Bounding Treatment Effects by Pooling Limited Information across Observations*

Sokbae Lee‡ Martin Weidner§

November 2021

Abstract

We provide novel bounds on average treatment effects (on the treated) that are valid under an unconfoundedness assumption. Our bounds are designed to be robust in challenging situations, for example, when the conditioning variables take on a large number of different values in the observed sample, or when the overlap condition is violated. This robustness is achieved by only using limited "pooling" of information across observations. Namely, the bounds are constructed as sample averages over functions of the observed outcomes such that the contribution of each outcome only depends on the treatment status of a limited number of observations. No information pooling across observations leads to so-called "Manski bounds", while unlimited information pooling leads to standard inverse propensity score weighting. We explore the intermediate range between these two extremes and provide corresponding inference methods. We show in Monte Carlo experiments and through an empirical application that our bounds are indeed robust and informative in practice.

*We are grateful for useful comments from seminar/conference participants at Cornell, Cowles Foundation, Harvard/MIT, Maryland, UC Davis, Vanderbilt, Virginia, and Wisconsin. This research was supported by the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice (grant numbers RES-589-28-0001, RES-589-28-0002 and ES/P008909/1), and by the European Research Council grants ERC-2014-CoG-646917-ROMIA and ERC-2018-CoG-819086-PANEDA.

‡Department of Economics, Columbia University. Email: sl3841@columbia.edu.

§Dept. of Economics & Nuffield College, Univ. of Oxford. Email: martin.weidner@economics.ox.ac.uk.

1 Introduction

In many applications, causal inference hinges on strong ignorability, namely unconfoundedness and overlap (see, e.g., Imbens and Rubin, 2015, for a recent monograph). The former condition is non-testable but requires that all confounders be used as covariates; the latter is a testable condition that may not be satisfied in practice.

The overlap condition has received increasing attention in the literature. In applications, it is not uncommon to have a situation where the estimated propensity scores are close to zero or one. This problem is referred to as limited overlap (e.g., Crump, Hotz, Imbens and Mitnik, 2009). The existence of limited overlap may change the asymptotic behavior of the estimators (e.g., Khan and Tamer, 2010; Hong, Leung and Li, 2019) and may necessitate using a more robust inference method (e.g., Rothe, 2017; Sasaki and Ura, 2021). D'Amour, Ding, Feller, Lei and Sekhon (2021) provide a cautionary tale on the overlap condition when high-dimensional covariates are adopted to make unconfoundedness more plausible.

There are several approaches in the literature to estimate treatment effects when facing limited overlap. Arguably, the most popular method is to focus on a subpopulation where the overlap condition holds (e.g., Crump, Hotz, Imbens and Mitnik, 2009; Yang and Ding, 2018). For example, Crump, Hotz, Imbens and Mitnik (2009) recommend a simple rule of thumb to drop all observations with estimated propensity scores outside the range $[\alpha, 1-\alpha]$ for some predetermined constant α, say $\alpha = 0.1$. Alternatively, Li, Morgan and Zaslavsky (2018) advocate the use of the so-called 'overlap weights' to define the average treatment effect. This amounts to assigning weights equal to one minus the propensity score for the treated units and equal to the propensity score for the control units. If the treatment effects are heterogeneous, both trimming and overlap weighting change the parameter of interest from the population average treatment effect. Without changing it, Nethery, Mealli and Dominici (2019) develop a Bayesian framework by extrapolating estimates from the overlap region to the non-overlap region via a spline model. However, identification by extrapolation is subject to model misspecification.

In this paper, we start with the observation that none of the aforementioned papers would work well if the overlap condition is not satisfied at the population level and it is a priori unknown where it fails. In that case, the population average treatment effect is not point-identified and one may resort to Manski (1989, 1990)'s bounds, provided that the support of outcome is bounded and known. However, it may not yield tight bounds if unconfoundess assumption is plauisible, while the overlap condition being the only source of identification failure. This paper provides a systematic method to explore this possibility.

Our contributions are two-fold. First, we provide novel population bounds on average treatment effects (on the treated) that are valid under an unconfoundedness assumption. Our bounds are applicable if the conditioning variables do not satisfy the overlap condition and take on a large number of different values in the observed sample. This robustness is achieved by only using limited "pooling" of information across observations. Namely, the bounds are constructed as the expectations of functions of the observed outcomes such that the contribution of each outcome only depends on the treatment status of a limited number of observations. No information pooling across observations leads to Manski (1989, 1990)'s bounds, while unlimited information pooling leads to standard inverse propensity score weighting. We explore the intermediate range between these two extremes by considering the setup where an applied researcher provides a reference propensity score. Our bounds are valid independent of the value of this reference propensity score, but if it happens to be close to the true propensity score, then our bounds are optimal in terms of expected width within the class of limited pooling bounds considered in this paper. The reference propensity score is therefore crucial to construct our novel treatment effect bounds uniquely, and it also allows to incorporate prior knowledge on the propensity score in a robust way.

Second, we develop estimation and inference methods for the population bounds we have established under the unconfoundedness assumption. A leading data scenario we analyze assumes that the observed covariates take on many different values, thereby implying that for each possible covariate value only a small number of individuals with similar value are observed. In this scenario, it is a statistically challenging problem to provide a valid confidence interval for the treatment effects, which we tackle in this paper.

An alternative approach to robust inference for treatment effects under unconfoundedness is provided by Armstrong and Kolesár (2021). In particular, their confidence intervals are asymptotically valid under a violation of the overlap condition, as long as the researcher specifies a Lipschitz bound on the conditional mean of the outcome variable. Their approach is distinct from and complementary to ours. The approach of Armstrong and Kolesár (2021) reduces to a matching estimator for the average treatment effect (e.g., Abadie and Imbens 2006, 2008, 2011) if the Lipschitz bound is chosen to be very large. Those matching estimators crucially require that for every observation we can find other observations with similar covariate values but opposite treatment status. This is not required in our approach. Crucially, we only pool information across observations with similar covariate values, but in contrast to Armstrong and Kolesár (2021) and matching estimators, we do so completely independent of the treatment status of the observations involved. This is the key difference

compared to those existing methods.

The remainder of the paper is organized as follows. In Section 2, we describe the setup and intuition behind our approach. In Section 3, we derive our novel bounds in a systematic way at the population level, and afterwards construct sample analogs in Section 4. Using those bounds we then provide asymptotically valid confidence intervals. We discuss how to cluster the covariate observations in Section 5. The results of Monte Carlo experiments are reported in Section 6. In Section 7, we use a well known dataset from Connors et al. (1996)'s study of the efficacy of right heart catheterization (RHC) to illustrate the practical usefulness of our approach. This dataset has been analyzed in the context of limited overlap in the literature (see, for example, Crump, Hotz, Imbens and Mitnik (2009), Rothe (2017), and Li, Morgan and Zaslavsky (2018) among others). We show in Monte Carlo experiments and through an empirical application that our bounds are indeed robust and informative in practice. Appendices includes all the proofs and technical derivations omitted from the main text.

2 Setup and Basic Bounds

We have treatment status $D \in \{0,1\}$, potential outcomes $Y(0), Y(1) \in \mathbb{R}$, and regressors $X \in \mathcal{X}$, where $\mathcal{X}$ is the range of X.[1] These are all random variables. The observed outcome is $Y = (1-D)Y(0) + D\,Y(1) \in \mathbb{R}$. Our main goal is to develop inference procedures for the average treatment effect (ATE) and the average treatment effect on the treated (ATT),

$$\text{ATE} := \mathbb{E}\left[Y(1) - Y(0)\right], \qquad \text{ATT} := \mathbb{E}\left[Y(1) - Y(0) \middle| D = 1\right].$$

It is also convenient to define

$$\tau(x) := \mathbb{E}\left[Y(1) - Y(0) \,\middle|\, X = x\right], \qquad \pi(x) := \frac{\mathbb{E}\left[D\,Y(1) - D\,Y(0) \,\middle|\, X = x\right]}{\mathbb{E}\left(D\right)}. \tag{1}$$

$\tau(x)$ is simply the conditional average treatment effect, and by the law of iterated expectations we have $\text{ATE} = \mathbb{E}[\tau(X)]$. Analogously we have $\text{ATT} = \mathbb{E}[\pi(X)]$. However, $\pi(x)$ is *not* the conditional ATT, but rather the contribution to ATT from $X = x$.[2]

The main assumptions that are maintained throughout this paper are the following.

[1] We assume throughout that expectations conditional on $X = x$ are well-defined for every $x \in \mathcal{X}$. This is why $\mathcal{X}$ is not the domain but the range of X.

[2] In this paper we focus on inference on ATE and ATT, but all our results immediately generalize to weighted treatment effects of the form $\mathbb{E}[h(X)\tau(X)]$ and $\mathbb{E}[h(X)\pi(x)]$, with known $h(x)$. For example, one can obtain treatment effects for specific target populations of interest by choosing the "tilting function" $h(x)$ appropriately.

Assumption 1.

(i) $Y(0), Y(1) \perp D \,\big|\, X$. *(unconfoundedness)*

(ii) There are known constants $a_{\min}, a_{\max} \in \mathbb{R}$ *such that* $a_{\min} \leq Y(d) \leq a_{\max}$, $d \in \{0,1\}$.

(iii) $\mathbb{E}(D) > 0$.

(iv) $(D_i, Y_i(0), Y_i(1), X_i)$ *are i.i.d. draws from the population distribution. We observe* D_i, $Y_i = Y_i(D_i)$, *and* X_i, *for* $i = 1, \ldots, n$.

Assumptions 1(i), (ii), (iii) are conditions on the population distribution. We usually do not write indices i for statements about the population. Assumptions 1(iv) specifies the sampling scheme, where sampling units are indicated by subscripts $i = 1, \ldots, n$. For most of our results we could replace the unconfoundedness assumption (i) by the weaker mean independence assumption $\mathbb{E}\left[Y(d) \,\middle|\, D,\, X\right] = \mathbb{E}\left[Y(d) \,\middle|\, X\right]$, for $d \in \{0,1\}$, but in practice, a convincing argument for mean independence usually also implies conditional independence.

Let

$$p(x) := \mathbb{E}\left(D \,\middle|\, X = x\right)$$

be the propensity score. Notice that our assumptions do *not* impose the overlap condition $0 < p(x) < 1$. All the treatment effects bounds derived in this paper are also valid if $p(x) = 0$ and $p(x) = 1$ for some $x \in \mathcal{X}$. Assumption 1(iii) rules out that $p(x) = 0$ for all $x \in \mathcal{X}$, because otherwise ATT is not well-defined.[3]

Note that ATE and ATT may not be point-identified under Assumption 1(i)-(iii), thereby implying that the standard estimators of ATE and ATT that require point-identification can be inconsistent. We will develop the bounds that are valid and immune to this kind of point-identification failure.

Let $\overline{\mathcal{P}}$ be the population distribution of $(Y(0), Y(1), X)$ such that every value of $X \in \mathcal{X}$ has positive mass or density. Let $p : \mathcal{X} \to [0,1]$ be any chosen propensity score. Then, there exists a population distribution $\mathcal{P}$ for $(D, Y(0), Y(1), X)$ such that $\mathbb{E}\left(D \,\middle|\, X\right) = p(X)$, and $(Y(0), Y(1), X)$ has distribution $\overline{\mathcal{P}}$, and Assumptions 1(i) is satisfied. One can construct $\mathcal{P}$ by drawing D, conditional on $Y(0), Y(1), X$, from a Bernoulli distribution with mean $p(X)$. Furthermore, if $\overline{\mathcal{P}}$ satisfies Assumption 1(ii), and p satisfies $\mathbb{E}_{\overline{\mathcal{P}}} p(X) > 0$, then $\mathcal{P}$ satisfies Assumption 1(i), (ii) and (iii).

[3]Our inference results on the ATE do not require Assumption 1(iii), but we may often impose all of Assumption 1 without making that distinction, both for notational convenience and because the case $\mathbb{E}(D) = 0$ is not very interesting anyway.

The reason for defining $\mathcal{P}$ formally in the last paragraph is that on multiple occasions in the paper we want to make probabilistic statements for specific chosen values of the propensity score $p : \mathcal{X} \to [0,1]$. We then always think of $\overline{\mathcal{P}}$ to be fixed, implying that the population distribution $\mathcal{P} = \mathcal{P}(p, \overline{\mathcal{P}})$ is uniquely determined by specifying p, and we write $\mathbb{E}_p$ for the expectation over $\mathcal{P}$. If the expectation is conditional on $X = x$, then we write $\mathbb{E}_{p(x)}$ for $\mathbb{E}_p$. However, most of the time it will not be necessary to be explicit about the value of p, and we then simply write $\mathbb{E}$ for the expected value.

2.1 Manski bounds

Assumptions 1(ii) imposes the outcomes to be bounded by the known constants $a_{\min}$ and $a_{\max}$. Using this one finds, for example, $Y(1) = (1-D)Y(1) + DY(1) \leq (1-D)\, a_{\max} + D\, Y$, where we used that $DY(1) = D\, Y$. In the same way one finds that

$$\begin{aligned} D\, a_{\min} + (1-D)\, Y \leq Y(0) \leq D\, a_{\max} + (1-D)\, Y, \\ (1-D)\, a_{\min} + D\, Y \leq Y(1) \leq (1-D)\, a_{\max} + D\, Y, \end{aligned} \tag{2}$$

which are upper and lower bounds on the unobserved potential outcomes in terms of observables. Thus, if we define

$$B^{(1)}(0,a) := a + (1-D)\,(Y-a), \qquad\qquad B^{(1)}(1,a) := a + D\,(Y-a),$$

then the inequalities in (2) can be written as $B^{(1)}(d, a_{\min}) \leq Y(d) \leq B^{(1)}(d, a_{\max})$, for $d \in \{0,1\}$. Combining those bounds for $d = 1$ and $d = 0$, and taking expectations gives

$$\mathbb{E}\left[B^{(1)}(1,a_{\min}) - B^{(1)}(0,a_{\max})\right] \leq \text{ ATE } \leq \mathbb{E}\left[B^{(1)}(1,a_{\max}) - B^{(1)}(0,a_{\min})\right]. \tag{3}$$

Similarly, if we define $C^{(1)}(a) := D\,(Y-a)$, then we have

$$\frac{\mathbb{E}\left[C^{(1)}(a_{\max})\right]}{\mathbb{E}\,(D)} \leq \text{ ATT } \leq \frac{\mathbb{E}\left[C^{(1)}(a_{\min})\right]}{\mathbb{E}\,(D)}. \tag{4}$$

The bounds in (3) and (4) are well-known, and we will denote those bounds on ATE and ATT as either Manski bounds (Manski 1989, 1990) or as first-order bounds, as indicated by the superscripts on $B^{(1)}$ and $C^{(1)}$.

2.2 Pooling information across observations

If we are unwilling to impose any additional conditions beyond Assumptions 1(ii), then the bounds in (3) and (4) are sharp. For example, the lower treatment effect bounds are sharp, because we could have $Y(1) = a_{\min}$ whenever $D = 0$ and $Y(0) = a_{\max}$ whenever $D = 1$.

However, if we additionally impose the unconfoundedness in Assumption 1(i) together with the overlap condition $0 < p(x) < 1$, for all $x \in \mathcal{X}$, then we have[4]

$$\text{ATE} = \mathbb{E}\left[\frac{D\,Y}{p(X)} - \frac{(1-D)\,Y}{1-p(X)}\right], \qquad \text{ATT} = \frac{1}{\mathbb{E}\,(D)}\,\mathbb{E}\left[D\,Y - \frac{p(X)\,(1-D)\,Y}{1-p(X)}\right]. \tag{5}$$

Thus, if the propensity score $p(x)$ can be point-identified, then both ATE and ATT are also point-identified. One possibility is to identify $p(x)$ by assuming a correctly specified parametric model for $p(x)$, but that is an additional strong assumption that we do not want to impose in this paper. The other possibility is to identify $p(x) = \mathbb{E}\left(D \,\middle|\, X = x\right)$ non-parametrically, but for that to work in practice it requires that for each value of $x \in \mathcal{X}$ we have many observations with X_i close to x, which for a finite sample size n is often not the case. Furthermore, if the overlap condition $0 < p(x) < 1$ is violated, then ATE and ATT are not point-identified, and if $p(x)$ takes on values very close to zero or one, then ATE and ATT are only weakly identified, implying that inference based on (5) will result in very noisy estimates.

For the discussion in this paper the Manski bounds in (3) and (4) represent one extreme for inference on treatment effects, where we do not pool any information across observations i. For example, the sample analog of the upper bound in (3) is given by $\frac{1}{n}\sum_i \left[B_i^{(1)}(1, a_{\max}) - B_i^{(1)}(0, a_{\min})\right]$, where the contribution $B_i^{(1)}(d,a) = a + \mathbb{1}\{D = d\}\,(Y_i - a)$ from each observation i is completely independent from the data of any other observations. The data requirements for validity of this inference approach are very weak, but the resulting bounds often are wide.

The other extreme for our discussion is trying to achieve point-identification based on (5), but that requires pooling a lot of information across observations in order to consistently estimate the propensity score. For example, if we choose a kernel estimator for $p(x)$ with kernel function $k_{ij} = k(\|X_i - X_j\|)$ and Euclidean norm $\|\cdot\|$, then we obtain an estimator $\widehat{\text{ATE}} = \frac{1}{n}\sum_i \left\{[\widehat{p}(X_i)]^{-1} D_i Y_i - [1-\widehat{p}(X_i)]^{-1}(1-D_i)Y_i\right\}$, with $\widehat{p}(X_i) = \left[\sum_j k_{ij} D_i\right] / \left[\sum_j k_{ij}\right]$. Here, the contribution to $\widehat{\text{ATE}}$ from each observation i necessarily depends on a large number of other observations j, because otherwise $\widehat{p}(X_i)$ cannot be consistent for $p(X_i)$. The data requirements for validity of this approach are quite strong, and the curse of dimensionality kicks in quickly as the dimension of X_i gets large.

The goal of this paper is to explore a balanced approach between these two extremes, where we pool some information across observations in order to obtain bounds on the average

[4]The formula for ATT in (5) may be lesser known. Notice that under Assumption 1(i) we have $\mathbb{E}\left\{[1-p(X)]^{-1} p(X)\,(1-D)\,Y \,\middle|\, X\right\} = p(X)\,\mathbb{E}\left[Y(0) \,\middle|\, X\right] = \mathbb{E}\left[D\,Y(0) \,\middle|\, X\right]$.

treatment effects, but using much less pooling than is required for consistent non-parametric point-estimation of $p(x)$. Consider, for example, the case where for two observations $i \neq j$ we have $X_i = X_j$, and define

$$\widetilde{B}^{(2)}_{ij}(1,a) := a + (2 - D_j)\, D_i\, (Y_i - a), \tag{6}$$

which gives

$$\overline{B}^{(2)}_{ij}(1,a) := \frac{1}{2}\left[\widetilde{B}^{(2)}_{ij}(1,a) + \widetilde{B}^{(2)}_{ji}(1,a)\right] = \begin{cases} a & \text{if } (D_i, D_j) = (0,0), \\ Y_j & \text{if } (D_i, D_j) = (0,1), \\ Y_i & \text{if } (D_i, D_j) = (1,0), \\ \frac{1}{2}[Y_i + Y_j] & \text{if } (D_i, D_j) = (1,1). \end{cases} \tag{7}$$

This last expression is a very natural generalization of the Manski bounds for $Y(1)$ in (2). Here, we only use the worst-case bounds a (which will be set to either $a_{\min}$ or $a_{\max}$) if both outcomes D_i and D_j are equal to zero, while in (2) we have to use the worst-case bounds whenever either $D_i = 0$ or $D_j = 0$. It is easy to see that, under Assumptions 1, we have

$$\mathbb{E}\left[\widetilde{B}^{(2)}_{ij}(1, a_{\min}) \,\middle|\, X_i = X_j = x\right] \leq \mathbb{E}\left[Y(1) \,\middle|\, X = x\right] \leq \mathbb{E}\left[\widetilde{B}^{(2)}_{ij}(1, a_{\max}) \,\middle|\, X_i = X_j = x\right]. \tag{8}$$

From (7) we see that this example is very closely related to matching estimators, where outcomes with $D_i \neq D_j$ and $X_i = X_j$ (or $X_i \approx X_j$) are matched mutually to obtain counterfactual outcomes. The key difference is that we do not impose $D_i \neq D_j$ here and therefore only obtain bounds.

The bounds in (8) are the simplest example for what we call second-order bounds in this paper, by which we mean that information is pooled across *two* observations to construct the bounds — notice that in (6) the treatment status D_j of observation j affects the observation i contribution to the bounds, and vice versa.

Analogous to (8) we find second-order bounds for $\mathbb{E}\left[Y_i(0) \middle| X_i = X_j\right]$ by transforming $D \mapsto 1 - D$ in all the expressions, which then also allows us to construct bounds on the ATE.

We also want to give a simple example for second-order ATT bounds here. For $i \neq j$ with $X_i = X_j$ we define $\widetilde{C}^{(2)}_{ij}(a) := D_i\,(Y_i - a) - D_j\,(1 - D_i)\,(Y_i - a)$, which under Assumptions 1

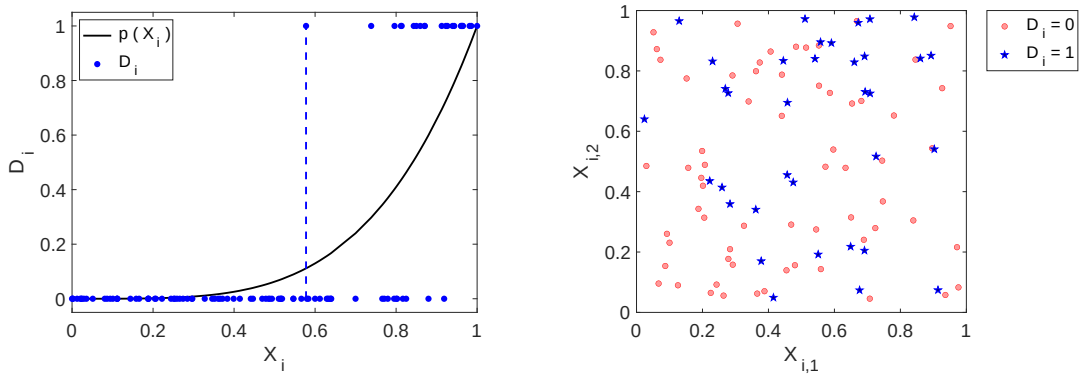

Figure 1: Two simple examples for samples of (X_i, D_i), $i = 1, \ldots, n$, with $n = 100$. For the example on the left we have one-dimensional $X_i \sim U[0,1]$ and $p(x) = x^4$. For the example on the right we have two-dimensional $X_i \sim U[0,1]^2$, and $p(x) = 0.3$.

implies that[5]

$$\frac{\mathbb{E}\left[\widetilde{C}_{ij}^{(2)}(a_{\max}) \,\middle|\, X_i = X_j = x\right]}{\mathbb{E}(D)} \leq \pi(x) \leq \frac{\mathbb{E}\left[\widetilde{C}_{ij}^{(2)}(a_{\min}) \,\middle|\, X_i = X_j = x\right]}{\mathbb{E}(D)}. \tag{9}$$

Section 3 discusses second- and higher-order bounds on ATE and ATT that can be obtained under Assumptions 1 more systematically.

