How Much Can We Ignore the Ignorability Assumption in Digital Platforms?

Omid Rafieian*
Cornell Tech and Cornell University

March 15, 2022

*Please address all correspondence to: or83@cornell.edu.

Abstract

Digital platforms deliver numerous interventions to their users. One of their main goals is to estimate the causal effect of these interventions. An ideal way to answer this question is to run a fully randomized experiment. However, the economic cost of such experiments is high, making alternative approaches based on observational data appealing to digital platforms. In this paper, we study the feasibility of using observational methods in the presence of algorithmic decision-making. A central assumption needed for observational studies is strong ignorability assumption, which requires the unconfoundedness of the treatment assignment, and an often-ignored part: overlap assumption that requires the assignment to be non-deterministic. Although the setting created by algorithmic decision-making satisfies the unconfoundedness assumption as the assignment rule is known, the overlap assumption is often violated because these algorithms generate deterministic recommendations. We theoretically show that the violation of overlap can substantially bias the estimates of the average treatment effect from observational data. We quantify this bias and discuss whether it is practically relevant in digital platforms. We then focus on advertising auctions to show the prevalence of this problem by showing an impossibility result in observational advertising effectiveness studies. Finally, we propose a novel solution based on machine learning methods used for matrix completion that allows us to recover the average treatment effect estimates if the underlying treatment space is low-rank. We use our algorithm to develop statistical tests to examine whether the lack of overlap results in substantial bias in the main estimates of the causal parameters.

Keywords: causal inference, machine learning, overlap assumption, unconfoundedness, digital platforms, observational methods

1 Introduction

Digital platforms deliver numerous interventions to their users every day. These interventions can take different forms, such as advertisements on websites, push notifications on mobile phones, etc. At the core of this large-scale delivery of interventions are two elements: data collection and algorithmic decision-making. Digital platforms collect a massive amount of data from the users of their basic information such as demographics and their behavioral characteristics such as their past browsing history. These data are then given as inputs to algorithms that can efficiently process them and make real-time decisions, allowing the platform to deliver interventions at a very large scale.

An important question that digital platforms and academic researchers want to know the answer to is the causal effect of these interventions. The gold standard in both research and practice is to use randomized controlled trials (RCT) where some users randomly receive the treatment, and some do not. This randomization, in turn, allows us to estimate the causal effect of an intervention. However, running fully randomized experiments is not always in the interest of the platform. The very reason why many digital platforms use algorithmic decision-making is to deliver optimal interventions, depending on their economic goal. From that perspective, experiments can come at the expense of this economic goal since the platform needs to assign a large group of users to sub-optimal interventions. Thus, it is crucially important for these platforms to estimate the causal effects of interventions with their existing observational data.

Both experimental and observational methods to estimate the causal effect of an intervention rely on a set of assumptions called strong ignorability of the treatment assignment. Strong ignorability assumption is a mix of two assumptions: (1) *unconfoundedness* of the treatment assignment, which states that conditional on observed covariates, assignment to the treatment is independent of potential outcomes, and (2) *overlap* or *positivity* of the treatment assignment, which assumes that the assignment to the treatment is probabilistic, that is the propensity score of the treatment is a probability strictly between zero and one. The part that is often violated in observational studies is the unconfoundedness assumption. That is, there are unobserved confounding factors that affect both the treatment assignment and the outcome of interest. The presence of confounding, therefore, hampers researchers' ability to draw causal inference from observational studies.

What is different in digital platforms is that the unconfoundedness assumption is more plausible than most settings. This is because the platform itself delivers the interventions to users. As such, given the output of the algorithm used for decision-making at the digital

platform, the assignment to a treatment is unconfounded. Even if the researcher does not have access to the algorithmic output but the data used for algorithmic decision-making, it is still possible to satisfy the unconfoundedness assumption by learning the underlying selection mechanism from data. This is increasingly an easier task with the development of methods that combine causal inference with machine learning methods to capture complex confoundedness in the data. Thus, the presence of the exact algorithm or high-dimensional data used for algorithmic decision-making serves as a strong motivation for using observational methods in the context of digital platforms.

What arises as an important challenge is an often-ignored part of the ignorability assumption: overlap or the requirement for the probabilistic assignment. Although algorithmic decision-making helps platforms better use their interventions, many of these algorithms only generate deterministic outputs. That is, one intervention will be shown with probability one, and the rest of the interventions have zero probability of being shown. A canonical example of such algorithmic decision-making is advertising auctions, where an ad impression is awarded to the ad with the highest bid (or some scoring output determined by the platform). In these cases, the overlap assumption is violated, which leaves us with no theoretical guarantee on the estimated treatment effects. In this paper, we consider the case for a digital platform whose context satisfies unconfoundedness assumption because the algorithmic outputs are readily available at the platform, but violates the overlap assumption to the deterministic assignment employed by the algorithms. To that end, we seek to answer the following set of research questions:

1. How does the lack of overlap bias the estimates of average treatment effect in observational studies that satisfy unconfoundedness assumption? Can the state-of-the-art model-based and model-free approaches overcome this challenge?

2. How likely is this lack of overlap to cause bias in the average treatment effect estimates from a practical standpoint?

3. What are the solutions to this problem, and under what assumptions do they work?

To answer these questions, we develop a simple framework that distinguishes between three regions in the data based on the treatment assignment: (1) probabilistic assignment, where the propensity score of the assignment is a number in the non-exclusive interval of $(0, 1)$, (2) deterministic assignment, where the treatment assignment happens deterministically with probability one, i.e., propensity score for the treatment is one, and (3) deterministic no-assignment, where the treatment assignment will not happen with probability one, i.e., the

propensity score for the treatment is zero. As such, the only region that satisfies the overlap assumption is the one with the probabilistic assignment. We further define three conditional average treatment effect (CATE) for each of these three regions to allow for the possibility that these estimands are different at the population level. This allows us to say something concrete and testable about the magnitude of bias in our treatment effect estimates that is caused by the lack of overlap.

Our theoretical analysis first shows that the conditional average treatment effect for the regions with deterministic assignment is unidentified. We then consider the case where we use the data from all the three regions with a known propensity score and examine how well we can estimate the average treatment effect in this case. This mimics the setting at digital platforms where the propensity scores are either known ex-ante or can be estimated accurately. We also focus on the state-of-the-art model-based and model-free approaches to estimate the average treatment effects such as double machine learning [Chernozhukov et al., 2018a] and causal forests [Athey et al., 2019] to ensure that a poor modeling choice does not drive the results of our analysis. Our analysis shows that all these methods can result in substantial bias due to the lack of overlap even when the propensity score is known. In cases where the propensity scores need to be estimated, this bias can be considerably larger.

On the bright side, our analysis shows that if the propensity scores are known, a large class of observational methods can recover the only identifiable causal estimand in the data, the conditional average treatment effect for the region with a probabilistic assignment. This finding allows us to quantify the magnitude of bias in a concrete manner and arrive at a critical practical insight: if the conditional average treatment effect of the probabilistic region is the same as the average treatment effect for the entire population, our observational estimates of the average treatment effects converge to the true population parameters. This is the key assumption under the trimming techniques that suggest dropping observations with the propensity scores close to zero or one. Thus, the important practical question is whether we should expect the conditional average treatment effect for the probabilistic region in our data to be the same as the average treatment effect for the entire population.

In principle, if the assignment probability is a function of the conditional average treatment effect for an observation, the lack of overlap likely results in large biases in the estimates of average treatment effects. The problem is that if the digital platform is also interested in optimizing the same causal estimand, the optimal strategy for them is to assign interventions based on their conditional average treatment effects [Shalit et al., 2017, Wager and Athey, 2018]. To show how it arises in real settings, we consider one of the most canonical cases

of algorithmic decision-making in digital platforms: advertising auctions. We show that observational data of digital advertising platforms suffer from an extreme case of lack of overlap. Digital platforms run standard auctions such as the second- of first-price auctions that deterministically allocate the impression to the ad with the highest bid. As such, all units in our data violate the overlap assumption, which results in the unidentifiability of the average treatment effect if the propensity scores are known. If the propensity scores are unknown and need to be estimated, the analysis can exhibit very large biases. This is particularly the case because, in digital advertising auctions, advertisers' equilibrium bid theoretically has a direct relationship with their conditional average treatment effect.

