Thesis Defense

Topics in Nonparametric Causal Inference

Matteo Bonvini

May 1, 2023

Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Edward H. Kennedy (Chair)
Sivaraman Balakrishnan
Zach Branson
Marco Carone (University of Washington)
Larry Wasserman

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Keywords: Continuous Treatments, Heterogenous Effects, Influence Function, Observational Studies, Semiparametric Models, Sensitivity Analysis

Abstract

We study several problems related to the identification and the efficient estimation of parameters arising in causal inference. In the first part of this thesis, we consider the problem of conducting sensitivity analysis to the no-unmeasured-confounding assumption in observational studies. Roughly speaking, confounders are variables that affect both the treatment receipt and the outcome. To estimate causal effects, all such variables must be measured and properly taken into account in the statistical analysis. This is an untestable assumption in the problems considered here because the treatment is not randomly assigned by the experimenter. Therefore, in these settings, gauging the impact of departures from this assumption on the causal effects' estimates is of great practical relevance. In one project, we develop a novel framework that bounds the average treatment effect (ATE) as a function of the proportion of units for which the treatment-outcome association is confounded. In other work, we propose and analyze a suite of models for obtaining bounds on certain causal effects when a marginal structural model is assumed.

In the second part of this thesis, we study the efficient estimation of two popular causal parameters: the dose-response function (DRF) and the level sets of the conditional ATE (CATE) curve. The DRF measures the expected outcome if everyone in the population takes a given treatment level. When the treatment is continuous, this parameter is a curve, viewed as a function of the infinitely many treatment values. We study several procedures to estimate the DRF and derive an estimator that, under certain conditions and to the best of our knowledge, achieves the lowest mean-square-error currently known in the literature. In a second paper, we derive the minimax optimal estimator of CATE level sets and provide upper bounds on the risk of other simpler estimation procedures. CATE level sets are a useful quantity to compute in many applications because they identify units with large treatment effects, which is the crucial information needed to optimally allocate the treatment.

Finally, in the third part of this thesis, we study the effects of reduced mobility on the number of Covid-19 deaths. We tackle this problem by specifying a marginal structural model motivated by an epidemic model. Our analysis finds that, for many US States and at the beginning of the pandemic, a decrease in mobility leads to significantly fewer deaths.

Contents

Acknowledgements

First, I would like to thank my advisor, Prof. Edward Kennedy, for his constant patience, invaluable help and generous support during my PhD studies. I count my years in Pittsburgh among the happiest of my life and this is certainly due to having him as my advisor. Learning about statistics and generally how to approach life from him has had an enourmous impact on me, which I will always be thankful for. I am also extremely grateful to Prof. Luke Keele for his guidance and generous support through the years. I would also like to express my sincere gratitude to my mentors and co-authors Profs. Larry Wasserman, Valérie Ventura, and Zach Branson. I feel very fortunate to have had the chance to learn and do research under their guidance. I am also very thankful for having had the possibility to meet and learn from Profs. Sivaraman Balakrishnan and Marco Carone, both directly and indirectly by reading their inspiring papers and lecture notes– thank you also for kindly accepting to be part of my thesis committee. Heartfelt thanks also to all the other faculty members, the staff and the amazing students for contributing to the vibrant and extraordinarily welcoming culture of our department. Finally, I would like to thank my family and my dear friends Alberto, Carlo Alberto, Chiara, Edo, Giacomo and Pietro for their support and the many happy moments we have lived together.

Chapter 1

Introduction

1.1 Motivation

Understanding the effect of a variable on an outcome is ultimately how science progresses and new knowledge is accumulated. Experiments are generally considered the gold-standard for this task because the randomization of the treatment ensures that, on average, any difference in outcomes between the treated and the untreated groups is due to the treatment status alone and not some other factors. However, experiments have also certain limitations, such as they may not be representative of the general population or very costly to conduct. Importantly, they might not be feasible for ethical reasons; for example, humans cannot be randomly assigned to smoking or to not going to college.

Building upon seminal work conducted at the beginning of the 20th century [Fisher, 1936, Neyman, 1923], from the 1970s researchers (e.g. Cochran and Rubin [1973], Cornfield et al. [1959], Robins [1986], Rubin [1974]) started to lay out sufficient and necessary conditions so that causal effects can be estimated even in observational studies, that is in any setting that is not a perfectly executed experiment. The standard assumption invoked to interpret observed associations as causal effects is the *no-unmeasured-confounding assumption*. It states that all confounders, roughly variables affecting both the outcome and the treatment receipt, have been correctly measured and included in the statistical analysis. In generality, it is impossible to test whether the measured covariates are sufficient to deconfound the treatment-outcome association, and thus justifying the veridicity of this assumption represents one of the main challenges in observational studies.

Another major challenge often encountered in observational studies is that the confounders must be correctly included in the statistical model for the data generating mechanism, typically in the form of covariates in a regression. With a large number of confounders, the precision with which the causal effect can be estimated quickly deteriorates, a phenomenon known as the *curse of dimensionality* in the nonparametric statistics literature. This is not the case in experiments because the probability of receiving treatment is known for each unit, so that

inverse-probability-weighted estimators are root-n consistent essentially under no conditions.

The goal of this thesis is to make progress on both of these challenges. In the first part, we propose several methods to gauge the impact of potential residual unmeasured confounding on the causal effects estimates. In the second part, we study estimation procedures designed to make the most efficient use of the sample and thus mitigate the issue of the curse of dimensionality. We conclude the thesis with an application of our methods to investigate the magnitude of the effect of reduced social mobility on Covid-19 deaths.

1.2 Overview of contributions

1.2.1 Chapters 2 and 3

In the first part of the thesis, we propose sensitivity models to gauge the impact of potential unmeasured confounding on the causal effects estimates. The holy grail of a sensitivity analysis in this context is to operationalize residual unmeasured confounding in a way that is both interpretable and general enough to capture plausible confounding mechanisms. Interpretability is key in sensitivity analyses because the researcher is required to judge whether the minimum amount of residual confounding needed to reduce the observed effect to the null value is too large to be plausible in the specific context of their application. If this is the case, the study's conclusions are deemed "robust."

In Chapter 2, we consider studies where the treatment is binary and does not vary with time and propose bounding the averaget treatment effect (ATE) as a function of the proportion of units for whom the treatment-outcome association is confounded. This proportion is not identified and is varied by the researcher as a sensitivity parameter: we argue that this model strikes a good balance between interpretability and generality. When this proportion equals zero, the bounds collapse to the point-identified ATE under no-unmeasured-confounding. When it equals one, the bounds are the so called "natural bounds" that are valid under no assumptions if the outcome is bounded and that are guaranteed to include the null value zero [Manski, 1990, Robins, 1989]. We also propose reporting a point and interval-estimate of the minimum proportion of "confounded units" such that the bounds include zero, which we view as a one-number summary of the study's robustness.

In Chapter 3, motivated by the application on Covid-19 reported in Chapater 6, we propose and analyze several methods to conduct sensitivity analysis when the treatment is multi-valued / continuous and potentially time-varying. We do so under the assumption of a marginal structural model (MSM) [Robins, 2000]. An MSM $g(a;\beta)$ is a parametric model that maps a given treatment sequence to the the expected outcome if everyone in the population takes that sequence. For example, if the treatment A can vary over T time-points, one can imagine counterfactual worlds where every unit in the population takes the same treatment sequence $\overline{a}_T = a_1, \ldots, a_T$ leading to the population expected outcome $\mathbb{E}(Y^{\overline{a}_T})$. An MSM specifies a model for the map $\overline{a}_T \mapsto \mathbb{E}(Y^{\overline{a}_T})$, e.g. $g(\overline{a}_T;\beta) = \beta_0 + \beta_1 \sum_{t=1}^{T} a_t$. Under the no-unmeasured-confounding assumption, the parameters of an MSM can be identified and estimated as the solution to a particular moment equation. In this work, we propose ways to bound the

parameter β or the model $g(\overline{a}_T; \beta)$ itself under several sensitivity models governing how residual unmeasured confounding acts on the observed distribution.

1.2.2 Chapters 4 and 5

In the second part of this thesis, we investigate the efficient estimation of popular causal parameters. While the estimands considered are motivated by causal inference applications, these two chapters are purely about statistical methodology. The following is a summary of the estimands considered together with the main challenges that we aim to address.

Chapter 4:

- *Estimand*: the set
$$\{x \in \mathbb{R}^d : \mu_1(x) - \mu_0(x) > \theta\},$$
for some user-specified cutoff θ and $\mu_a(x) = \mathbb{E}(Y \mid A = a, X = x)$, for $A \in \{0, 1\}$ and $X \in \mathbb{R}^d$. Under no-unmeasure-confounding, $\mu_1(x) - \mu_0(x)$ measures the conditional average treatment effect (CATE) for all units with covariates' value $X = x$ and the estimand is the **CATE (upper) level sets**. An important case is the upper level set at $\theta = 0$, which effectively identifies the portion of the covariates' space where the treatment effect is positive. If the goal is to maximize the mean outcome in the population, then the treatment should be allocated only to the units with covariates' values falling into this region.
- *Main challenges*: the parameter is set-valued and it depends on the complexity of $\mathbb{P}(A = 1 \mid X = x)$, $\mu_a(x)$ and the difference $\tau(x) = \mu_1(x) - \mu_0(x)$, which can be potentially of smaller complexity than the individual regression functions. The main challenge we tackle is to bound the risk of an estimator that simply thresholds an estimator of $\tau(x)$ and establish 1) that thresholding the minimax optimal estimator of $\tau(x)$ results in the minimax optimal estimator for $\tau(x)$'s level sets and 2) sufficient conditions such that $\tau(x)$'s level sets can be estimated with better accuracy than $\tau(x)$ itself. We also connect this statistical problem to the areas of classification, nonparametric regression and functional estimation.

Chapter 5:

- *Estimand*: the curve $a \mapsto \int \mathbb{E}(Y \mid A = a, X = x) d\mathbb{P}(x)$, where $A \in \mathbb{R}$ and $X \in \mathbb{R}^d$. Under no-unmeasured-confounding, this parameter representes the **dose-response function** (DRF), i.e., the expected outcome in the population if every units takes treatment value $A = a$.
- *Main challenges*: the DRF is a one-dimensional object that is the result of marginally integrating out all but one covariates of a $d + 1$-dim regression. The main challenge is to construct a flexible, nonparametric estimator whose risk is close to that of a one-dim regression rather than a $d + 1$-dim one.

1.2.3 Chapter 6

In the context of the Covid-19 pandemic, this chapter investigates the relationship between (anti)-mobility, as measured by the fraction of mobile devices that do not leave the immediate area of their home every week, and the number of Covid-19 deaths at the state-level. For each state, we thus observe a sequence $(A_1, Y_1), \ldots, (A_T, Y_T)$, where A_t and Y_t denote (anti)-mobility and Covid-19 deaths in week t, respectively. We consider data from Feb 15, 2020 to December 19, 2020.

During the beginning of the Covid-19 pandemic, many authors have studied the effect of mobility and interventions, e.g. lockdowns and school closures, on the number of Covid-19 cases and deaths. Many of the models proposed in the literature are generative in the sense that they try to model the infection or death processes as accurately as possible, typically using mechanistic models relating susceptible, infected and recovered (SIR) people via differential equations. These models are rooted in rigorous epidemiology theory, but they can make statistical inference intractable. To overcome this challenge, we start by considering a simple SIR model relating mobility, infections and deaths each week. From this model, by the g-formula [Robins, 1986] and under the assumption that there are no confounding variables except for previous deaths, we get an expression for the function mapping each treatment sequence (anti-mobility) to the expected number of deaths if everyone in the population follows that sequence. We then abandon the original SIR working model and simply interpret that map semiparametrically as a marginal structural model (MSM): we consider all distributions such that, if plugged into the g-formula, yield that particular MSM.

The parameters of an MSM can be identified as the solution to a moment condition. As such, inference can be carried out by standard Z-estimation theory. This is one of the main advantages of our approach. In the specific application considered, we find that, for many states, reduced mobility appears to decrease the number of deaths approximately four weeks later. Data availability in this study is limited, which has led us to make several simplifying assumptions, including the use of parsimonious semi-parametric models as well as the inclusion of only previous weeks' Covid-19 deaths as possible confounders. However, we also carry out several sensitivity analyses and find that the results are quite robust to the assumptions invoked.

Chapter 2

Sensitivity analysis via the proportion of unmeasured confounding

This chapter is taken from my work supervised by Edward H. Kennedy, which was published in the Journal of the American Statistical Association [Bonvini and Kennedy, 2020].

2.1 Introduction

In an experiment, the random assignment of the treatment to the units ensures that any measured and unmeasured factors are balanced between the treatment and control groups, thereby allowing the researcher to attribute any observed effect to the treatment. In observational studies, however, achieving such balance requires the untestable assumption that all confounders, roughly variables affecting both the treatment A and the outcome Y, are collected. To gauge the consequences of departures from the no-unmeasured-confounding assumption, a sensitivity analysis generally posits the existence of an unmeasured confounder U and varies either the U-A association or the U-Y association or both. The minimal strength of these associations that would drive the observed Y-A association to zero is often reported as a measure of the study's robustness to unmeasured confounding.

Since the seminal work of Cornfield et al. [1959] on the association between smoking and lung cancer, a plethora of sensitivity analysis frameworks have been proposed. Here, we mention a few of them and refer to Liu et al. [2013] and Richardson et al. [2014] for excellent reviews. In the context of matched studies, Rosenbaum's framework [Rosenbaum, 1987, 2002] is likely the most commonly used. It governs the U-A association via a parameter $\Gamma \geq 1$ by requiring that, within each pair, the ratio of the odds that unit 1 is treated to the odds that unit 2 is treated falls in the interval $[\Gamma^{-1}, \Gamma]$. The U-Y association is often left unrestricted or bounded as in Gastwirth et al. [1998]. More recently, Zhao et al. [2019] and Yadlowsky et al.

[2018] have proposed extensions to this framework that do not require matching.

In addition, Ding and VanderWeele [2016] and VanderWeele and Ding [2017] have derived a bounding factor for certain treatment effects in terms of two sensitivity parameters governing the U-A and U-Y relationships. Other authors have proposed modeling the distribution of U and the relationships $U - Y$ and $U - A$ directly [Imbens, 2003, Rosenbaum and Rubin, 1983], which has been recently extended to the case where the distribution of U is left unspecified by Zhang and Tchetgen Tchetgen [2019]. In the context of time-varying treatments, sensitivity analyses have been proposed for marginal structural models [Brumback et al., 2004] and cause-specific selection models [Rotnitzky et al., 2001].

In this paper, we propose a novel approach to sensitivity analysis based on a mixture model for confounding. We conceptualize that an unknown fraction ϵ of the units in the sample is arbitrarily confounded while the rest is not. The parameter ϵ is unknown and not estimable but can be varied as a sensitivity parameter. As discussed below, our model generalizes some relaxations to the no-unmeasured-confounding assumption that have been previously proposed in the literature. Furthermore, our framework yields a natural one-number summary of a study's robustness: the minimum proportion of confounded units such that bounds on the average treatment effect contain zero. All the code can be found in the Github repository matteobonvini/experiments-sensitivity-paper.

2.1.1 Motivation

The most widely adopted frameworks for sensitivity analysis generally assume that each unit in the sample could be subject to unmeasured confounding and then proceed by specifying the maximal extent of such confounding. However, just like a treatment effect can be heterogeneous, confounding, too, can differ between units. We propose a complementary approach: in some instances, the researcher may have failed to measure relevant confounders but may hope that there is a subset of units, possibly unknown, for whom the treatment is as good as randomized given the measured covariates.

As a toy example, suppose it is observed that adolescent alcohol drinking (treatment A) is positively associated with the occurrence of liver diseases (outcome Y). Suppose all confounders X have been recorded except for parental smoking, which could be associated with both A [Oliveira et al., 2019, Pengpid and Peltzer, 2019] and Y due to second-hand smoking [Lammert et al., 2013]. Previously proposed sensitivity analyses would check whether a small association between parental smoking and A or Y can explain away the observed A-Y association. Instead, we propose to leverage on the observation that parental smoking is a confounder only for units whose parents smoke at home. For instance, some parents may only smoke at work, in which case parental smoking would not have an effect on Y. The sample is thus composed of two groups: those units for which A is as good as randomized given X because they are not subject to second-hand smoking regardless of whether their parents smoke and those for which it is not. Depending on how prevalent the former group is, the observed A–Y association might be at least partially attributed to the effect of A. This toy example generalizes to other cases. For instance, if a confounder is measured with error, the observed covariates may be sufficient

to de-confound the treatment-outcome relationship only for an unknown subset of units. In such case, the sample can be thought of containing two groups: those units for whom the confounder was measured correctly, e.g. if the questionnaire on motivation or drugs usage was answered truthfully, and those for whom it was not.

The possibility that a sample comes from a mixture of distributions has been studied in great detail in statistics. In robust statistics, for example, it is assumed that a small unknown fraction of the sample comes from a "corrupted" or "contaminated" distribution that is not the target of inference (see Remark 2). In causal inference, unmeasured confounding takes the role of contamination. Borrowing the contaminated model from this literature, we conceptualize that an unknown fraction of the sample suffers from unmeasured confounding.

For example, consider Figure 2.1. In the shaded region of the space defined by the two observed covariates, the treatment is not assigned randomly; units with covariates' values falling in this region may have self-selected into the treatment arms and therefore estimating the effect of the treatment on their outcomes is impossible without making further, untestable assumptions. For brevity, we say these units are "confounded," while the other units are "unconfounded." Note that, except in special cases, some of which are discussed next, the region is not identifiable from the observed data. However, even if the region is not identifiable, its measure, termed ϵ in our model, might be specified or upper bounded using subject-matter knowledge. More generally, ϵ can be varied as a sensitivity parameter. In Figure 2.1, despite covering different sets of units, all three regions have the same mass, with approximately 20% of the points falling inside them. Given a value for ϵ, we show how to find the region yielding the most conservative inference.

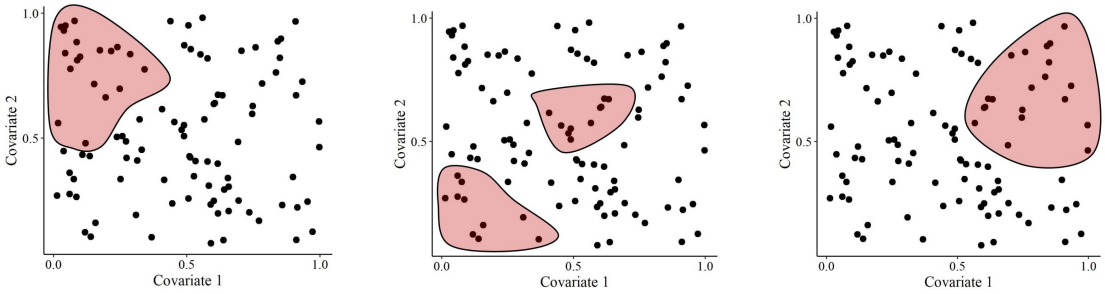


Figure 2.1: The shaded region represents the set of units for whom the treatment is not assigned randomly, even after conditioning on observed covariates. All three figures show approximately the same number of points falling within the "confounded region," albeit covering different sets of units. The probability ϵ that a unit falls within the region is our model's sensitivity parameter, here $\epsilon \approx 0.2$.

Special cases of our model have already been discussed in the literature when it is known who the confounded units are. For example, in introducing the selective ignorability framework, Joffe et al. [2010] discuss estimating the effect of erythropoietin alpha (EPO) on mortality using an observational database containing information on all subjects in the United States on hemodialysis. The treatment is thought to be unconfounded only after conditioning on

hematocrit, which, however, is not recorded for 10.6% of the subjects. Thus, one may view 10.6% of the sample as coming from a "confounded distribution." In addition, the differential effects framework proposed in Rosenbaum [2006], too, can be regarded as a special case of our model. Differential effects are treatment contrasts that are immune to certain types of biases called "generic biases." For example, suppose two treatments are under study. In certain cases, it is plausible that, while units might self select into either treatment arm, the choice of the treatment among units who take exactly one treatment is as good as random. Notice that this setup is a special case of our model: the confounded units are precisely those who are not taking any treatment or are taking a combination of both of them.

Finally, a standard instrumental variables (IV) setting, too, can be thought of as a case where a fraction of the units is unconfounded. For example, consider an experiment with binary treatment that suffers from units' non-compliance. The treatment assignment is randomized but the treatment received is not. For the units who complied with the experimental guidelines, the treatment received is equal to the treatment assigned, which is randomly assigned. Thus, the compliers can be considered the units for whom the treatment / outcome relationship is not confounded. In fact, in their detailed analysis of the binary IV model, Richardson and Robins [2010] propose a sensitivity analysis for the average treatment effect where the sensitivity parameter can be expressed as the proportion of compliers. For the observational setting considered in this paper, however, the instrument is never observed, thus, contrarily to a standard IV analysis, the sample contains no information regarding who the confounded units are. In this light, our contribution can also be regarded as an attempt to infer average treatment effects when it is plausible that nature is acting via an unobservable IV.

2.2 The Sensitivity Model

We suppose we are given an iid sample $(\mathbf{O}_1, \dots, \mathbf{O}_n) \sim \mathbb{P}$ with $\mathbf{O} = (\mathbf{X}, A, Y)$, for covariates $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^p$, a binary treatment $A \in \{0, 1\}$ and an outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$. We let Y^a denote the potential outcome that would have been observed had the treatment been set to $A = a$ [Rubin, 1974]. The goal is to estimate the Average Treatment Effect (ATE) defined as $\psi = \mathbb{E}(Y^1 - Y^0)$. To ease the notation, we let $\pi(a \mid \mathbf{X}) = \mathbb{P}(A = a \mid \mathbf{X})$,

$$\mu_a(\mathbf{X}) = \mathbb{E}\left(Y \mid A = a, \mathbf{X}\right), \quad \text{and} \quad \boldsymbol{\eta} = \{\pi(0 \mid \mathbf{X}), \pi(1 \mid \mathbf{X}), \mu_0(\mathbf{X}), \mu_1(\mathbf{X})\}.$$

Throughout, we assume that the following two assumptions hold

Assumption 1 (Consistency). $Y = AY^1 + (1 - A)Y^0$.

Assumption 2 (Positivity). $\mathbb{P}\{t \leq \pi(a \mid \mathbf{X}) \leq 1 - t\} = 1$ for some $t > 0$.

Both assumptions are standard in the causal inference literature. Consistency rules out any interference between the units, whereas positivity requires that each unit has a non-zero chance of receiving either treatment arm regardless of their covariates' values. It is well known that if, in addition to consistency and positivity, it also holds that $Y^a \perp\!\!\!\perp A \mid \mathbf{X}$ (no unmeasured confounding), then ψ can be point-identified as $\psi = \mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\}$. In this work, we

propose a sensitivity model that relaxes the no-unmeasured-confounding assumption while retaining both consistency and positivity. As a consequence of this relaxation, ψ is no longer point-identified but it can still be bounded.

Our model supposes that the observed distribution $\mathbb{P}$ is derived from a counterfactual distribution $\mathbb{Q}$ of $(\mathbf{X}, A, Y^1, Y^0)$ such that

$$\mathbb{Q} = \epsilon\mathbb{Q}_0 + (1-\epsilon)\mathbb{Q}_1 \tag{2.1}$$

where $\mathbb{Q}_0$ is a "confounded distribution" for which $A \not\perp\!\!\!\perp Y^a \mid \mathbf{X}$ and $\mathbb{Q}_1$ is an "unconfounded distribution" for which $A \perp\!\!\!\perp Y^a \mid \mathbf{X}$. In practice, it might be useful to think of each $\mathbb{Q}_i$ as potentially factoring according to $A \perp\!\!\!\perp Y^a \mid \mathcal{S}_i$, where $\mathcal{S}_i$ is a set of confounding variables such that $\mathcal{S}_1$ is measured but $\mathcal{S}_0 \setminus \mathcal{S}_1 \neq \emptyset$ is not. [1] [2]

The parameter $\epsilon \in \mathcal{E} \subseteq [0,1]$ governs the proportion of unmeasured confounding. It is unknown and not estimable but can be varied as a sensitivity parameter. Here, $\mathcal{E}$ is an interval that the user can specify. Although ψ cannot be point-identified for $\epsilon > 0$, it is possible to bound it as a function of ϵ. In particular, for $\epsilon = 1$, the familiar worst-case bounds are recovered. For an outcome bounded in $[0,1]$, these bounds have width equal to 1, which means that the sign of the treatment effect is not identified. Varying the sensitivity parameter to recover different identification regions has been proposed in other works, such as Richardson et al. [2014], Kennedy et al. [2019] and Díaz and van der Laan [2013a], albeit for different targets of inference or sensitivity models.

An equivalent formulation of our model (2.1) is one where there is a latent selection indicator $S \in \{0,1\}$, with $\mathbb{P}(S=1) = 1-\epsilon$, such that $A \not\perp\!\!\!\perp Y^a \mid \mathbf{X}, S=0$, but $A \perp\!\!\!\perp Y^a \mid \mathbf{X}, S=1$. The following lemma rewrites ψ in terms of S.

Lemma 1. *Let* $\lambda_a(\mathbf{X}) = \mathbb{E}(Y^a \mid A = 1-a, \mathbf{X}, S=0)$. *Under consistency* (1) *and positivity* (2)*, it holds that*

$$\psi = \mathbb{E}((1-S)[\{Y - \lambda_{1-A}(\mathbf{X})\}(2A-1)] + S\{\mathbb{E}(Y \mid A=1, \mathbf{X}, S=1) - \mathbb{E}(Y \mid A=0, \mathbf{X}, S=1)\})$$

All proofs can be found in the supplementary material. As shown in Lemma 1, ψ depends on three unobservable quantities: $\lambda_0(\mathbf{X})$, $\lambda_1(\mathbf{X})$ and S. The quantity $\lambda_1(\mathbf{X})$ ($\lambda_0(\mathbf{X})$) represents the average outcome for those control (treated) units subject to unmeasured confounding had they taken the treatment (control) instead. Without further assumptions, the observed distribution $\mathbb{P}$ would not impose any restrictions on $\lambda_0(\mathbf{X})$ or $\lambda_1(\mathbf{X})$ even if S was known.

[1] As pointed out by an anonymous reviewer, the mixture model (2.1) could be generalized to $\mathbb{Q} = \sum_{j=1}^{J} \epsilon_j \mathbb{Q}_j$, where each Q_j is a distribution on the counterfactuals capturing different degrees of the confounding. While richer sensitivity analyses can yield more nuanced conclusions, the large number of parameters whose plausibility range would need to be assessed ($J-1$ in this case) may hinder their applications in many settings.

[2] For instance, consider the toy example above, with $X = \emptyset$ and $Y, A, U \in \{0,1\}$ for simplicity. Suppose that $\mathbb{P}(U=1 \mid A) = \gamma_0 + \gamma_1 A$ and $\mathbb{Q}_s(Y^a = 1 \mid A, U) = \alpha_1 s + (1-s)(\alpha_2 + \alpha_3 U)$, for some constants γ and α. Then, $\mathbb{E}_{\mathbb{Q}_0}(Y^1 - Y^0) = \mathbb{E}_{\mathbb{Q}_1}(Y^1 - Y^0) = 0$ and $\mathbb{E}_{\mathbb{Q}_1}(Y^1 \mid A=1) - \mathbb{E}_{\mathbb{Q}_1}(Y^0 \mid A=0) = 0$, but $\mathbb{E}_{\mathbb{Q}_0}(Y^1 \mid A=1) - \mathbb{E}_{\mathbb{Q}_1}(Y^0 \mid A=0) = \alpha_3\gamma_1$, which is generally nonzero.

For any given ϵ, a sharp lower (upper) bound on ψ can be obtained by minimizing (maximizing) ψ in Lemma 1 over $\lambda_0(\mathbf{X})$, $\lambda_1(\mathbf{X})$ and S. Without imposing some restrictions on the distribution of S, the optimization step involves finding, and nonparametrically estimating, the optimal regression functions $\mathbb{E}(Y \mid A = a, \mathbf{X}, S = 1)$. Given a sample of n observations, this step would involve fitting regression functions on $\binom{n}{\lceil n\epsilon \rceil}$ different sub-samples of size $\lceil n\epsilon \rceil$, which is computationally very costly even for moderate sample sizes.

Instead, we proceed by requiring that $S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$; we call the resulting sensitivity model "X-mixture model". The assumption that $S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$ can be interpreted in at least three ways. First, one may hope that it holds exactly for the mechanism that generated the sample. For instance, it is trivially satisfied, for example, if S is just a possibly unknown, deterministic function of the observed covariates. An example satisfying this condition is given by the selected ignorability framework proposed in Joffe et al. [2010]: if the treatment is as good as randomized conditional on hematocrit (and possibly other observed covariates), then S could be an indicator of whether hematocrit is missing.

Even if it does not hold exactly, assuming $S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$ may be a close approximation to reality that one can use to make the problem computationally tractable. This second interpretation is in the same spirit as using parametric regression models in order to simplify a given problem, hoping that they will be a close approximation to the true regression function. Third, even if $S \not\perp\!\!\!\perp (Y, A) \mid \mathbf{X}$, the X-mixture model can help determining whether a study is not robust to unmeasured confounding. Because the bounds if no assumptions are made will be at least as wide as those under $S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$, if a study does not appear robust in the X-mixture model, it will not appear robust in the general case either. In the following theorem, we derive closed-form expressions for sharp bounds on ψ in the X-mixture model.

Theorem 1 (Bounds in X-mixture model)**.** *Suppose that assumptions 1 and 2 hold. Further suppose that*

$$S \perp\!\!\!\perp (A, Y) \mid \mathbf{X} \tag{A1}$$

and that $\mathbb{P}(Y \in [y_{min}, y_{max}]) = 1$*, for* y_{min}, y_{max} *finite. Choose* $\delta \in [0, 1]$ *such that*

$$L_a \equiv \delta\{y_{min} - \mu_a(\mathbf{X})\} \leq \lambda_a(\mathbf{X}) - \mu_a(\mathbf{X}) \leq \delta\{y_{max} - \mu_a(\mathbf{X})\} \equiv U_a \text{ with prob. 1} \tag{2.2}$$

for $a \in \{0, 1\}$*. Then, as a function of* ϵ*, sharp bounds on* ψ *are:*

$$\begin{aligned}\psi_l(\epsilon) &= \mathbb{E}\left[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\left\{g(\boldsymbol{\eta}) \leq q_\epsilon\right\} g(\boldsymbol{\eta})\right] - \epsilon\delta(y_{max} - y_{min}) \\ \psi_u(\epsilon) &= \mathbb{E}\left[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\left\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\right\} g(\boldsymbol{\eta})\right]\end{aligned}$$

where $g(\boldsymbol{\eta}) = \pi(0 \mid \mathbf{X})U_1 - \pi(1 \mid \mathbf{X})L_0$ *and* q_τ *is its* τ*-quantile.*

Theorem 1 yields the identification of sharp lower and upper bounds on ψ when it is suspected that $100\epsilon\%$ of the units in the sample are confounded and it is assumed that predicting whether a unit is confounded or not cannot be improved by conditioning on (Y, A). Relaxing condition (A1) to $S \perp\!\!\!\perp Y \mid (A, \mathbf{X})$ poses no additional challenges and it is discussed in

Appendix A.3. We refer to this relaxed version of the X-mixture model as the "XA-mixture model." Notably, it covers the differential effects framework of Rosenbaum [2006], as one could specify $S = \mathbb{1}(A_1 + A_2 = 1)$ for some binary treatment A_1 and A_2.

The bounds are in terms of the parameters ϵ and δ, as well as the regression functions $\pi(a \mid \mathbf{X})$ and $\mu_a(\mathbf{X})$, and they involve non-smooth transformations of unknown functions of $\mathbb{P}$. The parameter ϵ is our main sensitivity parameter and controls the proportion of unmeasured confounding in the sample. Parallely, δ controls the extent of unmeasured confounding among the $S = 0$ units, as it bounds the difference between the unobservable regression $\lambda_a(\mathbf{X})$ and the estimable regression $\mu_a(\mathbf{X})$. Notice that (2.2) always holds for $\delta = 1$. Setting $\delta < 1$ imposes an untestable assumption on the severity of the unmeasured confounding, which might be sensible if some knowledge on the confounding mechanism is available. Specifically, our parametrization is such that $\lambda_a(\mathbf{X})$ can be bounded by linear combinations of y_{min}, y_{max} and $\mu_a(\mathbf{X})$:

$$\delta y_{\text{min}} + (1-\delta)\mu_a(\mathbf{X}) \leq \lambda_a(\mathbf{X}) \leq \delta y_{\text{max}} + (1-\delta)\mu_a(\mathbf{X})$$

Unless otherwise specified, in what follows we consider $y_{\text{min}} = 0$, $y_{\text{max}} = 1$ and set $\delta = 1$, thus yielding

$$\begin{aligned}\psi_l(\epsilon) &= \mathbb{E}\left[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\left\{g(\boldsymbol{\eta}) \leq q_\epsilon\right\} g(\boldsymbol{\eta})\right] - \epsilon \\ \psi_u(\epsilon) &= \mathbb{E}\left[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\left\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\right\} g(\boldsymbol{\eta})\right]\end{aligned}$$

for $g(\boldsymbol{\eta}) = \pi(0 \mid \mathbf{X})\{1 - \mu_1(\mathbf{X})\} + \pi(1 \mid \mathbf{X})\mu_0(\mathbf{X})$. If Y is bounded, this choice does not impose any assumption since Y can be rescaled to be in $[0, 1]$. If Y is unbounded, Theorem 1 is not directly applicable, but a similar result can be derived if one is willing to assume that $|\lambda_a(\mathbf{X}) - \mu_a(\mathbf{X})| \leq \delta$ for $a \in \{0, 1\}$ and $\delta < \infty$. We leave further investigation of the unbounded case as future work. We conclude this section with four remarks aiming to shed some more light on the bounds derived in Theorem 1.

Remark 1. Suppose Y is bounded in $[0, 1]$ and take $\delta = 1$. The length of the bound is then

$$\Delta(\epsilon) = [\mathbb{E}\{g(\boldsymbol{\eta}) \mid g(\boldsymbol{\eta}) > q_{1-\epsilon}\} - \mathbb{E}\{g(\boldsymbol{\eta}) \mid g(\boldsymbol{\eta}) \leq q_\epsilon\} + 1]\epsilon$$

If S was known, the length of the bound would reduce to $\Delta(\epsilon) = \epsilon$. Thus, we can view the term $[\mathbb{E}\{g(\boldsymbol{\eta}) \mid g(\boldsymbol{\eta}) > q_{1-\epsilon}\} - \mathbb{E}\{g(\boldsymbol{\eta}) \mid g(\boldsymbol{\eta}) \leq q_\epsilon\}]\epsilon$ as the "cost" of not knowing who the confounded units are.

Remark 2. The conditional independence of S and Y considerably simplifies the optimization step. To see this, notice that $\mathbb{E}(Y \mid A = a, \mathbf{X}, S = 1) = \mu_a(\mathbf{X})$ if $S \perp\!\!\!\perp Y \mid A, \mathbf{X}$. In turn, this implies that ψ can be written as

$$\psi = \mathbb{E}[\Gamma(Y, A, \mathbf{X}) + S\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) - \Gamma(Y, A, \mathbf{X})\}]$$

where $\Gamma(Y, A, \mathbf{X}) = \{Y - \lambda_{1-A}(\mathbf{X})\}(2A - 1)$. Therefore, bounds on ψ can be derived from

bounds on $\mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) - \Gamma(Y, A, \mathbf{X}) \mid S = 1\}$, which fits the framework studied by Horowitz and Manski [1995]. In their work, the goal is to do inference about a distribution Q_1 using data Y such that $Y = ZY_1 + (1 - Z)Y_0$, with $Z \in \{0, 1\}$ and $Y_i \sim Q_i$. They discuss two models: the "contaminated sampling model", which assumes Z to be independent of Y_1, and the "corrupted sampling model", which does not make this assumption. If it is known that $\mathbb{P}(Z = 0) \leq \lambda$, they derive sharp bounds on the conditional expectation of Y_1 given some covariates $\mathbf{X}$ when contamination or corruption does not occur in $\mathbf{X}$. Our setup does not immediately fit this framework because corruption applies to all observed variables $(Y, A, \mathbf{X})$. However, if $S \perp\!\!\!\perp Y \mid A, \mathbf{X}$, the optimal solution for S can be found by considering only the marginal distribution of the one-dimensional random variable $\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) - \Gamma(Y, A, \mathbf{X})$. Following the terminology in Horowitz and Manski [1995], we may view the assumption that $S \perp\!\!\!\perp Y \mid A, \mathbf{X}$ as a compromise between contamination ($S \perp\!\!\!\perp (Y, A, \mathbf{X})$) and corruption (no assumption on S).

Remark 3. As pointed out by Robins [2002], many interesting sensitivity analyses make use of parameters that depend on the covariates collected. In turn, this might hinder the direct comparison of studies' robustness. For example, a study where many confounders have been properly taken into account might appear more sensitive to departures from the no-unmeasured-confounding assumption than a study that failed to control for any confounder. This could happen, for instance, if the effect estimate in the former study is closer to the null value than the estimate from the latter. This apparent paradox might arise because a sensitivity analysis measures departures from a weak or strong assumption depending on whether many or few observed confounders are collected. Our proposed sensitivity analysis hinges on ϵ, the proportion of unmeasured confounding, which depends on the covariates collected. As such, it might be subject to this paradox.

Remark 4. Section 4 of Rosenbaum [1987] contains a modification to the sensitivity analysis proposed in that paper, and briefly summarized in our introduction, that allows an unknown fraction β of the sample to suffer from arbitrarily confounding. While conceptually similar to the approach presented in this paper, their method relies on exact matching. In fact, if units are exactly matched on observed covariates, our sensitivity model recovers Rosenbaum's with $\beta = \epsilon$ and $\Gamma = 0$. However, exact matching is often infeasible due to the presence of continuous or high-dimensional covariates. Therefore, our work can be viewed as an extension to Rosenbaum's Section 4 model to the case where units are not matched on observed covariates.

2.2.1 One-number Summary of a Study's Robustness

In practice, one might want to report a one-number summary of how robust the estimated effect is to the number of confounded units. An example of such summary is the minimum proportion of confounded units ϵ_0 such that the bounds on ψ are no longer informative about the sign of the effect, i.e. that they contain zero. Larger values of ϵ_0 indicate that the estimated

effect is more robust to potential unmeasured confounding. Mathematically,

$$\epsilon_0 = \underset{\epsilon \in \mathcal{E}}{\operatorname{argmin}} \, \mathbb{1}[\operatorname{sgn}\{\psi_l(\epsilon)\} \neq \operatorname{sgn}\{\psi_u(\epsilon)\}]$$

where $\operatorname{sgn}(x)$ measures the sign of x, $\operatorname{sgn}(x) = -\mathbb{1}(x < 0) + \mathbb{1}(x > 0)$. Because $\psi_u(\epsilon = 1) - \psi_l(\epsilon = 1) = 1$, the minimum is guaranteed to be attained in $\mathcal{E} = [0, 1]$. Furthermore, under certain mild conditions, the bounds are continuous and strictly monotone in ϵ, hence ϵ_0 is generally the unique value such that $\psi_l(\epsilon_0) = 0$ or $\psi_u(\epsilon_0) = 0$. This motivates the moment condition $\psi_l(\epsilon_0)\psi_u(\epsilon_0) = 0$, which we use to construct a Z-estimator of ϵ_0.

Other authors have proposed one-number summaries of a study's robustness to unmeasured confounding. For example, the minimum value for Γ in Rosenbaum's framework and its extensions [Gastwirth et al., 1998, Rosenbaum, 1987, Yadlowsky et al., 2018, Zhao et al., 2019] such that the observed effect ceases to be statistically significant can be used as a summary of study's robustness to unmeasured confounding. Recently, Ding and VanderWeele [2016] and VanderWeele and Ding [2017] have introduced the E-Value, which measures the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the outcome and the treatment in order to "explain away" the observed effect of the treatment on the outcome. In order to derive the elegant formula for the E-Value, the unobserved confounder is assumed to be associated with the treatment and with the outcome in equal magnitude. Furthermore, the derivation makes use of a bounding factor that needs to be computed for each stratum of the covariates. Computing such bounding factor when the observed covariates are continuous or high-dimensional can be problematic. Moreover, their method requires additional approximations if the outcome is not binary. On the other hand, the one-number summary proposed here does not require any further assumption other than the restriction on S described above. Hence, we view these summary measures as complementary and the specific context would generally dictate which one is more appropriate.

2.3 Estimation & Inference

2.3.1 Proposed Estimators

There are at least two types of bias that can arise when estimating a causal effect using observational data: the bias arising from incorrectly assuming that all confounders have been collected and the statistical bias of the chosen estimator [Luedtke et al., 2015]. In Section 2.2, we constructed a model to probe the effects of the former bias. In this section, we propose estimators that aim to minimize the latter. Our estimators of the bounds are built using the efficient influence functions (IFs) and cross-fitting. IFs play a crucial role in nonparametric efficiency theory, as the variance of the efficient IF can be considered the nonparametric counterpart of the Cramer-Rao lower bound in parametric models. Furthermore, estimators constructed using the efficient IF have favorable properties, such as doubly-robustness or second-order bias. Here, we note that $\psi_l(\epsilon)$ and $\psi_u(\epsilon)$ do not possess an influence function, as they are not pathwise differentiable. However, certain terms appearing in their expressions, such as $\mathbb{E}\{\mu_a(\mathbf{X})\}$, are pathwise differentiable; as such, they can be estimated using IFs. For

terms that are not pathwise differentiable we resort to plug-in estimators. We refer to Bickel et al. [1993], van der Vaart [2002], Van der Laan et al. [2003], Tsiatis [2007], Chernozhukov et al. [2016] and others for detailed accounts on IFs and their use.

To ease the notation in this section, let

$$\nu(\mathbf{O};\boldsymbol{\eta}) = \frac{(2A-1)\{Y-\mu_A(\mathbf{X})\}}{\pi(A \mid \mathbf{X})} + \mu_1(\mathbf{X}) - \mu_0(\mathbf{X})$$

denote the uncentered influence function for the parameter $\mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\}$. Furthermore, let $\tau(\mathbf{O};\boldsymbol{\eta})$ denote the uncentered influence function for $\mathbb{E}\{g(\boldsymbol{\eta})\}$:

$$\tau(\mathbf{O};\boldsymbol{\eta}) = \frac{(1-2A)\{Y-\mu_A(\mathbf{X})\}}{\pi(A \mid \mathbf{X})/\pi(1-A \mid \mathbf{X})} + A\mu_0(\mathbf{X}) + (1-A)(1-\mu_1(\mathbf{X}))$$

and let

$$\begin{aligned}\varphi_l(\mathbf{O};\boldsymbol{\eta};q_\epsilon) &= \nu(\mathbf{O};\boldsymbol{\eta}) + \mathbb{1}\{g(\boldsymbol{\eta}) \leq q_\epsilon\}\tau(\mathbf{O};\boldsymbol{\eta}) - \epsilon \\ \varphi_u(\mathbf{O};\boldsymbol{\eta};q_{1-\epsilon}) &= \nu(\mathbf{O};\boldsymbol{\eta}) + \mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\}\tau(\mathbf{O};\boldsymbol{\eta})\end{aligned}$$

Then, it holds that $\psi_l(\epsilon) = \mathbb{E}\{\varphi_l(\mathbf{O};\boldsymbol{\eta};q_\epsilon)\}$ and $\psi_u(\epsilon) = \mathbb{E}\{\varphi_u(\mathbf{O};\boldsymbol{\eta};q_{1-\epsilon})\}$.

Following Robins et al. [2008], Zheng and Van Der Laan [2010] and Chernozhukov et al. [2016] among others, we use cross-fitting to allow for arbitrarily complex estimators of the nuisance functions $\boldsymbol{\eta}$ and q_τ in order to avoid empirical process conditions. Specifically, we split the data into B disjoint groups of size n/B and we let $K_i = k$ indicate that subject i is split into group k, for $k \in \{1,\ldots,B\}$. Notice that it is not required that the groups have equal size, for example each K_i could be drawn uniformly from $\{1,\ldots,B\}$. For simplicity, we proceed with having equal-size groups. We let $\mathbb{P}_n$ denote the empirical measure as $\mathbb{P}_n\{f(\mathbf{O})\} = \frac{1}{n}\sum_{i=1}^n f(\mathbf{O}_i)$ and $\mathbb{P}_n^k$ denote the sub-empirical measure as $\mathbb{P}_n^k\{f(\mathbf{O})\} = \sum_{i=1}^n f(\mathbf{O}_i)\mathbb{1}(K_i = k)/\sum_{i=1}^n \mathbb{1}(K_i = k)$. In addition, we let $\widehat{\boldsymbol{\eta}}_{-k}$ denote the estimator of $\boldsymbol{\eta}$ computed without using observations from fold $K = k$ and $\widehat{q}_{\tau,-k}$ denote the estimator of q_τ equal to the empirical quantile of $g(\widehat{\boldsymbol{\eta}}_{-k})$ solving $\mathbb{P}_n^k[\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-k}) \leq \widehat{q}_{\tau,-k}\}] = \tau + o_{\mathbb{P}}(n^{-1/2})$. Then, we estimate the bounds as

$$\widehat{\psi}_l(\epsilon) = \frac{1}{B}\sum_{k=1}^B \mathbb{P}_n^k[\nu(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}) + \mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-k}) \leq \widehat{q}_{\epsilon,-k}\}\tau(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k})] - \epsilon \equiv \mathbb{P}_n\left\{\varphi_l(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-K},\widehat{q}_{-K,\epsilon})\right\}$$

$$\widehat{\psi}_u(\epsilon) = \frac{1}{B}\sum_{k=1}^B \mathbb{P}_n^k[\nu(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}) + \mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-k}) > \widehat{q}_{1-\epsilon,-k}\}\tau(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k})] \equiv \mathbb{P}_n\left\{\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-K},\widehat{q}_{-K,1-\epsilon})\right\}$$

The computation of the estimators above is straightforward as it amounts to fitting regression functions on $B-1$ subsets of the data and evaluate the estimated functions at the values of the covariates on the corresponding test set. The use of cross-fitting lends itself naturally to the use of parallel computing as one can estimate the regression functions on different subsets of the data simultaneously. We incorporate this possibility in our implementation of the methods

in R. Moreover, it is worth noting that cross-fitting does not discard any data point in the estimation step, since each observation is used twice without overfitting: once for estimating the regression functions and once for estimating the expectation operator. In addition, because we are working under a fully nonparametric model, there exists only one influence function; therefore, our estimators of the pathwise differentiable terms are efficient in the sense that they asymptotically achieve the semiparametric efficiency bound.

Finally, while the estimators of the bounds discussed in this section have several attractive properties in terms of computational tractability and convergence rates, they might not be monotone in ϵ in finite samples. To remedy this, the estimators can be "rearranged" using the procedure described in Chernozhukov et al. [2009]. We apply this procedure in Section 2.4, although we find that the original, non-rearranged estimators achieve low bias and nominal uniform coverage as well.

2.3.2 Establishing Weak Convergence

To state asymptotic guarantees for the proposed estimators, we first make the following technical assumption:

Assumption 3 (Margin Condition)**.** The random variable $g(\boldsymbol{\eta})$ has absolutely continuous CDF and there exists $\alpha > 0$ such that for all $t > 0$ and $\tau \in \mathcal{E}$, it holds that $\mathbb{P}\left(|g\left(\boldsymbol{\eta}\right) - q_\tau| \leq t\right) \lesssim t^\alpha$ and $\mathbb{P}\left(|g\left(\boldsymbol{\eta}\right) - q_{1-\tau}| \leq t\right) \lesssim t^\alpha$.

Assumption 3 requires that there is not too much mass around any ϵ-quantile or $(1-\epsilon)$-quantile of $g(\boldsymbol{\eta})$, for $\epsilon \in \mathcal{E}$. It is essentially equivalent to the margin condition used in classification problems [Audibert and Tsybakov, 2007], optimal treatment regime settings [Luedtke and Van Der Laan, 2016, van der Laan and Luedtke, 2014], and other problems involving estimation of non-smooth functionals [Kennedy et al., 2019, 2020]. Notably it is satisfied for $\alpha = 1$ if, for instance, the density of $g(\boldsymbol{\eta})$ is bounded on $\mathcal{E}$. We give the main convergence theorem for $\widehat{\psi}_u(\epsilon)$. A similar statement holds for $\widehat{\psi}_l(\epsilon)$.

Theorem 2. *Let*

$$\widehat{\sigma}_u^2(\epsilon) = \mathbb{P}_n\{(\varphi_u(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-K}, \widehat{q}_{1-\epsilon,-K}) - \widehat{\psi}_u(\epsilon) - \widehat{q}_{1-\epsilon,-K}[\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-K}) > \widehat{q}_{1-\epsilon,-K}\} - \epsilon])^2\}$$

be the estimator of the variance function

$$\sigma_u^2(\epsilon) = \mathbb{E}\{(\varphi_u(\mathbf{O}; \boldsymbol{\eta}, q_{1-\epsilon}) - \psi_u(\epsilon) - q_{1-\epsilon}[\mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\} - \epsilon])^2\}$$

If assumptions 1, 2 and 3 hold, and the following conditions also hold:

1. $\mathbb{P}\left\{t \leq \widehat{\pi}(a \mid \mathbf{X}) \leq 1 - t\right\} = 1$ *for* $a = 0, 1$ *and some* $t > 0$.
2. $\sup_{\epsilon \in \mathcal{E}} \left|\frac{\widehat{\sigma}_u(\epsilon)}{\sigma_u(\epsilon)} - 1\right| = o_{\mathbb{P}}(1)$.
3. $\| \sup_{\epsilon \in \mathcal{E}} |\varphi_u(\mathbf{o}; \widehat{\boldsymbol{\eta}}, \widehat{q}_{1-\epsilon}) - \varphi_u(\mathbf{o}; \boldsymbol{\eta}, q_{1-\epsilon}) - q_{1-\epsilon}[\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}) > \widehat{q}_{1-\epsilon}\} - \mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\}]| \| =$

$o_{\mathbb{P}}(1)$.

4. $\left(\left\|g(\widehat{\boldsymbol{\eta}}) - g\left(\boldsymbol{\eta}\right)\right\|_{\infty} + \sup_{\epsilon\in\mathcal{E}}\left|\widehat{q}_{1-\epsilon} - q_{1-\epsilon}\right|\right)^{1+\alpha} = o_{\mathbb{P}}(n^{-1/2})$, *for* α *satisfying assumption 3.*

5. $\|\widehat{\pi}(1 \mid \mathbf{X}) - \pi(1 \mid \mathbf{X})\|$ $max_a\|\widehat{\mu}_a(\mathbf{X}) - \mu_a(\mathbf{X})\| = o_{\mathbb{P}}(n^{-1/2})$.

Then $\sqrt{n}\{\widehat{\psi}_u(\epsilon) - \psi_u(\epsilon)\}/\widehat{\sigma}_u(\epsilon) \rightsquigarrow \mathbb{G}(\epsilon)$ *in* $\ell^{\infty}(\mathcal{E})$, *with* $\mathcal{E} \subseteq [0,1]$, *where* $\mathbb{G}(\cdot)$ *is a mean-zero Gaussian process with covariance* $\mathbb{E}\left\{\mathbb{G}(\epsilon_1)\mathbb{G}(\epsilon_2)\right\} = \mathbb{E}\left\{\phi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon_1})\phi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon_2})\right\}$ *and*

$$\phi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon}) = \frac{\varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon}) - \psi_u(\epsilon) - q_{1-\epsilon}[\mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\} - \epsilon]}{\sigma_u(\epsilon)}.$$

Theorem 2 gives sufficient conditions so that the estimated curves tracing the lower and upper bounds as a function of ϵ converge to a Gaussian process. In turn, this enables the computation of confidence bands trapping the average treatment effect with any desired confidence level uniformly over ϵ. The first three conditions of the theorem are quite mild. Condition 1 is a positivity condition requiring that the estimator of the propensity score is bounded away from 0 and 1. Condition 2 requires uniform consistency of the variance estimator at any rate. Condition 3 holds if, in addition to satisfying the margin assumption 3, $g(\widehat{\boldsymbol{\eta}})$ and $\widehat{q}_{\tau}$ converge uniformly, in $\mathbf{x}$ and ϵ respectively, to the truth at any rate.

The key assumptions are conditions 4 and 5. While more restrictive than the first three, these conditions can be satisfied even if flexible machine learning tools are used. In fact, condition 5 only requires that the product of the L_2 errors in estimating $\pi(a \mid \mathbf{X})$ and $\mu_a(\mathbf{X})$ is of order $n^{-1/2}$, which means that, for example, each regression function can be estimated at the slower rate $n^{-1/4}$. A rate of convergence in L_{∞} norm of order $n^{-1/4}$ is also sufficient to satisfy condition 4 if the density of $g(\boldsymbol{\eta})$ is bounded because the margin assumption 3 would hold for $\alpha = 1$. A convergence rate of order $n^{-1/4}$ can be achieved if nonparametric smoothness, sparsity or other structural assumptions are imposed on the true regression functions. For instance, if a minimax optimal estimator is used, in order to satisfy condition 5, it is sufficient that the underlying regression functions belong to a β-Hölder class with smoothness parameter $\beta > p/2$, where p is the number of covariates. In addition, even in regimes of very large p, convergence at $n^{-1/4}$ rate can be achieved under structural assumptions such as additivity or sparsity [Farrell, 2015, Horowitz, 2009, Kandasamy and Yu, 2016, Raskutti et al., 2012, Yang and Tokdar, 2015]. Furthermore, such convergence rate can also be achieved if the regression functions belong to the class of cadlag functions with bounded variation norm [Benkeser and Van Der Laan, 2016, van der Laan, 2017]. We refer to Györfi et al. [2006] among others for additional convergence results.

Similarly to Kennedy [2018], we can use Theorem 2 and the multiplier bootstrap to construct uniform confidence bands covering the identification region $[\psi_l(\epsilon), \psi_u(\epsilon)]$. Placing $(1-\alpha/2)$ uniform confidence bands on each curve also yields a (conservative) $(1-\alpha)$ uniform confidence band for ψ. We also deploy the procedure of Imbens and Manski [2004] to construct bands covering just ψ that are valid pointwise. Details are provided in Appendix A.5.2. Constructing

uniformly valid bands covering ψ, as opposed to the whole identification region, is left for future research.

2.3.3 Estimation of the One-Number Summary ϵ_0

In our settings, a natural way to define ϵ_0 is via the moment condition $\psi_l(\epsilon_0)\psi_u(\epsilon_0) = 0$ and construct an estimator $\widehat{\epsilon}_0$ defined implicitly as the solution to the empirical moment condition

$$\mathbb{P}_n\{\varphi_l(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-K}, \widehat{q}_{\widehat{\epsilon}_0,-K})\}\mathbb{P}_n\{\varphi_u(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-K}, \widehat{q}_{1-\widehat{\epsilon}_0,-K})\} = o_{\mathbb{P}}(n^{-1/2}).$$

Standard results in Z-estimation theory (Theorem 3.3.1 in van der Vaart and Wellner [1996]) yield the following theorem.

Theorem 3. *Suppose that the CDF G of $g(\boldsymbol{\eta})$ is strictly increasing in neighborhoods of q_{ϵ_0} and $q_{1-\epsilon_0}$. Suppose assumptions 1, 2, 3 and conditions 1, 3, 4, 5 (and 3's and 4's counterpart for the lower bound) from Theorem 2 are satisfied with $\mathcal{E} = [0, 1]$. Then*

$$\sqrt{n}\,(\widehat{\epsilon}_0 - \epsilon_0) \rightsquigarrow N\left(0, [\psi_u(\epsilon_0)(q_{\epsilon_0} - 1) + \psi_l(\epsilon_0)q_{1-\epsilon_0}]^{-2}\, var\{\widetilde{\varphi}(\epsilon_0)\}\right)$$

provided that the denominator $\psi_u(\epsilon_0)(q_{\epsilon_0} - 1) + \psi_l(\epsilon_0)q_{1-\epsilon_0} \neq 0$, and where the unscaled influence function is

$$\widetilde{\varphi}(\epsilon_0) = \psi_u(\epsilon_0)[\varphi_l(\mathbf{O}; \boldsymbol{\eta}, q_{\epsilon_0}) - q_{\epsilon_0}\mathbb{1}\{g(\boldsymbol{\eta}) \leq q_{\epsilon_0}\}] + \psi_l(\epsilon_0)[\varphi_u(\mathbf{O}; \boldsymbol{\eta}, q_{1-\epsilon_0}) - q_{1-\epsilon_0}\mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon_0}\}].$$

Theorem 3 describes sufficient conditions so that $\widehat{\epsilon}_0$ is $\sqrt{n}$-consistent and asymptotically normally distributed. We require the same conditions as the ones required for Theorem 2, plus that the CDF of $g(\boldsymbol{\eta})$ is strictly increasing in neighborhoods of q_{ϵ_0} and $q_{1-\epsilon_0}$. The asymptotic normality of $\widehat{\epsilon}_0$ relies on the existence (and non-singularity) of the derivative of the map $\epsilon \mapsto \psi_l(\epsilon)\psi_u(\epsilon)$ at $\epsilon = \epsilon_0$. Calculating such derivative requires computing the derivative of the quantile function, which is why we require the CDF of $g(\boldsymbol{\eta})$ to be strictly increasing in the relevant neighborhoods. We expect all these conditions to be satisfied in practice in the presence of continuous covariates and enough smoothness or sparsity for the regression functions.[3] Asymptotic normality allows the straightforward calculation of a Wald-type confidence interval for ϵ_0 using a consistent estimate for the variance. We thus propose reporting both a point-estimate for ϵ_0 and $1 - \alpha$ confidence interval as a summary of the study's robustness to unmeasured confounding.[4]

[3] In principle, one could construct the empirical moment condition after performing the rearrangement procedure of Chernozhukov et al. [2009]. Whether or not the rearrangement is done, we expect the inference about ϵ_0 to be equivalent asymptotically and vary minimally in finite samples.

[4] In order to incorporate finite sampling uncertainty in sensitivity analyses, one-number summaries of a study's robustness are generally computed as the values of the sensitivity parameter(s) such that a α-level confidence interval for the ATE under no unmeasured confounding includes the null value. Choosing different αs to estimate the ATE with no residual confounding may then yield different conclusions regarding the study's robustness to unmeasured confounding, despite the latter being a separate inferential task. Constructing a confidence interval for ϵ_0 directly bypasses this issue.

2.4 Illustrations

2.4.1 Simulation Study

In this section, we report the results of the simulations we performed to investigate the finite-sample performance of our proposed estimators. We consider the following data generating mechanism:

$$
\begin{aligned}
&X_i \sim \text{TruncNorm}(\mu = 0, \sigma = 1, \text{lb} = -2, \text{ub} = 2) \text{ for } i \in \{1, 2\},\ U \sim \text{Bern}(0.5),\\
&S \mid X_1, X_2 \sim \text{Bern}\{\Phi(X_1)\},\\
&A \mid X_1, X_2, U, S \sim \text{Bern}[0.5\{\Phi(X_1) + 0.5S + (1 - S)U\}],\\
&Y^a \mid X_1, X_2, U, S, A \sim \text{Bern}\{0.25 + 0.5\Phi(X_1 + X_2) + (a - 0.5)r - 0.1U\},\\
&Y = AY^1 + (1 - A)Y^0,
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the CDF of a standard normal random variable. Notice that

$$
\mathbb{P}(A = 1 \mid X_1, X_2, S = 0) = \mathbb{P}(A = 1 \mid X_1, X_2, S = 1) = 0.5\Phi(X_1) + 0.25,
$$

thus this model satisfies the assumptions of Theorem 2 and it implies that $\mathbb{E}(Y^1 - Y^0) = r$. The random variable U acts as a binary unmeasured confounder; given the observed covariates $\mathbf{X}$, units with $S = 0$ and $U = 1$ are more likely to be treated and exhibit $Y = 0$ than those with $S = 0$ and $U = 0$. Therefore, under this setup, one would expect the treatment effect to be underestimated if the no-unmeasured-confounding assumption is (incorrectly) assumed to be true.[5]

We estimate the lower bound $\psi_l(\epsilon)$, the upper bound $\psi_u(\epsilon)$ and ϵ_0 using the methods outlined in Section 2.3.1. In particular, we use 5-fold cross-fitting to estimate the nuisance functions, fitting both generalized linear and additive models via the SuperLearner method [Van der Laan et al., 2007]. The performance of the proposed estimators is evaluated via integrated bias, root-mean-squared-error (RMSE), and coverage. These evaluation metrics offer insight into what sample size is required to achieve a good performance of the multiplier bootstrap, which relies on the convergence of the bounds' estimators to a Gaussian process.

$$
\widehat{\text{bias}} = \frac{1}{I}\sum_{i=1}^{I}\left|\frac{1}{J}\sum_{j=1}^{J}\{\widehat{\psi}_{l,j}(\epsilon_i) - \psi_{l,j}(\epsilon_i)\}\right|, \quad \widehat{\text{RMSE}} = \frac{1}{I}\sum_{i=1}^{I}\left[\frac{1}{J}\sum_{j=1}^{J}\{\widehat{\psi}_{l,j}(\epsilon_i) - T_j(\epsilon_i)\}^2\right]^{1/2}
$$

and suitably modified formulas for $\psi_u(\epsilon)$ and ϵ_0. We run $J = 500$ simulations across $I = 21$ values of ϵ equally spaced in $\mathcal{E} = [0, 0.2]$. To better estimate ϵ_0 we make the grid finer and consider 201 values of ϵ equally spaced in $\mathcal{E}$. To evaluate 95% uniform coverage, we say that the uniform band covers if it contains the true region $[\psi_l(\epsilon), \psi_u(\epsilon)]$ for all $\epsilon \in \mathcal{E}$. Finally, we assess bias and 95% coverage for ϵ_0.

[5] In the context of the toy example of Section 2.1.1, U and S indicate whether the parents are smokers and whether they would smoke at home respectively, X_1 and X_2 may be measures of the parents' education level and income respectively, A indicates adolescent alcohol consumption and Y indicates the occurrence of liver disease.

n	Bias ($\times 100$)			$\sqrt{n}\times$RMSE			Coverage ($\times 100$)	
	$\psi_l(\epsilon)$	$\psi_u(\epsilon)$	ϵ_0	$\psi_l(\epsilon)$	$\psi_u(\epsilon)$	ϵ_0	$[\psi_l(\epsilon), \psi_u(\epsilon)]$	ϵ_0
500	0.38	0.12	2.47	0.95	0.96	1.32	95.4	97.0
1000	0.51	0.14	1.59	0.95	0.95	1.45	93.2	95.6
5000	0.04	0.10	0.16	0.99	0.98	1.72	92.4	95.4
10000	0.05	0.09	0.07	0.95	0.96	1.75	93.6	94.8

Table 2.1: Simulation results across 500 simulations.

Table 2.1 shows the results of our simulation for $r = 0.05$. This set up is such that $\epsilon_0 = 0.041$. In addition, if no-unmeasured-confounding is erroneously thought to hold ($\epsilon = 0$), ψ is, on average, underestimated since $\mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\} \approx 0.023 < r$. This simple simulation setup exemplifies what our theory predicts. Even for moderate sample sizes, we achieve approximately correct nominal uniform coverage for the identification region and ϵ_0. Furthermore, the $\sqrt{n}\times$RMSE remains roughly constant as the sample size increases. Finally, in Section A.7 of the Appendix, we extend this simulation study to investigate how conservative our model would be if the true ϵ_0 is actually zero, i.e. there is no unmeasured confounding.

2.4.2 Application

In this section, we illustrate the proposed sensitivity analysis by reanalyzing the data from the study on Right Heart Catheterization (RHC) conducted by Connors et al. [1996].[6] The data consist of 5735 records from critically ill adult patients receiving care in an ICU for certain disease categories in one out of five US teaching hospitals between 1989 and 1994. For each patient, demographic variables, comorbitidies and diagnosis variables as well as several laboratory values were recorded. A total of 2184 patients underwent RHC within the first 24 hours in the ICU. Within 30 days of admission, 1918 patients died, approximately 38.00% and 30.64% of the treated and control groups respectively. After conditioning on the measured confounders, the authors concluded that patients treated with RHC had, on average, lower probability of surviving (30-day mortality: OR $= 1.24$, 95% CI $= [1.03, 1.49]$). Notably, sensitivity analyses targeting potential violations of the propensity score model suggested robustness of the study's conclusions to unmeasured confounding.

We investigate the effects of varying the proportion of confounded units while avoiding any parametric assumptions on the nuisance regression functions. One reason to believe that a fraction of the sample might be effectively unconfounded is the following. Suppose there are two types of surgeons: those who prefer performing RHC (R-surgeon) and those who don't (NR-surgeon). One might believe that the surgeon's preference for RHC is a valid instrument. Roughly, an instrument is a variable that is unconfounded, associated with the treatment receipt, and that affects the outcome only through the treatment. It appears plausible that a surgeon's preference for RHC would satisfy these conditions if, for instance, the efficacy of RHC was not well understood at the time the study was conducted. In fact, physicians' preferences

[6] Available at http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets.

for a treatment have been used as IVs before, see for example Hernán and Robins [2006] and Baiocchi et al. [2014] for reviews and discussions. Then, the patients who would undergo RHC if assigned to an R-surgeon but would not undergo RHC if assigned to a NR-surgeon represent the unconfounded unknown fraction of the sample.

Consider the group of patients who underwent RHC. A unit in this group can be either a "complier" or a "non-complier". She's a complier if she would not have undergone RHC if assigned to an NR-surgeon, whereas she's a non-complier if she would have undergone RHC regardless of the type of surgeon or only if assigned to a NR-surgeon. In many instances, these two types will differ in terms of observed covariates $\mathbf{X}$. However, for certain values $\mathbf{x}$ of $\mathbf{X}$, a unit might be either a complier or a non-complier with non-zero probability. In this scenario, our relaxed XA-model posits that the probability of survival conditional on receiving RHC is the same for a complier and a non-complier sharing the same $\mathbf{X} = \mathbf{x}$. Notice that this is not imposing any assumption on what would have happened to the non-complier had she not been treated. In fact, we derived the lower (upper) bound on the average effect of RHC by assuming that she would have certainly survived (died) had she not undergone RHC. This maximal conservativeness in deriving the bounds likely protects our conclusions from mild violations of our X- and XA-models.

To construct the curves tracing the bounds using the data, we estimate the nuisance regression functions via the cross-validation-based SuperLearner ensemble [Van der Laan et al., 2007], combining generalized additive models, random forests, splines, support vector machines as well as generalized linear models. We perform 5-fold cross-fitting. We also construct pointwise and uniform confidence bands. Results are reported in Figure 2.2.

In line with the results in Connors et al. [1996], if no-unmeasured-confounding holds, patients treated with RHC show a statistically significant decrease in 30-day survival rates. The risk difference equals -3.74% (95% CI $= [-6.00\%, -1.49\%]$). Under the X-mixture model, the bounds on the difference in survival rate would include zero if more than 4.89% (95% CI $= [1.50\%, 8.28\%]$) of the patients were confounded. The value reduces to 4.02% (95% CI $= [1.59\%, 6.45\%]$) under the relaxed XA-mixture model. Whether robustness to 5% of potentially confounded units is enough to attach a causal interpretation to the study's result largely depends on subject-matter knowledge. Earlier we have described ϵ_0 as the proportion of "non-compliers," but other interpretations are also possible. For instance, suppose it is known that, before deciding whether a patient undergoes RHC, most surgeons look at lab value v_1, but some may check lab value v_2 as well. Both values are correlated with survival, but only v_1 is measured. If reviewers of the study have an idea of how common it is for surgeons to check v_2 in addition to v_1, then they would be able to decide whether $\widehat{\epsilon} = 5\%$ is large or small. In the supplementary material, we consider varying δ, the parameter governing the severity of the unmeasured confounding. For instance, if $\delta = 0.5$ is thought to be reasonable, robustness would increase to 11.00% (95% CI $= [3.84\%, 18.16\%]$) under the X-mixture model.

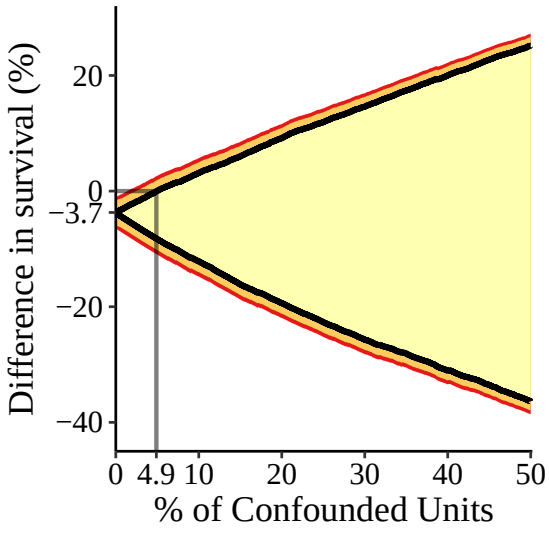
Finally, we refer the readers to Lin et al. [1998] and Altonji et al. [2008], among others, for additional sensitivity analyses applied to this dataset. In particular, in the context of Cox proportional hazard regression, and under certain simplifying assumptions, Lin et al. [1998]

derive that a confidence interval for the relative hazard of death would include 1 as long as the prevalence of a binary unmeasured confounder is at least 10% greater in the group that underwent RHC than in the control group. Using a probit model of mortality at day 90, Altonji et al. [2008] show that the observed positive association between mortality and RHC usage could be "explained away" if the correlation between the unmeasured factors determining RHC usage and mortality is approximately 0.15. In addition, in Section A.6.1 of the supplementary material, we apply the sensitivity analysis designed for linear models proposed in Cinelli and Hazlett [2020]. We find that an unmeasured confounder that explains 4.2+% of the variance in mortality not captured by RHC usage and the measured covariates and 4.2+% of the variance in RHC usage not captured by the measured covariates would be sufficient to drive the observed effect (≈ -0.04) to zero. Notice that these approaches are designed for specific models used in the primary analysis, whereas our framework is agnostic regarding modeling choices. Further, they assume that the treatment-outcome association may be confounded for every unit, while our sensitivity model captures departures from such homogeneity by allowing the treatment-outcome association to be unconfounded for an unknown subgroup of units.

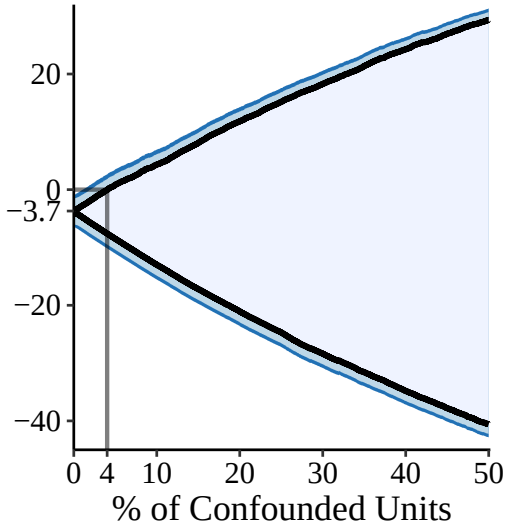
2.5 Discussion

In this paper, we propose a novel approach to sensitivity analysis in observational studies where the sensitivity parameter is the proportion of unmeasured confounding. A strength of our model is that it captures a rich form of unmeasured confounding heterogeneity. While even richer models may allow for a more flexible characterization of confounding heterogeneity, we believe our approach strikes a nice balance between complexity and transparency. In fact, it captures heterogeneity with just one, intuitive sensitivity parameter: an unknown fraction ϵ of the units can be arbitrarily confounded while the rest are not. The model is general enough to cover some relaxations to the no-unmeasured-confounding assumption already proposed in the literature. As ϵ is varied, lower and upper bounds on the ATE are derived under certain assumptions on the distribution of the confounded units. The parameter ϵ is interpretable and yields a natural one-number summary of a study's robustness to unmeasured confounding, namely the minimal proportion of confounding such that the bounds on the ATE contain zero. We provide sufficient conditions to construct both pointwise and uniform confidence bands around the curves tracing the lower and upper bounds on the ATE as a function of ϵ. We also describe the asymptotic normality of a Z-estimator of ϵ_0; we propose reporting an estimate of ϵ_0 together with a Wald-type confidence interval when discussing results from an observational study.

Several questions remain unanswered and could be the subject of future research. First, bounding the ATE under no restrictions on the distribution of the confounded units is currently computationally intractable. Therefore, the discovery of a clever way to compute the bounds in this setting would generalize the current version of our model. Second, generalizing the approach of Imbens and Manski [2004] to construct uniform confidence bands trapping the true ATE ψ, rather than the identification region $[\psi_l(\epsilon), \psi_u(\epsilon)]$, would allow far more precise



(a) X-mixture model ($S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$)



(b) XA-mixture model ($S \perp\!\!\!\perp Y \mid (A, \mathbf{X})$)

Figure 2.2: Estimated bounds on the Average Treatment Effect as a function of the proportion of confounded units ϵ assuming "worst-case" $\delta = 1$, with pointwise [Imbens and Manski, 2004] and uniform 95% confidence bands. Curves under the X-mixture model and under the XA-mixture model are shown along with estimates of ϵ_0 on the abscissa.

inference. Lastly, extensions to our model other than the one considered in Appendix A.4 would likely lead to a richer set of sensitivity models, ultimately allowing the user to gauge the effects of departures from the no-unmeasured-confounding assumption in more nuanced ways. For example, it would be interesting to extend our sensitivity model to accommodate

time-varying or continuous exposures, as well as to explore the possibility of tighter bounds by employing specific sensitivity analysis models to the confounded fraction of the sample.

2.6 Acknowledgments

The authors thank Sivaraman Balakrishnan, Colin Fogarty, Marshall Joffe, Alan Mishler, Pratik Patil and members of the Causal Group at Carnegie Mellon University for helpful discussions. Edward Kennedy gratefully acknowledges financial support from NSF Grant DMS1810979.

Chapter 3

Sensitivity analysis for marginal structural models

This chapter is taken from my work supervised by Larry Wasserman, Valérie Ventura and Edward H. Kennedy, which can be found on arXiv [Bonvini et al., 2022a].

3.1 Introduction

Marginal structural models (MSMs) [Robins, 1998, 2000, Robins et al., 2000] are a class of semiparametric model commonly used for causal inference. As is typical in causal inference, the parameters of the model are only identified under an assumption of no unmeasured confounding. Thus, it is important to quantify how sensitive the inferences are to this assumption. Most existing sensitivity analysis methods deal with binary point treatments. In contrast, in this paper we develop tools for assessing sensitivity for MSMs with both continuous (non-binary) and time-varying treatments.

For simplicity, consider the static treatment setting first. Extensions to time-varying treatments are described in Section 3.6. Suppose we have n iid observations $(Z_1, \ldots, Z_n)$, with $Z_i = (X_i, A_i, Y_i)$ from a distribution $\mathbb{P}$, where $Y \in \mathbb{R}$ is the outcome of interest, $A \in \mathbb{R}$ is a treatment (or exposure) and $X \in \mathbb{R}^d$ is a vector of confounding variables. Define the collection of counterfactual random variables (also called potential outcomes) $\{Y(a) : \ a \in \mathbb{R}\}$, where $Y(a)$ denotes the value that Y would have if A were set to a. The usual assumptions in causal inference are:

(A1) No interference: if $A = a$ then $Y = Y(a)$, meaning that a subject's potential outcomes only depend on their own treatment.

(A2) Overlap: $\pi(a|x) > 0$ for all x and a, where $\pi(a|x)$ is the density of A given $X = x$ (the *propensity score*). Overlap guarantees that all subjects have some chance of receiving each treatment level.

(A3) No unmeasured confounding: the counterfactuals $\{Y(a) : a \in \mathbb{R}\}$ are independent of A given the observed covariates X. This assumption means that the treatment is as good as randomized within levels of the measured covariates; in other words, there are no unmeasured variables U that affect both A and Y.

Under these assumptions, the causal mean $\mathbb{E}\{Y(a)\}$ is identified and equal to

$$\psi(a) \equiv \int \mu(x,a) d\mathbb{P}(x), \tag{3.1}$$

where $\mu(x,a) = \mathbb{E}[Y|X=x, A=a]$ is the outcome regression (causal parameters other than $\mathbb{E}\{Y(a)\}$, e.g., cumulative distribution functions, are identified similarly). Equation (3.1) is a special case of the g-formula [Robins, 1986].

A marginal structural model (MSM) is a semiparametric model assuming $\psi(a) = g(a;\beta)$ [Robins, 1998, 2000, Robins et al., 2000]. The MSM provides an interpretable model for the treatment effect and β can be estimated using simple estimating equations. The model is semiparametric in the sense that it leaves the data generating distribution unspecified except for the restriction that $\int \mu(x,a) d\mathbb{P}(x) = g(a;\beta)$. If g is mis-specified, one can regard $g(a;\beta)$ as an approximation to $\psi(a)$, in which case one estimates the value β_* that minimizes $\int (\psi(a) - g(a;\beta))^2 \omega(a) da$, where ω is a user provided weight function [Neugebauer and van der Laan, 2007].

In practice, there are often unmeasured confounders U so that assumption (A3) fails. This is especially true for observational studies where treatment is not under investigators' control, but it can also occur in experiments in the presence of non-compliance. In these cases, $\mathbb{E}\{Y(a)\}$ is no longer identified. We can still estimate the functional $\psi(a)$ in (3.1) but we no longer have $\mathbb{E}\{Y(a)\} = \psi(a)$. Sensitivity analysis methods aim to assess how much $\mathbb{E}\{Y(a)\}$ and the MSM parameter β will change when such unmeasured confounders U exist. In this paper, we will derive bounds for $\mathbb{E}\{Y(a)\} \equiv g(a;\beta)$, as well as for β, under varying amounts of unmeasured confounding.

We consider several sensitivity models for unmeasured confounding: a propensity-based model, an outcome-based model, and a subset confounding model, in which only a fraction of the population is subject to unmeasured confounding.

3.1.1 Related Work

Sensitivity analysis for causal inference began with Cornfield et al. [1959]. Theory and methods for sensitivity analysis were greatly expanded by Rosenbaum [1995]. Recently, there has been a flurry of interest in sensitivity analysis including Chernozhukov et al. [2021], Kallus et al. [2019], Scharfstein et al. [2021], Yadlowsky et al. [2018], Zhao et al. [2019], among others. We refer to Section 2 of Scharfstein et al. [2021] for a review. Most work deals with binary, static treatments.

The closest work to ours is Brumback et al. [2004], who study sensitivity for MSMs with binary treatments using parametric models for the sensitivity analysis. We instead consider

nonparametric sensitivity models, for continuous rather than binary treatments. While completing this paper, Dorn and Guo [2021] appeared on arXiv, who independently derived bounds on treatment effects for nonparametric causal models that are similar to our bounds in Section 3.4.1, Lemma 2. Here we treat MSMs rather than nonparametric causal models, with Lemma 2 being an intermediate step to our results.

3.1.2 Outline

We first treat the static treatment setting. In Section 3.2 we review MSMs. In Section 6.6.2 we introduce our three sensitivity analysis models. We find bounds for the MSM $g(a;\beta)$ and for its parameter β under propensity sensitivity in Section 3.4, under outcome sensitivity in Section 3.5 and under subset sensitivity in Appendix B.0.2. Then in Section 3.6, we extend our methods to the time series setting. We illustrate our methods on simulated data in Appendix B.0.1 and on observational data in Section 3.7. Section 6.7 contains concluding remarks. All proofs can be found in the Appendix.

3.1.3 Notation

We use the notation $\mathbb{P}[f(Z)] = \int f(z)d\mathbb{P}(z)$ and $\mathbb{U}[f(Z_1,Z_2)] = \int f(z_1,z_2)d\mathbb{P}(z_1,z_2)$ to denote expectations of a fixed function, and $\mathbb{P}_n[f(Z)] = n^{-1}\sum_{i=1}^n f(Z_i)$ and $\mathbb{U}_n[f(Z_1,Z_2)] = \{n(n-1)\}^{-1}\sum_{1\leq i\neq j\leq n}^n f(Z_i,Z_j)$ to denote their sample counterparts, where $\mathbb{U}_n$ is the usual U-statistic measure. We also let $\|f\|^2 = \int f^2(z)d\mathbb{P}(z)$ denote the $L^2(\mathbb{P})$ norm of f and $\|f\|_\infty = \sup_z |f(z)|$ denote the L^∞ or sup-norm of f. For $\beta \in \mathbb{R}^k$ we let $\|\beta\|$ denote the Euclidean norm. For $f(z_1,z_2)$ we let $S_2[f] = \{f(z_1,z_2)+f(z_2,z_1)\}/2$ be the symmetrizing function. Then $\mathbb{U}_n[f(Z_1,Z_2)] = \mathbb{U}_n[S_2[f(Z_1,Z_2)]]$.

3.1.4 Some Inferential Issues

Here we briefly discuss three issues that commonly arise in this paper when constructing confidence intervals.

The first is that we often have to estimate quantities of the form

$$\nu = \int\int f(x,a)\pi(a)da d\mathbb{P}(x)$$

where $\pi(a)$ is the marginal density of A. This is not a usual expected value since the integral is with respect to a product of marginals, $\pi(a)d\mathbb{P}(x)$, rather than the joint measure $\mathbb{P}(x,a)$. Then ν can be written as

$$\begin{aligned}\mathbb{U}[f(Z_1,Z_2)] &\equiv \int\int \frac{1}{2}\left[f(x_1,a_2)+f(x_2,a_1)\right] d\mathbb{P}(x_1,a_1)d\mathbb{P}(x_2,a_2)\\ &= \int\int g(z_1,z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2)\end{aligned}$$

where $Z_1 = (X_1,A_1,Y_1)$ and $Z_2 = (X_2,A_2,Y_2)$ are two independent draws and $g(z_1,z_2) =$

$S_2[f] \equiv (f(x_1, a_2) + f(x_2, a_1))/2$. Under certain conditions, the limiting distribution of $\sqrt{n}\{\mathbb{U}_n[\widehat{f}(Z_1, Z_2)] - \mathbb{U}[f(Z_1, Z_2)]\}$, where $\widehat{f}$ is an estimate of f, is the same as that of $\sqrt{n}(\mathbb{U}_n - \mathbb{U})[f(Z_1, Z_2)]$. More specifically, let $\alpha \in \mathbb{R}^k$, where k is the dimension of f. By Theorem 12.3 in Van der Vaart [2000],

$$\sqrt{n}(\mathbb{U}_n - \mathbb{U})[\alpha^T f(Z_1, Z_2)] \to N(0, 4\sigma^2),$$

where $\sigma^2 = \frac{1}{4}\alpha^T \Sigma \alpha$ and $\Sigma = \text{var}\left[\int S_2[f(Z_1, z_2)] d\mathbb{P}(z_2)\right]$. Therefore, by the Cramer-Wold device, $\sqrt{n}(\mathbb{U}_n - \mathbb{U})[f(Z_1, Z_2)] \rightsquigarrow N(0, \Sigma)$. Thus, $\sqrt{n}(\mathbb{U}_n - \mathbb{U})[S_2[f(Z_1, Z_2)]]$ has variance equal to the variance of the influence function of $\nu = \int\int f(x, a)\pi(a) da d\mathbb{P}(x)$ and thus it is efficient.

The second issue is that calculating the variances of these estimators can be cumbersome. Instead, we construct confidence intervals using the HulC [Kuchibhotla et al., 2021], which avoids estimating variances. The dataset is randomly split into $B = \log(2/\alpha)/\log 2$ subsamples ($B = 6$ when $\alpha = 5\%$) and the estimators are computed in each subsample. Then, the minimum (maximum) of the six estimates is returned as the lower (upper) end of the confidence interval.

The third issue is that many of our estimators depend on nuisance functions such as the outcome model $\mu(a, x)$ and the conditional density $\pi(a|x)$. To avoid imposing restrictions on the complexity of the nuisance function classes, we analyze estimators based on cross-fitting. That is, unless otherwise stated, the nuisance functions are assumed to be estimated from a different sample than the sample used to compute the estimator. Such construction can always be achieved by splitting the sample into k folds; using all but one fold for training the nuisance functions and the remaining fold to compute the estimator. Then, the roles of the folds can be swapped, thus yielding k estimates that are averaged to obtain a single estimate of the parameter. For simplicity, we will use $k = 2$, but our analysis can be easily extended to the case where multiple splits are performed.

3.2 Marginal Structural Models

In this section we review basic terminology and notation for marginal structural models. We focus for now on studies with one time point; we deal with time varying cases in Section 3.6. More detailed reviews can be found in Robins and Hernán [2009] and Hernán and Robins [2010]. Let

$$\mathbb{E}\{Y(a)\} \equiv \psi(a) = g(a; \beta), \quad \beta \in \mathbb{R}^k, \tag{3.2}$$

be a model for the expected outcome under treatment regime $A = a$. An example is the linear model $g(a; \beta) = b^T(a)\beta$ for some specified vector of basis functions $b(a) = [b_1(a), \ldots, b_k(a)]$. It can be shown that β in (3.2) satisfies the k-dimensional system of equations

$$\mathbb{E}\left[h(A)w(A, X)\{Y - g(A; \beta)\}\right] = 0 \tag{3.3}$$

for any vector of functions $h(a) = [h_1(a), \ldots, h_k(a)]$, where $w(a, x)$ can be taken to be either $1/\pi(a|x)$ or $\pi(a)/\pi(a|x)$, and $\pi(a)$ is the marginal density of the treatment A. The latter

weights are called stabilized weights and can lead to less variable estimators of β. We will use them throughout. The parameter β can be estimated by solving the empirical analog of (3.3), leading to the estimating equations

$$\mathbb{P}_n \left[h(A)\widehat{w}(A,X)\{Y - g(A;\beta)\} \right] = 0, \tag{3.4}$$

where $\widehat{w}(a,x) = \widehat{\pi}(a)/\widehat{\pi}(a|x)$, and $\widehat{\pi}(a|x)$ and $\widehat{\pi}(a)$ are estimates of $\pi(a|x)$ and $\pi(a)$. Under regularity conditions, including the correct specification of $\pi(a|x)$, confidence intervals based on $\sqrt{n}(\widehat{\beta} - \beta) \rightsquigarrow N(0, \sigma^2)$, where $\sigma^2 = M^{-1}\text{var}[h(A)w(A,X)\{Y - g(A;\beta)\}]M^{-1}$ and $M = \mathbb{E}\{h(A)\nabla_\beta g(A;\beta)^T\}$, will be conservative.

Under model (3.2), every choice of $h(a)$ leads to a $\sqrt{n}$-consistent, asymptotically Normal estimator of β, though different choices lead to different standard errors. If the MSM is linear, i.e. $g(a;\beta) = b(a)^T\beta$, a common choice of $h(a)$ is $h(a) = b(a)$. In this case, the solution to the estimating equation (3.4) can be obtained by weighted regression, $\widehat{\beta} = (B^T\mathbb{W}B)^{-1}B^T\mathbb{W}Y$, where B is the $n \times k$ matrix with elements $B_{ij} = b_j(A_i)$, $\mathbb{W}$ is diagonal with elements $\widehat{W}_i \equiv \widehat{w}(A_i, X_i)$ and $Y = (Y_1, \ldots, Y_n)$.

3.3 Sensitivity Models

We now describe three models for representing unmeasured confounding when treatments are continuous. Each model defines a class of distributions for (U, X, A, Y) where U represents unobserved confounders. Our goal is to find bounds on causal quantities, such as β or $g(a;\beta)$, as the distribution varies over these classes.

3.3.1 Propensity Sensitivity Model

In the case of binary treatments $A \in \{0,1\}$, a commonly used sensitivity model [Rosenbaum, 1995] is the odds ratio model

$$\Gamma(\gamma) = \left\{ \pi(a|x,u) : \frac{1}{\gamma} \leq \frac{\pi(1|x,u)}{\pi(0|x,u)} \frac{\pi(0|x,\widetilde{u})}{\pi(1|x,\widetilde{u})} \leq \gamma \text{ for all } u, \widetilde{u}, x \right\}$$

for $\gamma \geq 1$. When A is continuous, it is arguably more natural to work with density ratios, and so we define

$$\Pi(\gamma) = \left\{ \pi(a|x,u) : \frac{1}{\gamma} \leq \frac{\pi(a|x,u)}{\pi(a|x)} \leq \gamma, \int \pi(a|x,u)da = 1, \text{ for all } a, x, u \right\}. \tag{3.5}$$

We can think of $\Pi(\gamma)$ as defining a neighborhood around $\pi(a|x)$. This is related to the class in Tan [2006] but we consider density ratios rather than odds ratios. There are other constraints possible, such as $\int \pi(a|x,u)d\mathbb{P}(u|x) = \pi(a|x)$; we leave enforcing these additional constraints, which can yield more precise bounds, for future work.

3.3.2 Outcome Sensitivity Model

For an outcome-based sensitivity model, we define a neighborhood around $\mu(x,a)$ given by

$$\mathcal{M}(\delta) = \left\{ \mu(u,x,a) : \ |\Delta(a)| \leq \delta, \ \ \Delta(a) = \int [\mu(u,x,a) - \mu(x,a)] d\mathbb{P}(x,u) \right\},$$

which is the set of unobserved outcome regressions (on measured covariates, treatment, and unmeasured confounders) such that differences between unobserved and observed regressions differ by at most δ after averaging over measured and unmeasured covariates. We immediately have the simple nonparametric bound $\mathbb{E}\{\mu(a,X)\} - \delta \leq \mathbb{E}\{Y(a)\} \leq \mathbb{E}\{\mu(a,X)\} + \delta$. For a given $\Delta(a)$, is a known function, nonparametric bounds can be computed by regressing an estimate of $\Delta(A) + w(A,X)\{Y - \mu(A,X)\} + \int \mu(A,x) d\mathbb{P}(x)$ on A (see, e.g. Kennedy et al. [2017], Semenova and Chernozhukov [2021], Foster and Syrgkanis [2019], Bonvini and Kennedy [2022]). However, our main goal is not to bound $\mathbb{E}\{Y(a)\}$, but bound the parameters β of the MSM or the MSM itself. Finding these bounds under outcome sensitivity will require specifying an outcome model. For the propensity sensitivity model, we will also need an outcome model if we want doubly robust estimators of β.

3.3.3 Subset Confounding

Bonvini and Kennedy [2020] consider a model where only an unknown fraction ϵ of the population is subject to unobserved confounding. Specifically, suppose there exists a latent binary variable S such that $P(S=0) = \epsilon$ as well as $Y(a) \perp\!\!\!\perp A | X, S = 1$ and $Y(a) \perp\!\!\!\perp A | X, U, S = 0$. It follows that $P = (1-\epsilon)P_1 + \epsilon P_0$ where P_j is the distribution of (U, X, A, Y) given $S = j$. For the $S = 0$ group of units, we will control the extent of unmeasured confounding using either the outcome model or the propensity sensitivity model. This can be regarded as a type of contamination model.

Results under the propensity and outcome sensitivity confounding models are in the next two sections. Due to space restrictions, the results on subset confounding are in the appendix.