2.3 Lack of overlap and curse of dimensionality

Equation (5) shows that both ATE and ATT are point identified if, in addition to Assumption 1, the overlap condition $0 < p(x) < 1$ holds. However, estimating the conditional conditional expectation $p(x) = \mathbb{E}\left(D \,\middle|\, X = x\right)$ at finite sample can be challenging, and is subject to the curse of dimensionality for multi-dimensional covariates X (see e.g. Stone 1980). To illustrate those finite-sample challenges, consider the two examples in Figure 1, both of which have a sample size of $n = 100$.

For the example on the left-hand side we have drawn $X_i \in (0,1]$ from a uniform distribution and chosen $p(x) = x^4$. Thus, the overlap conditions $0 < p(x) < 1$ is satisfied for all

[5]One calculates

$$\frac{\mathbb{E}\left[\widetilde{C}_{ij}^{(2)}(a) \,\middle|\, X_i = X_j = x\right]}{\mathbb{E}(D)} = \frac{\mathbb{E}\left\{D[Y(1) - a] \,\middle|\, X = x\right\} - p(x)\left[1 - p(x)\right]\mathbb{E}\left[Y(0) - a \,\middle|\, X = x\right]}{\mathbb{E}(D)}$$
$$= \left[1 - p(x)\right]\pi(x) + p(x)\frac{\mathbb{E}\left\{D[Y(1) - a] \,\middle|\, X = x\right\}}{\mathbb{E}(D)},$$

which implies (9), because $D[Y(1) - a] = D(Y - a)$ provides our first-order bounds on ATT.

$x \in (0, 1]$, and ATE is point-identified. However, in the sample, we never observe $D_i = 1$ for values of X_i that are smaller than 0.58 (indicated by the dotted line), and aiming to provide a precise point-estimate for the ATE based on this sample is therefore obviously futile, unless some prior knowledge (e.g. a parametric model of the propensity score) is available. A plausible ATE inference approach in this example would be to use the Manski bounds in (2) for $X_i < 0.58$, and apply some matching or inverse propensity score weighting approach to point-estimate $\tau(X_i)$ for $X_i \geq 0.58$. The resulting confidence intervals for the ATE will be quite robust, and conceptually very similar to Armstrong and Kolesár (2021), except that we have replaced their Lipschitz bound on the conditional mean of Y_i by a worst-case bounds on Y_i.

This first example was quite simple, and the insights from the population analysis (i.e. point identification for $0 < p(x) < 1$, Manski bounds otherwise) could still be usefully employed there. However, if the covariates become multi-dimensional, then it is not equally easy to decide whether we have limited overlap (a propensity score close to zero or one) for any given value of the covariates. A simple illustration of this is the right hand side example in Figure 1, where $X_i \in [0, 1]^2$ is drawn from a uniform distribution and $p(x) = 0.3$ is constant. Thus, in terms of the identification analysis we have good overlap everywhere in this example, and ATE is point-identified. Nevertheless, we have relatively large regions in the covariate space for which not a single observation with $D_i = 1$ (blue stars) is available, for example, the region close to the origin $x = (0, 0)$. From just observing this sample of (X_i, D_i) it is, of course, impossible to know whether this is just a finite sample problem, or whether we truly have a lack of overlap ($p(x) = 0$ for x close to $(0, 0)$), because the total number of observations in those covariate regions with only $D_i = 0$ observations (red dots) is quite small. We have only drawn a two-dimensional example here, but this problem becomes more severe for larger covariate dimensions.

The bounds in this paper are designed to be useful in exactly those situations, where we do not know whether we have a lack of overlap in certain covariate regions or not, because the local sample size is too small. Our bounds allow the researcher to incorporate prior information on the propensity score, which may lead to point-identification if that prior information is correct and the overlap condition is satisfied, but gives robust confidence intervals in case the prior information on the propensity is not accurate (e.g. if the overlap condition is not satisfied).

3 Derivation of the new bounds

3.1 Second-order bounds

In the following we explain the derivation of our second-order ATE bounds in some detail. Afterwards, we explain more briefly how the same logic can be applied to obtain ATT bounds. Focusing on ATE for now, we want to generalize the Manski bounds in (3) by replacing $B^{(1)}(d,a)$ by

$$B^{(2)}(d,a) = a + \left[\lambda_0(d,X) + \lambda_1(d,X)\, p(X)\right] \mathbb{1}\left\{D = d\right\} (Y - a), \quad d \in \{0,1\}, \tag{10}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function,[6] and for every $d \in \{0,1\}$ and $x \in \mathcal{X}$ the coefficients $\lambda_0(d,x), \lambda_1(d,x) \in \mathbb{R}$ are non-random real numbers. Our goal is to choose those coefficients such that, for all $x \in \mathcal{X}$ and $d \in \{0,1\}$, we have

$$\mathbb{E}\left[B^{(2)}(d,a_{\min}) \,\middle|\, X = x\right] \leq \mathbb{E}\left[Y(d) \,\middle|\, X = x\right] \leq \mathbb{E}\left[B^{(2)}(d,a_{\max}) \,\middle|\, X = x\right]. \tag{11}$$

The $B^{(2)}(d,a)$ are linear functions of the unknown propensity score $p(X)$, that is, even after we have chosen the coefficients $\lambda_{0/1}(d,x)$ appropriately, the second order bounds that we discuss here are infeasible, because $p(X)$ is unknown. However, estimating those bounds turns out to be easier than estimating the expression for ATE in display (5), because the propensity score $p(X) = \mathbb{E}\left(D \,\middle|\, X\right)$ only enters linearly into the bounds. Consider, in particular, the case where for every observation $i \in \{1,\ldots,n\}$ we can find a matched observation $[i] \in \{1,\ldots,n\} \setminus \{i\}$ such that $X_i = X_{[i]}$. Sample analogs of $B^{(2)}(d,a)$ are then given by

$$\widehat{B}_i^{(2)}(d,a) = a + \left[\lambda_0(d,X_i) + \lambda_1(d,X_i)\, D_{[i]}\right] \mathbb{1}\left\{D_i = d\right\} (Y_i - a), \quad d \in \{0,1\}. \tag{12}$$

Then, under Assumption 1 we have $\mathbb{E}\left[D_{[i]} \,\middle|\, X_i = X_{[i]} = x\right] = p(x)$, implying that those sample analogs satisfy

$$\mathbb{E}\left[\widehat{B}_i^{(2)}(d,a) \,\middle|\, X_i = X_{[i]} = x\right] = \mathbb{E}\left[B^{(2)}(d,a) \,\middle|\, X = x\right], \qquad \text{for } d \in \{0,1\}. \tag{13}$$

Notice that $\widetilde{B}_{ij}^{(2)}(1,a)$ in (6) is an example of $\widehat{B}_i^{(2)}(1,a)$ with $\lambda_0(1,x) = 2$, $\lambda_1(1,x) = -1$, and $j = [i]$.

Having derived (11) and (13), we expect that we can take sample averages of the bounds $\widehat{B}_i^{(2)}(1,a_{\min}) - \widehat{B}_i^{(2)}(0,a_{\max})$ and $\widehat{B}_i^{(2)}(1,a_{\max}) - \widehat{B}_i^{(2)}(0,a_{\min})$ to obtain consistent upper and lower bounds for the ATE. This is indeed the case, and we properly discuss estimation of the

[6] Thus, we have $\mathbb{1}\{D = d\} = (1 - D)$ for $d = 0$ and $\mathbb{1}\{D = d\} = D$ for $d = 1$.

bounds in the following sections. The key takeaway from the discussion here is that estimating the second-order bounds is significantly easier than constructing a non-parametrically consistent estimate for $p(x)$ itself, because we only require one other observations $[i]$ with $X_i = X_{[i]}$ for each i, not many such observations for each i. In other words, the second-order bounds in (10) can be implemented by pooling information across only two observations.

Choosing the coefficients

We still need to choose the coefficients $\lambda_0(d,x), \lambda_1(d,x) \in \mathbb{R}$ in (10). If we find those coefficients such that the lower bound in (11) holds for all data generating processes (DGPs) that satisfy Assumption 1, then the upper bound in (11) also holds for such DGPs, because the problem of finding upper and lower bounds is symmetric under the transformation $Y \leftrightarrow -Y$ and $a_{\min} \leftrightarrow a_{\max}$. Also, if we find coefficients such that (11) holds for $d = 1$, then by applying the transformation $D \mapsto 1 - D$ and $p(X) \mapsto 1 - p(X)$ we also obtain coefficients that satisfy (11) for $d = 0$. Without loss of generality we therefore focus on finding $\lambda_{0/1}(1,x)$ such that the lower bound in (11) holds for $d = 1$.

Definition 1. *For a given value $x \in \mathcal{X}$ we say that the coefficients $\lambda_0(1,x), \lambda_1(1,x) \in \mathbb{R}$ and the corresponding bounds $B^{(2)}(1,a)$ defined in* (10) *are admissible if they satisfy* (11) *with $d = 1$, and if there do not exist alternative values $\lambda_0^*(1,x), \lambda_1^*(1,x) \in \mathbb{R}$ such that $B^*(1,a) := a + [\lambda_0^*(1,X) + \lambda_1^*(1,X)\, p(X)]\, D\,(Y - a)$ satisfies*

$$\mathbb{E}\left[B^{(2)}(1,a_{\min}) \,\middle|\, X = x\right] \leq \mathbb{E}\left[B^*(1,a_{\min}) \,\middle|\, X = x\right] \leq \mathbb{E}\left[Y(1) \,\middle|\, X = x\right], \tag{14}$$

for all DGPs that satisfy Assumption 1(i) and (ii), and where the first inequality in (14) *is strict for some of those DGPs.*

In other words, the coefficients are admissible if there is no alternative choice of coefficients that provide strictly better bounds in expectation. The following lemma characterizes all such admissible coefficients $\lambda_0(1,x), \lambda_1(1,x)$.

Lemma 1. *Let $x \in \mathcal{X}$. Let $\lambda_0(1,x), \lambda_1(1,x) \in \mathbb{R}$ be such that $B^{(2)}(1,a)$ defined in* (10) *satisfies $\mathbb{E}\left[B^{(2)}(1,a_{\min}) \,\middle|\, X = x\right] \leq \mathbb{E}\left[Y(1) \,\middle|\, X = x\right]$ for all DGPs that satisfy Assumption 1(i) and (ii). Assume furthermore that the coefficients $\lambda_0(1,x), \lambda_1(1,x) \in \mathbb{R}$ are admissible in the sense of Definition 1. Then, there exists $p_*(x) \in (0,1]$ such that*

$$\lambda_0(1,x) = \frac{2}{p_*(x)}, \qquad \lambda_1(1,x) = -\frac{1}{[p_*(x)]^2}.$$

The proof is provided in the appendix. The conclusion of Lemma 1 could equivalently have been written as $\lambda_0(1,x) \geq 2$ and $\lambda_1(1,x) = -[\lambda_0(1,x)]^2/4$. However, parameterizing the admissible coefficients in terms of $p_*(x) \in (0,1]$ turns out to be convenient in the following. Plugging the solution for the coefficients in Lemma 1 into (10) gives[7]

$$B^{(2)}(1,a) = a + w^{(2)}\big(p(x), p_*(x)\big)\, \frac{D(Y-a)}{p(x)},$$

where the weight function $w^{(2)} : [0,1] \times (0,1] \to (-\infty, 1]$ is given by

$$w^{(2)}(p, p_*) := 1 - \left(\frac{p_* - p}{p_*}\right)^2.$$

The expected value of the bounds therefore reads

$$\begin{aligned}\mathbb{E}\left[B^{(2)}(1,a) \,\middle|\, X = x\right] &= a + w^{(2)}\big(p(x), p_*(x)\big)\, \mathbb{E}\left[Y(1) - a \,\middle|\, X = x\right] \\ &= \Big[1 - w^{(2)}\big(p(x), p_*(x)\big)\Big]\, a + w^{(2)}\big(p(x), p_*(x)\big)\, \mathbb{E}\left[Y(1) \,\middle|\, X = x\right].\end{aligned}$$

Thus $\mathbb{E}\left[B^{(2)}(1,a) \,\middle|\, X = x\right]$ is a weighted average of $\mathbb{E}\left[Y(1) \,\middle|\, X = x\right]$ and the constant a (chosen to be $a_{\min}$ for the lower and $a_{\max}$ for the upper bound). If $p(x) = p_*(x)$, then we have $w^{(2)}(p(x), p_*(x)) = 1$ and therefore $\mathbb{E}\left[B^{(2)}(1,a) \,\middle|\, X = x\right] = \mathbb{E}\left[Y(1) \,\middle|\, X = x\right]$. Thus, if the true propensity score is equal to the value $p_*(x)$ that is chosen to construct $B^{(2)}(1,a)$, then the upper and lower bounds in (11) for $d = 1$ hold with equality. If $p(x) \neq p_*(x)$ then we have $w^{(2)}(p(x), p_*(x)) < 1$ and the upper and lower bounds in (11) then usually are not binding.

Figure 2 shows the weights $w^{(2)}(p, p_*)$ as a function of p for different values of p_*. For the Manski bounds we have $\mathbb{E}\left[B^{(1)}(1,a) \,\middle|\, X = x\right] = [1 - p(x)]\, a + p(x)\, \mathbb{E}\left[Y(1) \,\middle|\, X = x\right]$, that is the corresponding weight function is simply $w^{(1)} : p \mapsto p$, which is also shown in the figure. Our initial example in (6) corresponds to $p_*(x) = 1$, which is the only second-order weight function that strictly dominates the Manski bounds, independent from the value of the true propensity score (a larger weight implies a better bound). Since we only consider admissible second-order bounds in the sense of Definition 1, we have that none of the second-order bounds strictly dominates any of the other second-order bounds with a different value of $p_*(x)$. Therefore, the performance of the second-order bounds depends on whether the true $p(x)$ is close to the chosen $p_*(x)$. Notice also that for $p_*(x) < 0.5$ the weights $w^{(2)}\big(p(x), p_*(x)\big)$ become negative for large values of $p(x)$, implying that the lower (upper) bound in (11) can be smaller (larger) than $a_{\min}$ ($a_{\max}$).

[7]The formula for $B^{(2)}(1,a)$ here is not applicable for $p(x) = 0$, but the limit $p(x) \to 0$ is well-defined, see display (19) below.

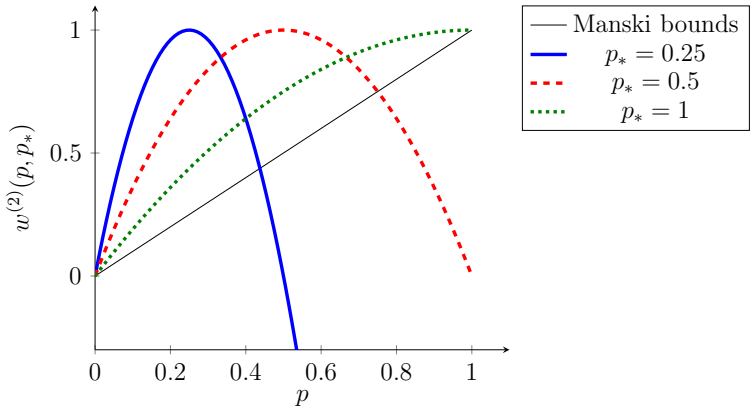

Figure 2: Weights $w^{(2)}(p, p_*)$ as a function of p, for different values of p_*.

Second-order ATT bounds

So far we have focused on the ATE bounds. We now apply the same argument to derive ATT bounds. We want to generalize the first-oder bounds in (4) by replacing $C^{(1)}(a)$ with

$$C^{(2)}(a) = D\,(Y-a) + [\lambda_0(X) + \lambda_1(X)\,p(X)]\;(1-D)\,(Y-a), \tag{15}$$

where the coefficients $\lambda_0(x), \lambda_1(x) \in \mathbb{R}$ need to be determined such that

$$\mathbb{E}\left[C^{(2)}(a_{\max}) \,\middle|\, X=x\right] \;\leq\; \mathbb{E}\left\{D\;[Y(1)-Y(0)] \;\middle|\; X=x\right\} \;\leq\; \mathbb{E}\left[C^{(2)}(a_{\min}) \,\middle|\, X=x\right], \tag{16}$$

which guarantees that (4) holds when replacing $C^{(1)}(a)$ by $C^{(2)}(a)$.[8] Analogous to Lemma 1 one finds that the set of admissible coefficients $\lambda_0(x), \lambda_1(x) \in \mathbb{R}$ is described by

$$\lambda_0(x) = \left[\frac{p_*(x)}{1-p_*(x)}\right]^2, \qquad\qquad \lambda_1(x) = -\frac{1}{[1-p_*(x)]^2}, \tag{18}$$

where $p_*(x) \in [0,1)$ can be chosen arbitrarily. Plugging those solutions for the coefficients back into (15) gives

$$C^{(2)}(a) = D\,(Y-a) - \widetilde{w}^{(2)}\big(p(x), p_*(x)\big)\,\frac{p(X)\,(1-D)\,(Y-a)}{1-p(X)},$$

[8] Analogous to $B^{(2)}(d,a)$ the bounds $C^{(2)}(a)$ presented here are still infeasible, because they depend on the unknown propensity score $p(x)$. However, for the case that observation i has a matched observation $[i] \neq i$ such that $X_i = X_{[i]}$ we have the sample analog

$$\widehat{C}_i^{(2)}(a) = D_i\,(Y_i-a) + \left[\lambda_0(X_i) + \lambda_1(X_i)\,D_{[i]}\right]\;(1-D_i)\,(Y_i-a), \tag{17}$$

and under Assumption 1, $\mathbb{E}\left[\widehat{C}_i^{(2)}(a) \,\middle|\, X_i = X_{[i]} = x\right] = \mathbb{E}\left[C^{(2)}(a) \,\middle|\, X=x\right]$.

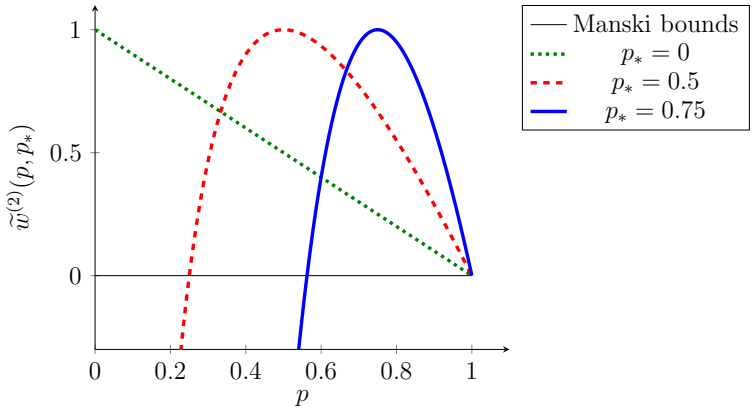

Figure 3: Weights $\widetilde{w}^{(2)}(p, p_*)$ as a function of p, for different values of p_*.

where the weight function $\widetilde{w}^{(2)} : [0,1] \times (0,1] \to (-\infty, 1]$ reads

$$\widetilde{w}^{(2)}(p, p_*) := 1 - \frac{1}{p} \left(\frac{p - p_*}{1 - p_*} \right)^2 .$$

Under Assumption 1 we calculate that

$$\begin{aligned} \frac{\mathbb{E}\left[C^{(2)}(a) \middle| X{=}x\right]}{\mathbb{E}\left(D\right)} &= \frac{p(x)}{\mathbb{E}\left(D\right)} \Big\{ \mathbb{E}\left[Y(1) - a \middle| X = x\right] - \widetilde{w}^{(2)}\big(p(x), p_*(x)\big)\, \mathbb{E}\left[Y(0) - a \middle| X = x\right] \Big\} \\ &= \Big[1 - \widetilde{w}^{(2)}\big(p(x), p_*(x)\big)\Big] \frac{\mathbb{E}\left[C^{(1)}(a) \middle| X = x\right]}{\mathbb{E}\left(D\right)} + \widetilde{w}^{(2)}\big(p(x), p_*(x)\big)\, \pi(x). \end{aligned}$$

Therefore, conditional on $X = x$, the second-order bounds on ATT are weighted averages between the first-order bounds and ATT(x). The weight $\widetilde{w}^{(2)}\big(p(x), p_*(x)\big)$ is equal to one if $p(x) = p_*(x)$. Thus, if $p(x) = p_*(x)$, then the second order bound on ATT holds with equality.

Figure 3 plots the weights $\widetilde{w}^{(2)}\big(p(x), p_*(x)\big)$ as a function of $p(x)$ for different values of $p_*(x)$. Here, the Manski bounds correspond to a weight $\widetilde{w}^{(2)}$ equal to zero. Only for $p_*(x) = 0$ are the second-order bounds uniformly better than the Manski bounds. However, whenever the true $p(x)$ is close to the chosen $p_*(x)$, then the second-order bounds improve on the Manski bounds.

