Pessimistic Data Integration for Policy Evaluation

Xiangkun Wu[*]
School of Mathematical Sciences
Zhejiang University
Hangzhou, China
12235031@zju.edu.cn

Ting Li[*]
School of Statistics and Data Science
Shanghai University of Finance and Economics
Shanghai, China
tingli@mail.shufe.edu.cn

Gholamali Aminian
The Alan Turing Institute
London, UK
gaminian@turing.ac.uk

Armin Behnamnia
Sharif University of Technology
Tehran, Iran
arminbehnamnia@gmail.com

Hamid R. Rabiee
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
rabiee@sharif.edu

Chengchun Shi [†]
Department of Statistics
London School of Economics and Political Science
London, UK
C.Shi7@lse.ac.uk

Abstract

This paper studies how to integrate historical control data with experimental data to enhance A/B testing, while addressing the distributional shift between historical and experimental datasets. We propose a pessimistic data integration method that combines two causal effect estimators constructed based on experimental and historical datasets. Our main idea is to conceptualize the weight function for this combination as a policy so that existing pessimistic policy learning algorithms are applicable to learn the optimal weight that minimizes the resulting weighted estimator's mean squared error. Additionally, we conduct comprehensive theoretical and empirical analyses to compare our method against various baseline estimators across five scenarios. Both our theoretical and numerical findings demonstrate that the proposed estimator achieves near-optimal performance across all scenarios.

1 Introduction

A/B testing is widely used by various technology companies such as Amazon, Google, Netflix, Uber, and Didi to evaluate the performance of new products, policies or treatments compared to existing controls. However, the effectiveness of such evaluations is often limited by short duration of online experiments. For instance, in ridesharing, most experiments last no more than two weeks [1]. Before conducting these experiments, companies usually have access to a substantial amount of historical data collected under the control policy. Recent work has demonstrated that integrating these historical control data with experimental data can largely improve the efficiency of A/B testing [2].

The primary challenge in data integration stems from the distributional shift between historical and experimental data, which can generally be categorized into three types: (i) covariate shift – the changes in the distribution of contextual covariates; (ii) policy shift – the changes in the behavior

[*]Equal contribution.
[†]Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

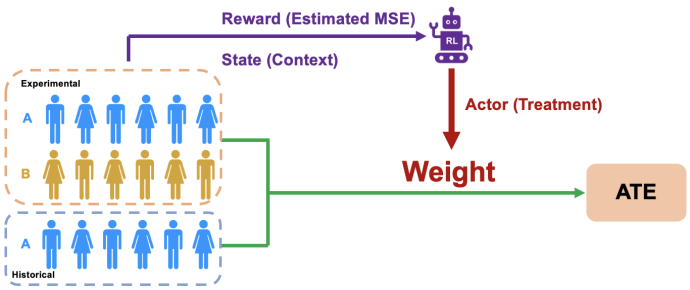

Figure 1: Workflow of the proposed estimator.

policy or propensity score; and (iii) posterior shift – the changes in the outcome distribution given covariates and treatment. Such distributional shifts can substantially bias the resulting treatment effect estimator and hinder the effective use of historical data for A/B testing.

This paper studies the integration of historical data to improve A/B testing, while addressing all three types of distributional shifts simultaneously. Our contributions are as follows.

Methodologically, we propose a weighted average treatment effect (ATE) estimator that optimally combines information from both experimental and historical datasets. Our main idea is the development of a pessimistic data integration approach that conceptualizes the weight function for data combination as a policy, which enables existing pessimistic policy learning algorithms to learn this optimal weight function; see Figure 1 for an illustration of our methodology.

Theoretically, we conduct a comprehensive comparative analysis to compare our proposed weighted ATE estimator against five baseline estimators across five different scenarios, reflecting differing degrees of posterior shift between experimental and historical datasets, and different levels of heavy-tailedness in reward residuals. Our analysis reveals that while each baseline estimator may perform optimally under certain scenarios, they often fail in others. In contrast, the proposed estimator is adaptive: it satisfies the *oracle* property across all scenarios, meaning that it achieves comparable performance to the optimal scenario-specific estimator, working effectively as if it knew the optimal weight function. Our theories are further supported by synthetic and real-world data analyses. Table 1 summarizes our theoretical and numerical findings, with detailed descriptions of the baseline estimators and scenarios provided in Sections 3.2 and 4.

2 Related work

Our paper is closely related to data integration, the pessimistic principle in offline policy learning and off-policy evaluation (OPE), which we elaborate below.

Data integration. Data integration is related to various fields in statistics and machine learning, ranging from the classical meta-analysis [3], to more recent advancements in transfer learning [4], and their applications to numerous downstream tasks such as multi-task learning [5], multimodal learning [6, 7], fusion learning [8], individualized treatment regime estimation [9, 10], and RL [11].

Our proposal is particularly related to those methods tailored for causal inference; see [12] and [13] for recent reviews. Based on their approach to handling distributional shifts, these methods can roughly be classified into two categories:

1. The first category addresses the covariate and policy shifts under the assumption of no posterior shift [14–19]. A primary example is given by Li et al. [20], who integrated experimental data with historical controls and developed an estimator that achieves the asymptotically smallest MSE.
2. The second category handles all three types of distributional shifts simultaneously [21–24]. Among these methods, a notable subset applied an ℓ_1-type penalty function for selecting external data [25, 26]. Recent advancements include federated causal inference approaches [27, 28], which employed such penalization to estimate ATE in a manner that preserves privacy.

Our approach differs from both categories in the following ways. Methodologically, we overcome several critical limitations of the aforementioned methods, resulting in a more flexible and practical

Table 1: MSEs of different ATE estimators across five scenarios. Green indicates that the estimator achieves the oracle property (its MSE is asymptotically equivalent to that of the optimal estimator). Yellow indicates that the estimator may generally have a high MSE but can attain the oracle property in some special cases. Red indicates that the estimator exhibits a generally large MSE.

Scenarios	EDO	HDB	MVE	CWE	NonPessi	Proposed
(i) Heavy-tailed historical rewards						
(ii) Heavy-tailed experimental rewards						
(iii)Small posterior shifts						
(iv) Moderate posterior shifts						
(v) Large posterior shifts						

solution. First, unlike the first category of methods which assumes no posterior shift, our proposal accounts for this shift. Second, whereas the second category of methods requires the posterior shift to be either zero or sufficiently large for consistent data selection – failing in intermediate regimes [2] – our method remains robust even with moderate posterior shifts (see Corollary 4).

In terms of applications, our work focuses on combining experimental and historical data for A/B testing, whereas most works either integrate data from multiple treatment centers or trials for meta analysis [29–32], or combine RCT and observational data to handle unmeasured confounding [33–35].

Pessimistic policy learning. The pessimistic principle is fundamental to most existing offline policy learning algorithms, which aim to learn an optimal policy from a pre-collected historical dataset. This principle originates from the seminal works of Swaminathan and Joachims [36, 37], who proposed a counterfactual risk minimizing approach that incorporates the uncertainty of a policy's value estimator as a penalty term to learn policies with lower-variance value estimates. It has been widely employed in contextual bandits [38–41], dynamic treatment regimes [42], RL [43–53], and more recently in the training of large language models [see e.g., 54–56] to prevent value function overestimation and encourage the learning policy to stay close to the behavior policy.

A recent proposal by Li et al. [2] applied the pessimistic principle to data integration for policy evaluation. In particular, they proposed to linearly combine policy value estimators computed from experimental and historical datasets using a weighted average. While their approach is closely related to ours, a key difference lies in their use of a fixed weight function for data integration. In contrast, our approach employs a covariate-dependent weighting, leading to a more accurate estimator than theirs, as demonstrated analytically in Sections 4 and empirically in Section 5.

Off-policy evaluation. There is a huge literature on OPE in bandits and RL; see [57] and [58] for reviews. The goal of OPE is to estimate the expected outcome of a target policy using offline data collected under a potentially different behavior policy. Existing approaches can be classified into four main categories: (i) Model-based methods that estimate a dynamic model (e.g., a Markov decision process) from offline data and compute the target policy's expected outcome via dynamic programming or Monte Carlo [59–62]; (ii) Direct methods that estimate the expected outcome by learning either a reward or value function from the offline data [63–73]; (iii) Importance sampling (IS) methods that reweight observed rewards using the IS ratio (the density ratio between target and behavior policies) [74–84]; (iv) Doubly or multiply robust methods that combine the estimated reward or value function from direct methods with the IS ratio from IS methods, and require only the reward/value function or the IS ratio to be consistent [85–99].

Many of these approaches have been recently adopted for A/B testing [see e.g., 100–111]. However, these works rely solely on experimental datasets, without leveraging historical datasets to improve policy evaluation.

More recently, [112] proposed a doubly robust (DR) estimator by combining data from multiple experimental studies. Their approach requires the target outcome and covariate distribution to match those in at least one experimental dataset. We also notice that there is an emerging line of research that studied how to improve OPE estimation by strategically combining multiple base OPE estimators to leverage their strengths [see e.g., 38, 113–115]. While similar in spirit to our combination of OPE estimators from experimental and historical datasets, the base estimators in these papers were derived from a single dataset (thus avoiding distributional shifts). On the contrary, we need to address the challenges posed by the distributional shift between experimental and historical datasets.

3 Pessimistic data integration

We first introduce two baseline OPE estimators (formally defined in Equations (1) and (2)). We next introduce our proposed estimator, which builds upon these baseline estimators.

3.1 Two baseline estimators

Suppose we are given an experimental dataset $\mathcal{D}^{(e)}$ and a historical dataset $\mathcal{D}^{(h)}$. During the experiment, at each time t, the decision maker observes certain contextual covariates (e.g., market features), and assigns an action between a new treatment strategy (denoted by 1) and a baseline control (0), resulting in a reward that measures the company's profit at that time. Thus, the experimental data can be summarized as a set of context-action-reward $O^{(e)} = (S^{(e)}, A^{(e)}, R^{(e)})$ triplets, which are assumed to be i.i.d. over time. Similarly, the historical dataset consists of another set of i.i.d. triplets $O^{(h)} = (S^{(h)}, A^{(h)}, R^{(h)})$, but differs in distribution from $\mathcal{D}^{(e)}$ in the following three aspects:

1. **Covariate shift**: the probability mass function of $S^{(e)}$ (denoted by p_e) might differ from that of S_h (denoted by p_h), leading to the IS ratio $\mu(s) = p_e(s)/p_h(s)$ generally deviating from 1.
2. **Policy shift**: actions in the historical dataset are exclusively generated under the control policy such that $A^{(h)} = 0$ almost surely, whereas actions in the experimental dataset are generated under both the control and the treatment for A/B testing.
3. **Posterior shift**: the reward function $r^{(e)}(a, s) = \mathbb{E}(R^{(e)}|A^{(e)} = a, S^{(e)} = s)$ in the experimental dataset might differ from that in the historical dataset (denoted by $r^{(h)}$).

Our objective lies in estimating the ATE – the difference between the expected reward under the treatment and that under the control, i.e.,

$$\text{ATE} = \mathbb{E}[r^{(e)}(1, S^{(e)}) - r^{(e)}(0, S^{(e)})],$$

using both experimental and historical datasets.

The first baseline estimator for ATE we introduce is the experimental-data-only (EDO) estimator, which uses exclusively the experimental dataset $\mathcal{D}^{(e)}$ to learn the ATE. This estimator is simple to describe: we construct two OPE estimators using $\mathcal{D}^{(e)}$ to estimate the expected outcomes under treatment 1 and 0, respectively, and compute their difference to obtain the ATE estimator. Notice that any OPE method discussed in Section 2 can be applied for estimation.

As a concrete example, consider the IS estimator with the estimating function $\psi_a^{(e)}(O^{(e)}) = \mathbb{I}(A = a)R^{(e)}/\pi(a|S^{(e)})$ where $\mathbb{I}(A = a)/\pi(a|S)$ denotes the IS ratio of the target policy over the behavior policy (i.e., propensity score) π in the experimental data. Using the change of measure theorem, it can be shown that $\psi_a^{(e)}(O^{(e)})$ is unbiased to the expected outcome under treatment a. This motivates the use of $\mathbb{E}_n[\psi_a^{(e)}(O^{(e)})]$ to estimate this expected outcome, leading to the following EDO estimator for the ATE,

$$\text{EDO} = \mathbb{E}_n[\psi_1^{(e)}(O^{(e)})] - \mathbb{E}_n[\psi_0^{(e)}(O^{(e)})], \tag{1}$$

where $\mathbb{E}_n$ denotes the empirical average over the offline dataset.

The second baseline estimator is the historical-data-based (HDB) estimator. Similar to EDO, it uses $\mathbb{E}_n[\psi_1^{(e)}(O^{(e)})]$ to estimate the expected outcome under the new treatment. For the control policy, the corresponding estimator $\mathbb{E}_n[\psi_0^{(h)}(O^{(h)})]$ with $\psi_0^{(h)}(O^{(h)}) = \mu(S^{(h)})R^{(h)}$ is constructed using solely the historical data. Here, the IS ratio $\mu(\cdot)$ denotes the density ratio of the probability mass/density function of $S^{(e)}$ over that of $S^{(e)}$. It depends only on the contextual variable, since the historical data is exclusively generated under the control policy, leading to a propensity score $\pi(A^{(h)}|S^{(h)})$ of 1 almost surely. Similarly, it can be shown that this estimator is unbiased to $\mathbb{E}[r^{(h)}(0, S^{(h)})]$. In summary, we have

$$\text{HDB} = \mathbb{E}_n[\psi_1^{(e)}(O^{(e)})] - \mathbb{E}_n[\psi^{(h)}(O^{(h)})]. \tag{2}$$

To conclude this section, we remark that there is a bias-variance trade-off between the two estimators. Specifically, the HDB estimator is generally biased due to the incorporation of historical data.

Table 2: A numerical example demonstrating the bias–variance trade-off (see Appendix A for details). As shown, EDO achieves the lowest bias, while HDB attains the lowest variance. The proposed estimator strikes a balance between bias and variance, resulting in the lowest overall MSE.

Method	**MSE(95% CI)**	**Bias(95% CI)**	**Variance(95% CI)**
EDO	1.701 (1.598–1.804)	**0.007** (-0.064–0.051)	1.701 (1.598–1.804)
HDB	2.372 (2.289–2.455)	1.400 (1.372–1.428)	**0.413** (0.388–0.438)
Proposed	**1.394** (1.312–1.476)	0.221 (0.170–0.272)	1.345 (1.262–1.428)

Although it addresses the covariate shift through the use of the IS ratio μ, the posterior shift from the experimental data is extremely challenging to correct, resulting in a non-negligible bias. In contrast, the EDO estimator, derived exclusively from the experimental data, remains asymptotically unbiased. On the other hand, HDB typically achieves lower variance by leveraging the historical data, which usually has a much larger sample size than the experimental data. Finally, our proposed estimator, which we introduce in the following section, effectively strikes a balance between bias and variance and outperforms both baseline estimators; see Table 2 for an illustration.

3.2 A pessimistic estimator for data integration

We begin with a summary of our proposal. Our approach is to linearly combine the two baseline estimators presented in Section 3.1 for data integration, while taking into account the posterior shift between the experimental and historical data. The key here is to determine the optimal weight (see e.g., Equation (3) below for the definition) for data combination. Our main idea is to transform this weight selection problem into offline policy learning. Specifically, we conceptualize the choice of weight as an 'action', which could vary as a function of the contextual information. This conceptualization effectively frames the weight selection as a policy learning problem where the goal is to identify an optimal policy that maximizes reward or minimizes cost, the latter of which corresponds to the MSE of the weighted ATE estimator. Figure 1 gives an overview of the proposed estimator pipeline. Adopting this perspective enables us to employ state-of-the-art pessimistic policy learning algorithms such as counterfactual risk minimization to effectively determine the weight.

We next detail our methodology. Similar to EDO and HDB, our estimator employs $\mathbb{E}_n[\psi_1^{(e)}(O^{(e)})]$ to estimate the mean outcome under the treatment policy. As for the control, it uses a weight function $w(s)$ to linearly combine the estimating functions used in EDO and HDB. Specifically, we define the following estimating function,

$$\psi_w(O^{(e)}, O^{(h)}) = w(S^{(e)})\psi_0^{(e)}(O^{(e)}) + [1 - w(S^{(h)})]\psi^{(h)}(O^{(h)}). \tag{3}$$

It is immediate to see that setting $w = 1$ recovers EDO's estimating function $\psi_0^{(e)}$ whereas setting $w = 0$ recovers the HDB's estimating function $\psi^{(h)}$. This leads to the following weighted ATE estimator,

$$\widehat{\mathrm{ATE}}(w) = \mathbb{E}_n[\psi_1^{(e)}(O^{(e)})] - \mathbb{E}_n[\psi_w(O^{(e)}, O^{(h)})], \tag{4}$$

where the second empirical average $\mathbb{E}_n$ is taken over all pairs of $(O^{(e)}, O^{(h)}) \in \mathcal{D}^{(e)} \times \mathcal{D}^{(h)}$.

It remains to identify the optimal weight function w^* that optimally balances the bias and variance of the ATE estimator to minimize its MSE. As mentioned earlier, we adopt an offline policy learning framework and view each value of w – bounded between 0 and 1 – as an arm in a contextual bandit model. Given that w is a function of the covariates, it defines a policy on this contextual space. The identification of w^* is thus equivalent to the identification of the optimal policy that minimizes the cost, which in our case equals the MSE.

Although the oracle MSE is unknown, it can be estimated from the offline data. Specifically, since EDO is derived solely from $\mathcal{D}_e$, it is expected to be asymptotically unbiased. Thus, its deviation from $\widehat{\mathrm{ATE}}(w)$ (i.e., $\widehat{\mathrm{bias}}(w) = \widehat{\mathrm{ATE}}(w) - \mathrm{EDO}$) can be used to measure $\widehat{\mathrm{ATE}}(w)$'s bias (denoted by $\mathrm{bias}(w)$). Additionally, its variance (denoted by $\mathrm{Var}(w)$) can be estimated using the sampling variance formula (see Appendix B.3 for details). Denote the resulting variance estimator by $\widehat{\mathrm{Var}}(w)$. Given a parametric function class $\mathcal{W}$, the optimal w^* can be estimated by minimizing

$$\widehat{\mathrm{MSE}}(w) = \widehat{\mathrm{bias}}^2(w) + \widehat{\mathrm{Var}}(w), \tag{5}$$

over $w \in \mathcal{W}$. Following the pessimistic principle, we instead minimize an upper bound of the estimated MSE given by

$$\widehat{\mathrm{MSE}}_U(w) = \widehat{\mathrm{bias}}_U^2(w) + \widehat{\mathrm{Var}}_U(w), \tag{6}$$

where $\widehat{\mathrm{bias}}_U$ and $\widehat{\mathrm{Var}}_U$ are required to upper bound the oracle bias and variance so that the following assumption is satisfied.

Assumption 1 (Coverage probability)**.** *Assume* $\mathbb{P}(\cap_{w\in\mathcal{W}}\{\widehat{\mathrm{bias}}_U(w) \geq |\mathrm{bias}(w)|\}) \geq 1-\alpha$ *and* $\mathbb{P}(\cap_{w\in\mathcal{W}}\{\widehat{\mathrm{Var}}_U(w) \geq \mathrm{Var}(w)\}) \geq 1-\alpha$ *for some* $0 < \alpha < 1$.

In practice, $\widehat{\mathrm{bias}}_U$ and $\widehat{\mathrm{Var}}_U$ can be constructed using concentration inequalities [116] or Wald-type confidence intervals [117]. We detail our implementation in Appendix B.4. We also remark that for clarity of presentation, we focus on IS estimators for ATE estimation in this section. However, our actual implementation employs DR estimators, which are known to be more efficient than IS with well-specified reward models [118]. The detailed formulas are relegated to Appendix B.2

Let $\widehat{w}$ denote the minimizer of (6), which yields our proposed estimator $\widehat{\mathrm{ATE}}(\widehat{w})$. To conclude, we remark that our framework unifies several baseline estimators through specific choices of $\widehat{w}$:

1. **EDO**: Setting $\widehat{w}$ to 1 recovers the experimental-data-only estimator;
2. **HDB**: Setting $\widehat{w}$ to 0 yields the historical-data-based estimator;
3. **MVE**: Omitting the bias term in (5) and minimizing (5) leads to the minimal-variance estimator in Li et al. [20];
4. **NonPessi**: Minimizing (5) as opposed to (6) produces the non-pessimistic estimator;
5. **CWE** (short for constant weight estimator): Restricting $\mathcal{W}$ to constant functions of the context and setting $\widehat{\mathrm{Var}}_U$ to $\widehat{\mathrm{Var}}$ result in the pessimistic estimator in Li et al. [2].

We will analytically compare these estimators in the following section.

4 Statistical properties and analytical comparisons

We first analyze the MSE of our proposed estimator. We next analytically compare it against other baseline estimators. Our analysis covers five different scenarios:

(i) **Heavy-tailed historical rewards**, where the reward residual $R^{(h)} - r(A^{(h)}, S^{(h)})$ exhibits substantial variability;

(ii) **Heavy-tailed experimental rewards**, where the reward residual from the control group exhibits substantial variability;

(iii) **Small posterior shifts**, where the bias due to posterior shift $\mathbb{E}[\psi^{(h)}(O^{(h)}) - \psi_0^{(e)}(O^{(e)})]$ is much smaller than the standard deviation of its estimator $\mathbb{E}_n[\psi^{(h)}(O^{(h)}) - \psi_0^{(e)}(O^{(e)})]$;

(iv) **Moderate posterior shifts**, with the bias being much larger than the estimator's standard deviation, yet falling within its high-confidence bound, making it undetectable from the data;

(v) **Large posterior shifts**, where the bias is larger than the upper confidence bound of $\mathbb{E}_n[\psi^{(h)}(O^{(h)}) - \psi_0^{(e)}(O^{(e)})]$, allowing it to be detected from the data.

See the formal definitions of these scenarios in Corollaries 1 – 4. While each of the aforementioned baseline estimators might be optimal in certain scenarios, they can perform poorly in others. In contrast, the proposed estimator is adaptive and robust: it performs comparably to the scenario-specific optimal estimators in most cases. See Table 1 for a summary.

