A PRACTICAL GUIDE OF OFF-POLICY EVALUATION FOR BANDIT PROBLEMS

A PREPRINT

Masahiro Kato[1], **Kenshi Abe**[1]*, **Kaito Ariu**[2]*, **Shota Yasui**[1]
[1]CyberAgent Inc.
masahiro_kato@cyberagent.co.jp
abe_kenshi@cyberagent.co.jp
yasui_shota@cyberagent.co.jp
[2] KTH
ariu@kth.se

October 26, 2020

ABSTRACT

Off-policy evaluation (OPE) is the problem of estimating the value of a target policy from samples obtained via different policies. Recently, applying OPE methods for *bandit problems* has garnered attention. For the theoretical guarantees of an estimator of the policy value, the OPE methods require various conditions on the target policy and policy used for generating the samples. However, existing studies did not carefully discuss the practical situation where such conditions hold, and the gap between them remains. This paper aims to show new results for bridging the gap. Based on the properties of the evaluation policy, we categorize OPE situations. Then, among practical applications, we mainly discuss the *best policy selection*. For the situation, we propose a meta-algorithm based on existing OPE estimators. We investigate the proposed concepts using synthetic and open real-world datasets in experiments.

1 Introduction

As an instance of sequential decision-making problems, the *multi-armed bandit* (MAB) problem has attracted significant attention in various applications, such as ad optimization (Narita et al., 2019), personalized medicine (Chow & Chang, 2011; Villar, 2018), and recommendation systems (Li et al., 2010, 2016). However, because we only observe an action chosen following a *behavior policy*, which yields *sample selection bias* in the observed dataset, it is not easy to evaluate a new policy (*evaluation policy*) from the log data. For solving this problem, *off-policy evaluation* (OPE) is proposed (Beygelzimer & Langford, 2009; Li et al., 2010; Dudík et al., 2011; Li et al., 2011; Wang et al., 2017; Narita et al., 2019; Bibaut et al., 2019; Kallus & Uehara, 2019; Oberst & Sontag, 2019).

As reviewed in this paper, OPE estimators require several conditions to achieve theoretical properties, such as *unbiasedness*, *consistency*, and *asymptotic normality*. For instance, we need to put an assumption such that an evaluation policy is independent of log data used for constructing the OPE because the correlation between the evaluation policy and an OPE estimator causes a biased result (Li et al., 2011; Kato et al., 2020b; Kallus & Uehara, 2020). However, although real-world applications often break the condition, existing studies did not pay much attention to which practical situation satisfies the condition.

In addition, we point out terminology problems in OPE. A confusing terminology may cause misuse of OPE methods. For example, Narita et al. (2019) and Saito et al. (2020) study methods for evaluating bandit policies, such as the Thompson sampling, but their experimental process has the potential to cause unexpected bias because the evaluation probability is not deterministic (Narita et al., 2019) and fail to reproduce trajectories of the bandit algorithms (Saito et al., 2020). Although they propose OPE methods for bandit policy, their methods estimate a weighted average

*Equal contributions.

of expected reward over a fixed probability of choosing an action, not a data-dependent bandit policy; that is, the definition of evaluation policy differs between their theoretical results and experiments. Thus, the confusing use of terminologies potentially misleads the applications of OPE.

Five contributions are made: (i) summarizing the potential concern and limitations of OPE methods; (ii) organizing the OPE terminologies; (iii) categorizing situations in which OPE methods are applicable; (iv) showing some new results bridging existing studies and practical situations. This paper is organized as follows. In Section 2 and 3, we formulate our OPE problem setting and review OPE methods with the attention to the potential theoretical concerns. In Section 4, based on cases of evaluation probabilities, we categorize the OPE methods. Built upon the categorization, we introduce a practical application of OPE in Section 5 and show the experimental results in Section 6.

2 Problem Setting

In this section, we describe the problem setting of OPE.

Data-Generating Process

Let A_t be an action in $\mathcal{A} = \{1, 2, \dots, K\}$, X_t be the *covariate* observed by the decision maker when choosing an action, and $\mathcal{X}$ be the space of covariate. Let us denote a random variable of a reward at period t as $Y_t = \sum_{a=1}^{K} \mathbb{1}[A_t = a] Y_t(a)$, where $Y_t(a) : \mathcal{A} \to \mathbb{R}$ is a potential outcome[2]. This setting is also called *bandit feedback*. Suppose that we have access to a dataset $\mathcal{S}_T = \left\{(X_t, A_t, Y_t)\right\}_{t=1}^{T}$ with the following data-generating process (DGP):

$$\left\{(X_t, A_t, Y_t)\right\}_{t=1}^{T} \sim p(x) p_t(a \mid x) p(y \mid a, x), \tag{1}$$

where $p(x)$ denotes the density of the covariate X_t, $p_t(a \mid x)$ denotes the probability of choosing an action a conditioned on a covariate x at period t, and $p(y \mid a, x)$ denotes the density of a reward Y_t conditioned on an action a and covariate x. We assume that $p(x)$ and $p(y \mid a, x)$ are invariant across periods, but $p_t(a \mid x)$ can take different values across periods.

Let us call a function determines the probability $p_t(a \mid x)$ a *behavior probability*. For example, Narita et al. (2019) uses a behavior probability defined as $\pi_t^{\mathrm{b}} : \mathcal{X} \to \Delta^K$, where $\Delta^K := \{(p_a) \in [0,1]^K \mid \sum_{a=1}^{K} p_a = 1\}$. We also define a behavior policy as a system that determines the behavior probability. In MAB problems, the behavior policy can be considered as a function of the trajectory. In most existing studies, a behavior policy and probability are not distinguished, and both are called a behavior policy.

Value of Evaluation Policy

We consider estimating the value of an *evaluation policy* using samples obtained under the behavior policy. Let a policy generating a probability of choosing an action $\pi^{\mathrm{e}} : \mathcal{X} \to \Delta^K$ be an evaluation policy. Similar to the behavior probability, we also call π^{e} as the evaluation probability. In most existing studies, a behavior/evaluation policy and probability are not clearly distinguished, and both are called a behavior/evaluation policy. However, distinguishing them is critical, as mentioned; if an evaluation policy depends on a trajectory generated from the evaluation policy itself, we need to reproduce the trajectory for evaluating the evaluation policy. For example, MAB algorithms carefully select an arm in each period and balance the trade-off between exploration and exploitation using self-generated trajectories to update the parameter. However, in OPE, we have a trajectory generated from a behavior policy, which is usually different from that of the evaluation policy and cannot know the true real-time behavior of the evaluation policy. We discuss such terminology-related problems in Appendix B.

The goal of OPE for an evaluation policy generating an evaluation probability $\pi^{\mathrm{e}}(a \mid x)$ is to estimate the expected reward from the evaluation probability $\pi^{\mathrm{e}}(a \mid x)$ defined as $R(\pi^{\mathrm{e}}) := \mathbb{E}\left[\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid x) Y_t(a)\right]$.

To identify the policy value $R(\pi^{\mathrm{e}})$, we assume overlaps of the distributions of policies and the boundedness of reward.

Assumption 1. There exists a constant C_1 such that $0 \leq \frac{\pi^{\mathrm{e}}(a|x)}{p_t(a|x)} \leq C_1$.

Assumption 2. There exists a constant C_2 such that $|Y_t| \leq C_2$.

[2]We can express the reward without using the potential reward variable. See Appendix B of Kato et al. (2020b)

In addition to the above assumptions, we also assume that an evaluation probability π^{e} is deterministic (Kallus & Uehara, 2019; Kato et al., 2020b). For each situation, a deterministic evaluation probability has different meanings, such as a random variable independent from $\mathcal{S}_T$, random variable constructed converging a time-invariant function as $T \to \infty$, or not a random variable. Note that the deterministic evaluation probability does not mean that the evaluation probability π^{e} chooses a certain action with probability 1. This assumption mainly implies that we cannot evaluate an evaluation probability constructed from samples correlated with the dataset $\mathcal{S}$. In Section 4, we consider practical situations with deterministic evaluation policies.

Besides, let us also define the Average Treatment Effect (ATE) between actions a and a'. Although breaking the definition of the evaluation probability, we can regard for all $x \in \mathcal{X}$, $\pi^{\mathrm{e},(1)}(a \mid x) = 1$ and $\pi^{\mathrm{e},(2)}(a' \mid x) = 1$ as a deterministic evaluation probability in ATE estimation and apply existing OPE estimators to estimate the value. In this paper, when referring to OPE, we allow estimations of ATE to be contained in the target of OPE estimators.

Notations: Let $\mathcal{S}_{t-1} \in \mathcal{M}_{t-1}$ be the history until $t-1$ period defined as $\mathcal{S}_{t-1} = \{X_{t-1}, A_{t-1}, Y_{t-1}, \ldots, X_1, A_1, Y_1\}$ with the space $\mathcal{M}_{t-1}$. Let us denote $\mathbb{E}[Y_t(a) \mid x]$ and $\mathrm{Var}(Y_t(a) \mid x)$ as $f^*(a,x)$ and $\nu^*(a,x)$, respectively. Let $\mathcal{F}$ be the classes of $f^*(a,x)$ and $\hat{f}_t(a, x \mid \mathcal{S}_{t-1})$ be an estimator of $f^*(a,x)$ constructed from $\mathcal{S}_{t-1}$. Let $\mathcal{N}(\mu, \mathrm{var})$ be the normal distribution with the mean μ and the variance var. We often denote a behavior probability as π_t^{b} or π^{b}, and an evaluation probability as π_t^{e} or π^{e}. Additionally, let $\|\mu(X,A,Y)\|_2$ be $\mathbb{E}[\mu^2(X,A,Y)]^{1/2}$ for the function μ.