3.4 Bounds under the Propensity Sensitivity Model

3.4.1 Preliminaries

In this section, we develop preliminary results needed to derive bounds under the propensity sensitivity model. A preliminary step in deriving bounds for the MSM is to first bound $\mathbb{E}\{Y(a)|X\}$ and it may be verified that $\mathbb{E}\{Y(a)|X\} = m(a,X)$ where $m(A,X) = \mathbb{E}\{Yv(Z)|A,X\}$ and

$$v(Z) \equiv \mathbb{E}\left\{ \frac{\pi(A|X)}{\pi(A|X,U)} \,\middle|\, A, X, Y \right\} \in \left[\gamma^{-1}, \gamma\right].$$

It is easy to see that $\mathbb{E}\{v(Z)|A,X\}=1$. So bounding $\mathbb{E}\{Y(a)|X\}$ is equivalent to bounding $m(a,X)=\mathbb{E}\{Yv(Z)|A,X\}$ as v varies over the set

$$\mathcal{V}(\gamma)=\left\{v(\cdot):\ \gamma^{-1}\leq v(z)\leq\gamma,\ \mathbb{E}\{v(Z)|X=x,A=a\}=1 \text{ for all } x,a\right\}. \tag{3.6}$$

Proposition 1. The following moment condition holds for the MSM:

$$\mathbb{E}\left\{h(A)\Big[\int m(A,x)d\mathbb{P}(x)-g(A;\beta)\Big]\right\}=\mathbb{U}\left[h(A_1)\{m(A_1,X_2)-g(A_1;\beta)\}\right]=0, \tag{3.7}$$

where (X_1,A_1) and (X_2,A_2) are two independent draws (see Section 3.1.4).

Notice that if $U=\emptyset$, then $v(Z)=1$ and $m(a,x)=\mu(a,x)=\mathbb{E}\{Y|A=a,X=x\}$. However, when there is residual unmeasured confounding, $m(a,x)$ does not equal $\mathbb{E}(Y|A=a,X=x)$ and in general cannot be identified. However, it can still be bounded under the propensity sensitivity model, as in the following lemma.

Lemma 2. *For $j\in\{\ell,u\}$ (corresponding to lower and upper bound) let $q_j(Y|A,X)$ denote the τ_j-quantile of Y given (A,X), where $\tau_\ell=1/(1+\gamma)$ and $\tau_u=\gamma/(1+\gamma)$. Define*

$$v_\ell(Z)=\gamma^{\text{sgn}\{q_\ell(Y|A,X)-Y\}} \quad \textit{and } v_u(Z)=\gamma^{\text{sgn}\{Y-q_u(Y|A,X)\}}.$$

Then $m_\ell(a,x)\leq m(a,x)\leq m_u(a,x)$, where $m_j(a,x)=\mathbb{E}\{Yv_j(Z)|A=a,X=x\}$, $j\in\{u,\ell\}$.

Now that we have bounds on $m(a,x)$, we turn to finding bounds on the MSM $g(a;\beta)$ and on its parameter β. We will use the notation $c_\ell=\gamma^{-1}$, $c_u=\gamma$, $S_j\equiv s(Z;q_j)=q_j(Y|A,X)+\{Y-q_j(Y|A,X)\}c_j^{\text{sgn}\{Y-q_j(Y|A,X)\}}$, $\kappa_j\equiv\kappa(A,X;q_j)=\mathbb{E}\{S_j|A,X\}$ and

$$\varphi_j(Z_1,Z_2)\equiv\varphi_j(Z_1,Z_2;w,q_j,\kappa_j)=w(A_1,X_1)\{s(Z_1;q_j)-\kappa(A_1,X_1;q_j)\}+\kappa(A_1,X_2;q_j). \tag{3.8}$$

Notice that

$$\mathbb{U}\{\kappa(A_1,X_2;q_j)\}=\int\int m_j(a,x)d\mathbb{P}(a)d\mathbb{P}(x), \tag{3.9}$$

since $\mathbb{E}\left[c_j^{\text{sgn}\{Y-q_j(Y|A,X)\}}|A,X\right]=1$.

3.4.2 Bounds on $g(a;\beta)$

Under the MSM $\mathbb{E}\{Y(a)\}=g(a;\beta)$, given the discussion in Section 3.4.1, we have that $\mathbb{E}\{Y(a)\}=\mathbb{E}\{m(a,X)\}$ if $Y(a)\perp\!\!\!\perp A|(X,U)$. This implies that $\mathbb{E}\{m_\ell(a,X)\}\leq g(a;\beta)\leq\mathbb{E}\{m_u(a,X)\}$, where m_ℓ and m_u are defined in Lemma 2. Thus, a straightforward way to bound $g(a;\beta)$ is to assume that the bounds follow a model similar to the model we assume under no unmeasured confounding, when $\mathbb{E}\{Y(a)\}=g(a;\beta)$ is identified. That is, we let $\mathbb{E}\{m_j(a,X)\}=g(a;\beta_j)$, $j\in\{u,\ell\}$, and estimate β_j by solving the empirical analog of the

moment condition:

$$\mathbb{E}\left\{h(A)\Big[\int m_j(A,x)d\mathbb{P}(x) - g(A;\beta_j)\Big]\right\} = 0, \quad j \in \{u,\ell\}. \tag{3.10}$$

Using (3.9) and the fact that the first term in (3.8) has conditional mean 0, we also have that $\mathbb{U}\left[h(A_1)\left\{\varphi_j(Z_1,Z_2) - g(A_1;\beta_j)\right\}\right] = 0$. Given an estimate of the function $\varphi_j(Z_1,Z_2)$ in (3.8), estimated from an independent sample, we estimate β_j by solving

$$\mathbb{U}_n\left[h(A_1)\left\{\widehat{\varphi}_j(Z_1,Z_2) - g(A_1;\widehat{\beta}_j)\right\}\right] = 0. \tag{3.11}$$

The following proposition provides the asymptotic distributions of $g(a;\widehat{\beta}_j)$, $j \in \{u,\ell\}$.

Proposition 2. Suppose the following conditions hold:

1. The function class $\mathcal{G}_l = \{a \mapsto h_l(a)g(a;\beta)\}$ is Donsker for every $l = \{1,\ldots,k\}$ with integrable envelop and $g(a;\beta)$ is a continuous function of β.
2. For $j \in \{\ell,u\}$, the map $\beta \mapsto \mathbb{U}\{h(A)[\varphi_j(Z_1,Z_2) - g(A_1;\beta)]$ is differentiable at all β with continuously invertible matrices $\dot{\Psi}_{\beta_0}$ and $\dot{\Psi}_{\widehat{\beta}}$, where $\dot{\Psi}_\beta = -\mathbb{E}\{h(A)\nabla^T g(a;\beta)\}$;
3. $\left\|\int S_2\left\{\widehat{\varphi}_j(Z_1,z_2) - \varphi_j(Z_1,z_2)\right\}d\mathbb{P}(z_2)\right\| = o_{\mathbb{P}}(1)$;
4. $\|w - \widehat{w}\|\|\kappa_j - \widehat{\kappa}_j\| + \|q_j - \widehat{q}_j\|^2 = o_{\mathbb{P}}(n^{-1/2})$, where φ_j, κ_j and q_j are defined in Section 3.4.1.

Then

$$\sqrt{n}(\widehat{\beta}_j - \beta_j) \rightsquigarrow N\left(0, 4\text{var}\{\dot{\Psi}_{\beta_j}^{-1}\phi_j(T;\beta_j)\}\right),$$

where $\phi_j(Z_1;\beta_j) = \int S_2 h(A_1)\left\{\varphi_j(Z_1,z_2) - g(A_1;\beta_j)\right\}d\mathbb{P}(z_2)$, and it follows that

$$\sqrt{n}\{g(a;\widehat{\beta}_j) - g(a;\beta_j)\} \rightsquigarrow N\left(0, 4\nabla g(a;\beta_j)^T\text{var}\{\dot{\Psi}_{\beta_j}^{-1}\phi_j(Z;\beta_j)\}\nabla g(a;\beta_j)\right), \quad j \in \{u,\ell\}.$$

The main requirement, in condition (d), to achieve asymptotic normality is that certain products of errors for estimating the nuisance functions are $o_{\mathbb{P}}(n^{-1/2})$. This can be achieved even if these functions are estimated at nonparametric rates, e.g. $n^{-1/4}$, under structural constraints such as smoothness or sparsity. We note that, strictly speaking, our estimator is not doubly robust since one needs to consistently estimate q_j for consistency. However, the dependence on the estimation error in $\widehat{q}_j$ is still second-order, in that it depends on the squared error.

3.4.3 Bounds on $g(a;\beta)$ when $g(a;\beta)$ is linear

When the MSM is linear, it is straightforward to bound $g(a;\beta) = b(a)^T\beta$ directly, without assuming that the bounds themselves follow parametric models $g(a;\beta_j)$. Let $h(A) = b(A)$ and

$Q = \mathbb{E}\{b(A)b(A)^T\}$. Then we can re-write $g(a;\beta) = b(a)^T Q^{-1}\mathbb{U}\left\{b(A_1)m(A_1,X_2)\right\}$. Let $\lambda^-(a,A) = \mathbb{1}\{b(a)^T Q^{-1} b(A) \leq 0\}$ and $\lambda^+(a,A) = \mathbb{1}\{b(a)^T Q^{-1} b(A) \geq 0\}$. Further define

$$g_j^s(a) = \mathbb{U}\{b(A_1)\lambda^s(a,A_1)\kappa(A_1,X_2;q_j)\}$$

for $s = \{-,+\}$ and $j = \{\ell, u\}$. Bounds on $g(a;\beta) = b(a)^T\beta$ are $g_\ell(a) \leq g(a;\beta) \leq g_u(a)$ where $g_\ell(a) = b(a)^T Q^{-1}\{g_\ell^+(a) + g_u^-(a)\}$ and $g_u(a) = b(a)^T Q^{-1}\{g_\ell^-(a) + g_u^+(a)\}$. That is, depending on the sign of $b(a)^T Q^{-1} b(A)$, we set $m(A,x) = m_\ell(A,x)$ or $m(A,x) = m_u(A,x)$. Let

$$f_j^s(Z_1,Z_2) = \lambda^s(A_1)\varphi_j(Z_1,Z_2).$$

We analyze the performance of estimators that construct $\widehat{f}_j^s(Z_1,Z_2)$ from a separate, independent sample and output $\widehat{g}_j(a_0) = b(a_0)^T\widehat{\beta}_j$, where

$$\widehat{\beta}_\ell = \underset{\beta\in\mathbb{R}^k}{\operatorname{argmin}}\,\mathbb{U}_n\left\{\widehat{f}_u^-(Z_1,Z_2) + \widehat{f}_\ell^+(Z_1,Z_2) - b(A_1)^T\beta\right\}^2$$
$$\widehat{\beta}_u = \underset{\beta\in\mathbb{R}^k}{\operatorname{argmin}}\,\mathbb{U}_n\left\{\widehat{f}_\ell^-(Z_1,Z_2) + \widehat{f}_u^+(Z_1,Z_2) - b(A_1)^T\beta\right\}^2.$$

The following proposition gives the limiting distribution of the estimated upper and lower bounds $\widehat{g}_j(a)$ for $g(a;\beta)$, $j \in \{u,\ell\}$.

Proposition 3. Suppose the following conditions hold:

1. $\left\|\int S_2\{\widehat{f}_j^s(Z_1,z_2) - f_j^s(Z_1,z_2)\}d\mathbb{P}(z_2)\right\| = o_{\mathbb{P}}(1)$;
2. $\|\widehat{q}_j - q_j\|^2 + \|\widehat{w} - w\|\|\widehat{\kappa}_j - \kappa_j\| = o_{\mathbb{P}}(n^{-1/2})$;
3. The density of $b(a)^T Q^{-1} b(A)$ is bounded.

Then

$$\sqrt{n}\{\widehat{g}_j(a) - g_j(a)\}$$
$$\rightsquigarrow N\left(0, 4b(a)^T \text{var}\left[Q^{-1}\int S_2 b(A_1)\{f_j^s(Z_1,z_2) - b^T(A_1)\beta_j\}d\mathbb{P}(z_2)\right]b(a)\right).$$

Another approach for getting bounds on $g(a;\beta)$ is to note that

$$\delta g(a;\beta)/\delta v = \sum_j b_j(a)\delta\beta_j/\delta v,$$

where δ is the functional derivative, and then apply the homotopy algorithm from Section 3.4.4.

3.4.4 Bounds on β

We now turn to finding approximate bounds on components of β rather than on $g(a;\beta)$. Suppose, to be concrete, that we want to upper bound β_1. (Lower bounds can be found similarly.) At this point, we re-name $\mathcal{V}(\gamma)$ in (3.6) as $\mathcal{V}_{\text{small}}(\gamma)$ and we define

$$\mathcal{V}_{\text{large}}(\gamma) = \left\{ v(\cdot) : \ \gamma^{-1} \le v(z) \le \gamma, \ \mathbb{E}[v(Z)] = 1 \right\}.$$

Bounds over $\mathcal{V}_{\text{large}}(\gamma)$ are conservative but, as we shall see, are easier to compute. Next we define two functionals. Let $F_1(v)$ be the value of b that solves

$$\int yh(a)w(a,x)v(z)\mathbb{P}(z) = \int h(a)w(a,x)g(a;b)v(z)\mathbb{P}(z)$$

and $F_2(v)$ be the value of b that solves

$$\int yh(a)w(a,x)v(z)\mathbb{P}(z) = \int h(a)w(a,x)g(a;b)\mathbb{P}(z).$$

At the true value v_* we have $\beta_{1*} = e^T F_1(v_*) = e^T F_2(v_*)$ where $e = (1,0,\ldots,0)$ and β_{1*} is the true value of β_1. But $F_1(v) \neq F_2(v)$ in general, and bounding $F_1(v)$ and $F_2(v)$ both lead to valid bounds for β_1. A quick summary of what will follow is this:

i. For $\mathcal{V}_{\text{small}}(\gamma)$, bounds based on F_1 and F_2 are equal, as stated in Lemma 3. These bounds require quantile regression.

ii. For $\mathcal{V}_{\text{large}}(\gamma)$, bounds based on F_1 and F_2 are different so we take their intersection. These bounds do not require quantile regression. In our experience, bounds based on F_1 are often tighter.

Lemma 3. *We have*

$$\inf_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_1(v) = \inf_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_2(v)$$
$$\sup_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_1(v) = \sup_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_2(v).$$

For $\mathcal{V}_{\text{large}}(\gamma)$, the bounds may differ.

We want to find v_γ such that $e^T F_k(v_\gamma) = \sup_{v\in\mathcal{V}} e^T F_k(v)$, for $k \in \{1,2\}$ and $\mathcal{V} \in \{\mathcal{V}_{\text{small}}, \mathcal{V}_{\text{large}}\}$.

Unless the MSM $g(a;\beta) = b^T(a)\beta$ is linear, determining the optimal v_γ is intractable, so we find an approximate bound. For example, to optimize over $\mathcal{V}_{\text{large}}(\gamma)$, we proceed as follows:

1. We will find a function v_γ that is a local maximum of $F_k(v)$.
2. We show that v_γ is defined by a fixed point equation $v_\gamma = L(v_\gamma)$.

3. We construct an increasing grid $\{\gamma_1, \gamma_2, \dots, \}$ where $\gamma_1 = 1$ and $\gamma_j = \gamma_{j-1} + \delta$. Then we take $v_{\gamma_j} \approx L(v_{\gamma_{j-1}})$.
4. In the limit, as $\delta \to 0$, this defines a sequence of functions $(v_\gamma : \gamma \geq 1)$ where each v_γ is a local optimizer in $\mathcal{V}_{\text{large}}(\gamma)$.

We refer to this as a *homotopy algorithm.* (An alternative approach based on gradient ascent is described Appendix B.1.2.) To make this precise, we need the functional derivative of $F_k(v)$ with respect to v. First we recall the definition of a functional derivative: if $G(v) \in \mathbb{R}$, we say that $\frac{\delta G(v)}{\delta v}$ is the functional derivative of $G(v)$ with respect to v in $L_2(\mathbb{P})$ if

$$\int \frac{\delta G(v)}{\delta v}(z) f(z) d\mathbb{P}(z) dz = \left[\frac{d}{d\epsilon} G[v + \epsilon f]\right]_{\epsilon=0}$$

for every function f. When $G(v) = (G_1(v), \dots, G_k(v))$ is vector valued, we define $\delta G/\delta v = (\delta G_1(v)/\delta v, \dots, \delta G_k(v)/\delta v)$.

Lemma 4 (Functional derivatives)**.** *We have*

$$\frac{\delta F_1(v)}{\delta v}(z) = \left\{\mathbb{E}\Big[v(Z)h(A)w(A,X)\nabla_\beta g(A;\beta)^T\Big]\right\}^{-1} h(a)(y - g(a;\beta))w(a,x), \quad (3.12)$$

$$\frac{\delta F_2(v)}{\delta v}(z) = \left\{\mathbb{E}\Big[h(A)w(A,X)\nabla_\beta g(A;\beta)^T\Big]\right\}^{-1} h(a)yw(a,x).$$

Notice that, unless $g(a;\beta)$ is linear in β, $\frac{\delta F_2(v)}{\delta v}(z)$ depends on $v(z)$ through $\nabla_\beta g(A;\beta)$ since β is implicitly a function of $v(z)$. We can now find the expression for the local optimizer v_γ from Step (a) above.

Lemma 5. *Suppose that for every v, $(\delta F_k(v)/\delta v)(Z)$ has a continuous distribution. There is a set of functions $(v_\gamma : \gamma \geq 1)$ such that:*

1. $v_\gamma \in \mathcal{V}_{\text{large}}(\gamma)$;

2. v_γ satisfies the fixed point equation

$$v_\gamma(z) = \gamma \mathbb{1}\Big[d_\gamma(z) \geq q_u(d_\gamma)\Big] + \gamma^{-1}\mathbb{1}\Big[d_\gamma(z) < q_u(d_\gamma)\Big] \quad (3.13)$$

where

$$d_\gamma = e^T \left(\frac{\delta F_k(v)}{\delta v}\bigg|_{v=v_\gamma}\right)$$

and $q_u(d_\gamma)$ is the $\tau_u = \gamma/(1+\gamma)$ quantile of $d_\gamma(Z)$. (This is a fixed point equation since d_γ on the right hand side is a function of v_γ.);

3. v_γ is a local maximizer of $e^T F_k(v)$, in the sense that, for all small $\epsilon > 0$, $e^T F_k(v_\gamma) \geq$

$e^T F_k(v) + O(\epsilon^2)$ *for any* $v \in \mathcal{V}_{\text{large}}(\gamma) \bigcap B(v_\gamma, \epsilon)$ *where, for any* v *and any* $\epsilon > 0$ *we define* $B(v, \epsilon) = \{f : \int (f - v)^2 d\mathbb{P}(z) \leq \epsilon^2\}$.

In practice, we compute v_γ sequentially using an increasing sequence of values of γ. Using (3.13) we approximate v_γ by $\gamma \mathbb{1}\{d_{\gamma-\delta}(z) \geq q_u(d_{\gamma-\delta})\} + \gamma^{-1}\mathbb{1}\{d_{\gamma-\delta}(z) < q_u(d_{\gamma-\delta})\}$ where $q_u(d_{\gamma-\delta})$ is the $\tau_u = \gamma/(1+\gamma)$ quantile of $d_{\gamma-\delta}(Z)$ and δ is a small positive number. The sample approximation to the functional derivative for observation i is

$$d_i = \frac{\partial \widehat{F}_1(v)}{\partial V_i} = \left\{ \frac{1}{n} \sum_j h(A_j) V_j \widehat{W}_j \nabla_\beta g(A_j; \widehat{\beta})^T \right\}^{-1} h(A_i) \widehat{W}_i (Y_i - g(A_i; \widehat{\beta})) \tag{3.14}$$

for F_1 and

$$d_i = \frac{\partial \widehat{F}_2(v)}{\partial V_i} = \left\{ \frac{1}{n} \sum_j h(A_j) \widehat{W}_j \nabla_\beta g(A_j; \widehat{\beta})^T \right\}^{-1} h(A_i) \widehat{W}_i Y_i \tag{3.15}$$

for F_2, where $V_i = v(X_i, A_i, Y_i)$. The algorithm is described in Appendix B.1.1. The lower bound on β_1 is obtained the same way, with (3.13) replaced by $v_\gamma(z) = \gamma^{-1}\mathbb{1}\{d_\gamma(z) \geq q_\ell(d_\gamma)\} + \gamma \mathbb{1}\{d_\gamma(z) < q_\ell(d_\gamma)\}$, where $\tau_\ell = 1/(1+\gamma)$. Getting confidence intervals for these bounds is challenging because we need to adjust the estimator with the influence function to make the bias second order, but their influence functions are very complicated; the details are in Appendix B.2.11.

For $\mathcal{V} = \mathcal{V}_{\text{small}}$, which imposes the stronger restriction $\mathbb{E}\{v(Z)|A, X\} = 1$, we replace $q_u(d_\gamma)$ in (3.13) with $q_u(d_\gamma|A, X)$, the conditional quantile of $d_\gamma(z)$ given (X, A). Then

$$d_\gamma(Z) = \mathbb{E}\{h(A)\nabla_\beta g(A; \beta)^T\}^{-1} h(A) w(A, X) Y \equiv T(A, X) Y,$$

so that the τth quantile of $d_\gamma(Z)$ given (A, X) can be expressed as

$$q_\tau(d_\gamma|A, X) = \begin{cases} T(A, X) q_\tau(Y|A, X) & \text{if } T(A, X) < 0, \\ T(A, X) q_{1-\tau}(Y|A, X) & \text{if } T(A, X) > 0, \end{cases}$$

where $q_\tau(Y|A, X)$ is the τth quantile of Y given (A, X). Then, to obtain an upper bound on β_1, $v_\gamma(Z)$ has to satisfy the fixed-point equation:

$$v_\gamma(z) = \mathbb{1}\{e^T T(a, x) \geq 0\} v_u(z) + \mathbb{1}\{e^T T(a, x) < 0\} v_\ell(z),$$

where $v_u(Z) = \gamma^{\text{sgn}\{Y - q_u(Y|A,X)\}}$ and $v_\ell(Z) = \gamma^{\text{sgn}\{q_\ell(Y|A,X) - Y\}}$ are defined in Lemma 2, and $T(a, x)$ depends on $v_\gamma(z)$ through β. Similarly, a lower bound on β_1 requires $v_\gamma(z)$ to satisfy

$$v_\gamma(z) = \mathbb{1}\{e^T T(a, x) \leq 0\} v_u(z) + \mathbb{1}\{e^T T(a, x) > 0\} v_\ell(z).$$

3.4.5 Bounds on β when $g(a;\beta)$ is linear

If $g(a;\beta) = b(a)^T\beta$, we can derive simpler bounds. In this case we have

$$F_1(v) = \int yw(a,x)v(z)M^{-1}(v)b(a)d\mathbb{P}(z), \quad F_2(v) = \int yw(a,x)v(z)M^{-1}b(a)d\mathbb{P}(z)$$

where $M(v) = \int w(a,x)v(z)b(a)b(a)^T d\mathbb{P}(z)$ and $M = \int w(a,x)b(a)b(a)^T d\mathbb{P}(z)$.

Lemma 6. *Let $f(z) = yw(a,x)e^T M^{-1}b(a)$. We have*

$$\inf_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_1(v) = \inf_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_2(v) = \int f(z)\underline{v}(z)dP(z),$$
$$\sup_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_1(v) = \sup_{v\in\mathcal{V}_{\text{small}}(\gamma)} e^T F_2(v) = \int f(z)\overline{v}(z)dP(z),$$

where

$$\overline{v}(Z) = \gamma\mathbb{1}\{f(Z) \geq q_u(f|A,X)\} + \gamma^{-1}\mathbb{1}\{f(Z) < q_u(f|A,X)\},$$
$$\underline{v}(z) = \gamma\mathbb{1}\{f(Z) \leq q_\ell(f|A,X)\} + \gamma^{-1}\mathbb{1}\{f(Z) > q_\ell(f|A,X)\},$$

and $q_u(f|A,X)$ and $q_\ell(f|A,X)$ are the $\tau_u = \gamma/(1+\gamma)$ and $\tau_\ell = 1/(1+\gamma)$ quantiles of $f(Z)$ given (X,A).

Again, for the class $\mathcal{V}_{\text{large}}(\gamma)$ the bounds can differ and one can construct examples where either of the two is tighter, so we use the intersection of the bounds from F_1 and F_2. Bounding $F_2(v)$ over $\mathcal{V}_{\text{large}}(\gamma)$ is straightforward as discussed in the following lemma.

Lemma 7. *Let $f(z) = yw(a,x)e^T M^{-1}b(a)$. Then*

$$\inf_{v\in\mathcal{V}_{\text{large}}(\gamma)} F_2(v) = F_2(\underline{v}), \qquad \sup_{v\in\mathcal{V}_{\text{large}}(\gamma)} F_2(v) = F_2(\overline{v}),$$

where

$$\overline{v}(Z) = \gamma\mathbb{1}\{f(Z) \geq q_u(f)\} + \gamma^{-1}\mathbb{1}(f(Z) < q_u(f)),$$
$$\underline{v}(z) = \gamma\mathbb{1}\{f(Z) \leq q_\ell(f)\} + \gamma^{-1}\mathbb{1}\{f(Z) > q_\ell(f)\},$$

and $q_u(f)$ and $q_\ell(f)$ are the $\tau_u = \gamma/(1+\gamma)$ and $\tau_\ell = 1/(1+\gamma)$ quantiles of $f(Z)$.

That is, we only need marginal quantiles for the bound on $F_2(v)$. We do not have a closed form expression for bounds on $F_1(v)$ over $\mathcal{V}_{\text{large}}(\gamma)$. Instead we use the homotopy algorithm. As in the general MSM case presented in Section 3.4.4, getting confidence intervals for the bounds of β_1 over $\mathcal{V}_{\text{large}}$ is challenging because their influence functions involve solving an integral equation.

3.4.6 Local (Small γ) Bounds on β

A fast, simple approach to bounding F_1 over $\mathcal{V}_{\text{large}}(\gamma)$ is based on a functional expansion of $F_1(v)$ around the function $v_0 = 1$, or alternatively, an expansion of $F_1(L)$ around the function $L_0 \equiv \log v_0 = 0$. In principle, this will lead to tight bounds only for γ near 1, but, in our examples, it leads to accurate bounds over a range of γ values; see Figures 3.3, B.1 and B.2. Note that we do not need local bounds based on F_2 because we have an exact expression in that case.

Let $L(z) = \log v(x, a, y)$. Our propensity sensitivity model is the set of functions L such that $||L||_\infty \leq \log \gamma$. Note that $\frac{\delta F_1}{\delta L}(z) = \frac{\delta F_1}{\delta v}(z)e^L = \frac{\delta F_1}{\delta v}(z)v(z)$. No unmeasured confounding corresponds to $v_0(z) = 1$, $L_0(z) = 0$ and $\gamma = 1$. Then $F_1(L) = F_1(L_0) + e^T \int (L - L_0)\frac{\delta F_1}{\delta L}(z)d\mathbb{P}(z) + O(\gamma - 1)^2 = F_1(L_0) + e^T \int L\frac{\delta F_1}{\delta v}(z)d\mathbb{P}(z) + O(\gamma - 1)^2$ where $F_1(L_0)$ is the value of β_1 assuming no unmeasured confounding. Now, by Holder's inequality, $\int L\frac{\delta F_1}{\delta v}(z)d\mathbb{P} \leq ||L - L_0||_\infty \int |\frac{\delta F_1}{\delta v}(z)d\mathbb{P}| \leq \log \gamma \int |\frac{\delta F_1}{\delta v}(z)d\mathbb{P}|$. So, up to order $O(\gamma - 1)^2$,

$$\beta_1(L_0) - \log \gamma \int \left| \frac{\delta F_1}{\delta v}(z)d\mathbb{P} \right| \leq F_1(L) \leq \beta_1(L_0) + \log \gamma \int \left| \frac{\delta F_1}{\delta v}(z)d\mathbb{P} \right|. \tag{3.16}$$

3.5 Bounds under the Outcome Sensitivity Model

Consider now the outcome sensitivity model from Section 3.3.2. Recall that $\mu(A, X, U) = \mathbb{E}(Y|A, X, U)$ is the outcome regression on treatment and both observed and unobserved confounders, and $\Delta(a) = \int \{\mu(a, x, u) - \mu(a, x)\}d\mathbb{P}(x, u)$ is the (integrated) difference between this regression and its observed counterpart, and $|\Delta(a)| \leq \delta$. If $Y(a) \perp\!\!\!\perp A|(X, U)$, then

$$\mathbb{E}\{Y(a)\} = \int \mu(a, x, u)d\mathbb{P}(x, u) = \Delta(a) + \int \mu(a, x)d\mathbb{P}(x).$$

We can write a corresponding MSM moment condition as

$$\mathbb{E}\left[h(A)\left\{\Delta(A) + \int \mu(A, x)d\mathbb{P}(x) - g(A; \beta)\right\}\right] = 0$$

so that β is identified under no unmeasured confounding whenever $\mathbb{E}\{h(A)\Delta(A)\} = 0$. Using an approach similar to Section 3.4.2, if we assume that the bounds $\int \mu(a, x)d\mathbb{P}(x) \pm \delta$ follow models $g(a; \beta_\ell)$ and $g(a; \beta_u)$, it is straightforward to estimate β_ℓ and β_u by solving the empirical, influence function based, bias-corrected analogs of the moment conditions

$$\mathbb{E}\left[h(A)\left\{\int \mu(A, x)d\mathbb{P}(x) + \delta - g(A; \beta_u)\right\}\right] = 0,$$
$$\mathbb{E}\left[h(A)\left\{\int \mu(A, x)d\mathbb{P}(x) - \delta - g(A; \beta_\ell)\right\}\right] = 0.$$

Inference can be performed as outlined in Proposition 2.

In the linear MSM case, we have

$$g(a;\beta) = b(a)^T\beta = b(a)^T Q^{-1}\mathbb{U}\left[h(A_1)\left\{\mu(A_1,X_2)+\Delta(A_1)\right\}\right],$$

where $Q = \mathbb{E}\{h(A)b(A)^T\}$. Therefore, valid bounds on $g(a;\beta)$ are

$$b(a)^T Q^{-1}\mathbb{U}\left\{b(A_1)\mu(A_1,X_2)\right\} \pm \delta\mathbb{E}|b(a)^T Q^{-1}b(A)|,$$

which we re-write as

$$\begin{aligned} g_\ell(a) &= b(a)^T Q^{-1}\mathbb{U}\left(b(A_1)\left[\mu(A_1,X_2) - \delta\,\mathrm{sgn}\left\{b(a)^T Q^{-1}b(A_1)\right\}\right]\right)\\ g_u(a) &= b(a)^T Q^{-1}\mathbb{U}\left(b(A_1)\left[\mu(A_1,X_2) + \delta\,\mathrm{sgn}\left\{b(a)^T Q^{-1}b(A_1)\right\}\right]\right), \end{aligned}$$

since $|b(a)^T Q^{-1}h(A)| = b(a)^T Q^{-1}\,\mathrm{sgn}\left\{b(a)^T Q^{-1}b(A)\right\}b(A)$.

Our estimators are

$$\widehat{g}_j(a) = b(a)^T\widehat{\beta}_j, \quad \widehat{\beta}_j = \underset{\beta\in\mathbb{R}^k}{\mathrm{argmin}}\,\mathbb{U}_n\left\{\widehat{\zeta}_j(Z_1,Z_2) - b(A_1)^T\beta\right\}^2, \quad j\in\{u,\ell\},$$

where

$$\begin{aligned} \zeta_\ell(Z_1,Z_2) &= w(A_1,X_1)\{Y_1-\mu(A_1,X_1)\} + \mu(A_1,X_2) - \delta\,\mathrm{sgn}\left\{b(a)^T Q^{-1}b(A_1)\right\},\\ \zeta_u(Z_1,Z_2) &= w(A_1,X_1)\{Y_1-\mu(A_1,X_1)\} + \mu(A_1,X_2) + \delta\,\mathrm{sgn}\left\{b(a)^T Q^{-1}b(A_1)\right\}. \end{aligned}$$

To simplify the analysis of our estimators and avoid imposing additional Donsker-type requirements on $\widehat{\mu}$ and $\widehat{\pi}$, we proceed by assuming that $\widehat{\zeta}_l$ and $\widehat{\zeta}_u$ are estimated on samples independent from that used to compute the U-statistic in the empirical risk minimization step. This means that, in finite samples, the matrix $\widehat{Q}$ appearing in $\widehat{\zeta}_j$ and $\widetilde{Q}$ arising from the minimization step (since $\widehat{\beta}_j = \widetilde{Q}^{-1}\mathbb{U}_n\{b(A_1)\widehat{\zeta}_j(Z_1,Z_2)\}$) will not be equal, even if they estimate the same matrix $Q = \mathbb{P}\{b(A)b(A)^T\}$. In particular, $\mathrm{sgn}\left\{b(a)^T Q^{-1}b(A_1)\right\}$ might not equal $\mathrm{sgn}\left\{b(a)^T\widetilde{Q}^{-1}b(A_1)\right\}$ and so $\widehat{g}_\ell(a)$ could be larger than $\widehat{g}_u(a)$. However, this is expected to occur with vanishing probability as the sample size increases.

Proposition 4. Assume that:

1. $e^T Q^{-1}h(A)$ has a bounded density with respect to the Lebesgue measure;
2. $\left\|\int S_2\{\widehat{\zeta}_j(Z_1,z_2) - \zeta_j(Z_1,z_2)\}d\mathbb{P}(z_2)\right\| = o_{\mathbb{P}}(1)$;
3. $\|w-\widehat{w}\|\|\mu-\widehat{\mu}\| = o_{\mathbb{P}}(n^{-1/2})$.

Then $\sqrt{n}\{\widehat{g}_j(a) - g_j(a)\} \rightsquigarrow N(0, 4\Sigma)$, for $j \in \{u, \ell\}$, where

$$\Sigma = b(a)^T \text{var}\left[Q^{-1}\int S_2 b(A_1)\{\zeta_j(Z_1, z_2) - b^T(A_1)\beta_j\} d\mathbb{P}(z_2)\right] b(a).$$

Bounds on a specific coordinate of β, say β_1, are straightforward to derive in the linear MSM case by replacing $b(a)^T$ with e^T in the bounds above. When $g(a;\beta)$ is not linear, bounds on β_1 can be obtained using a homotopy algorithm similar to that in Section 3.4.4. The algorithm uses the functional derivative of $\beta(\Delta)$ with respect to Δ in $L_2(\mathbb{P}(a))$:

$$\frac{\delta\beta(\Delta)}{\delta\Delta} = \mathbb{E}\left\{h(A)\nabla_\beta g(A;\beta)^T\right\}^{-1} h(A).$$

Another, exact but computationally expensive, approach is described in Appendix B.0.3.

3.6 Time Series

Now we extend the methods to time varying treatments. In this setting, we have data $(X_1, A_1), \ldots, (X_T, A_T, Y)$ on each subject, where X_t can include an intermediate outcome Y_t. We write $\overline{X}_t = (X_1, \ldots, X_t)$ and $\overline{A}_t = (A_1, \ldots, A_t)$. An intervention corresponds to setting $\overline{A}_T = \overline{a}_T = (a_1, \ldots, a_T)$ with corresponding counterfactual outcome $Y(\overline{a}_T)$. In this case, the assumption of no unmeasured confounding is expressed as $A_t \perp\!\!\!\perp Y(\overline{a}_T)|(\overline{A}_{t-1}, \overline{X}_t)$ for every $t \in \{1, \ldots, T\}$. Under this assumption, the g-formula [Robins, 1986] is

$$\mathbb{E}\{Y(\overline{a}_T)\} = \int \mu(\overline{a}_T, \overline{x}_T) \prod_{s=1}^{T} d\mathbb{P}(x_s|\overline{x}_{s-1}, \overline{a}_{s-1})$$

where $\mu(\overline{a}_T, \overline{x}_T) = \mathbb{E}(Y \mid \overline{X}_T = \overline{x}_T, \overline{A}_T = \overline{a}_T)$. As before, a MSM is a model $g(\overline{a}_T;\beta)$ for $\mathbb{E}\{Y(\overline{a}_T)\}$. A common example is $g(\overline{a}_T;\beta) = \beta_0 + \beta_1 \sum_{s=1}^{T} a_s$. For some user-specified function $h(\cdot)$ of the treatments, it can be shown that

$$\mathbb{E}\left[h(\overline{A}_T)W_T(\overline{A}_T, \overline{X}_T)\left\{Y - g(\overline{A}_T;\beta)\right\}\right] = 0, \text{ where } W_T(\overline{a}_T, \overline{x}_T) = \frac{\prod_{s=1}^{T}\pi(a_s|\overline{a}_{s-1})}{\prod_{s=1}^{T}\pi(a_s|\overline{x}_s, \overline{a}_{s-1})}.$$

3.6.1 Bounds on $g(\overline{a}_t;\beta)$ under Propensity Sensitivity Confounding

Let $\overline{U}_T = (U_1, \ldots, U_T)$ denote unobserved confounders. If $A_t \perp\!\!\!\perp Y_t(\overline{a}_t)|(\overline{A}_{t-1}, \overline{X}_t, \overline{U}_t)$ for all t, then the g-formula becomes

$$\mathbb{E}\{Y(\overline{a}_T)\} = \int \mathbb{E}(Y \mid \overline{A}_T = \overline{a}_T, \overline{X}_T = \overline{x}_T, \overline{U}_T = \overline{u}_T) \prod_{s=1}^{T} d\mathbb{P}(x_s, u_s|\overline{x}_{s-1}, \overline{u}_{s-1}, \overline{a}_{s-1}),$$

Define

$$v_T(Y, \overline{A}_T, \overline{X}_T) = \mathbb{E}\left\{ \frac{\prod_{s=1}^T \pi(A_s|\overline{X}_s, \overline{A}_{s-1})}{\prod_{s=1}^T \pi(A_s|\overline{X}_s, \overline{U}_s, \overline{A}_{s-1})} \mid Y, \overline{A}_T, \overline{X}_T \right\}$$

and note that we can rewrite $\mathbb{E}\{Y(\overline{a}_T)\}$ as

$$\mathbb{E}\{Y(\overline{a}_T)\} = \int \mathbb{E}\{Y v_T(Y, \overline{a}_T, \overline{x}_T)|\overline{A}_T = \overline{a}_T, \overline{X}_T = \overline{x}_T\} \prod_{s=1}^T d\mathbb{P}(x_s|\overline{x}_{s-1}, \overline{a}_{s-1}).$$

It can be shown that $\mathbb{E}\{v_T(Y, \overline{A}_T, \overline{X}_T)\} = 1$ and also that

$$\int \mathbb{E}\{v_T(Y, \overline{A}_T, \overline{X}_T) \mid \overline{A}_T, \overline{X}_T\} \prod_{s=2}^T d\mathbb{P}(x_s \mid \overline{x}_{s-1}, \overline{a}_{s-1}) = 1 \tag{3.17}$$

However, unless additional assumptions are invoked, it is not the case that $\mathbb{E}\{v_T(Y, \overline{A}_T, \overline{X}_T) \mid \overline{A}_T, \overline{X}_T\} = 1$. Getting bounds in the propensity sensitivity model enforcing $v_T(Y, \overline{A}_T, \overline{X}_T) \in [\gamma^{-1}, \gamma]$ and $\mathbb{E}\{v_T(Y, \overline{A}_T, \overline{X}_T)\} = 1$ is straightforward. For example, as shown in Section B.2.13 in the appendix, it holds that

$$\mathbb{E}\left[h(\overline{A}_T) W_T(\overline{A}_T, \overline{X}_T)\left\{Y v_T(Y, \overline{A}_T, \overline{X}_T) - g(\overline{A}_T; \beta)\right\}\right] = 0$$

In this light, methods based on the class $\mathcal{V}_{\text{large}}(\gamma)$ described in Sections 3.4.4 and 3.4.5 apply here as well with W_T replacing W and v_T replacing v. The local approach taken in Section 3.4.6 also applies. However, enforcing (3.17) appears more challenging and we leave it for future work.

3.6.2 Bounds under Outcome Sensitivity Confounding

Bounds for $g(\overline{a}_T; \beta)$ and for coordinates of β governed by the outcome sensitivity model can be derived in a similar fashion by extending the results in Section 3.5.

3.7 Examples

In this section we present a static treatment example and a time series example. The appendix also contains simple, proof of concept synthetic examples.

3.7.1 Effect of Mothers' Smoking on Infant Birthweight

We re-analyzed a dataset of births in Pennsylvania between 1989 and 1991, which has been used to investigate the causal effects of mothers' smoking behavior on infants birthweight. Previous analyses [Almond et al., 2005, Cattaneo, 2010], assuming no unmeasured confounders, found a negative effect of smoking on the infant's weight. Recently, Scharfstein et al. [2021] conducted a sensitivity analysis to the assumption of no unmeasured confounding by dichotomizing the

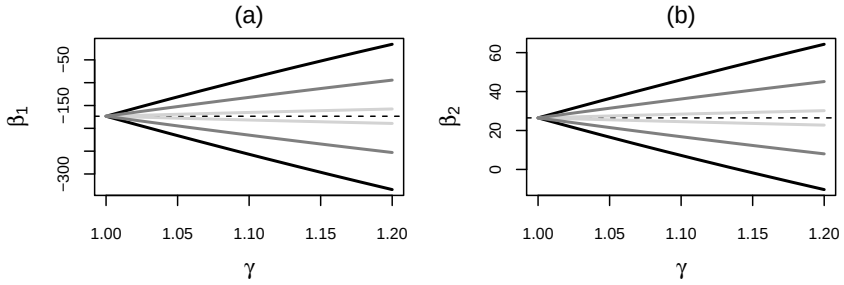


Figure 3.1: **Bounds for β_1 and β_2 for the birthweight dataset, assuming the MSM $g(a;\beta) = \beta_0 + \beta_1 a + \beta_2 a^2$.** *The dotted horizontal lines are at $\widehat{\beta}_1$ and $\widehat{\beta}_2$. The black bounds are from F_1 over $\mathcal{V}_{\text{large}}$ for the propensity model, found using the homotopy algorithm (Section 3.4.4). The local approximation to F_1 (Section 3.4.6) matched the black bounds closely (not shown), similar to the appendix examples. The quadratic term parameter loses significance at $\gamma \approx 1.11$ and the linear term at $\gamma \approx 1.25$. The dark and light grey bounds use the subset sensitivity model (Sections 3.3.3 and B.0.2), with $\epsilon = .5$ and $.1$, respectively. Bounds are all the narrower when ϵ is smaller, as expected.*

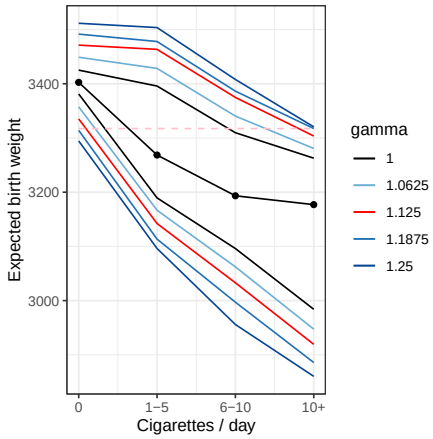
treatment into smoking vs non-smoking. In line with previous work, they found a negative effect of smoking on the child's weight, but also identified plausible values of their sensitivity parameters consistent with a null effect. They concluded that, while likely negative, the true effect of smoking on weight might be smaller than that estimated under no unmeasured confounding. We complement and expand on their analysis by considering sensitivity models that can accommodate MSMs; we reach similar conclusions, although we find the estimated effect to be less sensitive to the unmeasured confounding parametrized by our sensitivity models.

The dataset consists of a random subsample of $5,000$ observations from the original dataset that is available online.[1] The outcome is birthweight and the treatment is an ordered categorical variable taking six values corresponding to ranges $\{0, 1\text{-}5, 6\text{-}10, 11\text{-}15, 16\text{-}20, 21+\}$ of cigarettes smoked per day. There are 53 pre-treatment covariates including mother's and father's education, race, and age; mother's marital status and foreign born status; indicators for trimester of first prenatal care visit and mother's alcohol use.

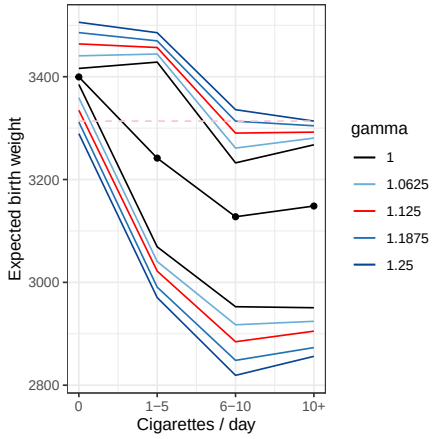
Figure 3.1 shows bounds on β_1 and β_2 for the quadratic MSM given by $g(a;\beta) = \beta_0 + \beta_1 a + \beta_2 a^2$ under propensity sensitivity, based on F_1 over $\mathcal{V}_{\text{large}}$ (with only six treatment values, we cannot fit a more complex parametric model). We estimated the propensity $\pi(a|x)$ via a log-linear neural net using the $\texttt{nnet}$ package for the R software, as in Cattaneo [2010]. The quadratic term parameter loses significance at $\gamma \approx 1.11$ and the linear term at $\gamma = 1.25$. Figure 3.1 also shows bounds on β_1 and β_2 under the subset sensitivity model with $\epsilon = .5$ and $.1$. As expected, there is much less sensitivity for small ϵ.

Recall from (3.5) that γ measures the change in the propensity score when U is dropped.

[1] https://github.com/mdcattaneo/replication-C_2010_JOE



(a) *Estimated bands for the linear MSM* $g(a;\beta) = \beta_0 + \beta_1 a + \beta_2 a^2$.



(b) *Estimated bands for the saturated MSM* $g(a;\beta) = \beta_0 + \sum_{j=1}^{4} \beta_j \mathbb{1}\{a_j \in j^{th}bin\}$.

Figure 3.2: **Pointwise 95%-confidence bands on the bounds for $\mathbb{E}\{Y(a)\} = g(a;\beta)$ under the propensity sensitivity model**, *where $a \in \{0, 1\text{–}5, 6\text{–}10, 10+\}$ cigarettes per day. The lines with dots are $g(a;\widehat{\beta})$.*

To determine if $\gamma = 1.25$ constitutes substantial confounding, we followed the ideas in Cinelli and Hazlett [2020] by assessing changes to the propensity score when observed confounders are dropped. Most authors drop one covariate at a time but with 53 covariates, we found that this caused almost no changes to the propensity score. Instead, we (i) dropped half of the covariates, and (ii) computed, for each data point, the ratio of propensity scores using all the covariates and the randomly chosen subset, and repeated (i, ii) 100 times. Each repeat yielded a distribution of propensity score ratios and we used the average of their 80th percentiles as a measure of substantial confounding. This value is $\gamma = 1.20$, so we conclude that the causal effect of smoking on infant birthweight remains significant even under substantial confounding. The next analysis confirms this conclusion.

Next, Figure 3.2 shows 95% point-wise confidence bands for the bounds on $g(a;\beta)$ under propensity sensitivity based on $\mathcal{V}_{\text{small}}$, assuming that the bounds are modeled as $g(a;\beta_\ell)$ and $g(a;\beta_u)$; see Proposition 2. Note that Figure 3.1 showed the bounds on β_1 and β_2 rather than confidence bands on these bounds, because confidence bands are difficult to obtain; see Sections 3.4.4 and 3.4.5. Figure 3.2a shows results for the quadratic MSM $g(a;\beta) = \beta_0 + \beta_1 a + \beta_2 a^2$, and as a safeguard against MSM mis-specification, Figure 3.2b shows the saturated parametric MSM fit. The black bands corresponding to $\gamma = 1$ assume no-unmeasured-confounding (so they are confidence bands for $g(a;\widehat{\beta})$) and increasing values of γ correspond to increasing amount of unmeasured confounding. We estimated the nuisance functions nonparametrically: the outcome model $\mu(x,a)$ and conditional quantiles $q_j(Y|a,x)$ were fitted assuming generalized additive models, with mother's and father's ages, education and birth order entering the model linearly, and number of prenatal care visits and months since last birth entering the model as smooth terms – we used the `mgcv` and `qgam` packages in R, respectively;

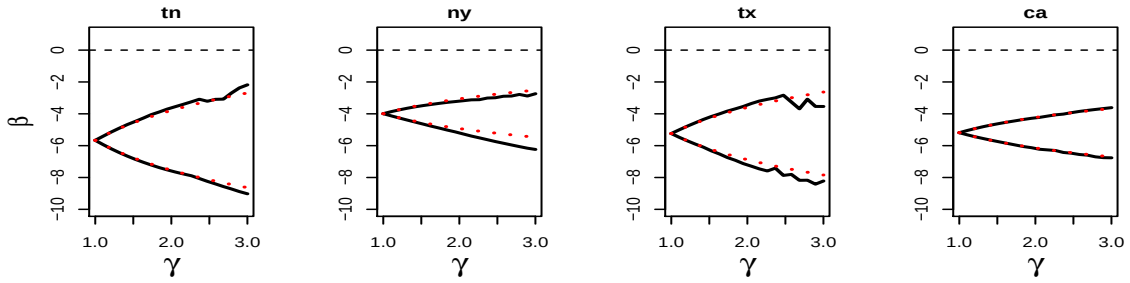


Figure 3.3: **Bounds on β in MSM (6.13) for the Covid data in four US states**. *The black bounds are from F_1 over $\mathcal{V}_{\text{large}}$ with increasing amount of unobserved confounding, under the propensity confounding model. The occasional lack of smoothness is due to estimating the quantile q from small samples ($n = 40$) in the homotophy algorithm (Section B.1.1). The dotted red bounds are the local approximations to F_1 (Section 3.4.6). The MSM coefficients remain significant under substantial unobserved confounding for the four states: mobility has a significant effect on Covid deaths.*

the propensity $\pi(a|x)$ was estimated via a log-linear neural net, as above. We constructed the 95% point-wise confidence bands relying on Proposition 2 and the Hulc method by Kuchibhotla et al. [2021]. For the Hulc, the sample needs to be split into six subsamples, but because of small sample sizes in some categories, we collapsed all regimes of 10+ cigarettes into one category, thereby reducing the number of treatment regimes to four. Consistent with Figure 3.1 and previous analyses [Almond et al., 2005, Cattaneo, 2010], we found a statistically significant negative relationship between smoking and birthweight under no-unmeasured-confounding. The relation ceases to be significant for $\gamma = 1.1875$ when the quadratic model is used and $\gamma = 1.25$ when the saturated model is used.

3.7.2 Effect of Mobility on Covid-19 Deaths

We revisit the analysis in Bonvini et al. [2022b] on the causal effects of mobility on deaths due to Covid-19 in the United States. In their paper, a sensitivity analysis to the no unmeasured confounding assumption was conducted under the propensity model without providing details. We provide details here.

The data consist of weekly observations, at the state level, on the number of Covid-19 deaths Y_t and a measure of mobility "proportion at home," A_t, which is the fraction of mobile devices that did not leave the immediate area of their home. The time period considered in the analysis is February 15 2020 (week 1) to November 15 2020 (week 40). We focus on four states, CA, FL, NY and TN, as representatives of four different evolutions of the pandemic; their observed time series of deaths are plotted as dots in Figure 3.4. We model each state separately so that differences between states do not act as confounders of the treatment/outcome relationship.

Our MSM is given by

$$g(\overline{a}_t, \beta, \nu) = \mathbb{E}[L_t(\overline{a}_t)] = \nu(t) + \beta M_t \tag{3.18}$$

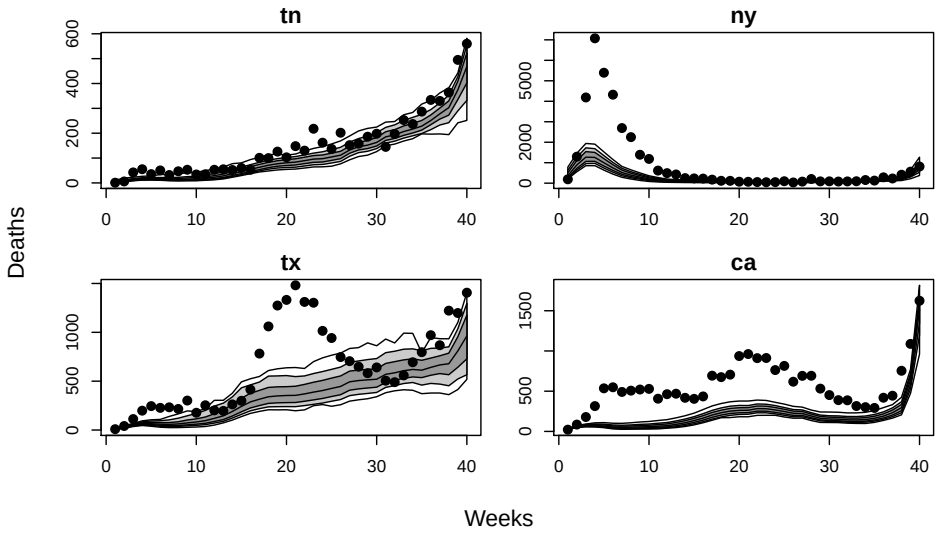


Figure 3.4: **Bounds for counterfactual deaths $\psi(a_T) = \mathbb{E}(Y(a_T))$ for the Covid data in four US states using MSM (6.13) in a hypothetical mobility scenario a_T corresponding to shifting the observed mobility pattern two weeks earlier.** *The bounds are from F_1 over $\mathcal{V}_{\text{large}}$ under propensity sensitivity, found using the homotopy algorithm. The shades correspond to $\gamma = 3$ (white), $\gamma = 2$ (light grey) and $\gamma = 1$ (no unmeasured confounding, dark grey). The black dots are observed deaths. Our analysis suggests that, even with substantial unobserved confounding, sheltering two weeks earlier would have saved lives, although only by a small number in TN, because the epidemic there started more mildly.*

where $\overline{a}_t = (a_1, \ldots, a_t)$, $L_t(\overline{a}_t)$ are log-counterfactual deaths, $L_t = \log(Y_t + 1)$, $M_t \equiv M(\overline{a}_t) = \sum_{s=1}^{t-\delta} a_s$, and $\delta = 4$ weeks is approximately the mean time from infection to death from Covid-19. The nuisance function $\nu(t)$ is assumed to be non-linear to capture changes in death incidence due to time varying variables other than mobility, for example probability of dying, which decreased over time due to better hospital treatment, number of susceptibles to Covid-19, which naturally decreased, and social distancing changes.

Figure 3.3 shows $\widehat{\beta}$ for the four states, along with lower and upper bounds under propensity sensitivity. The estimates are negative, as would be expected since higher A_s means that more people sheltered at home, and they remain negative even under substantial unobserved confounding.

Bonvini et al. [2022b] also estimated counterfactual deaths under three hypothetical mobility regimes $\overline{A}_t = (A_1, \ldots, A_t)$: "*start one week earlier*" and "*start two weeks earlier*", which shifts the observed mobility profiles back by one or two weeks with aim to assess Covid-19 infections if we had started sheltering in place one and two weeks earlier; and "*stay vigilant*", which halves the slope of the rapid decrease in stay at home mobility after the initial peak in week 9, when a large proportion of the population hunkered down after witnessing the situation in New York city. To save space, Figure 3.4 shows only the estimated counterfactual deaths and

bounds for the "start two weeks earlier" scenario. Bounds were computed on $g(\overline{a}_t; \beta)$ for each t using the homotopy algorithm on F_1 over $\mathcal{V}_{\text{large}}$, under propensity sensitivity (Section 3.4.4).

Bounds for β and $g(\overline{a}_t; \beta)$ under the outcome sensitivity model requires an outcome model, which we do not pursue here.

3.8 Conclusion

We have derived several sensitivity analysis methods for marginal structural models. Doing so may require additional modeling, for example, using quantile regression. We also saw that approximate, conservative bounds are possible without quantile regression.

We have focused on the traditional interventions corresponding to setting the treatment to a particular value. In a future paper, we address sensitivity analysis under stochastic interventions. Here we find that these interventions can lead to inference that is less sensitive to unmeasured confounding than traditional interventions.

One issue that always arises in sensitivity analysis is how to systematically choose ranges of values for the sensitivity parameters (e.g., γ, δ, ϵ, in our case). In the smoking example, we dropped large sets of observed confounders to provide a benchmark, but for the most part this is an open problem.

3.9 Acknowledgements

We thank Prof. Nicole Pashley for helpful discussions regarding the interpretation of the causal effect of mobility on deaths due to Covid-19. In particular, unmeasured confounding is not the only issue that needs to be addressed when interpreting our results. A potential complication is that there could be multiple versions of mobility, e.g. a person may move to go to work versus a bar. These different versions of mobility may affect the probability of dying due to Covid-19 differently, complicating the interpretation of the overall effect of reduced mobility on deaths. Conducting a sensitivity analysis to gauge the impact of multiple versions of the same treatment is an important avenue for future work.

Chapter 4

Minimax optimal subgroup identification

This chapter is a preliminary draft of my work supervised by Edward H. Kennedy and Luke J. Keele.

4.1 Introduction

Much empirical research focuses on estimating causal effects. One commonly estimated causal effect is the average treatment effect (ATE), which is the difference in average outcome if everyone in the population, versus no one, receives treatment. By definition, the ATE is an aggregate measure of treatment efficacy that does not capture any effect heterogeneity. An alternative measure of treatment effect is the conditional average treatment effect (CATE), which is the ATE restricted to a subpopulation of interest. The subpopulation is typically defined by the values of some *a priori* selected variables known as *effect modifiers*. One natural extension of the CATE is to estimate the set of units with treatment effects larger (or smaller) than some user-specified threshold. For example, when the threshold is zero, assigning treatment to only those units with a positive treatment effect is the optimal rule maximizing the mean outcome in the population (see, e.g, Robins [2004], Hirano and Porter [2009], Chakraborty and Moodie [2013], and Luedtke and Van Der Laan [2016]).

To consider this problem, informally, we define Y as the outcome, A as an indicator for treatment, and X as measured confounders and simultenously effect modifiers. Using these terms, the CATE $\tau(x)$ is equal to $\tau(x) = \mathbb{E}(Y \mid A = 1, X = x) - \mathbb{E}(Y \mid A = 0, X = x)$, and the ATE is $\mathbb{E}\{\tau(X)\}$. Our target of inference, the upper level set of the CATE at θ, is

$$\Gamma(\theta) = \{x \in \mathcal{X} : \tau(x) > \theta\}$$

We assume the level $\theta \in \mathbb{R}$ to be user-specified. For some estimator $\widehat{\tau}(x)$ of $\tau(x)$, we estimate

the level set, $\Gamma(\theta)$, with

$$\widehat{\Gamma}(\theta) = \{x \in \mathcal{X} : \widehat{\tau}(x) > \theta\}.$$

This estimator defines the set of study units with estimated CATEs that are greater than θ. Clearly, an estimator for $\Gamma(\theta)$ depends on an estimator for $\tau(x)$, and the performance of $\widehat{\Gamma}(\theta)$ will be affected by how well $\tau(x)$ can be estimated. Yet, we will show that estimating $\Gamma(\theta)$ can be an easier statistical problem than estimating $\tau(x)$ itself on its support. Intuitively, one needs to be able to estimate $\tau(x)$ accurately only in regions of the covariates' space where $\tau(x)$ is close to θ. Further, if $\tau(X)$ has a bounded density and a particular loss function is used, we will show that the convergence rate of $\widehat{\Gamma}(\theta)$ to $\Gamma(\theta)$ will generally be faster than that of $\widehat{\tau}(x)$ to $\tau(x)$.

Recent work has developed a number of proposals for CATE estimation with an emphasis on using nonparametric estimation methods borrowed from the machine learning (ML) literature [Athey and Imbens, 2016, Foster and Syrgkanis, 2019, Hahn et al., 2020, Imai and Ratkovic, 2013, Kennedy, 2020, Kennedy et al., 2022, Künzel et al., 2019, Nie and Wager, 2021, Semenova and Chernozhukov, 2021, Shalit et al., 2017, Wager and Athey, 2018]. In our work, we focus on a class of nonparametric estimators for the CATE that are embedded in a meta-learner framework that separates estimation of the CATE into a multi-step regression procedure. In the first step, a set of nuisance functions is estimated using flexible machine learning models. Then, in the second-stage, an estimate of $\tau(x)$ is computed using the previous nuisance function estimates as inputs. More specifically, we focus on two recently proposed estimators of $\tau(x)$: the DR-Learner analyzed in Kennedy [2020] and the Lp-R-Learner proposed in Kennedy et al. [2022]. The first one is a general estimation procedure based on a two-stage regression that can be computed using off-the-shelf software. The second is a more complicated estimator, which has been shown to be minimax optimal for an important set of models.

We merge this work on flexible estimation of CATEs with the extensive literature on nonparametric estimation of (upper) level sets. See, for examples, Qiao and Polonik [2019], Mammen and Polonik [2013], Chen et al. [2017], Rigollet and Vert [2009], Willett and Nowak [2007] and references therein. The main difference between our work and this research is that in our context the level set is defined by the difference of two regressions, the optimal estimation of which can be considerably more involved than that of either regression. Within this literature, our work is closest to Rigollet and Vert [2009], and we use their general framework to analyze the performance of our estimators.

Other streams of research closely related to our work are policy learning [Athey and Wager, 2021, Ben-Michael et al., 2022, Hirano and Porter, 2009] and contextual bandits [Gur et al., 2022]. In the policy learning literature, it is typically assumed that the best policy belongs to some well-behaved and interpretable class of decision rules. This is different from the route we take in this work; instead of restricting the complexity of the level set class, we restrict the complexity of the CATE function in nonparametric models. In addition, while one of the core goals of the literature on contextual bandits is to identify regions of the covariates space where the treatment effect is positive, this is usually done in settings where the probability of taking

a given action or receiving treatment, i.e., the propensity score, is under the experimenter's control and known. Instead, we consider observational studies where the propensity score is unknown. Finally, Reeve et al. [2021] has also considered a similar problem to the one discussed in this paper, but they require that the propensity score is known and the estimator appears to be more complicated.

4.1.1 Our contribution

The level set estimator that we study follows the plug-in principle and consists of simply thresholding an estimator of the CATE. To the best of our knowledge, how the properties of a CATE learner relate to those of the corresponding level set estimator has not been investigated in the literature yet. As such, our first goal is to derive the asymptotic properties of level set estimators depending on which estimator of the CATE is used. We calculate the risk for estimating $\Gamma(\theta)$ by thresholding a general estimator of the CATE required to satisfy a particular exponential inequality. Then, we specialize the results when the CATE is estimated with the DR-Learner or the Lp-R-Learner. Further, we show that if the Lp-R-Learner is used, the risk achieved is minimax optimal, under certain conditions. The optimality of thresholding the Lp-R-Learner for estimating CATE level sets had yet to be established. As an intermediate step for obtaining our main results, we derive exponential inequalities for CATE estimators based on linear smoothing, which might be of independent interest.

We establish the minimax optimal rate for estimating $\Gamma(\theta)$ in Hölder smoothness models where $\tau(x)$ and the nuisance functions have potentially different smoothness levels. Kennedy et al. [2022] have recently shown that, from a minimax optimality point of view, the parameter $\tau(x)$ shares features of a functional with nuisance components [Robins et al., 2009b, 2017b] and a standard nonparametric regression [Tsybakov, 2009]. Building upon their work and Rigollet and Vert [2009], we show that $\Gamma(\theta)$ behaves as a hybrid parameter not only exhibiting features similar to those of $\tau(x)$, but also those of a Bayes classifier. Effectively, we connect the problem of estimating CATE level sets to the domains of classification, nonparametric regression and functional estimation. We also briefly discuss the construction of confidence sets for $\Gamma(\theta)$ based on the distribution of $\sup_{x\in\mathcal{X}} |\widehat{\tau}(x) - \tau(x)|$. Finally, we illustrate our methods in simulations and with a real dataset used to study the effect of laparoscopic surgery for partial colectomy on mortality and complications.

4.2 Notation

We assume that X has at least one continuous component and denote the marginal CDF of X by $F(x)$, with corresponding density $f(x)$ with respect to the Lebesgue measure, which we assume to be uniformly bounded. We also let d denote the dimension of X and let $\mathcal{X}$ be the set of all $x \in \mathbb{R}^d$ such that $f(x) > 0$.

We define the nuisance functions:

$$\pi(X) = \mathbb{P}(A = 1 \mid X),\ \mu(X) = \mathbb{E}(Y \mid X),\ \mu_a(X) = \mathbb{E}(Y \mid A = a, X),$$
$$\text{and } \tau(X) = \mu_1(X) - \mu_0(X) = \mathbb{E}[\{\pi(X)\}^{-1} YA - \{1 - \pi(X)\}^{-1} Y(1 - A) \mid X].$$

Unless we need to keep track of constants, we will adopt the notation $a \lesssim b$ to mean that there exists a constant C such that $a \leq Cb$. We assume all the nuisance functions are uniformly bounded and $\pi(x)$ is also bounded away from 0 and 1. To keep the notation as light as possible, we will often write Γ to mean $\Gamma(\theta)$.

Let $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$, $|s| = \sum_{i=1}^d s_i$, $s! = s_1! \cdots s_d!$ and $D^s = \frac{\partial^{s_1 + \cdots s_d}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}$ be the differential operator. For $\beta > 0$, let $\lfloor \beta \rfloor$ denote the largest integer strictly less than β. Given $x \in \mathbb{R}^d$ and f a $\lfloor \beta \rfloor$-times continuously differentiable function, let

$$f_x(u) = \sum_{|s| \leq \lfloor \beta \rfloor} \frac{(u - x)^s}{s!} D^s f(x)$$

denote its Taylor polynomial approximation of order $\lfloor \beta \rfloor$ at $u = x$.

Definition 1 (locally Hölder-β function)**.** A function f is "β-smooth locally around a point $x_0 \in \mathcal{X}_0$" if it is $\lfloor \beta \rfloor$-times continuosly differentiable at x_0 and there exists a constant L such that

$$|f(x) - f_{x_0}(x)| \leq L \|x - x_0\|^\beta \text{ for all } x \in B(x_0, r), r > 0.$$

There are a few ways to measure the performance of $\widehat{\Gamma}(\theta)$, two of which are

- $d_\Delta(\widehat{\Gamma}, \Gamma) = \int_{\widehat{\Gamma} \Delta \Gamma} f(x) dx$, for $\widehat{\Gamma} \Delta \Gamma = (\widehat{\Gamma}^c \cap \Gamma) \cup (\widehat{\Gamma} \cap \Gamma^c)$ (set difference);
- $d_H(\widehat{\Gamma}, \Gamma) = \int_{\widehat{\Gamma} \Delta \Gamma} |\tau(x) - \theta| f(x) dx$ (penalized set difference).

The first one is simply the $\mathbb{P}_X$-measure of the set difference between $\widehat{\Gamma}$ and Γ. The second one is the $\mathbb{P}_X$-measure of the set difference simply with a smaller penalty assigned to errors made by including / excluding values of X for which the CATE is close to θ. In particular, whether or not x such that $\tau(x) = \theta$ is included or excluded from the set $\widehat{\Gamma}$ has no impact on the error measured by $d_H(\widehat{\Gamma}, \Gamma)$. If $\theta = 0$, this means that, according to this metric, it does not matter whether we assign treatment to units with zero treatment effect. We will focus on $d_H(\widehat{\Gamma}, \Gamma)$ and analyze the risk

$$\mathbb{E}\left\{ d_H(\widehat{\Gamma}, \Gamma) \right\} = \mathbb{E}\left\{ \int_{\widehat{\Gamma} \Delta \Gamma} |\tau(x) - \theta| f(x) dx \right\}, \tag{4.1}$$

which we represent in Figure 4.1 for the case $d = 1$ and $X \sim \text{Unif}(0, 1)$.

Remark 5. Willett and Nowak [2007] study estimation of the level sets of a function using dyadic trees. Their approach, adjusted to our settings, would prescribe finding $\widehat{\Gamma}(\theta)$ by

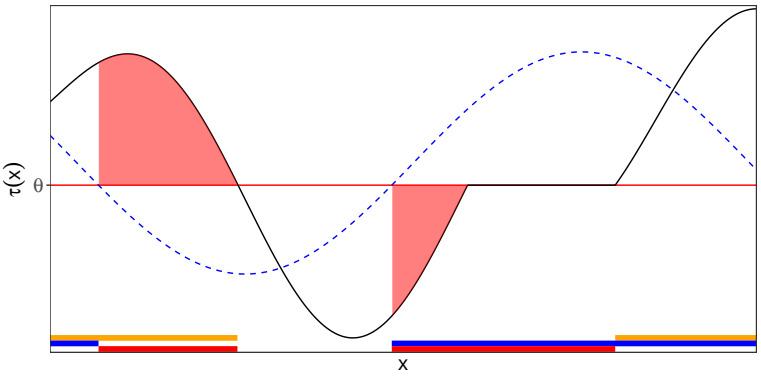


Figure 4.1: Representation of the loss in eq. (4.1) for $d = 1$ and X uniformly distributed. The solid line is $\tau(x)$, the dotted line is $\widehat{\tau}(x)$, and the shaded area equals $d_H(\widehat{\Gamma}, \Gamma)$. The orange portion of the x-axis represents $\Gamma(\theta)$, the blue one $\widehat{\Gamma}(\theta)$ and the red one $\widehat{\Gamma}\Delta\Gamma$.

minimizing an estimate of

$$R\{\overline{\Gamma}(\theta)\} \propto \int \{\theta - \tau(x)\}[\mathbb{1}\{x \in \overline{\Gamma}(\theta)\} - \mathbb{1}\{x \in \overline{\Gamma}^c(\theta)\}]f(x)dx$$

as the risk function. They show that minimizing $R\{\overline{\Gamma}(\theta)\}$ is equivalent to minimizing the excess risk, i.e.

$$R\{\widehat{\Gamma}(\theta)\} - R\{\Gamma(\theta)\} = \int_{\Gamma\Delta\widehat{\Gamma}} |\tau(x) - \theta| f(x)dx$$

which is equivalent to the loss $d_H(\widehat{\Gamma}, \Gamma)$ that we use in this paper. We leave the study of empirical risk minimizers for estimating CATE level sets for future work.

As described below, the performance of our estimators will depend crucially on the difficulty in estimating the CATE around the level θ. The intuition is that, to estimate $\Gamma(\theta)$, one needs to estimate the sign of $\tau(x) - \theta$ well and, in regions of the covariates' space where $\tau(x)$ is far from θ, estimating this sign well does not require estimating $\tau(x)$ precisely. On the contrary, for values of x such that $\tau(x)$ is close to θ, estimating $\tau(x)$ precisely plays an important role in determining the sign of $\tau(x) - \theta$. For example, $\tau(x)$ may be a very complex function far away from θ but, as long as it is well-behaved and easy to estimate close to θ, one may hope to still be able to estimate $\Gamma(\theta)$ well. In this respect, a typical example that we consider is when $\tau(x)$ is γ-smooth in a neightborhood around θ and γ'-smooth everywhere else.

4.3 Estimation

4.3.1 Estimand & setup

The goal of this section is to provide an upper bound on the risk (4.1) for generic CATE estimators. Following Rigollet and Vert [2009], we introduce a margin assumption governing the mass concentrated around the level set $\chi = \{x \in \mathcal{X} : \tau(x) = \theta\}$ encoded below.

Assumption 4. There exist positive constants ϵ_0 and c_0 such that such that, for all $\epsilon \in (0, \epsilon_0]$, it holds that $\mathbb{P}_X(0 < |\tau(X) - \theta| < \epsilon) \leq c_0 \epsilon^{\xi}$.

The margin condition (Assumption 4) can yield fast convergence rates when the performance is measured by the risk in eq. (4.1). Crucially, it can be shown to hold with exponent $\xi = 1$ as long as the density of $\tau(X)$ is bounded, which can be satisfied in many applications. The following two propositions are restatements of Lemmas 5.1 and 5.2 in Audibert and Tsybakov [2007] written for the problem considered here; we provide their proofs for completeness.

Proposition 5. Under Assumption 4, it holds that

$$\mathbb{E}\left\{d_H(\widehat{\Gamma}, \Gamma)\right\} \leq \mathbb{E}\left[\int_{\mathcal{X}} \mathbb{1}\left\{|\tau(x) - \theta| \leq \|\widehat{\tau} - \tau\|_\infty\right\} |\tau(x) - \theta| f(x) dx\right] \lesssim \mathbb{E}\left(\|\widehat{\tau} - \tau\|_\infty^{1+\xi}\right)$$

Proof. The proposition simply follows from the observation that

$$\begin{aligned}
\mathbb{1}\left\{x \in \widehat{\Gamma}(\theta) \Delta \Gamma(\theta)\right\} &= \left|\mathbb{1}\left\{\widehat{\tau}(x) - \theta > 0\right\} - \mathbb{1}\left\{\tau(x) - \theta > 0\right\}\right| \\
&\leq \mathbb{1}\left\{|\tau(x) - \theta| \leq |\widehat{\tau}(x) - \tau(x)|\right\} \\
&\leq \mathbb{1}\left\{|\tau(x) - \theta| \leq \|\widehat{\tau} - \tau\|_\infty\right\}
\end{aligned}$$

The second inequality follows by Lemma 1 in Kennedy et al. [2020]. □

Proposition 5 applies to any estimator $\widehat{\tau}$ of τ and links the error in estimating the upper level sets to the error in estimating τ. In particular, it is often the case that $\|\widehat{\tau} - \tau\|_\infty$ is of the same order of the pointwise error $|\widehat{\tau}(x) - \tau(x)|$ up to a log factor. In this sense, Proposition 5 would typically match the sharper result described in Lemma 8 up to a log factor provided that estimating $\tau(x)$ near the level θ is at least as difficult as estimating it on the entire domain. The next proposition links the level set estimator error to the L_p norm of the error in estimating τ. This proposition, however, appears to give results that match those in Lemma 8 only if the margin condition does not hold, i.e. $\xi = 0$.

Proposition 6. Under Assumption 4, it holds, for any $1 \leq p < \infty$:

$$\mathbb{E}\left\{d_H(\widehat{\Gamma}, \Gamma)\right\} \leq C_{\xi,p} \mathbb{E}\left\{\|\widehat{\tau} - \tau\|_p^{\frac{p(1+\xi)}{p+\xi}}\right\}$$

for some constant $C_{\xi,p}$ depending on p and ξ.

Proof. It holds that

$$
\begin{aligned}
d_H(\widehat{\Gamma}, \Gamma) &\leq \int \mathbb{1}\left\{|\tau(x) - \theta| \leq |\widehat{\tau}(x) - \tau(x)|\right\} \mathbb{1}\left\{0 < |\tau(x) - \theta| \leq t\right\} |\tau(x) - \theta| f(x) dx \\
&\quad + \int \mathbb{1}\left\{|\tau(x) - \theta| \leq |\widehat{\tau}(x) - \tau(x)|\right\} \mathbb{1}\left\{|\tau(x) - \theta| > t\right\} |\tau(x) - \theta| f(x) dx \\
&\leq \int \mathbb{1}\left\{|\tau(x) - \theta| \leq |\widehat{\tau}(x) - \tau(x)|\right\} \mathbb{1}\left\{0 < |\tau(x) - \theta| \leq t\right\} |\tau(x) - \widehat{\tau}(x)| f(x) dx \\
&\quad + \int \mathbb{1}\left\{|\tau(x) - \theta| \leq |\widehat{\tau}(x) - \tau(x)|\right\} \mathbb{1}\left\{|\tau(x) - \theta| > t\right\} |\tau(x) - \widehat{\tau}(x)| f(x) dx \\
&\lesssim \|\widehat{\tau} - \tau\|_p t^{\frac{\xi}{p}(p-1)} + \frac{\|\widehat{\tau} - \tau\|_p^p}{t^{p-1}}
\end{aligned}
$$

by Hölder's inequality. Minimizing the RHS over t yields the desired bound. □

Proposition 5 and 6 show that larger values of ξ make estimation of the upper level sets easier. However, as noted in Audibert and Tsybakov [2007], ξ cannot be too large or else the class of distributions satisfying the margin condition becomes small. This is particularly clear in smoothness models where $\tau(x)$ is γ-smooth around the cutoff in the sense of Definition 1. If $\tau(x)$ is smooth enough around the cutoff, it cannot jump away from the level θ too quickly. This means that the measure of the set where it stays close to the cutoff cannot be too small and thus ξ cannot be too large. In particular, following the proof of Proposition 3.4 in Audibert and Tsybakov [2007], $\xi \min(1, \gamma) \leq 1$ is necessary for $\tau(x)$ to cross θ in the interior of the support of the distribution of X, when this has a density bounded above and below away from zero.

The lemma below, which is essentially Lemma 3.1 in Rigollet and Vert [2009] and Theorem 3.1 in Audibert and Tsybakov [2007] adjusted for our setting, shows that if $\widehat{\tau}(x) - \tau(x)$ satisfies an exponential inequality, then the bound on the risk $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\}$ can be sharpened relative to the results presented in Propositions 5 and 6 above. Furthermore, the bound on the risk depends on how fast $\widehat{\tau}(x)$ converges to $\tau(x)$ for values of x near the cutoff $\tau(x) = \theta$.

Lemma 8. *Fix $\eta > 0, \Delta > 0$ and let $D(\eta) = \{x \in \mathcal{X} : |\tau(x) - \theta| \leq \eta\}$. Let a_n, b_n and δ_n be monotonically decreasing sequences. Suppose that*

1. *$b_n \leq c_1 (\log n)^{-1/\kappa_2 - \epsilon}$, with $\epsilon > 0$;*
2. *$a_n \geq c_2 n^{-\mu}$ for some positive constants c_2 and μ, and $a_n \leq b_n$;*

and that the following inequalities hold

$$\mathbb{P}\left(|\widehat{\tau}(x) - \tau(x)| > t\right) \leq c_3 e^{-c_4 (t/a_n)^{\kappa_1}} + c_5 \frac{\delta_n^{1+\xi}}{t^{1+\xi}} \tag{4.2}$$

for all $x \in D(\eta)$ *and* $c_a a_n < t < \Delta$, *and*

$$\mathbb{P}\left(|\widehat{\tau}(x) - \tau(x)| > t\right) \leq c_6 e^{-c_7 (t/b_n)^{\kappa_2}} + c_8 \frac{\delta_n^{1+\xi}}{t^{1+\xi}} \tag{4.3}$$

for all $x \notin D(\eta)$ *and* $c_b b_n < t < \Delta$.

for some constants $c_a, c_b, c_1, \ldots, c_8, \kappa_1$, *and* κ_2. *Then,* $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\} \lesssim a_n^{1+\xi} + (c_5 \vee c_8)\delta_n^{1+\xi} \log n$. *In particular, if* $c_5 = c_8 = 0$, *then* $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\} \lesssim a_n^{1+\xi}$.

The central requirement to apply Lemma 8 is that the estimator $\widehat{\tau}(x)$ must satisfy an exponential inequality. If this is the case, this lemma shows that, provided that $\widehat{\tau}(x)$ converges to $\tau(x)$ at a rate $(\log n)^{-\kappa}$ for some κ on the entire domain, the accuracy for estimating the CATE level set $\Gamma(\theta)$ is entirely determined by the rate for estimating $\tau(x)$ near the level θ. If it is hard to show that the estimator satisfies an exponential inequality, then one can resort to applying Propositions 5 and 6. Because of Lemma 8, we are left with the task of deriving concentration inequalities for $|\widehat{\tau}(x) - \tau(x)|$. We will do that for the case when $\widehat{\tau}(x)$ is either a DR-Learner or an Lp-R-Learner, which may be of independent interest.

4.3.2 Bound on estimation error using a DR-Learner

To start, we consider the DR-Learner proposed and analyzed by Kennedy [2020].

Definition 2 (DR-Learner algorithm based on linear smoothing)**.** Let D^n and Z^n be two independent samples of observations.

1. Using only observations in D^n, construct estimators $\widehat{\pi}(x) = \widehat{\mathbb{P}}(A = 1 \mid X - x)$ and $\widehat{\mu}_a(x) = \widehat{\mathbb{E}}(Y \mid A = a, X = x)$.

2. Using only observations in Z^n, construct

$$\widehat{\tau}(x) = \sum_{i=1}^{n} W_i(x; X^n)\widehat{\varphi}(Z_i), \quad \text{for}$$
$$\widehat{\varphi}(Z_i) = \frac{\{A - \widehat{\pi}(X_i)\}\{Y_i - \widehat{\mu}_A(X_i)\}}{\widehat{\pi}(X_i)\{1 - \widehat{\pi}(X_i)\}} + \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i),$$

some weights $W_i(x; X^n)$ and $X^n \subset Z^n$.

Let us define:

$$S(x;X^n)=\left\{\sum_{i=1}^n W_i^2(x;X^n)\right\}^{1/2}, \quad \widehat{b}(X_i)=\mathbb{E}\{\widehat{\varphi}(Z_i)-\varphi(Z_i)\mid X_i\} \quad \text{and}$$
$$\Delta(x;X^n)=\sum_{i=1}^n W_i(x;X^n)\tau(X_i)-\tau(x)$$

The quantity $\Delta(x;X^n)$ is the smoothing bias (conditional on $X_1,\ldots,X_n$) of the oracle estimator that has access to the true function $\varphi(Z_i)$. The quantity $\widehat{b}(x)$ expresses the bias resulting from having to estimate the nuisance functions. If $\pi(x)$ and $\widehat{\pi}(x)$ are bounded away from zero and one (positivity), it can be shown that

$$|\widehat{b}(x)|\lesssim|\{\pi(x)-\widehat{\pi}(x)\}[\{\mu_1(x)-\widehat{\mu}_1(x)\}+\{\mu_0(x)-\widehat{\mu}_0(x)\}]|$$

Remark 6. A major advantage of the DR-Learner framework, not necessarily based on linear smoothing, is that regressing an estimate of the pseudo-outcome $\widehat{\varphi}(Z)$ on $V=v$ yields an estimate of $\tau(v)=\mathbb{E}\{\mu_1(X)-\mu_0(X)\mid V=v\}$, i.e., the CATE function evaluated at effect modifiers V, which may differ from the covariates X needed to deconfound the treatment-outcome association. This is particularly useful when the dimension of X is much greater than that of V. Thus, if the goal is to compute the upper level sets of $\tau(v)$, with $v\neq x$, thresholding a DR-Learner is an attractive option.

We have the following exponential inequality.

Lemma 9. *Suppose $\widehat{\tau}(x)$ is a DR-Learner defined in* (2)*. Further suppose that*

1. $|\Delta(x;X^n)|\leq c_1 a_n$ *almost surely for a monotonically decreasing sequence a_n and constant c_1;*
2. $\mathbb{E}\{S^p(x;X^n)\}\leq s_n^p$ *for any $p>0$, a monotonically decreasing sequence s_n;*
3. $\mathbb{E}\left|\sum_{i=1}^n W_i(x;X^n)\widehat{b}(X_i)\right|^{1+\xi}\leq\delta_n^{1+\xi}$ *for a monotonically decreasing sequence δ_n;*
4. $\|\widehat{\varphi}-\varphi-\widehat{b}\|_\infty\leq c_2\|\varphi\|_\infty$ *for a constant c_2.*

Then, for any $t\geq 3c_1a_n$, it holds that

$$\mathbb{P}\left(|\widehat{\tau}(x)-\tau(x)|>t\right)\leq 2e^2\exp\left\{-\left(\frac{t}{12(c_2\vee 2)e\|\varphi\|_\infty s_n}\right)^2\right\}+3^{1+\xi}\left(\frac{\delta_n}{t}\right)^{1+\xi}$$

Lemma 9 provides an exponential inequality for the DR-Learner based on linear smoothing, which might be of independent interest. Conditions 1-3 are not really assumptions in the sense that they are simply used to state the inequality in a succint form. Depending on the weights of the linear smoother and the accuracy in estimating the nuisance functions, conditions 1-3

would be satisfied by different sequences a_n, s_n and δ_n. Condition 4 is a mild boundedness assumption. In the following example, we show how Lemmas 8 and 9 can be used to derive an upper bound on $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\}$, where $\widehat{\Gamma} = \{x \in \mathcal{X} : \widehat{\tau}(x) > \theta\}$ for $\widehat{\tau}(x)$ the DR-Learner based on local polynomial second stage regression.

Example 1 (DR-Learner with local polynomials). Suppose that $\tau(x)$ is γ-smooth locally around any $x \in D(\eta)$ in the sense of Definition 1 and it is γ'-smooth for any $x \notin D(\eta)$. Further suppose that $\widehat{\tau}(x)$ is based on local polynomial second stage regression and that all observations are bounded. That is, $W_i(x; X^n)$ are the weights of a local polynomial of degree $p = \lfloor \gamma \rfloor$. The calculations in Tsybakov [2009] (Section 1.6) show that, under mild regularity conditions:

$$S(x; X^n) \lesssim \frac{1}{\sqrt{nh^d}}, \quad T(x; X^n) = \sum_{i=1}^{n} |W_i(x; X^n)| \lesssim 1, \quad \text{and} \quad |\Delta(x; X^n)| \lesssim h^\gamma.$$

for $x \in D(\eta)$. Choosing h of order $n^{-1/(2\gamma+d)}$ yields that there exist constants c_1 and c_2 such that

$$S(x; X^n) \leq c_1 n^{-\gamma/(2\gamma+d)} \quad \text{and} \quad |\Delta(x; X^n)| \leq c_2 n^{-\gamma/(2\gamma+d)}$$

Typically, it will be the case that $W_i(x; X^n) = 0$ if $\|X_i - x\| > h$ so that by Jensen's inequality (since $u \mapsto |u|^{1+\xi}$ is convex):

$$\begin{aligned}
\mathbb{E}\left|\sum_{i=1}^{n} W_i(x; X^n)\widehat{b}(X_i)\right|^{1+\xi} &\leq \mathbb{E}\left[T^{1+\xi}(x; X^n) \cdot \left\{\frac{\sum_{i=1}^{n} |W_i(x; X^n)| \left|\widehat{b}(X_i)\right|}{T(x; X^n)}\right\}^{1+\xi}\right] \\
&\leq \mathbb{E}\left[T^{1+\xi}(x; X^n) \cdot \left\{\frac{\sum_{i=1}^{n} |W_i(x; X^n)| \left|\widehat{b}(X_i)\right|^{1+\xi}}{T(x; X^n)}\right\}\right] \\
&\leq \mathbb{E}\left\{T^{1+\xi}(x; X^n) \sup_{u:\|u-x\|\leq h} \left|\widehat{b}(u)\right|^{1+\xi}\right\} \\
&= \mathbb{E}\left\{T^{1+\xi}(x; X^n)\right\} \mathbb{E}\left\{\sup_{u:\|u-x\|\leq h} \left|\widehat{b}(u)\right|^{1+\xi}\right\} \\
&\lesssim \mathbb{E}\left\{\sup_{u:\|u-x\|\leq h} \left|\widehat{b}(u)\right|^{1+\xi}\right\}
\end{aligned}$$

where the last equality follows because $\widehat{b}(u)$ depends only on the observations in the training

sample, which is independent of X^n. By Lemma 9 and all $t \geq 3c_2 n^{-\gamma/(2\gamma+d)}$ and $x \in D(\eta)$:

$$\mathbb{P}\left(|\widehat{\tau}(x) - \tau(x)| > t\right) \lesssim \exp\left(-Ct^2 n^{-2\gamma/(2\gamma+d)}\right) + t^{-1-\xi}\mathbb{E}\left(\sup_{u:\|u-x\|\leq h} |\{\pi(u) - \widehat{\pi}(u)\}[\{\mu_1(u) - \widehat{\mu}_1(u)\} + \{\mu_0(u) - \widehat{\mu}_0(u)\}]|^{1+\xi}\right)$$

For $x \notin D(\eta)$, we have the same inequality with γ replaced by γ'. Thus, by Lemma 8, we have

$$\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\} \lesssim n^{-(1+\xi)\gamma/(2\gamma+d)} + \delta_n^{1+\xi} \log n$$

where δ_n satisfies

$$\mathbb{E}\left(\sup_{u:\|u-x\|\leq h} |\{\pi(u) - \widehat{\pi}(u)\}[\{\mu_1(u) - \widehat{\mu}_1(u)\} + \{\mu_0(u) - \widehat{\mu}_0(u)\}]|^{1+\xi}\right) \lesssim \delta_n^{1+\xi}.$$

4.3.3 Bound on estimation error using Lp-R-Learners

In this section, we derive an exponential inequality when $\widehat{\tau}(x)$ is the Lp-R-Learner. To describe the Lp-R-Learner estimator, we need to introduce some additional notation. We refer to the original paper Kennedy et al. [2022] for more details. In particular, the authors consider two different parametrizations of the data generating process: one based on (f, π, μ_0, τ), which we consider in our work, and one based on (f, π, μ, τ), where $\mu(x) = \mathbb{E}(Y \mid X = x)$ (see their Section 6). We expect that extending our analysis to cover the latter parametrization is straightforward.

Definition 3 (Lp-R-Learner)**.** Let F denote the CDF of X. For each covariate x_j, let $\rho(x_j) = [\rho_0(x_j), \ldots, \rho_{\lfloor\gamma\rfloor}(x_j)]$ be the first $(\lfloor\gamma\rfloor + 1)$ Legendre polynomials shifted to be orthonormal in $[0, 1]$. That is,

$$\rho_m(x_j) = \sum_{l=1}^{m} \theta_{lm} x_j^l, \text{ for } \theta_{lm} = (-1)^{l+m}\sqrt{2m+1}\binom{m}{l}\binom{m+l}{l}$$

Define $\rho(x)$ to be the tensor product containing all interactions of $\rho(x_1), \ldots, \rho(x_d)$ up to order $\lfloor\gamma\rfloor$. Thus, $\rho(x)$ has length $J = \binom{d+\lfloor\gamma\rfloor}{\lfloor\gamma\rfloor}$ and is orthonormal in $[0, 1]^d$. Finally, define

$\rho_h(x) = \rho(0.5 + (x - x_0)/h)$. The Lp-R-Learner $\widehat{\tau}(x_0)$ is defined as

$$\widehat{\tau}(x_0) = \rho_h^T(x_0)\widehat{Q}^{-1}\widehat{R}, \text{ where } K_h(x) = \mathbb{1}(2\|x - x_0\| \leq h)$$
$$\widehat{Q} = \mathbb{P}_n\left\{\rho_h(X)K_h(X)\widehat{\varphi}_{a1}(Z)\rho_h^T(X)\right\} + \mathbb{U}_n\left\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{a2}(Z_1, Z_2)K_h(X_2)\rho_h^T(X_1)\right\}$$
$$\widehat{R} = \mathbb{P}_n\left\{\rho_h(X_1)K_h(X)\widehat{\varphi}_{y1}(Z)\right\} + \mathbb{U}_n\left\{\rho_h(X_1)K_h(X_1)\widehat{\varphi}_{y2}(Z_1, Z_2)K_h(X_2)\right\}$$
$$\widehat{\varphi}_{a1}(Z) = A\{A - \widehat{\pi}(X)\}$$
$$\widehat{\varphi}_{a2}(Z_1, Z_2) = -\{A_1 - \widehat{\pi}(X_1)\}K_h(X_1)b_h^T(X_1)\widehat{\Omega}^{-1}b_h^T(X_2)A_2$$
$$\widehat{\varphi}_{y1}(Z) = \{Y - \widehat{\mu}_0(X)\}\{A - \widehat{\pi}(X)\}$$
$$\widehat{\varphi}_{y2}(Z_1, Z_2) = -\{A_1 - \widehat{\pi}(X_1)\}b_h^T(X_1)\widehat{\Omega}^{-1}b_h^T(X_2)\{Y_2 - \widehat{\mu}_0(X_2)\}$$
$$b_h(X) = b(0.5 + (x - x_0)/h)\mathbb{1}(2\|x - x_0\| \leq h)$$
$$\widehat{\Omega} = \int_{v\in[0,1]^d} b(v)b^T(v)d\widehat{F}(x_0 + h(v - 0.5))$$

for $b : \mathbb{R}^d \mapsto \mathbb{R}^k$ a basis vector of dimension k that should have good approximating properties for the nuisance function class. The nuisance functions $(\widehat{F}, \widehat{\pi}, \widehat{\mu}_0)$ are computed from a training sample D^n, independent of that used to calculate the empirical and U-statistic measures.

