

Online Appendix

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

A.1 Model-based Approaches to Estimate ATE

There are many model-based approaches one could use to estimate ATE from observational data. The traditional approach is to use a linear regression that projects the outcome on the treatment variable as well as other controls and estimates the average treatment effect. These methods work well if the confounding 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 for accounting for observed confounding. 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 to increase the flexibility and robustness of model-based methods to estimate ATE (Belloni et al. 2014, Chernozhukov et al. 2018). 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:

$$Y_i(w) = \mu_0(X_i) + \tau^*(X_i)w + \epsilon_i(w), \quad (\text{A.1})$$

where $\epsilon_i(w) = Y_i(w) - \mu_w(X_i)$ denotes the structural error term for any value of the treatment $w \in \{0, 1\}$, so that $\mathbb{E}[\epsilon_i(w) \mid X_i] = 0$ by construction. Unconfoundedness, $Y_i(w) \perp W_i \mid X_i$, further implies that conditioning on the realized treatment leaves this moment unchanged, that is, $\mathbb{E}[\epsilon_i(W_i) \mid X_i, W_i] = 0$. 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:

$$Y_i - m(X_i) = (W_i - \pi(X_i))\tau^*(X_i) + \epsilon_i(W_i), \quad (\text{A.2})$$

which holds because $m(X_i) = \mu_0(X_i) + \tau^*(X_i)\pi(X_i)$. This decomposition – 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:

$$\tau^*(\cdot) = \underset{\tau}{\operatorname{argmin}} \mathbb{E} \left[(Y_i - m(X_i) - (W_i - \pi(X_i))\tau(X_i))^2 \right]. \quad (\text{A.3})$$

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. 2018). We refer to this as the *partially linear* form of DML.

Chernozhukov et al. (2018) also study a second form, which they call the interactive regression model and which corresponds to the doubly robust or augmented inverse propensity score (AIPW) estimator. Instead of partialling the treatment out of the outcome, this approach models both conditional mean functions directly and corrects them with an inverse-propensity-weighted residual:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{N} \sum_{i=1}^N \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{W_i(Y_i - \hat{\mu}_1(X_i))}{\pi(X_i)} - \frac{(1 - W_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \pi(X_i)} \right]. \quad (\text{A.4})$$

The estimator is doubly robust in the sense that it remains consistent if either the outcome models $\mu_w(\cdot)$ or

the propensity score $\pi(\cdot)$ is correctly specified.

The two forms are advantageous in different circumstances, and the distinction matters for our setting. The partially linear form never divides by the propensity score, so it remains numerically stable and comparatively efficient even when $\pi(X_i)$ approaches 0 or 1. Its cost is that the single coefficient it recovers is the ATE only when the treatment effect is homogeneous. The AIPW form targets the unweighted ATE without any homogeneity restriction and is doubly robust, but it inherits the inverse-propensity weights of (A.5) below, so its variance grows without bound as overlap weakens. In short, the partially linear form is preferable when overlap is weak and effects are close to homogeneous, whereas AIPW is preferable when effects are heterogeneous and overlap is comfortable.

A.2 Model-free Approaches to Estimate ATE

We now discuss model-free approaches to estimate the ATE that directly use the realized outcomes without modeling them. The foundation for these approaches is the idea of importance sampling proposed by Horvitz and Thompson (1952) in their seminal paper. The idea is to weight each observation by its inverse propensity score, which gives us the following estimator for the ATE:

$$\hat{\tau}_{\text{IPS}} = \frac{1}{N} \left(\sum_{i=1}^N Y_i \left(\frac{W_i}{\pi(X_i)} - \frac{1 - W_i}{1 - \pi(X_i)} \right) \right), \quad (\text{A.5})$$

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 α_r fraction of observations with the probabilistic assignment. We note that trimming makes the estimator well-defined but not necessarily precise: the retained observations still carry weights as large as $1/\pi(X_i)$, so the variance of $\hat{\tau}_{\text{IPS}}$ can remain large even after trimming.

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. (2018), we know that if the treatment effect is homogeneous within the probabilistic region, that is, $\tau^*(X_i) = \tau_r$ for all $i \in \mathcal{I}_r$, then:

$$\operatorname{argmin}_{\tau} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} (Y_i - m(X_i) - (W_i - \pi(X_i)) \tau)^2 \xrightarrow{P} \tau_r. \quad (\text{A.6})$$

Under heterogeneity, the same objective converges instead to the overlap-weighted average

$$\theta_r = \frac{\mathbb{E}[\pi(X_i)(1 - \pi(X_i))\tau^*(X_i) \mid i \in \mathcal{I}_r]}{\mathbb{E}[\pi(X_i)(1 - \pi(X_i)) \mid i \in \mathcal{I}_r]}, \quad (\text{A.7})$$

which reduces to τ_r under homogeneity. For the purposes of this proposition the distinction is immaterial: both τ_r and θ_r are functionals of $\tau^*(\cdot)$ restricted to $\mathcal{I}_r$, so neither carries any information about the treatment effects in the deterministic region.