Final result for second-order bounds

In contrast to the Manski bounds discussed in the last section, the second-order bounds derived here are not unique. In order to implement those bounds we therefore require the researcher to provide a reference propensity score $p_* : \mathcal{X} \to (0,1)$, which can be postulated

or estimated. The resulting second-order bounds will be valid independent of the choice of $p_*(x)$, but the performance of the bounds depends on whether the true propensity score is close to $p_*(x)$ or not. For all our theoretical results we assume that $p_*(x)$ is non-random, for example, $p_*(x) = 1/2$.

Based on the results above, for every $a \in \mathbb{R}$, we define

$$B^{(2)}(0,a) := a + \frac{1 - 2\,p_*(X) + p(X)}{[1 - p_*(X)]^2}\,(1-D)\,(Y-a),$$

$$B^{(2)}(1,a) := a + \frac{2\,p_*(X) - p(X)}{[p_*(X)]^2}\,D\,(Y-a),$$

$$C^{(2)}(a) := D\,(Y-a) + \frac{[p_*(X)]^2 - p(X)}{[1 - p_*(X)]^2}\,(1-D)\,(Y-a). \tag{19}$$

From now on these are our definitions of the random variables $B^{(2)}(0,a), B^{(2)}(1,a), C^{(2)}(a) \in \mathbb{R}$. The expressions here are obtained from (10) and (15) by plugging in the solutions for the coefficients in Lemma 1 and display (18). The following proposition summarizes the main properties of these bounds.

Proposition 1. *Let Assumption 1(i), (ii), (iii) hold. Let $p_* : \mathcal{X} \to (0,1)$. Then,*

$$(a) \qquad \mathbb{E}\left[B^{(2)}(d, a_{\min})\right] \le \mathbb{E}\,Y(d) \le \mathbb{E}\left[B^{(2)}(d, a_{\max})\right], \quad d \in \{0,1\},$$

$$(b) \quad \mathbb{E}\left[B^{(2)}(1, a_{\min}) - B^{(2)}(0, a_{\max})\right] \le \text{ ATE } \le \mathbb{E}\left[B^{(2)}(1, a_{\max}) - B^{(2)}(0, a_{\min})\right],$$

$$(c) \qquad \frac{\mathbb{E}\left[C^{(2)}(a_{\max})\right]}{\mathbb{E}\,(D)} \le \text{ ATT } \le \frac{\mathbb{E}\left[C^{(2)}(a_{\min})\right]}{\mathbb{E}\,(D)}.$$

If, in addition, $p(x) = p_*(x)$, for all $x \in \mathcal{X}$, then all the inequalities in this proposition become equalities.

The proof is provided in the appendix. Notice that the result in the proposition also holds conditional on a particular regressor value, that is, under Assumption 1 we also have

$$\mathbb{E}\left[B^{(2)}(d, a_{\min}) \,\middle|\, X\right] \le \mathbb{E}\left[Y(d) \,\middle|\, X\right] \le \mathbb{E}\left[B^{(2)}(d, a_{\max}) \,\middle|\, X\right],$$

$$\mathbb{E}\left[B^{(2)}(1, a_{\min}) - B^{(2)}(0, a_{\max}) \,\middle|\, X\right] \le \tau(X) \le \mathbb{E}\left[B^{(2)}(1, a_{\max}) - B^{(2)}(0, a_{\min}) \,\middle|\, X\right],$$

$$\frac{\mathbb{E}\left[C^{(2)}(a_{\max}) \,\middle|\, X\right]}{\mathbb{E}\,(D)} \le \pi(X) \le \frac{\mathbb{E}\left[C^{(2)}(a_{\min}) \,\middle|\, X\right]}{\mathbb{E}\,(D)}, \tag{20}$$

again with equality everywhere if $p(x) = p_*(x)$. We prefer to formulate the proposition in an unconditional way, because ATE and ATT are our actual objects of interest. Again, we stress

that the bounds in Proposition 1 are infeasible (because dependent on the unknown propensity score), but feasible estimation of those bounds is discussed in the following sections. Focusing on those infeasible bounds in this section significantly simplifies the discussion, and it also allows for different estimation methods (depending on the DGP for X) to be considered in the following sections.

3.2 Higher-order bounds

The starting point for the second-order bounds were equations (10) and (15), where $p(x)$ enters linearly. Improved bounds can be obtained by letting $p(x)$ enter as higher order polynomials. For positive integers q we consider

$$B^{(q)}(d,a,\lambda) = a + \left\{\sum_{r=0}^{q-1} \lambda_r(d,X)\,[p(X)]^r\right\} \mathbb{1}\{D=d\}\,(Y-a), \quad d\in\{0,1\},$$
$$C^{(q)}(a,\lambda) = D\,(Y-a) - \left\{\sum_{r=0}^{q-1} \lambda_r(X)\,[p(X)]^r\right\} (1-D)\,(Y-a), \tag{21}$$

where we now make the dependence of $B^{(q)}$ and $C^{(q)}$ on the coefficients $\lambda_r(d,x), \lambda_r(x) \in \mathbb{R}$ explicit. The motivation for (21) is that, analogous to (12) and (17) above, we can construct unbiased estimates for $B^{(q)}(d,a,\lambda)$ and $C^{(q)}(a,\lambda)$ by replacing $[p(X)]^r$ with a product of treatment indicators D_i from r different observations i with the same (or similar) regressor values X_i. This is discussed in detail in the following section.

Motivated by our derivation of the second-order bounds we again choose a reference propensity score $p_*(x)$ to find unique solutions for the coefficients $\lambda_r(d,x)$ and $\lambda_r(x)$. Once we have chosen $p_*(x)$, then for the second-order bounds the coefficients are determined by the properties of the bounds summarized in Proposition 1 — namely, the bounds should be valid for all population distributions satisfying Assumption 1, and the bounds should be binding if $p(x) = p_*(x)$. However, for $q > 2$ those properties are not sufficient anymore to uniquely determine the coefficients, because we now have additional degrees of freedom in the higher-order polynomial coefficients. To make use of this additional flexibility and to obtain unique coefficients again we therefore demand the bounds to not only have good properties when $p(x) = p_*(x)$, but also when $p(x) \in [p_*(x) - \epsilon, p_*(x) + \epsilon]$, for small $\epsilon > 0$, that is, we want to have good performance in a small neighborhood around the reference propensity score $p_*(x)$.

Let $d \in \{0,1\}$ and $x \in \mathcal{X}$. Let $\Lambda(d,x)$ be the set of coefficients $\lambda(d,x) = (\lambda_0(d,x), \ldots,$

$\lambda_{q-1}(d,x)) \in \mathbb{R}^q$ that satisfy

$$\mathbb{E}\left[B^{(q)}(d, a_{\min}, \lambda) \,\middle|\, X = x\right] \leq \mathbb{E}\left[Y(d) \,\middle|\, X = x\right] \leq \mathbb{E}\left[B^{(2)}(d, a_{\max}, \lambda) \,\middle|\, X = x\right],$$

for all population distributions that satisfy Assumption 1. We then choose the optimal coefficients as the solution of the following minimax problem

$$\min_{\lambda(d,x)\in\Lambda(d,x)} \max_{\left\{p(x)\in[0,1]\,:\,\left|p(x)-p_*(x)\right|\leq\epsilon\right\}} \mathbb{E}_{p(x)}\left[B^{(q)}(d, a_{\max}, \lambda) - B^{(q)}(d, a_{\min}, \lambda) \,\middle|\, X = x\right], \quad (22)$$

where $\epsilon > 0$ is chosen to be infinitesimally small in order to make the solution unique. In (22) the $\mathbb{E}_{p(x)}$ refers to the expectation over $\mathcal{P} = \mathcal{P}(p, \overline{\mathcal{P}})$, where p is the chosen propensity score, and the distribution $\overline{\mathcal{P}}$ of $(Y(1), Y(2), X))$ is fixed, see Section 2 for details. Notice that $\mathbb{E}_{p(x)}\left[B^{(q)}(d, a_{\max}, \lambda) - B^{(q)}(d, a_{\min}, \lambda) \,\middle|\, X = x\right]$ is the width of the interval that we obtain for $\mathbb{E}_{p(x)}\left[Y(d) \,\middle|\, X = x\right]$, that is, among all the $\lambda(d,x) \in \Lambda(d,x)$ we choose the one that generates bounds that are closest to point-identification within a small neighborhood of $p_*(x)$.

An analogous minimax problem can be written down for the coefficients $\lambda_r(x)$ of $C^{(q)}(a, \lambda)$. Those optimality properties of our bounds are presented more formally in Proposition 2. Once we have solved for the optimal coefficients, then we obtain the following optimal $B^{(q)}(d, a, \lambda)$ and $C^{(q)}(a, \lambda)$, for integers $q \geq 1$,[9]

$$\begin{aligned}
B^{(q)}(0, a) &:= a + w^{(q)}\big(1 - p(X), 1 - p_*(X)\big)\, \frac{(1-D)(Y-a)}{1-p(X)}, \\
B^{(q)}(1, a) &:= a + w^{(q)}\big(p(X), p_*(X)\big)\, \frac{D(Y-a)}{p(X)}, \\
C^{(q)}(a) &:= D\,(Y-a) - \widetilde{w}^{(q)}\big(p(X), p_*(X)\big)\, \frac{p(X)\,(1-D)\,(Y-a)}{1-p(X)},
\end{aligned} \quad (23)$$

[9]Formally, for $p(x) = 1$ we have $B^{(q)}(0,a) = a + \left\{\frac{q - p_*(x)\,\mathbb{1}\{q \text{ is odd}\}}{1-p_*(x)}\right\}(1-D)(Y-a)$ and $C^{(q)}(a) = D\,(Y-a) + \left[\frac{q-1+p_*(x)\,\mathbb{1}\{q \text{ is even}\}}{1-p_*(x)}\right](1-D)(Y-a)$. For $p(x) = 0$ we have $a + \left\{\frac{q-[1-p_*(x)]\,\mathbb{1}\{q \text{ is odd}\}}{p_*(x)}\right\} D(Y-a)$. From the formulas in (23) we obtain those results for $p(x) = 1$ and $p(x) = 0$ as limits when $p(x) \to 1$ and $p(x) \to 0$. However, the details of those special cases do not actually matter, because e.g. for $p(x) = 1$ we also have $D = 1$ with probability one, and therefore $B^{(q)}(0,a) = a$ and $C^{(q)}(a) = D\,(Y-a)$.

where the weight functions are given by

$$w^{(q)}(p,p_*) := \begin{cases} 1-(1-p)\left(\dfrac{p_*-p}{p_*}\right)^{q-1} & \text{if } q \text{ is odd,} \\ 1-\left(\dfrac{p_*-p}{p_*}\right)^{q} & \text{if } q \text{ is even,} \end{cases}$$

$$\widetilde{w}^{(q)}(p,p_*) := \begin{cases} 1-\left(\dfrac{p-p_*}{1-p_*}\right)^{q-1} & \text{if } q \text{ is odd,} \\ 1-\dfrac{1}{p}\left(\dfrac{p-p_*}{1-p_*}\right)^{q} & \text{if } q \text{ is even.} \end{cases} \tag{24}$$

For $q=1$ and $q=2$ the formulas in (23) just give the same functions $B^{(q)}(d,a)$ and $C^{(q)}(a)$ that were already discussed above. It may not be obvious from those general formulas, but $B^{(q)}(d,a)$ and $C^{(q)}(a)$ are indeed polynomials of order $(q-1)$ in $p(X)$. For example, for $q=3$ we find

$$B^{(3)}(0,a) = a + \left\{1 + p(X)\,\frac{1+p(X)-2\,p_*(X)}{[1-p_*(X)]^2}\right\}(1-D)\,(Y-a),$$

$$B^{(3)}(1,a) = a + \left\{1 + [1-p(X)]\,\frac{2\,p_*(X)-p(X)}{[p_*(X)]^2}\right\}D\,(Y-a),$$

$$C^{(3)}(a) = D\,(Y-a) - p(X)\,\frac{1+p(X)-2\,p_*(X)}{[1-p_*(X)]^2}(1-D)\,(Y-a),$$

which are all second order polynomials in $p(X)$. We now want to formally state the optimality result for these bounds. For $\epsilon \geq 0$ and $p_*(x) \in (0,1)$, let $\mathcal{B}_\epsilon(p_*(x)) := \Big\{p(x) \in [0,1] \,\Big|\, |p(X)-p_*(X)| \leq \epsilon\Big\}$ be the ϵ-ball around $p_*(x)$.

Proposition 2. *Let Assumption 1(i), (ii), (iii) hold. Let $p_* : \mathcal{X} \to (0,1)$. Then:*

(i) For integers $q \geq 1$ we have

$$(a) \qquad \mathbb{E}\left[B^{(q)}(d,a_{\min})\right] \leq \mathbb{E}\,Y(d) \leq \mathbb{E}\left[B^{(q)}(d,a_{\max})\right], \quad d \in \{0,1\},$$

$$(b) \quad \mathbb{E}\left[B^{(q)}(1,a_{\min}) - B^{(q)}(0,a_{\max})\right] \leq \text{ ATE } \leq \mathbb{E}\left[B^{(q)}(1,a_{\max}) - B^{(q)}(0,a_{\min})\right],$$

$$(c) \qquad \frac{\mathbb{E}\left[C^{(q)}(a_{\max})\right]}{\mathbb{E}\,(D)} \leq \text{ ATT } \leq \frac{\mathbb{E}\left[C^{(q)}(a_{\min})\right]}{\mathbb{E}\,(D)}.$$

(ii) If $q > 1$ and $p(x) = p_*(x)$, for all $x \in \mathcal{X}$, then all the inequalities in part (i) of the proposition become equalities.

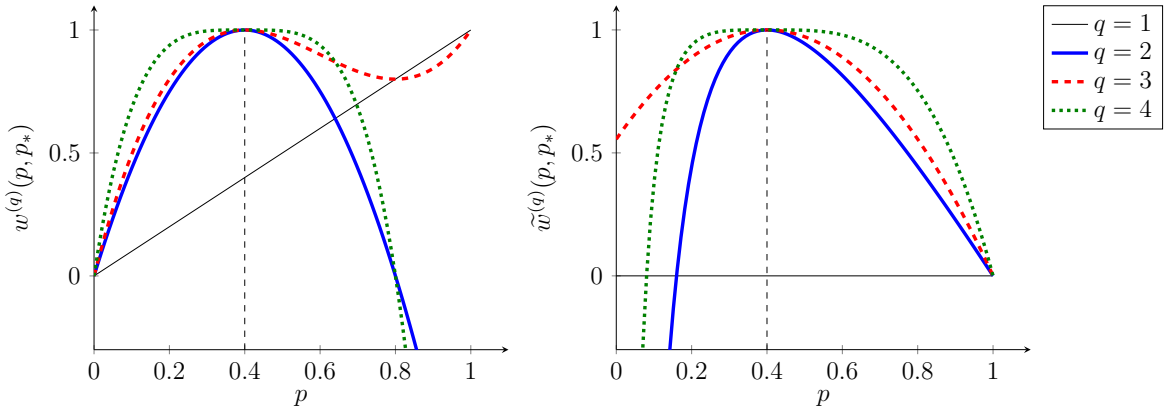

Figure 4: Weights $w^{(q)}(p, p_*)$ and $\widetilde{w}^{(q)}(p, p_*)$ as a function of p, for $p_* = 0.4$ and $q \in \{1, 2, 3, 4\}$.

(iii) Let $\lambda_r(d, x) \in \mathbb{R}$ and $\lambda_r(x) \in \mathbb{R}$ be such that $B^{(q)}(d, a, \lambda)$ and $C^{(q)}(a, \lambda)$ defined in (21) satisfy the inequalities in part (i) for all population distribution that satisfy Assumption 1. Then, for all $x \in \mathcal{X}$ there exists $\epsilon > 0$ such that for all $p(x) \in \mathcal{B}_\epsilon(p_*(x))$ and $d \in \{0, 1\}$ we have

$$\mathbb{E}_{p(x)}\left[B^{(q)}(d, a_{\max}) - B^{(q)}(d, a_{\min}) \,\middle|\, X = x\right] \leq \mathbb{E}_{p(x)}\left[B^{(q)}(d, a_{\max}, \lambda) - B^{(q)}(d, a_{\min}, \lambda) \,\middle|\, X = x\right],$$

and

$$\mathbb{E}_{p(x)}\left[C^{(q)}(a_{\min}) - C^{(q)}(a_{\max}) \,\middle|\, X = x\right] \leq \mathbb{E}_{p(x)}\left[C^{(q)}(a_{\min}, \lambda) - C^{(q)}(a_{\max}, \lambda) \,\middle|\, X = x\right].$$

That is, within a small neighborhood of $p_*(x)$, the expected width of the bounds in part (i) is smaller or equal to the expected width of any other set of valid q'th order bounds.

The proof is given in the appendix. To better understand the result of Proposition (2), consider the lower bound on $\mathbb{E}\left[Y(1) \,\middle|\, X\right]$, which is given by

$$\mathbb{E}\left[B^{(q)}(1, a_{\min}) \,\middle|\, X\right] = \left[1 - w^{(q)}(p(X), p_*(X))\right] a_{\min} + w^{(q)}(p(X), p_*(X))\, \mathbb{E}\left[Y(1) \,\middle|\, X\right].$$

Thus, $\mathbb{E}\left[B^{(q)}(1, a_{\min}) \,\middle|\, X = x\right]$ is a weighted average between $a_{\min}$ and $\mathbb{E}\left[Y(1) \,\middle|\, X = x\right]$. The weights always satisfy $w^{(q)}(p, p_*) \leq 1$, which together with $a_{\min} \leq Y(1)$ guarantees that $\mathbb{E}\left[B^{(q)}(1, a_{\min}) \,\middle|\, X = x\right] \leq \mathbb{E}\left[Y(1) \,\middle|\, X = x\right]$.

Figure 4 shows $w^{(q)}(p, p_*)$ as a function of p for $p_* = 0.4$ and different values of q. For $p = 0$ we always have $w^{(q)}(p, p_*) = 0$, because in that case we only have observations with

$D = 0$ for $X = x$, implying that we cannot learn anything about $Y(1)$ from the data. For $p = p_*$ we have $w^{(q)}(p, p_*) = 1$ for $q \geq 2$, that is, the lower bound is sharp in that case. For p close to p_* the weights are closer to one (implying that the bounds are sharper) the larger we choose q. For the k'th derivative of $w^{(q)}(p, p_*)$ at $p = p_*$ we have

$$\frac{\partial^k w^{(q)}(p_*, p_*)}{\partial^k p} = 0, \qquad \text{for} \begin{cases} k \in \{1, \ldots, q-2\} & \text{if } q \text{ is odd,} \\ k \in \{1, \ldots, q-1\} & \text{if } q \text{ is even,} \end{cases}$$

which explain why for p close to p_* the weights are closer to one the higher we choose q. However, if p is far away from p_*, then the weights $w^{(q)}(p, p_*)$ for $q \geq 2$ can be far away from one, and can even be smaller than $w^{(1)}(p)$, that is, the bounds can be worse than Manski bounds if p is far away from p_*.

The discussion for ATT bounds is analogous. In that case we have

$$\frac{\mathbb{E}\left[C^{(q)}(a)\middle|X\right]}{\mathbb{E}\left(D\right)} = \left[1 - \widetilde{w}^{(q)}\big(p(X), p_*(X)\big)\right] \frac{\mathbb{E}\left[C^{(1)}(a)\middle|X\right]}{\mathbb{E}\left(D\right)} + \widetilde{w}^{(q)}\big(p(X), p_*(X)\big)\, \pi(X),$$

that is, conditional on X, the ATT bounds are a linear combination between their Manski bounds and the true ATT contribution for X. Figure 4 also shows the weights $\widetilde{w}^{(q)}(p, p_*)$ as a function of p, for $p_* = 0.4$ and various values of q.

Remark 1. We have chosen to consider bounds that are optimal in a small neighborhood of a given reference propensity score $p_*(x)$. Alternative bounds can be constructed based on other optimality criteria. For example, subject to the bounds being valid for all population distribution that satisfy Assumption 1, one could minimize the expected width of the bounds under a chosen prior on the propensity score. From a frequentist perspective, it is ultimately a matter of taste what optimality criteria to use here. We find it convenient to parameterize the bounds in terms of the reference propensity score $p_*(x)$, because it is easy to interpret and leads to easy analytic formulas for the bounds.

Remark 2. Even the local optimality of our bounds needs to be interpreted carefully. This is because in part (iii) of Proposition 2 we only compare to other bounds of the form (21), and it is natural ask about the existence of other bounds, say for $\mathbb{E}\, Y(d)$, that are not of the form $\mathbb{E}\left[B^{(q)}(d, a, \lambda)\right]$. Such bounds indeed exist, and the most obvious example is the following: Let $B^{(q)}_{p_*}(d, a) := B^{(q)}(d, a)$ be as defined in (23), but with the dependence on p_* now made explicit. Let $\mathcal{P}_*$ be a set of functions $p_* : \mathcal{X} \to (0, 1)$. Then we have

$$\sup_{p_* \in \mathcal{P}_*} \mathbb{E}\left[B^{(q)}_{p_*}(d, a_{\min})\right] \leq \mathbb{E}\, Y(d) \leq \inf_{p_* \in \mathcal{P}_*} \mathbb{E}\left[B^{(q)}_{p_*}(d, a_{\max})\right]. \tag{25}$$

Thus, by forming intersections of the bounds discussed so far we can obtain new valid bounds, and those intersection bounds are generally tighter (see Chernozhukov, Lee and Rosen 2013). We do not consider such intersection bounds any further in this paper, and leave the question of constructing truly "optimal bounds" (in some sense) to future research.

Remark 3. Armstrong and Kolesár (2021) construct fixed-length confidence intervals, which are optimal in finite samples for normally distributed regression errors with known variance. Their method is valid asymptotically under a lack of overlap, provided that the researcher specifies a Lipschitz bound on how much the conditional mean of the outcome variable can change over the covariate space. Their approach is complementary to ours, in particular, they condition on the realizations of the treatment variable in their inference results, while our bounds are valid only after taking expectations over the realization of the treatment variable — as a result of this, our approach allows to incorporate prior information on (or prior estimates of) the propensity score, which can help to shrink the width of the confidence intervals significantly (to a point, if the prior is correct), while maintaining robustness over all possible data generating processes. If the researcher is willing to specify a Lipschitz bound on the conditional mean, then in principle, that information can also be incorporated into our bounds, that is, the two complementary approaches to robust confidence intervals could be fruitfully combined, but we leave that generalization to future research.

