Combining Experimental and Historical Data for Policy Evaluation

Ting Li [*1] **Chengchun Shi** [*2] **Qianglin Wen** [3] **Yang Sui** [1] **Yongli Qin** [4] **Chunbo Lai** [4] **Hongtu Zhu** [5]

Abstract

This paper studies policy evaluation with multiple data sources, especially in scenarios that involve one experimental dataset with two arms, complemented by a historical dataset generated under a single control arm. We propose novel data integration methods that linearly integrate base policy value estimators constructed based on the experimental and historical data, with weights optimized to minimize the mean square error (MSE) of the resulting combined estimator. We further apply the pessimistic principle to obtain more robust estimators, and extend these developments to sequential decision making. Theoretically, we establish non-asymptotic error bounds for the MSEs of our proposed estimators, and derive their oracle, efficiency and robustness properties across a broad spectrum of reward shift scenarios. Numerical experiments and real-data-based analyses from a ridesharing company demonstrate the superior performance of the proposed estimators.

1. Introduction

Motivation. This paper seeks to establish data-driven approaches for evaluating the effectiveness of a newly target policy against a conventional control. A basic approach relies solely on experimental data to formulate the treatment effect estimator, which we refer to as the experimental-data-only (EDO) estimator. However, the often limited sample size of experimental data prompts the need to incorporate auxiliary external datasets to enhance the precision of the treatment or policy effect estimator. We provide three illustrative examples to demonstrate this concept.

[*]Equal contribution [1]School of Statistics and Management, Shanghai University of Finance and Economics [2]Department of Statistics, London School of Economics and Political Science [3]Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University [4]Didi Chuxing [5]Department of Biostatistics, The University of North Carolina at Chapel Hill. Correspondence to: Hongtu Zhu <htzhu@email.unc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Example 1: A/B testing with historical data. A/B testing is frequently used in modern technology companies such as Amazon, eBay, Facebook, Google, LinkedIn, Microsoft, Netflix, Uber and Didi for comparing new products/strategies against existing ones (see Larsen et al., 2023, for a recent review). A common challenge in A/B testing is the limited experiment duration coupled with weak treatment effects. For example, in the ridesharing industry, experiments usually last no more than two weeks, and the effect sizes often range from 0.5% to 5% (Xu et al., 2018; Tang et al., 2019; Zhou et al., 2021; Qin et al., 2022). Yet, prior to these experiments, companies often have access to a substantial volume of historical data under the current policy. Leveraging such historical data can significantly improve the efficiency of A/B testing.

Example 2: Meta analysis. In medicine, data often spans across multiple healthcare institutions. A notable example is in the schizophrenia study that examined the efficacy of cognitive-behavioral therapy in early-stage schizophrenia patients (Tarrier et al., 2004). This study was a multicenter randomized controlled trial (RCT) executed in three treatment centers (Manchester, Liverpool, and North Nottinghamshire). Given that each center had a relatively small cohort, with fewer than 100 participants, pooling data from all sources becomes crucial to enhance causal learning.

Example 3: Combining observational data. RCTs are widely regarded as the benchmark for learning the causal impacts of interventions or treatments on specific outcomes (Imbens & Rubin, 2015). However, RCTs often face practical challenges such as high costs and time constraints, resulting in limited participant numbers. Conversely, observational data, derived from sources such as biobanks or electronic health records, boast larger sample sizes. Integrating both data presents a unique opportunity to improve the statistical learning efficiency (Colnet et al., 2020).

Challenge. The challenge of merging multiple data sources often lies in the distributional shifts that occur between them. In the domain of ridesharing and healthcare, datasets from different time frames frequently display temporal nonstationarity (Wan et al., 2021; Li et al., 2022; Wang et al., 2023b). In the schizophrenia study, variations across various treatment centers and ethnic groups introduce heterogeneity, leading to distributional shifts (Dunn & Bentall, 2007; Shi

Table 1. Summary of the properties of MSEs of our estimators.

Reward shift	Non-pessimistic estimator	Pessimistic estimator
Zero	Close to efficiency bound	Same order to oracle MSE
Small	Close to oracle MSE	Same order to oracle MSE
Moderate	May suffer a large MSE	Oracle property
Large	Oracle property	Oracle property

The **oracle** MSE denotes MSE of the oracle estimator that use the **best** weight to combine historical and experimental data whereas the **efficiency bound** is the smallest achievable MSE among a broad class of regular estimators (Tsiatis, 2006).

et al., 2018). Furthermore, the integration of RCTs with observational data introduces the potential for unmeasured confounding within the observational data (Pearl, 2009). Neglecting these distributional shifts would produce biased estimations of treatment effects.

Contributions. This paper focuses on the application of A/B testing with historical data. However, the methodologies and theoretical frameworks we develop are equally applicable to the other two examples discussed earlier. Our contributions are summarized as follows:

Methodologically, we propose several weighted estimators for data integration, including both pessimistic and non-pessimistic estimators, covering both non-dynamic settings (also referred to as contextual bandits in the OPE literature) and sequential decision making. We demonstrate the superior empirical performance of these estimators through simulations and real-data-based analyses [1].

Theoretically, we derive various statistical properties (e.g., efficiency, robustness and oracle property) of the proposed estimators across a wide range of scenarios, accommodating varying degrees of reward shift between the experimental data and the historical data in mean – from zero to small (the shift's order is much smaller than $n^{-1/2}$), moderate (the shift's order falls between $n^{-1/2}$ and $n^{-1/2}\sqrt{\log(n)}$), and large (the shift is substantially larger than $n^{-1/2}\sqrt{\log(n)}$), where n represents the effective sample size; see Table 1 for a summary. To the contrary, existing works impose more restrictive conditions. They either require the mean shift to be zero, or sufficiently large for clear detection (see e.g., Cheng & Cai, 2021; Han et al., 2021; Dahabreh et al., 2023; Li et al., 2023). In summary, our findings suggest that the non-pessimistic estimator tends to be effective in scenarios where the reward shift is minimal or substantial. In contrast, the pessimistic estimator demonstrates greater robustness, particularly in situations with moderately large reward shifts.

[1]R code implementing the proposed weighted estimators is available at https://github.com/tingstat/Data_Combination.

2. Related Work

Data integration in causal inference. There is a growing literature on combining randomized data with other sources of datasets; see Degtiar & Rose (2023) and Shi et al. (2023c) for reviews. These methods can be broadly classified into three categories, as outlined in the introduction:

1. The first category leverages historical datasets collected under the control (see e.g., Pocock, 1976; Cuffe, 2011; Viele et al., 2014; van Rosmalen et al., 2018; Schmidli et al., 2020; Cheng et al., 2023; Liu et al., 2023; Scott & Lewin, 2024). In particular, assuming no reward shift, Li et al. (2023) developed a semi-parametric efficient estimator whose MSE achieves the efficiency bound. In contrast, our methods are more flexible, allowing the reward shift to exist.
2. The second category is meta analysis where the external data is collected from different trials (Schmidli et al., 2014; DerSimonian & Laird, 2015; Hasegawa et al., 2017; Zhang et al., 2019; Steele et al., 2020; Lian et al., 2023; Rott et al., 2024). This category includes a notable subset of methods that apply ℓ_1-type penalty functions for selecting external data (Dahabreh et al., 2020; Han et al., 2021; 2023). However, their performance is sensitive to the choice of the tuning parameter, as shown in our numerical study (see Figure A2).
3. The last category incorporates observational data to enhance causal learning (see e.g., Hartman et al., 2015; Peysakhovich & Lada, 2016; Kallus et al., 2018; Athey et al., 2020; Gui, 2020; Yang et al., 2020a;b; Wu & Yang, 2022; Lee et al., 2023).

Remark 1. The first category of research is closely related to our work, while the focus of the last two categories differs from ours. Additionally, all aforementioned studies concentrate on the non-dynamic setting framework. Our research, however, broadens this perspective by investigating sequential decision making where treatments are assigned sequentially over time – a typical scenario studied in reinforcement learning (RL, Sutton & Barto, 2018).

Offline policy learning. The proposed pessimistic estimator is inspired by recent advancements in offline policy learning, which aims to learn an optimal policy from a pre-collected offline dataset without active exploration of the environment. Existing methods typically adopt the "pessimistic principle" to mitigate the discrepancy between the behavior policy that generates the offline data and the optimal policy. In contrast to the optimistic principle widely used in contextual bandits and online RL, the pessimistic principle favors actions whose values are less uncertain.

In non-dynamic settings, pessimistic algorithms can generally be categorized into value-based and policy-based methods. Value-based methods learn a conservative reward func-

tion to prevent overestimation and compute the greedy policy with respect to this estimated reward function (Buckman et al., 2020; Jin et al., 2021; Rashidinejad et al., 2021; Zhou et al., 2023). Conversely, policy-based methods directly search the optimal policy by either restricting the policy class to stay close to the behavior policy or optimizing the policy that maximizes an estimated lower bound of the reward function (Swaminathan & Joachims, 2015a;b; Wu & Wang, 2018; Kennedy, 2019; Aminian et al., 2022; Jin et al., 2022; Zhao et al., 2023).

Furthermore, the pessimistic principle has been extensively adopted in offline RL, accommodating more complex sequential settings (Kumar et al., 2019; 2020; Yu et al., 2020; Uehara & Sun, 2021; Xie et al., 2021; Bai et al., 2022; Rigter et al., 2022; Shi et al., 2022b; Yin et al., 2022; Zhou, 2023).

Off-policy evaluation (OPE). Finally, our work is closely related to OPE in contextual bandits and RL, which aims to estimate the mean outcome of a new target policy using data collected by a different policy (see Dudík et al., 2014; Uehara et al., 2022, for reviews). It has been recently employed to conduct A/B testing in sequential decision making (Bojinov & Shephard, 2019; Farias et al., 2022; Tang et al., 2022; Shi et al., 2023a;b; Li et al., 2024; Wen et al., 2024). Existing approaches in this field can generally be classified into three main groups:

1. **Direct methods**: these methods learn a reward or value function from offline data to estimate the policy value (Bradtke & Barto, 1996; Le et al., 2019; Feng et al., 2020; Luckett et al., 2020; Hao et al., 2021; Liao et al., 2021; Chen & Qi, 2022; Shi et al., 2022a; Bian et al., 2023; Uehara et al., 2024).

2. **Importance sampling (IS) methods**: this group employs the IS ratio to adjust the observed rewards, accounting for the discrepancy between the target policy and the behavior policy (Heckman et al., 1998; Hirano et al., 2003; Thomas et al., 2015; Liu et al., 2018; Dai et al., 2020; Wang et al., 2023a; Hu & Wager, 2023).

3. **Doubly robust (DR) methods**: these strategies integrate the principles of direct methods and IS (Tan, 2010; Dudík et al., 2011; van der Laan et al., 2011; Zhang et al., 2012; Jiang & Li, 2016; Thomas & Brunskill, 2016; Chernozhukov et al., 2018; Farajtabar et al., 2018; Oprescu et al., 2019; Shi et al., 2020; Uehara et al., 2020; Kallus & Uehara, 2022; Liao et al., 2022). Their validity relies on the consistency of either the direct method or IS, but not necessarily both. We refer to such a property as the double robustness property.

We note that none of the aforementioned work studied data integration, which is the central theme of this paper.

3. Estimators in Non-dynamic Setting

Summary. In this section, we present our newly developed non-pessimistic and pessimistic estimators, tailored for non-dynamic settings. These estimators differ in their approach to weighting historical and experimental data. The non-pessimistic estimator determines its weight by minimizing an estimated MSE, whereas the pessimistic estimator optimizes its weight to minimize a "pessimistic" version of the estimated MSE. We will discuss adaptations of these estimators for sequential decision making later.

Data. The offline data comprises an experimental dataset $\mathcal{D}_e$ and a historical dataset $\mathcal{D}_h$. In the experimental setting, the decision maker observes certain contextual information at each time point, denoted by S_e, and makes a choice, represented by A_e, between a baseline control policy $A_e = 0$ and a target policy $A_e = 1$, resulting in an immediate reward, R_e. Thus, the experimental data contains a set of i.i.d. context-action-reward triplets. In contrast, the historical data consists of i.i.d. context-reward pairs $O_h = (S_h, R_h)$ generated solely under the control policy.

Objective. Our objective is to estimate the difference between the mean outcome under the target policy and that under the control in the experimental data. This estimand is commonly referred to as the average treatment effect (ATE). Since no unmeasured confounders exists during the experiment, the ATE can be represented by $\tau_e = \mathbb{E}[r_e^*(1, S_e) - r_e^*(0, S_e)]$, where $r_e^*(a, s) = \mathbb{E}(R_e | A_e = a, S_e = s)$.

Two base estimators. We next introduce two base estimators for ATE. The first estimator is the EDO estimator $\widehat{\tau}_e$, which exclusively uses $\mathcal{D}_e$ to learn ATE. The second estimator $\widehat{\tau}_h$, on the other hand, incorporates $\mathcal{D}_h$ into the ATE estimation. Specifically, it uses $\mathcal{D}_e$ to estimate the target policy's value and $\mathcal{D}_h$ to estimate the control policy's value.

Mathematically, let O_e be a shorthand for the triplet (S_e, A_e, R_e) in the experimental data. We define an estimation function $\psi_e(\bullet)$ for O_e as follows

$$\sum_{a=0}^{1}(-1)^{a-1}\Big\{r_e(a, S_e) + \nu^a(A_e|S_e)[R_e - r_e(A_e, S_e)]\Big\},$$

where r_e represents our posit model for the reward function r_e^*. Moreover, ν^a denotes the model for the IS ratio $\mathbb{I}(A_e = a)/\mathbb{P}(A_e = a|S_e)$, where $\mathbb{I}(\bullet)$ is the indicator function and the denominator is the behavior policy (or the propensity score) that generates $\mathcal{D}_e$. The following definition gives the EDO estimator.

Definition 1 (**Experimental-data-only Estimator).** The doubly robust estimator based on the experimental data alone is defined as

$$\widehat{\tau}_e = \frac{1}{|\mathcal{D}_e|}\sum_{O_e \in \mathcal{D}_e} \psi_e(O_e). \tag{1}$$

Remark 2. $\widehat{\tau}_e$ covers a range of estimators including the direct method estimator, the IS estimator, and the DR estimator. In its general form, without any specific restrictions, $\widehat{\tau}_e$ functions as the doubly robust estimator. It can be easily verified that $\widehat{\tau}_e$ is consistent to ATE when either the reward function or the IS ratio is correctly specified. Meanwhile, $\widehat{\tau}_e$ can be simplified to the direct method estimator by setting $\nu^0 = \nu^1 = 0$, and to the IS estimator when r_e is set to 0.

Define $r_h^*(\bullet) = \mathbb{E}(R_h|S_h = \bullet)$ as the reward function in the historical data and let $\mu^*(\bullet)$ denote the density ratio of the probability mass/density function of S_e over that of S_h. The historical data distribution might differ from the experimental data distribution in the following two aspects:

1. **Reward shift**: the reward function $r_h^*(\bullet)$ might differ from $r_e^*(0, \bullet)$ conditional on the control.
2. **Covariate shift**: the distribution of S_e might differ from that of S_h, i.e., $\mu^* \neq 1$.

Let r_h and μ denote the posit models for r_h^* and μ^*. We next give the definition of historical-data-based estimator $\widehat{\tau}_h$

Definition 2 (**Historical-data-based Estimator).** The doubly robust estimator that uses the historical data to estimate the control policy's value is defined as

$$\widehat{\tau}_h = \frac{1}{|\mathcal{D}_e|} \sum_{O_e \in \mathcal{D}_e} \psi_{h,1}(O_e) - \frac{1}{|\mathcal{D}_h|} \sum_{O_h \in \mathcal{D}_h} \psi_{h,2}(O_h),$$

where $\psi_{h,1}(O_e) = r_e(1, S_e) + \nu^1(A_e|S_e)[R_e - r_e(A_e, S_e)] - r_h(S_e)$ and $\psi_{h,2}(O_h) = \mu(S_h)[R_h - r_h(S_h)]$.

