Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Sister Conferences Best Papers Track

What Makes Treatment Effects Identifiable? Characterizations and Estimators Beyond Unconfoundedness (Extended Abstract)*

Yang Cai[1] , **Alkis Kalavasis**[1] , **Katerina Mamali**[1] , **Anay Mehrotra**[2] and **Manolis Zampetakis**[1]
[1]Yale University
[2]Stanford University
{yang.cai, alkis.kalavasis, katerina.mamali, manolis.zampetakis}@yale.edu
anaymehrotra1@gmail.com

Abstract

Most of the widely used estimators of the *average treatment effect* (ATE) in causal inference rely on the assumptions of *unconfoundedness* and *overlap*. Unconfoundedness requires that the observed covariates account for all correlations between the outcome and treatment. Overlap requires the existence of randomness in treatment decisions for all individuals. Nevertheless, many types of studies frequently violate unconfoundedness or overlap, for instance, observational studies with deterministic treatment decisions – popularly known as Regression Discontinuity designs – violate overlap.

In this paper, we initiate the study of general conditions that enable the *identification* of the average treatment effect, extending beyond unconfoundedness and overlap. In particular, following the paradigm of statistical learning theory, we provide an interpretable condition that is sufficient and necessary for the identification of ATE. Moreover, this condition can be used to characterize other treatment effects (such as the average treatment effect on the treated (ATT)) as well. To illustrate the utility of our condition, we present several well-studied scenarios where our condition is satisfied and, hence, we prove that ATE can be identified in regimes that prior works could not capture. For example, under mild assumptions on the data distributions, this holds for sensitivity-analysis models of Tan [2006] and Rosenbaum [2002], and for the Regression Discontinuity design model of Thistlethwaite and Campbell [1960]. For each of these scenarios, we also show that, under natural additional assumptions, ATE can be estimated from finite samples.

We believe these findings open new avenues for bridging learning-theoretic insights and causal inference methodologies, particularly in observational studies with complex treatment mechanisms.

*Full version: Cai *et al.* [2025].

1 Introduction

Understanding cause and effect is a central goal in science and decision-making. Across disciplines, we ask: What is the effect of a new drug on disease rates? How does a policy impact growth? Is technology driving economic growth? *Causal inference* tackles such questions by disentangling correlation from causation. Unlike statistical learning, which predicts outcomes from data, causal inference estimates the effects of interventions that alter the data-generating process.

A fundamental challenge in causal inference is that we can never observe both potential outcomes for the same individual. For example, if a patient takes a medication and recovers, we do not know whether the patient would have recovered without it. Due to this *fundamental problem of causal inference,* causal effects must be inferred under certain assumptions [Holland, 1986].

To formalize this challenge, we consider the widely-used *potential outcomes model* introduced by Neyman [1990] (originally published in 1923) and later formalized by Rubin [1974]; see also [Hernan and Robins, 2023; Rosenbaum, 2002; Chernozhukov *et al.*, 2024]. Here, for a unit with *covariates* $X \in \mathbb{R}^d$, $Y(1)$ and $Y(0)$ denote potential outcomes under treatment and control, respectively. Since only the outcome $Y(T)$ corresponding to the assigned treatment T is observed, certain assumptions are needed to estimate the *average treatment effect* (ATE), defined as $\tau := \mathbb{E}[Y(1) - Y(0)]$, where $(T, Y(1), Y(0))$ are random variables whose distribution may depend on X. This framework underpins many modern causal inference methods – both practical and theoretical – and can capture many treatment effects, apart from τ. Two fundamental questions under this framework, studied since [Cochran, 1965; Rubin, 1974; Rubin, 1978; Heckman, 1979], are *identification* and *estimation*:

> **Q1:** Given infinite samples of the form $(X, T, Y(T))$, can we *identify* treatment effects?
>
> **Q2:** Given n samples $(X, T, Y(T))$, can we *estimate* treatment effects up to error $\varepsilon(n)$?

Due to the missingness in data (explained above), even the identification problem is unsolvable without making structural assumptions on the distribution of $(X, T, Y(T))$, which is a censored version of the (complete) data distribution $(X, T, Y(1), Y(0))$. The earliest and most widely used such assumptions are *unconfoundedness* and *overlap*.

