Quantifying the Internal Validity of Weighted Estimands*

Alexandre Poirier† Tymon Słoczyński‡

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. We also apply these tools to revisit a prominent study of the effects of unilateral divorce laws on female suicide.

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 version: October 5, 2025. This paper was presented at LMU Munich, University of Bonn, Brown University, Brandeis University, University of Pittsburgh, McMaster University, Boston College, University of Oslo, Norwegian School of Economics, 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, the 2024 BU/BC Greenline Econometrics Workshop, and the 2025 Econometrics Mini-Conference at the University of Iowa. We thank audiences at those seminars and conferences, as well as Stéphane Bonhomme, Greg Caetano, Brant Callaway, Kevin Chen, Clément de Chaisemartin, Joachim Freyberger, Christian Hansen, Peter Hull, Shakeeb Khan, Toru Kitagawa, Matt Masten, Tomasz Olma, Guillaume Pouliot, Jonathan Roth, Pedro Sant'Anna, Andres Santos, 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 goal 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 $\mathbb{P}(D = 1 \mid X)$ 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.

There are two main questions that this paper seeks to answer. The first is whether, and under what conditions, 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 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.[1] 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 restrictions on 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

[1] We use the term "internal validity" because it is often associated with the question of whether the probability limit of an estimator is equal to the parameter of interest. If the researcher reports estimates of $\mu(a, \tau_0)$ when $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$ is their ultimate target parameter, then the estimator is internally valid for the target when $\mathbb{P}(W^* = 1 \mid W_0 = 1) = 1$. If $\mathbb{P}(W^* = 1 \mid W_0 = 1) < 1$, then $\mu(a, \tau_0)$ will be more informative about the target as the value of $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ increases.

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 common requirement that the weights underlying a suitable estimand must be nonnegative. In a related contribution, Blandhol, Bonney, Mogstad, and Torgovitsky (2022) show that the lack of negative weights and level dependence 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 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 establish the nonstandard limiting distribution of this estimator and describe inference procedures for it in Appendix C.

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.

Implications for Empirical Practice

Besides theoretical interest, we argue that $\mathbb{P}(W^* = 1)$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1)$ represent a valuable diagnostic for empirical research. First, as stated above, our

initial results demonstrate that two popular 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) argue 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) suggest 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 collapse 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 $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$. This makes our measures a practical diagnostic tool to evaluate the robustness of weighted estimands to heterogeneous treatment effects of any form.

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), Aronow and Samii (2016), Słoczyński (2022), Chen (2024), Goldsmith-Pinkham,

Hull, and Kolesár (2024), and Humphreys (2025) 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 nonnegative. 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 over a subpopulation. While the lack of negative weights is essential for subpopulation existence when the CATE function is unrestricted, nonuniform weights, too, have a detrimental effect on subpopulation size. 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 little attention to date.

One exception is de Chaisemartin (2012, 2017), who focuses on the interpretation of the instrumental variables (IV) estimand. 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 our paper. Second, when the usual monotonicity assumption is violated, de Chaisemartin (2012, 2017) uses a specific restriction on the CATE function to reinterpret 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.

Another exception is Aronow and Samii (2016), who 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 fixed effects estimands and argue that it is problematic if "switchers" are a small subset of the sample. We similarly argue 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 briefly discuss our motivating example of the OLS estimand. 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 size of subpopulations associated with the estimand in (1.1). In Section 5, we revisit our motivating example from Section 2 and apply our theoretical results to additional examples of weighted estimands. 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 appendices contain our proofs as well as several additional results and derivations.

2 Motivating Example

Here we provide further discussion of the OLS estimand. 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 assumption.

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) under a restriction on

the propensity score. 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 a weighted estimand satisfying the representation in (1.1) 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 $W_0 = D$, $w_0(X) = p(X)$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D = 1, X] = \mathbb{E}[Y(1) - Y(0) \mid X]$, which follows from conditional independence, and let $\tilde{a}(X) = 1 - p(X)$. Then, we can write

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

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

In this section, we consider a general class of weighted estimands. 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 treatment effect heterogeneity.

3.1 Preliminaries

Recall the earlier setting where we let $D \in \{0,1\}$ denote a treatment variable, and let $(Y(0), Y(1))$ denote the corresponding potential outcomes. 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. 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.

Note that τ_0 is defined for all values of X such that $w_0(X) = \mathbb{P}(W_0 = 1 \mid X) > 0$.[2] 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 condition.

Assumption 3.1. 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 ensure the existence 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.

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

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

$W_0 = 1$ almost surely for simplicity. Then,

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

which are weights that sum to one. The representations in (1.1) and (3.1) are equivalent since 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 by the estimand relative to $\{X = x_{k'}\}$, when compared to their relative weights in the population. Similar algebra can be used to write the estimand as an integral when X is continuously distributed. We instead 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 only consider what we call "regular subpopulations," defined 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) $\mathbb{P}(W_0 = 1 \mid W^* = 1) = 1$;
2. (Conditional independence) $W^* \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$.

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

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

These subpopulations have positive masses and are subsets of $\{W_0 = 1\}$. The other substantive requirement is that they 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 $\text{supp}(X)$. 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 $x \in \text{supp}(X)$" can be constructed as $W^* = \mathbb{1}(U \leq a(X))$ where $U \sim \text{Unif}(0, 1)$ is independent from $(Y(1), Y(0), X)$.

The conditional independence requirement rules out subpopulations that directly depend on the potential outcomes such as $W^* = \mathbb{1}(Y(1) \geq Y(0))$, i.e., 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 $W^* \in \text{SP}(W_0)$. Suppose Assumption 3.1 holds. Let $\underline{w}^*(x) \coloneqq \mathbb{P}(W^* = 1 \mid X = x, W_0 = 1)$. Then,

1. $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1, X] = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X]$ when $\underline{w}^*(X) > 0$;
2. $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} = \mu(\underline{w}^*, \tau_0)$.

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 units, $W^* = D$. The second property allows us to write the average treatment 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 treatment effects for regular subpopulations 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. 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 functions τ_0 in the set $\mathcal{T}$. If the set $\mathcal{W}(a; W_0, \mathcal{T})$ is empty, then the 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 further examine several 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 $a_{\max} \coloneqq \sup(\text{supp}(a(X) \mid W_0 = 1))$ 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

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

A uniform (in $\mathcal{T}_{\text{all}}$) causal representation exists if and only if $a(X)$ is nonnegative when $W_0 = 1$. 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 treatment 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 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 $\tau_0 \in \mathcal{T}_{\text{all}}$.

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 well established. Blandhol, Bonney, Mogstad, and Torgovitsky (2022) show that it is equivalent to an estimand being "weakly causal," which means that it will match the sign of τ_0 whenever that sign 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 B.

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 $a(X)$. 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 τ_0 if and only if

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

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 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 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 X has finite support. 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 placing restrictions on $\mathcal{T}$ 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 values of $(a(0), a(1), a(2))$ 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 $\tau_0 \in \mathcal{T}_{\text{lin}}$. The result is stated for discrete X, but an estimand with negative weights can have a causal representation even when X has continuous components.