3 Preliminaries of OPE

We review three types of well-known estimators of $R(\pi^{\mathrm{e}})$. For simplicity, let us assume that the behavior probability π_t^{b} is known. In the case where the behavior probability π_t^{b} is unknown, we replace the behavior probability with its estimator. The first estimator is an inverse probability weighting (IPW) type estimator (Rubin, 1987; Hirano et al., 2003; Swaminathan & Joachims, 2015) given by $\frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X_t)\Gamma_t^{\mathrm{IPW}}(a;\pi_t^{\mathrm{b}})$, where

$$\Gamma_t^{\mathrm{IPW}}(a;\pi) = \Gamma^{\mathrm{IPW}}(a; X_t, A_t, Y_t, \pi) = \frac{\mathbb{1}[A_t = a]Y_t}{\pi(A_t \mid X_t)}.$$

Even though this estimator is unbiased when the behavior probability is known, it suffers from high variance. The second estimator is a direct method (DM) type estimator $\frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X_t)\Gamma_t^{\mathrm{DM}}(a;\hat{f})$, where

$$\Gamma_t^{\mathrm{DM}}(a;f) = \Gamma^{\mathrm{DM}}(a; X_t, f) = f(a, X_t),$$

and $\hat{f}(a, X_t)$ is an estimator of $f^*(a, X_t)$ (Hahn, 1998). This estimator is known to be weak against model misspecification of $f^*(a, X_t)$. The third estimator is an augmented inverse probability (AIPW) type estimator $\frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X_t)\Gamma_t^{\mathrm{AIPW}}(a;\pi_t^{\mathrm{b}},\hat{f})$, where

$$\begin{aligned}\Gamma_t^{\mathrm{AIPW}}(a;\pi,f) &= \Gamma^{\mathrm{AIPW}}(a; A_t, X_t, Y_t, \pi, f)\\ &= \frac{\mathbb{1}[A_t = a](Y_t - f(a, X_t))}{\pi(A_t \mid X_t)} + f(a, X_t).\end{aligned}$$

When replacing the behavior probability with its estimator, AIPW estimator is called doubly robust estimator (Robins et al., 1994; Chernozhukov et al., 2018).

When theoretically evaluating OPE estimators, we often check the unbiasedness, consistency, asymptotic normality, and concentration inequality. For the above estimators, unbiasedness and/or consistency are satisfied in general. On the other hand, there are various methods for guaranteeing the asymptotic normality of the above estimator for specific choices of π^{b} and $\hat{f}$. For brevity, we explain the asymptotic distribution of the AIPW and Adaptive AIPW (A2IPW) (Kato et al., 2020a) estimators for the cases in which the behavior probability is fixed and updated through periods, respectively. First, we consider the case where $\forall t,\ \pi_t^{\mathrm{b}}(a \mid x) = \pi^{\mathrm{b}}(a \mid x)$ in the DGP (1). In this case, by using some methods for deriving the asymptotic normality, such as double/debiased machine learning (DDM) proposed by Chernozhukov et al. (2018), we can derive the asymptotic distribution of AIPW type estimator. Let us define an AIPW estimator with an estimator $\hat{f}_T$ of f^* as $\widehat{R}_T^{\mathrm{AIPW}}(\pi^{\mathrm{e}}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a|X_t)\Gamma_t^{\mathrm{AIPW}}(a;\pi_t^{\mathrm{b}},\hat{f}_T)$. Then, under some conditions, we can show that $\sqrt{T}\left(\widehat{R}_T^{\mathrm{AIPW}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2_{\mathrm{AIPW}}(\pi^{\mathrm{b}}, \pi^{\mathrm{e}}))$, where $\sigma^2_{\mathrm{AIPW}}(\pi,\pi^{\mathrm{e}}) = \mathbb{E}\Big[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}\left(a,X_t\right)\right)^2\nu^*\left(a,X_t\right)}{\pi^{\mathrm{b}}(k|X_t)} + \Big(\sum_{k=1}^{K}\pi^{\mathrm{e}}(a,X_t)f^*(a,X_t) - R(\pi^{\mathrm{e}})\Big)^2\Big]$. Next, we consider the case where the behavior probability is time-dependent. When samples have dependency under a behavior probability $\pi_t^{\mathrm{b}}(a \mid x, \mathcal{S}_{t-1})$, one of useful OPE estimators is A2IPW estimator defined as $\widehat{R}_T^{\mathrm{A2IPW}}(\pi^{\mathrm{e}}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid$

$X_t)\Gamma_t^{\mathrm{AIPW}}(a;\pi_t^{\mathrm{b}},\hat{f}_{t-1})$, where $\hat{f}_{t-1}$ is a step-wise consistent estimator of f^* estimated only using samples $\mathcal{S}_{t-1}$. Then, for a converging behavior probability π_t^{b} such that $\pi_t^{\mathrm{b}}(a \mid x, \mathcal{S}_{t-1}) \xrightarrow{\mathrm{p}} \alpha(a \mid x)$ as $t \to \infty$, Kato et al. (2020a) showed that, under some regularity conditions, $\sqrt{T}\left(\widehat{R}_T^{\mathrm{A2IPW}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathrm{A2IPW}}(\pi, \pi^{\mathrm{e}})\right)$, where $\sigma^2_{\mathrm{A2IPW}}(\pi, \pi^{\mathrm{e}}) = \mathbb{E}\Big[\sum_{a=1}^K \frac{\left(\pi^{\mathrm{e}}\left(a, X_t\right)\right)^2 \nu^*\left(a, X_t\right)}{\alpha(k|X_t)} + \Big(\sum_{k=1}^K \pi^{\mathrm{e}}(a, X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\Big]$. The detailed explanations are in Appendix C. As explained above, under certain conditions, it is known that the AIPW type estimators achieve the efficiency bound (a.k.a semiparametric lower bound), which is the lower bound of the asymptotic MSE of OPE, among regular $\sqrt{T}$-consistent estimators (van der Vaart, 1998, Theorem 25.20). Let us note that for the best regret order bandit algorithms, we cannot use them as the behavior policy of OPE because the asymptotic variance including $1/\pi_t^{\mathrm{b}}$ diverges. We discuss in Appendix I.

4 Categorization based on Evaluation Policy

In this section, from the deterministic evaluation probability perspective, we categorize the situation where OPE is appropriate as follows: (a) OPE with a given fixed probability; (b) inter-temporal OPE (ITOPE); (c) sample splitting OPE (SSOPE); (d) on-policy off-policy evaluation (OPOPE). While an evaluation probability is given exogenously in case (a), it is constructed from samples in case (b)–(d). We illustrate the concepts of case (b)–(d) in Figure 1.

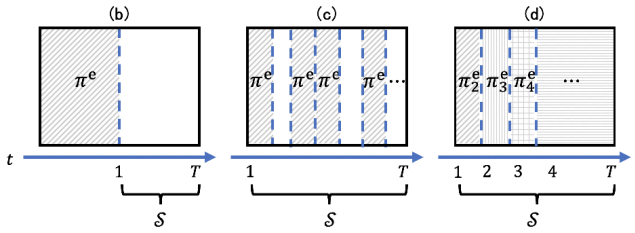

Figure 1: Illustrations of cases (b)–(d).

When a fixed policy is given as the case (a), we can naively apply existing OPE methods. However, such a situation is not common in practice because we often construct an evaluation probability from samples. In such cases, we practically consider two cases, (b) ITOPE and (c) SSOPE, where we separate training and evaluation datasets for constructing and evaluating an evaluation probability, respectively. In ITOPE, we have a dataset with time series $\{-T', \cdots -1, 0, 1, \ldots, T\}$, where $T' \in \mathbb{N}$, and allow the behavior policy to be updated at period $t = 1$ using samples obtained until that period. A defect of these approaches, we cannot use the whole dataset for constructing an evaluation probability. On the other hand, in the case (d) OPOPE, we can conduct OPE for a sequentially constructed evaluation probability owing to martingale property. In this case, although the whole dataset for constructing an evaluation probability is used for OPE, we only identify the average value of the sequence of policy value. We introduce the details as follows.

Remark 1 (Categorization based on Behavior Policy)**.** We also introduce the categorization based on behavior policy from perspectives of time-dependency scenario and the number of behavior policies. The details are shown in Appendix D. We show a new theoretical result for the case of multiple behavior policies by using the generalized method of moments (GMM), which enables us to construct an efficient estimator by weighting OPE estimators from the multiple behavior policies. We also show the experimental result using the real-world dataset released by Saito et al. (2020).

(a) OPE with a given probability: In this case, the probability π^{e} is data-independent and given exogenously. We include ATE estimation in this situation.

