# UNIVERSITY OF TORINO

M.Sc. in Stochastics and Data Science

Final dissertation


**Efficient Estimation of Survival Curves and Quantile Treatment Effects under Right-Censored Data: a Causal Machine Learning Approach.**

Supervisor: Matteo Giordano
Co-supervisor: Stijn Vansteelandt

Candidate: Edoardo Carroccetto

ACADEMIC YEAR 2024/2025

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Randomized controlled trials (RCTs) are widely regarded as the most reliable method for evaluating the causal effect of a binary treatment on a time-to-event outcome. In an RCT, participants are randomly assigned to either the treatment or control group and are observed over time. The treatment effect is typically assessed by comparing the proportion of participants in each group who experience the event by the end of the study. However, some participants' outcomes may be unknown due to factors such as study dropout or relocation, a phenomenon known as right-censoring of the event time. While RCTs are highly reliable, they are often impractical due to ethical concerns, logistical challenges, or financial constraints. In such cases, researchers turn to observational data, where treatment assignment is not randomized but determined by unknown mechanisms. Assessing the causal effect of a treatment using observational data is challenging due to confounding factors that influence both treatment selection and the outcome. As a result, any observed differences in outcomes between treated and untreated individuals may be attributed to these confounding variables rather than the treatment itself. If the available covariates are sufficiently rich to account for confounding in treatment assignment, treatment-censoring, and outcome-censoring relationships, it may still be possible to recover a valid causal effect.

In causal inference, the Quantile Treatment Effect (QTE) provides a comprehensive understanding of treatment effects. Unlike the Average Treatment Effect (ATE), which summarizes the mean difference between treatment groups, QTE examines how the treatment influences different points of the outcome distribution, offering insights into heterogeneity in effects. This approach is particularly useful in cases where treatment effects vary across

individuals, revealing disparities that would be masked by an average-based analysis.

A common approach for estimating QTE is quantile regression, which models conditional quantiles of the outcome distribution. However, in the presence of right-censoring (i.e., when the event time for certain participants remains unknown due to factors such as dropping out of the study or relocating), quantile regression becomes problematic because it does not naturally account for right-censored observations, leading to biased estimates.

To address this limitation, we use the survival estimator proposed by Westling, T. et al. (2023) to construct a survival-based method for estimating QTE. By leveraging survival functions, we construct a more accurate estimation of treatment effects across quantiles. This approach ensures that censoring is appropriately handled, making it more suitable for time-to-event data compared to standard quantile regression methods.

The remainder of this thesis is structured as follows: Chapter 2 outlines the statistical framework and introduces the causal parameters of interest. Chapter 3 examines the survival estimators proposed by Westling, T. et al. (2023), detailing their derivation through the influence function, their large-sample properties, and corresponding simulation studies. Chapter 4 introduces the quantile estimator, building upon the survival estimators from Chapter 3 and extending the analysis of large-sample properties.

# Chapter 2

# Statistical setting

## 2.1 Ideal and Observed data structure

We now define the ideal data structure in the context of temporal order. First, we consider a vector $\mathbf{L}$ of measured covariates taking values in $\mathcal{L} \subseteq \mathbb{R}^d$. After recording $\mathbf{L}$, and before time $t = 0$, we observe a binary exposure $A \in \{0,1\}$. Adopting the Neyman-Rubin potential outcomes framework (Neyman (1923), Rubin (1974)), let $T(a)$ denote the event time of interest under exposure assignment to exposure $A = a$. We assume that for $a \in \{0,1\}$, $T(a)$ takes value in $(0, \infty]$ meaning that, all patients start the study without having experienced the event of interest, and since we allow $T(a) = \infty$, some patients may never experience the event. Similarly, let $C(a)$ be the right-censoring time under exposure assignment $A = a$, where $C(a) \in [0, \infty]$. Since we allow $C(a) = 0$, patients may be censored immediately, for instance, if they are lost to follow-up just after the exposure $A$ is recorded. Define $\mathcal{O}_F := (\mathbf{L}, A, T(0), T(1), C(0), C(1))$ as the ideal data unit, with distribution function $P_{\mathcal{O}_F}$. We assume that each patient's potential event and censoring times are independent of all other patients exposures.
In practice, we do not observe the full potential outcomes $T(0)$, $T(1)$, $C(0)$, $C(1)$. Instead, the observed data are a coarsened version, where we only see the realized event or censoring time under the observed exposure. Let $T := T(A)$ and $C := C(A)$ represent the event and censoring time under the observed exposure $A$. Define $Y := \min\{T, C\}$ as the observed right-censored time and $\Delta := \mathbb{I}(T \leq C)$ as the event indicator. Thus the available data consist of $n$ independent and identically distributed observations $O_1, ..., O_n$ of the observed data unit $O := (\mathbf{L}, A, Y, \Delta)$, with distribution $P_0$, induced by $P_{\mathcal{O}_F}$ (formal derivation in Appendix A).

We denote summaries of $P_0$ with the subscript 0, for example,

$$E_0[f(O)] := E_{P_0}[f(O)],$$

and summaries of $P_{O_F}$ with subscript $O_F$. In cases where $f$ is a random function, the expectation $E_0[f(O)]$ should be understood as being taken with respect to the distribution of the random unit $O$, but not the function $f$.
In addition, we let $a \wedge b = \min\{a, b\}$, $\mathbb{P}_n$ be the empirical distribution corresponding to $O_1, O_2, \ldots, O_n$, and $Pf := \int f(o)\, dP(o)$ for any probability measure $P$ and $P$-measurable function $f$.
For a function $f : \mathbb{R} \to \mathbb{R}$ that is left-continuous at $x \in \mathbb{R}$, we define

$$f(x-) := \lim_{u \uparrow x} f(u).$$

## 2.2 Causal Assumptions

Randomized experiments are considered the gold standard for identifying and quantifying causal effects because the randomized assignment of treatment ensures exchangeability. In contrast, observational studies lack randomized treatment assignment, making causal interpretation more challenging due to potential confounding (Hernán and Robins (2020)).

When randomized experiments are not feasible, causal inference from observational data depends on the assumption that the study can be viewed as a conditionally randomized experiment. This requires meeting three key assumptions often referred to as identifiability conditions (Hernán and Robins (2020)):

- **Consistency**: the potential outcomes under treatment and control correspond to the observed outcomes.
  *Note: define the potential outcomes as* $Y^{a=1}$ *be the outcome variable that would have been observed under the treatment value* $a = 1$*, and* $Y^{a=0}$ *the outcome variable that would have been observed under the treatment value* $a = 0$.
  *Example where consistency does not hold:* If the definition of treatment is ambiguous or varies across subjects, then consistency is violated. For instance, if $A = 1$ refers to "receiving treatment", but the treatment can be administered in multiple ways (e.g., different dosages or frequencies), then $Y^{a=1}$ may not be well-defined.

- **Exchangeability**: there are no unmeasured confounders once we condition on the observed covariates.
  *Example where exchangeability does not hold:* Suppose a study examines the effect of a drug on heart disease, but smoking status, a strong confounder, is unmeasured. If smokers are more likely to take the drug and also have higher baseline risk of heart disease, failing to condition on smoking would introduce confounding bias.

- **Positivity**: each subject has a positive probability of receiving either treatment or control, given their covariates.
  *Example where positivity does not hold:* If a treatment is never given to individuals with certain characteristics (e.g., a chemotherapy drug not prescribed to patients with severe liver disease), then within that covariate stratum, the probability of receiving the treatment is zero, violating the positivity assumption.

In the case of censored data in an observational study, where the exposure $A$ is not randomized, the exposure-outcome, exposure-censoring, or outcome-censoring relationships are often confounded. To address this, we extend these assumptions with the following identification conditions, specific to fixed values of $\tau \in (0, \infty)$ and $a \in \{0,1\}$ (Westling, T. et al. (2023)):

(A1) **Restricted exchangeability** (exposure-outcome):

$$T(a)\mathbb{I}(T(a) \leq \tau) \perp\!\!\!\perp A \mid \mathbf{L} \tag{2.1}$$

The restriction to the event $T(a) \leq \tau$ allows aspects of the event mechanism occurring after time $\tau$ to depend on $A$, which is permitted if we only want to identify the survival probability up to time $\tau$.
*Example where exchangeability holds:* If a study on the effect of a drug on survival includes detailed patient characteristics (age, health status, genetics) that account for differences in treatment assignment, then given $\mathbf{L}$, treatment assignment is conditionally independent of survival time.

(A2) **Restricted exchangeability** (exposure-censoring):

$$C(a)\mathbb{I}(C(a) \leq \tau) \perp\!\!\!\perp A \mid \mathbf{L} \tag{2.2}$$

(A3) **Restricted exchangeability** (outcome-censoring):

$$T(a)\mathbb{I}(T(a) \leq \tau) \perp\!\!\!\perp C(a)\mathbb{I}(C(a) \leq \tau) \mid A = a, \mathbf{L} \tag{2.3}$$

This condition allows the event and censoring times to be dependent, as long as they are conditionally independent given $A$ and $\mathbf{L}$.
*Example where* $(A2)$ *and* $(A3)$ *hold:* If censoring is due to administrative reasons (e.g., study and date) rather than patient characteristics, and if all relevant patient characteristics are included in $\mathbf{L}$, then censoring is conditionally independent of both exposure and outcome given $\mathbf{L}$.

(A4) **Positivity trought strata:**

$$P_0(A = a \mid \mathbf{L}) > 0 \tag{2.4}$$

This condition ensures that all the covariate strata allow for both treatment and control groups, avoiding deterministic assignment.

(A5) **Positive uncensored probability**

$$P_{\mathcal{O}_F}(C(a) \geq \tau \mid \mathbf{L}) > 0 \tag{2.5}$$

This condition requires a positive probability of remaining uncensored in almost every stratum defined by $\mathbf{L}$, ensuring that no subset of individuals defined by $\mathbf{L}$ is deterministically censored before $\tau$, which could prevent estimation in those strata.

Consistency is fundamental in causal inference, ensuring that observed outcomes align with their respective potential outcomes under the received treatment. In standard identifiability conditions, it is explicitly stated to define counterfactual outcomes properly. However, in the censored data framework, assumptions $(A1)-(A5)$ focus on ensuring identifiability despite missing data due to censoring. Since these conditions already treat event and censoring times as counterfactual outcomes, consistency is implicitly assumed rather than explicitly stated. The primary concern in $(A1)-(A5)$ is establishing exchangeability and positivity under censoring, which ensures valid survival probability estimation despite unobserved data.

In conclusion, we can notice that this identification approach does not allow time-varying common causes of the event and censoring times, as the assumptions rely on conditional independence at fixed covariate values $\mathbf{L}$.
*Example where time-varying common causes are problematic:* suppose a study examines the effect of exercise on heart disease progression, but stress is a time-varying confounder that affects both exercise frequency and heart disease risk. If stress is not measured dynamically, standard methods that rely on baseline covariates may fail, leading to biased estimates of the causal effect.

## 2.3 Causal Parameters of Interest and Identification

We are interested in the causal effect of a binary treatment $A \in \{0,1\}$ on time-to-event outcomes. Specifically, we focus on two complementary causal estimands: the treatment-specific survival function and the quantile treatment effect (QTE). These parameters capture different aspects of the treatment effect, with the survival function providing a probabilistic characterization of event occurrence over time, while the QTE describes shifts in the distribution of event times due to treatment. Together, they offer a comprehensive understanding of how treatment influences both the timing and distribution of events.

### 2.3.1 Treatment-Specific Survival Function

The treatment-specific survival function $t \to \theta_{\mathcal{O}_F}(t,a) := P_{\mathcal{O}_F}(T(a) > t)$ for $a \in \{0,1\}$ and $t \in [0,\tau]$ (for some positive $\tau < \infty$) quantifies the population probability that an individual receiving treatment $A = a$ will not experience the event of interest by time $t$ (Westling, T. et al. (2023)).
Under the previous conditions, we can identify the causal parameter $\theta_{\mathcal{O}_F}(t,a)$ in terms of the distribution $P_0$ of the observes data unit.
Define

- $$F_{0,1}(t \mid a, \mathbf{l}) := P_0(Y \leq t, \Delta = 1 \mid A = a, \mathbf{L} = \mathbf{l}) \tag{2.6}$$
  represents the cumulative probability that an individual with treatment $A = a$ and covariates $\mathbf{L} = \mathbf{l}$ experience the event of interest by time $t$. The indicator $\Delta = 1$ ensures that only actual event occurences are counted.

- $$R_0(t \mid a, \mathbf{l}) := P_0(Y \geq t \mid A = a, \mathbf{L} = \mathbf{l}) \tag{2.7}$$
  represents the probability that an individual with treatment $A = a$ and coavariates $\mathbf{L} = \mathbf{l}$ is still at risk at time $t$, meaning they have neither experienced the event nor been censored before $t$.

- $$\Lambda_0(t \mid a, \mathbf{l}) := \int_0^t \frac{F_{0,1}(du \mid a, \mathbf{l})}{R_0(u \mid a, \mathbf{l})} \tag{2.8}$$
  represents the cumulative hazard function, which quantifies the accumulated risk of experiencing the event over time for individuals with

treatment $A = a$ and covariates $\mathbf{L} = \mathbf{l}$ (i.e. it describe how the event rate accumulates over time, adjusting for the number of individuals still at risk). It is defined as the integral of the instantaneous failure probability $F_{0,1}(du \mid a, \mathbf{l})$ divided by the risk set $R_0(u \mid a, \mathbf{l})$.

for each $(t, a, \mathbf{l})$ we have the following identification result.

**Proposition 2.3.1.** *If conditions* $(A1) - (A5)$ *hold for some* $a \in \{0,1\}$ *and* $\tau \in (0, \infty)$*, then*

$$P_{\mathcal{O}_F}(T(a) > t \mid \mathbf{L}) = S_0(t \mid a, \mathbf{L}) \quad P_0\text{-almost surely for all } t \in [0, \tau] \quad (2.9)$$

*where*

$$S_0(t \mid a, \mathbf{l}) := \prod_{(0,t]}^{st} \{1 - \Lambda_0(du \mid a, \mathbf{l})\} \quad (2.10)$$

*and so*

$$\theta_{\mathcal{O}_F}(t, a) = \theta_0(t, a) = \mathbb{E}_0[S_0(t \mid a, \mathbf{L})] \quad (2.11)$$

where $\prod^{\text{st}}$ denotes the Riemann-Stieltjes product integral (Gill and Johansen (1990)) and $S_0(t|a, \mathbf{L})$ represents the conditional survival probability given treatment status and covariates.
*Note: the Riemann-Stieltjes product integral is a generalization of the usual exponential function and is commonly used in survival analysis and stochastic processes. It provides a way to define cumulative effects in a multiplicative rather than an additive manner.*
Proprosition 2.3.1 is a combination of the G-formula from causal inference (Robins (1986), Gill and Robins (2001)) and the identification of a survival function in the context of dependent censoring (Bernan (1981), Dabrowska (1989)). The proof is provided in Appendix B.

### 2.3.2 Quantile Treatment Effect

The QTE, defined as the difference in quantiles of the potential event times under treatment and control, captures heterogeneity in treatment effects across different points in the event-time distribution. Unlike the survival function, which provides an aggregate probability of event occurrence, the QTE highlights how treatment modifies specific percentiles of the event time distribution, offering a nuanced view of its impact.

Under the assumptions made in the previous section, the QTE can be identified using the G-formula approach. Specifically, we define $Q_{\mathcal{O}_F}(p, a)$

as the $p$-th quantile of $T(a)$, which, under identification conditions, can be expressed in terms of observed data as:

$$Q_{\mathcal{O}_F}(p, a) = Q_0(p, a) = \inf\{t : \theta_0(t, a) \leq 1 - p\}. \tag{2.12}$$

The QTE at quantile $p$ is then given by:

$$\text{QTE}(p) = Q_0(p,1) - Q_0(p,0). \tag{2.13}$$

By considering the QTE alongside survival functions, we provide a more comprehensive causal analysis of treatment effects on time-to-event outcomes, accommodating both average and distributional treatment effects. This approach is particularly useful for assessing heterogeneity in treatment impact, as it reveals whether the treatment effect is uniform across individuals or varies depending on event-time quantiles.

# Chapter 3

# Estimation of Survival Curves

## Introduction

In this chapter, we focus on the non-parametric estimation of the treatment-specific G-computed survival functions $\{\theta_0(t, a) : t \in [0, \tau]\}$, a crucial task in survival analysis and various applied fields such as medical research, economics, and social sciences. The methods and results discussed in this chapter, are primarily drawn from the work of (Westling, T. et al. (2023)), where the non-parametric estimation is based on the efficient influence function (EIF), a powerful tool for understanding and estimating functionals of distributions.