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 $a(\cdot)$ 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.

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 $\mu(a, \tau_0)$.

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

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

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

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

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. The 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 $\overline{P}(a, W_0; \mathcal{T})$. We focus on 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

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

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.

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

$$\mu(a, \tau_0) = \frac{\mathbb{E}[a(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[a(X) \mid W_0 = 1]} = \frac{\mathbb{E}[\underline{w}^*(X)\tau_0(X) \mid W_0 = 1]}{\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]} = \mu(\underline{w}^*, \tau_0) \quad (4.1)$$

for all $\tau_0 \in \mathcal{T}_{\text{all}}$. Equation (4.1) 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.

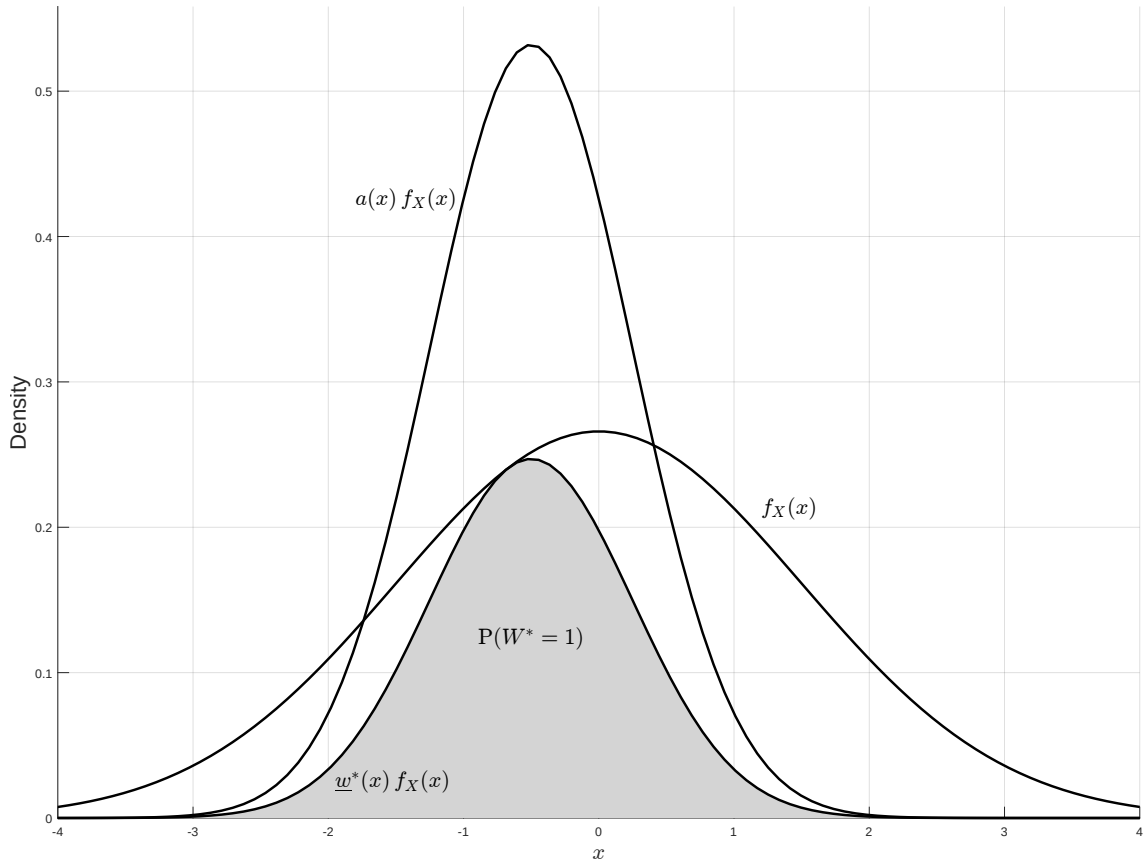
Several further comments about Theorem 4.1 are in order.

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

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

where $a(X)/a_{\max} \in [0, 1]$ almost surely when weights are nonnegative. From this

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



Note: X is a single continuously distributed covariate with density f_X.

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 $W_0 = 1$ and $a(X) = a_{\max}$ almost surely, 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 unconfoundedness, we also note that the average treatment effect on the treated (ATT) can be written as $\mathbb{E}[Y(1) - Y(0) \mid D = 1] = \mu(a, \tau_0)$, a weighted estimand with weights $a(X) = \mathbb{P}(D = 1 \mid X)$. This means that it 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 $W^* = \mathbb{1}(U \leq a(X)/a_{\max}) \cdot W_0$.

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 included in 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 (Meaning of internal validity and representativeness)**.** Suppose we consider $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ to be the target parameter, 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 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 example illustrates how the internal validity of $\mu(a, \tau_0)$ is entirely dependent on the target estimand. This is appropriate because we associate the term "internal validity" with the question of whether the probability limit of an estimator is equal to the parameter of interest. Thus, the degree of internal validity should naturally depend

on the parameter choice. On the other hand, the representativeness of the weighted estimand, as measured by the largest value of $\mathbb{P}(W^* = 1)$, is independent of the target parameter. In fact, it will be less than one unless $W^* = 1$ almost surely, or that the weighted estimand actually equals the ATE. If, in addition, we wish to operationalize the concept of external validity in our framework, we need to consider whether the weighted estimand can be written as the average treatment effect within a subpopulation of another, possibly arbitrary population. This extension is left for future work.

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

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 equation (3.1), $\mu(a, \tau_0)$ 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, when $a(1) = a(2)$, $\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = \mathbb{E}[a(X)]/\sup_{x\in\{1,2\}} a(x) = a(2)/a(2) = 1$.

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 weighted 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

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

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 $\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = \mathbb{E}[a(X)]/\left(\sup_{x\in\{1,2\}} a(x)\right) = (a(1)p_1 + a(2)p_2)/a(1) = 0.5$. The average effect in this subpopulation is given by

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

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 $\frac{\mathbb{P}(X=1|W^*=1)}{\mathbb{P}(X=2|W^*=1)} = 0.4/0.6 = \omega_1/\omega_2$, 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 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 weighted 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 quantify its degree of internal validity for the ATE. To simplify notation, assume that

$W_0 = 1$ almost surely and denote $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$ by $\overline{P}_{\text{all}}(a)$. Abstracting from sample uncertainty, we only assume knowledge of the weighted estimand and its internal validity. We can decompose the target estimand, here $\mathbb{E}[Y(1) - Y(0)]$, as

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

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 $\text{supp}(Y(1) - Y(0)) \subseteq [B_\ell, B_u]$, bounds for the target estimand are given by

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0)] \; \in \; & [\mu(a, \tau_0) \cdot \mathbb{P}(W^* = 1) + B_\ell \cdot (1 - \mathbb{P}(W^* = 1)), \\ & \mu(a, \tau_0) \cdot \mathbb{P}(W^* = 1) + B_u \cdot (1 - \mathbb{P}(W^* = 1))] \,. \end{aligned} \tag{4.2}$$

