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

Omid Rafieian*
Cornell Tech and Cornell University

Abstract

Digital platforms deliver numerous interventions to their users. One of platforms' 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. 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 population average treatment effect from observational data. 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. Using both synthetic and real data in the context of advertising, we demonstrate the performance of our algorithm and quantify the economic gains from using it for decision-making.

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

*The author thanks David Blei, Daria Dzyabura, Tesary Lin, Unnati Narang, Matt Osborne, Caio Waisman, and Hema Yoganarasimhan for detailed comments that have improved the paper. The author also thanks the participants of the 2023 UT Dallas FORMS conference and Temple University marketing seminars for their feedback. Please address all correspondence to: or83@cornell.edu.

1 Introduction

Algorithms help digital platforms scale the number of interventions they deliver to their users. One of the platforms' 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. Thus, it is 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.

What is different in digital platforms is that the unconfoundedness assumption is more plausible than in most settings. The treatment assignment rule is known as the platform itself delivers the interventions. However, the challenge in these settings comes from 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, thereby violating the overlap assumption.

In this paper, we consider the case of a digital platform whose context satisfies the 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. What are the consequences of overlap violation for estimating causal parameters such as the population average treatment effects? Can state-of-the-art causal machine learning methods estimate causal parameters under overlap violation?

2. How can we design an algorithm to overcome the challenges posed by the overlap violation? What are the required assumptions for this solution to work/

3. How likely is the problem of overlap violation in real application settings? How does the proposed solution perform in the presence of these challenges? What are the gains from using our algorithm for decision-making?

To address these questions, we develop a simple framework that categorizes data into three regions based on treatment assignment: (1) probabilistic assignment, where the propensity score is between

0 and 1, (2) deterministic assignment, where the treatment is assigned with probability 1, and (3) deterministic no-assignment, where the treatment is never assigned (propensity score is 0). Only the probabilistic assignment region satisfies the overlap assumption. We define conditional average treatment effects (CATE) for each region, allowing for potential differences at the population level.

Our theoretical analysis shows that without overlap, both model-based and model-free state-of-the-art methods for estimating the Average Treatment Effect (ATE) can only estimate the CATE for the probabilistic region and fail to estimate the population ATE. Specifically, when assignment probability depends on the conditional average treatment effect, the absence of overlap can lead to sizable differences between the CATE for the probabilistic region and the population ATE. We highlight the prevalence of this issue in algorithmic decision-making, demonstrating that even slight connections between heterogeneous treatment effects and deterministic assignment can cause systematic biases in population ATE estimates.

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 our algorithm, we set 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: if there are a few factors that collectively determine CATE for each study, we can exploit similarities across users and across treatments to identify those factors and impute CATEs for units that belong to overlap-violating regions. Once we complete the matrix for the formerly unidentified parts, we can correct the bias in the population ATE estimates.

We then use a calibrated simulation in the context of online advertising, where we micro-found the algorithmic ad allocation through advertising auctions. In this simulation, we are interested in measuring ad effectiveness for a series of ads on a population of users. We first theoretically show that the algorithmic ad allocation violates the overlap assumption because ads with lower bids will have a zero propensity score. We then demonstrate that the estimates for population ATE under state-of-the-art ATE estimation methods are largely biased. Notably, we show that our proposed algorithm correctly recovers the ATE for each ad. Further, we evaluate the targeting performance of our algorithm and show substantial economic gains for the advertising platform from using our

algorithm compared to a series of benchmarks. Together, our results demonstrate the superior performance of our algorithm compared to the existing benchmarks.

Finally, we conduct an empirical validation exercise based on the data from a leading in-app advertising platform in a large Asian country. We particularly focus on this platform due to its use of extensive randomization in ad allocation, which provides us with a ground-truth benchmark to validate the assumptions needed for our algorithm and evaluate the performance of our model. We first define an underlying CATE matrix for this application and show that this matrix is low-rank. We then introduce a counterfactual setting wherein an ad allocation algorithm is used to demonstrate how this intuitive allocation mechanism can lead to overlap violation. Interestingly, we find that the extent of the overlap violation is so severe that the CATE for the probabilistic region has the opposite sign from the population ATE. Despite this large discrepancy, we show that our algorithm can still recover the true ATE using the variation across ads. We further demonstrate the practical gains from using our algorithm for targeting and show substantial gains compared to the benchmarks. Together, our calibrated simulation and empirical application provide evidence for the performance of our model and the practical value it creates for managers and decision-makers.

In summary, our paper makes several contributions to the literature. Methodologically, we present a comprehensive study of the overlap assumption and theoretically characterize the context in digital platforms that use algorithmic decision-making. In particular, we propose a novel machine-learning solution that views the identification challenge as a missing data problem and combines heterogeneous treatment effect estimation with matrix completion to recover the treatment effects. From a substantive and practical viewpoint, 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 discuss empirical contexts where this problem may arise and present the value from using our algorithm using both synthetic experiments and real field data. Overall, our proposed algorithm is fairly general and can be applied to many contexts, specifically those in digital settings where platforms deliver numerous interventions that have common factors and satisfy the low-rank requirements. Thus, we expect our framework to be valuable for platforms in utilizing their existing observational data and researchers who access the data from such platforms.

2 Related Literature

Broadly, our paper relates to the causal inference literature that aims to estimate treatment effects (Neyman 1923, 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 the unconfoundedness assumption has received considerable attention in the literature, the overlap assumption has often been viewed as a more straightforward 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), extra assumptions that help recover causal estimands for overlap-violating regions (Nethery et al. 2019), 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 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 correct 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, Chernozhukov et al. 2018a, Athey et al. 2018) 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 2023), video advertising format (Rafieian et al. 2023), 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.

Third, 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 applying matrix completion models to exploit cross-study variation and recover the true causal parameters. Specifically, we bring the recent advancements in CATE estimation to the matrix completion problem to help utilize the rich information in the covariate space.

Finally, our paper relates to the stream of literature on advertising effectiveness and measurement (Lewis et al. 2011, Johnson et al. 2017a,b, Gordon et al. 2019, 2022). In particular, a stream of work in this domain has focused on the measurement problems even in the presence of randomized controlled trials, such as statistical power issues (Lewis and Rao 2015, Johnson et al. 2017b) or the compliance issue (Johnson et al. 2017a). Another series of papers have investigated the possibility of estimating true ad effectiveness measures by using observational methods (Lewis et al. 2011, Gordon et al. 2019, 2022). Our work extends this body of work in several ways. First, we bring the possibility of overlap violation as an explanation largely missing from the prior literature for the inability of observational methods to recover ad effectiveness. In particular, we develop a micro-founded model of algorithmic ad allocation and show – through a formal lemma and simulation – that numerous user-ad pairs have practically zero propensity scores, thereby violating the overlap assumption. Second, our work differs from these papers in proposing an algorithmic solution to the problem of overlap violation.

3 Overlap Violation in 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 observable 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.

In order to determine which intervention to deliver in each observation, digital platforms gen-

erally 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 direct 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 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

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

[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.

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:*

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

  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. Intuitively, this assumption ensures that the distribution of covariates under treatment fully overlaps with that of covariates under control.

The strong ignorability assumption serves as the foundation for studies of causal inference. The most common challenge in observational 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 the average treatment effect.

The key difference in digital platforms that employ algorithmic decision-making is that the assignment rule is often fully observable. That is, the platform can easily store the X_i used for algorithmic decision-making and the output of the algorithm $\pi(X_i)$, which is shown to be sufficient to satisfy the unconfoundedness assumption (Rosenbaum and Rubin 1983). Hence, observational studies on digital platforms do not suffer from the well-known confoundedness or endogeneity problem since there is no selection on unobservables. What makes these observational studies challenging is the commonly ignored part of the strong ignorability assumption, which requires the treatment assignment to be probabilistic. Although the 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.

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

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 and conditional average treatment effects (CATE) for each value of the vector of covariates.*

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 Theoretical 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 such as double machine learning perform in estimating the ATE in §A.1. Finally, we focus on model-free approaches such as importance sampling and theoretically derive their properties in §A.2.

3.2.1 Identification Challenge

In this section, we present a simple framework to illustrate how the violation of overlap poses challenge for estimating the population ATE. To do so, we first introduce a new notation that captures 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*: 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 by α_r.
- *Deterministic no-assignment region*: 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 receive the control. We denote the fraction of such observations in our data by α_0.
- *Deterministic assignment region*: 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 that certainly receive the treatment. We denote the fraction of such observations in our data by α_1.

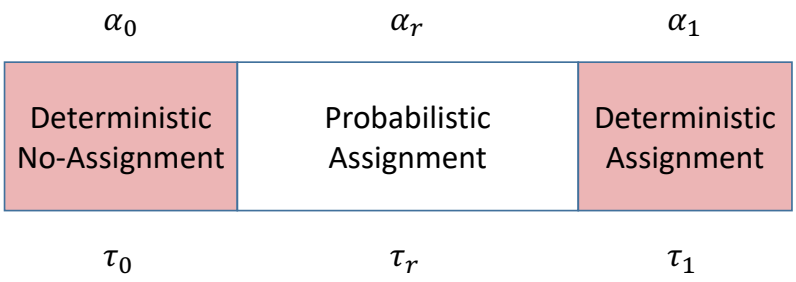


Figure 1. Different regions based on the type of assignment.

Now, we can define the average treatment effect as $\tau^* = \alpha_r\tau_r + \alpha_0\tau_0 + \alpha_1\tau_1$, where $\alpha_r + \alpha_0 + \alpha_1 = 1$. 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, since there is no residual variation in treatment. Thus, we can write the following lemma:

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

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. One argument is that state-of-the-art ATE estimation methods that combine flexible machine learning models with causal inference can capture very complex treatment assignment mechanisms and address potential selection issues.[4] A few prominent examples of these advanced methods are Double Machine Learning (Chernozhukov et al. 2018a) and Approximate Residual Balancing (Athey et al. 2018). Inspired by these developments, Gordon et al. (2022) test this possibility in the context of online advertising and consider both Double Machine Learning (DML) and Propensity Score Matching (PSM) methods as model-based and model-free benchmarks, respectively. The following proposition shows that state-of-the-art model-based and model-free approaches could only estimate the identifiable piece τ_r and fail to estimate the population ATE τ^*:

Proposition 1. *Suppose 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. Let $\hat{\tau}_{DML}$ and $\hat{\tau}_{IPS}$ denote the ATE estimate based on Double Machine Learning and Inverse Propensity Scoring estimators, respectively. Both these estimates converge to τ_r in probability, that is, $\hat{\tau}_{DML} \xrightarrow{p} \tau_r$ and $\hat{\tau}_{IPS} \xrightarrow{p} \tau_r$.*

Proof. See Web Appendix B.1. □

[4] Please see a summary of state-of-the-art model-based and model-free approaches to estimate treatment effects in Appendix B.

Lemma 1 and Proposition 1 highlight an important identification problem for state-of-the-art ATE estimation models that cannot be fixed with higher expressiveness and complexity of the machinery used in these models. The bright side, however, is that these methods are guaranteed to estimate the CATE for the probabilistic region, thereby allowing researchers to precisely set the scope of their interpretations. This is something we use later when developing our proposed solution to this problem.