Once we establish the existence and prevalence of the lack of overlap in observational studies involving digital platforms, we focus on the potential solutions for this problem. We propose a framework that formulates the unidentifiability of the conditional average treatment effect for the overlap-violating regions of the data as a missing data problem. As such, if we have a relatively large space of treatments that is low-rank, we can use matrix completion methods to impute the conditional average treatment effect for the overlap-violating regions. The intuition is that if there is a unit for which the assignment is deterministic and CATE is not estimable, we can use the estimates for other treatments for the same unit when the assignment is probabilistic and therefore CATE is estimable. Once we complete the matrix for the parts that are formerly unidentified, we can statistically test whether the magnitude of bias due to the violation of overlap is significantly different from zero.

We then extend this statistical test to a fully observational case, where the propensity scores are unknown. The main challenge in these settings is that we cannot easily distinguish the region with a probabilistic assignment from those with a deterministic assignment. As such, the classification rule is often arbitrary, depending on the sensitivity of the study. We develop a test that finds the magnitude of bias for different trimming thresholds. For example, for a 0.1 threshold, we consider the estimated propensities between 0 to 0.1 as a deterministic no-assignment, those between 0.1 and 0.9 as a probabilistic assignment, and those between 0.9 and 1 as a deterministic assignment. Once we run this procedure for every trimming threshold, we will get a curve that allows testing the magnitude of bias caused by the lack of overlap in our study.

Finally, we consider two design-based solutions to address the overlap violation caused by algorithmic decision-making. To that end, we consider two potential solutions that consider platforms' economic goals while designing assignment rules that satisfy the overlap assumption. First, digital platforms can leave a small portion of their total traffic for the purpose of

experimentation and use it to extract generalizable insights for the rest of their traffic. The second alternative is to employ a version of ϵ-greedy assignment where no observation is assigned deterministically to units, but with probability $1 - \epsilon$, and the rest remaining ϵ probability is distributed across other actions such that any action has a non-zero propensity score.

In sum, our paper makes several contributions to the literature. First, we identify an important challenge for the digital platforms that employ algorithmic decision-making. While most of the applied causal inference literature is focused on satisfying unconfoundedness using state-of-the-art causal machine learning methods, we show that the fundamental problem in digital platforms is, in fact, the overlap violation. We further quantify the bias caused by the violation of the overlap assumption and discuss when we should expect this bias to be higher. From a practical standpoint, our paper narrows down the modeling focus on advertising auctions and shows that the problem goes beyond a theoretical possibility. We prove an important impossibility result, which states that the causal estimands cannot be identified under observational studies on advertising platforms that run standard auctions. Finally, we propose a novel machine learning approach for matrix completion that can be used to correct for the biased caused due to the lack of overlap. Our approach only requires a large treatment space, which makes it easily applicable to digital platforms that deliver numerous interventions to their users. We further develop a statistical test that can be used to assess whether the lack of overlap is detrimental to an observational study.

2 Related Literature

Broadly, our paper relates to the causal inference literature that aims to estimate treatment effects [Neyman, 1923, Rubin, 1974, Imbens and Rubin, 2015]. Following the influential paper by [Rosenbaum and Rubin, 1983], much of this literature focuses on a set of assumptions known as the strong ignorability of the treatment assignment, which is a combination of two assumptions: unconfoundedness and overlap. While unconfoundedness has received considerable in the literature, overlap has often been viewed as an easier assumption to be satisfied in real settings. As such, less attention has been paid to the overlap assumption in prior studies on causal inference with a few notable exceptions that focus on various aspects of the overlap assumption such as studying sample trimming strategies [Crump et al., 2009, Ma and Wang, 2020, D'Amour et al., 2021] and quantifying the uncertainty in overlap-violating regions of observational data [Jesson et al., 2020]. Motivated by the context of algorithmic decision-making in digital platforms and the prevalent violation of this assumption in such contexts, we study the overlap assumption – how it arises and what theoretical implications it

has who want to estimate treatment effects. We contribute to this literature by characterizing the bias induced by the lack of overlap and identifying cases where the lack of overlap can be detrimental in the sense that the conventional solutions such as using more competent causal machine learning models and sample trimming do not solve the problem. We further add to this literature by proposing a machine learning approach based on matrix completion that imposes low-rank assumptions on the treatment space to help researchers test whether the lack of overlap can cause substantial bias in treatment effects and correct for this bias.

Second, our paper relates to the literature on the growing intersection of machine learning and causal inference. In recent years, a series of papers combined the insights from the causal inference literature with the flexibility and scalability of machine learning models in learning patterns from data to develop new methods to estimate causal estimands such as average treatment effect [Belloni et al., 2014, Hartford et al., 2017, Chernozhukov et al., 2018a, Athey et al., 2018, Shi et al., 2019] or conditional average treatment effect [Shalit et al., 2017, Athey et al., 2019, Chernozhukov et al., 2018b, Nie and Wager, 2021]. In marketing, many recent papers used these methods in a variety of application domains such as personalized promotions [Simester et al., 2020a,b], customer relationship management [Ascarza, 2018], personalized free-trial [Yoganarasimhan et al., 2022], ad targeting and sequencing [Rafieian and Yoganarasimhan, 2021, Rafieian, 2022], and personalized versioning [Goli et al., 2022b]. We add to this literature in two separate ways. First, we theoretically characterize the performance of causal machine learning methods when the overlap assumption is violated. Second, we propose a machine learning algorithm that exploit the similarities between the treatments in the treatment space and overcomes the issue of overlap violation under certain assumptions.

Finally, our paper contributes to the literature on the observational studies on advertising effectiveness. In the context TV advertising, Shapiro et al. [2021] consider a series of observational methods to account for the endogeneity of ad assignment and estimate ad effect for 288 brands. In the context of digital advertising, the endogeneity of ad assignment is more prevalent, which motivated empirical research on the possibility of causally estimating the advertising effectiveness. Lewis et al. [2011] were the one of the first to emphasize how inaccurate observational studies may be due to the presence of hidden confounding. While this insight has been echoed in other studies in different contexts [Blake et al., 2015, Gordon et al., 2019, 2022], other papers show more promising results on the ability of observational methods in recovering the causal ad effects [Hoban and Arora, 2018, Tunuguntla], thereby resulting in an overall lack of consensus in the literature about the possibility of recovering

the true ad effects. Yet, none of these papers examines this possibility from a theoretical standpoint. In particular, the problem of lack of overlap is ignored and the question boils down to whether observational methods can control for the confoundedness of ad assignment. Our main contribution to this stream of literature is to theoretically study how the lack of overlap may arise in these situations and how this problem can bias our analysis.

3 Algorithmic Decision-making

3.1 Problem Definition

We start with a formal definition of our problem. Consider a general case where a digital platform delivers interventions to observation units. The observation unit is often a user in digital platforms. When an observation unit is available to receive the intervention, the platform chooses from the set of all interventions, which is denoted by $\mathcal{W}$ in our problem. For example, this set can be the list of different ads to show to the user. For observation i, let W_i denote the intervention delivered to the user, and X_i denote the vector of observation characteristics from the super set $\mathcal{X}$. As customary in digital platforms, the vector of characteristics X_i is often high-dimensional with detailed information about the user such as demographics and past user history, as well as contextual factors such as the timestamp of the observation.

To determine which intervention to deliver in each observation, digital platforms generally use an algorithm that scalably uses the feature vector X_i and returns an intervention that optimizes the platform's objective. For any intervention $w \in \mathcal{W}$, we formalize this algorithmic policy as a function $\pi_w : \mathcal{X} \to [0, 1]$, where $\pi_w(X_i)$ determines the probability that the platform chooses intervention w in observation i. The function π_w is the same as the propensity score function in the causal inference literature. Digital platforms often have access to this function.