The width of these bounds is minimized when $\mathbb{P}(W^* = 1)$ is maximized, or when it equals the measure of internal validity for $\mu(a, \tau_0)$, given by $\overline{P}_{\text{all}}(a)$. The resulting bounds for $\mathbb{E}[Y(1) - Y(0)]$ can be written as

$$\left[\mu(a, \tau_0) \cdot \overline{P}_{\text{all}}(a) + B_\ell \cdot (1 - \overline{P}_{\text{all}}(a)), \; \mu(a, \tau_0) \cdot \overline{P}_{\text{all}}(a) + B_u \cdot (1 - \overline{P}_{\text{all}}(a))\right] \tag{4.3}$$

$$= \left[\mu(a, \tau_0) \cdot \frac{\mathbb{E}[a(X)]}{a_{\max}} + B_\ell \cdot \left(1 - \frac{\mathbb{E}[a(X)]}{a_{\max}}\right), \; \mu(a, \tau_0) \cdot \frac{\mathbb{E}[a(X)]}{a_{\max}} + B_u \cdot \left(1 - \frac{\mathbb{E}[a(X)]}{a_{\max}}\right)\right]. \tag{4.4}$$

If $\overline{P}_{\text{all}}(a) = 1$, it is easy to see that the estimand equals the ATE and that the bounds in (4.3) collapse to a point. However, the ATE is not uniquely determined from $(\mu(a, \tau_0), \overline{P}_{\text{all}}(a))$ when $\overline{P}_{\text{all}}(a) < 1$.

The width of these bounds is $(B_u - B_\ell) \cdot (1 - \overline{P}_{\text{all}}(a))$. Hence for fixed (B_ℓ, B_u), this width decreases linearly with $\overline{P}_{\text{all}}(a)$. Moreover, values of $\overline{P}_{\text{all}}(a)$ close to 1, or high degrees of internal validity, lead to narrow bounds. It is easy to obtain a sample analog of these bounds by combining estimators for $\mu(a, \tau_0)$ and our proposed estimator for $\overline{P}_{\text{all}}(a)$ 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 (4.2). We focus here on the case where we add a single piece of

additional information to $\mu(a, \tau_0)$, namely its internal validity, 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 can be 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:

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

where $\omega^+ \coloneqq \mathbb{E}[\max\{a(X), 0\}]/\mathbb{E}[a(X)]$ and $\omega^- \coloneqq \mathbb{E}[-\min\{a(X), 0\}]/\mathbb{E}[a(X)]$ 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 $a(\cdot)$. Thus, by Theorem 3.1, we can write

$$\mu(a, \tau_0) \;=\; \omega^+ \cdot \mathbb{E}[Y(1) - Y(0) \mid W^+ = 1] \;-\; \omega^- \cdot \mathbb{E}[Y(1) - Y(0) \mid W^- = 1], \quad (4.5)$$

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

$$\begin{aligned}&\mathbb{E}[Y(1) - Y(0) \mid W^+ = 1] \cdot \mathbb{P}(W^+ = 1) \;+\; \mathbb{E}[Y(1) - Y(0) \mid W^- = 1] \cdot \mathbb{P}(W^- = 1) \\ &+\; \mathbb{E}[Y(1) - Y(0) \mid W^+ + W^- = 0] \cdot \left[1 - \mathbb{P}(W^+ = 1) - \mathbb{P}(W^- = 1)\right]. \qquad (4.6)\end{aligned}$$

Substituting equation (4.5) in (4.6) 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

$$\begin{aligned}\mathbb{E}[Y(1)-Y(0)] \;\in\; &\left[\frac{\mathbb{P}(W^+=1)}{\omega^+}\cdot\mu(a,\tau_0)+\left(1-\frac{\mathbb{P}(W^+=1)}{\omega^+}\right)\cdot B_\ell,\right.\\ &\left.\frac{\mathbb{P}(W^+=1)}{\omega^+}\cdot\mu(a,\tau_0)+\left(1-\frac{\mathbb{P}(W^+=1)}{\omega^+}\right)\cdot B_u\right]. \qquad (4.7)\end{aligned}$$

This interval is similar to the interval in (4.2), 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 (4.7) 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 $\mu(\max\{a, 0\}, \tau_0)$ where $\max\{a, 0\} \geq 0$, and compute the value of ω^+. The resulting interval depends only on the ratio of the two quantities, which can be written as

$$\begin{aligned}\frac{\mathbb{P}(W^+=1)}{\omega^+} \;&\leq\; \frac{\overline{P}_{\text{all}}(\max\{a,0\})}{\omega^+} \;=\; \frac{\mathbb{E}[\max\{a(X),0\}]/\sup(\text{supp}(\max\{a(X),0\}))}{\mathbb{E}[\max\{a(X),0\}]/\mathbb{E}[a(X)]}\\ &=\; \mathbb{E}[a(X)]/a_{\max}.\end{aligned}$$

This last expression equals the level of internal validity of the original estimand $\mu(a, \tau_0)$ when it is assumed (perhaps incorrectly) to have nonnegative weights. It follows that the bounds in (4.4) are valid regardless of whether weights are nonnegative, because minimizing the length of the interval in (4.7) yields precisely the bounds in (4.4). 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

$$\overline{P}(a, W_0; \{\tau_0\}) = \sup_{W^* \in \mathcal{W}(a; W_0, \{\tau_0\})} \mathbb{P}(W^* = 1 \mid W_0 = 1), \qquad (4.8)$$

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 $\mu_0 = \mu(\underline{w}^*, \tau_0)$, where we let $\mu_0 \coloneqq \mu(a, \tau_0)$ to simplify the notation.

This condition is equivalent to $\mathbb{E}[(\tau_0(X) - \mu_0)\underline{w}^*(X) \mid W_0 = 1] = 0$, 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 (4.8) 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 $f_k \coloneqq \mathbb{P}(W^* = 1, X = x_k)$ and note that $f_k \in [0, p_k]$ where $p_k = \mathbb{P}(X = x_k)$.

We can write the above optimization problem as

$$\max_{(f_1,\ldots,f_K)\geq \mathbf{0}} \sum_{k=1}^{K} f_k \;\text{ s.t. }\; f_k \leq p_k \text{ for } k \in \{1, \ldots, K\} \;\text{ and }\; \sum_{k=1}^{K} (\tau_0(x_k) - \mu_0) f_k = 0,$$

a finite-dimensional linear program. This program has a feasible solution if $\tau_0(x_k) - \mu_0$ 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.

Algorithm 4.1 (Internal validity for fixed τ_0)**.** Without loss of generality, let $\tau_0(x_1) \leq \ldots \leq \tau_0(x_K)$.