- ▷ *Unconfoundedness* presumes that after conditioning on the value of the covariate X, the treatment random variable T is independent of the outcomes $Y(1)$ and $Y(0)$, *i.e.*, $T \perp (Y(0), Y(1)) \mid X$.
- ▷ *Overlap* requires that the probability of being assigned treatment conditional on the covariate X, *i.e.*, $\Pr[T{=}1 \mid X{=}x]$, is a quantity strictly between 0 and 1 for each x.

Unconfoundedness and overlap are essential for unbiased estimation of the average treatment effect and are widely studied across Statistics (*e.g.*, [Rosenbaum, 2002]) and many other disciplines, including Medicine (*e.g.*, [Rosenbaum and Rubin, 1983]), Economics (*e.g.*, [Athey and Imbens, 2017]), Political Science (*e.g.*, [Brunell and DiNardo, 2004]), Sociology (*e.g.*, [Morgan and Harding, 2006]), and other fields (*e.g.*, [Austin, 2008]). Despite their wide use across different disciplines, there are fundamental instances where unconfoundedness or overlap are easily violated:

- ▷ Unconfoundedness is often violated in *observational studies*, where treatments or exposures are not assigned by the researcher but observed in a natural setting. In a prospective cohort study, for example, individuals are followed over time to assess how exposures influence outcomes. A common violation arises when key confounders are unmeasured. For instance, in studying smoking's impact on health, omitting socioeconomic status (SES), which affects both smoking habits and health, can bias results, as lower SES correlates with higher smoking rates and poorer health, independent of smoking.
- ▷ Overlap is violated when certain covariate values make treatment assignments nearly deterministic. In a marketing study estimating the effect of personalized advertisements on purchases, covariates like demographics, browsing history, and preferences define a high-dimensional feature space. As this space grows, many user profiles either always or never receive the ad, leading to *lack of overlap* [D'Amour *et al.*, 2021]. Without comparable treated and untreated units, causal inference methods struggle to estimate counterfactual outcomes, yielding unreliable effect estimates.

These examples lead us to the following question:

Is identification and estimation of treatment effects possible in any meaningful setting without unconfoundedness or overlap?

This question is not new and can be traced back to at least the work of Rubin [1977], who recognized that, without substantial overlap between treatment and control groups, identification of treatment effects necessarily requires additional prior assumptions. To the best of our knowledge, the present work provides the first formal characterization of the precise assumptions required to identify treatment effects in scenarios lacking substantial overlap, unconfoundedness, or both.

2 Framework

The main conceptual contribution of this work is a *learning-theoretic* approach that enables a characterization of when identification and estimation of treatment effects are possible. Before presenting this approach, it is instructive to reconsider how unconfoundedness and overlap enable identification of the simplest and most widely used treatment effect – the average treatment effect: Given the observational study $\mathcal{D}$, which is a distribution over $(X, T, Y(0), Y(1))$, unconfoundedness and overlap put a strong constraint on $\mathcal{D}$: they require that $Y(t) \perp T \mid X{=}x$ for each $t \in \{0,1\}$ and $x \in \mathbb{R}^d$ and that the propensity scores $e(x) = \Pr[T{=}1 | X{=}x]$ are bounded away from 0 and 1. Under these assumptions, identification and estimation of ATE $\tau = \tau_{\mathcal{D}}$ is possible given censored samples $(X, T, Y(T))$ due to the following decomposition of $\tau_{\mathcal{D}}$ for a fixed $x \in \mathbb{R}^d$ (we integrate over the x-marginal to get $\tau_{\mathcal{D}}$):

$$\begin{aligned}&\mathbb{E}_{Y(0),Y(1)}\left[Y(1) - Y(0) \mid X{=}x\right] \\ &= \mathbb{E}_{Y(0),Y(1),T}\left[\tfrac{Y(1)\cdot T}{e(X)} - \tfrac{Y(0)\cdot(1-T)}{1-e(X)} \;\middle|\; X{=}x\right], \quad (1)\end{aligned}$$

where we use overlap to divide by $e(X), 1-e(X)$ and unconfoundedness to obtain the equation $\mathbb{E}[Y(1)\cdot T \mid X] = \mathbb{E}[Y(1) \mid X] \cdot \Pr[T{=}1 \mid X]$ and analogously for $Y(0)$. Note that all the quantities appearing in the RHS are identifiable and estimable[1] from the censored distribution $\mathcal{C}_{\mathcal{D}}$ [Rubin, 1978], which is defined over $(X, T, Y(T))$.

