive given W. For practical applications the positive assumption implies that we can only identify the conditional distribution of the process X given W on a time interval where the probability of censoring is strictly less than 100%. To achieve this, we assume in Corollary 3.1 that the

cumulative hazard function for the censoring distribution is bounded by a finite constant on a bounded interval conditional on the covariates. Lemma 2.2 and Corollary 3.1 below establish identifiability results for our setting. Let $\gamma_0 \colon \mathbb{R}_+ \times \mathcal{W} \to \mathbb{R}_+$ and $\gamma_1 \colon \mathcal{U} \times \mathcal{W} \to \mathbb{R}_+$ denote respectively the transition hazard functions for the censoring process for individuals in state 0 and state 1 with respect to the filtration $\mathcal{F}_t \vee \sigma\{N_C(t)\}$, and denote by $\tilde{\mathcal{F}}_t$ the filtration generated by $(\tilde{N}_{01}(t), \tilde{N}_{02}(t), \tilde{N}_{12}(t), \tilde{N}_C(t), W)$.

Lemma 2.2. *Under* CAR *the intensity processes of the observed counting processes and the censoring process with respect to $\tilde{\mathcal{F}}_t$ are given by*

$$\mathbb{E}(\tilde{N}_{0k}(\mathrm{d}t)|\tilde{\mathcal{F}}_{t-}) = h_{0k}(t, W)\tilde{Y}_0(t)\mathrm{d}t, \quad \textit{for} \quad k \in \{1, 2\},$$
$$\mathbb{E}(\tilde{N}_{12}(\mathrm{d}t)|\tilde{\mathcal{F}}_{t-}) = h_{12}(t, T_0, W)\tilde{Y}_1(t)\mathrm{d}t,$$

and

$$\mathbb{E}(\tilde{N}_C(\mathrm{d}t)|\tilde{\mathcal{F}}_{t-}) = \Big\{\gamma_0(t, W)\tilde{Y}_0(t) + \gamma_1(t, T_0, W)\tilde{Y}_1(t)\Big\}\,\mathrm{d}t.$$

Proof. The proof of the lemma is given in Appendix A. □

3 Target and nuisance parameters

In the context of the illness death model, parameters of interest are transition probabilities [e.g., Allignol et al., 2014] as well as state occupation probabilities [e.g., Datta and Satten, 2001]. Our methods require that the parameter is defined by a smooth functional of the distribution $Q \in \mathcal{Q}$ and identifiable under CAR via Lemma 2.2 from the censored data distribution $P \in \mathcal{P}$. Here and in what follows we consider as parameter of interest the probability of being in the illness state at time horizon $t > 0$, i.e., $Q(T_0 \leq t, T > t)$. We define the cumulative hazard functions for the possible transitions as

$$H_{0k}(t, w) = \int_0^t h_{0k}(s, w)\,\mathrm{d}s, \quad H_{12}(t, s, w) = \int_s^t h_{12}(u, s, w)\,\mathrm{d}u,$$

for $k \in \{1, 2\}$, and similarly we define the cumulative censoring hazard functions as

$$\Gamma_0(t, w) = \int_0^t \gamma_0(s, w)\,\mathrm{d}s, \quad \Gamma_1(t, s, w) = \int_s^t \gamma_1(u, s, w)\,\mathrm{d}u.$$

By μ we denote the marginal distribution of W, and we let $\nu = (H_{01}, H_{02}, H_{12}, \Gamma_0, \Gamma_1, \mu)$ denote the set of all nuisance parameters characterizing the distribution Q and the censoring mechanism. In the Appendix we show how Lemma 2.2 implies that H_{01}, H_{02}, and H_{12}, restricted to the interval $[0, t]$, are identifiable from the observed data distribution P under a positivity assumption and CAR, and how we can express the target parameter $Q(T_0 \leq t, T > t)$ as a functional of these nuisance parameters and μ. This gives the following identifiability result.

Corollary 3.1. *Assume* CAR *and that there is a fixed constant $K < \infty$ such that $\Gamma_0(t, w) < K$ and $\Gamma_1(t, u, w) < K$ for all $w \in \mathcal{W}$ and $u \in [0, t]$. Then we have that $Q(T_0 \leq t, T > t) = \Psi_t(P)$ where $\Psi_t : \mathcal{P} \to \mathbb{R}$ is defined as*

$$\Psi_t(P) = \tilde{\Psi}_t(\nu) = \int_{\mathcal{W}} \rho(t, 0, w; \nu)\mu(\mathrm{d}w), \tag{1}$$

where

$$\begin{aligned}&\rho(t,u,w;\nu)\\&=\int_u^t \exp\left(-\int_u^s \{h_{01}(z,w)+h_{02}(z,w)\}\,\mathrm{d}z-\int_s^t h_{12}(z,s,w)\,\mathrm{d}z\right)H_{01}(\mathrm{d}s,w).\end{aligned} \tag{2}$$

Proof. See Appendix A. □

By the identifiability result above we may now focus on estimating the parameter Ψ_t defined on $\mathcal{P}$ from samples $O=(\tilde{T}_0,\tilde{T},\Delta_0,\Delta,W)\sim P\in\mathcal{P}$. The asymptotic distribution of any regular asymptotically linear estimator is uniquely characterized by its influence function [see e.g., van der Vaart, 2000]. The influence function IF_t of an estimator $\hat{\Psi}_t$ of the parameter Ψ_t under the model $\mathcal{P}$ is a function of the data O and the probability measure $P\in\mathcal{P}$ such that for all $P\in\mathcal{P}$,

$$\hat{\Psi}_t-\Psi_t(P)=\frac{1}{n}\sum_{i=1}^n \mathrm{IF}_t(O_i;P)+o_P(n^{-1/2})\quad\text{and}\quad\mathbb{E}[\mathrm{IF}_t(O_i;P)]=0, \tag{3}$$

when $O_i\sim P$. For a fully non-parametric model, all regular asymptotically linear estimators have the same influence function. In this case the influence function is closely related to the Gâteaux derivative of Ψ_t and therefore we use the same notation for the influence function and the Gâteaux derivative. For more details about the relation between the Gâteaux derivative and the influence function see Bickel et al. [1993], Ichimura and Newey [2022], Chernozhukov et al. [2018], Hampel [1974], Huber [2004]. For the parameter of interest defined in equation (1) the Gâteaux derivative is given as follows.

Lemma 3.2. *Under CAR, the Gâteaux derivative of Ψ_t at $P\in\mathcal{P}$ in direction of the Dirac measure δ_O of an observation $O=(\tilde{T}_0,\tilde{T},\Delta_0,\Delta,W)$ is given by*

$$\mathrm{IF}_t(O;\nu)=\rho(t,0,W;\nu)+\varphi_t(O;\nu)-\tilde{\Psi}_t(\nu)$$

where

$$\begin{aligned}\varphi_t(O;\nu)=&\int_0^t\left[e^{-H_{12}(t,u,W)}-\rho(t,u,W;\nu)\right]\frac{\tilde{N}_{01}(\mathrm{d}u)-\tilde{Y}_0(u)H_{01}(\mathrm{d}u,W)}{e^{-\Gamma_0(u,W)}}\\&-\int_0^t\rho(t,u,W;\nu)\frac{\tilde{N}_{02}(\mathrm{d}u)-\tilde{Y}_0(u)H_{02}(\mathrm{d}u,W)}{e^{-\Gamma_0(u,W)}}\\&-e^{\Gamma_0(\tilde{T}_0,W)}\int_0^t e^{-[H_{12}(t,\tilde{T}_0,W)-H_{12}(u,\tilde{T}_0,W)]}\frac{\tilde{N}_{12}(\mathrm{d}u)-\tilde{Y}_1(u)H_{12}(\mathrm{d}u,\tilde{T}_0,W)}{e^{-\Gamma_1(u,\tilde{T}_0,W)}}.\end{aligned} \tag{4}$$

Proof. See Appendix B. □

The Gâteaux derivative above depends on $P\in\mathcal{P}$, and above we abuse notation slightly by letting this dependence be implicit through ν which is uniquely determined by P. To formally calculate the information bound for Ψ_t we would have to show that IF_t is also the path-wise (or Hadamard) derivative of Ψ_t, which is a stronger result than the above lemma [Bickel et al., 1993, van der Vaart, 2000] and can be shown in several ways, see for instance van der Vaart [1991], van der Laan and Robins [2003], Tsiatis [2006], Bickel et al. [1993], Ichimura and Newey [2022]. Below we use the Gâteaux derivative IF_t to motivate a class of estimators and then directly show that these estimators are asymptotically linear. For this purpose, Lemma 3.2 is sufficient.

4 Targeted estimation

We construct a class of asymptotically linear estimators for the parameter $\Psi_t(P)$ that has IF_t as its influence function. Assume that $\{O_i\}_{i=1}^n$ are n independent draws from the distribution P. Given estimators $\hat{\nu} = (\hat{H}_{01}, \hat{H}_{02}, \hat{H}_{12}, \hat{\Gamma}_0, \hat{\Gamma}_1, \hat{\mu})$ of all the nuisance parameters, we define the plug-in estimator based on equation (1):

$$\hat{\Psi}_t^0 = \int \rho(t, 0, w; \hat{\nu}) \hat{\mu}(\mathrm{d}w). \tag{5}$$

The superscript '0' is used to indicate that this is an initial estimator. Note that this estimator does not use the estimators of the cumulative censoring hazard functions $\hat{\Gamma}_0$ and $\hat{\Gamma}_1$. Denoting by $\mathbb{P}_n$ the empirical measure corresponding to $\{O_i\}_{i=1}^n$ and writing $\mathbb{P}_n[f(\cdot;\nu)] = n^{-1}\sum_{i=1}^n f(O_i;\nu)$, we define the one-step estimator

$$\hat{\Psi}_t = \hat{\Psi}_t^0 + \mathbb{P}_n[\mathrm{IF}_t(\cdot;\hat{\nu})]. \tag{6}$$

To motivate this estimator we provide two interpretations of the term $\mathbb{P}_n[\mathrm{IF}_t(\cdot;\hat{\nu})]$: It can be interpreted as an approximation of the first order asymptotic bias due to estimation of the nuisance parameter ν. This interpretation motivates the general strategy of adjusting the initial estimator $\hat{\Psi}_t^0$ by adding $\mathbb{P}_n[\mathrm{IF}_t(O;\hat{\nu})]$ and is referred to as one-step estimation or debiasing [Kraft and van Eeden, 1972, Bickel, 1975, Pfanzagl and Wefelmeyer, 1982, Fisher and Kennedy, 2021, Chernozhukov et al., 2018]. Alternatively, $\nu \mapsto \mathbb{P}_n[\mathrm{IF}_t(O;\nu)]$ can be interpreted as an empirical score function, which suggests that we should look for an estimator $\hat{\nu}$ that solves $\mathbb{P}_n[\mathrm{IF}_t(O;\nu)] = 0$. This interpretation motivates targeted minimum loss based estimation (TMLE) [van der Laan and Rubin, 2006, van der Laan and Rose, 2011], which substitutes an updated estimator $\hat{\nu}_*$ for the nuisance parameter estimator $\hat{\nu}$ such that $\mathbb{P}_n[\mathrm{IF}_t(O;\hat{\nu}_*)] = 0$. The two approaches are asymptotically equivalent, and for simplicity we focus on the one-step estimation approach in the following.

Theorem 4.2 below states that the one-step estimator in equation (6) is asymptotically linear and normally distributed under a set of conditions concerning the estimator of the nuisance parameter ν. To ensure these conditions, one approach is to specify (semi-)parametric models for the conditional transition hazard. To decrease the risk of model misspecification, our strategy is to use data-adaptive methods and to use cross-validation to select the best model from a class of candidate models. For example, in Section 5 we use a penalized Poisson regression approach where the penalty parameter is selected based on cross-validation. The asymptotic distribution of an estimator selected by cross-validation is difficult to analyze, and an important consequence of the following theorem is that we only need results on the rate of convergence of these estimators to make asymptotically valid inference.

We now state the assumptions needed to prove our main theorem. In the following we use the notation $P[f] = \int f(o)P(\mathrm{d}o)$, and we let $\|\cdot\|_{\mathcal{L}^2_P(\mathcal{Z})}$ denote the $\mathcal{L}^2$-norm of real-valued functions defined on $\mathcal{Z}$ with respect to the measure P, i.e., $\|f\|_{\mathcal{L}^2_P(\mathcal{Z})} = (P[f^2])^{1/2}$. We also define $\mathcal{U}_t := \mathcal{U} \cap [0,t]^2$ for any $t > 0$. Recall from Section 2 that we assumed the existence of intensities for $\mathbf{N}$ and N_C. Below we use h_{01}, h_{02}, h_{12}, γ_0, and γ_1 to denote the transition hazard functions corresponding to a given $P \in \mathcal{P}$ under consideration. Notably, we will assume that also the *estimators* of the cumulative hazard functions have densities with respect to Lebesgue measure, i.e., we assume that we can write

$$\begin{gathered}\hat{H}_{0k}(t,w) = \int_0^t \hat{h}_{0k}(s,w)\,\mathrm{d}s, \quad \text{for} \quad k \in \{1,2\}, \quad \hat{H}_{12}(t,s,w) = \int_s^t \hat{h}_{12}(u,s,w)\,\mathrm{d}u, \\ \hat{\Gamma}_0(t,w) = \int_0^t \hat{\gamma}_0(s,w)\,\mathrm{d}s, \quad \text{and} \quad \hat{\Gamma}_1(t,s,w) = \int_s^t \hat{\gamma}_1(u,s,w)\,\mathrm{d}u.\end{gathered} \tag{7}$$

In Section 5 we propose to estimate the nuisance parameters by estimating the hazard functions $(h_{01}, h_{02}, h_{12}, \gamma_0, \gamma_1)$ directly, and hence equation (7) will in that case be satisfied.

Assumption 4.1. Let $\mathcal{P}$ be a family of probability measures specifying a distribution for $O = (\tilde{T}_0, \tilde{T}, \Delta_0, \Delta, W)$. We assume the following for any $P \in \mathcal{P}$.

(i) There exists a constant $K \in (0, \infty)$ such that

$$\frac{1}{K} < h_{01}(u,w), h_{02}(u,w), \gamma_0(u,w) < K, \quad \text{for all} \quad (u,w) \in [0,t] \times \mathcal{W},$$

and

$$\frac{1}{K} < h_{12}(u,s,w), \gamma_1(u,s,w) < K, \quad \text{for all} \quad (u,s,w) \in \mathcal{U}_t \times \mathcal{W}.$$

(ii) Let $\hat{\nu} = (\hat{H}_{01}, \hat{H}_{02}, \hat{H}_{12}, \hat{\Gamma}_0, \hat{\Gamma}_1, \hat{\mu})$ denote the tuple of estimators of the nuisance parameter ν, where $\hat{\mu}$ denotes the empirical measure on $\mathcal{W}$. There exist $\hat{h}_{01}$, $\hat{h}_{02}$, $\hat{h}_{12}$, $\hat{\gamma}_0$, and $\hat{\gamma}_1$ so that equation (7) holds, and we have

$$\begin{aligned} \|\hat{h}_{0k} - h_{0k}\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{W})} &= o_P(n^{-1/4}), \quad \text{for} \quad k \in \{1,2\}, \\ \|\hat{\gamma}_0 - \gamma_0\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{W})} &= o_P(n^{-1/4}), \\ \|\hat{h}_{12} - h_{12}\|_{\mathcal{L}^2_{m\otimes\mu}(\mathcal{U}_t\times\mathcal{W})} &= o_P(n^{-1/4}), \quad \text{and} \\ \|\hat{\gamma}_1 - \gamma_1\|_{\mathcal{L}^2_{m\otimes\mu}(\mathcal{U}_t\times\mathcal{W})} &= o_P(n^{-1/4}), \end{aligned}$$

where m denotes Lebesgue measure. Furthermore, there exists a constant $C < \infty$ such that $\hat{H}_{01}$, $\hat{H}_{02}$, $\hat{H}_{12}$, $\hat{\Gamma}_0$, and $\hat{\Gamma}_1$ are uniformly bounded by C with probability tending to one.

(iii) There exists a Donsker class of functions $\mathcal{F} = \{f : \mathcal{U}_t \times \mathcal{W} \to \mathbb{R}_+\}$ such that $\mathrm{IF}_t(\cdot; \hat{\nu}) \in \mathcal{F}$ with probability tending to one.

Assumption (i) is a general assumption about the family $\mathcal{P}$. Assumption (ii) is specific to the choice of estimators of the nuisance parameters and states that we are able to estimate the transition hazards at a specific rate. Assumption (iii) also concerns the estimators of the cumulative transition hazards and will for instance hold if the estimators of the cumulative transition hazards belong to a class of functions with uniformly bounded sectional variation norm (see for instance [van der Vaart, 2000, van der Vaart and Wellner, 1996, Gill et al., 1995, van der Laan and Benkeser, 2018]). In Section 5 we propose highly data-adaptive estimators of the transition hazard functions and indicate how assumptions (ii) and (iii) can be established for such estimators.

Theorem 4.2. *Assume that Assumption 4.1 holds for the family $\mathcal{P}$ and the tuple of estimators $\hat{\nu}$. Then for any $P \in \mathcal{P}$ and with $\hat{\Psi}_t$ defined in equation (6), it holds that*

$$\hat{\Psi}_t - \Psi_t = (\mathbb{P}_n - P)[\mathrm{IF}_t(\cdot; \nu)] + o_P(n^{-1/2}), \tag{8}$$

i.e., $\hat{\Psi}_t$ is asymptotically linear with influence function $\mathrm{IF}_t(\cdot; \nu)$.

Proof. We have the expansion

$$\begin{aligned} \hat{\Psi}^0_t - \Psi_t(P) &= \mathbb{P}_n[\rho(t,0,\cdot;\hat{\nu})] - P[\rho(t,0,\cdot;\nu)] \\ &= \mathbb{P}_n[\rho(t,0,\cdot;\hat{\nu})] - P[\rho(t,0,\cdot;\nu)] \pm (\mathbb{P}_n - P)[\mathrm{IF}_t(\cdot;\hat{\nu})] \\ &= (\mathbb{P}_n - P)[\mathrm{IF}_t(\cdot;\hat{\nu})] - \mathbb{P}_n[\mathrm{IF}_t(\cdot;\hat{\nu})] + \mathrm{Rem}(P,\hat{\nu}), \end{aligned} \tag{9}$$

where the remainder term is by definition

$$\mathrm{Rem}(P, \hat{\nu}) = \mathbb{P}_n[\rho(t, 0, \cdot; \hat{\nu})] - P[\rho(t, 0, \cdot; \nu)] + P[\mathrm{IF}_t(\cdot; \hat{\nu})]. \tag{10}$$

In Appendix C we show that Assumptions 4.1 (i) and (ii) imply that

$$\mathrm{Rem}(P, \hat{\nu}) = o_P(n^{-1/2}), \tag{11}$$

see in particular Proposition C.1. Assumptions 4.1 (i) and (ii) also imply that $\| \mathrm{IF}_t(\cdot; \hat{\nu}) - \mathrm{IF}_t(\cdot; \nu)\|_{\mathcal{L}^2_P(\mathcal{U}_t \times \mathcal{W})} \xrightarrow{P} 0$, and thus Assumption 4.1 (iii) and lemma 19.24 in van der Vaart [2000] states that

$$(\mathbb{P}_n - P)[\mathrm{IF}_t(\cdot; \hat{\nu})] = (\mathbb{P}_n - P)[\mathrm{IF}_t(\cdot; \nu)] + o_P(n^{-1/2}). \tag{12}$$

Combining equations (9), (11), and (12) we obtain

$$\hat{\Psi}^0_t - \Psi_t(P) = (\mathbb{P}_n - P)[\mathrm{IF}_t(\cdot; \nu)] - \mathbb{P}_n[\mathrm{IF}_t(\cdot; \hat{\nu})] + o_P(n^{-1/2}).$$

Adding $\mathbb{P}_n[\mathrm{IF}_t(\cdot; \hat{\nu})]$ to both sides of the equation above and using the definition of the one-step estimator (equation (6)) gives the claimed asymptotic representation in equation (8). The final claim follows by the definition of an influence function (equation (3)). □

Before detailing our suggested estimation method for the conditional hazard functions (Section 5), we consider the special case when no baseline information is available. In this setting the most natural estimator of the transition rates from the healthy state to another state is the Nelson-Aalen estimator [Andersen et al., 2012]. To estimate the cause-specific cumulative hazard functions of the first transitions, H_{01}, H_{02}, and Γ_0, the corresponding Nelson-Aalen estimators are

$$\begin{aligned} \hat{\Gamma}_0(t) &= \int_0^t \mathbb{1}\left\{\sum_{i=1}^n \tilde{Y}_{0,i}(u) > 0\right\} \frac{\sum_{i=1}^n \tilde{N}_{C,i}(\mathrm{d}u)}{\sum_{i=1}^n \tilde{Y}_{0,i}(u)}, \quad \text{and} \\ \hat{H}_{0k}(t) &= \int_0^t \mathbb{1}\left\{\sum_{i=1}^n \tilde{Y}_{0,i}(u) > 0\right\} \frac{\sum_{i=1}^n \tilde{N}_{0k,i}(\mathrm{d}u)}{\sum_{i=1}^n \tilde{Y}_{0,i}(u)}, \end{aligned} \tag{13}$$

for $k \in \{1, 2\}$. In the remainder of this section the nuisance parameter is $\nu = (H_{01}, H_{02}, H_{12}, \Gamma_0, \Gamma_1)$ and the estimators $\hat{H}_{01}$, $\hat{H}_{02}$, and $\hat{\Gamma}_0$ are now assumed to be Nelson-Aalen estimators, while H_{12} and Γ_1 can for instance be estimated as proposed in Section 5. In this situation the one-step estimator given in equation (6) simplifies as shown in the following lemma.

Lemma 4.3. *If $\hat{H}_{01}$ and $\hat{H}_{02}$ are Nelson-Aalen estimators as defined in equation* (13) *the one-step estimator of equation* (6) *is given by:*

$$\begin{aligned} \hat{\Psi}_t = &\int_0^t e^{-\hat{H}_{12}(t,u) - \hat{H}_{01}(u) - \hat{H}_{02}(u)} \hat{H}_{01}(\mathrm{d}u) \\ &- \frac{1}{n} \sum_{i=1}^n e^{\hat{\Gamma}_0(\tilde{T}_{0,i})} \int_0^t e^{-[\hat{H}_{12}(t, \tilde{T}_{0,i}) - \hat{H}_{12}(u, \tilde{T}_{0,i})]} \frac{\tilde{N}_{12,i}(\mathrm{d}u) - \tilde{Y}_{1,i}(u) \hat{H}_{12}(\mathrm{d}u, \tilde{T}_{0,i})}{e^{-\hat{\Gamma}_1(u, \tilde{T}_{0,i})}}. \end{aligned} \tag{14}$$

Proof. First note that $\mathbb{P}_n[\rho(t, 0, \cdot; \hat{\nu})] = \tilde{\Psi}(\hat{\nu}) = \mathbb{P}_n[\tilde{\Psi}(\hat{\nu})]$. This also holds in the case with covariates as long as we use the empirical measure to estimate μ. It follows that $\mathbb{P}_n[\mathrm{IF}_t(\cdot; \nu)] = \mathbb{P}_n[\varphi_t(\cdot; \nu)]$ with φ_t defined in equation (4). Here we abuse notation slightly

because $\varphi_t(\cdot;\nu)$ is not a function of W in the present setting. The one-step estimator is given by

$$\begin{aligned}
\hat{\Psi}_t &= \tilde{\Psi}(\hat{\nu}) + \mathbb{P}_n[\varphi_t(\cdot;\hat{\nu})] \\
&= \int_0^t e^{-\hat{H}_{12}(t,u)-\hat{H}_{01}(u)-\hat{H}_{02}(u)} \hat{H}_{01}(\mathrm{d}u) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{e^{-\hat{H}_{12}(t,u)} - \rho(t,u;\hat{\nu})}{e^{-\hat{\Gamma}_0(u)}} \left[\tilde{N}_{01,i}(\mathrm{d}u) - \tilde{Y}_{0,i}(u)\hat{H}_{01}(\mathrm{d}u)\right] \\
&\quad - \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{\rho(t,u;\hat{\nu})}{e^{-\hat{\Gamma}_0(u)}} \left[\tilde{N}_{02,i}(\mathrm{d}u) - \tilde{Y}_{0,i}(u)\hat{H}_{02}(\mathrm{d}u)\right] \\
&\quad - \frac{1}{n}\sum_{i=1}^n e^{\hat{\Gamma}_0(\tilde{T}_{0,i})} \int_0^t \frac{e^{-[\hat{H}_{12}(t,\tilde{T}_{0,i})-\hat{H}_{12}(u,\tilde{T}_{0,i})]}}{e^{-\hat{\Gamma}_1(u,\tilde{T}_{0,i})}} \left[\tilde{N}_{12,i}(\mathrm{d}u) - \tilde{Y}_{1,i}(u)\hat{H}_{12}(\mathrm{d}u)\right],
\end{aligned}$$

where the second equality follows by definition. We claim that when $\hat{H}_{01}$ and $\hat{H}_{02}$ are Nelson-Aalen estimators the two middle terms in the above display are zero. This follows by a direct calculation as follows. For any function f, let $f^t(u) := f(u)\cdot \mathbb{1}\{u \le t\}$. We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \int_0^t f(u)\tilde{Y}_{0,i}(u)\hat{H}_{0k}(\mathrm{d}u) &= \frac{1}{n}\sum_{j=1}^n \sum_{i=1}^n \int_0^{\tilde{T}_{0,i}} f^t(u) \frac{\tilde{N}_{0k,j}}{\sum_{l=1}^n \tilde{Y}_{0,l}(u)} \\
&= \frac{1}{n}\sum_{j=1}^n \sum_{i=1}^n \frac{\Delta_j f^t(\tilde{T}_{0,j})\mathbb{1}\{\tilde{T}_{0,j} \le \tilde{T}_{0,i}\}}{\sum_{l=1}^n \tilde{Y}_{0,l}(\tilde{T}_{0,j})} \\
&= \frac{1}{n}\sum_{j=1}^n \Delta_j f^t(\tilde{T}_{0,j}) \sum_{i=1}^n \frac{\mathbb{1}\{\tilde{T}_{0,j} \le \tilde{T}_{0,i}\}}{\sum_{l=1}^n \mathbb{1}\{\tilde{T}_{0,j} \le \tilde{T}_{0,l}\}} \\
&= \frac{1}{n}\sum_{j=1}^n \Delta_j f^t(\tilde{T}_{0,j}) \\
&= \frac{1}{n}\sum_{i=1}^n \int_0^t f(u)\tilde{N}_{0k,i}(\mathrm{d}u).
\end{aligned}$$

Using this with $f(u) = [e^{-\hat{H}_{12}(t,u)} - \rho(t,u;\hat{\nu})]e^{\hat{\Gamma}_0(u)}$ and $f(u) = \rho(t,u;\hat{\nu})e^{\hat{\Gamma}_0(u)}$ gives the result. □

Using similar arguments as given in the proof of Theorem 4.2 we can establish asymptotic normality of the estimator given in equation (14). Unfortunately, equation (7) does not hold for the Nelson-Aalen estimator of the transition hazard functions, and hence this result is not a direct corollary of Theorem 4.2. In Appendix C.1 we demonstrate how we can use empirical process theory to prove an analogous version of Theorem 4.2 in a setting without baseline covariates.

5 Nuisance parameter estimation

To apply Theorem 4.2 we need estimators of h_{01}, h_{02}, h_{12}, γ_0, and γ_1. We suggest to use a penalized Poisson regression approach to model each of the conditional hazard functions, similar to the approach taken by Rytgaard et al. [2021]. A Poisson regression model is suitable when the conditional hazard function can be approximated by a piece-wise constant function of time and covariates. By letting the grid on which the hazard function is approximated be

sufficiently fine with increasing sample size we can approximate a large class of models. Another useful property is that when the nuisance parameter estimators $(\hat{h}_{01}, \hat{h}_{02}, \hat{h}_{12}, \hat{\gamma}_0, \hat{\gamma}_1)$ are piece-wise constant, the estimated martingale integrals in equation (6) are straightforward to calculate analytically (see Appendix D). We focus on estimating a conditional hazard function h defined on $\mathcal{U}_t \times \mathcal{W}$. Here we have the nuisance parameter h_{12} in mind but estimation of γ_1 proceeds in exactly the same way. Estimation of h_{01}, h_{02}, and γ_0 can be done in a similar but simpler way as these functions only have one time-varying argument. We approximate $\log h$ as a linear combination of basis functions indexed by some time grid $0 = t_0 < t_1 < \cdots < t_K = t$, i.e.,

$$\log h(u, s, w) = \sum_{k=1}^{K} \sum_{l=1}^{k} \mathbb{1}\{t_{k-1} \leq u < t_k, t_{l-1} \leq s < t_l\} h_{k,l}(w),$$

where $h_{k,l} \colon \mathcal{W} \to \mathbb{R}$ for all (k, l) with $1 \leq l \leq k \leq K$ and K is the number of intervals defining the grid. We then model each $h_{k,l}$ as a linear combination of p basis functions $f_j \colon \mathcal{W} \to \mathbb{R}$, $j = 1, \ldots, p$, so that

$$h_{k,l}(w) = \sum_{j=1}^{p} \beta_{k,l,j} f_j(w), \quad \text{for some} \quad (\beta_{k,l,1}, \ldots, \beta_{k,l,p})^T \in \mathbb{R}^p, \quad \text{with} \quad 1 \leq l \leq k \leq K.$$

One option is to use a partition $\mathcal{W}_1, \ldots, \mathcal{W}_p$ of the covariate space $\mathcal{W}$ to construct the basis functions $f_j(w) := \mathbb{1}\{w \in \mathcal{W}_j\}$, $j = 1, \ldots, p$. With this choice of basis functions we can index h by a collection of coefficients $\beta := \{ \beta_{k,l,j} : 1 \leq l \leq k \leq K, j \in \{1, \ldots, p\}\}$ and define h_β:

$$\log h_\beta(u, s, w) = \sum_{k=1}^{K} \sum_{l=1}^{k} \sum_{j=1}^{p} \beta_{k,l,j} \cdot \mathbb{1}\{t_{k-1} \leq u < t_k, t_{l-1} \leq s < t_l, w \in \mathcal{W}_j\}. \tag{15}$$

Estimation of h_β then reduces to estimation of the $pK(K+1)/2$ number of coefficients. When the covariate space $\mathcal{W}$ is discrete we can use the collection of all singletons as a partition, which corresponds to considering interactions of all order between the covariates and the time grid. We refer to this as the saturated model. When the number of covariates is not too large we can use this approach to approximate a completely non-parametric model. For a large number of baseline covariates this approach is not feasible, and we can instead consider all interactions up to some order.

To be more specific, we now consider estimation of the transition hazard function $h = h_{12}$ with a data set $\mathcal{D} = \{O_1, \ldots, O_n\}$ of observations $O_i \sim P \in \mathcal{P}$. Assuming h to be on the form in equation (15) the negative log-likelihood for β is equal to

$$\begin{aligned} &L(\beta; \mathcal{D}) \\ &= -\sum_{i=1}^{n} \Delta_{0,i} \mathbb{1}\{\eta_i = 1\} \left(\Delta_i \log h_\beta(\tilde{T}_i, T_{0,i}, W_i) - \int_{T_{0,i}}^{\tilde{T}_i} h_\beta(u, T_{0,i}, W_i) \, \mathrm{d}u \right) + C \\ &= -\sum_{k=1}^{K} \sum_{l=1}^{k} \sum_{j=1}^{p} \left(D_{k,l,j} \log h_\beta(t_{k-1}, t_{l-1}, w_j) - R_{k,l,j} h_\beta(t_{k-1}, t_{l-1}, w_j) \right) + C, \end{aligned} \tag{16}$$

where C is a constant depending only on the data and the censoring distribution, $w_j \in \mathcal{W}_j$ for each $j = 1, \ldots, p$,

$$D_{k,l,j} = \sum_{i=1}^{n} \Delta_{0,i} \mathbb{1}\{\eta_i = 1\} \mathbb{1}\Big\{ t_{k-1} \leq \tilde{T}_i < t_k, \Delta_i = 1, t_{l-1} \leq T_{0,i} < t_l, W_i \in \mathcal{W}_j \Big\},$$

and

$$R_{k,l,j} = \sum_{i=1}^{n} \Delta_{0,i} \mathbb{1}\{\eta_i = 1\} \mathbb{1}\left\{t_{k-1} \leq \tilde{T}_i, t_{l-1} \leq T_{0,i} < t_l, W_i \in \mathcal{W}_j\right\} \left([t_k \wedge \tilde{T}_i] - [t_{k-1} \vee T_{0,i}]\right).$$

The likelihood in equation (16) is then recognized as the likelihood of a Poisson model with event counts $D_{k,l,j}$, mean values $h_\beta(t_{k-1}, t_{l-1}, w_j)$ and offset given by $\log R_{k,l,j}$. When the time grid and partition of $\mathcal{W}$ are fairly coarse and we have a large data set, we can achieve a substantial amount of dimension reduction by accumulating the data into counts $D_{k,l,j}$ and accumulated risk time $R_{k,l,j}$. However, for finer grids and partitions the total number of coefficients to be estimated can easily be of the same or higher order than n. In these cases, optimizing the likelihood in (16) with respect to β might either be an ill-posed problem or highly unstable in practice. To alleviate this we suggest to penalize the coefficients $\{\beta_{k,l,j}\}$ and use e.g., the LASSO estimator $\hat{\beta}_\lambda$ [Tibshirani, 1996],

$$\hat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \left\{ L(\beta, \mathcal{D}) + \lambda \sum_{k,l,j} |\beta_{k,l,j}| \right\}.$$

Besides making the optimization problem well-defined the penalization parameter λ introduces a selection procedure. The solution will be attained at a value β where some $\beta_{k,l,j}$ coefficients are zero. To select a value for λ we use V-fold cross-validation with the loss function L given in equation (16). Letting $\hat{\lambda}$ denote the value selected by the cross-validation procedure and $\hat{\beta} := \hat{\beta}_{\hat{\lambda}}$, we define $h_{\hat{\beta}}$ as our estimator of h.

By using theoretical results for cross-validated LASSO estimators [Chetverikov et al., 2021] or general results for hyperparameter selection using cross-validation [van der Vaart et al., 2006, van der Laan and Dudoit, 2003] it is possible to establish Assumption 4.1 (ii) even when the number of grid points K and the partition size p increase with the sample size. A rigorous derivation of the convergence rates of $h_{\hat{\beta}}$ is beyond the scope of this paper. In the next section we examine the finite sample performance of the one-step estimator from equation (6) when we use $h_{\hat{\beta}}$ to estimate the conditional transition hazard functions.

6 Empirical study

To examine the finite sample performance of the one-step estimator and to demonstrate the importance of the debiasing step when using data-adaptive methods we conduct a simulation study for the setting without baseline covariates. We simulate data from an irreversible illness-death model observed on the interval from 0 to 10. All individuals who are still alive at time $t = 10$ are administratively censored and we also introduce an additional right-censoring scheme. The full data distribution is generated according to a piece-wise constant hazard model. Healthy individuals have a constant hazard of illness and a (different) constant hazard of dying. For individuals in the illness state the conditional hazard of dying is also constant, but depends on the time illness occurred. Specifically, the data are generated according to

$$h_{01}(u) = 0.3, \quad h_{02}(u) = 0.1, \quad \text{and} \quad h_{12}(u, s) = \sum_{i=1}^{10} \mathbb{1}\{i - 1 \leq s < i\} \beta_i,$$

for all $u \in [0, 10]$ and all $s \leq u$, where the β_i's are chosen as the equally spaced grid of 10 decreasing numbers with $\beta_1 = 0.3$ and $\beta_{10} = 0.01$. In addition to the administrative

censoring we simulate a censoring time with constant hazard that depends on the state. We use $\gamma_0(u) = 0.2$ and $\gamma_1(u, s) = \alpha$, for $\alpha \in \{0.02, 0.11, 0.20\}$, corresponding to three different censoring regimes with more or less state-dependent censoring.

We report results for estimation of the probability of being ill and alive at time $t = 9$, i.e., Ψ_9 with Ψ_t defined in equation (1). In each simulated data set we fit three estimators: The Aalen-Johansen estimator, the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the one-step estimator $\hat{\Psi}_t$ defined in equation (6). Recall from Lemma 4.3 that we only need to estimate the transition hazard h_{12} and γ_1 with the penalized Poisson regression approach from Section 5 to obtain the one-step estimator. To estimate these nuisance parameters we use the implementation of penalized Poisson regression available through the `R`-package `glmnet` [Friedman et al., 2010]. We use the time grid $0, 1, \ldots, 10$, and we use 10-fold cross-validation to select the penalization parameter. To calculate the Nelson-Aalen estimator we use the `R`-package `prodlim` [Gerds, 2019], and to calculate the Aalen-Johansen estimator we use the `R`-package `etm` [Allignol et al., 2011].

		Bias			SE			MSE		
Dependent censoring	n	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$
High ($\alpha = 0.02$)	200	-1.58	-3.59	-1.27	3.73	2.53	3.65	16.32	19.25	14.84
	500	-1.34	-2.84	-0.48	2.44	1.90	2.43	7.73	11.66	6.11
	1000	-1.46	-2.54	-0.43	1.70	1.54	1.81	5.00	8.81	3.43
	1500	-1.43	-2.22	-0.27	1.32	1.33	1.42	3.77	6.69	2.09
	5000	-1.39	-0.81	-0.10	0.76	0.85	0.81	2.50	1.37	0.67
	10000	-1.41	-0.46	-0.09	0.58	0.62	0.62	2.32	0.59	0.39
Medium ($\alpha = 0.011$)	200	-0.98	-2.77	-1.13	4.29	2.75	4.06	19.24	15.21	17.70
	500	-0.74	-2.72	-0.55	2.79	1.97	2.76	8.31	11.23	7.87
	1000	-0.87	-2.46	-0.46	1.96	1.68	1.95	4.58	8.85	4.01
	1500	-0.64	-2.07	-0.13	1.50	1.37	1.49	2.63	6.13	2.24
	5000	-0.65	-0.75	-0.02	0.91	0.94	0.93	1.25	1.44	0.86
	10000	-0.77	-0.45	-0.10	0.65	0.65	0.66	1.01	0.62	0.45
None ($\alpha = 0.2$)	200	-0.41	-1.62	-0.59	5.33	3.01	5.06	28.44	11.66	25.78
	500	0.21	-2.44	-0.23	3.52	2.23	3.31	12.38	10.89	10.94
	1000	-0.10	-2.30	-0.41	2.24	1.88	2.16	5.02	8.78	4.82
	1500	-0.00	-2.07	-0.20	1.83	1.42	1.78	3.32	6.31	3.21
	5000	0.01	-0.75	-0.06	1.10	1.04	1.09	1.20	1.64	1.18
	10000	-0.10	-0.44	-0.15	0.75	0.70	0.74	0.58	0.68	0.56

Table 1: The results of 200 simulations of the three estimators for different censoring regimes ('High', 'Medium', and 'Low') and sample sizes ($n \in \{200, 500, 1000, 1500, 5000, 10000\}$). The table shows the bias, standard error (SE), and mean squared error (MSE) of the Aalen-Johansen estimator (Aa-J), the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the debiased estimator $\hat{\Psi}_t$ defined in equation (6).

For each of the three censoring mechanisms we simulated a data set with n number of observations for $n \in \{200, 500, 1000, 1500, 5000, 10000\}$ and calculated the three estimators. This was repeated 200 times. The results are summarized in Table 1 in terms of bias, standard errors (SE), and mean squared errors (MSE), and Figure 2 shows the MSE against sample size. Figure 3 shows the distribution of the three estimators across the 200 simulated data sets.