Lastly, a fundamental question is whether our resulting estimates based on the state-of-the-art approaches are far from the true population ATE τ^*, and if so, whether this is consequential for decision-making. We can characterize the difference between these estimates from the true population ATE as follows:

$$|\tau^* - \hat{\tau}| \xrightarrow{p} |\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)|. \tag{5}$$

This equation highlights the fact that if the treatment effects for the deterministic regions are the same as the treatment effect for the probabilistic region, there will be no difference between τ_r and τ^*. However, it is easy to imagine scenarios where the difference in τ_1, τ_0, and τ_r creates a substantial difference in estimates of the average treatment effect. In fact, 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 for this difference. In the rest of this paper, we present application cases that show in which cases this difference is large and to what extent it leads to economic losses for managers who make decisions based on these estimates.

3.2.2 Practical Relevance

As discussed earlier, the difference between the identifiable piece τ_r and the population ATE τ^* can be arbitrarily large. An important question is whether this is just a theoretical possibility that is not practically important. In other words, do we expect the difference $|\alpha_0(\tau_0 - \tau_r) + \alpha_1(\tau_1 - \tau_r)|$ to be large in real settings? Part of the rationale for the trimming approaches that are 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, which is of interest to the researcher. We discuss two prime examples of such settings in practice:

- *Promotions*: In the context of promotions, many digital platforms use algorithmic scores and thresholding rules to assign users to promotions (Shi et al. 2022). That is, users with a score above a certain threshold will deterministically receive the promotion, which creates a case

of *deterministic assignment*. The rest of the users will either be assigned to the probabilistic assignment or even deterministic no-assignment, depending on the context. The algorithmic scores are generally measured using supervised learning models that use the responsiveness of users. As such, there is some positive correlation between CATEs and belonging to the deterministic assignment region.

- *Advertising Auctions*: Digital ads are sold through auctions. In such settings, advertisers place bids per impression and win only when their submitted bid is the highest among all bidders. The advertiser's submitted bid per impression for a user is a function of the CATE of that ad for the user (Waisman et al. 2019). The auction setting implies that an ad could never reach a certain user if the CATE for that user is too low because there will always be advertisers with higher bids for that user. This creates a form of *deterministic no-assignment*: some users in the control condition could have never seen the ad because of their low valuation for the advertiser. Therefore, there will be a negative correlation between CATEs and belonging to the deterministic no-assignment region.

The examples above characterize practical settings where deterministic assignment happens in a way that violates the homogeneity of treatment effects across regions, i.e., we have $\tau_0 \neq \tau_r \neq \tau_1$. In the examples above, we expect to have $\tau_0 \leq \tau_r \leq \tau_1$, and therefore a large bias in any observational approach to estimate the ATE. We now formalize this intuition in the following proposition:

Proposition 2. *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. See Web Appendix B.2. □

Proposition 2 is important because it shows that even a small correlation can link to a violation of $\tau_0 \neq \tau_r \neq \tau_1$. Later in our applications, we focus on realistic assignment policies and examine whether this issue could arise in practice. We further propose a solution and evaluate how our proposed solution performs in recovering the causal estimates and making decisions based on those estimates.

4 Proposed Algorithm

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 outputs of algorithms that

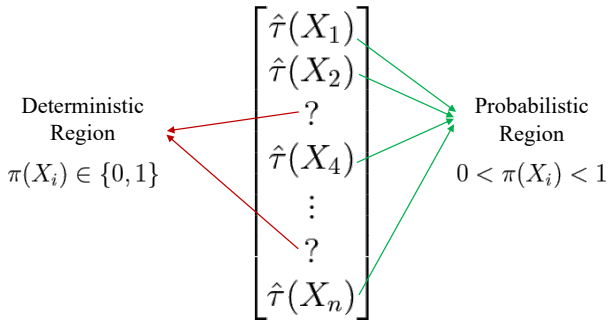


Figure 2. An illustration of the missing data problem due to the overlap violation.

are 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 explicitly state our assumptions and data requirements and propose a novel solution based on machine learning methods to overcome the challenge posed by the overlap violation.

4.1 Problem Definition: Overlap Violation 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:

$$\tau^*(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \mu_1(X_i) - \mu_0(X_i), \tag{6}$$

where $\mu_w(x)$ is the population function for potential outcomes conditional on x when assigned to treatment w. From a learning standpoint, if one of the two treatment states could have never been generated in the data, no model can estimate the corresponding μ function. For example, if a unit with covariates X_i could have never received the treatment, we have no observation in our data to estimate $\mu_1(X_i)$. As such, the problem caused by the lack of overlap is one of missing data. 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 question 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 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 has 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:*

$$\mathcal{T} = \begin{bmatrix} \tau^{(1)}(X_1) & \tau^{(2)}(X_1) & \dots & \tau^{(J)}(X_1) \\ \tau^{(1)}(X_2) & \tau^{(2)}(X_2) & \dots & \tau^{(J)}(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \tau^{(1)}(X_N) & \tau^{(2)}(X_N) & \dots & \tau^{(J)}(X_N) \end{bmatrix}, \tag{7}$$

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:

$$\tau^{(j)}(X_i) = \mathbb{E}[Y_i^{(j)}(1) - Y_i^{(j)}(0) \mid X_i]. \tag{8}$$

If the digital platform achieves the objective in Definition 4, it can recover the 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 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. Second, for each study, we need to have the same set of observation units that form rows in the matrix in Equation (7). As such, one user can be assigned to multiple treatments (e.g., both the movie recommendation and the 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 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 despite the overlap violation. To do so, we first need to know which elements we cannot estimate with existing

methods for CATE estimation. Therefore, we define the propensity matrix as follows:

$$\Pi = \begin{bmatrix} \pi^{(1)}(X_1) & \pi^{(2)}(X_1) & \dots & \pi^{(J)}(X_1) \\ \pi^{(1)}(X_2) & \pi^{(2)}(X_2) & \dots & \pi^{(J)}(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \pi^{(1)}(X_N) & \pi^{(2)}(X_N) & \dots & \pi^{(J)}(X_N) \end{bmatrix}, \tag{9}$$

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 are defined as rows where the propensity score is either zero or one. We know that the CATE 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 element of this matrix as follows:

$$F = \begin{bmatrix} \mathbb{1}(0 < \pi^{(1)}(X_1) < 1) & \mathbb{1}(0 < \pi^{(2)}(X_1) < 1) & \dots & \mathbb{1}(0 < \pi^{(J)}(X_1) < 1) \\ \mathbb{1}(0 < \pi^{(1)}(X_2) < 1) & \mathbb{1}(0 < \pi^{(2)}(X_2) < 1) & \dots & \mathbb{1}(0 < \pi^{(J)}(X_2) < 1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{1}(0 < \pi^{(1)}(X_N) < 1) & \mathbb{1}(0 < \pi^{(2)}(X_N) < 1) & \dots & \mathbb{1}(0 < \pi^{(J)}(X_N) < 1) \end{bmatrix}. \tag{10}$$

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 transforms the problem in Definition 4 into a matrix completion problem, where we have an estimated CATE matrix $\hat{\mathcal{T}}^{\text{incomplete}}$ and each element $[i, j]$ is defined as follows:

$$\hat{\mathcal{T}}_{i,j}^{\text{incomplete}} = \begin{cases} \hat{\tau}^{(j)}(X_i) & \text{if} F_{i,j} = 1 \\ ? & \text{if} F_{i,j} = 0 \end{cases} \tag{11}$$

As shown in Equation (11), 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 use existing 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 $\hat{\mathcal{T}}^{\text{complete}}$.[5] Algorithm 1 presents the details of our proposed approach.

[5] In Appendix C.1, we present the details on the SoftImpute algorithm proposed by Mazumder et al. (2010) as the algorithm we use for matrix completion in our applications.

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
$\hat{\mathcal{T}}^{\text{complete}} \leftarrow$ ***Complete***$(\hat{\mathcal{T}}^{\text{incomplete}})$

The output of this algorithm is a complete matrix $\hat{\mathcal{T}}^{\text{complete}}$ where all the elements are 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:

$$\hat{\tau}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathcal{T}}_{i,j}^{\text{complete}}. \tag{12}$$

If the matrix $\mathcal{T}$ is low-rank, $\hat{\tau}^{(j)}$ is a bias-corrected version of the ATE for treatment j. Further, we can use the imputed CATE estimates for targeting, as shown later in §5.3.3 and §6.3.2.

4.3 Assumptions and Identification

We now discuss the assumptions we need for the matrix completion approach to impute the missing entries in the CATE matrix. At a high level, our identification claim is that for each individual 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 individual. The following example helps illustrate the intuition behind our identification. Suppose that the treatment assignment in study j is deterministic for user i. As such, the CATE for this entry ($\tau_i^{(j)}$) cannot be identified using the data for study j. For each missing entry (i, j), there are some neighboring individuals (rows) i' in the same study (column) and neighboring studies (columns) j' in the same row that are non-missing. Hence, the ability of the algorithm to impute the missing entry depends on whether the information in the sub-matrix containing these neighboring individuals (rows) and studies (columns) can correctly impute the missing entry. Therefore, for this method to work, we need a systematic way to capture the similarities in the space of treatments. This is why

we use a matrix completion approach that has been widely used for collaborative filtering.

In our setting, we have an incomplete and noisy version of the true CATE matrix. The entries are noisy because the estimated CATE will have some errors. The identification task at hand is to identify the complete CATE matrix and estimate ATEs. To perform this task with standard matrix completion algorithms, we need assumptions on (1) the rank of the matrix, (2) the missingness pattern[6], and (3) noise in the observed entries. In the following parts, we present details on each assumption and discuss what they intuitively mean, when they are satisfied, and how we can test them.

4.3.1 Low Rank CATE Matrix

The fundamental assumption matrix completion methods require is the low-rank assumption, presented as follows:

Assumption 1. *The underlying CATE matrix $\mathcal{T}$ is low-rank; that is, for $R \ll \min(N, J)$, there exist two matrices $P_{N\times R}$ and $Q_{J\times R}$ such that $\mathcal{T} = PQ^T$.*

At a very high level, this assumption suggests that the user response exhibits some common patterns across different treatments. More specifically, Assumption 1 implies that CATE values across studies come from a linear combination of a few factors that are defined at the individual level. In that sense, this assumption is close to those commonly made in the structural economics literature that imposes a micro-foundation that allows only a few factors to drive user behavior. For example, in promotional treatments, we expect a few structural parameters to determine most of the treatment effects, such as users' price sensitivity, search cost, etc. We illustrate this insight formally in the following equation:

$$\mathcal{T} = \begin{bmatrix} \overbrace{ps(X_1)}^{\text{price sensitivity}} & \overbrace{sc(X_1)}^{\text{search cost}} & \dots \\ ps(X_2) & sc(X_2) & \dots \\ \vdots & \vdots & \ddots \\ ps(X_N) & sc(X_N) & \dots \end{bmatrix} \times \begin{bmatrix} \overbrace{w_1^{(1)}}^{\text{Study 1 Weights}} & \overbrace{w_1^{(2)}}^{\text{Study 2 Weights}} & \dots & \overbrace{w_1^{(J)}}^{\text{Study J Weights}} \\ w_2^{(1)} & w_2^{(2)} & \dots & w_2^{(J)} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix},$$

where factors include individual-level primitives such as price sensitivity and search cost that can be any complex function of covariates, and the linear weights determine how much these factors matter in driving the treatment effect for each study.[7]

[6] We use the missingness pattern interchangeably with the overlap violation, as we treat the entries with overlap violation as missing in our solution concept.