The Lp-R-Learner estimator is tailored to a particular smoothness model, which we describe next and adopt in this section and when discussing minimax optimality.

Definition 4 (Lp-R-Learner smoothness model)**.** Fix γ, γ', α and β. Recall that $D(\eta) = \{x \in \mathcal{X} : |\tau(x) - \theta| \leq \eta\}$, $\eta > 0$. We define $\mathcal{P}$ to be the collection of all distributions satisfying the following conditions:

1. $\tau(x)$ is γ-smooth locally around any $x \in D(\eta)$ in the sense of Definition 1;
2. $\tau(x)$ is γ'-smooth for any $x \notin D(\eta)$;
3. $\mu_0(x) - \widehat{\mu}_0(x)$ is β-smooth and $\pi(x) - \widehat{\pi}(x)$ is α-smooth for any $x \in \mathcal{X}$ [1];
4. $\epsilon \leq \pi(x) \leq 1 - \epsilon$ almost-surely, for some $\epsilon > 0$;
5. The eigenvalues of Q and Ω are bounded above and below away from zero.

Let $s = (\alpha + \beta)/2$ denote the average smoothness of the nuisance functions. Let $T = 1 + d/(4s) + d/(2\gamma)$ and $T' = 1 + d/(4s) + d/(2\gamma')$. For the model described in Definition 4, Kennedy et al. [2022] proved that, under certain regularity conditions, the pointwise risk

[1] In principle, $\mu_0(x)$ and $\pi(x)$ could have different smoothness levels depending on whether $x \in D(\eta)$ or not. This would not complicate the analysis conceptually but it would make the notation more involved. For simplicity, we treat the nuisance functions as having a smoothness level that does not vary across the covariates' space.

satisfies

$$\mathbb{E}|\widehat{\tau}(x)-\tau(x)|^2 \lesssim \begin{cases} n^{-2\gamma/(2\gamma+d)} & \text{if } x \in D(\eta) \text{ and } s \geq \frac{d/4}{1+d/(2\gamma)} \\ n^{-2/T} & \text{if } x \in D(\eta) \text{ and } s < \frac{d/4}{1+d/(2\gamma)} \\ n^{-2\gamma'/(2\gamma'+d)} & \text{if } x \notin D(\eta) \text{ and } s \geq \frac{d/4}{1+d/(2\gamma')} \\ n^{-2/T'} & \text{if } x \notin D(\eta) \text{ and } s < \frac{d/4}{1+d/(2\gamma')} \end{cases}$$

Crucially, Kennedy et al. [2022] shows that these rates are the minimax optimal rates for estimating $\tau(x)$ in this model. Notice that the rate $n^{-2\gamma/(2\gamma+d)}$ is the optimal rate for estimating a d-dimensional, γ-smooth regression function. It is referred to as the *oracle rate* because it is the fastest rate achievable by an infeasible estimator that has access to the true pseudo-outcomes $\varphi(Z_i)$ (see definition 2). In the next section, we show that $\widehat{\Gamma}(\theta)$ based on thresholding the Lp-R-Learner estimator of the CATE is minimax optimal for $\Gamma(\theta)$ in this model as well. We derive the following exponential inequality, which may be of independent interest.

Lemma 10. *Suppose the data generating mechanism satisfies the model described in Definition 4. Let* $dF^*(v) = dF(x_0 + h(v-0.5))$ *and* $\|g\|^2_{F^*} = \int g^2(v)dF^*(v)$. *Further suppose that:*

1. *The quantities* $y^2, \widehat{\pi}^2, \widehat{\mu}_0^2, \|\mu_0 - \widehat{\mu}_0\|_{F^*}, \|\widehat{Q}^{-1} - Q^{-1}\|$ *are all bounded above and* $\|dF/d\widehat{F}\|_\infty$, $\|\widehat{Q}\|$ *and* $\|\widehat{\Omega}\|$ *are bounded above and below away from zero;*
2. $\|dF/d\widehat{F} - 1\|_\infty \{\|\widehat{\pi} - \pi\|_{F^*}(h^\gamma + \|\widehat{\mu}_0 - \mu_0\|_{F^*})\} \lesssim n^{-1/T} \vee n^{-1/T'} \vee n^{-1/(1+d/(2\gamma))}$;
3. *The basis dictionary is suitable for approximating Hölder functions of order* s *in the sense that*

$$\left\| g - b^T \Omega^{-1} \int b(u)g(u)dF^*(u) \right\|_{F^*} \lesssim k^{-s/d}$$

if g *is* s*-smooth.*

Then there exist some constants C, c, c_r *and* Δ *so that, for all* $c_r r_n \leq t \leq \Delta$, *it holds that*

$$\mathbb{P}\left(|\widehat{\tau}(x) - \tau(x)| > t\right) \leq C \exp\left[-cmin\left\{\left(\frac{t}{r_n}\right)^2, \left(\frac{t}{r_n}\right)^{1/2}\right\}\right],$$

where

$$r_n = \begin{cases} n^{-\gamma/(2\gamma+d)} & \text{if } x \in D(\eta) \text{ and } s \geq \frac{d/4}{1+d/(2\gamma)} \\ n^{-1/T} & \text{if } x \in D(\eta) \text{ and } s < \frac{d/4}{1+d/(2\gamma)} \\ n^{-\gamma'/(2\gamma'+d)} & \text{if } x \notin D(\eta) \text{ and } s \geq \frac{d/4}{1+d/(2\gamma')} \\ n^{-1/T'} & \text{if } x \notin D(\eta) \text{ and } s < \frac{d/4}{1+d/(2\gamma')} \end{cases}$$

All the conditions listed in the lemma above are needed in the derivation of the convergence rate (in pointwise RMSE) of $\widehat{\tau}(x_0)$ as proven in Kennedy et al. [2022]. We note that condition 2

requires estimating the covariates' density sufficiently well. We refer the reader to the original paper for a detailed discussion of their interpretation. Next, we use Lemmas 8 and 10 to derive a bound on $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\}$ when $\widehat{\Gamma}$ is estimated using the Lp-R-Learner.

Corollary 1. *Under the setup of Lemma 10, it holds that* $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\} \lesssim r_n^{*1+\xi}$*, where*

$$r_n^* = \begin{cases} n^{-\gamma/(2\gamma+d)} & \text{if } s \geq \frac{d/4}{1+d/(2\gamma)} \\ n^{-1/T} & s < \frac{d/4}{1+d/(2\gamma)} \end{cases}$$

Proof. It is sufficient to apply Lemma 8 with $c_5 = c_8 = 0$. □

In the next section, we show that $r_n^{*1+\xi}$ is also the minimax rate for estimating the level set $\Gamma(\theta)$ in the model described by Definition 4 when the risk is $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\}$. We will derive the lower bound on the minimax risk in the low-smoothness regime ($s < \frac{d/4}{1+d/(2\gamma)}$), with the understanding that a similar construction yields the appropriate lower bound in the high-smoothness regime ($s \geq \frac{d/4}{1+d/(2\gamma)}$).

4.4 Minimax lower bound

Here, the goal is to lower bound the minimax risk, defined as:

$$\inf_{\widehat{\Gamma}} \sup_{p \in \mathcal{P}} \mathbb{E}_p\{d_H(\widehat{\Gamma}, \Gamma_p)\} = \inf_{\widehat{\Gamma}} \sup_{p \in \mathcal{P}} \mathbb{E}_p \left\{ \int_{\widehat{\Gamma} \Delta \Gamma_p} |\tau_p(x) - \theta| f_p(x) dx \right\}$$

where $\mathcal{P}$ is a set of distributions compatible with our assumptions. Calculating the minimax risk for estimating a given parameter is important for at least two reasons. First, it serves as a benchmark for comparing estimators. In particular, if the lower bound on the minimax risk matches the rate of an available estimator, then one can conclude that there is not another estimator that can improve upon the minimax optimal one, at least in terms of a worst-case analysis, without introducing additional assumptions. Conversely, if there are no estimators attaining a rate that matches the minimax lower bound, then one has to either construct a better estimator or tighten the upper or lower bound. In our setting, we show that a valid lower bound matches the upper bound of Corollary 1 up to constants, which therefore establishes the minimax rate for estimating Γ under the loss $d_H(\widehat{\Gamma}, \Gamma)$ in model 4. Furthermore, a tight minimax lower bound is helpful because it precisely characterizes the difficulty in estimating this parameter.

Theorem 4. *Suppose that* $\xi\gamma \leq d$*. Under assumption 4 and the smoothness model defined in 4, then*

$$\inf_{\widehat{\Gamma}} \sup_{p \in \mathcal{P}} \mathbb{E}_p\{d_H(\widehat{\Gamma}, \Gamma_p)\} \gtrsim r_n^{*1+\xi}, \text{ where } r_n^* = n^{-1/T} \text{ and } T = 1 + d/(4s) + d/(2\gamma).$$

when $s < \frac{d/4}{1+d/(2\gamma)}$.

As shown in Kennedy et al. [2022], the rate r_n^* is the minimax rate for estimating $\tau(x)$ (at a point and under the square loss) in the smoothness model encoded in Definition 4 in the low smoothness regime. Our result shows that the same estimator can be thresholded to yield an optimal estimator of the CATE level sets. The result in Theorem 4 aligns with that of Rigollet and Vert [2009], where r_n^* is replaced by the optimal minimax rate for estimating a γ-smooth density on a d-dimensional domain, i.e. $n^{-\gamma/(2\gamma+d)}$ (on the root-mean-square error scale).

The proof of Theorem 4 combines the construction of Rigollet and Vert [2009], Kennedy et al. [2022] and Assouad's lemma (specifically, we rely on Theorem 2.12 in Tsybakov [2009]). To derive a lower bound on the risk of some estimator, one needs to construct two worst-case distributions Q_1 and Q_2 such that Q_1 and Q_2 are similar enough so that one cannot perfectly determine whether a sample is from Q_1 or Q_2 but, at the same time, the value of the parameter at Q_1 is maximally separated from that at Q_2. To construct Q_1 and Q_2 one typically carefully designs fluctuations around the quantities that need to be estimated, in our case $\pi(x)$, $\mu_0(x)$ and $\tau(x)$. As shown in Figure 4.2, we place bumps on these functions of particular heights depending on the level of smoothness. Our construction extends that of Kennedy et al. [2022], which is localized in a neighborhood around $x = x_0$, to the entire domain of X. In particular, it can be used to show that the rate obtained in Kennedy et al. [2022] for the pointwise risk is also the minimax rate for the integrated risk $\int\{\widehat{\tau}(x) - \tau(x)\}^2 dF(x)$, which might be of independent interest. We refer to Kennedy et al. [2022] for additional details. Finally, the lower bound from Theorem 4 applies only to the case $\xi\gamma \leq d$. This condition also appears in the work of Audibert and Tsybakov [2007] and the more stringent condition $\xi\gamma \leq 1$ appears in the lower bound construction of Rigollet and Vert [2009]. To the best of our knowledge, deriving a tight lower bound without this condition is still an open problem.

4.5 Inference

In this section, we discuss a simple way to carry out inference when a DR-Learner is thresholded to estimate $\Gamma(\theta)$. Inspired by Mammen and Polonik [2013], we propose constructing two sets $\widehat{C}_l$ and $\widehat{C}_u$ of the form

$$\widehat{C}_l = \left\{x \in \mathbb{R}^d : \widehat{\sigma}^{-1}(x)\{\widehat{\tau}(x) - \theta\} > c_n(1-\alpha)\right\}$$
$$\widehat{C}_u = \left\{x \in \mathbb{R}^d : \widehat{\sigma}^{-1}(x)\{\widehat{\tau}(x) - \theta\} \geq -c_n(1-\alpha)\right\},$$

where $\widehat{\sigma}(x)$ is an estimate of the standard deviation of $\widehat{\tau}(x)$ and $c_n(1-\alpha)$ is some carefully chosen cutoff, depending on the $1-\alpha$ confidence level. The rationale for constructing such sets is outlined in the following lemma, which is written for level sets of some arbitrary function $f(x)$.

Lemma 11. *Let $\overline{\Lambda}(\theta) = \{x \in \mathcal{X} : f(x) \geq \theta\}$ and $\Lambda(\theta) = \{x \in \mathcal{X} : f(x) > \theta\}$. Let $\widehat{f}(x)$ be an estimator of $f(x)$ with some standard deviation $\widehat{\sigma}(x)$. Define the t-statistic $t_n =$*

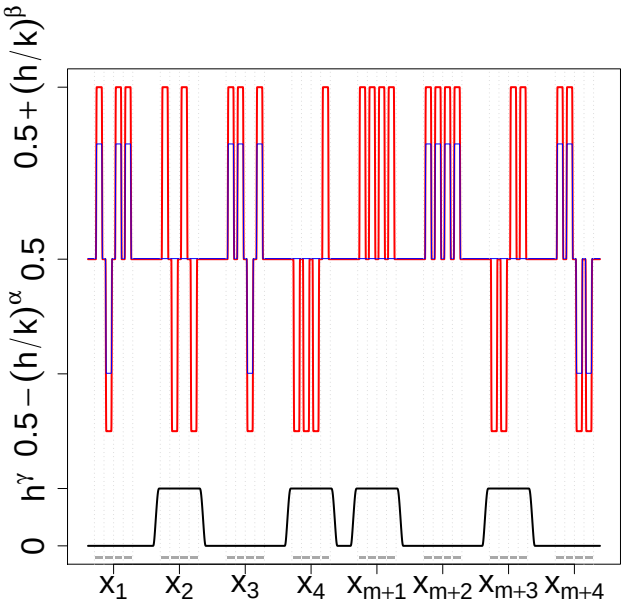


Figure 4.2: Lower bound construction for the case $d = 1$, $\theta = 0$ and $\alpha \geq \beta$. The solid black curve represents $\tau(x)$, the red curve represents $\mu_0(x)$ while the blue curve represents $\pi(x)$. Notice that if $\tau(x) > 0$ then $\pi(x) = 1/2$, whereas $\mu_0(x)$ is always fluctuated.

$\{\widehat{\sigma}(x)\}^{-1}\{\widehat{f}(x) - f(x)\}$. *Finally, define*

$$\widehat{C}_l = \{x \in \mathbb{R}^d : \{\widehat{\sigma}(x)\}^{-1}\{\widehat{f}(x) - \theta\} > t\}$$
$$\widehat{C}_u = \{x \in \mathbb{R}^d : \{\widehat{\sigma}(x)\}^{-1}\{\widehat{f}(x) - \theta\} \geq -t\}$$

Then, it holds that

$$\mathbb{P}\left(\overline{\Lambda}(\theta) \subseteq \widehat{C}_u \text{ and } \widehat{C}_l \subseteq \Lambda(\theta)\right) \geq \mathbb{P}\left(\|t_n\|_\infty \leq t\right).$$

Proof. Let x_0 be any member of $\overline{\Lambda}(\theta)$ and notice the following chain of implications

$$\|t_n\|_\infty \leq t \implies \widehat{\sigma}^{-1}(x_0)\{\widehat{f}(x_0) - f(x_0)\} \geq -t \implies \widehat{\sigma}^{-1}(x_0)\{\widehat{f}(x_0) - \theta\} \geq -t$$

because $f(x_0) \geq \theta$. This means that $x_0 \in \widehat{C}_u$ so that we conclude that $\mathbb{P}\left(\overline{\Lambda}(\theta) \subseteq \widehat{C}_u\right) \geq \mathbb{P}\left(\|t_n\|_\infty \leq t\right)$.

Similarly, let x_0 be any member of $\widehat{C}_l$ and notice that

$$\begin{aligned}\|t_n\|_\infty \leq t &\implies \widehat{\sigma}^{-1}(x_0)\{\widehat{f}(x_0) - \theta\} + \widehat{\sigma}^{-1}(x_0)\{\theta - f(x_0)\} \leq t \\ &\implies \widehat{\sigma}^{-1}(x_0)\{\theta - f(x_0)\} < 0\end{aligned}$$

because $\widehat{\sigma}^{-1}(x_0)\{\widehat{f}(x_0) - \theta\} > t$. Thus, $x_0 \in \Lambda(\theta)$ so that

$$\mathbb{P}\left(\widehat{C}_l \subseteq \Lambda(\theta)\right) \geq \mathbb{P}\left(\|t_n\|_\infty \leq t\right).$$

□

In light of Lemma 11, $\widehat{C}_l$ and $\widehat{C}_u$ act as $1-\alpha$ lower and upper confidence sets for $\Lambda(\theta)$ as long as $\|t_n\|_\infty \leq t$ with probability at least $1-\alpha$. Thus, constructing $\widehat{C}_l$ and $\widehat{C}_u$ to cover $\Gamma(\theta)$ effectively reduces to the problem of constructing uniform confidence bands around $\widehat{\tau}(x)$.

Constructing confidence regions for level sets based on the supremum of the function that is being thresholded is an example of confidence sets based on "vertical variation." An alternative route would be to construct confidence regions based on "horizontal variation," an example of which would be a confidence region based on approximating the distribution of the Hausdorff distance between the estimated set and the true set. We leave this for future work and refer to Qiao and Polonik [2019] and Chen et al. [2017] for more details regarding the differences between these approaches in the context of density estimation.

Semenova and Chernozhukov [2021] establish uniform confidence bands for a DR-Learner estimator of the CATE such that the second-stage regression is carried out via orthogonal series regression. One can therefore leverage their results (Theorem 3.5) to construct confidence sets for the CATE level sets based on Lemma 11. [2] Finally, in the context of dose-response estimation, Takatsu and Westling [2022] construct uniformly valid confidence bands for second-stage local linear smoothers where the outcome is estimated in a first-step. We expect their results to be useful in the setting considered here as well. We plan on including a more precise result on uniform inference for DR-Learners in an updated version of this work.

4.6 Small simulation experiment

The goal of this section is to evaluate the performance of the estimators and investigate the role of various aspects of the data generating processes in finite samples. First, we study the impact of the nuisance functions' estimation step on the coverage of the CATE upper level set. Our estimator of the upper level set will consist of thresholding a DR-Learner estimator of the CATE based on a parametric second-stage linear regression. Based on Lemma 9 and Example

[2] Estimating the quantile of $\|t_n\|_\infty$ typically requires that the smoothing bias for estimating the CATE converges to zero faster than the standard error (e.g., see condition (iv) in Theorem 3.5 in Semenova and Chernozhukov [2021]). This condition can be challenging to guarantee in applications, but we note that it is not required if one changes the target of inference to upper level sets of a "smoothed version" of the CATE function, i.e. a modified CATE function that can be estimated without smoothing bias. See also Chen et al. [2017].

1, we expect the performance of our estimator to deteriorate significantly when the product of the nuisance functions' error is greater than $o_{\mathbb{P}}(n^{-1/2})$.

Next, we investigate the impact of the parameters governing the margin assumption 4 on the error in estimating the upper level set. To simplify the simulation settings, we consider a smooth CATE with bounded density so that $\xi = 1$ in Assumption 4 holds and we increase the constant c_0 on the right hand side of the margin condition inequality. We expect that the larger the region of the covariates' space where the CATE is close to the threhsold the harder the estimation problem becomes.

In all simulation scenarios, we define $\theta = 0$, the sample size $n = 1000$, the number of bootstrap replications used in constructing the confidence regions $B = 10^5$ and the number of simulations $I = 500$. We enforce consistency by setting $Y = AY^1 + (1 - A)Y^0$. We approximate the space $[-1, 1]^2$ by a grid of points $x_1, \ldots, x_m$, which are equally spaced points (x_{1i}, x_{2j}) for $1 \leq i, j \leq 50$. Uniform coverage is computed relative to this approximation.

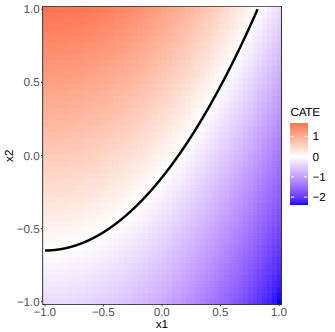
Setup 1A: Impact of the nuisance functions' estimation step. We generate data from the following model:

$$X_i \overset{iid}{\sim} \text{Unif}(-1, 1), \quad A \mid X_1, X_2 \sim \text{Bin}(\text{expit}(-1 + X_1 + X_2)),$$
$$Y^1 \mid X_1, X_2 \sim N(0.15 - X_1 - 0.5X_1^2 + X_2, 1), \quad Y^0 \mid X_1, X_2 \sim N(0, 1).$$

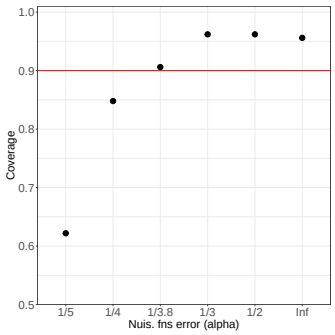
Notice that $\tau(x) = 0.15 - x_1 - 0.5x_1^2 + x_2$, which we assume is correctly specified in the second-stage regression. However, we construct the nuisance functions estimators $\widehat{\pi}$, $\widehat{\mu}_a$ by injecting Gaussian noise of order $n^{-1/c}$, for $c = \{0, 2, 3, 3.8, 4, 5\}$ in the true functions. For example, $\widehat{\pi}(x) = \text{expit}(x^T\widehat{\beta})$, where $\widehat{\beta} = [-1\ 1\ 1]^T + \mathcal{N}_3(n^{-1/c}, n^{-1/c}I_3)$. The case $c = 0$ refers to the case where we do not inject any noise. Figure 4.3a represents the simulation setup; the black solid line denotes the set of covariates' values where the CATE is zero.

Setup 1B: Impact of the parameters governing the margin assumption 4. We generate data as in Setup 1A except that $Y^1 \mid X_1, X_2 \sim N(\kappa(0.15 - X_1 - 0.5X_1^2 + X_2), 1)$, where $\kappa = \{0.1, 0.5, 1, 5, 10\}$. The parameter κ is meant to govern the size of the set $\{x \in \mathcal{X} : |\tau(x)| \leq \epsilon\}$ for some fixed ϵ; the smaller κ the larger this set is. We thus expect the performance of our estimator to deteriorate as κ decreases. To isolate the impact of varying κ on the performance of the estimators, we use the true nuisance functions, instead of the estimated ones, in the construction of the pseudo outcome. In other words, we gauge the impact of κ on the oracle estimator. We compute a monte-carlo approximation to $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\}$.

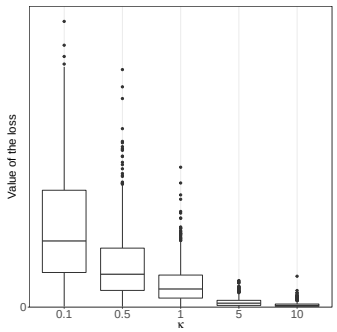
As shown in Figure 4.3b, in agreement with our theoretical results, the coverage of the CATE upper level sets starts to deteriorate as soon as the product of the errors in estimating the nuisance functions equals or exceeds the rate $n^{-1/4}$. Furthermore, Figure 4.3c shows simulation evidence that the estimation error as measured by the risk $\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\}$ decreases, i.e. the estimation problem becomes easier, if the size of covariates' space where the CATE is close to the level decreases. This too is in agreement with the results from the previous sections.



(a) True CATE (from Setup A, $\kappa = 1$) with corresponding level set $\chi(0) = \{x \in [-1,1]^2 : \tau(x) = 0\}$ (black line).



(b) Empirical coverage of $\Gamma(0)$ under setup A as a function the nuisance estimators' accuracy (c in n^{-c}).



(c) Value of the loss $d_H(\widehat{\Gamma}, \Gamma)$ as a function of κ in Setup B.

Figure 4.3: Simulation results

4.7 Data Analysis

Partial colon removal, also known as partial colectomy, is a medical procedure where a surgeon removes the diseased portion of the patient's colon and a small portion of surrounding healthy tissue. A partial colectomy serves as a treatment for various conditions including Crohn's disease, ulcerative colitis, and colon cancer. Surgery for appendicitis can be done in two ways. Traditionally, partial colectomy is done via open surgery (OS), which requires a long incision in the abdomen to gain access to the colon. The primary alternative to open surgery is laparoscopic surgery (LS) which is a surgical technique that uses a small incision and small narrow tubes. The surgeon pumps carbon dioxide through the tubes to inflate the organs and create more space for the procedure. Surgical instruments are inserted and used to remove part of the colon. LS is a minimally invasive colectomy and is designed to help patients recover more quickly and experience fewer surgical complications.

LS for partial colectomy has been widely evaluated in randomized controlled trials, observational studies and meta-analyses [Kannan et al., 2015, Kemp and Finlayson, 2008, Varela et al., 2008, Wu et al., 2022, 2010]. Across these various types of studies, results indicate that LS leads to better patient outcomes including lower morbidity and lower complications. However, it is also likely that the effect of LS varies from patient to patient. More specifically, there may be some patients for whom LS is particularly beneficial, and there may be other patients for whom it is harmful or ineffective. As an empirical application, we use level sets to characterize optimal treatment for LS for partial colectomy. In our analysis, we use a large observational data set and exploit the large sample size and rich set of covariates to better detect whether the effects of LS vary systematically with key patient characteristics.

We use a data set that merges the American Medical Association (AMA) Physician Masterfile with all-payer hospital discharge claims from New York, Florida and Pennsylvania in 2012-2013. The data include patient sociodemographic and clinical characteristics including indicators

for frailty, severe sepsis or septic shock, and 31 comorbidities based on Elixhauser indices [Elixhauser et al., 1998]. The data also include information on insurance type. Our primary outcomes are indicator variables for mortality and complications. In our data, there are 46,506 patients that underwent a partial colectomy. Among these patients, 20,133 underwent LS and 26,373 underwent OS.

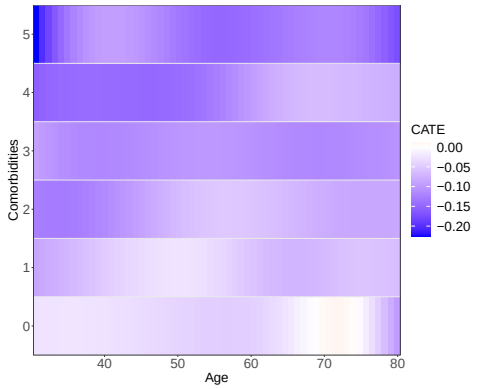
In Figure 4.4, we report the results from a very preliminary data analysis we have conducted. We plan on reporting our final data analysis in an updated version of this paper. The figure shows our DR-Learner estimate of the CATE function defined in terms of two effect modifiers, an aggregate measure of comorbidity and age. Age is approximately continuous ranging from 18 to 102 years old, whereas the measure of comorbidity is ordinal taking values in $0, 1, \ldots, 7, 8+$. The outcome is a binary indicator for whether a set of complications occured. To create Figure 4.4, we restrict the range of comorbidities to be between 0 and 5 and the range of age to be between 30 and 80. In the rest of the covariates' space, we observe too few data points. We deconfound the treatment / outcome association using all pre-treatment variables available. To estimate the nuisance functions, we use Random Forests implemented in the `ranger` R package with default parameters. We then estimate the CATE with a DR-Learner where the second stage regression is a spline regression with six degrees of freedom as implemented in the `splines` R package. We construct the confidence regions using the method of approximating the distribution of $\sup_x |\widehat{\tau}(x) - \tau(x)|$ described in Semenova and Chernozhukov [2021].

As shown in Figure 4.4a, our estimates of the CATE are negative in most of the covariates' space. This is consistent with the idea that laparoscopic surgery, being minally invasive, reduces the risk to develop complications. From this preliminary analysis, and in particular from Figure 4.4b, it appears that laparoscopic surgery significantly decreases the chance of complications for units with an average number of comorbities across many age groups, for units with no comorbidities and age between 50 and 60 and for units with a relatively large number of comorbidities 5 (4) and roughly age between 55 and 67 (51 and 63). Notice that these regions make up the complement of $\widehat{C}_u$ (with $\theta = 0$) as defined in Section 4.5. Therefore, their union is a region that, with high probability, is contained in, and thus potentially smaller than, the true region where the CATE is negative.

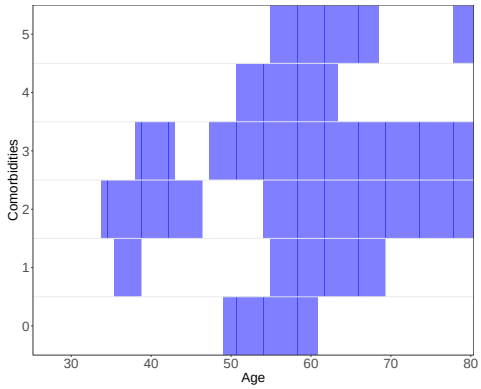
4.8 Conclusions

In this work, we have studied the convergence rates for estimating the upper level sets of the conditional average treatment effect (CATE). We have provided upper bounds on the error in estimating this parameter when either DR-Learners or Lp-R-Learners of the CATE are thresholded to yield estimators of the CATE level sets. Furthermore, we have shown that the estimator based on thresholding the Lp-R-Learner is minimax optimal in a particular smoothness model that allows the CATE and the nuisance functions to have different smoothness levels. We have also discussed a straightforward method to construct upper and lower confidence regions for the upper level set.

There are many questions that remain to be investigated. First, implementing the minimax



(a) Estimates of the CATE as a function of age and comorbidities.



(b) The blue regions represent covariates' values where, with high probability, the CATE function is negative.

Figure 4.4: Data analysis results

optimal Lp-R-Learner estimator of the CATE presents a few challenges. For example, it would be very useful to study how to choose the right values for the tuning parameters that would adapt to the unknown smoothness of the data generating process. In addition, when the covariates' dimension is large, this estimator requires substantial computational power. Second, our construction used to derive the minimax rates requires that the product of the parameter ξ governing the margin condition times the smoothness γ of the CATE is less than the dimension of the covariates. Establishing minimax optimality without imposing this assumption remains an open problem.

It would also be of interest to consider the estimation of related parameters. For example, one could estimate 1) the $\mathbb{P}_X$ measure of the CATE upper level set, which could potentially be estimated with even more precision than the upper level set and the CATE itself, 2) the boundary level set at $\theta = 0$, as well as 3) the CATE upper level sets under additional structural constraints, e.g. in cases where the covariates take values on a lower-dim manifold in $\mathbb{R}^d$.

Finally, an important avenue for future work is to consider estimators of CATE upper level sets that are based on empirical risk minimization, as opposed to the one we have considered in this work that consist of simply thresholding estimators of the CATE functions. This would naturally allow the user to pre-specify a family of candidate upper level sets, which can be chosen sufficiently regular, e.g. hyper-rectangles, to have a natural interpretation in the context of the application considered.

4.9 Acknowledgments

MB thanks Profs. Abhishek Ananth, Alejandro Sanchez Becerra, Ruoxuan Xiong, and Miles Lopes, Tudor Manole, and the participants at the QTM Seminar at Emory University for very helpful discussions.

Chapter 5

Fast convergence rates for dose-response estimation

This chapter is taken from my work supervised by Edward H. Kennedy, which can be found on arXiv [Bonvini and Kennedy, 2022].

5.1 Introduction

5.1.1 Notation & setup

Continuous or multi-valued treatments occur often in practice; time, distance traveled, or dosage of a drug are common examples. We study the problem of estimating the effect of a continuous treatment $A \in \mathcal{A} \subset \mathbb{R}$ on an outcome $Y \in \mathcal{Y} \subset \mathbb{R}$. Within the potential outcomes framework [Rubin, 1974], this effect is defined as the expectation of the potential outcome Y^a, which is the outcome observed if the subject takes treatment level $A = a$. In other words, the estimand represents the average outcome if everyone in the population had taken treatment level a. Because A is continuous, $\mathbb{E}(Y^a)$ is a curve, often referred to as the *dose-response function* (DRF). Under standard assumptions (see e.g. Kennedy et al. [2017]), the DRF takes the form of a partial mean:

$$\theta(t) = \mathbb{E}\{\mathbb{E}(Y \mid A = t, X)\} = \int \mathbb{E}(Y \mid A = t, X = x) d\mathbb{P}(x)$$

where $X \in \mathcal{X} \subset \mathbb{R}^d$ denotes measured confounders. Let $Z = (Y, A, X)$ be distributed according to some distribution $\mathbb{P}$ with density p with respect to the Lebesgue measure. The goal of this paper is to discuss new ways of estimating $\theta(a)$ using n iid copies of Z, which yield strong error guarantees, under weaker conditions, and fast rates of convergence. To simplify

the notation we define:

$$p(u) = \frac{d}{du}\mathbb{P}(U \leq u), \quad \pi(a \mid x) = \frac{p(a,x)}{p(x)}, \quad \mu(a,x) = \mathbb{E}(Y \mid A = a, X = x),$$

and $w(a,x) = p(a)/\pi(a \mid x)$. That is, $p(u)$ is the density of U at $U = u$, $\mu(a,x)$ is the outcome regression, and $\pi(a \mid x)$ is the conditional density of A given $X = x$. We will sometimes denote all the nuisance functions by $\eta = \{p(a), p(x), \mu(a,x), \pi(a \mid x)\}$. With this notation, we have

$$\theta(t) = \mathbb{E}\{\mu(t,X)\} = \mathbb{E}\{w(t,X)Y \mid A = t\}$$

For a kernel function $K(u)$, we let $K_{ht}(a) = h^{-1}K((a-t)/h)$. We use the notation $\mathbb{P}\{g(Z)\} = \int g(z)d\mathbb{P}(z)$ and $\mathbb{P}_n\{g(Z)\} = n^{-1}\sum_{i=1}^n g(Z_i)$ to denote means (given g) and sample means. Further, we let $\|f\|^2 = \int f^2(z)d\mathbb{P}(z) = \mathbb{P}\{f^2(Z)\}$ to denote the squared $L_2(\mathbb{P})$ norm.

Throughout the paper, we will rely on the following assumptions. Additional assumptions will be introduced as needed.

1. Positivity: $\pi(a \mid x)$ and its estimator $\widehat{\pi}(a \mid x)$ are bounded above and away from zero for all $a \in \mathcal{A}$ and $x \in \mathcal{X}$;
2. Boundedness: Y, A, $\widehat{\mu}(a,x)$ are uniformly bounded.

Notice that Positivity is enough to define $\theta(a) = \int \mu(a,x)d\mathbb{P}(x)$, but not enough to interpret $\theta(a)$ as the dose-response curve. To interpret $\theta(a)$ as the effect of A on Y, one needs to impose additional causal assumptions such as $Y^a \perp\!\!\!\perp A \mid X$ and $A = a \implies Y^a = Y$, i.e., no unmeasured confounding and consistency (e.g. no interference). This paper is about estimating $\theta(a)$, regardless of its interpretation, and we refer to Kennedy et al. [2017] and reference therein for more details on identification.

Finally, our focus will be on estimation of the dose-response in nonparametric models where the dose-response itself and the nuisance functions possess varying degrees of smoothness. In particular, we will distinguish between the smoothness levels of the dose response $a \mapsto \theta(a)$, the conditional density of the treatment given the measured confounders $(a,x) \mapsto \pi(a \mid x)$, and the outcome regression $(a,x) \mapsto \mu(a,x)$. We will further refine this distinction when introducing the m^{th}-order estimator in the sense that we will consider models where $a \mapsto \mu(a,x)$, $x \mapsto \mu(a,x)$, $a \mapsto \pi(a \mid x)$ and $x \mapsto \pi(a \mid x)$ may have different smoothness levels. Note that it is reasonable to expect the smoothness of $a \mapsto \mu(a,x)$ to match that of $a \mapsto \theta(a) = \int \mu(a,x)d\mathbb{P}(x)$ in most applications.

5.1.2 Literature review

Crucially, because A is continuous, the parameter $\theta(t)$ cannot be estimated at $\sqrt{n}$ rates in nonparametric models. Informally, in order to see this, notice that we can write $\theta(t) = \mathbb{E}\{w(t,X)Y \mid A = t\}$ and thus even if $w(t,X)Y$ was fully observed (e.g., in a randomized experiment), the best convergence rate attainable would be that of nonparametric regression.

In fact, in order to compare the performances of different estimators, it is useful to establish the regimes where they behave like the oracle estimator that has access to $w(t, X)Y$ and can regress it on A.

Definition 5 (Oracle rate)**.** Given an iid sample $Z_1, \ldots, Z_n$, let $\widetilde{\theta}(\cdot)$ be the (infeasible) estimator regressing $w(t, X)Y$ on A and let r_n be its error under some loss, e.g., the square loss at a point: $\mathbb{E}[\{\widetilde{\theta}(t) - \theta(t)\}^2]$. We refer to r_n as the *oracle rate*.

In nonparametric models that admit slow rates of convergence, two commonly employed strategies are either 1) to specify a *marginal structural model* $\theta(t) = m(t; \beta)$ [Robins et al., 2000] or 2) to change the target of inference from $\theta(t)$ to a *projection* of $\theta(t)$ onto a finite-dimensional model $g(t; \beta)$. We refer to Neugebauer and van der Laan [2007] and Ai et al. [2018] for discussions of efficient estimation in the latter case.

Another approach for estimation of dose-responses is to impose some nonparametric, structural assumptions on the curve itself. For instance, if it is known that the treatment cannot harm the patients, one may impose a monotonicity assumption [Westling and Carone, 2020, Westling et al., 2020]. Yet another approach is to choose a candidate estimator of $\theta(a)$ that minimizes a good estimate of the risk. The key insight is that, while $\theta(a)$ is generally not estimable at $\sqrt{n}$-rates, the integrated risk of a candidate estimator is [Díaz and van der Laan, 2013b, Van der Laan et al., 2003].

In the context of nonparametric estimation, Newey [1994] derives sufficient conditions under which a two-stage kernel estimator of $\theta(t)$ is asymptotically normal and unbiased. Their estimator is of the plug-in variety and takes the form $\widehat{\theta}(a) = n^{-1} \sum_{i=1}^{n} \widehat{\mu}(t, X_i)$, where $\widehat{\mu}(t, x)$ is a kernel-smoothed estimate of $\mu(t, x)$ depending on some bandwidth h. To achieve $\sqrt{nh}$-consistency and asymptotic unbiasedness, $\widehat{\mu}(t, x)$ has to be *undersmoothed*; i.e., h has to be chosen smaller than that minimizing the asymptotic mean-square-error of $\widehat{\mu}(t, x)$. Choosing the right amount of undersmoothing presents challenges in practice; see e.g., Section 5.7 in Wasserman [2006]. Starting from the estimator considered in Newey [1994], Flores [2007] develops plug-in-type estimators of the maximum of $\theta(a)$ and the value of a at which the maximum is attained. Galvao and Wang [2015] study estimation and testing of continuous treatment effects in general settings using inverse-probability-weighted estimators. Singh et al. [2020] analyze plug-in-type estimators of general causal functions based on reproducing kernel methods.

There exists another representation of the DRF that plays an important role in developing efficient estimators:

$$\theta(a) = \mathbb{E}\{\varphi(Z) \mid A = a\}, \text{ where } \varphi(Z) = w(A, X)\{Y - \mu(A, X)\} + \int \mu(A, x) d\mathbb{P}(x)$$

This representation motivates estimators that regress the *pseudo-outcome* $\varphi(Z)$ onto A. Because this pseudo-outcome depends on unknown nuisance functions, it needs to be estimated from the data; thus we refer to the regression of $\varphi(Z)$ on A as a second-stage regression. The crucial point is that $\varphi(Z)$ is such that an estimated regression $\widehat{\mathbb{E}}_n\{\widehat{\varphi}(Z) \mid A = a\}$ can behave like the

oracle $\widehat{\mathbb{E}}_n\{\varphi(Z) \mid A = a\}$ even when $\widehat{\varphi}(Z)$ converges to $\varphi(Z)$ at a rate that is slower than the convergence of $\widehat{\mathbb{E}}_n\{\varphi(Z) \mid A = a\}$ to $\theta(a)$. We conclude this section with a review of the use of this pseudo-outcome in the estimators proposed in Kennedy et al. [2017], Semenova and Chernozhukov [2017] and Colangelo and Lee [2020], as they are the ones most similar to the estimators considered in this article.

5.1.3 Review of existing doubly-robust estimators

In this section, we review a few estimation strategies that are doubly-robust and yield fast convergence rates in the sense that the upper bound on the risk is of the form: "oracle rate + second order, doubly-robust remainder terms," and thus yield consistent estimators when either $w(a, x)$ or $\mu(a, x)$, but not necessarily both, are consistently estimated. This is analogous to the case of treatment effects defined by categorical treatments, whereby estimators based on influence functions are doubly-robust and enjoy second-order error terms.

The estimators proposed in Semenova and Chernozhukov [2017] and Kennedy et al. [2017] are based on regressing an estimate of $\varphi(Z)$ onto A. The quantity $\varphi(Z)$ has a doubly robust remainder error, or equivalently, satisfies a Neyman-orthogonality condition in the sense that, for $\eta = \{p(A), \pi(A \mid X), \mu(A, X)\}$:

$$\partial r \mathbb{E}\{\varphi(Z; \eta_0 + r(\overline{\eta} - \eta)) \mid A = a\}|_{r=0} = 0 \text{ for all } a, \overline{\eta}.$$

This implies that the loss $\{\varphi(Z; \eta) - \theta(A)\}^2$ is universally Neyman-orthogonal in the sense that

$$\begin{aligned}
&\partial r_2 \partial r_1 \mathbb{E}[\varphi(Z; \eta_0 + r_2(\eta - \eta_0)) - \theta(A) - r_1\{\theta(A) - \overline{\theta}(A)\}]^2|_{r_1=r_2=0} \\
&= -2 \int \partial r \mathbb{E}\{\varphi(Z; \eta_0 + r(\eta - \eta_0)) \mid A = a\}_{r=0}\{\theta(a) - \overline{\theta}(a)\} d\mathbb{P}(a) \\
&= 0
\end{aligned}$$

for any $\theta, \overline{\theta}$.

Constructing estimators satisfying Neyman-orthogonality conditions has a long history in Statistics, albeit under different names. For example, in functional estimation, and, in particular, estimation of average treatment effects, estimators that are "Neyman-orthogonal," "bias-corrected," "augmented-inverse-probability-weighted," or, more generally, constructed according to the "double machine learning" framework are all based on first-order functional Taylor expansions, also known as von-Mises expansions [Kennedy, 2022]. In fact, underlying Neyman orthogonality is a first-order expansion of the target estimand $\psi(\mathbb{P})$, viewed as a function of the unknown distribution $\mathbb{P}$, around an estimator $\widehat{\mathbb{P}}$ of $\mathbb{P}$. If the derivative term, say $\psi'(\mathbb{P} - \widehat{\mathbb{P}}; \widehat{\mathbb{P}})$, exists then the estimator consisting of the (estimated) derivative term plus the initial estimator $\psi(\widehat{\mathbb{P}})$ should exhibit second-order error rates. For smooth functionals, the derivative term can be written as $\psi'(\mathbb{P} - \widehat{\mathbb{P}}; \widehat{\mathbb{P}}) = \int \phi(z; \widehat{\mathbb{P}}) d\mathbb{P}(z)$, where $\phi(z)$ is the influence function and is mean-zero. For more complex parameters, such as the dose-response curve, this representation is generally not possible. However, one may try to express the derivative as

an integral with respect to the conditional distribution of the observations given, for example, the treatment.

Kennedy et al. [2017] show that, when the second stage regression $\widehat{\mathbb{E}}_n(\widehat{\varphi}(Z) \mid A = a)$ is a local linear regression, then the oracle rate (the rate achievable if $\varphi(Z)$ was fully observed as defined in Definition 5) is attained as long as

$$\sup_{a:|a-t|\leq h} \|\widehat{\pi}(a \mid X) - \pi(a \mid X)\| \|\widehat{\mu}(a, X) - \mu(a, X)\| = o_{\mathbb{P}}(1/\sqrt{nh}), \tag{5.1}$$

where h is the bandwidth used in the second-stage regression. A similar requirement appears in Colangelo and Lee [2020]. Notice that this error term is *second order*, as it is a product of errors. It also reveals the *double-robustness* property of $\varphi(Z)$: consistency of $\widehat{\mathbb{E}}_n\{\widehat{\varphi}(Z) \mid A = a\}$ requires consistency of either $\pi(A \mid X)$ or $\mu(A, X)$ but not necessarily both.

Semenova and Chernozhukov [2017] studies an estimator of the same form $\mathbb{P}_n^*\{\widehat{\varphi}(Z) \mid A = a\}$ where $\mathbb{P}_n^*$ is a series estimator. Their estimator uses *cross-fitting*, whereby, for a given fold k, the nuisance functions are estimated on all folds but k, and $\mathbb{P}_n^*$ is computed using observations from k. This construction bypasses the need to impose Donsker conditions on the nuisance functions. We note that, relative to the results in Kennedy et al. [2017] and Colangelo and Lee [2020], those in Semenova and Chernozhukov [2017] appear to require that the product of root-mean-square-errors for estimating $\mu(a, x)$ and $\pi(a \mid x)$ is of smaller order than $1/\sqrt{nk}$, where k is the dimension of the basis (Assumptions 3.5 and 4.9). This is more stringent of a requirement than (5.1).

The approach taken by Colangelo and Lee [2020] is different in that instead of regressing $\widehat{\varphi}(Z)$ onto A, the estimator is

$$\widehat{\theta}(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{K_{ht}(A_i)\{Y_i - \widehat{\mu}(t, X_i)\}}{\widehat{\pi}(t \mid X_i)} + \widehat{\mu}(t, X_i) \right] \tag{5.2}$$

They motivate their estimator as being based on an approximate first-order influence function, which can be calculated as the Gateaux derivative with respect to smooth deviations from the true data-generating distribution as these deviations approach a distribution with point-mass at $A = t$ (see their Section 4). This estimator still enjoys second order rates, but, it is not immediately clear how it adapts to different level of smoothness of $\theta(a)$. That is, their error rates may be of the form "oracle + second-order terms" only in certain smoothness regimes of $\theta(a)$. This is in contrast to estimators based on regressing $\widehat{\varphi}(Z)$ on A, which would behave like an oracle, and thus adapt to the smoothness of $\theta(a)$, as long as the second-order remainder terms are negligible. Their analysis focuses on low-smoothness regimes; viewed as a function of a, they assume that the joint density of the observations is three-times continuously differentiable. Notice that this implies that both the outcome regression $a \mapsto \mu(a, x)$ and the conditional density $a \mapsto \pi(a \mid x)$ are three-times continuously differentiable. In practice, however, it could be that $a \mapsto \mu(a, x)$ and thus the dose-response curve are smoother than $a \mapsto \pi(a \mid x)$. Our m^{th}-order estimator is an extension of (5.2) and appears to track the smoothness of the dose-response only in cases when this is no-greater than the smoothness of $a \mapsto \pi(a \mid x)$,

which appears to be consistent with the results in Colangelo and Lee [2020].

5.1.4 Our contribution

Our contribution is mainly three-fold. We study three approaches to dose-response estimation: one based on estimators relying on approximate first-order influence functions and one based on higher-order corrections. For the first approach, we consider two estimation strategies. The first one is based on empirical loss minimization, which we view as a "global" method since it naturally estimates the curve on its entire support. Our approach specializes the results of Foster and Syrgkanis [2019] on empirical loss minimization with estimated outcomes to estimate dose-response functions under the square-loss. Importantly, we show that the resulting estimator is doubly-robust and give an explicit characterization of the remainder term. This, in turn, implies faster rates of convergence than those directly obtainable from the results in Foster and Syrgkanis [2019] whenever the treatment and the outcome models are estimated at different rates. The second one extends the DR-learner estimator of the conditional average treatment effect (proposed in Kennedy [2020]) to the continuous treatment effect setting. We view this as a "local" method since it estimates the dose-response at a specific point.

Next, we show how convergence rates can be substantially improved using kernel-smoothed, approximate higher order influence functions [Robins et al., 2008, 2009a, 2017a]. To the best of our knowledge, our higher order estimator is the first use of higher order influence functions to estimate a dose-response curve. Further, we are not aware of other estimators of the dose-response curve that exhibit convergence rates as fast as that of our higher order estimator, under similar assumptions on the data generating process.

Finally, extending the work of Bonvini et al. [2022a] on sensitivity analysis in marginal structural models, we describe a simple, yet flexible framework to gauge the impact of potential unmeasured confounders on the dose-response estimates. We analyze the performance of DR-Learner-based estimators of the bounds on the dose-response function derived under the sensitivity model.

5.2 Doubly-robust estimators

5.2.1 General doubly-robust estimation procedure

Here, we expand on the list of estimators enjoying second-order, doubly robust errors. We will show that extensions of the general procedure proposed in Foster and Syrgkanis [2019] and the DR-learner approach proposed by Kennedy [2020] in the context of conditional effects defined by binary treatments also yield estimators enjoying second-order and doubly-robust remainder terms. The work by Foster and Syrgkanis [2019] is rather general and already yields estimators that have second-order remainder terms, but their rates are in terms of $\|\widehat{\eta} - \eta\|_{\mathcal{F}}$ where $\|f\|_{\mathcal{F}}$ is a norm for the function spaces where all nuisance functions η live in. We apply their results to the dose-response settings and show that it is possible to obtain estimators that are also doubly-robust. Establishing the double-robustness property, i.e. that the second order remainder term is a *product of errors*, is particularly important when the estimators of

the nuisance functions converge at different rates, since the product of the errors would be of smaller order than the sum of the squared errors.

Let Z_1^n, Z_2^n and Z_3^n denote three independent samples. We will work with estimates of the pseudo-outcome $\varphi(Z_j)$ of the form

$$\widehat{\varphi}(Z_j) = \widehat{w}(A_j, X_j)\{Y_j - \widehat{\mu}(A_j, X_j)\} + \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}(A_j, X_i)$$

where $\widehat{\mu}(a, x)$ and $\widehat{w}(a, x)$ are estimated using observations in Z_1^n, the observations $(X_i)_{i=1}^n$ belong to Z_2^n and Z_j belongs to Z_3^n. An alternative approach, taken in Semenova and Chernozhukov [2017], is to consider only two samples, say Z_1^n and Z_2^n, and compute

$$\widehat{\varphi}(Z_j) = \widehat{w}(A_j, X_j)\{Y_j - \widehat{\mu}(A_j, X_j)\} + \frac{1}{n}\sum_{i\neq j}^{n} \widehat{\mu}(A_j, X_i)$$

for Z_j and $(X_i)_{i=1}^n$ in the same sample Z_2^n. We proceed by considering three separate samples to simplify the analysis of all our estimators, as we have $\widehat{\varphi}(Z_k) \perp\!\!\!\perp \widehat{\varphi}(Z_l) \mid (Z_1^n, Z_2^n)$ for $k \neq l$. The roles of Z_1^n, Z_2^n and Z_3^n can be swapped, which results in three estimators of $\theta(t)$. One can then take their average as the final estimator. From a sample of iid observations, it is possible to obtain separate independent samples simply by randomly split the data into sub-samples. To keep the notation as light as possible, we analyze the theoretical properties of the estimators based a single split into three subsamples. However, we expect the same arguments to hold when multiple splits are performed.

Our estimation procedure is summarized in the following algorithm. In the following two sections, we give error bounds for a procedure that generalizes Algorithm 5.1. In particular, the bounds apply to the problem of estimating some $\theta(u) \equiv \mathbb{E}\{f(Z) \mid U = u\}$, where U is some observed subset of Z and $f(Z)$ is not directly observable. The estimator is $\widehat{\theta}(u) = \widehat{\mathbb{E}}_n\{\widehat{f}(Z) \mid U = u\}$, where $\widehat{\mathbb{E}}_n(\cdot \mid U = u)$ is either an empirical risk minimizer or a linear smoother and it is computed from a sample independent of that used to construct $\widehat{f}(\cdot)$. One can see that Algorithm 5.1 fits exactly this framework where $f(Z) = \varphi(Z)$. The additional sample split considered in Algorithm 5.1 is not needed to derive the next two propositions but it is useful to derive the result in Lemma 12. Finally, both bounds on the risk will involve a particular bias term $\widehat{r}(u)$ that would need to be analyzed on a case-by-case basis

$$\widehat{r}(u) = \int \widehat{f}(z) d\mathbb{P}(z \mid U = u) - \theta_0(u).$$

To estimate a dose-response curve, we have $f(Z) = \varphi(Z)$ and we propose using $\widehat{\varphi}(Z)$ as an estimator of $\varphi(Z)$, as detailed in Algorithm 5.1. Lemma 12 below shows that $\widehat{r}(u)$ is second-order and doubly-robust.

Let Z_1^n, Z_2^n and Z_3^n denote three independent samples of n iid observations of $Z = (Y, A, X)$.

1. Nuisance training
   - Using only observations in Z_1^n, estimate $\mu(A, X)$ with $\widehat{\mu}(A, X)$ and $w(A, X)$ with $\widehat{w}(A, X)$;
   - Using only observations in Z_2^n, estimate $m(a) = \int \mu(a, x)p(x)dx$ with $\widehat{m}(a) = n^{-1}\sum_{i=1}^n \widehat{\mu}(a, X_i)$.
2. Pseudo-outcome construction: using observations in Z_3^n, construct the pseudo-outcome

$$\widehat{\varphi}(Z) = \widehat{w}(A, X)\{Y - \widehat{\mu}(A, X)\} + \widehat{m}(A)$$

3. Second stage regression, either of the following:
   (a) Empirical-risk-minimization: Define $\widehat{\theta}$ to be the empirical risk minimizer

$$\widehat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i\in Z_3^n} \{\widehat{\varphi}(Z_i) - \theta(A_i)\}^2$$

   where Θ is some function class.
   (b) DR-Learner: Define

$$\widehat{\theta}(t) = \frac{1}{n}\sum_{i\in Z_3^n} W_i(t; A^n)\widehat{\varphi}(Z_i)$$

   where $W_i(t; A^n)$ are weights depending on t and $A^n = (A_1, \ldots, A_n) \subset Z_3^n$.
4. (Optional) Cross-fitting: swap the role of Z_1^n, Z_2^n and Z_3^n and repeat steps 1 and 2. Use the average of the three estimators as an estimate of θ.

Figure 5.1: Algorithm to compute general doubly-robust estimators of the dose-response function.

5.2.2 Upper bound on the risk of the ERM-based estimator

We start by considering estimating $\theta(t)$ via empirical loss minimization as in Algorithm 5.1 (a). We view this as a "global" method, as we estimate the function on its entire support as opposed to local methods, such as the DR-Learner discussed next, whereby the dose-response is estimated at a specific point. The error bound we describe in this section will be on the L_2 loss and will be a specialization of the results described in Foster and Syrgkanis [2019] and Wainwright [2019]. Foster and Syrgkanis [2019] provides a general framework for doing empirical risk minimization in the presence of nuisance components that need to be estimated. Here, we take their approach and find that the oracle rate is achievable if $\mathbb{E}\|\widehat{r}\|^2$ is simply of smaller order. In particular, from Lemma 27, if the orthogonal signal $\varphi(Z)$ is used, $\widehat{r}$ consists of a product of errors, as opposed to simply being of second order, and thus the bound on the MSE of our procedure improves upon the bound from Foster and Syrgkanis [2019].

The next proposition provides a bound on the error incurred by an estimator that uses an estimated outcome $\widehat{f}(Z)$ in place of the true (unobservable) outcome $f(Z)$, when doing empirical risk minimization with the square loss to estimate a regression function $\mathbb{E}\{f(Z) \mid U = u\}$.

Proposition 7. Consider two independent samples, $D^n = (Z_{01}, \ldots, Z_{0n})$ and $Z^n = (Z_1, \ldots, Z_n)$, consisting of n iid copies of some generic observation Z distributed according to $\mathbb{P}$. Let U denote a generic variable such that $U \subset Z$. Let $\theta_0(u) \equiv \mathbb{E}\{f(Z) \mid U = u\}$ and suppose $\widehat{f}(\cdot)$ is constructed using only observations in D^n. Consider the estimator

$$\widehat{\theta} \equiv \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{\widehat{f}(Z_i) - \theta(U_i)\}^2.$$

Let $\theta^* = \operatorname{argmin}_{\theta \in \Theta} \|\theta - \theta_0\|$ and $\Theta^* = \{\theta - \theta^* : \theta \in \Theta\}$. Define the local Rademacher complexity:

$$\mathcal{R}_n(\Theta^*, \delta) = \mathbb{E}\left\{ \sup_{g \in \Theta^* : \|g\| \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(U_i) \right| \right\}$$

where $\epsilon_1, \ldots, \epsilon_n$ are iid Rademacher random variables, independent of the sample. Suppose Θ^* is star-shaped and $S \equiv \sup_{z \in \mathcal{Z}} |\widehat{f}(z)| \vee \sup_{\theta \in \Theta} \|\theta\|_\infty$ is finite. Let δ_n be any solution to $\mathcal{R}_n(\Theta^*, \delta) \leq \delta^2$ that satisfies

$$\delta_n^2 \gtrsim \frac{\log\log(n)}{n} \vee \frac{1}{2n}.$$

Then,

$$\mathbb{E}(\|\widehat{\theta} - \theta_0\|^2) \lesssim \|\theta^* - \theta_0\|^2 + \delta_n^2 + \mathbb{E}(\|\widehat{r}\|^2)$$

where $\|f\|^2 = \int f^2(z) d\mathbb{P}(z)$.

The error bound from Proposition 7 takes the form of an oracle rate plus a term involving $\widehat{r}$, which is controlled by Lemma 12 when $f(Z) = \varphi(Z)$ and $\widehat{f}(Z) = \widehat{\varphi}(Z)$.

The assumptions underlying Theorem 7 are rather mild. Appendix D in Foster and Syrgkanis [2019] and Chapters 13 and 14 in Wainwright [2019] describe common classes of functions for which the theorem applies, e.g., linear functions with constraints on the coefficients, functions satisfying Sobolev-type constraints or Reproducing Kernel Hilbert spaces. In order to apply Proposition 7, the class of functions considered has to be star-shaped. A class is star-shaped around the origin if, for any $g \in \mathcal{G}$ and $\alpha \in [0, 1]$, it is the case that $\alpha g \in \mathcal{G}$. Importantly, a convex set is star-shaped. If the star-shaped condition is not met, the statement of the theorem would hold for δ_n defined in terms of the star-hull of the function class. The boundedness assumption on $\widehat{\varphi}(Z)$ and Θ is used in various places in the proof, including in ensuring that the square-loss is globally Lipschitz; we expect this assumption to hold when the observations are bounded. Finally, the inequality involving δ_n should often be satisfied. For instance, $\delta_n^2 \geq 1/(2n)$ as long as Θ^* contains the constant function $\theta(u) = 1$.[1]

Example 2 (Orthogonal series, Examples 13.14 and 13.15 in Wainwright [2019]). Suppose $\theta(u)$ is α-times differentiable with $\theta^{(\alpha)}(u)$ satisfying $\int \{\theta^{(\alpha)}(u)\}^2 d\mathbb{P}(u) \leq B$ for some constant B. Let $\{p_j\}_{j=1}^{\infty}$ be an orthonormal basis of $L_2(\mathbb{P})$, such as the sine / cosine basis (see Belloni et al. [2015] for a discussion on different basis choices). Consider estimating θ_0 via ERM over the function class

$$\Theta(k, b) = \left\{ \theta_c(\cdot) : \sum_{j=1}^{k} p_j(\cdot) c_j, \ \sum_{j=1}^{k} c_j^2 \leq 1, \text{ and } |\theta_c(\cdot)| \ \leq b \right\}$$

Writing $\theta_0(u) = \sum_{j=1}^{\infty} p_j(u) c_{0j}$, we have

$$\theta^*(u) = \sum_{j=1}^{k} p_j(u) c_{0j} \quad \text{and} \quad \|\theta^* - \theta_0\|^2 = \sum_{j=k+1}^{\infty} c_{0j}^2.$$

It can be shown that $\|\theta^* - \theta_0\|^2 \leq k^{-2\alpha}$. Furthermore, the function class $\Theta^*(k) = \{\theta - \theta^*, \theta \in \Theta(k)\} = \Theta(k, 2b)$ is convex and thus star-shaped and can be shown to satisfy $\delta_n^2 \lesssim k/n$. Thus, Proposition 7 provides an upper bound of the mean-square error of the order

$$\mathbb{E}(\|\widehat{\theta} - \theta\|^2) \lesssim k^{-2\alpha} + \frac{k}{n} + \mathbb{E}(\|\widehat{r}\|^2)$$

[1] To see this, suppose that, for the sake of contradiction, $\delta_n < 1/\sqrt{2n}$. To start, because Θ^* is star-shaped, we have $g(U) = \delta_n \in \Theta^*$ because $\theta(U) = 1 \in \Theta^*$ and $\delta_n \in [0, 1]$. Then, $\|g\| = \delta_n$ so that

$$\mathcal{R}_n(\Theta^*, \delta_n) \geq \delta_n \mathbb{E}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \right| \right) \geq \frac{\delta_n}{\sqrt{2n}} > \delta_n^2$$

where the second inequality is an application of the Khintchine inequality. This is a contradiction because δ_n satisfies $\mathcal{R}_n(\Theta^*, \delta_n) \leq \delta_n^2$.

If k is chosen optimally, i.e. $k \sim n^{1/(2\alpha+1)}$, Proposition 7 shows that the oracle rate is attained as long as $\mathbb{E}(\|\widehat{r}\|^2)$ is of order $O(n^{-2\alpha/(2\alpha+1)})$.

5.2.3 Upper bound on the risk of the linear smoothing-based estimator

In this section, we consider a DR-Learner-style estimator (cf. Van der Laan [2006] and Kennedy [2020] for heterogeneous effects of binary treatments). The DR-Learning framework proposed and analyzed in Kennedy [2020] covers a broad class of second-stage estimators satisfying a stability condition, linear smoothers being one example. Here, for simplicity, we consider the case where the second-stage estimator in $\widehat{\theta}(t)$ is based on localized linear smoothing. As discussed in Example 3, regressing $\widehat{\varphi}(Z)$ on A via local polynomial regression represents an archetype of a localized DR-Learner. Kennedy et al. [2017] propose using generic learners to regress the estimated pseudo-outcome $\widehat{\varphi}(Z)$ on A but only analyze local linear estimators. Thus, our next proposition is an extension to their work, in the spirit of analyzing more general linear smoothers. Theorem 1 and Proposition 1 in Kennedy [2020] yield the following proposition.

Proposition 8. Consider two independent samples, $D^n = (Z_{01}, \ldots, Z_{0n})$ and $Z^n = (Z_1, \ldots, Z_n)$, consisting of n iid copies of some generic observation Z distributed according to $\mathbb{P}$. Let U denote a generic variable such that $U \subset Z$. Let $\theta_0(u) \equiv \mathbb{E}\{f(Z) \mid U = u\}$ and suppose $\widehat{f}(\cdot)$ is constructed using only observations in D^n. Consider the following estimator:

$$\widehat{\theta}(t) = n^{-1} \sum_{i=1}^{n} W_i(t; A^n) \widehat{f}(Z)$$

Further suppose that the following regularity conditions hold:

- Minimum variance: $\text{var}\{f(Z) \mid U = u\} \geq c > 0$ for all $u \in \mathcal{U}$ and some constant c;
- Consistency of nuisance estimators: $\sup_z |\widehat{f}(z) - f(z)| = o_{\mathbb{P}}(1)$;
- Localized weights: $n^{-1} \sum_{i=1}^{n} |W_i| \leq C$, for some constant C, and there exists a neighborhood N_t around $U = t$ such that $W_i(t; U^n) = 0$ if $U_i \notin N_t$.

Then, letting $\widetilde{\theta}(t) = n^{-1} \sum_{i=1}^{n} W_i(t; U^n) f(Z_i)$ denote the oracle estimator:

$$|\widehat{\theta}(t) - \theta_0(t)| \leq \left|\widetilde{\theta}(t) - \theta_0(t)\right| + \sup_{u \in N_t} |\widehat{r}(u)| + o_{\mathbb{P}}\left(\mathbb{E}\left[\left\{\widetilde{\theta}(t) - \theta_0(t)\right\}^2\right]\right).$$

As discussed in Kennedy [2020], the assumptions underlying Proposition 8 are easily satisfied for linear smoothers of the local polynomial regression variety. In particular, the weights of the local polynomial regression satisfies the assumptions (Tsybakov [2009], Lemma 1.3). This proposition follows from the results contained in Kennedy [2020] that apply to general linear smoothers, e.g. it does not require the weights to be localized. We work with localized weights to simplify the analysis of the point-wise risk.

Example 3. Suppose $\theta_0(t) \equiv \mathbb{E}\{f(Z) \mid A = t\}$ belongs to a Hölder class of order α and let $p = \lfloor \alpha \rfloor$. A DR-Learner can be based upon local polynomial regression of order p. The weights are

$$W_i(t; A^n) = s(t)^T \widehat{Q}^{-1} K_{ht}(A_i) s(A_i)^T,$$

where $K(\cdot)$ is a kernel function, $\widehat{Q} = \mathbb{P}_n\{s(A)s(A)^T\}$ and $s(a) = \left[1 \quad \frac{a-t}{h} \quad \dots \quad \left(\frac{a-t}{h}\right)^p\right]^T$. A standard calculation (see, for example, Tsybakov [2009]), yields that

$$\mathbb{E}\left[\left\{\widetilde{\theta}(t) - \theta_0(t)\right\}^2\right] = O(n^{-2\alpha/(2\alpha+1)})$$

This means that the oracle rate is attainable if $\sup_{u \in N_t} \widehat{r}^2(u) = O_{\mathbb{P}}(n^{-2\alpha/(2\alpha+1)})$, which is essentially the same requirement as for the estimator based on empirical-risk-minimization, see Example 2.

Remark 7. From the bound in Proposition 8, inference can be carried out in the oracle regime, i.e., under the assumption that $\sup_{u \in N_t} |\widehat{r}(u)|$ is of smaller order that $|\widetilde{\theta}(t) - \theta(t)|$. In particular, if this holds, all inference tools for standard local nonparametric regression can be used. For example, let the setup be as in Example 2. Let $\sigma^2(t)$ be asymptotic variance of $\widetilde{\theta}(t)$, $\widehat{\sigma}^2(t)$ its consistent estimator and $b(t)$ the asymptotic bias. Then, if $\sup_{u \in N_t} \widehat{r}^2(u) = o_{\mathbb{P}}((nh)^{-1/2})$, we have

$$\frac{\sqrt{nh}[\widehat{\theta}(t) - \theta(t) - b(t)]}{\widehat{\sigma}(t)} \rightsquigarrow N(0,1)$$

as shown, for instance, in Section 4 of Fan and Gijbels [2018]. Notice that, without undersmoothing or bias-correction, a Wald-type confidence interval based on the asymptotic statement above will cover the smoothed dose-response curve $\mathbb{E}\{\widetilde{\theta}(t)\}$, rather than $\theta(t)$ itself (see Section 5.7 in Wasserman [2006] for more discussion).

5.2.4 Bounding the conditional bias of $\widehat{\varphi}(Z)$

As outlined in Propositions 7 and 8, the analysis of the estimator based on empirical risk minimization and that of the one based on linear smoothing yield a bound on the MSE that is the oracle rate plus a term of the order of $\widehat{r}^2(t)$. We show that $\widehat{r}(t)$ for $\widehat{f}(Z) = \widehat{\varphi}(Z)$ is second-order, as outlined in the following lemma.

Lemma 12. *Let $\widehat{r}(t) = \mathbb{E}\{\widehat{\varphi}(Z) \mid A = t, D^n\} - \theta_0(t)$. It holds that*

$$|\widehat{r}(t)| \lesssim \|w - \widehat{w}\|_t \|\mu - \widehat{\mu}\|_t + |(\mathbb{P}_n - \mathbb{P})\{\widehat{\mu}(t, X)\}|$$

where $\|f\|_t^2 = \int f^2(z) d\mathbb{P}(z \mid A = t)$ and $\mathbb{P}_n$ denotes an average over observations in sample Z_2^n.

Proof. Recall that $\theta_0(t) = \mathbb{E}\{\varphi(Z) \mid A = t\}$. By Bayes' rule, we have

$$\mathbb{P}\{\widehat{\mu}(t, X)\} - \theta(t) = \int w(t, x)\{\widehat{\mu}(t, x) - \mu(t, x)\} d\mathbb{P}(x \mid A = t)$$

and

$$\mathbb{E}\{\widehat{\varphi}(Z) \mid A = t, D^n\} = \int \widehat{w}(t, x)\{\mu(t, x) - \widehat{\mu}(t, x)\} d\mathbb{P}(x \mid A = t) + \mathbb{P}_n\{\widehat{\mu}(t, X)\}$$

Adding and subtracting $\mathbb{P}\{\widehat{\mu}(t, X)\}$ and applying Cauchy-Schwarz yield the result. □

The result from Lemma 12 shows that $|\widehat{r}(t)|$ can be bounded by the product of the L_2 errors in estimating $w(a, x)$ and $\mu(a, x)$ plus a centered sample average, which would generally be of the smaller order $O_{\mathbb{P}}(n^{-1/2})$ if, for instance, the second moment of $\widehat{\mu}(t, X)$ (conditional on Z_1^n) is bounded. In this respect, this term is effectively asymptotically negligible in nonparametric models where the rate of convergence is of slower order than $n^{-1/2}$. Thus, the conditional bias of $\widehat{\varphi}(Z)$ is driven by the product of the errors incurred in estimating the nuisance functions; this product structure of the bias is important when the nuisance functions are estimated at different rates.

Standard results are generally calculated for $L_2(d\mathbb{P}(a, x))$ errors defined by the joint distribution of (A, X), for example

$$\|w - \widehat{w}\|^2 \equiv \int \{w(a, x) - \widehat{w}(a, x)\}^2 d\mathbb{P}(a, x).$$

In this case, optimal convergence rates for estimating $\mu(a, x)$ are well-understood for many classes. For instance, if $\mu(a, x)$ belongs to a Hölder class of order γ, then minimax-optimal convergence rates in $L_2(d\mathbb{P}(a, x))$ are of order $n^{-2\gamma/(2\gamma+d+1)}$. The bound from Lemma 12 is actually on an L_2 error with weight given by the conditional density of X given $A = t$. In most settings, we expect the more conventional rate based on $L_2(d\mathbb{P}(a, x))$ to match that based on $L_2(d\mathbb{P}(x \mid A = t))$. For example, Result 1 from Colangelo and Lee [2020] shows that the rate in $L_2(d\mathbb{P}(x \mid A = t)p(t))$ matches that for the point-wise risk (in (A, X)) under a mild boundeness assumption. Alternatively, we note that one can always upper bound (up to constants) $\|f\|_t$ by the supremum norm $\|f\|_\infty$ and the rate for estimating a regression function in L_∞ generally matches that for estimating the function in $L_2(d\mathbb{P}(a, x))$ up to log factors.

There are fewer results available for conditional density estimation compared to regression estimation. Recently Ai et al. [2018] have proposed a method to estimate $w(a, x)$ directly that, under certain conditions, exhibits a convergence rate in L_2 of order $n^{-2\gamma/(2\gamma+d+1)}$ if $w(a, x)$ is γ-smooth (see their Theorem 3). Alternatively, one can estimate $p(a)$ and $\pi(a \mid x)$ and compute their ratio to estimate $w(a, x)$. We refer to Colangelo and Lee [2020] for a discussion on ways to estimate $\pi(t \mid x)$. In particular, one approach is to estimate $\mathbb{E}\{G_{h_1 t}(A) \mid X = x\}$, where $G(u)$ is some kernel and h_1 some bandwidth of choice. As a third approach, because $w(a, x) = p(a)p(x)/p(a, x)$, one can estimate the marginals $p(x)$ and $p(a)$ and the joint density

$p(a, x)$ and take the ratio as an estimate of $w(a, x)$. Estimating a joint density of d variables that belongs to a Hölder-class of order γ can be done with error scaling as $n^{-2\gamma/(2\gamma+d)}$. Thus, the MSE of this ratio would trivially be upper bounded by the MSEs for estimating $p(a, x)$, $p(x)$ and $p(x)$, which would depend on their respective smoothness levels.

As investigated in more detail in the next section, an interesting setting is where $a \mapsto \mu(a, x)$ has a different smoothness level than $x \mapsto \mu(a, x)$, where we expect the former to match the smoothness of the dose-response $\theta(a)$ in many applications. This is an example of anisotropic regression. The optimal rate for estimating a d-dimensional regression in a Hölder class where each coordinate has its own level of smoothness γ_j is of order $n^{-2\gamma/(1+2\gamma)}$, where γ satisfies $\gamma^{-1} = \sum_{j=1}^{d} \gamma_j^{-1}$ [Bertin, 2004, Hoffman and Lepski, 2002]. If $\mu(a, x)$ is in an anisotropic Hölder class of order $(\alpha, b, \ldots, b)$, the rate simplifies to $n^{-2b/(2b+b/\alpha+d)}$, where $d = \dim(X)$. If the treatment A is categorical or α is much larger than b, the rate is essentially $n^{-2b/(2b+d)}$, i.e. the optimal rate for estimating a d-dimensional regression function that is b-smooth. In a similar fashion, we may think of $a \mapsto \pi(a \mid x)$ and $x \mapsto \pi(a \mid x)$ as having different smoothness levels; optimal convergence rates in this context typically depend too on the harmonic means of the smoothness levels of each coordinate [Efromovich, 2007].

Remark 8. Suppose the dose-response $\theta(a)$ belongs to a Hölder class of order α and that w and μ are s-smooth so that they can be estimated in L_2 at the rate $n^{-2s/(2s+d+1)}$. Estimators whose risk is of the form "oracle rate + a term of the same order as $\widehat{r}$" would behave like an oracle estimator that has access to the true nuisance functions as soon as $s \geq (d+1)/\{2(1+1/\alpha)\}$. We will show that this oracle efficiency bar can be lowered, under certain conditions, by a higher order estimator. See Remark 13.

Remark 9. We note that our discussion on the rates attained by the doubly-robust estimators discussed in Section 5.2 is driven by the bound computed in Lemma 12. If $\widehat{\mu}$ and $\widehat{w}$ are designed to optimally estimate μ and w, e.g. by selecting tuning parameters to minimize estimates of their MSEs, then generally the bound based on Cauchy-Schwarz is the best available. However, there are other techniques, such as particular forms of sample splitting coupled with undersmoothing, whereby the nuisance functions are estimated optimally with respect to the target of inference, and so the selected tuning parameters for the nuisance estimators may not minimize the MSEs with respect to the nuisance functions. This approach has favorable theoretical properties, see e.g. Kennedy [2020], although it can be challenging to implement in practice. We leave studying undersmoothing in the context of continuous treatments for future work.

Remark 10. Compared to Theorem 2 in Kennedy et al. [2017], Theorem 8 and Lemma 12 provide the same error bound, but under substantially weaker conditions. Sample-splitting circumvents the need to impose Donsker-type conditions on the nuisance functions' classes in the form of bounded uniform entropy integrals. Moreover, the use of local polynomial regression allows the estimator to track the smoothness of $\theta(a)$, thereby achieving the oracle rate in high smoothness regimes (provided that the remainder term is negligible).

5.3 Higher-order estimators

5.3.1 Preliminaries

Inspired by the seminal work of Robins et al. [2008, 2009a, 2017a], in this section we investigate the use of higher-order influence functions (HOIFs) to estimate continuous treatment effects. To the best of our knowledge, this is the first time HOIFs are used in this context. For an introduction to higher-order influence functions, we refer to the main papers [Robins et al., 2009a, 2017a] and give a brief overview here. Informally, an m^{th}-order estimator of a functional $\chi(p)$ (where p is the density of the observations) takes the form

$$\widehat{\chi}(p) = \chi(\widehat{p}) + \sum_{j=1}^{m} \mathbb{U}_n\{\widehat{\varphi}_j(Z_1, \ldots, Z_j)\} \tag{5.3}$$

where $\mathbb{U}_n$ is the U-statistic measure so that

$$\mathbb{U}_n\{\varphi_j(Z_1, \ldots, Z_j)\} = \frac{1}{n(n-1)\cdots(n-j+1)} \sum_{1 \leq i_1 \neq i_2 \ldots \neq i_j \leq n} \varphi_j(Z_{i1}, \ldots, Z_{ij}).$$

Letting $\mathbb{P}^j\{f(Z_1, \ldots, Z_j)\} = \int f(z_1, \ldots, z_j) d\mathbb{P}(z_1) \ldots d\mathbb{P}(z_j)$ denote the corresponding population measure, this implies an expansion:

$$\widehat{\chi}(p) - \chi(p) = \chi(\widehat{p}) - \chi(p) + \sum_{j=1}^{m} \mathbb{P}^j\{\widehat{\varphi}_j(Z_1, \ldots, Z_j)\} + \sum_{j=1}^{m} (\mathbb{U}_n - \mathbb{P}^j)\{\widehat{\varphi}_j(Z_1, \ldots, Z_j)\}$$

Following Robins et al. [2009a], van der Vaart [2014], Robins et al. [2017a], if φ_j is chosen such that $-\mathbb{P}^j\{\widehat{\varphi}_j(Z_1, \ldots, Z_j)\}$ acts as the j^{th}-order term in the functional Taylor expansion of $\chi(\widehat{p}) - \chi(p)$, then

$$\chi(\widehat{p}) - \chi(p) + \sum_{j=1}^{m} \mathbb{P}^j\{\widehat{\varphi}_j(Z_1, \ldots, Z_j)\} = O(d(p - \widehat{p})^{m+1})$$

for some distance $d(\cdot)$. The quantity φ_j is referred to as the j^{th}-order influence function of $\chi(p)$. Provided that

$$\text{var}\left[\sum_{j=1}^{m} (\mathbb{U}_n - \mathbb{P}^j)\{\widehat{\varphi}_j(Z_1, \ldots, Z_j)\}\right] = O(n^{-1}),$$

this calculation would suggest that $\widehat{\chi}(p)$ would always be root-n consistent if m is large enough. However, higher order influence functions do not exist for many functionals of interest, including the average treatment effect of a binary treatment. In our setting, the dose-response does not possess influence functions of *any order*, in nonparametric models. While this means that generally it is not possible to construct root-n consistent estimators,

we will show that estimators of the form (5.3) that employ approximate influence functions still enjoy favorable properties. The performance of the resulting estimators will be based on a careful bias-variance trade-off. We show that an m^{th}-order estimator of the dose-response can outperform the doubly-robust estimators from Section 5.2 under certain smoothness conditions. Our estimator is tailored to models where $a \mapsto \mu(a, x)$ and $a \mapsto \pi(a \mid x)$ are α-times and β-times continuously differentiable, respectively. However, our analysis suggests that this estimator can outperform the current state-of-the-art only when $\alpha \leq \beta$.

5.3.2 Notation

Before describing our m^{th}-order estimator of the dose-response, we need to introduce some notation. Let $K_{ht}(a)$ denote a kernel of order $l = \lfloor \alpha \wedge \beta \rfloor$ and $b(x)$ denote a vector of the first k terms of some orthonormal basis. Define

$$\Pi_{i,j} \equiv \Pi(x_i, x_j) = b(x_i)^T \Omega^{-1} b(x_j)$$

where, for $g(x) = \int K_{ht}(a) p(a, x) da$:

$$\Omega = \int b(x) b(x)^T g(x) dx$$

Thus, provided that $g(x)$ is positive and bounded away from zero and infinity, $\Pi_{i,j}$ is effectively the kernel of an orthogonal projection in $L_2(g)$ onto a k-dimensional subspace. That is, for some function $f(x)$, $\int \Pi(x_i, x) f(x) g(x) dx = b(x_i)^T \beta^*$, where β^* solves the minimization problem

$$\beta^* = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \int \left\{ f(x) - b(x)^T \beta \right\}^2 g(x) dx$$

The kernel $\Pi(x_i, x_j)$ has to be estimated in practice because $g(x)$ depends on the true density $p(a, x)$. When X is multivariate, the basis can be taken to be the tensor product basis. Following Robins et al. [2017a], by a slight abuse of notation, we will denote the projection operator associated with the kernel above using the same symbol Π. This way, we have $\Pi(f)(x_i) = \int \Pi(x_i, x) f(x) g(x) dx$.

Example 4. Suppose $X_i \in [a_i, b_i]$ for $i \in \{1, 2\}$, i.e., $X \in \mathcal{X} \subset \mathbb{R}^2$. Let $\widetilde{b}(u)$ be a k-dim vector of terms from an orthonormal basis in L_2 over the interval $[-1, 1]$. We may construct a generic element $b_u(x_1, x_2)$ of $b(x_1, x_2)$ as

$$b_u(x_1, x_2) = \frac{4}{(b_1 - a_1)(b_2 - a_2)\sqrt{g(x)}} \widetilde{b}_l \left(\frac{2x_1 - a_1 - b_1}{b_1 - a_1} \right) \widetilde{b}_m \left(\frac{2x_2 - a_2 - b_2}{b_2 - a_2} \right)$$

where l and m range over $\{1, \ldots, k\}$. By a change of variables, it can be seen that $b_u(a, x)$ is orthonormal in $L_2(g)$ so that the kernel $\Pi(x_{1i}, x_{2i}, x_{1j}, x_{2j})$ simplifies to

$$\Pi(x_{1i}, x_{2i}, x_{1j}, x_{2j}) = b(x_{1i}, x_{2i})^T b(x_{1j}, x_{2j}).$$

5.3.3 The estimator

In this section, we describe an estimator of $\theta(t) = \int \mu(t,x)p(x)dx$ based on approximate, m^{th}-order HOIFs. Define the first approximate influence function:

$$f_0(Z) = \frac{K_{ht}(A)\{Y - \mu(t,X)\}}{\pi(t \mid X)} + \mu(t,X)$$

and the functions

$$f_1(Z) = K_{ht}(A)\{Y - \mu(A,X)\}$$
$$f_2(Z) = \frac{K_{ht}(A)}{\pi(A \mid X)} - 1$$

The function $f_0(Z)$ is a sum of a residual term involving $Y - \mu(t,X)$ and the outcome model $\mu(t,X)$. If A was binary and $K_{ht}(A) = A$, $f_0(Z)$ would be exactly the influence function of $\int \mu(1,x)d\mathbb{P}(x)$, which equals $\mathbb{E}(Y^1)$ under standard causal assumptions. The terms $f_1(Z)$ and $f_2(Z)$ are kernel-weighted residuals; $f_2(Z)$ is a residual term in the sense that $\mathbb{E}\{K_{ht}(A)\pi(A \mid X) \mid X\} = 1$ whenever $\int K_{ht}(a)da = 1$.

The m^{th}-order estimator of $\theta(t)$ that we study is

$$\widehat{\theta}(t) = \mathbb{P}_n\{\widehat{f}_0(Z)\} + \sum_{j=2}^{m} \mathbb{U}_n\{\widehat{\varphi}_j(Z_1,\ldots,Z_j)\}$$

where

$$\varphi_j(Z_1,\ldots,Z_j) = (-1)^{j-1} \sum_{A \subset \{1,\ldots,j\}} (-1)^{j-|A|} \mathbb{E}\left\{\overline{\varphi}_j(Z_1,\ldots,Z_j) \mid Z_i, i \in A\right\}$$
$$\overline{\varphi}_j(Z_1,\ldots,Z_j) = f_1(Z_1)\Pi_{1,2}K_{ht}(A_2)\cdots\Pi_{j-2,j-1}K_{ht}(A_{j-1})\Pi_{j-1,j}f_2(Z_j)$$

are the m^{th}-order approximate influence functions. Notice that $\varphi_j(Z_1,\ldots,Z_j)$ is simply the degenerate version of $\overline{\varphi}_j(Z_1,\ldots,Z_j)$, which ensures that

$$\int \varphi_j(z_1,\ldots,z_j)d\mathbb{P}(z_i) = 0$$

for every i and $(z_l : l \neq i)$. In addition, it holds that

$$\begin{aligned}
&\int \Pi(x_{i-1},x_i)K_{ht}(a_i)\Pi(x_i,x_{i+1})d\mathbb{P}(z_i) \\
&= b(x_{i-1})^T\Omega^{-1}\int b(x_i)b(x_i)^T K_{ht}(a_i)d\mathbb{P}(z_i)\Omega^{-1}b(x_{i+1}) \\
&= \Pi(x_{i-1},x_{i+1})
\end{aligned}$$

and, by degeneracy of $f_1(z)$ and $f_2(z)$,

$$\int f_1(z_1)\Pi(x_1, x_2)d\mathbb{P}(z_1) = \int \Pi(x_{j-1}, x_j)f_2(z_j)d\mathbb{P}(z_j) = 0.$$

This means that the first few approximate HOIFs take a rather simple form:

$$\begin{aligned}
\varphi_2(Z_1, Z_2) &= -f_1(Z_1)\Pi_{1,2}f_2(Z_2) \\
\varphi_3(Z_1, Z_2, Z_3) &= f_1(Z_1)\Pi_{1,2}K_{ht}(A_2)\Pi_{2,3}f_2(Z_3) - f_1(Z_1)\Pi_{1,3}f_2(Z_3) \\
\varphi_4(Z_1, Z_2, Z_3, Z_4) &= -f_1(Z_1)\Pi_{1,2}K_{ht}(A_2)\Pi_{2,3}K_{ht}(A_3)\Pi_{3,4}f_2(Z_4) \\
&\quad + f_1(Z_1)\Pi_{1,2}K_{ht}(A_2)\Pi_{2,4}f_2(Z_4) + f_1(Z_1)\Pi_{1,3}K_{ht}(A_3)\Pi_{3,4}f_2(Z_4) \\
&\quad - f_1(Z_1)\Pi_{1,4}f_2(Z_4)
\end{aligned}$$

Remark 11. The estimator $\widehat{\theta}(t) = \mathbb{P}_n\{\widehat{f}_0(Z)\}$, corresponding to $m = 1$, is precisely the estimator studied in Colangelo and Lee [2020]. Thus, we may view the m^{th}-order estimator as a higher-order generalization of their approach.

Remark 12. The m^{th}-order estimator that we study has the same form as the m^{th}-order estimator of the functional $\psi = \int \mathbb{E}(Y \mid A = 1, X = x)p(x)dx$ studied in Robins et al. [2017a] (Section 8) except that terms of the form $Af(Z)$ for some function f of the observations are replaced by $K_{ht}(A)f(Z)$. In fact, the rate described in Theorem 5 is similar to that for ψ from Theorem 8.1 in Robins et al. [2017a] with n replaced by nh. Finally, Section 9 in Robins et al. [2017a] presents an estimator that is a modified version of that presented in Section 8.1 where certain terms in the influence functions are "cut out" to decrease the variances without increasing the bias. This results in a more complex estimator that exhibits a better, and in fact minimax optimal under certain conditions, bias-variance trade-off. We plan to apply this refinement to the dose-response settings in future work, with the idea of first calculating a candidate minimax lower bound.

We propose estimating all nuisance functions, namely $\pi(a \mid x)$, $\mu(a, x)$ and $g(x)$ using a separate independent sample D^n. Notice that $\Pi(x_i, x_j)$ can be estimated by $b(x_i)^T\widehat{\Omega}^{-1}b(x_j)$, where $\widehat{\Omega}$ is a suitable estimator of $\int b(x)b(x)^T g(x)dx$. The weight $g(x) = \int K_{ht}(a)p(a, x)da$ can be estimated as $\int K_{ht}(a)\widehat{p}(a, x)da$. However, for k sufficiently small, an attractive alternative is to use the empirical version of Ω, namely $\widehat{\Omega} = \mathbb{P}_n\{b(X)b(X)^T K_{ht}(A)\}$. See also Mukherjee et al. [2017] for an in-depth discussion of using the empirical counterpart of Ω for estimators based on higher-order influence functions.

5.3.4 Upper bound on the (conditional) risk

Here, we bound the risk of the estimator $\widehat{\theta}(t)$ conditional on the training sample D^n.

Theorem 5. *Suppose Assumptions 1-2 hold and the following assumptions also hold:*

1. *The functions $a \mapsto \mu(a, x)$ and $a \mapsto \pi(a \mid x)$ are α-times and β-times continuously differentiable with uniformly bounded derivatives, for any $x \in \mathcal{X}$;*

2. *The kernel K of order $l = \alpha \wedge \beta$ is uniformly bounded, supported in $[-1, 1]$ and satisfies $\int K(u)du = 1$ and $\int K_{ht}(a)p(a, x)da \in [\epsilon, M]$ for some $\epsilon > 0$, $M < \infty$ and all $x \in \mathcal{X}$.*

3. *The orthogonal projection kernel Π and its estimator $\widehat{\Pi}$ satisfy $\sup_x \Pi(x, x) \lesssim k$ and $\sup_x \widehat{\Pi}(x, x) \lesssim k$;*

4. *Boundedness: $\int K_{ht}(a)p(a, x)da / \int K_{ht}(a)\widehat{p}(a, x)da \in [\epsilon^{'}, M^{'}]$ for some $\epsilon^{'} > 0$, $M^{'} < \infty$ and all $x \in \mathcal{X}$; similarly the density of X is uniformly bounded.*

Then

$$\left|\mathbb{E}\{\widehat{\theta}(t) - \theta(t) \mid D^n\}\right| \lesssim \|(I - \Pi)(v)\|_g \|(I - \Pi)(q)\|_g + h^{\alpha\wedge\beta} + \|q\|_g \|v\|_g \|f\|_\infty^{m-1}$$

$$var\{\widehat{\theta}(t) \mid D^n\} \lesssim \sum_{j=1}^{m} \frac{k^{j-1}h^{-j}}{n(n-1)\cdots(n-j+1)}$$

where $v(x) = \mu(t, x) - \widehat{\mu}(t, x)$, $q(x) = 1/\widehat{\pi}(t \mid x) - 1/\pi(t \mid x)$ and $f(x) = \widehat{p}(t, x) - p(t, x)$, and $\|f\|_g^2 = \int f^2(x)g(x)dx$.

