Quantifying the Internal Validity of Weighted Estimands*

Alexandre Poirier† Tymon Słoczyński‡

March 19, 2025

Abstract

In this paper we study a class of weighted estimands, which we define as parameters that can be expressed as weighted averages of the underlying heterogeneous treatment effects. The popular ordinary least squares (OLS), two-stage least squares (2SLS), and two-way fixed effects (TWFE) estimands are all special cases within our framework. Our focus is on answering two questions concerning weighted estimands. First, under what conditions can they be interpreted as the average treatment effect for some (possibly latent) subpopulation? Second, when these conditions are satisfied, what is the upper bound on the size of that subpopulation, either in absolute terms or relative to a target population of interest? We argue that this upper bound provides a valuable diagnostic for empirical research. When a given weighted estimand corresponds to the average treatment effect for a small subset of the population of interest, we say its internal validity is low. Our paper develops practical tools to quantify the internal validity of weighted estimands.

Keywords: internal validity, ordinary least squares, representativeness, treatment effects, two-stage least squares, two-way fixed effects, weakly causal estimands, weighted estimands

JEL classification: C20, C21, C23, C26

*First arXiv draft: April 22, 2024. This paper was presented at LMU Munich, University of Bonn, Brown University, Brandeis University, University of Pittsburgh, McMaster University, the 2024 Annual Congress of the European Economic Association, the 2024 DC-MD-VA Econometrics Workshop, the 2024 Midwest Econometrics Group Meeting, the 2024 Southern Economic Association Meeting, the 2024 EC^2 Conference, and the 2024 BU/BC Greenline Econometrics Workshop. We thank audiences at those seminars and conferences, as well as Greg Caetano, Brant Callaway, Kevin Chen, Clément de Chaisemartin, Joachim Freyberger, Christian Hansen, Peter Hull, Toru Kitagawa, Matt Masten, Tomasz Olma, Guillaume Pouliot, Jonathan Roth, Pedro Sant'Anna, Alex Torgovitsky, and Daniel Wilhelm for helpful conversations and comments.

†Department of Economics, Georgetown University, alexandre.poirier@georgetown.edu

‡Department of Economics, Brandeis University, tslocz@brandeis.edu

1 Introduction

Estimating average treatment effects is an important objective in many areas of empirical research. Applied researchers usually believe that treatment effects are heterogeneous, which means that they vary across units. Yet, many researchers also favor using well-established estimation methods that were not originally designed with treatment effect heterogeneity in mind. These methods may be chosen because of their computational simplicity, comparability across studies, effectiveness at incorporating high-dimensional covariates, and other reasons. In turn, these methods often lead to estimands that can be represented as weighted averages of the underlying treatment effects of interest.

For example, consider a scenario where unconfoundedness holds given covariates X. Let treatment D be binary, $(Y(1), Y(0))$ be potential outcomes, and let $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$ be the conditional average treatment effect, or CATE, for covariate value X. Following Angrist (1998), if we additionally assume that $\mathbb{E}[D \mid X]$ is linear in X, the population regression of Y on a constant, treatment D, and covariates X yields a coefficient on D that can be written as

$$\beta_{\text{OLS}} = \frac{\mathbb{E}[\text{var}(D \mid X)\tau_0(X)]}{\mathbb{E}[\text{var}(D \mid X)]},$$

a weighted average of CATEs with nonnegative weights that integrate to 1. This parameter will be equal to the average treatment effect, $\mathbb{E}[Y(1) - Y(0)]$, if and only if $\text{var}(D \mid X)$ and $\tau_0(X)$ are uncorrelated.

In this paper we are concerned with a general class of *weighted estimands* that can be expressed as follows:

$$\mu(a, \tau_0) \coloneqq \frac{\mathbb{E}[a(X)w_0(X)\tau_0(X)]}{\mathbb{E}[a(X)w_0(X)]} = \frac{\mathbb{E}[a(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]}, \tag{1.1}$$

where $W_0 \in \{0, 1\}$ is an indicator for a subpopulation, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X]$ are the CATEs given covariates X in the same subpopulation W_0, $w_0(X) = \mathbb{P}(W_0 = 1 \mid X)$ is the probability of being in this subpopulation given X, and $a(X)$ is an identified weight function. The regression estimand above belongs to this class, which can be seen by letting $W_0 = 1$ with probability 1, and letting the weight function $a(X)$ be the conditional variance of treatment given covariates. Under some assumptions, this class also includes the two-stage least squares (2SLS) and two-way fixed effects (TWFE) estimands in instrumental variables and difference-in-differences settings, as well as many other parameters. Here, the leading cases of W_0 are compliers in the case of 2SLS and treated units in the case of TWFE. In the case of 2SLS, we would thus interpret $\tau_0(X)$ as the average treatment effect for compliers with covariates X.

There are two main questions that this paper seeks to answer. The first is whether, and under what circumstances, the estimand in (1.1) corresponds to an average treatment effect of the form $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$, where $W^* \in \{0, 1\}$ is an indicator for a (possibly latent) subpopulation of W_0. An affirmative answer to this question would endow a specific weighted estimand with some degree of validity as a causal parameter, given that it would then measure the average effect of treatment for a subset of all units.

The second and primary aim of this paper is to *quantify* the degree of validity of $\mu(a, \tau_0)$ as a causal parameter. To do this, we characterize the size, and the size relative to W_0, of subpopulations W^* associated with the estimand in (1.1). More plainly, we ask how large $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ can be in the representation $\mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$. If these probabilities can be large, the estimand

corresponds to the average treatment effect for a (relatively) large subpopulation, and when they are small, it corresponds to the average effect for a (relatively) small number of units. If $\mathbb{E}[Y(1)-Y(0) \mid W_0 = 1]$ is the target parameter, we interpret a large value of $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ as evidence of a high degree of *internal validity* of $\mu(a, \tau_0)$ with respect to the target. If $\mathbb{P}(W^* = 1)$, the corresponding marginal probability, is large, we say that $\mu(a, \tau_0)$ is highly *representative* of the underlying population.

The answer to our questions about subpopulation existence and size depends on the information we have about the CATE function, τ_0. Specifically, in one case, we may want to know whether $\mu(a, \tau_0)$ can be written as $\mathbb{E}[Y(1)-Y(0) \mid W^* = 1]$ for any choice of τ_0, or without any knowledge of this function. If this is the case, then we know that the interpretation of $\mu(a, \tau_0)$ as a causal parameter is robust to heterogeneous treatment effects of any form, including the most adversarial CATE functions. We can also answer the second question about the maximum values of $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ without needing to estimate or know the structure of the CATEs. In a second case, we may want to know how representative $\mu(a, \tau_0)$ is *given* knowledge of the CATE function. While the resulting maximum values of $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ are less useful as measures of robustness than in the first case—after all, if the researcher knows or estimates the entire CATE function, they can as well report any average of $\tau_0(X)$ that may be relevant—we consider this problem to be of independent theoretical interest. Additionally, if the researcher estimates and compares the maximum values of $\mathbb{P}(W^* = 1)$ or $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ in both cases, they can evaluate the importance of treatment effect heterogeneity for the interpretation of $\mu(a, \tau_0)$ in a given application.

In the first case, when the CATE function is unrestricted, we formally show that $\mu(a, \tau_0)$ can be written as the average treatment effect for a subpopulation of W_0 if and only if $a(X) \geq 0$ with probability 1 given $W_0 = 1$. The contrapositive of this statement is that the incidence of "negative weights," that is, $\mathbb{P}(a(X) < 0 \mid W_0 = 1) > 0$, implies that $\mu(a, \tau_0)$ cannot be represented as an average treatment effect for some subpopulation uniformly in τ_0. This result provides a novel justification for the commonly invoked requirement that the weights underlying a suitable estimand must all be positive. In a related contribution, Blandhol, Bonney, Mogstad, and Torgovitsky (2022) have shown that, for estimands that do not depend on potential outcome levels, the lack of negative weights is a sufficient and necessary condition for the weighted estimand to be "weakly causal," that is, to guarantee that the sign of τ_0 will be preserved whenever it is uniform across all units. We also provide simple expressions for the maxima of $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$. We show how knowledge of the estimand and of these expressions can be used to construct simple bounds on the target parameter. We also propose an analog estimator for our measure of internal validity. We then establish the nonstandard limiting distribution of this estimator, and describe inference procedures for it.

In the second case, when the CATE function is assumed to be known, we show that $\mu(a, \tau_0)$ can be written as an average treatment effect whenever it lies in the convex hull of CATE values, a weaker criterion than having nonnegative weights. The maximum values of $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ now depend on τ_0, and can be obtained via linear programming when X is discrete. We show the solution to this linear program also admits a closed-form expression even when the support of X includes discrete, continuous, and mixed components. This expression can be used to derive plug-in estimators.

Besides theoretical interest, we argue that $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ are a practically relevant measure that may be appealing to applied researchers. First, as stated above, our initial results demonstrate

that two commonly invoked criteria for weighted estimands—that they lack negative weights and that they lie in the convex hull of CATE values—are necessary and sufficient (under different assumptions) for the existence of their causal representation, that is, for $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ to be strictly positive. This suggests that researchers invoking these criteria are (indirectly) interested in whether their estimands can be represented as an average treatment effect for some subpopulation. If this is the case, it makes sense to better understand this implicit subpopulation, similar to how it is standard practice in instrumental variables settings to study the subpopulation of compliers. Relatedly, even though our main results concern subpopulation size, we also show that the distribution of covariates in the implicit subpopulation is identified. Thus, practitioners can examine whether this subpopulation has similar characteristics as the entire population, and report the associated sample statistics.

Second, we argue that when $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$ is the parameter of interest, it is reassuring for $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ to be large. For one, related claims have been made by other researchers. Mogstad and Torgovitsky (2024) assert that "[t]arget parameters that reflect larger subpopulations of the population of interest are more interesting than those that reflect smaller and more specific subpopulations." In a setting with multiple instrumental variables, van 't Hoff, Lewbel, and Mellace (2024) argue that the largest subpopulation of compliers is generally more interesting than other complier subpopulations. However, we also formalize this claim and show how to construct bounds on $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$ that only depend on $\mathbb{P}(W^* = 1 \mid W_0 = 1)$, $\mu(a, \tau_0)$, and a support restriction. The bounds are easy to compute and converge to a point as $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ approaches 1. Indeed, large values of $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$, our primary measures of interest, guarantee that the weighted estimand is not "too different" from $\mathbb{E}[Y(1) - Y(0)]$ and $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$.

Literature Review

This paper is related to a large literature studying weighted average representations of common estimands, including ordinary least squares (OLS), 2SLS, and TWFE in additive linear models. Some of the contributions to this literature include Angrist (1998), Humphreys (2009), Aronow and Samii (2016), Blandhol, Bonney, Mogstad, and Torgovitsky (2022), Słoczyński (2022), Chen (2024), and Goldsmith-Pinkham, Hull, and Kolesár (2024) for OLS; Imbens and Angrist (1994), Angrist and Imbens (1995), Kolesár (2013), Słoczyński (2020), and Blandhol, Bonney, Mogstad, and Torgovitsky (2022) for 2SLS; and de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2021), Sun and Abraham (2021), Athey and Imbens (2022), Caetano and Callaway (2023), Borusyak, Jaravel, and Spiess (2024), and Callaway, Goodman-Bacon, and Sant'Anna (2024) for TWFE.

A common view in much of this literature, attributable to Imbens and Angrist (1994), is that causal interpretability of weighted estimands requires all weights to be positive. For example, Sun and Abraham (2021) explicitly associate "reasonable weights" with weights that "sum to one and are non-negative." Blandhol, Bonney, Mogstad, and Torgovitsky (2022) show that the lack of negative weights and level dependence is necessary and sufficient for an estimand to be "weakly causal," that is, to guarantee sign preservation when all treatment effects have the same sign. In this paper we focus on the related problem of whether a weighted estimand can be written as the average treatment effect for some (possibly latent) subpopulation. While the lack of negative weights is essential in our framework when the CATE function is unrestricted, negative and nonuniform weights play a similar role when the CATE function is assumed to be known, at least as long as

the weighted estimand lies in the convex hull of CATE values. This point is related to the negative view of both negative and nonuniform weights in Callaway, Goodman-Bacon, and Sant'Anna (2024).

Some papers focus on weighted averages of heterogeneous treatment effects as legitimate targets in their own right rather than as probability limits of existing estimators. Hirano, Imbens, and Ridder (2003) introduce the class of weighted average treatment effects, which are a subclass of the more general class of estimands in (1.1). Li, Morgan, and Zaslavsky (2018) discuss the connection between weighted average treatment effects and implicit target subpopulations. However, the internal validity and representativeness of weighted estimands have received very little attention to date.

One exception is de Chaisemartin (2012, 2017), who revisits the interpretation of the instrumental variables (IV) estimand in the framework of Imbens and Angrist (1994). First, de Chaisemartin (2012) studies the size of the largest subpopulation whose average treatment effect is equal to that of compliers. While this question is similar to ours, the corresponding subpopulation size is not point identified, unlike in this paper. Our framework is also more general and includes instrumental variables settings as a special case. Second, when the usual monotonicity assumption is violated, de Chaisemartin (2012, 2017) reinterprets the IV estimand as the average treatment effect for a subset of compliers. In our framework, this result can be seen as an existence result in an intermediate case between the setting where τ_0 is unrestricted and where it is fixed. The specific homogeneity assumption considered by de Chaisemartin (2012, 2017) allows him to salvage the causal representation of the IV estimand despite the incidence of negative weights.

Another exception is Aronow and Samii (2016), who explicitly acknowledge that the OLS estimand, like the local average treatment effect of Imbens and Angrist (1994), corresponds to the average effect for a "highly specific subpopulation" rather than the entire population, and consequently is not necessarily representative of that population. Then, Aronow and Samii (2016) focus on whether mean covariate values are similar in the entire sample and in the "effective sample" used by OLS. We focus on the size of the implicit subpopulation, which is different and complementary. We also extend the results on mean covariate values to the entire distribution of covariates and to other weighted estimands besides OLS.

Yet another exception is Miller, Shenhav, and Grosz (2023), who focus on (one-way) fixed effects estimands and argue that it is problematic if "switchers," that is, fixed-effect groups with nonzero variation in treatment, are a small subset of the sample. They also recommend that applied researchers report the sample size when limited to "switcher groups." In this paper we build a general framework to study the internal validity and representativeness of weighted estimands, with the fixed effects estimand (equivalent to OLS) as a special case. We argue, similar to Miller, Shenhav, and Grosz (2023), that if a given weighted estimand corresponds to the average treatment effect for a small subpopulation, then it may not be an appropriate target parameter, unless that subpopulation is interesting in its own right.

Plan of the Paper

We organize the paper as follows. In Section 2, we provide a more detailed discussion of the OLS estimand, which is our motivating example. In Section 3, we develop our theoretical framework and examine the conditions under which the estimand in (1.1) has a causal representation as an average treatment effect over a population. In Section 4, we establish our main results on the absolute and relative size of subpopulations associated with the estimand in (1.1), which we propose as measures of representativeness and internal

validity of weighted estimands. In Section 5, we revisit our motivating example from Section 2 and apply our theoretical results to additional examples of weighted estimands. In particular, we study 2SLS with a binary instrument and TWFE under parallel trends assumptions. In Section 6, we briefly discuss estimation and inference for the proposed measures. In Section 7, we provide an empirical application to the effects of unilateral divorce laws on female suicide, as in Stevenson and Wolfers (2006) and Goodman-Bacon (2021). In Section 8, we conclude. The appendix contains our proofs as well as several additional results and derivations.

2 Motivating Example

Here we provide a more detailed discussion of the OLS estimand, which is our initial theoretical example. We postpone the discussion of the 2SLS and TWFE estimands to Section 5. In the initial example, we have a binary treatment $D \in \{0, 1\}$, potential outcomes $(Y(1), Y(0))$, covariate vector X, and realized outcome $Y = Y(D)$. We make the following two assumptions.

Assumption 2.1 (Unconfoundedness)**.** Let

1. Conditional independence: $(Y(1), Y(0)) \perp\!\!\!\perp D \mid X$;
2. Overlap: $p(X) \coloneqq \mathbb{P}(D = 1 \mid X) \in (0, 1)$ almost surely.

Following Angrist (1998), we can establish that β_{OLS}, the coefficient on D in the linear projection of Y on $(1, D, X)$, satisfies the representation in (1.1). The following proposition summarizes Angrist's (1998) result.

Proposition 2.1. Suppose Assumption 2.1 holds. Suppose $p(X)$ is linear in X. Then

$$\beta_{\text{OLS}} = \frac{\mathbb{E}[p(X)(1 - p(X)) \cdot \mathbb{E}[Y(1) - Y(0) \mid X]]}{\mathbb{E}[p(X)(1 - p(X))]}.$$

The linearity assumption can be removed if we instead regress Y on $(1, D, h(X))$ where $h(X)$ is a vector of functions of X such that $p(X)$ is in their linear span. The overlap assumption can also be weakened since it is not required for β_{OLS} to be defined.

Proposition 2.1 implies that we can write β_{OLS} as

$$\beta_{\text{OLS}} = \frac{\mathbb{E}[a(X)\tau_0(X)]}{\mathbb{E}[a(X)]},$$

where $a(X) = p(X)(1 - p(X))$ and $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$. Here we implicitly set $W_0 = 1$ with probability 1. Thus, the regression coefficient β_{OLS} is a weighted average of CATEs whose weights are $p(X)(1 - p(X))$. Note that $\beta_{\text{OLS}} = \text{ATE} \coloneqq \mathbb{E}[Y(1) - Y(0)]$ if and only if $a(X)$ and $\tau_0(X)$ are uncorrelated, which is the case, for example, when $p(X)$ or $\tau_0(X)$ is constant.

An alternative representation of this estimand can be obtained by focusing on the subpopulation of treated units, $D = 1$. Let $W_0 = D$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D = 1, X] = \mathbb{E}[Y(1) - Y(0) \mid X]$, which follows from conditional independence, and let $\tilde{a}(X) = 1 - p(X)$. Then, we can write

$$\beta_{\text{OLS}} = \frac{\mathbb{E}[(1 - p(X))w_0(X)\tau_0(X)]}{\mathbb{E}[(1 - p(X))w_0(X)]} = \frac{\mathbb{E}[\tilde{a}(X)w_0(X)\tau_0(X)]}{\mathbb{E}[\tilde{a}(X)w_0(X)]},$$

where $w_0(X) = \mathbb{P}(D = 1 \mid X) = p(X)$. Yet another representation can be obtained when focusing on the subpopulation of untreated units by letting $W_0 = 1 - D$. We omit details for brevity.

We will return to this example in Section 5 after establishing conditions under which weighted estimands have a causal representation (Section 3) and identifying the size of subpopulations that are represented by these estimands (Section 4).

3 Causal Representation of Weighted Estimands

We now consider a general class of weighted estimands. In this section, we show necessary and sufficient conditions for an estimand in this class to have a causal representation as an average treatment effect over a subpopulation. We provide these conditions under various assumptions—including no assumptions—on the heterogeneity of treatment effects.

Recall the earlier setting where we let $D \in \{0, 1\}$ denote a binary treatment variable, and let $(Y(1), Y(0))$ denote the corresponding potential outcomes under treatment and control, respectively. Let $X \in \text{supp}(X) \subseteq \mathbb{R}^{d_X}$ denote a d_X-vector of covariates, where $\text{supp}(\cdot)$ denotes the support of a random vector. We suppose that $(Y(1), Y(0), D, X)$ are drawn from a common population distribution $F_{Y(1),Y(0),D,X}$.

Let $W_0 \in \{0, 1\}$ be an indicator variable used to denote a subpopulation $\{W_0 = 1\}$ and let $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X]$ denote the conditional average treatment effect given X in that subpopulation. For example, this subpopulation can be the entire population by setting $W_0 = 1$ almost surely, in which case τ_0 denotes the usual CATE function. It can also denote the subpopulation of treated units by setting $W_0 = D$. In the presence of a binary instrument Z, the complier subpopulation is defined by setting $W_0 = \mathbb{1}(D(1) > D(0))$, where $D(1)$ and $D(0)$ are potential treatments. In this case, τ_0 denotes the conditional local average treatment effect or conditional LATE.

Note that τ_0 is defined for all values of X such that $w_0(X) = \mathbb{P}(W_0 = 1 \mid X) > 0$.[1] Throughout this paper, we assume that $\mathbb{P}(W_0 = 1) > 0$, so that this subpopulation has a positive mass, which avoids technical issues associated with conditioning on zero-probability events.

Also recall the weighted estimands of equation (1.1):

$$\mu(a, \tau_0) = \frac{\mathbb{E}[a(X)w_0(X)\tau_0(X)]}{\mathbb{E}[a(X)w_0(X)]} = \frac{\mathbb{E}[a(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]}.$$

The estimands we consider have the above representation and satisfy the following regularity conditions.

Assumption 3.1 (Regularity)**.** Let $\mathbb{E}[\tau_0(X)^2] < \infty$, $\mathbb{E}[a(X)^2] < \infty$, and $\mathbb{E}[a(X) \mid W_0 = 1] > 0$.

The first two restrictions are weak regularity assumptions that ensure the finiteness of the numerator of $\mu(a, \tau_0)$. We rule out $\mathbb{E}[a(X) \mid W_0 = 1] = 0$ since it implies the estimand does not exist. The estimand in (1.1) is unchanged if the sign of $a(X)$ is reversed, so $\mathbb{E}[a(X) \mid W_0 = 1] > 0$ is a sign normalization.

[1] While $\tau_0(X)$ is only defined when $w_0(X) > 0$, we set $\tau_0(X)w_0(X) = 0$ when $w_0(X) = 0$.

3.1 Alternative Representations of Weighted Estimands

The weighted estimands of equation (1.1) can also be written as a weighted sum when X is discrete, or an integral when X is continuous. In the discrete case, let $\text{supp}(X) = \{x_1, \ldots, x_K\}$, let $p_k \coloneqq \mathbb{P}(X = x_k) > 0$ for $k = 1, \ldots, K$, and assume $W_0 = 1$ almost surely for simplicity. Then,

$$\mu(a, \tau_0) = \sum_{k=1}^{K} \omega_k \tau_0(x_k) \qquad \text{where} \qquad \omega_k = \frac{a(x_k)p_k}{\sum_{l=1}^{K} a(x_l)p_l}, \tag{3.1}$$

which are weights that sum to one. The representations in (1.1) and (3.1) are equivalent as we can obtain $a(x_k)$ (up to scale) as the ratio ω_k/p_k, and ω_k is defined as a function of $\{(a(x_k), p_k)\}_{k=1}^{K}$ in equation (3.1).

From equation (3.1), we can see that $a(x_k)$ being constant ensures $\omega_k = p_k$, or that the estimand is the ATE. Moreover,

$$\frac{a(x_k)}{a(x_{k'})} = \frac{\omega_k}{\omega_{k'}} \Big/ \frac{p_k}{p_{k'}},$$

which is the ratio of the relative weights of covariate cells $\{X = x_k\}$ and $\{X = x_{k'}\}$ in the estimand $(\omega_k/\omega_{k'})$ and in the population $(p_k/p_{k'})$. The inequality $a(x_k) > a(x_{k'})$ indicates that covariate cell $\{X = x_k\}$ is overweighted relative to $\{X = x_{k'}\}$, when compared to their relative weights in the population.

Alternatively, consider the case where X is continuously distributed with density[2] f_X. Still assuming $W_0 = 1$ almost surely, we can write

$$\mu(a, \tau_0) = \int_{\text{supp}(X)} \omega(x)\tau_0(x)\, dx \qquad \text{where} \qquad \omega(x) = \frac{a(x)f_X(x)}{\int_{\text{supp}(X)} a(x)f_X(x)\, dx} \tag{3.2}$$

is a weight function that integrates to 1. We focus on the representation in equation (1.1) since it seamlessly accommodates discrete, continuous, and mixed covariates.

3.2 Regular Subpopulations

The first question we address is whether an estimand defined by (1.1) can be represented as $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$, where $W^* \in \{0, 1\}$ is binary and $\{W^* = 1\}$ characterizes a subpopulation of $\{W_0 = 1\}$. Formally, $\{W^* = 1\}$ forms a subpopulation if $\{W^* = 1\} \subseteq \{W_0 = 1\}$ or, equivalently, if $W^* \leq W_0$ almost surely.

We impose some structure on this problem by restricting how these subpopulations may be formed. We will consider what we call "regular subpopulations," which we define here.

Definition 3.1. Let $W^* \in \{0, 1\}$ such that $\mathbb{P}(W^* = 1) > 0$. Say $\{W^* = 1\}$ is a *regular subpopulation* of $\{W_0 = 1\}$ if

1. (Inclusion) $\mathbb{P}(W^* = 1 \mid W_0 = 0) = 0$;
2. (Conditional independence) $W^* \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$.

[2] With respect to the Lebesgue measure.

For convenience, we will abbreviate this as "W^* is a regular subpopulation of W_0". We denote the set of regular subpopulations of W_0 as

$$\mathrm{SP}(W_0) = \{W^* \in \{0,1\} : W^* \text{ is a regular subpopulation of } W_0\}.$$

We consider subpopulations with positive masses that are subsets of $\{W_0 = 1\}$. The second and main requirement is that regular subpopulations do not depend on potential outcomes when conditioning on X and the original population $W_0 = 1$. While this may seem restrictive, it allows for rich and natural classes of subpopulations. For example, consider the unconfoundedness restriction of Section 2 and let W_0 be the entire population, i.e. $\mathbb{P}(W_0 = 1) = 1$. In this case, regular subpopulations must satisfy $W^* \perp\!\!\!\perp (Y(1), Y(0)) \mid X$, or be unconfounded. Regular subpopulations include the population of all treated (or untreated) individuals, i.e. $W^* = D$ (or $W^* = 1 - D$), and any subpopulation characterized by a subset of $\mathrm{supp}(X)$. More generally, they include any subpopulation that can be described through a combination of (D, X, U) where U is independent from $(Y(1), Y(0), X)$. For example, a subpopulation characterized by "fraction $a(x)$ of units with covariate $X = x$ for all $x \in \mathrm{supp}(X)$" can be constructed as $W^* = \mathbb{1}(U \leq a(X))$ where $U \sim \mathrm{Unif}(0,1)$ is independent from $(Y(1), Y(0), X)$.

The conditional independence requirement rules out subpopulations that directly depend on the potential outcomes such as $W^* = \mathbb{1}(Y(1) \geq Y(0))$, the subpopulation of those who benefit from treatment. Note that $\mathbb{P}(W^* = 1 \mid X) = \mathbb{P}(Y(1) \geq Y(0) \mid X)$ and $\mathbb{P}(W^* = 1) = \mathbb{P}(Y(1) \geq Y(0))$ are not point-identified under unconfoundedness. Another way to view this requirement is that regular subpopulations are policy relevant in the sense that we could design a policy that targets a regular subpopulation. Indeed, a policy maker may observe X and can use U to randomly target a fraction of units with specific values of X, but cannot observe potential outcomes.

These particular subpopulations enjoy a number of useful properties. Two of them are characterized in the following proposition.

Proposition 3.1 (Properties of regular subpopulations)**.** Suppose that $\mathbb{P}(W_0 = 1) > 0$ and $W^* \in \mathrm{SP}(W_0)$.

1. Suppose $\mathbb{P}(W^* = 1 \mid W_0 = 1, X) > 0$. Then,

$$\mathbb{E}[Y(1) - Y(0) \mid W^* = 1, X] = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X]. \tag{3.3}$$

2. Suppose $\mathbb{E}[\tau_0(X)^2] < \infty$. Then,

$$\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} = \mu(\underline{w}^*, \tau_0), \tag{3.4}$$

   where $\underline{w}^*(x) \coloneqq \mathbb{P}(W^* = 1 \mid X = x, W_0 = 1)$.

The first part of this proposition shows that average effects within the original population W_0 and regular subpopulation W^* are the same when conditioning on X. For example, this holds under unconfoundedness for the subpopulation of treated individuals, $W^* = D$. The second property allows us to write the average effect for $W^* = 1$ using the same functional $\mu(\cdot, \cdot)$ that was used to characterize the class of estimands we analyze. This property will be used when studying the mapping between weighted estimands and average effects for regular subpopulations of W_0.

We conclude this subsection by showing that regular subpopulations are transitive, or that regular subpopulations of a regular subpopulation of W_0 are also regular subpopulations of W_0.

Lemma 3.1 (Transitivity of regular subpopulations)**.** Suppose W^* is a regular subpopulation of W' and that W' is a regular subpopulation of W_0. Then, W^* is a regular subpopulation of W_0.

3.3 Existence of a Causal Representation for Weighted Estimands

We now consider necessary and sufficient conditions for the weighted estimand $\mu(a, \tau_0)$ to be written as the average treatment effect within a regular subpopulation of W_0. As we will show, these conditions depend on what is assumed about the function $\tau_0 = \mathbb{E}[Y(1) - Y(0) \mid X = \cdot, W_0 = 1]$.

For example, if τ_0 is constant in X, then any weighted estimand satisfying (1.1) equals $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, the average treatment effect within population $\{W_0 = 1\}$. This is the case even when the sign of weight function $a(X)$ varies with X. However, if τ_0 is nonconstant, the existence of causal representations will depend on the weight function $a(X)$. Among other cases, we will consider the case where no restrictions are placed on function τ_0, and in this case, the existence of a causal representation of $\mu(a, \tau_0)$ will require the sign of $a(X)$ to be constant.

To formalize this, let $\mathcal{T}$ denote a class of functions such that $\tau_0 \in \mathcal{T}$ and define

$$\mathcal{W}(a; W_0, \mathcal{T}) = \{W^* \in \mathrm{SP}(W_0) : \mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1] \text{ for all } \tau_0 \in \mathcal{T}\}.$$

This is the set of regular subpopulations of W_0 such that the estimand $\mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ for all τ_0 functions in the set $\mathcal{T}$. If the set $\mathcal{W}(a; W_0, \mathcal{T})$ is empty, then estimand $\mu(a, \tau_0)$ cannot be written as an average treatment effect over a regular subpopulation of W_0 uniformly in $\tau_0 \in \mathcal{T}$. We use this set to formally define a notion of uniform causal representation.

Definition 3.2. A weighted estimand $\mu(a, \tau_0)$ has a *causal representation uniformly in* $\tau_0 \in \mathcal{T}$ if

$$\mathcal{W}(a; W_0, \mathcal{T}) \neq \emptyset.$$

Recall that $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \mu(\underline{w}^*, \tau_0)$ where $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid W_0 = 1, X)$, so $W^* \in \mathcal{W}(a; W_0, \mathcal{T})$ if $\mu(a, \tau_0) = \mu(\underline{w}^*, \tau_0)$ for all $\tau_0 \in \mathcal{T}$. We examine further two main cases for the set $\mathcal{T}$.

3.3.1 Existence Uniformly in τ_0

We begin by considering the largest class of functions in which τ_0 lies: the class of all functions, subject to the moment condition in Assumption 3.1 that ensures the existence of $\mu(a, \tau_0)$. We denote this class by

$$\mathcal{T}_{\text{all}} \coloneqq \{\tau_0 : \mathbb{E}[\tau_0(X)^2] < \infty\}.$$

In this function class, we show the existence of a causal representation is equivalent to the estimand's weights being nonnegative. In what follows, let $a_{\max} \coloneqq \sup(\mathrm{supp}(a(X) \mid W_0 = 1))$ be the essential supremum of $a(X)$ given $W_0 = 1$.