Once the intervention is delivered, the platform collects the outcome of interest Y_i for observation i. This outcome is defined based on the problem under the study. For example, this outcome would be clicks or sales for advertising. Following the potential outcomes framework, we define $Y_i(w)$ for each $w \in \mathcal{W}$ as the potential outcome we would have observed under intervention w. For simplicity and greater consistency with the causal inference literature, we focus our analysis on the binary case with one treatment and one control group.[1] As such, $W_i = 1$ means that observation i has received the treatment, whereas $W_i = 0$ refers to the case where where observation i has received the control. Hence, for each

[1]The results are easily generalizable to the case with multiple treatment levels.

observation i, there are two potential outcomes $Y_i(0)$ and $Y_i(1)$.

With this notation in place, we now define two estimands that researchers and practitioners often want to estimate as follows:

Definition 1. *The Average Treatment Effect (ATE) is denoted by τ^* and defined as follows:*

$$\tau^* = \mathbb{E}[Y_i(1) - Y_i(0)], \tag{1}$$

where the expectation is taken over the entire population.

The Conditional Average Treatment Effect (CATE) is the same as ATE conditional on a certain value of the covariate space. We denote CATE as $\tau^*(x)$ and define it as follows:

$$\tau^*(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]. \tag{2}$$

The prior literature on causal inference has proposed a wide variety of methods to estimate ATE and CATE [Imbens and Rubin, 2015]. These methods require a set of assumptions known as (1) *Stable Unit Treatment Value Assumption (SUTVA)*, and (2) *Strong Ignorability of Treatment Assignment*. SUTVA states that there is a single version of each treatment and the units do not interfere with each other. In digital settings where treatments are well-defined with a single version and a unit's treatment status and action is isolated in the sense that it does not change the treatment status of other units, SUTVA would be more plausible. In this paper, we consider the cases where SUTVA holds to exclusively focus on cases where the ignorability assumption is violated.[2]

The second set of assumptions is known as *Strong Ignorability* assumption, which is defined in the seminal paper by Rosenbaum and Rubin [1983] as follows:

Definition 2. *The assignment to treatment is strongly ignorable given the observed covariates X_i, if we have:*

- *Unconfoundedness: The potential outcomes are independent of the treatment assignment conditional on observed covariates:*

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X_i, \tag{3}$$

which is known as the unconfoundedness assumption and referred to with other names such as selection on observables, conditional exogeneity, etc.

[2] A series of recent studies show cases where SUTVA is violated in digital settings. Please see Goli et al. [2022a] for a great summary of these cases.

- *Overlap: The assignment to the treatment is probabilistic, that is:*

$$0 < \Pr(W_i = 1 \mid X_i) < 1, \tag{4}$$

where $\Pr(W_i = 1 \mid X_i)$ *is the same as the propensity score when* $w = 1$*, that is,* $\pi(X_i)$*.*[3] *This assumption is often referred to as the overlap or positivity assumption and guarantees that the assignment to the treatment is not deterministic.*

The strong ignorability assumption serves as the foundation for studies of causal inference. The most common challenge in these studies is often the unobservability of the assignment rule, which results in the confoundedness of the treatment. That is, there is an unobservable variable Z_i that affects both the treatment assignment and the outcome, thereby resulting in selection bias in the estimates of average treatment effect.

The key difference in digital platforms that employ algorithmic decision-making is that the assignment rule is often fully observable. That is, the platform can easily store the X_i used for algorithmic decision-making and the output of the algorithm $\pi(X_i)$, which is shown to be sufficient to satisfy the unconfoundedness assumption [Rosenbaum and Rubin, 1983]. Hence, observational studies on digital platforms do not suffer from the well-known confoundedness or endogeneity problem, since there is no selection on unobservables. What makes these observational studies challenging is the commonly ignored part of the strong ignorability assumption, which requires the treatment assignment to be probabilistic. Although probabilistic assignment is plausible in more traditional studies without algorithmic decision-making in the background, algorithms used by digital platforms to deliver interventions are often deterministic. That is, $\pi(X_i)$ can be equal to zero or one depending on X_i.

Our goal in this paper is to study the consequences of this lack of overlap in observational studies on digital platforms. As such, we consider a digital platform that uses data $\mathcal{D} = \{Y_i, W_i, X_i, \pi(X_i)\}$ to estimate the treatment effect estimands: ATE and CATE. To that end, we want to (1) quantify the magnitude of bias due to this overlap violation, (2) identify the link between this bias and the algorithm used by the platform, and (3) discuss potential solutions to overcome this problem.

3.2 Analysis

In this section, we theoretically analyze how the lack of overlap can lead to biased estimates of the average treatment effect (ATE). We start by showing the identification problem with the lack of overlap in observational data in §3.2.1. We then examine how the model-based

[3]For brevity, instead of $\pi_1(X_i)$, we use $\pi(X_i)$.

approaches like double machine learning performs in estimating the ATE in §3.2.2. Finally, we focus on model-free approaches such as importance sampling and theoretically derive their properties in §3.2.3.

3.2.1 Identification Challenge

It is well-known that the violation of the overlap assumption can bias the estimates of the ATE. In this section, we illustrate this point by presenting a simple framework that we can use for our subsequent analysis. To do so, we first introduce new notation that capture the difference between different parts of the covariate space. In particular, we focus on the conditional average treatment effect for three separate groups of observation units:

- Probabilistic assignment region ($0 < \pi(X_i) < 1$): For observations where $0 < \pi(X_i) < 1$, we define $\tau_r^* = \mathbb{E}[Y_i(1) - Y_i(0) \mid 0 < \pi(X_i) < 1]$, which is the average treatment effect for the observations that have a probabilistic assignment. We denote the fraction of such observations in our data as α_r.
- Deterministic assignment region ($\pi(X_i) = 1$): For observations where $\pi(X_i) = 1$, we define $\tau_1^* = \mathbb{E}[Y_i(1) - Y_i(0) \mid \pi(X_i) = 1]$, which is the average treatment effect for observations where the assignment to the treatment certainly happens. We denote the fraction of such observations in our data as α_1.
- Deterministic no-assignment region ($\pi(X_i) = 0$): For observations where $\pi(X_i) = 0$, we define $\tau_0^* = \mathbb{E}[Y_i(1) - Y_i(0) \mid \pi(X_i) = 0]$, which is the average treatment effect for observations that certainly do not receive the treatment. We denote the fraction of such observations in our data as α_0.

Now, we can define the average treatment effect as $\tau^* = \alpha_r\tau_r^* + \alpha_1\tau_1^* + \alpha_0\tau_0^*$, where $\alpha_r + \alpha_1 + \alpha_0 = 1$. This decomposition allows us to highlight where the deterministic assignment creates a problem. Suppose that the digital platform wants to use data $\mathcal{D}$ to estimate τ_1^*. The problem is that for this slice of the population, the treatment variable is perfectly correlated with the propensity score, that is, $W_i = \pi(X_i) = 1$. The same problem is present in identifying τ_0^*. Thus, we can write the following lemma:

Lemma 1. *The conditional average treatment effects τ_1^* and τ_0^* are unidentifiable given data $\mathcal{D}$.*

Proof. *There is no variation in the treatment variable to estimate $\tau_j^* = \mathbb{E}[Y_i(1) - Y_i(0) \mid \pi(X_i) = j]$ for $j \in \{0, 1\}$.*

In light of Lemma 1, the only identifiable piece of τ^* is τ_r^*. We now want to see how this identification problem manifests itself in both model-based and model-free approaches to

estimate causal estimands.

3.2.2 Model-based Approaches to Estimate ATE

There are many model-based approaches one could use to estimate ATE from observational data. The traditional approach is to use a linear regression that projects the outcome on the treatment variable as well as other controls and estimate the average treatment effect. These methods work well if the confoundedness in the treatment assignment is captured by a linear combination of covariates. However, in many high-dimensional settings, the assignment has more complex patterns, which makes linear controls inadequate in accounting for observed confoundedness. Further, the relationship between other covariates and the outcome can also follow a non-linear pattern. These limitations, in turn, attracted a growing body of work that brings machine learning methods to casual inference in order to increase flexibility and robustness of model-based methods to estimate ATE [Belloni et al., 2014, Hartford et al., 2017, Chernozhukov et al., 2018a, Shi et al., 2019]. Many of these methods are now considered as the state-of-the-art methods for estimating the ATE. Our goal is to quantify the magnitude of bias when we use these methods to estimate the causal estimands.