We see that the Aalen-Johansen estimator is biased when state-dependent censoring is present and unbiased when the censoring is independent. In accordance with results of Gunnes et al. [2007], also in our study the bias is seen to be fairly small even for a high degree of state-dependent censoring. Furthermore, the bias does not vary much with n when the sample size is bigger than 500. The plug-in estimator $\hat{\Psi}_t^0$ is also biased, but its

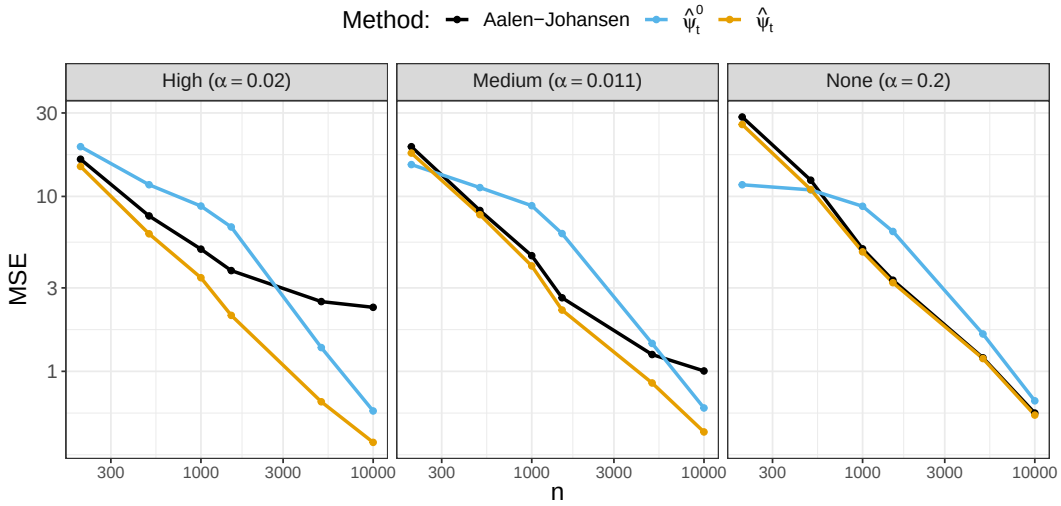


Figure 2: The mean squared error (MSE) estimated with 200 simulations of the three estimators for different censoring regimes ('High', 'Medium', and 'Low') plotted against sample size. Note the log scale on both axis. The estimators are the Aalen-Johansen estimator (Aa-J), the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the debiased estimator $\hat{\Psi}_t$ defined in equation (6).

bias decreases with n while it seems unaffected by the degree of state-dependent censoring. Notably, in small samples the bias of the plug-in estimator is higher than the bias of the Aalen-Johansen estimator even for a high degree of state-dependent censoring. The debiased estimator $\hat{\Psi}_t$ has the lowest bias for almost all sample sizes when there is a high or medium degree of state-dependent censoring, and it has approximately the same bias as the Aalen-Johansen estimator when censoring is independent. We also see that the standard errors of all three estimators are comparable for each n, though the plug-in estimator tends to have smaller standard errors for small sample sizes. This reflects the fact that penalization trades off bias for a decrease in variance.

A central motivation for targeted or debiased learning is that a naive plug-in approach will typically give poor performance for estimation of a low-dimensional target parameter when machine learning is used to estimate one or several high-dimensional nuisance parameters, see for instance Chernozhukov et al. [2018]. Our simulation results for the plug-in estimator confirm this phenomenon in the setting of the illness-death model. Our empirical studies further indicate that the bias due to model misspecification in finite samples can be substantially smaller than the bias due to penalization and hyperparameter selection. We also conclude that the debiased estimator is in this setting able to compensate for the bias caused by penalization and hyperparameter selection, and that it provides good estimates across all censoring regimes and sample sizes considered.

We also performed a simulation study when baseline covariates were present. The results are shown in Appendix E.

7 Analysis of the PROVA trial

To illustrate our methods we consider data from the PROVA trial [PROVA Study Group, 1991]. This randomized clinical trial included 286 patients and was initiated to investigate

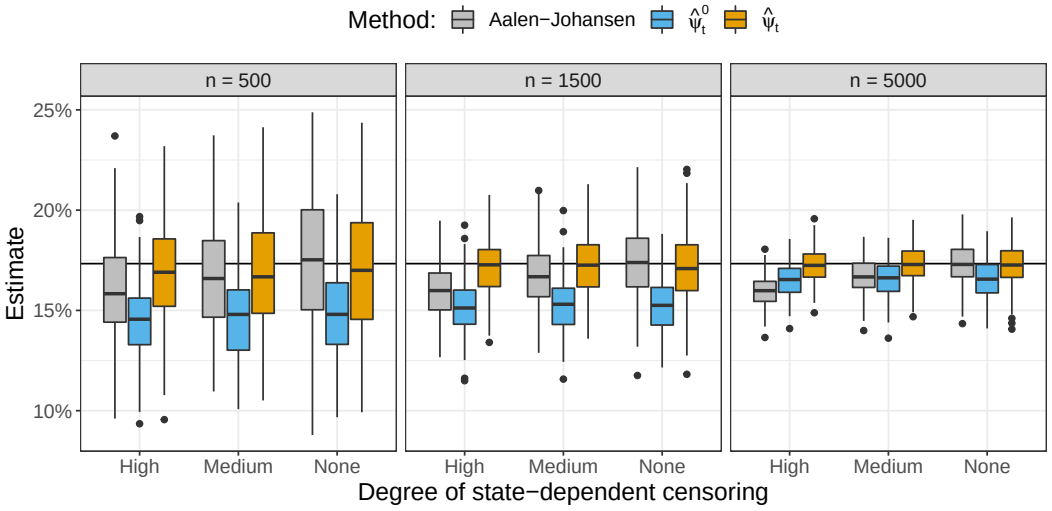


Figure 3: The results of 200 simulations of the three estimators for different censoring regimes and increasing number of samples, $n \in \{500, 1500, 5000\}$. The gray boxplots show the distribution of the Aalen-Johansen estimator, the blue boxplots show the distribution of the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the orange boxplots show the distribution of the debiased estimator $\hat{\Psi}_t$ defined in equation (6). The results are stratified according to both the number of samples n and the degree of state-dependent censoring. For the state-dependent censoring, 'High' is the regime with $\alpha = 0.02$, 'Medium' is $\alpha = 0.11$ and 'None' is $\alpha = 0.2$. The black line is the state occupation probability Ψ_t at time $t = 9$ of the data-generating distribution.

the effect of propanolol and/or sclerotherapy on the incidence of the first variceal bleeding among cirrhotic patients. We focus on the effect of receiving sclerotherapy compared to not receiving sclerotherapy on both the risk of variceal bleeding and death. At the time of randomization no patient had experienced variceal bleeding, and we consider this as the healthy state (state 0 in Figure 1). After randomization a patient can experience variceal bleeding, which we refer to as the illness state (state 1 in Figure 1). A patient can die both with or without experiencing variceal bleeding, so death is the absorbing state (state 2 in Figure 1). Patients in the study were censored either because they dropped out of the study (20 patients) or because the study ended (191 patients), see Figure 4. In addition a set of baseline variables was measured at the time of randomization: Age, prothrombin, bilirubin, sex, size of variceal size (grade 1-3), and whether or not the patient also received propanolol.

In this study the cause of censoring was not exclusively administrative, and hence it makes sense to use both the available baseline information as well as the time-dependent status of bleeding to model the censoring distribution. However, since there is no good understanding of how these variables would influence the censoring mechanism, (semi)-parametric models of these transitions are difficult to pre-specify. The risk of model misspecification can be avoided by using a data-adaptive model selection strategy that incorporates highly flexible models, such as the penalized Poisson regression approach described in Section 5.

To use the penalized Poisson regression approach we categorize each of the continuous variables age, prothrombin, and bilirubin into three groups determined by the quantiles of the empirical distribution at 33% and 66% probability. The time axis is discretized according to the empirical quantiles of $\tilde{T}_0$ at $10\%, 20\%, \ldots, 90\%$ probability. With this setup, the saturated Poisson regression models (which include all interactions) for the transitions hazards from state 0 contain $10 \times 3^4 \times 2^2 = 3240$ parameters, and the saturated models for

the transitions hazards from state 1 contain $55 \times 3^4 \times 2^2 = 17820$ parameters. For each transition hazard the penalty parameter λ is chosen by 10-fold cross-validation, which is repeated 5 times. The penalty parameter used for the final model is chosen as the average of the minimizer across these 5 repetitions, and the final model is estimated on the full data set with this penalty parameter. With estimates of the nuisance parameters h_{01}, h_{02}, h_{12}, γ_0, and γ_1 in hand, the one-step estimator Ψ_t defined in equation (6) is calculated using the algorithm given in Appendix D. The estimates of the nuisance parameters and the final one-step estimates are calculated separately in the two treatment arms.

The asymptotic representation of the estimator given in Theorem 4.2 yields point-wise confidence intervals for the state occupation probability. We can use this to conclude that the probability of being alive and having experienced a variceal bleeding 6 months after randomization is 2.11% lower for patients who received sclerotherapy, but that this difference is not statistically significant because the Wald-based confidence interval is $[-3.62\%, 7.84\%]$ which includes zero.

To give a more complete description of the effects of the treatment we also estimate the probability of being event-free, see Appendix B.1 for a brief outline of how an estimator of this parameter can be obtained. This provides us with estimates of all state occupation probabilities of the three possible states. The results are shown in Figure 5 as a function of time after randomization. We see that while treatment with sclerotherapy does not seem to have a significant effect on bleeding it instead seems to increase mortality.

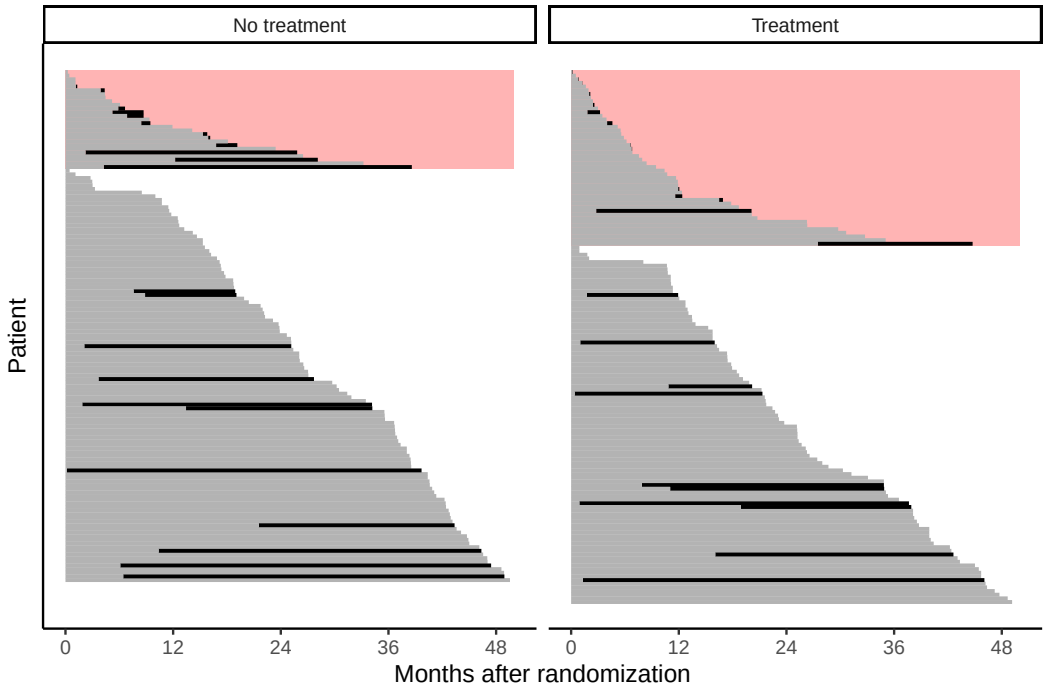


Figure 4: The data from the PROVA trial. Each line segment represents a patient. The gray line segments indicate that the patient is in the healthy state, and the black line segments indicate that the patient is in the illness state. The time of death was observed for the patients in the red area, and for these patients the end of the line segments indicate the time of death. The time of death was not observed for patients outside the red area, and for these patients the end of the line segments indicate the end of follow-up. Treatment means that sclerotherapy was given while no treatment means that no sclerotherapy was given.

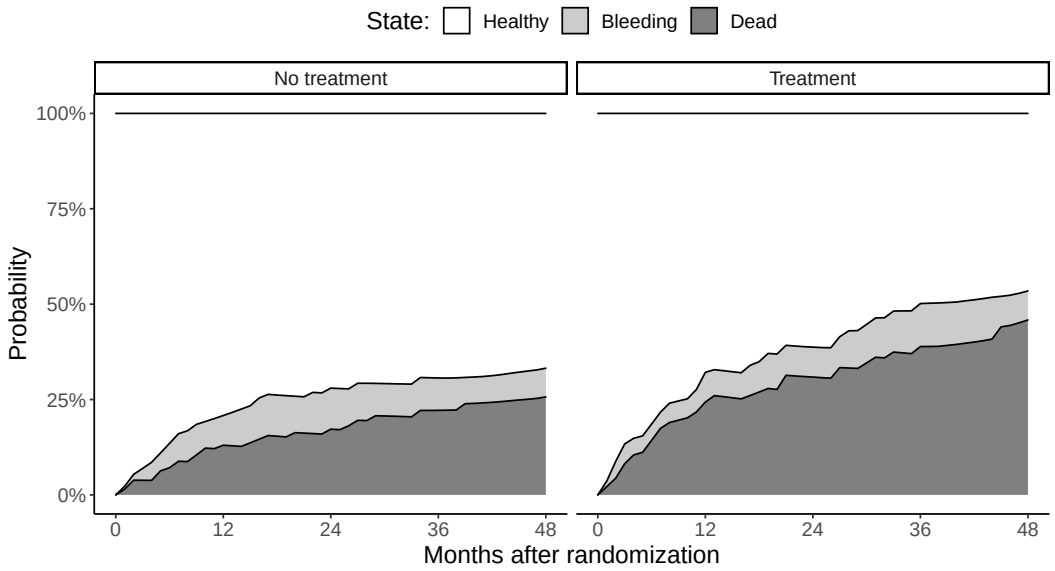


Figure 5: Estimated state occupation probabilities calculated from the PROVA trial data using the one-step estimators described in Section 4 and Appendix B.1. For each time point after randomization the dark-gray area indicates the probability of being dead, the light-gray area indicates the probability of being alive and having experienced variceal bleeding, and the white area indicates the probability of being alive and free of variceal bleeding. Treatment means that sclerotherapy was given while no treatment means that no sclerotherapy was given.

8 Discussion

In this paper we have derived a class of estimators of the state occupation probability in the irreversible illness-death model. We have established asymptotic normality of these estimators under a set of general conditions on the estimators of the transition hazards of the model, and we have derived a flexible data-adaptive penalized Poisson regression approach to estimate these hazard functions. Compared to inverse probability of censoring weighted (IPCW) estimators our proposed estimators do not rely on correctly specified (semi-)parametric models converging at the parametric rate $n^{-1/2}$. Instead, the validity of our proposed estimators relies on the convergence rate of the estimators of the transition hazards being $o_P(n^{-1/4})$. In particular, we can use non-parametric estimators of the hazard functions and data-adaptive model selection while still obtaining asymptotically valid confidence intervals for our parameter of interest, even when we do not know the exact asymptotic distribution of these hazard function estimators. Importantly, for an IPCW-based estimator to be valid we have to either make more restrictive assumptions about the censoring distribution than when using our class of estimators or use a suitably undersmoothed non-parametric estimator.

We have considered the setting where the covariates W was assumed to be measured only once at baseline. The censoring process was allowed to depend on W and on the whole history of the process X. We believe that this, together with the use of data-adaptive methods, will make the identifying assumption about the censoring mechanism plausible in many realistic settings. However, an even weaker assumption on the censoring mechanism is to allows it to also depend on a time-dependent covariate process. A natural extension of our work is thus to allow the transition hazards of the illness-death model to depend on other time-dependent processes than the process X itself. This has been considered for other models and in more generality, typically in discrete time [Murray and Tsiatis, 1996, Tsiatis,

2006, van der Laan and Robins, 2003]. Recently, Rytgaard et al. [2022] derived results for a class of targeted estimators for parameters of processes observed in continuous time. The illness-death model can be considered as a special case of their setting, and so these results will be useful to generalize our results to a setting with time-dependent covariates.

A Identifiability under CAR

In this section we prove Lemma 2.2 and Corollary 3.1.

Proof of Lemma 2.2. The lemma describes the intensity processes of the observed counting processes with respect to the filtration generated by the observed data. We show that the intensity process of $\tilde{N}_{01}(t)$ with respect to the filtration $\tilde{\mathcal{F}}_{t-}$ is given by $\tilde{Y}_0(t)h_{01}(t,W)$. When $\tilde{T}_0 < t$ then $\tilde{N}_{01}(\mathrm{d}t) = 0$, and when $\tilde{T}_0 \geq t$, i.e., $\tilde{Y}_0(t) = 1$, then the information in $\tilde{\mathcal{F}}_{t-}$ is equivalent to $\{\tilde{T}_0 \geq t, W\}$, hence

$$\mathbb{E}\left[\tilde{N}_{01}(\mathrm{d}t) \mid \tilde{\mathcal{F}}_{t-}\right] = \tilde{Y}_0(t)\mathbb{E}\left[\tilde{N}_{01}(\mathrm{d}t) \mid \tilde{T}_0 \geq t, W\right] = \frac{\tilde{Y}_0(t)}{\mathbb{E}(\tilde{Y}_0(t) \mid W)}\mathbb{E}\left[\tilde{N}_{01}(\mathrm{d}t) \,\middle|\, W\right]. \quad (17)$$

Similarly, we have

$$\mathbb{E}\left[\tilde{Y}_0(t)h_{01}(t,W)\,\mathrm{d}t \,\middle|\, \tilde{\mathcal{F}}_{t-}\right] = \frac{\tilde{Y}_0(t)}{\mathbb{E}[\tilde{Y}_0(t) \mid W]}\mathbb{E}\left[\tilde{Y}_0(t)h_{01}(t,W)\,\mathrm{d}t \,\middle|\, W\right]. \quad (18)$$

To prove the statement of Lemma 2.2 for $\tilde{N}_{01}$ we have to show equality of equations (17) and (18). Under CAR (Assumption 2.1) we have for all $z \geq t \geq s$ and w,

$$\mathbb{E}[1\{C < s\} \mid T = z, T_0 = t, W = w] = \int_0^s r(c \mid z, t, w)\,\mathrm{d}c = \int_0^s \tilde{r}(c \mid c, c, w)\,\mathrm{d}c,$$

which implies

$$\mathbb{E}[1\{C < s\} \mid T > t, T_0 = t, W = w] = \mathbb{E}[1\{C < s\} \mid T_0 \geq t, W = w]. \quad (19)$$

Using the equality in equation (19) we have

$$\begin{aligned}\mathbb{E}\left[\tilde{N}_{01}(\mathrm{d}t) \,\middle|\, W = w\right] &= \mathbb{E}[1\{C \geq t\} \mid T > t, T_0 = t, W = w]\mathbb{E}[Y_0(t) \mid W = w]h_{01}(t,w)\,\mathrm{d}t \\ &= \mathbb{E}[1\{C \geq t\} \mid T_0 \geq t, W = w]\mathbb{E}[Y_0(t) \mid W = w]h_{01}(t,w)\,\mathrm{d}t \\ &= \mathbb{E}[\tilde{Y}_0(t) \mid W = w]h_{01}(t,w)\,\mathrm{d}t.\end{aligned}$$

Substituting the previous display into equation (17) yields equality with equation (18).

For $\tilde{N}_{12}$ we should focus on the event $\tilde{T}_0 < t$, in which case we can treat $\tilde{T}_0$ as a baseline covariate and use the same arguments as above. The result for $\tilde{N}_C$ follows from the definition of the filtrations. □

Proof of Corollary 3.1. We need to show that the parameter $Q(T_0 \leq t, T > t)$ is identifiable from the observed data distribution P when we assume CAR and that Γ_0 and Γ_1 are uniformly bounded on the interval before time t. First we write

$$\begin{aligned}&Q(T_0 \leq t, T > t) \\ &= \int_{\mathcal{W}} Q(T_0 \leq t, T > t \mid W = w)\mu(\mathrm{d}w) \\ &= \int_{\mathcal{W}} \int_0^t Q(T > t \mid \eta = 1, T_0 = s, W = w)Q(T_0 \in \mathrm{d}s, \eta = 1 \mid W = w)\mu(\mathrm{d}w).\end{aligned}$$

The probability $Q(T > t \mid \eta = 1, T_0 = s, W = w)$ is a conditional survival function which we can write as

$$Q(T > t \mid \eta = 1, T_0 = s, W = w) = \exp\left(-\int_s^t h_{12}(z, s, w)\,\mathrm{d}z\right),$$

and by definition of a hazard we can write

$$Q(T_0 \in \mathrm{d}s, \eta = 1 \mid W = w) = \exp\left(-\int_0^s \{h_{01}(z, w) + h_{02}(z, w)\}\,\mathrm{d}z\right) H_{01}(\mathrm{d}s, w).$$

Combining the three previous displays we get

$$\begin{aligned}
&Q(T_0 \leq t, T > t) \\
&= \int_{\mathcal{W}} \int_0^t \exp\left(-\int_s^t h_{12}(z, s, w)\,\mathrm{d}z - \int_0^s \{h_{01}(z, w) + h_{02}(z, w)\}\,\mathrm{d}z\right) H_{01}(\mathrm{d}s, w)\mu(\mathrm{d}w),
\end{aligned}$$

which is equation (1) of Corollary 3.1. Let $H_{01|t}$ and $H_{02|t}$ denote, respectively, H_{01} and H_{02} restricted to the set $[0, t] \times \mathcal{W}$, and let $H_{12|t}$ denote H_{12} restricted to the set $\mathcal{U}_t \times \mathcal{W}$. We can then write

$$\begin{aligned}
&Q(T_0 \leq t, T > t) \\
&= \int_{\mathcal{W}} \int_0^t \exp\left(-\int_s^t h_{12}(z, s, w)\,\mathrm{d}z - \int_0^s \{h_{01}(z, w) + h_{02}(z, w)\}\,\mathrm{d}z\right) H_{01}(\mathrm{d}s, w)\mu(\mathrm{d}w) \\
&= \int_{\mathcal{W}} \int_0^t \exp\left(-H_{12}(t, s, w) - H_{01}(s, w) - H_{02}(s, w)\right) H_{01}(\mathrm{d}s, w)\mu(\mathrm{d}w) \\
&= \int_{\mathcal{W}} \int_0^t \exp\left(-H_{12|t}(t, s, w) - H_{01|t}(s, w) - H_{02|t}(s, w)\right) H_{01|t}(\mathrm{d}s, w)\mu(\mathrm{d}w),
\end{aligned}$$

and it now only remains to be argued that the parameters $H_{01|t}$, $H_{02|t}$, $H_{12|t}$, and μ can be identified from the observed data distribution. As the baseline covariates W are unaffected by the coarsening mechanism it is clear that μ is identifiable. Next, the assumption that Γ_0 is bounded by a finite constant on $[0, t]$ implies that

$$\int_0^s \frac{\tilde{N}_{0k}(\mathrm{d}u) - \tilde{Y}_0(u) H_{01|t}(\mathrm{d}u, W)}{P(\tilde{T}_0 \geq u \mid W)}$$

is well-defined for all $s \in [0, t]$. Lemma 2.2 implies that the above expression is a zero-mean martingale with respect to $\tilde{\mathcal{F}}_t$, in particular,

$$\begin{aligned}
0 &= \mathbb{E}_P\left[\int_0^s \frac{\tilde{N}_{0k}(\mathrm{d}u)}{P(\tilde{T}_0 \geq u \mid W)} - \int_0^s \frac{\tilde{Y}_0(u) H_{01|t}(\mathrm{d}u, W)}{P(\tilde{T}_0 \geq u \mid W)} \,\middle|\, \tilde{\mathcal{F}}_0\right] \\
&= \mathbb{E}_P\left[\int_0^s \frac{\tilde{N}_{0k}(\mathrm{d}u)}{P(\tilde{T}_0 \geq u \mid W)} - \int_0^s \frac{\tilde{Y}_0(u) H_{01|t}(\mathrm{d}u, W)}{P(\tilde{T}_0 \geq u \mid W)} \,\middle|\, W\right],
\end{aligned}$$

which gives

$$\begin{aligned}
\mathbb{E}_P\left[\int_0^s \frac{\tilde{N}_{0k}(\mathrm{d}u)}{P(\tilde{T}_0 \geq u \mid W)} \,\middle|\, W\right] &= \mathbb{E}_P\left[\int_0^s \frac{\tilde{Y}_0(u) H_{01|t}(\mathrm{d}u, W)}{P(\tilde{T}_0 \geq u \mid W)} \,\middle|\, W\right] \\
&= \int_0^s \frac{\mathbb{E}_P\left[\tilde{Y}_0(u) \,\middle|\, W\right] H_{01|t}(\mathrm{d}u, W)}{P(\tilde{T}_0 \geq u \mid W)} \\
&= \int_0^s H_{01|t}(\mathrm{d}u, W) \\
&= H_{01|t}(s, W),
\end{aligned}$$

for all $s \in [0, t]$. The left hand side involves only observable terms, and hence $H_{01|t}$ is identifiable. Similar arguments give that $H_{02|t}$ and $H_{12|t}$ are also identifiable. □

B Gâteaux derivative of the target parameter

In this section we prove Lemma 3.2. Let $P_\varepsilon = P + \varepsilon(\delta_O - P)$, where δ_O denotes the Dirac measure in $O = (\tilde{T}_0, \tilde{T}, \Delta_0, \Delta_1, W)$. For a function $f\colon \mathbb{R} \to \mathbb{R}$ that is differentiable at 0 we use the notation $\partial_\varepsilon f(\varepsilon) := f'(0)$. Define the operators $\kappa_{0l}\colon \mathcal{P} \to \mathcal{K}_0$,

$$\kappa_{0l}[P](t,w) := \int_0^t \frac{P(\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1, W = w)}{P(\tilde{T}_0 \geq u, W = w)}, \quad \text{for} \quad l \in \{1,2\}, \tag{20}$$

and $\kappa_{12}\colon \mathcal{P} \to \mathcal{K}_1$

$$\kappa_{12}[P](t,s,w) := \int_s^t \frac{P(\tilde{T}_0 \in \mathrm{d}u, \eta = 1, \Delta = 1, \tilde{T}_0 = s, W = w)}{P(\tilde{T}_0 \geq u, \eta = 1, \tilde{T}_0 = s, W = w)},$$

where $\mathcal{K}_0$ is the collection all cumulative transition hazard functions out of state 0, and $\mathcal{K}_1$ is the collection all cumulative transition hazard functions out of state 1. By Lemma 2.2, κ_{01}, κ_{02}, and κ_{12} identify the cumulative transition hazard H_{01}, H_{02}, and H_{12}, and thus we may write

$$\Psi_t(P_\varepsilon) = \int_{\mathcal{W}} \int_0^t \exp\{-\kappa_{01}[P_\varepsilon](s,w) - \kappa_{02}[P_\varepsilon](s,w) - \kappa_{12}[P_\varepsilon](t,s,w)\}\kappa_{01}[P_\varepsilon](\mathrm{d}s,w)P_\varepsilon(\mathrm{d}w)$$

Above we abuse notation slightly by letting $P(\mathrm{d}w)$ denote the marginal measure over $\mathcal{W}$. From this we derive

$$\begin{aligned}
&\partial_\varepsilon \Psi_t(P_\varepsilon) \\
&= \int_{\mathcal{W}} \int_0^t \exp\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\} \\
&\qquad \times \kappa_{01}[P](\mathrm{d}s,w)\{\partial_\varepsilon P_\varepsilon(\mathrm{d}w)\} \\
&\quad + \int_{\mathcal{W}} \int_0^t \exp\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\} \\
&\qquad \times \{\partial_\varepsilon \kappa_{01}[P_\varepsilon](\mathrm{d}s,w)\}P(\mathrm{d}w) \\
&\quad - \int_{\mathcal{W}} \int_0^t \{\partial_\varepsilon \kappa_{01}[P_\varepsilon](s,w) + \partial_\varepsilon \kappa_{02}[P_\varepsilon](s,w) + \partial_\varepsilon \kappa_{12}[P_\varepsilon](t,s,w)\} \\
&\qquad \times \exp\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\}\kappa_{01}[P](\mathrm{d}s,w)P(\mathrm{d}w).
\end{aligned} \tag{21}$$

As $\partial_\varepsilon P_\varepsilon(\mathrm{d}w) = (\delta_O - P)(\mathrm{d}w)$ and $\kappa_{01}(P) = H_{01}$, $\kappa_{02}(P) = H_{02}$, and $\kappa_{12}(P) = H_{12}$ we have that the first expression on the right hand side of equation (21) is equal to

$$\int_{\mathcal{W}} \rho(t,0,w;\nu)(\delta_0 - P)(\mathrm{d}w) = \rho(t,0,W;\nu) - \tilde{\Psi}_t(\nu).$$

To prove Lemma 3.2 it then only remains to show that

$$\begin{aligned}
&\varphi_t(O;\nu) \\
&= \int_{\mathcal{W}} \int_0^t \exp\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\}\{\partial_\varepsilon \kappa_{01}[P_\varepsilon](\mathrm{d}s,w)\} \\
&\qquad \times P(\mathrm{d}w) \\
&\quad - \int_{\mathcal{W}} \int_0^t \Big\{ (\partial_\varepsilon \kappa_{01}[P_\varepsilon](s,w) + \partial_\varepsilon \kappa_{02}[P_\varepsilon](s,w) + \partial_\varepsilon \kappa_{12}[P_\varepsilon](t,s,w)) \\
&\qquad \times \exp\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\}\kappa_{01}[P](\mathrm{d}s,w)\Big\}P(\mathrm{d}w),
\end{aligned} \tag{22}$$

where φ_t is defined in equation (4). Firstly we have for $l \in \{1,2\}$ that

$$
\begin{aligned}
&\partial_\varepsilon \kappa_{0l}[P_\varepsilon](s,w) \\
&= \partial_\varepsilon \int_0^s \frac{P_\varepsilon(\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1, W = w)}{P_\varepsilon(\tilde{T}_0 \geq u, W = w)} \\
&= \int_0^s \frac{[\delta_O - P](\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1, W = w)}{P(\tilde{T}_0 \geq u, W = w)} \\
&\quad - \frac{[\delta_O - P](\tilde{T}_0 \geq u, W = w) P(\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1, W = w)}{P(\tilde{T}_0 \geq u, W = w)^2} \\
&= \int_0^s \frac{\delta_O(\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1, W = w)}{P(\tilde{T}_0 \geq u, W = w)} \\
&\quad - \frac{\delta_O(\tilde{T}_0 \geq u, W = w) P(\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1, W = w)}{P(\tilde{T}_0 \geq u, W = w)^2} \\
&= \delta_W(w) \int_0^s \frac{1}{P(\tilde{T}_0 \geq u, W = w)} \Bigg\{ \mathbb{1}(\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1) \\
&\qquad\qquad - \frac{\mathbb{1}(\tilde{T}_0 \geq u) P(\tilde{T}_0 \in \mathrm{d}u, \eta = l, \Delta_0 = 1, W = w)}{P(\tilde{T}_0 \geq u, W = w)} \Bigg\} \\
&= \delta_W(w) \int_0^s \frac{1}{P(\tilde{T}_0 \geq u, W = w)} \left\{ \tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u) H_{0l}(\mathrm{d}u, w) \right\}.
\end{aligned}
\tag{23}
$$

This gives

$$
\begin{aligned}
&\int_{\mathcal{W}} \int_0^t \exp\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\} \\
&\qquad \times \{\partial_\varepsilon \kappa_{01}[P_\varepsilon](\mathrm{d}s, w)\} P(\mathrm{d}w) \\
&= \int_0^t \exp\{-\kappa_{01}[P](s,W) - \kappa_{02}[P](s,W) - \kappa_{12}[P](t,s,W)\} \\
&\qquad \times \int_0^{\mathrm{d}s} \frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u) H_{01}(\mathrm{d}u, W)}{P(\tilde{T}_0 \geq u \mid W = W)} \\
&= \int_0^t \frac{\exp\{-\kappa_{01}[P](u,W) - \kappa_{02}[P](u,W) - \kappa_{12}[P](t,u,W)\}}{\exp\{-H_{01}(u,W) - H_{02}(u,W) - \Gamma_0(u,W)\}} \\
&\qquad \times \left[\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u) H_{01}(\mathrm{d}u, W) \right] \\
&= \int_0^t \frac{\exp\{-\kappa_{12}[P](t,u,W)\}}{\exp\{-\Gamma_0(u,W)\}} \left[\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u) H_{01}(\mathrm{d}u, W) \right] \\
&= \int_0^t \exp\{-H_{12}(t,u,W)\} \frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u) H_{01}(\mathrm{d}u, W)}{\exp\{-\Gamma_0(u,W)\}}.
\end{aligned}
\tag{24}
$$

From equation (23) it also follows that

$$
\begin{aligned}
&\int_{\mathcal{W}} \int_0^t \partial_\varepsilon \kappa_{0l}[P_\varepsilon](s,w) \exp\left\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\right\} \\
&\qquad \times \kappa_{01}[P](\mathrm{d}s,w) P(\mathrm{d}w) \\
&= \int_0^t \int_0^s \left(\frac{\left[\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u) H_{0l}(\mathrm{d}u,W)\right]}{\exp\left\{-H_{01}(u,W) - H_{02}(u,W) - \Gamma_0(u,W)\right\}} \right) \\
&\qquad \times e^{-H_{01}(s,W) - H_{02}(s,W) - H_{12}(t,s,W)} H_{01}(\mathrm{d}s,w) \\
&= \int_0^t \int_u^t e^{-H_{01}(s,W) - H_{02}(s,W) - H_{12}(t,s,W)} H_{01}(\mathrm{d}s,w) \\
&\qquad \times \left(\frac{\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u) H_{0l}(\mathrm{d}u,W)}{\exp\left\{-H_{01}(u,W) - H_{02}(u,W) - \Gamma_0(u,W)\right\}} \right) \\
&= \int_0^t \rho(t,u,W;\nu) \frac{\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u) H_{0l}(\mathrm{d}u,W)}{\exp\left\{-\Gamma_0(u,W)\right\}},
\end{aligned}
\tag{25}
$$

where we used Fubini's theorem for the second equality, and for the third equality we used the definition of ρ from equation (2) and that $\int_u^s h_{0l}(z)\,\mathrm{d}z = H_{0l}(s) - H_{0l}(u)$. Next, a

calculation similar to the one in equation (23) gives

$$
\begin{aligned}
&\partial_\varepsilon \kappa_{12}[P_\varepsilon](t,s,w) \\
&\quad = \partial_\varepsilon \int_s^t \frac{P_\varepsilon(\tilde{T} \in \mathrm{d}u, \Delta = 1, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)}{P_\varepsilon(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)} \\
&\quad = \int_s^t \frac{[\delta_O - P](\tilde{T} \in \mathrm{d}u, \Delta = 1, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)} \\
&\qquad - \left(\frac{[\delta_O - P](\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)^2} \right. \\
&\qquad\qquad \left. \times P(\tilde{T} \in \mathrm{d}u, \Delta = 1, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w) \right) \\
&\quad = \int_s^t \frac{\delta_O(\tilde{T} \in \mathrm{d}u, \Delta = 1, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)} \\
&\qquad - \left(\frac{\delta_O(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)^2} \right. \\
&\qquad\qquad \left. \times P(\tilde{T} \in \mathrm{d}u, \Delta = 1, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w) \right) \\
&\quad = \delta_{W,\tilde{T}_0}(w,s) \int_s^t \frac{1}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)} \\
&\qquad \times \left\{ \mathbb{1}(\tilde{T} \in \mathrm{d}u, \Delta = 1, \Delta_0 = 1, \eta = 1) \right. \\
&\qquad\qquad - \left(\frac{\mathbb{1}(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1)}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)} \right. \\
&\qquad\qquad\qquad \left. \left. \times P(\tilde{T} \in \mathrm{d}u, \Delta = 1, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w) \right) \right\} \\
&= \delta_{W,\tilde{T}_0}(w,s) \int_s^t \frac{\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u) H_{12}(\mathrm{d}u, s, w)}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)},
\end{aligned}
$$

and so we have that

$$
\begin{aligned}
&\int_{\mathcal{W}} \int_0^t \partial_\varepsilon \kappa_{12}[P_\varepsilon](t,s,w) \exp\left\{-\kappa_{01}[P](s,w) - \kappa_{02}[P](s,w) - \kappa_{12}[P](t,s,w)\right\} \\
&\qquad \times \kappa_{01}[P](\mathrm{d}s,w) P(\mathrm{d}w) \\
&= \int_{\mathcal{W}} \int_0^t \delta_{W,\tilde{T}_0}(w,s) \\
&\qquad \times \int_s^t \frac{[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u) H_{12}(\mathrm{d}u,s,w)] e^{-H_{01}(s,w)-H_{02}(s,w)-H_{12}(t,s,w)}}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s, W = w)} \\
&\qquad \times H_{01}(\mathrm{d}s,w) P(\mathrm{d}w) \\
&= \int_0^t \delta_{\tilde{T}_0}(s) \int_s^t \frac{[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u) H_{12}(\mathrm{d}u,s,W)] e^{-H_{01}(s,W)-H_{02}(s,W)-H_{12}(t,s,W)}}{P(\tilde{T} \geq u, \Delta_0 = 1, \eta = 1, \tilde{T}_0 = s \mid W = W)} \\
&\qquad \times H_{01}(\mathrm{d}s,W) \qquad (26) \\
&= \int_0^t \delta_{\tilde{T}_0}(s) \int_s^t \frac{[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u) H_{12}(\mathrm{d}u,s,W)] e^{-H_{01}(s,W)-H_{02}(s,W)-H_{12}(t,s,W)}}{e^{-H_{01}(s,W)-H_{02}(s,W)-\Gamma_0(s,W)} h_{01}(s,W) e^{-H_{12}(u,s,W)-\Gamma_1(u,s,W)}} \\
&\qquad \times H_{01}(\mathrm{d}s,W) \\
&= \int_0^t \delta_{\tilde{T}_0}(s) \int_s^t \frac{[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u) H_{12}(\mathrm{d}u,s,W)] e^{-[H_{12}(t,s,W)-H_{12}(u,s,W)]}}{e^{-\Gamma_0(s,W)} h_{01}(s,W) e^{-\Gamma_1(u,s,W)}} \\
&\qquad \times H_{01}(\mathrm{d}s,W) \\
&= \int_{\tilde{T}_0}^t \frac{[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u) H_{12}(\mathrm{d}u,\tilde{T}_0,W)] e^{-[H_{12}(t,\tilde{T}_0,W)-H_{12}(u,\tilde{T}_0,W)]}}{e^{-\Gamma_0(\tilde{T}_0,W)} e^{-\Gamma_1(u,\tilde{T}_0,W)}} \\
&= e^{\Gamma_0(\tilde{T}_0,W)} \int_{\tilde{T}_0}^t e^{-[H_{12}(t,\tilde{T}_0,W)-H_{12}(u,\tilde{T}_0,W)]} \frac{\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u) H_{12}(\mathrm{d}u,\tilde{T}_0,W)}{e^{-\Gamma_1(u,\tilde{T}_0,W)}}.
\end{aligned}
$$

Combining equations (24), (25), and (26) gives equation (22).

B.1 Gâteaux derivative of the probability of being event-free

The probability of being event-free can be analyzed as a standard survival problem where the adverse event is illness or death. This is a well-studied problem for which the canonical gradient is well-known [e.g., Robins and Rotnitzky, 1992, van der Laan and Robins, 2003, van der Laan and Rose, 2011, Rytgaard et al., 2021]. In this section we briefly show how the calculations above can be used to construct a debiased estimator of this parameter. Using Lemma 2.2 and the definition of κ_{0l}, $l \in \{1,2\}$ (see equation (20)), we can show that the probability of being event free can be expressed as

$$
\Upsilon_t(P) = \int_{\mathcal{W}} \exp\left\{-\kappa_{01}[P](t,w) - \kappa_{02}[P](t,w)\right\} P(\mathrm{d}w).
$$

Below we calculate the Gâteaux derivative of this parameter. With this calculation in hand we can use the same arguments as given in Section 4 to define a debiased estimator of Υ_t. Finally, similar arguments as given in Appendix B below can then be used to establish a result akin to Theorem 4.2 for this parameter.