We now want to show that the RHS of Equation (A.6) is the same as what any method optimizing Equation (A.3) would estimate. We can write:

$$\begin{aligned}
\hat{\tau} &= \operatorname{argmin}_{\tau} \frac{1}{N} \sum_{i=1}^N (Y_i - m(X_i) - (W_i - \pi(X_i)) \tau)^2 \\
&= \operatorname{argmin}_{\tau} \frac{1}{N} \sum_{i \in \mathcal{I}_r} (Y_i - m(X_i) - (W_i - \pi(X_i)) \tau)^2 \\
&\quad + \sum_{i \notin \mathcal{I}_r} (Y_i - m(X_i) - (W_i - \pi(X_i)) \tau)^2 \\
&= \operatorname{argmin}_{\tau} \frac{1}{N} \sum_{i \in \mathcal{I}_r} (Y_i - m(X_i) - (W_i - \pi(X_i)) \tau)^2 + \sum_{i \notin \mathcal{I}_r} (Y_i - m(X_i))^2 \\
&= \operatorname{argmin}_{\tau} \frac{1}{N} \sum_{i \in \mathcal{I}_r} (Y_i - m(X_i) - (W_i - \pi(X_i)) \tau)^2 \\
&= \operatorname{argmin}_{\tau} \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} (Y_i - m(X_i) - (W_i - \pi(X_i)) \tau)^2,
\end{aligned} \tag{A.8}$$

where the second line is a simple decomposition based on the observations with probabilistic and deterministic assignment, the fourth line is because $W_i - \pi(X_i) = 0$ for observations with deterministic assignment, the fifth line drops the term $\sum_{i \notin \mathcal{I}_r} (Y_i - m(X_i))^2$ because it does not depend on τ , and the sixth line changes $1/N$ to $1/N_r$ because this rescaling does not affect the minimizer. Now if we combine the result of Equation (A.8) with that of Equation (A.6), 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 \(1952\)](#).

Finally, we note that the AIPW estimator in Equation (A.4) does not escape this limitation either. Unlike the partially linear form, AIPW converges to the unweighted τ_r without requiring homogeneity, since $\mathbb{E}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + W_i(Y_i - \hat{\mu}_1(X_i))/\pi(X_i) - (1 - W_i)(Y_i - \hat{\mu}_0(X_i))/(1 - \pi(X_i)) \mid X_i] = \tau^*(X_i)$. However, the inverse-propensity terms are undefined whenever $\pi(X_i) \in \{0, 1\}$, so AIPW must discard the overlap-violating observations exactly as the IPS estimator does. Its average is therefore taken over $\mathcal{I}_r$ alone, and it converges to τ_r rather than τ^* . All three estimators thus recover only the identifiable piece τ_r , which establishes the proposition. $\square$

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 1. *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:

$$\begin{aligned}
\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)] &= \mathbb{E}[Q_i\tau(X_i)] \\
&= \mathbb{E}[\mathbb{E}[Q_i\tau(X_i) \mid Q_i]] \\
&= \mathbb{E}[Q_i\mathbb{E}[\tau(X_i) \mid Q_i]] \\
&= P(Q_i = 1)(1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] + P(Q_i = 0)(0)\mathbb{E}[\tau(X_i) \mid Q_i = 0] \\
&= P(Q_i = 1)\mathbb{E}[\tau(X_i) \mid Q_i = 1] \\
&= P(\pi(X_i) = 1)\mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]
\end{aligned} \tag{A.9}$$

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:

$$\begin{aligned}
\tau_1 &= \mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1] \\
&= \frac{P(\pi(X_i) = 1) \mathbb{E}[\tau(X_i) \mid \pi(X_i) = 1]}{P(\pi(X_i) = 1)} \\
&= \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&\geq \frac{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]\mathbb{E}[\tau(X_i)]}{\mathbb{E}[\mathbb{1}(\pi(X_i) = 1)]} \\
&= \mathbb{E}[\tau(X_i)] \\
&= \tau^*,
\end{aligned} \tag{A.10}$$

where the fourth line comes from the fact that the two variables are positively correlated. $\square$

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 (Mazumder et al. 2010). 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. It 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.

We state the algorithm using the notation of §3.3. Let $\mathcal{T} \in \mathbb{R}^{N \times J}$ denote the CATE matrix with missing entries, and let $F \in \{0, 1\}^{N \times J}$ denote the feasibility matrix, with $F_{ij} = 1$ if entry $\mathcal{T}_{ij}$ is observed and $F_{ij} = 0$ otherwise. Write $\Omega = \{(i, j) : F_{ij} = 1\}$ for the set of observed entries, and let $P_\Omega(\cdot)$ denote the projection that retains the entries in Ω and sets the rest to zero, with $P_\Omega^\perp(\cdot)$ its complement. The goal is to approximate $\mathcal{T}$ by a matrix M of lower rank such that the missing values are imputed in a way that preserves the structure of the observed data. The optimization problem solved by SoftImpute is:

$$\widehat{M} = \underset{M}{\operatorname{argmin}} \frac{1}{2} \|P_\Omega(\mathcal{T} - M)\|_F^2 + \lambda \|M\|_*, \tag{A.11}$$

where $\|M\|_*$ is the nuclear norm of M , that is, the sum of its singular values, and $\lambda \geq 0$ is the regularization parameter that controls the trade-off between fit on the observed entries and the rank of the solution. The

nuclear norm is the tightest convex relaxation of the rank, so (A.11) is a convex problem and has a global optimum.

The algorithm relies on the soft-thresholding operator. For a fully observed matrix A with singular value decomposition $A = U\Sigma V^\top$ and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_K)$, define

$$\mathcal{S}_\lambda(A) = U\Sigma_\lambda V^\top, \quad \Sigma_\lambda = \text{diag}((\sigma_1 - \lambda)_+, \dots, (\sigma_K - \lambda)_+), \quad (\text{A.12})$$

where $(x)_+ = \max(x, 0)$. Cai et al. (2010) show that $\mathcal{S}_\lambda(A)$ is the exact solution of (A.11) when all entries are observed. SoftImpute exploits this by alternating between filling in the missing entries with the current estimate and applying $\mathcal{S}_\lambda$ to the completed matrix:

Algorithm 1 SoftImpute (Mazumder et al. 2010)

Require: Incomplete matrix $\mathcal{T}$, observed set Ω , regularization parameter λ , maximum rank r , tolerance ε