The assumptions underlying Theorem 5 are similar to those made in Propositions 7 and 8. The main difference is that the higher order estimator $\widehat{\theta}(t)$ is specifically designed for nonparametric models where $a \mapsto \mu(a, x)$ and $a \mapsto \pi(a \mid x)$ possess some smoothness, which we encode in condition 1. The second condition ensures that the kernel K accurately tracks the least smooth function between $a \mapsto \mu(a, x)$ and $a \mapsto \pi(a \mid x)$. A better estimator or a tighter bound would track just the smoothness of $\theta(a)$ or, at least, the smoothness of $a \mapsto \mu(a, x)$, as that should match the smoothness of $\theta(a)$ in most applications. We leave this for future work. In particular, we conjecture it might be possible to derive a tighter bound that would have, in place of the term $h^{\alpha\wedge\beta}$, terms of order $h^{\alpha\wedge(\beta+1)}$ plus terms of order $h^{\alpha\wedge\beta}(\|v\| + \|q\| + o(h))$. This refined bound would also not track the smoothness of the dose-response and thus we preferred the simpler and more interpretable bound in terms of $h^{\alpha\wedge\beta}$.

Because the higher order kernels can take negative values on sets of non-zero Lebesgue measure (see, e.g. Proposition 1.3 in Tsybakov [2009]), we require $g(x) = \int K_{ht}(a)p(a, x)da$ to be bounded away from zero since this is the weight used in the projection Π onto the finite space of dimension k. Condition 3 requires the kernels Π and $\widehat{\Pi}$ to be bounded on the diagonal. This would be satisfied, for instance, if the basis elements are bounded. Condition 4 is a mild regularity condition on the estimator $\widehat{p}(a, x)$.

We now discuss a few implications of Theorem 5, under the assumptions that 1) $\alpha \leq \beta$, i.e. $a \mapsto \pi(a \mid x)$ is smoother than $a \mapsto \mu(a, x)$, and 2) the dose-response is also α-smooth.

Remark 13. In order to understand the implications of Theorem 5, we consider the case where $x \mapsto \widehat{\mu}(t, x)$ and $x \mapsto \mu(t, x)$ are Hölder-γ_1 and $x \mapsto \widehat{\pi}(t \mid x)$ and $x \mapsto \pi(t \mid x)$ are Hölder-γ_2. Given an appropriate basis, suppose the approximation error satisfies

$$\|(I - \Pi)(v)\|_g \|(I - \Pi)(q)\|_g \lesssim k^{-(\gamma_1+\gamma_2)/d}.$$

Each term in the variance bound contributes a term of order $k^{j-1}/(nh)^j$. Therefore, if we choose $k \sim nh$ the variance is of order $(nh)^{-1}$. With this choice of k, the third term in the bias $\|q\|_g\|v\|_g\|f\|_\infty^{m-1}$ can be made arbitrarily small by choosing m large enough and thus it is negligible relative to the other terms. The bound on the MSE (conditional on D^n) of $\widehat{\theta}(t)$ is thus $O(k^{-2(\gamma_1+\gamma_2)/d} + h^{2\alpha} + (nh)^{-1})$.

This means that, if the average nuisance functions' smoothness satisfies $(\gamma_1+\gamma_2)/2 \geq d/4$ and $h \sim n^{-1/(2\alpha+1)}$, one obtains the rate $n^{-2\alpha/(2\alpha+1)}$. Thus, $\widehat{\theta}(t)$ behaves like the oracle estimator that uses the true nuisance regression functions if $\alpha \leq \beta$ and $(\gamma_1+\gamma_2)/2 \geq d/4$, provided that m is chosen large enough. If $s = \gamma_1 = \gamma_2$, this means that $\widehat{\theta}(t)$ is oracle efficient for $s \geq d/4$. To the best of our knowledge, no existing estimator of the dose-response is oracle-efficient in this regime.

In order to compare this result to that from Remark 8, consider the case where the error in estimating the nuisance functions is entirely driven by that in estimating $x \mapsto \mu(t,x)$ and $x \mapsto \pi(t \mid x)$. This would be the case, for example, if A is categorical. Then, for $s = \gamma_1 = \gamma_2$, $|\widehat{r}(a)| \lesssim n^{-2s/(2s+d)}$ and the estimators from Section 5.2 are oracle-efficient only in the regime $s \geq d/\{2(1+1/\alpha)\}$. Thus, higher-order corrections, at least in the case where $\alpha \leq \beta$, effectively lower the bar for oracle efficiency.

Remark 14. Suppose we use HOIFs of order $m = 2$, the nuisance functions' smoothness satisfies $(\gamma_1+\gamma_2)/2 < d/4$ and $w(a,x) = p(a)/\pi(a \mid x)$ and $1/\pi(a \mid x)$ are estimable at the same rate in L_2. In this regime, the estimators from Section 5.2 are not oracle efficient, so the rate is driven by $\widehat{r}$. Without further corrections, $\widehat{r}$ is bounded by the product of the MSEs for estimating w and μ, which is of bigger order than the term $\|v\|\|q\|\|f\|_\infty$, which is of the same order as $\|v\|_g\|q\|_g\|f\|_\infty$ because $\int K_{ht}(a)\pi(a \mid x)da$ is uniformly bounded. Suppose k and h are chosen optimally and so are of orders

$$k \sim (nh)^{2d/(d+2\gamma_1+2\gamma_2)} \quad \text{and} \quad h \sim n^{-2(\gamma_1+\gamma_2)/[\alpha\{2(\gamma_1+\gamma_2)+d\}+2(\gamma_1+\gamma_2)]}.$$

Then, the MSE of $\widehat{\theta}(t)$ is of order n^{-2r_2} for

$$r_2 = \left\{1 + \frac{d}{2(\gamma_1+\gamma_2)} + \frac{1}{\alpha}\right\}^{-1} \wedge \|v\|_g\|q\|_g\|f\|_\infty$$

Thus, if the first term in r_2 dominates the rate, then the rate obtained by the quadratic estimator $\widehat{\theta}(t)$ is a combination of the oracle rate $1/(2+1/\alpha)$ and the minimax rate for estimating the dose-response when A is categorical (i.e. some average treatment effect) in the non-root-n regime, namely n^{-2r_f}, for $r_f = [1 + d/\{2(\gamma_1+\gamma_2)\}]^{-1}$, which is recovered as $\alpha \to \infty$.

In Figure 5.2, we illustrate the rates obtained in this work as a function of $s = \gamma_1 = \gamma_2$. Here s refers to the smoothness of $x \mapsto \mu(a,x)$ and $x \mapsto \pi(a \mid x)$. For illustration, we set $\alpha = \beta = 2$ and $\dim(X) = 20$, where α is the smoothness of $a \mapsto \mu(a,x)$ and $a \mapsto \pi(a \mid x)$. In this setting, the optimal rate for estimating the anisotropic functions $\mu(a,x)$ and $\pi(a \mid x)$ is $n^{-2s/(2s+s/\alpha+d)}$. This is also the rate inherited by the plug-in estimator (black line) $\mathbb{P}_n\{\widehat{\mu}(a,X)\}$, without further corrections. The oracle rate is $n^{-2\alpha/(2\alpha+1)}$. The DR-Learner and the EMR-based estimator

(red line) achieve a rate of order $n^{-2\alpha/(2\alpha+1)} \vee n^{4s/(2s+s/a+d)}$. The blue line refers to the rate obtainable by the quadratic ($m = 2$) estimator under the assumption that the covariates density is estimated well enough so that the term $\|v\|_g\|q\|_g\|f\|_\infty$ is negligible, which is $n^{-2/\{1+d/(4s)+1/\alpha\}} \vee n^{-2\alpha/(2\alpha+1)}$; see Remark 14. Finally, as a reference value, we also plot the minimax lower bound for estimating the ATE, which is of order $n^{-2/\{1+d/(4s)\}} \vee n^{-1}$ [Robins et al., 2009b].

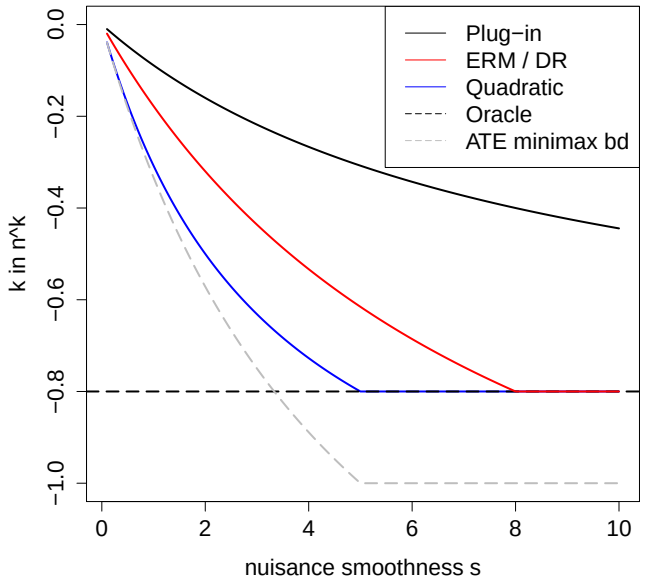


Figure 5.2: Illustration of the convergence rates in MSE for the estimator considered in this article, as a function of the smoothness $s = \gamma_1 = \gamma_2$. We take the smoothness of the dose-response to be $\alpha = 2$ and $\dim(X) = 20$.

For smooth functionals possessing a first-order influence function, efficient estimators based on the influence function are asymptotically equivalent. For instance, corrected plug-in estimators and TMLE may be different in finite samples but are asymptotically equivalent. In contrast, for functionals like the dose-response $\theta(t)$, which do not possess influence functions of any order, it is not clear whether estimators based on different approximations of the influence functions are equivalent asymptotically. This is true for higher order corrections as well, particularly for the choice of the projection kernel Π. For example, Π could be taken to represent a projection in $L_2(\int K_{ht}(a)p(a,x)da)$, as we have done in this work, or in $L_2(p(t,x))$. Using projections in $L_2(p(t,x))$ would avoid the assumption that $\int K_{ht}(a)p(a,x)da$ is positive and bounded away from zero, but at the expense of complicating the proof of the theorem, since the arguments made in the proof of Theorem 8.1 in Robins et al. [2017a] would need to adjusted to deal with issues such as $\int \Pi(x_{i-1},x_i)K_{ht}(a_i)\Pi(x_i,x_{i+1})d\mathbb{P}(z_i) \neq \Pi(x_{i-1},x_{i+1})$. Similarly, one may consider replacing $K_{ht}(a)$ with the weight function of a local polynomial regression. That

is, replacing $K_{ht}(a)$ with $s(t)^T Q^{-1} \widetilde{K}_{ht}(a) s(a)$, where $s(a) = \begin{bmatrix} 1 & (a-t) & \cdots & (a-t)^l \end{bmatrix}^T$, $Q = \int \widetilde{K}_{ht}(a) s(a) s(a)^T p(a) da$ and $\widetilde{K}(u)$ is a standard second-order kernel, such as the Epanechnikov. An approach conceptually similar to the DR-Learner may use projections in $L_2(p(x \mid t))$, although this may require a different analysis than what used to prove Theorem 5. Exploring the differences between these approaches is an important avenue for future work.

5.4 Sensitivity analysis to the no-unmeasured-confounding assumption

In this section, we briefly outline a simple pseudo-outcome regression method to carry out flexible, nonparametric sensitivity analysis to the no-unmeasured-confounding assumption, i.e., when $Y^a \not\perp\!\!\!\perp A \mid X$ so that $\int \mu(t,x) d\mathbb{P}(x)$ can no longer be interpreted as the dose-response curve. To the best of our knowledge, this is the first nonparametric sensitivity analysis method for continuous treatment effects. Bonvini et al. [2022a] propose an extension to Rosenbaum's sensitivity model for binary treatments as follows. Let U be such that $Y^a \perp\!\!\!\perp A \mid (X, U)$ and recall that $\mathbb{E}(Y^a) = \mathbb{E}\{Yp(a)/\pi(a \mid X, U) \mid A = a\}$. Let $\gamma \geq 1$ be a user-specified sensitivity parameter. Departures from the no-unmeasured-confounding assumption are parametrized by considering all densities of A given (X, U), $\pi(a \mid x, u)$, in the class

$$\Pi(\gamma) = \left\{ \pi(a \mid x, u) : \frac{1}{\gamma} \leq \frac{\pi(a \mid x, u)}{\pi(a \mid x)} \leq \gamma \right\}$$

When $\gamma = 1$, corresponding to the case when the measured covariates are sufficient to characterize the treatment selection process, one has the usual identification formula

$$\mathbb{E}(Y^a) = \mathbb{E}\{w(a, X) Y \mid A = a\} = \int \mu(a, x) d\mathbb{P}(x).$$

Lemma 2 in Bonvini et al. [2022a] shows that valid bounds on $\mathbb{E}(Y^a)$ under the sensitivity model $\Pi(\gamma)$ are

$$\theta_l(t; \gamma) = \int \mathbb{E}[Y \gamma^{\operatorname{sgn}\{q_l(t,x) - Y\}} \mid A = t, X = x] d\mathbb{P}(x)$$
$$\theta_u(t; \gamma) = \int \mathbb{E}[Y \gamma^{\operatorname{sgn}\{Y - q_u(t,x)\}} \mid A = t, X = x] d\mathbb{P}(x)$$

where $q_l(A, X)$ (resp. $q_u(A, X)$) is the $1/(1+\gamma)$ (resp. $\gamma/(1+\gamma)$)-quantile of Y given (A, X). In other words, for a given, user-specified γ, if $\pi(a \mid x, u) \in \Pi(\gamma)$, then $\theta_l(a; \gamma) \leq \mathbb{E}(Y^a) \leq \theta_u(a; \gamma)$.

A DR-Learner estimator of the bounds above can be computed by appropriately modifying the original pseudo-outcome $\varphi(Z) = w(A, X)\{Y - \mu(A, X)\} + \int \mu(A, x) d\mathbb{P}(x)$ and

regressing it onto A. For $j = \{l, u\}$, define

$$\varphi_j(Z;\gamma) \equiv \varphi_j(Z; w, \kappa_j, q_j, \gamma) = w(A,X)\{s_j(Z;q_j) - \kappa_j(A,X;q_j)\} + \int \kappa_j(A,x;q_j)d\mathbb{P}(x), \text{ for}$$

$$s_l(Z;q_l) = q_l(A,X) + \{Y - q_l(A,X)\}\gamma^{\text{sgn}\{q_l(A,X)-Y\}}$$

$$s_u(Z;q_u) = q_u(A,X) + \{Y - q_u(A,X)\}\gamma^{\text{sgn}\{Y-q_u(A,X)\}}$$

$$\kappa_j(A,X;q_j) = \mathbb{E}\{s_j(Z;q_j) \mid A, X\}$$

Following the sample splitting scheme whereby all nuisance functions are estimated on a separate, independent sample D^n, a DR-Learner estimator of $\theta_j(t;\gamma)$ regresses an estimate of $\varphi_j(Z;\gamma)$ onto A on the test set. For example, if the second stage regression is done via linear smoothing, then $\widehat{\theta}_j(t;\gamma) = n^{-1}\sum_{i=1}^n W_i(t;A_i)\widehat{\varphi}_j(Z;\gamma)$. It can be shown that $\varphi_j(Z)$ is just part of the influence function of $\int \theta_j(a;\gamma)d\mathbb{P}(a)$, which is a pathwise-differentiable parameter. Furthermore, $\varphi_l(Z;1) = \varphi_u(Z;1) = \varphi(Z)$.

The error analysis of the DR-Learners $\widehat{\theta}_l(t;\gamma)$ and $\widehat{\theta}_u(t;\gamma)$ follows from Propositions 7 and 8. In this light, it only remains to calculate $\mathbb{E}\{\widehat{\varphi}_j(Z;\gamma) - \varphi_j(Z;\gamma) \mid A = t, D^n\}$. We do so in the following lemma, proved in Appendix D.4.1, which plays the role of Lemma 12 in the no-unmeasured-confounding case.

Lemma 13. *Let* $\widehat{r}_j(t) = \mathbb{E}\{\widehat{\varphi}_j(Z;\gamma) - \varphi_j(Z;\gamma) \mid A = t, D^n\}$. *It holds that*

$$|\widehat{r}_j(t)| \lesssim \|w - \widehat{w}\|_t\|\kappa_j - \widehat{\kappa}_j\|_t + \|q_j - \widehat{q}_j\|_t^2 + |(\mathbb{P}_n - \mathbb{P})\widehat{\kappa}_j(t,X;\widehat{q}_j)|$$

The result of Lemma 13 is similar to that of Lemma 12, except that the upper bound on the conditional bias involve the additional term $\|q_j - \widehat{q}_j\|_t^2$. Thus, consistent estimation of the bounds relies on the consistency of the conditional quantiles estimators. The centered empirical average term is of order $O_{\mathbb{P}}(n^{-1/2})$, under mild boundedness conditions, and thus negligible in nonparametric models for which the convergence rate is slower than $n^{-1/2}$.

We conclude this section by establishing that $\varphi_j(Z;\gamma)$ satisfies the *doubly-valid* structure discovered by Dorn et al. [2021] in a similar sensitivity model for binary treatments. In particular, the bounds remain valid even if the conditional quantiles are not correctly specified. While Dorn et al. [2021] focused on binary treatments, their observation extends to the continuous treatment case as well, as summarized in the following proposition.

Proposition 9. Let $\overline{w}$, $\overline{\kappa}_l$, $\overline{\kappa}_u$, $\overline{q}_l$ and $\overline{q}_u$ be some fixed-functions such that all the expectations below are well defined. If either $\overline{\kappa}_j = \kappa_j(a,x;\overline{q}_j)$ or $\overline{w}(a,x) = w(a,x)$, but not necessarily both, then

$$\mathbb{E}\{\varphi_l(Z;\overline{w},\overline{\kappa}_l,\overline{q}_l,\gamma) \mid A = t\} \leq \theta_l(t;\gamma) \leq \theta_u(t;\gamma) \leq \mathbb{E}\{\varphi_u(Z;\overline{w},\overline{\kappa}_u,\overline{q}_u,\gamma) \mid A = t\}$$

Proof. If either $\overline{\kappa}_j = \kappa_j(a, x; \overline{q}_j)$ or $\overline{w}(a, x) = w(a, x)$, then

$$\begin{aligned}
&\mathbb{E}\{\varphi_l(Z; \overline{w}, \overline{\kappa}_l, \overline{q}_l, \gamma) \mid A = t\} \\
&= \int \left(\overline{q}_l(t, x) + \mathbb{E}[\{Y - \overline{q}_l(A, X)\}\gamma^{\text{sgn}\{\overline{q}_l(A,X)-Y\}} \mid A = t, X = x]\right) d\mathbb{P}(x) \\
&\mathbb{E}\{\varphi_u(Z; \overline{w}, \overline{\kappa}_u, \overline{q}_u, \gamma) \mid A = t\} \\
&= \int \left(\overline{q}_u(t, x) + \mathbb{E}[\{Y - \overline{q}_u(A, X)\}\gamma^{\text{sgn}\{Y-\overline{q}_u(A,X)\}} \mid A = t, X = x]\right) d\mathbb{P}(x)
\end{aligned}$$

The result follows because it holds that

$$\mathbb{E}\left[\gamma^{\text{sgn}\{q_l(A,X)-Y\}} \mid A, X\right] = \mathbb{E}\left[\gamma^{\text{sgn}\{Y-q_u(A,X)\}} \mid A, X\right] = 1$$

and, deterministically, that

$$\begin{aligned}
\{Y - \overline{q}_l(A, X)\}\gamma^{\text{sgn}\{\overline{q}_l(A,X)-Y\}} &\leq \{Y - \overline{q}_l(A, X)\}\gamma^{\text{sgn}\{q_l(A,X)-Y\}} \\
\{Y - \overline{q}_u(A, X)\}\gamma^{\text{sgn}\{Y-\overline{q}_u(A,X)\}} &\geq \{Y - \overline{q}_u(A, X)\}\gamma^{\text{sgn}\{Y-q_u(A,X)\}}
\end{aligned}$$

□

Proposition 3 establishes the doubly-valid structure of $\varphi_l(Z; \gamma)$ and $\varphi_u(Z; \gamma)$. Just like in the sensitivity model studied by Dorn et al. [2021] for binary treatments, the bounds on $\mathbb{E}(Y^a)$ remain valid even if the conditional quantiles are not correctly specified as long as either $w(a, x)$ or the second stage regression of $s_j(Z; \overline{q})$ onto (A, X) are.

In the next proposition, we provide the sample analog of Proposition 9 when the estimator of the bounds is a DR-Learner. Let $\overline{\kappa}_j(a, x) \equiv \kappa_j(a, x; \overline{q}_j) = \mathbb{E}\{s_j(Z; \overline{q}_j) \mid A = a, X = x\}$. Further, let $R_j^2(t)$ be the mean-square-error of an oracle estimator of $\theta_j(t; \gamma)$ regressing the pseudo-outcome $\overline{\varphi}_j(Z; w, \overline{\kappa}_j, \overline{q}_j, w, q_j)$ onto A, defined as

$$\begin{aligned}
\overline{\varphi}_u(Z; w, \overline{\kappa}_u, \overline{q}_u, w, q_u) = w(A, X)\{s_u(Z; \overline{q}_u) - \overline{\kappa}_u(A, X; \overline{q}_u)\} + \int \overline{\kappa}_u(A, x; \overline{q}_u) d\mathbb{P}(x) \\
- w(A, X)\{Y - \overline{q}_u(A, X)\}\left[\gamma^{\text{sgn}\{Y-\overline{q}_u(A,X)\}} - \gamma^{\text{sgn}\{Y-q_u(A,X)\}}\right] \\
\overline{\varphi}_l(Z; w, \overline{\kappa}_l, \overline{q}_l, w, q_l) = w(A, X)\{s_l(Z; \overline{q}_l) - \overline{\kappa}_l(A, X; \overline{q}_l)\} + \int \overline{\kappa}_l(A, x; \overline{q}_l) d\mathbb{P}(x) \\
- w(A, X)\{Y - \overline{q}_l(A, X)\}\left[\gamma^{\text{sgn}\{\overline{q}_l(A,X)-Y\}} - \gamma^{\text{sgn}\{q_l(A,X)-Y\}}\right]
\end{aligned}$$

It can be shown that $\mathbb{E}\{\overline{\varphi}_j(Z; w, \overline{\kappa}_j, \overline{q}_j, w, q_j) \mid A = t\} = \theta_j(t; \gamma)$ for $j = \{l, u\}$.

Proposition 10. Let $\widehat{\theta}_j(t; \gamma)$ be an DR-Learner estimator of $\theta_j(t; \gamma)$ based on linear smoothing (Sections 5.2 and 5.4). Further, let the sample splitting scheme be the same as in Figure 5.1 and assume that the following conditions hold:

1. If $T_i \leq V_i$ for all $i \in \{1, \ldots, n\}$, then the weights satisfy

$$\sum_{i=1}^{n} W_i(t; A^n) T_i \leq \sum_{i=1}^{n} W_i(t; A^n) V_i;$$

2. $\|\widehat{w} - w\|_\infty$, $\|\widehat{\kappa}_j - \overline{\kappa}_j\|_\infty$ and $\|\widehat{q}_j - \overline{q}_j\|_\infty$ are all $o_{\mathbb{P}}(1)$, where $\overline{q}_j(a, x)$ does not need to equal $q_j(a, x)$;

3. $\text{var}\{\overline{\varphi}_j(Z; w, \overline{\kappa}_j, \overline{q}_u, w, q_u) \mid A = a\} \geq c > 0$ for all $a \in \mathcal{A}$ and some constant c.

4. The outcome Y has a uniformly bounded conditional density given any values of (A, X);

5. The linear smoother weights $W_i(t; A^n)$ are localized as in Proposition 8 in a neighborhood N_t around $A = t$.

Then, the following inequalities hold

$$\widehat{\theta}_l(t; \gamma) \leq \theta_l(t; \gamma) + O_{\mathbb{P}}\left(R_l(t) + \sup_{a \in N_t} r_l(a)\right)$$
$$\widehat{\theta}_u(t; \gamma) \geq \theta_u(t; \gamma) + O_{\mathbb{P}}\left(R_u(t) + \sup_{a \in N_t} r_u(a)\right)$$

where, for $\|f\|_t^2 = \int f^2(z) d\mathbb{P}(z \mid A = t)$:

$$r_j(t) = \|\widehat{w} - w\|_t \|\widehat{\kappa}_j - \overline{\kappa}_j\|_t \; + \; \|\widehat{w} - w\|_t \|\widehat{q}_j - \overline{q}_j\|_t \; + \; |(\mathbb{P}_n - \mathbb{P})\widehat{\kappa}_j(t, X; \widehat{q}_j)|$$

Proposition 10 shows that, even if the conditional quantiles of Y given (A, X) are not well estimated, the estimators of the bounds can still converge to functions that contain the region $[\theta_l(t; \gamma), \theta_u(t; \gamma)]$ and, in this sense, are "valid bounds." The result holds under mild conditions. For instance, conditions 1 and 5 are a mild stability conditions on the second-stage linear smoother. Conditions 3 and 4 are mild regularity conditions on the data generating process and the nuisance functions' estimators. The speed at which $\widehat{\theta}_j(t; \gamma)$ converges to valid bounds depends on the structural properties of $\theta_j(t; \gamma)$, encoded in the oracle MSE $R_j^2(t)$, as well as the accuracy in estimating w and $\overline{\kappa}$. The proof of Proposition 10 extends the strategy of Dorn et al. [2021] to the case of non-root-n estimable parameters.

5.5 Small simulation experiment

We conduct a small simulation experiment to evaluate the performance of the first- and second-order estimators in finite samples. We generate data according to the following process

$$X \sim U(-1, 1), \quad A \sim \text{TruncNorm}(a_{\text{min}} = -1, a_{\text{max}} = 1, \text{ mean} = \kappa(x), \text{ sd} = 1),$$
$$\text{and} \quad Y \mid A, X \sim N(\xi(a, x), 0.25).$$

where $b(x) = \begin{bmatrix} b_1(x) & \dots & b_6(x) \end{bmatrix}^T$ are the first six, normalized Legendre polynomials and

$$\beta = \begin{bmatrix} 1, & 0.8, & 0.4, & 0.2, & 0.1, & 0.05 \end{bmatrix}^T$$
$$\kappa(x) = \frac{1}{3} b(x)^T \beta \quad \text{and} \quad \xi(a, x) = b(a)^T \beta + b(x)^T \beta$$

To estimate $\mu(a, x)$ and $\pi(a \mid x)$ while keeping tight control on the error incurred by the nuisance estimation step, we simulate estimators as

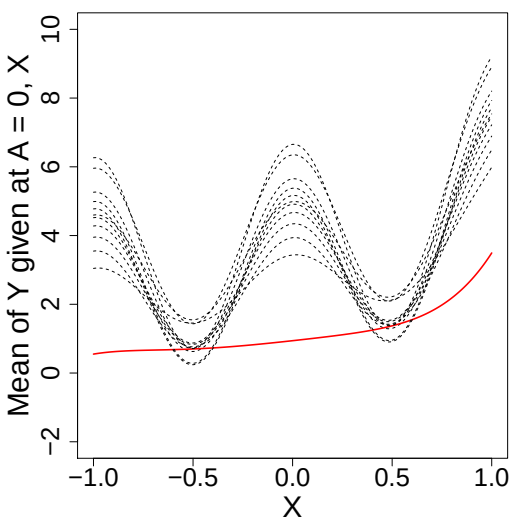
$$\widehat{\mu}(a, x) = \xi(a, x) + N(5n^{-1/\alpha}, n^{-1/\alpha}) \cdot \cos(2\pi x) + N(5n^{-1/\alpha}, n^{-1/\alpha}) \cdot \cos(2\pi a)$$
$$\widehat{\pi}(a \mid x) = \phi(a;\ \text{mean} = \kappa(x) + N(n^{-1/\alpha}, 0.5n^{-1/\alpha}) \cdot \cos(2\pi x),\ \text{sd} = 1)$$

for $n = 500$ (the sample size used), $\alpha = \{2, 4, 6, 8, 10, 15\}$ and where $\phi(a; \mu, \sigma^2)$ is the density of a truncated normal and the terms $N(\mu, \sigma)$ denote independent Normal random variables. The estimators are fluctuations of the true curves where the fluctuations scale as $n^{-1/\alpha}$. We estimate $p(a)$ as $n^{-1} \sum_{i=1}^{n} \widehat{\pi}(a \mid X_i)$.

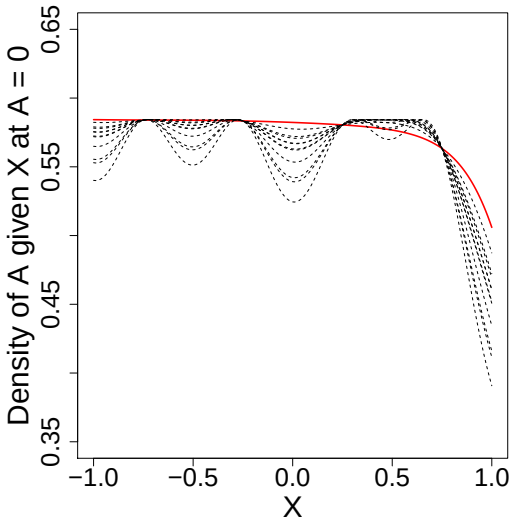
As an example of the ERM-based estimator, we consider orthogonal series regression, where the basis that we use is the Legendre polynomials basis. The number of terms ranges from 2 to 8. For the DR-Learner, we consider local linear regression with Gaussian kernel and bandwidth taking value in bw $= \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Finally, we consider first-order (the estimator of Colangelo and Lee [2020]) and second-order estimators based on the higher-order estimator construction. We use a Gaussian kernel for the term $K_{ht}(a)$, with bandwidth taking value in bw and the first eleven Legendre polynomials (normalized) as the basis in $\Pi(x_i, x_j)$. We estimate Ω by its empirical counterpart $\widehat{\Omega} = \mathbb{P}_n\{b(X)b(X)^T K_{ht}(A)\}$.

To compare the estimators' performance, we evaluate the dose-response $\theta(a) = b(a)^T\beta + \frac{1}{2}\int_{-1}^{1} b(x)^T \beta dx$ at 5 points equally spaced in $[-0.5, 0.5]$. At each point t, we approximate the mean-square-errors of the estimators by averaging their errors across 500 simulations. At each point t, we thus have one estimate of the MSEs for each tuning parameter value (number of basis or bandwidth value). To compare the estimators at each point, we consider the best-performing tuning parameter in terms of MSEs. In practice, this is not viable; potential alternatives would be to select the bandwidth via some form of cross-validation or simply to report a sequence of estimates for tuning parameter value. We finally compute a weighted mean of the MSEs with weight proportional to the density of A at t.

Figure 5.4 reports the results. We have included the MSEs for an oracle DR-Learner estimator that has access to the true nuisance functions to give a reference value. As expected, the performance of the estimators is similar when the error in the nuisance estimators is small. As the error increases, however, the second-order estimator performs better. Across the regimes for the nuisance errors that we considered, the first-order estimator performs better than either the one based on orthogonal series regression (ERM-based) or the one based on local polynomial regression (DR-Learner). In future work, it would be interesting to explore if this conclusion holds even when $\pi(a \mid x)$ and $\mu(a, x)$ have vastly different smoothness levels.



(a) True outcome model $x \mapsto \mathbb{E}(Y \mid A = 0, X = x)$ (red) and simulated estimator (black, dotted lines) when $\alpha = 8$.



(b) True conditional density $x \mapsto \pi(0 \mid X = x)$ (red) and simulated estimators (black, dotted lines) when $\alpha = 8$.

Figure 5.3: Examples of estimators of the true nuisance functions $\mu(a, x)$ and $\pi(a \mid x)$ and simulation results.

We conclude with a word of caution. In all the results contained in this work, we have not kept track of constant terms. While in asymptotic regimes, constants do not matter, in finite samples they might. Our simulated estimators would thus converge to the truth with the desired rate of order $n^{-1/\alpha}$ even if we consider fluctuations $cn^{-1/\alpha}$ for any constant c. Perhaps not surprisingly, we find that our simulation setup is sensitive to the choice of the constants multiplying the rate. In this sense, while encouraging, our limited simulation results should be interpreted with caution. We leave the design and implementation of larger simulation experiments to future work. We refer the reader to Li et al. [2005] for a comprehensive simulation study illustrating the superior performance of estimators based on higher order influence functions in the context of pathwise differentiable parameters.

5.6 Conclusions and future directions

In this work, we have explored the possibility of improving existing approaches to doubly-robust estimation of a dose-response curve by considering estimators based on DR-Learning framework and higher-order influence functions. We have shown that an estimator akin to the higher-order estimator of the average treatment effect described in Robins et al. [2017a] perform better than existing estimators, at least under certain smoothness conditions. In addition, we have specialized recent advancements on regression estimation with estimated outcomes to the dose-response settings and introduced two new doubly-robust estimators of the dose-response curve. A small simulation experiment has corroborated our theoretical results in finite samples.

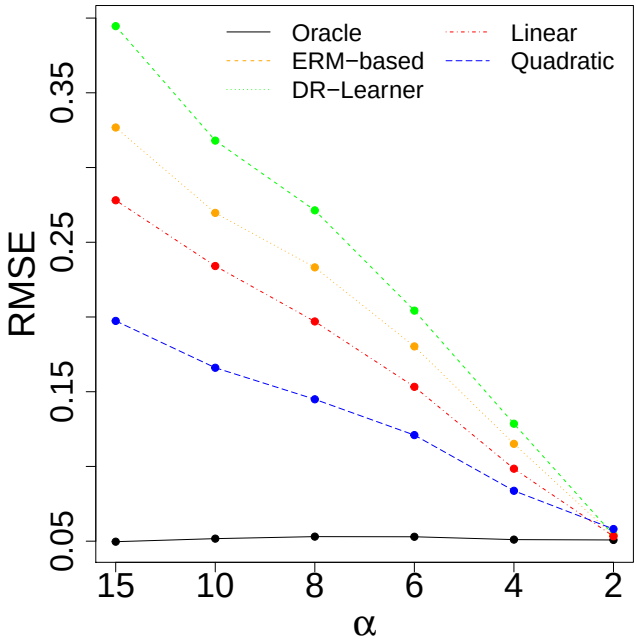


Figure 5.4: Estimated MSEs for different estimators of the dose-response across 500 simulations.

We have also described a flexible method to bound the causal dose-response function in the presence of unmeasured confounding.

Many open questions remain. First, and perhaps most importantly, a minimax lower bound for estimating the dose-response curve has not been described in the literature, to the best of our knowledge. Computing a lower bound on the risk of any estimator of this parameter is instrumental for understanding under what conditions, if any, the higher order estimator that we have proposed can be improved. Second, the higher-order estimator is currently not capable of tracking the smoothness of the dose-response when the conditional density of the treatment given the covariates, viewed as a function of the treatment alone, is less smooth than the dose-response itself. It is unclear if this stems from an intrinsic limitation of our estimator, the upper bound on the risk that we have computed is not tight enough or this is part of the minimax rate. A potential avenue for future research is to investigate the possibility of constructing a higher-order estimator that is based on regressions of some particular pseudo-outcomes onto A.

Finally, our results are about convergence of the estimators in mean-square-error. We leave the study of the inferential properties of the estimators discussed here for future work.

Chapter 6

Causal inference for the effect of mobility on Covid-19 deaths

This chapter is taken from my work supervised by Larry Wasserman, Valérie Ventura and Edward H. Kennedy, which was published in the Annals of Applied Statistics [Bonvini et al., 2022b].

6.1 Introduction

During a pandemic, it is reasonable to expect that reduced social mobility will lead to fewer deaths. But how do we quantify this effect? In this paper we combine ideas from mechanistic epidemic models with modern causal inference tools to answer this question using state level data on deaths and mobility. Our goal is not to provide definitive estimates for the effects but rather to develop some methods and highlight the challenges in doing causal inference for pandemics. We also show how a generative epidemic model motivates a semiparametric causal model.

We use state death data at the weekly level. The data are available at the daily county level but the weekly state level data are more reliable. Indeed, the data are subject to many reporting issues. It is not uncommon for a state to fail to report many deaths for a few days and then suddenly report a bunch of unreported deaths on a single day. The problems are worse at the county level. Also, there are many small counties with very little data. We find using weekly state level data to be a good compromise between the quantity and quality of the data. We also note that epidemic analyses, such as flu surveillance, are generally done at the weekly level.

Epidemics are usually modeled by using generative models, which fully specify the distribution of the outcome (deaths). The most common epidemic models relate exposure, infections, recoveries and deaths by way of a set of differential equations. The simplest version is the SIR model (susceptible, infected, recovered) but there are many flavors of the model. We review the basic model in Section 6.4.

Instead of a generative model, we use a marginal structural model (MSM) (Robins [2000], Robins et al. [2000]). An MSM is a semiparametric model that directly models the effect of mobility on death without specifying a generative model. Because it is semiparametric, it makes fewer assumptions than a generative model. However, our MSM is motivated by a modified SIR-type generative model.

We model deaths in each state separately to reduce confounding due to state differences. After obtaining model parameter estimates for each state, we will be interested in the causal question: what would happen if we set mobility to a certain value? For example, how many deaths would have occurred if mobility had been reduced earlier, or if people had remained more vigilant throughout? We follow standard causal language and refer to changing mobility as an intervention. A different notion of intervention would be a policy change like closing schools. In this case, mobility is a mediator meaning that the intervention affects the outcome through mobility. In this paper we focus on the effect of mobility on deaths and refer to hypothetically setting mobility to a certain value as an intervention. Providing estimates of the effect of mobility on deaths is valuable so that we can tell policy makers what mobility level they should aim for with their interventions. Analyzing the effect of interventions is also of interest but in this paper we focus on the effect of mobility on deaths.

We will see that the data provide evidence for an effect of mobility. But the data are very limited. As mentioned above, we use state-specific models with weekly resolution due to concerns about data quality and unmeasured confounding due to geographic differences. The result is that we have about 40 observations per state. With so little data, we are restricted to use fairly simple models. We do find significant causal effects but we conduct sensitivity analyses that show that the effects need to be interpreted cautiously. This sensitivity analysis includes assessing the impact of model assumptions and unobserved confounding.

Related Work. A number of researchers have considered modeling the effect of causal interventions (such as mobility and masks) on Covid-19. Notable examples are Unwin et al. [2020], Chang et al. [2020], and IHME [2020]. These authors develop very detailed epidemic models of the dynamics of the disease. One advantage of such an approach is that one can then consider the effects of a large array of potential interventions. Further, the models themselves are of great interest for understanding the dynamics of Covid-19. However, these models are very complex, and they involve a large number of parameters including parameters for various latent variables. Fitting such models and assessing uncertainty is challenging. Some authors take a Bayesian approach with informative priors. Others use heuristics such as reporting intervals based on using various settings of the parameters. To the best of our knowledge, it is not known how to get valid, frequentist confidence intervals in these complex models. This is not meant as a criticism of these papers but rather, this reflects the intrinsic difficulty of dealing with such models. Furthermore, when used for causal analysis, parametrically specified epidemic models are susceptible to a problem known as the null paradox which we discuss in Section 4.2.

In contrast, our goal is to make the model as simple as possible and to use standard estimating equation methods so that standard errors can be obtained fairly easily. We do

not claim that our approach is superior but we do believe that the model and the resulting confidence intervals are more transparent. Getting precise results from our simple model turns out to be challenging and raises doubts about the accuracy of published studies using highly complex models.

The papers by Chernozhukov et al. [2020] and Xiong et al. [2020] are much closer to ours. The authors of Chernozhukov et al. [2020] use a set of causal linear structural equations to model weekly cases as a function of social behavior (mobility) and social behavior as a function of policies. They model several policies simultaneously and they model all states simultaneously. They do obtain valid frequentist confidence intervals. Xiong et al. [2020] construct a measure of mobility inflow and using daily county level cases they fit a linear structural model to relate cases to mobility inflow. Our approach differs in several ways: we model deaths, we focus only on the effect of mobility, we model one state at a time, and we use a MSM rather than a generative model. By modeling within each state, we have much less data at our disposal, which makes modeling challenging. On the other hand, the threat of confounding due to state differences is reduced. By using a marginal structural model, our approach is semiparametric and so makes fewer assumptions. Unlike these authors, we focus on deaths instead of cases because we find the data on cases to be quite unreliable in general; for example, the availability of testing changed over time in various ways within and across states. Moreover, the data early in the pandemic are very important and this is when case data were least reliable. Also, we place a strong emphasis on sensitivity analysis. These analyses complement each other nicely.

Paper Outline. We describe the data in Section 6.2. In Section 6.3 we review some basics of causal inference. In Section 6.4 we construct the models that we will use and we explain how the models are fit in Section 6.5. The results are presented in Section 6.6. Concluding remarks are in Section 6.7.

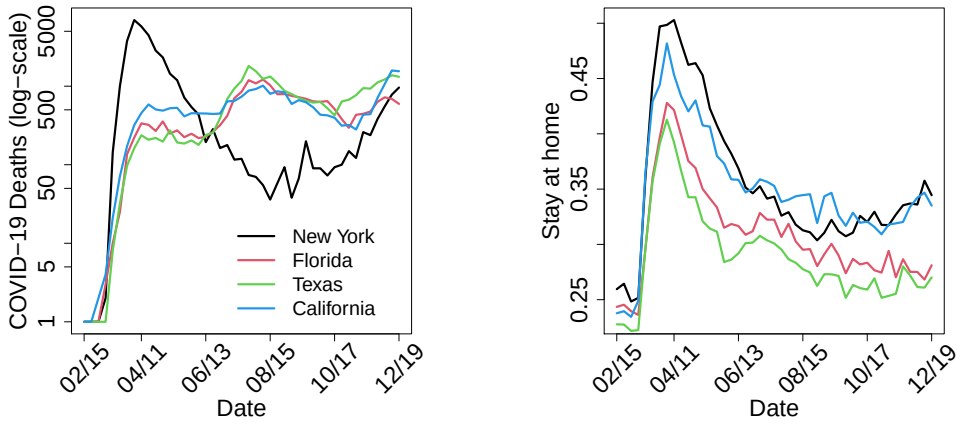
6.2 Data

As mentioned earlier, we model each state separately, at the weekly level. The data for each state have the form

$$(A_1, Y_1), \ldots, (A_T, Y_T)$$

where A_t is mobility on week t and Y_t is the number of deaths due to Covid-19 on week t. We obtained our data from CMU's Delphi group (cmu.covidcast.edu) which gets the death data from Johns Hopkins (https://coronavirus.jhu.edu) and the mobility data from Safegraph (safegraph.com). The data are from Feb 15 2020 to December 19 2020.

Figure 6.1 shows log deaths $L_t = \log(Y_t + 1)$ and "proportion at home" A_t which is one of the mobility measures, for four states. This is the fraction of mobile devices that did not leave the immediate area of their home. In this case, a higher value means less mobility so we can think of this measure as anti-mobility. This is the variable we will use throughout. In the rest of the paper we standardize mobility by subtracting A_1 from each value of A_t so that mobility starts at zero.



(a) Plot of log deaths versus time, from Feb 15 2020 to December 19 2020, for four populous states.

(b) Plot of anti-mobility measure "stay at home (percent)" versus week.

Figure 6.1: Plots of log deaths and anti-mobility across time.

6.3 Causal Inference

In this section, we briefly review basic ideas from causal inference. Consider weekly mobility and death data $(A_1, Y_1), \ldots, (A_T, Y_T)$ in one state. Define $\overline{A}_t = (A_1, \ldots, A_t)$ and $\overline{Y}_t = (Y_1, \ldots, Y_t)$ for $t \geq 1$.

Now consider the causal question: what would Y_t be if we set $\overline{A}_t$ equal to some value $\overline{a}_t = (a_1, \ldots, a_t)$? Let $Y_t^{\overline{a}_t}$ denote this counterfactual quantity. It is important to distinguish the observed data $(\overline{A}_T, \overline{Y}_T)$ from the collection of unobserved counterfactual random variables

$$\left\{ Y^{\overline{a}_T} : \overline{a}_T \in \mathbb{R}^T \right\},$$

which is an infinite collection of random vectors, one for each possible mobility trajectory $\overline{a}_T$. We make the usual consistency assumption that $\overline{Y}_T = \overline{Y}_T^{\overline{A}_T}$. To make sure this is clear, consider a simple case where a subject gets either treatment $A = 1$ or control $A = 0$. In this case, the random variables are (A, Y, Y^0, Y^1) and the consistency assumption is that the observed outcome Y satisfies $Y = Y^1$ if $A = 1$ and $Y = Y^0$ if $A = 0$.

Causal inference when the treatment varies over time is subtle. It may be tempting to simply regress Y_T on the past and get the regression coefficient for mobility. This strategy has serious problems because $\overline{Y}_{T-1}$ are both confounding and mediating variables. Indeed, previous deaths can affect both future mobility and future deaths, while also being affected by previous mobility. More precisely, a large number of deaths implies a large number of infections which can cause future infections which then cause future deaths, and a large number of deaths

might scare people into staying home. So we must adjust for past deaths. A common principle in epidemiology is to adjust for pre-treatment variables but not for post-treatment variables. But Y_s comes after A_{s-1} and before A_{s+1} making it both a pre-treatment and post-treatment variable. So how do we properly define the causal effect?

The solution is to use Robins' g-formula. Assuming for the moment that there are no other confounding variables except past deaths, Robins [1986] proved that the mean of $Y_t^{\overline{a}_t}$ is given by the g-formula:

$$\psi(\overline{a}_t) \equiv \mathbb{E}[Y_t^{\overline{a}_t}] = \int \cdots \int \mathbb{E}[Y_t|\overline{A}_t = \overline{a}_t, \overline{Y}_{t-1} = \overline{y}_{t-1}] \prod_{s=1}^{t-1} p(y_s|\overline{y}_{s-1}, \overline{a}_s)\, dy_s; \tag{6.1}$$

$\psi(\overline{a}_t)$ is the causal effect we seek to estimate. (We note that some authors denote $\mathbb{E}[Y_t^{\overline{a}_t}]$ by $\mathbb{E}[Y_t|\text{do}(\overline{a}_t)]$.) When there are other confounders X_t besides past deaths, the formula becomes

$$\psi(\overline{a}_t) \equiv \int \cdots \int \mathbb{E}[Y_t|\overline{A}_t = \overline{a}_t, \overline{Y}_{t-1} = \overline{y}_{t-1}, \overline{X}_{t-1} = \overline{x}_{t-1}] \times \prod_{s=1}^{t-1} p(y_s, x_s|\overline{y}_{s-1}, \overline{a}_s, \overline{x}_{s-1})\, dy_s\, dx_s.$$

Intuitively, the g-formula can be obtained as follows. The density of $(\overline{y}_t, \overline{a}_t)$ can be written as

$$p(\overline{y}_t, \overline{a}_t) = \prod_{s=1}^{t} p(y_s|\overline{y}_{s-1}, \overline{a}_s) p(a_s|\overline{a}_{s-1}, \overline{y}_{s-1}). \tag{6.2}$$

Now replace $p(a_s|\overline{a}_{s-1}, \overline{y}_{s-1})$ with a point mass at a_s (i.e. the A's are fixed, no longer random) and then find of the mean of Y_t from this new distribution. It will be useful later in the paper to bear in mind that $\psi(\overline{a}_t) \equiv \psi(\overline{a}_t, p)$ is a functional of the joint density p from (6.2).

For the causal effect $\psi(\overline{a}_t)$ to be identified we require three standard assumptions. These are: (1) there is no unmeasured confounding. Formally, this means that at each time, the treatment is independent of the counterfactuals given the past measured variables. (2) The distribution of treatment has a positive density. (3) Counterfactual consistency: If $\overline{A}_t = \overline{a_t}$ then $Y_t = Y^{\overline{a}_t}$. Later we add a fourth assumption, namely, that the dependence of mobility on the past satisfies a Markov condition.

The next question is: how do we estimate $\psi(\overline{a}_t)$? A natural idea is to plug-in estimates of all the unknown quantities in the g-formula which leads to

$$\widehat{\psi}(\overline{a}_t) \equiv \int \cdots \int \widehat{\mathbb{E}}[Y_t|\overline{A}_t = \overline{a}_t, \overline{Y}_{t-1} = \overline{y}_{t-1}] \prod_{s=1}^{t-1} \widehat{p}(y_s|\overline{y}_{s-1}, \overline{a}_s)\, dy_s. \tag{6.3}$$

As discussed in Robins [1989, 2000], Robins et al. [2000] there are a number of problems with this approach, called g-computation. If we plug-in nonparametric estimates, we quickly face

the curse of dimensionality. If we use parametric estimates, we encounter the null-paradox (Robins and Wasserman [1997]): there may be no setting of the parameters which can represent the case where there is no treatment effect, i.e., there is no setting of the parameters which makes $\psi(\overline{a}_t)$ a constant function of $\overline{a}_t$. We discuss the null paradox further in Section 4.2.

An alternative approach to estimating $\psi(\overline{a}_t)$ is to directly specify a parametric functional form $g(\overline{a}_t, \beta)$ for $\psi(\overline{a}_t)$. Such a model is called a marginal structural model (MSM). Robins et al. [2000] showed that β can be estimated by solving the following inverse-probability-weighted estimating equation:

$$\sum_t W_t \, h(\overline{A}_t)(Y_t - g(\overline{A}_t, \widehat{\beta})) = 0 \tag{6.4}$$

where

$$W_t = \prod_{s=1}^{t} \frac{\pi(A_s|\overline{A}_{s-1})}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1})} \tag{6.5}$$

and h is an arbitrary function. The choice of h affects the efficiency of the estimator but not its consistency. We discuss the choice of h in Section 6.5.

An MSM is a semiparametric model in the sense that it leaves the data generating process unspecified, subject to the restriction that the functional $\psi(\overline{a}_t)$ has a specific form. Specifically, let us write $\psi(\overline{a}_t)$ as $\psi(\overline{a}_t, p)$ to make it clear that $\psi(\overline{a}_t, p)$ depends on the joint density of the data $p(\overline{a}_T, \overline{y}_T)$ from (6.2). The generative model we are using is then

$$\mathcal{P} = \Big\{ p(\overline{a}_T, \overline{y}_T) : \text{ there exists } \beta \text{ such that } \psi(\overline{a}_t, p) = g(\overline{a}_t, \beta) \text{ for all } t \Big\}. \tag{6.6}$$

The model g is typically chosen to be interpretable. For example, suppose that $g(\overline{a}_t, \beta) = \beta_0 + \beta_1 \sum_s a_s$. Then the effect of the parameter settings is simple (i.e., mean outcomes only depend linearly on the amount of cumulative treatment), and the null (of no treatment effect) simply corresponds to $\beta_1 = 0$. It is important to keep in mind that $g(\overline{a}_t, \beta)$ is not a model for the entire data generating process, just for marginal treatment effects, i.e., how mean outcomes under different treatment sequences are connected. Marginal structural models are often chosen to be some arbitrary but simple parametric model. Instead, we choose to specify the marginal structural model $g(\overline{a}_t, \beta)$ by the following route: we tentatively specify a generative model and find a closed form formula $g(a, \beta)$ for $\psi(\overline{a}_t)$. We then drop the generative model and use $g(\overline{a}_t, \beta)$ as a MSM. We explain this in more detail in the next section.

Remark 15. *There is a difference between the standard MSM setup and the one we are considering that warrants mentioning. Typically one assumes access to n different time series $(Z_1, ..., Z_n)$, with each series $Z = \{(A_1, Y_1), ..., (A_T, Y_T)\} = (\overline{A}_T, \overline{Y}_T)$ observed for n different independent units (e.g., states). There, one could have a different estimating equation at each time, for example,*

$$\sum_i W_{ti} \, h_t(\overline{A}_{ti})(Y_{ti} - g_t(\overline{A}_{ti}, \widehat{\beta}_t)) = 0$$

where the i subscript denotes weights, treatments, outcomes, etc. for series i. If there are common

parameters across timepoints, then these estimating equations could be combined, for example by summing over time, or using a generalized method of moments approach, etc. However, we model states individually, and so do not assume different states are independent. This leaves us with one observation per state at each time, which we then combine across time (but only within state) to obtain estimating equation (6.4)*. This represents the trade-off between independence versus modeling assumptions (e.g., Markov assumptions in the weights, or linearity in $g(\cdot)$): the less we require of one, the more we require of the other.*

6.4 Models

Epidemics are often modeled using differential equations that describe the evolution of certain subgroups over time. Perhaps the most common is the SIR (Susceptible, Infected, Recovered) model (Kermack and McKendrick [1927], Brauer et al. [2012], Bjørnstad [2018]) described by the equations

$$\begin{aligned}\frac{dS_t}{dt} &= -\frac{\alpha I_t S_t}{N}\\ \frac{dI_t}{dt} &= \frac{\alpha I_t S_t}{N} - \gamma I_t\\ \frac{dR_t}{dt} &= \gamma I_t,\end{aligned}$$

where N is population size, S_t is the number of susceptibles, I_t is the number of infected, R_t are the removed (due to infection) at time t and $\alpha > \gamma$. Solving the second equation conditional on S_t yields $I_t = I_{t-1} e^{\int_{t-1}^{t} \alpha S_u/N - \gamma du}$, which can be discretized as

$$I_t \approx I_{t-1} e^{\alpha S_t/N - \gamma} \tag{6.7}$$

when $S_u \approx S_t$ for all $u \in (t-1, t)$. Without intervention, the epidemic grows exponentially, peaks when $S_t/N = \gamma/\alpha$ and then decays exponentially. There are numerous generalizations of this model including stochastic versions, discretized versions and models with more states besides S, I and R.

6.4.1 The Mobility Model

Our proposed MSM is

$$g(\overline{a}_t, \nu_0, \beth, f) = \sum_{s=1}^{t} f(s,t) e^{\nu_0(s) + \sum_{r=1}^{s} \beth(a_r)} \tag{6.8}$$

with nuisance functions f, ν_0 and $\beth$. The model is motivated by (6.7) as we now explain.

The basic idea of the SIR model is that there is a natural tendency for an epidemic to increase exponentially at the beginning. But there are also elements that reduce the epidemic such as the depletion of susceptible individuals due to infection. At the beginning of a pandemic,

reduction of susceptibles will play a negligible role. On the other hand, interventions like lockdowns, school closings etc can have a drastic effect. These considerations lead us to the following working model. We use this working model only to suggest a form for the MSM.

Let I_t denote *new* infections in week t. Let

$$\begin{aligned} A_t &\sim Q_t \\ I_t &= I_{t-1} e^{c_t + \beth(A_t)} + \delta_t \\ Y_t &= \sum_{s=1}^{t} f(s,t) I_s + \xi_t \end{aligned} \tag{6.9}$$

where Q_t is an arbitrary distribution depending on $(\overline{A}_{t-1}, \overline{I}_{t-1}, \overline{Y}_{t-1})$, δ_t and ξ_t are mean 0 random variables (independent of the other variables), $f(s,t)$ denotes the probability that someone infected at time s dies of COVID at time t, the parameter c_t is a positive number and $\beth$ is a smooth function. Notice that the infection process (second equation) has an exponential growth form as in (6.7), but we model the exponent directly as a function of mobility and time instead of stipulating a model for the susceptibles S_t. Here, c_t represents the evolution of the epidemic without intervention and $\beth(A_t)$ is the effect of mobility. We allow c_t to vary with t to make the model more general and to allow the spread of Covid-19 to depend on the availability of susceptibles. We write

$$f(s,t) = d(s) f_0(s,t) \tag{6.10}$$

where $d(s)$ is the probability that someone infected at time s will eventually die of COVID and $f_0(s,t)$ is the probability that someone infected at time s and who will eventually die, will die at time t. Following Unwin et al. [2020] we take $f_0(s,t)$, on the scale of days, to be the density of $T_1 + T_2$ where T_1 (time from infection to symptoms) is Gamma with mean 5.1 and coefficient of variation 0.86 and T_2 (time from symptoms to death) is Gamma with mean 18.8 and coefficient of variation 0.45. The resulting distribution can be accurately approximated by a Gamma with mean 23.9 days and coefficient of variation 0.40. Finally, we integrate this distribution over 7 day bins to get $f_0(s,t)$ on a weekly scale. A directed graph illustrating the model is given in Figure 6.2.

At this point, we might use (6.9) as our model. But the I_t's are not observed. Furthermore, a non-linear, sequentially specified parametric generative model can suffer from serious anomalies when used for causal inference. In particular, such a model can suffer from the *null paradox* (Robins [1986, 1989], Robins and Wasserman [1997]). This means that there may be no parameter values that satisfy both (i) Y_t is conditionally dependent on past values of A_s and (ii) the null hypothesis of no treatment effect holds. We explain this point in more detail in Section 6.4.2.

Instead, we apply the g-formula to the model specified by (6.9) to find $\mathbb{E}[Y_t^{\overline{a}_t}]$ and use the

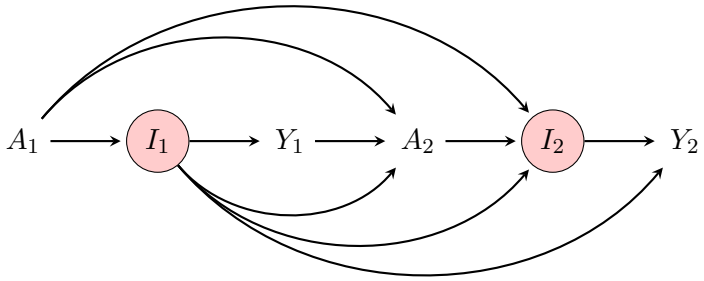


Figure 6.2: Directed graph illustrating the working model. Infections I_t are unobserved. We use this model to find the form $g(a;\beta)$ of the causal effect $\psi(a)$. But when we estimate β we use a semiparametric estimating equation approach; we do not fit the above model to the data.

resulting function as an MSM. This yields

$$\mathbb{E}[Y_t^{\overline{a}_t}] = \sum_{s=1}^{t} f(s,t) e^{\nu_0(s)+\sum_{r=1}^{s} \beth(A_r)} \equiv g(\overline{a}_t, \nu_0, \beth, f)$$

where $\nu_0(s) = \log I_1 + \sum_{r=1}^{s} c_r$. (We treat I_1 as an unknown parameter that is absorbed into ν_0.) Now we abandon the working model and just interpret $g(\overline{a}_t, \nu_0, \beth, f)$ directly as a model for the counterfactual $\mathbb{E}[Y^{\overline{a}_t}]$, that is, as an MSM. Put another way, we start with the model (6.9), find $g(\overline{a}_t, \nu_0, \beth, f) = \mathbb{E}[Y^{\overline{a}_t}]$, and then expand the model to include all joint distributions that satisfy $\mathbb{E}[Y_t^{\overline{a}_t}] = g(\overline{a}_t, \nu_0, \beth, f)$. This defines the model (6.6).

The MSM can be fit with the estimating equation (6.4), which corrects for confounding due to past deaths, not by modeling the entire conditional process, but by weighting by propensity weights W_t given by (6.5). This MSM approach allows us to be agnostic about whether it is our motivating model (6.9) that holds, or some other much more complicated data-generating process. In fact, one can go further and take a completely agnostic view, in which the marginal structural model is not assumed to be correct, but only viewed as an approximation to the true, and possibly very complex, underlying counterfactual mean [Neugebauer and van der Laan, 2007].

To summarize, our approach involves three steps.

1. Tentatively specify a working model for infections I_t.

2. Find the resulting functional form $g(a;\beta)$ for $\psi(a)$ using the g-formula. We use $g(a;\beta)$ as our MSM.

3. Drop the working model and fit the MSM semiparametrically without further assumptions on the data generating process.

It is important to emphasize that when we estimate the causal parameter β, we do not

assume any model for the epidemic process. Note that the model for I_t in step 1 is very flexible but it does assume that the mobility effect is additive. An alternative would be to use a more sophisticated epidemic model for $\mathbb{E}[I_t|\text{past}]$ in step 1. It would be interesting to do this and this would help unify the traditional approach to epidemic modeling with the MSM approach we are using. However, the implied function $g(a;\beta)$ would not be in closed form and it would be very hard to fit this model especially with only 40 observations.

6.4.2 The Null Paradox

To see how the null paradox works, consider a simple example with four time ordered variables (A_0, I_1, A_1, I_2) where A_0 and A_1 are mobility and I_1 and I_2 are number of infected, which we assume are observed. This is a snippet of the entire time series. A simple epidemic model is

$$\begin{aligned} A_0 &\sim p(a_0) \\ \log I_1 &= \beta_0 + \epsilon \\ A_1 &\sim p(a_1|I_1, A_0) \\ \log I_2 &= \theta_0 + \theta_1 A_0 + \theta_2 \log I_1 + \theta_3 A_1 + \delta \end{aligned}$$

where ϵ and δ are, say, mean 0 Normal random variables. This is meant to capture exponential growth of I_t (i.e. the SIR model at early times with no recovered individuals). By applying the g-formula, the causal effect of setting $A = (A_0, A_1)$ to $a = (a_0, a_1)$ is

$$\psi(a) = \mathbb{E}[\log I_2^a] = \theta_0 + \theta_1 a_0 + \theta_2 \beta_0 + \theta_3 a_1.$$

This means that, if we simulated the epidemic model with $A = (A_0, A_1)$ set to $a = (a_0, a_1)$, the mean of $\log I_2$ would precisely be $\theta_0 + \theta_1 a_0 + \theta_2 \beta_0 + \theta_3 a_1$. Suppose now that there is an unobserved variable U that affects I_1 and I_2. For example, U could represent the general health of the population. The variable U is not a confounder as it does not affect A_0 or A_1. The causal effect is still given by the g-formula with no change. Suppose now that neither A_0 or A_1 have a causal effect on I_2. The set up is shown in Figure 6.3. Despite the fact that A_0 and A_1 have no causal effect on I_2, it may be verified that I_2 is conditionally dependent on A_0 and A_1. (This follows since I_1 is a collider on the path I_2, U, I_1, A_0, A_1.) It follows that the maximum likelihood estimators $\widehat{\theta}_1$ and $\widehat{\theta}_3$ are not zero (and in fact converge to nonzero numbers in the large sample limit). The estimated causal effect is

$$\widehat{\psi}(a) = \widehat{\theta}_0 + \widehat{\theta}_1 a_0 + \widehat{\theta}_2 \widehat{\beta}_0 + \widehat{\theta}_3 a_1$$

and will therefore be a function of a even when a has no causal effect.

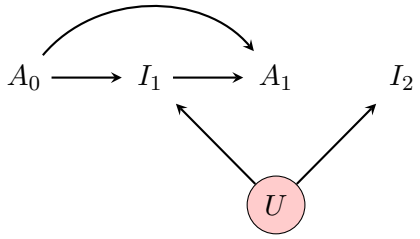


Figure 6.3: The null paradox. The directed graph is a snippet of the time series. Mobility is (A_0, A_1) and number of infected individuals is (I_1, I_2). The latent variable U is not a confounder as it has no arrows to mobility. Neither A_0 nor A_1 have a causal effect on I_2. The variable I_1 is a collider, meaning that two arrowheads meet at I_1. This implies that I_2 and (A_0, A_1) are dependent conditional on I_1. The estimate of the parameters that relate I_2 to (A_0, A_1) in the epidemic model will be non-zero even though there is no causal effect.

The details of the model were not important. A similar model is

$$\begin{aligned} A_0 &\sim p(a_0) \\ I_1 &\sim p(i_1|A_0) \\ A_1 &\sim p(a_1|I_1, A_0) \\ I_2 &\sim p(i_2|A_0, I_1, A_1) \end{aligned}$$

where $\mathbb{E}[I_2|A_0, I_1, A_1] = e^{\beta_0+\beta_1 A_0+\beta_2 A_1} I_1$. In this case

$$\mathbb{E}[I_2^a] = e^{\beta_0+\beta_1 a_0+\beta_2 a_1}\mathbb{E}[I_1|A_0].$$

The same argument shows that the estimate will be a function of a even when a has no causal effect.

6.4.3 Simplified Models

The MSM in (6.8) is not identified without further constraints. We will take $\beth(A_s) = \beta A_s$ so that

$$\mathbb{E}[Y_t^{\bar{a}_t}] = \sum_{s=1}^{t} f(s,t)e^{\beta\sum_{r=1}^{s} A_r+\nu_0(s)}.$$

Solving the estimating equation with this model is unstable and computationally prohibitive. Hence we make two approximations. First, we take $f_0(s,t)$ in (6.10) to be a point mass at $\delta = 4$ weeks (approximately its mean). Then we get

$$\mathbb{E}[Y_t^{\bar{a}_t}] = e^{d(t-\delta)+\nu_0(t-\delta)+\beta M_t}$$

where $M_t \equiv M(\overline{a}_t) = \sum_{s=1}^{t-\delta} a_s$. If we approximate $\log \mathbb{E}[Y_t^{\overline{a}_t}]$ with $\mathbb{E}[\log(Y_t^{\overline{a}_t})]$ we further obtain

$$\mathbb{E}[L_t^{\overline{a}_t}] = \log d(t-\delta) + \nu_0(t-\delta) + \beta M_t \tag{6.11}$$

where $L_t = \log(Y_t + 1)$. Note that $\partial \mathbb{E}[L_t^{\overline{a}_t}]/\partial a_s = \beta$ for any $s \leq t - \delta$ so β has a clear meaning. Finally, we take

$$\nu(t) \equiv \log d(t-\delta) + \nu_0(t-\delta) = \sum_{j=1}^{k} \beta_j \phi_j(t)$$

where $\phi_1, \ldots, \phi_k$ are orthogonal polynomials starting with $\phi_1(t) = t$. This model is easy to fit and will be used in Section 6.6. Note that the probability of dying $d(t)$ is allowed to change smoothly over time, which it likely did as hospitals were better prepared during the second wave. We have consistently found that using $k = 1$ leads to unreasonable results which means that the disease exponential growth changes with time other than through mobility. The method for choosing k is described in Section 6.6.2.

The model in (6.11) was used independently in Shi and Ban [2020] with $k = 1$. They used the model for curve fitting and they showed that this simple model fits the data surprisingly well. However, we find that making $\nu(t)$ non-linear (i.e. $k > 1$) is important.

We will also consider a different approach to fitting the model. Specifically, we will use deconvolution methods to estimate the unobserved infection process $I_1, \ldots, I_T$. The first equation in (6.9) implies $\mathbb{E}[I_t] = e^{\nu(t)+\beta \sum_s A_s}$ suggesting the MSM

$$\mathbb{E}[L_t^{\overline{a}_t}] = \nu(t) + \beta M_t$$

which is the same as (6.11) except that now $L_t = \log(I_t)$ and $M_t = \sum_{s=1}^{t} a_s$ rather than $M_t = \sum_{s=1}^{t-\delta} a_s$.

Remark 16. *We have regularized the model by restricting $\nu(t)$ to have a finite basis expansion. We also considered a different approach in which $\nu(t)$ is restricted to be increasing which seems a natural restriction if $\nu(t)$ is supposed to represent the growth of the pandemic in lieu of intervention. (This is valid only at the start of the pandemic; later in the pandemic, ν could be decreasing.) Using the methods in Liao and Meyer [2018], Meyer [2008, 2018] we obtained estimates and standard errors. The results were very similar to the results in Section 6.6.*

Counterfactual Estimands. Now we discuss some causal quantities that we can estimate from the model. Let $\overline{a}_t = (a_1, \ldots, a_t)$ be a mobility profile of interest. After fitting the model we will plot estimates and confidence intervals for counterfactual deaths

$$\theta_t = \exp\Big\{\mathbb{E}[L^{\overline{a}_t}]\Big\} \tag{6.12}$$

under mobility regime $\overline{a}_t$, $t = 1, \ldots, T$.

We will consider the following three interventions:

$$\begin{aligned}
&\text{Start one week earlier} \ : \ \overline{a}_T = (A_2, A_3, \ldots, , A_{T+1}) \\
&\text{Start two weeks earlier} \ : \ \overline{a}_T = (A_3, A_4, \ldots, A_{T+2}) \\
&\text{Stay vigilant} \ : \ \overline{a}_T = (A_1, A_2, \ldots, A_9, A_{10}, A_{10}, A_{11}, A_{11}, A_{12}, A_{12}, A_{13}, A_{13}, \ldots)
\end{aligned}$$

The first two interventions aim to assess COVID-19 infections if we had started sheltering in place one and two weeks earlier. The last intervention halves the slope of the rapid decrease in stay at home mobility after the initial peak in week 9 that is clearly visible in Fig.6.1. See Figure 6.6.

6.5 Fitting the Model

Now we discuss the method for estimating the model.

6.5.1 Fitting the Semiparametric Model

Recall the MSM

$$\mathbb{E}[L_t^{\overline{a}_t}] = \nu(t) + \beta M(\overline{a}_t) \tag{6.13}$$

where $\nu(t) = \beta_0 + \sum_{j=1}^{k} \beta_j \phi_j(t)$. We estimate $\nu(t)$ and β by solving the estimating equation

$$\sum_t h_t(\overline{a}_t) W_t [L_t - (\widehat{\nu}(t) + \widehat{\beta} M(\overline{a}_t))] = 0 \tag{6.14}$$

corresponding to (6.4). We discuss the estimation of the weights W_t in Section 6.5.2. As is often done for MSMs we choose

$$h_t(\overline{a}_t) = (1, \phi_1(t), \ldots, \phi_k(t), M(\overline{a}_t))^T$$

since solving the estimating equation then corresponds to using least squares with weights W_t. The estimating equation is then the derivative of the weighted sum of squares set to zero.

Recall from (6.12) that $\theta_t = e^{\psi(\overline{a}_t)} = e^{\nu(t) + \beta M(\overline{a}_t)}$ which we estimate by $\widehat{\theta}_t = e^{\widehat{\nu}(t) + \widehat{\beta} M(\overline{a}_t)}$. We obtain approximate confidence intervals using the delta method and the aymptotic normality of estimating equations estimators. The asymptotic variance is based on the heteroskedasticity and autocorrelation consistent HAC sandwich estimator (Newey and West [1987]).

6.5.2 Estimating the Stabilized Weights

To estimate the marginal structural model we need to estimate the stabilized weights

$$W_t = \prod_{s=1}^{t} \frac{\pi(A_s | \overline{A}_{s-1})}{\pi(A_s | \overline{A}_{s-1}, \overline{Y}_{s-1})};$$

see (6.5). One approach is to plug in estimates of the numerator and denominator densities into the formula for W_t. But estimating these densities is not easy and ratios of density estimates can be unstable. The problem is exacerbated when we multiply densities. Instead we use a moment-based approach as in Fong et al. [2018], Zhou and Wodtke [2018]. The idea is to estimate the vector of weights $W_1, \ldots, W_T$ by noting that they need to satisfy certain moment constraints. Our method is similar to the approach in Zhou and Wodtke [2018].

We rewrite $W_t = \prod_{s=1}^{t} V_s$ where

$$V_s \equiv V_s(\overline{A}_s, \overline{Y}_{s-1}) = \frac{\pi(A_s|\overline{A}_{s-1})}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1})}.$$

Let $\widetilde{h}_1(a_t)$ and $\widetilde{h}_2(y_{t-1})$ be arbitrary functions and define their centered versions by

$$\begin{aligned} h_1(a_t) &= \widetilde{h}_1(a_t) - \mu_t \\ h_2(y_{t-1}) &= \widetilde{h}_2(y_{t-1}) - \eta_t \end{aligned}$$

where the conditional means are

$$\begin{aligned} \mu_t &\equiv \mu_t(\overline{A}_{t-1}) = \mathbb{E}[\widetilde{h}_1(A_t)|\overline{A}_{t-1}] \\ \eta_t &\equiv \eta_t(\overline{A}_{t-\delta-1}, \overline{Y}_{t-2}) = \mathbb{E}[\widetilde{h}_2(Y_{t-1})|\overline{A}_{t-\delta-1}, \overline{Y}_{t-2}]. \end{aligned}$$

Weighted products of these functions have mean zero since

$$\begin{aligned} \mathbb{E}[h_1(A_t)h_2(Y_{t-1})W_t] &= \int \cdots \int h_1(a_t)h_2(y_{t-1})p(\overline{a}_t, \overline{y}_{t-1})W_t(\overline{a}_t, \overline{y}_{t-1})\, d\overline{a}_t\, d\overline{y}_{t-1} \\ &= \int \cdots \int h_1(a_t)h_2(y_{t-1})\pi(a_t|\overline{a}_{t-1}, \overline{y}_{t-1})p(y_{t-1}|\overline{a}_{t-1}, \overline{y}_{t-2})p(\overline{a}_{t-1}, \overline{y}_{t-2}) \\ &\qquad \times \frac{\pi(a_t|\overline{a}_{t-1})}{\pi(a_t|\overline{a}_{t-1}, \overline{y}_{t-1})} \left(\prod_{s=1}^{t-1} V_s\right) \; d\overline{a}_t\, d\overline{y}_{t-1} \\ &= \int \Bigg\{ \omega(\overline{y}_{t-2}, \overline{a}_{t-1}) \int h_1(a_t)\pi(a_t|\overline{a}_{t-1})da_t \\ &\qquad \times \int h_2(y_{t-1})p(y_{t-1}|\overline{a}_{t-1}, \overline{y}_{t-2})dy_{t-1} \Bigg\} d\overline{a}_{t-1}\, d\overline{y}_{t-2} \\ &= 0 \end{aligned}$$

from the definition of h_1 and h_2, where

$$\omega(\overline{y}_{t-2}, \overline{a}_{t-1}) = p(\overline{y}_{t-2}, \overline{a}_{t-1}) \prod_{s=1}^{t-1} V_s.$$

Thus, the weights are characterized by the moment constraints

$$\mathbb{E}[h_1(A_t)h_2(Y_{t-1})W_t] = 0. \tag{6.15}$$

As in Zhou and Wodtke [2018] we estimate the weights by finding W_t to satisfy

$$\mathbb{E}[h_1(A_t)h_2(Y_{t-1})W_t] = 0$$

for a set of functions h_1, h_2. This requires estimating these moments and estimating μ_t and η_t. To proceed, we make a Markov assumption, namely

$$\mathbb{E}[\widetilde{h}_1(A_t)|\overline{A}_{t-1}] = \mathbb{E}[\widetilde{h}_1(A_t)|A_{t-1}, \ldots, A_{t-\kappa}]$$

and

$$\mathbb{E}[\widetilde{h}_2(Y_{t-1})|\overline{A}_{t-\delta-1}, \overline{Y}_{t-2}] = \mathbb{E}[\widetilde{h}_2(Y_{t-1})|A_{t-1-\delta}, \ldots, A_{t-\kappa-\delta}, Y_{t-2}, \ldots, Y_{t-\kappa}]$$

for some κ. We will use $\kappa = 1$ in our analysis. Moreover, we assume homogeneity so that the functions μ_t and η_t do not depend on t. Under the homogeneous Markov assumption, μ_t and η_t can be estimated by regression. For example, if $\kappa = 1$, μ can be estimated by regressing $\widetilde{h}_1(A_2), \ldots, \widetilde{h}_1(A_T)$ on $A_1, \ldots, A_{T-1}$. (We tried both linear and nonparametric regression and obtained similar weights from each approach so we have used linear regression in our results.) The sample versions of the moment conditions (6.15) are then

$$\frac{1}{T}\sum_t H_{tj}W_t = 0$$

where

$$H_{tj} = (\widetilde{h}_{1j}(A_t) - \widehat{\mu}_j)(\widetilde{h}_{2j}(Y_{t-1}) - \widehat{\eta}_j)$$

and $\{(\widetilde{h}_{1j}, \widetilde{h}_{2j}) : \ j = 1, \ldots, J\}$ are a set of pairs of functions, $\widehat{\mu}_j$ is the estimate of

$$\mathbb{E}[\widetilde{h}_1(A_t)|A_{t-1}, \ldots, A_{t-\kappa}]$$

and $\widehat{\eta}_j$ is the estimate of $\mathbb{E}[\widetilde{h}_2(Y_{t-1})|A_{t-1-\delta}, \ldots, A_{t-\kappa-\delta}, Y_{t-2}, \ldots, Y_{t-\kappa}]$.

The moment conditions do not completely specify the weights. As in the above references we add a regularization term, in this case, $(1/2)\sum_t (W_t - 1)^2$ and we require $\sum_t W_t = T$. This leads to the following minimization problem: minimize $W_1, \ldots, W_T$ in

$$\frac{1}{2}\sum_t (1 - W_t)^2 + \lambda_0 \sum_t (W_t - T) + \sum_{j=1}^{J} \lambda_j \sum_t W_t H_{tj} \tag{6.16}$$

where the λ_j's are Lagrange multipliers. The solution to the minimization is

$$W = \mathbf{1} - H(H^T H)^{-1}[H^T \mathbf{1} - \mathbf{D}] \tag{6.17}$$

1. Choose the order κ of the Markov assumption.
2. Choose J pairs of functions $\left\{(\widetilde{h}_{1j}(a), \widetilde{h}_{1j}(y)) : \ j = 1, \dots J\right\}$.
3. Estimate $\mu_j = \mathbb{E}[\widetilde{h}_{1j}(A_t)|A_{t-\kappa}, \dots, A_{t-1}]$ and $\eta_j = \mathbb{E}[\widetilde{h}_{2j}(Y_{t-1})|A_{t-\kappa-\delta-1}, \dots, A_{t-\delta-1}, Y_{t-1-\kappa}, \dots, Y_{t-2}]$ by regression.
4. Compute the weights $W_1, \dots, W_n$ from (6.17).
5. Fit the model $L_t = \beta \sum_{i=1}^{t-\delta} A_s + \nu(t) + \epsilon_t$ using weighed least squares with weights $W_1, \dots, W_n$.

Figure 6.4: Steps for fitting the model.

where $W = (W_1, \dots, W_T)$, $\mathbf{1}$ is a vector of $1's$, $\mathbf{D} = (T, 0, \dots, 0)^T$ and

$$H = \begin{pmatrix} 1 & H_{11} & \cdots & H_{1N} \\ 1 & H_{21} & \cdots & H_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & H_{T1} & \cdots & H_{TN} \end{pmatrix}$$

and N is the total number of moment constraints. In our case we choose $h_{11}(a) = a$, $h_{12}(a) = a^2$, $h_{21}(y) = y$, $h_{22}(y) = y^2$.

To include other time varying confounders X_t one should replace $h_2(y_{t-1})$ with two functions:

$$h_2(y_{t-1}) = \widetilde{h}_2(y_{t-1}) - \mathbb{E}[\widetilde{h}_2(y_{t-1})|\overline{X}_{t-1}, \overline{A}_{t-1}, \overline{Y}_{t-2}]$$

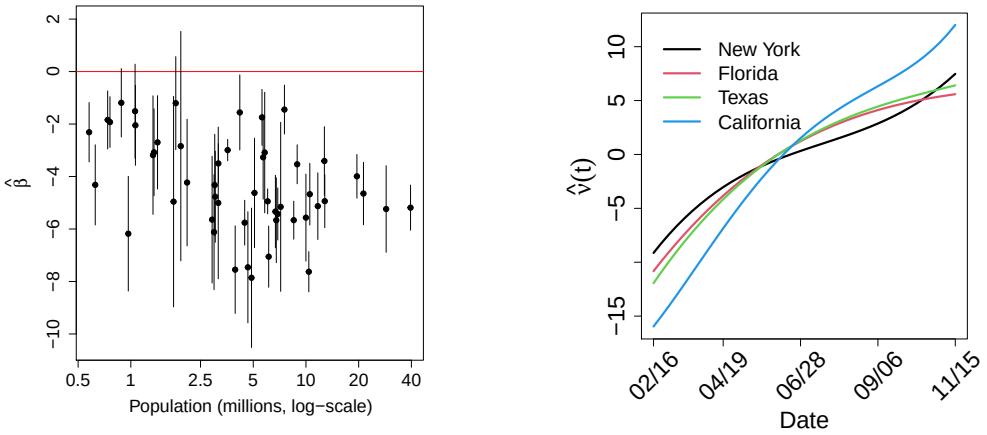
and

$$h_3(x_{t-1}) = \widetilde{h}_3(x_{t-1}) - \mathbb{E}[\widetilde{h}_3(x_{t-1})|\overline{X}_{t-2}, \overline{A}_{t-1}, \overline{Y}_{t-2}].$$

The steps for fitting the model are summarized in Figure 6.4. Note that we cannot include past infections as a confounder since this variable is not observed. We choose not to include past cases or hospitalizations because the former is terribly biased downward at the beginning of the epidemic, and reliable data for the second is difficult to obtain. We need to assume that adjusting for past deaths serves as an adequate surrogate for infections, cases and hospitalizations. We address the more general problem of unoberved confounding in Section 6.2.

6.6 Results

In this section we give results for the mobility measure 'proportion of people staying at home.' We begin by showing the results of fitting the MSM to each state. Then we report on various types of sensitivity analysis.



(a) Plot of $\widehat{\beta}$ and 95% confidence interval from the marginal structural model (6.13) for each state, versus state log population. A value of $\beta = -5$, for example, means that log deaths are reduced by 5 if A_s is increased by one percent at any time s.

(b) Plot of $\widehat{\nu}(t)$ for four populous states.

Figure 6.5: Estimates of the MSM parameters defined in (6.13).

6.6.1 Main Results

Figure 6.5a shows 95 percent confidence intervals for $\widehat{\beta}$ for each state from the marginal structural model in (6.13). We computed standard errors as if the weights were known, which results in valid but potentially conservative inference as long as the weight models are correctly specified [Tsiatis, 2007]. The estimates are mostly negative, as would be expected, since higher A_s means less mobility. Interestingly, we find that there turns out to be little confounding due to past deaths, as the fits with and without the estimated weights (not shown) are very similar. Nevertheless, we keep the weights in all the fits as a safeguard. In Section 6.6.2 we investigate this further by doing a sensitivity analysis.

Figure 6.5b shows the estimated smooth function $\widehat{\nu}(t)$ in (6.13) for four states. The functions are increasing with slopes tapering off as time goes on, and picking up again in NY and CA in late September, consistent with deaths rising at that time in these two states; see Figure 6.1. The shape of $\widehat{\nu}(t)$ is consistent with the usual epidemic dynamics where it is assumed that this component should initially grow (linearly with no interventions and with an infinite pool of susceptibles) on the log-scale at the start of the epidemic and then decrease. Some of the non-linearity probably reflects the fact that the probability $d(t)$ of dying decreases over time due to better hospital treatment, social distancing changes, and the number of susceptibles to COVID-19 decreases over time as recovered patients are likely immune for some period post-infection.

Next we consider counterfactual deaths $\theta_t = \exp(\mathbb{E}[L^{\bar{a}_T}])$ in (6.12) for the three mobility

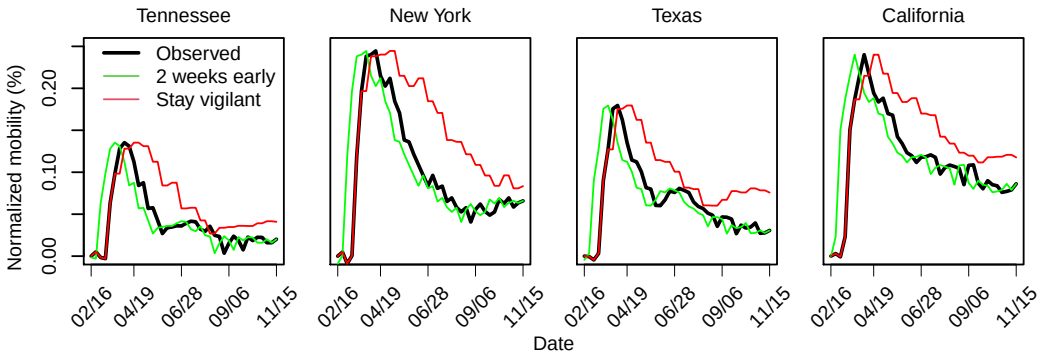


Figure 6.6: The observed mobility curves and hypothetical interventions for four states. Mobility has been standardized to have value 0 at the beginning of the series. All plots are on the same scale.

scenarios described at the end of section 4; two mobility scenarios are shown in Figure 6.6 for four states. Figure 6.7 shows the estimates and pointwise 95 percent confidence bands for θ_t for these four states. The plots for all states are in the Supplement [Bonvini et al., 2022b].

Finally, Figure 6.8 shows 95 percent confidence intervals for $\sum_t \exp\left(\mathbb{E}[L^{\bar{a}_t}]\right) - \sum_t Y_t$ and for $\left(\sum_t \exp\left(\mathbb{E}[L^{\bar{a}_t}]\right) - \sum_t Y_t\right) / \sum_t Y_t$ under the 'stay vigilant' scenario. We refer to these as total and relative excess deaths, where a negative excess means that lives would be saved. Of course, this number is larger for more populous states, although relative to the total number of observed deaths, all states small and large would have benefited equally from more sustained vigilance. Note that the confidence interval for New York (fourth from right) is very large. New York experienced the pandemic early and responded with large values of A_s so it is believable that further vigilance may not have a large effect.

We now compare our results to those in Unwin et al. [2020]. They use a sophisticated model of the epidemic dynamics so a direct comparison is difficult. They estimate a parameter R_t that measures how many individuals an infected person will infect. Using a Bayesian approach, they find a 95 percent posterior interval for the change in R_t for the U.S. when setting mobility to its maximum value is [26.5,77.0]. The log of the change in R_t is roughly equivalent to $-\beta$ in our setting. On the log scale, their interval is [3.3,4.3]. Our effect sizes are similar and slightly larger for the large states. For the middle sized states our effect estimates vary somewhat and are sometimes larger and sometimes smaller than theirs. Overall, the effect estimates are quite similar which is reassuring given how vastly different the methods are. Another point of comparison is Chernozhukov et al. [2020] who consider a very ambitious model which includes multiple policy interventions and multiple mobility measures (which they call behavior) simultaneously and the model is over all states. Their estimate of the mobility effect on log cases is -0.54 with a standard error of .19. Unlike Unwin et al. [2020], this estimate is very different from ours and we do not know why. They are using a different measure of mobility (they used Google mobility) which might have some effect. It is possible that some of the mobility effect might be absorbed into their policy effect which could happen if there is model misspecification.

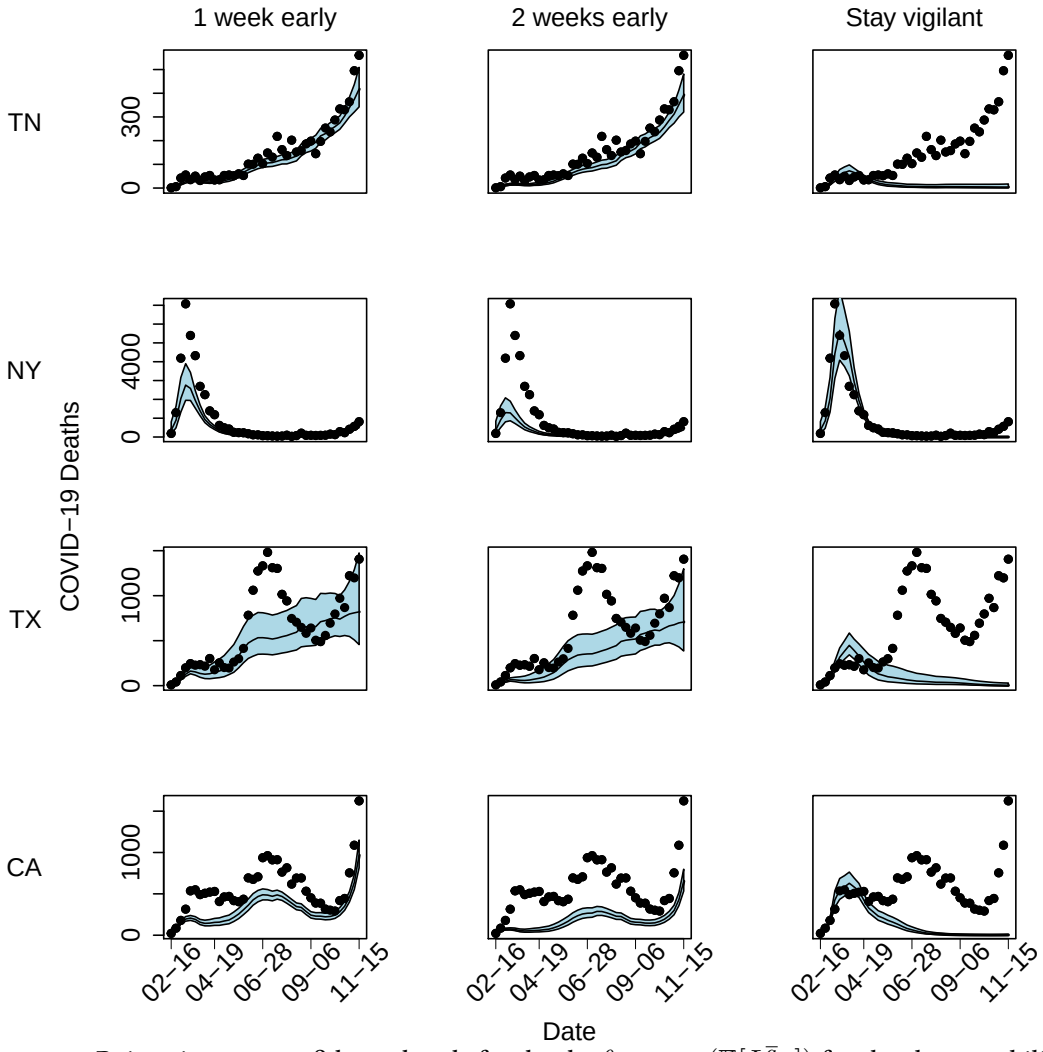


Figure 6.7: Pointwise 95% confidence bands for deaths $\theta_t = \exp(\mathbb{E}[L^{\overline{a}_T}])$ for the three mobility scenarios $\overline{a}_T$ described at the end of section 4; see also Figure 6.6. Each row is a different state. Each column is a different scenario, start one week early, start two weeks early and stay vigilant. The epidemic in NY started early so staying at home sooner had a large impact. The same is true for PA, IL, MI, NJ, MA. Staying home earlier would not have had as much impact in states such as TN that did not suffer the epidemic early. Staying more vigilant would have had a large impact except for New York. Some lack of fit in the early time period is evident in Texas, where counterfactual deaths exceed observed deaths under 'stay vigilant' where mobility has not yet been changed.