1. Set $(f_1, \ldots, f_K) = (p_1, \ldots, p_K)$.
2. If $\sum_{k=1}^{K} (\tau_0(x_k) - \mu_0) f_k = 0$, end the algorithm and report $\sum_{k=1}^{K} f_k$.
3. If $\sum_{k=1}^{K} (\tau_0(x_k) - \mu_0) f_k \neq 0$:
   (a) If $\sum_{k=1}^{K} (\tau_0(x_k) - \mu_0) f_k > 0$, let $k^* = \max\{k \in \{1, \ldots, K\} : f_k = p_k\}$ and set $f_{k^*} = \max\left\{0, -\sum_{k=1}^{k^*-1} (\tau_0(x_k) - \mu_0) p_k / (\tau_0(x_{k^*}) - \mu_0)\right\}$.
   (b) If $\sum_{k=1}^{K} (\tau_0(x_k) - \mu_0) f_k < 0$, let $k^* = \min\{k \in \{1, \ldots, K\} : f_k = p_k\}$ and set $f_{k^*} = \max\left\{0, -\sum_{k=k^*+1}^{K} (\tau_0(x_k) - \mu_0) p_k / (\tau_0(x_{k^*}) - \mu_0)\right\}$.
4. Go to step 2.

When μ_0 exceeds $\mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, this algorithm reduces the weights associated with smallest CATEs until μ_0 equals $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$ for some subpopulation. When $\mu_0 < \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, the same procedure is instead applied to the largest 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 is infinite, which generates difficulties in implementation. However, we show this program has an analytical solution even when X's components are allowed to be discrete, continuous, and mixed, as is often the case in practice.

Theorem 4.2. Let $\mu(a, \tau_0)$ be an estimand satisfying equation (1.1). Suppose Assumption 3.1 holds. If $\mu_0 \notin \mathcal{S}(\tau_0; W_0)$, then $\overline{P}(a, W_0; \{\tau_0\}) = 0$. If $\mu_0 \in \mathcal{S}(\tau_0; W_0)$, then

$$\overline{P}(a, W_0; \{\tau_0\}) = \begin{cases} \mathbb{P}(T_\mu \leq \alpha^+ \mid W_0 = 1) - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha^+) | W_0 = 1]}{\alpha^+} & \text{if } \mu_0 < E_0 \\ \mathbb{P}(T_\mu \geq \alpha^- \mid W_0 = 1) - \frac{\mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha^-) | W_0 = 1]}{\alpha^-} & \text{if } \mu_0 > E_0 \\ 1 & \text{if } \mu_0 = E_0, \end{cases} \tag{4.9}$$

where $T_\mu \coloneqq \tau_0(X) - \mu_0$, $E_0 \coloneqq \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1]$, $\alpha^+ \coloneqq \inf\{\alpha \in \mathbb{R} : \mathbb{E}[T_\mu \mathbb{1}(T_\mu \leq \alpha) \mid W_0 = 1] \geq 0\}$, and $\alpha^- \coloneqq \sup\{\alpha \in \mathbb{R} : \mathbb{E}[T_\mu \mathbb{1}(T_\mu \geq \alpha) \mid W_0 = 1] \leq 0\}$.

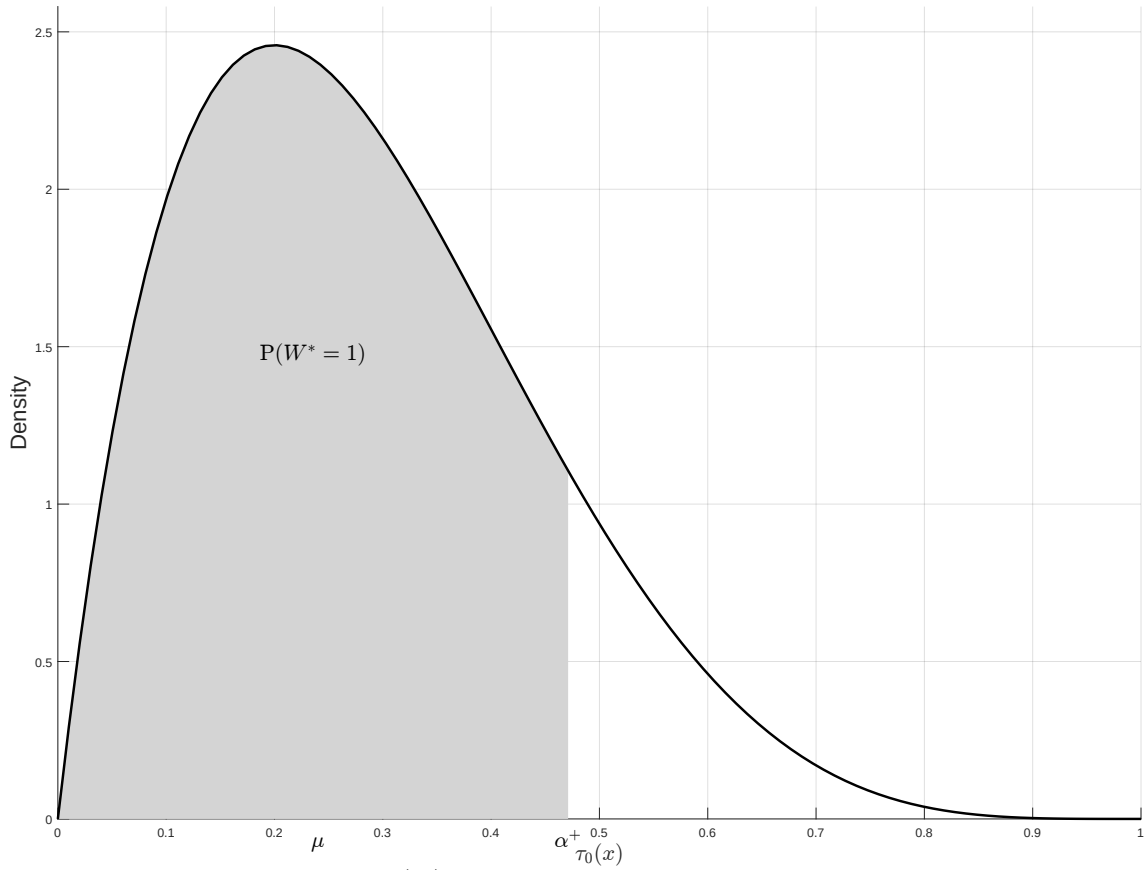
The computation of these bounds can be done using a linear programming algorithm when X is discrete, or through plug-in estimators of the terms in equation (4.9) regardless of the nature of the support of X.

To illustrate this theorem, let $\tau_0(X)$ be continuously distributed with support $[\underline{\tau}, \overline{\tau}]$, and suppose $\mu_0 \in [\underline{\tau}, \overline{\tau}]$. Without loss of generality, assume $E_0 \geq \mu_0$. If $E_0 = \mu_0$, then the estimand is perfectly representative of the population since it equals the average treatment effect over it. In the case where $E_0 > \mu_0$, the estimand is not representative of the entire population. We are searching for the largest subpopulation $\{W^* = 1\}$ such that $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1] = \mu_0$. Initializing W^* at W_0, removing the subpopulation with the largest values of $\tau_0(x)$ yields the steepest decrease in $\mathbb{E}[Y(1) - Y(0) \mid W^* = 1]$. Therefore, $\overline{P}(a, W_0; \{\tau_0\})$ is obtained by removing a subpopulation of the kind $W^-(\alpha) = \mathbb{1}(\tau_0(X) > \alpha) \cdot W_0$ for a given threshold α. This threshold is determined by the constraint

$$\mathbb{E}[\tau_0(X) \mid \tau_0(X) \leq \alpha] = \mu_0. \tag{4.10}$$

Thus, the remaining subpopulation W^* corresponds to $W^* = \mathbb{1}(\tau_0(X) \leq \alpha^*) \cdot W_0$ where α^* is the unique solution to (4.10). 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.

