A Matrix Completion Solution to the Problem of Ignoring the Ignorability Assumption

Omid Rafieian*
Cornell Tech and Cornell University
Click Here for Latest Version

March 1, 2023

*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 the 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. To address this issue, 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 space of treatment effects 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. Finally, we validate our theoretical results using synthetic data and discuss the implications.

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 push notifications on mobile phones, content recommendation on streaming platforms, etc. At the core of this large-scale delivery of interventions are two elements: data collection and algorithmic decision-making. Digital platforms collect massive amounts 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 platforms 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 identify and estimate the causal effect of an intervention. However, running fully randomized experiments is not always in the interest of the platform, because experiments can come at the expense of assigning 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, 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 output of the 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. For example, the promotion offered by a ride-sharing app is the deterministic output of an algorithm. 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 because of the deterministic assignment employed by the algorithms. To that end, we seek to answer the following sets 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 effects (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 an important insight: the magnitude of bias in the ATE estimate can be arbitrarily large if the overlap assumption is violated. We then carry out a series of analysis with simulated data to verify our theoretical findings. In particular, we consider three different scenarios: (1) known propensity scores, (2) unknown propensity scores under the full observability of covariates that influence propensity scores, and (3) unknown propensity scores under the partial observability of covariates that influence propensity scores. When propensity scores are known, we find that model-based (e.g., Double Machine Learning) and model-free (Inverse Propensity Scoring) approaches recover the conditional average treatment effect for the region with probabilistic assignment, but fail to recover the true average treatment effect for the population.

Next, we examine what will happen if the propensity scores are not known. We demonstrate that the ATE estimates will no longer converge to the CATE for the probabilistic region, even if the entire set of covariates that influence propensity scores are observed by the researcher. In particular, we find that even trimming approaches do not work properly in these settings. As expected, we find that this issue is exacerbated when the covariates that influence propensity scores are only partially observed. Together, our analysis of settings

where propensity scores need to be estimated highlight that even a data-rich environment where every factor in the algorithmic decision-making is observed can generate estimates with no proper interpretation. Thus, a data-rich environment is vastly insufficient for studies of platforms that engage in algorithmic decision-making.

We then focus on the prevalence of this problem and ask the following question: to what extent will this issue arise in practical contexts? 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 scores that are related to users' conditional average treatment effects (Shalit et al., 2017; Wager and Athey, 2018). For example, for each user, a ride-sharing app provides a promotion that is most profitable based on their algorithm. Similarly, a mobile app uses an algorithm to identify the notification that yields a better engagement outcome than all other actions. Therefore, the conditional average treatment effect of the probabilistic region is not the same as the average treatment effect for the entire population because the deterministic regions are likely selected from the tails of the CATE distribution.

Once we establish the existence and prevalence of the lack of overlap in observational studies involving digital platforms and the challenges it pose, 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. Although we cannot fix this problem with a single study at hand, we can potentially use the information across studies to help with this missing data problem. In particular, if we have multiple studies with different treatments (e.g., price discount in one study, and push notification for a loyalty program in another) whose individualized effects come from a low-rank space, we can use matrix completion methods to impute the conditional average treatment effect for the overlap-violating regions. In particular, we treatment CATE estimates from the overlap-violating regions as question marks in a matrix and only estimate CATE for units whose assignment is probabilistic. We then exploit the variation among those entries in the matrix to complete the matrix for the deterministic regions. The intuition for this approach is as follows: for a user i whose CATE is unidentifiable because their assignment to the treatment in a study is deterministic, we can exploit the variation in how similar users responded to similar treatments when their assignment to those treatments were probabilistic. Once we complete the matrix for the parts that are formerly unidentified, we can correct for

the bias in the ATE estimates.

Our proposed method further allows us to offer a test for cases 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.

Lastly, we deliver a series of simulation studies to establish the performance of our proposed algorithm. We consider a wide range of deterministic assignment problems that may arise in real settings. Each case corresponds to a specific missingness pattern in the estimated CATE matrix due to identifiability issues. To that end, we consider three specific types of missingness patterns: (1) random, (2) CATE-dependent, and (3) user-dependent. When missingness is at random, we show that both Double ML (or other conventional ATE estimation approaches) and our proposed method are able to recover ATE across studies. However, our proposed method has lower error as it exploits the variation across studies, which makes it suitable for small-data environments. In settings with CATE-dependent missingness, we simulate cases where observations with higher or lower CATEs are more likely to have a deterministic assignment or no-assignment. We show that the ATE estimates under conventional approaches such as Double ML are largely biased. However, our proposed method can reliably recover the true ATE. Finally, we consider two forms of user-dependent missingness where some users are more likely to have deterministic assignments or no-assignments. We show that as long as the data for some users are not entirely missing, our proposed algorithm can recover the true ATE. We further establish the boundary conditions of our algorithm.

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. Notably, 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 assumption has received considerable attention 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 for treatment effect estimates. 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 effects 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 exploits the similarities between the treatments in the treatment space and overcomes the issue of overlap violation under certain assumptions.

Finally, our paper relates to the literature on matrix completion. Although the popularity of these models stems from the Netflix Prize for movie recommendation (Bennett et al., 2007), the application of matrix completion models is much broader to any setting where the underlying structure of matrix with missing data is low-rank (Mazumder et al., 2010). The relevance and success of matrix completion models motivated a large stream of theoretical work that establish the main theoretical guarantees of these models (Candès and Recht, 2009; Candès and Tao, 2010; Recht, 2011; Gross, 2011; Negahban and Wainwright, 2011). Recent work has focused on the intersection of matrix completion and causal inference and found useful applications (Kallus et al., 2018; Athey et al., 2021; Agarwal et al., 2021). Our work adds to this literature by formulating the unidentifiability of the overlap-violating parts of data as a missing data problem and apply matrix completion models to exploit cross-study variation and recover the true causal parameters.

3 Algorithmic Decision-making

3.1 Problem Definition

We first formally define 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 characterize 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 can be clicks or usage for push notifications. 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 observation i has received the control. Hence, for each 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 vector. 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

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

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.

- *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

[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.

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

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 the lack of overlap in observational studies on digital platforms. As such, we can formally define the problem as follows:

Definition 3. *Consider a digital platform that uses data $\mathcal{D} = \{Y_i, W_i, X_i, \pi(X_i)\}$. The main estimands the platform wants to estimate are the average treatment effect (ATE) for the entire population, as well as the conditional average treatment effects (CATE) for each value of the vector of covarites.*

Following the formal definition of our problem in Definition 3, our primary goals in this paper are 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 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 as shown in Figure 1:

- 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.

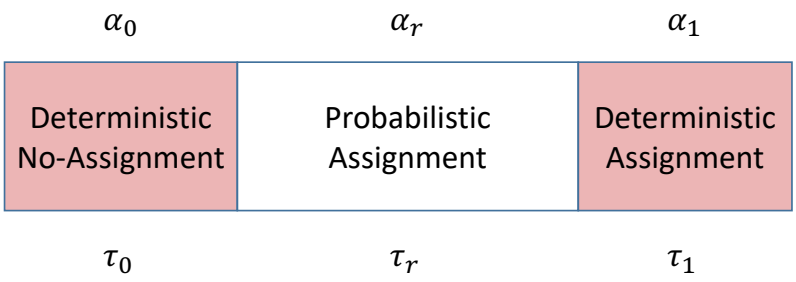


Figure 1: Different regions based on the type of assignment.

- 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 $\sum_{i \notin \mathcal{I}_r} \left(Y_i - m(X_i)\right)^2$ 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 simultaneously equal to zero. As such, for any constant c, we can find τ_0 and τ_1 such that $|\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)| = c$, which implies that we can have any magnitude of bias. □

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 overlap-violating 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 presumably large weights 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.*

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. An immediate fix for this problem on the platform's end is to log the propensity scores data. However, this practice is not quite trivial as it requires multiple teams to work together within the platform. Further, in many cases, researchers use these data sets without accessing the true propensity scores. A data-driven solution to this problem is to use sample trimming techniques 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). If the covariates needed to estimate propensity scores are all available, sample trimming can help recover τ_r. Yet, the estimated treatment effect can be far from the true ATE.