4 Implementation of the Bounds

In this section we construct sample analogs of the bounds in Proposition 2, and use those sample bounds to obtain asymptotically valid confidence intervals on the average treatment effects. The bounds constructed in this section are valid for both discrete and continuous covariates X_i. However, if the covariates are continuously distributed, then every observed value X_i is typically only observed once, in which case the bounds here simply become Manski worst-case bounds.

The interesting case, for the purpose of this section, is therefore the case where the set of possible covariate values $\mathcal{X}$ is discrete. However, we consider an asymptotic setting where the number of covariate values grows to infinity jointly with the total sample size. This is the challenging case from the perspective of treatment effect estimation, in particular when the average number observations available for each observed $x \in \mathcal{X}$ remains small.

In Section 5 we explain how the sample bounds for discrete covariate values from this section can be generalized to continuous covariate values via clustering, that is, by approxi-

mating the continuous set $\mathcal{X}$ with a finite set. In that way we obtain non-trivial bounds also for the case of continuous covariates.

4.1 Sample analogs of the bounds of Section 3

We require some additional notation to formulate the sample bounds. Let $D^{(n)} := (D_1, \ldots, D_n)$ and $X^{(n)} := (X_1, \ldots, X_n)$ be the observed samples of treatment status and covariates. Let $\mathcal{X}_* := \{X_i : i = 1, \ldots, n\} \subset \mathcal{X}$ be the set of actually observed covariate values in the sample, and let $m := |\mathcal{X}_*|$ be the cardinality of $\mathcal{X}_*$. As already mentioned above, in our asymptotic analysis we let $m \to \infty$ as $n \to \infty$. This implies that $\mathcal{X}_*$ changes with the sample size (we can allow $\mathcal{X}$ to change with n as well), but we do not make that explicit in our notation. For $x \in \mathcal{X}$ we define

$$\mathcal{N}(x) := \Big\{ i \in \{1, \ldots, n\} \,\Big|\, X_i = x \Big\},$$

the set of observations i for which the observed covariates value is equal to x.[10] Let $n(x) := |\mathcal{N}(x)|$ be number of observations with $X_i = x$, and let

$$n_0(x) := \sum_{i \in \mathcal{N}(x)} (1 - D_i), \qquad n_1(x) := \sum_{i \in \mathcal{N}(x)} D_i = n(x) - n_0(x)$$

be the number of observations with $X_i = x$, and $D_i = 0$ or $D_i = 1$, respectively.

To construct our sample bounds, we furthermore require the researcher to choose a "bandwidth parameter" $Q \in \{1, 2, 3, \ldots, \infty\}$. If $\max_{x \in \mathcal{X}_*} n(x)$ remains bounded as $n \to \infty$, then we can choose $Q = \infty$, which simplifies many of the expressions in this section, and the reader may think of this case as the baseline case which makes the connection to Section 3 most obvious.

For each covariate value $x \in \mathcal{X}_*$ we need to choose the order $q(x) \in \{1, 2, 3, \ldots\}$ of the bounds in Proposition 2 that we want to implement. To implement bounds of a certain order $q(x)$ we require at least that many observations for that covariate value, that is, we need to choose $q(x) \leq n(x)$. Choosing the maximal value $q(x) = n(x)$ is optimal from the perspective of expected width of the bounds, but it is not advisable in general since it can lead to upper and lower bound estimates with very large variance. In our implementation of the bounds we therefore choose

$$q(x) := \min\{Q, n(x)\}, \tag{26}$$

[10] $\mathcal{N}(x)$ is empty for $x \notin \mathcal{X}_*$.

that is, we choose the maximum order that satisfies both $q(x) \leq n(x)$ and $q(x) \leq Q$. In practice, we recommend choosing Q as small as $Q = 3$ or $Q = 4$, but the choice $Q = \infty$ gives some theoretical optimally properties for the expected width of the bounds (but usually at the cost of higher variance).

Having chosen the order $q(x)$ for each $x \in \mathcal{X}_*$, we then construct sample weights $\widehat{w}_0(x)$, $\widehat{w}_1(x)$, $\widehat{v}(x)$, which are functions of the chosen order $q(x)$, the chosen reference propensity score $p_*(x)$, and the values $n(x)$, $n_0(x)$, $n_1(x)$ obtained from the sample, such that

$$\begin{aligned}
\mathbb{E}\left[\widehat{w}_0(x) \,\middle|\, X^{(n)}\right] &= w^{(q(x))}\big(1 - p(x), 1 - p_*(x)\big), \\
\mathbb{E}\left[\widehat{w}_1(x) \,\middle|\, X^{(n)}\right] &= w^{(q(x))}\big(p(x), p_*(x)\big), \\
\mathbb{E}\left[\widehat{v}(x) \,\middle|\, X^{(n)}\right] &= p(x)\, \widetilde{w}^{(q(x))}\big(p(x), p_*(x)\big),
\end{aligned} \tag{27}$$

where $q(x)$ is given in (26), and the weight functions $w^{(q)}$ and $\widetilde{w}^{(q)}$ on the right hand side were defined in (24). Here and in the following, the dependence of those sample weights on $p_*(x)$ and $q(x)$ (and thereby on Q) is not made explicit, and the dependence on the sample $X^{(n)}$ and $D^{(n)}$ (through $n(x)$, $n_0(x)$, $n_1(x)$) is only indicated by the "hat". Explicit formulas for $\widehat{w}_0(x)$, $\widehat{w}_1(x)$, $\widehat{v}(x)$ are provided in the next subsection.

The natural sample analogs of the bounds $B^{(q)}(d, a)$ and $C^{(q)}(a)$ in the last section are then given by

$$\begin{aligned}
\widehat{B}_i(0, a) &:= a + \widehat{w}_0(X_i)\, \frac{n(X_i)\,(1 - D_i)\,(Y_i - a)}{\max\{1, n_0(X_i)\}}, \\
\widehat{B}_i(1, a) &:= a + \widehat{w}_1(X_i)\, \frac{n(X_i)\, D_i\,(Y_i - a)}{\max\{1, n_1(X_i)\}}, \\
\widehat{C}_i(a) &:= D_i\,(Y_i - a) - \widehat{v}(X_i)\, \frac{n(X_i)\,(1 - D_i)\,(Y_i - a)}{\max\{1, n_0(X_i)\}},
\end{aligned} \tag{28}$$

for $a \in \mathbb{R}$. In view of (27), the expressions in (28) are direct translations of the formulas in display (23), where the weights were replaced by sample weights, and the remaining occurrences of the unknown $1 - p(x)$ and $p(x)$ were replaced by their sample analogs $n_1(x)/n(x)$ and $n_0(x)/n(x)$, respectively. In all three expression of display (28) the maximum function in the denominator is only included to avoid a potentially zero denominator. However, $n_0(X_i) = 0$ implies $1 - D_i = 0$, and $n_1(X_i) = 0$ implies $D_i = 0$, that is, in all cases where the maximum function is required to avoid a zero denominator, the corresponding numerator is zero anyways. In particular, we could replaced $\max\{1, \ldots\}$ by $\max\{c, \ldots\}$ for any $c > 0$ without changing the sample bounds in (28) at all.

The sample analogs of the expectations over $B^{(q)}(d,a)$ and $C^{(q)}(a)$ are then given by

$$\overline{B}(d,a) := \frac{1}{n}\sum_{i=1}^{n} \widehat{B}_i(d,a), \qquad\qquad \overline{C}(a) := \frac{1}{n}\sum_{i=1}^{n} \widehat{C}_i(a), \tag{29}$$

and the final upper and lower sample bounds on the ATE read

$$\overline{L}^{(\mathrm{ATE})} := \overline{B}(1,a_{\min}) - \overline{B}(0,a_{\max}), \qquad \overline{U}^{(\mathrm{ATE})} := \overline{B}(1,a_{\max}) - \overline{B}(0,a_{\min}). \tag{30}$$

Similarly, the lower and upper sample bounds on $\mathrm{ATT}\cdot\mathbb{E}D$ are given by $\overline{C}(a_{\max})$ and $\overline{C}(a_{\min})$, respectively. To estimate the lower- and upper bounds on the ATT itself we still need to plug-in the sample analog of $\mathbb{E}D$, which gives

$$\overline{L}^{(\mathrm{ATT})} := \frac{\overline{C}(a_{\max})}{\frac{1}{n}\sum_{i=1}^{n} D_i}, \qquad\qquad \overline{U}^{(\mathrm{ATT})} := \frac{\overline{C}(a_{\min})}{\frac{1}{n}\sum_{i=1}^{n} D_i}. \tag{31}$$

In Section 4.3 we show that the sample bounds just constructed are unbiased and consistent estimates (as $m \to \infty$) of the corresponding population bounds from the last section, and we will also use those sample bounds to construct asymptotically valid confidence intervals for ATE and ATT.

4.2 Construction of the sample weights $\widehat{w}_0(x)$, $\widehat{w}_1(x)$, $\widehat{v}(x)$

A key ingredient of the sample bounds just introduced are the sample weights that satisfy (27), and which we want to define in this section. For ease of exposition we start with the simplest case $q(x) = n(x)$, which can be even or odd, and then generalize the formulas to the case $q(x) = \min\{Q, n(x)\}$ afterwards.

4.2.1 Case $q(x) = n(x)$ and $n(x)$ even

Let $q(x) = n(x)$, and assume that $n(x)$ is even. We consider $\widehat{w}_1(x)$ first. By setting

$$\widehat{w}_1(x) = 1 - \prod_{i \in \mathcal{N}(x)} \frac{p_*(x) - D_i}{p_*(x)} \tag{32}$$

and using that, under Assumption 1, we have $\mathbb{E}\left(D_i \,\middle|\, X^{(n)}\right) = p(X_i)$, we find that

$$\begin{aligned}\mathbb{E}\left[\widehat{w}_1(x) \,\middle|\, X^{(n)}\right] &= 1 - \prod_{i \in \mathcal{N}(x)} \frac{p_*(x) - p(X_i)}{p_*(x)} = 1 - \left(\frac{p_*(x) - p(x)}{p_*(x)}\right)^{q(x)} \\ &= w^{(q(x))}\big(p(x), p_*(x)\big),\end{aligned}$$

where we used that the set $\mathcal{N}(x)$ has $n(x) = q(x)$ elements, and the definition of the population weights in (24). Thus, $\widehat{w}_1(x)$ satisfies the desired result in (27). Finally, we can rewrite equation (32) as

$$\widehat{w}_1(x) := 1 - \left(\frac{p_*(x) - 1}{p_*(x)}\right)^{n_1(x)}, \tag{33}$$

which from now on will serve as our definition of $\widehat{w}_1(x)$ in the current case. By analogous arguments one obtains, for the current case of $q(x) = n(x)$ and $n(x)$ even, that

$$\widehat{w}_0(x) := 1 - \left(\frac{p_*(x)}{p_*(x) - 1}\right)^{n_0(x)},$$
$$\widehat{v}(x) := \frac{n_1(x)}{n(x)} - \left(\frac{p_*(x)}{p_*(x) - 1}\right)^{n_0(x)},$$

and one can easily verify that those expressions satisfy (27).

4.2.2 Case $q(x) = n(x)$ and $n(x)$ odd

For $q(x) = n(x)$ odd we have $w^{(q(x))}(p, p_*) = 1 - (1 - p)\left(\frac{p_* - p}{p_*}\right)^{q(x)-1}$ according to (24), and we then need to change (32) to

$$\widehat{w}_1(x) = 1 - \frac{1}{n(x)} \sum_{i \in \mathcal{N}(x)} (1 - D_i) \prod_{j \in \mathcal{N}(x) \setminus \{i\}} \frac{p_*(x) - D_j}{p_*(x)}. \tag{34}$$

Under Assumption 1, it is again easy to see that the approximate unbiasedness condition for $\widehat{w}_1(x)$ in (27) is satisfied here. In equation (34), the sum over i only gives a contribution for the $n(x) - n_1(x)$ instances where $D_i = 0$, in which case there still are $n_1(x)$ units $j \in \mathcal{N}(x) \setminus \{i\}$ with $D_j = 1$. We can therefore rewrite this equation as

$$\widehat{w}_1(x) := 1 - \frac{n(x) - n_1(x)}{n(x)} \left(\frac{p_*(x) - 1}{p_*(x)}\right)^{n_1(x)}, \tag{35}$$

which from now on is our definition of $\widehat{w}_1(x)$ for the case $q(x) = n(x)$ odd. By analogous arguments one obtains, for the current case, that

$$\widehat{w}_0(x) := 1 - \frac{n(x) - n_0(x)}{n(x)} \left(\frac{p_*(x)}{p_*(x) - 1}\right)^{n_0(x)},$$
$$\widehat{v}(x) := \frac{n_1(x)}{n(x)} - \frac{n(x) - n_0(x)}{n(x)} \left(\frac{p_*(x)}{p_*(x) - 1}\right)^{n_0(x)},$$

and one can again verify that those expressions satisfy (27).

4.2.3 General case $q(x) = \min\{Q, n(x)\}$

For $Q = \infty$ we have $q(x) = n(x)$, in which case all the required formulas for the sample weights are already provided in Subsections 4.2.1 and 4.2.2 above. The generalization to finite Q discussed in the following is not conceptually difficult, but it requires some combinatorial arguments. Remember that we choose the order $q(x)$ of the bounds according to (26). For even order $q(x) = q$, we generalize the formula for $\widehat{w}_1(x)$ in (32) as follows:

$$\widehat{w}_1(x) = 1 - \binom{n(x)}{q}^{-1} \sum_{\mathcal{S}_q} \prod_{i \in \mathcal{S}_q} \frac{p_*(x) - D_i}{p_*(x)}, \tag{36}$$

where the sum is over all subsets $\mathcal{S}_q \subset \mathcal{N}(x)$ with q elements. For odd order $q(x) = q$, we generalize the formula for $\widehat{w}_1(x)$ in (34) to

$$\widehat{w}_1(x) = 1 - \frac{1}{n(x)} \sum_{i \in \mathcal{N}(x)} (1 - D_i) \binom{n(x) - 1}{q - 1}^{-1} \sum_{\mathcal{S}_{q-1,i}} \prod_{j \in \mathcal{S}_{q-1,i}} \frac{p_*(x) - D_j}{p_*(x)}. \tag{37}$$

where the sum is over all subsets $\mathcal{S}_{q-1,i} \subset \mathcal{N}(x) \setminus \{i\}$ with $q - 1$ elements.

Under Assumption 1, it is again straightforward to verify that those formulas for $\widehat{w}_1(x)$ guarantee that $\mathbb{E}\left[\widehat{w}_1(x) \,\middle|\, X^{(n)}\right] = w^{(q(x))}\big(p(x), p_*(x)\big)$. If $q(x) < n(x)$, then alternative choices for the sample weight $\widehat{w}_1(x)$ exist that have the same conditional expectation – for example, instead of averaging over $\mathcal{S}_q$ and $\mathcal{S}_{q-1,i}$, one could randomly choose one subset of $q(x)$ observations out of the set $\mathcal{N}(x)$ and implement the formulas in Subsections 4.2.1 and 4.2.2 using only that subset of observations. To avoid that ambiguity in the definition of the sample weights we have chosen the formulas in (36) and (37) such that the binary treatment values D_i of all units $i \in \mathcal{N}(x)$ enter exchangeably into $\widehat{w}_1(x)$, that is, the sample weights remain unchanged if we swap the data of any two observations in the same cluster $\mathcal{N}(x)$. This requirement also guarantees that it is possible to rewrite $\widehat{w}_1(x)$ such that the D_i only enter through their summary statistics $n_1(x) = \sum_{i \in \mathcal{N}(x)} D_i$ and $n(x)$. Namely, one can rewrite (36) and (37) as

$$\widehat{w}_1(x) := 1 - \sum_{k=0}^{2\lfloor q(x)/2 \rfloor} \omega_{k,n_1(x),n(x),Q} \left(\frac{p_*(x) - 1}{p_*(x)} \right)^k, \tag{38}$$

where $\lfloor q(x)/2 \rfloor$ is the integer part of $q(x)/2$, and the combinatorial coefficients $\omega_{k,n_1(x),n(x),Q} \in [0, 1]$ are implicitly determined from (36) and (37), and one can show that

$$\omega_{k,n_1,n(x),Q} = \begin{cases} \dfrac{n(x) - n_1}{n(x)} \dbinom{n(x) - 1}{q - 1}^{-1} \dbinom{n_1}{k} \dbinom{n(x) - 1 - n_1}{q - 1 - k} & \text{if } q \text{ is odd,} \\ \dbinom{n(x)}{q}^{-1} \dbinom{n_1}{k} \dbinom{n(x) - n_1}{q - k} & \text{if } q \text{ is even,} \end{cases} \tag{39}$$

where $n_1 = n_1(x)$ and $q = q(x)$ also depend on x. Appendix B provides a derivation of this formula for $\omega_{k,n_1(x),n(x),Q}$. Implementing $\widehat{w}_1(x)$ via (38) and (39) is much faster than via (36) and (37), and can be done quickly also for relatively large values of $n(x)$ and $n_1(x)$.

Analogously we have

$$\widehat{w}_0(x) := 1 - \sum_{k=0}^{2\lfloor \min\{Q,n(x)\}/2 \rfloor} \omega_{k,n_0(x),n(x),Q} \left(\frac{p_*(x)}{p_*(x)-1} \right)^k,$$
$$\widehat{v}(x) := \frac{n_1(x)}{n(x)} - \sum_{k=0}^{2\lfloor \min\{Q,n(x)\}/2 \rfloor} \omega_{k,n_0(x),n(x),Q} \left(\frac{p_*(x)}{p_*(x)-1} \right)^k, \tag{40}$$

where the combinatorial coefficients $\omega_{k,n_0(x),n(x),Q} \in [0,1]$ are again those in (39), only the argument $n_1(x)$ was changed to $n_0(x)$. The equations in (38) and (40) provide general definitions of the sample weights that satisfy (27).

4.2.4 Discussion of the sample weights

We want to briefly discuss some properties of the sample weights, again mostly focusing on $\widehat{w}_1(x)$ for concreteness. If we choose $Q = \infty$, then the formula for $\widehat{w}_1(x)$ is given in (33) for even $n(x)$, and in (35) for odd $n(x)$. For $p_*(x) < \frac{1}{2}$ we have $\left|\frac{p_*(x)-1}{p_*(x)}\right| > 1$, implying that the absolute value of $\widehat{w}_1(x)$ grows exponentially with $n_1(x)$. Analogously, for $Q = \infty$ and $p_*(x) > \frac{1}{2}$ the absolute values of the weights $\widehat{w}_0(x)$ and $\widehat{v}(x)$ grow exponentially with $n_0(x)$. Only for $p_*(x) = 1/2$ are all the sample weights bounded, independent of the realization of $n_0(x)$ and $n_1(x)$.

Thus, for $Q = \infty$ the weights can take very large negative or positive values, potentially resulting in sample bounds for ATE and ATT with very large variance. This is the main reason why we introduce the bandwidth parameter Q, which in practice we recommend to set relative small, say $Q = 3$ or $Q = 4$. Once we have chosen a finite value of Q, then our sample weights in (38) and (40) are all bounded, independent of the realization of $n_0(x)$ and $n_1(x)$ — notice that the combinatorial coefficients $\omega_{k,n_{0/1}(x),n(x),Q}$ are all bounded between zero and one.

An interesting alternative way to guarantee that the weights $\widehat{w}_0(x)$ and $\widehat{w}_1(x)$ both remain bounded is to choose $Q = \infty$, but $p_*(x) = 1/2$ for all $x \in \overline{\mathcal{X}}$. That is not our leading recommendation, because in many applications one might prefer values of $p_*(x)$ different from $1/2$ to obtain better bounds. If the parameter of interest is ATT, then we can choose $Q = \infty$ and $\widehat{v}(x)$ will remain bounded as long as $p_*(x) \leq \frac{1}{2}$ for all $x \in \overline{\mathcal{X}}$. This could indeed be an interesting option in applications on ATT estimation. Nevertheless, the variance of the

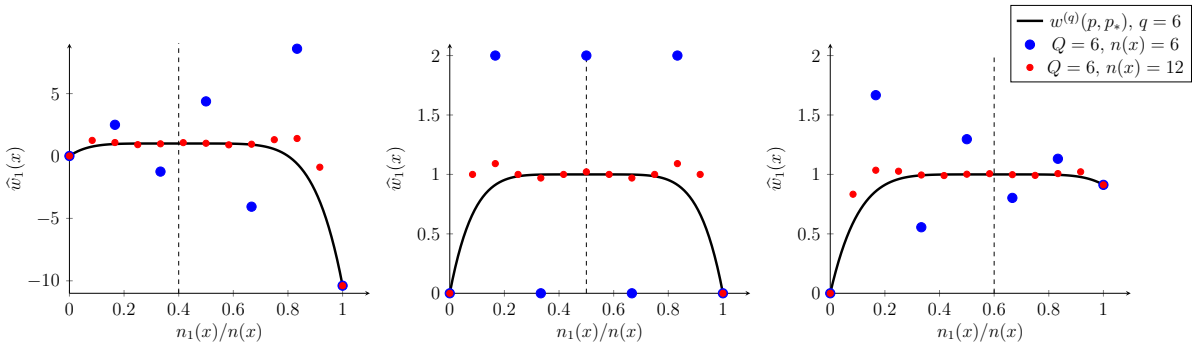

Figure 5: Sample Weights $\widehat{w}_1(x)$ plotted as a function of $n_1(x)/n(x)$ for $p_*(x) = 0.4$ (left), $p_*(x) = 0.5$ (middle) and $p_*(x) = 0.6$ (right). The corresponding population weights $w^{(q)}(p(x), p(x)_*)$ are also plotted as a function of $p(x)$.

bounds will usually be smaller when a finite value of Q is chosen. Furthermore, as illustrated in the following concrete examples for $\widehat{w}_1(x)$, only for finite Q do the sample weights converge to the population weights as $n(x) \to \infty$.