Remark 3. By addition and subtraction, $\widehat{\tau}_h$ is unbiased to the difference between $\psi_{h,1}(O_e) + r_h(S_e)$ and $r_h(S_e) + \psi_{h,2}(O_h)$. To elaborate $\widehat{\tau}_h$, it is crucial to understand these two estimating functions. The first function uses the experimental data to construct the doubly robust estimator for the value of the target policy, while the second function incorporates the historical data to construct the doubly robust estimator for the average outcome under the control policy. These two terms are unbiased estimators for $\mathbb{E}[r_e(1, S_e)]$ and $\mathbb{E}[r_h(S_e)]$, respectively, provided that either the density ratio or the reward function is correctly specified.

Additionally, the use of $r_h(S_e) + \psi_{h,2}(O_h)$ partially addresses the distributional shift between the experimental and historical data. Specifically, by using S_e instead of S_h in $r_h(S_e)$ and using the density ratio μ in $\psi_{h,2}$, it addresses the covariate shift, leading to an unbiased estimator toward $\mathbb{E}[r_h(S_e)]$ instead of $\mathbb{E}[r_h(S_h)]$. However, it introduces a potential bias equal to

$$b_h = \mathbb{E}[r_e(0, S_e)] - \mathbb{E}[r_h(S_e)]. \tag{2}$$

This parameter represents the mean reward shift between the experimental and historical data, serving as a pivotal metric for quantifying discrepancies between the two datasets. It is also equal to the bias of the ATE estimator $\widehat{\tau}_h$ which incorporates the historical data to estimate the control policy's outcome. A small value of b_h implies a relatively safe use of historical data to enhance the precision of the ATE estimator. Conversely, a large b_h suggests caution against using historical data due to the significant bias it introduces.

The proposed estimators. Both the proposed non-pessimistic and pessimistic estimators are formulated as linear combinations of the two base estimators $\widehat{\tau}_e$ and $\widehat{\tau}_h$.

Definition 3 (**Weighted Estimator).** The weighted estimator is defined as

$$\widehat{\tau}_w = w\widehat{\tau}_e + (1 - w)\widehat{\tau}_h$$

for some properly chosen weight $w \in [0, 1]$

The weight is selected to minimize the MSE of the resulting estimator. Specifically, for a given w, according to the bias-variance decomposition, we obtain

$$\text{MSE}(\widehat{\tau}_w) = \text{Bias}^2(\widehat{\tau}_w) + \text{Var}(\widehat{\tau}_w), \tag{3}$$

where the bias is proportional to $(1 - w)$, given by $-(1 - w)b_h$ according to (2) and the variance term equals $w^2\text{Var}(\widehat{\tau}_e) + (1-w)^2\text{Var}(\widehat{\tau}_h) + 2w(1-w)\text{Cov}(\widehat{\tau}_e, \widehat{\tau}_h)$. This yields a close-form expression for (3).

We aim to estimate the oracle weight w^* that minimizes (3). We first note that the variance/covariance terms can be consistently estimated using the sampling variance formula[2]. It remains to estimate the reward shift bias b_h.

The non-pessimistic estimator employs the unbiased estimator $\widehat{b}_h = \widehat{\tau}_e - \widehat{\tau}_h$ for estimating b_h. Through certain derivations, this approach yields the subsequent estimator for w^*:

$$\widehat{w} = \frac{\widehat{b}_h^2 + \widehat{\text{Var}}(\widehat{\tau}_h) - \widehat{\text{Cov}}(\widehat{\tau}_e, \widehat{\tau}_h)}{\widehat{\text{Var}}(\widehat{\tau}_e) + \widehat{b}_h^2 + \widehat{\text{Var}}(\widehat{\tau}_h) - 2\widehat{\text{Cov}}(\widehat{\tau}_e, \widehat{\tau}_h)}, \tag{4}$$

where the precise expressions for the estimated variance and covariance terms are detailed in Appendix D.

We next discuss the limitations of the non-pessimistic estimator. We draw a parallel with the "offline bandit" problem where each weight w represents an arm in the bandit framework. Here, the MSE of $\widehat{\tau}_w$ is analogous to the cost of selecting an arm, with the aim being to identify the optimal arm (weight) that minimizes this cost.

For each arm, the estimated cost, $\widehat{\text{MSE}}(\widehat{\tau}_w)$, is calculated by incorporating $\widehat{b}_h$ and the estimated variance/covariance values into (3). The non-pessimistic estimator employs

[2]For an i.i.d. average $\bar{X} = \sum_{i=1}^n X_i/n$, its variance can be consistently estimated by $\sum_{i=1}^n (X_i - \bar{X})^2/n(n-1)$.

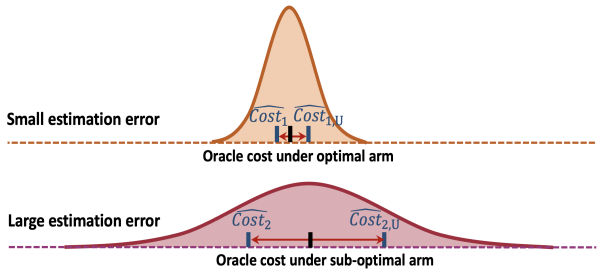

Figure 1. Distributions of estimated costs for optimal and sub-optimal arms. A key challenge arises when the estimated cost of a sub-optimal arm is inaccurately high, leading to failure of the greedy action selection method. To address this issue, we apply the pessimistic principle which takes into account the uncertainties inherent in these estimations. The estimates of the cost under the two arms are given by $\widehat{Cost}_k$ $(k = 1, 2)$ with their pessimistic versions $\widehat{Cost}_{k,U}$ $(k = 1, 2)$. By comparing the upper bounds of the estimated costs, we effectively identify the optimal arm.

a greedy action selection method, selecting the arm with the lowest estimated cost. The reliability of this estimator is largely dependent on a uniform consistency condition, which requires that the estimated costs uniformly converge to the actual costs across all arms. However, this condition is likely to be violated, if the estimated cost for any sub-optimal arm is inconsistent, as depicted in Figure 1.

In our framework, underestimating the absolute value of b_h leads to lower estimated MSEs for smaller weights. As a result, the weight $\widehat{w}$ chosen by the non-pessimistic estimator tends to be smaller than the ideal (oracle) value, resulting in a significant bias in $\widehat{\tau}_{\widehat{w}}$. This reveals the limitations of the non-pessimistic estimator, particularly when b_h is moderately large, as detailed in Table 1 and further elaborated in Section 5.1.

The pessimistic estimator addresses this limitation by incorporating the uncertainty of cost estimation. Instead of selecting the greedy arm with the lowest estimated cost, it selects the arm based on a more pessimistic cost estimate that upper bounds the oracle cost with a high probability. This method relaxes the stringent uniform consistency condition, as illustrated in Figure 1. Importantly, the consistency of the resulting estimator relies on the accurate estimation of the optimal arm's cost only.

In our setup, we compute an uncertainty quantifier U for the estimation error $\widehat{b}_h - b_h$. It satisfies the following condition,

$$\mathbb{P}(|\widehat{b}_h - b_h| \leq U) \leq 1 - \alpha, \tag{5}$$

for a given significance level $\alpha > 0$. In practice, U can be constructed using concentration inequalities or asymptotic normal approximation (Casella & Berger, 2021). We next use $(|\widehat{b}_h| + U)^2$ as a pessimistic estimator for b_h^2 and plug-in this estimator into the right-hand-side of (4) to determine the weight $\widehat{w}_U$. Under the event defined in (5), $(|\widehat{b}_h| + U)^2$ serves as a valid upper bound for b_h^2. This leads to the pessimistic estimator $\widehat{\tau}_{\widehat{w}_U}$. We show in Section 5.2 that this estimator is more robust than the non-pessimistic estimator, particularly when b_h is moderately large.

Confidence Interval. Under certain regularity conditions, each weighted estimator is asymptotically normal such that $(\widehat{\tau}_{\widehat{w}} - \tau_e)/\sqrt{Var(\widehat{\tau}_{\widehat{w}})} \xrightarrow{d} N(0, 1)$. This motivates us to consider the following Wald-type confidence interval for ATE

$$[\widehat{\tau}_{\widehat{w}} - \Phi^{-1}(1-\alpha)\sqrt{\widehat{Var}(\widehat{\tau}_{\widehat{w}})}, \widehat{\tau}_{\widehat{w}} + \Phi^{-1}(1-\alpha)\sqrt{\widehat{Var}(\widehat{\tau}_{\widehat{w}})}],$$

where Φ^{-1} is the inverse cumulative distribution function of a standard random variable, and the variance of $\widehat{\tau}_{\widehat{w}}$ is estimated based on the sampling variance formula.

4. Extension to Sequential Decision Making

We next briefly outline the extension of our methods to sequential decision making. To save space, more details are given in Appendix C. This extension aligns closely with our ridesharing example, where policy decisions are made sequentially over time, and past policies can influence future outcomes (Bojinov et al., 2023; Shi et al., 2023b). In this setting, the ATE is defined as the difference in expected cumulative rewards between the control and target policies.

The online experiment spans multiple days, with daily data summarized as sequences of state-action-reward triplets. Actions are binary, denoting either a baseline control or an experimental target policy. To account for day-to-day variations, we model the experiment as a time-varying Markov decision process. For ATE estimation, we employ the double RL estimator (Kallus & Uehara, 2020), leading to the development of the EDO estimator. The historical data comprises state-reward pair sequences from previous days under the control policy, forming the basis for our second estimator. This estimator, also doubly robust, is used to estimate the cumulative reward under the control policy, thereby facilitating ATE calculation. Building on the approaches outlined in Section 3, we apply both pessimistic and non-pessimistic strategies to integrate these base estimators.

5. Theoretical Properties

To simplify our theoretical analysis, this section examines a sample-split version of the proposed estimator in non-dynamic settings. Further extensions of our analysis to sequential decision making are detailed in Appendix C.

Our analysis compares three key estimators: a conceptual oracle estimator $\widehat{\tau}_{w^*}$, which utilizes the ideal w^* value, the EDO estimator detailed in (1), and the semi-parametrically efficient (SPE) estimator (Li et al., 2023) developed on the assumption of no reward shift. The EDO and SPE estimators represent two polar views of reward shift: the

EDO anticipates a notable divergence between $r_e(0, \bullet)$ and $r_h(\bullet)$, while the SPE assumes no difference.

Summary. Before delving into the technical details, we offer a concise summary of our theories:

- **Small** b_h: in scenarios where the reward shift is much smaller than the standard deviations of the doubly robust estimators, the SPE estimator achieves the best performance. However, our analysis shows that the MSEs of the proposed estimators closely approximate those of both the oracle and the SPE estimator.
- **Moderate** b_h: when the reward shift is comparable to or larger than the standard deviation terms, yet falls within the high confidence bounds of the estimation error, it remains uncertain which estimator (other than the oracle estimator) outperforms the rest. In these settings, the MSE of our pessimistic estimator is generally smaller than that of the non-pessimistic estimator.
- **Large** b_h: when the reward shift is much larger than the estimation error, both the EDO estimator and our estimators are equivalent to that of the oracle estimator. We refer to this equivalence as the oracle property.

5.1. Properties of the Non-pessimistic Estimator

We study a sample-split variant of our estimator, where the dataset is equally divided into two parts. The first half, labeled as $\mathcal{D}_e^{(1)} \cup \mathcal{D}_h^{(1)}$, is utilized to deduce the weight $\widehat{w}$. The second half, $\mathcal{D}_e^{(2)} \cup \mathcal{D}_h^{(2)}$, is then employed to construct the final doubly robust estimator $\widehat{\tau}$, leveraging the previously estimated weight. This sample-splitting approach removes the dependencies between the estimated weight and the dataset used in formulating the ATE estimator, considerably simplifying our theoretical analysis. It has been widely used in causal inference and OPE (see e.g., Luedtke & Van Der Laan, 2016; Chernozhukov et al., 2018; Kallus & Uehara, 2020; Bibaut et al., 2021; Shi et al., 2021). An alternative method involves swapping the roles of the data subsets $\mathcal{D}_e^{(1)} \cup \mathcal{D}_h^{(1)}$ and $\mathcal{D}_e^{(2)} \cup \mathcal{D}_h^{(2)}$ to generate a second estimator and then averaging both estimators to attain full efficiency. Nonetheless, this approach is not explored further in our paper for the sake of simplicity.

We impose the following assumptions.

Assumption 1 (Coverage)**.** *Let $\pi^*(a|s) = \mathbb{P}(A_e = a|S_e = s)$ be the propensity score. There exists a scalar $\epsilon > 0$ such that $\pi^*(a|s) \geq \epsilon$ and $\mu^*(s) \leq \epsilon^{-1}$ hold for any a and s.*

Assumption 2 (Boundedness)**.** *(i) There exists some constant $R_{\max}$ such that $\max(|R_e|, |R_h|) \leq R_{\max}$ holds almost surely. (ii) $\max(|r_e|, |r_h|)$ is upper bounded by $R_{\max}$. (iii) π and μ are lower bounded by ϵ.*

Assumption 3 (Doubly-robust Specification)**.** *Either the reward functions or the density ratios are correctly specified.*

Remark 4. The coverage condition in Assumption 1 is frequently imposed in the OPE literature (see e.g., Uehara et al., 2022). It is also referred to as the positivity assumption in the causal inference literature (Hernán & Robins, 2010).

Remark 5. The condition of bounded rewards in Assumption 2(i) is commonly imposed in RL (see e.g., Agarwal et al., 2019). Given the bounded nature of the reward and the density ratio/propensity score, it is reasonable to assume that the user-defined nuisance functions are similarly bounded, as detailed in Assumptions 2(ii) and 2(iii).

Remark 6. Assumption 3 reflects the double robustness property of the proposed estimator. Alternatively, this assumption can be replaced by requiring these nuisance functions to satisfy certain convergence rate conditions (Chernozhukov et al., 2018; Kallus & Uehara, 2020).

We begin by providing a non-asymptotic upper bound for MSE of the non-pessimistic estimator. Define $n_{\min} = \min\{|D_e|, |D_h|\}$ as the effective sample size.

Theorem 1 (MSE of the non-pessimistic estimator)**.** *Under Assumptions 1 – 3, the excess MSE of the non-pessimistic estimator compared to $\widehat{\tau}_{w^*}$, i.e., MSE$(\widehat{\tau}_{\widehat{w}})$ – MSE$(\widehat{\tau}_{w^*})$ can be upper bounded by*

$$\mathbb{E}[(1-w^*)^2 - (1-\widehat{w})^2](\widehat{b}_h^2 - b_h^2) + O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big). \quad (6)$$

The upper bound can be decomposed into two parts: the first one represents the error for estimating the mean reward shift b_h, whereas the second one upper bounds the errors for estimating the variance and covariance terms, namely $\mathrm{Var}(\widehat{\tau}_e)$, $\mathrm{Var}(\widehat{\tau}_h)$, and $\mathrm{Cov}(\widehat{\tau}_e, \widehat{\tau}_h)$.

We next compare this excess MSE against MSE$(\widehat{\tau}_{w^*})$. First, we observe that when MSE$(\widehat{\tau}_{w^*})$ is proportional to $R_{\max}^2/(\epsilon^2 n_{\min})$, the second term in (6) becomes negligible as $n_{\min}$ grows to infinity. Hence, it suffices to compare the first term in (6) in contrast to MSE$(\widehat{\tau}_{w^*})$. To elaborate the first term, we examine three scenarios previously introduced in this section, differentiated by the magnitude of b_h.

Small b_h. In this scenario, we assume $|b_h| \ll n_{\min}^{-1/2} R_{\max}/\epsilon$ and thus, the first term is asymptotically equivalent to

$$\mathrm{SEE}(\widehat{b}_h) = \mathbb{E}[(1-w^*)^2 - (1-\widehat{w})^2](\widehat{b}_h - b_h)^2. \quad (7)$$

We refer to this term as the spurious estimation error (SEE) of $\widehat{b}_h$, since it occurs due to the spurious correlation between $\widehat{w}$ and $\widehat{b}_h$. Theoretically, it is of the same order of magnitude as MSE$(\widehat{\tau}_{w^*})$. However, our empirical investigation reveals that it is considerably smaller than the oracle MSE, as illustrated in Figure 2.