- 1: Initialize $M_0 \leftarrow \mathbf{0}$ ▷ any starting value works; the problem is convex
- 2: **for** $t = 0, 1, 2, \dots$ **do**
- 3: Fill the missing entries with the current estimate:

$$Z_t \leftarrow P_\Omega(T) + P_\Omega^\perp(M_t).$$

- 4: Apply soft-thresholded SVD, retaining at most r components:

$$M_{t+1} \leftarrow \mathcal{S}_\lambda(Z_t).$$

- 5: **break** if $\frac{\|M_{t+1} - M_t\|_F^2}{\|M_t\|_F^2} < \varepsilon$
 - 6: **end for**
 - 7: **return** $\widehat{M} = M_{t+1}$
-

Each iteration is guaranteed not to increase the objective in (A.11), and the sequence $\{M_t\}$ converges to its global minimizer (Mazumder et al. 2010). Two parameters govern the solution. The regularization parameter λ controls the amount of shrinkage applied to the singular values: a larger λ produces more aggressive shrinkage and a lower-rank approximation, and any component with $\sigma_k \leq \lambda$ is set to zero. The maximum rank r caps the number of components retained in Step 4. The two are partially redundant, in that either can limit the rank of $\widehat{M}$, but they act differently: λ shrinks every retained singular value toward zero in addition to truncating, whereas r truncates without shrinking. We describe how we set both in Online Appendix C.2.

For the large matrices in our empirical application we use the alternating least squares variant of the algorithm, SoftImpute-ALS (Hastie et al. 2015), which replaces the full SVD in Step 4 with an alternating ridge regression on a rank- r factorization. It solves the same problem as Algorithm 1 but scales far better when N is large, since it never forms a dense $N \times J$ SVD.

The main advantages of the SoftImpute algorithm are summarized as follows:

- **Scalability:** SoftImpute can efficiently handle large matrices with many missing entries by exploiting the low-rank structure of the data, and its ALS variant scales to matrices with millions of rows.
- **Convexity:** The nuclear norm is a convex relaxation of the rank, so the objective has a global optimum and the iterates converge to it regardless of initialization.
- **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 λ 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. Alternatively, if there is no prior information available, we can use a validation procedure and use all different maximum ranks with a fixed λ and pick the one with the lowest hold-out RMSE.
- Let $\tilde{\mathcal{T}}_{[n \times m]}$ denote the following matrix:

$$\tilde{\mathcal{T}}_{i,j} = \begin{cases} \mathcal{T}_{i,j} & \text{if } (i, j) \in \Omega(\mathcal{T}) \\ 0 & \text{if } (i, j) \notin \Omega(\mathcal{T}) \end{cases} \quad (\text{A.13})$$

Use Singular Value Decomposition and set $\lambda_{\max}$ as the maximum singular value.

- Create a grid of Λ ranging from zero to $\lambda_{\max}$. Without loss of generality, let $\lambda_1 < \lambda_2 < \dots < \lambda_{\max}$.
- 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 $\Omega_{\text{train}}(\mathcal{T}) \cap \Omega_{\text{validation}}(\mathcal{T}) = \emptyset$, and $\Omega_{\text{validation}}(\mathcal{T})$ contains a random α fraction of all observed entries. In our applications, we set $\alpha = 0.1$.
- Start with λ_1 and run the SoftImpute algorithm on $\mathcal{T}_{\text{train}}$. Record the model output as $\hat{\mathcal{T}}_{\lambda_1}$ and its performance on the validation set of entries $\Omega_{\text{validation}}(\mathcal{T})$. For better computational performance, warm start the model with λ_t with the model in the previous step $\hat{\mathcal{T}}_{\lambda_{t-1}}$. Record the validation performance for all values of $\lambda \in \Lambda$.
- Select the best-performing λ^* based on the performance on the validation set as follows:

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

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:

$$\mathcal{T} = X B^T, \quad (\text{A.14})$$

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, $\mathcal{T} = X V U^T$, 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 than 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:

$$\Pr(a \text{ wins an impression}) = \mathbb{1}(a \geq A_r) \frac{A_r \binom{a-1}{A_r-1}}{A \binom{A_i-1}{A_r-1}} \quad (\text{A.15})$$

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

$$\begin{aligned} \Pr(a \text{ wins at least one impression}) &= 1 - \Pr(a \text{ wins no impression}) \\ &= 1 - (1 - \Pr(a \text{ wins an impression}))^{T_i} \\ &= 1 - \left(1 - \mathbb{1}(a \geq A_r) \frac{A_r \binom{a-1}{A_r-1}}{A \binom{A_i-1}{A_r-1}}\right)^{T_i} \end{aligned} \quad (\text{A.16})$$

□

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:

$$\tilde{\mathcal{T}} = X B^T, \quad (\text{A.17})$$

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

$$B = U V^T, \quad (\text{A.18})$$

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 σ_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:

$$\mathcal{G} = XG^T + 1, \quad (\text{A.19})$$

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

$$Y_i^{(a)} = \mathcal{G}_{i,a} + W_i^{(a)}\mathcal{T}_{i,a} + \epsilon_{i,a}, \quad (\text{A.20})$$

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 $\epsilon_{i,a} \sim \mathcal{N}(0, 2)$.

- *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.

