Personalized Incentive Alignment: Correcting Utility-Driven Selection Bias in A/B Tests

Jiachun Li MIT
Yang Meng UChicago
David Simchi-Levi MIT
Chonghuan Wang UT Dallas

Abstract

Although A/B testing is a powerful tool for estimating the average treatment effect (ATE), it often proves impractical in social or commercial settings because ethical and business constraints induce participant non-compliance. For example, patients may refuse assignment to less promising therapies, and users may choose whether to adopt a newly released feature based on personal preferences. In this work, we posit that participants act to maximize individual incentives. To capture this behavior, we adopt a utility-based random choice model that explicitly characterizes the identification bias introduced by self-selection and the estimation instability caused by feature imbalance. We then demonstrate how heterogeneous incentives generate both selection bias and inflated variance. Building on these insights, we design an optimal incentive mechanism that equalizes preference distributions across treatment arms, thereby achieving a more balanced covariate profile, lower variance, and a sharper identified set with minimal bias. Finally, we propose an online learning framework that adaptively identifies the optimal incentive scheme during the experiment and produces valid treatment-effect estimates. We validate our theoretical results through both simulation studies and field experiments.

1 INTRODUCTION

Causal inference, particularly the estimation of the average treatment effect (ATE), has become central to the regulatory evaluation of new therapeutics and to decision-making about feature launches on online platforms. Conventional A/B testings randomize participants into treatment and control groups, ensuring that the two groups are balanced in terms of observed and unobserved characteristics. These types of randomized control trials (RCTs) have many desired statistical properties and is now widely adopted in clinical research, public health interventions, and social sciences to evaluate the efficacy of treatments and interventions. However, in many cases RCTs are infeasible either due to ethical or practical restrictions, or due to unaffordable costs ([Johnston et al., 2006], [Haussmann et al., 2023]). Consider a common setting on large online platforms (e.g., Netflix) where the objective is to estimate expected watch time under alternative UX designs and recommendation algorithms. Rather than randomizing and compelling users to adopt different systems, the platform often releases a β version and invites users to opt in, thereby inducing self-selection based on their preferences.

The issue of estimating causal effects when individuals fail to adhere to their assigned treatment group is referred to as the non-compliance problem in econometrics and causal inference. This issue is a well-motivated example for the instrumental variable (IV) approach and has been extensively studied since the seminal work of ([Amemiya, 1974]). However, in this study, we depart from this line of work, since they typically require strong structural assumptions, or focus on other estimand like local average treament effect (LATE), while the primary focus of the present work is average treatment effect.

On the other hand, drawing from perspectives in economics and operations research ([McFadden, 1974], [Greene, 2012], [Train, 2009]), If non-compliance is systematic rather than arbitrary and individuals make rational choices, it can be modeled with random utility theory—e.g., multinomial logit (MNL)—to characterize non-compliance behavior. The resulting self-selection induces acquisition bias [Chen et al., 2021, Dubé et al., 2010], which carries over to the potential-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

outcomes framework because utilities correlate with expected outcomes. For instance, users opting into the $\beta-$version of a streaming service tend to be heavier, more loyal users, so observed gains (e.g., weekly viewing time) overstate the treatment effect for the population. Since the target is the population ATE, the standard unconfoundedness assumption fails and conventional estimators break down. This motivates our first research question:

What is the effect of self-selection process on the estimation of treatment effect, and how can we reduce such effect and provide consistent and stable estimation?

When all confounders are observed, we prove that propensity-score adjustment (via reweighting or stratification) corrects self-selection bias by conditioning on the treatment propensity. However, it can induce high variance under feature imbalance. In the streaming example, if only 5% of the β-group are light watchers, reweighting them heavily to achieve balance (e.g., 50/50) yields unstable estimates because few observations carry large weights. Thus, mitigating imbalance is essential for precise ATE estimation [Hainmueller, 2012, Imai and Ratkovic, 2014, Tan, 2010].

In practice, unobserved confounding is common, precluding point identification without knowledge of the joint potential-outcome distribution. We therefore turn to partial identification, recovering an interval that contains the true effect [Ji et al., 2023, Swanson et al., 2018]. In our setting—where confounders affect both utility and outcomes—we provide insights on (i) the magnitude of selection bias, (ii) its direction, and (iii) its relative contributions across treatment and control, and use these insights to derive optimal identification bounds.

The issues of self-selection bias and high variance stem from an **incentive misalignment** between the experimenter and the participants. Participants make choices based on their personal preferences or utilities, which often conflict with the experimenter's objective of achieving balanced and unbiased treatment assignment. To address this misalignment, we introduce external intervention mechanisms designed to realign incentives to achieve preference balance. By doing so, we can reduce bias and achieve a more stable and reliable estimation of treatment effects. We prove that incentivizing the most imbalanced features can greatly reduce the variance of estimation. Moreover, we also prove that balancing the preferences can reduce the length of partial identification interval with unobservable confounders, therefore showing that aligning incentives can provide more stable and accurate estimation of treatment effect.

While we demonstrate that aligning incentives can mitigate feature imbalance and selection bias, another inherent challenge emerges: the design of an effective incentive alignment mechanism is heavily dependent on the underlying utility-based choice model, which is often unknown. Consequently, a natural approach is to estimate the utility function during the experiment and subsequently adjust the incentive mechanism accordingly. Furthermore, in practice, the most straightforward way to balance preferences is by providing a bonus for the less preferred option. However, this approach can be quite costly, which leads to the consideration of a total bonus budget. The optimal allocation of this budget throughout the experimental process presents an additional layer of complexity. Therefore, it is natural to ask:

Can we adaptively design the incentive mechanism to achieve the optimal estimation accuracy under the budget constraint, such that no prior information about the utility function will not impact the estimation accuracy?

To address this issue, we adopt a **learning to incentivize mechanism** that provides non-asymptotic guarantees for the mean squared error (MSE) of treatment effect estimation. Adaptive experimentation has become widely utilized in A/B testing literature due to its ability to improve experimental efficiency by leveraging information from historical results ([Simchi-Levi and Wang, 2024],[Li et al., 2024]). However, most existing literature on adaptive experimentation directly determines treatment assignments, which is infeasible in our setting. In contrast, we propose an adaptive design of the incentivize mechanism that adjusts the choice probabilities for each experimental unit, a concept also explored in incentive-compatible online learning and bandit literature ([Mansour et al., 2015],[Gonen and Pavlov, 2007]). We adapt this framework to adaptive statistical inference with the objective of minimizing the MSE of average treatment effect (ATE) estimation.

Online experimentation often yields dependent observations that hinder standard bias–variance analysis ([Zhang et al., 2022, Dimakopoulou et al., 2021]). We address this by recasting variance minimization as a low-switching contextual bandit problem: the incentive policy is updated only $O(\log n)$ times, creating i.i.d. batches that both enable consistent estimation of the utility function and maintain a weakly dependent structure for inference. A total-budget constraint introduces further complexity. We develop a tailored procedure that exploits an "equal-derivative" characterization of the static optimum: we form confidence intervals for the utility and compute a pessimistic incentive via this structure, yielding a budget-feasible, near-optimal mechanism with infrequent updates. This low-switching design, together with our non-asymptotic

MSE bounds, provides new tools for analyzing non-compliance and self-selection in adaptive experiments.

Finally, we evaluate the mechanism through extensive simulations and a field experiment. Approximately 700 participants reported their genre preferences and then rated one of two AI-generated videos (romance or sci-fi); we offered additional advertising exposure as an incentive when they selected their preferred genre. Across both simulations and the field study, the incentive mechanism substantially reduces selection bias, improves covariate balance, and yields more stable treatment-effect estimates.

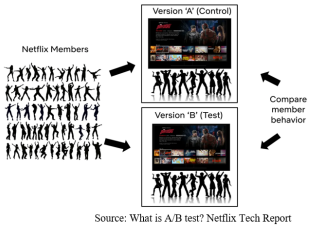

Figure 1: An example of self-selected experiment

2 PROBLEM FORMULATION

2.1 Self-selection Experiment

Classical causal inference method heavily relies on two assumptions: unconfoundedness and overlap condition. While these assumptions are satisfied in randomized control trials (RCTs), in many scenarios it is infeasible due to ethical or economical constraints. Instead of being randomly allocated to certain treatments, the participants of experiment self-select a treatment based on some underlying, possibly unobservable idiosyncratic factors. Specifically, each participant i is faced with a binary set of choice alternatives. Also, each participant i is endowed with a vector observable of features, denoted by $X_i \in \mathcal{X}$ that influence the participant's decision. We assume that the features $\{X_i\}_{i=1}^n$ are drawn i.i.d. from a distribution P_X over $\mathcal{X}$ and conditioned on X_i, the participant's decision process is governed by a utility function for treatment $j \in \{0,1\}$ defined as $U_i^{(j)}(X_i) = V^{(j)}(X_i) + \alpha_i^{(j)}$, where

- $V^{(j)}(X_i)$ is the systematic (or observable) component of the utility, possibly depending on the participant's characteristics, the nature of the treatment, and any relevant covariates.
- $\alpha_i^{(j)}$ is an idiosyncratic (or random) component capturing unobserved preference heterogeneity, measurement errors, or behavioral noise.

And the decision W_i is decided to maximize the utility $W_i = \arg\max_{j\in\{0,1\}} U_i^{(j)}(X_i)$. Notice that equivalently, we have

$$P(W_i = 1|X_i) = P(\text{gap}(X_i) + \alpha_i^{(1)} > \alpha_i^{(0)}), \qquad (1)$$

where $\text{gap}(X_i) = V^{(1)}(X_i) - V^{(0)}(X_i)$ is the utility gap between two treatments. While in general $V^{(1)}, V^{(0)}$ can not be identified in a choice model, the utility gap is uniquely determined. This unobserved preference factor α_i, is essentially in the same spirit as the "unobservable confounders" in causal inference literature, which affects both the treatment assignment W_i and the outcome. Therefore, it violates the most fundamental unconfoundedness condition. The key modeling choice is the distributional assumption on α_i. Commonly used families include:

(1) Logit models, assuming $\alpha_i^{(j)}$ are drawn i.i.d. from the Type I extreme value distribution (Gumbel distribution), which leads to a logistic probability of choosing alternative a.

(2) Probit models, assuming $\alpha_i^{(j)}$ are drawn i.i.d. from a Normal distribution, yielding a Gaussian specification for choice probabilities.

In this work, we focus on these two choice models due to their widespread use.

2.2 Data Generating Process

We draw i.i.d. samples $X_1, X_2, \ldots, X_n$ sequentially from distribution P_X on some compact set $\mathcal{X}$, where n is fixed and known to the experimenter. There are two groups that participants can choose, treatment and control group. For i th experiment participant, the experimenter observes the covariate X_i generated from P_X, and the experiment participant chooses a treatment $W_i \in \{0,1\}$ based on his own utility and choice model (1). Denote the propensity score $e_i(X_i) := \mathcal{P}(W_i = 1|X_i)$. We also assume that $\eta < e(X) < 1 - \eta$ for some constant η. The observed treatment outcome is denoted as $Y_i := Y^{(W_i)}(X_i)$. The potential outcome $Y^{(1)}(X)$ is generated from distribution $P_{Y^{(1)}|X}$ with expectation $\mu^{(1)}(X)$ and variance σ^2. Similarly, the outcome $Y^{(0)}(X)$ is generated from distribution $P_{Y^{(0)}|X}$ with expectation $\mu^{(0)}(X)$ and variance σ^2. We equivalently represent these outcomes as: $Y_1 := Y^{(W_1)}(X_i) = \mu^{(1)}(X_i) + \delta_i^{(1)}$ and $Y_0 := Y^{(W_0)}(X_i) = \mu^{(0)}(X_i) + \delta_i^{(0)}$, where $\delta_i^{(1)}, \delta_i^{(0)}$ represents random noise or error terms. We assume that the expectation $\mu^{(1)}(X)$, $\mu^{(0)}(X)$ belongs to a bounded function class $\mathcal{F}_\mu$. Finally, We use the notation $\mathcal{F}_i$ to represent all information collected for the first i experiment subjects: $\{(X_1, W_1, Y_1), ..., (X_i, W_i, Y_i)\}$. The causal effect of interest is the average treatment effect (ATE):

$$\tau := \mathbb{E}[\mu^{(1)}(X) - \mu^{(0)}(X)]$$

where the expectation is taken over P_X. After collecting all data from the experiment, the experimenter constructs an estimator $\hat{\tau}$ of τ which is a measurable

function that depends on collected data $\mathcal{F}_n$. Our goal is to minimize the expected square loss of $\hat{\tau}$, $\mathbb{E}[(\hat{\tau}-\tau)^2]$, where the expectation is taken over all data and treatment randomness.