Figure 5 plots the weights $\widehat{w}_1(x)$ for $Q = 6$, $n(x) \in \{6, 12\}$, and for three different values for the reference propensity score $p_*(x)$. The plot shows that as $n(x)$ becomes large the weights $\widehat{w}_1(x)$ as a function of $\widehat{p}(x) = n_1(x)/n(x)$ converge to the population weights $w^{(q)}(p(x), p(x)_*)$ as a function of $p(x)$. This, in particular, implies that $\widehat{w}_1(x)$ becomes a smooth function of $n_1(x)$ for large values of $n(x)$. However, for small $n(x) = Q = 6$ the weights $\widehat{w}_1(x)$ heavily fluctuate as a function of $n_1(x)$. Furthermore, for $p_*(x) < 0.5$ the weights $\widehat{w}_1(x)$ can take on very small and very large values (notice the different scale of the plot for $p_*(x) = 0.4$), but for $p_*(x) \geq 0.5$ the weights remain within the bounded interval $[0, 2]$.

4.3 Asymptotically valid confidence intervals

Remember that $m = |\mathcal{X}_*|$ is the number of different covariate values in our sample. Our treatment effect bounds are then based on weight functions that combine the observed treatment status D_i for observations $i \in \mathcal{N}(x)$ of the same covariate value $x \in \mathcal{X}_*$ in a non-linear way. However, across covariate values $x \in \mathcal{X}_*$, the bounds are just averages across independent observations. Given that the bounds have this structure, it is useful to think of m as our effective sample size, and of each $x \in \mathcal{X}_*$ as labelling one effective observation. It is therefore convenient to rewrite the sample bounds in (29) not as cross-sectional averages over $i \in \{1, \ldots, n\}$, but as sample averages over $x \in \mathcal{X}_*$. For that purpose, for $d \in \{0, 1\}$

and $a \in \mathbb{R}$, we define[11]

$$\widehat{B}_x(d,a) := \frac{1}{n(x)} \sum_{i \in \mathcal{N}(x)} \widehat{B}_i(d,a), \qquad \widehat{C}_x(a) := \frac{1}{n(x)} \sum_{i \in \mathcal{N}(x)} \widehat{C}_i(a), \tag{41}$$

which allows us to rewrite the sample bounds in (29) as

$$\overline{B}(d,a) := \frac{1}{n} \sum_{x \in \mathcal{X}_*} n(x)\, \widehat{B}_x(d,a), \qquad \overline{C}(a) := \frac{1}{n} \sum_{x \in \mathcal{X}_*} n(x)\, \widehat{C}_x(a).$$

Using the definitions of $\widehat{B}_i(d,a)$ and $\widehat{C}_i(a)$ in (28) we furthermore have

$$\begin{aligned} \widehat{B}_x(d,a) &= a + \widehat{w}_d(x)\, \left[\overline{Y}_x(d) - a\right], \\ \widehat{C}_x(a) &= \frac{n_1(x)}{n(x)}\, \left[\overline{Y}_x(1) - a\right] - \widehat{v}(x)\, \left[\overline{Y}_x(0) - a\right], \end{aligned} \tag{42}$$

where

$$\overline{Y}_x(d) := \begin{cases} \dfrac{1}{n_d(x)} \displaystyle\sum_{i \in \mathcal{N}(x)} \mathbb{1}\{D_i = d\}\, Y_i & \text{if } n_d(x) > 0, \\ \mathbb{E}\left[Y(d) \,\middle|\, X = x\right] & \text{if } n_d(x) = 0. \end{cases}$$

Notice that for $n_d(x) = 0$ we have $\widehat{w}_d(x) = 0$, and for $n_0(x) = 0$ we have $\widehat{v}(x) = 0$. Therefore, $\overline{Y}_x(d)$ only enters into the bounds in (42) when $n_d(x) > 0$. In that case, $\overline{Y}_x(d)$ is simply the average of the $n_d(x)$ observed outcomes Y_i for which $X_i = x$ and $D_i = d$. However, for our theoretical discussion it is useful to also define $\overline{Y}_x(d)$ for the case $n_d(x) = 0$, because with that definition we have that, under Assumption 1,

$$\mathbb{E}\left[\overline{Y}_x(d) \,\middle|\, D^{(n)},\, X^{(n)}\right] = \mathbb{E}\left[Y(d) \,\middle|\, X = x\right] \tag{43}$$

Equation (43) states that $\overline{Y}_x(d)$ is mean-independent of $D^{(n)}$ and $X^{(n)}$. The properties of $\widehat{w}_{0/1}(x)$ and $\widehat{v}(x)$ in display (27) together with (43) guarantee that the expected values of $\widehat{B}_x(d,a)$ and $\widehat{C}_x(a)$ are equal to the expectations of the population bounds $B^{(q)}(0,a)$ and $C^{(q)}(a)$ in Section 3.

Next, we want to show consistency of those sample bounds and use them to construct confidence intervals. For that purpose it is convenient to define

$$\theta^{(0)} := \mathbb{E}\, Y(0), \qquad \theta^{(1)} := \mathbb{E}\, Y(1), \qquad \theta^{(\mathrm{ATE})} := \mathrm{ATE}, \qquad \theta^{(\mathrm{ATT})} := \mathrm{ATT}, \tag{44}$$

[11] We are slightly abusing notation here, for example, $\widehat{B}_x(d,a)$ for $x = 1$ (assuming $1 \in \mathcal{X}_*$) is not the same as $\widehat{B}_i(d,a)$ for $i = 1$. However, it will always be clear from the subscript letter which object is meant.

which are the four parameters of interest that we focus on in this paper. For each of those parameters we have already introduced upper and lower bound estimates in (29), (30), (31). For $\theta^{(0)}$ and $\theta^{(1)}$ we now denote those bounds by

$$\overline{L}^{(d)} := \overline{B}(d, a_{\min}), \qquad \overline{U}^{(d)} := \overline{B}(d, a_{\max}), \qquad \text{where } d \in \{0, 1\}.$$

Using the above definitions we have, for $r \in \{0, 1, \text{ATE}\}$,

$$\overline{L}^{(r)} = \frac{1}{m} \sum_{x \in \mathcal{X}_*} L_x^{(r)}, \qquad \overline{U}^{(r)} = \frac{1}{m} \sum_{x \in \mathcal{X}_*} U_x^{(r)},$$

where

$$\begin{aligned} L_x^{(d)} &:= \frac{m\, n(x)}{n}\, \widehat{B}_x(d, a_{\min}), \qquad & L_x^{(\text{ATE})} &:= \frac{m\, n(x)}{n} \left[\widehat{B}_x(1, a_{\min}) - \widehat{B}_x(0, a_{\max})\right], \\ U_x^{(d)} &:= \frac{m\, n(x)}{n}\, \widehat{B}_x(d, a_{\max}), \qquad & U_x^{(\text{ATE})} &:= \frac{m\, n(x)}{n} \left[\widehat{B}_x(1, a_{\max}) - \widehat{B}_x(0, a_{\min})\right], \end{aligned}$$

for $d \in \{0, 1\}$. When evaluating the asymptotic variance of $\overline{L}^{(\text{ATT})}$ and $\overline{U}^{(\text{ATT})}$ we also need to account for the randomness of the denominator term $\frac{1}{n}\sum_{i=1}^n D_i$, and we therefore write those bounds as (see appendix C for details)

$$\begin{aligned} \overline{L}^{(\text{ATT})} &= \frac{\mathbb{E}\overline{C}(a_{\max})}{\mathbb{E}(D)} + \frac{1}{m} \sum_{x \in \mathcal{X}_*} L_x^{(\text{ATT})} + O_P(1/m), \\ \overline{U}^{(\text{ATT})} &= \frac{\mathbb{E}\overline{C}(a_{\min})}{\mathbb{E}(D)} + \frac{1}{m} \sum_{x \in \mathcal{X}_*} U_x^{(\text{ATT})} + O_P(1/m), \end{aligned} \tag{45}$$

where

$$\begin{aligned} L_x^{(\text{ATT})} &:= \frac{m\, n(x)\, \widehat{C}_x(a_{\max})}{\sum_{i=1}^n D_i} - \frac{m\, n\, n_1(x)\, \overline{C}(a_{\max})}{\left(\sum_{i=1}^n D_i\right)^2}, \\ U_x^{(\text{ATT})} &:= \frac{m\, n(x)\, \widehat{C}_x(a_{\min})}{\sum_{i=1}^n D_i} - \frac{m\, n\, n_1(x)\, \overline{C}(a_{\min})}{\left(\sum_{i=1}^n D_i\right)^2}. \end{aligned} \tag{46}$$

Theorem 1. *Let Assumption 1 hold, and assume that while $m \to \infty$ both Q and $\max_{x \in \mathcal{X}_*} n(x)$ are bounded, and $p_*(x)$ is bounded away from zero and one, uniformly over $x \in \mathcal{X}_*$. Let $r \in \{0, 1, \text{ATE}, \text{ATT}\}$. Then we have:*

(i) The sample bounds are $\sqrt{m}$ consistent for their expectations:

$$\overline{L}^{(r)} = \mathbb{E}\,\overline{L}^{(r)} + O_P(m^{-1/2}), \qquad \overline{U}^{(r)} = \mathbb{E}\,\overline{U}^{(r)} + O_P(m^{-1/2}).$$

(ii) The sample bounds are asymptotically normally distributed:

$$\overline{L}^{(r)} - \mathbb{E}\,\overline{L}^{(r)} \Rightarrow \mathcal{N}\left(0, \mathrm{Var}\left(\overline{L}^{(r)}\right)\right), \qquad \overline{U}^{(r)} - \mathbb{E}\,\overline{U}^{(r)} \Rightarrow \mathcal{N}\left(0, \mathrm{Var}\left(\overline{U}^{(r)}\right)\right),$$

(iii) The variances of the sample bounds satisfy:

$$\mathrm{Var}\left(\overline{L}^{(r)}\right) \leq \frac{\mathrm{SVar}\left(L_x^{(r)}\right) + o_P(1)}{m}, \qquad \mathrm{Var}\left(\overline{U}^{(r)}\right) \leq \frac{\mathrm{SVar}\left(U_x^{(r)}\right) + o_P(1)}{m},$$

where for $M_x \in \{L_x^{(r)}, U_x^{(r)}\}$ *we have*

$$\mathrm{SVar}\left(M_x\right) := \frac{1}{m}\sum_{x\in\mathcal{X}_*} M_x^2 - \left(\frac{1}{m}\sum_{x\in\mathcal{X}_*} M_x\right)^2.$$

Based on Theorem 1 one can construct valid confidence intervals for $\mathbb{E}\,Y(d)$, the ATE, and the ATT. Let $\widehat{\sigma}_L^{(r)} := \sqrt{\mathrm{SVar}\left(L_x^{(r)}\right)}$, $\widehat{\sigma}_U^{(r)} := \sqrt{\mathrm{SVar}\left(U_x^{(r)}\right)}$, and for confidence level $(1-\alpha) \in (0,1)$, define the confidence interval[12]

$$\mathrm{CI}_{\mathrm{basic}}^{(r)} := \left[\overline{L}^{(r)} - \frac{\widehat{\sigma}_L^{(r)}}{\sqrt{m}}\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right),\ \overline{U}^{(r)} + \frac{\widehat{\sigma}_U^{(r)}}{\sqrt{m}}\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]. \tag{47}$$

The following corollary states that $\mathrm{CI}_{\mathrm{basic}}^{(r)}$ covers the true parameter of interest with probability at least $1-\alpha$ in large samples.

Corollary 1. *Let* $\alpha \in (0,1)$*. Under the assumptions of Theorem 1 we have*

$$\lim_{n\to\infty} \Pr\left(\theta^{(r)} \in \mathrm{CI}_{\mathrm{basic}}^{(r)}\right) \geq 1-\alpha,$$

for $r \in \{0, 1, \mathrm{ATE}, \mathrm{ATT}\}$*.*

Thus, those confidence intervals $\mathrm{CI}_{\mathrm{basic}}^{(r)}$ are asymptotically valid, but they may be conservative for three reasons: (i) the true $\theta^{(r)}$ may be an interior point of the expected bounds, implying 100% coverage in large samples; (ii) we are using an upper bound estimate for the variance of the upper and lower bounds when constructing the confidence interval, and (iii) we are using Bonferroni inequalities when dividing the statistical problem into one-sided confidence interval constructions for the upper and lower bound — notice the $\alpha/2$ in both the upper and lower bound in (47).[13]

[12] Here, we use the convention $[a,b] = \emptyset$ if $a > b$.

[13] One could improve on those $\alpha/2$ critical values by adapting the methods in Imbens and Manski (2004) and Stoye (2009) to our case. However, we want to keep the confidence interval construction simple here, and there is also the more important issue that $\mathrm{CI}_{\mathrm{basic}}^{(r)}$ can be empty in our case, which we address using Stoye (2020).

Here, the issues (i) and (iii) are very typical for bound estimation, and (ii) is impossible to fully overcome in our setting, unless $n_d(x)$ are sufficiently large for all d and x. For example, if $n_d(x) = 1$, then only a single outcome Y_i is observed for which we have $D_i = d$ and $X_i = \overline{x}$, implying that unbiased estimation of the variance of that outcome is impossible, but since Y_i enters into $\overline{L}^{(r)}$ and $\overline{U}^{(r)}$ we can in general not expect to estimate the variances of these bounds consistently.[14]

We therefore believe that one needs to be content with conservative confidence intervals in our setting, and that our construction so far has the advantage of being relatively simple and robust. However, a potentially more severe problem in practice is that the confidence interval $\mathrm{CI}_{\mathrm{basic}}$ may be empty, that is, the lower bound may be larger than the upper bound, because nothing in our construction guarantees that $\overline{L}^{(r)}$ cannot be larger than $\overline{U}^{(r)}$ at finite sample. While our theory guarantees that this problem cannot occur asymptotically, it is still undesirable to have a potentially empty confidence interval in applications.

We therefore use the method in Stoye (2020) to obtain a valid confidence interval that is never empty. The general version of that method requires knowing the correlation ρ between $\overline{L}^{(r)}$ and $\overline{U}^{(r)}$, which we cannot estimate consistently in our setting (for the same reasons for which we can only obtain upper bounds on the variances of $\overline{L}^{(r)}$ and $\overline{U}^{(r)}$). We therefore apply Stoye (2020)'s method with $\rho = 1$, which corresponds to the worst case: Let

$$\widehat{\theta}_*^{(r)} := \frac{\widehat{\sigma}_U^{(r)}\,\overline{L}^{(r)} + \hat{\sigma}_L\,\overline{U}^{(r)}}{\hat{\sigma}_L^{(r)} + \widehat{\sigma}_U^{(r)}}, \qquad\qquad \widehat{\sigma}_*^{(r)} := \frac{2\,\widehat{\sigma}_L^{(r)}\,\widehat{\sigma}_U^{(r)}}{\widehat{\sigma}_L^{(r)} + \widehat{\sigma}_U^{(r)}},$$

and

$$\mathrm{CI}_*^{(r)} := \left[\widehat{\theta}_*^{(r)} - \frac{\widehat{\sigma}_*^{(r)}}{\sqrt{m}}\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right),\ \widehat{\theta}_*^{(r)} + \frac{\widehat{\sigma}_*^{(r)}}{\sqrt{m}}\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right],$$

and define the final confidence interval to be reported for θ as the union of $\mathrm{CI}_{\mathrm{basic}}$ and CI_*, that is,

$$\mathrm{CI}_\theta^{(r)} := \mathrm{CI}_{\mathrm{basic}}^{(r)} \cup \mathrm{CI}_*^{(r)}.$$

Then, by construction, CI_θ is never empty, because CI_* is never empty, and Corollary 1 implies that

$$\lim_{n\to\infty} \Pr\left(\theta^{(r)} \in \mathrm{CI}_\theta^{(r)}\right) \geq 1-\alpha,$$

[14] Another problem is that the true propensity scores $p(x)$ are unknown, rendering the distribution of the sample weights $\widehat{w}_d(x)$ also unknown.

We refer to Stoye (2020) for a further justification of this specific confidence interval construction. We have thus shown how to construct valid non-empty confidence intervals for all of those objects of interest.

Notice also that for the constructions of confidence intervals here we have assumed that $p_*(x)$ is non-random. If $p_*(x)$ is estimated, then the randomness of $p_*(x)$ should be accounted for when constructing those confidence intervals, either via an application of the delta method, or via a bootstrap procedure.

5 Clustering the covariate observations

The unconfoundedness Assumption 1(i) only provides a restriction on the observed data $\{(Y_i, D_i, X_i) : i = 1, \ldots, n\}$ if at least two observations $i \neq j$ are available with the same covariate value $X_i = X_j$. However, if the difference $X_i - X_j$ is small, then we might be willing to ignore this difference and apply the unconfoundedness assumption as if X_i and X_j were equal to each other. Alternatively, if the propensity score is known (or estimable), then instead of finding matching covariate observations ($X_i \approx X_j$) it would be sufficient to find matching propensity score observations ($p(X_i) \approx p(X_j)$), but a key motivation for the current paper is exactly that the propensity might be unknown (or not reliably estimable), implying that dimensional reduction based on the propensity score is not feasible.

The problem of finding matching observations in covariate space to make use of unconfoundedness is, of course, not specific to our paper, and the technical contribution of our population bounds in Section 3 and their sample versions in Section 4 is indeed independent of this matching problem. It is nevertheless a problem that we need to address here, because of its obvious practical relevance in applying our bounds, but the reader should not expect any substantial novelty or contribution in our solution to this matching problem.

To be clear, all the results on the sample bounds given in the last section are applicable to the case where every observed covariate value X_i is unique in our sample. However, we then have $q(X_i) = n(X_i) = 1$, for all $i = 1, \ldots, n$, implying that we just implement first-order (Manski) bounds, which do not require unconfoundedness to be valid. Our bounds are only novel bounds if we have $n(X_i) > 1$ for some i.

If $n(X_i) = 1$ for all observations, then the simplest way in practice to still make use of unconfoundedness is to make the covariates coarser by coordinate wise binning. For example, if the k'th regressor $X_{i,k}$ is age in days, then we may replace it by the coarser measure age in years $\overline{X}_{i,k}$, which is a non-injective function of $X_{i,k}$. By this coarsening, the researcher decides

to discard information, but in an interpretable way that often makes it possible to judge whether the discarded information was relevant for the analysis or not. One might then find it plausible that Assumption 1 is satisfied with X_i replaced by $\overline{X}_i$, in which case our bounds and theoretical results can be applied to the sample $\{(Y_i, D_i, \overline{X}_i) : i = 1, \ldots, n\}$. Approximations of this kind are very common in practical applications, and the approximation error created by the binning is typically ignored (both for the bias and for the variance of the resulting treatment effect estimators). Binning is a conventional and transparent method that is driven by researchers decisions.

Alternatively, one can use more agnostic and automated methods to either cluster the individuals based on their covariates X_i or to apply nearest neighbor matching techniques. We focus on clustering in the following, because it corresponds more closely to our implementation of the bounds discussed in the last section, but in principle our bounds could equally be implemented using nearest neighbor matching. By clustering we mean that the observed sample of covariates $X = (X_1, \ldots, X_n)$ is used to partition the set of observations $\{1, \ldots, n\}$ into m partitions such that any two observations i and j in the same partition have similar covariate values $X_i \approx X_j$. Once the observations are clustered, then we again apply our bounds with X_i replaced by a label $\overline{X}_i$ of the cluster identity of unit i — specifically, we use the average covariate value within each cluster as our label $\overline{X}_i$.

If we generate the clusters based on the full covariate sample $X = (X_1, \ldots, X_n)$, then this generates dependence in the resulting sample $(Y_i, D_i, \overline{X}_i)$ across i, because the clustering procedure itself depends on all the observed covariates, that is, Assumption 1(iv) is only approximately satisfied after replacing X_i by $\overline{X}_i$. This technical problem could be overcome by standard arguments such as sample splitting methods, which would ensure that construction of the clusters is independent of subsequent estimation and inference. However, a proper theoretical analysis of our treatment effects bounds after clustering would either require assumptions on the existence of a true unobserved clustering structure of the covariates or (if we think of clustering as an approximation device in the spirit of Bonhomme, Lamadon and Manresa 2021) smoothness assumptions on $\mathbb{E}(D_i|X_i = x)$ and $\mathbb{E}(Y_i|X_i = x)$ in x. We do work out those statistical implications of the clustering in this paper, but instead leave those problems for future research. Again, this is because we think of this covariate approximation problem to be quite orthogonal to the main contribution of this paper described in the previous sections.

The specific clustering method that we employ in our simulations and empirical application below is as follows: We studentize each of the observed covariates, and afterwards

use the Euclidian distance $\|X_i - X_j\|$ as our measure of closeness of observation i and j. Using this distance measure we then apply hierarchical, agglomerative clustering with complete linkage to the observed covariate sample $(X_1, \ldots, X_n)$. We refer to, e.g., Kaufman and Rousseeuw (2005) and Everitt, Landau, Leese and Stahl (2011) for an introductory treatment to this clustering method and, e.g., Müllner (2013) and Maechler, Rousseeuw, Struyf, Hubert and Hornik (2021) for software implementation, respectively. Hierarchical, agglomerative clustering starts with clusters consisting of singletons and joins clusters stepwise until reaching the one common cluster. A desired number of clusters can be obtained by cutting a 'tree' that is produced by hierarchical clustering. One method of hierarchical, agglomerative clustering differs from another in terms of inter-cluster distances. The method of complete linkage uses the maximum distance between any pair of covariates, one in one cluster, one in the other, and tends to find compact clusters (Everitt, Landau, Leese and Stahl, 2011, Chapter 4).

The only additional tuning parameter that we need to choose when applying the clustering method is the number of clusters, which we denote $m \in \{1, 2, \ldots, n\}$, and which exactly takes the role of m, the number of unique covariate values, in the last section. We want to chose a large value of m to guarantee that the X_i's in each cluster are relatively close to each other (small approximation error), which fits well with our large m asymptotic theory in the last section. In practice, we recommend setting the number of clusters as

$$m = \left\lceil \frac{n}{L} \right\rceil \tag{48}$$

for some constant L, say $L = 10$, and where $\lceil \cdot \rceil$ is the ceiling function. This ad hoc choice of m provides about L observations in each cluster on average. Keeping L fixed implies that $m \to \infty$ as $n \to \infty$, in line with Section 4.3.