To calculate the Gâteaux derivative of Υ_t at δ_O we first note that

$$\begin{aligned}
&\partial_\varepsilon \Upsilon_t(P_\varepsilon) \\
&= \exp\{-H_{01}(t,W) - H_{02}(t,W)\} - \int_{\mathcal{W}} \exp\{-H_{01}(t,w) - H_{02}(t,w)\} P(\mathrm{d}w) \\
&\quad - \int_{\mathcal{W}} \partial_\varepsilon \{\kappa_{01}[P_\varepsilon](t,w) + \kappa_{02}[P_\varepsilon](t,w)\} \exp\{-H_{01}(t,w) - H_{02}(t,w)\} P(\mathrm{d}w) \\
&= \exp\{-H_{01}(t,W) - H_{02}(t,W)\} \left(1 - \int_{\mathcal{W}} \partial_\varepsilon \{\kappa_{01}[P_\varepsilon](t,w) + \kappa_{02}[P_\varepsilon](t,w)\} P(\mathrm{d}w)\right) \\
&\quad - \int_{\mathcal{W}} \exp\{-H_{01}(t,w) - H_{02}(t,w)\} P(\mathrm{d}w).
\end{aligned}$$

We can then use equation (23) to write

$$\int_{\mathcal{W}} \partial_\varepsilon \{\kappa_{0l}[P_\varepsilon](t,w)\} P(\mathrm{d}w) = \int_0^t \frac{\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u) H_{0l}(\mathrm{d}u, W)}{\exp\{-H_{01}(u,w) - H_{02}(u,w) - \Gamma_0(u,W)\}},$$

for $l \in \{1,2\}$, and hence the Gâteaux derivative of Υ_t is

$$\begin{aligned}
&\exp\{-H_{01}(t,W) - H_{02}(t,W)\} \\
&\qquad \times \left(1 - \int_0^t \sum_{l=1}^{2} \frac{\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u) H_{0l}(\mathrm{d}u, W)}{\exp\{-H_{01}(u,w) - H_{02}(u,w) - \Gamma_0(u,W)\}}\right) - \Upsilon_t(P)
\end{aligned}$$

C Bounding the remainder term

In this section we show that Assumptions 4.1 (i) and (ii) imply that the remainder term defined in equation (10) is $o_P(n^{-1/2})$. Firstly, we define for notational convenience

$$\begin{aligned}
\hat{\varepsilon}_{0k} &= \|\hat{h}_{0k} - h_{0k}\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{W})}, \quad k \in \{1,2\}, \\
\hat{\varepsilon}_{0C} &= \|\hat{\gamma}_0 - \gamma_0\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{W})}, \\
\hat{\varepsilon}_{12} &= \|\hat{h}_{12} - h_{12}\|_{\mathcal{L}^2_{m\otimes\mu}(\mathcal{U}_t\times\mathcal{W})}, \\
\hat{\varepsilon}_{1C} &= \|\hat{\gamma}_1 - \gamma_1\|_{\mathcal{L}^2_{m\otimes\mu}(\mathcal{U}_t\times\mathcal{W})},
\end{aligned} \tag{27}$$

where m denotes Lebesgue measure. Recall also that the remainder is

$$\begin{aligned}
&\mathrm{Rem}(P, \hat{\nu}) \\
&= \mathbb{P}_n[\rho(t,0,\cdot;\hat{\nu})] - P[\rho(t,0,\cdot;\nu)] + P[\mathrm{IF}_t(\cdot;\hat{\nu})] \\
&= \mathbb{P}_n[\rho(t,0,\cdot;\hat{\nu})] - P[\rho(t,0,\cdot;\nu)] + P[\rho(t,0,\cdot;\hat{\nu})] \\
&\quad + P[\varphi(t,0,\cdot;\hat{\nu})] - P[\mathbb{P}_n[\rho(t,0,\cdot;\hat{\nu})]] \\
&= P[\rho(t,0,\cdot;\hat{\nu})] - P[\rho(t,0,\cdot;\nu)] + P[\varphi(t,0,\cdot;\hat{\nu})] \\
&= \mathbb{E}[\rho(t,0,W;\hat{\nu})] - \mathbb{E}[\rho(t,0,W;\nu)] \\
&\quad + \mathbb{E}\left[\int_0^t e^{-\hat{H}_{12}(t,u,W)} \frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{01}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right] \\
&\quad - \mathbb{E}\left[\int_0^t \rho(t,u,W;\hat{\nu}) \frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{01}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right] \\
&\quad - \mathbb{E}\left[\int_0^t \rho(t,u,W;\hat{\nu}) \frac{\tilde{N}_{02}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{02}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right] \\
&\quad - \mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T},W)} \int_0^t e^{-[\hat{H}_{12}(t,\tilde{T}_0,W) - \hat{H}_{12}(u,\tilde{T}_0,W)]} \frac{\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{H}_{12}(\mathrm{d}u,\tilde{T}_0,W)}{e^{-\hat{\Gamma}_1(u,\tilde{T}_0,W)}}\right],
\end{aligned} \tag{28}$$

where we here and in the following use $\mathbb{E}$ to denote expectation under P with respect to a sample $O \sim P$ when the estimators $\hat{\nu}$ are considered fixed, i.e., formally we here use $\mathbb{E}$ to denote the conditional expectation given $\{O_i\}_{i=1,\ldots,n}$ when $O \perp\!\!\!\perp \{O_i\}_{i=1,\ldots,n}$. We have the following result.

Proposition C.1. *If Assumption 4.1 (i) holds, the estimators $\hat{H}_{01}$, $\hat{H}_{02}$, $\hat{H}_{12}$, $\hat{\Gamma}_0$, and $\hat{\Gamma}_1$ are uniformly bounded by a fixed constant with probability tending to 1, equation (7) holds for these estimators, and all terms in equation (27) are $o_P(1)$, then*

$$\mathrm{Rem}(P, \hat{\nu}) = O_P\big\{(\hat{\varepsilon}_{C1} + \hat{\varepsilon}_{C0} + \hat{\varepsilon}_{12})\,\hat{\varepsilon}_{12} + (\hat{\varepsilon}_{01} + \hat{\varepsilon}_{02})\,(\hat{\varepsilon}_{01} + \hat{\varepsilon}_{02} + \hat{\varepsilon}_{12}) + \hat{\varepsilon}_{01}^2 + \hat{\varepsilon}_{02}^2 + \hat{\varepsilon}_{12}^2\big\}.$$

In particular, Assumptions 4.1 (i) and (ii) imply that $\mathrm{Rem}(P, \hat{\nu}) = o_P(n^{-1/2})$.

We divide the proof of Proposition C.1 into three lemmas, Lemmas C.2, C.3, and C.4 stated below. To ease the notation we will for a function $f\colon [0,t] \times \mathcal{W} \to \mathbb{R}$ write $\|f\|$ to mean $\|f\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{W})}$, and for a function $f\colon \mathcal{U}_t \times \mathcal{W} \to \mathbb{R}$ we write $\|f\|$ to mean $\|f\|_{\mathcal{L}^2_{m\otimes\mu}(\mathcal{U}_t\times\mathcal{W})}$. When we write $f_n \xrightarrow{\mathrm{P}} f$ we mean convergence in probability with respect to this norm, i.e., $\|f_n - f\| \xrightarrow{\mathrm{P}} 0$.

Lemma C.2. *Let μ be a σ-finite measure on $\mathcal{Z}$, $h\colon [0,t] \times \mathcal{Z} \to \mathbb{R}$ a measurable function, and define $H\colon [0,t] \times \mathcal{Z} \to \mathbb{R}$ as*

$$H(s,z) := \int_0^s h(u,z)\,\mathrm{d}u.$$

Then $\|H\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})} = O(\|h\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})})$.

Proof. This follows from Jensen's inequality:

$$\begin{aligned}
\|H\|^2_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})} &= \int_{\mathcal{Z}}\int_0^t H(s,z)^2 \,\mathrm{d}s\, \mu(\mathrm{d}z) \\
&= \int_{\mathcal{Z}}\int_0^t \left(s\int_0^s h(u,z)\frac{\mathrm{d}u}{s}\right)^2 \mathrm{d}s\, \mu(\mathrm{d}z) \\
&= \int_{\mathcal{Z}}\int_0^t s^2\left(\int_0^s h(u,z)\frac{\mathrm{d}u}{s}\right)^2 \mathrm{d}s\, \mu(\mathrm{d}z) \\
&\leq \int_{\mathcal{Z}}\int_0^t s^2\int_0^s h(u,z)^2\frac{\mathrm{d}u}{s}\,\mathrm{d}s\, \mu(\mathrm{d}z) \\
&= \int_{\mathcal{Z}}\int_0^t s\int_0^s h(u,z)^2\,\mathrm{d}u\,\mathrm{d}s\, \mu(\mathrm{d}z) \\
&\leq t^2\int_{\mathcal{Z}}\int_0^t h(u,z)^2\,\mathrm{d}u\, \mu(\mathrm{d}z) = t^2\|h\|^2_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})}.
\end{aligned}$$

□

Applying Lemma C.2 on $\hat{h}_{01} - h_{01}$ gives that $\|\hat{H}_{01} - H_{01}\| = O_P(\hat{\varepsilon}_{01})$, and likewise for H_{02}, H_{12}, Γ_0, and Γ_1.

Lemma C.3. *(a) For a measurable function $f\colon [0,t]\times\mathcal{W}\to\mathbb{R}$ it holds that*

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t f(u,W)[\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{h}_{0l}(u,W)\,\mathrm{d}u]\right] \\
&\quad= \mathbb{E}\left[\int_0^t f(u,W)e^{-H_{01}(u,W)-H_{02}(u,W)-\Gamma_0(u,W)}[h_{0l}(u,W) - \hat{h}_{0l}(u,W)]\,\mathrm{d}u\right],
\end{aligned}$$

for $l\in\{1,2\}$.

(b) For a measurable function $f\colon \mathcal{U}_t\times\mathcal{W}\to\mathbb{R}$ it holds that

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t f(u,\tilde{T}_0,W)[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{h}_{12}(u,\tilde{T}_0,W)\,\mathrm{d}u]\right] \\
&\quad= \mathbb{E}\Bigg[\mathbb{1}\{\Delta_0=1,\eta=1,\tilde{T}_0<t\} \\
&\qquad\times\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)e^{-H_{12}(u,\tilde{T}_0,W)-\Gamma_1(u,\tilde{T}_0,W)}[h_{12}(u,\tilde{T}_0,W) - \hat{h}_{12}(u,\tilde{T}_0,W)]\,\mathrm{d}u\Bigg].
\end{aligned}$$

Proof. For statement (a) we have by definition of the counting process $\tilde{N}_{0l}$ that

$$\int_0^t f(u,W)\tilde{N}_{0l}(\mathrm{d}u) = \Delta_0\mathbb{1}\{\eta=l\}\mathbb{1}\{\tilde{T}<t\}f(\tilde{T},W),$$

and so

$$\begin{aligned}
\mathbb{E}\left[\int_0^t f(u,W)\tilde{N}_{0l}(\mathrm{d}u)\,\middle|\,W\right] &= \mathbb{E}\left[\Delta_0\mathbb{1}\{\eta=l\}\mathbb{1}\{\tilde{T}<t\}f(\tilde{T},W)\,\middle|\,W\right] \\
&= \int_0^t f(u,W)e^{-H_{01}(u,W)-H_{02}(u,W)-\Gamma_0(u,W)}H_{0l}(\mathrm{d}u,W).
\end{aligned}$$

Thus the tower property gives that

$$\mathbb{E}\left[\int_0^t f(u,W)\tilde{N}_{0l}(\mathrm{d}u)\right] = \mathbb{E}\left[\int_0^t f(u,W)e^{-H_{01}(u,W)-H_{02}(u,W)-\Gamma_0(u,W)}H_{0l}(\mathrm{d}u,W)\right]. \quad (29)$$

Next, Fubini's theorem and the tower property gives that

$$\begin{aligned}
&E\left[\int_0^t f(u,W)\tilde{Y}_0(u)\hat{h}_{0l}(u,W)\,\mathrm{d}u\right] \\
&= \int_0^t E\left[f(u,W)\tilde{Y}_0(u)\hat{h}_{0l}(u,W)\right]\mathrm{d}u \\
&= \int_0^t \mathbb{E}\left[f(u,W)\hat{h}_{0l}(u,W)E\Big[\tilde{Y}_0(u)\;\Big|\;W\Big]\right]\mathrm{d}u \\
&= \int_0^t \mathbb{E}\left[f(u,W)\hat{h}_{0l}(u,W)e^{-H_{01}(u,W)-H_{02}(u,W)-\Gamma_0(u,W)}\right]\mathrm{d}u \\
&= \mathbb{E}\left[\int_0^t f(u,W)\hat{h}_{0l}(u,W)e^{-H_{01}(u,W)-H_{02}(u,W)-\Gamma_0(u,W)}\,\mathrm{d}u\right].
\end{aligned} \tag{30}$$

Combining equations (29) and (30) then gives the result.

For statement (b), let $\tilde{\eta}_t := \mathbb{1}\{\tilde{T}_0 < t, \Delta_0 = 1, \Delta_0\eta = 1\}$ and note that $\tilde{\eta}_t = \mathbb{1}\{\tilde{T}_0 < t, \Delta_0 = 1, \eta = 1\}$. We can write

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t f(u,\tilde{T}_0,W)[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{h}_{12}(u,\tilde{T}_0,W)\,\mathrm{d}u]\;\Big|\;\tilde{\eta}_t\right] \\
&\quad= \tilde{\eta}_t\mathbb{E}\left[\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{h}_{12}(u,\tilde{T}_0,W)\,\mathrm{d}u]\;\Big|\;\tilde{\eta}_t = 1\right], \\
&\quad= \tilde{\eta}_t\mathbb{E}\Bigg[\mathbb{E}\left[\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{h}_{12}(u,\tilde{T}_0,W)\,\mathrm{d}u]\;\Big|\;\tilde{T}_0,\tilde{\eta}_t = 1\right]\;\Big|\;\tilde{\eta}_t = 1\Bigg]
\end{aligned}$$

and then we can use the same arguments as we used to prove statement (a) to derive

$$\begin{aligned}
&\mathbb{E}\left[\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{h}_{12}(u,\tilde{T}_0,W)\,\mathrm{d}u]\;\Big|\;\tilde{T}_0,\tilde{\eta}_t = 1\right] \\
&\quad= \mathbb{E}\Bigg[\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)e^{-H_{12}(u,\tilde{T}_0,W)-\Gamma_1(u,\tilde{T}_0,W)} \\
&\qquad\qquad\times[h_{12}(u,\tilde{T}_0,W) - \hat{h}_{12}(u,\tilde{T}_0,W)]\,\mathrm{d}u\;\Big|\;\tilde{T}_0,\tilde{\eta}_t = 1\Bigg].
\end{aligned}$$

The tower property then gives

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t f(u,\tilde{T}_0,W)[\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{h}_{12}(u,\tilde{T}_0,W)\,\mathrm{d}u]\right] \\
&\quad= \mathbb{E}\Bigg[\tilde{\eta}_t\mathbb{E}\Bigg[\mathbb{E}\Bigg[\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)e^{-H_{12}(u,\tilde{T}_0,W)-\Gamma_1(u,\tilde{T}_0,W)} \\
&\qquad\qquad\times[h_{12}(u,\tilde{T}_0,W) - \hat{h}_{12}(u,\tilde{T}_0,W)]\,\mathrm{d}u\;\Big|\;\tilde{T}_0,\tilde{\eta}_t = 1\Bigg]\;\Big|\;\tilde{\eta}_t = 1\Bigg]\Bigg] \\
&\quad= \mathbb{E}\Bigg[\tilde{\eta}_t\mathbb{E}\Bigg[\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)e^{-H_{12}(u,\tilde{T}_0,W)-\Gamma_1(u,\tilde{T}_0,W)} \\
&\qquad\qquad\times[h_{12}(u,\tilde{T}_0,W) - \hat{h}_{12}(u,\tilde{T}_0,W)]\,\mathrm{d}u\;\Big|\;\tilde{\eta}_t\Bigg]\Bigg] \\
&\quad= \mathbb{E}\Bigg[\tilde{\eta}_t\int_{\tilde{T}_0}^t f(u,\tilde{T}_0,W)e^{-H_{12}(u,\tilde{T}_0,W)-\Gamma_1(u,\tilde{T}_0,W)}[h_{12}(u,\tilde{T}_0,W) - \hat{h}_{12}(u,\tilde{T}_0,W)]\,\mathrm{d}u\Bigg],
\end{aligned}$$

which by definition of $\tilde{\eta}_t$ is the wanted result. □

Lemma C.4. *Let μ be a measure on $\mathcal{Z}$, and let $\{f_n\}_{n\in\mathbb{N}}$ and $\{h_n\}_{n\in\mathbb{N}}$ be sequences of measurable real-valued random functions with domain $[0,t]\times\mathcal{Z}$ such that $\|h_n\|_\infty = O_P(1)$. Then*

$$\begin{aligned}&\int_{\mathcal{Z}}\int_0^t e^{h_n(u,z)} f_n(u,z)\,\mathrm{d}u\,\mu(\mathrm{d}z)\\ &= \int_{\mathcal{Z}}\int_0^t f_n(u,z)\,\mathrm{d}u\,\mu(\mathrm{d}z) + O_P\left\{\|h_n\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})}\cdot\|f_n\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})}\right\}.\end{aligned}$$

Proof. Using that $e^h \le 1 + he^h$ for all $h\in\mathbb{R}$ we can write

$$\begin{aligned}&\int_{\mathcal{Z}}\int_0^t e^{h_n(u,z)} f_n(u,z)\,\mathrm{d}u\,\mu(\mathrm{d}z) - \int_{\mathcal{Z}}\int_0^t f_n(u,z)\,\mathrm{d}u\,\mu(\mathrm{d}z)\\ &\le \int_{\mathcal{Z}}\int_0^t h_n(u,z)e^{h_n(u,z)} f_n(u,z)\,\mathrm{d}u\,\mu(\mathrm{d}z).\end{aligned}$$

Two applications of Hölder's inequality give that

$$\begin{aligned}\left|\int_{\mathcal{Z}}\int_0^t h_n(u,z)e^{h_n(u,z)} f_n(u,z)\,\mathrm{d}u\,\mu(\mathrm{d}z)\right| &= \|h_n e^{h_n} f_n\|_{\mathcal{L}^1_{m\otimes\mu}([0,t]\times\mathcal{Z})}\\ &\le \|e^{h_n}\|_\infty \|h_n f_n\|_{\mathcal{L}^1_{m\otimes\mu}([0,t]\times\mathcal{Z})}\\ &\le \|e^{h_n}\|_\infty \|h_n\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})}\|f_n\|_{\mathcal{L}^2_{m\otimes\mu}([0,t]\times\mathcal{Z})}.\end{aligned}$$

As $\|h_n\|_\infty$ is bounded in probability so is $\|e^{h_n}\|_\infty$, and from this the result follows. □

We are now ready to prove Proposition C.1

Proof of Proposition C.1. By Lemma C.3 (a) we have that

$$\begin{aligned}&\mathbb{E}\left[\int_0^t e^{-\hat{H}_{12}(t,u,W)}\frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{01}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right]\\ &= \mathbb{E}\left[\int_0^t e^{-\hat{H}_{01}(u,W)-\hat{H}_{02}(u,W)-\hat{H}_{12}(t,u,W)}\left[H_{01}(\mathrm{d}u,W) - \hat{H}_{01}(\mathrm{d}u,W)\right]\right]\end{aligned}$$

and hence by the definition of ρ (see equation (2)) the first three terms of the right hand side of equation (28) are equal to

$$\begin{aligned}&\mathbb{E}[\rho(t,0,W;\hat{\nu})] - \mathbb{E}[\rho(t,0,W;\nu)] + \mathbb{E}\left[\int_0^t e^{-\hat{H}_{12}(t,u,W)}\frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{01}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right]\\ &= \mathbb{E}\left[\int_0^t e^{-\hat{H}_{01}(u,W)-\hat{H}_{02}(u,W)-\hat{H}_{12}(t,u,W)}H_{01}(\mathrm{d}u,W)\right]\\ &\quad - \mathbb{E}\left[\int_0^t e^{-H_{01}(u,W)-H_{02}(u,W)-H_{12}(t,u,W)}H_{01}(\mathrm{d}u,W)\right] \qquad (31)\\ &= \mathbb{E}\left[\int_0^t \left(e^{[H_{01}-\hat{H}_{01}](u,W)+[H_{01}-\hat{H}_{02}](u,W)+[H_{12}-\hat{H}_{12}](t,u,W)} - 1\right)\right.\\ &\qquad \left.\times\, e^{-H_{01}(u,W)-H_{02}(u,W)-H_{12}(t,u,W)}H_{01}(\mathrm{d}u,W)\right]\end{aligned}$$

Next, using Lemma C.3 (a) and Fubini's theorem we have that

$$
\begin{aligned}
&\mathbb{E}\left[\int_0^t \rho(t,u,W;\hat{\nu})\frac{\tilde{N}_{0l}(\mathrm{d}u)-\tilde{Y}_0(u)\hat{H}_{0l}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right]\\
&=\mathbb{E}\left[\int_0^t \rho(t,u,W;\hat{\nu})e^{-\hat{H}_{01}(u,W)-\hat{H}_{02}(u,W)}[H_{0l}-\hat{H}_{0l}](\mathrm{d}u,W)\right]\\
&=\mathbb{E}\left[\int_0^t\int_u^t \exp\left(-\int_u^s\{\hat{h}_{01}(z,W)+\hat{h}_{02}(z,W)\}\,\mathrm{d}z-\int_s^t \hat{h}_{12}(z,s,W)\,\mathrm{d}z\right)\right.\\
&\qquad\left.\times\hat{H}_{01}(\mathrm{d}s,W)e^{-\hat{H}_{01}(u,W)-\hat{H}_{02}(u,W)}[H_{0l}-\hat{H}_{0l}](\mathrm{d}u,W)\right]\\
&=\mathbb{E}\left[\int_0^t\int_u^t e^{-\hat{H}_{01}(s,W)-\hat{H}_{02}(s,W)-\hat{H}_{12}(t,s,W)}\hat{H}_{01}(\mathrm{d}s,W)[H_{0l}-\hat{H}_{0l}](\mathrm{d}u,W)\right]\\
&=\mathbb{E}\left[\int_0^t\int_0^s [H_{0l}-\hat{H}_{0l}](\mathrm{d}u,W)e^{-\hat{H}_{01}(s,W)-\hat{H}_{02}(s,W)-\hat{H}_{12}(t,s,W)}\hat{H}_{01}(\mathrm{d}s,W)\right]\\
&=\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-\hat{H}_{01}(s,W)-\hat{H}_{02}(s,W)-\hat{H}_{12}(t,s,W)}\hat{H}_{01}(\mathrm{d}s,W)\right].
\end{aligned}
$$

Using Cauchy-Schwarz' inequality this implies that

$$
\begin{aligned}
&\mathbb{E}\left[\int_0^t \rho(t,u,W;\hat{\nu})\frac{\tilde{N}_{0l}(\mathrm{d}u)-\tilde{Y}_0(u)\hat{H}_{0l}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right]\\
&=\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-\hat{H}_{01}(s,W)-\hat{H}_{02}(s,W)-\hat{H}_{12}(t,s,W)}h_{01}(s,W)\,\mathrm{d}s\right]\\
&\quad+\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-\hat{H}_{01}(s,W)-\hat{H}_{02}(s,W)-\hat{H}_{12}(t,s,W)}[\hat{h}_{01}-h_{01}](s,W)\,\mathrm{d}s\right] \qquad (32)\\
&=\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-\hat{H}_{01}(s,W)-\hat{H}_{02}(s,W)-\hat{H}_{12}(t,s,W)}h_{01}(s,W)\,\mathrm{d}s\right]\\
&\quad+O_P\left\{\hat{\varepsilon}_{0l}\cdot\hat{\varepsilon}_{01},\right\},
\end{aligned}
$$

where we used Lemma C.2 and that $\hat{H}_{01}$, $\hat{H}_{02}$, $\hat{H}_{12}$ are uniformly bounded with probability tending to one for the last equality. Using Lemma C.4 we have

$$
\begin{aligned}
&\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-\hat{H}_{01}(s,W)-\hat{H}_{02}(s,W)-\hat{H}_{12}(t,s,W)}h_{01}(s,W)\,\mathrm{d}s\right]\\
&=\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{[H_{01}-\hat{H}_{01}](s,W)+[H_{02}-\hat{H}_{02}](s,W)+[H_{12}-\hat{H}_{12}](t,s,W)}\right.\\
&\qquad\left.\times e^{-H_{01}(s,W)-H_{02}(s,W)-H_{12}(t,s,W)}h_{01}(s,W)\,\mathrm{d}s\right]\\
&=\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-H_{01}(s,W)-H_{02}(s,W)-H_{12}(t,s,W)}h_{01}(s,W)\,\mathrm{d}s\right] \qquad (33)\\
&\quad+O_P\left\{\left\|[H_{0l}-\hat{H}_{0l}]h_{01}e^{-H_{01}-H_{02}-H_{12}}\right\|\right.\\
&\qquad\left.\times\left\|[H_{01}-\hat{H}_{01}]+[H_{02}-\hat{H}_{02}]+[H_{12}-\hat{H}_{12}]\right\|\right\}\\
&=\mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-H_{01}(s,W)-H_{02}(s,W)-H_{12}(t,s,W)}h_{01}(s,W)\,\mathrm{d}s\right]\\
&\quad+O_P\left\{\|H_{0l}-\hat{H}_{0l}\|\cdot\left(\|H_{01}-\hat{H}_{01}\|+\|H_{02}-\hat{H}_{02}\|+\|H_{12}-\hat{H}_{12}\|\right)\right\},
\end{aligned}
$$

where the last equality follows from the triangle inequality, and the assumptions that H_{01}, H_{02}, H_{12}, and h_{01} are uniformly bounded. Equations (32) and (33) and Lemma C.2 then gives

$$\begin{aligned}
&\mathbb{E}\left[\int_0^t \rho(t,u,W;\hat{\nu})\frac{\tilde{N}_{0l}(\mathrm{d}u)-\tilde{Y}_0(u)\hat{H}_{0l}(\mathrm{d}u,W)}{e^{-\hat{\Gamma}_0(u,W)}}\right] \\
&= \mathbb{E}\left[\int_0^t [H_{0l}-\hat{H}_{0l}](s,W)e^{-H_{01}(s,W)-H_{02}(s,W)-H_{12}(t,s,W)}h_{01}(s,W)\,\mathrm{d}s\right] \\
&\quad + O_P\{\hat{\varepsilon}_{0l}\cdot(\hat{\varepsilon}_{01}+\hat{\varepsilon}_{02}+\hat{\varepsilon}_{12})\}.
\end{aligned} \tag{34}$$

Next we have by Lemma C.3 (b) that

$$\begin{aligned}
&\mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T},W)}\int_0^t e^{-[\hat{H}_{12}(t,\tilde{T}_0,W)-\hat{H}_{12}(u,\tilde{T}_0,W)]}\frac{\tilde{N}_{12}(\mathrm{d}u)-\tilde{Y}_1(u)\hat{H}_{12}(\mathrm{d}u,\tilde{T}_0,W)}{e^{-\hat{\Gamma}_1(u,\tilde{T}_0,W)}}\right] \\
&= \mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T},W)-\hat{H}_{12}(t,\tilde{T}_0,W)}\int_0^t e^{\hat{H}_{12}(u,\tilde{T}_0,W)+\hat{\Gamma}_1(u,\tilde{T}_0,W)}\left[\tilde{N}_{12}(\mathrm{d}u)-\tilde{Y}_1(u)\hat{H}_{12}(\mathrm{d}u,\tilde{T}_0,W)\right]\right] \\
&= \mathbb{E}\Bigg[e^{\hat{\Gamma}_0(\tilde{T},W)-\hat{H}_{12}(t,\tilde{T}_0,W)}\mathbb{1}\{\Delta_0=1,\eta=1,\tilde{T}_0<t\} \\
&\qquad \times\int_0^t e^{\hat{H}_{12}(u,\tilde{T}_0,W)-H_{12}(u,\tilde{T}_0,W)+\hat{\Gamma}_1(u,\tilde{T}_0,W)-\Gamma_1(u,\tilde{T}_0,W)}[H_{12}-\hat{H}_{12}](\mathrm{d}u,\tilde{T}_0,W)\Bigg] \\
&= \mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T},W)-\hat{H}_{12}(t,\tilde{T}_0,W)}\mathbb{1}\{\Delta_0=1,\eta=1,\tilde{T}_0<t\}\int_0^t[H_{12}-\hat{H}_{12}](\mathrm{d}u,\tilde{T}_0,W)\right] \\
&\quad + O_P\{(\hat{\varepsilon}_{12}+\hat{\varepsilon}_{C1})\cdot\hat{\varepsilon}_{12}\},
\end{aligned}$$

where the last equality follows from Lemmas C.2 and C.4 and the triangle inequality. By the same arguments we have

$$\begin{aligned}
&\mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T},W)-\hat{H}_{12}(t,\tilde{T}_0,W)}\mathbb{1}\{\Delta_0=1,\eta=1,\tilde{T}_0<t\}\int_0^t[H_{12}-\hat{H}_{12}](\mathrm{d}u,\tilde{T}_0,W)\right] \\
&= \mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T},W)-\hat{H}_{12}(t,\tilde{T}_0,W)}\mathbb{1}\{\Delta_0=1,\eta=1,\tilde{T}_0<t\}[H_{12}-\hat{H}_{12}](t,\tilde{T}_0,W)\right] \\
&= \mathbb{E}\left[e^{\Gamma_0(\tilde{T},W)-H_{12}(t,\tilde{T}_0,W)}\mathbb{1}\{\Delta_0=1,\eta=1,\tilde{T}_0<t\}[H_{12}-\hat{H}_{12}](t,\tilde{T}_0,W)\right] \\
&\quad + O_P\{(\hat{\varepsilon}_{C0}+\hat{\varepsilon}_{12})\cdot\hat{\varepsilon}_{12}\} \\
&= \mathbb{E}\left[\int_0^t[H_{12}-\hat{H}_{12}](t,s,W)e^{-H_{12}(t,s,W)-H_{01}(s,W)-H_{02}(s,W)}h_{01}(s,W)\,\mathrm{d}s\right] \\
&\quad + O_P\{(\hat{\varepsilon}_{C0}+\hat{\varepsilon}_{12})\cdot\hat{\varepsilon}_{12}\},
\end{aligned}$$

where the last equality follows by calculating the conditional expectation given W and using the tower property. Thus we can write

$$\begin{aligned}
&\mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T},W)}\int_0^t e^{-[\hat{H}_{12}(t,\tilde{T}_0,W)-\hat{H}_{12}(u,\tilde{T}_0,W)]}\frac{\tilde{N}_{12}(\mathrm{d}u)-\tilde{Y}_1(u)\hat{H}_{12}(\mathrm{d}u,\tilde{T}_0,W)}{e^{-\hat{\Gamma}_1(u,\tilde{T}_0,W)}}\right] \\
&= \mathbb{E}\left[\int_0^t[H_{12}-\hat{H}_{12}](t,s,W)e^{-H_{12}(t,s,W)-H_{01}(s,W)-H_{02}(s,W)}h_{01}(s,W)\,\mathrm{d}s\right] \\
&\quad + O_P\{(\hat{\varepsilon}_{C1}+\hat{\varepsilon}_{C0}+\hat{\varepsilon}_{12})\cdot\hat{\varepsilon}_{12}\}.
\end{aligned} \tag{35}$$

Combining equations (31), (34), and (35) gives

$$\begin{aligned}
&\mathrm{Rem}(P,\hat{\nu}) \\
&= \mathbb{E}\Bigg[\int_0^t \Big(e^{[H_{01}-\hat{H}_{01}](u,W)+[H_{01}-\hat{H}_{02}](u,W)+[H_{12}-\hat{H}_{12}](t,u,W)} - 1 \\
&\qquad - \Big([H_{01}-\hat{H}_{01}](u,W) + [H_{01}-\hat{H}_{02}](u,W) + [H_{12}-\hat{H}_{12}](t,u,W)\Big)\Big) \\
&\qquad\qquad \times e^{-H_{01}(u,W)-H_{02}(u,W)-H_{12}(t,u,W)} H_{01}(\mathrm{d}u,W)\Bigg] \\
&+ O_P\{(\hat{\varepsilon}_{C1}+\hat{\varepsilon}_{C0}+\hat{\varepsilon}_{12})\,\hat{\varepsilon}_{12} + (\hat{\varepsilon}_{01}+\hat{\varepsilon}_{02})\,(\hat{\varepsilon}_{01}+\hat{\varepsilon}_{02}+\hat{\varepsilon}_{12})\}.
\end{aligned} \tag{36}$$

Using the power series expansion of e^x we have that

$$e^x - 1 - x = x^2 \sum_{k=0}^{\infty} \frac{x^k}{(k+2)!} \leq x^2 \sum_{k=0}^{\infty} \frac{x^k}{k!} = x^2 e^x,$$

and as $e^x - 1 - x$ is minimized at 0, we can write

$$\begin{aligned}
&\Bigg|\mathbb{E}\Bigg[\int_0^t \Big(e^{[H_{01}-\hat{H}_{01}](u,W)+[H_{01}-\hat{H}_{02}](u,W)+[H_{12}-\hat{H}_{12}](t,u,W)} - 1 \\
&\qquad - \Big([H_{01}-\hat{H}_{01}](u,W) + [H_{01}-\hat{H}_{02}](u,W) + [H_{12}-\hat{H}_{12}](t,u,W)\Big)\Big) \\
&\qquad\qquad \times e^{-H_{01}(u,W)-H_{02}(u,W)-H_{12}(t,u,W)} H_{01}(\mathrm{d}u,W)\Bigg]\Bigg| \\
&\leq \mathbb{E}\Bigg[\int_0^t \Big([H_{01}-\hat{H}_{01}](u,W) + [H_{01}-\hat{H}_{02}](u,W) + [H_{12}-\hat{H}_{12}](t,u,W)\Big)^2 \\
&\qquad\qquad \times e^{-\hat{H}_{01}(u,W)-\hat{H}_{02}(u,W)-\hat{H}_{12}(t,u,W)} H_{01}(\mathrm{d}u,W)\Bigg] \\
&= O_P\{\hat{\varepsilon}_{01}^2 + \hat{\varepsilon}_{02}^2 + \hat{\varepsilon}_{12}^2\},
\end{aligned}$$

where we used the triangle inequality, that $\hat{H}_{01}$, $\hat{H}_{02}$, $\hat{H}_{12}$, and H_{01} are assumed uniformly bounded with probability tending to 1, and Lemma C.2 for the last equality. From this calculation and equation (36) we thus finally obtain

$$\mathrm{Rem}(P,\hat{\nu}) = O_P\big\{(\hat{\varepsilon}_{C1}+\hat{\varepsilon}_{C0}+\hat{\varepsilon}_{12})\,\hat{\varepsilon}_{12} + (\hat{\varepsilon}_{01}+\hat{\varepsilon}_{02})\,(\hat{\varepsilon}_{01}+\hat{\varepsilon}_{02}+\hat{\varepsilon}_{12}) + \hat{\varepsilon}_{01}^2 + \hat{\varepsilon}_{02}^2 + \hat{\varepsilon}_{12}^2\big\}.$$

□

C.1 Bounding the remainder when using the Nelson-Aalen estimator

In this section we give an additional argument that is needed when we cannot assume that estimators of the cumulative hazard H_{01} and H_{02} are absolutely continuous, but are instead Nelson-Aalen estimators. Firstly, when we have no baseline covariates the remainder term

defined in equation (10) is equal to

$$
\begin{aligned}
&\mathrm{Rem}(P, \hat{\nu}) \\
&= \int_0^t e^{-\hat{H}_{01}(s)-\hat{H}_{02}(s)-\hat{H}_{12}(t,s)} \hat{H}_{01}(\mathrm{d}s) - \int_0^t e^{-H_{01}(s)-H_{02}(s)-H_{12}(t,s)} H_{01}(\mathrm{d}s) \\
&\quad + \mathbb{E}\left[\int_0^t e^{-\hat{H}_{12}(t,u)} \frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{01}(\mathrm{d}u)}{e^{-\hat{\Gamma}_0(u)}} \right] \\
&\quad - \mathbb{E}\left[\int_0^t \rho(t,u;\hat{\nu}) \frac{\tilde{N}_{01}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{01}(\mathrm{d}u)}{e^{-\hat{\Gamma}_0(u)}} \right] \\
&\quad - \mathbb{E}\left[\int_0^t \rho(t,u;\hat{\nu}) \frac{\tilde{N}_{02}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{02}(\mathrm{d}u)}{e^{-\hat{\Gamma}_0(u)}} \right] \\
&\quad - \mathbb{E}\left[e^{\hat{\Gamma}_0(\tilde{T})} \int_0^t e^{-[\hat{H}_{12}(t,\tilde{T}_0)-\hat{H}_{12}(u,\tilde{T}_0)]} \frac{\tilde{N}_{12}(\mathrm{d}u) - \tilde{Y}_1(u)\hat{H}_{12}(\mathrm{d}u,\tilde{T}_0)}{e^{-\hat{\Gamma}_1(u,\tilde{T}_0)}} \right].
\end{aligned} \tag{37}
$$

Similar to Lemma C.3, for any measurable function $f\colon [0,t] \to \mathbb{R}$ we have that

$$
\begin{aligned}
\mathbb{E}\left[\int_0^t f(u)[\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0(u)\hat{H}_{0l}(\mathrm{d}u)] \right] &= \mathbb{E}\left[\int_0^t f(u)\tilde{N}_{0l}(\mathrm{d}u) \right] - \int_0^t f(u)\mathbb{E}[\tilde{Y}_0(u)]\hat{H}_{0l}(\mathrm{d}u) \\
&= \int_0^t f(u) e^{-H_{01}(u)-H_{02}(u)-\Gamma_0(u)} [H_{0l} - \hat{H}_{0l}](\mathrm{d}u),
\end{aligned}
$$

for $l \in \{1,2\}$. We can use this calculation on the three terms in equation (37) involving $\tilde{N}_{0l}(\mathrm{d}u) - \tilde{Y}_0\hat{H}_{0l}(\mathrm{d}u)$. However, to bound these term we in the previous section used Cauchy-Schwarz' inequality, see in particular Lemma C.4 and equation (34). We cannot use the same argument here as we cannot write $\hat{H}_{0l} = \hat{h}_{0l} \cdot \mu$ with respect to some fixed measure μ. In the following lemma we show how we can instead control these terms using empirical process theory. With this result in hand, one can proceed to bound the remainder in equation (37) using the same arguments as given in the main section (Appendix C) above.

Below we use $\tilde{T}$ to denote an event time and Δ denotes whether a particular event of interest was observed or not. We use h to denote the cause-specific hazard of this event and P denotes the distribution of the observed data $(\tilde{T}, \Delta)$.