2.3 Incentivize Mechanism

Selection bias arises when participants self-select into groups based on their preferences, leading to imbalanced treatment allocation. While propensity score methods can correct for the observable bias, they often result in large estimation variance. Moreover, in our model the existence of idiosyncratic utility preference $\alpha_i^{(j)}$ serve as an unobserved confounder, which cannot be corrected and the causal effect will be **unidentifiable**. Therefore, we could only hope to find the sharp identified set of treatment effects. To simultaneously reduce **both the identified set and variance caused by self-selection**, we propose incentivizing participants to achieve a more balanced allocation across groups, which is known as encouragement design in econometrics literature. Suppose we have a total budget B to be distributed among n trial participants. The policy decides the bonus allocated to each participant denoted as $p(X_i)$ for $i = 1, ..., n$. For instance, if a participant with feature X_i prefers the treatment group ($\text{gap}(X_i) > 0$), the policy provides an incentive $p(X_i)$ to encourage them to join the control group. This shifts the propensity score for the treatment group closer to 1/2 by shifting the utility for control group as $V^{(0)}(X_i, p(X_i)) = V^{(0)}(X_i)+\beta(X_i)p(X_i)$, where the utility is a linear function w.r.p to the incentive with (possibly heterogeneous) sensitivity parameter $\beta(X_i)$. Or equivalently, the propensity score can be characterized by

$$e(X_i, p(X_i)) = P(W_i = 1 \mid X_i) = P(\text{gap}(X_i) - \beta(X_i)p(X_i) \mid X_i).$$

By subsidizing the less preferred choice, the policy aligns the incentives of the experimenter with those of the participants. This incentivization encourages participants to allocate more evenly across groups, achieving greater feature balance and reducing variance in the experiment. In this work, we make the assumption that the incentive only affects the outcome through treatment, i.e., w.p. 1

$$Y^{(W_i)}(X_i) = Y^{(W_i)}(X_i, p(X_i)).$$

This is the basic assumption in instrumental variable literature. In this sense, the incentive is also a valid instrument, which is in the same spirit as existing encouragement literature.

3 OPTIMAL INCENTIVIZE MECHANISM

3.1 Estimator

Assume that we have an experiment with a collected dataset $\{X_i, W_i, Y_i\}_{i=1}^n$. We employ the **Augmented Inverse Probability Weighting (AIPW) estimator**. Under the assumptions of conditional independence (ignorability), positivity (overlap), and correct specification of either the outcome model or the propensity score model, the true AIPW estimator, denoted as τ_{AIPW}, is defined as:

$$\tau_{\text{AIPW}} = \mathbb{E}\left[\mu^{(1)}(X) - \mu^{(0)}(X) + \tfrac{Y-\mu^{(1)}(X)}{e(X)}W - \tfrac{Y-\mu^{(0)}(X)}{1-e(X)}(1-W)\right],$$

where $e(X)$ is the true propensity score, and $\mu^{(1)}(X), \mu^{(0)}(X)$ are the true conditional expectations of the potential outcomes given covariates X. The AIPW estimator is known to be **unbiased** and **asymptotically efficient**, providing optimal variance among unbiased estimators under mild conditions. The variance of the AIPW estimator is given by:

$$\text{Var}(\tau_{\text{AIPW}}) = \tfrac{1}{n}\mathbb{E}\left[\tfrac{\sigma^2}{e(X)(1-e(X))} + \left(\mu^{(1)}(X) - \mu^{(0)}(X) - \tau\right)^2\right]. \tag{2}$$

The **plugged-in AIPW estimator**, denoted as $\hat{\tau}_{\text{AIPW}}$, replaces the true models with their estimated counterparts, $\hat{e}(X), \hat{\mu}^{(1)}(X)$, and $\hat{\mu}^{(0)}(X)$. The focus of this section is to develop optimal incentivize mechanism to minimize the variance in (2) and selection bias by mitigating the utility gap between alternative options assuming that the utility function is known.

3.2 Variance Reduction Through Covariate Balancing

We start with minimizing the variance by incentivizing trial participants to choose the less preferred treatment, thereby achieving a more balanced distribution of features. This balance helps reduce the first term in the variance expression. Thus, the question is: **What is the optimal incentive policy $p(X)$ that minimizes the variance under the given constraints?**

To address this, we frame the problem as an optimization task, with the decision variable $p(X)$ representing the incentive policy:

$$\min_{p(X)} \mathbb{E}[(\frac{\sigma^2}{e(X)(1-e(X))}) + (\mu^{(1)}(X) - \mu^{(0)}(X) - \tau)^2]. \tag{3}$$

Observe that only the first component is influenced by $p(X)$, we isolate this term and denote it as $F = \frac{\sigma^2}{e(X)(1-e(X))}$. The optimization problem can then be

re-expressed as:

$$\begin{aligned} \min_{p(X)} \quad & \int F(X, p(X)) P_X \, dX \\ \text{s.t.} \quad & n \, \mathbb{E}\,[p(X)] \le B, \\ & p(X) \ge 0, \quad \forall X. \end{aligned}$$

The first constraint ensures that the total expected spending over all participants does not exceed the budget B. The second constraint ensures the non-negativity of our incentive policy. By framing the problem this way, we focus on determining the optimal allocation of incentives that balances treatment assignment while adhering to resource constraints.

Next we will provide an outline of the solution to the above optimization problem. We first prove the property of $F(X, p(X))$ in the following lemma:

Lemma 3.1. *For each fixed X, the function $F(X, p(X)) = \frac{\sigma^2}{e(X)(1-e(X))}$, where $e(X)$ is the propensity score influenced by the incentive policy $p(X)$, is convex with respect to $p(X)$.*

The convexity of $F(X, p(X))$ ensures that the optimization problem for minimizing the variance of the AIPW estimator has a well-defined solution. We derive a unique closed-form solution based on the principle of equalizing the marginal rate of variance reduction per unit of incentive. We refer to this as **Equal Derivative Solution**, where the optimal incentive policy ensures that the decrease in variance with respect to the incentive, $\frac{dF}{dp}$, is balanced across participants up to a threshold. This solution leverages the convexity of F to guarantee the existence and uniqueness of a threshold parameter λ.

Theorem 3.2. *[Equal Derivative Solution] Under the convexity of F with respect to $p(X)$, there exists a unique threshold λ such that the optimal incentive policy $p^*(X)$ satisfies:*

- *If $\frac{dF}{dp} < -\lambda$, then $p^*(X) = g_\lambda(X) > 0$.*
- *If $\frac{dF}{dp} \ge -\lambda$, then $p^*(X) = 0$.*

Furthermore, if the sensitivity of incentive $\beta(X) \equiv \beta$ is homogeneous across features, we have:

- *If $e(X) < \eta \le \frac{1}{2}$ or $1 - \eta < e(X) \le 1$, the optimal policy adjusts the propensity score to $e^*(X, p^*(X)) = \eta$.*
- *If $\eta \le e(X) \le 1 - \eta$, no inventives are applied, and the propensity score remains $e^*(X, 0)$.*

The key insight in theorem 3.2 lies in targeting participants with the most imbalanced propensity scores, as incentivizing them yields the fastest variance reduction under the same budget constraints.

3.3 Selection Bias Mitigation via Preference Alignment

As discussed above, the ATE is identifiable only under conditional independence. In our self-selection setting, however, this assumption generally fails when unobserved preference factors $\alpha_i^{(1)}, \alpha_i^{(0)}$ also affect the outcome. For instance, in the live-streaming application, a viewer's watch time on another platform (e.g., YouTube) may shape their preference for the recommendation interface while remaining unobserved, thereby violating conditional independence. Mathematically, we can decompose the outcome of treatment group as (and similarly for control group)

$$\begin{aligned} Y^{(1)}(X_i, \alpha_i^{(1)}) &= g^{(1)}(X_i, \alpha_i^{(1)}) + \eta_i^{(1)} \\ &= f^{(1)}(X_i) + \text{bias}^{(1)}(X_i, \alpha_i^{(1)}) + \eta_i^{(1)}, \end{aligned} \tag{4}$$

which consists of an independent noise η_i, a baseline observable outcome and a confounding selection bias

$$f^{(1)}(X_i) := \mathbb{E}_{\alpha_i^{(1)}}[g^{(1)}(X_i, \alpha_i^{(1)})], \tag{5}$$

$$\begin{aligned} \text{bias}^{(1)}(X_i, \alpha_i^{(1)}) &= g^{(1)}(X_i, \alpha_i^{(1)}) \\ &\quad - \mathbb{E}_{\alpha_i^{(1)}}[g^{(1)}(X_i, \alpha_i^{(1)})]. \end{aligned} \tag{6}$$

Note that the bias term is never **observable**, thus there exists multiple different outcome functions that will lead to the same distribution of observation. To minimize structural assumption, we only assume that the selection bias term $\text{bias}^{(1)}(X_i, \alpha_i^{(1)})$ is ε- lipschitz continuous. And we have parallel notations and assumptions for the control group $Y^{(0)}$. The parameter ε represents the belief of the marginal of unobservable confounding effect. In particular, when $\varepsilon = 0$, it is equivalent to conditional independence assumption. Then we have the following characterization of the identified set of ATE:

Theorem 3.3. *The identified set of ATE is*

$$[\mathbb{E}_{\text{AIPW}}[Y^{(1)} - Y^{(0)}] - \text{bias}, \mathbb{E}_{\text{AIPW}}[Y^{(1)} - Y^{(0)}] + \text{bias}],$$

where the selection bias is

$$\text{bias} = \varepsilon \, \mathbb{E}_X \left[\text{Cov}\left[\alpha, H(\alpha, \text{gap}(X))\right] + \text{Cov}\left[\alpha, H(\alpha, -\text{gap}(X))\right] \right].$$

Here $H(\alpha, \beta) = \frac{\mathrm{P}(\alpha < \alpha' - \beta)}{\int_{\alpha'} \mathrm{P}(\alpha < \alpha' - \beta)}$ is a random variable transformation, where α, α' are i.i.d. copies from the same distribution as the unobservable utility factor.

Theorem 3.3 implies that the true ATE is only **partially identified**: it lies in a set centered at the AIPW estimand, with a radius that quantifies the magnitude of selection bias. Unlike ordinary estimation error, this bias is **irreducible**—it cannot be eliminated by

larger samples or alternative estimators, as it stems from unobserved confounding. Because $\text{bias}^{(1)}, \text{bias}^{(0)}$ are unobserved and not identifiable, our set provides a worst-case bound over all models consistent with the constraints, capturing the largest possible selection bias. The bound grows linearly in ε, matching the intuition that a larger confounding budget yields a proportionally wider ATE interval. With this explicit analytical characterization of selection bias, we (i) characterize how the utility gap governs the size of the identified set, and (ii) design incentivize mechanisms that optimally contract that set. Denote $\text{bias}(X) = \text{Cov}\left[\alpha, H(\alpha, \text{gap}(X))\right] + \text{Cov}\left[\alpha, H(\alpha, -\text{gap}(X))\right]$ as the selection bias for each feature X.

Proposition 3.4. *The selection bias shrinks with* $|\text{gap}(X)|$. *Moreover, the optimal incentivize mechanism also satisfies equal-derivative property:*

1. *If* $\frac{d\,\text{bias}}{dp} < -\lambda$, *then* $p^*(X) > 0$;
2. *If* $\frac{d\,\text{bias}}{dp} \geq -\lambda$, *then* $p^*(X) = 0$.

Furthermore, if $\beta(X) \equiv \beta$ *for all feature* X, *the optimal mechanism is the same as in theorem 3.2.*

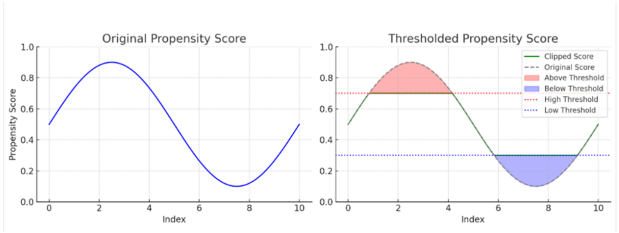

Figure 2: An Illustration of Equal Derivative Design

Thus, we have shown that **both variance and selection bias are driven by the utility gap** between the two choices, and that an optimal incentive mechanism can effectively reduce both.