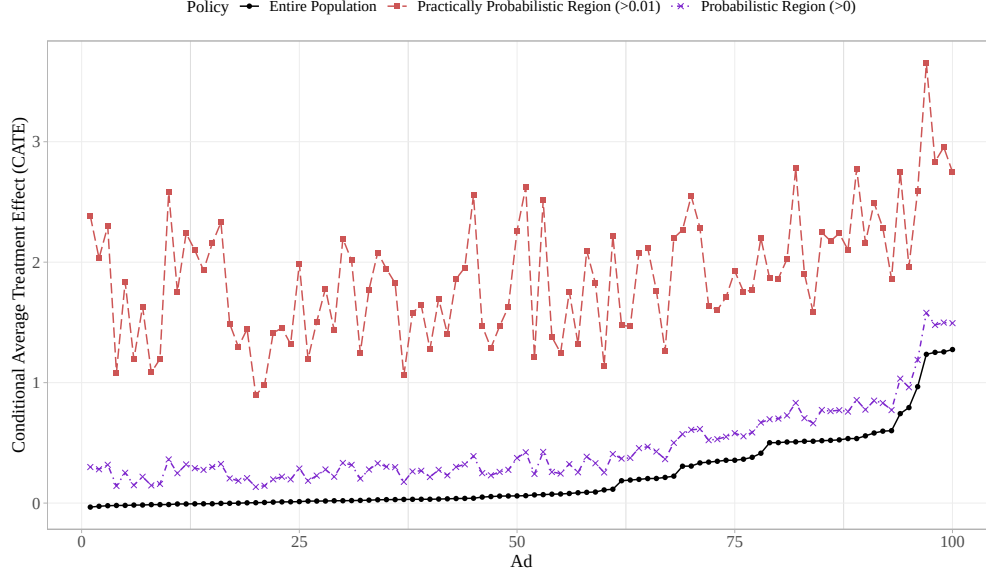


Figure A.1. Treatment effects for different regions of the data

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 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.

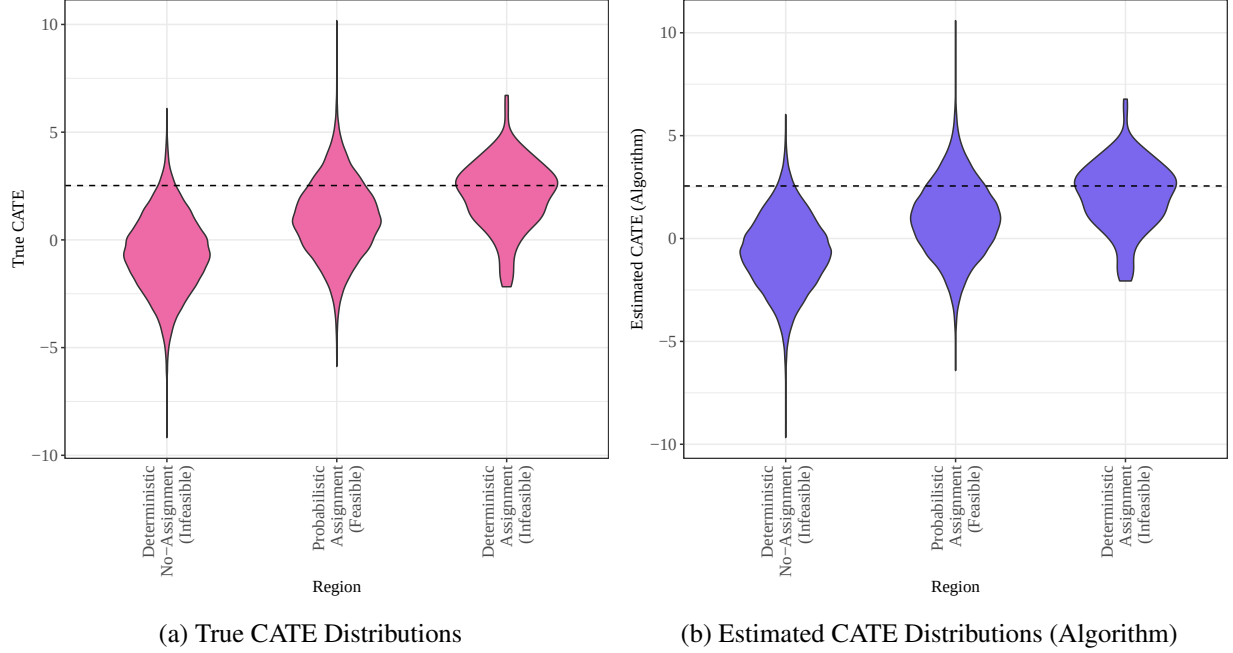


Figure A.2. Distributions of true and estimated CATE for different assignment regions

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_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_{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:

$$b_{rp} - \frac{1 - F(b_{rp})}{f(b_{rp})} = 0, \quad (\text{A.21})$$

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

$$\pi_i^{(a)} = 1 - \left(1 - \mathbb{1}(b_{i,a} \geq b_{rp}) p_a \prod_{a' > a} (1 - p_{a'}) \right)^{T_i}, \quad (\text{A.22})$$

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

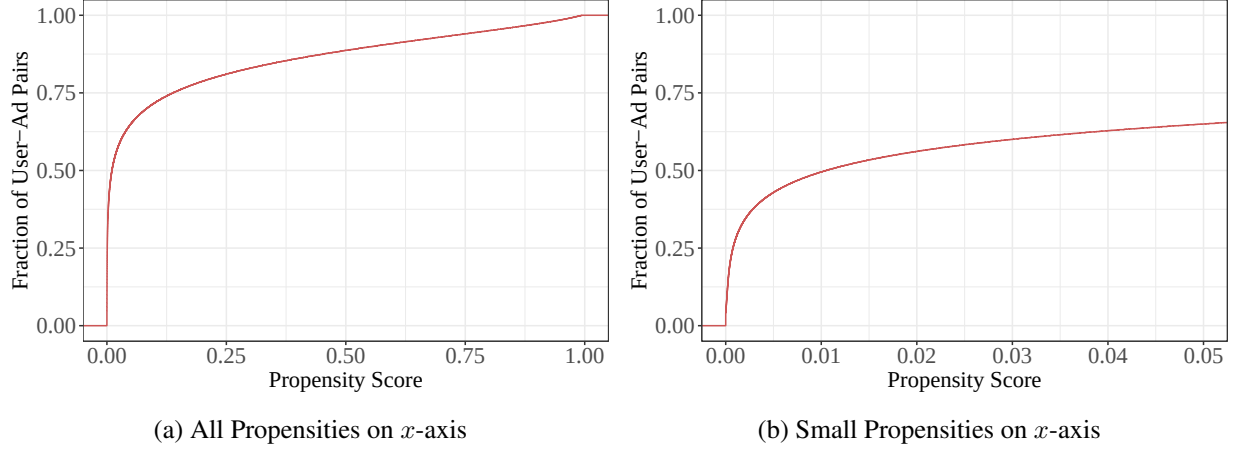


Figure A.3. Empirical CDF of the propensity scores for user-ad pairs in alternative ad allocation mechanism.