Lemma C.5. *Let $\hat{H}_n$ be the Nelson-Aalen estimator of the cause-specific cumulative hazard function H, and let $\mathcal{F}$ be a Donsker class of functions $f\colon [0,t] \to \mathbb{R}$. Assume that $\hat{f}_n \in \mathcal{F}$ with probability tending to one, and $\|\hat{f}_n\|_{\mathcal{L}^2_m([0,t])} \xrightarrow{P} 0$. Assume also that h is uniformly bounded away from 0 and infinity on $[0,t]$ and that $P(\tilde{T} > t) > 0$. Then*

$$
\int_0^t \hat{f}_n(u)[H - \hat{H}_n](\mathrm{d}u) = o_P(n^{-1/2}).
$$

Proof. Let $y(s) = P(\tilde{T} \geq s)$, $\tilde{F}(s) = P(\tilde{T} \leq s, \Delta = 1)$, $\hat{y}_n(s) = \frac{1}{n}\sum_{i=1}^n \tilde{Y}_i(s)$ and $\hat{\eta}(s) := \mathbb{1}\{\hat{y}_n(s) > 0\}$, where $\tilde{Y}_i(s)$ is the at risk indicator at time s. Then we can write

$$
\mathrm{d}H = \frac{\mathrm{d}\tilde{F}}{y} = \frac{\hat{\eta}\,\mathrm{d}\tilde{F}}{\hat{y}_n} + \left(\frac{1}{y} - \frac{\hat{\eta}}{\hat{y}_n} \right) \mathrm{d}\tilde{F},
$$

and so

$$\begin{aligned}
&\int_0^t \hat{f}_n(u)[H - \hat{H}_n](\mathrm{d}u) \\
&= \int_0^t \hat{f}_n(u) \left[\frac{\hat{\eta}(u)\tilde{F}(\mathrm{d}u)}{\hat{y}_n(u)} - \hat{H}_n(\mathrm{d}u) \right] + \int_0^t \hat{f}_n(u) \left(\frac{1}{y(u)} - \frac{\hat{\eta}(u)}{\hat{y}_n(u)} \right) \tilde{F}(\mathrm{d}u) \\
&= \int_0^t \frac{\hat{\eta}(u)\hat{f}_n(u)}{\hat{y}_n(u)} \left[\tilde{F}(\mathrm{d}u) - \left(\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\tilde{T}_i \in \mathrm{d}u, \Delta_i = 1\} \right) \right] \\
&\quad + \int_0^t \hat{f}_n(u) \left(\frac{1}{y(u)} - \frac{\hat{\eta}(u)}{\hat{y}_n(u)} \right) \tilde{F}(\mathrm{d}u).
\end{aligned}$$

If we define the functions $g_n \colon [0,t] \times \{0,1\} \to \mathbb{R}$ as $g_n(u,\delta) := \delta \hat{\eta}(u) \hat{f}_n(u) \hat{y}_n(u)^{-1}$, then

$$\sqrt{n} \int_0^t \frac{\hat{\eta}(u)\hat{f}_n(u)}{\hat{y}_n(u)} \left[\tilde{F}(\mathrm{d}u) - \left(\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\tilde{T}_i \in \mathrm{d}u, \Delta_i = 1\} \right) \right] = \sqrt{n}(P - \mathbb{P}_n)[g_n] = -\mathbb{G}_n[[g_n]],$$

where $\mathbb{G}_n[=]\sqrt{n}(\mathbb{P}_n - P)$ is the empirical process with respect to the observations $(\tilde{T}_i, \Delta_i)$. As $\hat{f}_n$ and $\hat{y}_n$ belong to Donsker classes with probability tending to one so does g_n. As $\hat{f}_n$ converges to 0 in probability with respect to the $\mathcal{L}^2_m([0,t])$-norm so does g_n, and by the assumptions about h and P this implies that also $g_n \xrightarrow{P} 0$ with respect to the $\mathcal{L}^2_P([0,t])$-norm. From this it follows by lemma 19.24 in van der Vaart [2000] that $\mathbb{G}_n[[g_n]] = o_P(1)$. Hence the result will follow if we can show that

$$\int_0^t \hat{f}_n(u) \left(\frac{1}{y(u)} - \frac{\hat{\eta}(u)}{\hat{y}_n(u)} \right) \tilde{F}(\mathrm{d}u) = o_P(n^{-1/2}). \tag{38}$$

With probability tending to 1, $\hat{\eta} = 1$ and thus

$$\int_0^t \hat{f}_n(u) \left(\frac{1}{y(u)} - \frac{\hat{\eta}(u)}{\hat{y}_n(u)} \right) \tilde{F}(\mathrm{d}u) = \int_0^t \hat{f}_n(u)\hat{\eta}(u) \left(\frac{1}{y(u)} - \frac{1}{\hat{y}_n(u)} \right) \tilde{F}(\mathrm{d}u)$$

with probability tending to 1. We can thus focus on controlling the right hand side, which we can write as

$$\begin{aligned}
\left| \int_0^t \hat{f}_n(u) \frac{\hat{\eta}(u)(\hat{y}_n(u) - y(u))}{y(u)\hat{y}_n(u)} \tilde{F}(\mathrm{d}u) \right| &= \left| \int_0^t \hat{f}_n(u) \frac{\hat{\eta}(u)(\hat{y}_n(u) - y(u))}{\hat{y}_n(u)} H(\mathrm{d}u) \right| \\
&= \left\| \hat{f}_n \frac{\hat{\eta}_n}{\hat{y}_n} (\hat{y}_n - y) h \right\|_{\mathcal{L}_m([0,t])}.
\end{aligned}$$

Now, by assumption there exist a constant $C < \infty$ such that $\|h\|_\infty \leq C$, where we use $\|\cdot\|_\infty$ to denote the supremum-norm on $[0,t]$. We can then use this and Hölder's inequality twice to write

$$\begin{aligned}
\left\| \hat{f}_n \frac{\hat{\eta}_n}{\hat{y}_n} (y - \hat{y}_n) h \right\|_{\mathcal{L}^1_m([0,t])} &\leq C \left\| \hat{f}_n \frac{\hat{\eta}_n}{\hat{y}_n} (y - \hat{y}_n) \right\|_{\mathcal{L}^1_m([0,t])} \\
&\leq C \left\| \hat{f}_n \frac{\hat{\eta}_n}{\hat{y}_n} \right\|_{\mathcal{L}^1_m([0,t])} \cdot \|y - \hat{y}_n\|_\infty \\
&\leq C \|\hat{f}_n\|_{\mathcal{L}^2_m([0,t])} \cdot \left\| \frac{\hat{\eta}_n}{\hat{y}_n} \right\|_{\mathcal{L}^2_m([0,t])} \cdot \|y - \hat{y}_n\|_\infty.
\end{aligned}$$

It follows from the strong law of large numbers and the assumption $y(t) = P(\tilde{T} > t) > 0$ that $\|\hat{\eta}_n / \hat{y}_n\|_\infty = O_P(1)$, and by Donsker's theorem $\|y - \hat{y}_n\|_\infty = O_P(n^{-1/2})$. Finally, we have by assumption that $\|\hat{f}_n\|_{\mathcal{L}^2_m([0,t])} = o_P(1)$, and thus

$$\left\| \hat{f}_n \frac{\hat{\eta}_n}{\hat{y}_n} (y - \hat{y}_n) h \right\|_{\mathcal{L}^1_m([0,t])} = o_P(1) \cdot O_P(1) \cdot O_P(n^{-1/2}) = o_P(n^{-1/2}),$$

which implies equation (38). □

D Iterative representation of some simple integrals

In this section we derive an algorithm that can be used to calculate the one-step estimator defined in equation (6) when we use estimators of the transition hazard functions that are piece-wise constant. This is for instance the case when we use the penalized Poisson regression approach described in Section 5. The final procedure is summarized in Algorithm 1. It is often possible to perform the computations needed in Algorithm 1 in a vectorized fashion, and hence each step of the algorithm can be done for all data points at once. In this section we suppress the dependence on baseline covariates W and the index i.

Let $0 = t_0 < t_1 < \cdots < t_K = \tau$ be a fixed grid on the interval $[0, \tau]$. Let $\mathcal{A}$ denote the class of functions $a\colon [0,\tau] \to \mathbb{R}$ that are constant on the each interval $[t_{k-1}, t_k)$ for $k = 1, \ldots, K$, and let $\mathcal{B}$ be the class of functions $b\colon [0,\tau]^2 \to \mathbb{R}$ that are constant on each square $[t_{k-1}, t_k) \times [t_{l-1}, t_l)$ for all $k = 1, \ldots, K$ and $l = 1, \ldots, K$. To calculate the estimator defined in (6) when we use nuisance parameter estimators that give piece-wise constant hazard functions, the main difficulty lies in calculating expressions of the form

$$f(s, r, t; a, b, h) = \int_s^r \exp\left(\int_s^u a(v)\,\mathrm{d}v + \int_u^t b(v, u)\,\mathrm{d}v\right) h(u)\,\mathrm{d}u, \tag{39}$$

with $0 \le s < r \le t \le \tau$, and $a, h \in \mathcal{A}, b \in \mathcal{B}$, or the form

$$g(r, t; c, d, f) = \int_0^r f(u, t, t) \exp\left(\int_0^u c(v)\,\mathrm{d}v\right) d(u)\,\mathrm{d}u, \tag{40}$$

with $0 < r \le t \le \tau$, and $c, d \in \mathcal{A}$, and where f is on the form in equation (39). For instance, to calculate the one-step estimator from equation (6) we need to calculate

$$\int_0^t e^{-\hat{H}_{12}(t,u)} \frac{\tilde{Y}_0(s)\hat{H}_{01}(\mathrm{d}u)}{e^{-\Gamma_0(u)}},$$

which we can write as

$$\begin{aligned}\int_0^t e^{-\hat{H}_{12}(t,u)+\Gamma_0(u)} \tilde{Y}_0(s)\hat{H}_{01}(\mathrm{d}u) &= \int_0^{\tilde{T}_0 \wedge t} \exp\left(\int_0^u \hat{\gamma}_0(v)\,\mathrm{d} - \int_u^t \hat{h}_{12}(v, u)\,\mathrm{d}v\right) \hat{h}_{01}(u)\,\mathrm{d}u \\ &= f(0, \tilde{T}_0 \wedge t, t; \hat{\gamma}_0, -\hat{h}_{12}, \hat{h}_{01}),\end{aligned}$$

We also need to calculate

$$\int_0^t \rho(t, u, \hat{\nu}) \frac{\tilde{Y}_0(s)\hat{H}_{01}(\mathrm{d}u)}{e^{-\Gamma_0(u)}},$$

which, by definition of ρ (see equation (2)), equals $g(\tilde{T}_0 \wedge t, t; \hat{\gamma}_0, \hat{h}_{01}, f_\rho)$ where $f_\rho(u, t, t) := \rho(t, u, \hat{\nu})$.

We first show the correctness of Algorithm 1, and afterwards we demonstrate how the expression $g(r, t; c, d, f)$ in equation (40) can be rewritten as several terms on the form given in equation (39). Thus Algorithm 1 can be used to calculate both expressions.

To derive Algorithm 1, first define for $k = 1, \ldots, K$,

$$l(k) = \mathbb{1}\{s < t_k, t_{k-1} < r\} \int_{s \vee t_{k-1}}^{r \wedge t_k} \exp\left(\int_s^u a(v)\,\mathrm{d}v + \int_u^t b(v, u)\,\mathrm{d}v\right) h(u)\,\mathrm{d}u. \tag{41}$$

Then define recursively

$$L(0) = 0, \quad L(k) = L(k-1) + l(k),$$

Algorithm 1: Iterative integral calculation

Input: Functions $a, h \in \mathcal{A}$ and $b \in \mathcal{B}$ and time points $0 \leq s < r \leq t \leq \tau$
Output: The value $f(s, r, t; a, b, h)$ defined in equation (39)

Initialize: $L \leftarrow 0$, $\Delta A \leftarrow 0$, $A \leftarrow 0$
for $k = 1, \ldots, K$ **do**
  $\delta \leftarrow \mathbb{1}\{s < t_k, t_{k-1} < r\}$
  if $\delta = 0$ **then**
    continue
  $A \leftarrow A + \Delta A$
  $h \leftarrow h(t_{k-1})$
  $a \leftarrow a(t_{k-1})$
  $b \leftarrow b(t_{k-1}, t_{k-1})$
  $B \leftarrow b \cdot [(t_k \wedge t) - (t_k \wedge r)] + \sum_{l=k}^{K-1} \mathbb{1}\{t_l < t\} b(t_l, t_{k-1})[(t_{l+1} \wedge t) - t_l]$
  $\Delta T \leftarrow (r \wedge t_k) - (s \vee t_{k-1})$
  $\Delta A \leftarrow a \cdot \Delta T$
  $\beta \leftarrow b \cdot \Delta T$
  $l \leftarrow h \cdot e^{A+B}$
  if $a \neq b$ **then**
    $l \leftarrow l \cdot \frac{e^{\Delta A} - e^{\beta}}{a - b}$
  else
    $l \leftarrow l \cdot e^{\Delta A} \cdot \Delta T$
  $L \leftarrow L + l$
return L

and note that $f(s, r, t; a, b, h) = L(K)$. For $u \in [t_{k-1}, t_k)$ we have that

$$\begin{aligned}
&\exp\left\{\int_s^u a(v)\,\mathrm{d}v + \int_u^t b(v, u)\,\mathrm{d}v\right\} \\
&= \exp\left\{\int_s^{t_{k-1} \vee s} a(v)\,\mathrm{d}v - a(t_{k-1})[t_{k-1} \vee s] + \int_{r \wedge t_k}^t b(v, t_{k-1})\,\mathrm{d}v + b(t_{k-1}, t_{k-1})[r \wedge t_k]\right\} \times \\
&\quad \exp\left\{[a(t_{k-1}) - b(t_{k-1}, t_{k-1})]\, u\right\},
\end{aligned}$$

and thus

$$\begin{aligned}
&\int_{s \vee t_{k-1}}^{r \wedge t_k} \exp\left(\int_s^u a(v)\,\mathrm{d}v + \int_u^t b(v, u)\,\mathrm{d}v\right) h(u)\,\mathrm{d}u \\
&= \exp\left\{\int_s^{t_{k-1} \vee s} a(v)\,\mathrm{d}v - a(t_{k-1})[t_{k-1} \vee s] + \int_{r \wedge t_k}^t b(v, t_{k-1})\,\mathrm{d}v + b(t_{k-1}, t_{k-1})[r \wedge t_k]\right\} \times \\
&\quad h(t_{k-1}) \int_{s \vee t_{k-1}}^{r \wedge t_k} \exp\left\{[a(t_{k-1}) - b(t_{k-1}, t_{k-1})]\, u\right\} \mathrm{d}u.
\end{aligned}$$

For $a(t_{k-1}) \neq b(t_{k-1}, t_{k-1})$ this equals

$$\begin{aligned}
&\exp\left\{\int_s^{t_{k-1} \vee s} a(v)\,\mathrm{d}v + \int_{r \wedge t_k}^t b(v, t_{k-1})\,\mathrm{d}v\right\} \times \\
&\quad \exp\left\{-a(t_{k-1})[t_{k-1} \vee s] + b(t_{k-1}, t_{k-1})[r \wedge t_k]\right\} \times \\
&\quad h(t_{k-1}) \frac{\exp\left\{[a(t_{k-1}) - b(t_{k-1}, t_{k-1})]\,(r \wedge t_k)\right\} - \exp\left\{[a(t_{k-1}) - b(t_{k-1}, t_{k-1})]\,(s \vee t_{k-1})\right\}}{a(t_{k-1}) - b(t_{k-1}, t_{k-1})} \\
&= \frac{h(t_{k-1})}{a(t_{k-1}) - b(t_{k-1}, t_{k-1})} \exp\left\{\int_s^{t_{k-1} \vee s} a(v)\,\mathrm{d}v + \int_{r \wedge t_k}^t b(v, t_{k-1})\,\mathrm{d}v\right\} \times \\
&\quad \exp\left\{a(t_{k-1})\,[(r \wedge t_k) - (s \vee t_{k-1})]\right\} - \exp\left\{b(t_{k-1}, t_{k-1})\,[(r \wedge t_k) - (s \vee t_{k-1})]\right\},
\end{aligned}$$

while for $a(t_{k-1}) = b(t_{k-1}, t_{k-1})$ we get

$$\begin{aligned}
&\exp\left\{\int_s^{t_{k-1}\vee s} a(v)\,\mathrm{d}v + \int_{r\wedge t_k}^{t} b(v, t_{k-1})\,\mathrm{d}v\right\} \times \\
&\quad \exp\left\{-a(t_{k-1})[t_{k-1}\vee s] + b(t_{k-1}, t_{k-1})[r\wedge t_k]\right\} h(t_{k-1})\,[r\wedge t_k - s\vee t_{k-1}] \\
&= \exp\left\{\int_s^{t_{k-1}\vee s} a(v)\,\mathrm{d}v + \int_{r\wedge t_k}^{t} b(v, t_{k-1})\,\mathrm{d}v\right\} \times \\
&\quad \exp\left\{a(t_{k-1})[r\wedge t_k - t_{k-1}\vee s]\right\} h(t_{k-1})\,[r\wedge t_k - s\vee t_{k-1}]\,.
\end{aligned}$$

Define for $k = 1, \ldots, K$,

$$\begin{aligned}
A(k) &= \int_s^{t_{k-1}\vee s} a(v)\,\mathrm{d}v, \\
B(k) &= \int_{r\wedge t_k}^{t} b(v, t_{k-1})\,\mathrm{d}v, \\
\Delta A(k) &= a(t_{k-1})\,[(r\wedge t_k) - (s\vee t_{k-1})]\,, \\
\beta(k) &= b(t_{k-1}, t_{k-1})\,[(r\wedge t_k) - (s\vee t_{k-1})]\,,
\end{aligned}$$

where we note that $B(k)$ is well-defined because $r > t_{k-1}$, so $v \geq t_{k-1}$. Then we may write

$$\begin{aligned}
l(k) = \mathbb{1}\{s < t_k, t_{k-1} < r\} h(t_{k-1}) e^{A(k)+B(k)} \times \\
\Big(\mathbb{1}\{a(t_{k-1}) \neq b(t_{k-1})\} \frac{e^{\Delta A(k)} - e^{\beta(k)}}{a(t_{k-1}) - b(t_{k-1}, t_{k-1})} + \\
\mathbb{1}\{a(t_{k-1}) = b(t_{k-1})\} e^{\Delta A(k)} [(r\wedge t_k) - (s\vee t_{k-1})]\Big).
\end{aligned}$$

Defining $\Delta A(0) = 0$ and $A(0) = 0$, for use in the above expression we can calculate $A(k)$ recursively as

$$A(k) = A(k-1) + \Delta A(k-1),$$

because this holds when $t_{k-1} < r$. Thus all terms except for $B(k)$ can be calculated recursively using only the values of the function h, a, and b when evaluated at t_{k-1} at each step. We calculate $B(k)$ as

$$B(k) = b(t_{k-1}, t_{k-1})[(t_k \wedge t) - (t_k \wedge r)] + \mathbb{1}\{k < K\} \sum_{l=k}^{K-1} \mathbb{1}\{t_l < t\} b(t_l, t_{k-1})[(t_{l+1} \wedge t) - t_l]$$

where we have again used that $r > t_{k-1}$. This demonstrate the correctness of the procedure given in Algorithm 1.

Finally we consider terms of the form in equation (40). Firstly, for any $r \in (u, t)$ we can write

$$f(u, t, t; a, b, h) = f(u, r, t; a, b, h) + e^{\int_u^r a(v)\,\mathrm{d}v} f(r, t, t; a, b, h),$$

and thus

$$
\begin{aligned}
&g(r,t;c,d,f(\cdot;a,b,h))\\
&=\int_0^r f(u,t,t;a,b,h)e^{\int_0^u c(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u\\
&=\int_0^r f(u,r,t;a,b,h)e^{\int_0^u c(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u\\
&\quad+\int_0^r e^{\int_u^r a(v)\,\mathrm{d}v}f(r,t,t;a,b,g)e^{\int_0^u c(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u\\
&=\int_0^r f(u,r,t;a,b,h)e^{\int_0^u c(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u\\
&\quad+f(r,t,t;a,b,h)\int_0^r e^{\int_0^u c(v)\,\mathrm{d}v+\int_u^r a(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u.\\
&=\int_0^r f(u,r,t;a,b,h)e^{\int_0^u c(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u\\
&\quad+f(r,t,t;a,b,h)e^{\int_0^r a(v)\,\mathrm{d}v}\int_0^r e^{\int_0^u [c-a](v)\,\mathrm{d}v}d(u)\,\mathrm{d}u.\\
&=\int_0^r f(u,r,t;a,b,h)e^{\int_0^u c(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u\\
&\quad+f(r,t,t;a,b,h)e^{f(0,r,t;0,0,a)}f(0,r,t;c-a,0,d).
\end{aligned}
\tag{42}
$$

The second expression can be calculated using Algorithm 1 so we now focus on the first expression,

$$
(*):=\int_0^r f(u,r,t;a,b,h)e^{\int_0^u c(v)\,\mathrm{d}v}d(u)\,\mathrm{d}u. \tag{43}
$$

Fubini's theorem gives that

$$
\begin{aligned}
(*)&=\int_0^r\int_u^r e^{\int_u^v a(z)\,\mathrm{d}z+\int_v^t b(z,v)\,\mathrm{d}z}h(v)\,\mathrm{d}v e^{\int_0^u c(z)\,\mathrm{d}z}d(u)\,\mathrm{d}u\\
&=\int_0^r\int_0^v e^{\int_u^v a(z)\,\mathrm{d}z+\int_v^t b(z,v)\,\mathrm{d}z}h(v)e^{\int_0^u c(z)\,\mathrm{d}z}d(u)\,\mathrm{d}u\,\mathrm{d}v\\
&=\int_0^r\int_0^v e^{\int_0^v a(z)\,\mathrm{d}z+\int_v^t b(z,v)\,\mathrm{d}z}h(v)e^{\int_0^u c(z)-a(z)\,\mathrm{d}z}d(u)\,\mathrm{d}u\,\mathrm{d}v\\
&=\int_0^r e^{\int_0^v a(z)\,\mathrm{d}z+\int_v^t b(z,v)\,\mathrm{d}z}h(v)\int_0^v e^{\int_0^u [c-a](z)\,\mathrm{d}z}d(u)\,\mathrm{d}u\,\mathrm{d}v,
\end{aligned}
$$

so if we define

$$
\begin{aligned}
H(v)&:=e^{\int_0^v a(z)\,\mathrm{d}z+\int_v^t b(z,v)\,\mathrm{d}z}h(v)\\
F(u)&:=e^{\int_0^u [c-a](z)\,\mathrm{d}z}d(u),
\end{aligned}
$$

we have

$$
(*)=\int_0^r H(v)\int_0^v F(u)\,\mathrm{d}u\,\mathrm{d}v.
$$

In the following we assume that $a(t_k)\neq c(t_k)$ for all $k=1,\ldots,K$. For any functions H and

F we may write

$$
\begin{aligned}
(*) &= \sum_{k=1}^{K} \left\{ \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v) \sum_{l=1}^{k} \int_{t_{l-1}}^{t_l \wedge v} F(u)\,\mathrm{d}u\,\mathrm{d}v \right\} \\
&= \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v) \int_{t_{k-1}}^{t_k \wedge v} F(u)\,\mathrm{d}u\,\mathrm{d}v \\
&\quad + \sum_{k=2}^{K} \left\{ \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v) \sum_{l=1}^{k-1} \int_{t_{l-1}}^{t_l \wedge v} F(u)\,\mathrm{d}u\,\mathrm{d}v \right\} \qquad (44) \\
&= \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v) \int_{t_{k-1}}^{v} F(u)\,\mathrm{d}u\,\mathrm{d}v \\
&\quad + \sum_{k=2}^{K} \left\{ \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \sum_{l=1}^{k-1} \int_{t_{l-1}}^{t_l} F(u)\,\mathrm{d}u \right\}.
\end{aligned}
$$

Now, consider the summands of the first sum of the right hand side. We have that

$$
\begin{aligned}
&\int_{t_{k-1}}^{t_k \wedge r} H(v) \int_{t_{k-1}}^{v} F(u)\,\mathrm{d}u\,\mathrm{d}v \\
&= \int_{t_{k-1}}^{t_k \wedge r} H(v) \int_{t_{k-1}}^{t_k \wedge r} F(u)\,\mathrm{d}u\,\mathrm{d}v - \int_{t_{k-1}}^{t_k \wedge r} H(v) \int_{v}^{t_k \wedge r} F(u)\,\mathrm{d}u\,\mathrm{d}v \\
&= \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \int_{t_{k-1}}^{t_k \wedge r} F(u)\,\mathrm{d}u - \int_{t_{k-1}}^{t_k \wedge r} H(v) \int_{v}^{t_k \wedge r} F(u)\,\mathrm{d}u\,\mathrm{d}v,
\end{aligned}
$$

and for $v \in (t_{k-1}, t_k \wedge r)$,

$$
\begin{aligned}
\int_{v}^{t_k \wedge r} F(u)\,\mathrm{d}u &= \int_{v}^{t_k \wedge r} e^{\int_0^u [c-a](z)\,\mathrm{d}z} d(u)\,\mathrm{d}u \\
&= e^{\int_0^{t_{k-1}} [c-a](z)\,\mathrm{d}z} d(t_{k-1}) \int_{v}^{t_k \wedge r} e^{\int_{t_{k-1}}^u [c-a](z)\,\mathrm{d}z}\,\mathrm{d}u \\
&= e^{\int_0^{t_{k-1}} [c-a](z)\,\mathrm{d}z} d(t_{k-1}) \int_{v}^{t_k \wedge r} e^{[c-a](t_{k-1})\cdot(u-t_{k-1})}\,\mathrm{d}u \\
&= \frac{e^{\int_0^{t_{k-1}} [c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \left(e^{[c-a](t_{k-1})\cdot(t_k \wedge r - t_{k-1})} - e^{[c-a](t_{k-1})\cdot(v-t_{k-1})} \right) \\
&= \frac{e^{\int_0^{t_k \wedge r} [c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} - \frac{e^{\int_0^{v} [c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})},
\end{aligned}
$$

and hence we can write

$$
\begin{aligned}
&\int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \int_{t_{k-1}}^{v} F(u)\,\mathrm{d}u \\
&= \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \int_{t_{k-1}}^{t_k\wedge r} F(u)\,\mathrm{d}u - \frac{e^{\int_0^{t_k\wedge r}[c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \\
&\quad + \int_{t_{k-1}}^{t_k\wedge r} H(v) \frac{e^{\int_0^{v}[c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})}\,\mathrm{d}v \\
&= \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \int_{t_{k-1}}^{t_k\wedge r} F(u)\,\mathrm{d}u - \frac{e^{\int_0^{t_k\wedge r}[c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \\
&\quad + \int_{t_{k-1}}^{t_k\wedge r} H(v) \frac{e^{\int_0^{v}[c-a](z)\,\mathrm{d}z} d(v)}{[c-a](v)}\,\mathrm{d}v \\
&= \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \int_{t_{k-1}}^{t_k\wedge r} F(u)\,\mathrm{d}u - \frac{e^{\int_0^{t_k\wedge r}[c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \\
&\quad + \int_{t_{k-1}}^{t_k\wedge r} e^{\int_0^{v} a(z)\,\mathrm{d}z + \int_v^{t} b(z,v)\,\mathrm{d}z} h(v) \frac{e^{\int_0^{v}[c-a](z)\,\mathrm{d}z} d(v)}{[c-a](v)}\,\mathrm{d}v \\
&= \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \int_{t_{k-1}}^{t_k\wedge r} F(u)\,\mathrm{d}u \\
&\quad - \frac{e^{\int_0^{t_k\wedge r}[c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \int_{t_{k-1}}^{t_k\wedge r} H(v)\,\mathrm{d}v \\
&\quad + \int_{t_{k-1}}^{t_k\wedge r} e^{\int_0^{v} c(z)\,\mathrm{d}z + \int_v^{t} b(z,v)\,\mathrm{d}z} \frac{h(v)d(v)}{[c-a](v)}\,\mathrm{d}v.
\end{aligned}
\tag{45}
$$

Combining equations (44) and (45) then gives

$$
\begin{aligned}
(*) &= \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v) \int_{t_{k-1}}^{v} F(u)\,\mathrm{d}u\,\mathrm{d}v \\
&\quad + \sum_{k=2}^{K} \left\{ \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \sum_{l=1}^{k-1} \int_{t_{l-1}}^{t_l} F(u)\,\mathrm{d}u \right\} \\
&= \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \int_{t_{k-1}}^{t_k \wedge r} F(u)\,\mathrm{d}u \\
&\quad - \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \frac{e^{\int_0^{t_k \wedge r} [c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \\
&\quad + \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} e^{\int_0^v c(z)\,\mathrm{d}z + \int_v^t b(z,v)\,\mathrm{d}z} \frac{h(v)d(v)}{[c-a](v)}\,\mathrm{d}v \\
&\quad + \sum_{k=2}^{K} \left\{ \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \sum_{l=1}^{k-1} \int_{t_{l-1}}^{t_l} F(u)\,\mathrm{d}u \right\} \\
&= \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \left\{ \sum_{l=1}^{k} \int_{t_{l-1}}^{t_l \wedge r} F(u)\,\mathrm{d}u - \frac{e^{\int_0^{t_k \wedge r} [c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \right\} \\
&\quad + \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} e^{\int_0^v c(z)\,\mathrm{d}z + \int_v^t b(z,v)\,\mathrm{d}z} \frac{h(v)d(v)}{[c-a](v)}\,\mathrm{d}v \\
&= \sum_{k=1}^{K} \mathbb{1}\{t_{k-1} < r\} \int_{t_{k-1}}^{t_k \wedge r} H(v)\,\mathrm{d}v \left\{ \int_{0}^{t_k \wedge r} F(u)\,\mathrm{d}u - \frac{e^{\int_0^{t_k \wedge r} [c-a](z)\,\mathrm{d}z} d(t_{k-1})}{[c-a](t_{k-1})} \right\} \\
&\quad + \int_0^r e^{\int_0^v c(z)\,\mathrm{d}z + \int_v^t b(z,v)\,\mathrm{d}z} \frac{h(v)d(v)}{[c-a](v)}\,\mathrm{d}v.
\end{aligned}
$$

Defining $\tilde{h}$ as the piece-wise constant function determined by

$$
\tilde{h}(t_k) = \int_0^{t_{k+1} \wedge r} F(u)\,\mathrm{d}u - \frac{d(t_k)}{[c-a](t_k)} e^{\int_0^{t_{k+1} \wedge r} [c-a](z)\,\mathrm{d}z}, \quad \text{for} \quad k = 0, \ldots, K-1, \tag{46}
$$

we then have

$$
\begin{aligned}
(*) &= \int_0^r H(v)\tilde{h}(v)\,\mathrm{d}v + \int_0^r e^{\int_0^v c(z)\,\mathrm{d}z + \int_v^t b(z,v)\,\mathrm{d}z} \frac{h(v)d(v)}{[c-a](v)}\,\mathrm{d}v \\
&= \int_0^r e^{\int_0^v a(z)\,\mathrm{d}z + \int_v^t b(z,v)\,\mathrm{d}z} h(v)\tilde{h}(v)\,\mathrm{d}v + \int_0^r e^{\int_0^v c(z)\,\mathrm{d}z + \int_v^t b(z,v)\,\mathrm{d}z} \frac{h(v)d(v)}{[c-a](v)}\,\mathrm{d}v \\
&= f(0, r, t; a, b, h \cdot \tilde{h}) + f\left(0, r, t; c, b, \frac{h \cdot d}{c-a}\right).
\end{aligned}
$$

Using this and equations (42) and (43) we conclude that

$$
\begin{aligned}
g(r, t; c, d, f(\cdot; a, b, h)) = &f(0, r, t; a, b, h \cdot \tilde{h}) + f\left(0, r, t; c, b, \frac{h \cdot d}{c-a}\right) \\
&+ f(r, t, t; a, b, h) e^{f(0,r,t;0,0,a)} f(0, r, t; c-a, 0, d),
\end{aligned} \tag{47}
$$

and hence we can use Algorithm 1 to also calculate g if we can evaluate $\tilde{h}$ defined in equation (46) efficiently in addition to a, b, h, c, and d. A simple calculation shows that

$$
\int_0^{t_{k+1} \wedge t} F(u)\,\mathrm{d}u = \int_0^{t_k} F(u)\,\mathrm{d}u + \frac{d(t_k)}{[c-a](t_k)} e^{\int_0^{t_{k+1} \wedge t} [c-a](z)\,\mathrm{d}z} - \frac{d(t_k)}{[c-a](t_k)} e^{\int_0^{t_k} [c-a](z)\,\mathrm{d}z},
$$

which implies

$$\begin{aligned}
\tilde{h}(t_k) &= \int_0^{t_{k+1}\wedge t} F(u)\,\mathrm{d}u - \frac{d(t_k)}{[c-a](t_k)} e^{\int_0^{t_{k+1}\wedge t}[c-a](z)\,\mathrm{d}z} \\
&= \int_0^{t_k} F(u)\,\mathrm{d}u - \frac{d(t_k)}{[c-a](t_k)} e^{\int_0^{t_k}[c-a](z)\,\mathrm{d}z} \\
&= \int_0^{t_k} F(u)\,\mathrm{d}u - \frac{d(t_k)}{[c-a](t_k)} e^{\int_0^{t_k}[c-a](z)\,\mathrm{d}z} \pm \frac{d(t_{k-1})}{[c-a](t_{k-1})} e^{\int_0^{t_k}[c-a](z)\,\mathrm{d}z} \\
&= \tilde{h}(t_{k-1}) + \left(\frac{d(t_{k-1})}{[c-a](t_{k-1})} - \frac{d(t_k)}{[c-a](t_k)} \right) e^{\int_0^{t_k}[c-a](z)\,\mathrm{d}z}.
\end{aligned}$$

The values $\int_0^{t_k}[c-a](z)\,\mathrm{d}z$ can be calculated recursively, and so also $\tilde{h}$ can be calculated recursively. Hence with Algorithm 1 and the relation in equation (47) we have an computationally efficient way of calculating g defined in equation (40).

E Additional simulations

To further examine the performance of the one-step estimator we conduct a simulation study when baseline covariates are present. In this setting all transition hazard functions are estimated with the method proposed in Section 5. We use the same data-generating mechanism as described in Section 6, except that we use a coarser time grid for computational reasons. The transition hazard function from state 1 to state 2 is now

$$h_{12}(u,s) = \sum_{i=1}^{5} \mathbb{1}\{2(i-1) \le s < 2i\}\beta_i,$$

where the β_i's are chosen as the equally spaced grid of 5 decreasing numbers with $\beta_1 = 0.3$ and $\beta_5 = 0.01$. The four other transition hazard functions remain the same. In addition to the censored state transition times we generate a set of binary baseline covariates which have no effect on the transition times or the censoring time. We consider the cases with 1, 3, and 5 baseline covariates. We consider only two censoring regimes, one with state-dependent censoring ($\alpha = 0.02$) and one with no state-dependent censoring ($\alpha = 0.2$), see Section 6. We simulate data with sample sizes 200, 500, 1000, and 5000. All simulations are repeated 1000 times.

We again calculate the Aalen-Johansen estimator, the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the one-step estimator $\hat{\Psi}_t$ defined in equation (6). We now use the coarser time grid $0, 2, \ldots, 8, 10$ for the penalized Poisson regression models but also include all interactions of all orders between the baseline covariates and the time grid.

The results are shown in Figures 6-7 and Tables 2-3. From this we draw the same conclusion as we did in Section 6.

We have formulas and estimators for the asymptotic variance of the Aalen-Johansen estimator and the one-step estimator, and thus we can also examine the coverage of Wald-based confidence intervals for these two estimators. When we use the data-adaptive estimation method from Section 5 we do not have a formula for the asymptotic variance of the plug-in estimator so there is no coverage to examine for this estimator. The asymptotic variance of the Aalen-Johansen estimator is estimated using the `R`-package `etm` [Allignol et al., 2011] which uses a Greenwood type estimator [Andersen et al., 2012]. The asymptotic variance for the one-step estimator can be estimated using the asymptotic linear expansion given by Theorem 4.2. Figure 8 shows the coverage of the Nelson-Aalen estimator and the one-step

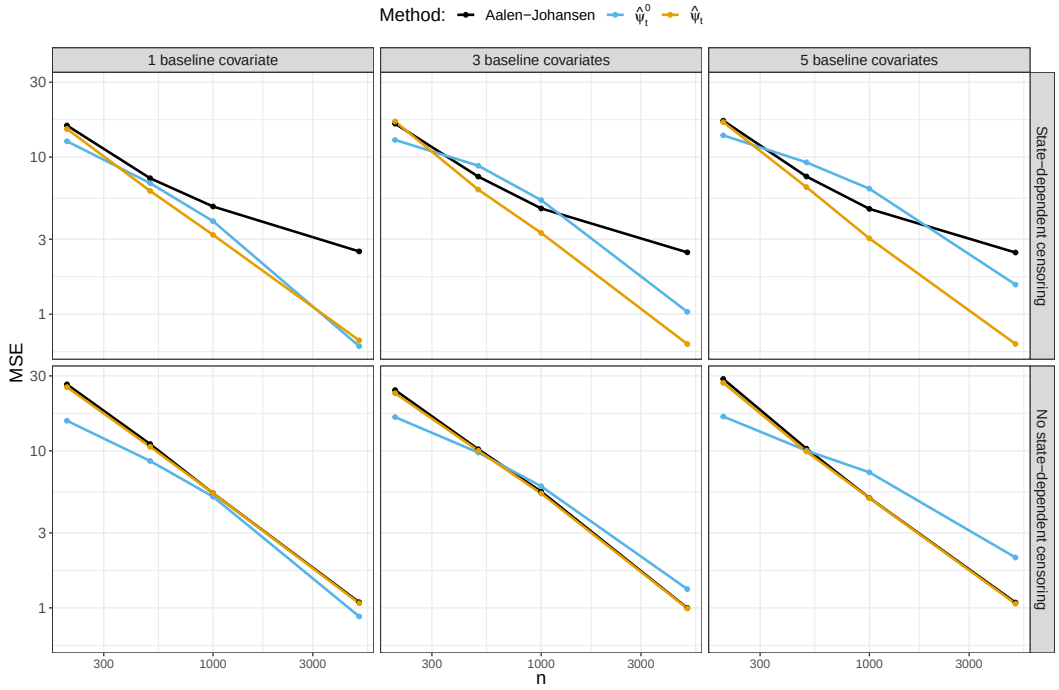


Figure 6: The mean squared error (MSE) estimated with 1000 simulations of the three estimators for two censoring regimes (with and without state-dependent censoring) and different number of baseline covariates (1, 3, and 5) plotted against sample size. Note the log scale on both axis. The estimators are the Aalen-Johansen estimator (Aa-J), the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the debiased estimator $\hat{\Psi}_t$ defined in equation (6).

estimator. We see that the one-step estimator has good coverage which is comparable to the coverage for the Nelson-Aalen estimator when there is no state-dependent censoring. When there is state-dependent censoring the coverage for the one-step estimator remains good while the coverage for the Nelson-Aalen estimator is poor and decreases to 0 with sample size due to bias.

