Aligning Incentives to Balance Covariates in Experiments with Selection Bias

Jiachun Li[1], **Yang Meng**[2], **David Simchi-Levi**[1], **Chonghuan Wang**[1]
[1]Laboratory for Information and Decision Systems, MIT, Cambridge, MA, USA
[2]Department of Computer Science, University of Chicago, Chicago, IL, USA
{jiach334, dslevi, chwang9}@mit.edu, ymeng3@uchicago.edu

Abstract

Estimating the average treatment effect (ATE) when participants can self-select into treatment or control groups based on their preferences can lead to significant selection bias and large variance of the estimation. We propose an incentivization framework that realigns participant preferences to balance covariates, thereby reducing bias and variance in treatment effect estimation. Our approach leverages incentive mechanisms solved under budget constraints to redistribute participants towards underrepresented groups. We provide theoretical guarantees for variance reduction using the Augmented Inverse Probability Weighting (AIPW) estimator and analyze the impact of unobserved confounders, showing that aligning incentives mitigates bias in treatment effect estimation. To achieve these goals, we introduce a low-switching learning-to-incentivize algorithm that dynamically adjusts incentives while adhering to resource constraints, achieving consistent and asymptotically efficient ATE estimation.

Introduction and Related Work

In estimating average treatment effect (ATE), conventional A/B testing 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)).

The problem of estimating causal effects when individuals do not adhere to their assigned treatment group is called non-compliance problem in econometrics/causal inference. Existing methods address this issue in various settings, but the effects they estimate, such as the intent-to-treat (ITT) effect or local-average-treatment-effect (LATE) derived from instrumental variable (IV) approaches, differ from the conventional average treatment effect (Wang and Tchetgen Tchetgen (2018), Swanson et al. (2018)). However, drawing from perspectives in economics and operations research (McFadden (1974), Greene (2012), Train (2009)), if we assume that non-compliance is not random and that individuals behave rationally in making choices, the pattern of non-compliance can be modeled mathematically. This approach utilizes random utility theory, which, while less conventional in econometrics, is widely employed in behavioral studies. Specifically, well-established random choice models, such as the Multinomial Logit (MNL) model, can be applied to systematically characterize non-compliance behavior. These models assume that participants select options based on their preferences and utilities, providing a structured framework to analyze and predict non-compliance patterns.

Our goal is to minimize the Mean Square Error (MSE) of our estimation of treatment effect, which is composed of bias and variance. The self-selection process introduces bias known as acquisition bias (Hernán et al. (2004), Rosenbaum and Rubin (1983), Heckman (1979), Imbens and Wooldridge (2009), Manski (1990), Stuart (2010)), since outcomes are correlated with utility due to the existence of confounding factors. Namely, the utility of certain choice might highly depends on the expected outcome. Therefore, one of the most fundamental assumption in causal inference literature, unconfoundedness, doesn't hold here, leading to failure of existing estimation procedures.

Under an ideal setting when all confounders are observed, we prove that the use of the propensity score adjustment method can correct for selection bias by reweighting or stratifying the participants based on the probability of being assigned to the treatment, fixing the bias caused by self-selection. While this approach effectively removes selection bias, it introduces a new challenge: high variance due to feature imbalance. In order to correct for bias, we are putting much more weight on rare observations, causing the estimation to be unstable. Addressing this feature imbalance remains critical for achieving more precise and reliable ATE estimates (Hainmueller (2012), Imai and Ratkovic (2014), Tan (2010)).

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 misalign-

ment, we introduce external intervention mechanisms designed to realign incentives to acheive feature 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, in most practical cases unobserved confounders exist and can never be identified, and we prove that under mild assumption on the confounding effect, we can prove that the estimation effect is systematically underestimated, and the bias is increasing with larger incentive gaps. In summary, we analyze the effect of incentives in bias and variance of treatment effect estimation, showing that incentive misalignment can leads to large variance and self-selection bias, proving optimal incentive mechanism to have a accurate and stable estimation. In addition, we propose adaptive experiments that begin without knowledge of the underlying model, iteratively estimate key parameters during the experiment, and adjust the incentive mechanism based on the updated model. This dynamic approach ensures that incentives are progressively optimized, achieving accurate and stable treatment effect estimation under practical constraints.

Problem Formulation

Notations

We draw i.i.d. samples $X_1, X_2, \ldots, X_n$ from a distribution P_X on a compact set $\mathcal{X}$. Each participant chooses a treatment $W_i \in \{0, 1\}$ (treatment or control) based on covariates X_i, with the propensity score defined as $e(X_i) := P(W_i = 1 \mid X_i)$. To ensure overlap, we assume $\eta < e(X) < 1-\eta$ for some constant $\eta > 0$, known as "overlap condition" (Imai and Ratkovic (2014), Petersen et al. (2012)).

The observed outcome is $Y_i = Y^{(W_i)}(X_i)$, with potential outcomes $Y^{(1)}(X)$ and $Y^{(0)}(X)$, which have expectations $\mu^{(1)}(X)$, $\mu^{(0)}(X)$ and variances $(\sigma^{(1)})^2$, $(\sigma^{(0)})^2$. These can be expressed as:

$$Y^{(1)}(X_i) = \mu^{(1)}(X_i) + \epsilon_i, Y^{(0)}(X_i) = \mu^{(0)}(X_i) + \epsilon_i,$$

where ϵ_i represents random noise. For simplicity, we assume constant variance. Participants make decisions based on utility functions $\alpha^{(1)}(X_i)$ and $\alpha^{(0)}(X_i)$, modeled linearly as:

$$\alpha^{(1)}(X_i) = V^{(1)}(X_i) + \epsilon_1, \alpha^{(0)}(X_i) = V^{(0)}(X_i) + \epsilon_0,$$

where $V^{(1)}(X_i) = \theta_1^\top X_i + c_1$ and $V^{(0)}(X_i) = \theta_0^\top X_i + c_0$. Based on the multinomial logit model, participants choose the treatment with the highest utility, leading to a Bernoulli-distributed treatment assignment:

$$e(X_i) = \frac{\exp(V^{(1)}(X_i))}{\exp(V^{(1)}(X_i)) + \exp(V^{(0)}(X_i))}.$$

We assume unconfoundedness, meaning potential outcomes $\{Y^{(1)}(X_i), Y^{(0)}(X_i)\}$ are conditionally independent of W_i given X_i:

Assumption 1. $\{Y^{(1)}(X_i), Y^{(0)}(X_i)\} \perp W_i \mid X_i$

Remark 1. The error term in utilities follows a Gumbel distribution. This random component reflects uncertainties such as unpredictability in future outcomes, variability in human behavior, or partial information, and it is independent of the outcome Y. Based on these reasons, humans will have a random decision based on a noisy evaluation of their utility. Thus, the source of randomness in human choice is uncorrelated with the realization of outcome. This independence ensures that the selection process, represented by W, does not confound the outcome Y.

The primary causal quantity of interest is the average treatment effect (ATE):

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

After collecting all data from the experiment, the experimenter constructs an estimator $\hat{\tau}$ of τ. Our goal is to minimize the expected square loss of $\hat{\tau}$, $E[(\hat{\tau} - \tau)^2]$, where the expectation is taken over all data and treatment randomness.

Incentivize through Utility Distortion

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 this bias, they often result in large estimation variance. To simultaneously reduce both bias and variance, we propose incentivizing participants to achieve a more balanced allocation across groups.

Consider a total budget B distributed among n participants. Our policy is the incentive allocated to each participant denoted as $p(X_i)$ for $i = 1, ..., n$. Here we assume that p is a deterministic policy based on X_i, meaning that each set of features corresponds to a unique incentive. The policy allocates incentives to participants choosing the less preferred treatment, effectively redistributing participants to balance the groups.

For a participant with features X_i who prefers the control group ($\alpha^{(1)}(X_i) \leq \alpha^{(0)}(X_i)$), an incentive $p(X_i)$ shifts the preference towards the treatment group. This adjusts the propensity score closer to 0.5, calculated as:

$$e(X_i) = P(W_i = 1 \mid X_i) \tag{1}$$

$$= \frac{\exp(V^{(1)}(X_i) + p(X_i))}{\exp(V^{(1)}(X_i) + p(X_i)) + \exp(V^{(0)}(X_i))}. \tag{2}$$

Estimator

Assume an experiment collects data $\{X_i, W_i, Y_i\}_{i=1}^n$. We use the Augmented Inverse Probability Weighting (AIPW) estimator. Under assumptions of conditional independence, positivity, and correct model specification, the true AIPW is:

$$\tau^*_{\text{AIPW}} = E\Bigg[\mu^{(1)}(X) - \mu^{(0)}(X) + \frac{Y - \mu^{(1)}(X)}{e(X)} W - \frac{Y - \mu^{(0)}(X)}{1 - e(X)}(1 - W)\Bigg]. \tag{3}$$

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 X. The AIPW estimator is known to be unbiased and asymptotically efficient, providing optimal variance among unbiased estimators under mild conditions. It has two key advantages: (1) it explicitly incorporates the propensity score to adjust for confounding, thereby reducing bias due to imbalanced treatment assignments; and (2) it leverages regression-based outcome models to correct for potential noise or bias in the observed outcomes, further enhancing robustness and efficiency. This dual incorporation of propensity scores and regression estimation provides resilience to model misspecification when either the propensity score model or the outcome model is correctly specified.