The chapter is structured into three main parts, each addressing key aspects of survival curve estimation. First, we introduce the process of estimating the survival curve itself, starting with the computation of the efficient influence function. We then propose a one-step cross-fitted estimator, a modern approach that improves the estimation of survival curves by reducing bias and enhancing accuracy. Additionally, we discuss methods for enforcing monotonicity in the estimated survival curve, ensuring that the curve behaves consistently with the expected properties of survival functions.

Next, we turn to the asymptotic properties of the estimator. In particular, we establish the consistency and asymptotic linearity of the survival curve estimator, under certain regularity conditions. These properties are crucial for understanding the performance of the estimator as the sample size grows and ensuring that the estimator converges to the true survival function.

Finally, we focus on the issue of statistical inference. In this section, we discuss the construction of valid confidence intervals for the survival curve

estimator, allowing for robust statistical inference in practical applications. These confidence intervals provide a means to quantify the uncertainty surrounding the estimated survival curve, facilitating hypothesis testing and making the results more interpretable in real-world scenarios.

## 3.1 Estimation of the Survival curves

### 3.1.1 Efficiency Calculation

The efficient influence function (EIF) is a fundamental concept in semiparametric estimation. It quantifies the sensitivity of a parameter estimate to small perturbations in the underlying distribution while satisfying the efficiency bound. Mathematically, the EIF is the canonical gradient of the parameter functional in the model's score space. It is a mean-zero function that characterizes the asymptotic behavior of estimators and plays a crucial role in constructing optimal estimators (Hines et al. (2022)).

Theorem 3.1.1, whose proof can be found in Appendix B, provides the efficient influence function of the survival probability.

Now, we will define some key quantities that will be useful throughout the proof of the theorem and the construction of the EIF. We begin by defining

$$\pi_0(a \mid \mathbf{l}) := P_0(A = a \mid \mathbf{L} = \mathbf{l}).$$

We use the fact that, for any $(a, \mathbf{l})$ such that $S_0(\tau- \mid a, \mathbf{l}) > 0$,

$$P_{\mathcal{O}_F}(C \geq t \mid A = a, \mathbf{L} = \mathbf{l}),$$

can be identified under the causal identifiability condition as:

$$P_{\mathcal{O}_F}(C \geq t \mid A = a, \mathbf{L} = \mathbf{l}) = G_0(t \mid a, \mathbf{l}) := \prod_{[0,t)}^{\text{st}} \{1 - H_0(du \mid a, \mathbf{l})\},$$

for any $t \in [0, \tau]$, where

$$H_0(u \mid a, \mathbf{l}) := \int_{[0,u]} \left( \frac{S_0(s- \mid a, \mathbf{l})}{S_0(s \mid a, \mathbf{l})} \right) \cdot \frac{F_{0,0}(ds \mid a, \mathbf{l})}{R_0(s \mid a, \mathbf{l})}.$$

*Note: the formula represents a cumulative hazard type function related to the treatment-specific survival analysis. It is composed by:*

1. $S_0(s \mid a, \mathbf{l})$*: the survival function, which gives the probability of surviving beyond time* $s$ *given treatment* $a$ *and covariates* $\mathbf{l}$*.*

2. $S_0(s- \mid a, \mathbf{l})$*: The left-limit of the survival function, capturing the probability just before* $s$*.*

3. $F_{0,0}(ds \mid a, \mathbf{l})$*: A cumulative distribution function (CDF) that describes the probability of the event occurring at time* $s$*.*

4. $R_0(s \mid a, \mathbf{l})$*: The risk set, representing the number of individuals still at risk just before time* $s$*.*

*The formula is an integral that accumulates contributions to the cumulative hazard, adjusting for the survival probability at different points in time.*

We emphasize that $G_0$ is defined as the left-continuous conditional survival function of $C$, whereas $S_0$ is the right-continuous conditional survival function of $T$.

Even when $t = \tau$, the identification remains valid as long as $S_0(t- \mid a, \mathbf{l}) > 0$, since the product integral is taken up to, but not including, $t$.

In the following theorem, we present the nonparametric efficient influence function of $\theta_0(t, a_0)$, where we use $a_0$ rather than $a$ to denote the exposure value of interest in order to avoid confusion between the values of the random variable $A$ and the specific $a_0$ at which we evaluate $\theta_0$.

**Theorem 3.1.1.** *If there exists* $\eta > 0$ *such that*

$$\min\left\{\pi_0(a_0 \mid \mathbf{l}), G_0(t \mid a_0, \mathbf{l})\right\} \geq \eta,$$

*for* $P_0$*-almost every* $\mathbf{l}$ *such that* $S_0(t \mid a_0, \mathbf{l}) > 0$*, then* $\theta_0(t, a_0)$ *is a pathwise differentiable parameter in a nonparametric model with efficient influence function*

$$\phi^*_{0,t,a_0} := \phi_{0,t,a_0} - \theta_0(t, a_0),$$

*where* $\phi_{0,t,a_0}(y, \delta, a, \mathbf{l})$ *equals*

$$S_0(t \mid a_0, \mathbf{l})\Bigg[1 - \frac{\mathbb{I}(a = a_0)}{\pi_0(a \mid \mathbf{l})}\Bigg(\frac{\mathbb{I}(y \leq t, \delta = 1)}{S_0(y \mid a, \mathbf{l}) \cdot G_0(y \mid a, \mathbf{l})} - \int_0^{\min(t,y)} \frac{\Lambda_0(du \mid a, \mathbf{l})}{S_0(u \mid a, \mathbf{l}) \cdot G_0(u \mid a, \mathbf{l})}\Bigg)\Bigg] \quad (3.1)$$

In the next sections, we explore how this efficient influence function can be leveraged to construct estimators with desirable statistical properties.

### 3.1.2 Cross-Fitted One-Step Estimator

The efficient influence function $\phi^*_{0,t,a_0}$ involves three variation-independent nuisance functions: $S_0$, $G_0$, and $\pi_0$.
We start noticing that since

$$S_0(t \mid a, \mathbf{l}) := \prod_{(0,t]}^{\text{st}} \{1 - \Lambda_0(du \mid a, \mathbf{l})\}$$

$\Lambda_0$ and $S_0$ are in one-to-one correspondence, then estimating $S_0$ provides an estimator of $\Lambda_0$ and vice versa. Given estimators $S_n$, $G_n$, and $\pi_n$ of $S_0, G_0$, and $\pi_0$, respectively, there are multiple possible asymptotically linear and efficient estimators of $\theta_0(t, a)$. We discuss the estimation of these functions in the following section.

Denoting by $\phi_{n,t,a}$ the function $\phi_{0,t,a}$ with $S_0, \Lambda_0, G_0$, and $\pi_0$ replaced by their respective estimators, the standard one-step estimator is given by $\mathbb{P}_n\phi_{n,t,a}$. This was the approach taken by (Hubbard et al. (2000), Bai et al. (2013)).
We employ a cross-fitted version of the one-step estimator, which we define below, because in observational studies, it is often impractical to specify correct parametric models for nuisance parameters a priori, necessitating the use of data-adaptive estimators. However, these estimators frequently do not fall within small function classes, making it challenging to satisfy both the following remainder term conditions necessarly for asymptotic linearity. The asymptotic linearity (a property prove in section 3.2.2) of estimators of this type depends on the nuisance estimators in two key ways:

- The second-order remainder term must be negligible, requiring nuisance parameters to converge rapidly to their true values.
- The empirical process remainder term must also be negligible, which can be ensured if the nuisance estimators belong to sufficiently small function classes with high probability.

Cross-fitting mitigates this issue by removing constraints on the complexity of nuisance estimators (see, e.g., Bickel (1982), Robins et al. (2008), Zheng and van der Laan (2011), Dìaz (2019)).

#### The cross-fitting algorithm

For a deterministic integer $K \in \{2,3,\ldots,\lfloor n/2 \rfloor\}$, we perform the following steps:

- Randomly partition the indices $\{1,2,\ldots,n\}$ into $K$ disjoint sets $\mathcal{V}_{n,1}$, $\mathcal{V}_{n,2}$, …, $\mathcal{V}_{n,K}$, with cardinalities $n_1$, $n_2$, …, $n_K$.
- Ensuring each set has a nearly equal number of elements, satisfying $|n_k - n/K| \leq 1$ for each $k$, and that the number of folds $K$ remains bounded as $n$ grows.
- For each $k \in \{1,2,\ldots,K\}$:
  - Define the training set $\mathcal{T}_{n,k} := \{O_i : i \notin \mathcal{V}_{n,k}\}$ for fold $k$.
  - Define $S_{n,k}, G_{n,k}, \pi_{n,k}$, and $\Lambda_{n,k}$ as nuisance estimators estimated using only observations from the training set $T_{n,k}$ and $\phi_{n,k,t,a}$ as the function $\phi_{0,t,a}$ in which these nuisance estimators have substituted their true counterparts.

The cross-fitted one-step estimator $\theta_n$ is then defined pointwise as:

$$\theta_n(t,a) := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_{n,k}}\phi_{n,k,t,a}(O_i).$$

The final cross-fitted one-step estimator $\theta_n(t,a)$ involves a summation over the folds and the observations in each fold. This averaging process is crucial because it aggregates the contributions from each fold's nuisance function estimator. By averaging over the folds, we ensure that no single fold dominates the estimation process, which helps reduce bias and variance.
Once the nuisance functions are estimated, $\theta_n(t,a)$ can be efficiently computed for various time points $t$ since the same nuisance function estimators are reused across different $t$ values.
*Note: The computation of the integral in $\phi_{n,k,t,a}$ depends on the form of $S_{n,k}$. If $S_{n,k}$ is a step function, the integral reduces to a summation. Otherwise, it can be approximated as a sum, a very fine grid can be used for this purpose with little impact on computational cost.*

### 3.1.3 Data Adaptive estimation of Nuisance Functions

In the proposed estimator, we need to estimate three nuisance parameters: the conditional survival functions $S_0$ and $G_0$ for the event time and censoring distributions, respectively, given exposure and covariates, as well as the propensity score $\pi_0$ of exposure given covariates.

To estimate the propensity score $\pi_0$, we employ a combination of parametric, semiparametric, and nonparametric methods through the Super-Learner algorithm (Breiman (1996); Van der Laan et al. (2007)). In Westling, T. et al. (2023), the authors propose an iterative SuperLearner algorithm that integrates multiple candidate nuisance estimators for the survival function $S_0$ and censoring function $G_0$. This algorithm is specifically designed to overcome the limitations of existing ensemble learning methods for right-censored data. Their approach addresses key limitations in the following ways:

- **Estimation over an interval**: Unlike traditional methods that estimate the survival function at a single fixed time point, their algorithm estimates the entire survival function over an interval. This provides a more comprehensive characterization of survival dynamics, especially in situations where the hazard is not constant over time.
- **Accommodating discrete and continuous event time distributions**: Many existing methods assume a fully discrete or fully continuous event time distribution. The proposed approach is more flexible, as it can accommodate both types of distributions as well as mixed types, making it more applicable to diverse real-world datasets.
- **Joint optimization of $S_0$ and $G_0$**: The algorithm optimizes both the survival function $S_0$ and censoring function $G_0$ simultaneously and iteratively. By doing so, it refines the estimation of both functions together, which has the potential to improve overall accuracy compared to methods that treat these functions separately.

Central to the ensemble method are representations of $S_0$ and $G_0$ as minimizer of oracle risk functions. For the following result, we define $\mathcal{C}_\tau$ as the set of functions from $[0,\tau] \times [0,1] \times \mathcal{W}$ to $[0,1]$.
Specifically, the minimization problems are:

$$S^* = \arg\min_{S \in \mathcal{C}_\tau} P_0 L_{S,G_0}, \quad G^* = \arg\min_{G \in \mathcal{C}_\tau} P_0 M_{G,S_0},$$

where $L_{S,G}$ and $M_{G,S}$ are the loss functions that measure the discrepancy between the estimated and true survival and censoring functions. These loss functions are defined as follows:

$$L_{S,G} : (\mathbf{l}, a, y, \delta) \to \int_0^\tau S(t \mid a, \mathbf{l}) \left[ S(t \mid a, \mathbf{l}) - 2 \left( 1 - \frac{\delta \mathbb{I}(y \leq t)}{G(y|a, \mathbf{l})} \right) \right] dt,$$

$$M_{S,G} : (\mathbf{l}, a, y, \delta) \to \int_0^\tau G(t \mid a, \mathbf{l}) \left[ G(t \mid a, \mathbf{l}) - 2 \left( 1 - \frac{(1-\delta)\mathbb{I}(y \leq t)}{S(y|a, \mathbf{l})} \right) \right] dt,$$

If conditions $(A1)-(A5)$ in supplementary material hold for each $a \in \{0,1\}$, then $S^*(t \mid a, \mathbf{l}) = S_0(t \mid a, \mathbf{l})$ for $P_0$-almost every $(a, \mathbf{l})$ and all $t \leq \tau$, and $G^*(t \mid a, \mathbf{l}) = G_0(t \mid a, \mathbf{l})$ for $P_0$-almost every $(a, \mathbf{l})$ and all $t \leq \tau$ such that $S_0(t- \mid a, \mathbf{l}) > 0$.
These loss functions ensure that the estimated survival and censoring functions align closely with the true functions, taking into account both the event times and the censoring indicator $\delta$.

### The iterative algorithm

Were $G_0$ known, an optimal weighted combination of $p$ candidate estimators $S_n^{(1)}, S_n^{(2)}, \ldots, S_n^{(p)}$ of $S_0$ could be found by minimizing the cross-validated empirical risk $\mathbb{P}_n L_{S,G_0}$ over $S$ in the set $\Pi_S$ of convex combinations:

$$\sum_{j=1}^{p} \alpha_j S_n^{(j)}$$

for $\alpha$ in the $p$-dimensional simplex.
Here, by cross-validated we mean that the sample is split into $K$ folds, candidate estimators are each trained holding out each fold, evaluated on the held-out fold, and these held-out evaluations are used to compute the empirical mean $P_n L_{S,G}$.
Were $S_0$ known, an analogous procedure could be used to find an optimal weighted combination of $q$ candidate estimators $G_n^{(1)}, G_n^{(2)}, \ldots, G_n^{(q)}$ in the set $\Pi_G$ of convex combinations:

$$\sum_{j=1}^{q} \alpha_j G_n^{(j)}$$

of $G_0$ for $\alpha$ in the $q$-dimensional simplex. Since $S_0$ and $G_0$ are not known in practice, we propose the following iterative strategy:

1. **Initialization**: Start by obtaining an initial estimator $G_{n,0}^*$ for the censoring function $G_0$ using a nonparametric procedure, such as a survival tree or kernel-based method.

2. **Iterative Updates**: At each iteration $k$, perform the following steps:

   - Compute the updated estimator for $S_{n,k}^*$ by minimizing the empirical loss function $P_n L_{S,G_{n,k-1}^*}$

- Compute the updated estimator for $G^*_{n,k}$ by minimizing the empirical loss function $P_n M_{G,S^*_{n,k}}$

3. **Convergence**: The algorithm terminates when the changes in $S^*_n$ and $G^*_n$ between iterations fall below a predefined threshold, indicating that the algorithm has converged.

To evaluate the integrals in the loss functions $L_{S,G}$ and $M_{G,S}$, we approximate them using Riemann sums over a fine grid. This numerical approximation allows us to compute the loss functions efficiently.

While obtaining the cross-validated estimates of the candidate learners' survival and censoring functions can be computationally expensive, these estimates are computed only once at the beginning. Therefore, the additional computational burden is minimal. Beyond that, the primary computational effort in this procedure is the optimization process to find the optimal convex combinations of the candidate learners. This optimization step is typically much less computationally demanding than obtaining the cross-validated estimates themselves. Thus, the overall computational cost of the proposed algorithm is not substantially higher than that of a standard SuperLearner approach.

### 3.1.4 Enforcing monotonicity of the estimator

The function $t \to \theta_0(t, a)$ is necessarily monotone nonincreasing for each $a \in \{0, 1\}$ and takes values in $[0, 1]$. However, the proposed estimator $\theta_n(t, a)$ is generally neither guaranteed to lie in $[0, 1]$ nor to be monotone in $t$ in any given sample. We ensure that our final estimator satisfies the above parameter constraints as follows.

- First, we construct $\theta_n(t, a)$ as defined above for each $t \in \mathcal{T}_n$, where $\mathcal{T}_n$ is the set of unique values of $Y_1, Y_2, \ldots, Y_n$.
- Second, for each $t \in \mathcal{T}_n$ and $a \in \{0,1\}$, we define

$$\theta^+_n(t, a) = \begin{cases} \theta_n(t, a) & \text{if } \theta_n(t, a) \in [0, 1] \\ 1 & \text{if } \theta_n(t, a) > 1 \\ 0 & \text{if } \theta_n(t, a) < 0 \end{cases}$$

- Next, for each $a \in 0, 1$, we define $\{\theta^\circ_n(t, a) : t \in T_n\}$ as the projection of $\{\theta^+_n(t, a) : t \in T_n\}$ onto the space of nonincreasing functions using isotonic regression.