Additionally, under the assumption that $r_h(s) = r_e(0, s)$ for all s — effectively resulting in $b_h = 0$ — the SPE estimator achieves the smallest MSE asymptotically, since it

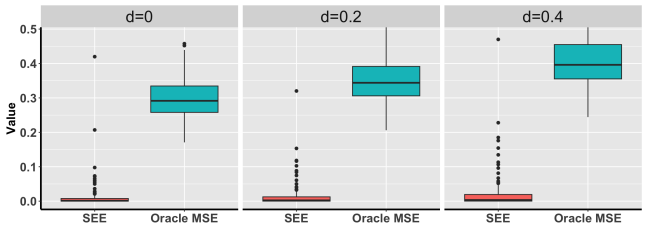

Figure 2. Boxplots of the SEE and the oracle MSE under the setting of Example 6.1 when the bias $b_h = 0$, and d indicates the difference of the conditional variance of the reward between the experimental data and historical data.

is tailored to minimize MSE under this assumption. Assuming all nuisance functions are correctly specified, and the proportionality assumption holds such that the ratio $\text{Var}(R_e|A_e = 0, S_e)/[\text{Var}(R_h|S_e)\mu(S_e)\pi(0|S_e)]$ remains constant irrespective of S_e, the SPE estimator is equivalent to the oracle estimator. Consequently, this suggests that the MSE of our non-pessimistic estimator asymptotically equals the efficiency bound augmented by a small spurious estimation error. We summarize these discussions below.

Corollary 1 (MSE with a small b_h)**.** *In the small b_h scenario, if MSE$(\widehat{\tau}_{w^*})$ is proportional to $R_{\max}^2/(\epsilon^2 n_{\min})$, then*

$$\left|\frac{MSE(\widehat{\tau}_{\widehat{w}}) - MSE(\widehat{\tau}_{w^*})}{MSE(\widehat{\tau}_{w^*})} - \frac{SEE(\widehat{b}_h)}{MSE(\widehat{\tau}_{w^*})}\right| \to 0,$$

as $n_{\min} \to \infty$. Additionally, when $r_h(\bullet) = r_e(0, \bullet)$, the proportionality assumption holds, and all nuisance functions are correctly specified, MSE$(\widehat{\tau}_{\widehat{w}})$ is asymptotically equivalent to the sum of the efficiency bound plus SEE$(\widehat{b}_h)$.

Large b_h. In this scenario, we require $|b_h| \gg n_{\min}^{-1/2}\sqrt{\log n_{\min}}R_{\max}/\epsilon$. Notice that the lower bound is aligned with the high confidence bound for the estimation error of $\widehat{b}_h$. Consequently, the reward shift is sufficiently large to be "detectable" from the data.

Under this condition, the b_h^2 term becomes the dominant factor in the MSE (3), leading to the optimal weight w^* approaching 1. Consequently, the EDO estimator is asymptotically equivalent to the oracle estimator whereas the SPE estimator is sub-optimal since it assumes a zero b_h.

In the large b_h scenario, the weight selected by the non-pessimistic estimator tends towards one, so that the excess MSE is of a small order. Hence, the MSE of the non-pessimistic estimator is asymptotically the same as that of the oracle estimator, achieving the oracle property.

Corollary 2 (Oracle property with a large b_h)**.** *In the large b_h scenario, both MSE$(\widehat{\tau}_{\widehat{w}})$/MSE$(\widehat{\tau}_{w^*})$ and MSE$(\widehat{\tau}_{\widehat{w}})$/MSE$(\widehat{\tau}_e)$ approach 1 as $n_{\min} \to \infty$.*

Moderate b_h. In this scenario, the magnitude of $|b_h|$ falls between $n_{\min}^{-1/2}R_{\max}/\epsilon$ and $n_{\min}^{-1/2}\sqrt{\log n_{\min}}R_{\max}/\epsilon$. This scenario is the most challenging, as it is not clear whether the SPE estimator or the EDO estimator will deliver superior performance.

To illustrate the issues of the non-pessimistic estimator, let us examine a scenario where $|b_h|$ significantly exceeds $n_{\min}^{-1/2}R_{\max}/\epsilon$, causing the optimal weight $w^* \to 1$. In this context, even though $|b_h|$ is considerably large, it remains within the high-confidence interval of the estimation error for $\widehat{b}_h - b_h$, which might not make $|b_h|$ adequately distinguishable from the data. As $w^* \to 1$, the dominant factor in the first part of (6) becomes $-\mathbb{E}(1-\widehat{w})^2(\widehat{b}_h^2 - b_h^2)$. Nevertheless, there is no guarantee that $\widehat{w}$ will converge to 1 with high confidence. This uncertainty introduces a significant excess MSE for the non-pessimistic estimator. See the numerical results in Section 6.

5.2. Robustness of the Pessimistic Estimator

The pessimistic estimator effectively mitigates the aforementioned limitation of the non-pessimistic estimator by incorporating the estimation error of $\widehat{b}_h$ into weight selection. To elaborate, we first provide a non-asymptotic upper bound for its MSE.

Theorem 2 (MSE of the pessimistic estimator)**.** *Under Assumptions 1 – 3 and (5), MSE$(\widehat{\tau}_{\widehat{w}_U})$ − MSE$(\widehat{\tau}_{w^*})$ can be upper bounded by*

$$(1-w^*)^2\mathbb{E}[(|\widehat{b}_h| + U)^2 - b_h^2] + O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big) + O(\alpha[b_h^2 + R_{\max}^2/\epsilon^2 n_{\min}]). \tag{8}$$

According to Theorem 2, the excess MSE of the pessimistic estimator can be decomposed into three parts:

1. **Estimation error of** $\widehat{b}_h$: the first term quantifies the estimation error for $\widehat{b}_h$. Unlike the non-pessimistic estimator where this term depends on the estimated weight, here it relies only on w^*. This distinction enhances the robustness of the estimator, particularly when b_h is moderately large.
2. **Estimation errors of the variance/covariance terms**: similar to the non-pessimistic estimator, the second term quantifies the estimation error for the variance/covariance terms.
3. **Type-I error**: the last term is directly proportional to the type-I error α which upper bounds the probability that the $|\widehat{b}_h - b_h|$ exceeds U. Notice that this term can be made sufficiently small by employing concentration inequalities without substantially increasing the estimation error associated with $\widehat{b}_h$.

To further illustrate the advantage of the pessimistic estimator, we consider the moderate b_h scenario. When

$|b_h| \gg n_{\min}^{-1/2} R_{\max}/\epsilon$, $\widehat{w}$ might not necessarily converge to 1 with high confidence. Hence, the non-pessimistic estimator can suffer from a large loss, due to the involvement of $\widehat{w}$ in (6). On the contrary, (8) depends solely on w^*, which results in a smaller excess loss. Indeed, the following corollary shows that the pessimistic estimator achieves the oracle property even when b_h is moderately large.

Corollary 3 (Oracle property of the pessimistic estimator)**.** *Suppose that U is proportional to the order $n_{\min}^{-1/2}\sqrt{\log n_{\min}} R_{\max}/\epsilon$ such that $\alpha = o(1/n_{\min})$, if further $b_h \gg n_{\min}^{-1/2}(\log n_{\min})^{1/6} R_{\max}/\epsilon$, then the pessimistic estimator achieves the oracle property.*

The condition $b_h \gg n_{\min}^{-1/2}(\log n_{\min})^{1/6} R_{\max}/\epsilon$ applies to both the moderate and large b_h scenarios. Hence, even in cases of moderate b_h, as long as it is much larger than $n_{\min}^{-1/2}(\log n_{\min})^{1/6} R_{\max}/\epsilon$, the oracle property is satisfied. This formally establishes the robustness of the pessimistic estimator in comparison to the non-pessimistic estimator.

6. Experiments

In this section, we investigate the finite sample performance of the proposed estimators. Comparison is made among the following ATE estimators:

- **NonPessi**: the proposed non-pessimistic estimator.
- **Pessi**: the proposed pessimistic estimator.
- **EDO**: the doubly robust estimator $\widehat{\tau}_e$ constructed based on the experimental data only (see (1)).
- **Lasso**: a weighted estimator $\widehat{\tau}_{Lasso} = w\widehat{\tau}_e + (1-w)\widehat{\tau}_h$ that linearly combines the ATE estimator $\widehat{\tau}_e$ based on experimental data and $\widehat{\tau}_h$ based on historical data, where the weight w is chosen to minimize the estimated variance of the final ATE estimator with the Lasso penalty (Cheng & Cai, 2021),
- **SPE**: the semi-parametrically efficient estimator proposed by Li et al. (2023) developed under the assumption of no reward shift between the experimental and historical data, i.e., $r_e(0, s) = r_h(s)$ for any s.

Notice that it remains unclear how to extend SPE in sequential decision making. Consequently, our implementation of SPE is confined to non-dynamic settings only. We compare the MSEs of the ATE estimators based on 100 simulation replications. Details about the data generating process can be found in Appendix A.

Example 6.1 (**Non-dynamic simulation**). We consider a non-dynamic setting where the sample size of the experimental data is $|\mathcal{D}_e| = 48$, and the sample size of the historical data is set to be $|\mathcal{D}_h| = m|\mathcal{D}_e|$ with $m \in \{1, 2, 3\}$. A deterministic switchback design is adopted to generate $\mathcal{D}_e$. We vary the mean reward shift b_h within the range from 0 to 1.5, incrementing by 0.1 at each step. We also vary the conditional variance of the reward and use d to characterize this difference (see Appendix A for its detailed definition).

Figure 3 visualizes the empirical means of the MSEs for different methods. According to our theory, the effectiveness of different estimators is determined by the magnitude of the reward shift. To validate our theory, we further classify b_h into different regimes as follows:

- **Small b_h regime**: $|b_h| \le c_1\sqrt{\text{Var}(\widehat{b}_h)}$;
- **Moderately large b_h regime**: $c_1 < \frac{|b_h|}{\sqrt{\text{Var}(\widehat{b}_h)}} \le c_2$;
- **Large b_h regime**: $|b_h| > c_2\sqrt{\text{Var}(\widehat{b}_h)}$.

According to our theoretical analysis, we set $c_1 = 1$ and $c_2 = \sqrt{\log(n_{\min})}$. This ensures that scenarios where variance dominates the bias are categorized within the small reward shift region. Conversely, when the bias exceeds the established high confidence bound, it is classified under the large reward shift regime.

We depict the boundaries between different regimes in Figure 3. It can be seen that in the small b_h regime, the SPE estimator is the top performer. However, the MSE of our proposed non-pessimistic estimator is close to that of the SPE estimator. As b_h grows to moderate levels, our pessimistic estimator achieves smaller or comparable MSEs compared to other alternatives. Finally, in the large b_h regime, our pessimistic estimator achieves comparable performance to the EDO estimator, both outperforming other estimators in terms of MSE. These findings establish a concrete link between our theories and empirical observations. Particularly, they numerically verify our theoretically identified optimal method within each respective regime. Additionally, the bottom panel of Figure 3 specifically reports the mean MSEs for methods excluding SPE, offering an in-depth comparison of the other estimators' performance. Here, Lasso is implemented with a carefully selected tuning parameter, which has been determined to yield reasonably good performance. However, as illustrated by additional numerical results in Figure A2 in the Appendix, this estimator is sensitive to the choice of the tuning parameter.

Example 6.2 (**Ridesharing-data based sequential simulation**). In this example, we build a simulation environment based on a real dataset collected from a ridesharing company. The experimental data lasts for $|\mathcal{D}_e| = 30$ days and is generated from a switchback design. We divide each day into $T = 24$ time intervals. The state variable consists of the number of order requests and the driver's total online time within each one-hour time interval. The reward is defined as the total income earned by the drivers within each time interval. To generate the historical data, we assume it

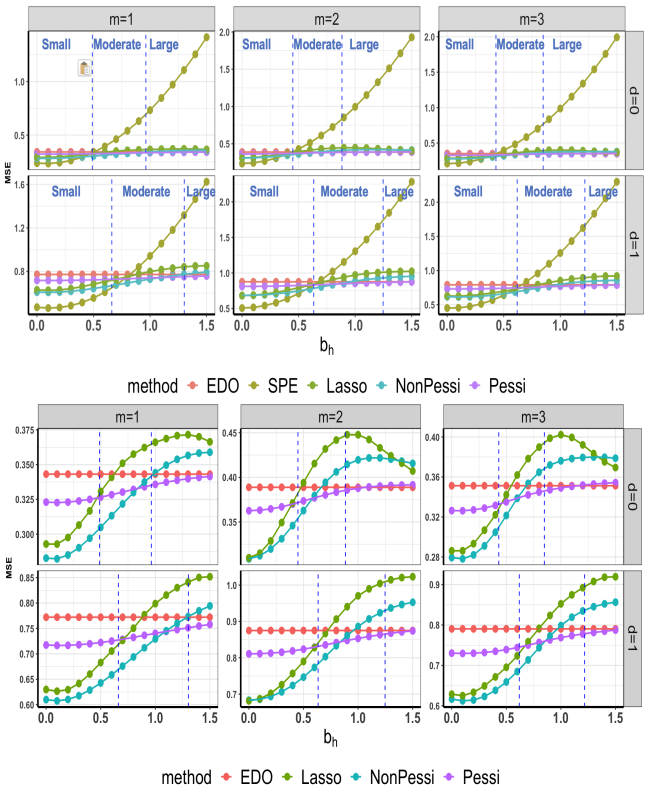

Figure 3. Empirical means of MSEs for different methods under the switchback design in Example 6.1. The top panel displays all the methods, whereas the bottom panel focuses on the area excluding the SPE method.

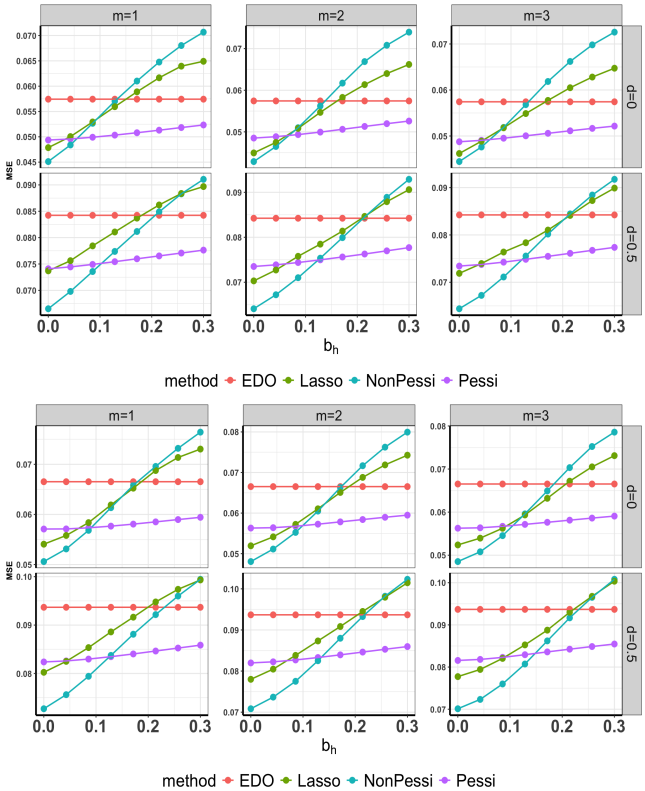

Figure 4. Empirical Means of MSEs for different methods in Example 6.2. The treatment effect ratios are equal to 5% (Top), 10% (Bottom), respectively.

contains another $|\mathcal{D}_h| = m|\mathcal{D}_e|$ days with $m \in \{1, 2, 3\}$, and set b_h as a linearly increasing sequence ranging from 0 to 0.3, consisting of eight values, and choose the conditional variance difference parameter d from the set $\{0, 0.5\}$. Refer to Appendix A for the detailed data generating process.

Figure 4 reports the empirical means of the MSEs of different estimators. The Lasso method is again implemented with a reasonably good tuning parameter. It can be seen that the proposed "NonPessi" estimator outperforms Lasso in most cases. When b_h is large, the two methods perform comparably. In contrast, our proposed "Pessi" showcases robustness in dealing with the distributional shift. When b_h increases, the MSEs of Lasso and "NonPessi" increase significantly, while "Pessi" has only a slight increase in the MSE. It is also consistently better than the EDO estimator, demonstrating the usefulness of data integration.

Example 6.3 (**Ridesharing-data based non-dynamic simulation**). We also conduct a real-data based non-dynamic simulation study to compare different estimators. The findings are very similar. To save space, we relegate the detailed results to Appendix A.