Figure 2: Characterizing a Representative Subpopulation When τ_0 Is Known



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

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 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.

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 $a_{\max} = \sup(\text{supp}(a(X) \mid W_0 = 1)) = \sup_{x \in \text{supp}(X \mid W_0 = 1)} a(x)$ in this section. This condition is satisfied when $a(\cdot)$ is

continuous or when X has finite support, among other cases. We also note that our assumption $a_{\max} < \infty$ holds trivially in every case considered below.

5.1 Unconfoundedness

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

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

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 $U \perp\!\!\!\perp (Y(1), Y(0), X)$ and $U \sim \text{Unif}(0,1)$. This is a subpopulation where units with 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 contained in $\{b, 1-b\}$ for some $b \in (0,1)$. 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 β_{OLS}, bounds on the support of $(Y(1), Y(0))$, and our measure of internal validity $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$. Following the expression in (4.4), bounds on the ATE are given by

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

Estimating these bounds requires the estimation of one additional quantity beyond the OLS estimand, which is the expectation of $\text{var}(D \mid X)$. The width of these bounds depends crucially on $B_u - B_\ell$, 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 consider an alternative representation of the estimand:

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

Applying Theorem 4.1 yields that

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

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 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. $\overline{P}(1-p, D; \mathcal{T}_{\text{all}})$ 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, the bound satisfies

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

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

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

where α^* satisfies $\mathbb{E}[\tau_0(X) \mid \tau_0(X) \leq \alpha^*] = \beta_{\text{OLS}}$. 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 $\mathbb{E}[\tau_0(X) \mid \tau_0(X) \geq \alpha]$ is large for large α.

5.2 Instrumental Variables

Now consider a 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)**.** Almost surely, the following hold:

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: $\mathbb{P}(Y(d, 0) = Y(d, 1) \mid X) = 1$ for $d \in \{0, 1\}$;
3. First stage: $\mathbb{P}(Z = 1 \mid X) \in (0, 1)$ and $\mathbb{P}(D(1) = 1 \mid X) \neq \mathbb{P}(D(0) = 1 \mid X)$;
4. Strong monotonicity: $\mathbb{P}(D(1) \geq D(0) \mid X) = 1$.

The first instrumental variables estimand we consider was originally studied by Angrist and Imbens (1995). In addition to Assumption 5.1, suppose that the model for X is saturated, with K possible combinations of covariate values, i.e.,

let $\operatorname{supp}(X) = \{x_1, \ldots, x_K\}$. 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 given by

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

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

Proposition 5.1. Suppose Assumption 5.1 holds. Suppose X is discrete with finite support. Then

$$\beta_{\text{2SLS}} = \frac{\mathbb{E}\left[\left|\operatorname{cov}(D, Z \mid X)\right| \cdot \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0), X] \mid D(1) > D(0)\right]}{\mathbb{E}[\left|\operatorname{cov}(D, Z \mid X)\right| \mid D(1) > D(0)]}.$$

Thus, β_{2SLS} satisfies the representation in (1.1) with $a_{\text{2SLS}}(X) = \left|\operatorname{cov}(D, Z \mid X)\right|$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0), X]$, and $W_0 = \mathbb{1}(D(1) > D(0))$. Note that $\beta_{\text{2SLS}} = \text{LATE} \coloneqq \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0)]$ if and only if $a_{\text{2SLS}}(X)$ is uncorrelated with $\tau_0(X)$ given $D(1) > D(0)$.

The practical limitation of focusing on β_{2SLS} is that applied researchers rarely create multiple interacted instruments (cf. Blandhol, Bonney, Mogstad, and Torgovitsky, 2022), which is how Z_S is defined and used 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 $W = (D, X)$. 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 5.2 (Rich covariates)**.** $\mathbb{P}(Z = 1 \mid X)$ is linear in X.

Under the instrument validity assumption and the rich covariates assumption, Słoczyński (2020) obtains the following representation of the "noninteracted" IV estimand.

Proposition 5.2. Suppose Assumptions 5.1 and 5.2 hold. Then

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

It follows that β_{IV} is a weighted estimand satisfying (1.1) with weights $a_{\text{IV}}(X) = \text{var}(Z \mid X)$, CATE function $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0), X]$, and where the average is again taken over the complier subpopulation, i.e., $W_0 = \mathbb{1}(D(1) > D(0))$.

Causal Representation and Internal Validity of 2SLS

First consider the estimand β_{2SLS}, which can be characterized as $\mu(a_{\text{2SLS}}, \tau_0)$. Since $a_{\text{2SLS}}(X) \geq 0$, there exists a subpopulation of $\{D(1) > D(0)\}$ such that β_{2SLS} is an average treatment effect over that subpopulation. The maximum size of that subpopulation is given by

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

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

Under the representation in Proposition 5.2, the IV estimand has the same W_0, but has $a_{\text{IV}}(X) = \text{var}(Z \mid X)$ instead. Here, $a_{\text{IV}}(X) \geq 0$ and

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

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 $\mathbb{P}(D(1) > D(0) \mid X)$ 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 $a_{\text{2SLS}}(X) = |\operatorname{cov}(D, Z \mid X)|$ is constant but $\mathbb{P}(D(1) > D(0) \mid X)$ is not. They are equally representative when $\mathbb{P}(D(1) > D(0) \mid X)$

is constant. In this case, the estimands are equal, so this is not unexpected.

5.3 Difference-in-Differences

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. By partitioned regression results, the coefficient on treatment indicator is

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

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 5.3 (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 β_{TWFE} as a weighted average.

Proposition 5.3. Suppose Assumption 5.3 holds. Then

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

where $a_{\text{TWFE}}(g,t) = 1 - \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s \mid G = g] - \mathbb{E}[D_t] + \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[D_s]$.

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

$$\beta_{\text{TWFE}} = \frac{\mathbb{E}[a_{\text{TWFE}}(X) \cdot \tau_0(X) \mid D = 1]}{\mathbb{E}[a_{\text{TWFE}}(X) \mid D = 1]},$$

where $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D = 1, G, P]$, $a_{\text{TWFE}}(X)$ is not generally nonnegative, and the average is taken over the treated units, i.e., $W_0 = D$.[3] A nonnegative weight function can be obtained under the assumption that $\tau_0(X)$ is 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 TWFE 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 5.3 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

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

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 $X = G$, $W_0 = D$, $\tau_0(X) = \mathbb{E}[Y(1) - Y(0) \mid D = 1, G]$, and the weight function $a_{\text{TWFE,H}}(G) \geq 0$. This weight function is derived in Appendix E, with additional comparisons with the weights in Goodman-Bacon (2021) in Appendix H.