*Note: Isotonic regression is a nonparametric technique used to fit a monotonic function to data while preserving the order of the observations. It minimizes a loss function, typically the squared error, subject to the constraint that the estimated function is nonincreasing. This ensures that the final estimator adheres to the expected monotonicity properties of the target function.*

- For any $t \in (0, \tau]$, we then define $\theta_n^\circ(t, a)$ as the evaluation of the right-continuous stepwise interpolation of $\{\theta_n^\circ(t, a) : t \in T_n\}$.

The projected estimator $\theta_n^\circ$ is guaranteed to be no farther from $\theta_0$ than $\theta_n$ in every finite sample, and if the true function is strictly decreasing, then the initial and projected estimators are asymptotically equivalent (Westling et al. (2020)). Therefore, in what follows, we focus on providing large-sample results for $\theta_n$, since results for the isotonized estimator $\theta_n^\circ$ are identical in view of the general results of (Westling et al. (2020)).

## 3.2 Large sample properties

### 3.2.1 Consistency of the Survival Estimator

In this section, we analyze the asymptotic properties of the estimator. Specifically, we establish conditions under which $\theta_n(t, a)$ is consistent for $\theta_0(t, a)$, both for a fixed $t$ and uniformly over $t$.

#### Assumptions

(B1) **Convergence of Estimated Functions:** there exist fixed limit functions $\pi_\infty$, $G_\infty$, and $S_\infty$ such that:

$$\max_k \mathbb{E}_0 \left[ \frac{1}{\pi_{n,k}(a \mid \mathbf{L})} - \frac{1}{\pi_\infty(a \mid \mathbf{L})} \right]^2 \xrightarrow{P} 0$$

$$\max_k \mathbb{E}_0 \left[ \sup_{u \in [0,t]} \left| \frac{1}{G_{n,k}(u \mid a, \mathbf{L})} - \frac{1}{G_\infty(u \mid a, \mathbf{L})} \right| \right]^2 \xrightarrow{P} 0$$

$$\max_k \mathbb{E}_0 \left[ \sup_{u \in [0,t]} \left| \frac{S_{n,k}(t \mid a, \mathbf{L})}{S_{n,k}(u \mid a, \mathbf{L})} - \frac{S_\infty(t \mid a, \mathbf{L})}{S_\infty(u \mid a, \mathbf{L})} \right| \right]^2 \xrightarrow{P} 0$$

With these conditions we require that the estimated functions converge in an appropriate sense to limit functions, which is used to control

certain empirical process terms.
*Note: the expectations are with respect to* $\mathbf{L}$*, and not with respect to the randomness of the nuisance estimators.*

(B2) **Bounded Propensity and Censoring Functions:** there exists a constant $\eta > 0$ such that, with high probability, for $P_0-$almost all $\mathbf{l}$:

$$\pi_{n,k}(a \mid \mathbf{l}) \geq \frac{1}{\eta} \quad \text{and} \quad \pi_{\infty}(a \mid \mathbf{l}) \geq \frac{1}{\eta}$$

$$G_{n,k}(t \mid a, \mathbf{l}) \geq \frac{1}{\eta} \quad \text{and} \quad G_{\infty}(t \mid a, \mathbf{l}) \geq \frac{1}{\eta}$$

This ensures that estimated functions are bounded away from zero in all subpopulation of patients defined by $\mathbf{L}$. In practice, truncation can be used to enforce this condition.
*Note: this assumption must hold for all different* $a$*.*

(B3) **Double Robustness Property:** for $P_0$-almost all $\mathbf{l}$, there exist measurable sets $\mathcal{S}_L, \mathcal{G}_L \subseteq [0, t]$ such that:

- $\mathcal{S}_L \cup \mathcal{G}_L = [0, t]$
- $\Lambda_0(u \mid a, \mathbf{l}) = \Lambda_{\infty}(u \mid a, \mathbf{l})$ for all $u \in \mathcal{S}_L$
- $G_0(u \mid a, \mathbf{l}) = G_{\infty}(u \mid a, \mathbf{l})$ for all $u \in \mathcal{G}_L$

If $\mathcal{S}_L$ is a strict subset of $[0, t]$, then $\pi_0(a \mid \mathbf{l}) = \pi_{\infty}(a \mid \mathbf{l})$. In combination with the first assumption, this implies that for almost all $(t, \mathbf{l})$, either $S_n(t \mid a, \mathbf{l})$ is consistent, or both $G_n(t \mid a, \mathbf{l})$ and $\pi_n(a \mid \mathbf{l})$ are consistent.
*Note: This is a form of double-robustness of the estimator of* $\theta_n$ *to estimation of the nuisances* $S_0$ *and* $(G_0, \pi_0)$*, because, in particular,* $\theta_n$ *is consistent when either* $S_n$ *is consistent everywhere or both* $G_n$ *and* $\pi_n$ *are consistent everywhere. However, is a relaxed form of doubly-robusness since none of the limit function need to be identically equal to their counterparts.*

(B4) **Stronger Uniform Consistency Condition:** a stronger condition on $S_{n,k}$ ensures uniform consistency:

$$\max_k \mathbb{E}_0 \left[ \sup_{u \in [0,t]} \sup_{v \in [0,u]} \left| \frac{S_{n,k}(u \mid a, \mathbf{L})}{S_{n,k}(v \mid a, \mathbf{L})} - \frac{S_{\infty}(u \mid a, \mathbf{L})}{S_{\infty}(v \mid a, \mathbf{L})} \right| \right]^2 \xrightarrow{P} 0.$$

This condition ensures that the convergence of $S_n$ to its limit holds uniformly over time, which is critical for achieving uniform consistency.

Given the previous assumption the following important result establish the consistency of the proposed survival estimator (the proof can be found in the supplamentary material of Westling, T. et al. (2023))

**Theorem 3.2.1.** *(Consistency) If conditions* $(B1) - (B3)$ *hold, then*

$$\theta_n(t,a) \xrightarrow{P} \theta_0(t,a)$$

*If condition* $(B4)$ *also holds, then*

$$\sup_{u\in[0,t]} |\theta_n(u,a) - \theta_0(u,a)| \xrightarrow{P} 0$$

### 3.2.2 Asymptotic linearity of the survival estimator

We now present additional conditions under which $\theta_n(t,a)$ is asymptotically linear for fixed $t$ and uniformly over $t$.

#### Assumptions

We define

$$r_{n,t,a,1} := \max_k \mathbb{E}_0 \left| \Big(\pi_{n,k}(a \mid \mathbf{L}) - \pi_0(a \mid \mathbf{L})\Big) \cdot \Big(S_{n,k}(t \mid a, \mathbf{L}) - S_0(t \mid a, \mathbf{L})\Big) \right|$$

and

$$r_{n,t,a,2} := \max_k \mathbb{E}_0 \left| S_{n,k}(t \mid a, \mathbf{L}) \cdot \int_0^t \left( \frac{G_0(u \mid a, \mathbf{L})}{G_{n,k}(u \mid a, \mathbf{L})} - 1 \right) \cdot \left( \frac{S_0}{S_{n,k}} - 1 \right) (du \mid a, \mathbf{L}) \right|$$

Based on these quantities we introduce additional conditions for asymptotic linearity

(B5) It hold that

$$r_{n,t,a,1} = o_P(n^{-\frac{1}{2}}) \quad \text{and} \quad r_{n,t,a,2} = o_P(n^{-\frac{1}{2}}) \tag{3.2}$$

this condition requires that the rates of convergence of $(S_n - S_0)(\pi_n - \pi_0)$ and $(S_n - S_0)(G_n - G_0)$ to zero be faster than $n^{-1/2}$.
*Note: one approach to satisfying this condition is to assume that these nuisance functions fall in known parametric or semiparametric families such that existing estimators achieve the stipulated rates.*

*In many cases, it is not possible to construct parametric or semiparametric models that are known to be correctly specified. In those cases condition* $(B5)$ *can be satisfied when using data-adaptive estimators since* $(S_n - S_0)(\pi_n - \pi_0)$ *and* $(S_n - S_0)(G_n - G_0)$ *can converge faster than* $n^{-1/2}$ *even if* $S_n$, $\pi_n$, *and/or* $G_n$ *converge slower than* $n^{-1/2}$. *This is a primary benefit of constructing an estimator based on the nonparametric efficient influence function. However, achieving these rates of convergence is not guaranteed when using data-adaptive estimators. Whether the rate is achieved depends on the dimension of the covariates, structure such as smoothness, additivity, or sparsity of* $S_0$, $G_0$ *and* $\pi_0$, *and the extent to which the nuisance estimators adapt to this structure. Since the true structure of these functions is often unknown, a suggested approach is to combine multiple candidate parametric, semiparametric, and nonparametric estimators using cross-validation, which has the potential to achieve the same rate as the best candidate estimator.*

(B6) It holds that

$$\sup_{u\in[0,t]} r_{n,u,a,1} = o_P(n^{-\frac{1}{2}}) \quad \text{and} \quad \sup_{u\in[0,t]} r_{n,u,a,2} = o_P(n^{-\frac{1}{2}})$$

We have the following result concerning the asymptotic linearity of $\theta_n(t,a)$.

**Theorem 3.2.2.** *(Asymptotic linearity) If conditions* $(B1)-(B2)$ *hold with* $S_\infty = S_0$, $G_\infty = G_0$ *and* $\pi_\infty = \pi_0$, *and condition* $(B5)$ *also holds, then*

$$\theta_n(t,a) = \theta_0(t,a) + \mathbb{P}_n\phi^*_{0,t,a} + o_P(n^{-1/2}). \tag{3.3}$$

*In particular,* $n^{1/2}[\theta_n(t,a) - \theta_0(t,a)]$ *then converges in distribution to a normal random variable with mean zero and variance*

$$\sigma_0^2(t,a) := P_0\phi^{*2}_{0,t,a}. \tag{3.4}$$

*If in addition conditions* $(B4)$ *and* $(B6)$ *also hold, then*

$$\sup_{u\in[0,t]} \left|\theta_n(u,a) - \theta_0(u,a) - \mathbb{P}_n\phi^*_{0,u,a}\right| = o_P(n^{-1/2}). \tag{3.5}$$

*In particular, the process* $\{n^{1/2}[\theta_n(u,a)-\theta_0(u,a)] : u \in [0,t]\}$ *then converges weakly as a process in the space* $\ell^\infty([0,t])$ *of uniformly bounded functions on* $[0,t]$ *to a tight mean-zero Gaussian process with covariance function*

$$(u,v) \mapsto P_0(\phi^*_{0,u,a}\phi^*_{0,v,a}). \tag{3.6}$$

## 3.3 Inference for the survival curves

Thanks to the asymptotic results established in the previous section, we can conduct valid inference for $\theta_0(t,0)$, $\theta_0(t,1)$ and their contrasts.
A common approach in statistical inference is to construct a Wald-type asymptotic $(1-\alpha)$-level confidence interval for $\theta_0(t,a)$. Specifically we propose,

$$\theta_n^\circ(t,a) \pm z_{1-\alpha/2} n^{-1/2} \sigma_n(t,a) \tag{3.7}$$

where $z_p$ denotes the $p$-quantile of the standard normal distribution and

$$\sigma_n^2(t,a) := \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{V}_{n,k}} [\phi_{n,k,t,a}(O_i) - \theta_n^\circ(t,a)]^2 \tag{3.8}$$

is a cross-fitted influence function-based estimator of the asymptotic variance $\sigma_0^2(t,a)$.
Since $\theta_0(t,a)$ represents a probability, it has been observed in classical settings (Anderson et al. (1982)) that constructing Wald-type confidence intervals on the logistic probability scale improves finite-sample coverage. To apply this approach, we define the functions

$$\text{expit}(x) := \frac{\exp(x)}{1+\exp(x)} \quad \text{for any } x \in \mathbb{R}, \tag{3.9}$$

and

$$\text{logit}(u) := \log(u) - \log(1-u) \quad \text{for any } u \in (0,1), \tag{3.10}$$

Using these transformations, we define

$$\tilde{\sigma}_n(t,a) := \frac{\sigma_n(t,a)}{\theta_n^\circ(t,a)[1-\theta_n^\circ(t,a)]}, \tag{3.11}$$

and propose the transformed Wald-type interval $[l_n(t,a), u_n(t,a)]$ given by

$$\text{expit}\left\{\text{logit}[\theta_n^\circ(t,a)] \pm z_{1-\alpha/2} n^{-1/2} \tilde{\sigma}_n(t,a)\right\} \tag{3.12}$$

for any $(t,a)$ for which $\theta_n^\circ(t,a) \in (0,1)$. For boundary cases:

- If $\theta_n^\circ(t,a) = 0$, we set

$$l_n(t,a) := 0 \quad \text{and} \quad u_n(t,a) := \min_s \{u_n(s,a) : u_n(s,a) > 0\},$$

- if $\theta_n^\circ(t,a) = 1$, we set

$$l_n(t,a) := \max_s \{l_n(s,a) : l_n(s,a) < 1\} \quad \text{and} \quad u_n(t,a) := 1$$

The endpoints of this interval are strictly contained between 0 and 1 for any $(t, a)$ such that $\theta_n^\circ(t, a) \in (0,1)$. If the pointwise statement of Theorem 3.2.2 holds for both $a = 0$ and $a = 1$, then

$$n^{1/2}\left[\theta_n(t,0) - \theta_0(t,0)\right] \text{ and } n^{1/2}\left[\theta_n(t,1) - \theta_0(t,1)\right] \tag{3.13}$$

converge jointly to a mean-zero bivariate normal distribution. This result allows us to apply the delta method to perform inference on causal effects of the form $h(\theta_0(t,0), \theta_0(t,1))$ for any differentiable $h$. Furthermore, this result extends to inference on functionals of the treatment-specific survival function using the functional delta method, which we explore in the following chapter.

Overall, this approach provides robust and well-calibrated inference for survival probabilities and causal contrasts, particularly in finite samples where naive Wald intervals may perform poorly.

## 3.4 Numerical studies

### Simulating Baseline Confounders

To simulate the baseline confounders, we generate a vector $\mathbf{L} = (L_1, L_2, L_3)$, where each component represents a continuous confounder. The process for generating these variables is as follows:

1. **Simulating $L_1$ (Age):** We simulate $L_1$ to represent age using a Beta distribution. Specifically, $L_1$ follows a scaled Beta distribution $\text{Beta}(1.1, 1.1)$ with a range from 20 to 80 years. The scaling ensures that the simulated values of $L_1$ are within the desired age range:

   $$L_1 \sim 20 + 60 \cdot \text{Beta}(1.1, 1.1)$$

   Here, $\text{Beta}(1.1, 1.1)$ produces a distribution skewed towards the center (i.e., the values are more likely to be closer to the middle of the range [20, 80]).

2. **Simulating $L_2$ (BMI):** Given that $L_1 = l_1$, we simulate $L_2$ using a Beta distribution with parameters that depend on $L_1$. This conditional simulation reflects the potential relationship between the confounders:

   $$L_2 \sim 18 + 32 \cdot \text{Beta}\left(1.5 + \frac{l_1}{20}, 6\right)$$

The parameters of the Beta distribution are influenced by $L_1$, where a larger value of $l_1$ (age) results in a higher shape parameter for the Beta distribution, thus shifting the distribution of $L_2$.

*Note: $L_2$ is meant to represent Body Mass Index (BMI). The choice of distribution reflects how BMI may vary with age. As individuals age, their BMI distribution can shift, potentially due to changes in lifestyle, metabolism, or health conditions. The conditional relationship between $L_2$ and $L_1$ allows the BMI to change in a realistic manner based on age, with older individuals having a broader distribution of BMI values.*

3. **Simulating $L_3$ (Risk Score for the Event of Interest):** Similarly, $L_3$ is simulated conditional on $L_1 = l_1$. In this case, we use a Beta distribution that adjusts based on the absolute difference between $L_1$ and 50. This simulates a potential relationship where $L_3$ is influenced by how far $L_1$ (age) is from 50:

$$L_3 \sim 10 \cdot \text{Beta}\left(1.5 + \frac{|l_1 - 50|}{20}, 3\right)$$

Here, the term $|l_1 - 50|$ introduces a non-linearity in the relationship between $L_1$ and $L_3$, emphasizing the difference from the midpoint value of 50 years.

*Note: $L_3$ represents a risk score for the event of interest, which could be something like the probability of experiencing a particular health outcome, such as a cardiovascular event, based on age. The risk score is assumed to vary with age in a non-linear manner, as the likelihood of certain events may change dramatically around midlife (e.g., around 50 years of age). The absolute difference $|l_1 - 50|$ models this non-linear effect, where the risk might increase or decrease more sharply as age moves further from 50, which is commonly observed in many health-related risk assessments. The Beta distribution helps model the skewed nature of risk scores, where extreme values (both low and high) might be less likely than more moderate values.*

The Beta distribution, Beta$(a, b)$, used in the simulations is parameterized by two shape parameters $a$ and $b$, with the mean given by $\frac{a}{a+b}$. In each case, we are tailoring the distribution to represent different patterns of confounding based on the values of $L_1$, the baseline confounder.