Example 6.4 (**Clinical-data based non-dynamic simulation**). The data for this experiment is sourced from the AIDS Clinical Trials Group Protocol 175, involving 2139 HIV-infected individuals. Participants were randomly assigned to one of four treatment groups: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine, or ddI monotherapy (Hammer et al., 1996). Following the analyses in (Lu et al., 2013) and (Shi et al., 2019), we use the CD4 count (cells/mm3) at 20 ± 5 weeks post-baseline as the continuous response variable, and consider three contextual variables: age (in years), homosexual activity (0=no, 1=yes), and hemophilia (0=no, 1=yes).

Based on this clinical dataset, we create a simulation environment to test our methodology. The findings from this clinical-data-based non-dynamic experiment align with the patterns observed in the above examples. Details about the experimental settings and figures summarizing the findings can be found in Appendix A.

Furthermore, we conduct an additional experiment in Example A.1 of Appendix A to evaluate the coverage probabilities of the confidence intervals (CIs). While maintaining nominal coverage, the pessimistic estimator yields narrower confidence intervals compared to the EDO estimator, indicating an improvement in efficiency by incorporating historical data. We also develop a hybrid procedure that chooses different methods according to the magnitude of the bias in Appendix B.

Acknowledgement

We thank the anonymous referees and the meta reviewer for their constructive comments, which have led to a significant improvement of the earlier version of this article. Li's research is partially supported by the National Science Foundation of China 12101388, CCF- DiDi GAIA Collaborative Research Funds for Young Scholars and Program for Innovative Research Team of Shanghai University of Finance and Economics. Shi's research is partially supported by an EPSRC grant EP/W014971/1.

Impact Statement

This paper introduces innovative methods for policy evaluation, particularly focusing on the integration of multiple data sources to enhance decision-making processes. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

Aminian, G., Vega, R., Rivasplata, O., Toni, L., and Rodrigues, M. Semi-counterfactual risk minimization via neural networks. *arXiv preprint arXiv:2209.07148*, 2022.

Athey, S., Chetty, R., and Imbens, G. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*, 2020.

Bai, C., Wang, L., Yang, Z., Deng, Z., Garg, A., Liu, P., and Wang, Z. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. *arXiv preprint arXiv:2202.11566*, 2022.

Bian, Z., Shi, C., Qi, Z., and Wang, L. Off-policy evaluation in doubly inhomogeneous environments. *arXiv preprint arXiv:2306.08719*, 2023.

Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A., and van Der Laan, M. Post-contextual-bandit inference. *Advances in Neural Information Processing Systems*, 34: 28548–28559, 2021.

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

Bojinov, I., Simchi-Levi, D., and Zhao, J. Design and analysis of switchback experiments. *Management Science*, 69 (7):3759–3777, 2023.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22:33–57, 1996.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations*, 2020.

Casella, G. and Berger, R. L. *Statistical inference*. Cengage Learning, 2021.

Chen, X. and Christensen, T. M. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015.

Chen, X. and Qi, Z. On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. In *International Conference on Machine Learning*, pp. 3558–3582. PMLR, 2022.

Cheng, D. and Cai, T. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021.

Cheng, Y., Wu, L., and Yang, S. Enhancing treatment effect estimation: A model robust approach integrating randomized experiments and external controls using the double penalty integration estimator. In *Uncertainty in Artificial Intelligence*, pp. 381–390. PMLR, 2023.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

Cuffe, R. L. The inclusion of historical control data may reduce the power of a confirmatory study. *Statistics in Medicine*, 30(12):1329–1338, 2011.

Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A., and Steingrimsson, J. A. Towards causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. *Epidemiology (Cambridge, Mass.)*, 31(3):334, 2020.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., and Steingrimsson, J. A. Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, 2023.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. Coindice: Off-policy confidence interval estimation. *Advances in Neural Information Processing Systems*, 33:9398–9411, 2020.

Degtiar, I. and Rose, S. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.

DerSimonian, R. and Laird, N. Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, 45:139–145, 2015.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 1097–1104, 2011.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Dunn, G. and Bentall, R. Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine*, 26(26):4719–4745, 2007.

Fan, C., Lu, W., Song, R., and Zhou, Y. Concordance-assisted learning for estimating optimal individualized treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1565–1582, 2017.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Farias, V., Li, A., Peng, T., and Zheng, A. Markovian interference in experiments. *Advances in Neural Information Processing Systems*, 35:535–549, 2022.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pp. 3102–3111. PMLR, 2020.

Gui, G. Combining observational and experimental data using first-stage covariates. *Available at SSRN 3662061*, 2020.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996.

Han, L., Hou, J., Cho, K., Duan, R., and Cai, T. Federated adaptive causal estimation (face) of target treatment effects. *arXiv preprint arXiv:2112.09313*, 2021.

Han, L., Shen, Z., and Zubizarreta, J. R. Multiply robust federated estimation of targeted average treatment effects. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvári, C., and Wang, M. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pp. 4074–4084. PMLR, 2021.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. From sate to patt: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society:SeriesA (Statistics in Society)*, 178:757–778, 2015.

Hasegawa, T., Claggett, B., Tian, L., Solomon, S. D., Pfeffer, M. A., and Wei, L.-J. 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.

Heckman, J. J., Ichimura, H., and Todd, P. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

Hernán, M. A. and Robins, J. M. Causal inference, 2010.

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

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

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Jin, Y., Ren, Z., Yang, Z., and Wang, Z. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *The Journal of Machine Learning Research*, 21 (1):6742–6804, 2020.

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

Kallus, N., Puli, A. M., and Shalit, U. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 31, 2018.

Kennedy, E. H. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Larsen, N., Stallrich, J., Sengupta, S., Deng, A., Kohavi, R., and Stevens, N. T. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician*, pp. 1–15, 2023.

Le, H., Voloshin, C., and Yue, Y. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., and Cai, J. Improving trial generalizability using observational studies. *Biometrics*, 79(2):1213–1225, 2023.

Li, M., Shi, C., Wu, Z., and Fryzlewicz, P. Testing stationarity and change point detection in reinforcement learning. *arXiv preprint arXiv:2203.01707*, 2022.

Li, T., Shi, C., Lu, Z., Li, Y., and Zhu, H. Evaluating dynamic conditional quantile treatment effects with applications in ridesharing. *Journal of the American Statistical Association*, (just-accepted):1–26, 2024.

Li, X., Miao, W., Lu, F., and Zhou, X.-H. Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403, 2023.

Lian, Q., Zhang, J., Hodges, J. S., Chen, Y., and Chu, H. Accounting for post-randomization variables in meta-analysis: A joint meta-regression approach. *Biometrics*, 79(1):358–367, 2023.

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

Liao, P., Qi, Z., Wan, R., Klasnja, P., and Murphy, S. A. Batch policy learning in average reward markov decision processes. *The Annals of Statistics*, 50(6):3364–3387, 2022.

Liu, J., Zhang, J., Mitchell, A., Fang, M., and Tian, L. Causal inference for longitudinal data based on historical controls. *Journal of Biopharmaceutical Statistics*, 33(3): 289–306, 2023.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Lu, W., Zhang, H. H., and Zeng, D. Variable selection for optimal treatment decision. *Statistical methods in medical research*, 22(5):493–504, 2013.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020.

Luedtke, A. R. and Van Der Laan, M. J. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44 (2):713, 2016.

Luo, S., Yang, Y., Shi, C., Yao, F., Ye, J., and Zhu, H. Policy evaluation for temporal and/or spatial dependent experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. 1–27, 2024.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pp. 4932–4941. PMLR, 2019.

Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009.

Peysakhovich, A. and Lada, A. Combining observational and experimental data to find heterogeneous treatment effects. *arXiv preprint arXiv:1611.02385*, 2016.

Pocock, S. J. The combination of randomized and historical controls in clinical trials. *Journal of chronic diseases*, 29 (3):175–188, 1976.

Qin, Z. T., Zhu, H., and Ye, J. Reinforcement learning for ridesharing: An extended survey. *Transportation Research Part C: Emerging Technologies*, 144:103852, 2022.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Rigter, M., Lacerda, B., and Hawes, N. Rambo-rl: Robust adversarial model-based offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 16082–16097, 2022.

Rott, K. W., Bronfort, G., Chu, H., Huling, J. D., Leininger, B., Murad, M. H., Wang, Z., and Hodges, J. S. Causally interpretable meta-analysis: Clearly defined causal effects and two case studies. *Research Synthesis Methods*, 15(1): 61–72, 2024.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

Schmidli, H., Häring, D. A., Thomas, M., Cassidy, A., Weber, S., and Bretz, F. Beyond randomized clinical trials: use of external controls. *Clinical Pharmacology & Therapeutics*, 107(4):806–816, 2020.

Scott, D. A. and Lewin, A. Borrowing from historical control data in a bayesian time-to-event model with flexible baseline hazard function. *arXiv preprint arXiv:2401.06082*, 2024.

Shi, C., Song, R., Lu, W., and Fu, B. 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.

Shi, C., Song, R., and Lu, W. On testing conditional qualitative treatment effects. *Annals of statistics*, 47(4):2348, 2019.

Shi, C., Lu, W., and Song, R. 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.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pp. 9580–9591. PMLR, 2021.

Shi, C., Zhang, S., Lu, W., and Song, R. 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, 2022a.

Shi, C., Wan, R., Song, G., Luo, S., Zhu, H., and Song, R. A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets. *The Annals of Applied Statistics*, 17(4):2701–2722, 2023a.

Shi, C., Wang, X., Luo, S., Zhu, H., Ye, J., and Song, R. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, 118(543):2059–2071, 2023b.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pp. 19967–20025. PMLR, 2022b.

Shi, X., Pan, Z., and Miao, W. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023c.

Steele, R. J., Schnitzer, M. E., and Shrier, I. Importance of homogeneous effect modification for causal interpretation of meta-analyses. *Epidemiology*, 31(3):353–355, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(1): 1731–1755, 2015a.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, 2015b.

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

Tang, X., Qin, Z., Zhang, F., Wang, Z., Xu, Z., Ma, Y., Zhu, H., and Ye, J. A deep value-network based approach for multi-driver order dispatching. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1780–1790, 2019.

Tang, Z., Duan, Y., Zhang, S., and Li, L. A reinforcement learning approach to estimating long-term treatment effects. *arXiv preprint arXiv:2210.07536*, 2022.

Tarrier, N., Lewis, S., Haddock, G., Bentall, R., Drake, R., Kinderman, P., Kingdon, D., Siddle, R., Everitt, J., Leadley, K., et al. Cognitive–behavioural therapy in first-episode and early schizophrenia: 18-month follow-up of a randomised controlled trial. *The British Journal of Psychiatry*, 184(3):231–239, 2004.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Tsiatis, A. A. *Semiparametric theory and missing data*. Springer, 2006.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2021.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Uehara, M., Shi, C., and Kallus, N. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Uehara, M., Kiyohara, H., Bennett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C., and Sun, W. Future-dependent value-based off-policy evaluation in pomdps. *Advances in Neural Information Processing Systems*, 36, 2024.

van der Laan, M. J., Rose, S., Zheng, W., and van der Laan, M. J. Cross-validated targeted minimum-loss-based estimation. *Targeted Learning: Causal Inference for Observational and Experimental Data*, pp. 459–474, 2011.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., and Lesaffre, E. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*, 27(10):3167–3182, 2018.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2014.

Wan, R., Zhang, S., Shi, C., Luo, S., and Song, R. Pattern transfer learning for reinforcement learning in order dispatching. *arXiv preprint arXiv:2105.13218*, 2021.

Wang, J., Qi, Z., and Wong, R. K. Projected state-action balancing weights for offline reinforcement learning. *The Annals of Statistics*, 51(4):1639–1665, 2023a.

Wang, J., Shi, C., and Wu, Z. A robust test for the stationarity assumption in sequential decision making. In *International Conference on Machine Learning*, pp. 36355–36379. PMLR, 2023b.

Wen, Q., Shi, C., Yang, Y., Tang, N., and Zhu, H. An analysis of switchback designs in reinforcement learning. *arXiv preprint arXiv:2403.17285*, 2024.

Wu, C.-F. J. et al. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.

Wu, H. and Wang, M. Variance regularized counterfactual risk minimization via variational divergence minimization. In *International Conference on Machine Learning*, pp. 5353–5362. PMLR, 2018.

Wu, L. and Yang, S. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*, pp. 904–926. PMLR, 2022.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021.

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J., Liu, C., Bian, W., and Ye, J. 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*, pp. 905–913, 2018.

Yang, S., Ding, P., et al. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020a.

Yang, S., Zeng, D., and Wang, X. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*, 2020b.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Zhang, J., Ko, C.-W., Nie, L., Chen, Y., and Tiwari, R. Bayesian hierarchical methods for meta-analysis combining randomized-controlled and single-arm studies. *Statistical Methods in Medical Research*, 28(5):1293–1310, 2019.

Zhao, P., Chambaz, A., Josse, J., and Yang, S. Positivity-free policy learning with observational data. *arXiv preprint arXiv:2310.06969*, 2023.

Zhou, F., Luo, S., Qie, X., Ye, J., and Zhu, H. Graph-based equilibrium metrics for dynamic supply–demand systems with applications to ride-sourcing platforms. *Journal of the American Statistical Association*, 116(536):1688–1699, 2021.

Zhou, W. Bi-level offline policy optimization with limited exploration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zhou, Y., Qi, Z., Shi, C., and Li, L. Optimizing pessimism in dynamic treatment regimes: A bayesian learning approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 6704–6721. PMLR, 2023.

A. Additional Experiment Results

In this section, we present details of the data-generating process for Section 6 and additional experiment results.

Example 6.1 **(Continued).** We consider the reward function as follows,

$$R_e = 10 + b_h + A_e + S_e + (2+d)\varepsilon_e, \quad R_h = 10 + S_h + \varepsilon_h,$$

where S_e, S_h's and $\varepsilon_e, \varepsilon_h$'s are from standard normal distribution $N(0,1)$. The sample size $|\mathcal{D}_e| = 48$ with a horizon of $T = 1$. The sample size of historical data is set to be $|\mathcal{D}_h| = m|\mathcal{D}_e|$ with $m \in \{1,2,3\}$. We consider the switchback design, which alternates the treatment and the control along the time. We set b_h to range over the set $\{0, 0.1, 0.2, \ldots 1.5\}$. A larger value of b_h indicates a greater difference in the average cumulative reward of the control between the historical data and the experimental data. Meanwhile, we use $d \in \{0,1\}$ to characterize the difference of the conditional variance of the reward between the historical data and the experimental data.

Example 6.2 **(Continued).** We perform the analysis based on the real dataset obtained from a prominent ridesharing company. The dataset covers the period from May 17th, 2019, to June 25th, 2019, with one-hour time units, resulting in a total of $T = 24$ hours per day, collected over the span of $N = 40$ days. To protect privacy, specific details about the company and cities are omitted, and all data, including states and rewards, are scaled to ensure privacy. In particular, the state variable consists of the number of order requests and the driver's total online time within each one-hour time interval. The reward is defined as the total income earned by the drivers within each time interval. Noting daily temporal trends in these variables, as shown in Figure A1, a time-varying Markov Decision Process (MDP) model is employed to understand the dataset dynamics. The dataset is based on the A/A experiments, in which a single order dispatch policy was consistently applied over time ($A_t = 0$ for all t). Similar to the data generating process in Luo et al. (2024), we create a simulation environment using the wild bootstrap method (Wu et al., 1986) based on this A/A dataset. In general, we assume the following time-varying linear models:

$$\begin{cases} R_t = \alpha_t + S_t^\top \beta_t + \gamma_t A_t + e_t, \\ S_{t+1} = \phi_t + \Phi_t S_t + \Gamma_t A_t + E_t, \end{cases} \tag{9}$$

where α_t and γ_t are real-valued scalars, β_t, ϕ_t, and Γ_t are vectors in the space $\mathbb{R}^d$, $\Phi_t \in \mathbb{R}^{d\times d}$, e_t is the time-dependent random noise, and E_t is the time-independent random error vector. Specifically, we first fit the data based on the linear models in (9) by setting $\gamma_t = \Gamma_t = 0$ and derive the estimates $\{\widehat{\alpha}_t\}_t$, $\{\widehat{\beta}_t\}_t$, $\{\widehat{\phi}_t\}_t$ and $\{\widehat{\Phi}_t\}_t$. We then calculate the residuals in the reward and state regression models based on these estimators as follows:

$$\widehat{e}_{i,t} = R_{i,t} - \widehat{\alpha}_t - S_{i,t}^\top \widehat{\beta}_t, \quad \widehat{E}_{i,t} = S_{i,t+1} - \widehat{\phi}_t - \widehat{\Phi}_t S_{i,t}. \tag{10}$$

To simulate data reflecting varied treatment effects, we introduce a treatment effect ratio λ, which is defined as the proportional impact on the average return of the baseline policy. Define treatment effect parameters $\widehat{\gamma}_t = \delta_1 \times (\sum_i R_{i,t}/(100\times N))$ and $\widehat{\Gamma}_t = \delta_2 \times (\sum_i S_{i,t}/(100\times N))$. We examine λ at 5% and 10% levels. Correspondingly, δ_1 and δ_2 are adjusted to ensure that both the direct and carryover effects increment by $\lambda/2$, cumulatively elevating the Average Treatment Effect (ATE) by λ.