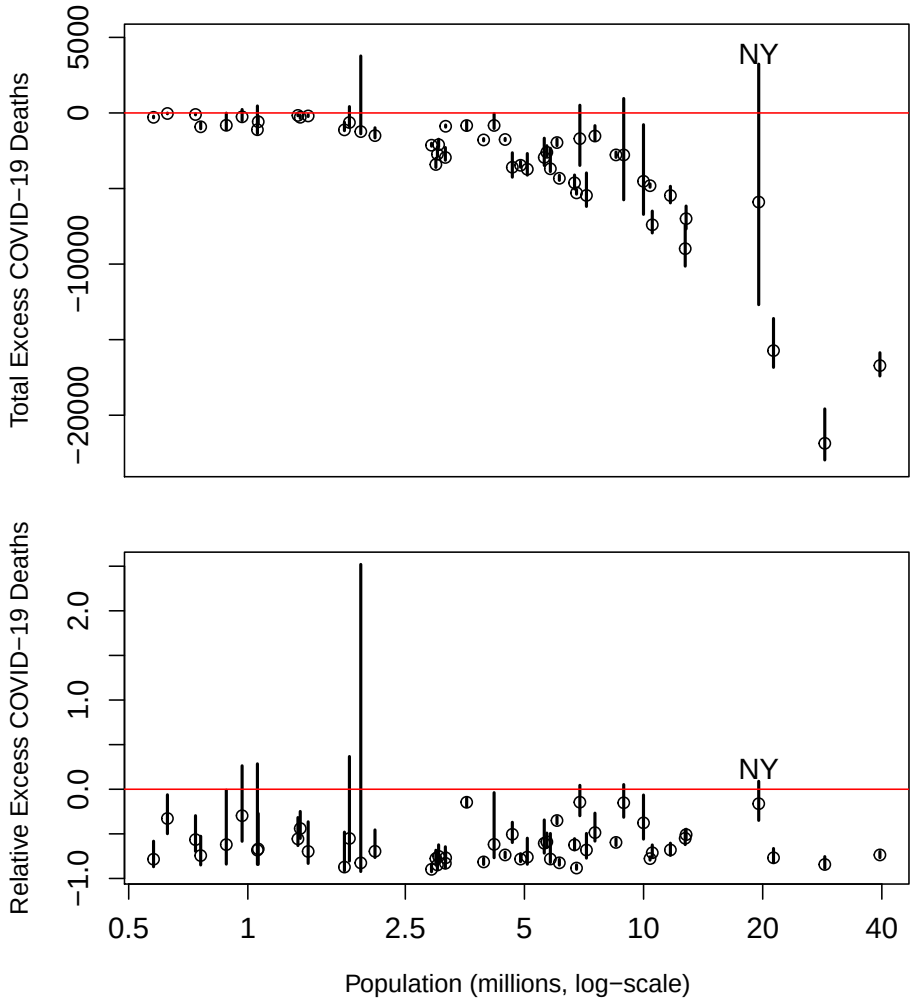


Figure 6.8: 95% confidence intervals for total excess deaths $\sum_t \exp\left(\mathbb{E}[L^{\bar{a}_t}]\right) - \sum_t Y_t$ (top) and relative excess deaths $\left(\sum_t \exp\left(\mathbb{E}[L^{\bar{a}_t}]\right) - \sum_t Y_t\right) / \sum_t Y_t$ (bottom) under the 'stay vigilant' scenario. The confidence intervals for NY (fourth from right) and a handful of other states include zero and suggests that staying more vigilant would not have significantly impacted the death toll. On the other hand, many states, small and large, could have reduced their death tolls by over a half.

6.6.2 Sensitivity Analysis

We have made a number of strong assumptions in our model. Our preference would be to weaken these assumptions and use nonparametric methods but the data are too limited to do so. Instead, we now assess the sensitivity of the results to various assumptions. We consider various perturbations of our analysis. These include: (1) changing the model/estimation method (we replace the MSM with a generative model), (2) assessing the Markov assumption (which was used to estimate the weights), (3) checking the accuracy of the point mass approximation (which was used in Section 6.4.3 to simplify the model) and (4) assessing sensitivity to unmeasured confounding (we have assumed that the only time varying confounders are past values of mobility and death).

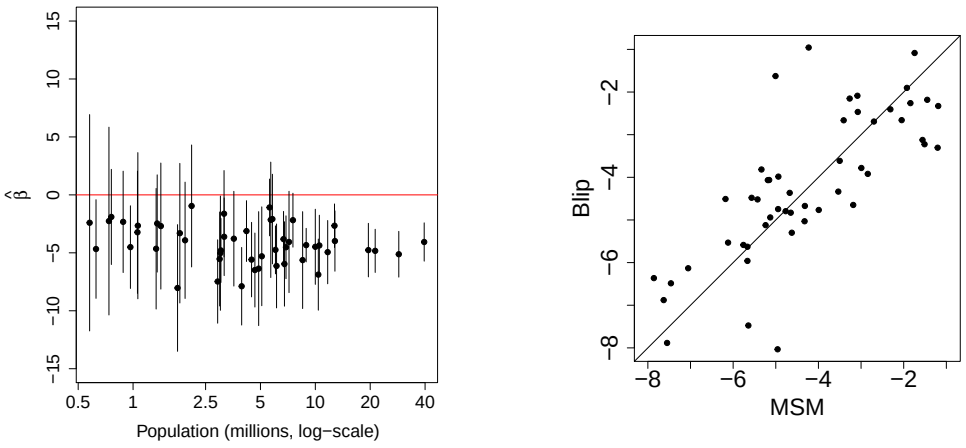
1. An Alternative Model. Here we compare the results from the MSM in (6.13) to the time series AR(1) model:

$$L_t = L_{t-1} + \beta A_{t-\delta} + r(t) + \epsilon_t \tag{6.18}$$

where $r(t)$ is a polynomial of degree $k-1$. This says that, apart from random error, L_t differs from L_{t-1} for two reasons, mobility $A_{t-\delta}$ and the natural increase $r(t)$ due to epidemic dynamics (at the start of the epidemic). If we apply the g-formula in (6.1) to this model, we find $\mathbb{E}[L_t^{\overline{a}_t}] = \beta M(\overline{a}_t) + \nu(t)$ where $\nu(t) = \sum_{s=1}^{t} r(s)$ is a polynomial of order k. Hence, this model is consistent with the MSM. In other words, this model is contained in the semiparametric model $\mathcal{P}$ defined in (6.6). This model resembles Robins' *blip models* (Robins [2000], Vansteelandt and Joffe [2014]) as it measures the effect of one blip of treatment $A_{t-\delta}$ so we will refer to (6.18) as the blip model. We will fit (6.18) by least squares. There are three reasons for fitting this model. First, it as a point of comparison for the MSM. Second, we are able to check residuals and model fit. Third, since it is a regression model, we can use AIC to choose the degree $k-1$ of $r(t)$. We also use this choice of k in the MSM. The degree k chosen by AIC is typically $k=1$ for small states and $k=3$ or $k=4$ for the larger states. A plot of the selected degree versus log population and versus log deaths is in the supplementary material [Bonvini et al., 2022b].

The left plot in Figure 6.9 shows the estimates of β and 95 percent confidence intervals for all the states from the blip model in (6.18), and the right plot compares the estimates of β from the MSM and blip models, where we see the similarity of the inferences. Since the blip model is a regression model, it makes sense to compare the observed data to the fits. Fig 6.10 shows the fitted values and the data for four states. The fit is not perfect but is reasonable. There are some large outliers in some states, mostly in the first few weeks of the pandemic where mobility A_t and log deaths L_t change rapidly. Because of this we also fitted a robust regression but the results did not change much.

2. The Markov Assumption. In Section 6.5.2, to estimate the weights, we have made the Markov assumption that $A_{t-\delta}$ is conditionally independent of the past given $(A_{t-1-\delta}, L_{t-1-\delta})$. We also assumed that L_t is conditionally independent of the past given $(A_{t-1-\delta}, L_{t-1})$. To assess



(a) Estimates of β from the blip model in (6.18) with 95% confidence intervals.

(b) Comparison of estimates of β from the blip model and the MSM in (6.13).

Figure 6.9: Estimates from the blip model compared with estimates from the MSM model.

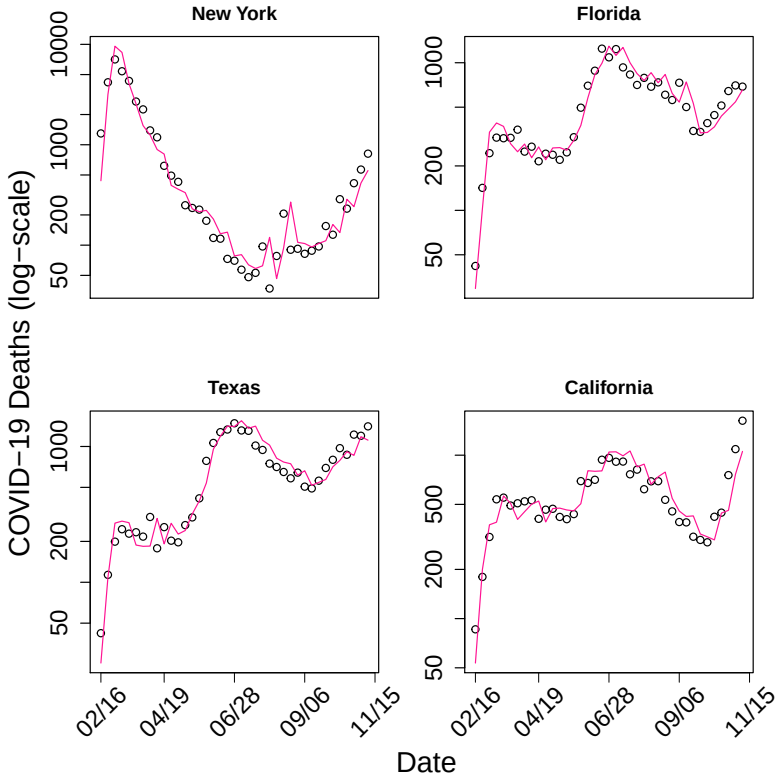


Figure 6.10: Observed log deaths in four states as functions of time with estimates (red) from the blip model in (6.18).

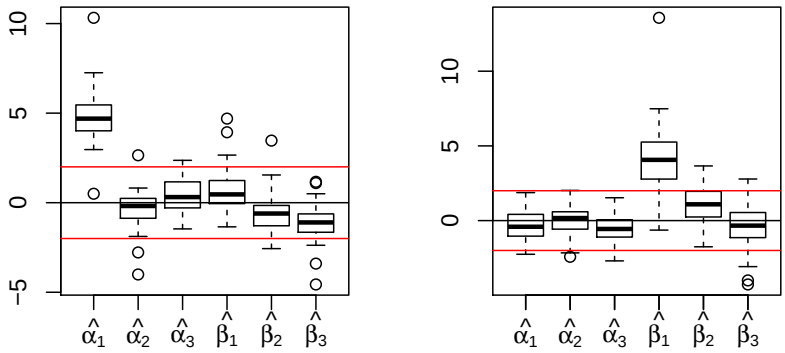


Figure 6.11: (Left) Boxplots across states of t-statistics for the parameters in the model for A_t as a function of the past. The horizontal red lines are at ± 2. Only $\widehat{\alpha}_1$ is consistently significantly different from zero across states, suggesting that the times series of at home mobility A_t is a memory one process. (Right) Same for Y_t. Only $\widehat{\beta}_1$ is consistently significantly different from zero across states, suggesting that the deaths times series Y_t is a memory one process.

this assumption, we fit the models

$$A_{t-\delta} = \alpha_0 + \alpha_1 A_{t-1-\delta} + \alpha_2 A_{t-2-\delta} + \alpha_3 A_{t-3-\delta} + \beta_1 L_{t-1-\delta} + \beta_2 L_{t-2-\delta} + \beta_3 L_{t-3-\delta} + \epsilon_t$$
$$L_t = \alpha_0 + \alpha_1 A_{t-\delta} + \alpha_2 A_{t-\delta-1} + \alpha_3 A_{t-\delta-2} + \beta_1 L_{t-1} + \beta_2 L_{t-2} + \beta_3 L_{t-3} + \delta_t.$$

Figure 6.11 shows boxplots of the t-statistics for these parameters. The evidence suggests that the first order Markov assumption is reasonable. The weak dependence of A_t on past values of Y_t is consistent with the weights W_t having almost no effect, i.e. there is little confounding due to past deaths. However, this assessment still assumes that the Markov assumption is homogeneous, that is, that the law of A_t given (A_{t-1}, Y_{t-1}) is constant over time. This assumption is not checkable without invoking further assumptions.

3. Point Mass Versus Deconvolution. Recall that in Section 6.4.3 we approximated $f_0(s,t)$ with a point mass at $t-\delta$ with $\delta = 4$. An alternative is to solve the estimating equation using g defined as in (6.8) but this is numerically very unstable. Yet another alternative to the point mass approximation is to estimate the number of infections I by deconvolution. From the number of infections, we can estimate the model parameters as in Section 6.5 without making the point mass approximation, using $\log(I)$ as the outcome variable. We infer $\widetilde{I}_t = d(t)I_t$ from the optimization:

$$\min_{I \geq 0} \|Y - F\widetilde{I}\|_2^2 + \lambda \sum_{r=2}^{T-1} (\widetilde{I}_r - \widetilde{I}_{r-1})^2, \tag{6.19}$$

where Y denotes the vector of weekly deaths and F is a matrix with (i,j)-entry equal to $f(i,j)$ if $j \leq i$ and zero otherwise; that is, F_{ij} is proportional to the probability of dying at

time j given that infection occurred at time i. The parameter λ is user-specified and represents a penalty imposed on non-smooth solutions. Because f is proportional to the density of a Gamma random variable, we have $F_{ii} = f(i,i) = 0$. To ensure nonzero elements on the diagonal of F, we remove the first row and last column (all zeros) from F and solve (6.19) using $Y = (Y_2, \ldots, Y_T)$, thus obtaining an estimate of $\widetilde{I} = (\widetilde{I}_1, \ldots, \widetilde{I}_{T-1})$. To enforce nonnegative values of I, we use the constrained optimization routine L-BFGS-B from optim in R. Using a penalty $\lambda = 1$, we report the inferred infections (up to proportionality) $\widehat{\widetilde{I}}$ (red line) for California, Florida, New York and Texas in Figure 6.12 along with the implied deaths computed as $F\widehat{\widetilde{I}}$. The latter match the observed deaths well, leading credence to this procedure. In Figure 6.13, we compare the estimates of β from the MSM using the point-mass approximation and those from the MSM using the estimates of infections from the deconvolution step. The estimates are in rough agreement as they lie near the diagonal.

4. Unmeasured Confounding. At time t, we treated $(A_1, Y_1), \ldots, (A_{t-1}, Y_{t-1})$ as confounders. Now suppose there is an unmeasured confounder U. We would like to assess $|\widehat{\beta}_U - \widehat{\beta}|$ where $\widehat{\beta}_U$ is the value of our estimate if we had access to U. This quantity is not identified and so any sensitivity analysis must invoke some extra assumption. Let $\Delta = |\widehat{\beta}_U - \widehat{\beta}|/\mathrm{se}(\widehat{\beta})$ denote the unobserved confounding on the standard error scale. So $\Delta = 0$ corresponds to no unmeasured confounding, $\Delta = 1$ corresponds to saying that the unmeasured confounding is the same size as the standard error, etc. For each state, we enlarge the confidence interval by $\Delta\,\mathrm{se}(\widehat{\beta})$. We can then ask: how large would Δ have to be so that the enlarged confidence interval would contain 0. Figure 6.14a shows this critical Δ. We see that for most states, it takes a fairly large Δ to lose statistical significance. A substantial number of medium to large states are quite robust to unmeasured confounding.

Adding other potential within state confounders would be desirable but, in a within state analysis, we can only accommodate time varying confounders. (A fixed confounder is a single variable with no replication and can only be used an across state analysis.) So far we do not have any within state time varying variables that would be expected to directly affect both A_t and Y_t. One could imagine that a variable like "the percentage of rural cases" could change over time and possibly affect both variables but we do not have such data.

Next we consider a second style of sensitivity analysis inspired by the approach in Rosenbaum [2010]. The effect of unmeasured confounding in our analysis is that the weights W_t are misspecified. If there are unobserved confounders U_t, then the correct weights are

$$\widetilde{W}_t = \prod_{s=1}^{t} \frac{\pi(A_s|\overline{A}_{s-1})}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1}, \overline{U}_{s-1})}$$

whereas we estimated the weights

$$W_t = \prod_{s=1}^{t} \frac{\pi(A_s|\overline{A}_{s-1})}{\pi(A_s|\overline{A}_{s-1}, \overline{Y}_{s-1})}.$$

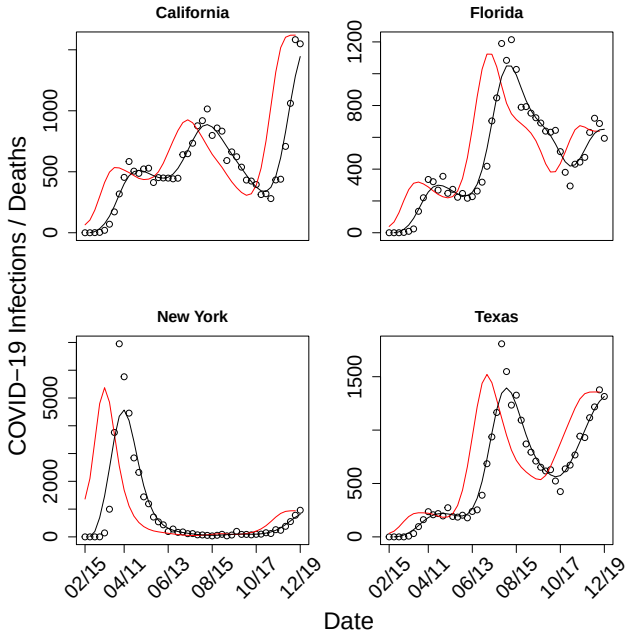


Figure 6.12: Inferred infections in four states. The red curve is $\widehat{\widetilde{I}}_t$, the estimate of the number of infections times the probability of dying if infected by Covid-19, $\widetilde{I}_t = d(t)I_t$. The black curve is deaths $F\widehat{\widetilde{I}}$ computed from the optimization with $\lambda = 1$ in (6.19), and the dots are the observed deaths.

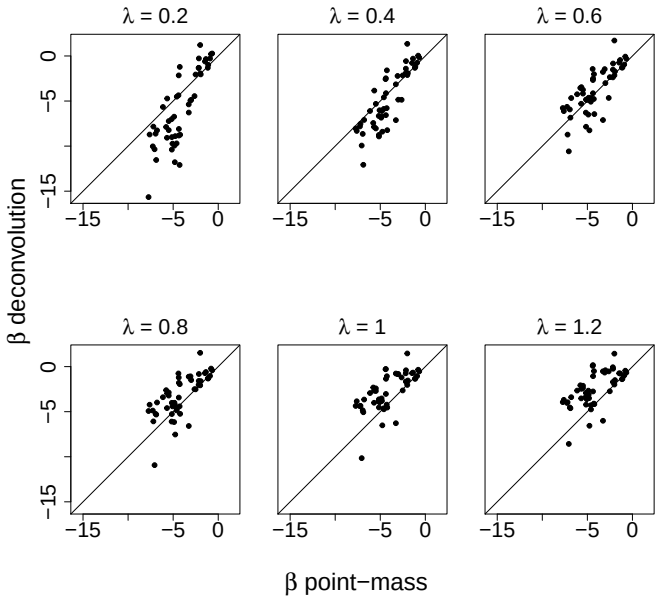
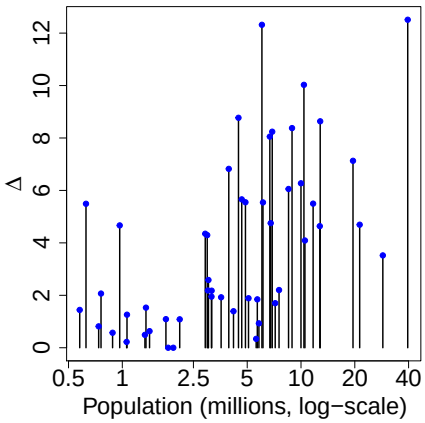
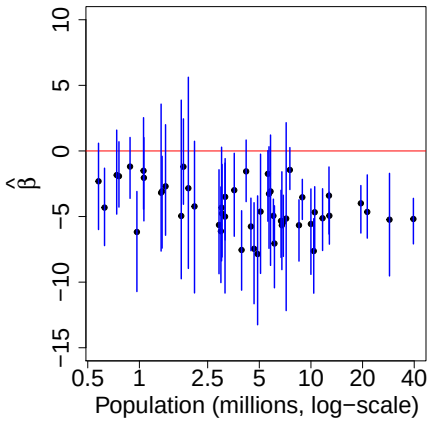


Figure 6.13: Comparison of estimates of $\widehat{\beta}$ from the MSM using the point-mass approximation versus using estimates of infections via deconvolution for different values of λ.



(a) Minimum value of Δ versus log-population for each state, such that unmeasured confounding of size $\Delta \, \mathrm{se}(\widehat{\beta})$ causes the confidence interval for β to contain 0. For most states, it takes a fairly large Δ to lose statistical significance.



(b) The blue line segments span the lower and upper bounds of $\widehat{\beta}$ over the weights $1/\Gamma \leq \widetilde{W}_t/W_t \leq \Gamma$ with $\Gamma = 3$. The black dots are the original point estimates. The effects for most large and medium states remain significant, indicating robustness to unmeasured confounding.

Figure 6.14: Unmeasured confounding sensitivity plots.

To assess this impact we find the maximum and minimum $\widehat{\beta}$ under the assumption that

$$\frac{W_t}{\Gamma} \leq \widetilde{W}_t \leq \Gamma W_t$$

for $t = 1, \ldots, T$ and some $\Gamma \geq 1$. Similar ideas for static, binary treatments have been considered in Yadlowsky et al. [2018], Zhao et al. [2019]. Figure 6.14b shows the bounds on $\widehat{\beta}$ using $\Gamma = 3$. Even with this fairly large value of Γ the effects for most large and medium states remain significant indicating robustness to unmeasured confounding. (Several methods for computing the bounds in this context can be found in Bonvini et al. [2022a].)

6.6.3 Across Versus Within States

We have focused on within state estimation. An alternative is to fit a model across states as well. Although we are skeptical of combining data over states we do so here for completeness. We fit the blip model (6.18) with common β and, rather than include state level covariates such as population size, proportion of residents in cities, etc., we use a fixed effect for each state. The resulting estimates of β and standard errors for $k = 1, 2, 3, 4$ are:

k	$\widehat{\beta}$	standard error
1	-5.20	0.27
2	-4.60	0.27
3	-3.82	0.34
4	-2.83	0.43

The estimates are consistent with the within state models. AIC chooses $k = 1$, which conflicts with the within state analysis which favors larger k for larger states. The likely reason is that combining states adds variability in the combined dataset since β's and $r(t)$'s are different between states, so there is less signal compared to the noise to estimate a more complicated relationship than a linear. A natural extension of this model is to use a random effects approach, although we do not pursue that here.

6.7 Discussion

Our approach to modeling the causal effect of mobility on deaths is to construct a marginal structural model whose parameters are estimated by solving an estimating equation. We model each state separately to reduce confounding due to state differences. Our approach has several advantages and disadvantages.

Our modeling assumptions are reasonable in the short term but not in the long term. Eventually, the effects of acquired immunity, masks, vaccinations etc might have to be accounted for by using a more complex form of ν. Also, the effect of mobility β could change with new variants.

Estimating the model parameters comes down to solving the estimating equation (6.14). Computing standard errors and confidence intervals is then straightforward. This is in contrast

to more traditional and Icarian epidemic modeling which requires estimating many parameters using grid searches or MCMC. Provably valid confidence intervals are elusive for those methods. On the other hand, the more detailed models might be more realistic and can capture effects that our simple model cannot capture. Moreover, our inferences are asymptotic in nature. When comparing exact Bayesian methods to approximate frequentist methods it is hard to argue that one approach is more valid than the other.

We believe that focusing on weekly data at the state level gives us the best chance of getting data of reasonable quality and helps avoid confounding related to state differences. Further, this allows the causal effect to vary between states. But this results in a paucity of data, a few dozen observations per state. This limits the complexity of the models we can fit and it requires that we make a homogeneous Markov assumption. A natural compromise worthy of future investigation would be to use some sort of random effects model to allow modeling all states simultaneously. This could also permit using data from other countries. At any rate, there is a tradeoff: within state analysis requires stronger modeling assumptions while analyzing all states together requires assuming independence and it assumes we can model all sources of between state confounding.

Detailed dynamic modeling versus the more traditional causal modeling done here (and in Chernozhukov et al. [2020]) represent two different approaches to causal inference for epidemics. It would be interesting to see a general comparison of these approaches, perhaps eventually leading to some sort of fusion of these ideas.

Finally, let us recap the null paradox. Any nonlinear, sequentially specified parametric model — which includes most epidemic models — has the following problem. There is no value of the parameters that allows both (i) the outcome is conditionally dependent on the intervention variable A and (ii) there is no causal effect of A. But, due to latent non-confounding variables U (see Figure 6.3), (i) and (ii) can both be true. This means that we would find a causal effect even if there is no such effect. We can in principle avoid the null paradox by using nonparametric models but then the model complexity explodes as T increases leading to the curse of dimensionality. Linear models avoid the null paradox but caution is still needed since the causal effect $\psi(a)$ involves complicated nonlinear functions of the regression parameters. Hence, the model is very difficult to interpret and the individual regression parameters do not have a causal interpretation. Also, most epidemic models are not linear.

The quickly growing literature on using sequentially specified epidemic models does include such models; see, for example, Bhatt et al. [2020], Scott et al. [2021], Unwin et al. [2020] and references therein. MSMs avoid the null paradox, and this is another reason for using MSMs (or some other semiparametric causal model such as structural nested models). In our case we motivated the MSM by starting with a sequentially specified model. This seems like a reasonable approach for using epidemic models to define an MSM but there may be other approaches as well.

Acknowledgments

The authors would like to thank Rob Tibshirani and the reviewers for providing helpful feedback on an earlier draft of the paper. Edward Kennedy gratefully acknowledges support from NSF Grant DMS1810979.

Chapter 7

Conclusions and future work

In this thesis, we have proposed several contributions to two of the main streams of causal inference research: identification and nonparametric functional estimation. In the first two chapters, we have proposed and analyzed a suite of methods to perform sensitivity analysis to the no-unmeasured-confounding assumption, which is the crucial, untestable assumption needed to identify popular causal effects in observational studies. In both chapters, we have exemplified our methods on real datasets.

In the second part of the thesis, we have proposed and analyzed new estimators of two commonly targeted causal estimands: the dose-response function and the level sets of the conditional average treatment effect (CATE). In each case, we have derived estimators achieving the best convergence rate currently known in the literature, for the models considered. In the CATE level sets problem, we have established that the rate of the proposed optimal estimator cannot be improved without introducing additional assumptions. In both problems, we have also analyzed the properties of other estimators that are arguably easier to implement in practice than the better performing ones.

In the third part of the thesis, we have conducted an analysis of the causal effects of reduced mobility on deaths due to Covid-19; we have focused on the early stages of the pandemic. We have proposed a semiparametric approach to causal inference in this challenging setting that is based on specifying a marginal structural model motivated by an epidemic model. We have complemented our main findings with sensitivity analyses showing that the results are not, for the most part, overly effected by deviations of the assumptions invoked in the main analysis.

Many open questions remain. The following is a summary of the most pressing questions motivated by this research that we would like to find an answer to.

Chapter 2:

- In deriving the bounds on the average treatment effect as a function of the proportion of units for which the treatment-outcome association is confounded, we have assumed that the indicator for whether a unit belongs to the "confounded" subgroup is idependent

of the outcome given the treatment and the covariates. This has allowed us to derive closed-form expressions for the bounds. Calculating the bounds without this untestable assumption appears to lead to prohibitive computational costs. Finding a clever algorithm to compute or approximate the bounds without relying on this independence assumption is an important open question.

- For each value of the sensitivity parameter $\epsilon \in [0, 1]$, our model returns bounds on the average treatment effect. We thus derive lower and upper curves as a function of the proportion ϵ. The seminal paper by Imbens and Manski [2004] shows that, if the goal is to carry out inference for a parameter that is only partially identified, there is a way to construct confidence intervals that 1) adapt to the length of the bounds and 2) can be considerably narrower than confidence intervals based on the intervals for the lower and upper end-points of the partial identification region. The method relies on the asymptotic normality of the estimators of the end-points of the identification interval. It remains an open question to adapt their construction to the case when the bounds are curves whose estimators converge weakly to a Gaussian Process.

Chapter 3:

- In the propensity sensitivity model that we have considered, all the bounds on the causal parameters that we have derived are valid but may not be sharp. That is, we have not been able to rule out that there exist valid bounds that are strictly narrower than ours for every data generating mechanism compatible with the assumptions made. Establishing the sharpness of our results or deriving tighter bounds would have important implications for the adoption of our methods in practice.
- In the non time-varying propensity sensitivity model, we have been able to derive bounds enforcing several constrains that the model on unobserved confounding needs to satisfy based on the observed distribution. However, the time-varying case presents more challenges and we have not been able to enforce all the constrains in this case. There is therefore the danger that our time-varying bounds are too lose in certain settings, which is an issue we plan to address in a revised version of the paper.

Chapters 4 and 5:

- In recent years, it has emerged that the optimal estimators of many causal effects in regimes of low smoothness are based on the theory of higher order influence functions (HOIF) [Robins et al., 2008, 2017a]. Our findings in this thesis also align with this recent trend as our optimal estimator of CATE level sets simply thresholds the optimal, HOIF-based estimator of the CATE derived in Kennedy et al. [2022]. We also conjecture that our HOIF-based estimator of the dose-response function is also optimal, at least under certain conditions. Unfortunately, while theoretically appealing, estimators based on HOIFs are rarely adopted in practice because of important limitations, including requiring a delicate choice of the tuning parameters and often a heavy computational cost. Mitigating these challenges to allow for more widely adoption of these methods would likely open the path for more precise inference in many domains of science.

- Specifically for the estimation of the dose-response parameter $a \mapsto \theta(a) \equiv \int \mathbb{E}(Y \mid A = a, X = x) d\mathbb{P}(x)$, $A \in \mathbb{R}$ and $X \in \mathbb{R}^d$, an important open question is to establish the minimax rate in models where $a \mapsto \theta(a)$ has its own smoothness level, which can be different from that of $(a, x) \mapsto \mathbb{E}(Y \mid A = a, X = x)$.

Bibliography

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002. A.3

Chunrong Ai, Oliver Linton, Kaiji Motegi, and Zheng Zhang. A unified framework for efficient estimation of general treatment models. *arXiv preprint arXiv:1808.04936*, 2018. 5.1.2, 5.2.4

Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005. 3.7.1, 3.7.1

Joseph G Altonji, Todd E Elder, and Christopher R Taber. Using selection on observed variables to assess bias from unobservables when evaluating swan-ganz catheterization. *American Economic Review*, 98(2):345–50, 2008. 2.4.2

Miguel A Arcones and Evarist Giné. Limit theorems for u-processes. *The Annals of Probability*, pages 1494–1542, 1993. B.2.12

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. 4.1

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. 4.1

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007. 2.3.2, 4.3.1, 4.3.1, 4.4

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014. 2.4.2

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015. 2, 19

Eli Ben-Michael, Kosuke Imai, and Zhichao Jiang. Policy learning with asymmetric utilities. *arXiv preprint arXiv:2206.10479*, 2022. 4.1

David Benkeser and Mark Van Der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016. 2.3.2

Karine Bertin. Asymptotically exact minimax estimation in sup-norm for anisotropic hölder classes. *Bernoulli*, 10(5):873–888, 2004. 5.2.4

Samir Bhatt, Neil Ferguson, Seth Flaxman, Axel Gandy, Swapnil Mishra, and James A Scott. Semi-mechanistic bayesian modeling of covid-19 with renewal processes. *arXiv preprint arXiv:2012.00394*, 2020. 6.7

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993. 2.3.1

Ottar N Bjørnstad. Epidemics. *Models and data using R: Springer International Publishing*, page 318, 2018. 6.4

Matteo Bonvini and Edward H Kennedy. Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–31, 2020. 2, 3.3.3, B.0.2

Matteo Bonvini and Edward H Kennedy. Fast convergence rates for dose-response estimation. *arXiv preprint arXiv:2207.11825*, 2022. 3.3.2, 5

Matteo Bonvini, Edward Kennedy, Valerie Ventura, and Larry Wasserman. Sensitivity analysis for marginal structural models. *arXiv preprint arXiv:2210.04681*, 2022a. 3, 5.1.4, 5.4, 6.6.2

Matteo Bonvini, Edward H Kennedy, Valerie Ventura, and Larry Wasserman. Causal inference for the effect of mobility on covid-19 deaths. *The Annals of Applied Statistics*, 16(4):2458–2480, 2022b. 3.7.2, 3.7.2, 6, 6.6.1, 6.6.2

Fred Brauer, Carlos Castillo-Chavez, and Carlos Castillo-Chavez. *Mathematical models in population biology and epidemiology*, volume 2. Springer, 2012. 6.4

Babette A Brumback, Miguel A Hernán, Sebastien JPA Haneuse, and James M Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine*, 23(5):749–767, 2004. 2.1, 3.1.1

Matias D Cattaneo. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154, 2010. 3.7.1, 3.7.1

Matias D Cattaneo, Rajita Chandak, Michael Jansson, and Xinwei Ma. Boundary adaptive local polynomial conditional density estimators. *arXiv preprint arXiv:2204.10359*, 2022. C.3

Bibhas Chakraborty and Erica E Moodie. Statistical methods for dynamic treatment regimes. *Springer-Verlag. doi*, 10:978–1, 2013. 4.1

Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, pages 1–6, 2020. 6.1

Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–1696, 2017. 4.1, 4.5, 2

Victor Chernozhukov, Ivan Fernandez-Val, and Alfred Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009. 2.3.1, 3

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney K Newey. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, 2016. 2.3.1

Victor Chernozhukov, Hiroyuki Kasahara, and Paul Schrimpf. Causal impact of masks, policies, behavior on early covid-19 pandemic in the us. *arXiv preprint arXiv:2005.14168*, 2020. 6.1, 6.6.1, 6.7

Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Omitted variable bias in machine learned causal models. *arXiv preprint arXiv:2112.13398*, 2021. 3.1.1

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020. (document), 2.4.2, 3.7.1, A.6.1

William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 417–446, 1973. 1.1

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020. 5.1.2, 5.1.3, 5.1.3, 5.2.4, 11, 5.5

Alfred F Connors, Theodore Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Jama*, 276(11):889–897, 1996. 2.4.2, A.6

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959. 1.1, 2.1, 3.1.1

Victor H de la Peña and Stephen J Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, pages 806–816, 1995. C.3

Iván Díaz and Mark J van der Laan. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics*, 9(2): 149–160, 2013a. 2.2

Iván Díaz and Mark J van der Laan. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference*, 1(2):171–192, 2013b. 5.1.2

Peng Ding and Tyler J VanderWeele. Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368, 2016. 2.1, 2.2.1

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *arXiv preprint arXiv:2102.04543*, 2021. 3.1.1

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021. 5.4, 5.4, 5.4, B.2.14, D.4.2

Sam Efromovich. Conditional density estimation in a regression setting. *The Annals of Statistics*, 35(6):2504–2535, 2007. 5.2.4

Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27, 1998. 4.7

Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*. Routledge, 2018. 7

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015. 2.3.2

Ronald Aylmer Fisher. Statistical methods for research workers. *Statistical methods for research workers.*, (6th Ed), 1936. 1.1

Carlos A Flores. Estimation of dose-response functions and optimal doses with a continuous treatment. *University of Miami, Department of Economics, November*, 2007. 5.1.2

Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018. 6.5.2

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019. 3.3.2, 4.1, 5.1.4, 5.2.1, 5.2.2, 5.2.2, D.1

Antonio F Galvao and Liang Wang. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 110(512): 1528–1542, 2015. 5.1.2

Joseph L Gastwirth, Abba M Krieger, and Paul R Rosenbaum. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):907–920, 1998. 2.1, 2.2.1

Evarist Giné, Rafał Latała, and Joel Zinn. Exponential and moment inequalities for u-statistics. In *High Dimensional Probability II*, pages 13–38. Springer, 2000. C.2, C.3, 25, C.3

Yonatan Gur, Ahmadreza Momeni, and Stefan Wager. Smoothness-adaptive contextual bandits. *Operations Research*, 2022. 4.1

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006. 2.3.2

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020. 4.1

Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pages 360–372, 2006. 2.4.2

Miguel A Hernán and James M Robins. *Causal inference*. CRC Boca Raton, FL, 2010. 3.2

Keisuke Hirano and Jack R Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77 (5):1683–1701, 2009. 4.1

M Hoffman and Oleg Lepski. Random rates in anisotropic regression (with a discussion and a rejoinder by the authors). *The Annals of Statistics*, 30(2):325–396, 2002. 5.2.4

Joel L Horowitz. *Semiparametric and nonparametric methods in econometrics*, volume 12. Springer, 2009. 2.3.2

Joel L Horowitz and Charles F Manski. Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302, 1995. 2, A.2

IHME. Modeling covid-19 scenarios for the united states. *Nature Medicine*, 2020. 6.1

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013. 4.1

Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003. 2.1

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004. 2.3.2, 2.5, 2.2, 7, A.5.2

Marshall M Joffe, Wei Peter Yang, and Harold I Feldman. Selective ignorability assumptions in causal inference. *The International Journal of Biostatistics*, 6(2), 2010. 2.1.1, 2.2

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019. 3.1.1

Kirthevasan Kandasamy and Yaoliang Yu. Additive approximations in high dimensional nonparametric regression via the salsa. In *International Conference on Machine Learning*, pages 69–78, 2016. 2.3.2

Umashankkar Kannan, Vemuru Sunil K Reddy, Amar N Mukerji, Vellore S Parithivel, Ajay K Shah, Brian F Gilchrist, and Daniel T Farkas. Laparoscopic vs open partial colectomy in elderly patients: Insights from the american college of surgeons-national surgical quality improvement program database. *World Journal of Gastroenterology*, 21(45):12843, 2015. 4.7

Jason A Kemp and Samuel RG Finlayson. Outcomes of laparoscopic and open colectomy: a national population-based comparison. *Surgical innovation*, 15(4):277–283, 2008. 4.7

Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, pages 1–12, 2018. 2.3.2, A.5.1, A.5.2, A.5.2

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020. 4.1, 4.3.2, 5.1.4, 5.2.1, 5.2.3, 5.2.3, 9, D.2, D.4.2

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022. 5.1.3

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. 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, 2017. 3.3.2, 5.1.1, 5.1.1, 5.1.2, 5.1.3, 5.1.3, 5.2.3, 10

Edward H Kennedy, Steve Harris, and Luke J Keele. Survivor-complier effects in the presence of selection on treatment, with application to a study of prompt icu admission. *Journal of the American Statistical Association*, 114(525):93–104, 2019. 2.2, 2.3.2

Edward H Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008–2030, 2020. 2.3.2, 4.3.1, A.5.1, B.2.14

Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Minimax rates for heterogeneous effect estimation. *arXiv preprint arXiv:*, 2022. 4.1, 4.1.1, 4.3.3, 4.3.3, 4.3.3, 4.4, 7, C.3, C.3.1, C.4, 26, C.4

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927. 6.4

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, 2008. A.5.1, A.5.3, B.2.12, B.2.14, B.2.14

Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman. The hulc: Confidence regions from convex hulls. *arXiv preprint arXiv:2105.14577*, 2021. 3.1.4, 3.7.1

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019. 4.1

Craig Lammert, Douglas L Nguyen, Brian D Juran, Erik Schlicht, Joseph J Larson, Elizabeth J Atkinson, and Konstantinos N Lazaridis. Questionnaire based assessment of risk factors for primary biliary cirrhosis. *Digestive and Liver Disease*, 45(7):589–594, 2013. 2.1.1

Lingling Li, Eric Tchetgen, J Robins, and A van der Vaart. Robust inference with higher order influence functions: Parts i and ii. In *Joint Statistical Meetings, Minneapolis, Minnesota*, 2005. 5.5

Xiyue Liao and Mary C Meyer. cgam: An r package for the constrained generalized additive model. *arXiv preprint arXiv:1812.07696*, 2018. 16

Danyu Y Lin, Bruce M Psaty, and Richard A Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, pages 948–963, 1998. 2.4.2

Weiwei Liu, S Janet Kuramoto, and Elizabeth A Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14 (6):570–580, 2013. 2.1

Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016. 2.3.2, 4.1

Alexander R Luedtke, Ivan Diaz, and Mark J van der Laan. The statistics of sensitivity analyses. 2015. 2.3.1

Enno Mammen and Wolfgang Polonik. Confidence regions for level sets. *Journal of Multivariate Analysis*, 122:202–214, 2013. 4.1, 4.5

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990. 1.2.1

Mary C Meyer. Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, 2(3):1013–1033, 2008. 16

Mary C Meyer. A framework for estimation and inference in generalized additive models with shape and order restrictions. *Statistical Science*, 33(4):595–614, 2018. 16

Rajarshi Mukherjee, Whitney K Newey, and James M Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017. 5.3.3

Romain Neugebauer and Mark van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007. 3.1, 5.1.2, 6.4.1

Whitney K Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, pages 233–253, 1994. 5.1.2

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913610. 6.5.1

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923. 1.1

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. 4.1

Luciano Machado Ferreira Tenório de Oliveira, Ana Raquel Mendes dos Santos, Breno Quintella Farah, Raphael Mendes Ritti-Dias, Clara Maria Silvestre Monteiro de Freitas, and Paula Rejane Beserra Diniz. Influence of parental smoking on the use of alcohol and illicit drugs among adolescents. *Einstein (São Paulo)*, 17(1), 2019. 2.1.1

Supa Pengpid and Karl Peltzer. Alcohol use and misuse among school-going adolescents in thailand: results of a national survey in 2015. *International journal of environmental research and public health*, 16(11):1898, 2019. 2.1.1

Wanli Qiao and Wolfgang Polonik. Nonparametric confidence regions for level sets: Statistical properties and geometry. *Electronic Journal of Statistics*, 13(1):985–1030, 2019. 4.1, 4.5

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427, 2012. 2.3.2

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Optimal subgroup selection. *arXiv preprint arXiv:2109.01077*, 2021. 4.1

Amy Richardson, Michael G Hudgens, Peter B Gilbert, and Jason P Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596, 2014. 2.1, 2.2

Thomas S Richardson and James M Robins. Analysis of the binary instrumental variable model. *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pages 415–444, 2010. 2.1.1

Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009. 4.1, 4.1.1, 4.3.1, 4.3.1, 4.4, 2, C.4

James M Robins. 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(9-12):1393–1512, 1986. 1.1, 1.2.3, 3.1, 3.6, 6.3, 6.4.1

James M Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989. 1.2.1, 6.3, 6.4.1

James M Robins. Marginal structural models. *Proceedings of the American Statistical Association*, pages 1–10, 1998. 3.1, 3.1

James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000. 1.2.1, 3.1, 3.1, 6.1, 6.3, 6.6.2

James M Robins. [covariance adjustment in randomized experiments and observational studies]: Comment. *Statistical Science*, 17(3):309–321, 2002. 3

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data*, pages 189–326. Springer, 2004. 4.1

James M Robins and Miguel A Hernán. Estimation of the causal effects of time-varying exposures. *Longitudinal Data Analysis*, 553:599, 2009. 3.2

James M Robins and L Wasserman. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 409–420. Morgan Kaufmann, 1997. 6.3, 6.4.1

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000. 3.1, 3.1, 5.1.2, 6.1, 6.3

James M Robins, Lingling Li, Eric Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008. 2.3.1, 5.1.4, 5.3.1, 7

James M Robins, Lingling Li, Eric Tchetgen, and Aad W van der Vaart. Quadratic semiparametric von mises calculus. *Metrika*, 69(2):227–247, 2009a. 5.1.4, 5.3.1, 5.3.1

James M Robins, Eric Tchetgen Tchetgen, Lingling Li, and Aad van der Vaart. Semiparametric minimax rates. *Electronic journal of statistics*, 3:1305, 2009b. 4.1.1, 5.3.4, 26, C.4

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Higher order estimating equations for high-dimensional models. *Annals of statistics*, 45(5):1951, 2017a. 5.1.4, 5.3.1, 5.3.1, 5.3.2, 12, 5.3.4, 5.6, 7, D.3, D.3.1, D.3.2

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017b. 4.1.1

Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987. 2.1, 4, 2.2.1

Paul R Rosenbaum. *Observational Studies*. Springer, 1995. 3.1.1, 3.3.1

Paul R Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002. 2.1

Paul R Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004. A.7

Paul R Rosenbaum. Differential effects and generic biases in observational studies. *Biometrika*, 93(3):573–586, 2006. 2.1.1, 2.2

Paul R Rosenbaum. *Design of observational studies*, volume 10. Springer, 2010. 6.6.2

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983. 2.1

Andrea Rotnitzky, Daniel Scharfstein, Ting-Li Su, and James M Robins. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, 57(1):103–113, 2001. 2.1

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974. 1.1, 2.2, 5.1.1

D. Scharfstein, R. Nabi, E. Kennedy, M. Huang, M. Bonvini, and M. Smid. Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv:2104.08300*, 2021. 3.1.1, 3.7.1

James A Scott, Axel Gandy, Swapnil Mishra, Samir Bhatt, Seth Flaxman, H Juliette T Unwin, and Jonathan Ish-Horowicz. Epidemia: An r package for semi-mechanistic bayesian modelling of infectious diseases using point processes. *arXiv preprint arXiv:2110.12461*, 2021. 6.7

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *arXiv preprint arXiv:1702.06240*, 2017. 5.1.2, 5.1.3, 5.1.3, 5.2.1

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021. 3.3.2, 4.1, 4.5, 2, 4.7

Bodhisattva Sen. A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University*, 2018. B.2.4, B.2.12

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017. 4.1

Yunfeng Shi and Xuegang Ban. Capping mobility to control covid-19: A collision-based infectious disease transmission model. *medRxiv*, 2020. 6.4.3

Rahul Singh, Liyuan Xu, and Arthur Gretton. Reproducing kernel methods for nonparametric and semiparametric treatment effects. *arXiv preprint arXiv:2010.04855*, 2020. 5.1.2

Jörg Stoye. More on confidence intervals for partially identified parameters. *Econometrica*, 77 (4):1299–1315, 2009. A.5.2

Kenta Takatsu and Ted Westling. Debiased inference for a covariate-adjusted regression function. *arXiv preprint arXiv:2210.06448*, 2022. 4.5

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006. 3.3.1

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007. 2.3.1, 6.6.1

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, NY, 2009. 4.1.1, 1, 4.4, 5.2.3, 3, 5.3.4, 4, C.4

H Juliette T Unwin, Swapnil Mishra, Valerie C Bradley, Axel Gandy, Thomas A Mellan, Helen Coupland, Jonathan Ish-Horowicz, Michaela AC Vollmer, Charles Whittaker, Sarah L Filippi, et al. State-level tracking of covid-19 in the united states. *Nature communications*, 11(1):1–9, 2020. 6.1, 6.4.1, 6.6.1, 6.7

Mark van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The international journal of biostatistics*, 13(2), 2017. 2.3.2

Mark J Van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006. 5.2.3

Mark J van der Laan and Alexander R Luedtke. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. 2014. 2.3.2, A.5.1

Mark J Van der Laan, MJ Laan, and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003. 2.3.1, 5.1.2

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007. 2.4.1, 2.4.2

Aad van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, pages 679–686, 2014. 5.3.1

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 3.1.4, A.5.1, 18

Aad W. van der Vaart. Semiparametric statistics. In *Lectures on Probability Theory and Statistics*, pages 331–457. Springer, 2002. 2.3.1

Aad W. van der Vaart and John A. Wellner. *Weak Convergence and Empirical Processes with Application to Statistics*. Springer Verlage, 1996. 2.3.3, A.5.1, A.5.3, A.5.3

Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017. 2.1, 2.2.1

Stijn Vansteelandt and Marshall Joffe. Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, 29(4):707–731, 2014. 6.6.2

Stijn Vansteelandt, Els Goetghebeur, Michael G Kenward, and Geert Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16(3): 953–979, 2006. A.5.2

J Esteban Varela, Massimo Asolati, Sergio Huerta, and Thomas Anthony. Outcomes of laparoscopic and open colectomy at academic centers. *The American Journal of Surgery*, 196(3): 403–406, 2008. 4.7

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. 4.1

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019. 5.2.2, 5.2.2, 2, D.1, D.1

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006. 5.1.2, 7

Ted Westling and Marco Carone. A unified study of nonparametric inference for monotone functions. *Annals of statistics*, 48(2):1001, 2020. 5.1.2

Ted Westling, Peter Gilbert, and Marco Carone. Causal isotonic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):719–747, 2020. 5.1.2

Rebecca M Willett and Robert D Nowak. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007. 4.1, 5

Jini Wu, Bo Li, Shiliang Tu, Boan Zheng, and Bingchen Chen. Comparison of laparoscopic and open colectomy for splenic flexure colon cancer: a systematic review and meta-analysis. *International Journal of Colorectal Disease*, pages 1–11, 2022. 4.7

Xiao-Jian Wu, Xiao-Sheng He, Xu-Yu Zhou, Jia Ke, and Ping Lan. The role of laparoscopic surgery for ulcerative colitis: systematic review with meta-analysis. *International journal of colorectal disease*, 25(8):949–957, 2010. 4.7

Chenfeng Xiong, Songhua Hu, Mofeng Yang, Weiyu Luo, and Lei Zhang. Mobile device data reveal the dynamics in a positive relationship between human mobility and covid-19 infections. *Proceedings of the National Academy of Sciences*, 117(44):27087–27089, 2020. 6.1

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect in the presence of unobserved confounders. *arXiv preprint arXiv:1808.09521*, 2018. 2.1, 2.2.1, 3.1.1, 6.6.2

Yun Yang and Surya T Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015. 2.3.2

Bo Zhang and Eric J Tchetgen Tchetgen. A semiparametric approach to model-based sensitivity analysis in observational studies. *arXiv preprint arXiv:1910.14130*, 2019. 2.1

Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019. 2.1, 2.2.1, 3.1.1, 6.6.2, A.4, A.4

Wenjing Zheng and Mark J Van Der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation.(uc berkeley division of biostatistics working paper 273). berkeley. *CA: University of California, Berkeley*, 2010. 2.3.1

Xiang Zhou and Geoffrey T Wodtke. Residual balancing weights for marginal structural models: with application to analyses of time-varying treatments and causal mediation. *arXiv preprint arXiv:1807.10869*, 2018. 6.5.2, 6.5.2

Appendices

Appendix A

Appendix for Chapter 2

A.1 Proof of Lemma 1

Notice that, because $A \perp\!\!\!\perp Y^a \mid \mathbf{X}, S = 1$, we have

$$\mathbb{E}\{(Y^1 - Y^0)S\} = \mathbb{E}[S\{\mathbb{E}(Y \mid A = 1, \mathbf{X}, S = 1) - \mathbb{E}(Y \mid A = 0, \mathbf{X}, S = 1)\}]$$

and, by the consistency assumption, it holds that

$$\begin{aligned}\mathbb{E}\left\{(Y^1 - Y^0)(1 - S)\right\} &= \mathbb{E}\left[(1 - S)\left\{(Y - Y^0)A + (Y^1 - Y)(1 - A)\right\}\right] \\ &= \mathbb{E}((1 - S)[\{Y - \lambda_{1-A}(\mathbf{X})\}(2A - 1)])\end{aligned}$$

Therefore, we conclude that

$$\psi = \mathbb{E}((1 - S)[\{Y - \lambda_{1-A}(\mathbf{X})\}(2A - 1)] + S\{\mathbb{E}(Y \mid A = 1, \mathbf{X}, S = 1) - \mathbb{E}(Y \mid A = 0, \mathbf{X}, S = 1)\})$$

as desired.

A.2 Proof of Theorem 1

Notice that (A1) is equivalent to $S \perp\!\!\!\perp A \mid \mathbf{X}$ and $S \perp\!\!\!\perp Y \mid \mathbf{X}, A$. Then, under (A1), we have that $\mathbb{E}\left(Y \mid \mathbf{X}, A = a, S\right) = \mu_a(\mathbf{X})$ and $\mathbb{P}\left(A = a \mid \mathbf{X}, S\right) = \pi(a \mid \mathbf{X})$. This means that the result in Lemma 1 simplifies to

$$\psi(S, \lambda_0, \lambda_1) = \mathbb{E}\left(\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + (1 - S)\left[\pi(0 \mid \mathbf{X})\left\{\lambda_1(\mathbf{X}) - \mu_1(\mathbf{X})\right\} - \pi(1 \mid \mathbf{X})\left\{\lambda_0(\mathbf{X}) - \mu_0(\mathbf{X})\right\}\right]\right)$$

The observed distribution $\mathbb{P}$ and the knowledge of S places no restrictions on $\lambda_0(\mathbf{X})$ and $\lambda_1(\mathbf{X})$. Recalling that δ is chosen such that

$$L_a \equiv \delta\{y_{\text{min}} - \mu_a(\mathbf{X})\} \leq \lambda_a(\mathbf{X}) - \mu_a(\mathbf{X}) \leq \delta\{y_{\text{max}} - \mu_a(\mathbf{X})\} \equiv U_a \text{ with prob. } 1$$

for $a \in \{0, 1\}$, we have that

$$\begin{gathered}\mathbb{E}\left\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + (1-S)g(\boldsymbol{\eta})\right\} - \epsilon\delta(y_{\max} - y_{\min}) \\ \leq \psi(S, \lambda_0, \lambda_1) \leq \\ \mathbb{E}\left\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + (1-S)g(\boldsymbol{\eta})\right\}\end{gathered}$$

where $g(\boldsymbol{\eta}) = \pi(0 \mid \mathbf{X})U_1 - \pi(1 \mid \mathbf{X})L_0$. These bounds are sharp for any given S.

Next, notice that $g(\boldsymbol{\eta}) : \mathcal{X}^p \to \mathbb{R}$ and $\mathbb{P}(S = 0) = \epsilon$. Thus, by Proposition 4 in Horowitz and Manski [1995], it holds that $\psi \in [\psi_l(\epsilon), \psi_u(\epsilon)]$ where

$$\begin{aligned}\psi_l(\epsilon) &= \mathbb{E}\left[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\left\{g(\boldsymbol{\eta}) \leq q_\epsilon\right\} g(\boldsymbol{\eta})\right] - \epsilon\delta(y_{\max} - y_{\min}) \\ \psi_u(\epsilon) &= \mathbb{E}\left[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\left\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\right\} g(\boldsymbol{\eta})\right]\end{aligned}$$

and these bounds are sharp.

A.3 Bounds in XA-mixture model

The restriction in (A1) can easily be weakened to

$$S \perp\!\!\!\perp Y \mid \mathbf{X}, A \tag{A2}$$

Under (A2), it still holds that $\mathbb{E}(Y \mid \mathbf{X}, A = a, S) = \mu_a(X)$, but $\pi(a \mid \mathbf{X}, S = 1)$ does not equal $\pi(a \mid \mathbf{X}, S = 0)$ necessarily. Therefore, the result in Lemma 1 simplifies only to

$$\psi(S, \lambda_0, \lambda_1) = \mathbb{E}(\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + (1-S)[(1-A)\{\lambda_1(\mathbf{X}) - \mu_1(\mathbf{X})\} - A\{\mu_0(\mathbf{X}) - \lambda_0(\mathbf{X})\}])$$

where $\lambda_a(\mathbf{X}) = \mathbb{E}(Y^a \mid A = 1 - a, \mathbf{X}, S = 0)$. Following the same line of reasoning as in the proof of Theorem 1, under consistency and positivity, sharp bounds on ψ are:

$$\begin{aligned}\psi_l(\epsilon) &= \mathbb{E}[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\{g(A, \boldsymbol{\eta}) \leq q_\epsilon\} g(A, \boldsymbol{\eta})] - \epsilon\delta(y_{\min} - y_{\max}) \\ \psi_u(\epsilon) &= \mathbb{E}[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \mathbb{1}\{g(A, \boldsymbol{\eta}) > q_{1-\epsilon}\} g(A, \boldsymbol{\eta})]\end{aligned}$$

where $g(A, \boldsymbol{\eta}) = (1-A)U_1 - AL_0$, q_τ is the τ-quantile of $g(A, \boldsymbol{\eta})$ and δ is chosen such that

$$L_a \equiv \delta\{y_{\min} - \mu_a(\mathbf{X})\} \leq \lambda_a(\mathbf{X}) - \mu_a(\mathbf{X}) \leq \delta\{y_{\max} - \mu_a(\mathbf{X})\} \equiv U_a \text{ with prob. 1.}$$

with $y_{\min}$ and $y_{\max}$ finite. The following lemma shows that the bounds assuming $S \perp\!\!\!\perp Y \mid \mathbf{X}, A$ are at least as wide as those assuming $S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$.

Lemma 14. *Let X, A be two random variables and let $\pi(X) = \mathbb{E}(A \mid X)$. Consider the functions:*

$$g_1(a, x) = af(x) \quad \textit{and} \quad g_2(x) = \pi(x)f(x)$$

for a measurable function f. Then, it holds that

$$\begin{aligned}\mathbb{E}\left[g_1(A,X)\mathbb{1}\left\{g_1(A,X) \leq q_{1\tau}\right\}\right] &\leq \mathbb{E}\left[g_2(X)\mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\}\right] \\ \mathbb{E}\left[g_1(A,X)\mathbb{1}\left\{g_1(A,X) > q_{1\tau}\right\}\right] &\geq \mathbb{E}\left[g_2(X)\mathbb{1}\left\{g_2(X) > q_{2\tau}\right\}\right] \end{aligned} \tag{A.1}$$

where $q_{i\tau}$ is the τ-quantile of $g_i(\cdot)$.

Proof. This lemma is essentially a restatement of the subadditivity property of expected shortfall [Acerbi and Tasche, 2002]. It is sufficient to note that

$$\mathbb{E}\left[g_2(X)\mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\}\right] = \mathbb{E}\left[g_1(A,X)\mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\}\right]$$

and that

$$\begin{aligned}\mathbb{E}\left(g_1(A,X)\left[\mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\} - \mathbb{1}\left\{g_1(A,X) \leq q_{1\tau}\right\}\right]\right) &\geq q_{1\tau}\mathbb{E}\left[\mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\} - \mathbb{1}\left\{g_1(A,X) \leq q_{1\tau}\right\}\right] \\ &= q_{1\tau}(\tau - \tau) \\ &= 0\end{aligned}$$

where the inequality follows because

$$\begin{cases} \mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\} - \mathbb{1}\left\{g_1(A,X) \leq q_{1\tau}\right\} \leq 0 & \text{if } g_1(A,X) \leq q_{1\tau} \\ \mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\} - \mathbb{1}\left\{g_1(A,X) \leq q_{1\tau}\right\} \geq 0 & \text{if } g_1(A,X) > q_{1\tau} \end{cases}$$

Inequality (A.1) follows by rearranging:

$$\begin{aligned}\mathbb{E}\left[g_1(A,X)\mathbb{1}\left\{g_1(A,X) > q_{1\tau}\right\}\right] &= \mathbb{E}\left(g_1(A,X)\left[1 - \mathbb{1}\left\{g_1(A,X) \leq q_{1\tau}\right\}\right]\right) \\ \mathbb{E}\left[g_2(X)\mathbb{1}\left\{g_2(X) > q_{2\tau}\right\}\right] &= \mathbb{E}\left(g_1(A,X)\left[1 - \mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\}\right]\right)\end{aligned}$$

so that

$$\begin{aligned}&\mathbb{E}\left(g_1(A,X)\left[\mathbb{1}\left\{g_2(X) > q_{2\tau}\right\} - \mathbb{1}\left\{g_1(A,X) > q_{1\tau}\right\}\right]\right) \\ &\quad = \mathbb{E}\left(g_1(A,X)\left[\mathbb{1}\left\{g_1(A,X) \leq q_{1\tau}\right\} - \mathbb{1}\left\{g_2(X) \leq q_{2\tau}\right\}\right]\right) \\ &\quad \leq 0\end{aligned}$$

as desired. □

From Lemma 14 we conclude that the lower bound (upper bound) under $S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$ is greater (smaller) than that under $S \perp\!\!\!\perp Y \mid A, \mathbf{X}$.

A.4 Extensions

In this section, we discuss one possible extension to our model, though we note that others are possible. The impact of unmeasured confounding U can be controlled by linking the true, unidentifiable propensity score $\mathbb{P}(A = a \mid \mathbf{X}, U, S = 0)$ to the estimable "pseudo-propensity

score" $\pi(a \mid \mathbf{X})$ via a sensitivity model of choice. For example, as proposed in Zhao et al. [2019], an extension to Rosenbaum's framework to non-matched data can be formulated by noting that, under consistency and positivity,

$$\mathbb{E}\left(Y^a\right) = \mathbb{E}\left\{\frac{Y\mathbb{1}\left(A=a\right)S}{\mathbb{P}(A=a \mid \mathbf{X}, S=1, Y^a)}\right\} + \mathbb{E}\left\{\frac{Y\mathbb{1}\left(A=a\right)(1-S)}{\mathbb{P}(A=a \mid \mathbf{X}, S=0, Y^a)}\right\} \tag{A.2}$$

and thus we can simply take the unobserved confounder U to be one of the potential outcomes. Next, notice that $\mathbb{P}(A=a \mid \mathbf{X}, S=1, Y^a) = \pi(a \mid \mathbf{X})$ under Assumption (A1) ($S \perp\!\!\!\perp (Y, A) \mid \mathbf{X}$), so that (A.2) simplifies to

$$\mathbb{E}\left(Y^a\right) = \mathbb{E}\left\{\frac{Y\mathbb{1}\left(A=a\right)S}{\pi(a \mid \mathbf{X})}\right\} + \mathbb{E}\left\{\frac{Y\mathbb{1}\left(A=a\right)(1-S)}{\mathbb{P}(A=a \mid \mathbf{X}, S=0, Y^a)}\right\}$$

Let $\pi_a(\mathbf{x}, y) = \mathbb{P}(A=a \mid \mathbf{X}=\mathbf{x}, S=0, Y^a=y)$. Noting that $\mathbb{P}(A=a \mid \mathbf{X}, S=0) = \pi(a \mid \mathbf{X})$ under Assumption (A1), the impact of unmeasured confounding can be governed by requiring $\pi_a(\mathbf{x}, y)$ to be an element of the following sensitivity model

$$\mathcal{E}\left(\Lambda\right) = \left\{\Lambda^{-1} \leq \text{OR}\left\{\pi_a(\mathbf{x}, y), \pi(a \mid \mathbf{x})\right\} \leq \Lambda, \text{ for all } \mathbf{x} \in \mathcal{X},\ y \in [0,1],\ a \in \{0,1\}\right\} \tag{A.3}$$

where $\Lambda \geq 1$ and $\Lambda = 1$ corresponds to the unconfounded case. Model (A.3) can be conveniently reformulated on the logit scale. Let

$$\begin{aligned} &g(a \mid \mathbf{x}) = \text{logit}\{\pi(a \mid \mathbf{x})\}, \quad g_a(\mathbf{x}, y) = \text{logit}\{\pi_a(\mathbf{x}, y)\} \\ &h(\mathbf{x}, y) = g(a \mid \mathbf{x}) - g_a(\mathbf{x}, y), \quad \pi^{(h)}(\mathbf{x}, y) = \left[1 + \exp\left\{h(\mathbf{x}, y) - g(a \mid \mathbf{x})\right\}\right]^{-1} \end{aligned}$$

and write

$$\mathcal{E}\left(\Lambda\right) = \left\{\pi^{(h)}(\mathbf{x}, y) :\ h \in \mathcal{H}(\Lambda)\right\}, \text{ where } \mathcal{H}(\Lambda) = \{h : \mathcal{X} \times [0,1] \to \mathbb{R} \text{ and } \|h\|_\infty \leq \log \Lambda\} \tag{A.4}$$

From (A.4), we rewrite $\mathbb{E}\left(Y^a\right)$ as

$$\mathbb{E}\left(Y^a\right) = \mathbb{E}\left(\frac{SY\mathbb{1}\left(A=a\right)}{\pi(a \mid \mathbf{X})} + (1-S)Y\mathbb{1}\left(A=a\right)\left[1 + \exp\left\{h(\mathbf{X}, Y)\right\}\exp\left\{-g(a \mid \mathbf{X})\right\}\right]\right) \tag{A.5}$$

where $\exp\left\{h(\mathbf{X}, Y)\right\} \in [\Lambda^{-1}, \Lambda]$. Bounds on ψ can then be computed following the same line of reasoning as in Theorem 1, where $\exp\left\{h(\mathbf{X}, Y)\right\}$ takes the role of $\lambda_a(\mathbf{X})$. Convergence statements for estimators of (A.5) can be derived using standard arguments for convergence of inverse propensity score-weighted estimators together with the arguments made in proving Theorem 2. However, we expect the conditions for $\sqrt{n}$-consistency and asymptotic normality to be stronger than those assumed in Theorem 2. Moreover, note that, if $\mathbb{P}(S=1) = 0$, as in Zhao et al. [2019], expression (A.5) can be bounded and estimated via a stabilized IPW (SIPW) and a suitable linear program. In our model, because $\mathbb{P}(S=1) \geq 0$, optimization of a SIPW is

harder due to the integer nature of S and beyond the scope of this paper.

A.5 Technical Proofs

A.5.1 Proof of Theorem 2

Before proceeding with the proof of Theorem 2, we report a lemma used below. It can be found in Kennedy et al. [2020] (Lemma 1) or in the proof of Lemma 2 in van der Laan and Luedtke [2014].

Lemma 15. *Let $\widehat{f}$ and f take any real values. Then*

$$|\mathbb{1}(\widehat{f} > 0) - \mathbb{1}(f > 0)| \leq \mathbb{1}(|f| \leq |\widehat{f} - f|)$$

Proof. This follows since

$$|\mathbb{1}(\widehat{f} > 0) - \mathbb{1}(f > 0)| = \mathbb{1}(\widehat{f}, f \text{ have opposite sign})$$

and if $\widehat{f}$ and f have opposite sign then

$$|\widehat{f}| + |f| = |\widehat{f} - f|$$

which implies that $|f| \leq |\widehat{f} - f|$. Therefore, whenever $|\mathbb{1}(\widehat{f} > 0) - \mathbb{1}(f > 0)| = 1$, it must also be the case that $\mathbb{1}(|f| \leq |\widehat{f} - f|) = 1$, which yields the result. □

The proof of Theorem 2 is similar to that of Theorem 3 in Kennedy [2018], with the main difference being that the influence function of the estimator proposed is not a smooth function of the sensitivity parameter ϵ. Fortunately, we can exploit the fact that the bounds are monotone in ϵ to establish convergence to a Gaussian process. We prove the result for the upper bound, as the case for the lower bound follows analogously. We also proceed by assuming Y is bounded in $[0, 1]$.

Let $\|f\|_{\mathcal{E}} = \sup_{\epsilon \in \mathcal{E}} |f(\epsilon)|$ denote the supremum norm over $\mathcal{E} \subseteq [0, 1]$, a known interval. Let $\lambda_{1-\epsilon}$ be shorthand notation for $\mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\}$. Similarly, let τ and ν be shorthand notations for the uncentered influence functions of $\mathbb{E}\{g(\boldsymbol{\eta})\}$ and $\mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\}$ respectively, so that

$$\tau = \frac{(1 - 2A)\{Y - \mu_A(\mathbf{X})\}}{\pi(A \mid \mathbf{X})/\pi(1 - A \mid \mathbf{X})} + A\mu_0(\mathbf{X}) + (1 - A)(1 - \mu_1(\mathbf{X}))$$
$$\nu = \frac{(2A - 1)\{Y - \mu_A(\mathbf{X})\}}{\pi(A \mid \mathbf{X})} + \mu_1(\mathbf{X}) - \mu_0(\mathbf{X})$$

Define the following processes:

$$\begin{aligned}
\widehat{\Psi}_n(\epsilon) &= \sqrt{n}\{\widehat{\psi}_u(\epsilon) - \psi_u(\epsilon)\}/\widehat{\sigma}_u(\epsilon) \\
\widetilde{\Psi}_n(\epsilon) &= \sqrt{n}\{\widehat{\psi}_u(\epsilon) - \psi_u(\epsilon)\}/\sigma_u(\epsilon) \\
\Psi_n(\epsilon) &= \mathbb{G}_n([\varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon}) - \lambda_{1-\epsilon}q_{1-\epsilon} - \{\psi_u(\epsilon) - \epsilon q_{1-\epsilon}\}]/\sigma_u(\epsilon)) \\
&= \mathbb{G}_n([\overline{\varphi}_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon}) - \{\psi_u(\epsilon) - \epsilon q_{1-\epsilon}\}]/\sigma_u(\epsilon)) \\
&= \mathbb{G}_n\{\phi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\}
\end{aligned}$$

where $\overline{\varphi}_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon}) = \nu + \lambda_{1-\epsilon}(\tau - q_{1-\epsilon})$ and $\mathbb{G}_n(\cdot) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ denotes the empirical process on the full sample.

We also let $\mathbb{G}(\cdot)$ denote the mean-zero Gaussian process with covariance

$$\mathbb{E}\left\{\phi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon_1})\phi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon_2})\right\}.$$

We will show that $\Psi_n(\cdot) \rightsquigarrow \mathbb{G}(\cdot)$ in $\ell^\infty(\mathcal{E})$ and that $\|\widehat{\Psi}_n - \Psi_n\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$.

To show that $\Psi_n(\cdot) \rightsquigarrow \mathbb{G}(\cdot)$ in $\ell^\infty(\mathcal{E})$, notice that $\overline{\varphi}_u(\cdot;\boldsymbol{\eta}, q_{1-\epsilon}) : \mathcal{E} \to [-M, M]$, for some $M < \infty$, consists of a sum of a bounded, constant function plus a product of two monotone functions. Specifically, consider $s(\cdot;\boldsymbol{\eta}, \epsilon) : \mathcal{E} \mapsto [-S, S]$, defined as $s(\cdot;\boldsymbol{\eta}, \epsilon) = \nu$, $f(\cdot;\boldsymbol{\eta}, \epsilon) : \mathcal{E} \mapsto \{0, 1\}$, defined as $f(\cdot;\boldsymbol{\eta}, \epsilon) = \lambda_{1-\epsilon}$, and $h(\cdot;\boldsymbol{\eta}, \epsilon) : \mathcal{E} \mapsto [-H, H]$, defined as $h(\cdot;\boldsymbol{\eta}, \epsilon) = \tau - q_{1-\epsilon}$. Then, $\overline{\varphi}_u(\cdot;\boldsymbol{\eta}, q_{1-\epsilon}) = s(\cdot;\boldsymbol{\eta}, \epsilon) + f(\cdot;\boldsymbol{\eta}, \epsilon)h(\cdot;\boldsymbol{\eta}, \epsilon)$. The fact that $s(\cdot;\boldsymbol{\eta}, \epsilon)$ and $h(\cdot;\boldsymbol{\eta}, \epsilon)$ are uniformly bounded follows by the assumptions that $\mathbb{P}\{t \leq \pi(a \mid \mathbf{X}) \leq 1 - t\} = 1$, for some $t > 0$ and $a \in \{0, 1\}$, and that the outcome Y is bounded.

Then we define the class $\mathcal{F}_\eta$ where $\overline{\varphi}_u(\cdot;\boldsymbol{\eta}, q_{1-\epsilon})$ takes value in

$$\mathcal{F}_\eta = \{\nu + \lambda_{1-\epsilon}(\tau - q_{1-\epsilon}) : \epsilon \in \mathcal{E}\}.$$

$\mathcal{F}_\eta$ is contained in the sum of $\mathcal{F}_{\eta,0}$ and the pairwise product $\mathcal{F}_{\eta,1} \cdot \mathcal{F}_{\eta,2}$, where $\mathcal{F}_{\eta,0} = \{\nu : \epsilon \in \mathcal{E}\}$ (constant function class), $\mathcal{F}_{\eta,1} = \{\lambda_{1-\epsilon} : \epsilon \in \mathcal{E}\}$ and $\mathcal{F}_{\eta,2} = \{\tau - q_{1-\epsilon} : \epsilon \in \mathcal{E}\}$.

By, for example, Theorem 2.7.5 in van der Vaart and Wellner [1996], the class of bounded monotone functions possesses a finite bracketing integral, and in particular, for $w \in \{0, 1, 2\}$:

$$\log N_{[]}\left(\delta, \mathcal{F}_{\eta,w}, L_2(\mathbb{P})\right) \lesssim \frac{1}{\delta}$$

Furthermore, because $\mathcal{F}_{\eta,0}$, $\mathcal{F}_{\eta,1}$ and $\mathcal{F}_{\eta,2}$ are uniformly bounded:

$$\log N_{[]}\left(\delta, \mathcal{F}_\eta, L_2(\mathbb{P})\right) \lesssim 3 \log N_{[]}\left(\frac{\delta}{2}, \mathcal{F}_{\eta,1}, L_2(\mathbb{P})\right) \lesssim \frac{1}{\delta}$$

by, for instance, Lemma 9.24 in Kosorok [2008]. Thus, by for example Theorem 19.5 in Van der Vaart [2000], $\mathcal{F}_\eta$ is Donsker.

Next, we prove the statement that $\|\widehat{\Psi}_n - \Psi_n\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$. First, we notice that

$$\begin{aligned}\|\widehat{\Psi}_n - \Psi_n\|_{\mathcal{E}} &= \|(\widetilde{\Psi}_n - \Psi_n)\sigma_u/\widehat{\sigma}_u + \Psi_n\left(\sigma_u - \widehat{\sigma}_u\right)/\widehat{\sigma}_u\|_{\mathcal{E}} \\ &\leq \|\widetilde{\Psi}_n - \Psi_n\|_{\mathcal{E}}\|\sigma_u/\widehat{\sigma}_u\|_{\mathcal{E}} \ + \ \|\sigma_u/\widehat{\sigma}_u - 1\|_{\mathcal{E}}\|\Psi_n\|_{\mathcal{E}} \\ &\lesssim \|\widetilde{\Psi}_n - \Psi_n\|_{\mathcal{E}} \ + \ o_{\mathbb{P}}(1)\end{aligned}$$

where the last inequality follows because $\|\widehat{\sigma}_u/\sigma_u - 1\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$ by assumption and $\|\Psi_n\|_{\mathcal{E}} = O_{\mathbb{P}}(1)$ by, for example, Theorem 2.14.2 in van der Vaart and Wellner [1996] since $\mathcal{F}_\eta$ possesses a finite bracketing integral.

Let $N = n/B$ be the number of samples in any group $k = 1, \ldots, B$, and denote the empirical process over group k units by $\mathbb{G}_n^k = \sqrt{N}(\mathbb{P}_n^k - \mathbb{P})$. Then, we have

$$\begin{aligned}\widetilde{\Psi}_n(\epsilon) - \Psi_n(\epsilon) &= \frac{\sqrt{n}}{\sigma_u(\epsilon)}\{\widehat{\psi}_u(\epsilon) - \psi_u(\epsilon)\} - \mathbb{G}_n\{\widetilde{\varphi}_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\} \\ &= \frac{\sqrt{n}}{B\sigma_u(\epsilon)}\sum_{k=1}^{B}\left[\mathbb{P}_n^k\{\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon})\} - \psi_u(\epsilon) - (\mathbb{P}_n - \mathbb{P})\{\overline{\varphi}(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\}\right] \\ &= \frac{\sqrt{n}}{B\sigma_u(\epsilon)}\sum_{k=1}^{B}\left[\mathbb{P}_n^k\{\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - \varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\} + (\mathbb{P}_n - \mathbb{P})\left(\lambda_{1-\epsilon}q_{1-\epsilon}\right)\right] \\ &= \frac{\sqrt{n}}{B\sigma_u(\epsilon)}\sum_{k=1}^{B}\Bigg[\frac{1}{\sqrt{N}}\mathbb{G}_n^k\{\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - \varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\} \\ &\qquad + \mathbb{P}\{\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - \varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\} + (\mathbb{P}_n - \mathbb{P})(\lambda_{1-\epsilon}q_{1-\epsilon})\Bigg]\end{aligned}$$

where we used the facts that

$$\psi_u(\epsilon) = \mathbb{P}\left\{\varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\right\} \quad \text{and} \quad \sum_{k=1}^{B}\mathbb{P}_n^k\left\{\varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\right\} = \sum_{k=1}^{B}\mathbb{P}_n\left\{\varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\right\}$$

The term $\mathbb{P}\left\{\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - \varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\right\}$ can be decomposed in

$$\begin{aligned}\mathbb{P}\{\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - \varphi_u(\mathbf{O};\boldsymbol{\eta}, q_{1-\epsilon})\} = \mathbb{P}\{&\widehat{\nu}_{-k} - \nu + \widehat{\lambda}_{-k,1-\epsilon}\left(\widehat{\tau}_{-k} - \tau\right) + (\tau - q_{1-\epsilon})(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon}) \\ &+ q_{1-\epsilon}(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon})\}\end{aligned}$$

Notice that $\epsilon + o_{\mathbb{P}}(n^{-1/2}) = \mathbb{P}_n^k(\widehat{\lambda}_{-k,1-\epsilon}) = \mathbb{P}\left(\lambda_{1-\epsilon}\right)$, so that

$$\begin{aligned} o_{\mathbb{P}}(n^{-1/2}) &= \sum_{k=1}^{B} \mathbb{P}_n^k(\widehat{\lambda}_{-k,1-\epsilon}) - \mathbb{P}(\lambda_{1-\epsilon}) \\ &= \sum_{k=1}^{B} \left(\mathbb{P}_n^k - \mathbb{P}\right)(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon}) + \mathbb{P}(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon}) + (\mathbb{P}_n - \mathbb{P})(\lambda_{1-\epsilon}) \end{aligned}$$

where we used again the fact that $\sum_{k=1}^{B} \mathbb{P}_n^k(\lambda_{1-\epsilon}) = \sum_{k=1}^{B} \mathbb{P}_n(\lambda_{1-\epsilon})$. Thus, we have that

$$\sum_{k=1}^{B} \mathbb{P}\{q_{1-\epsilon}(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon})\} = -\sum_{k=1}^{B} (\mathbb{P}_n^k - \mathbb{P})\{q_{1-\epsilon}(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon})\} - (\mathbb{P}_n - \mathbb{P})(q_{1-\epsilon}\lambda_{1-\epsilon}) + o_{\mathbb{P}}(n^{-1/2})$$

Therefore, we rewrite $\widetilde{\Psi}_n(\epsilon) - \Psi_n(\epsilon)$ as

$$\begin{aligned} \widetilde{\Psi}_n(\epsilon) - \Psi_n(\epsilon) = \frac{\sqrt{n}}{B\sigma_u(\epsilon)} \sum_{k=1}^{B} \Bigg(& \frac{1}{\sqrt{N}} \mathbb{G}_n^k\{\varphi_u(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - q_{1-\epsilon}\widehat{\lambda}_{-k,1-\epsilon} - \overline{\varphi}_u(\mathbf{O}; \boldsymbol{\eta}, q_{1-\epsilon})\} \\ & + \mathbb{P}\{\widehat{\nu}_{-k} - \nu + \widehat{\lambda}_{-k,1-\epsilon}(\widehat{\tau}_{-k} - \tau) + (\tau - q_{1-\epsilon})(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon})\} \Bigg) \\ \equiv B_{n,1}(\epsilon) + B_{n,2}(\epsilon) + o_{\mathbb{P}}(1) \end{aligned}$$

Next, we show that $\|B_{n,1}\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$ and $\|B_{n,2}\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$, which completes the proof.

For $B_{n,1}(\epsilon)$, notice that, because B is fixed regardless of n, we have that

$$\begin{aligned} \|B_{n,1}\|_{\mathcal{E}} &= \sup_{\epsilon \in \mathcal{E}} \left| \frac{1}{\sqrt{B}\sigma_u(\epsilon)} \sum_{k=1}^{B} \mathbb{G}_n^k\{\varphi_u(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - q_{1-\epsilon}\widehat{\lambda}_{-k,1-\epsilon} - \overline{\varphi}_u(\mathbf{O}; \boldsymbol{\eta}, q_{1-\epsilon})\} \right| \\ &\lesssim \max_k \sup_{f \in \mathcal{F}_n^k} |\mathbb{G}_n(f)| \end{aligned}$$

where we define the class $\mathcal{F}_n^k = \mathcal{F}_{\widehat{\boldsymbol{\eta}}_{-k}} - \mathcal{F}_{\boldsymbol{\eta}}$, where $\mathcal{F}_{\widehat{\boldsymbol{\eta}}_{-k}} = \{\widehat{\nu}_{-k} + \widehat{\lambda}_{-k,1-\epsilon}(\widehat{\tau}_{-k} - q_{1-\epsilon}) : \epsilon \in \mathcal{E}\}$ and $\mathcal{F}_\eta = \{\overline{\varphi}_u(\cdot; \boldsymbol{\eta}, \epsilon) : \epsilon \in \mathcal{E}\}$ as above. Viewing $\widehat{\boldsymbol{\eta}}_{-k}$ as fixed given the training data $D_0^k = \{\mathbf{O}_i : K_i \neq k\}$, by Theorem 2.14.2 in van der Vaart and Wellner [1996], we have that

$$\mathbb{E}\left\{ \sup_{f \in \mathcal{F}_n^k} |\mathbb{G}_n(f)| \mid D_0^k \right\} \lesssim \left\| F_n^k \right\| \int_0^1 \sqrt{1 + \log N_{[]}\left(\delta \left\| F_n^k \right\|, \mathcal{F}_n^k, L_2(\mathbb{P})\right)} \, d\delta$$

where F_n^k is an envelop of the class $\mathcal{F}_n^k$. Given the training data, the function class where $\widehat{\lambda}_{-k,1-\epsilon}$ lives can be expressed as $\{\mathbb{1}(u > q), q \in \mathcal{Q}\}$, where $\mathcal{Q}$ is the set of all quantile functions, which in this case is a subset of the class of all bounded, monotone functions because $g(\boldsymbol{\eta}_{-k})$ is bounded for any k. Therefore, by the same line of argument as above, the class $\mathcal{F}_n^k$ is contained in unions and products of classes of uniformly bounded, monotone functions. As

such, it satisfies

$$\log N_{[]}\left(\delta \left\|F_n^k\right\|, \mathcal{F}_n^k, L_2(\mathbb{P})\right) \lesssim \frac{1}{\delta \left\|F_n^k\right\|}$$

If we take

$$F_n^k(\mathbf{o}) = \sup_{\epsilon\in\mathcal{E}} |\varphi_u(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-k}, \widehat{q}_{-k,1-\epsilon}) - q_{1-\epsilon}\widehat{\lambda}_{-k,1-\epsilon} - \overline{\varphi}_u(\mathbf{o}; \boldsymbol{\eta}, q_{1-\epsilon})|$$

then $\|F_n^k\| = o_{\mathbb{P}}(1)$ by assumption. The bracketing integral is finite for any fixed $\boldsymbol{\eta}$, but here $\mathcal{F}_n^k$ depends on n through $\widehat{\boldsymbol{\eta}}_{-k}$, hence concluding that the LHS is $o_{\mathbb{P}}(1)$ requires further analysis. Letting $C_n^k = \|F_n^k\|$, we have that

$$\begin{aligned}\left\|F_n^k\right\| \int_0^1 \sqrt{1 + \log N_{[]}\left(\delta \left\|F_n^k\right\|, \mathcal{F}_n^k, L_2(\mathbb{P})\right) d\delta} &\lesssim C_n^k \int_0^1 \sqrt{1 + \frac{1}{\delta C_n^k}} d\delta \\ &= \sqrt{C_n^k(C_n^k+1)} + \frac{1}{2}\log\left\{1 + 2C_n^k\left(1 + \sqrt{1 + \frac{1}{C_n^k}}\right)\right\}\end{aligned}$$

which goes to zero as $C_n^k \to 0$. Hence, we conclude that $\sup_{f\in\mathcal{F}_n^k} |\mathbb{G}_n(f)| = o_{\mathbb{P}}(1)$ for each k. Because B is finite, this implies that $\|B_{n,1}\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$ as desired.

For $B_{n,2}(\epsilon)$, first notice that

$$\begin{aligned}\mathbb{P}(\widehat{\nu}_{-k} - \nu) &\lesssim \mathbb{P}\left[\{\pi(1 \mid \mathbf{X}) - \widehat{\pi}(1 \mid \mathbf{X})\}\left\{\frac{\mu_1(\mathbf{X}) - \widehat{\mu}_1(\mathbf{X})}{\widehat{\pi}(1 \mid \mathbf{X})} + \frac{\mu_0(\mathbf{X}) - \widehat{\mu}_0(\mathbf{X})}{1 - \widehat{\pi}(1 \mid \mathbf{X})}\right\}\right] \\ &\lesssim \|\widehat{\pi}(1 \mid \mathbf{X}) - \pi(1 \mid \mathbf{X})\| \max_a \|\widehat{\mu}_a(\mathbf{X}) - \mu_a(\mathbf{X})\|\end{aligned}$$

by an application of the Cauchy-Schwartz inequality.

Next, similar calculations yield

$$\begin{aligned}\sup_{\epsilon\in\mathcal{E}} \mathbb{P}\{\widehat{\lambda}_{-k,1-\epsilon}(\widehat{\tau}_{-k} - \tau)\} &\leq \mathbb{P}(|\widehat{\tau}_{-k} - \tau|) \\ &\lesssim \|\widehat{\pi}(1 \mid \mathbf{X}) - \pi(1 \mid \mathbf{X})\| \left(\max_a \|\widehat{\mu}_a(\mathbf{X}) - \mu_a(\mathbf{X})\|\right)\end{aligned}$$

where the first inequality follows because $\sup_{\epsilon\in\mathcal{E}} |\widehat{\lambda}_{-k,1-\epsilon}| \leq 1$.

Finally, we have

$$\begin{aligned}\mathbb{P}\{(\tau - q_{1-\epsilon})(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon})\} &\lesssim \mathbb{P}[(\widehat{\lambda}_{1-\epsilon} - \lambda_{1-\epsilon})\{g(\boldsymbol{\eta}) - q_{1-\epsilon}\}] \\ &\leq \mathbb{P}[|g(\boldsymbol{\eta}) - q_{1-\epsilon}| \, |\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}) - \widehat{q}_{1-\epsilon} > 0\} - \mathbb{1}\{g(\boldsymbol{\eta}) - q_{1-\epsilon} > 0\}|] \\ &\leq \mathbb{P}[|g(\boldsymbol{\eta}) - q_{1-\epsilon}| \, \mathbb{1}\{|g(\boldsymbol{\eta}) - q_{1-\epsilon}| \leq |g(\boldsymbol{\eta}) - g(\widehat{\boldsymbol{\eta}})| + |\widehat{q}_{1-\epsilon} - q_{1-\epsilon}|\}] \\ &\lesssim (\|g(\widehat{\boldsymbol{\eta}}) - g(\boldsymbol{\eta})\|_\infty + |\widehat{q}_{1-\epsilon} - q_{1-\epsilon}|)^{1+\alpha}\end{aligned}$$

where the third inequality follows by Lemma 15 and the last inequality follows by the margin

condition (assumption (3)).

Therefore, we have that

$$\frac{\|B_{n,2}\|_{\mathcal{E}}}{\sqrt{n}} \lesssim \|\widehat{\pi}(1 \mid \mathbf{X}) - \pi(1 \mid \mathbf{X})\| \max_a \|\widehat{\mu}_a(\mathbf{X}) - \mu_a(\mathbf{X})\| + \left(\|g(\widehat{\boldsymbol{\eta}}) - g(\boldsymbol{\eta})\|_\infty + \sup_{\epsilon \in \mathcal{E}} |\widehat{q}_{1-\epsilon} - q_{1-\epsilon}| \right)^{1+\alpha}$$

where the RHS is $o_{\mathbb{P}}(n^{-1/2})$ by assumption.

A.5.2 Construction of Uniform Confidence Bands

In this section, we propose the construction of $1-\alpha$ confidence bands capturing ψ uniformly in ϵ. For any given ϵ, confidence intervals for ψ can be constructed in at least two ways. One way is to construct a confidence interval for the identification region $[\psi_l(\epsilon), \psi_u(\epsilon)]$. Another way is to construct a confidence interval for ψ directly [Imbens and Manski, 2004, Stoye, 2009, Vansteelandt et al., 2006]. The former approach yields a conservative confidence interval for ψ, particularly for larger values of ϵ for which the identification interval is wider. To see this, notice that, unless the length of the interval is of the same order as the sampling variability, the true parameter ψ can be close to either the lower bound or the upper bound, but not to both. Thus, the confidence interval in regimes of large ϵ is practically one-sided. Here, we provide confidence bands for the identification region that are valid uniformly over ϵ. These bands also serve as conservative uniform bands for the true ψ curve. We also provide the code to construct bands covering just $\psi(\epsilon)$, as in Imbens and Manski [2004], that are valid pointwise. We leave the construction of bands covering just $\psi(\epsilon)$ that are valid uniformly over ϵ for future research.

Let sample analogues of the variance functions of the bounds at ϵ be

$$\widehat{\sigma}_u^2(\epsilon) = \mathbb{P}_n([\varphi_u(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-K}, \widehat{q}_{1-\epsilon,-K}) - \mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-K}) > \widehat{q}_{1-\epsilon,-K}\}\widehat{q}_{1-\epsilon,-K} - \widehat{\psi}_u(\epsilon) + \epsilon \widehat{q}_{1-\epsilon,-K}]^2)$$
$$\widehat{\sigma}_l^2(\epsilon) = \mathbb{P}_n([\varphi_l(\mathbf{O}; \widehat{\boldsymbol{\eta}}_{-K}, \widehat{q}_{\epsilon,-K}) - \mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-K}) \leq \widehat{q}_{\epsilon,-K}\}\widehat{q}_{\epsilon,-K} - \widehat{\psi}_l(\epsilon) + \epsilon \widehat{q}_{\epsilon,-K}]^2).$$

To construct asymptotically valid $(1-\alpha)$-uniform bands of the form

$$\widehat{\mathrm{CI}}(\epsilon; c_\alpha, d_\alpha) = \left[\widehat{\psi}_l(\epsilon) - c_\alpha \frac{\widehat{\sigma}_l(\epsilon)}{\sqrt{n}}, \widehat{\psi}_u(\epsilon) + d_\alpha \frac{\widehat{\sigma}_u(\epsilon)}{\sqrt{n}} \right], \tag{A.6}$$

we need to find the critical values c_α and d_α such that

$$\mathbb{P}\left[\sup_{\epsilon \in \mathcal{E}} \left\{ \frac{\widehat{\psi}_l(\epsilon) - \psi_l(\epsilon)}{\widehat{\sigma}_l(\epsilon)/\sqrt{n}} \right\} \leq c_\alpha \text{ and } \sup_{\epsilon \in \mathcal{E}} \left\{ \frac{\psi_u(\epsilon) - \widehat{\psi}_u(\epsilon)}{\widehat{\sigma}_u(\epsilon)/\sqrt{n}} \right\} \leq d_\alpha \right] \geq 1 - \alpha + o(1)$$

In particular, we propose choosing c_α and d_α such that

$$\mathbb{P}\left[\sup_{\epsilon\in\mathcal{E}}\left\{\frac{\widehat{\psi}_l(\epsilon)-\psi_l(\epsilon)}{\widehat{\sigma}_l(\epsilon)/\sqrt{n}}\right\}\leq c_\alpha\right]=\mathbb{P}\left[\sup_{\epsilon\in\mathcal{E}}\left\{\frac{\psi_u(\epsilon)-\widehat{\psi}_u(\epsilon)}{\widehat{\sigma}_u(\epsilon)/\sqrt{n}}\right\}\leq d_\alpha\right]=1-\frac{\alpha}{2}+o(1), \tag{A.7}$$

essentially allowing the lower (upper) bound estimate to be greater (smaller) than the true lower (upper) bound with probability equal to $\alpha/2$. In light of the result in Theorem 2, c_α and d_α can be found by approximating the distribution of the supremum of the respective Gaussian processes. Similarly to Kennedy [2018], we use the multiplier bootstrap to approximate these distributions. A key advantage of this approximating method is its computational efficiency, as it does not require refitting the nuisance functions estimators.