[7] It is worth emphasizing that the CATE in each study being a linear function of a few factors is not in contrast with the fact that CATE estimators are often designed to flexibly capture the underlying relationship between covariates and

When is low-rank assumption more reasonable? In general, a greater commonality in the structure of different studies makes the low-rank assumption more suitable. For example, suppose one is interested in how much each user likes a movie recommendation. In that case, it is reasonable to assume that a few factors can largely explain the variation in users' taste in movies, as is commonly assumed in recommender systems. However, if studies are completely unrelated, the low-rank assumption will be less realistic. In other words, more than only a few factors determine the treatment effects across all studies. However, the homogeneity of studies is a condition that is likely satisfied in most digital platforms as interventions likely share some common characteristics. In Appendix C.2, we present more structural reasons why this assumption is widely applied in practice, particularly in recommendation systems. Later in §5.1, we discuss why this assumption likely holds in our application context: online advertising.

Can we test the low-rank assumption? A key advantage of using low-rank methods is that it is testable. To perform this test, one could use the estimated CATE matrix with missing entries. The procedure is simple. The first step is to split the observed entries into training and validation. We can then use a matrix completion algorithm (e.g., SoftImpute) to complete the matrix using any rank r. We can then choose the best-performing low-rank approximation of the matrix on the validation set. Importantly, we can verify if the underlying matrix is not low-rank. Further, it is important to note that even if the underlying matrix is not low-rank, our method does as well as the existing approaches. Later in §6, we present an empirical application in the context of mobile in-app advertising and validate the low-rank assumption in that context.

4.3.2 Missingness Patterns

The second set of assumptions for matrix completion to work relates to the missingness pattern. In our setting, feasibility matrix F produces the missingness pattern in the CATE matrix. Most of the prior theoretical literature on matrix completion assumes fully random missingness to derive theoretical results on the recovery of the matrix (Candès and Recht 2009, Mazumder et al. 2010, Chen et al. 2019). More recent papers extend these theoretical results to specific non-random missingness patterns (Ma and Chen 2019, Athey et al. 2021, Agarwal et al. 2021). In general, a common factor in all this literature is to assume that the missingness pattern does not affect the identification of factors. We present the following informal assumption and refer the reader to Agarwal et al. (2021) for formal details we need for the missingness pattern:

Assumption 2. *For each missing entry in the CATE matrix, there are enough neighboring rows and columns in the feasibility matrix F with observed entries to identify the factors.*

treatment effects. The factors can still be very complex functions of user characteristics that need flexible learners to be identified.

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 row, there is no way to recover the parameters for that row. This issue may arise in settings where a user is very responsive to interventions and deterministically receives the intervention across all studies. However, even in such cases, one could verify whether there are such missingness patterns. In our calibrated simulation and empirical validation exercises in §5 and §6, we consider realistic and challenging missingness patterns to provide validity to this assumption.

Can we test the missingness assumption? A simple way to test whether this assumption is reasonable is to use the feasibility matrix for a simulated low-rank approximation. That is, one could simulate the underlying matrix using a low-rank assumption and induce missingness according to matrix F. If the algorithm accurately completes the matrix by identifying the low-rank factors, this provides evidence supporting Assumption 2.

4.3.3 Noise in Estimated CATE Matrix

Finally, since our task at hand is completing a noisy matrix, we must impose some structure on the noise added to entries. In general, we have $\mathcal{T} = \hat{\mathcal{T}} + E$, where E is the error in CATE estimates. We impose the following assumption on the noise in the CATE matrix:

Assumption 3. *The error E in the matrix is independent of the underlying missingness pattern F.*

This assumption suggests that there is no systematic error in CATE estimates that is correlated with the missingness pattern. For example, if our CATE estimates are upward biased for the feasible region, this would bias the estimates from our matrix completion approach. It is important to note that this assumption is satisfied as long as the CATE estimate is unbiased and consistent for the feasible entries. We are not worried about this assumption since we use unbiased CATE estimators like Causal Forests in our applications.

5 Application: Online Advertising

We focus on online advertising as a prominent application case for our proposed algorithm. Estimating ad effectiveness at the population or individual level has been a longstanding goal in both research and practice. However, this task is challenging because individuals' assignment to ads is not random. Advertising platforms use auctions in conjunction with advanced algorithms to allocate ads. Given that ad revenues account for the majority of total revenues for major advertising platforms such as Meta or Google, these platforms are often reluctant to run experiments or induce large-scale randomization in their auctions, as it is well-known that randomized allocation reduces

auction revenues (Myerson 1981). Thus, using observational methods to recover ad effectiveness is of great value to all parties involved, such as the advertising platform and advertisers.

In this section, we first define the CATE matrix in §5.1, which is our target estimand. We then describe the algorithmic allocation in common advertising auctions and show the violation of overlap as a result of this algorithmic allocation in §5.2. Next, we present results from our calibrated simulations in §5.3 and demonstrate the relationship between CATE and overlap violation, the performance of our algorithm in terms of accuracy compared to existing benchmarks, and the economic gains from decision-making based on our algorithm.

5.1 Estimation Target: CATE Matrix

We start by defining the estimation target in the advertising application. Definition 4 presents a general characterization of the problem. In this section, we want to define all elements in that definition for the online advertising application. In this application, there are N users indexed by i, and each study corresponds to an advertising campaign indexed by a, with a total of A campaigns. The treatment variable $W_i^{(a)}$ indicates whether or not user i is exposed to ad a. Each ad campaign a defines a conversion outcome, which could be a click, app install, website visit, or actual purchase of the advertised product, depending on the campaign objective. The outcome of interest $Y_i^{(a)}$ is user i's conversion outcome for ad campaign a. Each user has a vector of characteristics X_i that are user-level characteristics used for targeting.[8] For each pair of unit i and ad campaign a, we can formulate the Conditional Average Treatment Effect (CATE) as follows:

$$\tau^{(a)}(X_i) = \mathbb{E}[Y_i^{(a)}(1) - Y_i^{(a)}(0) \mid X_i], \tag{13}$$

which helps us define the CATE matrix $\mathcal{T}_{[N\times A]}$. The CATE matrix is the ultimate target for advertising platforms and advertisers as it allows them to target ads at the individual level. It further enables estimating Average Treatment Effect (ATE) for each ad, which is often a key objective for the platform, advertisers, and researchers (Lewis et al. 2011, Gordon et al. 2019, 2022).

Our algorithm requires an important low-rank assumption on the CATE matrix. We now justify why this assumption is reasonable for our application setting. First, it is useful to consider the opposite case where the CATE matrix is full-rank. Intuitively, it means that the individual-level treatment effect for each ad is independent of those for all other ads. That is, the treatment effect for each ad tells us nothing about the treatment effects for other ads. However, we expect the treatment effect for one ad to be informative about another ad. For example, if a smartwatch ad has

[8] One could easily change the unit of observation to a targeting profile rather than a user, characterized by the mixture of all covariates. This allows for changes in the targeting profiles of a certain user. We use this approach in our empirical application in §6.

a high treatment effect on a user, we expect the ad for a mobile health app to have a high treatment effect on that user. Similarly, if an ad for a right-wing news channel is effective for a user, we expect the ad for a left-wing news channel to be ineffective. These are just simple examples where the CATE from one ad is likely informative about that for another ad, thereby violating a full-rank assumption. Secondly, the prevalence of using matrix factorization models for ad targeting and click-through rate prediction tasks offers field evidence for the validity of the low-rank assumption in this setting (Menon et al. 2011, Juan et al. 2016, Choi et al. 2020). Later in §6, we validate this assumption in the context of mobile in-app advertising.

5.2 Algorithmic Allocation through Auctions

As discussed earlier, obtaining the CATE matrix is the ultimate goal for advertising platforms, advertisers, and researchers. However, estimating this matrix is challenging because the assignment of users to ads is not random. What determines a user's assignment to an ad is the auction run by the advertising platform. In this section, we describe the ad allocation process in the most commonly used advertising auctions and characterize propensity scores in our application.

We define $\pi^{(a)}(X_i)$ as the propensity score for ad a to be shown to user i, which is a function of users' observable characteristics X_i. Suppose there is an impression opportunity for user i. We index these impression opportunities by t. The advertising platform runs an auction to allocate an ad to this impression. There are a total of A candidates to participate in the auction for this impression. For computational reasons, most major advertising platforms include a randomly drawn subset of size A_r from all A ads to compete in the auction for the focal impression (Gordon et al. 2022).[9] For the set of participating ads, the auction requests bids from advertisers to award the impression to the one with the highest bid, as in both second- and first-price auctions, which are the most commonly used auctions by advertising platforms. Let $\mathcal{A}$ denote the full set of ads and $\mathcal{A}_{i,t}^{(r)}$ denote the subset of size A_r from these ads selected to participate in the auction for impression t of user i. Let $b_{i,t,a}$ denote the bid submitted by advertiser a in impression t of user i. The winning ad for this impression, denoted by $a_{i,t}^*$, will be determined as follows:

$$a_{i,t}^* = \arg\max_{a \in \mathcal{A}_{i,t}^{(r)}} b_{i,t,a} \tag{14}$$

We now link this to the treatment assignment for each ad a. Let T_i denote the total number of impression opportunities for user i. If ad a is selected in at least one of T_i impressions shown to

[9] This random sub-sampling from the full set of candidates is a source of randomization in auctions. Even if the platform does not use this direct form of bidder sub-sampling, the number of bidders participating in an auction is a fraction of all bidders because of reasons such as budget exhaustion and budget pacing, some of which are exploited as sources of random variation in ad exposure (Gui et al. 2021).

user i, then we have $W_i^{(a)} = 1$. Therefore, given T_i and bids submitted by advertisers, we can calculate the propensity score $\pi^{(a)}(X_i)$ for each user i. In particular, for fixed bid profiles for each user, the following lemma offers a closed-form relationship for propensity scores:

Lemma 2. *Suppose that each bidder a submits a bid $b_{i,a}$ for each one of user i's impressions, such that $b_{i,1} \leq b_{i,2} \leq \cdots \leq b_{i,A}$, without loss of generality. For user i, ad a's propensity score is determined as follows:*

$$\pi_i^{(a)} = 1 - \left(1 - \mathbb{1}(a \geq A_r) \frac{A_r \binom{a-1}{A_r-1}}{A \binom{A-1}{A_r-1}}\right)^{T_i} \tag{15}$$

Proof. See Appendix B.3. □

An immediate corollary of Lemma 2 is that the overlap assumption is violated because the propensity score is equal to zero if $a < A_r$, which is the deterministic no-assignment region. Likewise, for higher bids, the probability quickly converges to one as T_i increases, creating a deterministic assignment region. Now, we ask whether this overlap violation is consequential for observational methods that aim to estimate Average Treatment Effects (ATE).