To structure the experimental ($|\mathcal{D}_e| = 30$ days) and historical ($|\mathcal{D}_h| = m|\mathcal{D}_e|$ days) datasets with $m \in \{1,2,3\}$, we employ an alternating time interval design with a 3-hour span for the experimental dataset, and a global control for the historical dataset (i.e., $A_{i,t} = 0$ for all i,t). We introduce i.i.d. standard Gaussian noise $\{\xi_i\}_{i=1}^{|D_e|}$ or $\{\xi_i\}_{i=1}^{|D_h|}$ for each dataset. For each day i, a random integer from set I (where $I \in \{1,2,\ldots,N\}$) is selected to determine the initial state $S_{I,1}$. For the historical dataset, the b-th bootstrap sample of rewards and states is generated following specific equations.

$$\begin{cases} \widehat{R}_{i,t}^b = \widehat{\alpha}_t + (\widehat{S}_{i,t}^b)^\top \widehat{\beta}_t + \widehat{\gamma}_t A_{i,t} + \xi_i^b \hat{e}_{i,t}, \\ \widehat{S}_{i,t+1}^b = \widehat{\phi}_t + \widehat{\Phi}_t \widehat{S}_{i,t}^b + \widehat{\Gamma}_t A_{i,t} + \xi_i^b \widehat{E}_{i,t}, \end{cases}$$

with the estimated $\{\widehat{\alpha}_t\}_t$, $\{\widehat{\beta}_t\}_t$, $\{\widehat{\phi}_t\}_t$, $\{\widehat{\Phi}_t\}_t$, the specified $\{\widehat{\gamma}_t\}_t$ and $\{\widehat{\Gamma}_t\}_t$, and the error residuals given by $\{\xi_i^b \widehat{e}_{i,t} : 1 \le t \le T\}$ and $\{\xi_i^b \widehat{E}_{i,t} : 1 \le t \le T\}$, respectively. In the experimental dataset, we exclusively generate the b-th Bootstrap sample of reward with shifted mean parameter b_h and standard deviation parameter d, according to the equation,

Algorithm 1 Bootstrap-assisted procedure.

Input: Real data $\{(S_{it}, R_{it}) : 1 \le i \le N; 1 \le t \le T\}$, the adjustment parameters for the ratios (δ_1, δ_2), the assignment of actions, the bootstrapped sample size ($n = |D_e|$ or $n = |D_h|$, where $|D_h| = m|D_e|$), shifted mean parameter b_h and standard deviation parameter d, random seed, the number of replications $B = 200$.

1: **Initialization:** Calculating the least square estimates $\{\hat{\alpha}\}_t$, $\left\{\hat{\beta}_t\right\}_t$, $\left\{\hat{\phi}_t\right\}_t$, $\left\{\hat{\Phi}_t\right\}_t$ in the model (9), treatment effect parameters $\{\hat{\gamma}_t\}_t$ and $\left\{\hat{\Gamma}_t\right\}_t$ and the residuals of reward model and state regression model by the (10)

2: **for** $1 \le b \le B$ **do**

3: Sampling the number of days n from $\{1, \cdots, N\}$ with replacement, and generating the i.i.d. normal random variables $\xi_i^b \sim N(0, 1)$;

4: Generating the pseudo rewards $\left\{\hat{R}_{i,t}^b\right\}_{i,t}$ and states $\left\{\hat{S}_{i,t}^b\right\}_{i,t}$ according to the following equations,

$$\hat{R}_{i,t}^b = [1, (\hat{S}_{i,t}^b)^\top, A_{i,t}] \begin{pmatrix} \hat{\alpha}_t \\ \hat{\beta}_t \\ \hat{\gamma}_t \end{pmatrix} + D_i b_h + (1 + D_i d)\xi_i^b \hat{e}_{i,t}, \quad \hat{S}_{i,t+1}^b = [\hat{\phi}_t, \hat{\Phi}_t, \hat{\Gamma}_t] \begin{pmatrix} 1 \\ \hat{S}_{i,t}^b \\ A_{i,t} \end{pmatrix} + \xi_i^b \hat{E}_{i,t},$$

where $D_i = 1$ corresponds to index in the experimental dataset, while $D_i = 0$ corresponds to index in the historical dataset

5: Calculating the set of estimators $\left\{\text{ATE}^b\right\}_b$ by proposed methods and state-of-the-art methods.

6: **end for**

Output: The empirical means of MSEs of different ATE estimators.

$\widehat{R}_{i,t}^b = \widehat{\alpha}_t + (\widehat{S}_{i,t}^b)^\top \widehat{\beta}_t + \widehat{\gamma}_t A_{i,t} + b_h + (1 + d)\xi_i^b \widehat{e}_{i,t}$, and we continue to generate states based on the settings of historical data. A summary of the bootstrap-assisted procedure is provided in Algorithm 1.

Furthermore, to explore the Lasso method's efficacy across various tuning parameters, we select a set of ℓ_1-tuning parameters $\{0.8, 1.2, 1.6, 2, 4\}$. Additionally, we take the mean difference b_h from a sequence ranging from 0 to 0.6 in increments, comprising 8 values. Figure A2 illustrates the performance comparison between varying ℓ_1-tuning parameters and the EDO method. We observe that for smaller b_h values, Lasso methods with lower tuning parameters outperform the EDO method. However, their performance declines as b_h increases. In contrast, Lasso methods with larger tuning parameters maintain a consistent efficiency, comparable to the EDO method, regardless of b_h values.

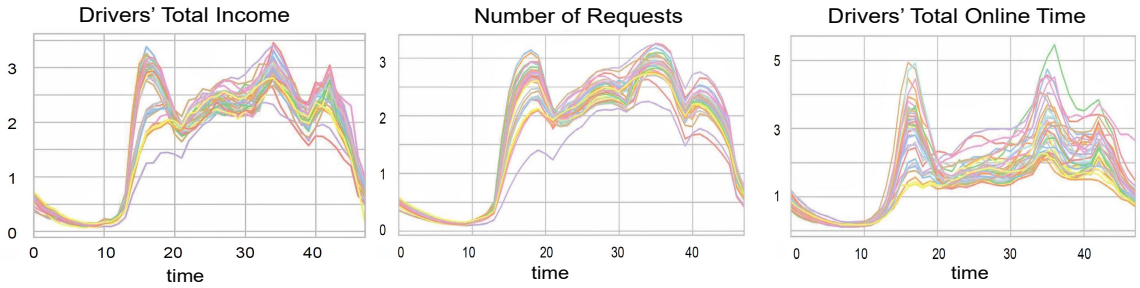

Figure A1. Visual representations of scaled states and rewards in one city across 40 days, comprising drivers' total income, the number of requests, and drivers' total online time. Each line represents data from a specific day.

Example 6.3 **(Continued).** The data generating process is adapted from the approach outlined in Example 6.2. Crucially, although the data is inherently sequential, we adapt it to a non-dynamic setting by treating each day as an independent instance with $T = 1$. This adjustment involves maintaining the same policy daily and defining the daily average total income of drivers over all time intervals as the reward. The state variable is represented by the number of order requests and the drivers' total online time during the initial time interval. The sample size of the experimental data is $|\mathcal{D}_e| = 30$ and that of the historical data is $|D_h| = m|D_e|$ with $m \in \{1, 2, 3\}$, and we take the mean difference b_h from an arithmetic sequence ranging from 0 to 0.6, with a length of 8, and the conditional variance difference $d \in \{0, 0.5\}$ to explore different scenarios.

Figure A3 illustrates the performance of our proposed estimators, EDO, SPE, and Lasso with a reasonably good tuning

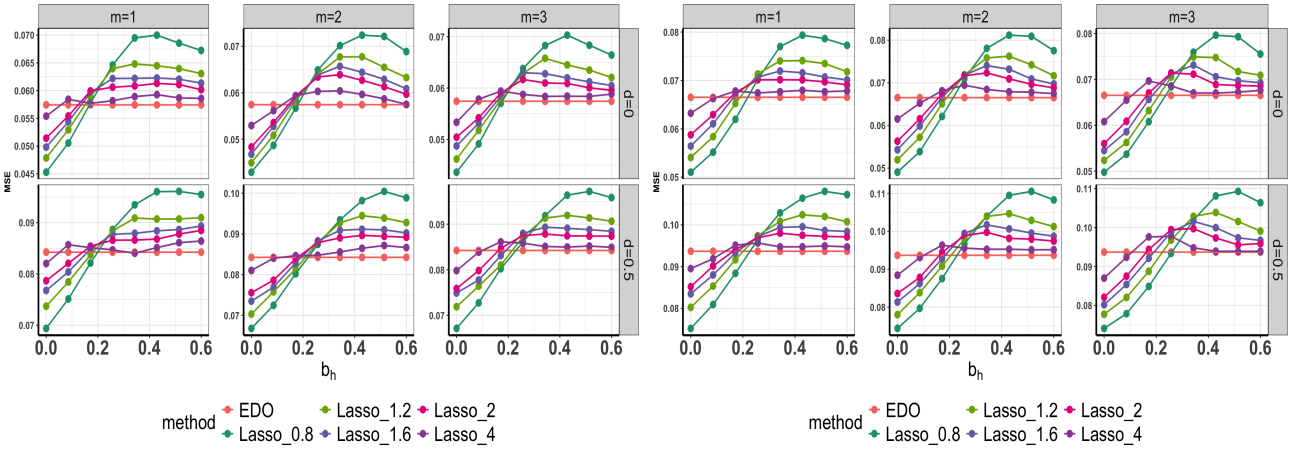

Figure A2. Empirical Means of MSEs for Lasso with different tuning parameters under the switchback design in Example 6.2. The treatment effect ratio parameters are equal to 5% (Left) and 10% (Right), respectively.

parameter. The outcomes align with those observed in Example 6.1, demonstrating comparable insights. Meanwhile, Figure A4 depicts the performance of these methods, excluding SPE, further validating our theoretical insights and highlighting the resilience of the "Pessi" estimator. Notably, as the treatment effect size grows, the "Pessi" approach exhibits increased stability across different values of b_h.

various methods, excluding SPE, under the switchback design as detailed in Example 6.3. The figures display results for treatment effect ratio parameters set at 5% (left) and 10% (right) respectively.

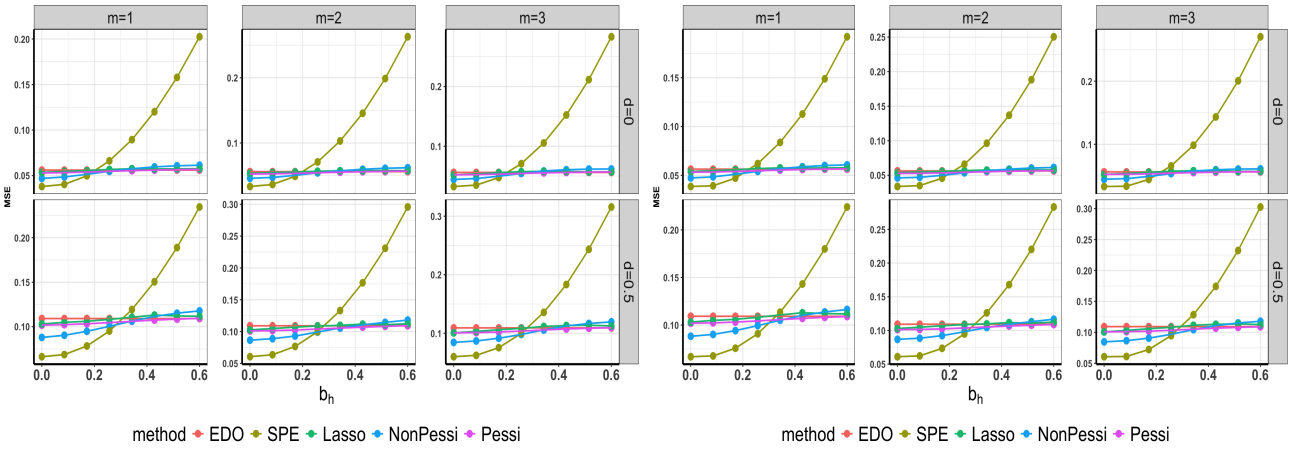

Figure A3. Empirical means of MSEs for various methods under the switchback design as detailed in Example 6.3. The figures display results for treatment effect ratio parameters set at 5% (left) and 10% (right), respectively.

Example 6.4 **(Continued).** We apply the proposed method to data from the AIDS Clinical Trials Group Protocol 175 (ACTG175), involving 2139 HIV-infected individuals. In this study, participants were randomly assigned to one of four treatment groups: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine, or ddI monotherapy (Hammer et al., 1996). The CD4 count (cells/mm3) at 20 ± 5 weeks post-baseline is chosen to be the continuous response variable R. Based on the significant factors identified in Lu et al. (2013), Fan et al. (2017) and Shi et al. (2019), the state variables are chosen to be age (in years), homosexual activity (0=no, 1=yes), and hemophilia (0=no, 1=yes).

We focus on two groups of patients receiving treatments ZDV+ddI (with a sample size of 522) and ZDV+zal (with a sample size of 524). To construct a simulation environment of these data, we use various nonlinear models to fit the data, taking

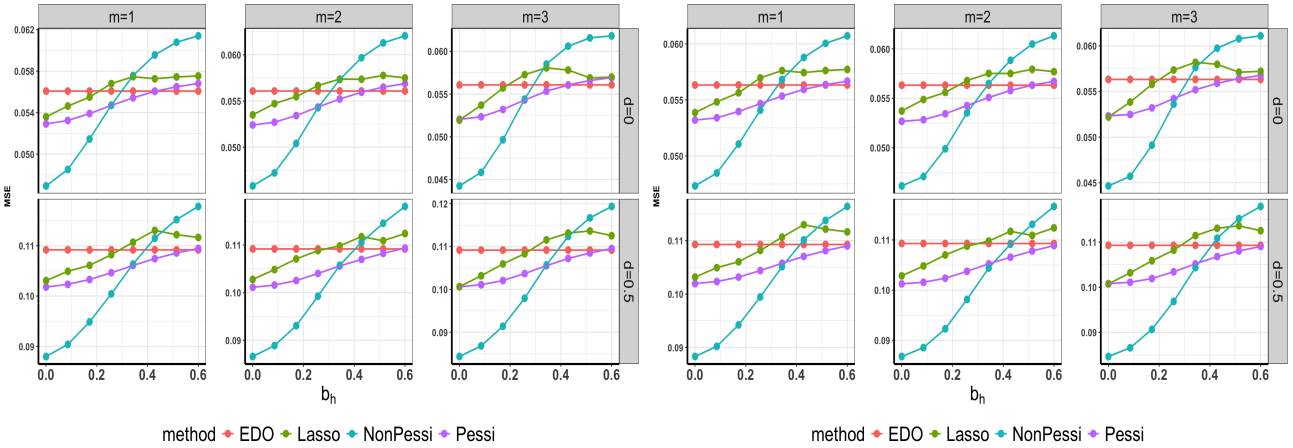

Figure A4. Empirical means of MSEs for various methods, excluding SPE, under the switchback design in Example 6.3. The treatment effect ratio parameters are set to 5% (Left) and 10% (Right), respectively.

into account the complex relationships between the state variables and the response. The goodness-of-fit for each model is assessed through examining their residual plots. This rigorous evaluation process has led us to select the following nonlinear model:

$$R = f(S_1, S_2, S_3) + \gamma A S_1 + \epsilon, \tag{11}$$

where $f(S_1, S_2, S_3) = (1 + \beta_1 S_1)^2 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_1 S_2 + \beta_5 S_1 S_3$. In this model, R represents the CD4 count, S_1, S_2, S_3 correspond to age, homosexual activity, and hemophilia, and A indicates whether a patient is receiving ZDV+ddI ($A = 1$) or ZDV+zal ($A = 0$).