(b) ITOPE: In this case, we allow behavior and evaluation policies to be time-dependent. In ITOPE, for a time series $t = -T', \ldots, -1, 0, 1, \ldots, T$, we construct an evaluation probability using past samples obtained until $t = 0$. Then, we estimate the policy value using samples from $t = 1, 2, \ldots, T$, $\mathcal{S}$. For representing the dependency, let us denote an evaluation probability as $\pi^{\mathrm{e}}_{0:-T'}(a \mid x, \mathcal{S}_{0:-T'})$ and behavior probability as $\pi^{\mathrm{b}}_{0:-T'}(a \mid x, \mathcal{S}_{0:-T'})$, where $\mathcal{S}_{0:-T'} = \{X_0, A_0, Y_0, X_{-1}, A_{-1}, Y_{-1}, \ldots, X_{-T'}, A_{-T'}, Y_{-T'}\}$. In this case, we can estimate the policy value defined as $R'_{0:-T'}(\pi^{\mathrm{e}}) := \mathbb{E}\left[\sum_{a=1}^K \pi^{\mathrm{e}}(a \mid x, \mathcal{S}_{0:-T'}) Y_t(a) \mid \mathcal{S}_{0:-T'}\right]$. This situation is similar to *progressive*

validation (Blum et al., 1999). In this case, an AIPW type estimator $\widehat{R}^{\mathrm{AIPW}'}(\pi^{\mathrm{e}}_{0:-T'}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{0:-T'})\Gamma^{\mathrm{AIPW}}_t(a; \pi^{\mathrm{b}}_{0:-T'}, \hat{f}_{0:-T'})$ has the asymptotic normality as shown in the following theorem, where $\hat{f}_{0:-T'}$ is an estimator of f^* constructed from $\mathcal{S}_{0:-T'}$.

Theorem 1. *Suppose that*

(i) *Pointwise convergence in probability of $\hat{f}_{0:-T'}$, i.e., for all $x \in \mathcal{X}$ and $a \in \{1, 2, \ldots, K\}$, $\hat{f}_{0:-T'}(a, x) - f^*(a, x) \xrightarrow{\mathrm{p}} 0$ as $T' \to \infty$;*

(ii) *There exists a constant C_3 such that $|\hat{f}_{0:-T'}| \leq C_3$.*

Then, under Assumption 1 and 2,

$$\sqrt{T}\left(\widehat{R}^{\mathrm{AIPW}'}\left(\pi^{\mathrm{e}}_{0:-T'}\right) - R'(\pi^{\mathrm{e}}_{0:-T'})\right) \to \mathcal{N}\left(0, \sigma^{2\dagger}_{\mathrm{AIPW}}\right),$$

where $\sigma^{2\dagger}_{\mathrm{AIPW}}$ is the variance of an AIPW estimator with an evaluation probability $\pi^{\mathrm{e}}_{0:-T}$.

The proof is in Appendix E. The pointwise convergence assumption of $\hat{f}_{0:-T'}$ is for technical ease of proof.

(c) SSOPE: In SSOPE, we randomly separate the dataset $\mathcal{S}$ into multiple subsets. We consider estimating an evaluation policy using a subset and estimate the value using the other subsets. If samples are i.i.d., the estimated evaluation policy is also independent of the evaluation datasets.

(d) OPOPE: In OPOPE, we construct an evaluation probability using past samples obtained until period $t-1$ to evaluate the value at each period t. In this case, we can estimate the average policy values. Let us denote the evaluation policies as $\pi^{\mathrm{e}}_t(a \mid x, \mathcal{S}_{t-1})$, and assume that $\pi^{\mathrm{e}}_t \xrightarrow{\mathrm{p}} \pi^{\mathrm{e}}$, where π^{e} is a time invariant probability. Let us define an Average Policy Value AIPW (AP) estimator as $\widehat{R}^{\mathrm{AP}}\left(\{\pi^{\mathrm{e}}_t\}_{t=1}^{T}\right) = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}_t(a \mid X_t, \mathcal{S}_{t-1})\Gamma^{\mathrm{AIPW}}_t(a; \pi^{\mathrm{b}}, \hat{f}_{t-1})$, where note that $\hat{f}_t$ is an estimator of f^* constructed from $\mathcal{S}_{t-1}$. Then, by using the AIPW estimator, we can show the following theorem.

Theorem 2. *Suppose that*

(i) *Pointwise convergence in probability of $\hat{f}_{t-1}$ and π^{e}_t, i.e., for all $x \in \mathcal{X}$ and $a \in \{1, 2, \ldots, K\}$, $\hat{f}_{t-1}(a, x) - f^*(a, x) \xrightarrow{\mathrm{p}} 0$ and $\pi^{\mathrm{e}}_t(a \mid x, \mathcal{S}_{t-1}) - \pi^{\mathrm{e}}(a \mid x) \xrightarrow{\mathrm{p}} 0$, where $\pi^{\mathrm{e}} : \mathcal{A} \times \mathcal{X} \to (0, 1)$;*

(ii) *There exists a constant C_3 such that $|\hat{f}_{t-1}| \leq C_3$.*

Then, under Assumption 1 and 2,

$$\sqrt{T}\left(\widehat{R}^{\mathrm{AP}}\left(\{\pi^{\mathrm{e}}_t\}_{t=1}^{T}\right) - \frac{1}{T}\sum_{t=1}^{T}R(\pi^{\mathrm{e}}_t)\right) \to \mathcal{N}\left(0, \sigma^2_{\mathrm{AP}}\right),$$

where σ^2_{AP} is the variance of an AIPW estimator with an evaluation probability π^{e}.

The proof is in Appendix G. In practice, we face such a situation when estimating an evaluation probability from small samples with gathering more samples via bandit process. However, let us note that we cannot estimate the true bandit policy itself because we cannot reproduce the trajectory generated by the target bandit policy Appendix I.

Remark 2 (Estimation of $R(\pi^{\mathrm{e}})$)**.** Can we estimate $R(\pi^{\mathrm{e}})$? To this aim, we need to show $\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}R(\pi^{\mathrm{e}}_t) - R(\pi^{\mathrm{e}}) = \mathrm{o}(1)$. This condition requires π^{e}_t converges to π^{e} with the $\mathrm{o}(1/\sqrt{T})$ order. However, standard regulation estimators achieve at least $\mathrm{O}(1/\sqrt{T})$ order.

Remark 3 (Inadequateness of off-policy bandit evaluation)**.** In general, it is inadequate to evaluate bandit policies through OPE. The objective of typical bandit policies is to minimize cumulative regret. The value of the regret reflects the transient cumulative cost incurred to reach the optimal stationary policy. To evaluate such a policy, we need to evaluate the policy that also explicitly depends on the history $\mathcal{S}'_{t-1}$, which is different from the history of the log data $\mathcal{S}_{t-1}$ and improbable to reproduce. Gilotte et al. (2018) also pointed out this inadequateness in their experiments, but did not pursue this topic further.

5 Best Evaluation Policy Selection

In this section, we consider a case study of the best evaluation policy selection (BEPS), which is one of the most important applications of OPE. For ease of discussion, we only consider an unbiased, consistent, and asymptotically normal estimator under some conditions with deterministic evaluation probability. For example, AIPW and IPW estimators satisfy the requirements. Let us assume that there is a set of L-candidate evaluation policies, $\Lambda = \left\{\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(L)}\right\}$, which construct evaluation probabilities $\Pi^{\mathrm{e}} = \{\pi_{\mathcal{S}}^{\mathrm{e},\lambda^{(1)}}, \pi_{\mathcal{S}}^{\mathrm{e},\lambda^{(2)}}, \ldots, \pi_{\mathcal{S}}^{\mathrm{e},\lambda^{(L)}}\}$ from a dataset $\mathcal{S}$. In this section, for a given dataset $\mathcal{S}_T$, our goal is to select the best evaluation policy λ^* with the highest expected reward, i.e., $\lambda^* = \arg\max_{\lambda\in\Lambda} R(\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda})$. For achieving this goal, we propose three approaches: (I) estimating the policy value using samples $\mathcal{S}_T$, which also constructs the evaluation probabilities; (II) estimating the policy value using an evaluation dataset independent of the samples constructing the evaluation probability; (III) evaluation the policy value using cross-validation. We refer to these three approaches as In-Sample OPE (ISOPE), OPE with Evaluation Dataset (OPE2D), and Off-Policy Cross-Validation (OPCV). Among them, only OPE2D enables us to estimate the best λ^* with keeping unbiasedness, consistency, and asymptotic normality. In ISOPE, we cannot guarantee the unbiasedness and asymptotic normality. In OPCV, instead of $R(\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda})$, we choose the best evaluation policy based on an unbiased, consistent, and asymptotically normal estimator of $R(\pi_{\mathcal{S}'_T}^{\mathrm{e},\lambda})$, where $\mathcal{S}'_T$ is a subset of $\mathcal{S}_T$. Finally, we discuss the criteria of the BEPS when multiple OPE estimators are given.

Remark 4 (Off-Policy Learning (OPL))**.** As a closely related method of BEPS, we introduce OPL, which learns the optimal policy that maximizes the expected reward. Let us define the optimal policy π^* as $\pi^* = \arg\max_{\pi\in\Pi} R(\pi)$, where Π is a policy class. By applying each of OPE estimators $\widehat{R}(\pi)$, we define an estimator for the optimal policy as $\hat{\pi} = \arg\max_{\pi\in\mathcal{H}} \widehat{R}(\pi)$, where $\mathcal{H}$ is a hypothesis set. In this case, if $\mathcal{H}$ includes π^*, we can regard the convergence point of the probability $\hat{\pi}$ as a deterministic one, π^*.