3.2.3 Model-free Approaches to Estimate ATE

In §3.2.2, we show that model-based approaches to estimate ATE fail to recover the true ATE. However, one could argue that the bias comes from outcome modeling. To address this issue, we discuss model-free approaches to estimate the ATE that directly use the realized outcomes without modeling them. 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 wherein 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:*

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

Proof. The proof is straightforward and directly follows from the fact that we can only use non-deterministic propensity scores. As a result, we only focus on the observations in the probabilistic region. Therefore, the proof directly follows Horvitz and Thompson (1952). □

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 Simulation Experiments

In this section, we conduct simulation experiments with the general case of algorithmic decision-making as presented in section 3. In these cases, the goal is to find the effect of treatment W_i on Y_i. However, the assignment to W_i is through an algorithm π that can be partially deterministic. That is, given the vector of covariates X_i, the assignment probability $\pi(X_i)$ can be zero or one in some observations. We consider two interesting cases of this general problem. First, in §3.3.1, we consider the case where the algorithm output is known to the researcher. We show how different approaches perform in these scenarios. Next, in §3.3.2, we focus on the case where the algorithmic output is not known, but can be estimated using a high-dimensional set of covariates.

3.3.1 Known Propensity Scores

We begin with the case where the platform has the following data: $\mathcal{D} = \{Y_i, W_i, X_i, \pi(X_i)\}$. This means that the platform has direct access to the propensity scores needed to orthogonalize the treatment and potential outcomes and satisfy the unconfoundedness assumption (Rosenbaum and Rubin, 1983). To generate data in our simulation experiments, we use CATE parameters $\{\tau_0, \tau_1, \tau_r\}$ and their corresponding proportions $\{\alpha_0, \alpha_1, \alpha_r\}$ as follows:

- *Step 1*: For each observation i, we take a draw to determine whether it belongs to the probabilistic assignment region ($0 < \pi(X_i) < 1$) with probability α_r, deterministic no-assignment region ($\pi(X_i) = 0$) with probability α_0, and deterministic assignment region ($\pi(X_i) = 1$) with probability α_1. If i belongs to the probabilistic region, we draw a random probability value from the Uniform distribution, $\pi(X_i) \sim \mathcal{U}(0, 1)$.
- *Step 2*: We use $\pi(X_i)$ values to simulate the treatment variable W_i.
- *Step 3*: We generate structural error terms $\epsilon_i \sim \mathcal{N}(0, \sigma)$.
- *Step 4*: We use appropriate CATE value from $\{\tau_0, \tau_1, \tau_r\}$ to calculate the outcome as follows:

$$Y_i = W_i\Big(\mathbb{1}\big(\pi(X_i) = 0\big)\tau_0 + \mathbb{1}\big(\pi(X_i) = 1\big)\tau_1 + \mathbb{1}\big(0 < \pi(X_i) < 1\big)\tau_r\Big) + \epsilon_i. \tag{14}$$

We now simulate data under different sets of parameters and estimate the Average Treatment Effect (ATE) using the following approaches:

1. Plain Mean Difference (MD): In this model, we simply estimate the the mean difference between treated and control groups. Equivalently, we can regress Y_i on W_i without controlling for $\pi(X_i)$. We denote this estimate with $\hat{\tau}_{\text{MD}}$.

2. Double Machine Learning (DML): In this model, we follow the Double ML procedure in Chernozhukov et al. (2018a), where we first use a model to fit Y_i and W_i separately using $\pi(X_i)$ and X_i, and then regress the residuals of Y_i on those of W_i.[4] The model we use here is a Random Forest with maximum depth of 5 aggregated over 500 trees. We denote the ATE estimate under this model by $\hat{\tau}^{\text{RF}}_{\text{DML}}$.

3. Inverse Propensity Score (IPS): In this approach, we directly use the actual propensity scores and estimate the ATE using Equation (12). It is worth noting that we drop

[4]In this example, we drop X_i because all the propensity scores are given (Rosenbaum and Rubin, 1983). The results will not change if we include X_i in our analysis.

N	$\{\tau_0, \tau_1, \tau_r\}$	$\{\alpha_0, \alpha_1, \alpha_r\}$	*True ATE*	*Estimated ATE*		
			(τ^*)	$(\hat{\tau}_{\text{MD}})$	$(\hat{\tau}^{\text{RF}}_{\text{DML}})$	$(\hat{\tau}_{\text{IPS}})$
10^5	$\{1, 8, 2\}$	$\{0.25, 0.25, 0.50\}$	3.25	5.0196	2.0122	1.9888
10^5	$\{1, 8, 2\}$	$\{0.50, 0.50, 0.00\}$	4.50	7.9715	NA	NA
10^5	$\{1, 6, 2\}$	$\{0.40, 0.10, 0.50\}$	2.00	3.1221	1.9997	1.9725
10^5	$\{-1, -8, 2\}$	$\{0.10, 0.10, 0.80\}$	0.70	0.0007	2.0039	2.0728
10^3	$\{1, 8, 2\}$	$\{0.25, 0.25, 0.50\}$	3.25	5.2453	2.5337	3.0368

Table 1: Estimates of Average Treatment Effects (ATE) using different estimators when propensity scores are known. Data are simulated using the corresponding parameters and $\sigma = 3$.

observations with deterministic assignment because the denominator would be zero in these cases. We denote this ATE estimate by $\hat{\tau}_{\text{IPS}}$.

We present the results of our simulation experiments in Table 1. Each row in our table presents one simulation experiment and the first two columns show the parameters needed to simulate data. A few noteworthy patterns emerge from Table 1. First, as expected, in all instances, ATE estimates from the plain regression are largely biased. Second, we notice that both DML and IPS estimators are generally able to recover the true CATE for the probabilistic region (τ_r), which is equal to 2 in all instances. This result confirms the theoretical results presented in Propositions 1 and 2. The exception is the second row where these estimators cannot identify any estimate because the proportion of the probabilistic region is zero. This confirms the theoretical result in Lemma 1.

Third, we examine the magnitude of bias in estimating the ATE. As indicated in Corollary 1, this magnitude is $|\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)|$. We confirm this theoretical results using the main parameters and estimates in Table 1. The only case in which there is no bias is the third row, where $|\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)| = 0$, or alternatively, where $\tau_r = \tau^*$. It is important to notice that the reason is that our DML and IPS estimates are able to recover τ_r, which happened to be the same in this instance as τ^*. Therefore, while these estimators can recover the true τ_r, their estimates of τ^* remain largely biased.

In the fourth row of Table 1, we consider a case where only a small fraction of observations have deterministic assignment. The results in this row show that even a small fraction of deterministic assignment in the data can largely bias the estimates of ATE, depending on how the CATE in these deterministic regions is different from that in the probabilistic region. Finally, in the fifth row of Table 1, we focus on a sample with fewer observations and simulate

data with the same parameters as the first row. As expected, we find that DML and IPS estimates are not as accurate and the noise in the data created some small sample biases.

3.3.2 Estimated Propensity Scores Using High-Dimensional Covariate Space

We now turn to the case where the propenstiy scores are not known, but can be estimated using a high-dimensional covariate space. This case is common when researchers use digital platforms' internal data, where the algorithmic output is not stored but the full input of the algorithm is observed, which allows researchers to estimate the assignment function. Formally, it means that although $\pi(X_i)$ is not directly available, the full set of covariates X_i is available to the researcher. As such, we can estimate $\hat{\pi}(X_i)$ using data.

As discussed earlier in the paper in Corollary 2, we no more have guarantees in cases with estimated propensity scores that our ATE estimates will recover the CATE for the probabilistic region in our data (τ_r). In the case where propensity scores are known, we have $W_i = \hat{\pi}(X_i)$, which allows both model-free and model-based approaches to ignore observations with deterministic assignment. However, when propensity scores are estimated $W_i - \hat{\pi}(X_i)$ is not exactly zero for deterministic assignment regions, so the estimator will no more ignore these observations in the data. In this section, we want to see how different models perform under simulated data for cases where propensity scores have to be estimated.