After obtaining the estimators of the unknown parameters $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4, \widehat{\beta}_5, \widehat{\gamma})$, we calculate the fitted values $\widehat{f}(S_1, S_2, S_3)$ by plugging in the estimates along with the estimated residuals $\widehat{\epsilon}$. This enables us to generate the experimental data and the historical data similar to the bootstrap-assisted procedure in Algorithm 1 (Appendix A, page 15),

$$\begin{aligned} R_e &= \widehat{f}(S_{e,1}, S_{e,2}, S_{e,3}) + \widehat{\gamma} A_e S_{e,1} + b_h + (1+d)\widehat{\epsilon}_e, \\ R_h &= \widehat{f}(S_{h,1}, S_{h,2}, S_{h,3}) + \widehat{\epsilon}_h, \end{aligned}$$

where the sets $(S_{e,1}, S_{e,2}, S_{e,3}, \widehat{\epsilon}_e)$ and $(S_{h,1}, S_{h,2}, S_{h,3}, \widehat{\epsilon}_h)$ are sampled from the state variables and the estimated residuals with replacement based on (11). Furthermore, $A_e = 1$ and $A_e = 0$ are generated with equal probability, b_h is varied from 0 to 0.15 in increments to produce a series of eight values, and $d \in \{0, 1\}$. The selected sample size is $|\mathcal{D}_e| = 200$ and $|\mathcal{D}_h| = m|\mathcal{D}_e|$ with $m \in \{1, 2, 3\}$.

Figure A5 presents the empirical means of MSEs for all methods in the clinical-data based experiment. It further validates the effectiveness of the proposed methods, with findings aligning closely with those in Section 6.

Example A.1 (Inference for the pessimistic method). We further conduct an additional experiment to evaluate the coverage probabilities of the confidence intervals (CIs) based on our pessimistic estimator and compare them against the CIs based on the EDO and Lasso estimators. Data settings mirror that in Example 6.1, except that $m = 2$ is fixed and $|\mathcal{D}_e| = 100$. The left panel of Figure A6 displays the coverage probability of the 5% confidence intervals across the three methods. Our findings indicate that the CIs based on both the pessimistic and EDO estimators achieve the expected nominal coverage (e.g., 0.95). In contrast, the CI based on the Lasso estimator exhibits significant undercoverage. Additionally, while maintaining nominal coverage, the pessimistic estimator yields narrower confidence intervals compared to the EDO estimator, indicating an improvement in efficiency by incorporating historical data.

B. A hybrid procedure

If we have prior knowledge about the reward shift b_h, we can introduce a hybrid procedure that leverages the strengths of each method within their optimal ranges. It consists of the following steps. For given thresholds c_1, c_2,

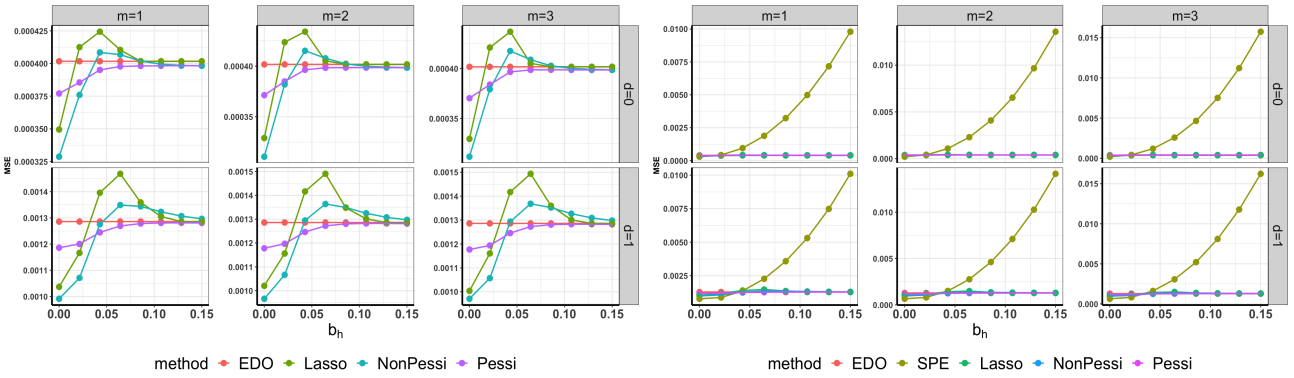

Figure A5. Empirical means of MSEs for various methods in the clinical-data based experiment in Example 6.4.

- If $|b_h| \leq c_1\sqrt{Var(\widehat{b}_h)}$, the bias is expected to be small and the SPE estimator (Li et al., 2023) is adopted.
- If $c_1\sqrt{Var(\widehat{b}_h)} < |b_h| \leq c_2\sqrt{Var(\widehat{b}_h)}$, indicating moderate bias, the proposed pessimistic method is then applied.
- If $|b_h| > c_2\sqrt{Var(\widehat{b}_h)}$, signifying substantial bias and the EDO estimator $\widehat{\tau}_e$ is employed.

According to our theoretical analysis, we can choose $c_1 = 1$ and $c_2 = \sqrt{\log(n_{\min})}$. This ensures that scenarios, where variance dominates the bias, are categorized within the small reward shift region. Conversely, when the bias exceeds the established high confidence bound, it is classified under the large reward shift regime.

We conduct an additional experiment that includes the hybrid procedure, under the assumption that there is pre-existing knowledge about the specific regime to which the data corresponds. The experimental parameters are aligned with those outlined in Example 6.1, as detailed on page 7. Figure A7 presents the empirical means of MSEs of all methods, with dotted vertical lines depicting the boundaries. These empirical results formally verify our theoretical assertions, demonstrating the superiority of the optimal method in each identified region. Additionally, the findings demonstrate that the hybrid procedure, when applied with prior knowledge of the data regime, consistently outperforms the rest of the methods.

In the absence of prior knowledge regarding b_h, estimating the regime to which the data belongs becomes necessary to implement the hybrid method. However, this estimation introduces additional variability. Consequently, our analysis reveals that the hybrid method, under these conditions, does not consistently outperform other methods in all scenarios. While optimizing the hybrid method for scenarios lacking prior regime knowledge presents a valuable avenue for future exploration, such an endeavor falls outside the scope of this current paper.

C. Extension to Sequential Decision Making

Let $O_{e,t}$ be a shorthand of $(S_{e,t}, A_{e,t}, R_{e,t})$ and denote $\{O_{e,t}\}$ containing all the data from $t = 1, \ldots, T$ in the experimental data. The EDO estimator can be represented by

$$\widehat{\tau}_e = \frac{1}{|\mathcal{D}_e|} \sum_{\{O_{e,t}\} \in \mathcal{D}_e} \psi_e(\{O_{e,t}\}),$$

where $\psi_e(\{O_{e,t}\}) = \sum_{a=0}^{1}(-1)^{a-1}\{V_{e,1}^a(S_{e,1}) + \sum_{t=1}^{T}\mu_t^a(A_{e,t}, S_{e,t})[R_{e,t} + V_{e,t+1}^a(S_{e,t+1}) - V_{e,t}^a(S_{e,t})]\}$, $\mu_t^1(A_{e,t}, S_{e,t})$ and $\mu_t^0(A_{e,t}, S_{e,t})$ are, respectively, the density ratio of the state-action pair of time t under the treatment and control policy under the behavior policy of the experimental data, and $V_{e,t}^a(s) = \sum_{k=t}^{T}\mathbb{E}(R_{e,k}|S_{e,t} = s)$ is the value function for the experimental data. The estimator based on the historical data can be constructed as

$$\widehat{\tau}_h = \frac{1}{|\mathcal{D}_e|} \sum_{O_{e,1} \in \mathcal{D}_e} \psi_{h,1}(O_{e,1}) - \frac{1}{|\mathcal{D}_h|} \sum_{\{O_{h,t}\} \in \mathcal{D}_h} \psi_{h,2}(\{O_{h,t}\}),$$

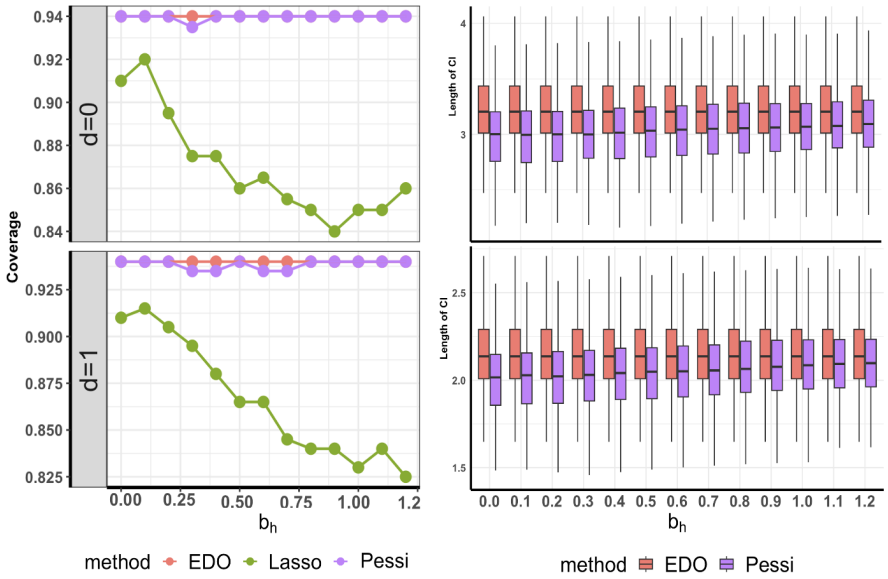

Figure A6. Coverage probabilities of the 95% confidence interval (left panel) and the boxplots of the length of the confidence intervals (right panel).

where $\psi_{h,1}(O_{e,1}) = V^1_{e,1}(S_{e,1}) + \sum_{t=1}^T \mu^1_t(A_{e,t}, S_{e,t})[R_{e,t} + V^1_{e,t+1}(S_{e,t+1}) - V^1_{e,t}(S_{e,t})] - V_{h,1}(S_e)$, $\psi_{h,2}(\{O_{h,t}\}) = \mu^h_t(S_{h,t})[R_{h,t} + V_{h,t+1}(S_{h,t+1}) - V_{h,t}(S_{h,t})]$, and $V_{h,t}(\bullet) = \sum_{k=t}^T \mathbb{E}(R_{h,k}|S_{h,t} = \bullet)$ is the value function for the historical data. The proposed estimator can be represented by $\widehat{\tau}_w = w\widehat{\tau}_e + (1-w)\widehat{\tau}_h$. In sequential decision-making, the bias caused by the reward shift is given by $b_h = \mathbb{E}[V^0_{e,1}(S_e)] - \mathbb{E}[V_{h,1}(S_e)]$.

The construction of $\widehat{\tau}_e$ and $\widehat{\tau}_w$ for the sequential decision making have similar patterns as that in Section 3. Following the methodologies developed in Section 3, we adopt similar pessimistic and non-pessimistic strategies to determine the weight.

We next give the theoretical properties of the proposed estimators for sequential decision making. Similar to the analysis in the non-dynamic setting, we compare the proposed estimator with the oracle estimator $\widehat{\tau}_{w^*}$. Before that, we impose the following conditions that extend the Assumptions 1-3 to the sequential setting.

Assumption 4. *There exists some constant $\varepsilon > 0$ such that the true density ratios $\mu^{a,*}_t(a,s) \geq \epsilon$ and $\mu^{h,*}_t(s) \geq \epsilon$ for any t, a, s.*

Assumption 5. *(i) There exists some constant $R_{\max}$ such that $|R_{e,t}| \leq R_{\max}$ and $|R_{h,t}| \leq R_{\max}$ for $1 \leq t \leq T$ almost surely. (ii) $|V^a_{e,t}|$ and $V_{h,t}$ are bounded by $(T+1-t)R_{\max}$. (iii) μ^a_t and μ^h_t are lower bounded by ε.*

Assumption 6. *(i) Either μ^a_t or $V^a_{e,t}$ is correctly specified. (ii) Either μ^h_t or $V_{h,t}$ is correctly specified.*

The following theorem provides a non-asymptotic upper bound for MSE of the non-pessimistic estimator.

Theorem 3 (MSE of the non-pessimistic estimator)**.** *Under Assumptions 4-6, the excess MSE of the non-pessimistic estimator in sequential decision making satisfies*

$$MSE(\widehat{\tau}_{\widehat{w}}) - MSE(\widehat{\tau}_{w^*}) \leq \mathbb{E}\Big[(1-w^*)^2 - (1-\widehat{w})^2\Big](\widehat{b}_h^2 - b_h^2) + O\Big(\frac{T^2 R^2_{\max}}{\varepsilon^2 n^{3/2}_{\min}}\Big).$$

Compared to the MSE of the non-pessimistic estimator in Theorem 1, the first term is about the estimation error for the mean shift, and the second term is about the estimation errors for the variance and variance terms, which is inflated by a factor of T^2 due to the sequential setting.

Corollary 4 (b_h Conditions)**.** *(i) If $|b_h| \ll n^{-1/2}_{\min} T R_{\max}/\epsilon$ and $MSE(\widehat{\tau}_{w^*})$ is proportional to $T^2 R^2_{\max}/(\varepsilon^2 n_{\min})$, then*

$$\Big|\frac{MSE(\widehat{\tau}_{\widehat{w}}) - MSE(\widehat{\tau}_{w^*})}{MSE(\widehat{\tau}_{w^*})} - \frac{SEE(\widehat{b}_h)}{MSE(\widehat{\tau}_{w^*})}\Big| \to 0,$$

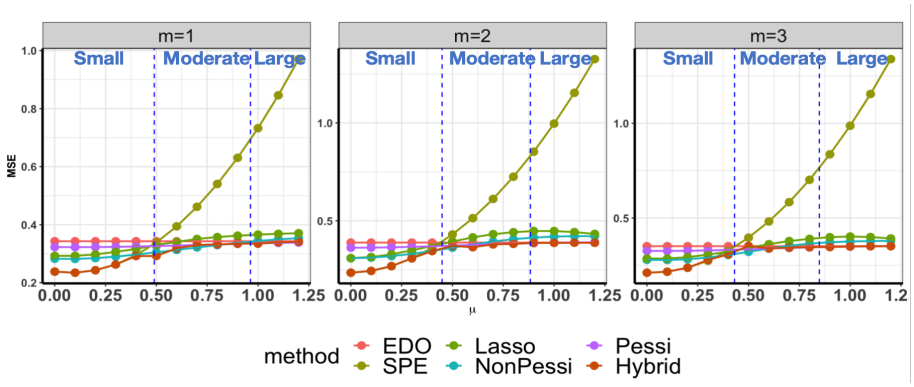

Figure A7. Empirical means of the MSEs across all the methods including the hybrid procedure. The vertical dotted blue lines denote the boundaries.

as $n_{\min} \to \infty$*, where SEE*$(\widehat{b}_h) = \mathbb{E}[(1-w^*)^2 - (1-\widehat{w})^2](\widehat{b}_h - b_h)^2$.

(ii) If $|b_h| \gg n_{\min}^{-1/2}\sqrt{\log(n_{\min})}TR_{\max}/\epsilon$*, both MSE*$(\widehat{\tau}_{\widehat{w}})$*/MSE*$(\widehat{\tau}_{w^*})$ *and MSE*$(\widehat{\tau}_{\widehat{w}})$*/MSE*$(\widehat{\tau}_e)$ *approach 1 as* $n_{\min} \to \infty$.

Part (i) of Corollary 4 gives the excess MSE against MSE$(\widehat{\tau}_{w^*})$ for a small b_h. Part (ii) of Corollary 4 shows the oracle property of the proposed non-pessimistic estimator for a large b_h.