The variance of the AIPW estimator is:

$$\text{Var}(\tau^*_{\text{AIPW}}) = \frac{1}{n} E\Bigg[\frac{\sigma^2}{e(X)(1-e(X))} + \big(\mu^{(1)}(X) - \mu^{(0)}(X) - \tau\big)^2\Bigg]. \quad (4)$$

where $e(X)$ is the true propensity score, and $\mu^{(1)}(X), \mu^{(0)}(X)$ are the true conditional means of the potential outcomes.

In practice, the true models $e(X), \mu^{(1)}(X)$, and $\mu^{(0)}(X)$ are unknown and must be estimated. The plugged-in AIPW estimator is:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\hat{\mu}^{(1)}(X_i) - \hat{\mu}^{(0)}(X_i) + \frac{Y_i - \hat{\mu}^{(1)}(X_i)}{\hat{e}(X_i)} W_i - \frac{Y_i - \hat{\mu}^{(0)}(X_i)}{1-\hat{e}(X_i)}(1-W_i)\Bigg). \quad (5)$$

To simplify the analysis, we focus on the true AIPW estimator τ^*_{AIPW} in the next two sections. We will analyze its properties under ideal conditions where the true mdoels are unknown. In the last section, we address the practical case of $\hat{\tau}_{\text{AIPW}}$, showing its convergence to τ^*_{AIPW} as $n \to \infty$, given sufficient data and appropriate model assumptions.

Optimal Incentivize Mechanism to Balance Covariates

If we are given a model $\alpha^{(1)}, \alpha^{(0)}, \mu^{(1)}, \mu^{(0)}, P_X$, the variance of our ATE estimation by the AIPW estimator τ^*_{AIPW} is given by:

$$\text{Var}(\tau^*_{\text{AIPW}}) = \frac{1}{n} E\Bigg[\frac{\sigma^2}{e(X)(1-e(X))} + \big(\mu^{(1)}(X) - \mu^{(0)}(X) - \tau\big)^2\Bigg]. \quad (6)$$

Our goal is to minimize this variance by incentivizing trial participants to choose the less preferred treatment, thereby achieving a more balanced distribution of features in the treatment and control groups. This balance helps reduce the first term in the variance expression, which depends on the propensity score $e(X)$. 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:

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

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:

$$\min_{p(X)} \int F(X, p(X)) P_X \, dX$$
$$\text{s.t.} \int p(X) P_X \, dX \le \frac{B}{n}, p(X) \ge 0, \forall X. \quad (7)$$

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 incentivize 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 1. *The function $F(X, p(X)) = \frac{\sigma^2}{e(X)(1-e(X))}$, where $e(X)$ depends on $p(X)$, 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. Convexity implies that as the propensity score $e(X)$ approaches its optimal value (e.g. 0.5 under symmetric variance assumptions), the variance decrease. However, the rate of decrease slows as the propensity score nears the optimal value.

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 1. *(Equal Derivative Solution) Under the convexity of F, 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 $p^*(X) = 0$.*

By the monotonicity of $e(X)$, there exists a threshold η such that:

- If $e(X) < \eta \leq \frac{1}{2}$ or $1-\eta < e(X) \leq 1$, the policy adjusts $e^*(X, p^*(X)) = \eta$.
- If $\eta \leq e(X) \leq 1-\eta$, no incentives are applied ($e^*(X,0) = e(X)$).

The key insight in theorem 1 lies in targeting participants with the most imbalanced propensity scores, as incentivizing them yields the fastest variance reduction under the same budget constraints. Specifically, the objective is to prioritize individuals whose preferences are strongly skewed toward one group, adjusting their incentives to achieve a more balanced allocation.

Bias Analysis with Unobserved Confounders

In section 2, we relied on the unconfoundedness assumption, which posits that all relevant confounders are observed, and thus, treatment assignment is conditionally independent of the potential outcomes. While this assumption is standard in causal inference, it is often considered too strong and untestable in practice, as it requires all confounders to be identified and measured (Imbens and Rubin (2015), Hernán and Robins (2020)). Furthermore, the unconfoundedness assumption is inherently unidentifiable because there is no empirical test to confirm whether all confounders have been accounted for.

To relax this assumption, we now consider a more realistic scenario where most, but not all, confounders are observed. Specifically, we assume that even after conditioning on the observed covariates X, there remain unmeasured factors that influence both treatment selection and potential outcomes. Mathematically, the utilities associated with treatment selection are modeled as:

$$\alpha^{(1)}(X_i) = V^{(1)}(X_i) + \epsilon_1,$$
$$\alpha^{(0)}(X_i) = V^{(0)}(X_i) + \epsilon_0$$

where ϵ_1 and ϵ_0 represent unobserved factors. As a result, the utilities $\alpha^{(1)}(X_i)$ and $\alpha^{(0)}(X_i)$ remain correlated with the potential outcomes $Y_1(X_i)$ and $Y_0(X_i)$, introducing residual confounding even after conditioning on X.

Equivalently, we can express the outcome as:

$$Y_i(X,\alpha) = f_i(X,\alpha) + \eta = \tilde{f}_i(X) + g_i(X,\alpha) + \eta,$$

where:

- $\tilde{f}_i(X) = E_\alpha[f_i(X,\alpha)]$ is the marginalized outcome, which depends only on X. This term represents the part of the outcome that is observable and can be estimated.
- $g_i(X,\alpha) = f_i(X,\alpha) - \tilde{f}_i(X)$ is the deviation caused by α, capturing the residual dependence of the outcome on the unobserved confounders. By construction, $g_i(X,\alpha)$ satisfies $E_\alpha[g_i(X,\alpha)] = 0$.

Intuitively, $\tilde{f}_i(X)$ accounts for the main effects of the observed covariates X, while $g_i(X,\alpha)$ quantifies the impact of the unobserved confounders α on the outcome. Since we expect that the remaining unobserved confounders have only a small impact on the outcome, we assume $g_i(X,\alpha)$ is small in magnitude. As the symmetry of α and β in the utility function, we will assume $g_1 = g_0 = g$ throughout this section.

This type of analysis, which involves quantifying and accounting for the residual impact of unobserved confounders, falls under the framework of partial identification or sensitivity analysis. Such approaches are useful for assessing the robustness of causal conclusions when full identification of treatment effects is infeasible.(Imbens and Rubin (2015))

To simplify the analysis, we fix X and occasionally omit it in the notation for brevity. We define the utility gap between the two choices as:

$$\text{gap}(X) = V^{(1)}(X) - V^{(0)}(X),$$

where the gap is conditioned on a given X. Without loss of generality, we assume $\text{gap}(X) > 0$.

Throughout this section, we analyze the bias under the true AIPW estimator and the estimations of the potential outcomes $\bar{Y}_1$ and $\bar{Y}_0$ can be expressed as:

$$\widehat{Y}_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\mu^{(1)}(X_i) + \frac{Y_i - \mu^{(1)}(X_i)}{e(X_i)}W_i\right),$$

$$\widehat{Y}_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\mu^{(0)}(X_i) + \frac{Y_i - \mu^{(0)}(X_i)}{1-e(X_i)}(1-W_i)\right).$$

These expressions represent the estimates of the potential outcomes based on the observed data and the true propensity scores $e(X_i)$, as well as the outcome models $\mu^{(1)}(X)$ and $\mu^{(0)}(X)$.

The selection bias associated with feature X can be characterized in the following lemma:

Lemma 2. *The selection bias in the estimated potential outcomes is given by:*

$$E[\widehat{Y}_1 - Y_1] = E_X\left[E\big(g(X,\alpha) \mid \alpha + gap(X) > \beta\big)\right],$$

$$E[\widehat{Y}_0 - Y_0] = E_X\left[E\big(g(X,\alpha) \mid \alpha - gap(X) > \beta\big)\right],$$

where α and β are independently drawn from a Gumbel distribution.

As a special case, if $gap(X) = 0$ and $g_1(X,\alpha) = g_0(X,\alpha)$ for all X, then the AIPW estimator $\hat{\tau}$ is unbiased, i.e.,

$$E[\hat{\tau}] = \tau.$$

The condition $g_1(X,\alpha) = g_0(X,\alpha)$ is necessary to achieve unbiasedness when $\text{gap}(X) = 0$. This is because $\alpha + \text{gap}(X)$ and $\alpha - \text{gap}(X)$ are symmetric expressions when $\text{gap}(X) = 0$. For $E[\widehat{Y}_1 - Y_1]$ and $E[\widehat{Y}_0 - Y_0]$ to vanish simultaneously when $\text{gap}(X) = 0$, the conditional expectations $E[g_1(X,\alpha) \mid \alpha > \beta]$ and $E[g_0(X,\alpha) \mid \alpha > \beta]$ must be equal for all X. This requires $g_1(X,\alpha)$ and $g_0(X,\alpha)$ to have identical functional forms (and distributions) for any given X. Intuitively, the symmetry between $\alpha + \text{gap}(X)$ and $\alpha - \text{gap}(X)$ ensures that the selection processes for the treatment and control groups are mirror images when the utility gap vanishes. If $g_1 \neq g_0$, differences in how unobserved confounders α influence treatment and control groups will introduce residual bias, even when $\text{gap}(X) = 0$.

This result highlights two critical factors in mitigating selection bias. First, aligned incentives, represented by

$\text{gap}(X) = 0$, eliminate preference-driven selection bias by making participants indifferent between the treatment and control groups. Second, symmetry in the residual confounder effects ($g_1 = g_0$) ensures that any unobserved confounder effects are balanced between the two groups. When these conditions are satisfied, the AIPW estimator remains unbiased, even in the presence of unobserved confounders. This lemma underscores that incentive misalignment, quantified through the utility gap $\text{gap}(X)$, is the fundamental source of selection bias.