4 LEARNING TO INCENTIVIZE: ADAPTIVE INCENTIVE DESIGN ALLOCATION

In section 3, we show the optimal incentivize mechanism to mitigate selection bias and reduce estimation variance given a fixed utility-based choice model. In particular, we show that the **Equal Derivative Solution** provides the optimal allocation mechanism. However, in most real-world settings, it is neither feasible nor reasonable to assume that customer utilities or outcomes models are fully known before conducting experiments. Instead, these models must be learned during the course of the experiment. This necessitates the use of **adaptive experimental designs**, where participants' observed behaviours and responses dynamically inform subsequent incentive allocations. For brevity, we assume that $\beta(X) \equiv \beta$ and logit choice model in this section, since the optimal mechanism for bias reduction and variance minimization coincides in this case. In the general case, the two objectives need not coincide; a weighted combination can be optimized to yield a tight confidence interval, and the experimental design and analysis extend straightforwardly. We propose the following adaptive learning-to-incentivize algorithm 1 that

(1) Learns the underlying choice and outcome models during the experiment, and

(2) Constructs a estimation $\hat{\tau}$ that approximates the optimal variance V^*: $\text{Var}(\hat{\tau}) \leq (1 + o(1))V^*$ and exhibits minimal potential selection bias denoted as bias^*.

Algorithm 1 Low-Switching Learning-to-Incentivize Algorithm

Input: Total sample size n, number of batches K, initial batch-length coefficient c
Output: Final unbiased estimator $\hat{\tau}$
Initialization: Set initial policy $p(X) = 0$
for $k \leftarrow 1$ **do**
    Set batch length $n_k \leftarrow c\sqrt{n}$
    Collect n_k samples using policy $p(X) = 0$
    Estimate $\text{gap}(X)$ using logistic regression.
end for
for $k \leftarrow 2$ **to** K **do**
    Set batch length $n_k \leftarrow 2^{k-2}\, c\sqrt{n}$
    Solve the optimization problem with proportional budget to compute estimated policy $\hat{p}(X)$
    Compute conservative incentive allocation policy $p^{\text{LB}}(X) = \hat{p}(X) - 2O\left(\frac{1}{\sqrt{n_k}}\right)$
    Collect n_k samples using policy $p^{\text{LB}}(X)$
    Update model parameters β, $\text{gap}(X, p(X))$ via logistic regression.
end for
Compute final unbiased estimator $\hat{\tau}$ based on all collected data

We propose a **low-switching learning-to-incentivize algorithm**, which operates over $\log(n)$ sequential batches, dynamically refining the incentive policy to balance exploration and exploitation. Below, we outline the key steps and theoretical guarantees of the algorithm. In the first batch, no incentives are offered ($p = 0$). This phase serves as a pretraining phase to have an initial estimate of the choice model for better incentive allocation mechansim in latter batches.

In the following batches, we can solve the optimal incentive allocation design problem with our estimated models. With standard non-parametric regression oracle (which we describe in appendix), we have the following guarantee on approximating the optimal in-

centive allocation mechanism.

Lemma 4.1. *Under appropriate regularity conditions, the estimated mechanism $\hat{p}(X)$ at each batch k with length n_k is close to the true optimal allocation mechanism $p^*(X)$, with a high degree of confidence:*

$$\mathbb{E}[(p^*(X) - \hat{p}(X))^2] \leq C_1 n_k^{-\alpha_1},$$

where α_1 is the nonparametric convergence rate for estimating the utility function $V^{(1)}(X), V^{(0)}(X)$ and utility gap $\text{gap}(X)$.

To ensure conservativeness, we always select the lower bound of the confidence interval for incentives at batch k: $\hat{p}^{\text{LB}}(X) = \hat{p}(X) - 2C_1 n_k^{-\alpha}$. We apply the conservative incentive policy $\hat{p}^{\text{LB}}(X)$, run experiments, collect data, and re-estimate models. This process is repeated for each batch until all batches are completed. Our algorithm is designed with a conservative strategy to ensure that the total budget is respected. Specifically:

Proposition 4.2. *With high probability, the algorithm will not exceed the allocated budget during the course of the experiment.*

After completing all batches, we guarantee that the output estimator $\hat{\tau}$ satisfies the following: 1. The estimator $\hat{\tau}$ exhibits minimal selection bias. 2. The variance of $\hat{\tau}$ is close to the theoretical lower bound V^*. In particular, we have the following guarantee:

Theorem 4.3. *After completing the experiment, the mean square error of the estimator $\hat{\tau}$ satisfies:*

$$\mathbb{E}[(\hat{\tau} - \tau)^2] \leq (1 + o(1))V^*,$$

where $\tau = \mathbb{E}_{\text{AIPW}}[Y^{(1)} - Y^{(0)}]$ is the average treatment effect assuming unconfoundedness, and also the midpoint in the identified set in section 3.

We can also demonstrate a central limit theorem, showing that the estimator $\hat{\tau}$ converges to the normal distribution with the smallest achievable variance V^*. Combining everything together, we can provide a valid confidence interval considering both bias and variance:

Theorem 4.4. *A valid $1-\alpha$ confidence interval for the average treatment effect is $\left[\hat{\tau} - z_{1-\alpha/2}\frac{V^*}{\sqrt{n}} - \text{bias}^*, \hat{\tau} + z_{1-\alpha/2}\frac{V^*}{\sqrt{n}} + \text{bias}^*\right]$, where z is the quantile for standard normal distribution.*

The resulting confidence interval achieves (asymptotically) minimal length among valid $(1-\alpha)$ intervals in our class of procedures, which is the paper's main contribution. Beyond the usual variance term, the length features an additional component quantifying selection bias. And V^*, bias^* can be estimated by plugging in the approximated utility function. This component yields a principled method for analyzing and testing the average treatment effect under self-selection, and also informs an incentive-allocation rule designed to minimize inferential uncertainty. We further highlight the practical efficiency of the proposed low-switching learning-to-incentivize framework. Implementation requires only few queries to a standard offline regression oracle. In addition, data within each batch are i.i.d., which markedly reduces dependence and analytical complexity when characterizing the sampling behavior of τ and could be of independent interests in analyzing adaptive experimentation with potential constraints.

5 FIELD EXPERIMENT ON AI-GENERATED VIDEOS RATING

In this section, we provide numerical results supporting the derived theoretical insights. Due to the limit of space, we only provide the results of the more interesting real world experiment, and a very detailed simulation analysis under different settings is contained in appendix. We recruit 683 participants on Prolific and ask them to evaluate two AI-generated short videos—one in the fantasy genre and one in sci-fi—on overall quality. Before viewing, we collect five key covariates for each person: 1) Age range, 2) Gender, 3) Enjoyment of imaginative or emotional content, 4) Curiosity about scientific or technological topics, 5) Preferred movie genre. All the data and code can be found in https://github.com/papersubmission1319/paper-code.

Sci-Fi Fantasy

Figure 3: Two AI-generated Videos

Participants then indicate which video they would choose to watch; however, we have each participant view and rate both videos, thereby observing both factual and counterfactual outcomes. As expected, stated genre preference strongly predicts choice, so we implement a genre-targeted incentive: participants preferring sci-fi watched an extra 20-second advertisement before the sci-fi video, and similarly for those preferring fantasy.

First, we assess covariate balance before and after introducing this incentive. The table shows that our genre-targeted bonus improved overlap by roughly 3–5 percentage points. Although this increase in balance is modest, the reductions in treatment-control imbalance can yield substantially more stable and accurate aver-

Table 1: Preference Balance in Incentive and No-Incentive Data

Dataset	Choice	Pr(Sci-Fi)	Pr(Fantasy)
With Incentive	Fantasy	0.381	0.619
	Sci-Fi	0.756	0.244
No Incentive	Fantasy	0.333	0.667
	Sci-Fi	0.787	0.213

age treatment effect estimates, as we demonstrate below. We begin by examining the no-incentive condition, comparing six estimators: 1. DIM: Naive difference-in-means estimator. 2. PSM: Propensity-score matching estimator. 3. IPW-Logit: Inverse-propensity-weighted estimator using a logistic regression choice model. 4. IPW-RF: IPW estimator using a random forest choice model. 5. AIPW-Logit: Augmented IPW (doubly robust) with logistic choice model and random forest outcome model. 6. AIPW-RF: AIPW with random forest choice and outcome models.

For all AIPW variants, outcome regressions are fit by random forest. Table 2 presents the estimated average treatment effects along with variance estimates obtained from 200 bootstrap replications.

Table 2: Estimator Results with Bias and ATE Variance Without Incentive

Method	Y1 (b)	Y0 (b)	ATE (b)	Var
True	3.6 (+0.00)	3.7 (+0.00)	-0.12 (+0.00)	—
Naive	3.6 (+0.07)	3.9 (+0.19)	-0.23 (-0.11)	0.017
PSM_ATT	3.6 (+0.07)	4.1 (+0.45)	-0.50 (-0.38)	0.018
IPW_Logit	2.2 (-1.3)	1.7 (-2.0)	0.59 (+0.71)	0.049
AIPW_Logit	3.6 (+0.0090)	4.0 (+0.35)	-0.46 (-0.34)	0.013
IPW_RF	2.9 (-0.63)	2.9 (-0.83)	0.076 (+0.20)	0.041
AIPW_RF	3.6 (+0.014)	3.8 (+0.11)	-0.22 (-0.095)	0.018

The "true" rating refers to the overall mean across all 364 participants. We observe a clear positive selection bias for both videos: participants who favor a given genre are not only more likely to watch that video but also tend to award it higher scores than the general population. Pure IPW methods suffer from excessive variability and often over-adjust, leading to unstable estimates. For the sci-fi video, the AIPW estimator with a logistic-regression propensity model achieves the lowest bias and variance. However, it does not fully correct the bias for fantasy ratings. In contrast, the AIPW estimator that uses a random-forest propensity model is the only method to substantially reduce bias in the fantasy arm—though some residual bias remains.

Next, we evaluate estimator performance under the incentivized design. Of the 319 participants, 155 chose Fantasy and 164 chose Sci-Fi. Again true rating refers to the overall mean across all 319 participants. We apply the same six estimators as in the no-incentive case, and report results in Table 3; variances are again estimated via 200 bootstrap replications.

Compared to the no-incentive condition in Table 2, the incentivized design in Table 3 yields uniformly lower variance and reduced bias across all estimators. For example, IPW_Logit's ATE bias falls from +0.709 to +0.253, variance from 0.0493 to 0.0147. And even the naive DM estimator's variance decreases from 0.0173 to 0.0125. **These dramatic, real-world gains mirror our simulation findings—where AIPW paired with a low-switching incentive policy consistently delivered the most accurate and stable ATE estimates—and confirm that aligning incentives to rebalance covariates can robustly stabilize propensity-based estimators in practice.**

Table 3: Estimator Results with Bias and ATE Variance (Incentive Data)

Method	Y1 (b)	Y0 (b)	ATE (b)	Var
True	3.4 (+0.00)	3.6 (+0.00)	-0.25 (+0.00)	—
Naive	3.5 (+0.14)	3.7 (+0.027)	-0.14 (+0.11)	0.013
PSM_ATT	3.5 (+0.14)	3.5 (-0.096)	-0.014 (+0.24)	0.029
IPW_Logit	3.5 (+0.10)	3.5 (-0.15)	0.0030 (+0.25)	0.015
IPW_RF	3.2 (-0.23)	3.2 (-0.44)	-0.035 (+0.22)	0.035
AIPW_Logit	3.5 (+0.075)	3.6 (-0.037)	-0.14 (+0.11)	0.015
AIPW_RF	3.5 (+0.097)	3.6 (-0.023)	-0.13 (+0.12)	0.016

We also compare the importance of different factors in making choice with and without incentive. The significance of preferred movie genre reduces dramatically, which again shows the effectiveness of incentives. More details can be found in appendix.

6 CONCLUSION

In this paper, we study causal inference in self-selected experiments. We derive sharp partial-identification bounds for the average treatment effect and characterize the associated variance, showing that incentive misalignment is the primary driver of selection bias and covariate imbalance. Building on this insight, we design an optimal incentive mechanism that minimizes both variance and identification set length. We further develop a low-switching learning-to-incentivize framework that attains the same optimal accuracy without a priori knowledge of utilities or the choice model, and we provide valid inference procedures with guaranteed coverage under self-selection. Simulations and a real-world experiment corroborate the theoretical results.

References

Takeshi Amemiya. The nonlinear two-stage least-squares estimator. *Journal of econometrics*, 2(2): 105–110, 1974.

Ningyuan Chen, Anran Li, and Kalyan Talluri. Reviews and self-selection bias with operational implications. *Management Science*, 67(12):7472–7492, 2021.

Thomas Cook, Alan Mishler, and Aaditya Ramdas. Semiparametric efficient inference in adaptive experiments. In *Causal Learning and Reasoning*, pages 1033–1064. PMLR, 2024.

Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34:1939–1951, 2021.

Jean-Pierre H Dubé, Günter J Hitsch, and Pradeep K Chintagunta. Tipping and concentration in markets with indirect network effects. *Marketing Science*, 29 (2):216–249, 2010.

Rica Gonen and Elan Pavlov. An incentive-compatible multi-armed bandit mechanism. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, pages 362–363, 2007.