When no constraints are put on $\mathcal{D}$, identification of ATE is impossible in general [Imbens and Rubin, 2015]. Without unconfoundedness, propensity scores $\Pr[T{=}1 \mid X{=}x]$ are not sufficient to identify the distribution of T, which can also depend on the outcomes $Y(0)$ and $Y(1)$ (conditioned on $X{=}x$). Instead, writing $p_t(x, y) = \Pr[T{=}t \mid X{=}x, Y(t){=}y]$, we can decompose the expression of $\tau_{\mathcal{D}}$ for a fixed $x \in \mathbb{R}^d$ as follows:

$$\begin{aligned}&\mathop{\mathbb{E}}_{Y(0),Y(1)}\left[Y(1) - Y(0) \mid X{=}x\right] \\ &= \mathbb{E}\left[\frac{Y(1)\cdot T}{p_1(X, Y(1))} \;\middle|\; X{=}x\right] \\ &\quad - \mathbb{E}\left[\frac{Y(0)\cdot(1-T)}{p_0(X, Y(0))} \;\middle|\; X{=}x\right].\end{aligned}$$

If unconfoundedness holds, then we could recover (1) since then T would not depend on $Y(1), Y(0)$ given X. However, unlike the previous decomposition of Equation (1), the above equation always holds and crucially utilizes the *generalized propensity scores* with $t \in \{0,1\}$.[2] Unfortunately, these generalized propensity scores, in contrast to the standard propensity scores, are not always identifiable from data. To understand when these are identifiable, we need to consider the *joint distribution of covariates and outcomes* $\mathcal{D}_{X,Y(t)}$ for each $t \in \{0,1\}$.

To this end, we adopt an approach inspired by statistical learning theory [Valiant, 1984; Vapnik, 1999; Blumer *et al.*, 1989; Hastie *et al.*, 2013; Anthony and Bartlett, 1999; Alon *et al.*, 1997; Lugosi, 2002; Massart and Nédélec, 2006;

[1] We remark that the problem of estimating the propensity scores $e(x) = \Pr[T{=}1 | X{=}x]$ is identical to the classical problem of learning probabilistic concepts [Kearns and Schapire, 1994].

[2] Observe that we need some overlap condition to divide by $p_0(\cdot)$ and $p_1(\cdot)$ in the above equation. In our main results, however, we do *not* follow this decomposition and will *not* need such overlap conditions.

Vapnik, 2006; Bousquet *et al.*, 2003; Bousquet, 2003]. We introduce *concept classes* for the two key quantities derived by the above discussion p_t and $\mathcal{D}_{X,Y(t)}$ (for each $t \in \{0,1\}$) that will *place some restrictions* on the observational study $\mathcal{D}$ towards understanding which conditions enable identification and estimation. In the remainder of the paper, we assume that all distributions are continuous and have a density. (All results also extend to discrete domains by replacing densities by probability mass functions.)

We are interested in the structure of two concept classes: the class of generalized propensity scores $\mathbb{P} \subseteq \{p\colon \mathbb{R}^d \times \mathbb{R} \to [0,1]\}$ and the class of covariate-outcome distributions $\mathbb{D} \subseteq \Delta(\mathbb{R}^d \times \mathbb{R})$. As in classical statistical learning theory, having fixed the concept classes, our next step is to restrict the underlying distribution $\mathcal{D}$ to be *realizable* with respect to the pair of concept classes $(\mathbb{P}, \mathbb{D})$. An observational study $\mathcal{D}$ is said to be realizable with respect to the concept class pair $(\mathbb{P}, \mathbb{D})$ if the generalized propensity scores $p_0(\cdot), p_1(\cdot)$ induced by $\mathcal{D}$ belong to $\mathbb{P}$ and $\mathcal{D}_{X,Y(t)} \in \mathbb{D}$ for each $t \in \{0,1\}$. This learning-theoretic framework is quite expressive. For instance, it can capture unconfoundedness and overlap[3] by letting $\mathbb{D}$ be the set of all distributions over $\mathbb{R}^d \times \mathbb{R}$, denoted by $\mathbb{D}_{\text{all}}$, and restricting $\mathbb{P}$ to be the following class

$$\mathbb{P}_{\text{OU}}(c) \coloneqq \{p\colon \mathbb{R}^d \times \mathbb{R} \to [0,1] | \\ \forall_{x,y,z},\ p(x,y) = p(x,z) \text{ and } c < p(x,y) < 1-c\}. \quad (2)$$

That is, $\mathcal{D}$ satisfies unconfoundedness and c-overlap if and only if it is realizable with respect to the pair of classes $(\mathbb{P}_{\text{OU}}(c), \mathbb{D}_{\text{all}})$.