The following result provides an upper bound on the magnitude of the selection bias:

Lemma 3. *Assume the confounding effect $g(X, \alpha)$ is ε_X-Lipschitz for some (arguably small) $\varepsilon_X > 0$. Then the selection bias can be bounded as:*

$$|E[\hat{\tau}] - \tau| \leq E_X[\varepsilon_X |gap(X)|].$$

Building on this, suppose we know only that $g(X, \alpha)$ is ε-Lipschitz for all X. In this case, the best possible partial identification interval for the true treatment effect τ is:

$$[\tau - \varepsilon E_X|\text{gap}(X, p^*(X))|,\ \tau + \varepsilon E_X|\text{gap}(X, p^*(X))|]\,,$$

where $p^*(X)$ is the optimal Equal Derivative solution designed to minimize variance. This shows that our incentivize mechanism is not only the best in variance reduction, but also for minimizing the bias range. In our specific setting, the confounder affects both the utility and the outcome, and these two are often highly correlated (Szklo and Nieto (2014)). This motivates the following assumption:

Assumption 2 (Monotonic Confounding)**.** *For each X and for both the treatment and control groups, the confounding effect $g_i(X, \alpha)$ is monotonic in α.*

This assumption reflects a natural setting where the confounding variable α has a consistent directional effect on the outcome Y. Under this mild and arguably reasonable assumption, we demonstrate that aligning incentives reduces or at least controls the scale of selection bias. First, We show that the estimation bias is a monotonic function of utility gap, and the treatment effect is always over-estimated for the preferred treatment.

Proposition 1. *The conditional bias for feature X is:*

$$bias(X) = \widehat{Y}_1(X) - \widehat{Y}_0(X) - (Y_1(X) - Y_0(X)) > 0.$$

If $gap(X) = 0$ and $g_1 = g_0$, the bias vanishes, i.e., $bias(X) = 0$. Moreover, the estimation bias is an increasing function of the utility gap. Specifically, there exists a monotonically increasing function t such that:

$$bias(X) = t(gap(X)).$$

This result establishes that aligning incentives, thereby reducing the utility gap, decreases the self-selection bias for each feature X. As a corollary, if the preference structure is consistent across the entire population, i.e., $\text{gap}(X) > 0$ or $\text{gap}(X) < 0$ for all X, aligning incentives reduces selection bias across the entire population. To illustrate this relationship more explicitly, we consider the following simple yet illustrative case:

Lemma 4. *If $g_0(X, \alpha) = g_1(X, \alpha) = \varepsilon\alpha - E[\varepsilon\alpha]$, then:*

$$bias(X) = \varepsilon gap(X).$$

This lemma demonstrates that when the deviation functions are linear in α, the selection bias scales directly with the utility gap $\text{gap}(X)$.

Learn to Incentivize: Adaptive Experiments to Reduce Variance

In this section, we return to the unconfoundedness assumption, which underpins much of our analysis. As discussed previously, while this assumption may be strong and untestable in practice, it provides a framework for identifying treatment effects. Additionally, when unconfoundedness cannot be fully assumed, the methods of partial identification from the last section remain applicable and can provide bounds on the treatment effect estimates.

Recall that our primary goal is to both reduce the selection bias and minimize estimation variance by incentivizing experiment participants appropriately. If we assume that all underlying models $\alpha^{(1)}(X), \alpha^{(0)}(X), \mu^{(1)}(X), \mu^{(0)}(X), P_X$ are known prior to the experiment, the **Equal Derivative Solution** provides the optimal allocation mechanism, denoted by $p^*(X)$, with a corresponding propensity score $e^*(X)$. This solution guarantees that the output estimator achieves the lowest possible variance, as established by the following theorem.

Theorem 2. *(Static Policy Optimal Variance)*
If the models $\alpha^{(1)}(X), \alpha^{(0)}(X), \mu^{(1)}(X), \mu^{(0)}(X), P_X$ are known, then the optimal estimation accuracy is achieved by the AIPW estimator:

$$Var\left(\hat{\tau}^*_{AIPW}\right) = v^*.$$

The theorem is given by the efficiency of AIPW estimator and also the optimiality of equal derivative solution design. Here v^* represents the minimal variance achievable under a static allocation policy, demonstrating the efficiency of static allocation when all models are known. 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 experiments that use adaptive budget allocation mechanisms, where the incentive policy $p(X)$ can depend on prior observations or history, we establish a fundamental lower bound on the estimation variance for any unbiased estimator. This result defines the statistical limit of the problem and highlights the constraints imposed by the available resources.

To formalize the variance limit in the context of budget constraints, we define $v^*(B)$ as the minimal achievable variance for unbiased estimation when allocating a total budget B across participants. Importantly, $v^*(B)$ is defined with respect to a single participant's allocation mechanism, scaled appropriately for n participants.

Theorem 3. *(Statistical Limit for Adaptive Policies) For any adaptive budget allocation mechanism or adaptive algorithm running on n participants satisfying the resource constraint $E_X\left(\sum_{i=1}^{n} p_i(X)\right) \leq B$, the output estimator $\hat{\tau}$ satisfies:*

$$Var(\hat{\tau}) \geq \frac{1}{n} v^*\left(\frac{B}{n}\right).$$

This inequality demonstrates that no allocation mechanism—adaptive or static—can achieve an estimation variance below the limit using the Equal Derivative Solution, serving as a benchmark against which all experimental designs can be evaluated.

In practical settings, the models $\alpha^{(1)}(X), \alpha^{(0)}(X), \mu^{(1)}(X), \mu^{(0)}(X), P_X$ are typically unknown. This raises a critical question:
Can we design an adaptive experiment that:

1. Learns the underlying models $\alpha^{(1)}(X), \alpha^{(0)}(X), \mu^{(1)}(X), \mu^{(0)}(X), P_X$ during the experiment, and
2. Constructs an (asymptotically) unbiased estimation $\hat{\tau}$ that approximates the optimal variance v^*:

$$Var(\hat{\tau}) \leq (1 + o(1))v^*$$

We give an affirmative answer by proposing the following low switching learning-to-incentivize algorithm.

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 $k = 1$ **do**
  Set batch length $m_k = c\sqrt{n}$.
  Collect m_k samples using policy $p(X) = 0$.
  Estimate initial model parameters $\hat{\theta}_1^{(1)}, \hat{\theta}_0^{(1)}, \hat{c}_1^{(1)}, \hat{c}_0^{(1)}$ via logistic regression.
end for
for $k = 2$ to K **do**
  Set batch length $m_k = 2^{k-2} c\sqrt{n}$.
  Solve the optimization problem with proportional budget to compute estimated policy $\hat{p}(X)$.
  Compute conservative policy $p^{\mathrm{LB}}(X) = \hat{p}(X) - 2O\left(\frac{1}{\sqrt{m_k}}\right)$.
  Collect m_k samples using policy $p(X) = p^{\mathrm{LB}}(X)$.
  Update model parameters $\hat{\theta}_1^{(k)}, \hat{\theta}_0^{(k)}, \hat{c}_1^{(k)}, \hat{c}_0^{(k)}$ via logistic regression.
end for
Compute final unbiased estimator $\hat{\tau}$ based on all collected data.

We propose an estimation which is consistent and its mean square error approaches the optimal variance through a **low-switching learning-to-incentivize algorithm**. This algorithm operates over $\log(n)$ sequential batches, dynamically refining the incentive policy to balance exploration and exploitation. Note that in the optimization step, we adopt a proportional budget approach, where the input budget is scaled by the fraction of samples remaining (e.g., if the total budget is B and n_1 samples remain, we allocate $\frac{n_1}{n}B$ as input) to ensure a conservative design.

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 two important purposes:
1. Conservative Budget Usage: This zero-price policy acts as a buffer to ensure that we do not overspend before the experiment concludes.
2. Pretraining Phase: Since no prior information about the model is available at the start, this batch allows us to collect unbiased data for rough model estimation.

Using the data (X_i, Y_i, W_i) collected in this phase, we estimate the underlying distribution P_X, which governs the covariates. A uniform convergence result provides a guarantee for the accuracy of this estimation (van der Vaart and Wellner (1996)):

Lemma 5. *(Uniform Convergence of Empirical Distribution). 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.

Assume:

1. *The function $p(x)$ is Lipschitz continuous, satisfying $|p'(x)| \leq L$.*
2. *K is a smooth, symmetric kernel such that:*

$$\int K(u)\, du = 1, \int uK(u)\, du = 0, \int u^2 K(u)\, du = \kappa.$$

3. *The kernel K is uniformly bounded: $|K(u)| \leq M$.*

Then, 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 3 (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 $E[Y \mid X = x] = \mu(x)$). 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 $\hat{\mu} : \mathcal{X} \to R$ approximating $\mu(x)$, such that with probability $1-\delta$ (for $\delta = \frac{1}{n^4}$):

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

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

Using the data (X_i, W_i), where $X_i \sim P_X$ represents the features and $W_i \in \{0,1\}$ denotes the treatment assignment, we estimate the parameters of the utility functions $\alpha^{(1)}(X)$ and $\alpha^{(0)}(X)$, which follow linear models:

$$\alpha^{(1)}(X) = \theta_1^\top X + c_1, \quad \alpha^{(0)}(X) = \theta_0^\top X + c_0.$$

The parameter set $\phi = (\theta_1, c_1, \theta_0, c_0)$ is estimated using logistic regression or similar methods. Importantly, the low-switching design ensures that the data within each batch is collected i.i.d., facilitating accurate parameter estimation. We make the following assumption for an oracle that guarantees estimation accuracy for logistic regression, which is already well established (Hosmer et al. (2013)):

Assumption 4 (Oracle for Parameter Estimation)**.** *Let $(X_1, W_1), \ldots, (X_m, W_m)$ denote a batch of i.i.d. data, where $X_i \sim P_X$ and the probability of treatment assignment follows:*

$$P(W_i = 1 \mid X_i) = e(X_i) = \frac{\exp(\alpha^{(1)}(X_i))}{\exp(\alpha^{(1)}(X_i)) + \exp(\alpha^{(0)}(X_i))}$$

Assume the batch size $m \ge C_3 \log n$ for a sufficiently large constant C_3. Then, there exists an estimation oracle that takes $(X_1, W_1), \ldots, (X_m, W_m)$ as input and outputs an estimated parameter set $\hat{\phi} = (\hat{\theta}_1, \hat{c}_1, \hat{\theta}_0, \hat{c}_0)$, such that with probability $1-\delta$ (for $\delta = \frac{1}{n^4}$):

1. *The squared error for $\hat{\theta}_1$ satisfies:*

$$E_{X\sim P_X}\left[\|\theta_1 - \hat{\theta}_1\|^2\right] \le C_4 \frac{d\sigma^2}{m^\alpha}\log\left(\frac{1}{\delta}\right),$$

   where $\alpha = 1$, d is the dimension of θ_1 and σ^2 represents the variance of the noise in the data.