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, $b_{i,a} \geq b_{rp}$, (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 participates in the auction. As such, we can write:

$$\Pr(a \text{ wins an impression}) = \mathbb{1}(b_{i,a} \geq b_{rp}) p_a \prod_{a' > a} (1 - p_{a'}) \quad (\text{A.23})$$

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

$$\begin{aligned} \Pr(a \text{ wins at least one impression}) &= 1 - \Pr(a \text{ wins no impression}) \\ &= 1 - (1 - \Pr(a \text{ wins an impression}))^{T_i} \\ &= 1 - \left(1 - \mathbb{1}(b_{i,a} \geq b_{rp}) p_a \prod_{a' > a} (1 - p_{a'}) \right)^{T_i} \end{aligned} \quad (\text{A.24})$$

□

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 50% of these impression-ad pairs have propensity scores lower than 0.01, 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 great job in recovering the treatment effect estimates. In particular, we find that the RMSE for ATE and CATE estimates are 0.025 and 0.169, respectively. When evaluating the gains from our algorithm, we find that $\text{OracleRatio}_{(0.1)}$ for our algorithm is 0.998, implying that targeting the top 10% of impressions based on our algorithm recovers 99.8% 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.

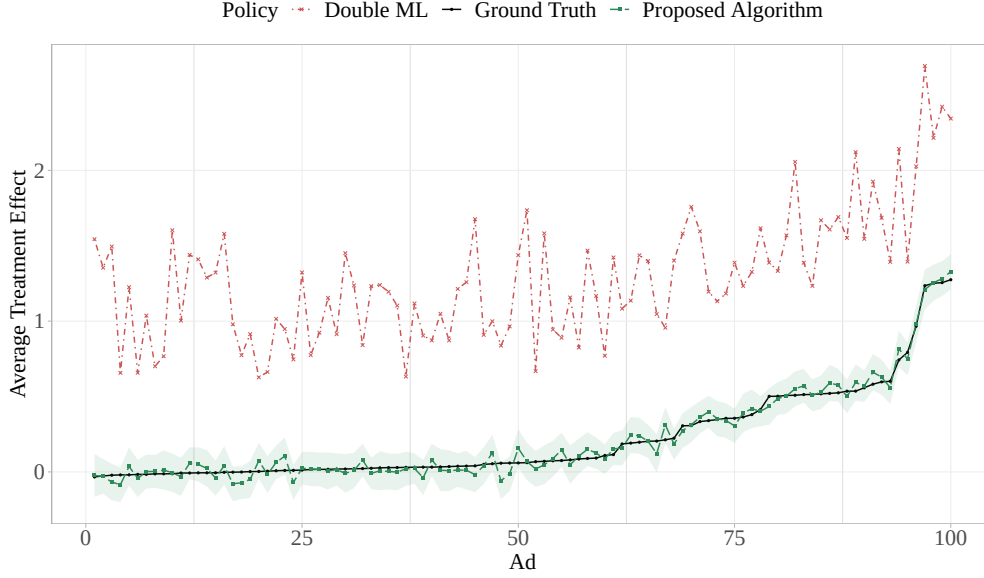


Figure A.4. The performance of the proposed algorithm and Double ML in recovering ATE in alternative ad allocation mechanism.

