# A Matrix Completion Solution to the Problem of Ignoring the Ignorability Assumption in Advertising Measurement

**Abstract**

Digital platforms increasingly rely on algorithms for decision-making. A canonical success story is online advertising, where platforms use algorithmic auctions to allocate ads efficiently and maximize revenue. Another pillar of these platforms' long-term success is accurate causal measurement of advertising effectiveness. While randomized experiments remain the gold standard, their high cost makes observational approaches appealing to both platforms and advertisers. In this paper, we show that the algorithmic nature of ad allocation fundamentally challenges causal measurement by violating the overlap assumption required for valid estimation. Even when treatment assignment is unconfounded, the deterministic structure of auctions implies that many users receive ads with near-zero or near-one probability, leading to biased estimates of ad effectiveness. We develop a novel framework that interprets this lack of overlap as a missing data problem and leverages variation across multiple advertising campaigns using matrix completion to recover causal estimates. Using both simulated and real advertising data, we demonstrate that our approach accurately recovers ad effects at aggregate and individual levels and substantially improves targeting efficiency, offering a scalable solution for causal advertising measurement in algorithmic markets. We conclude with a discussion of managerial implications and potential extensions to other algorithmic decision-making settings.

**Keywords**: causal inference, machine learning, overlap assumption, advertising measurement, digital platforms, observational methods

## 1 Introduction

Digital platforms often rely on algorithms to deliver interventions (e.g., ads, promotions). A canonical success story of algorithmic decision-making is online advertising auctions that determine how ads are allocated across impressions. These auctions serve two main objectives of advertising platforms: (1) automated and scalable delivery of ads, and (2) strong revenue guarantees by following foundational economic principles established in auction theory (Ostrovsky and Schwarz 2023). The key to the long-term success of this algorithmic ecosystem is better advertising measurement at the aggregate and individual levels, which helps platforms and advertisers make better decisions.

Causal advertising measurement has been a longstanding goal in both marketing research and practice (Lodish et al. 1995, Gordon et al. 2019). The gold standard for advertising measurement is to run fully randomized experiments known as Randomized Controlled Trials (RCTs). However, the economic cost of it for advertising platforms is high, making experimentation infeasible in many contexts. In particular, because ad revenues account for the majority of total revenues for major advertising platforms such as Meta and 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). Therefore, it is valuable for these platforms to estimate the causal effects of ads using observational data.

The core challenge in using observational methods for advertising measurement is that the ads are usually targeted, making their assignment to users non-random. In the context of TV advertising, advertisers can target their ads at the Designated Market Area (DMA) level, which creates selection problems in measuring advertising effectiveness (Shapiro et al. 2021). Targeting capabilities are substantially more fine-grained in the context of digital advertising, making the underlying selection problems more challenging. As a consequence, a convergent finding that emerges from the prior literature is that observational methods are vastly inadequate in measuring the true ad effectiveness (Lewis et al. 2011, Gordon et al. 2022).

The failure of observational methods in causal advertising measurement comes from the violation of a well-known assumption needed for both experimental and observational methods to work: strong ignorability of the treatment assignment (Rosenbaum and Rubin 1983). The strong ignorability assumption is a mix of two assumptions: (1) *unconfoundedness* of the treatment assignment, which intuitively means that there is no relevant variable hidden from researchers that influence the treatment assignment, and (2) *overlap* or *positivity* of the treatment assignment, which assumes that the treatment assignment is not deterministic, that is, the probability of users being exposed to ads (propensity score) is strictly between 0 and 1.

What is different in digital advertising 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. In online advertising, ads are allocated through auctions that are largely deterministic and induce only partial and local randomization in assignment. Therefore, even if this context satisfies the unconfoundedness assumption because the algorithmic outputs are readily available to the platform, it violates the overlap assumption because of the deterministic assignment employed by the algorithms.

In this paper, we consider a digital advertising platform that seeks to estimate the causal effect of ads

at the population and individual levels for better decision-making and targeting. We aim to achieve the following goals in our research agenda. First, we study the consequences of overlap violation in causal ad measurement and evaluate whether the state-of-the-art causal machine learning approaches can overcome the estimation challenges. Second, we examine how prevalent the overlap violation is in the context of algorithmic ad allocation through auctions. Third, we design an algorithm that overcomes the challenges posed by the overlap violation. Lastly, we evaluate the performance of our proposed algorithm in recovering the causal parameters and quantify the gains from using them for decision-making.

To accomplish our objectives, we first develop a simple generic 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 Group Average Treatment Effects (GATE) 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 GATE for the probabilistic region and fail to estimate the population ATE. Specifically, when assignment probability depends on the heterogeneous treatment effect, the absence of overlap can lead to sizable differences between the GATE for the probabilistic region and the population ATE. We highlight the prevalence of this issue in advertising auctions because advertisers' optimal bids are directly related to the user-level advertising effectiveness, which leads to systematic biases in measuring aggregate advertising effectiveness.

Once we establish the existence and prevalence of the lack of overlap in observational studies involving digital advertising platforms and the challenges it poses, we focus on devising a way to estimate Conditional Average Treatment Effects (CATE) for units with deterministic assignment. We propose a framework that formulates the unidentifiability of CATE for regions with deterministic treatment (ad) assignment that violate the overlap assumption as a missing data problem. Although we cannot fix this problem with a single advertising campaign at hand, we can potentially use the information across advertising campaigns to help with this missing data problem. In particular, if we have multiple advertising campaigns with different ads (e.g., smartwatch ad in one study and mobile health app ad in another) whose individualized effects come from a low-rank space, we can use matrix completion methods to impute the CATE 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 the 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 advertising campaign, we can exploit similarities across users and across ads 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 estimate the CATE for the missing entries of the matrix and correct the bias in the population ATE estimates.

We use a calibrated simulation in the context of online advertising, where we micro-found the algo-

rithmic 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 receive a propensity score of 0. 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 a typical allocation mechanism can lead to overlap violation. Interestingly, we find that the extent of the overlap violation is so severe that the GATE 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 strong 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 in advertising measurement. 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 advertising measurements at the aggregate and individual levels. From a substantive and practical viewpoint, we identify an important challenge for advertising platforms that employ algorithmic ad allocation. 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 advertising platforms is, in fact, the overlap violation. We further discuss empirical contexts where this problem may arise and demonstrate the value of our algorithm through both synthetic experiments and real field data, as it allows for better targeting of interventions. Notably, our framework is fairly general and can be extended to other settings within and beyond advertising, where platforms deliver numerous interventions that have common factors and satisfy the low-rank requirements but the treatment effects are not identifiable for entries in the matrix. 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. 2020), 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). In particular, a handful of studies in the prior literature have used the variation across studies as a useful source of variation for the estimation task at hand (Zantedeschi et al. 2017, Gordon et al. 2023). 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 a series of simulations–that numerous user-ad pairs have practically 0 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 that allows managers to implement personalized targeting of interventions.

## 3 Methodological Framework

We now present our methodological framework for a general class of problems in which a decision-maker delivers interventions to observational units and seeks to measure the causal effect of these interventions for better decision-making. Although the focus of this paper is on advertising measurement, we follow the convention in applied causal inference literature and keep our methodological framework generic to facilitate extensions to broader domains—an objective we revisit later when discussing broader implications.

In this section, we start with a theoretical analysis of overlap violation and an impossibility result in causal measurement using a single study in §3.1. Next, in §3.2, motivated by the impossibility result in a single-study setting, we extend to a multi-study setting, formally define our problem and formulate the estimation problem under overlap violation as a missing data problem. We then present our proposed algorithm based on matrix completion in §3.3, discuss the identification assumptions required for its validity in §3.4, and outline evaluation metrics for assessing its performance in §3.5.

### 3.1 Causal Measurement Under Overlap Violation: A Theoretical Analysis

Consider a general setting in which a decision-maker delivers interventions to observation units. An observation unit can be defined at any level at which interventions occur (e.g., online users, targeting profiles, cities). As a running example throughout this section, we focus on an advertising platform that delivers ads to users online, and later examine two applications: one where the user is the unit of observation, and another where the impression is the unit of observation, in Sections 4 and 5, respectively. When an observation unit is available to receive the intervention, the decision-maker 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 that unit, and $X_i$ denote the vector of observable characteristics from the super set $\mathcal{X}$. As customary in digital settings such as online advertising, 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, decision-makers generally rely on a policy function. This policy function–which is increasingly determined algorithmically in digital settings–takes the feature vector $X_i$ as input and outputs a probability distribution over possible interventions. For any intervention $w \in \mathcal{W}$, we define the policy as a function $\pi_w : \mathcal{X} \to [0, 1]$, where $\pi_w(X_i)$ represents the probability that the platform assigns intervention $w$ to unit $i$. The function $\pi_w$ corresponds to the propensity score in the causal inference literature. In the context of online advertising, the policy function is determined by the auctions used by the advertising platform to algorithmically allocated ads. Because this process is internal, the decision-maker often has direct access to the policy function $\pi_w(\cdot)$, which is an assumption we maintain throughout the paper.

Once the intervention is delivered, the decision-maker collects the outcome of interest $Y_i$ for unit $i$. In advertising measurement, this outcome can be a user's clicks or conversion decision. 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 causal estimands that researchers and practitioners often want to estimate. The first causal estimand is *Average Treatment Effect (ATE)* that we denote by $\tau^*$ and define it as follows:

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

where the expectation is taken over the entire population. The second causal estimand, *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 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

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

assumption is violated.[2]

The second set of assumptions is known as *Strong Ignorability* assumption, which is defined in the seminal paper by Rosenbaum and Rubin (1983). The assignment to treatment is *strongly ignorable* given the observed covariates $X_i$, if we have:

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

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

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

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

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

  where $\Pr(W_i = 1 \mid X_i)$ is the same as the propensity score when $w = 1$, that is, $\pi(X_i)$.[3] This assumption is often referred to as the *overlap* or *positivity* assumption and guarantees that the assignment to the treatment is not deterministic. Intuitively, this assumption ensures that the distribution of covariates under treatment fully overlaps with that of covariates under control.

**What is different in settings with algorithimic decision-making?** 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 unobserved 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 (e.g., online advertising) is that the assignment rule is often fully observable. That is, the platform can store the $X_i$ used for algorithmic decision-making or at least 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 can address 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 largely deterministic. That is, $\pi(X_i)$ can be equal to 0 or 1 depending on $X_i$.

We now present our formal problem definition in a single-study setting:

**Problem 1** (Single-Study Setting)**.** *Consider a decision-maker that uses data* $\mathcal{D} = \{Y_i, W_i, X_i, \pi(X_i)\}$*. The main estimands the decision-maker 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.*

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

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

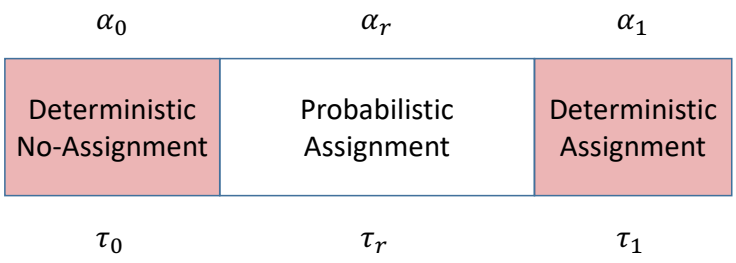



Figure 1. Different regions based on the type of assignment.

Our primary goals in this section are to (1) theoretically examine the possibility of causal measurement under overlap violation, (2) quantify the magnitude of bias due to this overlap violation, and (3) identify the link between this bias and the algorithm used by the platform to shed light on the prevalence of this problem in real-world applications.

### 3.1.1 Impossibility of Causal Measurement Under Overlap Violation in a Single-Study Setting

In this section, we theoretically analyze how the lack of overlap can lead to biased estimates of the average treatment effect (ATE). 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 group average treatment effect (GATE) 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 $\alpha_r$.
- *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 $\alpha_0$.
- *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 $\alpha_1$.

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 $\tau_1$. 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 $\tau_0$, since there is no residual variation in treatment. Thus, we can write the following lemma stating our main impossibility result:

**Lemma 1.** *The group average treatment effects $\tau_1$ and $\tau_0$ are unidentifiable given data $\mathcal{D}$.*

### 3.1.2 Bias Analysis

In light of Lemma 1, the only identifiable piece of $\tau^*$ is $\tau_r$. 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 $\tau_r$ and fail to estimate the population ATE $\tau^*$:

**Proposition 1.** *Suppose there is a decision-maker 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 0 and 1 for parts of the population. Let $\widehat{\tau}_{DML}$ and $\widehat{\tau}_{IPS}$ denote the ATE estimate based on Double Machine Learning and Inverse Propensity Scoring estimators, respectively. Both these estimates converge to $\tau_r$ in probability, that is, $\widehat{\tau}_{DML} \xrightarrow{p} \tau_r$ and $\widehat{\tau}_{IPS} \xrightarrow{p} \tau_r$.*

*Proof.* See Web Appendix B.1. □

Lemma 1 and Proposition 1 highlight an important identification problem and impossibility result 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 treatment effects for the probabilistic region under unconfoundedness, 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 $\tau^*$, 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:

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

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 $\tau_r$ and $\tau^*$. However, it is easy to imagine scenarios where the difference in $\tau_1$, $\tau_0$, and $\tau_r$ creates a substantial difference in estimates of the average treatment effect. In fact, for any constant $c$, we can find $\tau_0$ and $\tau_1$ 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.

### 3.1.3 Practical Relevance

As discussed earlier, the difference between the identifiable piece $\tau_r$ and the population ATE $\tau^*$ 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

[4] See a summary of state-of-the-art model-based and model-free approaches to estimate treatment effects in Web Appendix A.

settings? Part of the rationale for the trimming approaches that are widely used in the literature is that $\tau_0$ and $\tau_1$ are not different from $\tau_r$. Here we ask the following question: is this homogeneity assumption (i.e., $\tau_0 = \tau_r = \tau_1$) correct in the advertising measurement problem?

Intuitively, to the extent that $\pi(x)$ is a function of $\tau^*(x)$, we expect $\tau_0$ and $\tau_1$ to be different from $\tau_r$. The following proposition formalizes this insight:

**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 $\tau_0 \neq \tau_r \neq \tau_1$. We now examine how this issue can arise in the advertising measurement problem, where the goal is to use observational methods to measure the average treatment effect of digital ads on users (Gordon et al. 2019, 2022). Digital ads are sold through auctions, where 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. 2025). The auction setting implies that an ad may never reach certain users if their CATE is too low, since other advertisers will always outbid it. This results in a form of *deterministic no-assignment*: some users in the control condition could never have 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, which can lead to an over-estimation of the ATE using state-of-the-art observational methods. Later in §4, we micro-found the ad allocation process to study the canonical advertising effectiveness measurement problem and illustrate how the issue of overlap violation could severely bias the treatment effect estimates.

### 3.2 Problem Definition: Overlap Violation as a Missing Data Problem

In the previous section, we presented the challenge decision-makers face due to the lack of overlap in observational studies. For example, observational methods lead to biased estimates of advertising effectiveness, using the data of a single ad campaign. However, the presence of multiple ad campaigns in the platform context suggests the possibility of transferring information between campaigns. Motivated by the availability of data from other interventions and A/B tests in platform settings, in this section, we re-formulate our problem for the multi-study setting, which allows us to build a novel solution based on matrix completion methods to overcome the challenge posed by the overlap violation.

In light of our impossibility result in Lemma 1, the fundamental problem with the deterministic assignment is one of identification: we know that with the current set of assumptions, the parameters $\tau_1$ and $\tau_0$

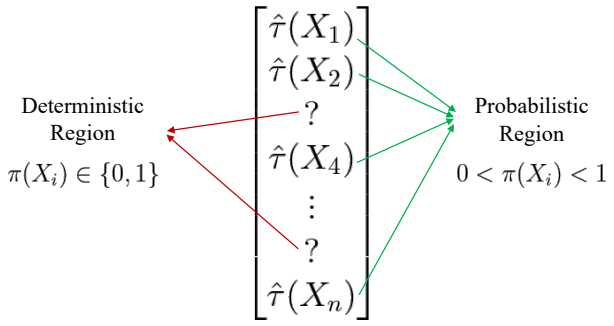



Figure 2. An illustration of the missing data problem due to the overlap violation.

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 never have been generated in the data, no model can estimate the corresponding $\mu$ function. For example, if a user with covariates $X_i$ could never have seen an ad (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 numerous different treatments (e.g., ad campaigns in Facebook) at a large scale. Motivated by this insight, we define the problem of the digital platform as follows:

**Problem 2** (Multi-study Setting)**.** *Consider a decision-maker 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 $i^{\text{th}}$ observation is either 0 or 1, i.e., $W_i^{(j)} \in \{0, 1\}$. For each study $j$, the decision-maker 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 decision-maker'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 decision-maker achieves the objective in Problem 2, it can recover the average treatment effect for the treatment in each study.

**Data Requirements:** Problem 2 highlights a few important data requirements. First, treatments in different studies can be different. For example, the treatment in study $j$ can be an ad for a smartwatch and the treatment in study $k$ can be an ad for a mobile health app. One could imagine this as different interventions the platform made over time, such as different ad campaigns in an advertising platform, or different in-app interventions (e.g., free coins) in a gaming app. 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 unit can be assigned to multiple treatments (e.g., both the smartwatch ad and mobile health app ad in the example above). Lastly, we require having data on multiple interventions that induce sufficient probabilistic assignment. It is important to stress that this requirement is not excessive in the context of digital platforms. For example, advertising platforms often deliver numerous ads to their users that induce some probabilistic assignment, either through direct experimentation and A/B testing or natural experiments.