2. *Similar guarantees hold for $\hat{\theta}_0, \hat{c}_1$, and $\hat{c}_0$, with appropriately defined constants C_5, C_6, C_7.*

Remark 2. The parameter estimation guarantees rely on the properties of P_X, which governs the covariate distribution. The batch size m must be sufficiently large to achieve these guarantees, ensuring accurate parameter recovery under the linear utility model.

In the second batch, we can solve the optimization problem with our estimated models to derive an estimated incentivize mechanism $\hat{p}(X)$.

Lemma 6. *Under appropriate regularity conditions and given a sufficiently large batch size n_i (where i denotes the batch index), the estimated mechanism $\hat{p}(X)$ is close to the true optimal allocation mechanism $p^*(X)$, with a high degree of confidence:*

$$|p^*(X) - \hat{p}(X)| \le \frac{1}{\sqrt{n_i}}.$$

To ensure conservativeness, we always select the lower bound of the confidence interval for incentives: $\hat{p}^{\mathrm{LB}}(X) = \hat{p}(X) - 2\frac{1}{\sqrt{n}}$. This choice satisfies:

$$p^*(X) - 3\frac{1}{\sqrt{n}} \le \hat{p}^{\mathrm{LB}}(X) \le p^*(X).$$

The next lemma tells the approximation error of the propensity score.

Lemma 7. *Given the adjusted incentive mechanism $\hat{p}^{LB}$, the resulting adjusted propensity score $e^{LB}(X)$ is guaranteed to approximate the optimal propensity score $e^*(X)$ with high confidence:*

$$|e^*(X) - e^{LB}(X)| \le \frac{1}{\sqrt{n_i}}.$$

We then apply the conservative incentive policy $\hat{p}^{\mathrm{LB}}(X)$, 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 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 $\hat{\tau}$ is consistent. 2. The variance of $\hat{\tau}$ is close to the theoretical lower bound v^*:

Theorem 4. *After completing the experiment, the mean square error of the estimator $\hat{\tau}$ satisfies:*

$$E[(\hat{\tau}-\tau)^2] \le (1 + O(n^{-\frac{1}{2}\alpha}))v^*.$$

Proposition 3. *By Central Limit Theorem, the estimator satisfies:*

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

where v^* is the minimum variance achievable.

Conclusion

In this work, we proposed a novel approach to address selection bias and covariate imbalance in experimental designs where participants self-select into treatment or control groups. Through the integration of random utility models and the application of incentivization mechanisms, we demonstrated that aligning participants' incentives with the experimental objectives effectively minimizes bias and variance in treatment effect estimation. Our theoretical framework, supported by the derivation of optimal incentive policies and robust estimation techniques, highlights the importance of balancing feature distributions while adhering to practical constraints such as budget limitations. Furthermore, we extended our analysis to settings with unobserved confounders, providing insights into bias dynamics and partial identification of treatment effects. Finally, we introduced adaptive experimental designs that iteratively estimate key parameters and refine incentive mechanisms in real time, ensuring that treatment effect estimation becomes progressively more accurate and stable as the experiment progresses.

References

Greene, W. H. (2012). *Econometric Analysis* (7th ed.). Pearson Education.

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

Haussmann, M., T.-M. S. Le, V. Halla-aho, et al. (2023). Estimating treatment effects from single-arm trials via latent-variable modeling. *arXiv preprint arXiv:2311.03002*.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Hernán, M. A., S. Hernández-Díaz, and J. M. Robins (2004). A structural approach to selection bias. *Epidemiology 15*(5), 615–625.

Hernán, M. A. and J. M. Robins (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Logistic Regression: A Self-Learning Text* (3rd ed.). Springer.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 243–263.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 5–86.

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

Li, J., D. Simchi-Levi, and Y. Zhao (2024). Optimal adaptive experimental design for estimating treatment effect. *arXiv preprint arXiv:2410.05552*.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press.

Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research 21*(1), 31–54.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science 25*(1), 1–21.

Swanson, S. A., M. A. Hernán, M. Miller, J. M. Robins, and T. S. Richardson (2018). 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.

Szklo, M. and F. J. Nieto (2014). *Epidemiology: Beyond the Basics* (3rd ed.). Jones & Bartlett Learning.

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

Train, K. E. (2009). *Discrete Choice Methods with Simulation* (2nd ed.). Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Wager, S. (2024). Causal inference: A statistical learning approach.

Wang, L. and E. J. Tchetgen Tchetgen (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(3), 531–550.

Appendix

Proof of Lemma 1

Assume $e(X) < 0.5$ WLOG. We aim to prove that the function $F(X, p(X)) = f(e(X, p(X)))$, where $f(e) = -\frac{\sigma^2}{e(1-e)}$, satisfies: 1. $F'(X, p(X)) < 0$ (non-increasing) for $e(X, p(X)) < 0.5$, and 2. $F''(X, p(X)) > 0$ (convexity) for $e(X, p(X)) < 0.5$, where $e(X, p(X))$ is the softmax-based probability for the treatment group.

Define the propensity score $e(X, p(X))$ as:

$$e(X, p(X)) = \frac{\exp(u_1(X) + p(X))}{\exp(u_1(X) + p(X)) + \exp(u_0(X))}.$$

Here, $u_1(X)$ and $u_0(X)$ are fixed utility components for the treatment and control groups, while $p_1(X)$ is the incentive adjustment applied to the treatment group. Introduce the notation:

$$A = \exp(u_1(X) + p(X)), \quad B = \exp(u_0(X)).$$

Thus, the propensity score simplifies to:

$$e(X, p(X)) = \frac{A}{A + B}.$$

The first derivative of $e(X, p(X))$ with respect to $p(X)$ is computed as:

$$\frac{d}{dp(X)} e(X, p(X)) = \frac{\frac{dA}{dp(X)}(A + B) - A\frac{d(A+B)}{dp(X)}}{(A + B)^2}.$$

Since $\frac{dA}{dp(X)} = A$ and $\frac{d(A+B)}{dp(X)} = A$, we have:

$$\frac{d}{dp(X)} e(X, p(X)) = \frac{AB}{(A + B)^2}.$$

This is always positive for $A, B > 0$.

The second derivative is:

$$\frac{d^2}{dp(X)^2} e(X, p(X)) = \frac{\frac{d}{dp(X)} AB \cdot (A + B)}{(A + B)^3} - \frac{2AB \cdot \frac{d}{dp(X)}(A + B)}{(A + B)^3}.$$

Using $\frac{d}{dp(X)} AB = AB$ and $\frac{d}{dp(X)}(A + B) = A$, this simplifies to:

$$\frac{d^2}{dp(X)^2} e(X, p(X)) = \frac{AB\left[(A + B) - 2A\right]}{(A + B)^3}.$$

When $e(X, p(X)) < 0.5$, we have $A < B$, which implies $(A + B) - 2A = B - A > 0$. Thus:

$$\frac{d^2}{dp(X)^2} e(X, p(X)) > 0.$$

This proves that $e(X, p(X))$ is convex with respect to $p(X)$.

Now consider $f(e) = -\frac{\sigma^2}{e(1-e)}$, where $e = e(X, p(X))$.

Using the chain rule and computing the derivative of $f(e)$, we have:

$$f'(e) = \frac{\sigma^2}{e^2} - \frac{\sigma^2}{(1-e)^2}.$$

When $e(X, p(X)) < 0.5$, we know $e < 1 - e$, so $e^2 < (1 - e)^2$. This implies:

$$f'(e) < 0.$$

Thus, $f(e(X, p(X)))$ is non-increasing for $e(X, p(X)) < 0.5$.

The second derivative is:

$$f''(e) = -\frac{2\sigma^2}{e^3} - \frac{2\sigma^2}{(1-e)^3}.$$

Since $0 < e < 1$, both terms are positive, so:

$$f''(e) > 0.$$

This proves that $f(e(X, p(X)))$ is convex for all $0 < e(X, p(X)) < 1$.

Finally, consider the composition $F(X, p(X)) = f(e(X, p(X)))$. Using the above results:
1. $e(X, p(X))$ is convex with respect to $p(X)$,
2. $f(e) = -\frac{\sigma^2}{e(1-e)}$ is convex and non-increasing for $e(X, p(X)) < 0.5$. The composition rule for convex functions states that if $f(e)$ is convex and non-decreasing, or in this case non-increasing and strictly positive, then $F(X, p(X)) = f(e(X, p(X)))$ is convex whenever $e(X, p(X))$ is convex.