We begin by introducing a boundedness assumption and presenting an upper bound for the MSE of the proposed estimator. To simplify the theoretical analysis, we follow [2] and study a sample-split version of the ATE estimator where half of the data triplets in $\mathcal{D}^{(e)}$ and $\mathcal{D}^{(h)}$ are used to estimate $\widehat{w}$ by solving (6), while the remaining half are used to construct the ATE estimator in (4).

Assumption 2 (ATE boundedness)**.** *There exists some constant $B > 0$ such that both the absolute values of ATE and our estimator $\widehat{\mathrm{ATE}}(\widehat{w})$ are upper bounded by B.*

Theorem 1 (MSE of the proposed estimator)**.** *Under Assumptions 1 and 2, we have for any $w \in \mathcal{W}$, $\mathrm{MSE}(\widehat{\mathrm{ATE}}(\widehat{w})) - \mathrm{MSE}(\widehat{\mathrm{ATE}}(w))$ can be bounded by:*

$$\mathbb{E}[\widehat{\mathrm{bias}}_U^2(w) - \mathrm{bias}^2(w)] + \mathbb{E}[\widehat{\mathrm{Var}}_U(w) - \mathrm{Var}(w)] + O(\alpha B^2). \tag{7}$$

Theorem 1 is generic in that it applies to any OPE estimator – direct, IS or DR – used to learn the ATE, provided that Assumptions 1 and 2 are satisfied. We also remark that Assumption 2 is mild. In practice, the size of the ATE is typically very small in A/B testing [102, 119, 108, 107]. Equation (7) upper bounds the difference in MSE between the proposed ATE estimator and any weighted estimator with a fixed weight function w. Under the realizability assumption [see e.g., 120] where $w^* \in \mathcal{W}$, setting $w = w^*$ in Theorem 1 leads to an upper bound on the difference between the MSE of our estimator and that of the optimal weighted estimator. According to (7), this upper bound can be decomposed into three parts: the first two terms quantify the estimation errors of the squared bias and the two variances respectively, and the last term, being proportional to α, represents the probability of under-coverage – the probability that $\widehat{\mathrm{bias}}_U$ or $\widehat{\mathrm{Var}}_U$ fails to upper bound the oracle bias or variance.

Notice that through the use of concentration inequalities, the last term can be made arbitrarily small without largely inflating the estimation errors of the bias and variance. As for the first two terms, a key observation is that the bias and variance upper bounds in these terms depend on the weight function only through a fixed w, rather than the estimated weight $\widehat{w}$. This arises from the pessimistic principle, which, in policy learning, ensures that the regret of the estimated policy depends only on the reward estimation error under the optimal action, rather than under the *estimated* optimal action [121, 122]. In our setting, this principle is crucial for enabling the proposed pessimistic estimator to achieve adaptivity. To elaborate, we impose the following conditions.

Assumption 3 (Coverage)**.** *The probability mass functions of both $(A^{(e)}, S^{(e)})$ and $S^{(h)}$ are bounded from below by some constant $\epsilon > 0$.*

Assumption 4 (Additive noise)**.** *Assume $R^{(h)} = r^{(h)}(0, S^{(h)}) + \epsilon^{(h)}$ for some mean-zero random error $\epsilon^{(h)}$ independent of $S^{(h)}$. Similarly, assume $R^{(e)} = r^{(e)}(A^{(e)}, S^{(e)}) + \epsilon^{(e)}_{A^{(e)}}$ for mean-zero random errors $\epsilon^{(e)}_0$ and $\epsilon^{(e)}_1$ independent of $S^{(e)}$ and $A^{(e)}$.*

Assumption 5 (Reward function boundedness)**.** *The reward functions $r^{(h)}$ and $r^{(e)}$ are uniformly bounded in absolute value by some constant $r_{\max} > 0$.*

The coverage and boundedness assumptions are commonly imposed in RL and OPE [see e.g., 123, 124]. Note that the boundedness condition applies only to the *reward function*, not to the *reward* itself. The reward – being the sum of the reward function and the residual – can be unbounded due to the potential heavy-tailedness of the residual. The additive noise assumption is widely imposed in machine learning and statistics [see e.g., 125, 126]. Under this assumption, we use $\sigma^{(h)}$ and $\sigma^{(e)}$ to denote the standard deviations of $\epsilon^{(h)}$ and $\epsilon^{(e)}_0$, respectively. These standard deviations are used to characterize the tails of these error distributions. Specifically, in the first two scenarios with heavy-tailed reward residuals, $\sigma^{(h)}$ and $\sigma^{(e)}$ can be substantially large. These two cases naturally favor EDO and HDB as optimal estimators, respectively, since they avoid incorporating heavy-tailed rewards for ATE estimation. In the last three scenarios, we measure the posterior shift by the reward difference $b(s) = r^{(h)}(0, s) - r^{(e)}(0, s)$. When $|b(s)|$ is small so that variance dominates the squared bias, MVE is asymptotically optimal since it is designed for variance minimization. With moderate-to-large values of $|b(s)|$, EDO becomes again the optimal estimator as it avoids bias by excluding the historical dataset from the ATE estimation. The following corollaries demonstrate that our proposed estimator performs comparably to these optimal estimators across all scenarios, maintaining robustness with either heavy-tailed reward residuals or posterior shift.

Corollary 1 (Scenario (i))**.** *Assume Assumptions 1 – 5 hold. Let $\delta = |\mathcal{D}^{(h)}|/|\mathcal{D}^{(e)}|$ denote the ratio between the sample sizes of the two datasets. Then with heavy-tailed historical rewards where $\sigma^{(h)} \gg [\epsilon^{-1}\sqrt{\delta}(\sigma^{(e)} + r_{\max})]$, $\omega^*(s) \to 1$ for any s so that EDO becomes the asymptotically optimal estimator. By setting w in Theorem 1 to 1, the difference in MSE between the proposed estimator and EDO is*

$$\mathbb{E}[\widehat{\mathrm{Var}}_U(\mathrm{EDO}) - \mathrm{Var}(\mathrm{EDO})] + O(\alpha B^2), \tag{8}$$

which is much smaller than MSE(EDO) itself under mild conditions specified in Appendix C.4.

Corollary 2 (Scenario (ii))**.** *Assume Assumptions 1 – 5 hold. Then with heavy-tailed experimental rewards where* $\sigma^{(e)} \gg [\epsilon^{-1/2}(\sigma^{(h)}\delta^{-1/2} + \sqrt{|\mathcal{D}^{(e)}|}r_{\max})]$*,* $\omega^*(s) \to 0$ *for any* s *so that HDB becomes the asymptotically optimal estimator. Additionally, the difference in MSE between the proposed estimator and HDB is much smaller than MSE(HDB) itself under mild conditions specified in Appendix C.5.*

Corollary 3 (Scenario (iii))**.** *Assume Assumptions 1 – 5 hold. Then with small posterior shifts such that* $|b(s)| \ll \min(\sigma^{(e)}/\sqrt{|\mathcal{D}^{(e)}|}, \sigma^{(h)}/\sqrt{|\mathcal{D}^{(h)}|})$*, MVE achieves the smallest MSE. Additionally, the difference in MSE between the proposed estimator and MVE is much smaller than MSE(MVE) itself under certain conditions specified in Appendix C.6.*

Corollary 4 (Scenarios (iv) and (v))**.** *Assume Assumptions 1 – 5 hold and that either* $b(s) > 0$ *for all* s *or* $b(s) < 0$ *for all* s*. Then with either moderate posterior shifts such that*

$$\frac{\sigma^{(e)} + r_{\max}}{\sqrt{\epsilon|\mathcal{D}^{(e)}|}} + \frac{\sigma^{(h)} + r_{\max}}{\sqrt{\epsilon|\mathcal{D}^{(h)}|}} \ll |b(s)| = O\Big(\frac{\sigma^{(e)} + r_{\max}}{\sqrt{\epsilon|\mathcal{D}^{(e)}|}}\sqrt{\log|\mathcal{D}^{(e)}|} + \frac{\sigma^{(h)} + r_{\max}}{\sqrt{\epsilon|\mathcal{D}^{(h)}|}}\sqrt{\log|\mathcal{D}^{(h)}|}\Big)$$

for any s*, or large posterior shifts such that*

$$|b(s)| \gg \Big(\frac{\sigma^{(e)} + r_{\max}}{\sqrt{\epsilon|\mathcal{D}^{(e)}|}}\sqrt{\log|\mathcal{D}^{(e)}|} + \frac{\sigma^{(h)} + r_{\max}}{\sqrt{\epsilon|\mathcal{D}^{(h)}|}}\sqrt{\log|\mathcal{D}^{(h)}|}\Big),$$

for any s*,* $\omega^*(s) \to 1$ *for any* s *so that EDO becomes the asymptotically optimal estimator. Additionally, the difference in MSE between the proposed estimator and EDO is upper bounded by* (8)*, which is much smaller than MSE(EDO) itself under mild conditions specified in Appendix C.7.*

Corollaries 1 – 4 upper bound the excess MSE of the proposed estimator over the scenario-specific optimal estimators across Scenarios (i)-(v). Importantly, the excess MSEs in (i), (iv) and (v) are independent of $\sigma^{(h)}$ or $b(s)$, which demonstrates our estimator's robustness when these parameters become (moderately) large in their respective scenarios. Furthermore, these corollaries establish the *oracle* property of our estimator: it asymptotically achieves the same MSE as the optimal estimator for each scenario, working efficiently as if it knew the underlying scenario.

We next compare against the baseline estimators mentioned in Section 3.2 analytically.

- **EDO**: According to Corollaries 1 and 4, EDO is asymptotically optimal in Scenarios (i), (iv) and (v). However, it underperforms our estimator in Scenarios (ii) and (iii), where incorporating historical data yields more accurate ATE estimation.
- **HDB**: As demonstrated in Corollary 2, HDB is asymptotically optimal in Scenario (ii). However, unlike the proposed estimator, it generally fails in Scenarios (i), (iii), (iv) and (v).
- **MVE**: Corollary 3 shows that MVE is asymptotically optimal in Scenario (iii). However, it suffers from a large bias in Scenarios (iv) and (v), due to the posterior shifts.
- **NonPessi**: Similar to Corollaries 3 and 4, we can show that NonPessi is asymptotically optimal in scenarios (iii) and (v) when the posterior shift is either small or large. However, it is not optimal for moderate shifts in Scenario (iv). This is because without adopting the pessimistic principle, its excess MSE depends on the estimation errors $\widehat{\mathrm{bias}}(\widehat{w})$ and $\widehat{\mathrm{Var}}(\widehat{w})$ at the estimated weight $\widehat{w}$. Although the optimal population-level weight $w^* \to 1$ with moderate posterior shifts, the estimated $\widehat{w}$ may not, since the bias is not large enough to be detectable [2]. Similarly, in the first two scenarios with heavy-tailed rewards, NonPessi – unlike the pessimistic estimator – can suffer from a large MSE when $\sigma^{(e)}$ and $\sigma^{(h)}$ are large, yet not sufficiently so to be detected from the data.
- **CWE**: While [2] showed that CWE is optimal in Scenarios (iv) and (v), it differs from our method in two ways: (a) it restricts $\mathcal{W}$ to constant weight functions, and (b) it applies the pessimistic principle only partially – to upper bound the squared bias term but not the variance. (a) leads to its sub-optimality in Scenario (iii), where the optimal weight for MVE is typically context-dependent rather than being constant. Similar to NonPessi, (b) makes CWE sub-optimal in the first two scenarios.

We have so far focused exclusively on the estimation of the ATE. To conclude this section, we remark that our proposal also accommodates valid inference for A/B testing. By employing sample-splitting

and doubly robust ATE estimation, valid p-values can be readily obtained when combined with. Specifically, we use one half of the data to estimate the weight function $\widehat{w}$ and nuisance functions (the reward and density ratio), and the other half to construct the ATE estimator. Following [127], one can show that the resulting ATE estimator is asymptotically normal under suitable regularity conditions. As a result, standard z-tests based on normal approximation can be used to obtain valid p-values. Our numerical studies in Section 5 confirm that the resulting p-values remain valid across all experimental settings.

5 Numerical experiments

In this section, we evaluate the finite-sample performance of the proposed estimator, comparing it against **EDO**, **MVE**, **CWE**, and **NonPessi** (introduced in Section 3.2). We exclude **HDB** as it performs similarly or worse than MVE in our experiments. Instead, we include **LASSO**, proposed by Cheng and Cai [25], which selects weights by minimizing the estimated variance of the ATE estimator with a Lasso penalty. MSEs are computed over 100 simulation replications. Details of the data generating process are provided in Appendix A.

Example 5.1 (**Synthetic-data simulation).** We design settings to cover all five scenarios mentioned in Section 4 and Table 1. Specifically, we model the difference between reward functions $b(s)$ as $\mu_{\text{diff}} \times d(s)$ for some nonzero function d and a scalar parameter $\mu_{\text{diff}} \in [0, 5]$ controlling the degree of posterior shift. When $\mu_{\text{diff}} = 0$, the reward functions are identical, indicating no posterior shift. Increasing μ_{diff} leads to larger shifts. This covers Scenarios (iii) – (v), which range from small to moderate to large posterior shifts. We also consider two forms for the function d: (i) a piecewise function of the context variable, and (ii) a linear function of the context variable, resulting in piecewise and linear shifts, respectively. Finally, we allow the reward residuals to follow either a normal distribution (light-tailed) or a Student's t-distribution with 6 degrees of freedom (heavy-tailed). This covers the first two scenarios.

The top panels of Figure 2 report the MSEs of all ATE estimators under piecewise shifts, while the middle panels exclude MVE, which exhibits large MSEs even under moderate to large posterior shifts, to allow for a clearer comparison of the remaining estimators. It can be seen that when μ_{diff} is small, MVE achieves the lowest MSEs, and the proposed method performs comparably. As μ_{diff} increases, the proposed estimator outperforms all baseline alternatives in most cases. Notably, even when μ_{diff} is large – where EDO is expected to perform best – our proposal still achieves lower MSEs under normally distributed experimental reward residuals. This benefits from its use of a context-adaptive weight function. When the reward difference $b(s)$ is negative for some contexts and positive for others, a properly chosen context-adaptive weight can still incorporate historical data to reduce variance while effectively cancelling out bias. In contrast, CWE and LASSO, which rely on constant weights, tend to converge toward EDO's performance.

Bottom panels of Figure 2 show similar trends under linear shifts (excluding MVE). We also consider a nonlinear form of $d(s)$ and conduct additional experiments in Appendix A, which confirm similar patterns under nonlinear posterior shifts. Finally, we remark that the LASSO estimator is implemented using a carefully chosen tuning parameter to ensure competitive performance. Additional results in Appendix A (Figure 8) reveal LASSO's sensitivity to this hyperparameter.

Example 5.2 (**Ridesharing-data-based simulation).** In this example, we construct a simulation environment based on a real-world A/A dataset collected from a ridesharing platform. The contextual information includes two variables: the total online time of drivers and the number of order requests across one day. The reward is defined as the total daily income earned by each driver. We first learn the outcome model using these variables from this A/A dataset, and then generate synthetic experimental and historical data based on this model, following scenarios similar to those in Example 5.1. Results reported in Figures 3 and additional results in Appendix A (Figures 9-12) align with the findings in Example 5.1, where the proposed estimator achieves the lowest MSEs in most cases.

Additionally, we conduct a clinical data–based simulation in Example A.1 (Appendix). The results exhibit patterns consistent with those in Examples 5.1–5.2, where the proposed estimator outperforms competing methods in most cases. Furthermore, we assess the inference procedure by testing the nullity of the ATE and comparing it with those based on EDO and CWE. As shown in Table 3, all three methods adequately control the Type I error under the null (ATE = 0), while the proposed test demonstrates higher power under the alternative.

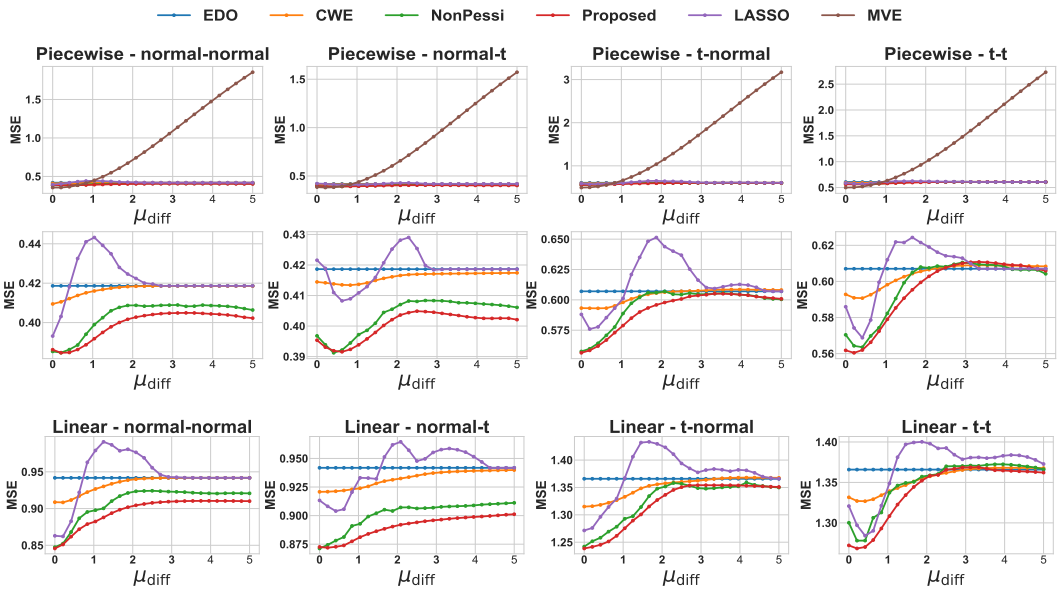

Figure 2: MSEs in Example 5.1. Top panels show all estimators under piecewise shifts; middle panels zoom in without MVE; bottom panels present results for linear shifts excluding MVE.

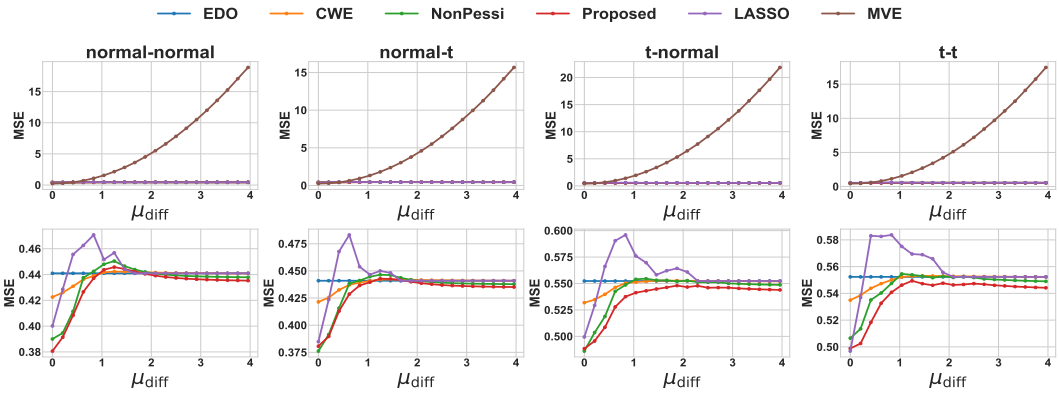

Figure 3: MSEs of ATE estimators in Example 5.2. Top: all estimators; Bottom: exclude MVE.

Discussion

In this work, we study how to integrate historical control data with experimental data to enhance A/B testing. We proposed a covariate-dependent weighting scheme that treats the weight as a policy and learns it by minimizing a pessimistic upper bound on the estimator's MSE. We establish MSE bounds for the resulting estimator. We evaluate it against competitive baselines across five representative settings. Both our theoretical analysis and empirical results demonstrate greater robustness to heavy-tailed rewards and near-optimal handling of diverse posterior shifts.

Several directions merit future exploration. The proposed method focuses on a non-dynamic setting, whereas in many practical applications, treatments are sequential and may influence future outcomes. A natural extension is to accommodate dynamic settings with carryover effects by explicitly modeling the underlying dynamics. Our theoretical analysis also relies on bounded-reward assumptions, which could be relaxed to handle unbounded outcomes. Beyond IS and DR estimators for the ATE, the proposed framework can be extended to more general off-policy evaluation and reinforcement learning objectives, including least-squares temporal-difference methods and fitted Q-evaluation.

Acknowledgments

Xiangkun Wu's research is supported by the National Key Research and Development Program of China (Grant No. 2024YFC2511003). Ting Li's research is partially supported by the National Natural Science Foundation of China (No. 12571304), the Shanghai Pujiang Program (No. 24PIC030), CCF-DiDi GAIA Collaborative Research Funds and the Program for Innovative Research Team of Shanghai University of Finance and Economics.

References

[1] Zhe Xu, Zhixin Li, Qingwen Guan, Dingshui Zhang, Qiang Li, Junxiao Nan, Chunyang Liu, Wei Bian, and Jieping Ye. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 905–913, 2018.

[2] Ting Li, Chengchun Shi, Qianglin Wen, Yang Sui, Yongli Qin, Chunbo Lai, and Hongtu Zhu. Combining experimental and historical data for policy evaluation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28630–28656, 2024.

[3] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[4] Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

[5] Lu Tang and Peter XK Song. Fused lasso approach in regression coefficients clustering–learning parameter heterogeneity in data integration. *Journal of Machine Learning Research*, 17(113):1–23, 2016.

[6] Fei Xue and Annie Qu. Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, 116(536):1914–1927, 2021.

[7] Qi Xu and Annie Qu. Representation retrieval learning for heterogeneous data integration. *arXiv preprint arXiv:2503.09494*, 2025.

[8] Jieli Shen, Regina Y Liu, and Min-ge Xie. i fusion: Individualized fusion learning. *Journal of the American Statistical Association*, 115(531):1251–1267, 2020.

[9] Chengchun Shi, Rui Song, Wenbin Lu, and Bo Fu. Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):681–702, 2018.