### Logit Model for Treatment Assignment

After generating the baseline confounders, we define a logistic model to determine the probability of treatment assignment, denoted as $P_0(A = 1 \mid L = l)$. The logit function for this probability is specified as follows:

$$\text{logit}(P_0(A = 1 \mid L = l)) = -1 + \log\left(1 + \exp\left(-20 + \frac{l_1}{10}\right) + \exp\left(-3 + \frac{l_3}{2}\right)\right)$$

Here, the logit function represents the natural logarithm of the odds of receiving treatment $A = 1$ given the confounders $L = l$. This model accounts for non-linear relationships between the confounders and treatment assignment:

- **Effect of $L_1$ (age):** The term $\exp\left(20 + \frac{l_1}{10}\right)$ models the influence of age on treatment probability. Since this term is inside a logarithmic transformation, it implies a steep increase in treatment probability as age ($L_1$) increases, particularly at higher values.

- **Effect of $L_3$ (risk score):** The term $\exp\left(-3 + \frac{l_3}{2}\right)$ captures the impact of the risk score on treatment assignment. As $L_3$ increases, the exponential function ensures a rapid increase in the odds of receiving treatment, meaning individuals with higher risk scores are more likely to be assigned treatment.

- **Baseline adjustment:** The constant term $-1$ shifts the overall log-odds of treatment assignment, ensuring that the probability remains within a reasonable range. The inclusion of $\log(1+\cdot)$ prevents extremely high probabilities while maintaining the non-linearity induced by the exponential terms.

This model effectively captures complex dependencies between age, risk factors, and treatment assignment. The use of exponential and logarithmic transformations ensures that treatment probability remains bounded while allowing for steep changes based on individual characteristics.

### Generating Event and Censoring Times

We considered two processes for generating event and censoring times:

**Process 1: Proportional Hazards Model**

- **Event Time $T$:** We simulated the time-to-event variable $T$ using a Weibull distribution with a shape parameter of 1.5. The hazard function for the event time is defined as follows:

$$\begin{aligned}\lambda_0(t \mid a, \mathbf{l}) = t^{0.5} \exp\{&-5.02 - 0.36a + 0.02(l_1 - 50) + 0.015a(l_1 - 50) \\ &+ 0.05(l_2 - 30) - 0.025a(l_2 - 30) \\ &+ 0.2(l_3 - 5) - 0.15a(l_3 - 5) \\ &+ 0.01(l_1 - 50)(l_3 - 5)\}\end{aligned} \tag{3.14}$$

  The hazard function models how the risk of the event occurring at time $t$ depends on treatment status ($a$) and the baseline confounders ($l_1, l_2, l_3$). Specifically:

  - **Effect of $l_1$ (age):** The term $\exp(0.02(l_1 - 50))$ captures the effect of age, shifting the hazard function as age increases. The interaction term $0.015a(l_1 - 50)$ allows the treatment effect to vary with age.
  - **Effect of $l_2$ (BMI):** The term $\exp(0.05(l_2 - 30))$ models the direct influence of BMI on event risk. The interaction term $-0.025a(l_2 - 30)$ suggests that treatment might mitigate the effect of BMI on the hazard.
  - **Effect of $l_3$ (risk score):** The term $\exp(0.2(l_3 - 5))$ indicates that a higher risk score increases the hazard, while the interaction $-0.15a(l_3 - 5)$ suggests that treatment reduces this effect.
  - **Nonlinear interaction:** The term $0.01(l_1 - 50)(l_3 - 5)$ captures the joint effect of age and risk score, introducing additional complexity into the hazard function.

  This hazard function ensures that the event probability varies realistically with age, BMI, and risk score while allowing treatment to have differential effects based on patient characteristics.

- **Censoring Time $C$:** To account for right-censoring, we simulated the censoring time $C$ using a Weibull distribution with a similar hazard function:

$$\begin{aligned}h_0(t \mid a, \mathbf{l}) = t^{0.5} \exp\{&-4.87 - 0.4a + 0.02(l_1 - 50) - 0.015a(l_1 - 50) \\ &+ 0.05(l_2 - 30) - 0.025a(l_2 - 30) \\ &+ 0.2(l_3 - 5) - 0.15a(l_3 - 5) \\ &+ 0.01(l_1 - 50)(l_3 - 5)\}\end{aligned} \tag{3.15}$$

The censoring hazard follows a similar structure but differs in key parameters:

- A slightly higher baseline hazard ($-4.87$ vs. $-5.02$), meaning censoring generally occurs slightly earlier than the event.
- A stronger treatment effect ($-0.4a$ vs. $-0.36a$), implying treated individuals tend to have longer censoring times.
- The interaction term between age and treatment has an opposite sign ($-0.015a(w_1 - 50)$ instead of $+0.015a(w_1 - 50)$), suggesting that censoring is less common for older treated individuals.

To ensure realistic observation windows, censoring times were truncated at $C = 24$, meaning any censoring time exceeding this threshold was set to 24. This reflects a maximum follow-up period, ensuring that individuals are not observed indefinitely.

**Process 2: Non-Proportional Hazards Model**

- **Censoring Time $C$:** We simulated $C$ from an exponential distribution with a rate parameter:

$$\lambda_C = \exp\left[\theta_1 + 0.3a + \log\left(1 + \exp\left(\frac{30 - l_1}{3}\right)\right) + \left(\frac{l_3 - 5}{4}\right)^2\right]$$

The censoring time was truncated at $C = 24$.

- **Event Time $T$:** Given $A = 1$ and $\mathbf{L} = \mathbf{l}$, we simulated $T$ from the survival function:

$$S_0(t \mid 1, \mathbf{l}) := S_0(\psi(t, \mathbf{l}) \mid 0, \mathbf{l})$$

where $S_0(t \mid 0, \mathbf{l})$ is the survival function of an exponential distribution with rate:

$$\exp\left(\theta_1 - \frac{|w_1 - 60|}{10} - 2\log(l_2) + \frac{l_3}{2}\right)$$

We define $\psi(t, \mathbf{l})$ as:

$$\psi(t, \mathbf{L}) = \begin{cases} t - \frac{t^2}{2r}\left[1 - \delta(\mathbf{l})\right], & 0 \leq t < r \\ \frac{r}{2}\left[1 + \delta(\mathbf{l})\right] + (t - r)\delta(\mathbf{l}), & r \leq t < r + \tau(\mathbf{l}) \\ t + \frac{r}{2}\left[1 + \delta(\mathbf{l})\right] + \tau(\mathbf{l})\delta(\mathbf{l}) & \\ \quad - \left[r + \tau(\mathbf{l})\right]\left[2 - \delta(\mathbf{l})\right] & t \geq r + \tau(\mathbf{l}) \\ \quad + \frac{1}{t}\left[1 - \delta(\mathbf{l})\right]\left[r + \tau(\mathbf{l})\right]^2, & \end{cases} \tag{3.16}$$

where we have set $r = 1.5$ months, the initial period where treatment is becoming effective.
The covariate-dependent treatment durability $\tau(\mathbf{l})$ (i.e. how long the treatment remain effective) is given by:

$$\tau(\mathbf{l}) := \exp\left(2 - \frac{1}{2}\log\left(1 + \exp\left(\frac{l_1 - 55}{5}\right)\right) - \frac{1}{10}\log\left(1 + \exp\left(\frac{l_2 - 30}{3}\right)\right)\right)$$

and ensures treatment lasts longer for younger patients and shorter for older ones. The covariate-dependent maximal treatment effectiveness $\delta(\mathbf{l})$ (i.e. how much the treatment reduces hazard) is:

$$\delta(\mathbf{l}) := \text{expit}\left(\theta_2 + \frac{1}{2}\log\left(1 + \exp\left(\frac{l_1 - 55}{5}\right)\right) + \frac{1}{4}\log\left(1 + \exp\left(\frac{l_2 - 30}{3}\right)\right)\right)$$

Eventually, we set

$$\theta_0 = -4.81, \quad \theta_1 = -6.28, \quad \theta_2 = -1.64$$

To ensure:

- Average censoring rate: $E_0[P(C = 12 \mid A = 0, \mathbf{L})] = 0.2$
- Average observed event rate: $E_0[P(T \leq C \mid A = 0, \mathbf{L})] = 0.15$
- Counterfactual risk ratio at $t = 12$ is 0.7 (i.e. treatment reduces risk by 30

### Simulation Setup

The simulation consists of creating 1000 datasets using the specified processes for $n = 250, 500, \ldots, 1500$.
For each dataset, the propensity score (i.e., the probability of receiving

| Algorithm Name | Algorithm Description |
|---|---|
| SL.mean | Marginal mean |
| SL.glm | Main-terms logistic regression |
| SL.gam | Main-terms generalized additive model |
| SL.earth | Multivariate adaptive regression splines |
| SL.xgboost | Extreme gradient boosting |

Table 3.1: Algorithms used for the estimation of the propensity score.

treatment given observed covariates) was estimated using the SuperLearner algorithm with the methods listed in Table 3.1.

For each dataset, the conditional survival curves were estimated using four different methods:

1. **A correctly specified Cox proportional hazards model**: This model correctly represents the true relationship between covariates and survival time.

2. **An incorrectly specified proportional hazards model with main terms only**: This is a misspecified model that does not fully capture the relationships in the data.

3. **Survival random forest**: A machine learning-based approach to estimating survival curves.

4. **The iterative SuperLearner**: Described in Chapter 3.1.3, this method combines parametric survival models, semiparametric proportional hazards models, generalized additive Cox models, and survival random forest.
   *Note: For the iterative process, we used 5-fold cross-validation, survival random forest as the initial estimator, and limited the recursive procedure to fifteen iterations.*

To evaluate the estimators, the accuracy of each candidate estimator $S_n$ of $S_0$ was measured using the **root mean squared error (RMSE)**:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[S_n(t \mid A_i, \mathbf{L}_i) - S_0(t \mid A_i, \mathbf{L}_i)\right]^2}$$

for each $t \in \{0.5, 1, \ldots, 12\}$. The same process was applied to the estimator of $G_0$.

Subsequently, counterfactual survival curves (i.e., survival under different treatment conditions) were estimated using two methods:

1. **G-computation**: This method estimated the survival function using the empirical distribution of covariates.

2. **Cross-fitted one-step estimator**:(5-fold) This method used different combinations of conditional survival estimators.

Finally, for each method, the following quantities were recorded:

1. **Estimated survival probability** for control and treatment groups at $t = 12$.

2. **Estimated risk ratio** at $t = 12$, comparing survival probabilities between groups.

3. **Confidence intervals** for the survival probability at $t = 12$.

4. **Uniform confidence bands** over the time range [0,12], showing the range of uncertainty across different time points.

### Simulation results

Before delving into the actual results of the numerical studies, we need to make two important observations:

1. In the first simulation scenario, correctly specified proportional hazards models produce conditional survival estimators that converge at the parametric rate of $n^{-1/2}$, satisfying condition (B5). The proposed SuperLearner library includes a correctly specified proportional hazards estimator. Therefore, if the method successfully selects this estimator from the candidate library, it should also achieve the required rate of convergence.

2. In the second scenario, the proportional hazards estimators are inconsistent. The censoring time follows a generalized additive proportional hazards model, and the proposed SuperLearner library includes a corresponding estimator. Thus, if the proposed estimator adapts appropriately, it should attain the necessary rate of convergence. However, the event time is generated by a complex nonproportional hazards mechanism that is not captured by standard semiparametric survival models. As a result, we did not include a correctly specified semiparametric estimator in the SuperLearner library, making the rate of convergence of the estimator uncertain.

With these considerations in mind, we now turn to the actual results of the numerical studies.
Figure 3.1 display the average RMSE over 1000 simulations and over $t \in \{0.5, 1, \ldots, 12\}$ for the four estimators of the conditional survival of event and censoring and for the two simulation settings.

In the proportional hazards setting (left two panels), the correctly specified Cox proportional hazards estimator with interactions achieved the lowest RMSE among the four estimators. The proposed SuperLearner exhibited slightly higher RMSE for event survival and a similar RMSE for censoring survival, suggesting that it effectively selected the correctly specified estimator from the candidate library. The RMSEs of the incorrectly specified Cox estimator and the survival random forest (RF) were larger, though the RMSE of RF decreased as $n$ increased.

In the nonproportional hazards setting (right two panels), the SuperLearner estimator achieved the lowest RMSE among the four estimators for both event and censoring survival. The RMSEs of both Cox model estimators did not decrease significantly with $n$ since neither was correctly specified. While the RMSE of the survival random forest (RF) was slightly higher than that of the SuperLearner for event survival, it performed significantly worse for censoring survival.
Overall, we conclude that the iterative SuperLearner successfully adapted to the proportional hazards setting when that model was appropriate and effectively leveraged alternative candidate estimators to outperform the proportional hazards models in scenarios where the data did not follow a proportional hazards structure.

Figure 3.2 illustrates the properties of the estimators and confidence intervals for the control survival curve.

We begin by examining the results for the scenario where the data were generated from a proportional hazards model with interactions (left column). The biases of the proposed method (first row), using both the Cox model with interactions and the G-computed Cox estimator with interactions, remained within Monte Carlo error of zero across all sample sizes. The iterative SuperLearner, used for nuisance estimation, exhibited bias of less than 0.5% for all sample sizes. In contrast, the biases of the G-computation Cox estimator without interactions and the G-computation survival random

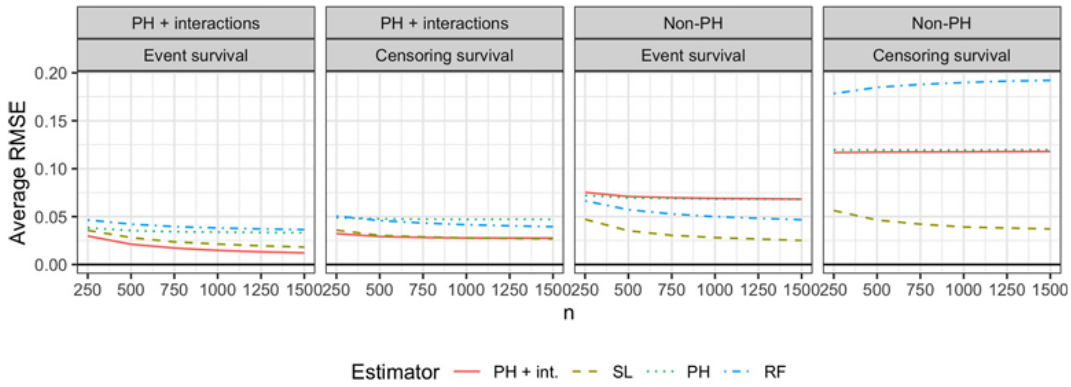


Figure 3.1: The average root mean squared error (RMSE), as defined in the text, of the four conditional survival estimators as a function of sample size $n$. The four panels correspond to the two simulation settings and to which conditional survival is being considered. In the legend, "PH" refers to the Cox proportional hazards model, "int." refers to interactions, "RF" refers to survival random forest, and "SL" refers to our iterative SuperLearner.

forest exceeded 1%. Notably, the bias of the former remained relatively constant as $n$ increased, suggesting inconsistency. This outcome was expected, as the true conditional survival curves contained interactions that this estimator failed to capture.

Estimators based on the one-step method exhibited greater standard deviation (second row) compared to those based on G-computation. Consequently, the G-computation approach using a correctly specified Cox model estimator yielded the lowest mean squared error. However, this result depends heavily on the correct specification of the Cox model. The pointwise coverage (third row) of confidence intervals constructed using our method closely matched the nominal 95% level for all sample sizes of 500 and larger. The uniform coverage of the proposed method (bottom row) ranged between 90% and 93%, indicating slight undercoverage even at large sample sizes. We attribute this to poor coverage at small values of $t$, highlighting the challenge of constructing equi-precision confidence bands when the standard deviation is small at the boundary.