To answer this question, we focus on advertisers' bids as a key factor influencing the propensity scores and deterministic regions. Theoretically, advertisers' bids are functions of how much they value an impression. For example, in a second-price auction, theory suggests that advertisers bid their valuations. The value of an impression is closely tied to the treatment effect of an ad for the user, known as CATE. Although advertisers do not necessarily know the true CATE for a user, it is reasonable to assume that they have an imperfect version of this signal based on data and modeling tools they have in place (Waisman et al. 2019). Together, for each ad, a lower CATE for a user leads to a lower bid submitted by the ad, which, in turn, leads to a higher possibility of belonging to the deterministic no-assignment region. Therefore, we expect the overlap violation to bias the estimates for population ATE in this context.

Finally, from a practical point-of-view, one could argue that if CATE values are low for the deterministic no-assignment region, not identifying those values may not be important for advertisers or platforms as they are looking for users with higher CATE. However, it is important to note that advertisers do not have a perfect CATE estimate for each user. As a result, there can be numerous high-CATE users within the group with deterministic no-assignment, leading to missed opportunities for both advertisers and platforms. We demonstrate these missed opportunities in our results in §5.3.3 and quantify the economic gains from identifying CATEs through our algorithm.

5.3 Results from Calibrated Simulations

We present the results from calibrated simulations in the online advertising context. In particular, we calibrate three important details in our simulations. First, we use the deciles for lift estimates from randomized controlled trials presented in Table 4 of Gordon et al. (2022) to set the ATE for our studies. Second, we assume an underlying low-rank CATE matrix consistent with applications of matrix factorization models in ad targeting.[10] Third, we use the micro-founded algorithmic ad allocation presented in §5.2 to generate ad assignments that mimic reality. Calibrating the details of our application setting ensures that the challenges imposed on our algorithm and benchmarks are realistic. We carry out large-scale simulations with $N = 100,000$ users and $A = 100$ ad campaigns and present all the simulation details in Appendix D.1.

In this section, we first present results on the overlap violation in our application context in §5.3.1. Next, in §5.3.2, we present the results on the performance of our algorithm compared to existing benchmarks in terms of estimation accuracy. Finally, in §5.3.3, we demonstrate how the increased estimation accuracy from our algorithm translates into economic gains for the advertising platform and advertisers.

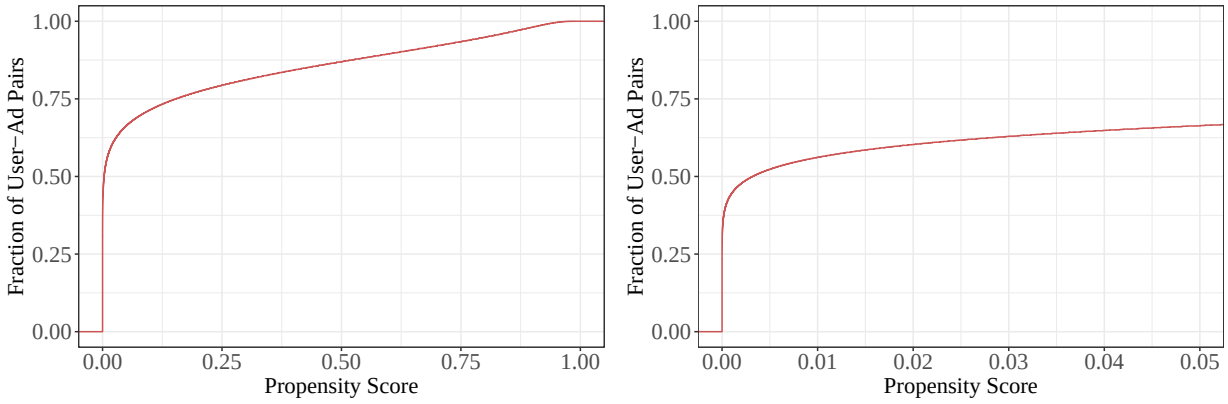
5.3.1 Overlap Violation

We now show the results from our simulation to examine the extent to which the algorithmic ad allocation violates the overlap assumption. As discussed earlier, advertisers' bid for a user $b_{i,a}$ is a function of their own estimate of CATE, denoted by $\tilde{\tau}^{(a)}(X_i)$. It is worth noting that $\tilde{\tau}^{(a)}(X_i)$ is not a fully calibrated estimate of the true CATE $\tau^{(a)}(X_i)$ because of issues such as modeling error or X_i observability, but it is highly correlated with it. In our simulation, we assume a correlation of 0.5 between $\tilde{\tau}^{(a)}(X_i)$ and $\tau^{(a)}(X_i)$. We consider a second-price auction, so we use the conventional assumption that bidders submit their valuation as bid: $b_{i,a} = \tilde{\tau}^{(a)}(X_i)$ (Waisman et al. 2019).[11] For each user, we simulate the number of impression opportunities T_i as a random draw from a Poisson distribution with parameter 25. We set $A_r = 10$, which means that the platform randomly draws 10 bidders in each impression to participate in the auction. We then run the auctions for all impressions to generate the data and determine the propensity scores.

Figure 3 visualizes the empirical Cumulative Density Function (CDF) for the propensity scores for all pairs of users and ads. Theoretically, we have established that for $(A_r - 1)/A = 0.09$ fraction of all pairs must have a propensity score exactly equal to zero. However, as shown in Figure 3a,

[10] The CATE matrix being low-rank is the fundamental assumption our proposed algorithm requires. We impose this assumption in our calibrated simulation to demonstrate the performance of our proposed algorithm. Later, in §6, we relax this assumption to test the validity of this assumption in an empirical setting.

[11] One could easily relax this assumption to allow for other factors to affect the advertiser's bid and to extend to cases for other auction such as first-price auctions. To the extent that bid is a function of valuation, the positive association between the true CATEs and bids will remain.



(a) All Propensities on x-axis

(b) Small Propensities on x-axis

Figure 3. Empirical CDF of the propensity scores for user-ad pairs in online advertising application.

the vast majority of user-ad pairs have a very small propensity score. In particular, when we zoom into the x-axis in Figure 3b, we find that the propensity score is lower than 0.01 for over 50% of user-ad pairs. These instances practically violate the overlap assumption as it is extremely hard for a model to estimate CATE for an observation with such low propensity scores. Our results show that the propensity scores produced in the online advertising setting violate the overlap assumption and pose challenges to algorithms that aim to estimate treatment effects using observational data. In Appendix D.2, we present the CATE for different regions of the data in each study to examine whether the CATE for the overlap-violating region is systematically different from that for the probabilistic region.

5.3.2 Performance of the Proposed Algorithm

We now examine the performance of our proposed algorithm in overcoming the challenges posed by the overlap violation relative to the existing benchmarks. In particular, we examine how accurately each method estimates ATE and CATE, as we have the underlying oracle ATE and CATE values. We start by comparing the performance of our model with the Double ML method in recovering ATE and show the patterns in Figure 4. As shown in this figure, the Double ML approach is not successful in recovering the true ATE. Combining our theoretical propositions in §3 and the discussion in the previous section on the overlap violation problem in our application setting, we argue that this is because there are many observations in the deterministic no-assignment with a lower CATE, on average. As a result, ignoring those points leads to overestimating the ATE. On the other hand, we find that our proposed algorithm performs well despite the presence of overlap violation. This is because our algorithm attempts to recover the full distribution of CATEs by systematically using estimable CATE in other distributions and using the collective information to

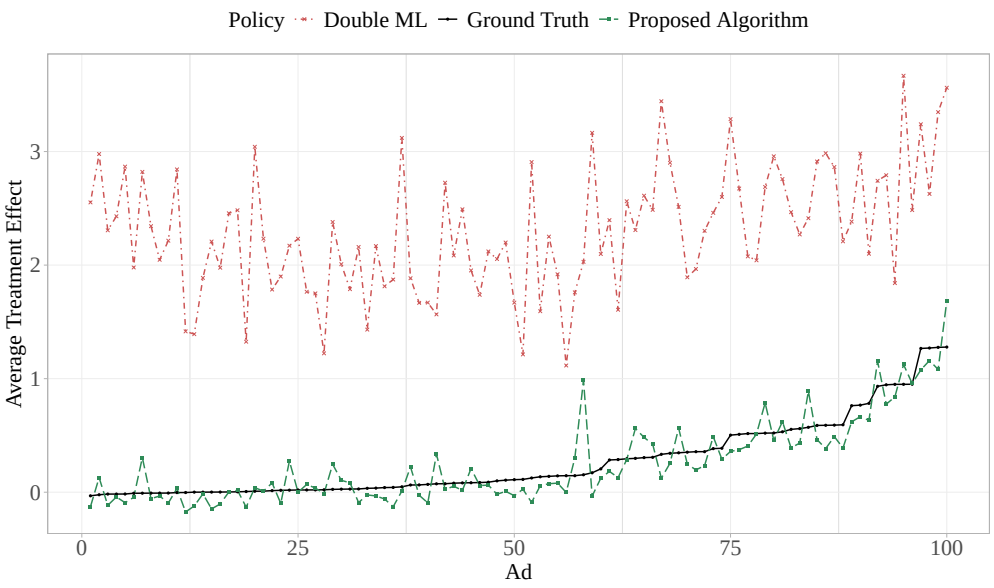


Figure 4. The performance of the proposed algorithm and Double ML in recovering ATE.

impute the missing values of CATEs in the overlap-violating regions.

We then extend our analysis in two important directions. First, we present the performance of different methods in terms of Root Mean Squared Error (RMSE) using the oracle ATE and CATE as the true targets. We aggregate these performance metrics over all ads in our study. Second, we consider a richer set of models, including different versions of our algorithm with mis-specified rank and estimated propensity scores. We present the results of this practice in Table 1. A few important insights emerge from the results in this table. First, our proposed algorithm performs better than Double ML in recovering ATE and CATE. For ATE, the results in the first two rows show a numerical equivalent of Figure 4. Our results are not surprising for CATE, given that benchmarks like DML cannot estimate CATE. For that reason, in §5.3.3, we propose better ways to evaluate the CATE performance of our algorithm in targeting and personalization and compare it with stronger benchmarks.

In the second part of Table 1, we focus on one of the key specifications of our algorithm: rank. The underlying rank for the CATE matrix is 10. We note that our cross-validation procedure described in §4.3.1 correctly identifies this rank without the specific knowledge of the true underlying rank. However, we still want to see how the algorithm would perform with different rank specifications. We focus on two mis-specified rank cases: (1) lower than optimal rank ($= 5$) and (2) higher than optimal rank ($= 20$). Intuitively, the model with a lower-than-optimal rank is often too simple, which leads to biased estimates of entries in the matrix, whereas the model with a higher-than-optimal rank can be too complex, which leads to higher variance. In Table 1, we see a worse performance by models with mis-specified ranks. However, we note that both models still

Method	RMSE for ATE	RMSE for CATE
Algorithm	0.163	1.761
Double ML	2.079	3.351
Algorithm (Rank = 5)	1.443	2.679
Algorithm (Rank = 20)	0.247	1.827
Algorithm (Incorrectly Estimated Propensities)	1.313	2.449
Algorithm (Correctly Estimated Propensities)	0.614	1.133

Table 1. Performance of our algorithm and benchmarks in recovering oracle ATE and CATE.

outperform the Double ML benchmark, implying that the algorithm is able to exploit the cross-ad variation even with the mis-specified rank.