[10] Xinlei Chen, Victor B Talisa, Xiaoqing Tan, Zhengling Qi, Jason N Kennedy, Chung-Chou H Chang, Christopher W Seymour, and Lu Tang. Federated learning of robust individualized decision rules with application to heterogeneous multihospital sepsis population. *The Annals of Applied Statistics*, 19(2):1270–1291, 2025.

[11] Zhuangdi Zhu, Kaixiang Lin, Anil K Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (11):13344–13362, 2023.

[12] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524, 2023.

[13] Xu Shi, Ziyang Pan, and Wang Miao. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023.

[14] Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pages 5247–5256. PMLR, 2021.

[15] Sijia Li and Alex Luedtke. Efficient estimation under data fusion. *Biometrika*, 110(4): 1041–1054, 2023.

[16] Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596, 2023.

[17] Haotian Wang, Haoxuan Li, Wenjing Yang, Hao Zou, Wanrong Huang, and Kun Kuang. Estimating treatment effect across heterogeneous data sources: An instrumental variable approach. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024.

[18] Chenyin Gao, Shu Yang, Mingyang Shan, Wenyu Ye, Ilya Lipkovich, and Douglas Faries. Improving randomized controlled trial analysis via data-adaptive borrowing. *Biometrika*, 112 (2):asae069, 2025.

[19] Xueqing Liu, Qinwei Yang, Zhaoqing Tian, Ruocheng Guo, and Peng Wu. Optimal policy adaptation under covariate shift. *arXiv preprint arXiv:2501.08067*, 2025.

[20] Xinyu Li, Wang Miao, Fang Lu, and Xiao-Hua Zhou. Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403, 2023.

[21] Ruoxuan Xiong, Allison Koenecke, Michael Powell, Zhu Shen, Joshua T Vogelstein, and Susan Athey. Federated causal inference in heterogeneous observational data. *Statistics in Medicine*, 42(24):4418–4439, 2023.

[22] Ying Sheng, Jing Qin, and Chiung-Yu Huang. Sequential data integration under dataset shift. *Technometrics*, 66(4):662–670, 2024.

[23] Peng Wu, Shanshan Luo, and Zhi Geng. On the comparative analysis of average treatment effects estimation via data combination. *Journal of the American Statistical Association*, pages 1–12, 2025.

[24] Yi Liu, Alexander W Levis, Ke Zhu, Shu Yang, Peter B Gilbert, and Larry Han. Targeted data fusion for causal survival analysis under distribution shift. *arXiv preprint arXiv:2501.18798*, 2025.

[25] David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021.

[26] Issa J Dahabreh, Sarah E Robertson, Lucia C Petito, Miguel A Hernán, and Jon A Steingrimsson. Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, 79(2):1057–1072, 2023.

[27] Larry Han, Zhu Shen, and Jose Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. *Advances in Neural Information Processing Systems*, 36:70453–70482, 2023.

[28] Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai. Federated adaptive causal estimation (face) of target treatment effects. *Journal of the American Statistical Association*, (Accepted), 2025.

[29] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials revisited. *Contemporary clinical trials*, 45:139–145, 2015.

[30] Takahiro Hasegawa, Brian Claggett, Lu Tian, Scott D Solomon, Marc A Pfeffer, and Lee-Jen Wei. The myth of making inferences for an overall treatment efficacy with data from multiple comparative studies via meta-analysis. *Statistics in Biosciences*, 9:284–297, 2017.

[31] Qinshu Lian, Jing Zhang, James S Hodges, Yong Chen, and Haitao Chu. Accounting for post-randomization variables in meta-analysis: A joint meta-regression approach. *Biometrics*, 79(1):358–367, 2023.

[32] Kollin W Rott, Gert Bronfort, Haitao Chu, Jared D Huling, Brent Leininger, Mohammad Hassan Murad, Zhen Wang, and James S Hodges. Causally interpretable meta-analysis: Clearly defined causal effects and two case studies. *Research Synthesis Methods*, 15(1):61–72, 2024.

[33] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.

[34] Piersilvio De Bartolomeis, Javier Abad Martinez, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR, 2024.

[35] Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):362–388, 2025.

[36] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1): 1731–1755, 2015.

[37] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *Advances in Neural Information Processing Systems*, 28, 2015.

[38] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

[39] Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pages 11362–11371. PMLR, 2021.

[40] Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using l_p confidence sets. *Advances in Neural Information Processing Systems*, 35:20974–20987, 2022.

[41] Olivier Jeunen and Bart Goethals. Pessimistic decision-making for recommender systems. *ACM Transactions on Recommender Systems*, 1(1):1–27, 2023.

[42] Yunzhe Zhou, Zhengling Qi, Chengchun Shi, and Lexin Li. Optimizing pessimism in dynamic treatment regimes: A bayesian learning approach. In *International Conference on Artificial Intelligence and Statistics*, pages 6704–6721. PMLR, 2023.

[43] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

[44] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

[45] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

[46] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021.

[47] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025. PMLR, 2022.

[48] Xiaohong Chen, Zhengling Qi, and Runzhe Wan. Steel: Singularity-aware reinforcement learning. *arXiv preprint arXiv:2301.13152*, 2023.

[49] Wenzhuo Zhou. Bi-level offline policy optimization with limited exploration. *Advances in Neural Information Processing Systems*, 36:55022–55035, 2023.

[50] Yue Wang, Zhongchang Sun, and Shaofeng Zou. A unified principle of pessimism for offline reinforcement learning under model mismatch. *Advances in Neural Information Processing Systems*, 37:9281–9328, 2024.

[51] Danyang Wang, Chengchun Shi, Shikai Luo, and Will Wei Sun. Pessimistic causal reinforcement learning with mediators for confounded offline data. *arXiv preprint arXiv:2403.11841*, 2024.

[52] Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 541–549. PMLR, 2024.

[53] Jin Zhu, Xin Zhou, Jiaang Yao, Gholamali Aminian, Omar Rivasplata, Simon Little, Lexin Li, and Chengchun Shi. Semi-pessimistic reinforcement learning. *arXiv preprint arXiv:2505.19002*, 2025.

[54] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[55] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

[56] Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

[57] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, pages 485–511, 2014.

[58] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

[59] Josiah Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[60] Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning*, pages 2366–2375. PMLR, 2019.

[61] Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

[62] Shuguang Yu, Shuxing Fang, Ruixin Peng, Zhengling Qi, Fan Zhou, and Chengchun Shi. Two-way deconfounder for off-policy evaluation in causal reinforcement learning. *Advances in Neural Information Processing Systems*, 37:78169–78200, 2024.

[63] Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

[64] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.

[65] Wei Luo, Yeying Zhu, and Debashis Ghosh. On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104(1):51–65, 2017.

[66] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[67] Yihao Feng, Tongzheng Ren, Ziyang Tang, and Qiang Liu. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pages 3102–3111. PMLR, 2020.

[68] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

[69] Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391, 2021.

[70] Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793, 2022.

[71] Andrew Bennett, Nathan Kallus, Miruna Oprescu, Wen Sun, and Kaiwen Wang. Efficient and sharp off-policy evaluation in robust markov decision processes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[72] Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in pomdps. *Advances in Neural Information Processing Systems*, 36, 2024.

[73] Zeyu Bian, Chengchun Shi, Zhengling Qi, and Lan Wang. Off-policy evaluation in doubly inhomogeneous environments. *Journal of the American Statistical Association*, 120(550): 1102–1114, 2025.

[74] James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

[75] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[76] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[77] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

[78] Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in Neural Information Processing Systems*, 33:9398–9411, 2020.

[79] Mark Rowland, Anna Harutyunyan, Hado Hasselt, Diana Borsa, Tom Schaul, Rémi Munos, and Will Dabney. Conditional importance sampling for off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 45–55. PMLR, 2020.

[80] Muhammad Faaiz Taufiq, Arnaud Doucet, Rob Cornish, and Jean-Francois Ton. Marginal density ratio for off-policy evaluation in contextual bandits. *Advances in Neural Information Processing Systems*, 36:52648–52691, 2023.

[81] Yuchen Hu and Stefan Wager. Off-policy evaluation in partially observed markov decision processes under sequential ignorability. *The Annals of Statistics*, 51(4):1561–1585, 2023.

[82] David M Bossens and Philip S Thomas. Low variance off-policy evaluation with state-based importance sampling. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 871–883. IEEE, 2024.

[83] Shuze Liu and Shangtong Zhang. Efficient policy evaluation with offline data informed behavior policy design. In *International Conference on Machine Learning*, pages 32345–32368. PMLR, 2024.

[84] Hongyi Zhou, Josiah P Hanna, Jin Zhu, Ying Yang, and Chengchun Shi. Demystifying the paradox of importance sampling with an estimated history-dependent behavior policy in off-policy evaluation. In *Forty-second International Conference on Machine Learning*, 2025.

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

[86] Pedro HC Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122, 2020.

[87] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104, 2011.

[88] Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

[89] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[90] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

[91] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

[92] Chengchun Shi, Wenbin Lu, and Rui Song. Breaking the curse of nonregularity with subagging—inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21(176):1–67, 2020.

[93] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[94] Chengchun Shi, Runzhe Wan, Victor Chernozhukov, and Rui Song. Deeply-debiased off-policy interval estimation. In *International conference on machine learning*, pages 9580–9591. PMLR, 2021.

[95] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

[96] Peng Liao, Zhengling Qi, Runzhe Wan, Predrag Klasnja, and Susan A Murphy. Batch policy learning in average reward markov decision processes. *Annals of statistics*, 50(6):3364, 2022.

[97] Yang Xu, Chengchun Shi, Shikai Luo, Lan Wang, and Rui Song. Quantile off-policy evaluation via deep conditional generative learning. *arXiv preprint arXiv:2212.14466*, 2022.

[98] Jeonghwan Lee and Cong Ma. Off-policy estimation with adaptively collected data: the power of online learning. *Advances in Neural Information Processing Systems*, 37:133908–133947, 2024.

[99] Chengchun Shi, Jin Zhu, Ye Shen, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, 119(545):273–284, 2024.

[100] Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528): 1665–1682, 2019.

[101] Peter W Glynn, Ramesh Johari, and Mohammad Rasouli. Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems*, 33:15054–15064, 2020.

[102] Vivek Farias, Andrew Li, Tianyi Peng, and Andrew Zheng. Markovian interference in experiments. *Advances in Neural Information Processing Systems*, 35:535–549, 2022.

[103] Chengchun Shi, Runzhe Wan, Ge Song, Shikai Luo, Hongtu Zhu, and Rui Song. A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets. *The Annals of Applied Statistics*, 17(4):2701–2722, 2023.

[104] Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, 118(543):2059–2071, 2023.

[105] Ting Li, Chengchun Shi, Jianing Wang, Fan Zhou, et al. Optimal treatment allocation for efficient policy evaluation in sequential decision making. *Advances in Neural Information Processing Systems*, 36:48890–48905, 2023.

[106] Ting Li, Chengchun Shi, Zhaohua Lu, Yi Li, and Hongtu Zhu. Evaluating dynamic conditional quantile treatment effects with applications in ridesharing. *Journal of the American Statistical Association*, 119(547):1736–1750, 2024.

[107] Ke Sun, Linglong Kong, Hongtu Zhu, and Chengchun Shi. Arma-design: Optimal treatment allocation strategies for a/b testing in partially observable time series experiments. *arXiv preprint arXiv:2408.05342*, 2024.

[108] Ruoxuan Xiong, Alex Chin, and Sean J Taylor. Data-driven switchback experiments: Theoretical tradeoffs and empirical bayes designs. *arXiv preprint arXiv:2406.06768*, 2024.

[109] Shikai Luo, Ying Yang, Chengchun Shi, Fang Yao, Jieping Ye, and Hongtu Zhu. Policy evaluation for temporal and/or spatial dependent experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):623–649, 2024.

[110] Jinjuan Wang, Qianglin Wen, Yu Zhang, Xiaodong Yan, and Chengchun Shi. A two-armed bandit framework for a/b testing. *arXiv preprint arXiv:2507.18118*, 2025.

[111] Qianglin Wen, Chengchun Shi, Niansheng Tang, Hongtu Zhu, et al. Unraveling the interplay between carryover effects and reward autocorrelations in switchback experiments. In *Forty-second International Conference on Machine Learning*, 2025.

[112] Yonghan Jung and Alexis Bellot. Efficient policy evaluation across multiple different experimental datasets. *Advances in Neural Information Processing Systems*, 37:136361–136392, 2024.

[113] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

[114] Christina Yuan, Yash Chandak, Stephen Giguere, Philip S Thomas, and Scott Niekum. Sope: Spectrum of off-policy estimators. *Advances in Neural Information Processing Systems*, 34: 18958–18969, 2021.

[115] Allen Nie, Yash Chandak, Christina Yuan, Anirudhan Badrinath, Yannis Flet-Berliac, and Emma Brunskill. Opera: Automatic offline policy evaluation with re-weighted aggregates of multiple estimators. *Advances in Neural Information Processing Systems*, 37:103652–103680, 2024.

[116] Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

[117] George Casella and Roger Berger. *Statistical inference*. CRC press, 2024.

[118] Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[119] Susan Athey, Peter J Bickel, Aiyou Chen, Guido W Imbens, and Michael Pollmann. Semiparametric estimation of treatment effects in randomised experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1615–1638, 2023.

[120] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[121] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[122] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196, 2022.

[123] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[124] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for dynamics and control*, pages 486–489. PMLR, 2020.

[125] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1): 2009–2053, 2014.

[126] Chengchun Shi, Yunzhe Zhou, and Lexin Li. Testing directed acyclic graph via structural, supervised and generative adversarial learning. *Journal of the American Statistical Association*, 119(547):1833–1846, 2024.

[127] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[128] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes: with applications to statistics*, pages 16–28. Springer, 1996.

[129] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564, 2014.

[130] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, pages 2786–2819, 2013.

[131] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the problem setting, the proposed method, and the scope of the contributions. They outline the motivation (addressing distributional shift in data integration, particularly in the context of A/B testing), introduce our method (a pessimistic data integration approach), and summarize the main findings (near-optimal performance across five scenarios). These claims are substantiated by both theoretical analysis and empirical evaluations presented in the main body of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:We provide a detailed discussion in Appendix D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: In Section 4, we present the full set of assumptions required by our analysis and discuss their practicality and commonality. We first derive a general expression for the MSE of our proposed method and then provide performance results for five specific scenarios in the form of corollaries. All assumptions and theorems are properly numbered and referenced. Complete and rigorous proofs are provided in Appendix C, where we also elaborate on how each assumption is used within the corresponding theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We propose a new algorithm, which is described in detail in Section 3.2. Additional implementation details are provided in Appendix B. To further enhance reproducibility, we include the full source simulation code in the supplementary material, covering both the algorithm implementation and the experimental setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide full access to the simulation code and experimental scripts in the supplementary material, along with detailed instructions for reproducing the main experimental results. A 'README' file is included to guide users through the process. For the semi-synthetic experiment based on private data, we are unfortunately unable to release the dataset due to privacy constraints. However, we provide detailed descriptions of the experimental setup, processing procedures, and key results in Section 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer:[Yes]

   Justification: We provide comprehensive experimental details, including hyperparameter settings, and training configurations. These are described in Section 5, with additional implementation and tuning details provided in Appendix A and in the supplementary code. We also specify the choice of optimizer, the number of repetitions, and how parameters were selected to ensure fairness across baselines. This information allows readers to fully interpret and evaluate the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[Yes]

   Justification: In the toy example2, we report 95% confidence intervals derived using the Central Limit Theorem, based on repeated simulations and the estimated standard errors. For the simulation and real-data-based simulation experiments, we report the mean MSE over 100 independent runs. This is motivated by the Law of Large Numbers, aiming to ensure statistical reliability and reduce the impact of randomness, rather than relying on a small number of trials. Details of the error estimation process are provided in Appendix 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were conducted on a high-performance computing node equipped with dual AMD EPYC 7742 64-Core Processors (128 logical cores). No GPUs were used. Each simulation experiment with 100 replications are typically completed within 90 minutes. We have provided the above details in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research complies fully with the NeurIPS Code of Ethics. No human subjects or personally identifiable information are involved. The real-data-based simulation is based on private structured data that cannot be released due to legal and privacy constraints, but all usage complies with institutional and data protection regulations.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work contributes to methodological advancements in data integration and policy evaluation, particularly in the context of A/B testing under distributional shift. Potential positive societal impacts include enabling more efficient and statistically sound experimentation in fields such as healthcare, education, and online platforms. One potential risk is the misuse of historical data integration in sensitive domains, which could amplify existing biases if the data distribution is not properly accounted for.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: This paper does not involve the release of models or datasets that carry a high risk of misuse. All simulations are based on synthetic or private structured data, and no pretrained generative models or scraped datasets are used.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer:[Yes]

Justification: While not explicitly mentioned in the main text, we use standard Python packages such as NumPy,SciPy ,and scikit-learn in our implementation. All usage complies with the respective license terms.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper introduces a new method and provides implementation code for reproducibility. However, it does not release any new dataset, pretrained model, or software asset intended for reuse beyond the experimental context. Therefore, we consider that no new asset has been introduced.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: This paper does not involve crowdsourcing or research with human subjects. We simply propose a new algorithm and evaluate it using simulation data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve research with human subjects or crowdsourcing. We propose a new algorithm and evaluate it using simulated data, which do not require IRB or equivalent approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer:[NA]

    Justification: The core methodology and experiments in this paper do not involve any large language models (LLMs) as important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

Appendix Contents

A Experimental Settings and Additional Results

In this section, we present details of the data generating process for the toy example in Table 2 and Section 5, and additional experimental results. All experiments were conducted on a high-performance computing node equipped with dual AMD EPYC 7742 64-Core Processors (128 logical cores). Each experiment with 100 replications can be typically completed within 90 minutes.

Toy Example in Table 2. We define the reward functions for the experimental and historical data as follows

$$R_e = \begin{cases} 10 + 3S_e + 6\epsilon_e, & \text{if } A_e = 0 \\ 10 + A_e + 3S_e + 2\epsilon_e, & \text{if } A_e = 1 \end{cases}, \quad R_h = 11.4 + S_h + 6\epsilon_h,$$

where the state variables S_h, S_e and the errors ϵ_h, ϵ_e are independently drawn from the standard normal distribution $N(0,1)$. In the experimental data, actions are evenly split between the two treatment groups, with half assigned to 1 and the other half to 0.

Example 5.1 (Continued). In this example, we consider five posterior shift scenarios as follows.

(1) **Piecewise Shifts:** The reward functions for the historical and experimental data are:

$$R_e = 10 + A_e + S_e + 2\epsilon_e, \quad R_h = 10 + d_\mu(S_h)\mu_{\text{diff}} + S_h + (2 + d_S(S_h))\epsilon_h$$

where

$$d_S(S) = \begin{cases} -1, & S < -1 \\ -1, & -1 \le S < 0 \\ 2, & S > 0 \end{cases}, \quad d_\mu(S) = \begin{cases} 0, & S < -1 \\ 1, & -1 \le S < 0 \\ 1, & S > 0 \end{cases},$$

and S_e and S_h are sampled from $N(0,1)$. The noise terms ϵ_e and ϵ_h follow four distribution combinations: **normal-normal**, **normal-t**, **t-normal**, and **t-t**, where **normal** is $N(0,1)$ and **t** is the t-distribution with 6 degrees of freedom.

The parameter μ_{diff} captures distributional differences between datasets, discretized into 25 values from 0 to 5. In the experimental dataset, the action A_e alternates deterministically between 0 and 1 – 0 for the first, third, and fifth samples, and 1 for the second, fourth, and sixth samples. The shift magnitude depends on the state S through $d_\mu(S)$, and the noise variance varies with the state via $d_S(S)$.

(2) **Linear Shifts:** The reward functions for the two datasets are

$$R_e = 10 + A_e + S_e + 3\epsilon_e, \quad R_h = 10 + \mu_{\text{diff}} + \mu_{\text{diff}} S_h + 3\epsilon_h,$$

where S_e, S_h, ϵ_e, ϵ_h, and A_e follow the same configurations to Setting (1).

(3) **Cosine Shifts:** The reward functions are given by

$$R_e = 10 + A_e + \cos(S_e) + 3\epsilon_e, \quad R_h = 10 + \mu_{\text{diff}} + \mu_{\text{diff}} \cos(S_h) + S_h + 3\epsilon_h.$$

All other configurations remain the same to Setting (1).

(4) **Quadratic Shifts:** The reward functions are given by

$$R_e = 10 + A_e + S_e^2 + 3\epsilon_e, \quad R_h = 10 + \mu_{\text{diff}} + \mu_{\text{diff}} S_e^2 + S_h + 3\epsilon_h.$$

All other configurations remain the same.

(5) **Absolute Shifts:** The reward functions are given by

$$R_e = 10 + A_e + |S_e| + 3 \cdot \epsilon_e, \quad R_h = 10 + \mu_{\text{diff}} + \mu_{\text{diff}} |S_h| + |S_h| + 3\epsilon_h.$$

All other configurations remain the same.

Figure 4 shows the empirical MSEs of all estimators, including MVE, under linear shifts. When μ_{diff} is small, MVE performs well due to minimal bias in the historical data. However, as μ_{diff} grows, this bias increases sharply, causing MVE's MSE to rise substantially. Figures 5–7 present the empirical MSEs under cosine, quadratic, and absolute shift settings. The results remain consistent across varying levels of posterior shift. The MVE method shows marginal gains only when μ_{diff} is very small; however, its performance quickly deteriorates as μ_{diff} increases, because it ignores the bias from the historical data, resulting in substantially large MSE.

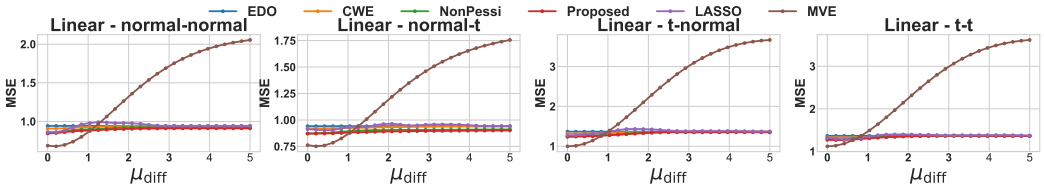

Figure 4: Empirical means of MSE for various estimators, including MVE, under the piecewise shifts and linear shifts.