We present a general framework to study model-based approaches. Let $\mu_w(x) = \mathbb{E}[Y_i(w) \mid X_i = x]$ denote the underlying population model for the conditional potential outcomes for any w. We can write:

$$Y_i(w) = \mu_0(X_i) + \tau^*(X_i)w + \epsilon_i(w), \tag{5}$$

where $\epsilon_i(w)$ denotes the structural error term for any value of the treatment $w \in \{0, 1\}$. Unconfoundedness implies that $\mathbb{E}[\epsilon_i(W_i) \mid X_i, W_i] = 0$. We further define function m as the conditional mean function such that $m(x) = \mathbb{E}[Y \mid X = x]$. We can now write the following decomposition:

$$Y_i - m(X_i) = (W_i - \pi(X_i))\,\tau^*(X_i) + \epsilon_i(W_i), \tag{6}$$

which holds because $m(X_i) = \mu_0(X_i) + \tau^*(X_i)\pi(X_i)$. This decomposition – which is first proposed by Robinson [1988] for estimating partially linear models – serves as a foundation for model-based approaches to estimated ATE or CATE that use machine learning models for causal inference. The key insight is that we can use machine learning models to flexibly learn nuisance functions $m(X_i)$ and $\pi(X_i)$, and then feed these estimates into an objective function to estimate causal estimands. We can define this objective function as follows:

$$\tau^*(\cdot) = \underset{\tau}{\operatorname{argmin}}\, \mathbb{E}\left[\left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau(X_i)\right)^2\right]. \tag{7}$$

The double machine learning (DML) approach estimates both nuisance functions using machine learning models and then estimate the ATE using a version of the objective function above, where there is only one $\tau(X_i)$ for the population [Chernozhukov et al., 2018a]. A series of methods use this decomposition to estimate heterogeneous treatment effects or CATE by using random forests [Athey et al., 2019], or more broadly any loss minimization method [Nie and Wager, 2021, Chernozhukov et al., 2018b]. We now use this objective function to prove the following proposition:

Proposition 1. *Suppose that there is a digital platform that has access to data* $\mathcal{D} = \{Y_i, W_i, X_i, \pi(X_i)\}$*, where* $\pi(X_i)$ *is known, but takes values zero and one for parts of the population. The estimated average treatment effect (ATE)* $\hat{\tau}$ *under any method that uses the objective function in Equation 7 converges to* τ_r^* *in probability, that is:*

$$\hat{\tau} \xrightarrow{p} \tau_r^* \tag{8}$$

Proof. *Let* $\mathcal{I}_r$ *denote the set of observations that have probabilistic assignment. We denote the total number of these observations by* N_r*. From Chernozhukov et al. [2018a], we know that:*

$$\underset{\tau}{\operatorname{argmin}} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \xrightarrow{p} \tau_r^*. \tag{9}$$

We now want to show that the RHS of Equation (9) *is the same as what any methods optimizing Equation* (7) *would estimate. We can write:*

$$\begin{aligned}
\hat{\tau} &= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&\quad + \sum_{i \notin \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 + \sum_{i \notin \mathcal{I}_r} \left(Y_i - m(X_i)\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2,
\end{aligned} \tag{10}$$

where the second line is a simple decomposition based on the observations with probabilistic and deterministic assignment, the fourth line is because $W_i - \pi(X_i) = 0$ *for observations with deterministic assignment, the fifth line drops the term* $\sum_{i \notin \mathcal{I}_r} (Y_i - m(X_i))^2$ *because it is invariant of* τ*, and the sixth line changes* $1/N$ *to* $1/N_r$ *because it is invariant of* τ*. Now if we combine the result of Equation* (10) *with that of Equation* (9)*, the proof is complete.*

This proposition shows that methods such as double machine learning or causal forests estimate τ_r^* as the ATE when the propensity is known. As such, to the extent that τ_r^* is different from τ^*, the estimate for the ATE would be biased. Given that τ_r^* appears in the equation for τ^*, the question is if there is any bound for the magnitude of bias. In light of Proposition 1, we know that the magnitude of bias is $|\tau^* - \hat{\tau}| \xrightarrow{p} |(\alpha_r - 1)\tau_r^* + \alpha_1 \tau_1^* + \alpha_0 \tau_0^*|$ such that $\alpha_r + \alpha_1 + \alpha_0 = 1$, which allows us to further simplify this expression to the following:

$$|\tau^* - \hat{\tau}| \xrightarrow{p} |\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)|. \tag{11}$$

This simplification highlights the fact that if the treatment effect for the deterministic regions is the same as the treatment effect for the probabilistic region, there will be no bias. However, it is easy to imagine scenarios where the difference in τ_1^*, τ_0^*, and τ_r^* creates substantial bias in estimates of the average treatment effect. We formalize this intuition in the following corollary :

Corollary 1. *The magnitude of bias can be any arbitrary amount if either* α_0 *or* α_1 *is non-zero.*

Proof. *The proof is simple based on unidentifiability of* τ_1^* *and* τ_0^**, in conjunction with the fact that* α_0 *and* α_1 *are not both equal to zero. As such, for any constant* c*, we can find* τ_0 *and* τ_1 *such that* $|\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)| = c$*.*

While Corollary 1 shows that the bias can be of any magnitude if we have deterministic assignment in our population, the bright side is that all the methods that use the objective function in Equation (7) are able to recover the only identifiable part of τ^*. That is, the presence of deterministic assignment for parts of the population does not result in biased estimates of the region with the probabilistic assignment. Hence, the researcher can rely on the estimates as consistent estimators of the true population parameters for the region with the probabilistic assignment. However, it is important to notice that this is only the case when propensity scores are known, which allows the optimizer to ignore those elements of the objective function because $W_i - \pi(X_i) = 0$. The situation would be very different if the propensity scores are to be estimated. This is because $W_i - \pi(X_i)$ in the objective function

will no more be zero but a very small number, which may largely bias the ATE estimate as the optimizer attempts to minimize the loss in Equation (7) by assigning a very large weight to τ^*. The following corollary summarizes this point:

Corollary 2. *If the propensity scores are not known, the estimated average treatment effect (ATE) $\hat{\tau}$ under any method that uses the objective function in Equation 7 no more converges to τ_r^* in probability.*

While this issue likely does not exist for platforms that have access to the true propensity scores, researchers who use data from platforms but do not access to the true propensity scores need to overcome this challenge. The problem is exacerbated as the logarithmic loss function often used to estimate the propensity scores never estimates zero or one as the predicted outcome. This is the motivation for sample trimming based on the estimated propensity scores, where the researcher drops the observations where the estimated propensity score is very close to zero or one [Crump et al., 2009, Ma and Wang, 2020, D'Amour et al., 2021].

3.2.3 Model-free Approaches to Estimate ATE

We now discuss model-free approaches to estimate the ATE. The foundation for these approaches is the idea of importance sampling proposed by Horvitz and Thompson [1952] in their seminal paper. The idea is to weight each observation by their inverse propensity score, which gives us the following estimator for the ATE:

$$\hat{\tau}_{\text{IPS}} = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i \left(\frac{W_i}{\pi(X_i)} - \frac{1-W_i}{1-\pi(X_i)}\right)\right), \tag{12}$$

where the first term $W_i/\pi(X_i)$ weights the observations that received the treatment by the inverse probability of that assignment, and the second term $(1-W_i)/(1-\pi(X_i))$ weights the observations that did not receive the treatment. This estimator estimates the average treatment effect by subtracting an estimate of what would happened if everyone had received the control from an estimate of what would have happened if everyone had received the treatment. It is a model-free approach because we do not need any model of the outcome to estimate our causal estimand.