### 3.3 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 0 or 1. We know that the CATE is unidentified for these units. Thus, we define a feasibility matrix $F$ that takes value 1 only when the assignment is probabilistic; that is, the propensity score is strictly between 0 and 1. 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 any consistent 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 Problem 2 into a matrix completion problem, where we have an estimated CATE matrix $\widehat{\mathcal{T}}^{\text{incomplete}}$ and each element $[i, j]$ is defined as follows:

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

As shown in Equation (A.32), $F$ determines the question marks in our matrix completion task. We now have an incomplete matrix $\widehat{\mathcal{T}}^{\text{incomplete}}$, 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. In our algorithm, we use the SoftImpute algorithm proposed by Mazumder et al. (2010) as the algorithm we use for matrix completion, which combines Singular Value Decomposition (SVD) with soft thresholding to obtain a low-rank approximation of the incomplete matrix. This procedure exploits the similarities in the joint space of units and treatments. We present more details on the SoftImpute algorithm in Web Appendix C.1. It is worth noting that we use SoftImpute because of its computational performance, but a researcher can use different algorithms depending on the features of the problem (e.g., missingness pattern). One could perform a generic model selection procedure using validation data to select that best-performing matrix completion model from the host of methods available (Koren et al. 2021). Once we complete the matrix, we denote it by $\widehat{\mathcal{T}}^{\text{complete}}$. Algorithm 1 presents the details of our proposed approach.

**Algorithm 1** Matrix Completion for CATE Estimation

**Input:** $\mathcal{D}_T$ ▷ From Problem 2
**Output:** $\widehat{\mathcal{T}}^{\text{complete}}$

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

The output of this algorithm is a complete matrix $\widehat{\mathcal{T}}^{\text{complete}}$ where all the elements are imputed. This complete CATE 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:

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

If the matrix $\mathcal{T}$ is low-rank, $\widehat{\tau}^{(j)}$ 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 §4.3.3 and §5.4.2.

Two important questions about the proposed algorithm is how to select the optimal rank of the matrix, and how to quantify uncertainty around the estimates. We discuss both in the following sections.

### 3.3.1 Validation Procedure for Optimal Rank Selection

Any matrix completion method has a set of tuning parameters that control the rank of the final imputed matrix. In the SoftImpute algorithm we use, there are two key parameters: (1) the regularization parameter $\lambda$, which controls the rank of the estimated matrix by regularizing its nuclear norm, as discussed in Web Appendix C.1, and (2) the maximum rank, which is the maximum allowable rank of the matrix. In many settings, researchers either use domain knowledge to set the maximum rank or automatically set it as the number of columns.

To tune $\lambda$, we use a validation procedure to obtain the best-performing value by taking the following steps. We first build a grid of candidate $\lambda$ values. To select the maximum $\lambda$ in the grid, we use the highest singular value of the matrix $\widehat{\mathcal{T}}_{i,j}^{\text{incomplete}}$ when the missing entries are replaced with 0s. Second, for model selection, we split the observed entries into training and validation sets, holding out an $\alpha$ fraction of all observed entries in the validation set. We train models based on different values of $\lambda$ in our grid and select the one with the best performance on the validation set. Importantly, this validation procedure also helps verify the validity of the low-rank assumption. For example, if the best-performing matrix has a high rank, we take that as evidence that the low-rank assumption may not be appropriate for the problem. We stress that even in this case, the algorithm still acts as a bias-reduction tool and performs no worse than the conventional method described in §3.1.2. In Web Appendix C.2, we provide more details on the validation procedure for the SoftImpute algorithm.

### 3.3.2 Uncertainty Quantification

The presence of noisy entries and missingness introduce uncertainty in the imputed entries of the matrix. Quantifying uncertainty has been a long-standing challenge in the matrix completion literature, and only recently have researchers developed tools to address it analytically (Chen et al. 2019, Zhao and Udell 2020, Agarwal et al. 2021, Gui et al. 2023). However, these asymptotic theories often make strong assumptions about missingness patterns and only apply to a specific class of matrix completion methods. As such, we use re-sampling methods to account for the statistical fluctuations in the imputed matrix, similar in spirit to the approach proposed in Athey et al. (2021). In particular, we leverage the uncertainty in the CATE estimates from the first stage, resample these parameters from their estimated distribution, construct the incomplete matrix, and then complete it using Algorithm 1. To see how this procedure works, let $\widehat{\Theta}_j(\cdot)$ denote the distribution of estimated CATE parameters for study $j$. The details of our procedure are presented below:

- Step 1: Draw $\widetilde{\theta}_j \sim \widehat{\Theta}_j(\cdot)$ for each study $j$.

- Step 2: Build an incomplete CATE matrix $\widetilde{\mathcal{T}}^{\text{incomplete}}$ as follows:

$$\widetilde{\mathcal{T}}_{i,j}^{\text{incomplete}} = \begin{cases} \widehat{\tau}^{(j)}(X_i; \widetilde{\theta}_j) & \text{if} F_{i,j} = 1 \\ ? & \text{if} F_{i,j} = 0 \end{cases} \tag{13}$$

  This approach ensures that any perturbation in drawn parameters affect all the entries in a single column. Approaches that rely on independent entry-wise draws ignore the within-study dependence, so our procedure is advantageous.
- Step 3: Complete CATE using the same matrix completion procedure in Algorithm 1. Define the complete matrix as $\widetilde{\mathcal{T}}^{\text{complete}}$.
- Step 4: Repeat this process $B$ times and build confidence intervals for entries.

Repeating this process multiple times allows us to build confidence intervals for any imputed entry $\widehat{\mathcal{T}}_{i,j}^{\text{complete}}$, as well as the column-wise average $\widehat{\tau}^{(j)}$ that recovers the ATE. The uncertainty measured stems from both the sampling uncertainty in the matrix entries and the uncertainty induced by the matrix completion algorithm due to the missingness structure. As such, like most standard econometrics models, our procedure quantifies uncertainty given the validity of modeling assumptions. However, it is important to emphasize that the structural epistemic uncertainty from the modeling assumptions is not supposed to be captured with this procedure. For example, if the underlying matrix is full rank and all columns are independent of each other, a low-rank approximation is not well-suited for the problem. As such, the confidence interval in this example will not capture the true parameter. This situation parallels the case with OLS estimation, where the confidence intervals for a coefficient do not capture the true value if the model exhibits omitted variable bias. In all these cases, the validity of modeling assumptions must be assessed separately. For our algorithm, we will specifically discuss how one could assess the validity of these assumptions in §3.4.

### 3.4 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 unit in an overlap-violating region ($F_{i,j} = 0$), if we have enough cross-study variation, we can exploit the similarities in the data to impute the conditional average treatment effect for that individual unit. 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 individual units (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 individual units (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[5], 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.

### 3.4.1 Low Rank CATE Matrix

The key 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 units' responses exhibit common patterns across different treatments. Theoretically, the low-rank assumption is what guarantees the possibility of fully recovering a matrix from only a fraction of its entries using convex optimization (Candès and Recht 2009). A more practical interpretation of the assumption is that the original matrix can be well-approximated by a low-rank matrix. In this sense, the low-rank assumption can be viewed less as an assumption and more as a hyperparameter that can be optimally chosen in a data-driven manner.

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 unit 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 govern user behavior. For example, in promotional ad campaigns, 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 unit-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.[6]

**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 finds an ad relevant. In that case, it is reasonable to assume that a few factors can largely explain the variation in users' ad preferences, as is commonly assumed in recommender systems.

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

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

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. In general, the homogeneity of studies is a condition that is likely satisfied in most digital platforms as interventions likely share some common characteristics. In Web Appendix C.3, we present more structural reasons why this assumption is widely applied in practice, particularly in recommendation systems. Later in §4.1, 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 the assumption is testable and transparent. Recall the validation procedure in §3.3, where we directly select the best-performing low-rank approximation of the matrix on the validation set. This allows us to assess whether the underlying matrix is indeed low-rank. Moreover, it is important to note that even if the underlying matrix is not low-rank, our method performs at least as well as existing approaches. In §5, we present an empirical application in the context of mobile in-app advertising and validate the low-rank assumption in that setting.

### 3.4.2 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.*

Intuitively, the missingness pattern needs to be such that we can jointly exploit the similarities between units 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 unit is very responsive to interventions and therefore deterministically receives the intervention across all studies. However, even in such cases, one could empirically verify whether such missingness patterns exist. In our calibrated simulation and empirical validation exercises in §4 and §5, we consider realistic and challenging missingness patterns to provide validity to this assumption.

Theoretically, we need Assumption 2 to ensure that we can identify factors from the observed entries. That is, there must be enough variation in the observed data to estimate all entries of the low-rank matrix. We acknowledge that this is a strong assumption, but we are not concerned about it in our application for two reasons. First, platforms run numerous fully randomized experiments, in which all rows are feasible for the CATE estimation task. Including these experiments populates the observed matrix by creating columns that are fully observed, thereby ensuring identification. Second, there are simple ways to test whether this assumption is reasonable. One approach is to use the feasibility matrix for a simulated low-rank approximation. Specifically, one could simulate the underlying matrix using a low-rank structure and induce

missingness according to matrix $F$. If the algorithm is able to accurately complete the matrix by identifying the low-rank factors, this provides evidence in support of Assumption 2.

#### 3.4.3 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 $\mathcal{T} = \widehat{\mathcal{T}} + E$, 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. Theoretically, this ensures that the systematic noise in observed entries does not bias the objective function. 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 in our applications.

### 3.5 Algorithm Evaluation

Algorithm 1 estimates all entries in the CATE matrix. One way to evaluate our algorithm is to assess the accuracy of these estimates when ground-truth values are available. Since CATE estimates can be used to develop personalized policies, another way to evaluate the algorithm's performance is by examining its targeting effectiveness. We present the evaluation strategies for these two components as follows:

- **Accuracy Performance:** Our algorithm returns the complete matrix $\widehat{\mathcal{T}}^{\text{complete}}$. If we know the ground truth CATE matrix $\mathcal{T}$, we can define the Root Mean Squared Error of both CATE and ATE estimates as follows:

$$\texttt{RMSE}_{\text{CATE}}(\widehat{\mathcal{T}}^{\text{complete}}; \mathcal{T}) = \sqrt{\frac{1}{NJ}\sum_{i=1}^{N}\sum_{j=1}^{J}(\widehat{\mathcal{T}}_{i,j}^{\text{complete}} - \mathcal{T}_{i,j})^2} \tag{14}$$

$$\texttt{RMSE}_{\text{ATE}}(\widehat{\mathcal{T}}^{\text{complete}}; \mathcal{T}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\mathcal{T}}_{i,j}^{\text{complete}} - \frac{1}{N}\sum_{i=1}^{N}\mathcal{T}_{i,j}\right)^2\right)} \tag{15}$$

  The RMSE measures approximate the average deviation of our estimates from the truth.

- **Targeting Performance:** A more managerial measure that we can use is the targeting performance of the personalized policy developed based on our estimated matrix $\widehat{\mathcal{T}}^{\text{complete}}$. This would be a proxy for the real economic gains from using our proposed algorithm. To develop this measure, we simply measure the average gain from assigning the top $\alpha$ fraction of units to treatment based on a model, compared to giving the control condition to the whole population. We define this measure using the function $\texttt{Gain}_{(\alpha)}$ as follows:

$$\texttt{Gain}_{(\alpha)}(\widehat{\mathcal{T}}^{\text{complete}}; \mathcal{T}) = \frac{1}{NJ}\sum_{j=1}^{J}\sum_{i=1}^{N}\mathbb{1}\left(\widehat{\mathcal{T}}_{i,j} \geq F^{-1}_{\widehat{\mathcal{T}}_j}(1-\alpha)\right)\mathcal{T}_{i,j}, \tag{16}$$

where the indicator function only selects the top $\alpha$ fraction of estimated CATE for any study $j$. We further define similar but normalized measure that compares the average gain from the model to the that from the oracle (first-best) that knows the true CATE values. We define this using the $\texttt{OracleRatio}_{(\alpha)}$ as follows:

$$\texttt{OracleRatio}_{(\alpha)} = \frac{\texttt{Gain}_{(\alpha)}(\widehat{\mathcal{T}}^{\text{complete}}; \mathcal{T})}{\texttt{Gain}_{(\alpha)}(\mathcal{T}; \mathcal{T})} \tag{17}$$

The Oracle Ratio measure is always less than or equal to 1 and has a straightforward modeling interpretation: it indicates the percentage of the gap between the oracle and the all-control policy that is explained by the algorithm.

In both our applications in §4 and §5, we use both measures to evaluate the performance of our algorithm.

## 4 Application: Online Advertising

Advertising measurement 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 machine learning algorithms that tune auction weights to allocate ads. Given that ad revenues account for the majority of total revenues for major advertising platforms such as Meta and 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, including both the advertising platform and the advertisers.

In this section, we first define the CATE matrix in §4.1, which is our target estimand. We then describe algorithmic allocation in common advertising auctions and show how this allocation leads to a violation of the overlap assumption in §4.2. Next, we present results from our calibrated simulations in §4.3. Lastly, we present a series of robustness checks and sensitivity analyses in §4.4.

### 4.1 Estimation Target: CATE Matrix

We start by defining the estimation target in the advertising application. Problem 2 presents a general characterization of the problem. In this section, we want to define all elements in that problem 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 $a$, over the course of potentially many impressions shown to user $i$.[7] 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

[7] This is consistent with the recent experimental literature on advertising (Lewis et al. 2011, Gordon et al. 2019, 2022) where we define units to be users as opposed to the ad response modeling where we are interested in the effects at the impression level.

[8] One could 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 §5.

the Conditional Average Treatment Effect (CATE) as follows:

$$\tau^{(a)}(X_i) = \mathbb{E}[Y_i^{(a)}(1) - Y_i^{(a)}(0) \mid X_i], \tag{18}$$

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, as suggested by the prior advertising literature (Zantedeschi et al. 2017, Gordon et al. 2023). For example, if a smartwatch ad has 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 §5, we validate this assumption in the context of mobile in-app advertising.

### 4.2 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. Consider a superset $\mathcal{A}_i$ of size $A_i$ of ads who are interested in participating in the auction for user $i$'s impressions. For each impression, only a subset of these ads are available due to their budget decisions and the auctioneer's computational reasons (Kim et al. 2024). We consider a setting where the platform randomly draws a subset $\mathcal{A}_{i,t}^{(r)}$ of size $A_r$ to include in the auction for impression $t$ of user $i$.[9] For the set

[9]This random sub-sampling from the full set of candidates is the key 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). Since the specific form of bidder sub-sampling is a stylized abstraction, we perform robustness checks

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 $b_{i,t,a}$ 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:

$$a^*_{i,t} = \arg\max_{a \in \mathcal{A}^{(r)}_{i,t}} b_{i,t,a} \tag{19}$$

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 user $i$, then we have $W^{(a)}_i = 1$. 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:*

$$\pi^{(a)}_i = 1 - \left(1 - \mathbb{1}(a \geq A_r)\frac{A_r\binom{a-1}{A_r-1}}{A\binom{A_i-1}{A_r-1}}\right)^{T_i} \tag{20}$$

*Proof.* See Web Appendix D.1. □

An immediate corollary of Lemma 2 is that the overlap assumption is violated because the propensity score is equal to 0 if $a < A_r$, which is the deterministic no-assignment region. Likewise, for higher bids, the probability quickly converges to 1 as $T_i$ increases, creating a deterministic assignment region. It is worth emphasizing that this insight does not come from the fixed size of $A_r$. Later in §4.4, we relax the assumption on fixed $A_r$ and impose a different micro-structure on ad's availability based on their budget pacing and reserve pricing and arrive at the same insights. 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. 2025). 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

and sensitivity analysis in §4.4 to ensure the validity of our algorithm.

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 §4.3.3 and quantify the economic gains from identifying CATEs through our algorithm by evaluating its targeting performance.

### 4.3 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.[10] Second, we assume an underlying low-rank CATE matrix consistent with applications of matrix factorization models in ad targeting.[11] Third, we use the micro-founded algorithmic ad allocation presented in §4.2 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 $N = 50,000$ users with $D$-dimensional covariates $X_{[N\times D]}$ coming from $\mathcal{N}(0, \sigma_x)$, and $A = 100$ ad campaigns. We present all the simulation details in Web Appendix D.2.

In this section, we first present results on the overlap violation in our application context in §4.3.1. Next, in §4.3.2, we present the results on the performance of our algorithm compared to existing benchmarks in terms of estimation accuracy. Finally, in §4.3.3, we demonstrate how the increased estimation accuracy from our algorithm translates into economic gains from better targeting.