We first describe our simulation procedure, which is largely similar to our procedure in the previous section. The only difference comes from the fact that we now involve X_i in determining the propensity scores and outcomes. Therefore, we need to functions l and g for determining the propensity scores and the nuisance part of the outcome respectively. Ideally, we want these functions to be complex enough to reflect the practice in digital platforms that use algorithmic decision-making. For this purpose, we train two XGBoost models on a subset of normalized data from Rafieian (2022), with users' leave and click decisions as the outcome in these models to determine functions l and g for our simulation practice. Please notice that the data set used is arbitrary and one could use any data to obtain complex functions. With these functions, we proceed with our simulation as follows:

- *Step 1:* For each observation i, we first calculate the probability $l(X_i)$ using the pre-trained l function.
- *Step 2:* Given α_0 and α_1, we first sort the $l(X_i)$ values and allocate the bottom α_0 fraction to deterministic no-assignment region ($\pi(X_i) = 0$) and the top α_1 fraction to deterministic assignment region ($\pi(X_i) = 1$). For the middle α_r fraction values of $l(X_i)$, we re-scale the values with the existing minimum and maximum in these values. Let $l_{\min}$ and $l_{\max}$ denote the minimum and maximum of the values of $l(X_i)$ for the probabilistic region. We use

the following propensity scores for the observations in this region:

$$\pi(X_i) = \frac{l(X_i) - l_{\min}}{l_{\max} - l_{\min}}, \tag{15}$$

where i is in the probabilistic region.

- *Step 3:* We use $\pi(X_i)$ values to simulate the treatment variable W_i.
- *Step 4:* We generate structural error terms $\epsilon_i \sim \mathcal{N}(0, \sigma)$.
- *Step 5:* We first calculate $g(X_i)$ as the nuisance part of the outcome and then use appropriate CATE value from $\{\tau_0, \tau_1, \tau_r\}$ to calculate the outcome as follows:

$$Y_i = g(X_i) + W_i\Big(\mathbb{1}\big(\pi(X_i) = 0\big)\tau_0 + \mathbb{1}\big(\pi(X_i) = 1\big)\tau_1 + \mathbb{1}\big(0 < \pi(X_i) < 1\big)\tau_r\Big) + \epsilon_i. \tag{16}$$

We now simulate data using the same set of parameters as Table 1 and estimate the ATE using the following models:

1. Plain Mean Difference (MD): Like previous section, we regress Y_i on W_i without controlling for any other factor. This is the simple mean difference between the groups and we denote it by $\hat{\tau}_{\text{MD}}$.

2. Regression with Controls: The second estimator we consider is an OLS model where we regress Y_i on W_i and X_i. Since X_i determines the propensity scores, we have unconfoundedness conditional on X_i, so this mimics the practice where the researcher controls for all the confounding factors in a regression model. We denote the ATE estimate under this approach by $\hat{\tau}_{\text{OLS}}$.

3. Double Machine Learning (DML): Like previous section, we use a DML model where we first use a model to fit Y_i and W_i separately using X_i, and then regress the residuals of Y_i on those of W_i. We use two different learners: (1) Random Forests, and (2) XGBoost. The corresponding ATE estimate in each case is denoted by $\hat{\tau}^{\text{RF}}_{\text{DML}}$ and $\hat{\tau}^{\text{XGB}}_{\text{DML}}$.

4. Inverse Propensity Score (IPS) Estimator: We first estimate the propensity scores using an XGBoost model that predicts W_i using X_i. We then trim the sample based on the propensity scores that are below 0.05 or above 0.95 and estimate the ATE using the IPS estimator in Equation (12). We denote this estimate by $\hat{\tau}^{\text{XGB}}_{\text{IPS}}$.

Overall, this gives us five different models to estimate the ATE. We present the results in Table 2. There are a few important insights from this table. First, we find that models are

N	$\{\tau_0, \tau_1, \tau_r\}$	$\{\alpha_0, \alpha_1, \alpha_r\}$	*True ATE*	*Estimated ATE*				
			(τ^*)	$(\hat{\tau}_{\text{MD}})$	$(\hat{\tau}_{\text{OLS}})$	$(\hat{\tau}^{\text{RF}}_{\text{DML}})$	$(\hat{\tau}^{\text{XGB}}_{\text{DML}})$	$(\hat{\tau}^{\text{XGB}}_{\text{IPS}})$
10^5	$\{1, 8, 2\}$	$\{0.25, 0.25, 0.50\}$	3.25	5.2831	4.1834	3.6415	1.5884	2.3930
10^5	$\{1, 8, 2\}$	$\{0.50, 0.50, 0.00\}$	4.50	7.9965	7.9631	7.9202	5.2997	5.6950
10^5	$\{1, 6, 2\}$	$\{0.40, 0.10, 0.50\}$	2.00	3.4704	3.0880	2.7649	1.3277	1.9784
10^5	$\{-1, -8, 2\}$	$\{0.10, 0.10, 0.80\}$	0.70	−0.5836	0.4588	0.8912	1.0512	0.8167
10^3	$\{1, 8, 2\}$	$\{0.25, 0.25, 0.50\}$	3.25	5.2882	4.1905	3.6464	1.5587	2.3268

Table 2: Estimates of Average Treatment Effects (ATE) using different estimators when propensity scores have to be estimated with the full set of covariates. Data are simulated using the corresponding parameters and $\sigma = 3$.

not able to recover neither the true ATE (τ^*) nor the true CATE for the probabilistic region (τ_r). Even when $\tau^* = \tau_r$, most models fail to correctly estimate these parameters. The only exception is the IPS estimator, which can be due to the fact that propensity estimates are very accurate in this case.

Next, we examine the magnitude of bias in the estimates by comparing the estimates to the true ATE (τ^*). As expected, the plain mean difference estimator is the most biased since it does not control for any confounding in the data. The OLS estimator performs better than the plain model but still exhibits larger bias compared to more flexible techniques. This is because the OLS estimator only accounts for the linear relationships that the covariates have with the treatment and outcome. Hence, models that can capture the complexities of these relationships like DML and IPS perform better.

Despite smaller bias of DML and IPS, it is worth emphasizing that the estimates obtained by these models are still largely biased. More importantly, unlike the case with known propensities, these models fail to recover the CATE for the probabilistic region (i.e., τ_r) even when we have all the covariates that determine the true propensity function. This is an important result because in many cases the argument is that having access to all the covariates that the platforms have would allow the researchers to estimate the treatment effects from observational data. However, our results in Table 2 highlight that even the most advanced models are not reliable when propensity scores have to be estimated, even if all the required variables are available.

To illustrate this point further, we note that the instances in Table 2 are ideal in the sense that the researcher has access to all the inputs that determine propensity scores. In reality, we only have access to a subset of covariates used by the algorithm. We simulate this

N	$\{\tau_0, \tau_1, \tau_r\}$	$\{\alpha_0, \alpha_1, \alpha_r\}$	*True ATE*	*Estimated ATE*				
			(τ^*)	$(\hat{\tau}_{\text{MD}})$	$(\hat{\tau}_{\text{OLS}})$	$(\hat{\tau}^{\text{RF}}_{\text{DML}})$	$(\hat{\tau}^{\text{XGB}}_{\text{DML}})$	$(\hat{\tau}^{\text{XGB}}_{\text{IPS}})$
10^5	$\{1, 8, 2\}$	$\{0.25, 0.25, 0.50\}$	3.25	5.2831	4.8005	4.5412	4.2153	4.2367
10^5	$\{1, 8, 2\}$	$\{0.50, 0.50, 0.00\}$	4.50	7.9965	8.0364	7.9818	5.7698	7.5732
10^5	$\{1, 6, 2\}$	$\{0.40, 0.10, 0.50\}$	2.00	3.4704	3.3208	3.1748	2.9796	2.8311
10^5	$\{-1, -8, 2\}$	$\{0.10, 0.10, 0.80\}$	0.70	-0.5836	-0.1248	0.1167	0.1742	0.2705
10^3	$\{1, 8, 2\}$	$\{0.25, 0.25, 0.50\}$	3.25	5.2882	4.7254	4.4349	4.0739	4.1607

Table 3: Estimates of Average Treatment Effects (ATE) using different estimators when propensity scores have to be estimated with a subset of covariates used by the algorithm to determine propensity scores. Data are simulated using the corresponding parameters and $\sigma = 3$.

case by generating the data using the procedure mentioned above, but only use a subset of covariates to estimate the propensity scores. In this case, we expect to have coarser estimates of propensity scores. We present the results of this practice in Table 3. Like Table 2, we find that the models recover neither τ^* nor τ_r. However, almost all the estimates are farther away from their corresponding true parameters, compared to the estimates in Table 2.