Theorem 3.1. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds and that $a_{\max} < \infty$. Then, $\mu(a, \tau_0)$ has a causal representation uniformly in $\tau_0 \in \mathcal{T}_{\text{all}}$ if and only if

$$\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) = 1. \tag{3.5}$$

A uniform (in $\mathcal{T}_{\text{all}}$) causal representation exists if and only if $a(X)$ is nonnegative on the support of $X \mid \{W_0 = 1\}$. To give some intuition on why the sign of $a(X)$ must be nonnegative with probability 1, we present a contradiction that occurs when $a(X)$ can be negative. Suppose that $\mathbb{P}(a(X) < 0 \mid W_0 = 1) > 0$ and consider the "adversarial" CATE function $\tau^-(X) = \mathbb{1}(a(X) < 0)$. This CATE function is nonnegative for all X, and implies a positive average effect only for units with negative weights. However, it yields a strictly negative weighted estimand, $\mu(a, \tau^-) = \mathbb{E}[a(X)\mathbb{1}(a(X) < 0) \mid W_0 = 1]/\mathbb{E}[a(X) \mid W_0 = 1] < 0$. Clearly, $\mu(a, \tau^-)$ cannot be the average treatment effect for any subpopulation of W_0, because averaging a nonnegative CATE function over any subpopulation cannot yield a negative average.

Conversely, if $a(X) \geq 0$, our proof constructively defines a regular subpopulation W^* for which the average effect is equal to the weighted estimand $\mu(a, \tau_0)$ uniformly in $\tau_0 \in \mathcal{T}_{\text{all}}$. Let

$$W^* = \mathbb{1}\left(U \leq \frac{a(X)}{a_{\max}}\right) \cdot W_0,$$

where $U \sim \text{Unif}(0, 1) \perp\!\!\!\perp (Y(1), Y(0), X, W_0)$. This is a regular subpopulation of W_0 for which the probability of inclusion, conditional on X and $W_0 = 1$, is proportional to $a(X)$. We can also interpret $\mu(a, \tau_0)$ as the average effect of an intervention in which units with covariate value X are treated randomly with probability $a(X)/a_{\max}$ given $W_0 = 1$. From this construction, we can see that $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid W_0 = 1, X)$ is proportional to $a(X)$, and therefore $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \mu(\underline{w}^*, \tau_0) = \mu(a, \tau_0)$ uniformly in τ_0.

The condition $a_{\max} < \infty$ restricts our attention to subpopulations with positive mass. We note that $a_{\max}$ is bounded above in each of our theoretical examples in sections 2 and 5, implying that this condition trivially holds in these cases.

As mentioned earlier, the condition that weights are nonnegative is commonly invoked and was shown in Blandhol, Bonney, Mogstad, and Torgovitsky (2022) to be equivalent to an estimand being "weakly causal," that is, to match the sign of τ_0 whenever that is the same across all units. Thus, in the class of weighted estimands we consider, estimands have a causal representation uniformly in $\mathcal{T}_{\text{all}}$ if and only if they are weakly causal. This connection is formally established in Appendix A.

3.3.2 Existence for a Given τ_0

We now provide an existence result that requires the causal representation to exist only for the *given* τ_0, rather than uniformly for τ_0 in the larger set $\mathcal{T}_{\text{all}}$. The following result depends on the CATE function τ_0 in the population, whereas Theorem 3.1's condition depended only on the weight function $a(X)$ and the nature of the covariates' support. Thus, the distribution of the potential outcomes will have an impact on the existence of a causal representation given τ_0. Using the notation from Definition 3.2, a causal representation exists if and only if $\mathcal{W}(a; W_0, \{\tau_0\}) \neq \emptyset$. The following theorem characterizes this existence.

Theorem 3.2. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds. Then,

$\mu(a, \tau_0)$ has a causal representation for the given τ_0 if and only if

$$\mu(a, \tau_0) \in \mathcal{S}(\tau_0; W_0) \coloneqq \{t \in \mathbb{R} : \mathbb{P}(\tau_0(X) \leq t \mid W_0 = 1) > 0 \text{ and } \mathbb{P}(\tau_0(X) \geq t \mid W_0 = 1) > 0\}. \tag{3.6}$$

The existence condition in Theorem 3.2 is weaker than the one in Theorem 3.1 since we no longer require this representation to be valid for any CATE function, but rather just for the one that is identified from the population. The necessary and sufficient condition in this theorem is rather weak, since it only requires that the estimand is in the convex hull of the support of the CATEs. This means $\mu(a, \tau_0)$ has a causal representation even with negative weights, as long as there are CATEs smaller and greater than $\mu(a, \tau_0)$. We can see this support condition holds for all $\tau_0 \in \mathcal{T}_{\text{all}}$ if and only if $\mu(a, \tau_0)$ is in the support of $\tau_0(X)$ for any $\tau_0 \in \mathcal{T}_{\text{all}}$. This is precisely the case when the weights $a(X)$ are nonnegative since it guarantees $\inf(\text{supp}(\tau_0(X) \mid W_0 = 1)) \leq \mu(a, \tau_0) \leq \sup(\text{supp}(\tau_0(X) \mid W_0 = 1))$ for any function τ_0.

3.3.3 Existence in Intermediate Cases

Analyzing the causal representation of an estimand under no restrictions on τ_0 could be viewed as unnecessarily conservative in some settings. At the other extreme, assuming knowledge of τ_0 may be unrealistic, especially in scenarios where X has many components which makes the estimation of τ_0 more challenging. For example, some shape constraints may be known to hold for τ_0. In some economic applications one may posit that τ_0 is monotonic or convex in some components of X, or positive/negative over a subset of $\text{supp}(X \mid W_0 = 1)$. In these cases, the existence of a causal representation may occur under weaker conditions than those in Theorem 3.1, but stronger than those in Theorem 3.2. In particular, one may be able to relax the requirement that $a(X) \geq 0$ without requiring that τ_0 be completely known to the researcher. The following proposition shows this is the case when $\tau_0(X)$ is assumed to be linear in X.

Proposition 3.2. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds and define

$$\mathcal{T}_{\text{lin}} = \{\tau_0 \in \mathcal{T}_{\text{all}} : \tau_0(X) = c + d'X : (c, d) \in \mathbb{R}^{1+d_X}\}.$$

Then, $\mu(a, \tau_0)$ has a causal representation uniformly in $\tau_0 \in \mathcal{T}_{\text{lin}}$ if and only if

$$\frac{\mathbb{E}[a(X)X \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} \in \text{conv}(\text{supp}(X \mid W_0 = 1)),$$

where $\text{conv}(\cdot)$ denotes the convex hull.

The above proposition shows that restricting the class of CATE functions τ_0 belongs to may remove the requirement that $a(X) \geq 0$ for the existence of a uniform causal representation for an estimand. In particular, the requirement here is that $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]}$ lies in the convex hull of the support of X given $W_0 = 1$. When X is scalar, this consists of an interval. This condition does not require $a(X)$ be nonnegative. For example, if $\text{supp}(X) = \{0, 1, 2\}$ and $W_0 = 1$ almost surely, then any combination of $(a(0), a(1), a(2))$ values such that $\mathbb{E}[a(X)X]/\mathbb{E}[a(X)] \in [0, 2] = \text{conv}(\text{supp}(X))$ implies a causal representation. Let $\mathbb{P}(X = x) = 1/3$ for $x \in \{0, 1, 2\}$ and $(a(0), a(1), a(2)) = (1, -1, 1)$. Here units with $X = 1$ have a negative weight, but $\mathbb{E}[a(X)X]/\mathbb{E}[a(X)] = 1 \in [0, 2]$, implying that the corresponding weighted estimand has a causal

representation uniformly in $\tau_0 \in \mathcal{T}_{\text{lin}}$.

We consider another class of CATE functions that restricts their heterogeneity. For $K \geq 0$, let

$$\mathcal{T}_{\text{BD}}(K) = \left\{ \tau_0 \in \mathcal{T}_{\text{all}} : \sup_{x,x' \in \text{supp}(X|W_0=1)} |\tau_0(x) - \tau_0(x')| \leq K \right\}.$$

This function class uniformly bounds differences of the CATE function. When $K = 0$, the CATE function is constant, and thus equal to $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$. When $K > 0$, CATEs may differ in value, but the maximum discrepancy between two CATEs is bounded above by K. We show that restricting the CATEs to satisfy this bounded difference assumption does not remove the requirement that $a(X)$ be nonnegative, unless $K = 0$, in which case all a functions yield a causal representation uniformly in $\mathcal{T}_{\text{BD}}(0)$. We formalize this in the next proposition.

Proposition 3.3. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds. Then, $\mu(a, \tau_0)$ has a causal representation uniformly in $\mathcal{T}_{\text{BD}}(K)$ when $K > 0$ if and only if

$$\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) = 1.$$

The estimand $\mu(a, \tau_0)$ has a causal representation uniformly in $\mathcal{T}_{\text{BD}}(0)$ for any $a(\cdot)$.

To understand this proposition, consider the adversarial CATE function $\tau^-(X) = K \cdot \mathbb{1}(a(X) < 0)$, a member of $\mathcal{T}_{\text{BD}}(K)$, and assume $\mathbb{P}(a(X) \geq 0) < 1$. Then we obtain the same contradiction we discussed after Theorem 3.1, where the CATE is nonnegative for all covariate values but the estimand is negative, implying that it cannot be written as an average effect over a subpopulation.

These last two propositions show that the impact of restrictions on τ_0 on the requirement that $a(X)$ be nonnegative critically depends on the nature of these restrictions. Generalizations to additional or empirically motivated function classes are left for future work.

4 Quantifying the Internal Validity of Weighted Estimands

Many estimands will admit causal representations, but their associated subpopulations $\{W^* = 1\}$ will generally differ. Also, a weighted estimand may not always correspond to the *target estimand* a researcher is interested in. For example, a researcher may be interested in setting $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, the average effect in population $\{W_0 = 1\}$, as the target parameter. In general, this parameter differs from estimand $\mu(a, \tau_0)$.

However, the set of subpopulations corresponding to a weighted estimand can be used to understand how representative the weighted estimand is of the target. For example, we may seek estimands for which $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ attains values closest to 1, since they have a higher degree of internal validity with respect to the target $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$. At one extreme, an estimand for which $\mathbb{P}(W^* = 1 \mid W_0 = 1) = 1$ would be deemed to have the highest degree of internal validity for this target parameter since it would equal $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$.

We convert this interpretation in a formal measure of internal validity that we define here.

Definition 4.1 (Internal validity)**.** Let

$$\overline{P}(a, W_0; \mathcal{T}) = \sup_{W^* \in \mathcal{W}(a; W_0, \mathcal{T})} \mathbb{P}(W^* = 1 \mid W_0 = 1) \tag{4.1}$$

denote the measure of *internal validity* of weighted estimand $\mu(a, \tau_0)$ over function class $\mathcal{T}$.

Formally, $\overline{P}(a, W_0; \mathcal{T})$ is the sharp upper bound on $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ for any regular subpopulation W^* of W_0 such that the weighted estimand $\mu(a, \tau_0)$ has a causal representation as the average treatment effect over subpopulation W^*. Note that we set $\overline{P}(a, W_0; \mathcal{T}) = 0$ when $\mathcal{W}(a; W_0, \mathcal{T})$ is empty. This object depends on the chosen function class $\mathcal{T}$, as did Theorems 3.1 and 3.2 in the previous section. Given the above terminology and assuming that $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$ is the target, we call $\overline{P}(a, W_0; \mathcal{T})$ a measure of the internal validity of estimand $\mu(a, \tau_0)$, and we use this definition in the remainder of the paper.

We can also compute the maximum value of $\mathbb{P}(W^* = 1)$ across $W^* \in \mathcal{W}(a; W_0, \mathcal{T})$, which measures the largest share of the entire population for which the weighted estimand has a causal representation. We refer to this measure as a measure of representativeness. The measures of internal validity and representativeness are the same when $W_0 = 1$ almost surely.

Definition 4.2 (Representativeness)**.** Let

$$\overline{P}(a, W_0; \mathcal{T}) \cdot \mathbb{P}(W_0 = 1) = \sup_{W^* \in \mathcal{W}(a; W_0, \mathcal{T})} \mathbb{P}(W^* = 1) \tag{4.2}$$

denote the measure of *representativeness* of weighted estimand $\mu(a, \tau_0)$ over function class $\mathcal{T}$.

Note that $\mathbb{P}(W^* = 1) = \mathbb{P}(W^* = 1 \mid W_0 = 1) \cdot \mathbb{P}(W_0 = 1)$ since W^* is a subpopulation of W_0. This maximum value of $\mathbb{P}(W^* = 1)$ gives the internal validity of the weighted estimand with respect to target estimand $\mathbb{E}[Y(1) - Y(0)]$, the average treatment effect in the population from which the sample is drawn. Our measures of internal validity and representativeness are closely linked and a subpopulation will maximize $\mathbb{P}(W^* = 1)$ if and only if it maximizes $\mathbb{P}(W^* = 1 \mid W_0 = 1)$. We will also show how to use these measures to obtain simple bounds on target estimands.

We now derive explicit expressions for $\overline{P}(a, W_0; \mathcal{T})$. As earlier, we break down our results in two cases, the first being when τ_0 is unrestricted.

4.1 Quantifying Internal Validity Uniformly in τ_0

Without imposing any restrictions on the CATE function, except for the existence of second moments, the maximum value that $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ can achieve is given by the following theorem.

Theorem 4.1. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds and that $a_{\max} < \infty$. If $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) = 1$, then

$$\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[a(X) \mid W_0 = 1]}{a_{\max}}. \tag{4.3}$$

If $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) < 1$, then $\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = 0$.

Here we see that the maximum size of a subpopulation characterizing the estimand $\mu(a, \tau_0)$ depends on $a(X)$ through two terms: its conditional mean in the numerator, and its supremum $a_{\max}$ in the denominator. This bound can be computed at what Imbens and Rubin (2015) call the "design stage" of the study, that is, without any knowledge of the conditional distribution of the outcome. The bound depends solely on the weight function $a(\cdot)$ and the distribution of $X \mid \{W_0 = 1\}$.

To understand the supremum's role in this expression, let $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid X, W_0 = 1)$ and note that $\mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ is equivalent to writing

$$\mu(a, \tau_0) = \frac{\mathbb{E}[a(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} = \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} = \mu(\underline{w}^*, \tau_0) \tag{4.4}$$

for all $\tau_0 \in \mathcal{T}_{\text{all}}$. Equation (4.4) holding for all τ_0 requires $\underline{w}^*(X)$ to be exactly proportional to $a(X)$. While the range of $a(X)$ is unconstrained, $\underline{w}^*(X)$ must lie in $[0, 1]$ to be a valid conditional probability. Since we seek to maximize $\mathbb{P}(W^* = 1 \mid W_0 = 1) = \mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]$, we let $\underline{w}^*(X)$ be the largest multiple of $a(X)$ that lies in $[0, 1]$ with probability 1, which is defined below:

$$W^* = \mathbb{1}\left(U \leq \frac{a(X)}{a_{\max}}\right) \cdot W_0 \qquad \text{and} \qquad \underline{w}^*(X) = \frac{a(X)}{a_{\max}}.$$

Here, $U \sim \text{Unif}(0, 1)$ and $U \perp\!\!\!\perp (Y(1), Y(0), X, W_0)$. This population places relatively more weight on units with larger values of $a(X)$. Specifically, the population $\{W^* = 1\}$ contains a random subset of $\{W_0 = 1\}$ where the probability of inclusion is proportional to $a(X)$. Thus, units with larger values of $a(X)$ are more likely to be included in W^*. All units in $\{W_0 = 1\}$ with X such that $a(X) = a_{\max}$ are included in W^*, whereas no units where $a(X) = 0$ are included.

The construction of this subpopulation is illustrated in Figure 1 for the case where x is continuous and where we omit the conditioning on $W_0 = 1$ for simplicity. We seek to maximize $\mathbb{P}(W^* = 1) = \int \underline{w}^*(x) f_X(x)\, dx$ with the requirement that $\underline{w}^*(x) \leq 1$ (or, equivalently, $\underline{w}^*(x) f_X(x) \leq f_X(x)$) and that $\underline{w}^*(x)$ is a multiple of $a(x)$. In the figure, we see that $a_{\max} > 1$ and thus the largest multiple of $a(x)$ that is weakly smaller than 1 is illustrated by the gray curve. The area under this curve is precisely $\mathbb{P}(W^* = 1)$. Note that the area under $f_X(x)$ is one, so closer alignment of the gray curve and the density $f_X(x)$ corresponds to more representative estimands.

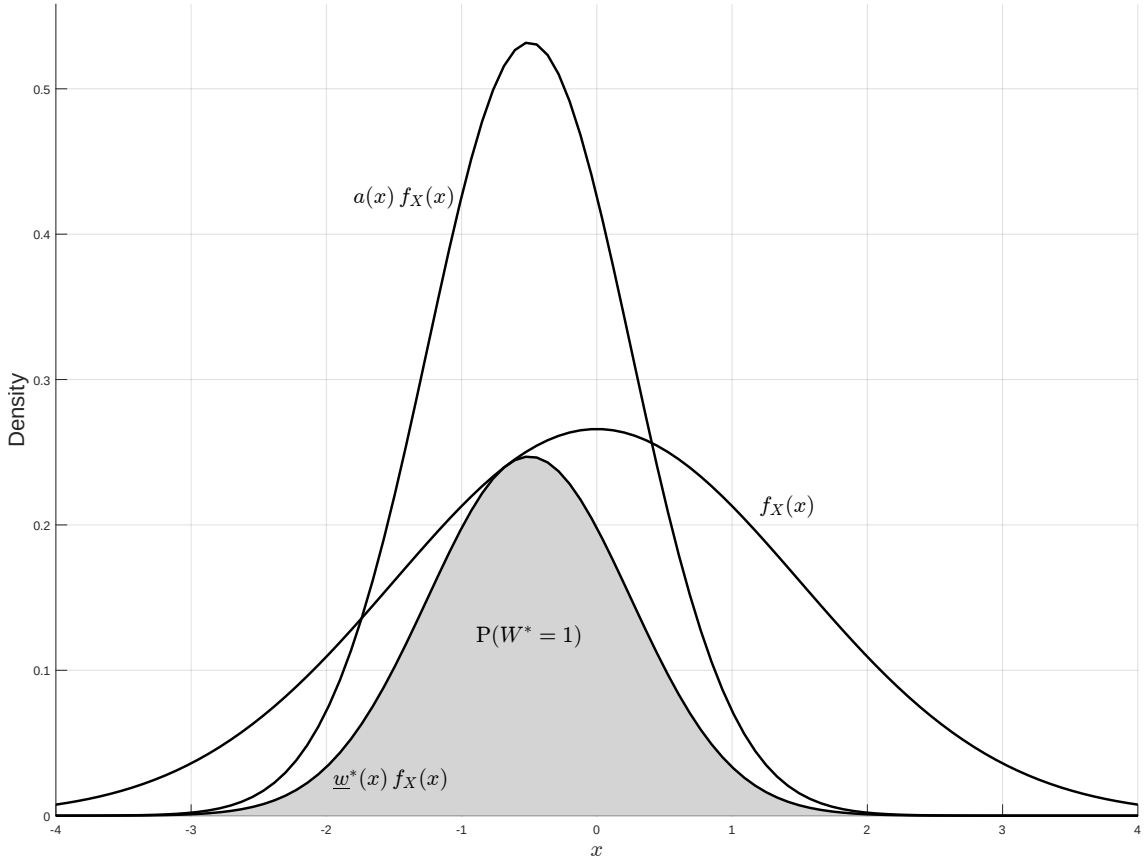
Several further comments about Theorem 4.1 are in order.

Remark 4.1 (Estimands and their corresponding interventions)**.** We note that estimand $\mu(a, \tau_0)$ is invariant to the scale of $a(\cdot)$ and thus can be written as

$$\mu(a, \tau_0) = \frac{\mathbb{E}\left[\frac{a(X)}{a_{\max}}\tau_0(X) \mid W_0 = 1\right]}{\mathbb{E}\left[\frac{a(X)}{a_{\max}} \mid W_0 = 1\right]},$$

where $a(X)/a_{\max} \in [0, 1]$ almost surely when weights are nonnegative. From this representation, we can link estimand $\mu(a, \tau_0)$ with an intervention where fraction $a(X)/a_{\max}$ of units with covariate value X and $W_0 = 1$ are treated. For example, if $W_0 = 1$ almost surely and $a(X) = a_{\max}$, then $\mu(a, \tau_0)$ is the average treatment effect, $\mathbb{E}[Y(1) - Y(0)]$, and it measures the average effect of treatment among all units. Under

Figure 1: Characterizing a Representative Subpopulation Uniformly in $\mathcal{T}_{\text{all}}$



Notes: X is a single continuously distributed covariate.

unconfoundedness, we also note that the average treatment effect on the treated (ATT) can be written as

$$\mathbb{E}[Y(1) - Y(0) \mid D = 1] = \frac{\mathbb{E}[\mathbb{P}(D = 1 \mid X) \cdot \tau_0(X)]}{\mathbb{E}[\mathbb{P}(D = 1 \mid X)]}.$$

Thus, a weighted estimand with weights $a(X) = \mathbb{P}(D = 1 \mid X)$ can be interpreted either as the effect of an intervention where fraction $\mathbb{P}(D = 1 \mid X)$ of units with covariate X are treated or as the effect of an intervention where all treated units are treated. In our setting, we can interpret any weighted estimand with nonnegative weights as the effect of treatment for a feasible intervention defined only in terms of X, W_0, and independent noise $U \sim \text{Unif}(0, 1)$ via $W^* = \mathbb{1}(U \leq a(X)/a_{\max}) \cdot W_0$. However, some of these estimands correspond to interventions that are more likely to be of interest to researchers, such as the ATE or ATT.

Remark 4.2 (Uniqueness of representative subpopulations)**.** It is also worth noting that the subpopulation maximizing the level of internal validity is generally not unique. The population $W^* = \mathbb{1}\left(U \leq a(X)/a_{\max}\right) \cdot W_0$ will generally change if U is replaced by another draw from a uniform distribution. The probability (conditional on X) of any unit being part of W^* does not change with the draw of U, but whether any given unit is part of subpopulation $\{W^* = 1\}$ cannot be determined.

Remark 4.3 (Distributional characteristics of representative subpopulations)**.** We can generally identify

distributional characteristics of units within the population W^*. For example, when W_0 is set to 1 almost surely, we can write the average values of $g(X)$ among subpopulation $\{W^* = 1\}$ as

$$\mathbb{E}[g(X) \mid W^* = 1] = \frac{\mathbb{E}[g(X)\underline{w}^*(X)]}{\mathbb{E}[\underline{w}^*(X)]} = \frac{\mathbb{E}[g(X)a(X)]}{\mathbb{E}[a(X)]},$$

a simple function of weights $a(\cdot)$ and the marginal distribution of X. We can recover the average covariate values in $\{W^* = 1\}$ by setting $g(X) = X$, or the entire distribution by considering $g(X) = \mathbb{1}(X \leq x)$ for all $x \in \mathbb{R}^{d_X}$. Reporting the average covariate values of units within and outside of W^* can be of interest when assessing the representativeness of $\mu(a, \tau_0)$.

Remark 4.4 (Defining the target estimand from the weighted estimand)**.** Suppose we consider $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ to be the target estimand, where $W^* = \mathbb{1}\left(U \leq a(X)/a_{\max}\right) \cdot W_0$ is the subpopulation for which $\mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ uniformly in τ_0. For example, in the case of the OLS estimand in Section 2, this consists of a subpopulation where the probability of inclusion is proportional to $\text{var}(D \mid X)$, the conditional variance of treatment. If this subpopulation is the target, it would be reasonable to infer that the measure of internal validity of the estimand $\mu(a, \tau_0)$ is the maximum value of 1. Indeed, this is the case because the estimand can be written as

$$\mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$$

and $\{W^* = 1\}$ is trivially the largest regular subpopulation of $\{W^* = 1\}$. This illustrates how the internal validity of $\mu(a, \tau_0)$ is entirely dependent on the target estimand. On the other hand, the representativeness of the weighted estimand, as measured by the largest value of $\mathbb{P}(W^* = 1)$, is independent of this target. In this case, it is less than one unless $W^* = 1$ almost surely, or that the weighted estimand actually equals the ATE. Theorem 4.4 below can also be applied to obtain this intuition.

We now consider a simple example to give further intuition for Theorem 4.1.

4.1.1 Illustrative Example: A Single Binary Covariate

Consider an estimand $\mu(a, \tau_0)$ where $W_0 = 1$ almost surely, $a(X) \geq 0$, and where X is binary with support $\text{supp}(X) = \{1, 2\}$. Let $p_x = \mathbb{P}(X = x)$ for $x \in \{1, 2\}$. As in Section 3.1, the weighted estimand can be written as a linear combination of the two CATEs:

$$\mu(a, \tau_0) = \frac{a(1)p_1}{\mathbb{E}[a(X)]}\tau_0(1) + \frac{a(2)p_2}{\mathbb{E}[a(X)]}\tau_0(2) \coloneqq \omega_1\tau_0(1) + \omega_2\tau_0(2).$$

Let $\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$ be the target estimand, which can be written as

$$\text{ATE} = p_1\tau_0(1) + p_2\tau_0(2).$$

If $a(1) = a(2)$, the relative weights placed on $\{X = 1\}$ and $\{X = 2\}$ by the estimand are equal to p_1/p_2, the ratio of the weights placed by the ATE. Therefore, the estimand equals the ATE and thus clearly has the maximum degree of internal validity with respect to the ATE. Applying Theorem 4.1, we can directly see

that

$$\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[a(X)]}{\sup_{x\in\{1,2\}} a(x)} = \frac{a(1)p_1 + a(2)p_2}{\max\{a(1), a(2)\}} = \frac{a(2)(p_1+p_2)}{a(2)} = 1$$

when $a(1) = a(2)$.

However, when $a(1) \neq a(2)$, the estimand's weights differ from (p_1, p_2), the population weights for the two covariate cells. For concreteness, let $(p_1, p_2) = (0.2, 0.8)$ and $(a(1), a(2)) = (0.24, 0.09)$, where the latter correspond, for example, to the OLS weights of Proposition 2.1 when the propensity score is $(p(1), p(2)) = (0.4, 0.1)$. In this case, $(\omega_1, \omega_2) = (0.4, 0.6)$ and thus

$$\mu(a, \tau_0) = 0.4 \cdot \tau_0(1) + 0.6 \cdot \tau_0(2) \qquad \text{and} \qquad \text{ATE} = 0.2 \cdot \tau_0(1) + 0.8 \cdot \tau_0(2).$$

Relative to the ATE, the estimand overrepresents the population with $X = 1$ and underrepresents the population with $X = 2$. The largest subpopulation $\{W^* = 1\}$ that causally represents the estimand can be constructed by combining subsets of the subpopulations defined by $\{X = 1\}$ and $\{X = 2\}$. Specifically, let

$$W^* = \mathbb{1}(X = 1) + \mathbb{1}\left(U \leq \frac{a(2)}{a(1)}, X = 2\right) = \mathbb{1}(X = 1) + \mathbb{1}\left(U \leq \frac{3}{8}, X = 2\right),$$

where $U \sim \text{Unif}(0, 1)$ is independent of $(Y(1), Y(0), X)$. This is a regular subpopulation that contains all units with $X = 1$ and three eighths of units with $X = 2$, selected uniformly at random. Therefore

$$\mathbb{P}(W^* = 1 \mid X = 1) = \underline{w}^*(1) = 1 \qquad \text{and} \qquad \mathbb{P}(W^* = 1 \mid X = 2) = \underline{w}^*(2) = \frac{3}{8},$$

which yields $\mathbb{P}(W^* = 1) = p_1\underline{w}^*(1) + p_2\underline{w}^*(2) = 0.5$. The same quantity can be obtained from Theorem 4.1, which implies that $\mathbb{E}[a(X)]/\left(\sup_{x\in\{1,2\}} a(x)\right) = (a(1)p_1 + a(2)p_2)/a(1) = 0.5$. The average treatment effect in this subpopulation is given by

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] &= \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X)]}{\mathbb{P}(W^* = 1)} \\
&= 2\left(1 \cdot \tau_0(1) \cdot 0.2 + 3/8 \cdot \tau_0(2) \cdot 0.8\right) \\
&= 0.4 \cdot \tau_0(1) + 0.6 \cdot \tau_0(2),
\end{aligned}$$

which equals $\mu(a, \tau_0)$ for any choice of τ_0. Note that the relative weights placed on $\{X = 1\}$ and $\{X = 2\}$ in subpopulation $\{W^* = 1\}$ are given by

$$\frac{\mathbb{P}(X = 1 \mid W^* = 1)}{\mathbb{P}(X = 2 \mid W^* = 1)} = \frac{0.4}{0.6} = \frac{\omega_1}{\omega_2},$$

matching the ratio of the weights on $\{X = 1\}$ and $\{X = 2\}$ assigned by the estimand. The subpopulation $\{W^* = 1\}$ cannot expand while preserving this ratio since it already includes all units with $X = 1$. Therefore, W^* is the largest subpopulation for which $\mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ for any τ_0.

4.2 Using the Measure of Internal Validity to Bound Average Effects

The subpopulation size in Theorem 4.1 can be used to bound the target estimand $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$. Consider a scenario where only the baseline estimand $\mu(a, \tau_0)$, which we assume has a causal representation uniformly in $\mathcal{T}_{\text{all}}$, and its internal validity are known. For example, this could be the case if a researcher uses a weighted estimand (e.g., OLS) and reports the measure we propose in Definition 4.1 to measure its degree of internal validity for the ATE. To simplify notation, assume that $W_0 = 1$ almost surely. Abstracting from sample uncertainty, we only assume knowledge of the weighted estimand $\mu(a, \tau_0)$ and its internal validity, given by $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$. We can decompose the target estimand, here the ATE, as

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1] \cdot \mathbb{P}(W^* = 1) + \mathbb{E}[Y(1) - Y(0) \mid W^* = 0] \cdot (1 - \mathbb{P}(W^* = 1))$$

for a $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{all}})$. If we have knowledge of bounds for the treatment effect $\mathbb{E}[Y(1) - Y(0) \mid W^* = 0]$, e.g. from the support of the potential outcomes, we can obtain bounds on $\mathbb{E}[Y(1) - Y(0)]$. For example, if $\text{supp}(Y(1) - Y(0)) = [B_\ell, B_u]$, bounds for the target estimand are given by

$$\left[\mu(a, \tau_0) \cdot \mathbb{P}(W^* = 1) + B_\ell \cdot (1 - \mathbb{P}(W^* = 1)),\ \mu(a, \tau_0) \cdot \mathbb{P}(W^* = 1) + B_u \cdot (1 - \mathbb{P}(W^* = 1))\right]. \tag{4.5}$$

The width of those bounds is minimized when $\mathbb{P}(W^* = 1)$ is maximized, or when it equals the measure of representativeness for $\mu(a, \tau_0)$, given by $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$. The resulting bounds are

$$\left[\mu(a, \tau_0) \cdot \overline{P}(a, W_0; \mathcal{T}_{\text{all}}) + B_\ell \cdot (1 - \overline{P}(a, W_0; \mathcal{T}_{\text{all}})),\ \mu(a, \tau_0) \cdot \overline{P}(a, W_0; \mathcal{T}_{\text{all}}) + B_u \cdot (1 - \overline{P}(a, W_0; \mathcal{T}_{\text{all}}))\right]. \tag{4.6}$$

If $\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = 1$, it is easy to see that the estimand equals the ATE and that the bounds in (4.6) collapse to a point. However, the ATE is not uniquely determined from $(\mu(a, \tau_0), \overline{P}(a, W_0; \mathcal{T}_{\text{all}}))$ when $\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) < 1$.