Finally, we focus on the case where propensity scores are not known but need to be estimated and consider two separate cases: (1) incorrectly estimated propensity scores, where only a subset of important covariates determining propensity scores is used, and (2) correctly estimated propensity scores, where all covariates are used to ensure that we have a calibrated propensity model. A few interesting results emerge from our analysis. First, as expected, we find that the model with incorrectly estimated propensity scores performs worse than the one with correctly estimated propensity scores in all metrics. Second, we find that the algorithm with correctly estimated propensity scores performs worse in terms of ATE but better in terms of CATE. Further investigation shows that the algorithm with correctly estimated propensities has a higher bias but lower variance than the model with known propensity scores, which explains its relative performance in recovering ATE and CATE. The trade-off suggests room for improvement by developing hybrid approaches based on known and estimated propensity scores to achieve the right balance in the bias-variance trade-off.

5.3.3 Economic Gains from the Proposed Algorithm

As discussed earlier, the overlap problem exists because observations in the deterministic no-assignment region have lower CATE, on average. In our application setting, the overlap violation for a user-ad pair is directly linked to the low value of that user for the advertiser. Hence, one could argue that although our algorithm helps recover the ATE, it does not offer any practical value to firms (e.g., advertising platforms and advertisers) who want to use this information for decision-making because they would just ignore the overlap-violating due to its lower CATE on average. In this section, we examine the economic gains from decision-making based on our algorithm and ask the following questions: Are there any consequences in ignoring the overlap-violating part of the data for the firm that wants to target? To what extent could using our algorithm improve outcomes

for advertising platforms and advertisers?

To answer these questions, we turn to a personalization exercise where we compare the performance of different algorithms in selecting the targeting population. This allows us to measure outcomes under targeting based on different algorithms and examine the real economic gains from our proposed algorithm. We take the allocation in the data as the benchmark and examine how each model re-allocates users to ads while keeping the total number of treated users the same. If ignoring the overlap-violating region is not consequential for real targeting decisions, we should not observe any meaningful differences in the outcome. We perform this exercise to test whether there are any gains from our algorithm.

Let K_a denote the number of users in the data exposed to ad a, that is, $K_a = \sum_{i=1}^{N} W_i^{(a)}$. In our exercise, we want to see how each model re-allocates these K_a treated users. For example, our algorithm may change the allocation based on the CATE estimates obtained. We then measure the gains as the difference between the targeted policy and the baseline where no one is exposed to the ad. In particular, we are interested in comparing the targeting performance of our proposed algorithm two policies: (1) *Data*, where the allocation is based on submitted bids and the underlying auction environment, and (2) *Bids*, where advertisers are allowed to target K_a users for whom they have the highest bids. Comparing the targeting performance of our algorithm with these two benchmarks helps us quantify the economic gains from our algorithm.

We present the results of our exercise in Table 2. The first column represents the average gains from each targeting policy and the second column presents the ratio of average gain to the targeting model based on oracle CATE values, i.e., the first-best performance. Interestingly, we find that the average gains from our model are substantially higher than those based on bids and the algorithmic allocation in the data. In other words, by not ignoring the overlap-violating region in the data, our model is able to generate almost double the targeting effectiveness. This finding implies that there are numerous targeting opportunities in the overlap-violating region despite it having a lower CATE on average. Compared with the first-best oracle performance, we find that our algorithm recovers almost 97% of that performance. We further examine the performance of alternative specifications of our algorithm, where rank is mis-specified or propensity scores are not known. Interestingly, we find that in all these instances, the algorithm still outperforms the two benchmarks based on bids and data.

6 Empirical Validation Exercise

In this section, we work with real data from mobile in-app advertising and present empirical validation for our algorithm. In particular, we seek to provide validity to the low-rank assumption needed for our matrix completion algorithm. Therefore, we need a setting with clearly defined CATE es-

Method	Average Gains	Oracle Ratio
Algorithm	4.183	0.967
Bid	2.295	0.532
Data	2.011	0.466
Algorithm (Rank = 5)	3.453	0.800
Algorithm (Rank = 20)	4.119	0.955
Algorithm (Incorrectly Estimated Propensities)	2.812	0.652
Algorithm (Correctly Estimated Propensities)	4.153	0.963

Table 2. Economic gains from targeting based on different models.

timands across multiple studies that are identifiable from the data so we can correctly recover the underlying CATE matrix and test whether it is low-rank. In this section, we present our setting and data in §6.1. Next, in §6.2, we present the details of our empirical framework, target estimand, and our identification strategy to estimate this target. Finally, we present validation results in §6.3.

6.1 Setting and Data

The empirical setting of our problem is mobile in-app advertising, an industry that has exhibited sustainable growth over the past decade. We use impression-level data from a leading mobile in-app advertising network from a large Asian country with over 85% market share around the time of this study. We observe over a billion ad impressions in this data set. The sample we use is identical to Rafieian (2023), so we refer the reader to that paper for details on the sampling. In this sample, we observe 6,357,389 impressions from 327 distinct ads. For each impression in this sample, the covariates we observe are demographic features that contain the province, latitude, longitude, smartphone brand, mobile service provider (MSP), and connectivity type, as well as a number of historical and session-level features that are defined based on the past short- and long-term history of the user (e.g., the variety of past ads seen, number of impressions).

Besides the scale and richness of the data, a few key features of our setting and data make it ideal for our study. First, unlike the standard practice in advertising auctions that produce limited randomization in ad allocation, this platform uses a quasi-proportional auction wherein each bidder has a probability of winning proportional to each advertiser's quality-adjusted bid. That is, if ad a's quality-adjusted bid is q_a, its probability of winning in an auction is $q_a / \sum_{j \in \mathcal{A}} q_j$, where $\mathcal{A}$ is the set of participating ads in that auction. Second, the targeting provision for advertisers is limited such that they can only target ads based on broad targeting categories that are all observable to the researcher. Therefore, as shown formally in Proposition 1 of Rafieian (2023), observed covariates fully determine the distribution of propensity scores.

6.2 Empirical Framework

6.2.1 Effect of Focal Ad vs. Native Ad

We start by defining the causal effect of interest. In our setting, each impression is characterized with a targeting profile X_i, and is allocated to an ad from the set of participating ads in that auction, denoted by $\mathcal{A}_i$. We are interested in the causal effect of showing a focal ad a^* as the treatment relative to a native ad $a^{(n)}$ that the platform can serve at any time.[12] As such, the assignment to the focal ad in each study is the treatment condition ($W_i^{(*)} = 1$) and the control condition represents the assignment to the native ad ($W_i^{(*)} = 0$). Let $Y_i^{(*)}(w)$ denote the potential click outcome for the targeting profile X_i upon receiving condition $w \in \{0, 1\}$. We can define the CATE for ad a for targeting profile X_i as follows:

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

In our empirical analysis, we are interested in recovering the parameter presented in Equation (16) and present results that show our ability to recover this parameter in different settings. A few points are worth noting in defining our target estimand. First, our unit of population is a targeting profile rather than a user. This implies that one user can be represented in targeting profiles, which is more consistent with the approach in contextual bandits literature where individual contexts arrive, and one user's context can evolve. Second, the click outcome that we focus on in this exercise is a key conversion metric because ads are all mobile apps, and the click is very closely linked with the app install action. Further, the platform runs a pay-per-click auction and each click directly adds to the platform revenues. Third, we use the causal estimand above primarily because it is well-defined, but we note that the comparison with the native ad is a reasonable benchmark for the platform to evaluate the engagement created by an ad.

6.2.2 Definition of CATE Matrix

The causal effect of the focal ad a^* relative to the native ad $a^{(n)}$ constitutes only one study. For our algorithm, we need to define a CATE matrix that helps with the estimation of the causal effects in our focal study. To define this CATE matrix, we define the same causal effect as in Equation (16) for other ads in our inventory. That is, for any ad $a \in \mathcal{A}$, we define the following CATE:

$$\tau^{(a)}(X_i) = \mathbb{E}[Y_i^{(a)}(1) - Y_i^{(a)}(0) \mid X_i], \tag{17}$$

[12] It is worth emphasizing that in the data, the native ad participates in an auction and has no advantage over other ads. However, it never runs out of budget, which is why we focus on it as the control condition.

where $Y_i^{(a)}(1)$ and $Y_i^{(a)}(0)$ are potential outcomes for cases where ad a and native ad $a^{(n)}$ are shown in the impression with targeting profile X_i. This allows us to form the CATE matrix, where rows are different targeting profiles and columns are different ads in our data. The low-rank assumption in this setting indicates that the information in CATE from other ads could inform us about the CATE for the focal ad.

6.2.3 Identification Strategy

We now discuss our identification strategy for estimating CATE for all targeting profiles across ads. For each study, two ads can be shown in the impression: ad a and native ad $a^{(n)}$. We use the randomness induced by the quasi-proportional auction as our main identification strategy. Let q_a and q_n denote the quality-adjusted bids for the focal ad a and the native ad $a^{(n)}$ in a given impression. If both ads a and $a^{(n)}$ participate in the auction for that impression, their corresponding winning probabilities will be $q_a/(\sum_{j\in\mathcal{A}} q_j)$ and $q_n/(\sum_{j\in\mathcal{A}} q_j)$. This implies that the ratio only depends on q_a and q_n. Therefore, for the set of impressions where both ads a or $a^{(n)}$ participate and one of them wins, we have both unconfoundedness and overlap assumption satisfied because we know that the probability of treatment assignment (ad a) is $q_a/(q_a + q_n)$ and the probability of control assignment (native ad $a^{(n)}$) is $q_n/(q_a + q_n)$. Given the infrequent updating of bids by advertisers and quality scores by the platform, the proportions remain largely stable for any impression where both ads participate, and one wins the auction, which greatly stabilizes the estimation of causal parameters.

It is worth emphasizing that if one ad does not participate in the auction for an ad, the assignment probability will be zero, which violates the overlap assumption. Therefore, it is crucial for our sample construction to filter out impressions in which either ad does not participate. We know that there are only two reasons for an ad not participating in the auction for an impression: (1) the ad specifically excludes a targeting category in that impression (e.g., smartphone brand), or (2) the ad is not available due to budget exhaustion. As such, we use a two-step sampling approach for sample construction. First, we only focus on ads that do not exclude any targeting category. Second, when estimating CATE for those ads, we exclude the impressions won by the native ad (control condition) for which ad a was unavailable. This ensures that the sample we use to estimate CATE for each ad mimics a randomized experiment and satisfies both overlap and unconfoundedness.

6.2.4 Empirical Estimation of the Underlying CATE Matrix

We now discuss our approach to estimating the underlying CATE matrix. We present the steps in our empirical approach as follows:

- *Sampling Ads:* We first identify the set of ads that do not exclude any targeting category in

their targeting decisions. From all 327 ads in our data, 123 do not use targeting. Of these 123 ads, we focus on those that at least have 2000 impressions in our data, so we have reasonable statistical power in our estimation. This gives us a sample of 59 ads other than the native ad used as the control condition. These ads represent the columns in our CATE matrix.