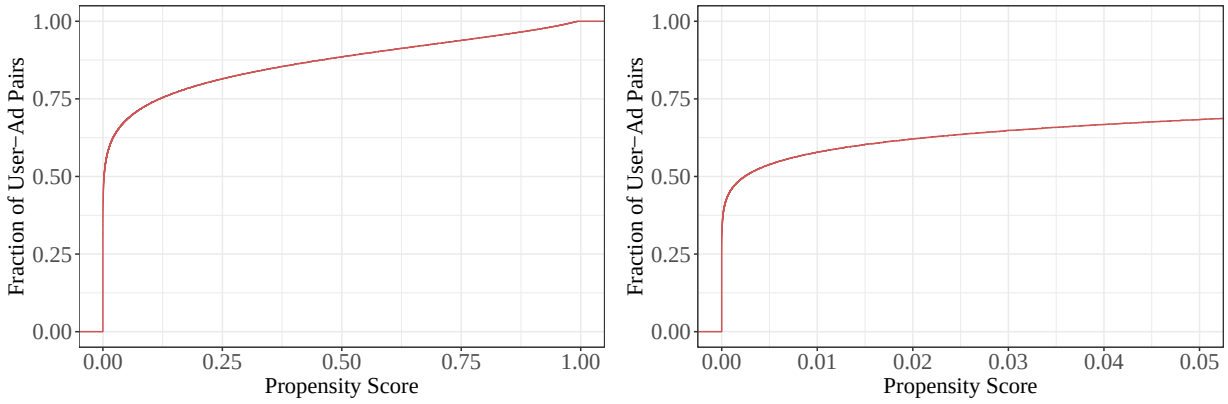
#### 4.3.1 Overlap Violation

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. 2025).[12] For each user, we simulate the number of impression opportunities $T_i$ as a random draw from a discrete uniform distribution from 1 to 50. We assume the same set of ads competing for all impressions, that is, $A_i = A = 100$, and 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.

[10] It is worth emphasizing that calibrating the treatment effects naturally poses a challenge for our algorithm due to the low signal-to-noise ratio in advertising experiments and the fact that the ATE is close to zero for many ads. Therefore, strong performance of our algorithm in this setting would suggest even better performance in settings with stronger treatment signals. We verify this by running simulations with higher magnitudes of treatment effects.

[11] The low-rank structure of the CATE matrix is the fundamental assumption underlying our proposed algorithm. We impose this assumption in our calibrated simulation to demonstrate the algorithm's performance. Later, in §5, we relax this assumption to test its validity in an empirical setting.

[12] 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 or auctions with some form of quality scoring. 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.

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 0. However, as shown in Figure 3a, 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 Web Appendix D.3, we present the GATE for different regions of the data in each study to see how far is the feasible information (probabilistic regions) from the true ATE.

#### 4.3.2 Performance of the Proposed Algorithm

We now examine the performance of our proposed algorithm in overcoming the challenges posed by the overlap violation compared to 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 fails to recover the true ATE. Combining our theoretical propositions in §3.1 and the discussion in the previous section on the overlap violation problem in our application setting, we argue that this is because many observations lie in deterministic no-assignment regions with, on average, lower CATE values. As a result, ignoring these points leads to an overestimation of the ATE. In contrast, 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 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 as described in §3.5. Second, we consider

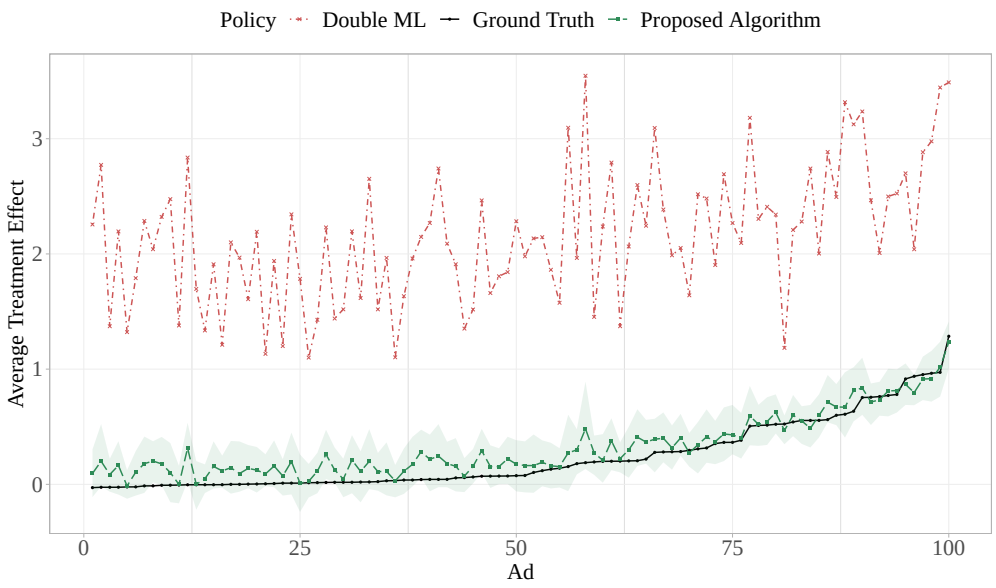



Figure 4. The performance of the proposed algorithm and Double ML in recovering ATE.

| **Method** | **RMSE for ATE** | **RMSE for CATE** |
|---|---|---|
| Algorithm | 0.153 | 1.178 |
| Double ML | 1.969 | 3.342 |
| Algorithm (Rank = 5) | 0.852 | 2.343 |
| Algorithm (Rank = 20) | 0.276 | 1.505 |
| Algorithm (Incorrectly Estimated Propensities) | 1.425 | 2.589 |
| Algorithm (Correctly Estimated Propensities) | 0.773 | 1.244 |

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

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 §4.3.3, we use better ways to evaluate the CATE performance of our algorithm by demonstrating its targeting performance and economic gains, as suggested in §3.5.

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 the cross-validation procedure described in §3.4.1 correctly identifies this rank–without prior knowledge of the true value–by fine-tuning the regularization parameter $\lambda$. However, we still want to see how the algorithm would perform with different rank specifications. We focus on two mis-specified rank cases without cross-validation: (1) a lower than optimal rank ($= 5$) and (2) a 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, models with mis-specified ranks perform worse. However, we note that both models still outperform the Double ML benchmark, suggesting that the algorithm can still exploit cross-ad variation, even with a 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 that determine propensity scores is used, and (2) correctly estimated propensity scores, where all covariates are used to ensure that we have a calibrated propensity model. Two 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 both metrics. Second, we find that although the algorithm with correctly estimated propensity scores performs well, it still underperforms compared to the version with known propensity scores across both metrics.

#### 4.3.3 Economic Gains from the Proposed Algorithm

As discussed earlier, the overlap problem arises because observations in the deterministic no-assignment region tend to have lower CATE values. In our application, overlap violations for a user-ad pair are directly tied to the user's low value to the advertiser. One might therefore argue that, while our algorithm recovers both ATE and CATE, it offers limited practical value to firms—such as advertising platforms and advertisers—since they are likely to disregard the overlap-violating region due to its lower average CATE. In this section, we examine the economic value of decision-making based on our algorithm by addressing the following questions: What are the consequences of ignoring the overlap-violating portion of the data for firms aiming to target effectively? And to what extent can our algorithm improve outcomes for advertising platforms and advertisers?

To answer these questions, we conduct a personalization exercise in which we compare the performance of different algorithms in selecting the targeting population. This allows us to evaluate outcomes under various targeting strategies and assess the economic gains from our proposed algorithm. We follow the evaluation procedure described in §3.5, using the functions $\texttt{Gain}(\alpha)$ and $\texttt{OracleRatio}(\alpha)$. We set $\alpha = 0.1$, meaning each model selects the top 10% of CATE estimates within each study. To assess whether the gains from our algorithm are substantial, we compare it against two benchmarks: (1) Data, where targeting is based on the top 10% of users most likely to be assigned to each ad, as determined by their propensity scores; and (2) Bids, where advertisers target the top 10% of users for whom they have the highest bids. If ignoring the overlap-violating region has no meaningful impact on targeting decisions, we should not observe significant performance differences between our model and these benchmarks. This exercise enables us to quantify the economic gains attributable to our algorithm.

We present the results of this exercise in Table 2. The first column reports the average gains from each targeting policy, while the second column shows the ratio of each policy's average gain to that of the oracle CATE-based targeting model—that is, the first-best performance. Notably, the average gains from our model are substantially higher than those achieved using bids or the algorithmic allocation observed in the data. In other words, by not ignoring the overlap-violating region, our model achieves nearly double the targeting effectiveness. This finding suggests that there are considerable targeting opportunities within the

| **Method** | $\texttt{Gain}_{(0.1)}$ | $\texttt{OracleRatio}_{(0.1)}$ |
|---|---|---|
| Algorithm | 0.474 | 0.981 |
| Bid | 0.254 | 0.525 |
| Data | 0.226 | 0.467 |
| Algorithm (Rank = 5) | 0.390 | 0.807 |
| Algorithm (Rank = 20) | 0.461 | 0.954 |
| Algorithm (Incorrectly Estimated Propensities) | 0.300 | 0.622 |
| Algorithm (Correctly Estimated Propensities) | 0.460 | 0.951 |

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

overlap-violating region, despite its lower average CATE. We further explore this insight in Web Appendix D.4 by visualizing the distribution of CATE values across different regions. Relative to the oracle benchmark, our algorithm recovers approximately 98% of the first-best performance. We also evaluate alternative specifications of our algorithm, including cases where the rank is mis-specified or propensity scores are unknown. Importantly, in all of these cases, our mis-specified algorithm continues to outperform the two benchmarks based on bids and data, even though its performance is weaker than that of the well-specified version.

### 4.4 Robustness Checks and Sensitivity Analysis

In this section, we conduct robustness checks and sensitivity analyses to examine how the performance of our algorithm is affected by changes in various aspects of the data-generating process and to establish its boundary conditions. We begin with a robustness check focused on the auction mechanism used by the platform to allocate ads. In our main analysis, we consider a combinatorial case in which the platform randomly draws a fixed number of ads from the inventory to include in the auction. In Web Appendix D.5.1, we relax the assumption of a fixed auction size $A_r$ and consider an alternative setting in which each ad's availability is determined by its budget pacing decisions and the auction's reserve pricing. We find that all of our qualitative results hold in this alternative environment, with only minor quantitative differences, suggesting that our main setting is a good approximation of settings where the probabilistic assignment stems from advertisers' budget-pacing decisions.

In particular, we conduct sensitivity analyses on four key components: (1) the size of the advertising sample in each auction, $A_r$; (2) the rank of the underlying CATE matrix; (3) the variance of the underlying CATE matrix; and (4) the correlation between advertisers' CATE values and their bids. The full results are presented in Web Appendix D.5.2. We begin by increasing $A_r$, which induces more deterministic assignment and makes the feasibility matrix sparser. As expected, the performance of our algorithm deteriorates as the assignment becomes more deterministic. However, it still outperforms conventional methods across all evaluation metrics, underscoring the value of transferring information across studies. Next, we vary the rank of the underlying CATE matrix. Higher-rank matrices pose greater challenges for the matrix completion algorithm, so we expect better performance in low-rank scenarios. Our findings confirm this intuition: as the

underlying rank increases, performance declines, but the algorithm continues to outperform conventional methods by leveraging information in the observed entries to recover the signal in the missing ones. Third, we assess the effect of increasing the variance of the underlying CATE matrix. As expected, estimation error increases with variance. Interestingly, however, the targeting performance of our algorithm remains robust, even at high levels of variance. This is because the value of targeting grows as the CATE matrix becomes more heterogeneous (Rafieian and Zuo 2024), which helps offset the decline in estimation accuracy. Finally, we vary the correlation between advertisers' CATE values and their bids. In all cases, our algorithm consistently outperforms the benchmarks.

## 5 Empirical Validation Exercise

In this section, we use real data from mobile in-app advertising to provide empirical validation for our algorithm. Our primary goal is to relax the rank assumption and assess its validity in an applied setting. We also aim to demonstrate the performance and economic value our algorithm delivers in environments with overlap violations. We begin by describing the empirical context in §5.1. Next, in §5.2, we outline the empirical framework, define the target estimand, and present our identification strategy. We report validation results in §5.3 and evaluate the performance of our algorithm in §5.4.

### 5.1 Setting and Data

The empirical setting for our study is mobile in-app advertising, an industry that has experienced sustained growth over the past decade. We use impression-level data from a leading mobile in-app advertising network in a large Asian country, which held over 85% market share at the time of the study. The dataset includes over one billion ad impressions. Our sample is identical to that used in Rafieian (2023); we refer readers to that paper for detailed information on sampling. In this sample, we observe 6,357,389 impressions from 327 distinct ads displayed within a messenger app. For each impression, we observe a rich set of covariates, including demographic features such as province, latitude, longitude, smartphone brand, mobile service provider (MSP), and connectivity type. We also observe several historical and session-level features constructed from both short- and long-term user activity (e.g., the variety of past ads seen, and the number of past impressions). Table 3 presents summary statistics on user behavior within the messenger app.[13] The data reveal substantial heterogeneity in user behavior: the median user participated in 8 sessions, was exposed to 37 ad impressions across 10 distinct ads, and made 1 click.

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

[13] This corresponds to Table 2 in Rafieian (2023), which includes all user impressions. Table 3, by contrast, is restricted to impressions within the messenger app only.

| Variable | Mean | SD | Min | Median | Max |
|---|---|---|---|---|---|
| **Number of Sessions** | 15.71 | 20.64 | 1.00 | 8.00 | 253.00 |
| **Number of Impressions Seen** | 91.25 | 159.87 | 1.00 | 37.00 | 4655.00 |
| **Variety of Ads Seen** | 13.47 | 11.79 | 1.00 | 10.00 | 114.00 |
| **Number of Clicks Made** | 1.38 | 2.08 | 0.00 | 1.00 | 19.00 |
| **Click-through Rate (CTR)** | 0.03 | 0.06 | 0.00 | 0.01 | 1.00 |

Table 3. Summary statistics of the user behavior in the messenger app.

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.

## 5.2 Empirical Framework

### 5.2.1 Effect of Focal Ad vs. Platform Ad

We start by defining the causal effect of interest. In our setting, each impression is characterized by 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 platform ad $a^{(p)}$ that the platform can serve at any time.[14] 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 platform 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:

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

In our empirical analysis, we aim to recover the parameter defined in Equation (21) and present results demonstrating our ability to estimate this parameter across various settings. A few points are worth noting in defining our target estimand. First, our unit of analysis is a targeting profile rather than a user. This implies that a single user can be represented by multiple targeting profiles, which aligns more closely with the contextual bandits literature, where individual contexts arrive dynamically and a user's context can evolve over time. Second, the click outcome we focus on serves as a key conversion metric. Since all ads are for mobile apps, clicks are closely tied to app installs. Moreover, the platform operates a pay-per-click auction, meaning each click directly contributes to platform revenue. Third, we use the causal estimand above primarily because it is well-defined. However, the comparison with the platform ad also serves as a meaningful benchmark for assessing the level of engagement an ad generates.

### 5.2.2 Definition of CATE Matrix

The causal effect of the focal ad $a^*$ relative to the platform ad $a^{(p)}$ 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 (21) for other ads in our

[14] It is worth emphasizing that in the data, the platform 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.

inventory. That is, for any ad $a \in \mathcal{A}$, we define the following CATE:

$$\tau^{(a)}(X_i) = \mathbb{E}[Y_i^{(a)}(1) - Y_i^{(a)}(0) \mid X_i], \tag{22}$$

where $Y_i^{(a)}(1)$ and $Y_i^{(a)}(0)$ are potential outcomes for cases where ad $a$ and platform ad $a^{(p)}$ 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.

### 5.2.3 Identification Strategy

We now discuss our identification strategy for estimating the CATE of each ad $a$ as a function of targeting profiles. Consider the task of estimating the CATE for ad $a$. We need to select impressions that (1) satisfy the causal inference assumptions, such as overlap and unconfoundedness, and (2) are either allocated to ad $a^*$ (treatment condition) or the platform ad $a^{(p)}$ (control condition). We use the randomness induced by the quasi-proportional auction as our main identification strategy. Let $q^{(*)}$ and $q^{(p)}$ denote the quality-adjusted bids for the focal ad $a^*$ and the platform ad $a^{(p)}$ for a given impression. If both ads $a^*$ and $a^{(p)}$ participate in the auction for that impression, their corresponding winning probabilities will be $q^{(*)}/(\sum_{j\in\mathcal{A}} q^{(j)})$ and $q^{(p)}/(\sum_{j\in\mathcal{A}} q^{(j)})$. This implies that the ratio only depends on $q^{(*)}$ and $q^{(p)}$. Therefore, for the set of impressions where both ads $a^*$ and $a^{(p)}$ participate and one of them wins, we have both the unconfoundedness and overlap assumptions satisfied because we know that the non-deterministic probability of treatment assignment (assignment to ad $a^*$) is $q^{(*)}/(q^{(*)} + q^{(p)})$ and the probability of control assignment (platform ad $a^{(p)}$) is $q^{(p)}/(q^{(*)} + q^{(p)})$. 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; this greatly stabilizes the estimation of causal parameters.

In summary, to estimate CATE for each ad $a$, we can select a sample of impressions allocated to either ad $a$ or the platform ad $a^{(p)}$ where both ads participated in the auction (to satisfy overlap), and estimate CATE $\widehat{\tau}^{(a)}(\cdot)$ using any CATE estimator as a function of the targeting profile $x$. However, it is worth emphasizing that the estimates are not accurate for all targeting profiles. In particular, if one ad does not participate in the auction for an impression with targeting profile $x'$, the assignment probability will be 0, which violates the overlap assumption, indicating that our estimate $\widehat{\tau}^{(a)}(x')$ no longer has statistical properties such as unbiasedness and consistency. Therefore, it is crucial not to use our CATE estimates for the overlap-violating regions. The bright side in our empirical setting is that we can identify the overlap-violating impressions for any ad $a$, because 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. We ensure that the sample we use to estimate CATE for each ad $a$ satisfies the overlap and unconfoundedness assumption, and we avoid predicting CATE for overlap-violating regions.

