Nonparametric inference on monotone functions, with applications to observational studies

Theodore Westling

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Marco Carone, Chair

Jon Wellner

Peter Gilbert

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Nonparametric inference on monotone functions,
with applications to observational studies

Theodore Westling

Chair of the Supervisory Committee:
Assistant Professor Marco Carone
Department of Biostatistics

In this dissertation, we study general strategies for constructing nonparametric monotone function estimators in two broad statistical settings. In the first setting, a sensible initial estimator of the monotone function of interest is available, but may fail to be monotone. We study the correction of such an estimator obtained via projection onto the space of functions monotone over a finite grid in the domain. We demonstrate that this corrected estimator is always at least as good in supremum norm as the initial estimator, and provide conditions under which the two estimators are asymptotically equivalent. In the second setting, a sensible estimator of the primitive of the function of interest is available. In this setting, estimators considered in the literature have often been of so-called Grenander type, being representable as the left derivative of the greatest convex minorant of the primitive estimator. We provide general conditions for consistency and pointwise convergence in distribution of a class of generalized Grenander-type estimators. This broad class allows the minorization operation to be performed on a data-dependent transformation of the domain. We use our general results from the second setting to perform a detailed study of generalized Grenander-type estimation of a monotone covariate-adjusted regression curve, which describes the effect of a continuous exposure on an outcome while adjusting for potential confounders. In particular, we show how our results can be used to conduct doubly-robust inference for this parameter.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

Monotone function: A function $f : I \to \mathbb{R}$ satisfying either $f(x) \leq f(y)$ for all $x, y \in I$ such that $x \leq y$ or $f(x) \geq f(y)$ for all $x, y \in I$ such that $x \leq y$, where $I \subseteq \mathbb{R}$

NPMLE: Nonparametric maximum likelihood estimator

$\mathbb{R}$: The real numbers

$\mathbb{R}^d$: The d-dimensional Euclidean space

$\ell^\infty(\mathcal{X})$: The Banach space of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\sup_{x \in \mathcal{X}} |f(x)| < \infty$ endowed with supremum norm $\|f\|_{\mathcal{X}} := \sup_{x \in \mathcal{X}} |f(x)|$

$L_2(P)$: For P a measure on a measurable space $(\mathcal{X}, \Omega)$, the Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\int f^2 \, dP < \infty$, endowed with inner product $< f, g >:= \int fg \, dP$

$\text{GCM}_I(G)$: The Greatest Convex Minorant (GCM) of a function G over an interval $I \subseteq \mathbb{R}$, defined pointwise as $G(x) := \sup_{F \in \mathcal{F}_{G,I}} F(x)$, where $\mathcal{F}_{G,I}$ is the set of convex functions F on I such that $F \leq G$

$\text{LCM}_I(G)$: The Least Concave Minorant (LCM) of a function G over an interval $I \subseteq \mathbb{R}$, defined pointwise as $G(x) := \inf_{F \in \mathcal{F}_{G,I}} F(x)$, where $\mathcal{F}_{G,I}$ is the set of concave functions F on I such that $F \geq G$

G^-: The generalized inverse mapping of a monotone function $G : I \to \mathbb{R}$, defined pointwise as $G^-(x) := \inf\{x \in I : G(u) \geq x\}$

$\partial_- G$: The left derivative function of a left-differentiable function G

$\mathbb{P}_n$: The empirical probability measure corresponding to a sample of size n placing mass $1/n$ at each observation

$\mathbb{G}_n$: The empirical process corresponding to a sample of size n, which is defined as the signed measure $n^{1/2}[\mathbb{P}_n - P_0]$

$I_S(x)$: The indicator function that x is in the set S

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Marco Carone. I have been fortunate to work with a statistician with such clear and broad vision, yet also so careful and insightful; a statistician who can appreciate the mathematical beauty in statistics while remaining committed to doing scientifically impactful research. You managed to push me to set my sights high and to work outside my comfort zone, while still being supportive and understanding. I have no doubt that I will be benefitting from your training, preparation, wisdom, and guidance for years to come, and I aspire to one day be even half the advisor you have been. I am lucky to be a leaf on your academic tree.

I want to thank Peter Gilbert for being the best RA advisor and scientific collaborator a PhD student could ask for. I have learned an enormous amount from you about statistical methods in causal inference and vaccine trials. You have also demonstrated to me that it is possible for a statistician to be a leader in a global medical research effort. I want to thank you in particular for your patience as Marco and I worked out the general theory of Chapters 2 and 3; my hope is that we and others can continue applying this general theory to important problems in vaccine trials for many years.

I would also like to thank the other members of my PhD committee: Jon Wellner for several helpful discussions about monotone function estimation, for pointing out some crucial references, and for feedback on the first draft of my dissertation; Ali Shojaie for serving on my committee on short notice; and Emily Williams for being my GSR and for providing helpful feedback during my general and final exams.

There are many others at the University of Washington whose wisdom I benefited from over the past five years. I want to thank Tyler McCormick for invaluable mentorship, en-

thusiasm, and RA support in the early years of my PhD, and Elena Erosheva for advice and guidance as my academic advisor and supervisor while I was the CSSS Consultant. I would like to thank the Department of Statistics and CSSS staff, including Eileen Heimer, Ellen Reynolds, Kristine Chan, Mee-Ling Hon, Tracy Pham, Kris Shaw, Asa Sourdiffe, and Vickie Graybeal, for helping me with the administrative, financial, and technical aspects of getting a PhD. Special thanks to Eileen for friendly conversation and for taking care of Matilda when I was away. I would also like to thank the HIV Vaccine Trials Network and Fred Hutchinson Cancer Research Center for RA support while writing my dissertation.

Finally, I would like to thank all the people in my life whose love and support I relied upon each and every day. Mom, Dad, and Plato – you were always there for me, either on the phone or in person. Well, Plato wasn't much help on the phone, but that's not his fault. I have your love and upbringing to thank for everything in my life, and I cannot begin to express my gratitude for all that you have given me. I want to thank my life partner and future wife for your steadfast support and for always reminding me of what is most important in life, and I want to thank our fuzzy friends Baxter and Hope for endless entertainment and antics, and for always putting a smile on my face. Last but far from least, I would like to thank all of my friends in Seattle and elsewhere who brought me cheer and laughter – Amit Meir, Rebecca Ferrell, Sam Wang, Tyler Boomsauce, Jonah Ostroff, Kyle Loh, Alden and Raechel Timme, Alex Hofstetter, Jay Garlapati, Alex Tank, Jason Xu, Jessica Godwin, Hannah Director, Sam Frizell, Emin Topalovic, Zahan Malkani, Aaron Kalb, Amanda Hillsberg, Matt Junge, Ashley Hufnagle, Angelica Meinhofer, Felipe Barrientos, and Maria Cristina Ramos, to name a few.

DEDICATION

To Mom and Dad.

Chapter 1

INTRODUCTION

1.1 Definitions and basic facts

1.1.1 Empirical processes

For an arbitrary set T, we define $\ell^\infty(T)$ as the Banach space of bounded functions $f : T \to \mathbb{R}$ endowed with supremum norm $\|f\|_T := \sup_{t\in T} |f(t)|$. For a metric space $\mathbb{D}$, we define $C_b(\mathbb{D})$ as the space of bounded continuous maps from $\mathbb{D}$ to $\mathbb{R}$.

A *stochastic process* indexed by T is a collection $X := \{X(t) : t \in T\}$ of random variables $X(t)$ defined on the same measureable space $(\Omega, \mathcal{B})$. We say X has *bounded sample paths* if $\|X(\cdot, \omega)\|_T < \infty$ for every ω. We can consider a stochastic process with bounded sample paths X a map from Ω to $\ell^\infty(T)$. A sequence $\{X_i : i \in \mathbb{N}\}$ of stochastic processes on T with bounded sample paths *converges weakly* in $\ell^\infty(T)$ to a random element X of $\ell^\infty(T)$ if $E^*[h(X_n)] \to E[h(X)]$ as $n \to \infty$ for every $h \in C_b(\ell^\infty(T))$, and we write $X_n \rightsquigarrow X$. Here, E^* is the *outer expectation*, which permits processes X_n to not be Borel-measurable elements of $\ell^\infty(T)$. For more detail on outer expectation and weak convergence, we refer the interested reader to Chapter 1 of van der Vaart and Wellner (1996).

Let $\{O_i : i \in \mathbb{N}\}$ be a sequence of independent observations from a probability measure P_0 on a measurable space $(\mathcal{O}, \mathcal{B})$. For $n \in \mathbb{N}$, define $\mathbb{P}_n$ as the *empirical probability measure* that puts mass $1/n$ at each observation $O_1, \ldots, O_n$. For a function $f : \mathcal{O} \to \mathbb{R}$, define $Pf := \int f \, dP$ for any probability measure P, and $\mathbb{G}_n f$ as $\mathbb{G}_n := n^{1/2}(\mathbb{P}_n - P_0)f$. Given a class of functions $\mathcal{F}$ from $\mathcal{O} \to \mathbb{R}$, $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ forms a stochastic process indexed by $\mathcal{F}$, and is called the *empirical process on* $\mathcal{F}$. A random element X of a metric space $\mathbb{D}$ is *tight*

if for every $\varepsilon > 0$ there is a compact $K \subseteq \mathbb{D}$ such that $P(X \in K) \geq 1 - \varepsilon$. A class $\mathcal{F}$ is called P_0-*Donsker* if $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges weakly in $\ell^\infty(\mathcal{F})$ to a tight Borel-measurable limit process. A detailed study of conditions under which a class is P_0-Donsker, in addition to many other results about empirical processes, can be found in van der Vaart and Wellner (1996).

1.1.2 Statistical models and efficiency theory of pathwise differentiable functionals

A *statistical model* $\mathcal{M}$ is a collection of probability measures on a measurable space $(\mathcal{O}, \mathcal{B})$. We say a measure P is *dominated* by another measure μ if $\mu(S) = 0$ implies $P(S) = 0$, and we say $\mathcal{M}$ is dominated by μ if P is dominated by μ for every $P \in \mathcal{M}$. For a measure P on $(\mathcal{O}, \mathcal{B})$, we define $L_2(P)$ as the Hilbert space of square-integrable functions $g : \mathcal{O} \to \mathbb{R}$. We define $L_2^0(P)$ as the linear subspace of functions satisfying $Pg = 0$.

A one-dimensional path $\{P_t : 0 \leq t < \tau\}$ through a distribution P_0 in a model $\mathcal{M}$ dominated by μ is called *differentiable in quadratic mean*, also known as *Hellinger differentiable*, with *score function* $g \in L_2^0(P_0)$, if

$$\lim_{t \to 0} \int \left[\frac{p_t^{1/2} - p_0^{1/2}}{t} - \tfrac{1}{2} g p_0^{1/2} \right]^2 d\mu = 0,$$

where $p_t := \frac{dP_t}{d\mu}$ is the density of P_t with respect to μ. The collection of all such scores over one-dimensional paths in $\mathcal{M}$ through P_0 is called the *tangent set* of $\mathcal{M}$ at P_0, denoted $T_{\mathcal{M}}^\circ(P_0)$. The closure of the linear span of $T_{\mathcal{M}}^\circ(P_0)$ is called the *tangent space* of $\mathcal{M}$ at P_0, denoted $T_{\mathcal{M}}(P_0)$. A statistical model $\mathcal{M}$ is *finite dimensional* or *parametric* at P_0 if $T_{\mathcal{M}}(P_0) \simeq \mathbb{R}^d$ for some $d < \infty$.

A *functional* or *parameter* ψ on $\mathcal{M}$ is simply a map from $\mathcal{M}$ to some other space $\mathcal{R}$. A *Euclidean* parameter is a parameter with $\mathcal{R} = \mathbb{R}^k$ for some $1 \leq k < \infty$. A Euclidean parameter $\psi : \mathcal{M} \to \mathbb{R}^k$ is called *pathwise differentiable* at P_0 relative to $\mathcal{M}$ if there exists a continuous, linear map $\psi_0' : T_{\mathcal{M}}(P_0) \to \mathbb{R}^k$ such that for all $g \in T_{\mathcal{M}}(P_0)$ and Hellinger

differentiable paths $\{P_t : 0 \leq t < \tau\}$ with score g,

$$\lim_{t \to 0} \left\| \frac{\psi(P_t) - \psi(P_0)}{t} - \psi_0'(g) \right\|_2 = 0,$$

where $\| \cdot \|_2$ denotes the Euclidean norm in $\mathbb{R}^k$. Any function $D_0 := D_{P_0} \in L_2^0(P_0)$, which may depend on summaries of P_0, such that $\psi_0'(g) = P_0[D_0 g]$ for all $g \in T_{\mathcal{M}}(P_0)$ is called an *influence function* for estimating $\psi_0 := \psi_{P_0}$. If ψ is pathwise differentiable at P_0 relative to $\mathcal{M}$, then by the Reisz representation theorem, there exists a unique influence function contained in $T_{\mathcal{M}}(P_0)$. This function D_0^* is called the *efficient influence function* for estimating ψ under P_0 relative to $\mathcal{M}$, and can be found by projecting any influence function D_0 on to $T_{\mathcal{M}}(P_0)$.

If $\mathcal{M}$ consists of all probability measures on $(\mathcal{O}, \mathcal{B})$, then $T_{\mathcal{M}}(P_0) = L_2^0(P_0)$. This can be seen by defining $p_t := c(t)^{-1} \text{expit}(2tg)\, p_0$ for any $g \in L_2^0(P_0)$, where $c(t) := \int \text{expit}(2tg)\, p_0$ and $\text{expit}(x) := (1 + e^{-x})^{-1}$. Then $\{P_t : 0 \leq t < \tau\}$ is a one-dimensional path through P_0 with score g. Certain types of restrictions to a model do not change the tangent space $T_{\mathcal{M}}(P_0)$, such as requiring that $|Ph| < \infty$ for every $P \in \mathcal{M}$ and some fixed function h. Therefore, we define a *nonparametric model* as any model $\mathcal{M}$ such that $T_{\mathcal{M}}(P_0) = L_2^0(P_0)$. In a nonparametric model there is only one influence function, which is necessarily the efficient influence function. This is because there is only one influence function contained in $T_{\mathcal{M}}(P_0)$, which for a nonparametric model is the entirety of $L_2^0(P_0)$. We say that $\mathcal{M}$ is a *semiparametric model* if $T_{\mathcal{M}}(P_0)$ is infinite-dimensional, but a strict subset of $L_2^0(P_0)$. For more on tangent spaces, efficiency theory, and many worked examples, see Bickel et al. (1998).

1.1.3 Estimators and asymptotic linearity

An *estimator* ψ_n of the evaluation of a parameter $\psi : \mathcal{M} \to \mathcal{R}$ at P_0 based on the sample $O_1, \ldots, O_n$ is a measureable map from $\mathcal{O}^n$ to $\mathcal{R}$. An estimator ψ_n is said to be *weakly consistent* for ψ_0 if $(\mathcal{R}, \| \cdot \|_{\mathcal{R}})$ is a metric space and $\|\psi_n - \psi_0\|_{\mathcal{R}} \xrightarrow{\text{P}} 0$. An estimator ψ_n of a Euclidean parameter $\psi_0 \in \mathbb{R}^k$ is *asymptotically linear* if there exists a function $D_0 : \mathcal{O} \to \mathbb{R}^k$ such that $P_0 D_0 = 0$, $P_0 D_0^2 < \infty$, and $\psi_n = \psi_0 + \mathbb{P}_n D_0 + R_n$, where $R_n = \text{o}_{\text{P}}(n^{-1/2})$. D_0 is then

called the influence function of ψ_n. Asymptotic linearity of ψ_n implies that ψ_n is weakly consistent for ψ_0 (by the Weak Law of Large Numbers) and that $n^{1/2}(\psi_n - \psi_0) \xrightarrow{\mathrm{d}} N_d(0, P_0 D_0^2)$ (by the Central Limit Theorem). If a parameter is pathwise differentiable at P_0 relative to $\mathcal{M}$ and there exists a consistent estimator of an influence function D_0, then there exist asymptotically linear estimators of the parameter with influence function D_0 (Pfanzagl, 1982). An estimator ψ_n of a pathwise differentiable parameter ψ_0 is *asymptotically efficient* with respect to $\mathcal{M}$ if it is asymptotically linear with influence function equal to the efficient influence function D_0^* for estimating ψ_0 with respect to $\mathcal{M}$.

Multiple general-purpose methods exist for constructing asymptotically linear (and efficient) estimators of pathwise differentiable parameters. Suppose that for each n, D_n is an estimator of an influence function D_0 based on the observations $O_1, \ldots, O_n$. By rearranging terms, we can always write

$$\psi_n = \psi_0 + \mathbb{P}_n D_0 + n^{-1/2}\mathbb{G}_n(D_n - D_0) + \psi_n - \psi_0 + P_0 D_n - \mathbb{P}_n D_n \ . \tag{1.1}$$

Then ψ_n would be asymptotically linear if $\mathbb{G}_n(D_n - D_0) \xrightarrow{\mathrm{P}} 0$ and $\psi_n - \psi_0 + P_0 D_n - \mathbb{P}_n D_n = o_{\mathrm{P}}(n^{-1/2})$. The first statement is true if D_n is contained in a P_0-Donsker class of functions with probability tending to one and $P_0(D_n - D_0)^2 \xrightarrow{\mathrm{P}} 0$ (van der Vaart, 1998, Lemma 19.24). The truth of the second statement depends on the form of ψ_n. We introduce two general constructions of ψ_n for which the second statement can frequently be shown to hold.

The first general construction we consider is the *one-step construction*. Suppose that ψ_0 and D_0 only depend on P_0 through a summary $\pi(P_0)$, and write $\psi(\pi) := \psi(\pi(P))$ and $D_\pi := D_{\pi(P)}$. We do not lose any generality by this assumption, since we can always take $\pi(P) = P$. Let π_n be an estimator of $\pi_0 := \pi(P_0)$. Then $\psi(\pi_n)$ is referred to as a *plug-in* estimator. Also let $D_n := D_{\pi_n}$. A one-step estimator is then defined as $\psi_n := \psi(\pi_n) + \mathbb{P}_n D_{\pi_n}$. We then have

$$\psi_n - \psi_0 + P_0 D_n - \mathbb{P}_n D_n = \psi(\pi_n) - \psi(\pi_0) + P_0 D_{\pi_n} \ .$$

This expression is known as a *second-order remainder*, whose form depends on the parameter

mapping and influence function. It can often be shown to be $o_P(n^{-1/2})$ provided the estimator π_n is tending to π_0 at a fast enough rate in an appropriate sense. We will make use of the one-step construction throughout this dissertation.

The second general approach we consider is *estimating equations-based estimators*. In many cases, the influence function has the form $D_0 = D_{\psi_0,\eta_0}$; i.e., D_P depends on P through ψ and an additional nuisance parameter η. Since ψ_0 then solves the equation $Z_0(\psi) = 0$ for $Z_0(\psi) := P_0 D_{\psi,\eta_0}$, the estimating equations approach then defines ψ_n as a solution to the equation $Z_n(\psi) = 0$ for $Z_n(\psi) := \mathbb{P}_n D_{\psi,\eta_n}$, where η_n is an estimator of η_0. With $D_n := D_{\psi_n,\eta_n}$, we then have

$$\psi_n - \psi_0 + P_0 D_n - \mathbb{P}_n D_n = \psi_n - \psi_0 + P_0 D_{\psi_n,\eta_n} \ .$$

The results in Chapter 3.3 of van der Vaart and Wellner (1996) can be used to study the asymptotics of estimating equations-based estimators. We also note that if $D_{\psi,\eta} = D^{(1)}_\eta - \psi$ for a function $D^{(1)}_\eta$ only depending on P through η, then $\psi_n = \mathbb{P}_n D^{(1)}_{\eta_n}$ solves the resulting estimating equation. In this case, the estimating equations-based estimator coincides exactly with a one-step estimator.

We also mention a third general approach to constructing asymptotically linear estimators based on *targeted maximum likelihood/minimum loss-based estimation*, frequently abbreviated as TMLE. As before, suppose that $\psi(P)$ and D_P only depend on P through summaries $\pi(P)$, and suppose π_n is an initial estimator of π_0. TMLE works by finding, via specific fluctuations of π_n, a sensible estimator π^*_n such that $\mathbb{P}_n D_{\pi^*_n} = 0$. Then, the targeted estimator is defined as $\psi_n := \psi(\pi^*_n)$. Therefore, TMLE estimators have the benefit of being plug-in estimators, so that they necessarily satisfy probabilistic bounds on ψ, unlike one-step and estimating equations-based estimators. Since we do not explicitly use TMLE in this dissertation, we do not provide more detail here, but instead refer the interested reader to van der Laan and Rose (2011).

1.2 Background and motivation

1.2.1 Empirical risk minimization

In many scientific settings, investigators are interested in learning about a function known to be monotone, either due to probabilistic constraints or in view of existing scientific knowledge. The statistical treatment of nonparametric monotone function estimation has a long and rich history. In the literature, most monotone function estimators have been constructed via empirical risk minimization. Specifically, suppose that $O_1, \ldots, O_n$ are independent observations on a measurable space $(\mathcal{O}, \mathcal{B})$ drawn from a common probability measure P_0. Let $\theta_0 : I \to \mathbb{R}$ be a function that we wish to estimate, where $I \subseteq \mathbb{R}$ is an interval, and suppose it is known that $\theta_0 \in \Theta_{\mathbb{R}}$, the space of non-decreasing functions on $\mathbb{R}$. Suppose we have access to a *loss function* $L : \Theta \times \mathcal{O} \to \mathbb{R}$ such that

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \int L(\theta, o)\, dP_0(o) \ .$$

We then call $R_0(\theta) := \int L(\theta, o)\, dP_0(o)$ the *oracle risk* of a candidate function θ. Since P_0 is not known in practice, neither is R_0, and hence we cannot base estimation of θ_0 on R_0. Instead, we define the *empirical risk* as $R_n(\theta) := \int L(\theta, o)\, d\mathbb{P}_n(o)$. We then take

$$\theta_n := \underset{\theta \in \Theta}{\operatorname{argmin}}\, R_n(\theta) = \underset{\theta \in \Theta}{\operatorname{argmin}} \int L(\theta, o)\, d\mathbb{P}_n(o) \ ,$$

and we say that θ_n is an *empirical risk minimizer* corresponding to the loss function L.

The above definitions of loss functions and empirical risk minimizers were used extensively by Vladimir Vapnik in the statistical learning literature (see, e.g., Vapnik, 1992, 1999, 2013). They are separate from, and not to be confused with, the definitions of loss and risk from statistical decision theory (see, e.g., Ferguson, 2014).

We note that it is not always immediately clear that a minimizer of $R_n(\theta)$ exists; often, proving that θ_n exists, in addition to providing an algorithm for finding it, is a substantial

task. Additionaly, in many cases, there is not a unique minimizer of $R_n(\theta)$, so that it would be more accurate to say $\theta_n \in \text{argmin}_{\theta \in \Theta} R_n(\theta)$. In such cases, we typically define a suitable convention that uniquely identifies θ_n.

Many monotone function estimators in the literature have been constructed via empirical risk minimization. This is because many of the parameters treated in the literature on monotone function estimation can be viewed as an index of the statistical model, in the sense that the model space is in bijection with the product space corresponding to the parameter of interest and an additional variation-independent parameter. In such cases, identifying an appropriate loss function is often easy, and a risk minimization representation is therefore usually available. However, the fact that empirical risk minimizers under no more than a monotonicity constraint have useful properties is often quite remarkable. We comment more on this in certain examples below.

In this dissertation, we are motivated by observational studies. Sampling complications such as informative treatment attribution, informative missingness, and informative censorship are common in observational studies. In the presence of such complications, it is frequently the case that identifying the parameter of interest with the observed data distribution requires adjustment for a set of observed covariates. Appropriate loss functions for such observed-data parameters may not be readily available, and if they are available, they may depend on unknown summaries of the data-generating distribution. This motivates us to study methods of estimating monotone functions that (1) do not rely on empirical risk minimization, and (2) permit data-adaptive estimation of unknown nuisance parameters.

Before outlining the contributions of this dissertation, we illustrate the preceding discussion in three examples: estimation of a distribution function, estimation of a monotone density function, and estimation of a monotone regression function. In each of these examples, we first discuss a classical estimation problem where an appropriate loss function is easy to find, and where the resulting empirical risk minimizer has been previously studied. We then illustrate the challenges that can arise in observational studies by considering identification of the parameter in the presence of a *coarsened at random* observed data structure

(Heitjan and Rubin, 1991; Gill et al., 1997).

1.2.2 Estimation of a distribution function

The simplest and most well-studied monotone parameter is a distribution function. Suppose that $O = Y$, so that $Y_1, \ldots, Y_n$ are independent real-valued observations drawn from a probability measure P_0. We note that the distribution function $\theta_0 : \mathbb{R} \to [0, 1]$ defined as $\theta_0(x) := P_0(Y \leq x)$ is necessarily non-decreasing on $\mathbb{R}$. For this parameter, the log-likelihood loss $L(\theta, y) := -\log d\theta(y)$, yields the nonparametric maximum likelihood estimator (NPMLE) θ_n, which can be shown to be equal to $\theta_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(Y_i)$, and is known as the *empirical distribution function*.

Suppose now that, in addition to Y, we observe a binary treatment A. Imagine that each unit in the population has two potential outcomes $Y(0)$ and $Y(1)$ were they to receive treatment $A = 0$ and $A = 1$, respectively. We wish to estimate the distribution function of $Y(1)$, but for each unit, we only observe the potential outcome $Y = Y(A)$, where A is the treatment that the unit actually received. If $Y(1)$ and A are independent, then $P(Y(1) \leq x) = P(Y \leq x \mid A = 1) := \theta_0(x)$. The log-likelihood loss can still be used for this conditional distribution function, and yields the NPMLE

$$\theta_n(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(Y_i)A_i}{\frac{1}{n}\sum_{i=1}^{n} A_i} ,$$

known as the *empirical conditional distribution function* of Y corresponding to $A = 1$.

If $Y(1)$ and A are not independent, then $P(Y(1) \leq x)$ is no longer equal to the conditional distribution of Y given $A = 1$. As we explain in Chapter 2, if $Y(1)$ and A are *conditionally* independent given a vector of observed covariates $W \in \mathbb{R}^d$, then $P(Y(1) \leq x)$ is equal to the so-called *G-computed distribution function* $\theta_0(x) := E_{P_0}[P_0(Y \leq x \mid A = 1, W)]$ (Robins, 1986). The outer expectation is taken with respect to the marginal distribution of W, and the inner probability is with respect to the conditional distribution of Y given A and W. Due to the outer average over the marginal distribution of W, it is more challenging to find

an appropriate loss function for the G-computed distribution function θ_0 than it was for the conditional distribution function of Y given $A = 1$, and it is thus difficult to specify an empirical risk minimizer of θ_0. We return to this example and provide a monotone estimator of θ_0 in Chapter 2.

1.2.3 Estimation of a monotone density function

Let the observed data be $O = T$, where T is an event time taking values in $\mathbb{R}^+$. Our parameter of interest is now the density function of T: $\theta_0(x) := \frac{d}{dx} P_0(T \leq x)$. We can again use the log-likelihood loss function, which now takes the form $L(\theta, t) = -\log \theta(t)$. Grenander (1956) showed that the resulting NPMLE of a monotone density function exists, and characterized it in terms of least concave majorants. This is now commonly referred to as Grenander's estimator. We return to this example in Chapter 3. Prakasa Rao (1969) developed the asymptotic theory of Grenander's estimator using its characterization as an empirical risk minimizer.

We note that it is *a priori* surprising that an empirical risk minimizer of a density function over the very large space of monotone functions has good statistical properties. Indeed, without the monotonicity constraint, the empirical risk minimizer of a density function degenerates to the "density" with point mass $1/n$ at each observation. This is not a proper density function with respect to Lebesgue measure, so it is not useful as an estimator of θ_0. The fact that empirical risk minimizers for challenging parameters such as the density function exist under the relatively mild monotonicity constraint may be seen as one of the motivations for incorporating knowledge of monotonicity into the estimation procedure when such knowledge is available.

An alternative approach to nonparametric estimation of a density function is kernel smoothing. However, kernel smoothing requires the choice of both a kernel function and a bandwidth. Grenander's estimator requires no such choice. This is a recurring theme for estimators of monotone parameters, and provides another motivation for incorporating knowledge of a monotonicity constraint.

Estimation of a monotone density can be extended to the case of independently right-censored data. Suppose now that the observed data are $O = (Y, \Delta)$, where $Y = \min\{T, C\}$ and $\Delta = I_{[0,C]}(T)$ for $C \in \mathbb{R}^+$ a right-censoring variable. Our parameter of interest θ_0 is still the density function of T, but since T is not directly observed, more work must be done to identify θ_0 using the observed data. If T and C are independent random variables, then θ_0 can be identified as $\theta_0(x) = -\frac{d}{dx} S_0(x)$ for

$$S_0(x) = \prod_{t \leq x} \left[1 - \frac{F_{1,0}(dt)}{S_{Y,0}(t)} \right] ,$$

where $F_{1,0}(t) := P_0(Y \leq t, \Delta = 1)$ is the subdistribution function of Y corresponding to $\Delta = 1$ and $S_{Y,0}(t) := P_0(Y \geq t)$ is the conditional proportion-at-risk at time t. Here and throughout, $\prod$ denotes the product-integral transform (Gill and Johansen, 1990). The NPMLE in this context was derived in Laslett (1982) and McNichols and Padgett (1982), and its asymptotic behavior was studied in Huang and Zhang (1994).

In some settings, it is not reasonable to expect that T and C are independent, in which case the above identification does not hold. Instead, if T and C are *conditionally* independent given $W \in \mathbb{R}^d$, we can still identify $\theta_0(x)$ as $-\frac{d}{dx} S_0(x)$, but now $S_0(x)$ is identified as:

$$S_0(x) = E_{P_0} \left\{ \prod_{t \leq x} \left[1 - \frac{F_{1,0}(dt \mid W)}{S_{Y,0}(t \mid W)} \right] \right\} , \tag{1.2}$$

where $F_{1,0}(t \mid w) := P_0(Y \leq t, \Delta = 1 \mid W = w)$ is the conditional subdistribution function of Y given $W = w$ corresponding to $\Delta = 1$ and $S_{Y,0}(t \mid w) := P_0(Y \geq t \mid W = w)$ is the conditional proportion-at-risk at time t given $W = w$. The identification in (1.2) was described by Beran (1981) in an unpublished technical report, and was used to provide an estimator of a survival function with conditionally independent right-censoring by Dabrowska (1989). We call (1.2) a *G-computed product integral*. As with the G-computed distribution function, it is not immediately clear what an appropriate loss function is for this parameter, and if we could find one, it would likely depend on unknown nuisance parameters that

would need to be estimated from the data. We will return to this discussion and provide an estimator of θ_0 under conditionally independent right-censoring in Chapter 3.

1.2.4 Estimation of a monotone regression function

Here, the observed data consist of $O = (Y, X)$, where $Y \in \mathbb{R}$ is an outcome and $X \in \mathbb{R}$ is a treatment. The parameter of interest is the regression of Y on X, $\theta_0(x) := E_{P_0}[Y \mid X = x]$. For this parameter, the least-squares loss function $L(\theta, o) = [y - \theta(x)]^2$ may be used. Brunk (1970) characterized the resulting empirical risk minimizer, which is known as the *least-squares isotonic regression*, or simply isotonic regression, of Y on X.

Again, it is surprising that a least-squares estimator under just a monotonicity constraint yields a useful estimator. Without the monotonicity constraint, the least-squares estimator at any x is the average of the observed Y-values corresponding to this x, if they exist, and is undefined otherwise. This estimator is not consistent at any x such that $P(X = x) = 0$. However, under a monotonicity constraint, the least-squares estimator exists, and again does not require choice of a kernel function or bandwidth.

Similarly as in the first example, suppose now that each unit in the population possesses a potential outcome $Y(x)$ for every possible value x of the treatment, representing the outcome the unit would have if their treatment were set to level x. We are interested in estimating $E[Y(x)]$, but as before, we only observe the outcome $Y = Y(X)$. If $Y(x)$ and X are independent, then $E[Y(x)] = E[Y \mid X = x]$, and least-squares isotonic regression may be used to estimate $E[Y(x)]$ using the observed data. However, if $Y(x)$ and X are only independent given a vector of observed covariate $W \in \mathbb{R}^d$, then we instead have $E[Y(x)] = E_{P_0}[E_{P_0}[Y \mid X = x, W]] := \theta_0(x)$, the G-computed regression function. For this parameter, (Kennedy et al., 2017) demonstrated that

$$L_0(\theta, o) := \left[\frac{y - \mu_0(x, w)}{g_0(x, w)} + \int \mu_0(x, w) Q_0(dw) - \theta(x) \right]^2$$

forms a proper loss function for θ_0. Here, $\mu_0(x, w) := E_{P_0}[Y \mid X = x, W = w]$ is the outcome

regression implied by P_0, $g_0(x, w) := [\frac{d}{dx} P_0(X \leq x \mid W = w)]/[\frac{d}{dx} P_0(X \leq x)]$, and Q_0 is the marginal distribution of W implied by P_0. We have indexed L by 0 to indicate that the loss function depends on these three unknown summaries of P_0. In order to define an empirical risk minimizer based on this loss function, we would need to plug in estimators μ_n, g_n, and Q_n for these nuisance parameters. Since we do not typically have access to correctly-specified parametric models for these nuisance parameters, we will want to permit these estimators to use flexible learning strategies. However, doing so will make existing approaches to providing asymptotic theory for the resulting estimator not applicable. We return to nonparametric estimation of this parameter briefly in Chapter 3, and again in more detail in Chapter 4.

1.3 Contribution and organization of the dissertation

In this dissertation, we study general procedures for constructing an estimator of a monotone function $\theta_0 := \theta_{P_0}$ on an interval $I \subseteq \mathbb{R}$. The general procedures we study do not rely on empirical risk minimization, so their use does not depend on having access to an appropriate loss function for the parameter of interest. They can therefore be used for parameters arising from observational studies, including the parameters described above.

We study two types of general procedures, designed for use in two distinct settings. In the first setting, we have access to an initial estimator θ_n of θ_0, but this initial estimator may fail to be monotone. This occurs frequently, for example, when the estimator is constructed pointwise over the domain, and each point in the domain requires a separate adjustment due to the sampling complications present in the study. In Chapter 2, we propose correcting such an initial estimator θ_n by projecting it into the space of monotone functions over a finite grid in the domain. This is a simple correction procedure, but like any post-estimation correction procedure, *a priori* it carries the risk of disturbing the statistical properties of the initial estimator. Our main contribution in Chapter 2 is the establishment of general results guaranteeing that the corrected estimator is always *at least as good* as the initial estimator, in a sense that we make precise. Under additional conditions, we establish rates at which the difference between the initial and corrected estimators goes to zero in probability. We apply

our general results to the specific problem of estimation and inference for a G-computed distribution function introduced above.

We primarily have in mind in Chapter 2 initial estimators that are constructed to be *uniformly asymptotically linear*, meaning for each $x \in I$, there exists a function $D_{0,x} : \mathcal{O} \mapsto \mathbb{R}$, possibly depending on P_0, such that $P_0 D_{0,x} = 0$, $P_0 D_{0,x}^2 < \infty$ and

$$\theta_n(x) = \theta_0(x) + \mathbb{P}_n D_{0,x} + R_{n,x} \tag{1.3}$$

for a remainder term $R_{n,x}$ with $n^{1/2} \sup_{x \in I} |R_{n,x}| = \mathrm{o_P}(1)$. Uniform asymptotic linearity of θ_n is powerful because it immediately implies a variety of asymptotic properties of θ_n, including pointwise weak consistency, uniform weak consistency (if the class $\{D_{0,x} : x \in I\}$ is P_0-Glivenko Cantelli), pointwise asymptotic normality, and weak convergence as a process indexed by I (if the class $\{D_{0,x} : x \in I\}$ is P_0-Donsker).

Uniformly asymptotically linear estimators can be constructed when $\theta_P(x)$ is pathwise differentiable at P_0 for each $x \in I$ and there is a uniformly consistent estimator of a set of influence functions $\{D_{x,0} : x \in I\}$. Furthermore, there exist multiple general-purpose methods for constructing asymptotically linear estimators of pathwise differentiable parameters, as we discussed above. However, since such estimators perform a correction separately for each point in the domain, there is no guarantee that these estimators are monotone.

In the second setting we consider, which we address in Chapter 3, an initial nonparametric estimator of θ_0 may not be readily available, but a nonparametric estimator of a primitive function of θ_0 *is* available. Primarily, we have in mind in this setting situations where $P \mapsto \theta_P(x)$ is not pathwise differentiable with respect to $\mathcal{M}$, but a primitive function $\Gamma_P(x)$ is pathwise differentiable, and hence non- or semiparametric methods may be used to construct a uniformly asymptotically linear estimator Γ_n of $\Gamma_0 := \Gamma_{P_0}$. For instance, the density function of a continuous random variable is not a pathwise differentiable parameter in a nonparametric model, and therefore asymptotically linear estimators of a density do not exist. However, the primitive function of a density is a cumulative distribution function,

which often is pathwise differentiable. Similarly, the hazard function of a continuous random variable is not pathwise differentiable in a nonparametric model, but a cumulative hazard function typically is pathwise differentiable.

We define a broad class of *generalized Grenander-type* estimators based on differentiating the greatest convex minorant or least concave majorant of an estimator of a primitive function of θ_0. This broad class allows the minorization or majoratization operation to be performed on a data-dependent transformation of the domain, which has two benefits. First, this generalization allows many existing estimators from the literature to be represented in our class. Second, allowing a data-dependent transformation of the domain can provide simplified estimation procedures in more complicated problems that have not yet been considered. We provide general conditions for consistency and pointwise convergence in distribution of estimators in this class. Additionally, we provide simpler conditions and more concrete distributional theory in the important case that the primitive estimator and data-dependent transformation function are uniformly asymptotically linear.

We use our general results in the context of three well-studied problems, and show that we readily recover classical results established separately in each case. Specifically, we consider estimation of monotone density, hazard, and regression functions. In the first two cases, we study both fully observed data and independently right-censored data. We also consider extensions of the three classical monotone problems above to more complex settings in which covariates must be accounted for, because either the censoring process or the treatment allocation mechanism are informative, as is typical in observational studies. Specifically, we derive novel estimators of monotone density and hazard functions for use when the survival data are subject to right-censoring that may depend on covariates, and a novel estimator of a monotone dose-response curve for use when the relationship between the exposure and outcome is confounded by recorded covariates. Unlike for their classical analogues, in these more difficult problems, nonparametric estimation of the primitive function involves nuisance functions for which flexible estimation strategies (e.g., machine learning) must be employed.

In Chapter 4, we conduct a detailed study of the final parameter considered in Chap-

ter 3: the monotone G-computed regression function. G-computed regression functions, introduced above, describe the effect of a continuous exposure on an outcome while adjusting for potential confounders. Classical methods for estimating such curves often rely on overly restrictive parametric assumptions, which carry a significant risk of model misspecification. It is therefore of interest to estimate and draw inference about these curves without such strong modeling assumptions. Nonparametric estimation in this context is challenging because in a nonparametric model these curves cannot be estimated at regular rates and available estimators will generally be sensitive to the selection of certain tuning parameters.

In Chapter 4, we show that if the covariate-adjusted regression curve is monotone, nonparametric estimation and inference is possible without the need to select tuning parameters and under minimal smoothness conditions. Our proposed estimation procedure generalizes the classical least-squares isotonic regression estimator of a monotone regression function. We use the general results provided in Chapter 3 to describe theoretical properties of our proposed estimator, including its irregular asymptotic limit distribution and the potential for doubly-robust inference. We also illustrate its practical performance via numerical studies, and use our method to assess the effect of BMI on immune response in HIV vaccine trials.

Finally, Chapter 5 presents a discussion of the work and some potential future directions.

Chapter 2

CORRECTING AN ESTIMATOR OF A MULTIVARIATE MONOTONE FUNCTION WITH ISOTONIC REGRESSION

2.1 Introduction

2.1.1 Background

In this chapter, we consider estimation of a possibly multivariate monotone function in the first setting introduced in Chapter 1: an initial estimator of the monotone function is available, and may have several desirable statistical properties, yet fail to be monotone. This often occurs when this estimator is obtained through the pointwise application of a statistical procedure over the domain of the function. For instance, we may be interested in estimating a G-computed distribution function θ_0, defined pointwise as $\theta_0(t) := E_{P_0}[P_0(Y \leq t \mid A = 1, W)]$, over its domain $\mathbb{R}$. Here, Y represents an outcome, A is a binary exposure, and W is a vector of observed confounders of the exposure-outcome relationship. The map $t \mapsto \theta_0(t)$ is necessarily monotone. As we discuss in Section 2.4, many asymptotically efficient estimators of $\theta_0(t)$ are constructed pointwise for each t, and due to random variation, need not be monotone as a function of t.

This thesis is primarily focused on univariate monotone functions, but in this chapter, we consider the more general setting of multivariate monotone functions. For instance, suppose now that X is a categorical exposure taking values in $\mathcal{K} = \{1, \ldots, K\}$ and define $\theta_0(x, t) = E_{P_0}[P_0(Y \leq t \mid X = x, W)]$, over its domain $\mathbb{R} \times \mathcal{K}$. As before, the map $t \mapsto \theta_0(x, t)$ is necessarily monotone for each fixed x. In some scientific contexts, it may also be known that $x \mapsto \theta_0(x, t)$ is monotone for each t, in which case θ_0 is a bivariate component-wise monotone function. Again, estimators constructed pointwise for each x and t are not necessarily guaranteed to be monotone in either component.

In many situations, failure of an estimator to be monotone is a serious problem. This is most apparent if the monotonicity constraint is probabilistic in nature – that is, there does not exist a distribution such that the parameter mapping is not monotone. This is the case, for instance, if θ_0 is a univariate distribution function. In such settings, returning an estimate which fails to be monotone is nonsensical, like returning an estimate of a probability outside the interval $[0, 1]$. However, even if the monotonicity constraint is based on scientific rather than probabilistic knowledge, failure of an estimator to be monotone can be problematic. For example, if the parameter of interest represents average height or weight in children as a function of age, clinical collaborators and other scientists would be confused by a non-monotone estimate. Finally, as we will see, there are often finite-sample benefits to be gained from incorporating the monotonicity constraint.

Whenever this phenomenon occurs, it is natural to seek an estimator that respects the monotonicity constraint but nevertheless remains close to the initial estimator, which may otherwise be known to have good statistical properties. A monotone estimator can be constructed by projecting the initial estimator onto the space of monotone functions with respect to some norm. A common choice is the L_2-norm, which amounts to using multivariate isotonic regression to correct the initial estimator. Chernozhukov et al. (2009) demonstrated that, for each $p \in [1, \infty]$, this corrected estimator is closer to the truth in L_p-norm and that confidence bands similarly obtained generally have smaller L_p-size while having no smaller L_p-coverage. Therefore, the estimator obtained via isotonic regression inherits the L_p-rates of convergence of the initial estimator, and confidence bands obtained in this fashion are valid, though possibly conservative.

2.1.2 Contribution and organization of the chapter

In this chapter, we discuss correcting an initial estimator of a multivariate monotone function by computing the isotonic regression of the estimator over a finite grid in the domain, and interpolating between grid points. We consider correcting an initial confidence band by the same procedure applied to the upper and lower limits. We provide three main results

regarding this correction procedure:

1. we demonstrate that the corrected estimator is *at least as good* than the initial estimator in the sense that (a) its uniform loss over the grid used for projecting is less than or equal to that of the initial estimator for every sample, (b) its uniform loss over the entire domain is less than or equal to the initial estimator asymptotically, and (c) the corrected confidence band contains the true function on the grid used for projecting whenever the initial band does with no cost to average or uniform widths of the band;

2. we provide general sufficient conditions under which the uniform difference between the initial and corrected estimators is $o_P(r_n^{-1})$ for a generic sequence $r_n \to \infty$; and

3. we provide simpler sufficient conditions for uniform asymptotic equivalence at the rate $r_n = n^{1/2}$ when the initial estimator is uniformly asymptotically linear.

We note here that the projection approach is not the only possible correction procedure. Chernozhukov et al. (2009) and Chernozhukov et al. (2010) studied a correction based on monotone rearrangements and compared the two procedures. However, monotone rearrangements do not generalize to the multivariate setting as naturally as projections – for example, Chernozhukov et al. (2009) proposed averaging a variety of possible multivariate monotone rearrangements to obtain a final monotone estimator. By contrast, the projection over the space of multivariate monotone functions is unique in both univariate and multivariate contexts. Additionally, projections are easy to understand, rely on familiar tools, and have minimal squared-error impact on the initial estimator among all candidate monotone corrections.

Daouia and Park (2013) also proposed a correction procedure that consists of taking a convex combination of upper and lower monotone envelope functions. They demonstrated that their estimator is asymptotically equivalent to the correction based on projecting in supremum norm. Therefore, application of their results requires studying the properties of the projection correction. This is another motivation for the results we provide here.

The chapter proceeds as follows. In the next section we formally define our statistical setting. In Section 4.3, we present our results concerning properties of the projected estimator. In Section 2.4, we check our conditions for covariate-adjusted cumulative incidence curves, both without censoring and with conditionally independent right-censoring. Finally, in Section 2.5, we present a numerical study.

2.2 Correction procedure

Let $\mathcal{M}$ be a statistical model of probability measures on a probability space $(\mathcal{O}, \mathcal{B})$. Let $\theta : \mathcal{M} \to \ell^\infty(\mathcal{T})$ be a parameter of interest on $\mathcal{M}$, where $\mathcal{T} = [0,1]^d$ is the d-dimensional unit cube and $\ell^\infty(\mathcal{T})$ is the Banach space of bounded functions from $\mathcal{T}$ to $\mathbb{R}$ equipped with supremum norm $\|\cdot\|_\mathcal{T}$. We have specified this particular $\mathcal{T}$ for simplicity, but the results established here apply to any bounded rectangular domain $\mathcal{T} \subset \mathbb{R}^d$. For each $P \in \mathcal{M}$, denote by θ_P the evaluation of θ at P and note that θ_P is a bounded real-valued function on $\mathcal{T}$. For any $t \in \mathcal{T}$, denote by $\theta_P(t) \in \mathbb{R}$ the evaluation of θ_P at t.

For any vector $t \in \mathbb{R}^d$ and $1 \leq j \leq d$, denote by z_j the j^{th} component of z. Define the partial order $\leq$ on $\mathbb{R}^d$ by setting $t \leq t'$ if and only if $t_j \leq t'_j$ for each $1 \leq j \leq d$. Note that $t < t'$ if and only if $t \leq t'$ and $t \neq t'$, which holds if and only if $t_j \leq t'_j$ for each j and $t_j < t'_j$ for at least one j. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (component-wise) monotone non-decreasing if $t \leq t'$ implies that $f(t) \leq f(t')$. Denote $\|t\| = \max_{1 \leq j \leq d} |t_j|$ for any vector $t \in \mathbb{R}^d$. Additionally, denote by $\mathbf{\Theta} \subset \ell^\infty(\mathcal{T})$ the convex set of bounded monotone non-decreasing functions from $\mathcal{T}$ to $\mathbb{R}$. For concreteness we focus on non-decreasing functions, but all results established here apply equally to non-increasing functions.

Let $\mathcal{M}_0 = \{P \in \mathcal{M} : \theta_P \in \mathbf{\Theta}\} \subseteq \mathcal{M}$ and suppose that $\mathcal{M}_0$ is nonempty. Generally, this inclusion is strict only if, rather than being implied by the rules of probability, the monotonicity constraint stems at least in part from prior scientific knowledge. Also, define $\mathbf{\Theta}_0 = \{\theta \in \mathbf{\Theta} : \theta = \theta_P, P \in \mathcal{M}\} \subseteq \mathbf{\Theta}$. We are primarily interested in settings where $\mathbf{\Theta}_0 = \mathbf{\Theta}$, since in this case there is no additional knowledge about θ encoded by $\mathcal{M}$, and in particular there is no danger of yielding a corrected estimator that is compatible with no $P \in \mathcal{M}$.

Suppose that observations $O_1, O_2, \ldots, O_n$ are sampled independently from an unknown distribution $P_0 \in \mathcal{M}_0$, and that we wish to estimate $\theta_0 := \theta_{P_0}$ based on these observations. Suppose that, for each $t \in \mathcal{T}$, we have access to an estimator (i.e. a measurable function of $O_1, O_2, \ldots, O_n$) $\theta_n(t)$ of $\theta_0(t)$. (Note that the assumption that the data are independent and identically distributed is not necessary for our Theorems 1 and 2 below.)

The central premise of this chapter is that $\theta_n(t)$ may have desirable statistical properties for each t or even uniformly as element of $\ell^\infty(\mathcal{T})$, but that θ_n as an element of $\ell^\infty(\mathcal{T})$ may not fall in $\boldsymbol{\Theta}$ for any finite n or even with probability tending to one. Our goal is to provide a corrected estimator θ_n^* which necessarily falls in $\boldsymbol{\Theta}$, and yet retains the statistical properties of θ_n. A natural way to accomplish this is to define θ_n^* as the closest element of $\boldsymbol{\Theta}$ to θ_n in some norm on $\mathcal{T}$. Ideally, we would prefer to take θ_n^* to minimize $\|\theta - \theta_n\|_\infty$ over $\theta \in \boldsymbol{\Theta}$. However, this is not tractable for two reasons.

The first reason this solution is not tractable is that optimization over the entirety of $\mathcal{T}$ is an infinite-dimensional optimization problem, and is hence frequently not possible to perform in practice. To resolve this issue, for each n, we let $\mathcal{T}_n = \{t_1, t_2, \ldots, t_{m_n}\} \subseteq \mathcal{T}$ be a possibly random finite rectangular lattice in $\mathcal{T}$ over which we will perform the optimization, and define $\|\cdot\|_{\mathcal{T}_n}$ as the supremum norm over $\mathcal{T}_n$. Let $\omega_n = \sup_{t \in \mathcal{T}} \min_{s \in \mathcal{T}_n} \|t - s\|$ be the mesh of $\mathcal{T}_n$ in $\mathcal{T}$, which may be random if $\mathcal{T}_n$ is random. While it is now computationally feasible to define $\theta_{n,\infty}^*$ as a minimizer over $\theta \in \boldsymbol{\Theta}$ of the finite-dimensional objective function $\|\theta - \theta_n\|_{\mathcal{T}_n}$, this objective function is challenging due to its non-differentiability. Instead, we will define

$$\theta_n^* \in \operatorname*{argmin}_{\theta \in \boldsymbol{\Theta}} \sum_{t \in \mathcal{T}_n} \{\theta(t) - \theta_n(t)\}^2 \ . \tag{2.1}$$

The squared-error objective function is smooth in its arguments. Furthermore, in dimension $d = 1$, θ_n^* thus defined is simply the isotonic regression of θ_n on the grid $\mathcal{T}_n$, the solution to which has a closed form representation as the greatest convex minorant of the so-called cumulative sum diagram. Furthermore, since $\|\theta_n^* - \theta_n\|_{\mathcal{T}_n} \geq \|\theta_{n,\infty}^* - \theta_n\|_{\mathcal{T}_n}$, all of our results apply to $\theta_{n,\infty}^*$ as well as θ_n^*.

We allow $\mathcal{T}_n$ to be random because in some circumstances, it may be natural to choose the grid based on the data. For instance, suppose the computational cost of estimating $\theta_n(t_k)$ at each t_k is large and the computational cost is linear in the number of grid points m_n at which the estimation is to be performed. If the user has a fixed computational budget, then the number of grid points used in practice may depend on the computational burden of estimating each $\theta_n(t_k)$, which may itself depend on features of the observed data.

We note that θ_n^* is only uniquely defined on $\mathcal{T}_n$. To completely characterize θ_n, we must monotonically interpolate function values between elements of $\mathcal{T}_n$. We will permit any monotonic interpolation that satisfies a mild condition. By the definition of a rectangular lattice, every $t \in \mathcal{T}$ can be assigned a hyper-rectangle whose vertices $\{s_1, s_2 \ldots, s_{2^d}\}$ are elements of $\mathcal{T}_n$ and whose interior has empty intersection with $\mathcal{T}_n$. If multiple such hyper-rectangles exist for t, such as when t lies on the boundary of two or more such hyper-rectangles, one can be assigned arbitrarily. We will then assume that, for $t \notin \mathcal{T}_n$, $\theta_n^*(t) = \sum_k \lambda_k(t)\theta_n^*(s_k)$ for $\sum_k \lambda_k(t) = 1$ and $0 \leq \lambda_k(t) \leq 1$ for each $k = 1, 2, \ldots, 2^d$. In other words, we assume that $\theta_n^*(t)$ is a convex combination of the values of θ_n^* on the vertices of the hyper-rectangle containing t. A simple interpolation approach consists of setting $\theta_n^*(t) = \theta_n^*(t')$ with t' the element of $\mathcal{T}_n$ closest to t, and choosing any such element if there are multiple elements of $\mathcal{T}_n$ equally close to t. This particular scheme satisfies our requirement.