The following lemma asserts that, for ξ and ζ iid Rademacher random variables, the suprema of the following multiplier processes

$$\sqrt{n}\mathbb{P}_n(\zeta[\varphi_l(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-K},\widehat{q}_{\epsilon;-K})-\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-K})\leq\widehat{q}_{\epsilon,-K}\}\widehat{q}_{\epsilon,-K}-\widehat{\psi}_l(\epsilon)+\epsilon\widehat{q}_{\epsilon;-K}]/\widehat{\sigma}_l(\epsilon))$$
$$\sqrt{n}\mathbb{P}_n(\xi[\widehat{\psi}_u(\epsilon)-\epsilon\widehat{q}_{1-\epsilon;-K}-\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-K},\widehat{q}_{1-\epsilon;-K})+\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-K})>\widehat{q}_{1-\epsilon,-K}\}\widehat{q}_{1-\epsilon,-K}]/\widehat{\sigma}_u(\epsilon))$$

are valid approximations to their counterparts in (A.7).

Lemma 16. *Conditional on the sample, let $\widehat{c}_\alpha$ and $\widehat{d}_\alpha$ denote the $(1-\alpha/2)$-quantiles of*

$$\sup_{\epsilon\in\mathcal{E}}\sqrt{n}\mathbb{P}_n(\zeta[\varphi_l(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-K},\widehat{q}_{\epsilon;-K})-\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-K})\leq\widehat{q}_{\epsilon,-K}\}\widehat{q}_{\epsilon,-K}-\widehat{\psi}_l(\epsilon)+\epsilon\widehat{q}_{\epsilon;-K}]/\widehat{\sigma}_l(\epsilon))$$
$$\sup_{\epsilon\in\mathcal{E}}\sqrt{n}\mathbb{P}_n(\xi[\widehat{\psi}_u(\epsilon)-\epsilon\widehat{q}_{1-\epsilon;-K}-\varphi_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-K},\widehat{q}_{1-\epsilon;-K})-\mathbb{1}\{g(\widehat{\boldsymbol{\eta}}_{-K})>\widehat{q}_{1-\epsilon,-K}\}\widehat{q}_{1-\epsilon,-K}]/\widehat{\sigma}_u(\epsilon))$$

respectively, where $(\zeta_1,\ldots,\zeta_n)$ and $(\xi_1,\ldots,\xi_n)$ are iid Rademacher random variables independent of the sample. Then, under the same conditions of Theorem 2, it holds that

$$\mathbb{P}\{[\psi_l(\epsilon),\psi_u(\epsilon)]\subseteq\widehat{CI}(\epsilon;\widehat{c}_\alpha,\widehat{d}_\alpha),\ \textit{for all}\ \epsilon\in\mathcal{E}\}\geq 1-\alpha+o(1)$$

Proof. Together with an application of the Bonferroni correction, the proof of Theorem 4 in Kennedy [2018] can be used here. □

A.5.3 Proof of Theorem 3

Recall the following map used to define ϵ_0:

$$\Psi(\epsilon)=\psi_l(\epsilon)\psi_u(\epsilon)=\mathbb{P}\left\{\varphi_l(\mathbf{O};\boldsymbol{\eta};q_\epsilon)\right\}\mathbb{P}\left\{\varphi_u(\mathbf{O};\boldsymbol{\eta};q_{1-\epsilon})\right\}$$

where

$$\varphi_l(\mathbf{O};\boldsymbol{\eta},q_\epsilon)=\nu(\mathbf{O};\boldsymbol{\eta})+\tau(\mathbf{O};\boldsymbol{\eta})\kappa_\epsilon-\epsilon,\quad \varphi_u(\mathbf{O};\boldsymbol{\eta},q_{1-\epsilon})=\nu(\mathbf{O};\boldsymbol{\eta})+\tau(\mathbf{O};\boldsymbol{\eta})\lambda_{1-\epsilon},$$

$\kappa_\epsilon = \mathbb{1}\{g(\boldsymbol{\eta}) \leq q_\epsilon\}$ and $\lambda_{1-\epsilon} = \mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon}\}$. The corresponding empirical version, which makes use of cross-fitting, is:

$$\widehat{\Psi}_n(\epsilon) = \frac{1}{B}\sum_{k=1}^{B} \mathbb{P}_n^k\left\{\widehat{\varphi}_l(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k},\widehat{q}_{-k,\epsilon})\right\}\mathbb{P}_n^k\left\{\widehat{\varphi}_u(\mathbf{O};\widehat{\boldsymbol{\eta}}_{-k},\widehat{q}_{-k,1-\epsilon})\right\}$$

where $\mathbb{P}_n^k$ is the empirical measure over fold k, defined as in Section 2.3.1.

The moment condition defining ϵ_0 is $\Psi(\epsilon_0) = 0$, since at $\epsilon = \epsilon_0$ either the lower bound or the upper bound is equal to 0 and both are uniformly bounded so that the product is 0. Furthermore, the lower and upper bound curves are monotone in ϵ; if the bounds are continuous and strictly monotone in a neighborhood of ϵ_0, then the moment condition will be satisfied by a unique value in $[0,1]$. In practice, we would estimate ϵ_0 by ϵ_n solving the empirical moment condition $\widehat{\Psi}_n(\epsilon_n) = o_{\mathbb{P}}(n^{-1/2})$.

Theorem 3 follows from a direct application of Theorem 3.3.1 in van der Vaart and Wellner [1996]. Therefore, our proof consists of checking that the following conditions hold:

1. $\sqrt{n}(\widehat{\Psi}_n - \Psi)(\epsilon_0) \rightsquigarrow N(0, \text{var}\{\widetilde{\varphi}(\mathbf{O};\boldsymbol{\eta},\epsilon_0)\})$, where

   $$\widetilde{\varphi}(\mathbf{O};\boldsymbol{\eta},\epsilon) = \psi_u(\epsilon)[\nu(\mathbf{O};\boldsymbol{\eta}) + \kappa_\epsilon\{\tau(\mathbf{O};\boldsymbol{\eta}) - q_\epsilon\} - \epsilon] + \psi_l(\epsilon)[\nu(\mathbf{O};\boldsymbol{\eta}) + \lambda_{1-\epsilon}\{\tau(\mathbf{O};\boldsymbol{\eta}) - q_{1-\epsilon}\}]$$

2. $\sqrt{n}(\widehat{\Psi}_n - \Psi)(\epsilon_n) - \sqrt{n}(\widehat{\Psi}_n - \Psi)(\epsilon_0) = o_{\mathbb{P}}\left(1 + \sqrt{n}\,|\epsilon_n - \epsilon_0|\right)$
3. The map $\epsilon \mapsto \Psi(\epsilon)$ is differentiable at $\epsilon = \epsilon_0$.
4. ϵ_n is such that $\widehat{\Psi}_n(\epsilon_n) = o_{\mathbb{P}}(n^{-1/2})$ and $\epsilon_n \xrightarrow{p} \epsilon_0$.

We will follow the same notation as for the proof of Theorem 2. In particular, let $\|f\|_{\mathcal{E}} = \sup_{\epsilon\in\mathcal{E}}|f(\epsilon)|$ denote the supremum norm over $\mathcal{E}$. We proceed with considering $\mathcal{E} = [0,1]$.

Proof of Statement 1

We actually prove the following stronger result:

$$\|\sqrt{n}(\widehat{\Psi}_n - \Psi) - \sqrt{n}(\mathbb{P}_n - \mathbb{P})\widetilde{\varphi}\|_{\mathcal{E}} = o_{\mathbb{P}}(1),$$

for $\widetilde{\varphi}(\cdot;\boldsymbol{\eta},\epsilon)$ living in a Donsker class. This is useful in establishing the other conditions.

First, we claim that the function $\widetilde{\varphi}(\cdot;\boldsymbol{\eta},\epsilon)$ lives in a Dosker class. To see this, notice that

$$\widetilde{\varphi}(\mathbf{O};\boldsymbol{\eta},\epsilon) = \psi_u(\epsilon)\overline{\varphi}_l(\mathbf{O};\boldsymbol{\eta},\epsilon) + \psi_l(\epsilon)\overline{\varphi}_u(\mathbf{O};\boldsymbol{\eta},\epsilon)$$

where $\overline{\varphi}_l(\mathbf{O};\boldsymbol{\eta},\epsilon) = \nu(\mathbf{O};\boldsymbol{\eta}) + \kappa_\epsilon\{\tau(\mathbf{O};\boldsymbol{\eta}) - q_\epsilon\} - \epsilon$ and $\overline{\varphi}_u(\mathbf{O};\boldsymbol{\eta},\epsilon) = \nu(\mathbf{O};\boldsymbol{\eta}) + \lambda_{1-\epsilon}\{\tau(\mathbf{O};\boldsymbol{\eta}) - q_{1-\epsilon}\}$. In the proof of Theorem 2, we showed that $\overline{\varphi}_u(\cdot;\boldsymbol{\eta},\epsilon)$ lives in a Donsker class because its class can be constructed via sums and products of classes of uniformly bounded, monotone functions. Therefore, following a similar logic, we conclude that $\widetilde{\varphi}(\cdot;\boldsymbol{\eta},\epsilon)$ lives in a Donsker class as well.

Next, we argue that $\|\sqrt{n}(\widehat{\Psi}_n - \Psi) - \sqrt{n}(\mathbb{P}_n - \mathbb{P})\widetilde{\varphi}\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$. A bit of algebra reveals that

$$\begin{aligned}\widehat{\Psi}_n(\epsilon) - \Psi(\epsilon) - (\mathbb{P}_n - \mathbb{P})\widetilde{\varphi} = \frac{1}{B}\sum_{k=1}^{B}\Big[&(\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{l,-k} - \overline{\varphi}_l)(\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{u,-k} - \overline{\varphi}_u)\\ &+ (\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{l,-k} - \overline{\varphi}_l)\{T_1 - T_2 + (\mathbb{P}_n - \mathbb{P})(\varphi_u)\}\\ &+ (\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{u,-k} - \overline{\varphi}_u)\{V_1 - V_2 + (\mathbb{P}_n - \mathbb{P})(\varphi_l)\}\\ &+ (\mathbb{P}_n - \mathbb{P})(\varphi_u)(V_1 - V_2) + (\mathbb{P}_n - \mathbb{P})(\varphi_l)(T_1 - T_2)\\ &+ (\mathbb{P}_n - \mathbb{P})(\varphi_u)(\mathbb{P}_n - \mathbb{P})(\varphi_l)\\ &+ T_1V_1 - T_1V_2 - T_2V_1 + T_2V_2\\ &+ \mathbb{P}(\varphi_u)\{(\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{l,-k} - \overline{\varphi}_l) + V_1\}\\ &+ \mathbb{P}(\varphi_l)\{(\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{u,-k} - \overline{\varphi}_u) + T_1\}\ \Big]\end{aligned}$$

where

$$\begin{aligned}&\widehat{\overline{\varphi}}_{l,-k} - \overline{\varphi}_l = \widehat{\varphi}_{l-k} - q_\epsilon\widehat{\kappa}_{-k,\epsilon} - \varphi_{l-k} + q_\epsilon\kappa_\epsilon\\ &\widehat{\overline{\varphi}}_{u,-k} - \overline{\varphi}_u = \widehat{\varphi}_{u-k} - q_{1-\epsilon}\widehat{\lambda}_{-k,1-\epsilon} - \varphi_u + q_{1-\epsilon}\lambda_{1-\epsilon}\\ &V_1 = \mathbb{P}\{\widehat{\nu}_{-k} - \nu + \widehat{\kappa}_{-k,\epsilon}(\widehat{\tau}_{-k} - \tau) + (\tau - q_\epsilon)(\widehat{\kappa}_{-k,\epsilon} - \kappa_\epsilon)\}\\ &T_1 = \mathbb{P}\{\widehat{\nu}_{-k} - \nu + \widehat{\lambda}_{-k,1-\epsilon}(\widehat{\tau}_{-k} - \tau) + (\tau - q_{1-\epsilon})(\widehat{\lambda}_{-k,1-\epsilon} - \lambda_{1-\epsilon})\}\\ &V_2 = (\mathbb{P}_n - \mathbb{P})(q_\epsilon\kappa_\epsilon) \quad \text{and} \quad T_2 = (\mathbb{P}_n - \mathbb{P})(q_{1-\epsilon}\lambda_{1-\epsilon})\end{aligned}$$

As shown in the proof of Theorem 2, under the conditions of the theorem, it holds that

$$\left\|\frac{1}{B}\sum_{k=1}^{B}(\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{l,-k} - \overline{\varphi}_l)\right\|_{\mathcal{E}} = o_{\mathbb{P}}(n^{-1/2}), \quad \left\|\frac{1}{B}\sum_{k=1}^{B}(\mathbb{P}_n^k - \mathbb{P})(\widehat{\overline{\varphi}}_{u,-k} - \overline{\varphi}_u)\right\|_{\mathcal{E}} = o_{\mathbb{P}}(n^{-1/2}),$$

$$\|V_1\|_{\mathcal{E}} = o_{\mathbb{P}}(n^{-1/2}), \quad \|T_1\|_{\mathcal{E}} = o_{\mathbb{P}}(n^{-1/2}), \quad \|V_2\|_{\mathcal{E}} = O_{\mathbb{P}}(n^{-1/2}), \quad \text{and} \quad \|T_2\|_{\mathcal{E}} = O_{\mathbb{P}}(n^{-1/2}).$$

Therefore, by an application of the triangle inequality, it holds that

$$\|\sqrt{n}(\widehat{\Psi}_n - \Psi) - \sqrt{n}(\mathbb{P}_n - \mathbb{P})\widetilde{\varphi}\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$$

In particular,

$$\sqrt{n}(\widehat{\Psi}_n - \Psi)(\epsilon_0) \rightsquigarrow N(0, \text{var}\{\widetilde{\varphi}(\mathbf{O}; \boldsymbol{\eta}, \epsilon_0)\})$$

by Slutsky's theorem.

Proof of Statement 2

Because in the proof of Statement 1 we have argued that

$$\|\sqrt{n}(\widehat{\Psi}_n - \Psi) - \sqrt{n}(\mathbb{P}_n - \mathbb{P})\widetilde{\varphi}\|_{\mathcal{E}} = o_{\mathbb{P}}(1)$$

to prove Statement 2, it is sufficient to show

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}(\mathbf{O};\boldsymbol{\eta},\epsilon_n)\} - \sqrt{n}(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}(\mathbf{O};\boldsymbol{\eta},\epsilon_0)\} = o_{\mathbb{P}}(1 + \sqrt{n}\,|\epsilon_n - \epsilon_0|) \tag{A.8}$$

Because $\widetilde{\varphi}(\cdot;\boldsymbol{\eta},\epsilon)$ lives in a Donsker class and $\epsilon_n \xrightarrow{p} \epsilon_0$ (proved below in the proof of Statement 4), by Lemma 3.3.5 in van der Vaart and Wellner [1996], in order to prove (A.8) it is sufficient to show that

$$\mathbb{P}\{\widetilde{\varphi}(\epsilon) - \widetilde{\varphi}(\epsilon_0)\}^2 \to 0 \quad \text{as} \quad \epsilon \to \epsilon_0$$

We have that

$$\begin{aligned}\mathbb{P}\{\widetilde{\varphi}(\epsilon) - \widetilde{\varphi}(\epsilon_0)\}^2 &= \mathbb{P}[\psi_l(\epsilon)\{\nu + \lambda_{1-\epsilon}(\tau - q_{1-\epsilon})\} - \psi_l(\epsilon_0)\{\nu + \lambda_{1-\epsilon_0}(\tau - q_{1-\epsilon_0})\} \\ &\quad + \psi_u(\epsilon)\{\nu + \kappa_\epsilon(\tau - q_\epsilon) - \epsilon\} - \psi_u(\epsilon_0)\{\nu + \kappa_{\epsilon_0}(\tau - q_{\epsilon_0}) - \epsilon_0\}]^2 \\ &= \mathbb{P}\,(D_l + D_u)^2\end{aligned}$$

Notice that we can write

$$\begin{aligned}D_l &= \{\psi_l(\epsilon) - \psi_l(\epsilon_0)\}\{\nu + \lambda_{1-\epsilon}(\tau - q_{1-\epsilon})\} \\ &\quad + \psi_l(\epsilon_0)\{(\lambda_{1-\epsilon} - \lambda_{1-\epsilon_0})(\tau - q_{1-\epsilon}) + \lambda_{1-\epsilon_0}(q_{1-\epsilon_0} - q_{1-\epsilon})\} \\ D_u &= \{\psi_u(\epsilon) - \psi_u(\epsilon_0)\}(\nu + \kappa_\epsilon(\tau - q_\epsilon) - \epsilon_n) \\ &\quad + \psi_u(\epsilon_0)\{(\kappa_\epsilon - \kappa_{\epsilon_0})(\tau - q_\epsilon) + \kappa_{\epsilon_0}(q_{\epsilon_0} - q_\epsilon) - (\epsilon - \epsilon_0)\}\end{aligned}$$

Then, we have

$$\begin{aligned}\mathbb{P}(D_l^2) &\lesssim \mathbb{P}|\lambda_{1-\epsilon} - \lambda_{1-\epsilon_0}| \;+\; |q_{1-\epsilon_0} - q_{1-\epsilon}| \;+\; |\psi_l(\epsilon) - \psi_l(\epsilon_0)| \\ \mathbb{P}(D_u^2) &\lesssim \mathbb{P}|\kappa_\epsilon - \kappa_{\epsilon_0}| \;+\; |q_{\epsilon_0} - q_\epsilon| \;+\; |\psi_u(\epsilon) - \psi_u(\epsilon_0)| \;+\; |\epsilon - \epsilon_0| \\ \mathbb{P}(D_l D_u) &\lesssim\; |\psi_l(\epsilon) - \psi_l(\epsilon_0)| \;+\; |\psi_u(\epsilon) - \psi_u(\epsilon_0)| \;+\; \mathbb{P}|\lambda_{1-\epsilon} - \lambda_{1-\epsilon_0}| \;+\; \mathbb{P}|\kappa_\epsilon - \kappa_{\epsilon_0}| \\ &\quad + |q_{1-\epsilon_0} - q_{1-\epsilon}| \;+\; |q_{\epsilon_0} - q_\epsilon| \;+\; |\epsilon - \epsilon_0|\end{aligned}$$

Next, notice

$$\begin{aligned}\mathbb{P}|\kappa_\epsilon - \kappa_{\epsilon_0}| \;&\leq \mathbb{P}[\mathbb{1}\{|g(\boldsymbol{\eta}) - q_{\epsilon_0}| \;\leq |q_{\epsilon_0} - q_\epsilon|\}] \lesssim |q_{\epsilon_0} - q_\epsilon|^\alpha \\ \mathbb{P}|\lambda_{1-\epsilon} - \lambda_{1-\epsilon_0}| \;&\leq \mathbb{P}[\mathbb{1}\{|g(\boldsymbol{\eta}) - q_{1-\epsilon_0}| \;\leq |q_{1-\epsilon_0} - q_{1-\epsilon}|\}] \lesssim |q_{1-\epsilon_0} - q_{1-\epsilon}|^\alpha\end{aligned}$$

for some $\alpha > 0$. The first inequalities rely on Lemma 15. The last step hinges on the fact that the density of $g(\boldsymbol{\eta})$ satisfies the margin condition 3 for some $\alpha > 0$.

Moreover, we have

$$|\psi_l(\epsilon) - \psi_l(\epsilon_0)| \lesssim \mathbb{P}|\kappa_\epsilon - \kappa_{\epsilon_0}| + |\epsilon - \epsilon_0| \quad \text{and} \quad |\psi_u(\epsilon) - \psi_u(\epsilon_0)| \lesssim \mathbb{P}|\lambda_{1-\epsilon} - \lambda_{1-\epsilon_0}|$$

since $\mathbb{P}(|g(\boldsymbol{\eta})| \leq 1) = 1$.

We have assumed that the CDF of $g(\boldsymbol{\eta})$ is continuous and strictly increasing in neighborhoods of q_{ϵ_0} and $q_{1-\epsilon_0}$, thus the quantile function is continuous in neighborhoods of ϵ_0 and $1 - \epsilon_0$ as well, allowing us to conclude that, for $\alpha > 0$

$$|q_{\epsilon_0} - q_\epsilon|^\alpha \to 0 \quad \text{and} \quad |q_{1-\epsilon_0} - q_{1-\epsilon}|^\alpha \to 0 \quad \text{as} \quad \epsilon \to \epsilon_0.$$

Then, it follows that $\mathbb{P}\{\widetilde{\varphi}(\epsilon) - \widetilde{\varphi}(\epsilon_0)\}^2 \to 0$ as $\epsilon \to \epsilon_0$.

Proof of Statement 3

To prove Statement 3, notice that

$$\psi_l(\epsilon)\psi_u(\epsilon) = \left[\mathbb{E}\left\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\right\} + \int_0^{q_\epsilon} t dG(t) - \epsilon\right]\left[\mathbb{E}\left\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\right\} + \int_{q_{1-\epsilon}}^{1} t dG(t)\right]$$

Because we have assumed that the quantile function of $g(\boldsymbol{\eta})$ is differentiable in neighborhoods of ϵ_0 and $1 - \epsilon_0$, by "Leibniz integral rule," it holds that

$$\Psi'(\epsilon_0) = \frac{d}{d\epsilon}\psi_l(\epsilon)\psi_u(\epsilon)\bigg|_{\epsilon=\epsilon_0} = \psi_u(\epsilon_0)(q_{\epsilon_0} - 1) + \psi_l(\epsilon_0)q_{1-\epsilon_0}$$

which we have assumed to be nonzero. Notice that in calculating the derivative, we used the fact that $\int t dG(t) = \int t f(t) dt$ with f being the density of $g(\boldsymbol{\eta})$, which we have assumed to exist.

Proof of Statement 4

We have that $\Psi_n(\epsilon_n) = o_{\mathbb{P}}(n^{-1/2})$ by definition. Furthermore, we have shown that

$$\|\Psi_n - \Psi\|_{\mathcal{E}} = \|(\mathbb{P}_n - \mathbb{P})\{\widetilde{\varphi}(\mathbf{O}; \boldsymbol{\eta}, \epsilon)\}\|_{\mathcal{E}} + o_{\mathbb{P}}(n^{-1/2}) = o_{\mathbb{P}}(1)$$

where the last equality follows because $\widetilde{\varphi}(\cdot; \boldsymbol{\eta}, \epsilon)$ is Donsker and thus Glivenko-Cantelli.

We now show that $\psi_l(\epsilon)$ and $\psi_u(\epsilon)$ are strictly monotone. First, for $\epsilon_1 < \epsilon_2$, we have

$$\begin{aligned}
\psi_l(\epsilon_1) - \psi_l(\epsilon_2) &= \mathbb{E}(g(\boldsymbol{\eta})[\mathbb{1}\{g(\boldsymbol{\eta}) \leq q_{\epsilon_1}\} - \mathbb{1}\{g(\boldsymbol{\eta}) \leq q_{\epsilon_2}\}]) - (\epsilon_1 - \epsilon_2) \\
&= -\mathbb{E}\{g(\boldsymbol{\eta}) \mid q_{\epsilon_1} < g(\boldsymbol{\eta}) < q_{\epsilon_2}\}\mathbb{P}\{q_{\epsilon_1} < g(\boldsymbol{\eta}) < q_{\epsilon_2}\} - (\epsilon_1 - \epsilon_2) \\
&= -\mathbb{E}\{g(\boldsymbol{\eta}) \mid q_{\epsilon_1} < g(\boldsymbol{\eta}) < q_{\epsilon_2}\}(\epsilon_2 - \epsilon_1) + (\epsilon_2 - \epsilon_1) \\
&> 0
\end{aligned}$$

where we used the facts that $\mathbb{P}\{0 < g(\boldsymbol{\eta}) < 1\} = 1$ and $\mathbb{P}\{q_{\epsilon_1} < g(\boldsymbol{\eta}) < q_{\epsilon_2}\} = \epsilon_2 - \epsilon_1$ (continuity of $g(\boldsymbol{\eta})$), that $\mathbb{1}\{g(\boldsymbol{\eta}) \leq q_{\epsilon_1}\} \leq \mathbb{1}\{g(\boldsymbol{\eta}) \leq q_{\epsilon_2}\}$ (monotonicity of quantile function) and that

$$\mathbb{1}\{g(\boldsymbol{\eta}) \leq q_{\epsilon_1}\} - \mathbb{1}\{g(\boldsymbol{\eta}) \leq q_{\epsilon_2}\} = -1 \iff q_{\epsilon_1} < g(\boldsymbol{\eta}) < q_{\epsilon_2}$$

Similarly, we note that, for $\epsilon_1 < \epsilon_2$, we have

$$\begin{aligned}\psi_u(\epsilon_1) - \psi_u(\epsilon_2) &= \mathbb{E}(g(\boldsymbol{\eta})[\mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon_1}\} - \mathbb{1}\{g(\boldsymbol{\eta}) > q_{1-\epsilon_2}\}]) \\ &= -\mathbb{E}\{g(\boldsymbol{\eta}) \mid q_{1-\epsilon_2} < g(\boldsymbol{\eta}) < q_{1-\epsilon_1}\}(\epsilon_2 - \epsilon_1) \\ &< 0\end{aligned}$$

using the same logic as before. Thus, we conclude that, under the assumption that $g(\boldsymbol{\eta})$ is a continuous random variable, both $\psi_l(\epsilon)$ and $\psi_u(\epsilon)$ are continuous and strictly monotone. Therefore, the value ϵ_0 satisfying $\Psi(\epsilon_0) = 0$ must be unique. Furthermore, we have assumed (to derive a finite asymptotic variance of ϵ_n) that $\Psi^{'}(\epsilon_0) \neq 0$, thus a first-order Taylor expansion of $\Psi(\epsilon_n)$ around ϵ_0

$$\Psi(\epsilon_n) = \Psi^{'}(\epsilon_0)(\epsilon_n - \epsilon_0) + o(|\epsilon_n - \epsilon_0|)$$

suffices to conclude that $|\Psi(\epsilon_n)| \to 0$ implies $|\epsilon_n - \epsilon_0| \to 0$ for any sequence $\epsilon_n \in \mathcal{E}$. In other words, under the assumptions of the theorem, the identifiability condition of ϵ_0 is satisfied. Then, by an application of Theorem 2.10 in Kosorok [2008], we conclude that $|\epsilon_n - \epsilon_0| = o_{\mathbb{P}}(1)$ as desired.

A.6 Additional Data Analysis

In this section, we provide additional analysis of the data from Connors et al. [1996]. In Figure A.1, we consider values of δ smaller than 1, and notice that the bounds would start to include zero for larger values of ϵ. For instance, under the X-mixture model, if $\delta = 1/2$ is used, the results appear to be robust for up to 11.00% (95% CI $= [3.84\%, 18.16\%]$) of confounded units in the sample. A value of $\delta = 1/2$ requires that the counterfactual mean outcomes satisfy:

$$\frac{\mu_a(\mathbf{X})}{2} \leq \mathbb{E}(Y^a \mid A = 1 - a, \mathbf{X}, S = 0) \leq \frac{1}{2} + \frac{\mu_a(\mathbf{X})}{2} \text{ with prob. 1.}$$

for $a \in \{0, 1\}$, thereby restricting $\mathbb{E}(Y^a \mid A = 1 - a, \mathbf{X}, S = 0)$ to be in an interval of length 1/2 instead of the worst-case interval of length 1. Robustness is up to 8.99% (95% CI $= [3.78\%, 14.20\%]$) confounded units if the XA-mixture model is considered instead.

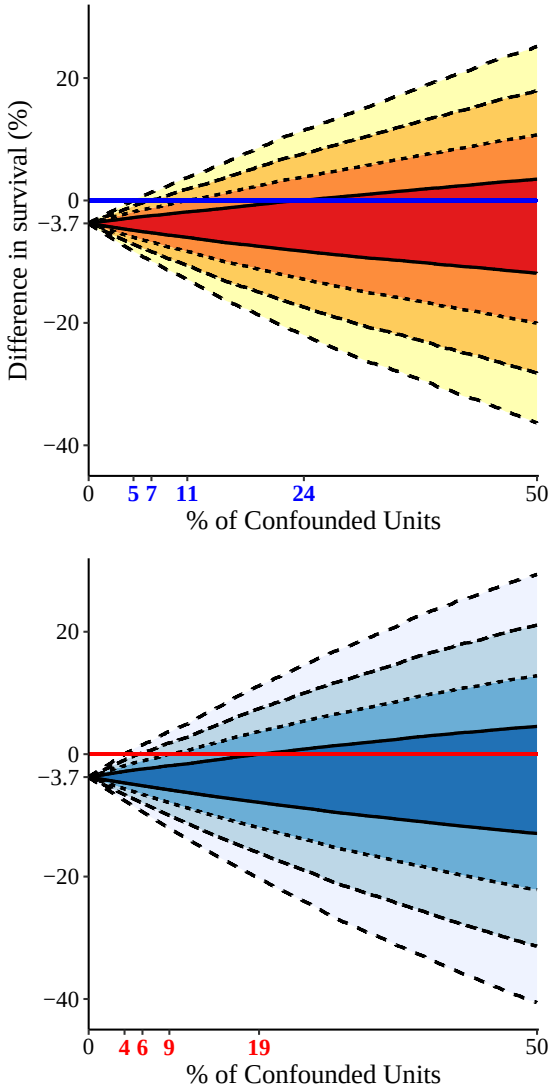


Figure A.1: Estimated bounds on the average treatment effect as a function of the proportion of confounded units ϵ and the parameter $\delta \in \{0.25, 0.5, 0.75, 1\}$, which governs the amount of confounding among the $S = 0$ units. Darker shades correspond to smaller values of δ. Bolded labels on the abscissa represent estimates of ϵ_0 for corresponding values of δ. Uniform and pointwise confidence intervals are not shown for the sake of clarity.

A.6.1 Results using the sensitivity model from Cinelli and Hazlett [2020]

In this section, we briefly report the results from applying the sensitivity analysis for linear models discussed in Cinelli and Hazlett [2020]. Because their model is appropriate only for causal effect estimates computed using OLS, we fit a linear model for 30-day survival regressed all baseline covariates, the treatment and no interactions. If the model is accurate and there is

no residual confounding, RHC usage appears to decrease the probability of 30-day survival by 0.042 ($95\%\text{CI} = [-0.07, -0.02]$). However, it is sufficient that an unmeasured confounder explains 1.7% of the outcome variance not already captured by the treatment and the covariates and 1.7% of the treatment variance not already captured by the covariates to make the effect not statistically significance at the 0.05-level (Table A.1). As shown in Figure A.2, the observed

Outcome: survival at day 30

Treatment:	Est.	S.E.	t-value	$R^2_{Y\sim D\|\mathbf{X}}$	$RV_{q=1}$	$RV_{q=1,\alpha=0.05}$
RHC usage	-0.042	0.013	-3.261	0.2%	4.2%	1.7%
df = 5658				*Bound (2x* $\mathrm{dnr1}$*)*: $R^2_{Y\sim Z\|\mathbf{X},D}$ = 3.7%, $R^2_{D\sim Z\|\mathbf{X}}$ = 1%		
df = 5658				*Bound (2x* $\mathrm{is\_miss\_adld3p}$*)*: $R^2_{Y\sim Z\|\mathbf{X},D}$ = 6%, $R^2_{D\sim Z\|\mathbf{X}}$ = 1.6%		

Table A.1: Summary of the effect estimate under no unmeasured confounding as well as assessment of the estimate's robustness using some key covariates as benchmarks.

effect would cease to be significance also if there is an unmeasured confounder with explanatory power that is 2 times greater than that of the variable $\mathrm{dnr1}$ (an indicator for whether there was a "do not resuscitate" order when the patient was admitted on day 1) or 2 times greater than that of the variable indicating that $\mathrm{adld3p}$ (ADL) is missing.

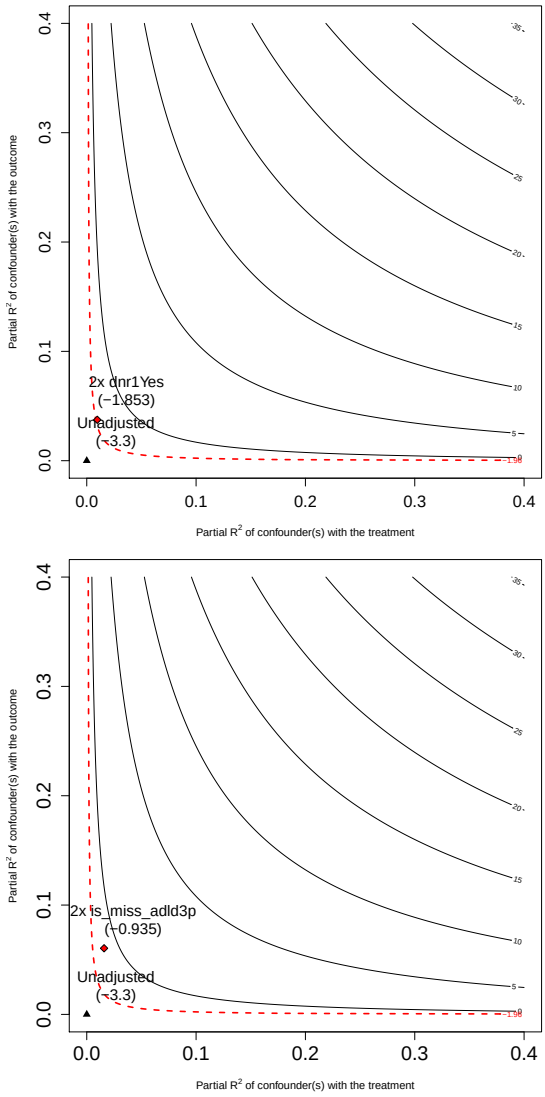


Figure A.2: Sensitivity contour plots in the partial R^2 scale with benchmark bounds of the t-value.

A.7 Simulations regarding power

In this section, we conduct a brief simulation to investigate how conservative inference based on ϵ_0 is when all the confounders have been measured so that the true ϵ is actually zero. The bounds on the ATE τ depends on three fundamental quantities, $\mu_a(\mathbf{X}) = \mathbb{E}(Y \mid A = a, \mathbf{X})$ for $a = 0, 1$ and $\pi(\mathbf{X}) = \mathbb{P}(A = 1 \mid \mathbf{X})$. Let $Q_g(p, \delta)$ be the quantile function of either $g(\boldsymbol{\eta})$ as defined in Theorem 1 (X-model) or $g(A, \boldsymbol{\eta})$ as defined in Section A.3 (XA-model). When

$\epsilon = 0$, $\tau = \mathbb{E}\{\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\}$ and the bounds can be written as

$$\psi_l(\epsilon, \delta) = \tau + \int_0^{\epsilon} Q_g(p, \delta)dp - \epsilon\delta(y_{\max} - y_{\min}) \quad \text{and} \quad \psi_u(\epsilon, \delta) = \tau + \int_{1-\epsilon}^{1} Q_g(p, \delta)dp.$$

Therefore, we may define the design sensitivity [Rosenbaum, 2004] as $\widetilde{\epsilon}$ solving

$$\tau + \int_0^{\widetilde{\epsilon}} Q_g(p, \delta)dp - \widetilde{\epsilon}\delta(y_{\max} - y_{\min}) = 0 \text{ if } \tau \geq 0 \quad \text{and} \quad \tau + \int_{1-\widetilde{\epsilon}}^{1} Q_g(p, \delta)dp = 0 \text{ if } \tau < 0.$$

Thus, $\widetilde{\epsilon}$ depends on τ and the quantile function $Q_g(p, \delta)$, which itself depends on τ through the functions $\mu_a(\mathbf{X})$. Without knowing $Q_g(p, \delta)$, one can get crude bounds on $\widetilde{\epsilon}$ as

$$\begin{aligned} &\frac{\tau}{\delta(y_{\max} - y_{\min}) - Q_g(0, \delta)} \leq \widetilde{\epsilon} \leq \frac{\tau}{\delta(y_{\max} - y_{\min}) - Q_g(1, \delta)} && \text{if } \tau \geq 0 \\ &\frac{|\tau|}{Q_g(1, \delta)} \leq \widetilde{\epsilon} \leq \frac{|\tau|}{Q_g(0, \delta)} && \text{if } \tau < 0 \end{aligned}$$

since, for example, $\int_0^{\epsilon} Q_g(p, \delta)dp \geq \epsilon Q_g(0, \delta)$. The derivatives of the bounds are

$$\frac{d}{d\epsilon}\psi_l(\epsilon, \delta) = Q_g(\epsilon, \delta) - \delta(y_{\max} - y_{\min}) \quad \text{and} \quad \frac{d}{d\epsilon}\psi_u(\epsilon, \delta) = Q_g(1 - \epsilon, \delta).$$

so that the rate at which they widen crucially depends on $Q_g(p, \delta)$. For example, let the data be generated as in the simulation setup of Section 2.4.1 except for

$$Y^a \mid X_1, X_2, U, S, A \sim \text{Bern}\{(1 - a)/2 + aB_{\alpha}^{-1} \circ TN(X_1)\},$$

where $B_{\alpha}^{-1}(\cdot)$ is the quantile function of a Beta$(\alpha, 1)$ random variable and $TN(\cdot)$ is the CDF of X_1, a truncated normal random variable in $[-2, 2]$. Therefore, $\mu_0(X_1, X_2) = 1/2$, $\mu_1(X_1, X_2) \sim \text{Beta}(\alpha, 1)$, and $\tau = \alpha/(\alpha + 1) - 1/2$. As shown in Figure A.3, as τ increases $\widetilde{\epsilon}$ increases, although the relationship is nonlinear. In fact, different values of α also affects the skewness of the distribution of $g(\boldsymbol{\eta})$, for example. Finally, $\widetilde{\epsilon}$ is always greater under the X-model than under the XA-model because the bounds under the latter are at least as wide as those under the former model.

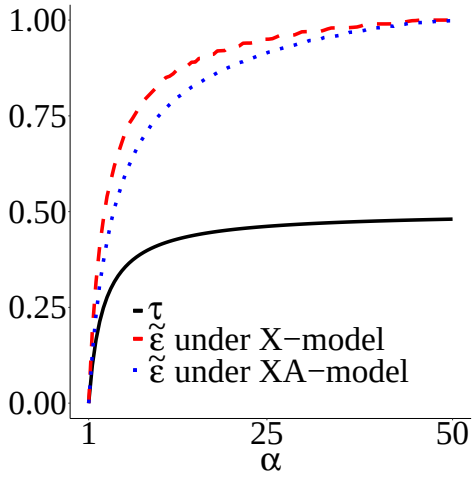


Figure A.3: The values of τ and $\widetilde{\epsilon}$ under either the X-model or the XA-model are shown as a function of α.

Appendix B

Appendix for Chapter 3

B.0.1 Synthetic Examples

As a proof of concept, we consider a simple simulated example. We take $n = 100$, $A_1, \ldots, A_n \sim N(0,1)$ and $Y_i = \beta A_i + \epsilon_i$ where $\epsilon_i \sim N(0,1)$. Figure B.1a shows the propensity sensitivity bounds for $\beta = 3$ based on the homotopy algorithm and using the local approximation; the local method is an excellent approximation. We also conducted a few simulations using very small sample sizes where the exact solution can be computed by brute force. We found that the homotopy method was indistinguishable from the exact bound. Figure B.1b shows the bounds using the outcome sensitivity approach.

Now we look at the effect of bounding over $\mathcal{V}_{\text{large}}$ using F_1 and F_2. The example in Figure B.2 shows that the propensity sensitivity bounds from F_2 (green) are wider than bounds from F_1 (black). In this case we used $n = 1000$,1 $X \sim N(0,1)$, $A = X + N(0,1)$, $Y = \beta A + 2X + N(0,1)$, with $\beta = 3$. Conversely, the example in Figure B.2b shows that the propensity sensitivity bounds from F_2 (green) are narrower than bounds from F_1 (black). Here we used $n = 1000$ with: $U \sim \text{Unif}(.5, 1)$, $A = 3 - U$, $Y = 5U$ and $Y = 2.5U + .25N(0,1)$. The red dotted lines are the local approximations to the F_1 bounds, which are very good in these two examples as well. Our experience is that usually F_1 gives tighter bounds.

B.0.2 Subset Confounding

Recall that, under this model, an unknown proportion of the population is subject to unobserved confounding. Suppose S is such that $Y(a) \perp\!\!\!\perp A | X, S = 1$ but $Y(a) \perp\!\!\!\perp A | X, U, S = 0$ where U is not observed. That is, $S = 0$ represents the subset with unmeasured confounding and $S = 1$ represents the subset with no unmeasured confounding. This is a sensitivity model proposed by Bonvini and Kennedy [2020] in the case of binary treatments. Here, we extend this framework to multivalued treatments and MSMs under the propensity sensitivity model

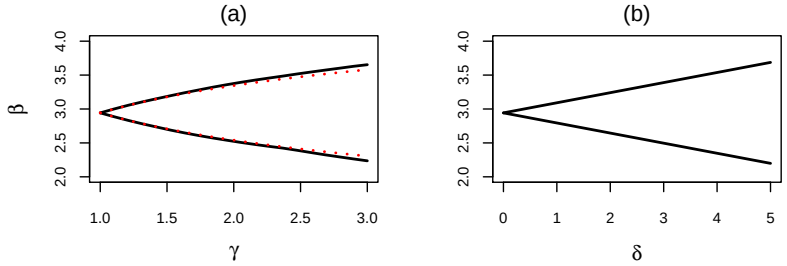


Figure B.1: **(a) Propensity sensitivity model bounds and (b) Outcome sensitivity model bounds for β in the MSM $g(a;\beta) = \beta a$ with $\beta = 3$,** for a simulated example. In (a), the black bounds in are from F_1 over $\mathcal{V}_{\text{large}}$, obtained by the homotopy algorithm (Section 3.4.4), and their local approximations (Section 3.4.6) are in dotted red.

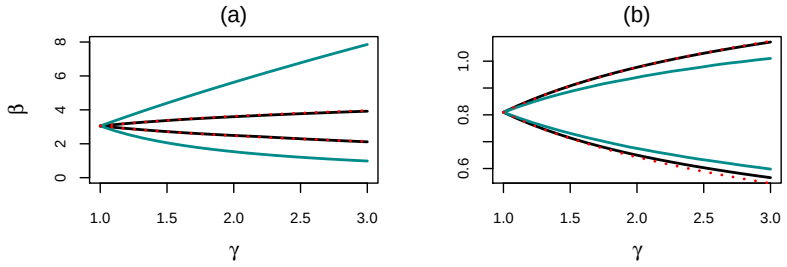


Figure B.2: **Propensity sensitivity model bounds from F_1 (black) and F_2 (green) over $\mathcal{V}_{\text{large}}$,** in simulated examples. Bounds from F_2 are computationally easier to obtain than bounds from F_1. The local approximations to F_1, which are also simple to obtain, are in dotted red. (a) In this example, bounds from F_2 are wider than bounds from F_1, so we use the latter. (b) In this other example, the reverse is true. We do not know in advance which bounds to use.

and the outcome sensitivity model. To start, we define the propensity model in this case to be

$$\gamma^{-1} \leq \frac{\pi(a|x,u,S=0)}{\pi(a|x,S=0)} \leq \gamma \quad \text{for all } a,x,u.$$

Let

$$v_0(Z) = \mathbb{E}\left\{ \frac{\pi(A|X,S=0)}{\pi(A|X,S=0,U)} \,\middle|\, Y,A,X,S=0 \right\}$$

and notice that $\mathbb{E}\{v_0(Z)|A,X,S=0\} = 1$ and $v_0(Z) \in [\gamma^{-1},\gamma]$. Essentially, we can repeat the same calculations as in the non-contaminated model, this time simply applied to the $S=0$ group.

The same argument used in proving Lemma 2 yields that

$$\begin{aligned}\mathbb{E}\{Y\gamma^{\text{sgn}\{q_\ell(Y|a,x,S=0)-Y\}}|A=a,X=x,S=0\}\\ \leq \mathbb{E}\{Yv_0(Z)|A=a,X=x,S=0\}\leq\\ \mathbb{E}\{Y\gamma^{\text{sgn}\{Y-q_u(Y|a,x,S=0)\}}|A=a,X=x,S=0\}\end{aligned}$$

where $q_\ell(Y|a,x,S=0)$ and $q_u(Y|a,x,S=0)$ are the $\tau_\ell = 1/(\gamma+1)$ and $\tau_u = \gamma/(1+\gamma)$ quantiles of the distribution of $Y|(A,X,S=0)$. As in Bonvini and Kennedy [2020], we make the simplifying assumption that $S \perp\!\!\!\perp Y|A,X$. This way,

$$\begin{aligned}m_\ell(a,x)\equiv\mathbb{E}\{Y\gamma^{\text{sgn}\{q_\ell(Y|a,x)-Y\}}|A=a,X=x\}\\ \leq \mathbb{E}\{Yv_0(Z)|A=a,X=x,S=0\}\leq\\ \mathbb{E}\{Y\gamma^{\text{sgn}\{Y-q_u(Y|a,x)\}}|A=a,X=x\}\equiv m_u(a,x)\end{aligned}$$

where now the quantile are those of the distribution of $Y|(A,X)$. Notice that these are the usual bounds in the non-contaminated model.

We can then compute the bounds on $\mathbb{E}\{Y(a)|X\}$. First notice that,

$$\begin{aligned}\mathbb{E}\{Y(a)S|X\} &= \mathbb{E}(Y|A=a,X,S=1)\mathbb{P}(S=1|X) && (Y(a)\perp\!\!\!\perp A|X,S=1)\\ &= \mu(a,X)\mathbb{P}(S=1|X). && (Y\perp\!\!\!\perp S|A,X)\end{aligned}$$

This means that $\mathbb{E}\{Y(a)S\}=\mathbb{E}\{S\mu(a,X)\}$. Next notice that

$$\begin{aligned}\mathbb{E}\{Y(a)(1-S)|X\} &= \mathbb{E}\{Y(a)|X,S=0\}\mathbb{P}(S=0|X)\\ &= \mathbb{E}\{Y\alpha(a,X,U,S=0)|A=a,X,S=0\}\mathbb{P}(S=0|X)\\ &= \mathbb{E}\{Yv_0(Z)|A=a,X,S=0\}\mathbb{P}(S=0|X).\end{aligned}$$

Therefore,

$$\mathbb{E}\{Y(a)|X\}=\mu(a,X)\mathbb{P}(S=1|X)+\mathbb{E}\{Yv_0(Z)|A=a,X,S=0\}\mathbb{P}(S=0|X)$$

so that

$$\mathbb{E}\{Y(a)\}=\mathbb{E}\left\{\mu(a,X)\right\}+\mathbb{E}\left[(1-S)\left\{\mathbb{E}\{Yv_0(Z)|A=a,X,S=0\}-\mu(a,X)\right\}\right]$$

which implies, for $r_j(a,X)=m_j(a,X)-\mu(a,X)$ and $j\in\{l,u\}$:

$$\mathbb{E}\{\mu(a,X)+(1-S)r_\ell(a,X)\}\leq\mathbb{E}\{Y(a)\}\leq\mathbb{E}\{\mu(a,X)+(1-S)r_u(a,X)\}.$$

Let $t_{\epsilon,l}(a)$ the ϵ-quantile of $r_\ell(a,X)$ and $t_{\epsilon,u}$ be the $(1-\epsilon)$-quantile of $r_u(a,X)$. Further, let $\lambda_\ell(a,x)=\mathbb{1}\{r_u(a,x)\leq t_{\epsilon,l}(a)\}$ and $\lambda_u(a,x)=\mathbb{1}\{r_u(a,x)>t_{\epsilon,u}(a)\}$. Under the

assumption that $\mathbb{P}(S=0)=\epsilon$, we bound $\mathbb{E}\{Y(a)\}$ by further optimizing over S as

$$\mathbb{E}\{Y(a)\} \geq \mathbb{E}\{\mu(a,X)\} + \mathbb{E}[\lambda_\ell(a,X) r_\ell(a,X)] \equiv \theta_\ell(a)$$
$$\mathbb{E}\{Y(a)\} \leq \mathbb{E}\{\mu(a,X)\} + \mathbb{E}[\lambda_u(a,X) r_u(a,X)] \equiv \theta_u(a).$$

As discussed in Section 3.4.2, one option to estimate the bounds is to assume that they follow some parametric models $g(a;\beta_\ell)$ and $g(a;\beta_u)$. Define

$$f_\mu(Z_1,Z_2) = W(A_1,X_1)\{Y_1 - \mu(A_1,X_1)\} + \mu(A_1,X_2)$$
$$f_{\Delta,j}(Z_1) = W(A_1,X_1)\left[\{s_j(Z;q_j) - \kappa(A_1,X_1;q_j)\} - Y_1 + \mu(A_1,X_1)\right]$$
$$f_{r,j}(Z_1,Z_2) = \kappa(A_1,X_2;q_j) - \mu(A_1,X_2)$$
$$f_j(Z_1,Z_2) = f_\mu(Z_1,Z_2) + \lambda_j(A_1,X_1) f_{\Delta,j}(Z_1) + \lambda_j(A_1,X_2) f_{r,j}(Z_1,Z_2).$$

Then, if it is assumed that $\theta_j(a) = g(a;\beta_j)$, the following moment condition holds

$$\mathbb{U}\left[h(A_1)\left\{f_j(Z_1,Z_2) - g(A_1;\beta_j)\right\}\right] = 0.$$

In this respect, we define $\widehat{\beta}_j$ to solve $\mathbb{U}_n\left[h(A_1)\left\{\widehat{f}_j(Z_1,Z_2) - g(A_1;\widehat{\beta}_j)\right\}\right] = 0$. We estimate the nuisance functions on a separate sample D^n independent from the sample Z^n used to evaluate the U-statistic. However, in the proof of the proposition below, we require that $\widehat{t}_{\epsilon,j}(a)$ satisfies

$$\frac{1}{n}\sum_{i\in Z^n} \mathbb{1}\{\widehat{r}_j(a,X_i) > \widehat{t}_{\epsilon,j}(a)\} = \epsilon + o_{\mathbb{P}}(n^{-1/2}) \text{ for all } a\in\mathcal{A} \text{ and } j=\{l,u\}.$$

In other words, we estimate all nuisance functions on a separate, training sample except for $a \mapsto t_{1-\epsilon}(a)$, which is estimated on the same sample used to estimate the moment condition. This helps with controlling the bias due to the presence of the indicator at the expense of an additional requirement on the complexity of the class where $a \mapsto \widehat{t}_{\epsilon,j}(a)$ belongs to. We have the following proposition.

Proposition 11. Suppose

1. The function class $\mathcal{G}_l = \{a \mapsto h_l(a) g(a;\beta)\}$ is Donsker for every $l=\{1,\dots,k\}$ with integrable envelop and $g(a;\beta)$ is a continuous function of β;
2. The map $\beta \mapsto \mathbb{U}\{h(A_1) f_j(Z_1,Z_2) - g(A_1;\beta)\}$ is differentiable at all β with continuosly invertible matrices $\dot{\Psi}_{\beta_0}$ and $\dot{\Psi}_{\widehat{\beta}}$, where $\dot{\Psi}_\beta = -\mathbb{E}\{h(A)\nabla^T g(A;\beta)\}$;
3. The function class $\mathcal{T}$ where $a \mapsto \widehat{t}_{\epsilon,j}(a)$ and $a \mapsto t_{\epsilon,j}(a)$ belong to is VC-subgraph;
4. For any $a\in\mathcal{A}$ and $x\in\mathcal{X}$, $r_j(a,X) - t_{\epsilon,j}(a)$, $r_j(A,x) - t_{\epsilon,j}(A)$ and $r_j(A,X) - t_{\epsilon,j}(A)$ have bounded densities;

5. The following holds

$$\|\widehat{q}_u - q_u\| = o_{\mathbb{P}}(n^{-1/4}),\ \|r_j - \widehat{r}_j\|_\infty + \left\|t_{\epsilon,j} - \widehat{t}_{\epsilon,j}\right\|_\infty = o_{\mathbb{P}}(n^{-1/4})$$
$$\left(\|r_j - \widehat{r}_j\|_\infty + \left\|t_{\epsilon,j} - \widehat{t}_{\epsilon,j}\right\|_\infty + \|\widehat{w} - w\|\right)\left(\|\widehat{\mu} - \mu\| + \|\kappa_j - \widehat{\kappa}_j\|\right) = o_{\mathbb{P}}(n^{-1/2}).$$

Then,

$$\sqrt{n}(\widehat{\beta}_j - \beta_j) \rightsquigarrow N(0, 4\Sigma)$$

where $\Sigma = \text{var}\left[\dot{\Psi}_{\beta_j}^{-1} \int S_2 h(A_1)\left\{f_j(Z_1, z_2) - t_{\epsilon,j}(A_1)\lambda_j(A_1, x_2) - g(A_1; \beta_j)\right\} d\mathbb{P}(z_2)\right]$.

To get bounds on some coordinate of β, say β_1, one may proceed by homotopy as in the non-contaminated model. In the linear MSM case, i.e. $g(a;\beta) = b(a)^T\beta$, bounds on β_1 that enforce the restriction $\mathbb{E}\{v_0(Z)|A, X, S = 0\} = 1$ would be

$$l_\gamma = \int \min\left\{e^T M^{-1} b(a)\theta_u(a), e^T M^{-1} b(a)\theta_\ell(a)\right\} d\mathbb{P}(a)$$
$$u_\gamma = \int \max\left\{e^T M^{-1} b(a)\theta_u(a), e^T M^{-1} b(a)\theta_\ell(a)\right\} d\mathbb{P}(a)$$

where $M = \mathbb{E}\{b(A)b(A)^T\}$ and we set $h(A) = b(A)$. A similar statement to Proposition 3 can be derived using the influence function established in proving Proposition 11.

Remark: *If we make the stronger assumption that S is independent of (X, A, Y) then $p_0(x, a, y) = p_1(x, a, y)$. In this case it is easy to see that $\mathbb{E}[h(A)(Y - g(A, \beta))w(A, X)((1 - \epsilon) + \epsilon v(Z))] = 0$. All the previous methods can then be used with v replaced with $(1-\epsilon)+\epsilon v(Z)$.*

Now we use the outcome sensitivity model on the confounded subpopulation. We will assume that $S \perp\!\!\!\perp Z$. The distribution is

$$(1 - \epsilon)p(u, x, a)p(y|x, a) + \epsilon p(u, x, a)p(y|u, x, a).$$

The moment condition is

$$
\begin{aligned}
0 &= \int b(a)(y - b^T(a)\beta)w(u,x,a)dP(u,x,a) \\
&= (1-\epsilon)\int b(a)(y - b^T(a)\beta)w(u,x,a)p(u,x,a)p(y|x,a) \\
&\quad + \epsilon \int b(a)(y - b^T(a)\beta)w(u,x,a)p(u,x,a)p(y|u,x,a) \\
&= (1-\epsilon)\int b(a)(y - b^T(a)\beta)w(x,a)p(x,a)p(y|x,a) \\
&\quad + \epsilon \int b(a)(y - b^T(a)\beta)w(u,x,a)p(u,x,a)p(y|u,x,a) \\
&= (1-\epsilon)\int b(a)(y - b^T(a)\beta)w(x,a)p(x,a)p(y|x,a) \\
&\quad + \epsilon \int b(a)(y - \mu(x,a))w(u,x,a)p(u,x,a)p(y|u,x,a) \\
&\quad + \epsilon \int b(a)(\mu(x,a) - b^T(a)\beta)w(u,x,a)p(u,x,a)p(y|u,x,a) \\
&= (1-\epsilon)\int b(a)(y - b^T(a)\beta)w(x,a)p(x,a)p(y|x,a) \\
&\quad + \epsilon \underbrace{\int b(a)(\mu(u,x,a) - \mu(x,a))w(u,x,a)p(u,x,a)}_{\Xi} \\
&\quad + \epsilon \int b(a)(\mu(x,a) - b^T(a)\beta)w(x,a)p(x,a) \\
&= (1-\epsilon)\int b(a)(y - b^T(a)\beta)w(x,a)p(x,a)p(y|x,a) \\
&\quad + \epsilon \int b(a)(\mu(x,a) - b^T(a)\beta)w(x,a)p(x,a)p(y|x,a) + \epsilon\Xi \\
&= \mathbb{E}\left[b(A)\Big((1-\epsilon)Y + \epsilon\mu(X,A)\Big)w(X,A)\right] - \underbrace{\Big(\mathbb{E}[b(A)b^T(A)w(X,A)]\Big)}_{\Omega}\beta + \epsilon\Xi,
\end{aligned}
$$

where $\Xi = \int b(a)(\mu(u,x,a) - \mu(x,a))w(u,x,a)p(u,x,a)$ and $\Omega = \mathbb{E}[b(A)b^T(A)w(X,A)]$. Therefore

$$
\beta = \Omega^{-1}\mathbb{E}\left[b(A)\Big((1-\epsilon)Y + \epsilon\mu(X,A)\Big)w(X,A)\right] + \epsilon\Omega^{-1}\Xi
$$

and $\beta_1 = e^T\beta$, where $e = (1,0,\ldots,0)$. Let r be the first row of Ω^{-1} and let $f(a) = \sum_j r_j b_j(a)$.

Then

$$\begin{aligned} e^T\Omega^{-1}\Xi = r^T\Xi &= \int (\sum_j r_j b_j(a))(\mu(u,x,a) - \mu(x,a))w(u,x,a)p(u,x,a) \\ &\leq \delta \int f(a)I(f(a) > 0)\pi(a) - \delta \int f(a)I(f(a) < 0)\pi(a) \\ &= \delta \int f(a)(2I(f(a) > 0) - 1)\pi(a). \end{aligned}$$

Similarly,

$$\begin{aligned} e^T\Omega^{-1}\Xi \geq \delta \int f(a)I(f(a) < 0)\pi(a) - \delta \int f(a)I(f(a) > 0)\pi(a) \\ - \delta \int f(a)(2I(f(a) > 0) - 1)\pi(a). \end{aligned}$$

Therefore,

$$\beta_1^* - \delta \int f(a)(2I(f(a) > 0) - 1)\pi(a) \leq \beta_1 \leq \beta_1^* + \delta \int f(a)(2I(f(a) > 0) - 1)\pi(a)$$

where

$$\beta_1^* = \mathbb{E}\left[b(A)\Big((1-\epsilon)Y + \epsilon\mu(X,A)\Big)w(X,A)\right].$$

B.0.3 Bounds for β under the outcome sensitivity confounding model when the MSM is not linear

Say the MSM is not linear. Since $g(a;\beta) = \mathbb{E}\{Y(a)\} = \int\int yp(y|u,x,a)dP(x,u)$, we have

$$\begin{aligned} 0 &= \int\int\int h(a)(y - g(a;\beta))p(y|u,x,a)\pi(a)dydP(u,x) \\ &= \int\int h(a)(\mu(u,x,a) - g(a;\beta))\pi(a)dP(u,x) \\ &= \int\int h(a)(\mu(u,x,a) - \mu(x,a))\pi(a)dP(u,x) + \int\int h(a)(\mu(x,a) - g(a;\beta))\pi(a)dP(u,x) \\ &= \int\int h(a)(\mu(u,x,a) - \mu(x,a))\pi(a)dP(u,x) + \int\int h(a)(\mu(x,a) - g(a;\beta))\pi(a)dP(x) \\ &= \int\int h(a)(\mu(u,x,a) - \mu(x,a))\pi(a)dP(u,x) \\ &\quad + \int\int \frac{h(a)(\mu(x,a) - g(a;\beta))\pi(a)}{\pi(a|x)}\pi(a|x)dP(x) \\ &= \int h(a)\xi(a)\pi(a)da + \mathbb{E}[h(A)(\mu(X,A) - g(A;\beta))w(A,X)]. \end{aligned}$$

Let $C = \{\mathbb{E}[h_1(A)\xi(A)], \ldots, \mathbb{E}[h_k(A)\xi(A)] : \ -\delta \leq \xi(a) \leq \delta\}$. For each vector $t \in C$, let $\beta(t)$ solve $\mathbb{E}[h(A)(\mu(X, A) - g(A; \beta))w(A, X)] = t$. Then

$$\inf_{t \in C} e^T \beta(t) \leq \beta_j \leq \sup_{t \in C} e^T \beta(t).$$

These bounds can be found numerically by solving for $\beta(t)$ over a grid on C.

B.1 Algorithms

B.1.1 Homotopy Algorithm

Input: grid $\{\gamma_1, \ldots, \gamma_N\}$ where $\gamma_1 = 1$ and $\gamma_1 < \cdots < \gamma_N$.

1. Let $\widehat{\beta}$ be the solution of $\sum_i h(A_i)(Y_i - g(A_i; \widehat{\beta}))\widehat{W}_i = 0$. Let $u_1 = \ell_1 = e^T \widehat{\beta}$.
2. For $j = 2, \ldots, N$:
   (a) Let $d_{j,i} \equiv d_{\gamma_{j-1},i}$ from (3.14) evaluated at $v = v_{\gamma_j - 1}$.
   (b) Set $V_i = \gamma_j^{-1} I(d_{j,i} \leq q) + \gamma_j I(d_{j,i} > q)$ where q is the $\gamma_j/(1 + \gamma_j)$ quantile of $d_{j,1}, \ldots, d_{j,n}$. Let $\widehat{\beta}$ be the solution of $\sum_i h(A_i)(Y_i - g(A_i; \widehat{\beta}))\widehat{W}_i V_i = 0$. Set $u_j = e^T \widehat{\beta}$.
   (c) Set $V_i = \gamma_j I(d_{j,i} \leq q) + \gamma_j^{-1} I(d_{j,i} > q)$ where q is the $1/(1 + \gamma_j)$ quantile of $d_{j,1}, \ldots, d_{j,n}$. Let $\widehat{\beta}$ be the solution of $\sum_i h(A_i)(Y_i - g(A_i; \widehat{\beta}))\widehat{W}_i V_i = 0$. Set $\ell_j = e^T \widehat{\beta}$.
3. Return $(\ell_1, u_1), \ldots, (\ell_N, u_N)$.

B.1.2 Bounds on β by Coordinate Ascent

Another approach we consider is coordinate ascent where we maximize (or minimize) $\widehat{\beta}_1(v)$ over each coordinate v_i in turn. It turns out that this is quite easy since $\widehat{\beta}_1(v)$ is strictly monotonic in each v_i for many models so we need only compare the estimate at the two values $v_i = \gamma$ and $v_i = 1/\gamma$. Furthermore, in the linear case, getting the estimate after changing one coordinate v_i can be done quickly using a Sherman-Morrison rank one update.

The coordinate ascent approach will lead to a local optimum but it will depend on the ordering of the data so we repeat the algorithm using several random orderings. The homotopy method instead uses the last solution as a starting point for the new solution. This makes the homotopy method faster but, in principle, the coordinate ascent approach could explore a wider set of possible solutions. For simplicity, the only restriction we enforce is $1/\gamma \leq v_i \leq \gamma$. In practice, we find that the solutions are very similar.

Lemma 17. *Suppose that the function $\widehat{\beta}(v)$ is strictly monotonic in each coordinate v_i. The*

maximizer and minimizer occur at corners of the cube $[1/\gamma, \gamma]^n$. *We have that*

$$\frac{\partial \widehat{\beta}_1(v)}{\partial v_j} = \frac{1}{W_i} e^T \Big\{ (X^T \mathbb{W} X)^{-1} [S_i - (R_i R_i^T)\widehat{\beta}] \Big\}$$

where $W_i = 1/\pi(A_i|X_i)$, $\mathbb{W}$ *is diagonal with* $\mathbb{W}_{ii} = W_i$, $R_i = (X_{i1}, \ldots, X_{id})^T$ *and* $S_i = R_i Y_i$. *Also,*

$$H_{ij} \equiv \frac{\partial^2 \widehat{\beta}_1}{\partial v_i \partial v_j} = -e^T (X^T \mathbb{W} X)^{-1} \left\{ (R_i R_i^T) \frac{\partial \widehat{\beta}}{\partial v_i} + (R_j R_j^T) \frac{\partial \widehat{\beta}}{\partial v_j} \right\}.$$

The proof is straightforward and is omitted.

Coordinate Ascent

1. Input: Data (B, A, Y), where B is the $n \times k$ matrix with elements $B_{ij} = b_j(A_i)$, weights $W_i = 1/\pi(A_i|X_i)$ and grid $\{\gamma_1, \ldots, \gamma_N\}$ with $\gamma_1 = 1$.
2. Let $\widehat{\beta} = (B^T \mathbb{W} B)^{-1} B^T \mathbb{W} Y$ where $\mathbb{W}$ is diagonal with $\mathbb{W}_{ii} = W_i$. Set $\overline{\beta}_1(1) = \underline{\beta}_1(1) = \widehat{\beta}_1$.
3. Now move $v_i = 1$ to $v_i = \gamma_2$ or $v_i = 1/\gamma_2$, whichever makes $\widehat{\beta}_1$ larger:
   (a) Let $G = (B^T \mathbb{W} B)$. For each i let

$$u_i = e^T \left(G_i^{-1} - \frac{\Delta_i G_i^{-1} r_i r_i^T G^{-1}}{1 + \Delta_i r_i^T A_i^{-1} r_i} \right) (B^T \mathbb{W} + \Delta_i e_i e_i^T) Y \quad \text{flip 1 to } \gamma_2$$

$$\ell_i = e^T \left(G_i^{-1} - \frac{\delta_i G^{-1} r_i r_i^T G^{-1}}{1 + \delta_i r_i^T G^{-1} r_i} \right) (B^T \mathbb{W} + \delta_i e_i e_i^T) Y \quad \text{flip 1 to } 1/\gamma_2$$

   where $\Delta_i = \gamma_2 - 1$ and $\delta_i = \frac{1}{\gamma_2} - 1$.
   (b) If $u_i \geq \ell_i$: set $v_i = \gamma_2$ and $I_i = 1$. Else, set $v_i = 1/\gamma_2$ and $I_i = 0$.
4. For $j = 3, \ldots, N$: Try flipping each v_i to $1/alpha_i$.
   (a) Let $v_i = \gamma_j I_i + \gamma_j^{-1}(1 - I_i)$.
   (b) Let $\mathbb{W}_{ii} = v_i$ and $\widehat{\beta} = (B^T \mathbb{W} B)^{-1} B^T \mathbb{W} Y$.
   (c) Let $A = (B^T \mathbb{W} B)$,

$$t_i = e^T \left(A_i^{-1} - \frac{\Delta_i A_i^{-1} r_i r_i^T A_i^{-1}}{1 + \Delta_i r_i^T A_i^{-1} r_i} \right) (B^T \mathbb{W} + \Delta_i e_i e_i^T) Y.$$

   where $\Delta_i = 1/v_i - v_i$. If $t_i > \widehat{\beta}_1$ let $v_i = 1/v_i$. Let $\mathbb{W}_{ii} = v_i$. Let $\widehat{\beta} = (B^T \mathbb{W} B)^{-1} B^T \mathbb{W} Y$. Let $\overline{\beta}_1(\gamma_j) = \widehat{\beta}_1$.

B.2 Technical proofs

B.2.1 Proof of Proposition 1

Let $\alpha(u,x,a) = \pi(a|x)/\pi(a|x,u)$. Recall that $v(X,A,Y) = \mathbb{E}\{\alpha(U,X,A) \mid X,A,Y\}$ and $w(A,X) = \pi(A)/\pi(A \mid X)$. We have

$$\begin{aligned}
0 &= \mathbb{E}\left[h(A)w(A,X)\{Y - g(A;\beta)\}\alpha(U,X,A)\right] \\
&= \mathbb{E}\left[h(A)w(A,X)\{Y - g(A;\beta)\}v(X,A,Y)\right] \\
&= \mathbb{E}\left\{h(A)Yv(X,A,Y) - h(A)w(A,X)g(A;\beta)v(X,A,Y)\right\} \\
&= \mathbb{E}\left[h(A)w(A,X)\mathbb{E}\{Yv(X,A,Y)|X,A\} - h(A)w(A,X)g(A;\beta)\mathbb{E}\{v(X,A,Y)|X,A\}\right] \\
&= \mathbb{E}\left\{h(A)w(A,X)m(X,A) - h(A)w(A,X)g(A;\beta)\right\} \\
&= \int\int \{h(a)m(x,a)d\mathbb{P}(x) - h(a)g(a;\beta)\}d\mathbb{P}(x)\pi(a)da \\
&= \mathbb{E}\left[h(A)\left\{\int m(A,x)d\mathbb{P}(x) - g(A;\beta)\right\}\right]. \ \square
\end{aligned}$$

B.2.2 Proof of Lemma 2

We will prove the result for the upper bound. The proof for the lower bound follows analogously. We have $v_u(Z) \in [\gamma^{-1},\gamma]$ and we can check that $\mathbb{E}\{v_u(Z)|A,X\} = 1$. Indeed

$$\begin{aligned}
\mathbb{E}\{v_u(Z)|A,X\} &= \gamma\mathbb{P}\left(Y > q_u(Y|A,X)|A,X\right) + \frac{1}{\gamma}\mathbb{P}\left(Y \leq q_u(Y|A,X)|A,X\right) \\
&= \gamma\left(1 - \frac{\gamma}{1+\gamma}\right) + \frac{1}{\gamma}\cdot\frac{\gamma}{1+\gamma} = 1,
\end{aligned}$$

because $q_u(A,X)$ is the $\gamma/(1+\gamma)$-quantile of the conditional distribution of Y given (A,X). Let $v(Z)$ be any function contained in $[\gamma^{-1},\gamma]$ such that $\mathbb{E}\{v(Z)|A,X\} = 1$. We have

$$\begin{cases}
v_u(Z) - v(Z) \geq 0 & \text{if } Y > q_u(Y|A,X) \\
v_u(Z) - v(Z) \leq 0 & \text{if } Y \leq q_u(Y|A,X)
\end{cases}$$

Therefore, $Y\{v_u(Z) - v(Z)\} \geq q_u(Y|A,X)\{v_u(Z) - v(Z)\}$ so that

$$\mathbb{E}\left\{Y\{v_u(Z) - v(Z)\}|A,X\right\} \geq q_u(Y|A,X)\mathbb{E}\{v_u(Z) - v(Z)|A,X\} = 0$$

as desired.

B.2.3 Proof of Proposition 2

This proposition follows directly from Lemma 23, except that we need to show the validity of condition 4. This condition holds under the assumption of Proposition 2 because

$$\mathbb{U}\left[h_l(A_1)\{\widehat{\varphi}_j(Z_1,Z_2)-\varphi_j(Z_1,Z_2)\}\right]=\int h_l(a)\widehat{w}(a,x)\{\kappa(a,x;\widehat{q}_j)-\kappa(a,x;q_j)\}d\mathbb{P}(a,x)$$
$$+\int\{w(a,x)-\widehat{w}(a,x)\}\{\widehat{\kappa}(a,x;\widehat{q}_j)-\kappa(a,x;q_j)\}d\mathbb{P}(a,x)$$

Therefore, by Cauchy-Schwarz and Lemma 21:

$$\left|\mathbb{U}\left[h_l(A_1)\{\widehat{\varphi}_j(Z_1,Z_2)-\varphi_j(Z_1,Z_2)\}\right]\right|\lesssim\|q_j-\widehat{q}_j\|^2+\|w-\widehat{w}\|\|\widehat{\kappa}_j-\kappa_j\|.$$

B.2.4 Proof of Proposition 3

We apply Lemma 23 to the moment condition

$$\Psi_n(\beta)=\mathbb{U}_n\left[b(A)\left\{\widehat{f}_j^s(Z_1,Z_2)-b(A_1)^T\beta\right\}\right]=o_{\mathbb{P}}(n^{-1/2})$$

The function class $\mathcal{G}_l=\left\{a\mapsto b_l(a)b(a)^T\beta,\beta\in\mathbb{R}^k\right\}$ is Donsker since its a finite dimensional vector space (Lemma 7.15 in Sen [2018]). Thus, it remains to check condition 4. We have

$$\left|\mathbb{U}\left\{\widehat{f}_j^s(Z_1,Z_2)-f_j^s(Z_1,Z_2)\right\}\right|\lesssim\|q_j-\widehat{q}_j\|^2+\|w-\widehat{w}\|\|\kappa_j-\widehat{\kappa}_j\|+\sup_a\left|b_0^T(\widehat{Q}-Q)h(a)\right|^2$$
$$=o_{\mathbb{P}}(n^{-1/2})$$

by assumption and because the last term is $O_{\mathbb{P}}(n^{-1})=o_{\mathbb{P}}(n^{-1/2})$ by Lemma 19.

B.2.5 Proof of Lemma 3

For $v\in\mathcal{V}_{\text{small}}(\gamma)$ we have

$$\int\int\int h(a)w(a,x)g(a;b)v(z)dP(z)=\int\int h(a)w(a,x)g(a;b)[\int v(z)p(y|x,a)]\pi(a|x)dP(x)$$
$$=\int\int h(a)w(a,x)g(a;b)\pi(a|x)dP(x)$$
$$=\int\int h(a)w(a,x)g(a;b)dP(z)$$

since $\mathbb{E}[v(Z)|X,A]=1$. Therefore $F_1=F_2$ and the result follows.

B.2.6 Proof of Lemma 4

Let $Z = (A, X, Y)$ and $p(z)$ denote its density. From the moment condition, we have that $F_2(v)$ satisfies

$$\int h(a)w(a,x)v(z)yd\mathbb{P}(z) = \int h(a)g(a; F_2(v))w(a,x)d\mathbb{P}(z)$$

Let $m : \mathcal{F} \mapsto \mathbb{R}$ be a generic functional taking as input a function f. The functional derivative of m with respect to $f(z)$, denoted $\frac{\delta}{\delta f}m$, satisfies

$$\frac{d}{d\epsilon}m(f + \epsilon\eta)|_{\epsilon=0} = \int \frac{\delta}{\delta f}m(z)\eta(z)d\mathbb{P}(z)$$

for any function η. Letting

$$\nabla_\beta g(A:\beta) = \begin{bmatrix} \frac{d}{d\beta_1}g(A;\beta) \\ \vdots \\ \frac{d}{d\beta_k}g(A;\beta) \end{bmatrix}, \quad \frac{d}{d\epsilon}F_2(v+\epsilon\eta) = \begin{bmatrix} \frac{d}{d\epsilon}F_{2,1}(v+\epsilon\eta) \\ \vdots \\ \frac{d}{d\epsilon}F_{2,k}(v+\epsilon\eta) \end{bmatrix},$$
$$\text{and} \quad \frac{\delta}{\delta v}F_2(v) = \begin{bmatrix} \frac{\delta}{\delta v}F_{2,1}(v) \\ \vdots \\ \frac{\delta}{\delta v}F_{2,k}(v) \end{bmatrix},$$

and taking the functional derivative with respect to $v(z)$ on both sides of the expression above yields

$$\begin{aligned}
&\frac{d}{d\epsilon}\int h(a)w(a,x)\{v(z)+\epsilon\eta(z)\}yp(z)dz|_{\epsilon=0} = \int h(a)w(a,x)y\eta(z)p(z)dz \\
&\implies \frac{\delta}{\delta v}\int h(a)w(a,x)v(z)yp(z)dz = h(a)w(a,x)y \\
&\frac{d}{d\epsilon}\int h(a)w(a,x)g(a;\beta(v+\epsilon\eta))p(z)dz|_{\epsilon=0} \\
&= \int h(a)w(a,x)\nabla_\beta g(a;\beta)^T p(z)dz\frac{d}{d\epsilon}\beta(v+\epsilon\eta)|_{\epsilon=0} \\
&\implies \frac{\delta}{\delta v}\int h(a)w(a,x)g(a;\beta(v))p(z)dz = \mathbb{E}\left\{h(A)w(A,X)\nabla_\beta g(A;\beta)^T\right\}\frac{\delta\beta(v)}{\delta v}
\end{aligned}$$

Thus, we conclude that the functional derivative of $\beta(v)$ with respect to v satisfies

$$\frac{\delta F_2(v)}{\delta v} = \mathbb{E}\left\{h(A)w(A,X)\nabla_\beta g(A;\beta)^T\right\}^{-1}h(a)w(a,x)y$$

as desired. A similar calculation yields $\frac{\delta F_1(v)}{\delta v}$.

B.2.7 Proof of Lemma 5

Property 1: This is clear.
Property 2: Define a map $F : \mathcal{V}(\gamma) \to \mathcal{V}(\gamma)$ by

$$F(v) = \gamma I(d_v > q_v) + \frac{1}{\gamma} I(d_v < q_v)$$

where $d_v(z) = \delta\beta/\delta v(z)$ and q_v is the $\gamma/(1+\gamma)$ quantile of $d_v(Z)$. We want to show that there is a fixed point $v = L(v)$. Define the metric m by $m(v_1, v_2) = \sqrt{\int (v_1(z) - v_2(z))^2 d\mathbb{P}(z)}$. The set of functions $\mathcal{V}(\gamma)$ is a nonempty, closed, convex set. It is easy to see that $L : \mathcal{V}(\gamma) \to \mathcal{V}(\gamma)$ is continuous, that is, $m(v_n, v) \to 0$ implies $L(v_n) \to L(v)$. According to Schauder's fixed point theorem there exists a fixed point v_γ so that $L(v_\gamma) = v_\gamma$.
Property 3: Let $v \in \mathcal{V}(\gamma) \bigcap B(v_\gamma, \epsilon)$. Then $\beta(v) = \beta(v_\gamma) + \int (v(z) - v_\gamma(z)) d_\gamma(z) d\mathbb{P}(z) + O(\epsilon^2)$. The linear functional $\int (v(z) - v_\gamma(z)) d_\gamma(z) d\mathbb{P}(z)$ is maximized over $\mathcal{V}(\gamma)$ by choosing $v = \gamma I(d_\gamma(z) > t) + \gamma^{-1} I(d_\gamma(z) < t)$. The condition $\int v(z) d\mathbb{P}(z) = 1$ implies that $t = q$. So $\int (v(z) - v_\gamma(z)) d_\gamma(z) d\mathbb{P}(z)$ is maximized by $v = v_\gamma$ and hence $\int (v(z) - v_\gamma(z)) d_\gamma(z) d\mathbb{P}(z) \leq 0$. Thus $\beta(v_\gamma) \geq \beta(v) + O(\epsilon^2)$. □

B.2.8 Proof of Lemma 6.

The fact that F_1 and F_2 yield the same bounds follows from Lemma 6. Now $F_2(v) = \int v(z) q(z) dP(z)$ where $q(z) = yw(a, x) M^{-1} b(a)$ which is a linear functional. The form of the maximizer amd minimizer follows by the same argument as in the proof of Lemma 2.

B.2.9 Proof of Lemma 7.

Since $F_2(v)$ is a linear functional of v, The form of the maximizer and minimizer follows by the same argument as in the proof of Lemma 2.

B.2.10 Proof of Lemma 4

Consider the upper bound. We apply Lemma 23 to the moment condition

$$\Psi_n(\widehat{\beta}) = \mathbb{U}_n \left[b(A_1) \left\{ \widehat{\zeta}_u(Z_1, Z_2) - b(A_1)^T \widehat{\beta} \right\} \right] = o_{\mathbb{P}}(n^{-1/2})$$

where $\zeta_u(Z_1, Z_2) = w(A_1, X_1) \{Y_1 - \mu(A_1, X_1)\} + \mu(A_1, X_2) + \delta \operatorname{sgn} \left\{ b(a_0)^T Q^{-1} b(A_1) \right\}$ and a_0 is a fixed value of a that we want to distinguish from the dummy a in the function class $\mathcal{G}_l = \left\{ a \mapsto b_l(a) b(a)^T \beta, \beta \in \mathbb{R}^k \right\}$. Notice that $\mathcal{G}_l$ is Donsker and we have $\dot{\Psi}_{\beta_0} = Q = \mathbb{E}\{b(A) b^T(A)\}$. Next notice that, by virtue of the statement of Lemma 23:

$$\begin{aligned} \widehat{g}_u(a_0) - g_u(a_0) &= b(a_0)^T (\widehat{\beta} - \beta_u) = b(a_0)^T Q^{-1} \mathbb{U} b(A_1) \left\{ \widehat{\zeta}_u(Z_1, Z_2) - \zeta_u(Z_1, Z_2) \right\} \\ &+ b(a_0)^T Q^{-1} (\mathbb{U}_n - \mathbb{U}) b(A_1) \left\{ \zeta_u(Z_1, Z_2) - b^T(A_1) \beta_u \right\} + o_{\mathbb{P}}(n^{-1/2}). \end{aligned}$$

Next, we have

$$\begin{aligned}
&\left|\mathbb{U}\left\{b(a_0)^T Q^{-1} b(A_1)\widehat{\zeta}_u(Z_1,Z_2) - \zeta_u(Z_1,Z_2)\right\}\right| \\
&\lesssim \sup_a |b(a_0)^T Q^{-1} b(a)| \|w-\widehat{w}\| \|\mu - \widehat{\mu}\| \\
&\quad + \left|\mathbb{P}\left(b(a_0)^T Q^{-1} b(A)\left[\operatorname{sgn}\left\{b(a_0)^T \widehat{Q}^{-1} b(A)\right\} - \operatorname{sgn}\left\{b(a_0)^T Q^{-1} b(A)\right\}\right]\right)\right|.
\end{aligned}$$

By assumption the first term is $o_{\mathbb{P}}(n^{-1/2})$. By Lemma 22, the last term is upper bounded by a constant multiple of

$$\sup_a \left|b(a_0)^T(\widehat{Q}^{-1} - Q^{-1})b(a)\right|^2$$

which is $O_{\mathbb{P}}(n^{-1})$ by Lemma 19.

B.2.11 Influence Function for $\beta(v_\gamma)$

The parameter is $\psi = \beta(v)$ where v is given by the fixed point equation

$$v(z) = \gamma - \left(\gamma - \gamma^{-1}\right) I(d(z) - q < 0).$$

Now v is a function of p and z and d is a function of v and z so we will write $v = v(p,z)$ and $d = d(v(p), z)$ and

$$v(p,s) = \gamma - \left(\gamma - \gamma^{-1}\right) I(d(v(p),z) - q < 0).$$

The influence function is not well-defined beacause of the presence of the indicator function. So we approximate v by

$$v(p,z) = \gamma - \left(\gamma - \gamma^{-1}\right) S(d(v(p),z) - q)$$

where S is any smooth approximation to the indicator function. In general, the influence function $\varphi(z)$ of a parameter ψ is relate to the $L_2(P)$ functional derivative by $\varphi(z) = (1/p(z))\delta\psi(z)/\delta p$. We then have

$$\frac{\delta\beta(v(p))}{\delta p} = \int \frac{\delta\beta(v(p))}{\delta v(p,s)} \frac{\delta v(p,z)}{\delta p} d\mathbb{P}(z) = \int d_\gamma(z) \frac{\delta v(p,z)}{\delta p} d\mathbb{P}(z).$$

Now

$$\begin{aligned}
\frac{\delta v(p,z)}{\delta p}(Z) &= -\left(\gamma - \gamma^{-1}\right) S'(d(v(p),z) - q)\left(\frac{\delta d(v(p),z)}{\delta p}(Z) - \frac{\delta q}{\delta p}(Z)\right) \\
&= -\left(\gamma - \gamma^{-1}\right) S'(d(v(p),z) - q)\left(\int \frac{\delta d(v(p),z)}{\delta v(p,t)}(Z) \frac{\delta v(p,t)}{\delta p}(Z) d\mathbb{P}(t) - \frac{\delta q}{\delta p}(Z)\right)
\end{aligned}$$

Note that $\delta v/\delta p$ appears on both sides and so the influence function involves solving an integral equation.

We still need to find $\delta d(v(p),z)/\delta v(p,t)(Z)$ and $\frac{\delta q}{\delta p}(Z)$. We may write the formula for $d(z)$ as

$$d_\gamma(z)\int h(a)g'(a,\beta)r(x,a,y)=h(a)yw(a,x)v(z)$$

where $r=p(x)\pi(a)p(y|x,a)$ and $W=\pi(a)/\pi(a|x)$. Note that

$$\mathring{r}=\delta_a p(x)p(y|x,a)+\delta_x\pi(a)p(y|x,a)+\pi(a)\delta_{xa}\frac{\delta_y-p(y|x,a)}{\pi(a|x)}-p(x)\pi(a)p(y|x,a)$$

and

$$\mathring{W}=\frac{\delta_a p(x)+\delta_x\pi(a)-W\delta_{xa}}{p(x)\pi(a|x)}-W$$

where $\mathring{r}$ means the influence function of r etc. So

$$\mathring{d}\int h(a)g'r+d(z)\int h(a)\mathring{g}'r+d(z)\int h(a)g'\mathring{r}=h(a)y\mathring{W}v(z)$$

and therefore

$$\mathring{d}=(\int h(a)g'r)^{-1}h(a)y\mathring{W}v(z)-d(z)\int h(a)\mathring{g}'r-d(z)\int h(a)g'\mathring{r}\quad\text{and}$$
$$\frac{\delta d(v(p),z)}{\delta v(p,t)}(Z)=\frac{\mathring{d}(Z)}{p(Z)}.$$

To find $\mathring{q}$ note that $F(q,p)=\gamma/(1+\gamma)$ where $F(t,p)=P(d(Z)\le t)$. So $f(q)\mathring{q}+\mathring{F}=0$, which implies $\mathring{q}=-\mathring{F}/f(q)$. Now

$$F(t,p)=\int I(d(z,p)\le t)p(z)dz\ \text{ and }\ \mathring{F}(t,p)=I(d_\gamma(z)\le t)-\int I(d_\gamma(z)=t)\mathring{d}_\gamma(z)p(z)dz,$$

so that $\mathring{F}(q,p)=I(d_\gamma(z)\le q)-\int I(d_\gamma(z)=q)\mathring{d}_\gamma(z)p(z)dz$. Hence

$$\mathring{q}=-\frac{I(d_\gamma(z)\le q)-\int I(d_\gamma(z)=q)\mathring{d}_\gamma(z)p(z)dz}{f(q)}\quad\text{and}\quad\frac{\delta q}{\delta p}(Z)=\frac{\mathring{q}(Z)}{p(Z)}.$$

Finally,

$$\mathring{g}'(a,\beta)=p(z)\frac{\delta g'(a,\beta)}{\delta p}=p(z)\int\frac{\delta g'(a,\beta)}{\delta v_\gamma}\frac{\delta v_\gamma}{\delta p}=p(z)\int d_\gamma(z)\mathring{v}(z)dz.$$

B.2.12 Proof of Proposition 11

We will prove the proposition in two steps:

1. We show that Lemma 23 yields that

$$\widetilde{\beta}_j - \beta_j = -\dot{\Psi}_{\beta_j}^{-1}(\mathbb{U}_n - \mathbb{U})h(A_1)\{f(Z_1, Z_2) - g(A_1; \beta_j)\} + o_{\mathbb{P}}(n^{-1/2})$$

   where $\widetilde{\beta}_j$ solves:

$$\mathbb{U}_n h(A_1)\left[\widetilde{f}(Z_1, Z_2) - g(A_1; \widetilde{\beta})\right] = o_{\mathbb{P}}(n^{-1/2}), \text{ where}$$
$$\widetilde{f}(Z_1, Z_2) = \widehat{f}_\mu(Z_1, Z_2) + \lambda_j(A_1, X_1)\widehat{f}_\Delta(Z_1) + \lambda(A_1, X_2)\widehat{f}_r(Z_1, Z_2).$$

   That is, $\widetilde{\beta}$ solves the original moment condition except that the estimator of the indicator term is replaced with the true indicator , e.g. $\widehat{\lambda}_u(a, x) = \mathbb{1}\{\widehat{r}_u(a, x) > \widehat{t}_{\epsilon,u}\}$ is replaced by $\lambda_u(a, x) = \mathbb{1}\{r_u(a, x) > t_{\epsilon,u}\}$.

2. We show that

$$\widehat{\beta}_j - \widetilde{\beta}_j = -\dot{\Psi}_{\widehat{\beta}_j}^{-1}(\mathbb{U}_n - \mathbb{U})h(A_1)t_{\epsilon,j}(A_1)\lambda_j(A_1, X_2) + o_{\mathbb{P}}(n^{-1/2}).$$

From these statements, it follows by Lemma 18 and Slutsky's theorem, that

$$\sqrt{n}(\widehat{\beta}_j - \beta_j) \rightsquigarrow N(0, 4\Sigma)$$

because, by the continuous mapping theorem, $\dot{\Psi}_{\widehat{\beta}}^{-1} \xrightarrow{p} \dot{\Psi}_{\beta_j}^{-1}$ since $\widehat{\beta} \xrightarrow{p} \beta_j$.

Step 1

Because $\widetilde{f}(Z_1, Z_2)$ is fixed given the training sample, we can apply Lemma 23. In particular, all the conditions of the lemma are satisfied by assumption and by noticing that

$$\left|\mathbb{U}\left\{\widehat{f}_\mu(Z_1, Z_2) - f_\mu(Z_1, Z_2)\right\}\right| \lesssim \|w - \widehat{w}\|\|\mu - \widehat{\mu}\|$$

and

$$\left|\mathbb{U}\left[\lambda_j(A_1, X_1)\widehat{f}_\Delta(Z_1) + \lambda(A_1, X_2)\left\{\widehat{f}_r(Z_1, Z_2) - f_r(Z_1, Z_2)\right\}\right]\right|$$
$$\lesssim \|w - \widehat{w}\|\left(\|\kappa_j - \widehat{\kappa}_j\| + \|\mu - \widehat{\mu}\|\right) + \|q_j - \widehat{q}_j\|^2.$$

Therefore, condition 4 in Lemma 23 is satisfied as well under the assumption that the nuisance functions are estimated with enough accuracy.