### 5.2.4 Empirical Estimation of the Underlying CATE Matrix

We now discuss our approach to estimating the underlying CATE matrix. Figure 5 outlines the steps in our estimation procedure. We present these steps in our empirical approach as follows:

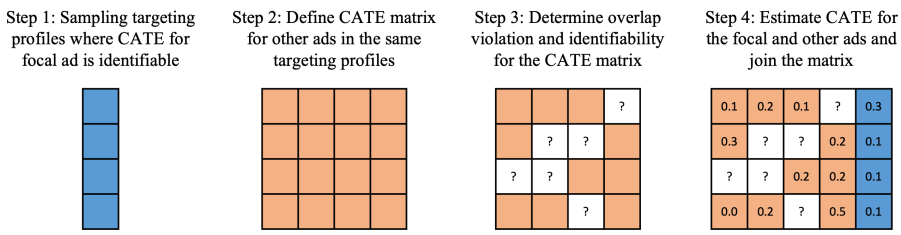
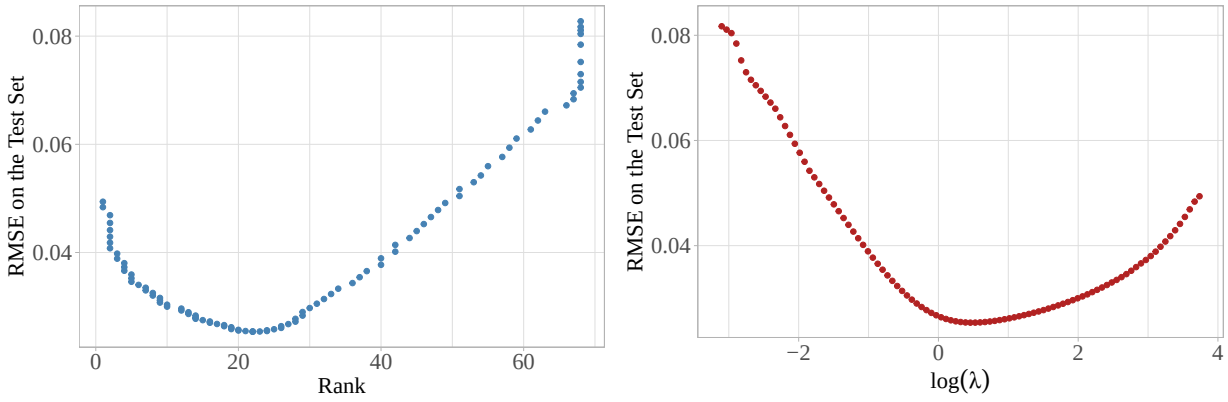



Figure 5. Step-by-step procedure for estimation of the underlying CATE matrix

- *Step 1:* We first sample targeting profiles to estimate the CATE for the focal ad, as defined in Equation (21). To do this, we draw a random sample of 100,000 impressions from the full set of impressions awarded to either the focal ad or the platform ad, conditional on both having a non-zero propensity score. This ensures we can consistently estimate the CATE for these impressions and establish a ground truth. In Web Appendix E.1, we present a complete list of features used to characterize a targeting profile.
- *Step 2:* In the second step, we sample ads and construct the CATE matrix. To do this, we select ads from the full set of 327 ads that have at least 10,000 impressions in our data, ensuring reasonable statistical power for our estimation. We include 68 ads in addition to the focal and platform ads, resulting in a total of 70 ads that collectively account for 94% of all impressions observed in our data. More detailed information on the cumulative share of impressions allocated to the top ads is provided in Web Appendix E.2.
- *Step 3:* We determine the feasibility of CATE estimation for each ad in the sample of targeting profiles. Naturally, there are impression-ad pairs for which the overlap assumption is violated in our data, due to targeting decisions. As discussed in §5.2.3, our CATE estimates for an ad are not valid in an impression if that ad could never have been shown in those impressions (deterministic assignment). We replace entries where the overlap assumption for an ad is violated with question marks to ensure we only use estimates with proven consistency. This sets the scope for our analysis and highlights the presence of both probabilistic and deterministic assignment regions in real applications. In Web Appendix E.3, we visualize the missingness pattern in our CATE matrix.
- *Step 4:* In the final step, we estimate CATE for each ad and predict CATEs for feasible entries from the previous step. To estimate CATE for each ad $a$, we first draw a sample of impressions allocated to either ad $a$ or the platform ad $a^{(p)}$ that satisfy the overlap assumption. As such, the estimation sample for each ad is different from the set of targeting profiles selected for the focal ad. However, we can use the resulting estimated function $\widehat{\tau}^{(a)}(\cdot)$ for each ad $a$ to predict CATE for the feasible entries in the CATE matrix. To estimate CATE for the focal ad, we use the sample of targeting profiles illustrated in Figure 5. For CATE estimation, we use R-learner with XGBoost to estimate the nuisance functions with a two-fold cross-validation (Nie and Wager 2021). The reason for this modeling choice was the performance of XGBoost R-learner in our context. In Web Appendix E.4, we provide more details on the sample size and the CATE estimation procedure for each ad and present results on the distribution of CATE estimates for the focal ad.




(a) Performance with different ranks

(b) Performance with different $\lambda$

Figure 6. RMSE of low-rank approximation of the estimated CATE matrix on the test data using SoftImpute

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 use a rich set of features for CATE estimation to capture the high-dimensionality of real application settings.

### 5.3 Validation of Low-Rank Assumption

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. A simple way to test this assumption is to split the entries in the CATE matrix into training and test entries and apply our SoftImpute algorithm to find the best low-rank approximation. If the rank of the best approximated matrix is high, we determine that the underlying matrix is not low-rank. We hold out 10% of the entries for testing the performance and train the SoftImpute algorithm on the remaining 90% of the entries, with maximum rank equal to 69 (full rank) and using a grid of 100 values for the regularization parameter $\lambda$. Figure 6 shows the performance of matrix factorization approaches with different ranks (Figure 6a) and regularization parameters (Figure 6b). As shown in Figure 6a, the best-performing model on the test set has a rank of 22, offering validation for the low-rank assumption in our actual empirical setting.

The ability to select the rank optimally through a validation procedure offers great convenience for researchers. For example, if the best rank in the exercise above is high (e.g., 60), it indicates that our algorithm's ability to recover the true causal parameters is limited, as CATEs across studies are not strongly related. Another key benefit of the validation procedure is that it allows researchers to incorporate their desired objectives and qualitatively assess the quality of the low-rank approximation. For instance, in our exercise, one could set a threshold for RMSE based on the overall uncertainty in CATE estimates and evaluate whether the RMSE of the best-performing model is acceptable. Similarly, a manager could set an objective threshold for targeting performance to qualitatively assess how well the low-rank approximation meets their goals.

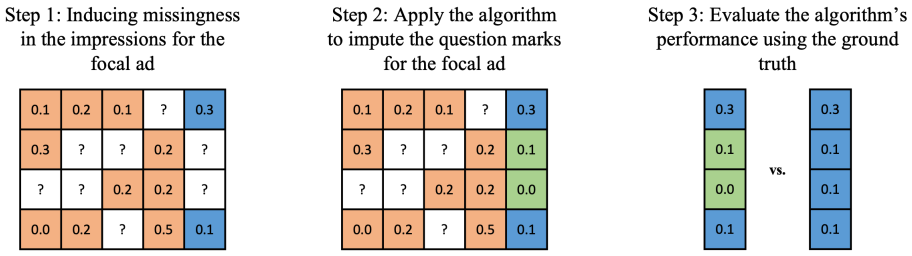



Figure 7. Step-by-step procedure for evaluation of our algorithm

### 5.4 Performance of Our Algorithm

We now turn to the study of the focal ad and evaluate our algorithm to provide an empirical proof of concept. First, we describe our evaluation procedure in §5.4.1. Next, in §5.4.2, we present the results on our algorithm's performance in this empirical exercise.

#### 5.4.1 Evaluation Procedure

Our goal is to evaluate how our algorithm performs when the overlap assumption is violated for the focal ad. We leverage the fact that we have ground-truth estimates for the focal ad, given the unique setting of our empirical exercise. This allows us to induce realistic missingness in the CATE estimates for the focal study and then apply our algorithm. Figure 7 illustrates the steps we take to evaluate the algorithm's performance, which are summarized as follows:

- Step 1 induces a missingness pattern in the CATE estimates for the focal ad. As discussed earlier in §4, algorithmic ad allocation can produce a mixture of deterministic and probabilistic patterns due to the deterministic nature of commonly used auctions (e.g., second-price).[15] We consider a similar scenario to induce realistic missingness patterns. In standard auctions used by advertising platforms, the platform estimates the CTR for each ad and requests a bid-per-click from each advertiser. The auction then allocates the impression to the ad with the higher quality-adjusted bid, calculated as the product of the bid and the estimated CTR. Let $\widehat{\text{CTR}}_i^{(*)}$ and $\widehat{\text{CTR}}_i^{(p)}$ denote the platform's estimated CTRs for the focal ad and platform ad in impression $i$, respectively. In our data, both ads submit the same bid-per-click, so we assume they have the same bid in the simulated auction format.[16] As such, we consider an auction environment where the impression is allocated to the ad with the higher estimated CTR.

  We use XGBoost to estimate the CTR for each ad in each impression $i$ and measure the difference $\widehat{\delta}_i = \widehat{\text{CTR}}_i^{(*)} - \widehat{\text{CTR}}_i^{(p)}$ between them. For assignment, we assume that if $-0.005 < \widehat{\delta}_i < 0.005$, the assignment is probabilistic; otherwise, the impression is deterministically allocated to the ad with the higher estimated CTR. This assumption is reasonable because many advertising platforms use posterior sampling approaches—such as Thompson Sampling—to estimate CTR, where draws from the posterior distribution can lead to probabilistic assignments (Rashid et al. 2025). In Web Appendix E.5, we visualize the distribution of $\widehat{\delta}$, which results in 39.3% of impressions falling in the probabilistic region and the

[15] Even in a probabilistic auction, such as the one used in our empirical setting, this issue arises due to advertisers' targeting and budget decisions, resulting in missing entries in the CATE matrix, as shown in Figure 5.

[16] We acknowledge that this assumption may not hold in practice; however, our goal in this section is to provide a proof of concept, not to conduct a full counterfactual analysis.

remaining 60.7% in the deterministic region.

- Step 2 applies a matrix completion algorithm as in Algorithm 1 to complete the CATE matrix. Naturally, this process completes all the entries in the matrix, but we are only interested in the completed entries for the focal ad. We follow the validation procedure to choose the optimal regularization parameter $\lambda$ and complete the matrix using SoftImpute.
- Step 3 uses the imputed CATE values for the focal ad and compares them with the ground-truth estimates using the evaluation metrics introduced in §3.5. In particular, we use the RMSE of the treatment effect estimates (ATE and CATE) to evaluate accuracy performance, and $\texttt{Gain}(\alpha)$ and $\texttt{OracleRatio}(\alpha)$ to examine targeting performance.

### 5.4.2 Results

We now present the results on the performance of our algorithm. We compare it against three benchmarks: (1) *Double Machine Learning (DML)*, which estimates the GATE for the probabilistic region of the focal ad; (2) *Ad Allocation Algorithm*, which estimates causal parameters using $\widehat{\delta}_i = \widehat{\text{CTR}}_i^{(*)} - \widehat{\text{CTR}}_i^{(p)}$; and (3) *Ground Truth*, which uses CATE estimates assuming all impressions fall within the probabilistic region and serves as our oracle.

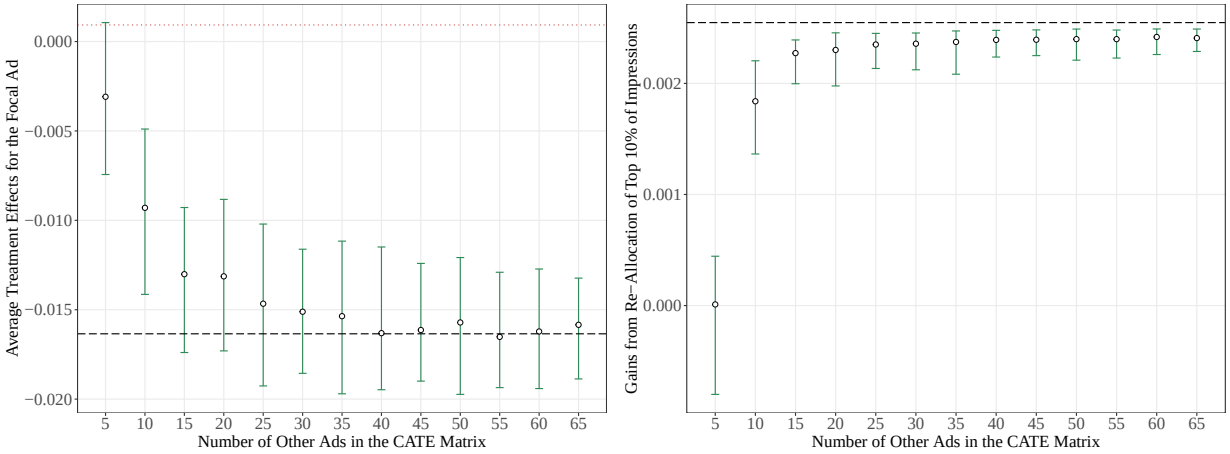
We present our results in Table 4. We begin by focusing on the ATE estimates provided by all models. The ground truth estimate for the ATE of the focal ad in the set of targeting profiles in our study is $-0.01634$, which indicates that showing the focal ad results in $0.01634$ lower CTR compared to the platform ad, on average.[17] Our algorithm's ATE estimate is $-0.01679$, demonstrating its great performance in recovering the true ATE. However, we find that benchmarks fail to correctly estimate the true ATE. Notably, we find that *Double ML* is largely biased and even misses the sign of the ATE. The *Ad Allocation Algorithm* performs better, producing an ATE estimate of $-0.01860$, but still misses the true value. The superior performance of our algorithm compared to the scores from the ad allocation algorithm becomes even more apparent in the second column of Table 4, where we measure the RMSE of the CATE estimates.

We then focus on the targeting performance and the economic gains from our algorithm using $\texttt{Gain}_{(0.1)}$, which measures the gain from a targeted re-allocation of top 10% of impressions. Using the ground-truth estimates, we find that re-allocating the top 10% results in a $0.00255$ unit increase in CTR for a sample of 100,000 impressions—equivalent to a $10.89\%$ increase, given the platform ad's baseline CTR of $0.02338$. Re-allocation based on our algorithm's estimated CATEs yields similar performance, with a $0.00253$ unit increase in CTR, corresponding to a $10.81\%$ improvement over full allocation to the platform ad. In contrast, targeting the top 10% of impressions using $\widehat{\delta}$ values from the ad allocation algorithm leads to a smaller gain of $0.00101$, or $4.33\%$. The fourth column shows the Oracle Ratio of the gains from each model compared to the ground truth and highlights the remarkable performance of our algorithm in recovering 99.24% of gains. In Web Appendix E.6, we explore alternative re-allocation strategies and show that our algorithm can deliver even greater gains. Overall, our findings underscore a key managerial insight: while algorithmic decision-making improves outcomes, it often overlooks high-value opportunities in regions of deterministic assignment. Managers can use pre-existing experimental data to construct CATE factors and recover these

[17] To put this number in perspective, the CTR for the platform ad in our sample is $0.02338$, so an ATE of $-0.01634$ translates to a ratio of $-0.01634/0.02338 = -0.6988$, equivalent to a $-69.88\%$ lower CTR.

| **Method** | ATE | $\texttt{RMSE}_{\text{CATE}}$ | $\texttt{Gain}_{(0.1)}$ | $\texttt{OracleRatio}_{(0.1)}$ |
|---|---|---|---|---|
| Proposed Algorithm | $-0.01679$ | 0.00377 | 0.00253 | 99.24% |
| Double ML | 0.00093 | – | – | – |
| Ad Allocation Algorithm ($\widehat{\delta}$) | $-0.0183$ | 0.03711 | 0.00101 | 39.75% |
| Ground Truth | $-0.01634$ | 0 | 0.00255 | 100% |

Table 4. Performance of different models in terms of estimation and targeting.




(a) Accuracy Performance in Estimating ATE

(b) Targeting Performance in Improving Outcomes

Figure 8. Performance of our algorithm with different number of ads in the CATE matrix

missed opportunities.

To better understand the strong performance of our algorithm, we perform a simple analysis: we vary the number of other ads in the CATE matrix. Our analysis in Table 4 uses 68 other ads. We first order the ads and then progressively add 5 ads at a time to re-run our algorithm. We present the results from this exercise in Figure 8. Figure 8a shows the performance of our algorithm in recovering the true ATE. The dotted red line is the Double ML estimate, and the dashed black line is the ground truth estimate. We follow the uncertainty quantification procedure in §3.3 and present the median estimate and the 95% confidence intervals, using 100 draws for each case. As shown in Figure 8a, our algorithm performs better in terms of both accuracy and precision as we include more ads.[18] However, we notice that the performance stabilizes after including only 35–40 ads. Figure 8b illustrates the same pattern for the targeting performance of our algorithm. The value created by our algorithm stabilizes with around 40 ads, while adding more ads helps reduce the uncertainty around the estimates.

We further examine the robustness of our algorithm by adding irrelevant ads. We randomly generate CATE estimates for 20 pseudo-ads. Theoretically, adding independent columns to a matrix will increase the rank of that matrix. As such, we expect our algorithm to learn the increased rank and tune out the irrelevant columns added to the matrix. As shown in Web Appendix E.7, we find that the best rank approximation

[18] As shown in Figure 8a, the estimate does not capture the true parameter with only a few ads, which is expected since our resampling method does not capture the epistemic structural uncertainty about the sufficiency of information from other ads.

selected by our validation procedure is 34, and that our algorithm performs remarkably well as it effectively ignores the information in these new columns. On the contrary, when we include only the set of 20 irrelevant ads, we find that our algorithm fails to recover the true ATE, because the set of irrelevant ads does not provide any useful information.