Finally, for each n, we let $\ell_n(t) \leq u_n(t)$ denote lower and upper endpoints of a confidence band for $\theta_0(t)$. We then define ℓ_n^* and u_n^* as the corrected versions of ℓ_n and u_n, using the same projection and interpolation procedure defined above for obtaining θ_n^* from θ_n.

2.3 Properties of the projected estimator

2.3.1 Basic properties

The projected estimator θ_n^* is the isotonic regression of θ_n over the grid $\mathcal{T}_n$. Hence, many existing finite-sample results on isotonic regression can be used to deduce properties of θ_n^*. Theorem 1 below collects a few of these properties, building upon the results of Barlow et al.

(1972) and Chernozhukov et al. (2009).

Theorem 1. *(i) For every n, $\|\theta_n^* - \theta_0\|_{\mathcal{T}_n} \leq \|\theta_n - \theta_0\|_{\mathcal{T}_n}$.*

(ii) If $\omega_n = o_P(1)$ and θ_0 is uniformly continuous on $\mathcal{T}$, then $\|\theta_n^* - \theta_0\|_{\mathcal{T}} \leq \|\theta_n - \theta_0\|_{\mathcal{T}} + o_P(1)$.

(iii) If $\sup\{|\theta_0(t) - \theta_0(s)| : t, s, \in \mathcal{T}, \|t - s\| \leq \delta\} = o(\|\delta\|^\alpha)$ as $\delta \to 0$ for $\alpha > 0$, then $\|\theta_n^* - \theta_0\|_{\mathcal{T}} \leq \|\theta_n - \theta_0\|_{\mathcal{T}} + o_P(\omega_n^\alpha)$.

(iv) The event $\{\forall t \in \mathcal{T}_n : \theta_0(t) \in [\ell_n(t), u_n(t)]\}$ is contained in the event $\{\forall t \in \mathcal{T}_n : \theta_0(t) \in [\ell_n^*(t), u_n^*(t)]\}$.

(v) $\sum_{t \in \mathcal{T}_n}(u_n^*(t) - \ell_n^*(t)) = \sum_{t \in \mathcal{T}_n}(u_n(t) - \ell_n(t))$ and $\|u_n^* - \ell_n^*\|_{\mathcal{T}_n} \leq \|u_n - \ell_n\|_{\mathcal{T}_n}$.

Theorem 1 (i) says that the estimation error of of θ_n^* over the grid $\mathcal{T}_n$ is never worse than that of θ_n, and (ii) and (iii) say that the estimation error of θ_n^* on all of $\mathcal{T}$ is asymptotically no worse than the estimation error of θ_n in supremum norm. Similarly, (iv) says that the isotonized band $[\ell_n^*, u_n^*]$ never has worse coverage than the original band over $\mathcal{T}_n$, and (v) says that the potential increase in coverage comes at no cost to the average or supremum width of the bands over $\mathcal{T}_n$.

2.3.2 Asymptotic equivalence

While powerful, Theorem 1 does not rule out the possibility that θ_n^* performs strictly better, even asymptotically, than θ_n, or that the band $[\ell_n^*, u_n^*]$ is asymptotically strictly more conservative than $[\ell_n, u_n]$. In order to construct confidence intervals or bands with correct asymptotic coverage, a stronger result is needed: it must be that $\|\theta_n^* - \theta_n\|_{\mathcal{T}} = o_P(r_n^{-1})$, where r_n is a diverging sequence such that $r_n\|\theta_n - \theta_0\|_{\mathcal{T}}$ is converging in distribution to a non-degenerate limit distribution. Then, we would have that $r_n\|\theta_n^* - \theta_0\|_{\mathcal{T}}$ converges in distribution to this same limit, and hence confidence bands constructed based on approximating this limit distribution would have correct coverage when centered around θ_n^*, as we discuss more below.

We consider the following conditions on the true function θ_0 and the initial estimator θ_n:

(A1) there exists a deterministic sequence r_n tending to infinity such that, for all $\delta > 0$,

$$\sup_{\|t-s\|<\delta/r_n} |r_n \{\theta_n(t) - \theta_0(t)\} - r_n \{\theta_n(s) - \theta_0(s)\}| \xrightarrow{\mathrm{P}} 0;$$

(A2) there exist $0 < K_0 \leq K_1 < \infty$ such that $K_0\|t-s\| \leq |\theta_0(t) - \theta_0(s)| \leq K_1\|t-s\|$ for all $t, s \in \mathcal{T}$.

Condition (A1) is related to, but slightly weaker than, uniform stochastic equicontinuity (van der Vaart and Wellner, 1996, p. 37). (A1) would follow if, in particular, the process $\{r_n(\theta_n(t) - \theta_0(t)) : t \in \mathcal{T}\}$ would converge weakly to a tight limit process in the space $\ell^\infty(\mathcal{T})$. However, weak convergence to a tight limit is not necessary for (A1) to hold. This is important for application of our results to kernel smoothing-type estimators, which typically do not converge weakly to a tight limit, but for which condition (A1) can nevertheless often be shown to hold.

Condition (A2) may be thought of as constraining the *lower modulus of monotonicity* of θ_0, and is slightly more restrictive than a requirement for strict monotonicity. If, for instance, θ_0 is differentiable, then (A2) is satisfied with $\alpha = 1$ if all first-order partial derivatives of θ_0 are bounded away from zero and above. If instead some partial derivatives are zero, but only at isolated points, (A2) may still be satisfied with some $\alpha < 1$.

Based on these conditions, we have the following result.

Theorem 2. *If (A1)–(A2) hold and* $\omega_n = \mathrm{o_P}(r_n^{-1})$*, then* $\|\theta_n^* - \theta_n\|_{\mathcal{T}} = \mathrm{o_P}(r_n^{-1})$.

This result indicates that the projected estimator is uniformly asymptotically equivalent to the original estimator in supremum norm at the rate r_n. In addition to conditions (A1)–(A2), Theorem 2 requires that the mesh ω_n of $\mathcal{T}_n$ tends to zero in probability faster than r_n^{-1}. Since $\mathcal{T}_n$ is chosen by the user, this is not a problem in practice.

The left-hand side of the inequality in condition (A2) excludes, for instance, situations in which θ_0 is differentiable with null derivative over an interval. In such cases, θ_n^* may have

strictly smaller variance on these intervals than θ_n because θ_n^* will pool estimates across the flat region while θ_n may not. Hence, it is possible that θ_n^* may asymptotically improve on θ_n, so that θ_n^* and θ_n are not asymptotically equivalent at the rate r_n.

Our proof of Theorem 2 follows from three lemmas, which we state here. The first lemma controls the size of deviations in θ_n over small neighborhoods.

Lemma 1. *If (A1) and (A2) hold and* $b_n = \mathrm{o_P}(r_n^{-1})$, *then* $\sup_{\|t-s\| \leq b_n} |\theta_n(t) - \theta_n(s)| = \mathrm{o_P}(r_n^{-1})$.

The second lemma controls the size of neighborhoods over which violations in monotonicity can occur. Henceforth, we define $h_n = \sup \{\|t - s\| : s, t \in \mathcal{T}, s \leq t, \theta_n(t) \leq \theta_n(s)\}$.

Lemma 2. *If (A1) and (A2) hold, then* $h_n = \mathrm{o_P}(r_n^{-1})$.

Our final lemma bounds the maximal absolute deviation between θ_n^* and θ_n over the grid $\mathcal{T}_n$ in terms of the supremal deviations of θ_n over neighborhoods smaller than h_n.

Lemma 3. *The inequality* $\max_{t \in \mathcal{T}_n} |\theta_n^*(t) - \theta_n(t)| \leq \sup_{\|s-t\| \leq h_n} |\theta_n(s) - \theta_n(t)|$ *holds.*

2.3.3 Construction of confidence bands

Suppose there exists a fixed function $\gamma_\alpha : \mathcal{T} \to \mathbb{R}$ such that ℓ_n and u_n satisfy:

$$\begin{aligned} \|r_n(\theta_n - \ell_n) - \gamma_\alpha\|_{\mathcal{T}} &\xrightarrow{\mathrm{P}} 0 \\ \|r_n(u_n - \theta_n) - \gamma_\alpha\|_{\mathcal{T}} &\xrightarrow{\mathrm{P}} 0 \\ P_0\big(\forall t \in \mathcal{T} : r_n|\theta_n(t) - \theta_0(t)| \geq \gamma_\alpha(t)\big) &\longrightarrow 1 - \alpha \ . \end{aligned} \tag{2.2}$$

As an example of a confidence band that satisfies display (2.2), suppose $\sigma_0(t) > 0$ is a scaling function and c_α is a fixed constant such that, as n tends to infinity,

$$P_0\left(r_n \left\| \frac{\theta_n - \theta_0}{\sigma_0} \right\|_{\mathcal{T}} \geq c_\alpha\right) \longrightarrow 1 - \alpha \ .$$

If $\sigma_n(t)$ is an estimator of $\sigma_0(t)$ satisfying $\|\sigma_n - \sigma_0\|_{\mathcal{T}} \xrightarrow{\mathrm{P}} 0$ and $c_{\alpha,n}$ is an estimator of c_α satisfying $c_{\alpha,n} \xrightarrow{\mathrm{P}} c_\alpha$, then the Wald-type band with lower and upper endpoints $\ell_n(t) = \theta_n(t) - c_{\alpha,n} r_n^{-1} \sigma_n(t)$ and $u_n(t) = \theta_n(t) + c_\alpha r_n^{-1} \sigma_n(t)$ satisfies display (2.2) with $\gamma_\alpha := c_\alpha \sigma_0$. However, (2.2) can also be satisfied for other types of bands, such as bands constructed with a consistent bootstrap procedure.

Conditions (A1) and (A2) also imply that, under (2.2), the corrected confidence band $[\ell_n^*, u_n^*]$ have the same asymptotic coverage as the original band $[\ell_n, u_n]$, as we state in our next result.

Corollary 1. *If* (2.2) *holds then* $[\ell_n, u_n]$ *has asymptotic coverage* $1 - \alpha$*. If conditions (A1) and (A2) also hold,* γ_0 *is uniformly continuous on* $\mathcal{T}$*, and* $\omega_n = \mathrm{o_P}(r_n^{-1})$*, then* $[\ell_n^*, u_n^*]$ *also has asymptotic coverage* $1 - \alpha$*.*

We also note that Theorem 2 immediately implies that Wald-type confidence bands constructed around θ_n have the same asymptotic coverage if they are constructed around θ_n^* instead.

2.3.4 Special case: asymptotically linear estimators

Suppose now that the initial estimator θ_n is uniformly asymptotically linear: for each $t \in \mathcal{T}$, there exists a function $D_{0,t}^* : \mathcal{O} \mapsto \mathbb{R}^d$ depending on P_0 such that $\int D_{0,t}^*(o)\, dP_0(o) = 0$, $\int D_{0,t}^{*2}(o)\, dP_0(o) < \infty$ and

$$\theta_n(t) = \theta_0(t) + \frac{1}{n} \sum_{i=1}^{n} D_{0,t}^*(O_i) + R_{n,t} \tag{2.3}$$

for a remainder term $R_{n,t}$ with $n^{1/2} \sup_{t \in \mathcal{T}} |R_{n,t}| = \mathrm{o_P}(1)$. The function $D_{0,t}^*$ is the influence function of $\theta_n(t)$ under sampling from P_0. It is desirable for θ_n to have representation (2.3) because this then immediately implies its uniform weak consistency as well as the pointwise asymptotic normality of elements in the process $\theta_{n,0} = \{n^{1/2} \{\theta_n(t) - \theta_0(t)\} : t \in \mathcal{T}\}$. If in addition the collection $\{D_{0,t}^* : t \in \mathcal{T}\}$ of influence functions forms a P_0-Donsker class,

the process $\theta_{n,0}$ converges weakly in $\ell^\infty(\mathcal{T})$ to a Gaussian process with covariance function $\Sigma_0 : (t,s) \mapsto \int D_{0,t}(o) D_{0,s}(o)\, dP_0(o)$. Uniform asymptotic confidence bands based on θ_n can then be formed by using appropriate quantiles from any suitable approximation of the distribution of the maximum of the limiting Gaussian process.

We introduce two additional conditions:

(A1.i) the collection $\{D^*_{0,t} : t \in \mathcal{T}\}$ of influence curves is a P_0-Donsker class;

(A1.ii) Σ_0 is uniformly continuous in the sense that $\limsup_{\|t-s\|\to 0} |\Sigma_0(s,t) - \Sigma_0(t,t)| = 0$.

Whenever θ_n is uniformly asymptotically linear, Theorem 2 can be shown to hold under (A1.i), (A1.ii) and (A2), as implied by the theorem below. The validity of (A1.i) and (A1.ii) can be assessed by scrutinizing the influence function $D^*_{0,t}$ of $\theta_n(t)$ for each $t \in \mathcal{T}$. This fact renders the verification of these conditions very simple once uniform asymptotic linearity has been established.

Theorem 3. *For any estimator θ_n satisfying* (2.3)*, (A1.i) and (A1.ii) together imply (A1).*

2.4 Applications of the general theory

2.4.1 Estimation of a G-computed distribution function

We first demonstrate the use of Theorem 3 in the particular problem where we wish to draw inference on a G-computed distribution function. Suppose that the data unit is the vector $O = (Y, A, W)$, where Y is an outcome, $A \in \{0,1\}$ is a treatment indicator, and W is a vector of baseline covariates. The observed data consist of independent draws $X_1, X_2, \ldots, X_n$ from $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is an unrestricted nonparametric model.

For $P \in \mathcal{M}$, we define the parameter value θ_P pointwise as

$$\theta_P(t) := E_P \left\{ P\left(Y \le t \mid A = 1, W\right) \right\},$$

the G-computed distribution function of Y evaluated at t. We are interested in estimating $\theta_0 := \theta_{P_0}$. This parameter is often of interest because, under certain causal identification conditions, θ is the distribution function of the counterfactual outcome $Y(1)$ defined by the intervention that deterministically sets treatment $A = 1$ (Robins, 1986; Gill and Robins, 2001). When there are unmeasured confounders of the relationship between A and Y, the parameter may no longer have a causal interpretation but could still be appealing as a covariate-adjusted marginal summary of the conditional distribution of Y given $A = 1$.

For each t, the parameter $P \mapsto \theta_P(t)$ is pathwise differentiable in the nonparametric model, and its nonparametric efficient influence function at $P \in \mathcal{M}$ is given by

$$o = (y, a, w) \;\mapsto\; D^*_{P,t}(o) = \frac{a}{g_P(w)} \left\{ I_{[0,t]}(y) - \bar{Q}_P(t \mid w) \right\} + \bar{Q}_P(t \mid w) - \theta_P(t) \;,$$

where $g_P(w) = P(A = 1 \mid W = w)$ is the propensity score and

$$\bar{Q}_P(t \mid w) = P\left(Y \leq t \mid A = 1, W = w\right)$$

is the conditional treatment-specific distribution function, as implied by P (van der Laan and Robins, 2003). Given estimators g_n and $\bar{Q}_n$ of $g_0 = g_{P_0}$ and $\bar{Q}_0 = \bar{Q}_{P_0}$, respectively, several approaches can be used to construct, for each t, an asymptotically linear estimator of $\theta_0(t)$ with influence function $D^*_{0,t} := D^*_{P_0,t}$. For example, the use of either optimal estimating equations or the one-step correction procedure leads to the doubly-robust estimator

$$\theta_n(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i}{g_n(W_i)} \left\{ I_{[0,t]}(Y_i) - \bar{Q}_n(t \mid W_i) \right\} + \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n(t \mid W_i)$$

as discussed in detail in van der Laan and Robins (2003). Under certain conditions on g_n and $\bar{Q}_n$, including consistency at fast enough rates, $\theta_n(t)$ is asymptotically efficient relative to $\mathcal{M}$. In this case, $\theta_n(t)$ satisfies (2.3) with influence function $D^*_{0,t}$. However, there is no guarantee that θ_n is monotone.

In the context of this example, we can identify simple sufficient conditions under which conditions (A1) and (A2), and hence the asymptotic equivalence of the initial and isotonized estimators of the G-computed distribution function, are guaranteed. Specifically, we find that this is the case provided (i) there exists some $\eta > 0$ such that $g_0(W) \geq \eta$ almost surely under P_0, and (ii) there exist non-negative real-valued functions K_1, K_2 such that, under P_0, $K_1(W)$ has non-zero probability of being strictly positive and $K_2(W)$ has finite second moment, and additionally, $K_1(w)|t-s| \leq |\bar{Q}_0(t \mid w) - \bar{Q}_0(t \mid w)| \leq K_2(w)|t-s|$ for all $t, s \in \mathcal{T}$.

2.4.2 Estimation of a G-computed survival function with right-censored data

We now extend the methodology presented in the previous section to estimation and inference on G-computed survival curves with possibly informative censoring. Suppose that T and C are positive event and censoring times, respectively, and that the data structure is $O = (Y, \Delta, A, W)$, where $Y = \min\{T, C\}$, $\Delta = I_{[0,C]}(T)$, $A \in \{0, 1\}$ is a baseline treatment indicator, and W is a vector of baseline covariates. As before, the data consist of independent draws $O_1, O_2, \ldots, O_n$ from $P_0 \in \mathcal{M}$, where $\mathcal{M}$ is a nonparametric model.

For $P \in \mathcal{M}$, the parameter of interest is defined pointwise as

$$\theta_P(t) = E_P \left\{ \prod_{u \leq t} \left[1 - \frac{F_{1,P}(du \mid w)}{S_{Y,P}(u \mid w)} \right] \right\},$$

where $F_{1,P}(t \mid w) = P(Y \leq t, \Delta = 1 \mid A = 1, W = w)$ is the conditional sub-distribution function of Y given $A = 1$ and $W = w$ corresponding to $\Delta = 1$, and $S_{Y,P}(t \mid w) = P(Y \geq t \mid A = 1, W = w)$ is the conditional proportion-at-risk at time t given $A = 1$ and $W = w$. As discussed in Scharfstein and Robins (2002), if T and C are conditionally independent given W and $P(C \geq t \mid A = 1, W = w) \geq \tau$ for P-almost every w, then $\theta_P(t) = E_P\{P(T > t \mid A = 1, W)\}$, the G-computed survival function of T. Furthermore, under additional causal identifiability conditions, this G-computed survival function is equal to the survival function of the counterfactual event time $T(1)$ corresponding to the intervention that sets $A = 1$.

As before, $P \mapsto \theta_P(t)$ is pathwise differentiable in the nonparametric model, and its nonparametric efficient influence function at $P \in \mathcal{M}$ has the form $D^*_{P,t} = D_{P,t} - \theta_P(t)$, where $D_{P,t}$ is given by

$$\frac{a}{g_P(w)} S_P(t \mid w) \left[-\frac{\delta I_{[0,t]}(y)}{S_P(y \mid w) G_P(y- \mid w)} + \int_0^{y \wedge t} \frac{\Lambda_P(du \mid w)}{S_P(u \mid w) G_P(u- \mid w)} \right] + S_P(t \mid w)$$

with $S_P(t \mid w)$ and $G_P(t \mid w)$ the conditional survival functions of T and C, respectively, and $\Lambda_P(t \mid w)$ the conditional cumulative hazard function of T implied by P, all evaluated at t and given $A = 1$ and $W = w$. A simple one-step estimator of $\theta_0(t)$ is given by $\theta_n(t) = \frac{1}{n} \sum_{i=1}^n D_{n,t}(O_i)$, where $D_{n,t}$ is obtained by substituting S_n, G_n, and g_n for S_P, G_P, and g_P respectively, in $D_{P,t}$. Conditions under which such a one-step estimator satisfies (2.3) with $D^*_{0,t} := D^*_{P_0,t}$ are provided in Hubbard et al. (2000).

Condition (A2) is satisfied when the hazard function of T is bounded above and below away from zero, which is very reasonable in many settings. Additionally, it is straightforward to see that, if the conditional hazard $\lambda_0(t \mid w)$ corresponding to $S_0(t \mid w)$ exists and is bounded above by M for Q_0-a.e. w and all $t \in [0, 1]$, and $G_0(t \mid w)$ and $g_0(w)$ are bounded away from zero uniformly in w and t, then (A1.i) and (A1.ii) hold.

2.5 Simulation study

We conducted a simulation study to validate our theoretical results in the context of the first example studied in Section 2.4. For samples sizes $n \in \{100, 250, 500, 750, 1000\}$, we generated 1000 random datasets as follows. We first simulated a bivariate covariate W with independent components W_1 and W_2, respectively distributed as a Bernoulli variate with success probability 0.5 and a uniform variate on $(-1, 1)$. Given $W = (w_1, w_2)$, treatment A was simulated from a logistic model with $P_0(A = 1 \mid W_1 = w_1, W_2 = w_2) = \text{expit}(0.5 + w_1 - 2w_2)$. Given $(W_1, W_2) = (w_1, w_2)$ and $A = a$, Y was simulated as the inverse-logistic transformation of a normal variate with mean $0.5 + a - 4w_2$ and variance 0.3.

For each simulated dataset, we estimated $\theta_0(t)$ for each unique Y_i between 0.1 and 0.9.

To do so, we used the estimator described above, with propensity score and conditional treatment-specific distribution function estimated using correctly specified parametric models. We recorded whether the initial curve had any monotonicity violation as well as the maximal absolute differences between (i) the initial and and projected estimates, (ii) the initial estimate and the truth, and (iii) the projected estimate and the truth. We also recorded the maximal widths of the initial and projected confidence bands.

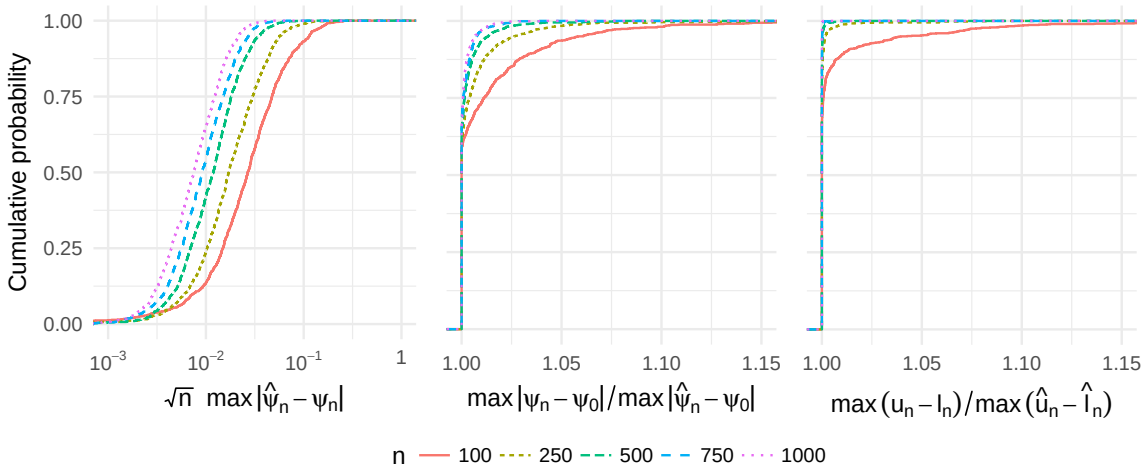

Figure 2.1: G-computed distribution function simulation results. Each plot shows cumulative distributions of a particular quantity over 1000 simulated datasets for each value of n. Left panel: maximal absolute difference between the initial and projected estimators over the grid used for projecting, scaled up by $\sqrt{n}$. Middle panel: ratio of the maximal absolute difference between the initial estimator and the truth and the maximal absolute difference between the projected estimator and the truth. Right panel: ratio of the maximal width of the initial confidence band and the maximal width of the projected confidence band.

On the left exhibit of Figure 2.1, the empirical distribution of the scaled maximum absolute discrepancy between θ_n and θ_n^* is depicted for all sample sizes studied. This plot confirms that the discrepancy between these two estimators indeed decreases faster than $n^{-1/2}$, as our theory suggests. The right exhibit displays for different sample sizes the empirical distribution function of the ratio between the maximum discrepancy between θ_n and θ_0 and that of θ_n^* and θ_0. This plot confirms that θ_n^* is always at least as close to θ_0 as θ_n over $\mathcal{T}_n$. It also

Table 2.1: Coverage of 95% confidence bands for the true counterfactual survival function.

n	100	250	500	750	1000
Initial band	90.6	91.8	94.3	93.3	94.2
Monotone band	90.8	91.8	94.3	93.3	94.2

indicates that even in small samples there is often only a negligible difference between θ_n^* and θ_n. However, in some cases, the maximum discrepancy between θ_n and θ_0 can be up to 15% larger than that between θ_n^* and θ_0. The empirical coverage of uniform 95% influence function-based confidence bands based on θ_n and of bands obtained by isotonic regression are provided in Table 2.1. The coverage of the isotonic band is essentially the same as the initial band.

Chapter 3

A UNIFIED STUDY OF GENERALIZED GRENANDER-TYPE ESTIMATORS OF MONOTONE FUNCTIONS

3.1 Introduction

3.1.1 Background

In this chapter, we address the second general setting introduced in Chapter 1. Specifically, we are primarily motivated by monotone parameters which are not pathwise differentiable relative to the nonparametric model, and hence cannot be estimated nonparametrically at the rate $n^{-1/2}$.

It is a simple fact that the primitive of a non-decreasing function is convex. This observation serves as motivation to consider as an estimator of the function of interest the derivative of the greatest convex minorant (GCM) of an estimator of its primitive function. In the literature on monotone function estimation, many estimators obtained as empirical risk minimizers can alternatively be represented as the left derivative of the GCM of some primitive estimator. This is because the definition of the GCM is intimately tied to the necessary and sufficient conditions for optimization of certain risk functionals over the convex cone of monotone functions (see, e.g., Chapter 2 of Groeneboom and Jongbloed, 2014). In particular, Grenander's NPMLE of a monotone density equals the left derivative of the GCM of the empirical distribution function. In the recent literature, estimators obtained in this fashion have thus been referred to as being of *Grenander-type*. Leurgans (1982) is an early example of a general study of Grenander-type estimators for a class of regression problems.

In a seminal paper, Groeneboom (1985) introduced an approach to studying GCMs based on an inversion operation. This approach has facilitated the theoretical study of certain Grenander-type estimators without the need to utilize their representation as empirical risk

minimizers. For example, under the assumption of independent right-censoring, Huang and Wellner (1995) used this approach to derive large-sample properties of a monotone hazard function estimator obtained by differentiating the GCM of the Nelson-Aalen estimator of the cumulative hazard function. This general strategy was also used by van der Vaart and van der Laan (2006), who derived and studied an estimator of a covariate-marginalized survival curve based on current-status data, including possibly high-dimensional and time-varying covariates. More recently, there has been interest in deriving general results for Grenander-type estimators applicable to a variety of cases. For instance, Durot (2007), Durot (2012) and Lopuhaä and Musta (2016) derived limit results for the estimation error of Grenander-type estimators under L_p, supremum and Hellinger norms, respectively. Durot et al. (2013) studied the problem of testing the equality of generic monotone functions with K independent samples. Durot and Lopuhaä (2014), Beare and Fang (2017) and Lopuhaä and Musta (2018a) studied properties of the least concave majorant of an arbitrary estimator of the primitive function of a monotone parameter. The monograph of Groeneboom and Jongbloed (2014) also summarizes certain large-sample properties for these estimators. Despite the growing body of work on the study of Grenander-type estimators, to the best of our knowledge, general pointwise distributional results are not currently available for these estimators.

3.1.2 Contribution and organization of the chapter

In this chapter, we wish to address the following three key objectives:

1. to provide a unified framework for studying a large class of nonparametric monotone function estimators that implies classical results but also applies in more complicated, modern applications;

2. to derive tractable sufficient conditions under which estimators in this class are known to be consistent and have a non-degenerate limit distribution upon proper centering and scaling;

3. to illustrate the use of this general framework to construct targeted estimators of monotone parameters that are possibly complex summaries of the observed data distribution, and whose estimation may require the use of data-adaptive estimators of nuisance functions.

Our first objective is to introduce a class of monotone estimators that allow the greatest convex minorization process to be performed on a possibly data-dependent transformation of the domain. For many monotone estimators in the literature, the greatest convex minorization is performed on a transformation of the domain. A strategic domain transformation can lead to significant benefits in practice, including in some cases the elimination of the need to estimate challenging nuisance parameters. Unfortunately, to our knowledge, existing results for general Grenander-type estimators do not apply in a straightforward manner in cases in which a data-dependent transformation of the domain has been used. We will define a class that permits such transformations, and demonstrate both how this class encompasses many existing estimators in the literature and how a transformation can be strategically selected in novel problems.

Our second goal is to derive sufficient conditions on the estimator of the primitive function and domain transformation that imply consistency and pointwise convergence in distribution of the monotone function estimator. We pay special attention to parameters for which asymptotically linear estimators of the primitive and transformation functions can be constructed – in such cases, relatively straightforward sufficient conditions can be developed, and the limit distribution has a simpler form. To our knowledge, the broad class of estimators we consider in this chapter has not previously been studied in a unified manner, and hence, general results for this class do not currently exist. In fact, as mentioned above, general results on pointwise convergence in distribution also do not appear to be available for the smaller class of Grenander-type estimators. Because our goal is to provide conditions that can be readily checked, we demonstrate the utility of our general results for three groups of examples – estimation of monotone density, hazard and regression functions – and show that

our results coincide with established results in these settings.

Our third goal is to discuss and illustrate Grenander-type estimation in cases in which nonparametric estimation of the primitive function requires estimation of challenging nuisance parameters. In this sense, our work follows the lead of van der Vaart and van der Laan (2006), whose setting is of this type. More generally, such primitive functions arise frequently, for example, when the observed data unit represents a coarsened version of an ideal data structure, and the coarsening occurs randomly conditional on observed covariates (Heitjan and Rubin, 1991). In our general results, we provide sufficient conditions that can be readily applied to such primitive estimators. To demonstrate the application of our theory in such cases, we consider extensions of the three classical monotone problems above to more complex settings in which covariates must be accounted for, because either the censoring process or the treatment allocation mechanism are informative, as is typical in observational studies. Specifically, we derive novel estimators of monotone density and hazard functions for use when the survival data are subject to right-censoring that may depend on covariates, and a novel estimator of a monotone dose-response curve for use when the relationship between the exposure and outcome is confounded by recorded covariates. Unlike for their classical analogues, in these more difficult problems, nonparametric estimation of the primitive function involves nuisance functions for which flexible estimation strategies (e.g., machine learning) must be employed. As van der Vaart and van der Laan (2006) was able to achieve in a particular problem, our general framework explicitly allows the integration of such strategies while still yielding estimators with a tractable limit theory.

The remainder of the chapter is organized as follows. In Section 3.2, we define the class of estimators we consider and briefly introduce our three working examples. In Section 3.3, we present our most general results for the consistency and convergence in distribution of our class of estimators. We provide refined results, including simpler sufficient conditions and distributional results, for the special case in which the primitive and transformation estimators are asymptotically linear in Section 3.4. In Section 3.5, we apply our general theory in three examples, both for classical parameters and for the novel extensions we

consider. In Section 3.6, we provide results from simulation studies that evaluate the validity of the theory in two examples.

3.2 Generalized Grenander-type estimators

3.2.1 Statistical setup and definitions

Throughout, we make use of the following definitions. For intervals $I, J \subseteq R$, define $\ell^\infty(I)$ as the space of bounded, real-valued functions on I, $\mathcal{D}_I \subset \ell^\infty(I)$ as the subset of non-decreasing and càdlàg (right-continuous with left-hand limits) functions on I, and $\mathcal{D}_{I,J} \subset \mathcal{D}_I$ as the further subset of functions whose range is contained in J. The GCM operator $\mathrm{GCM}_I : \ell^\infty(I) \to \ell^\infty(I)$ is defined for any $G \in \ell^\infty(I)$ as the pointwise supremum over all convex functions $H \leq G$ on I. We note that $\mathrm{GCM}_I(G)$ is necessarily convex. For $G \in \mathcal{D}_I$, we denote by G^- the generalized inverse mapping $x \mapsto \inf\{u \in I : G(u) \geq x\}$, and for a left-differentiable G, we denote by $\partial_- G$ the left derivative of G.

Suppose $O_1, O_2, \ldots, O_n$ are observations sampled independently from an unknown distribution P_0 contained in a nonparametric model $\mathcal{M}$. We denote by O a prototypical data unit, and by $\mathcal{O}(P)$ the support of O under an arbitrary $P \in \mathcal{M}$. We set $\mathcal{O} := \cup_{P \in \mathcal{M}} \mathcal{O}(P)$. We are interested in making inference about an unknown function $\theta_0 \in \mathcal{D}_I$ determined by P_0 for an interval $I \subseteq \mathbb{R}$. We denote the endpoints of I by $a_I := \inf I$ and $b_I := \sup I$. We define the primitive function Θ_0 of θ_0 pointwise for each $x \in I$ as $\Theta_0(x) := \int_{a_I}^x \theta_0(u)du$, where if $a_I = -\infty$ we assume the integral exists.

In its simplest formulation, a Grenander-type estimator of θ_0 is given by $\partial_- \mathrm{GCM}_I(\Theta_n)$ for some estimator Θ_n of Θ_0. However, as a critical step in unifying classical estimators and constructing procedures with possibly improved properties, we wish to allow the GCM procedure to be performed on a possibly data-dependent transformation of the domain I. To do so, we first define for any interval $J \subseteq \mathbb{R}$ the operator $\mathrm{Iso}_J : \ell^\infty(J) \times \mathcal{D}_{I,J} \to \ell^\infty(I)$ as $\mathrm{Iso}_J(\Psi, \Phi) = (\partial_- \mathrm{GCM}_J(\Psi)) \circ \Phi$ for each $\Psi \in \ell^\infty(I)$ and $\Phi \in \mathcal{D}_{I,J}$. Set $J = [0, u_0]$, with $u_0 \in (0, \infty)$ possibly depending on P_0, and suppose that a domain transform $\Phi_0 \in \mathcal{D}_{I,J}$ is

chosen. We may then consider the domain-transformed parameter $\psi_0 := \theta_0 \circ \Phi_0^-$, which has primitive Ψ_0 defined pointwise as $\Psi_0(t) := \int_0^t \psi_0(u)du$ for $t \in (0, u_0]$. As with θ_0 and Θ_0, ψ_0 is non-decreasing and Ψ_0 is convex. Thus, it must be true that $\mathrm{Iso}_{[0,u_0]}(\Psi_0, \Phi_0)(x) = \theta_0(x)$ for each $x \in I$ at which θ_0 is left-continuous and such that $\Phi_0(u) < \Phi_0(x)$ for all $u < x$. This observation motivates us to consider estimators of θ_0 of the form $\mathrm{Iso}_{[0,u_n]}(\Psi_n, \Phi_n)$, where Ψ_n, Φ_n and u_n are estimators of Ψ_0, Φ_0 and u_0, respectively. We refer to any such estimator as being of the *generalized Grenander-type*. This class, of course, contains the standard Grenander-type estimators: setting $\Psi_n = \Theta_n$ and $\Phi_n = \mathrm{Id}$ for Id the identity mapping yields $\theta_n = \partial\_\mathrm{GCM}_I(\Theta_n)$. We note that, in this formulation, we require the domain $[0, u_0]$ over which the GCM is performed to be bounded, but not so for the domain I of θ_0.

Defining $\Gamma_0 := \Psi_0 \circ \Phi_0$, suppose that Γ_n be an estimator of Γ_0. In this work, we study the properties of a generic generalized Grenander-type estimator θ_n of θ_0 of the form

$$\theta_n := \mathrm{Iso}_{[0,u_n]}(\Gamma_n \circ \Phi_n^-, \Phi_n) \ . \tag{3.1}$$

Specifically, our goal is to provide sufficient conditions on the pair (Γ_n, Φ_n) under which θ_n is consistent, and under which a suitable standardization of θ_n converges in distribution to a nondegenerate limit.

We note that estimators taking form (3.1) constitute a more restrictive class than the set of all estimators of the form $\mathrm{Iso}_{[0,u_n]}(\Psi_n, \Phi_n)$ for arbitrary Ψ_n. Our focus on this slightly less general form is motivated by two reasons. First, as we will see in various examples, Γ_0 often has a simpler form than Ψ_0, and in such cases, it may be significantly easier to verify required regularity conditions for Γ_n and to derive limit distribution properties based on Γ_n rather than Ψ_n. Second, many celebrated monotone estimators in the literature follow this particular form. This can be seen by noting that, if Φ_n is a right-continuous step function with jumps at points $x_1, x_2, \ldots, x_m$, then for each $x \in I$ the estimator $\theta_n(x)$ given in (3.1) equals the slope at $\Phi_n(x)$ of the greatest convex minorant of the diagram of points $\{(\Phi_n(x_j), \Gamma_n(x_j)) : j = 0, 1, \ldots, m\}$, where $x_0 = a_I$. We highlight well-known examples of

estimators of this type below. In brief, we sacrifice a little generality for a substantial gain in the ease of application of our results, both for well-known and novel monotone estimators. Nevertheless, conditions on the pair (Ψ_n, Φ_n) under which consistency and distributional results hold for θ_n can be derived similarly.

3.2.2 Examples

Before proceeding to our main results, we briefly discuss the several examples we will use to illustrate how our framework allows us to not only obtain results on classical estimators in the monotone estimation literature directly, but also tackle more complex problems for which no estimators are currently available. These examples will be studied extensively in Section 3.5.

Example 1: monotone density function

Suppose that T is a univariate positive random variable with non-decreasing density function f_0. The simplest scenario consists of estimating this density using data sampled independently from f_0. Thus, in this case, the observed data unit is $O := T$ with distribution P_0 with density function f_0. The parameter of interest is then $\theta_0 := f_0$, the density function of O with support I. Taking Φ_0 to be the identity function, we get that $\psi_0 = \theta_0$. Both Ψ_0 and Θ_0 here represent the distribution function F_0 of T. Taking Ψ_n to be the empirical distribution function, Φ_n the identity map, $\Gamma_n := \Psi_n$ and $u_n := \max_i O_i$, the estimator $\theta_n := \text{Iso}_{[0,u_n]}(\Gamma_n, \Phi_n)$ is precisely the NPMLE of θ_0, that is, Grenander's estimator. Here, the transformation Φ_0 plays no role. If T is right-censored by an independent random censoring time C, the observed data unit is then $O := (Y, \Delta)$ with $Y := \min(T, C)$ and $\Delta := I(T \leq C)$ with distribution P_0 implied by the true marginal distributions of T and C. In this case, θ_0 can be identified from the observed data distribution using the product-limit transform, and a natural estimator θ_n of θ_0 can be obtained by taking Ψ_n to be the Kaplan-Meier estimator of the distribution function Ψ_0. This produces the estimator studied by Huang and Wellner (1995).

In Section 3.5, we will extend estimation of a monotone density function to the setting in which the data are subject to possibly informative right-censoring. Specifically, we will only require T and C to be independent conditionally upon a vector W of baseline covariates. We will study the estimator defined by differentiating the GCM of a one-step estimator of Ψ_0. As we will discuss, estimation of Ψ_0 in this context requires estimation of nuisance functions. We will use our general results to provide conditions on the nuisance estimators that imply consistency and distributional results for θ_n.

Example 2: monotone hazard function

Suppose now that T is a univariate positive random variable with non-decreasing hazard function λ_0. In this example, we are interested in $\theta_0 := \lambda_0$. Setting $S_0 := 1 - F_0$ to be the survival function of T, we note that $\Gamma_0(u) = \int_0^u f_0(v)/S_0(v)\Phi_0(dv)$, and so, taking Φ_0 to satisfy $\Phi_0(dv) = S_0(v)dv$ makes $\Gamma_0 = F_0$. The restricted mean lifetime function $\Phi_0(u) = \int_0^u S_0(v)dv$ satisfies this condition. Using this transformation, the estimator of the monotone hazard function θ_0 only requires estimation of F_0.

In Section 3.5, we again extend estimation of a monotone hazard function to allow the data to be subject to possibly informative right-censoring using the same one-step estimator Γ_n of $\Gamma_0 = F_0$ that will be introduced in Example 1 and the data-dependent transformation $\Phi_n(u) = \int_0^u [1-\Gamma_n(v)]dv$. We will show that, once the simpler details regarding the estimation of a monotone density are established, the asymptotic properties of this estimator of a monotone hazard are obtained essentially for free.

Example 3: monotone regression function

As our last example, we study estimation of a non-decreasing regression function. In the simplest setup, the data unit is $O := (Y, A)$ and we are interested in $\theta_0 : x \mapsto E_0(Y \mid A = x)$. Assume without loss of generality that the data are sorted according to the observed values of A. Taking I to be the support of A and Φ_0 to be the marginal distribution function of A, we have that $\psi_0(u) = E_0[Y \mid \Phi_0(A) = u]$ for each $u \in [0, 1]$, and $\Gamma_0(x) = E_0\left[Y I_{(-\infty,x]}(A)\right]$ for

each $x \in I$. Thus, $\Gamma_n(x) := \frac{1}{n}\sum_{i=1}^n Y_i I_{(-\infty,x]}(A_i)$ and $\Phi_n(x) := \frac{1}{n}\sum_{i=1}^n I_{(-\infty,x]}(A_i)$ are natural nonparametric estimators of $\Gamma_0(x)$ and $\Phi_0(x)$, respectively. Then, $\theta_n := \mathrm{Iso}_{[0,1]}(\Gamma_n, \Phi_n)$ is the classical monotone least-squares estimator of θ_0. Since Φ_n is a step function with jumps at the observed values of A, $\theta_n(x)$ is equal to the left-hand slope of the GCM at $\Phi_n(x)$ of the so-called *cusum diagram* $\{(\Phi_n(A_k), \Gamma_n(A_k)) : k = 0, 1, \ldots, n\} = \{(\frac{k}{n}, \frac{S_k}{n}) : k = 0, 1, \ldots, n\}$, where we let $A_0 = -\infty$, $S_0 = 0$ and $S_k = \sum_{i=1}^k Y_i$ for $k \geq 1$.

In Section 3.5, we will consider an extension to estimation of a covariate-marginalized regression function for use when the relationship between exposure and outcome of interest is confounded. Specifically, we will consider the data unit $O := (Y, A, W)$ with W representing a vector of potential confounders, and focus on $\theta_0 : x \mapsto E_0\left[E_0\left(Y \mid A = x, W\right)\right]$. Under untestable causal identifiability conditions, $\theta_0(x)$ is the mean of the counterfactual outcome $Y(x)$ obtained by setting exposure at level $A = x$. This parameter plays a critical role in causal inference, particularly when the available data are obtained from an observational study and the exposure assignment process may be informative. To tackle this more complex parameter, we will again transform the domain using the marginal distribution function of A, and we will consider a one-step estimator of $\Gamma_0 : x \mapsto E_0\left[Y I_{(-\infty,x]}(A) f_0(A)/g_0(A \mid W)\right]$, where f_0 is the marginal density of A and g_0 the conditional density of A given W.

3.3 General results

We begin with our first set of results on the large-sample properties of θ_n. Our goal is to establish conditions under which consistency and pointwise convergence in distribution hold. First, we provide general results on the consistency of θ_n, both pointwise and uniformly. We note that the results of Durot (2007), Durot (2012) and Lopuhaä and Musta (2016) imply conditions for consistency of general Grenander-type estimators. However, because the objective of their work is to establish distributional theory for a global discrepancy between the estimated and true monotone function, the conditions they require are stronger than needed for consistency alone. Also, their work is restricted to Grenander-type estimators, without data-dependent transformations of the domain.

Below, we refer to the sets $I_n := \{z \in I : z = \Phi_n^-(u), u \in [0, u_n]\}$ and $I_{n,\beta} := \{x \in I : 0 \leq \Phi_0(x-\beta) \leq \Phi_0(x+\beta) \leq u_n\}$ for $\beta \geq 0$.

We begin by stating two lemmas we will require – proofs are provided in the Appendix. The first lemma is a generalization of the switch relation first introduced in Groeneboom (1985) and discussed in detail on page 296 of van der Vaart and Wellner (1996), on page 64 of van der Vaart and van der Laan (2006), in Groeneboom and Jongbloed (2014) and in Balabdaoui et al. (2011). For brevity, throughout, we will refer to van der Vaart and Wellner (1996) as VW.

Lemma 4. *Let Φ and Γ be functions from a closed interval $I \subseteq \mathbb{R}$ to $[a,b] \subset \mathbb{R}$, where Φ is nondecreasing and càdlàg , Γ and $\Psi := \Gamma \circ \Phi^-$ are lower semi-continuous, and $\{a,b\} \subset \Phi(I)$. Let ψ be the left derivative of the GCM $\bar{\Psi}$ of Ψ and $\theta := \psi \circ \Phi$. Then, for any $c \in \mathbb{R}$ and $x \in I$ with $\Phi(x) \in (a,b)$, $\theta(x) > c$ if and only if $\sup \operatorname{argmax}_{v \in I^*} \{c\Phi(v) - \Gamma(v)\} < \Phi^-(\Phi(x))$, where $I^* := I \cap \Phi^-([a,b]) = \{x \in I : x = \Phi^-(u), u \in [a,b]\}$.*

The switch relation requires Ψ to be lower semi-continuous. If Ψ_n is not so, it can be replaced by its greatest lower semi-continuous minorant. As argued in van der Vaart and van der Laan (2006), this only possibly changes the GCM at the endpoints of the interval and has no effect on asymptotic properties (e.g., weak convergence of Ψ_n). In the second lemma, pointwise and uniform finite-sample tail bounds are provided. These tail bounds are not sharp but suffice to derive consistency results in broad generality. Simpler tail bounds can be derived in the absence of a transformation Φ_0.

Lemma 5. *Suppose that $|\Phi_0^-(u) - x| \leq \gamma(|u - \Phi_0(x)|)$ for all $u \in [0, u_0]$ and a continuous, strictly increasing function $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ with $\gamma(0) = 0$, and that Φ_0 is strictly increasing and continuous on $[x-\delta, x+\delta] \subset \Phi_0^{-1}([0,u_n])$. Let $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ be a non-decreasing function satisfying $\lim_{z \downarrow 0} \omega(z) = \omega(0) = 0$, and suppose $|\theta_0(u) - \theta_0(x)| \leq \omega(|u-x)|)$. Define $c(\delta,\eta) := \gamma^{-1}(\delta \wedge \omega^-(\eta))$ and $r(\delta,\eta) := \int_0^{c(\delta,\eta)/2} [\eta - \omega(\gamma(u))]\, du$. Then, for any $\eta > 0$ and*

$x \in I$ such that $\Phi_n(x), \Phi_0(x) \in (0, u_n)$,

$$P_0\left(|\theta_n(x) - \theta_0(x)| > \eta\right) \leq P_0\left(A_{n,1}(\eta) > r(\delta/2, \eta/2)\right) + P_0\left(A_{n,2} \geq c(\delta/2, \eta/2)\right)$$

with $A_{n,1}(\eta) := 2\|\Gamma_n - \Gamma_0\|_{\infty,I_n} + (2|\theta_0(x)| + \eta)\|\Phi_n - \Phi_0\|_{\infty,I_n}$ and $A_{n,2} := 4\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]}$. If Φ_0 is strictly increasing and continuous on I, $|\Phi_0^{-1}(u) - \Phi_0^{-1}(v)| \leq \gamma(|u - v|)$ for all $u, v \in [0, u_0]$, and $|\theta_0(u) - \theta_0(v)| \leq \omega(|u - v|)$ for all $u, v \in I$, then, for any $\eta, \beta > 0$,

$$P_0\left(\|\theta_n - \theta_0\|_{\infty,I_{n,\beta}} > \eta\right) \leq P_0\left(B_{n,1}(\eta) > r(\beta/2, \eta/2)\right) + P_0\left(B_{n,2} \geq c(\beta/2, \eta/2)\right)$$

with $B_{n,1}(\eta) := 2\|\Gamma_n - \Gamma_0\|_{\infty,I_n} + (2\|\theta_0\|_{\infty,I} + \eta)\|\Phi_n - \Phi_0\|_{\infty,I_n}$ and $B_{n,2} := 4\|\Phi_n - \Phi_0\|_{\infty,I}$.

Theorem 4 (Weak consistency)**.** *(1) Suppose θ_0 is continuous at $x \in I$ and, for some $\delta > 0$ such that $[x-\delta, x+\delta] \subset \Phi_0^{-1}(J_0)$, Φ_0 is strictly increasing and continuous on $[x-\delta, x+\delta]$. If $\|\Gamma_n - \Gamma_0\|_{\infty,I_n}$, $\|\Phi_n - \Phi_0\|_{\infty,I_n}$ and $\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]}$ tend to zero in probability, then $\theta_n(x) \xrightarrow{\mathrm{P}} \theta_0(x)$.*

(2) Suppose θ_0 and Φ_0 are uniformly continuous on I, and Φ_0 is strictly increasing on I. If $\|\Gamma_n - \Gamma_0\|_{\infty,I_n}$ and $\|\Phi_n - \Phi_0\|_{\infty,I}$ tend to zero in probability, then $\|\theta_n - \theta_0\|_{\infty,I_{n,\beta}} \xrightarrow{\mathrm{P}} 0$ for each fixed $\beta > 0$.

We note that in part 1 of Theorem 4, we require uniform convergence of Γ_n and Φ_n to obtain a pointwise result for θ_n – this will also be the case for Theorem 5 below. This is because the GCM is a global procedure, and so, the value of $\theta_n(x_1)$ depends on $\Gamma_n(x_2)$ even for x_2 not near x_1. Without uniform consistency of Γ_n, θ_n may indeed fail to be pointwise consistent. Also, we note that in part 1 of Theorem 4, we require that $\Gamma_n - \Gamma_0$ and $\Phi_n - \Phi_0$ tend to zero uniformly over the set I_n. This requirement stems from the fact that θ_n only depends on Γ_n through the composition $\Gamma_n \circ \Phi_n^-$, and so, values of Γ_n only matter at points in the range of Φ_n^-. In part 1, we also require that $\Phi_n - \Phi_0$ tend to zero uniformly in a neighborhood of x, while in part 2, we require that $\Phi_n - \Phi_0$ tend to zero uniformly over I.

These requirements allow us to obtain results for x values that are possibly outside I_n for all n. In many applications, it may be the case that $\Gamma_n - \Gamma_0$ and $\Phi_n - \Phi_0$ both tend to zero in probability uniformly over I, which implies convergence over I_n.

The weak conditions required for Theorem 4 are especially important for the extensions of the classical parameters that we consider in Section 3.5. The estimators we propose often require estimating difficult nuisance parameters, such as conditional hazard, density and mean functions. While under mild conditions it is typically possible to construct uniformly consistent estimators of these nuisance parameters, ensuring a given local or uniform rate of convergence often requires additional knowledge about the true function. Thus, Theorem 4 is useful for guaranteeing consistency under weak conditions.

We now provide lower bounds on the convergence rate of θ_n, both pointwise and uniformly, depending on (a) the uniform rates of convergence of Γ_n and Φ_n, and (b) the moduli of continuity of θ_0 and Φ_0^-.

Theorem 5 (Rates of convergence)**.** *Let $x \in I$ be given. Suppose that, for some $\delta > 0$, $[x-\delta, x+\delta] \subset \Phi_0^{-1}([0,u_0])$ and Φ_0 is strictly increasing and continuous on $[x-\delta, x+\delta]$. Let r_n be a fixed sequence such that $r_n\|\Gamma_n - \Gamma_0\|_{\infty,I_n}$, $r_n\|\Phi_n - \Phi_0\|_{\infty,I_n}$ and $r_n\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]}$ are bounded in probability.*

(1) If there exist $K_1(x), K_2(x) \in [0,\infty)$ and $\alpha_1, \alpha_2 \in (0,1]$ such that $|\theta_0(u) - \theta_0(x)| \leq K_1(x)|u-x|^{\alpha_1}$ for all $u \in I$ and $|\Phi_0^-(u) - \Phi^-(x)| \leq K_2(x)|u-x|^{\alpha_2}$ for all $u \in [0,u_0]$, then

$$r_n^{\frac{\alpha_1\alpha_2}{1+\alpha_1\alpha_2}}\left[\theta_n(x) - \theta_0(x)\right] = \mathrm{O_P}(1)\ .$$

(2) If θ_0 is constant on $[x-\delta, x+\delta]$, then $r_n\left[\theta_n(x) - \theta_0(x)\right] = \mathrm{O_P}(1)$.

Let r_n be a fixed sequence such that $r_n\|\Gamma_n - \Gamma_0\|_{\infty,I_n}$ and $r_n\|\Phi_n - \Phi_0\|_{\infty,I}$ are bounded in probability, and suppose that Φ_0 is strictly increasing on I.

(3) If there exist $K_1, K_2 \in [0,\infty)$ and $\alpha_1, \alpha_2 \in (0,1]$ such that $|\theta_0(u) - \theta_0(v)| \leq K_1|u-v|^{\alpha_1}$

for all $u, v \in I$ *and* $|\Phi_0^-(u) - \Phi_0^-(v)| \leq K_2|u-v|^{\alpha_2}$ *for all* $u, v \in [0, u_0]$*, then*

$$r_n^{\frac{\alpha_1\alpha_2}{1+\alpha_1\alpha_2}} \|\theta_n - \theta_0\|_{\infty, I_{n,\beta_n}} = \mathrm{O_P}(1)$$

for any (possibly random) positive real sequence β_n *such that* $\beta_n r_n^{1/(1+\alpha_1\alpha_2)} \xrightarrow{\mathrm{P}} \infty$*.*

We note here that the uniform results only cover subintervals of the interval over which the GCM procedure is performed. This should not be surprising given the poor behavior of Grenander-type estimators at the boundary of the GCM interval, as discussed, for example, in Woodroofe and Sun (1993), Kulikov and Lopuhaä (2006) and Balabdaoui et al. (2011). Various boundary corrections have been proposed – applying these in our general framework is an interesting avenue for future work.