The clustering algorithm then delivers the partition $\{1, \ldots, n\} = \mathcal{N}_1 \cup \mathcal{N}_2 \cup \ldots \cup \mathcal{N}_m$. As mentioned earlier, we label clusters by their average covariate value, that is, for all $g \in \{1, \ldots, m\}$ and $i \in \mathcal{N}_g$ we define

$$\overline{X}_i := \frac{1}{|\mathcal{N}_g|} \sum_{j \in \mathcal{N}_g} X_j,$$

and we let $\overline{\mathcal{X}} = \left\{ \overline{X}_i \,:\, i \in \{1, \ldots, n\} \right\}$ be the set of all those cluster averages. The algorithm guarantees that no two clusters have the same average covariate value, implying that $\overline{X}_i$ uniquely identifies the cluster membership of observation i, and that $\left|\overline{\mathcal{X}}\right| = m$. For $\overline{x} \in \overline{\mathcal{X}}$ the corresponding cluster is denoted by

$$\mathcal{N}(\overline{x}) := \left\{ i \in \{1, \ldots, n\} \,\middle|\, \overline{X}_i = \overline{x} \right\}.$$

Let $n(\overline{x}) := |\mathcal{N}(\overline{x})|$ be number of observations with $\overline{X}_i = \overline{x}$, that is, the number of observations in that cluster. Notice that observations without any close covariate match will become their own cluster, that is, $n(\overline{x}) = 1$ is explicitly allowed for here.

Once those definitions are in place, then the construction of our sample bounds is exactly as described in Section 4, we just replace X_i by $\overline{X}_i$, $\mathcal{X}_*$ by $\overline{\mathcal{X}}$, $\mathcal{N}(x)$ by $\mathcal{N}(\overline{x})$, etc.

6 Monte Carlo Experiments

In this section, we report results of Monte Carlo experiments. The scalar covariate X is randomly generated from Unif$[-3, 3]$. The binary treatment variable D is then obtained from the following two models:

$$\text{(DGP A)} \quad \mathbb{E}[D|X] = p_0(X) = 0.5,$$
$$\text{(DGP B)} \quad \mathbb{E}[D|X] = p_0(X) = 0.75 \times \mathbb{1}\{X \geq 2\} + 0.5 \times \mathbb{1}\{|X| < 2\} + 1 \times \mathbb{1}\{X \leq -2\}.$$

To generate the outcome variable, define

$$Y_d^* = d + 1 - p_0(X) + V_d,$$

where $V_d \sim N(0, 1)$, $d \in \{0, 1\}$, and (V_1, V_0) are independent of (D, X). Finally, the observed outcome variable is generated by

$$Y = D\mathbb{1}\{Y_1^* > 0\} + (1 - D)\mathbb{1}\{Y_0^* > 0\}.$$

To study the effect of misspecification and the lack of overlap, we take $p_*(x) = 0.5$. That is, under DGP A, the model is correctly specified and the overlap condition is satisfied; whereas, under DGP B, the model is misspecified and the overlap condition is not satisfied. When $X \leq -2$, $p_0(X) = 1$ in DGP B. By simulation design, $a_{\min} = 0$ and $a_{\max} = 1$. In the Monte Carlo experiments, we focus on the ATT.

Define $\widehat{p} = n^{-1} \sum_{i=1}^n D_i$. We consider the following point estimators:

$$\widehat{\text{ATT}}_0 = (n\widehat{p})^{-1} \sum_{i=1}^n D_i \left[\mathbb{1}\{Y_{1i}^* > 0\} - \mathbb{1}\{Y_{0i}^* > 0\}\right],$$
$$\widehat{\text{ATT}}_* = (n\widehat{p})^{-1} \sum_{i=1}^n \left\{D_i - \frac{p_*(X_i)}{1 - p_*(X_i)}(1 - D_i)\right\} Y_i.$$

Here, $\widehat{\text{ATT}}_0$ is an infeasible oracle estimator of ATT, whereas $\widehat{\text{ATT}}_*$ is an estimator using the parametric propensity score $p_*(\cdot)$. We also consider the nearest neighbor estimator of

ATT:

$$\widehat{\mathrm{ATT}}_{\mathrm{NN}} = (n\widehat{p})^{-1} \sum_{i=1}^{n} D_i \left[Y_i - \widehat{Y}_{0i} \right],$$

where $\widehat{Y}_{0i}$ the nearest neighbor estimator of $\mathbb{E}[Y|X = X_i, D = 0]$.

Table 1 summarizes Monte Carlo results. The oracle estimator refers to $\widehat{\mathrm{ATT}}_0$, the reference propensity score (RPS) estimator is $\widehat{\mathrm{ATT}}_*$, NN is $\widehat{\mathrm{ATT}}_{\mathrm{NN}}$, and [LB$q$, UB$q$] corresponds to the q-order bound estimator using the method described in Section 4. The number m of clusters is chosen by (48) with $L = 10$. The sample size was $n = 1,000$ and the number of simulation replications was $1,000$.

In DGP A, the oracle, RPS, NN, LB2, UB2, LB3, and UB3 estimators all have almost the same mean and median. However, the Manski bounds (LB1 and UB1) are wide because the unconfounded assumption is not used in that case. In DGP B, the NN estimator ($\widehat{\mathrm{ATT}}_{\mathrm{NN}}$) does not work at all because the overlap condition is violated. In fact, the ATT is not point identified in this case. As a result, the standard deviation of the NN estimator is large. If we look at the RPS estimator that uses the misspecified propensity score, its mean is outside the average of our bound estimates. That is, 0.492 is larger than the averages of UB2 and UB3 (0.448 and 0.438). The average lower bound of LB3 is larger than that of LB2 but is smaller than the average of the oracle estimator. The simulation results from DGP B show that our approach does not require the overlap condition and improves the parametric estimator when it is misspecified. The Manski bounds are again much more conservative because they do not exploit the unconfoundedness assumption. Our bound estimators assume the unconfoundedness condition but not the overlap condition; hence, our bound approach can be viewed as a compromise between the point identified ATT under strong ignorability and Manski's worst case bounds.

6.1 Additional Monte Carlo Experiments: Inference

In this subsection, we report additional Monte Carlo experiments that focus on finite sample performance of our proposed methods. We consider both continuous and discrete X. The former is randomly drawn from Unif$[-3, 3]$) and the latter is generated by $X = \mathrm{round}(10 \times \mathrm{Unif}[-3, 3])/10$. That is, X is a discrete uniform random variable on the discrete support $[-3, -2.9, \ldots, 2.9, 3]$. The rest of the simulation design is the same as before, and we focus on ATT as well.

Panels I and II in Table 2 summarizes the results of Monte Carlo experiments when the distribution of X is discrete. In the columns heading 'Coverage', we report the Monte Carlo

Table 1: Monte Carlo Results

	Mean	Median	St.Dev.	Min	Max
DGP A					
Oracle	0.244	0.243	0.024	0.143	0.310
RPS	0.241	0.244	0.052	0.084	0.400
NN	0.240	0.239	0.034	0.143	0.365
LB1	-0.066	-0.065	0.012	-0.108	-0.034
UB1	0.932	0.934	0.012	0.892	0.964
LB2	0.247	0.247	0.030	0.156	0.337
UB2	0.239	0.237	0.033	0.145	0.358
LB3	0.248	0.247	0.029	0.146	0.337
UB3	0.240	0.239	0.030	0.157	0.353
DGP B					
Oracle	0.281	0.281	0.022	0.185	0.352
RPS	0.492	0.493	0.034	0.370	0.587
NN	0.236	0.168	0.131	0.041	0.495
LB1	-0.069	-0.069	0.012	-0.113	-0.032
UB1	0.901	0.902	0.012	0.859	0.935
LB2	0.238	0.238	0.026	0.164	0.325
UB2	0.448	0.448	0.030	0.350	0.557
LB3	0.273	0.274	0.026	0.198	0.363
UB3	0.438	0.438	0.029	0.340	0.526

Notes: The oracle estimator refers to the infeasible estimator using observations $\mathbb{1}\{Y_{1i}^* > 0\}$ and $\mathbb{1}\{Y_{0i}^* > 0\}$, the RPS estimator is the estimator using the reference propensity score $p_*(x) = 0.5$, NN is the nearest neighbor estimator of ATT, [LBq, UBq] corresponds to the q-order bound estimator. The sample size was $n = 1,000$ and the number of simulation replications was $1,000$.

coverage proportion that the true value of ATT is included in either sample analog bounds or inference bounds.In the columns heading 'Non-Empty Interval', we report the Monte Carlo proportion of the cases that the resulting interval is non-empty. In the columns heading 'Avg. Length', we show the average length of the confidence interval when it is not empty. The inference bounds are constructed by applying the method described in Section 4.3. We first discuss the results for DGP A. In this scenario, the ATT is point-identified and the lower bound equals the upper bound; thus, the sample lower bound can be easily larger than the sample upper bound, resulting in frequent occurrence of empty intervals. However, the inference bounds are never empty and provides good coverage results. In DGP B, there is no surprising result. The bounds are wide enough to cover the true value in every Monte Carlo repetition. This is because the ATT is only partially identified in DGP B.

Panels III and IV in Table 2 summarizes the results of Monte Carlo experiments when the distribution of X is continuous. Overall, the results are similar to the discrete X case for DGP A. However, there is a rather surprising result with $Q = 4$ for DGP B. In this case, the inference bounds include the true value only 339 out 1000. This suggests that the clustering estimators with a large value of Q may lead to severe estimation bias and size distortion, possibly due to the bias from the clustering method.

7 An Empirical Example

In this section, we apply our methods to Connors et al. (1996)'s study of the efficacy of right heart catheterization (RHC), which is a diagnostic procedure for directly measuring cardiac function in critically ill patients. This dataset has been subsequently used in the context of limited overlap by Crump, Hotz, Imbens and Mitnik (2009), Rothe (2017), and Li, Morgan and Zaslavsky (2018) among others. The dataset is publicly available on the Vanderbilt Biostatistics website at `https://hbiostat.org/data/`.

In this example, the dependent variable is 1 if a patient survived after 30 days of admission, and 0 if a patient died within 30 days. The binary treatment variable is 1 if RHC was applied within 24 hours of admission, and 0 otherwise. The sample size was $n = 5735$, and 2184 patients were treated with RHC. There are a large number of covariates: Hirano and Imbens (2001) constructed 72 variables from the dataset and the same number of covariates were considered in both Crump, Hotz, Imbens and Mitnik (2009) and Li, Morgan and Zaslavsky (2018) and 50 covariates were used in Rothe (2017). In our exercise, we constructed the same 72 covariates. For the purpose of illustrating our methodology, we assume that the

Table 2: Monte Carlo Results: Inference

Q	Coverage		Non-Empty Interval		Avg. Length	
	Sample Analogs	Inference Bounds	Sample Analogs	Inference Bounds	Sample Analogs	Inference Bounds
Panel I. DGP A with a Discrete Covariate						
1	1.000	1.000	1.000	1.000	1.000	1.379
2	0.124	0.976	0.449	1.000	0.019	0.129
3	0.067	0.968	0.408	1.000	0.010	0.118
4	0.031	0.969	0.445	1.000	0.006	0.115
Panel II. DGP B with a Discrete Covariate						
1	1.000	1.000	1.000	1.000	0.982	1.293
2	0.999	1.000	1.000	1.000	0.274	0.469
3	0.996	1.000	1.000	1.000	0.238	0.428
4	0.511	0.992	1.000	1.000	0.151	0.343
Panel III. DGP A with a Continuous Covariate						
1	1.000	1.000	1.000	1.000	0.998	1.251
2	0.090	0.985	0.371	1.000	0.021	0.141
3	0.043	0.978	0.270	1.000	0.011	0.128
4	0.053	0.981	0.299	1.000	0.014	0.141
Panel IV. DGP B with a Continuous Covariate						
1	1.000	1.000	1.000	1.000	0.970	1.185
2	0.940	0.999	1.000	1.000	0.210	0.379
3	0.529	0.993	1.000	1.000	0.164	0.334
4	0.003	0.349	0.963	1.000	0.046	0.242

Notes: The nominal coverage probability is 0.95. The sample size was $n = 1,000$ and the number of simulation replications was $1,000$.

unconfoundedness assumption holds in this example.[15]

In this section, we focus on ATT. We first estimate ATT by the normalized inverse probability weighted estimator[16]:

$$\widehat{\text{ATT}}_{\text{PS}} := \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i} - \frac{\sum_{i=1}^n (1-D_i) W_i Y_i}{\sum_{i=1}^n (1-D_i) W_i},$$

where $W_i := \widehat{p}(X_i)/[1-\widehat{p}(X_i)]$ and $\widehat{p}(X_i)$ is the estimated propensity score for observation i based on a logit model with all 72 covariates being added linearly as in the aforementioned papers. The estimator $\widehat{\text{ATT}}_{\text{PS}}$ requires that the assumed propensity score model is correctly specified and the overlap condition is satisfied. The resulting estimate is $\widehat{\text{ATT}}_{\text{PS}} = -0.0639$.[17]

We now turn to our methods. We take the reference propensity score to be $\widehat{p}_{\text{RPS}}(X_i) = n^{-1}\sum_{i=1}^n D_i$ for each observation i. That is, we assign the sample proportion of the treated to the reference propensity scores uniformly for all observations. Of course, this is likely to be misspecified; however, it has the advantage that $1/\widehat{p}_{\text{RPS}}(X_i)$ is never close to 0 or 1. The resulting inverse reference-propensity-score weighted ATT estimator is[18]

$$\widehat{\text{ATT}}_{\text{RPS}} := \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i} - \frac{\sum_{i=1}^n (1-D_i) Y_i}{\sum_{i=1}^n (1-D_i)} = -0.0507.$$

None of the covariate values in the observed sample are identical among patients (that is, $n(X_i) = 1$ for all observations here). We therefore implement the clustering method described in Section 5. As recommended in Section 5, we choose the number m of clusters by (48): $m = \lceil \frac{n}{L} \rceil$ with $L = 5, 10, 20$. In addition, we consider $Q = 1, \ldots, 4$.

Table 3 reports estimation results of ATT bounds for selected values of L and Q. When $Q = 1$, our estimated bounds correspond to Manski bounds, which includes zero and is wide with the interval length of almost one in all cases of L. Our bounds with $Q = 1$ are different across L because we apply hierarchical clustering before obtaining Manski bounds. With $Q = 2$, the bounds shrink so that the estimated upper bound is zero for all cases of L; with $Q = 3$, they shrink even further so that the upper end point of the 95% confidence

[15] Bhattacharya, Shaikh and Vytlacil (2008, 2012) raise the concern that catheterized and noncatheterized patients may differ on unobserved dimensions and propose different bounds using a day of admission as an instrument for RHC.

[16] See, e.g., equation (3) and discussions in Busso, DiNardo and McCrary (2014) for details of the normalized inverse probability weighted ATT estimator.

[17] The unnormalized ATT estimate is -0.0837 using the same propensity scores.

[18] When the sample proportion is used as the propensity score estimator, there is no difference between unnormalized and normalized versions of ATT estimates. In fact, it is simply the mean difference between treatment and control groups.

Table 3: ATT Bounds: Right Heart Catheterization Study

L	Q	LB	UB	CI-LB	CI-UB
5	1	-0.638	0.282	-0.700	0.330
	2	-0.131	-0.000	-0.174	0.033
	3	-0.034	-0.048	-0.076	-0.007
	4	-0.006	-0.073	-0.079	-0.006
10	1	-0.664	0.307	-0.766	0.376
	2	-0.169	0.004	-0.216	0.039
	3	-0.077	-0.039	-0.117	-0.006
	4	-0.049	-0.057	-0.090	-0.016
20	1	-0.675	0.316	-0.843	0.430
	2	-0.178	-0.005	-0.238	0.034
	3	-0.099	-0.046	-0.149	-0.007
	4	-0.065	-0.060	-0.112	-0.017

Notes: LB and UB correspond to the lower and upper bound estimates, where CI-LB and CI-UB represent the lower and upper 95% confidence interval estimates. Estimates are shown for selected values of $L = 5, 10, 20$ and $Q = 1, \ldots, 4$.

interval excludes zero. Among three different values of L, the case of $L = 5$ gives the tightest confidence interval but in this case, the lower bound is larger than the upper bound, indicating that the estimates might be biased. In view of that, we take the bound estimates with $L = 10$ as our preferred estimates $[-0.077, -0.039]$ with the 95% confidence interval $[-0.117, 0.006]$. When $Q = 4$, the lower bound estimates exceed the upper bound estimates with $L = 5, 10$. However, the estimates with $L = 20$ give an almost identical confidence interval to our preferred estimates. It seems that the pairs of $(L, Q) = (10, 3)$ and $(L, Q) = (20, 4)$ provide reasonable estimates.

The study of Connors et al. (1996) offered a conclusion that RHC could cause an increase in patient mortality. Based on our preferred estimates, we can exclude large beneficial effects with confidence. This conclusion is based solely on the unconfoundedness condition, but not on the overlap condition, nor on the correct specification of the logit model. Overall, our estimates seem to be consistent with the qualitative findings in Connors et al. (1996) under the maintained assumption that the unconfoundedness assumption holds.

References

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Abadie, A. and G. W. Imbens (2008). On the failure of the bootstrap for matching estimators. *Econometrica 76*(6), 1537–1557.

Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Armstrong, T. B. and M. Kolesár (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica 89*(3), 1141–1177.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2008). Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization. *American Economic Review: Papers and Proceedings 98*(2), 351–56.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012). Treatment effect bounds: An application to Swan-Ganz catheterization. *Journal of Econometrics 168*(2), 223–243.

Bonhomme, S., T. Lamadon, and E. Manresa (2021). Discretizing unobserved heterogeneity. *Econometrica*. Forthcoming.

Busso, M., J. DiNardo, and J. McCrary (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics 96*(5), 885–897.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: estimation and inference. *Econometrica 81*(2), 667–737.

Connors, Alfred F., J., T. Speroff, N. V. Dawson, C. Thomas, J. Harrell, Frank E., D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, J. Fulkerson, William J., H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus (1996). The Effectiveness of Right Heart Catheterization in the Initial Care of Critically III Patients. *JAMA 276*(11), 889–897.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

D'Amour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics 221*(2), 644–654.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011). *Cluster Analysis* (5th ed.). John Wiley & Sons.

Hirano, K. and G. W. Imbens (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology 2*(3-4), 259–278.

Hong, H., M. P. Leung, and J. Li (2019). Inference on finite–population treatment effects under limited overlap. *Econometrics Journal 23*, 32–47.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kaufman, L. and P. J. Rousseeuw (2005). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Khan, S. and E. Tamer (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica 78*(6), 2021–2042.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association 113*(521), 390–400.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik (2021). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.1. `https://CRAN.R-project.org/package=cluster`.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human resources*, 343–360.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review 80*(2), 319–323.

Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software 53*(9), 1–18.

Nethery, R. C., F. Mealli, and F. Dominici (2019). Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *Annals of Applied Statistics 13*(2), 1242–1267.

Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica 85*(2), 645–660.

Sasaki, Y. and T. Ura (2021). Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory*. `https://doi.org/10.1017/S0266466621000025`, forthcoming.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics 8*(6), 1348–1360.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica 77*(4), 1299–1315.

Stoye, J. (2020). A simple, short, but never-empty confidence interval for partially identified parameters. arXiv:2010.10484, [econ.EM], `https://arxiv.org/abs/2010.10484`.

Yang, S. and P. Ding (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika 105*(2), 487–493.

A Proofs for Section 3

A.1 Proofs of the main text results in Section 3

Proof of Lemma 1. Let $\lambda_0(1,x), \lambda_1(1,x) \in \mathbb{R}$ be such that the conditions stated in the lemma are satisfied. Under Assumption 1(i) we have

$$\mathbb{E}\left[B^{(2)}(1,a_{\min}) \,\middle|\, X=x\right] = a_{\min} + \left[\lambda_0(1,x)\,p(x) + \lambda_1(1,x)\,p^2(x)\right]\mathbb{E}\left[Y(1) - a_{\min} \,\middle|\, X=x\right].$$

By Assumption 1(ii) we must have $\mathbb{E}\left[Y(1) - a_{\min} \,\middle|\, X=x\right] \geq 0$. Now, consider a DGP for which $\mathbb{E}\left[Y(1) - a_{\min} \,\middle|\, X=x\right] > 0$ for a particular value of x. Then, the requirement that $\mathbb{E}\left[B^{(2)}(1,a_{\min}) \,\middle|\, X=x\right] \leq \mathbb{E}\left[Y(1) \,\middle|\, X=x\right]$ for that DGP can equivalently be written as

$$\lambda_0(1,x)\,p(x) + \lambda_1(1,x)\,p^2(x) \leq 1, \tag{A.1}$$

which needs to hold for all $x \in \mathcal{X}$, and for DGP's with arbitrary values $p(x) \in [0,1]$. For every $x \in \mathcal{X}$ we now consider three possible cases:

Case 1: Consider the case where (A.1) holds with equality for some value $p_*(x) \in [0,1)$, that is,

$$\lambda_0(1,x)\, p_*(x) + \lambda_1(1,x)\, p_*^2(x) = 1. \tag{A.2}$$

We must have $p_*(x) \neq 0$, because otherwise (A.2) is violated. Furthermore, (A.1) still needs to hold for all $p(x) \in [0,1]$, which implies that the polynomial $p \mapsto \lambda_0(1,x)\, p + \lambda_1(1,x)\, p^2$ is maximized at $p_*(x)$. Since $p_*(x) \in (0,1)$ this maximum does not appear at the boundary, and therefore the FOC of the maximization problem must hold, which read

$$\lambda_0(1,x) + 2\,\lambda_1(1,x)\, p_*(x) = 0. \tag{A.3}$$

Solving (A.2) and (A.3) for $\lambda_{0/1}(1,x)$ gives

$$\lambda_0(1,x) = \frac{2}{p_*(x)}, \qquad\qquad \lambda_1(1,x) = -\frac{1}{[p_*(x)]^2},$$

which is the conclusion of the lemma.

Case 2: Consider the case where (A.1) holds with equality for $p_*(x) = 1$, that is,

$$\lambda_0(1,x) + \lambda_1(1,x) = 1. \tag{A.4}$$

If $\lambda_1(1,x) = -1$, then we have $\lambda_0(1,x) = -2$ and the conclusion of the lemma is satisfied. If $\lambda_1(1,x) < -1$, then (A.1) is violates for $p(x) = 1 - \epsilon$, with $\epsilon > 0$ sufficiently small, and this possibility is therefore ruled out by our assumptions.