In the absence of full overlap, a drawback of this approach becomes immediately apparent. For observations with deterministic assignment, the denominator in one of the terms is zero, which makes the overall estimator undefined. The conventional solution is to use sample trimming whereby we drop observations with a deterministic assignment. As a result, this

approach only relies on the α_r fraction of observations with the probabilistic assignment. We can show the following proposition:

Proposition 2. *Suppose that there is a digital platform that has access to data* $\mathcal{D} = \{Y_i, W_i, X_i, \pi(X_i)\}$*, where* $\pi(X_i)$ *is known, but takes values zero and one for parts of the population. The ATE estimator based on Equation* (12) *that drops observations with a deterministic assignment converges in probability to* τ_r^**, that is:*

$$\hat{\tau}_{IPS} \xrightarrow{p} \tau_r^* \tag{13}$$

Similar to Proposition 1, Proposition 2 guarantees that the Inverse Propensity Scoring (IPS) estimator recovers the treatment effect for the probabilistic region. As such, Corollary 1 holds for this proposition too, indicating that the bias is a function of two unidentifiable elements τ_1^* and τ_0^*. The equivalent of Corollary 2 here is that when propensity scores are not known, trimming can become a non-trivial task because propensity scores very close to zero or one can result in very large inverse weights, thereby heavily influencing the performance of the estimator. This is why a body of work focuses on data-driven and robust rules for finding the trimming threshold [Crump et al., 2009, Ma and Wang, 2020].

3.3 Discussion

So far, we showed that the lack of overlap can be detrimental in observational studies that aim to estimate causal parameters. We first emphasized that the lack of overlap is prevalent in digital platforms that use algorithms for decision-making. We considered a generic case of an algorithm that creates three regions based on its outcome: (1) $0 < \pi(X_i) < 1$, (2) $\pi(X_i) = 1$, and (3) $\pi(X_i) = 0$. We showed that even if we have the exact outputs of the function $\pi(\cdot)$ that automatically satisfies unconfoundedness, the lack of overlap can bias the estimates of the state-of-the-art model-based and model-free approaches. We further highlighted that the problem can be exacerbated if the propensity scores are now known and the researcher has to estimate them from the data. Thus, in observation studies on digital platforms, there is a possibility for substantial bias in the main estimates.

An important discussion to be had is whether this is just a theoretical possibility that is not practically important. In other words, do we expect the bias term $|\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)|$ to be large in real settings? Part of the rationale for the trimming approaches widely used in the literature is that τ_0 and τ_1 are not different from τ_r. To the extent that $\pi(x)$ is a function of $\tau^*(x)$, we expect τ_0 and τ_1 to be different from τ_r. The problem is that in many cases, the objective function in the algorithm used by the digital platform incorporates the CATE that

is of interest to the researcher. For example, suppose that there is a ride-hailing app that wants to offer promotions to users with the objective to maximize the demand. As such, the platform offers promotion to users for whom the effect of promotion on demand is higher, such that users who have a significant and positive CATE of promotion on demand certainly receive the treatment and users who have a significant and negative CATE of promotion on demand never receive the treatment. Now, if a researcher wants to use this data to study the effect of promotion on demand, we expect that $\tau_0 \leq \tau_r \leq \tau_1$, hence a large bias in any observational approach to estimate the ATE.

It is worth noting that the default is that $\tau_0 \neq \tau_r \neq \tau_1$, unless we have a strong reason to believe that $\tau_0 = \tau_r = \tau_1$. In particular, we expect the algorithmic decision-making in digital platforms to at least implicitly use the information in $\tau^*(\cdot)$. In the next section, we focus on the case of advertising auctions, which is one of the most well-known cases of algorithmic decision-making in digital platforms.

4 Case of Advertising Auctions

Digital advertising is one of the most successful cases of algorithmic decision-making. Advertising revenues constitute the dominant revenue share of two of the largest companies globally: Google and Facebook. From its early days, advertising platforms used auctions to sell their advertising spaces. As such, the assignment of a user to an ad is determined by an auction. In this section, we first present a general framework for the auction environment in digital advertising and the allocation (assignment) rule in §4.1. We then discuss the challenges in front of the observational studies that intend to quantify ad effectiveness and present theoretical results on such studies in §4.2.

4.1 Assignment Mechanism in Advertising Auction

Ad impressions are sold through auctions. For a single ad impression, platforms generally use first- or second-price auctions. For more complicated settings such as search advertising where there are multiple items, platforms extend these auctions to Generalized Second Price (GSP) or Vickrey–Clarke–Groves (VCG) auctions. In all these auctions, advertisers participate in an auction for an impression and submit a bid, which is a signal of their willingness-to-pay for the item being sold, whether it is an impression or a click. These auctions often combine advertisers' bids with the data available at the platform to determine which ad to serve at each impression. Our goal in this section is to better understand the specifics of this environment and translate it into the potential outcomes framework used in §3.

We present a very general framework to characterize the auction environment. Let b_{ia}

denote advertiser a's bid for impression i. As common in most advertising auctions, the auction applies a scoring function $f(\cdot)$ to these bids given the information available for impression i. For example, the scoring output for advertiser a in impression i is $f(b_{ia}; X_i)$. A well-known example of scoring is the quality score in pay-per-click environments where advertiser's bid only reflects their willingness-to-pay for click and there may be advertisers with very high bid per click but low expected revenue at the impression level because they have a very low click-through rate (CTR). As such, platforms adjust for the probability of click in their scoring function such that $f(b_{ia}; X_i) = b_{ia} \times CTR(X_i, a)$, where $CTR(X_i, a)$ is the probability of click on ad a when shown in impression i with the vector of characteristics X_i. In our study, we do not impose any specific structure on the scoring function. In cases where the auctioneer solely relies on the advertiser's bid, f would be the identity function.

The auctioneer then computes the scoring output for all the ads and runs an auction with a specific allocation rule that returns a probability for each ad participating in the auction. Let $q(a; f(\mathbf{b_i}))$ denote the probability that ad a wins the auction given the vector of scoring outputs $f(\mathbf{b_i})$, where $\sum_a q(a; f(\mathbf{b_i})) = 1$. We now present the following definition to distinguish between deterministic and probabilistic auctions:

Definition 3. *An auction is deterministic when there is only one bidder who wins the item conditional on the item being sold. In other words, all bidders except one have $q(a; f(\mathbf{b_i})) = 0$. An auction is probabilistic, if there some level of randomness in the allocation function.*

Most common auctions used by advertising platforms are deterministic. For example, in the second-price auction, the impression is sold to the advertiser with the highest scoring output.

With these preliminaries about the auction, we now characterize our problem within the potential outcomes framework. Suppose that we have N observation units and one focal ad. For each observation unit i, there is an assignment status, which denotes whether the unit is assigned to the treatment (focal ad) or not, that is, $W_i = 1$ when the observation is assigned to our focal ad and $W_i = 0$ when the observation is not assigned to the focal ad. Consistent with our problem formulation in §3.1, we use $Y_i(w)$ to denote the potential outcomes (e.g., sales, visits), X_i to denote the vector of characteristics (e.g., demographics, browsing history), and $\pi(X_i)$ to denote the propensity score.

The part that is specific to the context of digital advertising is the propensity function, which is determined by the auction and advertisers' equilibrium bidding behavior. Advertisers submit bids based on their private valuations. Fundamentally, an advertiser's private valuation is the value they receive from being shown in an observation unit (impression) as compared to

not being shown in that observation unit. What is unique in the context of digital advertising is that this private valuation is the same as the CATE for that observation. That is, the value of being shown in an observation unit to an advertiser is the amount that this exposure casually shifts the advertiser's outcome of interest $Y_i(1) - Y_i(0)$. [Waisman et al., 2019] use this observation to theoretically derive the equilibrium bidding in the second-price auction: if CATE is greater than zero, the advertiser bids CATE. Given the goal of our study, we want to know how CATE relates to the propensity function. We present the following proposition for a large class of auctions:

Lemma 2. *In any standard auction where the item is allocated to the bidder with the highest scoring output, $\pi(X_i)$ is increasing in $\tau^*(X_i)$.*

Proof. *In any standard auction, we know that a higher bid results in a higher probability of allocation. We also know that the equilibrium bid is increasing in the private valuation. Therefore, the allocation probability is increasing in the private valuation, which implies that the propensity function $\pi(X_i)$ is increasing in the private valuation $\tau^*(X_i)$.*

The result of Lemma 2 emphasizes a strong link between the assignment function and CATE. In light of our discussion in §3.3, this imposes important challenges in estimating the average treatment effect (ATE), particularly when there are regions where the assignment is deterministic. In the next section, we examine how prevalent the issue of overlap violation is in digital advertising auctions.