When the experimental data follow a normal distribution, our method consistently outperforms all baselines over the entire range of $\mu_{\text{diff}} \in [0, 5]$, regardless of the data generating process of the historical dataset. It achieves a much smaller MSE compared to EDO and Pessi when μ_{diff} is small. Although its MSE increases as μ_{diff} grows, our method still maintains a clear advantage over other baselines. This gain arises from the use of a non-constant, learned weight function that adapts to distributional shifts, in contrast to EDO's fixed weight.

When the experimental data follow a t-distribution, our method continues to perform well, particularly when $\mu_{\text{diff}} < 2$. As μ_{diff} increases, the performance of all methods converges to that of EDO.

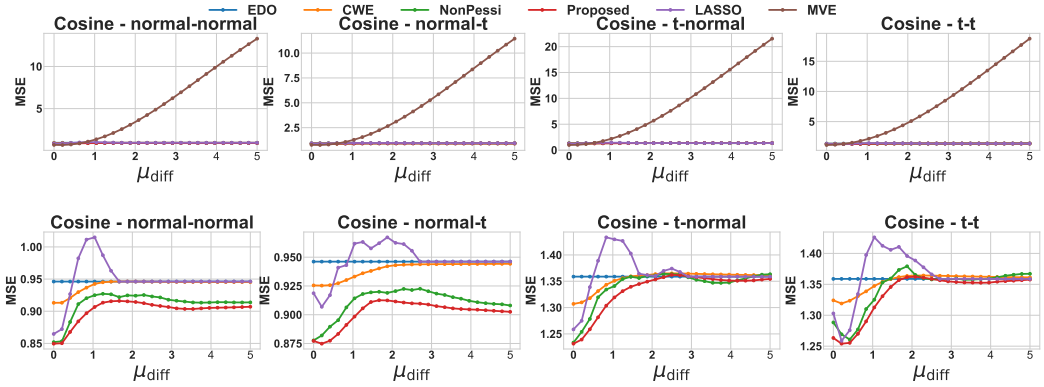

Figure 5: Empirical means of MSEs in Example 5.1 under cosine shifts. The top panel shows all estimators; the bottom panel zooms in by excluding the MVE method.

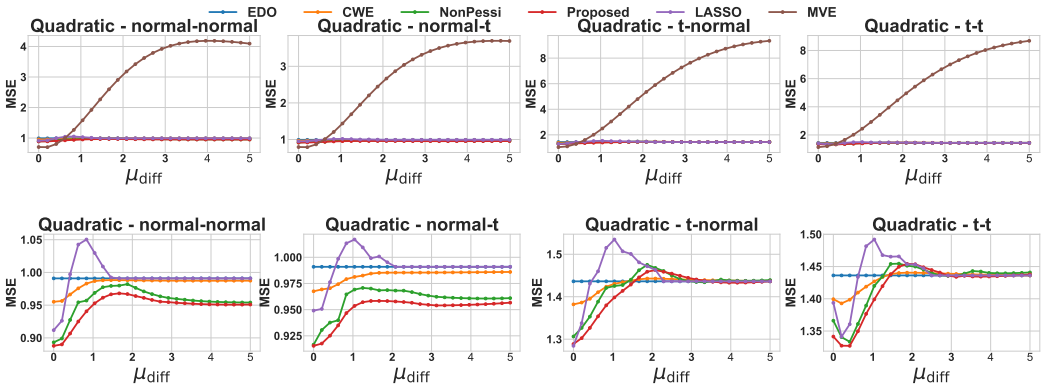

Figure 6: Empirical means of MSEs in Example 5.1 under quadratic shifts. The top panel shows all estimators; the bottom panel zooms in by excluding the MVE method.

Sensitivity Analysis of the Lasso Tuning Parameter. We investigate the performance of the Lasso estimator across a range of tuning parameters $\lambda \in \{0.1, 0.2, \ldots, 1.0\}$, using the same data generating process as in the piecewise shifts setting. Figure 8 reports the performance of Lasso and EDO across varying λ values. For small values of μ_{diff}, Lasso with a small λ outperforms EDO. However, its

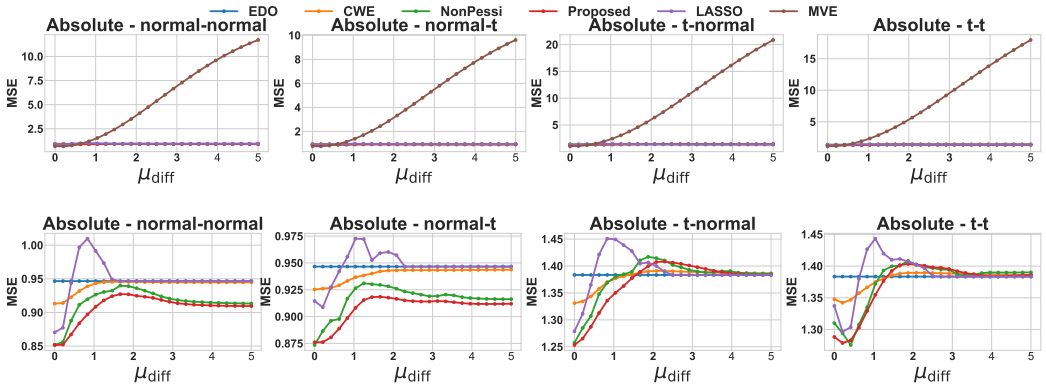

Figure 7: Empirical means of MSEs in Example 5.1 under absolute shifts. The top panel shows all estimators; the bottom panel zooms in by excluding the MVE method.

performance deteriorates as μ_{diff} increases. In contrast, Lasso with a large λ performs comparably to EDO.

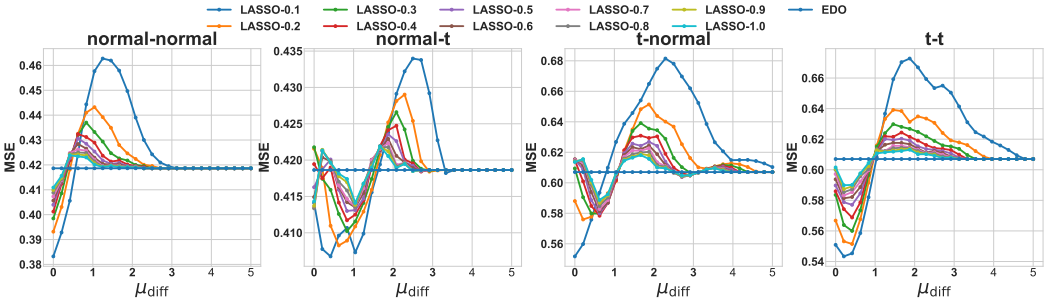

Figure 8: Hyperparameter Sensitivity Analysis

Example 5.2 (Continued). This example uses a real-world A/A experiment dataset from a leading ridesharing platform. The reward R represents daily driver income, and the contextual variables S_1 and S_2 denote the number of ride requests and total online time during the first hour of the day, respectively.

For privacy concerns, company and city identifiers are omitted, and all variables are scaled. The contextual features S_1 and S_2 are normalized to have unit standard deviation. We fit the following linear model:

$$R = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \epsilon,$$

and obtain the estimates $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$. Based on the estimated coefficients, we generate the experimental and historical datasets as follows:

$$R_e = \widehat{\beta}_0 + A_e + \widehat{\beta}_1 S_{e1} + \widehat{\beta}_2 S_{e2} + \delta\epsilon_e,$$
$$R_h = \widehat{\beta}_0 + \widehat{\beta}_1 S_{h1} + \widehat{\beta}_2 S_{h2} + \mu_{\text{diff}} + \mu_{\text{diff}} \cdot (S_{h1} + S_{h2})/20 + \delta\epsilon_h,$$

where S_{e1}, S_{h1} are sampled from $N(\mu_1, 1)$ and S_{e2}, S_{h2} from $N(\mu_2, 1)$, with μ_1 and μ_2 being the empirical means of S_1 and S_2 from the real dataset. To ensure privacy, we do not report μ_1 and μ_2 individually, but their sum lies between 10 and 20. The action A_e is binary, assigned deterministically: even-indexed samples receive $A_e = 1$, odd-indexed samples $A_e = 0$. The noise terms ϵ_e and ϵ_h follow four combinations: normal-normal, normal-t, t-normal, and t-t, where the t-distribution has 6 degrees of freedom. The experimental dataset contains $|\mathcal{D}_e| = 48$ samples, and the historical dataset has $|\mathcal{D}_h| = m \cdot |\mathcal{D}_e|$, with $m \in \{1, 2, 3\}$. A noise scaling constant $\delta \in \{1, 2, 3\}$ controls the noise magnitude.

Figure 9 reports the empirical MSEs across methods, showing consistent patterns with the $m = 1$ case in the main text. Across all settings, our method consistently achieves strong performance.

When the experimental data is heavy-tailed, it significantly outperforms non-pessimistic baselines with lower MSEs across all μ_{diff} values. With small posterior shifts, it clearly outperforms both EDO and Pessi; in the moderate shift regime, it outperforms Lasso; and under large posterior shifts, it remains stable and performs slightly better than EDO.

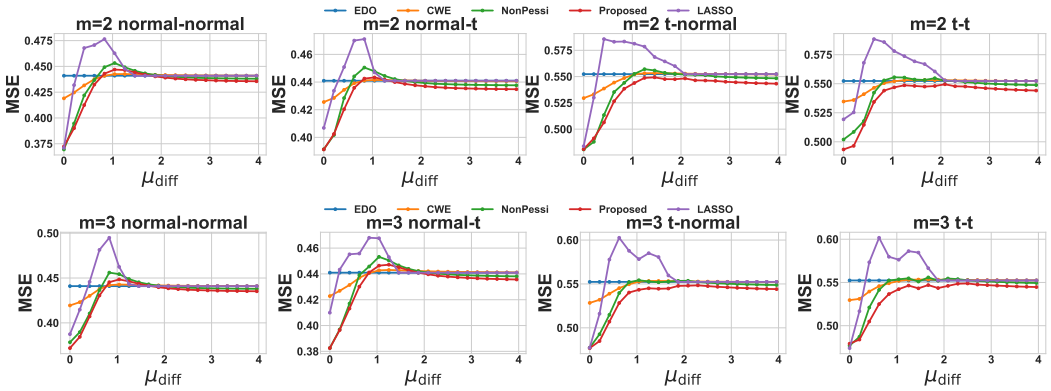

Figure 9: Empirical means of MSEs of various methods with $\delta = 1$ in Example 5.2 for $m = 2$ (top) and $m = 3$ (bottom).

Figure 10 shows the results under varying noise magnitudes. As expected, the MSE of all methods increases with higher residual variance (characterized by δ), reflecting the impact of noise on estimation accuracy. Nevertheless, our method consistently outperforms all baselines across the full range of μ_{diff} values.

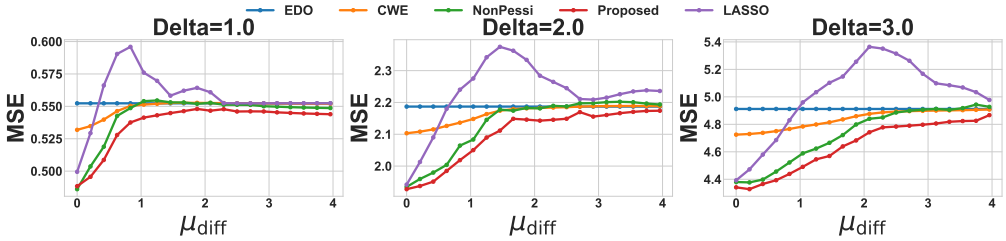

Figure 10: Empirical means of MSEs of various methods with $m = 1$ in Example 5.2 across different δs.

Figure 11 examines the impact of treatment assignment under varying probabilities $\mathbb{P}(A_e = 1) =$ prob, with prob $\in \{0.3, 0.5, 0.7\}$, in the normal-normal setting with $m = \delta = 1$. The results show consistent performance across all methods, with our approach remaining robust and effective under different assignment probabilities.

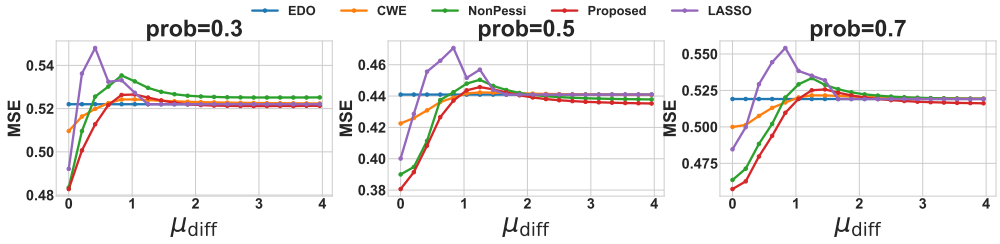

Figure 11: Empirical means of MSEs of various methods with $m = 1, \delta = 1$ in Example 5.2 in different prob scales.

Figure 12 examines the performance of various methods as sample size increases, focusing on the t-normal noise setting. As expected, MSEs decrease with larger sample sizes due to reduced variance. Notably, our method consistently performs well across all subplots and levels of μ_{diff}.

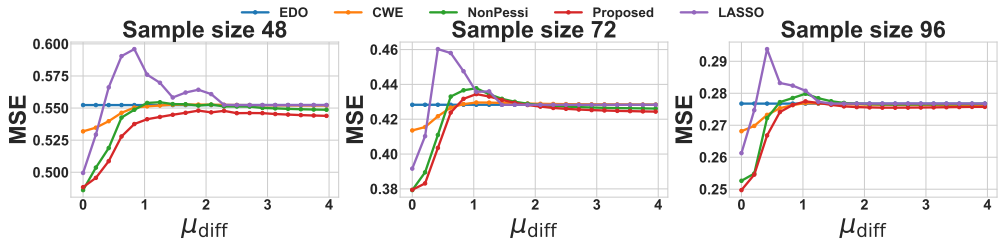

Figure 12: Empirical means of MSEs of various methods with $m = 1, \delta = 1$ in Example 5.2 across $|\mathcal{D}_e| = 48, 72, 96$.

Example A.1 (Clinical-data-based simulation). In this section, we construct a simulation environment based on the public real-world dataset ACTG175, which consists of 2,139 HIV-positive individuals randomized to four treatments. We focus on comparing ZDV+ddI ($n = 522$) and ZDV+zal ($n = 524$), treating them as actions 1 and 0, respectively. The outcome of interest is the rescaled CD4 count, and we consider three covariates: age (S_1), homosexual activity (S_2), and hemophilia (S_3). We construct a simulator similar to Example 5.2, and generate both experimental and historical data for evaluation.

Specifically, the outcome model is specified as:

$$R = f(S) = \beta_0 + \beta_1 S_1^2 + \beta_2 S_1 + \beta_3 S_2 + \beta_4 S_3 + \gamma A,$$

We use the data to fit the model.Using fitted parameters, we generate synthetic outcomes based on real data:

$$R_e = f(S_e) + 0.8\delta\epsilon_e, \quad R_h = f(S_h) + \mu_d + 0.05\mu_d S_{h,1} + 0.8\delta\epsilon_h,$$

where ϵ_e is drawn from $\mathcal{N}(0,1)$ and ϵ_h from either $\mathcal{N}(0,1)$ or the heavy-tailed t_6 distribution; alternatively, ϵ_e is drawn from t_9 and ϵ_h from either $\mathcal{N}(0,1)$ or t_9. This setting yields four possible scenarios, corresponding to all combinations of ϵ_e and ϵ_h being sampled from a standard normal or from a heavy-tailed distribution. Here, S_1' represents covariates generated from a normal distribution fitted using the empirical mean and variance of all observed S_1. Here, S_2' and S_3' are sampled from Bernoulli distributions with parameters set to the empirical means of S_2 and S_3, respectively. We estimate the variance parameter δ as the average of the squared residuals from the fitted model. In experimental data, treatment assignments are randomized ($A \sim \text{Bernoulli}(0.5)$), whereas in historical data they are fixed at $A = 0$. Both datasets contain 48 samples. We vary $\mu_d \in [0, 5]$ (25 points) and compare the empirical average MSE over 100 simulations. Results in Figure 13 reveal that the proposed estimator achieves the lowest MSEs in most cases.

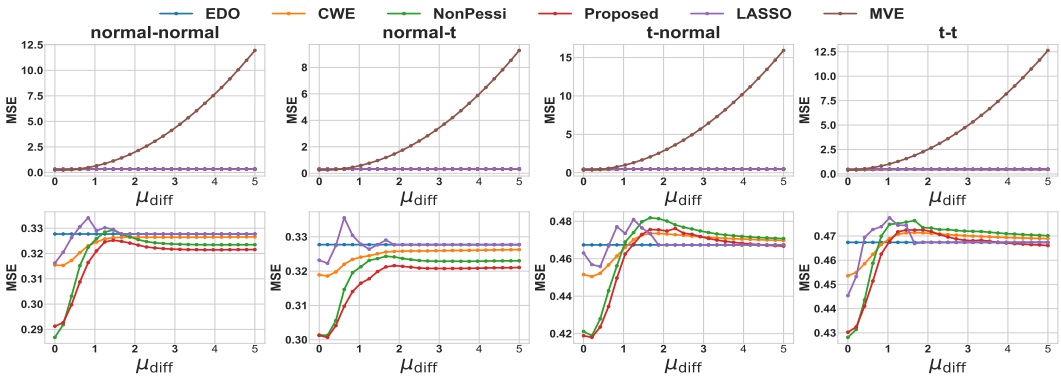

Figure 13: MSEs of ATE estimators in Clinical example. Top: all estimators; Bottom: exclude MVE.

We also consider the following hypothesis testing problem:

$$H_0 : \text{ATE} = 0 \quad \text{vs.} \quad H_1 : \text{ATE} \neq 0.$$

To conduct valid inference, we estimate the ATE using doubly robust procedures and compute p-values using a Wald test. To compare the proposed method against EDO and Pessi, we first set the true ATE to 0 and conduct 1,000 Monte Carlo simulations to estimate the Type I error rate of each method at the 5% significance level. To assess empirical power, we gradually increase the signal strength by setting the ATE to $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. We further examine three levels of posterior shifts: small, moderate, and large. Table 3 reports the empirical rejection rates. Under the null where ATE equals zero, all three methods control Type I error. Under the alternative, the proposed test is more powerful than the two competitors.

Table 3: Empirical Type-I error rates and power under different bias levels. The best results in each row are highlighted in bold.

ATE	Metric	Small Posterior Shifts			Moderate Posterior Shifts			Large Posterior Shifts		
		EDO	CWE	Proposed	EDO	CWE	Proposed	EDO	CWE	Proposed
0.0	Type-I Error	0.030	0.032	**0.024**	0.030	0.035	**0.029**	0.030	0.030	**0.027**
0.5	Power	0.081	**0.097**	0.093	0.081	0.103	**0.114**	0.081	0.084	**0.088**
1.0	Power	0.267	0.310	**0.320**	0.267	0.311	**0.335**	0.267	0.276	**0.286**
1.5	Power	0.546	0.592	**0.611**	0.546	0.586	**0.613**	0.546	0.558	**0.572**
2.0	Power	0.810	0.847	**0.867**	0.810	0.835	**0.863**	0.810	0.815	**0.825**
2.5	Power	0.955	0.967	**0.974**	0.955	0.965	**0.970**	0.955	0.960	**0.962**
3.0	Power	0.994	0.996	**0.997**	0.994	0.996	**0.997**	0.994	0.994	**0.994**

B Implementation Details

In this section, we provide detailed procedures for implementing our proposal based on importance sampling estimator (Section B.1) and doubly robust estimator (Section B.2).

B.1 Importance Sampling Method

Recall that $\psi_a^{(e)}(O^{(e)}) = \mathbb{I}(A = a)R^{(e)}/\pi(a|S^{(e)})$, and $\psi^{(h)}(O^{(h)}) = \mu(S^{(h)})R^{(h)}$. The proposed weighted ATE estimator is given by (4), and can be obtained as follows.

- Step 1: estimate $\psi_a^{(e)}(O^{(e)})$ and $\psi^{(h)}(O^{(h)})$ using the methodology detailed in B.5 and B.6
- Step 2: Estimate $w(S)$ using Section B.7 by minimizing the upper bound of the estimated MSE derived in B.4.
- Step 3: Obtain the weighted ATE estimator by plugging in the unknown terms with their estimates in (4).

To simplify notation, we define

$$Z^{(e)}(w) = \psi_1^{(e)}(O^{(e)}) - w(S^{(e)})\psi_0^{(e)}(O^{(e)}), \qquad Z^{(h)}(w) = (1 - w(S^{(h)}))\psi^{(h)}(O^{(h)}). \tag{9}$$

The weighted ATE estimator for a given weight function w can thus be written as

$$\widehat{\mathrm{ATE}}(w) = \mathbb{E}_n(Z^{(e)}(w)) - \mathbb{E}_n(Z^{(h)}(w)).$$

B.2 Doubly Robust Method

We present two baseline DR estimators before introducing the proposed weighted DR estimator. The first baseline estimator is constructed using only experimental data,

$$\tau_{dr}^{(e)} = \frac{1}{|\mathcal{D}_e|} \sum_{O_e \in \mathcal{D}_e} [\psi_{dr,1}^{(e)}(O^{(e)}) - \psi_{dr,0}^{(e)}(O^{(e)})],$$

where the estimating function $\psi_{\mathrm{dr},a}^{(e)}$ for $a \in \{0, 1\}$ is given by

$$\psi_{\mathrm{dr},a}^{(e)}(O^{(e)}) = \frac{\mathbb{I}(A^{(e)} = a)}{\pi(a \mid S^{(e)})} \left[R^{(e)} - r^{(e)}(A^{(e)}, S^{(e)})\right] + r^{(e)}(a, S^{(e)}).$$

It can be shown that $\psi^{(e)}_{\mathrm{dr},a}(O^{(e)})$ is unbiased to the mean outcome under treatment a as long as either the propensity score model π or the outcome model $r^{(e)}$ is correctly specified. This yields the doubly robustness property.