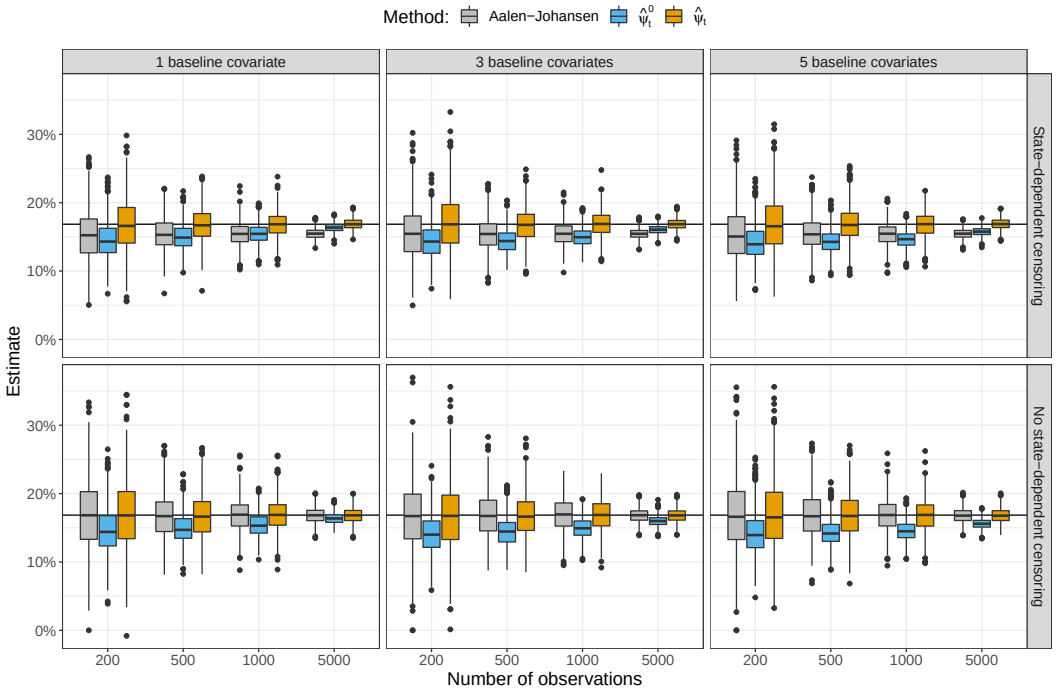


Figure 7: The results of 1000 simulations of the three estimators for different censoring regimes, increasing number of baseline covariates, and increasing number of samples. The gray boxplots show the distribution of the Aalen-Johansen estimator, the blue boxplots show the distribution of the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the orange boxplots show the distribution of the debiased estimator $\hat{\Psi}_t$ defined in equation (6). The black line is the state occupation probability Ψ_t at time $t = 9$ of the data-generating distribution.

		Bias			SE			MSE		
Baseline covariates	n	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$
1	200	-1.65	-2.33	-0.17	3.63	2.68	3.89	15.90	12.62	15.11
	500	-1.43	-1.86	-0.03	2.30	1.84	2.47	7.34	6.83	6.09
	1000	-1.44	-1.40	-0.03	1.67	1.40	1.79	4.85	3.91	3.20
	5000	-1.39	-0.48	0.03	0.76	0.63	0.83	2.51	0.63	0.68
3	200	-1.33	-2.56	0.06	3.82	2.51	4.11	16.33	12.86	16.86
	500	-1.47	-2.42	-0.09	2.31	1.71	2.49	7.51	8.82	6.21
	1000	-1.40	-1.88	0.02	1.66	1.34	1.81	4.71	5.32	3.29
	5000	-1.38	-0.79	0.02	0.75	0.63	0.80	2.48	1.03	0.65
5	200	-1.48	-2.74	-0.04	3.86	2.51	4.09	17.06	13.76	16.70
	500	-1.37	-2.52	0.00	2.38	1.71	2.54	7.53	9.26	6.44
	1000	-1.43	-2.20	-0.04	1.63	1.21	1.74	4.68	6.29	3.04
	5000	-1.39	-1.08	0.04	0.74	0.61	0.80	2.47	1.54	0.65

Table 2: The results of 1000 simulations of the three estimators when there is state-dependent censoring for different numbers of baseline covariates (1, 3, and 5) and sample sizes ($n \in \{200, 500, 1000, 5000\}$). The table shows the bias, standard error (SE), and mean squared error (MSE) of the Aalen-Johansen estimator (Aa-J), the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the debiased estimator $\hat{\Psi}_t$ defined in equation (6).

		Bias			SE			MSE		
Baseline covariates	n	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$	Aa-J	$\hat{\Psi}_t^0$	$\hat{\Psi}_t$
1	200	-0.11	-2.23	-0.10	5.15	3.25	5.05	26.48	15.54	25.45
	500	-0.15	-1.95	-0.18	3.32	2.19	3.25	11.04	8.62	10.61
	1000	-0.02	-1.45	-0.01	2.32	1.74	2.32	5.37	5.10	5.38
	5000	-0.04	-0.51	-0.04	1.04	0.79	1.03	1.09	0.88	1.07
3	200	-0.12	-2.77	-0.14	4.93	2.95	4.82	24.34	16.39	23.26
	500	-0.10	-2.36	-0.12	3.20	2.04	3.16	10.23	9.75	10.01
	1000	0.10	-1.90	0.06	2.34	1.53	2.31	5.50	5.94	5.35
	5000	-0.04	-0.87	-0.05	1.00	0.75	1.00	1.00	1.32	0.99
5	200	-0.05	-2.69	0.02	5.36	3.04	5.21	28.67	16.52	27.08
	500	-0.08	-2.53	-0.10	3.21	1.91	3.15	10.32	10.03	9.90
	1000	-0.01	-2.29	-0.04	2.24	1.43	2.24	5.01	7.29	5.00
	5000	-0.07	-1.25	-0.08	1.04	0.73	1.03	1.08	2.09	1.06

Table 3: The results of 1000 simulations of the three estimators when there is no state-dependent censoring for different numbers of baseline covariates (1, 3, and 5) and sample sizes ($n \in \{200, 500, 1000, 5000\}$). The table shows the bias, standard error (SE), and mean squared error (MSE) of the Aalen-Johansen estimator (Aa-J), the plug-in estimator $\hat{\Psi}_t^0$ defined in equation (5), and the debiased estimator $\hat{\Psi}_t$ defined in equation (6).

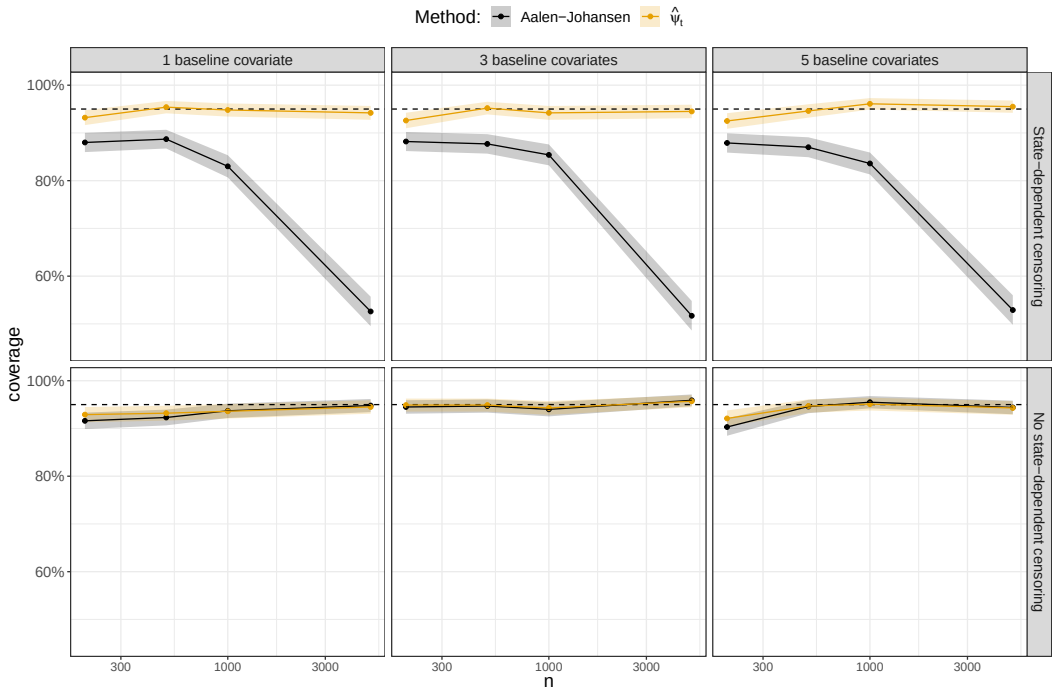


Figure 8: The coverage of Wald-based confidence intervals estimated with 1000 simulations of the Aalen-Johansen estimator and the debiased estimator $\hat{\Psi}_t$ defined in equation (6). Results are shown for two censoring regimes (with and without state-dependent censoring) and different number of baseline covariates (1, 3, and 5) plotted against sample size. The confidence bands around the estimated coverage are calculated using the empirical standard deviation of the 1000 Monte Carlo samples of the coverage.

References

A. Allignol, M. Schumacher, and J. Beyersmann. Empirical transition matrix of multi-state models: The etm package. *Journal of Statistical Software*, 38(4):1–15, 2011. URL `https://www.jstatsoft.org/v38/i04/`.

A. Allignol, J. Beyersmann, T. Gerds, and A. Latouche. A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 20(4):495–513, 2014.

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

P. J. Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL `https://doi.org/10.1111/ectj.12097`.

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021.

S. Datta and G. A. Satten. Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models. *Statistics & probability letters*, 55(4):403–411, 2001.

S. Datta and G. A. Satten. Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics*, 58(4):792–802, 2002.

S. Datta, G. A. Satten, and S. Datta. Nonparametric estimation for the three-stage irreversible illness–death model. *Biometrics*, 56(3):841–847, 2000.

J. de Uña-Álvarez and L. Meira-Machado. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics*, 71(2): 364–375, 2015.

A. Fisher and E. H. Kennedy. Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2):162–172, 2021.

E. Fix and J. Neyman. A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biology*, 23(3):205–241, 1951.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `https://www.jstatsoft.org/v33/i01/`.

T. A. Gerds. *prodlim: Product-Limit Estimation for Censored Event History Analysis*, 2019. URL `https://CRAN.R-project.org/package=prodlim`. R package version 2019.11.13.

R. D. Gill, M. J. van der 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.

R. D. Gill, M. J. van der Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

D. V. Glidden. Robust inference for event probabilities with non-Markov event data. *Biometrics*, 58(2):361–368, 2002.

N. Gunnes, Ø. Borgan, and O. O. Aalen. Estimating stage occupation probabilities in non-Markov models. *Lifetime data analysis*, 13(2):211–240, 2007.

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

P. J. Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

C. H. Kraft and C. van Eeden. Asymptotic efficiencies of quick methods of computing efficient estimates based on ranks. *Journal of the American Statistical Association*, 67 (337):199–202, 1972.

N. Maltzahn, R. Hoff, O. O. Aalen, I. S. Mehlum, H. Putter, and J. M. Gran. A hybrid landmark Aalen-Johansen estimator for transition probabilities in partially non-Markov multi-state models. *Lifetime data analysis*, 27(4):737–760, 2021.

L. Meira-Machado, J. De Una-Alvarez, and C. Cadarso-Suarez. Nonparametric estimation of transition probabilities in a non-Markov illness–death model. *Lifetime Data Analysis*, 12(3):325–344, 2006.

S. Murray and A. A. Tsiatis. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*, pages 137–151, 1996.

M. L. Petersen and M. J. van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418, 2014.

J. Pfanzagl and W. Wefelmeyer. *Contributions to a general asymptotic statistical theory*. Springer, 1982.

T. PROVA Study Group. Prophylaxis of first hemorrhage from esophageal varices by sclerotherapy, propranolol or both in cirrhotic patients: a randomized multicenter trial. *Hepatology*, 14(6):1016–1024, 1991.

H. Putter and C. Spitoni. Non-parametric estimation of transition probabilities in non-Markov multi-state models: the landmark Aalen–Johansen estimator. *Statistical methods in medical research*, 27(7):2081–2092, 2018.

J. M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer, 1992.

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*, 2022.

H. C. W. Rytgaard, F. Eriksson, and M. van der Laan. Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation, 2021. URL `https://arxiv.org/abs/2106.11009`.

E. Sverdrup. Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. *Scandinavian Actuarial Journal*, 1965(3-4):184–211, 1965.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A. C. Titman. Transition probability estimates for non-Markov multi-state models. *Biometrics*, 71(4):1034–1041, 2015.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2006.

M. J. van der Laan and D. Benkeser. Highly adaptive lasso (hal). In *Targeted Learning in Data Science*, pages 77–94. Springer, 2018.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Bepress, 2003.

M. J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

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

M. J. van der Laan and S. Rose. *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer, 2018.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

A. W. van der Vaart. On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991.

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

A. W. van der Vaart. On Robins' formula. *Statistics & Decisions*, 22(3):171–200, 2004.

A. W. 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, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.

J. Xu, J. D. Kalbfleisch, and B. Tai. Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics*, 66(3):716–725, 2010.

Manuscript II

Title	The state learner – a super learner for right-censored data
Authors	Anders Munch and Thomas A. Gerds
Status	In preparation

The state learner – a super learner for right-censored data

Anders Munch and Thomas A. Gerds

Abstract

In survival analysis, prediction models are needed as stand-alone tools and in applications of causal inference to estimate nuisance parameters. The super learner is a machine learning algorithm which combines a library of prediction models into a meta learner based on cross-validated loss. Unfortunately, the commonly used partial likelihood loss is not suited for super learning, and inverse probability of censoring weighted loss functions require a pre-specified estimator of the censoring distribution. To relax this, we introduce the state learner, a new super learner for survival analysis, which evaluates the loss based on the observed data simultaneously for two libraries of prediction models, one for the event time distribution and one for the censoring distribution. We establish an oracle inequality for the state learner and investigate its performance through numerical experiments.

1 Introduction

A super learner is a machine learning algorithm that combines a finite set of learners into a meta learner by estimating prediction performance in hold-out samples using a pre-specified loss function [van der Laan et al., 2007]. When the aim is to make a prediction model, super learners typically combine strong learners, such as Cox regression models and random survival forests [Gerds and Kattan, 2021]. While the general idea of combining strong learners based on cross-validation data stems from earlier work [Wolpert, 1992, Breiman, 1996], the name super learner is justified by an oracle inequality [van der Laan and Dudoit, 2003, van der Vaart et al., 2006]: The super learner is guaranteed to perform almost as well as the model which minimizes the expected performance, i.e., the model we would select if we could evaluate the prediction performance in an infinite hold-out sample.

We are concerned with the choice of the loss function for super learning in survival analysis. Existing super learner algorithms for right-censored data use partial log-likelihood loss or inverse probability of censoring weighted loss [Polley and van der Laan, 2011, Keles et al., 2004, Golmakani and Polley, 2020, Westling et al., 2021]. The use of the partial log-likelihood loss restricts the class of learners and excludes for example simple Kaplan-Meier based learners and also more complex random survival forest algorithms. For this reason Golmakani and Polley [2020] restrict their learners to Cox proportional hazard models. A lesser known fact is that a super learner constructed with the negative partial log-likelihood loss implicitly depends on the censoring distribution (Appendix A). A disadvantage of inverse probability of censoring weighted loss functions is that they requires a pre-specified model for the censoring distribution. Westling et al. [2021] tackle this challenge by iterating between super learning of the censoring distribution and the event time distribution.

In this article we define the state learner, a new super learner for right-censored data, which simultaneously evaluates the loss for learners of the event time distribution and the censoring distribution. The loss function which is used to define the state learner is only based on observable quantities. The state learner can be applied to all types of survival estimators, works in the presence of competing risks, and does not require a single pre-specified estimator

of the conditional censoring distribution. To analyze the theoretical properties of the state learner we focus on the so-called discrete super learner which 'combines' the library of learners by picking the one that minimizes the cross-validated loss. The state learner uses separate libraries to model each competing event and the censoring distribution. We show that the oracle selector of the state learner is consistent if all libraries contain a consistent learner and prove a finite sample oracle inequality.

The state learner can be used to select a model which predicts the probability of an event based on covariates in the presence of competing risks. Another application is in targeted learning where conditional event probabilities occur as high-dimensional nuisance parameters which need to be estimated at a certain rate [van der Laan and Rose, 2011, Rytgaard et al., 2021, Rytgaard and van der Laan, 2022]. We show how a targeted estimator can be obtained from the state learner, and that a second order product structure for the asymptotic bias term of the targeted estimator is retained when the state learner is used to estimate nuisance parameters.

The article is organized as follows. We introduce our notation and framework in Section 2. In Section 3 we define super learning in general with right-censored data, and in Section 4 we introduce the state learner. Section 5 provides theoretical guarantees for the state learner. In Section 6 we discuss the use of the state learner in the context of targeted learning. We report a numerical study in Section 7, and analyze a prostate cancer data set in Section 8. Finally, we relate the state learner to existing approaches in Section 9 and discuss some limitations of our proposal. In Appendix A, we present a formal result stating that the oracle according to the partial log-likelihood loss (always) depends on the censoring distribution. Appendices B and C contain proofs.

2 Notation and framework

In a competing risk framework [Andersen et al., 2012], let T be a time to event variable, $D \in \{1, 2\}$ the cause of the event, and $X \in \mathcal{X}$ a vector of baseline covariates taking values in a bounded subset $\mathcal{X} \subset \mathbb{R}^p$, $p \in \mathbb{N}$. Let $\tau < \infty$ be the maximal length of follow-up. We use $\mathcal{Q}$ to denote the collection of all probability measures on $[0, \tau] \times \{1, 2\} \times \mathcal{X}$ such that $(T, D, X) \sim Q$ for some unknown $Q \in \mathcal{Q}$. For $j \in \{1, 2\}$, the cause-specific conditional cumulative hazard functions are defined by $\Lambda_j \colon [0, \tau] \times \mathcal{X} \to \mathbb{R}_+$ such that

$$\Lambda_j(t \mid x) = \int_0^t \frac{Q(T \in \mathrm{d}s, D = j \mid X = x)}{Q(T \geq s \mid X = x)}.$$

For ease of notation we assume throughout that $\Lambda_j(\cdot \mid x)$ is continuous for all x and j. We denote by S the conditional event-free survival function,

$$S(t \mid x) = \exp\left\{-\Lambda_1(t \mid x) - \Lambda_2(t \mid x)\right\}.$$

Let $\mathcal{M}$ denote the space of all conditional cumulative hazard functions on $[0, \tau] \times \mathcal{X}$. Any distribution $Q \in \mathcal{Q}$ can be characterized by

$$\begin{aligned} Q(\mathrm{d}t, j, \mathrm{d}x) = &\left\{S(t- \mid x)\Lambda_1(\mathrm{d}t \mid x)H(\mathrm{d}x)\right\}^{\mathbb{1}\{j=1\}} \\ &\left\{S(t- \mid x)\Lambda_2(\mathrm{d}t \mid x)H(\mathrm{d}x)\right\}^{\mathbb{1}\{j=2\}}, \end{aligned}$$

where $\Lambda_j \in \mathcal{M}$ for $j = 1, 2$ and H is the marginal distribution of the covariates.

We consider the usual right-censored setting in which we observe data $O = (\tilde{T}, \tilde{D}, X)$, where $\tilde{T} = \min(T, C)$ for a right-censoring time C, $\Delta = \mathbb{1}\{T \leq C\}$, and $\tilde{D} = \Delta D$. Let $\mathcal{P}$

denote a set of probability measures on the sample space $\mathcal{O} = [0, \tau] \times \{0, 1, 2\} \times \mathcal{X}$ such that $O \sim P$ for some unknown $P \in \mathcal{P}$. We assume that the event time and the censoring time are conditionally independent given covariates, $T \perp\!\!\!\perp C \mid X$. This implies that any distribution $P \in \mathcal{P}$ is characterized by a distribution $Q \in \mathcal{Q}$ and a conditional cumulative hazard function for C given X [c.f., Begun et al., 1983, Gill et al., 1997]. We use $\Gamma \in \mathcal{M}$ to denote the conditional cumulative hazard function for censoring. We assume that $\Gamma(\cdot \mid x)$ is continuous for all x, and let $G(t \mid x) = \exp\{-\Gamma(t \mid x)\}$ denote the survival function of the conditional censoring distribution. In our setting with competing risks, this yields

$$\begin{aligned} P(\mathrm{d}t, j, \mathrm{d}x) &= \{G(t- \mid x)S(t- \mid x)\Lambda_1(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=1\}} \\ &\quad \{G(t- \mid x)S(t- \mid x)\Lambda_2(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=2\}} \\ &\quad \{G(t- \mid x)S(t- \mid x)\Gamma(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=0\}} \\ &= \{G(t- \mid x)Q(\mathrm{d}t, j, \mathrm{d}x)\}^{\mathbb{1}\{j\neq 0\}} \\ &\quad \{G(t- \mid x)S(t- \mid x)\Gamma(\mathrm{d}t \mid x)H(\mathrm{d}x)\}^{\mathbb{1}\{j=0\}}. \end{aligned} \tag{1}$$

Hence, we may write $\mathcal{P} = \{P_{Q,\Gamma} : Q \in \mathcal{Q}, \Gamma \in \mathcal{G}\}$ for some $\mathcal{G} \subset \mathcal{M}$. We also have

$$P(\tilde{T} > t \mid X = x) = S(t \mid x)G(t \mid x) = \exp\{-\Lambda_1(t \mid x) - \Lambda_2(t \mid x) - \Gamma(t \mid x)\}.$$

We further assume that there exists $\kappa < \infty$ such that $\Lambda_j(\tau- \mid x) < \kappa$, for $j \in \{1, 2\}$, and $\Gamma(\tau- \mid x) < \kappa$ for almost all $x \in \mathcal{X}$. Note that this implies that $G(\tau- \mid x)$ is bounded away from zero for almost all $x \in \mathcal{X}$. Under these assumptions, the conditional cumulative hazard functions Λ_j and Γ can be identified from P by

$$\Lambda_j(t \mid x) = \int_0^t \frac{P(\tilde{T} \in \mathrm{d}s, \tilde{D} = j \mid X = x)}{P(\tilde{T} \geq s \mid X = x)}, \tag{2}$$

$$\Gamma(t \mid x) = \int_0^t \frac{P(\tilde{T} \in \mathrm{d}s, \tilde{D} = 0 \mid X = x)}{P(\tilde{T} \geq s \mid X = x)}. \tag{3}$$

Thus, we can consider Λ_j and Γ as operators which map from $\mathcal{P}$ to $\mathcal{M}$.

3 Super learning with right-censored survival data

A super learner estimates a parameter Ψ which can be identified from the observed data distribution $P \in \mathcal{P}$. In this section, to introduce the concept of super learning, we simply consider estimation of the function Λ_j. The parameter $\Psi : \mathcal{P} \to \mathcal{M}$ is then identified via equation (2) by $\Psi(P) = \Lambda_j$.

As input to the super learner we need a sample $\mathcal{D}_n = \{O_i\}_{i=1}^n$ of i.i.d. observations from some unknown $P \in \mathcal{P}$ and a finite collection of candidate learners $\mathcal{A}$. Each learner $a \in \mathcal{A}$ is a map $a \colon \mathcal{O}^n \to \mathcal{M}$ which takes a data set as input and returns an estimate $a(\mathcal{D}_n) \in \mathcal{M}$ of Λ_j. In what follows, we use the short-hand notation $P[f] = \int f(o)P(\mathrm{d}o)$. A super learner evaluates the performance of $a \in \mathcal{A}$ using a loss function $L \colon \mathcal{M} \times \mathcal{O} \to \mathbb{R}_+$ by estimating the expected loss $P[L(a(\mathcal{D}_n), \cdot)]$ using cross-validation. Specifically, the expected loss of $a \in \mathcal{A}$ is estimated by splitting the data set $\mathcal{D}_n$ into K disjoint approximately equally sized subsets $\mathcal{D}_n^1, \mathcal{D}_n^2, \ldots, \mathcal{D}_n^K$ and then calculating the cross-validated loss

$$\hat{R}_n(a; L) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|\mathcal{D}_n^k|} \sum_{O_i \in \mathcal{D}_n^k} L\big(a(\mathcal{D}_n^{-k}), O_i\big), \quad \text{with} \quad \mathcal{D}_n^{-k} = \mathcal{D}_n \setminus \mathcal{D}_n^k.$$

The subset $\mathcal{D}_n^{-k}$ is referred to as the k'th training sample, while $\mathcal{D}_n^k$ is referred to as the k'th test or hold-out sample. The discrete super learner is defined as

$$\hat{a}_n = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \hat{R}_n(a; L).$$

The final estimator of $\Psi(P) = \Lambda_j$ is then the selected learner applied to the full data set, i.e., $\hat{a}_n(\mathcal{D}_n)$. The oracle learner is defined as the learner that minimizes the average loss according to the data-generating distribution P, i.e.,

$$\tilde{a}_n = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \tilde{R}_n(a; L), \quad \text{with} \quad \tilde{R}_n(a; L) = \frac{1}{K} \sum_{k=1}^{K} P\big[L\big(a(\mathcal{D}_n^{-k}), \cdot\big)\big].$$

Note that both the discrete super learner and the oracle learner depend on the library of learners and on the number of folds K, and that the oracle learner is a function of the data and the unknown data-generating distribution. These dependencies are suppressed in the notation.

4 The state learner

The problem with most existing super learners for right-censored data is that they depend on a pre-specified estimator of the censoring distribution. The main idea of the state learner is to jointly use learners of Λ_1, Λ_2, and Γ, and the relations in equation (1), to learn a feature of the observed data distribution P. The discrete state learner ranks a tuple of learners of $(\Lambda_1, \Lambda_2, \Gamma)$ based on how well they jointly model the observed data. To formally introduce the state learner, we define the multi-state process

$$\eta(t) = \mathbb{1}\{\tilde{T} \leq t, \tilde{D} = 1\} + 2\,\mathbb{1}\{\tilde{T} \leq t, \tilde{D} = 2\} + 3\,\mathbb{1}\{\tilde{T} \leq t, \tilde{D} = 0\}, \quad \text{for} \quad t \in [0, \tau].$$

Note that at time t, we observe that each observation is in one of four mutually exclusive states (Figure 1). The conditional distribution of $\eta(t)$ given $X = x$ is determined by the

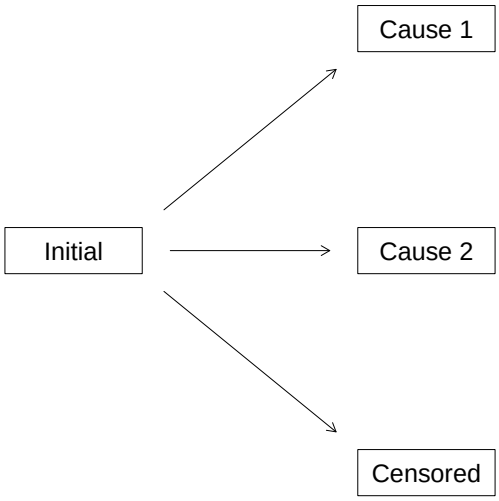


Figure 1: Illustration of the multi-state process η used by the state learner. Note that 'censored' is a state, hence the process is always observed at any time.

function

$$F(t, k, x) = P(\eta(t) = k \mid X = x), \quad \text{for all} \quad t \in [0, \tau], k \in \{0, 1, 2, 3\}, x \in \mathcal{X}. \tag{4}$$

The function F describes the conditional state occupation probabilities corresponding to the observed multi-state process η.

We propose to construct a super learner for F, i.e., the target of this super learner is $\Psi(P) = F$ where the parameter is identified through equation (4). Because each quadruple $(\Lambda_1, \Lambda_2, \Gamma, H)$ characterizes a $P \in \mathcal{P}$ which in turn determines (F, H), a learner for F can be constructed from learners of Λ_1, Λ_2, and Γ as follows:

$$\begin{aligned}
F(t,1,x) &= P(\tilde{T} \leq t, \Delta = 1 \mid X = x) = \int_0^t e^{\{-\Lambda_1(s|x)-\Lambda_2(s|x)-\Gamma(s|x)\}} \Lambda_1(\mathrm{d}s \mid x), \\
F(t,2,x) &= P(\tilde{T} \leq t, \Delta = 2 \mid X = x) = \int_0^t e^{\{-\Lambda_1(s|x)-\Lambda_2(s|x)-\Gamma(s|x)\}} \Lambda_2(\mathrm{d}s \mid x), \\
F(t,3,x) &= P(\tilde{T} \leq t, \Delta = 0 \mid X = x) = \int_0^t e^{\{-\Lambda_1(s|x)-\Lambda_2(s|x)-\Gamma(s|x)\}} \Gamma(\mathrm{d}s \mid x), \\
F(t,0,x) &= P(\tilde{T} > t \mid X = x) = 1 - F(t,1,x) - F(t,2,x) - F(t,3,x).
\end{aligned} \tag{5}$$

The state learner requires three libraries of learners, $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{B}$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ contain learners of the conditional cause-specific cumulative hazard functions of the event time distribution Λ_1 and Λ_2, respectively, and $\mathcal{B}$ contains learners of the conditional cumulative hazard function of the censoring distribution. Based on the Cartesian product of libraries of learners for $(\Lambda_1, \Lambda_2, \Gamma)$ we construct a library $\mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$ of learners for F:

$$\mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B}) = \{\varphi_{a_1,a_2,b} : a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2, b \in \mathcal{B}\},$$

where in correspondance with the relations in equation (5),

$$\begin{aligned}
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t,1,x) &= \int_0^t e^{\{-a_1(\mathcal{D}_n)(s|x)-a_2(\mathcal{D}_n)(s|x)-b(\mathcal{D}_n)(s|x)\}} a_1(\mathcal{D}_n)(\mathrm{d}s \mid x), \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t,2,x) &= \int_0^t e^{\{-a_1(\mathcal{D}_n)(s|x)-a_2(\mathcal{D}_n)(s|x)-b(\mathcal{D}_n)(s|x)\}} a_2(\mathcal{D}_n)(\mathrm{d}s \mid x), \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t,3,x) &= \int_0^t e^{\{-a_1(\mathcal{D}_n)(s|x)-a_2(\mathcal{D}_n)(s|x)-b(\mathcal{D}_n)(s|x)\}} b(\mathcal{D}_n)(\mathrm{d}s \mid x), \\
\varphi_{a_1,a_2,b}(\mathcal{D}_n)(t,0,x) &= 1 - \sum_{j=1}^{3} \varphi_{a_1,a_2,b}(\mathcal{D}_n)(t,j,x).
\end{aligned}$$

To evaluate how well a function F predicts the observed multi-state process we use the integrated Brier score $\bar{B}_\tau(F,O) = \int_0^\tau B_t(F,O)\,\mathrm{d}t$, where B_t is the Brier score [Brier et al., 1950] at time $t \in [0,\tau]$,

$$B_t(F,O) = \sum_{j=0}^{3} \left(F(t,j,X) - \mathbb{1}\{\eta(t) = j\}\right)^2.$$

As described in Section 3, each learner $\varphi_{a_1,a_2,b}$ in the library $\mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$ is evaluated using the cross-validated loss,

$$\hat{R}_n(\varphi_{a_1,a_2,b}; \bar{B}_\tau) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{D}_n^k|} \sum_{O_i \in \mathcal{D}_n^k} \bar{B}_\tau\big(\varphi_{a_1,a_2,b}(\mathcal{D}_n^{-k}), O_i\big),$$

and the discrete state learner is

$$\hat{\varphi}_n = \underset{(a_1,a_2,b) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{B}}{\operatorname{argmin}} \hat{R}_n(\varphi_{a_1,a_2,b}; \bar{B}_\tau).$$

5 Theoretical results for the state learner

In this section we establish theoretical guarantees for the state learner. Proposition 5.1 can be derived from the fact that the integrated Brier score (also called the continuous ranked probability score) is a strictly proper scoring rule [Gneiting and Raftery, 2007]. This implies that if we minimize the average loss of the integrated Brier score, we recover the parameters of the data-generating distribution. In particular, the oracle of a state learner will be consistent if the library of learners contains at least one learner that is consistent for estimation of F. Recall that the function F implicitly depends on the data generating probability measure $P \in \mathcal{P}$ but that this was suppressed in the notation. We now make this dependence explicit by writing F_0 for the function which is obtained by substituting a specific $P_0 \in \mathcal{P}$ for P in equation (5).

Proposition 5.1. *For $P_0 \in \mathcal{P}$ define*

$$F^* = \operatorname*{argmin}_F P_0[\bar{B}_\tau(F, \cdot)],$$

where the minimum is taken over all F, such that F is a conditional state occupation probability function for some measure P as defined in equation (4). Then $F^*(t, j, \cdot) = F_0(t, j, \cdot)$ H-almost surely for any $j \in \{0, 1, 2, 3\}$ and almost any $t \in [0, \tau]$.

Proof. See Appendix B. □

We establish a finite sample oracle result for the state learner. Our Corollary 5.2 is in essence a special case of a general cross-validation result by van der Vaart et al. [2006]. We assume that we split the data into equally sized folds, and for simplicity of presentation we take n to be such that $|\mathcal{D}_n^{-k}| = n/K$ with K fixed. We will allow the number of learners to grow with n and write $\mathcal{F}_n = \mathcal{F}(\mathcal{A}_{1,n}, \mathcal{A}_{2,n}, \mathcal{B}_n)$ as short-hand notation and to emphasize the dependence on n. In the following we let $\|\cdot\|_P$ denote the norm

$$\|F\|_P = \left\{ \sum_{j=0}^{3} \int_{\mathcal{X}} \int_0^\tau F(t, j, x)^2 \, \mathrm{d}t H(\mathrm{d}x) \right\}^{1/2}. \tag{6}$$

Corollary 5.2. *For all $P_0 \in \mathcal{P}$, $n \in \mathbb{N}$, $k \in \{1, \ldots, K\}$, and $\delta > 0$,*

$$\mathbb{E}_{P_0}\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_0\|_{P_0}^2\right] \leq (1 + 2\delta)\mathbb{E}_{P_0}\left[\|\tilde{\varphi}_n(\mathcal{D}_n^{-k}) - F_0\|_{P_0}^2\right] + (1 + \delta)16K\tau\left(13 + \frac{12}{\delta}\right)\frac{\log(1 + |\mathcal{F}_n|)}{n}.$$

Proof. See Appendix B. □

Corollary 5.2 has the following asymptotic consequences.

Corollary 5.3. *Assume that $|\mathcal{F}_n| = O(n^q)$, for some $q \in \mathbb{N}$ and that there exists a sequence $\varphi_n \in \mathcal{F}_n$, $n \in \mathbb{N}$, such that $\mathbb{E}_{P_0}\left[\|\varphi_n(\mathcal{D}_n^{-k}) - F_0\|_{P_0}^2\right] = O(n^{-\alpha})$, for some $\alpha \leq 1$.*

(i) If $\alpha = 1$ then $\mathbb{E}_{P_0}\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_0\|_{P_0}^2\right] = O(\log(n)n^{-1})$.

(ii) If $\alpha < 1$ then $\mathbb{E}_{P_0}\left[\|\hat{\varphi}_n(\mathcal{D}_n^{-k}) - F_0\|_{P_0}^2\right] = O(n^{-\alpha})$.

Proof. See Appendix B. □

6 Targeted learning

For ease of presentation, in this section we discuss the special case without competing risks, i.e., where $D = 1$ for all subjects and hence $\tilde{D} = \Delta$. By abusing notation we now also simply write Λ for the cumulative hazard function instead of Λ_1. In the two-state survival setting with baseline covariates, the distribution $P \in \mathcal{P}$ is characterized by (Λ, Γ, H).

Features of the observed data distribution $P \in \mathcal{P}$ are rarely of interest. We are instead interested in a parameter $\theta \colon \mathcal{Q} \to \Theta$ that expresses a property of the uncensored population governed by the measure $Q \in \mathcal{Q}$. The parameter space Θ can be a subset of $\mathbb{R}^d$ or a subset of a function space, such as $\mathcal{M}$. In causal inference, θ could be an average treatment effect on the survival probability [Rytgaard et al., 2021] in which case $\Theta = [-1, 1]$. Under the assumption of conditional independent censoring and positivity, θ is identifiable from $\mathcal{P}$ which means that there exists an operator $\Psi \colon \mathcal{P} \to \Theta$ such that $\theta(Q) = \Psi(P_{Q,\Gamma})$ for all $\Gamma \in \mathcal{M}$. According to equation (1) and equation (28) in Appendix C, any P is characterized by both (Λ, Γ, H) and (F, H), which means that there exist operators $\tilde{\Psi}$ and $\bar{\Psi}$ such that

$$\tilde{\Psi}(\Lambda_0, \Gamma_0, H_0) = \Psi(P_0) = \bar{\Psi}(F_0, H_0).$$

Hence, any parameter Ψ defined on $\mathcal{P}$ can be estimated using either estimators $(\hat{\Lambda}_n, \hat{\Gamma}_n, \hat{H}_n)$ or $(\hat{F}_n, \hat{H}_n)$. When we use the state learner the final estimator of the target parameter is $\bar{\Psi}(\hat{\varphi}_n(\mathcal{D}_n), \hat{H}_n)$, where $\hat{H}_n$ can often be taken to be the marginal empirical measure.

For the rest of this section we consider the case where $\Theta = \mathbb{R}$, and Ψ is estimable by a regular asymptotically linear estimator. Specifically, we assume that it is possible to construct an estimator $\tilde{\Psi}(\hat{\Lambda}_n, \hat{\Gamma}_n, H_n)$ such that the the following expansion holds:

$$\tilde{\Psi}(\hat{\Lambda}_n, \hat{\Gamma}_n, H_n) - \Psi(P) = \mathbb{P}_n[\psi_P] + \mathrm{Rem}(\hat{\Lambda}_n, \hat{\Gamma}_n, P) + o_P(n^{-1/2}), \tag{7}$$

where $\mathbb{P}_n$ is the empirical measure of a sample $\{O_i\}_{i=1}^n$, ψ_P a zero-mean function with $P[\psi_P^2] < \infty$, and $\mathrm{Rem}(\hat{\Lambda}_n, \hat{\Gamma}_n, P)$ a second order remainder term [Bickel et al., 1993, Fisher and Kennedy, 2021]. The remainder term $\mathrm{Rem}(\hat{\Lambda}_n, \hat{\Gamma}_n, P)$ is typically dominated by terms of the form

$$P\left[\int_0^\tau w_n(s, \cdot)\hat{M}_{1,n}(s \mid \cdot)\hat{M}_{2,n}(\mathrm{d}s \mid \cdot)\right], \tag{8}$$

where $(\hat{M}_{1,n}, \hat{M}_{2,n})$ is any of the four combinations of $\hat{M}_{1,n} \in \{[\Gamma - \hat{\Gamma}_n], [\Lambda - \hat{\Lambda}_n]\}$ and $\hat{M}_{2,n} \in \{[\Gamma - \hat{\Gamma}_n], [\Lambda - \hat{\Lambda}_n]\}$, and w_n is some data-dependent function with domain $[0, \tau] \times \mathcal{X}$ [van der Laan and Robins, 2003]. In particular, the estimator $\hat{\Psi}_n$ is asymptotically linear with influence function ψ_P if the 'products' of the estimation errors $\hat{M}_{1,n}$ and $\hat{M}_{2,n}$ in equation (8) are $o_P(n^{-1/2})$. The $o_P(n^{-1/2})$ bound on the products of estimation errors is typically a weaker condition than requiring Γ and Λ to be estimated independently at rate $n^{-1/2}$. This is particularly important when data-adaptive estimation methods are used as these methods rarely provide $n^{-1/2}$-rate convergence for Γ and Λ independently.

Proposition 6.1 implies that if equation (8) holds for the estimator $\tilde{\Psi}(\hat{\Lambda}_n, \hat{\Gamma}_n, H_n)$, then a similar product structure holds for the estimator $\bar{\Psi}(\hat{F}_n, H_n)$. We state the result for the special case that $\hat{M}_{1,n} = \Gamma_0 - \hat{\Gamma}_n$ and $\hat{M}_{2,n} = \Lambda_0 - \hat{\Lambda}_n$, but similar results holds for any combinations of $\Gamma_0 - \hat{\Gamma}_n$ and $\Lambda_0 - \hat{\Lambda}_n$.