Before proceeding to our results, we introduce some further terminology. Given classes $(\mathbb{P}, \mathbb{D})$, we are particularly interested in the generalized propensity scores in $\mathbb{P}$ and covariate-outcome distributions in $\mathbb{D}$ that induce a valid observational study. To this end, we say that a tuple $(p, \mathcal{P}) \in \mathbb{P} \times \mathbb{D}$ is *compatible* with classes $(\mathbb{P}, \mathbb{D})$ (henceforth, just compatible) if there exists another tuple $(\widehat{p}, \widehat{\mathcal{P}}) \in \mathbb{P} \times \mathbb{D}$ such that setting $(p_0, p_1, \mathcal{D}_{X,Y(0)}, \mathcal{D}_{X,Y(1)}) = (p, \widehat{p}, \mathcal{P}, \widehat{\mathcal{P}})$ (or *equivalently* the re-ordered assignment $(p_0, p_1, \mathcal{D}_{X,Y(0)}, \mathcal{D}_{X,Y(1)}) = (\widehat{p}, p, \widehat{\mathcal{P}}, \mathcal{P})$) defines a valid observational study, *i.e.*, a valid distribution over $(X, T, Y(0), Y(1))$.

3 Main Results on Identification

We say that a certain treatment effect $\eta_{\mathcal{D}}$ is *identifiable* from the censored distribution $\mathcal{C}_{\mathcal{D}}$ when $(\mathbb{P}, \mathbb{D})$ satisfy some Condition C, if there is a mapping f such that $f(\mathcal{C}_{\mathcal{D}}) = \eta_{\mathcal{D}}$ for any observational study $\mathcal{D}$ realizable with respect to $(\mathbb{P}, \mathbb{D})$ that satisfies C. In other words, if $\eta_{\mathcal{D}_1} \neq \eta_{\mathcal{D}_2}$ then it should be $\mathcal{C}_{\mathcal{D}_1} \neq \mathcal{C}_{\mathcal{D}_2}$. Having set the stage, we now ask our first main question:

Which conditions on $(\mathbb{P}, \mathbb{D})$ characterize the average treatment effect's identifiability?

[3]We refer to overlap as c-overlap: for some absolute constant $c \in (0, 1/2)$, $c < p_0(x,y), p_1(x,y) < 1-c$.

As a first contribution, we identify a condition on the classes $(\mathbb{P}, \mathbb{D})$ that will be crucial for the results on the identification of ATE that proceed.

Condition 1 (Identifiability Condition)**.** *The concept classes $(\mathbb{P}, \mathbb{D})$ satisfy the Identifiability Condition if for any tuples $(p, \mathcal{P}), (q, \mathcal{Q}) \in \mathbb{P} \times \mathbb{D}$ that are compatible with $(\mathbb{P}, \mathbb{D})$, at least one of the following holds:*

1. ***(Equivalence of Outcome Expectations)*** $\mathbb{E}_{(x,y)\sim\mathcal{P}}[y] = \mathbb{E}_{(x,y)\sim\mathcal{Q}}[y]$
2. ***(Distinction of Covariate Marginals)*** $\mathcal{P}_X \neq \mathcal{Q}_X$
3. ***(Distinction under Censoring)*** *There exist $x \in \text{supp}(\mathcal{P}_X)$ and $y \in \mathbb{R}$ such that $p(x,y)\mathcal{P}(x,y) \neq q(x,y)\mathcal{Q}(x,y)$.*

Observe that in Condition 1 we only focus on tuples that are compatible. This is due to the fact that incompatible tuples of $\mathbb{P} \times \mathbb{D}$ cannot be part of a realization of any valid observational study and hence properties of incompatible tuples are not relevant to our characterizations.