The second estimator incorporates historical data into the ATE estimation and is defined as

$$\tau^{(h)}_{\mathrm{dr}} = \frac{1}{|\mathcal{D}^{(e)}|}\sum_{O^{(e)}\in\mathcal{D}^{(e)}}\psi^{(e)}_{\mathrm{dr},1}(O^{(e)}) - \frac{1}{|\mathcal{D}^{(e)}|}\sum_{S^{(e)}\in\mathcal{D}^{(e)}} r^{(h)}(0,S^{(e)}) - \frac{1}{|\mathcal{D}^{(h)}|}\sum_{O^{(h)}\in\mathcal{D}^{(h)}}\psi^{(h)}_{\mathrm{dr}}(O^{(h)}),$$

where the estimating function $\psi^{(h)}_{\mathrm{dr}}(O^{(h)})$ is given by

$$\psi^{(h)}_{\mathrm{dr}}(O^{(h)}) = \mu(S^{(h)})\left[R^{(h)} - r^{(h)}(0,S^{(h)})\right].$$

Next, given a weight function w, we define the weighted DR estimator as:

$$\begin{aligned}\widehat{\mathrm{ATE}}_{\mathrm{dr}}(w) &= \frac{1}{|\mathcal{D}^{(e)}|}\sum_{O^{(e)}\in\mathcal{D}^{(e)}}\psi^{(e)}_{\mathrm{dr},1}(O^{(e)}) - \frac{1}{|\mathcal{D}^{(e)}|}\sum_{O^{(e)}\in\mathcal{D}^{(e)}} w(S^{(e)})\psi^{(e)}_{\mathrm{dr},0}(O^{(e)}) \\ &- \frac{1}{|\mathcal{D}^{(e)}|}\sum_{S^{(e)}\in\mathcal{D}^{(e)}}\left(1-w(S^{(e)})\right) r^{(h)}(0,S^{(e)}) \\ &- \frac{1}{|\mathcal{D}^{(h)}|}\sum_{O^{(h)}\in\mathcal{D}^{(h)}}\left(1-w(S^{(h)})\right)\psi^{(h)}_{\mathrm{dr}}(O^{(h)}). \end{aligned} \tag{10}$$

Similar to IS, we define $Z^{(e)}_{\mathrm{dr}}(w)$ and $Z^{(h)}_{\mathrm{dr}}(w)$ as

$$Z^{(e)}_{dr}(w) = \psi^{(e)}_{\mathrm{dr},1}(O^{(e)}) - w(S^{(e)})\psi^{(e)}_{\mathrm{dr},0}(O^{(e)}) - (1-w(S^{(e)}))r^{(h)}(0,S^{(e)}) \tag{11}$$

$$Z^{(h)}_{dr}(w) = (1-w(S^{(h)}))\psi^{(h)}_{dr}(O^{(h)}). \tag{12}$$

Using these notations, the final ATE estimator is given by

$$\widehat{\mathrm{ATE}_{dr}}(w) = \mathbb{E}_n(Z^{(e)}_{dr}(w)) - \mathbb{E}_n(Z^{(h)}_{dr}(w)).$$

To summarize, our DR ATE estimator can be constructed through the following steps.

- Step 1: estimate $\psi^{(e)}_{\mathrm{dr},a}(O^{(e)})$ for $a \in \{0,1\}$, $r^{(h)}(0,S^{(e)})$and $\psi^{(h)}_{\mathrm{dr}}(O^{(h)})$ using the method in B.5and B.6.
- Step 2: Estimate $w(S)$ using Section B.7 by minimizing the upper bounded of the estimated MSE B.4.
- Step 3: Obtain the weighted ATE estimator by plugging in the unknown terms with their estimates in (10).

B.3 Explicit Form of $\mathrm{Var}(w)$ and its Estimator $\widehat{\mathrm{Var}}(w)$

In this part, we present the detailed expression for the variance term $\mathrm{Var}(w)$ in (5), and its estimator $\widehat{\mathrm{Var}}(w)$. Since the experimental and historical dataset are mutually independent, and samples within each dataset are independently and identically distributed, $\mathrm{Var}(w)$ can be written as:

$$\mathrm{Var}(w) = \mathrm{Var}(\widehat{\mathrm{ATE}}(w)) = \frac{\mathrm{Var}(Z^{(e)}(w))}{|\mathcal{D}^{(e)}|} + \frac{\mathrm{Var}(Z^{(h)}(w))}{|\mathcal{D}^{(h)}|}, \tag{13}$$

Therefore, to estimate $\mathrm{Var}(w)$, it suffices to estimate $\mathrm{Var}(Z^{(e)}(w))$ and $\mathrm{Var}(Z^{(h)}(w))$.

Their estimators $\widehat{\mathrm{Var}}(Z^{(e)}(w))$ and $\widehat{\mathrm{Var}}(Z^{(h)}(w))$ can be obtained using the standard sample variance formula,

$$\widehat{\mathrm{Var}}(Z^{(e)}(w)) := \frac{1}{|\mathcal{D}^{(e)}|-1}\sum_{i=1}^{|\mathcal{D}^{(e)}|}\left(Z^{(e)}_i(w) - \mathbb{E}_n[Z^{(e)}(w)]\right)^2,$$

$$\widehat{\mathrm{Var}}(Z^{(h)}(w)) := \frac{1}{|\mathcal{D}^{(h)}|-1}\sum_{j=1}^{|\mathcal{D}^{(h)}|}\left(Z^{(h)}_j(w) - \mathbb{E}_n[Z^{(h)}(w)]\right)^2,$$

where $\mathbb{E}_n[Z^{(e)}(w)]$ and $\mathbb{E}_n[Z^{(h)}(w)]$ are defined as:

$$\mathbb{E}_n[Z^{(e)}(w)] = \frac{1}{|\mathcal{D}^{(e)}|}\sum_{i=1}^{|\mathcal{D}^{(e)}|} Z_i^{(e)}(w), \quad \text{and} \quad \mathbb{E}_n[Z^{(h)}(w)] = \frac{1}{|\mathcal{D}^{(h)}|}\sum_{j=1}^{|\mathcal{D}^{(h)}|} Z_j^{(h)}(w).$$

This yields the variance estimator

$$\widehat{\mathrm{Var}}(w) := \frac{\widehat{\mathrm{Var}}(Z^{(e)}(w))}{|\mathcal{D}^{(e)}|} + \frac{\widehat{\mathrm{Var}}(Z^{(h)}(w))}{|\mathcal{D}^{(h)}|}. \tag{14}$$

B.4 Explicit Form of $\widehat{\mathrm{MSE}}_U(w)$

In this section, we derive $\widehat{\mathrm{MSE}}_U(w)$ in (6). It consists of the two terms: $\widehat{\mathrm{Var}}_U(w)$ and $\widehat{\mathrm{bias}}_U(w)$. We seek these two terms that satisfy the coverage probability in Assumption 1. In what follows, we take the bias term as an example and present three approaches for its construction: one based on empirical process theory [128], another based on Markov inequality and Bonferroni's inequality, and a third based on the multiplier bootstrap [129].

For the first one, define the following function classes:

$$\mathcal{F}^{(e)} = \left\{ f_w^{(e)}(o) := (1 - w(s))\psi^{(e)}(o) : w \in \mathcal{W} \right\},$$

$$\mathcal{F}^{(h)} = \left\{ f_w^{(h)}(o) := (1 - w(s))\psi^{(h)}(o) : w \in \mathcal{W} \right\}.$$

Then, $\widehat{\mathrm{bias}}(w) - \mathrm{bias}(w)$ is the difference between two empirical processes indexed by w. If the function class $\mathcal{W}$ satisfies certain complexity properties (e.g., being a VC class), one can apply empirical process theory to construct a uniform upper bound U such that

$$\mathbb{P}\left(\sup_{w \in \mathcal{W}} |\widehat{\mathrm{bias}}(w) - \mathrm{bias}(w)| \leq U \right) \geq 1 - \alpha, \tag{15}$$

or

$$\mathbb{P}\left(\sup_{w \in \mathcal{W}} \frac{|\widehat{\mathrm{bias}}(w) - \mathrm{bias}(w)|}{\widehat{\sigma}(w)} \leq U \right) \geq 1 - \alpha, \tag{16}$$

where $\widehat{\sigma}(w)$ is a consistent estimator of the asymptotic variance of $\widehat{\mathrm{bias}}(w)$. Accordingly, we can set:

- $\widehat{\mathrm{bias}}_U(w) = U + |\widehat{\mathrm{bias}}(w)|$ in the unnormalized case;
- $\widehat{\mathrm{bias}}_U(w) = \widehat{\sigma}(w)U + |\widehat{\mathrm{bias}}(w)|$ in the normalized case.

More specifically, when the error terms $\epsilon^{(e)}$, $\epsilon^{(h)}$ are sub-Gaussian, we may set:

$$U = \frac{c r_{\max}}{\epsilon}\sqrt{\frac{v \log n_{\max} + \log(1/\alpha)}{n_{\min}}},$$

where $c > 0$ is a constant, v denotes the VC dimension of the function class $\mathcal{W}$, $n_{\max} = \max(|\mathcal{D}^{(e)}|, |\mathcal{D}^{(h)}|)$ and $n_{\min} = \min(|\mathcal{D}^{(e)}|, |\mathcal{D}^{(h)}|)$. This can also be extended to heavy-tailed errors [see e.g., 129, Section 5].

Alternatively, under the following finite-hypothesis-class assumption, $\widehat{\mathrm{bias}}_U$ and $\widehat{\mathrm{Var}}_U$ can be constructed based on asymptotic normality and Bonferroni's inequality.

Assumption 6 (Finite hypothesis class)**.** *The number of elements in $\mathcal{W}$ is finite.*

Assumption 6 is commonly employed in machine learning to simplify the theoretical analysis [see e.g., 123].

Assumption 7. *The error terms $\epsilon_0^{(e)}$, $\epsilon_1^{(e)}$ and ϵ_h are assumed to have finite eighth moments.*

Assumption 7 is standard in high-order moment analysis and is commonly adopted in the literature on finite-sample concentration .The class of distributions with a finite eighth moment is very broad. We define $\widehat{\mathrm{Var}}_U\big(Z^{(e)}(w_i)\big) = \widehat{\mathrm{Var}}\big(Z^{(e)}(w_i)\big) + U^{(e)}(w_i)$ as the upper bound for $\mathrm{Var}\big(Z^{(h)}(w_i)\big)$ and $\widehat{\mathrm{Var}}_U\big(Z^{(h)}(w_i)\big) = \widehat{\mathrm{Var}}\big(Z^{(h)}(w_i)\big) + U^{(h)}(w_i)$ as the upper bound for $\mathrm{Var}\big(Z^{(e)}(w_i)\big)$. We consider a set $\mathcal{W}$ with K elements. Since the samples are independent and identically distributed (i.i.d.) and the experimental dataset $\mathcal{D}^{(e)}$ is independent of the historical dataset $\mathcal{D}^{(h)}$, we have:

$$\begin{aligned}
\widehat{\mathrm{Var}}_U(w_i) &= \frac{\widehat{\mathrm{Var}}_U\big(Z^{(e)}(w_i)\big)}{|\mathcal{D}^{(e)}|} + \frac{\widehat{\mathrm{Var}}_U\big(Z^{(h)}(w_i)\big)}{|\mathcal{D}^{(h)}|} \\
&= \frac{\widehat{\mathrm{Var}}\big(Z^{(e)}(w_i)\big) + U^{(e)}(w_i)}{|\mathcal{D}^{(e)}|} + \frac{\widehat{\mathrm{Var}}\big(Z^{(h)}(w_i)\big) + U^{(h)}(w_i)}{|\mathcal{D}^{(h)}|} \\
&= \widehat{\mathrm{Var}}(w_i) + \frac{U^{(e)}(w_i)}{|\mathcal{D}^{(e)}|} + \frac{U^{(h)}(w_i)}{|\mathcal{D}^{(h)}|}.
\end{aligned}$$

Assume there exist positive constants $\alpha_1, \alpha_2, \alpha_3, \alpha_4 > 0$, and define $\alpha_{\mathrm{var}} := \alpha_1 + \alpha_2$ and $\alpha_{\mathrm{bias}} := \alpha_3 + \alpha_4$. For each $i \in \{1, \ldots, K\}$, we construct a confidence interval via Markov's inequality applied to the fourth moment:

$$\mathbb{P}\left(\mathrm{Var}\Big(Z^{(e)}(w_i)\Big) \geq \widehat{\mathrm{Var}_{\mathrm{U}}}\Big(Z^{(e)}(w_i)\Big)\right) \leq \frac{\mathbb{E}\left[\Big(\mathrm{Var}\big(Z^{(e)}(w_i)\big) - \widehat{\mathrm{Var}}\big(Z^{(e)}(w_i)\big)\Big)^4\right]}{U^{(e)}(w_i)^4} = \frac{\alpha_1}{K}. \quad (17)$$

$$\mathbb{P}\left(\mathrm{Var}\Big(Z^{(h)}(w_i)\Big) \geq \widehat{\mathrm{Var}}_U\Big(Z^{(h)}(w_i)\Big)\right) \leq \frac{\mathbb{E}\left[\Big(\mathrm{Var}\big(Z^{(h)}(w_i)\big) - \widehat{\mathrm{Var}}\big(Z^{(h)}(w_i)\big)\Big)^4\right]}{U^{(h)}(w_i)^4} = \frac{\alpha_2}{K}. \quad (18)$$

Direct calculations lead to:

$$\begin{aligned}
& \mathbb{P}\Big(\mathrm{Var}(w_i) \geq \widehat{\mathrm{Var}}_U(w_i)\Big) \\
\leq\ & \mathbb{P}\Big(\mathrm{Var}\Big(Z^{(e)}(w_i)\Big) \geq \widehat{\mathrm{Var}}\Big(Z^{(e)}(w_i)\Big) + U^{(e)}(w_i)\Big) \\
& \bigcup \mathrm{Var}\Big(Z^{(h)}(w_i)\Big) \geq \widehat{\mathrm{Var}}\Big(Z^{(h)}(w_i)\Big) + U^{(h)}(w_i)\Big) \\
\leq\ & \mathbb{P}\Big(\mathrm{Var}\Big(Z^{(e)}(w_i)\Big) \geq \widehat{\mathrm{Var}}\Big(Z^{(e)}(w_i)\Big) + U^{(e)}(w_i)\Big) \\
+\ & \mathbb{P}\Big(\mathrm{Var}\Big(Z^{(h)}(w_i)\Big) \geq \widehat{\mathrm{Var}}\Big(Z^{(h)}(w_i)\Big) + U^{(h)}(w_i)\Big) \\
\leq\ & \frac{\alpha_1}{K} + \frac{\alpha_2}{K} = \frac{\alpha_{var}}{K},
\end{aligned} \quad (19)$$

where the first inequality follows from the relationship between the sets, and the second inequality follows from the probability of the union bound.

We can write $\widehat{\mathrm{bias}}_U(w_i)$ as:

$$\widehat{\mathrm{bias}}_U(w_i) = |\widehat{\mathrm{bias}}(w_i)| + \Big(U_b^{(e)}(w_i) + U_b^{(h)}(w_i)\Big) \quad (20)$$

This bound can also be derived via the fourth-moment version of Markov's inequality,

$$\begin{aligned}
& \mathbb{P}\left(|\mathrm{bias}(w_i) - \widehat{\mathrm{bias}}(w_i)| \le U_b^{(e)}(w_i) + U_b^{(h}(w_i)\right) \\
\le\ & \frac{\mathbb{E}\left[\left((1 - w(S^{(e)})\psi_0^{(e)}(O^{(e)}) - \mathbb{E}(1 - w(S^{(e)})\psi_0^{(e)}(O^{(e)})\right)^4\right]}{U_b^{(e}(w_i)^4} \\
+\ & \frac{\mathbb{E}\left[\left((1 - w(S^{(h)})\psi_0^{(h)}(O^{(h)}) - \mathbb{E}(1 - w(S^{(h)})\psi_0^{(h)}(O^{(h)})\right)^4\right]}{U_b^{(h}(w_i)^4} \\
=\ & \frac{\alpha_3}{K} + \frac{\alpha_4}{K} = \frac{\alpha_{bias}}{K}.
\end{aligned} \tag{21}$$

According to Bonferroni's inequality, we have

$$P\left(\bigcap_{i=1}^{K}\left\{\widehat{\mathrm{bias}}_U(w_i) \ge |\mathrm{bias}(w_i)|\right\}\right) \ge 1 - \sum_{w=1}^{K} P\left(\widehat{\mathrm{bias}}_U(w_i) < |\mathrm{bias}(w_i)|\right) \ge 1 - \alpha_{bias}. \tag{22}$$

$$P\left(\bigcap_{i=1}^{K}\left\{\widehat{\mathrm{Var}}_U(w_i) \ge \mathrm{Var}(w_i)\right\}\right) \ge 1 - \sum_{w=1}^{K} P\left(\widehat{\mathrm{Var}}_U(w_i) < \mathrm{Var}(w_i)\right) \ge 1 - \alpha_{Var}. \tag{23}$$

Then, combining (22) and (23) and setting $\alpha = \alpha_{\mathrm{var}} + \alpha_{\mathrm{bias}}$, we obtain the claim. This is the procedure we adopt to construct $\widehat{\mathrm{bias}}_U$ and $\widehat{\mathrm{Var}}_U$ in Corollaries 1–4

Finally, one may use the high-dimensional multiplier bootstrap to construct $\widehat{\mathrm{bias}}_U$ [129, 130]. Under mild regularity conditions, the distribution of the supremum in (15) or (16) converges to that of a Gaussian process [129, Theorem 2.1]. Furthermore, the supremum of this Gaussian process can be approximated via the multiplier bootstrap, which enables us to set to set U to the α-quantile of the supremum of the following bootstrapped process

$$\sup_{w\in\mathcal{W}}\left[\frac{1}{|\mathcal{D}^{(e)}|}\sum_{i=1}^{|\mathcal{D}^{(e)}|}(1 - w(S^{(e)}))\psi^{(e)}(O_i^{(e)})g_i - \frac{1}{|\mathcal{D}^{(h)}|}\sum_{j=1}^{|\mathcal{D}^{(h)}|}(1 - w(S^{(h)}))\psi^{(h)}(O_j^{(h)})g_{|\mathcal{D}^{(e)}|+j}\right],$$

in the unnormalized case, and

$$\sup_{w\in\mathcal{W}}\frac{1}{\widehat{\sigma}(w)}\left[\frac{1}{|\mathcal{D}^{(e)}|}\sum_{i=1}^{|\mathcal{D}^{(e)}|}(1 - w(S^{(e)}))\psi^{(e)}(O_i^{(e)})g_i - \frac{1}{|\mathcal{D}^{(h)}|}\sum_{j=1}^{|\mathcal{D}^{(h)}|}(1 - w(S^{(h)}))\psi^{(h)}(O_j^{(h)})g_{|\mathcal{D}^{(e)}|+j}\right],$$

in the normalized case, where g_is are i.i.d. standard Gaussian variables. Repeating the process over multiple bootstrap samples provides an empirical estimate of the quantile.

Similar to the first approach, the uniform bound can then be defined as:

$$\widehat{\mathrm{bias}}_U(w) = U + |\widehat{\mathrm{bias}}(w)| \quad \text{or} \quad \widehat{\mathrm{bias}}_U(w) = U + \widehat{\sigma}(w)|\widehat{\mathrm{bias}}(w)|.$$

As for the double robust estimator, to obtain $\widehat{\mathrm{Var}}_U(w)$, we only need to replace $Z^{(e)}(w)$ and $Z^{(h)}(w)$ with $Z_{\mathrm{dr}}^{(e)}(w)$ and $Z_{\mathrm{dr}}^{(h)}(w)$. The bias upper bound can be similarly established.

In our implementation, we find that the uniform upper bound over $w \in \mathcal{W}$ tends to be overly conservative. Therefore, instead of enforcing a global bound, we adopt pointwise, non-uniform upper bounds for $\widehat{\mathrm{Var}}_U(w)$ and $\widehat{\mathrm{bias}}_U(w)$, computed individually at each w based on normal approximation.

B.5 Estimation of Nuisance Function

Accurate estimation of the propensity score $\pi(a \mid S^{(e)})$ is essential for both DR and IS. In our implementation, we estimate this nuisance function via logistic regression. The outcome functions $r^{(h)}(0, S^{(h)})$, $r^{(e)}(0, S^{(e)})$, and $r^{(e)}(1, S^{(e)})$ can be flexibly estimated using a variety of regression models, including basis function expansions, random forests, and neural networks.

B.6 Estimation of $\mu(S)$

To estimate the density ratio $\mu(S)$, we adopt a moment-matching approach. Specifically, we seek a function $\mu(S)$ such that the following moment conditions are satisfied:

$$\mathbb{E}[\mu(S^{(e)})\,\Phi_k(S^{(e)})] = \mathbb{E}[\Phi_k(S^{(h)})], \quad \text{for } k = 1, \ldots, K,$$

where $\Phi_k(S)$ denotes the k-th test function.

Let $\Phi(S) = [\Phi_1(S), \ldots, \Phi_K(S)]^\top$ be the corresponding feature map. We approximate the density ratio $\mu(S)$ by a linear model of the form $\mu(S) \approx \Phi(S)^\top \gamma$, and estimate the coefficient vector $\gamma \in \mathbb{R}^K$ by solving a sample moment-matching equation between the historical and experimental datasets.

$$\frac{1}{|\mathcal{D}_h|} \sum_{S_i^{(h)} \in \mathcal{D}_h} \Phi(S_i^{(h)}) \Phi(S_i^{(h)})^\top \gamma = \frac{1}{|\mathcal{D}_e|} \sum_{S_j^{(e)} \in \mathcal{D}_e} \Phi(S_j^{(e)}).$$

In practice, the feature function $\Phi(\bullet)$ can be set to polynomials, splines, or neural network features. Under mild or negligible covariate shift, the density ratio can be simplified to 1.