In summary, our analysis shows that our algorithm can effectively exploit similarities between ads to overcome a severe case of overlap violation, where the Double ML estimator even misses the sign of the true estimates. We stress that the applicability of our algorithm is context-specific. For example, if CATEs are independent across ads, our algorithm cannot use any cross-study information that would be helpful for its primary task. However, many empirical settings exhibit similarity patterns and correlation structures akin to our problem, making our algorithm suitable for a wide range of applications. In Web Appendix §E.8, we further investigate the correlation structure between CATE estimates in our study and demonstrate that, despite similarities, these ads differ significantly based on their weights on various factors in our setting. Furthermore, a key feature of our algorithm is that its parameters can be optimally set for each context. Thus, one could specifically test whether the conditions required for our algorithm's applicability are satisfied.

## 6 Managerial Implications

Our work offers several implications for managers and practitioners. Clearly, the most direct set of implications from our work are for advertising platforms and advertisers who are interested in advertising measurement without using a randomized experiment. As discussed in the paper, advertising auctions only create partial randomization in the allocation of ads, leading to major identification issues in ad measurement. Our algorithm first estimates the identifiable pieces and then exploits the similarities in the joint space of units and ads to recover ad measurement for unidentifiable parts. Advertising platforms and advertisers can use our proposed algorithm to improve their ad measurement in settings where experimentation is not feasible or too costly.

More broadly, the application of our proposed algorithm extends beyond the advertising measurement problem. 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 unit level and face identification issues due to a violation of ignorability assumptions required for consistent estimation of causal parameters. Specifically, most settings with algorithmic delivery of interventions fall under this broad category, as these are settings where one could manipulate the assignment to an intervention, but there are already algorithms in place to facilitate this assignment problem.

The first class is the set of problems where the platform has full control over the assignment policy. Examples of these cases include assignment of promotions (e.g., Uber's promotion assignment for future rides), push notifications (e.g., Fitbit's notification on body activity), and rewards (e.g., gaming app offering free coin to some users). In these cases, the decisions are often made by algorithms that induce limited randomization at the local level, because of the arbitrariness in cutoffs used (Shi et al. 2022) or the use of bandit algorithms that explore actions (Swaminathan and Joachims 2015). In Web Appendix F, we present simulations in such cases where the probabilistic assignment comes from the posterior uncertainty in algorithmic scores, and highlight the value of our algorithm.

The second class includes settings where the platform determines the allocation rule but does not fully control the final assignment because other agents influence the outcome. Examples include an advertising platform selecting which ad to show, a social network curating a news feed from user-generated content, or a platform ranking products for a search query. In such cases, the platform builds an algorithm that incorporates agent inputs—such as bids, content, or availability—into the assignment process. Consequently, the platform only partially controls the assignment. These settings generate both deterministic and probabilistic regions. A prime example of the application of this class of problems is online advertising auctions.

Our algorithm creates value in these cases in two distinct ways. First, our algorithm creates value in settings with no randomization through experimental design, but where algorithmic allocation induces a mixture of probabilistic and deterministic assignments. In such cases, the algorithm enables platforms to take advantage of their existing observational data to refine their targeting algorithms. Our analyses in §4 and §5 show how the proposed approach allows platforms to use their existing data in advertising auctions to improve causal measurement and targeting decisions. Second, our algorithm creates value by reducing experimentation costs even when experimentation is feasible. Platforms can use prior experimental data to learn CATE factors in a low-rank space, which in turn allows them to estimate CATEs for new interventions with limited randomization in allocation. The value of this feature is more pronounced in the first class of problems where the platform has full control over the treatment assignment, because experimentation is more feasible and platforms can more easily achieve the first set of benefits offered by our algorithm through small-scale experimentation. Together, our algorithm offers an important tool for digital platforms and managers implementing algorithmic decision-making.

Lastly, we stress that our algorithm should not be seen as a replacement for experimentation and A/B testing and firms should use experimentation whenever it is cost-effective. 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.

## 7 Conclusion

Digital platforms increasingly rely on algorithmic decision-making. We study a canonical example—online advertising auctions—and examine the problem of causal advertising measurement, a longstanding goal in marketing. While randomized experiments are the gold standard, they are often prohibitively expensive, motivating observational approaches that use existing platform data. Such methods rely on strong ignorability of treatment assignment, which requires both unconfoundedness and overlap. While much of the prior literature focused on the former, the latter received considerably less attention. We show that in advertising platforms, overlap frequently fails due to the deterministic nature of algorithmic allocation, leading to an impossibility result for causal measurement in single-ad settings. We recast this identification failure as a missing data problem and propose a matrix completion approach for multi-ad environments. Using calibrated simulations and field data, we demonstrate that our method reliably recovers population-level ad

effects and delivers substantial gains in targeting performance relative to existing approaches.

In summary, this paper makes several contributions to the literature. Methodologically, we introduce a novel machine learning approach that frames the identification challenge as a missing data problem, combining heterogeneous treatment effect estimation with matrix completion to recover treatment effects. Our proposed algorithm is fairly general and applies to a variety of contexts beyond advertising measurement. From a substantive and practical perspective, we show that our algorithm corrects estimates of causal parameters and results in substantial gains in targeting performance. In particular, our algorithm is useful as an observational tool for causal measurement and targeting aimed at reducing experimentation costs, and it is applicable when the platform observes multiple treatments for the same units and the space of treatment effects is low-rank, conditions that apply to most major advertising platforms.

Nevertheless, our paper has limitations that open up fruitful avenues for future research. First, as emphasized earlier, absent cost considerations, experimentation is always preferable to our algorithm. Future work could investigate how our approach can be integrated with A/B testing and establish theoretical and empirical regret bounds reflecting its improved sample efficiency. Second, given the applied focus of our paper, we employ a general matrix completion approach that must be chosen via model selection. This generality, however, prevents us from providing sharp theoretical guarantees. Future research could extend our work by deriving statistical guarantees for specific matrix completion methods. Finally, our framework assumes that CATEs are linear functions of underlying latent factors. Future work could relax this linearity assumption and develop nonlinear extensions, for example using variational autoencoders (VAEs).

**Disclosure Statement**

Author(s) have no competing interests to declare.

## 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.

B. R. Gordon, R. Moakler, and F. Zettelmeyer. Predictive incrementality by experimentation (pie) for ad measurement. *arXiv preprint arXiv:2304.06828*, 2023.

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.

Y. Gui, R. Barber, and C. Ma. Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36:4820–4844, 2023.

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.

S. Kim, M. Dornfeld, and T. Zhang. Introducing a global retrieval ranking model in the ads funnel. Reddit Engineering blog post, 2024. URL https://www.reddit.com/r/RedditEng/comments/1d2wfsd/introducing_a_global_retrieval_ranking_model_in/. Accessed: 2025-06-12.

Y. Koren, S. Rendle, and R. Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142, 2021.

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.

L. M. Lodish, M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson, and M. E. Stevens. How tv advertising works: A meta-analysis of 389 real world split cable tv advertising experiments. *Journal of marketing research*, 32(2):125–139, 1995.

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.

M. Ostrovsky and M. Schwarz. Reserve prices in internet advertising auctions: A field experiment. *Journal of Political Economy*, 131(12):3352–3376, 2023.

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 and S. Zuo. Personalization, algorithmic dependence, and learning. *Cornell SC Johnson College of Business Research Paper (forthcoming)*, 2024.

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

M. Rashid, O. Rafieian, and S. Ghili. Auctions meet bandits: An empirical analysis. *arXiv preprint arXiv:2508.21162*, 2025.

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.

B. T. Shapiro, G. J. Hitsch, and A. E. Tuchman. Tv advertising effectiveness and profitability: Generalizable results from 288 brands. *Econometrica*, 89(4):1855–1879, 2021.

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, 2020.

A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.

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, and C. Carrion. Online causal inference for advertising in real-time bidding auctions. *Marketing Science*, 44(1):176–195, 2025.

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

D. Zantedeschi, E. M. Feit, and E. T. Bradlow. Measuring multichannel advertising response. *Management Science*, 63(8):2706–2728, 2017.

Y. Zhao and M. Udell. Matrix completion with quantified uncertainty through low rank gaussian copula. *Advances in Neural Information Processing Systems*, 33:20977–20988, 2020.

# Web 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:

$$Y_i(w) = \mu_0(X_i) + \tau^*(X_i)w + \epsilon_i(w), \tag{A.23}$$

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:

$$Y_i - m(X_i) = (W_i - \pi(X_i))\,\tau^*(X_i) + \epsilon_i(W_i), \tag{A.24}$$

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:

$$\tau^*(\cdot) = \underset{\tau}{\operatorname{argmin}}\, \mathbb{E}\left[\left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau(X_i)\right)^2\right]. \tag{A.25}$$

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:

$$\widehat{\tau}_{\text{IPS}} = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i \left(\frac{W_i}{\pi(X_i)} - \frac{1 - W_i}{1 - \pi(X_i)}\right)\right), \tag{A.26}$$

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 0, 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 $\alpha_r$ fraction of observations with the probabilistic assignment.

## B Proofs for Propositions

### 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:

$$\underset{\tau}{\operatorname{argmin}} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \xrightarrow{p} \tau_r. \tag{A.27}$$

We now want to show that the RHS of Equation (A.27) is the same as what any methods optimizing Equation (A.25) would estimate. We can write:

$$\begin{aligned}
\widehat{\tau} &= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&\quad + \sum_{i \notin \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 + \sum_{i \notin \mathcal{I}_r} \left(Y_i - m(X_i)\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2 \\
&= \underset{\tau}{\operatorname{argmin}} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} \left(Y_i - m(X_i) - (W_i - \pi(X_i))\,\tau\right)^2,
\end{aligned} \tag{A.28}$$

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 $\tau$, and the sixth line changes $1/N$ to $1/N_r$ because it is invariant of $\tau$. Now if we combine the result of Equation (A.28) with that of Equation (A.27), 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:

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)] &= \mathbb{E}[Q_i\tau(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Q_i\tau(X_i) \mid Q_i]] \\
&= \mathbb{E}[Q_i\mathbb{E}[\tau(X_i) \mid Q_i]] \\
&= P(Q_i = 1)(1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] + P(Q_i = 0)(0)\mathbb{E}[\tau(X_i) \mid Q_i = 0] \\
&= P(Q_i = 1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] \\
&= P(\pi(X_i) = 1)\mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]
\end{aligned} \tag{A.29}$$

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:

$$\begin{aligned}
\tau_1 &= \mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1] \\
&= \frac{P\left(\pi(X_i) = 1\right)\mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]}{P\left(\pi(X_i) = 1\right)} \\
&= \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&\geq \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]\mathbb{E}[\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&= \mathbb{E}[\tau(X_i)] \\
&= \tau^*,
\end{aligned} \tag{A.30}$$

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

## 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:

$$\widehat{\mathbf{M}} = \arg\min_{\mathbf{M}} \frac{1}{2}\|\mathbf{W} \odot (\mathbf{X} - \mathbf{M})\|_F^2 + \lambda\|\mathbf{M}\|_*, \tag{A.31}$$

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, $\lambda$ 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 $\lambda$ controls the amount of shrinkage applied to the singular values, which determines the rank of the resulting matrix. A larger $\lambda$ 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 Validation Procedure

In this section, we present step-by-step details on our validation procedure to set the hyper-parameters for the SoftImpute algorithm. Suppose we have an incomplete matrix $\mathcal{T}_{[n\times m]}$. As discussed in §3.3, the SoftImpute algorithm requires two parameters: (1) regularization parameter $\lambda$ that controls the rank of estimated matrix by regularizing the nuclear norm of the matrix, and (2) maximum rank, which is the maximum allowable rank of the matrix. In our setting, matrix $F_{[n\times m]}$ determines the missingness pattern in matrix $\mathcal{T}_{[n\times m]}$. We further define $\Omega(\mathcal{T})$ as the set of $(i,j)$ pairs that are observed in matrix $\mathcal{T}_{[n\times m]}$, that is, $\Omega(\mathcal{T}) = \{(i,j) \mid F_{i,j} = 1\}$. We take the following steps to tune the hyper-parameters of the model.

- Set the maximum rank based on the domain knowledge. If there is no prior information is available, set the maximum rank as $\min(n, m)$, which is the maximum possible rank of matrix $\mathcal{T}_{[n\times m]}$.
- Let $\widetilde{\mathcal{T}}_{[n\times m]}$ denote the following matrix:

$$\widetilde{\mathcal{T}}_{i,j} = \begin{cases} \mathcal{T}_{i,j} & \text{if } (i,j) \in \Omega(\mathcal{T}) \\ 0 & \text{if } (i,j) \notin \Omega(\mathcal{T}) \end{cases} \tag{A.32}$$

  Use Singular Value Decomposition and set $\lambda_{\max}$ as the maximum singular value.
- Create a grid of $\Lambda$ ranging from zero to $\lambda_{\max}$. Without loss of generality, let $\lambda_1 < \lambda_2 < \cdots < \lambda_{\max}$.
- Split the set of observed entries $\Omega(\mathcal{T})$ into two groups $\Omega_{\text{train}}(\mathcal{T})$ and $\Omega_{\text{validation}}(\mathcal{T})$, such that $\Omega_{\text{train}}(\mathcal{T}) \cup \Omega_{\text{validation}}(\mathcal{T}) = \Omega(\mathcal{T})$ and $\Omega_{\text{train}}(\mathcal{T}) \cap \Omega_{\text{validation}}(\mathcal{T}) = \varnothing$, and $\Omega_{\text{validation}}(\mathcal{T})$ contains a random $\alpha$ fraction of all observed entries. In our applications, we set $\alpha = 0.1$.
- Start with $\lambda_1$ and run the SoftImpute algorithm on $\mathcal{T}_{\text{train}}$. Record the model output as $\widehat{\mathcal{T}}_{\lambda_1}$ and its performance on the validation set of entries $\Omega_{\text{validation}}(\mathcal{T})$. For better computational performance, warm start the model with $\lambda_t$ with the model in the previous step $\widehat{\mathcal{T}}_{\lambda_{t-1}}$. Record the validation performance for all values of $\lambda \in \Lambda$.
- Select the best-performing $\lambda^*$ based on the performance on the validation set as follows:

$$\lambda^* = \underset{\lambda\in\Lambda}{\operatorname{argmin}}\, \mathcal{L}_{\text{validation}}(\widehat{\mathcal{T}}_{\lambda})$$

  Please note that the loss function $\mathcal{L}$ can be anything based on the researcher's objective. A conventional choice is the RMSE of estimated entries in the validation set.

### C.3 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:

$$\mathcal{T} = XB^T, \tag{A.33}$$

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 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 0, 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 $A_r/A_i$. 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 $A_i - 1$. As such, the probability of ad $a \geq A_r$ winning an impression is given as follows:

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

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

$$\Pr(a\text{wins an impression}) = \mathbb{1}(a \geq A_r)\frac{A_r\binom{a-1}{A_r-1}}{A\binom{A_i-1}{A_r-1}} \tag{A.34}$$

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

$$\begin{aligned}\Pr(a\text{wins at least one impression}) &= 1 - \Pr(a\text{wins no impression}) \\ &= 1 - (1 - \Pr(a\text{wins an impression}))^{T_i} \\ &= 1 - \left(1 - \mathbb{1}(a \geq A_r)\frac{A_r\binom{a-1}{A_r-1}}{A\binom{A_i-1}{A_r-1}}\right)^{T_i}\end{aligned} \tag{A.35}$$

□

### D.2 Simulation Details

In this section, we present the details of our calibrated simulation exercise. We first define a few preliminaries. As described in §4, in our simulation, we have $N = 50,000$ and $A = 100$. We define the covariate matrix as $X_{N\times D}$ where $D$ is the dimensionality of the covariate space. Elements in $X$ come from a Normal distribution $\mathcal{N}(0, \sigma_x)$, where we set $\sigma_x = 0.5$. 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:

$$\tilde{\mathcal{T}} = XB^T, \tag{A.36}$$

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

$$B = UV^T, \tag{A.37}$$

  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 $\mathcal{N}(0, \sigma_u)$ and $\mathcal{N}(0, \sigma_v)$. We set $\sigma_u = 0.5$ and $\sigma_v = 0.5$. To control the variance in CATE within studies, one could vary $\sigma_v$. Since both $U$ and $V$ are mean zero in each element, the mean of each column in matrix $\tilde{\mathcal{T}}$ is equal to 0. 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:

$$\mathcal{G} = XG^T + 1, \tag{A.38}$$

where entries in the weight matrix $G_{A\times D}$ all come from $\mathcal{N}(0, 0.5)$. The addition of the term 1 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 $\Pi$ 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:

$$Y_i^{(a)} = \mathcal{G}_{i,a} + W_i^{(a)}\mathcal{T}_{i,a} + \epsilon_{i,a}, \tag{A.39}$$

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.3 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 §3.1.3, 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 0 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.