Finally, if $\lambda_1(1,x) > -1$, then we consider the alternative coefficients

$$\lambda_0^*(1,x) := 2, \qquad\qquad \lambda_1^*(1,x) := -1.$$

One can easily verify that $\lambda_0^*(1,x)\, p(x) + \lambda_1^*(1,x)\, p^2(x) \leq 1$, for all $p(x) \in [0,1]$, which implies that the alternative bounds

$$B^*(d,a) := a + [\lambda_0^*(d,X) + \lambda_1^*(d,X)\, p(X)]\ \mathbb{1}\{D = d\}\ (Y - a), \quad d \in \{0,1\},$$

are valid bounds, in the sense that they satisfy (11). Furthermore, we have

$$\begin{aligned}
&\lambda_0(1,x)\, p(x) + \lambda_1(1,x)\, p^2(x) < \lambda_0^*(1,x)\, p(x) + \lambda_1^*(1,x)\, p^2(x), && \text{for } p(x) \in (0,1),\\
&\lambda_0(1,x)\, p(x) + \lambda_1(1,x)\, p^2(x) = \lambda_0^*(1,x)\, p(x) + \lambda_1^*(1,x)\, p^2(x), && \text{for } p(x) \in \{0,1\},
\end{aligned}$$

which implies that (14) holds for all DGPs that satisfy Assumption 1(i) and (ii), and where the first inequality in (14) is strict for all DGPs with $p(x) \in (0,1)$ and $\mathbb{E}\big[Y(1) - a_{\min} \,\big|\, X = x\big] > 0$. This implies that $\lambda_0(1,x)$ and $\lambda_1(1,x)$ are not admissible, which violates the assumptions of the lemma, and $\lambda_1(1,x) > -1$ is therefore ruled out by our assumptions.

Case 3: Consider the case where (A.1) never holds with equality for any $p(x) \in [0,1]$. We define

$$p_{\max}(x) := \operatorname*{argmax}_{p \in [0,1]} \left[\lambda_0(1,x)\, p + \lambda_1(1,x)\, p^2\right].$$

If $\lambda_1(1,x) < 0$ and $\lambda_0(1,x) \in [0, -2\,\lambda_1(1,x)]$, then we have $p_{\max}(x) = -\frac{\lambda_0(1,x)}{2\,\lambda_1(1,x)}$. Otherwise we have a boundary solution, either $p_{\max}(x) = 0$ or $p_{\max}(x) = 1$. We furthermore define

$$p_*(x) := \begin{cases} p_{\max}(x) & \text{if } p_{\max}(x) > 0, \\ 1 & \text{if } p_{\max}(x) = 0, \end{cases}$$

and we consider the alternative coefficients

$$\lambda_0^*(1,x) := \frac{2}{p_*(x)}, \qquad\qquad \lambda_1^*(1,x) := -\frac{1}{[p_*(x)]^2}.$$

One can easily verify that

$$\lambda_0^*(1,x)\, p(x) + \lambda_1^*(1,x)\, p^2(x) \leq 1,$$

for all $p(x) \in [0,1]$. This implies that the alternative bounds

$$B^*(d,a) := a + \left[\lambda_0^*(d,X) + \lambda_1^*(d,X)\, p(X)\right] \mathbb{1}\{D = d\}\ (Y - a), \quad d \in \{0,1\},$$

are valid bounds, in the sense that they satisfy (11).

Furthermore, we have

$$\lambda_0(1,x)\, p(x) + \lambda_1(1,x)\, p^2(x) \leq \lambda_0^*(1,x)\, p(x) + \lambda_1^*(1,x)\, p^2(x),$$

for all $p(x) \in [0,1]$. This implies that (14) holds for all DGPs that satisfy Assumption 1(i) and (ii), that is, the alternative bounds never perform worse in expectation than the original bounds.

Finally, we are considering the case where $\lambda_0(1,x)\, p(x) + \lambda_1(1,x)\, p^2(x) < 1$ for all $p(x)$, and by construction we have $\lambda_0^*(1,x)\, p_*(x) + \lambda_1^*(1,x)\, p_*^2(x) = 1$. This implies that

$$\lambda_0(1,x)\, p_*(x) + \lambda_1(1,x)\, p_*^2(x) < \lambda_0^*(1,x)\, p_*(x) + \lambda_1^*(1,x)\, p_*^2(x).$$

Therefore, the first inequality in (14) is strict for DGPs with $p(x) = p_*(x)$ and $\mathbb{E}\big[Y(1) - a_{\min} \,\big|\, X = x\big] > 0$.

This implies that $\lambda_0(1, x)$ and $\lambda_1(1, x)$ are not admissible, which violates the assumptions of the lemma. Thus, the current case is ruled out be the assumptions of the lemma and need not be considered further. ■

Proof of Proposition 1. This proposition is the special case $q = 2$ of part (i) and (ii) of Proposition 2. We therefore refer to the proof of Proposition 2 below. ■

Before presenting the proof of Proposition 2 it is useful to provide two intermediate lemmas. Those lemmas explain the properties of the weight functions $w^{(q)}(p, p_*)$ and $\widetilde{w}^{(q)}(p, p_*)$ that were defined in the main text, and are crucial for the proof of part (iii) of Proposition 2.

Lemma 2. *Let $q \in \{1, 2, \ldots\}$. For $\lambda = (\lambda_0, \ldots, \lambda_{q-1}) \in \mathbb{R}^q$ and $p \in [0, 1]$ we define $v(p, \lambda) := \sum_{r=0}^{q-1} \lambda_r \, p^{r+1}$, and for $p \in (0, 1]$ we define $\widetilde{v}(p, \lambda) := \sum_{r=0}^{q-1} \lambda_r \, p^{r-1} \, (1 - p)$. Let $p_* \in (0, 1)$. Then, the functions $w^{(q)}(p, p_*)$ and $\widetilde{w}^{(q)}(p, p_*)$ defined in (24) are the unique solutions to the following optimization problems.*

(i) The solution to the optimization problem

$$\overline{\lambda} = \underset{\lambda \in \mathbb{R}^q}{\operatorname{argmin}} \left| \frac{\partial^{q-1} v(p_*, \lambda)}{\partial^q p} \right| \qquad \begin{aligned} &\textit{subject to} \quad && v(p_*, \lambda) = 1, \\ &\textit{and} && \frac{\partial^k v(p_*, \lambda)}{\partial^k p} = 0, \;\; \textit{for } k \in \{1, \ldots, q-2\}, \\ &\textit{and} && v(p, \lambda) \leq 1, \;\; \textit{for } p \in [0, 1], \end{aligned}$$

satisfies

$$v(p_*, \overline{\lambda}) = w^{(q)}(p, p_*).$$

(ii) The solution to the optimization problem

$$\widetilde{\lambda} = \underset{\lambda \in \mathbb{R}^q}{\operatorname{argmin}} \left| \frac{\partial^{q-1} \widetilde{v}(p_*, \lambda)}{\partial^q p} \right| \qquad \begin{aligned} &\textit{subject to} \quad && \widetilde{v}(p_*, \lambda) = 1, \\ &\textit{and} && \frac{\partial^k \widetilde{v}(p_*, \lambda)}{\partial^k p} = 0, \;\; \textit{for } k \in \{1, \ldots, q-2\}, \\ &\textit{and} && \widetilde{v}(p, \lambda) \leq 1, \;\; \textit{for } p \in (0, 1], \end{aligned}$$

satisfies

$$\widetilde{v}(p_*, \widetilde{\lambda}) = \widetilde{w}^{(q)}(p, p_*).$$

The proof of Lemma 2 is provided in Appendix A.2. For the statement of the next lemma, remember that for $p_* \in (0,1)$ and $\epsilon > 0$ we defined $\mathcal{B}_\epsilon(p_*)$ to be the ϵ-ball around p_*.

Lemma 3. *Let $q \in \{1,2,\ldots\}$ and $p_* \in (0,1)$. For $\lambda = (\lambda_0, \ldots, \lambda_{q-1}) \in \mathbb{R}^q$ let $v(p,\lambda)$ and $\widetilde{v}(p,\lambda)$ be as defined in Lemma 2.*

(i) Let $\lambda \in \mathbb{R}^q$ be such for all $p \in [0,1]$ we have $v(p,\lambda) \leq 1$. Then, there exists $\epsilon > 0$ such that for all $p \in \mathcal{B}_\epsilon(p_*)$ we have

$$v(p,\lambda) \leq w^{(q)}(p, p_*).$$

(ii) Let $\lambda \in \mathbb{R}^q$ be such for all $p \in (0,1]$ we have $\widetilde{v}(p,\lambda) \leq 1$. Then, there exists $\epsilon > 0$ such that for all $p \in \mathcal{B}_\epsilon(p_*)$ we have

$$\widetilde{v}(p,\lambda) \leq \widetilde{w}^{(q)}(p, p_*).$$

The proof of Lemma 3 is provided in Appendix A.2.

Proof of Proposition 2. # Part (i): Under Assumption 1(i) we find for the bounds defined in (23) that

$$\begin{aligned}
\mathbb{E}\left[B^{(q)}(0,a) - a \,\middle|\, X = x\right] &= w^{(q)}\big(1-p(x), 1-p_*(x)\big)\mathbb{E}\left[Y(0) - a \,\middle|\, X = x\right], \\
\mathbb{E}\left[B^{(q)}(1,a) - a \,\middle|\, X = x\right] &= w^{(q)}\big(p(x), p_*(x)\big)\ \mathbb{E}\left[Y(1) - a \,\middle|\, X = x\right], \\
\mathbb{E}\left[C^{(q)}(a) \,\middle|\, X = x\right] &= p(x)\Big\{\mathbb{E}\left[Y(1) - a \,\middle|\, X = x\right] \\
&\qquad - \widetilde{w}^{(q)}\big(p(x), p_*(x)\big)\,\mathbb{E}\left[Y(0) - a \,\middle|\, X = x\right]\Big\}.
\end{aligned}$$

From the definition of the weight functions in (24) we have

$$w^{(q)}\big(1-p(x), 1-p_*(x)\big) \leq 1, \qquad \widetilde{w}^{(q)}\big(1-p(x), 1-p_*(x)\big) \leq 1.$$

Assumption 1(ii) guarantees that, for $d \in \{0,1\}$,

$$\mathbb{E}\left[Y(d) - a_{\min} \,\middle|\, X = x\right] \geq 0, \qquad \mathbb{E}\left[Y(d) - a_{\max} \,\middle|\, X = x\right] \leq 0.$$

Combining the results in the last three displays we find that

$$\begin{aligned}
\mathbb{E}\left[B^{(q)}(d, a_{\min}) - a_{\min} \,\middle|\, X = x\right] &\leq \mathbb{E}\left[Y(d) - a_{\min} \,\middle|\, X = x\right], \\
\mathbb{E}\left[B^{(q)}(d, a_{\max}) - a_{\max} \,\middle|\, X = x\right] &\geq \mathbb{E}\left[Y(d) - a_{\max} \,\middle|\, X = x\right],
\end{aligned}$$

and therefore

$$\mathbb{E}\left[B^{(q)}(d, a_{\min}) \,\middle|\, X\right] \leq \mathbb{E}\left[Y(d) \,\middle|\, X\right] \leq \mathbb{E}\left[B^{(q)}(d, a_{\min}) \,\middle|\, X\right]. \tag{A.5}$$

Taking the expectation over X gives the results of part (i)(a) of the proposition, and part (i)(b) immediately follows from that.

Similarly, we find

$$\begin{aligned}
\mathbb{E}\left[C^{(q)}(a_{\min}) \,\middle|\, X = x\right] &\geq p(x)\, \left\{\mathbb{E}\left[Y(1) - a_{\min} \,\middle|\, X = x\right] - \mathbb{E}\left[Y(0) - a_{\min} \,\middle|\, X = x\right]\right\} \\
&= p(x)\, \mathbb{E}\left[Y(1) - Y(0) \,\middle|\, X = x\right], \\
\mathbb{E}\left[C^{(q)}(a_{\max}) \,\middle|\, X = x\right] &\leq p(x)\, \left\{\mathbb{E}\left[Y(1) - a_{\max} \,\middle|\, X = x\right] - \mathbb{E}\left[Y(0) - a_{\max} \,\middle|\, X = x\right]\right\} \\
&= p(x)\, \mathbb{E}\left[Y(1) - Y(0) \,\middle|\, X = x\right],
\end{aligned}$$

and therefore

$$\frac{\mathbb{E}\left[C^{(q)}(a_{\max}) \,\middle|\, X = x\right]}{\mathbb{E}(D)} \leq \tau(x) \leq \frac{\mathbb{E}\left[C^{(q)}(a_{\min}) \,\middle|\, X = x\right]}{\mathbb{E}(D)}, \tag{A.6}$$

where $\tau(x)$ is defined in display (1) of the main text. Taking the expectation over X gives the results of part (i)(c) of the proposition.

Part (ii): From the definition of the weight functions in (24) we find that for $p(x) = p_*(x)$ we have

$$w^{(q)}\big(1 - p(x), 1 - p_*(x)\big) = 1, \qquad \widetilde{w}^{(q)}\big(1 - p(x), 1 - p_*(x)\big) = 1.$$

By the same arguments as in part (i) of the proof we therefore find that (A.5) and (A.6) hold with equality, and all the inequalities in part (i) of the proposition then also hold with equality.

Part (iii): Define

$$\begin{aligned}
v^{(q)}(p, x) &:= \sum_{r=0}^{q-1} \lambda_r(1, x)\, p^{r+1}, \\
\widetilde{v}^{(q)}(p, x) &:= -\sum_{r=0}^{q-1} \lambda_r(x)\, p^{r-1}\, (1 - p).
\end{aligned}$$

The bounds in (21) can then be written as

$$\begin{aligned}
B^{(q)}(1, a, \lambda) &= a + v^{(q)}(p(X), X)\, \frac{D\,(Y - a)}{p(X)}, \\
C^{(q)}(a, \lambda) &= D\,(Y - a) - \widetilde{v}^{(q)}(p(X), X)\, \frac{p(X)\,(1 - D)\,(Y - a)}{1 - p(X)}.
\end{aligned}$$

Thus, $v^{(q)}(p,1,x)$ and $\widetilde{v}^{(q)}(p,x)$ take exactly the roles of $w^{(q)}\big(p,p_*(x)\big)$ and $\widetilde{w}^{(q)}\big(p,p_*(x)\big)$ in (23). By the same arguments as in the proof of Lemma 1 and in part (i) of the proof of the current proposition we therefore find that these bounds are valid (in the sense of satisfying the inequalities in part (i) of this proposition) for all DGP's that satisfy Assumption 1(i) and (ii) if and only if we have for all $x \in \mathcal{X}$ and $p \in [0,1]$ (or $p \in (0,1]$ for $\widetilde{v}$) that

$$v^{(q)}(p,x) \leq 1, \qquad\qquad \widetilde{v}^{(q)}(p,x) \leq 1.$$

Thus, $v^{(q)}(p,x)$ and $\widetilde{v}^{(q)}(p,x)$ satisfy all conditions on $v(p,\lambda)$ and $\widetilde{v}(p,\lambda)$ in Lemma 3. there exists $\epsilon > 0$ such that for all $p(x) \in \mathcal{B}_\epsilon(p_*(x))$ we have

$$w^{(q)}(p,p_*) - v^{(q)}(p,x) \geq 0, \qquad \text{and} \qquad \widetilde{w}^{(q)}(p,p_*) - \widetilde{v}^{(q)}(p,x) \geq 0. \tag{A.7}$$

Using this together with

$$\begin{aligned}&\mathbb{E}_{p(x)}\left[B^{(q)}(1,a) - B^{(q)}(1,a,\lambda)\,\middle|\, X=x\right] \\ &\quad = \left[w^{(q)}\big(p(x),p_*(x)\big) - v^{(q)}(p,x)\right]\,\mathbb{E}_{p(x)}\left[Y(1)-a\,\middle|\, X=x\right],\end{aligned}$$

and $\mathbb{E}_{p(x)}\left[Y(1)-a_{\min}\,\middle|\, X=x\right] > 0$, and $\mathbb{E}_{p(x)}\left[Y(1)-a_{\max}\,\middle|\, X=x\right] < 0$ we obtain that

$$\begin{aligned}\mathbb{E}_{p(x)}\left[B^{(q)}(1,a_{\min}) - B^{(q)}(1,a_{\min},\lambda)\,\middle|\, X=x\right] \geq 0,\\ \mathbb{E}_{p(x)}\left[B^{(q)}(1,a_{\max}) - B^{(q)}(1,a_{\max},\lambda)\,\middle|\, X=x\right] \leq 0,\end{aligned}$$

where everywhere $p(x) \in \mathcal{B}_\epsilon(p_*(x))$ to guarantee that (A.7) holds. From this we find that

$$\begin{aligned}&\mathbb{E}_{p(x)}\left[B^{(q)}(d,a_{\max}) - B^{(q)}(d,a_{\min})\,\middle|\, X=x\right] \\ &\qquad\qquad \leq \mathbb{E}_{p(x)}\left[B^{(q)}(d,a_{\max},\lambda) - B^{(q)}(d,a_{\min},\lambda)\,\middle|\, X=x\right]\end{aligned}$$

holds for $d=1$. The same result for $d=0$ follows by applying the transformation $Y \leftrightarrow 1-Y$ and $p(x) \leftrightarrow 1-p(x)$.

Similarly, we have

$$\mathbb{E}_{p(x)}\left[C^{(q)}(a) - C^{(q)}(a,\lambda)\,\middle|\, X=x\right] = -p(x)\left[\widetilde{w}^{(q)}\big(p(x),p_*(x)\big) - \widetilde{v}^{(q)}(p,x)\right]\mathbb{E}_{p(x)}\left[Y(0)-a\,\middle|\, X=x\right],$$

and therefore, for $p(x) \in \mathcal{B}_\epsilon(p_*(x))$, we find that

$$\mathbb{E}\left[C^{(q)}(a_{\min}) - C^{(q)}(a_{\min},\lambda)\,\middle|\, X=x\right] \leq 0, \mathbb{E}\left[C^{(q)}(a_{\max}) - C^{(q)}(a_{\max},\lambda)\,\middle|\, X=x\right] \quad \geq 0,$$

which implies that

$$\begin{aligned}&\mathbb{E}_{p(x)}\left[C^{(q)}(a_{\min}) - C^{(q)}(a_{\max})\,\middle|\, X=x\right] \\ &\qquad\qquad \leq \mathbb{E}_{p(x)}\left[C^{(q)}(a_{\min},\lambda) - C^{(q)}(a_{\max},\lambda)\,\middle|\, X=x\right].\end{aligned}$$

This concludes the proof of the proposition. ■

A.2 Proofs of intermediate lemmas

Proof of Lemma 2. # Part (i) for q even: Since $w^{(q)}(p, p_*) = 1 - \left(\frac{p_* - p}{p_*}\right)^q$ is a q'th order polynomial in p and satisfies $w^{(q)}(0, p_*) = 0$ we can find coefficients $\overline{\lambda}$ such that $v(p_*, \overline{\lambda}) = w^{(q)}(p, p_*)$. Furthermore, from the definition of $w^{(q)}(p, p_*)$ it is straightforward to verify that

$$\begin{aligned} w^{(q)}(p_*, p_*) &= 1, \\ w^{(q)}(p, p_*) &\leq 1, && \text{for } p \in [0, 1], \\ \frac{\partial^k w^{(q)}(p_*, p_*)}{\partial^k p} &= 0, && \text{for } k \in \{1, \ldots, q-1\}. \end{aligned}$$

This shows that $\overline{\lambda}$ with $v(p_*, \overline{\lambda}) = w^{(q)}(p, p_*)$ satisfies the optimization problem in part (i) of the lemma with objective function $\left|\frac{\partial^{q-1} v(p_*, \lambda)}{\partial^{q-1} p}\right|$ equal to zero at the optimum. Since the objective function is non-negative this indeed must be a minimizer. The solution is unique, because $v(p_*, \lambda) = 1$ and $\frac{\partial^k v(p_*, \lambda)}{\partial^k p} = 0$, for $k \in \{1, \ldots, q-1\}$, is a system of q linear equations in q unknowns λ that has a unique solution.

Part (i) for q odd: The optimization problem has $q - 1$ linear equality constraints:

$$\begin{aligned} v(p_*, \lambda) &= 1, \\ \frac{\partial^k v(p_*, \lambda)}{\partial^k p} &= 0, && \text{for } k \in \{1, \ldots, q-2\}. \end{aligned}$$

Any solution $\lambda = \lambda(\kappa)$ to this system of equations satisfies

$$v(p, \lambda) = 1 - (1 - \kappa\, p) \left(\frac{p_* - p}{p_*}\right)^{q-1},$$

where $\kappa \in \mathbb{R}$ is one remaining degree of freedom that is not determined from those equality constraints. For this solution we have

$$v(1, \lambda) = 1 - (1 - \kappa) \left(\frac{p_* - 1}{p_*}\right)^{q-1},$$

and the constraint $v(1, \lambda) \leq 1$ therefore requires that $\kappa \leq 1$. It is easy to check that for $\kappa \leq 1$ we also have $v(p, \lambda) \leq 1$ for all other $p \in [0, 1]$. We furthermore find

$$\left|\frac{\partial^{q-1} v(p_*, \lambda)}{\partial^q p}\right| = (q-1)!\, \frac{|1 - \kappa p_*|}{p_*^{q-1}}.$$

Minimizing this over $\kappa \leq 1$ gives the optimal value at the boundary point $\overline{\kappa} = 1$. We have therefore shown that the unique solution to the minimization problem is given by

$$v(p, \overline{\lambda}) = 1 - (1 - p) \left(\frac{p_* - p}{p_*}\right)^{q-1} = w^{(q)}(p, p_*).$$

Part (ii) for q even: Since $p\,\widetilde{w}^{(q)}(p,p_*) = p - \left(\frac{p-p_*}{1-p_*}\right)^q$ is a q'th order polynomial in p and satisfies $\widetilde{w}^{(q)}(1,p_*) = 0$ we can find coefficients $\widetilde{\lambda}$ such that $\widetilde{v}(p_*,\widetilde{\lambda}) = \widetilde{w}^{(q)}(p,p_*)$. Furthermore, from the definition of $\widetilde{w}^{(q)}(p,p_*)$ it is straightforward to verify that

$$\begin{aligned} \widetilde{w}^{(q)}(p_*,p_*) &= 1, \\ \widetilde{w}^{(q)}(p,p_*) &\leq 1, && \text{for } p \in (0,1], \\ \frac{\partial^k \widetilde{w}^{(q)}(p_*,p_*)}{\partial^k p} &= 0, && \text{for } k \in \{1,\ldots,q-1\}. \end{aligned}$$

This shows that $\widetilde{\lambda}$ with $\widetilde{v}(p_*,\widetilde{\lambda}) = \widetilde{w}^{(q)}(p,p_*)$ satisfies the optimization problem in part (i) of the lemma with objective function $\left|\frac{\partial^{q-1}\widetilde{v}(p_*,\lambda)}{\partial^{q-1}p}\right|$ equal to zero at the optimum. Since the objective function is non-negative this indeed must be a minimizer. The solution is unique, because $\widetilde{v}(p_*,\lambda) = 1$ and $\frac{\partial^k \widetilde{v}(p_*,\lambda)}{\partial^k p} = 0$, for $k \in \{1,\ldots,q-1\}$, is a system of q linear equations in q unknowns λ that has a unique solution.