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

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

representation for β_{TWFE} uniformly in $\tau_0 \in \mathcal{T}_{\text{all}}$.[4] 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

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

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, calculations yield that

$$\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}}) = \min \left\{ \frac{2\mathbb{P}(G = 2) + \omega \mathbb{P}(G = 3)}{2\mathbb{P}(G = 2) + \mathbb{P}(G = 3)}, \frac{\omega^{-1} 2\mathbb{P}(G = 2) + \mathbb{P}(G = 3)}{2\mathbb{P}(G = 2) + \mathbb{P}(G = 3)} \right\},$$

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, β_{TWFE} is perfectly representative of the ATT if and only if $\omega = 1$, which occurs if and only if $\mathbb{P}(G = 3) = 2/3$. Thus, $\overline{P}(a_{\text{TWFE,H}}, D; \mathcal{T}_{\text{all}}) = 1$ when $\mathbb{P}(G = 3) = 2/3$ and the internal validity of β_{TWFE} declines as $|\mathbb{P}(G = 3) - 2/3|$ increases. This is due to the weight function $a_{\text{TWFE,H}}(g)$ being constant in g if and only if $\mathbb{P}(G = 3) = 2/3$. As before, constant weights imply that the weighted estimand equals the average treatment effect over $\{W_0 = 1\}$.

6 Estimation and Inference

We now consider estimation and inference for our measures of internal validity and representativeness. We focus our attention on the case when $\mathcal{T} = \mathcal{T}_{\text{all}}$ and briefly discuss the case where $\mathcal{T} = \{\tau_0\}$ in Appendix C.

To measure internal validity, we seek to estimate

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

Suppose we observe a random sample of size n, $\{(W_i, X_i)\}_{i=1}^n$, where W_i is a set of

[4] In the context of Proposition 5.4, τ_0 is a function of G only, thus $\mathcal{T}_{\text{all}}$ denotes the set of all functions of G with finite second moments. Note that this is a strict subset of all "time-heterogeneous" conditional average treatment effects, $\mathbb{E}[Y(1) - Y(0) \mid D = 1, G, P]$.

variables needed to estimate $a(\cdot)$ 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 consider the following analog estimator of $\overline{P}(a, W_0; \mathcal{T}_{\text{all}})$:

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

Here c_n is a tuning parameter that converges to 0 as n diverges. We 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 $\widehat{a}(\cdot)$ and $\widehat{w}_0(\cdot)$. Estimating $a_{\max} = \sup(\text{supp}(a(X) \mid W_0 = 1))$ is more delicate. In some of our examples, this supremum is known or can be bounded above without using data. For example, the OLS estimand under unconfoundedness has weights $a(X) = \text{var}(D \mid X)$ which are bounded above by $1/4$. If X is continuously distributed, then $1/2$ may lie in the support of $p(X)$, and thus we may avoid the estimation of $a_{\max}$. The IV estimand of Section 5 has weights $a_{\text{IV}}(X) = \text{var}(Z \mid X)$ which are similarly bounded above by $1/4$. Similarly, $a_{\text{2SLS}}(X) = |\text{cov}(D, Z \mid X)| \leq 1/4$ by 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 $\sup_{x \in \text{supp}(X|W_0=1)} \widehat{a}(x)$ will be consistent for $a_{\max}$ 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 $a(\cdot)$ satisfy this requirement.

In Appendix C, we prove the consistency of $\widehat{\overline{P}}$ and derive its limiting distribution. We also provide a step-by-step bootstrap algorithm that can be employed to conduct inference and prove its validity. This bootstrap approach is based on Fang and Santos (2019) and is nonstandard, but yields valid inferences even when $a_{\max}$ is estimated, as opposed to standard bootstrap approaches, such as the empirical bootstrap.

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 D.C.) reformed their law by

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

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, as computed by the National Center for Health Statistics. The treatment is whether the state allowed unilateral divorce in a given year. Following Goodman-Bacon (2021), our sample omits Alaska and Hawaii. We also omit eight further states which 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 standard assumptions. Using the approach of Callaway and Sant'Anna (2021), we obtain an estimate of –10.220 with a p-value of 0.001. The approach of Wooldridge (2025) produces an estimate of –5.530 and a p-value of 0.138. 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 $\mathbb{P}(W^* = 1 \mid D = 1)$, 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 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

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

A. Estimates of the effects of unilateral divorce laws		
TWFE	ATT	
	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 data are a panel of the 1964–1996 U.S. The outcome is the state- and year-specific female suicide rate (per million women). The treatment is whether the state allowed unilateral divorce in a given year. The sample includes D.C. but excludes the states excluded by Goodman-Bacon (2021) as well as eight additional always-treated states. The measures of internal validity "given τ_0" require an estimate of the CATE function, which we obtain using the approach of Wooldridge (2025).

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-time 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 (2025), and other recent papers, each of which explicitly targets the ATT.

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 over a subpopulation. 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 our main results, we derived the sharp upper bound on the size of that subpopulation under different assumptions on treatment effect heterogeneity, which offers a practical tool to quantify the internal validity of weighted estimands.

References

Angrist, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 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, 431–442.

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

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

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, 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, 200–230.

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

de Chaisemartin, C. (2012): "All You Need Is LATE," working paper.

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

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

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

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

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

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

Humphreys, M. (2025): "Bounds on the Fixed Effects Estimand in the Presence of Heterogeneous Assignment Propensities," *Journal of Causal Inference*, 13, 20240040.

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

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018): "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113, 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, 1523–1566.

Mogstad, M., and A. Torgovitsky (2024): "Instrumental Variables with Un-

observed 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, 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, 267–288.

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

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

Wooldridge, J. M. (2025): "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators," *Empirical Economics*, forthcoming.

A Proofs of Key Results

This appendix contains the proofs for our key results in Sections 3 and 4. Proofs for other results, including Theorem 4.2, can be found in the Supplemental Appendix.

Proof of Proposition 3.1. To show the first claim, note that the equation $\mathbb{E}[W^*(Y(1)-Y(0)) \mid W_0 = 1, X] = \mathbb{E}[Y(1) - Y(0) \mid W_0 = 1, X] \cdot \underline{w}^*(X)$ holds since $W^* \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$ by Definition 3.1. Since $\underline{w}^*(X) > 0$, we can obtain

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

where the third equality holds from W^* being a subpopulation of W_0.

The proposition's second claim is established below:

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

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

The first equality follows from W^* being a subpopulation of W_0, the second from the law of iterated expectations and $W^* \leq W_0$, and the third from equation (A.1). $\square$

Proof of Theorem 3.1. ($\Longrightarrow$) First, suppose there exists $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{all}})$. By Proposition 3.1, we have that $\mu(a, \tau_0) - \mu(\underline{w}^*, \tau_0) = 0$ for all $\tau_0 \in \mathcal{T}_{\text{all}}$. Let $\tau^*(X) = \frac{a(X)}{\mathbb{E}[a(X)|W_0=1]} - \frac{\underline{w}^*(X)}{\mathbb{P}(W^*=1|W_0=1)}$. We have $\mathbb{E}[a(X)^2] < \infty$ by Assumption 3.1 and $\mathbb{E}[\underline{w}^*(X)^2] \leq 1$, which holds by construction. Hence, $\mathbb{E}[\tau^*(X)^2] < \infty$, and thus $\tau^* \in \mathcal{T}_{\text{all}}$.