Step 2

Define $\widetilde{\lambda}_\ell(a, x) = \mathbb{1}\{\widehat{r}_\ell(a, x) \leq t_{\epsilon,l}(a)\}$, $\widetilde{\lambda}_u(a, x) = \mathbb{1}\{\widehat{r}_u(a, x) > t_{\epsilon,u}(a)\}$. First notice that, by construction of $\widehat{t}_{\epsilon,j}(a)$, for every $a \in \mathcal{A}$:

$$o_{\mathbb{P}}(n^{-1/2}) = \mathbb{P}_n\widehat{\lambda}_j(a, X) - \mathbb{P}\lambda_j(a, X)$$

where $\mathbb{P}_n\widehat{\lambda}_j(a,X)$ is the sample average over the test sample used to construct the U-statistics. In this light, $\mathbb{U}_n h_l(A_1)t_{\epsilon,j}(A_1)\widehat{\lambda}_j(A_1,X_2) - \mathbb{U}h_l(A_1)t_{\epsilon,j}(A_1)\lambda_j(A_1,X_2) = o_{\mathbb{P}}(n^{-1/2})$ and

$$
\begin{aligned}
o_{\mathbb{P}}(n^{-1/2}) &= (\mathbb{U}_n - \mathbb{U})\left[h(A_1)t_{\epsilon,j}(A_1)\left\{\widehat{\lambda}_j(A_1,X_2) - \widetilde{\lambda}_j(A_1,X_2)\right\}\right] \\
&\quad + (\mathbb{U}_n - \mathbb{U})\left[h(A_1)t_{\epsilon,j}(A_1)\left\{\widetilde{\lambda}_j(A_1,X_2) - \lambda_j(A_1,X_2)\right\}\right] \\
&\quad + (\mathbb{U}_n - \mathbb{U})\left[h(A_1)t_{\epsilon,j}(A_1)\lambda_j(A_1,X_2)\right] + \mathbb{U}\left[t_{\epsilon,j}(A_1)h_l(A_1)\left\{\widehat{\lambda}_j(A_1,X_2) - \lambda_j(A_1,X_2)\right\}\right].
\end{aligned}
$$

Notice that the middle term involving $\widetilde{\lambda}_j(A_1,X_2) - \lambda_j(A_1,X_2)$ is an empirical process term of a fixed function given the training sample. Therefore, by Lemma 20, it is $o_{\mathbb{P}}(n^{-1/2})$ because

$$
\begin{aligned}
&\int \left|S_2\left\{\widetilde{\lambda}_j(a_1,x_2) - \lambda_j(a_1,x_2)\right\}\right| d\mathbb{P}(z_2) \\
&\leq \int \mathbb{1}\left\{|r_j(a_1,x_2) - t_{\epsilon,j}(a_1)| \leq \|\widehat{r}_j - r_j\|_\infty\right\} d\mathbb{P}(x_2) \\
&\quad + \int \mathbb{1}\left\{|r_j(a_2,x_1) - t_{\epsilon,j}(a_2)| \leq \|\widehat{r}_j - r_j\|_\infty\right\} d\mathbb{P}(a_2) \\
&\lesssim \|\widehat{r}_j - r_j\|_\infty \\
&= o_{\mathbb{P}}(1)
\end{aligned}
$$

because the densities of $r_j(a,X) - t_{\epsilon,j}(a)$ and $r_j(A,x) - t_{\epsilon,j}(A)$ are assumed to be bounded for any a and x. In this respect, we have

$$
\begin{aligned}
o_{\mathbb{P}}(n^{-1/2}) = (\mathbb{U}_n - \mathbb{U})&\left[h(A_1)t_{\epsilon,j}(A_1)\left\{\widehat{\lambda}_j(A_1,X_2) - \widetilde{\lambda}_j(A_1,X_2)\right\}\right] \\
&+ (\mathbb{U}_n - \mathbb{U})\left[h(A_1)t_{\epsilon,j}(A_1)\lambda_j(A_1,X_2)\right] \\
&+ \mathbb{U}\left[t_{\epsilon,j}(A_1)h_l(A_1)\left\{\widehat{\lambda}_j(A_1,X_2) - \lambda_j(A_1,X_2)\right\}\right]
\end{aligned}
$$

Because both $\widehat{\beta}$ and $\widetilde{\beta}$ solve empirical moment conditions, we have

$$
o_{\mathbb{P}}(n^{-1/2}) = \mathbb{U}_n\left\{\widehat{f}(Z_1,Z_2) - \widetilde{f}(Z_1,Z_2)\right\} + \mathbb{P}_n h(A)\left\{g(A;\widetilde{\beta}) - g(A;\widehat{\beta})\right\}
$$

and, in light of the observations above, we can subtract the $o_{\mathbb{P}}(n^{-1/2})$ term to obtain

$$\begin{aligned}
o_{\mathbb{P}}(n^{-1/2}) = &(\mathbb{P}_n - \mathbb{P})h(A_1)\widehat{f}_{\Delta}(Z_1)\left\{\widehat{\lambda}_j(A_1, X_1) - \widetilde{\lambda}_j(A_1, X_1)\right\} \\
&+ (\mathbb{U}_n - \mathbb{U})\left[h(A_1)\left\{\widehat{\lambda}_j(A_1, X_2) - \widetilde{\lambda}_j(A_1, X_2)\right\}\left\{\widehat{f}_r(Z_1, Z_2) - t_{\epsilon,j}(A_1)\right\}\right] \\
&- (\mathbb{U}_n - \mathbb{U})\left[t_{\epsilon,j}(A_1)h(A_1)\lambda_j(A_1, X_2)\right] \\
&+ \mathbb{P}\left[h(A_1)\left\{\widehat{\lambda}_j(A_1, X_1) - \lambda_j(A_1, X_1)\right\}\widehat{f}_{\Delta}(Z_1)\right] \\
&+ \mathbb{U}\left[h(A_1)\left\{\widehat{\lambda}_j(A_1, X_2) - \lambda_j(A_1, X_2)\right\}\left\{\widehat{f}_r(Z_1, Z_2) - t_{\epsilon,j}(A_1)\right\}\right] \\
&+ (\mathbb{P}_n - \mathbb{P})h(A)\left\{g(A; \widetilde{\beta}) - g(A; \widehat{\beta})\right\} + \dot{\Psi}_{\widehat{\beta}}(\widetilde{\beta} - \widehat{\beta}) + o(\|\widetilde{\beta} - \widehat{\beta}\|)
\end{aligned}$$

where we used the identity

$$\mathbb{P}_n h(A)\left\{g(A; \widetilde{\beta}) - g(A; \widehat{\beta})\right\} = (\mathbb{P}_n - \mathbb{P})h(A)\left\{g(A; \widetilde{\beta}) - g(A; \widehat{\beta})\right\} + \dot{\Psi}_{\widehat{\beta}}(\widetilde{\beta} - \widehat{\beta}) + o(\|\widetilde{\beta} - \widehat{\beta}\|).$$

Next, we claim that, conditioning on the training sample D^n and thus viewing $\widehat{f}_{\Delta}(z)$ and $\widehat{r}_j(a, x)$ as fixed functions, the function class

$$\mathcal{F} = \left\{f(z) = h_j(a)\widehat{f}_{\Delta}(z)\mathbb{1}\left\{\widehat{r}_j(a, x) - t_{\epsilon,j}(a) > 0\right\}, t_{\epsilon,j}(a) \in \mathcal{T}\right\}$$

is VC-subgraph. The subgraph $\mathcal{C}_q$ of $f_t(z) \equiv \mathbb{1}\left\{\widehat{r}_j(a, x) - t_{\epsilon,j} > 0\right\}$ is the collection of sets (z, c) in $\mathcal{Z} \times \mathbb{R}$ such that $f_t(z) \geq c$. For a given $t \equiv t_{\epsilon,j}$, let $S_0(t)$ be the collection of all z such that $\widehat{r}_j(a, x) - t_{\epsilon,j}(a) \leq 0$. Then, we have that the subgraph of $f_t(z)$ is

$$S_0(t) \times (-\infty, 0] \cup S_0^c(t) \times (-\infty, 1]$$

By Lemma 7.19 (iii) in Sen [2018], $S_0(t)$ is a VC set whenever $\widehat{r}_j(a, x) - t_{\epsilon,j}(a)$ is VC-subgraph, which is the case since $t_{\epsilon,j}(a)$ is VC-subgraph by assumption and $\widehat{r}_j(a, x)$ is a fixed function (given the training data). This then yields that the subgraph of $f_t(z)$ is a VC-set. Because $\mathcal{F}$ consists of products of VC-subgraph functions and $h_l(a)\widehat{f}_{\Delta}(z)$, a fixed function, we conclude that $\mathcal{F}$ itself is a VC-subgraph class. This means that the process $\sqrt{n}(\mathbb{P}_n - \mathbb{P})f$, $f \in \mathcal{F}$, is stochastically equicontinuous relative to $\rho(f_1, f_2) = [\text{var}\{f_1(Z) - f_2(Z)\}]^{1/2} \leq \|f_1 - f_2\|$. Thus,

$$(\mathbb{P}_n - \mathbb{P})h(A_1)\widehat{f}_{\Delta}(Z_1)\left\{\widehat{\lambda}_j(A_1, X_1) - \widetilde{\lambda}_j(A_1, X_1)\right\} = o_{\mathbb{P}}(n^{-1/2})$$

because, using the assumption that $r_j(A,X) - t_{\epsilon,j}(A)$ has a bounded density:

$$\begin{aligned}
&\int \left[h_j(a)\widehat{f}_\Delta(z)\left\{\widehat{\lambda}_j(a,x) - \widetilde{\lambda}_j(a,x)\right\}\right]^2 d\mathbb{P}(z) \lesssim \int \left|\widehat{\lambda}_j(a,x) - \widetilde{\lambda}_j(a,x)\right| d\mathbb{P}(a,x) \\
&\leq \int \mathbb{1}\left\{|\widehat{r}_j(a,x) - t_{\epsilon,j}(a)| \leq |\widehat{t}_{\epsilon,j}(a) - q(a)|\right\} d\mathbb{P}(a,x) \\
&\leq \int \mathbb{1}\left\{|r_j(a,x) - t_{\epsilon,j}(a)| \leq \|\widehat{r}_j - r_j\|_\infty + \|\widehat{t}_{\epsilon,j} - t_{\epsilon,j}\|_\infty\right\} d\mathbb{P}(a,x) \\
&\lesssim \|\widehat{r}_j - r_j\|_\infty + \|\widehat{t}_{\epsilon,j} - t_{\epsilon,j}\|_\infty \\
&= o_{\mathbb{P}}(1)
\end{aligned}$$

To analyze the empirical U-process, we rely on Arcones and Giné [1993]. In particular, by their Theorem 4.9 applied in conjuction with their Theorem 4.1, the process $\sqrt{n}(\mathbb{U}_n - \mathbb{U})f$, for

$$f \in \mathcal{F} = \left\{f(z_1,z_2) \mapsto h_l(a_1)\{\widehat{f}_r(z_1,z_2) - t_{\epsilon,j(a_1)}\}\mathbb{1}\left\{\widehat{r}_j(a_1,x_2) - \bar{t}_{\epsilon,j}(a_1) > 0\right\}, \bar{t}_{\epsilon,j}(a_1) \in \mathcal{T}\right\}$$

is stochastically equicontinuous, relative to the norm

$$\rho^2(f_1,f_2) = \int \left[\int S_2\left\{f_1(z_1,z_2) - f_2(z_1,z_2)\right\} d\mathbb{P}(z_2)\right]^2 d\mathbb{P}(z_1),$$

if, for instance, the class $\mathcal{F}$ is VC-subgraph. This is indeed the case under the assumption that $\mathcal{T}$ is a VC-subgraph class. Let $\widetilde{t}_{\epsilon,j}(z_1,z_2) \equiv \bar{t}_{\epsilon,j}(a_1)$ and $\mathcal{C}_t$ the subgraph of $a \mapsto \bar{t}_{\epsilon,j}(a)$. Then the subgraph of $\widetilde{t}$ is simply $\mathcal{Z} \cap \mathcal{C}_t \times \mathcal{Z}$, which is still a VC set. Then, as argued earlier, $\mathcal{F}$ consists of functions that are products of VC-subgraph classes and thus it is VC-subgraph. This concludes our proof that

$$(\mathbb{U}_n - \mathbb{U})\left[h(A_1)\left\{\widehat{\lambda}_j(A_1,X_2) - \widetilde{\lambda}_j(A_1,X_2)\right\}\left\{\widehat{f}_r(Z_1,Z_2) - t_{\epsilon,j}(A_1)\right\}\right] = o_{\mathbb{P}}(n^{-1/2})$$

since

$$\left|\int S_2\left\{\widehat{\lambda}_j(a_1,x_2) - \widetilde{\lambda}_j(a_1,x_2)\right\} d\mathbb{P}(z_2)\right| \lesssim \|\widehat{r}_j - r_j\|_\infty + \|\widehat{t}_{\epsilon,j} - t_{\epsilon,j}\|_\infty = o_{\mathbb{P}}(1).$$

Next, we have by Cauchy-Schwarz

$$\begin{aligned}
&\left|\mathbb{P}\left[h_l(A)\left\{\widehat{\lambda}_j(A,X) - \lambda_j(A,X)\right\}\widehat{w}(A,X)\left\{\kappa(A,X;\widehat{q}_j) - \widehat{\kappa}(A,X;\widehat{q}_j) - \mu(A,X) - \widehat{\mu}(A,X)\right\}\right]\right| \\
&\lesssim \int \left|\widehat{\lambda}_j(a,x) - \lambda_j(a,x)\right| d\mathbb{P}(a,x)\left(\|\kappa_j - \widehat{\kappa}_j\| + \|\widehat{q} - q\|^2 + \|\mu - \widehat{\mu}\|\right) \\
&\lesssim \left(\|\widehat{r}_j - r_j\|_\infty + \|\widehat{t}_{\epsilon,j} - t_{\epsilon,j}\|_\infty\right)\left(\|\kappa_j - \widehat{\kappa}_j\| + \|\widehat{q} - q\|^2 + \|\mu - \widehat{\mu}\|\right) \\
&= o_{\mathbb{P}}(n^{-1/2})
\end{aligned}$$

by assumption. This concludes our proof that $\mathbb{P}h(A_1)\left\{\widehat{\lambda}_j(A_1,X_1) - \lambda_j(A_1,X_1)\right\}\widehat{f}_\Delta(Z_1) =$

$o_{\mathbb{P}}(n^{-1/2})$.

Next, we have

$$\begin{aligned}
&\left|\mathbb{U}\left[h_l(A_1)\left\{\widehat{\lambda}_j(A_1,X_2)-\lambda_j(A_1,X_2)\right\}\left\{\widehat{f}_r(Z_1,Z_2)-f_r(Z_1,Z_2)\right\}\right]\right| \\
&=\left|\int h_l(A_1)\left\{\widehat{\lambda}_j(a,x)-\lambda_j(a,x)\right\}\{\widehat{\kappa}(a,x;\widehat{q}_j)-\kappa(a,x;\widehat{q})-\widehat{\mu}(a,x)-\mu(a,x)\}\,d\mathbb{P}(a)d\mathbb{P}(x)\right| \\
&\lesssim(\|\widehat{r}_j-r_j\|_\infty+\|\widehat{t}_{\epsilon,j}-t_{\epsilon,j}\|_\infty)(\|\widehat{\kappa}_j-\kappa_j\|+\|\widehat{\mu}-\mu\|+\|\widehat{q}-q\|^2)
\end{aligned}$$

and

$$\begin{aligned}
&\left|\mathbb{U}\left[h_l(A_1)\left\{\widehat{\lambda}_j(A_1,X_2)-\lambda_j(A_1,X_2)\right\}\{f_r(Z_1,Z_2)-t_{\epsilon,j}(A_1)\}\right]\right| \\
&=\left|\int h_l(a)\{\widehat{\lambda}_j(a,x)-\lambda_j(a,x)\}\{r_j(a,x)-t_{\epsilon,j}(a)\}d\mathbb{P}(a)d\mathbb{P}(x)\right| \\
&\lesssim\int\mathbb{1}\{|r_j(a,x)-t_{\epsilon,j}(a)|\le\|\widehat{r}_j-r_j\|_\infty+\|t_{\epsilon,j}-\widehat{t}_{\epsilon,j}\|_\infty\}\{r_j(a,x)-t_{\epsilon,j}(a)\}d\mathbb{P}(a)d\mathbb{P}(x) \\
&\le\left(\|\widehat{r}_j-r_j\|_\infty+\|t_{\epsilon,j}-\widehat{t}_{\epsilon,j}\|_\infty\right)\int\mathbb{P}\left(|r(a,X)-t_{\epsilon,j}(a)|\le\|\widehat{r}_j-r_j\|_\infty+\|t_{\epsilon,j}-\widehat{t}_{\epsilon,j}\|_\infty\right)d\mathbb{P}(a) \\
&\lesssim\|\widehat{r}_j-r_j\|_\infty^2+\|t_{\epsilon,j}-\widehat{t}_{\epsilon,j}\|_\infty^2
\end{aligned}$$

This concludes our proof that

$$\mathbb{U}\left[h_l(A_1)\left\{\widehat{\lambda}_j(A_1,X_2)-\lambda_j(A_1,X_2)\right\}\left\{\widehat{f}_r(Z_1,Z_2)-t_{\epsilon,j}(A_1)\right\}\right]=o_{\mathbb{P}}(n^{-1/2})$$

Statement 2 now follows if we can show that

$$(\mathbb{P}_n-\mathbb{P})h(A)\left\{g(A;\widetilde{\beta}_j)-g(A;\widehat{\beta}_j)\right\}=o_{\mathbb{P}}(n^{-1/2})$$

which is the case if $\|\widehat{\beta}_j-\widetilde{\beta}_j\|\le\|\widehat{\beta}_j-\beta_j\|+\|\widetilde{\beta}_j-\beta_j\|=o_{\mathbb{P}}(1)$ because $g(A;\beta)$, $\beta\in\mathbb{R}^k$ is a Donsker class. We can show consistency of $\widehat{\beta}_j$ for β_j by relying on Theorem 2.10 in Kosorok [2008] as done in the proof of Statement 1 of Lemma 23. Let $\widehat{\Psi}_n(\beta)=\mathbb{U}_nh(A_1)\{\widehat{f}_j(Z_1,Z_2)-g(A_1;\beta)\}$ and $\Psi(\beta)=\mathbb{U}h(A_1)\{f_j(Z_1,Z_2)-g(A_1;\beta_j)\}$. First, we need to show that $\|\Psi(\beta_n)\|\to0$ implies $\|\beta_n-\beta_j\|\to0$ for any sequence $\beta_n\in\mathbb{R}^k$. This is accomplished as in the proof of Lemma 23 by differentiability of $\Psi(\beta):\mathbb{R}^k\to\mathbb{R}^k$ and invertibility of its Jacobian matrix:

$$\Psi(\beta_n)=\dot{\Psi}_{\beta_j}(\beta_n-\beta_j)+o(\|\beta_n-\beta_j\|)\implies\|\beta_n-\beta_j\|\{1+o(1)\}\lesssim\|\Psi(\beta_n)\|\to0.$$

Second, we need to show that $\sup_{\beta\in\mathbb{R}^k}\|\Psi_n(\beta)-\Psi(\beta)\|=o_{\mathbb{P}}(1)$, which is the case since

$$\begin{aligned}
\Psi_n(\beta)-\Psi(\beta)=&(\mathbb{U}_n-\mathbb{U})h(A_1)\{\widehat{f}_j(Z_1,Z_2)-f_j(Z_1,Z_2)\}+(\mathbb{U}_n-\mathbb{U})f_j(Z_1,Z_2) \\
&+\mathbb{U}h(A_1)\{\widehat{f}_j(Z_1,Z_2)-f_j(Z_1,Z_2)\}+(\mathbb{P}_n-\mathbb{P})h(A)g(A;\beta)
\end{aligned}$$

All the terms above are $o_{\mathbb{P}}(1)$ by the arguments made in proving the previous steps and because

$g(a;\beta)$, $\beta \in \mathbb{R}^k$ is Donsker and thus Glivenko-Cantelli. This concludes our proof that

$$\widehat{\beta}_j - \widetilde{\beta}_j = -\dot{\Psi}^{-1}_{\widehat{\beta}_j}(\mathbb{U}_n - \mathbb{U})h(A_1)t_{\epsilon,j}(A_1)\lambda_j(A_1, X_2) + o_{\mathbb{P}}(n^{-1/2})$$

B.2.13 Moment condition in the time-varying case

We assume that $Y(\overline{a}_T) \perp\!\!\!\perp A_t \mid \overline{A}_{t-1}, \overline{X}_t, \overline{U}_t$. Then, we have, for $p(\cdot)$ denoting generically a density:

$$\begin{aligned}
&\mathbb{E}\left[h(\overline{A}_T)W_T(\overline{A}_T, \overline{X}_T)\left\{Y v_T(Y, \overline{A}_T, \overline{X}_T) - g(\overline{A}_T;\beta)\right\}\right] \\
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^T \pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \left\{y v_T(y, \overline{a}_T, \overline{x}_T) - g(\overline{a}_T;\beta)\right\} p(y, \overline{a}_T, \overline{x}_T) dy d\overline{a}_T d\overline{x}_T \\
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^T \pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \left\{ y \int \frac{\prod_{s=1}^T \pi(a_s \mid \overline{a}_{s-1}\overline{x}_s)}{\prod_{s=1}^T \pi(a_s \mid \overline{a}_{s-1}, \overline{x}_s, \overline{u}_s)} d\mathbb{P}(\overline{u}_T \mid \overline{a}_T, \overline{x}_T, y) - g(\overline{a}_T;\beta) \right\} \\
&\qquad \times p(y, \overline{a}_T, \overline{x}_T) dy d\overline{a}_T d\overline{x}_T \\
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^T \pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \left\{ \int y \frac{\prod_{s=1}^T \pi(a_s \mid \overline{a}_{s-1}\overline{x}_s)}{\prod_{s=1}^T \pi(a_s \mid \overline{a}_{s-1}, \overline{x}_s, \overline{u}_s)} d\mathbb{P}(\overline{u}_T, y \mid \overline{a}_T, \overline{x}_T) - g(\overline{a}_T;\beta) \right\} \\
&\qquad \times p(\overline{a}_T, \overline{x}_T) d\overline{a}_T d\overline{x}_T \\
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^T \pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \left\{ \int \mathbb{E}(Y^{\overline{a}_T} \mid \overline{a}_T, \overline{x}_T, \overline{u}_T) \frac{\prod_{s=1}^T \pi(a_s \mid \overline{a}_{s-1}\overline{x}_s)}{\prod_{s=1}^T \pi(a_s \mid \overline{a}_{s-1}, \overline{x}_s, \overline{u}_s)} d\mathbb{P}(\overline{u}_T \mid \overline{a}_T, \overline{x}_T) \right. \\
&\qquad\qquad \left. - g(\overline{a}_T;\beta) \right\} p(\overline{a}_T, \overline{x}_T) d\overline{a}_T d\overline{x}_T
\end{aligned}$$

Next, because $Y^{\overline{a}_T} \perp\!\!\!\perp A_T \mid \overline{X}_T, \overline{U}_T, \overline{A}_{T-1}$ and by Bayes' rule, we can further simplify:

$$
\begin{aligned}
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^{T}\pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \\
&\quad \times \left\{ \int \mathbb{E}(Y^{\overline{a}_T} \mid \overline{a}_{T-1}, \overline{x}_T, \overline{u}_T) \frac{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}\overline{x}_s)}{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}, \overline{x}_s, \overline{u}_s)} d\mathbb{P}(\overline{u}_T \mid \overline{a}_{T-1}, \overline{x}_T) \right. \\
&\qquad \left. - g(\overline{a}_T; \beta) \right\} p(\overline{a}_T, \overline{x}_T) d\overline{a}_T d\overline{x}_T \\
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^{T}\pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \\
&\quad \times \left\{ \int \mathbb{E}(Y^{\overline{a}_T} \mid \overline{a}_{T-1}, \overline{x}_T, \overline{u}_{T-1}) \frac{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}\overline{x}_s)}{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}, \overline{x}_s, \overline{u}_s)} d\mathbb{P}(\overline{u}_{T-1} \mid \overline{a}_{T-1}, \overline{x}_T) \right. \\
&\qquad \left. - g(\overline{a}_T; \beta) \right\} p(\overline{a}_T, \overline{x}_T) d\overline{a}_T d\overline{x}_T \\
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \\
&\quad \times \left\{ \int \mathbb{E}(Y^{\overline{a}_T} \mid \overline{a}_{T-1}, \overline{x}_T, \overline{u}_{T-1}) \frac{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}\overline{x}_s)}{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}, \overline{x}_s, \overline{u}_s)} d\mathbb{P}(\overline{u}_{T-1} \mid \overline{a}_{T-1}, \overline{x}_T) \right. \\
&\qquad \left. - g(\overline{a}_T; \beta) \right\} p(\overline{a}_{T-1}, \overline{x}_T) d\overline{a}_T d\overline{x}_T \\
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{x}_s, \overline{a}_{s-1})} \\
&\quad \times \left\{ \int \mathbb{E}(Y^{\overline{a}_T} \mid \overline{a}_{T-1}, \overline{x}_{T-1}, \overline{u}_{T-1}) \frac{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}\overline{x}_s)}{\prod_{s=1}^{T-1}\pi(a_s \mid \overline{a}_{s-1}, \overline{x}_s, \overline{u}_s)} d\mathbb{P}(\overline{u}_{T-1} \mid \overline{a}_{T-1}, \overline{x}_{T-1}) \right. \\
&\qquad \left. - g(\overline{a}_T; \beta) \right\} p(\overline{a}_{T-1}, \overline{x}_{T-1}) d\overline{a}_T d\overline{x}_{T-1}
\end{aligned}
$$

Repeating this calculation $T-1$ times, we arrive at

$$
\begin{aligned}
&= \int \frac{h(\overline{a}_T)\pi(\overline{a}_T)}{\pi(a_1 \mid x_1)} \left\{ \int \mathbb{E}(Y^{\overline{a}_T} \mid a_1, x_1, u_1) \frac{\pi(a_1 \mid x_1)}{\pi(a_1 \mid, x_1, u_1)} d\mathbb{P}(u_1 \mid a_1, x_1) - g(\overline{a}_T; \beta) \right\} p(a_1, x_1) d\overline{a}_T dx_1 \\
&= \int h(\overline{a}_T)\pi(\overline{a}_T) \left\{ \int \mathbb{E}(Y^{\overline{a}_T} \mid x_1, u_1) d\mathbb{P}(u_1 \mid x_1) - g(\overline{a}_T; \beta) \right\} p(x_1) d\overline{a}_T dx_1 \\
&= \int h(\overline{a}_T)\pi(\overline{a}_T) \left\{ \mathbb{E}(Y^{\overline{a}_T}) - g(\overline{a}_T; \beta) \right\} d\overline{a}_T = 0
\end{aligned}
$$

B.2.14 Additional useful lemmas

Lemma 18 (Theorem 12.3 in Van der Vaart [2000]). *Let $h(z_1, z_2)$ be a symmetric function of two variables and $\mathbb{E}\{h^2(Z_1, Z_2)\} < \infty$. Then,*

$$\sqrt{n}(\mathbb{U}_n - \mathbb{U})h(Z_1, Z_2) \rightsquigarrow N\left(0, 4var\{h_1(Z_1)\}\right)$$

where $h_1(Z_1) = \int h(Z_1, z_2)d\mathbb{P}(z_2)$.

Lemma 19 (Rudelson LLN for Matrices, Lemma 6.2 in Belloni et al. [2015]). *Let $Q_1, \ldots, Q_n$ be a sequence of independent symmetric, nonnegative $k \times k$-matrix valued random variables with $k \geq 2$ such that $Q = \mathbb{P}_n\{\mathbb{E}(Q_i)\}$ and $\|Q_i\| \leq M$ a.s.. Then, for $\widehat{Q} = \mathbb{P}_n Q$:*

$$\mathbb{E}\|\widehat{Q} - Q\| \lesssim \frac{M \log k}{n} + \sqrt{\frac{M\|Q\| \log k}{n}}.$$

Lemma 20. *Let $\widehat{h}(z_1, z_2)$ be a symmetric function estimated on a separate training sample D^n and*

$$\widehat{\Delta}(z_1) = \int \{\widehat{h}(z_1, z_2) - h(z_1, z_2)\}d\mathbb{P}(z_2).$$

If $\mathbb{E}\left[\left\{\widehat{h}(Z_1, Z_2) - h(Z_1, Z_2)\right\}^2 | D^n\right] < \infty$, then

$$(\mathbb{U}_n - \mathbb{U})\left\{\widehat{h}(Z_1, Z_2) - h(Z_1, Z_2)\right\} = O_{\mathbb{P}}\left(\frac{\|\widehat{\Delta}\|}{\sqrt{n}}\right).$$

Proof. We have

$$\mathbb{E}\left[(\mathbb{U}_n - \mathbb{U})\left\{\widehat{h}(Z_1, Z_2) - h(Z_1, Z_2)\right\} | D^n\right] = 0$$

because U-statistics are unbiased and $\widehat{h}(z_1, z_2)$ is a fixed function given D^n.
Let $\theta = \int f(z_1, z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2)$. The variance of a U-statistic with symmetric kernel f

satisying $\mathbb{E}f^2(Z_1, Z_2) < \infty$ is

$$\begin{aligned}
&\text{var}\{\mathbb{U}_n f(Z_1, Z_2)\} \\
&= \binom{n}{2}^{-2} \sum_{1\le i<j\le n} \sum_{1\le k<l\le n} \int \{f(z_i, z_j) - \theta\}\{f(z_k, z_l) - \theta\} d\mathbb{P}(z_i) d\mathbb{P}(z_j) d\mathbb{P}(z_k) d\mathbb{P}(z_l) \\
&= \binom{n}{2}^{-2} \binom{n}{2} \cdot 2 \cdot (n-1) \text{var}\left\{\int f(Z_1, z_2) d\mathbb{P}(z_2)\right\} + \binom{n}{2}^{-2} \binom{n}{2} \text{var}\{f(Z_1, Z_2)\} \\
&= \frac{4}{n} \text{var}\left\{\int f(Z_1, z_2) d\mathbb{P}(z_2)\right\} + o(n^{-1}) \\
&\le \frac{4}{n} \mathbb{E}\left[\left\{\int f(Z_1, z_2) d\mathbb{P}(z_2)\right\}^2\right] + o(n^{-1}).
\end{aligned}$$

Substituting $f(z_1, z_2) = \widehat{h}(z_1, z_2) - h(z_1, z_2)$ into the expression above, we get

$$\text{var}\left[(\mathbb{U}_n - \mathbb{U})\{\widehat{h}(Z_1, Z_2) - h(Z_1, Z_2)\} | D^n\right] \le \frac{4\|\widehat{\Delta}\|^2}{n} + o(n^{-1}).$$

The result then follows from Chebyshev's inequality. □

Lemma 21. *For $j = \{\ell, u\}$, let $s_j(Z; q_j) = q_j(Y|A, X) + \{Y - q_j(A, X)\}c_j^{\text{sgn}\{Y - q_j(Y|A,X)\}}$, where $c_\ell = \gamma^{-1}$, $c_u = \gamma$, $q_\ell(Y|A, X)$ is the $1/(1+\gamma)$-quantile of Y given (A, X), $q_u(Y|A, X)$ is the $\gamma/(1+\gamma)$-quantile of Y given (A, X) and $\kappa(A, X; q_j) = \mathbb{E}\{s(Z; q_j)|A, X\}$. Then, the following holds:*

1. *The map $q \mapsto s(Z; q)$ is Lipschitz;*
2. *The first and second derivatives of $q \mapsto \kappa(a, x; q)$ are*

$$\begin{aligned}
\frac{d}{dq}\kappa(A, X; q) &= 1 - c_j^{-1} \int_{-\infty}^{q} f(y|A = a, X = x)dy - c_j \int_{q}^{\infty} f(y|A = a, X = x)dy \\
\frac{d^2}{dq^2}\kappa(A, X; q) &= -c_j^{-1} f(q|A = a, X = x) + c_j f(q|A = a, X = x);
\end{aligned}$$

3. *The first derivative of $q \mapsto \kappa(a, x; q)$ vanishes at the true quantile $q_j(Y|A = a, X = x)$.*

Proof. All three statements were noted by Dorn et al. [2021]. To prove the first one, let $q_1 < q_2$ without loss of generality and notice that if either $y < q_1 < q_2$ or $q_1 < q_2 < y$,

$$|s(Z; q_1) - s(Z; q_2)| = \left|q_1 - q_2 + (q_2 - q_1)c_j^{\text{sgn}\{y - q_1\}}\right| \le (1 + \gamma)|q_1 - q_2|$$

because $y - q_1$ and $y - q_2$ agree on the sign. If $q_1 < y < q_2$, $|y - q_1| \le |q_1 - q_2|$ and

$|y - q_2| \leq |q_1 - q_2|$ so that

$$\begin{aligned}|s(Z;q_1) - s(Z;q_2)| &= \left|q_1 - q_2 + (y - q_1)c_j^{\text{sgn}\{y-q_1\}} + (y - q_2)c_j^{\text{sgn}\{y-q_2\}}\right| \\ &\leq (1 + \gamma^{-1} + \gamma)|q_1 - q_2|\end{aligned}$$

The second statement follows from an application of Leibniz rule of integration and the third by noticing that

$$\begin{aligned}\frac{d}{dq}\kappa(A, X; q)\Big|_{q=q_\ell} &= 1 - \gamma \cdot \frac{1}{1+\gamma} - \gamma^{-1}\left(1 - \frac{1}{1+\gamma}\right) = 0 \\ \frac{d}{dq}\kappa(A, X; q)\Big|_{q=q_u} &= 1 - \gamma^{-1} \cdot \frac{\gamma}{1+\gamma} - \gamma\left(1 - \frac{\gamma}{1+\gamma}\right) = 0.\end{aligned}$$

□

Lemma 22. *Let $f(A)$ be a fixed function of the random variable A with density upper bounded by B and $g(A)$ be any other fixed function. Then,*

$$\left|\int [\mathbb{1}\{g(a) \leq 0\} - \mathbb{1}\{f(a) \leq 0\}f(a)d\mathbb{P}(a)]\right| \leq 2B\|f - g\|_\infty^2.$$

Proof. By Lemma 1 in Kennedy et al. [2020],

$$|\mathbb{1}\{g(a) \leq 0\} - \mathbb{1}\{f(a) \leq 0\}| \leq \mathbb{1}\{|f(a)| \leq |f(a) - g(a)|\} \leq \mathbb{1}\{|f(a)| \leq \|f - g\|_\infty\}.$$

Therefore,

$$\begin{aligned}&\left|\int [\mathbb{1}\{g(a) \leq 0\} - \mathbb{1}\{f(a) \leq 0\}f(a)d\mathbb{P}(a)]\right| \leq \int \mathbb{1}\{|f(a)| \leq \|f - g\|_\infty\}|f(a)|d\mathbb{P}(a) \\ &\leq \|f - g\|_\infty \int \mathbb{1}\{|f(a)| \leq \|f - g\|_\infty\}d\mathbb{P}(a) \\ &= \|f - g\|_\infty \mathbb{P}\left(-\|f - g\|_\infty \leq f(A) \leq \|f - g\|_\infty\right) \\ &\leq 2B\|f - g\|_\infty^2.\end{aligned}$$

□

Lemma 23. *Let $\widehat{f}(z_1, z_2)$ be a function estimated on a separate independent sample and $g(A;\beta)$ be some parametric model indexed by $\beta \in \mathcal{B} \subset \mathbb{R}^k$. For some finite collection of known functions $h_1(A), \ldots, h_k(A)$, define*

$$\begin{aligned}\Psi_{n,l}(\beta) &= \mathbb{U}_n\left[h_l(A_1)\left\{\widehat{f}(Z_1, Z_2) - g(A_1;\beta)\right\}\right] \\ \Psi_l(\beta) &= \mathbb{U}\left[h_l(A_1)\left\{f(Z_1, Z_2) - g(A_1;\beta)\right\}\right].\end{aligned}$$

and let $\Psi_n(\beta) = [\Psi_{n,1}(\beta), \ldots, \Psi_{n,k}(\beta)]$ *and* $\Psi(\beta)$ *be defined similarly. Let* $\widehat{\beta}_n$ *and* β_0 *be the solutions to* $\Psi_n(\widehat{\beta}) = o_{\mathbb{P}}(n^{-1/2})$ *and* $\Psi(\beta_0) = 0$*, respectively, with* β_0 *in the interior of* $\mathcal{B}$*. Suppose that*

1. $\left\| \int S_2 \left\{ \widehat{f}(Z_1, z_2) - f(Z_1, z_2) \right\} d\mathbb{P}(z_2) \right\| = o_{\mathbb{P}}(1)$;
2. *The function class* $\mathcal{G} = \left\{ a \mapsto h_l(a) g(a; \beta), \beta \in \mathbb{R}^k \right\}$ *is Donsker for every* $l = \{1, \ldots, k\}$ *with integrable envelop and* $g(a; \beta)$ *is a continuous function of* β*;*
3. *The function* $\beta \mapsto \Psi(\beta)$ *is differentiable at all* β *with continuously invertible matrices* $\dot{\Psi}_{\beta_0}$ *and* $\dot{\Psi}_{\widehat{\beta}}$*, where* $\dot{\Psi}_\beta = -\mathbb{E}\left\{ h(A) \nabla_\beta^T g(A; \beta) \right\}$.
4. $max_l \left| \mathbb{U}\left[h_l(A_1) \left\{ \widehat{f}(Z_1, Z_2) - f(Z_1, Z_2) \right\} \right] \right| = o_{\mathbb{P}}(1)$.

Then,

1. $\|\widehat{\beta} - \beta\| = o_{\mathbb{P}}(1)$;
2. $\widehat{\beta} - \beta = \dot{\Psi}_{\widehat{\beta}}^{-1} \mathbb{U} h(A_1) \left\{ \widehat{f}(Z_1, Z_2) - f(Z_1, Z_2) \right\} - \dot{\Psi}_{\beta_0}^{-1} (\mathbb{U}_n - \mathbb{U}) h(A_1) \{ f(Z_1, Z_2) - g(A_1; \beta) \} + o_{\mathbb{P}}(n^{-1/2})$;
3. *In particular, if* $\dot{\Psi}_{\widehat{\beta}}^{-1} \mathbb{U}\left[h(A_1) \left\{ \widehat{f}(Z_1, Z_2) - f(Z_1, Z_2) \right\} \right] = o_{\mathbb{P}}(n^{-1/2})$*, then*

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \rightsquigarrow -\dot{\Psi}_{\beta_0}^{-1} N(0, 4\Sigma)$$

where

$$\Sigma = \mathbb{E}\left[\int S_2 h(A_1) \{ f(Z_1, z_2) - g(A_1; \beta_0) \} \, d\mathbb{P}(z_2) \right]^2.$$

Proof. Statement 1 follows from Theorem 2.10 in Kosorok [2008]. We need to verify the two conditions of the theorem, namely:

1. $\|\Psi(\beta_n)\| \to 0$ implies $\|\beta_n - \beta_0\| \to 0$ for any sequence $\beta_n \in \mathbb{R}^k$;
2. $\sup_{\beta \in \mathbb{R}^k} \|\Psi_n(\beta) - \Psi(\beta)\| = o_{\mathbb{P}}(1)$.

By differentiability of $\Psi(\beta) : \mathbb{R}^k \to \mathbb{R}^k$,

$$\Psi(\beta_n) = \dot{\Psi}(\beta_0)(\beta_n - \beta_0) + o(\|\beta_n - \beta_0\|) \implies \beta_n - \beta_0 + o(\|\beta_n - \beta_0\|) = \dot{\Psi}^{-1}(\beta_0) \Psi(\beta_n)$$

Therefore, $\|\beta_n - \beta_0\|\{1 + o(1)\} \lesssim \|\Psi(\beta_n)\| \to 0$. In addition,

$$\begin{aligned}\Psi_n(\beta) - \Psi(\beta) &= (\mathbb{U}_n - \mathbb{U})\left[h(A_1)\left\{\widehat{f}(Z_1, Z_2) - f(Z_1, Z_2)\right\}\right] \\ &\quad + (\mathbb{U}_n - \mathbb{U})h(A_1)f(Z_1, Z_2) + \mathbb{U}\left[h(A_1)\left\{\widehat{f}(Z_1, Z_2) - f(Z_1, Z_2)\right\}\right] \\ &\quad + (\mathbb{P}_n - \mathbb{P})h(A)g(A; \beta) \\ &= o_{\mathbb{P}}(1) + (\mathbb{P}_n - \mathbb{P})h(A; \beta)g(A; \beta)\end{aligned}$$

Because a Donsker class is also Glivenko-Cantelli, $\sup_{\beta \in \mathbb{R}^k} |(\mathbb{P}_n - \mathbb{P})h_l(A)g(A; \beta)| = o_{\mathbb{P}}(1)$. Therefore, since k is fixed, $\sup_{\beta \in \mathbb{R}^k} \|(\mathbb{P}_n - \mathbb{P})h(A)g(A; \beta)\| = o_{\mathbb{P}}(1)$.

To prove Statement 2, we apply Theorem 2.11 in Kosorok [2008] to the "debiased" moment condition

$$\begin{aligned}\widetilde{\Psi}_n(\widetilde{\beta}) &= \mathbb{U}_n\left[h(A_1)\left\{\widehat{f}(Z_1, Z_2) - g(A_1; \widetilde{\beta})\right\}\right] - \mathbb{U}\left[h(A_1)\left\{\widehat{f}(Z_1, Z_2) - f(Z_1, Z_2)\right\}\right] \\ &= o_{\mathbb{P}}(n^{-1/2})\end{aligned}$$

By the same reasoning used to derive statement 1, we have that $\|\widetilde{\beta} - \beta_0\| = o_{\mathbb{P}}(1)$. Next, we have

$$\begin{aligned}\sqrt{n}(\widetilde{\Psi}_n - \Psi)(\beta_0) &= \sqrt{n}(\mathbb{U}_n - \mathbb{U})h(A_1)\{f(Z_1, Z_2) - g(A_1; \beta_0)\} \\ &\quad + \sqrt{n}(\mathbb{U}_n - \mathbb{U})h(A_1)\left\{\widehat{f}(Z_1, Z_2) - f(Z_1, Z_2)\right\}\end{aligned}$$

Thus, by condition 1 and Lemma 20 together with Lemma 18, $\sqrt{n}(\widetilde{\Psi}_n - \Psi)(\beta_0) \rightsquigarrow N(0, 4\Sigma)$. Condition 2.12 in Theorem 2.11 in Kosorok [2008] requires that

$$\left\|\sqrt{n}(\mathbb{P}_n - \mathbb{P})\left\{h(A)g(A; \widetilde{\beta}) - h(A)g(A; \beta_0)\right\}\right\| = o_{\mathbb{P}}\left(1 + \sqrt{n}\|\widetilde{\beta} - \beta_0\|\right)$$

Because each function class $\mathcal{G}_l = \{a \mapsto h_l(a)g(a; \beta), \beta \in \mathbb{R}^k\}$ is Donsker, the process $\sqrt{n}(\mathbb{P}_n - \mathbb{P})f$, $f \in \mathcal{G}$, is stochastically equicontinuous relative to the norm $\rho^2(f_1, f_2) = \text{var}(f_1 - f_2) \leq \|f_1 - f_2\|^2$. In this respect, because $\|\widetilde{\beta} - \beta_0\| = o_{\mathbb{P}}(1)$ and k is fixed, the condition above is satisfied. Therefore, we conclude that

$$\widetilde{\beta} - \beta_0 = -\dot{\Psi}_{\beta_0}^{-1}(\mathbb{U}_n - \mathbb{U})h(A_1)\{f(Z_1, Z_2) - g(A_1; \beta_0)\} + o_{\mathbb{P}}(n^{-1/2}).$$

Finally, because $h_j(a)g(a; \beta)$ belongs to a Donsker class and $\|\widetilde{\beta} - \widehat{\beta}\| \leq \|\widetilde{\beta} - \beta_0\| + \|\widehat{\beta} - \beta_0\| =$

$o_{\mathbb{P}}(1)$:

$$\begin{aligned}
o_{\mathbb{P}}(n^{-1/2}) &= \left\{\Psi_n(\widehat{\beta}) - \widetilde{\Psi}_n(\widetilde{\beta})\right\} \\
&= (\mathbb{P}_n - \mathbb{P})h(A)\left\{g(A;\widetilde{\beta}) - g(A;\widehat{\beta})\right\} + \mathbb{P}h(A)\left\{g(A;\widetilde{\beta}) - g(A;\widehat{\beta})\right\} \\
&\quad + \mathbb{U}\left[h(A_1)\left\{\widehat{f}(Z_1,Z_2) - f(Z_1,Z_2)\right\}\right] \\
&= o_{\mathbb{P}}(n^{-1/2}) + o(\|\widetilde{\beta} - \widehat{\beta}\|) + \dot{\Psi}_{\widehat{\beta}}(\widetilde{\beta} - \widehat{\beta}) + \mathbb{U}\left[h(A_1)\left\{\widehat{f}(Z_1,Z_2) - f(Z_1,Z_2)\right\}\right]
\end{aligned}$$

Rearranging, we have

$$\widehat{\beta} - \widetilde{\beta} = \dot{\Psi}_{\widehat{\beta}}^{-1}\mathbb{U}\left[h(A_1)\left\{\widehat{f}(Z_1,Z_2) - f(Z_1,Z_2)\right\}\right] + o_{\mathbb{P}}(n^{-1/2}).$$

This concludes our proof that

$$\begin{aligned}
\widehat{\beta} - \beta_0 = &-\dot{\Psi}_{\beta_0}^{-1}(\mathbb{U}_n - \mathbb{U})h(A_1)\{f(Z_1,Z_2) - g(A_1;\beta_0)\} \\
&+ \dot{\Psi}_{\widehat{\beta}}^{-1}\mathbb{U}\left[h(A_1)\left\{\widehat{f}(Z_1,Z_2) - f(Z_1,Z_2)\right\}\right] + o_{\mathbb{P}}(n^{-1/2}).
\end{aligned}$$

□

Appendix C

Appendix for Chapter 4

C.1 Proof of Lemma 8

Write

$$\mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma)\} = \mathbb{E}\int_{\widehat{\Gamma}^c \cap \Gamma} |\tau(x) - \theta| f(x)dx + \mathbb{E}\int_{\widehat{\Gamma} \cap \Gamma^c} |\tau(x) - \theta| f(x)dx$$

Let $\alpha_n = c_\alpha a_n$ and $\beta_n = c_\beta (\log n)^{-\min\{(1+\xi)^{-1}, \epsilon\}}$, where $c_\alpha = 2c_a$ and

$$c_\beta = \max[c_b, c_1\{\mu(1+\xi)/c_7\}^{1/\kappa_2}].$$

Consider the first term and write

$$\begin{aligned}
&\widehat{\Gamma}^c \cap \Gamma = \{x \in \mathcal{X} : \widehat{\tau}(x) \leq \theta \text{ and } \tau(x) > \theta\} = A_1 \cup A_2 \cup A_3 \\
&A_1 = \{x \in \mathcal{X} : \widehat{\tau}(x) \leq \theta \text{ and } \theta < \tau(x) \leq \theta + \alpha_n\} \\
&A_2 = \{x \in \mathcal{X} : \widehat{\tau}(x) \leq \theta \text{ and } \theta + \alpha_n < \tau(x) \leq \theta + \beta_n\} \\
&A_3 = \{x \in \mathcal{X} : \widehat{\tau}(x) \leq \theta \text{ and } \tau(x) > \theta + \beta_n\}
\end{aligned}$$

Let n_0 be the integer such that for all $n \geq n_0$, it holds that

$$\alpha_n < \beta_n < \min(\eta, \epsilon_0, \Delta)$$

For the proof we assume that the sample size n exceeds n_0.

We have $A_1 \subseteq \{x \in \mathcal{X} : 0 < |\tau(x) - \theta| \leq \alpha_n\}$. By Assumption 4 (margin condition), we have

$$\mathbb{E}\int_{A_1} |\tau(x) - \theta| f(x)dx \leq \alpha_n \mathbb{P}(0 < |\tau(X) - \theta| \leq a_n) \leq c_0 \alpha_n^{1+\xi}$$

Next, let $J_n = \lfloor \log_2\{\beta_n/\alpha_n\} \rfloor + 1$. Notice that $\beta_n/\alpha_n \lesssim n^\mu$ so that $J_n \lesssim \log n$. Partition A_2

as $A_2 = \cup_{j=1}^{J_n} A_2 \cap \mathcal{V}_j$, where

$$\mathcal{V}_j = \{x \in \mathcal{X} : \widehat{\tau}(x) \leq \theta \text{ and } 2^{j-1}\alpha_n < \tau(x) - \theta \leq 2^j \alpha_n\}$$

We have

$$\mathbb{E}\int_{A_2} |\tau(x) - \theta| f(x) dx = \sum_{i=1}^{J_n} \mathbb{E}\int_{A_2 \cap \mathcal{V}_j} |\tau(x) - \theta| f(x) dx$$

and

$$\mathcal{V}_j \subset \{x \in \mathcal{X} : |\widehat{\tau}(x) - \tau(x)| > 2^{j-1}\alpha_n \text{ and } |\tau(x) - \theta| < 2^j \alpha_n\} \cap \mathcal{D}(\min(\eta, \epsilon_0))$$

To see why this is the case, consider a x^* such that $\tau(x^*) \in (\theta + 2^{j-1}\alpha_n, \theta + 2^j \alpha_n]$ and $\widehat{\tau}(x^*) \leq \theta$. Clearly, x^* satisfies $\tau(x^*) \in [\theta - 2^j a_n, \theta + 2^j \alpha_n]$. Notice that

$$\tau(x^*) - 2^{j-1}\alpha_n > \theta + 2^{j-1}\alpha_n - 2^{j-1}\alpha_n = \theta \geq \widehat{\tau}(x^*)$$

for any j. The claim follows because we have shown that $\widehat{\tau}(x^*) < \tau(x^*) - 2^{j-1}\alpha_n$ and thus x^* is in the larger set.

For any $j \geq 1$, we have that $2^{j-1}\alpha_n > c_a a_n$ and

$$\begin{aligned}
\mathbb{E}\int_{A_2 \cap \mathcal{V}_j} |\tau(x) - \theta| f(x) dx &\leq \|f\|_\infty 2^j \alpha_n \int_{\mathcal{X}} \mathbb{P}\left(|\widehat{\tau}(x) - \tau(x)| > 2^{j-1}\alpha_n\right) \mathbb{1}\{0 < |\tau(x) - \theta| < 2^j \alpha_n\} dx \\
&\leq \|f\|_\infty c_0 2^{j(1+\xi)} \alpha_n^{1+\xi} \left\{ c_3 \exp(-c_4 2^{(j-1)\kappa_1} c_\alpha^{\kappa_1}) + c_5 \frac{\delta_n^{1+\xi}}{2^{(j-1)(1+\xi)} \alpha_n^{1+\xi}} \right\} \\
&= \|f\|_\infty c_3 c_0 2^{j(1+\xi)} \alpha_n^{1+\xi} \exp(-c_4 2^{(j-1)\kappa_1} c_\alpha^{\kappa_1}) + c_5 c_0 2^{1+\xi} \delta_n^{1+\xi}
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\int_{A_2} |\tau(x) - \theta| f(x) dx &= \sum_{j=1}^{J_n} \mathbb{E}\int_{A_2 \cap \mathcal{V}_j} |\tau(x) - \theta| f(x) dx \\
&\leq \|f\|_\infty c_3 c_0 \alpha_n^{1+\xi} \sum_{j=1}^{J_n} 2^{j(1+\xi)} \exp\left\{ -c_4 \left(\frac{c_\alpha}{2}\right)^{\kappa_1} 2^{j\kappa_1} \right\} + J_n c_5 c_0 2^{1+\xi} \delta_n^{1+\xi} \\
&\lesssim a_n^{1+\xi} + \delta_n^{1+\xi} \log n
\end{aligned}$$

The last inequality follows, because for any $a, b, c > 0$, $\sum_{j=1}^{\infty} 2^{aj} \exp(-b2^{jc}) < \infty$. In fact, for any α, there exists a constant C such that $(1/e^b)^x \leq C x^{-\alpha}$ for any $x \geq 1$. Let j_0 be large

enough so that $2^{jc-1} \geq j \geq 1$ for all $j \geq j_0$. Then, for some constant C and $x_j = 2^{jc-1}$:

$$2^{aj}\left(\frac{1}{e^b}\right)^{2^{jc-1}} = 2^{a/c}x_j^{a/c}\left(\frac{1}{e^b}\right)^{x_j} \leq C \text{ for all } j \geq j_0.$$

Then, we have

$$\begin{aligned}\sum_{j=1}^{\infty} 2^{aj}\exp(-b2^{jc}) &= \sum_{j=1}^{j_0-1} 2^{aj}\exp(-b2^{jc}) + \sum_{j=j_0}^{\infty} 2^{aj}\exp(-b2^{jc-1})\exp(-b2^{jc-1}) \\ &\leq \sum_{j=1}^{j_0-1} 2^{aj}\exp(-b2^{jc}) + C\sum_{j=j_0}^{\infty}\exp(-b2^{jc-1}) \\ &< \infty\end{aligned}$$

Finally, we have that $\beta_n > c_b b_n$ so that

$$\begin{aligned}&\mathbb{E}\int_{A_3}|\tau(x)-\theta|f(x)dx \leq \int_{\mathcal{X}}|\tau(x)-\theta|\mathbb{P}\left(|\widehat{\tau}(x)-\tau(x)|>\beta_n\right)f(x)dx \\ &\leq \int_{\mathcal{X}}|\tau(x)-\theta|f(x)dx\left(c_6\exp\left[-c_7\left(\frac{c_\beta}{c_1}\right)^{\kappa_2}\left\{\frac{(\log n)^{1/\kappa_2+\epsilon}}{(\log n)^{\min\{(1+\xi)^{-1},\epsilon\}}}\right\}^{\kappa_2}\right]+\frac{c_8}{c_\beta^{1+\xi}}\delta_n^{1+\xi}\log n\right) \\ &\lesssim \exp\left\{-c_7\left(\frac{c_\beta}{c_1}\right)^{\kappa_2}\log n\right\}+c_8\delta_n^{1+\xi}\log n \\ &= \exp\left[-\max\{c_7(c_b/c_1)^{\kappa_2},\mu(1+\xi)\}\log n\right]+c_8\delta_n^{1+\xi}\log n \\ &\leq n^{-(1+\xi)\mu}+c_8\delta_n^{1+\xi}\log n \\ &\lesssim a_n^{1+\xi}+c_8\delta_n^{1+\xi}\log n\end{aligned}$$

The bound on $\mathbb{E}\int_{\widehat{\Gamma}\cap\Gamma^c}|\tau(x)-\theta|f(x)dx$ follows similarly.

C.2 Proof of Lemma 9

By definition, we have $\widehat{\tau}(x) = \sum_{i=1}^n W_i(x;X^n)\widehat{\varphi}(Z_i)$. Define $\overline{\tau}(x;X^n) = \sum_{i=1}^n W_i(x;X^n)\tau(X_i)$ and recall that $\Delta(x;X^n) = \overline{\tau}(x;X^n) - \tau(x)$. Finally, let $\widehat{b}(X_i) = \mathbb{E}\{\widehat{\varphi}(Z_i) - \varphi(Z_i) \mid X_i, D^n\}$. We start from the decomposition

$$\begin{aligned}\widehat{\tau}(x)-\tau(x) = &\sum_{i=1}^n W_i(x;X^n)\varphi(Z_i) - \overline{\tau}(x;X^n) + \sum_{i=1}^n W_i(x;X^n)\{\widehat{\varphi}(Z_i)-\varphi(Z_i)-\widehat{b}(X_i)\} \\ &+\Delta(x;X^n)+\sum_{i=1}^n W_i(x;X^n)\widehat{b}(X_i).\end{aligned}$$

Given (D^n, X^n), the last two terms are constants, whereas the first two are mean-zero. We have

$$\begin{aligned}
&\mathbb{P}\left(|\widehat{\tau}(x) - \tau(x)| > t\right) \\
&\leq \mathbb{E}\left\{\mathbb{P}\left(\left|\sum_{i=1}^{n} W_i(x; X^n)\varphi(Z_i) - \overline{\tau}(x; X^n)\right| > \frac{t}{3} - \frac{|\Delta(x; X_n)|}{2} \mid X^n\right)\right\} \\
&\quad + \mathbb{E}\left[\mathbb{P}\left(\left|\sum_{i=1}^{n} W_i(x; X^n)\{\widehat{\varphi}(Z_i) - \varphi(Z_i) - \widehat{b}(X_i)\}\right| > \frac{t}{3} - \frac{|\Delta(x; X_n)|}{2} \mid D^n, X^n\right)\right] \\
&\quad + \inf_{p>0}\left(\frac{3}{t}\right)^p \mathbb{E}\left|\sum_{i=1}^{n} W_i(x; X^n)\widehat{b}(X_i)\right|^p
\end{aligned}$$

As in equation 2.16 in Giné et al. [2000], we have, for $p \geq 2$:

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^{n} W_i(x; X^n)\{\varphi(Z_i) - \tau(X_i)\}\right|^p \mid X^n\right] \\
&\leq 2^p(p-1)^{p/2}\mathbb{E}\left(\left[\sum_{i=1}^{n} W_i^2(x; X^n)\{\varphi(Z_i) - \tau(X_i)\}^2\right]^{p/2} \mid X^n\right) \\
&\leq (4\|\varphi\|_\infty)^p p^{p/2} S^p(x; X^n)
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=1}^{n} W_i(x; X^n)\varphi(Z_i) - \overline{\tau}(x; X^n)\right|^p\right] &\leq (4\|\varphi\|_\infty)^p p^{p/2}\mathbb{E}\{S^p(x; X^n)\} \\
&\equiv (4\|\varphi\|_\infty)^p p^{p/2} s_n^p
\end{aligned}$$

The following lemma, which can be found for instance in Giné et al. [2000] (eq. 3.2, page 14), shows that an exponential inequality follows if all the moments $\mathbb{E}|\widehat{\tau}(x) - \tau(x)|^p$ are properly controlled.

Lemma 24. *Let X be some random variable such that $\mathbb{E}|X|^p \leq a_n^p p^{p/\alpha}$, for all $p \geq p_0$ and some fixed p_0. Then,*

$$\mathbb{P}\left(|X| > t\right) \leq e^{p_0} \exp\left\{-\left(\frac{t}{a_n e}\right)^\alpha\right\}$$

Proof. For t such that $p = (te^{-1}a_n^{-1})^\alpha \geq p_0$, we have the bound $t^{-p}\mathbb{E}|X|^p \leq e^{-p}$. For all t, we thus have $\mathbb{P}(|X| > t) \leq e^{p_0 - p}$, since for values of t for which $p < p_0$, $e^{p_0-p} > 1$ is a valid bound. □

In light of Lemma 24, we have for all $t \geq 3|\Delta(x; X^n)|$:

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^{n} W_i(x; X^n)\varphi(Z_i) - \overline{\tau}(x; X^n)\right| > \frac{t}{3} - \frac{|\Delta(x; X_n)|}{2} \mid X^n\right) \\
&\leq \mathbb{P}\left(\left|\sum_{i=1}^{n} W_i(x; X^n)\varphi(Z_i) - \overline{\tau}(x; X^n)\right| > \frac{t}{6} \mid X^n\right) \\
&\leq e^2 \exp\left\{-\left(\frac{t}{24e\|\varphi\|_\infty s_n}\right)^2\right\}
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^{n} W_i(x; X^n)\{\widehat{\varphi}(Z_i) - \varphi(Z_i) - \widehat{b}(X_i)\}\right| > \frac{t}{3} - \frac{|\Delta(x; X_n)|}{2} \mid D^n, X^n\right) \\
&\leq e^2 \exp\left\{-\left(\frac{t}{12c_2\|\varphi\|_\infty e s_n}\right)^2\right\}
\end{aligned}$$

under the assumption that $\|\widehat{\varphi} - \varphi - \widehat{b}\|_\infty \leq c_2\|\varphi\|_\infty$. Therefore, we conclude that

$$\mathbb{P}\left(|\widehat{\tau}(x) - \tau(x)| > t\right) \leq 2e^2 \exp\left\{-\left(\frac{t}{12(c_2 \vee 2)e\|\varphi\|_\infty s_n}\right)^2\right\} + 3^{1+\xi}\left(\frac{\delta_n}{t}\right)^{1+\xi}$$

for all $t \geq 3|\Delta(x; X^n)|$.

C.3 Proof of Lemma 10

Recall that the Lp-R-Learner can be written as $\widehat{\tau}(x_0) = \rho_h(x_0)^T \widehat{Q}^{-1}\widehat{R}$, where

$$\widehat{Q} = \mathbb{U}_n\{\widehat{f}_1(Z_1, Z_2)\} \quad \text{and} \quad \widehat{R} = \mathbb{U}_n\{\widehat{f}_2(Z_1, Z_2)\}$$

for some functions $\widehat{f}_1$ and $\widehat{f}_2$ described in Definition 3. Define $\tau_h(x_0) = \rho_h(x_0)^T Q^{-1} R$ to be the projection parameter. By proposition 4 in Kennedy et al. [2022], it holds that

$$|\tau_h(x_0) - \tau(x_0)| \leq \begin{cases} c_1 h^\gamma & \text{if } x_0 \in D(\eta) \\ c_2 h^{\gamma'} & \text{if } x_0 \notin D(\eta) \end{cases}$$

for some constants c_1 and c_2. Let $\widehat{S} = \widehat{R} - \widehat{Q}Q^{-1}R$. By Proposition 6 in Kennedy et al. [2022], we have, under the conditions of the theorem, for $J = \binom{d+\lfloor\gamma\rfloor}{\lfloor\gamma\rfloor}$ and a constant c_3:

$$\{\widehat{\tau}(x_0) - \tau_h(x_0)\}^2 \leq c_3 \sum_{j=1}^{J} \widehat{S}_j^2$$

Therefore, if $x_0 \in D(\eta)$ and $t \geq 2c_1 h^\gamma$:

$$\begin{aligned}
&\mathbb{P}\left(|\widehat{\tau}(x_0) - \tau(x_0)| > t\right) \leq \mathbb{P}\left(|\widehat{\tau}(x_0) - \tau_h(x_0)| > t - c_1 h^\gamma\right) \\
&\leq \mathbb{P}\left(\{\widehat{\tau}(x_0) - \tau_h(x_0)\}^2 > \frac{t^2}{4}\right) \\
&\leq \sum_{j=1}^{J} \mathbb{P}\left(\left|\widehat{S}_j\right| > \frac{t}{2\sqrt{c_3 J}}\right)
\end{aligned}$$

Next, we bound

$$\mathbb{P}\left(\left|\widehat{S}_j\right| > \frac{t}{2\sqrt{c_3 J}}\right)$$

Recall that $\widehat{S}_j = \mathbb{U}_n\{g(Z_1, Z_2)\}$, where

$$\begin{aligned}
&g(Z_1, Z_2) = f_{2j}(Z_1, Z_2) - [\widehat{Q}Q^{-1}R]_j \\
&f_{2j}(Z_1, Z_2) = \rho_{hj}(X_1)K_h(X_1)\widehat{\varphi}_{y1}(Z_1) + \rho_{hj}(X_1)K_h(X_1)\widehat{\varphi}_{y2}(Z_1, Z_2)K_h(X_2) \\
&\varphi_{y1}(Z_1) = \{Y - \mu_0(X_1)\}\{A - \pi(X_1)\} \\
&\varphi_{y2}(Z_1, Z_2) = -\{A_1 - \pi(X_1)\}b_h^T(X_1)\Omega^{-1}b_h(X_2)\{Y_2 - \mu_0(X_2)\} \\
&\Omega = \mathbb{E}\{b_h(X)b_h(X)^T\}, \text{ for } K_h(x) = h^{-d}\mathbb{1}(2\|x - x_0\| \leq h).
\end{aligned}$$

It will be useful to write $\widehat{S}_j$ as a sum of degenerate U-statistics, as follows:

$$\widehat{S}_j = \mathbb{U}_n g(Z_1, Z_2) = \mathbb{U}_n\{g_D(Z_1, Z_2)\} + \mathbb{P}_n\{g_1(Z_1)\} + \mathbb{P}_n\{g_2(Z_2)\} + \int g(z_1, z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2)$$

where

$$\begin{aligned}
&g_D(Z_1, Z_2) = g(Z_1, Z_2) - \int g(z_1, Z_2)d\mathbb{P}(z_1) - \int g(Z_1, z_2)d\mathbb{P}(z_2) + \int g(z_1, z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2) \\
&g_1(Z_1) = \int g(Z_1, z_2)d\mathbb{P}(z_2) - \int g(z_1, z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2) \\
&g_2(Z_2) = \int g(z_1, Z_2)d\mathbb{P}(z_1) - \int g(z_1, z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{S}_j\right| > \frac{t}{2\sqrt{c_3 J}}\right) \\
&\leq \mathbb{P}\left(|\mathbb{U}_n\{g_D(Z_1, Z_2)\}| + |\mathbb{P}_n\{g_1(Z_1)\}| + |\mathbb{P}_n\{g_2(Z_2)\}| > \frac{t}{2\sqrt{c_3 J}} - \left|\int g(z_1, z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2)\right|\right)
\end{aligned}$$

By proposition 9 in Kennedy et al. [2022], there exists a constant c_4 such that

$$\left| \int g(z_1, z_2) d\mathbb{P}(z_1) d\mathbb{P}(z_2) \right| \leq c_4 \left(\frac{k}{h^d} \right)^{-2s/d}$$

Therefore, for $t \geq 4\sqrt{c_3 J} c_4 \left(\frac{k}{h^d}\right)^{-2s/d}$:

$$\mathbb{P}\left(\left| \widehat{S}_j - S_j \right| > \frac{t}{2\sqrt{2c_3 J}} \right) \leq \mathbb{P}\left(|\mathbb{U}_n\{g_D(Z_1, Z_2)\}| + |\mathbb{P}_n\{g_1(Z_1)\}| + |\mathbb{P}_n\{g_2(Z_2)\}| > \frac{t}{4\sqrt{c_3 J}} \right)$$
$$\leq \mathbb{P}\left(|\mathbb{U}_n\{g_D(Z_1, Z_2)\}| > \frac{t}{12\sqrt{c_3 J}} \right) + \mathbb{P}\left(|\mathbb{P}_n\{g_1(Z_1)\}| > \frac{t}{12\sqrt{c_3 J}} \right)$$
$$+ \mathbb{P}\left(|\mathbb{P}_n\{g_2(Z_2)\}| > \frac{t}{12\sqrt{c_3 J}} \right)$$

The second and third terms can be analyzed by Bernstein's inequality. For the first term, we use a concentration inequality for U-statistics derived in Giné et al. [2000] and restated below. See also Cattaneo et al. [2022] for a similar use of this lemma.

Lemma 25 (Equation 3.5 in Giné et al. [2000]). *Let $f_{ij}(z_i, \widetilde{z}_j)$ be the kernel of a degenerate and decoupled second order U-statistic. Define*

$$A = max_{1 \leq i,j \leq n} \sup_{z, \widetilde{z}} |f_{ij}(z, \widetilde{z})|, \quad B^2 = max \left[\sup_{\widetilde{z}} \sum_{i=1}^{n} \mathbb{E}\left\{ f_{ij}^2(Z_i, \widetilde{z}) \right\}, \sup_{z} \sum_{j=1}^{n} \mathbb{E}\left\{ f_{ij}^2(z, \widetilde{Z}_j) \right\} \right]$$
$$C^2 = \sum_{1 \leq i,j \leq n} \mathbb{E}\left\{ f_{ij}^2(Z_i, \widetilde{Z}_j) \right\}$$

where $\{Z_i, 1 \leq i \leq n\}$ are independent random variables and $\{\widetilde{Z}_j, 1 \leq j \leq n\}$ are independent copies of Z_i. Then, for a universal constant K, the following holds

$$\mathbb{P}\left(\left| \sum_{i,j} f_{ij}(z_i, \widetilde{z}_j) \right| > t \right) \leq K \exp\left[-\frac{1}{K} min\left\{ \frac{t}{C}, \left(\frac{t}{B}\right)^{2/3}, \left(\frac{t}{A}\right)^{1/2} \right\} \right]$$

Lemma 25 is for decoupled U-statistics, however, because of the result in de la Peña and Montgomery-Smith [1995], as noted in Giné et al. [2000] (pages 15 and 20), the same conclusion holds for regular, undecoupled U-statistics simply with K replaced by a different constant. Thus, we will apply Lemma 25 without performing the additional decoupling step or introducing a different constant. In particular, we apply Lemma 25 with $f_{ij}(z_i, z_j) = \{n(n-1)\}^{-1} g_D(z_i, z_j)$

if $i \neq j$ and zero otherwise; that is:

$$\mathbb{P}\left(\left|\frac{1}{n(n-1)}\sum_{i\neq j} g_D(z_i,z_j)\right| > t \mid D^n\right) \leq K\exp\left[-\frac{1}{K}\min\left\{\frac{t}{C},\left(\frac{t}{B}\right)^{2/3},\left(\frac{t}{A}\right)^{1/2}\right\}\right],$$

where $A = \dfrac{\sup_{z_1,z_2}|g_D(z_1,z_2)|}{n(n-1)}$,

$$B^2 = \max\left[\sup_{z_2}\frac{\mathbb{E}\left\{g_D^2(Z_1,z_2)\mid D^n\right\}}{n^2(n-1)}, \sup_{z_1}\frac{\mathbb{E}\left\{g_D^2(z_1,Z_2)\mid D^n\right\}}{n^2(n-1)}\right], \text{ and}$$

$$C^2 = \frac{\mathbb{E}\left\{g_D^2(Z_1,Z_2)\mid D^n\right\}}{n(n-1)}$$

C.3.1 Bound on $\mathbb{P}\left(|\mathbb{U}_n g_D(Z_1,Z_2)| > \frac{t}{12\sqrt{c_3 J}} \mid D^n\right)$

First, notice that, under the assumption that $\|\widehat{Q}\| \lesssim 1$ and $\|Q^{-1}\| \lesssim 1$,

$$\begin{aligned}
[\widehat{Q}Q^{-1}R]_j &\lesssim \max_j |R_j| \\
&= \max_j \left|\int \rho_{hj}(x)K_h(x)\text{var}(A \mid X=x)\tau(x)f(x)dx\right| \\
&= \max_j \left|\int \rho_j(1/2+v)K(v)\text{var}(A \mid X = x_0+vh)\tau(x_0+vh)f(x_0+vh)dv\right| \\
&\leq \max_j \sup_v |\rho_j(1/2+v)\text{var}(A \mid X=x_0+vh)\tau(x_0+vh)f(x_0+vh)| \left|\int K(v)dv\right| \\
&\lesssim 1
\end{aligned}$$

and

$$\begin{aligned}
\left|\int g(z_1,z_2)d\mathbb{P}(z_1)\right| &\leq \left|\int f_{2j}(z_1,z_2)d\mathbb{P}(z_1)\right| + [\widehat{Q}Q^{-1}R_j]_j \\
&\lesssim h^{-d}\mathbb{1}(2\|x_2-x_0\|\leq h)\left|\widehat{\Pi}\left(\frac{dF}{d\widehat{F}}(\pi-\widehat{\pi})\right)(x_2)\right| + 1 \lesssim h^{-d}
\end{aligned}$$

$$\begin{aligned}
\left|\int g(z_1,z_2)d\mathbb{P}(z_2)\right| &\leq \left|\int f_{2j}(z_1,z_2)d\mathbb{P}(z_2)\right| + [\widehat{Q}^{-1}QR_j]_j \\
&\lesssim h^{-d}\mathbb{1}(2\|x_1-x_0\|\leq h)\left|1-\widehat{\Pi}\left(\frac{dF}{d\widehat{F}}(\mu-\widehat{\mu}_0)\right)(x_1)\right| + 1 \lesssim h^{-d}
\end{aligned}$$

and recall that by Proposition 9 in Kennedy et al. [2022],

$$\left|\int g(z_1,z_2)d\mathbb{P}(z_1,z_2)\right| \leq c_4\left(\frac{k}{h^d}\right)^{-2s/d}$$

for some constant c_4.

Term A.

We have

$$\begin{aligned}\sup_{z_1,z_2}|g_D(z_1,z_2)| &\leq 4\sup_{z_1,z_2}|g(z_1,z_2)| \\ &\lesssim \sup_{z_1,z_2}|\rho_{hj}(x_1)K_h(x_1)\widehat{\varphi}_{y1}(z_1)+\rho_{hj}(x_1)K_h(x_1)\widehat{\varphi}_{y2}(z_1,z_2)K_h(x_2)| \\ &\lesssim kh^{-2d}\end{aligned}$$

Thus, there exists a constant c_A such that $A \lesssim \frac{k}{n(n-1)h^{2d}} \leq c_A\sqrt{\frac{k}{n(n-1)h^{2d}}}$ since $kh^{-2d}n^{-2} \to 0$.

Term B.

We have

$$\begin{aligned}&f_{2j}^2(z_1,z_2) \\ &\lesssim h^{-2d}\mathbb{1}(2\|x_1-x_0\|\leq h) \\ &\quad + h^{-2d}\mathbb{1}(2\|x_1-x_0\|\leq h)h^{-2d}\mathbb{1}(2\|x_2-x_0\|\leq h)b_h^T(X_1)\widehat{\Omega}^{-1}b_h(X_2)b_h^T(X_2)\widehat{\Omega}^{-1}b_h(X_1)\end{aligned}$$

and, for instance, for $dF^*(v) = dF(x_0+h(v-0.5))$:

$$\begin{aligned}&\int f_{2j}^2(z_1,z_2)d\mathbb{P}(z_1) \lesssim h^{-d}\int_{v:\|v-0.5\|\leq 0.5} dF^*(v) \\ &\quad + h^{-3d}\mathbb{1}(2\|x_2-x_0\|\leq h)b_h(x_2)^T\widehat{\Omega}^{-1}\int_{v:\|v-0.5\|\leq 0.5} b(v)b(v)^T dF^*(v)\widehat{\Omega}^{-1}b_h(x_2) \\ &\lesssim \frac{k}{h^{3d}}\end{aligned}$$

and similarly for $\int f_{2j}^2(z_1,z_2)d\mathbb{P}(z_2)$. Therefore,

$$\sup_{z_2}\int g^2(z_1,z_2)d\mathbb{P}(z_1) \lesssim \frac{k}{h^{3d}} + [\widehat{Q}^{-1}QR_j]_j^2 \lesssim \frac{k}{h^{3d}}$$

and similarly for $\sup_{z_1}\int g^2(z_1,z_2)d\mathbb{P}(z_2)$. Furthermore,

$$\begin{aligned}&\sup_{z_2}\left|\int g(z_1,z_2)d\mathbb{P}(z_1)\right| \lesssim h^{-d}, \quad \sup_{z_1}\left|\int g(z_1,z_2)d\mathbb{P}(z_2)\right| \lesssim h^{-d}, \\ &\text{and} \quad \left|\int g(z_1,z_2)d\mathbb{P}(z_1)d\mathbb{P}(z_2)\right| \lesssim \left(\frac{k}{h^d}\right)^{-2s/d}.\end{aligned}$$

Therefore,

$$\sup_{z_2} \int g_D^2(z_1, z_2) d\mathbb{P}(z_1) \lesssim \frac{k}{h^{3d}} \quad \text{and} \quad \sup_{z_1} \int g_D^2(z_1, z_2) d\mathbb{P}(z_2) \lesssim \frac{k}{h^{3d}}.$$

Thus, for some constant c_B, we have

$$B \lesssim \frac{1}{\sqrt{nh^d}} \cdot \sqrt{\frac{k}{n(n-1)h^{2d}}} \leq c_B \sqrt{\frac{k}{n(n-1)h^{2d}}}$$

Term C.

$$\begin{aligned}
&\int f_{2j}^2(z_1, z_2) d\mathbb{P}(z_1) d\mathbb{P}(z_2) \lesssim h^{-d} \int_{v:\|v-0.5\|\leq 0.5} dF^*(v) \\
&\quad + h^{-2d} \int_{v:\|v-0.5\|\leq 0.5} b_h(v)^T \widehat{\Omega}^{-1} \int_{v:\|v-0.5\|\leq 0.5} b(v) b(v)^T dF^*(v) \widehat{\Omega}^{-1} b_h(v) dF^*(v) \\
&\lesssim \frac{k}{h^{2d}}
\end{aligned}$$

so that

$$\int g^2(z_1, z_2) d\mathbb{P}(z_1) d\mathbb{P}(z_2) \lesssim \frac{k}{h^{2d}}$$

Furthermore,

$$\left| \int g(z_1, z_2) d\mathbb{P}(z_1) \right| \lesssim h^{-d} \quad \text{and} \quad \left| \int g(z_1, z_2) d\mathbb{P}(z_2) \right| \lesssim h^{-d}$$

Therefore,

$$\begin{aligned}
\int g_D^2(z_1, z_2) d\mathbb{P}(z_1) d\mathbb{P}(z_2) &\lesssim \int g^2(z_1, z_2) d\mathbb{P}(z_1) d\mathbb{P}(z_2) + \int \left\{ \int g(z_1, z_2) d\mathbb{P}(z_1) \right\}^2 d\mathbb{P}(z_2) \\
&\quad + \int \left\{ \int g(z_1, z_2) d\mathbb{P}(z_2) \right\}^2 d\mathbb{P}(z_1) \\
&\lesssim \frac{k}{h^{2d}}
\end{aligned}$$

This means that $C \leq c_C \sqrt{\frac{k}{n(n-1)h^{2d}}}$ for some constant c_C.

To recap, we have derived that for some constant K, which now includes c_A, c_B and c_C:

$$\mathbb{P}\left(|\mathbb{U}_n g_D(Z_1, Z_2)| > \frac{t}{2\sqrt{c_3 J}} \mid D^n\right)$$
$$\leq K \exp\left[-\frac{1}{K}\min\left\{\frac{t}{\sqrt{\frac{k}{n(n-1)h^{2d}}}}, \left(\frac{t}{\sqrt{\frac{k}{n(n-1)h^{2d}}}}\right)^{2/3}, \left(\frac{t}{\sqrt{\frac{k}{n(n-1)h^{2d}}}}\right)^{1/2}\right\}\right],$$

for $t \geq 4\sqrt{c_3 J} c_4 \left(\frac{k}{h^d}\right)^{-2s/d}$.

C.3.2 Bound on $\mathbb{P}\left(|\mathbb{P}_n g_1(Z_1)| > \frac{t}{12\sqrt{c_3 J}} \mid D^n\right)$

By Bernstein's inequality, we have

$$\mathbb{P}\left(|\mathbb{P}_n g_1(Z_1)| > \frac{t}{12\sqrt{c_3 J}} \mid D^n\right) \leq 2\exp\left(-\frac{n(t/\{12\sqrt{c_3 J}\})^2/2}{\mathbb{E}\{g_1^2(Z_1) \mid D^n\} + \sup_{z_1}(3)^{-1}|g_1(z_1)|t}\right)$$

We have

$$\sup_{z_1}\left|\int g(z_1, z_2) d\mathbb{P}(z_2)\right| \leq c_5 h^{-d}, \quad \text{and} \quad \left|\int g(z_1, z_2) d\mathbb{P}(z_1) d\mathbb{P}(z_2)\right| \leq c_4 \left(\frac{k}{h^d}\right)^{-2s/d}.$$

Moreover,

$$\int \left\{\int g(z_1, z_2) d\mathbb{P}(z_2)\right\}^2 d\mathbb{P}(z_1) \leq c_6 h^{-d}$$

Therefore, we have

$$\mathbb{E}\{g_1^2(Z_1) \mid D^n\} \leq 2c_6 h^{-d} + 2c_4 \left(\frac{k}{h^d}\right)^{-4s/d} \leq c_7 h^{-d} \quad \text{and}$$
$$\sup_{z_1}|g_1(z_1)| \leq c_5 h^{-d} + c_4 \left(\frac{k}{h^d}\right)^{-2s/d} \leq c_8 h^{-d}$$

since $k/h^d \to \infty$. Therefore, we conclude that for all $t \leq 3c_7/c_8$ and $c_9 = 1/(4 \cdot 12^2 c_3 J c_7)$:

$$\mathbb{P}\left(|\mathbb{P}_n g_1(Z_1)| > \frac{t}{12\sqrt{c_3 J}} \mid D^n\right) \leq 2\exp\left(-c_9 n h^d t^2\right)$$

C.3.3 Bound on $\mathbb{P}\left(|\mathbb{P}_n g_2(Z_2)| > \frac{t}{12\sqrt{c_3 J}} \mid D^n\right)$

A similar to the one above yields that there exist constants c_{10} and Δ_{10} such that

$$\mathbb{P}\left(|\mathbb{P}_n g_2(Z_2)| > \frac{t}{12\sqrt{c_3 J}} \mid D^n\right) \leq 2\exp\left(-c_{10} n h^d t^2\right) \text{ for all } t \leq \Delta_{10}$$

C.3.4 Final step

To conclude, notice that

$$\sqrt{\frac{k}{n(n-1)h^{2d}}} = \sqrt{\frac{k}{n^2 h^{2d}} + \frac{k}{n^2(n-1)h^{2d}}} \lesssim \sqrt{\frac{k}{n^2 h^{2d}}} \equiv a_n$$

We have obtained that, for $x_0 \in D(\eta)$, there exists constants K, Δ, c_{11} and c_{12} such that, for all $c_{11}\max\left\{h^\gamma, \left(\frac{k}{h^d}\right)^{-2s/d}\right\} \leq t \leq \Delta$, it holds that:

$$\begin{aligned}\mathbb{P}\left(|\widehat{\tau}(x_0) - \tau(x_0) > t| \mid D^n\right) \leq K\exp\left[-\frac{1}{K}\min\left\{\frac{t}{a_n}, \left(\frac{t}{a_n}\right)^{2/3}, \left(\frac{t}{a_n}\right)^{1/2}\right\}\right] \\ + 4\exp\left(-c_{12} n h^d t^2\right)\end{aligned}$$

The optimal choice of k and h depends on the values of γ, s and d. In particular, we distinguish two cases.

Case I: $s \geq \frac{d/4}{1+d/2\gamma}$. Set

$$h = n^{-1/(2\gamma+d)} \text{ and } k = nh^d = n^{2\gamma/(2\gamma+d)} \implies a_n = \frac{1}{\sqrt{nh^d}} = n^{-\gamma/(2\gamma+d)}.$$

In this case, we have for all t such that $c_a a_n \leq t \leq \Delta$ for some constant c_a:

$$\mathbb{P}\left(|\widehat{\tau}(x_0) - \tau(x_0) > t| \mid D^n\right) \leq K\exp\left[-\frac{1}{K}\min\left\{\left(\frac{t}{a_n}\right)^2, \frac{t}{a_n}, \left(\frac{t}{a_n}\right)^{2/3}, \left(\frac{t}{a_n}\right)^{1/2}\right\}\right]$$

for some constant K.

Case II: $s < \frac{d/4}{1+d/2\gamma}$. Define $T = 1 + d/(2\gamma) + d/(4s)$, where we recall $s = (\alpha+\beta)/2$ (α = smoothness of $\pi, \widehat{\pi}$ and β = smoothness of $\mu, \widehat{\mu}$). Set

$$\begin{aligned}&h = n^{-1/(T\gamma)} \text{ and } k = n^{\{d/(2s) - d/\gamma\}/T} \\ &\implies a_n = n^{-1/T} \text{ and } a_n \geq \frac{1}{\sqrt{nh^d}} = n^{d/(2T\gamma)-1/2} \text{ because } s < \frac{d/4}{1+d/2\gamma}.\end{aligned}$$

Thus, in this case too, we have for all t such that $c_a a_n \leq t \leq \Delta$ for some constant c_a:

$$\mathbb{P}\left(|\widehat{\tau}(x_0) - \tau(x_0) > t| \mid D^n\right) \leq K \exp\left[-\frac{1}{K}\min\left\{\left(\frac{t}{a_n}\right)^2, \frac{t}{a_n}, \left(\frac{t}{a_n}\right)^{2/3}, \left(\frac{t}{a_n}\right)^{1/2}\right\}\right]$$

for some constant K.

If $x_0 \notin D(\eta)$, the same arguments use to prove the $x_0 \in D(\eta)$ case hold simply with γ replaced by γ'. The final rate would be $a'_n = n^{-1/T'}$, where $T' = 1 + d/(2\gamma') + d/(4s)$.

C.4 Proof of Theorem 4

We prove the case when $\alpha \geq \beta$, since the case $\alpha < \beta$ can be proved in a symmetric way by swapping the perturbations of μ_0 and π in a way analogous to that presented in Kennedy et al. [2022]. We proceed as follows

1. Let $z = (y, a, x) \in \{0,1\}^2 \times [0,1]^d$ and $f(x)$ the density of x with respect to the Lebesgue measure. We consider a data generating process such that each observation follows a distribution with density $p_{\omega,\lambda}(z)$, where $\omega \in \Omega = \{0,1\}^m$ and $\lambda = \{0,1\}^{2mk}$, for some k and with prior $\overline{\omega}$ on λ. The sample of n independent observations has thus density

$$p^n_{\omega} \equiv p^n_{\omega}(z_1, \ldots, z_n) = \int \prod_{i=1}^{n} p_{\omega,\lambda}(z_i) d\overline{\omega}(\lambda)$$

   Depending on ω, the density will have $\tau_\omega(x) = \mu_1(x) - \mu_0(x)$ fluctuated. For each density, the λ vector will govern the fluctuations of $\pi(x)$ and $\mu_0(x)$ and will not generally interact with ω. Notice that we parametrize the density by (π, μ_0, τ), so that $\mu_1(x) = \tau(x) + \mu_0(x)$. Crucially, we will establish that $p_{\omega,\lambda}(z)$ belongs to $\mathcal{P}$, the set of all densities compatible with assumptions 4 and 4, so that we have

$$\inf_{\widehat{\Gamma}} \sup_{p \in \mathcal{P}} \mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma_p)\} \geq \inf_{\widehat{\Gamma}} \max_{\omega} \mathbb{E}_{p^n_{\omega}}\{d_H(\widehat{\Gamma}, \Gamma_\omega)\}$$

   where Γ_ω is the true upper level set when the data is sampled from p^n_{ω}.

2. Under the margin assumption 4, we rely Proposition 2.1 in Rigollet and Vert [2009] to obtain

$$d_H(\widehat{\Gamma}, \Gamma_\omega) \gtrsim \left[\int_{(\widehat{\Gamma} \Delta \Gamma_\omega) \cap \{\tau_\omega(x) \neq \theta\}} f(x) dx\right]^{(1+\xi)/\xi}$$

3. We construct a vector $\widehat{\omega}$ such that

$$\int_{(\widehat{\Gamma} \Delta \Gamma_\omega) \cap \{\tau_\omega(x) \neq \theta\}} f(x) dx \geq \frac{1}{2} \int_{(\Gamma_{\widehat{\omega}} \Delta \Gamma_\omega) \cap \{\tau_\omega(x) \neq \theta\}} f(x) dx$$

and show that

$$\int_{(\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega)\cap\{\tau_\omega(x)\neq\theta\}} f(x)dx = 2\text{Leb}_d\{\mathcal{S}_{hk}(x_1)\}\rho(\widehat{\omega},\omega)$$

where $\rho(\widehat{\omega},\omega) = \sum_{i=1}^n \mathbb{1}(\widehat{\omega}_i \neq \omega_i)$ is the Hamming distance and $\mathcal{S}_{hk}(x_1)$ is a particular set defined below.

At this point we have the following chain of inequalities:

$$\begin{aligned}
\inf_{\widehat{\Gamma}} \sup_{p\in\mathcal{P}} \mathbb{E}\{d_H(\widehat{\Gamma},\Gamma_p)\} &\geq \inf_{\widehat{\Gamma}} \max_\omega \mathbb{E}_{p_\omega^n}\{d_H(\widehat{\Gamma},\Gamma_\omega)\} \\
&\gtrsim \inf_{\widehat{\Gamma}} \max_\omega \mathbb{E}_{p_\omega^n}\left(\left[\int_{(\widehat{\Gamma}\Delta\Gamma_\omega)\cap\{\tau_\omega(x)\neq\theta\}} f(x)dx\right]^{(1+\xi)/\xi}\right) \\
&\geq \inf_{\widehat{\Gamma}} \max_\omega \left(\mathbb{E}_{p_n^\omega}\left[\int_{(\widehat{\Gamma}\Delta\Gamma_\omega)\cap\{\tau_\omega(x)\neq\theta\}} f(x)dx\right]\right)^{(1+\xi)/\xi} \\
&\geq \frac{1}{2}\inf_{\widehat{\omega}} \max_\omega \left(\mathbb{E}_{p_n^\omega}\left[\int_{(\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega)\cap\{\tau_\omega(x)\neq\theta\}} f(x)dx\right]\right)^{(1+\xi)/\xi} \\
&= [\text{Leb}_d\{\mathcal{S}_{hk}(x_1)\}]^{(1+\xi)/\xi} \inf_{\widehat{\omega}} \max_\omega \left[\mathbb{E}_{p_n^\omega}\{\rho(\widehat{\omega},\omega)\}\right]^{(1+\xi)/\xi}
\end{aligned}$$

4. By Theorem 2.12 in Tsybakov [2009], if the Hellinger distance satisfies $H^2(p_{\omega'}^n, p_\omega^n) \leq 1$ for any ω', ω such that $\rho(\omega',\omega) = 1$, then

$$\inf_{\widehat{\omega}} \max_\omega \mathbb{E}_{p_\omega^n}\rho(\widehat{\omega},\omega) \geq m\left(\frac{1}{2} - \frac{\sqrt{3}}{4}\right)$$

We show that $\text{Leb}_d\{\mathcal{S}_{hk}(x_1)\} = (h/2)^d$ so that, putting everything together, we have

$$\inf_{\widehat{\Gamma}} \sup_{p\in\mathcal{P}} \mathbb{E}\{d_H(\widehat{\Gamma},\Gamma_p)\} \gtrsim (h^d m)^{(1+\xi)/\xi}$$

Choosing $h = O\left(n^{-1/(T\gamma)}\right)$ and $m = O\left(h^{-d+\gamma\xi}\right)$ yields the desired rate, where $T = 1 + d/(4s) + d/(2\gamma)$.

5. We verify that choosing $h = O\left(n^{-1/(T\gamma)}\right), m = O\left(h^{-d+\gamma\xi}\right)$ and $k = O\left(n^{d(\gamma-2s)/(2s\gamma T)}\right)$ yields $H^2(p_{\omega'}^n, p_\omega^n) \leq 1$ for any ω', ω such that $\rho(\omega',\omega) = 1$.

Step 1: Construction of fluctuated densities

Let $x_1, \ldots, x_{2m}$ denote a grid of $[0,1]^d$, for some m to be chosen later, and $\mathcal{C}_h(x_i)$ a cube with side h centered at x_i. Let $\mathcal{C}_{h/k^{1/d}}(m_{ji})$, $j \in \{1,\ldots,k\}$, be a partition of the cube $\mathcal{C}_h(x_i)$ into k, equally-sized cubes with midpoints $m_{1i}, \ldots, m_{ki}$. Then, for $\lambda \in \{-1,1\}^{2mk}$ and $\omega \in \{0,1\}^m$,

define the functions

$$\tau_\omega(x) = \theta + h^\gamma \sum_{i=1}^{m} \left[\omega_i B\left(\frac{x - x_i}{h}\right) + (1-\omega_i) B\left(\frac{x - x_{i+m}}{h}\right)\right]$$

$$\mu_{0\lambda}(x) = \frac{1}{2} + \left(\frac{h}{k^{1/d}}\right)^\beta \sum_{i=1}^{2m}\sum_{j=1}^{k} \lambda_{ij} B\left(\frac{x - m_{ji}}{h/2k^{1/d}}\right) - \frac{\tau_\omega(x)}{2}$$

$$\pi_{\lambda\omega}(x) = \frac{1}{2} + \left(\frac{h}{k^{1/d}}\right)^\alpha \sum_{i=1}^{m}\sum_{j=1}^{k} \left\{ (1-\omega_i)\lambda_{ij} B\left(\frac{x - m_{ji}}{h/2k^{1/d}}\right) + \omega_i \lambda_{i+mj} B\left(\frac{x - m_{ji+m}}{h/2k^{1/d}}\right)\right\}$$

$$f(x) = c_{hm}\left[1 - \sum_{i=1}^{2m} \mathbb{1}\{x \in \mathcal{C}_{2h}(x_i)\} + \sum_{i=1}^{2m} \mathbb{1}\{x \in \mathcal{S}_{hk}(x_i)\}\right]$$

$$\mathcal{S}_{hk}(x_i) = \cup_{j=1}^{k} \mathcal{C}_{h/2k^{1/d}}(m_{ji})$$

where $c_{hm} = \{1 - 2(2^d - 2^{-d})h^d m\}^{-1}$ and $B(u)$ is an infinitely differentiable function such that $B(u) = 1$ for $u \in [-1/2, 1/2]^d$ and $B(u) = 0$ for $u \notin [-1, 1]^d$. For example, $B\left(\frac{x-m_{ji}}{h/2k^{1/d}}\right) = 0$ for any $x \notin \mathcal{C}_{h/k^{1/d}}(m_{ji})$ and $B\left(\frac{x-m_{ji}}{h/2k^{1/d}}\right) = 1$ for any $x \in \mathcal{C}_{h/2k^{1/d}}(m_{ji})$.