Proposition 6.1. *Assume that $w(s, x) \leq c$, $F(s, 0, x) \geq 1/c$ and $\hat{F}_n(s, 0, x) \geq 1/c$ for some $c > 0$ for all $s \in [0, \tau]$ and $x \in \mathcal{X}$. Then there are real-valued uniformly bounded functions*

w_n^a, w_n^b, w_n^c, and w_n^d with domain $[0,\tau]^2 \times \mathcal{X}$ such that

$$
\begin{aligned}
&P_0\left[\int_0^\tau w(s,\cdot)\left\{\Gamma_0(s,\cdot) - \hat{\Gamma}_n(s,\cdot)\right\}[\Lambda_0 - \hat{\Lambda}_n](\mathrm{d}s,\cdot)\right] \\
&= P_0\left[\int_0^\tau \int_0^s w_n^a(s,u,\cdot)[F_0 - \hat{F}_n](u-,0,\cdot)[F_0 - \hat{F}_n](s-,0,\cdot)F_0(\mathrm{d}u,2,\cdot)F_0(\mathrm{d}s,1,\cdot)\right] \\
&\quad + P_0\left[\int_0^\tau \int_0^s w_n^b(s,u,\cdot)[F_0 - \hat{F}_n](u-,0,\cdot)F_0(\mathrm{d}u,2,\cdot)[F_0 - \hat{F}_n](\mathrm{d}s,1,\cdot)\right] \\
&\quad + P_0\left[\int_0^\tau \int_0^s w_n^c(s,u,\cdot)[F_0 - \hat{F}_n](\mathrm{d}u,2,\cdot)[F_0 - \hat{F}_n](s-,0,\cdot)F_0(\mathrm{d}s,1,\cdot)\right] \\
&\quad + P_0\left[\int_0^\tau \int_0^s w_n^d(s,u,\cdot)[F_0 - \hat{F}_n](\mathrm{d}u,2,\cdot)[F_0 - \hat{F}_n](\mathrm{d}s,1,\cdot)\right].
\end{aligned}
$$

Proof. See Appendix C. □

7 Numerical experiments

We compare the state learner with two other discrete super learners and an oracle selector in a simulation study without a competing event. The two other super learners are based on inverse probability of censoring weighted (IPCW) Brier scores [Graf et al., 1999, Gerds and Schumacher, 2006], and we refer to these as IPCW super learners. These super learners depend on an estimator of the censoring distribution, and we consider IPCW super learners that use either the Kaplan-Meier estimator (IPCW(KM)) or a correctly specified Cox model (IPCW(Cox)) to estimate the censoring distribution. Both IPCW super learners are fitted using the `R`-package `riskRegression` [Gerds et al., 2023]. Each discrete super learner provides a learner for the cumulative hazard function for the outcome of interest, and from this a risk prediction model can be obtained. We measure performance of each super learner in terms of the Brier score of the provided risk prediction model at a specific time horizon. The Brier score is approximated using a large ($n = 20,000$) independent data set of uncensored data. The oracle selector uses the large data set of uncensored event times to select the learner with the lowest expected Brier score. The expected Brier score of the oracle selector serves as a lower benchmark value. For all super learners we split the data into five folds for training and testing.

Note that given a learner for the cumulative hazard function of the outcome event, we can typically use the same method to construct a learner of the cumulative hazard function of the censoring distribution. This would typically work by training the learner on the data set $\mathcal{D}_n^c$, where $\mathcal{D}_n^c = \{O_i^c\}_{i=1}^n$ with $O_i^c = (\tilde{T}_i, 1-\Delta_i, X_i)$. When we say that we use a learner for the cumulative hazard function of the outcome to learn the cumulative hazard function of the censoring time, we mean that the learner is trained on $\mathcal{D}_n^c$.

Scenario 1 We first generate a simple dataset to demonstrate that an IPCW super learner can perform poorly when the censoring model is misspecified. We start by generating a binomial baseline covariate A with success probability 30%. We then generate outcome and censoring variables according to a Cox-Weibull distribution with hazard rates of approximately 0.5 and 2.5, respectively. We use a library $\mathcal{A}$ consisting of two learners, the (marginal) Kaplan-Meier estimator and the Kaplan-Meier estimator stratified on the binary covariate. For the state learner we use the same learners to construct the library $\mathcal{B}$.

The results are shown in Figure 2. We see that the IPCW super learner based on a misspec-

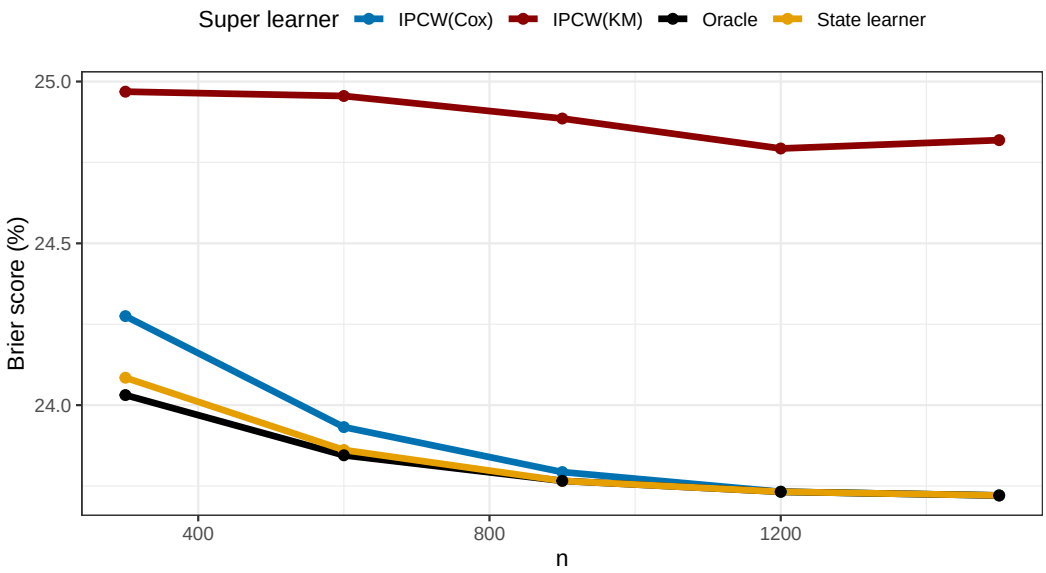


Figure 2: Results for scenario 1 of the simulation study. For the learner selected by each of the four discrete super learners, the Brier score calculated in a large independent data set without censoring is plotted against sample size. The results are based on 200 repetitions.

ified censoring model (IPCW(KM)) has a larger Brier score than the other super learners and that its performance does not improve much with sample size. The performance of the state learner is comparable to or slightly better than the IPCW super learner based on a correctly pre-specified censoring model. The performance of both the IPCW(Cox) super learner and the state learner are close to that of the oracle for most sample sizes.

Scenario 2 We next generate data according to a more complex distribution motivated from a real dataset in which censoring depends on the baseline covariates. We simulate data based on a prostate cancer study described in [Kattan et al., 2000]. The outcome of interest was the time to tumor recurrence, and five baseline covariates were used to predict outcome: prostate-specific antigen (PSA, ng/mL), Gleason score sum (GSS, values between 6 and 10), radiation dose (RD), hormone therapy (HT, yes/no) and clinical stage (CS, six values). The study was designed such that a patient's radiation dose depended on when the patient entered the study [Gerds et al., 2013]. This in turn implied that the time of censoring depended on the radiation dose. The data were re-analyzed in [Gerds et al., 2013] where a sensitivity analysis was conducted based on simulated data. We use the same simulation setup, where event and censoring times are generated according to parametric Cox-Weibull models estimated from the original data, and the covariates are generated according to either marginal Gaussian normal or binomial distributions estimated from the original data [c.f., Gerds et al., 2013, Section 4.6]. We use the library consisting of the nine learners described in Table 1. For the state learner we use the same library to learn the censoring distribution.

The results are shown in Figure 3. We see that the Brier score of the IPCW super learner based on a misspecified censoring model (IPCW(KM)) decreased with sample size, but is larger than that of the other super learners for all sample sizes. The state learner and the IPCW super learner based on a correctly pre-specified censoring model demonstrate similar performance for all sample sizes. The performance of both the IPCW(Cox) super learner and the state learner approaches the benchmark provided by the oracle selector for large

Family	Model	Description
Marginal	`KM`	The Kaplan-Meier estimator
Cox	`Cox`	All five covariates included with additive effects
	`Cox strata CS`	Cox model stratified on CS
	`Cox strata HT`	Cox model stratified on HT
	`Cox spline`	PSA and RD modeled with splines
Penalized Cox	`Lasso`	Cox model with L_1-norm penalty
	`Ridge`	Cox model with L_2-norm penalty
	`Elastic`	Cox model with L_1- and L_2-norm penalty
Random forest	`RF`	Random forest with 50 trees and default settings

Table 1: Overview of the nine learners used in scenario 2 of the simulation study. The Kaplan-Meier estimator was fitted using the package `prodlim` [Gerds, 2019]. All Cox models included all five covariates in the model and were fitted using the package `survival` [Therneau, 2022]. All penalized Cox models included all five covariates as linear predictors and were fitted using the package `glmnet` [Simon et al., 2011, Friedman et al., 2010]. The random forest was fitted with the package `randomForestSRC` [Ishwaran and Kogalur, 2023].

sample sizes.

8 Real data application

The original prostate cancer data analyzed by Kattan et al. [2000], which we introduced in Section 7, include a competing event in the form of death without tumor recurrence. To illustrate our method we fit the state learner to the original data set consisting of 1,042 patients. We consider death without tumor recurrence and recurrence of tumor as two competing events of interest. We include the fives learner `KM`, `Cox strata CS`, `Lasso`, `Elastic`, and `RF` which are described in Table 1. We use the same library of learners to learn Λ_1, Λ_2, and Γ. In this case, Λ_1 denotes the cause-specific cumulative hazard function of tumor recurrence, and Λ_2 denotes the cause-specific cumulative hazard function of death without tumor recurrence.

This gives a library consisting of $5^3 = 125$ learners for the conditional state occupation probability function F defined in equation (4). We use five folds for training and testing the models, and we repeat training and evaluation five times with different splits. The integrated Brier score for all learners are shown in Figure 4, and the top 10 combinations of learners are displayed in Table 2. We see that the prediction performance is mostly affected by the choice of learner for the censoring distribution. Several combinations of learners give similar performance as measured by the integrated Brier score, as long as a random forest is used to model the censoring distribution.

9 Discussion

We have proposed a new super learner that can be used with right-censored data and competing events. In this section, we compare our proposal to existing super learners and discuss avenues for further research.

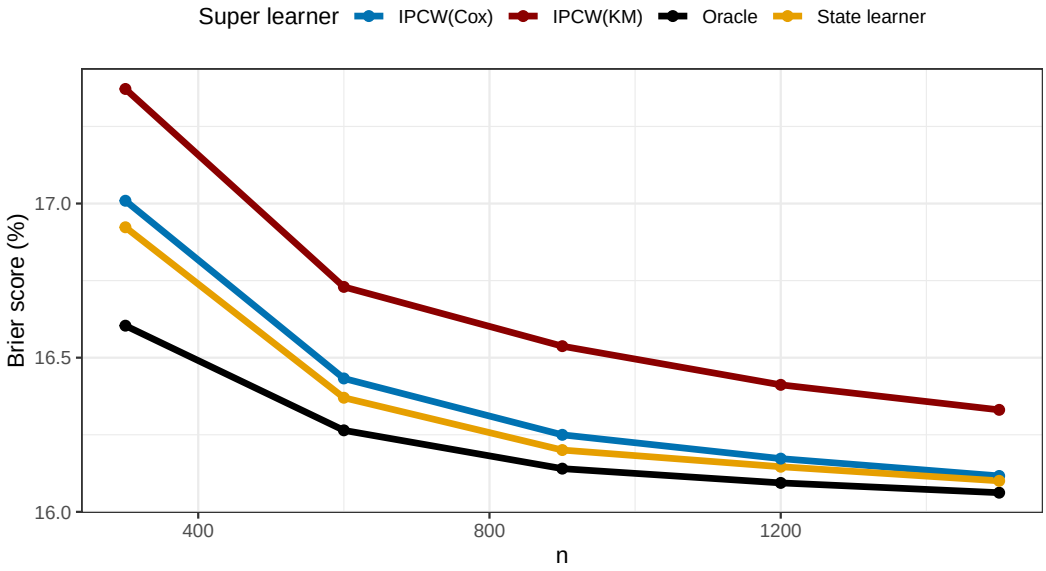


Figure 3: Results for scenario 2 of the simulation study. For the learner selected by each of the four discrete super learners, the Brier score calculated in a large independent data set without censoring is plotted against sample size. The results are based on 200 repetitions.

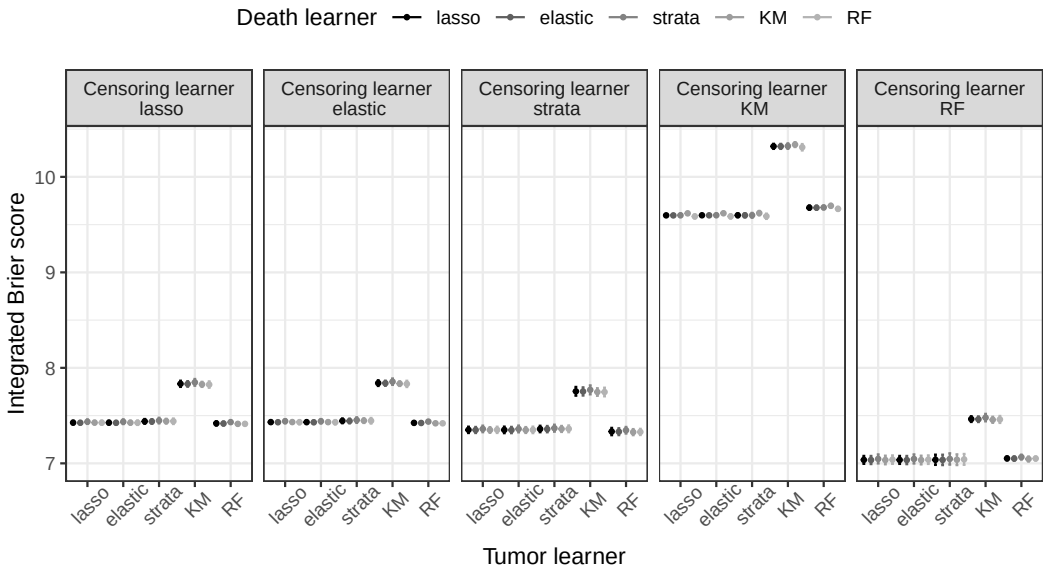


Figure 4: The results of applying the 125 combinations of learners to the prostate cancer data set. The learners are `KM` (KM), `Cox strata CS` (strata), `Lasso` (lasso), `Elastic` (elastic), and `RF` (RF) as described Table 1. The error bars are based on five repetitions using different splits. We refer to learners of Λ_1, Λ_2, and Γ as 'Tumor learner', 'Death learner', and 'Censoring learner', respectively.

Tumor learner	Death learner	Censoring learner	Integrated Brier score
Elastic	Elastic	RF	7.03 ± 0.02
Elastic	KM	RF	7.03 ± 0.02
Lasso	Elastic	RF	7.04 ± 0.02
Lasso	KM	RF	7.04 ± 0.02
Elastic	Lasso	RF	7.04 ± 0.02
Cox strata CT	Elastic	RF	7.04 ± 0.03
Lasso	Lasso	RF	7.04 ± 0.02
Cox strata CT	Lasso	RF	7.04 ± 0.03
Cox strata CT	KM	RF	7.04 ± 0.03
Elastic	RF	RF	7.04 ± 0.02

Table 2: The 10 best performing models in terms of integrated Brier score. The reported standard errors are based on five repetitions using different splits. The models are described in Table 1. We refer to learners of Λ_1, Λ_2, and Γ as 'Tumor learner', 'Death leaner', and 'Censoring learner', respectively.

9.1 Existing super learners for right-censored data

Machine learning based on right-censored data commonly use the negative partial log-likelihood as loss function [e.g., Li et al., 2016, Yao et al., 2017, Lee et al., 2018, Katzman et al., 2018, Gensheimer and Narasimhan, 2019, Lee et al., 2021, Kvamme and Borgan, 2021]. However, this loss function is unsuited for super learning, because many canonical survival learners (e.g., the Kaplan-Meier estimator, random survival forest, and semi-parametric Cox models) provide cumulative hazard functions that are piece-wise constant in the time argument, and hence assign zero probability to event times not observed in the training data. This implies that when data are observed in continuous time, any of these learners will almost surely have infinite loss in any independent hold-out sample according to the negative partial log-likelihood loss. When a proportional hazards model is assumed, the baseline hazard function can be profiled out of the likelihood to give a new partial log-likelihood loss [Cox, 1972], which has been suggested as a loss function for super learning [Golmakani and Polley, 2020, Verweij and van Houwelingen, 1993]. While this allows the library of learners to include Cox' proportional hazard models, the drawback is that the library is in fact *only* allowed to include these models. The advantage of the state learner is that it does not require evaluation of the density of $F(\cdot, j, x)$ and does not assume a particular semi-parametric structure for Λ_j but can be used with any library of learners.

Another approach for super learning with right-censored data is to use an inverse probability of censoring weighted (IPCW) loss function [Graf et al., 1999, van der Laan and Dudoit, 2003, Molinaro et al., 2004, Keles et al., 2004, Hothorn et al., 2006, Gerds and Schumacher, 2006, Gonzalez Ginestet et al., 2021]. An IPCW loss function is attractive because the associated risk does not depend on the censoring distribution but describes a feature of the population of interested governed by the measure $Q \in \mathcal{Q}$. Similar results can be obtained using censoring unbiased transformations [Fan and Gijbels, 1996, Steingrimsson et al., 2019] or pseudo-values [Andersen et al., 2003, Mogensen and Gerds, 2013]. All these methods rely on an estimator of the censoring distribution, and their drawback is that this estimator has to be pre-specified. When the data-generating mechanism is complex and not well-understood, pre-specification of the censoring distribution is a challenge. The advantage of using the state learner is that a censoring distribution need not be pre-specified but is selected automatically based on the provided library $\mathcal{B}$.

To the best of our knowledge, the only existing attempt at avoiding the need to pre-specify a censoring model is a recent proposal suggested independently by Han et al. [2021] and

Westling et al. [2021]. The authors do not consider competing risks but suggest to iterate between learning Λ and Γ using IPCW loss functions and select the final learner when the iterative procedure has converged. No general theoretical guarantees exist for this procedure, but it would be interesting to compare its performance to that of the state learner in a simulation study.

In a non-survival targeted learning setting Sun et al. [2022], based on ideas presented by Robins et al. [2007], proposed to use as a model selection criteria the so-called 'doubly robustness' property that some targeted estimators enjoy. In many special cases the remainder term $\text{Rem}(\hat{\Lambda}_n, \hat{\Gamma}_n, P)$ defined in equation (7) has the property that

$$\text{Rem}(\Lambda^*, \Gamma_0, P_0) = \text{Rem}(\Lambda_0, \Gamma^*, P_0) = 0, \quad \text{for any } \Lambda^*, \Gamma^* \in \mathcal{M}. \tag{9}$$

The property in equation (9) is known as 'doubly robustness' because it ensures that $\tilde{\Psi}(\hat{\Lambda}_n, \hat{\Gamma}_n, H_n)$ defined in equation (7) is consistent if just one of Λ or Γ is estimated consistently [van der Laan and Robins, 2003, Bang and Robins, 2005, Kang and Schafer, 2007]. For $a \in \mathcal{A}$ and $b \in \mathcal{B}$, define $\hat{\Psi}_n^k(a, b) = \tilde{\Psi}(a(\mathcal{D}_n^k), b(\mathcal{D}_n^k), \mathbb{P}_n^k)$ where $\mathbb{P}_n^k$ is the empirical measure of $\mathcal{D}_n^k$. Sun et al. [2022] suggest to select $a \in \mathcal{A}$ as the minimizer of the estimated 'fluctuation pseudo-risk',

$$\hat{R}_{\text{f}}(a) = \max_{b_1, b_2 \in \mathcal{B}} \frac{1}{K} \sum_{k=1}^{K} \left\{ \hat{\Psi}_n^k(a, b_1) - \hat{\Psi}_n^k(a, b_2) \right\}^2.$$

The idea is that the doubly robustness property (equation (9)) implies that $\hat{R}_{\text{f}}(a)$ will be zero (asymptotically) if the learner a correctly estimates Λ. The authors establish finite sample results guarantying that the selected learner a will be robust against changing the learner of the other nuisance parameter across $\mathcal{B}$. It seems unclear to what extend this robustness property also guarantees consistency of the estimator of the target parameters when we use the fluctuation risk $\hat{R}_{\text{f}}$ to select the learners of the nuisance parameters. It would be interesting to explore the performance of this procedure in a survival setting and compare it to targeted estimators obtained using the state learner as outlined in Section 6.

9.2 A performance measure of interest

A major advantage of the state learner is that performance of each combination of learners is measured in terms of observable quantities. This means that no additional nuisance parameters need to be estimated to evaluate the loss. The drawback with this approach is that we are rarely interested in features of the observed data distribution when the data are right-censored. The finite sample oracle inequality in Corollary 5.2 concerns the function F, which is a feature of $P \in \mathcal{P}$, while what we are typically interested in is Λ_j or S, which are features of $Q \in \mathcal{Q}$. We emphasize that while the state learner provides us with estimates of Λ_j and Γ based on libraries $\mathcal{A}_j$ and $\mathcal{B}$, performance are not assessed directly for these parameters, but only jointly for estimation of the parameter F. For settings without a competing risk, our numerical studies suggest that measuring performance with respect to estimation of F also leads to good performance for estimation of S. Further research on this topic, both numerical and theoretical, is warranted.

In the context of targeted learning, a drawback of the state learner is that the doubly robustness property defined in equation (9) seem to be lost because we only use a single nuisance parameter estimator. As defined here, however, the state learner is build using libraries for the conditional cause-specific cumulative hazard functions, so some doubly robustness might be preserved.

9.3 Implementation

Our proposed super learner can be implemented with a broad library of learners using existing software, for instance the `R`-package `riskRegression` [Gerds et al., 2023]. Furthermore, while the library $\mathcal{F}(\mathcal{A}_1, \mathcal{A}_2, \mathcal{B})$ consists of $|\mathcal{A}_1||\mathcal{A}_2||\mathcal{B}|$ learners, as long as we have sufficient memory we need only fit $|\mathcal{A}_1| + |\mathcal{A}_2| + |\mathcal{B}|$ learners in each fold. To evaluate the performance of each learner we need to perform $|\mathcal{A}_1||\mathcal{A}_2||\mathcal{B}|$ operations to calculate the integrated Brier score in each hold-out sample, one for each combination of the fitted models, but these operations are often negligible compared to fitting the models. Hence the state learner is essentially no more computationally demanding than any procedure that uses super learning to learn Λ_1, Λ_2, and Γ separately. While our proposal is based on constructing the library $\mathcal{F}$ from libraries for learning Λ_1, Λ_2, and Γ, it would also be of interest to consider learners that estimate F directly.

In our numerical studies, we only considered learners of Λ_j and Γ that provide cumulative hazard functions which are piece-wise constant in the time argument. This simplifies the calculation of F as the integrals in equation (5) reduce to sums. When Λ_j or Γ are absolutely continuous in the time argument, calculating F is more involved, but we expect that a good approximation can be achieved by discretization. In the future, we intend to investigate the performance of the state learner when using a broader library of learners in more comprehensive simulation studies.

A The oracle according to the partial log-likelihood loss

In this section we examine the oracle according to the partial log-likelihood loss and its relation to the censoring distribution. For simplicity we assume the special case without competing risks so that $\tilde{D} = \Delta$, and we abuse notation by writing $\Lambda = \Lambda_1$. We also simplify matters considerably by considering a setting with no baseline covariates, which is of course a severe limitation. However, as our Proposition A.1 is a 'negative result', we find that a generalization is perhaps not so interesting. We also simplify the exposition by consider an oracle with respect to a fixed collection of learners $\mathcal{R} \subset \mathcal{M}$ that does not depend on data. The collection $\mathcal{R}$ should be interpreted as the limit of a library of learners, which we might informally write as $\mathcal{R} = \{a(\mathcal{D}_\infty) : a \in \mathcal{A}\}$.

Let $\mathcal{M}_b \subset \mathcal{M}$ be the collection of all absolutely continuous cumulative hazard functions $\Lambda\colon [0, \tau] \to \mathbb{R}_+$ such that $0 < \Lambda(s) < \infty$ for all $s \in [0, \tau)$. As we have no baseline covariates, the measure $P_{\Lambda,\Gamma}$ is the distribution of $O = (\tilde{T}, \Delta) = (T \wedge C, \mathbb{1}\{T \leq C\})$ when T and C are independent draws from the measures induced by $\Lambda \in \mathcal{M}_b$ and $\Gamma \in \mathcal{M}_b$, respectively. The negative partial log-likelihood for the parameter $\Lambda \in \mathcal{M}_b$ is

$$L^{\mathrm{pl}}(\Lambda, O) = \int_0^{\tilde{T}} \lambda(s)\,\mathrm{d}s - \Delta \log \lambda(\tilde{T}), \quad \text{where} \quad \Lambda(\mathrm{d}s) = \lambda(s)\,\mathrm{d}s.$$

While L^{pl} does not depend on the censoring distribution, the oracle typically will. This is because the oracle is defined with respect to the data-generating distribution P which depends on the censoring distribution [Hjort, 1992, Whitney et al., 2019]. Recall that the Kullback-Leibler divergence between the probability measures P_1 and P_2 is defined as

$$D_{\mathrm{KL}}(P_1 \,||\, P_2) = P_1\left[\log \frac{p_1}{p_2}\right], \quad \text{where} \quad P_1 = p_1 \cdot \nu, \quad \text{and} \quad P_2 = p_2 \cdot \nu,$$

for some σ-finite measure ν such that $\{P_1, P_2\} \ll \nu$. Let ℓ denote the partial likelihood for

Γ. We observe that

$$\begin{aligned}
&P_{\Lambda_0,\Gamma_0}[L^{\mathrm{pl}}(\Lambda,\cdot)] \\
&= P_{\Lambda_0,\Gamma_0}[L^{\mathrm{pl}}(\Lambda,\cdot)] \pm P_{\Lambda_0,\Gamma_0}[\log \ell(\Gamma_0,\cdot) - L^{\mathrm{pl}}(\Lambda_0,\cdot)] \\
&= P_{\Lambda_0,\Gamma_0}[\log \ell(\Gamma_0,\cdot) - L^{\mathrm{pl}}(\Lambda_0,\cdot) - \{\log \ell(\Gamma_0,\cdot) - L^{\mathrm{pl}}(\Lambda,\cdot)\}] + P_{\Lambda_0,\Gamma_0}[L^{\mathrm{pl}}(\Lambda_0,\cdot)] \\
&= D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_0} \,||\, P_{\Lambda,\Gamma_0}) + P_{\Lambda_0,\Gamma_0}[L^{\mathrm{pl}}(\Lambda_0,\cdot)],
\end{aligned}$$

hence the oracle according to L^{pl} under P_{Λ_0,Γ_0} is equivalent to the minimizer of $\Lambda \mapsto D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_0} \,||\, P_{\Lambda,\Gamma_0})$. For any $\Gamma \in \mathcal{M}_b$ and sub-family $\mathcal{R} \subset \mathcal{M}_b$ define the oracle relative to Γ and $\mathcal{R}$ as

$$\tilde{a}(\Gamma, \mathcal{R}) = \operatorname*{argmin}_{\Lambda \in \mathcal{R}} D_{\mathrm{KL}}\left(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda,\Gamma}\right). \tag{10}$$

We can now state the main result of this section.

Proposition A.1. *Let $\Lambda_0 \in \mathcal{M}_b$, $\Gamma \in \mathcal{M}_b$, and $\mathcal{R}^* \subset \mathcal{M}_b$ be such that $\Lambda_0 \notin \mathcal{R}^*$ and $P_{\Lambda_0,\Gamma}[|L^{\mathrm{pl}}(\Lambda_0,\cdot) - L^{\mathrm{pl}}(\tilde{a}(\Gamma,\mathcal{R}^*),\cdot)|] < \infty$. Then there exist $\Lambda' \in \mathcal{M}_b$ and $\Gamma' \in \mathcal{M}_b$ such that*

$$\tilde{a}(\Gamma, \mathcal{R}^* \cup \{\Lambda'\}) \neq \tilde{a}(\Gamma', \mathcal{R}^* \cup \{\Lambda'\}).$$

Proof. In the following let $\Lambda_* = \tilde{a}(\Gamma, \mathcal{R})$. The proposition follows if we can find a $\Lambda' \in \mathcal{M}_b$ such that

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda',\Gamma}) > D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_*,\Gamma}), \tag{11}$$

and a $\Gamma' \in \mathcal{M}_b$ such that

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma'} \,||\, P_{\Lambda',\Gamma'}) < D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma'} \,||\, P_{\Lambda_*,\Gamma'}). \tag{12}$$

In the following we use $\mathrm{d}P_{\Lambda,\Gamma}$ to denote the density of $P_{\Lambda,\Gamma}$ with respect to the product of Lebesgue and counting measure, and we use $\mathbb{E}_{\Lambda,\Gamma}$ to denote the expectation under $P_{\Lambda,\Gamma}$. We can then for any $\Lambda \in \mathcal{M}_b$ write the Kullback-Leibler divergence as

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_0} \,||\, P_{\Lambda,\Gamma_0}) = \mathbb{E}_{\Lambda_0,\Gamma_0}\left[\Delta \log \frac{\lambda_0(\tilde{T})}{\lambda(\tilde{T})}\right] - \mathbb{E}_{\Lambda_0,\Gamma_0}\left[\Lambda_0(\tilde{T}) - \Lambda(\tilde{T})\right], \tag{13}$$

where we use λ to denote the hazard corresponding to the cumulative hazard Λ. Because the likelihood for the observed data O factorises we have

$$\begin{aligned}
\mathbb{E}_{\Lambda_0,\Gamma}\left[\left|\log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right|\right] &= \mathbb{E}_{\Lambda_0,\Gamma}\left[\left|\log \frac{\ell(\Lambda_0,O)}{\ell(\Lambda_*,O)}\right|\right] \\
&= P_{\Lambda_0,\Gamma}\left[\left|L^{\mathrm{pl}}(\Lambda_0,\cdot) - L^{\mathrm{pl}}(\Lambda_*,\cdot)\right|\right] < \infty,
\end{aligned} \tag{14}$$

where the last inequality follows because we assumed that $\Gamma \in \mathcal{M}_b$ was chosen such that $P_{\Lambda_0,\Gamma}[|L^{\mathrm{pl}}(\Lambda_0,\cdot) - L^{\mathrm{pl}}(\Lambda_*,\cdot)|] < \infty$.

To find Λ' and Γ' that satisfy equations (11) and (12) we construct parametric families $\{\Lambda_\beta\} \subset \mathcal{M}_b$ and $\{\Gamma_\alpha\} \subset \mathcal{M}_b$, and show that we can find Λ' and Γ' within these families. First we pick a time $t_1 \in (0,\tau)$ such that

$$\mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \leq u\} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] \geq c > 0, \tag{15}$$

for all $u \in [t_1,\tau)$. This is possible because equation (14) and dominated convergence allows us to conclude that

$$\lim_{u \uparrow \tau} \mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \leq u\} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] = D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_*,\Gamma}) > 0,$$

where the inequality at the end holds because $\Lambda_* \neq \Lambda_0$ by assumption. Next, we pick some $t_2 \in (t_1, \tau)$ and $\varepsilon > 0$ such that $t_2 + \varepsilon < \tau$, and define the function

$$h_\beta(s) = \begin{cases} \beta^{-1} & \text{if } u \in [t_2, t_2 + \varepsilon] \\ 1 & \text{if } u \notin [t_2, t_2 + \varepsilon] \end{cases}, \quad \text{for all} \quad \beta > 1.$$

Let Λ_β be the cumulative hazard function with derivative $\lambda_0 h_\beta$. Since the hazard of Λ_0 and Λ_β are identical on $[0, t_2) \cup (t_2 + \varepsilon, \infty)$, equation (13) yields for any $\Gamma \in \mathcal{M}_b$

$$\begin{aligned} D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_\beta,\Gamma}) &= \mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \in [t_2, t_2+\varepsilon]\}\Delta \log \frac{\lambda_0(\tilde{T})}{\lambda_\beta(\tilde{T})}\right] \\ &\quad - \mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \geq t_2\}(\Lambda_0(\tilde{T}) - \Lambda_\beta(\tilde{T}))\right] \\ &= \log(\beta)\mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \in [t_2, t_2+\varepsilon]\}\Delta\right] \\ &\quad - \mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \geq t_2\}(\Lambda_0(\tilde{T}) - \Lambda_\beta(\tilde{T}))\right]. \end{aligned} \tag{16}$$

For $s \in [t_2, t_2 + \varepsilon]$ we have

$$\Lambda_0(s) - \Lambda_\beta(s) = \int_{t_2}^{s} \left\{\lambda_0(u) - \beta^{-1}\lambda_0(u)\right\} \mathrm{d}u = (1 - \beta^{-1})[\Lambda_0(s) - \Lambda_0(t_2)],$$

and for $s > t_2 + \varepsilon$ we have

$$\Lambda_0(s) - \Lambda_\beta(s) = \int_{t_2}^{t_2+\varepsilon} \left\{\lambda_0(u) - \beta^{-1}\lambda_0(u)\right\} \mathrm{d}u = (1 - \beta^{-1})[\Lambda_0(t_2 + \varepsilon) - \Lambda_0(t_2)],$$

so

$$\begin{aligned} &\mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \geq t_2\}(\Lambda_0(\tilde{T}) - \Lambda_\beta(\tilde{T}))\right] \\ &\qquad = (1 - \beta^{-1})\Big\{\mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} \in [t_2, t_2+\varepsilon]\}\Lambda_0(\tilde{T})] \\ &\qquad\quad + \mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} > t_2 + \varepsilon\}\Lambda_0(t_2 + \varepsilon)] \\ &\qquad\quad - \mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} \geq t_2\}\Lambda_0(t_2)]\Big\}. \end{aligned} \tag{17}$$

As Λ_0 is non-decreasing it holds that

$$\begin{aligned} \mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} \in [t_2, t_2+\varepsilon]\}\Lambda_0(\tilde{T})] &\leq \Lambda_0(t_2 + \varepsilon)\mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} \in [t_2, t_2+\varepsilon]\}] \\ &\leq \Lambda_0(t_2 + \varepsilon)P_{\Lambda_0,\Gamma}(\tilde{T} \geq t_2). \end{aligned} \tag{18}$$

Using again that Λ_0 is non-decreasing and positive we have

$$\begin{aligned} &\left|\mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} > t_2 + \varepsilon\}\Lambda_0(t_2 + \varepsilon)] - \mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} \geq t_2\}\Lambda_0(t_2)]\right| \\ &\leq \max\left\{\mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} > t_2 + \varepsilon\}\Lambda_0(t_2 + \varepsilon)], \mathbb{E}_{\Lambda_0,\Gamma}[\mathbb{1}\{\tilde{T} \geq t_2\}\Lambda_0(t_2)]\right\} \\ &= \max\left\{P_{\Lambda_0,\Gamma}(\tilde{T} > t_2 + \varepsilon)\Lambda_0(t_2 + \varepsilon), P_{\Lambda_0,\Gamma}(\tilde{T} \geq t_2)\Lambda_0(t_2)\right\} \\ &\leq \Lambda_0(t_2 + \varepsilon)P_{\Lambda_0,\Gamma}(\tilde{T} \geq t_2). \end{aligned} \tag{19}$$

Equations (17), (18), and (19) then imply

$$\begin{aligned} \left|\mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \geq t\}(\Lambda_0(\tilde{T}) - \Lambda_\beta(\tilde{T}))\right]\right| &\leq (1 - \beta^{-1})2\Lambda_0(t_2 + \varepsilon)P_{\Lambda_0,\Gamma}(\tilde{T} \geq t_2) \\ &\leq 2\Lambda_0(t_2 + \varepsilon)P_{\Lambda_0,\Gamma}(\tilde{T} \geq t_2), \end{aligned} \tag{20}$$

where we used that $(1-\beta^{-1}) \in (0,1)$. In the following we use O-notation to bound a function in ε close to zero, i.e., $f(\varepsilon) = O(g(\varepsilon))$ means that there exist $\varepsilon_0 > 0$ and $M < \infty$ such that

$$|f(\varepsilon)| \leq Mg(\varepsilon) \quad \text{for all} \quad \varepsilon < \varepsilon_0.$$

Thus by equations (16) and (20) we can write

$$\begin{aligned} &D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_\beta,\Gamma}) \\ &\quad = \log(\beta) P_{\Lambda_0,\Gamma}(\tilde{T} \in [t_2, t_2+\varepsilon], \Delta = 1) + O\Big\{\Lambda_0(t_2+\varepsilon) P_{\Lambda_0,\Gamma}(\tilde{T} \geq t_2)\Big\}, \end{aligned} \tag{21}$$

for any $\Gamma \in \mathcal{M}_b$ and $\beta > 1$. As $\Lambda_0 \in \mathcal{M}_b$ and $\Gamma \in \mathcal{M}_b$ it follows that $P_{\Lambda_0,\Gamma}(\tilde{T} \in [t_2, t_2 + \varepsilon], \Delta = 1) > 0$ and thus equation (21) implies

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_\beta,\Gamma}) \to \infty, \quad \text{when} \quad \beta \to \infty.$$

By equation (14) we have that $D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_*,\Gamma}) < \infty$, which together with the previous display implies that we can find $\beta_2 > 1$ such that

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_{\beta_2},\Gamma}) > D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma} \,||\, P_{\Lambda_*,\Gamma}). \tag{22}$$

If we choose the model Λ' to be Λ_{β_2}, equation (22) shows that equation (11) holds, providing the first part of the proof. The next step is to construct the censoring mechanism Γ'. To do this, let F be the probability measure on $[0,\tau]$ uniquely determined by the cumulative hazard function Γ and let f denote the Lebesgue density of F. Define $\bar{\alpha} = \{F(C \in [t_1, t_2) \mid C \geq t_1)\}^{-1}$ and the function $f_\alpha \colon [0,\tau] \to \mathbb{R}_+$ as

$$f_\alpha(s) = \begin{cases} f(s) & \text{if } s < t_1 \\ \alpha f(s) & \text{if } s \in [t_1, t_2) \\ c(\alpha) f(s) & \text{if } s \geq t_2 \end{cases}, \quad \text{with} \quad c(\alpha) = \frac{F(C \geq t_1) - \alpha F(C \in [t_1, t_2))}{F(C \geq t_2)},$$

for $\alpha \in [1, \bar{\alpha}]$. Note that $\bar{\alpha} > 1$ because F has support on $[0,\tau]$. By construction, for any $\alpha \in [1,\bar{\alpha}]$, f_α defines a density, and we let F_α denote the measure with this density and Γ_α the corresponding cumulative hazard function. Note that $\Gamma_\alpha \in \mathcal{M}_b$ for all $\alpha \in [1,\bar{\alpha})$ while $\Gamma_{\bar{\alpha}} \notin \mathcal{M}_b$ because $f_{\bar{\alpha}}$ is equal to 0 on $(t_2, \tau]$ and thus $F_{\bar{\alpha}}$ does not have support on all of $[0,\tau]$. By Lemma A.2 below,

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_\alpha} \,||\, P_{\Lambda_*,\Gamma_\alpha}) \longrightarrow D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_{\bar{\alpha}}} \,||\, P_{\Lambda_*,\Gamma_{\bar{\alpha}}}) \quad \text{when} \quad \alpha \uparrow \bar{\alpha}. \tag{23}$$

As the measures $P_{\Lambda_0,\Gamma}$ and $P_{\Lambda_0,\Gamma_{\bar{\alpha}}}$ agree on $(0,t_1)$, and the measures $P_{\Lambda_*,\Gamma}$ and $P_{\Lambda_*,\Gamma_{\bar{\alpha}}}$ agree on $(0,t_1)$, equation (15) implies that the measures $P_{\Lambda_0,\Gamma_{\bar{\alpha}}}$ and $P_{\Lambda_*,\Gamma_{\bar{\alpha}}}$ must be different. In particular, we have that $D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_{\bar{\alpha}}} \,||\, P_{\Lambda_*,\Gamma_{\bar{\alpha}}}) > 0$, and hence equation (23) implies that we can find an $\alpha' \in (1,\bar{\alpha})$ such that

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_\alpha} \,||\, P_{\Lambda_*,\Gamma_\alpha}) \geq c' > 0, \quad \text{for all} \quad \alpha \in (\alpha', \bar{\alpha}). \tag{24}$$

By construction of Γ_α we have that $P_{\Lambda_0,\Gamma_\alpha}(\tilde{T} \geq t_2) \to 0$ when $\alpha \uparrow \bar{\alpha}$, and hence equation (21) gives

$$\begin{aligned} &D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_\alpha} \,||\, P_{\Lambda_{\beta_2},\Gamma_\alpha}) \\ &= \log(\beta_2) P_{\Lambda_0,\Gamma_\alpha}(\tilde{T} \in [t_2, t_2+\varepsilon_2], \Delta = 1) + O\Big\{\Lambda_0(t_2+\varepsilon_2) P_{\Lambda_0,\Gamma_\alpha}(\tilde{T} \geq t_2)\Big\} \\ &\longrightarrow 0, \quad \text{when} \quad \alpha \uparrow \bar{\alpha}, \end{aligned} \tag{25}$$

where we used that $P_{\Lambda_0,\Gamma_\alpha}(\tilde{T} \in [t_2, t_2+\varepsilon_2], \Delta = 1) \leq P_{\Lambda_0,\Gamma_\alpha}(\tilde{T} \geq t_2)$. Finally, equations (24) and (25) imply that we can find $\alpha_2 \in (\alpha', \bar{\alpha})$ such that

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_{\alpha_2}} \,||\, P_{\Lambda_{\beta_2},\Gamma_{\alpha_2}}) < D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_{\alpha_2}} \,||\, P_{\Lambda_*,\Gamma_{\alpha_2}}).$$

Choosing the alternative censoring mechanism Γ' to be Γ_{α_2} the inequality above gives the second part of the proof. □

To complete the proof it now remains to prove the following lemma.