(I) ISOPE

Let us consider constructing an OPE estimator from $\mathcal{S}_T$ for estimating a policy value of Π^{e}. In this case, as pointed out in the existing studies, in general, we cannot construct an unbiased and asymptotic normal estimator, but can construct a consistent estimator if $\|\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda} - \pi^{\mathrm{e}*}\|_2 = \mathrm{o}_p(1)$ for a data-independent evaluation probability $\pi^{\mathrm{e}*}$. Let us consider an AIPW estimator $\widehat{R}_T^{\mathrm{AIPW}}(\pi^{\mathrm{e}}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a|X_t)\Gamma_t^{\mathrm{AIPW}}(a;\pi_t^{\mathrm{b}},\hat{f}_T)$. Then, $\widehat{R}_T^{\mathrm{AIPW}}(\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda}) - R(\pi^{\mathrm{e}*})$ is equal to $\frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\left(\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda}(a|X_t) - \pi^{\mathrm{e}*}(a|X_t)\right)\Gamma_t^{\mathrm{AIPW}}(a;\pi_t^{\mathrm{b}},\hat{f}_T) + \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}*}(a|X_t)\Gamma_t^{\mathrm{AIPW}}(a;\pi_t^{\mathrm{b}},\hat{f}_T) - R(\pi^{\mathrm{e}*})$. If $\|\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda} - \pi^{\mathrm{e}*}\|_2 = \mathrm{o}_p(1)$, we can easily show $\widehat{R}_T^{\mathrm{AIPW}}(\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda}) \xrightarrow{\mathrm{p}} R(\pi^{\mathrm{e}*})$. However, for guaranteeing the asymptotic normality, it is known that we need to impose stronger restrictions for $\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda}$ (Kallus & Uehara, 2020). Although this estimator has an asymptotic bias, which does not disappear with $\sqrt{T}$-order (Chernozhukov et al., 2018), this estimator is still useful owing to the consistency. ISOPE does not belong to the previous categorization because the evaluation probability is not deterministic. However, as a special case of ISOPE, if we use OPOPE of case (d) in Section 4, we can guarantee the unbiasedness to the average of the policy value, but it requires a complicated procedure.

(II) OPE2D

For estimating the policy values of Π^{e}, let us consider constructing an OPE estimator from $\mathcal{S}'$, which is independent of $\mathcal{S}_T$. In this case, we can estimate the value of $R(\pi_{\mathcal{S}_T}^{\mathrm{e},\lambda^{(m)}})$ without loss of the asymptotic normality. This case requires us exogenous dataset $\mathcal{S}'$ but is most desirable from the theoretical perspective. OPE2D is closely related with the case (b) and (c) in the categorizations of Section 4.

Remark 5. In this case, we can also apply hypothesis testing for selecting an evaluation policy. We introduce hypothesis testing framework in Appendix H.

(III) OPCV

Cross-validation (CV) is a standard method for measuring the performance of methods. For simplicity, the dataset $\mathcal{S}_T$ is randomly separated into two datasets $\mathcal{S}_{T^{(1)}}^{(1)}$ and $\mathcal{S}_{T^{(2)}}^{(2)}$, where note that sample sizes are $T^{(1)}$ and $T^{(2)}$, respectively. Then, using the two datasets, we construct evaluation probabilities using $\mathcal{S}_{T^{(1)}}^{(1)}$ and estimate the policy value by using $\mathcal{S}_{T^{(2)}}^{(2)}$. Next, we construct evaluation probabilities using $\mathcal{S}_{T^{(2)}}^{(2)}$ and estimate the policy value by using $\mathcal{S}_{T^{(1)}}^{(1)}$. Finally,

Table 1: Experimental results of OPE2D with various OPE estimators using mnist and pendigits datasets. The best and worst methods are highlighted in red and blue, respectively.

mnist	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00239	0.00683	0.00665	0.03627	0.00020	0.00080	0.00248	0.00733	0.00093	0.00542	0.00649	0.03575	0.00677	0.03567	0.00181	0.00751
0.4	0.00006	0.00043	0.00430	0.01283	0.00005	0.00040	0.00059	0.00314	0.00005	0.00040	0.00116	0.00608	0.00146	0.00641	0.00306	0.01100
0.0	0.00000	0.00000	0.00690	0.04216	0.00000	0.00000	0.00000	0.00000	0.00437	0.03457	0.00545	0.04072	0.00540	0.04070	0.00108	0.01076

pendigits	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00170	0.00375	0.00448	0.00603	0.00080	0.00208	0.00059	0.00164	0.00113	0.00296	0.00230	0.00453	0.00244	0.00405	0.00145	0.00329
0.4	0.00075	0.00212	0.00535	0.00745	0.00082	0.00247	0.00073	0.00203	0.00093	0.00251	0.00460	0.00696	0.00530	0.00678	0.00088	0.00253
0.0	0.00061	0.00325	0.00073	0.00368	0.00024	0.00232	0.00024	0.00232	0.00057	0.00337	0.00073	0.00368	0.00084	0.00384	0.00054	0.00318

we calculate an average of the policy value estimators and choose an evaluation probability based on the averaged OPE estimator. In this case, we can estimate the policy values of $R(\pi^{\mathrm{e},\lambda^{(m)}}_{\mathcal{S}^{(1)}_{T^{(1)}}})$ and $R(\pi^{\mathrm{e},\lambda^{(m)}}_{\mathcal{S}^{(2)}_{T^{(2)}}})$ with asymptotic normality for $m \in \{1, 2, \ldots, M\}$ although we do estimate the policy value of $R(\pi^{\mathrm{e},\lambda^{(m)}}_{\mathcal{S}_T})$. OPCV belongs to the case (c) in the categorizations of Section 4.

Remark 6 (Evaluation without an exogenous dataset)**.** Note that, in both ISOPE and OPCV, we cannot estimate the policy value $\pi^{\mathrm{e},\lambda}_{\mathcal{S}_T}$ for an algorithm λ with keeping desirable theoretical properties. Although it is difficult to prepare an exogenous dataset as OPE2D, ISOPE and OPCV are realistic alternative approaches in practice.

Policy Selection Criteria for Multiple OPE Estimators

Let us assume that there is a set of E OPE estimators, $\mathcal{E} = \left\{\widehat{R}^{(1)}, \ldots, \widehat{R}^{(E)}\right\}$. Unlike the standard CV, in OPE, we cannot directly obtain the target metrics of the algorithm, $\frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K} Y_i(a, X_t)$. Thus, we need to use various OPE estimators approximating $\mathbb{E}\big[\sum_{a=1}^{K} \pi^{\mathrm{e}}(a, X_t)Y_t(a)\big]$. Then, which estimator should we use in the case with several OPE estimators? We raise the following three criteria of OPE estimators for BEPS.

Low asymptotic variance estimator: First, we consider choosing an OPE estimator with the lowest asymptotic variance. The lower bound of the asymptotic variance is given as semiparametric lower bound (Appendix C), and the asymptotic variances of several estimators, such as an AIPW estimator, are known to achieve this lower bound (Narita et al., 2019). Therefore, from this perspective, a hopeful OPE estimator is an estimator with the lowest asymptotic variance. However, this approach is not useful if several OPE estimators achieve the lower bound.

Weighted estimator: Second, we consider a new OPE estimator $\sum_e^E w_e \widehat{R}^{(E)}$ from $\mathcal{E}$ by weighting them using a weight w_e such that $\sum_e^E w_e = 1$. A candidate for this weight is $w_e = 1/E$; that is, constructing a new estimator as the average of $\mathcal{E}$. Another candidate is using e-th element of $(I_E^\top \Sigma I_E)^{-1} I_E^\top \Sigma$ as the weight $w_e$if they asymptotically follow the multivariate normal distribution with covariance matrix Σ (Hamilton, 1994).

Minimax criterion: The final criterion is a minimax approach. In this criterion, for a dataset $\mathcal{S}$, we select the best evaluation policy as

$$\lambda^* = \arg\min_{\lambda\in\Lambda} \max_{\widehat{R}\in\mathcal{E}} -\widehat{R}\left(\pi^{\mathrm{e},\lambda}_{\mathcal{S}}\right). \tag{2}$$

In addition, we consider the following two-player zero-sum game; the player 1 decides the evaluation policy given as a linear combination of $\pi^{\mathrm{e},\lambda^{(1)}}_{\mathcal{S}}, \ldots, \pi^{\mathrm{e},\lambda^{(L)}}_{\mathcal{S}}$; the player 2 constructs the estimator given as a linear combination of $\hat{R}^{(1)}, \ldots, \hat{R}^{(E)}$; The goal of player1/player2 is to maximize/minimize the estimated expected reward of the evaluation policy. Formally, the minmax strategy of player 1 is

$$p^* = \arg\max_{p\in\Delta^L} \min_{w\in\Delta^E} \sum_{j=1}^{E} -w_j \hat{R}^{(j)}(\pi^{\mathrm{e}}_p), \tag{3}$$

where $\pi^{\mathrm{e}}_p = \sum_{l=1}^{L} p_l \pi^{\mathrm{e},\lambda^{(l)}}_{\mathcal{S}}$ and note that $\Delta^D := \{(p_d) \in [0,1]^K \mid \sum_{d=1}^{D} p_d = 1\}$. Under some mild assumptions, we can compute p^* via linear programming:

Table 2: Experimental results of ISOPE with various OPE estimators using mnist and pendigits datasets. The best and worst methods are highlighted in red and blue, respectively.