Theorem 4 (MSE of the pessimistic estimator)**.** *Under Assumptions 4-6, the excess MSE of the pessimistic estimator in sequential decision making satisfies*

$$MSE(\widehat{\tau}_{\widehat{w}_U}) - MSE(\widehat{\tau}_{w^*}) \le (1-w^*)^2\mathbb{E}[(|\widehat{b}_h| + U)^2 - b_h^2] + O\Big(\frac{T^2R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big) + O(\alpha[b_h^2 + T^2R_{\max}^2/\epsilon^2 n_{\min}]).$$

Furthermore, if $b_h \gg n_{\min}^{-1/2}(\log n_{\min})^{1/6}R_{\max}T/\epsilon$ *and* U *is proportional to the order* $n_{\min}^{-1/2}\sqrt{\log n_{\min}}R_{\max}T/\epsilon$*, then the pessimistic estimator achieves the oracle property.*

Theorem 4 gives the excess MSE of the pessimistic estimator, which is also related to the estimation error of the bias, estimation errors of the variance and covariance terms, as well as the type I error α. It further establishes the oracle property of the estimator for moderate and large b_h cases.

The proofs of Theorem 3, Corollary 4 and Theorem 4 are similar to proofs under the non-dynamic setting, we omit here.

D. Implementation Details

In this section, we present the construction of variance and covariance terms in (4) and the estimation for sequence decision making.

Construction of variance and covariance terms in (4). We can obtain $\widehat{\tau}_e$ and $\widehat{\tau}_h$ by plugging in the unknown terms with their estimates

$$\widehat{\tau}_e = \frac{1}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}\widehat{\psi}_e(O_e), \quad \widehat{\tau}_h = \frac{1}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}\widehat{\psi}_{h,1}(O_e) - \frac{1}{|\mathcal{D}_h|}\sum_{O_h\in\mathcal{D}_h}\widehat{\psi}_{h,2}(O_h).$$

Then we can obtain $\widehat{\mathrm{Var}}(\widehat{\tau}_e)$, $\widehat{\mathrm{Var}}(\widehat{\tau}_h)$, and $\widehat{\mathrm{Cov}}(\widehat{\tau}_e, \widehat{\tau}_h)$ by the corresponding sample variances and sample covariance,

$$\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\tau}_e) &= \frac{1}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}(\widehat{\psi}_e(O_e) - \widehat{\tau}_e)^2,\\
\widehat{\mathrm{Var}}(\widehat{\tau}_h) &= \frac{1}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}(\widehat{\psi}_{h,1}(O_e) - \mathbb{P}_e\widehat{\psi}_{h,1}(O_e))^2 + \frac{1}{|\mathcal{D}_h|}\sum_{O_h\in\mathcal{D}_h}(\widehat{\psi}_{h,2}(O_h) - \mathbb{P}_h\widehat{\psi}_{h,2}(O_h))^2,\\
\widehat{\mathrm{Cov}}(\widehat{\tau}_e, \widehat{\tau}_h) &= \frac{1}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}(\widehat{\psi}_e(O_e) - \widehat{\tau}_e)(\widehat{\psi}_{h,1}(O_e) - \mathbb{P}_e\widehat{\psi}_{h,1}(O_e)),
\end{aligned}$$

where $\mathbb{P}_e\widehat{\psi}_{h,1}(O_e) = |\mathcal{D}_e|^{-1}\sum_{O_e\in\mathcal{D}_e}\psi_{h,1}(O_e)$ and $\mathbb{P}_h\widehat{\psi}_{h,2}(O_h) = |\mathcal{D}_h|^{-1}\sum_{O_h\in\mathcal{D}_h}\psi_{h,2}(O_h)$.

Estimation for sequence decision making. The estimator $\widehat{\tau}_e$ can be constructed using the existing estimation methods for doubly robust estimators, such as (Kallus & Uehara, 2020; Jiang & Li, 2016) by plugging in the estimates $\{\widehat{V}^a_{e,t}\}$ of the value function and estimates $\widehat{\mu}^a_t$ of the marginal density ratio. The estimator $\widehat{\tau}_h$ can be constructed similarly by plugging the estimates of the value function, and the estimates of the μ^h_t. Next, we present the estimation of μ^h_t.

Recall that $\mu^h_t(S_t)$ is the ratio of the density of the state S_t under the control policy in the experimental dataset to that in the historical dataset. It is noteworthy that for any function $f(S_t)$, we have

$$\mathbb{E}^h\Big[\mu^h_t(S_t)f(S_t)\Big] = \mathbb{E}\Big[\mu^0_t(A_{t-1},S_{t-1})f(S_t)\frac{I(A_t=0)}{\pi(A_t|S_t)}\Big],$$

where $\mathbb{E}^h$ is the expectation taken over the historical data, and $\pi(A_t|S_t)$ is the behavior policy that generates the experimental data. Based on the above estimating equation, after we obtain μ^0_t, $\mu^h_t(S_t)$ can be estimated recursively. Specifically, we propose to approximate $\mu^h_t(S_t)$ using linear seievs (Chen & Christensen, 2015; Chen & Qi, 2022; Bian et al., 2023; Shi et al., 2023b) such that $\widehat{\mu}^h_t(S_t) = \Phi(S_t)^\top\widehat{\gamma}_t$ for some basis function $\Phi(S_t)$. For $k = 1,\ldots,T$, taking $f(S_k) = \Phi(S_k)$, $\widehat{\gamma}_k$ is determined by solving

$$\frac{1}{|\mathcal{D}_h|}\sum_{\{O_{h,t}\}\in\mathcal{D}_h}\sum_{t=k}^{T}\Big(\Phi(S_{h,t})^\top\gamma_k\Phi(S_{h,t})\Big) = \frac{1}{|\mathcal{D}_e|}\sum_{\{O_{e,t}\}\in\mathcal{D}_e}\sum_{t=k}^{T}\left(\widehat{\mu}^0_t(A_{e,t-1},S_{e,t-1})\Phi(S_{e,t})\frac{I(A_{e,t}=0)}{\pi(A_{e,t}|S_{e,t})}\right).$$

If we have additional information that the distribution of the state variable under the control policy in the experimental data is the same as that of the state variable in the historical data, we can directly set $\mu^h_t(S_t)$=1 in the sequential setting and $\mu(S_t) = 1$ in the non-dynamic setting.

E. Proofs of the Theorems in Section 5.1

E.1. Notations and Auxiliary Lemmas

We begin by listing the notations used throughout the proof:

- $\mathcal{D}_e$: the experimental data with size;
- $\mathcal{D}_e^{(1)}$: the data subset to learn the weight;
- $\mathcal{D}_e^{(2)}$: the data subset to construct ATE estimator;
- $\mathcal{D}_h$: the historical data with size;
- $\mathcal{D}_h^{(1)}$: the data subset to learn the weight;
- $\mathcal{D}_h^{(2)}$: the data subset to construct ATE estimator;
- $O_e = (S_e, A_e, R_e)$: a context-action-reward triplet in $\mathcal{D}_e$;
- $O_h = (S_h, R_h)$: a context-reward pair in $\mathcal{D}_h$;
- $\pi^*(a|s)$: the behavior policy that generates $\mathcal{D}_e$;
- $r^*_e(a,s)$: $\mathbb{E}(R_e|A_e = a, S_e = s)$;
- $\tau^*(s)$: $r_e(1,s) - r_e(0,s)$;
- $r^*_h(s)$: $\mathbb{E}(R_h|S_h = s)$;
- r_e, r_h, π and μ: the nuisance functions used to construct the doubly robust estimator;
- $\mu(s)$: the density ratio of the probability
- $\widehat{\tau}_e$: the experimental-data-only estimator;
- ψ_e: the estimation function for $\widehat{\tau}_e$;
- $\widehat{\tau}_h$: the estimator which incorporates the historical data;
- $\psi_{h,1}$ and $\psi_{h,2}$: the two estimation functions for $\widehat{\tau}_h$;
- b_h: the mean shift $\mathbb{E}[r^*_e(0,S_e) - r^*_h(S_e)]$;
- $\widehat{b}_h$: the estimated bias;
- $\widehat{\mathrm{Var}}, \widehat{\mathrm{MSE}}, \widehat{\mathrm{Cov}}$: the empirical variance, MSE and covariance estimator.

We will sometimes write $\psi_e(O_e;\pi,r_e)$, $\psi_{h,1}(O_e;\pi,r_e)$ and $\psi_{h,2}(O_h;\mu,r_h)$ by $\psi_e(O_e)$, $\psi_{h,1}(O_e)$ and $\psi_{h,2}(O_h)$ when the nuisance functions used are clear from the context.

We next introduce the following auxiliary lemmas:

Lemma 1. *Under Assumption 3, for a given weight w, the MSE of the resulting doubly robust estimator $\widehat{\tau}_w$ is given by*

$$\frac{2w^2}{|D_e|}Var(\psi_e(O_e;\pi,r_e)) + \frac{2(1-w)^2}{|D_e|}Var(\psi_{h,1}(O_e;\pi,r_e)) + \frac{2(1-w)^2}{|D_h|}Var(\psi_{h,2}(O_h;\mu,r_h)) + \frac{4w(1-w)}{|D_e|}Cov(\psi_e(O_e;\pi,r_e),\psi_{h,1}(O_e;\pi,r_e)) + (1-w)^2 b_h^2.$$

Lemma 2. *Under Assumptions 1-3, we have*

$$\mathbb{E}|\widehat{Var}(\psi_e(O_e;\pi,r_e)) - Var(\psi_e(O_e;\pi,r_e))| = O\Big(\frac{R_{\max}^2}{\sqrt{|D_e|}\epsilon^2}\Big),$$

$$\mathbb{E}|\widehat{Var}(\psi_{h,1}(O_e;\pi,r_e)) - Var(\psi_{h,1}(O_e;\pi,r_e))| = O\Big(\frac{R_{\max}^2}{\sqrt{|D_e|}\epsilon^2}\Big),$$

$$\mathbb{E}|\widehat{Var}(\psi_{h,2}(O_h;\pi,Q)) - Var(\psi_{h,2}(O_h;\mu,r_h))| = O\Big(\frac{R_{\max}^2}{\sqrt{|D_h|}\epsilon^2}\Big),$$

$$\mathbb{E}|\widehat{Cov}(\psi_e(O_e;\pi,r_e),\psi_{h,1}(O_e;\pi,r_e)) - Cov(\psi_e(O_e;\pi,r_e),\psi_{h,1}(O_e;\pi,r_e))| = O\Big(\frac{R_{\max}^2}{\sqrt{|D_e|}\epsilon^2}\Big).$$

Lemma 3. *Under conditions of Lemma 2, notice that $n_{\min} = \min\{|D_e|,|D_h|\}$, we have*

$$\mathbb{E}|\widehat{w} - w^*| \quad \leq \quad O\left(\frac{R_{\max}^2 n_{\min}^{-1/2}\epsilon^{-2} + R_{\max}^2\epsilon^{-2} + |b_h| R_{\max}^2 n_{\min}^{1/2}\epsilon^{-2}}{R_{\max}^2\epsilon^{-2} + nb_h^2}\right).$$

Proof of Theorem 1. Since $\widehat{w}$ minimizes the empirical MSE, it follows that $\widehat{\text{MSE}}(\widehat{\tau}_{\widehat{w}}) \leq \widehat{\text{MSE}}(\widehat{\tau}_{w^*})$, which leads to

$$\widehat{\text{MSE}}(\widehat{\tau}_{\widehat{w}}) - \text{MSE}(\widehat{\tau}_{w^*}) \leq \widehat{\text{MSE}}(\widehat{\tau}_{w^*}) - \text{MSE}(\widehat{\tau}_{w^*}).$$

This allows us to upper bound the excess MSE $\text{MSE}(\widehat{\tau}_{\widehat{w}}) - \text{MSE}(\widehat{\tau}_{w^*})$ by the sum of $\text{MSE}(\widehat{\tau}_{\widehat{w}}) - \widehat{\text{MSE}}(\widehat{\tau}_{\widehat{w}})$ and $\widehat{\text{MSE}}(\widehat{\tau}_{w^*}) - \text{MSE}(\widehat{\tau}_{w^*})$. Let $\widetilde{\text{MSE}}$ denote a version of $\widehat{\text{MSE}}$ by replacing the sampling variance/covariance estimator with its oracle value. It follows from Lemmas 1 and 2 that

$$\begin{aligned}\mathbb{E}[\widetilde{\text{MSE}}(\widehat{\tau}_{\widehat{w}}) - \widehat{\text{MSE}}(\widehat{\tau}_{\widehat{w}})] &\leq \frac{2}{|D_e|}\mathbb{E}[|\widehat{\text{Var}}(\psi_e(O_e)) - \text{Var}(\psi_e(O_e))| + |\widehat{\text{Var}}(\psi_{h,1}(O_e)) - \text{Var}(\psi_{h,1}(O_e))|] \\ &+ \frac{2}{|D_h|}\mathbb{E}[|\widehat{\text{Var}}(\psi_{h,2}(O_e)) - \text{Var}(\psi_{h,2}(O_e))|] + \frac{4}{|D_e|}\mathbb{E}|\widehat{\text{Cov}}(\psi_e,\psi_{h,1}) - \text{Cov}(\psi_e,\psi_{h,1})| = O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big).\end{aligned} \tag{12}$$

Similarly, we can show that

$$\mathbb{E}[\widehat{\text{MSE}}(\widehat{\tau}_{w^*}) - \widetilde{\text{MSE}}(\widehat{\tau}_{w^*})] = O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big). \tag{13}$$

Consequently, it remains to bound $\mathbb{E}[\text{MSE}(\widehat{\tau}_{\widehat{w}}) - \widetilde{\text{MSE}}(\widehat{\tau}_{\widehat{w}})]$ and $\mathbb{E}[\widetilde{\text{MSE}}(\widehat{\tau}_{w^*}) - \mathbb{E}\text{MSE}(\widehat{\tau}_{w^*})]$. By definition, we have

$$\begin{aligned}\mathbb{E}[\text{MSE}(\widehat{\tau}_{\widehat{w}}) - \widetilde{\text{MSE}}(\widehat{\tau}_{\widehat{w}})] &= \mathbb{E}(1-\widehat{w})^2(b_h^2 - \widehat{b}_h^2), \\ \mathbb{E}[\widetilde{\text{MSE}}(\widehat{\tau}_{w^*}) - \text{MSE}(\widehat{\tau}_{w^*})] &= \mathbb{E}(1-w^*)^2(\widehat{b}_h^2 - b_h^2).\end{aligned}$$

The proof is hence completed. □

Proof of Corollary 1. We first prove the first part. According to (6) and the definition of SEE in (7), one can derive that

$$\begin{aligned}\Big|\text{MSE}(\widehat{\tau}_{\widehat{w}}) - \text{MSE}(\widehat{\tau}_{w^*}) - \text{SEE}(\widehat{b}_h)\Big| &\leq \Big|\mathbb{E}[(1-w^*)^2 - (1-\widehat{w})^2][(\widehat{b}_h^2 - b_h^2) - (\widehat{b}_h - b_h)^2] + O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big)\Big| \\ &\leq \Big|\mathbb{E}[(1-w^*)^2 - (1-\widehat{w})^2][2(\widehat{b}_h - b_h)b_h]\Big| + O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big) \\ &\ll O\Big(\frac{R_{\max}^2}{\epsilon^2 n}\Big) + O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big),\end{aligned}$$

where the last inequality follows from that $\mathbb{E}|\widehat{b}_h - b_h|$ can be upper bounded by $O(n_{\min}^{-1/2} R_{\max}/\epsilon)$ and the condition $|b_h| \ll n_{\min}^{-1/2} R_{\max}/\epsilon$. Since $\text{MSE}(\widehat{\tau}_{w^*})$ is proportional to $R_{\max}^2/\epsilon^2 n_{\min}$, it is easy to see that

$$\Big|\frac{\text{MSE}(\widehat{\tau}_{\widehat{w}}) - \text{MSE}(\widehat{\tau}_{w^*}) - \text{SEE}(\widehat{b}_h)}{\text{MSE}(\widehat{\tau}_{w^*})}\Big| \to 0$$

as $n_{\min} \to 0$.