We also note that, in Theorem 5, when θ_0 and Φ_0 are locally or globally Lipschitz, then $\alpha_1 = \alpha_2 = 1$ and the resulting rate is $\mathrm{O_P}(r_n^{-1/2})$, which yields $\mathrm{O_P}(n^{-1/4})$ when $r_n = n^{1/2}$. This rate is slower than the rate $n^{-1/3}$ that is often achievable for pointwise convergence when θ_0 and Φ_0 are differentiable at x and the primitive estimator converges at rate $n^{-1/2}$, as we discuss below. However, the assumptions in Theorems 5 are significantly weaker than typically required for the $n^{-1/3}$ rate of convergence: they constrain the supremum norm of the estimation error rather than its modulus of continuity, and hold when the true function is Lipschitz but not differentiable. Our results also cover situations in which θ_0 or Φ_0 are in Hölder classes. The rates provided by Theorem 5 should thus be seen as lower bounds on the true rate, for use when less is known about the properties of the estimation error or of the true functions. The distributional results we provide below recover the usual rates under stronger conditions.

For a fixed sequence r_n of positive real numbers, we now study the pointwise convergence in distribution of $r_n[\theta_n(x) - \theta_0(x)]$ at a fixed interior point $x \in I$ at which both θ_0 and Φ_0 have a strictly positive derivative. Writing $\Gamma_{n,0} := \Gamma_n - \Gamma_0$ and $\Phi_{n,0} := \Phi_n - \Phi_0$, we define

the localized process

$$W_{n,x} : u \mapsto r_n^2 \left\{ \Gamma_{n,0}(x + ur_n^{-1}) - \Gamma_{n,0}(x) - \theta_0(x) \left[\Phi_{n,0}(x + ur_n^{-1}) - \Phi_{n,0}(x) \right] \right\}$$

and introduce the following conditions:

(B1) for each $M > 0$, $\{W_{n,x}(u) : |u| \leq M\}$ converges weakly in $\ell^\infty[-M, M]$ to a tight limit process $\{W_x(u) : |u| \leq M\}$ with almost surely upper semi-continuous sample paths and stationary increments;

(B2) $\left\{ W_x(u) + \frac{1}{2}\theta_0'(x)\Phi_0'(x)u^2 : u \in \mathbb{R} \right\}$ almost surely possesses a unique and finite point of minimum;

(B3) there exist $\alpha \in (1, 2)$, $\delta^* > 0$, and a sequence $f_n : \mathbb{R}^+ \to \mathbb{R}^+$ such that $u \mapsto u^{-\alpha} f_n(r_n u)$ is decreasing, $f_n(1) = O(1)$, and $E_0 \left[\sup_{|u| \leq r_n \delta} |W_{n,x}(u)| \right] \leq f_n(r_n \delta)$ for all large n and $\delta \leq \delta^*$.

In addition, we introduce conditions on the uniform convergence of estimators Φ_n and Γ_n:

(B4) $r_n E_0 \left[\sup_{|v| < \delta} |\Phi_n(x + v) - \Phi_0(x + v)| \right] \longrightarrow 0$ for some $\delta > 0$;

(B5) $\|\Gamma_{n,0} - \theta_0(x) \cdot \Phi_{n,0}\|_{\infty, I_n} \xrightarrow{\mathrm{P}} 0$.

These conditions suffice to apply the argmin continuous mapping theorem to the limit process.

Theorem 6 (Convergence in distribution)**.** *If x is a fixed interior point of I at which θ_0 is differentiable and Φ_0 is continuously differentiable, both with positive derivative, conditions (B1)–(B5) imply that*

$$r_n \left[\theta_n(x) - \theta_0(x) \right] \xrightarrow{\mathrm{d}} -\theta_0'(x) \operatorname*{argmin}_{u \in \mathbb{R}} \left\{ W_x(u) + \tfrac{1}{2}\theta_0'(x)\Phi_0'(x)u^2 \right\}.$$

If $W_x = [\kappa_0(x)]^{1/2} W_0$ *with* W_0 *a standard two-sided Brownian motion process satisfying* $W_0(0) = 0$, *then* $r_n [\theta_n(x) - \theta_0(x)] \xrightarrow{\mathrm{d}} \tau_0(x) Z$ *with* $\tau_0(x) := [4\theta_0'(x)\kappa_0(x)/\Phi_0'(x)^2]^{1/3}$ *and* $Z := \operatorname{argmin}_{u \in \mathbb{R}} \{W_0(u) + u^2\}$.

The latter limit distribution is referred to as a scaled Chernoff distribution, since Z is said to follow the standard Chernoff distribution. This distribution appears prominently in classical results in nonparametric monotone function estimation and has been extensively studied (e.g., Groeneboom and Wellner, 2001). It can also be defined as the distribution of the slope at zero of $\mathrm{GCM}_{\mathbb{R}}(W(x) + x^2)$, where W is a standard Brownian motion.

Suppose W_x^0 is the limit process that arises when no domain transformation is used in the construction of a generalized Grenander-type estimator, that is, when both Φ_0 and Φ_n are taken to be the identity map. In this case, under (B1)–(B5), Theorem 6 indicates that

$$r_n [\theta_n(x) - \theta_0(x)] \xrightarrow{\mathrm{d}} -\theta_0'(x) \operatorname*{argmin}_{u \in \mathbb{R}} \left\{ W_x^0(u) + \tfrac{1}{2}\theta_0'(x) u^2 \right\}.$$

It is natural to ask how this limit distribution compares to the one obtained using a non-trivial transformation Φ_0. In particular, does using Φ_0 change the pointwise distributional results for θ_n? The answer is of course negative whenever W_x and $\Phi_0'(x) W_x^0$ are equal in distribution, since the multiplicative factor $\Phi_0'(x)$ then does not change the value of the minimizer defining the limit distribution. A more detailed discussion of this question and lower-level conditions are provided in the next section.

3.4 Refined results for asymptotically linear primitive and transformation estimators

3.4.1 Distributional results

So far, we have not assumed any particular form for estimators Γ_n and Φ_n, and our results can be applied broadly. However, it is common for these estimators to be linear or asymptotically linear, and in such cases, simpler, lower-level conditions for consistency of θ_n and convergence in distribution of $r_n[\theta_n(x) - \theta_0(x)]$ can be derived. Below, we write Pf to denote $\int f(o) dP(o)$

for any probability measure P and P-integrable function $f : \mathcal{O} \to \mathbb{R}$. We also use $\mathbb{P}_n$ to denote the empirical distribution of independent observations $O_1, O_2, \ldots, O_n$ from P_0 so that $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(O_i)$ for any $f : \mathcal{O} \to \mathbb{R}$.

Suppose that, as estimators of Γ_0 and G_0, respectively, Γ_n and Φ_n are uniformly asymptotically linear over I, in the sense that there exist functions $D^*_{x,0} : \mathcal{O} \to \mathbb{R}$ and $L^*_{x,0} : \mathcal{O} \to \mathbb{R}$ depending on P_0 such that, for each $x \in I$, $P_0 D^*_{x,0} = P_0 L^*_{x,0} = 0$ and both $P_0 D^{*2}_{x,0}$ and $P_0 L^{*2}_{x,0}$ are finite, and

$$\Gamma_n(x) - \Gamma_0(x) = \mathbb{P}_n D^*_{x,0} + H_{x,n} \quad \text{and} \quad \Phi_n(x) - \Phi_0(x) = \mathbb{P}_n L^*_{x,0} + R_{x,n} \ , \tag{3.2}$$

where $H_{x,n}$ and $R_{x,n}$ are stochastic remainder terms. Objects $D^*_{x,0}$ and $L^*_{x,0}$ are referred to as the influence functions of $\Gamma_n(x)$ and $\Phi_n(x)$, respectively, under sampling from P_0.

Assessing consistency and uniform consistency of θ_n is straightforward in this important case. For example, if the classes $\{D^*_{x,0} : x \in I\}$ and $\{L^*_{x,0} : x \in I\}$ are P_0-Donsker classes, and the remainder terms satisfy $n^{1/2}\sup_{x\in I}|H_{x,n}|$ and $n^{1/2}\sup_{x\in I}|R_{x,n}|$ tend to zero in probability, then $n^{1/2}\|\Gamma_n - \Gamma_0\|_{\infty,I}$ and $n^{1/2}\|\Phi_n - \Phi_0\|_{\infty,I}$ are both bounded in probability, and Theorems 4 and 5 can be directly applied with $r_n = n^{1/2}$ provided the required conditions on θ_0 and Φ_0 hold. As such, we focus here on deriving a refined version of Theorem 6 for use whenever Γ_n and Φ_n are uniformly asymptotically linear estimators.

It is reasonable to expect the linear terms $\mathbb{P}_n D^*_{x,0}$ and $\mathbb{P}_n L^*_{x,0}$ to drive the behavior of the standardized difference $r_n[\theta_n(x) - \theta_0(x)]$ in Theorem 6. The natural rate here is $r_n = n^{1/3}$, for which Kim and Pollard (1990) provide intuition. Our first goal in this section is to provide sufficient conditions for weak convergence of the process $\{n^{1/6}\mathbb{G}_n g_{x,n^{-1/3}u} : |u| \le M\}$, where $\mathbb{G}_n$ is the empirical process $n^{1/2}(\mathbb{P}_n - P_0)$ and we define the localized difference function $g_{x,v} := D^*_{x+v,0} - D^*_{x,0} - \theta_0(x)(L^*_{x+v,0} - L^*_{x,0})$. Kim and Pollard (1990) also provide detailed conditions for weak convergence of processes of this type. Building upon their results, we are able to provide simplified sufficient conditions for convergence in distribution of $n^{1/3}[\theta_n(x) - \theta_0(x)]$ when Γ_n and Φ_n are uniformly asymptotically linear estimators.

We begin by introducing conditions we will refer to. First, we define $\mathcal{G}_{x,R} := \{g_{x,u} : |u| \leq R\}$ and suppose that $\mathcal{G}_R$ has envelope function $G_{x,R}$. The first two conditions concern the size of $\mathcal{G}_{x,R}$ for small R in terms of bracketing or uniform entropy numbers, which for completeness we define here – see van der Vaart and Wellner (1996) for a comprehensive treatment. Denote by $\|G\|_{P,2} = [P(G^2)]^{1/2}$ the $L_2(P)$ norm of a given P-square-integrable function $G : \mathcal{O}(P) \to \mathbb{R}$. The bracketing number $N_{[]}(\varepsilon, \mathcal{G}, L_2(P))$ of a class $\mathcal{G}$ with respect to the $L_2(P)$ norm is the smallest number of ε-brackets needed to cover $\mathcal{G}$, where an ε-bracket is any set of functions $\{f : \ell \leq f \leq u\}$ with ℓ and u such that $\|\ell - u\|_{P,2} < \varepsilon$. The covering number $N(\varepsilon, \mathcal{G}, L_2(Q))$ of $\mathcal{G}$ with respect to the $L_2(Q)$ norm is the smallest number of ε-balls in $L_2(Q)$ required to cover $\mathcal{G}$. The uniform covering number is the supremum of $N(\varepsilon\|G\|_{2,Q}, \mathcal{G}, L_2(Q))$ over all discrete probability measures Q such that $\|G\|_{2,Q} > 0$, where G is an envelope function for $\mathcal{G}$. We consider conditions on the size of $\mathcal{G}_{x,R}$:

(C1) for some constants $C > 0$ and $0 \leq V < 2$, either (C1a) $\log N_{[]}(\varepsilon\|G_{x,R}\|_{P_0,2}, \mathcal{G}_{x,R}, L_2(P_0)) \leq C\varepsilon^{-V}$ or (C1b) $\log \sup_Q N(\varepsilon\|G_{x,R}\|_{Q,2}, \mathcal{G}_{x,R}, L_2(Q)) \leq C\varepsilon^{-V}$ for all $\varepsilon \in (0, 1]$ and R small enough;

(C2) $P_0 G_{x,R}^2 = O(R)$, and for all $\eta > 0$, $P_0 G_{x,R}^2\{RG_{x,R} > \eta\} = o(R)$, as $R \to 0$.

Condition (C1) replaces the notion of *uniform manageability* of the class $\mathcal{G}_{x,R}$ for small R as defined in Kim and Pollard (1990), whereas condition (C2) directly corresponds to their condition (vi). Since bounds on the bracketing and uniform entropy numbers have been derived for many common classes of functions, condition (C1) can be readily checked in practice. Together, conditions (C1) and (C2) ensure that $\mathcal{G}_{x,R}$ is a relatively small class, and this helps to establish the weak convergence of the localized process $\{W_{n,x}(u) : |u| \leq M\}$.

As in Kim and Pollard (1990), to guarantee that the covariance function of this localized process stabilizes, it suffices that $\delta^{-1} \sup_{|u-v|<\delta} P_0(g_{x,u} - g_{x,v})^2$ be bounded for small enough $\delta > 0$ and that, up to a scaling factor possibly depending on x, $\sigma_{x,\alpha}(u, v) := \alpha^{-1} P_0[(g_{x,\alpha u} - P_0 g_{x,\alpha u})(g_{x,\alpha v} - P_0 g_{x,\alpha v})]$ tend to the covariance function $\sigma^2(u, v)$ of a two-sided Brownian

motion as $\alpha \to 0$. Below, we provide simple conditions that imply these two statements for a broad class of settings that includes our examples.

The covariance function of the Gaussian process to which $\{\mathbb{G}_n[D^*_{t,0} - \theta_0(x)L^*_{t,0}] : t\}$ converges weakly is defined pointwise as $\Sigma_0(s,t) := P_0[D^*_{s,0} - \theta_0(x)L^*_{s,0}][D^*_{t,0} - \theta_0(x)L^*_{t,0}]$. The behavior of Σ_0 near (x,x) dictates the covariance of the local limit process W_x and hence the scale parameter $\kappa_0(x)$. If Σ_0 is differentiable in (s,t) at (x,x), it follows that $\kappa_0(x) = 0$ and θ_n converges at a faster rate, although possibly with an asymptotic bias. When instead scaled Chernoff asymptotics apply, the covariance function can typically be written as

$$\Sigma_0(s,t) = \Sigma^*_0(s,t) + \iint_{-\infty}^{s\wedge t} A_0(s,t,v,w)H_0(dv,w)Q_0(dw) \tag{3.3}$$

for some functions $\Sigma^*_0 : I \times I \to \mathbb{R}$, $A_0 : I \times I \times I \times \mathcal{W} \to \mathbb{R}$ and $H_0 : I \times \mathcal{W} \to \mathbb{R}$ depending on P_0, where Q_0 is a probability measure induced by P_0 on some measurable space $\mathcal{W}$. In this representation, Σ^*_0 is taken to be the differentiable portion of the covariance function, which does not contribute to the scale parameter. The second summand is not differentiable at (x,x) and makes $\sigma_{x,\alpha}(u,v)$ tend to a non-zero limit. We consider cases in which Σ^*_0, A_0 and H_0 satisfy the following conditions:

(C3) Representation (3.3) holds, and for some $\delta > 0$, setting $B_\delta(x) := (x-\delta, x+\delta)$, it is also true that:

(C3a) Σ^*_0 is symmetric in its arguments and continuously differentiable on $B_\delta(x)$;

(C3b) A_0 is symmetric in its first two arguments, and $s \mapsto A_0(s,t,v,w)$ is differentiable for Q_0-almost every w and each $s,t,v \in B_\delta(x)$, with derivative $A'_0(s,t,v,w)$ continuous in s,t,v each in $B_\delta(x)$ for Q_0-almost every w and satisfying the boundedness condition

$$\iint_{-\infty}^{x+\delta} \sup_{s,t\in B_\delta(x)} |A'_0(s,t,v,w)| H_0(dv,w)Q_0(dw) < \infty \ ;$$

(C3c) $v \mapsto A_0(x, x, v, w)$ is continuous at $v = x$ uniformly in w over the support of Q_0;

(C3d) $v \mapsto H_0(v, w)$ is nondecreasing for all w and differentiable at each $v \in B_\delta(x)$, with derivative $H_0'(v, w)$ continuous at $v = x$ uniformly in w over the support of Q_0.

Representation (3.3) is deliberately broad to encompass a wide variety of parameters. Nevertheless, in many settings, the covariance function can be considerably simplified, leading then to simpler conditions in (C3). For instance, when W is a vector of covariates over which marginalization is performed to compute the parameter, Q_0 typically plays the role of the marginal distribution of W under P_0. In classical problems in which there is no adjustment for covariates, this feature of representation (3.3) is not needed and indeed vanishes. In other settings, $A_0(s, t, v, w)$ depends on v and w but not on s and t.

Finally, we must ensure that the stochastic remainder terms $H_{x,n}$ and $R_{x,n}$ arising the asymptotic linear representations of Γ_n and Φ_n do not contribute to the limit distribution. Defining $\tilde{H}_{u,n} := H_{x+u,n} - H_{x,n}$, $\tilde{R}_{u,n} := R_{x+u,n} - R_{x,n}$ and $K_n(\delta) := n^{2/3} \sup_{|u| \leq \delta n^{-1/3}} |\tilde{H}_{u,n} - \theta_0(x)\tilde{R}_{u,n}|$, we consider the following conditions for the asymptotic negligibility of these remainder terms:

(C4) $K_n(\delta) \xrightarrow{\mathrm{P}} 0$ for each fixed $\delta > 0$;

(C5) for some $\alpha \in (1, 2)$, $\delta \mapsto \delta^{-\alpha} E_0 [K_n(\delta)]$ is decreasing for all δ small enough and n large enough.

Condition (C4) guarantees that the remainder terms do not contribute to the weak convergence of $\{W_{n,x}(u) : |u| \leq M\}$, and condition (C5) guarantees that the remainder terms satisfy condition (B3).

Combining the conditions above, we can state the following master theorem for pointwise convergence in distribution when the monotone estimator is based upon asymptotically linear primitive estimators:

Theorem 7. *Suppose that, at an interior point $x \in I$, θ_0 is differentiable and Φ_0 is continuously differentiable with positive derivative. Suppose also that Γ_n and Φ_n are asymptotically linear estimators of Γ_0 and Φ_0, respectively, uniformly over I, such that conditions (C1)–(C5) and (B4)–(B5) hold. Then*

$$r_n \left[\theta_n(x) - \theta_0(x)\right] \xrightarrow{\mathrm{d}} \tau_0(x) Z \ ,$$

where Z follows the standard Chernoff distribution, and $\tau_0(x) := \left[4\theta_0'(x)\kappa_0(x)/\Phi_0'(x)^2\right]^{1/3}$ is a scale factor involving $\kappa_0(x) := \int A_0(x,x,x,w) H_0'(x,w) Q_0(dw)$.

3.4.2 Effect of domain transform on limit distribution

As was done briefly after Theorem 6, it is natural to compare the limit distribution obtained by Theorem 7 when a transformation of the domain is used and when it is not. We will consider $\theta_n := \mathrm{Iso}_{[0,u_n]}(\Theta_n, \mathrm{Id})$, the estimator obtained by directly isotonizing an estimator Θ_n of the primitive function Θ_0 without use of a domain transformation. Denoting by Φ_0 a candidate non-decreasing transformation function, and letting $\Gamma_0 := \Psi_0 \circ \Phi_0$ be as described in Section 3.2, we will also consider $\theta_n^* := \mathrm{Iso}_{[0,u_n^*]}(\Gamma_n \circ \Phi_n^-, \Phi_n)$, where Γ_n and Φ_n are estimators of Γ_0 and Φ_0, respectively. Suppose $\Theta_n(x), \Gamma_n(x)$ and $\Phi_n(x)$ are each asymptotically linear estimators of their respective targets with influence functions $M_{x,0}^*$, $D_{x,0}^*$ and $L_{x,0}^*$, respectively, under sampling from P_0.

We wish to compare the scale parameters $\kappa_0(x)$ and $\kappa_0^*(x)$ arising from the use of the distinct estimators $\theta_n(x)$ and $\theta_n^*(x)$. To do so, we can use expression (C3) to examine the covariance obtained in both cases. However, it appears difficult to say much without having more specific forms for the involved influence functions. Unfortunately, it also appears difficult to characterize these influence functions generally since they depend inherently on the parameter of interest θ_0, and we wish to remain agnostic to the form of θ_0. Nevertheless, in our next result, we describe a class of problems, characterized by the generated influence functions and regularity conditions on these, in which domain transformation has no effect

on the limit distribution of the generalized Grenander-type estimator.

Theorem 8. *Suppose conditions (C1)–(C5) hold for* (Θ_n, Id) *and* (Γ_n, Φ_n)*, and the observed data unit can be partitioned as* $O = (U, Z)$ *with* $U \in \mathbb{R}^+$*. Suppose that the influence functions can be expressed as*

$$\begin{aligned}
M^*_{x,0} &: (u,z) \mapsto I_{[0,x]}(u) M^{(1)}_{x,0}(u,z) + M^{(2)}_{x,0}(u,z) \ , \\
L^*_{x,0} &: (u,z) \mapsto I_{[0,x]}(u) L^{(1)}_{x,0}(u,z) + L^{(2)}_{x,0}(u,z) \ , \\
D^*_{x,0} &: (u,z) \mapsto I_{[0,x]}(u) \Phi_0'(u) M^{(1)}_{x,0}(u,z) + D^{(2)}_{x,0}(u,z) + \int_0^x \theta_0(v) L^*_{dv,0}(u,z) \ ,
\end{aligned}$$

and satisfy the smoothness conditions stated in the Appendix. Suppose that the density function h_0 *of the conditional distribution of* U *given* Z *exists and is continuous in a neighborhood of* x *uniformly over the support of the marginal distribution* $Q_{Z,0}$ *of* Z*. Then, it follows that*

$$\begin{aligned}
\kappa_0(x) &= \int \left[M^{(1)}_{x,0}(x,z) \right]^2 h_0(x \mid z) Q_{Z,0}(dz) \quad \textit{and} \\
\kappa^*_0(x) &= \left[\Phi_0'(x) \right]^2 \int \left[M^{(1)}_{x,0}(x,z) \right]^2 h_0(x \mid z) Q_{Z,0}(dz) \ .
\end{aligned}$$

Consequently, $n^{1/3} \left[\theta_n(x) - \theta_0(x) \right]$ *and* $n^{1/3} \left[\theta^*_n(x) - \theta_0(x) \right]$ *have the same limit distribution.*

In all the examples we study in Section 3.5, the conditions of Theorem 8 apply. The forms of $M^*_{x,0}$ and $L^*_{x,0}$ arise naturally in a wide variety of settings because the parameters considered involve a primitive function. The supposed form of $D^*_{x,0}$ may seem restrictive at first glance but is in fact expected given the forms of $M^*_{x,0}$ and $L^*_{x,0}$. We now provide a heuristic justification for the forms of these influence functions.

In a nonparametric model, the efficient influence function $M^*_{x,0}$ of $\Theta_0(x)$ is the unique element of $L^2_2(P_0)$ such that, for each $s_0 \in L^0_2(P_0)$,

$$\left. \frac{\partial}{\partial \varepsilon} \Theta_{P_\varepsilon}(x) \right|_{\varepsilon = 0} = \int_0^x \left. \frac{\partial}{\partial \varepsilon} \theta_{P_\varepsilon}(u) \right|_{\varepsilon = 0} du = P_0(M^*_{x,0} s_0)$$

for each regular, one-dimensional parametric path $\{P_\varepsilon\} \subset \mathcal{M}$ through P_0 at $\varepsilon = 0$ and with

score s_0 at $\varepsilon = 0$. In the presence of a transformation depending on P_0, the efficient influence function $L^*_{x,0}$ of $\Phi_0(x)$ similarly satisfies $\frac{\partial}{\partial\varepsilon}\Phi_{P_\varepsilon}(x)\big|_{\varepsilon=0} = P_0(L^*_{x,0}s_0)$. The nonparametric efficient influence function $D^*_{x,0}$ of $\Gamma_0(x)$ thus satisfies

$$P_0(D^*_{x,0}s_0) = \frac{\partial}{\partial\varepsilon}\Gamma_{P_\varepsilon}(x)\bigg|_{\varepsilon=0} = \int_0^x \frac{\partial}{\partial\varepsilon}\theta_{P_\varepsilon}(v)\bigg|_{\varepsilon=0}\Phi_0(dv) + \int_0^x \theta_0(v)\,\frac{\partial}{\partial\varepsilon}\Phi_{P_\varepsilon}(dv)\bigg|_{\varepsilon=0}.$$

The first term typically contributes $o \mapsto I_{[0,x]}(u)M^{(1)}_{x,0}(o)\Phi_0'(u) + D^{(2)}_{x,0}(o)$ to the form of $D^*_{x,0}$. $M^{(1)}_{x,0}$ in this term is deliberately the same as in the influence function of $\Theta_0(x)$. The second term is equal to

$$\int_0^x \theta_0(v)P_0\left(L^*_{dv,0}s_0\right) = P_0\left[s_0\int_0^x \theta_0(v)L^*_{dv,0}\right],$$

so that this term contributes $o \mapsto \int_0^x \theta_0(v)L^*_{dv,0}(o)$ to the form of $D^*_{x,0}$. Hence, we get the general form

$$D^*_{x,0}(o) = I_{[0,x]}(u)D^{(1)}_{s,0}(o)\Phi_0'(u) + D^{(2)}_{x,0}(o) + \int_0^x \theta_0(v)L^*_{dv,0}(o)\ .$$

3.4.3 Negligibility of remainder terms

In some applications, the estimators Γ_n and Φ_n may be linear rather than simply asymptotically linear. In such situations, the remainder terms $H_{x,n}$ and $R_{x,n}$ are identically zero, and conditions (C4) and (C5) are trivially satisfied. Otherwise, these conditions must be verified. While in general the exact form of these remainder terms depends upon the specific parameter under consideration and estimators used, it is frequently the case that part of the remainder is an empirical process term arising from the estimation of nuisance functions appearing in the influence functions $D^*_{x,0}$ and $L^*_{x,0}$, as we illustrate below with one particular construction. To facilitate the verification of conditions (C4) and (C5) for these empirical process terms, we outline sufficient conditions in terms of uniform entropy and bracketing numbers.

In this subsection, we assume that $\Gamma_0(x)$ and $\Phi_0(x)$ arise as the evaluation at P_0 of

maps from $\mathcal{M}$ to $\mathbb{R}$, and denote by $\Gamma_P(x)$ and $\Phi_P(x)$ the evaluation of these maps at an arbitrary $P \in \mathcal{M}$. Let $\pi = \pi(P)$ be a summary of P, and suppose that $\Gamma_P(x)$, $\Phi_P(x)$ and the nonparametric efficient influence functions of $P \mapsto \Gamma_P(x)$ and $P \mapsto \Phi_P(x)$ at P each only depend on P through π. Denote these efficient influence functions by $D_x^*(\pi)$ and $L_x^*(\pi)$, respectively. Since $\mathcal{M}$ is nonparametric, it must be that $D_{x,0}^* = D_x^*(\pi_0)$ and $L_{x,0}^* = L_x^*(\pi_0)$ for $\pi_0 := \pi(P_0)$. To emphasize the fact that $\Gamma_P(x)$ and $\Phi_P(x)$ depend on P only through π, we will use the symbols $\Gamma_\pi(x)$ and $\Phi_\pi(x)$ to refer to $\Gamma_P(x)$ and $\Phi_P(x)$, respectively.

Under regularity conditions, the so-called *one-step estimators*

$$\Gamma_n(x) := \Gamma_{\pi_n}(x) + \mathbb{P}_n D_x^*(\pi_n) \quad \text{and} \quad \Phi_n(x) := \Phi_{\pi_n}(x) + \mathbb{P}_n L_x^*(\pi_n) \tag{3.4}$$

are asymptotically linear and efficient estimators of $\Gamma_0(x)$ and $\Phi_0(x)$, even when π_n is a data-adaptive (e.g., machine learning) estimator of π_0 (e.g., Pfanzagl, 1982). van der Vaart and van der Laan (2006) pioneered the use of such one-step estimators in the context of nonparametric monotone function estimation. When this one-step construction is used, it can be shown that the remainder terms have the form $H_{x,n} = H_{x,n}^{(1)} + H_{x,n}^{(2)}$ and $R_{x,n} = R_{x,n}^{(1)} + R_{x,n}^{(2)}$, where $H_{x,n}^{(1)} := (\mathbb{P}_n - P_0)\left[D_x^*(\pi_n) - D_x^*(\pi_0)\right]$ and $R_{x,n}^{(1)} := (\mathbb{P}_n - P_0)\left[L_x^*(\pi_n) - L_x^*(\pi_0)\right]$ are empirical process terms, and $H_{x,n}^{(2)}$ and $R_{x,n}^{(2)}$ are so-called *second-order* remainder terms arising from linearization of the corresponding parameter. Similar representations exist when other constructive approaches, such as gradient-based estimating equations methodology (e.g., van der Laan and Robins, 2003; Tsiatis, 2007) and targeted maximum likelihood estimation (e.g., van der Laan and Rose, 2011), are used. As we will see in the examples of Section 3.5, these second-order terms can usually be shown to be asymptotically negligible provided π_n tends to π_0 fast enough in some appropriate norm. Here, we provide conditions on π_n that ensure that the contribution of $H_{x,n}^{(2)} - \theta_0(x) R_{x,n}^{(2)}$ to $K_n(\delta)$ satisfies conditions (C4) and (C5).

A primary benefit of decomposing the remainder terms as above is that the empirical process terms can be controlled using empirical process theory, a strategy also used in van der Vaart and van der Laan (2006). In particular, we can provide conditions under which

$H_{x,n}^{(1)}$ and $R_{x,n}^{(1)}$ satisfy conditions (C4) and (C5). Defining $g_{x,u}(\pi) := [D_{x+u}^*(\pi) - D_x^*(\pi)] - \theta_0(x)[L_{x+u}^*(\pi) - L_x^*(\pi)]$, the relevant contribution of these empirical process terms to $K_n(\delta)$ is

$$K_n^{(1)}(\delta) := n^{1/6} \sup_{|u| \leq \delta} \left| \mathbb{G}_n \left[g_{x,un^{-1/3}}(\pi_n) - g_{x,un^{-1/3}}(\pi_0) \right] \right| .$$

Suppose that π_n falls in a semimetric space $(\mathcal{P}, \rho)$, with probability tending to one, and that $G_{x,\mathcal{P},R}$ is an envelope function for $\mathcal{G}_{x,\mathcal{P},R} := \{g_{x,u}(\pi) : |u| \leq R, \pi \in \mathcal{P}\}$. We consider the following the conditions:

(D1) for some constants $C > 0$ and $V > -1$, either

(D1a) $\log N_{[]}(\varepsilon \|G_{x,\mathcal{P},R}\|_{P_0,2}, \mathcal{G}_{x,\mathcal{P},R}, L_2(P_0)) \leq C\varepsilon^{2V}$, or

(D1b) $\log \sup_Q N(\varepsilon \|G_{x,\mathcal{P},R}\|_{Q,2}, \mathcal{G}_{x,\mathcal{P},R}, L_2(Q)) \leq C\varepsilon^{2V}$

for all $\varepsilon \in (0, 1]$ and R small enough;

(D2) $P_0 G_{x,\mathcal{P},R}^2 = O(R)$, and for all $\eta > 0$, $P_0 G_{x,\mathcal{P},R}^2 \{RG_{x,\mathcal{P},R} > \eta\} = o(R)$, as $R \to 0$;

(D3) $P_0 \left[g_{x,u}(\pi) - g_{x,v}(\pi) \right]^2 = O(|u - v|)$ uniformly for $\pi \in \mathcal{P}$, and

$$\frac{P_0 \left[g_{x,u}(\pi_1) - g_{x,u}(\pi_2) \right]^2}{\rho(\pi_1, \pi_2)^2} = O(|u|)$$

uniformly for $\pi_1, \pi_2 \in \mathcal{P}$ and $u \in I$;

(D4) there exists some $\bar{\pi} \in \mathcal{P}$ such that $\rho(\pi_n, \bar{\pi}) \xrightarrow{\mathrm{P}} 0$.

Our next result states that, under these conditions, the remainder term $K_n^{(1)}(\delta)$ stated above is asymptotically negligible in the sense of conditions (C4) and (C5).

Theorem 9. *Suppose that, with probability tending to one, $\pi_n \in \mathcal{P}$ and conditions (D1)–(D4) hold. Then, $K_n^{(1)}(\delta)$ satisfies conditions (C4)–(C5).*

We note that conditions (D1) and (D2) together imply conditions (C1) and (C2). As such, if conditions (D1) and (D2) have been verified, there is no need to also verify conditions (C1) and (C2).

Theorem 9 is established with the help of the following simple result, which is a useful generalization of van der Vaart (1998) Theorem 19.24.

Lemma 6. *Let* $\{V_n(u, f) : u \in \mathcal{U}, f \in \mathcal{F}\}$ *be a sequence of stochastic processes indexed by* $\mathcal{U} \times \mathcal{F}$*, where* $(\mathcal{U}, d_1)$ *and* $(\mathcal{F}, d_2)$ *are semi-metric spaces. Let* ρ *be the corresponding product semi-metric on* $\mathcal{U} \times \mathcal{F}$*. Suppose* V_n *are asymptotically uniformly* ρ*-equicontinuous in the sense of VW and* $d_2(f_n, f_0)$ *tends to zero in probability. Then,* $\sup_{u \in \mathcal{U}} |V_n(u, f_n) - V_n(u, f_0)|$ *tends to zero in probability.*

3.5 Applications of the general theory

In this section, we demonstrate the use of our general results for the three examples introduced in Section 3.2: estimation of monotone density, hazard and regression functions. For each of these functions of interest, we consider various levels of complexity of the relationship between the observed and ideal data units. This allows us to illustrate that our general results (i) coincide with classical results in the simpler cases that have already been studied, and (ii) suggest novel estimation procedures with well-understood inferential properties even in the context of complex problems that do not appear to have been previously studied. Below, we focus on distributional results for the various estimators considered.

3.5.1 Example 1: monotone density function

Let $\theta_0 := f_0$ be the density function of an event time T with support $I := [0, u_0]$, and suppose that f_0 is known to be non-decreasing on I. We will not use any transformation in this example, so we take Φ_0 and Φ_n to be the identity map. Thus, $\psi_0 = \theta_0$ also corresponds to the density function of T, and $\Psi_0 = \Theta_0 = \Gamma_0$ to its distribution function. Below, we consider various data settings that increase in complexity. In the first setting, available

observations are independent draws from the distribution of interest. In the second, each of these observations are subject to independent right-censoring. In the third, the right-censoring mechanism is allowed to be informative – only conditional independence of the event and censoring times given a vector of observed covariates is assumed. The first two cases have been studied in the literature – for these, we wish to verify that our general results coincide with results already established. The third case is more difficult and does not seem to have been studied before. Our work in this setting not only highlights the generality of the theory in Sections 3.3 and 3.4, but also yields novel methodology.

No censoring

In this simple case, we have that $O = T$ is directly sampled from a distribution with density function θ_0. We wish to estimate $\theta_0(x)$ for some $x \in I$ at which θ_0 is differentiable with positive derivative. Let Γ_n be the empirical distribution function based on $O_1, O_2, \ldots, O_n$. This estimator is linear with influence function $D^*_{x,0} : t \mapsto I_{(-\infty,x]}(t) - \Gamma_0(x)$, and so, the local process involves the class of functions $\mathcal{G}_{x,R} := \{t \mapsto I_{(x,x+u]}(t) : |u| \leq R\}$, which is Vapnik-Červonenkis. Its covering number is thus polynomial in ϵ with respect to the natural envelope $G_{x,R} : t \mapsto I_{[0,R]}(|t - x|)$, thus satisfying (B1b). Since θ_0 is differentiable at x, $G_{x,R}$ satisfies (C2). The asymptotic covariance function of $n^{1/2}(\Gamma_n - \Gamma_0)$ is $\Sigma_0 : (s,t) \mapsto \Gamma_0(s \wedge t) - \Gamma_0(s)\Gamma_0(t)$. Thus, we have that $\Sigma^*_0(s,t) = \Theta_0(s)\Theta_0(t)$, $A_0(s,t,y,w) = \theta_0(y)$ and $H_0(y,w) = y$. It is easy to verify that (C3) is satisfied and furthermore that $\kappa_0(x) = \theta_0(x)$. Conditions (C4) and (C5) are trivially satisfied because of the linearity of Γ_n. Theorem 7 yields that $n^{1/3}[\theta_n(x) - \theta_0(x)] \xrightarrow{\text{d}} \tau_0(x)Z$, where $\tau_0(x) = [4f'_0(x)f_0(x)]^{1/3}$ and Z follows the standard Chernoff distribution. This finding agrees with classical results for the Grenander estimator (Prakasa Rao, 1969).

Independent censoring

Now, suppose that C is a positive random variable independent of T, and that the observed data unit $O = (Y, \Delta)$, where $Y = \min(T, C)$ and $\Delta = I(T \leq C)$. The NPMLE of a monotone

density function based on independently right-censored data was obtained in Laslett (1982) and McNichols and Padgett (1982), and distributional results were derived in Huang and Zhang (1994). Huang and Wellner (1995) considered an estimator θ_n obtained by differentiating the GCM of the Kaplan-Meier estimator of the distribution function. While this is not the NPMLE, Huang and Wellner (1995) showed that it is asymptotically equivalent to the NPMLE, and it is an attractive estimator because it is simple to construct and reduces to the Grenander estimator if T is fully observed, that is, if $C \geq T$ almost surely.

Since Ψ_0 is the distribution function $F_0 = 1 - S_0$ with S_0 denoting the survival function of T, it is natural to consider $\Psi_n := 1 - S_n$, where S_n is the Kaplan-Meier estimator of S_0. It is well known that $n^{1/2}(S_n - S_0)$ converges weakly in $\ell^\infty([0,\tau])$ to a tight zero-mean Gaussian process as long as $G_0(\tau) > 0$ and $S_0(\tau) < 1$, where G_0 denotes the survival function of C. Denoting by Λ_0 the cumulative hazard function corresponding to S_0, the influence function of the Kaplan-Meier estimator $S_n(x)$ is known to be the nonparametric efficient influence function

$$D^*_{0,x} : (y,\delta) \mapsto S_0(x)\left[-\frac{\delta I_{[0,x]}(y)}{S_0(y)G_0(y)} + \int_0^{y\wedge x} \frac{\Lambda_0(du)}{G_0(u)S_0(u)}\right]$$

and so, the local difference $g_{x,u}$ can be written as

$$(y,\delta) \mapsto \frac{-[S_0(x+u) - S_0(x)]\delta I_{[0,x+u]}(y)}{S_0(y)G_0(y)} - \frac{S_0(x)\delta I_{(x,x+u]}(y)}{S_0(y)G_0(y)} + \int_{v<y} \frac{I_{(x,x+u]}(v)}{S_0(v)G_0(v)}\Lambda_0(dv) \ .$$

The class of functions $\mathcal{G}_{x,R}$ is a Lipschitz transformation of the classes $\{u \mapsto S_0(x+u) - S_0(x) : |u| \leq R\}$ and $\{t \mapsto I_{(x,x+u]}(t) : |u| \leq R\}$, and hence satisfies (C1). An envelope function for $\mathcal{G}_{x,R}$ is given by

$$G_{x,R} : (y,\delta) \mapsto \frac{\delta I_{[0,R]}(|y-x|)}{S_0(y)G_0(y)}(1 + KR) + \int_{x-R}^{x+R} \frac{I_{[0,y)}(v)}{S_0(v)G_0(v)}\Lambda_0(dv) \ .$$

It is easy to see that (C2) is satisfied if S_0 and G_0 are positive in a neighborhood of x. The covariance function is given by $\Sigma_0 : (s,t) \mapsto \int_0^{s\wedge t} \frac{S_0(s)S_0(t)}{S_0(u)G_0(u)}\Lambda_0(du)$. Display (3.3) is thus satisfied with $A_0(s,t,v,w) = [S_0(t)S_0(s)]/[G_0(v)S_0(v)]$ and $H_0(v,w) = \Lambda_0(v)$. Con-

dition (C3) is satisfied if θ_0 is positive and continuous in a neighborhood of x. We then get $\kappa_0(x) = [S_0(x)/G_0(x)]\lambda_0(x) = f_0(x)/G_0(x)$, so that the scale parameter is $\tau_0(x) = [4f_0'(x)f_0(x)/G_0(x)]^{1/3}$. This agrees with the results of Huang and Wellner (1995).

It remains to scrutinize the conditions arising from the remainder term $H_{x,n}$. If G_n is the Kaplan-Meier estimator of G_0, it is always true that $\mathbb{P}_n D^*_{n,x} = 0$, where $D^*_{n,x}$ is the estimator of $D^*_{0,x}$ obtained by replacing S_0 and G_0 by S_n and G_n, respectively. It is easy to verify that $H_{x,n}$ can be decomposed as $H^{(1)}_{x,n} + H^{(2)}_{x,n}$, where $H^{(1)}_{x,n} := (\mathbb{P}_n - P_0)(D^*_{n,x} - D^*_{0,x})$ is the usual empirical process term and

$$H^{(2)}_{x,n} := S_n(x) \int_0^x \frac{S_0(u)}{S_n(u)} \left[\frac{G_0(u)}{G_n(u)} - 1\right] (\Lambda_n - \Lambda_0)(du)$$

is the second-order remainder term. The local remainder emanating from $H^{(1)}_{x,n}$ can be studied using results from Section 3.4.3. We instead focus on the local remainder $K^{(2)}_n(\delta) := n^{2/3} \sup_{|u| \leq \delta n^{-1/3}} |H^{(2)}_{n,x+u} - H^{(2)}_{n,x}|$, which can be bounded as

$$\begin{aligned} K^{(2)}_n(\delta) \;\leq\; & n^{2/3} \int_0^{x+\delta n^{-1/3}} \frac{S_0(u)}{S_n(u)} \left|\frac{G_0(u)}{G_n(u)} - 1\right| |(\Lambda_n - \Lambda_0)(du)| \sup_{|u| \leq \delta n^{-1/3}} |S_n(x+u) - S_n(x)| \\ & + n^{2/3} \int_{x-\delta n^{-1/3}}^{x+\delta n^{-1/3}} \frac{S_0(u)}{S_n(u)} \left|\frac{G_0(u)}{G_n(u)} - 1\right| |(\Lambda_n - \Lambda_0)(du)| \;. \end{aligned}$$

Writing $|S_n(x+u) - S_n(x)| = n^{-1/2}\{n^{1/2}[S_n(x+u) - S_0(x+u)] - n^{1/2}[S_n(x) - S_0(x)]\} + [S_0(x+u) - S_0(x)]$, we note that $\sup_{|u| \leq \delta n^{-1/3}} |S_n(x+u) - S_n(x)| = \mathrm{o_P}(n^{-1/2}) + \mathrm{O_P}(n^{-1/3})$ in view of the weak convergence of $n^{1/2}(S_n - S_0)$ in a neighborhood of x. Using that $\int_a^b |f(u)||g(du)| \leq \sup_{u \in [a,b]} |f(u)| \|g\|_{TV,[a,b]}$ with $\|\cdot\|_{TV,[a,b]}$ denoting the total variation norm over $[a,b]$, and that $\|\Lambda_n - \Lambda_0\|_{TV,[a,b]} \leq \|\Lambda_n\|_{TV,[a,b]} + \|\Lambda_0\|_{TV,[a,b]} = [\Lambda_n(b) - \Lambda_n(a)] + [\Lambda_0(b) - \Lambda_0(a)]$ in view of the monotonicity of Λ_n and Λ_0, we find that

$$K^{(2)}_n(\delta) \;=\; \left[\mathrm{o_P}(n^{-1/6}) + \mathrm{O_P}(1)\right] \mathrm{O_P}(n^{-1/6}) \;=\; \mathrm{O_P}(n^{-1/6}) \;,$$

which is sufficient to establish conditions (C4) and (C5).

Conditionally independent censoring

In many cases, the censoring mechanism may be informative but still independent of the event time process conditionally on a vector of recorded covariates. For simplicity, we only consider the case in which these covariates are defined at baseline although the case of time-varying covariates can be tackled similarly. The observed data unit is now $O = (Y, \Delta, W)$, and we assume that T and C are independent given W. As long as $P_0(\Delta = 1 \mid W)$ is bounded away from zero almost surely, the survival function S_0 of T can be identified pointwise in terms of the distribution P_0 of O via the product-limit transform

$$S_0(x) = \int \prod_{t \leq x} \left[1 - \frac{F_{1,0}(dt, w)}{S_{Y,0}(t \mid w)} \right] Q_0(dw) \ ,$$

where $F_{1,0}(t \mid w) := P_0(Y \leq t, \Delta = 1 \mid W = w)$ is the conditional subdistribution function of Y given $W = w$ corresponding to $\Delta = 1$, $S_{Y,0}(t \mid w) := P_0(Y \geq t \mid W = w)$ is the conditional proportion-at-risk at time t given $W = w$, and Q_0 is the marginal distribution of W under P_0. This constitutes an example of coarsening at random, as described in Heitjan and Rubin (1991) and Gill et al. (1997). Estimation of the marginal survival function S_0 in the context of conditionally independent censoring has been studied before by Hubbard et al. (2000), Scharfstein and Robins (2002) and Zeng (2004), among others.

In this context, the nonparametric efficient influence function $D^*_{0,x}$ of $S_0(x)$ has the form $D_{0,x} - S_0(x)$, where $D_{0,x}$ is given by

$$(y, \delta, w) \mapsto -S_0(x \mid w) \left[\frac{\delta I_{(-\infty, x]}(y)}{S_0(y \mid w) G_0(y \mid w)} - \int_0^{y \wedge x} \frac{\Lambda(du \mid w)}{S_0(u \mid w) G_0(u \mid w)} \right] + S_0(x \mid w)$$

with $S_0(x \mid w)$ and $G_0(x \mid w)$ the conditional survival functions of T and C, respectively, at x given $W = w$, and $\Lambda_0(x \mid w)$ is the conditional cumulative hazard function of T at x given $W = w$. A simple one-step estimator of $\Gamma_0(x)$ is given by $\Gamma_n(x) := 1 - \mathbb{P}_n D_{n,x}$, where $D_{n,x}$ is obtained by substituting S_n and G_n for S_0 and G_0, respectively, in $D_{0,x}$. Conditions (C1) and (C2) are satisfied under Lipschitz conditions on S_0 and G_0 uniformly over w. The

asymptotic covariance function is given by

$$\Sigma_0(s,t) = \int S_0(t \mid w) S_0(s \mid w) Q_0(dw) - S_0(t) S_0(s) + \iint_0^{s \wedge t} \frac{S_0(t \mid w) S_0(s \mid w)}{G_0(y \mid w) S_0(y \mid w)} \Lambda_0(dy \mid w) Q_0(dw) \,,$$

and so, we find that display (3.3) holds with $\Sigma_0^*(s,t) = \int S_0(t \mid w) S_0(s \mid w) Q_0(dw) - S_0(t) S_0(s)$, $A_0(s,t,v,w) = [S_0(t \mid w) S_0(s \mid w)]/[G_0(v \mid w) S_0(v \mid w)]$ and $H_0(v,w) = \Lambda_0(v \mid w)$. Thus, condition (C3) holds, and we get $\kappa_0(x) = \int [f_0(x \mid w)/G_0(x \mid w)] Q_0(dw)$, where $f_0(x \mid w)$ is the conditional density of T at x given $W = w$. It follows directly then that the Chernoff scale factor is

$$\tau_0(x) = \left[4 f_0'(x) \int \frac{f_0(x \mid w)}{G_0(x \mid w)} Q_0(dw) \right]^{1/3} ,$$

which reduces to the scale factor of Huang and Wellner (1995) when T and C are independent.

The remainder term $H_{x,n}$ again has the form $H_{x,n}^{(1)} + H_{x,n}^{(2)}$ with $H_{x,n}^{(1)} := (\mathbb{P}_n - P_0)(D_{n,x}^* - D_{0,x}^*)$ and

$$H_{x,n}^{(2)} := \int S_n(x \mid w) \int_0^x \frac{S_0(y \mid w)}{S_n(y \mid w)} \left[\frac{G_0(y \mid w)}{G_n(y \mid w)} - 1 \right] (\Lambda_n - \Lambda_0)(dy \mid w) Q_0(dw) \,.$$

Once more, we focus on $H_{x,n}^{(2)}$. Writing $S_n^{(0)} := S_n - S_0$, if S_n and G_n are bounded away from zero in a neighborhood of x with probability tending to one, and if S_0 is Lipschitz in x uniformly in w, then the term $K_n^{(2)}(\delta) := n^{2/3} \sup_{|u| \leq \delta n^{-1/3}} |H_{x+u,n}^{(2)} - H_{x,n}^{(2)}|$ is bounded by a constant multiple of

$$n^{2/3} \left\{ \left[E_0 \sup_{|u-x| \leq \epsilon} |S_n^{(0)}(u \mid W) - S_n^{(0)}(x \mid W)|^2 \right]^{1/2} + \delta n^{-1/3} \right\} \times \left[E_0 \sup_{u \leq x+\epsilon} |G_n(u \mid W) - G_0(u \mid W)|^2 \right]^{1/2}$$

with probability tending to one. Control of $K_n^{(2)}(\delta)$ is highly dependent on the behavior of

S_n and G_n. If, for instance, $S_n - S_0$ and $G_n - G_0$ uniformly tend to zero in probability at rates faster than $n^{-1/3}$, then conditions (C4) and (C5) are satisfied. This is not a restrictive requirement if W only has few components – in such cases, many nonparametric smoothing-based estimators satisfy such rates. Otherwise, semiparametric estimators building upon additional structure (e.g., additivity on an appropriate scale) could be used. Alternatively, for higher-dimensional W, estimators of the form $S_n(x \mid w) = \exp\left[-\int_0^x \lambda_n(u \mid w)du\right]$ with λ_n an estimator of the conditional hazard λ_0 may be worth considering. For such S_n, we find that $K_n^{(2)}(\delta)$ is bounded by a constant multiple of

$$\delta n^{1/3}\left[E_0 \sup_{u \leq x+\epsilon} |\lambda_n(u \mid W) - \lambda_0(u \mid W)|^2 E_0 \sup_{u \leq x+\epsilon} |G_n(u \mid W) - G_0(u \mid W)|^2\right]^{1/2}$$

with probability tending to one, and so, we require that the product of the convergence rates of $\lambda_n - \lambda_0$ and $G_n - G_0$ to be faster than $n^{-1/3}$. In practice, with a moderate or high-dimensional covariate vector W, it seems desirable to leverage multiple candidate estimators using ensemble learning (e.g., van der Laan et al., 2007; van der Laan and Rose, 2011).

3.5.2 Example 2: monotone hazard function

We now consider estimation of $\theta_0 := \lambda_0$, the hazard function of T. The most obvious approach to tackle this problem would be to consider an identity transformation as in the previous example. The primitive function of interest is then the cumulative hazard function Λ_0, which can be expressed as the negative logarithm of the survival function S_0 and estimated naturally using any asymptotically linear estimator of S_0, for example. The conditions of Theorem 6 and 7 can then be directly verified. An alternative, more expeditious approach consists of taking the domain transform Φ_0 to be the restricted mean mapping $u \mapsto \int_0^u S_0(v)dv$. In such case, Γ_0 is simply the cumulative distribution function F_0. This particular choice of domain transformation for estimating a monotone hazard function therefore yields the same parameter Γ_0 as for estimating a monotone density with the identity transform. Denoting by S_n the estimator of the survival function S_0 based on the available data, the resulting

generalized Grenander-type estimator θ_n is defined by taking $\Gamma_n := 1 - S_n$ and setting Φ_n to be $u \mapsto \int_0^u S_n(v)dv$. As the result below suggests, when this special domain transform is used, we can leverage some of the work performed above in analyzing the Grenander-type estimator of a monotone density function under the various right-censoring schemes considered. We recall that Id denotes the identity function.

Theorem 10. *Suppose that* $E_0\left[\sup_{u\in I_n}|S_n(u) - S_0(u)|\right] = o(r_n^{-1})$ *and set* $\Gamma_n := 1 - S_n$. *If the pair* (Γ_n, Id) *satisfies conditions (B1)–(B3), then the pair* (Γ_n, Φ_n) *with* $\Phi_n : u \mapsto \int_0^u S_n(v)dv$ *necessarily satisfies conditions (B1)–(B5). Thus, for* $\theta_n := \mathrm{Iso}_{[0,u_n]}(\Gamma_n \circ \Phi_n^-, \Phi_n)$, *this implies that*

$$r_n\left[\theta_n(x) - \theta_0(x)\right] \xrightarrow{\mathrm{d}} -\theta_0'(x) \operatorname*{argmin}_{u\in\mathbb{R}} \left\{W_x(u) + \tfrac{1}{2}\theta_0'(x)S_0(x)u^2\right\} \ .$$

If $W_x = [\kappa_0(x)]^{1/2}W_0$ *for* W_0 *a two-sided Brownian motion, then* $r_n\left[\theta_n(x) - \theta_0(x)\right] \xrightarrow{\mathrm{d}} \tau_0(x)Z$, *where* Z *is the standard Chernoff distribution and* $\tau_0(x) := \left[4\theta_0'(x)\kappa_0(x)/S_0(x)^2\right]^{1/3}$.