Thus, $F(X, p(X))$ is both convex and non-increasing for $e(X, p(X)) < 0.5$. The analogous results hold for $e(X, p(X)) > 0.5$.

Proof of Theorem 1

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

$$\min_{p(X)} \int F(X, p(X)) P_X \, dX$$

$$\text{s.t.} \int p(X) P_X \, dX \leq \frac{B}{n}, p(X) \geq 0, \forall X. \tag{8}$$

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

$$\begin{aligned}\mathcal{L}(p, \lambda, \mu) = &\int F(X, p(X)) P_X dX \\ &+ \int \mu(X)(-p(X)) \, dX \\ &+ \lambda \left(\int p(X) P_X \, dX - C \right). \end{aligned} \tag{9}$$

where $\lambda \geq 0$ and $\mu(X) \geq 0$ 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 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:

$$p(X) = 0.$$

Proof of Lemma 6

Utility function estimation To bound the error between the true function $\alpha^{(1)}(X)$ and its estimate $\hat{\alpha}^{(1)}(X)$, consider the following setup:

The true function is given by $\alpha^{(1)}(X) = \theta_1^\top X + c_1$, while the estimated function is $\hat{\alpha}^{(1)}(X) = \hat{\theta}_1^\top X + \hat{c}_1$.

The difference between the two functions is:

$$\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X) = (\theta_1^\top X + c_1) - (\hat{\theta}_1^\top X + \hat{c}_1),$$

which simplifies to:

$$\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X) = (\theta_1 - \hat{\theta}_1)^\top X + (c_1 - \hat{c}_1).$$

Taking the norm of this difference, we have:

$$\|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| = \|(\theta_1 - \hat{\theta}_1)^\top X + (c_1 - \hat{c}_1)\|.$$

The term $(\theta_1 - \hat{\theta}_1)^\top X$ is a scalar, so its norm can be bounded using the Cauchy-Schwarz inequality:

$$\|(\theta_1 - \hat{\theta}_1)^\top X\| \leq \|\theta_1 - \hat{\theta}_1\|\|X\|.$$

If we assume that X is bounded by a constant $M > 0$, i.e., $\|X\| \leq M$, this simplifies to:

$$\|(\theta_1 - \hat{\theta}_1)^\top X\| \leq M\|\theta_1 - \hat{\theta}_1\|.$$

In assumption 4, we assume that the parameter θ_1 is estimated with an error bounded by a small constant $\epsilon > 0$:

$$\|\theta_1 - \hat{\theta}_1\| \leq \epsilon.$$

Combining this with the previous inequality, we obtain:

$$\|(\theta_1 - \hat{\theta}_1)^\top X\| \leq M\epsilon.$$

Now, consider the full error between $\alpha^{(1)}(X)$ and $\hat{\alpha}^{(1)}(X)$:

$$\|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| = \|(\theta_1 - \hat{\theta}_1)^\top X + (c_1 - \hat{c}_1)\|.$$

Using the triangle inequality, we split this into two terms:

$$\|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| \leq \|(\theta_1 - \hat{\theta}_1)^\top X\| + \|c_1 - \hat{c}_1\|.$$

Substituting the bound for $\|(\theta_1 - \hat{\theta}_1)^\top X\|$, we get:

$$\|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| \leq M\epsilon + \|c_1 - \hat{c}_1\|.$$

A similar argument applies to the second function, $\alpha^{(0)}(X)$, where:

$$\alpha^{(0)}(X) = \theta_0^\top X + c_0, \quad \hat{\alpha}^{(0)}(X) = \hat{\theta}_0^\top X + \hat{c}_0.$$

By following the same steps, we obtain:

$$\|\alpha^{(0)}(X) - \hat{\alpha}^{(0)}(X)\| \leq M\epsilon + \|c_0 - \hat{c}_0\|.$$

Propensity Score Estimation Denote $\|e(X) - \hat{e}(X)\|, \|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\|$, and$\|\alpha^{(0)}(X) - \hat{\alpha}^{(0)}(X)\|$ as the errors in the propensity score and utility functions. Without loss of generality, we assume that the price policy $p(X) = 0$ in this analysis.

We approximate the change in $e(X)$ using the first-order Taylor expansion, then take the absolute value. By triangle inequality:

$$\begin{aligned}\|e'(X) - e(X)\| \leq &\left|\frac{\partial e}{\partial \alpha^{(1)}}\right| \|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| \\ &+ \left|\frac{\partial e}{\partial \alpha^{(0)}}\right| \|\alpha^{(0)}(X) - \hat{\alpha}^{(0)}(X)\|.\end{aligned} \tag{10}$$

Here, $e(X)$ is given by:

$$e(X) = \frac{\exp(\alpha^{(1)}(X))}{\exp(\alpha^{(1)}(X)) + \exp(\alpha^{(0)}(X))}.$$

First, calculate $\frac{\partial e}{\partial \alpha^{(1)}}$. Using the quotient rule:

$$\frac{\partial e}{\partial \alpha^{(1)}} = \frac{\partial}{\partial \alpha^{(1)}}\left(\frac{\exp(\alpha^{(1)})}{\exp(\alpha^{(1)}) + \exp(\alpha^{(0)})}\right).$$

Let $f = \exp(\alpha^{(1)})$ and $g = \exp(\alpha^{(1)}) + \exp(\alpha^{(0)})$. Then:

$$\frac{\partial e}{\partial \alpha^{(1)}} = \frac{f' \cdot g - f \cdot g'}{g^2}.$$

Substituting $f' = \exp(\alpha^{(1)})$, $g' = \exp(\alpha^{(1)})$, and $g = \exp(\alpha^{(1)}) + \exp(\alpha^{(0)})$, we have:

$$\frac{\partial e}{\partial \alpha^{(1)}} = \frac{\exp(\alpha^{(1)})\exp(\alpha^{(0)})}{(\exp(\alpha^{(1)}) + \exp(\alpha^{(0)}))^2}$$

Simplify:

$$\frac{\partial e}{\partial \alpha^{(1)}} = e(X)(1 - e(X)).$$

Next, calculate $\frac{\partial e}{\partial \alpha^{(0)}}$. Similarly:

$$\frac{\partial e}{\partial \alpha^{(0)}} = -\frac{\exp(\alpha^{(1)})\exp(\alpha^{(0)})}{(\exp(\alpha^{(1)}) + \exp(\alpha^{(0)}))^2}.$$

This simplifies to:

$$\frac{\partial e}{\partial \alpha^{(0)}} = -e(X)(1 - e(X)).$$

Thus, the total change in $e(X)$ is:

$$\begin{aligned}\|e'(X) - e(X)\| &\le \left|\frac{\partial e}{\partial \alpha^{(1)}}\right| \|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| \\ &\quad + \left|\frac{\partial e}{\partial \alpha^{(0)}}\right| \|\alpha^{(0)}(X) - \hat{\alpha}^{(0)}(X)\| \\ &\le e(X)(1-e(X))\big(\|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| \\ &\quad + \|\alpha^{(0)}(X) - \hat{\alpha}^{(0)}(X)\|\big). \qquad (11)\end{aligned}$$

Since:

$$\|\alpha^{(1)}(X) - \hat{\alpha}^{(1)}(X)\| + \|\alpha^{(0)}(X) - \hat{\alpha}^{(0)}(X)\| \le \epsilon_1 + \epsilon_2.$$

Thus:

$$\|e(X) - \hat{e}(X)\| \le e(X)(1-e(X))(\epsilon_1 + \epsilon_2).$$

If $\epsilon_1 = \epsilon_2 = \epsilon$, then:

$$\|e(X) - \hat{e}(X)\| \le e(X)(1-e(X))(2\epsilon).$$

The maximum value of $e(X)(1-e(X))$ occurs at $e(X) = 0.5$, where $e(X)(1-e(X)) = 0.25$. Thus:

$$\|e(X) - \hat{e}(X)\| \le 0.25 \cdot 2\epsilon = 0.5\epsilon.$$

Incentive policy estimation

The decision variable $p(X)$ is determined by matching the speed of decrease in variance, given by the gradient:

$$\frac{\partial F(X, p(X))}{\partial p(X)} = -\lambda.$$

We know:

$$\begin{aligned}e(X, p(X)) &= \frac{\exp(\alpha^{(1)})\exp(p(X))}{\exp(\alpha^{(1)} + p(X)) + \exp(\alpha^{(0)})} \\ &= \frac{\exp(\alpha^{(1)})\exp(p(X))}{\exp(p(X)) + \exp(\alpha^{(0)} - \alpha^{(1)})} \\ &= \frac{e(X)\exp(p(X))}{e(X)\exp(p(X)) + (1-e(X))}.\end{aligned}$$

and $\|e(X) - \hat{e}(X)\| \le \epsilon$.