B.7 Estimation of the weight function $w(S)$

In our implementation, we parameterized $w(S)$ using a logistic model,

$$w(S) = \frac{1}{1 + e^{-\theta^\top S}},$$

which ensures that $w(S) \in (0, 1)$ for all S. Alternatively, a neural network can be employed. Given a parameterized w, the pessimistic objective function derived in Section B.4 can be optimized using gradient-based methods to learn the parameters.

C Proofs of the Theorems and Corollaries

In this section, we present the proofs of Theorem 1 and Corollaries 1 – 4. Recall that Theorem 1 applies to any ATE estimator, while Corollaries 1 – 4 are specific to the IS estimator.

C.1 Proof of Theorem 1

Proof of Theorem 1. To facilitate the analysis, we define the following events:

$$A := \bigcap_{w \in \mathcal{W}} \left\{ \widehat{\text{bias}}_U(w) \geq |\text{bias}(w)| \right\}, \quad B := \bigcap_{w \in \mathcal{W}} \left\{ \widehat{\text{Var}}_U(w) \geq \text{Var}(w) \right\}, \quad C := A \cap B.$$

For a given w, we define $\text{MSE}(w)$ as the MSE of the ATE estimator $\widehat{\text{ATE}}(w)$. Since the estimated weight $\widehat{w}$ itself is random, $\text{MSE}(\widehat{w})$ is a random variable. This MSE is well-defined due to the use of sample splitting which ensures that the ATE estimator is independent of $\widehat{w}$. The MSE of our proposed estimator is given by the expected value of $\text{MSE}(\widehat{w})$,

$$\text{MSE}\big(\widehat{\text{ATE}}(\widehat{w})\big) = \mathbb{E}[\text{MSE}(\widehat{w})]\,,$$

where the expectation on the right-hand-side (RHS) is taken with respect to the randomness of $\widehat{w}$.

We decompose the difference in MSE into two parts:

$$\begin{aligned} \text{MSE}\big(\widehat{\text{ATE}}(\widehat{w})\big) - \text{MSE}\big(\widehat{\text{ATE}}(w)\big) = &\underbrace{\mathbb{E}\left[(\text{MSE}(\widehat{w}) - \text{MSE}(w)) \cdot 1_C\right]}_{M_1} \\ &+ \underbrace{\mathbb{E}\left[(\text{MSE}(\widehat{w}) - \text{MSE}(w)) \cdot 1_{C^c}\right]}_{M_2}. \end{aligned} \tag{24}$$

Bounding M_1: Under Assumption 1, on event C, we have:

$$\text{bias}^2(\widehat{w}) + \text{Var}(\widehat{w}) \leq \widehat{\text{bias}}_U^2(\widehat{w}) + \widehat{\text{Var}}_U(\widehat{w}).$$

Moreover, since $\widehat{w}$ minimizes the pessimistic objective defined in (6), it follows that

$$\widehat{\mathrm{bias}}_U^2(\widehat{w}) + \widehat{\mathrm{Var}}_U(\widehat{w}) \leq \widehat{\mathrm{bias}}_U^2(w) + \widehat{\mathrm{Var}}_U(w).$$

Thus, M_1 can be bounded as follows:

$$\begin{aligned} M_1 = \quad & \mathbb{E}\left[\left(\mathrm{bias}^2(\widehat{w}) + \mathrm{Var}(\widehat{w}) - \mathrm{bias}^2(w) - \mathrm{Var}(w)\right) \cdot 1_C\right] \\ \leq \quad & \mathbb{E}\left[\left(\widehat{\mathrm{bias}}_U^2(w) - \mathrm{bias}^2(w) + \widehat{\mathrm{Var}}_U(w) - \mathrm{Var}(w)\right) \cdot 1_C\right] \\ \leq \quad & \mathbb{E}\left[\widehat{\mathrm{bias}}_U^2(w) - \mathrm{bias}^2(w)\right] + \mathbb{E}\left[\widehat{\mathrm{Var}}_U(w) - \mathrm{Var}(w)\right]. \end{aligned} \tag{25}$$

Bounding M_2: We next bound the term associated with the complement event C^c. By the definition of the mean squared error (MSE), we have:

$$\mathrm{MSE}(\widehat{\mathrm{ATE}}(\widehat{w})) = \mathbb{E}\left[\left(\widehat{\mathrm{ATE}}(\widehat{w}) - \mathrm{ATE}\right)^2\right] \leq \mathbb{E}\left[\left(|\widehat{\mathrm{ATE}}(\widehat{w})| + |\mathrm{ATE}|\right)^2\right] = O(B^2),$$

where the first inequality follows from the triangle inequality, and the second follows from Assumption 2. Similarly, we have $\mathrm{MSE}(w) \leq O(B^2)$ for any $w \in \mathcal{W}$.

Using this and the union bound on probabilities:

$$\mathbb{P}(C^c) = \mathbb{P}(A^c \cup B^c) \leq \mathbb{P}(A^c) + \mathbb{P}(B^c) \leq 2\alpha.$$

Hence, we bound term M_2 as:

$$M_2 = \mathbb{E}\left[(\mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(w)) \cdot 1_{C^c}\right] \leq 2B^2 \cdot \mathbb{P}(C^c) \leq O(\alpha B^2). \tag{26}$$

Plugging the bounds for terms M_1 in (25) and M_2 in (26) into (24), the result follows. □

C.2 Supporting Lemmas

Lemma 1 (MSE decomposition)**.** *For a given weight function w, the MSE of the weighted estimator can be decomposed as:*

$$\mathrm{MSE}(w) = \frac{\mathrm{Var}(Z^{(e)}(w))}{|\mathcal{D}^{(e)}|} + \frac{\mathrm{Var}(Z^{(h)}(w))}{|\mathcal{D}^{(h)}|} + \left(\mathbb{E}\left[(1 - w(S^{(e)}))b(S^{(e)})\right]\right)^2. \tag{27}$$

We interpret the three terms on the right-hand-side of (27) *as follows: (1) variance from the experimental data, (2) variance from the historical data, and (3) the squared bias introduced by incorporating historical data.*

Proof of Lemma 1. We decompose the MSE into the sum of variance and squared bias $\mathrm{MSE}(w) = \mathrm{Var}(w) + \mathrm{bias}^2(w)$. We have already derived the closed-form expression of the variance term in (13). As for the bias, we note that:

$$\begin{aligned} & \mathbb{E}[(1 - w(S^{(e)}))\psi_0^{(e)}(O^{(e)})] - \mathbb{E}[(1 - w(S^{(h)}))\psi^{(h)}(O^{(h)})] \\ =& \mathbb{E}\left[(1 - w(S^{(e)})) \cdot \mathbb{E}[(r^{(e)}(0, S^{(e)}) + \epsilon_0^{(e)} \mid S^{(e)}]\right] \\ -& \mathbb{E}\left[(1 - w(S^{(h)}) \cdot \frac{p_e(S^{(e)})}{p_h(S^{(h)})} \cdot \mathbb{E}[r^{(h)}(0, S^{(h)}) + \epsilon^{(h)} \mid S^{(h)}]\right] \\ =& \mathbb{E}\left[(1 - w(S^{(e)}))(r^{(e)}(0, S^{(e)}) - r^{(h)}(0, S^{(e)}))\right] = \mathbb{E}\left[(w(S^{(e)}) - 1) \cdot b(S^{(e)})\right]. \end{aligned} \tag{28}$$

The second equality follows from Assumption 4 that $\epsilon_0^{(e)}$ is independent of $S^{(e)}$ with $\mathbb{E}[\epsilon_0^{(e)} \mid S^{(e)}] = 0$, and that $\epsilon^{(h)}$ is independent of $S^{(h)}$ with $\mathbb{E}[\epsilon^{(h)} \mid S^{(h)}] = 0$. Combining (13) and (28) completes the proof. □

Lemma 2 (Variance from Experiment Data)**.** *For a given weight function w, the variance of $Z^{(e)}(w)$ defined in* (9) *is given by*

$$\begin{aligned} \mathrm{Var}\left(Z^{(e)}(w)\right) =& \mathbb{E}\left[\frac{r^{(e)}(1, S^{(e)})^2 + (\sigma_1^{(e)})^2}{\pi(1 \mid S^{(e)})}\right] + \mathbb{E}\left[\frac{w(S^{(e)})^2(r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2)}{\pi(0 \mid S^{(e)})}\right] \\ & - \left(\mathbb{E}[r^{(e)}(1, S^{(e)})] - \mathbb{E}[w(S^{(e)})r^{(e)}(0, S^{(e)})]\right)^2. \end{aligned} \tag{29}$$

Proof. Direct calculation leads to

$$\begin{aligned}\mathrm{Var}\left(Z^{(e)}(w)\right) &= \mathrm{Var}(\psi_1^{(e)}(O^{(e)})) + \mathrm{Var}(w(S^{(e)})\psi_0^{(e)}(O^{(e)}) \\ &- 2\mathrm{Cov}(\psi_1^{(e)}(O^{(e)}), w(S^{(e)})\psi_0^{(e)}(O^{(e)})).\end{aligned}$$

We proceed to compute each term on the RHS, respectively.

1. Variance of $\psi_1^{(e)}(O^{(e)})$: We use the law of total variance:

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid S)] + \mathrm{Var}(\mathbb{E}[X \mid S]).$$

In our case, $X = \psi_1^{(e)}(O^{(e)})$. According to Assumption 4, we have:

$$\mathbb{E}[\psi_1^{(e)}(O^{(e)}) \mid S^{(e)}] = r^{(e)}(1, S^{(e)}), \quad \mathbb{E}[(\psi_1^{(e)}(O^{(e)}))^2 \mid S^{(e)}] = \frac{r^{(e)}(1, S^{(e)})^2 + (\sigma_1^{(e)})^2}{\pi(1 \mid S^{(e)})}.$$

It follows from the total variance formula that

$$\mathrm{Var}(\psi_1^{(e)}(O^{(e)})) = \mathbb{E}\left[\frac{r^{(e)}(1, S^{(e)})^2 + (\sigma_1^{(e)})^2}{\pi(1 \mid S^{(e)})}\right] + \mathrm{Var}(r^{(e)}(1, S^{(e)})) - \mathbb{E}[r^{(e)}(1, S^{(e)})^2]. \quad (30)$$

2. Variance of $w(S^{(e)})\psi_0^{(e)}(O^{(e)})$: Under Assumption 4, direct calculation yields

$$\mathbb{E}[w(S^{(e)})\psi_0^{(e)}(O) \mid S^{(e)}] = w(S^{(e)})r^{(e)}(0, S^{(e)}),$$

$$\mathbb{E}[w(S^{(e)})^2\psi_0^{(e)}(O)^2 \mid S^{(e)}] = \frac{w(S^{(e)})^2(r^{(e)}(0, S^{(e)})^2 + {\sigma_0^{(e)}}^2)}{\pi(0 \mid S^{(e)})}.$$

Here, we use $\sigma_1^{(e)}$ to denote the standard deviation of $\epsilon_1^{(e)}$, and $\sigma_0^{(e)}$ to denote the standard deviation of $\epsilon_0^{(e)}$.

We next apply the total variance formula, which leads to

$$\begin{aligned}\mathrm{Var}(w(S^{(e)})\psi_0^{(e)}(O^{(e)})) = \mathbb{E}&\left(w(S^{(e)})^2 \cdot \frac{r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2}{\pi(0 \mid S^{(e)})}\right) \\ &+ \mathrm{Var}(w(S^{(e)})r^{(e)}(0, S^{(e)})) - \mathbb{E}\left(w(S^{(e)})^2 r^{(e)}(0, S^{(e)})^2\right)\end{aligned} \quad (31)$$

3. Covariance term: Since $\psi_1^{(e)}(O^{(e)})$ is nonzero only when $A = 1$ and $\psi_0^{(e)}(O^{(e)})$ only when $A = 0$, their supports are disjoint; hence $\mathbb{E}[\psi_1^{(e)}(O^{(e)})\,\psi_0^{(e)}(O^{(e)})] = 0$. Therefore, their covariance simplifies to

$$\begin{aligned}\mathrm{Cov}(\psi_1^{(e)}(O^{(e)}), w(S^{(e)})\psi_0^{(e)}(O^{(e)})) &= -\mathbb{E}[\psi_1^{(e)}(O^{(e)})] \cdot \mathbb{E}[w(S^{(e)})\psi_0^{(e)}(O^{(e)})] \\ &= -\mathbb{E}[r^{(e)}(1, S^{(e)})] \cdot \mathbb{E}[w(S^{(e)})r^{(e)}(0, S^{(e)})].\end{aligned} \quad (32)$$

Combining (30)–(32) completes the proof of the lemma.

□

Lemma 3 (Variance from Historical Data)**.** *For a given weight function w, the variance of $Z^{(h)}(w)$ defined in* (9) *is given by*

$$\mathrm{Var}\left(Z^{(h)}(w)\right) = \mathrm{Var}\left((1 - w(S^{(h)}))\mu(S^{(h)})r^{(h)}(0, S^{(h)})\right) + \mathbb{E}\left[(1 - w(S^{(h)}))^2\mu(S^{(h)})^2\right](\sigma^{(h)})^2.$$

Proof. By definition,

$$(1 - w(S^{(h)}))\psi^{(h)}(S^{(h)}) = (1 - w(S^{(h)}))\mu(S^{(h)})\left(r^{(h)}(0, S^{(h)}) + \epsilon^{(h)}\right).$$

Under Assumption 4, the conditional mean of $\epsilon^{(h)}$ given $S^{(h)}$ is zero. We obtain that

$$\begin{aligned}\mathrm{Var}\left((1-w(S^{(h)}))\cdot\psi^{(h)}(S^{(h)})\right) &= \mathrm{Var}\left((1-w(S^{(h)})\mu(S^{(h)})r^{(h)}(0,S^{(h)})\right)\\ &+ \mathrm{Var}\left((1-w(S^{(h)})\mu(S^{(h)})\epsilon^{(h)}\right).\end{aligned}$$

As for the second term, by the variance formula we have

$$\begin{aligned}&\mathbb{E}\left[\left((1-w(S^{(h)}))\mu(S^{(h)})\epsilon^{(h)}\right)^2\right]-\left(\mathbb{E}\left[(1-w(S^{(h)}))\mu(S^{(h)})\epsilon^{(h)}\right]\right)^2\\ &=\mathbb{E}\left[\left((1-w(S^{(h)}))\mu(S^{(h)})\epsilon^{(h)}\right)^2\right] \quad (\text{since } \mathbb{E}[\epsilon^{(h)}]=0 \text{ and } \epsilon^{(h)}\text{is dependent of } S^{(h)})\\ &=\mathbb{E}\left[(1-w(S^{(h)}))^2\mu(S^{(h)})^2\right]\cdot\mathbb{E}\left[(\epsilon^{(h)})^2\right]=\mathbb{E}\left[(1-w(S^{(h)}))^2\mu(S^{(h)})^2\right](\sigma^{(h)})^2.\end{aligned}$$

□

C.3 Preliminaries for the Proofs of Corollaries

We first derive the closed-form expression of the MSEs of EDO and HDB.

The MSE of EDO estimator: When $w=1$, the weighted ATE estimator becomes EDO, and its MSE simplifies to,

$$\begin{aligned}\mathrm{MSE(EDO)}=\frac{1}{|\mathcal{D}^{(e)}|}\Big(&\mathbb{E}\left[\frac{r^{(e)}(1,S^{(e)})^2+(\sigma_1^{(e)})^2}{\pi(1\mid S^{(e)})}\right]+\mathbb{E}\left[\frac{r^{(e)}(0,S^{(e)})^2+(\sigma_0^{(e)})^2}{\pi(0\mid S^{(e)})}\right]\\ &-\left(\mathbb{E}[r^{(e)}(1,S^{(e)})]-\mathbb{E}[r^{(e)}(0,S^{(e)})]\right)^2\Big).\end{aligned}\tag{33}$$

The MSE of HDB estimator: When $w=0$, the weighted estimator becomes HDB, and its MSE can be expressed as,

$$\begin{aligned}\mathrm{MSE(HDB)}=&\frac{1}{|\mathcal{D}^{(e)}|}\left(\mathbb{E}\left[\frac{r^{(e)}(1,S^{(e)})^2+(\sigma_1^{(e)})^2}{\pi(1\mid S^{(e)})}\right]-\left(\mathbb{E}[r^{(e)}(1,S^{(e)})]\right)^2\right)\\ &+\frac{1}{|\mathcal{D}^{(h)}|}\left(\mathrm{Var}(\mu(S^{(h)})r^{(h)}(0,S^{(h)}))+\mathbb{E}[\mu(S^{(h)})^2]\cdot(\sigma^{(h)})^2\right)+\left(\mathbb{E}[b(S^{(e)})]\right)^2.\end{aligned}\tag{34}$$

Next, we notice that although the MSE of the weighted estimator varies with the weight function w, certain components of the MSE remain constant with respect to w. To focus on the terms that vary with w, we define a new loss function $\mathcal{L}(w)$ as follows:

$$\begin{aligned}\mathcal{L}(w)=&\frac{1}{|\mathcal{D}^{(e)}|}\left(\mathbb{E}\left[\frac{w(S^{(e)})^2\cdot(r^{(e)}(0,S^{(e)})^2+(\sigma_0^{(e)})^2)}{\pi(0\mid S^{(e)})}\right]\right.\\ &\left.+2\mathbb{E}[r^{(e)}(1,S^{(e)})]\cdot\mathbb{E}[w(S^{(e)})r^{(e)}(0,S^{(e)})]-\left(\mathbb{E}[w(S^{(e)})r^{(e)}(0,S^{(e)})]\right)^2\right)\\ &+\frac{1}{|\mathcal{D}^{(h)}|}\mathrm{Var}((1-w(S^{(h)}))\mu(S^{(h)})r^{(h)}(0,S^{(h)}))\\ &+\frac{1}{|\mathcal{D}^{(h)}|}\mathbb{E}[(1-w(S^{(h)}))^2\mu(S^{(h)})^2]\cdot(\sigma^{(h)})^2+\left(\mathbb{E}\left[(1-w(S^{(e)}))b(S^{(e)})\right]\right)^2,\\ &=\frac{1}{|\mathcal{D}^{(e)}|}(\mathcal{L}_1+\mathcal{L}_2-\mathcal{L}_3)+\frac{1}{|\mathcal{D}^{(h)}|}(\mathcal{L}_4+\mathcal{L}_5)+\mathcal{L}_6,\end{aligned}\tag{35}$$

where $\mathcal{L}_1,\ldots,\mathcal{L}_6$ denote the above six terms, respectively.

Remark. To simplify the proof of the corollaries, we consider the case where the contextual variables $S^{(e)}$,$S^{(h)}$ are discrete. Our results can be extended to settings with continuous contextual variables.

C.4 Proof of Corollary 1

Optimality of EDO. We prove the optimality by contradiction. Suppose the minimal MSE is achieved by a weight function w that does not converge to one. Then, there must exist a non-empty set $\mathcal{S}_1 := \{S : 1 - w(S) \geq \Delta, \Delta > 0\}$. Under this assumption, we derive a lower bound of $\mathcal{L}_5$ in (35):

$$\begin{aligned}\mathcal{L}_5 &= \sum_{S^{(h)} \in \mathcal{S}} (1 - w(S^{(h)}))^2 \mu(S^{(h)})^2 p_h(S^{(h)}) \cdot (\sigma^{(h)})^2 \\ &= \sum_{S^{(h)} \in \mathcal{S}_1} (1 - w(S^{(h)}))^2 \cdot \frac{(p_e(S^{(e)}))^2}{p_h(S^{(h)})} \cdot (\sigma^{(h)})^2,\end{aligned} \tag{36}$$

where the third equality follows from the definition of $\mu(S^{(h)} = \frac{p_e(S^{(e)})}{p_h(S^{(h)})}$.

As for other terms in the objective function $\mathcal{L}(w)$ in (35), under Assumption 5, it is easy to show that

$$\begin{aligned}|\mathcal{L}_2| &\leq 2|\mathbb{E}[r^{(e)}(1, S^{(e)})] \cdot \mathbb{E}[w(S^{(e)}) r^{(e)}(0, S^{(e)})]| \leq 2r_{max}^2, \\ \mathcal{L}_3 &= \left(\mathbb{E}[w(S^{(e)}) r^{(e)}(0, S^{(e)})]\right)^2 \leq r_{max}^2,\end{aligned}$$

leading to $\mathcal{L}_2 - \mathcal{L}_3 \geq -3r_{max}^2$. Since $\mathcal{L}_4$ and $\mathcal{L}_6$ are always non-negative, we have

$$\begin{aligned}\mathcal{L}(w) &\geq \frac{1}{|\mathcal{D}^{(e)}|}(\mathcal{L}_1 + \mathcal{L}_2 - \mathcal{L}_3) + \frac{1}{|\mathcal{D}^{(h)}|}\mathcal{L}_5 \geq \frac{1}{|\mathcal{D}^{(e)}|}(\mathcal{L}_1 - 3r_{max}^2) + \frac{1}{|\mathcal{D}^{(h)}|}\mathcal{L}_5 \\ &= \frac{1}{|\mathcal{D}^{(e)}|}\mathbb{E}\left[\frac{w(S^{(e)})^2 \cdot (r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2)}{\pi(0 \mid S^{(e)})}\right] - \frac{O(r_{\max}^2)}{|\mathcal{D}^{(e)}|} + \frac{1}{|\mathcal{D}^{(h)}|}\mathcal{L}_5\end{aligned}$$

Consider the EDO estimator. By setting $w = 1$, its objective function $\mathcal{L}(1)$ becomes

$$\frac{1}{|\mathcal{D}^{(e)}|}\left(\mathbb{E}\left[\frac{r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2}{\pi(0 \mid S^{(e)})}\right] + 2\mathbb{E}[r^{(e)}(1, S^{(e)})]\mathbb{E}[r^{(e)}(0, S^{(e)})] - \left(\mathbb{E}[r^{(e)}(0, S^{(e)})]\right)^2\right)$$

and it is smaller than

$$\frac{1}{|\mathcal{D}^{(e)}|}\left(\mathbb{E}\left[\frac{r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2}{\pi(0 \mid S^{(e)})}\right] + 2\left|\mathbb{E}[r^{(e)}(1, S^{(e)})]\mathbb{E}[r^{(e)}(0, S^{(e)})]\right| + \left(\mathbb{E}[r^{(e)}(0, S^{(e)})]\right)^2\right)$$

Futhermore,it can be bounded by

$$\frac{1}{|\mathcal{D}^{(e)}|}\left(\mathbb{E}\left[\frac{r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2}{\pi(0 \mid S^{(e)})}\right] + 3r_{\max}^2\right).$$

where the first inequality follows from the triangle inequality, and the second holds by the condition $|r^{(e)}(\cdot, \cdot)| < r_{\max}$ in Assumption 5. Thus,

$$\mathcal{L}(w) - \mathcal{L}(1) \geq G_1 + \frac{O(r_{\max}^2)}{|\mathcal{D}^{(e)}|}, \tag{37}$$

where

$$\begin{aligned}G_1 &= \frac{1}{|\mathcal{D}^{(h)}|}\sum_{S^{(h)} \in \mathcal{S}_1} (1 - w(S^{(h)}))^2 \cdot \frac{(p_e(S^{(e)}))^2}{p_h(S^{(h)})} \cdot (\sigma^{(h)})^2 \\ &- \frac{1}{|\mathcal{D}^{(e)}|}\sum_{S^{(e)} \in \mathcal{S}_1} (1 - w(S^{(e)})^2) \cdot p_e(S^{(e)}) \frac{r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2}{\pi(0 \mid S^{(e)})} \\ &= \sum_{S^{(e)} \in \mathcal{S}_1} \frac{(1 - w(S^{(e)}))^2 p_e^2(S^{(e)})}{|\mathcal{D}^{(h)}| p_h(S^{(h)})}\left((\sigma^{(h)})^2 - \delta \frac{r^{(e)}(0, S^{(e)})^2 + (\sigma_0^{(e)})^2}{\pi(0 \mid S^{(e)})} \frac{p_h(S^{(h)})}{p_e(S^{(e)})}\right)\end{aligned}$$

where the first equality follows using similar arguments to (36), and the second equality follows from the definition of $\delta = |\mathcal{D}^{(h)}|/|\mathcal{D}^{(e)}|$. Under Assumptions 3 and 5, it is easy to show that the condition $\sigma^{(h)} \gg \epsilon^{-1}\sqrt{\delta}(r_{\max} + \sigma_0^{(e)})$ implies $G_1 > 0$, which in turn ensures that $\mathcal{L}(w) - \mathcal{L}(1) > 0$ for any $w \neq 1$.