Denote by $T_{(j)}$ the j^{th} order statistic of $\{T_1, T_2, \ldots, T_n\}$ and define $T_{(0)} := 0$. When there is no censoring, the choice (Γ_n, Φ_n) prescribed above yields that Γ_n is the empirical distribution function and Φ_n is defined pointwise as $\Phi_n(x) := \frac{1}{n}\sum_{i=1}^n \min(T_{(i)}, x)$, which is strictly increasing on $[0, T_{(n)}]$. Therefore, $\theta_n(x)$ is the left derivative at $\Phi_n(x)$ of the GCM of the graph of $\{(\Phi_n(T_{(k)}), \Gamma_n(T_{(k)})) : k = 0, 1, \ldots, n\} = \{((\frac{n-k}{n})T_{(k)} + \frac{1}{n}\sum_{i=1}^k T_{(i)}, \frac{k}{n}) : k = 0, 1, \ldots, n\}$. This is the NPMLE of a non-decreasing hazard function with uncensored data (see, e.g., Chapter 2.6 of Groeneboom and Jongbloed, 2014).

We note that in Section 3.5.1 we verified (B1)–(B3) for each of three right-censoring schemes for $\Theta_n = 1 - S_n$ and Φ_0 and Φ_n both equal to the identity. Thus, to use Theorem 10, it would suffice to verify that $E_0\left[\sup_{u\in I_n}|S_n(u) - S_0(u)|\right]$ tends to zero faster than $n^{-1/3}$. This is straightforward given the weak convergence of $n^{1/2}(S_n - S_0)$. Thus, the above theorem provides distributional results for monotone hazard function estimators in each right-censoring scheme considered, as summarized below:

(i) when there is no censoring, we find $\tau_0(x) = [4\lambda_0'(x)\lambda_0(x)/S_0(x)]^{1/3}$, which agrees with Prakasa Rao (1970);

(ii) when there is independent right-censoring, we find that

$$\tau_0(x) = \{4\lambda_0'(x)\lambda_0(x)/[G_0(x)S_0(x)]\}^{1/3} ,$$

which agrees with Huang and Wellner (1995);

(iii) when there is conditionally independent right-censoring, an important setting that does not seem to have been previously studied in the literature, we find that

$$\begin{aligned}\tau_0(x) &= \left[\frac{4\lambda_0'(x)}{S_0(x)^2}\int \frac{f_0(x \mid w)}{G_0(x \mid w)} Q_0(dw)\right]^{1/3} \\ &= \left\{\frac{4\lambda_0'(x)\lambda_0(x)}{G_0(x)S_0(x)}\left[\frac{G_0(x)}{f_0(x)}\int \frac{f_0(x \mid w)}{G_0(x \mid w)} Q_0(dw)\right]\right\}^{1/3} .\end{aligned}$$

If either T or C are independent of W, the unadjusted Kaplan-Meier estimator is consistent for the true marginal survival function of T, and so, unadjusted estimators of the density and hazard functions are consistent. In these cases, we may then ask how the asymptotic distributions of the adjusted and unadjusted estimators compare. Since all limit distributions are of the scaled Chernoff type, it suffices to compare the scale factors arising from the different estimators. The second expression in (iii) is helpful to assess the impact of unnecessary covariate adjustment. If C and W are independent, then $G_0(x \mid w) = G_0(x)$ for each w, and so, the scale factors in (ii) and (iii) are identical. If T and W are dependent, so that $f_0(x \mid w) = f_0(x)$ for each w, but C and W are not, then the scale factor in (iii) is generally larger than that the scale factor in (ii). In summary, when using an adjusted rather than unadjusted estimator of the hazard function, there may only be a penalty in asymptotic efficiency when adjusting for covariates that C depends on but T does not. The relative loss of efficiency is given by $\left\{\int [G_0(x)/G_0(x \mid w)] Q_0(dw)\right\}^{1/3}$.

3.5.3 Example 3: monotone regression function

We finally consider estimation of a monotone regression function. We first focus on the simple case in which the association between the outcome and exposure of interest is not confounded. In such cases, the parameter of interest is the conditional mean of the outcome given exposure level, and the standard least-squares isotonic regression estimators can be used. We show that our general theory covers this classical case. We then consider the case in which the relationship between outcome and exposure is confounded but the confounders of this relationship have been recorded. In this more challenging case, we consider the marginalization (or standardization) of the conditional mean outcome given exposure level and confounders over the marginal confounder distribution. We study this problem using results from Section 3.4, which allow us to provide theory for a novel estimator proposed for this important case.

No confounding

In the standard least-squares isotonic regression problem, we observe independent replicates of $O := (A, Y)$, where $Y \in \mathbb{R}$ is an outcome and $A \in \mathbb{R}$ is the exposure of interest. We are interested in the conditional mean function $\theta_0 := \mu_0$, where $\mu_0(x) := E_0(Y \mid A = x)$ is the mean outcome at exposure level x. The primitive function of θ_0 can be written as $\Theta_0(t) = E_0\left[YI_{(-\infty,t]}(A)/f_0(A)\right]$ for each t, where f_0 is the marginal density of A. The corresponding primitive parameter at x is pathwise differentiable with nonparametric efficient influence function $(a, y) \mapsto yI_{(-\infty,x]}(a)/f_0(a) - \Theta_0(x)$. An obvious approach to estimation of θ_0 consists of constructing an asymptotically linear estimator of Θ_0 – this involves nonparametric estimation of the nuisance density f_0 – and differentiating the GCM of the resulting curve – this involves selecting the interval over which the GCM is calculated.

By using a domain transformation, it is possible to avoid both the need for nonparametric density estimation and the choice of isotonization interval. Let Φ_0 be the marginal distribution function of A. With this transformation, we note that $\Psi_0(t) = E_0\left[YI_{(-\infty,t]}(\Phi_0(A))\right]$ and

$\Gamma_0(t) = E_0\left[Y I_{(-\infty,t]}(A)\right]$ for each t. This suggests taking Φ_n to be the empirical distribution function based on $A_1, A_2, \ldots, A_n$ and $\Gamma_n(x) := \frac{1}{n}\sum_{i=1}^n Y_i I_{(-\infty,x]}(A_i)$. The resulting estimator $\theta_n(x)$ is precisely the well-known least-squares isotonic regression estimator of $\theta_0(x)$.

Because both Γ_n and Φ_n are linear estimators, these estimators do not generate second-order remainder terms to analyze. The influence functions of Γ_n and Φ_n are, respectively, $D^*_{0,x} : (a,y) \mapsto y I_{(-\infty,x]}(a) - \Gamma_0(x)$ and $L^*_{0,x} : (a,y) \mapsto I_{(-\infty,x]}(a) - \Phi_0(x)$, and so, we find the localized difference function to be $g_{x,u} : (a,y) \mapsto [y - \theta_0(x)] I_{(x,x+u]}(a) - [\Gamma_0(x+u) - \Gamma_0(x)] + \theta_0(x)[\Phi_0(x+u) - \Phi_0(x)]$. The second and third summands are constant as functions of (a,y) and Lipschitz in u with a constant envelope function. Hence, they easily satisfy conditions (C1) and (C2). The first summand is the fixed function $(a,y) \mapsto [y - \theta_0(x)]$ multiplied by an element of the class $\{v \mapsto I_{(x,x+u]}(v) : u > 0\}$. This class has been studied for the Grenander estimator of a monotone density function; in particular, it is known to possess polynomial covering numbers. The natural envelope for the class generated by the first summand is thus $(a,y) \mapsto |y - \theta_0(x)| I_{[0,R]}(|a - x|)$, which satisfies (C1) and (C2) if, in a neighborhood of x, the conditional variance function, defined pointwise as $\sigma_0^2(t) := \mathrm{Var}_0(Y \mid A = t)$, is bounded and Φ_0 possesses a positive, continuous density. In such cases, Theorem 7 holds. Through straightforward calculations, we find that

$$\Sigma_0(s,t) = -\left[\Gamma_0(s) - \theta_0(x)\Phi_0(s)\right]\left[\Gamma_0(t) - \theta_0(x)\Phi_0(t)\right] + E_0\left\{1_{(-\infty,s\wedge t]}(A)\left[Y - \theta_0(x)\right]^2\right\} .$$

The first summand is continuously differentiable at (x,x) since each of θ_0 and Φ_0 are continuously differentiable at x. The second summand can be expressed as $\int_{-\infty}^{s\wedge t}\{\sigma_0^2(u) + [\theta_0(u) - \theta_0(x)]^2\}\Phi_0(du)$. We thus confirm that display (3.3) holds with $A_0(s,t,v,w) = \sigma_0^2(v) + [\theta_0(v) - \theta_0(x)]^2$ and $H_0 = \Phi_0$. Provided σ_0^2 is continuous at x and Φ_0 is continuously differentiable at x, condition (C3) holds. As such, we obtain that $n^{1/3}\left[\theta_n(x) - \theta_0(x)\right]$ has a scaled Chernoff distribution with scale parameter

$$\tau_0(x) = \left[\frac{4\mu_0'(x)\sigma_0^2(x)}{f_0(x)}\right]^{1/3}$$

coinciding with the classical results of Brunk (1970).

Confounding by recorded covariates

We now consider a scenario in which the relationship between outcome Y and exposure A is confounded by a vector W of recorded covariates. The observed data unit is thus $O := (W, A, Y)$. A more relevant estimand in this scenario might be the marginalized regression function $\theta_0 := \nu_0$ with $\nu_0(x)$ defined as $E_0 \left[E_0 \left(Y \mid A = x, W \right) \right]$. We note that $\nu_0(x)$ can be interpreted as a causal dose-response curve if (a) W includes all confounders of the relationship between A and Y, and (b) the probability of observing an individual subject to exposure level x is positive in P_0-almost every stratum defined by W. In many scientific settings, it may be known that the causal dose-response curve is monotone in exposure level.

We again consider transformation by the marginal distribution function of A. In other words, we set $\Phi_0(x) := P_0(A \leq x)$ and take $\Phi_n(x) := \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(A_i)$ for each x. We then have that

$$\Gamma_0(x) = E_0 \left[\frac{Y I_{(-\infty, x]}(A)}{g_0(A, W)} \right] = \iint I_{(-\infty, x]}(a) \mu_0(a, w) \Phi_0(da) Q_0(dw) \ ,$$

where g_0 is the density ratio $(a, w) \mapsto f_0(a \mid w) / f_0(a)$, with $f_0(a \mid w)$ denoting the conditional density function of A at a given $W = w$ and $f_0(a)$ is the marginal density function of A at a as before, and μ_0 is the regression function $(a, w) \mapsto E_0(Y \mid A = a, W = w)$. While in this case the domain transform does not eliminate the need to estimate nuisance functions, it nevertheless results in a procedure for which there is no need to choose the interval over which the GCM is calculated.

Setting $\eta_0(x, w) := \int I_{(-\infty, x]}(a) \mu_0(a, w) \Phi_0(da)$ for each x and w, the nonparametric efficient influence function of $\Gamma_0(x)$ is

$$(w, a, y) \mapsto I_{(-\infty, x]}(a) \left[\frac{y - \mu_0(a, w)}{g_0(a, w)} + \theta_0(a) \right] + \eta_0(x, w) - 2\Gamma_0(x) \ .$$

Suppose that μ_n and g_n denote estimators of μ_0 and g_0, respectively. If the empirical distributions Φ_n and Q_n based on $A_1, A_2, \ldots, A_n$ and $W_1, W_2, \ldots, W_n$, respectively, are used as estimators of Φ_0 and Q_0, it is not difficult to show that

$$\Gamma_n(x) := \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(A_i) \left[\frac{Y_i - \mu_n(A_i, W_i)}{g_n(A_i, W_i)} + \frac{1}{n} \sum_{j=1}^{n} \mu_n(A_i, W_j) \right]$$

is a one-step estimator of $\Gamma_0(x)$, and that it is asymptotically efficient under regularity conditions on the nuisance estimators μ_n and g_n.

Conditions (C1)–(C5) can be verified with routine albeit tedious work. Here, we focus on condition (C3), which allows us to obtain the scale parameter of the limit distribution, and on condition (C4), which requires that the nuisance estimators converge sufficiently fast. For condition (C4), we focus as before on the second-order remainder term $H_{x,n}^{(2)}$ given by

$$\begin{aligned} &\iint_{-\infty}^{x} \left[\mu_n(u,w) - \mu_0(u,w)\right] \left[\frac{g_n(u,w)}{g_0(u,w)} - 1\right] \Phi_0(du) Q_0(dw) \\ &\quad - \iint_{-\infty}^{x} \mu_n(u,w) \left(\Phi_n - \Phi_0\right)(du)(Q_n - Q_0)(dw) \ . \end{aligned}$$

The contribution to $K_n(\delta)$ of the first summand above is bounded above by

$$2\delta n^{1/3} \sup_{|x-u| \leq \delta n^{-1/3}} \left[f_0(u) \left\{ E_0 \left[\mu_n(u,W) - \mu_0(u,W)\right]^2 E_0 \left[\frac{g_0(u,W)}{g_n(u,W)} - 1 \right]^2 \right\}^{1/2} \right] .$$

which implies that condition (C4) is satisfied if, for some $\epsilon > 0$,

$$\sup_{|x-u| \leq \epsilon} E_0 \left[\mu_n(u,W) - \mu_0(u,W)\right]^2 \sup_{|x-u| \leq \epsilon} E_0 \left[\frac{g_0(u,W)}{g_n(u,W)} - 1 \right]^2 = o_P\left(n^{-1/3}\right) .$$

The contribution of the second summand to $K_n(\delta)$ can easily be controlled using empirical process theory. To scrutinize condition (C5), the relevant portion of the covariance function

$\Sigma_0(s,t)$ is given by

$$\iint_0^{s\wedge t} \left\{ \frac{\sigma_0^2(a,w)}{g_0(a,w)} + [\theta_0(a) - \theta_0(x)]^2 \right\} \Phi_0(da) Q_0(dw) \ ,$$

where $\sigma_0^2 : (a,w) \mapsto \mathrm{Var}_0(Y \mid A = a, W = w)$ denotes the conditional variance function of Y given A and W. Under certain smoothness conditions, we have that $\kappa_0(x) = f_0(x)^2 \int [\sigma_0^2(x,w)/f_0(x \mid w)]\, Q_0(dw)$, from which we find that the scale parameter of the limit Chernoff distribution to be

$$\tau_0(x) = \left\{ 4\nu_0'(x) \int \left[\frac{\sigma_0^2(x,W)}{f_0(x \mid W)} \right] Q_0(dw) \right\}^{1/3} .$$

The marginalized and marginal regression functions exactly coincide – that is, $\nu_0 = \mu_0$ – if, for example, (i) Y and W are conditionally independent given A, or (ii) A and W are independent. It is natural then to ask how the limit distribution of estimators of these two parameters compare under scenarios (i) and (ii), when the parameters in fact agree with each other. In scenario (i), the scale parameter obtained based on the estimator accounting for potential confounding reduces to

$$\begin{aligned} \tau_{0,red}(x) = & \left\{ 4\mu_0'(x)\sigma_0^2(x) \int \frac{Q_0(dw)}{f_0(x \mid w)} \right\}^{1/3} \\ & \geq \left\{ \frac{4\mu_0'(x)\sigma_0^2(x)}{\int f_0(x \mid w) Q_0(dw)} \right\}^{1/3} = \left\{ \frac{4\mu_0'(x)\sigma_0^2(x)}{f_0(x)} \right\}^{1/3} \end{aligned}$$

by Jensen's inequality. Thus, if Y and W are conditionally independent given A, in which case there is no need to adjust for potential confounders, the marginal isotonic regression estimator has a more concentrated limit distribution than the marginalized isotonic regression estimator. In scenario (ii), the scale parameter of the estimator accounting for potential confounding reduces to

$$\tau_{0,red}(x) \ = \ \left\{ \frac{4\mu_0'(x)}{f_0(x)} \int \sigma_0^2(x,W) Q_0(dw) \right\}^{1/3} \ \leq \ \left\{ \frac{4\mu_0'(x)\sigma_0^2(x)}{f_0(x)} \right\}^{1/3}$$

given that $\int \sigma_0^2(x,w)Q_0(dw) \leq \sigma_0^2(x)$ by the law of total variance. Thus, if A and W are independent, the marginal isotonic regression estimator has a less concentrated limit distribution than the marginalized isotonic regression estimator. In both scenarios (i) and (ii), the difference in concentration between the limit distributions of the two estimators varies with the amount of dependence between A and W. We note that these observations are analogous to those obtained in linear regression.

3.6 Simulation study

In this section, we report results from a small simulation study conducted to illustrate the large-sample results derived in Sections 3.3 and 3.4. Here, we consider Examples 1 and 2 from Section 3.5, namely estimation of a monotone density and hazard functions. Since the purpose of studying the cases without censoring or with independent censoring was to verify our general results in previously studied settings, our simulation is focused on the novel and more difficult scenario in which censoring is only conditionally independent. Through our simulation study, we wish to assess how well the finite-sample distribution of $n^{1/3}\left[\theta_n(x) - \theta_0(x)\right]$ approximates the limit distributions derived in the previous section.

Conditionally on a single covariate W distributed uniformly on the interval $(-1, +1)$, we consider the event and censoring times T and C to be independent and to each follow a Weibull distribution. Specifically, we take the conditional distribution of T given $W = w$ to be a Weibull distribution with shape parameter 4 and scale parameter $\exp(\alpha_0 + \alpha_1 w)$, while we take the conditional distribution of C given $W = w$ to be a Weibull distribution with shape parameter 2 and scale parameter $\exp(\beta_0 + \beta_1 w)$. We perform simulations under four distinct settings: (i) both T and C depend on W; (ii) only T depends on W; (iii) only C depends on W; and (iv) neither T nor C depend on W. To achieve this, in settings (i), (ii), (iii) and (iv), we set the vector $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ of parameters to be $(0.25, -0.375, 0.25, -0.75)$, $(0.25, -0.375, 1, 0)$, $(0.25, 0, 0.25, -0.75)$ and $(0.25, 0, 1, 0)$, respectively. We note that T and C follow proportional hazards models conditionally on W, and that the marginal density and hazard functions of T are monotone over the interval $[0, 1]$.

We used the generalized Grenander-type estimators proposed in the previous section to estimate the marginal density and hazard functions of T over $[0, 1]$ in each of the four simulation settings. First, we employed a naive procedure based on the Kaplan-Meier estimator of S_0, and second, we used a one-step procedure based on estimating the underlying conditional event and censoring hazard functions using a Cox model with single covariate W as main term only. We note that our goal differs from recent work on estimating a monotone baseline hazard (e.g., Lopuhaä and Nane, 2013a,b; Lopuhaä and Musta, 2017, 2018b). Our interest is in the marginal distribution of T rather than the conditional distribution of T given $W = 0$. Additionally, in principle, other consistent estimators of the conditional distributions of T and C given W could be used instead of Cox model-based estimators without changing the asymptotic results, as discussed in the previous section.

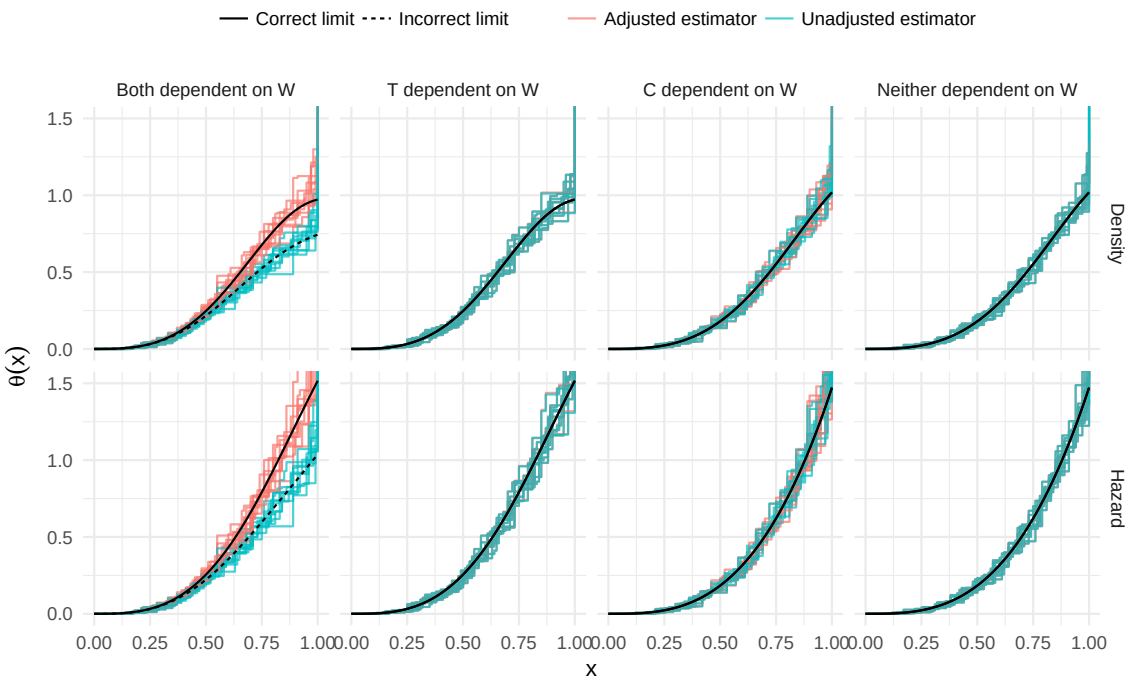

Figure 3.1: Estimated monotone density and hazard functions based on 10 realizations of datasets including 5000 right-censored observations. Solid black lines are the true density and hazard functions. Dotted black lines indicate limit of unadjusted estimators.

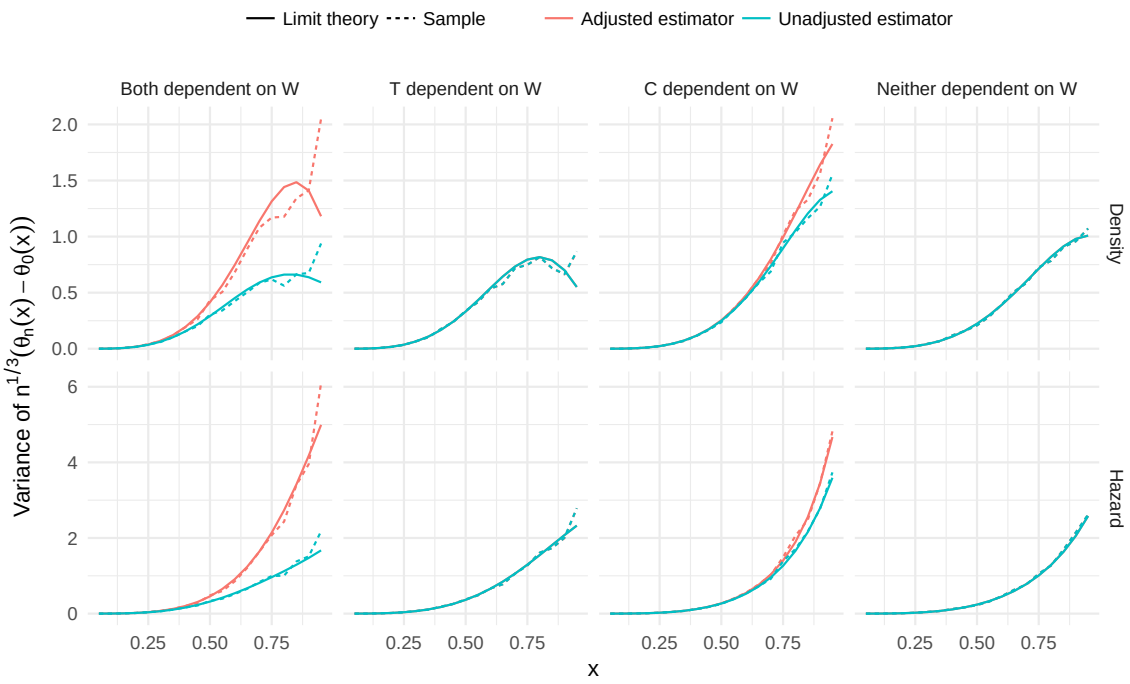

Figure 3.2: Empirical variance over 1000 simulations of the standardized monotone density and hazard estimators and theoretical variance of the corresponding Chernoff limit distribution.

The true density and hazard functions are plotted in Figure 3.1 along with an overlay of ten realizations of the estimator based on the naive and one-step procedures for estimating the marginal survival function S_0 based on random samples of size $n = 5000$. Realizations of the estimator based on the one-step procedure track the true marginal density and hazard functions of T over all four simulation settings, as expected. Realizations of the estimator based on the naive procedure also track the true marginal density and hazard functions of T for settings (ii) through (iv), since in each of these settings T and C are independent. However, in setting (i), the estimator based on the naive procedure is inconsistent. The limit of the estimators of the marginal density and hazard functions can be derived to be

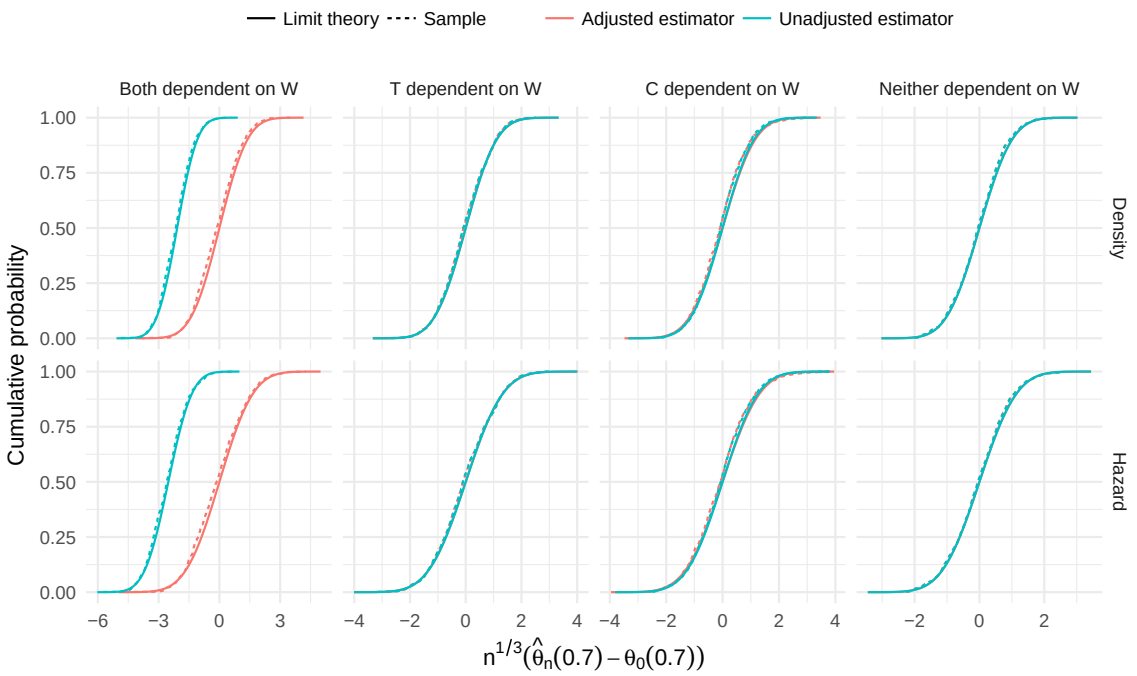

Figure 3.3: Sampling distribution over 1000 simulations of the monotone density and hazard estimators at $x = 0.7$ and the corresponding theoretical scaled Chernoff limit distribution.

the density and hazard functions, respectively, corresponding to the survival function

$$t \mapsto \exp\left[-\int_0^t \frac{\int f_0(u \mid w) G_0(u \mid w) Q_0(dw)}{\int S_0(u \mid w) G_0(u \mid w) Q_0(dw)}\, du\right].$$

These density and hazard functions are shown as black dotted lines in Figure 3.1.

In Figure 3.2, the empirical variance over 1000 simulations of $n^{1/3}\left[\theta_n(x) - \theta_0(x)\right]$ for $n = 5000$ is compared to the corresponding theoretical variances based on the limit theory we have presented in Section 3.5, for values of x between 0 and 1 and under the four considered scenarios. The sampling variance of the estimator appears close to the theoretical large-sample variance, except for x values near the upper boundary of the isotonizing interval. As expected, estimators based on the naive and one-step procedures have nearly identical sampling variances when only T is dependent on W (second column) and when neither T nor

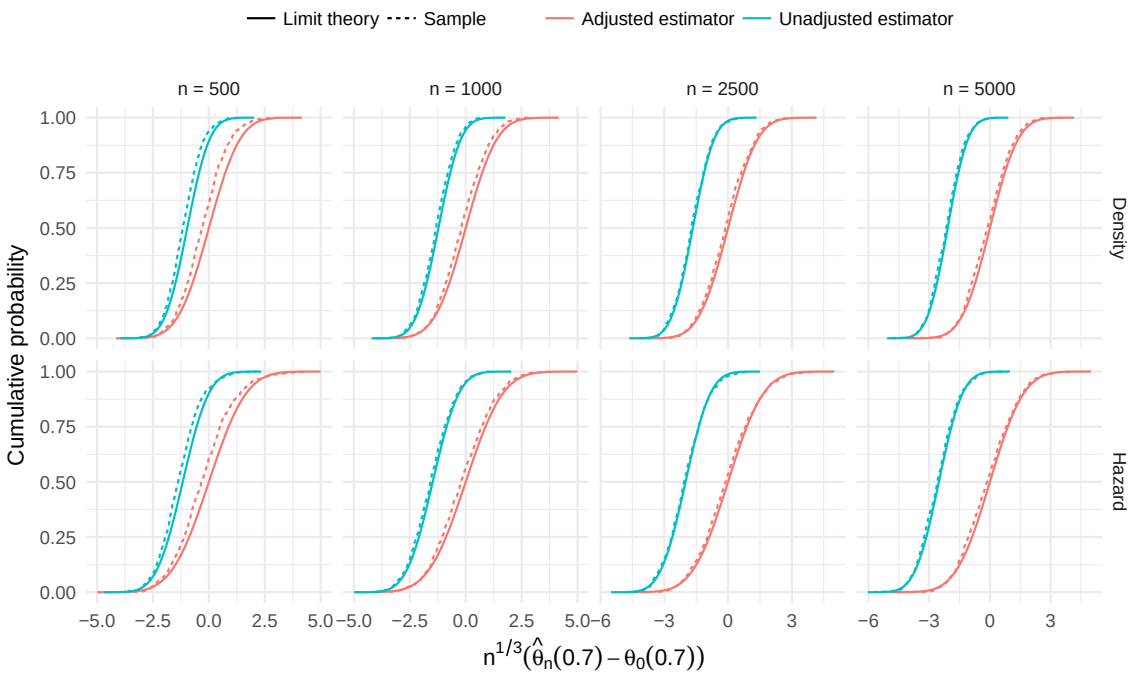

Figure 3.4: Sampling distribution over 1000 simulations of the monotone density and hazard estimators at $x = 0.7$ and the corresponding theoretical scaled Chernoff limit distribution. These figures are based on the scenario wherein T and C depend on W.

C are dependent on W (fourth column), but the sampling variance of the estimator based on the naive procedure is smaller than that based on the one-step procedure when only C is dependent on W (third column).

The empirical sampling distribution over 1000 simulations of $n^{1/3}\left[\theta_n(0.7) - \theta_0(0.7)\right]$ for $n = 5000$ is compared in Figure 3.3 to the theoretical scaled Chernoff limit distributions under the four different scenarios. In all situations, the sampling distribution approximates the theoretical limit. In the left-most columns, the bias of the estimator based on the naive procedure is evident. In Figure 3.4, the empirical sampling distribution of the estimators in settings where both T and C are dependent on W is plotted against the theoretical scaled Chernoff limit distribution for four different values of sample size n. At $n = 500$, the estimators are moderately biased downward, but as n increases, this bias vanishes.

Chapter 4

CAUSAL ISOTONIC REGRESSION

4.1 Introdocution

4.1.1 Motivation and literature review

In Chapter 3, we introduced the G-computed regression parameter $\theta_P(a) = E_P[E_P[Y \mid A = a, W]]$. (In this chapter, we change our notation from X to A and x to a in order to be consistent with the causal inference literature.) We proposed a generalized Grenander-type estimator of θ_0, and considered the key points of the asymptotic theory of this estimator. In this chapter, we again consider the G-computed regression parameter, but we go in to much greater detail about the properties of the proposed estimator.

We begin by motivating the G-computed regression function from the perspective of causal inference. Questions regarding the causal effect of an exposure on an outcome are ubiquitous in science. Many methods for statistical inference of causal effects address binary treatments. Specifically, suppose that a treatment A can either be received in full, corresponding to $A = 1$, or not at all, corresponding to $A = 0$, and that we are interested in the causal effect of receiving A on an outcome. In the language of potential outcomes, we let $Y(1)$ denote a unit's potential outcome if they were to receive the treatment, and $Y(0)$ their outcome if not (Neyman, 1923; Rubin, 1974). Throughout this article, we will make the Stable Unit Treatment Value Assumption: each unit's potential outcome is independent of all other units' treatment assignments, and there is only one version of the treatment (Cox, 1958). We may then aim to estimate $m(1) := E[Y(1)]$, the average over the population of the outcome were all units assigned to receive treatment, and $m(0) := E[Y(0)]$, the same were all units assigned to not receive treatment. The difference $m(1) - m(0)$, known as the average treatment effect, measures the average of the unit-level additive causal effect $Y(1) - Y(0)$.

The fundamental challenge of causal inference is that for each unit we only observe the outcome corresponding to a single treatment value. If A is the treatment actually received for a unit, we observe $Y := Y(A)$ corresponding to the potential outcome of this received treatment. For each $a \in \{0, 1\}$, whether and how we can obtain a valid estimator of $m(a)$ depends on the relationship between A and $Y(a)$. The simplest scenario is that A and $Y(a)$ are independent, in which case

$$m(a) := E[Y(a)] = E[Y(a) \mid A = a] = E[Y(A) \mid A = a] = E[Y \mid A = a] := \rho_P(a) \quad (4.1)$$

as long as $P(A = a) > 0$. We note that ρ_P is a mapping of the distribution P of the observed data, and we say that the causal quantity $m(a)$ is *identified with* the observed data mapping $\rho_P(a)$. We can estimate $\rho_P(a)$ with the observed data by averaging the observed values of Y in the cohort of units with $A = a$.

Independence of A and $Y(a)$ is a strong assumption, and typically only holds if treatment allocation is controlled by the researcher, as is true in randomized trials. In observational studies, A and $Y(a)$ are often dependent because there are common causes of A and $Y(a)$, which are also called *confounders* of the treatment-outcome relationship. In such situations, $m(a)$ does not generally equal $\rho_P(a)$, and simply conditioning on $A = a$ will not typically yield valid estimates or inference. However, if we observe *all* confounders, we may still be able to identify $m(a)$. Specifically, if A and $Y(a)$ are *conditionally* independent given an observed vector of covariates W, known as *unconfoundedness*, and additionally $P(A = a \mid W = w) \geq \tau > 0$ for all w in the support $\mathcal{W}$ of W, known as *positivity*, then

$$\begin{aligned} E[Y(a)] &= E[E[Y(a) \mid W]] = E[E[Y(a) \mid A = a, W]] \\ &= E[E[Y(A) \mid A = a, W]] = E[E[Y \mid A = a, W]]. \end{aligned} \quad (4.2)$$

The first equality follows from the tower property, the second from unconfoundedness and positivity, and the fourth from the assumption that $Y = Y(A)$, which is sometimes referred

to as *consistency*. The resulting identification $m(a) = \theta_P(a) := E[E[Y \mid A = a, W]]$ is known as the *backdoor formula*, *G-computation*, the *G-formula*, or *standardization* (Robins, 1986; Gill and Robins, 2001).

The G-formula (4.2) provides us with a way to estimate the causal quantity $m(a)$ using the observed data in the presence of observed confounding. If W is supported on a finite domain $\mathcal{W}$, we may estimate $\theta_P(a)$ by computing the average value of Y within each strata $(A = a, W = w)$, for w ranging over $\mathcal{W}$, then averaging these quantities according to the marginal distribution of W. However, if $\mathcal{W}$ is infinite, or its cardinality is large enough relative to the sample size that there are no or few observations in some strata of $\mathcal{W}$, such an approach is not feasible. If we have access to a correctly specified parametric model for the outcome regression function $\mu_P(a, w) := E[Y \mid A = a, W = w]$, we can use maximum likelihood estimation to obtain a statistically efficient estimator of $\theta_P(a)$ (Imbens, 2004). Alternatively, if we have access to a correctly specified parametric model for the propensity score $\pi_P(a \mid w) := P(A = a \mid W = w)$, we can use an inverse probability weighted estimator (Rosenbaum and Rubin, 1983).

Correctly-specified parametric models are not often available in practice, and using a mis-specified parametric model can lead to substantial estimation bias and invalid statistical inference. Hence, it is desirable to have estimators that do not rely on having correctly specified parametric models. Many nonparametric estimators for $\theta_P(a)$ utilize the fact that it is a *pathwise differentiable* parameter with respect to the nonparametric model, which we do not define here but instead refer the interested reader to Pfanzagl (1982), van der Vaart (1991), and Bickel et al. (1998). The pathwise differentiability of $\theta_P(a)$, and additional regularity conditions, imply that it is possible to construct asymptotically linear estimators of $\theta_P(a)$ that converge at a $n^{-1/2}$ rate, where n is the sample size, and moreover that it is possible to characterize the optimal asymptotic efficiency of any regular and asymptotically linear estimator of $\theta_P(a)$. Such a characterization was provided by Hahn (1998). Augmented inverse probability weighted estimators (van der Laan and Robins, 2003) and targeted maximum likelihood estimators (TMLE) (van der Laan and Rose, 2011) are two examples that achieve

the nonparametric efficiency bound when both the outcome regression and propensity score estimators converge to their true counterparts faster than $n^{-1/4}$. Furthermore, these estimators are *doubly-robust*, meaning they are consistent if *either* the outcome regression or propensity score estimators are consistent (Scharfstein et al., 1999; Bang and Robins, 2005).

So far, we have discussed causal inference with binary treatments. However, in many applied settings treatments of interest are continuous, in the sense that they can take any value in an interval $\mathcal{A} \subseteq \mathbb{R}$. In this case, we can again imagine the potential outcome $Y(a)$ under assigning a unit to treatment level $A = a$. Now, there are an infinite number of potential outcomes for each unit, and the corresponding average outcomes $m(a) := E[Y(a)]$ under assignment of the entire population to treatment $A = a$ over $a \in I$ forms a function from $\mathcal{A}$ to $\mathbb{R}$, which is sometimes called the *treatment effect curve* or *causal dose-response curve*.

Many of the principles stated above for identifying $m(a)$ with the observed data A and $Y = Y(A)$ in the case of a discrete treatment have direct analogues when A is continuous. If A and $Y(a)$ are independent and the conditional density of A exists and is positive in a neighborhood of $a \in I$, then $m(a) = \rho_P(a) = E[Y \mid A = a]$. If A and $Y(a)$ are conditionally independent given W and the conditional density of A given $W = w$ is bounded away from zero uniformly over $w \in \mathcal{W}$ and over a neighborhood of $a \in \mathcal{A}$, then the G-formula $m(a) = \theta_P(a) = E[E[Y \mid A = a, W]]$ still holds (Gill and Robins, 2001).

Although identification of $m(a)$ is similar when A is discrete and continuous, nonparametric estimation of the resulting identified parameters $\rho_P(a)$ and $\theta_P(a)$ is quite distinct in these two settings. The fundamental difference is that, when A is continuous, $\rho_P(a)$ and $\theta_P(a)$ are no longer pathwise differentiable parameters in the nonparametric model. Thus, many of the general tools for constructing asymptotically linear estimators, including estimating equations, one-step estimation, and TMLE, can not be immediately applied to nonparametric estimation of $\rho_P(a)$ and $\theta_P(a)$, and we do not expect to obtain estimators with $n^{-1/2}$ rates of convergence. Instead, we must turn to alternative estimation techniques.

Since $\rho_P(a)$ is a marginal regression function, it can be estimated nonparametrically

using a variety of strategies from nonparametric regression (see, e.g., Györfi et al., 2006). The Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964) and local polynomial estimators (Fan, 1996) are examples of nonparametric regression based on kernel smoothing. These approaches rely on assumed smoothness of $\rho_P(a)$ as a function of a, require selection of a kernel function and bandwidth, and the resulting estimators typically converge pointwise at the rate $n^{-2/5}$ to a normal distribution whose mean depends on the derivatives of $\rho_P(a)$.

In the presence of observed confounding, Kennedy et al. (2017) recently proposed a local linear estimator of $\theta_P(a)$ with many of the same properties as the local linear estimator of $\rho_P(a)$, including the $n^{-2/5}$ rate of convergence to a normal distribution with nonzero mean. Furthermore, Kennedy et al. (2017) showed that their estimator is doubly-robust. Additional relevant works include Rubin and van der Laan (2006) and Díaz Muñoz and van der Laan (2011), who discussed data-adaptive estimation of $\theta_P(a)$ but did not consider doubly-robustness or tools for inference, Robins (2000), who considered estimation of $\theta_P(a)$ with parametric models, Neugebauer and van der Laan (2007), who considered estimation and inference for a projection of $\theta_P(a)$ on to a parametric working model, and van der Laan et al. (2018), who presented a general framework for estimation and inference of non-pathwise differentiable parameters and addressed $\theta_P(a)$ as an example.

An alternative approach to nonparametric estimation of non-pathwise differentiable functions is to leverage shape constraints of the true function. For instance, if $m(a)$ is known to be monotone non-decreasing and $m(a) = \rho_P(a)$, then the isotonic regression of Y on A is the function minimizing the empirical risk corresponding to the squared-error loss function over all non-decreasing functions (Barlow et al., 1972). This estimator does not require selecting a kernel function or bandwidth, is invariant to strictly monotone transformations of A, and converges at the rate $n^{-1/3}$ to a non-normal, but log-concave and mean zero, limiting distribution (Brunk, 1970). We note that estimators of many other monotone parameters have been proposed and studied, including estimators of monotone density and hazard functions (see, e.g., Grenander, 1956, Prakasa Rao, 1970). Many of these estimators are presented and their properties studied in Groeneboom and Jongbloed (2014).

4.1.2 Contribution and organization of the chapter

In this chapter, we consider nonparametric estimation of the G-computed dose-response curve $\theta_P : a \mapsto E_P[E_P[Y \mid A = a, W]]$ with a continuous exposure A under the assumption that $\theta_0 := \theta_{P_0}$ is monotone as a function of a. This problem can be seen as an extension of classical isotonic regression to the setting where the exposure-outcome relationship is confounded by recorded covariates. Specifically, we aim to:

1. propose an estimator of $\theta_0(a)$ that extends the isotonic regression estimator;

2. investigate finite-sample and asymptotic properties of our estimator, including invariance to increasing and invertible transformations of A, doubly-robust consistency, and convergence at the rate $n^{-1/3}$ to a mean zero limit distribution; and

3. propose a method of performing doubly-robust pointwise inference for $\theta_0(a)$.

Our asymptotic results use the general results for generalized Grenander-type estimators developed in Chapter 3. In Chapter 3, we introduced the estimator we consider here as an example of our general approach to generalized Grenander-type estimators. We used our results to highlight particular elements of the asymptotic properties of this estimator, including sufficient conditions on the nuisance parameters for convergence in distribution of the monotone estimator, and we derived the scale parameter arising in the limit distribution under correct specification of both nuisance parameters. This comprises a small subset of the results we provide here. In particular, we provide a rigorous treatment of both consistency and convergence in distribution, including doubly-robust properties of the estimator, and a practical method for constructing pointwise confidence intervals.

We also note again the work of van der Vaart and van der Laan (2006), who studied Grenander-type estimation of a survival function from current status data with dependent censoring. To our knowledge, they were the first to consider Grenander-type estimation in a setting where constructing an efficient estimator of the primitive function requires estimating

unknown nuisance functions, and to derive sufficient conditions on the rates of convergence of the nuisance functions that guarantee $n^{-1/3}$-rate convergence in distribution of the monotone estimator.

Monotonicity of $\theta_P(a)$ is not always an appropriate assumption, but is in many circumstances reasonable for at least certain parts of the exposure domain. For example, in biomedical applications many drugs are known to have an increasing average effect on an outcome within a range of doses. Similarly, in public health, certain exposures, such as exercise and healthy eating, are known to have increasing average effects on health outcomes, while other behaviors such as smoking and drug abuse are known to have decreasing average effects. In economics and sociology, personal characteristics such as income, net worth, and years of education are known to have monotone relationships with some outcomes. In situations where prior scientific knowledge indicates that the dose-response curve is monotone, it is often of interest to know the exact values of the dose-response curve. For instance, we may wish to know how much of a drug is necessary to achieve a certain average response, or how much exercise per week is needed to reduce BMI by some amount over a certain period of time. These are the types of settings in which the methods we develop here can be applied.

The remainder of the chapter is organized as follows. In Section 4.2, we concretely define our estimator. In Section 4.3, we study finite-sample and asymptotic properties of our estimator. In Section 4.4, we propose two methods for performing pointwise inference for $\theta_P(a)$. In Section 4.5, we perform numerical studies to assess the performance of our estimator in finite samples, and in Section 4.6, we use our estimator to assess the effect of BMI on immune response in HIV vaccine trials. Proofs of all theorems are provided in the Appendix.

4.2 Proposed approach

4.2.1 Definitions and statistical setting

We observe n observations $O_1, \ldots, O_n$ sampled independently from a distribution P_0. A prototypical observation is composed of $O = (Y, A, W)$, where Y is a response, A is a continuous exposure, and W is a vector of covariates. We let $\mathcal{O} = \mathcal{Y} \times \mathcal{A} \times \mathcal{W}$ be the support of P_0, where $\mathcal{Y} \subseteq \mathbb{R}$, $\mathcal{A} \subseteq \mathbb{R}$ is an interval, and $\mathcal{W} \subseteq \mathbb{R}^p$.

Our parameter mapping of interest is the real-valued function $\{a \mapsto \theta_P(a) : a \in \mathcal{A}\}$, where $\theta_P(a) = E_P[E_P[Y \mid A = a, W]]$ and $\theta_0(a) := \theta_{P_0}(a)$. Throughout, we index parameter mappings by 0 for the evaluation of the parameter at P_0. We consider the setting where $\theta_0(a)$ is known to be a non-decreasing function of a on $\mathcal{A}$, and where the true marginal distribution function of A, $F_0(a) := P_0(A \leq a)$, is strictly increasing and continuous on $\mathcal{A}$. We denote by $\mathcal{M}$ the set of all probability distributions P supported on $\mathcal{O}$ satisfying these two conditions.

As discussed at length in the introduction, under untestable causal assumptions $\theta_0(a)$ is equal to the counterfactual mean of Y under assignment of the study population to exposure $A = a$ (Gill and Robins, 2001). Even if such assumptions cannot be expected to hold or the scientific question is not causal in nature, $\theta_0(a)$ is often of more interest than the marginal regression function $\rho_0(a) := E_0[Y \mid |A = a]$ when W is associated with both A and Y.

When $P_0(A = a) = 0$, $P \mapsto \theta_P(a)$ is not a pathwise differentiable parameter with respect to the nonparametric model at P_0 (Díaz Muñoz and van der Laan, 2011). As a result, many statistical methods for targeting Euclidean parameters in nonparametric and semiparametric models, such as estimating equations-based estimators, one-step estimators, generalized Newton-Raphson estimators, and targeted maximum likelihood estimators, cannot be used directly to target $\theta_P(a)$.

4.2.2 Review of isotonic regression

Since our estimator of $\theta_0(a)$ builds upon isotonic regression, we start with a review of the classical least-squares isotonic regression estimator of $\rho_0(a)$. The isotonic regression of Y on

A is the function ρ_n which minimizes the least-squares criterion $\sum_{i=1}^{n}[Y_i - \rho(A_i)]^2$ over all monotone non-decreasing functions ρ. It can be obtained via the Pool Adjacent Violators Algorithm (Ayer et al., 1955; Barlow et al., 1972), and it also has a representation using greatest convex minorants (GCMs). The GCM of a bounded function f on an interval $[a, b]$ is defined as the supremum over all convex functions g such that $g \leq f$. Then $\rho_n(a)$ can be shown to equal the left derivative, evaluated at $F_n(a)$ for F_n the empirical distribution function of $A_1, \ldots, A_n$, of the GCM over the interval $[0, 1]$ of the function obtained by linear interpolation of the *cusum diagram*

$$\left\{ \left(\frac{i}{n}, \frac{1}{n} \sum_{j=0}^{i} Y^*_{(j)} \right) : i = 0, 1, \ldots, n \right\},$$

where $Y^*_{(0)} := 0$ and $Y^*_{(j)}$ is the Y-value corresponding to the jth sorted A-value for $j \geq 1$.

The isotonic regression estimator ρ_n has many attractive properties. First, unlike smoothing-based estimators, isotonic regression does not require the choice of a kernel function, bandwidth, or any other tuning parameter. Second, it is invariant to increasing and invertible transformations of A. Specifically, if $H : \mathcal{A} \to \mathbb{R}$ is an increasing and invertible function and ρ^*_n is the isotonic regression of $Y_1, \ldots, Y_n$ on $H(A_1), \ldots, H(A_n)$, then $\rho^*_n = \rho_n \circ H^{-1}$. Third, ρ_n is uniformly consistent on any strict sub-interval of $\mathcal{A}$. Fourth,

$$n^{1/3}[\rho_n(a) - \rho_0(a)] \xrightarrow{\mathrm{d}} \left[4\rho_0'(a)\sigma_0^2(a)/f_0(a)\right]^{1/3} \mathbb{W}$$

for any a in the interior of $\mathcal{A}$ at which: $\rho_0'(a)$, $f_0(a) := F_0'(a)$, and $\sigma_0^2(a) := E_0[(Y - \rho_0(a))^2 \mid |A = a]$ exist and are positive and continuous in a neighborhood of a. Here, $\mathbb{W} := \mathrm{argmax}_{u \in \mathbb{R}}\{Z_0(u) - u^2\}$, where Z_0 denotes a two-sided Brownian motion originating from zero, and is said to follow *Chernoff's distribution*. The properties of Chernoff's distribution are well-studied (Chernoff, 1964; Groeneboom and Wellner, 2001); in particular it is mean-zero and has moments of all orders, is log-concave, and is well-approximated by a $N(0, 0.52)$ random variable. It appears frequently in the limit distribution of estimators of monotonicity-

constrained parameters.

4.2.3 Definition of our estimator

For $P \in \mathcal{M}$, we define $\mu_P(a, w) := E_P[Y \mid A = a, W = w]$ as the outcome regression function and $g_P(a, w) := \pi_P(a \mid w)/f_P(a)$, where $\pi_P(a \mid w) := \frac{d}{da}P(A \leq a \mid W = w)$ is the conditional density function of A given W and f_P is the marginal density function of A. We then define the *pseudo-outcome* $\xi_{\mu,g,Q}(y, a, w)$ as

$$\xi_{\mu,\pi,Q}(y, a, w) := \frac{y - \mu(a, w)}{g(a, w)} + \int \mu(a, w')Q(dw') \ .$$

As noted by Kennedy et al. (2017), $E_{P_0}[\xi_{\mu,g,Q_0}(Y, A, W) \mid A = a] = \theta_0(a)$ if *either* $\mu = \mu_0$ or $g = g_0$. They used this fact to motivate defining their estimator $\theta_{n,h}(a)$ as the local linear regression with bandwidth $h > 0$ of the pseudo-outcomes

$$\xi_{\mu_n,g_n,Q_n}(Y_1, A_1, W_1), \ldots, \xi_{\mu_n,g_n,Q_n}(Y_n, A_n, W_n)$$

on $A_1, \ldots, A_n$, where μ_n is an estimator of μ_0, g_n is an estimator of g_0, and Q_n is the empirical marginal distribution function of $W_1, \ldots, W_n$. The asymptotics of this nonparametric regression problem are non-standard since the pseudo-outcomes are dependent when the nuisance functions μ_n and g_n are estimated from the data. Nevertheless, Kennedy et al. (2017) showed that their estimator is consistent if either μ_n or g_n is consistent. They also showed that if both nuisance functions converge to their true counterparts fast enough, $h_n^* \sim n^{-1/5}$, and additional regularity conditions hold, then $n^{2/5}[\theta_{n,h_n^*}(a) - \theta_0(a)] \xrightarrow{\text{d}} N(b_0(a), v_0(a))$, where $b_0(a)$ is an asymptotic bias term depending on the second derivative of θ_0.