mnist	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.01519	0.01420	0.00406	0.00812	0.00294	0.00556	0.00745	0.01130	0.00768	0.01143	0.00572	0.01008	0.00554	0.00924	0.01140	0.01304
0.4	0.02942	0.02778	0.00362	0.00763	0.00384	0.00631	0.01048	0.01306	0.00899	0.01056	0.00344	0.00614	0.00383	0.00627	0.02645	0.02717
0.0	0.08167	0.08505	0.00121	0.01213	0.00000	0.00000	0.02718	0.04801	0.02191	0.04385	0.00121	0.01213	0.00121	0.01213	0.08106	0.08541
pendigits	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00375	0.00751	0.00380	0.00550	0.00317	0.00537	0.00299	0.00507	0.00319	0.00536	0.00413	0.00752	0.00440	0.00731	0.00319	0.00536
0.4	0.00199	0.00568	0.00470	0.00754	0.00129	0.00355	0.00163	0.00451	0.00163	0.00451	0.00439	0.00738	0.00489	0.00730	0.00199	0.00568
0.0	0.00079	0.00496	0.00079	0.00496	0.00010	0.00066	0.00037	0.00272	0.00037	0.00272	0.00079	0.00496	0.00079	0.00496	0.00079	0.00496

Table 3: Experimental results of OPCV with various OPE estimators using mnist and pendigits datasets. The best and worst methods are highlighted in red and blue, respectively.

mnist	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.02055	0.02191	0.02318	0.02461	0.02386	0.02410	0.02296	0.02148	0.02162	0.02272	0.02197	0.02147	0.02281	0.02097	0.02204	0.02426
0.4	0.01728	0.02157	0.01701	0.02108	0.01653	0.02091	0.01708	0.02168	0.01599	0.02116	0.01588	0.02099	0.01595	0.02019	0.01726	0.02104
0.0	0.05796	0.07521	0.05330	0.07467	0.05369	0.07449	0.05559	0.07465	0.05531	0.07450	0.05217	0.07364	0.05350	0.07421	0.05736	0.07541
pendigits	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00689	0.01216	0.01872	0.02768	0.00910	0.01761	0.00946	0.01803	0.00903	0.01759	0.01120	0.02113	0.01103	0.02094	0.01672	0.02655
0.4	0.01384	0.02227	0.02459	0.03017	0.01889	0.02723	0.01395	0.02228	0.01773	0.02610	0.02002	0.02736	0.01965	0.02617	0.01903	0.02705
0.0	0.00078	0.00397	0.00722	0.04067	0.00088	0.00406	0.00078	0.00397	0.00088	0.00406	0.00616	0.03941	0.00681	0.03985	0.00078	0.00397

Theorem 3. *Suppose that each estimator can be written as* $\hat{R}(\pi) = \frac{1}{T}\sum_{t=1}^{T}\langle \pi(X_t), \Gamma_t\rangle$ *for* $\hat{R}(\pi) \in \mathcal{E}$*, where* T *is the sample size and* Γ_t *is a component of* $\hat{R}(\pi)$*. Then,* p^* *is the solution of the following problem:*

$$\max_{z\in\mathbb{R}, p\in\Delta^L} z \quad \text{s.t.} \sum_{l=1}^{L} p_l \hat{R}(\pi_{\mathcal{S}}^{\lambda^{(l)}}) \geq z, \ \forall \hat{R} \in \mathcal{E}.$$

Such a minimax criterion is also adapted in existing machine learning studies (Wang et al., 2015). This criterion is conservative and aims to avoid the worst result.

6 Experiments

In this section, we empirically investigate our arguments.

Stochastic Policy Bias

We investigate what causes by estimating policy value based on a dataset used for constructing evaluation probability. We generate an artificial pair of covariate and reward $(X_t, Y_t(1), Y_t(2), Y_t(3))$. The covariate X_t is a 10 dimensional vector generated from the standard normal distribution. For $a \in \{1,2,3\}$, the potential outcome $Y_t(a)$ is 1 if a is chosen by following a probability defined as $p(a \mid x) = \frac{\exp(g(a,x))}{\sum_{a'}^{3}\exp(g(a',x))}$, where $g(1,x) = \sum_{d=1}^{10} X_{t,d}$, $g(2,x) = \sum_{d=1}^{10} W_d X_{t,d}^2$, and $g(3,x) = \sum_{d=1}^{10} W_d |X_{t,d}|$, where W_d is uniform randomly chosen from $\{-1,1\}$. Let us generate three datasets, $\mathcal{S}^{(1)}_{T_{(1)}}$, $\mathcal{S}^{(2)}_{T_{(2)}}$, and $\mathcal{S}^{(3)}_{T_{(3)}}$, where $\mathcal{S}^{(m)}_{T_{(m)}} = \{(X_t^{(m)}, Y_t^{(m)}(1), Y_t^{(m)}(2), Y_t^{(m)}(3))\}_{t=1}^{T^{(m)}}$. Firstly, we train an evaluation probability $\pi^{\mathrm{e}(1)}$ by solving a prediction problem between $X_t^{(1)}$ and $Y_t^{(1)}(1), Y_t^{(1)}(2), Y_t^{(1)}(3)$ using the dataset $\mathcal{S}^{(1)}_{T_{(1)}}$. Then, we apply the evaluation policy $\pi^{\mathrm{e}(1)}$ on the independent dataset $\mathcal{S}^{(2)}_{T_{(2)}}$, and artificially construct bandit data $\{(X_t^{(m)}, A_t^{(m)}, Y_t^{(m)})\}_{t=1}^{T^{(m)}}$, where A_t is a chosen action from the evaluation policy and $Y_t^{(m)} = \sum_{a=1}^{3} \mathbb{1}[A_t^{(m)} = a] Y_t^{(m)}(a)$. Then, we set the true policy value $R(\pi^{\mathrm{e}(1)})$ as $\frac{1}{T^{(2)}}\sum_{t=1}^{T^{(2)}} Y_t^{(m)}$. Next, using the datasets $\mathcal{S}^{(1)}_{T_{(1)}}$ and $\mathcal{S}^{(3)}_{T_{(3)}}$, we estimate the policy value as $\widehat{R}^{(1)}(\pi^{\mathrm{e}(1)}) = \frac{1}{T^{(1)}}\sum_{t=1}^{T^{(1)}}\sum_{a=1}^{3}\pi^{\mathrm{e}(1)}(a \mid X_t) Y_t^{(1)}(a)$ and $\widehat{R}^{(3)}(\pi^{\mathrm{e}(1)}) = \frac{1}{T^{(3)}}\sum_{t=1}^{T^{(3)}}\sum_{a=1}^{3}\pi^{\mathrm{e}(1)}(a \mid X_t) Y_t^{(3)}(a)$, respectively. Here, the estimator $\widehat{R}^{(1)}(\pi^{\mathrm{e}(1)})$ violates the assumption of deterministic evaluation policy because $\pi^{\mathrm{e}(1)}$ and $\mathcal{S}^{(1)}_{T_{(1)}}$ are correlated. Let us define estimation errors as $\mathrm{error1} = R(\pi^{\mathrm{e}(1)}) - \widehat{R}^{(1)}(\pi^{\mathrm{e}(1)})$ and $\mathrm{error2} = R(\pi^{\mathrm{e}(1)}) - \widehat{R}^{(3)}(\pi^{\mathrm{e}(1)})$, respectively. We conduct two cases of experiments by changing the sample size. The first case is $T^{(1)} = T^{(3)} = 100$, and the second case is $T^{(1)} = T^{(3)} = 1000$.

In both cases, we set $T^{(2)} = 10000$. We plot the distributions of these errors in Figure 2. If there is no bias, the error concentrates around on 0. As expected, we can check that error1 is biased, i.e., $\widehat{R}^{(1)}(\pi^{\mathrm{e}(1)})$ is biased. The results are shown in Figure 2. These results imply that evaluating evaluation probability using samples used for constructing the probability may cause a serious bias even if we observe potential outcomes, and the bias reduces as the sample size increases.

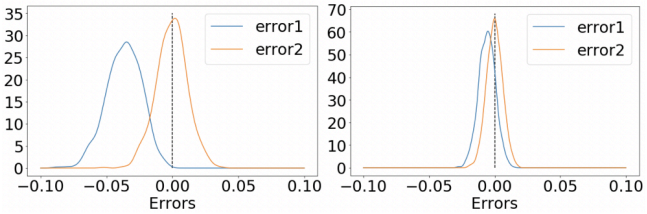

Figure 2: The error distributions. We smoothed the histograms using kernel density estimation. The left graph is the results with sample sizes $T^{(1)} = T^{(3)} = 100$. The right graph is the results with sample sizes $T^{(1)} = T^{(3)} = 1000$.

OPE2D, ISOPE, and OPCV with Various OPE Estimators

Following Dudík et al. (2011) and Farajtabar et al. (2018), we evaluate the proposed estimators using classification datasets by transforming them into contextual bandit data. From the LIBSVM repository, we use the pendigits, mnist, sensorless, and connect-4 datasets [3]. Then, we make a deterministic policy π_d by training a logistic regression classifier on the historical data. We construct three different behavior policies as mixtures of π^d and the uniform random policy π^u by changing a mixture parameter α, i.e., $\pi^{\mathrm{b}} = \alpha\pi^d + (1-\alpha)\pi^u$. The candidates of the mixture parameter α are $\{0.7, 0.4, 0.0\}$ as Kallus & Uehara (2019). We create a set of 6 evaluation policies, $\mathcal{E}$, as outputs of linear logistic regression, support vector machine (SVM) with linear kernel, SVM with the polynomial kernel, SVM with RBF kernel, and random forest. In each experiment, from a dataset, we generated sample 1000 samples for training a behavior policy, 1000 samples for creating evaluation policies, 1000 samples for OPE, and 2000 samples for calculating the true policy value. The more details of the experiment are shown in Appendix J.