The width of these bounds is $(B_u - B_\ell) \cdot (1 - \overline{P}(a, W_0; \mathcal{T}_{\text{all}}))$. Hence for fixed (B_ℓ, B_u), this width decreases linearly with $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$. Moreover, values of $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$ close to 1, or high degrees of internal validity, lead to narrow bounds. For example, if outcomes are binary with support $\{0, 1\}$, then $[B_\ell, B_u] = [-1, 1]$ and knowing the values of $(\mu(a, \tau_0), \overline{P}(a, W_0; \mathcal{T}_{\text{all}}))$ constrains the ATE to lie in

$$\left[\mu(a, \tau_0) \cdot \overline{P}(a, W_0; \mathcal{T}_{\text{all}}) - (1 - \overline{P}(a, W_0; \mathcal{T}_{\text{all}})),\ \mu(a, \tau_0) \cdot \overline{P}(a, W_0; \mathcal{T}_{\text{all}}) + (1 - \overline{P}(a, W_0; \mathcal{T}_{\text{all}}))\right].$$

In this case, the bounds are centered at the estimand multiplied by our measure of internal validity, while the width of the bounds equals $2 \cdot (1 - \overline{P}(a, W_0; \mathcal{T}_{\text{all}}))$. It is easy to obtain a sample analog of these bounds by combining estimators for $\mu(a, \tau_0)$ and our proposed estimator for $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$ from Section 6 below.

We note that bounds on $\mathbb{E}[Y(1) - Y(0)]$ may be tightened by assuming knowledge of other aspects of the joint distribution of $(Y(1), Y(0), D, X)$. For example, if knowledge of $a(\cdot)$ is assumed, additional constraints on $\mathbb{E}[Y(1) - Y(0)]$ can help narrow the bounds given in (4.6). We focus here on the case where we add a single piece of additional information to $\mu(a, \tau_0)$, namely the internal validity of $\mu(a, \tau_0)$, and how simple bounds can be obtained from the estimand and our proposed measure. We leave refinements of such bounds under different information sets to future work.

Bounding Average Effects with Negative Weights

Now consider a case where the weighted estimand $\mu(a, \tau_0)$ has weights that are sometimes negative, i.e., $\mathbb{P}(a(X) < 0) > 0$. We continue to assume that $W_0 = 1$ almost surely. In this case, we know that $\mu(a, \tau_0)$ does not have a causal representation uniformly in $\tau_0 \in \mathcal{T}_{\text{all}}$ and thus we cannot write $\mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ uniformly in τ_0. However, simple algebra reveals that an estimand with negative weights can be written as a weighted difference of two nonnegatively weighted estimands:

$$\begin{aligned}\mu(a, \tau_0) &= \frac{\mathbb{E}[(a(X)\mathbb{1}(a(X) \geq 0) + a(X)\mathbb{1}(a(X) \leq 0))\tau_0(X)]}{\mathbb{E}[a(X)]} \\ &= \omega^+ \cdot \mu(a \cdot \mathbb{1}(a \geq 0), \tau_0) - \omega^- \cdot \mu((-a) \cdot \mathbb{1}(a \leq 0), \tau_0),\end{aligned}$$

where $\omega^+ \coloneqq \mathbb{E}[a(X)\mathbb{1}(a(X) \geq 0)]/\mathbb{E}[a(X)]$ and $\omega^- \coloneqq \mathbb{E}[-a(X)\mathbb{1}(a(X) \leq 0)]/\mathbb{E}[a(X)]$ are both nonnegative, and $\omega^+ - \omega^- = 1$. We note that $(\omega^+, \omega^-) = (1, 0)$ when the estimand's weights are nonnegative, so this decomposition can be obtained regardless of the sign of a. Thus, by Theorem 3.1, we can write

$$\mu(a, \tau_0) \;=\; \omega^+ \cdot \mathbb{E}[Y(1) - Y(0) \mid W^+ = 1] \;-\; \omega^- \cdot \mathbb{E}[Y(1) - Y(0) \mid W^- = 1], \tag{4.7}$$

where W^+ and W^- characterize two disjoint, regular subpopulations. As above, suppose we want to bound $\mathbb{E}[Y(1) - Y(0)]$, the average treatment effect. Using the law of iterated expectations, we can write

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0)] \;=\; &\mathbb{P}(W^+ = 1) \cdot \mathbb{E}[Y(1) - Y(0) \mid W^+ = 1] \;+\; \mathbb{P}(W^- = 1) \cdot \mathbb{E}[Y(1) - Y(0) \mid W^- = 1] \\ &+\; \big[1 - \mathbb{P}(W^+ = 1) - \mathbb{P}(W^- = 1)\big] \cdot \mathbb{E}[Y(1) - Y(0) \mid W^+ + W^- = 0].\end{aligned} \tag{4.8}$$

Substituting equation (4.7) in (4.8) and assuming that $\mathbb{E}[Y(1) - Y(0) \mid W^- = 1]$ and $\mathbb{E}[Y(1) - Y(0) \mid W^+ + W^- = 0]$ lie in $[B_\ell, B_u]$ yields

$$\mathbb{E}[Y(1) - Y(0)] \;\leq\; \frac{\mathbb{P}(W^+ = 1)}{\omega^+} \cdot \mu(a, \tau_0) \;+\; \left(1 - \frac{\mathbb{P}(W^+ = 1)}{\omega^+}\right) \cdot B_u$$

as an upper bound for the ATE. A lower bound is obtained by replacing B_u with B_ℓ. Thus, the ATE lies in the interval

$$\left[\frac{\mathbb{P}(W^+ = 1)}{\omega^+} \cdot \mu(a, \tau_0) + \left(1 - \frac{\mathbb{P}(W^+ = 1)}{\omega^+}\right) \cdot B_\ell,\; \frac{\mathbb{P}(W^+ = 1)}{\omega^+} \cdot \mu(a, \tau_0) + \left(1 - \frac{\mathbb{P}(W^+ = 1)}{\omega^+}\right) \cdot B_u\right]. \tag{4.9}$$

This interval is similar to the interval in (4.6), but the latter is only valid when weights are nonnegative. These intervals are identical when weights are nonnegative because $\omega^+ = 1$ and $\mathbb{P}(W^+ = 1) = \mathbb{P}(W^* = 1)$ in that case. In order to compute the interval in (4.9) and minimize its length, one needs to maximize the value of $\mathbb{P}(W^+ = 1)$, which corresponds to the level of internal validity of the estimand $\mu(a \cdot \mathbb{1}(a \geq 0), \tau_0)$, which has nonnegative weights, and compute the value of ω^+. The interval depends only on the ratio of the two quantities, which can be written as

$$\frac{\mathbb{P}(W^+ = 1)}{\omega^+} \leq \frac{\overline{P}(a \cdot \mathbb{1}(a \geq 0), W_0; \mathcal{T}_{\text{all}})}{\omega^+} = \frac{\mathbb{E}[a(X)\mathbb{1}(a(X) \geq 0)]/\sup(\text{supp}(a(X)\mathbb{1}(a(X) \geq 0)))}{\mathbb{E}[a(X)\mathbb{1}(a(X) \geq 0)]/\mathbb{E}[a(X)]} = \frac{\mathbb{E}[a(X)]}{a_{\max}}.$$

This last expression equals the level of internal validity for the original estimand $\mu(a, \tau_0)$ when it is assumed (perhaps incorrectly) to have nonnegative weights. Thus, procedures used to quantify the internal validity of weighted estimands can also be used on estimands with negative weights if the goal is to bound average effects. As mentioned earlier, these bounds do not make use of the entire distribution of (Y, D, X), but simply of the original estimand $\mu(a, \tau_0)$ and of $\mathbb{E}[a(X)]/a_{\max}$, the expression for the level of internal validity under nonnegative weights.

4.3 Quantifying Internal Validity Given τ_0

We can also ask how internally valid a weighted estimand can be, given knowledge of the CATE function. In this case, the object of interest is

$$\overline{P}(a, W_0; \{\tau_0\}) = \sup_{W^* \in \mathcal{W}(a; W_0, \{\tau_0\})} \mathbb{P}(W^* = 1 \mid W_0 = 1), \tag{4.10}$$

where τ_0 is a given CATE function. Since τ_0 is known, the condition $W^* \in \mathcal{W}(a; W_0, \{\tau_0\})$ can be written as

$$\mu(a, \tau_0) = \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]}$$

or, equivalently,

$$\mathbb{E}[(\tau_0(X) - \mu(a, \tau_0))\underline{w}^*(X) \mid W_0 = 1] = 0, \tag{4.11}$$

where $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid W_0 = 1, X)$. Equation (4.11) is a linear constraint on the conditional probability of being in subpopulation W^*. Additionally, the objective function $\mathbb{P}(W^* = 1 \mid W_0 = 1) = \mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]$ is linear in $\underline{w}^*$. Thus, the optimization in (4.10) can be cast as a linear program. To see this, consider as an example the case where $W_0 = 1$ almost surely and where X is discrete with finite support, i.e. $\text{supp}(X) = \{x_1, \ldots, x_K\}$. Let $f_k \coloneqq \mathbb{P}(W^* = 1, X = x_k)$ denote the probability of being in subpopulation W^* and having covariate value x_k. Note that $f_k \in [0, p_k]$ where $p_k = \mathbb{P}(X = x_k)$.

We can write the above optimization problem as

$$\max_{(f_1, \ldots, f_K) \geq \mathbf{0}} \sum_{k=1}^{K} f_k \qquad \text{such that } f_k \leq p_k \text{ for } k \in \{1, \ldots, K\},$$
$$\text{and } \sum_{k=1}^{K} (\tau_0(x_k) - \mu(a, \tau_0)) f_k = 0,$$

a finite-dimensional linear program. This program has a feasible solution if $\tau_0(x_k) - \mu(a, \tau_0)$ is not strictly positive or strictly negative for all k, meaning that the weighted estimand lies in the convex hull of CATE values, which is precisely stated in the condition for Theorem 3.2. While there exist many methods for solving linear programs, the value function can be obtained through an algorithm that is simple to describe analytically.

Let $\mathbf{f} = (f_1, \ldots, f_K)$, $\mathbf{p} = (p_1, \ldots, p_K)$, and $\mathbf{t}_\mu = (\tau_0(x_1) - \mu(a, \tau_0), \ldots, \tau_0(x_K) - \mu(a, \tau_0))$. Without loss

of generality, assume that $\tau_0(x_1) - \mu(a, \tau_0) \leq \tau_0(x_2) - \mu(a, \tau_0) \leq \cdots \leq \tau_0(x_K) - \mu(a, \tau_0)$.

1. Set $\mathbf{f} = \mathbf{p}$.
2. If $\mathbf{t}_\mu' \mathbf{f} = 0$, end the algorithm and report $\sum_{k=1}^K f_k$.
3. If $\mathbf{t}_\mu' \mathbf{f} \neq 0$:
   (a) If $\mathbf{t}_\mu' \mathbf{f} > 0$, let $k^* = \max\{k \in \{1, \ldots, K\} : f_k = p_k\}$ and set $f_{k^*} = \max\left\{0, \frac{-\sum_{k=1}^{k^*-1}(\tau_0(x_k) - \mu(a,\tau_0))p_k}{\tau_0(x_{k^*}) - \mu(a,\tau_0)}\right\}$.
   (b) If $\mathbf{t}_\mu' \mathbf{f} < 0$, let $k^* = \min\{k \in \{1, \ldots, K\} : f_k = p_k\}$ and set $f_{k^*} = \max\left\{0, \frac{-\sum_{k=k^*+1}^{K}(\tau_0(x_k) - \mu(a,\tau_0))p_k}{\tau_0(x_{k^*}) - \mu(a,\tau_0)}\right\}$.
4. Go to step 2.

When $\mu(a, \tau_0)$ exceeds $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, this algorithm reduces the weights associated with largest CATEs until $\mu(a, \tau_0)$ equals $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ for some subpopulation. When $\mu(a, \tau_0) < \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, the same procedure is instead applied to the smallest CATEs. The support assumption of Theorem 3.2 guarantees that this algorithm ends.

When X is not discretely supported, the problem can still be cast as a linear program, but its dimension may be infinite, which generates difficulties in implementation. However, we show this program has an analytical solution for the size of the subpopulation of interest, $\mathbb{P}(W^* = 1 \mid W_0 = 1)$. The following theorem gives its expression and is valid when vector X has an arbitrary kind of distribution, with continuous, discrete, and mixed components, as is often the case in empirical applications.

Theorem 4.2. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds. Let $T_\mu = \tau_0(X) - \mu(a, \tau_0)$. If $\mu(a, \tau_0) \in \mathcal{S}(\tau_0; W_0)$,

$$
\begin{aligned}
&\overline{P}(a, W_0; \{\tau_0\}) \\
&= \begin{cases}
\mathbb{P}(T_\mu \leq \alpha^+ \mid W_0 = 1) - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+} \\
\qquad \text{where } \alpha^+ = \inf\{\alpha \in \mathbb{R} : \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha) \mid W_0 = 1] \geq 0\} \text{ if } \mu(a, \tau_0) < \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1] \\
\mathbb{P}(T_\mu \geq \alpha^- \mid W_0 = 1) - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha^-) \mid W_0 = 1]}{\alpha^-} \\
\qquad \text{where } \alpha^- = \sup\{\alpha \in \mathbb{R} : \mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha) \mid W_0 = 1] \leq 0\} \text{ if } \mu(a, \tau_0) > \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1] \\
1 \qquad \text{if } \mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1].
\end{cases}
\end{aligned}
\tag{4.12}
$$

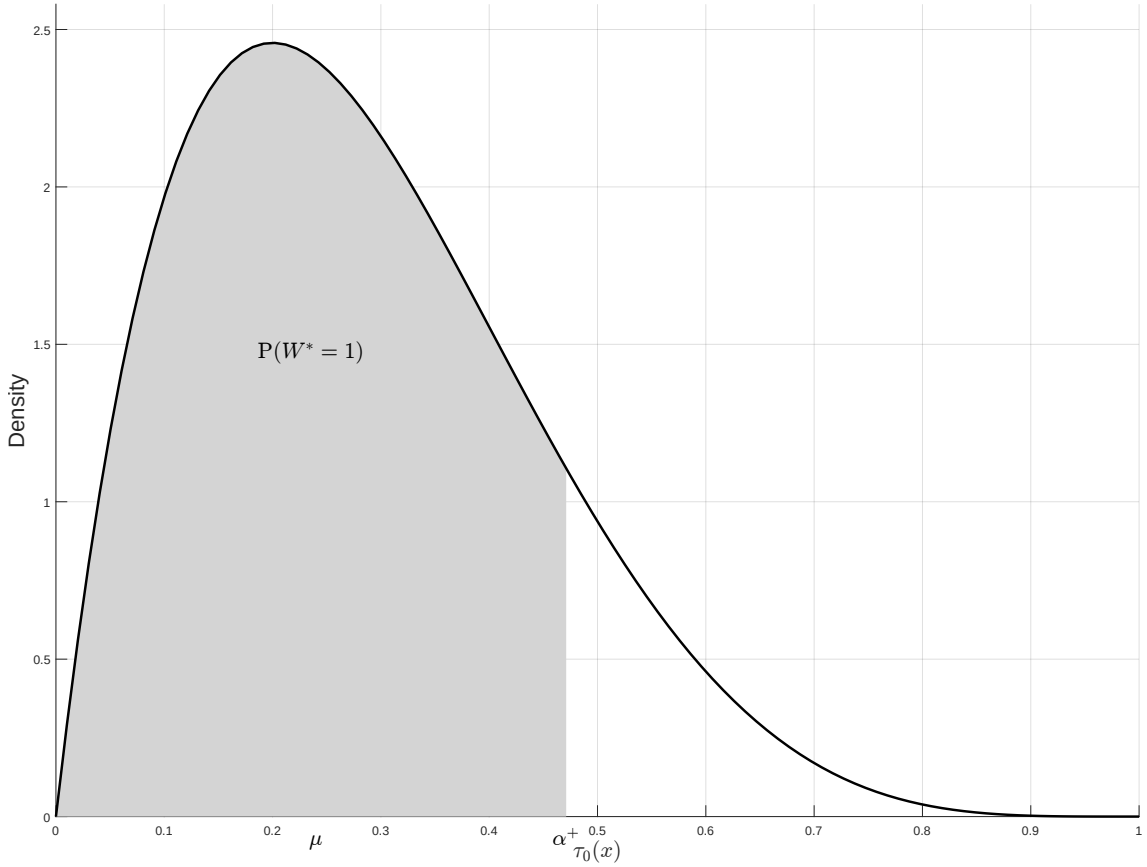
If $\mu(a, \tau_0) \notin \mathcal{S}(\tau_0; W_0)$, then $\overline{P}(a, W_0; \{\tau_0\}) = 0$.

The computation of these bounds can be done using a linear programming algorithm when X is discrete, or through plug-in estimators of the terms in equation (4.12) regardless of the nature of the support of X.

In this setting, the value $\overline{P}(a, W_0; \{\tau_0\})$ is larger when the truncated subpopulations are smaller. In particular, this is the case when there are a few units with extreme values of τ_0 whose removal has a large impact on the estimand, but a small impact on the share of the population.

Theorem 4.2 can also be illustrated visually. In Figure 2, the probability density function of $\tau_0(X)$ is drawn. In this figure, it is assumed that $\tau_0(X)$ is continuously distributed, that $W_0 = 1$ almost surely, and

Figure 2: Characterizing a Representative Subpopulation When τ_0 Is Known



Notes: The figure assumes that $\tau_0(X)$ is continuously distributed, that $W_0 = 1$ almost surely, and that $\mu < \mathbb{E}[\tau_0(X)] = \text{ATE}$.

that $\mu < \mathbb{E}[\tau_0(X)] = \text{ATE}$. The representative subpopulation is obtained by trimming away covariate values that correspond to $\tau_0(X) \geq \alpha^+$, where α^+ is determined by the equation $\mathbb{E}[\tau_0(X) \mid \tau_0(X) \leq \alpha^+] = \mu$. The size of the shaded area is the measure of internal validity.

4.3.1 Illustrative Example: A Single Continuous Covariate

To illustrate the previous theorem, let X be a continuously distributed covariate with support $[x_L, x_U]$. Let $W_0 = 1$ almost surely, and let $\mu \coloneqq \mu(a, \tau_0)$ denote the estimand. Also, for simplicity assume that $\tau_0(x)$ is increasing in x and that $\tau_0(x_L) < \mu < \tau_0(x_U)$. Without loss of generality, let $\mathbb{E}[Y(1) - Y(0)] \geq \mu$. If $\mathbb{E}[Y(1) - Y(0)] = \mu$, then the estimand is perfectly representative of the population since it equals the average treatment effect over it. Now consider the case where $\mathbb{E}[Y(1) - Y(0)] > \mu$. In this case, the estimand is larger than the ATE so it is not representative of the entire population.

We are seeking the largest subpopulation $\{W^* = 1\}$ such that $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \mu$. Equivalently, we can seek the *smallest* subpopulation $\{W^- = 1\}$ such that, when it is removed from the original population, the average treatment effect equals μ. The subpopulation $\{W^- = 1\}$ is related to $\{W^* = 1\}$ via $W^- + W^* = 1$. We will search for W^- such that $\mathbb{E}[Y(1) - Y(0) \mid W^- = 0] = \mu$ and $\mathbb{P}(W^- = 1)$ is minimized.

Since the ATE exceeds μ, we must have that $\mathbb{E}[\tau_0(X) \mid W^- = 1] > \mu$. For a given subpopulation of size $\mathbb{P}(W^- = 1)$, removing the subpopulation with the largest values of $\tau_0(x)$ yields the largest decrease in $\mathbb{E}[Y(1) - Y(0) \mid W^- = 0]$. Therefore, $\overline{P}(a, W_0; \{\tau_0\})$ is obtained by removing a subpopulation of the kind

$$W^-(\alpha) = \mathbb{1}(\tau_0(X) > \alpha)$$

for a given threshold α. This subpopulation consists of units whose CATEs exceed α. This threshold is determined by the constraint $\mathbb{E}[\tau_0(X) \mid W^-(\alpha) = 0] = \mu$, or

$$\mathbb{E}[\tau_0(X) \mid \tau_0(X) \leq \alpha] = \mu. \tag{4.13}$$

This constraint states that once units with covariate values satisfying $\tau_0(X) > \alpha$ are removed, the average treatment effect for the remaining units equals the weighted estimand. Since we assumed τ_0 is increasing, W^* corresponds to all units with covariate values below threshold $\delta = \tau_0^{-1}(\alpha)$. Therefore,

$$W^* = \mathbb{1}(X \leq \delta^+),$$

where $\delta^+ = \tau_0^{-1}(\alpha^+)$ and α^+ is the unique solution to (4.13). How large $\mathbb{P}(W^* = 1)$ is depends on the size of this fraction of removed units. A smaller subpopulation needs to be removed when $\tau_0(X)$ has a long right tail, and a larger subpopulation needs to be removed when the distribution of $\tau_0(X)$ in this right tail is shorter.

4.4 Generalizing to Subpopulations

The results from the previous two sections can be generalized to cases where we hold the average effect for a subset of W_0 as the object of interest. Let W' be a regular subpopulation of W_0 and $\mathbb{E}[Y(1) - Y(0) \mid W' = 1]$ be the target parameter. In analogy with the previous sections, we ask *(i)* when does there exist a regular subpopulation W^* of W' such that $\mu(a, \tau_0)$ can be written as the average treatment effect over W^* and *(ii)* how representative of W' is this estimand.

For concreteness, consider the example of Section 2 where $W_0 = 1$. We can let $W' = D$, the treated subpopulation, and have the ATT as the target parameter. This section's results can be used to show that there exists a uniform causal representation of the OLS estimand as an average effect for a subpopulation of the treated units. These results also allow us to compute the internal validity of the OLS estimand for this target subpopulation.

We first present a result showing conditions for the existence of such population W^* in the case where τ_0 is unrestricted, and when τ_0 is fixed. In what follows, we let $\underline{w}'(X) = \mathbb{P}(W' = 1 \mid X, W_0 = 1)$.

Theorem 4.3. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds and that W' is a regular subpopulation of W_0.

1. (Uniformly in τ_0) Suppose $\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W_0 = 1)) < \infty$.[3] Then, there exists $W^* \in \mathcal{W}(a; W', \mathcal{T}_{\text{all}})$ if and only if $\mathbb{P}(a(X) \geq 0 \mid W' = 1) = 1$.

[3] We use the conventions $0/0 = 1$ and $a/0 = \infty$ when $a > 0$.

2. (Given τ_0) There exists $W^* \in \mathcal{W}(a; W', \{\tau_0\})$ if and only if $\mu(a, \tau_0) \in \mathcal{S}(\tau_0; W')$.

Note that letting $W' = W_0$ yields Theorems 3.1 and 3.2 as special cases since W_0 is a regular subpopulation of itself by definition.

The condition $\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W_0 = 1)) < \infty$ rules out uniform causal representations if there are covariate values for which $a(x)$ is strictly positive but are not represented in W', i.e. $\mathbb{P}(W' = 1 \mid X = x, W_0 = 1) = 0$ when $a(x) > 0$. For example, if W' is a subpopulation of W_0 defined by a subset of covariate values such that $W' = \mathbb{1}(X \in \mathcal{X}_0) \cdot W_0$, then W' is a regular subpopulation but will fail the above inequality if $a(X) > 0$ for $X \in \text{supp}(X \mid W_0 = 1) \setminus \mathcal{X}_0$. In the case where τ_0 is known, the existence of a representative subpopulation of W' is equivalent to the estimand being in the convex hull of $\text{supp}(\tau_0(X) \mid W' = 1)$.

Regarding the internal validity of a weighted estimand, we generalize Theorems 4.1 and 4.2 and ask how large $\mathbb{P}(W^* = 1 \mid W' = 1)$ can be given that W^* is a regular subpopulation of W', which is itself a regular subpopulation of W_0.

Theorem 4.4. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds and that W' is a regular subpopulation of W_0.

1. (Uniformly in τ_0) Suppose $\sup\left(\text{supp}\left(a(X)/\underline{w}'(X) \mid W_0 = 1\right)\right) < \infty$. If $\mathbb{P}(a(X) \geq 0 \mid W' = 1) = 1$, then

$$\overline{P}(a, W'; \mathcal{T}_{\text{all}}) = \mathbb{E}[a(X) \mid W_0 = 1] \cdot \frac{\mathbb{P}(W_0 = 1)}{\mathbb{P}(W' = 1)} \cdot \inf\left(\text{supp}\left(\frac{\underline{w}'(X)}{a(X)} \mid W' = 1\right)\right). \tag{4.14}$$

   If $\mathbb{P}(a(X) \geq 0 \mid W' = 1) < 1$, then $\overline{P}(a, W'; \mathcal{T}_{\text{all}}) = 0$.

2. (Given τ_0) If $\mu(a, \tau_0) \in \mathcal{S}(\tau_0; W')$, then

$$\begin{aligned}
&\overline{P}(a, W'; \{\tau_0\}) \\
&= \begin{cases}
\mathbb{P}(T_\mu \leq \alpha^+ \mid W' = 1) - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W' = 1]}{\alpha^+} \\
\qquad \text{where } \alpha^+ = \inf\{\alpha \in \mathbb{R} : \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha) \mid W' = 1] \geq 0\} \text{ if } \mu(a, \tau_0) < \mathbb{E}[Y(1) - Y(0) \mid W' = 1] \\
\mathbb{P}(T_\mu \geq \alpha^- \mid W' = 1) - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha^-) \mid W' = 1]}{\alpha^-} \\
\qquad \text{where } \alpha^- = \sup\{\alpha \in \mathbb{R} : \mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha) \mid W' = 1] \leq 0\} \text{ if } \mu(a, \tau_0) > \mathbb{E}[Y(1) - Y(0) \mid W' = 1] \\
1 \qquad \text{if } \mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W' = 1].
\end{cases}
\end{aligned} \tag{4.15}$$

   If $\mu(a, \tau_0) \notin \mathcal{S}(\tau_0; W')$, then $\overline{P}(a, W'; \{\tau_0\}) = 0$.

We can also use this result to obtain bounds on $\mathbb{P}(W^* = 1)$ or $\mathbb{P}(W^* = 1 \mid W_0 = 1)$. This can be done by noting that W^* is a subpopulation of both W' and W_0, so

$$\mathbb{P}(W^* = 1) = \mathbb{P}(W^* = 1 \mid W' = 1) \cdot \mathbb{P}(W' = 1) \quad \text{and} \quad \mathbb{P}(W^* = 1 \mid W_0 = 1) = \frac{\mathbb{P}(W^* = 1 \mid W' = 1) \cdot \mathbb{P}(W' = 1)}{\mathbb{P}(W_0 = 1)}.$$

Thus, bounds on these probabilities are trivially obtained from the bound on $\mathbb{P}(W^* = 1 \mid W' = 1)$ from Theorem 4.4 and knowledge of $(\mathbb{P}(W' = 1), \mathbb{P}(W_0 = 1))$.

5 Applications to Common Estimands

Here we consider three identification strategies where commonly used estimands follow the structure of equation (1.1). We show how the results in sections 3 and 4 apply in each of these cases. For simplicity, we assume that $a_{\max} = \sup(\text{supp}(a(X) \mid W_0 = 1)) = \sup_{x \in \text{supp}(X \mid W_0 = 1)} a(x)$ for the remainder of the paper. This condition is satisfied when $a(\cdot)$ is continuous or when X has finite support, for example. We also note that our assumption $a_{\max} < \infty$ holds trivially in every case considered below.

5.1 Unconfoundedness

5.1.1 Causal Representation and Internal Validity of OLS

In Section 2, we provided the expression for the coefficient on D in a population regression of Y on $(1, D, X)$:

$$\beta_{\text{OLS}} = \frac{\mathbb{E}[p(X)(1 - p(X))\tau_0(X)]}{\mathbb{E}[p(X)(1 - p(X))]}.$$

Suppose the target estimand is the average treatment effect, i.e. $W_0 = 1$ almost surely. By Theorem 3.1, there exists a regular subpopulation W^* such that β_{OLS} equals the average treatment effect over W^* since the weight function $a(X) = p(X)(1 - p(X))$ is nonnegative. By Theorem 4.1, the upper bound on the size of subpopulation W^* is given by

$$\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[p(X)(1 - p(X))]}{\sup_{x \in \text{supp}(X)} p(x)(1 - p(x))}.$$

A corresponding subpopulation W^* can be written as

$$W^* = \mathbb{1}\left(U \leq \frac{p(X)(1 - p(X))}{\sup_{x \in \text{supp}(X)} p(x)(1 - p(x))}\right)$$

where $U \sim \text{Unif}(0, 1) \perp\!\!\!\perp (Y(1), Y(0), X)$. This is a subpopulation where units which have a larger variation in treatment given their covariate values are more likely to be included. The size of this subpopulation is largest when $\text{var}(D \mid X) = p(X)(1 - p(X))$ is constant, in which case $\mathbb{P}(W^* = 1 \mid X) = 1$. This is the case if and only if $p(X)$ has support equal to $\{b, 1 - b\}$ for some $b \in (0, 1)$. This is implied by $D \perp\!\!\!\perp X$, or random assignment. It can also be achieved if there exists a partition of $\text{supp}(X)$ where $\mathbb{P}(D = 1 \mid X) = b$ on one element and $\mathbb{P}(D = 1 \mid X) = 1 - b$ on its complement. Whenever $\text{var}(p(X)(1 - p(X))) > 0$, $\{W^* = 1\}$ will be a strict subpopulation.

The size of this subpopulation is the expectation of $\text{var}(D \mid X)$ divided by its maximum value. There are a few ways this expression can be further simplified or bounded. Its numerator is bounded above by $\text{var}(D) = \mathbb{P}(D = 1) \cdot \mathbb{P}(D = 0)$, which is particularly simple to estimate. As for the denominator, it is a nonsmooth functional of $p(\cdot)$. However, if X is continuously distributed, it may be likely that $p(X)$ is continuously distributed and thus that $1/2 \in \text{supp}(p(X))$. If this is the case, $\sup_{x \in \text{supp}(X)} p(x)(1 - p(x)) = 1/4$. Combining these two approximations yields

$$\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) \leq 4 \cdot \mathbb{P}(D = 1) \cdot \mathbb{P}(D = 0),$$

when the support of $p(X)$ includes $1/2$. This bound is trivial when $\mathbb{P}(D=1)=1/2$, but is informative when the unconditional treatment probability is close to 0 or 1. For example, if $\mathbb{P}(D=1)=0.1$, the OLS estimand cannot causally represent more than 36% of the population. This is consistent with the result in Słoczyński (2022) that the OLS estimand is more similar to the ATE when $\mathbb{P}(D=1)$ is close to $1/2$.

When $1/2 \in \text{supp}(p(X))$, we can also compute bounds on the ATE derived from the OLS estimand, bounds on the support of $(Y(1), Y(0))$, and our measure of representativeness $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$. Following the expression in (4.6), bounds on the ATE are given by

$$[(\beta_{\text{OLS}} - B_\ell) \cdot 4\mathbb{E}[\text{var}(D \mid X)] + B_\ell,\ (\beta_{\text{OLS}} - B_u) \cdot 4\mathbb{E}[\text{var}(D \mid X)] + B_u]\,.$$

Estimating these bounds requires the estimation of one additional quantity beyond the OLS estimator, which is the expectation of $\text{var}(D \mid X)$. The width of these bounds depends crucially on $B_u - B_\ell$, or the width of the support for unit-level treatment effects.