In our setting, θ_0 is known to be monotone. Therefore, instead of using a local linear regression to estimate the conditional mean of the pseudo-outcomes, we will define our estimator as the isotonic regression of the pseudo-outcomes on $A_1, \ldots, A_n$. Using the GCM representation of isotonic regression stated in the previous section, we can then summarize

our estimation procedure as follows:

1. Construct estimators μ_n, g_n of the outcome regression and $g_0 = \pi_0/f_0$.

2. Use these estimators to construct $\Gamma_n(a)$ for each a in the unique values of $A_1, \ldots, A_n$, where

$$\Gamma_n(a) := \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,a]}(A_i)\frac{Y_i - \mu_n(A_i, W_i)}{g_n(A_i, W_i)} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{(-\infty,a]}(A_i)\mu_n(A_i, W_j) \ . \quad (4.3)$$

3. Compute the GCM $\overline{\Psi}_n$ of the set of points $\{(i/n, \Gamma_n(i/n)) : i = 0, \ldots, n\}$ over $[0, 1]$.

4. Define $\theta_n(a)$ as the left derivative of $\overline{\Psi}_n$ evaluated at $F_n(a)$, for F_n the empirical distribution function of $A_1, \ldots, A_n$.

Much the same as Kennedy et al. (2017), while our estimator θ_n can be defined as an isotonic regression of a set of data-dependent pseudo-outcomes, the asymptotic properties of our estimator do not follow in a straightforward manner from classical results for isotonic regression because the pseudo-outcomes depend on the estimators μ_n, g_n, and Q_n, which themselves depend on all the observations $O_1, \ldots, O_n$. However, it turns out that our estimator is of the so-called *generalized Grenander-type*, and therefore we can use the general asymptotic results of Chapter 3 to study its asymptotic properties.

We first demonstrate that our estimator is a generalized Grenander-type estimator. We define $\psi_P := \theta_P \circ F_P^{-1}$, and we note that since θ_P and F_P^{-1} are increasing, ψ_P is as well. Therefore, the primitive function $\Psi_P(t) := \int_0^t \psi_P(u)\, du = \int_{-\infty}^{F_P^{-1}(t)} \theta_P(v)\, dF_P(v)$ is a convex function. Next, define $\Gamma_P := \Psi_P \circ F_P$, so that $\Gamma_P(a) = \int_{-\infty}^{a} \theta_P(u)\, F_P(du) = \int_{-\infty}^{a} \mu_P(u, w)\, F_P(du)\, Q_P(dw)$.

The parameter $\Gamma_P(a)$ is pathwise differentiable at P_0 in $\mathcal{M}$ for each a, and its nonparametric efficient influence function is given by

$$\phi^*_{P,a}(O) := I_{(-\infty,a]}(A)\frac{Y - \mu_P(A, W)}{g_P(A, W)} + \int_{-\infty}^{a} \mu_P(u, W)\, F_P(du) + I_{(-\infty,a]}(A)\theta_P(A) - 2\Gamma_P(a) \ .$$

Denoting $P_n = (\mu_n, g_n, F_n, Q_n)$, a one-step estimator of $\Gamma(a)$ is

$$\Gamma_n(a) = \Gamma_{P_n}(a) + \frac{1}{n}\sum_{i=1}^{n} \phi^*_{P_n,a}(O_i) \ .$$

Since $\int_{-\infty}^{a} \theta_{P_n}(u)\, F_n(du) = \Gamma_{P_n}(a)$, this one-step estimator is exactly equal to the estimator defined in (4.3). We then define $\Psi_n := \Gamma_n \circ F_n^-$ for F_n^- the empirical quantile function of A, as our estimator of Ψ_0, and ψ_n as the left derivative of the GCM of Ψ_n. We then have that $\theta_n = \psi_n \circ F_n$ is equivalent to the estimator defined in steps 1–4 above.

This representation allows us to use the general results of Chapter 3 to study the asymptotic behavior of our causal isotonic regression estimator θ_n. In the next section, we will discuss sufficient conditions on μ_n and g_n that yield consistency of $\theta_n(a)$, and convergence in distribution of $n^{1/3}[\theta_n(a) - \theta_0(a)]$. In practice, we recommend leveraging multiple candidate estimators using cross-validation to estimate μ_n and g_n (see, e.g. van der Laan et al., 2007), as we discuss more below.

If $\theta_0(a)$ were only known to be monotone on a fixed sub-interval $\mathcal{A}_0 \subset \mathcal{A}$, we would define $F_P(a) := P(A \leq a \mid A \in \mathcal{A}_0)$ as the marginal distribution function restricted to $\mathcal{A}_0$, and F_n as its empirical counterpart. Similarly, $I_{(-\infty,a]}(A_i)$ in (4.3) would be replaced with $I_{(-\infty,a]}(A_i) I_{\mathcal{A}_0}(A_i)$. In all other respects, our estimation procedure would remain the same.

4.3 Properties of the estimator

4.3.1 Generalization of least-squares isotonic regression

We have claimed that our estimator generalizes the least-squares isotonic regression estimator, and we start this section by explaining precisely what we mean. If it is known that (i) A is independent of W, then $g_0(a, w) = 1$ for all a in the support of A, so that we may take $g_n(a, w) = 1$ for all a and w. If, furthermore, it is known that (ii) Y is independent of W given A, then we may take an estimator μ_n satisfying $\mu_n(a, w) = \mu_n(a)$ for all a, w. Inserting $g_n = 1$ and any such μ_n in to (4.3), we can see that $\Gamma_n(a) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,a]}(A_i) Y_i$,

and it follows that $\theta_n(a) = \rho_n(a)$ for each a. Hence, in this case, our estimator reduces to least-squares isotonic regression.

We note that $\theta_0 = \rho_0$ if *either* of (i) or (ii) hold, but our estimator only reduces to the least-squares isotonic regression if *both* (i) and (ii) hold. This can be explained as follows. Suppose we believe that (i) holds, so we take $g_n = 1$, but we are wrong and in fact A is not independent of W. Then $\theta_0 \neq \rho_0$, so the marginal isotonic regression is not a consistent estimator of θ. However, by the doubly-robust properties of our estimator discussed below, if we correctly specify μ_n, our estimator θ_n will still be consistent for θ even though we have mis-specified g_n. Alternatively, suppose we believe (ii) holds, so we take $\mu_n(a, w) = \mu_n(a)$, but we are wrong and in fact Y is not conditionally independent of W given A. Then, $\theta_0 \neq \rho_0$, and as long as we correctly specify g_n, our estimator is still consistent for θ_0.

4.3.2 Invariance to invertible transformations of A

A second important feature of our estimator θ_n is that, as with the isotonic regression estimator, it is invariant to any increasing and invertible transformation of A. We let $V = H(A)$ for some increasing and invertible function $H : \mathcal{A} \to \mathbb{R}$, and define $\theta^*(v_0) := E_0[E_0[Y \mid V = v_0, W]] = \theta_0 \circ H^{-1}(v_0)$. We note that θ^* is also a monotone function. Let $\mu_0^*(v, w) = E_0[Y \mid V = v, W = w]$ and $g_0^*(v, w) = [\frac{d}{dv} P_0(V \leq v \mid W = w)]/[\frac{d}{dv} P_0(V \leq v)]$, and let μ_n^*, g_n^* be estimators of μ_0^* and g_0^*, respectively. If we apply the estimation procedure defined in the previous section to the data (Y, V, W), we arrive at the estimator $\theta_n^*(v_0) = \psi_n^* \circ F_n^*$, where $F_n^* = F_n \circ H^{-1}$ is the empirical distribution function of $V_1, \ldots, V_n$ and ψ_n^* is the left derivative of the GCM of $\Psi_n^* = \Gamma_n^* \circ F_n^{*-}$ for

$$\begin{aligned}
\Gamma_n^*(v_0) &= \frac{1}{n} \sum_{i=1}^{n} \left[I_{(-\infty, v_0]}(V_i) \frac{Y_i - \mu_n^*(V_i, W_i)}{g_n^*(V_i, W_i)} + \int_{-\infty}^{v_0} \mu_n^*(v, W_i)\, F_n^*(dv) \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[I_{(-\infty, H^{-1}(v_0)]}(A_i) \frac{Y_i - \mu_n^*(H(A_i), W_i)}{g_n^*(H(A_i), W_i)} + \int_{-\infty}^{H^{-1}(v_0)} \mu_n^*(H(a), W_i)\, F_n(da) \right] .
\end{aligned}$$

If it were the case that $\mu_n^*(H(a), w) = \mu_n(a, w)$ and $g_n^*(H(a), w) = g_n(a, w)$, then we would have $\Gamma_n^*(v_0) = \Gamma_n(H^{-1}(v_0))$, so that $\Psi_n^* = \Gamma_n \circ H^{-1} \circ H \circ F_n = \Psi_n$, and thus $\theta_n^* = \theta_n \circ H^{-1}$. In words, our estimator θ_n of θ would be invariant to H.

We can ensure that $\mu_n^*(H(a), w) = \mu_n(a, w)$ and $g_n^*(H(a), w) = g_n(a, w)$ in the following simple manner. Let $U_i := F_n(A_i)$, which is also equal to $F_n^*(V_i)$, and let $\bar{\mu}_n(u, w)$ be an estimator of the conditional mean of Y given $U_i = u, W_i = w$. Then let $\mu_n(a, w) = \bar{\mu}_n(F_n(a), w)$ and $\mu_n^*(v, w) = \bar{\mu}_n(F_n^*(v), w)$. Similarly, let $\bar{g}_n(u, w)$ be an estimator of the conditional density of $U = u$ given $W = w$, and let $g_n(a, w) = \bar{g}_n(F_n(a), w)$ and $g_n^*(v, w) = \bar{g}_n(F_n^*(v), w)$.

Invariance to increasing and invertible transformations of A is a desirable property because the scale of a continuous exposure is statistically arbitrary. For instance, if A is temperature, it should not matter whether A is measured in degrees Fahrenheit, degrees Celsius, or Kelvin, or if A is height, it should not matter whether A is measured in meters, feet, or some other unit of length. The parameters θ and θ^* encode exactly the same information about the effect of A on Y, after adjusting for W, so it is natural that an estimator should not depend on the scale of the exposure. More discussion of invariance can be found in Chapter 3 of Lehmann and Casella (2006).

4.3.3 Consistency

Next, we provide sufficient conditions for consistency of θ_n. Our conditions require controlling the uniform entropy of certain classes of functions. For a uniformly bounded class of functions $\mathcal{F}$, a finite discrete probability measure Q, and $\varepsilon > 0$, the covering number of $\mathcal{F}$ relative to the $L_2(Q)$ metric, written $N(\varepsilon, \mathcal{F}, L_2(Q))$, is defined as the smallest number of $L_2(Q)$-balls of radius less than or equal to ε needed to cover $\mathcal{F}$. The uniform entropy of $\mathcal{F}$ is then defined as $\log \sup_Q N(\varepsilon, \mathcal{F}, L_2(Q))$, where the supremum is taken over all finite discrete probability measures Q. For a thorough treatment of covering numbers and their role in empirical process theory, see, e.g., van der Vaart and Wellner (1996).

We now state our three sufficient conditions for consistency of θ_n.

(E1) There exist positive and finite constants c, δ, K_μ, K_g, and k_g, and a $V < 2$ such that, almost surely as $n \to \infty$, μ_n is contained in a class of functions $\mathcal{F}^{(\mu)}$ and g_n in a class $\mathcal{F}^{(g)}$ satisfying:

(a) $|\mu| \leq K_\mu$ for all $\mu \in \mathcal{F}^{(\mu)}$,

(b) $k_g \leq g \leq K_g$ for all $g \in \mathcal{F}^{(g)}$,

(c) for all $\varepsilon \leq \delta$, $\log \sup_Q N(\varepsilon, \mathcal{F}^{(\mu)}, L_2(Q)) \leq c\varepsilon^{-V/2}$, and

(d) for all $\varepsilon \leq \delta$, $\log \sup_Q N(\varepsilon, \mathcal{F}_n^{(g)}, L_2(Q)) \leq c\varepsilon^{-V}$.

(E2) There exist $\mu_\infty \in \mathcal{F}^{(\mu)}$ and $g_\infty \in \mathcal{F}^{(g)}$ such that $(F_0 \times Q_0)(\mu_n - \mu_\infty)^2 \xrightarrow{\mathrm{P}} 0$ and $(F_0 \times Q_0)(g_n - g_\infty)^2 \xrightarrow{\mathrm{P}} 0$.

(E3) There exist subsets $\mathcal{S}_1, \mathcal{S}_2$, and $\mathcal{S}_3$ of $\mathcal{A} \times \mathcal{W}$ such that $(F_0 \times Q_0)(\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3) = 1$ and where:

(a) $\mu_\infty(a, w) = \mu_0(a, w)$ for all $(a, w) \in \mathcal{S}_1$,

(b) $g_\infty(a, w) = g_0(a, w)$ for all $(a, w) \in \mathcal{S}_2$, and

(c) $\mu_\infty(a, w) = \mu_0(a, w)$ and $g_\infty(a, w) = g_0(a, w)$ for all $(a, w) \in \mathcal{S}_3$.

Under these three conditions we have the follow result.

Theorem 11 (Consistency)**.** *Suppose conditions (E1)–(E3) hold. Then* $\theta_n(a) \xrightarrow{\mathrm{P}} \theta_0(a)$ *for any* a *such that* $F_0(a) \in (0, 1)$*,* θ *is continuous at* a*, and* F_0 *is strictly increasing in a neighborhood of* a*. If* θ *is uniformly continuous and* F_0 *is strictly increasing on* $\mathcal{A}$ *then* $\sup_{a \in \mathcal{A}_0} |\theta_n(a) - \theta_0(a)| \xrightarrow{\mathrm{P}} 0$ *for any bounded strict subinterval* $\mathcal{A}_0$ *of* $\mathcal{A}$*.*

We note that in the pointwise statement of Theorem 11, $F_0(a)$ is required to be in the interior of $[0, 1]$, and similarly, the uniform statement of Theorem 11 only covers bounded strict subintervals of $\mathcal{A}$. This is due to the well-known issues with Grenander-type estimators at the boundaries of the domains. Various corrections to this issue have been proposed, and

it would be interesting to consider these in future work (see, e.g., Woodroofe and Sun, 1993; Balabdaoui et al., 2011; Kulikov and Lopuhaä, 2006).

Condition (E1) requires that μ_n and g_n are eventually contained in uniformly bounded function classes of manageable sizes. This condition is easily satisfied if, for instance, $\mathcal{F}^{(\mu)}$ and $\mathcal{F}^{(g)}$ are parametric classes. However, it is also satisfied for many types of semiparametric function classes. Uniform entropy bounds for many specific types of classes may be found in Chapter 2.6 of van der Vaart and Wellner (1996).

There is an asymmetry between the entropy requirements for $\mathcal{F}^{(\mu)}$ and $\mathcal{F}^{(g)}$ in items (c) and (d) of (E1). This is due to the term $\iint_{-\infty}^{a} \mu_n(a,w)\, F_n(da)\, Q_n(dw)$ in $\Gamma_n(a)$. To control this term, we use an upper bound from the theory of empirical U-processes of the form $\int_0^1 \log \sup_Q N(\varepsilon, \mathcal{F}^{(\mu)}, L_2(Q))\, d\varepsilon$ (Nolan and Pollard, 1987). Compare this to the uniform entropy integral $\int_0^1 [\log \sup_Q N(\varepsilon, \mathcal{F}, L_2(Q))]^{1/2}\, d\varepsilon$, which can be used for ordinary empirical processes indexed by a uniformly bounded class $\mathcal{F}$.

Cross-validating the observations used to train the nuisance estimators μ_n and g_n and the observations at which these estimators are evaluated in the definition of Γ_n would likely allow us to avoid the entropy conditions of (E1). However, while the asymptotic theory for such cross-validated estimators has been established for empirical processes indexed by a finite set (see, e.g., Zheng and van der Laan, 2011), it is has not, to our knowledge, been established for empirical processes indexed by infinite sets, as we would need in our results. We leave further considerations along these lines to future work.

Condition (E2) requires that μ_n and g_n are tending to limit functions μ_∞ and g_∞, and condition (E3) requires that at least one of $\mu_\infty(a,w) = \mu_0(a,w)$ or $g_\infty(a,w) = g_0(a,w)$ for $(F_0 \times Q_0)$-almost every (a,w). If either (i) $\mathcal{S}_1$ and $\mathcal{S}_3$ are null sets or (ii) $\mathcal{S}_2$ and $\mathcal{S}_3$ are null sets, then condition (E3) is known simply as *doubly-robustness* of the estimator θ_n to the two nuisance functions μ_0 and g_0, meaning we achieve consistency as long as either μ_n tends to μ_0 or g_n tends to g_0. Doubly-robust estimators are at this point a mainstay of causal inference and have been studied for over two decades (see, e.g., Robins et al., 1994; Rotnitzky et al., 1998; Scharfstein et al., 1999; van der Laan and Robins, 2003; Neugebauer

and van der Laan, 2005; Bang and Robins, 2005). However, (E3) is considerably more general than classical doubly-robustness, as it allows neither μ_n nor g_n to be correctly specified over the whole domain, as long as at least one of μ_n or g_n is correctly specified for almost every point in the domain. This generalization is similar to recent work on 2^K-*robust*, also known as *sequentially doubly robust*, estimators in longitudinal studies, where there are 2^K possible ways to achieve consistency for a longitudinal study with K time points (Luedtke et al., 2017; Rotnitzky et al., 2017). In our setting, there are an infinite number of ways to achieve consistency.

4.3.4 Convergence in distribution

Next, we study convergence in distribution of $n^{1/3}[\theta_n(a) - \theta_0(a)]$ for fixed a. We first define for any square-integrable functions $h_1, h_2 : \mathcal{A} \times \mathcal{W} \to \mathbb{R}$, $a \in \mathcal{A}$, $\varepsilon > 0$, and $\mathcal{S} \subseteq \mathcal{A} \times \mathcal{W}$:

$$d(h_1, h_2; a, \varepsilon, \mathcal{S})^2 := \sup_{|a-a| \leq \varepsilon} E_{Q_0} \left\{ I_{\mathcal{S}}(a, W) \left[h_1(a, W) - h_2(a, W) \right]^2 \right\}. \tag{4.4}$$

We also define $\sigma_0^2(a, w) := E_{P_0} \left\{ [Y - \mu_0(A, W)]^2 | A = a, W = w \right\}$ as the conditional variance of Y given $A = a$ and $W = w$. We now state two more conditions we will require:

(E4) There exists $\varepsilon^* > 0$ such that each of the following holds:

(a) $d(\mu_n, \mu_\infty; a, \varepsilon^*, \mathcal{S}_1) = \mathrm{o_P}(n^{-1/3})$;

(b) $d(\mu_n, \mu_\infty; a, \varepsilon^*, \mathcal{S}_2) \xrightarrow{\mathrm{P}} 0$;

(c) $d(g_n, g_\infty; a, \varepsilon^*, \mathcal{S}_1) \xrightarrow{\mathrm{P}} 0$;

(d) $d(g_n, g_\infty; a, \varepsilon^*, \mathcal{S}_2) = \mathrm{o_P}(n^{-1/3})$; and

(e) $d(\mu_n, \mu_\infty; a, \varepsilon^*, \mathcal{S}_3) d(g_n, g_\infty; a, \varepsilon^*, \mathcal{S}_3) = \mathrm{o_P}(n^{-1/3})$.

(E5) The functions $F_0, \mu_0, \mu_\infty, g_0, g_\infty$, and σ_0^2 are continuously differentiable in a in a neighborhood of $a = a$, uniformly over $w \in \mathcal{W}$.

Under (E1)–(E5), we have the following result regarding convergence in distribution of $n^{1/3}[\theta_n(a) - \theta_0(a)]$ for fixed a in the interior of $\mathcal{A}$.

Theorem 12 (Convergence in distribution)**.** *If conditions (E1)–(E5) hold for a such that* $0 < F_0(a) < 1$*, then*

$$n^{1/3}[\theta_n(a) - \theta_0(a)] \xrightarrow{\mathrm{d}} \left[4\frac{\theta_0'(a)\kappa_0(a)}{f_0(a)}\right]^{1/3} \mathbb{W}$$

where $\mathbb{W}$ *is Chernoff's distribution and*

$$\kappa_0(a) = E_{Q_0}\left\{E_{P_0}\left[\left(\frac{Y - \mu_\infty(A,W)}{g_\infty(A,W)} + \theta_\infty(A) - \theta_0(A)\right)^2 \middle| A = a, W\right] g_0(a,W)\right\}.$$

We remind the reader that Chernoff's distribution is defined as the point of maximum of a standard two-sided Brownian motion originating from zero minus a quadratic. We note the similarity of the limit distribution in Theorem 12 to the limit distribution of $n^{1/3}[\rho_n(a) - \rho_0(a)]$. In particular, as noted above, when either (i) Y and W are independent given A or (ii) A is independent of W, θ is equal to the marginal regression function ρ_0, so we can directly compare the limit distributions of θ_n and the marginal isotonic regression estimator ρ_n. Under correct specification of μ_0 and g_0, the limit distribution of ρ_n is more concentrated than that of θ_n in scenario (i), and less concentrated than that of θ_n in scenario (ii). This is analogous to results in linear regression, where including a covariate that is uncorrelated with the outcome inflates the standard error of the coefficient corresponding to A, while including a covariate that is correlated with the outcome but uncorrelated with A deflates the standard error of the coefficient corresponding to A.

Theorem 12 highlights an advantage gained by the monotonicity assumption over smoothing methods. As is often true with kernel smoothed estimators, the limit theory of Kennedy et al. (2017) has an asymptotic bias term depending on the second derivative of θ at a, meaning proper confidence intervals require under-smoothing. Since it is not known how to properly under-smooth for this problem, the confidence intervals ultimately used provide

asymptotically correct coverage for a *smoothed* parameter rather than the true parameter. In contrast, our estimator has a limit theory centered at the truth and hence avoids these complications. Additionally, smoothing methods require that θ be twice continuously differentiable at a, while Theorem 12 only requires a single continuous derivative.

Condition (E4) requires that, on the set $\mathcal{S}_1$ where μ_n is correctly specified but g_n is not, μ_n is converging to it's limit at a rate faster than $n^{-1/3}$ uniformly in a neighborhood of a, and similarly for g_n on the set $\mathcal{S}_2$. On the set $\mathcal{S}_3$ where both μ_n and g_n are correctly specified, (E4) only requires that the product of their rates of convergence is faster than $n^{-1/3}$. Hence, it is possible to attain doubly-robust inference for θ_n, meaning we can construct asymptotically valid confidence intervals and tests for θ based on θ_n as long as at least one of the nuisance estimators is converging faster than $n^{-1/3}$ to the true nuisance parameter, even if the other nuisance estimator is mis-specified. Furthermore, as with consistency, we allow that neither μ_n nor g_n is correctly specified everywhere, as long as at least one of them is correctly specified everywhere.

We contrast this possibility for doubly-robust inference with asymptotically linear estimators of pathwise differentiable parameters, in which case many standard estimators possess doubly-robust consistency, but *not* generally doubly-robust convergence in distribution. This is because doubly-robust convergence in distribution in the asymptotically linear setting would require that one of the nuisance parameters be converging at a rate faster than $n^{-1/2}$, which is not even achieved in correctly specified regular parametric models. The potential for doubly-robust convergence in distribution in our setting is produced by the weaker $o_P(n^{-1/3})$ convergence rate required for the correctly-specified nuisance parameter in (E4a) and (E4d). This is at least a feasible rate of convergence, though we note that attaining such a rate of convergence when W has two or more continuous components is only possible with specialized knowledge of the structural forms of g_0 or μ_0. If such additional knowledge is not available, (A4e) is much more realistic as long as both μ_n and g_n are correctly specified – for instance, these rates can be achieved in arbitrary dimension under the assumption of bounded variation norm (van der Laan and Benkeser, 2018). We note that specialized meth-

ods for doubly-robust convergence in distribution have been developed based on targeting higher-order bias terms (see, e.g., Benkeser et al., 2017).

If $\mu_\infty = \mu_0$, $\kappa_0(a)$ can be made arbitrarily small by taking $g_\infty(a, w)$ to infinity for all w. This can be explained as follows. Setting $g_\infty(a, w)$ to infinity implies that $g_n(a, w)$ is tending to infinity, and $g_n(a, w)$ equal to infinity corresponds to $\Gamma_n(a) = \iint_{-\infty}^{a} \mu_n(a, w)\, F_n(da)\, Q_n(dw)$. With this choice of Γ_n, the isotonic estimator θ_n we have provided is exactly equal to the isotonic regression of the plug-in estimator $a \mapsto \int \mu_n(a, w)\, Q_n(dw)$. Now, when $g_\infty \neq g_0$ is incorrectly specified, in order for the result of Theorem 12 to hold, $\int \mu_n(a, w)\, Q_0(dw)$ must be converging to $\theta_0(a)$ in probability for all a and at a rate faster than $n^{-1/3}$ uniformly in neighborhood of a, which implies, under the entropy condition (E1), the same for $\int \mu_n(a, w)\, Q_n(dw)$. Finally, since the isotonized estimator $\theta_n(a)$ is no farther from $\theta_0(a)$ than $\int \mu_n(a, w)\, Q_n(dw)$ is (Chernozhukov et al., 2009), $\theta_n(a)$ converges in probability to $\theta_0(a)$ at a rate faster than $n^{-1/3}$. Hence, with this choice of Γ_n, $n^{1/3}[\theta_n(a) - \theta_0(a)]$ converges in distribution to 0.

The preceding paragraph implies that an isotonized version of a plug-in estimator of θ_0 with a correctly specified estimator μ_n converging fast enough will converge faster than the estimator we have proposed. However, such an estimator μ_n is not possible without some additional prior knowledge about the form of μ_0, such as additional smoothness or a correctly specified semiparametric model. Furthermore, even if such additional knowledge were available, asymptotically unbiased inference in such models can still be a significant challenge. In contrast, Theorem 12 provides means to conduct asymptotically unbiased inference with much weaker smoothness or structural knowledge of μ_0.

4.3.5 Comparison with an un-transformed primitive

Our estimator of θ_n is based on first transforming the domain by the empirical distribution function of the exposure A, estimating the primitive function on this transformed scale, and finally transforming back to the original scale. An alternative procedure could omit the transformation step. That is, let $a_-, a_+ \in \mathbb{R}$ be fixed, and define $\Theta_P(a) = \int_{a_-}^{a} \theta_P(u)\, du$. For

$a \leq a_+$, the nonparametric efficient influence function of $\Theta_P(a)$ is

$$I_{(a_-,a]}(A)\frac{Y-\mu_0(A,W)}{p_0(A \mid W)} + \int_{a_-}^{a} \mu_0(u,W)\,du - \Theta(a) \ .$$

As with Γ_n, we can construct a one-step estimator $\Theta_n(a)$ of $\Theta(a)$. An alternative estimator $\bar{\theta}_n(a)$ can then be defined as the left derivative of the GCM of Θ_n over $[a_-, a_+]$.

It is natural to ask how $\bar{\theta}_n$ compares to the estimator θ_n we have studied thus far. First, we note that, unlike θ_n, $\bar{\theta}_n$ neither generalizes the classical isotonic regression estimator nor is invariant to monotone transformations of A. Additionally, utilizing the transformation F_0 fixes $[0,1]$ as the interval over which the GCM should be performed. If $\mathcal{A}$ is known to be a bounded set, $[a_-, a_+]$ can be taken as the endpoints of $\mathcal{A}$, but otherwise we need to choose the domain $[a_-, a_+]$ in defining $\bar{\theta}_n$.

Turning to an asymptotic analysis, using the results of Chapter 3, we could establish conditions very similar to (E1)–(E5) under which $n^{1/3}[\bar{\theta}_n(a) - \theta_0(a)] \xrightarrow{\mathrm{d}} [4\theta_0'(a)\bar{\kappa}_0(a)]^{1/3}\mathbb{W}$ for

$$\bar{\kappa}_0(a) = E_{Q_0}\left\{ E_{P_0}\left[\frac{(Y-\mu_\infty(A,W))^2}{\pi_\infty(A \mid W)^2} \middle| A = a, W \right] \pi_0(a \mid W) \right\} \ ,$$

where π_∞ is the limit of an estimator π_n of the conditional density function of A given W.

Let $[4\tau_0(a)]^{1/3}$ be the limit scaling factor of $n^{1/3}[\theta_n(a) - \theta_0(a)]$ and $[4\bar{\tau}_0(a)]^{1/3}$ be the limit scaling factor of $n^{1/3}[\bar{\theta}_n(a) - \theta_0(a)]$. If $\mu_\infty = \mu_0$ and $g_\infty(a,w) = p_\infty(a,w)/f_0(a)$, then $\tau_0(a) = \bar{\tau}_0(a)$, so that $n^{1/3}[\theta_n(a) - \theta_0(a)]$ and $n^{1/3}[\bar{\theta}_n(a) - \theta_0(a)]$ are converging to the same limit distribution. However, if $g_\infty = p_\infty/f_0 = g_0$ but $\mu_\infty \neq \mu_0$, the two estimators are not asymptotically equivalent. Letting $\sigma_0^2(a,w) = E_{P_0}[(Y-\mu_\infty(a,W))^2 \mid A = a, W = w]$, we have

$$\begin{aligned}
\tau_0(a) &= \theta_0'(a)\left\{ E_{Q_0}\left[\frac{\sigma_0^2(a,W)}{p_0(a \mid W)} \right] - \frac{(E_{Q_0}[\mu_\infty(a,W)] - \theta_0(a))^2}{f_0(a)} \right\} \\
\bar{\tau}_0(a) &= \theta_0'(a) E_{Q_0}\left[\frac{\sigma_0^2(a,W)}{p_0(a \mid W)} \right] \ .
\end{aligned}$$

Hence, when the outcome regression model is mis-specified we can potentially gain efficiency by utilizing the transformation, and the relative gain in efficiency is proportional to the degree of mis-specification of μ_∞. We note that this does not contradict Theorem 8 since if μ_n is mis-specified then Γ_n is not an asymptotically linear estimator.

4.3.6 Nuisance function estimation

Theorem 12 requires the estimator μ_n of the outcome regression function and the estimator g_n of the "standardized propensity" (the conditional density of A given W divided by the marginal density of A) be converging fast enough to their true counterparts in the discrepancy defined in (4.4). The required rate of convergence depends on whether both nuisance parameters are correctly specified, in which case the product of their rates needs to be $o_P(n^{-1/3})$, or just one is correctly specified, in which case its rate alone needs to be $o_P(n^{-1/3})$. These requirements would both be satisfied, for instance, if a correct parametric model were available for either of the nuisance parameters, but the premise of this article is that such correctly specified models are rarely available outside of randomized trials.

There are also various semiparametric methods that are able to attain these rates of convergence, including, for instance, additive models and index models (Horowitz, 2009). Nonparametric methods, such as kernel smoothing, are attractive because they do not rely on correct specification of a semiparametric model; however, as the number of continuous covariates contained in W increases, kernel smoothing methods require additional smoothness of the true function and the use of higher-order kernels to achieve the required rates of convergence. Under a mild bounded variation constraint, the nuisance parameters can be estimated at the rate $o_P(n^{-1/4})$ for an arbitrary number of continuous components of W (van der Laan and Benkeser, 2018), which, as discussed above, is sufficient if both nuisance parameters are correctly specified.

In practice, we recommend leveraging multiple candidate parametric, semiparametric, and nonparametric models using cross-validation. For estimation of the outcome regression estimator μ_n, this can be accomplished using a standard SuperLearner for a binary or con-

tinuous outcome, depending on whether Y is binary or continuous (van der Laan et al., 2007). For estimation of g_n, we recommend the following SuperLearner procedure. Define $U_0 := F_0(X)$, and recall that g_0 is equal to the conditional density of U_0 given W. Our estimator g_n is defined as an estimator of the conditional density of $U := F_n(X)$ given W. As with any cross-validation procedure, we begin by splitting the observations randomly in to V nearly-equally-sized folds. For each $v \in \{1, \ldots, V\}$, we call the observations in fold v the *test set* for fold v, and the remaining observations with fold not equal to v the *training set* for fold v.

First, for each $v \in \{1, \ldots, V\}$ we use any standard estimators of a conditional density function to estimate the conditional density of U given W in the training set for fold v, and obtain the out-of-sample predicted conditional densities in the test set for fold v. For instance, these standard estimators may include parametric models, kernel conditional density estimators, or other non- or semiparametric estimators. We may also choose to use the *location-scale* method employed in Kennedy et al. (2017), wherein first a SuperLearner is used to estimate the conditional mean of U given W, then a SuperLearner is used to estimate the conditional variance of U given W using the squared residuals of the first SuperLearner, and finally a univariate kernel density estimator is applied to the standardized residuals of the result.

Second, we use a modification of the method introduced in Díaz Muñoz and van der Laan (2011) as an additional nonparametric estimator. Let $B \geq 2$ be a pre-defined maximal number of bins. For each fold v and number of bins $b \in \{2, \ldots, B\}$, we split the unit interval $[0, 1]$ in to b sub-intervals $\{(0 = u_0, u_1], (u_1, u_2], \ldots, (u_{b-1}, u_b = 1]\}$ using the empirical quantiles of the values of A in the training set for fold v. Then, for each $j \in \{1, \ldots, b\}$, we estimate the probability that $U \in (u_{j-1}, u_j]$ using any number of standard algorithms for estimating a conditional mean with a binary outcome. For each such algorithm, we scale the estimated probabilities to sum to one. Then, we define the estimated conditional density at (u, w) as the estimated probability that $U \in (u_{j-1}, u_j]$ given $W = w$ divided by $u_j - u_{j-1}$, where $(u_{j-1}, u_j]$ is the interval containing u.

We call the collection of all conditional density estimators we consider from these two steps our *library*. We denote by K the total number of estimators in our library, and for each $k \in \{1, \ldots, K\}$, we let $g_{n,k}(U_i, W_i)$ be the out-of-sample estimate for estimator k and observation i. We obtain our final estimator by minimizing the empirical risk corresponding to the log-likelihood loss over all candidate estimators formed by convex combinations of the V estimators in our library. That is, we take $g_n := \sum_{k=1}^{K} \lambda_{n,k}^* g_{n,k}$, where

$$\lambda_n^* := \underset{\lambda \in \Lambda_K}{\operatorname{argmin}} \sum_{i=1}^{n} \log \sum_{k=1}^{K} \lambda_k g_{n,k}(U_i, W_i) \tag{4.5}$$

for Λ_K the K-dimensional simplex.

4.4 Pointwise asymptotic inference

4.4.1 Wald-type confidence intervals

Let $\tau_0(a) = \theta_0'(a)\kappa_0(a)/f_0(a)$. If $\tau_n(a)$ were a consistent estimator of $\tau_0(a)$, a Wald-type $1-\alpha$ level asymptotic confidence interval for $\theta_0(a)$ would be

$$\left[\theta_n(a) - \{4\tau_n(a)/n\}^{1/3} q_{\alpha/2}, \quad \theta_n(a) + \{4\tau_n(a)/n\}^{1/3} q_{1-\alpha/2}\right],$$

where q_p denotes the p-the quantile of $\mathbb{W}$. The quantiles of Chernoff's distribution $\mathbb{W}$ have been numerically computed and tabulated on a fine grid (Groeneboom and Wellner, 2001), and in particular are readily available in the statistical programming language `R`. Hence, we can construct approximate Wald-type intervals for $\theta_0(a)$ with correct asymptotic coverage given a consistent estimator of $\tau_0(a)$.

First, we note that we can write $\theta_0'(a)/f_0(a) = \psi_0'(F_0(a))$, where $\psi_0 = \theta_0 \circ F_0^{-1}$. Estimating ψ_0' rather than θ_0' and f_0 preserves our theme of providing an estimator invariant to monotone transformations of A. Hence, estimation of τ_0 can be decomposed in to (i) estimation of ψ_0' and (ii) estimation of κ.

For (i), we recall that θ_n is defined as $\psi_n \circ F_n$, and that ψ_n as defined in Section 4.2 is a

step function, and hence cannot be used directly to provide an estimator of ψ_0'. Instead, one approach to estimating ψ_0' is to smooth the function ψ_n and to define ψ_n' as the derivative of this smoothed estimator. We have found that a local quadratic kernel smoother of the points $\{(u_j, \psi_n(u_j)) : j = 1, \dots K\}$ for u_j the midpoints of the jump points of ψ_n, works well in practice.

4.4.2 Inference under correct nuisance specification

The remainder of this section is concerned with task (ii): estimation of $\kappa_0(a)$. First, we suppose that g_n and μ_n are correctly specified, so that $g_\infty = g_0$ and $\mu_\infty = \mu_0$. In this case, $\kappa_0(a) = E_{Q_0}[\sigma_0^2(a, W)/g_0(a, W)]$. One approach to estimating $\kappa_0(a)$ in this case is to use machine learning techniques to estimate the conditional expectation of $Z := [Y - \mu_n(A, W)]^2$ given A and W, which gives an estimator $\sigma_n^2(a, w)$ of $\sigma_0^2(a, w)$. We can then define a plug-in estimator of $\kappa_0(a)$ as

$$\kappa_{n,\mathrm{PI}}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sigma_n^2(a, W_i)}{g_n(a, W_i)}.$$

In practice, we will typically want to use an ensemble method such as SuperLearner (van der Laan et al., 2007) to increase the chance that we correctly specify the functional form of σ_0^2. If we do manage to correctly specify σ_n^2, then $\kappa_{n,\mathrm{PI}}(a, w)$ is a consistent estimator of $\kappa_0(a)$ under correct specification of μ_n and g_n.

In the special case of a binary outcome, $\sigma_0^2(a, w) = \mu_0(a, w)(1 - \mu_0(a, w))$, so we can set $\sigma_n^2 = \mu_n(1 - \mu_n)$. Hence, for a binary outcome with both models correctly specified, we do not have to perform any additional modeling to consistently estimate $\kappa_0(a)$.

4.4.3 Inference under possibly incorrect nuisance specification

As noted above, Theorem 12 provides a limit distribution even if one of the nuisance estimators is not correctly specified, as long as the other estimator is converging fast enough to the truth. It would be desirable to capitalize on this result by providing an estimator of $\tau_0(a)$ that is doubly-robust. Since ψ_n is a doubly-robust estimator of ψ_0, our estimator ψ_n' of the

ψ_0' is also doubly-robust. However, the estimator of $\kappa_0(a)$ that we described above assumed that both μ_n and g_n were correctly specified.

In order to define an estimator of κ that is consistent even if one of μ_∞ or g_∞ is not equal to its true counterpart, we note that $\kappa_0(a) = \lim_{h \downarrow 0} h^{-1} E_{P_0}\left[K\left(\frac{F_0(A)-F_0(a)}{h}\right)\eta(Y,A,W)\right]$ for any bounded, nonnegative kernel function K with bounded support and integrating to one, where

$$\eta(y,a,w) = \left[\frac{y-\mu_\infty(a,w)}{g_\infty(a,w)} + \theta_\infty(a) - \theta_0(a)\right]^2.$$

Letting $\theta_{\mu_n}(a) = \int \mu_n(a,w)\, Q_n(dw)$, we define $\kappa_{n,h}(a) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{F_n(A_i)-F_n(a)}{h}\right)\eta_n(Y_i,A_i,W_i)$ for

$$\eta_n(y,a,w) = \left[\frac{y-\mu_n(a,w)}{g_n(a,w)} + \theta_{\mu_n}(a) - \theta_n(a)\right]^2.$$

Then, under conditions (E1)–(E5) we have that $\kappa_{n,h_n}(a) \xrightarrow{\mathrm{P}} \kappa_0(a)$ by standard kernel smoothing arguments for any sequence $h_n \to 0$. In particular, $\kappa_{n,h_n}(a) \xrightarrow{\mathrm{P}} \kappa_0(a)$ under the very general form of doubly-robustness specified by condition (E3).

To choose the bandwidth h based on the data, we will use a method motivated by the mean integrated square error criterion from kernel smoothing. We define $\gamma := \int \kappa_0(a)\, F_0(da)$, and note that $\gamma = E_{P_0}[\eta(Y,A,W)]$. Thus, we also have $\gamma = \mathrm{argmin}_\gamma\, E_{P_0}\{[\eta(Y,A,W)-\gamma]^2\}$. We thus define $h_n^* := \mathrm{argmin}_h \frac{1}{n}\sum_{i=1}^n [\eta_n(Y,A,W) - \gamma_n(h)]^2$ for $\gamma_n(h) := \int \kappa_{n,h}(a)\, F_n(da)$. We further improve this method by cross-validating the observations used to compute $\kappa_{n,h}(a)$ and those used to compute the risk. Hence, our doubly-robust estimator of $\kappa_0(a)$ is $\kappa_{n,\mathrm{DR}}(a) := \kappa_{n,h_n^*}(a)$.

We conclude this section with two final remarks regarding this doubly-robust estimator of $\kappa_0(a)$. First, we note that as with our estimator of $\psi_0'(u_0)$, our estimator of $\kappa_0(a)$ only depends on A and a through the order statistics $F_n(A)$ and $F_n(a)$. Hence, as before, our estimator is invariant to monotone transformations of A. Second, we note that if $\mu_n(a,w) = \mu_n(a)$ does not depend on w and $g_n = 1$, the doubly-robust estimator of $\kappa_0(a)$ converges to the marginal conditional variance $\mathrm{Var}(Y \mid A = a)$, which is the scale parameter for regular

isotonic regression. Hence, our doubly-robust estimator of $\kappa_0(a)$ provides asymptotically valid coverage for the marginal regression function in the same setting in which our estimator reduces to the marginal isotonic regression function.

4.5 Numerical studies

In this section we perform numerical experiments to assess the performance of the estimator of $\theta_0(a)$ we have introduced and the two methods of pointwise inference we have proposed.

We simulated data as follows. First, we generated $W \in \mathbb{R}^4$ as independent standard normal deviates. Next, we generated U given W from the conditional distribution corresponding to the conditional density function $g_0(u, w) = \lambda(w) + 2(1 - \lambda(w))u$ for $\lambda(w) = 0.1 + 1.8 \times \text{logit}^{-1}(\beta^T w)$, and we then defined A as the standard normal quantile function of U. Since U thus defined is marginally uniform, A is marginally standard normal. Finally, conditionally upon A and W, we simulated Y as a Bernoulli random variate with conditional mean function given by $\mu_0(a, w) = \text{logit}^{-1}\left(\gamma_1^T \bar{w} + \gamma_2^T \bar{w} a + \gamma_3 a^2\right)$, where $\bar{w}$ denotes $(1, w)$. We set $\beta = (-1, -1, 1, 1)^T$, $\gamma_1 = (-1, -1, -1, 1, 1)^T$, $\gamma_2 = (3, -1, -1, 1, 1)^T$ and $\gamma_3 = 3$.

We used the causal isotonic regression estimator θ_n to estimate the true covariate-adjusted dose-response curve, which requires specifying nuisance estimators μ_n and g_n. We considered four combinations of nuisance estimators. First, we correctly specified both nuisance estimators by taking μ_n and g_n as maximum likelihood estimators of μ_0 and g_0. Next, we kept μ_n correctly specified, but incorrectly specified g_n by setting $g_n(a, w) = 1$ for all a and w. Third, we correctly specified g_n, but and incorrectly specified μ_n by setting $\mu_n(a, w) = 0$ for all a and w. Finally, we incorrectly specified both μ_n and g_n by setting $\mu_n = 0$ and $g_n = 1$. We constructed pointwise confidence intervals in each setting using both the plug-in and doubly-robust estimators of κ_0.

As noted previously, in the first three settings, we expect θ_n to be consistent for θ_0. We also expect the doubly-robust estimator of κ_0 to provide intervals with asymptotically correct coverage rates of θ_0 for each of the first three settings, but only expect the plug-in estimator of κ_0 to provide intervals with asymptotically correct coverage rates of θ_0 in the first setting.

In the fourth setting, θ_n reduces to the marginal isotonic regression estimator ρ_n, which is consistent for the marginal regression function ρ_0. In this case, we expect the doubly-robust estimator of κ to provide asymptotically correct coverage of ρ_0.

The left panel of Figure 4.1 shows a single causal isotonic regression estimate based on a sample of size $n = 5000$ with μ_n and g_n correctly specified. Also shown are asymptotic 95% pointwise confidence intervals constructed using the doubly-robust estimator of κ_0. The right panel shows the marginal isotonic regression estimate of the same data and 95% asymptotic confidence intervals. The true causal and marginal regression curves are shown in red. We note that $\theta_0(a) \neq \rho_0(a)$ for $a \neq 0$, since both Y and A are dependent on W. Therefore, the marginal isotonic regression estimator will not be consistent for the true causal parameter. In this data-generating setting, the causal effect of A on Y is larger in magnitude than the marginal effect of A on Y, in the sense that $\theta_0(a)$ has larger variation than $\rho_0(a)$.

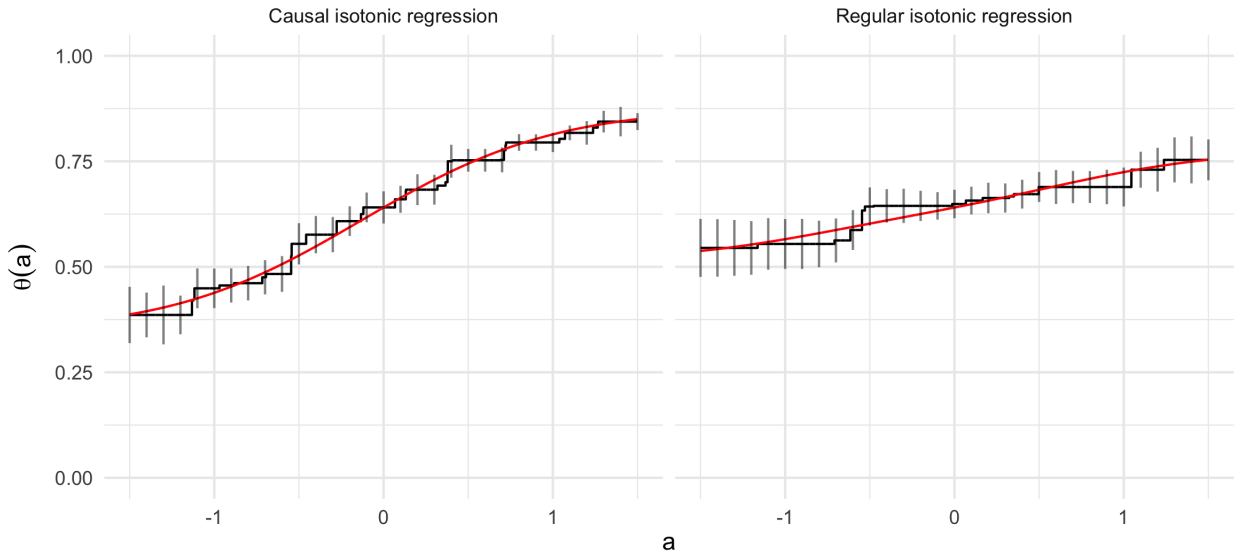

Figure 4.1: Causal isotonic regression estimate with correctly specified μ_n and g_n (left) and regular isotonic regression estimate (right). Pointwise 95% confidence intervals constructed using the doubly-robust estimator are shown as vertical bars and the true functions are shown in red.

We performed 1000 simulations each with $n = 500, 1000, 2500, 5000,$, and 10000 obser-

vations. Figure 4.2 shows the root mean squared error (RMSE) of the four estimators for different values a over these 1000 simulations and for each of these sample sizes. We observe that, for $a = 1$, correctly specified μ_n yields smaller RMSE than incorrectly specified μ_n when g_n is correctly specified, and that correctly specified g_n yields smaller RMSE than incorrectly specified g_n when μ_n is correctly specified. The improvement in RMSE in the former case is substantially larger than the improvement in the latter case.

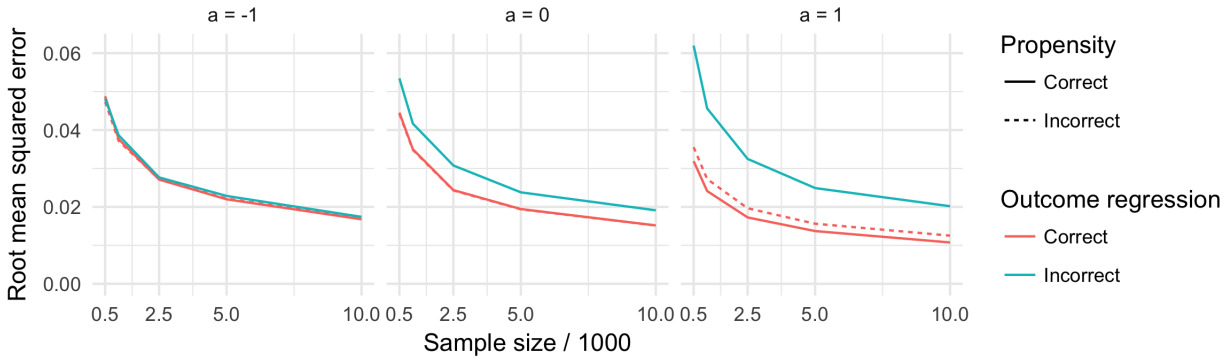

Figure 4.2: Root mean squared errors (RMSE) for the estimator proposed here for different values of a and different specifications of the outcome regression μ_n and the propensity g_n over one thousand data sets simulated as described in the text. In the left-most panel, all three lines overlap. In the middle panel, the two red lines overlap.

Figure 4.3 shows the coverage of nominal 95% pointwise confidence intervals for a range of values of a. As expected, the coverage improves as n grows, especially for values of a in the tails of the marginal standard normal distribution of A. Under correct specification, the plug-in method attains close to nominal coverage rates for a between -1 and 1 by $n = 1000$, and is anti-conservative otherwise. When the propensity is incorrectly specified, the plug-in method still performs well in this example, although we would not expect this to always be the case. Coverage is not shown when $\mu_n = 0$, i.e. when the outcome regression is incorrectly specified, because $\kappa_{n,\mathrm{PI}} = 0$ by necessity in this case and hence coverage is exactly zero.

Figure 4.3 also shows the coverage of nominal 95% pointwise doubly-robust confidence intervals. The doubly-robust method attains close to nominal coverage for large samples

under all specification settings. Unsurprisingly, compared to the plug-in method, it requires larger sample sizes to achieve good coverage, especially for values of a in the tails of the marginal distribution of A. In this example, the doubly-robust method appears to perform better under mis-specification of μ_n because setting μ_n to a constant removes all randomness from μ_n. The bottom-right panel shows the coverage of the marginal regression function $\rho_0(a)$. In studying the bottom-right panel, we also observe that the doubly-robust method obtains correct asymptotic coverage of $\rho_0(a)$ when $\mu_n = 0$ and $g_n = 1$. We emphasize that the bottom right panel does not display coverage of these confidence intervals for the causal parameter θ – the coverage of the intervals for the causal parameter would not be close to the nominal 95%.

Figure 4.4 presents boxplots of the estimator $\psi_n'(a)$ of the true derivative $\psi_0'(a)$ for each of the nuisance model specifications. The estimators are taken to the one-third power because that is what appears in the estimator of the pointwise confidence intervals. The estimators are roughly centered around the truth (shown in red).

Figure 4.5 shows histograms of the plug-in estimator of $\kappa_0(a)$ for μ_n correctly specified. The estimators are taken to the one-third power because that is what appears in the estimator of the pointwise confidence intervals. The estimators are centered around the truth (show in red) when g_n is correctly specified, but are biased when g_n is misspecified.

Figure 4.6 shows histograms of the doubly-robust estimator of $\kappa_0(a)$. Once again, the estimators are taken to the one-third power because that is what appears in the estimator of the pointwise confidence intervals. The estimators are roughly centered around the truth, which is shown in red, in all settings, though the variance of these estimators is larger than it is for the plug-in estimator.

We also conducted simulations using machine learning estimators for μ_n and g_n. To correctly specify μ_n and g_n, we used the methods described in Section 4.3.6 applied to the four covariates $W_1, \ldots, W_4$. To incorrectly specify μ_n and g_n, we again used the methods described in Section 4.3.6, but applied only to the covariates W_1 and W_2. Due to computational limitations, we performed 1000 simulations only for the sample size $n = 1000$.