3.4 Discussion

In light of our theoretical analysis, we know that the lack of overlap can substantially bias the estimates of the average treatement effects. Now, an important question 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. Here we ask the following question: is this homogeneity assumption (i.e., $\tau_0 = \tau_r = \tau_1$) correct in digital platforms?

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 is directly influenced by 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 some users who have a significant and positive CATE of promotion on demand certainly receive the treatment and some 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 $\tau_0 \leq \tau_r \leq \tau_1$, and therefore a large bias in any observational approach to estimate the ATE. We now formalize this intuition in the following proposition:

Proposition 3. *Let $\tau(X_i)$ denote the CATE for observation unit i. We have:*

1. *If $\tau(X_i)$ and belonging to the deterministic assignment region (i.e., $\mathbb{1}(\pi(X_i) = 1)$ are positively correlated, then we have $\tau_1 \geq \tau^*$.*
2. *If $\tau(X_i)$ and belonging to the deterministic no-assignment region (i.e., $\mathbb{1}(\pi(X_i) = 0)$ are negatively correlated, then we have $\tau_0 \leq \tau^*$.*

Proof. For the proof, we only show the first one, since the second one follows the same logic. We start by proving the following lemma:

Lemma 2. *We have $\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)] = P(\pi(X_i) = 1)\mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]$.*

For brevity in our proof, we first define $Q_i = \mathbb{1}(\pi(X_i) = 1)$. We can now write:

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)] &= \mathbb{E}[Q_i\tau(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Q_i\tau(X_i) \mid Q_i]] \\
&= \mathbb{E}[Q_i\mathbb{E}[\tau(X_i) \mid Q_i]] \\
&= P(Q_i = 1)(1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] + P(Q_i = 0)(0)\mathbb{E}[\tau(X_i) \mid Q_i = 0] \\
&= P(Q_i = 1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] \\
&= P(\pi(X_i) = 1)\mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]
\end{aligned} \tag{17}$$

Now, we use this lemma to prove that if $\tau(X_i)$ and belonging to the deterministic assignment region (i.e., $\mathbb{1}(\pi(X_i) = 1)$ are positively correlated, then we have $\tau_1 \geq \tau^*$. We can write:

$$\begin{aligned}
\tau_1 &= \mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1] \\
&= \frac{P\left(\pi(X_i) = 1\right)\mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]}{P\left(\pi(X_i) = 1\right)} \\
&= \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&\geq \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]\mathbb{E}[\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&= \mathbb{E}[\tau(X_i)] \\
&= \tau^*,
\end{aligned} \tag{18}$$

where the fourth line comes from the fact that the two variables are positively correlated. □

Proposition 3 is important because it shows that even a small correlation can link to a violation of $\tau_0 \neq \tau_r \neq \tau_1$. Therefore, unless we have a strong reason to believe that $\tau_0 = \tau_r = \tau_1$, the assumption is that the equality does not hold. In particular, we expect the algorithmic decision-making in digital platforms to at least implicitly use the information in $\tau^*(\cdot)$.

4 Observational Solution to Overlap Violation

In the previous section, we presented the challenge digital platforms face due to the lack of overlap in observational studies. The problem stems from the deterministic output of algorithms that is used for decision-making in these platforms. Our theoretical analysis shows the extent to which observational methods can produce largely biased and inconsistent estimates of the average treatment effect when the overlap assumption is violated.

In this section, we seek to find an observational solution to this challenge. As such, our goal is to use the existing data to recover the average treatment effects. Of course, 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 this section, we explicitly state our assumptions and data requirements and discuss a novel solution based on the machine learning methods. We first formally define our problem in §4.1, where we present the identification problem caused by the lack of overlap as a missing data problem. Next, in §4.2, we present our solution to the problem and the assumptions under which this solution works. Finally, in §4.3, we present a series of simulated experiments to show how our model performs under different scenarios.

4.1 Lack of Overlap as a Missing Data Problem

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 cannot be identified because there is no variation in the treatment variable when accounting for the propensity score. In general, we can write the conditional average treatment effect as follows:

$$\tau^*(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \mu_1(X_i) - \mu_0(X_i), \tag{19}$$

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

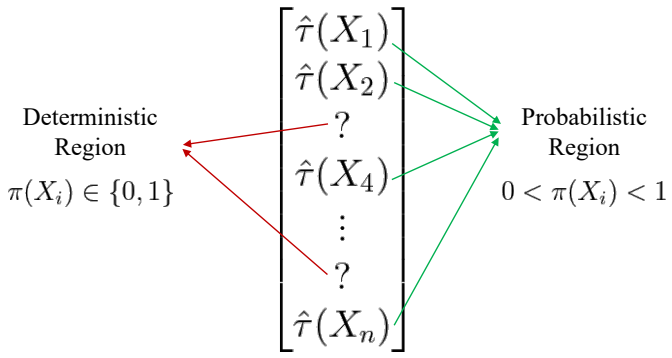


Figure 2: An illustration of the missing data problem due to the overlap violation.

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. That is, for a single treatment, the vector of CATE estimates has missing values for observations in the deterministic regions. Figure 2 visualizes this insight, where the CATE estimates are question marks for observations where the overlap assumption is violated.

We now turn to the question of what variation would allow us to impute these question marks. From our earlier results, we know that with only the data of a single treatment, it is not possible to identify these questions marks. However, we argue that having the data on a set of other treatments for the same set of observation units (e.g., users) can potentially help. That is, instead of exploiting the within-study variation, we can exploit between-study variation. Such a setting is quite common among digital platforms that deliver different treatments at a large scale. Motivated by this insight, we define the problem of the digital platform as follows:

Definition 4. *Consider a digital platform that have data from multiple studies indexed by j from 1 to J. Each study involves a binary treatment variable denoted by $W^{(j)}$, where the value for the i^{th} observation is either zero or one, i.e., $W_i^{(j)} \in \{0, 1\}$. For each study j, the platform has the data $\mathcal{D}^{(j)} = \{Y_i^{(j)}, W_i^{(j)}, X_i, \pi^{(j)}(X_i)\}$, which collectively makes the data*

$\mathcal{D}_T = \bigcup_{j=1}^{J} \mathcal{D}^{(j)}$. *The platform's goal is to recover the following matrix:*

$$\mathcal{T} = \begin{bmatrix} \tau^{(1)}(X_1) & \tau^{(2)}(X_1) & \dots & \tau^{(J)}(X_1) \\ \tau^{(1)}(X_2) & \tau^{(2)}(X_2) & \dots & \tau^{(J)}(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \tau^{(1)}(X_N) & \tau^{(2)}(X_N) & \dots & \tau^{(J)}(X_N) \end{bmatrix}, \tag{20}$$

where $\tau^{(j)}(X_i)$ *is the CATE from the treatment in study* j *for observation unit* i*. Formally, we can define this estimand as follows:*

$$\tau^{(j)}(X_i) = \mathbb{E}[Y_i^{(j)}(1) - Y_i^{(j)}(0) \mid X_i]. \tag{21}$$

If the digital platform achieves the objective in Definition 4, it can recover average treatment effect for the treatment in each study.

A few points are worth noting about the setting and data requirements presented in Definition 4. First, treatments in different studies can be completely different. For example, the treatment in study j and k can be whether a user receives a certain movie recommendation and whether a user receives a free-trial offer. One could imagine this as different interventions the platform made over time.[5] Second, for each study, we need to have the same set of observation units that form rows in the matrix in Equation 20. As such, one user can be assigned to multiple treatments (e.g., both movie recommendation and free-trial in the example above). Third, it is important to emphasize that this data requirement is not excessive as companies often run numerous different treatments over a short period of time.

4.2 Solution Concept

We now present our solution to the problem presented in Definition 4. We first propose the algorithm used for obtaining all the CATE values in Equation (20) in §4.2.1. We then discuss the assumptions that we need for identification in §4.2.2. Finally, in §4.2.3, we propose a test based on our proposed algorithm that allows researchers to examine the extent to which the lack of overlap in their study biases their main estimates of interest.