Alternatively, we can assess the internal validity of β_{OLS} with respect to an alternative estimand such as $\mathbb{E}[Y(1) - Y(0) \mid D=1]$, the average treatment effect on the treated. In this case, we can write

$$\beta_{\text{OLS}} = \frac{\mathbb{E}[(1-p(X))w_0(X)\tau_0(X)]}{\mathbb{E}[(1-p(X))w_0(X)]},$$

where $w_0(X) = \mathbb{P}(D=1 \mid X) = p(X)$, the propensity score, and $\tilde{a}(X) = \mathbb{P}(D=0 \mid X) = 1 - p(X)$. Applying Theorem 4.1 yields that

$$\overline{P}(\tilde{a}, D; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[1-p(X) \mid D=1]}{\sup_{x \in \text{supp}(X \mid D=1)}(1-p(x))} = \frac{\mathbb{E}[p(X)(1-p(X))]}{\mathbb{P}(D=1) \cdot \sup_{x \in \text{supp}(X \mid D=1)}(1-p(x))}$$

is the largest value that $\mathbb{P}(W^*=1 \mid D=1)$ can take. Once again, this bound depends only on the propensity score and the distribution of X. This bound can also be obtained by using Theorem 4.4 and setting $W_0 = 1$ and $W' = D$. This subpopulation satisfies

$$\mathbb{P}(W^*=1 \mid X, D=1) = \frac{1-p(X)}{1 - \inf_{x \in \text{supp}(X \mid D=1)} p(X)}$$

so units with smaller propensity scores are more likely to be included in W^*, given that they are treated. $\overline{P}(a, D; \mathcal{T}_{\text{all}})$ is maximized at 1 when $p(X)$ is constant, or if $D \perp\!\!\!\perp X$. In this case, $\mathbb{P}(W^*=1 \mid D=1)=1$ and $\mathbb{P}(W^*=1) = \mathbb{P}(D=1)$.

If $p(X)$ takes values close to 0, this bound equals

$$\overline{P}(a, D; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[p(X)(1-p(X))]}{\mathbb{P}(D=1)} \leq \frac{\mathbb{P}(D=1) \cdot \mathbb{P}(D=0)}{\mathbb{P}(D=1)} = \mathbb{P}(D=0).$$

This suggests that the OLS estimand is more representative of the ATT when the fraction of untreated units is larger. This again echoes the results in Słoczyński (2022) on the relationship between $\mathbb{P}(D=1)$ and the interpretation of the OLS estimand.

We can also assess the internal validity of β_{OLS} given $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$. For simplicity, assume that $\tau_0(X)$ has a continuous distribution and, without loss of generality, assume that ATE $> \beta_{\text{OLS}}$. Then,

using Theorem 4.2, we obtain

$$\overline{P}(a, W_0; \{\tau_0\}) = \mathbb{P}(\tau_0(X) \leq b^*),$$

where b^* satisfies $\mathbb{E}[\tau_0(X) \mid \tau_0(X) \leq b^*] = \beta_{\text{OLS}}$. The quantity $\overline{P}(a, W_0; \{\tau_0\})$ is largest when the least amount of trimming needs to be applied. This is the case when the trimmed values are largest, or when $\mathbb{E}[\tau_0(X) \mid \tau_0(X) \geq b]$ is large when b is near the maximum of $\text{supp}(\tau_0(X))$. More concretely, if some covariate values in $\text{supp}(X)$ have large CATEs, then considering a subpopulation that removes only a small subset of $\text{supp}(X)$—those corresponding to large CATE values—can allow $\mathbb{E}[\tau_0(X) \mid \tau_0(X) \leq b^*]$ and β_{OLS} to be equal.

5.2 Instrumental Variables

5.2.1 Relevant Results on 2SLS

Consider an endogenous binary treatment $D \in \{0, 1\}$ and a binary instrument $Z \in \{0, 1\}$. Potential treatments, denoted by $(D(1), D(0))$, are linked to the realized treatment through Z, that is, $D = D(Z)$. Potential outcomes, $Y(d, z)$ for $d, z \in \{0, 1\}$, may depend on both D and Z in the absence of an exclusion restriction. Let $Y = Y(D, Z)$ be the realized outcome. As before, let X denote covariates. We make the following assumptions.

Assumption 5.1 (Instrument validity)**.** We have

1. Exogeneity: $(Y(0,0), Y(1,0), Y(0,1), Y(1,1), D(1), D(0)) \perp\!\!\!\perp Z \mid X$;
2. Exclusion: $\mathbb{P}(Y(d, 0) = Y(d, 1)) = 1$ for $d \in \{0, 1\}$;
3. First stage: $e(X) \coloneqq \mathbb{P}(Z = 1 \mid X) \in (0, 1)$ and $\mathbb{P}(D(1) = 1 \mid X) \neq \mathbb{P}(D(0) = 1 \mid X)$ almost surely.

We also make one of the following two nested monotonicity assumptions.

Assumption 5.2 (Strong monotonicity)**.** $\mathbb{P}(D(1) \geq D(0) \mid X) = 1$ almost surely.

Assumption 5.3 (Weak monotonicity)**.** There exists a subset of the support of X such that $\mathbb{P}(D(1) \geq D(0) \mid X) = 1$ on it and $\mathbb{P}(D(1) \leq D(0) \mid X) = 1$ on its complement.

The first instrumental variables estimand we consider was originally studied by Angrist and Imbens (1995). In addition to Assumptions 5.1 and 5.3, suppose that the model for X is saturated, with K possible combinations of covariate values, i.e. let $\text{supp}(X) = \{x_1, \ldots, x_K\}$. Let $X_S = (1, \mathbb{1}(X = x_1), \ldots, \mathbb{1}(X = x_{K-1}))$ and $Z_S = (Z, Z \cdot \mathbb{1}(X = x_1), \ldots, Z \cdot \mathbb{1}(X = x_{K-1})) = ZX_S$, where Z_S is the constructed instrument vector. The estimand in Angrist and Imbens (1995) is the following 2SLS estimand:

$$\beta_{\text{2SLS}} \coloneqq \left[\left(\mathbb{E}\left[W_S' Q_S\right] \left(\mathbb{E}\left[Q_S' Q_S\right]\right)^{-1} \mathbb{E}\left[Q_S' W_S\right] \right)^{-1} \mathbb{E}\left[W_S' Q_S\right] \left(\mathbb{E}\left[Q_S' Q_S\right]\right)^{-1} \mathbb{E}\left[Q_S' Y\right] \right]_1,$$

where $W_S = (D, X_S)$, $Q_S = (Z_S, X_S)$, and $[\cdot]_k$ denotes the kth element of the corresponding vector. This estimand has been studied by Angrist and Imbens (1995), Kolesár (2013), Słoczyński (2020), and Bland-

hol, Bonney, Mogstad, and Torgovitsky (2022), and the specific representation in Proposition 5.1 follows Słoczyński (2020).

Proposition 5.1. Suppose Assumptions 5.1 and 5.3 hold. Suppose X is discrete with finite support. Then

$$\begin{aligned}\beta_{\text{2SLS}} &= \frac{\mathbb{E}\left[e(X)(1-e(X))\cdot\mathbb{P}\left(D(1)\neq D(0)\mid X\right)^2\cdot\mathbb{E}[Y(1)-Y(0)\mid D(1)\neq D(0),X]\right]}{\mathbb{E}[e(X)(1-e(X))\cdot\mathbb{P}(D(1)\neq D(0)\mid X)^2]}\\ &= \frac{\mathbb{E}\left[|\operatorname{cov}(D,Z\mid X)|\cdot\mathbb{E}[Y(1)-Y(0)\mid D(1)\neq D(0),X]\mid D(1)\neq D(0)\right]}{\mathbb{E}[|\operatorname{cov}(D,Z\mid X)|\mid D(1)\neq D(0)]}.\end{aligned}$$

This means that we can write β_{2SLS} as

$$\beta_{\text{2SLS}} = \frac{\mathbb{E}[a(X)w_0(X)\tau_0(X)]}{\mathbb{E}[a(X)w_0(X)]},$$

where $W_0 = \mathbb{1}(D(1)\neq D(0))$ is the population of compliers and defiers, $w_0(X) = \mathbb{P}(D(1)\neq D(0)\mid X)$, $\tau_0(X) = \mathbb{E}[Y(1)-Y(0)\mid D(1)\neq D(0),X]$, and $a(X) = e(X)(1-e(X))\cdot\mathbb{P}(D(1)\neq D(0)\mid X) = |\operatorname{cov}(D,Z\mid X)|$, a nonnegative weight function. Note that $\beta_{\text{2SLS}} = \text{LATE} \coloneqq \mathbb{E}[Y(1)-Y(0)\mid D(1)\neq D(0)]$ if and only if $a(X)$ is uncorrelated with $\tau_0(X)$ given $D(1)\neq D(0)$.

The practical limitation of focusing on β_{2SLS} is that applied researchers rarely create additional instruments by interacting the original instrument with covariates (cf. Blandhol, Bonney, Mogstad, and Torgovitsky, 2022), which is how Z_S is constructed to obtain β_{2SLS} above. A more practically relevant estimand is the "noninteracted" IV estimand,

$$\beta_{\text{IV}} \coloneqq \left[\left(\mathbb{E}\left[Q'W\right]\right)^{-1}\mathbb{E}\left[Q'Y\right]\right]_1,$$

where $Q = (Z,X)$ and $W = (D,X)$. To introduce one of the representations of β_{IV} below, define

$$c(X) \coloneqq \operatorname{sign}\Big(\mathbb{P}\left[D(1)\geq D(0)\mid X\right] - \mathbb{P}\left[D(1)\leq D(0)\mid X\right]\Big),$$

where $\operatorname{sign}(\cdot)$ is the sign function. We also make the following "rich covariates" assumption on the instrument propensity score, which is implied by the saturated specification in Proposition 5.1.

Assumption 5.4 (Rich covariates)**.** $e(X)$ is linear in X.

Under the instrument validity assumption, the rich covariates assumption, and either monotonicity assumption, Słoczyński (2020) obtains the following representations for the "noninteracted" IV estimand.

Proposition 5.2. Suppose Assumptions 5.1, 5.3, and 5.4 hold. Then

$$\begin{aligned}\beta_{\text{IV}} &= \frac{\mathbb{E}[c(X)\cdot e(X)(1-e(X))\cdot\mathbb{P}(D(1)\neq D(0)\mid X)\cdot\mathbb{E}[Y(1)-Y(0)\mid D(1)\neq D(0),X]]}{\mathbb{E}[c(X)\cdot e(X)(1-e(X))\cdot\mathbb{P}(D(1)\neq D(0)\mid X)]}\\ &= \frac{\mathbb{E}[c(X)\cdot\operatorname{var}(Z\mid X)\cdot\mathbb{E}[Y(1)-Y(0)\mid D(1)\neq D(0),X]\mid D(1)\neq D(0)]}{\mathbb{E}[c(X)\cdot\operatorname{var}(Z\mid X)\mid D(1)\neq D(0)]}.\end{aligned}$$

Suppose Assumptions 5.1, 5.2, and 5.4 hold instead. Then

$$\beta_{\text{IV}} = \frac{\mathbb{E}[e(X)(1-e(X))\cdot\mathbb{P}(D(1)>D(0)\mid X)\cdot\mathbb{E}[Y(1)-Y(0)\mid D(1)>D(0),X]]}{\mathbb{E}[e(X)(1-e(X))\cdot\mathbb{P}(D(1)>D(0)\mid X)]}$$

$$= \frac{\mathbb{E}[\text{var}(Z \mid X) \cdot \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0), X] \mid D(1) > D(0)]}{\mathbb{E}[\text{var}(Z \mid X) \mid D(1) > D(0)]}.$$

It follows that we can write β_{IV} as

$$\beta_{\text{IV}} = \frac{\mathbb{E}[a(X) w_0(X) \tau_0(X)]}{\mathbb{E}[a(X) w_0(X)]}$$

under either monotonicity assumption. Under weak monotonicity, $W_0 = \mathbb{1}(D(1) \neq D(0))$, $w_0(X) = \mathbb{P}(D(1) \neq D(0) \mid X)$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D(1) \neq D(0), X]$, and possibly negative weights $a(X) = c(X) \cdot \text{var}(Z \mid X)$. Under strong monotonicity, $W_0 = \mathbb{1}(D(1) > D(0))$, $w_0(X) = \mathbb{P}(D(1) > D(0) \mid X)$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0), X]$, and nonnegative weights $a(X) = \text{var}(Z \mid X) \geq 0$.

5.2.2 Causal Representation and Internal Validity of 2SLS

Consider again the setting of Section 5.2.1. The estimand β_{2SLS} can be characterized as $\mu(a_{\text{2SLS}}, \tau_0)$ for $a_{\text{2SLS}}(X) = |\text{cov}(D, Z \mid X)|$, $W_0 = \mathbb{1}(D(1) \neq D(0))$, $w_0(X) = \mathbb{P}(D(1) \neq D(0) \mid X)$, and $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D(1) \neq D(0), X]$.

Since $a_{\text{2SLS}}(X) \geq 0$, there exists a subpopulation of $\{D(1) \neq D(0)\}$ such that β_{2SLS} is an average treatment effect over that subpopulation. The maximum size of that subpopulation is given by

$$\overline{P}(a_{\text{2SLS}}, W_0; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[a_{\text{2SLS}}(X) \mid W_0 = 1]}{\sup_{x \in \text{supp}(X \mid W_0 = 1)} a_{\text{2SLS}}(x)} = \frac{\mathbb{E}[|\text{cov}(D, Z \mid X)| \mid D(1) \neq D(0)]}{\sup_{x \in \text{supp}(X \mid D(1) \neq D(0))} |\text{cov}(D, Z \mid X = x)|}.$$

The maximum value of $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ is obtained when $|\text{cov}(D, Z \mid X)|$ does not depend on X. This occurs, for example, when the instrument and the fraction of units for which $D(1) \neq D(0)$ are independent of X. In this case, we have that $\beta_{\text{2SLS}} = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, the average effect of treatment in the complier and defier subpopulation.

Under weak monotonicity, the IV estimand has the same W_0 and $w_0(X)$, but has $a_{\text{IV}}(X) = \text{sign}(\mathbb{P}(D(1) \geq D(0) \mid X) - \mathbb{P}(D(1) \leq D(0) \mid X)) \cdot \text{var}(Z \mid X) = c(X) \cdot \text{var}(Z \mid X)$ instead. If $\mathbb{P}(c(X) = -1) > 0$, then weights are negative with positive probability and there does not exist a causal representation for the estimand β_{IV} that is uniform in $\tau_0 \in \mathcal{T}_{\text{all}}$. However, there will exist a causal representation given τ_0 if the support condition of Theorem 3.2 holds, i.e. if β_{IV} lies in the support of $\tau_0(X)$.

If we assume strong monotonicity (Assumption 5.2), then $a_{\text{IV}}(X) = \text{var}(Z \mid X) \geq 0$ and

$$\overline{P}(a_{\text{IV}}, W_0; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[a_{\text{IV}}(X) \mid W_0 = 1]}{\sup_{x \in \text{supp}(X \mid W_0 = 1)} a_{\text{IV}}(x)} = \frac{\mathbb{E}[\text{var}(Z \mid X) \mid D(1) > D(0)]}{\sup_{x \in \text{supp}(X \mid D(1) > D(0))} \text{var}(Z \mid X = x)}.$$

Here the internal validity of the IV estimand is maximized when $\text{var}(Z \mid X)$ is constant, which occurs when Z is independent of X. In this case, β_{IV} equals LATE. The quantities $\overline{P}(a_{\text{IV}}, W_0; \mathcal{T}_{\text{all}})$ and $\overline{P}(a_{\text{2SLS}}, W_0; \mathcal{T}_{\text{all}})$ are not ranked uniformly in the distributions of $(D(1), D(0), X, Z)$ as there are data-generating processes that make each of these two quantities larger than the other. For example, if $\text{var}(Z \mid X)$ is constant but $\mathbb{P}(D(1) \neq D(0) \mid X)$ is not, then $\overline{P}(a_{\text{2SLS}}, W_0; \mathcal{T}_{\text{all}}) < \overline{P}(a_{\text{IV}}, W_0; \mathcal{T}_{\text{all}})$. This scenario is plausible if Z is randomly assigned and X is a vector of pre-assignment characteristics. This inequality is reversed if $a_{\text{2SLS}}(X) = |\text{cov}(D, Z \mid X)|$ is constant but $\mathbb{P}(D(1) \neq D(0) \mid X)$ is not. They are equally representative

when $\mathbb{P}(D(1) \neq D(0) \mid X)$ is constant. In this case, the estimands are equal, so this is not unexpected.

5.3 Difference-in-Differences

5.3.1 Relevant Results on TWFE

Now suppose units are observed for T periods and, for $t \in \{1, \ldots, T\}$, denote binary treatment by $D_t \in \{0, 1\}$, potential outcomes $(Y_t(1), Y_t(0))$, and realized outcome $Y_t = Y_t(D_t)$. We assume units are untreated prior to period $G \in \{2, 3, \ldots, T\} \cup \{+\infty\}$, receive the treatment in period G, and remain treated thereafter. We assume no units are treated in the first time period. This may include a group that remains untreated throughout, for which $G = +\infty$. Thus, $D_t = \mathbb{1}(G \leq t)$. The panel is balanced, that is, no group appears or disappears over time.

The two-way fixed effects estimand is often used in this setting, and consists of regressing the outcome on the treatment indicator, group indicators, and period indicators. Using partitioned regression results, the coefficient on treatment indicator is

$$\beta_{\text{TWFE}} \coloneqq \frac{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\ddot{D}_t Y_t\right]}{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\ddot{D}_t^2\right]},$$

where $\ddot{D}_t = D_t - \frac{1}{T}\sum_{s=1}^{T} D_s - \mathbb{E}[D_t] + \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s]$.

We assume a version of parallel trends most similar to the one in de Chaisemartin and D'Haultfœuille (2020).

Assumption 5.5 (Difference-in-differences)**.** We have

1. $\text{supp}(G) = \{2, 3, \ldots, T\} \cup \{+\infty\}$;
2. For all $t \in \{2, \ldots, T\}$ and $g, g' \in \text{supp}(G)$, we have that $\mathbb{E}[Y_t(0) - Y_{t-1}(0) \mid G = g] = \mathbb{E}[Y_t(0) - Y_{t-1}(0) \mid G = g']$.

We use a proposition that is essentially a special case of Theorem 1 in de Chaisemartin and D'Haultfœuille (2020) to obtain a representation of the two-way fixed effects estimand as a weighted average.

Proposition 5.3. Suppose Assumption 5.5 holds. Then

$$\beta_{\text{TWFE}} = \frac{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big[\big(1 - \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s \mid G] - \mathbb{E}[D_t] + \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s]\big) \cdot \mathbb{P}(D_t = 1 \mid G) \cdot \mathbb{E}[Y_t(1) - Y_t(0) \mid G, D_t = 1]\Big]}{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big[\big(1 - \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s \mid G] - \mathbb{E}[D_t] + \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s]\big) \cdot \mathbb{P}(D_t = 1 \mid G)\Big]}.$$

We show the above representation satisfies equation (1.1) by introducing an auxiliary variable P that is uniformly distributed on $\{1, \ldots, T\}$ independently from $\{(Y_t(0), Y_t(1), G)\}_{t=1}^{T}$. This *period* variable denotes the time period and we use it to define $(Y(1), Y(0), Y, D) \coloneqq (Y_P(1), Y_P(0), Y_P, D_P)$, which are potential outcomes, the realized outcome, and treatment at random period P, respectively.

Letting $X = (G, P)$, this means we can write β_{TWFE} as

$$\beta_{\text{TWFE}} = \frac{\mathbb{E}[a(X)w_0(X)\tau_0(X)]}{\mathbb{E}[a(X)w_0(X)]},$$

where $W_0 = D$, $w_0(X) = \mathbb{P}(D = 1 \mid G, P) = D \in \{0, 1\}$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D = 1, G, P]$, and the weight function $a_{\text{TWFE}}(X) = 1 - \mathbb{P}(D = 1 \mid G) - \mathbb{P}(D = 1 \mid P) + \mathbb{P}(D = 1)$ is not generally nonnegative.[4] A nonnegative weight function can be obtained under the assumption that group-level average treatment effects are constant over time. This property was described in de Chaisemartin and D'Haultfœuille (2020, Appendix 3.1) and Goodman-Bacon (2021, Section 3.1.1), and the resulting representations of the two-way fixed effects estimand are given in their Theorem S2 and equation (16), respectively. The following proposition yields a simple expression for the weights in our setting.

Proposition 5.4. Suppose Assumption 5.5 holds and that $\mathbb{E}[Y_t(1) - Y_t(0) \mid D = 1, G] = \mathbb{E}[Y_s(1) - Y_s(0) \mid D = 1, G]$ for any $s, t \in \{1, \ldots, T\}$. Then

$$\beta_{\text{TWFE}} = \frac{\mathbb{E}[a_{\text{TWFE,H}}(G) \cdot \mathbb{P}(D = 1 \mid G) \cdot \mathbb{E}[Y(1) - Y(0) \mid D = 1, G]]}{\mathbb{E}[a_{\text{TWFE,H}}(G) \cdot \mathbb{P}(D = 1 \mid G)]},$$

where $a_{\text{TWFE,H}}(g) = \mathbb{P}(D = 0 \mid G = g) \cdot (\mathbb{P}(D = 0 \mid P \geq g) + \mathbb{P}(D = 1 \mid P < g)) \geq 0$ for $g \in \{2, \ldots, T\}$.

As is the case of the representation in Proposition 5.3, the two-way fixed effects estimand in Proposition 5.4 satisfies the representation in (1.1), with $X = G$, $W_0 = D$, $w_0(X) = \mathbb{P}(D = 1 \mid G)$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D = 1, G]$, and the weight function $a(X) = a_{\text{TWFE,H}}(G) \geq 0$. This weight function is derived in the proof of the proposition, and we show it is equivalent to equation (16) in Goodman-Bacon (2021) in Appendix G.

5.3.2 Causal Representation and Internal Validity of TWFE

We now consider the weights obtained in Proposition 5.4 under its assumptions. These weights are nonnegative and therefore Theorem 3.1 guarantees the existence of a causal representation for β_{TWFE} uniformly in $\tau_0 \in \mathcal{T}_{\text{all}}$. Using Theorem 4.1, the internal validity of β_{TWFE} relative to target parameter $\mathbb{E}[Y(1) - Y(0) \mid D = 1]$ is given by

$$\begin{aligned}\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}}) &= \frac{\mathbb{E}[a_{\text{TWFE,H}}(G)w_0(G)]}{\mathbb{E}[w_0(G)] \cdot \sup_{g \in \text{supp}(G \mid D=1)} a_{\text{TWFE,H}}(g)} \\ &= \frac{\sum_{g=2}^{T} \text{var}(D \mid G = g) \cdot (\mathbb{P}(D = 0 \mid P \geq g) + \mathbb{P}(D = 1 \mid P < g)) \cdot \mathbb{P}(G = g)}{\mathbb{P}(D = 1) \cdot \sup_{g \in \{2, \ldots, T\}} \mathbb{P}(D = 0 \mid G = g) \cdot (\mathbb{P}(D = 0 \mid P \geq g) + \mathbb{P}(D = 1 \mid P < g))}.\end{aligned}$$

Due to the absorbing nature of the treatment in our setting, all expressions involving the distribution of D given P or G can be derived as a function of the marginal distribution of G. Therefore, $\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}})$ depends only on $\{\mathbb{P}(G = g)\}_{g \in \{2, \ldots, T\}}$.

To give some intuition, consider the case where $T = 3$ and therefore $G \in \{2, 3, +\infty\}$. In this case,

[4] Here, $\tau_0(X)$ is what Callaway and Sant'Anna (2021) call "the group-time average treatment effect."

$\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}})$ will simply be a function of $(\mathbb{P}(G=2), \mathbb{P}(G=3))$. Calculations yield that

$$\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}}) = \begin{cases} \frac{2\mathbb{P}(G=2)+\omega\mathbb{P}(G=3)}{2\mathbb{P}(G=2)+\mathbb{P}(G=3)} & \text{if } \omega < 1 \\ \frac{\omega^{-1}2\mathbb{P}(G=2)+\mathbb{P}(G=3)}{2\mathbb{P}(G=2)+\mathbb{P}(G=3)} & \text{if } \omega > 1 \\ 1 & \text{if } \omega = 1, \end{cases}$$

where

$$\omega = \frac{4 - 2\mathbb{P}(G=2) - 4\mathbb{P}(G=3)}{2 - 2\mathbb{P}(G=2) - \mathbb{P}(G=3)} = \frac{a_{\text{TWFE,H}}(2)}{a_{\text{TWFE,H}}(3)}.$$

Therefore, the TWFE estimand is perfectly representative of the ATT if and only if $\omega = 1$.

With some algebra, we can see that $\omega < 1$ if and only if $\mathbb{P}(G=3) > 2/3$, while $\omega > 1$ if and only if $\mathbb{P}(G=3) < 2/3$. We can also see that $\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}})$ is equal to 1 when $\mathbb{P}(G=3) = 2/3$, and that the internal validity of the TWFE estimand declines as $|\mathbb{P}(G=3) - 2/3|$ increases. This is due to weight function $a_{\text{TWFE,H}}(g)$ being constant in g if and only if the fraction of units treated in the third period is $2/3$. As before, constant weights imply that the weighted estimand equals the average treatment effect over $\{W_0 = 1\}$, which corresponds to the treated subpopulation here.

6 Estimation and Inference

We now consider the estimation and inference for our measures of internal validity and representativeness.

We will focus our attention on the case when $\mathcal{T} = \mathcal{T}_{\text{all}}$. Estimation and inference for $\overline{P}(a, W_0; \{\tau_0\})$ is related to the question of estimation and inference in linear programs with estimated constraints. See Andrews, Roth, and Pakes (2023), Cox and Shi (2023), Fang, Santos, Shaikh, and Torgovitsky (2023), and Cho and Russell (2024) for recent advances on this topic.

To measure internal validity, we seek to estimate

$$\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = \frac{\mathbb{E}[a(X) \mid W_0 = 1]}{a_{\max}} = \frac{\mathbb{E}[a(X) w_0(X)]}{\mathbb{E}[w_0(X)] \cdot a_{\max}}. \tag{6.1}$$

Suppose we observe a random sample of size n, $\{(W_i, X_i)\}_{i=1}^n$, where W_i are a set of variables that allow the estimation of $a(\cdot)$ and $w_0(\cdot)$. For example, under unconfoundedness we can let $W_i = D_i$ since the distribution of (D, X) is sufficient to identify $a(\cdot)$; the outcome's distribution does not affect $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$. In our instrumental variables examples, we let $W_i = (D_i, Z_i)$.

Assuming the existence of estimators for $a(\cdot)$ and $w_0(\cdot)$, we can propose the following analog estimator for $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$. Start by noting that we can estimate $\mathbb{E}[a(X) \mid W_0 = 1]$ via

$$\frac{\frac{1}{n}\sum_{i=1}^n \widehat{a}(X_i)\widehat{w}_0(X_i)}{\frac{1}{n}\sum_{i=1}^n \widehat{w}_0(X_i)},$$

which will be consistent under standard conditions on $\widehat{a}$ and $\widehat{w}_0$. However, estimating $a_{\max}$, the essential supremum of $a(X)$ given $W_0 = 1$, is more delicate. In some of our examples, this supremum is known or can

be bounded above without needing the use of data. For example, the OLS estimand under unconfoundedness has weight function $a(X) = p(X)(1-p(X))$ which is naturally bounded above by $1/4$. If X is continuously distributed, it is possible that $1/4$ lies in the support of $a(X)$, and thus we may side-step the estimation of this term. The IV estimand of Section 5.2.1 has weights $a_{\text{IV}}(X) = \text{var}(Z \mid X)$ which are also naturally bounded above by $1/4$. Similarly, $a_{\text{2SLS}}(X) = |\operatorname{cov}(D, Z \mid X)| \leq \sqrt{\text{var}(D \mid X)\operatorname{var}(Z \mid X)} \leq 1/4$ using the Cauchy–Schwarz inequality. If knowledge of $a_{\max}$ is not assumed, but $\text{supp}(X \mid W_0 = 1)$ is known and $a(x)$ is continuous[5], then one could use $\sup_{x \in \text{supp}(X \mid W_0 = 1)} \widehat{a}(x)$ as an estimator for $a_{\max}$. This estimator will be consistent when $\widehat{a}(x)$ is consistent for $a(x)$ uniformly in $x \in \text{supp}(X \mid W_0 = 1)$. Many parametric and nonparametric estimators for $a(\cdot)$ can be shown to satisfy this requirement.

In the case when $\text{supp}(X \mid W_0 = 1)$ is not known a priori, one can also estimate it. We focus the rest of this section on one particularly common case, where X is discretely distributed. In this case, $a(x)$ and $w_0(x)$ can usually be estimated "cell-by-cell" and be $\sqrt{n}$-consistent with a limiting Gaussian distribution. In this case, we will see that inference on $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$ is generally nonstandard and, as a result, most common bootstrap procedures fail. When X has a continuous component, there are many different estimation approaches and the limiting distribution of estimators of $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$ will vary with the type of estimators (parametric/nonparametric), rate of convergence, type of knowledge assumed for $\text{supp}(X \mid W_0 = 1)$, etc.

We consider the following simple plug-in estimator which does not require knowledge of $\text{supp}(X \mid W_0 = 1)$:

$$\widehat{\overline{P}} = \frac{\frac{1}{n}\sum_{i=1}^n \widehat{a}(X_i)\widehat{w}_0(X_i)}{\frac{1}{n}\sum_{i=1}^n \widehat{w}_0(X_i) \cdot \max_{i:\widehat{w}_0(X_i) > c_n} \widehat{a}(X_i)},$$

where c_n is a tuning parameter that converges to 0 as $n \to \infty$. Note that this tuning parameter is absent when w_0 is known, such as under unconfoundedness: see Section 2. This estimator does not assume knowledge of the support of X given $W_0 = 1$, but it can also be implemented by taking the maximum over $\text{supp}(X \mid W_0 = 1)$ when it is known.

Let $\text{supp}(X) = \{x_1, \ldots, x_K\}$ and denote by $p_j = \mathbb{P}(X = x_j)$ the frequency of cell j and let $\widehat{p}_j = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i = x_j)$ denote its sample frequency. Let $\widehat{\theta} = (\widehat{\mathbf{a}}, \widehat{\mathbf{w}}_0, \widehat{\mathbf{p}})$ where $\widehat{\mathbf{a}} = (\widehat{a}(x_1), \ldots, \widehat{a}(x_K))$, $\widehat{\mathbf{w}}_0 = (\widehat{w}_0(x_1), \ldots, \widehat{w}_0(x_K))$, and $\widehat{\mathbf{p}} = (\widehat{p}_1, \ldots, \widehat{p}_K)$. Let $\theta = (\mathbf{a}, \mathbf{w}_0, \mathbf{p})$ denote their population counterparts.

We make the following assumptions on the behavior of the preliminary estimators.

Assumption 6.1 (Preliminary estimators)**.** Let

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathbb{Z}$$

where $\mathbb{Z} \coloneqq (\mathbb{Z}_{\mathbf{a}}, \mathbb{Z}_{\mathbf{w}_0}, \mathbb{Z}_X) \in \mathbb{R}^{3K}$ has a Gaussian distribution.

The above assumption is often satisfied when X has finite support since estimators for $a(x_j)$ and $w_0(x_j)$ can be obtained using only the observations for which $X_i = x_j$. Note that the limiting distribution of $\mathbb{Z}_{\mathbf{w}_0}$ may be degenerate. For example, if $W_0 = 1$ a.s., then $\widehat{w}_0(x) = w_0(x) = 1$ for all $x \in \text{supp}(X)$ and thus $\mathbb{Z}_{\mathbf{w}_0} = \mathbf{0}_K$ almost surely, where $\mathbf{0}_K$ is a K-vector of zeros.