We now examine the results for the scenario where the data were generated from a nonproportional hazards model (right column of Figure 3.2). In this case, the bias (first row) of the proposed method, using the iterative SuperLearner for nuisance estimators, remained within Monte Carlo error of zero across all sample sizes. This result is somewhat surprising, given that

the SuperLearner library does not contain a correctly specified estimator for the conditional survival of the event.
The G-computation Cox estimators and the proposed method using the Cox estimator with interactions for nuisance estimation were biased, as these nuisance estimators were inconsistent in this setting (see Figure 3.1). The G-computation random forest exhibited the largest bias, though it decreased with increasing sample size. Notably, the G-computation random forest had the smallest standard deviation among the estimators considered (second row), leading to comparable mean squared errors across estimators. The pointwise coverage of our estimator (third row) was excellent for all sample sizes considered. It is particularly surprising that the proposed estimator maintained good coverage when using the Cox estimator for conditional survival, given that these estimators were inconsistent. Typically, we would expect coverage to deteriorate when the nuisance estimators are inconsistent, and we anticipate that this effect may become more pronounced at larger sample sizes. The uniform coverage of our estimator (bottom row) was slightly below the nominal level for small sample sizes but improved and reached the nominal level as the sample size increased.

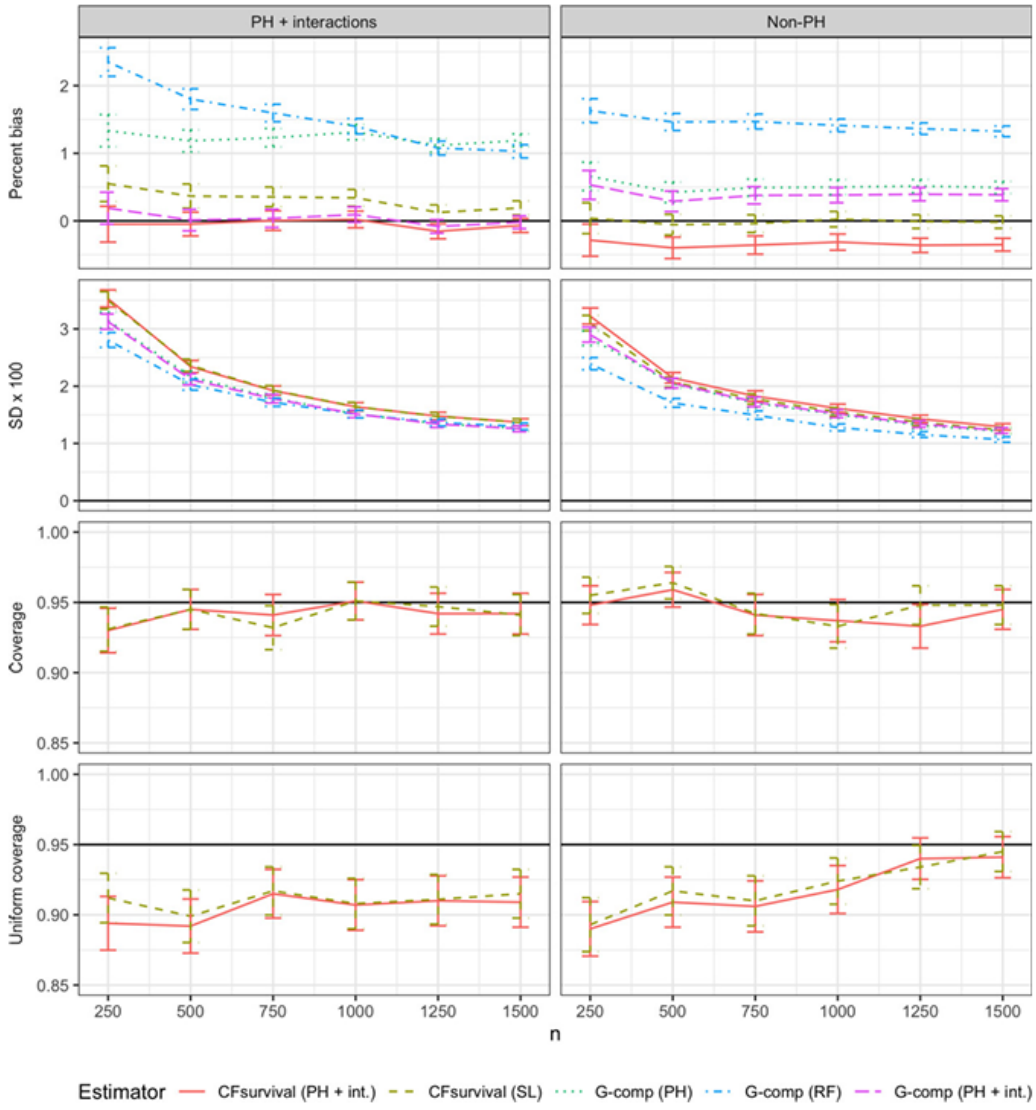


Figure 3.2: Properties of five of the estimators of the counterfactual control survival as a function of sample size. Columns correspond to the two simulation settings. From top to bottom, the rows contain: percent bias, standard deviation, pointwise coverage, and uniform coverage. The first three rows correspond to inference at time $t = 12$. "CFsurvival" is the method developed here, and "G-comp" is G-computation. Parentheticals indicate the estimator used for the conditional survival(s), with shorthand defined in the Figure 1 caption. Vertical bars represent 95% confidence intervals taking into account uncertainty due to conducting a finite number of simulations.

# Chapter 4

# Estimation of Quantiles and Quantile Treatment Effects

## Introduction

In this chapter, we explore the non-parametric estimation of quantiles and the quantile treatment effect (QTE) using the survival definition, which offers a robust approach to understanding the heterogeneity in treatment effects across different quantiles of the outcome distribution. This non-parametric method enables us to capture how treatment influences various subgroups, providing a more comprehensive perspective than average treatment effects. Unlike conventional approaches that often rely on regression techniques, our methodology leverages the survival function to estimate quantiles, offering a more direct and intuitive way to assess the distributional impact of a treatment.

We begin by introducing the non-parametric estimation of quantiles within the survival framework, followed by an examination of the quantile treatment effect (QTE). The second part of the chapter focuses on the asymptotic properties that ensure the consistency and normality of these estimators. Finally, we address inference-related considerations, discussing methods for constructing valid confidence intervals while maintaining the asymptotic properties of the estimators.

## 4.1 Estimation of Quantiles and Quantile Treatment Effect

In the previous chapter, we introduced an estimator for the conditional survival function $\hat{\theta}_n^\circ(t, a)$, where $A$ represents the treatment. In Proposition 2.3.1, we demonstrated that the estimator derived from the coarsened data structure is equivalent to the estimator obtained from the ideal data structure. Building on this, as discussed in Section 2.3.2, the same reasoning extends to marginal quantile estimation, since it is defined through the survival function. This equivalence further extends to the estimation of the quantile treatment effect, reinforcing the consistency and applicability of our approach across different estimation frameworks. We start recalling the following definition of quantile

$$Q_0(p, a) = \inf\{t : \theta_0(t, a) \leq 1 - p\}. \tag{4.1}$$

thus marginal quantile estimate $\hat{Q}_n(p, a)$ is obtained by finding the smallest $t$ such that:

$$\hat{\theta}_n^\circ(t, a) \leq 1 - p. \tag{4.2}$$

Since $\hat{\theta}_n^\circ(t, a)$ is a step function, it does not always attain exactly the value $1-p$. To address this, we employ linear interpolation, a widely used method that provides a simple yet effective approximation of the quantile.

The linear interpolation procedure is as follows:

1. Identify two consecutive time points $t_{\text{low}}$ and $t_{\text{high}}$ such that:

$$\hat{\theta}_n^\circ(t_{\text{low}}, a) > 1 - p \quad \text{and} \quad \hat{\theta}_n^\circ(t_{\text{high}}, a) \leq 1 - p.$$

2. Compute the estimate $\hat{Q}_n(p, a)$ using the interpolation formula:

$$\hat{Q}_n(p, a) = t_{\text{low}} + \left( \frac{(1 - p) - \hat{\theta}_n^\circ(t_{\text{low}}, a)}{\hat{\theta}_n^\circ(t_{\text{high}}, a) - \hat{\theta}_n^\circ(t_{\text{low}}, a)} \right) (t_{\text{high}} - t_{\text{low}}).$$

Alternative approaches, such as higher-order interpolation and kernel smoothing, exist. However, linear interpolation is chosen for the following reasons:

1. **Computational efficiency**: Linear interpolation is computationally simple, requiring only basic arithmetic operations, which is particularly advantageous in large-scale datasets.

2. **Preservation of monotonicity**: Linear interpolation preserves monotonicity, ensuring a meaningful and unique estimate for $\hat{Q}_n(p, a)$.

3. **Local approximation**: It relies only on two neighboring points, making it a localized approximation that does not assume unnecessary global trends.

Given these considerations, linear interpolation represents a balance between simplicity, efficiency, and accuracy, making it the preferred method in most cases. Although higher-order methods or kernel smoothing could potentially offer more precision, they come with increased computational cost and complexity, which are not necessary for this context.

## 4.2 Large Sample Properties

In this section, we introduce a change of notation to reflect the fact that the results we are presenting are not limited to the estimator previously proposed, but instead are more general in nature.

Quantiles are traditionally defined using the cumulative distribution function (CDF) as follows:

$$F^{-1}(p, a) = Q_0(p, a) = \inf\{t : F(t, a) \geq p\}$$

Since $S(t, a) = 1 - F(t, a)$, we can alternatively use the definition we introduced earlier:

$$S^{-1}(p, a) = Q_0(p, a) = \inf\{t : \theta_0(t, a) \leq 1 - p\} = \inf\{t : S(t, a) \leq 1 - p\}$$

In this section, we will use both definitions to ensure that the asymptotic properties derived for the survival function estimator also extend to the Marginal Quantile Treatment Effects (MQTE) and the Quantile Treatment Effect (QTE).

### 4.2.1 Consistency of Quantile Estimators and Quantile Treatment Effect (QTE)

The following result is a key step in understanding how quantile estimators behave as the sample size increases. For a more detailed explanation and the proof of this result, please refer to the Appendix B.

**Lemma 4.2.1** (Weak Convergence of Quantile Estimators)**.** *For any sequence of survival functions $(S_n)$, weak convergence $S_n \to S$ implies weak convergence of the corresponding quantile functions:*

$$S_n^{-1} \to S^{-1}.$$

This lemma ensures that if the sequence of estimated survival functions $S_n(t)$ converges weakly to the true survival function $S(t)$, then the estimated quantile function $S_n^{-1}(p)$ also converges weakly to the true quantile function $S^{-1}(p)$.
Building on this result, we note that the estimator $\hat{\theta}_n^{\circ}(t, a)$, which estimates the true parameter $\theta_0(t, a)$, under the previous assumptions is also weakly convergent. Consequently, we can deduce that the corresponding quantile estimator $\hat{Q}_n(p, a)$ also converges weakly to the true quantile function $Q_0(p, a)$, as expressed by the following:

$$\hat{Q}_n(p, a) \to Q_0(p, a).$$

This result guarantees that our quantile estimates become increasingly accurate as the sample size grows, reinforcing the reliability of the estimator in large samples.

### Consistency of the Quantile Treatment Effect (QTE)

Next, we examine the consistency of the Quantile Treatment Effect (QTE), which quantifies the effect of treatment at different quantiles of the outcome distribution.

**Lemma 4.2.2** (Consistency of the Quantile Treatment Effect (QTE))**.** *Let $\hat{S}_n^{-1}(p,0)$ and $\hat{S}_n^{-1}(p,1)$ be the estimators for the quantiles under control and treatment, respectively. If $S_n(t,0)$ and $S_n(t,1)$ converge weakly to their true survival functions, then the QTE estimator satisfies:*

$$\hat{QTE}(p) = \hat{S}_n^{-1}(p,1) - \hat{S}_n^{-1}(p,0) \to QTE(p),$$

*where the true Quantile Treatment Effect is given by:*

$$QTE(p) = S^{-1}(p,1) - S^{-1}(p,0).$$

*Thus, the estimator for the QTE is consistent at the given quantile.*

The consistency of $\hat{\text{QTE}}(p)$ follows from the weak convergence of the estimated survival functions.

Since $\hat{S}_n^{-1}(p,1)$ and $\hat{S}_n^{-1}(p,0)$ converge weakly to their true counterparts $S^{-1}(p,1)$ and $S^{-1}(p,0)$, their difference also converges in probability to the true QTE.
As a result, given that the estimators $\hat{\theta}_n^{\circ}(t,0)$ and $\hat{\theta}_n^{\circ}(t,1)$ weakly converge to $\theta_0(t,0)$ and $\theta_0(t,1)$ the estimator of the QTE is consistent.
Therefore, $\hat{\text{QTE}}(p)$ serves as an asymptotically valid measure of the distributional impact of treatment, reinforcing its reliability for policy evaluation and causal inference

### 4.2.2 Asymptotic Normality of Quantile Estimators

We now focus on establishing the asymptotic normality of quantile estimators. Since the relationship

$$S(t,a) = 1 - F(t,a)$$

holds, we present the theoretical results primarily using the cumulative distribution function rather than the survival function.
We aim to derive the asymptotic distribution of the quantile function by leveraging the asymptotic normality of the survival function estimator. This derivation relies on the application of the delta method. However, since the quantile function is a functional transformation of the survival function, we employ a more general extension known as the functional delta method, which allows us to handle transformations of stochastic processes in an infinite-dimensional setting.

**Theorem 4.2.3** (Functional Delta method)**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be normed linear spaces. Let $\psi : \mathbb{D}_\psi \subset \mathbb{D} \to \mathbb{E}$ be Hadamard differentiable at $\theta$ tangentially to $\mathbb{D}_0$. Let $T_n : \Omega_n \to \mathbb{D}_\psi$ be maps such that $r_n(T_n - \theta) \to T$ for some sequence of numbers $r_n \to \infty$ and a random element $T$ that takes its values in $\mathbb{D}_0$. Then $r_n(\psi(T_n) - \psi(\theta)) \to \psi'_\theta(T)$. If $\psi'_\theta$ is defined and continuous on the whole space $\mathbb{D}$, then we also have $r_n(\psi(T_n) - \psi(\theta)) = \psi'_\theta(r_n(T_n - \theta)) + o_p(1)$.*

The application of the functional delta method requires that the functional in question satisfies the condition of Hadamard differentiability

**Definition 4.2.4.** A map $\psi : \mathbb{D}_\psi \to \mathbb{E}$, defined on a subset $\mathbb{D}_\psi$ of a normed space $\mathbb{D}$ that contains $\theta$, is called **Hadamard differentiable** at $\theta$ if there exists a continuous linear map $\psi'_\theta : \mathbb{D} \to \mathbb{E}$ such that

$$\left| \frac{\psi(\theta + th_t) - \psi(\theta)}{t} - \psi'_\theta(h) \right|_{\mathbb{E}} \to 0 \quad .$$

as $t \to 0$, for every sequence $h_t \to h$ □

see Appendix D for more detailes.
To establish this property and ensure the validity of the method, we first need to prove that the functional meets the necessary criteria. To do so, we need to give some definitions and prove the following two lemmas.
For a nondecreasing function $F \in D[a,b], [a,b] \subset [-\infty, \infty]$, and a fixed $p \in \mathbb{R}$, let $\psi(F) \in [a,b]$ be an arbitrary point in $[a,b]$ such that

$$F(\psi(F)-) \leq p \leq F(\psi(F))$$

and the natural domain $\mathbb{D}_\psi$ of the resulting map $\psi$ is the set of all nondecreasing $F$ such that there exists a solution to the pair of inequalities.

**Lemma 4.2.5** (Hadamard Differentiability of a Single Quantile)**.** *Let $F \in \mathbb{D}_\psi$ be differentiable at a point $\xi_p \in (a,b)$ such that $F(\xi_p) = p$, with positive derivative. Then $\psi : \mathbb{D}_\psi \subset D[a,b] \to \mathbb{R}$ is Hadamard-differentiable at $F$ tangentially to the set of functions $h \in \mathcal{D}[a,b]$ that are continuous at $\xi_p$, with derivative*

$$\psi'_F(h) = -\frac{h(\xi_p)}{F'(\xi_p)}$$

The proof from van der Vaart A.W. (1998) can be found in Appendix C.
*Note: An analogous result for the survival function can be derived by considering a perturbation $h$ of the survival function, leading to the following outcome:*

$$\psi'_S(h) = \frac{h(\xi_p)}{S'(\xi_p)}$$

Now, instead of a single quantile we can consider the quantile function $F \to (F^{-1}(p))_{p_1<p<p_2}$, for fixed numbers $0 \leq p_1 \leq p_2 \leq 1$.
Given an interval $[a,b] \subset \mathbb{R}$, let $\mathbb{D}_1$ be the set of all restrictions of distribution functions on $\mathbb{R}$ to $[a,b]$, and let $\mathbb{D}_2$ be the subset of $\mathbb{D}_1$ of distribution functions of measures that give mass 1 to $(a,b]$.