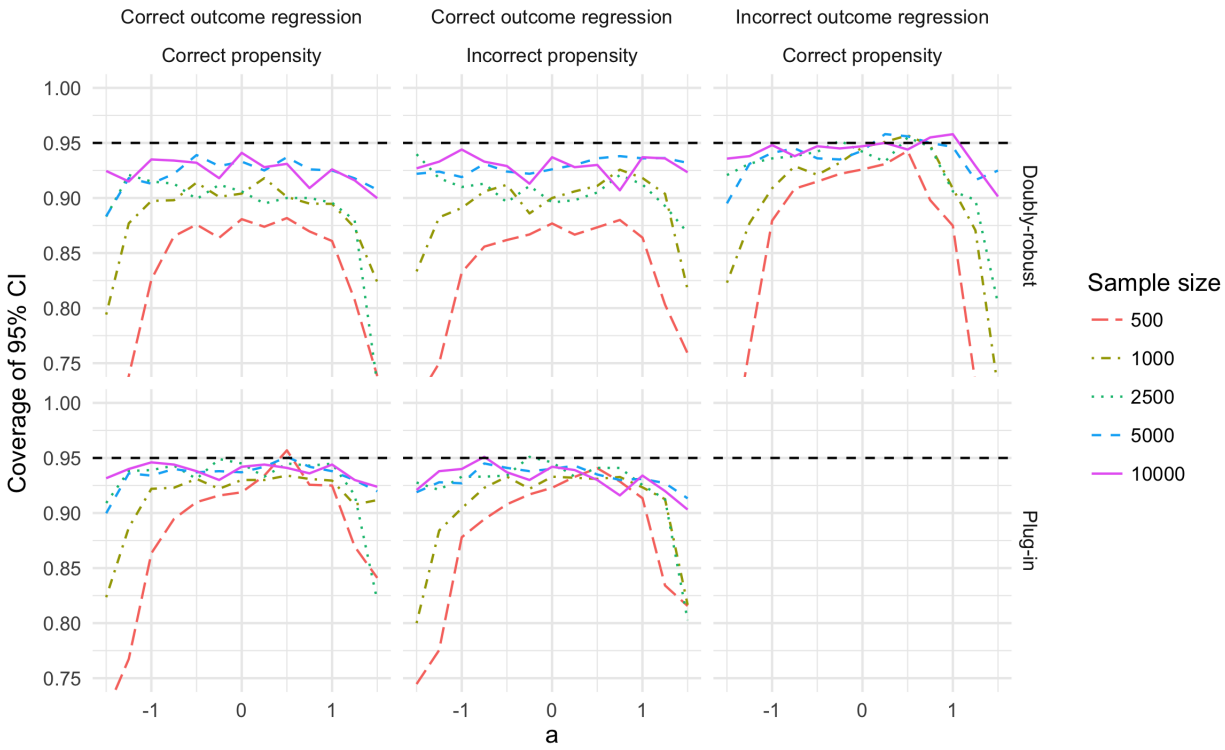

Figure 4.3: Coverage of pointwise 95% doubly-robust (top row) and plug-in (bottom row) confidence intervals for different values of a over one thousand data sets simulated as described in the text. Columns indicate whether the models for μ and g were correctly specified. Correct specifications used parametric models; μ was mis-specified by setting it to zero, g was incorrectly specified by setting it to one. Coverage is zero for the plug-in method with incorrect μ (lower left panel). Black dashed lines indicate the nominal coverage rate.

Figure 4.7 shows the coverage of nominal 95% confidence intervals using both the plug-in and doubly-robust methods. The plug-in intervals achieve very close to nominal coverage under correct specification, and also achieve surprisingly good coverage rates when the propensity is incorrectly specified. The plug-in intervals are somewhat conservative when the outcome regression is incorrectly specified. The doubly-robust method is anti-conservative under both correct specification and when the propensity is incorrectly specified, with coverage rates mostly between 90% and 95%. It achieves good coverages rates when the outcome regression is incorrectly specified. These results suggest that the doubly-robust intervals may

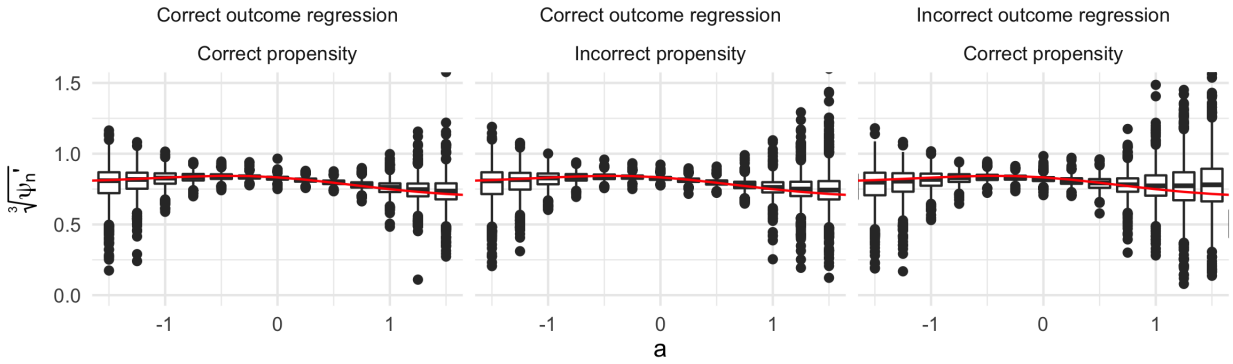

Figure 4.4: Distribution of the estimator $\psi_n'(a)$ of $\psi_0'(a)$ for different values of a and specifications of the nuisance parameters over one thousand data sets simulated as described in the text with $n = 5000$ observations. Note that the bottom-right panel corresponds to regular isotonic regression. Red lines show the true values $\psi_0'(a)$.

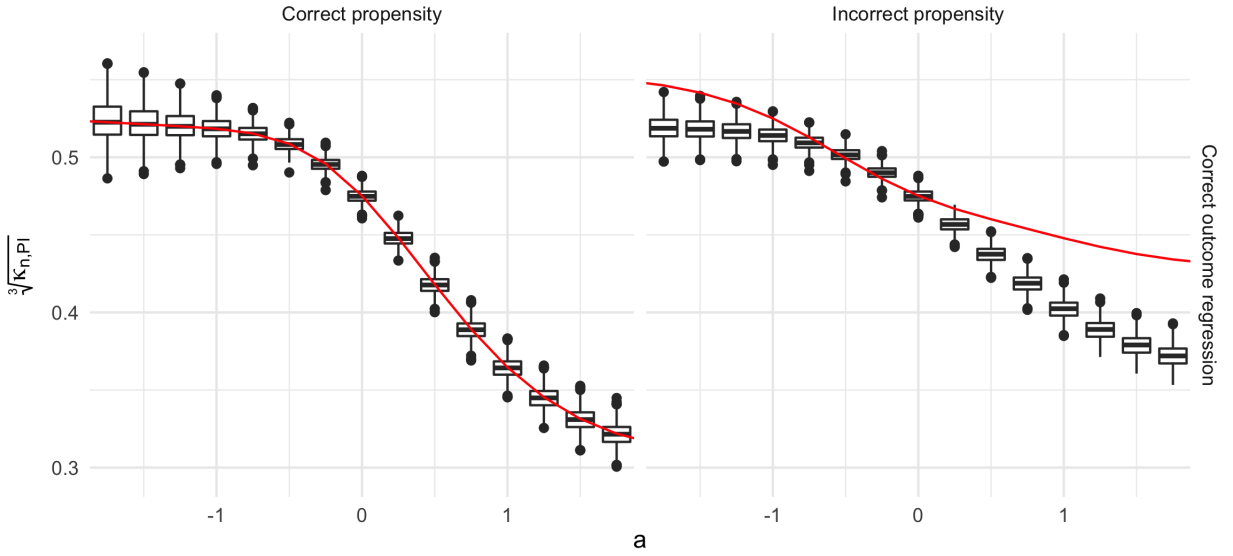

Figure 4.5: Distribution of the plug-in estimator $\kappa_n(a)$ of $\kappa_0(a)$ for different values of a and four specifications of the nuisance parameters over one thousand data sets simulated as described in the text with $n = 5000$ observations. Note that the bottom-right panel corresponds to regular isotonic regression. Red lines show the true values $\kappa_0(a)$.

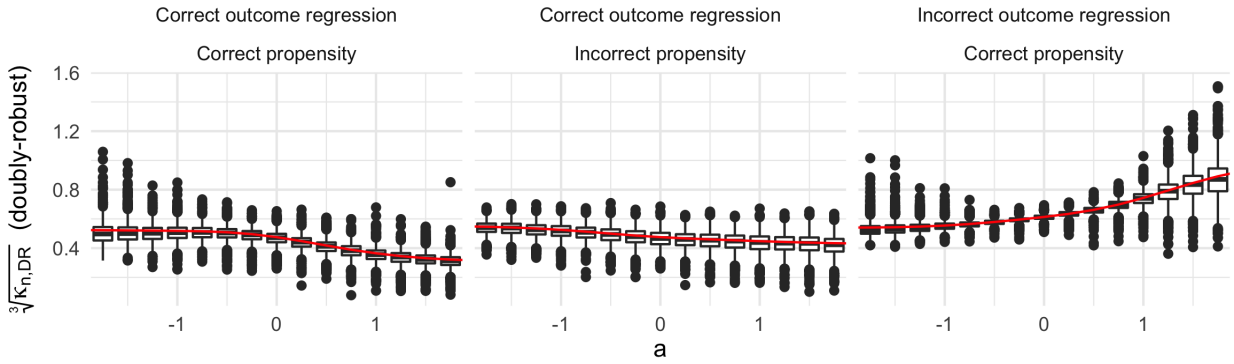

Figure 4.6: Distribution of the doubly-robust estimator $\kappa_n(a)$ of $\kappa_0(a)$ for different values of a and four specifications of the nuisance parameters over one thousand data sets simulated as described in the text. Note that the bottom-right panel corresponds to regular isotonic regression. Red lines show the true values $\kappa_0(a)$.

require larger sample sizes to achieve good coverage when machine learning estimators are used for μ_n and g_n. On the other hand, the plug-in intervals appear to be relatively robust to moderate mis-specification of nuisance parameters in finite samples.

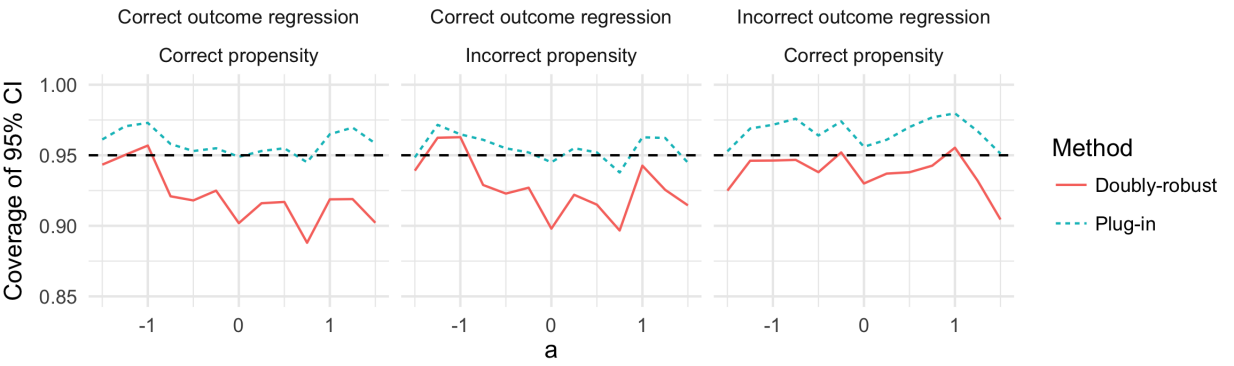

Figure 4.7: Coverage of pointwise 95% doubly-robust and plug-in confidence intervals using machine learning estimators and $n = 1000$ observations. Columns indicate whether the models for μ and g were correctly specified. Black dashed lines indicate the nominal coverage rate.

4.6 Effect of BMI on T-cell response in HIV vaccine trials

In this section we use the methods we have presented to assess the effect of body mass index (BMI) on CD4+ and CD8+ T-cell responses in a collection of clinical trials of candidate HIV vaccines. We pooled data from the vaccine arms of 11 phase I/II clinical trials, all conducted by the HIV Vaccine Trials Network (HVTN). Ten of these trials were previously studied in the analysis presented in Jin et al. (2015), and descriptions of the trials are contained therein. The final trial in our pooled analysis is HVTN 100, which randomized 210 participants to receive four doses of the ALVAC-HIV vaccine (vCP1521). The ALVAC-HIV vaccine, in combination with an AIDSVAX boost, was found to have statistically significant vaccine efficacy against HIV-1 in the RV-144 trial conducted in Thailand (Rerks-Ngarm et al., 2009).

CD4+ and CD8+ T-cell responses were measured in all 11 trials using validated intracellular cytokine staining (ICS) at HVTN laboratories. These responses were converted to binary indicators of whether there was a significant change from baseline using the method described in Jin et al. (2015). We analyze these binary responses at the first post-final-vaccination visit, which was scheduled for either two or four weeks after the final vaccination, depending on the trial. After accounting for missing responses from a small number of participants, our analysis data sets consisted of sample sizes of $n = 439$ participants for the analysis of CD4+ responses and $n = 462$ participants for the analysis of CD8+ responses.

We are interested in the effect of BMI on the probability of having a positive CD4+ or CD8+ response. In Jin et al. (2015), the authors examined the relationship between BMI and T-cell responses in two ways. First, they performed a marginal analysis comparing the CD4+ and CD8+ response rates among low (BMI < 25) medium ($25 \leq$ BMI < 30) and high (BMI ≥ 30) BMI participants. They found a monotonically decreasing trend across these three categories, and found that low BMI participants had a statistically significantly higher response rate than high BMI participants using Fisher's exact test. Second, they performed a logistic regression of the binary CD4+ and CD8+ responses against sex, age, BMI (not discretized), vaccination dose, and number of vaccinations. In this adjusted analysis they

found a statistically significant effect of BMI on CD4+ response rate after adjusting for the other covariates (OR: 0.92; 95% CI: 0.86, 0.98; p=0.007).

We assessed the effect of BMI on T-cell response using our estimator θ_n of the covariate-adjusted dose-response function θ under the assumption that θ is monotone decreasing. We adjusted for sex, age, vaccination dose, number of vaccinations, and study. We estimated the nuisance parameters μ_0 and g_0 using the methods described in Section 4.3.6. Finally, we formed pointwise confidence intervals used the doubly-robust estimator $\kappa_{n,\mathrm{DR}}$.

Figure 4.8 presents the estimated probability of a positive CD4+ T-cell response (left panel) and the estimated probability of a positive CD8+ T-cell response (right panel) as a function of BMI. Pointwise 95% confidence intervals are shown as vertical bars, and the marginal distributions of BMI for responders and non-responders are shown as box plots. We find that BMI had a greater absolute effect on CD4+ T-cell response probability than it did on CD8+ response probability. Additionally, the change in probability for CD4+ appears to be largest for BMI < 20.

We estimate that the probability of having a positive CD4+ T-cell response, after adjusting for potential confounders, is 0.57 (95% CI: 0.42, 0.72) for a BMI of 20, 0.50 (0.44, 0.56) for a BMI of 25, 0.45 (0.35, 0.54) for a BMI of 30, and 0.33 (0.28, 0.38) for a BMI of 35. We estimate that the probability of having a positive CD8+ T-cell response, after adjusting for potential confounders, is 0.27 (0.19, 0.34) for a BMI of 20, 0.19 (0.14, 0.23) for a BMI of 25, 0.19 (0.17, 0.21) for a BMI of 30, and 0.19 (0.17, 0.21) for a BMI of 35.

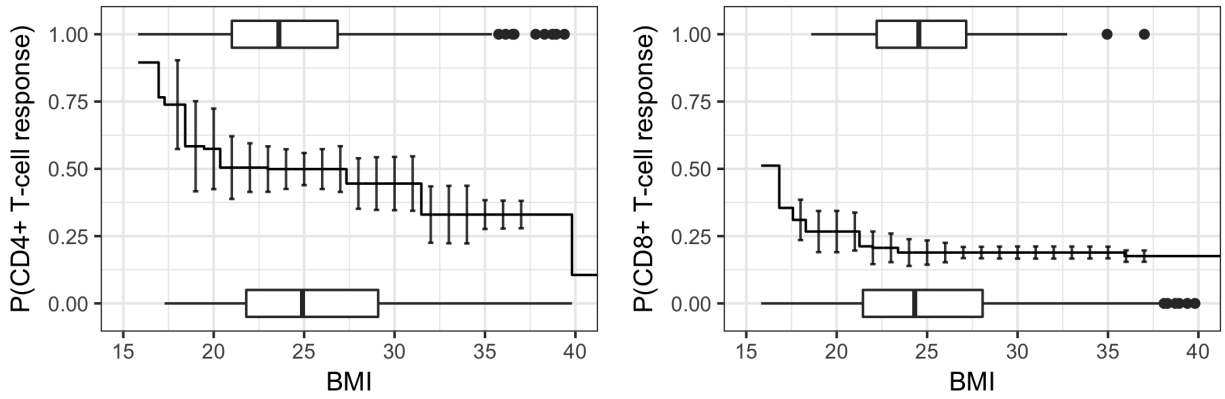

Figure 4.8: Estimated probabilities of CD4+ (left panel) and CD8+ (right panel) T-cell response as a function of BMI, adjusted for sex, age, number of vaccinations received, vaccine dose, and study. Vertical bars indicate pointwise 95% Wald-type confidence intervals using the doubly-robust inferential method.

Chapter 5

DISCUSSION AND FUTURE WORK

In this dissertation, we have considered two general approaches to nonparametric estimation of a monotone function. Our first approach applies to settings in which an initial estimator of the function is available. However, many such estimators of function-valued parameters in nonparametric and semiparametric models are not guaranteed to respect shape constraints on the true function. A simple and general solution to this problem is to project the initial estimator on to the constrained parameter space over a grid whose mesh goes to zero fast enough with sample size. However, this introduces the possibility that the projected estimator has different properties than the original estimator. In Chapter 2, we provided results that the projected estimator is generically no worse than the initial estimator, and that if the true function is upper and lower Lipschitz and the initial estimator possesses a type of uniform asymptotic equicontinuity, the projected estimator is uniformly asymptotically equivalent to the initial estimator.

We also provided especially simple sufficient conditions for this latter result when the initial estimator is uniformly asymptotically linear. This was our primary motivating setting upon initially studying this problem, and remains our main setting of interest for the application of our results. This is because when the function of interest is pathwise differentiable, there are multiple existing nonparametric techniques for constructing asymptotically linear estimators thereof. When the parameter of interest is not pathwise differentiable, asymptotically linear estimation is not possible. While there are many nonparametric techniques for estimating non-pathwise differentiable parameters, such as kernel smoothing, penalized optimization, and sieves, many of these methods leverage assumed smoothness of the underlying function, so the resulting estimators are also smooth. In these cases, the general theory we

developed in Chapter 2 can still often be used to demonstrate that the corrected estimator is asymptotically equivalent to the initial estimator, but in practice, we have found that initial estimators which are smooth are much less likely to exhibit monotonicity violations than initial estimators that are not smooth.

An additional reason that we focused on the asymptotically linear case is that constructing confidence intervals and uniform confidence bands for non-pathwise differentiable functions with correct asymptotic coverage is a very challenging problem. For instance, in kernel smoothing problems, the kernel smoothed estimator with bandwidth equal to the optimal bandwidth for minimizing the asymptotic integrated mean square error possesses an asymptotic bias. This means that the centered and appropriately rescaled estimator converges in distribution to a limit distribution that generally has non-zero mean. As a result, standard methods of constructing pointwise confidence intervals or uniform confidence bands, such as Wald's method and the nonparametric bootstrap, generally have asymptotic coverage that is strictly smaller than the nominal coverage. While the correction procedure we have proposed would not make the coverage worse, it also would not solve this under-coverage problem.

We studied the application of our results in two examples: a G-computed distribution function, for use in understanding the effect of a binary exposure on an outcome when the treatment-outcome relationship is confounded by recorded covariates, both with uncensored data and possibly informatively right-censored data. In numerical studies, we found that the projected estimator can yield benefits in small samples, but that these benefits diminish rapidly as n grows.

While our theory applies to monotone functions of arbitrary dimension $d \geq 1$, we only conducted numerical studies for univariate functions. We suspect that the size of the finite-sample benefits yielded by the correction procedure increases with dimension. Our reasoning is that the relative size, measured in metric entropy, of the set of component-wise monotone functions to the set of bounded variation functions is decreasing exponentially with d. Conducting numerical experiments for $d > 1$ is an area for future work.

Our theoretical results do not give the exact asymptotic behavior of the projected esti-

mator or projected confidence band when the true function possesses flat regions, or even when the the true function has a non-linear lower modulus of continuity. While there is no danger that confidence regions or tests based on the corrected band would be worse than the initial bands in such cases, it may be possible to construct tighter regions or more powerful tests at the same level. This would also be an interesting topic for future research.

Our results from Chapter 2 raise at least two other directions of future work. First, the same correction procedure can be used for other shape constraints, such as spaces of convex or log-convex functions or the space of multivariate distribution functions. It would be interesting to know whether our results extend to these other cases, and especially whether there is an underlying theory that could be developed. However, our proof techniques were strongly tied to monotonicity, and hence alternative approaches would likely be needed.

One potential avenue for a general approach is suggested by the literature on projections in Hilbert spaces. Suppose $\mathbf{\Psi}$ is a convex subset of a Hilbert space $\mathbb{H}$. Define $\text{Proj}_{\mathbf{\Psi}} : \mathbb{H} \to \mathbf{\Psi}$ as the metric projection on to $\mathbf{\Psi}$, defined as $\text{Proj}_{\mathbf{\Psi}}(x) := \text{argmin}_{\psi \in \mathbf{\Psi}} \|x - \psi\|_{\mathbb{H}}$. Zarantonello (1971) showed that $\text{Proj}_{\mathbf{\Psi}}$ is Hadamard directionally differentiable at any $\psi_0 \in \mathbf{\Psi}$, and that its derivative map is given by projection on to the *tangent cone* to $\mathbf{\Psi}$ at ψ_0, defined as $T_{\mathbf{\Psi}}(\psi_0) := \overline{\{t(\psi - \psi_0) : t \geq 0, \psi \in \mathbf{\Psi}\}}$. (The *directional* aspect of the differentiability is due to the fact that this derivative map need not be linear.) If ψ_0 is an *inner point* of $\mathbf{\Psi}$, meaning that $T_{\mathbf{\Psi}}(\psi_0) = \mathbb{H}$, then the derivative map is the identity. Suppose $\mathbf{\Psi}$ is a convex shape constrained class in the Hilbert space $\mathbb{H} := L_2([0,1]^d)$ of square-integrable real-valued functions on the d-dimensional unit cube $[0,1]^d$. Then the directional differentiability of $\text{Proj}_{\mathbf{\Psi}}$ and the delta method imply that if $r_n(\psi_n - \psi_0)$ converges weakly in $\mathbb{H}$ to a tight limit process $\mathbb{W}$, then $r_n(\psi_n^* - \psi_0)$ also converges weakly in $\mathbb{H}$ to $\text{Proj}_{T_{\mathbf{\Psi}}(\psi_0)}(\mathbb{W})$, where $\Psi_n^* := \text{Proj}_{\mathbf{\Psi}}(\psi_n)$ is the corrected estimator. The limit is the same, and $r_n\|\psi_n^* - \psi_n\|_{\mathbb{H}} = o_P(1)$, if ψ_0 is an inner point of $\mathbf{\Psi}$.

The result in the preceding paragraph provides a limit distribution even when ψ_0 is not an inner point of $\mathbf{\Psi}$, unlike our results in Chapter 2. However, in other respects, the conditions in the previous paragraph are stronger and the result is weaker than the results we provided

in Chapter 2. In particular, $r_n\|\psi_n^* - \psi_n\|_{\mathbb{H}} = o_P(1)$ does not imply that $r_n\|\psi_n^* - \psi_n\|_\infty = o_P(1)$, and the latter is needed to construct pointwise and uniform confidence bands for Ψ_0. However, strengthening the result to the uniform norm appears challenging. Additionally, in the preceding paragraph, we defined the corrected estimator Ψ_n^* as the projection of ψ_n over the entire domain $[0,1]^d$. This is an infinite-dimensional optimization problem, and hence is frequently not possible in practice. Instead, it would be preferable to define the projection over a finite grid in the domain and to interpolate elsewhere, as we did in Chapter 2.

The second question raised by the results of Chapter 2 is whether the efficiency theory of pathwise differentiable functionals in shape-constrained models can be characterized. Specifically, suppose $\mathcal{M}$ is a statistical model of probability measures on a probability space $(\mathcal{O}, \mathcal{B})$. Let $\mathcal{F}$ be a subset of a Banach space $(\mathbf{B}, \|\cdot\|_B)$ and $\psi : \mathcal{M} \to \mathcal{F}$ be a parameter mapping. Let $\boldsymbol{\Psi} \subset \mathcal{F}$ represent a subset of $\mathcal{F}$ within which the true parameter ψ_0 is known to lie. Define $\mathcal{M}_{\boldsymbol{\Psi}} := \psi^{-1}(\boldsymbol{\Psi})$ as the sub-model of $\mathcal{M}$ induced by the shape constraint $\boldsymbol{\Psi}$. Hence, the shape constraint $\boldsymbol{\Psi}$ on the functional ψ induces the shape-constrained model $\mathcal{M}_{\boldsymbol{\Psi}}$.

As we mentioned in Chapter 1, tangent spaces play a crucial role in the efficiency theory of pathwise differentiable functionals. In particular, the efficient influence function ϕ_0 of a pathwise differentiable parameter on $\mathcal{M}$ only depends on $\mathcal{M}$ through the tangent space $T_{\mathcal{M}}(P_0)$. Hence, if two models possess the same tangent space at P_0, then the efficient influence function for estimating any pathwise differentiable parameter is the same.

It is natural, then, to ask whether the tangent space $T_{\mathcal{M}_{\boldsymbol{\Psi}}}(P_0)$ of a shape-constrained model can be characterized in terms of the model $\mathcal{M}$, the shape-constrained function ψ, and the shape constraint $\boldsymbol{\Psi}$. This would make it possible to find the efficient influence function of any pathwise differentiable parameter under the shape constraint on ψ. In particular, if we could find sufficient conditions under which $T_{\mathcal{M}_{\boldsymbol{\Psi}}}(P_0) = T_{\mathcal{M}}(P_0)$, then under these conditions the shape constraint would have no effect on the efficiency theory of pathwise differentiable functionals on the model, so that incorporating the shape constraint in to an estimation procedure would not yield first-order asymptotic gains.

In Chapter 3, we studied a broad class of estimators of monotone functions based on

differentiating the greatest convex minorant of a preliminary estimator of a primitive parameter. A novel aspect of the class we have considered is its allowance for the primitive parameter to involve a possibly data-dependent transformation of the domain. The class we defined is useful because it generalizes classical approaches for simple monotone functions, including density, hazard and regression functions, facilitates the integration of flexible, data-adaptive learning techniques, and allows valid asymptotic statistical inference. We provided general asymptotic results for estimators in this class and have also derived refined results for the important case wherein the primitive estimator is uniformly asymptotically linear. We have proposed novel estimators of extensions of classical monotone parameters that deal with common sampling complications, and described their large-sample properties using our general results.

Our primary goal in Chapter 3 was to establish general theoretical results that can be applied to study many specific estimators, and as such, there are numerous potential applications of our results. In Chapter 4, we used our general results from Chapter 3 to study estimation of one such particular parameter, namely a monotone dose-response function with informative treatment allocation. We used the G-formula to identify the true counterfactual mean. This identification requires that we observe all confounders between the exposure and outcome. Instrumental variables are an alternative identification strategy that is sometimes useful when the exposure-outcome relationship is likely to be confounded by unobserved variables. Estimation of a monotone dose-response function identified via instrumental variables would be an interesting area of future research.

Extension of the parameter we study in Chapter 4 to right-censored data would also be natural. Specifically, suppose $T \in \mathbb{R}^+$ is an event time of interest, A is a binary treatment indicator, $X \in \mathbb{R}$ is a continuous covariate, measured at baseline, and $W \in \mathbb{R}^d$ is a vector of other baseline covariates. A parameter of interest in some studies is the G-computed conditional risk function $\theta_P(x) := E_P[P(T \leq t_0 \mid X = x, A = a_0, W)]$. For instance, in vaccine trials, T represents time to infection by a disease of interest, A represents assignment to the vaccine or placebo arm of the trial, and X is a potential *correlate of risk* of infection.

$\theta_P(x)$ then measures the chance of infection under assignment to arm a_0 and $X = x$. For certain types of covariates X, it may be known that θ_P is monotone. In observational studies, estimation of θ_P is often further complicated by the fact that T may be subject to right-censorship.

Two closely related parameters are risk difference and risk ratio functions. In some situations, the conditional risk function under assignment to a_0 and $X = x$ may not necessarily be monotone, but the conditional risk difference function

$$\theta_P(x) := E_P[P(T \leq t_0 \mid X = x, A = 1, W)] - E_P[P(T \leq t_0 \mid X = x, A = 0, W)]$$

or conditional risk ratio function

$$\theta_P(x) := \frac{E_P[P(T \leq t_0 \mid X = x, A = 1, W)]}{E_P[P(T \leq t_0 \mid X = x, A = 0, W)]}.$$

In vaccine trials, these parameters measure the extent to which X is a *correlate of protection* of the vaccine against infection. For instance, for some diseases, the more of a particular antibody is present in a patient's blood prior to vaccination (often due to previous exposure to the disease), the more the vaccine elicits an immune response, and hence the better the vaccine is able to confer protection from the disease. In studies of other types of treatments, a patient's health prior to treatment, measured for instance by BMI or blood pressure, is predictive of the success or failure of the treatment.

In addition to other monotone parameters of interest, there are also a multitude of useful properties and modifications of Grenander-type estimators that have been studied in the literature and whose extension to our class would be important. For instance, kernel smoothing of a Grenander-type estimator yields a monotone estimator that possesses many of the properties of usual kernel smoothing estimators, including possibly faster convergence to a normal distribution (e.g., Mukerjee, 1988; Mammen, 1991; Groeneboom et al., 2010). The asymptotic distribution of the supremum norm error of Grenander-type estimators has also

been derived (e.g., Durot, 2012), and extending this result to our class would refine further our results. Asymptotic results at the boundaries of the domain and corrections for poor behavior there have been developed and would further enhance the utility of these methods (e.g., Woodroofe and Sun, 1993; Balabdaoui et al., 2011; Kulikov and Lopuhaä, 2006).

There have also been various proposals for constructing asymptotically valid pointwise confidence intervals for Grenander-type estimators without the need to compute the complicated scale parameters appearing in their limit distribution. In regular statistical problems, the bootstrap is one of the most widely used such methods; unfortunately, the nonparametric bootstrap is known to fail for Grenander-type estimators (e.g., Kosorok, 2008; Sen et al., 2010). However, these articles have demonstrated that the m-out-of-n bootstrap is valid for Grenander-type estimators, and that bootstrapping smoothed Grenander-type estimators can also be an effective strategy for performing inference. Asymptotically pivotal distributions based on likelihood ratios have also been used to avoid the need to estimate nuisance parameters in the limit distribution and to provide a basis for improved finite-sample inference (e.g., Banerjee and Wellner, 2001; Banerjee, 2005a,b, 2007; Groeneboom and Jongbloed, 2015). Considering these strategies in our setting would be particularly interesting.

In Chapter 4, we defined a generalized Grenander-type estimator of a monotone covariate-marginalized regression function. Under untestable causal assumptions, including that all confounders of the exposure-outcome relationship are observed, our parameter of interest in Chapter 4 can be interpreted as a causal dose-response function. We showed that our estimator generalizes the least-squares estimator of a monotone marginal regression function, and that it retains many of the attractive properties of said estimator, including invariance to scale changes of the exposure. We also used our general results from Chapter 3 to provide sufficient conditions on the estimators of the outcome regression and exposure propensity functions that imply that our estimator is consistent and that our estimator minus the true G-computed regression function converges in distribution pointwise at the rate $n^{-1/3}$ to a non-standard limit distribution. Our sufficient conditions allow for the possibility of both doubly-robust consistency and doubly-robust convergence in distribution, and furthermore

the type of doubly-robustness that we allow is substantially more flexible than standard notions of doubly-robustness.

Our convergence in distribution result in Chapter 4 is doubly-robust in the sense that it does not require both nuisance parameter μ_n and g_n to be correctly specified. However, if one of the nuisance parameters is not correctly specified, then the other parameter needs to be converging at a rate faster than $n^{-1/3}$ uniformly over a neighborhood of the point of interest. This is a strong requirement, and in general requires that the estimator exploit special structure of the true function, such as multiple continuous derivatives or additivity over the components of W. In the setting where A is binary, recent work has shown that an asymptotically linear estimator of $\theta_0(a)$ can be constructed in such a way that the estimator converges at the rate $n^{-1/2}$ to a normal distribution as long as at least one of the nuisance functions is correctly specified and both estimators convergence to their respective limits at a rate faster than $n^{-1/4}$ (Benkeser et al., 2017). This rate of convergence can be guaranteed for arbitrarily high-dimensional W under the relatively mild assumption of bounded sectional variation (van der Laan and Benkeser, 2018). It would be interesting to consider estimating the primitive parameter that arises in our study of the monotone-dose response curve in Chapter 4 using such a doubly-robust asymptotically linear estimator. We conjecture that the resulting monotone estimator would converge in distribution to a mean-zero limit as long as one of the nuisance estimators were correctly specified and both estimators were converging at a rate faster than $n^{-1/6}$ to their respective limits.

BIBLIOGRAPHY

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 641–647.

Balabdaoui, F., H. Jankowski, M. Pavlides, A. Seregin, and J. Wellner (2011). On the Grenander estimator at zero. *Statistica Sinica 21*(2), 873.

Banerjee, M. (2005a). Likelihood ratio tests under local alternatives in regular semiparametric models. *Statistica Sinica 15*(3), 635–644.

Banerjee, M. (2005b). Likelihood ratio tests under local and fixed alternatives in monotone function problems. *Scandinavian Journal of Statistics 32*(4), 507–525.

Banerjee, M. (2007). Likelihood based inference for monotone response models. *The Annals of Statistics 35*(3), 931–956.

Banerjee, M. and J. A. Wellner (2001). Likelihood ratio tests for monotone functions. *The Annals of Statistics 29*(6), 1699–1731.

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

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression.* Wiley New York.

Beare, B. K. and Z. Fang (2017). Weak convergence of the least concave majorant of estimators for a concave distribution function. *Electron. J. Statist. 11*(2), 3841–3870.

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

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1998). *Efficient and adaptive estimation for semiparametric models*, Volume 2. Springer New York.

Brunk, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, London, pp. 177–197. Cambridge Univ. Press.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics 16*(1), 31–41.

Chernozhukov, V., I. Fernandez-Val, and A. Galichon (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika 96*(3), 559–575.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Cox, D. (1958). *Planning of Experiments*. New York: Wiley.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics 17*(3), 1157–1167.

Daouia, A. and B. U. Park (2013). On projection-type estimators of multivariate isotonic functions. *Scandinavian Journal of Statistics 40*(2), 363–386.

Díaz Muñoz, I. and M. J. van der Laan (2011). Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics 7*(1), 1–20.

Durot, C. (2007). On the $\mathbb{L}_p$ -error of monotonicity constrained estimators. *The Annals of Statistics 35*(3), 1080–1104.

Durot, C. (2012). The limit distribution of the L_∞-error of Grenander-type estimators. *The Annals of Statistics 40*(3), 1578–1608.

Durot, C., P. Groeneboom, and H. P. Lopuhaä (2013). Testing equality of functions under monotonicity constraints. *Journal of Nonparametric Statistics 25*(4), 939–970.

Durot, C. and H. P. Lopuhaä (2014). A kiefer-wolfowitz type of result in a general setting, with an application to smooth monotone estimation. *Electron. J. Statist. 8*(2), 2479–2513.

Fan, J. (1996). *Local Polynomial Modelling and Its Applications*. Boca Raton: CRC Press.

Ferguson, T. S. (2014). *Mathematical statistics: A decision theoretic approach*. Academic Press.

Gill, R. D. and S. Johansen (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics 18*(4), 1501–1555.

Gill, R. D. and J. M. Robins (2001). Causal inference for complex longitudinal data: The continuous case. *The Annals of Statistics 29*(6), 1785–1811.

Gill, R. D., M. J. Van Der Laan, and J. M. Robins (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In D. Lin (Ed.), *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294. Springer, New York.

Grenander, U. (1956). On the theory of mortality measurement. II. *Scandinavian Actuarial Journal 39*, 125–153.

Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II*, Belmont, CA, pp. 539–555. Wadsworth.

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

Groeneboom, P. and G. Jongbloed (2015). Nonparametric confidence intervals for monotone functions. *The Annals of Statistics 43*(5), 2019–2054.

Groeneboom, P., G. Jongbloed, and B. I. Witte (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *The Annals of Statistics 38*(1), 352–387.

Groeneboom, P. and J. A. Wellner (2001). Computing chernoff's distribution. *Journal of Computational and Graphical Statistics 10*(2), 388–400.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–331.

Heitjan, D. F. and D. B. Rubin (1991). Ignorability and coarse data. *The Annals of Statistics 19*(4), 2244–2253.

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer Series in Statistics. New York: Springer.

Huang, J. and J. A. Wellner (1995). Estimation of a monotone density or monotone hazard under random censoring. *Scandinavian Journal of Statistics 22*(1), 3–33.

Huang, Y. and C.-H. Zhang (1994). Estimating a monotone density from censored observations. *The Annals of Statistics 22*(3), 1256–1274.

Hubbard, A. E., M. J. van der Laan, and J. M. Robins (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.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics 86*(1), 4–29.

Jin, X., C. Morgan, X. Yu, S. DeRosa, G. D. Tomaras, D. C. Montefiori, J. Kublin, L. Corey, M. C. Keefer, N. H. V. T. Network, et al. (2015). Multiple factors affect immunogenicity of dna plasmid hiv vaccines in human clinical trials. *Vaccine 33*(20), 2347–2353.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(4), 1229–1245.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics 18*, 191–219.

Kosorok, M. R. (2008). Bootstrapping the grenander estimator. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, Volume 1 of *Collections*, pp. 282–292. Institute of Mathematical Statistics.

Kulikov, V. N. and H. P. Lopuhaä (2006). The behavior of the NPMLE of a decreasing density near the boundaries of the support. *The Annals of Statistics 34*(2), 742–768.

Laslett, G. M. (1982). The survival curve under monotone density constraints with applications to two-dimensional line segment processes. *Biometrika 69*(1), 153–160.

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

Leurgans, S. (1982). Asymptotic distributions of slope-of-greatest-convex-minorant estimators. *The Annals of Statistics 10*(1), 287–296.

Lopuhaä, H. P. and E. Musta (2016). A central limit theorem for the Hellinger loss of Grenander type estimators. *ArXiv e-prints*.

Lopuhaä, H. P. and E. Musta (2017). Isotonized smooth estimators of a monotone baseline hazard in the Cox model. *Journal of Statistical Planning and Inference 191*, 43 – 67.

Lopuhaä, H. P. and E. Musta (2018a). The distance between a naive cumulative estimator and its least concave majorant. *Statistics & Probability Letters 139*, 119 – 128.

Lopuhaä, H. P. and E. Musta (2018+b). Smoothed isotonic estimators of a monotone baseline hazard in the Cox model. *Scandinavian Journal of Statistics*, to appear.

Lopuhaä, H. P. and G. F. Nane (2013a). An asymptotic linear representation for the Breslow estimator. *Communications in Statistics - Theory and Methods 42*(7), 1314–1324.

Lopuhaä, H. P. and G. F. Nane (2013b). Shape constrained non-parametric estimators of the baseline distribution in Cox proportional hazards model. *Scandinavian Journal of Statistics 40*(3), 619–646.

Luedtke, A. R., O. Sofrygin, M. J. van der Laan, and M. Carone (2017). Sequential Double Robustness in Right-Censored Longitudinal Models. *ArXiv e-prints*.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics 19*(2), 724–740.

McNichols, D. and W. Padgett (1982). Maximum likelihood estimation of unimodal and decreasing densities based on arbitrarily right-censored data. *Communications in Statistics - Theory and Methods 11*(20), 2259–2270.

Mukerjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics 16*(2), 741–750.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications 9*(1), 141–142.

Neugebauer, R. and M. van der Laan (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference 129*(1-2), 405–426.

Neugebauer, R. and M. van der Laan (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference 137*(2), 419 – 434.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). *Statistical Science 5*, 465–480.

Nolan, D. and D. Pollard (1987). U-Processes: Rates of Convergence. *Ann. Statist. 15*(2), 780–799.

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory*. Springer.

Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 31*(1), 23–36.

Prakasa Rao, B. L. S. (1970). Estimation for distributions with monotone failure rate. *The Annals of Mathematical Statistics 41*(2), 507–519.

Rerks-Ngarm, S., P. Pitisuttithum, S. Nitayaphan, J. Kaewkungwal, J. Chiu, R. Paris, N. Premsri, C. Namwat, M. de Souza, E. Adams, et al. (2009). Vaccination with AL-VAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine 361*(23), 2209–2220.

Robertson, T., F. Wright, and R. Dykstra (1988). *Order Restricted Statistical Inference*. Wiley, New York.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9), 1393 – 1512.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, New York, NY, pp. 95–133. Springer New York.

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

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rotnitzky, A., J. Robins, and L. Babino (2017). On the multiply robust estimation of the mean of the g-functional. *ArXiv e-prints*.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association 93*(444), 1321–1339.

Rubin, D. and M. J. van der Laan (2006). Extending marginal structural models through local, penalized, and additive learning. Working Paper 212, Division of Biostatistics, University of California at Berkeley, Berkeley, California.

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

Scharfstein, D. O. and J. M. Robins (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika 89*(3), 617–634.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Sen, B., M. Banerjee, and M. Woodroofe (2010). Inconsistency of bootstrap: The Grenander estimator. *The Annals of Statistics 38*(4), 1953–1977.

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

van der Laan, M. J. and D. Benkeser (2018). Highly Adaptive Lasso (HAL). In M. J. van der Laan and S. Rose (Eds.), *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pp. 77–94. Springer.

van der Laan, M. J., A. Bibaut, and A. R. Luedtke (2018). CV-TMLE for nonpathwise differentiable target parameters. In M. J. van der Laan and S. Rose (Eds.), *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*, pp. 455–481. Springer.

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

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

van der Laan, M. J. and S. Rose (2011). *Targeted learning: causal inference for observational and experimental data*. Springer-Verlag New York.

van der Vaart, A. (1991). On differentiable functionals. *The Annals of Statistics 19*(1), 178–204.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3. Cambridge university press.

van der Vaart, A. W. and M. J. van der Laan (2006). Estimating a survival distribution with current status data and high-dimensional covariates. *The International Journal of Biostatistics 2*(1).

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pp. 831–838.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks 10*(5), 988–999.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 26*(4), 359–372.

Woodroofe, M. and J. Sun (1993). A penalized maximum likelihood estimate of f(0+) when f is non-increasing. *Statistica Sinica 3*(2), 501–515.

Zarantonello, E. H. (1971). Projections on convex sets in hilbert space and spectral theory: Part i. projections on convex sets: Part ii. spectral theory. In E. H. Zarantonello (Ed.), *Contributions to Nonlinear Functional Analysis*, pp. 237 – 424. Academic Press.

Zeng, D. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *The Annals of Statistics 32*(4), 1533–1555.

Zheng, W. and M. J. van der Laan (2011). Cross-validated targeted minimum loss based estimation. In M. van der Laan and S. Rose (Eds.), *Targeted learning: causal inference for observational and experimental data*, Chapter 27, pp. 459–473. New York: Springer-Verlag New York.

Appendix A

PROOF OF RESULTS FROM CHAPTER 2

Proof of Theorem 1. Statement (i) follows from Corollary B to Theorem 1.6.1 of Robertson et al. (1988). For (ii) and (iii), note that for every $t \in T$ we have by assumption

$$|\theta_n^*(t) - \theta_0(t)| \leq \sum_k \lambda_k(t)|\theta_n^*(s_k) - \theta_0(s_k)| + \sum_k \lambda_k(t)|\theta_0(s_k) - \theta_0(t)|,$$

where $\sum_k \lambda_k(t) = 1$, $s_k \in \mathcal{T}_n$ for each k, and $\|s_k - t\| \leq 2\omega_n$. By (i), the first term is bounded above by $\sup_{s\in\mathcal{T}_n} |\theta_n(s) - \theta_0(s)|$. The second term is bounded above by $\gamma(2\omega_n)$, where we define $\gamma(\delta) = \sup\{|\theta_0(t) - \theta_0(s)| : t, s, \in T, \|t - s\| \leq \delta\}$. If θ_0 is uniformly continuous on T, $\gamma(\delta) \to 0 = \gamma(0)$ as $\delta \to 0$, so that $\gamma(2\omega_n) \to_p 0$ if $\omega_n \to_p 0$. If $\gamma(\delta) = o(\delta^\alpha)$ as $\delta \to 0$, then $\gamma(2\omega_n) = \mathrm{o_P}(\omega_n^\alpha)$.

Statement (iv) follows from the proof of Proposition 3 of Chernozhukov et al. (2009), which applies to any order-preserving monotonization procedure. For the first statement of (v), note that $\sum_{t\in\mathcal{T}_n} u_n^*(t) = \sum_{t\in\mathcal{T}_n} u_n(t)$, and similarly for ℓ_n^*, by their definition as minimizers of the least-squares criterion function.

The second statement of (v) follows from a slight modification of Robertson et al. (1988) Theorem 1.6.1. As stated, the result says that $\sum_{t\in\mathcal{T}_n} G(\theta^*(t) - \theta(t)) \leq \sum_{t\in\mathcal{T}_n} G(\theta(t) - \psi(t))$ for any convex function $G : \mathbb{R} \to \mathbb{R}$ and monotone function ψ, where θ^* is the isotonic regression of θ over $\mathcal{T}_n$. Their proof can be modified in a straightforward and minimal way to show that it also holds that $\sum_{t\in\mathcal{T}_n} G(\theta_1^*(t) - \theta_2^*(t)) \leq \sum_{t\in\mathcal{T}_n} G(\theta_1(t) - \theta_2(t))$, where now θ_1^* and θ_2^* are the isotonic regressions of θ_1 and θ_2 over $\mathcal{T}_n$, respectively. As in Corollary B, taking $G(x) = |x|^p$ and letting $p \to \infty$ yields $\|\theta_1^* - \theta_2^*\|_{\mathcal{T}_n} \leq \|\theta_1 - \theta_2\|_{\mathcal{T}_n}$. Applying this with $\theta_1 = u_n$ and $\theta_2 = \ell_n$ yields the second part of (v). □

Proof of Lemma 1. By the triangle inequality, $|\theta_n(t) - \theta_n(s)| \leq |\{\theta_n(t) - \theta_0(t)\} - \{\theta_n(s) - \theta_0(s)\}| + |\theta_0(t) - \theta_0(s)|$. The first term is $o_P(r_n^{-1})$ by (A1). The second term is $o_P(r_n^{-1})$ by (A2). □

Proof of Lemma 2. Let $\epsilon > 0$ and $\eta_n = \epsilon/r_n$. Suppose that $h_n > \eta_n$. The there exist $s, t \in \mathcal{T}$ with $s < t$ and $\|t - s\| > \eta_n$ such that $\theta_n(s) \geq \theta_n(t)$. We claim that there must also exist $s^*, t^* \in \mathcal{T}$ with $s^* < t^*$ and $\|t^* - s^*\| \in [\eta_n/2, \eta_n]$ such that $\theta_n(s^*) \geq \theta_n(t^*)$. To see this, let $J = \lfloor \|t - s\|/(\eta_n/2) \rfloor - 1$, and note that $J \geq 1$. Define $t_j = s + (j\eta_n/2)(t - s)/\|t - s\|$ for $j = 0, 1, \ldots, J$, and set $t_{J+1} = t$. Thus, $t_j < t_{j+1}$ and $\|t_{j+1} - t_j\| \in [\eta_n/2, \eta_n]$ for each $j = 0, \ldots, J$. Since then $\sum_{j=0}^{J} \{\theta_n(t_{j+1}) - \theta_n(t_j)\} = \theta_n(t) - \theta_n(s) \leq 0$, it must be that $\theta_n(t_{j+1}) \leq \theta_n(t_j)$ for at least one j. This proves the claim.

We now have that $h_n > \eta_n$ implies that there exist $s, t \in \mathcal{T}$ with $s < t$ and $\|t - s\| \in [\eta_n/2, \eta_n]$ such that $\theta_n(s) \geq \theta_n(t)$. This further implies that

$$\{\theta_n(t) - \theta_0(t)\} - \{\theta_n(s) - \theta_0(s)\} \leq -(\theta_0(t) - \theta_0(s)) \leq -K_0\|t - s\| \leq -K_0\eta_n/2$$

by condition (A2). Finally, this allows us to write

$$P\left(h_n > \epsilon/r_n\right) \;\leq\; P\left(\sup_{\|t-s\| \leq \epsilon/r_n} |r_n\{\theta_n(t) - \theta_0(t)\} - r_n\{\theta_n(s) - \theta_0(s)\}| \geq \epsilon\right).$$

By condition (A1), this probability tends to zero for every $\epsilon > 0$, which completes the proof. □

Proof of Lemma 3. For any $t \in \mathcal{T}_n$, Theorem 1.4.4 of Robertson et al. (1988) gives the representation

$$\theta_n^*(t) = \max_{U \in \mathcal{U}_t} \min_{L \in \mathcal{L}_t} \theta_n(U \cap L) = \min_{L \in \mathcal{L}_t} \max_{U \in \mathcal{U}_t} \theta_n(U \cap L),$$

where, for any finite set $S \subseteq \mathcal{T}_n$, $\theta_n(S)$ is defined as $|S|^{-1} \sum_{s \in S} \theta_n(s)$. The sets U range over the collection $\mathcal{U}_t$ of upper sets of $\mathcal{T}_n$ containing t, where $U \subseteq \mathcal{T}_n$ is called an upper set if and only if $t_1 \in U, t_2 \in \mathcal{T}_n$ and $t_1 \leq t_2$ implies $t_2 \in U$. The sets L range over the collection $\mathcal{L}_t$ of lower sets of $\mathcal{T}_n$ containing t, where $L \subseteq \mathcal{T}_n$ is called a lower set if and only if $t_1 \in L, t_2 \in \mathcal{T}_n$ and $t_2 \leq t_1$ implies $t_2 \in L$.

Let $U_t = \{s : s \geq t\}$ and $L_t = \{s : s \leq t\}$. First, suppose there exists $L_0 \in \mathcal{L}_t$ and $s_0 \in L_0$ with $s_0 > t$ and $\|t - s_0\| > h_n$. Then, we claim that there exists another lower set $L_0' \in \mathcal{L}_t$ such that $\theta_n(U_t \cap L_0) > \theta_n(U_t \cap L_0')$. If $\theta_n(U_t \cap L_0) > \theta_n(t) = \theta_n(U_t \cap L_t)$, then $L_0' = L_t$ satisfies the claim. Otherwise, if $\theta_n(U_t \cap L_0) \leq \theta_n(t)$, let $L_0' = L_0 - \{s : s > t, \|t - s\| > h_n\}$. One can verify that $L_0' \in \mathcal{L}_t$, and since $s_0 \in L_0 \setminus L_0'$, L_0' is a strict subset of L_0. Furthermore, by definition of h_n, $\theta_n(s) > \theta_n(t)$ for all $s > t$ such that $\|t - s\| > h_n$, and since $\theta_n(U_t \cap L_0) \leq \theta_n(t)$, removing these elements from L_0 can only reduce the average, so that $\theta_n(U_t \cap L_0') < \theta_n(U_t \cap L_0)$. This establishes the claim. By an analogous argument, we can show that if there exists $U_0 \in \mathcal{U}_t$ and $s_0 \in U_0$ with $s_0 < t$ and $\|t - s_0\| > h_n$, then there exists another upper set $U_0' \in \mathcal{U}_t$ such that $\theta_n(U_0 \cap L_t) < \theta_n(U_0' \cap L_t)$.