- *Sampling Targeting Profiles:* Because we are interested in the causal effect of the focal ad, we draw a random sample of 50,000 targeting profiles in impressions allocated to either the focal ad or the native ad.[13] For each targeting profile, we observe a rich set of covariates, including demographic features such as location information as well as historical features. Please see a complete list of features used to characterize a targeting in Appendix E.1. These 50,000 impressions represent the rows in our CATE matrix.
- *CATE Estimation:* For each ad a, we first draw a sample of impressions allocated to either ad a or the native ad n. We know that the native ad was always available for all auctions, but each ad a may be unavailable for some auctions. Therefore, we drop the impressions allocated to the native ad for which the focal ad a did not participate in the auction. This ensures our sample for each study satisfies the assumptions needed to estimate CATE given our identification strategy in §6.2.3. We then use Causal Forest to estimate CATE for each ad, using a two-fold cross-validation to avoid overfitting (Athey et al. 2019). Once we have the CATE estimated, we estimate the CATE for the set of targeting profiles in each row of our CATE matrix. We then add these estimates to column a of our matrix. Repeating the process for each ad (study) recovers the CATE matrix.

It is worth emphasizing that the purpose of this exercise is to test the validity of our low-rank assumption in a real setting that does not impose a low-rank assumption. Consistent with this goal, we specifically use Causal Forest that take high-dimensional set of features and capture flexible relationships between these variables.

6.3 Validation Results

6.3.1 Rank of the CATE Matrix

In our calibrated simulation, we assume that the underlying CATE matrix is low-rank based on domain-specific justifications and the wide use of this assumption in practice, particularly in the context of ad targeting. One of the main purposes of our empirical validation exercise is to test this assumption in an actual setting. We can use Singular Value Decomposition (SVD) for any given rank and obtain a low-rank approximation of the estimated CATE matrix. We can then examine how well the low-rank approximation performs. We do this for rank one up to 30 and measure

[13] It is worth noting that the CATE parameters for other ads are well-defined and one could accurately estimate CATE for these targeting profiles from the data, even if their corresponding impressions are only awarded to the focal and native ads.

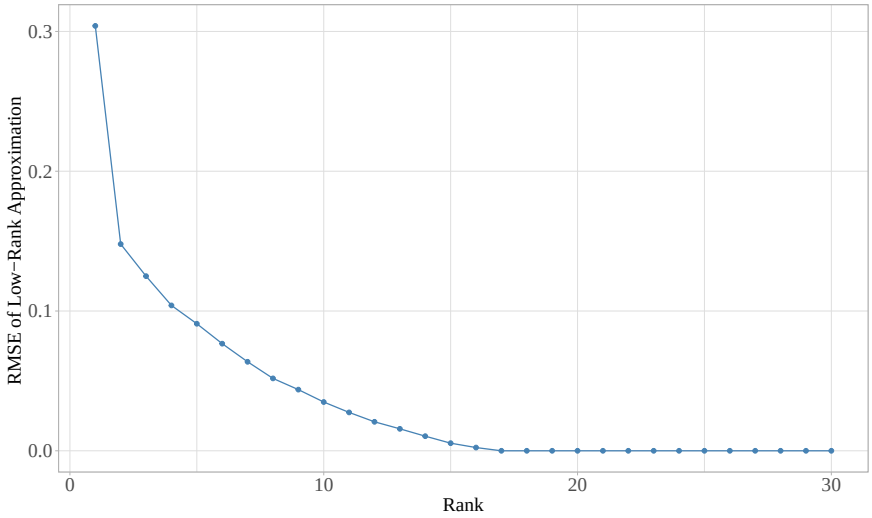


Figure 5. RMSE of low-rank approximation of the estimate CATE matrix using Singular Value Decomposition

the RMSE of the low-rank approximation. We present the resulting graph in Figure 5. This figure shows that the RMSE quickly reduces as we increase the rank for the low-rank approximation and almost becomes equal to zero at the rank-17 approximation. This finding validates our low-rank assumption in an actual empirical setting.

6.3.2 Evaluation of Our Algorithm

We now turn to the study for the focal ad and evaluate our algorithm to provide an empirical proof-of-concept. Our estimation results indicate that the Average Treatment Effect of the focal ad a^* relative to the native ad $a^{(n)}$ is -0.0102, which translates into a lift of -43.26%.[14] That is, on average, showing the focal ad leads to roughly 43% fewer clicks. In this section, we first present a simple case of algorithmic ad allocation that results in overlap violation. We then use our algorithm to recover ATE and CATE estimates and propose personalized policies. Since we have the true estimated CATEs from our data, we can examine the performance of our algorithm and quantify gains from targeting based on our algorithm.

We consider a simple case of algorithmic ad allocation between the two ads in our study: the focal ad a^* and the native ad $a^{(n)}$. In standard auctions used by advertising platforms, the platform estimates the CTR for each ad and requests bid-per-click from each advertiser. The auction then allocates the impression to the ad with the higher quality-adjusted bid, which is simply the product of the bid and the estimated CTR. Let $\widehat{\text{CTR}}_i^{(*)}$ and $\widehat{\text{CTR}}_i^{(n)}$ denote the platform's estimates for the focal ad and native ad in impression i, respectively. In our data, we note that both ads submit the

[14] The CTR for the native ad is 0.0237, so an ATE of -0.0102 translates into $-0.0102/0.0237 = -0.4326$ as lift.

same bid-per-click, so we make the assumption that they have the same bid in the new auction format.[15] As such, the impression will be allocated to the ad with the higher estimated CTR. We use XGBoost to estimate CTR for each ad in each impression i and measure the difference $\hat{\delta}_i = \widehat{\text{CTR}}_i^{(*)} - \widehat{\text{CTR}}_i^{(n)}$ between them. For the assignment, we assume that if $-0.005 < \hat{\delta}_i < 0.005$, the assignment will be probabilistic. Otherwise, the impression will be deterministically allocated to the ad with a higher estimated CTR. This assumption is reasonable because many advertising platforms use posterior sampling approaches, such as Thompson Sampling, for estimating CTR, where draws from the posterior distribution can lead to a probabilistic assignment.

In this simple setting with algorithmic allocation, we find that 42% of impressions are in the probabilistic region, and the rest are in the deterministic region. When measuring the CATE for the probabilistic region, we find an interesting sign flip: the CATE for the probabilistic region is 0.0057, which translates into a 24.21% lift. This means that all the state-of-the-art methods described in §3 will, at best, recover this parameter. Now, we apply our algorithm to impute CATE for the overlap-violating region. We consider four versions of our algorithm, based on the missingness of the CATE matrix for other studies: (1) *Incomplete for Low CATE*, (2) *Incomplete for High CATE*, (3) *Incomplete for High/Low CATE*, and (4) *Complete*. We present the details of these missingness patterns in Appendix E.2. The first three missingness patterns could happen if algorithmic ad allocation is used for other studies, whereas the complete CATE matrix for other studies is more consistent with the empirical exercise we have done, which is estimating CATE for other ads for the focal impressions and targeting profiles. We demonstrate the performance of these algorithms in their recovery of the true ATE and lift, and the CTR from their targeting policy. As in §5.3.3, we use the ad allocation algorithm described above as the benchmark to see if the algorithm already addresses the problem. We further use the oracle performance, which is the first-best performance achievable by any targeting model.

We present our results in Table 3. We first focus on columns on ATE and lift and find that all versions of our algorithm do a great job of recovering the true ATE and lift. As discussed earlier, we find that conventional methods to estimate ATE (e.g., DML) could even miss the direction of the ATE and lift, as they could only use the probabilistic region. Finally, we note that estimation based on the predictive ad allocation algorithm correctly identifies the sign but is biased by estimating a -77%. Next, we quantify gains from targeting based on each model. The targeting model in each case is simple: only if the estimated CATE is greater than zero, the impression is allocated to the focal ad. The last two columns of Table 3 show the CTR under each policy and measure the lift compared to the policy that allocates all impressions to the native ad. We find a near-optimal performance by our algorithm in targeting with around 36% improvement in CTR, which is sub-

[15] We acknowledge that this may not be true, but our goal here is to present a proof-of-concept in this section, not delivering a complete counterfactual analysis.

Method	ATE	Lift	Targeted CTR	Targeted Lift
Algorithm (Incomplete for High CATE)	−0.0114	−48.22%	0.0321	35.56%
Algorithm (Incomplete for Low CATE)	−0.0103	−43.30%	0.0322	36.14%
Algorithm (Incomplete for High/Low CATE)	−0.0108	−45.44%	0.0322	36.08%
Algorithm (Complete CATE)	−0.0106	−44.83%	0.0324	36.91%
Ad Allocation Algorithm ($\hat{\delta}$)	−0.0183	−77.41%	0.0249	5.12%
Conventional Methods (DML)	0.0057	24.21%	–	–
Oracle	−0.0102	−43.26%	0.0324	37.06%

Table 3. Performance of different models in terms of estimation and targeting.

stantially higher than the 5% improvement based on the ad allocation algorithm described above. This finding highlights a key insight: although algorithmic decision-making improves outcomes, there is further room for improvement that our algorithm can achieve. In particular, we find that the algorithm based on the complete CATE matrix for other ads performs better than the rest. This finding suggests that managers should use pre-existing experimental data to construct CATE factors to achieve better performance from our algorithm.

7 Managerial Implications

Our work has several implications for managers and practitioners. At a conceptual level, our algorithm is useful for any decision-maker who is interested in measuring the causal effect of interventions at the population or individual user level. Although the gold standard to estimate quantities is to run a randomized controlled trial, the cost of experimentation is often high, thereby making observational methods more appealing. Most settings with an algorithmic delivery of interventions fall under this category, as these are settings where one could manipulate the receipt of an intervention, but there are already algorithms in place to facilitate this assignment problem. For example, if using an algorithm for promotion assignment leads to desirable outcomes, a manager may view experimentation as a waste of resources. Similarly, in online advertising, the standard advertising auctions have remarkable revenue properties for the platform, so the platform is interested in avoiding costly experimentation if possible.

In our work, we argue that the use of algorithmic decision-making leads to a violation of the overlap assumption, which can prevent managers from estimating the causal parameters. Specifically, we propose an algorithm that helps managers achieve better estimates of important causal parameters without inducing too much randomization in their interventions. Our algorithm is applicable to a wide variety of cases where algorithmic decision-making is used. In particular, we describe two large classes of settings where our algorithm can create value:

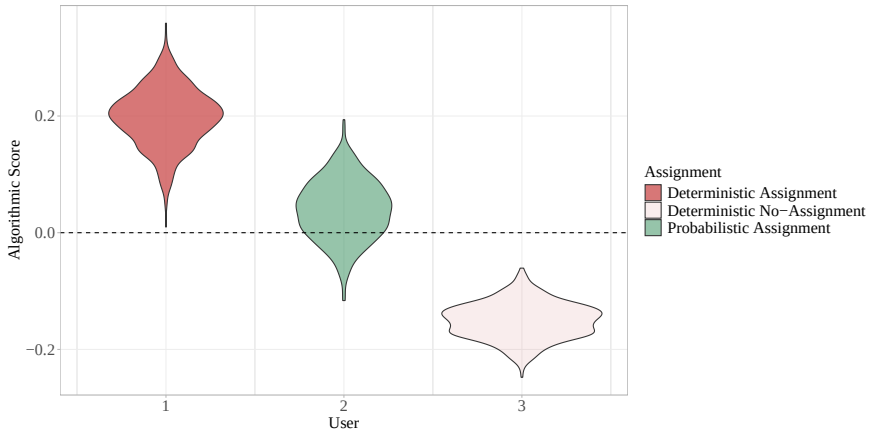


Figure 6. Assignment based on the posterior distribution of algorithmic score

- The first class is for settings where the platform has full control over the assignment policy. In these cases, algorithmic scores inform the treatment assignment. The intervention with the highest algorithmic score will be selected. This naturally creates a setting where sizable portions of the population are assigned to deterministic and probabilistic regions. The reason for the existence of probabilistic assignment in these settings is the residual uncertainty in the posterior distribution of these algorithmic scores. Figure 6 illustrates this point with a case where a threshold rule is used to assign individuals to an intervention: only if the posterior distribution of algorithmic score is entirely above or below the cutoff, the assignment will be deterministic. An important question is whether one would care about cases of deterministic assignment or no-assignment, if the algorithm is certain that they should or should not receive the treatment, respectively. The reason is that all models contain a certain degree of mis-specification that makes it possible for the existence of many high-CATE instances in the deterministic no-assignment region, and many low-CATE instances in the deterministic assignment region. Our algorithm helps re-estimate and re-prioritize these instances, which could lead to substantial gains as shown in §5.3.3. Practical examples of such settings are promotion assignment (e.g., Uber's promotion for future rides) and push notifications (e.g., Fitbit's notification on body activity).
- The second class is for settings where the platform controls the allocation rule, but other agents could also influence the assignment outcomes. Examples include an advertising platform trying to allocate an ad, or a social network trying to personalize the news feed using content generated by agents, or a platform determining the product ranking for a user query. In all these settings, the platform designs an algorithm that determines assignment based on the inputs from agent. Hence the platform does not fully control the assignment. These settings also introduce both probabilistic and deterministic regions. The source of randomization in these

cases are unexpected events related to the agents involved, e.g., an advertiser's budget runs out, or a new product enters a market and it shifts product rankings. Similarly, the algorithmic allocation rule can also introduce some randomization for computational reasons or exploration. We considered a similar setting in §5 and demonstrated that although randomization exists in users' allocation to ads, there is still a sizable overlap-violating region. Our algorithm can be useful in these settings as one could clearly set the scope of the CATE estimation and let the matrix completion algorithm determine the missing parts. To the extent that there is similarity in treatment effects on population units', our algorithm helps systematically capture the relationship between interventions and overcome the challenge posed by deterministic assignment. Like the first class of problems, the value of our algorithm comes from refining the target and personalized policy, which can lead to substantial gains.

Together, our algorithm offers an important tool for digital platforms and managers implementing algorithmic decision-making. In particular, our algorithm creates value for decision-makers in settings where experimentation is costly or infeasible, resulting in imperfect randomization. As such, it is more useful in the second class of problems described above, because it is easier to induce small-scale randomization in the first class of problems given the full control the platform has over the treatment assignment.

Lastly, we stress that our algorithm should not be seen as a replacement for experimentation and A/B testing. Rather, it can be used as a complement with the purpose of reducing experimentation costs. A very useful application of our algorithm is when a platform uses experimentation to build the CATE matrix and reliably estimate the underlying factors for it. Using our algorithm with the knowledge of the underlying factors will further allow the experimenter to reduce the experimentation cost. Even in settings where the cost of experimentation is relatively low, our algorithm enables practitioners to better use existing databases suitable for our application.

8 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 costs. 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 and 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 we expect the bias caused by the lack of overlap to be large in most digital platforms. 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. Across a series of calibrated simulations and empirical exercises, 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: the 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 of treatment effects is low-ranked. Our proposed algorithm can be used by digital platforms to utilize their existing observational data and by researchers who access a platform's data that suffer from the issue of deterministic assignment.

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.

H. Choi, C. F. Mela, S. R. Balseiro, and A. Leary. Online display advertising markets: A literature review and future directions. *Information Systems Research*, 31(2):556–575, 2020.

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

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

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

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

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

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

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

G. Gui, H. Nair, and F. Niu. Auction throttling and causal inference of online advertising effects. *arXiv preprint arXiv:2112.15155*, 2021.

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.

G. A. Johnson, R. A. Lewis, and E. I. Nubbemeyer. Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research*, 54(6):867–884, 2017a.

G. A. Johnson, R. A. Lewis, and D. H. Reiley. When less is more: Data and power in advertising experiments. *Marketing Science*, 36(1):43–53, 2017b.

Y. Juan, Y. Zhuang, W.-S. Chin, and C.-J. Lin. Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM conference on recommender systems*, pages 43–50, 2016.

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.

R. A. Lewis and J. M. Rao. The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973, 2015.

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

W. Ma and G. H. Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *Advances in neural information processing systems*, 32, 2019.

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.

A. K. Menon, K.-P. Chitrapura, S. Garg, D. Agarwal, and N. Kota. Response prediction using collaborative filtering with hierarchies and side-information. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 141–149, 2011.

R. B. Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

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.

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

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. *Marketing Science*, 42(5):910–933, 2023.

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

O. Rafieian, A. Kapoor, and A. Sharma. Multi-objective personalization of the length and skippability of video advertisements. *Available at SSRN 4394969*, 2023.

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.

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.

A. Shi, D. Zhang, T. Chan, H. Hu, and B. Zhao. Using algorithmic scores to measure the impacts of targeting promotional messages. *Available at SSRN*, 2022.

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.

M. Udell and A. Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

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

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

Online Appendix

A Overview of State-of-the-Art Approaches to Estimate ATE

A.1 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 estimates 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 causal inference in order to increase the flexibility and robustness of model-based methods to estimate ATE (Belloni et al. 2014, Chernozhukov et al. 2018a). Many of these methods are now considered state-of-the-art methods for estimating the ATE.

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:

$$Y_i(w) = \mu_0(X_i) + \tau^*(X_i)w + \epsilon_i(w), \tag{A.18}$$

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:

$$Y_i - m(X_i) = (W_i - \pi(X_i))\,\tau^*(X_i) + \epsilon_i(W_i), \tag{A.19}$$

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 estimate 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:

$$\tau^*(\cdot) = \underset{\tau}{\operatorname{argmin}}\, \mathbb{E}\left[\left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau(X_i)\right)^2\right]. \tag{A.20}$$

The double machine learning (DML) approach estimates both nuisance functions using machine learning models and then estimates 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.2 Model-free Approaches to Estimate ATE

We now 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 its inverse propensity score, which gives us the following estimator for the ATE:

$$\hat{\tau}_{\text{IPS}} = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i \left(\frac{W_i}{\pi(X_i)} - \frac{1-W_i}{1-\pi(X_i)}\right)\right), \tag{A.21}$$

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 have 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.

B Proofs

B.1 Proof of Proposition 1

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:

$$\underset{\tau}{\operatorname{argmin}} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \xrightarrow{p} \tau_r. \tag{A.22}$$

We now want to show that the RHS of Equation (A.22) is the same as what any methods optimizing Equation (A.20) would estimate. We can write:

$$\begin{aligned}
\hat{\tau} &= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&\quad + \sum_{i \notin \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 + \sum_{i \notin \mathcal{I}_r} \left(Y_i - m(X_i)\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2,
\end{aligned} \tag{A.23}$$

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 (A.23) with that of Equation (A.22), the proof is complete for DML.

We now turn to the IPS estimator. 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). □

B.2 Proof of Proposition 2

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 3. *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:

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)] &= \mathbb{E}[Q_i \tau(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Q_i \tau(X_i) \mid Q_i]] \\
&= \mathbb{E}[Q_i \mathbb{E}[\tau(X_i) \mid Q_i]] \\
&= P(Q_i = 1)(1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] + P(Q_i = 0)(0)\mathbb{E}[\tau(X_i) \mid Q_i = 0] \\
&= P(Q_i = 1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] \\
&= P(\pi(X_i) = 1)\mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]
\end{aligned} \tag{A.24}$$

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:

$$\begin{aligned}
\tau_1 &= \mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1] \\
&= \frac{P\left(\pi(X_i) = 1\right) \mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]}{P\left(\pi(X_i) = 1\right)} \\
&= \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&\geq \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]\mathbb{E}[\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&= \mathbb{E}[\tau(X_i)] \\
&= \tau^*,
\end{aligned} \tag{A.25}$$

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

B.3 Proof for Lemma 2

Proof. We first calculate the probability of each ad a winning an impression. For each ad a such that $a < A_r$, we use the *pigeonhole principle* and show that the probability of a winning an impression is zero, because there is always one ad among A_r selected ones with a higher bid that a. Now, if $a \geq A_r$, we first need a to be selected as one of A_r ads, which has a probability of A_r/A. Conditional on a being selected, the probability that a is the highest bid is the probability that all of $A_r - 1$ ads are selected from all $a - 1$ ones with bids lower than a. We now this probability is equal the number of combinations of $A_r - 1$ from $a - 1$, divided by all possible size $A_r - 1$ combinations from the remaining ads, which is the number of combinations of $A_r - 1$ from $A - 1$. As such, the probability of ad $a \geq A_r$ winning an impression is given as follows:

$$\left(\frac{A_r}{A}\right)\left(\frac{\binom{a-1}{A_r-1}}{\binom{A-1}{A_r-1}}\right).$$

Using the equation above, we can write the probability of any a winning as follows:

$$\Pr(a\text{wins an impression}) = \mathbb{1}(a \geq A_r)\frac{A_r\binom{a-1}{A_r-1}}{A\binom{A-1}{A_r-1}} \tag{A.26}$$

Now, we can calculate the probability of a winning at least one of T_i impressions. We can write:

$$\begin{aligned}
\Pr(a\text{wins at least one impression}) &= 1 - \Pr(a\text{wins no impression}) \\
&= 1 - (1 - \Pr(a\text{wins an impression}))^{T_i} \\
&= 1 - \left(1 - \mathbb{1}(a \geq A_r)\frac{A_r\binom{a-1}{A_r-1}}{A\binom{A-1}{A_r-1}}\right)^{T_i}
\end{aligned} \tag{A.27}$$

□

C Supplementary Materials for the Proposed Algorithm

C.1 SoftImpute Algorithm

The SoftImpute algorithm is a matrix completion technique that is widely used to fill in missing values in large datasets by exploiting low-rank structure in the data. The method was introduced as an efficient way to handle incomplete data by iteratively approximating the missing entries of the matrix while maintaining a low-rank approximation.