The next theorem shows the consistency of $\widehat{\overline{P}}$ and establishes the limiting distribution of this estimator. To simplify the exposition, we use $\overline{P}$ to denote $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$ in what follows.

[5]Note that $a(x)$ is trivially continuous on finite support.

Theorem 6.1 (Consistency and asymptotic distribution)**.** Suppose Assumption 6.1 holds. Suppose $c_n = o(1)$ and $c_n\sqrt{n} \to \infty$ as $n \to \infty$. Suppose $\overline{P} \neq 0$. Then, $\widehat{\overline{P}}$ is consistent for $\overline{P}$ and

$$\sqrt{n}(\widehat{\overline{P}} - \overline{P}) \xrightarrow{d} \psi(\mathbb{Z}),$$

where ψ is a continuous mapping defined in equation (6.2).

We now characterize the mapping ψ. Let $\mathbb{Z}_{\mathbf{a}}(j)$ denote the jth element of $\mathbb{Z}_{\mathbf{a}}$ and similarly define $\mathbb{Z}_{\mathbf{w}_0}(j)$ and $\mathbb{Z}_X(j)$ for $j \in \{1, \ldots, K\}$. Let

$$\begin{aligned}\psi(\mathbb{Z}) \;=\; & \sum_{j=1}^{K} \frac{w_0(x_j)p_j}{\mathbb{P}(W_0 = 1)a_{\max}} \mathbb{Z}_{\mathbf{a}}(j) \;-\; \frac{\mathbb{E}[a(X) \mid W_0 = 1]}{a_{\max}^2} \max_{j \in \Psi_{\mathcal{X}^+}} \mathbb{Z}_{\mathbf{a}}(j) \\ & + \sum_{j=1}^{K} \frac{(a(x_j) - \mathbb{E}[a(X) \mid W_0 = 1])p_j}{\mathbb{P}(W_0 = 1)a_{\max}} \mathbb{Z}_{\mathbf{w}_0}(j) \;+\; \sum_{j=1}^{K} \frac{(a(x_j) - \mathbb{E}[a(X) \mid W_0 = 1])w_0(x_j)}{\mathbb{P}(W_0 = 1)a_{\max}} \mathbb{Z}_X(j), \quad (6.2)\end{aligned}$$

where $\Psi_{\mathcal{X}^+} = \{j \in \{1, \ldots, K\} : a(x_j) = a_{\max}\}$.

The mapping ψ is linear if and only if $a(x)$ is maximized at a unique value $x \in \text{supp}(X \mid W_0 = 1)$, and nonlinear if multiple values maximize $a(x)$. The linearity of this mapping crucially affects the choice of inference procedure. When ψ is linear, the limiting distribution of $\widehat{\overline{P}}$ is Gaussian and common bootstrap procedures, such as the nonparametric bootstrap, are valid whenever they are valid for $\widehat{\theta}$.

However, when $a(x)$ is maximized at more than one value, the limiting distribution of $\widehat{\overline{P}}$ is nonlinear in $\mathbb{Z}$ and non-Gaussian because it depends on the maximum of two Gaussian variables. In these cases, it can be shown (see Theorem 3.1 in Fang and Santos (2019)) that standard bootstrap approaches are invalid. However, using the fact that the estimand $\overline{P}$ can be written as a Hadamard directionally differentiable mapping of θ implies that alternative bootstrap procedures, such as those proposed by Hong and Li (2018) and Fang and Santos (2019), can be applied to obtain valid inferences on $\overline{P}$.

We propose a bootstrap procedure that can be applied regardless of the linearity of ψ. In order to show its validity, we assume that the limiting distribution $\mathbb{Z}$ can be approximated via a bootstrap procedure.

Assumption 6.2 (Bootstrap for first-step estimators)**.** Let $\mathbb{Z}^* \coloneqq (\mathbb{Z}^*_{\mathbf{a}}, \mathbb{Z}^*_{\mathbf{w}_0}, \mathbb{Z}^*_X) \in \mathbb{R}^{3K}$ be a random vector such that $\mathbb{Z}^* \overset{p}{\rightsquigarrow} \mathbb{Z}$, where $\overset{p}{\rightsquigarrow}$ denotes convergence in probability conditioning on the data used to compute $\widehat{\theta}$.

This assumption is easily satisfied when $\widehat{\mathbf{p}}$ are sample frequencies, $(\widehat{\mathbf{a}}, \widehat{\mathbf{w}}_0)$ are cell-by-cell estimators that are asymptotically linear and Gaussian, and when $\mathbb{Z}^*$ is the distribution of these estimators under a standard bootstrap approach. For example, under the nonparametric bootstrap we can let $\mathbb{Z}^*_X(j) = \sqrt{n}(\widehat{p}^*_j - \widehat{p}_j)$ where $\widehat{p}^*_j = \frac{1}{n}\sum_{i=1}^{N} \mathbb{1}(X^*_i = x_j)$, where $(X^*_1, \ldots, X^*_n)$ are drawn from the empirical distribution of $(X_1, \ldots, X_n)$.

Theorem 6.2 (Bootstrap validity)**.** Suppose the assumptions of Theorem 6.1 hold and that Assumption 6.2 holds. Then,

$$\widehat{\psi}(\mathbb{Z}^*) \overset{p}{\rightsquigarrow} \psi(\mathbb{Z})$$

as $n \to \infty$, where $\widehat{\psi}$ is defined in equation (6.3).

We now characterize the mapping $\widehat{\psi}$. Let $\mathbb{Z}^*_{\mathbf{a}}(j)$ denote the jth element of $\mathbb{Z}^*_{\mathbf{a}}$ and similarly define $\mathbb{Z}^*_{\mathbf{w}_0}(j)$. Let

$$\begin{aligned}\widehat{\psi}(\mathbb{Z}^*) \;=\; & \sum_{j=1}^{K} \frac{\widehat{w}_0(x_j)\widehat{p}_j}{\widehat{\mathbb{P}}(W_0=1)\widehat{a}_{\max}}\mathbb{Z}^*_{\mathbf{a}}(j) \;-\; \frac{\widehat{\mathbb{E}}[a(X)\mid W_0=1]}{\widehat{a}^2_{\max}} \max_{j\in\widehat{\Psi}_{\mathcal{X}^+}} \mathbb{Z}^*_{\mathbf{a}}(j) \\ & + \sum_{j=1}^{K} \frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])\widehat{p}_j}{\widehat{\mathbb{P}}(W_0=1)\widehat{a}_{\max}}\mathbb{Z}^*_{\mathbf{w}_0}(j) \;+\; \sum_{j=1}^{K} \frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])\widehat{w}_0(x_j)}{\widehat{\mathbb{P}}(W_0=1)\widehat{a}_{\max}}\mathbb{Z}^*_{X}(j), \quad (6.3)\end{aligned}$$

where

$$\widehat{\mathbb{E}}[a(X)\mid W_0=1] = \frac{\frac{1}{n}\sum_{i=1}^n \widehat{a}(X_i)\widehat{w}_0(X_i)}{\frac{1}{n}\sum_{i=1}^n \widehat{w}_0(X_i)}, \qquad \widehat{\mathbb{P}}(W_0=1) = \frac{1}{n}\sum_{i=1}^n \widehat{w}_0(X_i), \qquad \widehat{a}_{\max} = \max_{i:\widehat{w}_0(X_i)>c_n} \widehat{a}(X_i).$$

The set $\widehat{\Psi}_{\mathcal{X}^+}$ is defined as all elements $j \in \{1,\dots,K\}$ for which $\widehat{a}(x_j)$ is within ξ_n of the maximal value, where ξ_n is a positive sequence satisfying $\xi_n = o(1)$ and $\xi_n\sqrt{n} \to \infty$ as $n \to \infty$.

Formally,

$$\widehat{\Psi}_{\mathcal{X}^+} = \left\{ k \in \{1,\dots,K\} : \widehat{a}(x_k) \geq \max_{i:\widehat{w}_0(X_i)>c_n} \widehat{a}(X_i) - \xi_n \right\}.$$

The proof of Theorem 6.2 shows that this set consistently estimates $\Psi_{\mathcal{X}^+}$ and thus satisfies the conditions of Theorem 3.2 in Fang and Santos (2019).

The bootstrap procedure is valid whether the limiting distribution is Gaussian or not. If we assume $a(x)$ is maximized at a single value, standard bootstrap procedures can also be used to approximate the limiting distribution of $\widehat{\overline{P}}$.

We propose the following bootstrap procedure to compute a one-sided $(1-\alpha)$ confidence interval for $\overline{P}$.

Algorithm 6.1 (One-sided confidence interval for $\overline{P}$)**.** We compute the confidence interval in three steps:

1. Compute $\widehat{\theta}$ and $\widehat{\overline{P}}$ using the random sample $\{(W_i, X_i)\}_{i=1}^n$;
2. For bootstrap samples $b = 1,\dots,B$, compute $\widehat{\theta}^{*,b} = (\widehat{\mathbf{a}}^{*,b}, \widehat{\mathbf{w}}_0^{*,b}, \widehat{\mathbf{p}}^{*,b})$ and $\mathbb{Z}^{*,b} = \sqrt{n}(\widehat{\theta}^{*,b} - \widehat{\theta})$;
3. Compute $\widehat{q}_\alpha$, the α quantile of $\widehat{\psi}(\mathbb{Z}^{*,b})$, and report the interval $\left[0, \widehat{\overline{P}} - \widehat{q}_\alpha/\sqrt{n}\right]$.

We can also view these inferential problems through the lens of intersection or union bounds. For example, we can write

$$\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = \inf_{x\in\text{supp}(X|W_0=1)} \frac{\mathbb{E}[a(X)w_0(X)]}{\mathbb{E}[w_0(X)]\cdot a(x)} \coloneqq \inf_{x\in\text{supp}(X|W_0=1)} \overline{P}(x).$$

Computing a one-sided confidence interval for $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$ of the kind $[0, \widehat{\overline{P}}^+]$ can be cast as doing inference on intersection bounds. Chernozhukov, Lee, and Rosen (2013) offer methods for such problems. Equivalently, the computation of a one-sided confidence interval $[\widehat{\overline{P}}^-, 1]$ is related to inferential questions in union bounds: see Bei (2024). We leave all details for future work.

7 Empirical Application

In this section, we implement the proposed tools in an application to the effects of unilateral divorce laws in the U.S. on female suicide, as in Stevenson and Wolfers (2006). Between 1969 and 1985, 37 states (including the District of Columbia) reformed their law by enabling each spouse to seek divorce without the other spouse's consent. Stevenson and Wolfers (2006) argue that these "unilateral" or "no-fault" divorce laws reduced female suicide, domestic violence, and spousal homicide. The results on female suicide are also replicated by Goodman-Bacon (2021), whose analysis we follow here.

Our sample consists of 41 states observed over the 1964–1996 period. The outcome of interest is the state- and year-specific female suicide rate (per million women), as computed by the National Center for Health Statistics (NCHS). The treatment is whether the state allowed unilateral divorce in a given year. Following Goodman-Bacon (2021), our sample omits Alaska and Hawaii. Additionally, we omit Louisiana, Maryland, North Carolina, Oklahoma, Utah, Vermont, Virginia, and West Virginia. These eight states (and Alaska) had unilateral divorce laws preceding 1964 and are therefore always treated within our timeframe.

Panel A of Table 1 reports our baseline estimates of the average effects of unilateral divorce laws on female suicide. After we drop the eight always-treated states, the TWFE estimate, –0.604, becomes much smaller in absolute value than the corresponding estimate in Goodman-Bacon (2021), –3.080. Unlike that estimate, ours is also statistically insignificant, with p-value = 0.819.

The conclusion changes, however, when we explicitly target the average treatment effect on the treated (ATT), that is, the average effect for the largest subpopulation for which such an effect is identified under

Table 1: Internal Validity of the TWFE Estimand of the Effects of Unilateral Divorce Laws

A. Estimates of the effects of unilateral divorce laws

	ATT	
TWFE	Callaway and Sant'Anna	Wooldridge
–0.604	–10.220	–5.530
(2.622)	(3.086)	(3.650)

B. Internal validity of the TWFE estimand based on Proposition 5.3

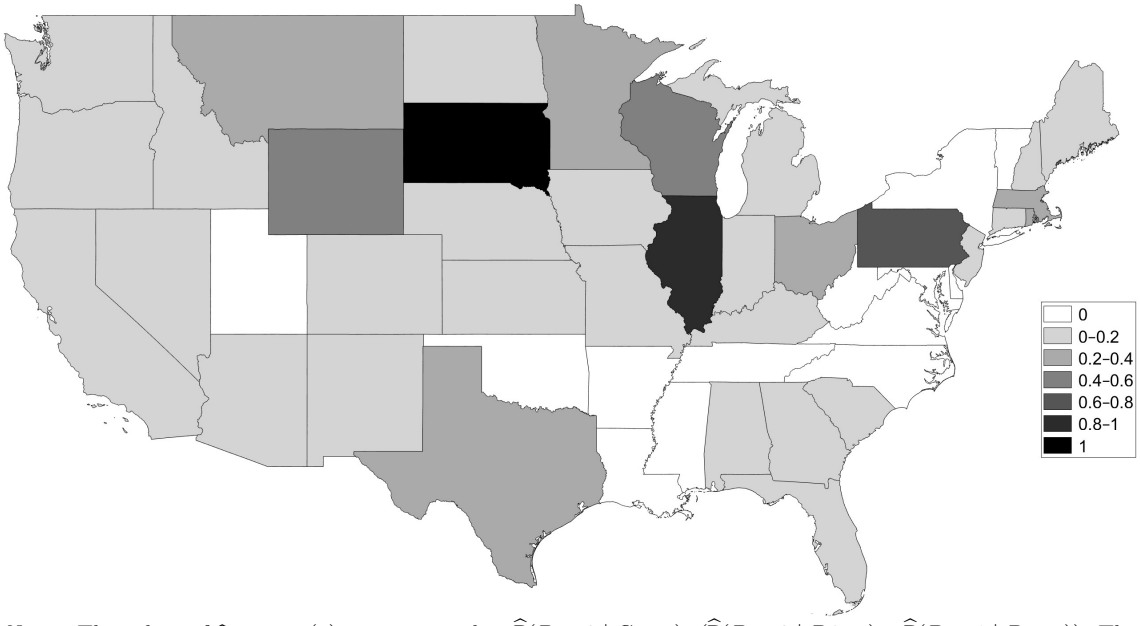
	uniformly in τ_0	given τ_0
$\widehat{\mathbb{P}}(W^* = 1)$	0	0.3873
$\widehat{\mathbb{P}}(W^* = 1 \mid D = 1)$	0	0.6216

C. Internal validity of the TWFE estimand based on Proposition 5.4

	uniformly in τ_0	given τ_0
$\widehat{\mathbb{P}}(W^* = 1)$	0.1400	0.4802
$\widehat{\mathbb{P}}(W^* = 1 \mid D = 1)$	0.2246	0.7707

Notes: The dataset is Goodman-Bacon (2021)'s panel of the 1964–1996 U.S. The outcome is the state- and year-specific female suicide rate (per million women), as computed by the National Center for Health Statistics (NCHS). The treatment is whether the state allowed unilateral divorce in a given year. The sample includes the District of Columbia but excludes Alaska and Hawaii, as in Goodman-Bacon (2021), as well as Louisiana, Maryland, North Carolina, Oklahoma, Utah, Vermont, Virginia, and West Virginia, which had unilateral divorce laws preceding 1964. The measures of internal validity "uniformly in τ_0" are based on Theorem 4.1, while those "given τ_0" are based on Theorem 4.2.

Figure 3: The Distribution of $\widehat{a}_{\text{TWFE,H}}(g)/(\sup_{g\in\text{supp}(G)}\widehat{a}_{\text{TWFE,H}}(g))$ Based on Proposition 5.4



Notes: The values of $\widehat{a}_{\text{TWFE,H}}(g)$ are computed as $\widehat{\mathbb{P}}(D=0 \mid G=g)\cdot(\widehat{\mathbb{P}}(D=0 \mid P\geq g)+\widehat{\mathbb{P}}(D=1 \mid P<g))$. The largest value is obtained for South Dakota. The "never-treated" states (Arkansas, Delaware, Mississippi, New York, and Tennessee) have an imputed value of 0 for $\widehat{a}_{\text{TWFE,H}}(g)$ since they are not part of the treated subpopulation and do not contribute to the TWFE estimand. The "always-treated" states (Louisiana, Maryland, North Carolina, Oklahoma, Utah, Vermont, Virginia, and West Virginia) have values of $\widehat{a}_{\text{TWFE,H}}(g)$ displayed as 0 since these states are dropped from our sample.

standard assumptions. Using the approach of Wooldridge (2021), we obtain an estimate of –5.530 with a p-value of 0.138. The approach of Callaway and Sant'Anna (2021) produces an estimate of –10.220 and a p-value of 0.001. These estimates are more strongly suggestive of a causal effect of unilateral divorce laws than the TWFE estimate.

While the TWFE estimate and the two estimates of the ATT are quite different, this paper focuses on another implication of the nonuniformity of the TWFE weight function. We ask: How representative of the underlying population is the TWFE estimand? What is the internal validity of this estimand if we are interested in the treated subpopulation? Panel B of Table 1 reports our estimates of $\mathbb{P}(W^*=1)$ and $\mathbb{P}(W^*=1 \mid D=1)=\mathbb{P}(W^*=1 \mid W_0=1)$, based on the representation of the TWFE estimand in de Chaisemartin and D'Haultfœuille (2020), revisited in our Proposition 5.3. First, because the weights on some of the group-time average treatment effects are negative, the TWFE estimand does not have a causal interpretation uniformly in τ_0, $\widehat{\mathbb{P}}(W^*=1)=\widehat{\mathbb{P}}(W^*=1 \mid D=1)=0$. Second, when we estimate the CATE function and use these estimates in constructing the bounds, we conclude that the TWFE estimand corresponds to the average treatment effect for at most 62.16% of the treated units or 38.73% of the entire population.

Panel C of Table 1 revisits these questions on the basis of the representation of the TWFE estimand in Proposition 5.4. Here, we assume that group-level average treatment effects are constant over time, which eliminates the problem of negative weights. Indeed, we now conclude that the TWFE estimand has a causal interpretation uniformly in τ_0, even if it is still not particularly representative of the underlying population

or the treated subpopulation. Our estimates of $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid D = 1)$ are equal to 14.00% and 22.46%, respectively. When we use the estimated CATE function in constructing the bounds, these estimates increase to 48.02% and 77.07%. This is obviously much more than our initial estimate of 0, but still substantially less than 1, guaranteed in the case of $\mathbb{P}(W^* = 1 \mid D = 1)$ when using the estimation methods in Callaway and Sant'Anna (2021), Wooldridge (2021), and other recent papers, each of which explicitly targets the ATT.

Figure 3 provides another illustration of the weight function in Proposition 5.4 and the corresponding measures of internal validity. Here, each state across the contiguous U.S. is associated with its value of $\widehat{a}_{\text{TWFE,H}}(g)/(\sup_{g \in \text{supp}(G)} \widehat{a}_{\text{TWFE,H}}(g))$. This value is maximized at 1 for South Dakota. Because $\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}}) = \mathbb{E}[a_{\text{TWFE,H}}(G) \mid D = 1]/(\sup_{g \in \text{supp}(G)} a_{\text{TWFE,H}}(g))$, our estimate of this parameter, reported as 22.46% in Table 1, can be obtained as a weighted mean of the nonzero values in Figure 3 with weights equal to each state's length of exposure to the treatment.

8 Conclusion

In this paper, we studied the representativeness and internal validity of a class of weighted estimands, which includes the popular OLS, 2SLS, and TWFE estimands in additive linear models. We examined the conditions under which such estimands can be written as the average treatment effect for some (possibly latent) subpopulation. In our main result, we derived the sharp upper bound on the size of that subpopulation. We consider this bound to be a valuable diagnostic for empirical research. When a given estimand can be shown to correspond to the average treatment effect for a large subset of the population of interest, we say its internal validity is high. In an application to the effects of unilateral divorce laws in the U.S. on female suicide, as in Stevenson and Wolfers (2006) and Goodman-Bacon (2021), we showed that the TWFE estimand has a low degree of internal validity (assuming that the treated subpopulation is of interest), even when we assume away the existence of negative weights. Because this result is then necessarily driven by the nonuniformity of the TWFE weight function, it corroborates the negative view of both negative and nonuniform weights in Callaway, Goodman-Bacon, and Sant'Anna (2024).

References

Andrews, I., J. Roth, and A. Pakes (2023): "Inference for Linear Conditional Moment Inequalities," *Review of Economic Studies*, 90(6), 2763–2791.

Angrist, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66(2), 249–288.

Angrist, J. D., and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430), 431–442.

Aronow, P. M., and C. Samii (2016): "Does Regression Produce Representative Estimates of Causal Effects?," *American Journal of Political Science*, 60(1), 250–267.

Athey, S., and G. W. Imbens (2022): "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 226(1), 62–79.

Bei, X. (2024): "Inference on Union Bounds with Applications to DiD, RDD, Bunching, and Structural Counterfactuals," working paper, Duke University.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When Is TSLS Actually LATE?," NBER Working Paper No. 29709.

Borusyak, K., X. Jaravel, and J. Spiess (2024): "Revisiting Event-Study Designs: Robust and Efficient Estimation," *Review of Economic Studies*, 91(6), 3253–3285.

Caetano, C., and B. Callaway (2023): "Difference-in-Differences with Time-Varying Covariates in the Parallel Trends Assumption," arXiv preprint arXiv:2202.02903.

Callaway, B., A. Goodman-Bacon, and P. H. C. Sant'Anna (2024): "Difference-in-Differences with a Continuous Treatment," NBER Working Paper No. 32117.

Callaway, B., and P. H. C. Sant'Anna (2021): "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 225(2), 200–230.

Chen, J. (2024): "Potential Weights and Implicit Causal Designs in Linear Regression," arXiv preprint arXiv:2407.21119.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

Cho, J., and T. M. Russell (2024): "Simple Inference on Functionals of Set-Identified Parameters Defined by Linear Moments," *Journal of Business & Economic Statistics*, 42(2), 563–578.

Cox, G., and X. Shi (2023): "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," *Review of Economic Studies*, 90(1), 201–228.

de Chaisemartin, C. (2012): "All You Need Is LATE," working paper, CREST and Paris School of Economics.

——— (2017): "Tolerating Defiance? Local Average Treatment Effects without Monotonicity," *Quantitative Economics*, 8(2), 367–396.

de Chaisemartin, C., and X. D'Haultfœuille (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110(9), 2964–96.

Fang, Z., and A. Santos (2019): "Inference on Directionally Differentiable Functions," *Review of Economic Studies*, 86(1), 377–412.

Fang, Z., A. Santos, A. M. Shaikh, and A. Torgovitsky (2023): "Inference for Large-Scale Linear Systems with Known Coefficients," *Econometrica*, 91(1), 299–327.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2024): "Contamination Bias in Linear Regressions," *American Economic Review*, 114(12), 4015–4051.

Goodman-Bacon, A. (2021): "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 225(2), 254–277.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Hong, H., and J. Li (2018): "The Numerical Delta Method," *Journal of Econometrics*, 206(2), 379–394.

Humphreys, M. (2009): "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities," working paper, Columbia University.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Imbens, G. W., and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York.

Kolesár, M. (2013): "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity," working paper, Princeton University.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018): "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113(521), 390–400.

Miller, D. L., N. Shenhav, and M. Grosz (2023): "Selection into Identification in Fixed Effects Models, with Application to Head Start," *Journal of Human Resources*, 58(5), 1523–1566.

Mogstad, M., and A. Torgovitsky (2024): "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects," in *Handbook of Labor Economics, Vol. 5*, ed. by C. Dustmann, and T. Lemieux, pp. 1–114. Elsevier, Amsterdam.

Słoczyński, T. (2020): "When Should We (Not) Interpret Linear IV Estimands as LATE?," arXiv preprint arXiv:2011.06695.

——— (2022): "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *Review of Economics and Statistics*, 104(3), 501–509.

Stevenson, B., and J. Wolfers (2006): "Bargaining in the Shadow of the Law: Divorce Laws and Family Distress," *Quarterly Journal of Economics*, 121(1), 267–288.

Sun, L., and S. Abraham (2021): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 225(2), 175–199.

van 't Hoff, N., A. Lewbel, and G. Mellace (2024): "Limited Monotonicity and the Combined Compliers LATE," working paper, University of Southern Denmark.

Wooldridge, J. M. (2021): "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," working paper, Michigan State University.

Appendix

This appendix is organized as follows: Appendix A establishes the connection between weakly causal estimands and uniform causal representations in $\mathcal{T}_{\text{all}}$, Appendix B contains proofs for Section 3, Appendix C contains proofs for Section 4, Appendix D contains proofs for Section 5, Appendix E contains proofs for Section 6, Appendix F contains proofs for Appendix A, and Appendix G contains additional results and derivations regarding difference-in-differences and associated weighted estimands.

A Connection Between Weakly Causal Estimands and Uniform Causal Representations in $\mathcal{T}_{\text{all}}$

We now establish equivalence between *weakly causal* estimands as defined in Blandhol, Bonney, Mogstad, and Torgovitsky (2022) (henceforth, BBMT) and estimands that have uniform causal representations as in Theorem 3.1. As in BBMT, consider the case where X has finite support and, as in this paper, assume the treatment is binary. We also abstract from choice groups denoted by G in BBMT.

Since X has finite support, let $\text{supp}(X \mid W_0 = 1) = \{x_1, \ldots, x_K\}$ and let $\tau \coloneqq (\tau(x_1), \ldots, \tau(x_K)) \in \mathbb{R}^K$ be the collection of CATEs. For $d \in \{0, 1\}$ let $\nu_d(x) \coloneqq \mathbb{E}[Y(d) \mid X = x, W_0 = 1]$ denote the *average structural function* (ASF) which also conditions on $W_0 = 1$, let $\nu_d \coloneqq (\nu_d(x_1), \ldots, \nu_d(x_K)) \in \mathbb{R}^K$, and let $\mathcal{M} \subseteq \mathbb{R}^{2K}$ be a set of possible ASFs such that $(\nu_0, \nu_1) \in \mathcal{M}$. We now state the definition of weakly causal estimands from BBMT (see Definition WC) in our setting which features binary treatments.

Definition A.1. The estimand β is *weakly causal* if the following statements are true for all $(\nu_0, \nu_1) \in \mathcal{M}$:

1. If $\nu_1 - \nu_0 \geq \mathbf{0}_K$,[6] then $\beta \geq 0$.
2. If $\nu_1 - \nu_0 \leq \mathbf{0}_K$, then $\beta \leq 0$.

Thus, an estimand is weakly causal if all CATEs having the same sign implies the estimand also has that sign. Whether an estimand satisfies this condition also depends on $\mathcal{M}$, the set of allowed ASFs. To compare weak causality to our result on uniform causal representations, we consider $\mathcal{M}_{\text{all}} \coloneqq \mathbb{R}^{2K}$, the unrestricted set of ASFs. The corresponding unrestricted set of CATE functions, which we denoted by $\mathcal{T}_{\text{all}}$, allows τ to be any vector in $\mathbb{R}^K$. With these choices, we can show these two definitions are equivalent.

Proposition A.1. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1), and let $W_0 = 1$ almost surely. Suppose Assumption 3.1 holds and that $a_{\max} < \infty$. Then $\mu(a, \tau_0)$ is weakly causal with $\mathcal{M} = \mathcal{M}_{\text{all}}$ if and only if it has a causal representation uniformly in $\mathcal{T}_{\text{all}}$.

The proof of this proposition hinges on the equivalence, under some conditions, of weakly causal estimands and estimands with nonnegative weights, as in Proposition 4 of BBMT. Also, as shown in Theorem 3.1, estimands with nonnegative weights have a uniform causal representation in $\mathcal{T}_{\text{all}}$. Therefore, a weighted estimand has nonnegative weights if and only if it is weakly causal and if and only if it has a causal representation uniformly in $\mathcal{T}_{\text{all}}$. Thus, a weakly causal estimand admits a regular subpopulation W^* such that the estimand measures the average effect of treatment over that subpopulation.

B Proofs for Section 3

Proof of Proposition 3.1. We begin by showing the first claim of the proposition. The equation

$$\mathbb{E}[(Y(1) - Y(0))W^* \mid W_0 = 1, X] = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X]\mathbb{P}(W^* = 1 \mid W_0 = 1, X) \tag{B.1}$$

[6] Vector inequalities hold if they hold component-wise.

holds since $W^* \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$, which holds by Definition 3.1. Since $\mathbb{P}(W^* = 1 \mid W_0 = 1, X)$ is assumed positive, we can divide both sides of equation (B.1) by it and obtain

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X] &= \frac{\mathbb{E}[(Y(1) - Y(0))W^* \mid W_0 = 1, X]}{\mathbb{P}(W^* = 1 \mid W_0 = 1, X)} \\ &= \mathbb{E}[Y(1) - Y(0) \mid W^* = 1, W_0 = 1, X] \\ &= \mathbb{E}[Y(1) - Y(0) \mid W^* = 1, X],\end{aligned}$$

where the second equality holds by definition, and the third holds from W^* being a subpopulation of W_0.