Thus, we must have $\mu(a, \tau^*) - \mu(\underline{w}^*, \tau^*) = 0$. Expanding this equality yields $0 = \mu(a, \tau^*) - \mu(\underline{w}^*, \tau^*) = \mathbb{E}[\tau^*(X)^2 \mid W_0 = 1]$, which implies that

$$
1 = \mathbb{P}(\tau^*(X) = 0 \mid W_0 = 1) = \mathbb{P}\left(a(X) = \frac{\underline{w}^*(X)\mathbb{E}[a(X) \mid W_0 = 1]}{\mathbb{P}(W^* = 1 \mid W_0 = 1)} \mid W_0 = 1\right). \tag{A.2}
$$

Note that $\underline{w}^*(X) \geq 0$ and $\mathbb{P}(W^* = 1 \mid W_0 = 1) > 0$ by construction. Also, $\mathbb{E}[a(X) \mid W_0 = 1] > 0$ by Assumption 3.1. Therefore, $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) = 1$.

($\Longleftarrow$) Second, suppose that $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) = 1$. Let $U \sim \text{Unif}(0, 1)$ where $U \perp\!\!\!\perp (Y(1), Y(0), X, W_0)$, and define $W^* = \mathbb{1}\left(U \leq a(X)/a_{\max}\right) \cdot W_0$. We verify W^* is a regular subpopulation of W_0. First we see that $\mathbb{P}(W^* = 1) > 0$ because

$$
\begin{aligned}
\mathbb{P}(W^* = 1) &= \mathbb{E}\left[\mathbb{P}\left(U \leq a(X)/a_{\max} \mid W_0 = 1, X\right) \mid W_0 = 1\right] \cdot \mathbb{P}(W_0 = 1) \\
&= \mathbb{E}\left[a(X)/a_{\max} \mid W_0 = 1\right] \cdot \mathbb{P}(W_0 = 1) > 0.
\end{aligned}
$$

The first equality follows from $\mathbb{P}(W^* = 1 \mid W_0 = 0) = 0$ and $U \perp\!\!\!\perp (X, W_0)$, the second equality from $a(X)/a_{\max} \in [0, 1]$ almost surely given $W_0 = 1$, and the inequality from $\mathbb{P}(W_0 = 1)$ and $\mathbb{E}[a(X) \mid W_0 = 1]$ being positive by assumption, and by $a_{\max} < \infty$. That W^* satisfies the two properties of Definition 3.1 holds immediately. Therefore, W^* is a regular subpopulation of W_0.

Finally, let $\underline{w}^*(X) = \mathbb{P}(W^* = 1 \mid X, W_0 = 1) = a(X)/a_{\max}$. Using Proposition 3.1,

we can see that for a given $\tau_0 \in \mathcal{T}_{\text{all}}$

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

where the last equality follows from the scale invariance of the estimands. Since $\tau_0 \in \mathcal{T}_{\text{all}}$ was arbitrary, we have that $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{all}})$. □

Proof of Theorem 3.2. To simplify the notation in the proof, we let $\mu_0 \coloneqq \mu(a, \tau_0)$.

($\Longrightarrow$) First, suppose $\mu_0 \notin \mathcal{S}(\tau_0; W_0)$ and suppose there exists $W^* \in \mathcal{W}(a; W_0, \{\tau_0\})$. Since $\mu_0 \notin \mathcal{S}(\tau_0; W_0)$, we can without loss of generality suppose that $\mathbb{P}(\tau_0(X) > \mu_0 \mid W_0 = 1) = 1$. Since $W^* \in \mathcal{W}(a; W_0, \{\tau_0\})$, we can write by Proposition 3.1

$$\mu_0 = \mu(\underline{w}^*, \tau_0) \geq \mu(\underline{w}^*, \mu_0) = \mu_0. \tag{A.3}$$

The inequality is strict unless $\mathbb{E}[\underbrace{(\tau_0(X) - \mu_0)}_{>0 \text{ w.p.1}} \underbrace{\underline{w}^*(X)}_{\in[0,1]} \mid W_0 = 1] = 0$ holds. This holds if $\mathbb{P}((\tau_0(X) - \mu_0)\underline{w}^*(X) = 0 \mid W_0 = 1) = 1$, which in turns occurs if and only if $\mathbb{P}(\underline{w}^*(X) = 0 \mid W_0 = 1) = 1$. This implies $\mathbb{P}(W^* = 1 \mid W_0 = 1) = 0$ and $\mathbb{P}(W^* = 1) = \mathbb{P}(W^* = 1 \mid W_0 = 1) \cdot \mathbb{P}(W_0 = 1) = 0$, a contradiction of $W^* \in \mathcal{W}(a, W_0, \{\tau_0\})$. Therefore, the inequality in (A.3) is strict and yields $\mu_0 > \mu_0$, a contradiction. Thus, $\mathcal{W}(a; W_0, \{\tau_0\}) = \emptyset$ when $\mu_0 \notin \mathcal{S}(\tau_0; W_0)$.

($\Longleftarrow$) Second, suppose $\mu_0 \in \mathcal{S}(\tau_0; W_0)$. Let $\mathcal{X}^- = \{x \in \text{supp}(X) : \tau_0(x) \leq \mu_0\}$ and $\mathcal{X}^+ = \{x \in \text{supp}(X) : \tau_0(x) \geq \mu_0\}$. By $\mu_0 \in \mathcal{S}(\tau_0; W_0)$, $\mathbb{P}(X \in \mathcal{X}^- \mid W_0 = 1) > 0$ and $\mathbb{P}(X \in \mathcal{X}^+ \mid W_0 = 1) > 0$.

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

$$W^*(u) = (\mathbb{1}(U > u, X \in \mathcal{X}^-) + \mathbb{1}(U \leq u, X \in \mathcal{X}^+)) \cdot W_0.$$

We show $W^*(u)$ is a regular subpopulation of W_0 for all $u \in [0, 1]$. We can see that $W^*(u) \in \{0, 1\}$, that $W^*(u) \perp\!\!\!\perp (Y(1), Y(0)) \mid X, W_0 = 1$, and that $\mathbb{P}(W_0 = 1 \mid W^*(u) = 1) = 1$. To show that $W^*(u)$ characterizes a regular subpopulation of W_0, we also show that it is nonzero with positive probability:

$$\mathbb{P}(W^*(u) = 1 \mid W_0 = 1) = (1 - u)\mathbb{P}(X \in \mathcal{X}^- \mid W_0 = 1) + u\mathbb{P}(X \in \mathcal{X}^+ \mid W_0 = 1) > 0$$

for all $u \in [0, 1]$, which implies $\mathbb{P}(W^*(u) = 1) > 0$ by $\mathbb{P}(W_0 = 1) > 0$. Hence,

$W^*(u) \in \mathrm{SP}(W_0)$ for all $u \in [0,1]$. For $u \in [0,1]$, we have that $\underline{w}^*(X;u) \coloneqq \mathbb{P}(W^*(u) = 1 \mid X, W_0 = 1) = (1-u)\mathbb{1}(X \in \mathcal{X}^-) + u\mathbb{1}(X \in \mathcal{X}^+)$. Therefore, using Proposition 3.1,