Claim 1: $\|e(X, p(X)) - \hat{e}(X, p(X))\| \le \epsilon$

To approximate this, we use a first-order Taylor expansion:

$$\|e(X, p(X)) - \hat{e}(X, p(X))\| \le \left|\frac{\partial e(X, p(X))}{\partial e(X)}\right| \cdot \|e(X) - \hat{e}(X)\|.$$

Now, we compute the derivative:

$$\frac{\partial e(X, p(X))}{\partial e(X)} = \frac{e(X)\exp(2p(X)) + \exp(p(X))(1-e(X))^2}{(e(X)\cdot\exp(p(X)) + (1-e(X)))^2}.$$

By Taylor expansion of the exponential function, $\exp(p(X)) \le 1 + p(X) + \frac{p(X)^2}{2}$ and $\exp(2p(X)) \le 1 + 2p(X) + 2p(X)^2$, we can then bound the numerator as:

$$e(X)(1+2p(X)+2p(X)^2) + \left(1 + p(X) + \frac{p(X)^2}{2}\right)(1-e(X))^2.$$

and bound the denominator as:

$$\left(e(X) + (1-e(X)) + e(X)p(X) + \frac{e(X)p(X)^2}{2}\right)^2.$$

Combining together, when $p(X)$ is small, the derivative can be simplified to:

$$\frac{\partial e(X, p(X))}{\partial e(X)} \le e(X) + (1-e(X))^2 < 1.$$

Since $\|e(X) - \hat{e}(X)\| \le \epsilon$, it follows that:

$$\|e(X, p(X)) - \hat{e}(X, p(X))\| \le \epsilon.$$

Claim proved.

Claim 2:

We want to show that if $\|e(X) - \hat{e}(X)\| < \epsilon$, then $|F(\hat{e}(X)) - F(e(X))| < \epsilon$ for the function $F(e(X)) = \frac{\sigma}{e(X)(1-e(X))}$.

To do this, we use the first-order Taylor expansion:

$$|F(\hat{e}(X)) - F(e(X))| \le |F'(e(X))| \cdot \|\hat{e}(X) - e(X)\|.$$

To ensure that $|F(\hat{e}(X)) - F(e(X))| < \epsilon$ whenever $\|e(X) - \hat{e}(X)\| < \epsilon$, it's sufficient to have $|F'(e(X))|$ bounded.

The derivative $F'(e(X))$ can be computed as:

$$F'(e(X)) = \frac{\sigma(1-2e(X))}{e(X)^2(1-e(X))^2}.$$

Since $\eta < e(X) < 1-\eta$, we can bound $F'(e(X))$ by:

$$|F'(e(X))| = \left|\frac{\sigma(1-2e(X))}{e(X)^2(1-e(X))^2}\right| \le \frac{|\sigma|\cdot|1-2e(X)|}{e(X)^2(1-e(X))^2}.$$

Using $|1-2e(X)| \le \max(1-2\eta, 2\eta-1)$, we get:

$$|F'(e(X))| \le \frac{|\sigma|\cdot\max(1-2\eta, 2\eta-1)}{\eta^4}.$$

Claim proved.

If the two functions $F(e(X), p(X))$ and $F(\hat{e}(X), p(X))$ are close, then their gradients will also be close by Mean Value Theorem, provided that the functions are sufficiently smooth. Specifically:

$$\left|\frac{\partial F(e(X), p(X))}{\partial p} - \frac{\partial F(\hat{e}(X), p(X))}{\partial p}\right| \le L\,\epsilon,$$

for some constant L depending on the smoothness of the gradient.

By Implicit Function Theorem, small changes in the gradient result in small changes in $p(X)$ and $\hat{p}(X)$. Specifically, the difference between $p(X)$ and $\hat{p}(X)$ is bounded by:

$$|p(X) - \hat{p}(X)| \le \frac{L\,\epsilon}{|\partial^2 F/\partial p^2|}.$$

Here, $|\partial^2 F/\partial p^2|$ is the second derivative of F with respect to $p(X)$, which we assume is bounded away from 0 for stability.

Proof of Lemma 7

Each optimal policy $p^*(x)$ corresponds to an optimal allocation function $e^*(x)$, taking into account the entire distribution of x. Our policy $\hat{p}^{LB}(x)$ corresponds to an allocation function $e^{LB}(x)$.

To show that $|p^{(\mathrm{LB})} - p^*| < \epsilon$ implies $|e^{(\mathrm{LB})} - e^*| < \epsilon$,

where

$$e^* = \frac{\exp(\alpha_1 + p^*)}{\exp(\alpha_1 + p^*) + \exp(\alpha_0)},$$

$$e^{(\text{LB})} = \frac{\exp(\alpha_1 + p^{(\text{LB})})}{\exp(\alpha_1 + p^{(\text{LB})}) + \exp(\alpha_0)}.$$

The difference can be written as:

$$|e^{(LB)} - e^*| = \left| \frac{\exp(\alpha^{(1)} + p^{(LB)})}{\exp(\alpha^{(1)} + p^{(LB)}) + \exp(\alpha^{(0)})} - \frac{\exp(\alpha^{(1)} + p^*)}{\exp(\alpha^{(1)} + p^*) + \exp(\alpha^{(0)})} \right|.$$

Let:

$x_1 = \exp(\alpha^{(1)} + p^{(LB)}), x_2 = \exp(\alpha^{(1)} + p^*), y = \exp(\alpha^{(0)}).$

Then:

$$e^{(LB)} = \frac{x_1}{x_1 + y}, \quad e^* = \frac{x_2}{x_2 + y}.$$

The difference can be expressed as:

$$|e^{(LB)} - e^*| = \left| \frac{x_1}{x_1 + y} - \frac{x_2}{x_2 + y} \right|.$$

$$|e^{(LB)} - e^*| = \left| \frac{x_1(x_2 + y) - x_2(x_1 + y)}{(x_1 + y)(x_2 + y)} \right|.$$

$$|e^{(LB)} - e^*| = \frac{y|x_1 - x_2|}{(x_1 + y)(x_2 + y)}.$$

Since $x_1 = \exp(\alpha^{(1)} + p^{(LB)})$ and $x_2 = \exp(\alpha^{(1)} + p^*)$, we have:

$$|x_1 - x_2| = \left| \exp(\alpha^{(1)} + p^{(LB)}) - \exp(\alpha^{(1)} + p^*) \right|.$$

$$|x_1 - x_2| = \exp(\alpha^{(1)}) \left| \exp(p^{(LB)}) - \exp(p^*) \right|.$$

Using the Mean Value Theorem on $\exp(p)$, which has a derivative of $\exp(p)$, we get:

$$\left| \exp(p^{(LB)}) - \exp(p^*) \right| \leq \exp(\xi)|p^{(LB)} - p^*|,$$

where ξ is between $p^{(LB)}$ and p^*. Since $\exp(\xi) \leq \max(\exp(p^{(LB)}), \exp(p^*))$, we can bound $|x_1 - x_2|$ as:

$$|x_1 - x_2| \leq \exp(\alpha^{(1)}) \max(\exp(p^{(LB)}), \exp(p^*))|p^{(LB)} - p^*|.$$

Substitute this bound for $|x_1 - x_2|$ into the expression for $|e^{(LB)} - e^*|$:

$$|e^{(LB)} - e^*| \leq \frac{\exp(\alpha^{(1)}) \max(\exp(p^{(LB)}), \exp(p^*))}{y} \times |p^{(LB)} - p^*|.$$

The denominator $(x_1 + y)(x_2 + y)$ is positive and can be bounded below:

$$(x_1 + y)(x_2 + y) \geq y^2,$$

because $x_1, x_2 \geq 0$.

Thus:

$$|e^{(LB)} - e^*| \leq \frac{\exp(\alpha^{(1)}) \max(\exp(p^{(LB)}), \exp(p^*))|p^{(LB)} - p^*|}{y}$$

Since $|p^{(LB)} - p^*| < \epsilon$, we have:

$$|e^{(LB)} - e^*| \leq C \cdot \epsilon,$$

where C is a constant depending on $\exp(\alpha^{(1)})$, $\exp(\alpha^{(0)})$, and the range of $p^{(LB)}$ and p^*. By choosing ϵ small enough, this guarantees that $|e^{(LB)} - e^*| < \epsilon$. This completes the proof.

Proof for Proposition 2

Under batch i, we draw samples $\{X_{i1}, \ldots, X_{in_i}\}$ and assign the policy $p^*(X_{ij})$ to each sample X_{ij}. Let B_i^* be the spending under policy $p^*(X_{ij})$ in batch i, and B_i be the spending under conservative policy $\hat{p}^{LB}(X_{ij})$ in batch i, and B be the expected spending per sample under the distribution of X.

Consider the base case: The budget spent in Batch 1 is B_1, and no incentives are provided in the first batch ($B_1 = 0$). Therefore:

$$nB - B_1 \geq (n - n_1)B$$

because $B_1 = 0$ where n_1 is the number of samples in Batch 1. This satisfies the budget constraint for the first batch.

Inductive hypothesis: Assume that after k batches, the budget satisfies:

$$nB - B_1 - B_2 - \cdots - B_k \geq (n - n_1 - n_2 - \cdots - n_k)B.$$

Inductive step: For the $(k + 1)$-th batch, we have the remaining budget:

$$nB - B_1 - B_2 - \cdots - B_{k+1} = nB - B_2 - \cdots - B_{k+1}$$

Since for each B_i, $B_i \leq n_i B + \sqrt{n_i}B$ (proved below in 1), then we can bound the above by:

$$nB - B_2 - \cdots - B_{k+1} \geq nB - n_2 B - \sqrt{n_2}B - \cdots - n_{k+1}B - \sqrt{n_{k+1}}B.$$

Since $\sum_{i=2}^{k+1} \sqrt{n_i}B \leq n_1 B = c\sqrt{n}B$ for large enough n_1 (proved below in 2):

$$nB - n_2 B - \sqrt{n_2}B - \cdots - n_{k+1}B - \sqrt{n_{k+1}}B \geq (n - n_1 - n_2 - \cdots - n_{k+1})B.$$

Thus, the inductive step holds.

By induction, after k batches, the remaining budget satisfies:

$$nB - B_1 - B_2 - \cdots - B_k \geq (n - n_1 - n_2 - \cdots - n_k)B,$$

which ensures that the total spending will always remain within the budget nB.