We now show the second claim of the proposition. Equation (3.4) can be obtained as follows:

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] &= \mathbb{E}[Y(1) - Y(0) \mid W^* = 1, W_0 = 1] \\ &= \frac{\mathbb{E}[W^*(Y(1) - Y(0)) \mid W_0 = 1]}{\mathbb{E}[W^* \mid W_0 = 1]} \\ &= \frac{\mathbb{E}[\mathbb{E}[W^*(Y(1) - Y(0)) \mid X, W_0 = 1] \mid W_0 = 1]}{\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W_0 = 1) \mid W_0 = 1]} \\ &= \frac{\mathbb{E}[\mathbb{P}(W^* = 1 \mid W_0 = 1, X)\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X] \mid W_0 = 1]}{\mathbb{E}[\mathbb{P}(W^* = 1 \mid W_0 = 1, X) \mid W_0 = 1]} \\ &= \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} \\ &= \mu(\underline{w}^*, \tau_0).\end{aligned}$$

The first equality follows from W^* being a subpopulation of W_0, the second from the definition of conditional expectation and $\mathbb{P}(W^* = 1) > 0$, the third from the law of iterated expectations and $W^* = 1$ implying $W_0 = 1$, the fourth from $W^* \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$, and the fifth and sixth follow immediately. □

Proof of Lemma 3.1. We verify that W^* satisfies the two conditions in Definition 3.1. Condition 1 holds since $\mathbb{P}(W^* = 1 \mid W_0 = 0) = \mathbb{P}(W^* = 1, W' = 1 \mid W_0 = 0) \leq \mathbb{P}(W' = 1 \mid W_0 = 0) = 0$. The first equality follows from $W^* \leq W'$ almost surely and the second from $W' \in \mathrm{SP}(W_0)$. To verify condition 2, note that

$$\begin{aligned}\mathbb{P}(W^* = 1 \mid Y(1), Y(0), X, W_0 = 1) &= \mathbb{P}(W^* = 1, W' = 1 \mid Y(1), Y(0), X, W_0 = 1) \\ &= \mathbb{P}(W^* = 1 \mid Y(1), Y(0), X, W' = 1, W_0 = 1)\mathbb{P}(W' = 1 \mid Y(1), Y(0), X, W_0 = 1) \\ &= \mathbb{P}(W^* = 1 \mid Y(1), Y(0), X, W' = 1)\mathbb{P}(W' = 1 \mid Y(1), Y(0), X, W_0 = 1) \\ &= \mathbb{P}(W^* = 1 \mid X, W' = 1)\mathbb{P}(W' = 1 \mid X, W_0 = 1) \\ &= \mathbb{P}(W^* = 1 \mid X, W' = 1, W_0 = 1)\mathbb{P}(W' = 1 \mid X, W_0 = 1) \\ &= \mathbb{P}(W^* = 1, W' = 1 \mid X, W_0 = 1) \\ &= \mathbb{P}(W^* = 1 \mid X, W_0 = 1).\end{aligned}$$

The first and seventh line follow from $W^* \leq W'$ almost surely. The second and sixth line follow from factoring conditional probabilities. The third and fifth line follow from $W' \leq W_0$ almost surely. The fourth line follows from $W' \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$ and $W^* \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W' = 1$. □

Proof of Theorem 3.1. This theorem follows as a special case of the first part of Theorem 4.3 when $W' = W_0$. This is because W_0 is trivially a regular subpopulation of W_0, and because the condition

$$\sup(\mathrm{supp}(a(X)/\underline{w}'(X) \mid W_0 = 1)) < \infty$$

is equivalent to $a_{\max} < \infty$. This equivalence follows from $\underline{w}'(X) = \mathbb{P}(W_0 = 1 \mid X, W_0 = 1) = 1$ almost surely. □

Proof of Theorem 3.2. This is a special case of the second part of Theorem 4.3 where we set $W' = W_0$. □

Proof of Proposition 3.2. First, we suppose there exists $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{lin}})$. Therefore, by Proposition 3.1, we have that $\mu(a, \tau_0) - \mu(\underline{w}^*, \tau_0) = 0$ for all $\tau_0 \in \mathcal{T}_{\text{lin}}$ where $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid X, W_0 = 1)$. Therefore,

$$\begin{aligned}
0 &= \mu(a, \tau_0) - \mu(\underline{w}^*, \tau_0) \\
&= \frac{\mathbb{E}[a(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} - \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} \\
&= \frac{\mathbb{E}[a(X)(c + d'X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} - \frac{\mathbb{E}[\underline{w}^*(X)(c + d'X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} \\
&= d' \left(\frac{\mathbb{E}[a(X)X \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} - \frac{\mathbb{E}[\underline{w}^*(X)X \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} \right)
\end{aligned}$$

for all $d \in \mathbb{R}^{d_X}$, which implies that $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]} = \frac{\mathbb{E}[\underline{w}^*(X)X|W_0=1]}{\mathbb{E}[\underline{w}^*(X)|W_0=1]}$. Letting $u(x) = \underline{w}^*(x)/\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]$, we have that

$$\frac{\mathbb{E}[\underline{w}^*(X)X \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} = \int_{\text{supp}(X|W_0=1)} xu(x)\, dF_{X|W_0=1}(x),$$

a convex combination of x values in $\text{supp}(X \mid W_0 = 1)$ because $\int_{\text{supp}(X|W_0=1)} u(x)\, dF_{X|W_0=1}(x) = 1$. Therefore, $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]} \in \text{conv}(\text{supp}(X \mid W_0 = 1))$.

Now suppose that $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]} \in \text{conv}(\text{supp}(X \mid W_0 = 1))$. Then, we can write $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]}$ as $\int_{\text{supp}(X|W_0=1)} xu(x)\, dF_{X|W_0=1}(x)$ for some function $u(x) \geq 0$ satisfying $\int_{\text{supp}(X|W_0=1)} u(x)\, dF_{X|W_0=1}(x) = 1$. Let

$$W^* = \mathbb{1}\left(U \leq \frac{u(X)}{\sup(\text{supp}(u(X) \mid W_0 = 1))} \right) \cdot W_0,$$

where $U \sim \text{Unif}(0,1) \perp\!\!\!\perp (X, Y(1), Y(0), W_0)$. Then W^* is a regular subpopulation of W_0 and $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid X, W_0 = 1) = \frac{u(X)}{\sup(\text{supp}(u(X)|W_0=1))}$ since $\frac{u(X)}{\sup(\text{supp}(u(X)|W_0=1))} \in [0,1]$ with probability 1 given $W_0 = 1$. Therefore, for all $\tau_0(x) = c + d'x \in \mathcal{T}_{\text{lin}}$, we have that

$$\begin{aligned}
\mu(\underline{w}^*, \tau_0) &= \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} \\
&= \frac{\mathbb{E}\left[\frac{u(X)}{\sup(\text{supp}(u(X)|W_0=1))} \tau_0(X) \mid W_0 = 1 \right]}{\mathbb{E}\left[\frac{u(X)}{\sup(\text{supp}(u(X)|W_0=1))} \mid W_0 = 1 \right]} \\
&= \frac{\mathbb{E}\left[u(X)\tau_0(X) \mid W_0 = 1\right]}{\mathbb{E}\left[u(X) \mid W_0 = 1\right]} \\
&= \frac{\mathbb{E}\left[u(X)(c + d'X) \mid W_0 = 1\right]}{\mathbb{E}\left[u(X) \mid W_0 = 1\right]} \\
&= c + d' \frac{\mathbb{E}\left[u(X)X \mid W_0 = 1\right]}{\mathbb{E}\left[u(X) \mid W_0 = 1\right]} \\
&= c + d' \frac{\mathbb{E}\left[a(X)X \mid W_0 = 1\right]}{\mathbb{E}\left[a(X) \mid W_0 = 1\right]} \\
&= \frac{\mathbb{E}\left[a(X)(c + d'X) \mid W_0 = 1\right]}{\mathbb{E}\left[a(X) \mid W_0 = 1\right]} \\
&= \mu(a, \tau_0).
\end{aligned}$$

Therefore, by Proposition 3.1 we have that $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{lin}})$. □

Proof of Proposition 3.3. We consider the $K > 0$ case first and the $K = 0$ case second.

Case 1: $K > 0$.

First, let $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) < 1$. We will show that $\mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K)) = \emptyset$ by way of contradiction.

Suppose there is a $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K))$ and let $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid X, W_0 = 1) \in [0, 1]$. Let $\tau^*(x) = K \cdot \mathbb{1}(a(x) < 0)$. This definition implies $\tau^* \in \mathcal{T}_{\text{BD}}(K)$. Since $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K))$ we must have that $\mu(a, \tau_0) = \mu(\underline{w}^*, \tau_0)$ for all $\tau_0 \in \mathcal{T}_{\text{BD}}(K)$. Since $\tau^* \in \mathcal{T}_{\text{BD}}(K)$,

$$\begin{aligned}
0 &= \mu(a, \tau^*) - \mu(\underline{w}^*, \tau^*) \\
&= \mathbb{E}\left[\left(\frac{a(X)}{\mathbb{E}[a(X) \mid W_0 = 1]} - \frac{\underline{w}^*(X)}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]}\right) \tau^*(X) \mid W_0 = 1\right] \\
&= K \cdot \mathbb{E}\left[\left(\frac{a(X)}{\mathbb{E}[a(X) \mid W_0 = 1]} - \frac{\underline{w}^*(X)}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]}\right) \mathbb{1}(a(X) < 0) \mid W_0 = 1\right]. \qquad \text{(B.2)}
\end{aligned}$$

The right-hand side of (B.2) is K times the expectation of a nonpositive function. This follows from $a(X)\mathbb{1}(a(X) < 0) \leq 0$, $\underline{w}^*(X)\mathbb{1}(a(X) < 0) \geq 0$ by $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid X, W_0 = 1) \geq 0$, $\mathbb{E}[a(X) \mid W_0 = 1] > 0$ by Assumption 3.1, and $\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1] = \mathbb{P}(W^* = 1 \mid W_0 = 1) > 0$ by $W^* \in \text{SP}(W_0)$. Since $K > 0$, equation (B.2) implies the nonpositive function must equal 0 with probability 1 given $W_0 = 1$:

$$\begin{aligned}
1 &= \mathbb{P}\left(\left(\frac{a(X)}{\mathbb{E}[a(X) \mid W_0 = 1]} - \frac{\underline{w}^*(X)}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]}\right) \mathbb{1}(a(X) < 0) = 0 \mid W_0 = 1\right) \\
&= \mathbb{P}\left(a(X) = \frac{\mathbb{E}[a(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} \cdot \underline{w}^*(X) \mid a(X) < 0, W_0 = 1\right),
\end{aligned}$$

where the second equality follows from $\mathbb{P}(a(X) < 0 \mid W_0 = 1) > 0$. This implies $a(X)$ equals a positive multiple of $\underline{w}^*(X)$, a nonnegative quantity, with probability 1 given $\{a(X) < 0, W_0 = 1\}$, an event with positive probability that implies $a(X)$ is strictly negative. This is a contradiction and therefore $\mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K)) = \emptyset$.

Second, suppose $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) = 1$. By Theorem 3.1, $\mathcal{W}(a; W_0, \mathcal{T}_{\text{all}}) \neq \emptyset$. Since $\mathcal{T}_{\text{BD}}(K) \subseteq \mathcal{T}_{\text{all}}$, we have that $\mathcal{W}(a; W_0, \mathcal{T}_{\text{all}}) \subseteq \mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K))$. Therefore, $\mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K)) \neq \emptyset$, which means that $\mu(a, \tau_0)$ has a causal representation uniformly in $\tau_0 \in \mathcal{T}_{\text{BD}}(K)$.

Case 2: $K = 0$.

When $K = 0$, the function class $\mathcal{T}_{\text{BD}}(K)$ is the set of all constant functions. In this case, $\tau_0(X) = t_0$, where $t_0 \in \mathbb{R}$ denotes a constant. Thus $\mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(0)) \neq \emptyset$ for all weight functions $a(\cdot)$ since $W_0 \in \text{SP}(W_0)$ and because $\mu(a, \tau_0) = \mathbb{E}[a(X)t_0 \mid W_0 = 1]/\mathbb{E}[a(X) \mid W_0 = 1] = t_0 = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$ for all $a(\cdot)$. □

C Proofs for Section 4

Proof of Theorem 4.1. This is a special case of the first part of Theorem 4.4 where we set $W' = W_0$. □

We begin with a technical lemma that we use to prove the subsequent theorems.

Lemma C.1. Suppose Assumption 3.1 holds. Let $T_\mu = \tau_0(X) - \mu$. Then, for any $W' \in \text{SP}(W_0)$, we have that

1. The functions $\alpha \mapsto \mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W' = 1]$ and $\alpha \mapsto \mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha) \mid W' = 1]$ are left-continuous.
2. The functions $\alpha \mapsto \mathbb{E}[T_\mu \mathbb{1}(T_\mu > \alpha) \mid W' = 1]$ and $\alpha \mapsto \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha) \mid W' = 1]$ are right-continuous.

Proof of Lemma C.1. The function $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W' = 1]$ is left-continuous if for any strictly increasing sequence $\alpha_n \nearrow \alpha$ we have that $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha_n) \mid W' = 1] \to \mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W' = 1]$. To see this holds, note that $f_n(t) \coloneqq t\mathbb{1}(t < \alpha_n) \to t\mathbb{1}(t < \alpha)$ since $t\mathbb{1}(t < \alpha_n) = 0$ for all $t \geq \alpha$, and $t\mathbb{1}(t < \alpha_n) = t$ whenever $t < \alpha$ for sufficiently large n. The random variable $|T_\mu \mathbb{1}(T_\mu < \alpha_n)|$ is dominated by $|T_\mu|$ and $\mathbb{E}[|T_\mu| \mid W' = 1] < \infty$ by Assumption 3.1 and by $\mathbb{P}(W' = 1) > 0$. Therefore, by dominated convergence, $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha_n) \mid W' = 1] \to \mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W' = 1]$ hence $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W' = 1]$ is left-continuous. The function $\mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha) \mid W' = 1]$ is also left-continuous because $\mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha) \mid W' = 1] = \mathbb{E}[T_\mu \mid W' = 1] - \mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W' = 1]$. The lemma's second claim can be similarly shown by considering a sequence $\alpha_n \searrow \alpha$. □

Proof of Theorem 4.2. To simplify the notation in the proof, let $\mu \coloneqq \mu(a, \tau_0)$. We break down this proof into four cases.

Case 1: $\mu \in \mathcal{S}(\tau_0; W_0)$ and $\mu < \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$

We want to maximize $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ over the subpopulations W^* in $\mathcal{W}(a; W_0, \{\tau_0\})$. Recall that $W^* \in \mathcal{W}(a; W_0, \{\tau_0\})$ if $\mu = \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X)|W_0=1]}{\mathbb{E}[\underline{w}^*(X)|W_0=1]}$ and $W^* \in \mathrm{SP}(W_0)$ hold, where $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid X, W_0 = 1)$. Therefore,

$$\begin{aligned}
\overline{P}(a, W_0; \{\tau_0\}) &= \max_{W^* \in \mathcal{W}(a; W_0, \{\tau_0\})} \mathbb{P}(W^* = 1 \mid W_0 = 1) \\
&= \max_{W^* \in \{0,1\}: \mu = \mathbb{E}[\underline{w}^*(X)\tau_0(X)|W_0=1]/\mathbb{E}[\underline{w}^*(X)|W_0=1], W^* \in \mathrm{SP}(W_0)} \mathbb{P}(W^* = 1 \mid W_0 = 1) \\
&\leq \max_{W^* \in \{0,1\}: \mu = \mathbb{E}[\underline{w}^*(X)\tau_0(X)|W_0=1]/\mathbb{E}[\underline{w}^*(X)|W_0=1]} \mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W_0 = 1) \mid W_0 = 1] \\
&= \max_{\underline{w}^*: \mu = \mathbb{E}[\underline{w}^*(X)\tau_0(X)|W_0=1]/\mathbb{E}[\underline{w}^*(X)|W_0=1], \underline{w}^*(X) \in [0,1]} \mathbb{E}[\underline{w}^*(X) \mid W_0 = 1].
\end{aligned}$$

We will first show a closed-form expression for an upper bound for $\overline{P}(a, W_0; \{\tau_0\})$. Then, we will show that this upper bound can be attained by a corresponding $W^+ \in \mathcal{W}(a; W_0, \{\tau_0\})$, and therefore it equals $\overline{P}(a, W_0; \{\tau_0\})$.

Before proceeding, let $\alpha^+ = \inf\{\alpha \in \mathbb{R} : R(\alpha) \geq 0\}$ where $R(\alpha) = \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha) \mid W_0 = 1]$. By construction, $\alpha^+ \geq 0$. By $\mu < \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$ we also have that $\alpha^+ < +\infty$. By Lemma C.1, R is a right-continuous function, and therefore $R(\alpha^+) = \lim_{\alpha \searrow \alpha^+} R(\alpha) \geq 0$. We now claim that $\alpha^+ > 0$. To show this claim, assume $\alpha^+ = 0$. Then, $0 \leq R(\alpha^+) = R(0) = \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq 0) \mid W_0 = 1] \leq 0$, which implies $\mathbb{P}(\tau_0(X) = \mu \mid W_0 = 1) = 1$. This is ruled out by the assumption that $\mu > \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1] = \mathbb{E}[\tau_0(X) \mid W_0 = 1]$. Therefore, $\alpha^+ > 0$.

We now show an upper bound for $\overline{P}(a, W_0; \{\tau_0\})$. For all $\underline{w}^*$ such that $\mu = \mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]/\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]$ and $\underline{w}^*(X) \in [0, 1]$, we have that

$$\begin{aligned}
\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1] &= \frac{\mathbb{E}[\underline{w}^*(X)\alpha^+ \mid W_0 = 1]}{\alpha^+} \\
&= \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^+ - T_\mu) \mid W_0 = 1]}{\alpha^+} + \frac{\mathbb{E}[\underline{w}^*(X)T_\mu \mid W_0 = 1]}{\alpha^+} \\
&= \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^+ - T_\mu) \mid W_0 = 1]}{\alpha^+} \\
&= \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^+ - T_\mu)\mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+} + \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^+ - T_\mu)\mathbb{1}(T_\mu > \alpha^+) \mid W_0 = 1]}{\alpha^+} \\
&\leq \frac{\mathbb{E}[1 \cdot (\alpha^+ - T_\mu)\mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+} + \frac{\mathbb{E}[0 \cdot (\alpha^+ - T_\mu)\mathbb{1}(T_\mu > \alpha^+) \mid W_0 = 1]}{\alpha^+} \\
&= \mathbb{E}[\mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1] - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+} \\
&\coloneqq P^+.
\end{aligned}$$

The third equality follows from $\mu = \mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]/\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]$. The inequality follows from $\{0,1\}$ being lower/upper bounds for $\underline{w}^*(X)$.

Therefore, $\overline{P}(a, W_0; \{\tau_0\}) \leq P^+$. We now show that this inequality is binding by finding $W^+ \in \mathcal{W}(a; W_0, \{\tau_0\})$ such that $\mathbb{P}(W^+ = 1 \mid W_0 = 1) = P^+$.

Let

$$\underline{w}^+(X) = \begin{cases} 1 & \text{if } T_\mu < \alpha^+ \\ 1 - \frac{R(\alpha^+)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \neq 0)}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)} & \text{if } T_\mu = \alpha^+ \\ 0 & \text{if } T_\mu > \alpha^+. \end{cases}$$

This function is bounded above by 1 because $R(\alpha^+) \geq 0$ and $\alpha^+ > 0$. To show $\underline{w}^+$ is bounded below by 0, consider cases where $\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)$ or $R(\alpha^+)$ equal and differ from 0. If $\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) = 0$ or $R(\alpha^+) = 0$, then $\underline{w}^+(X) \in \{0,1\} \subseteq [0,1]$ and it is therefore bounded below by 0. If $\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) > 0$ and $R(\alpha^+) > 0$, then

$$\begin{aligned} 1 - \frac{R(\alpha^+)}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)} &= \frac{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) - \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)} \\ &= \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu = \alpha^+) \mid W_0 = 1] - \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)} \\ &= \frac{-\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha^+) \mid W_0 = 1]}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)}. \end{aligned}$$

By the definition of α^+ as an infimum, we must have that $R(\alpha^+ - \varepsilon) < 0$ for all $\varepsilon > 0$, implying that $R(\alpha)$ is discontinuous at α^+. By Lemma C.1, $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W_0 = 1]$ is left-continuous and satisfies $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha) \mid W_0 = 1] \leq R(\alpha)$. Therefore, since $R(\alpha^+ - \varepsilon) < 0$ for all $\varepsilon > 0$, we must have that $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha^+ - \varepsilon) \mid W_0 = 1] < 0$ for all $\varepsilon > 0$. Letting $\varepsilon \searrow 0$ yields that $\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha^+) \mid W_0 = 1] \leq 0$. Therefore $-\mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha^+) \mid W_0 = 1]/(\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)) \geq 0$ and $\underline{w}^+(X) \geq 0$ in that case as well.

Next, we compute

$$\begin{aligned} \mathbb{E}[\underline{w}^+(X) \mid W_0 = 1] &= \mathbb{E}[\mathbb{1}(T_\mu < \alpha^+) \mid W_0 = 1] \\ &\quad + \left(1 - \frac{R(\alpha^+)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \neq 0)}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)}\right) \mathbb{E}[\mathbb{1}(T_\mu = \alpha^+) \mid W_0 = 1] \\ &= \mathbb{E}[\mathbb{1}(T_\mu < \alpha^+) \mid W_0 = 1] + \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \\ &\quad - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+} \mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \neq 0) \\ &= \mathbb{P}(T_\mu \leq \alpha^+ \mid W_0 = 1) - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1]}{\alpha^+} \\ &= P^+. \end{aligned}$$

The indicator function disappears in the third equality because $\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) = 0$ implies $\mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1] = 0$ as shown above.

Finally we verify that $\frac{\mathbb{E}[\underline{w}^+(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^+(X) \mid W_0 = 1]} = \mu$. This condition is equivalent to $\mathbb{E}[\underline{w}^+(X) T_\mu \mid W_0 = 1] = 0$, which we verify here:

$$\begin{aligned} \mathbb{E}[\underline{w}^+(X) T_\mu \mid W_0 = 1] &= \mathbb{E}[T_\mu \mathbb{1}(T_\mu < \alpha^+) \mid W_0 = 1] \\ &\quad + \left(1 - \frac{R(\alpha^+)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \neq 0)}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)}\right) \mathbb{E}[T_\mu 1(T_\mu = \alpha^+) \mid W_0 = 1] \\ &= \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) \mid W_0 = 1] - \frac{R(\alpha^+)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \neq 0)}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)} \alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \end{aligned}$$

$$= R(\alpha^+) - R(\alpha^+)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \neq 0)$$
$$= R(\alpha^+)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) = 0).$$

Therefore, $\mathbb{E}[\underline{w}^+(X)T_\mu \mid W_0 = 1]$ equals 0 when $R(\alpha^+) = 0$. When $R(\alpha^+) > 0$, we have that $\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) > 0$ as shown earlier. So $\mathbb{E}[\underline{w}^+(X)T_\mu \mid W_0 = 1]$ is also equal to 0 in this case.

We conclude the proof by showing that $\underline{w}^+(X)$ corresponds to $\mathbb{P}(W^+ = 1 \mid X, W_0 = 1)$ for some $W^+ \in \mathrm{SP}(W_0)$. Let $U \sim \mathrm{Unif}(0,1)$ satisfy $U \perp\!\!\!\perp (Y(1), Y(0), X, W_0)$ and define

$$W^+ = \left(\mathbb{1}(T_\mu < \alpha^+) + \mathbb{1}\left(T_\mu = \alpha^+, U \leq 1 - \frac{R(\alpha^+)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1) \neq 0)}{\alpha^+ \mathbb{P}(T_\mu = \alpha^+ \mid W_0 = 1)}\right)\right) \cdot W_0.$$

By construction, $W^+ \in \{0,1\}$, $\mathbb{P}(W^+ = 1 \mid X, W_0 = 1) = \underline{w}^+(X)$, $\mathbb{P}(W^+ = 1 \mid W_0 = 0) = 0$, and $W^+ \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$. Also, since $\mu \in \mathcal{S}(\tau_0; W_0)$, $\mathbb{P}(T_\mu \leq 0 \mid W_0 = 1) = \mathbb{P}(\tau_0(X) \leq \mu \mid W_0 = 1) > 0$. Since $\alpha^+ > 0$ we have that $\mathbb{P}(W^+ = 1 \mid W_0 = 1) \geq \mathbb{P}(T_\mu < \alpha^+ \mid W_0 = 1) \geq \mathbb{P}(T_\mu \leq 0 \mid W_0 = 1) > 0$. Therefore W^+ is a regular subpopulation of W_0 for which $\mathbb{P}(W^+ = 1 \mid W_0 = 1) = P^+$, hence P^+ is the maximum.

Case 2: $\mu \in \mathcal{S}(\tau_0; W_0)$ **and** $\mu > \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$

As in case 1,

$$\overline{P}(a, W_0; \{\tau_0\}) \leq \max_{\underline{w}^*: \mu = \mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]/\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1], \underline{w}^*(X) \in [0,1]} \mathbb{E}[\underline{w}^*(X) \mid W_0 = 1].$$

Before proceeding, let $\alpha^- = \sup\{\alpha \in \mathbb{R} : L(\alpha) \leq 0\}$ where $L(\alpha) = \mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha) \mid W_0 = 1]$. By construction, $\alpha^- \leq 0$ and by $\mu > \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$ we have that $\alpha^- > -\infty$. By Lemma C.1, L is a left-continuous function, and therefore $L(\alpha^-) = \lim_{\alpha \nearrow \alpha^-} L(\alpha) \leq 0$. Similarly to case 1, we can show that $\alpha^- < 0$.

We now show an upper bound for $\overline{P}(a, W_0; \{\tau_0\})$. For all $\underline{w}^*$ such that $\mu = \mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]/\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]$ and $\underline{w}^*(X) \in [0,1]$, we have that

$$\begin{aligned}
\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1] &= \frac{\mathbb{E}[\underline{w}^*(X)\alpha^- \mid W_0 = 1]}{\alpha^-} \\
&= \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^- - T_\mu) \mid W_0 = 1]}{\alpha^-} + \frac{\mathbb{E}[\underline{w}^*(X)T_\mu \mid W_0 = 1]}{\alpha^-} \\
&= \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^- - T_\mu) \mid W_0 = 1]}{\alpha^-} \\
&= \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^- - T_\mu)\mathbb{1}(T_\mu \geq \alpha^-) \mid W_0 = 1]}{\alpha^-} + \frac{\mathbb{E}[\underline{w}^*(X)(\alpha^- - T_\mu)\mathbb{1}(T_\mu < \alpha^-) \mid W_0 = 1]}{\alpha^-} \\
&\leq \frac{\mathbb{E}[1 \cdot (\alpha^- - T_\mu)\mathbb{1}(T_\mu \geq \alpha^-) \mid W_0 = 1]}{\alpha^-} + \frac{\mathbb{E}[0 \cdot (\alpha^- - T_\mu)\mathbb{1}(T_\mu < \alpha^-) \mid W_0 = 1]}{\alpha^-} \\
&= \mathbb{E}[\mathbb{1}(T_\mu \geq \alpha^-) \mid W_0 = 1] - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha^-) \mid W_0 = 1]}{\alpha^-} \\
&\coloneqq P^-,
\end{aligned}$$

which follows a similar argument as above. This implies $\overline{P}(a, W_0; \{\tau_0\}) \leq P^-$. We now show that this inequality is an equality by finding $W^- \in \mathcal{W}(a; W_0, \{\tau_0\})$ such that $\mathbb{P}(W^- = 1 \mid W_0 = 1) = P^-$.

Let

$$\underline{w}^-(X) = \begin{cases} 1 & \text{if } T_\mu > \alpha^- \\ 1 - \frac{L(\alpha^-)\mathbb{1}(\mathbb{P}(T_\mu = \alpha^- \mid W_0 = 1) \neq 0)}{\alpha^- \mathbb{P}(T_\mu = \alpha^- \mid W_0 = 1)} & \text{if } T_\mu = \alpha^- \\ 0 & \text{if } T_\mu < \alpha^-. \end{cases}$$

The rest of the proof symmetrically follows the one for the previous case.

Case 3: $\mu \in \mathcal{S}(\tau_0; W_0)$ **and** $\mu = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$

Note that $W^* = W_0$ is the largest regular subpopulation of W_0. Since $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, we have that $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ is trivially maximized at 1.

Case 4: $\mu \notin \mathcal{S}(\tau_0; W_0)$

By Theorem 3.2, there does not exist a regular subpopulation W^* satisfying $\mu = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ and therefore the supremum equals 0 by its definition. □

Proof of Theorem 4.3.

Part 1: $\mathcal{T} = \mathcal{T}_{\text{all}}$

($\Longrightarrow$) First, suppose there exists $W^* \in \mathcal{W}(a; W', \mathcal{T}_{\text{all}})$. Using the law of iterated expectations and letting $\underline{w}'(X) = \mathbb{P}(W' = 1 \mid W_0 = 1, X)$, we can write

$$
\begin{aligned}
\mu(a, \tau_0) &= \frac{\mathbb{E}[a(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} \\
&= \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\tau_0(X)\underline{w}'(X) \mid W_0 = 1\right]}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\underline{w}'(X) \mid W_0 = 1\right]} \\
&= \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\tau_0(X)W' \mid W_0 = 1\right]}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}W' \mid W_0 = 1\right]} \\
&= \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\tau_0(X) \mid W' = 1, W_0 = 1\right]\mathbb{P}(W' = 1 \mid W_0 = 1)}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} \mid W' = 1, W_0 = 1\right]\mathbb{P}(W' = 1 \mid W_0 = 1)} \\
&= \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X] \mid W' = 1\right]}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} \mid W' = 1\right]} \\
&= \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\mathbb{E}[Y(1) - Y(0) \mid W' = 1, X] \mid W' = 1\right]}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} \mid W' = 1\right]} \\
&\coloneqq \mu'\left(\frac{a}{\underline{w}'}, \tau_0\right).
\end{aligned}
$$

The second equality is valid due to $\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W_0 = 1)) < \infty$. The third and fourth follow from the law of iterated expectations, and the fifth from W' being a subpopulation of W_0. The second to last line follows from Proposition 3.1 and from $\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W_0 = 1)) < \infty$ implying $\mathbb{P}(\underline{w}'(X) > 0 \mid W_0 = 1) > 0$.

Similarly, we can write $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \mu'(\underline{w}^*, \tau_0)$ where $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid W' = 1, X)$. Therefore, by Proposition 3.1, we have that $\mu'\left(\frac{a}{\underline{w}'}, \tau_0\right) - \mu'(\underline{w}^*, \tau_0) = 0$ for all $\tau_0 \in \mathcal{T}_{\text{all}}$.

Let $\tau^*(X) = \frac{a(X)/\underline{w}'(X)}{\mathbb{E}[a(X)/\underline{w}'(X) \mid W' = 1]} - \frac{\underline{w}^*(X)}{\mathbb{P}(W^* = 1 \mid W' = 1)}$. We have $\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W_0 = 1)) < \infty$ and $\mathbb{E}[\underline{w}^*(X)^2] \leq 1$ by construction. Hence, $\mathbb{E}[\tau^*(X)^2] < \infty$ and $\tau^* \in \mathcal{T}_{\text{all}}$.

Thus, we must have $\mu'\left(\frac{a}{\underline{w}'}, \tau^*\right) - \mu'(\underline{w}^*, \tau^*) = 0$. Expanding this equality yields

$$
0 = \mu'\left(\frac{a}{\underline{w}'}, \tau^*\right) - \mu'(\underline{w}^*, \tau^*)
$$

$$= \mathbb{E}\left[\tau^*(X)\left(\frac{a(X)/\underline{w}'(X)}{\mathbb{E}[a(X)/\underline{w}'(X) \mid W'=1]} - \frac{\underline{w}^*(X)}{\mathbb{P}(W^*=1 \mid W'=1)}\right) \mid W'=1\right]$$
$$= \mathbb{E}[\tau^*(X)^2 \mid W'=1].$$

This implies that $\mathbb{P}(\tau^*(X)=0 \mid W'=1)=1$. In turn, this implies that

$$\mathbb{P}\left(a(X) = \frac{\underline{w}'(X)\underline{w}^*(X)\mathbb{E}[a(X)/\underline{w}'(X) \mid W'=1]}{\mathbb{P}(W^*=1 \mid W'=1)} \mid W'=1\right) = 1. \tag{C.1}$$

We have that $\underline{w}^*(X) \geq 0$ and $\underline{w}'(X) \geq 0$ almost surely, and $\mathbb{P}(W^*=1 \mid W'=1) > 0$ by the assumption that $W^* \in \mathcal{W}(a; W', \mathcal{T}_{\text{all}})$. Also, $\mathbb{E}[a(X)/\underline{w}'(X) \mid W'=1] = \mathbb{E}[a(X) \mid W_0=1]/\mathbb{P}(W'=1 \mid W_0=1) > 0$ by Assumption 3.1 and by W' being a regular subpopulation of W_0. Therefore, $\mathbb{P}(a(X) \geq 0 \mid W'=1)=1$.