For each dataset, we prepare a set of the following policy value estimators: IPW, DM, and AIPW. For the DM estimator, we estimate f^* by linear and kernel ridge regression and denote them DM LR and DM KR, respectively. For the AIPW estimator, we estimate f^* by kernel ridge regression. For selecting an OPE estimator from the set of OPE estimators, we introduce some criteria shown in Section 5. First, we create a new OPE estimator by weighting the OPE estimators with equal weight (MEAN). We also applied minimax criteria of (2) and (3) and denote them Minimax and Mix, respectively. Finally, for an evaluation probability class Π, we select an OPE estimator as $\hat{\pi^{\mathrm{e}}} = \arg\min_{\pi^{\mathrm{e}}\in\Pi} \min_{\widehat{R}\in\mathcal{E}} -\widehat{R}(\pi^{\mathrm{e}})$ and call it Maxmax. As a performance metric, we use the regret defined as $R(\pi^{\mathrm{e}*}) - R(\hat{\pi^{\mathrm{e}}})$, where $\pi^{\mathrm{e}'} = \arg\max_{\pi^{\mathrm{e}}\in\mathcal{E}} R(\pi^{\mathrm{e}})$ and $\hat{\pi^{\mathrm{e}}}$ is a chosen evaluation probability by a criterion. We conduct 100 trials and calculate the means (Mean) and standard deviations (SD) of the regrets. The results are shown in Table 1.

We also show experimental results of ISOPE and OPCV with multiple OPE estimators. We show them in Tables 2 and 3, respectively. For ISOPE experiment, unlike OPE2D experiment, we use the same dataset for both constructing and evaluating the evaluation probability. For OPCV experiment, we generated 1000 samples for training a behavior policy, 2000 samples for 2 fold CV, and 2000 samples for calculating the true policy value. After selecting the best method via CV, we reconstruct an evaluation probability using the whole 2000 samples and measure its regret. The other settings are identical to the OPE2D experiments.

Note that, only in OPE2D experiments, we can directly evaluate the target policy value. Although the minimax criteria fail to select the best policy in OPE2D experiments, they work well in OPCV experiments. As the IPW estimator achieves both the best and worst performance, it is not easy to choose one OPE estimator. Among the proposed criteria, MEAN works stably in these experiments.

7 Conclusion

This study presented a guide for OPE. In OPE, there are various technical problems, but some existing studies tend to ignore them. In this paper, we organized the terminologies of OPE for avoiding the confusion in OPE applications

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

and proposed several methods with new theoretical results. Among the applications, we focus on the BEPS problem and proposed several criteria for the practice of OPE. Finally, we showed empirical investigations of problems related to OPE.

Ethics Statement

Although OPE methods are expected to solve various applications, they may also return biased results by the abuse of them. This paper casts concern on this problem and shows a guide for the correct usage. Because OPE methods are potentially extensible to ethically sensitive areas, such as medicine, the OPE methods need to be applied with attention to possible issues that may arise.

Acknowledgement

Kaito Ariu was supported by the Nakajima Foundation Scholarship.

References

Beygelzimer, A. and Langford, J. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 129–138, 2009.

Bibaut, A., Malenica, I., Vlassis, N., and Van Der Laan, M. More efficient off-policy evaluation through regularized targeted learning. In *International Conference on Machine Learning*, pp. 654–663, 2019.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Blum, A., Kalai, A., and Langford, J. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 203–208, 1999.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018.

Chow, S.-C. and Chang, M. *Adaptive design methods in clinical trials*. CRC press, 2011.

Dudík, M., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. *In Proceedings of the 35th International Conference on Machine Learning*, pp. 1447–1456, 2018.

Gilotte, A., Calauzènes, C., Nedelec, T., Abraham, A., and Dollé, S. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 198–206, 2018.

Greene, W. *Econometric Analysis*. Pearson Education, 2003.

Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331, 1998.

Hahn, J., Hirano, K., and Karlan, D. Adaptive experimental design using the propensity score. *Journal of Business and Economic Statistics*, 29(1):96–108, 2011.

Hall, P., Heyde, C., Birnbaum, Z., and Lukacs, E. *Martingale Limit Theory and Its Application*. Communication and Behavior. Elsevier Science, 2014.

Hamilton, J. *Time series analysis*. Princeton Univ. Press, 1994.

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.

Kallus, N. and Uehara, M. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 3325–3334, 2019.

Kallus, N. and Uehara, M. Efficient evaluation of natural stochastic policies in offline reinforcement learning. arXiv:2006.03886, 2020.

Kato, M. Confidence interval for off-policy evaluation from dependent samples via bandit algorithm: Approach from standardized martingales. arXiv:2006.06982, 2020.

Kato, M., Ishihara, T., Honda, J., and Narita, Y. Adaptive experimental design for efficient treatment effect estimation: Randomized allocation via contextual bandit algorithm. arXiv:2002.05308, 2020a.

Kato, M., Uehara, M., and Yasui, S. Off-policy evaluation and learning for external validity under a covariate shift. arXiv:2002.11642, 2020b.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Li, S., Karatzoglou, A., and Gentile, C. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 539–548, 2016.

Loeve, M. *Probability Theory*. Graduate Texts in Mathematics. Springer, 1977.

Narita, Y., Yasui, S., and Yata, K. Efficient counterfactual learning from bandit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4634–4641, 2019.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4881–4890, 2019.

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., et al. Batched bandit problems. *The Annals of Statistics*, 44(2): 660–681, 2016.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

Saito, Y., Aihara, S., Matsutani, M., and Narita, Y. A large-scale open dataset for bandit algorithms. arXiv:2008.07146, 2020.

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

van der Laan, M. J. The construction and analysis of adaptive group sequential designs. U.C. Berkeley Division of Biostatistics Working Paper Series, 2008.

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.

Villar, S. S. Bandit strategies evaluated in the context of clinical trials in rare life-threatening diseases. *Probability in the engineering and informational sciences*, 32(2):229–245, 2018.

Wang, H., Xing, W., Asif, K., and Ziebart, B. D. Adversarial prediction games for multivariate losses. In *Advances in Neural Information Processing Systems 28*, pp. 2728–2736, 2015.

Wang, Y.-X., Agarwal, A., and Dudik, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3589–3597, 2017.

Yang, Y. and Zhu, D. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.

A Preliminaries

Mathematical Tools

Definition 1. [Uniformly Integrable, Hamilton (1994), p. 191] A sequence $\{A_t\}$ is said to be uniformly integrable if for every $\epsilon > 0$ there exists a number $c > 0$ such that

$$\mathbb{E}[|A_t| \cdot I[|A_t \geq c|]] < \epsilon$$

for all t.

Proposition 1. [Sufficient Conditions for Uniformly Integrable, Hamilton (1994), Proposition 7.7, p. 191] (a) Suppose there exist $r > 1$ and $M < \infty$ such that $\mathbb{E}[|A_t|^r] < M$ for all t. Then $\{A_t\}$ is uniformly integrable. (b) Suppose there exist $r > 1$ and $M < \infty$ such that $\mathbb{E}[|b_t|^r] < M$ for all t. If $A_t = \sum_{j=-\infty}^{\infty} h_j b_{t-j}$ with $\sum_{j=-\infty}^{\infty} |h_j| < \infty$, then $\{A_t\}$ is uniformly integrable.

Proposition 2 (L^r Convergence Theorem, Loeve (1977))**.** Let $0 < r < \infty$, suppose that $\mathbb{E}\big[|a_n|^r\big] < \infty$ for all n and that $a_n \xrightarrow{\text{p}} a$ as $n \to \infty$. The following are equivalent:

(i) $a_n \to a$ in L^r as $n \to \infty$;

(ii) $\mathbb{E}\big[|a_n|^r\big] \to \mathbb{E}\big[|a|^r\big] < \infty$ as $n \to \infty$;

(iii) $\big\{|a_n|^r, n \geq 1\big\}$ is uniformly integrable.

Martingale Limit Theorems

Proposition 3. [Weak Law of Large Numbers for Martingale, Hall et al. (2014)] Let $\{S_n = \sum_{i=1}^n X_i, \mathcal{H}_t, t \geq 1\}$ be a martingale and $\{b_n\}$ a sequence of positive constants with $b_n \to \infty$ as $n \to \infty$. Then, writing $X_{ni} = X_i \mathbb{1}[|X_i| \leq b_n]$, $1 \leq i \leq n$, we have that $b_n^{-1} S_n \xrightarrow{\text{p}} 0$ as $n \to \infty$ if

(i) $\sum_{i=1}^n P(|X_i| > b_n) \to 0$;

(ii) $b_n^{-1} \sum_{i=1}^n \mathbb{E}[X_{ni} \mid \mathcal{H}_{t-1}] \xrightarrow{\text{p}} 0$, and;

(iii) $b_n^{-2} \sum_{i=1}^n \big\{\mathbb{E}[X_{ni}^2] - \mathbb{E}\big[\mathbb{E}\big[X_{ni} \mid \mathcal{H}_{t-1}\big]\big]^2\big\} \to 0$.

Remark 7. The weak law of large numbers for martingale holds when the random variable is bounded by a constant.

Proposition 4. [Central Limit Theorem for a Martingale Difference Sequence, Hamilton (1994), Proposition 7.9, p. 194] Let $\{X_t\}_{t=1}^{\infty}$ be an n-dimensional vector martingale difference sequence with $\overline{X}_T = \frac{1}{T} \sum_{t=1}^T X_t$. Suppose that

(a) $\mathbb{E}[X_t^2] = \sigma_t^2$, a positive value with $(1/T) \sum_{t=1}^T \sigma_t^2 \to \sigma^2$, a positive value;

(b) $\mathbb{E}[|X_t|^r] < \infty$ for some $r > 2$;

(c) $(1/T) \sum_{t=1}^T X_t^2 \xrightarrow{p} \sigma^2$.