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

By construction, $\tau_0(X)\mathbb{1}(X \in \mathcal{X}^-) \leq \mu_0\mathbb{1}(X \in \mathcal{X}^-)$ and $\tau_0(X)\mathbb{1}(X \in \mathcal{X}^+) \geq \mu_0\mathbb{1}(X \in \mathcal{X}^+)$ almost surely. Therefore, $\mathbb{E}[Y(1) - Y(0) \mid W^*(0) = 1] \leq \mu_0 \leq \mathbb{E}[Y(1) - Y(0) \mid W^*(1) = 1]$. 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 $\mu_0 = \mathbb{E}[Y(1) - Y(0) \mid W^*(u^*) = 1]$ and $W^*(u^*) \in \mathcal{W}(a; W_0, \{\tau_0\})$. □

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

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

for all $d \in \mathbb{R}^{d_X}$, which implies $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]} = \frac{\mathbb{E}[\underline{w}^*(X)X|W_0=1]}{\mathbb{E}[\underline{w}^*(X)|W_0=1]}$. Let $u(x) = \underline{w}^*(x)\mathbb{P}(X = x \mid W_0 = 1)/\mathbb{E}[\underline{w}^*(X) \mid W_0 = 1]$. We have that $\frac{\mathbb{E}[\underline{w}^*(X)X|W_0=1]}{\mathbb{E}[\underline{w}^*(X)|W_0=1]} = \sum_{x \in \mathrm{supp}(X|W_0=1)} xu(x)$, a convex combination of values in $\mathrm{supp}(X \mid W_0 = 1)$ because $u(\cdot) \geq 0$ and $\sum_{x \in \mathrm{supp}(X|W_0=1)} u(x) = 1$. Thus, $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]} \in \mathrm{conv}(\mathrm{supp}(X \mid W_0 = 1))$.

($\Longleftarrow$) Second, suppose that $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]} \in \mathrm{conv}(\mathrm{supp}(X \mid W_0 = 1))$. By convexity, we can write $\frac{\mathbb{E}[a(X)X|W_0=1]}{\mathbb{E}[a(X)|W_0=1]}$ as $\sum_{x \in \mathrm{supp}(X|W_0=1)} xu(x)$ for some $u(\cdot) \geq 0$ satisfying $\sum_{x \in \mathrm{supp}(X|W_0=1)} u(x) = 1$. Let $W^* = \mathbb{1}\left(U \leq u(X)/\max(\mathrm{supp}(u(X) \mid W_0 = 1))\right) \cdot W_0$, where $U \sim \mathrm{Unif}(0,1) \perp\!\!\!\perp (X, Y(1), Y(0), W_0)$. Then W^* is a regular subpopulation of W_0 and $\underline{w}^*(X) = \frac{u(X)}{\max(\mathrm{supp}(u(X)|W_0=1))}$ since $\frac{u(X)}{\max(\mathrm{supp}(u(X)|W_0=1))} \in [0,1]$ with probability 1 given $W_0 = 1$. Therefore, for all $\tau_0(x) = c + d'x \in \mathcal{T}_{\text{lin}}$, we have that

$$\begin{aligned}\mu(\underline{w}^*, \tau_0) &= \mu(u/\max(\mathrm{supp}(u(X) \mid W_0 = 1)), \tau_0) \\ &= \mu(u, \tau_0) \\ &= \mathbb{E}\left[u(X)(c + d'X) \mid W_0 = 1\right]/\mathbb{E}\left[u(X) \mid W_0 = 1\right]\end{aligned}$$

$$
\begin{aligned}
&= c + d'\mathbb{E}\left[a(X)X \mid W_0 = 1\right] / \mathbb{E}\left[a(X) \mid W_0 = 1\right] \\
&= \mathbb{E}\left[a(X)(c + d'X) \mid W_0 = 1\right] / \mathbb{E}\left[a(X) \mid W_0 = 1\right] \\
&= \mu(a, \tau_0).
\end{aligned}
$$

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: $K > 0$.
($\Longrightarrow$) 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 $\tau^*(x) = K \cdot \mathbb{1}(a(x) < 0)$, which implies $\tau^* \in \mathcal{T}_{\text{BD}}(K)$. We have that $\mu(a, \tau^*) = K\mathbb{E}[a(X)\mathbb{1}(a(X) < 0) \mid W_0 = 1]/\mathbb{E}[a(X) \mid W_0 = 1]$. $\mathbb{E}[a(X)\mathbb{1}(a(X) < 0) \mid W_0 = 1]$ is strictly negative because it is weakly negative and because $\mathbb{E}[a(X)\mathbb{1}(a(X) < 0) \mid W_0 = 1] = 0$ implies $\mathbb{P}(a(X) = 0 \mid a(X) < 0, W_0 = 1) = 1$, a contradiction of $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) < 1$. Since $K > 0$ and $\mathbb{E}[a(X) \mid W_0 = 1] > 0$, we conclude that $\mu(a, \tau^*) < 0$. However, $\mu(\underline{w}^*, \tau^*) \geq 0$ since $\tau^* \geq 0$. By Proposition 3.1, $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K))$ implies $0 > \mu(a, \tau^*) = \mu(\underline{w}^*, \tau^*) \geq 0$, a contradiction. Therefore $\mathcal{W}(a; W_0, \mathcal{T}_{\text{BD}}(K)) = \emptyset$.

($\Longleftarrow$) 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 any $a(\cdot)$. □

Proof of Theorem 4.1. First, suppose $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) = 1$. From Theorem 3.1, there exists $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{all}})$. From derivations in the proof of Theorem 3.1 (see equation (A.2)) we have that $\mathbb{P}(C \cdot a(X) = \underline{w}^*(X) \mid W_0 = 1) = 1$ for some positive constant $C > 0$. Since $\underline{w}^*(X) \leq 1$, we must have $C \cdot a(X) \leq 1$ almost surely given $W_0 = 1$. Thus C is bounded above by $\inf(\text{supp}(1/a(X) \mid W_0 = 1)) = 1/a_{\max}$, which is

strictly positive by assumption. Therefore,

$$\mathbb{P}(W^* = 1 \mid W_0 = 1) = \mathbb{E}[\underline{w}^*(X) \mid W_0 = 1] = \mathbb{E}[C \cdot a(X) \mid W_0 = 1] \leq \frac{\mathbb{E}\left[a(X) \mid W_0 = 1\right]}{a_{\max}}.$$

This upper bound is sharp because it is attained by setting $W^* = \mathbb{1}\left(U \leq a(X)/a_{\max}\right) \cdot W_0$ and noting that $W^* \in \mathcal{W}(a; W_0, \mathcal{T}_{\text{all}})$ from the proof of Theorem 3.1.

Second, suppose $\mathbb{P}(a(X) \geq 0 \mid W_0 = 1) < 1$. By Theorem 3.1, $\mathcal{W}(a; W_0, \mathcal{T}_{\text{all}}) = \emptyset$ and therefore $\overline{P}(a, W_0; \mathcal{T}_{\text{all}}) = 0$. □