4.2.1 Proposed Algorithm

Before we present our algorithm, we need to define some model preliminaries. As mentioned earlier, 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

[5]If studies were concurrent, there is the possibility of interference. In our study, we assume no interference.

with the conventional methods to estimate CATE. Therefore, we define the propensity matrix as follows:

$$\Pi = \begin{bmatrix} \pi^{(1)}(X_1) & \pi^{(2)}(X_1) & \dots & \pi^{(J)}(X_1) \\ \pi^{(1)}(X_2) & \pi^{(2)}(X_2) & \dots & \pi^{(J)}(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \pi^{(1)}(X_N) & \pi^{(2)}(X_N) & \dots & \pi^{(J)}(X_N) \end{bmatrix}, \tag{22}$$

where each element $\Pi_{i,j}$ denotes the propensity score for the treatment in study j for unit i, i.e., $\Pi_{i,j} = \pi^{(j)}(X_i) = \Pr(W_i^{(j)} = 1 \mid X_i)$. 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 strictly between zero and one. As such, we can write each elements of this matrix as follows:

$$F = \begin{bmatrix} \mathbb{1}(0 < \pi^{(1)}(X_1) < 1) & \mathbb{1}(0 < \pi^{(2)}(X_1) < 1) & \dots & \mathbb{1}(0 < \pi^{(J)}(X_1) < 1) \\ \mathbb{1}(0 < \pi^{(1)}(X_2) < 1) & \mathbb{1}(0 < \pi^{(2)}(X_2) < 1) & \dots & \mathbb{1}(0 < \pi^{(J)}(X_2) < 1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{1}(0 < \pi^{(1)}(X_N) < 1) & \mathbb{1}(0 < \pi^{(2)}(X_N) < 1) & \dots & \mathbb{1}(0 < \pi^{(J)}(X_N) < 1) \end{bmatrix}. \tag{23}$$

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 $\tau^{(j)}(X_i)$. However, if $F_{i,j} = 1$, we can use conventional CATE estimators to estimate $\tau^{(j)}(X_i)$, because $\pi^{(j)}(X_i)$ is probabilistic and the setting satisfies the unconfoundedness assumption. Therefore, F determines what is identifiable and transform the problem in Definition 4 into a matrix completion problem where we have an estimated CATE matrix $\hat{\mathcal{T}}^{\text{incomplete}}$, where each element $[i, j]$ is defined as follows:

$$\hat{\mathcal{T}}_{i,j}^{\text{incomplete}} = \begin{cases} \hat{\tau}^{(j)}(X_i) & \text{if} F_{i,j} = 1 \\ ? & \text{if} F_{i,j} = 0 \end{cases} \tag{24}$$

As shown in Equation (24), F determines the question marks in our matrix completion task. We now have an incomplete matrix $\hat{\mathcal{T}}^{\text{incomplete}}$, where the incomplete elements are the overlap-violating regions. If the underlying matrix $\mathcal{T}$ is low-rank, we can then 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 $\hat{\mathcal{T}}^{\text{complete}}$. Algorithm 1 presents the details of our proposed approach.

Algorithm 1 Matrix Completion for CATE Estimation

Input: $\mathcal{D}_T$ ▷ From Definition 4
Output: $\hat{\mathcal{T}}^{\text{complete}}$

$F \leftarrow \mathbb{1}(0 < \Pi < 1)$
for $j = 1 \rightarrow J$ **do**
  $\hat{\tau}^{(j)} \leftarrow$ ***learnCATE***$(Y_i^{(j)}, W_i^{(j)}, \{X_i, \pi^{(j)}(X_i)\})$ ▷ Can be any CATE learner
  for $i = 1 \rightarrow N$ **do**
    $\hat{\mathcal{T}}_{i,j}^{\text{incomplete}} \leftarrow \hat{\tau}^{(j)}(X_i)$
    if $F_{i,j} = 0$ **then**
      $\hat{\mathcal{T}}_{i,j}^{\text{incomplete}} \leftarrow ?$
    end if
  end for
end for
$\hat{\mathcal{T}}^{\text{complete}} \leftarrow$ ***softImpute***$(\hat{\mathcal{T}}^{\text{incomplete}})$

The output of this algorithm is a complete matrix $\hat{\mathcal{T}}^{\text{complete}}$ where all the elements have been imputed. This complete matrix can then be used to estimate the ATE from the data. For each treatment in study j, we can recover the average treatment effect as follows:

$$\hat{\tau}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathcal{T}}_{i,j}^{\text{complete}}. \tag{25}$$

If the matrix $\mathcal{T}$ is low-rank, $\hat{\tau}^{(j)}$ is a bias-corrected version of the ATE for treatment j. Under some regularity assumptions on the matrix and a uniformly random missingness pattern, we can use the recent advances in the literature to quantify the uncertainty around the imputed entries in the data (Chen et al., 2019).

4.2.2 Assumptions and Identification

We now discuss the assumptions that we need for matrix completion approach to recover the true average treatment effects. At a high level, our identification claim is that for each observation unit in an overlap-violating region ($F_{i,j} = 0$), if we have enough cross-study variation, we can exploit the similarities in the data to impute the conditional average treatment effect for that observation unit. The following example helps illustrate the intuition. Suppose that treatment j 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 j is unidentifiable for this unit. Now, suppose that there is another treatment j' that

has a probabilistic assignment for unit i, so we can estimate the CATE of j' for unit i. 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 j' for unit i to impute the CATE of j for unit i. Similarly, if there is another unit i' that has a probabilistic assignment for treatment j and is very similar to unit i for most treatments, we can use this similar unit's CATE to impute the missing entry for unit i.

The simple example above is just to illustrate what kind of variation we use in our method. However, such exact similarities may be difficult to find in reality, especially if we have to search on a case-by-case basis. 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.

Now, we ask the question of when matrix completion is suitable for this practice. In general, we need assumptions on two aspects of the matrix: the rank of the matrix and the missingness patterns. Candès and Recht (2009) discuss the assumptions needed for exact low-rank matrix completion in detail. Our case differs from the exact low-rank matrix completion as we do have the true CATEs, but a noisy estimate of the CATE matrix for the feasible regions. As such, the formal requirements we have are analogous to Chen et al. (2019) who discuss the problem of statistical inference in noisy matrix completion.

The most important requirement is that we need the CATE matrix $\mathcal{T}$ to be low-rank. The low-rank requirement intuitively means that the user response exhibits some common patterns across different treatments in different studies. The result of such an environment is that we can exploit similarities across users ans across treatments and decompose the matrix at a reasonably low computational cost. For example, if the set of studies involve promotional treatments, we expect the CATE in most of them to depend on the price elasticity of a user. Similarly, if a digital platform runs notifications and each notification is a different treatment, we expect such similarities, because users who are responsive to one notification are more likely to respond to another notification. In particular, within the context of the same digital platform, it is reasonable to assume that interventions share some common characteristics, and more importantly, that the heterogeneity in the user response depends on a few broad user-level characteristic. The following remark justifies why the low-rank assumption is reasonable in these domains:

Remark 1. *Let $X_{N\times D}$ denote the covariate matrix where each row represents a user and each column represents a covariate. The CATE from treatment j for unit i is $\tau^{(j)}(X_i)$, which is a function of the covariates. For each treatment j, there is a D-dimensional vector of*

coefficients $\beta^{(j)}$ that determine the CATE value such that $\tau^{(j)}(X_i) = \beta^{(j)} X_i^T$. This linear approximation is reasonable as D can be large. Now, we can write the CATE matrix $\mathcal{T}$ as follows:

$$\mathcal{T} = XB^T, \tag{26}$$

where B is a $J \times D$ matrix where each column is the vector of coefficients for CATE for a specific treatment. Now, if the underlying heterogeneity is mainly driven by a few factors, the coefficients for most covariates become zero and only a few coefficients are important. A few instances of these important factors are price elasticity or age that are shown to be explain most of the heterogeneity in treatment effects. If only a few factors explain the heterogeneity in treatment effects, this means that the effective dimension of B is not D, but a number considerably smaller. Therefore, the CATE matrix $\mathcal{T}$ is, by construction, a low-rank matrix.