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 60, but the RMSE reduces for values of A_r higher than 60. The reason for this pattern is that for $A_r > 60$, a tiny fraction of entries in the matrix are in the feasible region (less than 1%) 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 RMSE_{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 σ_v defined in Online 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	
	RMSE _{ATE}	RMSE _{CATE}	Gain _(0.1)	OracleRatio _(0.1)	RMSE _{ATE} ^{DML}	OracleRatio _(0.1) ^{Bid}
<i>Panel 1: Varying bidder-subsampling parameter A_r</i>						
$A_r = 10$	0.055	0.289	0.512	0.995	1.316	0.521
$A_r = 20$	0.124	0.532	0.507	0.986	1.718	0.521
$A_r = 30$	0.342	1.115	0.489	0.951	1.946	0.521
$A_r = 40$	0.341	1.585	0.454	0.882	2.126	0.521
$A_r = 50$	0.446	1.975	0.411	0.798	2.266	0.521
$A_r = 60$	0.450	2.221	0.375	0.729	2.397	0.521
$A_r = 70$	0.397	2.392	0.342	0.665	2.509	0.521
$A_r = 80$	0.358	2.575	0.298	0.580	2.632	0.521
$A_r = 90$	0.367	2.725	0.261	0.507	2.645	0.521
<i>Panel 2: Varying the rank of the CATE matrix</i>						
Rank = 5	0.052	0.239	0.356	0.991	0.904	0.533
Rank = 10	0.055	0.289	0.512	0.995	1.316	0.521
Rank = 15	0.060	0.355	0.631	0.996	1.615	0.517
Rank = 20	0.065	0.430	0.710	0.995	1.829	0.516
Rank = 25	0.068	0.529	0.799	0.995	2.021	0.519
Rank = 30	0.100	0.715	0.858	0.993	2.188	0.515
Rank = 35	0.168	0.902	0.942	0.992	2.381	0.514
Rank = 40	0.229	1.082	0.984	0.990	2.517	0.512
Rank = 45	0.266	1.270	1.065	0.988	2.725	0.511
Rank = 50	0.341	1.416	1.109	0.987	2.864	0.512
<i>Panel 3: Varying the variance of CATE within study</i>						
$\sigma_v = 0.25$	0.044	0.250	0.264	0.984	0.658	0.540
$\sigma_v = 0.50$	0.055	0.289	0.512	0.995	1.316	0.521
$\sigma_v = 0.75$	0.060	0.337	0.758	0.997	1.965	0.515
$\sigma_v = 1.00$	0.067	0.396	1.005	0.998	2.627	0.512
$\sigma_v = 1.25$	0.091	0.470	1.251	0.998	3.294	0.509
$\sigma_v = 1.50$	0.109	0.539	1.497	0.998	3.963	0.508
$\sigma_v = 1.75$	0.123	0.605	1.743	0.998	4.632	0.506
$\sigma_v = 2.00$	0.137	0.690	1.989	0.998	5.276	0.506
$\sigma_v = 2.25$	0.160	0.789	2.235	0.998	5.927	0.506
$\sigma_v = 2.50$	0.165	0.872	2.481	0.998	6.575	0.505
<i>Panel 4: Varying the correlation between bid and CATE</i>						
$\rho(b, \tau) = 0.1$	0.044	0.291	0.512	0.995	0.279	0.137
$\rho(b, \tau) = 0.2$	0.045	0.288	0.512	0.995	0.533	0.233
$\rho(b, \tau) = 0.3$	0.046	0.290	0.512	0.995	0.786	0.329
$\rho(b, \tau) = 0.4$	0.049	0.289	0.512	0.995	1.051	0.425
$\rho(b, \tau) = 0.5$	0.055	0.289	0.512	0.995	1.316	0.521
$\rho(b, \tau) = 0.6$	0.063	0.293	0.512	0.994	1.573	0.616
$\rho(b, \tau) = 0.7$	0.068	0.296	0.512	0.994	1.831	0.712
$\rho(b, \tau) = 0.8$	0.076	0.300	0.512	0.994	2.098	0.808
$\rho(b, \tau) = 0.9$	0.097	0.311	0.511	0.994	2.365	0.904