Since for any weight function w, the objective $\mathcal{L}(w)$ differs from $\mathrm{MSE}(w)$ only by a constant independent of w.This in turn ,

$$\mathrm{MSE}(w) > \mathrm{MSE}(1).$$

which leads to a contraction with the property of the optimal weight $w^* = \arg\min_w \mathrm{MSE}(w)$. Therefore,

$$w^* \to 1 \quad \text{for all } S.$$

□

Oracle property. We next show that the difference in MSE between our proposed estimator and MSE(EDO) is smaller than MSE(EDO) itself.

Define the ℓ_p-norm of $\epsilon_a^{(e)}$ (for $a \in \{0,1\}$)and $\epsilon^{(h)}$ as

$$\|\epsilon_a^{(e)}\|_p := \left(\mathbb{E}\left[|\epsilon_a^{(e)}|^p\right]\right)^{1/p}, \quad \text{for } p = 2, 4, 8.$$

$$\|\epsilon^{(h)}\|_p := \left(\mathbb{E}\left[|\epsilon^{(h)}|^p\right]\right)^{1/p}, \quad \text{for } p = 2, 4, 8.$$

Under Assumption 7, we show the proposed estimator achieves the oracle property. By (33), assumptions 3 and 5, $\mathrm{MSE(EDO)}$ has such a lower bound.

$$\mathrm{MSE(EDO)} = \mathrm{MSE}(1) = \frac{\mathrm{Var}(Z^{(e)}(1))}{|\mathcal{D}^{(e)}|} = \Omega\left(\frac{(\sigma_1^{(e)} + \sigma_0^{(e)})^2}{|\mathcal{D}^{(e)}|}\right). \tag{38}$$

According to Theorem 1, we can deduce that

$$\begin{aligned} \mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(1) &\leq \mathbb{E}[\widehat{\mathrm{bias}}_U^2(1) - \mathrm{bias}^2(1)] + \mathbb{E}[\widehat{\mathrm{Var}}_U(1) - \mathrm{Var}(1)] + O(\alpha B^2) \\ &= \mathbb{E}[\widehat{\mathrm{Var}}_U(1) - \mathrm{Var}(1)] + O(\alpha B^2). \end{aligned} \tag{39}$$

We next focus on orders of the above two terms, respectively.

By definition,we have

$$\mathbb{E}\left[\widehat{\mathrm{Var}}_U(1) - \mathrm{Var}(1)\right] = \frac{1}{|\mathcal{D}^{(e)}|} \cdot \mathbb{E}\left[\widehat{\mathrm{Var}}_U\left(Z^{(e)}(1)\right) - \mathrm{Var}\left(Z^{(e)}(1)\right)\right].$$

For analytical convenience, we express the empirical variance estimator of $Z^{(e)}(1)$ as

$$\widehat{\mathrm{Var}}\left(Z^{(e)}(1)\right) = \frac{1}{|\mathcal{D}^{(e)}|} \sum_{i \in \mathcal{D}^{(e)}} \left(Z_i^{(e)}(1) - \mu\right)^2, \quad \text{where } \mu = \mathbb{E}\left[Z^{(e)}(1)\right].$$

Then $\mathbb{E}[\widehat{\mathrm{Var}}\left(Z^{(e)}(1)\right)] = \mathrm{Var}\left(Z^{(e)}(1)\right)$.

According to (18),we need to calculate $\mathbb{E}\left[\widehat{\mathrm{Var}}\left(Z^{(e)}(1)\right) - \mathrm{Var}\left(Z^{(e)}(1)\right)\right]^4$. Through a standard moment expansion, we obtain:

$$\begin{aligned} \mathbb{E}\left[\left(\widehat{\mathrm{Var}}\left(Z^{(e)}(1)\right) - \mathrm{Var}(Z^{(e)}(1))\right)^4\right] &= O\left(\frac{1}{|\mathcal{D}^{(e)}|^3} \cdot \mathbb{E}\left[\left((Z_i^{(e)}(1) - \mu)^2 - \mathrm{Var}(Z^{(e)}(1))\right)^4\right]\right) \\ &\quad + O\left(\frac{1}{|\mathcal{D}^{(e)}|^2} \cdot \left(\mathrm{Var}\left((Z_i^{(e)}(1) - \mu)^2\right)\right)^2\right). \end{aligned} \tag{40}$$

$\mathbb{E}\left[\left((Z_i^{(e)}(1) - \mu)^2 - \mathrm{Var}(Z^{(e)}(1))\right)^4\right]$ and $\left(\mathrm{Var}\left((Z_i^{(e)}(1) - \mu)^2\right)\right)^2$ are of the same order, as their leading terms involve the eighth moments of $\epsilon_0^{(e)}$ and $\epsilon_1^{(e)}$. For the remaining terms, under

Assumptions 3 and 5, they can be uniformly bounded the same order of $r_{\max}$ and ϵ. Therefore, both expressions share the same asymptotic order. So the second term is the dominant one, and the overall expression is of order $O\left(\frac{\mathrm{Var}\left((Z^{(e)}(1)-\mu)^2\right)^2}{|\mathcal{D}^{(e)}|^2}\right)$. If we choose U as

$$U = O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}| \cdot \mathrm{Var}^{1/2}\left((Z^{(e)}(1)-\mu)^2\right)}{|\mathcal{D}^{(e)}|^{1/4}}\right), \quad \alpha = O\left(|\mathcal{D}^{(e)}|^{-1}\log^{-1}|\mathcal{D}^{(e)}|\right),$$

we can get:

$$\left|\widehat{\mathrm{Var}}\left(Z^{(e)}(1)\right) - \mathrm{Var}\left(Z^{(e)}(1)\right)\right| = O(U) = O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}|\mathrm{Var}^{1/2}\left((Z_i^{(e)}(1)-\mu)^2\right)}{|\mathcal{D}^{(e)}|^{1/4}}\right).$$

Therefore, we obtain the following bound on the expected deviation between the upper-bound estimator and the true variance of the EDO estimator:

$$\mathbb{E}\left[\widehat{\mathrm{Var}}_U(1) - \mathrm{Var}(1)\right] = O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{5/4}} \cdot \mathrm{Var}^{1/2}\left((Z^{(e)}(1)-\mu)^2\right)\right). \tag{41}$$

We use the fact that lower-order moments can be controlled by higher-order moments. Specifically, we have

$$\begin{aligned}\mathrm{Var}^{1/2}\left(\left(Z^{(e)}(1)-\mu\right)^2\right) &\leq O\left(\mathbb{E}\left[Z^{(e)}(1)^4\right]^{1/2}\right) = O\left(\|Z^{(e)}(1)\|_4^2\right) \\ &\leq O\left(\|\psi_1^{(e)}(O^{(e)})\|_4^2\right) + O\left(\|\psi_0^{(e)}(O^{(e)})\|_4^2\right).\end{aligned} \tag{42}$$

The first equality follows directly from the definition of the ℓ_4-norm, while the final inequality is a consequence of the triangle inequality for norms and the definition in (9). According to Assumptions 7, 2, and 3, we obtain the following bound:

$$\begin{aligned}\mathrm{Var}^{1/2}\left(\left(Z^{(e)}(1)-\mu\right)^2\right) &\leq O\left(\|\psi_1^{(e)}(O^{(e)})\|_4^2 + \|\psi_0^{(e)}(O^{(e)})\|_4^2\right) \\ &\leq O\left(\frac{(r_{\max}+\sigma_1^{(e)}+\sigma_0^{(e)})^2}{\epsilon}\right).\end{aligned} \tag{43}$$

Furthermore, we have:

$$O(\alpha B^2) = O\left(|\mathcal{D}^{(e)}|^{-1}\log|\mathcal{D}^{(e)}|^{-1}B^2\right). \tag{44}$$

We treat $r_{\max}$, $\sigma_0^{(e)}$, $\sigma_1^{(e)}$, and ϵ as constants, and consider the asymptotic regime where $|\mathcal{D}^{(e)}| \to \infty$. Plugging (41), (43), (42), and (44) into (39), we have

$$\mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(1) \leq O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{5/4}} \cdot \frac{(r_{\max}+\sigma_1^{(e)}+\sigma_0^{(e)})^2}{\epsilon} + |\mathcal{D}^{(e)}|^{-1}\log|\mathcal{D}^{(e)}|^{-1}B^2\right).$$

Comparing it with the MSE of the EDO estimator given in (38), one can deduce that

$$\frac{\mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(1)}{\mathrm{MSE}(1)} \to 0$$

as $|\mathcal{D}^{(e)}| \to \infty$. We conclude that the gap between our proposed method and the EDO baseline vanishes at a faster rate than the MSE of the EDO estimator itself. □

C.5 Proof of Corollary 2

Oracle property. We prove the corollary by contradiction. Suppose that optimal $w(S)$ does not converge uniformly to zero. Then, there must exist a non-empty set $\mathcal{S}_2 := \{S : w(S) \geq \Delta, \Delta > 0\}$.

Under this assumption, we can derive a lower bound for the objective $\mathcal{L}(w^*)$ (35) as follows:

$$\begin{aligned}
\mathcal{L}(w) &\geq \frac{1}{|\mathcal{D}^{(e)}|}\left(\sum_{S^{(e)}\in\mathcal{S}}\frac{w(S^{(e)})^2(\sigma_0^{(e)})^2}{\pi(0\mid S^{(e)})}p_e(S^{(e)}) - O(r_{\max}^2)\right) \\
&\quad + \frac{1}{|\mathcal{D}^{(h)}|}\left(\mathbb{E}[(1-w(S^{(h)}))^2\mu(S^{(h)})^2]\cdot(\sigma^{(h)})^2\right) \\
&= \frac{1}{|\mathcal{D}^{(e)}|}\left(\sum_{S^{(e)}\in\mathcal{S}_2}\frac{w(S^{(e)})^2(\sigma_0^{(e)})^2}{\pi(0\mid S^{(e)})}p_e(S^{(e)}) - O(r_{\max}^2)\right) \\
&\quad + \frac{1}{|\mathcal{D}^{(h)}|}\left(\mathbb{E}[(1-w(S^{(h)}))^2\mu(S^{(h)})^2]\cdot(\sigma^{(h)})^2\right)
\end{aligned}$$

Here, the first inequality holds since $\mathcal{L}_4$ and $\mathcal{L}_6$ in (35) are always non-negative, and the remaining terms can be upper bounded by $O(r_{\max}^2)$ according to Assumption 5. On the other hand, we can derive an upper bound for $\mathcal{L}(0)$, which corresponds to assigning zero weights to all historical data. In this case, the objective reduces to the variance from historical data and the squared bias(35):

$$\begin{aligned}
\mathcal{L}(0) &= \frac{1}{|\mathcal{D}^{(h)}|}\left(\mathrm{Var}(\mu(S^{(h)})r^{(h)}(0,S^{(h)})) + \mathbb{E}[\mu(S^{(h)})^2]\cdot(\sigma^{(h)})^2\right) + \left(\mathbb{E}[b(S^{(e)})]\right)^2 \\
&= \frac{1}{|\mathcal{D}^{(h)}|}\left(\mathbb{E}[\mu(S^{(h)})r^{(h)}(0,S^{(h)})]^2 - (\mathbb{E}[\mu(S^{(h)})r^{(h)}(0,S^{(h)})])^2\right) \\
&+ \frac{1}{|\mathcal{D}^{(h)}|}\mathbb{E}[\mu(S^{(h)})^2](\sigma^{(h)})^2 + \left(\mathbb{E}[b(S^{(e)})]\right)^2. \\
&\leq \frac{1}{|\mathcal{D}^{(h)}|}\left(\mathbb{E}[\mu(S^{(h)})r^{(h)}(0,S^{(h)})]^2 + \mathbb{E}[\mu(S^{(h)})^2]\cdot(\sigma^{(h)})^2\right) + O(r_{max}^2) \\
&\leq \frac{1}{|\mathcal{D}^{(h)}|}\left(O\frac{(r_{max}^2)}{\epsilon} + \mathbb{E}[\mu(S^{(h)})^2]\cdot(\sigma^{(h)})^2\right) + O(r_{max}^2)
\end{aligned}$$

The reasoning behind this conclusion is as follows. The squared bias term $(\mathbb{E}[b(S^{(e)})])^2$ can be bounded by $O(r_{\max}^2)$. Similarly, the expectation term $\mathbb{E}[\mu(S^{(h)})r^{(h)}(S^{(h)})]$ is bounded by $O(r_{\max})$, which implies that its square is of order $O(r_{\max}^2)$ as well.

Direct calculations lead to

$$\begin{aligned}
\mathcal{L}(w^*) - \mathcal{L}(0) \quad &\geq \quad \frac{1}{|\mathcal{D}^{(e)}|}\sum_{S^{(e)}\in\mathcal{S}_2}\frac{w(S^{(e)})^2(\sigma_0^{(e)})^2}{\pi(0\mid S^{(e)})}p_e(S^{(e)}) - O(r_{\max}^2) \\
&- \quad \frac{1}{|\mathcal{D}^{(h)}|}\mathbb{E}[(1-(1-w(S^{(h)}))^2)\mu(S^{(h)})^2](\sigma^{(h)})^2 \\
&= \quad \frac{1}{|\mathcal{D}^{(e)}|}\sum_{S^{(e)}\in\mathcal{S}_2}\frac{w(S^{(e)})^2(\sigma_0^{(e)})^2}{\pi(0\mid S^{(e)})}p_e(S) - O(r_{\max}^2) \\
&- \quad \frac{1}{|\mathcal{D}^{(h)}|}\sum_{S^{(h)}\in\mathcal{S}_2}\left[1-(1-w(S^{(h)}))^2\right]\frac{(p_e(S^{(e)}))^2}{p_h(S^{(h)})}\cdot(\sigma^{(h)})^2 \\
&= \quad \sum_{S^{(e)}\in\mathcal{S}_2}\frac{w(S)^2p_e(S^{(e)})}{|\mathcal{D}^{(e)}|\pi(0\mid S^{(e)})}\cdot \\
&\qquad \left((\sigma_0^{(e)})^2 - \frac{(1-(1-w(S^{(e)}))^2)p_e(S^{(e)})}{w(S^{(h)})^2p_h(S^{(h)})\delta}\pi(0\mid S^{(e)})(\sigma^{(h)})^2\right) - O(r_{\max}^2),
\end{aligned}$$

where the first equality directly follows from the calculation of expectation, and the second equality uses $\mu(S) = \frac{p_e(S^{(e)})}{p_h(S^{(h)})}$ and $\delta = |\mathcal{D}^{(h)}|/|\mathcal{D}^{(e)}|$.

According to the assumption

$$\sigma^{(e)} \gg \epsilon^{-1/2}\left(\frac{\sigma^{(h)}}{\sqrt{\delta}} + \sqrt{|\mathcal{D}^{(e)}|}\cdot r_{\max}\right),$$

the order of the first term in $\mathcal{L}(w) - \mathcal{L}(0)$ dominates the term of order $\mathrm{O}(r_{\max}^2)$, leading to

$$\mathcal{L}(w) - \mathcal{L}(0) > 0$$

for any w. Hence, the lower bound of $\mathcal{L}(w)$ strictly exceeds the upper bound of $\mathcal{L}(0)$. This contradicts the optimality of w^*, since $w^* = \arg\min_w \mathcal{L}(w)$. Therefore, the assumption must be false, and it follows that $w^* \to 0$ for all S. In the following, we discuss additional conditions under which the gap between the MSE of our proposed method and that of the HDB estimator becomes negligible relative to the MSE of the HDB estimator itself.

□

Optimality of HDB. At first,We give the additional assumption:

Assumption 8. *The variance $\sigma_1^{(e)}$ satisfies:*

$$\sigma_1^{(e)} \gg \max\left\{\log^{1/4}|\mathcal{D}^{(e)}||\mathcal{D}^{(e)}|^{1/4}(\sigma_0^{(e)} + r_{\max}),\ \log^{1/4}|\mathcal{D}^{(h)}||\mathcal{D}^{(h)}|^{1/4}\sqrt{\frac{|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(h)}|}}(\sigma^{(h)} + r_{\max})\right\}.$$

We begin by recalling the asymptotic order of the mean squared error (MSE) of the HDB estimator given in (34):

$$\begin{aligned}\mathrm{MSE}(\mathrm{HDB}) &= \mathrm{bias}^2(0) + \frac{\mathrm{Var}\left(\psi_1^{(e)}(O^{(e)})\right)}{|\mathcal{D}^{(e)}|} + \frac{\mathrm{Var}\left(\psi_0^{(h)}(O^{(h)})\right)}{|\mathcal{D}^{(h)}|} \\ &= \mathrm{bias}^2(0) + \Omega\left(\frac{(\sigma_1^{(e)})^2}{|\mathcal{D}^{(e)}|}\right) + \Omega\left(\frac{(\sigma^{(h)})^2}{|\mathcal{D}^{(h)}|}\right).\end{aligned} \tag{45}$$

The difference between our proposed method and the MSE of the HDB estimator can be decomposed according to Theorem 1,

$$\mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(\mathrm{HDB}) \le \mathbb{E}\left[\widehat{\mathrm{bias}}_U^2(0) - \mathrm{bias}^2(0)\right] + \mathbb{E}\left[\widehat{\mathrm{Var}}_U(\mathrm{HDB}) - \mathrm{Var}(\mathrm{HDB})\right] + O(\alpha B^2).$$

For the variance difference component, we have:

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathrm{Var}}_U(\mathrm{HDB}) - \mathrm{Var}(\mathrm{HDB})\right] = &\frac{1}{|\mathcal{D}^{(e)}|}\cdot\mathbb{E}\left[\widehat{\mathrm{Var}}\left(\psi_1^{(e)}(O^{(e)})\right) - \mathrm{Var}\left(\psi_1^{(e)}(O^{(e)})\right)\right] \\ &+ \frac{1}{|\mathcal{D}^{(h)}|}\cdot\mathbb{E}\left[\widehat{\mathrm{Var}}_U\left(\psi_0^{(h)}(O^{(h)})\right) - \mathrm{Var}\left(\psi_0^{(h)}(O^{(h)})\right)\right].\end{aligned}$$

Similar to the derivation of (41), according to the fourth-moment version of Markov's inequality(17) ,(18),and assumption8, we can take:

$$U_1 = O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}|\cdot\mathrm{Var}^{1/2}\left((\psi_1^{(e)}(O^{(e)}) - \mu_1)^2\right)}{|\mathcal{D}^{(e)}|^{1/4}}\right) \quad \alpha_1 = O\left(|\mathcal{D}^{(e)}|^{-1}\log^{-1}|\mathcal{D}^{(e)}|\right),$$

$$U_2 = O\left(\frac{\log^{1/4}|\mathcal{D}^{(h)}|\cdot\mathrm{Var}^{1/2}\left((\psi_0^{(h)}(O^{(h)}) - \mu_2)^2\right)}{|\mathcal{D}^{(h)}|^{1/4}}\right) \quad \alpha_2 = O\left(|\mathcal{D}^{(h)}|^{-1}\log^{-1}|\mathcal{D}^{(h)}|\right),$$

where $\mu_1 = \mathbb{E}\left[\psi_1^{(e)}(O^{(e)})\right]$ and $\mu_2 = \mathbb{E}\left[\psi_0^{(h)}(O^{(h)})\right]$ denote the corresponding population means.