Let $L^* \in \operatorname{argmin}_{L \in \mathcal{L}_t} \theta_n(U_t \cap L)$ and $U^* \in \operatorname{argmax}_{U \in \mathcal{U}_t} \theta_n(U \cap L_t)$. Then, we have that

$$\begin{aligned}
\theta_n^*(t) = \max_{U \in \mathcal{U}_t} \min_{L \in \mathcal{L}_t} \theta_n(U \cap L) &\geq \min_{L \in \mathcal{L}_t} \theta_n(U_t \cap L) = \theta_n(U_t \cap L^*) \\
\theta_n^*(t) = \min_{L \in \mathcal{L}_t} \max_{U \in \mathcal{U}_t} \theta_n(U \cap L) &\leq \max_{U \in \mathcal{U}_t} \theta_n(U \cap L_t) = \theta_n(U^* \cap L_t) \ .
\end{aligned}$$

Hence, $\theta_n(U_t \cap L^*) \leq \theta_n^*(t) \leq \theta_n(U^* \cap L_t)$. By the above argument, $\theta_n(U_t \cap L^*) \geq \inf\{\theta_n(s) : s \geq t, \|t - s\| \leq h_n\}$ and $\theta_n(U^* \cap L_t) \leq \sup\{\theta_n(s) : s \leq t, \|t - s\| \leq h_n\}$. Therefore,

$$\inf\{\theta_n(s) - \theta_n(t) : \|t - s\| \leq h_n\} \leq \theta_n^*(t) - \theta_n(t) \leq \sup\{\theta_n(s) - \theta_n(t) : \|t - s\| \leq h_n\},$$

and thus, $|\theta_n^*(t) - \theta_n(t)| \leq \sup\{|\theta_n(s) - \theta_n(t)| : \|t - s\| \leq h_n\}$. Taking the maximum over $t \in \mathcal{T}_n$ yields the claim. □

Proof of Theorem 2. By construction, for each $t \in \mathcal{T}$, we can write

$$|\theta_n^*(t) - \theta_n(t)| \;\leq\; \Sigma_{j=1}^{2^d} \lambda_j(t) |\theta_n^*(s_j) - \theta_n(s_j)| + \Sigma_{j=1}^{2^d} \lambda_j(t) |\theta_n(s_j) - \theta_n(t)| \;,$$

where $s_j \in \mathcal{T}_n$ and $\|s_j - t\| \leq 2\omega_n$ for all t, s_j by definition. Thus, since $\sum_j \lambda_j(t) = 1$,

$$\sup_{t \in \mathcal{T}} |\theta_n^*(t) - \theta_n(t)| \;\leq\; \max_{t \in \mathcal{T}_n} |\theta_n^*(t) - \theta_n(t)| + \sup_{\|s-t\| \leq 2\omega_n} |\theta_n(s) - \theta_n(t)| \;.$$

By Lemma 3, the first summand is bounded above by $\sup_{\|s-t\| \leq h_n} |\theta_n(s) - \theta_n(t)|$, which is $o_P(r_n^{-1})$ by Lemmas 1 and 2. The second summand is $o_P(r_n^{-1})$ by Lemma 1. □

Proof of Corollary 1. For the first statement, we note that $\ell_n(t) \leq \theta_0(t) \leq u_n(t)$ if and only if

$$\{r_n[\theta_n(t) - \ell_n(t)] - \gamma_0(t)\} + \gamma_0(t) \geq r_n[\theta_n(t) - \theta_0(t)] \geq -\gamma_0(t) - \{r_n[u_n(t) - \theta_n(t)] - \gamma_0(t)\} \;.$$

Therefore, by (2.2),

$$P\bigg(\forall t \in \mathcal{T} : \ell_n(t) \leq \theta_0(t) \leq u_n(t)\bigg) \to 1 - \alpha \;.$$

For the second statement, we let $\delta > 0$, and we note that

$$\begin{aligned}
&\sup_{\|t-s\| \leq \delta / r_n} |r_n\{\ell_n(t) - \theta_0(t)\} - r_n\{\ell_n(s) - \theta_0(s)\}| \\
&\qquad \leq \sup_{\|t-s\| \leq \delta / r_n} |r_n\{\theta_n(t) - \theta_0(t)\} - r_n\{\theta_n(s) - \theta_0(s)\}| + 2\|r_n(\theta_n - \ell_n) - \gamma_0\|_{\mathcal{T}} \\
&\qquad\quad + \sup_{\|t-s\| \leq \delta / r_n} |\gamma_0(t) - \gamma_0(s)|.
\end{aligned}$$

The first term tends to zero in probability by (A1), the second by (2.2), and the third by the assumed uniform continuity of γ_0. An analogous decomposition holds for u_n. Therefore,

we can apply Theorem 2 with u_n and ℓ_n in place of θ_n to find that $\|\ell_n^* - \ell_n\|_{\mathcal{T}} = o_P(r_n^{-1})$ and $\|u_n^* - u_n\|_{\mathcal{T}} = o_P(r_n^{-1})$. □

Proof of Theorem 3. Let $\epsilon, \delta, \eta > 0$. By (2.3) and since $\sup_{t\in\mathcal{T}} |R_{n,t}| = o_P(n^{-1/2})$,

$$n^{1/2} \left|\{\theta_n(t) - \theta_0(t)\} - \{\theta_n(s) - \theta_0(s)\}\right| \;\leq\; |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| + o_P(1) \;.$$

Condition (A1.ii) implies that $\{\phi_{0,t} : t \in \mathcal{T}\}$ is uniformly mean-square continuous, in the sense that

$$\lim_{h\to 0} \sup_{\|t-s\|\leq h} \int \{\phi_{0,s}(x) - \phi_{0,t}(x)\}^2 \, dP_0(x) = 0 \;.$$

Since T is totally bounded in $\|\cdot\|$, this also implies that $\{\phi_{0,t} : t \in \mathcal{T}\}$ is totally bounded in the $L_2(P_0)$ metric. This, in addition to (A1.i), implies that $\{\mathbb{G}_n\phi_{0,t} : t \in \mathcal{T}\}$ converges weakly in $\ell^\infty(\mathcal{T})$ to a Gaussian process $\mathbb{G}$ with covariance function Σ_0. Furthermore, (A1.ii) implies that this limit process is a tight element of $\ell^\infty(\mathcal{T})$. By Theorem 1.5.4 of van der Vaart and Wellner (1996), $\{\mathbb{G}_n\phi_{0,t} : t \in \mathcal{T}\}$ is asymptotically tight. By Theorem 1.5.7 of van der Vaart and Wellner (1996), $\{\mathbb{G}_n\phi_{0,t} : t \in \mathcal{T}\}$ is thus asymptotically uniformly mean-square equicontinuous in probability, in the sense that there exists some $\delta_0 = \delta_0(\epsilon, \eta) > 0$ such that

$$\limsup_{n\to\infty} P\left\{ \sup_{\rho(s,t)<\delta_0} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon \right\} \;<\; \eta \;,$$

with $\rho(s,t) = [\int \{\phi_{0,t}(x) - \phi_{0,s}(x)\}^2 dP_0(x)]^{1/2}$. By (A1.ii), $\sup_{\|t-s\|\leq h} \rho(t,s) < \delta_0$ for some $h > 0$. Hence, for all n large enough, both $\delta n^{-1/2} \leq h$ and $P\{\sup_{\rho(s,t)<\delta_0} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon\} < \eta$, so that

$$P\left\{ \sup_{\|t-s\|\leq \delta n^{-1/2}} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon \right\} \;\leq\; P\left\{ \sup_{\rho(t,s)<\delta_0} |\mathbb{G}_n(\phi_{0,t} - \phi_{0,s})| > \epsilon \right\} \;<\; \eta \;,$$

which completes the proof. □

Appendix B

PROOF OF RESULTS FROM CHAPTER 3

Proof of Lemma 4. Since $\theta(x) = \psi(\Phi(x))$ and $\psi = -\partial_{-}\mathrm{LCM}_{[a,b]}(-\Gamma \circ \Phi^{-})$, where LCM is the least concave minorant operator, and $-\Gamma \circ \Phi^{-}$ is by assumption upper semi-continuous, the standard switch relation (e.g., Lemma 4.1 of van der Vaart and van der Laan, 2006, Lemma 3.2 of Groeneboom and Jongbloed, 2014) implies that $\theta(x) > c$ if and only if $\sup \mathrm{argmax}_{u\in[a,b]} \{cu - \Gamma(\Phi^{-}(u))\} < \Phi(x)$. We note that the set of maximizers is closed because $cu - \Gamma(\Phi^{-}(u))$ is upper semi-continuous.

If $c \neq 0$, the argmax can only contain elements in the range of Φ, since on intervals where Φ^{-} is constant, the function can be made larger by taking u to one end of the interval – which end of the interval depends on the sign of c. We have used here the fact that a and b are by assumption in the range of Φ. If $c = 0$, taking sup of the argmax ensures that the result will be at the right end of an interval. This shows that

$$\sup \operatorname*{argmax}_{u\in[a,b]} \left\{cu - \Gamma(\Phi^{-}(u))\right\} < \Phi(x) \quad \text{iff} \quad \sup \operatorname*{argmax}_{u\in J^*} \left\{cu - \Gamma(\Phi^{-}(u))\right\} < \Phi(x) \ ,$$

where $J^* := [a, b] \cap \mathrm{range}(\Phi)$. Let $\hat{u} = \sup \mathrm{argmax}_{u\in J^*} \{cu - \Gamma(\Phi^{-}(u))\}$. Because Φ^{-} is strictly increasing on range(Φ) and hence $\Phi^{-} = \Phi^{-1}$ on range(Φ), and furthermore, $u \in$ range(Φ) if and only if $\Phi(\Phi^{-}(u)) = u$, for every $u \in J^*$ there is a unique $v \in I^*$ such that $v = \Phi^{-}(u)$ and $\Phi(v) = u$. Let $\hat{v} \in I^*$ be such an element corresponding to $\hat{u}$. Then, we have that $\hat{u} < \Phi(x)$ if and only if $\Phi(\hat{v}) < \Phi(x)$, $c\Phi(\hat{v}) - \Gamma(\hat{v}) \geq c\Phi(v) - \Gamma(v)$ for all $v \in I^*$, and for any $v \in I^*$ such that equality holds, $v < \hat{v}$. Finally, $\Phi(\hat{v}) < \Phi(x)$ if and only if $\hat{v} < \Phi^{-}(\Phi(x))$ since $\hat{v} \in I^*$ and Φ is right-continuous and non-decreasing. It follows that $\theta(x) > c$ if and

only if

$$\sup \operatorname*{argmax}_{v \in I^*} \{c\Phi(v) - \Gamma(v)\} < \Phi^-(\Phi(x)) \ . \qquad \square$$

Proof of Lemma 5. First, note that $\omega^-(\eta) > 0$ for any $\eta > 0$ by right-continuity of ω at $\eta = 0$. Thus, $\omega(v) < \eta$ for all $v < \omega^-(\eta)$, so that $\omega(\gamma(u)) < \eta$ for all $u < \gamma^{-1}(\omega^-(\eta))$. It is straightforward to see that $c(\delta/2, \eta/2)/2 < \gamma^{-1}(\omega^-(\eta))$, which implies that $r(\delta, \eta) > 0$ for all $\delta, \eta > 0$.

Let $\rho_\eta(d) := \int_0^d [\eta - \omega(\gamma(u))]\, du$. Since Φ_0 is continuous and strictly increasing at x, we have that $\Phi_0^-(\Phi_0(x)) = x$. Recall that $\psi_0 := \theta_0 \circ \Phi_0^-$. Setting $\tilde{\omega} = \omega \circ \gamma$, the moduli of continuity of θ_0 and of Φ_0^- at x imply that $|\psi_0(u) - \psi_0(t)| \leq \tilde{\omega}(|u - t|)$ for all u and $t = \Phi_0(x)$. Note that $\tilde{\omega}^- = \gamma^{-1} \circ \omega^-$.

First, suppose $x \in I_n$ so that $\Phi_n^-(\Phi_n(x)) = x$. Define the functions $R_{n,\eta,x}(u) := \Gamma_n(u) - \Gamma_0(u) - [\eta + \theta_0(x)][\Phi_n(u) - \Phi_0(u)]$ and

$$h_{\eta,t}(u) := [\eta + \psi_0(t)]\Phi_0(u) - \Gamma_0(u) = \int_0^{\Phi_0(u)} [\eta + \psi_0(t) - \psi_0(v)] dv \ .$$

By Lemma 1, we have that $\theta_n(x) - \theta_0(x) > \eta$ holds if and only if

$$\sup \operatorname*{argmax}_{u \in I_n} \{[\eta + \theta_0(x)]\Phi_n(u) - \Gamma_n(u)\} < x \ \text{ iff } \ \sup \operatorname*{argmax}_{u \in I_n} \{h_{\eta,t}(u) - R_{n,\eta,x}(u)\} < x$$
$$\text{iff } \sup_{u \in I_n : u \leq x - \epsilon} \{h_{\eta,t}(u) - R_{n,\eta,x}(u)\} > \sup_{u \in I_n : x - \epsilon \leq u} \{h_{\eta,t}(u) - R_{n,\eta,x}(u)\}$$

for some $\epsilon > 0$. Note that $\sup_{u \in I_n : u \leq x - \epsilon} \{h_{\eta,t}(u) - R_{n,\eta,x}(u)\} \leq h_{\eta,t}(x) + \sup_{u \in I_n, u < x} \{-R_{n,\eta,x}(u)\}$ since $h_{\eta,t}(u)$ is non-decreasing for $u \leq x$. Let $v^+_{n,\eta,t} := \sup\{v \in I_n : v \geq x, \tilde{\omega}(\Phi_0(v) - t) \leq \eta\}$. Then, we can write that $\sup_{u \in I_n : x - \epsilon \leq u} \{h_{\eta,t}(u) - R_{n,\eta,x}(u)\} \geq h_{\eta,t}(v^+_{n,\eta,t}) + \inf_{x \leq u} \{-R_{n,\eta,x}(u)\}$. Hence, we have that $\theta_n(x) - \theta_0(x) > \eta$ implies that

$$h_{\eta,t}(x) + \sup_{u \leq x} \{-R_{n,\eta,x}(u)\} > h_{\eta,t}(v^+_{n,\eta,t}) + \inf_{x \leq u} \{-R_{n,\eta,x}(u)\}$$

$$\text{iff } \; X_{n,\eta} > \int_t^{\Phi_0(v^+_{n,\eta,t})} [\eta + \psi_0(t) - \psi_0(u)]du$$

with the latter statement implying that $X_{n,\eta} > \int_0^{\Phi_0(v^+_{n,\eta,t})-t} [\eta - \tilde{\omega}(u)]du = \rho_\eta(\Phi_0(v^+_{n,\eta,t}) - t)$, where we set $X_{n,\eta} := \sup_{u\in I_n, u<x} -R_{n,\eta,x}(u) + \sup_{u\in I_n, x\le u} R_{n,\eta,x}(u)$. An analogous argument for the opposite tail with $v^-_{n,\eta,t} := \inf\{v \in I_n : v \le x, \tilde{\omega}(t - \Phi_0(v)) \le \eta\}$ shows that $\theta_n(x) - \theta_0(x) < -\eta$ implies that

$$Y_{n,\eta} \ge \int_{\Phi_0(v^-_{n,\eta,t})}^t [\eta + \psi_0(u) - \psi_0(t)]du \ ,$$

which implies that $Y_{n,\eta} \ge \int_0^{t-\Phi_0(v^-_{n,\eta,t})} [\eta - \tilde{\omega}(u)]du = \rho_\eta(t - \Phi_0(v^-_{n,\eta,t}))$, where we have set $Y_{n,\eta} := \sup_{u\in I_n: u\ge x} -R_{n,-\eta,x}(u) + \sup_{u\in I_n: u\le x} R_{n,-\eta,x}(u)$.

Now, we have that $\max(X_{n,\eta}, Y_{n,\eta}) \le 2\|\Gamma_n - \Gamma_0\|_{\infty,I_n} + 2(\eta + |\theta_0(x)|)\|\Phi_n - \Phi_0\|_{\infty,I_n} =: Z_{n,\eta}$. Let $d_{n,\eta}(t) := [\Phi_0(v^+_{n,\eta,t}) - t] \wedge [t - \Phi_0(v^-_{n,\eta,t})]$. Then, since $\eta \ge \tilde{\omega}(u)$ for $u \le d_{n,\eta}(t)$, $d \mapsto \rho_\eta(d)$ is nondecreasing for $d \le d_{n,\eta}(t)$, and hence, $\{|\theta_n(x) - \theta_0(x)| > \eta\} \subseteq \{Z_{n,\eta} \ge \rho_\eta(d_{n,\eta}(t))\}$. Intuitively, since Φ_0 is strictly increasing and continuous, if Φ_n is uniformly close to Φ_0, then $\Phi_0(v^+_{n,\eta,t}) - t$ and $t - \Phi_0(v^-_{n,\eta,t})$ should be close to $\tilde{\omega}^-(\eta)$ with high probability. Therefore, we use the law of total probability with the event $\{d_{n,\eta}(t) < c(\delta,\eta)/2\}$ to see that

$$\{Z_{n,\eta} \ge \rho_\eta(d_{n,\eta}(t))\} \subseteq \{Z_{n,\eta} \ge r(\delta,\eta)\} \cup \{d_{n,\eta}(t) < c(\delta,\eta)/2\}\,.$$

Now, $d_{n,\eta}(t) < c(\delta,\eta)/2$ implies that either $\Phi_0(v^+_{n,\eta,t}) - t < c(\delta,\eta)/2$ or $t - \Phi_0(v^-_{n,\eta,t}) < c(\delta,\eta)/2$. Suppose the former. Then, for all $v \in I_n$ such that $\Phi_0(v) \ge t + c(\delta,\eta)/2$, it must be true that $\tilde{\omega}(\Phi_0(v) - t) > \eta$ and hence $\Phi_0(v) - t \ge \tilde{\omega}^-(\eta)$. Thus, there is no $v \in I_n$ such that $\Phi_0(v) \in t + [c(\delta,\eta)/2, \tilde{\omega}^-(\eta))$, which includes the interval $t + [c(\delta,\eta)/2, c(\delta,\eta))$. Note that $\Phi_0^-(t + \gamma^{-1}(\delta)) \le x + \delta$, which implies that $\Phi_0(x+\delta) \ge t + \gamma^{-1}(\delta) \ge t + c(\delta,\eta)$, and thus Φ_0 is strictly increasing and continuous on $[\Phi_0^{-1}(t + c(\delta,\eta)/2), \Phi_0^{-1}(t + c(\delta,\eta))]$. Hence, there is no $v \in I_n$ also contained in $[\Phi_0^{-1}(t + c(\delta,\eta)/2), \Phi_0^{-1}(t + c(\delta,\eta))]$. Suppose instead that $t - \Phi_0(v^-_{n,\eta,t}) < c(\delta,\eta)/2$. Then, by similar reasoning, there is no $v \in I_n$ also

contained in $[\Phi_0^{-1}(t-c(\delta,\eta)), \Phi_0^{-1}(t-c(\delta,\eta)/2)]$. Since $\Phi_0(x-\delta) \geq 0$ and $\Phi_0(x+\delta) \leq u_n$ by assumption, this implies that Φ_n is constant on at least one of these intervals. Since Φ_0 is strictly increasing and continuous on the intervals, we then have that the supremum distance between Φ_n and Φ_0 on one of these intervals is at least $c(\delta,\eta)/4$. We have now shown that if $x \in I_n$, then

$$\{d_{n,\eta}(t) < c(\delta,\eta)/2\} \subseteq \left\{\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} \geq c(\delta,\eta)/4\right\}.$$

Now, if $x \notin I_n$, then since ψ_n is the left-derivative of $\bar{\Psi}_n$ and Φ_n is right-continuous, we have $\theta_n(x) = \theta_n(x_n)$ for $x_n := \Phi_n^-(\Phi_n(x))) < x$. Hence, we have that

$$\begin{aligned}\{|\theta_n(x) - \theta_0(x)| > \eta\} &\subseteq \{|\theta_n(x_n) - \theta_0(x_n)| > \eta/2\} \cup \{\theta_0(x) - \theta_0(x_n) > \eta/2\} \\ &\subseteq \{|\theta_n(x_n) - \theta_0(x_n)| > \eta/2, x - x_n < \delta/2\} \\ &\qquad \cup \{|\theta_0(x_n) - \theta_0(x)| > \eta/2, x - x_n < \delta/2\} \cup \{x - x_n \geq \delta/2\}.\end{aligned}$$

Because by assumption $\Phi_n(x) \in (0, u_n)$, and so, $x_n \in I_n$, we can use the above inclusion on the first event with δ replaced by $\delta/2$. For the second term, we note that

$$\{|\theta_0(x_n) - \theta_0(x)| > \eta/2, x - x_n < \delta/2\} \subseteq \{\omega^-(\eta/2) \leq x - x_n < \delta/2\}.$$

Hence, we have that

$$\{|\theta_n(x) - \theta_0(x)| > \eta\} \subseteq \{|\theta_n(x_n) - \theta_0(x_n)| > \eta/2, x - x_n < \delta/2\} \cup \{x - x_n \geq \omega^-(\eta/2) \wedge \delta/2\}.$$

The second event implies that Φ_n is constant on $[x - \omega^-(\eta/2) \wedge \delta/2, x]$, and since Φ_0 is strictly increasing and continuous there, it implies that

$$\begin{aligned}\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} &\geq \left(t - \Phi_0(x - \omega^-(\eta/2) \wedge \delta/2)\right)/2 \\ &\geq \gamma^{-1}(\omega^-(\eta/2) \wedge \delta/2)/2 = c(\delta/2, \eta/2)/2\ .\end{aligned}$$

Therefore, we find that $\{|\theta_n(x) - \theta_0(x)| > \eta\}$ is contained in

$$\begin{aligned}\left\{Z_{n,\eta/2} \geq r(\delta/2, \eta/2)\right\} \cup \left\{\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} \geq c(\delta/2, \eta/2)/4\right\} \\ \cup \left\{\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} \geq c(\delta/2, \eta/2)/2\right\} \\ = \left\{Z_{n,\eta/2} \geq r(\delta/2, \eta/2)\right\} \cup \left\{\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} \geq c(\delta/2, \eta/2)/4\right\}.\end{aligned}$$

The pointwise inequality follows.

The uniform inequality follows from the pointwise inclusions. We note that $\sup_{x\in I_{n,\beta}} |\theta_n(x) - \theta_0(x)| > \eta$ implies there is an $x \in I_{n,\beta}$ such that $|\theta_n(x) - \theta_0(x)| > \eta$. Thus, we have that $\{\sup_{x\in I_{n,\beta}} |\theta_n(x) - \theta_0(x)| > \eta\}$ is contained in $\{\exists x \in I_{n,\beta} : |\theta_n(x) - \theta_0(x)| > \eta\}$, which can be decomposed as

$$\begin{aligned}\Big\{\exists x \in I_{n,\beta} : |\theta_n(x) - \theta_0(x)| > \eta, \Phi_n(x) \in (0, u_n)\Big\} \\ \cup \Big\{\exists x \in I_{n,\beta} : |\theta_n(x) - \theta_0(x)| > \eta, \Phi_n(x) \notin (0, u_n)\Big\}.\end{aligned}$$

Since the moduli of continuity are assumed to hold for all x, and by construction, for every $x \in I_{n,\beta}$, $\Phi_0(x - \beta)$ and $\Phi_0(x + \beta)$ are in $[0, u_n]$, the pointwise inclusion can be applied to the first event with $\delta = \beta$. For the second event, note that $\Phi_0(x - \beta), \Phi_0(x + \beta) \in [0, u_n]$ and $\Phi_n(x) \notin (0, u_n)$ imply that $|\Phi_n(x) - \Phi_0(x)| \geq \gamma^{-1}(\beta)$ and

$$\begin{aligned}\{\|\Phi_n - \Phi_0\|_{\infty,I} \geq c(\beta/2, \eta/2)/4)\} \cup \left\{\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} > \gamma^{-1}(\beta)\right\} \\ = \left\{\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} \geq c(\beta/2, \eta/2)/4\right\}.\end{aligned}$$

□

Proof of Theorems 4 and 5. For part 1 of Theorem 4 and parts 1 and 2 of Theorem 5, we use the pointwise tail bound in Lemma 5 with different choices of ω and γ. Since $u_n \to_{\mathrm{P}} u_0$, $\Phi_0(x) \in (0, u_0)$ and $\Phi_n(x) \to_{\mathrm{P}} \Phi_0(x)$, with probability tending to one, $\Phi_n(x) \in (0, u_n)$ and $[x - \delta', x + \delta'] \subset \Phi_0^{-1}([0, u_n])$ for some $\delta' > 0$. For part 2 of Theorem 4 and part 3 of Theorem 5, we use instead the uniform tail bound. We note that for any $\delta, \eta > 0$, $c(\delta, \eta) > 0$

and $r(\delta, \eta) > 0$, which we show in the proof of Lemma 5.

For part 1 of Theorem 4, we take $\omega(v) := [\theta_0(x+v) - \theta_0(x)] \vee [\theta_0(x) - \theta_0(x-v)]$, which is a valid choice since θ_0 is non-decreasing and continuous at x. Since Φ_0 is continuous and strictly increasing in a neighborhood of x, so is Φ_0^-. Since $[0, u_0]$ is bounded, such an invertible γ exists. By the pointwise tail bound in Lemma 5, both $A_{n,1}(\eta)$ and $A_{n,2}$ are $\mathrm{o_P}(1)$ by assumption, and the result follows.

For part 1 of Theorem 5, we consider the pointwise tail bound with $\eta = \eta_n := \eta_0 r_n^{-\alpha_1\alpha_2/(\alpha_1\alpha_2+1)}$. By assumption, $\omega(v) := K_1(x)v^{\alpha_1}$ and $\gamma(v) := K_2(x)v^{\alpha_2}$ are valid choices. Since $\delta > 0$ and $\eta_n \to 0$, $c(\delta/2, \eta_n/2) \sim \eta_n^{1/(\alpha_1\alpha_2)} = \eta_0 r_n^{-1/(1+\alpha_1\alpha_2)}$ for large n. Thus, the second term of the upper bound is $P_0(4r_n\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} \geq \eta_0^2\eta_n^{-1})$ for large n. Since $r_n\|\Phi_n - \Phi_0\|_{\infty,[x-\delta,x+\delta]} = \mathrm{O_P}(1)$ and $\eta_n = o(1)$ by assumption, this term tends to zero. Because $r(\delta/2, \eta_n/2) \sim r_n^{-1}$, the first term of the upper bound is bounded for any $\eta_0 > 0$ as $\|\Gamma_n - \Gamma_0\|_{\infty,I_n}$ and $\|\Phi_n - \Phi_0\|_{\infty,I_n}$ are both $\mathrm{O_P}(r_n^{-1})$ by assumption.

For part 2 of Theorem 5, we take $\omega(v) := 0$ for $v \leq \delta$ since θ_0 is constant on $[x-\delta, x+\delta]$. As before, since Φ_0 is continuous and strictly increasing in a neighborhood of x, such an invertible γ exists. Letting $\eta = \eta_n := \eta_0 r_n^{-1}$, we have $c(\delta/2, \eta_n/2) = \gamma^{-1}(\delta/2) > 0$ for all n, so the second term of the upper bound tends to zero. Since $r(\delta/2, \eta_n/2) \sim r_n^{-1}$, the first term of the upper bound is bounded.

For part 2 of Theorem 4, since it is uniformly continuous, θ_0 admits a uniform modulus of continuity, which we choose as ω. Since Φ_0 is strictly increasing and continuous, Φ_0^- is well-defined and continuous. Since $[0, u_0]$ is compact, Φ_0^{-1} is uniformly continuous and possesses a continuous and invertible uniform modulus of continuity, which we choose as γ. Thus, $c(\beta/2, \eta/2) > 0$ and $r(\beta/2, \eta/2) > 0$ for any $\beta, \eta > 0$, and so, both terms in the uniform upper bound tend to zero by assumption.

For part 3 of Theorem 5, we consider the uniform tail bound with $\eta = \eta_n := \eta_0 r_n^{-\alpha_1\alpha_2/(\alpha_1\alpha_2+1)}$. By assumption, $\omega(v) := K_1 v^{\alpha_1}$ and $\gamma(v) = K_2 v^{\alpha_2}$ are valid choices. With probability tending to one, $\beta_n > r_n^{-1/(1+\alpha_1\alpha_2)}$, $c(\beta_n/2, \eta_n/2) \sim r_n^{-1/(1+\alpha_1\alpha_2)}$, and so, $r_n c(\beta_n/2, \eta_n/2)$ tends to $+\infty$ in probability. Thus, the second term in the upper bound

tends to zero. Since $r(\beta_n/2, \eta_n/2) \sim r_n^{-1}$, the first term in the upper bound is bounded for any $\eta_0 > 0$. □

Proof of Theorem 6. We note that $r_n[\theta_n(x) - \theta_0(x)] > \eta$ if and only if $\theta_n(x) > \theta_0(x) + r_n^{-1}\eta$, which, by Lemma 4, occurs if and only if $\sup \operatorname{argmax}_{v \in I_n} \{[\theta_0(x) + r_n^{-1}\eta]\Phi_n(v) - \Gamma_n(v)\} < \Phi_n^-(\Phi_n(x))$. The latter event occurs if and only if

$$\sup \operatorname*{argmax}_{v \in r_n(I_n - x)} \left\{ \left[\theta_0(x) + r_n^{-1}\eta\right] \Phi_n(x + r_n^{-1}v) - \Gamma_n(x + r_n^{-1}v) \right\} < r_n \left[\Phi_n^-(\Phi_n(x)) - x\right].$$

Since adding terms not depending on v and scaling by constants does not affect the value of the maximizer, we note that the left-hand side of the inequality above equals

$$\sup \operatorname*{argmax}_{v \in r_n(I_n - x)} \{H_{n,x,\eta}(v) + R_{n,1}(v) + R_{n,2}(v) + R_{n,3}(v)\}$$

for $H_{n,x,\eta}(v) := -W_{n,x}(v) + [\eta\Phi_0'(x)]\, v - \left[\frac{1}{2}\theta_0'(x)\Phi_0'(x)\right] v^2$ and remainder terms

$$\begin{aligned}
R_{n,1}(v) &:= r_n\eta \left[\Phi_n(x + r_n^{-1}v) - \Phi_0(x + r_n^{-1}v)\right]; \\
R_{n,2}(v) &:= r_n\eta \left[\Phi_0(x + r_n^{-1}v) - \Phi_0(x) - \Phi_0'(x)(r_n^{-1}v)\right]; \\
R_{n,3}(v) &:= -r_n^2 \left[\Gamma_0(x + r_n^{-1}v) - \theta_0(x)\Phi_0(x + r_n^{-1}v) - \Gamma_0(x) \right. \\
&\qquad \left. + \theta_0(x)\Phi_0(x) - \tfrac{1}{2}\Phi_0'(x)\theta_0'(x)(r_n^{-1}v)^2\right].
\end{aligned}$$

By Slutsky's Theorem, $\{H_{n,x,\eta}(v) : |v| \le M\}$ converges weakly to $\{H_{x,\eta}(v) : |v| \le M\}$ in $\ell^\infty[-M, M]$ for every $M > 0$ for $H_{x,\eta}(v) := -W_x(v) + [\eta\Phi_0'(x)]\, v - \left[\frac{1}{2}\theta_0'(x)\Phi_0'(x)\right] v^2$. By the uniform consistency of Φ_n to Φ_0 at rate faster than r_n^{-1} in a neighborhood of x, $\sup_{|v| \le M} |R_{n,1}(v)| = o_P(1)$ for all $M > 0$. Continuous differentiability of Φ_0 at x gives $\sup_{|v| \le M} |R_{n,2}(v)| = o(1)$ for all $M > 0$. Define $M_{0,x}(u) := [\Gamma_0(x + u) - \theta_0(x)\Phi_0(x + u)] - [\Gamma_0(x) - \theta_0(x)\Phi_0(x)]$. Clearly, $M_{0,x}(0) = 0$, and $M_{0,x}'(u) = \Phi_0'(x + u)[\theta_0(x + u) - \theta_0(x)]$ for u in a neighborhood of 0, so that $M_{0,x}'(0) = 0$. Furthermore, $M_{0,x}'(u)/u \to \Phi_0'(x)\theta_0'(x)$

as $u \to 0$ by the differentiability of θ_0 and continuity of Φ_0' at x, and so, $M_{0,x}(u)$ is twice differentiable at $u = 0$ with $M_{0,x}''(0) = \Phi_0'(x)\theta_0'(x)$. Hence, by Taylor's theorem, $M_{0,x}(u) = \frac{1}{2}\Phi_0'(x)\theta_0'(x)u^2 + h_2(u)u^2$, where $h_2(u) \to 0$ as $u \to 0$, from which it follows that $\sup_{|v|\leq M} |R_{n,3}(v)| = \mathrm{o}(1)$ for all $M > 0$. In view of these findings, writing $R_n := R_{n,1} + R_{n,3} + R_{n,3}$, we have that $\{H_{n,x,\eta}(v) + R_n(v) : |v| \leq M\}$ converges weakly to $\{H_{x,\eta}(v) : |v| \leq M\}$ for every $M > 0$. Since there is a neighborhood of x in which Φ_0 is strictly increasing and Φ_n^- is uniformly consistent, $r_n(I_n - x) \to \mathbb{R}$ in probability. Therefore, the argmax continuous mapping theorem (Theorem 3.2.2 of VW) implies that

$$\hat{v}_n(x,\eta) := \sup \operatorname*{argmax}_{v \in r_n(I_n - x)} \{H_{n,x,\eta}(v) + R_n(v)\} \xrightarrow{\mathrm{d}} \sup \operatorname*{argmax}_{v \in \mathbb{R}} \{H_{x,\eta}(v)\} =: \hat{v}(x,\eta)$$

as long as $\hat{v}_n(x,\eta) = \mathrm{O_P}(1)$. This fact is established in Lemma 7 below. Since $r_n \sup_{|u-x|\leq\delta} |\Phi_n(u) - \Phi_0(u)| = \mathrm{o_P}(1)$ by assumption and Φ_0 is continuously differentiable at x with positive derivative, $r_n[\Phi_n^-(\Phi_n(x)) - x] = \mathrm{o_P}(1)$. Thus, we find that

$$P_0\left(r_n\left[\theta_n(x) - \theta_0(x)\right] > \eta\right) = P_0\left(\hat{v}_n(x,\eta) - r_n\left[\Phi_n^-(\Phi_n(x)) - x\right] < 0\right) \to P_0\left(\hat{v}(x,\eta) < 0\right) \ .$$

Using properties of the argmax functional and the stationary increments of W_x, we have that

$$\begin{aligned} P_0\left(\hat{v}(x,\eta) < 0\right) &= P_0\left(-\theta_0'(x) \operatorname*{argmin}_{u\in\mathbb{R}} \left\{W_x(u + \eta/\theta_0'(x)) + \tfrac{1}{2}\theta_0'(x)\Phi_0'(x)u^2\right\} > \eta\right) \\ &= P_0\left(-\theta_0'(x) \operatorname*{argmin}_{u\in\mathbb{R}} \left\{W_x(u) + \tfrac{1}{2}\theta_0'(x)\Phi_0'(x)u^2\right\} > \eta\right) \end{aligned}$$

for each η, and the result follows by the Portmanteau Theorem. Finally, if $W_x = [\kappa_0(x)]^{1/2} W_0$ for W_0 a standard two-sided Brownian motion, a standard argument (see Problem 3.2.5 of VW) shows that

$$\operatorname*{argmin}_{u\in\mathbb{R}} \left\{W_x(u) + \tfrac{1}{2}\theta_0'(x)\Phi_0'(x)u^2\right\} \overset{d}{=} \left\{\frac{2[\kappa_0(x)]^{1/2}}{\theta_0'(x)\Phi_0'(x)}\right\}^{2/3} \operatorname*{argmin}_{u\in\mathbb{R}} \left\{W_0(u) + u^2\right\}. \qquad \square$$

Lemma 7. *Under the conditions of Theorem 3,* $\sup \operatorname{argmax}_{v \in r_n(I_n - x)} \{H_{n,x,\eta}(v) + R_n(v)\} = \mathrm{O_P}(1)$.

Proof of Lemma 7. To establish that $\hat{v}_n = \mathrm{O_P}(1)$, we use Theorem 3.2.5 of VW. Write $r_n^{-1}\hat{v}_n(x,\eta) = \sup \operatorname*{argmax}_{v \in I_n - x} M_{n,x}(v)$ for

$$M_{n,x}(v) := -[\Gamma_n(x+v) - \Gamma_n(x)] + \theta_0(x)[\Phi_n(x+v) - \Phi_n(x)] + \eta r_n^{-1}\Phi_n(x+v) \ .$$

Defining $M_{0,x}(v) := [\Gamma_0(x+v) - \theta_0(x)\Phi_0(x+v)] - [\Gamma_0(x) - \theta_0(x)\Phi_0(x)]$, we have that $-M_{0,x}$ is twice differentiable at $u = 0$ with negative second derivative, so that $-M_{0,x}(v) \leq -cu^2$ for v in a neighborhood of 0 and some $c > 0$. The next requirement concerns the modulus of continuity of $M_{n,x}(v) - M_{0,x}(v)$, which we can write as

$$\begin{aligned} & E_0 \left[\sup_{|v|<\delta} |(M_{n,x} - M_{0,x})(v) - (M_{n,x} - M_{0,x})(0)| \right] \\ & \quad = \ E_0 \left[\sup_{|v|<\delta} \left| \eta r_n^{-1}[\Phi_n(x+v) - \Phi_n(x)] - r_n^{-2} W_{n,x}(r_n v) \right| \right] \\ & \quad \leq \ r_n^{-2} E_0 \left[\sup_{|u|<\delta r_n} |W_{n,x}(u)| \right] + |\eta| r_n^{-1} E_0 \left[\sup_{|v|\leq\delta} |\Phi_n(x+v) - \Phi_n(x))| \right] . \end{aligned}$$

By assumption, the first term is bounded by $r_n^{-2} f_n(r_n\delta)$. Taking differences with Φ_0, and since Φ_0 is continuously differentiable at x, we can find δ small enough such that for any $\epsilon > 0$ the second term is bounded up to a constant by $r_n^{-1}|\eta|(\epsilon r_n^{-1} + \delta)$. We thus have that the above expression is bounded up to a constant by

$$\tilde{f}_n(\delta) := r_n^{-2} f_n(r_n\delta) + r_n^{-1}|\eta|(\epsilon r_n^{-1} + \delta) \ .$$

By assumption, $\delta \mapsto \delta^{-\alpha}\tilde{f}_n(\delta)$ is decreasing for some $\alpha \in (1,2)$. Additionally, $r_n^2 \tilde{f}_n(r_n^{-1}) = f_n(1) + |\eta|(1+\epsilon) = \mathrm{O}(1)$. If we can establish that $r_n^{-1}\hat{v}_n(x,\eta) = \mathrm{o_P}(1)$, we will have checked all the conditions of Theorem 3.2.5 of VW, yielding $r_n^{-1}\hat{v}_n(x,\eta) = \mathrm{O_P}(r_n^{-1})$ and hence $\hat{v}_n(x,\eta) =$

$O_P(1)$.

Simplifying further, we have that $r_n^{-1}\hat{v}_n(x,\eta) = -x + \sup \operatorname{argmax}_{v\in I_n} \tilde{M}_{n,x}(v)$, where $\tilde{M}_{n,x}(v) := -\Gamma_n(v) + \Phi_n(v)[\theta_0(x) + r_n^{-1}\eta]$. Setting

$$\begin{aligned} h_x(v) &:= \psi_0(t)\Phi_0(v) - \Gamma_0(v) = \int_0^{\Phi_0(v)} [\psi_0(t) - \psi_0(u)]du \\ R_{n,x}(v) &:= \eta r_n^{-1}\Phi_n(v) + \theta_0(x)[\Phi_n(v) - \Phi_0(v)] - [\Gamma_n(v) - \Gamma_0(v)] \ , \end{aligned}$$

write $\tilde{M}_{n,x}(v) = h_x(v) + R_{n,x}(v)$. Note that $\sup_{v\in I_n} |R_{n,x}(v)| = o_P(1)$, and that h_x is unimodal and maximized at $v = x$, but x may not be in I_n for any n. Define $x_n^+ := \inf\{v \in I_n : v \geq x\}$ and let $\epsilon > 0$. Then, $\sup \operatorname{argmax}_{v\in I_n} \tilde{M}_{n,x}(v) \leq x - \epsilon$ implies that $h_x(x-\epsilon) + \sup_{v\in I_n : v<x} R_{n,x}(v) > h_x(x_n^+) + \inf_{v\in I_n : v\geq x} R_{n,x}(v)$, which in turn implies that

$$2 \sup_{v\in I_n} |R_{n,x}(v)| + \int_t^{\Phi_0(x_n^+)} [\psi_0(v) - \psi_0(t)]dv > \int_{\Phi_0(x-\epsilon)}^t [\psi_0(t) - \psi_0(v)]dv \ .$$

Since Φ_0 and ψ_0 are differentiable with positive derivative at x and t, respectively, for all $\epsilon > 0$, $\int_{\Phi_0(x-\epsilon)}^t [\psi_0(t) - \psi_0(v)]dv =: \delta_\epsilon > 0$. Additionally, by the boundedness of ψ_0,

$$\int_t^{\Phi_0(x_n^+)} [\psi_0(v) - \psi_0(t)]dv \leq c\left[\Phi_0(x_n^+) - \Phi_0(x)\right]$$

for some $c < \infty$. We claim that $x_n^+ \xrightarrow{P} x$. To see this, first note that $x_n^+ > x + \epsilon'$ implies that $\Phi_n(x) = \Phi_n(x+\epsilon')$. Hence, for all $0 \leq u \leq \delta \wedge \epsilon'$, we have that

$$[\Phi_n(x) - \Phi_0(x)] - [\Phi_n(x+u) - \Phi_0(x+u)] = \Phi_0(x+u) - \Phi_0(x) \geq c'u$$

for some $c' > 0$, again using that Φ_0 is differentiable with positive derivative at x. This implies that $0 < (\delta \wedge \epsilon)c' \leq 2\sup_{|u|<\delta} |\Phi_n(x+u) - \Phi_0(x+u)|$, the probability of which goes to zero for any $\epsilon > 0$. Hence, $x_n^+ \xrightarrow{P} x$, and so, $\Phi_0(x_n^+) \xrightarrow{P} \Phi_0(x)$ by the Continuous Mapping

Theorem. We have shown that

$$P_0\left(\sup \operatorname*{argmax}_{v\in I_n} \tilde{M}_{n,x}(v) - x \le -\epsilon\right) \le P_0(\mathrm{o_P}(1) > \delta_\epsilon) \ ,$$

which goes to 0 for each $\epsilon > 0$. The argument for the opposite tail probability is completely analogous, and hence $\sup \operatorname{argmax}_{v\in I_n} \tilde{M}_{n,x}(v) \xrightarrow{\mathrm{P}} x$. □

Proof of Theorem 7. We use Theorems 2.11.22 and 2.11.23 of VW to show weak convergence of $W_{n,x}$ to $[\kappa_0(x)]^{1/2}W_0$. In their notation, $f_{n,u} = n^{1/6}g_{x,un^{-1/3}}$ and $\mathcal{F}_{n,M} = \{f_{n,u} : |u| \le M\} = n^{1/6}\mathcal{G}_{x,Mn^{-1/3}}$ with envelope $F_{n,M} = n^{1/6}G_{Mn^{-1/3}}$. Thus, we have that $P_0F_{n,M}^2 = n^{1/3}P_0G_{x,Mn^{-1/3}}^2 = n^{1/3}\,\mathrm{O}(Mn^{-1/3}) = \mathrm{O}(1)$ for each $M > 0$ by (C2). For any $\epsilon > 0$ and $\eta > 0$, $R^{-1}P_0G_{x,R}^2\{G_{x,R} > \eta(MR)^{-1}\} < M\epsilon$ for all R small enough, so that after some rearrangement, for all n large enough,

$$P_0F_{n,M}^2\{F_{n,M} > \eta n^{1/2}\} < \epsilon \ .$$

In the case of Theorem 2.11.23, we will use the first possibility of (C1a) to establish the convergence of the bracketing entropy integral:

$$\begin{aligned}
&\int_0^{\delta_n} \left[\log N_{[]}(\varepsilon\|F_{n,M}\|_{P_0,2}, \mathcal{F}_{n,M}, L_2(P_0))\right]^{1/2} d\varepsilon \\
&\quad = \int_0^{\delta_n} \left[\log N_{[]}(\varepsilon n^{1/6}\|G_{x,Mn^{-1/3}}\|_{P_0,2}, n^{1/6}\mathcal{G}_{x,Mn^{-1/3}}, L_2(P_0))\right]^{1/2} d\varepsilon \\
&\quad = \int_0^{\delta_n} \left[\log N_{[]}(\varepsilon\|G_{x,Mn^{-1/3}}\|_{P_0,2}, \mathcal{G}_{x,Mn^{-1/3}}, L_2(P_0))\right]^{1/2} d\varepsilon \\
&\quad = \mathrm{O}\left(\int_0^{\delta_n} \varepsilon^{-V/2} d\varepsilon\right) \ = \ \mathrm{O}\left(\frac{\delta_n^{-V/2+1}}{-V/2+1}\right) \ \to \ 0
\end{aligned}$$

for all $\delta_n \to 0$. The calculation for the uniform entropy integral using the second possibility (C1b) to establish Theorem 2.11.22 is identical.

We now show that (C3) implies that, for all δ small enough, $\sup_{|u-v|<\delta} P_0(g_{x,u} - g_{x,v})^2 = \mathrm{O}(\delta)$ and that $\alpha^{-1}[P_0(g_{x,\alpha u}g_{x,\alpha v}) - P_0g_{x,\alpha u}P_0g_{x,\alpha v}] \to \sigma^2(u,v)\kappa_0(x)$ as $\alpha \to 0$, where

$\sigma^2(u,v) := (u \wedge v) - uI_{(-\infty,0)}(u) - vI_{(-\infty,0)}(v)$ is the covariance of a two-sided Brownian motion. Then we will have that

$$\sup_{|s-t|<\delta_n} P_0(f_{n,s} - f_{n,t})^2 = n^{1/3} \sup_{|u-v|<\delta_n n^{-1/3}} P_0(g_{x,u} - g_{x,v})^2 = \mathrm{O}\left(n^{1/3}\delta_n n^{-1/3}\right) = \mathrm{O}\left(\delta_n\right) \to 0$$

for all $\delta_n \to 0$ and that $P_0 f_{n,u} f_{n,v} - P_0 f_{n,u} P_0 f_{n,v} = n^{1/3} P_0 g_{x,un^{-1/3}} g_{x,vn^{-1/3}} - n^{1/3} P_0 g_{x,un^{-1/3}} P_0 g_{x,vn^{-1/3}}$ tends to $\sigma^2(u,v)\kappa_0(x)$; both of these statements are conditions of Theorems 2.11.22 and 2.11.23 of VW.

Writing $s := x+u$ and $t := x+v$, we can show that $P_0(g_{x,u} - g_{x,v})^2 = \Sigma_0(s,s) - 2\Sigma_0(s,t) + \Sigma_0(t,t)$. Hence, for the first claim, it is sufficient to show that $|\Sigma_0(s,s) - \Sigma_0(s,t)| = \mathrm{O}(|s-t|)$ for all s,t in a neighborhood of x. By assumption, Σ_0^* is continuously differentiable at (x,x), which implies that $|\Sigma_0^*(s,s) - \Sigma_0^*(s,t)| = \mathrm{O}(|s-t|)$ for s,t in a neighborhood of x. We can decompose $\iint_{-\infty}^{s\wedge t} A_0(s,t,u,w) H_0(du,w) Q_0(dw)$ as $\bar{\Sigma}_0(s,t) + \tilde{\Sigma}_0(s,t)$, where we set

$$\begin{aligned}
\bar{\Sigma}_0(s,t) &:= \iint_{-\infty}^{x} A_0(s,t,u,w) H_0(du,w) Q_0(dw), \\
\tilde{\Sigma}_0(s,t) &:= \iint_{x}^{s\wedge t} A_0(s,t,u,w) H_0(du,w) Q_0(dw) \ .
\end{aligned}$$

By (C3b), $\bar{\Sigma}_0$ is continuously differentiable at (x,x), which implies that $|\bar{\Sigma}_0(s,s) - \bar{\Sigma}_0(s,t)| = \mathrm{O}(|s-t|)$ for s,t in a neighborhood of x. For $\tilde{\Sigma}_0$, we have that $|\tilde{\Sigma}_0(s,t) - \tilde{\Sigma}_0(s,s)|$ is bounded above by

$$\begin{aligned}
&\iint_{x}^{s} \left|A_0(s,s,u,w) - A_0(s,t,u,w)\right| H_0(du,w) Q_0(dw) \\
&\qquad + \iint_{s}^{s\wedge t} \left|A_0(s,t,u,w)\right| H_0(du,w) Q_0(dw) \ .
\end{aligned}$$

Continuous differentiability of A_0 around (x,x) implies that the first summand is bounded above by

$$|s-t| \iint_{x}^{s} \sup_{s,t\in B_\delta(x)} |A_0'(s,t,u,w)| H_0(du,w) Q_0(dw)$$

for s, t close enough to x, which is bounded up to a constant by $|s-t|$ by assumption. Boundedness of A_0 and continuity of H_0 around x for all w yields the same for the second term.