($\Longleftarrow$) Second, suppose that $\mathbb{P}(a(X) \geq 0 \mid W'=1)=1$ and fix $\tau_0 \in \mathcal{T}_{\text{all}}$. Let $\underline{w}'(X) = \mathbb{P}(W'=1 \mid X, W_0=1)$. As in the first part of the proof, recall that

$$\mu(a,\tau_0) = \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\mathbb{E}[Y(1)-Y(0) \mid W'=1, X] \mid W'=1\right]}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} \mid W'=1\right]}$$
$$\coloneqq \mu'\left(\frac{a}{\underline{w}'}, \tau_0\right).$$

Let $U \sim \text{Unif}(0,1)$ where $U \perp\!\!\!\perp (Y(1), Y(0), X, W_0, W')$, and define

$$W^* = \mathbb{1}\left(U \leq \frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1))}\right) \cdot W'.$$

We verify that W^* is a regular subpopulation of W'. First, we see that $\mathbb{P}(W^*=1) > 0$ because

$$\begin{aligned}
\mathbb{P}(W^*=1) &= \mathbb{P}(W^*=1 \mid W'=1)\mathbb{P}(W'=1) + \mathbb{P}(W^*=1 \mid W'=0)\mathbb{P}(W'=0) \\
&= \mathbb{P}\left(\mathbb{1}\left(U \leq \frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1))}\right) \cdot W' = 1 \mid W'=1\right)\mathbb{P}(W'=1) \\
&= \mathbb{E}\left[\mathbb{P}\left(U \leq \frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1))} \mid X, W'=1\right) \mid W'=1\right]\mathbb{P}(W'=1) \\
&= \mathbb{E}\left[\frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1))} \mid W'=1\right]\mathbb{P}(W'=1) \\
&= \frac{\mathbb{E}[a(X)/\underline{w}'(X) \mid W'=1]\mathbb{P}(W'=1)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1))} \\
&= \frac{\mathbb{E}[a(X) \mid W_0=1]\mathbb{P}(W_0=1)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1))} \\
&> 0.
\end{aligned}$$

The second equality follows from $\mathbb{P}(W^*=1 \mid W'=0)=0$, which holds by the construction of W^*. The fourth equality holds from $U \perp\!\!\!\perp (X, W')$ and from $\frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1))} \in [0,1]$ almost surely given $W'=1$. The sixth equality holds from $\mathbb{E}[a(X)/\underline{w}'(X) \mid W'=1] = \mathbb{E}[a(X) \mid W_0=1]/\mathbb{P}(W'=1 \mid W_0=1)$ and from W' being a subpopulation of W_0. To establish the final inequality, recall that $\mathbb{P}(W'=1)$ and $\mathbb{E}[a(X) \mid W_0=1]$ are positive by assumption. Also $\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W'=1)) \leq \sup(\text{supp}(a(X)/\underline{w}'(X) \mid W_0=1)) < \infty$ since $\text{supp}(a(X)/\underline{w}'(X) \mid W'=1)$ is a subset of $\text{supp}(a(X)/\underline{w}'(X) \mid W_0=1)$. That W^* satisfies the two properties of Definition 3.1 holds immediately. Therefore, W^* is a regular subpopulation of W'.

Finally, let

$$\begin{aligned}\underline{w}^*(X) &= \mathbb{P}(W^* = 1 \mid X, W' = 1) \\ &= \mathbb{P}\left(U \leq \frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W' = 1))} \mid X, W' = 1\right) \\ &= \frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W' = 1))}.\end{aligned}$$

Letting $\mathbb{E}[Y(1) - Y(0) \mid X, W' = 1] = \tau_0(X) \in \mathcal{T}_{\text{all}}$ and using Proposition 3.1, we can see that

$$\begin{aligned}\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] &= \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W' = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W' = 1]} \\ &= \frac{\mathbb{E}\left[\frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X)|W'=1))}\tau_0(X) \mid W' = 1\right]}{\mathbb{E}\left[\frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X)|W'=1))} \mid W' = 1\right]} \\ &= \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\tau_0(X) \mid W' = 1\right]}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} \mid W' = 1\right]} \\ &= \frac{\mathbb{E}[a(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} \\ &= \mu(a, \tau_0).\end{aligned}$$

Since $\tau_0 \in \mathcal{T}_{\text{all}}$ was arbitrary, we have that $W^* \in \mathcal{W}(a; W', \mathcal{T}_{\text{all}})$ which concludes the proof.

Part 2: $\mathcal{T} = \{\tau_0\}$

To simplify the notation in the proof, we let $\mu \coloneqq \mu(a, \tau_0)$.

($\Longrightarrow$) First, let $\mu \notin \mathcal{S}(\tau_0; W')$ and suppose there exists $W^* \in \mathcal{W}(a; W', \{\tau_0\})$. Since $\mu \notin \mathcal{S}(\tau_0; W')$, we can without loss of generality suppose that $\mathbb{P}(\tau_0(X) \leq \mu \mid W' = 1) = 0$, which implies $\mathbb{P}(\tau_0(X) > \mu \mid W' = 1) = 1$. Since $W^* \in \mathcal{W}(a; W', \{\tau_0\})$, we can write by Proposition 3.1

$$\begin{aligned}\mu &= \mathbb{E}[Y(1) - Y(0) \mid W^* = 1] \\ &= \frac{\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1)\mathbb{E}[Y(1) - Y(0) \mid X, W' = 1] \mid W' = 1]}{\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1]} \\ &= \frac{\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1)\tau_0(X) \mid W' = 1]}{\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1]} \\ &\geq \frac{\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1)\mu \mid W' = 1]}{\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1]} = \mu. \qquad \text{(C.2)}\end{aligned}$$

The inequality is strict unless $\mathbb{E}[\underbrace{(\tau_0(X) - \mu)}_{>0 \text{ w.p.1}} \underbrace{\mathbb{P}(W^* = 1 \mid X, W' = 1)}_{\in[0,1]} \mid W' = 1] = 0$ holds. This holds if $\mathbb{P}((\tau_0(X) - \mu)\mathbb{P}(W^* = 1 \mid X, W' = 1) = 0 \mid W' = 1) = 1$, which in turns occurs if and only if $\mathbb{P}(\mathbb{P}(W^* = 1 \mid X, W' = 1) = 0 \mid W' = 1) = 1$. This implies $\mathbb{P}(W^* = 1 \mid W' = 1) = 0$ and $\mathbb{P}(W^* = 1) = \mathbb{P}(W^* = 1 \mid W' = 1)\mathbb{P}(W' = 1) = 0$, a contradiction of $W^* \in \mathcal{W}(a, W', \{\tau_0\})$. Therefore, the inequality in (C.2) is strict and yields $\mu > \mu$, a contradiction. Therefore, $\mathcal{W}(a; W', \{\tau_0\}) = \emptyset$ when $\mu \notin \mathcal{S}(\tau_0; W')$.

($\Longleftarrow$) Second, let $\mu \in \mathcal{S}(\tau_0; W')$. Let

$$\mathcal{X}^- = \{x \in \text{supp}(X) : \tau_0(x) \leq \mu\} \qquad \text{and} \qquad \mathcal{X}^+ = \{x \in \text{supp}(X) : \tau_0(x) \geq \mu\}.$$

By $\mu \in \mathcal{S}(\tau_0; W')$, $\mathbb{P}(X \in \mathcal{X}^- \mid W' = 1) = \mathbb{P}(\tau_0(X) \leq \mu \mid W' = 1) > 0$. Similarly, $\mathbb{P}(X \in \mathcal{X}^+ \mid W' = 1) > 0$.

Let $U \sim \text{Unif}(0,1)$ where $U \perp\!\!\!\perp (Y(1), Y(0), X, W', W_0)$. For $u \in [0,1]$, let

$$W^*(u) = (\mathbb{1}(U > u, X \in \mathcal{X}^-) + \mathbb{1}(U \leq u, X \in \mathcal{X}^+)) \cdot W'.$$

We can see that $W^*(u) \in \{0,1\}$, that $W^*(u) \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W' = 1$, and that $\mathbb{P}(W^*(u) = 1 \mid W' = 0) = 0$. To show that $W^*(u)$ characterizes a regular subpopulation of W', we also show that it is nonzero with positive probability:

$$\begin{aligned}
\mathbb{P}(W^*(u) = 1 \mid W' = 1) &= \mathbb{P}(\mathbb{1}(U > u, X \in \mathcal{X}^-) + \mathbb{1}(U \leq u, X \in \mathcal{X}^+) = 1 \mid W' = 1) \\
&= \mathbb{P}(U > u, X \in \mathcal{X}^- \mid W' = 1) + \mathbb{P}(U \leq u, X \in \mathcal{X}^+ \mid W' = 1) \\
&= (1-u)\mathbb{P}(X \in \mathcal{X}^- \mid W' = 1) + u\mathbb{P}(X \in \mathcal{X}^+ \mid W' = 1) \\
&> 0
\end{aligned}$$

for all $u \in [0,1]$. Therefore, $\mathbb{P}(W^*(u) = 1) = \mathbb{P}(W^*(u) = 1 \mid W' = 1)\mathbb{P}(W' = 1) > 0$ by $\mathbb{P}(W' = 1) > 0$. Hence, $W^*(u) \in \text{SP}(W')$ for all $u \in [0,1]$.

For a given u, we have that $\underline{w}^*(X;u) \coloneqq \mathbb{P}(W^*(u) = 1 \mid X, W' = 1) = (1-u)\mathbb{1}(X \in \mathcal{X}^-) + u\mathbb{1}(X \in \mathcal{X}^+)$. Therefore, using Proposition 3.1,

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0) \mid W^*(u) = 1] &= \frac{\mathbb{E}[\underline{w}^*(X;u)\mathbb{E}[Y(1) - Y(0) \mid X, W' = 1] \mid W' = 1]}{\mathbb{E}[\underline{w}^*(X;u) \mid W' = 1]} \\
&= \frac{\mathbb{E}[((1-u)\mathbb{1}(X \in \mathcal{X}^-) + u\mathbb{1}(X \in \mathcal{X}^+))\tau_0(X) \mid W' = 1]}{\mathbb{E}[((1-u)\mathbb{1}(X \in \mathcal{X}^-) + u\mathbb{1}(X \in \mathcal{X}^+)) \mid W' = 1]} \\
&= \frac{(1-u)\mathbb{E}[\mathbb{1}(X \in \mathcal{X}^-)\tau_0(X) \mid W' = 1] + u\mathbb{E}[\mathbb{1}(X \in \mathcal{X}^+)\tau_0(X) \mid W' = 1]}{(1-u)\mathbb{P}(X \in \mathcal{X}^- \mid W' = 1) + u\mathbb{P}(X \in \mathcal{X}^+ \mid W' = 1)}.
\end{aligned}$$

By construction, $\tau_0(X)\mathbb{1}(X \in \mathcal{X}^-) \leq \mu\mathbb{1}(X \in \mathcal{X}^-)$ and $\tau_0(X)\mathbb{1}(X \in \mathcal{X}^+) \geq \mu\mathbb{1}(X \in \mathcal{X}^+)$ almost surely. Therefore,

$$\mathbb{E}[Y(1) - Y(0) \mid W^*(0) = 1] = \frac{\mathbb{E}[\mathbb{1}(X \in \mathcal{X}^-)\tau_0(X) \mid W' = 1]}{\mathbb{P}(X \in \mathcal{X}^- \mid W' = 1)} \leq \frac{\mathbb{E}[\mathbb{1}(X \in \mathcal{X}^-)\mu \mid W' = 1]}{\mathbb{P}(X \in \mathcal{X}^- \mid W' = 1)} = \mu$$

and

$$\mathbb{E}[Y(1) - Y(0) \mid W^*(1) = 1] = \frac{\mathbb{E}[\mathbb{1}(X \in \mathcal{X}^+)\tau_0(X) \mid W' = 1]}{\mathbb{P}(X \in \mathcal{X}^+ \mid W' = 1)} \geq \frac{\mathbb{E}[\mathbb{1}(X \in \mathcal{X}^+)\mu \mid W' = 1]}{\mathbb{P}(X \in \mathcal{X}^+ \mid W' = 1)} = \mu.$$

By the continuity of $\mathbb{E}[Y(1) - Y(0) \mid W^*(u) = 1]$ in u and the intermediate value theorem, there exists $u^* \in [0,1]$ such that $\mu = \mathbb{E}[Y(1) - Y(0) \mid W^*(u^*) = 1]$ and $W^*(u^*) \in \mathcal{W}(a; W', \{\tau_0\})$. □

Proof of Theorem 4.4.

Part 1: $\mathcal{T} = \mathcal{T}_{\text{all}}$

First suppose $\mathbb{P}(a(X) \geq 0 \mid W' = 1) = 1$. From Theorem 4.3, there exists $W^* \in \mathcal{W}(a; W', \mathcal{T}_{\text{all}})$. Written differently, we have that

$$\frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)}\tau_0(X) \mid W' = 1\right]}{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} \mid W' = 1\right]} = \mu(a, \tau_0) = \mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \frac{\mathbb{E}\left[\mathbb{P}(W^* = 1 \mid X, W' = 1)\tau_0(X) \mid W' = 1\right]}{\mathbb{E}\left[\mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1\right]}$$

for all $\tau_0 \in \mathcal{T}_{\text{all}}$. From derivations in the proof of Theorem 4.3 (see equation (C.1)) we have that $\mathbb{P}(C \cdot \frac{a(X)}{\underline{w}'(X)} = \mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1) = 1$ for some positive constant $C > 0$.

Since $\mathbb{P}(W^* = 1 \mid X, W' = 1) \leq 1$ almost surely given $W' = 1$, we must have $C \cdot a(X)/\underline{w}'(X) \leq 1$ almost

surely given $W' = 1$. This means C is bounded above by $\inf(\text{supp}(\underline{w}'(X)/a(X) \mid W' = 1))$, which is strictly positive by assumption. Therefore,

$$\begin{aligned}
\mathbb{P}(W^* = 1 \mid W' = 1) &= \mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1] \\
&\leq \inf\left(\text{supp}\left(\frac{\underline{w}'(X)}{a(X)} \mid W' = 1\right)\right) \mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} \mid W' = 1\right] \\
&= \inf\left(\text{supp}\left(\frac{\underline{w}'(X)}{a(X)} \mid W' = 1\right)\right) \frac{\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} W'\right]}{\mathbb{P}(W' = 1)} \\
&= \inf\left(\text{supp}\left(\frac{\underline{w}'(X)}{a(X)} \mid W' = 1\right)\right) \\
&\quad \cdot \mathbb{E}\left[\mathbb{E}\left[\frac{a(X)}{\underline{w}'(X)} W' \mid X, W_0 = 1\right] \mid W_0 = 1\right] \frac{\mathbb{P}(W_0 = 1)}{\mathbb{P}(W' = 1)} \\
&= \inf\left(\text{supp}\left(\frac{\underline{w}'(X)}{a(X)} \mid W' = 1\right)\right) \cdot \mathbb{E}[a(X) \mid W_0 = 1] \frac{\mathbb{P}(W_0 = 1)}{\mathbb{P}(W' = 1)}.
\end{aligned}$$

The fifth line follows from W' being a subpopulation of W_0, and the last line follows from the law of iterated expectations.

This upper bound is sharp because it is attained by setting

$$W^* = \mathbb{1}\left(U \leq \frac{a(X)/\underline{w}'(X)}{\sup(\text{supp}(a(X)/\underline{w}'(X) \mid W' = 1))}\right) \cdot W'$$

and noting that $W^* \in \mathcal{W}(a; W', \mathcal{T}_{\text{all}})$ from the proof of Theorem 4.3.

Now suppose $\mathbb{P}(a(X) \geq 0 \mid W' = 1) < 1$. By Theorem 4.3, $\mathcal{W}(a; W', \mathcal{T}_{\text{all}}) = \emptyset$ and therefore $\overline{P}(a, W'; \mathcal{T}_{\text{all}})$ is zero.

Part 2: $\mathcal{T} = \{\tau_0\}$

In this case, we seek to maximize $\mathbb{P}(W^* = 1 \mid W' = 1)$ subject to $W^* \in \mathcal{W}(a; W', \{\tau_0\})$. Using Proposition 3.1, we can write $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ as $\mathbb{E}[\tau_0(X)\mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1]/\mathbb{E}[\mathbb{P}(W^* = 1 \mid X, W' = 1) \mid W' = 1]$. The result then follows from a direct application of Theorem 4.2 that replaces W_0 by W' in its statement and $\mathbb{P}(W^* = 1 \mid X, W_0 = 1)$ by $\mathbb{P}(W^* = 1 \mid X, W' = 1)$ in the proofs. □

D Proofs for Section 5

Proof of Proposition 5.3. We begin by noting that

$$\begin{aligned}
\beta_{\text{TWFE}} &= \frac{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ddot{D}_t Y_t]}{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ddot{D}_t^2]} \\
&= \frac{\sum_{t=1}^{T} \mathbb{E}[\ddot{D}_t Y_t \mid P = t]\mathbb{P}(P = t)}{\sum_{t=1}^{T} \mathbb{E}[\ddot{D}_t^2 \mid P = t]\mathbb{P}(P = t)} \\
&= \frac{\mathbb{E}[\ddot{D}Y]}{\mathbb{E}[\ddot{D}^2]},
\end{aligned}$$

where the second equality follows from the uniform distribution of P which is independent from $(\ddot{D}_t, Y_t)$ for all $t \in \{1, \ldots, T\}$. The third equality follows from defining $\ddot{D} \coloneqq \ddot{D}_P$. We also note that

$$\ddot{D}_P = D_P - \frac{1}{T}\sum_{s=1}^{T} D_s - \sum_{t=1}^{T} \mathbb{E}[D_t]\mathbb{1}(P = t) + \sum_{s=1}^{T} \mathbb{E}[D_s]\mathbb{E}[\mathbb{1}(P = s)]$$

$$= D - \frac{1}{T}\sum_{s=1}^{T} \mathbb{1}(G \leq s) - \mathbb{E}[D \mid P] + \mathbb{E}[D]$$
$$= D - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D].$$

The third equality follows from $\mathbb{E}[D \mid G] = \mathbb{E}[\mathbb{1}(G \leq P) \mid G] = \frac{1}{T}\sum_{s=1}^{T} \mathbb{1}(G \leq s) = \frac{1}{T}\sum_{s=1}^{T} D_s$. We break down the rest of this proof into four steps.

Step 1: Splitting the Numerator in Two

We have that

$$\mathbb{E}[\ddot{D}Y] = \mathbb{E}[\ddot{D}(Y(0) + D(Y(1) - Y(0)))] = \mathbb{E}[\ddot{D}\mathbb{E}[Y(0) \mid G, P]] + \mathbb{E}[\ddot{D}D\mathbb{E}[Y(1) - Y(0) \mid G, P]].$$

The first equality follows from $Y = Y(0) + D(Y(1) - Y(0))$ and the second from iterated expectations and $\mathbb{E}[D \mid G, P] = D$.

Step 2: First Numerator Term

We have that

$$\mathbb{E}[\ddot{D}\mathbb{E}[Y(0) \mid G, P]] = \mathbb{E}[\ddot{D}\theta(G, P)] = \mathbb{E}[\ddot{D}\ddot{\theta}(G, P)],$$

where $\theta(G, P) = \mathbb{E}[Y(0) \mid G, P]$. The second equality follows by properties of projections and from defining $\ddot{\theta}(G, P)$ as follows:

$$\begin{aligned}\ddot{\theta}(G, P) &\coloneqq \theta(G, P) - \mathbb{E}[\theta(G, P) \mid G] - \mathbb{E}[\theta(G, P) \mid P] + \mathbb{E}[\theta(G, P)] \\ &= \mathbb{E}[Y(0) \mid G, P] - \mathbb{E}[Y(0) \mid G] - \mathbb{E}[Y(0) \mid P] + \mathbb{E}[Y(0)].\end{aligned}$$

Then, we note that

$$\begin{aligned}\ddot{\theta}(g', t') &= \mathbb{E}[Y(0) \mid G = g', P = t'] - \mathbb{E}[Y(0) \mid G = g'] - \mathbb{E}[Y(0) \mid P = t'] + \mathbb{E}[Y(0)] \\ &= \mathbb{E}[Y_{t'}(0) \mid G = g'] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_t(0) \mid G = g'] - \sum_{g \in \mathcal{G}} \left(\mathbb{E}[Y_{t'}(0) \mid G = g] - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_t(0) \mid G = g] \right) \mathbb{P}(G = g) \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{t'}(0) - Y_t(0) \mid G = g'] - \sum_{g \in \mathcal{G}} \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{t'}(0) - Y_t(0) \mid G = g] \right) \mathbb{P}(G = g).\end{aligned}$$

Assumption 5.5.2 implies that for any pair $t, t' \in \{1, \ldots, T\}$ and any $g' \in \mathcal{G}$

$$\mathbb{E}[Y_{t'}(0) - Y_t(0) \mid G = g'] = \mathbb{E}[Y_{t'}(0) - Y_t(0)].$$

This can be shown for $t' > t$ by writing $\mathbb{E}[Y_{t'}(0) - Y_t(0) \mid G = g'] = \sum_{s=t+1}^{t'} \mathbb{E}[Y_s(0) - Y_{s-1}(0) \mid G = g'] = \sum_{s=t+1}^{t'} \mathbb{E}[Y_s(0) - Y_{s-1}(0)] = \mathbb{E}[Y_{t'}(0) - Y_t(0)]$. Similar derivations show this holds for $t' < t$. The case where $t' = t$ is trivial. Therefore,

$$\begin{aligned}\theta(g', t') &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{t'}(0) - Y_t(0) \mid G = g'] - \sum_{g \in \mathcal{G}} \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{t'}(0) - Y_t(0) \mid G = g] \right) \mathbb{P}(G = g) \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{t'}(0) - Y_t(0)] - \sum_{g \in \mathcal{G}} \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[Y_{t'}(0) - Y_t(0)] \right) \mathbb{P}(G = g) \\ &= 0\end{aligned}$$

for all $(g', t') \in \mathcal{G} \times \{1, \ldots, T\}$, which implies $\mathbb{E}[\ddot{D}\mathbb{E}[Y(0) \mid G, P]] = 0$.

Step 3: Second Numerator Term

We can write

$$\begin{aligned}
\mathbb{E}[\ddot{D}D\mathbb{E}[Y(1) - Y(0) \mid G, P]] &= \mathbb{E}[(D - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D\mathbb{E}[Y(1) - Y(0) \mid G, P]] \\
&= \mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D\mathbb{E}[Y(1) - Y(0) \mid G, P]] \\
&= \mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])\mathbb{E}[D(Y(1) - Y(0)) \mid G, P]] \\
&= \mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])\mathbb{E}[Y(1) - Y(0) \mid G, P, D = 1]\mathbb{P}(D = 1 \mid G, P)].
\end{aligned}$$

The first equality is by definition, the second by $D^2 = D$, the third by $\mathbb{E}[D \mid G, P] = D$, and the fourth by the law of total probability.

Step 4: Denominator

In this step, we show that

$$\begin{aligned}
\mathbb{E}[\ddot{D}^2] &= \mathbb{E}[(D - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D] \\
&= \mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D] \\
&= \mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])\mathbb{P}(D = 1 \mid G, P)].
\end{aligned}$$

The first line is obtained from properties of linear projections, the second from $D^2 = D$, and the third from $D = \mathbb{P}(D = 1 \mid G, P)$.

We conclude the proof by noting the equivalence of integrating over the distribution of P and averages over time periods, which shows the equivalence between

$$\beta_{\text{TWFE}} = \frac{\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])\mathbb{E}[Y(1) - Y(0) \mid G, P, D = 1]\mathbb{P}(D = 1 \mid G, P)]}{\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])\mathbb{P}(D = 1 \mid G, P)]}$$

and the expression in Proposition 5.3. □

Proof of Proposition 5.4. Proposition 5.3 and $\mathbb{P}(D = 1 \mid G, P) = D$ yielded

$$\beta_{\text{TWFE}} = \frac{\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D\mathbb{E}[Y(1) - Y(0) \mid G, P, D = 1]]}{\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D]}.$$

Since $\mathbb{E}[Y(1) - Y(0) \mid G, P, D = 1] = \mathbb{E}[Y(1) - Y(0) \mid G, D = 1]$ by assumption, we can use the law of iterated expectations to obtain

$$\beta_{\text{TWFE}} = \frac{\mathbb{E}[\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D \mid G]\mathbb{E}[Y(1) - Y(0) \mid G, D = 1]]}{\mathbb{E}[\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D \mid G]]}.$$

We now calculate the conditional expectation $\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D \mid G = g]$ for $g \in \mathcal{G}$. If $g = +\infty$, then this conditional expectation is 0 by construction, so we focus on the case where $g \in \{2, \ldots, T\}$. For these derivations, we let $F_G(p) \coloneqq \mathbb{P}(G \leq p)$ denote the cdf of G at p.

$$\begin{aligned}
&\mathbb{E}[(1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D])D \mid G = g] \\
&= \mathbb{E}[D \mid G = g]\left(1 - \mathbb{E}[D \mid G = g] - \frac{\mathbb{E}[\mathbb{E}[D \mid P]D \mid G = g]}{\mathbb{E}[D \mid G = g]} + \mathbb{E}[D]\right) \\
&= \mathbb{E}[D \mid G = g]\left(1 - \mathbb{E}[\mathbb{1}(G \leq P) \mid G = g] - \frac{\mathbb{E}[F_G(P)\mathbb{1}(G \leq P) \mid G = g]}{\mathbb{E}[\mathbb{1}(G \leq P) \mid G = g]} + \mathbb{E}[\mathbb{E}[\mathbb{1}(G \leq P) \mid P]]\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[D \mid G=g]\left(1 - \mathbb{E}[\mathbb{1}(g \le P)] - \frac{\mathbb{E}[F_G(P)\mathbb{1}(g \le P)]}{\mathbb{E}[\mathbb{1}(g \le P)]} + \mathbb{E}[F_G(P)]\right) \\
&= \mathbb{E}[D \mid G=g]\,(1 - \mathbb{E}[\mathbb{1}(g \le P)] - \mathbb{E}[F_G(P) \mid g \le P] \\
&\quad + \mathbb{E}[F_G(P) \mid g \le P]\mathbb{E}[\mathbb{1}(g \le P)] + \mathbb{E}[F_G(P) \mid g > P]\mathbb{E}[\mathbb{1}(g > P)]) \\
&= \mathbb{E}[D \mid G=g]\mathbb{E}[\mathbb{1}(g > P)](1 + \mathbb{E}[F_G(P) \mid g > P] - \mathbb{E}[F_G(P) \mid g \le P]) \\
&= \mathbb{E}[D \mid G=g](1 - \mathbb{E}[D \mid G=g])(1 + \mathbb{E}[D \mid g > P] - \mathbb{E}[D \mid g \le P]) \\
&= \mathbb{P}(D=1 \mid G=g)\mathbb{P}(D=0 \mid G=g)(\mathbb{P}(D=1 \mid P < g) + \mathbb{P}(D=0 \mid P \ge g)).
\end{aligned}
$$

The first equality follows from $\mathbb{E}[D \mid G=g] > 0$ for $g \in \{2, \ldots, T\}$. The second follows from $D = \mathbb{1}(G \le P)$ and the law of iterated expectations, the third from $G \perp\!\!\!\perp P$, the fourth from definitions of conditional expectations and the law of iterated expectations, the fifth from combining terms, the sixth from the law of iterated expectations again, and the last line is obtained by the fact that $D \in \{0, 1\}$. The representation in Proposition 5.4 follows. □

E Proofs for Section 6

We use the following lemma in the proof of Theorem 6.1.

Lemma E.1. Let $\theta = (\theta(1), \ldots, \theta(K)) \in \mathbb{R}^K$ and define the mapping $\phi : \mathbb{R}^K \to \mathbb{R}$ by $\phi(\theta) = \max_{j \in \{1,\ldots,K\}} \theta(j)$. Then, ϕ is Hadamard directionally differentiable (HDD) for all $\theta \in \mathbb{R}^K$ tangentially to $\mathbb{R}^K$ with directional derivative at θ in direction $h \in \mathbb{R}^K$ equal to

$$
\phi'_\theta(h) = \max_{j \in \arg\max_{k \in \{1,\ldots,K\}} \theta(k)} h(j).
$$

Proof of Lemma E.1. Let $h_n \to h \in \mathbb{R}^K$ and $t_n \searrow 0$ as $n \to \infty$. Then,

$$
\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = \frac{\max_{k \in \{1,\ldots,K\}}(\theta(k) + t_n h_n(k)) - \max_{k \in \{1,\ldots,K\}} \theta(k)}{t_n}.
$$

Let $\Theta_{\max} = \{j \in \{1, \ldots, K\} : \theta(j) = \max_{k \in \{1,\ldots,K\}} \theta(k)\}$ and let $j_{\max}$ be an element of $\Theta_{\max}$. Then, $\max_{k \in \{1,\ldots,K\}} \theta(k) = \theta(j_{\max})$ and thus

$$
\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = \max\left\{\frac{\theta(1) - \theta(j_{\max})}{t_n} + h_n(1), \ldots, \frac{\theta(K) - \theta(j_{\max})}{t_n} + h_n(K)\right\}.
$$

For each $j \in \Theta_{\max}$, $(\theta(j) - \theta(j_{\max}))/t_n + h_n(j) = h_n(j) \to h(j)$. For each $j \notin \Theta_{\max}$, $(\theta(j) - \theta(j_{\max}))/t_n \to -\infty$ since $\theta(j) - \theta(j_{\max}) < 0$ and $t_n \searrow 0$. Therefore, by continuity of the maximum operator in its arguments,

$$
\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \to \max_{j \in \Theta_{\max}} h(j).
$$

□

Proof of Theorem 6.1. We begin by showing the consistency of $\widehat{\overline{P}}$ for $\overline{P}$.

Part 1: Consistency

The estimators $(\frac{1}{n}\sum_{i=1}^n \widehat{a}(X_i)\widehat{w}_0(X_i), \frac{1}{n}\sum_{i=1}^n \widehat{w}_0(X_i))$ are consistent for $(\mathbb{E}[a(X)w_0(X)], \mathbb{E}[w_0(X)])$ since their components are assumed consistent by Assumption 6.1. More explicitly, we can write

$$
\frac{1}{n}\sum_{i=1}^n \widehat{a}(X_i)\widehat{w}_0(X_i) = \sum_{j=1}^K \widehat{a}(x_j)\widehat{w}_0(x_j)\widehat{p}_j \xrightarrow{p} \sum_{j=1}^K a(x_j)w_0(x_j)p_j = \mathbb{E}[a(X)w_0(X)]
$$

by the continuous mapping theorem. The consistency of $\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_0(X_i)$ for $\mathbb{E}[w_0(X)]$ is similarly established.

We now consider the maximum term in the denominator. We can write

$$\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i) = \max_{x:\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i=x)>0,\widehat{w}_0(x)>c_n}\widehat{a}(x).$$

Let $\mathcal{X}^+ = \{x \in \text{supp}(X) : w_0(x) > 0\}$, which equals $\text{supp}(X \mid W_0 = 1)$, and let $\widehat{\mathcal{X}}^+ = \{x : \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i = x) > 0, \widehat{w}_0(x) > c_n\}$. We first show that $\mathbb{P}(\widehat{\mathcal{X}}^+ = \mathcal{X}^+) \to 1$ as $n \to \infty$. To see this, first consider $x_j \in \mathcal{X}^+$. Then,

$$\begin{aligned}\mathbb{P}(x_j \in \widehat{\mathcal{X}}^+) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i = x_j) > 0 \cap \widehat{w}_0(x_j) > c_n\right)\\ &\geq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i = x_j) > 0\right) + \mathbb{P}\left(\widehat{w}_0(x_j) > c_n\right) - 1.\end{aligned}$$

The above inequality was obtained from

$$\mathbb{P}(A \cap B) = 1 - \mathbb{P}(A^c \cup B^c) \geq 1 - (\mathbb{P}(A^c) + \mathbb{P}(B^c)) = \mathbb{P}(A) + \mathbb{P}(B) - 1,$$

where A and B are Borel sets.