Part (ii) for q odd: The optimization problem has $q-1$ linear equality constraints:

$$\begin{aligned} \widetilde{v}(p_*,\lambda) &= 1, \\ \frac{\partial^k \widetilde{v}(p_*,\lambda)}{\partial^k p} &= 0, && \text{for } k \in \{1,\ldots,q-2\}. \end{aligned}$$

Any solution $\lambda = \lambda(\kappa)$ to this system of equations satisfies

$$\widetilde{v}(p,\lambda) = 1 - \left(\kappa + \frac{1-\kappa}{p}\right)\left(\frac{p-p_*}{1-p_*}\right)^{q-1},$$

where $\kappa \in \mathbb{R}$ is one remaining degree of freedom that is not determined from those equality constraints. For this solution we have

$$\lim_{p\to 0} \widetilde{v}(p,\lambda) = \begin{cases} \infty & \text{if } \kappa > 1, \\ 1 - \left(\frac{p_*}{1-p_*}\right)^{q-1} & \text{if } \kappa = 1 \\ -\infty & \text{if } \kappa < 1. \end{cases}$$

and the constraint $\widetilde{v}(p,\lambda) \leq 1$ for all $p \in (0,1]$ therefore requires that $\kappa \leq 1$. It is easy to check that for $\kappa \leq 1$ this inequality is indeed satisfied for all $p \in (0,1]$. We furthermore find

$$\left|\frac{\partial^{q-1} v(p_*,\lambda)}{\partial^q p}\right| = \frac{(q-1)!}{(1-p_*)^{q-1}}\left|\kappa + \frac{1-\kappa}{p_*}\right|.$$

Minimizing this over $\kappa \leq 1$ gives the optimal value at the boundary point $\widetilde{\kappa} = 1$. We have therefore shown that the unique solution to the minimization problem is given by

$$\widetilde{v}(p,\widetilde{\lambda}) = 1 - \left(\frac{p-p_*}{1-p_*}\right)^{q-1} = \widetilde{w}^{(q)}(p,p_*).$$

■

Proof of Lemma 3. # Part (i): We define the non-negative integer K and the positive number C as follows: If $v(p_*, \lambda) \neq 1$, then we set $K = 0$ and $C = 1 - v(p_*, \lambda)$. Otherwise, let K be the smallest integer such that

$$\frac{\partial^K v(p_*, \lambda)}{\partial^K p} \neq 0,$$

and set

$$C = -\frac{\partial^K v(p_*, \lambda)}{\partial^K p}.$$

It must be the case that K is even and that $C > 0$, because otherwise the assumption $v(p, \lambda) \leq 1$, for all $p \in [0, 1]$, would be violated. A Taylor expansion of $v(p, \lambda)$ around $p = p_*$ gives

$$v(p, \lambda) = 1 - C\,(p - p_*)^K + O\left(|p - p_*|^{K+1}\right). \tag{A.8}$$

Next, let $q_* = q$ if q is even, and let $q_* = q - 1$ if q is odd. We have $w^{(q)}(p_*, p_*) = 1$, and

$$\frac{\partial^k w^{(q)}(p_*, p_*)}{\partial^k p} = 0, \qquad \text{for all } k \in \{1, \ldots, q_* - 1\}.$$

Therefore, a Taylor expansion of $w^{(q)}(p, p_*)$ around $p = p_*$ gives

$$w^{(q)}(p, p_*) = 1 + O\left(|p - p_*|^{q_*}\right). \tag{A.9}$$

If $K < q_*$, then (A.8) and (A.9) imply that

$$v(p, \lambda) = w^{(q)}(p, p_*) - C\,(p - p_*)^K + O\left(|p - p_*|^{K+1}\right).$$

Since $C > 0$ and K is even, there must then exist $\epsilon > 0$ such that for all $p \in \mathcal{B}_\epsilon(p_*)$ we have $v(p, \lambda) \leq w^{(q)}(p, p_*)$.

If $K = q_*$ and q is even, then $v(p, \lambda)$ satisfies $v(p_*, \lambda) = 1$ and $\frac{\partial^k v(p_*, \lambda)}{\partial^k p} = 0$, for all $k \in \{1, \ldots, q - 1\}$. This is exactly the system of q linear equations in q unknowns λ whose solution is $\overline{\lambda}$. In that case, we therefore have $v(p, \lambda) = w^{(q)}(p, p_*)$, and the statement of the lemma holds for any $\epsilon > 0$.

If $K = q_*$ and q is odd, then $v(p, \lambda)$ satisfies all the constraints in the optimization problem in part (i) of Lemma 2. If $v(p, \lambda)$ is the solution to this optimization problem, then we again have $v(p, \lambda) = w^{(q)}(p, p_*)$, and the statement of the lemma holds for any $\epsilon > 0$. Otherwise, $v(p, \lambda)$ is not the solution to this optimization problem, which implies that

$$C = \frac{\partial^K v(p_*, \lambda)}{\partial^K p} > \frac{\partial^K w^{(q)}(p_*, p_*)}{\partial^K p} =: c > 0.$$

In that case, analogous to (A.8) we have

$$w^{(q)}(p, p_*) = 1 - c\,(p - p_*)^K + O\left(|p - p_*|^{K+1}\right),$$

and therefore

$$v(p, \lambda) = w^{(q)}(p, p_*) - (C - c)\,(p - p_*)^K + O\left(|p - p_*|^{K+1}\right).$$

Since $C - c > 0$ and K is even, there must again exist $\epsilon > 0$ such that for all $p \in \mathcal{B}_\epsilon(p_*)$ we have $v(p, \lambda) \leq w^{(q)}(p, p_*)$. We have therefore shown that the desired result holds in all possible cases.

Part (ii): The proof of $\widetilde{v}(p, \lambda) \leq \widetilde{w}^{(q)}(p, p_*)$ is analogous, using that $\widetilde{w}^{(q)}(p, p_*)$ is the solution to the optimization problem in part (ii) of Lemma 2. ■

B Derivation of the sample weights in Section 4

Here, we want to discuss where the formulas in (38) and (40) for $\widehat{w}_{0/1}(x)$ and $\widehat{v}(x)$ come from, and why the ω coefficients need to be chosen according to (39).

Consider first the case where $q(x) = \min\{Q, n(x)\}$ is even, in which case $\widehat{w}_1(x)$ is given by (36). As explained in the main text, this formula for $\widehat{w}_1(x)$ guarantees that the conditional expectation of $\widehat{w}_1(x)$ is given by (27), but for the purpose of practical implementation we want to express $\widehat{w}_1(x)$ not in terms of individual observations D_i, but in terms of the summary statistics $n(x)$ and $n_1(x)$. For simplicity, we only write q instead of $q(x)$ in the following. We can rewrite the expression for $\widehat{w}_1(x)$ in (36) as

$$\begin{aligned}
\widehat{w}_1(x) &= 1 - \binom{n(x)}{q}^{-1} \sum_{\mathcal{S}_q} \left(\frac{p_*(x) - 1}{p_*(x)}\right)^{n_1(\mathcal{S}_q)} \\
&= 1 - \binom{n(x)}{q}^{-1} \sum_{k=0}^{q} \underbrace{\left(\sum_{\mathcal{S}_q} \mathbb{1}\{n_1(\mathcal{S}_q) = k\}\right)}_{=:\,\alpha_{k, n_1(x), n(x), q}} \left(\frac{p_*(x) - 1}{p_*(x)}\right)^k,
\end{aligned}$$

where $n_1(\mathcal{S}_q)$ is the number of observations $i \in \mathcal{S}_q$ with $D_i = 1$, and $\alpha_{k, n_1(x), n(x), q} \in \{1, 2, \ldots\}$ is the number of subsets $\mathcal{S}_q$ for which we have $n_1(\mathcal{S}_q) = k$. By standard combinatorial

arguments one finds that[19]

$$\alpha_{k,n_1(x),n(x),q} = \binom{n_1(x)}{k}\binom{n(x)-n_1(x)}{q-k}. \tag{B.10}$$

We therefore obtain the definition of $\widehat{w}_1(x)$ in (38) by setting

$$\omega_{k,n_1(x),n(x),Q} = \binom{n(x)}{q}^{-1} \alpha_{k,n_1(x),n(x),q},$$

for $q = \min\{Q, n(x)\}$ even, and combining the last two displays gives the formulas for ω in (39) for that case. Since $\alpha_{k,n_1(x),n(x),q} \leq \binom{n(x)}{q}$ it follows that $\omega_{k,n_1(x),n(x),Q} \in [0,1]$. The combinatorial argument for the case that $q(x) = \min\{Q, n(x)\}$ odd is analogous, as are the derivations for $\widehat{w}_0(x)$ and $\widehat{v}(x)$, which give the same result for $\omega_{k,n_{0/1}(x),n(x),Q}$.

C Proofs for Section 4.3

Display (44) in the main text defined the parameters of interest $\theta^{(r)}$, which are labeled by the index $r \in \{0, 1, \text{ATE}, \text{ATT}\}$. Our lower and upper bound estimates for $r \in \{0, 1, \text{ATE}\}$ can be written as simple sample averages over $x \in \mathcal{X}_*$,

$$\overline{L}^{(r)} = \frac{1}{m}\sum_{x\in\mathcal{X}_*} L_x^{(r)}, \qquad\qquad \overline{U}^{(r)} = \frac{1}{m}\sum_{x\in\mathcal{X}_*} U_x^{(r)},$$

with $L_x^{(r)}$ and $U_x^{(r)}$ defined in the main text. By contrast, the lower and upper bound estimates $\overline{L}^{(\text{ATT})}$ and $\overline{U}^{(\text{ATT})}$ defined in (31) take the form of a ratio of sample averages, with numerator and denominator given by

$$\overline{C}(a) = \frac{1}{n}\sum_{i=1}^n \widehat{C}_i(a) = \frac{1}{m}\sum_{x\in\mathcal{X}_*} \frac{m\, n(x)\, \widehat{C}_x(a)}{n},$$
$$\frac{1}{n}\sum_{i=1}^n D_i = \frac{1}{m}\sum_{x\in\mathcal{X}_*} \frac{m\, n_1(x)}{n}.$$

Since we assume $\mathbb{E}(D) > 0$, and our assumptions also guarantee $\frac{1}{n}\sum_{i=1}^n [D_i - \mathbb{E}(D)] = O_P(1/\sqrt{n})$, we can apply the delta method to find

$$\frac{1}{\frac{1}{n}\sum_{i=1}^n D_i} = \frac{1}{\mathbb{E}(D)} + \frac{\mathbb{E}(D) - \frac{1}{n}\sum_{i=1}^n D_i}{[\mathbb{E}(D)]^2} + O_P(1/n)$$

[19] We can generate all subsets $\mathcal{S}_q \subset \mathcal{N}(x)$ with q elements and $n_1(\mathcal{S}_q) = k$ by first choosing k of the $n_1(x)$ units in $\mathcal{N}(x)$ with $D_i = 1$, which gives the factors $\binom{n_1(x)}{k}$, and secondly choosing $q-k$ of the $n(x) - n_1(x)$ units in $\mathcal{N}(x)$ with $D_i = 0$, which gives the factor $\binom{n(x)-n_1(x)}{q-k}$ in (B.10). Here, we use the standard convention for the binomial coefficient that $\binom{a}{b} = 0$ for all integers $b > a \geq 0$, but $\binom{0}{0} = 1$.

and therefore

$$
\begin{aligned}
&\frac{\overline{C}(a)}{\frac{1}{n}\sum_{i=1}^{n} D_i} \\
&= \frac{\mathbb{E}\overline{C}(a)}{\mathbb{E}(D)} + \frac{\frac{1}{m}\sum_{x\in\mathcal{X}_*} \frac{m\, n(x)\, \widehat{C}_x(a)}{n} - \mathbb{E}\overline{C}(a)}{\mathbb{E}(D)} + \frac{[\mathbb{E}\overline{C}(a)][\mathbb{E}(D) - \frac{1}{n}\sum_{i=1}^{n} D_i]}{\left[\mathbb{E}(D)\right]^2} + O_P(1/n) \\
&= \frac{\mathbb{E}\overline{C}(a)}{\mathbb{E}(D)} + \frac{\frac{1}{m}\sum_{x\in\mathcal{X}_*} \frac{m\, n(x)\, \widehat{C}_x(a)}{n}}{\mathbb{E}(D)} - \frac{\left[\mathbb{E}\overline{C}(a)\right]\left[\frac{1}{m}\sum_{x\in\mathcal{X}_*} \frac{m\, n_1(x)}{n}\right]}{\left[\mathbb{E}(D)\right]^2} + O_P(1/n) \\
&= \frac{\mathbb{E}\overline{C}(a)}{\mathbb{E}(D)} + \frac{1}{m}\sum_{x\in\mathcal{X}_*}\left[\frac{m\, n(x)\, \widehat{C}_x(a)}{n\, \mathbb{E}(D)} - \frac{m\, n_1(x)\, \left[\mathbb{E}\overline{C}(a)\right]}{n\, \left[\mathbb{E}(D)\right]^2}\right] + O_P(1/n). \qquad \text{(C.11)}
\end{aligned}
$$

This shows that the influence function of the ratio $\frac{\overline{C}(a)}{\frac{1}{n}\sum_{i=1}^{n} D_i}$ is given by $\frac{m\, n(x)\, \widehat{C}_x(a)}{n\, \mathbb{E}(D)} - \frac{m\, n_1(x)\, \left[\mathbb{E}\overline{C}(a)\right]}{n\, \left[\mathbb{E}(D)\right]^2}$. When using this influence function to calculate the asymptotic variance of the ration, then $\mathbb{E}(D)$ and $\mathbb{E}\overline{C}(a)$ need to again be replace by their consistent estimates $\frac{1}{n}\sum_{i=1}^{n} D_i$ and $\overline{C}(a)$, and after that replacement we obtain

$$
\frac{\overline{C}(a)}{\frac{1}{n}\sum_{i=1}^{n} D_i} = \frac{\mathbb{E}\overline{C}(a)}{\mathbb{E}(D)} + \frac{1}{m}\sum_{x\in\mathcal{X}_*}\left[\frac{m\, n(x)\, \widehat{C}_x(a_{\max})}{\sum_{i=1}^{n} D_i} - \frac{m\, n\, n_1(x)\, \overline{C}(a_{\max})}{\left(\sum_{i=1}^{n} D_i\right)^2}\right] + O_P(1/m), \qquad \text{(C.12)}
$$

which is exactly the expression for $\overline{L}^{(\mathrm{ATT})}$ and $\overline{U}^{(\mathrm{ATT})}$ given in (45) and (46) of the main text. We have thus derived the expressions for $L_x^{(\mathrm{ATT})}$ and $U_x^{(\mathrm{ATT})}$ given in the main text. We are now ready to prove Theorem 1.

Proof of Theorem 1. # Preliminaries: Since we assume that Q and $\max_{x\in\mathcal{X}_*} n(x)$ are bounded, and that $p_*(x)$ is bounded away from zero and one we have

$$
\max_{x\in\mathcal{X}_*} \left|L_x^{(r)}\right| = O(1), \qquad \text{(C.13)}
$$

for $r \in \{0, 1, \mathrm{ATE}, \mathrm{ATT}\}$.

Consider $r \in \{0, 1, \mathrm{ATE}\}$. In that case we have

$$
\overline{L}^{(r)} = \frac{1}{m}\sum_{x\in\mathcal{X}_*} L_x^{(r)}.
$$

Using that $L_x^{(r)}$ is independent across $x \in \mathcal{X}_*$ and that, according to (C.13), the $L_x^{(r)}$ are uniformly bounded, we find that

$$
\begin{aligned}
\mathrm{Var}\left(\overline{L}^{(r)}\right) &= \frac{1}{m^2}\sum_{x\in\mathcal{X}_*} \mathrm{Var}\left(L_x^{(r)}\right) = O(1/m), \\
\overline{L}^{(r)} - \mathbb{E}\,\overline{L}^{(r)} &\Rightarrow \mathcal{N}\left(0, \mathrm{Var}\left(\overline{L}^{(r)}\right)\right),
\end{aligned}
$$

where the first line is standard property of the variance of the sum of independent random variables, and in the second line we applied Lyapunov's CLT. From $\operatorname{Var}\left(\overline{L}^{(r)}\right) = O(1/m)$ we find, by an application of Markov's inequality, that

$$\overline{L}^{(r)} - \mathbb{E}\,\overline{L}^{(r)} = O_P(m^{-1/2}).$$

Finally, we compute

$$\begin{aligned} m \operatorname{Var}\left(\overline{L}^{(r)}\right) &= \frac{1}{m}\sum_{x\in\mathcal{X}_*} \operatorname{Var}\left(L_x^{(r)}\right) \\ &= \frac{1}{m}\sum_{x\in\mathcal{X}_*} \mathbb{E}\left[\left(L_x^{(r)}\right)^2\right] - \frac{1}{m}\sum_{x\in\mathcal{X}_*} \left[\mathbb{E}\left(L_x^{(r)}\right)\right]^2 \\ &\leq \frac{1}{m}\sum_{x\in\mathcal{X}_*} \mathbb{E}\left[\left(L_x^{(r)}\right)^2\right] - \left[\frac{1}{m}\sum_{x\in\mathcal{X}_*} \mathbb{E}\left(L_x^{(r)}\right)\right]^2, \end{aligned}$$

where in the last step we used that $\frac{1}{m}\sum_x (a_x)^2 \geq \left(\frac{1}{m}\sum_x a_x\right)^2$, which holds for any $a_x \in \mathbb{R}$ according to Jensen's inequality. Using again that $L_x^{(r)}$ is uniformly bounded and independent across x we have

$$\begin{aligned} \frac{1}{m}\sum_{x\in\mathcal{X}_*} \mathbb{E}\left[\left(L_x^{(r)}\right)^2\right] &= \frac{1}{m}\sum_{x\in\mathcal{X}_*} \left(L_x^{(r)}\right)^2 + O_P(1/\sqrt{m}), \\ \frac{1}{m}\sum_{x\in\mathcal{X}_*} \mathbb{E}\left(L_x^{(r)}\right) &= \frac{1}{m}\sum_{x\in\mathcal{X}_*} L_x^{(r)} + O_P(1/\sqrt{m}). \end{aligned}$$

Combining the results of the last two displays gives

$$\operatorname{Var}\left(\overline{L}^{(r)}\right) \leq \frac{\operatorname{SVar}\left(L_x^{(r)}\right) + O_P(1/\sqrt{m})}{m}.$$

We have thus shown part (i), (ii), (iii) of the theorem for $\overline{L}^{(r)}$ and $r \in \{0, 1, \text{ATE}\}$. The proof for $\overline{U}^{(r)}$ and $r \in \{0, 1, \text{ATE}\}$ is analogous.

Next, consider $r = \text{ATT}$. We can rewrite (C.11) for $a = a_{\max}$ as

$$\overline{L}^{(\text{ATT})} = \frac{\mathbb{E}\overline{C}(a_{\max})}{\mathbb{E}(D)} + \frac{1}{m}\sum_{x\in\mathcal{X}_*} \widetilde{L}_x^{(\text{ATT})} + O_P(1/n),$$

where

$$\widetilde{L}_x^{(\text{ATT})} := \frac{m\, n(x)\, \widehat{C}_x(a_{\max})}{n\, \mathbb{E}(D)} - \frac{m\, n_1(x)\, \left[\mathbb{E}\overline{C}(a_{\max})\right]}{n\, \left[\mathbb{E}(D)\right]^2}.$$

Using this expansion of $\overline{L}^{(\mathrm{ATT})}$ we then derive the results

$$\overline{L}^{(\mathrm{ATT})} - \mathbb{E}\,\overline{L}^{(\mathrm{ATT})} = O_P(m^{-1/2}),$$

and

$$\frac{\overline{L}^{(\mathrm{ATT})} - \mathbb{E}\,\overline{L}^{(\mathrm{ATT})}}{\left[\mathrm{Var}\left(\overline{L}^{(\mathrm{ATT})}\right)\right]^{1/2}} \Rightarrow \mathcal{N}(0,1),$$

and

$$\mathrm{Var}\left(\overline{L}^{(r)}\right) \leq \frac{\mathrm{SVar}\left(\widetilde{L}_x^{(r)}\right) + O_P(1/\sqrt{m})}{m}$$

in the same way as for $r \in \{0, 1, \mathrm{ATE}\}$ above. However, $\mathrm{SVar}\left(\widetilde{L}_x^{(r)}\right)$ is infeasible, because $\mathbb{E}(D)$ and $\mathbb{E}\overline{C}(a)$ are unknown. Replacing those expectations with their consistent estimates gives

$$\mathrm{SVar}\left(\widetilde{L}_x^{(r)}\right) = \mathrm{SVar}\left(L_x^{(r)}\right) + O_P(1/\sqrt{m}).$$

This concludes the proof. ■