4.2 Estimation of Advertising Effectiveness

As discussed earlier, the assignment to the focal ad is determined by the auction run by the advertising platform. In this section, we investigate the implications of the underlying auction for estimating advertising effectiveness.

4.2.1 Single Treatment Opportunity Per Unit

We start with the case where we have N observation units. For each observation unit, there is only one opportunity to receive the treatment variable (focal ad). As such, the propensity function π is equivalent to the allocation function q used by the auction. Like before, there are potential outcomes $Y_i(0)$ and $Y_i(1)$ and a vector of characteristics X_i. Our goal is to estimate the average treatment effect using observational data. The following proposition presents an important theoretical result in this case:

Proposition 3. *In a standard advertising auction with a single treatment opportunity per unit where the observation unit (impression) is allocated to the ad with the highest scoring*

output, the treatment assignment is fully deterministic. That is, we either have $\pi(X_i) = 0$ *or* $\pi(X_i) = 1$.

Proof. *Suppose that there are* k *bidders competing with the focal ad with a bid profile of* $(b_{i1}, b_{i2}, \ldots, b_{ik})$ *and the corresponding scoring output of* $(f(b_{i1}; X_i), f(b_{i2}, X_i), \ldots, f(b_{ik}; X_i))$. *Let* c *denote the maximum of the scoring outputs, i.e.,* $c = \max_j f(b_{ij}; X_i)$. *The allocation function of a standard advertising auction for ad* a *is equivalent to* $q(a; f(\mathbf{b_i})) = \Pr(f(b_{ia}; X_i) > c)$. *Thus, the propensity score is equal to zero for cases where* $f(b_{ia}; X_i) < c$ *and equal to one for cases where* $f(b_{ia}; X_i) > c$.

Proposition 3 shows that in standard auctions that are widely adopted in the online advertising industry, the assignment mechanism only generates two regions: (1) deterministic assignment, and (2) deterministic no-assignment. In light of Lemma 1, we know that the CATE for the deterministic regions is unidentified. Thus, we can write the following corollary:

Corollary 3. *Suppose that there is an advertising platform that runs a standard auction to allocate ads and has access to data* $\mathcal{D} = \{Y_i, W_i, X_i, \pi(X_i)\}$, *where* $\pi(X_i)$ *is known. In a setting where there is a single opportunity for each observation unit to receive the treatment, the average treatment effect (ATE) is unidentified.*

This corollary presents a statistical impossibility result for estimating advertising effectiveness using observational studies. In the case where propensity scores are all known, the researcher immediately realizes this point since no statistical package is able to estimate advertising effectiveness given the perfect collinearity between the assignments and propensity scores. However, in many cases, researchers do not have access to the actual propensity scores and need to estimate these values. As discussed earlier, the common procedure is to use a predictive algorithm to obtain estimates of $\hat{\pi}(X_i)$. The better the algorithm is, the closer it predicts the assignment probability to its true deterministic value. However, it rarely predicts the value to be exactly zero or one. As a result, the observational method to estimate the ATE uses some probabilistic variation in the treatment assignment to estimate the treatment effect. Further, in light of Lemma 2, we know that the focal ad is likely shown in observations with higher CATE. Hence, we can write the following corollary on the estimated advertising effectivenes under observational methods:

Corollary 4. *If the propensity score function* $\pi(\cdot)$ *is not known, the estimated advertising effectiveness under observational methods exhibit large bias.*

4.2.2 Multiple Treatment Opportunities Per Unit

The case in §4.2.1 only focused on a single impression opportunity for every user. However, in most observational studies, the observation unit is defined as a user [Lewis et al., 2011, Hoban and Arora, 2018, Gordon et al., 2019, 2022]. In these cases, the observation unit (user) has multiple opportunities to be assigned to the treatment over the course of his availability. That is, for any user i, there are t_i impressions for which the platform runs an auction to determine which ad to show. We can define the treatment as a binary variable indicating whether each user i has seen the focal ad at least in one of the t_i impression(s). Let $W_{i,j}$ denote whether the focal ad is shown in the j^{th} impression for user i. We can the define W_i as follows:

$$W_i = \mathbb{1}(\sum_{j=1}^{t_i} W_{i,j} \geq 1). \tag{14}$$

Now, the question is whether this new environment allows for the probabilistic assignment. Intuitively, the auctions running for user i change across t_i impressions as some of the competitors may exit the market and the contextual factors may change. This intuition motivates the use of observational methods that can capture unconfoundedness like the ones discussed in §3.2.2 and 3.2.3. The following proposition examines whether such aggregation overcome the overlap violation:

Proposition 4. *In an observational setting where each unit has multiple opportunities (impressions) to be assigned to the treatment at least once through a standard auction mechanism where each impression is awarded to the ad with the highest scoring output, the treatment assignment W_i is deterministic.*

Proof. *We can define the propensity score in this new environment as follows:*

$$\Pr(W_i = 1 \mid X_i) = 1 - \prod_{j=1}^{t_i}(1 - \Pr(W_{i,j} = 1 \mid X_{i,j})), \tag{15}$$

where $\prod_{j=1}^{t_i}(1 - \Pr(W_{i,j} = 1 \mid X_{i,j}))$ is the probability that the focal ad loses the auction for all t_i impressions for user i. For each impression j, $1 - \Pr(W_{i,j} = 1 \mid X_{i,j})$ is either zero or one in light of Proposition 3. Hence, $\prod_{j=1}^{t_i}(1 - \Pr(W_{i,j} = 1 \mid X_{i,j}))$ is binary, which implies that the propensity score $\Pr(W_i = 1 \mid X_i)$ is binary. Thus, the assignment is deterministic and the proof is complete.

Proposition 4 is important because it shows that the case for multiple treatment opportunities is the same as the case for single treatment opportunity in having the overlap violation.

Therefore, all the corollaries under the single treatment opportunity scenario follows under the multiple treatment opportunities scenarios.

5 Solutions to Overlap Violation

So far, we presented the challenge imposed by the lack of overlap in observational studies involving digital platforms. The problem stems from the deterministic output of algorithms that are used for decision-making in these platforms. Overall, our theoretical analysis showed how biased and inconsistent observational methods can be when the overlap assumption is violated, thereby suggesting researchers to be cautious when applying observational methods to digital platforms' data, as such data often violate the overlap assumption.

In this section, we aim to discuss the solutions – what can a digital platform do to overcome this challenge? We focus on this question from two different perspectives. First, we consider observational methods to overcome the lack of overlap. Given the fact that the treatment effect estimands are unidentifiable under the current set of assumptions, we can only overcome this issue by imposing further assumptions. In §5.1, we explicitly state these assumptions and present a solution based on matrix completion that works under some form of low-rank assumptions. The second perspective that we take is more design-based, where we discuss the potential solutions for the platforms to avoid running into overlap problems. That is, how can a platform design its algorithms used for decision-making to achieve overlap without hurting its broader objective, whether it is to maximize revenue, engagement, etc. We present these design-based solutions in §5.2.

5.1 Observational Solution Based on Matrix Completion

As discussed earlier, the fundamental problem with the deterministic assignment is one of identification. In light of Lemma 1, we know that with the current set of assumptions, the parameters τ_1^* and τ_0^* can not be identified because there is no variation in the treatment variable beyond the propensity score. In general, we can write the conditional average treatment effect as follows:

$$\tau^*(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \mu_1(X_i) - \mu_0(X_i), \tag{16}$$

where $\mu_w(x)$ is the population function for potential outcomes conditional on x when assigned to treatment w. Form a learning standpoint, if one of the two treatment states could have never been generated in the data, no model can estimate the corresponding μ function. For example, if a unit with covariates X_i could have never received the treatment, we have no

observation in our data to estimate $\mu_1(X_i)$. As such, the problem caused by the lack of overlap is one of missing data.

To overcome this missing data challenge, we turn to one of the well-known solutions to this problem: matrix completion. We argue that if we have enough treatments w in our treatment space that are implemented for our population through some varying propensity function π_w, we can exploit the similarity patterns between the treatments to impute our estimands of interest in unidentifiable regions of data if the underlying model is low-rank. The following example helps illustrate the intuition. Suppose that treatment w has deterministic assignment and no-assignment regions. For example, this treatment has a zero propensity to be shown in unit i of our data, so CATE of w is unidentifiable for this unit. We further suppose that there is another treatment w' that has a probabilistic assignment for unit i, so we can estimate the CATE of w' for unit i. Now, if the two treatments exhibit very similar patterns for the units where they can both feasibly estimate the CATE, we can use the CATE of w' for unit i to impute the CATE of w for unit i.