**Lemma 4.2.6** (Hadamard Differentiability of the Entire Quantile Function)**.** *We consider two cases:*

1. ***Finite Support Case:*** *Let $0 < p_1 < p_2 < 1$, and let $F$ be continuously differentiable on the interval*

$$[a,b] = [F^{-1}(p_1) - \varepsilon, F^{-1}(p_2) + \varepsilon]$$

*for some $\varepsilon > 0$, with a strictly positive derivative $f$. Then, the inverse map*

$$G \mapsto G^{-1}$$

*as a mapping*

$$\mathbb{D}_1 \subset D[a,b] \to \ell^\infty[p_1, p_2]$$

*is Hadamard differentiable at $F$ tangentially to $\mathcal{C}[a,b]$.*

2. ***Unbounded Support Case:*** *Suppose $F$ has compact support $[a,b]$ and is continuously differentiable on its support with a strictly positive derivative $f$. Then, the inverse map*

$$G \mapsto G^{-1}$$

*as a mapping*

$$\mathbb{D}_2 \subset D[a,b] \to \ell^\infty(0,1)$$

*is Hadamard differentiable at $F$ tangentially to $\mathcal{C}[a,b]$.*

*In both cases, the derivative is given by the mapping*

$$h \mapsto -\left(\frac{h}{f}\right) \circ F^{-1}.$$

The proof from van der Vaart A.W. (1998) can be found in Appendix C. Thus, the asymptotic normality of an estimator for a distribution function directly implies the asymptotic normality of the corresponding quantile estimators.

In particular, since the estimator of the survival function satisfies asymptotic linearity, we have:

$$\sqrt{n}(\hat{S}_n(t,a) - S(t,a)) = \sqrt{n}\mathbb{P}_n\phi^*_{0,t,a} + o_p(1).$$

From the Hadamard differentiability result, we recall that the quantile function is differentiable, and for a small perturbation $h$ of $S$, its derivative is given by:

$$\psi'_S(h) = \frac{h(Q_0(p))}{S'(Q_0(p),a)}.$$

Applying this result to the perturbation $h = \hat{S}_n - S$ and using a first-order Taylor expansion, we obtain:

$$\begin{aligned}\sqrt{n}(\hat{Q}_n(p,a) - Q_0(p,a)) &= \frac{\sqrt{n}(\hat{S}_n(Q_0(p,a)) - S(Q_0(p,a)))}{S'(Q_0(p,a),a)} + o_p(1) \\ &= \frac{\sqrt{n}\mathbb{P}_n\phi^*_{0,Q_0(p,a),a}}{S'(Q_0(p,a),a)} + o_p(1).\end{aligned}$$

Since the leading term is an average of i.i.d. random variables, the asymptotic variance of $\hat{Q}_n(p)$ follows as:

$$\begin{aligned}\mathrm{Var}(\sqrt{n}\hat{Q}_n(p,a)) &= \mathrm{Var}\left(\frac{\sqrt{n}\mathbb{P}_n\phi^*_{0,Q_0(p,a),a}}{S'(Q_0(p,a),a)}\right)\\ &= \frac{\mathrm{Var}(\sqrt{n}\mathbb{P}_n\phi^*_{0,Q_0(p,a),a})}{S'(Q_0(p,a),a)^2}\\ &= \frac{P_0(\phi^*_{0,Q_0(p,a),a})^2}{S'(Q_0(p,a),a)^2}.\end{aligned}$$

*Note: In general, if $\mathbb{P}_n$ represents an empirical measure based on i.i.d. observations, then for any function $\phi$:*

$$Var(\sqrt{n}\mathbb{P}_n\phi) = P_0(\phi^2),$$

*where $P_0$ denotes the expectation under the true probability measure.*

Dividing both sides by $n$, we obtain the asymptotic variance of $\hat{Q}_n(p,a)$ itself:

$$\tilde{\sigma}^2_{0,a} = \mathrm{Var}(\hat{Q}_n(p,a)) = \frac{P_0(\phi^*_{0,Q_0(p,a),a})^2}{nS'(Q_0(p,a),a)^2}$$

*It is important to note that the variance of $\hat{Q}_n(p,a)$ depends on the sample size $n$, specifically shrinking at a rate of $1/n$. This is a natural and expected property in asymptotic statistics:*

- The estimator $\hat{Q}_n(p,a)$ is *consistent*, meaning it converges to the true quantile $Q_0(p,a)$ as $n \to \infty$.
- The presence of $1/n$ indicates that, as the sample size increases, the variance of $\hat{Q}n(p,a)$ decreases, meaning that the estimator becomes more precise.
- The quantity $\frac{P_0(\phi^*0,Q_0(p,a),a)^2}{S'(Q_0(p,a),a)^2}$ represents the *asymptotic variance* when scaled by $n$.

Thus, the shrinking variance reflects the increasing accuracy of the quantile estimator with larger samples. This behavior is fundamental in statistical inference, ensuring that the estimator becomes more reliable as data accumulates.

## 4.3 Inference for Quantile Estimates

Having established the methodology for estimating quantiles and the quantile treatment effect (QTE) in the previous sections, we now turn our attention to the important issue of inference. In this section, we explore methods for constructing valid confidence intervals, all while ensuring that the inference remains valid even in finite samples.

### 4.3.1 Constructing Confidence Intervals for Quantiles

#### Asymptotic confidence interval

A common approach, as we presented in the previous chapter, is to construct a Wald-type asymptotic $(1-\alpha)-$level confidence interval for $Q_0(p,a)$. Specifically we propose

$$\hat{Q}_n(p,a) \pm z_{1-\alpha/2} \cdot \tilde{\sigma}_{n,a}$$

where $z_p$ is the critical value from the standard normal distribution corresponding to the desired confidence level (e.g., for 95%, $z_{0.025} \approx 1.96$), and the estimated variance is:

$$\frac{\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{V}_{n,k}}[\phi_{n,k,Q_0(p,a),a}(O_i) - \theta_n^{\circ}(t,a)]^2}{nS'(\hat{Q}(p,a),a)^2}$$

from which we can derive the estimated standard error.

#### Bootstrap confidence interval

While the standard asymptotic approach provides a solid foundation for inference, it may not always be suitable in finite samples or when the asymptotic assumptions do not hold perfectly. In such cases, the bootstrap method offers a valuable alternative, relying less on asymptotic results.

Various methods exist for computing the standard error of an estimator, such as Greenwood's formula, pointwise confidence intervals, or log-log transformations. These methods are useful when the distribution of the estimator is well-defined. However, in our case, determining the distribution is challenging, which makes the bootstrap method an attractive option.

The bootstrap is a widely used statistical tool that quantifies uncertainty around an estimator without relying on specific distributional assumptions. However, for the bootstrap to yield reliable results, a few key assumptions must hold:

1. **The observed sample is representative of the population**: The original sample should accurately reflect the underlying population from which it was drawn.

2. **Independence and identical distribution of observations**: This assumption is necessary because the bootstrap method resamples independent observations to create new datasets.

3. **The sample size should be sufficiently large**: While the bootstrap does not require a fixed sample size, it works best with a sufficiently large sample to capture the population's variability.

4. **The estimator should be consistent**: As the sample size increases, the estimator should converge to the true population parameter.

5. **No strong assumptions on the underlying distribution**: The bootstrap is non-parametric and does not require specific distributional assumptions about the data.

The assumption of consistency is particularly important. Since $\hat{\theta}_n^\circ(t, a)$ is a consistent estimator of the survival function $\theta_0(t, a)$, the quantile estimator $\hat{Q}_n(p, a)$ will also be consistent. This is because, as the sample size grows, the survival function converges to the true survival function. Consequently, the quantile estimate, based on the survival estimate, converges to the true quantile. This convergence holds because finding the quantile is essentially a root-finding problem based on the survival function, and the root of the consistent survival function will also be consistent.

#### The Bootstrap Algorithm

The bootstrap provides a practical approach to estimate the standard error for quantile estimators. The algorithm proceeds as follows:

1. **Resample:** From the observed dataset of size $n$, draw $B$ bootstrap samples, each of size $n$, with replacement. Each bootstrap sample represents a possible realization of the population under the observed data.

2. **Recompute the estimator:** For each bootstrap sample, compute the quantile estimator $\hat{Q}_n(p, a)^{(b)}$ for $b = 1, 2, \ldots, B$.

3. **Compute the standard error:** The bootstrap standard error is given by:

$$\mathrm{SE}(\hat{Q}_n(p,a)) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{Q}_n(p,a)^{(b)} - \bar{Q}_n(p,a)\right)^2}$$

where $\bar{Q}_n(p,a) = \frac{1}{B}\sum_{b=1}^{B}\hat{Q}_n(p,a)^{(b)}$ is the mean of the bootstrap estimates.

This bootstrap-based standard error quantifies the variability of the quantile estimator and can be used for further inference, such as hypothesis testing or constructing confidence intervals as in our case. Using the computed standard error, a $(1-\alpha)$-confidence interval for the quantile estimator can be constructed as:

$$\hat{Q}_n(p,a) \pm z_{1-\alpha/2} \cdot \mathrm{SE}(\hat{Q}_n(p,a))$$

where $z_p$ is the critical value from the standard normal distribution corresponding to the desired confidence level.

The bootstrap method allows for a non-parametric estimation of the uncertainty surrounding quantile estimates, making it a powerful tool when the theoretical distribution of the estimator is not easily determined or does not follow traditional parametric assumptions.

# Chapter 5

# Conclusion

In this work, we build upon the framework introduced by Westling, T. et al. (2023) to develop an estimator for a key quantity in causal inference: the Quantile Treatment Effect (QTE). Our approach constructs an estimator for the marginal quantiles derived from survival curves. A key contribution of this work is the extension of asymptotic properties from the survival function to the quantile function, and subsequently to the QTE. By rigorously establishing these properties, we provide a more robust theoretical foundation for quantile-based causal inference, enhancing the applicability and interpretability of QTE estimators in practice.

One key limitation of our approach is that the survival function is estimated based on a time threshold $\tau$ due to the presence of right-censored data. As a result, certain quantiles—particularly larger ones—may not be estimated accurately. However, for smaller quantiles, our method remains reliable. Addressing this limitation in future research could improve the robustness of QTE estimation.

For future work, we suggest expanding the simulation studies to further evaluate the performance of the estimator, particularly in the context of quantile extensions. This would help refine the methodology and provide deeper insights into its practical applications.

Ultimately, our findings contribute to the growing body of research in causal inference by providing a theoretically sound and practically viable method for quantile estimation. As the demand for more nuanced causal analysis continues to rise, we hope that our work will serve as a foundation for future advancements in this field.

# Appendices

# Appendix A

# Derivation of the Coarsened Data Distribution

Marginalization allows us to transition from the ideal data distribution, $P_{\mathcal{O}_F}$, which includes unobserved components such as potential outcome and censoring times, to the observed data distribution, $P_0$, which reflects the quantities directly available for analysis.

## Ideal Data Distribution

The ideal data distribution is expressed as:

$$P_{\mathcal{O}_F}(\mathbf{L}, A, T(0), T(1), C(0), C(1)) = P_L(\mathbf{L}) \cdot P(A \mid \mathbf{L}) \tag{A.1}$$
$$\cdot P(T(0), T(1), C(0), C(1) \mid \mathbf{L}) \tag{A.2}$$

where:

- $P_L(\mathbf{L})$ is the marginal distribution of the covariates $\mathbf{L}$.
- $P(A \mid \mathbf{L})$ represents the treatment assignment mechanism.
- $P(T(0), T(1), C(0), C(1) \mid \mathbf{L})$ describes the joint distribution of the potential event and censoring times, conditional on $\mathbf{L}$.

## Observed Data Distribution via Marginalization

The marginal distribution of the observed data unit $O = (\mathbf{L}, A, Y, \Delta)$ is obtained by integrating out the unobserved components $T(0)$, $T(1)$, $C(0)$,

and $C(1)$ from the ideal data distribution:

$$P_0(\mathbf{L}, A, Y, \Delta) = \int P_{\mathcal{O}_F}(\mathbf{L}, A, T(0), T(1), C(0), C(1)) \, d(T(0), T(1), C(0), C(1)) \tag{A.3}$$

However, this integration must respect the coarsening rules that map the unobserved variables to the observed quantities:

$$Y = \min(T(A), C(A)), \quad \Delta = \mathbb{I}(T(A) \leq C(A)),$$

where $T(A)$ and $C(A)$ are the event and censoring times corresponding to the treatment assignment $A$, and $\mathbb{I}(\cdot)$ is the indicator function.

## Enforcing Coarsening Rules via Dirac Delta Functions

To enforce these constraints during marginalization, we introduce Dirac delta functions. The observed data distribution is then given by:

$$\begin{aligned} P_0(\mathbf{L}, A, Y, \Delta) = \int P_{\mathcal{O}_F}(\mathbf{L}, A, T(0), T(1), C(0), C(1)) \\ \cdot \delta(Y - \min(T(A), C(A))) \\ \cdot \delta(\Delta - \mathbb{I}(T(A) \leq C(A))) \\ d(T(0), T(1), C(0), C(1)). \end{aligned} \tag{A.4}$$

The Dirac delta functions $\delta(\cdot)$ enforce the constraints by ensuring that only values of $T(0), T(1), C(0), C(1)$ consistent with the observed values $Y$ and $\Delta$ contribute to the integral.

## Final Expression for the Observed Data Distribution

After marginalization, the observed data distribution is expressed as:

$$P_0(\mathbf{L}, A, Y, \Delta) = P_L(\mathbf{L}) \cdot P(A \mid \mathbf{L}) \cdot P(Y, \Delta \mid \mathbf{L}, A). \tag{A.5}$$

Here, the term $P(Y, \Delta \mid \mathbf{L}, A)$ encapsulates the effect of the coarsening mechanism, effectively summarizing how the event and censoring times contribute to the observed data.

# Appendix B

# Proof of Theorems of Chapter 3

Before proceeding with the proofs, we clarify our notation to avoid possible confusion. Below, we use $a_0$ to denote the fixed exposure level of interest, while reserving $a$ to represent a possible realization of the exposure random variable $A$.

*Proof of Proposition 2.3.1.* Condition $(A1)$ and $(A4)$ imply that

$$\begin{aligned} P_{\mathcal{O}_F}(T(a_0) > t \mid \mathbf{L} = \mathbf{l}) &= P_{\mathcal{O}_F}(T(a_0) > t \mid A = a_0, \mathbf{L} = \mathbf{l}) \\ &= P_{\mathcal{O}_F}(T > t \mid A = a_0, \mathbf{L} = \mathbf{l}) \end{aligned}$$

for all $t \in (0, \tau]$ and $P_0$-almost every $\mathbf{l}$, since $\mathbb{I}(T(a_0) > t)$ is a measurable function of $T(a_0)\mathbb{I}(T(a_0) \leq \tau)$ for $t \leq \tau$.
Therefore, using the tower property,

$$\theta_{\mathcal{O}_F}(t, a_0) = P_{\mathcal{O}_F}(T(a_0) > t) = \mathbb{E}_0[P_{\mathcal{O}_F}(T > t \mid A = a_0, \mathbf{L})].$$

Define

$$S_{\mathcal{O}_F}(t \mid a_0, \mathbf{l}) := P_{\mathcal{O}_F}(T > t \mid A = a_0, \mathbf{L} = \mathbf{l}).$$

Since $T$ is a positive random variable, by the following theorem of (Gill and Johansen (1990))

**Theorem B.0.1.** *Let $\Lambda$ be a nonnegative measure on $(0, \tau]$ that is finite on $(0, s]$ and satisfies $\Lambda(\{s\}) < 1$ for all $s < \tau$, and suppose that $\Lambda$ satisfies one of the following conditions:*

*(a)* $\Lambda((0, \tau)) < \infty$ *and* $\Lambda(\{\tau\}) = 1$,

*(b)* $\Lambda((0, \tau)) = \infty$ *and* $\Lambda(\{\tau\}) = 0$.