William H. Greene. *Econometric Analysis*. Pearson Education, 7th edition, 2012.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Marvin Haussmann, Thanh-Mai Solange Le, Ville Halla-aho, et al. Estimating treatment effects from single-arm trials via latent-variable modeling. *arXiv preprint arXiv:2311.03002*, 2023.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 243–263, 2014. doi: 10.1111/rssb.12027.

Wenlong Ji, Lihua Lei, and Asher Spector. Model-agnostic covariate-assisted inference on partially identified causal effects. *arXiv preprint arXiv:2310.08115*, 2023.

S. Claiborne Johnston, Jason D. Rootenberg, Sudha Katrak, W. Scott Smith, and John S. Elkins. Effect of a us national institutes of health programme of clinical trials on public health and costs. *The Lancet*, 367(9519):1319–1327, 2006.

Jiachun Li, David Simchi-Levi, and Yunxiao Zhao. Optimal adaptive experimental design for estimating treatment effect. *arXiv preprint arXiv:2410.05552*, 2024.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, 2015.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

David Simchi-Levi and Chonghuan Wang. Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Management Science*, 2024.

Sonja A. Swanson, Miguel A. Hernán, Megan Miller, James M. Robins, and Thomas S. Richardson. Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018.

Zhiqiang Tan. Bounded, efficient, and doubly robust estimation with inverse weighting. *Biometrika*, 97(3): 661–682, 2010.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition, 2009.

Kelly W Zhang, Lucas Janson, and Susan A Murphy. Statistical inference after adaptive sampling for longitudinal data. *arXiv preprint arXiv:2202.07098*, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Yes]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [No]

Supplementary Materials

A Necessary Tools

In section 4, we need standard statistical regression oracle, which we formalize here.

Assumption A.1. *(Oracle for Distribution Estimation).*
Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples drawn from an unknown continuous distribution P_X with a probability density function $p(x)$. The Kernel Density Estimate (KDE) of $p(x)$ is given by:

$$\hat{P}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where K is a symmetric kernel function and $h > 0$ is the bandwidth parameter. We assume that for any $\epsilon > 0$, the uniform deviation of $\hat{P}_n(x)$ from $p(x)$ is bounded as:

$$\Pr\left(\sup_{x \in \mathcal{X}} \left|\hat{P}_n(x) - p(x)\right| \geq \epsilon\right) \leq C \exp\left(-cnh\epsilon^2\right),$$

where C and c are constants that depend on K, h, and the smoothness of $p(x)$.

Next, we estimate the conditional expectation function. The following assumption guarantees the accuracy of this estimation:

Assumption A.2 (Oracle for Expectation Function Estimation)**.** *Let $(X_1, Y_1), \ldots, (X_m, Y_m)$ denote a batch of i.i.d. data drawn from the joint distribution $P_X \cdot P_{Y|X}$, where $X \sim P_X$ and the conditional expectation is $\mathbb{E}[Y \mid X = x] = \mu(x)$). Assume that the batch size satisfies $m \geq C_3 \log n$ for a sufficiently large constant C_3.*

We posit the existence of a regression oracle that takes $(X_1, Y_1), \ldots, (X_m, Y_m)$ as input and outputs an estimated function $\hat{\mu} : \mathcal{X} \to \mathbb{R}$ approximating $\mu(x)$, such that with probability $1 - \delta$ (for $\delta = \frac{1}{n^4}$):

$$\mathbb{E}_{X \sim P_X}\left[(\mu(X) - \hat{\mu}(X))^2\right] \leq C_4 \frac{\sigma^2}{m^\alpha} \log\left(\frac{1}{\delta}\right),$$

where $C_4 > 0$ is a constant, α depends on the complexity of outcome function class, and σ^2 represents the variance of the noise in the data.

Besides, we also need a logistic regression oracle estimating the utility gap.

Assumption A.3 (Oracle for Utility Gap Estimation)**.** *Let $(X_1, W_1), \ldots, (X_m, W_m)$ denote a batch of i.i.d. data drawn from the joint distribution $P_X \cdot P_{W|X}$, where $X \sim P_X$ and the conditional expectation is $\mathbb{P}[W = 1 \mid X = x] = \frac{1}{1+\exp(-\operatorname{gap}(X))}$). Assume that the batch size satisfies $m \geq C_3 \log n$ for a sufficiently large constant C_3.*

We posit the existence of a regression oracle that takes $(X_1, W_1), \ldots, (X_m, W_m)$ as input and outputs an estimated function $\hat{\mu} : \mathcal{X} \to \mathbb{R}$ approximating $\mu(x)$, such that with probability $1 - \delta$ (for $\delta = \frac{1}{n^4}$):

$$\mathbb{E}_{X \sim P_X}\left[(\operatorname{gap}(X) - \widehat{\operatorname{gap}}(X))^2\right] \leq C_4 \frac{\sigma^2}{m^{\alpha_1}} \log\left(\frac{1}{\delta}\right),$$

where $C_4 > 0$ is a constant, α_1 depends on the complexity of the utility function class $V^{(1)}(X), V^{(0)}(X)$, and σ^2 represents the variance of the noise in the data. We also have $\mathbb{E}[(\hat{\beta} - \beta)^2] \leq C_5 \frac{\sigma^2}{m^{\alpha_1}} \log\left(\frac{1}{\delta}\right)$.

B Proof of Theoretical Results

B.1 Proof of Lemma 3.1

Let gap > 0 and set $s = p - \text{gap}$. Define $q(s) = \mathbb{P}(\varepsilon > s)$ and

$$f(s) = \frac{1}{q(s)\big(1 - q(s)\big)} = \Psi(p) \quad (\text{since } s = p - \text{gap}).$$

Because s is affine in p, convexity of f in s implies convexity of Ψ in p. We prove this for logistic and Gaussian ε.

Logistic noise. If $\varepsilon \sim \text{Logistic}(0, 1)$, then $q(s) = \frac{1}{1+e^s}$ and $1 - q(s) = \frac{e^s}{1+e^s}$. Hence

$$q(s)\big(1 - q(s)\big) = \frac{e^s}{(1 + e^s)^2} \quad \Rightarrow \quad f(s) = \frac{(1 + e^s)^2}{e^s} = e^s + e^{-s} + 2.$$

Differentiate: $f''(s) = e^s + e^{-s} > 0$ for all s, so f is globally strictly convex, hence Ψ is convex in p.

Gaussian noise. If $\varepsilon \sim \mathcal{N}(0, 1)$, let $\phi(s) = \frac{1}{\sqrt{2\pi}} e^{-s^2/2}$ and $\Phi(s) = \int_{-\infty}^{s} \phi$. Then $q(s) = 1 - \Phi(s)$ and

$$D(s) := q(s)\big(1 - q(s)\big) = \Phi(s)\big(1 - \Phi(s)\big), \qquad f(s) = \frac{1}{D(s)}.$$

We show D is log-concave; then $-\log D$ is convex, so $f = e^{-\log D}$ is log-convex and thus convex.

Log-concavity of Φ and $1 - \Phi$. Set $m(s) = \phi(s)/\Phi(s)$ (where $\Phi > 0$) and $r(s) = \phi(s)/(1 - \Phi(s))$ (where $1 - \Phi > 0$). Note $(\log \Phi)' = m$ and $(\log(1 - \Phi))' = -r$. Using $\phi'(s) = -s\phi(s)$ and quotient rule,

$$m'(s) = \frac{\phi'(s)\Phi(s) - \phi(s)\Phi'(s)}{\Phi(s)^2} = -\frac{\phi(s)}{\Phi(s)^2}\big(s\,\Phi(s) + \phi(s)\big) \le 0,$$

where the bracket is ≥ 0 because: if $s \ge 0$ then $s\Phi \ge 0$; if $s < 0$, Mills' bound $\Phi(s) < \phi(s)/|s|$ gives $s\Phi > -\phi$. Similarly,

$$r'(s) = \frac{\phi'(s)(1 - \Phi(s)) + \phi(s)\Phi'(s)}{\big(1 - \Phi(s)\big)^2} = \frac{\phi(s)}{(1 - \Phi(s))^2}\big(\phi(s) - s(1 - \Phi(s))\big) \ge 0,$$

since for $s > 0$ Mills' bound $1 - \Phi(s) < \phi(s)/s$ yields $\phi - s(1 - \Phi) > 0$, and for $s < 0$ we have $-s(1 - \Phi) > 0$. Thus $(\log \Phi)'' = m' \le 0$ and $(\log(1 - \Phi))'' = -r' \le 0$, so both Φ and $1 - \Phi$ are log-concave.

Log-concavity of D and convexity of f. The sum of concave functions is concave, hence $\log D = \log \Phi + \log(1 - \Phi)$ is concave on $\mathbb{R}$. Therefore D is log-concave, so $-\log D$ is convex; consequently $f = e^{-\log D}$ is log-convex and hence convex.

B.2 Proof of Theorem 3.2

We want to minimize the variance lower bound through our decision variable $p(X)$:

$$\begin{aligned} \min_{p(X)} \quad & \int F(X, p(X)) P_X \, dX \\ \text{s.t.} \quad & \int p(X) P_X \, dX \le \frac{B}{n}, \\ & p(X) \ge 0, \quad \forall X. \end{aligned}$$

The Lagrangian $\mathcal{L}(p, \lambda, \mu)$ is:

$$\mathcal{L}(p, \lambda, \mu) = \int \left[F(X, p(X)) P_X\right] dX + \int \mu(X)(-p(X))\, dX + \lambda \left(\int p(X) P_X \, dX - C \right),$$

where $\lambda \ge 0$ and $\mu(X) \ge 0$ are dual variables.
The Karush-Kuhn-Tucker (KKT) conditions provide the necessary conditions for optimality. These include:

1. Stationarity
$$\frac{\partial F(X,p(X))}{\partial p(X)} = \mu(X) - \lambda, \quad \forall X.$$

2. Primal Feasibility:
$$\int p(X)P_X \, dX \leq C, \quad p(X) \geq 0 \quad \forall X.$$

3. Dual Feasibility:
$$\lambda \geq 0, \quad \mu(X) \geq 0 \quad \forall X.$$

4. Complementary Slackness:
$$\lambda \left(\int p(X)P_X \, dX - C \right) = 0, \quad \mu(X)p(X) = 0 \quad \forall X.$$

The Euler-Lagrange equation gives the necessary conditions for $p(X)$ to be a stationary point of the functional. The general form for a functional $\int L(X, p(X), p'(X)) \, dX$ is:
$$\frac{\partial L}{\partial p} - \frac{d}{dX}\left(\frac{\partial L}{\partial p'(X)}\right) = 0.$$

In this case, since the Lagrangian does not depend on $p'(X)$, the Euler-Lagrange equation simplifies to:
$$\frac{\partial}{\partial p(X)} \left[F(X,p(X))P_X + \lambda p(X)P_X + \mu(X)p(X) \right] = 0.$$

Differentiating term by term with respect to $p(X)$ gives:
$$\frac{\partial F(X,p(X))}{\partial p} P_X + \lambda P_X + \mu(X) = 0.$$

If $p(X) > 0$, complementary slackness implies $\mu(X) = 0$. With $P_X > 0$, this reduces to:
$$\frac{\partial F(X,p(X))}{\partial p} = -\lambda.$$

Using this, we define $p(X) = g_{-\lambda}(X)$, representing the incentive policy that achieves the target rate of decrease in $F(X, p(X))$.

Next, we analyze the behavior of $F(X, p(X))$ based on the convexity properties established earlier. The conditions $F'(X, p(X)) < 0$ and $F''(X, p(X)) > 0$ ensure that $F(X, p(X))$ is non-increasing and convex with respect to $p(X)$, thus any solution to the optimality condition
$$\frac{\partial F(X,p(X))}{\partial p} = -\lambda$$
corresponds to the global minimum of the functional.

To interpret the optimal solution, consider two cases based on the rate of variance reduction:

Case 1: $F'(X, p(X)) < -\lambda$
If the marginal rate of variance reduction is sufficiently fast (i.e., $F'(X, p(X)) < -\lambda$), budget is allocated until the rate of decrease slows to match $-\lambda$. The corresponding optimal policy is:
$$p(X) = g_{-\lambda}(X).$$

If the total budget is not fully utilized $\left(\int p(X)P_X \, dX < C\right)$, the KKT conditions imply $\lambda = 0$, indicating no resource scarcity. In this case, we set $e(X) = 0.5$, achieving the maximum reduction in variance.

Case 2: $F'(X, p(X)) \geq -\lambda$
If the marginal rate of variance reduction is already too slow (i.e., $F'(X, p(X)) > -\lambda$), no additional budget is allocated:

$$p(X) = 0.$$

.

In particular, when $\beta(X) \equiv \beta$ is a constant, the derivative for different feature X only depends on $\text{gap}(X)$, therefore, the optimal solution always prioritize the most imbalanced feature, which is illustrated in 2.