To gain some intuition for Condition 1, consider two observational studies $\mathcal{D}_1$ and $\mathcal{D}_2$ which correspond to the pairs $(p, \mathcal{P})$ and $(q, \mathcal{Q})$ respectively, where $\mathcal{P}$ and $\mathcal{Q}$ are distributions of $(X, Y(1))$. Assume that the true observational study $\mathcal{D}$ is either $\mathcal{D}_1$ or $\mathcal{D}_2$. Given the censored distribution $\mathcal{C}_{\mathcal{D}}$, we want to identify $\mathbb{E}_{\mathcal{D}}[Y(1)]$. First, suppose that the tuples $(p, \mathcal{P}), (q, \mathcal{Q})$ satisfy Requirement 1 in Condition 1. Then we are done since we only care about the expected outcomes $\mathbb{E}_{\mathcal{D}}[Y(1)] = \mathbb{E}_{(x,y)\sim\mathcal{P}}[y] = \mathbb{E}_{(x,y)\sim\mathcal{Q}}[y]$ which are the same under both distributions. Next, let us assume that Requirement 1 is violated and, hence, the expected treatment outcome is different between the null and the alternative hypothesis. In this case, if Requirement 2 is satisfied, then we can distinguish $\mathcal{P}$ and $\mathcal{Q}$ from $\mathcal{C}_{\mathcal{D}}$ (by comparing $\mathcal{P}_X$ and $\mathcal{Q}_X$ to the covariate marginal of $\mathcal{C}_{\mathcal{D}}$) and, hence, distinguish between $\mathcal{D}_1$ and $\mathcal{D}_2$. Finally, if both Requirements 1 and 2 fail but Requirement 3 holds, then we can use it to identify whether the true observational study $\mathcal{D}$ corresponds to $(p, \mathcal{P})$ or $(q, \mathcal{Q})$ as follows: First, we check if both $\int p(x,y)\mathcal{P}(x,y)\mathrm{d}x\mathrm{d}y$ and $\int q(x,y)\mathcal{Q}(x,y)\mathrm{d}x\mathrm{d}y$ are identical to $\Pr_{\mathcal{D}}[T{=}1]$. If this is not true, then we are done, as if $\int p(x,y)\mathcal{P}(x,y)\mathrm{d}x\mathrm{d}y \neq \Pr_{\mathcal{D}}[T{=}1]$, then $\mathcal{D}$ does not correspond to $(p, \mathcal{P})$. Otherwise, if this is true, then, in particular, $\int p(x,y)\mathcal{P}(x,y)\mathrm{d}x\mathrm{d}y = \int q(x,y)\mathcal{Q}(x,y)\mathrm{d}x\mathrm{d}y = \Pr_{\mathcal{D}}[T{=}1]$. Now, we can distinguish between $(p, \mathcal{P})$ and $(q, \mathcal{Q})$ by checking which one of $\frac{p(\cdot)\mathcal{P}(\cdot)}{\Pr_{\mathcal{D}}[T=1]}$ and $\frac{q(\cdot)\mathcal{Q}(\cdot)}{\Pr_{\mathcal{D}}[T=1]}$ matches the density of $(X, T{=}1, Y(1))$ in the *censored* distribution. This is where we use Requirement 3, due to which along with the previous observation, the densities $\frac{p(\cdot)\mathcal{P}(\cdot)}{\Pr_{\mathcal{D}}[T=1]}$ and $\frac{q(\cdot)\mathcal{Q}(\cdot)}{\Pr_{\mathcal{D}}[T=1]}$ are *not* identical in at least one point (x, y). Our first result shows Condition 1 fully characterizes ATE.

Theorem 1 (Identification of ATE)**.** *The average treatment effect $\tau_{\mathcal{D}}$ is identifiable from the censored distribution $\mathcal{C}_{\mathcal{D}}$ for any observational study $\mathcal{D}$ realizable with respect to $(\mathbb{P}, \mathbb{D})$ if and only if $(\mathbb{P}, \mathbb{D})$ satisfy Condition 1.*

The above result adds to classical identifiability conditions in Statistics (*e.g.*, [Everitt and Hand, 2013; Teicher, 1963]),

Statistical Learning Theory[4] (*e.g.*, [Angluin, 1980; Angluin, 1988]), and Econometrics (*e.g.*, [Manski, 1990; Athey and Haile, 2002]). To the best of our knowledge, this is the first tight characterization of ATE identification in observational studies.