For the second claim, we first note that the contribution of $\bar{\Sigma}_0$ to $\frac{1}{\alpha}[P_0(g_{x,\alpha u}g_{x,\alpha v}) - P_0 g_{x,\alpha u} P_0 g_{x,\alpha v}] = \frac{1}{\alpha}\left[\Sigma_0(x+\alpha u, x+\alpha v) - \Sigma_0(x+\alpha u, x) - \Sigma_0(x, x+\alpha v) + \Sigma_0(x,x)\right]$ is

$$\begin{aligned}\tfrac{1}{\alpha}\left[\bar{\Sigma}_0(x+\alpha u, x+\alpha v) - \bar{\Sigma}_0(x,x)\right] - \tfrac{1}{\alpha}\left[\bar{\Sigma}_0(x+\alpha u, x) - \bar{\Sigma}_0(x,x)\right] \\ - \tfrac{1}{\alpha}\left[\bar{\Sigma}_0(x, x+\alpha v) - \bar{\Sigma}_0(x,x)\right],\end{aligned}$$

which, due to the differentiability of $\bar{\Sigma}_0$, tends to $(u+v)\bar{\Sigma}_0'(x,x) - u\bar{\Sigma}_0'(x,x) - v\bar{\Sigma}_0'(x,x) = 0$ as $\alpha \to 0$. Similarly, Σ_0^* does not contribute to the limit. The contribution of $\tilde{\Sigma}_0$ therefore determines the limit entirely. For any fixed r and w, we note that

$$\frac{1}{\alpha}\int_x^{x+\alpha r} A_0(x,x,u,w) H_0(du,w) \longrightarrow r A_0(x,x,x,w) H_0'(x,w)$$

as $\alpha \to 0$ by the continuous differentiability of $u \mapsto H_0(u,w)$ at $u = x$ and the continuity of $u \mapsto A_0(x,x,u,w)$. Since the continuity of $x \mapsto A_0(x,x,x,w)H_0'(x,w)$ is uniform in w and these functions are Q_0-integrable, by the Dominated Convergence Theorem, for any fixed r, we have that

$$\frac{1}{\alpha}\iint_x^{x+\alpha r} A_0(x,x,u,w) H_0(du,w) Q_0(dw) \longrightarrow r \int A_0(x,x,x,w) H_0'(x,w) Q_0(dw)\ .$$

We then find that $\frac{1}{\alpha}[\tilde{\Sigma}_0(x+\alpha u, x+\alpha v) - \tilde{\Sigma}_0(x+\alpha u, x) - \tilde{\Sigma}_0(x, x+\alpha v) + \tilde{\Sigma}_0(x,x)]$ can be written, up to a remainder term tending to zero as $\alpha \to 0$, as

$$\begin{aligned}\frac{1}{\alpha}\iint \left[I_{(x,x+\alpha(u\wedge v))}(y) - I_{(-\infty,0)}(u) I_{(x,x+\alpha u)}(y) - I_{(-\infty,0)}(v) I_{(x,x+\alpha v)}(y)\right] \\ \times A_0(x,x,y,w) H_0(dy,w) Q_0(dw)\end{aligned}$$

limiting to $\left[u \wedge v - I_{(-\infty,0)}(u)u - I_{(-\infty,0)}(v)v\right] \int A_0(x,x,x,w) H_0'(x,w) Q_0(dw)$, the claimed

covariance. The remainder term we left out can be expressed as

$$\begin{aligned}\frac{1}{\alpha}\iint_x^{x+\alpha(u\wedge v)} &\left[A_0(x+\alpha u, x+\alpha v, y, w) - A_0(x,x,y,w)\right] H_0(dy,w)Q_0(dw) \\ &- I_{(-\infty,0)}(u)\frac{1}{\alpha}\iint_x^{x+\alpha u}\left[A_0(x+\alpha u, x, y, w) - A_0(x,x,y,w)\right] H_0(dy,w)Q_0(dw) \\ &- I_{(-\infty,0)}(v)\frac{1}{\alpha}\iint_x^{x+\alpha v}\left[A_0(x, x+\alpha v, y, w) - A_0(x,x,y,w)\right] H_0(dy,w)Q_0(dw)\ .\end{aligned}$$

For α small enough the absolute value of each inner difference is bounded by $\alpha(|u| \vee |v|)|A_0'(x,x,y,w)|$. Since $y \mapsto A_0'(x,x,y,w)$ is continuous and $y \mapsto H_0(y,w)$ is differentiable in a neighborhood of x uniformly in w, for α small enough, the absolute value of the remainder is bounded up to a constant by

$$\iint \left[I_{(x,x+\alpha(u\wedge v))}(y) + I_{(-\infty,0)}(u)I_{(x,x+\alpha u)}(y) + I_{(-\infty,0)}(v)I_{(x,x+\alpha v)}(y)\right] H_0(dy,w)Q_0(dw)\ .$$

Since $y \mapsto H_0'(y,w)$ is bounded near x uniformly in w, this bound tends to zero as $\alpha \to 0$. This, in addition to condition (C4), proves (B1). Since $\theta_0'(x)$ and $\Phi_0'(x)$ are assumed positive, (B2) is also satisfied. For (B3), we note that

$$E_0\left[\sup_{|u|\le\delta n^{1/3}}|\mathbb{G}_n f_{n,u}|\right] = n^{1/6}E_0\left[\sup_{|u|\le\delta n^{1/3}}|\mathbb{G}_n g_{x,un^{-1/3}}|\right] = \mathrm{O}\left(\delta^{1/2}n^{1/6}\right)$$

for all n large enough is also implied by assumption (C1) and Theorems 2.14.1 and 2.14.2 of VW. The remainder term satisfies (B3) by condition (C5). □

Regularity conditions and proof of Theorem 8. Regularity conditions for Theorem 8 include that $(s,t,u,z) \mapsto M_{s,0}^{(1)}(u,z)M_{t,0}^{(1)}(u,z)$ and $(s,t,u,z) \mapsto L_{s,0}^{(1)}(u,z)L_{t,0}^{(1)}(u,z)$ satisfy (C3b) and (C3c), and that the following maps are continuously differentiable in (s,t) in a neighborhood of (x,x):

$$(s,t) \mapsto E_0\left[I_{[0,s]}(U)M_{s,0}^{(1)}(O)M_{t,0}^{(2)}(O)\right],\ (s,t) \mapsto E_0\left[I_{[0,s]}(U)M_{s,0}^{(1)}(O)D_{t,0}^{(2)}(O)\Phi_0'(U)\right],$$

$$(s,t) \mapsto E_0\left[I_{[0,s]}(U)M^{(1)}_{s,0}(O)L^{(2)}_{t,0}(O)\Phi_0'(U)\right],\ (s,t) \mapsto E_0\left[I_{[0,s]}(U)L^{(1)}_{s,0}(O)D^{(2)}_{t,0}(O)\right],$$
$$(s,t) \mapsto E_0\left[I_{[0,s]}(U)L^{(1)}_{s,0}(O)L^{(2)}_{t,0}(O)\right],\ (s,t) \mapsto E_0\left[M^{(2)}_{s,0}(O)M^{(2)}_{t,0}(O)\right],$$
$$(s,t) \mapsto E_0\left[D^{(2)}_{s,0}(O)D^{(2)}_{t,0}(O)\right],\ (s,t) \mapsto E_0\left[D^{(2)}_{s,0}(O)L^{(2)}_{t,0}(O)\right],$$
$$(s,t) \mapsto E_0\left[L^{(2)}_{s,0}(O)L^{(2)}_{t,0}(O)\right].$$

We first examine the covariance arising from the use of Θ_n and the identity transformation. Writing $H_0 : (u,z) \mapsto P_0(U \le u \mid Z = z)$, we have that $\Sigma_0(s,t) = P_0(M^*_{s,0}M^*_{t,0})$ is equal to

$$\begin{aligned}
&\int \left[I_{[0,s]}(u)M^{(1)}_{s,0}(u,z) + M^{(2)}_{s,0}(u,z)\right]\left[I_{[0,t]}(u)M^{(1)}_{t,0}(u,z) + M^{(2)}_{t,0}(u,z)\right]P_0(du,dz)\\
&= \iint_0^{s\wedge t} M^{(1)}_{s,0}(u,z)M^{(1)}_{t,0}(u,z)H_0(du,z)Q_0(dz)\\
&\qquad + \int \Big[I_{[0,s]}(u)M^{(1)}_{s,0}(u,z)M^{(2)}_{t,0}(u,z)\\
&\qquad\qquad + I_{[0,t]}(u)M^{(1)}_{t,0}(u,z)M^{(2)}_{s,0}(u,z) + M^{(2)}_{t,0}(u,z)M^{(2)}_{s,0}(u,z)\Big]P_0(du,dz)\ .
\end{aligned}$$

By assumption, the second summand plays the role of $\Sigma^*_0(s,t)$ and satisfies (C3a). The first summand satisfies (C3b) and (C3c) with $A_0(s,t,u,z) = M^{(1)}_{s,0}(u,z)M^{(1)}_{t,0}(u,z)$ by assumption, and $H_0(u,z)$ satisfies (C3d) with $H_0'(u,z) = h_0(u|z)$ equal to the conditional density of U given $Z = z$. Therefore, the scale factor for the Chernoff distribution in Theorem 4 is equal to $[4\theta_0'(x)\kappa_0(x)]^{1/3}$, where

$$\kappa_0(x) = \int \left[M^{(1)}_{x,0}(x,z)\right]^2 h_0(x \mid z)Q_{Z,0}(dz).$$

We then examine the covariance arising from the use of Γ_n and transformation Φ_n. Using integration by parts, we find that $D^*_{s,0}(o) - \theta_0(x)L^*_{s,0}(o)$ is equal to $I_{[0,s]}(u)\Upsilon_{1,s,x}(u,z) +$

$\Upsilon_{2,s,x}(u,z)$, where

$$\Upsilon_{1,s,x} : (u,z) \mapsto M^{(1)}_{s,0}(u,z)\Phi_0'(u) - \int_u^s L^{(1)}_{v,0}(u,z)\theta_0(dv) + [\theta_0(s) - \theta_0(x)]\, L^{(1)}_{s,0}(u,z) \ ,$$
$$\Upsilon_{2,s,x} : (u,z) \mapsto D^{(2)}_{s,0}(u,z) - \int_0^s L^{(2)}_{v,0}(u,z)\theta_0(dv) + [\theta_0(s) - \theta_0(x)]\, L^{(2)}_{s,0}(u,z) \ .$$

The covariance $\Sigma_0(s,t) = P_0[D^*_{s,0} - \theta_0(x)L^*_{s,0}][D^*_{t,0} - \theta_0(x)L^*_{t,0}]$ can then be written as the sum $\Sigma_{0,1}(s,t) + \Sigma_{0,2}(s,t) + \Sigma_{0,3}(s,t) + \Sigma_{0,4}(s,t)$ of all cross-product terms. The sum $\Sigma_{0,2} + \Sigma_{0,3} + \Sigma_{0,4}$ constitutes Σ^*_0, where the summands are defined pointwise as $\Sigma_{0,2}(s,t) = \iint I_{[0,s]}(u)\Upsilon_{1,s,x}(u,z)\Upsilon_{2,t,x}(u,z)P_0(du,dz)$, $\Sigma_{0,3}(s,t) = \Sigma_{0,2}(t,s)$ and $\Sigma_{0,4}(s,t) = \iint \Upsilon_{2,s,x}(u,z)\Upsilon_{2,t,x}(u,z)P_0(du,dz)$. By assumption, each of these expressions is continuously differentiable in (s,t) in a neighborhood of (x,x). Finally, we have $\Sigma_{0,1}(s,t) = \iint I_{[0,s\wedge t]}(u)\Upsilon_{1,s,x}(u,z)\Upsilon_{1,t,x}(u,z)H_0(du,z)Q_{Z,0}(dz)$. The product $\Upsilon_{1,s,x}(u,z)\Upsilon_{1,t,x}(u,z)$ forms $A_0(s,t,u,z)$, which satisfies (C3b) and (C3c) by assumption. Hence, in this case, the scale parameter is $[4\theta_0'(x)\kappa_0^*(x)/\Phi_0'(x)^2]^{1/3}$ in view of Theorem 7, where

$$\kappa_0^*(x) = \int \left[M^{(1)}_{x,0}(x,z)\Phi_0'(x)\right]^2 h_0(x \mid z)Q_{z,0}(dz) = \Phi_0'(x)^2\kappa_0(x) \ .$$

Thus, the scale factor obtained coincides with that obtained with Θ_n and identity transformation. □

Proof of Theorem 9. Let $\mathcal{F}_{x,n,\delta} := \{n^{1/6}g_{x,un^{-1/3}}(\pi) : |u| \le \delta, \pi \in \mathcal{P}\} = n^{1/6}\mathcal{G}_{x,\mathcal{P},\delta n^{-1/3}}$, which has envelope $F_{x,n,\delta} = n^{1/6}G_{x,\mathcal{P},\delta n^{-1/3}}$. We first show that the process $\{\mathbb{G}_n n^{1/6} g_{x,u/n^{1/3}}(\pi) : |u| \le \delta, \pi \in \mathcal{P}\}$ is asymptotically $\bar{\rho}$-equicontinuous using Theorems 2.11.1 and 2.11.9 of VW, where $\bar{\rho}$ is the product semimetric. We begin by assessing display (2.11.21) of VW. For the first line, we note that $P_0F^2_{x,n,\delta} = n^{1/3}P_0G^2_{x,\mathcal{P},\delta n^{-1/3}} \le c\delta$ for all n large enough, so $P_0F^2_{x,n,\delta} = \mathrm{O}(1)$ as $n \to \infty$ for all fixed δ. For the second line, we have, for any $\eta, \epsilon > 0$,

$$P_0F^2_{x,n,\delta}\{F_{x,n,\delta} > \eta n^{1/2}\} = n^{1/3}P_0G^2_{x,\mathcal{P},\delta n^{-1/3}}\{G_{x,\mathcal{P},\delta n^{-1/3}} > \eta n^{1/3}\}$$

$$= \delta(\delta n^{-1/3})^{-1} P_0 G^2_{x,\mathcal{P},\delta n^{-1/3}} \{G_{x,\mathcal{P},\delta n^{-1/3}} > (\delta\eta)(\delta n^{-1/3})^{-1}\} \ ,$$

which gives $P_0 F^2_{x,n,\delta}\{F_{x,n,\delta} > \eta n^{1/2}\} \leq \delta\epsilon'$ with $\epsilon' := \delta\eta$ for n large enough. Next, we must show that

$$\sup \left\{ n^{1/3} P_0 \left[g_{x,un^{-1/3}}(\pi_1) - g_{x,vn^{-1/3}}(\pi_2) \right]^2 : |u - v| < \delta_n, \rho(\pi_1, \pi_2) < \delta_n \right\} \longrightarrow 0$$

as $n \to \infty$ for all $\delta_n \downarrow 0$. We can bound the square root of $P_0 \left[g_{x,un^{-1/3}}(\pi_1) - g_{x,vn^{-1/3}}(\pi_2) \right]^2$ by

$$\left\{ P_0 \left[g_{x,un^{-1/3}}(\pi_1) - g_{x,vn^{-1/3}}(\pi_1) \right]^2 \right\}^{1/2} + \left\{ P_0 \left[g_{x,vn^{-1/3}}(\pi_1) - g_{x,vn^{-1/3}}(\pi_2) \right]^2 \right\}^{1/2} .$$

By assumption, for all n large enough and up to a multiplicative constant, the first summand is bounded up by $(|u-v|n^{-1/3})^{1/2}$, and the second summand, by $\rho(\pi_1, \pi_2)(|v|n^{-1/3})^{1/2}$. Thus, we find that

$$n^{1/3} P_0 \left[g_{x,un^{-1/3}}(\pi_1) - g_{x,vn^{-1/3}}(\pi_2) \right]^2 = \mathrm{O}\left(\left[|u - v| + |v| \rho(\pi_1, \pi_2) \right]^2 \right) ,$$

uniformly over u, v, π_1 and π_2, which satisfies the requirement. Under (D1a), for any $\delta > 0$ and n large enough, we have that

$$\begin{aligned}
&\int_0^t \left[\sup_Q \log N(\varepsilon \|F_{x,n,\delta}\|_{P_0,2}, \mathcal{F}_{x,n,\delta}, L_2(Q)) \right]^{1/2} d\varepsilon \\
&\quad = \int_0^t \left[\sup_Q \log N(\varepsilon n^{1/6} \|G_{x,\mathcal{P},\delta/n^{1/3}}\|_{P_0,2}, n^{1/6} \mathcal{G}_{x,\mathcal{P},\delta/n^{1/3}}, L_2(Q)) \right]^{1/2} d\varepsilon \\
&\quad = \int_0^t \sup_Q \left[\log N(\varepsilon \|G_{x,\mathcal{P},\delta n^{-1/3}}\|_{P_0,2}, \mathcal{G}_{x,\mathcal{P},\delta n^{-1/3}}, L_2(Q)) \right]^{1/2} d\varepsilon \\
&\quad = \mathrm{O}\left(\int_0^t \varepsilon^V d\varepsilon \right) \ = \ \frac{t^{V+1}}{V+1} \ \to \ 0
\end{aligned}$$

as $t \to 0$ since $V > -1$. An identical analysis holds under (D1b). We have thus verified

the conditions of Theorems 2.11.1 or 2.11.9 of VW, and hence, $\{\mathbb{G}_n n^{1/6} g_{x,un^{-1/3}}(\pi) : |u| \leq \delta, \pi \in \mathcal{P}\}$ is asymptotically $\bar{\rho}$-equicontinuous. Using (D4) and Lemma 4 (stated and proved in the Supplementary Material), we obtain the first statement of the theorem. For the second statement, we use Theorem 2.14.1 and 2.14.2 of VW to obtain that

$$E_0 \left\{ \sup_{|u|\leq\delta,\pi^*\in\mathcal{P}} \left| \mathbb{G}_n n^{1/6} \left[g_{x,un^{-1/3}}(\pi^*) - g_{x,un^{-1/3}}(\pi) \right] \right| \right\} = \mathrm{O}\left(\|F_{x,n,\delta}\|_{P_0,2} \right) = \mathrm{O}\left(\delta^{1/2}\right). \quad \square$$

Proof of Theorem 10. We need to verify conditions (B1)–(B5) for the pair (Γ_n, Φ_n). Let $W^*_{n,x}$ be local process for this pair, which we can write pointwise as

$$\begin{aligned} W^*_{n,x}(u) &= r_n^2 \left\{ \left[S_n(x + ur_n^{-1}) - S_0(x + ur_n^{-1}) \right] - \left[S_n(x) - S_0(x) \right] \right\} \\ &\quad - \theta_0(x) r_n^2 \left\{ \left[\Phi_n(x + ur_n^{-1}) - \Phi_0(x + ur_n^{-1}) \right] - \left[\Phi_n(x) - \Phi_0(x) \right] \right\} \\ &= W_{n,x}(u) - \theta_0(x) r_n^2 \int_x^{x+ur_n^{-1}} \left[S_n(v) - S_0(v) \right] dv \ , \end{aligned}$$

where $W_{n,x}$ is the local process for the pair (Γ_n, Id). We can rewrite the second term as

$$\theta_0(x) r_n^{-1} \int_0^u W_{n,x}(v) dv - u r_n \theta_0(x) \left[S_n(x) - S_0(x) \right].$$

Because for each $M > 0$ we have that $\{W_{n,x}(u) : |u| \leq M\}$ converges weakly in $\ell^\infty[-M, M]$ by (B1), so does $\left\{ \int_0^u W_{n,x}(v) dv : |u| \leq M \right\}$ by the continuous mapping theorem. The latter process is thus uniformly asymptotically negligible when multiplied by r_n^{-1}. The second term is also negligible since $S_n(x) - S_0(x) = \mathrm{o_P}(r_n^{-1})$. It follows then that $W^*_{n,x}$ and $W_{n,x}$ converge weakly to the same limit in $\ell^\infty[-M, M]$ and so, conditions (B1) and (B2) are automatically satisfied for $W^*_{n,x}$. The above expansion gives that $\sup_{|u|\leq\delta r_n} |W^*_{n,x}(u)|$ has mean bounded above by

$$E_0 \left[\sup_{|u|\leq\delta r_n} |W_{n,x}(u)| \right] + \theta_0(x) r_n^{-1} E_0 \left[\sup_{|u|\leq\delta r_n} \left| \int_0^u W_{n,x}(v) dv \right| \right] + \delta r_n \theta_0(x) E_0 \left[r_n |S_n(x) - S_0(x)| \right],$$

itself bounded by $f_n(r_n\delta) + \theta_0(x)\delta f_n(r_n\delta) + \theta_0(x)r_n\delta$ since $\left|\int_0^u W_{n,x}(v)\,dv\right| \leq |u| \sup_{|v|\leq|u|} |W_{n,x}(v)|$. This expression satisfies (B3) since $\delta f_n(r_n\delta) \leq f_n(r_n\delta)$ for each $\delta \leq 1$. Condition (B4) is satisfied since

$$E_0\left\{\sup_{|v|\leq\delta}\left|\int_0^{x+v}[S_n(u) - S_0(u)]\,du\right|\right\} = \mathrm{O}\left(E_0\left[\sup_{u\leq x+\delta}|S_n(u) - S_0(u)|\right]\right).$$

This is similarly true for (B5). □

Proof of Lemma 6. The result follows immediately upon noting that

$$\left\{\sup_{u\in\mathcal{U}}|V_n(u, f_n) - V_n(u, f_0)| > \epsilon\right\} \subseteq \left\{\sup_{\rho((u,f),(v,g))<\delta}|V_n(u, f) - V_n(v, g)| > \epsilon\right\} \cup \Big\{d_2(f_n, f_0) \geq \delta\Big\}.$$

□

Appendix C

PROOF OF RESULTS FROM CHAPTER 4

First-order expansions of the primitive estimator

We will say $a \lesssim b$ if there exists a $c < \infty$ such that $a \leq cb$. Before proving our main results, we perform a first-order expansion of $\Gamma_n(a)$. We define

$$\begin{aligned}
\phi_{n,a}(y,a,w) &:= I_{(-\infty,a]}(a)\left(\frac{y-\mu_n(a,w)}{g_n(a,w)} + \int \mu_n(a,\tilde{w})\,Q_n(d\tilde{w})\right) + \int_{-\infty}^{a} \mu_n(\tilde{a},w)\,F_n(d\tilde{a}) \\
&\quad - \iint_{-\infty}^{a} \mu_n(\tilde{a},\tilde{w})\,F_n(d\tilde{a})\,Q_n(d\tilde{w}) \\
\phi_{\mu,g,a}(y,a,w) &:= I_{(-\infty,a]}(a)\left(\frac{y-\mu(a,w)}{g(a,w)} + \int \mu(a,\tilde{w})\,Q_0(d\tilde{w})\right) + \int_{-\infty}^{a} \mu(\tilde{a},w)\,F_0(d\tilde{a}) \\
&\quad - \iint_{-\infty}^{a} \mu(\tilde{a},\tilde{w})\,F_0(d\tilde{a})\,Q_0(d\tilde{w})\ ,
\end{aligned}$$

so that $\Gamma_n(a) = \mathbb{P}_n\phi_{n,a}$. Letting $\phi_{\infty,a} := \phi_{\mu_\infty,g_\infty,a}$, we have

$$P_0\phi^*_{\infty,a} = \iint_{-\infty}^{a} (\mu_\infty(a,w) - \mu_0(a,w))\left(1 - \frac{g_0(a,w)}{g_\infty(a,w)}\right) F_0(da)\,Q_0(dw) + \Gamma(a) = \Gamma(a)$$

by (E3). Thus, with $\phi^*_{\infty,a} := \phi_{\infty,a} - \Gamma(a)$, we have the first-order expansion $\Gamma_n(a) - \Gamma(a) = \mathbb{P}_n\phi^*_{\infty,a} + R_{n,a}$, where

$$R_{n,a} := (\mathbb{P}_n - P_0)(\phi_{n,a} - \phi_{\infty,a}) + P_0\phi_{n,a} - \Gamma(a)\ .$$

We decompose $R_{n,a}$ in to three remainder terms, $\sum_{j=1}^{3} R_{n,a,j}$, for

$$R_{n,a,1} := \iint_{-\infty}^{a} (\mu_n(a,w) - \mu_0(u,w))\left(1 - \frac{g_0(a,w)}{g_n(a,w)}\right) dF_0(a)\,dQ_0(w)$$

$$R_{n,a,2} := (\mathbb{P}_n - P_0)(\phi_{n,a} - \phi_{\infty,a})$$
$$R_{n,a,3} := \iint_{-\infty}^{a} \mu_n(a,w)\,(F_n - F_0)(da)(Q_n - Q_0)(dw)\ . \tag{C.1}$$

Furthermore the last term can be written as

$$\frac{1}{2n^2}\sum_{i\neq j}\gamma_{\mu_n,a}(O_i,O_j) + \frac{1}{n}\int I_{(-\infty,a]}(a)\mu_n(a,w)\,(\mathbb{P}_n - P_0)(da,dw) + \frac{1}{n}E_{P_0}[I_{(-\infty,a]}(A)\mu_n(A,W)]\ ,$$

where

$$\begin{aligned}\gamma_{\mu,a}(o_i,o_j) := {} & I_{(-\infty,a]}(a_i)\mu(a_i,w_j) + I_{(-\infty,a]}(a_j)\mu(a_j,w_i) \\ & - \int \left[I_{(-\infty,a]}(a_i)\mu(a_i,w) + I_{(-\infty,a]}(a_j)\mu(a_j,w)\right]\,Q_0(dw) \\ & - \int_{-\infty}^{a}\left[\mu(a',w_i) + \mu(a',w_j)\right]\,F_0(da') + 2\int I_{(-\infty,a]}(a')\mu(a',w)\,F_0(da')\,Q_0(dw)\ .\end{aligned}$$

Before moving to the proofs, we state two lemmas that we will use. Lemma 8 below says that the entropy of a uniformly bounded class over a product space, when marginalized over one component of the product space with respect to a fixed probability measure, is bounded above by the entropy of the original class.

Lemma 8. *Let $\mathcal{F}$ be a uniformly bounded class of functions $f : \mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathbb{R}$ with $|f| \leq K < \infty$ for all $f \in \mathcal{F}$. Let R be a fixed probability measure on $\mathcal{Z}_2$, and define $\mathcal{F}^* = \{z_1 \mapsto \int f(z_1,z_2)\,R(dz_2) : f \in \mathcal{F}\}$. Then*

$$\sup_Q N(\varepsilon K, \mathcal{F}^*, L_2(Q)) \leq \sup_Q N(\varepsilon K/2, \mathcal{F}, L_2(Q)).$$

Proof. The statement follows immediately from Lemma 5.2 of van der Vaart and van der Laan (2006) by taking $r = s = t = 2$. □

The second lemma concerns so-called *degenerate U-processes*, and is a slight simplification of Theorem 6 of Nolan and Pollard (1987). A P_0-degenerate U-process for a class of functions

$\mathcal{F}$ is defined as a sum of the form $\{S_n(f) : f \in \mathcal{F}\}$, where

$$S_n(f) := \sum_{1 \leq i \neq j \leq n} f(O_i, O_j) \ ,$$

and where each $f \in \mathcal{F}$ is a function from $\mathcal{O} \times \mathcal{O} \to \mathbb{R}$ satisfying: (i) f is symmetric in its arguments, meaning that $f(o, o') = f(o', o)$ for all $o, o' \in \mathcal{O}$ and (ii) $\int f(o, o') P_0(do') = 0$ for all $o \in \mathcal{O}$. For such processes, we have the following result.

Lemma 9. *Suppose $\{S_n(f) : f \in \mathcal{F}\}$ be a P_0-degenerate U-process. If F is an envelope function for $\mathcal{F}$, then*

$$\frac{1}{[n(n-1)]^{1/2}} E_0 \left[\sup_{f \in \mathcal{F}} |S_n(f)| \right] \lesssim \int_0^1 \left[1 + \log \sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \, d\varepsilon \right] \|F\|_{P_0 \times P_0, 2} \ .$$

Proof. We let $\mathbb{T}_n f := \frac{1}{n(n-1)} \sum_{i \neq j} f(O_i, O_j)$. We also define $\theta_n := \frac{1}{4} \sup_{f \in \mathcal{F}} \|f\|_{\mathbb{T}_n, 2}$, $\tau_n := \|F\|_{\mathbb{T}_n, 2}$, and $J_n(s) := \int_0^s \log N(\varepsilon, \mathcal{F}, d_{\mathbb{T}_n, 2, F}) \, d\varepsilon$, where

$$d_{\mathbb{T}_n, 2, F}(f, g) := \left[\mathbb{T}_n (f - g)^2 / \mathbb{T}_n F^2 \right]^{1/2} = \|f - g\|_{\mathbb{T}_n, 2} / \|F\|_{\mathbb{T}_n, 2} \ .$$

Theorem 6 of Nolan and Pollard (1987) then says that

$$\frac{1}{[n(n-1)]^{1/2}} E_0 \left[\sup_{f \in \mathcal{F}} |S_n(f)| \right] \lesssim E_0 \left[\theta_n + \tau_n J_n(\theta_n / \tau_n) \right] \ .$$

Now we note that

$$J_n(s) = \int_0^s \log N(\varepsilon \|F\|_{\mathbb{T}_n, 2}, \mathcal{F}, L_2(\mathbb{T}_n)) \, d\varepsilon \leq \int_0^s \sup_Q \log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \, d\varepsilon \ ,$$

where the supremum is taken over all finite, discrete Q such that $QF > 0$. Next, since $\theta_n \leq \tau_n$, we have

$$E_0 \left[\theta_n + \tau_n J_n(\theta_n / \tau_n) \right] \leq E_0[\tau_n] \left[1 + \int_0^1 \sup_Q \log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \, d\varepsilon \right] \ .$$

Finally, by Jensen's inequality $E_0[\tau_n] \leq \|F\|_{P_0 \times P_0,2}$.

□

Proof of Theorem 11

We use Theorem 4 from Chapter 3 for the both the pointwise and uniform consistency statements. Since F_n is the empirical distribution function, $\sup_{a \in \mathcal{A}} |F_n(a) - F_0(a)| \xrightarrow{\mathrm{P}} 0$ by the Glivenko-Cantelli Theorem. Hence we only need to show that $\sup_{a \in \mathcal{A}} |\Gamma_n(a) - \Gamma(a)| \xrightarrow{\mathrm{P}} 0$.

We first establish that $\{\phi^*_{\infty,a} : a \in \mathcal{A}\}$ is a P_0-Donsker class. The class $\{I_{(-\infty,a]}(a) : a \in \mathcal{A}\}$ is a VC class and hence also P_0-Donsker. Since μ_∞ is a bounded, fixed function, $\{I_{(-\infty,a]}(a)\mu_\infty(a,w) : a \in \mathcal{A}\}$ is also P_0-Donsker, which implies that $\{\int_{-\infty}^a \mu_\infty(a,w)F_0(da) : a \in \mathcal{A}\}$ is P_0-Donsker by Lemma 8. Hence, by permanence properties of Donsker classes, $\{\phi^*_{\infty,a} : a \in \mathcal{A}\}$ is a P_0-Donsker class and $\sup_{a \in \mathcal{A}} |\mathbb{P}_n \phi^*_{\infty,a}| = \mathrm{O_P}(n^{-1/2})$.

For the first remainder term, we write

$$\begin{aligned}
\sup_{a \in \mathcal{A}} |R_{n,a,1}| \leq & \int_{\mathcal{O}_1} |\mu_n(a,w) - \mu_\infty(u,w)| \left|1 - \frac{g_0(a,w)}{g_n(a,w)}\right| (F_0 \times Q_0)(da,dw) \\
& + \int_{\mathcal{O}_2} |\mu_n(a,w) - \mu_0(u,w)| \left|1 - \frac{g_\infty(a,w)}{g_n(a,w)}\right| (F_0 \times Q_0)(da,dw) \\
& + \int_{\mathcal{O}_3} |\mu_n(a,w) - \mu_\infty(u,w)| \left|1 - \frac{g_\infty(a,w)}{g_n(a,w)}\right| (F_0 \times Q_0)(da,dw) \\
\leq & \left[(F_0 \times Q_0)(\mu_n - \mu_\infty)^2 (F_0 \times Q_0)(1 - g_0/g_n)^2\right]^{1/2} \\
& + \left[(F_0 \times Q_0)(\mu_n - \mu_0)^2 (F_0 \times Q_0)(1 - g_\infty/g_n)^2\right]^{1/2} \\
& + \left[(F_0 \times Q_0)(\mu_n - \mu_\infty)^2 (F_0 \times Q_0)(1 - g_\infty/g_n)^2\right]^{1/2} .
\end{aligned}$$

By assumption, $(F_0 \times Q_0)(\mu_n - \mu_\infty)^2 = \mathrm{o_P}(1)$, and since g_n is almost surely eventually bounded uniformly above and below away from zero, $(F_0 \times Q_0)(1 - g_\infty/g_n)^2 = \mathrm{o_P}(1)$ as well. Furthermore $(F_0 \times Q_0)(1 - g_0/g_n)^2 = \mathrm{O_P}(1)$ and $(F_0 \times Q_0)(\mu_n - \mu_0)^2 = \mathrm{O_P}(1)$ by since μ_n, g_n, μ_0, and g_0 are all bounded for n large enough. Hence $\sup_{a \in \mathcal{A}} |R_{n,a,1}| = \mathrm{o_P}(1)$.

Next we analyze $R_{n,a,2}$. We define

$$\phi'_{\mu,g,a} := I_{(-\infty,a]}(a)\left[\frac{y-\mu(a,w)}{g(a,w)} + \int \mu(a,\tilde{w})\, Q_0(d\tilde{w})\right] + \int_{-\infty}^{a} \mu(\tilde{a},w)\, F_0(d\tilde{a}) \ ,$$

and we note that $R_{n,a,2} = (\mathbb{P}_n - P_0)(\phi'_{\mu_n,g_n,a} - \phi'_{\mu_\infty,g_\infty,a})$. We then define the stochastic process $\{\mathbb{G}_n\phi'_{\mu,g,a} : \mu \in \mathcal{F}^{(\mu)}, g \in \mathcal{F}^{(g)}, a \in \mathcal{A}\}$. We will use Lemma 6 from Chapter 3 to establish that $\sup_{a\in\mathcal{A}} |\sqrt{n}R_{n,a,2}| = o_P(1)$. In their notation, we set $\mathcal{U} := \mathcal{A}$, equipped with the usual Euclidean norm, and $\mathcal{F} = \mathcal{F}^{(\mu)} \times \mathcal{F}^{(g)}$, equipped with the product $L_2(P_0)$ semi-metric $d((\mu,g),(\mu',g')) = [P_0(\mu-\mu')^2]^{1/2} + [P_0(g-g')^2]^{1/2}$. Application of this result requires showing that the process is uniformly asymptotically ρ-equicontinuous for ρ the product semi-metric. This would be implied if the class $\{\phi'_{\mu,g,a} : \mu \in \mathcal{F}^{\mu}, g \in \mathcal{F}^{(g)}, a \in \mathcal{A}\}$ were P_0-Donsker. Note that condition (E1) implies that $\mathcal{F}^{(\mu)}$ and $\mathcal{F}^{(g)}$ are P_0-Donsker classes by van der Vaart and Wellner (1996) Theorem 2.5.2. Since $\{I_{(-\infty,a]}(a) : a \in \mathcal{A}\}$ is a P_0-Donsker (established above), by Lemma 8 below the classes $\{\int I_{(-\infty,a]}(a)\mu(a,\tilde{w})\, Q_0(d\tilde{w}) : \mu \in \mathcal{F}^{(\mu)}, a \in \mathcal{A}\}$ and $\{\int_{-\infty}^{a} \mu(\tilde{a},w)\, F_0(d\tilde{a}) : \mu \in \mathcal{F}^{(\mu)}, a \in \mathcal{A}\}$ are also P_0-Donsker. Since $\mathcal{F}^{(g)}$ is bounded below, the class $\{I_{(-\infty,a]}(a)[y-\mu(a,w)]/g(a,w) : \mu \in \mathcal{F}^{(\mu)}, g \in \mathcal{F}^{(g)}, a \in \mathcal{A}\}$ is also P_0-Donsker. Adding together these terms yields that the original class is P_0-Donsker. The second requirement of Lemma 6 is satisfied by assumption.

Finally, we turn to $R_{n,a,3}$, which has three sub-components. The second sub-component is an ordinary empirical process involving function classes discussed in the preceding paragraph. Using these results yields the second component to be $O_P(n^{-3/2})$. The third sub-component is a bias term which, in view of the uniform boundedness of μ_n, is $O_P(n^{-1})$. The first sub-component is a P_0-degenerate U-process as defined above, to which we will apply Lemma 9. The function $\gamma_{\mu_n,a}(o_i,o_j)$ is contained in the class $\left\{(a_1,w_1,a_2,w_2) \mapsto \gamma_{\mu,a}(a_1,w_1,a_2,w_2) : a \in \mathcal{A}, \mu \in \mathcal{F}^{(\mu)}\right\}$. As we discuss in more detail below, by Lemma 8, Lemma 5.1 of van der Vaart and van der Laan (2006), and condition (E1), this class has uniform entropy bounded up to a constant by $-\log\varepsilon + \varepsilon^{-V/2}$ relative to a constant envelope. Therefore, Lemma 9 implies

that

$$E_0\left[\sup_{\mu\in\mathcal{F}^{(\mu)},a\in\mathcal{A}}\left|\sum_{i\neq j}\gamma_{\mu,a}(O_i,O_j)\right|\right]\lesssim[n(n-1)]^{1/2}\ .$$

Therefore, the first sub-component of $R_{n,a,3}$ is $\mathrm{O_P}(n^{-1})$. Thus, we have $\sup_{a\in\mathcal{A}}|R_{n,a,3}| = \mathrm{O_P}(n^{-1})$.

We have now controlled all three remainder terms, and hence have successfully showed that under (E1)–(E3), $\sup_{a\in\mathcal{A}}|\Gamma_n(a)-\Gamma(a)|\xrightarrow{\mathrm{P}}0$.

Proof of Theorem 12

We will use Theorem 7 of Chapter 3 to establish our Theorem 12. In what follows we check their conditions (C1)–(C5) and (B4)–(B5).

Conditions (C1) and (C2). Define $I_{a,u}(a) = I_{(-\infty,a+u]}(a) - I_{(-\infty,a]}(a)$ and $g_{a,u}(o) = [\phi^*_{\infty,a+u}(o) - \phi^*_{\infty,a}(o)] - \theta_0(a)I_{a,u}(a)$. Since F_0 is by assumption strictly increasing at a, we then have

$$\begin{aligned}g_{a,u}(o) = I_{a,u}(a)\left[\frac{y-\mu_\infty(a,w)}{g_\infty(a,w)}+\theta_\infty(a)-\theta_0(a)\right]+\int I_{a,u}(v)\mu_\infty(v,w)\,F_0(dv)\\ -[\Gamma_\infty(a+u)-\Gamma_\infty(a)]-[\Gamma(a+u)-\Gamma(a)]+[F_0(a+u)-F_0(a)]\ ,\end{aligned}$$

where we are defining $\theta_\infty(a) = E_{Q_0}[\mu_\infty(a,W)]$ and $\Gamma_\infty = \int_{-\infty}^{a}\theta_\infty(a)\,F_0(da)$.

The class $\mathcal{I}_R = \{I_{a,u} : |u|\le R\}$ is a VC class of functions by a slight extension of example 2.6.1 of van der Vaart and Wellner (1996). Its envelope function is $J_{a,u}(a) = I_{[0,R]}(|a-a|)$, and hence $\sup_Q \log N(\varepsilon\|J_R\|_{Q,2},\mathcal{I}_R,L_2(Q))\lesssim -\log(\varepsilon)$ by Theorem 2.6.7 of van der Vaart and Wellner (1996). The class $\{\int I_{a,u}(v)\mu(v,w)\,F_0(dv) : |u|\le R\}$ thus satisfies the same inequality by Lemma 8. The classes $\{\Gamma_\infty(a+u)-\Gamma_\infty(a) : |u|\le R\}$, $\{\Gamma(a+u)-\Gamma(a) : |u|\le R\}$ and $\{F_0(a+u)-F_0(a) : |u|\le R\}$ are sets of constants not depending on the data, bounded up to a constant by R for R small enough since Γ and F_0 are continuously differentiable in a neighborhood of a, and hence also have uniform entropy bounded up to a constant by $-\log(\varepsilon)$.

Finally, the class $\mathcal{G}_R$ is a linear combination of the above classes, and hence by van der Vaart and van der Laan (2006) Lemma 5.1, $\mathcal{G}_R$ satisfies $\sup_Q \log N(\varepsilon \|G_R\|_{Q,2}, \mathcal{G}_R, L_2(Q)) \lesssim -\log(\varepsilon)$ as well. This satisfies condition (C1).

Since Γ, Γ_∞, and F_0 are continuously differentiable in a neighborhood of a, an envelope function for the class $\mathcal{G}_R = \{g_{a,u} : |u| \leq R\}$ is

$$G_R(o) = J_{a,R}(a) \left| \frac{y - \mu_\infty(a,w)}{g_\infty(a,w)} + \theta_\infty(a) - \theta_0(a) \right| + \int J_{a,R}(v) |\mu_\infty(v,w)| f_0(v)\, dv + K_1 R$$

for some K_1. Using the triangle inequality on $\|G_R\|_{P_0,2}$, for the first term we have

$$E_{P_0}\left\{ J_{a,R}(A) \frac{[Y - \mu_\infty(A,W)]^2}{g_\infty(A,W)^2} \right\} = E_{P_0}\left\{ J_{a,R}(A) \frac{\sigma_0^2(A,W) + [\mu_\infty(A,W) - \mu_0(A,W)]^2}{g_\infty(A,W)^2} \right\} \leq K_2 R$$

for some K_2 by the boundedness of σ_0^2, $1/g_\infty$, μ_∞, μ_0 and the conditional density π_0 in a neighborhood of a uniformly over Q_0-a.e. w. Similarly bounds hold for the other terms, yielding $P_0 G_R^2 \lesssim R$ for all R small enough as required.

For the second requirement of (C2), we note that for all R small enough and some constants k_1, k_2, k_3, $G_R \leq J_{a,R}(|y|/k_1 + k_2) + k_3 R$. By assumption and properties of probability densities, for all R small enough and for all $\varepsilon > 0$ there is a c such that $P_0[J_{a,R}(A)|Y| > c] < \varepsilon$. This implies that for any $\eta > 0$, $P_0 G_R^2 I_{(\eta/R,\infty)}(G_R) < \varepsilon R$ for all R small enough.

Condition (C3). Next we need to study the covariance $\Sigma(s,t) = P_0[\phi^*_{\infty,s} - \theta_0(a)\gamma^*_s][\phi^*_{\infty,t} - \theta_0(a)\gamma^*_t]$ for s, t near a, and where $\gamma^*_s := I_{(-\infty,s]}(a) - F_0(s)$, where we may ignore any terms in the covariance which are continuously differentiable in a neighborhood of (a,a). We thus have

$$\phi^*_{\infty,s} - \theta_0(a)\gamma^*_s = [\phi_{\infty,s} - \Gamma_\infty(s) - \Gamma(s)] - \theta_0(a)[I_{(-\infty,s]}(a) - F_0(s)]\ .$$

Expanding $\Sigma(s,t)$, since Γ_∞, Γ and F_0 are continuously differentiable in a neighborhood of a, it is straightforward to see that we may focus on

$$
\begin{aligned}
&E_{P_0}\left[\phi_{\infty,s} - \theta_0(a) I_{(-\infty,s]}(a)\right]\left[\phi_{\infty,t} - \theta_0(a) I_{(-\infty,t]}(a)\right] \\
&\quad = E_{P_0}\left\{ I_{(-\infty,s\wedge t]}(A)\left[\frac{Y-\mu_\infty(A,W)}{g_\infty(A,W)} + \theta_\infty(A) - \theta_0(a)\right]^2 \right\} \\
&\qquad + E_{P_0}\left\{ I_{(-\infty,s]}(A)\left[\frac{\mu_0(A,W)-\mu_\infty(A,W)}{g_\infty(A,W)} + \theta_\infty(A) - \theta_0(a)\right]\right\}\int_{-\infty}^{t} \mu_\infty(a,W)\, F_0(da) \\
&\qquad + E_{P_0}\left\{ I_{(-\infty,t]}(A)\left[\frac{\mu_0(A,W)-\mu_\infty(A,W)}{g_\infty(A,W)} + \theta_\infty(A) - \theta_0(a)\right]\right\}\int_{-\infty}^{s} \mu_\infty(a,W)\, F_0(da) \\
&\qquad + E_{Q_0}\left\{ \int_{-\infty}^{s} \mu_\infty(a,W)\, F_0(da) \int_{-\infty}^{t} \mu_\infty(a,W)\, F_0(da)\right\}.
\end{aligned}
$$

The bottom three lines are continuously differentiable for (s,t) in a neighborhood of (a,a) since $\mu_\infty, \mu_0, g_\infty, g_0$ are all continuous in a neighborhood of a, uniformly over Q_0-a.e. w. Hence they do not contribute to the scale parameter of the limit.

By Fubini's theorem, the first line can be rewritten as:

$$
\int_{-\infty}^{s\wedge t}\int E_{P_0}\left\{\left[\frac{Y-\mu_\infty(A,W)}{g_\infty(A,W)} + \theta_\infty(a) - \theta_0(a)\right]^2 \middle| A=a, W=w\right\} g_0(a,w)\, Q_0(dw)\, F_0(da)\ .
$$

By (E5), this satisfies condition (B3), and hence the limit distribution is $[4\theta_0'(a)\tilde{\kappa}_0(a)/f_0(a)^2]^{1/3}\mathbb{W}$, where

$$
\tilde{\kappa}_0(a) = E_{Q_0}\left\{ E_{P_0}\left[\left(\frac{Y-\mu_\infty(A,W)}{g_\infty(A,W)} + \theta_\infty(A) - \theta_0(A)\right)^2 \middle| A=a, W=w\right] g_0(a,W)\right\} f_0(a).
$$

Therefore, $[4\theta_0'(a)\tilde{\kappa}_0(a)/f_0(a)^2]^{1/3} = [4\theta_0'(a)\kappa_0(a)/f_0(a)]^{1/3}$, for $\kappa_0(a)$ as defined in the statement of Theorem 12.

Conditions (C4) and (C5). Next, we need to show that the remainder is negligible. Denote

$$K_{n,j}(\delta) = n^{2/3} \sup_{|u| \leq \delta/n^{1/3}} |R_{n,a+u,j} - R_{n,a,j}| .$$

For each j we need to show that $K_{n,j}(\delta) \xrightarrow{\mathrm{P}} 0$ for all δ small enough and $E[K_{n,j}(\delta)]/\delta^{\beta}$ is decreasing in δ for some $1 < \beta < 2$ and for all δ small enough and n large enough.

By Fubini's theorem and simple supremum bounds, for all n large enough and δ small enough, $K_{n,1}(\delta)$ is bounded up to a constant by

$$\begin{aligned}
&\delta n^{1/3} \sup_{|a-a| \leq \varepsilon^*} E_{Q_0} \left[|\mu_n(a,W) - \mu_0(a,W)| \, |g_n(a,W) - g_0(a,W)| \right] \\
&\quad = \delta n^{1/3} \sup_{|a-a| \leq \varepsilon^*} E_{Q_0} \left[I_{\mathcal{S}_1}(a,W) \, |\mu_n(a,W) - \mu_\infty(a,W)| \, |g_n(a,W) - g_0(a,W)| \right] \\
&\qquad + \delta n^{1/3} \sup_{|a-a| \leq \varepsilon^*} E_{Q_0} \left[I_{\mathcal{S}_2}(a,W) \, |\mu_n(a,W) - \mu_0(a,W)| \, |g_n(a,W) - g_\infty(a,W)| \right] \\
&\qquad + \delta n^{1/3} \sup_{|a-a| \leq \varepsilon^*} E_{Q_0} \left[I_{\mathcal{S}_3}(a,W) \, |\mu_n(a,W) - \mu_\infty(a,W)| \, |g_n(a,W) - g_\infty(a,W)| \right] \\
&\quad \lesssim \delta n^{1/3} \left\{ d(\mu_n, \mu_\infty; a, \varepsilon^*, \mathcal{S}_1) + d(g_n, g_\infty; a, \varepsilon^*, \mathcal{S}_2) + d(\mu_n, \mu_\infty; a, \varepsilon^*, \mathcal{S}_3) d(g_n, g_\infty; a, \varepsilon^*, \mathcal{S}_3) \right\} .
\end{aligned}$$

Hence, under conditions (E4a), (E4c), and (E4d), $K_{n,1}(\delta) \xrightarrow{\mathrm{P}} 0$ for each δ. Furthermore, $E\left[K_{n,1}(\delta)\right]/\delta^{\beta}$ is decreasing in δ for any $1 < \beta < 2$ by the assumed uniform boundedness of μ_n, g_n, μ_∞, g_∞, μ_0, and g_0.

We will use Theorem 9 of Chapter 3 to establish negligibility of the empirical process term $K_{n,2}(\delta)$, which requires checking conditions (D1)–(D4). Let $\omega = (\mu, g)$, which is contained in the product class $\mathcal{P} = \mathcal{F}^{(\mu)} \times \mathcal{F}^{(g)}$ almost surely for all n large enough. We equip $\mathcal{P}$ with the semi-metric

$$d^*(\omega_1, \omega_2) := d(\mu_1, \mu_2; a, \varepsilon^*, \mathcal{A} \times \mathcal{W}) + d(g_1, g_2; a, \varepsilon^*, \mathcal{A} \times \mathcal{W}) .$$

Next, we define

$$\mathcal{G}_R = \left\{ s_u(\mu, g) : |u| \leq R, \mu \in \mathcal{F}^{(\mu)}, g \in \mathcal{F}^{(g)} \right\},$$

where

$$s_u(\mu, g)(y, a, w) = I_{a,u}(a) \left[\frac{y - \mu(a, w)}{g(a, w)} + E_{Q_0}[\mu(a, W)] \right] + E_{F_0}[I_{a,u}(A)\mu(A, w)] \ .$$

We let G_R be the envelope function for $\mathcal{G}_R$ obtained by combining the assumed uniform bounds on $\mathcal{F}^{(\mu)}$ and $\mathcal{F}^{(g)}$ and the natural envelope for $I_{a,u}(a)$. Specifically, we have $G_R(y, a, w) = I_{[0,R]}(|a - a|) [a|y| + b]$ for some $a, b < \infty$. For all R small enough and some $V < 1$, $\mathcal{G}_R$ is a Lipschitz transformation of the classes:

- $\mathcal{F}^{(\mu)}$, which possesses uniform entropy bounded up to a constant by ε^{-V};
- $\mathcal{F}^{(g)}$, which similarly possesses uniform entropy bounded up to a constant by ε^{-V};
- $\{a \mapsto E_{Q_0}[\mu(a, W)] : \mu \in \mathcal{F}^{(\mu)}\}$, which possesses uniform entropy bounded up to a constant by ε^{-V} by Lemma 8;
- $\{I_{a,u} : |u| \leq R\}$, which possesses polynomial covering number;
- $\{w \mapsto E_{F_0}[I_{a,u}(A)\mu(A, w)] : \mu \in \mathcal{F}^{(\mu)}, |u| \leq R\}$, which has uniform entropy bounded up to a constant by $-\log \varepsilon + \varepsilon^{-V}$ by van der Vaart and van der Laan (2006) Lemmas 5.1 and our Lemma 8);
- $\{E_{F_0}[I_{a,u}(A)\mu_\infty(A, w)] : |u| \leq R\}$, which possesses polynomial covering number;
- the singleton class $\{y\}$, with covering number equal to one.

Hence by Lemma 5.1 of van der Vaart and van der Laan (2006), the L_2 covering number of $\mathcal{G}_R$ relative to G_R is bounded up to a constant by $-\log \varepsilon + \varepsilon^{-V} + \varepsilon^{-V/2}$, and since $V < 2$, $\int_0^1 [\log \sup_Q N(\varepsilon \|G_R\|_{Q,2}, \mathcal{G}_R, L_2(Q))]^{1/2} \, d\varepsilon$ is with probability tending to one uniformly bounded above for all R small enough. This establishes (D1).

Existence of the conditional variance of Y given A, W in a neighborhood of a and the positivity of f_0 in a neighborhood of a yields $P_0 G_R^2 \leq cR$ and that for any ε there exists ε' such that $P_0[G_R^2 I_{(\varepsilon'/R,\infty)}(G_R)] \leq \varepsilon R$ for all R small enough. Hence condition (D2) is satisfied.

Turning to (D3), we note that

$$\begin{aligned}\left\{P_0[s_u(\mu,g) - s_v(\mu,g)]^2\right\}^{1/2} &\leq \left\{\int \left[\int_{a+v}^{a+u} \mu(a,w)\, F_0(da)\right]^2 Q_0(dw)\right\}^{1/2} \\ &+ \left\{\int_{a+v}^{a+u} \iint \left[\frac{y-\mu(a,w)}{g(a,w)} + E_{Q_0}[\mu(a,W)]\right]^2 g_0(a,w)\, P_0(dy \mid a,w)\, Q_0(dw)\, F_0(da)\right\}^{1/2},\end{aligned}$$

and by the finite conditional second moment of Y given A, W, the boundedness of g_0, the uniform boundedness of μ, g, and the positivity of f_0 near a, we have $P_0[s_u(\mu,g) - s_v(\mu,g)]^2 \lesssim |u-v|$ for all u, v in a neighborhood of 0. Similarly,

$$\begin{aligned}\left\{P_0[s_u(\mu_1,g_1) - s_u(\mu_2,g_2)]^2\right\}^{1/2} &\leq \left\{\int \left[\int_a^{a+v} \{\mu_1(a,w) - \mu_2(a,w)\}\, F_0(da)\right]^2 dQ_0(w)\right\}^{1/2} \\ &+ \left\{\int_a^{a+v} \left[\int \{\mu_1(a,w) - \mu_2(a,w)\}\, Q_0(dw)\right]^2 F_0(da)\right\}^{1/2} \\ &+ \left\{\int_a^{a+v} \iint \left[\frac{y-\mu_2(a,w)}{g_1(a,w)g_2(a,w)}\{g_2(a,w) - g_1(a,w)\}\right]^2 \right. \\ &\qquad \left. \times\, P_0(dy \mid a,w) g_0(a,w)\, Q_0(dw)\, F_0(da)\right\}^{1/2} \\ &+ \left\{\int_a^{a+v} \iint \left[\{g_1(a,w)g_2(a,w\}^{-1}\{\mu_1(a,w) - \mu_2(a,w)\}\right]^2 Q_0(dw)\, F_0(da)\right\}^{1/2}.\end{aligned}$$

We find that for v small enough this is bounded up to a constant by

$$\sqrt{|v|}\left[\left\{\sup_{|a-a|\leq\varepsilon^*} \left[E_{Q_0}\{\mu_1(a,W) - \mu_2(a,W)\}^2\right]^{1/2} + \sup_{|a-a|\leq\varepsilon^*} \left[E_{Q_0}\{g_1(a,W) - g_2(a,W)\}^2\right]^{1/2}\right\}.$$

as required. Finally, (D4) is satisfied by assumption.

For $K_{n,3}(\delta)$, we first note that (C4) is already satisfied by the work we did in the proof of Theorem 11, since

$$n^{2/3} \sup_{|u|\leq\delta/n^{1/3}} |R_{n,a+u,3} - R_{n,a,3}| \leq 2n^{2/3} \sup_{a\in\mathcal{A}} |R_{n,a,3}| = \mathrm{O_P}(n^{-1/6}).$$

We check (C5) for each of the three sub-components of $K_{n,3}(\delta)$ defined by the three sub-components of $R_{n,a,3}$. Due to the assumed boundedness of $|\mu_n|$, the contribution of the third component is bounded for all δ small enough up to a constant (not depending on δ or n) by $n^{-1/3}P_0(|A-a| \leq \delta/n^{1/3}) \lesssim n^{-2/3}\delta$, which satisfies (C5). For the second component, by Lemma 6, $E\left[\sup_{|u|\leq\delta/n^{1/3}} |\mathbb{G}_n I_{a,u}\mu_n|\right] \lesssim \sqrt{\delta}$, and hence the expectation of the second component is bounded up to a constant by $\sqrt{\delta}/n$ for all δ small enough and n large enough, which is also sufficient for (C5).

The first component requires controlling $\sum_{i\neq j} \gamma^*_{\mu_n,a,u}(O_i, O_j)$, where we define

$$\begin{aligned}\gamma^*_{\mu,a,u}(o_i, o_j) = {} & I_{a,u}(a_i)\mu(a_i, w_j) + I_{a,u}(a_j)\mu(a_j, w_i) \\ & - \int \left[I_{a,u}(a_i)\mu(a_i, w) + I_{a,u}(a_j)\mu(a_j, w)\right] Q_0(dw) \\ & - \int I_{a,u}(a)\left[\mu(a, w_i) + \mu(a, w_j)\right] F_0(da) + 2\iint I_{a,u}(a)\mu(a, w)F_0(da)Q_0(dw)\ .\end{aligned}$$

The function $\gamma^*_{\mu_n,a,u}$ falls in the class

$$\mathcal{H}_\delta = \left\{\gamma^*_{\mu,a,u} : |u| \leq \delta, \mu \in \mathcal{F}^{(\mu)}\right\}\ .$$

Thus, $\{\sum_{i\neq j} \gamma^*(O_i, O_j) : \gamma^* \in \mathcal{H}_\delta\}$ is a P_0-degenerate U-process. By a similar logic to that used above, the class H_δ possesses uniform entropy $\log\sup_Q N(\varepsilon\|H_\delta\|_{Q,2}, \mathcal{H}_\delta, L_2(Q))$ bounded up to a constant by $-\log\varepsilon + \varepsilon^{-V/2}$ relative to the envelope

$$H_\delta(a_1, w_1, a_2, w_2) = 2K_\mu I_{[0,\delta]}(|a_1 - a|) + 2K_\mu I_{[0,\delta]}(|a_2 - a|) + 4K_\mu P_0(|A-a| \leq \delta).$$

Since $-V/2 > -1$ and $\|H_\delta\|_{P_0 \times P_0, 2} \lesssim \sqrt{\delta}$, Lemma 9 yields that

$$n^{2/3} E_0 \left[\sup_{\gamma^* \in \mathcal{H}_\delta} \left| \frac{1}{n^2} \sum_{i \neq j} \gamma^*(O_i, O_j) \right| \right] \lesssim n^{-1/3} \sqrt{\delta}$$

for all δ small enough. Hence (C5) is satisfied for this U-process term.

Conditions (B4) and (B5). Condition (B4) is trivially satisfied since our transformation is the empirical distribution function. Condition (B5) was established in the proof of Theorem 11 under our conditions (E1)–(E3).

We have now checked all the conditions of Theorem 7 and verified that we have the stated limit distribution in the course of checking condition (C3). This concludes the proof.