*Then, defining*

$$S(t) = \prod_{(0,t]}^{st} (1 - d\Lambda),$$

*where $T$ is a random variable with upper support endpoint $\tau$, $S$ is the survival function of $T$. Conversely, if $T$ is a positive random variable with survival function $S$ and upper support endpoint $\tau$ satisfying either of the following conditions:*

*(a') $S(\tau-) > 0$,*

*(b') $S(\tau-) = 0$,*

*then $\Lambda$, defined by*

$$\Lambda((0,t]) = -\int_0^t \frac{S(du)}{S(u-)} = -\int_0^t d(S-1)$$

*has the properties described above, where in the second expression $S$ is interpreted as the multiplicative interval function*

$$S(s,t) = \frac{S(t)}{S(s)}.$$

we can write:

$$S_{\mathcal{O}_F}(t \mid a_0, \mathbf{l}) = \prod_{(0,t]}^{\text{st}} \{1 - \Lambda_{\mathcal{O}_F}(du \mid a_0, \mathbf{l})\},$$

where

$$\Lambda_{\mathcal{O}_F}(t \mid a_0, \mathbf{l}) = -\int_0^t \frac{S_{\mathcal{O}_F}(du \mid a_0, \mathbf{l})}{S_{\mathcal{O}_F}(u- \mid a_0, \mathbf{l})}.$$

By the definition of $Y$, $T$, and $C$, we have

$$\begin{aligned} R_0(t \mid a_0, \mathbf{l}) &= P_0(Y \geq t \mid A = a_0, \mathbf{L} = \mathbf{l}) \\ &= P_{\mathcal{O}_F}(T \geq t, C \geq t \mid A = a_0, \mathbf{L} = \mathbf{l}) \\ &= P_{\mathcal{O}_F}(T(a_0) \geq t, C(a_0) \geq t \mid A = a_0, \mathbf{L} = \mathbf{l}). \end{aligned}$$

for all $t$. Using $(A3)$, we obtain

$$R_0(t \mid a_0, \mathbf{l}) = S_{\mathcal{O}_F}(t- \mid a_0, \mathbf{l}) \cdot G_{\mathcal{O}_F}(t- \mid a_0, \mathbf{l}),$$

for each $t \in (0, \tau]$, where

$$G_{\mathcal{O}_F}(t \mid a_0, \mathbf{l}) := P_{\mathcal{O}_F}(C(a_0) > t \mid a_0, \mathbf{l}).$$

Similarly,

$$\begin{aligned} F_{0,1}(t \mid a_0, \mathbf{l}) &= P_0(Y \leq t, \Delta = 1 \mid A = a_0, \mathbf{L} = \mathbf{l}) \\ &= P_{\mathcal{O}_F}(\min\{T, C\} \leq t, T \leq C \mid A = a_0, \mathbf{L} = \mathbf{l}) \\ &= P_{\mathcal{O}_F}(T \leq t, T \leq C \mid A = a_0, \mathbf{L} = \mathbf{l}) \\ &= \int_{u \in (0,t]} \int_{v \geq u} P_{\mathcal{O}_F}(du, dv \mid A = a_0, \mathbf{L} = \mathbf{l}). \end{aligned}$$

By $(A3)$, we have

$$\begin{aligned} P_{\mathcal{O}_F}(u, v \mid A = a_0, \mathbf{L} = \mathbf{l}) &= P_{\mathcal{O}_F}(T \leq u, C \leq v \mid A = a_0, \mathbf{L} = \mathbf{l}) \\ &= P_{\mathcal{O}_F}(T(a_0) \leq u, C(a_0) \leq v \mid a_0, \mathbf{l}) \\ &= [1 - S_{\mathcal{O}_F}(u \mid a_0, \mathbf{l})] \cdot [1 - G_{\mathcal{O}_F}(v \mid a_0, \mathbf{l})], \end{aligned}$$

for each $u, v \in (0, \tau]$. It follows that

$$\begin{aligned} F_{0,1}(t \mid a_0, \mathbf{l}) &= \int_{u \in (0,t]} \int_{v \geq u} [1 - S_{\mathcal{O}_F}(u \mid a_0, \mathbf{l})] \cdot [1 - G_{\mathcal{O}_F}(v \mid a_0, \mathbf{l})] \, dv \, du \\ &= \int_{u \in (0,t]} [1 - S_{\mathcal{O}_F}(u \mid a_0, \mathbf{l})] \cdot \left( \int_{v \geq u} [1 - G_{\mathcal{O}_F}(v \mid a_0, \mathbf{l})] \, dv \right) du \\ &= \int_{u \in (0,t]} [1 - S_{\mathcal{O}_F}(u \mid a_0, \mathbf{l})] \cdot G_{\mathcal{O}_F}(u \mid a_0, \mathbf{l}) \, du \\ &= - \int_{u \in (0,t]} G_{\mathcal{O}_F}(u- \mid a_0, \mathbf{l}) \cdot S_{\mathcal{O}_F}(du \mid a_0, \mathbf{l}) \, du \end{aligned}$$

for each $t \in (0, \tau]$ Thus,

$$F_{0,1}(dt \mid a_0, \mathbf{l}) = G_{\mathcal{O}_F}(t- \mid a_0, \mathbf{l}) \cdot S_{\mathcal{O}_F}(dt \mid a_0, \mathbf{l}).$$

for each $t \in (0, \tau]$. Using $(A2)$ and $(A5)$, since $G_{\mathcal{O}_F}(t- \mid a_0, \mathbf{l}) > 0$ for $P_0$-almost every $\mathbf{l}$ and $t \in [0, \tau]$, we conclude that

$$\begin{aligned} \Lambda_{\mathcal{O}_F}(t \mid a_0, \mathbf{l}) &= - \int_{(0,t]} \frac{S_{\mathcal{O}_F}(du \mid a_0, \mathbf{l})}{S_{\mathcal{O}_F}(u- \mid a_0, \mathbf{l})} \\ &= - \int_{(0,t]} \frac{G_{\mathcal{O}_F}(u- \mid a_0, \mathbf{l}) \cdot S_{\mathcal{O}_F}(du \mid a_0, \mathbf{l})}{G_{\mathcal{O}_F}(u- \mid a_0, \mathbf{l}) \cdot S_{\mathcal{O}_F}(u- \mid a_0, \mathbf{l})} \\ &= \int_{(0,t]} \frac{F_{0,1}(du \mid a_0, \mathbf{l})}{R_0(u \mid a_0, \mathbf{l})}. \end{aligned}$$

for each $t \in (0, \tau]$, completing the proof. □

*Proof of Theorem 3.1.1.* Let $\{P_\epsilon : |\epsilon| \leq \delta\}$ be a suitably smooth and bounded Hellinger differentiable path with $P_{\epsilon=0} = P_0$ and score function $\dot{l}_0$ at $\epsilon = 0$.

For a distribution $P$ of $(\mathbf{L}, A, Y, \Delta)$, we let $Q$ be the marginal distribution of $\mathbf{L}$ as implied by $P$. We then have, under appropriate boundedness conditions, that

$$\begin{aligned}\frac{\partial}{\partial\epsilon}\theta_\epsilon(t,a_0)\Big|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\int S_\epsilon(t \mid a_0,\mathbf{l})dQ_\epsilon(\mathbf{l})\Big|_{\epsilon=0} \\ &= \int \frac{\partial}{\partial\epsilon}S_\epsilon(t \mid a_0,\mathbf{l})\Big|_{\epsilon=0} dQ_0(\mathbf{l}) + \int S_0(t \mid a_0,\mathbf{l})\dot{l}_0(\mathbf{l})dQ_0(\mathbf{l})\end{aligned}$$

The second term contributes $S_0(t \mid a_0, \mathbf{l})$ to the efficient influence function. By definition, the integrand in the first term is

$$\frac{\partial}{\partial\epsilon}\prod_{(0,t]}^{\text{st}}\{1-\Lambda_\epsilon(du \mid a_0,\mathbf{l})\}\Big|_{\epsilon=0}$$

By the following Theorem (Gill and Johansen (1990)),

**Theorem B.0.2** (Compact differentiability of the product-integral with respect to the supremum norm))**.** *Consider the product-integral as a mapping $\mathcal{P}$ from the set of additive interval functions on $(0,\tau]$ with variation bounded by the constant $c$ to the space of interval functions on $(0,\tau]$, both domain and range endowed with the supremum norm. Let $\alpha$ be given and define*

$$\mu = \mathcal{P}(\alpha) = \prod_{(0,t]}^{st}(1-d\alpha)$$

*Then $\mathcal{P}$ is compactly differentiable at $\alpha$ with derivative $d\mathcal{P}(\alpha)$ given by*

$$(d\mathcal{P}(\alpha)\cdot h)(s,t) = \int_{(s,t]}\mu(s,u-)h(du)\mu(u,t),$$

*where the integral with respect to $h$ is defined by the integration-by-parts formula.*

the product integral map

$$H \mapsto S_H(t) := \prod_{(0,t]}^{\text{st}}\{1+H(du)\}$$

is Hadamard differentiable relative to the supremum norm with derivative

$$\alpha \to S_H(t)\int_0^t \frac{S_H(u-)}{S_H(u)}\alpha(du).$$

at $H$. Therefore, by the chain rule, we obtain

$$\frac{\partial}{\partial\epsilon}\prod_{(0,t]}^{\text{st}}\{1-\Lambda_\epsilon(du\mid a_0,\mathbf{l})\}\Big|_{\epsilon=0} = -S_0(t\mid a_0,\mathbf{l})\int_0^t \frac{S_0(u-\mid a_0,\mathbf{l})}{S_0(u\mid a_0,\mathbf{l})}\frac{\partial}{\partial\epsilon}\Lambda_\epsilon(du\mid a_0,\mathbf{l})\Big|_{\epsilon=0}$$

Now, since we can write

$$\begin{aligned}\frac{\partial}{\partial\epsilon}\Lambda_\epsilon(t\mid a_0,\mathbf{l})\Big|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\int_0^t R_\epsilon(u\mid a_0,\mathbf{l})^{-1}F_{\epsilon,1}(du\mid a_0,\mathbf{l})\Big|_{\epsilon=0}\\ &= \int_0^t R_\epsilon(u\mid a_0,\mathbf{l})^{-1}\frac{\partial}{\partial\epsilon}F_{\epsilon,1}(du\mid a_0,\mathbf{l})\Big|_{\epsilon=0}\\ &\quad -\int_0^t \frac{\partial}{\partial\epsilon}R_\epsilon(u\mid a_0,\mathbf{l})\Big|_{\epsilon=0} R_0(u\mid a_0,\mathbf{l})^{-2}F_0(du\mid a_0,\mathbf{l})\end{aligned}$$

we have

$$\frac{\partial}{\partial\epsilon}\Lambda_\epsilon(du\mid a_0,\mathbf{l})\Big|_{\epsilon=0} = \frac{\frac{\partial}{\partial\epsilon}F_{\epsilon,1}(du\mid a_0,\mathbf{l})\big|_{\epsilon=0}}{R_0(u\mid a_0,\mathbf{l})} - \frac{\frac{\partial}{\partial\epsilon}R_\epsilon(u\mid a_0,\mathbf{l})\big|_{\epsilon=0}F_0(du\mid a_0,\mathbf{l})}{R_0(u\mid a_0,\mathbf{l})^2}$$

In addition

$$\begin{aligned}\frac{\partial}{\partial\epsilon}F_{\epsilon,1}(u\mid a_0,\mathbf{l})\Big|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}P_\epsilon(Y\le u,\Delta=1\mid A=a_0,\mathbf{L}=\mathbf{l})\Big|_{\epsilon=0}\\ &= \frac{\partial}{\partial\epsilon}\int\int \mathbb{I}(y\le u,\delta=1)P_\epsilon(dy,d\delta\mid a_0,\mathbf{l})\Big|_{\epsilon=0}\\ &= \int\int \mathbb{I}(y\le u,\delta=1)\dot{l}_0(y,\delta\mid a_0,\mathbf{l})P_0(dy,d\delta\mid a_0,\mathbf{l})\Big|_{\epsilon=0}\end{aligned}$$

so that

$$\frac{\partial}{\partial\epsilon}F_{\epsilon,1}(u\mid a_0,\mathbf{l})\Big|_{\epsilon=0} = \int_\delta \mathbb{I}(\delta=1)\dot{l}_0(u,\delta\mid a_0,\mathbf{l})P_0(du,d\delta\mid a_0,\mathbf{l})$$

In a similar manner, we find

$$\frac{\partial}{\partial\epsilon}R_\epsilon(u\mid a_0,\mathbf{l})\Big|_{\epsilon=0} = \int\int \mathbb{I}(y\ge u)\dot{l}_0(y,\delta\mid a_0,\mathbf{l})P_0(dy,d\delta\mid a_0,\mathbf{l})$$

Therefore,

$$\frac{\partial}{\partial\epsilon}\int\int\prod_{(0,t]}^{\text{st}}\{1-\Lambda_\epsilon(du\mid a_0,\mathbf{l})\}\,dQ_0(\mathbf{l})\Big|_{\epsilon=0} =$$
$$\int\int\int -\mathbb{I}(y\le t,\delta=1)\frac{S_0(t\mid a_0,\mathbf{l})S_0(y-\mid a_0,\mathbf{l})}{S_0(y\mid a_0,\mathbf{l})R_0(y\mid a_0,\mathbf{l})}\dot{l}_0(y,\delta\mid a_0,\mathbf{l})P_0(dy,d\delta\mid a_0,\mathbf{l})dQ_0(\mathbf{l})$$
$$+\int\int\int\int\mathbb{I}(u\le t,u\le y)\frac{S_0(t\mid a_0,\mathbf{l})S_0(u-\mid a_0,\mathbf{l})}{S_0(u\mid a_0,\mathbf{l})R_0(u\mid a_0,\mathbf{l})^2}\dot{l}_0(y,\delta\mid a_0,\mathbf{l})$$
$$\times P_0(dy,d\delta\mid a_0,\mathbf{l})F_0(du\mid a_0,\mathbf{l})dQ_0(\mathbf{l})$$
$$=\int\int\int -\mathbb{I}(y\le t,\delta=1)\frac{S_0(t\mid a_0,\mathbf{l})S_0(y-\mid a_0,\mathbf{l})}{S_0(y\mid a_0,\mathbf{l})R_0(y\mid a_0,\mathbf{l})}\times\dot{l}_0(y,\delta\mid a_0,\mathbf{l})P_0(dy,d\delta\mid a_0,\mathbf{l})dQ_0(\mathbf{l})$$
$$+\int\int\int S_0(t\mid a_0,\mathbf{l})\int_0^{\min(t,y)}\frac{S_0(u-\mid a_0,\mathbf{l})}{S_0(u\mid a_0,\mathbf{l})R_0(u\mid a_0,\mathbf{l})^2}$$
$$\times F_0(du\mid a_0,\mathbf{l})\dot{l}_0(y,\delta\mid a_0,\mathbf{l})P_0(dy,d\delta\mid a_0,\mathbf{l})dQ_0(\mathbf{l})$$

obtaining as result

$$\mathbb{E}_0\left[S_0(t\mid A,\mathbf{L})\frac{\mathbb{I}(A=a_0)}{\pi_0(a_0\mid\mathbf{L})}\left\{H_0(t\wedge Y,A,\mathbf{L})-\frac{\mathbb{I}(Y\le t,\Delta=1)S_0(Y-\mid A,\mathbf{L})}{S_0(Y\mid A,\mathbf{L})R_0(Y\mid A,\mathbf{L})}\right\}\dot{l}_0(Y,\Delta\mid a_0,\mathbf{L})\right]$$

where

$$H_0(u,a,\mathbf{l}):=\int_0^u\frac{S_0(u-\mid a,\mathbf{l})F_0(du\mid a,\mathbf{l})}{S_0(u\mid a,\mathbf{l})R_0(u\mid a,\mathbf{l})^2}$$

Now, we note that

$$\mathbb{E}_0\left[\frac{\mathbb{I}(Y\le t,\Delta=1)S_0(Y-\mid A,\mathbf{L})}{S_0(Y\mid A,\mathbf{L})R_0(Y\mid A,\mathbf{L})}\mid A=a,\mathbf{L}=\mathbf{l}\right]=\int_0^t\frac{S_0(y-\mid a,\mathbf{l})F_0(dy\mid a,\mathbf{l})}{S_0(y\mid a,\mathbf{l})R_0(y\mid a,\mathbf{l})}$$

and

$$\mathbb{E}_0[H_0(t\wedge Y,A,\mathbf{L})\mid A=a,\mathbf{L}=\mathbf{l}]=$$
$$=\int\int_0^t\mathbb{I}(u\le y)\frac{S_0(u-\mid a,\mathbf{l})F_0(du\mid a,\mathbf{l})}{S_0(u\mid a,\mathbf{l})R_0(u\mid a,\mathbf{l})^2}P_0(dy\mid a,\mathbf{l})$$
$$=\int_0^t P_0(Y\ge u\mid A=a,\mathbf{L}=\mathbf{l})\frac{S_0(u-\mid a,\mathbf{l})F_0(du\mid a,\mathbf{l})}{S_0(u\mid a,\mathbf{l})R_0(u\mid a,\mathbf{l})^2}$$
$$=\int_0^t\frac{S_0(u-\mid a,\mathbf{l})F_0(du\mid a,\mathbf{l})}{S_0(u\mid a,\mathbf{l})R_0(u\mid a,\mathbf{l})}.$$

since $P_0(Y\ge u\mid A=a,\mathbf{L}=\mathbf{l})=R_0(u\mid a,\mathbf{l})$ by definition. Therefore,

$$\mathbb{E}_0\left[H_0(t\wedge Y,A,\mathbf{L})-\frac{\mathbb{I}(Y\le t,\Delta=1)S_0(Y-\mid A,\mathbf{L})}{S_0(Y\mid A,\mathbf{L})R_0(Y\mid A,\mathbf{L})}\mid A,\mathbf{L}\right]=0$$