The second set of requirements for matrix completion to work involves the missingness pattern. Intuitively, the missingness pattern needs to be such that we can jointly exploit the similarities between users and between treatments. As such, if the data are missing for an entire column, there is no way to recover the parameters for that column. Likewise, if the data are entirely missing for a row, the matrix completion approach cannot exploit the similarities in any ways. Thus, although the missingness pattern can be non-random, a few entries are needed for each row and each column.

4.2.3 Statistical Test 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 a simple statistical test based on our algorithm to examine whether the lack of overlap can cause bias in estimating the ATE when propensity scores are known.

In this case, we 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 ($\pi^{(j)}(x) = 0$). As such, we can run our algorithm to first estimate CATE for the probabilistic regions and build the incomplete matrix $\hat{\mathcal{T}}^{\text{incomplete}}$. We can then obtain the completed matrix $\hat{\mathcal{T}}^{\text{complete}}$ using our algorithm. Hence, for each column j in matrix $\hat{\mathcal{T}}^{\text{complete}}$, we have three corresponding

groups of elements. This allows us to statistically test if $\tau_r^{(j)} = \tau_0^{(j)}$ and $\tau_r^{(j)} = \tau_1^{(j)}$, which is the underlying rationale behind 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 reject any of these tests, it means that the lack of overlap can fundamentally bias the estimates of average treatment effects.[6]

4.3 Simulation Experiments

In this section, we deliver a series of simulation experiments using synthetic data to validate our proposed algorithm. We consider a variety of cases that reflect real-world scenarios. Each scenario corresponds to a certain missingness pattern (random or non-random) and the extent of missingness. To show how our proposed algorithm performs, we need to make a ground truth CATE matrix $\mathcal{T}$ with N rows that represent users and J columns that represent studies. As discussed earlier in Remark 1, we can define the ground truth CATE matrix as the product of the covariate matrix $X_{N\times D}$ and the transpose of the coefficient matrix $B_{J\times D}$ as follows:

$$\mathcal{T} = XB^T. \tag{27}$$

We further decompose matrix B to control the rank of the CATE matrix for our experiment. We define two matrices $U_{J\times R}$ and $V_{D\times R}$ where we have:

$$B = UV^T, \tag{28}$$

where R controls the rank of the $N \times J$ CATE matrix. Intuitively, we can interpret XV as the CATE for R principal components that define a CATE for the treatment in a specific study through some weights. These weights are specified for each of the J studies in matrix U. Together, XVU^T gives us the underlying CATE matrix $\mathcal{T}$, which is low-rank.

Our goal in the simulation experiments is to generate data $\mathcal{D}_T$, as defined in Definition 4. This data set is the union of data sets corresponding to each study j. To generate $\mathcal{D}_T$, we need three inputs: (1) CATE matrix $\mathcal{T}$ that determines the treatment effect for each observation, (2) propensity matrix Π as defined in (22), and (3) nuisance matrix $\mathcal{G}$ that determines the relationship between covariates and the outcome. We can use these three

[6]It is worth emphasizing that these tests only use the point estimates for CATEs. A more robust approach in conducting these tests is to incorporate the uncertainty in our estimates. For that purpose, we can use the approach proposed by Chen et al. (2019) that build confidence intervals for the imputed entries of a noisy matrix.

inputs and simulate $\mathcal{D}_T^{\text{sim}}$ using the following procedure:

- *Step 1:* We use Π to simulate $W_i^{(j)}$ for each unit i in each study j.
- *Step 2:* With the treatment variable realized, we can simulate the outcome as follows:

$$Y_i^{(j)} = \mathcal{G}_{i,j} + W_i^{(j)} \mathcal{T}_{i,j} + \epsilon_{i,j}, \tag{29}$$

  where $\mathcal{G}_{i,j}$ is the nuisance part of the outcome, $W_i^{(j)} \mathcal{T}_{i,j}$ is the treatment effect given (if any), and $\epsilon_{i,j} \sim \mathcal{N}(0,1)$.
- *Step 3:* For each study j, we can construct data set $\tilde{\mathcal{D}}^{(j)} = \{Y_i^{(j)}, W_i^{(j)}, X_i, \pi^{(j)}(X_i)\}$. The union of $\tilde{\mathcal{D}}^{(j)}$ for all j's will give us the $\mathcal{D}_T^{\text{sim}}$.

For our simulations, we use $N = 1000$, $D = 50$, and $J = 100$. We set the rank of the CATE matrix as $R = 10$ and generate two random matrices $X_{N \times D}$ and $B_{J \times D}$, where each element of each matrix comes from $\mathcal{N}(0,1)$. We generate another coefficient matrix $G_{J \times D}$ from $\mathcal{N}(0,1)$ to generate the nuisance matrix $\mathcal{G}$ as follows:

$$\mathcal{G} = XG^T. \tag{30}$$

What varies across our simulation experiments is the missingness patterns that is operationalized by Π. We formalize this matrix in the following sections. Once we generate the data $\mathcal{D}_T^{\text{sim}}$ in a specific condition, we can apply our proposed method in Algorithm 1. Because we want to compare the performance of our proposed method with conventional methods such as Double ML, we use R-learner as our CATE estimator that directly uses Robinson's decomposition (Robinson, 1988).

In our simulation scenarios, we distinguish between two forms of missingness: random and non-random. Random missingness is similar to cases where platforms choose a very small but random subsample of their users and run the experiment. In these cases, we expect all the conventional methods such as Double ML to work well. Non-random missingness scenarios are those where we expect to see the difference between our proposed method and Double ML. Intuitively, these non-random cases are similar to adversarial missingness patterns where the task of estimating ATE is very challenging. We expect our method to perform better than Double ML methods in these cases.

4.3.1 Random Missingness Pattern

We start with the most well-known missingness pattern used in matrix completion problems: missing-completely-at-random. In this scenario, each entry in the matrix has a uniform

probability of being missing. As such, each element in our feasibility matrix takes value one with probability p, and zero with probability $1-p$. Specifically, we can write the propensity scores as follows:

$$\Pi_{i,j} = \begin{cases} 0 & \text{with prob } (1-p)/2 \\ 1/2 & \text{with prob } p \\ 1 & \text{with prob } (1-p)/2 \end{cases} \tag{31}$$

A random missingness of the elements in a matrix resembles the key intuition behind trimming approaches: if the overlap-violating regions are selected at random, then conventional models can recover the average treatment effect (ATE) as discussed earlier in §3.4. However, we want to consider this case as a starting point to compare the performance of our algorithm with that of the conventional approaches. We consider four different values for p: 0.2, 0.4, 0.6, and 0.8. We apply our Double ML and our proposed algorithm to the data to estimate the Average Treatment Effect for each study j.

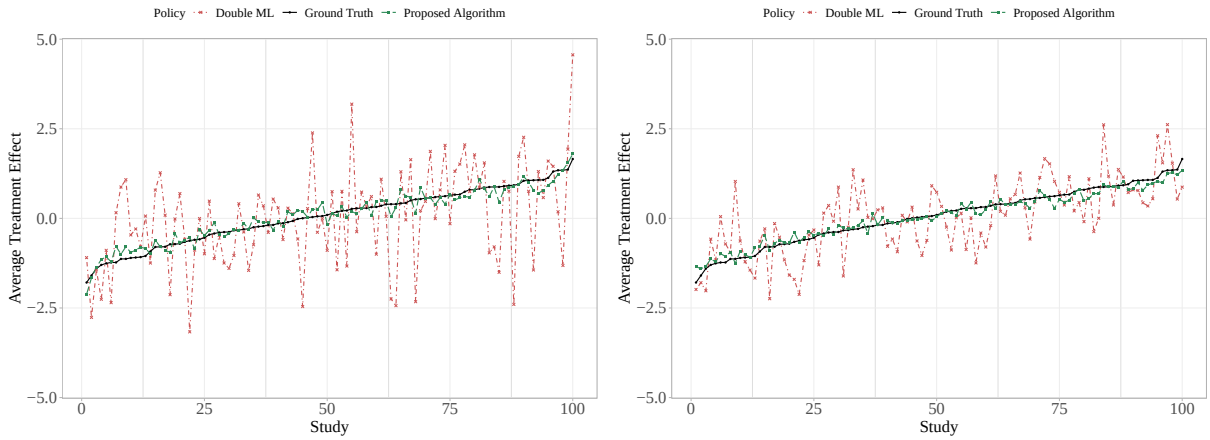
Figure 5 shows four figures, each corresponding to a certain p. The x-axis presents studies as sorted by their true Average Treatment Effect (ATE). As shown in these figures, conventional approaches like DML can all recover the true ATE, as expected. However, our matrix completion algorithm is more accurate even in these cases due to the fact that it uses data from other studies, especially in cases where a large portion of the elements of the CATE matrix is missing ($p = 0.2$).