B.3 Proof of Theorem 3.3

The selection bias for treatment group $Y^{(1)}$ can be characterized as

$$\begin{aligned}\text{bias}^{(1)} &= \mathbb{E}_X[g^{(1)}(X, \alpha_1)|W = 1] - E_X[g^{(1)}(X, \alpha_1)] \\ &= \mathbb{E}_X[g^{(1)}(X, \alpha_1)|\alpha_1 + \text{gap}(X) \geq \alpha_0] - E_X[g^{(1)}(X, \alpha_1)].\end{aligned}$$

Definition. Let $\alpha^{(1)}, \alpha^{(0)} \overset{i.i.d.}{\sim} f_\alpha$, $\Delta = \alpha^{(1)} - \alpha^{(0)}$, and for any threshold k set

$$\mu(k) := \Pr(\Delta > k) > 0, \qquad H(x, k) := \frac{F_\alpha(x - k)}{\mu(k)}.$$

Then $x \mapsto H(x, k)$ is strictly increasing and $\mathbb{E}[H(\alpha^{(1)}, k)] = 1$.

Shift as a covariance. For any ε-Lipschitz g with $g(0) = 0$,

$$\Delta_g(k) := \mathbb{E}\big[g(\alpha^{(1)}) \mid \Delta > k\big] - \mathbb{E}\big[g(\alpha^{(1)})\big] = \int g(x)\big(H(x, k) - 1\big) f_\alpha(x)\, dx = \text{Cov}\big(g(\alpha^{(1)}), H(\alpha^{(1)}, k)\big). \tag{1}$$

Monotone-kernel (bang–bang) inequality. For increasing $x \mapsto H(x, k)$ with $\mathbb{E}[H(\alpha^{(1)}, k)] = 1$,

$$\big|\,\text{Cov}\big(g(\alpha^{(1)}), H(\alpha^{(1)}, k)\big)\big| \leq \varepsilon\ \text{Cov}\big(\alpha^{(1)}, H(\alpha^{(1)}, k)\big), \tag{2}$$

and equality is attained by the extremizer $g^*(x) = \varepsilon x$. Moreover,

$$\text{Cov}\big(\alpha^{(1)}, H(\alpha^{(1)}, k)\big) = \mathbb{E}\big[\alpha^{(1)} \mid \Delta > k\big] - \mathbb{E}[\alpha^{(1)}]. \tag{3}$$

Consequences (all in terms of H). From above equations, we have

$$\boxed{\big|\Delta_g(k)\big| \leq \varepsilon\ \text{Cov}\big(\alpha^{(1)}, H(\alpha^{(1)}, k)\big) = \varepsilon\Big(\mathbb{E}[\alpha^{(1)} \mid \alpha^{(1)} - \alpha^{(0)} > k] - \mathbb{E}[\alpha^{(1)}]\Big)\ ,}$$

and

$$\boxed{\sup_{\text{Lip}(g) \leq \varepsilon, \mathbb{E}[g]=0} \Delta_g(k) = \varepsilon\ \text{Cov}\big(\alpha^{(1)}, H(\alpha^{(1)}, k)\big).}$$

Similarly, we can prove that

$$\boxed{|\,\text{bias}^{(0)}\,| \leq \varepsilon\, \mathbb{E}_X[\text{Cov}\big(\alpha^{(0)}, H(\alpha^{(0)}, -\,\text{gap}(X))\big)]}.$$

So the max selection bias is achieved by setting $g^{(1)}(X, \alpha^{(1)}) = \varepsilon\alpha^{(1)}, g^{(0)}(X, \alpha^{(0)}) = -\varepsilon\alpha^{(0)}$.

B.4 Proof of Proposition 3.4

We first proof that the bias term

$$\varepsilon\left[\mathbb{E}_X[\text{Cov}\big(\alpha^{(0)}, H(\alpha^{(0)}, -\,\text{gap}(X))\big)] + \mathbb{E}_X[\text{Cov}\big(\alpha^{(1)}, H(\alpha^{(1)}, -\,\text{gap}(X))\big)]\right]$$

is convex in $\mathrm{gap}(X)$. Let $u_1, u_2 \overset{i.i.d.}{\sim} f_u$, $\Delta = u_1 - u_2$, and

$$C(k) := \mathrm{Cov}\big(u_1, H(u_1,k)\big) = \mathbb{E}[u_1 \mid \Delta > k] - \mathbb{E}[u_1], \qquad H(x,k) := \frac{F_u(x-k)}{\mu},\ \mu = \Pr(\Delta > k) > 0.$$

Thus convexity of $C(k)$ reduces to convexity of $k \mapsto \mathbb{E}[u_1 \mid \Delta > k]$.

Normal case. Assume $u_1, u_2 \sim N(0,1)$. Then $\Delta \sim N(0,2)$ and

$$\mathbb{E}[u_1 \mid \Delta = d] = \frac{\mathrm{Cov}(u_1, \Delta)}{\mathrm{Var}(\Delta)}\, d = \tfrac{1}{2} d, \quad \Longrightarrow \quad \mathbb{E}[u_1 \mid \Delta > k] = \tfrac{1}{2}\mathbb{E}[\Delta \mid \Delta > k].$$

With $\sigma_\Delta = \sqrt{2}$, $\alpha = k/\sigma_\Delta$, and the inverse Mills ratio $\lambda(\alpha) = \dfrac{\phi(\alpha)}{1-\Phi(\alpha)}$,

$$C(k) = \mathbb{E}[u_1 \mid \Delta > k] - \mathbb{E}[u_1] = \frac{1}{\sqrt{2}}\,\lambda\!\left(\frac{k}{\sqrt{2}}\right).$$

Since $\lambda'(\alpha) = \lambda(\alpha)(\lambda(\alpha) - \alpha)$ and $\lambda(\alpha) > \alpha$ for all α, we have $\lambda''(\alpha) > 0$, hence

$$C''(k) = \frac{1}{2\sqrt{2}}\,\lambda''\!\left(\frac{k}{\sqrt{2}}\right) \geq 0,$$

so $C(k)$ is convex in k.

Logit (Gumbel) case. Let $u_i = -\log E_i$ with $E_i \overset{i.i.d.}{\sim} \mathrm{Exp}(1)$, so $\Delta = \log(E_2/E_1)$ is Logistic$(0,1)$. Writing $R = E_2/E_1$,

$$\mathbb{E}[u_1 \mid R = r] = \log(1+r) - (1-\gamma), \qquad \mathbb{E}[u_1 \mid \Delta = d] = \mathrm{softplus}(d) - (1-\gamma).$$

Hence

$$\mathbb{E}[u_1 \mid \Delta > k] = \log(1+e^k) + \gamma, \qquad \mathbb{E}[u_1] = \gamma,$$

and therefore

$$C(k) = \log(1+e^k) = \mathrm{softplus}(k), \quad C''(k) = \sigma(k)\big(1-\sigma(k)\big) > 0, \quad \sigma(k) = \frac{1}{1+e^{-k}}.$$

Thus $C(k)$ is strictly convex in k for the logit case as well.

Since the bias term $C(k) + C(-k)$ is a convex function and also increasing with k, we know that the selection bias reduces as the utility gap k goes to 0. Also, following the same argument as we derive in theorem 3.2, the optimal solution follows an equal derivative properties. Moreover, if $\beta(X) \equiv \beta$ is a constant, then the derivative only depends on the utility gap, so it's optimal to always incentivize the most imbalanced feature, which coinsides with the case of variance minimization.

B.5 Proof of Lemma 4.1

Goal. Fix a batch k of size n_k. Let $p^\star(\cdot)$ be the optimal incentive policy for the population problem in Section 3 (Equal-Derivative characterization), and let $\widehat{p}(\cdot)$ be the plug-in policy obtained by replacing $\mathrm{gap}(\cdot)$ with $\widehat{\mathrm{gap}}(\cdot)$ in the same characterization and choosing the threshold $\widehat{\lambda}$ to satisfy the budget constraint. We prove

$$\mathbb{E}\big\|\widehat{p} - p^\star\big\|^2_{L^2(P_X)} \lesssim n_k^{-\alpha_1}.$$

Ingredients. (i) **Oracle rate for the utility gap.** By Assumption A.3 (logistic regression oracle), for the batch-k sample of size n_k,

$$\mathbb{E}\big[(\widehat{\mathrm{gap}}(X) - \mathrm{gap}(X))^2\big] \leq C_{\mathrm{gap}}\, n_k^{-\alpha_1} \quad \text{(up to a } \log n \text{ factor absorbed into the constant).}$$

(ii) **Monotone-Kernel convexity and Equal-Derivative rule.** For fixed X, writing $e(X,p) = \sigma(\mathrm{gap}(X) - \beta(X)p)$ and $F(X,p) = \sigma^2/\big(e(1-e)\big)$, Lemma 3.1 shows F is convex in p; Theorem 3.2 gives the *Equal-Derivative* optimality: there exists a unique $\lambda^\star \geq 0$ such that

$$\partial_p F\big(X, p^\star(X)\big) = -\lambda^\star \quad \text{whenever } p^\star(X) > 0, \quad \text{and } p^\star(X) = 0 \text{ otherwise},$$

and the budget binds when $\lambda^\star > 0$. Both statements continue to hold for the plug-in $\widehat{p}$ with $\widehat{\lambda}$. (iii) **Overlap.** $e(X) \in [\eta, 1-\eta]$ for some $\eta \in (0, \frac{1}{2})$ (Section 2.2), ensuring uniform curvature and Lipschitz constants in what follows.

Step 1: Pointwise stability of the optimality equation. Fix X and abbreviate $g = \mathrm{gap}(X)$, $\hat{g} = \widehat{\mathrm{gap}}(X)$, $\beta = \beta(X)$. Define

$$\Psi(g,\lambda) \;:=\; \operatorname*{arg\,min}_{p \geq 0} \Big\{ F(g,p) \;+\; \lambda\, p \Big\}, \qquad F(g,p) := \frac{\sigma^2}{e(g,p)\big(1-e(g,p)\big)},$$

with $e(g,p) = \sigma(g - \beta p)$. By convexity of F in p and KKT, $\Psi(g,\lambda)$ is characterized (when positive) by the stationarity equation $\partial_p F(g,p) = -\lambda$ and complementary slackness at 0.

A direct differentiation using $e'_p(g,p) = -\beta\,\sigma(g-\beta p)\big(1-\sigma(g-\beta p)\big)$ and the product/quotient rule yields

$$\partial^2_{pp} F(g,p) \;=\; \underbrace{\frac{c_1\,\beta^2}{\big(e(1-e)\big)^3}}_{\geq c_0 > 0 \text{ under } e \in [\eta, 1-\eta]}, \qquad \big|\partial^2_{pg} F(g,p)\big| \;\leq\; \frac{C_1\,|\beta|}{\big(e(1-e)\big)^3} \;\leq\; C_2,$$

for universal constants c_0, C_1, C_2 depending only on $(\eta, \sigma^2, \|\beta\|_\infty)$. Hence by the (one-dimensional) Implicit Function Theorem applied to $\Phi(g,p,\lambda) := \partial_p F(g,p) + \lambda = 0$, the solution map $(g,\lambda) \mapsto \Psi(g,\lambda)$ is *locally Lipschitz* with

$$\Big|\partial_g \Psi(g,\lambda)\Big| \;=\; \frac{\big|\partial^2_{pg} F(g,p)\big|}{\partial^2_{pp} F(g,p)} \;\leq\; \frac{C_2}{c_0} \;:=\; L_g, \qquad \Big|\partial_\lambda \Psi(g,\lambda)\Big| \;=\; \frac{1}{\partial^2_{pp} F(g,p)} \;\leq\; \frac{1}{c_0} \;:=\; L_\lambda.$$

The same bounds hold globally once we note the overlap keeps e in a compact subinterval of $(0,1)$, so the constants do not blow up.

Step 2: Budget equation and stability of the dual variable. Let $G(\lambda; g) := \mathbb{E}[\Psi(g(X), \lambda)]$ denote expected spending at threshold λ. By envelope/IFT and the positivity of $\partial^2_{pp} F$,

$$\frac{\partial}{\partial \lambda} G(\lambda; g) = \mathbb{E}[\partial_\lambda \Psi(g(X), \lambda)] \;\leq\; -\,\underline{c} \quad \text{for some } \underline{c} > 0,$$

i.e., G is strictly decreasing and Lipschitz in λ with a slope bounded away from 0. Let $\lambda^\star$ and $\widehat{\lambda}$ solve $G(\lambda^\star; g) = B/n$ and $G(\widehat{\lambda}; \hat{g}) = B/n$ (binding case). Then the mean value theorem and the Lipschitz-in-g bound for Ψ give

$$|\widehat{\lambda} - \lambda^\star| \;\leq\; \underline{c}^{-1}\,\big|G(\widehat{\lambda}; \hat{g}) - G(\lambda^\star; g)\big| \;\leq\; \underline{c}^{-1}\,\mathbb{E}\big|\Psi(\hat{g}(X), \lambda^\star) - \Psi(g(X), \lambda^\star)\big| \;\leq\; \underline{c}^{-1} L_g\,\mathbb{E}\big|\hat{g}(X) - g(X)\big|.$$

By Cauchy–Schwarz, $\mathbb{E}|\hat{g} - g| \leq \sqrt{\mathbb{E}(\hat{g}-g)^2} \;\lesssim\; n_k^{-\alpha_1/2}$.