### D.4 Visualization of CATE Distributions for Different Assignment Regions

A major finding in §4.3.3 is that our algorithm performs remarkably well in finding the good targeting opportunities. This goes against the argument that the deterministic regions only involve opportunities whose assignment is obvious given their value. That is, the deterministic assignment region only involves high-value opportunities that the manager never wants to miss, whereas the deterministic no-assignment region

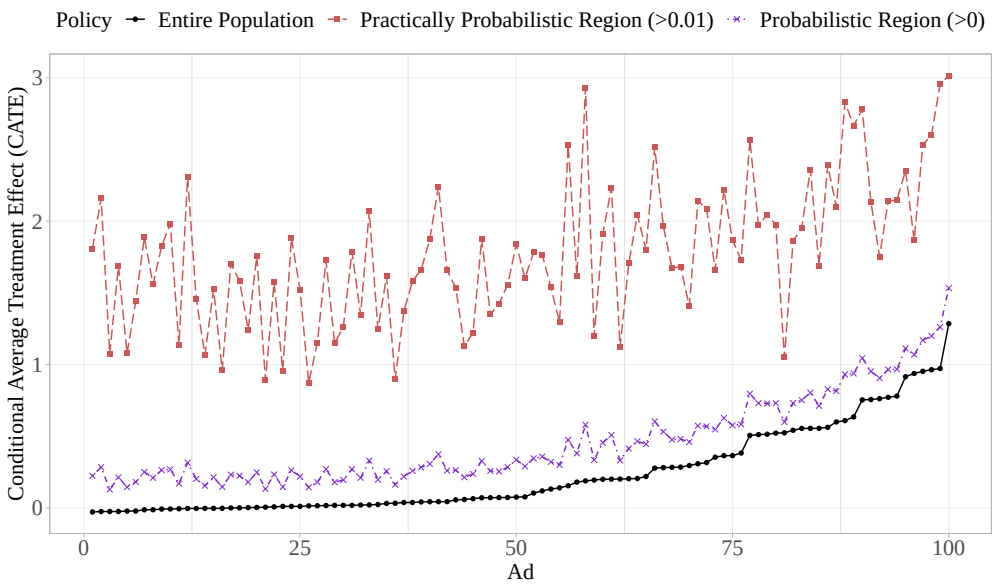



Figure A.1. Treatment effects for different regions of the data

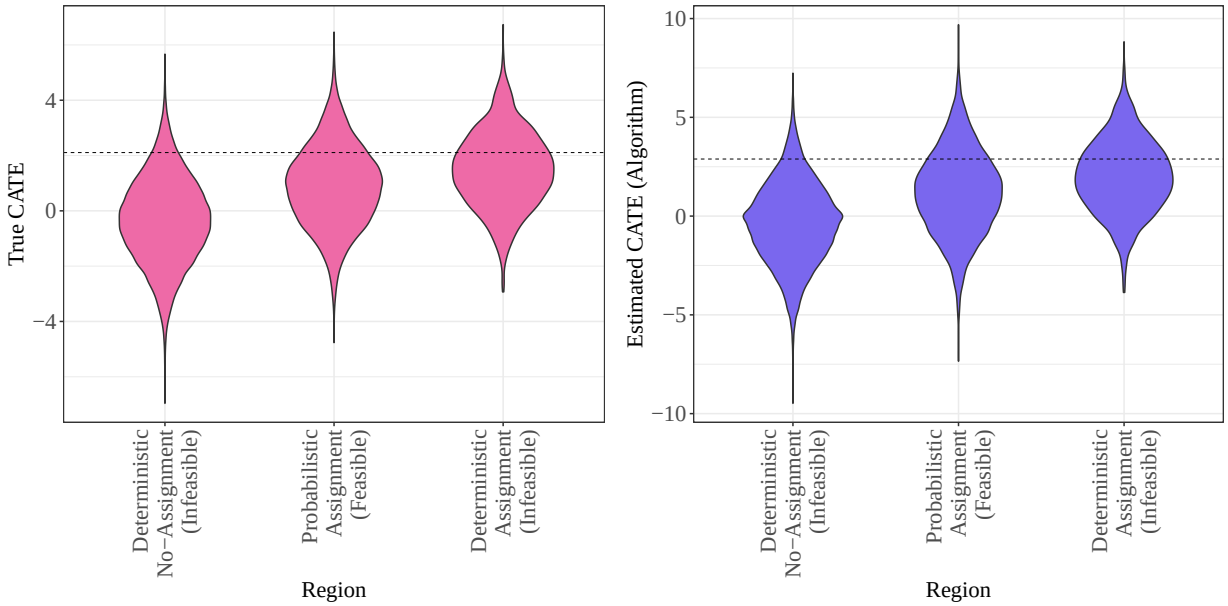
only includes poor-performing units that the manager never wants to target. We visualize this insight in Figure A.2 by showing the true and estimated CATE distributions of three different assignment regions: (1) practically deterministic no-assignment region that has a propensity score below 0.1, whose CATE is infeasible to be estimated using conventional methods, (2) probabilistic assignment region that has a propensity score between 0.1 and 0.9, whose CATE can be estimated using consistent CATE estimators, and (3) practically deterministic assignment region that has a propensity score above 0.9 and belongs to the infeasible region for CATE estimation. Please note all presented results are qualitatively the same if we use different trimming thresholds, such as 0.01, 0.05 or 0.20.

Figure A.2a shows the distributions of true CATE values different assignment regions. The dashed line in Figure A.2a is the targeting threshold for the top 10% of units. As shown in this figure, there are instances in the deterministic no-assignment region that are used for targeting, whereas there are instances in the deterministic assignment region that should not be used in targeting. The same pattern holds when we use the completed CATE based on our algorithm, as illustrated in Figure A.2b. Although the mean of the CATE distribution is lower for the deterministic no-assignment region, there are still numerous targeting opportunities in that region. The gains of our algorithm come from correctly detecting these targeting opportunities. It is worth clarifying that the longer tail of deterministic no-assignment and probabilistic assignment is due to the fact that there are substantially more instances in these regions in our simulation exercise.

## D.5 Robustness Checks and Sensitivity Analysis

### D.5.1 Robustness to Alternative Auctions

In our main analysis, we consider a combinatorial case where the platform draws a number of ads at random from the inventory to include in the auction. The rationale for this form of random bidder sub-sampling is the presence of advertisers' budget-pacing strategies and computational bottlenecks of the platform (Kim et al. 2024). Although our setup with bidder sub-sampling captures the essence of this rationale, it may appear as a stylized abstraction made for theoretical convenience. To show that our results are not driven by the specifics of the imposed structure on bidder sub-sampling, we relax the assumption on the fixed size of




(a) True CATE Distributions

(b) Estimated CATE Distributions (Algorithm)

Figure A.2. Distributions of true and estimated CATE for different assignment regions

$A_r$ and consider an alternative auction environment wherein each ad's availability depends on their budget pacing decision and the auction's reserve pricing. As such, each ad $a$ has a budget-pacing rule that translates to a probabilistic availability of ad $a$ for the $p_a$ fraction of impressions. The platform sets an optimal reserve price $b_{\text{rp}}$ for all impressions based on the distribution of bids in a direct revelation mechanism, according to Myerson (1981), so it satisfies the following condition:

$$b_{\text{rp}} - \frac{1 - F(b_{\text{rp}})}{f(b_{\text{rp}})} = 0, \tag{A.40}$$

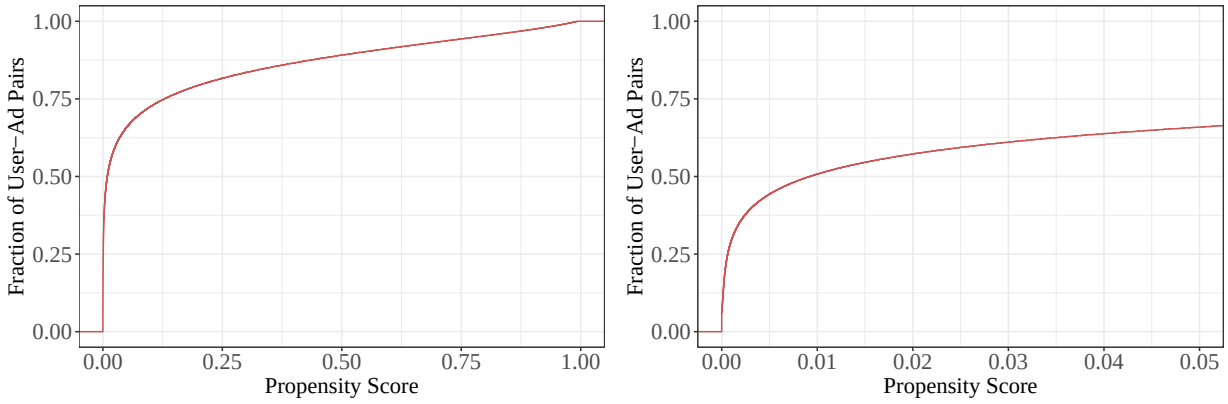
where $F(\cdot)$ is the distribution of valuations (bids in truthful mechanisms) and $f(\cdot)$ is its density function. With the setting defined, we can write the following lemma about the propensity scores:

**Lemma 4.** *Suppose that each bidder $a$ submits 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:*

$$\pi_i^{(a)} = 1 - \left(1 - \mathbb{1}(b_{i,a} \geq b_{rp}) p_a \prod_{a' > a} (1 - p_{a'})\right)^{T_i}, \tag{A.41}$$

*where $T_i$ is the total number of impressions shown to user $i$.*

*Proof.* We start by measuring the probability that ad $a$ wins an impression. For this event to happen, we need three independent events to happen at the same time: (1) we need ad $a$'s bid to be greater than or equal to the reserve price, that is, $b_{i,a} \geq b_{\text{rp}}$, (2) we need ad $a$ to participate in the impression, which happens with probability $p_a$, and (3) we need no other ad with a higher bid to participate in the auction for that impression, which is equal to $\prod_{a' > a}(1 - p_{a'})$ that measures the probability that no ad $a'$ with a higher bid than ad $a$




(a) All Propensities on $x$-axis

(b) Small Propensities on $x$-axis

Figure A.3. Empirical CDF of the propensity scores for user-ad pairs in alternative ad allocation mechanism.

participates in the auction. As such, we can write:

$$\Pr(a \text{ wins an impression}) = \mathbb{1}(b_{i,a} \geq b_{\text{rp}}) p_a \prod_{a' > a} (1 - p_{a'}) \tag{A.42}$$

Now, we can calculate the probability of $a$ winning at least one of $T_i$ impressions as follows:

$$\begin{aligned} \Pr(a\text{wins at least one impression}) &= 1 - \Pr(a\text{wins no impression}) \\ &= 1 - (1 - \Pr(a\text{wins an impression}))^{T_i} \\ &= 1 - \left(1 - \mathbb{1}(b_{i,a} \geq b_{\text{rp}}) p_a \prod_{a' > a} (1 - p_{a'})\right)^{T_i} \end{aligned} \tag{A.43}$$

□

We now generate the data with the same primitives as the simulation in §4.3. For simplicity, we use $p_a = 0.1$ for all ads. We begin by showing the empirical CDF of the propensity scores across ads. Figure A.3 shows an empirical CDF analogous to the one in Figure 3. As shown in Figure A.3a, the propensity scores for the vast majority of impression-ad pairs is very small. When we zoom in Figure A.3b, we find that nearly 70% of these impression-ad pairs have propensity scores lower than 0.05, with a sizable portion of these pairs being exactly at 0. Together, these results suggest that the alternative ad allocation mechanism generates similar patterns to the one used in the main text.

We then focus on the performance of our algorithm when applied to this example. We present the results in Figure A.4. As shown in this figure, the conventional methods like Double ML are largely biased in estimating ATE. However, our algorithm does a remarkable job in recovering the treatment effect estimates. In particular, we find that the RMSE for ATE and CATE estimates are 0.148 and 3.891, respectively. When evaluating the gains from our algorithm, we find that $\texttt{OracleRatio}_{(0.1)}$ for our algorithm is 0.985, implying that targeting the top 10% of impressions based on our algorithm recovers 98.5% of the first-best performance. In summary, our results demonstrate the robustness of our algorithm to alternative ad allocation mechanism that relaxes the fixed size of bidder sub-sampling.

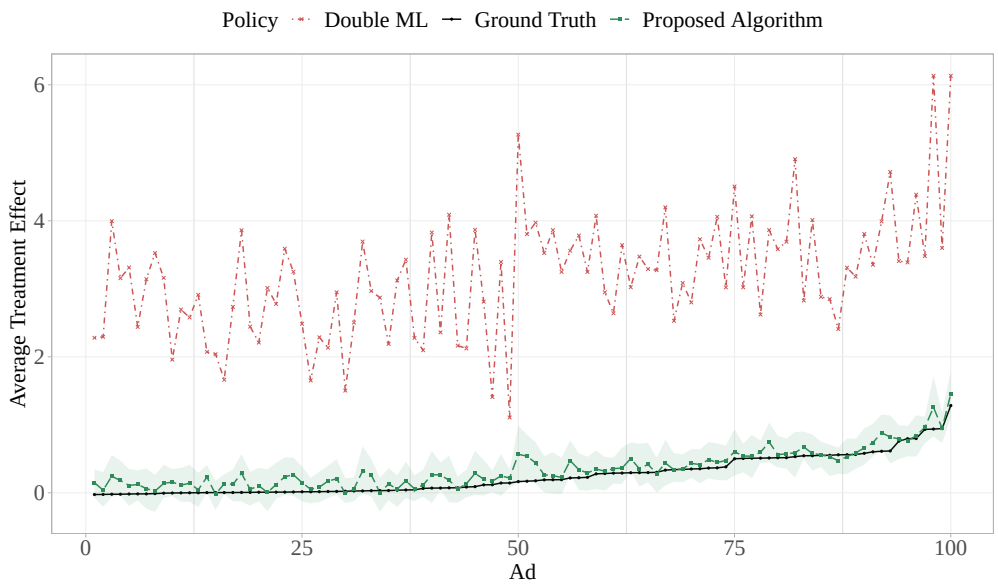



Figure A.4. The performance of the proposed algorithm and Double ML in recovering ATE in alternative ad allocation mechanism.

### D.5.2 Sensitivity Analysis

In this section, we run different sensitivity analysis for our main simulation to find the boundary conditions of the algorithm. We start with varying $A_r$, such that it becomes a larger share of the total number of ads. We know that in this situation, the assignment will become more deterministic, as more ads with lower bids will be assigned to the deterministic no-assignment region. The results of this simulation exercise are presented in the first panel in Table A1. We expect to observe a deterioration in the performance of our algorithm as $A_r$ increases. In terms of ATE estimation accuracy, we find that the performance gets worse as we increase $A_r$ from 10 to 50, but the RMSE reduces for values of $A_r$ higher than 50. The reason for this pattern is that for $A_r > 50$, a tiny fraction of entries in the matrix are in the feasible region (less than 5%) and the matrix completion algorithm basically returns values closer to zero for imputed entries. As such, the ATE across studies becomes a flat line close to zero, that incidentally performs well in terms of $\texttt{RMSE}_{\text{ATE}}$, although it has no discriminative power. The poor performance of our algorithm with higher values of $A_r$ is better demonstrated by the measures of targeting performance. Together, we conclude that extreme cases of deterministic pattern is a boundary condition for our algorithm, as there will be very limited signal in these cases for the matrix completion algorithm.

In the second panel of Table A1, we vary the rank of the underlying CATE matrix from 5 to 50. Naturally, we expect a poorer performance as we increase the matrix rank, because the algorithm does not have enough degrees of freedom for recovering a high-rank decomposition. As expected, the performance gets worse as we increase the rank of the underlying CATE matrix. However, we note that even in high-rank scenarios, the algorithm substantially outperforms the benchmarks. This is because the information in observed entries transfers valuable signal to missing entries of the matrix.

In the third panel, we vary the noise of the CATE values within each study through the parameter $\sigma_v$ defined in Web Appendix D.2. We find that the RMSE measures increase with the CATE variance, which is expected as the target is more noisy. We further find that the average gains from targeting increases as we increase the variance of CATE. This is also expected because sampling from the top 10% is more valuable if the distribution has higher variance, holding the mean fixed. Interestingly, we find the oracle ratio to be stable across different noise levels of CATE values. This suggests that the two forces – lower estimation

accuracy and higher value of targeting – cancel each other out.