The simple example above is just to illustrate what kind of variation we use in our method. However, such similarities may be rare in reality, especially if treatments are different from each.[4] Therefore, for this method to work, we need a more systematic way to capture the similarities in the space of treatments. That is why we use a matrix completion approach that has been widely used for collaborative filtering.

In this section, we start by describing our model preliminaries in §5.1.1. We then present our matrix completion algorithm in §5.1.2. Finally, in §5.1.3, we consider the case where the propensity scores are unknown and present a simple procedure based on our algorithm that tests whether the lack of overlap in our data causes large bias in observational studies.

5.1.1 Model Preliminaries

We use the same potential outcomes framework terminology for our modeling framework, but we depart from considering a single treatment. Instead, we focus on a J-dimensional treatment space $\mathcal{W}$ such that $\mathcal{W} = \{w_1, w_2, \ldots, w_J\}$. Our observation units may have received

[4]It is more plausible to see such similarities between two treatments if they are fundamentally similar, such as two versions of the same ad.

some of these treatments or not. We can define Ω as the treatment matrix as follows:

$$\Omega = \begin{bmatrix} \mathbb{1}(W_1 = w_1) & \mathbb{1}(W_1 = w_2) & \dots & \mathbb{1}(W_1 = w_J) \\ \mathbb{1}(W_2 = w_1) & \mathbb{1}(W_2 = w_2) & \dots & \mathbb{1}(W_2 = w_J) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{1}(W_N = w_1) & \mathbb{1}(W_N = w_2) & \dots & \mathbb{1}(W_N = w_J) \end{bmatrix}, \tag{17}$$

where the element in the i^{th} row and j^{th} column is a binary variable indicating whether the observation unit i has received intervention w_j or not. For brevity, we refer to the treatment w_j as the j^{th} treatment or treatment j henceforth. An important point to clarify here is that a unit can be assigned to multiple treatments. Such a context can arise in a variety of applications such as cases where the treatments are assigned at different times to the same population, or cases where units had multiple opportunities to receive the treatment. For each treatment, there are two potential outcomes $Y_i(1^{(j)})$ and $Y_i(0^{(j)})$, where the superscript (j) refers to the treatment and control condition when studying the effect of treatment j. We denote CATE of treatment j for unit i by $\tau_j^*(X_i)$ and define it as follows:

$$\tau_j^*(X_i) = \mathbb{E}[Y_i(1^{(j)}) - Y_i(0^{(j)}) \mid X_i] \tag{18}$$

This allows us to define the conditional average treatment effect (CATE) matrix $\mathcal{T}$ as follows:

$$\mathcal{T} = \begin{bmatrix} \tau_1^*(X_1) & \tau_2^*(X_1) & \dots & \tau_J^*(X_1) \\ \tau_1^*(X_2) & \tau_2^*(X_2) & \dots & \tau_J^*(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \tau_1^*(X_N) & \tau_2^*(X_N) & \dots & \tau_J^*(X_N) \end{bmatrix}, \tag{19}$$

where each element $[i, j]$ in matrix $\mathcal{T}$ is the CATE of the treatment j for unit i. The goal of our algorithm is to estimate CATE for all the elements in the matrix in spite of the overlap violation. To do so, we first need to know which elements we cannot estimate with the conventional methods to estimate CATE. Therefore, we define the propensity matrix as follows:

$$\Pi = \begin{bmatrix} \pi_1(X_1) & \pi_2(X_1) & \dots & \pi_J(X_1) \\ \pi_1(X_2) & \pi_2(X_2) & \dots & \pi_J(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1(X_N) & \pi_2(X_N) & \dots & \pi_J(X_N) \end{bmatrix}, \tag{20}$$

where each element $\Pi_{i,j}$ denotes the propensity score for the treatment j for unit i, i.e., $\Pi_{i,j} = \pi_j(X_i) = \Pr(W_i = w_j \mid X_i)$. As such, the deterministic regions for each treatment is defined as rows where the propensity score is either zero or one. We know that the conditional average treatment effect is unidentified for these units. Thus, we define a feasibility matrix F that takes value one only when the assignment is probabilistic, that is, the propensity score is between zero and one. As such, we can write each elements of this matrix as follows:

$$F_{i,j} = \mathbb{1}(0 < \pi_j(X_i) < 1). \tag{21}$$

The feasibility matrix F determines the scope of our CATE estimation. That is, if for treatment j in unit i, we have $F_{i,j} = 0$, Lemma 1 implies that we cannot identify $\tau_j^*(X_i)$. As such, F determines the question marks in our matrix completion task.

5.1.2 Algorithm for Imputing Overlap-violating Regions

We now present our algorithm to impute the values for the non-feasible regions of the matrix. Our algorithm has two steps presented as follows:

- *Step 1:* Upon feasibility, estimate CATE of treatment j in unit i. If $F_{i,j} = 1$, we can use conventional methods to estimate $\tau_j^*(X_i)$ since the assignment $\pi_j(X_i)$ is probabilistic and we have unconfoundedness. Conversely, if $F_{i,j} = 0$, we cannot estimate $\tau_j^*(X_i)$. Thus, we can construct matrix $\hat{\mathcal{T}}_1$ where the elements are either the estimated CATE denoted by $\hat{\tau}_j(X_i)$ or question marks in cases where CATE is not identified.
- *Step 2:* We now an incomplete matrix $\hat{\mathcal{T}}_1$, where the incomplete elements are the overlap-violating regions. If the underlying model is low-ranked, we can use conventional matrix decomposition techniques to impute the question marks. This procedure exploits the similarities in the joint space of units and treatments. We denote this new completed matrix by $\hat{\mathcal{T}}_2$ where the subscript denotes the second step in our algorithm.

The output of this algorithm is a complete matrix $\hat{\mathcal{T}}_2$ where all the elements have some values. This complete matrix can then be used to estimate the ATE from the data. While the average value of each column j in $\hat{\mathcal{T}}_1$ estimates τ_r for treatment j, the average value of column j in $\hat{\mathcal{T}}_2$ estimates τ^* if the underlying matrix $\mathcal{T}$ is low-rank. In the next section, we use this intuition to develop a test for the existence of bias.

5.1.3 Statistical Tests for the Existence of Bias

An important use of our matrix completion algorithm is to test whether the lack of overlap can cause bias in the estimates of average treatment effect. From Corollary 1, recall that the

bias term from ignoring the overlap is $|\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)|$. Hence, if $\tau_r = \tau_0 = \tau_1$, there would be no bias in our estimates if propensity scores are known. In fact, the motivation behind sample trimming in the prior literature is the fact that the τ_0 and τ_1 are no different from τ_r. In this section, we present statistical tests based on our algorithm to examine whether the lack of overlap can cause bias in estimating the ATE in two separate scenarios based on when propensity scores are (1) known, and (2) unknown.

We begin with the case where propensity scores are known. In this case, we can run our algorithm to obtain the completed matrix $\hat{\mathcal{T}}_2$. We also have the propensity matrix Π, which helps us distinguish between the three assignment possibilities for treatment j: (1) probabilistic assignment ($0 < \pi_j(x) < 1$), (2) deterministic assignment ($\pi_j(x) = 1$), and (3) deterministic no-assignment ($\pi_j(x) = 0$). Hence, for each column j in matrix $\hat{\mathcal{T}}_2$, we have three corresponding groups of elements. This allows us to statistically test if $\tau_r = \tau_0$ and $\tau_r = \tau_1$. This allows us to test the hypothesis underlying conventional trimming techniques to address the overlap issue. We can further test if the bias term is significantly different from zero. The hypothesis test in that case would be $\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r) = 0$. If we fail to reject any of these tests, it means that the lack of overlap can fundamentally bias the estimates of average treatment effects.