Step 3: Assemble the plug-in error bound. Using the global Lipschitz continuity of Ψ in both arguments and the projection $[\cdot]_+$ at the boundary,

$$\big|\widehat{p}(X) - p^\star(X)\big| = \big|\Psi(\hat{g}(X), \widehat{\lambda}) - \Psi(g(X), \lambda^\star)\big| \;\leq\; L_g\,|\hat{g}(X) - g(X)| \;+\; L_\lambda\,|\widehat{\lambda} - \lambda^\star|.$$

Square, take expectations, and apply $(a+b)^2 \leq 2a^2 + 2b^2$ together with the bound on $|\widehat{\lambda} - \lambda^\star|$:

$$\mathbb{E}\big(\widehat{p}(X) - p^\star(X)\big)^2 \;\leq\; 2L_g^2\,\mathbb{E}(\hat{g} - g)^2 \;+\; 2L_\lambda^2\,\underline{c}^{-2} L_g^2\,\big(\mathbb{E}|\hat{g} - g|\big)^2 \;\lesssim\; \mathbb{E}(\hat{g} - g)^2.$$

Finally integrate over P_X (the expectation above is already w.r.t. $X \sim P_X$) to obtain

$$\mathbb{E}\big\|\widehat{p} - p^\star\big\|^2_{L^2(P_X)} \;\lesssim\; \mathbb{E}(\hat{g} - g)^2 \;\lesssim\; n_k^{-\alpha_1},$$

where the last step invokes Assumption A.3. This completes the proof.

B.6 Proof of Proposition 4.2

Goal. Let batch k contain the index set $\mathcal{I}_k$ with $|\mathcal{I}_k| = n_k$. At time $t \in \mathcal{I}_k$, the (plug-in) policy chooses an offer with probability $\widehat{p}_t := \widehat{p}(X_t) \in [0,1]$ and the realized spending indicator is $Z_t \sim \text{Bernoulli}(\widehat{p}_t)$ (conditionally independent given $\{X_t\}$). Define the realized and expected batch expenditures

$$\widehat{S}_k := \sum_{t \in \mathcal{I}_k} Z_t, \qquad \bar{S}_k := \mathbb{E}[\widehat{S}_k \mid \{X_t\}_{t \in \mathcal{I}_k}] = \sum_{t \in \mathcal{I}_k} \widehat{p}_t.$$

We prove a high–probability concentration bound

$$\left|\widehat{S}_k - \bar{S}_k\right| \;\le\; C\Big(\sqrt{n_k \log n} \;+\; \log n\Big) \quad \text{for all } k,$$

and deduce that if the *first* batch holds a reserve of order $c\sqrt{N}$ with $c = \mathcal{O}(\log n)$ (here $N = \sum_k n_k$), then with high probability the cumulative spending will not exceed the budget over the entire run.

Step 1: Batchwise concentration (Bernstein/Freedman). Condition on $\{X_t\}_{t \in \mathcal{I}_k}$ and write

$$V_k \;:=\; \mathrm{Var}\big(\widehat{S}_k \mid \{X_t\}\big) = \sum_{t \in \mathcal{I}_k} \widehat{p}_t(1 - \widehat{p}_t) \;\le\; \sum_{t \in \mathcal{I}_k} \widehat{p}_t \;\le\; n_k.$$

Since $Z_t - \widehat{p}_t \in [-1, 1]$ are conditionally independent and mean-zero, Bernstein's inequality yields, for any $x > 0$,

$$\Pr\left(|\widehat{S}_k - \bar{S}_k| \;\ge\; \sqrt{2V_k x} \;+\; \frac{2x}{3} \;\middle|\; \{X_t\} \right) \;\le\; 2e^{-x}.$$

Choosing $x = \log n$ and using $V_k \le n_k$,

$$\Pr\left(|\widehat{S}_k - \bar{S}_k| \;\le\; C_1\sqrt{n_k \log n} + C_2 \log n \;\text{ for all } k \;\middle|\; \{X_t\} \right) \;\ge\; 1 - 2\sum_k n^{-1}.$$

Unconditioning and absorbing the $\sum_k$ into the constant (or imposing a mild $K \le n$), there exists $C > 0$ s.t.

$$\Pr\Big(|\widehat{S}_k - \bar{S}_k| \;\le\; C\big(\sqrt{n_k \log n} + \log n\big) \;\text{ for all } k \Big) \;\ge\; 1 - \mathcal{O}(n^{-1}). \tag{7}$$

(The same bound follows from Freedman's inequality for martingales if one prefers a conditional martingale difference formulation.)

Step 2: From batchwise to cumulative control. Let K be the number of batches and $N = \sum_{k=1}^{K} n_k$. Summing equation 7 and using Cauchy–Schwarz,

$$\sum_{k=1}^{K} \left|\widehat{S}_k - \bar{S}_k\right| \;\le\; C\Big(\sum_{k=1}^{K} \sqrt{n_k \log n} \;+\; K \log n\Big) \;\le\; C\Big(\sqrt{K}\,\sqrt{\textstyle\sum_k n_k}\,\sqrt{\log n} \;+\; K \log n\Big).$$

Hence, with probability at least $1 - \mathcal{O}(n^{-1})$,

$$\sum_{k=1}^{K} \left|\widehat{S}_k - \bar{S}_k\right| \;\le\; C\Big(\sqrt{KN \log n} \;+\; K \log n\Big). \tag{8}$$

Step 3: A sufficient reserve to avoid overspend. Let the total budget over the horizon be B, and suppose the planned (expected) spending per batch meets the budget exactly, i.e. $\sum_k \bar{S}_k \le B$ (binding in expectation). Allocate an initial *reserve* R_1 during the first batch (by slightly under-spending in expectation there), and run the remaining batches at their nominal planned expectations. Then by equation 8, a sufficient condition to avoid overspending is

$$R_1 \;\ge\; C\Big(\sqrt{KN \log n} \;+\; K \log n\Big).$$

In particular, since $K \le N$ and typically $N \gg \log n$, it is enough to take

$$R_1 \;=\; c\sqrt{N} \qquad \text{with} \qquad c \;=\; \mathcal{O}(\log n),$$

which dominates both terms on the right-hand side of equation 8 for moderate K (e.g. $K = \mathcal{O}(\log n)$ or any subpolynomial K). Therefore, by making the *first* batch sufficiently long and holding back a reserve of size $c\sqrt{N}$ with $c = \mathcal{O}(\log n)$, we obtain that the realized cumulative spending does not exceed the total budget with high probability.

B.7 Proof of Theorem 4.3

We define the following three estimators. We will use the AIPW estimator to estimate the treatment effect, which is defined as

$$\widehat{\tau}_1^X = \sum_{i=1}^{n} \frac{1}{n}\left(\widehat{\mu}^{(1)}(X_i) - \widehat{\mu}^{(0)}(X_i) + \frac{Y_i - \widehat{\mu}^{(1)}(X_i)}{\hat{e}_i(X_i, p(X_i))}W_i - \frac{Y_i - \widehat{\mu}^{(0)}(X_i)}{1 - \hat{e}_i(X_i, p(X_i))}(1 - W_i)\right), \tag{9}$$

where we use the superscript X to emphasize the existence of covariates, and $\widehat{\mu}^{(1)}(X)$, $\widehat{\mu}^{(0)}(X)$ as the estimation for outcome function of treatment and control $\mu^{(1)}(X)$, $\mu^{(0)}(X)$. Similarly, we can define the intermediary and optimal estimator as

$$\widehat{\tau}_2^X = \sum_{i=1}^{n} \frac{1}{n}\left(\mu^{(1)}(X_i) - \mu^{(0)}(X_i) + \frac{Y_i - \mu^{(1)}(X_i)}{e_i(X_i, p(X_i)}W_i - \frac{Y_i - \mu^{(0)}(X_i)}{1 - e_i(X_i, p(X_i))}(1 - W_i)\right), \tag{10}$$

$$\widehat{\tau}^{*X} = \sum_{i=1}^{n} \frac{1}{n}\left(\mu^{(1)}(X_i) - \mu^{(0)}(X_i) + \frac{Y_i - \mu^{(1)}(X_i)}{e^*(X_i, p^*(X_i))}W_i - \frac{Y_i - \mu^{(0)}(X_i)}{1 - e^*(X_i, p^*(X_i))}(1 - W_i)\right), \tag{11}$$

where the optimal propensity score $e^*(X) = \sigma^{(1)}(X)/(\sigma^{(1)}(X) + \sigma^{(0)}(X))$ and the estimator $\widehat{\tau}^{*X}$ achieves optimal variance v^* using the optimal incentive mechanism p^*. Throughout the proof, we will condition on the proved high probability event that $p(X_i)$ in the learning-to-incentivize algorithm will never run out of budget. We can divide the mean square error of the AIPW estimator into three parts: model estimation, propensity score optimization, and cross term:

$$\begin{aligned}
\mathbb{E}\left(\widehat{\tau}_1^X - \tau\right)^2 - \mathbb{E}\left(\widehat{\tau}^{*X} - \tau\right)^2 &= \mathbb{E}\left(\widehat{\tau}_1^X - \widehat{\tau}_2^X + \widehat{\tau}_2^X - \tau\right)^2 - \mathbb{E}\left(\widehat{\tau}^{*X} - \tau\right)^2 \\
&= \underbrace{\mathbb{E}\left(\widehat{\tau}_1^X - \widehat{\tau}_2^X\right)^2}_{\text{model estimation}} + \underbrace{\mathbb{E}\left(\widehat{\tau}_2^X - \tau\right)^2 - \mathbb{E}\left(\widehat{\tau}^{*X} - \tau\right)^2}_{\text{propensity score optimization}} \\
&\quad + \underbrace{2\mathbb{E}\left((\widehat{\tau}_1^X - \widehat{\tau}_2^X)(\widehat{\tau}_2^X - \tau)\right)}_{\text{cross-term}}.
\end{aligned} \tag{12}$$

We will prove the following two lemmas, which will then lead to the desired result.

Lemma B.1.

$$\mathbb{E}\left(\widehat{\tau}_2^X - \tau\right)^2 - \mathbb{E}\left(\widehat{\tau}^{*X} - \tau\right)^2 \le O(\frac{1}{n^{\alpha_1/2}})\mathbb{E}\left(\widehat{\tau}^{*X} - \tau\right)^2 \tag{13}$$

Proof of Lemma B.1

We have the following decomposition

$$\mathbb{E}\left(\widehat{\tau}_2^X - \tau\right)^2 - \mathbb{E}\left(\widehat{\tau}^{*X} - \tau\right)^2 = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\frac{\sigma^2}{e_i(X_i)} + \sum_{i=1}^{n}\frac{\sigma^2}{1 - e_i(X_i)} - n\left(\frac{\sigma^2}{e^*(X_i)} + \frac{\sigma^2}{1 - e^*(X_i)}\right)\right). \tag{14}$$

As proved in lemma 4.1, we have in batch k, $\mathbb{E}[(e_i(X_i, p(X_i)) - e^*(X_i, p^*(X_i)))^2] \le O(n_{k-1}^{-\alpha_1})$. Therefore, the

total regret in (14) can be bounded as

$$\begin{aligned}
&\frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\frac{\sigma^2}{e_i(X_i)}+\sum_{i=1}^{n}\frac{\sigma^2}{1-e_i(X_i)}-n\left(\frac{\sigma^2}{e^*(X_i)}+\frac{\sigma^2}{1-e^*(X_i)}\right)\right)\\
\leq&\frac{1}{n^2}\mathbb{E}\left(\sum_{k=1}^{O(\log(n))}O\left(\frac{\sigma^2}{e^*(X_i,p^*(X_i))(1-e^*(X_i,p^*(X_i)))}\right)n_k\cdot n_{k-1}^{-\alpha_1/2}\right)\\
\leq&\widetilde{O}(\frac{1}{n^{\alpha_1/2}})\mathbb{E}\left(\widehat{\tau}^{*X}-\tau\right)^2.
\end{aligned}\tag{15}$$

As we have $\sum_{k=1}^{O(\log n)} n_k\frac{1}{\sqrt{n_{k-1}}}\leq O(\log n)\sqrt{n}$. □

We then have the following lemma for bounding the model estimation error term.