Then $\sqrt{T}\, \overline{X}_T \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2)$.

B OPE Terminologies

In this section, we introduce OPE terminologies used in this paper.

Behavior probability and behavior policy: First, we distinguish the behavior probability and behavior policy. We call a probability choosing an action, p_t, a *behavior probability* and a system generating the behavior probability the *behavior policy*. For example, the contextual bandit algorithm is a behavior policy, which returns a behavior probability based on a trajectory. By distinguishing them, we can clarify the goal of OPE; that is, the goal is to estimate the expected reward given a behavior probability.

Evaluation probability and evaluation policy: As well as the behavior policy and behavior probability, at an evaluation step, we call a probability choosing an action an *evaluation probability* and a system generating the evaluation probability the *evaluation policy*.

True bandit policy and pseudo bandit policy evaluations: Because an MAB algorithm controls a trajectory for balancing exploration and exploitation trade-off, an exact evaluation of a bandit policy requires the generation of the trajectory. Let us define the true bandit policy evaluation as an evaluation of an MAB algorithm with reproducing the trajectory generated from the MAB algorithm. However, in general, it is not easy to produce such a trajectory. Although Narita et al. (2019) and Saito et al. (2020) attempt to evaluate bandit policies, they fail to conduct the *true bandit policy evaluation*. In their experiments, they run several MAB algorithms on a fixed trajectory. We call such approaches *pseudo bandit policy evaluation* because they do not estimate the expected reward given by the bandit policy, unlike the true bandit policy evaluation. On the other hand, as far as we know, OPE for the true bandit policy evaluation has not been proposed and is an intractable problem.

C Asymptotic Normality of OPE Estimators

In general, for guaranteeing the asymptotic normality, we need to impose the Donsker's condition on $\hat{f}$. However, by using cross-fitting of DDM, we can show the asymptotic normality only with the convergence rate condition of an estimator of f^* (Chernozhukov et al., 2018). In 2-fold DDM, we first separate the dataset into two subsets. Then, we construct an estimator $\hat{f}$ of f^* from one of the datasets and OPE estimator from the other dataset using the estimator $\hat{f}$. When using the AIPW type estimator, we can show the asymptotic normality only by imposing convergence rate conditions on $\hat{f}$.

Next, we consider the case where the behavior probability is time-dependent. When samples have dependency under a behavior probability $\pi_t^{\mathrm{b}}(a \mid x, \mathcal{S}_{t-1})$, one of useful OPE estimators is Adaptive AIPW (A2IPW) estimator defined as $\widehat{R}_T^{\mathrm{A2IPW}}(\pi^{\mathrm{e}}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X_t)\Gamma_t^{\mathrm{AIPW}}(a; \pi_t^{\mathrm{b}}, \hat{f}_t)$, where $\hat{f}_t$ is a step-wise consistent estimator of f^* estimated only using samples $\mathcal{S}_{t-1}$. Then, for a converging behavior probability π_t^{b} such that $\pi_t^{\mathrm{b}}(a \mid x, \mathcal{S}_{t-1}) \xrightarrow{\mathrm{p}} \alpha(a \mid x)$ as $t \to \infty$, Kato et al. (2020a) showed the following proposition.

Proposition 5 (Asymptotic normality of A2IPW estimator, (Kato et al., 2020a))**.** Suppose that

(i) Pointwise convergence in probability of $\hat{f}_{t-1}$ and π_t, i.e., for all $x \in \mathcal{X}$ and $k \in \mathbb{N}$, $\hat{f}_{t-1}(k, x) - f^*(a, x) \xrightarrow{\mathrm{p}} 0$ and $\pi_t(a \mid x, \mathcal{S}_{t-1}) - \alpha(a \mid x) \xrightarrow{\mathrm{p}} 0$, where $\alpha : \mathcal{A} \times \mathcal{X} \to (0, 1)$;

(ii) There exists a constant C_3 such that $|\hat{f}_{t-1}| \leq C_3$.

Then, under Assumptions 1 and 2, we have $\sqrt{T}\left(\widehat{R}_T^{\mathrm{A2IPW}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$, where

$$\sigma^2_{\mathrm{A2IPW}}(\pi, \pi^{\mathrm{e}}) = \mathbb{E}\left[\sum_{k=0}^{1}\frac{\nu^*(k, X_t)}{\alpha(k \mid X_t)} + \left(f^*(1, X_t) - f^*(0, X_t) - R(\pi^{\mathrm{e}})\right)^2\right]. \tag{4}$$

Let us note that the consistency of f_t requires different theoretical analysis from the standard case because the samples are not i.i.d. For some specific bandit process, existing studies show consistencies of parametric and nonparametirc estimators (Yang & Zhu, 2002).

Remark 8 (Semiparametric Lower Bound)**.** The lower bound of the variance is defined for an estimator of an parameter of interest under some posited models of the DGP. If this posited model is a parametric model, it is equal to Cramér-Rao lower bound. When this posited model is non or semiparametric model, we can still define a corresponding Cramér-Rao lower bound Bickel et al. (1998). The semiparametric lower bound of the DGP defined in (1) with $\pi_1^{\mathrm{b}}(a \mid x, \mathcal{S}_0) = \pi_2^{\mathrm{b}}(a \mid x, \mathcal{S}_1) = \cdots = \pi_T^{\mathrm{b}}(a \mid x, \mathcal{S}_{T-1}) = \pi^{\mathrm{b}}(a \mid x)$ is given as follows (Narita et al., 2019):

$$\sigma^2_{\mathrm{OPT}}(\pi^{\mathrm{b}}, \pi^{\mathrm{e}}) = \mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a \mid X)\right)^2\nu^*(a, X_t)}{\pi^{\mathrm{b}}(a \mid X_t)} + \left(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X)f^*(a, X_t) - \theta_0\right)^2\right].$$

D Categorization based on behavior probability

In this section, we introduce a categorization based on behavior probability.

Categorization based on the Scenarios of Behavior Policy Update

For time-dependency of behavior policy generating a behavior probability $\pi_t^{\mathrm{b}}(a \mid x)$, we consider the following three cases. The first is the case where the behavior probability is time-invariant, i.e., $\pi_1^{\mathrm{b}}(a \mid x, \mathcal{S}_0) = \pi_2^{\mathrm{b}}(a \mid x, \mathcal{S}_1) = \cdots = \pi_T^{\mathrm{b}}(a \mid x, \mathcal{S}_{T-1}) = \pi^{\mathrm{b}}(a \mid x)$ (Li et al., 2011). In the second and third cases, we consider a behavior policy updates the behavior probabilities at period t using past information $\mathcal{S}_{t-1}$. Following Kato (2020), we classify the such behavior policies into two patterns, *sequential update policy* and *batch update policy*. Let us assume that a behavior probability only depends on the past information $\mathcal{S}_{t-1}$, and let $\pi^{\mathrm{b}} : \mathcal{A} \times \mathcal{X} \times \mathcal{M}_{t-1} \to (0, 1)$ be a time-dependent behavior probability. In sequential update policy, the probability $\pi^{\mathrm{b}}(a \mid x, \mathcal{S}_{t-1})$ is updated at each period (van der Laan, 2008; Kato et al., 2020a). In batch update policy, after using a fixed probability $\pi^{\mathrm{b}}(a \mid x, \mathcal{S}_{t-1})$ for some periods without updating, the probability $\pi^{\mathrm{b}}(a \mid x, \mathcal{S}_{t-1})$ is updated (Hahn et al., 2011; Narita et al., 2019). Although the sequential update is standard in MAB problems, we often apply batch update in industrial applications such as ad-optimization (Perchet et al., 2016; Narita et al., 2019). In this paper, without loss of generality, when discussing the case where a behavior probability is time-dependent, we only consider sequential update policy.

Categorization based on the Number of Behavior Probabilities

The next categorization is based on the number of behavior probabilities. For simplicity, let us consider the case where the behavior probabilities are time-invariant. We can extend the result to the case where the behavior probabilities are time-dependent without loss of generality. The concepts are illustrated in Figure 3.

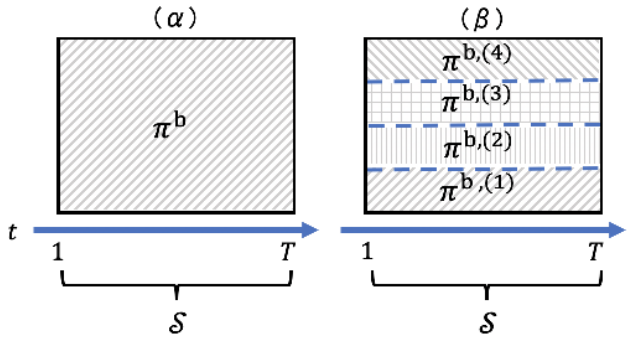

Figure 3: Illustrations of cases (α) and (β). In (β) Multiple behavior probabilities, we assume that there are multiple behavior algorithm on a time-series.

(α) **One behavior probability:** When there is one behavior probability, we use the standard OPE estimator.