4 Applications and Estimation of ATE

For Condition 1 to be useful given the other existing conditions (such as unconfoundedness and overlap), it needs to capture interesting examples *not captured by existing conditions*. Below, we revisit several well-studied scenarios in causal inference or their generalizations and, for each, provide identification results based on our results.

Scenario I: Unconfoundedness and Overlap. At the end of Section 2, we mentioned that our framework can capture unconfoundedness and overlap. Identification in this scenario is standard and can also be deduced using Theorem 1. Estimation is also standard [Imbens and Rubin, 2015].

Scenario II: Overlap without Unconfoundedness. Next, we consider observational studies $\mathcal{D}$ which satisfy c-overlap for some $c \in (0, 1/2)$ but may not satisfy unconfoundedness. We are going to use our framework to characterize the subset of these studies $\mathcal{D}$ for which ATE is identifiable. Since overlap holds with some parameter $c \in (0, 1/2)$, it restricts the concept class $\mathbb{P}$ to be $\mathbb{P}_{\mathrm{O}}(c)$ where $c < p(x, y) < 1 - c$ for any (x, y) and $p \in \mathbb{P}_{\mathrm{O}}(c)$. This case generalizes several models studied in the causal inference literature [Tan, 2006; Rosenbaum, 2002; Rosenbaum, 1987; Kallus and Zhou, 2021]; see the discussion below. Under this scenario, we can ask: which conditions should the covariate-outcome distributions $\mathbb{D}$ satisfy for τ to be identifiable, *i.e.*, for which observational studies realizable by $(\mathbb{P}_{\mathrm{O}}(c), \mathbb{D})$ is the ATE identifiable?

Informal Theorem 1. *Assume that for any pair $\mathcal{P}, \mathcal{Q} \in \mathbb{D}$, either Item 1 or 2 of Condition 1 holds or there exist $x \in \mathrm{supp}(\mathcal{P}_X)$, $y \in \mathbb{R}$ such that $\mathcal{P}(x, y) \notin (\frac{c}{1-c}, \frac{1-c}{c}) \cdot \mathcal{Q}(x, y)$. Then, $\tau_{\mathcal{D}}$ is identifiable from the censored distribution $\mathcal{C}_{\mathcal{D}}$ for any observational study $\mathcal{D}$ realizable with respect to $(\mathbb{P}_{\mathrm{O}}(c), \mathbb{D})$. Moreover, this condition is necessary.*

The above condition for identification is quite similar to Condition 1 and is satisfied by setting the outcome marginal of $\mathcal{P} \in \mathbb{D}$ to be, *e.g.*, Gaussian, Pareto, or Laplace, and letting the x-marginal $\mathcal{P}_X$ be unrestricted. This captures important practical models where the outcomes are modeled as a generalized linear model with Gaussian noise.

Connections to Prior Work. Since we do not require unconfoundedness in any form, the requirements on the generalized propensity score class $\mathbb{P}_{\mathrm{O}}(c)$, in this scenario, are very mild and are already satisfied by most existing frameworks that relax unconfoundedness while retaining overlap. The restriction on the propensity score class $\mathbb{P}_{\mathrm{O}}(c)$ relaxes the model of Tan [2006] and the odds-ratio model of Rosenbaum [2002], both widely used in sensitivity analysis.

[4]This mirrors language identification, whose characterizations also concern *pairs* of languages [Angluin, 1980].

Scenario III: Unconfoundedness without Overlap. We now consider the setting where overlap may fail but unconfoundedness holds. To rule out degenerate cases where everyone (or no one) receives the treatment, we assume that *some* nontrivial subset of covariates satisfies overlap. Concretely, there is a set $S \subseteq \mathbb{R}^d$ with Lebesgue measure $\mathrm{vol}(S) \geq c$ such that for each $(x, y) \in S \times \mathbb{R}$, we have $c < p_0(x, y)$, $p_1(x, y) < 1 - c$. This is already significantly weaker than the usual c-overlap assumption, which demands the previous inequalities *pointwise* for every $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. We relax it further into the notion of *c-weak-overlap* by removing the upper bound on $p_0(x, y)$ and $p_1(x, y)$. Then, we define the class $\mathbb{P} = \mathbb{P}_{\mathrm{U}}(c)$ which captures both unconfoundedness and c-weak-overlap.