Lemma B.2. $\mathbb{E}\left(\tau_1^X-\tau_2^X\right)^2\leq\widetilde{O}(n^{-(1+\alpha)})\leq\widetilde{O}(n^{-\alpha})v^*$.

Proof of Lemma B.2

We will use data splitting and cross fitting to reduce the correlations between data. In particular, cross-fitting first splits the data (at random) into two halves $\mathcal{I}_1$ and $\mathcal{I}_2$, and within each half, we are running an independent low swithcing learning-to-incentivize algorithm (1), then uses an estimator

$$\begin{aligned}
\hat{\tau}_1^X=\frac{|\mathcal{I}_1|}{n}\hat{\tau}_1^{\mathcal{I}_1,X}+\frac{|\mathcal{I}_2|}{n}\hat{\tau}_1^{\mathcal{I}_2,X},\quad\hat{\tau}^{\mathcal{I}_1}=\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\Big(&\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\hat{\mu}_{(0)}^{\mathcal{I}_2}(X_i)\\
&+W_i\frac{Y_i-\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}-(1-W_i)\frac{Y_i-\hat{\mu}_{(0)}^{\mathcal{I}_2}(X_i)}{1-\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}\Big),
\end{aligned}\tag{16}$$

where in batch k in $\mathcal{I}_1$, the $\hat{\mu}_{(w)}^{\mathcal{I}_2}(\cdot)$ and $\hat{e}^{\mathcal{I}_2}(\cdot)$ are estimates of $\mu_{(w)}(\cdot)$ and $e(\cdot)$ obtained using only the half-sample $\mathcal{I}_2$ in batch $k-1$, and $\hat{\tau}^{\mathcal{I}_2}$ is defined analogously (with the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ swapped). In other words, $\hat{\tau}^{\mathcal{I}_1}$ is a treatment effect estimator on $\mathcal{I}_1$ that uses $\mathcal{I}_2$ to estimate its nuisance components, and vice-versa.

To do so, we first note that we can write

$$\hat{\tau}_2^X=\frac{|\mathcal{I}_1|}{n}\hat{\tau}_2^{\mathcal{I}_1,X}+\frac{|\mathcal{I}_2|}{n}\hat{\tau}_2^{\mathcal{I}_2,X}\tag{17}$$

analogously to (16) (because $\hat{\tau}_2^X$ uses oracle nuisance components, the crossfitting construction doesn't change anything for it). Moreover, we can decompose $\hat{\tau}^{\mathcal{I}_1}$ itself as

$$\begin{aligned}
\hat{\tau}_1^{\mathcal{I}_1,X}&=\hat{Y}_{(1)}^{\mathcal{I}_1,1}-\hat{Y}_{(0)}^{\mathcal{I}_1}\\
\hat{Y}_{(1)}^{\mathcal{I}_1,1}&=\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)+W_i\frac{Y_i-\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}\right)
\end{aligned}$$

etc., and define $\hat{Y}_{(0)}^{\mathcal{I}_1,2}$ and $\hat{Y}_{(1)}^{\mathcal{I}_1,2}$ analogously. Given this buildup, in order to verify lemma (B.2), it suffices to show that

$$\mathbb{E}\left(\hat{Y}_{(1)}^{\mathcal{I}_1,1}-\hat{Y}_{(1)}^{\mathcal{I}_1,2}\right)^2\leq\widetilde{O}(n^{-(1+\alpha)}).\tag{18}$$

etc., across folds and treatment statuses.

We now study the term in (18) by decomposing it as follows:

$$
\begin{aligned}
&\hat{Y}_{(1)}^{\mathcal{I}_1,1} - \hat{Y}_{(1)}^{\mathcal{I}_{1,2}} \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))} - \mu_{(1)}(X_i) - W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i, p(X_i))}\right) \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i, p(X_i))}\right)\right) \\
&\quad + \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} W_i\left(\left(Y_i - \mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))} - \frac{1}{e(X_i, p(X_i))}\right)\right) \\
&\quad - \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} W_i\left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))} - \frac{1}{e(X_i, p(X_i))}\right)\right).
\end{aligned}
$$

Now, we can verify that these are small for different reasons. For the first term, we intricately use the fact that, thanks to our double machine learning construction, $\hat{\mu}_{(w)}^{\mathcal{I}_2}$ can effectively be treated as deterministic. And we will abbreviate $e(X_i, p(X_i))$ as $e(X_i)$ for simplicity. Thus after conditioning on $\mathcal{I}_2$, the summands used to build this term become mean-zero and independent (2nd and 3rd equalities below).

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right)^2 \middle| \mathcal{I}_2\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{1}{|\mathcal{I}_1|^2}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)^2\left(1 - \frac{W_i}{e(X_i)}\right)^2\right.\right. \\
&\qquad \left.\left. + \sum_{i,j\in\mathcal{I}_1}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_j) - \mu_{(1)}(X_j)\right)\left(1 - \frac{W_j}{e(X_j)}\right)|\,\mathcal{I}_2\right]\right] \qquad (19) \\
&= \frac{1}{|\mathcal{I}_1|^2}\mathbb{E}\left[\sum_{i\in\mathcal{I}_1}\mathrm{Var}\left[\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\middle|\,\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|^2}\mathbb{E}\left[\sum_{i\in\mathcal{I}_1}\mathbb{E}\left[\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)^2\left(\frac{1}{e(X_i)} - 1\right)\middle|\,\mathcal{I}_2\right]\right] \\
&\leq \frac{1}{\eta|\mathcal{I}_1|^2}\mathbb{E}\left[\sum_{i\in I_1}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)^2\right] = \frac{\tilde{O}(1)}{n^{1+\alpha}},
\end{aligned}
$$

where the third inequality holds as all the cross term has expectation 0 for $i \neq j$. and for each term $\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)$ has mean 0. And the last equality holds since with high probability we can argue that in batch k, we have $\mathbb{E}\left[\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)^2\right] \leq O(n_{k-1}^{-\alpha})$ for all batches k and data i in batch k from $t_{k-1}+1$ to t_k. Therefore, we have

$$
\mathbb{E}\left[\sum_{i\in I_1}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i)\right)^2\right] \leq \sum_{k=1}^{O(\log n)} n_k O(n_{k-1}^{-\alpha}) \leq \widetilde{O}(n^{1-\alpha}). \qquad (20)
$$

Similarly, we can prove that the second term

$$
\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} W_i\left(\left(Y_i - \mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))} - \frac{1}{e(X_i, p(X_i))}\right)\right)\right)^2\right] \leq \frac{\widetilde{O}(1)}{n^2}. \qquad (21)
$$

Now for the third term, we have

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}W_i\left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}-\frac{1}{e(X_i,p(X_i))}\right)\right)\right)^2\right]\\&\leq\frac{O(\log n)}{|\mathcal{I}_1|^2}\sum_{k=1}^{O(\log n)}\mathbb{E}\left(\sum_{i\in\mathcal{I}_1,\text{ batch k}}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}-\frac{1}{e(X_i,p(X_i))}\right)\right)^2.\end{aligned}\tag{22}$$

Now within batch k, we know by Cauchy-Schwarz that

$$\begin{aligned}&\mathbb{E}\left[\sum_{\{i:i\in\mathcal{I}_1,\text{ batch k}\}}\left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)\right)\right]^2\\&\leq\mathbb{E}\left[\sum_{\{i:i\in\mathcal{I}_1,\text{ batch k}\}}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\mu_{(1)}(X_i)\right)^2\right]\\&\quad\times\mathbb{E}\left[\sum_{\{i:i\in\mathcal{I}_1,\text{ batch k}\}}\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)^2\right]=\widetilde{O}_P\left(n_k^{1-\alpha}\right).\end{aligned}\tag{23}$$

Therefore we know that

$$\frac{O(\log n)}{|\mathcal{I}_1|^2}\sum_{k=1}^{O(\log n)}\mathbb{E}\left(\sum_{i\in\mathcal{I}_1,\text{ batch k}}\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)-\mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}-\frac{1}{e(X_i,p(X_i))}\right)\right)^2\leq\frac{O(\text{polylog}(n))}{n^{1+\alpha}}.\tag{24}$$

Combining everything together, we have the conclusion in (18). □

Finally, since we prove that $\mathbb{E}\left(\widehat{\tau}_2^X-\tau\right)^2\leq(1+O(\frac{1}{n^{-\alpha_1/2}}))v^*$, and $\mathbb{E}\left(\widehat{\tau}_1^X-\widehat{\tau}_2^X\right)^2\leq O(n^{-\alpha})v^*$, by Cauchy-Schwarz, we can bound the cross term as

$$\mathbb{E}\left((\widehat{\tau}_1^X-\widehat{\tau}_2^X)(\widehat{\tau}_2^X-\tau)\right)\leq O(n^{-\min(\alpha_1/4,\alpha/2)})v^*.$$

And this completes the proof of theorem 4.3.

B.8 Proof of Theorem 4.4

Using martinagle central limit theorem and checking the necessary condition in [Cook et al., 2024], the estimator satisfies:

$$\sqrt{n}(\hat{\tau}-\tau)\Rightarrow N(0,V^*),$$

where V^* is the minimum variance achievable. The results follow directly by considering both bias and variance term.

C Numerical Results

In this section, we conduct a comprehensive suite of simulation experiments: Section C.1 presents results under a no-confounder design with three nonlinear outcome models (polynomial, sigmoidal, and sinusoidal), comparing the naive difference-in-means (DIM), inverse-propensity-weighting (IPW) estimators, and doubly robust AIPW in terms of ATE bias and variance; Section C.3 then evaluates estimator performance under unobserved confounders and deliberate choice-model misspecification using the same outcome specifications.

C.1 Without Unobserved Confounder

We begin by evaluating our incentivized experiment through a series of simulation studies. Specifically, we examine two settings—one with unobserved confounders and one without—and impose a deliberately imbalanced

choice probability to reflect the primary focus of this paper. Throughout, the underlying utility-based choice model remains fixed: each feature vector $X_i \in \{0,1\}^4$ has

$$X_{i1}, X_{i3} \sim \text{Bernoulli}(0.8), \quad X_{i2}, X_{i4} \sim \text{Bernoulli}(0.2).$$

Without loss of generality, we let the latent utilities for treatments 0 and 1 be

$$U_{i0} = \varepsilon_{i0}, \quad U_{i1} = X_i^\top \theta^* + \varepsilon_{i1},$$

where $\varepsilon_{i0}, \varepsilon_{i1}$ are i.i.d. standard Gumbel noise and $\theta^* = (3, -3, 3, -3)^\top$. Consequently, the probability of assignment to treatment 1 is

$$\mathbb{P}(W_i = 1 \mid X_i) = \frac{\exp\big(X_i^\top \theta^*\big)}{1 + \exp\big(X_i^\top \theta^*\big)}.$$

C.2 Without Unobserved Confounders

When incentive gaps are unbalanced, even in the simplest no-confounder setting, propensity-score–based estimators (e.g. IPW) become highly unstable and can underperform the Difference-in-Means (DIM) estimator. We now introduce three distinct outcome definitions:

1. Polynomial:
$$Y_0 = (X^T\gamma_0)^2 + X^T\theta^* + \eta_0, \qquad Y_1 = (X^T\gamma_1)^2 + X^T\theta^* + \eta_1,$$
2. Sigmoid:
$$Y_0 = \sigma(X^T\theta^*) + \sigma(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sigma(X^T\theta^*) + \sigma(X^T\gamma_1) + \eta_1,$$
3. Sine:
$$Y_0 = \sin(X^T\theta^*) + \sin(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sin(X^T\theta^*) + \sin(X^T\gamma_1) + \eta_1,$$

where $\sigma(X) = \frac{1}{1+e^{-x}}$ is the sigmoid function, and η_1, η_0 are i.i.d. standard Gaussian noise. γ_1, γ_0 are uniformly generated from $[0,1]^4$.

We generate 20 independent pairs (γ_0, γ_1), each drawn uniformly from $[0,1]^4$. For each pair, we simulate an experiment with $n = 1000$ units, drawing treatments and outcomes as described above. We then repeat the entire sampling and estimation procedure 100 times per (γ_0, γ_1) to obtain empirical estimates of bias and variance for each estimator. Table 4 reports the aggregated results across all parameter realizations.

In settings with large class imbalance, penalized logistic regression introduces substantial shrinkage bias, which can destabilize estimates of the underlying choice model. This reflects a classic bias–variance trade-off: although IPW estimators are unbiased in theory when all confounders are correctly specified, their high variability can lead them to underperform even a simple DIM estimator in practice. Augumented-IPW (AIPW), or "doubly robust," estimators mitigate this volatility by combining outcome regression with propensity-score weighting, yielding more reliable estimates under model misspecification. Nonetheless, when the true outcomes exhibit strong nonlinearity (e.g., polynomial relationships), AIPW still incurs nontrivial bias in estimating Y_0, and—across all three simulation scenarios—its variance remains markedly larger than that of the DIM estimator.