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

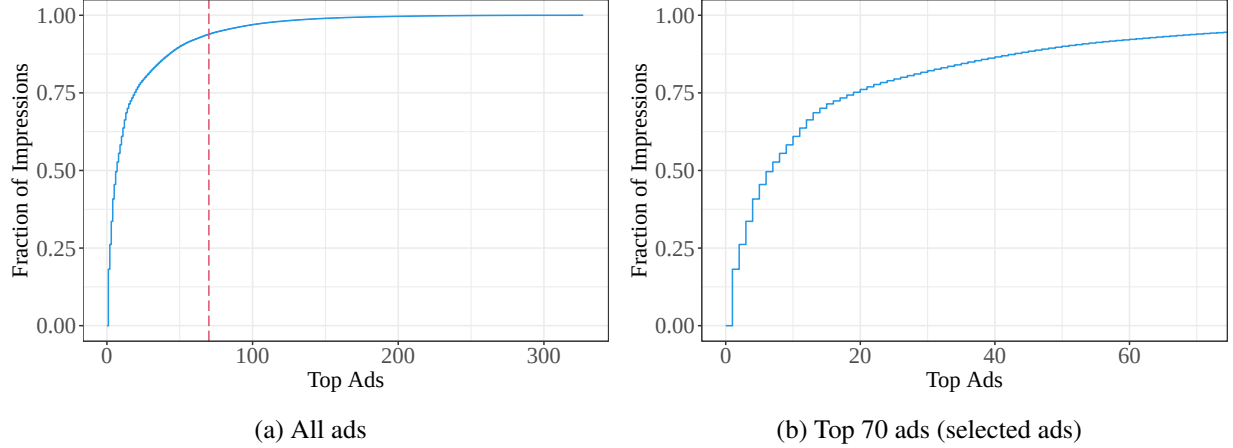


Figure A.5. Cumulative share of impressions generated by top K ads

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

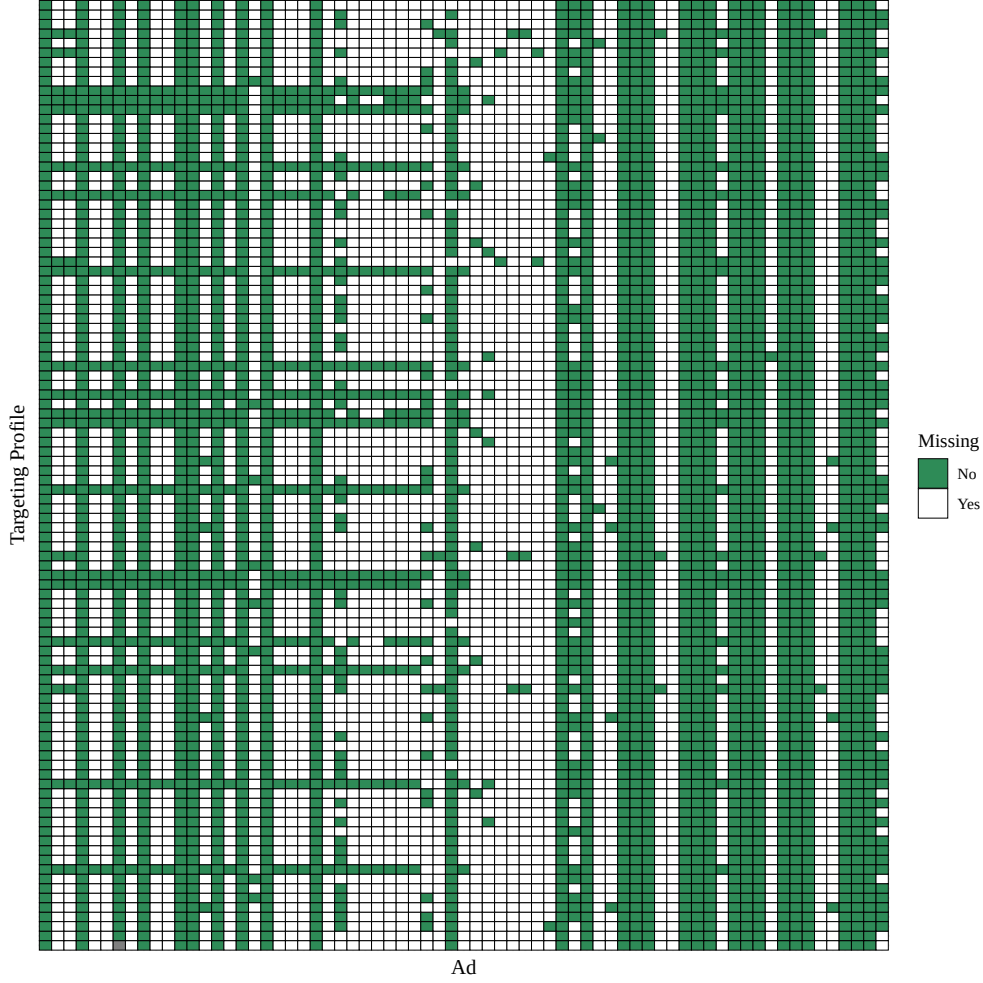


Figure A.6. Missingness pattern in a sub-sample of the CATE matrix

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

$$\hat{\tau}^{(a)}(\cdot) = \underset{\tau \in \mathcal{H}}{\operatorname{argmin}} \sum_{i \in \mathcal{I}_a} \left(\left(Y_i^{(a)} - \hat{m}^{-k_a(i)}(X_i) \right) - \left(W_i^{(a)} - \hat{e}^{-k_a(i)}(X_i) \right) \tau^{(a)}(X_i) \right)^2 + \lambda \Omega_{\mathcal{H}}(\tau), \quad (\text{A.25})$$

where $\hat{m}^{-k_a(i)}(\cdot)$ and $\hat{e}^{-k_a(i)}(\cdot)$ 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 λ .

- Once we estimate the CATE function $\hat{\tau}^{(a)}(\cdot)$ 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.

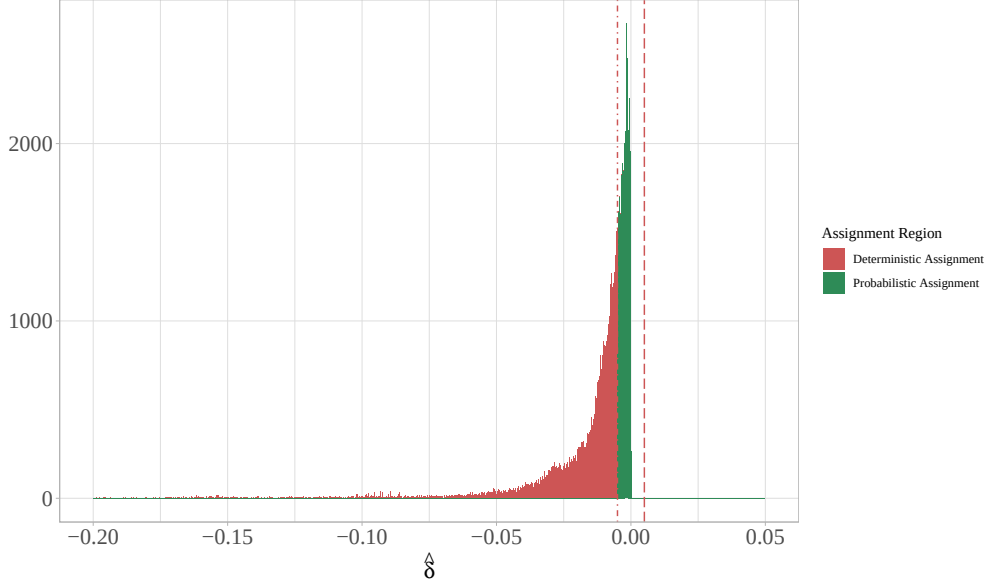


Figure A.7. Assignment regions based on the distribution of $\hat{\delta}$

Method	Gain ₊	OracleRatio ₊
Proposed Algorithm	0.00296	98.80%
Ad Allocation Algorithm ($\hat{\delta}$)	0.00015	4.92%
Ground Truth	0.00300	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 $\hat{\delta}$

Figure A.7 shows the distribution of $\hat{\delta}$ 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 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 $\hat{\delta}$ are largely negative, so less than 10% of them are selected for re-allocation.

Method	ATE	RMSE _{CATE}	Gain _(0.1)	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.