4.3.2 CATE-Dependent Missingness: Case of Algorithmic Decision-Making

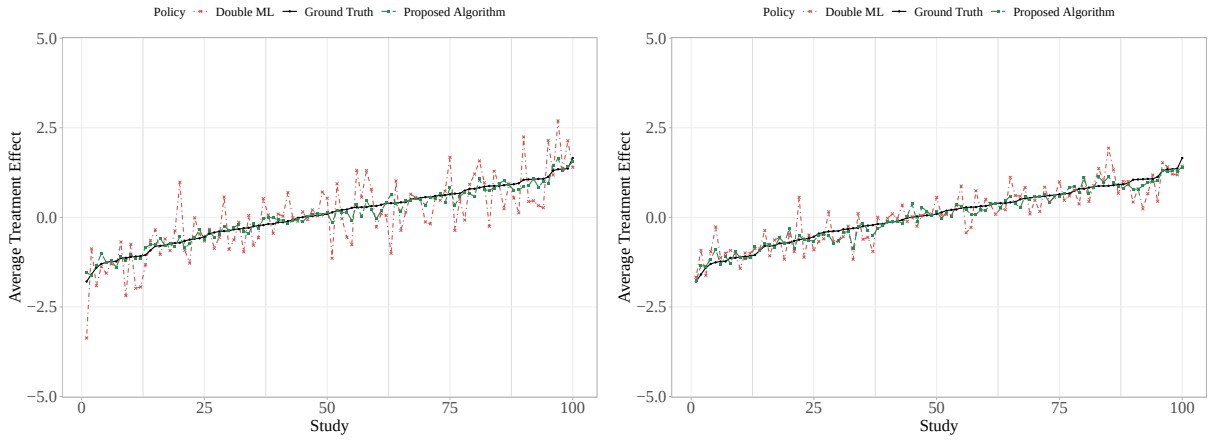
In real-world scenarios, we do not expect to have a random missingness pattern. As discussed in §3.4, we expect the overlap-violating regions to be correlated with the CATE, in the context of algorithmic decision-making. This case is more troublesome as the conventional approaches can be arbitrarily biased. Our goal is to see how our proposed approach performs under these scenarios. Before we present the scenarios, we define a score variable $s_i^{(j)}$ as follows:

$$s_i^{(j)} = \frac{\sum_{i=1}^{N} \mathbb{1}(\tau^{(j)}(X_i) > \tau^{(j)}(X_k))}{N}, \tag{32}$$

which determines the percentile of CATE for user i in study j in the distribution of CATEs in study j. The score $s_i^{(j)}$ shows the relative position of observation i in study j in contributing to a higher ATE in that study. We use this score variable to design different scenarios where higher or lower CATEs are systematically missing. To that end, we consider three separate types of missingness as follows:



(a) $p = 0.2$ (b) $p = 0.4$



(c) $p = 0.6$ (d) $p = 0.8$

Figure 3: The performance of our proposed algorithm and conventional observational method when missingness is at random with a uniform probability. Each figure represents a level of missingness.

- *Right-tail missingness of CATE:* The first scenario is a one-sided missingness pattern such that elements with higher CATE are more likely to be missing. In this case, we want higher CATE values to have a higher probability of being missing. Using the definition of score variable $s_i^{(j)}$ in (32), we can define the propensity scores as follows:

$$\Pi_{i,j} = \begin{cases} 1/2 & \text{with prob } 1 - s_i^{(j)} \\ 1 & \text{with prob } s_i^{(j)} \end{cases} \tag{33}$$

  Figure 4a shows the performance of different methods in recovering the true ATE. As

expected, we find that the conventional approaches (Double ML) largely underestimate the true ATE because the overlap assumption is more likely to violated for observations with higher CATE. However, we find that our proposed algorithm can recover the true ATE accurately by exploitng the cross-study variation.

- *Left-tail missingness of CATE:* The second scenario is again a one-sided CATE-dependent missingness, but this time the missingness happens at the lower end of the CATE distribution. We use the same insight for constructing the missingness pattern based on the score variable as before. We can write:

$$\Pi_{i,j} = \begin{cases} 0 & \text{with prob } 1 - s_i^{(j)} \\ 1/2 & \text{with prob } s_i^{(j)} \end{cases} \tag{34}$$

  In this case, the lower the score is, there is a higher chance that the propensity score is equal to zero. Figure 4b compares the performance of our proposed algorithm with that of the conventional approaches for this scenario. Like the previous scenario, the conventional approaches fail to recover the true ATE. However, this time, these models overestimate ATE because the lower end of the CATE distribution violates the overlap assumption, and is therefore missing. Our proposed algorithm, on the other hand, recovers the true ATE.
- *Alternating one-sided missingness of CATE:* The third CATE-dependent missingness is a mix of the first two, where we have a one-sided missingness for each study, but the direction alternates. That is, for some studies, we have right-tail missingness of CATE, whereas for some other studies, we have left-tail missingness of CATE. Figure 4c shows the results of this simulation. Our proposed method is able to recover the true ATE in this scenario, whereas the Double ML approach exhibits a high magnitude of bias in recovering ATE across studies.
- *Two-sided missingness of CATE:* The fourth CATE-dependent missingness that we consider is a two-sided missingness, where the observations from both ends of the CATE distribution are more likely to violate the overlap assumption. Like the previous scenarios, this scenario also frequently arises in the context of algorithmic decision-making. We operationalize this scenario by assigning a higher probability of missingness to observation

on both ends of the CATE distribution as follows:

$$\Pi_{i,j} = \begin{cases} 0 & \text{with prob } (1 - s_i^{(j)})/2 \\ 1/2 & \text{with prob } 1/2 \\ 1 & \text{with prob } s_i^{(j)}/2 \end{cases} \tag{35}$$

In this scenario, higher scores have a higher probability of being in a deterministic assignment, whereas lower scores have a higher probability of being in a deterministic no-assignment. Figure 4d shows the performance of Double ML in this case and compares it with the performance of our proposed method. As shown in this figure, although Double ML produce largely biased estimate in either direction, our proposed method is able to recover the true ATE in this scenario.

4.3.3 User-Dependent Missingness

We now discuss a different type of missingness that depends on users. That is, the data for some users is more sparse than others. If this missingness is at random, neither conventional approaches nor our proposed algorithm have any problem in recovering the ATE. However, if the missingness probability is different for users who are more or less responsive to interventions (higher or lower average CATE across studies), it is not clear how different methods will perform. To do so, we first define the overall sensitivity of the user to interventions as follows:

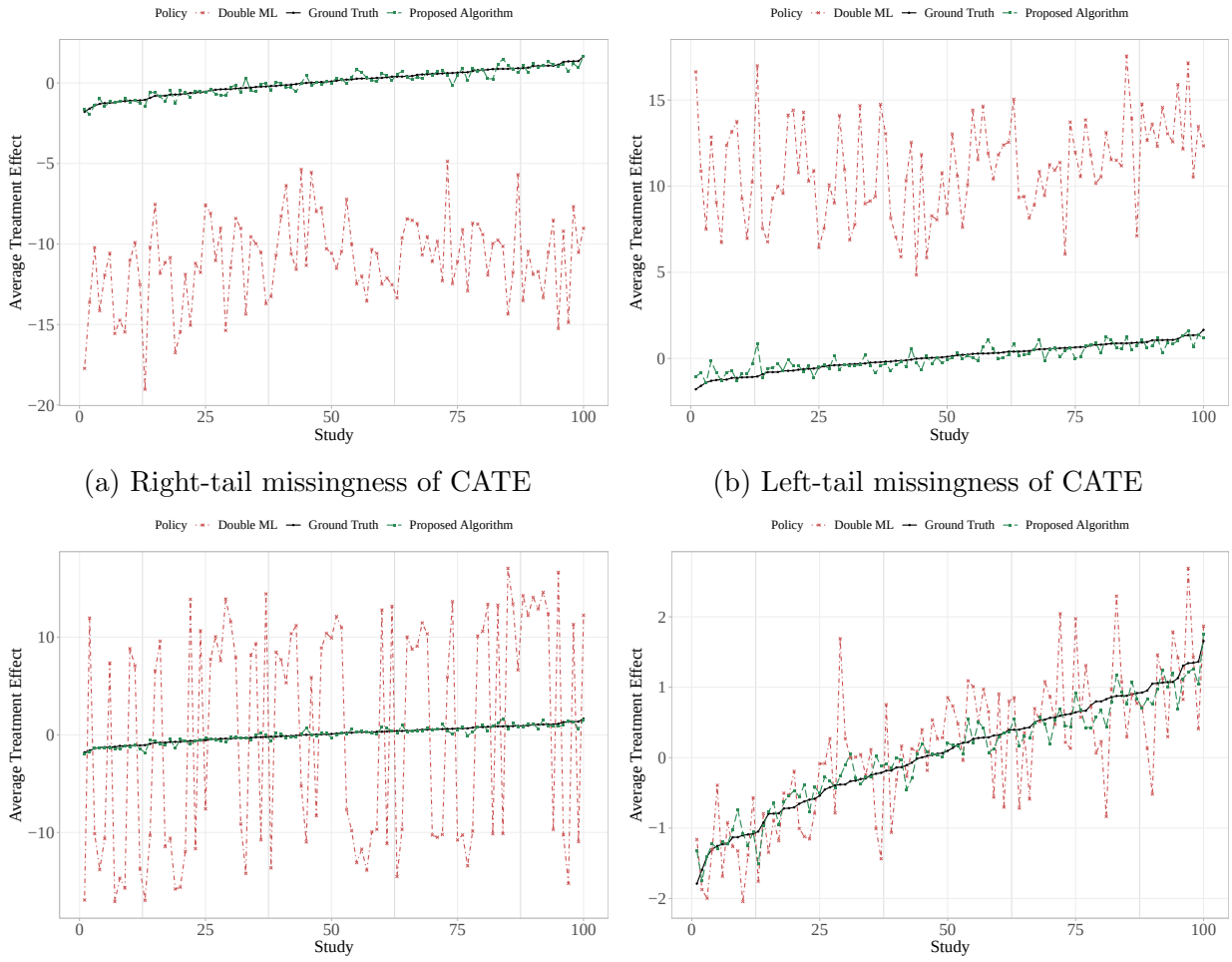
$$\tilde{\tau}_i = \frac{\sum_{j=1}^{J} \tau^{(j)}(X_i)}{J}. \tag{36}$$

We use this notion of sensitivity to create two different user-dependent missingness patterns through matrix F as follows:

- The first scenario we consider is a probabilistic user-dependent missingness. For each user, we define a notion of relative sensitivity, which is the absolute value of their sensitivity divided by the maximum absolute value of all user sensitivities. The higher the relative sensitivity, the more likely the user is to have sparse data. We operationalize this insight through matrix F as follows:

$$\Pi_{i,j} = \begin{cases} \mathbb{1}(\tilde{\tau}_i > 0) & \text{with prob } \left(\frac{|\tilde{\tau}_i|}{\max_k |\tilde{\tau}_k|}\right)^{\lambda} \\ 1/2 & \text{with prob } 1 - \left(\frac{|\tilde{\tau}_i|}{\max_k |\tilde{\tau}_k|}\right)^{\lambda} \end{cases} \tag{37}$$

where λ can be set to manipulate the extent of the missingness and $\mathbb{1}(\tilde{\tau}_i > 0)$ determines



(a) Right-tail missingness of CATE

(b) Left-tail missingness of CATE

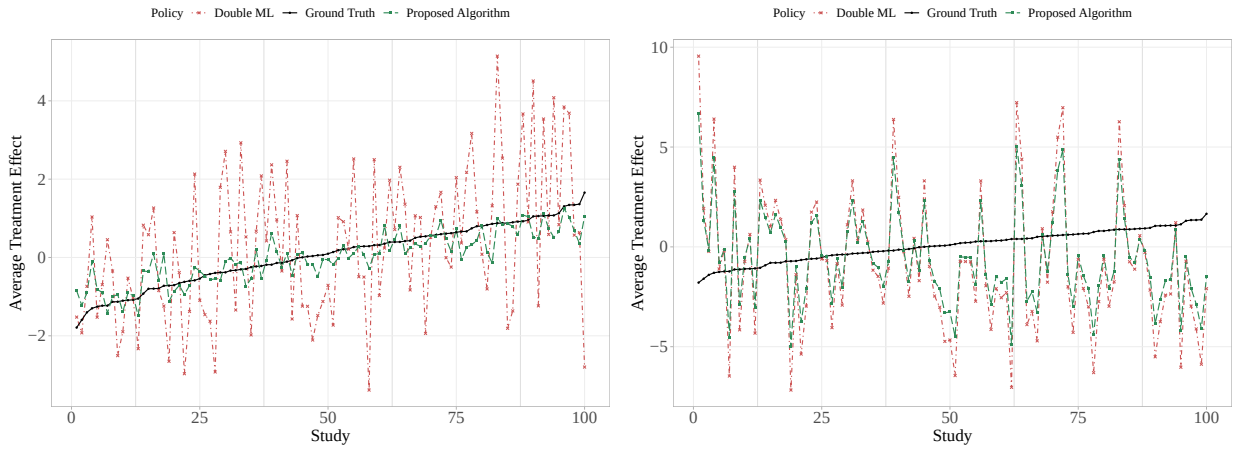
(c) Alternating one-sided missingness of CATE

(d) Two-sided missingness of CATE

Figure 4: The performance of our proposed algorithm and conventional observational method when missingness depends on CATE.

whether the user will be assigned to the deterministic no-assignment or deterministic assignment. We use $\lambda = 1/10$ that generates 80% overlap-violating entries in our data. We present the results from this simulation experiment in Figure 5a. As shown in this figure, Double ML method fails to recover the ATE with such missingness patterns. However, our proposed algorithm can successfully recover the true ATE, despite the extent of the missingness pattern. This is an important finding as the extent of sparsity is systematically different across users.

- The second scenario we consider is a case where some observation units consistently violate the overlap assumption in all studies. As such, the data are entirely missing for some rows. As discussed earlier, this would not be an issue for the task of recovering ATE



(a) Probabilistic missingness of sensitive users (b) Deterministic missingness of sensitive users

Figure 5: The performance of our proposed algorithm and conventional observational method in scenarios with user-dependent missingness.

if the missingness of user data is at random. However, if more sensitive users are more likely to be missing in our data, this can create issues. We operationalize this type of missingness as follows:

$$\Pi_{i,j} = \begin{cases} 0 & \text{if } \tilde{\tau}_i \geq c \\ 1/2 & \text{if } c \geq \tilde{\tau}_i \geq -c \\ 1 & \text{if } \tilde{\tau}_i \leq -c \end{cases} \tag{38}$$

where c is a constant reference value. We set this reference value at the cutoff for top 15% of the sensitivity values. This simulation is important because we normally expect the matrix completion approach to fail in circumstances of complete missingness of some rows. Figure 5b shows the results from this simulation experiment and verifies this limitation. As shown in this Figure, both conventional approaches and our proposed method fail to recover the true ATE. In fact, their performances are largely similar, which comes from the fact that with entire missingness of the entries for the user, the matrix completion approach does not have a better way for imputation than a simple observed mean imputation for observed entries.

5 Conclusion

Digital platforms use algorithmic decision-making to deliver interventions to their users at a very large scale. An important goal for both practitioners and academic researchers is to identify 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. Lastly, we formulate the identification problem caused by the lack of overlap as a missing data problem and propose a matrix completion 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. 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

A. Agarwal, M. Dahleh, D. Shah, and D. Shen. Causal matrix completion. *arXiv preprint arXiv:2109.15154*, 2021.

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.

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

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.

J. Bennett, S. Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, 2007.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Y. Chen, J. Fan, C. Ma, and Y. Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.

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.

D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

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. 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.

N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariates via matrix factorization. *Advances in neural information processing systems*, 31, 2018.

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

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

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.

B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*,

12(12), 2011.

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.

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. 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.

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