We have that $\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i = x_j) > 0) \to 1$ since $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i = x_j) \xrightarrow{p} p_j > 0$. We also have that $\mathbb{P}(\widehat{w}_0(x_j) > c_n) = \mathbb{P}(\widehat{w}_0(x_j) - c_n > 0) \to 1$ because $\widehat{w}_0(x_j) - c_n \xrightarrow{p} w_0(x_j) > 0$ by $c_n = o(1)$ and $w_0(x_j) > 0$, which follows from $x_j \in \mathcal{X}^+$. Therefore, $\mathbb{P}(x_j \in \widehat{\mathcal{X}}^+) \geq \mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i = x_j) > 0) + \mathbb{P}(\widehat{w}_0(x_j) > c_n) - 1 \to 1$ as $n \to \infty$.

Now let $x_j \notin \mathcal{X}^+$. Then

$$\begin{aligned}\mathbb{P}(x_j \notin \widehat{\mathcal{X}}^+) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(X_i = x_j) = 0 \cup \widehat{w}_0(x_j) \leq c_n\right)\\ &\geq \mathbb{P}(\widehat{w}_0(x_j) \leq c_n)\\ &= \mathbb{P}(\sqrt{n}\widehat{w}_0(x_j) \leq \sqrt{n}c_n).\end{aligned}$$

By Assumption 6.1, $\sqrt{n}\widehat{w}_0(x_j) = \sqrt{n}(\widehat{w}_0(x_j) - w_0(x_j)) \xrightarrow{d} \mathbb{Z}_{\mathbf{w}_0}(j) = O_p(1)$, since $w_0(x_j) = 0$ for $x_j \notin \mathcal{X}^+$. Also $\sqrt{n}c_n \to +\infty$ by the theorem assumption. Therefore, $\mathbb{P}(\sqrt{n}\widehat{w}_0(x_j) \leq \sqrt{n}c_n) \to 1$ and $\mathbb{P}(x_j \notin \widehat{\mathcal{X}}^+) \to 1$ as $n \to \infty$. Because of this,

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{X}}^+ = \mathcal{X}^+) &= \mathbb{P}\left(\bigcap_{x_j\in\mathcal{X}^+}\{x_j \in \widehat{\mathcal{X}}^+\} \cap \bigcap_{x_j\notin\mathcal{X}^+}\{x_j \notin \widehat{\mathcal{X}}^+\}\right)\\ &= 1 - \mathbb{P}\left(\bigcup_{x_j\in\mathcal{X}^+}\{x_j \notin \widehat{\mathcal{X}}^+\} \cup \bigcup_{x_j\notin\mathcal{X}^+}\{x_j \in \widehat{\mathcal{X}}^+\}\right)\\ &\geq 1 - \left(\sum_{j:x_j\in\mathcal{X}^+}\mathbb{P}(x_j \notin \widehat{\mathcal{X}}^+) + \sum_{j:x_j\notin\mathcal{X}^+}\mathbb{P}(x_j \in \widehat{\mathcal{X}}^+)\right)\\ &\to 1 - \left(\sum_{j:x_j\in\mathcal{X}^+}0 + \sum_{j:x_j\notin\mathcal{X}^+}0\right)\\ &= 1.\end{aligned}$$

Using this, we obtain

$$\mathbb{P}\left(\max_{x\in\widehat{\mathcal{X}}^+}\widehat{a}(x)=\max_{x\in\mathcal{X}^+}\widehat{a}(x)\right)\geq\mathbb{P}(\widehat{\mathcal{X}}^+\in\mathcal{X}^+)\to 1,$$

which yields

$$\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)=\max_{x\in\widehat{\mathcal{X}}^+}\widehat{a}(x)=\max_{x\in\mathcal{X}^+}\widehat{a}(x)+o_p(1).$$

By the consistency of $\widehat{\mathbf{a}}$ for $\mathbf{a}$, the continuity of the maximum operator, and the continuous mapping theorem, $\max_{x\in\mathcal{X}^+}\widehat{a}(x)\xrightarrow{p}\max_{x\in\mathcal{X}^+}a(x)$. Because $\mathcal{X}^+=\text{supp}(X\mid W_0=1)$ is a finite set of points, we also have that $\max_{x\in\mathcal{X}^+}a(x)=\sup(\text{supp}(a(X)\mid W_0=1))$. Another application of the continuous mapping theorem suffices to show that $\widehat{\overline{P}}$ is consistent for $\overline{P}$.

Part 2: Asymptotic Distribution

We first establish the joint limiting distribution of terms (i) $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n\widehat{a}(X_i)\widehat{w}_0(X_i)-\mathbb{E}[a(X)w_0(X)])$, (ii) $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n\widehat{w}_0(X_i)-\mathbb{E}[w_0(X_i)])$, and (iii) $\sqrt{n}(\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)-\max_{x\in\mathcal{X}^+}a(x))$. The terms (i) and (ii) can be expanded as follows using a first-order expansion:

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n\widehat{a}(X_i)\widehat{w}_0(X_i)-\mathbb{E}[a(X)w_0(X)]\right)&=\sqrt{n}\left(\sum_{j=1}^K\widehat{a}(x_j)\widehat{w}_0(x_j)\widehat{p}_j-\sum_{j=1}^K a(x_j)w_0(x_j)p_j\right)\\&=\sum_{j=1}^K\left(w_0(x_j)p_j\sqrt{n}(\widehat{a}(x_j)-a(x_j))+a(x_j)p_j\sqrt{n}(\widehat{w}_0(x_j)-w_0(x_j))+a(x_j)w_0(x_j)\sqrt{n}(\widehat{p}_j-p_j)\right)+o_p(1)\end{aligned}\tag{E.1}$$

and

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n\widehat{w}_0(X_i)-\mathbb{E}[w_0(X)]\right)&=\sqrt{n}\left(\sum_{j=1}^K\widehat{w}_0(x_j)\widehat{p}_j-\sum_{j=1}^K w_0(x_j)p_j\right)\\&=\sum_{j=1}^K\left(p_j\sqrt{n}(\widehat{w}_0(x_j)-w_0(x_j))+w_0(x_j)\sqrt{n}(\widehat{p}_j-p_j)\right)+o_p(1).\end{aligned}\tag{E.2}$$

For term (iii), we use the expansion

$$\sqrt{n}\left(\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)-\max_{x\in\mathcal{X}^+}a(x)\right)=\sqrt{n}\left(\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)-\max_{x\in\mathcal{X}^+}\widehat{a}(x)\right)\tag{E.3}$$

$$+\sqrt{n}\left(\max_{x\in\mathcal{X}^+}\widehat{a}(x)-\max_{x\in\mathcal{X}^+}a(x)\right).\tag{E.4}$$

The term in (E.3) is of order $o_p(1)$ because

$$\mathbb{P}\left(\sqrt{n}\left(\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)-\max_{x\in\mathcal{X}^+}\widehat{a}(x)\right)=0\right)=\mathbb{P}\left(\max_{x\in\widehat{\mathcal{X}}^+}\widehat{a}(x)=\max_{x\in\mathcal{X}^+}\widehat{a}(x)\right)\geq\mathbb{P}(\widehat{\mathcal{X}}^+\in\mathcal{X}^+)\to 1$$

as shown above.

The term in (E.4) can be analyzed using Theorem 2.1 in Fang and Santos (2019), which generalizes the delta method to the class of Hadamard directionally differentiable functions. Using Lemma E.1, we have

that

$$\sqrt{n}\left(\max_{x\in\mathcal{X}^+}\widehat{a}(x)-\max_{x\in\mathcal{X}^+}a(x)\right)=\max_{x_j\in\arg\max_{x\in\mathcal{X}^+}a(x)}\sqrt{n}\left(\widehat{a}(x_j)-a(x_j)\right)+o_p(1)$$
$$\coloneqq\max_{j\in\Psi_{\mathcal{X}^+}}\sqrt{n}\left(\widehat{a}(x_j)-a(x_j)\right)+o_p(1). \tag{E.5}$$

Combining the expressions in (E.1), (E.2), and (E.5) with the delta method yields

$$\begin{aligned}
\sqrt{n}(\widehat{\overline{P}}-\overline{P}) &= \frac{1}{\mathbb{P}(W_0=1)\max_{x\in\mathcal{X}^+}a(x)}\sum_{j=1}^{K}\big(w_0(x_j)p_j\sqrt{n}(\widehat{a}(x_j)-a(x_j))+a(x_j)p_j\sqrt{n}(\widehat{w}_0(x_j)-w_0(x_j))\\
&\quad +a(x_j)w_0(x_j)\sqrt{n}(\widehat{p}_j-p_j)\big)\\
&\quad -\frac{\mathbb{E}[a(X)\mid W_0=1]}{\mathbb{P}(W_0=1)\max_{x\in\mathcal{X}^+}a(x)}\sum_{j=1}^{K}\big(p_j\sqrt{n}(\widehat{w}_0(x_j)-w_0(x_j))+w_0(x_j)\sqrt{n}(\widehat{p}_j-p_j)\big)\\
&\quad -\frac{\mathbb{E}[a(X)\mid W_0=1]}{\max_{x\in\mathcal{X}^+}a(x)^2}\max_{j\in\Psi_{\mathcal{X}^+}}\sqrt{n}\left(\widehat{a}(x_j)-a(x_j)\right)+o_p(1)\\
&= \psi(\sqrt{n}(\widehat{\mathbf{a}}-\mathbf{a}),\sqrt{n}(\widehat{\mathbf{w}}_0-\mathbf{w}_0),\sqrt{n}(\widehat{\mathbf{p}}-\mathbf{p}))+o_p(1)\\
&\xrightarrow{d}\psi(\mathbb{Z})
\end{aligned}$$

by the continuity of ψ and Assumption 6.1. □

Proof of Theorem 6.2. We verify the validity of the bootstrap by appealing to Theorem 3.2 in Fang and Santos (2019). We show that their Assumption 4 holds by showing the mapping $\widehat{\psi}$ satisfies $|\widehat{\psi}(h')-\widehat{\psi}(h)|\leq C_n\|h'-h\|$ for any $h',h\in\mathbb{R}^{3K}$ and for $C_n=O_p(1)$, and by showing that $\widehat{\psi}(h)\xrightarrow{p}\psi(h)$ for all $h\in\mathbb{R}^{3K}$.

Let $h=(h_1,h_2,h_3)$ and $h'=(h_1',h_2',h_3')$.

$$\begin{aligned}
|\widehat{\psi}(h')-\widehat{\psi}(h)| &\leq \left|\sum_{j=1}^{K}\frac{\widehat{w}_0(x_j)\widehat{p}_j}{\widehat{\mathbb{P}}(W_0=1)\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)}(h_1'(j)-h_1(j))\right|\\
&\quad +\left|\frac{\widehat{\mathbb{E}}[a(X)\mid W_0=1]}{\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)^2}\left(\max_{j\in\widehat{\Psi}_{\mathcal{X}^+}}h_1'(j)-\max_{j\in\widehat{\Psi}_{\mathcal{X}^+}}h_1(j)\right)\right|\\
&\quad +\left|\sum_{j=1}^{K}\frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])\widehat{p}_j}{\widehat{\mathbb{P}}(W_0=1)\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)}(h_2'(j)-h_2(j))\right|\\
&\quad +\left|\sum_{j=1}^{K}\frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])\widehat{w}_0(x_j)}{\widehat{\mathbb{P}}(W_0=1)\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)}(h_3'(j)-h_3(j))\right|\\
&\quad \leq \left(\sum_{j=1}^{K}\frac{\widehat{w}_0(x_j)^2\widehat{p}_j^2}{\widehat{\mathbb{P}}(W_0=1)^2\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)^2}\right)^{1/2}\|h_1'-h_1\| \qquad \text{(E.6)}\\
&\quad +\frac{|\widehat{\mathbb{E}}[a(X)\mid W_0=1]|}{\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)^2}\left|\max_{j\in\widehat{\Psi}_{\mathcal{X}^+}}h_1'(j)-\max_{j\in\widehat{\Psi}_{\mathcal{X}^+}}h_1(j)\right| \qquad \text{(E.7)}\\
&\quad +\left(\sum_{j=1}^{K}\frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])^2\widehat{p}_j^2}{\widehat{\mathbb{P}}(W_0=1)^2\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)^2}\right)^{1/2}\|h_2'-h_2\| \qquad \text{(E.8)}
\end{aligned}$$

$$+\left(\sum_{j=1}^{K}\frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])^2\widehat{w}_0(x_j)^2}{\widehat{\mathbb{P}}(W_0=1)^2\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)^2}\right)^{1/2}\|h_3'-h_3\|, \tag{E.9}$$

where we applied the Cauchy–Schwarz inequality several times. Note that the maximum function is Lipschitz with Lipschitz constant one and therefore

$$\begin{aligned}\left|\max_{j\in\widehat{\Psi}_{\mathcal{X}^+}}h_1'(j)-\max_{j\in\widehat{\Psi}_{\mathcal{X}^+}}h_1(j)\right| &\le \sum_{j\in\widehat{\Psi}_{\mathcal{X}^+}}|h_1'(j)-h_1(j)|\\ &\le \sum_{j=1}^{K}|h_1'(j)-h_1(j)|\\ &\le \sqrt{K}\|h_1'-h_1\|. \end{aligned}\tag{E.10}$$

Combining equations (E.6)–(E.9) with the consistency of $(\widehat{\mathbf{a}},\widehat{\mathbf{w}}_0,\widehat{\mathbf{p}})$ established in Theorem 6.1 shows that $|\widehat{\psi}(h')-\widehat{\psi}(h)|\le C_n\|h'-h\|$ for any $h',h\in\mathbb{R}^{3K}$ and for $C_n=O_p(1)$. Therefore, by Remark 3.4 in Fang and Santos (2019), showing $\widehat{\psi}(h)\xrightarrow{p}\psi(h)$ for all $h\in\mathbb{R}^{3K}$ suffices.

Thus we now consider the consistency of the different components of $\widehat{\psi}(h)$. Applying Theorem 6.1, we can show that

$$\begin{aligned}\sum_{j=1}^{K}\frac{\widehat{w}_0(x_j)\widehat{p}_j}{\widehat{\mathbb{P}}(W_0=1)\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)}h_1(j) &= \sum_{j=1}^{K}\frac{w_0(x_j)p_j}{\mathbb{P}(W_0=1)\sup_{x\in\mathcal{X}^+}a(x)}h_1(j)+o_p(1),\\ \frac{\widehat{\mathbb{E}}[a(X)\mid W_0=1]}{\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)^2} &= \frac{\mathbb{E}[a(X)\mid W_0=1]}{\sup_{x\in\mathcal{X}^+}a(x)^2}+o_p(1),\\ \sum_{j=1}^{K}\frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])\widehat{p}_j}{\widehat{\mathbb{P}}(W_0=1)\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)}h_2(j) &= \sum_{j=1}^{K}\frac{(a(x_j)-\mathbb{E}[a(X)\mid W_0=1])p_j}{\mathbb{P}(W_0=1)\sup_{x\in\mathcal{X}^+}a(x)}h_2(j)+o_p(1),\\ \sum_{j=1}^{K}\frac{(\widehat{a}(x_j)-\widehat{\mathbb{E}}[a(X)\mid W_0=1])\widehat{w}_0(x_j)}{\widehat{\mathbb{P}}(W_0=1)\max_{i:\widehat{w}_0(X_i)>c_n}\widehat{a}(X_i)}h_3(j) &= \sum_{j=1}^{K}\frac{(a(x_j)-\mathbb{E}[a(X)\mid W_0=1])w_0(x_j)}{\mathbb{P}(W_0=1)\sup_{x\in\mathcal{X}^+}a(x)}h_3(j)+o_p(1).\end{aligned}$$

It remains to show that $\max_{j\in\widehat{\Psi}_{\mathcal{X}^+}}h_1(j)=\max_{j\in\Psi_{\mathcal{X}^+}}h_1(j)+o_p(1)$. This holds if the set $\widehat{\Psi}_{\mathcal{X}^+}$ is consistent for $\Psi_{\mathcal{X}^+}$, which we establish here. Let $k\in\Psi_{\mathcal{X}^+}$. Then,

$$\begin{aligned}\mathbb{P}(k\in\widehat{\Psi}_{\mathcal{X}^+}) &= \mathbb{P}\left(\widehat{a}(x_k)\ge\max_{j\in\widehat{\mathcal{X}}^+}\widehat{a}(x_j)-\xi_n\right)\\ &= \mathbb{P}\left(\sqrt{n}(\widehat{a}(x_k)-\max_{j\in\widehat{\mathcal{X}}^+}\widehat{a}(x_j))\ge-\sqrt{n}\xi_n\right)\\ &= \mathbb{P}\left(\sqrt{n}(\max_{j\in\mathcal{X}^+}\widehat{a}(x_j)-\max_{j\in\widehat{\mathcal{X}}^+}\widehat{a}(x_j))\ge-\sqrt{n}\xi_n\right).\end{aligned}$$

The third equality follows from $k\in\Psi_{\mathcal{X}^+}$. By the proof of Theorem 6.1, $\sqrt{n}(\max_{j\in\mathcal{X}^+}\widehat{a}(x_j)-\max_{j\in\widehat{\mathcal{X}}^+}\widehat{a}(x_j))=o_p(1)$. Since $-\sqrt{n}\xi_n\to-\infty$, $\mathbb{P}(k\in\widehat{\Psi}_{\mathcal{X}^+})\to\mathbb{P}(0\ge-\infty)=1$ when $k\in\Psi_{\mathcal{X}^+}$.

Now suppose that $k\notin\Psi_{\mathcal{X}^+}$. Then,

$$\mathbb{P}(k\in\widehat{\Psi}_{\mathcal{X}^+})=\mathbb{P}(\widehat{a}(x_k)\ge\max_{j\in\widehat{\mathcal{X}}^+}\widehat{a}(x_j)-\xi_n)$$

$$\to \mathbb{P}(a(x_k) \geq \max_{j \in \mathcal{X}^+} a(x_j) - 0) = 0,$$

where the last equality holds from $k \notin \Psi_{\mathcal{X}^+}$. Therefore, $\mathbb{P}(\widehat{\Psi}_{\mathcal{X}^+} = \Psi_{\mathcal{X}^+}) \to 1$ as $n \to \infty$. This implies $\widehat{\psi}(h) \xrightarrow{p} \psi(h)$, which concludes the proof. □

F Proofs for Appendix A

Proof of Proposition A.1. By Theorem 3.1, $\mu(a, \tau_0)$ has a uniform causal representation in $\mathcal{T}_{\text{all}}$ if and only if $a(x_k) \geq 0$ for $k \in \{1, \ldots, K\}$. Therefore, it is sufficient to show the equivalence between weakly causal estimands and estimands with nonnegative weights. A similar result was shown in Proposition 4 of BBMT, but we nevertheless provide a proof here to account for the slight differences in notation.

Suppose $\mu(a, \tau_0)$ is weakly causal. Let $\nu_1 = (\mathbb{1}(a(x_1) < 0), \ldots, \mathbb{1}(a(x_K) < 0))$ and $\nu_0 = \mathbf{0}_K$. Trivially, $(\nu_1, \nu_0) \in \mathcal{M}_{\text{all}}$ and $\tau^- \coloneqq \nu_1 - \nu_0 \in \mathcal{T}_{\text{all}}$, where $\tau^- \geq \mathbf{0}_K$. Since $\mu(a, \tau_0)$ is weakly causal, $\mu(a, \tau^-) \geq 0$ where

$$\mu(a, \tau^-) = \frac{\mathbb{E}[a(X)\tau^-(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} = \frac{1}{\mathbb{E}[a(X) \mid W_0 = 1]} \sum_{k=1}^{K} a(x_k)\mathbb{1}(a(x_k) < 0)\mathbb{P}(X = x_k \mid W_0 = 1) \geq 0.$$

This implies $a(x_k) \geq 0$ for all $k \in \{1, \ldots, K\}$. Thus, $\mu(a, \tau_0)$ has a uniform causal representation in $\mathcal{T}_{\text{all}}$.

Now suppose $\mu(a, \tau_0)$ has a uniform causal representation in $\mathcal{T}_{\text{all}}$, or that $a(x_k) \geq 0$ for $k = 1, \ldots, K$. Then, for any $(\nu_0, \nu_1) \in \mathcal{M}_{\text{all}}$ such that $\tau \coloneqq \nu_1 - \nu_0 \geq \mathbf{0}_K$, we have that

$$\mu(a, \tau) = \frac{1}{\mathbb{E}[a(X) \mid W_0 = 1]} \sum_{k=1}^{K} a(x_k)\tau(x_k)\mathbb{P}(X = x_k \mid W_0 = 1) \geq 0.$$

The inequality holds because $a(x_k)$ and $\tau(x_k)$ are nonnegative for all $k \in \{1, \ldots, K\}$. This last inequality is reversed if we instead assume that $\tau \leq \mathbf{0}_K$. Thus, $\mu(a, \tau_0)$ is weakly causal. □

G Difference-in-Differences

Goodman-Bacon (2021) provides the following representation of the two-way fixed effects estimand under the assumption that group-level average treatment effects are constant over time:

$$\beta_{\text{TWFE}} = \sum_{k:\ \text{var}(D|G=k)>0} \left[\sum_{j=1}^{k-1} \sigma_{jk}^k + \sum_{j=k+1}^{K} \sigma_{kj}^k \right] \cdot \mathbb{E}[Y(1) - Y(0) \mid G = k, D = 1],$$

where

$$\sigma_{jk}^k = \frac{\mathbb{P}(G = j) \cdot \mathbb{P}(G = k) \cdot \mathbb{P}(D = 1 \mid G = k) \cdot \big[\mathbb{P}(D = 1 \mid G = j) - \mathbb{P}(D = 1 \mid G = k)\big]}{\text{var}(D^{\perp(G_{t_1}, \ldots, G_{t_{K-1}}, P_1, \ldots, P_T)})}$$

and

$$\sigma_{kj}^k = \frac{\mathbb{P}(G = j) \cdot \mathbb{P}(G = k) \cdot \big[1 - \mathbb{P}(D = 1 \mid G = k)\big]\big[\mathbb{P}(D = 1 \mid G = k) - \mathbb{P}(D = 1 \mid G = j)\big]}{\text{var}(D^{\perp(G_{t_1}, \ldots, G_{t_{K-1}}, P_1, \ldots, P_T)})}.$$

Here $A^{\perp B}$ is used to denote the residual in the linear projection of A on $(1, B)$. It is also the case that $\sum_{k:\ \mathrm{var}(D|G=k)>0} \sum_{l>k} \left(\sigma^k_{kl} + \sigma^l_{kl}\right) = 1$.[7] When we compare this representation with Proposition 5.4, that is,

$$\begin{aligned}\beta_{\mathrm{TWFE}} &= \frac{\mathbb{E}[a_{\mathrm{TWFE,H}}(G) \cdot \mathbb{P}(D=1 \mid G) \cdot \mathbb{E}[Y(1) - Y(0) \mid G, D=1]]}{\mathbb{E}[a_{\mathrm{TWFE,H}}(G) \cdot \mathbb{P}(D=1 \mid G)]} \\ &= \frac{\sum_{k:\ \mathrm{var}(D|G=k)>0} \mathbb{P}(G=k) \cdot a_{\mathrm{TWFE,H}}(k) \cdot \mathbb{P}(D=1 \mid G=k) \cdot \mathbb{E}[Y(1) - Y(0) \mid G=k, D=1]}{\sum_{k:\ \mathrm{var}(D|G=k)>0} \mathbb{P}(G=k) \cdot a_{\mathrm{TWFE,H}}(k) \cdot \mathbb{P}(D=1 \mid G=k)},\end{aligned}$$

it becomes clear that, for each group k other than the always treated and the never treated,

$$\begin{aligned}a_{\mathrm{TWFE,H}}(k) \cdot \mathbb{P}(D=1 \mid G=k) &= \sum_{j=1}^{k-1} \mathbb{P}(G=j) \cdot \mathbb{P}(D=1 \mid G=k) \cdot \big[\mathbb{P}(D=1 \mid G=j) - \mathbb{P}(D=1 \mid G=k)\big] \\ &+ \sum_{j=k+1}^{K} \mathbb{P}(G=j) \cdot \big[1 - \mathbb{P}(D=1 \mid G=k)\big] \big[\mathbb{P}(D=1 \mid G=k) - \mathbb{P}(D=1 \mid G=j)\big],\end{aligned}$$

and this, in turn, implies that

$$\begin{aligned}a_{\mathrm{TWFE,H}}(k) = &\sum_{j=1}^{k-1} \mathbb{P}(G=j) \cdot \big[\mathbb{P}(D=1 \mid G=j) - \mathbb{P}(D=1 \mid G=k)\big] \\ &+ \sum_{j=k+1}^{K} \mathbb{P}(G=j) \cdot \big[\mathbb{P}(D=1 \mid G=k) - \mathbb{P}(D=1 \mid G=j)\big] \cdot \frac{1 - \mathbb{P}(D=1 \mid G=k)}{\mathbb{P}(D=1 \mid G=k)}. \qquad \text{(G.1)}\end{aligned}$$

G.1 Equivalence of Weight Functions

We now show that the weights obtained in equation (G.1) are equivalent to those in Proposition 5.4. First, we rewrite the weights in (G.1) as follows:

$$\begin{aligned}a_{\mathrm{TWFE,H}}(k) = &\sum_{j=1}^{k-1} \mathbb{P}(G=j) \cdot \big[\mathbb{P}(D=1 \mid G=j) - \mathbb{P}(D=1 \mid G=k)\big] \\ &+ \sum_{j=k+1}^{K} \mathbb{P}(G=j) \cdot \big[\mathbb{P}(D=1 \mid G=k) - \mathbb{P}(D=1 \mid G=j)\big] \cdot \frac{1 - \mathbb{P}(D=1 \mid G=k)}{\mathbb{P}(D=1 \mid G=k)} \\ &= \mathbb{P}(D=1, G<k) - \mathbb{P}(G<k)\mathbb{E}[D \mid G=k] + (1 - \mathbb{E}[D \mid G=k])\mathbb{P}(G>k) \\ &\quad - \frac{1 - \mathbb{E}[D \mid G=k]}{\mathbb{E}[D \mid G=k]} \mathbb{P}(D=1, G>k) \\ &= \mathbb{P}(D=1, G<k) - \mathbb{P}(G<k)\mathbb{E}[D \mid G=k] + \mathbb{P}(G>k) - \mathbb{E}[D \mid G=k]\mathbb{P}(G>k) \\ &\quad - \frac{1}{\mathbb{E}[D \mid G=k]} \mathbb{P}(D=1, G>k) + \mathbb{P}(D=1, G>k) \\ &= \mathbb{P}(D=1, G \neq k) - \mathbb{E}[D \mid G=k]\mathbb{P}(G \neq k) + \mathbb{P}(G>k) \left(1 - \frac{\mathbb{E}[D \mid G>k]}{\mathbb{E}[D \mid G=k]}\right) \\ &= (\mathbb{E}[D] - \mathbb{E}[D \mid G=k]\mathbb{P}(G=k)) - \mathbb{E}[D \mid G=k]\mathbb{P}(G \neq k) + \mathbb{P}(G>k) \left(1 - \frac{\mathbb{E}[D \mid G>k]}{\mathbb{E}[D \mid G=k]}\right)\end{aligned}$$

[7] The result in Goodman-Bacon (2021) technically also includes a weight σ_{kU} attached to the contrast between group k and the never-treated group. We subsume this weight under σ^k_{kj}, and likewise subsume the weight on the contrast with the always-treated group under σ^k_{jk}.

$$= \mathbb{E}[D] - \mathbb{E}[D \mid G = k] + \mathbb{P}(G > k) \left(1 - \frac{\mathbb{E}[D \mid G > k]}{\mathbb{E}[D \mid G = k]}\right).$$

For $k \in \{2, \ldots, T\}$, the weights in Proposition 5.4 are equal to

$$\mathbb{E}[1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D] \mid G = k] = 1 - \mathbb{E}[D \mid G = k] - \mathbb{E}[D \mid P \geq k] + \mathbb{E}[D], \tag{G.2}$$

because they are the average of the weights in Proposition 5.3 conditional on $G = k$. The proof of Proposition 5.4 explicitly shows that

$$\mathbb{E}[1 - \mathbb{E}[D \mid G] - \mathbb{E}[D \mid P] + \mathbb{E}[D] \mid G = k] = \mathbb{P}(D = 0 \mid G = k) \cdot (\mathbb{P}(D = 0 \mid P \geq k) + \mathbb{P}(D = 1 \mid P < g)).$$

Let us look at the difference between the weights in (G.1) and (G.2). Fix $k \in \{2, \ldots, T\}$ and write

$$\begin{aligned}
&(1 - \mathbb{E}[D \mid G = k] - \mathbb{E}[D \mid P \geq k] + \mathbb{E}[D]) - \left(\mathbb{E}[D] - \mathbb{E}[D \mid G = k] + \mathbb{P}(G > k) \left(1 - \frac{\mathbb{E}[D \mid G > k]}{\mathbb{E}[D \mid G = k]}\right)\right) \\
&= 1 - \mathbb{E}[D \mid P \geq k] - \mathbb{P}(G > k) + \frac{\mathbb{E}[D\mathbb{1}(G > k)]}{\mathbb{E}[D \mid G = k]} \\
&= \mathbb{E}[\mathbb{1}(G \leq k)] - \frac{\mathbb{E}[D\mathbb{1}(P \geq k)]}{\mathbb{E}[\mathbb{1}(P \geq k)]} + \frac{\mathbb{E}[D\mathbb{1}(G > k)]}{\mathbb{E}[\mathbb{1}(k \leq P)]} \\
&= \frac{1}{\mathbb{E}[\mathbb{1}(k \leq P)]} \left(F_G(k)\mathbb{E}[\mathbb{1}(k \leq P)] + \mathbb{E}[D\mathbb{1}(G > k)] - \mathbb{E}[D\mathbb{1}(P \geq k)]\right) \\
&= \frac{1}{\mathbb{E}[\mathbb{1}(k \leq P)]} \left(F_G(k)\mathbb{E}[\mathbb{1}(k \leq P)] + \mathbb{E}[\mathbb{1}(k < G \leq P)] - \mathbb{E}[\mathbb{E}[D \mid P]\mathbb{1}(P \geq k)]\right) \\
&= \frac{1}{\mathbb{E}[\mathbb{1}(k \leq P)]} \left(F_G(k)\mathbb{E}[\mathbb{1}(k \leq P)] + \mathbb{E}[\mathbb{E}[\mathbb{1}(k < G \leq P) \mid P]] - \mathbb{E}[F_G(P)\mathbb{1}(P \geq k)]\right) \\
&= \frac{1}{\mathbb{E}[\mathbb{1}(k \leq P)]} \left(F_G(k)\mathbb{E}[\mathbb{1}(k \leq P)] + \mathbb{E}[(F_G(P) - F_G(k))\mathbb{1}(P \geq k)] - \mathbb{E}[F_G(P)\mathbb{1}(P \geq k)]\right) \\
&= \frac{1}{\mathbb{E}[\mathbb{1}(k \leq P)]} \left(F_G(k)\mathbb{E}[\mathbb{1}(k \leq P)] + \mathbb{E}[F_G(P)\mathbb{1}(P \geq k)] - F_G(k)\mathbb{E}[\mathbb{1}(P \geq k)] - \mathbb{E}[F_G(P)\mathbb{1}(P \geq k)]\right) \\
&= 0.
\end{aligned}$$

Therefore, the weights in Proposition 5.4 and equations (G.1) and (G.2) are all equal to one another.