For the second part, as Li et al. (2023) selected the weight to attain the efficiency bound, it is sufficient to prove that the weight proposed in Li et al. (2023), under the constant potential outcome mean assumption and the proportionality assumption, is equivalent to the oracle weight w^*. To see this, the weight for each S_e proposed by Li et al. (2023) has the form

$$\begin{aligned} w_{EB}(S_e) &= \frac{|D_e|\mu(S_e)\pi^*(0|S_e)}{|D_e|\mu(S_e)\pi^*(0|S_e) + |D_h|\text{Var}(R_e|A_e=0,S_e)/\text{Var}(R_h|S_h)} \\ &= \frac{|D_e|}{|D_e| + |D_h|\text{Var}(R_e|A_e=0,S_e)/\text{Var}(R_h|S_h)\mu(S_e)\pi^*(0|S_e)}. \end{aligned}$$

Meanwhile, under the condition that $r_h(\bullet) = r_e(0,\bullet)$, $\widehat{\tau}_w$ can be rewritten as

$$\begin{aligned} \widehat{\tau}_w &= \frac{w}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}\psi_e(O_e) + \frac{1-w}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}\psi_{h,1}(O_e) - \frac{1-w}{|\mathcal{D}_h|}\sum_{O_h\in\mathcal{D}_h}\psi_{h,2}(O_h) \\ &= \frac{1}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}\Big\{r_e(1,S_e) + \nu^1(A_e|S_e)[R_e - r_e(A_e,S_e)] - r_e(0,S_e)\Big\} \\ &\quad -\frac{w}{|\mathcal{D}_e|}\sum_{O_e\in\mathcal{D}_e}\nu^0(A_e|S_e)[R_e - r_e(A_e,S_e)] - \frac{1-w}{|\mathcal{D}_h|}\sum_{O_h\in\mathcal{D}_h}\psi_{h,2}(O_h). \end{aligned}$$

Then the optimal w^* is minimizing the following variance

$$w^2|\mathcal{D}_e|^{-1}\mathbb{E}[\text{Var}(R_e|A_e=0,S_e)/\pi^*(0|S_e)] + (1-w)^2|\mathcal{D}_e|^{-1}\mathbb{E}[\text{Var}(R_h|S_h)\mu(S_e)].$$

It follows directly

$$w^* = \frac{|D_e|}{|D_e| + |D_h|\mathbb{E}[\text{Var}(R_e|A_e=0,S_e)/\pi^*(0|S_e)]/\mathbb{E}[\text{Var}(R_h|S_h)\mu(S_e)]}.$$

It is easy to see that $w_{EB}(S_e) = w^*$ under the proportionality assumption. □

Proof of Corollary 2. For $|b_h| \gg n_{\min}^{-1/2}\sqrt{\log n_{\min}} R_{\max}/\varepsilon$, we can have

$$w^* = \frac{\text{Var}(\widehat{\tau}_h) + b_h^2 - \text{Cov}(\widehat{\tau}_e,\widehat{\tau}_h)}{\text{Var}(\widehat{\tau}_e) + \text{Var}(\widehat{\tau}_h) + b_h^2 - 2\text{Cov}(\widehat{\tau}_e,\widehat{\tau}_h)} \to 1$$

as $n_{\min} \to \infty$. This implies that $\text{MSE}(\widehat{\tau}_{w^*})/\text{MSE}(\widehat{\tau}_e) \to 1$ as $n_{\min} \to \infty$. It suffices to show that $\widehat{w} \xrightarrow{p} 1$ as $n_{\min} \to \infty$. From Lemma 3, we can see that if $|b_h| \gg n_{\min}^{-1/2}\sqrt{\log(n_{\min})} R_{\max}/\epsilon$, then

$$\mathbb{E}|\widehat{w} - w^*| \ll O(R_{\max}\epsilon/\sqrt{\log(n_{\min})}).$$

Therefore, it is easy to deduce that $\widehat{w} \xrightarrow{p} w^*$ as $n_{\min} \to \infty$ by applying the Markov inequality. This completes the proof. □

Proof of Theorem 2. Let $\overline{\text{MSE}}$ denote a version of $\widehat{\text{MSE}}$ by replacing the bias with its oracle value. The difference between the MSE under $\widehat{w}_{\text{U}}$ and that under w^* can be decomposed into

$$\text{MSE}(\widehat{w}_{\text{U}}) - \text{MSE}(w^*) = \text{MSE}(\widehat{w}) - \overline{\text{MSE}}(\widehat{w}) + \overline{\text{MSE}}(\widehat{w}) - \overline{\text{MSE}}(w^*) + \overline{\text{MSE}}(w^*) - \text{MSE}(w^*). \tag{14}$$

Similar to the derivations of (13), we can obtain

$$\mathbb{E}[\text{MSE}(\widehat{w}) - \overline{\text{MSE}}(\widehat{w})] = O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big), \quad \mathbb{E}[\overline{\text{MSE}}(\widehat{w}) - \overline{\text{MSE}}(w^*)] = O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big).$$

Denote $U(w) = (1-w)^2[(|\widehat{b}_h| + U)^2 - b_h^2]$. We have that with probability $1-\alpha$,

$$\overline{\text{MSE}}(\widehat{w}) - \overline{\text{MSE}}(w^*) \le \widehat{\overline{\text{MSE}}}_{\text{UB}}(\widehat{w}) - \overline{\text{MSE}}(w^*) \le \widehat{\overline{\text{MSE}}}_{\text{UB}}(w^*) - \overline{\text{MSE}}(w^*) \le U(w^*), \tag{15}$$

where the first inequality follows from the definition of U, and the second inequality follows from the minimization of $\widehat{\overline{\text{MSE}}}_{\text{UB}}$ at $\widehat{w}$. Hence, with probability $1-\alpha$,

$$\text{MSE}(\widehat{w}_{\text{U}}) - \text{MSE}(w^*) \le \mathbb{E}((1-w^*)^2[(|\widehat{b}_h| + U)^2 - b_h^2]) + O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big).$$

According to (3) and Lemma 2 Meanwhile, the order of $\text{MSE}(\widehat{w}_{\text{U}})$ and $\text{MSE}(w^*)$ can be given by $O([b_h^2 + R_{\max}^2/\epsilon^2 n_{\min}])$. Therefore, by decomposing

$$\text{MSE}(\widehat{w}_{\text{U}}) - \text{MSE}(w^*) = \text{MSE}(\widehat{w}_{\text{U}}) - \text{MSE}(w^*)\mathbb{I}(b_h^2 \le (|\widehat{b}_h| + U)^2) + \text{MSE}(\widehat{w}_{\text{U}}) - \text{MSE}(w^*)\mathbb{I}(b_h^2 > (|\widehat{b}_h| + U)^2),$$

we can directly have

$$\text{MSE}(\widehat{w}_{\text{U}}) - \text{MSE}(w^*) \le \mathbb{E}((1-w^*)^2[(|\widehat{b}_h| + U)^2 - b_h^2]) + O\Big(\frac{R_{\max}^2}{\epsilon^2 n_{\min}^{3/2}}\Big) + O(\alpha[b_h^2 + R_{\max}^2/\epsilon^2 n_{\min}]).$$

This completes the proof. □

Proof of Corollary 3. According to the definition of w^*, we can have $1 - w^* = O(n^{-1}R_{\max}^2\epsilon^{-2}/(b_h^2 + n^{-1}R_{\max}^2\epsilon^{-2})$. Notice that $\text{MSE}(\widehat{\tau}_{w^*}) = O(b_h^2 + n^{-1}R_{\max}^2\epsilon^{-2})$ and with probability $1-\alpha$, $\mathbb{E}[(|\widehat{b}_h| + U)^2 - b_h^2] = O(U^2)$ according to (5), if $b_h \gg n_{\min}^{-1/2}\log^{1/3} n_{\min} R_{\max}/\epsilon$, then $1 - w^* = O(n^{-1}b_h^{-2})$, $\text{MSE}(\widehat{\tau}_{w^*}) = O(b_h^2)$ and

$$\frac{(1-w^*)^2\mathbb{E}[(|\widehat{b}_h| + U)^2 - b_h^2]}{\text{MSE}(\widehat{\tau}_{w^*})} = O\Big(\frac{U^2}{n^2 b_h^6}\Big) \to 0$$

as $n_{\min} \to \infty$ under the condition U is proportional to the order $n_{\min}^{-1/2}\sqrt{\log n_{\min}}R_{\max}/\epsilon$ and $b_h \gg n_{\min}^{-1/2}(\log n_{\min})^{1/6}R_{\max}/\epsilon$.

Meanwhile, $\alpha[b_h^2 + R_{\max}^2/\epsilon^2 n_{\min}]/\text{MSE}(\widehat{\tau}_{w^*}) = o(n_{\min}^{-1})$, $\alpha[b_h^2 + R_{\max}^2/\epsilon^2 n_{\min}]/\text{MSE}(\widehat{\tau}_{w^*}) \to 0$ as $n_{\min} \to \infty$. It leads to the following

$$\frac{\text{MSE}(\widehat{\tau}_{\widehat{w}_U}) - \text{MSE}(\widehat{\tau}_{w^*})}{\text{MSE}(\widehat{\tau}_{w^*})} \to 0.$$

This completes the proof. □

F. Auxiliary Lemmas

In this section, we provide the proofs of lemmas used in Section E.

Proof of Lemma 1. The proof of Lemma 1 is straightforward. The key is to observe that both $\widehat{\tau}_e$ and $\widehat{\tau}_h$ are unbiased under Assumption 3, due to the doubly robust property. This allows us to decompose the MSE into the sum of the squared bias term b_h^2 and the variance term $w^2\text{Var}(\widehat{\tau}_e) + (1-w)^2\text{Var}(\widehat{\tau}_h) + 2w(1-w)\text{Cov}(\widehat{\tau}_e, \widehat{\tau}_h)$. Additionally, the variance terms $\text{Var}(\widehat{\tau}_h)$ and $\text{Cov}(\widehat{\tau}_e, \widehat{\tau}_h)$ can be further simplified due to the independence between the experimental and historical datasets. This leads to the desired conclusion. □

Proof of Lemma 2. We focus on bounding the expected absolute difference between $\text{Var}(\psi_e(O_e;\pi,r_e))$ and its sampling variance estimator $\widehat{\text{Var}}(\psi_e(O_e;\pi,r_e))$ in this section. Other bounds can be similarly established.

Without loss of generality, we assume the sampling variance formula scales with $1/n$ instead of $1/(n-1)$. The error bounds are asymptotically the same when the formula scales with $1/(n-1)$. By definition, we can upper bound $\mathbb{E}|\widehat{\text{Var}}(\psi_e(O_e;\pi,r_e)) - \text{Var}(\psi_e(O_e;\pi,r_e))|$ into the sum of $\mathbb{E}|\widehat{\tau}_e^2 - \tau^2|$ and

$$\mathbb{E}\Big|\frac{2}{|D_e|}\sum_{O_e\in\mathcal{D}_{e,2}} \psi_e^2(O_e;\pi,Q) - \mathbb{E}\psi_e^2(O_e;\pi,Q)\Big|. \tag{16}$$

With some calculations, we have

$$\mathbb{E}|\widehat{\tau}_e^2 - \tau^2| = \mathbb{E}|\widehat{\tau}_e - \tau||\widehat{\tau}_e + \tau| \le \mathbb{E}|\widehat{\tau}_e - \tau|^2 + 2\tau\mathbb{E}|\widehat{\tau}_e - \tau| \le \mathbb{E}|\widehat{\tau}_e - \tau|^2 + 2\tau\sqrt{\mathbb{E}|\widehat{\tau}_e - \tau|^2}. \tag{17}$$

According to Lemma 1, the MSE of $\widehat{\tau}_e$ equals $2\text{Var}(\psi_e(O_e;\pi,r_e))/|D_e|$ which can be upper bounded by $O(R_{\max}^2/(|D_e|\epsilon^2))$ under Assumptions 1 and 2. Additionally, notice that τ can be upper bounded by $R_{\max}$. This allows us to bound $\mathbb{E}|\widehat{\tau}_e^2 - \tau^2|$ by $O(R_{\max}^2/(|D_e|\epsilon^2)) + O(R_{\max}/\sqrt{|D_e|}\epsilon)$.

We next upper bound (16). An application of Cauchy-Schwarz inequality yields that (16) can be upper bounded by $\sqrt{\text{Var}(2\sum_{O_e\in\mathcal{D}_{e,2}}\psi_e^2(O_e;\pi,Q)/|D_e|)} = 2\sqrt{\text{Var}(\psi_e^2(O_e;\pi,Q))/|D_e|} = O(R_{\max}^2/\sqrt{|D_e|}\epsilon^2)$. Since $\epsilon \le 1$, the difference $\mathbb{E}|\widehat{\text{Var}}(\psi_e(O_e;\pi,r_e)) - \text{Var}(\psi_e(O_e;\pi,r_e))|$ can be upper bound by $O(R_{\max}^2/\sqrt{|D_e|}\epsilon^2)$. □

Proof of Lemma 3. Denote $\widehat{F}_1 = \widehat{b}_h^2 + \widehat{\text{Var}}(\widehat{\tau}_h) - \widehat{\text{Cov}}(\widehat{\tau}_e,\widehat{\tau}_h)$, $\widehat{F}_2 = \widehat{\text{Var}}(\widehat{\tau}_e) + \widehat{b}_h^2 + \widehat{\text{Var}}(\widehat{\tau}_h) - 2\widehat{\text{Cov}}(\widehat{\tau}_e,\widehat{\tau}_h)$, and $F_1 = b_h^2 + \text{Var}(\widehat{\tau}_h) - \text{Cov}(\widehat{\tau}_e,\widehat{\tau}_h)$, $F_2 = \text{Var}(\widehat{\tau}_e) + b_h^2 + \text{Var}(\widehat{\tau}_h) - 2\text{Cov}(\widehat{\tau}_e,\widehat{\tau}_h)$. According to the formulations of $\widehat{w}$ and w^*, and $F_1 < F_2$, it is easy to derive that,

$$|\widehat{w} - w^*| = \left|\frac{\widehat{F}_1}{\widehat{F}_2} - \frac{F_1}{F_2}\right| = \left|\frac{(F_1-\widehat{F}_1)F_2 + F_1(F_2-\widehat{F}_2)}{\widehat{F}_2F_2}\right| \le \frac{\left|(\widehat{F}_1 - F_1)\right| + \left|(F_2-\widehat{F}_2)\right|}{F_2}\left|\frac{F_2}{\widehat{F}_2}\right|.$$

It is easy to derive that

$$|\widehat{w} - w^*| \quad\le\quad \frac{|\widehat{b}_h^2 - b_h^2| + 2|\widehat{\text{Cov}}(\widehat{\tau}_e,\widehat{\tau}_h) - \text{Cov}(\widehat{\tau}_e,\widehat{\tau}_h)| + |\widehat{\text{Var}}(\widehat{\tau}_e) - \text{Var}(\widehat{\tau}_e)| + |\widehat{\text{Var}}(\widehat{\tau}_h) - \text{Var}(\widehat{\tau}_h)|}{\text{Var}(\widehat{\tau}_e) + b_h^2 + \text{Var}(\widehat{\tau}_h) - 2\text{Cov}(\widehat{\tau}_e,\widehat{\tau}_h)}.$$

It follows from Lemma 2 and the derivation of (17) that

$$\mathbb{E}|\widehat{b}_h^2 - b_h^2| \le O\Big(\frac{R_{\max}^2}{\epsilon^2|D_e|}\Big) + O\Big(\frac{b_hR_{\max}}{\epsilon\sqrt{|D_e|}}\Big) \quad \mathbb{E}|\widehat{\text{Var}}(\widehat{\tau}_e) - \text{Var}(\widehat{\tau}_e)| \le O\Big(\frac{R_{\max}^2}{\epsilon^2|D_e|^{3/2}}\Big),$$
$$\mathbb{E}|\widehat{\text{Var}}(\widehat{\tau}_h) - \text{Var}(\widehat{\tau}_h)| \le O\Big(\frac{R_{\max}^2}{\epsilon^2|D_h|^{3/2}}\Big), \quad \mathbb{E}|\widehat{\text{Cov}}(\widehat{\tau}_e,\widehat{\tau}_h) - \text{Cov}(\widehat{\tau}_e,\widehat{\tau}_h)| \le O\Big(\frac{R_{\max}^2}{\epsilon^2|D_e|^{3/2}}\Big).$$

Then it is easy to derive that

$$\mathbb{E}|\widehat{w} - w^*| \quad\le\quad O\left(\frac{R_{\max}^2 n_{\min}^{-1/2}\epsilon^{-2} + R_{\max}^2\epsilon^{-2} + b_hR_{\max}n_{\min}^{1/2}\epsilon^{-1}}{R_{\max}^2\epsilon^{-2} + n_{\min}b_h^2}\right)$$

This completes the proof. □