1. Proof of $B_i \leq B_i^* \leq n_i B + \sqrt{n_i}B$

In lemma 5, we have shown that the empirical distribution $\hat{p}_X$ approximates the true distribution p_X at batch i. Then $E_{\hat{p}_X}[p^*(X_{ij})] = B$. Since each X_i are drawn i.i.d. and policy $p^*(X_{ij})$ is fixed within this batch, $p^*(X_{ij})$ is also i.i.d. Then we can apply the Law of Large Numbers, as the em-

pirical average converges to the expected spending:

$$B_i^* = \sum_{j=1}^{n_i} p^*(X_{ij}) \le \left(1 + \frac{1}{\sqrt{n_i}}\right) n_i B = n_i B + \sqrt{n_i} B$$

By our design $B_i \le B_i^*$, thus $B_i \le B_i^* \le n_i B + \sqrt{n_i} B$.

2. Proof of $\sum_{i=2}^{k+1} \sqrt{n_i} B \le n_1 B = c\sqrt{n} B$ for large enough n_1
The total contribution of $\sqrt{n_i} B$ over batches $i = 2$ to k can be bounded by $\sum_{i=2}^{k} \sqrt{n_i}$..

Since the total number of batches is $\log(n)$, and $\sqrt{n_i} \le \sqrt{n}$, we can write:

$$\sum_{i=2}^{k} \sqrt{n_i} B \le \log(n)\sqrt{n} B.$$

For the constraint $\log(n)\sqrt{n} \le c\sqrt{n}$, we choose $c = M \log(n)$, where M is a sufficiently large constant. This ensures:

$$\sum_{i=2}^{k} \sqrt{n_i} B \le n_1 B = c\sqrt{n} B.$$

Proof of Lemma 2

The bias of AIPW estimator is

$$\begin{aligned}
E[\widehat{Y}_1] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mu^{(1)}(X_i) + \frac{Y_i - \mu^{(1)}(X_i)}{e(X_i)} W_i\right)\right] \\
&= E\left[\mu^{(1)}(X) + \frac{Y - \mu^{(1)}(X)}{e(X)} W\right] \\
&= E\left[\frac{YW}{e(X)}\right] \\
&= E_X\left[E[Y|W=1]\frac{P(W=1|X)}{e(X)} + 0\frac{P(W=0|X)}{e(X)}\right] \\
&= E_X\left[E[\tilde{f}_1(X) + g_1(X,\alpha)|\alpha + \texttt{gap}(X) > \beta]\right] \\
&= E_X\left[E[g_1(X,\alpha)|\alpha + \texttt{gap}(X) > \beta]\right] + E_X[\tilde{f}_1(X)] \\
&= E_X\left[E[g_1(X,\alpha)|\alpha + \texttt{gap}(X) > \beta]\right] + E(Y_1).
\end{aligned} \tag{12}$$

The second equality comes from i.i.d. data generating process. The fourth one is by definition of expectation, and the fifth one is given by the characterization of random choice model $P(W = 1|X)$. The sixth holds since $\tilde{f}_1(X)$ is independent of α and β. Therefore, we have proved the bias of treatment group. Similar analysis holds for control group. The last claim holds due to the symmetry of α and β and g_1, g_0.

Proof of Lemma 3

We will prove in lemma 4 that for each fix X, we have

$$\begin{aligned}
&E[g(X,\alpha)|\alpha + \texttt{gap}(X) > \beta] - E[g(X,\alpha)|\alpha - \texttt{gap}(X) > \beta] \\
&= \frac{\int \big(g(X, \alpha + \texttt{gap}(X)) - g(X,\alpha)\big) f_1(\alpha)\, d\alpha}{\int f_1(\alpha)\, d\alpha}, \\
&\text{where} \quad f_1(\alpha) = \exp\big(-\exp(-(\alpha + \texttt{gap}(X)))\big) \\
&\quad \times \exp\big(-(\alpha + \exp(-\alpha))\big).
\end{aligned} \tag{13}$$

Since we also have $E(g(X,\alpha)) = 0$, we know that $E[g(X,\alpha)|\alpha + \texttt{gap}(X) > \beta] \le \frac{\varepsilon_X \texttt{gap}(X)}{1+\exp(\texttt{gap}(X))}$. Similarly we can prove that $E[g(X,\alpha)|\alpha - \texttt{gap}(X) > \beta] \ge -\frac{\varepsilon_X \texttt{gap}(X)\exp(\texttt{gap}(X))}{1+\exp(\texttt{gap}(X))}$. Therefore, we have for any X, $E[\widehat{Y}_1(X) - \widehat{Y}_0(X) - (Y_1(X) - Y_0(X))] \le \varepsilon_X \texttt{gap}(X)$. The result holds by taking expectation with respect to X.

Proof of Lemma 4

By definition of Gumbel distribution, we have

$$\begin{aligned}
\text{bias}(X) &= E[g(X,\alpha)|\alpha + \texttt{gap}(X) > \beta] \\
&\quad - E[g(X,\alpha)|\alpha - \texttt{gap}(X) > \beta] \\
&= \frac{\int \big(g(X, \alpha + \texttt{gap}(X)) - g(X,\alpha)\big) f_1(\alpha)\, d\alpha}{\int f_1(\alpha)\, d\alpha}, \\
\text{where} \quad f_1(\alpha) &= \exp\big(-\exp(-(\alpha + \texttt{gap}(X)))\big) \\
&\quad \times \exp\big(-(\alpha + \exp(-\alpha))\big).
\end{aligned} \tag{14}$$

take $g(X,\alpha) = \varepsilon\alpha - E(\varepsilon\alpha)$, we know that it equals ε.

Proof of Theorem 3

In theorem 4 of (Li et al. (2024)), it is proved that for any possible adaptive algorithm ALG and any unbiased estimator $\hat{\tau}$, the variance

$$\begin{aligned}
\text{Var}(\hat{\tau}) \ge E_{P_X}\Bigg(&\big(\mu^{(1)}(x) - \mu^{(0)}(x) - E_X\big[\mu^{(1)}(x) - \mu^{(0)}(x)\big]\big)^2 \\
&+ \frac{\sigma^{(1)}(x)^2}{e^{\widehat{\text{ALG}}}(x)} + \frac{\sigma^{(0)}(x)^2}{1 - e^{\widehat{\text{ALG}}}(x)}\Bigg).
\end{aligned}$$

where $\widehat{\text{ALG}}$ is defined as

$$\begin{aligned}
e^{\widehat{ALG}}(X) &= P^{\widehat{ALG}}(W = 1 \mid X) \\
&= \frac{1}{n}\sum_{t=1}^{n} \int P(W = 1 \mid X, \mathcal{F}_{t-1})\, dP^{ALG}(\mathcal{F}_{t-1}).
\end{aligned}$$

and $\mathcal{F}_{t-1}$ is the filtration up to time $t-1$. By convexity, one can check that the pricing strategy $p^{\widehat{ALG}}(X)$ which gives the corresponding $e^{\widehat{ALG}}$ satisfies the budget constraint:

$$E_X(p^{\widehat{ALG}}(X)) \le \frac{B}{n}.$$

Therefore, we know that the variance given by applying $p^{\widehat{ALG}}$ throughout the n periods, which gives the variance of $\hat{\tau}$, is no smaller than $\frac{1}{n} v^*(\frac{B}{n})$.

Proof of Theorem 4

Following the proof in Li et al. (2024), we will define the following three estimators. We will use the AIPW estimator to estimate the treatment effect, which is defined as

$$\begin{aligned}\widehat{\tau}_1^X = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\widehat{\mu}^{(1)}(X_i) - \widehat{\mu}^{(0)}(X_i) \\ &+ \frac{Y_i - \widehat{\mu}^{(1)}(X_i)}{{}_i(X_i, p(X_i))}W_i \\ &- \frac{Y_i - \widehat{\mu}^{(0)}(X_i)}{1 -{}_i\,(X_i, p(X_i))}(1-W_i)\Bigg).\end{aligned} \tag{15}$$

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

$$\begin{aligned}\widehat{\tau}_2^X = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\mu^{(1)}(X_i) - \mu^{(0)}(X_i) \\ &+ \frac{Y_i - \mu^{(1)}(X_i)}{e_i(X_i, p(X_i))}W_i \\ &- \frac{Y_i - \mu^{(0)}(X_i)}{1 - e_i(X_i, p(X_i))}(1-W_i)\Bigg).\end{aligned} \tag{16}$$

$$\begin{aligned}\widehat{\tau}^{*X} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\mu^{(1)}(X_i) - \mu^{(0)}(X_i) \\ &+ \frac{Y_i - \mu^{(1)}(X_i)}{e^*(X_i, p^*(X_i))}W_i \\ &- \frac{Y_i - \mu^{(0)}(X_i)}{1 - e^*(X_i, p^*(X_i))}(1-W_i)\Bigg).\end{aligned} \tag{17}$$

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:

$$\begin{aligned}&E\left(\widehat{\tau}_1^X - \tau\right)^2 - E\left(\widehat{\tau}^{*X} - \tau\right)^2 \\ &= E\left(\widehat{\tau}_1^X - \widehat{\tau}_2^X + \widehat{\tau}_2^X - \tau\right)^2 - E\left(\widehat{\tau}^{*X} - \tau\right)^2 \\ &= \underbrace{E\left(\widehat{\tau}_1^X - \widehat{\tau}_2^X\right)^2}_{\text{model estimation}} \\ &\quad + \underbrace{E\left(\widehat{\tau}_2^X - \tau\right)^2 - E\left(\widehat{\tau}^{*X} - \tau\right)^2}_{\text{propensity score optimization}} \\ &\quad + \underbrace{2E\left((\widehat{\tau}_1^X - \widehat{\tau}_2^X)(\widehat{\tau}_2^X - \tau)\right)}_{\text{cross-term}}.\end{aligned} \tag{18}$$