In the next phase, we deploy our low-switching incentivization policy to achieve substantially more accurate ATE estimates. We implement a simple two-stage design with a single policy switch. As established in our theoretical analysis, the optimal policy is characterized by a threshold λ: given estimates $\hat{p}(X)$ of the propensity score and $\hat{\theta}$ of the utility parameters, we assign a bonus $s^*(X)$ such that

$$\begin{cases} \text{Give bonus } s^*(X) \text{ to treatment 1, such that } \frac{e^{X^T\hat{\theta}}+s^*(X)}{1+e^{X^T\hat{\theta}}+s^*(X)} = \lambda, & \text{if } \dfrac{e^{X^T\hat{\theta}}}{1+e^{X^T\hat{\theta}}} \le \lambda, \\ \text{Give bonus } s^*(X) \text{ to treatment 1, such that } \frac{e^{X^T\hat{\theta}}}{1+s^*(X)+e^{X^T\hat{\theta}}} = 1-\lambda, & \text{if } \dfrac{e^{X^T\hat{\theta}}}{1+e^{X^T\hat{\theta}}} \ge 1-\lambda, \\ s^*(X) = 0, & \text{otherwise.} \end{cases}$$

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	ATE variance
Polynomial	DIM	4.044098	0.485364	-3.558735	0.051760
	IPW	2.063882	0.048878	-2.015004	1.217703
	AIPW	0.345920	0.039557	-0.306363	0.166380
Sigmoid	DIM	0.449822	0.054365	-0.395457	0.007994
	IPW	0.544242	-0.012095	-0.556337	0.112538
	AIPW	-0.020972	0.003569	0.024541	0.098769
Sine	DIM	-0.013673	0.007796	0.005877	0.008633
	IPW	0.187496	-0.011561	-0.199057	0.042525
	AIPW	-0.079291	0.001283	0.080574	0.109859

Table 4: Comparison of three estimators under three outcome models. Red text highlights, for each outcome, the smallest absolute bias in ATE, Y_1, and Y_0, and the smallest ATE variance.

In our simulations, we fix the threshold $\lambda = 0.4$ (other choices yield qualitatively similar results) and draw a total of $n = 1000$ units. We allocate $n_1 = 300$ to the first stage: after observing these 300 samples, we fit the choice model via penalized logistic regression and use the fitted parameters to construct the incentive function $s^*(X)$. Because penalized logistic regression tends to understate the true assignment imbalance, the post-incentive allocation will not be perfectly balanced at λ, but it still suffices to dramatically reduce selection bias. As Table 5 shows (with the best results highlighted in red), our incentivized design enables the AIPW estimator to achieve substantially lower bias—and markedly smaller variance—than in the no-incentive setting, across all three outcome models. From Table 5, we observe that the naive DIM estimator continues to exhibit nontrivial selection bias even under incentivization, although its bias is substantially reduced compared to the no-incentive case. More importantly, all three estimators achieve uniformly better performance once incentives are introduced. This improvement stems from our low-switching incentivization policy to correct the original treatment-assignment imbalance, and we can stabilize the propensity-score weights and shrink both bias and variance across all estimators.

Outcome	Method	ATE bias	Y1 bias	Y0 bias	Variance
Polynomial	DIM	2.083975	0.505646	-1.578329	0.105427
	IPW	1.295498	0.063204	-1.232294	0.084574
	AIPW	0.023256	0.021915	-0.001342	0.013452
Sigmoid	DIM	0.232880	0.052415	-0.180465	0.006985
	IPW	0.327998	-0.006939	-0.334937	0.010912
	AIPW	0.003361	0.001430	-0.001930	0.012843
Sine	DIM	-0.057995	0.004677	0.062673	0.006599
	IPW	0.086523	-0.010082	-0.096605	0.007499
	AIPW	0.001363	-0.000660	-0.002023	0.013045

Table 5: Comparison of three estimators under three outcome models. Red text highlights, for each outcome, the smallest absolute bias in ATE, Y_1, and Y_0, and the smallest ATE variance.

Comparing Tables 4 and 5 reveals that our two-stage incentivization policy substantially reduces both bias and variance across all estimators, with the largest gains for IPW and AIPW. In the polynomial outcome, AIPW's ATE bias falls from 0.3459 to 0.0233 and its variance from 0.1664 to 0.0135; IPW's bias decreases from 2.0639 to 1.2955 and its variance from 1.2177 to 0.0846; by contrast, DIM's bias is halved (from 4.0441 to 2.0840) while its variance increases only modestly. Similar improvements occur for the sigmoid model and for the sine model. These comparisons confirm that aligning treatment-assignment incentives to improve covariate distribution sharply enhances the stability and accuracy of propensity-based estimators, driving AIPW biases to near zero with variances comparable to or below those of DIM.

C.3 Robustness to Model Mis-specification and Unobserved Confounders

In the previous section, we studied estimator behavior when all confounders are observed. To assess performance under hidden bias, we now introduce an unobserved confounder. Consider the same choice model, for each covariate we let

$$U_{i0} = \varepsilon_{i0}, \quad U_{i1} = X_i^\top \theta^* + \varepsilon_{i1},$$

with $\varepsilon_{i0}, \varepsilon_{i1} \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0,1)$, and assign

$$W_i = \mathbf{1}\{U_{i1} > U_{i0}\}.$$

We then generate outcomes using the same three functional forms as before—continuous, sigmoidal, and sinusoidal:

1. Continuous:
$$Y_0 = U_0 + (X^T\gamma_0)^2 + X^T\gamma_0 + \eta_0, \qquad Y_1 = U_1 + (X^T\gamma_1)^2 + X^T\gamma_1 + \eta_1,$$
2. Sigmoid:
$$Y_0 = \sigma(U_0) + \sigma(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sigma(U_1) + \sigma(X^T\gamma_1) + \eta_1,$$
3. Sine:
$$Y_0 = \sin(U_0) + \sin(X^T\gamma_0) + \eta_0, \qquad Y_1 = \sin(U_1) + \sin(X^T\gamma_1) + \eta_1,$$

Here, $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the logistic function, and $\eta_0, \eta_1 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are Gaussian noise terms. The vectors γ_0, γ_1 are drawn uniformly from $[0,1]^4$.

We first repeat the no-incentive experiments under this confounded design, employing the same estimators as in the previous section. Table 6 reports the results. In all three outcome scenarios, the simple DIM estimator attains the lowest variance and the smallest ATE bias, outperforming both IPW and AIPW. This result reflects that when confounders drive imbalanced assignment probabilities, propensity-score weights become highly variable, inflating the variance while the DIM estimator avoids weight instability.

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	ATE variance
Continuous	DIM	-0.645	0.681	1.326	0.061
	IPW	-1.240	0.177	-1.417	1.633
	AIPW	-3.361	0.196	3.557	0.324
Sigmoid	DIM	-0.089	0.067	0.156	0.008
	IPW	0.228	0.012	-0.216	0.118
	AIPW	-0.287	0.035	0.322	0.097
Sine	DIM	-0.107	0.027	0.134	0.013
	IPW	0.366	0.023	-0.343	0.063
	AIPW	0.272	0.053	-0.219	0.131

Table 6: Comparison of three estimators across outcome models. For each outcome, the smallest absolute ATE bias, Y_1 bias, Y_0 bias, and the smallest ATE variance are highlighted in red.

Next, we apply our two-stage, single-switch incentivization policy to the confounded scenario. As before, we use $n_1 = 300$ initial observations to fit the penalized logistic model and then construct the bonus function $s^*(X)$ at threshold $\lambda = 0.4$. Table 7 presents the resulting ATE bias and variance for each estimator across all three outcome models.

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	ATE variance
Continuous	DIM	-0.163	0.013	0.176	0.020
	IPW	0.311	-0.034	-0.345	0.028
	AIPW	-0.022	0.011	0.033	0.010
Sigmoid	DIM	0.013	0.044	0.031	0.004
	IPW	0.187	-0.004	-0.556	0.009
	AIPW	-0.025	0.008	0.033	0.010
Sine	DIM	-0.014	0.001	0.015	0.007
	IPW	0.155	-0.005	-0.159	0.010
	AIPW	0.019	0.012	-0.007	0.015

Table 7: Comparison of three estimators across outcome models. For each outcome, the smallest absolute ATE bias, Y_1 bias, Y_0 bias, and the smallest ATE variance are highlighted in red.

Comparing Table 3 (no incentives) with Table 7 (with incentives) demonstrates the dramatic impact of our two-stage policy on bias and variance under confounding. In particular, it highlights

the exceptional robustness of our incentivization policy when paired with the AIPW estimator. In the continuous outcome, AIPW's ATE bias plummets from -3.361 to -0.022(variance $0.324 \to 0.010$). For the sigmoidal model, bias improves from -0.287 to -0.025 (variance $0.097 \to 0.010$), and in the sinusoidal case from 0.272 to 0.019 (variance $0.131 \to 0.015$). These improvements confirm that incentivization restores covariate overlap, stabilizes propensity-score weights, and enables all estimators—especially AIPW—to achieve near-oracle accuracy even with hidden confounding.

Overall, these results underscore two key points. First, in the presence of unobserved confounders and an imbalanced covariate distribution, the simple DIM estimator exhibits very large bias, and propensity-score–based methods (IPW and even doubly robust AIPW) become highly unstable—often performing worse than DIM. Second, introducing our targeted incentivization to restore overlap significantly reduces both bias and variance for all three estimators.

In our final experiment, we assess the robustness of the incentivization policy under deliberate choice-model misspecification. Specifically, we generate treatment assignments according to

$$U_{i0} = \varepsilon_{i0}, \quad U_{i1} = X_i^\top \theta^* + \varepsilon_{i1}, \quad W_i = \mathbf{1}\{U_{i1} > U_{i0}\},$$

where $\varepsilon_{i0}, \varepsilon_{i1} \sim \mathcal{N}(0,1)$ are standard Gaussian noise (instead of the Gumbel noise assumed by the logistic model). We nevertheless fit a logistic regression to estimate the propensity scores, thereby introducing misspecification. Remarkably, thanks to the doubly-robust property of AIPW and our incentive-driven covariate balancing, the AIPW estimator continues to achieve the lowest bias and variance in almost every scenario. Table 8 reports the detailed results.

Outcome	Method	ATE bias	Y_1 bias	Y_0 bias	Variance
Polynomial	DIM	2.277	0.520	-1.757	0.156
	IPW	1.467	0.063	-1.403	0.030
	AIPW	0.013	0.020	0.006	0.012
Sigmoid	DIM	0.232	0.052	-0.179	0.008
	IPW	0.384	0.007	-0.391	0.007
	AIPW	0.004	0.002	0.006	0.011
Sine	DIM	-0.035	0.007	0.042	0.007
	IPW	0.107	0.009	-0.116	0.006
	AIPW	-0.006	0.001	0.006	0.011

Table 8: Comparison of three estimators across outcome models. For each outcome, the smallest absolute ATE bias, Y_1 bias, Y_0 bias, and the smallest ATE variance are highlighted in red.

D More on Real World Experiment

Table 9 presents the choice-model estimates under both logistic-regression and random-forest specifications. In each case, stated genre preference emerges as the strongest predictor of video selection, followed by curiosity about science and technology.

Table 9: Choice Model Estimates: Logistic Regression vs. Random Forest

Feature	**Logit Coef.**	**RF Importance**
Age 26–30	−0.0242	0.0051
Age 31–40	−0.0498	0.0050
Age 40+	−0.0292	0.0054
Age Under 18	+0.0249	0.0005
Sex: Male	+0.2928	0.0239
Enjoyment of emotional materials	+0.0529	0.0252
Curiosity about science/technology progress	+0.3266	0.0457
Preference (Sci-Fi=1)	**+5.9469**	**0.8892**

Table 10 presents the fitted choice models under logistic regression and random forest. While genre preference remains the dominant predictor, its relative importance falls sharply after incentivization—from 0.88 to 0.25 in

the random forest and from 5.95 to 1.53 in the logistic model—confirming that our policy effectively attenuates its influence on treatment assignment.

Table 10: Choice Model Estimates: Logistic Regression Coefficients & Random Forest Importances

Feature	**Logit Coefficient**	**RF Importance**
Age 26–30	−0.4374	0.0405
Age 31–40	−0.9849	0.0560
Age 40+	−1.2820	0.0701
Sex: Male	+0.9047	0.1558
Enjoyment of imaginative or emotional materials	+0.0411	0.2155
Curiosity about scientific or technological progress	+0.3802	0.2116
Preference (Sci-Fi = 1)	**+1.5336**	**0.2506**