Furthermore, the order of $\mathbb{E}\left[\widehat{\mathrm{Var}}_U(\mathrm{HDB}) - \mathrm{Var}(\mathrm{HDB})\right]$ can be derived as follows:

$$\begin{aligned}\mathbb{E}\left[\widehat{\mathrm{Var}}_U(\mathrm{HDB}) - \mathrm{Var}(\mathrm{HDB})\right] = &O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}|\cdot\mathrm{Var}^{1/2}\left((\psi_1^{(e)}(O^{(e)}) - \mu_1)^2\right)}{|\mathcal{D}^{(e)}|^{5/4}}\right) \\ &+O\left(\frac{\log^{1/4}|\mathcal{D}^{(h)}|\cdot\mathrm{Var}^{1/2}\left((\psi_0^{(h)}(O^{(h)}) - \mu_2)^2\right)}{|\mathcal{D}^{(h)}|^{5/4}}\right).\end{aligned} \tag{46}$$

For the bias difference component, we directly compute the decomposition as:

$$\mathbb{E}\left[\widehat{\text{bias}}_U^2(0) - \text{bias}^2(0)\right] = \mathbb{E}\left[\left(\widehat{\text{bias}}_U(0) - \text{bias}(0)\right)^2\right] + 2 \cdot \text{bias}(0) \cdot \mathbb{E}\left[\widehat{\text{bias}}_U(0) - \text{bias}(0)\right].$$

More precisely, the bias estimation error can be decomposed into two components,

$$\frac{1}{|\mathcal{D}^{(h)}|}\sum_{i=1}^{|\mathcal{D}^{(h)}|}\psi_{0,i}^{(h)}(O^{(h)}) - \mathbb{E}[\psi_0^{(h)}(O^{(h)})] \quad \frac{1}{|\mathcal{D}^{(e)}|}\sum_{j=1}^{|\mathcal{D}^{(e)}|}\psi_{0,j}^{(e)}(O^{(e)}) - \mathbb{E}[\psi_0^{(e)}(O^{(e)})],$$

each of which can be controlled via (40):

$$U_{b1} = O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}| \cdot \text{Var}^{1/2}\left(\psi_0^{(e)}(O^{(e)})\right)}{|\mathcal{D}^{(e)}|^{1/4}}\right) \quad \alpha_3 = O\left(|\mathcal{D}^{(e)}|^{-1}\log^{-1}|\mathcal{D}^{(e)}|\right)$$

$$U_{b2} = O\left(\frac{\log^{1/4}|\mathcal{D}^{(h)}| \cdot \text{Var}^{1/2}\left(\psi_0^{(h)}(O^{(h)})\right)}{|\mathcal{D}^{(h)}|^{1/4}}\right) \quad \alpha_4 = O\left(|\mathcal{D}^{(h)}|^{-1}\log^{-1}|\mathcal{D}^{(h)}|\right).$$

Denote $U_b = U_{b1} + U_{b2}$, it is easy to deduce that $\text{bias}(0) \ll U_b$. Therefore, we obtain:

$$U_b = O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}|\,\text{Var}^{1/2}\big(\psi_0^{(e)}(O^{(e)})\big)}{|\mathcal{D}^{(e)}|^{1/4}}\right) + O\left(\frac{\log^{1/4}|\mathcal{D}^{(h)}|\,\text{Var}^{1/2}\big(\psi_0^{(h)}(O^{(h)})\big)}{|\mathcal{D}^{(h)}|^{1/4}}\right).$$

Next,we can get:

$$\begin{aligned}\mathbb{E}\Big[\widehat{\text{bias}}_U^2(0) - \text{bias}^2(0)\Big] &= O(U_b^2) + O\big(\text{bias}(0)\cdot U_b\big) = O(U_b^2)\\ &= O\left(\frac{\log^{1/2}|\mathcal{D}^{(e)}|\,\text{Var}\big(\psi_0^{(e)}(O^{(e)})\big)}{|\mathcal{D}^{(e)}|^{1/2}}\right) + O\left(\frac{\log^{1/2}|\mathcal{D}^{(h)}|\,\text{Var}\big(\psi_0^{(h)}(O^{(h)})\big)}{|\mathcal{D}^{(h)}|^{1/2}}\right). \quad (47)\end{aligned}$$

By summing all α_i's, we obtain

$$\alpha = \sum_{i=1}^{4}\alpha_i = O\left(|\mathcal{D}^{(e)}|^{-1}\log^{-1}|\mathcal{D}^{(e)}| + |\mathcal{D}^{(h)}|^{-1}\log^{-1}|\mathcal{D}^{(h)}|\right). \quad (48)$$

If Assumption 8 holds, namely,

$$\begin{aligned}\sigma_1^{(e)} \gg \max\Bigg\{ &\log^{1/4}|\mathcal{D}^{(e)}| \cdot |\mathcal{D}^{(e)}|^{1/4}(\sigma_0^{(e)} + r_{\max})\epsilon^{-1/2},\\ &\log^{1/4}|\mathcal{D}^{(h)}| \cdot |\mathcal{D}^{(h)}|^{1/4}\sqrt{\frac{|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(h)}|}}(\sigma^{(h)} + r_{\max})\epsilon^{-1/2}\Bigg\}. \quad (49)\end{aligned}$$

This condition is equivalent to:

$$\frac{(\sigma_1^{(e)})^2}{|\mathcal{D}^{(e)}|\epsilon} \gg \max\left\{\frac{\log^{1/2}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{1/2}}(\sigma_0^{(e)} + r_{\max})^2\epsilon^{-1},\ \frac{\log^{1/2}|\mathcal{D}^{(h)}|}{|\mathcal{D}^{(h)}|^{1/2}}\cdot(\sigma_0^{(h)} + r_{\max})^2\epsilon^{-1}\right\}.$$

Furthermore, we know that:

$$\text{Var}\left(\psi_0^{(e)}(O^{(e)})\right) \le O(\frac{(\sigma_0^{(e)} + r_{\max})^2}{\epsilon}), \text{Var}\left(\psi_0^{(h)}(O^{(h)})\right) \le O(\frac{(\sigma^{(h)} + r_{\max})^2}{\epsilon}). \quad (50)$$

By combining the bounds in (45), (47), (48), and (50), under Assumption 8, and comparing it with the order in (45), we have:

$$\begin{aligned}\text{MSE}(\widehat{w}) - \text{MSE}(0) \quad \ll \quad O\Bigg(&\frac{\log^{1/4}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{5/4}}\cdot\frac{(r_{\max} + \sigma_1^{(e)})^2}{\epsilon} + \frac{\log^{1/4}|\mathcal{D}^{(h)}|}{|\mathcal{D}^{(h)}|^{5/4}}\cdot\frac{(r_{\max} + \sigma^{(h)})^2}{\epsilon}\\ &+ \left(\frac{B^2}{\log|\mathcal{D}^{(e)}||\mathcal{D}^{(e)}|} + \frac{B^2}{\log|\mathcal{D}^{(h)}||\mathcal{D}^{(h)}|}\right) + (47)\Bigg),\end{aligned}$$

which is much smaller than $\mathrm{MSE}(0)$ such that

$$\frac{\mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(0)}{\mathrm{MSE}(0)} \to 0$$

This completes the proof. □

C.6 Proof of Corollary 3

Oracle property. Recall that in (27),

$$\mathrm{MSE}(w) = \frac{\mathrm{Var}(Z^{(e)}(w))}{|\mathcal{D}^{(e)}|} + \frac{\mathrm{Var}(Z^{(h)}(w))}{|\mathcal{D}^{(h)}|} + \left(\mathbb{E}\left[(1 - w(S^{(e)})b(S^{(e)})\right]\right)^2.$$

Since Minimum Variance Estimator (MVE) is designed to minimize total variance, it suffices to show that the squared bias (the third term) is negligible compared to the variance terms (the first two). In this case, the bias contributes asymptotically little to the overall error.

Therefore, our analysis focuses on deriving a nontrivial lower bound for the variance component. For the variance contribution from the experimental data, we have the lower bound from Lemma 2, and Assumptions 3 and 5, one can derive that

$$\frac{\mathrm{Var}(Z^{(e)}(w))}{|\mathcal{D}^{(e)}|} \geq \frac{1}{|\mathcal{D}^{(e)}|} \cdot \mathbb{E}\left[\frac{w(S^{(e)})^2}{\pi(0 \mid S^{(e)})}\right] \cdot (\sigma_0^{(e)})^2.$$

This inequality holds by regrouping the terms involving $r^{(e)}(0, S^{(e)})$ and $r^{(e)}(1, S^{(e)})$, and applying a basic inequality that ensures the non-negativity of the remaining components. Similarly, for the variance contribution from the historical data, we obtain the following lower bound from Lemma 3 and Assumptions 3 and 5,

$$\frac{\mathrm{Var}(Z^{(h)}(w))}{|\mathcal{D}^{(h)}|} \geq \frac{1}{|\mathcal{D}^{(h)}|} \cdot \mathbb{E}\left[(1 - w(S^{(h)}))^2 \mu(S^{(h)})^2\right] \cdot (\sigma^{(h)})^2.$$

Note that both $\mathbb{E}\left[\frac{w(S^{(e)})^2}{\pi(0|S^{(e)})}\right]$ and $\mathbb{E}\left[(1 - w(S^{(h)}))^2 \mu(S^{(h)})^2\right]$ can not be simultaneously zero for any nontrivial weight function w. Therefore, the total variance is lower bounded by:

$$\frac{\mathrm{Var}(Z^{(e)}(w))}{|\mathcal{D}^{(e)}|} + \frac{\mathrm{Var}(Z^{(h)}(w))}{|\mathcal{D}^{(h)}|} \geq \min\left\{\frac{(\sigma^{(e)})^2}{|\mathcal{D}^{(e)}|}, \frac{(\sigma^{(h)})^2}{|\mathcal{D}^{(h)}|}\right\}, \quad \text{for any } w.$$

For the bias term, the condition

$$|b(S^{(e)})| \ll \min\left(\frac{\sigma^{(e)}}{\sqrt{|\mathcal{D}^{(e)}|}}, \frac{\sigma^{(h)}}{\sqrt{|\mathcal{D}^{(h)}|}}\right),$$

implies

$$\left(\mathbb{E}\left[(1 - w(S^{(e)}))b(S^{(e)})\right]\right)^2 \ll \min\left\{\frac{(\sigma^{(e)})^2}{|\mathcal{D}^{(e)}|}, \frac{(\sigma^{(h)})^2}{|\mathcal{D}^{(h)}|}\right\}, \quad \text{for any } w.$$

In this scenario, minimizing MSE is equivalent to minimizing the total variance. Therefore, the optimal weight w^* is given by:

$$w^* = \arg\min_w \left(\frac{\mathrm{Var}(Z^{(e)}(w)}{|\mathcal{D}^{(e)}|} + \frac{\mathrm{Var}(Z^{(h)}(w)}{|\mathcal{D}^{(h)}|}\right), \tag{51}$$

which corresponds exactly to the MVE formulation. This completes the proof of the corollary. □

Optimality of MVE. The conditions required here are identical to those stated in Assumption 8. In this setting, the order of the MSE of the MVE estimator is given by

$$\mathrm{MSE}(\mathrm{MVE}) = \mathrm{MSE}(\mathrm{w}^*) = \Omega\left(\frac{\sigma_0^{(e)} + \sigma_1^{(e)})^2}{|\mathcal{D}^{(e)}|}\right) + \Omega\left(\frac{(\sigma^{(h)})^2}{|\mathcal{D}^{(h)}|}\right). \tag{52}$$

We can derive the following MSE gap by plugging in w^* in (51) in Theorem 1,

$$\mathrm{MSE}(\widehat{\mathrm{w}}) - \mathrm{MSE}(\mathrm{w}^*) \leq \mathbb{E}\left[\widehat{\mathrm{bias}}_U^2(w^*) - \mathrm{bias}^2(w^*)\right] + \mathbb{E}\left[\widehat{\mathrm{Var}}_U(w^*) - \mathrm{Var}(w^*)\right] + O(\alpha B^2).$$

For the bias component, we further obtain:

$$\mathbb{E}\left[\widehat{\mathrm{bias}}_U^2(w^*) - \mathrm{bias}^2(w^*)\right] = O(U_b^2) + O\left(\mathrm{bias}(w^*) \cdot U_b\right)$$

by using the 4-th moment Markov equality similarly to derivations in Sections C.4-C.5, where

$$U_b = O\left(\frac{\log^{1/4}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{1/4}}\sqrt{\left(\frac{r_{\max}^2 + (\sigma_0^{(e)})^2}{\epsilon}\right)} + \frac{\log^{1/4}|\mathcal{D}^{(h)}|}{|\mathcal{D}^{(h)}|^{1/4}}\sqrt{\left(\frac{r_{\max}^2 + (\sigma^{(h)})^2}{\epsilon}\right)}\right), \tag{53}$$

and $\alpha_1 = O\left(|\mathcal{D}^{(e)}|^{-1}\log^{-1}|\mathcal{D}^{(e)}| + |\mathcal{D}^{(h)}|^{-1}\log^{-1}|\mathcal{D}^{(h)}|\right)$.
Under the small shift assumption, the bias term satisfies $|\mathrm{bias}(w^*)| \ll U_b$. Therefore, we obtain:

$$\begin{aligned}\mathbb{E}\Big[\widehat{\mathrm{bias}}_U^2(w^*) - \mathrm{bias}^2(w^*)\Big] = O\Big(&\frac{\log^{1/2}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{1/2}}\left(\frac{r_{\max}^2 + (\sigma_0^{(e)})^2}{\epsilon}\right) \\ &+ \frac{\log^{1/2}|\mathcal{D}^{(h)}|}{|\mathcal{D}^{(h)}|^{1/2}}\left(\frac{r_{\max}^2 + (\sigma^{(h)})^2}{\epsilon}\right)\Big)\end{aligned} \tag{54}$$

For the second term in MSE gap, similar to the derivation for (41), it is easy to deduce that

$$\begin{aligned}&\mathbb{E}\left[\widehat{\mathrm{Var}}_U(w^*) - \mathrm{Var}(w^*)\right] \\ =\ & O(U) = \left(\frac{(r_{\max} + \sigma_0^{(e)} + \sigma_1^{(e)})^2}{\epsilon \cdot |\mathcal{D}^{(e)}|^{5/4}} \cdot \log^{1/4}|\mathcal{D}^{(e)}| + \frac{(r_{\max} + \sigma^{(h)})^2}{\epsilon \cdot |\mathcal{D}^{(h)}|^{5/4}} \cdot \log^{1/4}|\mathcal{D}^{(h)}|\right),\end{aligned} \tag{55}$$

$\alpha_2 = O\left(|\mathcal{D}^{(e)}|^{-1}\log^{-1}|\mathcal{D}^{(e)}| + |\mathcal{D}^{(h)}|^{-1}\log^{-1}|\mathcal{D}^{(h)}|\right)$. Simple calculations show that the decay rate of $\mathbb{E}\left[\widehat{\mathrm{Var}}_U(w^*) - \mathrm{Var}(w^*)\right]$ is faster than that in $\mathrm{MSE}(w^*)$ in (52). The term

$$\alpha B^2 = (\alpha_1 + \alpha_2)B^2 = O\left(B^2|\mathcal{D}^{(e)}|^{-1}\log^{-1}|\mathcal{D}^{(e)}| + B^2|\mathcal{D}^{(h)}|^{-1}\log^{-1}|\mathcal{D}^{(h)}|\right) \tag{56}$$

which also vanishes faster than the leading MSE terms in (52).

Given the additional assumptions that

$$\frac{(\sigma_1^{(e)})^2}{|\mathcal{D}^{(e)}|\epsilon} \gg \max\left\{\frac{\log^{1/2}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{1/2}}(\sigma_0^{(e)} + r_{\max})^2\epsilon^{-1},\ \frac{\log^{1/2}|\mathcal{D}^{(h)}|}{|\mathcal{D}^{(h)}|} \cdot (\sigma^{(h)} + r_{\max})^2\epsilon^{-1}\right\},$$

which is asymptotically smaller than

Furthermore, by summing all three gap terms in (54)-(56),

$$\begin{aligned}&\mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(w^*) = O\Bigg(\frac{\log^{1/4}|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|^{5/4}}\frac{(r_{\max} + \sigma_0^{(e)} + \sigma_1^{(e)})^2}{\epsilon} + \\ &\frac{\log^{1/4}|\mathcal{D}^{(h)}|}{|\mathcal{D}^{(h)}|^{5/4}}\frac{(r_{\max} + \sigma^{(h)})^2}{\epsilon} + \left(\frac{B^2}{\log|\mathcal{D}^{(e)}||\mathcal{D}^{(e)}|} + \frac{B^2}{\log|\mathcal{D}^{(h)}||\mathcal{D}^{(h)}|}\right) + (54)\Bigg).\end{aligned}$$

It follows directly that

$$\frac{\mathrm{MSE}(\widehat{w}) - \mathrm{MSE}(\mathrm{MVE})}{\mathrm{MSE}(\mathrm{MVE})} \to 0$$

as the sample size goes to infinity. This completes the proof. □

C.7 Proof of Corollary 4

Oracle property. We prove the corollary by contradiction, considering both the **moderate shift** and **large shift** cases.

(1)**Moderate shift.** We define the set $\mathcal{S}_3 := \{S : 1 - w(S) \geq \frac{\Delta}{\epsilon}, \Delta > 0\}$. We consider the experiment-related component for the objective function in (35),

According to the condition $|r^{(e)}(0, S^{(e)})| \leq r_{\max}$, $w(S^{(e)}) \in [0, 1]$, $\pi(0 \mid S^{(e)}) \geq \epsilon$ for any $S^{(e)}$ and triangle inequality, we obtain the upper bound

$$|\mathcal{L}_{\text{exp}}(w)| \leq \frac{1}{|\mathcal{D}^{(e)}|} O\left(\frac{r_{\max}^2 + (\sigma_0^{(e)})^2}{\epsilon}\right).$$

Similarly, the historical-related component can be bounded by

$$\begin{aligned}\mathcal{L}_{\text{his}}(w) &= \frac{1}{|\mathcal{D}^{(h)}|}\left(\mathcal{L}_4 + \mathcal{L}_5\right) \\ &= \frac{1}{|\mathcal{D}^{(h)}|}\ \text{Var}\left((1 - w(S^{(h)}))\,\mu(S^{(h)})\,r^{(h)}(0, S^{(h)})\right) \\ &+ \frac{1}{|\mathcal{D}^{(h)}|}\,\mathbb{E}\big[(1 - w(S^{(h)}))^2\,\mu(S^{(h)})^2\big]\ (\sigma^{(h)})^2 \leq \frac{1}{|\mathcal{D}^{(h)}|}\,O\Big(\frac{r_{\max}^2 + (\sigma^{(h)})^2}{\epsilon}\Big).\end{aligned}$$

For the bias item:

$$\mathcal{L}_6 = \left|\mathbb{E}\left[(1 - w(S^{(e)}))b(S^{(e)})\right]\right|^2 \geq (\sum_{S^{(e)} \in \mathcal{S}_3} p_e(S^{(e)})\frac{\Delta}{\epsilon} \cdot |b(S^{(e)})|)^2 \geq (\sum_{S^{(e)} \in \mathcal{S}_3} \Delta \cdot |b(S^{(e)})|)^2,$$

where the inequality follows from the fact that $b(S^{(e)})$ is sign-consistent over $\mathcal{S}$, either non-negative or non-positive for all $S^{(e)} \in \mathcal{S}$, and the coefficient of $|b(S^{(e)}|)$ is strictly positive in $\mathcal{S}_3$.

As for $\mathcal{L}(1)$, it can be upper bounded as:

$$\mathcal{L}(1) \leq O\left(\frac{r_{\max}^2 + (\sigma_1^{(e)})^2}{\epsilon|\mathcal{D}^{(e)}|}\right).$$

In the moderate case, for any $S^{(e)}$,we have $|b(S^{(e)})| \gg \frac{1}{\sqrt{\epsilon}}\left(\frac{\sigma^{(e)} + r_{\max}}{\sqrt{|\mathcal{D}^{(e)}|}} + \frac{\sigma^{(h)} + r_{\max}}{\sqrt{|\mathcal{D}^{(h)}|}}\right)$, so

$$\mathcal{L}_6 \geq (\sum_{S^{(e)} \in \mathcal{S}_3} \Delta \cdot |b(S^{(e)})|)^2 \gg \frac{1}{|\mathcal{D}^{(e)}|} \cdot \left(\frac{r_{\max}^2 + (\sigma_0^{(e)})^2}{\epsilon}\right) + \frac{1}{|\mathcal{D}^{(h)}|} \cdot \left(\frac{r_{\max}^2 + (\sigma^{(h)})^2}{\epsilon}\right),$$

then the bias dominates the total MSE. Compare the order of $\mathcal{L}(w)$ with $\mathcal{L}(1)$, we have

$$\mathcal{L}(w) = \mathcal{L}_{\text{exp}}(w) + \mathcal{L}_{\text{his}}(w) + \mathcal{L}_6 \gg \mathcal{L}(1),$$

for any $w \neq 1$,which leads to a contradiction. Hence $\text{MSE}(w)$ in w, the optimal weight function $w^* = \arg\min_w \text{MSE}(w)$ satisfies

$$w \to 1 \quad \text{for all } S.$$

(2) **Large shift.** The proof follows the same reasoning as in the moderate shift case by using the condition

$$(\sum_{S^{(e)} \in \mathcal{S}_3} \Delta \cdot |b(S^{(e)})|)^2 \geq \frac{\log|\mathcal{D}^{(e)}|}{|\mathcal{D}^{(e)}|}\left(\frac{r_{\max}^2 + (\sigma_0^{(e)})^2}{\epsilon}\right) + \frac{|\log\mathcal{D}^{(h)}|}{|\mathcal{D}^{(e)}|}\left(\frac{r_{\max}^2 + (\sigma^{(h)})^2}{\epsilon}\right).$$

In this case, where the EDO method remains optimal, the gap between the MSE of our proposed estimator and that of EDO becomes asymptotically negligible. This completes the proof. □

Optimality of EDO. **Conditions for small MSE difference** Since the assumption and proof strategy are the same as those used in C.4 , we omit the detailed proof of this result for brevity. □

D Limitation

The current paper considers settings without carryover effects where each action affects only its immediate reward and does not influence future outcomes. However, in many real-world applications, treatments are sequentially assigned over time, and such carryover effects can arise [104, 131]. This represents a potential limitation of our work, as it does not account for such effects. In scenarios with carryover effects, the weight function for data integration may depend not only on contextual variables but also vary over time. Determining an optimal time-dependent weight function remains an open question, which we leave for future research.