$P_0-$almost surely. This implies by properties of score functions and the tower property that:

$$\frac{\partial}{\partial \epsilon} \int \prod_{(0,t]}^{\text{st}} \{1 - \Lambda_\epsilon(du \mid a_0, \mathbf{l})\} \, dQ_0(\mathbf{l}) \Big|_{\epsilon=0}$$
$$= \mathbb{E}_0 \left[ S_0(t \mid A, \mathbf{L}) \frac{\mathbb{I}(A = a_0)}{\pi_0(a_0 \mid \mathbf{L})} \left\{ H_0(t \wedge Y, A, \mathbf{L}) - \frac{\mathbb{I}(Y \leq t, \Delta = 1) S_0(Y- \mid A, \mathbf{L})}{S_0(Y \mid A, \mathbf{L}) R_0(Y \mid A, \mathbf{L})} \right\} \dot{l}_0(Y, \Delta, A,$$

Combining these results, we find that the uncentered influence function is

$$o \to S_0(t \mid a_0, \mathbf{l}) \left[ 1 - \frac{\mathbb{I}(a = a_0)}{\pi_0(a_0 \mid \mathbf{l})} \left\{ \frac{\mathbb{I}(y \leq t, \delta = 1) S_0(y- \mid a_0, \mathbf{l})}{S_0(y \mid a_0, \mathbf{l}) R_0(y \mid a_0, \mathbf{l})} \right. \right.$$
$$\left. \left. + \int_0^{t \wedge y} \frac{S_0(u- \mid a_0, \mathbf{l}) F_0(du \mid a_0, \mathbf{l})}{S_0(u \mid a_0, \mathbf{l}) R_0(u \mid a_0, \mathbf{l})^2} \right\} \right]$$

By our calculation above, the mean of the term in curly brackets is zero, and so, the mean of the entire expression is $\mathbb{E}_0[S_0(t \mid a_0, \mathbf{L})] = \theta_0(t, a_0)$. We note that

$$\frac{F_0(du \mid a_0, \mathbf{l})}{R_0(u \mid a_0, \mathbf{l})} = \Lambda_0(du \mid a_0, \mathbf{l})$$

and that $R_0(u \mid a_0, \mathbf{l}) = S_0(u- \mid a_0, \mathbf{l}) G_0(u \mid a_0, \mathbf{l})$, so that, as claimed, the above is equal to

$$S_0(t \mid a_0, \mathbf{l}) \left[ 1 - \frac{\mathbb{I}(a = a_0)}{\pi_0(a_0 \mid \mathbf{L})} \left\{ \frac{\mathbb{I}(y \leq t, \delta = 1) S_0(y- \mid a_0, \mathbf{l})}{S_0(y \mid a_0, \mathbf{l}) S_0(y- \mid a_0, \mathbf{l}) G_0(y \mid a_0, \mathbf{l})} \right. \right.$$
$$\left. \left. + \int_0^{t \wedge y} \frac{S_0(u- \mid a_0, \mathbf{l}) \Lambda_0(du \mid a_0, \mathbf{l})}{S_0(u \mid a_0, \mathbf{l}) S_0(u- \mid a_0, \mathbf{l}) G_0(u \mid a_0, \mathbf{l})} \right\} \right]$$
$$= S_0(t \mid a_0, \mathbf{l}) \left[ 1 - \frac{\mathbb{I}(a = a_0)}{\pi_0(a_0 \mid \mathbf{l})} \left\{ \frac{\mathbb{I}(y \leq t, \delta = 1)}{S_0(y \mid a_0, \mathbf{l}) G_0(y \mid a_0, \mathbf{l})} \right. \right.$$
$$\left. \left. + \int_0^{t \wedge y} \frac{\Lambda_0(du \mid a_0, \mathbf{l})}{S_0(u \mid a_0, \mathbf{l}) G_0(u \mid a_0, \mathbf{l})} \right\} \right]$$

□

# Appendix C

# Proof of Theorems of Chapter 4

*Proof of Lemma 4.2.1.* Let $S_n(t)$ and $S(t)$ be monotone, non-increasing functions, and let the quantile function $S_n^{-1}(p)$ be the smallest $t$ such that $S_n(t) \leq 1-p$. Similarly, $S^{-1}(p)$ is the smallest $t$ such that $S(t) \leq 1-p$. The weak convergence $S_n \to S$ implies that for large $n$, $S_n(t)$ is close to $S(t)$, and the quantile functions behave in a similar manner.
To establish the weak convergence of $S_n^{-1}(p)$, we use a bounding argument:

- If $t > S^{-1}(p)$, then $S(t) > 1-p$, and since $S_n \to S$, we have $S_n(t) > 1-p$ for sufficiently large $n$, so:

$$S_n^{-1}(p) \leq S^{-1}(p) + \delta \quad \text{with } \delta > 0$$

- If $t < S^{-1}(p)$, then $S(t) \leq 1-p$, and for large $n$, $S_n(t) \leq 1-p$, so:

$$S_n^{-1}(p) \geq S^{-1}(p) - \delta \quad \text{with } \delta > 0$$

Combining these bounds, we obtain:

$$|S_n^{-1}(p) - S^{-1}(p)| \to 0$$

Thus, the quantile function $S_n^{-1}(p)$ weakly converges to $S^{-1}(p)$, as desired. □

*Proof of Leamma 4.2.4.* Let $h_t \to h$ uniformly on $[a, b]$ for a function $h$ that is continuous at $\xi_p$. Write $\xi_{pt}$ for $\psi(F + th_t)$. By the definition of $\psi$, for every $\varepsilon_t > 0$,

$$(F + th_t)(\xi_{pt} - \varepsilon_t) \leq p \leq (F + th_t)(\xi_{pt})$$

Choose $\varepsilon_t$ positive and such that $\varepsilon_t = o(t)$. Because the sequence $h_t$ converges uniformly to a bounded function, it is uniformly bounded. Conclude that

$$F(\xi_{pt} - \varepsilon_t) + O(t) \leq p \leq F(\xi_{pt}) + O(t).$$

By assumption, the function $F$ is monotone and bounded away from $p$ outside any interval $(\xi_p - \varepsilon, \xi_p + \varepsilon)$ around $\xi_p$. To satisfy the preceding inequalities, the numbers $\xi_{pt}$ must be to the right of $\xi_p - \varepsilon$ eventually, and the numbers $\xi_{pt} - \varepsilon_t$ must be to the left of $\xi_p + \varepsilon$ eventually. In other words, $\xi_{pt} \to \xi_p$. By the uniform convergence of $h_t$ and the continuity of the limit,

$$h_t(\xi_{pt} - \varepsilon_t) - h(\xi_p) \to 0 \quad \text{for every } \varepsilon_t \to 0.$$

Using this and Taylor's formula on the preceding display yields:

$$\begin{aligned} p + (\xi_{pt} - \xi_p)F'(\xi_p) - o(\xi_{pt} - \xi_p) + O(\varepsilon_t) \\ + th(\xi_p) - o(t) \leq p \\ \leq p + (\xi_{pt} - \xi_p)F'(\xi_p) + o(\xi_{pt} - \xi_p) + O(\varepsilon_t) \\ + th(\xi_p) + o(t) \end{aligned}$$

Conclude first that $\xi_{pt} - \xi_p = O(t)$. Next, use this to replace the $o(\xi_{pt} - \xi_p)$ terms in the display by $o(t)$ terms and conclude that:

$$\frac{(\xi_{pt} - \xi_p)}{t} \to -\left(\frac{h}{F'}\right)(\xi_p).$$

□

*Proof of lemma 4.2.5.* we analyze the two possible scenarios

- Because the function $F$ has a positive density, it is strictly increasing on an interval $[\xi_{p_1'}, \xi_{p_2'}]$ that strictly contains $[\xi_{p_1}, \xi_{p_2}]$. Then in $[p_1', p_2']$ the quantile function $F^{-1}$ is the ordinary inverse of $F$ and is (uniformly) continuous and strictly increasing.
  Let $h_t \to h$ uniformly on $[\xi_{p_1'}, \xi_{p_2'}]$ for a continuous function $h$. By the proof of lemma 4.2.3, $\xi_{p_i t} \to \xi_{p_i}$ and hence every $\xi_{pt}$ for $p_1 \leq p \leq p_2$ is contained in $[\xi_{p_1'}, \xi_{p_2'}]$ eventually. The remainder of the proof is the same as the proof of the preceding lemma.

- Let $h_t \to h$ uniformly in $D[a, b]$, where $h$ is continuous and $F + th_t$ is contained in $\mathcal{D}_2$ for all $t$. Abbreviate $F^{-1}(p)$ and $(F + th_t)^{-1}(p)$ to $\xi_p$ and $\xi_{pt}$, respectively. Because $F$ and $F + th_t$ are concentrated on $(a, b]$

by assumption, we have $a < \xi_{pt}, \xi_p \leq b$ for all $0 < p < 1$. Thus the numbers $\varepsilon_{pt} = \min(t^2, (\xi_{pt} - a))$ are positive, whence, by definition,

$$(F + th_t)(\xi_{pt} - \varepsilon_{pt}) \leq p \leq (F + th_t)(\xi_{pt})$$

By the smoothness of $F$ we have $F(\xi_p) = p$ and $F(\xi_{pt} - \varepsilon_{pt}) = F(\xi_{pt}) + O(\varepsilon_{pt})$, uniformly in $0 < p < 1$. It follows that

$$-th(\xi_{pt}) + o(t) \leq F(\xi_{pt}) - F(\xi_p) \leq -th(\xi_{pt} - \varepsilon_{pt}) + o(t)$$

The $o(t)$ terms are uniform in $0 < p < 1$. The far left side and the far right side are $O(t)$; the expression in the middle is bounded above and below by a constant times $|\xi_{pt} - \xi_p|$. Conclude that $|\xi_{pt} - \xi_p| = O(t)$, uniformly in $p$. Next, the lemma follows by the uniform differentiability of $F$.

□

# Appendix D

# Gateaux, Hadamard and Fréchet differentiability

There are several possible ways to define the differentiability of a map $\psi : \mathbb{D} \to \mathbb{E}$ between normed spaces. We discuss three main notions of differentiability: Gateaux, Hadamard, and Fréchet differentiability.

### Gateaux Differentiability

A map $\psi$ is said to be **Gateaux differentiable** at $\theta \in \mathbb{D}$ if, for every fixed direction $h \in \mathbb{D}$, there exists an element $\psi'_\theta(h) \in \mathbb{E}$ such that

$$\psi(\theta + th) - \psi(\theta) = t\psi'_\theta(h) + o(t),$$

which means more formally that

$$\left| \frac{\psi(\theta + th) - \psi(\theta)}{t} - \psi'_\theta(h) \right|_{\mathbb{E}} \to 0 \quad \text{as } t \to 0. \tag{D.1}$$

This notion of differentiability is also called *directional differentiability* because, for every possible direction $h$, the derivative $\psi'_\theta(h)$ measures the infinitesimal change in the function value along that direction. However, Gateaux differentiability is often too weak for statistical applications, particularly for the delta method, necessitating a stronger notion of differentiability. *Note: The main reason is that the Gateaux derivative may fail to be linear or continuous, which is crucial for the delta method to work. In particular, the delta method relies on approximating functionals of estimators using a first-order expansion, which requires a differentiability notion that ensures continuity and linearity of the derivative map, such as Hadamard or Frèchet differentiability.*

### Hadamard Differentiability

A map $\psi : \mathbb{D}_\psi \to \mathbb{E}$, defined on a subset $\mathbb{D}_\psi$ of a normed space $\mathbb{D}$ that contains $\theta$, is called **Hadamard differentiable** at $\theta$ if there exists a continuous linear map $\psi'_\theta : \mathbb{D} \to \mathbb{E}$ such that

$$\left| \frac{\psi(\theta + th_t) - \psi(\theta)}{t} - \psi'_\theta(h) \right|_{\mathbb{E}} \to 0 \quad \text{as } t \to 0, \text{ for every sequence } h_t \to h. \tag{D.2}$$

The key difference between Hadamard and Gateaux differentiability is that, in the former, the directions $h_t$ are allowed to vary with $t$, whereas in the latter, they must remain fixed.

*Note:* The above definition assumes that $\psi'_\theta : \mathbb{D} \to \mathbb{E}$ exists as a linear operator on the entire space $\mathbb{D}$. If $\psi'_\theta$ is only well-defined on a subset $\mathbb{D}_0 \subset \mathbb{D}$ and we restrict $h_t \to h$ to those limits $h \in \mathbb{D}_0$, then $\psi$ is said to be Hadamard differentiable *tangentially* to $\mathbb{D}_0$.

### Fréchet Differentiability

For completeness, we introduce a third, stronger form of differentiability. The map $\psi : \mathbb{D}_\psi \to \mathbb{E}$ is called **Fréchet differentiable** at $\theta$ if there exists a continuous linear map $\psi'_\theta : \mathbb{D} \to \mathbb{E}$ such that

$$\|\psi(\theta + h) - \psi(\theta) - \psi'_\theta(h)\|_{\mathbb{E}} = o(\|h\|), \quad \text{as } \|h\| \to 0.$$

Fréchet differentiability implies both Gateaux and Hadamard differentiability, but the converse is not necessarily true. In statistical applications, Fréchet differentiability may fail, while Hadamard differentiability often holds, making the latter the most relevant concept for practical purposes.

# Bibliography


Anderson, J. R., Bernstein, L., and Pike, M. C. (1982), *Approximate Confidence Intervals for Probabilities of Survival and Quantiles in Life-Table Analysis*, Biometrics, 38, 407–416.

Bai, X., Tsiatis, A. A., and O'Brien, S. M. (2013), *Doubly-RobustEstimators of Treatment-Specific Survival Distributions in Observational Studies with Stratified Sampling*, Biometrics, 69, 830–839.

Bernan, R. (1981) *Nonparamatric Regression with Randomly Censored Survival Data*, Technical report.

Bickel, P. J. (1982), *On Adaptive Estimation*, The Annals of Statistics, 10, 647–671.

Breiman, L. (1996), *Stacked Regressions*, Machine Learning, 24, 49-64.

Dabrowska, D. M. (1989), *Uniform Consistency of the Kernel Conditional Kaplain-Meier Estimate*, The Annals of Statistics, 17, 1157-1167.

Díaz,I.(2019a), *Machine Learning in the Estimation of Causal Effects: Targeted Minimum Loss-based Estimation and Double/Debiased Machine Learning*, Biostatistics, 21, 353–358.

Gill, R.D. and Johansen, S. (1990) *A Survey of Product-Integration with a View Toward Application in Survival Analysis*, The Annals of Statistics, 18, 1501-1555.

Gill, R. D. and Robins, J. M. (2001) *Causal Inference for Complex longitudinal data: The continuous case*, The annuals of Statistics, 29, 1785-1811.

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

Hines, O., Dukes, O., Karla Diaz-Ordaz & Stijn Vansteelandt (2022), *Demystifying Statistical Learning Based on Efficient Influence Functions*, The American Statistician,76:3, 292-304, DOI:10.1080/00031305.2021.2021984

Hubbard, A. E., van der Laan, M. J., and Robins, J. M. (2000), *Nonparametric Locally Efficient Estimation of the Treatment Specific Survival Distribution with Right Censored Data and Covariates in Observational Studies*, IMA Volumes in Mathematics and Its Applications, 116,135–178.

Neyman, J. (1923), *On the Application of Probability Theory to Agricultural Experiments. Essay on Principles*, Statistical Science, 5, 465 - 480.

Robins, J. (1986) *A New Approach to Causal Inference in mortality studies with a Sustained Exposure Period - Application to control of the Healty worker Survivor Effect*, Mathematical Modelling, 7, 1393-1512.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008), *Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals*, Probability and Statistics: Essays in Honor of David A. Freedman, eds. D. Nolan, and T. Speed, pp. 335–421, Beachwood, OH: Institute of Mathematical Statistics.

Rubin, D. B. (1974) *Estimating causal effects of treatments in randomized and nonrandomized studies*, Journal of Educational Psychology, 66(5):688-701.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007) *Super Learner*, Statistical Application in Genetics and Molecular Biology, 6, 1-23.

van der Vaart, A.W. (1998) *Asymptotic Statistics*Cambridge Series in Statistical and Probabilistic Mathematics

Westling, T., van der Laan, M. J., and Carone, M. (2020), *Correcting an Estimator of a Multivariate Monotone Function with Isotonic Regression*, Electronic Journal of Statistics, 14, 3032–3069.

Westling, T. et al. (2023) *Inference for Treatment-Specific Survival Curves Using Machine Learning*,Journal of the American Statistical Association, 119(546), pp. 1541–1553.

Zheng, W., and van der Laan, M. J. (2011), *Cross-Validated Targeted Minimum-Loss Based Estimation*,

Targeted Learning: Causal Inference for Observational and Experimental Data, eds. M. van der Laanand and S. Rose, pp. 459–473, New York: Springer-Verlag.