Scenarios with unconfoundedness but without full overlap frequently arise in practice. Classic examples include regression discontinuity designs [Cook, 2008] and observational studies with extreme propensity scores [Crump *et al.*, 2009; Li *et al.*, 2018; Khan and Ugander, 2024; Kalavasis *et al.*, 2024]; see further discussion after Informal Theorem 2. As before, we ask which conditions on $\mathbb{D}$ enable identification of ATE, *i.e.*, for which observational studies realizable with respect to $(\mathbb{P}_{\mathrm{U}}(c), \mathbb{D})$, can one identify the ATE?

Informal Theorem 2. *Assume that for any pair $\mathcal{P}, \mathcal{Q} \in \mathbb{D}$, either Item 1 or 2 of Condition 1 holds or there is no set $S \subseteq \mathbb{R}^d$ with $\mathrm{vol}(S) \geq c$ such that $\mathcal{P}(x, y) = \mathcal{Q}(x, y)$ for $(x, y) \in S \times \mathbb{R}$. Then $\tau_{\mathcal{D}}$ is identifiable from the censored distribution $\mathcal{C}_{\mathcal{D}}$ for any $\mathcal{D}$ realizable with respect to $(\mathbb{P}_{\mathrm{U}}(c), \mathbb{D})$. Moreover, this condition is necessary.*

We would like to stress that the above characterization has a novel conceptual connection with an important field of statistics, called *truncated statistics* [Galton, 1897; Cohen, 1991]. The main task in truncated statistics concerns *extrapolation:* given a true density $\mathcal{D}$ over some domain X and a set $S \subseteq X$, the question is whether the structure of $\mathcal{D}$ can be identified from *truncated* samples, *i.e.*, samples from the conditional density of $\mathcal{D}$ on S. The condition of the above result requires the pairs $\mathcal{P}, \mathcal{Q}$ to be distinguishable on any set of the form $S \times \mathbb{R}$ (where S has sufficient volume). Roughly speaking, this condition holds for any family $\mathbb{D}$ whose elements $\mathcal{P}$ can be extrapolated given samples from their truncations to full-dimensional sets — which is well-studied and provides us with multiple applications [Kontonis *et al.*, 2019; Daskalakis *et al.*, 2021; Lee *et al.*, 2024; Lee *et al.*, 2026].

Connections to Prior Work. This scenario captures two important and practical settings. First, as mentioned before, it captures regression discontinuity (RD). These designs were introduced by Thistlethwaite and Campbell [1960], were independently discovered in many fields [Cook, 2008], and have found applications in various contexts from education to public health to labor economics. To the best of our knowledge in RD designs, ATE is only known to be identifiable under strong linearity assumptions on the expected outcomes [Hahn *et al.*, 2001]. Due to that, recent work focuses on identifying certain local treatment effects [Imbens and Lemieux, 2008]. In contrast, Informal Theorem 2 enables us to achieve identification under much weaker restrictions.

Ethical Statement

There are no ethical issues.

References

[Alon *et al.*, 1997] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-Sensitive Dimensions, Uniform Convergence, and Learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[Angluin, 1980] Dana Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45(2):117–135, 1980.

[Angluin, 1988] Dana Angluin. *Identifying Languages from Stochastic Examples*. Yale University. Department of Computer Science, 1988.

[Anthony and Bartlett, 1999] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[Athey and Haile, 2002] Susan Athey and Philip A. Haile. Identification of standard auction models. *Econometrica*, 70(6):2107–2140, 2002.

[Athey and Imbens, 2017] Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, May 2017.

[Austin, 2008] Peter C. Austin. A Critical Appraisal of Propensity-Score Matching in the Medical Literature Between 1996 and 2003. *Statistics in Medicine*, 27(12):2037–2049, 2008.

[Blumer *et al.*, 1989] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[Bousquet *et al.*, 2003] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

[Bousquet, 2003] Olivier Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 55:371–389, 2003.

[Brunell and DiNardo, 2004] Thomas L. Brunell and John DiNardo. A Propensity Score Reweighting Approach to Estimating the Partisan Effects of Full Turnout in American Presidential Elections. *Political Analysis*, 12(1):28–45, 2004.

[Cai *et al.*, 2025] Yang Cai, Alkis Kalavasis, Katerina Mamali, Anay Mehrotra, and Manolis Zampetakis. What makes treatment effects identifiable? Characterizations and estimators beyond unconfoundedness. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of the Thirty-Eighth Conference on Learning Theory (COLT 2025)*, volume 291 of *Proceedings of Machine Learning Research*, pages 755–756. PMLR, 2025.