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

Lemma 8.

$$E\left(\widehat{\tau}_2^X - \tau\right)^2 - E\left(\widehat{\tau}^{*X} - \tau\right)^2 \le O(\frac{1}{\sqrt{n}})E\left(\widehat{\tau}^{*X} - \tau\right)^2 \tag{19}$$

Proof of Lemma 8

As proved in lemma 3 in (Li et al. (2024)), we have

$$\begin{aligned}&E\left(\widehat{\tau}_2^X - \tau\right)^2 - E\left(\widehat{\tau}^{*X} - \tau\right)^2 \\ &= \frac{1}{n^2}E\Bigg(\sum_{i=1}^{n}\frac{\sigma^2}{e_i(X_i)} + \sum_{i=1}^{n}\frac{\sigma^2}{1-e_i(X_i)} \\ &\quad - n\left(\frac{\sigma^2}{e^*(X_i)} + \frac{\sigma^2}{1-e^*(X_i)}\right)\Bigg).\end{aligned} \tag{20}$$

As proved in lemma 7, we have in batch k, $\|e_i(X_i, p(X_i)) - e^*(X_i, p^*(X_i))\| \le O(\frac{1}{\sqrt{n_{k-1}}})$. Therefore, the total regret in (20) can be bounded as

$$\begin{aligned}&\frac{1}{n^2}E\Bigg(\sum_{i=1}^{n}\frac{\sigma^2}{e_i(X_i)} + \sum_{i=1}^{n}\frac{\sigma^2}{1-e_i(X_i)} \\ &\quad - n\left(\frac{\sigma^2}{e^*(X_i)} + \frac{\sigma^2}{1-e^*(X_i)}\right)\Bigg) \\ &\le \frac{1}{n^2}E\Bigg(\sum_{k=1}^{O(\log(n))}O\Bigg(\frac{\sigma^2}{e^*(X_i, p^*(X_i))} \\ &\quad \times \frac{1}{1-e^*(X_i, p^*(X_i))}\Bigg)n_k\frac{1}{\sqrt{n_{k-1}}}\Bigg) \\ &\le \widetilde{O}\left(\frac{1}{\sqrt{n}}\right)E\left(\widehat{\tau}^{*X} - \tau\right)^2.\end{aligned} \tag{21}$$

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

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

Lemma 9. $E\left(\tau_1^X - \tau_2^X\right)^2 \le \widetilde{O}(n^{-(1+\alpha)}) \le \widetilde{O}(n^{-\alpha})v^*$.

Proof of Lemma 9

We will follow the proof in Li et al. (2024) and also Wager (2024). 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 switching learning-to-incentivize algorithm (1), then uses an estimator

$$\begin{aligned}\hat{\tau}_1^X &= \frac{|\mathcal{I}_1|}{n}\hat{\tau}_1^{\mathcal{I}_1,X} + \frac{|\mathcal{I}_2|}{n}\hat{\tau}_1^{\mathcal{I}_2,X}, \\ \hat{\tau}^{\mathcal{I}_1} &= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\Bigg(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \hat{\mu}_{(0)}^{\mathcal{I}_2}(X_i) \\ &+ W_i\frac{Y_i - \hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}^{\mathcal{I}_2}(X_i)}{1-\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))}\Bigg),\end{aligned} \tag{22}$$

where in batch k in $_1$, the $\hat{\mu}^{\mathcal{I}_2}_{(w)}(\cdot)$ 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

$$\hat{\tau}^X_2 = \frac{|\mathcal{I}_1|}{n}\hat{\tau}^{\mathcal{I}_1,X}_2 + \frac{|\mathcal{I}_2|}{n}\hat{\tau}^{\mathcal{I}_2,X}_2 \tag{23}$$

analogously to (22) (because $\hat{\tau}^X_2$ uses oracle nuisance components, the crossfitting construction doesn't change anything for it). Moreover, we can decompose $\hat{\tau}^{\mathcal{I}_1}$ itself as $\hat{\tau}^{\mathcal{I}_1,X}_1 = \hat{Y}^{\mathcal{I}_1,1}_{(1)} - \hat{Y}^{\mathcal{I}_1}_{(0)}$

$$\hat{Y}^{\mathcal{I}_1,1}_{(1)} = \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) + W_i\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))}\right) \text{ etc.,}$$

and define $\hat{Y}^{\mathcal{I}_1,2}_{(0)}$ and $\hat{Y}^{\mathcal{I}_1,2}_{(1)}$ analogously. Given this buildup, in order to verify lemma (9), it suffices to show that

$$E\left(\hat{Y}^{\mathcal{I}_1,1}_{(1)} - \hat{Y}^{\mathcal{I}_1,2}_{(1)}\right)^2 \leq \widetilde{O}(n^{-1+\alpha}). \tag{24}$$

etc., across folds and treatment statuses. We now study the term in (24) by decomposing it as follows:

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

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, $\hat{\mu}^{\mathcal{I}_2}_{(w)}$ 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).

$$\begin{aligned}
&E\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right)^2\right] \\
&= E\left[E\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right)^2\middle|\mathcal{I}_2\right]\right] \\
&= E\Bigg[\frac{1}{|\mathcal{I}_1|^2}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)^2\left(1 - \frac{W_i}{e(X_i)}\right)^2 \\
&\qquad + \sum_{i,j\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right) \\
&\qquad\quad \times\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_j) - \mu_{(1)}(X_j)\right)\left(1 - \frac{W_j}{e(X_j)}\right)\Bigg|\mathcal{I}_2\Bigg]\Bigg] \\
&= \frac{1}{|\mathcal{I}_1|^2}E\left[\sum_{i\in\mathcal{I}_1}\mathrm{Var}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\middle|\mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|^2}E\left[\sum_{i\in\mathcal{I}_1}E\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)^2\left(\frac{1}{e(X_i)} - 1\right)\middle|\mathcal{I}_2\right]\right] \\
&\leq \frac{1}{\eta\,|\mathcal{I}_1|^2}E\left[\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)^2\right] \\
&= \frac{\tilde{O}(1)}{n^{1+\alpha}}.
\end{aligned} \tag{25}$$

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

$$\begin{aligned}
E\left[\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)^2\right] &\leq \sum_{k=1}^{O(\log n)} n_k\, O(n^{-\alpha}_{k-1}) \\
&\leq \widetilde{O}(n^{1-\alpha}).
\end{aligned} \tag{26}$$

Similarly, we can prove that the second term

$$\begin{aligned}
&E\Bigg[\Bigg(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} W_i\Bigg(\left(Y_i - \mu_{(1)}(X_i)\right) \\
&\qquad \times\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i, p(X_i))} - \frac{1}{e(X_i, p(X_i))}\right)\Bigg)\Bigg)^2\Bigg] \\
&\leq \frac{\widetilde{O}(1)}{n^2}.
\end{aligned} \tag{27}$$

Now for the third term, we have

$$
\begin{aligned}
&E\Bigg[\Bigg(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} W_i\Bigg(\big(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\big)\\
&\qquad\times\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}-\frac{1}{e(X_i,p(X_i))}\right)\Bigg)\Bigg)^2\Bigg]\\
&\le \frac{O(\log n)^2}{|\mathcal{I}_1|^2}\sum_{k=1}^{O(\log n)} E\Bigg[\Bigg(\sum_{i\in\mathcal{I}_1,\text{ batch }k}\big(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\big)\\
&\qquad\times\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}-\frac{1}{e(X_i,p(X_i))}\right)\Bigg)^2\Bigg].
\end{aligned}
\tag{28}
$$

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

$$
\begin{aligned}
&E\left[\sum_{\{i:i\in\mathcal{I}_1,\text{ batch k}\}}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)\right)\right]^2\\
&\le E\left[\sum_{\{i:i\in\mathcal{I}_1,\text{ batch k}\}}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)^2\right]\\
&\quad\times E\left[\sum_{\{i:i\in\mathcal{I}_1,\text{ batch k}\}}\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)^2\right]=\widetilde{O}_P\left(n_k^{1-\alpha}\right).
\end{aligned}
\tag{29}
$$

Therefore we know that

$$
\begin{aligned}
&\frac{O(\log n)^2}{|\mathcal{I}_1|^2}\sum_{k=1}^{O(\log n)} E\Bigg[\Bigg(\sum_{i\in\mathcal{I}_1,\text{ batch }k}\big(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\big)\\
&\qquad\times\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i,p(X_i))}-\frac{1}{e(X_i,p(X_i))}\right)\Bigg)^2\Bigg]\\
&\le\frac{O(\text{polylog}(n))}{n^{1+\alpha}}.
\end{aligned}
\tag{30}
$$

Combining everything together, we have the conclusion in (24). □

Finally, since we prove that $E\left(\widehat{\tau}^X_2-\tau\right)^2 \le (1+O(\frac{1}{\sqrt{n}}))v^*$, and $E\left(\widehat{\tau}^X_1-\widehat{\tau}^X_2\right)^2\le O(n^{-\alpha})v^*$, by Cauchy-Schwarz, we can bound the cross term as

$$E\left((\widehat{\tau}^X_1-\widehat{\tau}^X_2)(\widehat{\tau}^X_2-\tau)\right)\le O(n^{-\frac{\alpha}{2}})v^*.$$

And this completes the proof of theorem 4.