Lastly, in the fourth panel, we vary the correlation between bid and CATE values to see how our algorithm performs with varying levels of imperfect linkage between bids and CATEs. We find a stable and strong performance by our algorithm across all cases. A few points are worth noting in our results. First, when the correlation between bid and CATE is weak, the deterministic assignment patterns create less of an issue, which makes the conventional methods like DML perform better, as the GATE for the probabilistic region is close to the ATE for the population. However, even in those settings, our algorithm performs substantially better, highlighting its power in learning parameters across studies. Second, we find that even in the case with a 0.9 correlation between bid and CATE values, the targeting performance of our algorithm remains higher than that of the targeting based on bids, as shown in the last column of the table. This finding is important because it suggests even in settings where advertisers have near perfect signal about the CATE, our algorithm is able to offer significant value.

| Initial Parameter | Algorithm Performance | | | | Benchmark Performance | |
|---|---|---|---|---|---|---|
| | $\texttt{RMSE}_{\text{ATE}}$ | $\texttt{RMSE}_{\text{CATE}}$ | $\texttt{Gain}_{(0.1)}$ | $\texttt{OracleRatio}_{(0.1)}$ | $\texttt{RMSE}^{\text{DML}}_{\text{ATE}}$ | $\texttt{OracleRatio}^{\text{Bid}}_{(0.1)}$ |
| | *Panel 1: Varying bidder-subsampling parameter $A_r$* | | | | | |
| $A_r = 10$ | 0.153 | 1.178 | 0.474 | 0.981 | 1.969 | 0.525 |
| $A_r = 20$ | 0.344 | 1.439 | 0.467 | 0.926 | 2.179 | 0.524 |
| $A_r = 30$ | 0.594 | 1.830 | 0.407 | 0.834 | 2.187 | 0.528 |
| $A_r = 40$ | 0.650 | 2.068 | 0.349 | 0.751 | 2.187 | 0.522 |
| $A_r = 50$ | 0.768 | 2.494 | 0.338 | 0.663 | 2.474 | 0.527 |
| $A_r = 60$ | 0.690 | 2.616 | 0.304 | 0.600 | 2.557 | 0.522 |
| $A_r = 70$ | 0.611 | 2.698 | 0.271 | 0.528 | 2.593 | 0.525 |
| $A_r = 80$ | 0.555 | 2.729 | 0.222 | 0.442 | 2.640 | 0.526 |
| $A_r = 90$ | 0.355 | 2.670 | 0.133 | 0.277 | 2.641 | 0.523 |
| | *Panel 2: Varying the rank of the CATE matrix* | | | | | |
| Rank $= 5$ | 0.073 | 0.879 | 0.369 | 0.987 | 1.524 | 0.534 |
| Rank $= 10$ | 0.153 | 1.178 | 0.474 | 0.981 | 1.969 | 0.525 |
| Rank $= 15$ | 0.268 | 1.443 | 0.566 | 0.970 | 2.391 | 0.524 |
| Rank $= 20$ | 0.483 | 1.870 | 0.681 | 0.956 | 2.938 | 0.517 |
| Rank $= 25$ | 0.826 | 2.341 | 0.748 | 0.936 | 3.307 | 0.513 |
| Rank $= 30$ | 1.056 | 2.753 | 0.786 | 0.920 | 3.533 | 0.516 |
| Rank $= 35$ | 1.264 | 3.133 | 0.830 | 0.908 | 3.745 | 0.517 |
| Rank $= 40$ | 1.608 | 3.610 | 0.876 | 0.894 | 4.050 | 0.515 |
| Rank $= 45$ | 1.809 | 3.980 | 0.905 | 0.883 | 4.298 | 0.509 |
| Rank $= 50$ | 2.011 | 4.399 | 0.973 | 0.877 | 4.584 | 0.510 |
| | *Panel 3: Varying the variance of CATE within study* | | | | | |
| $\sigma_v = 0.25$ | 0.071 | 0.609 | 0.254 | 0.979 | 1.009 | 0.545 |
| $\sigma_v = 0.50$ | 0.153 | 1.178 | 0.474 | 0.981 | 1.969 | 0.525 |
| $\sigma_v = 0.75$ | 0.236 | 1.790 | 0.738 | 0.981 | 3.138 | 0.518 |
| $\sigma_v = 1.00$ | 0.319 | 2.384 | 0.961 | 0.981 | 4.137 | 0.512 |
| $\sigma_v = 1.25$ | 0.379 | 2.862 | 1.163 | 0.980 | 5.051 | 0.509 |
| $\sigma_v = 1.50$ | 0.419 | 3.642 | 1.457 | 0.981 | 6.333 | 0.507 |
| $\sigma_v = 1.75$ | 0.473 | 4.176 | 1.690 | 0.981 | 7.312 | 0.507 |
| $\sigma_v = 2.00$ | 0.553 | 4.875 | 1.948 | 0.980 | 8.444 | 0.506 |
| $\sigma_v = 2.25$ | 0.609 | 5.506 | 2.185 | 0.980 | 9.542 | 0.506 |
| $\sigma_v = 2.50$ | 0.609 | 6.175 | 2.507 | 0.981 | 10.928 | 0.504 |
| | *Panel 4: Varying the correlation between bid and CATE* | | | | | |
| $\rho(b, \tau) = 0.1$ | 0.094 | 1.271 | 0.463 | 0.983 | 0.409 | 0.131 |
| $\rho(b, \tau) = 0.2$ | 0.081 | 1.391 | 0.515 | 0.983 | 0.873 | 0.237 |
| $\rho(b, \tau) = 0.3$ | 0.092 | 1.252 | 0.470 | 0.982 | 1.209 | 0.337 |
| $\rho(b, \tau) = 0.4$ | 0.119 | 1.307 | 0.523 | 0.982 | 1.737 | 0.435 |
| $\rho(b, \tau) = 0.5$ | 0.153 | 1.178 | 0.474 | 0.981 | 1.969 | 0.525 |
| $\rho(b, \tau) = 0.6$ | 0.210 | 1.123 | 0.477 | 0.979 | 2.354 | 0.623 |
| $\rho(b, \tau) = 0.7$ | 0.315 | 1.153 | 0.499 | 0.975 | 2.847 | 0.713 |
| $\rho(b, \tau) = 0.8$ | 0.496 | 1.105 | 0.484 | 0.973 | 3.071 | 0.811 |
| $\rho(b, \tau) = 0.9$ | 0.907 | 1.402 | 0.503 | 0.963 | 3.521 | 0.906 |

Table A1. Sensitivity analysis based on different initial parameters for the data-generating process.

# 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)

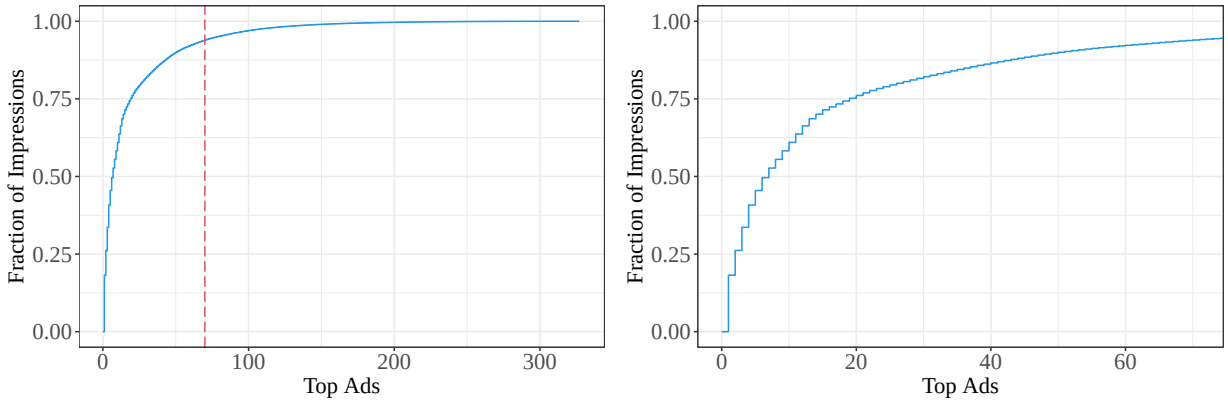
- Latitude
- Longitude
- Province
- Smartphone Brand
- Connectivity Type
- Mobile Service Provider
- The total number of impressions the user has seen prior to the current session
- The total number of clicks user the user has made prior to the current session
- The total number of impressions the user 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 Selection of Ads

As discussed in §5.2.4, we include ads that have at least 10,000 impressions to ensure that we have sufficient statistical power for CATE estimation. As such, we select 68 ads other than the focal ad and the platform ad. Figure A.5 shows the cumulative share of all impressions allocated to top ads. Figure A.5a visualizes these cumulative shares for all ads, whereas Figure A.5b zooms into the top 70 that are used in our study. Notably, these 70 ads generate 94% of all impressions in our main sample.

## E.3 Missingness Patterns Induced by Deterministic Assignment

We define a CATE matrix that has 100,000 rows and 69 columns, corresponding to 69 ads. Although all impressions satisfy the overlap assumption for the focal ad, some impressions may violate this assumption and belong to the deterministic region for some ads. To ensure the accuracy of our CATE estimates for




(a) All ads

(b) Top 70 ads (selected ads)

Figure A.5. Cumulative share of impressions generated by top $K$ ads

other ads, we filter impression-ad pairs that violate the overlap assumption. We visualize the missingness pattern induced by overlap violation in Figure A.6 for a small sub-sample of size 100 from our CATE matrix, where the feasible regions are shown with green and the infeasible entries are shown in blank. As shown in this figure, some ads are entirely available for all impressions, whereas other ads have substantially sparser feasibility patterns, reflecting the practical setting where algorithmic decision-making leads to both probabilistic and deterministic assignment regions.

### E.4 Details of CATE Estimation

As discussed in §5.2.4, we use a separate sample for each ad to estimate CATEs. For each ad $a$, the sample we use for CATE estimation must satisfy the fundamental causal inference assumptions: (1) SUTVA, (2) overlap, and (3) unconfoundedness. In our setting, we assume SUTVA as possibility of interference between users based on treatment assignment is reasonably low. We ensure overlap assumption is satisfied by selecting impressions where both ad $a$ and the platform ad are participating the its corresponding auction. Since the auction is probabilistic, this ensures that all participating ads will have a non-zero propensity of winning the auction. For the unconfoundedness assumption, we rely on the quasi-proportional auction used to allocated ads: we know that for the sample of impressions awarded to either ad $a$ or the platform ad where both ads are eligible to be shown, the propensity scores will be proportional to these ads' quality-adjusted bids. In particular, what gives us further estimation stability is the persistence of this proportionality due to a lack of bid-changing in our data, which makes the propensity fixed across impressions. We present the steps in CATE estimation and forming the CATE matrix as follows:

- For any non-focal ad $a$, we draw a sample of all impressions that are allocated to either ad $a$ or the platform ad, conditional on both being eligible for these impressions. This gives us a sample of treated and control impressions for each ad $a$. We denote the size of this sample by $N_a$. Given the varying frequency of ads shown in Figure A.5, we observe significant heterogeneity in $N_a$, with minimum, median and maximum being 12,591, 86,335, and 1,616,640 impressions. To reduce the computational burden, if $N_a > 100,000$, we draw a random sample of size 100,000 impressions for CATE estimation. We denote the set of impressions for each ad $a$ as $\mathcal{I}_a$. For the focal ad, we use the sample of 100,000 impressions (targeting profiles) illustrated in Figure 5 used for the CATE matrix.
- For any ad $a$, we use the sample of impressions $\mathcal{I}_a$ to estimate CATE using R-learner with XGBoost. Let $W_i^{(a)}$ denote the binary variable that takes the value 1 when the ad $a$ is shown and the value 0 when

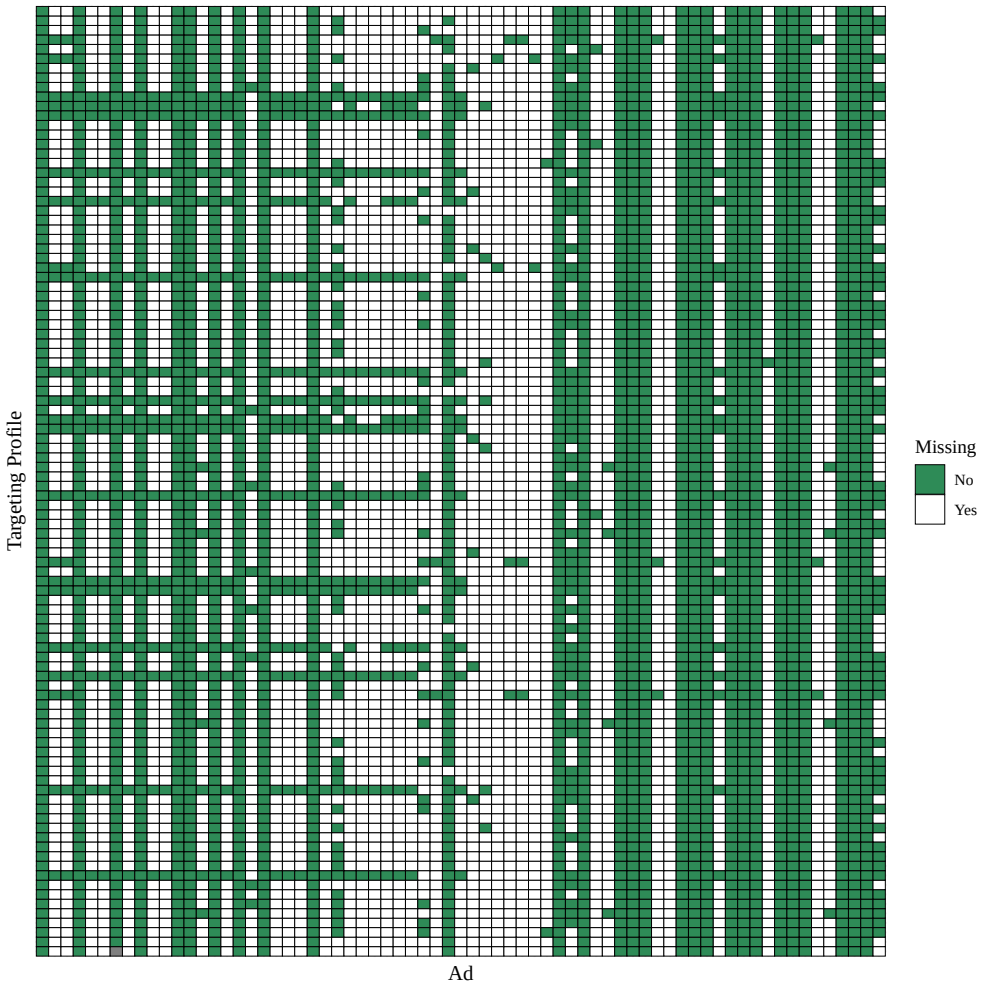



Figure A.6. Missingness pattern in a sub-sample of the CATE matrix

the platform ad is shown. We write the R-loss objective function as follows:

$$\widehat{\tau}^{(a)}(\cdot) = \underset{\tau \in \mathcal{H}}{\operatorname{argmin}} \sum_{i \in \mathcal{I}_a} \left( \left( Y_i^{(a)} - \widehat{m}_{\text{xgb}}^{-k_a(i)}(X_i) \right) - \left( W_i^{(a)} - \widehat{e}_{\text{xgb}}^{-k_a(i)}(X_i) \right) \tau^{(a)}(X_i) \right)^2 + \lambda \Omega_{\mathcal{H}}(\tau), \tag{A.44}$$

where $\widehat{m}_{\text{xgb}}^{-k_a(i)}(\cdot)$ and $\widehat{e}_{\text{xgb}}^{-k_a(i)}(\cdot)$ are the cross-fitted nuisance functions for $\mathbb{E}[Y_i^{(a)} \mid X_i]$ and $\mathbb{E}[W_i^{(a)} \mid X_i]$ using XGBoost as the machine learning method, respectively, and the superscript $-k_a(i)$ denotes that the fold containing observation $i$ was not used in learning the predictive model for observation $i$, and $\Omega_{\mathcal{H}}(\tau)$ denote the complexity of function $\tau \in \mathcal{H}$ that we want to regularize using the regularization parameter $\lambda$.

- Once we estimate the CATE function $\widehat{\tau}^{(a)}(\cdot)$ for each ad $a$, we predict CATE for entries in the CATE matrix. As such, the input for each entry in row $i$ is the targeting profile in impression $i$, denoted by $X_i$. It is worth emphasizing that although all the impressions in the CATE matrix are allocated to either the focal ad or the platform ad, we can still consistently estimate the CATE for other ads, as long as though ads could have been shown (non-zero propensity score) in an impression.

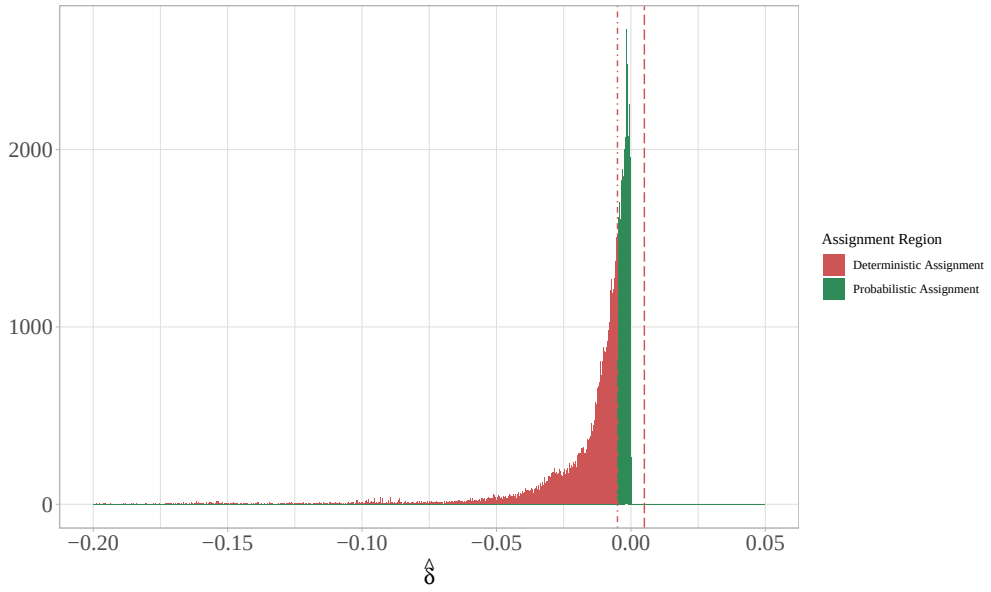



Figure A.7. Assignment regions based on the distribution of $\widehat{\delta}$

| **Method** | $\texttt{Gain}_+$ | $\texttt{OracleRatio}_+$ |
|---|---|---|
| Proposed Algorithm | 0.00296 | 98.80% |
| Ad Allocation Algorithm ($\widehat{\delta}$) | 0.00015 | 4.92% |
| Ground Truth | 0.00230 | 100.00% |

Table A2. Targeting performance of different models when re-allocating impressions based on the positivity of CATE estimates

### E.5 Algorithmic Ad Allocation: Distribution of $\widehat{\delta}$

Figure A.7 shows the distribution of $\widehat{\delta}$ to induce the missingness pattern in the focal ad, as illustrated in §5.4.1. As shown in this figure, the majority of CTR estimates predict a lower value for the focal ad compared to the platform ad. However, there are some values in the probabilistic range $[-0.005, 0.005]$ that induce some local randomization in ad allocation. We find that 69.7% are in the deterministic no-assignment region and 39.3% of impressions are in the probabilistic assignment region. Our algorithm can only use the ones in the probabilistic region to impute the ones in the deterministic region.