Lemma A.2. *Let Γ_α and $\bar{\alpha}$ be as defined above. Then*

$$D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_\alpha} \,||\, P_{\Lambda_*,\Gamma_\alpha}) \longrightarrow D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_{\bar{\alpha}}} \,||\, P_{\Lambda_*,\Gamma_{\bar{\alpha}}}) \quad \textit{when} \quad \alpha \uparrow \bar{\alpha}.$$

Proof. For the measure F, let $\bar{F}$ denote the survivor function and use the factorisation of the likelihood to write

$$\begin{aligned}
D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_\alpha} \,||\, P_{\Lambda_*,\Gamma_\alpha}) &= \mathbb{E}_{\Lambda_0,\Gamma_\alpha}\left[\log \frac{\mathrm{d}P_{\Lambda_0,\Gamma_\alpha}}{\mathrm{d}P_{\Lambda_*,\Gamma_\alpha}}(\tilde{T},\Delta)\right] \\
&= \mathbb{E}_{\Lambda_0,\Gamma_\alpha}\left[\log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] \\
&= \mathbb{E}_{\Lambda_0,\Gamma_\alpha}\left[\frac{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] \\
&= \mathbb{E}_{\Lambda_0,\Gamma}\left[\frac{\{f_\alpha(\tilde{T})\}^{1-\Delta}\{\bar{F}_\alpha(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right].
\end{aligned}$$

From this we obtain

$$\begin{aligned}
&D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_\alpha} \,||\, P_{\Lambda_*,\Gamma_\alpha}) \\
&= \mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} < t_2\}\frac{\{f_\alpha(\tilde{T})\}^{1-\Delta}\{\bar{F}_\alpha(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] \\
&\quad + \mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \geq t_2\}\frac{\{f_\alpha(\tilde{T})\}^{1-\Delta}\{\bar{F}_\alpha(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right].
\end{aligned} \tag{26}$$

Consider the first term of the right hand side of equation (26). By construction of f_α we have

$$\begin{aligned}
&\left|\mathbb{1}\{\tilde{T} < t_2\}\frac{\{f_\alpha(\tilde{T})\}^{1-\Delta}\{\bar{F}_\alpha(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right| \\
&\quad \leq \frac{\alpha}{\bar{F}(t_2)}\left|\log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right| \\
&\quad \leq \frac{\bar{\alpha}}{\bar{F}(t_2)}\left|\log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right|,
\end{aligned} \tag{27}$$

where we used that $\bar{F}_\alpha$ is bounded by 1 and $1/\bar{F}$ is non-decreasing. Because F has support on $(0,\tau)$ we have that $\bar{F}(t_2) > 0$, and thus it follows from equations (14) and (27) and dominated convergence that

$$\begin{aligned}
&\mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} < t_2\}\frac{\{f_\alpha(\tilde{T})\}^{1-\Delta}\{\bar{F}_\alpha(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] \\
&\quad \longrightarrow \mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} < t_2\}\frac{\{f_{\bar{\alpha}}(\tilde{T})\}^{1-\Delta}\{\bar{F}_{\bar{\alpha}}(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] \\
&\quad = \mathbb{E}_{\Lambda_0,\Gamma_{\bar{\alpha}}}\left[\mathbb{1}\{\tilde{T} < t_2\} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right] \\
&\quad = \mathbb{E}_{\Lambda_0,\Gamma_{\bar{\alpha}}}\left[\mathbb{1}\{\tilde{T} < t_2\} \log \frac{\mathrm{d}P_{\Lambda_0,\Gamma_{\bar{\alpha}}}}{\mathrm{d}P_{\Lambda_*,\Gamma_{\bar{\alpha}}}}(\tilde{T},\Delta)\right], \\
&\quad = \mathbb{E}_{\Lambda_0,\Gamma_{\bar{\alpha}}}\left[\log \frac{\mathrm{d}P_{\Lambda_0,\Gamma_{\bar{\alpha}}}}{\mathrm{d}P_{\Lambda_*,\Gamma_{\bar{\alpha}}}}(\tilde{T},\Delta)\right] \\
&\quad = D_{\mathrm{KL}}(P_{\Lambda_0,\Gamma_{\bar{\alpha}}} \,||\, P_{\Lambda_*,\Gamma_{\bar{\alpha}}}),
\end{aligned}$$

when $\alpha \uparrow \bar{\alpha}$. The second to last equality follows because $P_{\Lambda_0,\Gamma_{\bar{\alpha}}}$ has support on $(0, t_2)$, and the last equality follows by definition. Thus the result follows if we can show that the second term of the right hand side of equation (26) goes to zero when $\alpha \uparrow \bar{\alpha}$. To see this, first note that

$$\frac{\{f_\alpha(\tilde{T})\}^{1-\Delta}\{\bar{F}_\alpha(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}} = \frac{\{c(\alpha)f(\tilde{T})\}^{1-\Delta}\{c(\alpha)\int_{\tilde{T}}^{\tau} f(s)\,\mathrm{d}s\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\int_{\tilde{T}}^{\tau} f(s)\,\mathrm{d}s\}^{\Delta}} = c(\alpha).$$

This gives

$$\begin{aligned}&\left|\mathbb{E}_{\Lambda_0,\Gamma}\left[\mathbb{1}\{\tilde{T} \geq t_2\}\frac{\{f_\alpha(\tilde{T})\}^{1-\Delta}\{\bar{F}_\alpha(\tilde{T})\}^{\Delta}}{\{f(\tilde{T})\}^{1-\Delta}\{\bar{F}(\tilde{T})\}^{\Delta}}\log\frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right]\right| \\ &\quad < c(\alpha)\mathbb{E}_{\Lambda_0,\Gamma}\left[\left|\log\frac{\mathrm{d}P_{\Lambda_0,\Gamma}}{\mathrm{d}P_{\Lambda_*,\Gamma}}(\tilde{T},\Delta)\right|\right] \longrightarrow 0,\end{aligned}$$

when $\alpha \uparrow \bar{\alpha}$ by equation (14) and the definition of $c(\alpha)$. This proves the lemma. □

B Theoretical guarantees for the state learner

In this section we provide proofs of the results stated in Section 5.

Define $\bar{B}_{\tau,0}(F, o) = \bar{B}_\tau(F, o) - \bar{B}_\tau(F_0, o)$ and $R_0(F) = P_0[\bar{B}_{\tau,0}(F, \cdot)]$.

Lemma B.1. *$R_0(F) = \|F - F_0\|_{P_0}^2$, where $\|\cdot\|_{P_0}$ is defined in equation (6).*

Proof. For any $t \in [0, \tau]$ and $k \in \{0, 1, 2\}$ we have

$$\begin{aligned}&\mathbb{E}_{P_0}\left[(F(t,k,X) - \mathbb{1}\{\eta(t) = k\})^2\right] \\ &= \mathbb{E}_{P_0}\left[(F(t,k,X) - F_0(t,k,X) + F_0(t,k,X) - \mathbb{1}\{\eta(t) = k\})^2\right] \\ &= \mathbb{E}_{P_0}\left[(F(t,k,X) - F_0(t,k,X))^2\right] + \mathbb{E}_{P_0}\left[(F_0(t,k,X) - \mathbb{1}\{\eta(t) = k\})^2\right] \\ &\quad + 2\mathbb{E}_{P_0}[(F(t,k,X) - F_0(t,k,X))(F_0(t,k,X) - \mathbb{1}\{\eta(t) = k\})] \\ &= \mathbb{E}_{P_0}\left[(F(t,k,X) - F_0(t,k,X))^2\right] + \mathbb{E}_{P_0}\left[(F_0(t,k,X) - \mathbb{1}\{\eta(t) = k\})^2\right],\end{aligned}$$

where the last equality follows from the tower property. Hence, using Fubini, we have

$$P[\bar{B}_\tau(F, \cdot)] = \|F - F_0\|_{P_0}^2 + P_0[\bar{B}_\tau(F_0, \cdot)].$$

□

Proof of Proposition 5.1. The result follows from Lemma B.1. □

In the following, let Θ denote the function space consisting of all conditional state occupation probability functions for some measure P as defined in equation (4).

Proof of Corollary 5.2. First note that minimising the loss $\bar{B}_\tau$ is equivalent to minimising the loss $\bar{B}_{\tau,0}$, so the discrete super learner and oracle according to $\bar{B}_\tau$ and $\bar{B}_{\tau,0}$ are identical. By Lemma B.1, $R_0(F) \geq 0$ for any $F \in \Theta$, and so using Theorem 2.3 from [van der Vaart et al., 2006] with $p = 1$, we have that for all $\delta > 0$,

$$\begin{aligned}\mathbb{E}_{P_0}\left[R_0(\hat{\varphi}_n(\mathcal{D}_n^{-k}))\right] \leq &(1 + 2\delta)\mathbb{E}_{P_0}\left[R_0(\tilde{\varphi}_n(\mathcal{D}_n^{-k}))\right] \\ &+ (1 + \delta)\frac{16K}{n}\log(1 + |\mathcal{F}_n|)\sup_{F\in\Theta}\left\{M(F) + \frac{v(F)}{R_0(F)}\left(\frac{1}{\delta} + 1\right)\right\}\end{aligned}$$

where for each $F \in \Theta$, $(M(F), v(F))$ is some Bernstein pair for the function $o \mapsto \bar{B}_{\tau,0}(F, o)$. As $\bar{B}_{\tau,0}(F, \cdot)$ is uniformly bounded by τ for any $F \in \Theta$, it follows from section 8.1 in [van der Vaart et al., 2006] that $(\tau, 1.5P_0[\bar{B}_{\tau,0}(F, \cdot)^2])$ is a Bernstein pair for $\bar{B}_{\tau,0}(F, \cdot)$. Now, for any $a, b, c \in \mathbb{R}$ we have

$$\begin{aligned}(a-c)^2 - (b-c)^2 &= (a-b+b-c)^2 - (b-c)^2 \\ &= (a-b)^2 + (b-c)^2 + 2(b-c)(a-b) - (b-c)^2 \\ &= (a-b)\{(a-b) + 2(b-c)\} \\ &= (a-b)\{a+b-2c\},\end{aligned}$$

so using this with $a = F(t,k,x)$, $b = F_0(t,k,x)$, and $c = \mathbb{1}\{\eta(t) = k\}$, we have by Jensen's inequality

$$\begin{aligned}&P_0[\bar{B}_{\tau,0}(F,\cdot)^2] \\ &\leq 2\tau \mathbb{E}_{P_0}\left[\sum_{k=0}^{3}\int_0^\tau \left\{(F(t,k,X) - \mathbb{1}\{\eta(t)=k\})^2 - (F_0(t,k,X) - \mathbb{1}\{\eta(t)=k\})^2\right\}^2 \mathrm{d}t\right] \\ &= 2\tau \mathbb{E}_{P_0}\left[\sum_{k=0}^{3}\int_0^\tau (F(t,k,X) - F_0(t,k,X))^2 \right. \\ &\qquad\qquad \left. \times \{F(t,k,X) + F_0(t,k,X) - 2\mathbb{1}\{\eta(t)=k\}\}^2 \mathrm{d}t\right] \\ &\leq 8\tau \mathbb{E}_{P_0}\left[\sum_{k=0}^{3}\int_0^\tau (F(t,k,X) - F_0(t,k,X))^2 \mathrm{d}t\right]. \\ &= 8\tau \|F - F_0\|_{P_0}^2.\end{aligned}$$

Thus when $v(F) = 1.5P_0[\bar{B}_{\tau,0}(F,\cdot)^2]$ we have by Lemma B.1

$$\frac{v(F)}{R_0(F)} = 1.5\frac{P_0[\bar{B}_{\tau,0}(F,\cdot)^2]}{P_0[\bar{B}_{\tau,0}(F,\cdot)]} \leq 12\tau,$$

and so using the Bernstein pairs $(\tau, 1.5P_0[\bar{B}_{\tau,0}(F,\cdot)^2])$ we have

$$\sup_{F\in\Theta}\left\{M(F) + \frac{v(F)}{R_0(F)}\left(\frac{1}{\delta}+1\right)\right\} \leq \tau\left(13 + \frac{12}{\delta}\right),$$

For all $\delta > 0$ we thus have

$$\begin{aligned}\mathbb{E}_{P_0}\left[R_0(\hat{\varphi}_n(\mathcal{D}_n^{-k}))\right] \leq &(1+2\delta)\mathbb{E}_{P_0}\left[R_0(\tilde{\varphi}_n(\mathcal{D}_n^{-k}))\right] \\ &+ (1+\delta)\log(1+|\mathcal{F}_n|)\tau\frac{16K}{n}\left(13+\frac{12}{\delta}\right),\end{aligned}$$

and then the final result follows from Lemma B.1. □

Proof of Corollary 5.3. By definition of the oracle and Lemma B.1, $\mathbb{E}_{P_0}\left[\|\tilde{\varphi}_n(\mathcal{D}_n^{-k}) - F_0\|_{P_0}^2\right] \leq \mathbb{E}_{P_0}\left[\|\varphi_n(\mathcal{D}_n^{-k}) - F_0\|_{P_0}^2\right]$ for all $n \in \mathbb{N}$. The results then follows from Corollary 5.2. □

C The state learner with targeted learning

In this section we give an explicit formula for obtaining of Λ and Γ from F and show that a product structure is preserved when the estimator $\bar{\Psi}(\hat{F}_n, \hat{H}_n)$ is used instead of $\tilde{\Psi}(\hat{\Lambda}_n, \hat{\Gamma}_n, \hat{H}_n)$.

By equations (2) and (3) and the definition of F, we have

$$\Lambda_0(t,x) = \int_0^t \frac{F_0(\mathrm{d}s,1,x)}{F_0(s-,0,x)}, \quad \text{and} \quad \Gamma_0(t,x) = \int_0^t \frac{F_0(\mathrm{d}s,2,x)}{F_0(s-,0,x)}, \tag{28}$$

where we use a subscript to denote that all parameters are associated to the same measure P_0.

Proof of Proposition 6.1. For notational convenience we suppress X in the following. The final result can be obtained by adding the argument X to all functions and averaging. We use the relations from equation (28) to write

$$\begin{aligned}
&\int_0^\tau w(s)\left\{\Gamma(s) - \hat{\Gamma}_n(s)\right\}[\Lambda - \hat{\Lambda}_n](\mathrm{d}s) \\
&= \int_0^\tau w(s)\left\{\int_0^s \frac{F(\mathrm{d}u,2)}{F(u-,0)} - \int_0^s \frac{\hat{F}_n(\mathrm{d}u,2)}{\hat{F}_n(u-,0)} -\right\}\left[\frac{F(\mathrm{d}s,1)}{F(s-,0)} - \frac{\hat{F}_n(\mathrm{d}s,1)}{\hat{F}_n(s-,0)}\right] \\
&= \int_0^\tau w(s)\Bigg\{\int_0^s \left(\frac{1}{F(u-,0)} - \frac{1}{\hat{F}_n(u-,0)}\right)F(\mathrm{d}u,2) \\
&\qquad\qquad + \int_0^s \frac{1}{\hat{F}_n(u-,0)}\left[F(\mathrm{d}u,2) - \hat{F}_n(\mathrm{d}u,2)\right]\Bigg\} \\
&\qquad\quad \times \left[\left(\frac{1}{F(s-,0)} - \frac{1}{\hat{F}_n(s-,0)}\right)F(\mathrm{d}s,1) + \frac{1}{\hat{F}_n(s-,0)}\left(F(\mathrm{d}s,1) - \hat{F}_n(\mathrm{d}s,1)\right)\right] \\
&= \int_0^\tau \int_0^s w(s)\left(\frac{1}{F(u-,0)} - \frac{1}{\hat{F}_n(u-,0)}\right)\left(\frac{1}{F(s-,0)} - \frac{1}{\hat{F}_n(s-,0)}\right)F(\mathrm{d}u,2)F(\mathrm{d}s,1) \\
&\quad + \int_0^\tau \int_0^s w(s)\left(\frac{1}{F(u-,0)} - \frac{1}{\hat{F}_n(u-,0)}\right)\frac{F(\mathrm{d}u,2)}{\hat{F}_n(u-,0)}\left(F(\mathrm{d}s,1) - \hat{F}_n(\mathrm{d}s,1)\right) \\
&\quad + \int_0^\tau \int_0^s \frac{w(s)}{\hat{F}_n(u-,0)}\left[F(\mathrm{d}u,2) - \hat{F}_n(\mathrm{d}u,2)\right]\left(\frac{1}{F(s-,0)} - \frac{1}{\hat{F}_n(s-,0)}\right)F(\mathrm{d}s,1) \\
&\quad + \int_0^\tau \int_0^s \frac{w(s)}{\hat{F}_n(u-,0)}\left[F(\mathrm{d}u,2) - \hat{F}_n(\mathrm{d}u,2)\right]\frac{1}{\hat{F}_n(s-,0)}\left(F(\mathrm{d}s,1) - \hat{F}_n(\mathrm{d}s,1)\right).
\end{aligned}$$

Consider the first term on the right hand side. Defining

$$w_n^*(t) = \left(F(t-,0) - \hat{F}_n(t-,0)\right)\left(\frac{1}{F(t-,0)} - \frac{1}{\hat{F}_n(t-,0)}\right),$$

we can write

$$\begin{aligned}
&\int_0^\tau \int_0^s w(s)\left(\frac{1}{F(u-,0)} - \frac{1}{\hat{F}_n(u-,0)}\right)\left(\frac{1}{F(s-,0)} - \frac{1}{\hat{F}_n(s-,0)}\right)F(\mathrm{d}u,2)F(\mathrm{d}s,1) \\
&= \int_0^\tau \int_0^s w(s)w_n^*(u)\left(F(u-,0) - \hat{F}_n(u-,0)\right) \\
&\qquad\qquad \times w_n^*(s)\left(F(s-,0) - \hat{F}_n(s-,0)\right)F(\mathrm{d}u,2)F(\mathrm{d}s,1) \\
&= \int_0^\tau \int_0^s w_n^a(s,u)\left(F(u-,0) - \hat{F}_n(u-,0)\right)\left(F(s-,0) - \hat{F}_n(s-,0)\right)F(\mathrm{d}u,2)F(\mathrm{d}s,1),
\end{aligned}$$

where we have defined $w_n^a(s,u) = w(s)w_n^*(s)w_n^*(u)$. By assumption, $w_n^a(s,u)$ is uniformly bounded. The same approach can be applied to the three remaining terms which gives the result. □

References

P. K. Andersen, J. P. Klein, and S. Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

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

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

J. M. Begun, W. J. Hall, W.-M. Huang, and J. A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, 11(2):432–452, 1983.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

L. Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

G. W. Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Routledge, 1996.

A. Fisher and E. H. Kennedy. Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2):162–172, 2021.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01. URL `https://www.jstatsoft.org/v33/i01/`.

M. F. Gensheimer and B. Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

T. A. Gerds. *prodlim: Product-Limit Estimation for Censored Event History Analysis*, 2019. URL `https://CRAN.R-project.org/package=prodlim`. R package version 2019.11.13.

T. A. Gerds and M. W. Kattan. *Medical risk prediction models: with ties to machine learning*. CRC Press, 2021.

T. A. Gerds and M. Schumacher. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006.

T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–2184, 2013.

T. A. Gerds, J. S. Ohlendorff, and B. Ozenne. *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks*, 2023. URL `https://CRAN.R-project.org/package=riskRegression`. R package version 2023.03.22.

R. D. Gill, M. J. van der Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

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

M. K. Golmakani and E. C. Polley. Super learner for survival data prediction. *The International Journal of Biostatistics*, 16(2):20190065, 2020.

P. Gonzalez Ginestet, A. Kotalik, D. M. Vock, J. Wolfson, and E. E. Gabriel. Stacked inverse probability of censoring weighted bagging: A case study in the infcarehiv register. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(1):51–65, 2021.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

X. Han, M. Goldstein, A. Puli, T. Wies, A. Perotte, and R. Ranganath. Inverse-weighted survival games. *Advances in Neural Information Processing Systems*, 34, 2021.

N. L. Hjort. On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique*, pages 355–387, 1992.

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

H. Ishwaran and U. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2023. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 3.2.2.

J. D. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 2007.

M. W. Kattan, M. J. Zelefsky, P. A. Kupelian, P. T. Scardino, Z. Fuks, and S. A. Leibel. Pretreatment nomogram for predicting the outcome of three-dimensional conformal radiotherapy in prostate cancer. *Journal of clinical oncology*, 18(19):3352–3359, 2000.

J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018.

S. Keles, M. van der Laan, and S. Dudoit. Asymptotically optimal model selection method with right censored outcomes. *Bernoulli*, 10(6):1011–1037, 2004.

H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, 2021.

C. Lee, W. Zame, J. Yoon, and M. van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

Y. Li, K. S. Xu, and C. K. Reddy. Regularized parametric regression for high-dimensional survival analysis. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 765–773. SIAM, 2016.

U. B. Mogensen and T. A. Gerds. A random forest approach for competing risks based on pseudo-values. *Statistics in medicine*, 32(18):3102–3114, 2013.

A. M. Molinaro, S. Dudoit, and M. J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1): 154–177, 2004.

E. C. Polley and M. J. van der Laan. Super learning for right-censored data. In M. J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 249–258. Springer, 2011.

J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.

H. C. Rytgaard and M. J. van der Laan. Targeted maximum likelihood estimation for causal inference in survival and competing risks analysis. *Lifetime Data Analysis*, pages 1–30, 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*, 2021.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5): 1–13, 2011. doi: 10.18637/jss.v039.i05. URL `https://www.jstatsoft.org/v39/i05/`.

J. A. Steingrimsson, L. Diao, and R. L. Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(525):370–383, 2019.

B. Sun, Y. Cui, and E. T. Tchetgen. Selective machine learning of the average treatment effect with an invalid instrumental variable. *Journal of Machine Learning Research*, 23 (204):1–40, 2022.

T. M. Therneau. *A Package for Survival Analysis in R*, 2022. URL `https://CRAN.R-project.org/package=survival`. R package version 3.4-0.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, 2003.

M. J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

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

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.

P. J. Verweij and H. C. van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–2314, 1993.

T. Westling, A. Luedtke, P. Gilbert, and M. Carone. Inference for treatment-specific survival curves using machine learning. *arXiv preprint arXiv:2106.06602*, 2021.

D. Whitney, A. Shojaie, and M. Carone. Comment: Models as (deliberate) approximations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 34(4):591, 2019.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

J. Yao, X. Zhu, F. Zhu, and J. Huang. Deep correlational learning for survival prediction from multi-modality data. In *International conference on medical image computing and computer-assisted intervention*, pages 406–414. Springer, 2017.

Manuscript III

Title	Estimating conditional hazard functions and densities with the highly-adaptive lasso
Authors	Anders Munch, Mark J. van der Laan, Thomas A. Gerds, and Helene Rytgaard
Status	In preparation

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

Anders Munch, Mark J. van der Laan, Thomas A. Gerds, and Helene Rytgaard

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 emphasize the difference between the empirical risk minimizer and the highly-adaptive lasso and show that the empirical risk minimizer is either not well-defined or not consistent for estimation of conditional hazard functions and densities. Under an additional smoothness assumption we formally show that the highly-adaptive lasso achieves the same convergence rate as has been shown to hold for the empirical risk minimizer in settings where the latter is well-defined. Finally, we show that our results are 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 $\mathcal{D}^1_M$ be the space of càdlàg functions $f\colon [0,1] \to \mathbb{R}$ with variation norm bounded by $M < \infty$. For a suitable loss function $L\colon \mathcal{D}^1_M \times [0,1] \to \mathbb{R}$ and sample dataset $\{X_i\}_{i=1}^n$ of iid. observations $X_i \sim P$, we consider estimation of the parameter

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

The empirical risk minimizer estimates f^* by minimizing $\mathbb{P}_n[L(f,\cdot)]$ over $\mathcal{D}^1_M$, where $\mathbb{P}_n$ is the empirical measure of the sample dataset $\{X_i\}_{i=1}^n$. The highly-adaptive lasso (HAL) estimator [van der Laan, 2017] estimates f^* by minimizing $\mathbb{P}_n[L(f,\cdot)]$ over the data-dependent function class

$$\mathcal{D}^1_{M,n} = \left\{ f\colon [0,1] \to \mathbb{R} : f(x) = \beta_0 + \sum_{i=1}^n \beta_i \mathbb{1}\{X_i \le x\}, \beta_i \in \mathbb{R}, \sum_{i=0}^n |\beta_i| \le M \right\}.$$

In this article, we consider the multivariate version of the estimation problem in equation (1) and the corresponding multivariate HAL estimator. Previous work has studied the convergence rates of the empirical risk minimizer [van der Laan, 2017, Benkeser and van der Laan, 2016, Bibaut and van der Laan, 2019, Fang et al., 2021] but not of the HAL estimator directly. The HAL estimator can be interpreted as a highly data-adaptive sieve estimator [Grenander, 1981, Geman, 1981, Geman and Hwang, 1982, Walter and Blum, 1984]. We show that the empirical risk minimizer and the HAL estimator are not in general the same, and we study the asymptotic convergence rates for the HAL estimator. Our asymptotic results for the HAL estimator are particularly important for estimation of (conditional) densities and hazard functions, because the empirical risk minimizer is either not well-defined or not consistent for these parameters. For least-squares regression, the empirical risk minimizer is well-defined [Fang et al., 2021] but is only equal to the HAL estimator in the univariate case.

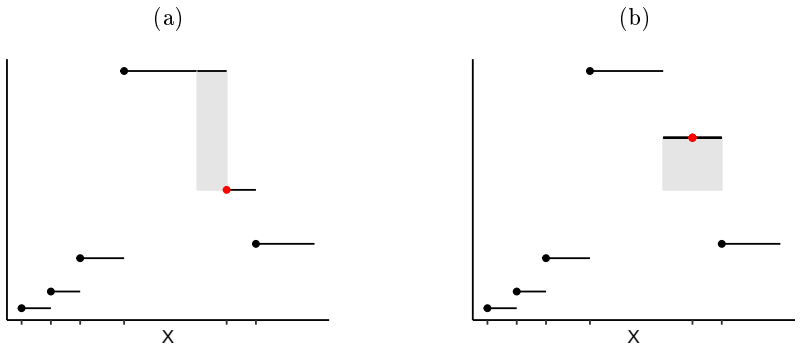


Figure 1: Illustration of two càdlàg densities, where the ticks at the x-axis denote observed data points, and the y-coordinates of the black dots denote the likelihood given to these points by the densities. Panel (a) shows a given càdlàg density, while panel (b) shows an adjusted density that has the 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.

Neuhaus [1971] generalized the concept of a càdlàg function to the multivariate setting; our Definition 2.1 is an equivalent definition used by, e.g., Czerebak-Morozowicz et al. [2008] and Ferger [2015]. A useful generalization of the variation norm of a function to the multivariate setting is the sectional variation norm. This norm was introduced by [Gill et al., 1995, van der Laan, 2017], and is closely related to the Hardy-Krause variation of a function [Krause, 1903, Hardy, 1906, Owen, 2005, Aistleitner and Dick, 2015]. The Hardy-Krause variation is identical to the sectional variation norm, except that a constant function $f(x) = c$ has zero Hardy-Krause variation while its sectional variation norm is $|c|$. Informally, the sectional variation norm measures how much a function fluctuates without taking into account where in the domain the fluctuations happen. This poses a challenge when the loss $L(f, x)$ in equation (1) depends on the whole function f and not just on the value $f(x)$. For instance, 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 (Figure 1). A similar issue occurs with right-censored data in survival analysis, and our Proposition 5.1 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.

Targeted learning [van der Laan and Rose, 2011] and debiased machine learning [Chernozhukov et al., 2018] rely on non-parametric estimators of infinite-dimensional nuisance parameters, such as regression functions, conditional densities, and conditional hazard functions. To guarantee valid statistical inference, we need to estimate these 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 for regression functions when the nuisance parameter is assumed to belong to the class of càdlàg functions with uniformly bounded sectional variation norm [van der Laan, 2017, Bibaut and van der Laan, 2019]. Our work extends this result to settings that include estimation of (conditional) densities and hazard functions. In addition, our work is relevant for estimation of regression functions, because 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 (when it is defined) needs a number of basis function that is potentially of order n^d [Fang et al., 2021].

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 approaches use boosting [Schmid and Hothorn, 2008, Hothorn, 2020, Lee et al., 2021]. To the best of our knowledge, estimation of conditional hazard functions over the class of càdlàg functions with uniformly bounded sectional variation norm has not been studied theoretically before. Fang et al. [2021] considered estimation of regression functions over the class of functions with uniformly bounded Hardy-Krause variation, but assumed only that the functions are right-continuous in each coordinate. Aistleitner and Dick [2015] showed that the class of right-continuous functions with bounded Hardy-Krause variation is in one-to-one correspondence with finite signed measures on $[0,1]^d$. Using their results we formally show that the same correspondence holds for càdlàg functions with uniformly bounded sectional variation norm (Proposition 2.3). Thus the class of functions $\mathcal{D}_M^d$ considered here is identical to the function class considered by Fang et al. [2021], except for a constant baseline value. In particular, for a coordinate-wise right-continuous function, having bounded sectional variation norm implies that the function is càdlàg. The reverse statement is not true; for instance, almost any trajectory of a Brownian motion is a continuous (hence càdlàg) function with infinite sectional variation norm.

In Section 2 we introduce our notation and the space 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 the HAL estimator. In Section 4 we derive, for a general loss function, 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. 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.

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 $a \in \{0,1\}$ 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 2.1 (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 $\mathbf{a} = \mathbf{1}$.

We use $\mathcal{D}^d$ to denote the collection of all càdlàg functions with domain $[0,1]^d$. The content of Defintion 2.1 is illustrated in Figure 2 for $d = 2$.

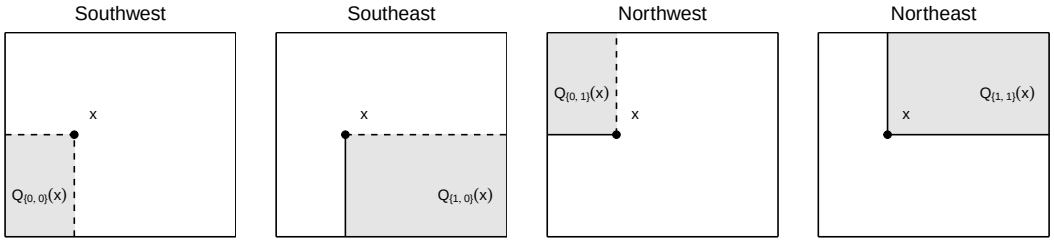


Figure 2: The four quadrants $Q_{\{0,0\}}(\mathbf{x})$, $Q_{\{1,0\}}(\mathbf{x})$, $Q_{\{0,1\}}(\mathbf{x})$, and $Q_{\{1,1\}}(\mathbf{x})$ spanned by the point $\mathbf{x} \in [0,1]^2$ and each of the vertices in the unit square. A sequence which is contained in one of the quadrants and converges to 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'.

A bit of notation is needed to formally define the section of a càdlàg function and the sectional variation norm. Our notation resembles that of Fang et al. [2021] but we focus on multivariate càdlàg functions. 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 $\mathrm{Im}(\pi_s) = s$, 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

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

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

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

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

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

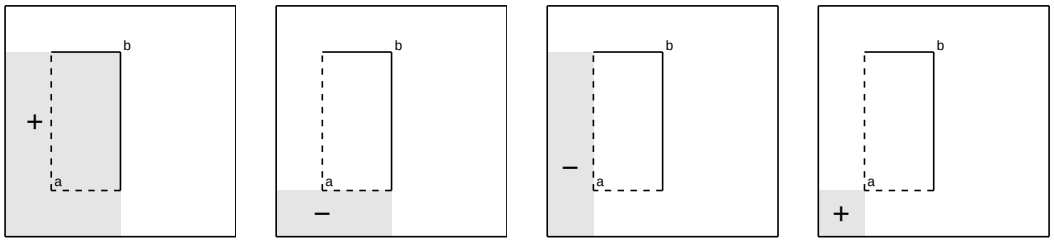


Figure 3: The area of the box $(\mathbf{a}, \mathbf{b}]$ can be calculated by first calculating the gray area in the leftmost figure, subtracting the gray areas in the two middle figures, and then adding the gray area in the rightmost figure.

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 2.2), and the second puts $\mathcal{D}_M^d$ into a one-to-one correspondence with finite signed measures (Proposition 2.3).

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

Proposition 2.2. *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 $\|\mu\|_{\text{TV}} = \mu_+([0, 1]^d) + \mu_-([0, 1]^d)$ denote the total variation norm of the measure μ.

Proposition 2.3. *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 in Aistleitner and Dick [2015]. We use their result in our proof in Appendix A which is for càdlàg functions. □

Based on Proposition 2.3 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 2.4. *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 2.2 showed that $\mathcal{D}_M^d$ is the closure of the piece-wise constant functions $\mathcal{R}_M^d$. Proposition 2.5 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 2.5. *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

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

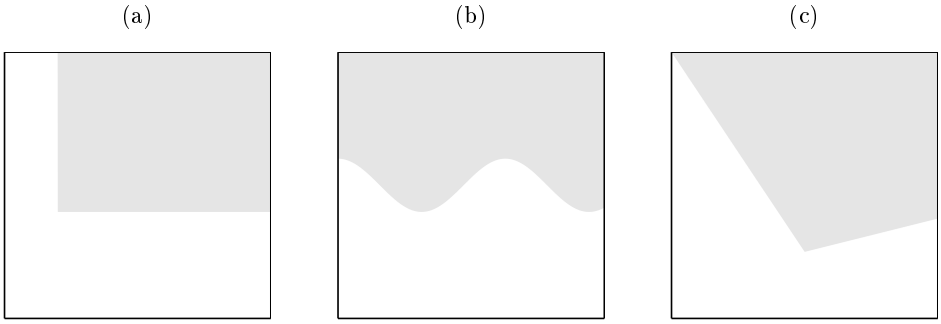


Figure 4: Illustration of functions $f\colon [0,1]^2 \to \mathbb{R}$ that are 0 on the white area and 1 on the shaded area. The function in panel (a) is càdlàg. The functions in panels (b) and (c) are not càdlàg.

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 2.4, 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 2.2, any $f_{\beta,n}$ is an element of $\mathcal{D}^d$ and we have

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

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

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

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

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

We refer to any minimizer as a HAL estimator.

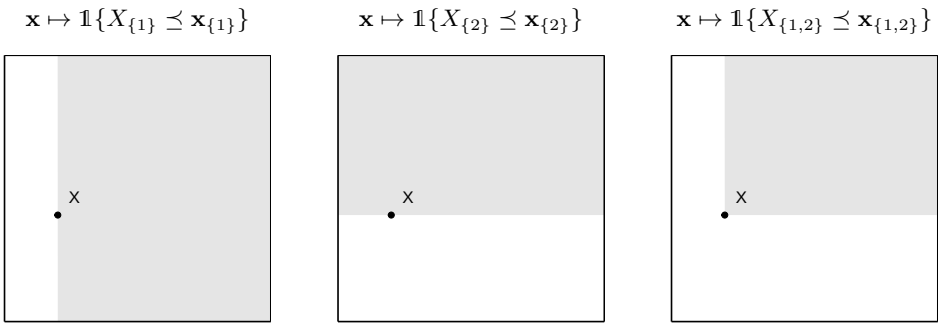


Figure 5: Examples of the basis functions that are used to construct the HAL estimator for $d = 2$.

4 Convergence rates using an approximate minimizer

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 4.2 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 $s \subset \{1, \ldots, d\}$ 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 4.1 (Smoothness of the loss function)**.** For a loss function L define the function space $\mathcal{L}_M = \{L(f, \cdot) : f \in \mathcal{D}^d_M\}$. There exist constants $C < \infty$, $\eta > 0$, and $\kappa \in \mathbb{N}$ such that the following conditions hold.

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

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

(iii) For any $\varepsilon > 0$, $\inf\{P[L(f,\cdot) - L(f^*,\cdot)] : f \in \mathcal{D}_M^d, \|f - f^*\|_\lambda \geq \varepsilon\} > 0$.

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

Assumptions 4.1 (ii) and (iii) are standard assumptions [e.g., van der Vaart and Wellner, 1996]. Some general conditions on the loss functions can be given to ensure that Assumption 4.1 (iv) holds, see for instance Lemma 4 in Appendix B in [Bibaut and van der Laan, 2019].

Assumption 4.2 (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 $\|\mathbb{1}_{(\mathbf{0}_s, \mathbf{1}_s]} \, \mathrm{d}f_s^* / \, \mathrm{d}P_s\|_\infty \leq C$ for all $s \in \mathcal{S}$.

Assumption 4.2 (ii) is substantial, as it imposes an additional smoothness condition on f^*. For instance, if P is dominated by Lebesgue measure, Assumption 4.2 (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.

Theorem 4.3. *If Assumptions 4.1 and 4.2 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 define

$$\Gamma_n(\delta) = \sup_{\|f - f^*\|_\lambda < \delta} |\mathbb{G}_n[L(f,\cdot) - L(f^*,\cdot)]|, \tag{10}$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ is the empirical process. We apply theorem 3.2.5 from van der Vaart and Wellner [1996] to a HAL estimator $\hat{f}_n$. This yields that $\hat{f}_n$ converges to f^* at rate r_n if Assumption 4.1 (ii) holds together with the following three conditions.

(a) For each $n \in \mathbb{N}$, there exists a function $\varphi_n \colon (0, \infty) \to \mathbb{R}$ such that $\delta \mapsto \varphi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and

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

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

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

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

Lemma B.1 in Appendix B shows that condition (a) holds for the rate $r_n = n^{1/3} \log(n)^{-2(d-1)/3}$. Lemma 4.4 shows that condition (b) holds for this rate. Condition (c) follows from Assumptions 4.1 (iii) and (iv), Lemma 4.4, and Lemma B.2 in Appendix B. □

The key component of Lemma 4.4 is to construct an auxiliary function f_n^* which belongs to $\mathcal{D}_{M,n}^d$ with high probability and is close to the target parameter f^*.