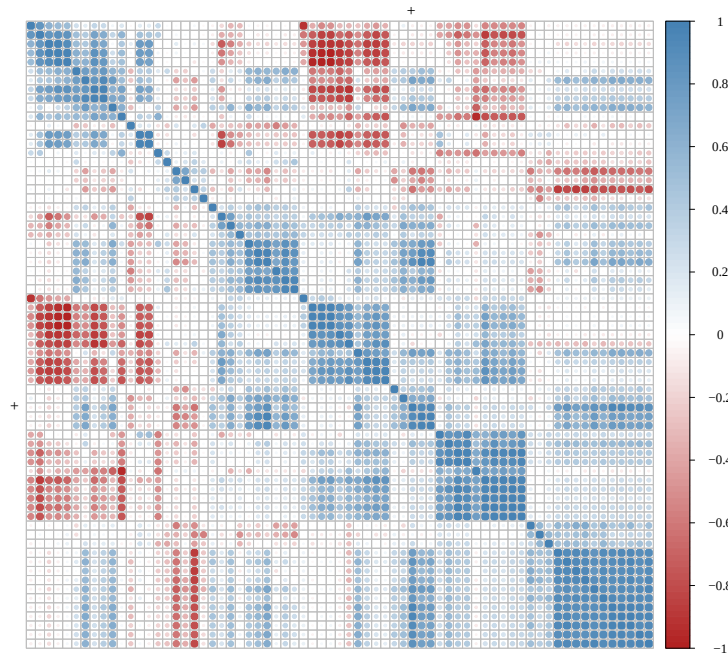
E.7 Inclusion of Irrelevant Ads

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.

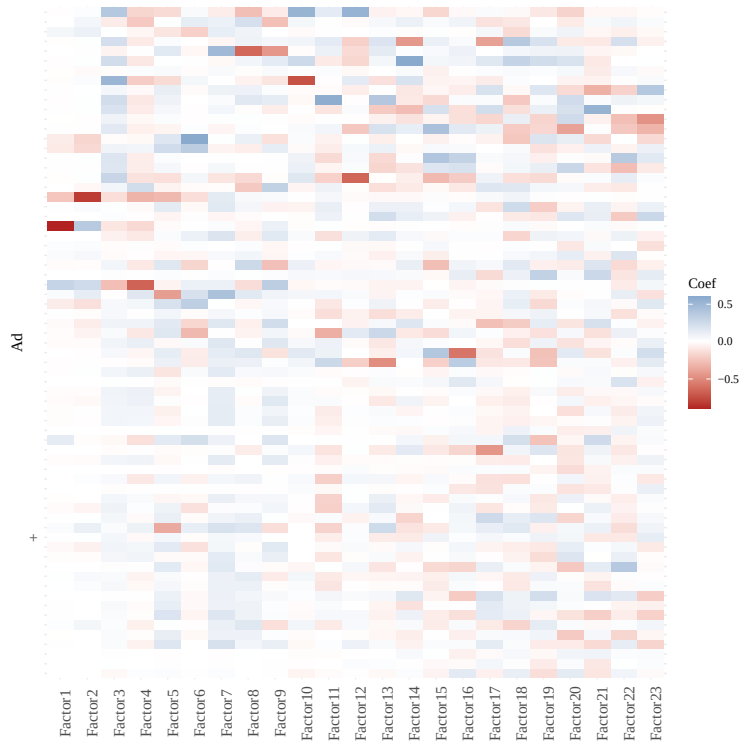
We present the results in Table A3. As shown in this table, adding 20 irrelevant ads does not have a major impact on the algorithm. We see a slightly better ATE estimation performance and a slightly worse CATE estimation and targeting performance compared to the algorithm with only 68 original ads. On 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 to ignore irrelevant 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.¹ 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

¹Please note that the rank of the matrix using the entries in training is 22, but when estimated with the same λ 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:

$$\pi_i^{(j)} = \Pr(S_{i,j} + \omega_{i,j} \geq c) \quad (\text{A.26})$$

In our simulation exercise, we set $c = 1$ and the correlation between algorithmic scores and true CATEs to be 0.5, and draw ω 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 great 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 $\text{OracleRatio}_{(0.1)}$ for our algorithm to be 0.997, highlighting its strong performance in estimating CATEs and learning to prioritize.

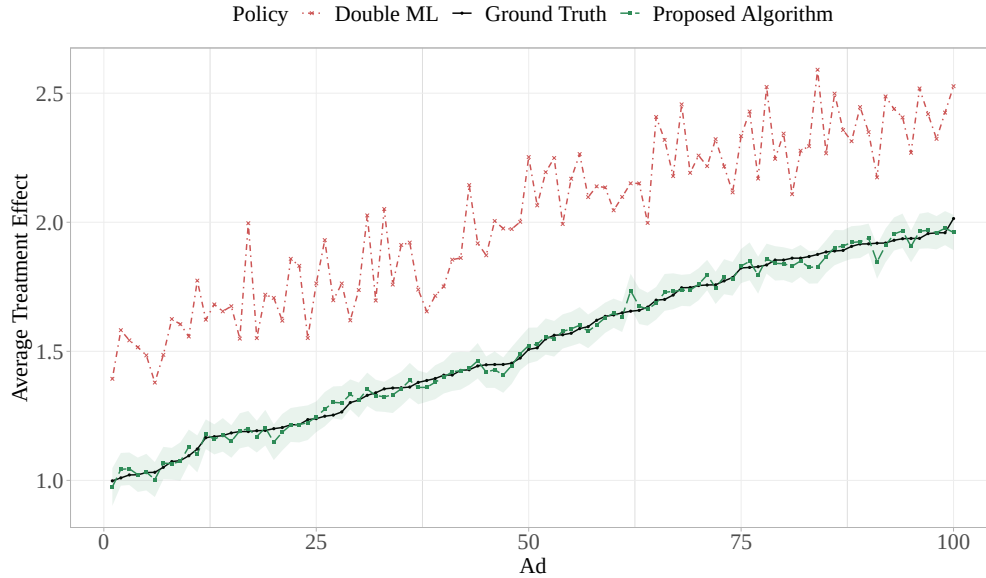


Figure A.10. The performance of the proposed algorithm and Double ML in recovering ATE in settings with cutoff-based algorithmic decision-making.

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.

References

- 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.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
- 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.
- T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.
- 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.
- 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.
- 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.
- R. B. Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

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