### E.6 Alternative Form of Personalized Re-allocation

In §5.4.2, we focus on re-allocating impressions to only top 10% of impressions. As an alternative approach in this section, we focus on another form of documenting gains from our algorithm: instead of top 10% re-allocation, we re-allocate to positive CATE estimates under each model, a common approach in developing personalized policies. We denote this metric by $\texttt{Gain}_+$ and present the results for it in Table A2. Our results from this exercise show even greater gains from using our algorithm. We find that the gains are higher for both our proposed algorithm and the ground truth because they can re-allocate more than 10% of impressions. The reason for the poor performance of the ad allocation algorithm is that the estimates of $\widehat{\delta}$ are largely negative, so less than 10% of them are selected for re-allocation.

| **Method** | ATE | $\texttt{RMSE}_{\text{CATE}}$ | $\texttt{Gain}_{(0.1)}$ | $\texttt{OracleRatio}_{(0.1)}$ |
|---|---|---|---|---|
| Proposed Algorithm (Only 68 Original Ads) | $-0.01679$ | 0.00377 | 0.00253 | 99.24% |
| Proposed Algorithm (68 Original Ads + 20 Irrelevant Ads) | $-0.01639$ | 0.00424 | 0.00250 | 98.10% |
| Proposed Algorithm (Only 20 Irrelevant Ads) | 0.00093 | 0.03741 | $-0.00158$ | $-62.23\%$ |
| Ground Truth | $-0.01634$ | 0 | 0.00255 | 100% |

Table A3. Performance of our proposed algorithm when adding irrelevant ads.

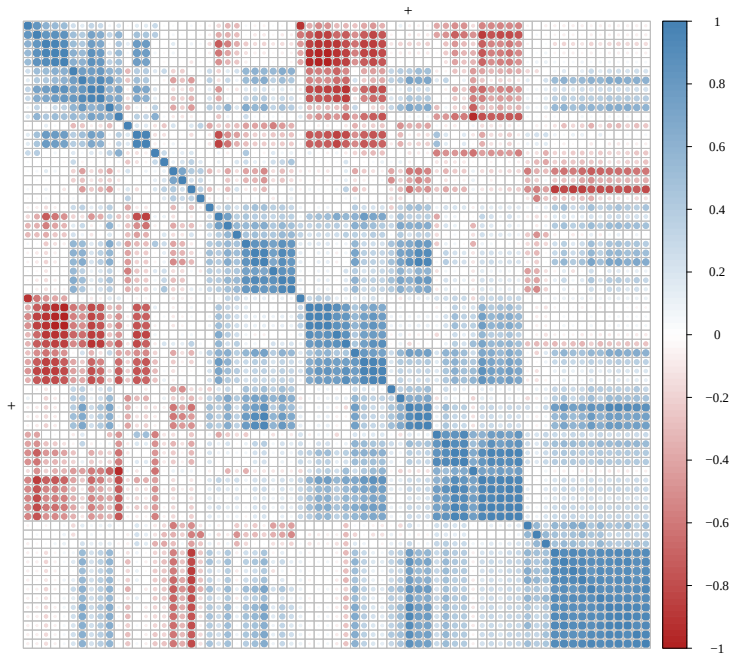
### E.7 Inclusion of Irrelevant Ads

The main idea behind our algorithm is using the information across studies to impute CATE for the overlap-violating or missing entries. In our main analysis, we find that our algorithm is able to exploit the information across studies and accurately impute entries. In this section, we include irrelevant columns in the CATE matrix to see how it affects the performance of our algorithm. Theoretically, we know that if we concatenate two matrices of rank $r_1$ and $r_2$, the rank of the combined matrix is at most $r_1 + r_2$. As such, we expect our matrix completion algorithm to learn the new rank of the combined matrix. In particular, our algorithm can effectively ignore irrelevant information. To test that, we create 20 random-generated columns all coming from standard Normal distribution and add these columns to the CATE matrix. We then run our algorithm in two settings with (1) CATE matrix containing all 68 original ads as well as 20 irrelevant ads, and (2) CATE matrix only containing the irrelevant ads. We find the best-performing matrix decompositions based on the validation procedure. The best-performing matrix for the first scenario with both original ads and irrelevant ads is of rank 34, whereas the rank for the matrix with irrelevant ads is 19. This suggests that our algorithm can verify the increased rank in a data-driven manner.

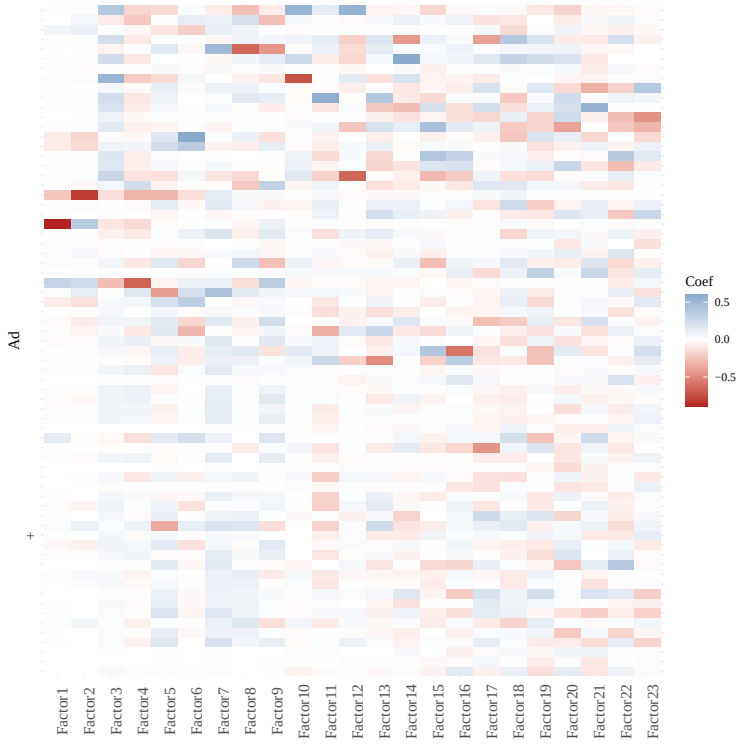
We present the results in Table A3. As shown in this table, adding 20 irrelevant ads does not have a major impact on the algorithm. We see a slightly better ATE estimation performance and a slightly worse CATE estimation and targeting performance compared to the algorithm with only 68 original ads. On the contrary, when we only include the 20 irrelevant ads, the algorithm fails in recovering the causal parameters and performs poorly in targeting. Overall, our results highlight an important feature of matrix completion algorithms, which is their ability in ignoring irrelvant information.

### E.8 Similarities and Differences Among the Ads in the CATE Matrix

One of the requirements for our algorithm to work is the commonality among studies, which satisfies the low-rank assumption. That is, there are fewer factors that collectively explain CATE in a larger number of studies. In our study, we select a set of mobile in-app ads to empirically test whether the low-rank assumption is satisfied and evaluate the performance of our algorithm. Although our ad sampling procedure is agnostic to the commonality between ads, the strong performance of our algorithm suggests that the 69 ads in our CATE matrix share some common factors. In this section, we aim to understand the correlation structure in the CATE matrix and the distinctiveness each ad has. We first present the correlation matrix in Figure A.8a. As shown in this figure, many pairs of ads exhibit high correlations in both directions, though most observed correlations are positive. Figure A.8a automatically cluster groups of ads that have more similar pairwise correlations. Although we do not have the ad content data, one could speculate that ads clustered together likely belong to the same category, such as mobile gaming app, or health. The plus sign in this figure shows the focal ad, which has a positive correlation with 51 ads, and a negative one with 17 ads.




(a) Correlation Matrix




(b) Factor Weights

Figure A.8. Similarities and differences among ads in the CATE matrix

Although the correlation matrix in Figure A.8a illustrates the presence of a correlation structure between

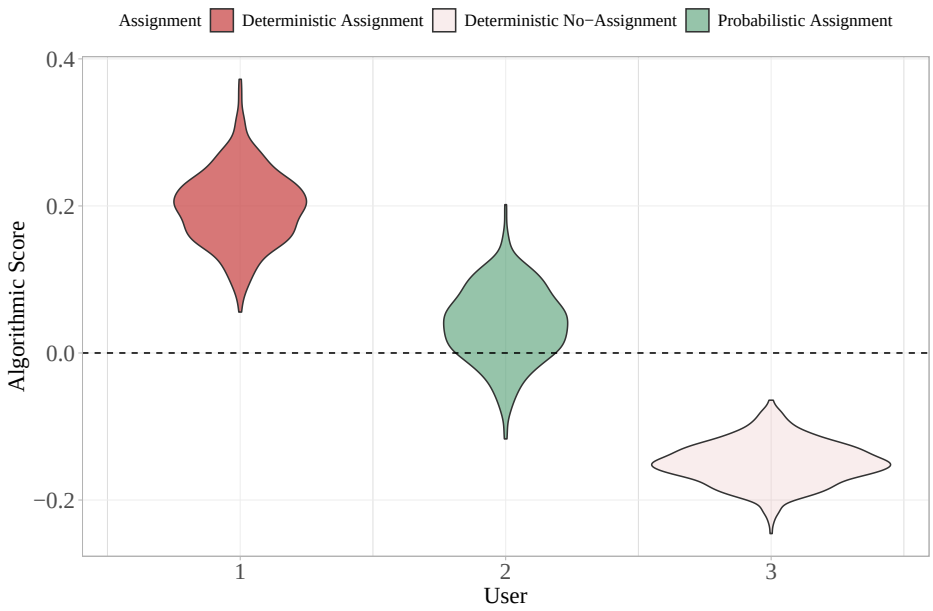



Figure A.9. Probabilistic and deterministic assignment regions under algorithmic decision-making

ads, it also shows major differences between ads. To further examine the distinctiveness of ads, we visualize the weights each ad has for the CATE factors. To do so, we take the CATE matrix completed by our algorithm and perform Principal Component Analysis (PCA). We then visualize the factor loadings for the 23 factors in our study in Figure A.8b, using a heatmap.[19] As before, we use a plus sign to refer to the focal ad. Figure A.8b illustrates how each ad relies distinctive factors, highlighting the differences among ads. In summary, both figures in Figure A.8 suggest that ads in our study are not unusually similar, although the commonalities among them satisfies the low-rank assumption and allows our model to perform well.

## F Settings with the Platform's Full Control over the Assignment

In this section, we consider a setting where the platform has full control over the assignment policy. In many of these cases, algorithmic scores determine the treatment assignment. The intervention with the highest algorithmic score is 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 A.9 illustrates this point with a case where a threshold rule is used to assign individuals to an intervention: only if the posterior distribution of the algorithmic score is entirely above or below the cutoff shown by the dashed line, the assignment becomes deterministic. Practical examples of such settings include promotion assignment (e.g., Uber's promotion for future rides) and push notifications (e.g., Fitbit's notification on body activity). In these settings, uncertainty in algorithmic scores creates exogenous variation near the decision threshold. There are two ways to conceptualize this exogenous variation near the decision threshold:

- First, in many settings, platforms use bandit algorithms that are explicitly probabilistic (e.g., Linear Thompson Sampling). In these cases, the algorithm draws from the posterior distribution of parameters for algorithmic scores. Since the distribution is known at any point, one could measure the propensity

[19] Please note that the rank of the matrix using the entries in training is 22, but when estimated with the same $\lambda$ on the full set of observed entries, the rank is 23. Here we use all entries, which is why we have 23 factors and their corresponding weights in Figure A.8b.

score.

- Second, even in cases where the function determining the algorithmic score is deterministic, there is still some inherent uncertainty in parameters learned in that model, which creates exogenous variation around the cutoffs. Hence, even though the function generating algorithmic scores is deterministic in these cases, the argument is that adding or removing just one or two data points during training could shift a unit's score slightly above or below the cutoff. This variation is conceptually analogous to that in Regression Discontinuity designs, where units close to the threshold form a natural experiment. Such variation has been leveraged in algorithmic decision-making contexts to identify Local Average Treatment Effects (LATE) of various interventions (Shi et al. 2022).

As discussed in the paper, although a key application of our algorithm is advertising auctions and the mixture of probabilistic and deterministic assignment patterns created by these auctions, our algorithm can be applied more broadly to cases where the platform has full control over the assignment policy. In this section, we study the application of our algorithm in one such scenario where a platform uses algorithmic scores to assign users to promotions. As highlighted in Figure A.9, platforms often use cutoff-based decision-making where the unit is assigned to a treatment if its algorithmic score is above a certain threshold. For example, if the algorithmic score for user responsiveness to a promotion is above a certain number, the platform assigns that user to the promotion. In these setting, the randomness in intervention assignment comes from the uncertainty in algorithmic scores.

Consider a promotion targeting problem where there are $N$ users indexed by subscript $i$ and $J$ different promotions indexed by $j$. For any pair of user $i$ and promotion $j$, the platform measures an algorithmic score $S_{i,j}$, which demonstrates the responsiveness of user $i$ to promotion $j$. As discussed earlier, this score has residual uncertainty, which we capture with an additive term $\omega_{i,j}$ that comes from a distribution $F(\cdot)$, such that $S_{i,j} + \omega_{i,j}$ characterizes the all possibilities in the full posterior distribution of the algorithmic score. If the algorithmic score $S_{i,j}$ is greater than a threshold $c$, then promotion $j$ will be assigned to user $i$. As such, the propensity score $\pi_i^{(j)}$ can be determined as follows:

$$\pi_i^{(j)} = \Pr(S_{i,j} + \omega_{i,j} \geq c) \tag{A.45}$$

In our simulation exercise, we set $c = 1$ and the correlation between algorithmic scores and true CATEs to be 0.5, and draw $\omega$ from the uniform distribution $U[-2, 2]$. Further, we draw the ATEs from the uniform distribution $U[1, 2]$. We keep all other primitives of our simulation the same as the primitives in our main simulation exercise in §4.3. We generate the data and examine the performance of our algorithm compared to the benchmarks. Figure A.10 shows the performance of our algorithm compared to the Double ML benchmark. As shown in this figure, our algorithm does a remarkable job in recovering the true ATEs, whereas the Double ML produces largely biased estimates. When evaluating the targeting performance of our algorithm, we find the $\texttt{OracleRatio}_{(0.1)}$ for our algorithm to be 0.997, highlighting its strong performance in estimating CATEs and learning to prioritize.

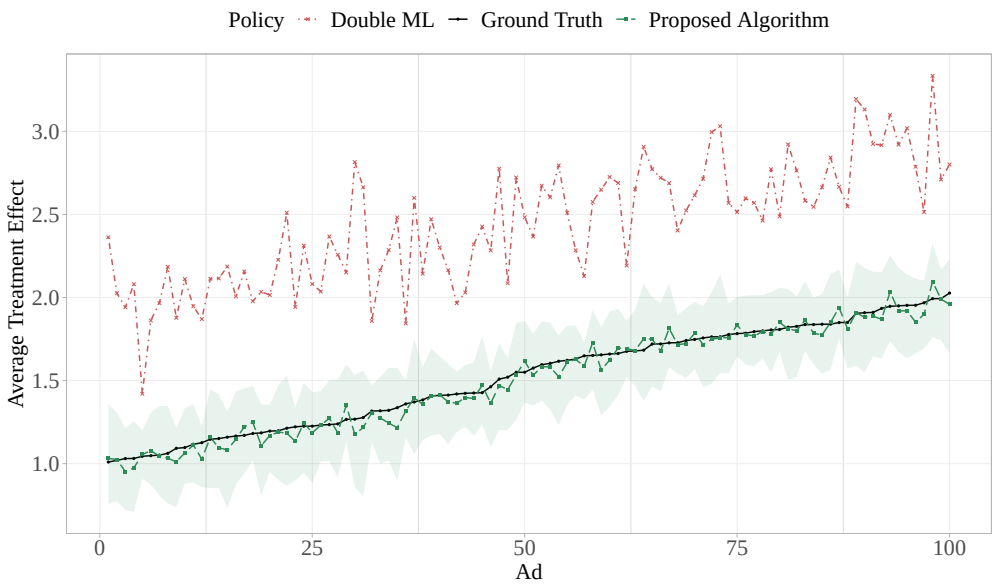



Figure A.10. The performance of the proposed algorithm and Double ML in recovering ATE in settings with cutoff-based algorithmic decision-making.

In summary, our simulation exercise in this section shows that our algorithm extends beyond the auction environment to other settings with targeted allocation and algorithmic decision-making. Although the platform can induce small-scale randomization in settings with full control over the assignment, it is worth emphasizing that such design-based solutions are not useful for the existing data. Many firms have data sets where the logged policy has a mixture of probabilistic and deterministic assignment. Our algorithm creates great value in these cases as it allows firms to better utilize their existing data and make more informed decisions.