Lemma 4.4. *If Assumptions 4.1 and 4.2 hold, and $\hat{f}_n$ is a HAL estimator as defined in equation (9), then*

$$\mathbb{P}_n[L(\hat{f}_n,\cdot)] \leq \mathbb{P}_n[L(f^*,\cdot)] + O_P\Big(n^{-2/3} (\log n)^{4(d-1)/3}\Big). \tag{11}$$

Proof. By Assumption 4.2 (ii) we can define the random function

$$f_n^*(\mathbf{x}) = f^*(\mathbf{0}) + \sum_{s \in \mathcal{S}} \int_{(\mathbf{0}_s, \mathbf{x}_s]} \frac{\mathrm{d}f_s^*}{\mathrm{d}P_s} \, \mathrm{d}\mathbb{P}_{s,n}, \tag{12}$$

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 $\{X_{s,i}\}_{i=1}^n$. The function f_n^* has the following useful properties,

$$\|f_n^* - f^*\|_\lambda = O_P(n^{-1/2}) \quad \text{and} \quad P\big(f_n^* \in \mathcal{D}_{M,n}^d\big) \longrightarrow 1. \tag{13}$$

To see this, we use the representation given by Proposition 2.4 and Assumption 4.2 (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 the first statement of equation (13). To show the second statement in equation (13), note that

$$f_n^*(\mathbf{x}) = f^*(\mathbf{0}) + \sum_{s \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(\mathbf{0}_s, \mathbf{1}_s]}(X_{s,i}) \frac{\mathrm{d}f_s^*}{\mathrm{d}P_s}(X_{s,i}) \mathbb{1}\{X_{s,i} \preceq \mathbf{x}_s\}. \tag{14}$$

Equation (14) shows that $f_n^* \in \mathcal{D}_n^d$, and by equation (8)

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

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 the last equality follows from the properties of the Jordan-Hahn decomposition and the fact that P_s is a positive measure. By Assumption 4.2 (ii) and the law of large numbers this implies that $\|f_n^*\|_v \xrightarrow{P} \|f^*\|_v$. As $\|f^*\|_v < M$ by Assumption 4.2 (i) it follows that $P(\|f_n^*\|_v < M) \to 1$, which shows the second statement in equation (13).

Let now $r_n = n^{1/3} \log(n)^{-2(d-1)/3}$ and define the indicator variable

$$\eta_n = \mathbb{1}\big\{f_n^* \in \mathcal{D}_{M,n}^d, \|f_n^* - f^*\|_\lambda < r_n^{-1}\big\}.$$

Note that equation (13) implies that $P(\eta_n = 1) \to 1$ and [Schuler et al., 2023, Lemma 2] yields

$$(1 - \eta_n) = o_P(a_n^{-1}) \quad \text{for any sequence } \{a_n\}_{n=1}^\infty \subset \mathbb{R}. \tag{15}$$

When $\eta_n = 1$ we have $f_n^* \in \mathcal{D}_{M,n}^d$ an thus by definition of $\hat{f}_n$ we have $\eta_n \mathbb{P}_n[L(\hat{f}_n, \cdot)] \leq \eta_n \mathbb{P}_n[L(f_n^*, \cdot)]$. We obtain

$$\begin{aligned}
&\mathbb{P}_n[L(\hat{f}_n, \cdot)] - \mathbb{P}_n[L(f^*, \cdot)] \\
&\quad = \eta_n \mathbb{P}_n[L(\hat{f}_n, \cdot) - L(f^*, \cdot)] + (1 - \eta_n)\mathbb{P}_n[L(\hat{f}_n, \cdot) - L(f^*, \cdot)] \\
&\quad \leq \eta_n \mathbb{P}_n[L(f_n^*, \cdot) - L(f^*, \cdot)] + (1 - \eta_n)\mathbb{P}_n[L(\hat{f}_n, \cdot) - L(f^*, \cdot)] \\
&\quad = \eta_n \mathbb{P}_n[L(f_n^*, \cdot) - L(f^*, \cdot)] + O_P(1 - \eta_n),
\end{aligned} \tag{16}$$

where we used Assumption 4.1 (i) for the last equality. Similarly, when $\eta_n = 1$ then $\|f_n^* - f^*\|_\lambda < r_n^{-1}$ so we can write

$$\begin{aligned}
|\eta_n \mathbb{P}_n[L(f_n^*, \cdot) - L(f^*, \cdot)]| &= \eta_n \left(\left| n^{-1/2} \mathbb{G}_n[L(f_n^*, \cdot) - L(f^*, \cdot)] + P[L(f_n^*, \cdot) - L(f^*, \cdot)] \right| \right) \\
&\leq \eta_n n^{-1/2} \Gamma_n(r_n^{-1}) + \eta_n P[L(f_n^*, \cdot) - L(f^*, \cdot)] \\
&\leq n^{-1/2} \Gamma_n(r_n^{-1}) + \eta_n P[L(f_n^*, \cdot) - L(f^*, \cdot)] \\
&\leq n^{-1/2} \Gamma_n(r_n^{-1}) + \eta_n r_n^{-2},
\end{aligned}$$

where Γ_n was defined in equation (10) and we used Assumption 4.1 (ii) for the last inequality. Thus, by equations (15) and (16) we have

$$\mathbb{P}_n[L(\hat{f}_n, \cdot)] - \mathbb{P}_n[L(f^*, \cdot)] \leq n^{-1/2} \Gamma_n(r_n^{-1}) + O_P(r_n^{-2}).$$

By Markov's inequality, Assumption 4.1 (i) and (iv), and Lemma B.1 in Appendix B, we have

$$n^{-1/2} \Gamma_n(r_n^{-1}) = O_P(r_n^{-2}), \tag{17}$$

which proves the lemma. □

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,

$$\log \alpha(t, \mathbf{w}) = f(t, \mathbf{w}), \quad \text{with} \quad f \in \mathcal{D}_M^d. \tag{18}$$

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],

$$L^{\mathrm{pl}}(f, O) = \int_0^{\tilde{T}} e^{f(u,W)} \, \mathrm{d}u - \Delta f(\tilde{T}, W). \tag{19}$$

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 5.1 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 5.1. *Let $f^\circ \in \mathcal{D}_M^1$ be given. If there exists a $j \in \{1, \ldots, n' - 1\}$ such that $f^\circ(\tilde{T}_{(j)}) > f^\circ(\tilde{T}_{(j+1)})$, then*

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

Proof. See Appendix C.1. □

Proposition 5.1 implies that any estimator $\hat{f}_n \in \mathcal{D}_M^1$ 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 5.2 on the other hand shows that a HAL estimator can be found as the solution to a convex optimization problem.

Proposition 5.2. *Let $f_{\beta,n}$ be the data-dependent model defined in equation (7). The problem*

$$\min_{\|\beta\|_1 \leq M} \mathbb{P}_n[L^{\mathrm{pl}}(f_{\beta,n}, \cdot)], \tag{20}$$

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

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

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

$$P(\tilde{T} = 1 \mid W = \mathbf{w}) = P(\tilde{T} = 1, \Delta = 0 \mid W = \mathbf{w}) = \exp\left\{-\int_{[0,1)} \gamma_0(s, \mathbf{w})\, \mathrm{d}s\right\}. \tag{21}$$

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 (21). 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 5.3. *Let P be a distribution such that $\|\gamma_P\|_\infty < \infty$, $\|\omega_P\|_\infty < \infty$, and $f_P \in \mathcal{D}_M^d$. Then the following holds.*

(a) For any $\varepsilon > 0$, $\inf\{P[L(f, \cdot) - L(f_P, \cdot)] : f \in \mathcal{D}_M^d, \|f - f_P\|_\lambda \geq \varepsilon\} > 0$.

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

Proof. These properties 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 5.4. *Let P be a distribution such that $\|\gamma_P\|_\infty < \infty$, $\|\omega_P\|_\infty < \infty$, and let $\hat{f}_n$ be a HAL estimator based on the negative partial log-likelihood loss defined in equation (19). If Assumption 4.2 holds, then*

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

Proof. Corollary 5.4 follows from Theorem 4.3 and Lemma 5.3. 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. [2021], 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.

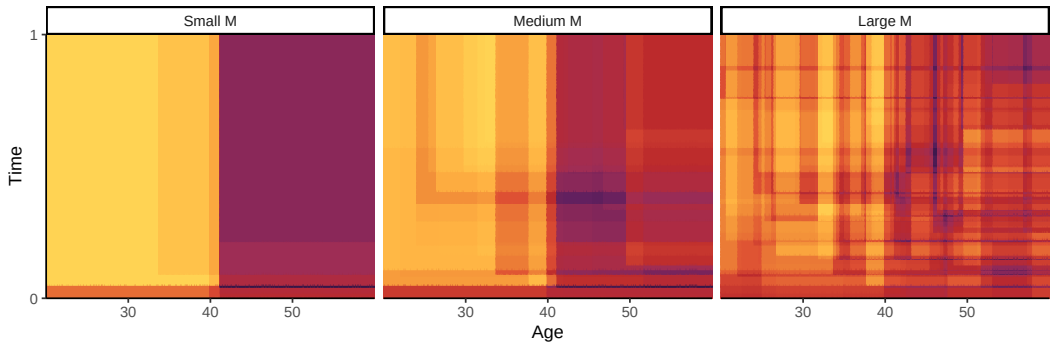


Figure 6: The HAL estimator of the hazard function for the treated group based on a sample size of 200 from the simulated study with darker values corresponding to higher values of the hazard function. Estimates are shown for three different values of the sectional variation norm (M).

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

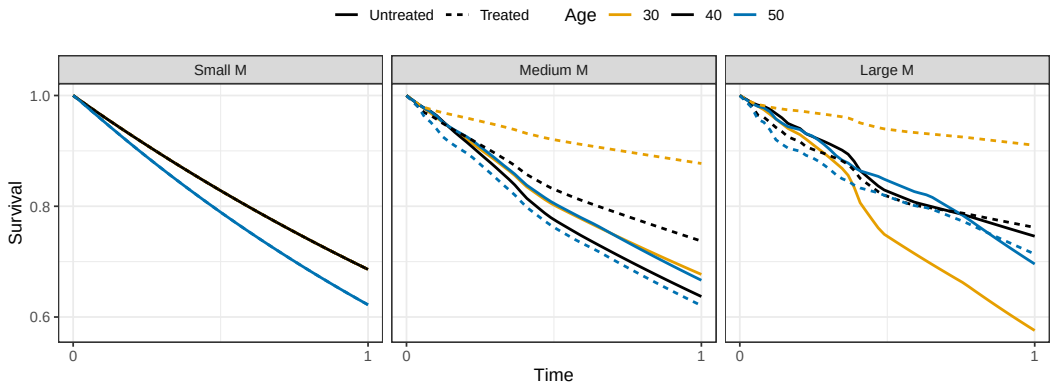


Figure 7: Estimates of the survival function derived from the HAL estimator stratified on treatment and three different age values based on a sample of 200 patients from the simulated study. Estimates are shown for three different values of the sectional variation norm (M).

the conditional density as an element of

$$\mathcal{P}_M^d = \left\{ p \colon [0,1]^d \to \mathbb{R}_+ \;\middle|\; \log p(u, \mathbf{w}) = f(u, \mathbf{w}) - \log \left(\int_0^1 e^{f(z, \mathbf{w})} \, \mathrm{d}z \right), f \in \mathcal{D}_M^d \right\}. \tag{22}$$

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

$$\hat{p}_n \in \operatorname*{argmin}_{p \in \mathcal{P}_{M,n}^d} \mathbb{P}_n[-\log p]. \tag{23}$$

Proposition 6.1 shows that a HAL estimator is well-defined and can be found as the solution to a convex optimization problem.

Proposition 6.1. *Define the set of indices* $\mathcal{I} = \{\{1\} \cup s : s \subset \{2, \ldots, d\}\}$ *and let*

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

The problem

$$\min_{\|\beta\|_1 \leq M} \mathbb{P}_n\big[\bar{L}(g_{\beta,n}, \cdot)\big], \quad \textit{with} \quad \bar{L}(g, O) = \log \left(\int_0^1 e^{g(z, W)} \, \mathrm{d}z \right) - g(U, W), \tag{24}$$

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

$$p_{\hat{\beta},n} \in \operatorname*{argmin}_{p \in \mathcal{P}_{M,n}^d} \mathbb{P}_n[-\log p], \quad \textit{where} \quad \log p_{\hat{\beta},n}(u, \mathbf{w}) = g_{\hat{\beta},n}(u, \mathbf{w}) - \log \left(\int_0^1 e^{g_{\hat{\beta},n}(z, \mathbf{w})} \, \mathrm{d}z \right).$$

Proof. See Appendix C.2 □

Proposition 6.1 shows that the HAL estimator defined in equation (23) 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 6.2. *Let P be a distribution such that $\|\omega_P\|_\infty < \infty$ and $p_P \in \mathcal{P}_M^d$, and let $\hat{p}_n$ be a HAL estimator as defined in equation (23). If Assumption 4.2 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 6.2 follows from Theorem 4.3. 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 (23) as a 'HAL density parametrization'. We compare these two estimators in Figure 8, where we have fitted both estimators to a simulated univariate dataset. 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 (22) 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

$$L^{\text{se}}(f, O) = (f(X) - Y)^2. \tag{25}$$

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

Corollary 7.1. *Let P be a distribution such that $\|\omega_P\|_\infty < \infty$ 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 (25). If Assumption 4.2 holds, then*

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

Proof. We show that Assumption 4.1 holds for the squared error loss, and so Corollary 7.1 follows from Theorem 4.3. First note that because the squared error loss is a strictly proper

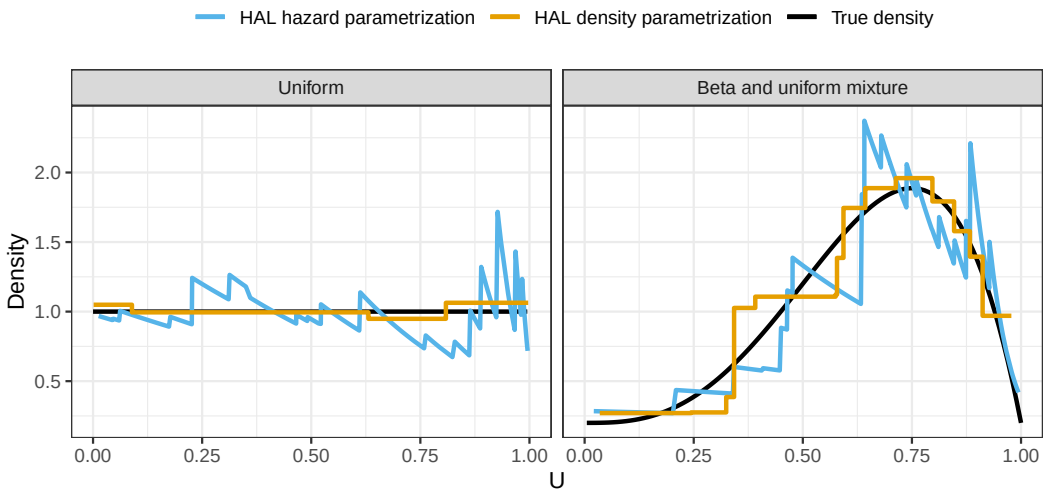


Figure 8: Two different density estimators under two different data-generating distributions. We generated 200 samples from a uniform distribution (left panel) and a mixture of a beta distribution and a uniform distribution (right panel). The 'HAL hazard parametrization' refers to a density estimator 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 (23).

scoring rule [Gneiting and Raftery, 2007], the assumption that $f_P \in \mathcal{D}_M^d$ implies that $f^* = f_P$ a.e. Conditions 4.1 (i)-(iii) hold by the definition of the squared error loss and the assumption that Y and ω_P are bounded. Condition 4.1 (iv) 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

In this paper we have demonstrated that an empirical risk minimizer over the class of càdlàg functions with bounded sectional variation norm is in general different from a HAL estimator. We have derived the asymptotic convergence rates directly for the HAL estimator. In particularly, our work now rigorously justifies the use of a HAL estimator for estimation of conditional hazard functions and densities with uniformly bounded sectional variation norm. As discussed in Sections 4, our main result relies on the smoothness assumption 4.2 (ii). We conjecture that it is possible to construct a sieve estimator using a finer grid than the one used by a HAL estimator, in such a way that the sieve estimator remains well-defined in general settings while achieving the same rate of convergence without imposing any other assumptions than $f^* \in \mathcal{D}_M^d$. From a practical perspective, an interesting question is how much we can reduce the number of basis functions and still expect to achieve fast

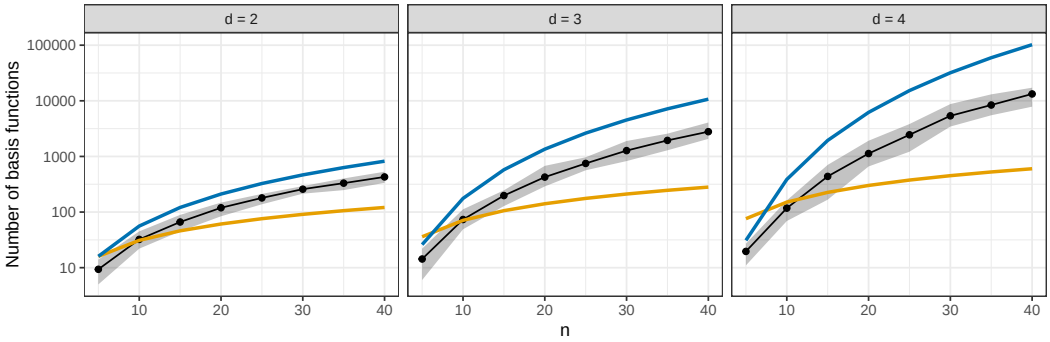


Figure 9: The black line is the average number of basis functions needed to calculate the empirical risk minimizer with ribbons denoting the 2.5%- and 97.5%-quantiles based on 200 simulations of uniformly distributed covariates. The number of observations is denoted by n and the dimension by d. The blue 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.

convergence, and whether we need to impose additional smoothness assumptions for this to hold. Schuler et al. [2023] recently demonstrated that a good approximation to the HAL estimator can be achieved using a gradient tree boosting algorithm which is computationally scalable.

As mention in Section 7, for the special case of least-squares regression, an empirical risk minimizer exists. In addition, the empirical risk minimizer converges at the same rate as the HAL estimator without imposing Assumption 4.2 [Bibaut and van der Laan, 2019, Fang et al., 2021]. We do not know whether this assumption is necessary for ensuring convergence of the HAL estimator. The dramatic reduction in the number of needed basis functions indicate that there might be a price to pay when using the HAL estimator instead of the empirical risk minimizer. It would be interesting to investigate if the HAL estimator remains consistent at the same or a slower rate when the additional smoothness assumption 4.2 (ii) fails 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 5.1 does not rule out 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

We use the following theorem from Aistleitner and Dick [2015] to prove the results from Section 2.

Theorem A.1 (Theorem 3 in [Aistleitner and Dick, 2015]). *(a) If $f\colon [0,1]^d \to \mathbb{R}$ is right-continuous in each of its coordinates and $\|f\|_v < \infty$ then there exists a unique signed Borel measure μ_f on $[0,1]^d$ such that*

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

and $\|\mu_f\|_{\mathrm{TV}} = \|f\|_v$*, where* $\|\cdot\|_{\mathrm{TV}}$ *denote the total variation norm of a measure.*

(b) If μ *is a signed Borel measure on* $[0,1]^d$ *with* $\|\mu_f\|_{\mathrm{TV}} < \infty$ *then there exists a unique function* f_μ *that is right-continuous in each of its coordinates such that*

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

We start by proving the following two lemmas.

Lemma A.2. *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 $f_n \in \mathcal{D}_M^d$ for all $n \in \mathbb{N}$, so we must have $\|f\|_v \leq M$. □

Lemma A.3. *Let* f *be a function that is right-continuous in each of its coordinates with* $\|f\|_v \leq M$*. There exists a sequence* $\{f_n\} \subset \mathcal{R}_M^d$ *such that* $\|f - f_n\|_\infty \to 0$ *for* $n \to \infty$*.*

Proof. By Theorem A.1 (a) 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 A.1 and because $P^+ \perp P^-$ we have

$$M \geq \|f\|_v = \|\mu_f\|_{\mathrm{TV}} = \alpha \|P^+\|_{\mathrm{TV}} + \beta \|P^-\|_{\mathrm{TV}} = \alpha + \beta. \tag{26}$$

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

$$p_n^+ \perp p_n^-, \quad \forall n \in \mathbb{N}, \tag{27}$$

and

$$\|f_n^+ - F^+\|_\infty \longrightarrow 0 \quad \text{and} \quad \|f_n^- - F^-\|_\infty \longrightarrow 0. \tag{28}$$

Note that f_n^+ is a linear combination of the indicator functions $\{\mathbb{1}_{[\mathbf{x}_i,\mathbf{1}]}\}_{i=1}^n$, where $\{\mathbf{x}_i\}_{i=1}^n$ 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 (26) and (27),

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

so $f_n \in \mathcal{R}_M^d$ for all $n \in \mathbb{N}$. Equation (28) gives that $\|f_n - f\|_\infty \to 0$ which concludes the proof. □

Proof of Proposition 2.2. 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 $\mathcal{R}_M^d \subset \mathcal{D}_M^d$. It thus follows from Lemma A.2 that $\overline{\mathcal{R}_M^d} \subset \mathcal{D}_M^d$. As any $f \in \mathcal{D}_M^d$ is right-continuous in each of its coordinates, the reverse inclusion follows from Lemma A.3. □

Proof of Proposition 2.3. 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 A.1 (a). For the second statement, we know by Theorem A.1 (b) 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 A.3, f_μ can be approximated uniformly by a sequence of functions $f_n \in \mathcal{D}_M^d$. Lemma A.2 then implies that $f_\mu \in \mathcal{D}_M^d$. □

Proof of Proposition 2.4. 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 2.3 we can write

$$f(\mathbf{x}) = \mu_f([\mathbf{0}, \mathbf{x}]) = \mu_f(\{\mathbf{0}\}) + \sum_{s \in \mathcal{S}} \mu_f(A(\mathbf{x}; s)) \tag{29}$$

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

$$f_s(\mathbf{x}) = \mu_{f_s}([\mathbf{0}_s, \mathbf{x}]), \quad \text{for all} \quad \mathbf{x} \in [0, 1]^{|s|}. \tag{30}$$

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

$$\mu_f(A(\mathbf{x}; s)) = \mu_{f_s}((\mathbf{0}_s, \mathbf{x}_s]). \tag{31}$$

By equations (29) and (31) 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 2.5. 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

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

implies $f \in \mathcal{R}_M^d$. 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 $f \in \mathcal{R}_M^d$, and thus we have proved the initial claim. The proposition now follows by noting that this implies that if $f \notin \mathcal{R}_M^d$, 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 $f \notin \mathcal{R}_M^d$, 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 Lemmas from empirircal process theory

Recall the notation $\mathcal{L}_M = \{L(f, \cdot) : f \in \mathcal{D}_M^d\}$ for a loss function $L \colon \mathcal{D}_M^d \times \mathcal{O} \to \mathbb{R}_+$, and the assumption

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

Lemma B.1. *Let*

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

If (B) holds and $\|L(f,\cdot)\|_\infty < C$*, then for all* $n \in \mathbb{N}$*,*

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

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

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

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

$$\mathbb{E}_P^*[\Gamma_n(\delta)] \lesssim J_{[]}(\delta, \mathcal{L}_M, \|\cdot\|_P) \left(1 + \frac{J_{[]}(\delta, \mathcal{L}_M, \|\cdot\|_P)}{\delta^2\sqrt{n}} C\right). \tag{32}$$

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

$$\log N_{[]}(\varepsilon, \mathcal{L}_M, \|\cdot\|_P) \leq \log C + \kappa \log N_{[]}(\varepsilon/C, \mathcal{D}_M^d, \|\cdot\|_\lambda) \lesssim \varepsilon^{-1} |\log(\varepsilon)|^{2(d-1)}, \tag{33}$$

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 (33) to obtain

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

and so equation (32) gives

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

which was the first statement of the lemma. For the second statement, set $\delta = r_n^{-1}$ and

obtain for all $n \geq 3$,

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

□

Lemma B.2. *Assume that (B) holds and that for any $\varepsilon > 0$,*

$$\inf\{P[L(f, \cdot) - L(f^*, \cdot)] : f \in \mathcal{D}_M^d, \|f - f^*\|_\lambda \geq \varepsilon\} > 0. \tag{34}$$

If $\mathbb{P}_n[L(\hat{f}_n, \cdot)] \leq \mathbb{P}_n[L(f^*, \cdot)] + o_P(1)$ then $\|\hat{f}_n - f^*\|_\lambda \xrightarrow{P^*} 0$.

Proof. Proposition 1 in [Bibaut and van der Laan, 2019] and Theorem 2.4.1 in [van der Vaart and Wellner, 1996] together with Assumption (B) imply that $\mathcal{L}_M$ is a Glivenko-Cantelli class of functions. The result then follows from Corollary 3.2.3 in [van der Vaart and Wellner, 1996]. □

C Additional proofs

C.1 Right-censored data

Proof of Proposition 5.1. Let $f^\circ \in \mathcal{D}_M^1$ be a function and $j \in \{1, \ldots, n-1\}$ 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 $\mathbb{P}_n[L^{\mathrm{pl}}(\check{f}, \cdot)] < \mathbb{P}_n[L^{\mathrm{pl}}(f^\circ, \cdot)]$ when L^{pl} is the negative log-likelihood defined in equation (19). 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

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

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

Proof of Proposition 5.2. 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 $j = 1, \ldots, n'$, 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 $\beta \mapsto \mathbb{P}_n[L^{\mathrm{pl}}(f_{\beta,n}, \cdot)]$ is convex, and as $\mathcal{B}_M$ is convex it follows that the problem in (20) 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. □

Proof of Lemma 5.3. 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.,

$$\begin{aligned}
p_f(\mathbf{w}, t, \delta) &= \left(e^{f(t,\mathbf{w})} \exp\left\{-\int_0^t [e^{f(s,\mathbf{w})} + \gamma_P(s, \mathbf{w})]\,\mathrm{d}s\right\}\right)^{\delta} \\
&\quad \times \left(\gamma_P(t, \mathbf{w})^{\mathbb{1}_{[0,1)}(t)} \exp\left\{-\int_0^t [e^{f(s,\mathbf{w})} + \gamma_P(s, \mathbf{w})]\,\mathrm{d}s\right\}\right)^{1-\delta} \\
&= \left(e^{f(t,\mathbf{w})}\right)^{\delta} \exp\left\{-\int_0^t e^{f(s,\mathbf{w})}\,\mathrm{d}s\right\} \\
&\quad \times \left(\gamma_P(t, \mathbf{w})^{\mathbb{1}_{[0,1)}(t)}\right)^{1-\delta} \exp\left\{-\int_0^t \gamma_P(s, \mathbf{w})\,\mathrm{d}s\right\}, \\
&=: q_f(\mathbf{w}, t, \delta) g(\mathbf{w}, t, \delta),
\end{aligned} \tag{40}$$

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

$$\begin{aligned}
D_{\mathrm{KL}}(P_{f_0} \,||\, P_f) &= \int \log \frac{p_{f_0}}{p_f} p_{f_0}\,\mathrm{d}\nu \\
&= \int_{[0,1]^d \times \{0,1\}} \left[\int_0^t e^{f(s,\mathbf{w})}\,\mathrm{d}s - \delta f(t, \mathbf{w}) \right. \\
&\qquad \left. - \left(\int_0^t e^{f_0(s,\mathbf{w})}\,\mathrm{d}s - \delta f_0(t, \mathbf{w})\right)\right] p_{f_0}(\mathbf{w}, t, \delta)\,\mathrm{d}\nu(\mathbf{w}, t, \delta) \\
&= P_{f_0}[L^{\mathrm{pl}}(f, \cdot)] - P_{f_0}[L^{\mathrm{pl}}(f_0, \cdot)],
\end{aligned} \tag{41}$$

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

$$\begin{aligned} D_{\mathrm{KL}}(P_{f_0} \,||\, P_f) &\geq \int \left(\sqrt{p_{f_0}} - \sqrt{p_f}\right)^2 \mathrm{d}\nu \\ &\geq \left(\|(\sqrt{p_{f_0}} + \sqrt{p_f})^2\|_\infty\right)^{-1} \int (p_{f_0} - p_f)^2 \,\mathrm{d}\nu \\ &\geq \left(4e^M(\|\gamma_P\|_\infty \vee 1)\right)^{-1} \int (p_{f_0} - p_f)^2 \,\mathrm{d}\nu \\ &= \left(4e^M(\|\gamma_P\|_\infty \vee 1)\right)^{-1} \int (p_{f_0} - p_f)^2 \,\mathrm{d}\nu. \end{aligned} \tag{42}$$

Letting $S_f(t, \mathbf{w}) = \exp\left\{-\int_0^t e^{f(s,\mathbf{w})}\,\mathrm{d}s\right\}$ we have

$$\begin{aligned} \int (p_{f_0} - p_f)^2 \,\mathrm{d}\nu &= \int g^2 \,(q_{f_0} - q_f)^2 \,\mathrm{d}\nu \\ &\geq \int_{[0,1]^{d-1}\times[0,1)\times\{1\}} g^2 \,(q_{f_0} - q_f)^2 \,\mathrm{d}\nu \\ &\geq e^{\|\gamma_P\|_\infty} \int_{[0,1]^{d-1}} \int_0^t (e^{f_0} S_{f_0} - e^f S_f)^2 \,\mathrm{d}\lambda \otimes P_W \\ &\geq \frac{e^{-\|\gamma_P\|_\infty}}{\|\omega_P\|_\infty} \|e^{f_0} S_{f_0} - e^f S_f\|_\lambda^2. \end{aligned} \tag{43}$$

Note that when f and f_0 are uniformly bounded, then $\|e^{f_0} S_{f_0} - e^f S_f\|_\lambda \to 0$ implies $\|f_0 - f\|_\lambda \to 0$. Hence, because $\frac{e^{-\|\gamma_P\|_\infty}}{\|\omega_P\|_\infty} > 0$ by the assumptions about γ_P and ω_P, equations (41)-(43) imply that for any sequence of function $\{f_n\} \subset \mathcal{D}_M^d$,

$$\left\{P_{f_0}[L^{\mathrm{pl}}(f_n, \cdot)] - P_{f_0}[L^{\mathrm{pl}}(f_0, \cdot)]\right\} \longrightarrow 0 \quad \text{implies} \quad \|f_n - f_0\|_\lambda \longrightarrow 0, \tag{44}$$

when $n \to \infty$. To see that this implies statement (a), note that if statement (a) were false there would exist an $\varepsilon > 0$ and a sequence of functions $\{f_n\} \subset \mathcal{D}_M^d$ such that $P_{f_0}[L^{\mathrm{pl}}(f_n, \cdot)] - P_{f_0}[L^{\mathrm{pl}}(f_0, \cdot)] \to 0$ and $\|f_0 - f_n\|_\lambda \geq \varepsilon$ for all $n \in \mathbb{N}$. However, this is not possible by equation (44).

To show statement (b) 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 (40) we obtain

$$\begin{aligned} D_{\mathrm{KL}}(P_{f_0} \,||\, P_f) &\leq \int \frac{g^2 (q_{f_0} - q_f)^2}{q_f g} \,\mathrm{d}\nu \\ &= \int \frac{g (q_{f_0} - q_f)^2}{q_f} \,\mathrm{d}\nu \\ &\leq \exp\{M + e^{-M}\}(\|\gamma_P\|_\infty \vee 1) \int (q_{f_0} - q_f)^2 \,\mathrm{d}\nu, \end{aligned} \tag{45}$$

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 $[0,1]^{d-1} \times \{1\} \times \{0\}$ is a null set under ν, we can write

$$\begin{aligned} \int (q_{f_0} - q_f)^2 \,\mathrm{d}\nu = &\int_{[0,1]^{d-1}\times[0,1)\times\{0,1\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu \\ &+ \int_{[0,1]^{d-1}\times\{1\}\times\{0\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu. \end{aligned} \tag{46}$$

If we use $\mathbb{E}$ to denote expectation of W under P_W, we can write the first term on the right hand side of equation (46) as

$$
\begin{aligned}
&\int_{[0,1]^{d-1}\times[0,1)\times\{0,1\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu \\
&= \mathbb{E}\left[\int_0^1 \left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(\exp\left\{f_0(t,W) - \int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{f(t,W) - \int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2 \mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^1 \left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(\exp\left\{f_0(t,W) - \int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{f(t,W) - \int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2 \mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^1 \left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left\{\left(e^{f_0(t,W)} - e^{f(t,W)}\right)\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\}\right.\right. \\
&\qquad\qquad \left.\left. + \left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right) e^{f(t,W)}\right\}^2 \mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^1 \left(1 + e^{2f(t,W)}\right)\left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(e^{f_0(t,W)} - e^{f(t,W)}\right)^2 \exp\left\{-2\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\}\mathrm{d}t\right] \\
&\quad - \mathbb{E}\left[\int_0^1 \left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)\left(e^{f_0(t,W)} - e^{f(t,W)}\right)\right. \\
&\qquad\qquad \left. \times \exp\left\{f(t,W) - \int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\}\mathrm{d}t\right] \\
&\lesssim \mathbb{E}\left[\int_0^1 \left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(e^{f_0(t,W)} - e^{f(t,W)}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(\exp\left\{-\int_0^t e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^t e^{f(s,W)}\,\mathrm{d}s\right\}\right)\left(e^{f_0(t,W)} - e^{f(t,W)}\right)\mathrm{d}t\right],
\end{aligned}
$$

where we used that f_0 and f are uniformly bounded. Using Taylor expansions of $x \mapsto e^x$

around 0, Jensen's inequality, and Cauchy-Schwarz' inequality we then obtain

$$
\begin{aligned}
&\int_{[0,1]^{d-1}\times[0,1)\times\{0,1\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu \\
&\lesssim \mathbb{E}\left[\int_0^1 \left(\int_0^t (e^{f_0(s,W)} - e^{f(s,W)})\,\mathrm{d}s\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(e^{f_0(t,W)} - e^{f(t,W)}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \int_0^t \left(e^{f_0(s,W)} - e^{f(s,W)}\right)\mathrm{d}s \left(e^{f_0(t,W)} - e^{f(t,W)}\right)\mathrm{d}t\right] \\
&\lesssim \mathbb{E}\left[\int_0^1 \left(e^{f_0(s,W)} - e^{f(s,W)}\right)^2 \mathrm{d}s\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(e^{f_0(t,W)} - e^{f(t,W)}\right)^2 \mathrm{d}t\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \int_0^t \left(e^{f_0(s,W)} - e^{f(s,W)}\right)\mathrm{d}s \left(e^{f_0(t,W)} - e^{f(t,W)}\right)\mathrm{d}t\right] \\
&\lesssim \mathbb{E}\left[\int_0^1 \left(e^{f_0(s,W)} - e^{f(s,W)}\right)^2 \mathrm{d}s\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(e^{f_0(t,W)} - e^{f(t,W)}\right)^2 \mathrm{d}t\right] \\
&\quad + \left(\sqrt{\mathbb{E}\left[\int_0^1 \left\{\int_0^t \left(e^{f_0(s,W)} - e^{f(s,W)}\right)\mathrm{d}s\right\}^2 \mathrm{d}t\right]}\right. \\
&\qquad \left. \times \sqrt{\mathbb{E}\left[\int_0^1 \left\{\left(e^{f_0(t,W)} - e^{f(t,W)}\right)\right\}^2 \mathrm{d}t\right]}\right) \\
&\lesssim \mathbb{E}\left[\int_0^1 \left(e^{f_0(s,W)} - e^{f(s,W)}\right)^2 \mathrm{d}s\right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(e^{f_0(t,W)} - e^{f(t,W)}\right)^2 \mathrm{d}t\right] \\
&\quad + \sqrt{\mathbb{E}\left[\int_0^1 \left(e^{f_0(s,W)} - e^{f(s,W)}\right)^2 \mathrm{d}s\right]} \sqrt{\mathbb{E}\left[\int_0^1 \left\{\left(e^{f_0(t,W)} - e^{f(t,W)}\right)\right\}^2 \mathrm{d}t\right]} \\
&= \mathbb{E}\left[\int_0^1 \left(e^{f_0(t,W)} - e^{f(t,W)}\right)^2 \mathrm{d}t\right] \\
&\lesssim \mathbb{E}\left[\int_0^1 \left(f_0(t,W) - f(t,W)\right)^2 \mathrm{d}t\right] = \|f_0 - f\|^2_{\lambda\otimes P_W}.
\end{aligned}
\tag{47}
$$

The second term on the right hand side of equation (46) is

$$
\int_{[0,1]^{d-1}\times\{1\}\times\{0\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu = \mathbb{E}\left[\left(\exp\left\{-\int_0^1 e^{f_0(s,W)}\,\mathrm{d}s\right\} - \exp\left\{-\int_0^1 e^{f(s,W)}\,\mathrm{d}s\right\}\right)^2\right],
$$

and thus using similar arguments we obtain

$$
\begin{aligned}
\int_{[0,1]^{d-1}\times\{1\}\times\{0\}} (q_{f_0} - q_f)^2 \,\mathrm{d}\nu &\lesssim \mathbb{E}\left[\int_0^1 \left(f_0(s,W) - f(s,W)\right)^2 \mathrm{d}s\right] \\
&= \|f_0 - f\|^2_{\lambda\otimes P_W}.
\end{aligned}
\tag{48}
$$

Equations (41) and (45)-(48) then give statement (b). □

Proof of Corollary 5.4. First note that because condition 4.2 (i) is assumed to hold, $f_P = f^*$ a.e. by Lemma 5.3 (a). Thus Corollary 5.4 follows from Theorem 4.3 if we can show that Assumption 4.1 is true. Condition 4.1 (i) follows by definition of the loss function and (ii)-(iii) follow from Lemma 5.3 as γ_P and ω_P are assumed uniformly bounded. It thus only remains to show 4.1 (iv). 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 (21), $\Delta = \Delta \mathbb{1}\{\tilde{T} < 1\}$ 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

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

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

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

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 6.1. 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

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

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 (24) 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

$$\mathbb{P}_n[-\log p] = \mathbb{P}_n[\bar{L}(\log p, \cdot)] \quad \text{for any} \quad p \in \mathcal{P}^d_{M,n}. \tag{49}$$

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 $O \in [0,1]^d$,

$$\begin{aligned}\bar{L}(f+a, O) &= \log\left(\int_0^1 e^{f(s,W)+a(s,W)}\,\mathrm{d}s\right) - (f(U,W) - a(U,W)) \\ &= \log\left(e^{a(0,W)}\int_0^1 e^{f(s,W)}\,\mathrm{d}s\right) - (f(U,W) - a(0,W)) \\ &= \bar{L}(f, O).\end{aligned} \tag{50}$$

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 (49), 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 6.2. Define the log-density $f_1^*(u,\mathbf{w}) = f^*(u,\mathbf{w}) - \log\left(\int_0^1 e^{f^*(z,\mathbf{w})}\,\mathrm{d}z\right)$ and note that $f_1^* \in \mathcal{P}^d_M$. Equations (49) and (50) 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 $p_P \in \mathcal{P}^d_M$ by assumption. For any HAL estimator $\hat{p}_n$ we can write $\log\hat{p}_n(u,\mathbf{w}) = g_{\hat{\beta},n}(u,\mathbf{w}) - \log\left(\int_0^1 e^{g_{\hat{\beta},n}(z,\mathbf{w})}\,\mathrm{d}z\right)$, for some solution $\hat{\beta}$ to the problem (24). By equation (50), $g_{\hat{\beta},n}$ is a HAL estimator for the loss $\bar{L}$ as defined in equation (9). To prove Corollary 6.2 it suffices to show that

$$\|g_{\hat{\beta},n} - f^*\|_\lambda = o_P(n^{-1/3}\log(n)^{2(d-1)/3}). \tag{51}$$

We show that Assumption 4.1 holds for $\bar{L}$, which imply that equation (51) is true by Theorem 4.3. Condition 4.1 (i) holds because all $f \in \mathcal{D}^d_{M,n}$ are uniformly bounded, and conditions 4.1 (ii)-(iii) hold by properties of the Kullback-Leibler divergence because we assume that ω_P is uniformly bounded [Gibbs and Su, 2002]. Condition 4.1 (iv) is established by the same arguments used in the proof of Corollary 5.4. □

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.

D. Benkeser and M. van der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016.

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.

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. W. Rytgaard, F. Eriksson, and M. van der Laan. Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *arXiv preprint arXiv:2106.11009*, 2021.

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. 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.