We now turn to a more complicated case where propensity scores are unknown and need to be estimated from the data. In this case, we first need to estimate the propensity matrix. Once we have $\hat{\Pi}$, we can perform our matrix completion algorithm to obtain $\hat{\mathcal{T}}_2$. As discussed earlier, when we estimate propensity scores, it is challenging to identify the three assignment regions in the data. As such, we use different classification rules based on a parameter η such that $0 < \eta < 0.5$ to obtain a bias curve with the following algorithmic procedure:

- Based on the value η, we create three regions such that (1) the probabilistic assignment is the set of elements in column j of matrix where $\eta < \hat{\pi}_j(x) < 1 - \eta$, (2) the deterministic assignment region is the set of elements in column j of matrix where $\hat{\pi}_j(x) \geq 1 - \eta$, and (3) the deterministic no-assignment region is the set of elements in column j of matrix where $\hat{\pi}_j(x) \leq \eta$.
- Using the three assignment regions, we compute the bias term $\hat{\alpha_0}(\hat{\tau}_0 - \hat{\tau}_r) + \hat{\alpha}_1(\hat{\tau}_1 - \hat{\tau}_r)$. We can compute a confidence interval using a conventional t-statistics or bootstrap.
- We use all the values of η and the corresponding bias term with confidence interval to obtain the bias curve.

The resulting bias curve is a function of the parameter η and allows us to track test whether the bias term is zero given any value of η. Not only does this approach allow us to statistically

test whether the bias term is zero at a given threshold η, it also allows us to see the overall pattern in our propensity scores and how it interacts with the bias term. Presenting the bias curve can enhance the transparency in an observational study.

5.2 Design-based Solutions

As discussed earlier, the fundamental problem that results in the violation of the overlap assumption is the deterministic nature of algorithms used for decision-making at platforms. That is, by design, the algorithms used do not generate randomization in the data where the assignment to treatment is probabilistic. An important consideration here is the platform's objective: the reason why platforms do not induce large-scale randomization in assigning interventions is that randomization hurts the economic goal of the platform. There is often an optimal intervention for any observation unit, and platforms want to implement that intervention to achieve better outcomes. To that end, we need to consider solutions that embed some form of randomization in platform's decision-making that helps them achieve overlap in their data without hurting their economic objectives. We present the following two solutions and discuss how they help platforms with their inference goals:

1. *Small-scale experimentation:* Many platforms leave a portion of their total traffic to implement full randomization and use the resulting data for research. Suppose that there are N observations in the data. An γ fraction of these observations is used as an experiment where the platform randomizes the interventions. For example, if there are two experimental conditions, the platform can implement a randomized experiment on this small portion of their data and implement the potentially optimal intervention for the rest of observations. As such, the propensity score for an impression i is either zero or one in $(1-\gamma)N$ observations and is 0.5 in γN observations. This enables the platform to γN portion of their data to draw generalizable inference and apply it to other parts of data. However, it is important to notice that only γ fraction of observations satisfy the overlap assumption, and all the unidentifiability results for deterministic assignment apply to the remaining $1-\gamma$ fraction of observations.

2. *ϵ-greedy approach:* Another approach that is advocated in bandits and reinforcement learning literature is to adopt some for of ϵ-greedy strategy that implements the policy suggested by the algorithm with a $1-\epsilon$ probability, and other sub-optimal policies with a collective probability of ϵ. The platform can vary ϵ depending on their economic goals. As an example, suppose there is an auction that decides which ad to show to the user from a set of $K+1$ ads. If ad a is the winner of the auction, it will be

shown with probability $1-\epsilon$, and any other ad with probability ϵ/K. It is worth noting that this approach satisfies the overlap assumption for all the observations as $1 < \Pr(W_i = w \mid X_i) < 1$ for any w. As such, the researcher can use all the observation for research purposes. However, the sensitivity of analysis increases as the propensity weights are either too small or large, resulting in a large variance in the parameter estimates, which requires massive amounts of data to overcome this statistical power issue.

The choice of which design-based solution to use depends on the platform's scale, algorithm in place, and its broad objective.

6 Conclusion

Digital platforms use algorithmic decision-making to deliver interventions to their users at a very large scale. An important research question to both practitioners and academic researchers is the causal effect of such interventions. The gold standard answer to this question is to run randomized experiments. However, these experiments are often too costly, thereby giving rise to observational methods that use platforms' existing data without incurring experimentation cost. We examine this problem using the well-established potential outcomes framework [Holland, 1986]. Observational studies generally require an important assumption called strong ignorability of the treatment assignment which comprises two parts: unconfoundedness of the treatment assignment and overlap. While much of the prior applied an methodological literature focused on the former, the latter received considerably less attention. We show that in digital platforms, this is in fact the overlap assumption that is not satisfied because the output of algorithmic recommendations is often deterministic. We theoretically show that the lack of overlap can be detrimental to the validity of an observational study. We quantify the bias term and argue that in most digital platforms, we expect the bias caused by the lack of overlap to be large. In particular, we focus on advertising auction and show that average treatment effect is unidentified in the observational studies on digital platforms. Lastly, we formulate the identification problem caused by the lack of overlap as a missing data problem and turn to a solution that is often considered for such challenges. We show that if the platform has data on many treatments for the same units of population and the space of treatment effects is low-rank, we can recover the true average treatment effect.

There are several contributions that our paper makes to the literature. First, we present a comprehensive study of overlap violation in observational studies. We show how the lack of

overlap can bias the estimates of average treatment effects from observational studies that ignore this assumption. Second, our paper provides important insights to practitioners. We show that the data from digital platforms that use algorithms to make decisions suffer from an often ignored part of the ignorability assumption: overlap assumption. We show that this problem is generally prevalent in digital platforms. In particular, we consider the case of advertising platforms that run auctions to determine what ad to show to users, and show an important impossibility result: we find that with the platform's data, the average treatment effect is unidentified. Finally, we provide a solution to this problem that can correct the bias caused by the lack of overlap if the platform has access to the data for numerous interventions and the underlying space is low-ranked.

References

E. Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98, 2018.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

T. Blake, C. Nosko, and S. Tadelis. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1):155–174, 2015.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018a.

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018b.

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

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

A. Goli, A. Lambrecht, and H. Yoganarasimhan. A bias correction approach for interference in ranking experiments. *Available at SSRN 4021266*, 2022a.

A. Goli, D. G. Reiley, and H. Zhang. Personalized versioning: Product strategies constructed from experiments on pandora. Working Paper, 2022b.

B. R. Gordon, F. Zettelmeyer, N. Bhargava, and D. Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225, 2019.

B. R. Gordon, R. Moakler, and F. Zettelmeyer. Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *arXiv preprint arXiv:2201.07055*, 2022.

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

P. Hoban and N. Arora. Measuring display advertising response using observational data: The impact of selection biases. *Available at SSRN 3264871*, 2018.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

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

A. Jesson, S. Mindermann, U. Shalit, and Y. Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33: 11637–11649, 2020.

R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web*, pages 157–166, 2011.

X. Ma and J. Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

J. Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

O. Rafieian. Optimizing user engagement through adaptive ad sequencing. Technical report, Working paper, 2022.

O. Rafieian and H. Yoganarasimhan. Targeting and privacy in mobile advertising. *Marketing Science*, 2021.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

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

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

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

B. T. Shapiro, G. J. Hitsch, and A. E. Tuchman. Tv advertising effectiveness and profitability:

Generalizable results from 288 brands. *Econometrica*, 89(4):1855–1879, 2021.

C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

D. Simester, A. Timoshenko, and S. I. Zoumpoulis. Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Science*, 66(8):3412–3424, 2020a.

D. Simester, A. Timoshenko, and S. I. Zoumpoulis. Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science*, 66 (6):2495–2522, 2020b.

S. Tunuguntla. Display ad measurement using observational data: A reinforcement learning approach. *working paper*.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 0(0):1–15, 2018. doi: 10.1080/01621459.2017.1319839.

C. Waisman, H. S. Nair, C. Carrion, and N. Xu. Online causal inference for advertising in real-time bidding auctions. *arXiv preprint arXiv:1908.08600*, 2019.

H. Yoganarasimhan, E. Barzegary, and A. Pani. Design and evaluation of optimal free trials. *Management Science*, 2022.