Also notice that $f(x) = c_{hm}$ for any $x \notin \cup_{i=1}^{2m} \mathcal{C}_{2h}(x_i) \equiv \mathcal{X}_0$. In this region $\mathcal{X}_0$, $\tau_\omega(x) = \theta$, $\mu_{0\lambda}(x) = (1-\theta)/2$ and$\pi_{\omega\lambda}(x) = 1/2$. Therefore, the density of each observation can be written as

$$\begin{aligned} p_{\omega,\lambda}(z) = c_{hm} \sum_{i=1}^{m} & \mathbb{1}\{x \in \mathcal{S}_{hk}(x_i)\}\{\omega_i p_{i\lambda}(z) + (1-\omega_i) q_{i\lambda}(z)\} \\ & + \mathbb{1}\{x \in \mathcal{S}_{hk}(x_{i+m})\}\{(1-\omega_i) p_{i\lambda}(z) + \omega_i q_{i\lambda}(z)\} \\ & + \frac{1}{4} c_{hm} \mathbb{1}\{x \in \mathcal{X}_0\}\{1 + (2y-1)(2a-1)\theta\} \end{aligned}$$

where, for $s = (\alpha + \beta)/2$,

$$p_{i\lambda}(z) = \frac{1}{4} + (y - 1/2)\left(\frac{h}{k^{1/d}}\right)^\beta \sum_{j=1}^{k} \lambda_{ij} B\left(\frac{x - m_{ji}}{h/2k^{1/d}}\right) + (2a-1)(2y-1)\left\{\frac{\theta}{4} + \frac{h^\gamma}{4} B\left(\frac{x - x_i}{h}\right)\right\}$$

$$\begin{aligned} q_{i\lambda}(z) = \frac{1}{4} + & \left[(a - 1/2)\left(\frac{h}{k^{1/d}}\right)^\alpha + (y - 1/2)\left\{\left(\frac{h}{k^{1/d}}\right)^\beta + \theta\left(\frac{h}{k^{1/d}}\right)^\alpha\right\}\right] \sum_{j=1}^{k} \lambda_{ij} B\left(\frac{x - m_{ji}}{h/2k^{1/d}}\right) \\ & + (2a-1)(2y-1)\left\{\frac{\theta}{4} + \left(\frac{h}{k^{1/d}}\right)^{2s} \sum_{j=1}^{k} B\left(\frac{x - m_{ji}}{h/2k^{1/d}}\right)^2\right\} \end{aligned}$$

It is possible to verify that $\tau_\omega(x)$ is γ-smooth, $\mu_{0\lambda}(x)$ is β-smooth and $\pi_{\lambda\omega}(x)$ is α-smooth. To

verify the margin condition, notice that

$$\begin{aligned}\int_{x\in\mathcal{X}:0<|\tau_\omega(x)-\theta|<\epsilon} f(x)dx &= c_{hm}\sum_{i=1}^{m}\text{Leb}_d\{x\in\mathcal{S}_{hk}(x_i)\cap 0<|\tau_\omega(x)-\theta|<\epsilon\}\\ &= c_{hm}m\text{Leb}_d\{x\in\mathcal{S}_{hk}(x_1)\cap 0<|\tau_\omega(x)-\theta|<\epsilon\}\\ &= \frac{c_{hm}}{2^d}h^d m\mathbb{1}(\epsilon>h^\gamma)\\ &= \frac{c_{hm}}{2^d}h^{\gamma\xi}\mathbb{1}(\epsilon^\xi>h^{\xi\gamma})\\ &\lesssim \epsilon^\xi\end{aligned}$$

because the choice of h and m ensures that c_{hm} is upper bounded by a constant.

Step 2: Proposition 2.1 in Rigollet and Vert [2009]

By Proposition 2.1 in Rigollet and Vert [2009], under the margin assumption 4, we have

$$d_H(G_1,G_2)^{\xi/(1+\xi)}\gtrsim\int_{G_1\Delta G_2\cap\tau(x)\neq\theta}f(x)dx$$

for any G_1 and G_2.

Step 3: Reduction from $\widehat{\Gamma}$ to $\Gamma_{p_{\widehat{\omega}}}$

Define

$$\widehat{\omega}_i=\begin{cases}0 & \text{if Leb}_d\{\widehat{\Gamma}\cap\mathcal{S}_{hk}(x_i)\}<\text{Leb}_d\{\widehat{\Gamma}\cap\mathcal{S}_{hk}(x_{i+m})\}\\ 1 & \text{otherwise}\end{cases}$$

and notice that

$$\begin{aligned}\int_{(\widehat{\Gamma}\Delta\Gamma_\omega)\cap\tau_\omega(x)\neq\theta} f(x)dx &= c_{hm}\sum_{i=1}^{m}\text{Leb}_d[\widehat{\Gamma}\Delta\Gamma_\omega\cap\{\mathcal{S}_{hk}(x_i)\cup\mathcal{S}_{hk}(x_{i+m})\}]\\ &\geq \sum_{i=1}^{m}\text{Leb}_d[\widehat{\Gamma}\Delta\Gamma_\omega\cap\{\mathcal{S}_{hk}(x_i)\cup\mathcal{S}_{hk}(x_{i+m})\}]\\ &\geq \frac{1}{2}\sum_{i=1}^{m}\text{Leb}_d[\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega\cap\{\mathcal{S}_{hk}(x_i)\cup\mathcal{S}_{hk}(x_{i+m})\}]\end{aligned}$$

The second inequality follows because $c_{hm}\geq 1$. To see why the third inequality holds, first consider the case where $\omega_i=1$. If $\widehat{\omega}_i=1$, then

$$0=\text{Leb}_d[\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega\cap\{\mathcal{S}_{hk}(x_i)\cup\mathcal{S}_{hk}(x_{i+m})\}]\leq\text{Leb}_d[\widehat{\Gamma}\Delta\Gamma_\omega\cap\{\mathcal{S}_{hk}(x_i)\cup\mathcal{S}_{hk}(x_{i+m})\}]$$

If $\widehat{\omega}_i = 0$, then it means that $\text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_i)\} < \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_{i+m})\}$ so that

$$\begin{aligned}
&\text{Leb}_d[\widehat{\Gamma}\Delta\Gamma_\omega \cap \{\mathcal{S}_{hk}(x_i) \cup \mathcal{S}_{hk}(x_{i+m})\}] = \text{Leb}_d\{\widehat{\Gamma}^c \cap \mathcal{S}_{hk}(x_i)\} + \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_{i+m})\} \\
&= \text{Leb}_d\{\mathcal{S}_{hk}(x_i)\} - \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_i)\} + \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_{i+m})\} \\
&> \text{Leb}_d\{\mathcal{S}_{hk}(x_i)\} \\
&= \frac{1}{2}\text{Leb}_d[\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega \cap \{\mathcal{S}_{hk}(x_i) \cup \mathcal{S}_{hk}(x_{i+m})\}]
\end{aligned}$$

Similarly, consider the case where $\omega_i = 0$. If $\widehat{\omega}_i = 0$, then as before

$$0 = \text{Leb}_d[\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega \cap \{\mathcal{S}_{hk}(x_i) \cup \mathcal{S}_{hk}(x_{i+m})\}] \leq \text{Leb}_d[\widehat{\Gamma}\Delta\Gamma_\omega \cap \{\mathcal{S}_{hk}(x_i) \cup \mathcal{S}_{hk}(x_{i+m})\}]$$

If $\widehat{\omega}_i = 1$, then it means that $\text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_i)\} \geq \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_{i+m})\}$ so that

$$\begin{aligned}
&\text{Leb}_d[\widehat{\Gamma}\Delta\Gamma_\omega \cap \{\mathcal{S}_{hk}(x_i) \cup \mathcal{S}_{hk}(x_{i+m})\}] = \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_i)\} + \text{Leb}_d\{\widehat{\Gamma}^c \cap \mathcal{S}_{hk}(x_{i+m})\} \\
&= \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_i)\} + \text{Leb}_d\{\mathcal{S}_{hk}(x_{i+m})\} - \text{Leb}_d\{\widehat{\Gamma} \cap \mathcal{S}_{hk}(x_{i+m})\} \\
&\geq \text{Leb}_d\{\mathcal{S}_{hk}(x_{i+m})\} \\
&= \frac{1}{2}\text{Leb}_d[\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega \cap \{\mathcal{S}_{hk}(x_i) \cup \mathcal{S}_{hk}(x_{i+m})\}]
\end{aligned}$$

We have

$$\sum_{i=1}^{m} \text{Leb}_d[\Gamma_{\widehat{\omega}}\Delta\Gamma_\omega \cap \{\mathcal{S}_{hk}(x_i) \cup \mathcal{S}_{hk}(x_{i+m})\}] = 2\text{Leb}_d\{\mathcal{S}_{hk}(x_1)\} \sum_{i=1}^{m} \mathbb{1}(\widehat{\omega}_i \neq \omega_i)$$

Therefore, we have that, for any $\widehat{\Gamma}$, it holds that

$$\max_{\omega\in\Omega}\mathbb{E}_\omega \left\{ \int_{(\widehat{\Gamma}\Delta\Gamma_\omega)\cap\tau_\omega(x)\neq\theta} f(x)dx \right\} \geq \text{Leb}_d\{\mathcal{S}_{hk}(x_1)\} \inf_{\widehat{\omega}} \max_{\omega\in\Omega}\mathbb{E}_{p_\omega^n}\rho(\widehat{\omega}, \omega)$$

Step 4: Final bound

By Theorem 2.12 in Tsybakov [2009], if we can show that $H^2(p_\omega^n, p_{\omega'}^n) \leq 1$ for any $\omega, \omega' \in \Omega$ such that $\rho(\omega', \omega) = 1$, then

$$\inf_{\widehat{\omega}} \max_{\omega\in\Omega}\mathbb{E}_\omega\rho(\widehat{\omega}, \omega) \geq m\left(\frac{1}{2} - \frac{\sqrt{3}}{4}\right)$$

Thus,

$$\inf_{\widehat{\Gamma}} \sup_{p\in\mathcal{P}} \mathbb{E}\{d_H(\widehat{\Gamma}, \Gamma_p)\} \gtrsim (mh^d)^{(1+\xi)/\xi}$$

because $\text{Leb}_d\{\mathcal{S}_{hk}(x_1)\} = 2^{-d}h^d$. By choosing $h = O\left(n^{-1/(T\gamma)}\right)$ and $m = O\left(h^{-d+\gamma\xi}\right)$, where

$$T = 1 + d/(4s) + d/(2\gamma),$$

we get a lower bound of order $n^{-(1+\xi)/T}$ as desired.

Step 5: Verification of upper bound on Hellinger distance

Lemma 26 (Robins et al. [2009b], Kennedy et al. [2022]). *Let P_λ and Q_λ denote distributions indexed by a vector $\lambda = (\lambda_1, \ldots, \lambda_k)$, and let $\mathcal{Z} = \cup_{j=1}^k \mathcal{Z}_j$ denote a partition of the sample space. Assume:*

1. *$P_\lambda(\mathcal{Z}_j) = Q_\lambda(\mathcal{Z}_j) = p_j$ for all λ, and*
2. *the conditional distributions $\mathbb{1}_{\mathcal{Z}_j} dP_\lambda/p_j$ and $\mathbb{1}_{\mathcal{Z}_j} dQ_\lambda/p_j$ do not depend on λ_l for $l \neq j$, and only differ on partitions $j \in S \subseteq \{1, \ldots, k\}$.*

For a prior distribution $\overline{\omega}$ over λ, let $\overline{p} = \int p_\lambda d\overline{\omega}(\lambda)$ and $\overline{q} = \int q_\lambda d\overline{\omega}(\lambda)$, and define

$$\delta_1 = max_{j\in S} \sup_\lambda \int_{\mathcal{Z}_j} \frac{(p_\lambda - \overline{p})^2}{p_\lambda p_j} d\nu$$
$$\delta_2 = max_{j\in S} \sup_\lambda \int_{\mathcal{Z}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda p_j} d\nu$$
$$\delta_3 = max_{j\in S} \sup_\lambda \int_{\mathcal{Z}_j} \frac{(\overline{q} - \overline{p})^2}{p_\lambda p_j} d\nu$$

for a dominating measure ν. If $\overline{p}/p_\lambda \leq b < \infty$ and $np_j max(1, \delta_1, \delta_2) \leq b$ for all j, then

$$H^2\left(\int P_\lambda^n d\overline{\omega}(\lambda), Q_\lambda^n d\overline{\omega}(\lambda)\right) \leq Cn \sum_{j\in S} p_j \left\{n\left(max_{j\in S} p_j\right)(\delta_1\delta_2 + \delta_2^2) + \delta_3\right\}$$

for a constant C only depending on b.

It remains to verify that, given our choices of h, m, and k, it holds that $H^2(p_\omega^n, p_{\omega'}^n) \leq 1$ for any $\omega, \omega' \in \Omega$ such that $\rho(\omega', \omega) = 1$.

Following similar calculations as Kennedy et al. [2022], we will rely on their Lemma 2 (from Robins et al. [2009b] and restated above in Lemma 26) to derive a bound on the Hellinger distance.

Let us partition the space according to $\mathcal{Z}_{ji} = \mathcal{C}_{h/2k^{1/d}}(m_{ji}) \times \{0,1\}^2$, $j \in \{1, \ldots, k\}$ and $i \in \{1, \ldots, 2m\}$ and $\mathcal{Z}_0 = ([0,1]^d \times \{0,1\}^2)/\left(\cup_i \cup_j \mathcal{Z}_{ji}\right)$. On $\mathcal{Z}_0$, we have for any ω:

$$p_\omega(z) = \frac{1}{4} c_{hm} \mathbb{1}(x \in \mathcal{X}_0)\{1 + (2y-1)(2a-1)\theta\} \text{ for any } z \in \mathcal{Z}_0,$$

so that $\int_{\mathcal{Z}_0} p_\omega(z)dz = c_{hm}(1 - 2^{d+1}mh^d)$.

Next, notice that, because $\rho(\omega', \omega) = 1$, the densities $p^n_{\omega'}$ and p^n_ω differ only on two cubes, which, without loss of generality, we take to be $\mathcal{C}_{2h}(x_1)$ and $\mathcal{C}_{2h}(x_{m+1})$. This corresponds to a difference in the first coordinate of ω. To keep things clear, let $\omega_1 = (1, \omega_2, \ldots, \omega_m)$ and $\omega_0 = (0, \omega_2, \ldots, \omega_m)$. This way, we have

$$
\begin{aligned}
&p_{\omega_1,\lambda}(z) = \\
&\quad c_{hm}\mathbb{1}\{x \in \mathcal{S}_{hk}(x_1)\}\left[\frac{1}{4} + (y-1/2)\left(\frac{h}{k^{1/d}}\right)^\beta \sum_{j=1}^k \lambda_{1j} B\left(\frac{x - m_{j1}}{h/2k^{1/d}}\right) + (2a-1)(2y-1)\left\{\frac{\theta}{4} + \frac{h^\gamma}{4} B\left(\frac{x - x_1}{h}\right)\right\}\right] \\
&\quad + c_{hm}\sum_{i=2}^m \mathbb{1}\{x \in \mathcal{S}_{hk}(x_i)\}\{\omega_i p_{i\lambda}(z) + (1-\omega_i)q_{i\lambda}(z)\} + \mathbb{1}\{x \in \mathcal{S}_{hk}(x_{i+m})\}\{(1-\omega_i)p_{i\lambda}(z) + \omega_i q_{i\lambda}(z)\} \\
&\quad + c_{hm}\mathbb{1}\{x \in \mathcal{S}_{hk}(x_{1+m})\}\left(\frac{1}{4} + \left[(a-1/2)\left(\frac{h}{k^{1/d}}\right)^\alpha + (y-1/2)\left\{\left(\frac{h}{k^{1/d}}\right)^\beta + \theta\left(\frac{h}{k^{1/d}}\right)^\alpha\right\}\right]\sum_{j=1}^k \lambda_{1+mj} B\left(\frac{x - m_{j1+m}}{h/2k^{1/d}}\right)\right. \\
&\qquad\qquad \left. +(2a-1)(2y-1)\left\{\frac{\theta}{4} + \left(\frac{h}{k^{1/d}}\right)^{2s}\sum_{j=1}^k B\left(\frac{x - m_{j1+m}}{h/2k^{1/d}}\right)^2\right\}\right) \\
&\quad + \frac{1}{4}c_{hm}\mathbb{1}\{x \in \mathcal{X}_0\}\{1 + (2y-1)(2a-1)\theta\}
\end{aligned}
$$

and $p_{\omega_0,\lambda}(z)$ similarly defined. We will apply Lemma 26 with $P_\lambda(z) = p_{\omega_1,\lambda}(z)$ and $Q_\lambda(z) = p_{\omega_0,\lambda}(z)$.

First, notice that, for any (i, j) and vector λ, we have

$$\int_{\mathcal{Z}_{ji}} p_{\omega_1,\lambda}(z)dz = \int_{\mathcal{Z}_{ji}} p_{\omega_0,\lambda}(z)dz = \frac{c_{hm}h^d}{2^d k} \equiv p_{ji}$$

Further, λ_{ij} only affects the densities in $\mathcal{Z}_{ji}$, so the second condition in the lemma is satisfied.

Furthermore, because $p_{\omega_1,\lambda}(z)$ only differs from $p_{\omega_0,\lambda}(z)$ on $2k$ elements of the partition, it holds that

$$\sum_{(ij)\in S} p_{ji} = \sum_{j=1}^k p_{1j} + p_{1+mj} \lesssim h^d$$

Therefore, provided we can verify the other assumptions of Lemma 26, the Hellinger distance is upper bounded by

$$H^2\left(\int p^n_{\omega_1,\lambda}d\overline{\omega}(\lambda), p^n_{\omega_0,\lambda}d\overline{\omega}(\lambda)\right) \lesssim n^2h^d\left(\max_{(ji)\in S} p_{ji}\right)(\delta_1\delta_2 + \delta_2^2) + nh^d\delta_3$$

We take $\overline{\omega}(\lambda)$ to be a uniform prior on λ so that $\lambda_j = \{-1, 1\}$ independently and with equal

probability. Then,

$$\overline{p}_i(z) \equiv \int p_{i\lambda}(z)d\overline{\omega}(\lambda) = \frac{1}{4} + (2a-1)(2y-1)\left\{\frac{\theta}{4} + \frac{h^\gamma}{4}B\left(\frac{x-x_i}{h}\right)\right\}$$

$$\overline{q}_i(z) \equiv \int q_{i\lambda}(z)d\overline{\omega}(\lambda) = \frac{1}{4} + (2a-1)(2y-1)\left[\frac{\theta}{4} + \left(\frac{h}{k^{1/d}}\right)^{2s}\sum_{j=1}^{k} B\left(\frac{x-m_{ji}}{h/2k^{1/d}}\right)^2\right]$$

$$\overline{p}_{-1}(z) \equiv c_{hm}\sum_{i=2}^{m} \mathbb{1}\{x \in \mathcal{S}_{hk}(x_i)\}\{\omega_i\overline{p}_i(z) + (1-\omega_i)\overline{q}_i(z)\}$$
$$+ \mathbb{1}\{x \in \mathcal{S}_{hk}(x_{i+m})\}\{(1-\omega_i)\overline{p}(z) + \omega_i\overline{q}_i(z)\} + \frac{1}{4}c_{hm}\mathbb{1}\{x \in \mathcal{X}_0\}\{1 + (2y-1)(2a-1)\theta\}$$

$$\overline{p}_{\omega_1}(z) \equiv \int p_{\omega_1,\lambda}(z)d\overline{\omega}(\lambda) = \overline{p}_{-1}(z) + c_{hm}\left[\mathbb{1}\{x \in \mathcal{S}_{hk}(x_1)\}\overline{p}_1(z) + \mathbb{1}\{x \in \mathcal{S}_{hk}(x_{1+m})\}\overline{q}_{1+m}(z)\right]$$

$$\overline{p}_{\omega_0}(z) \equiv \int p_{\omega_0,\lambda}(z)d\overline{\omega}(\lambda) = \overline{p}_{-1}(z) + c_{hm}\left[\mathbb{1}\{x \in \mathcal{S}_{hk}(x_1)\}\overline{q}_1(z) + \mathbb{1}\{x \in \mathcal{S}_{hk}(x_{1+m})\}\overline{p}_{1+m}(z)\right]$$

Next, we bound

$$\delta_1 \equiv \max_{(ij)\in S}\sup_{\lambda}\int_{\mathcal{Z}_{ji}} \frac{\{p_{\omega_1,\lambda}(z) - \overline{p}_{\omega_1}(z)\}^2}{p_{\omega_1,\lambda}(z)p_{ji}}dz$$

In the following, we use the bound $(a+b)^2 \leq 2a^2 + 2b^2$ and the fact that $\beta \leq \alpha$ and $k \geq 1$. We have

$$\int_{\mathcal{Z}_{ji}} \frac{\{p_{\omega_1,\lambda}(z) - \overline{p}_{\omega_1}(z)\}^2}{p_{\omega_1,\lambda}(z)p_{ij}}dz \lesssim \left(\frac{h}{k^{1/d}}\right)^{2\beta}\frac{c_{hm}^2}{p_{ij}}\int\sum_{a,y}\frac{\mathbb{1}\{x \in \mathcal{C}_{h/2k^{1/d}}(m_{ji})\}}{p_{\omega_1,\lambda}(z)}dz$$

$$\int_{\mathcal{Z}_{ji+m}} \frac{\{p_{\omega_1,\lambda}(z) - \overline{p}_{\omega_1}(z)\}^2}{p_{\omega_1,\lambda}(z)p_{ij}}dz \lesssim \left(\frac{h}{k^{1/d}}\right)^{2\beta}\frac{c_{hm}^2}{p_{i+mj}}\int\sum_{a,y}\frac{\mathbb{1}\{x \in \mathcal{C}_{h/2k^{1/d}}(m_{ji+m})\}}{p_{\omega_1,\lambda}(z)}dz$$

Let ϵ and $\overline{\epsilon}$ be such that

$$\min\left\{\left(\frac{1-|\theta|}{4} - \frac{1}{2}\left(\frac{h}{k^{1/d}}\right)^\beta - \frac{1+|\theta|}{2}\left(\frac{h}{k^{1/d}}\right)^\alpha - \left(\frac{h}{k^{1/d}}\right)^{2s}\right),\right.$$
$$\left.\left(\frac{1-|\theta|}{4} - \frac{1}{2}\left(\frac{h}{k^{1/d}}\right)^\beta - \frac{h^\gamma}{4}\right)\right\} \geq \epsilon$$

$$\overline{\epsilon} = \frac{1+|\theta|}{4} + \max\left\{\frac{h^\gamma}{4}, \left(\frac{h}{k^{1/d}}\right)^{2s}\right\}$$

Then, $p_{\omega_1,\lambda}(z) \geq c_{hm}\epsilon$ and

$$\delta_1 \leq \frac{1}{\epsilon}\left(\frac{h}{k^{1/d}}\right)^{2\beta} \quad \text{and} \quad \frac{\overline{p}_{\omega_1}(z)}{p_{\omega_1,\lambda}(z)} \leq 1 \vee \frac{\overline{\epsilon}}{\epsilon}.$$

Next, suppose $h^{\gamma-2s} = 4k^{-2s/d}$. We have

$$\begin{aligned}\delta_2 &\equiv \max_{ij} \sup_{\lambda} \int_{\mathcal{Z}_{ji}} \frac{\{p_{\omega_1,\lambda}(z) - p_{\omega_0,\lambda}(z)\}^2}{p_{1\lambda}(z)p_{ji}} dz \\ &= \max_{j} \sup_{\lambda} \int_{\mathcal{Z}_{j1}} \frac{\{p_{\omega_1,\lambda}(z) - p_{\omega_0,\lambda}(z)\}^2}{p_{\omega_1,\lambda}(z)p_{j1}} dz\end{aligned}$$

by symmetry and

$$\int_{\mathcal{Z}_{j1}} \frac{\{p_{\omega_1,\lambda}(z) - p_{\omega_0,\lambda}(z)\}^2}{p_{\omega_1,\lambda}(z)p_{j1}} dz \lesssim \frac{1}{\epsilon}\left(\frac{h}{k^{1/d}}\right)^{2\alpha}$$

because for any i:

$$B\left(\frac{x - m_{ji}}{h/2k^{1/d}}\right) = B\left(\frac{x - x_i}{h}\right) = 1 \quad \text{if } x \in \cup_j \mathcal{C}_{h/2k^{1/d}}(m_{ji})$$

Finally, again if $h^{\gamma-2s} = 4k^{-2s/d}$,

$$\begin{aligned}&\overline{p}_{\omega_0}(z) - \overline{p}_{\omega_1}(z) = \\ &= c_{hm}\mathbb{1}\{x \in \mathcal{S}_{hk}(x_1)\}(2a-1)(2y-1)\left\{\left(\tfrac{h}{k^{1/d}}\right)^{2s}\sum_{j=1}^{k} B\left(\tfrac{x - m_{j1}}{h/2k^{1/d}}\right)^2 - \tfrac{h^\gamma}{4} B\left(\tfrac{x - x_1}{h}\right)\right\} \\ &- c_{hm}\mathbb{1}\{x \in \mathcal{S}_{hk}(x_{1+m})\}(2a-1)(2y-1)\left\{\left(\tfrac{h}{k^{1/d}}\right)^{2s}\sum_{j=1}^{k} B\left(\tfrac{x - m_{j1+m}}{h/2k^{1/d}}\right)^2 - \tfrac{h^\gamma}{4} B\left(\tfrac{x - x_{1+m}}{h}\right)\right\} \\ &= 0\end{aligned}$$

By Lemma 26, we conclude that

$$H^2(p^n_{\omega_1}, p^n_{\omega_0}) \leq \frac{C}{\epsilon^2}\frac{n^2h^{2d}}{k}\left\{\left(\frac{h}{k^{1/d}}\right)^{4s} + \left(\frac{h}{k^{1/d}}\right)^{4\alpha}\right\} \leq \frac{2C}{\epsilon^2} \cdot n^2 h^{2d+4s} k^{-1-4s/d}$$

because $\alpha \geq s$. This bound on the Hellinger distance actually holds for any $p^n_{\omega'}$ and p^n_{ω} such that $\rho(\omega', \omega) = 1$. Finally, recall that $\xi\gamma \leq d$ by assumption. Set

$$m = c_m h^{-d+\xi\gamma}, \quad k = c_k n^{\frac{d}{2s}\cdot\frac{\gamma-2s}{\gamma}\cdot\frac{1}{1+d/(4s)+d/(2\gamma)}}, \quad h = \left(4c_k^{-\frac{2s}{d}}\right)^{\frac{1}{\gamma-2s}} n^{-\frac{1}{\gamma\{1+d/(4s)+d/(2\gamma)\}}}$$

Notice that this choice enforces $h^{\gamma-2s} = 4k^{-2s/d}$. Therefore,

$$n^2 h^{2d+4s} k^{-1-4s/d} = \left(4c_k^{-\frac{2s}{d}}\right)^{\frac{2d+4s}{\gamma-2s}} \cdot c_k^{-1-4s/d} \cdot n^2 \cdot n^{-\frac{2d+4s}{\gamma\{1+d/(4s)+d/(2\gamma)\}}} \cdot n^{-\frac{d}{2s}\cdot\frac{\gamma-2s}{\gamma}\cdot\frac{1+4s/d}{1+d/(4s)+d/(2\gamma)}}$$
$$= \left(4c_k^{-\frac{2s}{d}}\right)^{\frac{2d+4s}{\gamma-2s}} \cdot c_k^{-1-4s/d}$$

We can choose c_k large enough so that $k \geq 1$, $h \leq 1$ and the leading constant in the equation above is less than $\epsilon^2/(2C)$. This way, the bound on the Hellinger distance is less than or equal to 1. Finally, we verify that c_{hm} is finite. We need $2m$ disjoint cubes with sides equal to $2h$, so m needs to satisfy $m \leq (2h)^{-d}/2$ or, equivalently, $c_m h^{\gamma\xi} \leq 2^{-d-1}$. Because we can choose $h \leq 1$, choosing $c_m = 2^{-d-2}(2^d - 2^{-d})^{-1}$ satisfies this requirement and yields

$$c_{hm} = \{1 - 2(2^d - 2^{-d})h^d m)\}^{-1} \leq \{1 - 2(2^d - 2^{-d})c_m)\}^{-1} = \left\{1 - \frac{1}{2^{d+1}}\right\}^{-1}$$
$$\implies 1 \leq c_{hm} \leq \frac{4}{3}$$

Appendix D

Appendix for Chapter 5

D.1 Proof of Proposition 7

Suppose we observe two samples of n iid observations from $\mathbb{P}$, say D^n and Z^n. Denote $Z_1, \ldots, Z_n$ the observations in Z^n that are iid copies of a generic random variable Z. Let U denote a generic random variable such that $U \subset Z$. For example, in the dose-response settings, $Z = (Y, A, X)$ and $U = A$. Let $\theta_0(u) = \mathbb{E}\{f(Z) \mid U = u\}$ denote the true regression function that needs to be estimated. Recall that $\|f\|^2 = \int f(z)^2 d\mathbb{P}(z) = \mathbb{P}\{f(Z)^2\}$ and $\theta^* \in \operatorname{argmin}_{\theta \in \Theta} \|\theta - \theta_0\|^2$, a fixed function. Let $\widehat{f}(\cdot)$ denote an estimate of $f(\cdot)$ constructed using only observations in D^n. Let $\mathbb{P}_n$ denote the empirical average over sample Z^n. The estimator of $\theta_0(\cdot)$ considered is

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \{\widehat{f}(Z_i) - \theta(U_i)\}^2 \equiv \operatorname*{argmin}_{\theta \in \Theta} \mathbb{P}_n \{\widehat{f}(Z) - \theta(U)\}^2.$$

Finally, let $\widehat{r}(u) = \mathbb{E}\{\widehat{f}(Z) \mid U = u, D^n\} - \theta_0(u)$.

The statement of the theorem follows after proving

$$\mathbb{E}(\|\widehat{\theta} - \theta_0\|^2 \mid D^n) \lesssim \|\widehat{r}\|^2 + \|\theta^* - \theta_0\|^2 + \delta_n^2 \tag{D.1}$$

Our proof is a specialization of that of Theorem 3 in Foster and Syrgkanis [2019]. A useful reference for the arguments made in their proof is Chapter 14 in Wainwright [2019]. To prove (D.1), we need two lemmas.

Lemma 27. *The following inequality holds:*

$$\|\widehat{\theta} - \theta_0\|^2 \leq 8\|\widehat{r}\|^2 + 3\|\theta^* - \theta_0\|^2 - 2(\mathbb{P}_n - \mathbb{P}) \left[\{\widehat{f}(Z) - \widehat{\theta}(U)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2\right]$$

Proof. Notice that

$$\mathbb{P}\left[\{\widehat{f}(Z)-\widehat{\theta}(U)\}^2-\{\widehat{f}(Z)-\theta^*(U)\}^2\right]=-2\mathbb{P}[\widehat{r}(U)\{\widehat{\theta}(U)-\theta^*(U)\}]+\|\widehat{\theta}-\theta_0\|^2-\|\theta^*-\theta_0\|^2$$

By the AM-GM inequality we have, for any $\kappa>0$[1]:

$$\begin{aligned}2\widehat{r}(U)\{\widehat{\theta}(U)-\theta^*(U)\}&=2\widehat{r}(U)\{\widehat{\theta}(U)-\theta_0(U)\}+2\widehat{r}(U)\{\theta_0(U)-\theta^*(U)\}\\&\leq\frac{2}{\kappa}\widehat{r}(U)^2+\kappa\{\widehat{\theta}(U)-\theta_0(U)\}^2+\kappa\{\theta^*(U)-\theta_0(U)\}^2\end{aligned}$$

By monotonicity of integration, it follows that

$$-2\mathbb{P}[\widehat{r}(U)\{\widehat{\theta}(U)-\theta^*(U)\}]\geq-\frac{2\|\widehat{r}\|^2}{\kappa}-\kappa\|\widehat{\theta}-\theta_0\|^2-\kappa\|\theta^*-\theta_0\|^2$$

Rearranging and choosing $\kappa=1/2$, we have

$$\|\widehat{\theta}-\theta_0\|^2\leq 8\|\widehat{r}\|^2+3\|\theta^*-\theta_0\|^2+2\mathbb{P}\left[\{\widehat{f}(Z)-\widehat{\theta}(U)\}^2-\{\widehat{f}(Z)-\theta^*(U)\}^2\right]$$

Because $\mathbb{P}_n[\{\widehat{f}(Z)-\widehat{\theta}(U)\}^2-\{\widehat{f}(Z)-\theta^*(U)\}^2]\leq 0$ since $\theta^*\in\Theta$ and $\widehat{\theta}$ is a minimizer, we also have

$$\|\widehat{\theta}-\theta_0\|^2\leq 8\|\widehat{r}\|^2+3\|\theta^*-\theta_0\|^2-2(\mathbb{P}_n-\mathbb{P})\left[\{\widehat{f}(Z)-\widehat{\theta}(U)\}^2-\{\widehat{f}(Z)-\theta^*(U)\}^2\right]$$

as desired. □

Lemma 28. *For some constant L, let*

$$\mathcal{E}=\left\{\exists\theta\in\Theta:\|\theta-\theta^*\|\geq\delta_n\cap\left|(\mathbb{P}_n-\mathbb{P})\left[\{\widehat{f}(Z)-\theta(U)\}^2-\{\widehat{f}(Z)-\theta^*(U)\}^2\right]\right|\geq L\delta_n\|\theta-\theta^*\|\right\}$$

Under the conditions of Proposition 7, $\mathbb{P}\left(\mathcal{E}\mid D^n\right)\leq c_1\exp(-c_2n\delta_n^2)$ for some constants c_1 and c_2.

Proof. Consider the sets

$$\mathcal{S}_m=\left\{\theta\in\Theta:2^{m-1}\delta_n\leq\|\theta-\theta^*\|\leq 2^m\delta_n\right\}$$

Because $\sup_{\theta\in\Theta}\|\theta\|_\infty\leq S$, $\|\theta-\theta^*\|\leq 2S$ for any $\theta\in\Theta$, which implies that any θ such that $\|\theta-\theta^*\|\geq\delta_n$ must belong to some $\mathcal{S}_m$ for $m\in\{1,\ldots,M\}$, where $M\leq\log_2(2S/\delta_n)$. By a

[1]For any x,y and $\kappa>0$,

$$\left(\frac{x}{\sqrt{2\kappa}}-y\sqrt{\frac{\kappa}{2}}\right)^2\geq 0\implies\frac{x^2}{2\kappa}+y^2\frac{\kappa}{2}\geq xy$$

union bound,

$$\begin{aligned}
\mathbb{P}(\mathcal{E} \mid D^n) &\leq \sum_{m=1}^{M} \mathbb{P}(\mathcal{E} \cap \mathcal{S}_m \mid D^n) \\
&\leq \sum_{m=1}^{M} \mathbb{P}\left(\exists \theta \in \Theta : \|\theta - \theta^*\| \leq 2^m \delta_n \right. \\
&\qquad \left. \cap \left| (\mathbb{P}_n - \mathbb{P}) \left[\{\widehat{f}(Z) - \theta(U)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2 \right] \right| \geq 2^{m-1} L \delta_n^2 \mid D^n \right) \\
&\leq \sum_{m=1}^{M} \mathbb{P}(Z_n(2^m \delta_n) \geq 2^{m-1} L \delta_n^2 \mid D^n)
\end{aligned}$$

where we define

$$Z_n(r) = \sup_{\theta \in \Theta : \|\theta - \theta^*\| \leq r} \left| (\mathbb{P}_n - \mathbb{P}) \left[\{\widehat{f}(Z) - \theta(U)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2 \right] \right|$$

Under the conditions of the proposition, we have

$$\sup_{\theta \in \Theta} \sup_{z \in \mathcal{Z}} \left| \{\widehat{f}(z) - \theta(u)\}^2 - \{\widehat{f}(z) - \theta^*(u)\}^2 \right| \leq 8S^2$$

and

$$\left[\{\widehat{f}(z) - \theta(u)\}^2 - \{\widehat{f}(z) - \theta^*(u)\}^2 \right]^2 \leq 16 S^2 \{\theta(u) - \theta^*(u)\}^2$$

Thus, we have

$$\sigma^2(r) \equiv \sup_{\theta : \|\theta - \theta^*\| \leq r} \mathbb{P}\left(\left[\{\widehat{f}(Z) - \theta(U)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2 \right]^2 \right) \leq 16 S^2 r^2$$

By Theorem 3.27 in Wainwright [2019] and subsequent discussion, viewing $\widehat{f}(\cdot)$ as fixed given D^n, we have

$$\mathbb{P}\left(Z_n(r) \geq \mathbb{E}\{Z_n(r) \mid D^n\} + u \mid D^n \right) \leq 2 \exp\left(- \frac{n u^2}{8e[16 S^2 r^2 + 16 S^2 \mathbb{E}\{Z_n(r) \mid D^n\}] + 32 S^2 u} \right)$$

Next, we bound $\mathbb{E}\{Z_n(r) \mid D^n\}$. By a symmetrization argument, for ϵ a vector of iid Rademacher random variables independent of Z_n and D^n, it holds that

$$\mathbb{E}\{Z_n(r) \mid D^n\} \leq 2 \mathbb{E}_{Z, \epsilon} \left(\sup_{\theta \in \Theta : \|\theta - \theta^*\| \leq r} \left| \mathbb{P}_n \left(\epsilon \left[\{\widehat{f}(Z) - \theta(U)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2 \right] \right) \right| \mid D^n \right)$$

The Ledoux-Talagrand contraction inequality (see also pages 147 and 474 in Wainwright [2019]) yields that, for non-random $x_i \in \mathcal{X}$, a class $\mathcal{F}$ of real-valued functions and a L-Lipschitz

function $\phi : \mathbb{R} \to \mathbb{R}$, the following holds

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{\phi(f(x_i)) - \phi(f^*(x_i))\}\right|\right) \leq 2L\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{f(x_i) - f^*(x_i)\}\right|\right)$$

where $f^* : \mathcal{X} \to \mathbb{R}$ is any function.

Under the boundedness conditions of our proposition, we have

$$\left|\{\widehat{f}(z) - \theta_1(u)\}^2 - \{\widehat{f}(z) - \theta_2(u)\}^2\right| \leq 4S|\theta_1(u) - \theta_2(u)|$$

for any $z \in \mathcal{Z}$. Thus, the square-loss in this case is $4S$-Lipschitz for any $z \in \mathcal{Z}$. By the contraction inequality above, we have

$$\begin{aligned}
&\mathbb{E}_\epsilon\left(\sup_{\theta\in\Theta:\|\theta-\theta^*\|\leq r}\left|\mathbb{P}_n\left(\epsilon\left[\{\widehat{f}(Z) - \theta(u)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2\right]\right)\right|\right) \\
&\leq 2\mathbb{E}_\epsilon\left(\sup_{\theta\in\Theta:\|\theta-\theta^*\|\leq r}|\mathbb{P}_n[\epsilon\{\theta(U) - \theta^*(U)\}]|\right)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{Z,\epsilon}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*\|\leq r}\left|\mathbb{P}_n\left(\epsilon\left[\{\widehat{f}(Z) - \theta(U)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2\right]\right)\right| \,\middle|\, D^n\right) \\
&\leq 8S\mathbb{E}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*\|\leq r}|\mathbb{P}_n[\epsilon\{\theta(U) - \theta^*(U)\}]| \,\middle|\, D^n\right) \\
&\equiv 8S\mathcal{R}_n(\Theta^*, r)
\end{aligned}$$

Next, we have assumed Θ^* to be star-shaped; by Lemma 13.6 in Wainwright [2019] the function $r \mapsto \mathcal{R}_n(\Theta^*, r)/r$ is non-increasing. Therefore, because δ_n solves $\mathcal{R}_n(\Theta^*, \delta) \leq \delta^2$, we also have:

$$\mathcal{R}_n(\Theta^*, r) \leq r\delta_n \quad \text{for all } r \geq \delta_n.$$

Therefore, we conclude that $\mathbb{E}\{Z_n(r) \mid D^n\} \leq 16Sr\delta_n$ for all $r \geq \delta_n$.

Putting everything together, we have derived that

$$\mathbb{P}\left(Z_n(r) \geq 16Sr\delta_n + u \mid D^n\right) \leq 2\exp\left(-\frac{nu^2}{8e(16S^2r^2 + 16^2S^3r^2) + 32S^2u}\right)$$

Let $L = 34S$; specializing this bound to our setting with $r = 2^m\delta_n$ and $u = S2^m\delta_n^2$, we have

$$\mathbb{P}\left(Z_n(2^m\delta_n) \geq L \cdot 2^{m-1}\delta_n^2 \mid D^n\right) \leq 2\exp\left(-\frac{n\delta_n^2}{8e(16+16^2S)+32S}\right)$$

since $2^{-m} \leq 1$ for any $m \geq 1$. Finally,

$$\begin{aligned}\mathbb{P}(\mathcal{E} \mid D^n) &\leq \sum_{m=1}^{M} \mathbb{P}\left(Z_n(2^m\delta_n) \geq L2^{m-1}\delta_n^2 \mid D^n\right) \\ &\leq 2\exp\left(-\frac{n\delta_n^2}{8e(16+16^2S)+32S} + \ln M\right)\end{aligned}$$

Recall that $M \leq \log_2(2S/\delta_n) \leq \log_2(2S\sqrt{2n})$ because we have assumed $\delta_n \geq 1/\sqrt{2n}$. Therefore, if

$$\delta_n^2 \geq \frac{2\ln\{\log_2(2S\sqrt{2n})\}\{8e(16+16^2S)+32S\}}{n}$$

we can conclude

$$\mathbb{P}(\mathcal{E} \mid D^n) \leq 2\exp\left(-\frac{n\delta_n^2}{16e(16+16^2S)+64S}\right)$$

as desired. □

D.1.1 Proof of Equation (D.1)

Notice that Lemma 28 implies that, with probability at least $1 - c_1\exp(-c_2n\delta_n^2)$, either of the following two events occur:

1. Event 1:

$$\|\widehat{\theta} - \theta^*\| \leq \delta_n \implies \|\widehat{\theta} - \theta_0\| \leq \delta_n + \|\theta^* - \theta_0\| \implies \|\widehat{\theta} - \theta_0\|^2 \leq 2\delta_n^2 + 2\|\theta^* - \theta_0\|^2$$

2. Event 2:

$$\begin{aligned}&\left|(\mathbb{P}_n - \mathbb{P})\left[\{\widehat{f}(Z) - \widehat{\theta}(U)\}^2 - \{\widehat{f}(Z) - \theta^*(U)\}^2\right]\right| \leq L\delta_n\|\widehat{\theta} - \theta^*\| \\ &\leq \frac{L^2\delta_n^2}{\kappa} + \frac{\kappa\|\widehat{\theta} - \theta_0\|^2}{2} + \frac{\kappa\|\theta^* - \theta_0\|^2}{2}\end{aligned}$$

   for any $\kappa > 0$.

Because of the result from Lemma 27, Event 2 (with $\kappa = 1/2$) implies

$$\|\widehat{\theta} - \theta_0\|^2 \leq 16\|\widehat{r}\|^2 + 7\|\theta^* - \theta_0\|^2 + 8L^2\delta_n^2$$

This means that there exists a constant C such that

$$\mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 \leq C\left(\|\widehat{r}\|^2+\|\theta^*-\theta_0\|^2+\delta_n^2\right) \mid D^n\right) \geq 1-c_1\exp(-c_2 n\delta_n^2)$$

Let $t_0 = C\left(\|\widehat{r}\|^2+\|\theta^*-\theta_0\|^2+\delta_n^2\right)$. This implies that

$$\begin{aligned}
\mathbb{E}\left(\|\widehat{\theta}-\theta_0\|^2 \mid D^n\right) &= \int_0^\infty \mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 > t \mid D^n\right) dt \\
&= \int_0^{t_0} \mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 > t \mid D^n\right) dt + \int_{t_0}^\infty \mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 > t \mid D^n\right) dt \\
&= \int_0^{t_0} \mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 > t \mid D^n\right) dt + \int_0^\infty \mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 > t_0+t \mid D^n\right) dt \\
&\leq t_0 + \int_0^\infty c_1\exp(-c_3 n t)dt \\
&= t_0 + \frac{c_1}{c_3 n}
\end{aligned}$$

as desired. The last inequality holds because $\mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 > t \mid D^n\right) \leq 1$ and because, whenever δ_n satisfies $\mathcal{R}_n(\delta_n;\Theta^*)/\delta_n \leq \delta_n$, then so does $\delta_n' = \sqrt{\delta_n^2 + t/C} > \delta_n$. This means that we can write

$$t_0 + t = C\left(\|\widehat{r}\|^2+\|\theta^*-\theta_0\|^2+{\delta_n'}^2\right)$$

Thus,

$$\mathbb{P}\left(\|\widehat{\theta}-\theta_0\|^2 > t_0+t \mid D^n\right) \leq c_1\exp\{-c_2 n(\delta_n^2+t/C)\} \leq c_1\exp(-c_3 n t)$$

as Lemma 28 holds for any δ_n' that solves $\mathcal{R}_n(\delta;\Theta^*)/\delta \leq \delta$.

D.2 Proof of Proposition 8

The proof of this theorem is based on Proposition 1 and Theorem 1 from Kennedy [2020]. Their Theorem 1 together with consistency of $\widehat{f}(z)$ yields that

$$\begin{aligned}
|\widehat{\theta}(t)-\theta_0(t)| \leq |\widetilde{\theta}(t)-\theta_0(t)| &+ \left|\frac{1}{n}\sum_{i=1}^n W_i(t;U^n)\mathbb{E}\{\widehat{f}(Z_i)-f(Z_i) \mid D^n, U_i\}\right| \\
&+ o_{\mathbb{P}}\left(\mathbb{E}\left[\left\{\widetilde{\theta}(t)-\theta_0(t)\right\}^2\right]\right)
\end{aligned}$$

Under the assumptions of Proposition 8 (localized weights):

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i(t;U^n)\widehat{r}(U_i)\right| \leq \frac{1}{n}\sum_{i=1}^{n} |W_i(t;U^n)|\,|\widehat{r}(U_i)|\mathbb{1}\{U_i \in N(t)\} \lesssim \sup_{a\in N(t)} |\widehat{r}(u)|$$

as desired.

D.3 Proof of Theorem 5

The proof of this theorem essentially follows from that of Theorem 8.1 in Robins et al. [2017a], with the main difference that our estimator has $K_{ht}(a)$ in place of $\mathbb{1}(A = t)$ so that our analysis will need to keep track of terms of order $O(h^{\alpha\wedge\beta})$.

To simplify the notation, we let $v(a,x) = \mu(a,x) - \widehat{\mu}(a,x), h(a,x) = 1/\pi(a \mid x)$, $q(a,x) = \widehat{h}(a,x) - h(a,x)$, $v(x) = v(t,x)$, $q(x) = q(t,x)$, and $g(x) = \int K_{ht}(a)p(a,x)da$. Also we define $\|f\|_g^2 = \int f^2(x)g(x)dx$.

Before computing bias and variance of our estimator, we state some useful facts about orthogonal projections. More general versions of these statements can be found in the excellent supplementary material to Robins et al. [2017a]. First, recall the definition of the orthogonal projection and its kernel in our context. For $g(x) = \int K_{ht}(a)p(a,x)da$, $\widehat{g}(x) = \int K_{ht}(a)\widehat{p}(a,x)da$ and some function f:

$$\Pi(f)(x_i) = \int \Pi_{i,j} f(x_j)g(x_j)dx_j = b(x_i)^T\Omega^{-1}\int b(x_j)f(x_j)g(x_j)dx_j$$
$$\widehat{\Pi}(f)(x_i) = \int \widehat{\Pi}_{i,j} f(x_j)\widehat{g}(x_j)dx_j = b(x_i)^T\widehat{\Omega}^{-1}\int b(x_j)f(x_j)\widehat{g}(x_j)dx_j$$
$$\Omega = \int b(u)b(u)^T g(u)du \text{ and } \widehat{\Omega} = \int b(u)b(u)^T\widehat{g}(u)du.$$

- Fact 1. Orthogonal projections do not increase length: for any function f and projection in $L_2(\mu)$,

  $$\|\Pi(f)\|^2 = \int \Pi(f)(x)\Pi(f)(x)d\mu = \int \Pi(f)(x)f(x)d\mu \leq \|\Pi(f)\|\|f\| \implies \|\Pi(f)\| \leq \|f\|$$

  by Cauchy-Schwarz, where $\|f\|^2 = \int f^2 d\mu$ and

  $$\Pi(f)(x) = b(x)^T\left\{\int b(u)b(u)^T d\mu\right\}^{-1}\int b(u)f(u)d\mu.$$

- Fact 2. Let w denote some positive and bounded weight function and Π_w and Π projections in $L_2(\mu)$ onto some fixed k-dimensional space L spanned by $b_1(x),\ldots,b_k(x)$,

with weights w and 1 respectively. Then, for any $l \in L$, we have $\Pi_w(l) = \Pi(l)$:

$$l(x) = b(x)^T\beta = \Pi_w(l)(x) = b(x)^T\left\{\int b(u)b(u)^T w(u)d\mu\right\}^{-1}\int b(u)b(u)^T\beta w(u)d\mu$$
$$= b(x)^T\left\{\int b(u)b(u)^T d\mu\right\}^{-1}\int b(u)b(u)^T\beta d\mu = \Pi(l)(x)$$

where $\beta \in \mathbb{R}^k$ is some vector of coefficients.

- Fact 3. Useful identities:

$$\int \Pi(x_i,x_j)\Pi(x_j,x_k)g(x_j)dx_j = b(x_i)^T\Omega^{-1}\int b(x_j)b(x_j)^T g(x_j)dx_j\Omega^{-1}b(x_k) = \Pi(x_i,x_k)$$
$$\int \widehat{\Pi}(x_i,x_j)\widehat{\Pi}(x_j,x_k)\widehat{g}(x_j)dx_j = b(x_i)^T\widehat{\Omega}^{-1}\int b(x_j)b(x_j)^T\widehat{g}(x_j)dx_j\widehat{\Omega}^{-1}b(x_k) = \widehat{\Pi}(x_i,x_k)$$
$$\int \Pi(x_i,x_j)\widehat{\Pi}(x_j,x_k)g(x_j)dx_j = b(x_i)^T\Omega^{-1}\int b(x_j)b(x_j)^T g(x_j)dx_j\widehat{\Omega}^{-1}b(x_k) = \widehat{\Pi}(x_i,x_k)$$

D.3.1 Bias

We will divide the proof of the bias bound in several steps:

1. Prove that, for some functions r_1 and r_2 (defined in the proof) and

$$T = -\int r_1(x_1)\Pi(x_1,x_2)r_2(x_2)g(x_1)g(x_2)dx_1dx_2$$

the following holds

$$\left|\int \widehat{\varphi}_1(z)d\mathbb{P}(z) - \theta(t) + T\right| \lesssim \|(I-\Pi)(v)\|_g\|(I-\Pi)(q)\|_g + h^{\alpha\wedge\beta}$$

2. Prove that

$$\sum_{j=2}^m \int \widehat{\varphi}_j(z_1,\ldots,z_j)d\mathbb{P}(z_1,\ldots,z_j) - T$$
$$= (-1)^{m-1}\int r_1(x_1)(\widehat{\Pi}_{1,2}-\Pi_{1,2})\cdots(\widehat{\Pi}_{m-1,m}-\Pi_{m-1,m})r_2(x_m)g(x_1)\cdots g(x_m)dx_1\cdots dx_m$$
$$\equiv T_2$$

3. Prove that

$$|T_2| \lesssim \|r_1\|_g\|r_2\|_g\|\widehat{s}-1\|_\infty^{m-1} \lesssim \left(\|v\|_g\|q\|_g + h^{\alpha\wedge\beta}\right)\|\widehat{s}-1\|_\infty^{m-1}$$

where $\widehat{s} = g/\widehat{g}$.

Step 1

Let us define

$$\Delta_1(x) \equiv \int K_{ht}(a)\{v(a,x) - v(t,x)\}\pi(a \mid x)da$$
$$\Delta_2(x) \equiv \int K_{ht}(a)\{\mu(a,x) - \mu(t,x)\}\pi(a \mid x)da$$
$$\Delta_3(x) \equiv \int K_{ht}(a)\{\widehat{h}(a,x) - \widehat{h}(t,x)\}\pi(a \mid x)da$$
$$\Delta_4(x) \equiv \int K_{ht}(a)\pi(a \mid x)da - \pi(t \mid x)$$
$$\Delta_5(x) \equiv h(t,x) - \frac{1}{\int K_{ht}(a)\pi(a \mid x)da} = \frac{\Delta_4(x)}{\pi(t \mid x)\{\pi(t \mid x) + \Delta_4(x)\}}$$

We have

$$\int \widehat{\varphi}_1(z)d\mathbb{P}(z) - \theta(t) = \int K_{ht}(a)\widehat{h}(t,x)\{\mu(a,x) - \widehat{\mu}(t,x)\}\pi(a \mid x)dap(x)dx - \int v(x)p(x)dx$$
$$= \int K_{ht}(a)\widehat{h}(t,x)\{\mu(a,x) - \widehat{\mu}(t,x)\}\pi(a \mid x)dap(x)dx - \int \frac{v(x)}{\int K_{ht}(a)\pi(a \mid x)da}g(x)dx$$
$$= \int v(x)q(x)g(x)dx + \int v(x)\Delta_5(x)g(x)dx + \int \widehat{h}(t,x)\Delta_2(x)p(x)dx$$

Let

$$r_1(x) = v(x) + \frac{\Delta_1(x)}{\int K_{ht}(a)\pi(a \mid x)da}, \text{ and } r_2(x) = q(x) + \Delta_5(x) + \frac{\Delta_3(x)}{\int K_{ht}(a)\pi(a \mid x)da}$$

and notice that

$$\int K_{ht}(a)\{y - \widehat{\mu}(a,x)\}d\mathbb{P}(z \mid x) = r_1(x)\int K_{ht}(a)\pi(a \mid x)da$$
$$\int \{K_{ht}(a)\widehat{h}(a,x) - 1\}d\mathbb{P}(z \mid x) = r_2(x)\int K_{ht}(a)\pi(a \mid x)da$$

Define

$$T \equiv -\int r_1(x_1)\Pi(x_1,x_2)r_2(x_2)g(x_1)dx_1g(x_2)dx_2$$
$$= -\int v(x)\Pi(q)(x)g(x)dx - \int \frac{\Delta_1(x)}{\int K_{ht}(a)\pi(a \mid x)da}\Pi(r_2)(x)g(x)dx$$
$$-\int \Pi(v)(x)\left\{\Delta_5(x) + \frac{\Delta_3(x)}{\int K_{ht}(a)\pi(a \mid x)da}\right\}g(x)dx$$

Therefore,

$$\int \widehat{\varphi}_1(z)d\mathbb{P}(z) + T - \theta(t) = \int v(x)(I-\Pi)(q)(x)g(x)dx + \Delta$$

where Δ groups all the terms involving Δ_j together:

$$\begin{aligned}\Delta = &\int v(x)\Delta_5(x)g(x)dx + \int \widehat{h}(t,x)\Delta_2(x)p(x)dx \\ &- \int \frac{\Delta_1(x)}{\int K_{ht}(a)\pi(a \mid x)da}\Pi(r_2)(x)g(x)dx \\ &- \int \Pi(v)(x)\left\{\Delta_5(x) + \frac{\Delta_3(x)}{\int K_{ht}(a)\pi(a \mid x)da}\right\}g(x)dx\end{aligned}$$

The term Δ is controlled under the smoothness assumptions of the theorem, while

$$\left|\int v(x)(I-\Pi)(q)(x)g(x)dx\right| \leq \|(I-\Pi)(v)\|_g \|(I-\Pi)(q)\|_g$$

by Cauchy-Schwarz. In particular, we have assumed that $a \mapsto \mu(a,x)$, $a \mapsto \widehat{\mu}(a,x)$ are α-times continuously differentiable and $a \mapsto h(a,x)$, $a \mapsto \widehat{h}(a,x)$ (or, equivalently, $\pi(a \mid x)$ and $\widehat{\pi}(a \mid x)$) are β-times continuously differentiable. Thus, we have

$$\begin{aligned}v(a,x) &= \sum_{j=0}^{\alpha-1} v^{(j)}(t,x)\frac{(a-t)^j}{j!} + v^{(\alpha)}(t+\tau_1(a-t),x)\frac{(a-t)^\alpha}{\alpha!} \\ \mu(a,x) &= \sum_{j=0}^{\alpha-1} \mu^{(j)}(t,x)\frac{(a-t)^j}{j!} + \mu^{(\alpha)}(t+\tau_2(a-t),x)\frac{(a-t)^\alpha}{\alpha!} \\ \pi(a \mid x) &= \sum_{j=0}^{\beta-1} \pi^{(j)}(t \mid x)\frac{(a-t)^j}{j!} + \pi^{(\beta)}(t+\tau_3(a-t) \mid x)\frac{(a-t)^\beta}{\beta!} \\ h(a,x) &= \sum_{j=0}^{\beta-1} h^{(j)}(t,x)\frac{(a-t)^j}{j!} + h^{(\beta)}(t+\tau_4(a-t),x)\frac{(a-t)^\beta}{\beta!}\end{aligned}$$

for some $\tau_1, \tau_2, \tau_3, \tau_4 \in [0,1]$. Then, for example, we have the following

$$\begin{aligned}
\Delta_1(x) &\equiv \int K_{ht}(a)\{v(a,x) - v(t,x)\}\pi(a \mid x)da \\
&= \sum_{i=1}^{\alpha-1}\sum_{j=0}^{\beta-1} h^{i+j}\frac{v^{(i)}(t,x)}{i!}\frac{\pi^{(j)}(t \mid x)}{j!}\int_0^1 u^{i+j}K(u)du \\
&\quad + \sum_{j=1}^{\alpha-1} h^{\beta+j}\frac{v^{(j)}(t,x)}{\beta!j!}\int_0^1 u^{\beta+j}K(u)\pi^{(\beta)}(t+\tau_3 uh \mid x)du \\
&\quad + \sum_{j=0}^{\beta-1} h^{\alpha+j}\frac{\pi^{(j)}(t \mid x)}{\alpha!j!}\int_0^1 v^{(\alpha)}(t+\tau_1 uh, x)u^{\alpha+j}K(u)du \\
&\quad + \frac{h^{\alpha+\beta}}{\alpha!\beta!}\int_0^1 v^{(\alpha)}(t+\tau_1 uh, x)u^{\alpha+\beta}K(u)\pi^{(\beta)}(t+\tau_3 uh \mid x)du
\end{aligned}$$

so that $\|\Delta_1\|_\infty \lesssim h^{(\beta+1)\wedge\alpha}$. Similarly, $\|\Delta_2\|_\infty \lesssim h^{(\beta+1)\wedge\alpha}$ and

$$\begin{aligned}
\Delta_3(x) &\equiv \int K_{ht}(a)\{\widehat{h}(a,x) - \widehat{h}(t,x)\}\pi(a \mid x)da \\
&= \sum_{i=1}^{\beta-1}\sum_{j=0}^{\beta-1} h^{i+j}\frac{\widehat{h}^{(i)}(t,x)}{i!}\frac{\pi^{(j)}(t \mid x)}{j!}\int_0^1 u^{i+j}K(u)du \\
&\quad + \sum_{j=1}^{\beta-1} h^{\beta+j}\frac{\widehat{h}^{(j)}(t,x)}{\beta!j!}\int_0^1 u^{\beta+j}K(u)\pi^{(\beta)}(t+\tau_3 uh \mid x)du \\
&\quad + \sum_{j=0}^{\beta-1} h^{\beta+j}\frac{\pi^{(j)}(t \mid x)}{\beta!j!}\int_0^1 \widehat{h}^{(\beta)}(t+\tau_1 uh, x)u^{\beta+j}K(u)du \\
&\quad + \frac{h^{2\beta}}{\beta!\beta!}\int_0^1 v^{(\beta}(t+\tau_1 uh, x)u^{2\beta}K(u)\pi^{(\beta)}(t+\tau_3 uh \mid x)du
\end{aligned}$$

Therefore, $\|\Delta_3\|_\infty \lesssim h^\beta$ and, similarly, $\|\Delta_4\|_\infty \lesssim h^\beta$ and $\|\Delta_5\|_\infty \lesssim h^\beta$. In this light, it holds that $\|\Delta_j\|_\infty \lesssim h^{\alpha\wedge\beta}$, for $j = 1,2,3,4,5$. This concludes our proof that

$$\left|\int \widehat{\varphi}_1(z)d\mathbb{P}(z) + T - \theta(t)\right| \lesssim \|(I-\Pi)(v)\|_g\|(I-\Pi)(q)\|_g + h^{\alpha\wedge\beta}$$

Step 2

We will show that

$$\begin{aligned}
T_2 &= \sum_{j=2}^{m} \int \widehat{\varphi}_j(z_1, \ldots, z_j) d\mathbb{P}(z_1, \ldots, z_j) - T \\
&= \sum_{j=2}^{m} \int \widehat{\varphi}_j(z_1, \ldots, z_j) d\mathbb{P}(z_1, \ldots, z_j) + \int r_1(x) \Pi(x_1, x_2) r_2(x_2) g(x_1) g(x_2) dx_1 dx_2 \\
&= (-1)^{m-1} \int r_1(x_1)(\widehat{\Pi}_{1,2} - \Pi_{1,2}) \cdots (\widehat{\Pi}_{m-1,m} - \Pi_{m-1,m}) r_2(x_m) g(x_1) \cdots g(x_m) dx_1 \cdots dx_m
\end{aligned}$$

The result is clearly true for $m = 2$, so we proceed by induction. Relative to the m^{th} term, the term $m + 1$ receives the contribution from

$$\begin{aligned}
&\int \widehat{\varphi}_{m+1}(z_1, \ldots, z_{m+1}) d\mathbb{P}(z_1, \ldots, z_m) \\
&= (-1)^m \sum_{i=0}^{m-1} \binom{m-1}{i} (-1)^i \int r_1(x_1) \widehat{\Pi}_{1,2} \cdots \widehat{\Pi}_{m-i,m-i+1} r_2(x_{m-i+1}) g(x_1) \cdots g(x_{m-i+1}) dx_1 \cdots dx_{m-i+1} \\
&\equiv (-1)^m T_3
\end{aligned}$$

Thus to prove the claim we need to show that

$$\begin{aligned}
T_3 &= \int r_1(x_1)(\widehat{\Pi}_{1,2} - \Pi_{1,2}) \cdots (\widehat{\Pi}_{m-1,m} - \Pi_{m-1,m}) r_2(x_m) g(x_1) \cdots g(x_m) dx_1 \cdots dx_m \\
&\quad + \int r_1(x_1)(\widehat{\Pi}_{1,2} - \Pi_{1,2}) \cdots (\widehat{\Pi}_{m,m+1} - \Pi_{m,m+1}) r_2(x_{m+1}) g(x_1) \cdots g(x_{m+1}) dx_1 \cdots dx_{m+1} \\
&\equiv T_4 + T_5
\end{aligned}$$

Notice that T_4 can be written as a sum of terms of the form

$$B_l = (-1)^{m-1-l} \int r_1(x_1) B_{1,2} \cdots B_{m-1,m} r_2(x_m) g(x_1) \cdots g(x_m) dx_1 \cdots dx_m$$

where $B_{i,j}$ equals either $\widehat{\Pi}_{i,j}$ or $\Pi_{i,j}$ and l denotes the number of terms in the product for which $B_{i,j} = \widehat{\Pi}_{i,j}$. Similarly, T_5 is a sum of terms of the form

$$C_l = (-1)^{m-l} \int r_1(x_1) B_{1,2} \cdots B_{m,m+1} r_2(x_{m+1}) g(x_1) \cdots g(x_{m+1}) dx_1 \cdots dx_{m+1}$$

Fact 3 is the reason why we only need to keep track of the number of $\widehat{\Pi}_{i,j}$ terms and not specifically which B_{ij} equals $\Pi_{i,j}$ or $\widehat{\Pi}_{i,j}$. In fact, for $B_{ij} = \Pi_{i,j}$ or $\widehat{\Pi}_{i,j}$, we have

$$\int \Pi(x_{j-1}, x_j) B(x_j, x_{j+1}) g(x_j) dx_j = B_{j-1,j+1} = \int B(x_{j-1}, x_j) \Pi(x_j, x_{j+1}) g(x_j) dx_j$$

In this light, we can simplify as

$$B_l = (-1)^{m-1-l} \int r_1(x_1)\widehat{\Pi}_{1,2}\cdots\widehat{\Pi}_{l,l+1}r_2(x_{l+1})dg(x_1)\cdots dg(x_{l+1}) \text{ for } l \geq 1$$
$$B_0 = (-1)^{m-1} \int r_1(x_1)\Pi_{1,2}r_2(x_2)dg(x_1)d(x_2)$$
$$C_l = (-1)^{m-l} \int r_1(x_1)\widehat{\Pi}_{1,2}\cdots\widehat{\Pi}_{l,l+1}r_2(x_{l+1})dg(x_1)\cdots dg(x_{l+1}) \text{ for } l \geq 1$$
$$C_0 = (-1)^{m} \int r_1(x_1)\Pi_{1,2}r_2(x_2)dg(x_1)d(x_2) = -B_0$$

For $l \in \{1, \ldots, m-1\}$, we have $C_l = -B_l$. Thus, we have reached

$$T_4 = -C_0 - \sum_{l=1}^{m-1} \binom{m-1}{l} C_l \quad \text{and} \quad T_5 = C_0 + \sum_{l=1}^{m-1} \binom{m}{l} C_l + C_m$$

and this implies

$$T_4 + T_5 = \sum_{l=1}^{m-1} \left\{ \binom{m}{l} - \binom{m-1}{l} \right\} C_l + C_m = \sum_{l=1}^{m-1} \binom{m-1}{l-1} C_l + C_m = T_3$$

as desired. We have thus shown that

$$T_2 = (-1)^{m-1} \int r_1(x_1)(\widehat{\Pi}_{1,2} - \Pi_{1,2})\cdots(\widehat{\Pi}_{m-1,m} - \Pi_{m-1,m})r_2(x_m)g(x_1)\cdots g(x_m)dx_1\cdots dx_m$$

Step 3

We need to show that $|T_2| \leq \|r_1\|_g \|r_2\|_g \|\widehat{s} - 1\|_\infty^{m-1}$. This statement is essentially a direct consequence of Lemma 13.7 in the Supplementary material to Robins et al. [2017a]. For the sake of completeness, we give a proof here that is less general (and more verbose) than that in Robins et al. [2017a], although it uses the same arguments.

Define $\widehat{s} = g(x)/\widehat{g}(x)$ and let $M_{\widehat{s}}$ denoting multiplication by $\widehat{s}$. We have

$$\int (\widehat{\Pi}_{m-1,m} - \Pi_{m-1,m})r_2(x_m)g(x_m)dx_m = \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)(x_{m-1})$$

Continuing with this calculation, we get

$$T_2 = (-1)^{m-1} \int r_1(x) \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)^{m-1} (r_2)(x)g(x)dx$$

Let $\|f\|^2_{2,\widehat{g}} = \int f^2(u)\widehat{g}(u)du$ and bound $|T_2|$ as

$$|T_2| \lesssim \|r_1\|_g \left\| \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)^{m-1}(r_2) \right\|_{2,\widehat{g}}$$

Define

$$l(x) = \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)^{m-2}(r_2)(x) \equiv b(x)^T\beta$$

We can write $l(x)$ as a linear combination of the truncated basis because both $\widehat{\Pi}$ and Π project a function onto the same finite dimensional subspace. Notice that we can view Π as a weighted projection in $L_2(\widehat{g})$ with weight $\widehat{s}$, i.e.

$$\begin{aligned}\Pi(f)(x) &= b(x)^T\Omega^{-1}\int b(u)f(u)g(u)du \\ &= b(x)^T\left\{\int b(u)b(u)\widehat{s}(u)\widehat{g}(u)du\right\}^{-1}\int b(u)f(u)\widehat{s}(u)\widehat{g}(u)du\end{aligned}$$

Therefore, by Fact 2, we have

$$\begin{aligned}\left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)^{m-1}(r_2)(x) &= \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(l)(x) \\ &= \int \widehat{\Pi}(x,u)\{\widehat{s}(u)-1\}l(u)\widehat{g}(u)du \\ &= \widehat{\Pi}\left((\widehat{s}-1)l\right)(x)\end{aligned}$$

By Fact 1, we have

$$\left\| \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)^{m-1}(r_2) \right\|_{2,\widehat{g}} = \left\|\widehat{\Pi}\left((\widehat{s}-1)l\right)\right\|_{2,\widehat{g}} \leq \|(\widehat{s}-1)l\|_{2,\widehat{g}} \leq \|\widehat{s}-1\|_\infty \|l\|_{2,\widehat{g}}$$

Repeating this argument $m-3$ times applied to $\|l\|_{2,\widehat{g}}$, we obtain

$$\left\| \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)^{m-1}(r_2) \right\|_{2,\widehat{g}} \leq \|\widehat{s}-1\|^{m-2}_\infty \left\| \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2) \right\|_{2,\widehat{g}}$$

Furthermore,

$$\begin{aligned}\left\| \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2) \right\|^2_{2,\widehat{g}} &= \int \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)(x) \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)(x)\widehat{g}(x)dx \\ &= \int \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)(x)\Pi\left(r_2\right)(x)\{\widehat{s}(x)-1\}\widehat{g}(x)dx\end{aligned}$$

The second line follows because $\left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)$ belongs to the finite dimensional subspace

and can be expressed as $b(x)^T\beta$ for some β. Therefore,

$$\begin{aligned}
&\int \left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)(x)\widehat{\Pi}M_{\widehat{s}}(r_2)(x)\widehat{g}(x)dx \\
&= \beta^T \int b(x)\widehat{\Pi}M_{\widehat{s}}(r_2)(x)\widehat{g}(x)dx \\
&= \beta^T \int b(x)b(x)^T\widehat{\Omega}^{-1}\int b(u)\widehat{s}(u)r_2(u)\widehat{g}(u)du\widehat{g}(x)dx \\
&= \beta^T \int b(u)r_2(u)g(u)du \\
&= \beta^T \int b(x)b(x)^T\Omega^{-1}\int b(u)r(u)g(u)du\widehat{s}(x)\widehat{g}(x)dx \\
&= \int \beta^T b(x)\Pi(r_2)(x)\widehat{s}(x)\widehat{g}(x)dx
\end{aligned}$$

By Cauchy-Schwarz:

$$\left\|\left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)\right\|_{2,\widehat{g}}^2 \leq \left\|\left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)\right\|_{2,\widehat{g}} \|\Pi(r_2)(\widehat{s}-1)\|_{2,\widehat{g}},$$

implying

$$\left\|\left(\widehat{\Pi}M_{\widehat{s}} - \Pi\right)(r_2)\right\|_{2,\widehat{g}} \lesssim \|\widehat{s}-1\|_\infty \|r_2\|_g.$$

This then yields

$$|T_2| \lesssim \|r_1\|_g \|r_2\|_g \|\widehat{s}-1\|_\infty^{m-1}.$$

The bounds on the terms involving Δ_j derived in Step 1 finally yield the result:

$$|T_2| \lesssim \left(\|v\|_g\|q\|_g + h^{\alpha\wedge\beta}\right)\|\widehat{s}-1\|_\infty^{m-1}$$

D.3.2 Variance

The proof of the variance bound follows as in Robins et al. [2017a]. Because, for two random variables U_1 and U_2, $\text{var}(U_1 + U_2) \leq 2\text{var}(U_1) + 2\text{var}(U_2)$, and because m is fixed, we have:

$$\begin{aligned}
&\text{var}\left\{\mathbb{P}_n\widehat{f}_0(Z) + \sum_{j=2}^{m}\mathbb{U}_n\widehat{\varphi}_j(Z_1,\ldots,Z_j) \mid D^n\right\} \\
&\leq 2\text{var}\left\{\mathbb{P}_n\widehat{f}_0(Z) \mid D^n\right\} + 2\text{var}\left\{\sum_{j=2}^{m}\mathbb{U}_n\widehat{\varphi}_j(Z_1,\ldots,Z_j) \mid D^n\right\} \\
&\lesssim \text{var}\left\{\mathbb{P}_n\widehat{f}_0(Z) \mid D^n\right\} + P^n\left[\left\{\sum_{j=2}^{m}\mathbb{U}_n\widehat{\varphi}_j(Z_1,\ldots,Z_j)\right\}^2\right] \\
&\lesssim \text{var}\left\{\mathbb{P}_n\widehat{f}_0(Z) \mid D^n\right\} + \sum_{j=2}^{m}P^n\left[\{\mathbb{U}_n\widehat{\varphi}_j(Z_1,\ldots,Z_j)\}^2\right]
\end{aligned}$$

Because, given D^n, $\mathbb{P}_n\widehat{f}_0(Z)$ is a sample average of independent observations, we have

$$\text{var}\left\{\mathbb{P}_n\widehat{f}_0(Z) \mid D^n\right\} \leq \frac{1}{n}\mathbb{P}\{\widehat{f}_0^2(Z)\} \lesssim \frac{1}{nh}$$

since $\int K_{ht}^2(a)\pi(a \mid x)da = h^{-1}\int K^2(u)\pi(uh+t)du \lesssim h^{-1}$. By Lemma 14.1 in Robins et al. [2017a], the following holds

$$P^n\left[\{\mathbb{U}_n\widehat{\varphi}_j(Z_1\ldots,Z_j)\}^2\right] \leq 2j\left(1 + \left\|\frac{p}{\widehat{p}}\right\|_\infty\right)^{2j}\widehat{P}^n\left[\{\mathbb{U}_n\widehat{\varphi}_j(Z_1\ldots,Z_j)\}^2\right]$$

Because $\widehat{\varphi}_j$ is degenerate relative to $\widehat{P}$, we also have

$$\begin{aligned}
\widehat{P}^n\left[\{\mathbb{U}_n(\widehat{\varphi}_j(Z_1\ldots,Z_j))\}^2\right] &\lesssim \frac{1}{n(n-1)\cdots(n-j+1)}\widehat{P}^j\left[\{\widehat{\varphi}_j(Z_1\ldots,Z_j)\}^2\right] \\
&\lesssim \frac{1}{n(n-1)\cdots(n-j+1)}\widehat{P}^j\left[\left\{f_1(Z_1)\widehat{\Pi}_{1,2}K_{ht}(A_2)\cdots\widehat{\Pi}_{j-1,j}f_2(Z_j)\right\}^2\right]
\end{aligned}$$

Notice that

$$\left\{f_1(Z_1)\widehat{\Pi}_{1,2}K_{ht}(A_2)\cdots\widehat{\Pi}_{j-1,j}f_2(Z_j)\right\}^2 \lesssim K_{ht}^2(A_1)\widehat{\Pi}_{1,2}^2K_{ht}^2(A_2)\cdots\widehat{\Pi}_{j-1,j}^2K_{ht}^2(A_j)$$

since $h(a,x)$, Y and $\widehat{\mu}(a,x)$ are uniformly bounded by assumption. Next, because

$$\int K_{ht}^2(a)\pi(a \mid x)da = h^{-1}\int K^2(u)\pi(uh+t \mid x)du \lesssim h^{-1},$$

we have

$$\begin{aligned}
&\widehat{P}^j\left[\left\{f_1(Z_1)\widehat{\Pi}_{1,2}K_{ht}(A_2)\cdots\widehat{\Pi}_{j-1,j}f_2(Z_j)\right\}^2\right] \\
&\lesssim h^{-j}\int \widehat{\Pi}_{1,2}^2\cdots\widehat{\Pi}_{j-1,j}^2 p(x_1)\cdots p(x_j)dx_1\cdots dx_j \\
&= h^{-j}\int \widehat{\Pi}_{1,2}^2\cdots\widehat{\Pi}_{j-1,j}^2\left\{\prod_{l=1}^{j}\frac{p(x_l)}{\widehat{g}(x_l)}\right\}\widehat{g}(x_1)\cdots\widehat{g}(x_j)dx_1\cdots dx_j \\
&\leq \left\{\sup_x \frac{p(x)}{\widehat{g}(x)}\right\}^j h^{-j}\int \widehat{\Pi}_{1,2}^2\cdots\widehat{\Pi}_{j-1,j}^2\widehat{g}(x_1)\cdots\widehat{g}(x_j)dx_1\cdots dx_j \\
&\lesssim h^{-j}\int \widehat{\Pi}_{1,2}^2\cdots\widehat{\Pi}_{j-1,j}^2\widehat{g}(x_1)\cdots\widehat{g}(x_j)dx_1\cdots dx_j
\end{aligned}$$

Next, notice that

$$\int \widehat{\Pi}^2(x_i,x_j)\widehat{g}(x_j)dx_j = \int b(x_i)^T\widehat{\Omega}^{-1}b(x_j)b(x_j)^T\widehat{\Omega}^{-1}b(x_i)\widehat{g}(x_j)dx_j = \widehat{\Pi}(x_i,x_i)$$

We bound each term from $i=j$ to $i=3$ as

$$\begin{aligned}
\int \widehat{\Pi}_{i-2,i-1}^2\widehat{\Pi}_{i-1,i}^2\widehat{g}(x_{i-1})\widehat{g}(x_i)dx_{i-1}dx_i &= \int \widehat{\Pi}_{i-2,i-1}^2\widehat{\Pi}_{i-1,i-1}\widehat{g}(x_{i-1})dx_{i-1} \\
&\leq \sup_x \widehat{\Pi}(x,x)\int \widehat{\Pi}_{i-2,i-1}^2\widehat{g}(x_{i-1})dx_{i-1}
\end{aligned}$$

This leads to

$$\int \widehat{\Pi}_{1,2}^2\cdots\widehat{\Pi}_{j-1,j}^2\widehat{g}(x_1)\cdots\widehat{g}(x_j)dx_1\cdots dx_j \leq \left\{\sup_x \widehat{\Pi}(x,x)\right\}^{j-2}\int \widehat{\Pi}_{1,2}^2\widehat{g}(x_1)dx_1\widehat{g}(x_2)dx_2$$

Finally, without loss of generality, let $b(x)$ be scaled so that $\widehat{\Omega}$ is the identity matrix. This way, we immediately have

$$\int \widehat{\Pi}_{1,2}^2\widehat{g}(x_1)\widehat{g}(x_2)dx_1dx_2 = \int b(x)^Tb(x)\widehat{g}(x)dx = \sum_{i=1}^{k}\int b_i^2(x)\widehat{g}(x)dx = k$$

because the basis is orthonormal. Because m is fixed and does not grow with n and $\sup_x \widehat{\Pi}(x,x) \lesssim k$, this yields the bounds in the statement of the theorem.

D.4 Proofs of claims from Section 5.4

D.4.1 Proof of Lemma 13

We prove the result for the upper bound, as that for the lower bound can be proven with a similar argument. By Leibniz rule of integration, the derivative of the map $\overline{q}(a,x) \mapsto \mathbb{E}\{s_u(Z;\overline{q}) \mid A=a, X=x)$ is

$$\begin{aligned}
&\frac{d}{d\overline{q}}\left\{\overline{q} + \frac{1}{\gamma}\int_{-\infty}^{\overline{q}} (y-\overline{q}) f(y \mid A=a, X=x)dy + \gamma\int_{\overline{q}}^{\infty}(y-\overline{q})f(y \mid A=a, X=x)dy\right\} \\
&= 1 - \frac{1}{\gamma}\mathbb{P}\left(Y \leq \overline{q} \mid A=a, X=x\right) - \gamma\mathbb{P}\left(Y \geq \overline{q} \mid A=a, X=x\right)
\end{aligned}$$

Similarly, the second derivative is

$$\begin{aligned}
&\frac{d^2}{d\overline{q}^2}\left\{\overline{q} + \frac{1}{\gamma}\int_{-\infty}^{\overline{q}} (y-\overline{q}) f(y \mid A=a, X=x)dy + \gamma\int_{\overline{q}}^{\infty}(y-\overline{q})f(y \mid A=a, X=x)dy\right\} \\
&= \frac{d}{d\overline{q}}\left\{1 - \frac{1}{\gamma}\int_{-\infty}^{\overline{q}} f(y \mid A=a, X=x)dy - \gamma\int_{\overline{q}}^{\infty} f(y \mid A=a, X=x)dy\right\} \\
&= -\frac{1}{\gamma}f(\overline{q} \mid A=a, X=x) + \gamma f(\overline{q} \mid A=a, X=x) \\
&= O(1)
\end{aligned}$$

Notice that the first derivative vanishes at the true quantile $\overline{q}(a,x) = q_u(a,x)$. Therefore, by a second order Taylor expansion, it holds that

$$|\mathbb{E}\{s(Z;\widehat{q}_u) - s(Z;q_u) \mid A=a, X=x\}| \lesssim \{\widehat{q}_u(a,x) - q_u(a,x)\}^2$$

Next, notice that

$$\begin{aligned}
\widehat{r}_u(t) &= \int \widehat{w}(t,x)[\mathbb{E}\{s(Z;\widehat{q}_u) \mid A=t, X=x\} - \widehat{\kappa}_u(t,x)]d\mathbb{P}(x \mid A=t) \\
&\quad + \int w(t,x)\{\widehat{\kappa}_u(t,x) - \kappa_u(t,x)\}d\mathbb{P}(x \mid A=t) + (\mathbb{P}_n - \mathbb{P})\widehat{\kappa}_u(t,X;\widehat{q}_u) \\
&= \int \widehat{w}(t,x)[\mathbb{E}\{s(Z;\widehat{q}_u) \mid A=t, X=x\} - \kappa_u(t,x)]d\mathbb{P}(x \mid A=t) \\
&\quad + \int \{w(t,x) - \widehat{w}(t,x)\}\{\widehat{\kappa}_u(t,x) - \kappa_u(t,x)\}d\mathbb{P}(x \mid A=t) + (\mathbb{P}_n - \mathbb{P})\widehat{\kappa}_u(t,X;\widehat{q}_u)
\end{aligned}$$

The bound then follows by the Cauchy-Schwarz inequality.

D.4.2 Proof of Proposition 10

We prove the result for the upper bound, as the proof for the lower bound is analogous. Let $\widehat{\mathbb{E}}(\cdot \mid A = t)$ denote the second-stage regression based on linear smoothing. Define

$$\begin{aligned}
&\widetilde{\varphi}_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u, w, q_u) \\
&= \varphi_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u) - w(A, X)[Y - \widehat{q}_u(A, X)] \left[\gamma^{\text{sgn}\{Y - \widehat{q}_u(A,X)\}} - \gamma^{\text{sgn}\{Y - q_u(A,X)\}}\right],
\end{aligned}$$

where

$$\varphi_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u) = \widehat{w}(A, X)\{s_u(Z; \widehat{q}_u) - \widehat{\kappa}_u(A, X; \widehat{q}_u)\} + \frac{1}{n}\sum_{i=1}^{n} \widehat{\kappa}_u(A, X_i; \widehat{q}_u)$$

We have $\varphi_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u) \geq \widetilde{\varphi}_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u, w, q_u)$ and, deterministically by assumption,

$$\widehat{\mathbb{E}}\{\varphi_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u) \mid A = t\} \geq \widehat{\mathbb{E}}\{\widetilde{\varphi}_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u, w, q_u) \mid A = t\}.$$

Let

$$\begin{aligned}
\overline{\varphi}_u(Z; w, \overline{\kappa}_u, \overline{q}_u, q_u) &= \widetilde{\varphi}_u(Z; w, \overline{\kappa}_u, \overline{q}_u, q_u) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\kappa}_u(A, X_i; \widehat{q}_u) + \int \overline{\kappa}_u(A, x; \overline{q}_u) d\mathbb{P}(x) \\
&= w(A, X)\{s_u(Z; \overline{q}_u) - \overline{\kappa}(A, X; \overline{q}_u)\} + \int \overline{\kappa}(A, x; \overline{q}_u) d\mathbb{P}(x) \\
&\quad - w(A, X)[Y - \overline{q}_u(A, X)] \left[\gamma^{\text{sgn}\{Y - \overline{q}_u(A,X)\}} - \gamma^{\text{sgn}\{Y - q_u(A,X)\}}\right]
\end{aligned}$$

and notice that, because $\mathbb{E}[\gamma^{\{Y - q_u(t,x)\}} \mid A = t, X = x] = 1$:

$$\begin{aligned}
&\mathbb{E}\left\{\overline{\varphi}_u(Z; w, \overline{\kappa}_u, \overline{q}_u, q_u) \mid A = t, X = x\right\} \\
&= \int \overline{\kappa}(t, x; \overline{q}_u) d\mathbb{P}(x) - w(t, x)\mathbb{E}\left[\{Y - \overline{q}_u(t, x)\}\gamma^{\text{sgn}\{Y - \overline{q}_u(t,x)\}} \mid A = t, X = x\right] \\
&\quad + w(t, x)\mathbb{E}\left[Y\gamma^{\text{sgn}\{Y - q_u(t,x)\}} \mid A = t, X = x\right] - w(t, x)\overline{q}_u(t, x) \\
&= \int \overline{\kappa}(t, x; \overline{q}_u) d\mathbb{P}(x) - w(t, x)\overline{\kappa}(t, x; \overline{q}_u) + w(t, x)\mathbb{E}\left[Y\gamma^{\text{sgn}\{Y - q_u(t,x)\}} \mid A = t, X = x\right]
\end{aligned}$$

Therefore, $\mathbb{E}\left\{\overline{\varphi}_u(Z; w, \overline{\kappa}_u, \overline{q}_u, q_u) \mid A = t\right\} = \theta_u(t; \gamma)$.

By the reasoning in Kennedy [2020] and used to prove Proposition 8, one has

$$\begin{aligned}
&\widehat{\mathbb{E}}\{\widetilde{\varphi}_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u, q_u) \mid A = t\} - \theta_u(t; \gamma) \\
&= \widehat{\mathbb{E}}\{\widetilde{\varphi}_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u, q_u) \mid A = t\} - \mathbb{E}\{\overline{\varphi}_u(Z; w, \overline{\kappa}_u, \overline{q}_u, q_u) \mid A = t\} \\
&= O_{\mathbb{P}}\left(R_u(t)\right) + \frac{1}{n}\sum_{i=1}^{n} W_i(t; A^n)\mathbb{E}\left\{\widetilde{\varphi}_u(Z; \widehat{w}, \widehat{\kappa}_u, \widehat{q}_u, w, q_u) - \overline{\varphi}_u(Z; w, \overline{\kappa}_u, \overline{q}_u, w, q_u) \mid A_i, D^n\right\}
\end{aligned}$$

provided that $\sup_z |\widetilde{\varphi}_u(z;\widehat{w},\widehat{\kappa}_u,\widehat{q}_u,q_u) - \overline{\varphi}_u(z;w,\overline{\kappa}_u,\overline{q}_u,q_u)| = o_{\mathbb{P}}(1)$. This is the case, because $\widehat{w}$ is consistent for w, $\widehat{\kappa}_u$ is consistent for $\overline{\kappa}_u$ and $\widehat{q}$ is consistent for $\overline{q}$.

Next, recall that

$$\begin{aligned}\theta_u(A;\gamma) &= \mathbb{E}\{\overline{\varphi}_u(Z;w,\overline{\kappa}_u,\overline{q}_u,w,q_u) \mid A\} \\ &= \int W(A,x)\mathbb{E}\{Y\gamma^{\text{sgn}\{Y-q_u(A,X)\}} \mid A, X=x\}d\mathbb{P}(x \mid A)\end{aligned}$$

so that, because $\mathbb{E}[\gamma^{\text{sgn}\{Y-q_u(A,X)\}} \mid A, X] = 1$:

$$\begin{aligned}&\mathbb{E}\left(w(A,X)[Y-\widehat{q}_u(A,X)]\gamma^{\text{sgn}\{Y-q_u(A,X)\}} - \overline{\varphi}_u(Z;w,\overline{\kappa}_u,\overline{q}_u,w,q_u) \mid A_i, D^n\right) \\ &= -\int w(A_i,x)\widehat{q}_u(A_i,x)d\mathbb{P}(x \mid A_i)\end{aligned}$$

In turns, this means that

$$\begin{aligned}&\mathbb{E}\left(-w(A,X)[Y-\widehat{q}_u(A,X)]\left[\gamma^{\text{sgn}\{Y-\widehat{q}_u(A,X)\}} - \gamma^{\text{sgn}\{Y-q_u(A,X)\}}\right]\right. \\ &\qquad \left. - \overline{\varphi}_u(Z;w,\overline{\kappa}_u,\overline{q}_u,w,q_u) \mid A_i, D^n\right) \\ &= -\mathbb{E}\{w(A_i,X)s_u(Z;\widehat{q}_u) \mid A_i, D^n)\end{aligned}$$

yielding

$$\begin{aligned}\widehat{b}(A_i) &\equiv \mathbb{E}\{\widetilde{\varphi}_u(Z;\widehat{w},\widehat{\kappa}_u,\widehat{q}_u,w,q_u) - \overline{\varphi}_u(Z;w,\overline{\kappa}_u,\overline{q}_u,w,q_u) \mid A_i, D^n\} \\ &= \int\{\widehat{w}(A_i,x)-w(A_i,x)\}[\mathbb{E}\{s_u(Z;\widehat{q}_u) \mid A_i,x\} - \widehat{\kappa}_u(A_i,x;\widehat{q}_u)]d\mathbb{P}(x \mid A_i) \\ &\quad + (\mathbb{P}_n - \mathbb{P})\widehat{\kappa}_u(A_i,X;\widehat{q}_u) \\ &= \int\{\widehat{w}(A_i,x)-w(A_i,x)\}[\mathbb{E}\{s_u(Z;\widehat{q}_u) \mid A_i,x\} - \overline{\kappa}_u(A_i,x;\overline{q}_u)]d\mathbb{P}(x \mid A_i) \\ &\quad + \int\{\widehat{w}(A_i,x)-w(A_i,x)\}\{\overline{\kappa}_u(A_i,x;\overline{q}_u) - \widehat{\kappa}_u(A_i,x;\widehat{q}_u)\}d\mathbb{P}(x \mid A_i) \\ &\quad + (\mathbb{P}_n - \mathbb{P})\widehat{\kappa}_u(A_i,X;\widehat{q}_u)\end{aligned}$$

As shown in Dorn et al. [2021] (Lemma 5), the map $q \mapsto s_u(Z;q)$ is Lipschitz. Therefore, by Cauchy-Schwarz:

$$\left|\widehat{\mathbb{E}}\{\widehat{b}(A) \mid A=t\}\right| \lesssim \sup_{a\in N_t}\left[\|\widehat{w}-w\|_a\{\|\widehat{q}_u-\overline{q}_u\|_a + \|\widehat{\kappa}_u-\overline{\kappa}_u\|_a\} + |(\mathbb{P}_n-\mathbb{P})\widehat{\kappa}_u(a,X;\widehat{q}_u)|\right]$$