The algorithm is based on the concept of matrix factorization and is particularly useful when the underlying data matrix is assumed to have a low-rank structure, which means that much of the variation in the data can be captured by a few latent factors. SoftImpute achieves this by using singular value thresholding to shrink the singular values of the data matrix, thereby inducing a low-rank approximation. The algorithm is defined as follows:

Let $\mathbf{X} \in \mathbb{R}^{N \times J}$ be the data matrix with missing entries. The goal is to approximate $\mathbf{X}$ by a matrix $\mathbf{M}$ of lower rank such that the missing values are imputed in a way that preserves the structure of the original data.The optimization problem solved by SoftImpute can be formulated as:

$$\hat{\mathbf{M}} = \arg\min_{\mathbf{M}} \frac{1}{2} \|\mathbf{W} \odot (\mathbf{X} - \mathbf{M})\|_F^2 + \lambda \|\mathbf{M}\|_*, \tag{A.28}$$

where $\mathbf{W} \in \{0,1\}^{N \times J}$ is an indicator matrix, with $w_{ij} = 1$ if x_{ij} is observed and $w_{ij} = 0$ otherwise, $\odot$ denotes the element-wise product, $\|\mathbf{M}\|_*$ is the nuclear norm of the matrix $\mathbf{M}$, which is the sum of its singular values, λ is the regularization parameter that controls the trade-off between imputation accuracy and the rank of the matrix. SoftImpute proceeds by iteratively solving the following steps:

Algorithm 2 SoftImpute Algorithm

1: Initialize the missing values in $\mathbf{X}$ with zeros or the column means to form $\mathbf{M}_0$.
2: **while** not converged **do**
3: Perform Singular Value Decomposition (SVD) on the current matrix estimate $\mathbf{M}_t$:

$$\mathbf{M}_t = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

4: Apply soft-thresholding to the singular values $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}_\lambda = \max(\mathbf{\Sigma} - \lambda, 0).$$

5: Update the matrix $\mathbf{M}_{t+1}$ using the soft-thresholded singular values:

$$\mathbf{M}_{t+1} = \mathbf{U}\mathbf{\Sigma}_\lambda\mathbf{V}^\top.$$

6: Replace the missing entries in $\mathbf{X}$ with the corresponding values from $\mathbf{M}_{t+1}$.
7: **end while**

The algorithm iteratively reduces the objective function and converges when the change in the matrix $\mathbf{M}$ between iterations is below a specified tolerance level. The regularization parameter λ controls the amount of shrinkage applied to the singular values, which determines the rank of the resulting matrix. A larger λ results in more aggressive shrinkage and a lower-rank approximation. There are a few key advantages in the SoftImpute algorithm summarized as follows:

- **Scalability**: SoftImpute can efficiently handle large matrices with many missing entries by exploiting the low-rank structure of the data.
- **Flexibility**: The nuclear norm regularization helps in controlling overfitting and provides smooth low-rank approximations.
- **Simplicity**: The algorithm is easy to implement and can be combined with other methods for improved imputation accuracy.

C.2 Intuition Behind the Low-Rank Assumption

More generally, we can view the low-rank assumption in our setting through the structure of the CATE matrix. 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:

$$\mathcal{T} = XB^T, \tag{A.29}$$

where B is a $J \times D$ matrix where each column is the vector of coefficients for CATE for a specific treatment. For the low-rank assumption to be satisfied, we need matrix B to be low-rank. If the studies have similar characteristics, we expect weights in each row of B to be correlated, thereby making the matrix low-rank. Suppose there are two matrices $U_{J\times R}$ and $V_{D\times R}$ such that $B = UV^T$. In this case, $\mathcal{T} = XVU^T$, where XV maps the high-dimensional covariates into R factors, and U contains the weights for these factors in the different studies.

Apart from structural reasons for the suitability of low-rank assumption in the context of digital platforms, the insights from the prior literature suggest that the low-rank assumption performs remarkably well in a wide range of domains, especially when large-scale matrices are available. This insight is formally characterized in Udell and Townsend (2019) who show that under general conditions that the function generating the high dimensional $N \times J$ matrix is analytic piece-wise, the rank grows as $O(log(N + J))$.

D Supplementary Materials for the Calibrated Simulation

D.1 Simulation Details

In this section, we present the details of our calibrated simulation exercise. We first define a few preliminaries. As described in §5, in our simulation, we have $N = 100,000$ and $A = 100$. We define the covariate matrix as $X_{N\times D}$ where D is the dimensionality of the covariate space. We present a step-by-step procedure as follows:

- *Defining the base for CATE matrix:* The base for the underlying CATE matrix is given by the following equation:

$$\tilde{\mathcal{T}} = XB^T, \tag{A.30}$$

  where $B_{J\times D}$ is the coefficient matrix that is low-rank in the following way:

$$B = UV^T, \tag{A.31}$$

  where $U_{J\times R}$ and $V_{D\times R}$ are two matrices that make matrix B rank-R. All the entries in these matrices come from $\mathcal{N}(0, 0.5)$. This ensures that the mean of each column in matrix $\tilde{\mathcal{T}}$ is equal to zero. We now sample from the lift deciles provided from Gordon et al. (2022) to determine the ATE for each column and add the ATE to all entries in that column. This gives us the CATE matrix $\mathcal{T}_{[N\times A]}$.
- *Defining the bid matrix:* The bid matrix is something that is imperfectly correlated with the CATE matrix $\mathcal{T}_{[N\times A]}$. We define this matrix as $\mathcal{B}_{[N\times A]}$, such that the correlation between each column a in $\mathcal{T}$ and $\mathcal{B}$ is equal to 0.5.
- *Determining propensity scores:* Based on the bids defined at the user-level in the previous step and Lemma 2, we calculate each user-ad pair's propensity score. This defines the propensity matrix $\Pi_{[N\times A]}$ in our study.
- *Defining the nuisance matrix:* The nuisance matrix $\mathcal{G}_{[N\times A]}$ determines the relationship between covariates and the outcome. We define the nuisance matrix as a product of $X_{N\times D}$ and a weight matrix $G_{A\times D}$ as follows:

$$\mathcal{G} = XG^T + 1, \tag{A.32}$$

  where entries in the weight matrix $G_{A\times D}$ all come from $\mathcal{N}(0, 0.5)$. The addition of the term one is only to ensure that reported lifts are the same as ATEs.

We can now use all these primitives to simulate the data for our calibrated simulation exercise:

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

$$Y_i^{(a)} = \mathcal{G}_{i,a} + W_i^{(a)}\mathcal{T}_{i,a} + \epsilon_{i,a}, \tag{A.33}$$

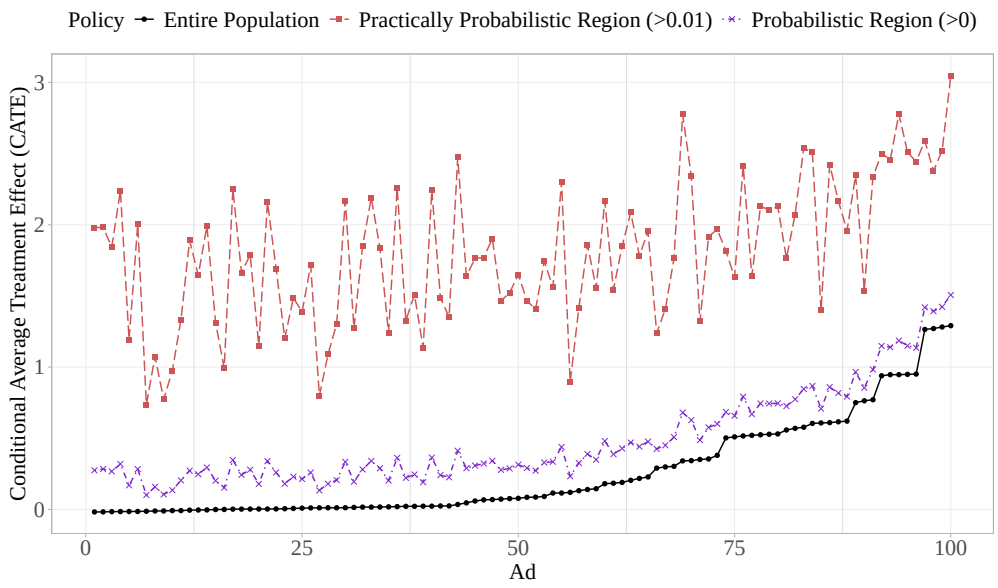


Figure A.1. Treatment effects for different regions of the data

where $\mathcal{G}_{i,a}$ is the nuisance part of the outcome, $W_i^{(a)}\mathcal{T}_{i,a}$ is the treatment effect given (if any), and $\epsilon_{i,a} \sim \mathcal{N}(0, 0.5)$.

- *Step 3:* For each study a, we can construct data set $\tilde{\mathcal{D}}^{(a)} = \{Y_i^{(a)}, W_i^{(a)}, X_i, \pi^{(a)}(X_i)\}$. The union of $\tilde{\mathcal{D}}^{(a)}$ for all a's will give us the $\mathcal{D}_T^{\text{sim}}$.

D.2 Relationship Between Propensity Scores and CATE

Next, we examine whether the lack of overlap induced by algorithmic ad allocation poses challenges for ATE estimation using observational data. As discussed in earlier in §3.2.2, if the CATE for the probabilistic region is different from ATE, all state-of-the-art methods will fail to recover the true ATE. We define the probabilistic region in our data in two ways: (1) *probabilistic region* where the propensity score is greater than zero and satisfies the weak overlap assumption, and (2) *practically probabilistic region* where the propensity score is greater than $\eta = 0.01$ and satisfies the strict overlap assumption. We define the *practically probabilistic region* because this is the region that empirical models can effectively use to learn treatment effect estimands. We then use the true CATE matrix and calculate the CATE for each region. Figure A.1 shows the true CATEs for these three regions. As shown in this figure, the CATE for both probabilistic regions is higher than the treatment effect for the population. This indicates that user-ad pairs with lower CATEs are systematically more likely to violate the overlap assumption. In particular, the correlation between CATE and propensity scores for all user-ad pairs is 0.37, which highlights the challenge posed on the observational methods to recover treatment effect estimands: the overlap-satisfying part of the data selected.

E Supplementary Materials for the Empirical Validation Exercise

E.1 Complete List of Features Used for Targeting Profiles

For each impression or targeting profile, we observe the following variables: (1) Latitude, (2) Longitude, (3) Province, (4) Smartphone Brand, (5) Connectivity Type, (6)

- Latitude
- Longitude
- Province
- Smartphone Brand
- Connectivity Type
- Mobile Service Provider
- The total number of impressions the user has seen prior to the current session
- The total number of clicks user the user has made prior to the current session
- The total number of impressions the user has seen prior to the current session in the top app
- The total number of clicks user the user has made prior to the current session in the top app
- The number of times the user has seen at exposure number t in prior sessions
- The number of times the user has clicked at exposure number t in prior sessions
- The length of last session (in number of exposures) that the user was exposed to prior to the current session
- The average length of sessions (in number of exposures) that the user was exposed to prior to the current session
- The gap or free time (in minutes) the user has had between her last session and the current session
- The average gap or free time (in minutes) the user has had between any two consecutive prior sessions
- The total number of distinct ads that the user has seen prior to the current session
- The Gini-Simpson index for ads that the user has seen prior to the current session
- The Shannon entropy of ad frequencies that the user has seen prior to the current session
- The total number of impressions the user has seen in the current session
- The total number of clicks user the user has made in the current session
- The total number of distinct ads that the user has seen within the current session
- The total number of consecutive changes of ads the user has experience in the current session
- The Gini-Simpson index for ads that the user has seen in the current session
- The Shannon entropy of ad frequencies that the user has seen in the current session

E.2 Details of Incomplete

In §6.3.2, we considered four versions of our algorithm based on the missingness pattern in the CATE matrix. We now present the details on each:

1. *Incomplete for Low CATE*: In this form of missingness, entries for other ads that have lower than median values of CATE are missing.

2. *Incomplete for Higher CATE*: In this form of missingness, entries for other ads that have higher than median values of CATE are missing.

3. *Incomplete for High/Low CATE*: In this form of missingness, entries for other ads that have higher than third quartile or lower than first quartile of the CATE distribution are missing. That is, the inter-quartile range of CATE values satisfy overlap.

4. *Complete*: In this setting, there is no missingness in the matrix of other ads.