[Chernozhukov *et al.*, 2024] Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. Applied causal inference powered by ml and ai, 2024.

[Cochran, 1965] William G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

[Cohen, 1991] A Clifford Cohen. *Truncated and Censored Samples: Theory and Applications*. CRC press, 1991.

[Cook, 2008] Thomas D. Cook. "waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654, 2008. The regression discontinuity design: Theory and applications.

[Crump *et al.*, 2009] Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika*, 96(1):187–199, 01 2009.

[Daskalakis *et al.*, 2021] Constantinos Daskalakis, Vasilis Kontonis, Christos Tzamos, and Emmanouil Zampetakis. A statistical taylor theorem and extrapolation of truncated densities. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1395–1398. PMLR, Aug 2021.

[D'Amour *et al.*, 2021] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

[Everitt and Hand, 2013] Brian Everitt and David J. Hand. *Finite Mixture Distributions*. Springer Science & Business Media, 2013.

[Galton, 1897] Francis Galton. An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1897.

[Hahn *et al.*, 2001] Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

[Hastie *et al.*, 2013] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2013.

[Heckman, 1979] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[Hernan and Robins, 2023] Miguel A. Hernan and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC monographs on statistics & applied probability. Taylor & Francis, 2023.

[Holland, 1986] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[Imbens and Lemieux, 2008] Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008. The regression discontinuity design: Theory and applications.

[Imbens and Rubin, 2015] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[Kalavasis *et al.*, 2024] Alkis Kalavasis, Anay Mehrotra, and Manolis Zampetakis. Smaller confidence intervals from ipw estimators via data-dependent coarsening (extended abstract). In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2767–2767. PMLR, 30 Jun–03 Jul 2024.

[Kallus and Zhou, 2021] Nathan Kallus and Angela Zhou. Minimax-Optimal Policy Learning Under Unobserved Confounding. *Management Science*, 67(5):2870–2890, 2021.

[Kearns and Schapire, 1994] Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[Khan and Ugander, 2024] Samir Khan and Johan Ugander. Doubly-robust and heteroscedasticity-aware sample trimming for causal inference. *Biometrika*, page asae053, 10 2024.

[Kontonis *et al.*, 2019] Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. Efficient Truncated Statistics with Unknown Truncation. *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1578–1595, 2019.

[Lee *et al.*, 2024] Jane H. Lee, Anay Mehrotra, and Manolis Zampetakis. Efficient statistics with unknown truncation, polynomial time algorithms, beyond gaussians. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 988–1006, 2024.

[Lee *et al.*, 2026] Jane H. Lee, Anay Mehrotra, and Manolis Zampetakis. Smoothed analysis of learning from positive samples, 2026. Accepted for presentation at the ACM Symposium on Theory of Computing (STOC 2026). https://arxiv.org/abs/2504.10428.

[Li *et al.*, 2018] Fan Li, Laine E. Thomas, and Fan Li. Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*, 188(1):250–257, 09 2018.

[Lugosi, 2002] Gábor Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer, 2002.

[Manski, 1990] Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

[Massart and Nédélec, 2006] Pascal Massart and Élodie Nédélec. Risk Bounds for Statistical Learning. *The Annals of Statistics*, 34(5):2326 – 2366, 2006.

[Morgan and Harding, 2006] Stephen L. Morgan and David J. Harding. Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. *Sociological Methods & Research*, 35(1):3–60, 2006.

[Neyman, 1990] Jerzy Splawa Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

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

[Rosenbaum, 1987] Paul R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

[Rosenbaum, 2002] Paul R. Rosenbaum. *Observational Studies*. Springer Series in Statistics. Springer, 2002.

[Rubin, 1974] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[Rubin, 1977] Donald B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26, 1977.

[Rubin, 1978] Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

[Tan, 2006] Zhiqiang Tan. A Distributional Approach for Causal Inference Using Propensity Scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[Teicher, 1963] Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 1265–1269, 1963.

[Thistlethwaite and Campbell, 1960] Donald L. Thistlethwaite and Donald T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960.

[Valiant, 1984] Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 11 1984.

[Vapnik, 1999] Vladimir N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[Vapnik, 2006] Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Science & Business Media, 2006.