(β) **Multiple behavior probabilities:** Let us separate the dataset into M subsets, $\{\mathcal{S}_{T^{(m)}}^{(m)}\}_{m=1}^{M}$. For each group m, we run a behavior probability $\pi^{\mathrm{b},(m)}$ as

$$\mathcal{S}_{T_{(m)}}^{(m)} = \left\{\left(X_t^{(m)}, A_t^{(m)}, Y_t^{(m)}\right)\right\}_{t^{(m)}=1}^{T^{(m)}} \overset{\text{i.i.d.}}{\sim} p(x)\pi^{\mathrm{b},(m)}(a \mid x)p(y \mid a, x),$$

where $T^{(m)}$ is the sample size of the dataset $\mathcal{S}_{T^{(m)}}^{(m)}$. When there are M behavior probabilities $\{\pi^{\mathrm{b},(m)}\}_{m=1}^{M}$ for a time series, we can consider two approaches based on the existing OPE estimators.

First, if a sample is uniform randomly grouped into one of M groups, we can define the behavior probability as $\pi^{\mathrm{b,mix}} = \frac{1}{M}\sum_{m=1}^{M}\pi^{\mathrm{b},(m)}$. Then, we apply the standard OPE estimators using the created behavior probability $\pi^{\mathrm{b,mix}}$. If we use an AIPW estimator under regularity conditions, the asymptotic variance is given as

$$\sigma_{\mathrm{MS}}^2 = \mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a, X_t)\right)^2\nu^*(a, X_t)}{\pi^{\mathrm{b,mix}}(a \mid X_t)} + \left(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a, X_t)f^*(a, X_t) - R(\pi^{\mathrm{e}})\right)^2\right],$$

Second, we consider using the generalized method of moments (GMM) for M datasets with different behavior probabilities Hamilton (1994). For ease of discussion, we use an AIPW type estimator. Let D_T be an M dimensional vector $(D_T^{(1)} \ \cdots \ D_T^{(M)})^\top$, where $D_T^{(m)}$ is $\frac{1}{T^{(m)}}\sum_{t=1}^{T}\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X_t)\Gamma_t^{\mathrm{AIPW}}\left(a; \pi^{\mathrm{b},(m)}, \hat{f}_T\right)\mathbb{1}\left[(X_t, A_t, Y_t) \in \mathcal{S}_T^{(m)}\right]$,

where $\hat{f}_T$ is a consistent estimator of f^* w.r.t. the sample size T. Then, we estimate $R(\pi^{\mathrm{e}})$ by $\widehat{R}^{\mathrm{GMM}}(\pi^{\mathrm{e}}) = \arg\min_{R\in\mathbb{R}}(D_T - I_M R)^\top W (D_T - I_M R)$, where I_M is an M dimensional vector $I_M = (1\ 1\ \cdots\ 1)$ and W is an $M \times M$ dimensional positive semidefinite weight matrix. The solution is analytically obtained as $\widehat{R}^{\mathrm{GMM}}_W(\pi^{\mathrm{e}}) = (I_M^\top W I_M)^{-1} I_M^\top W D_T$. Let us assume that $D_T^{(m)}$ converges to the asymptotic normal distribution with the asymptotic variance $\sigma^2_{(m)}$. Then, for the variance $\sigma^2_{(m)}$, when using a weight function W^* such that $(I_M^\top W^* I_M)^{-1} I_M^\top W^* = \left(\frac{1}{\sigma^2_{(1)}} / \sum_{m'=1}^M \frac{1}{\sigma^2_{(m')}}, \ldots, \frac{1}{\sigma^2_{(M)}} / \sum_{m'=1}^M \frac{1}{\sigma^2_{(m')}} \right)^\top$, the asymptotic distribution is given as follows.

Theorem 4 (Asymptotic variance under a stratified sampling)**.** *Suppose that there are M independent datasets, $\left\{\mathcal{S}_T^{(m)}\right\}_{m=1}^M$. If $D_T \xrightarrow{\mathrm{d}} (I_M R(\pi^{\mathrm{e}}), \Sigma)$, where Σ is an M dimensional diagonal matrix with m-th diagonal element $\sigma^2_{(m)}$, then $\widehat{R}^{\mathrm{GMM}}_W(\pi^{\mathrm{e}})$ asymptotically follows normal distribution with mean $R(\pi^{\mathrm{e}})$ and the variance $\sigma^2_{\mathrm{SS}} = \left\{ \sum_{m=1}^M \frac{1}{\sigma^2_{(m)}} \right\}^{-1}$, where*

$$\sigma^2_{(m)} = \frac{T}{T^{(m)}} \mathbb{E}\left[\sum_{a=1}^K \frac{\left(\pi^{\mathrm{e}}(a, X_t)\right)^2 \nu^*(a, X_t)}{\pi^{\mathrm{b},(m)}(a \mid X_t)} + \Big(\sum_{a=1}^K \pi^{\mathrm{e}}(a, X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}}) \Big)^2 \right].$$

For example, $D_T \xrightarrow{\mathrm{d}} (I_M R(\pi^{\mathrm{e}}), \Sigma)$ can be shown when using DDM. Note that, for $m \neq m'$, the covariance is 0, i.e.,

$$\mathbb{E}\left[\Gamma_t^{\mathrm{AIPW}}\left(a; \pi^{\mathrm{b},(m)}, \hat{f}_T\right) \mathbb{1}\left[(X_t, A_t, Y_t) \in \mathcal{S}_T^{(m)}\right] \Gamma_t^{\mathrm{AIPW}}\left(a; \pi^{\mathrm{b},(m')}, \hat{f}_T\right) \mathbb{1}\left[(X_t, A_t, Y_t) \in \mathcal{S}_T^{(m')}\right] \right] = 0.$$

The difference of the asymptotic variance comes from the difference of DGPs. The first case is called a mixture sampling, and the second case is called a stratified sampling. When considering an estimator $\widehat{R}^{\mathrm{GMM}}(\pi^{\mathrm{e}}) = \arg\min_{R\in\mathbb{R}}(D_T - I_M R)^\top W (D_T - I_M R)$, the first case is a special case of the estimator of the second case with a weight function $W' = (T^{(1)}, \ldots, T^{(M)})^\top$. For the estimator such as $\widehat{R}^{\mathrm{GMM}}(\pi^{\mathrm{e}})$, when using the weight W^* defined above, the asymptotic variance is minimized Hamilton (1994). Therefore, $\sigma^2_{\mathrm{MS}} \geq \sigma^2_{\mathrm{SS}}$ holds.

Experiments of GMM: Saito et al. (2020) released a dataset for OPE, which is a logged bandit feedback collected on a large-scale fashion e-commerce platform, ZOZOTOWN. The data is generated from Bernoulli Thompson Sampling (BTS) and random policies. The action space is $\mathcal{A} = \{1, 2, \ldots, 80\}$. The dimension of feature vector is 27 in BTS policy data and 26 in random policy data. The sample size of BTS policy data is $12,357,200$ and that of random policy data is $1,374,327$. Let an evaluation policy be $\pi^{\mathrm{e}}(a \mid X_t) = a / \sum_{a'=1}^{80} a'$. Let us make an evaluation dataset $\left\{\mathcal{S}^{BTS}_{T^{BTS}}, \mathcal{S}^{random}_{T^{random}}\right\}$, where $\mathcal{S}^{BTS}_{T^{BTS}}$ is generated from the BTS policy data, $\mathcal{S}^{random}_{T^{random}}$ is generated from the random policy data, and T^{BTS} and T^{random} denote the sample sizes, respectively. For the dataset, we construct four estimators: AIPW estimator only using $\mathcal{S}^{BTS}_{T^{BTS}}$ (BTSAIPW), AIPW estimator only using $\mathcal{S}^{random}_{T^{random}}\}$ (RAIPW), AIPW estimator using mixed behavior policy π^{mix} defined in Section 4 (MAIPW), and AIPW based GMM estimator $\widehat{R}^{\mathrm{GMM}}_W(\pi^{\mathrm{e}})$ (GMM) defined in Section 4. We calculate the true value of the evaluation policy by using $1,000,000$ data generated from random policy data, which is not used for the evaluation data. For tree dataset with sample sizes $T^{BTS} = T^{random} = 100,000$, $T^{BTS} = T^{random} = 150,000$, and $T^{BTS} = T^{random} = 200,000$, we calculate the root MSEs (RMSEs) between the estimators and the true value and the standard deviations (SD). The result is shown in Table 4. Among the estimators, the GMM estimator achieves the lowest MSEs.

Table 4: Experimental results of OPE from multiple behavior policy. The best performing method is highlighted in bold, where * denotes that there is a 95% significant difference compared with the suboptimal method in t-test.

sample size	150,000		150,000		150,000	
	RMSE	SD	RMSE	SD	RMSE	SD
BTSAIPW	0.00151	0.00128	0.00136	0.00105	0.00129	0.00079
RAIPW	0.00026	0.00018	0.00026	0.00016	0.00026	0.00015
MAIPW	0.00069	0.00064	0.00058	0.00053	0.00055	0.00040
GMM	**0.00021**	0.00015	**0.00019***	0.00014	**0.00018**	0.00013

E Proof of Theorem 1

For $Z_{t,0:-T'} = \sum_{a=1}^{K} \pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \Gamma^{\mathrm{AIPW}}_t(a; \pi^{\mathrm{b}}_{0:-T'}, \hat{f}_{t-1}) - R(\pi^{\mathrm{e}}_{0:-T'})$, we show

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} Z_{t,0:-T'}\right) \to \mathcal{N}\left(0, \sigma^{2\dagger}\right),$$

where note that

$$\Gamma^{\mathrm{AIPW}}_t(a; \pi^{\mathrm{b}}, f) = \frac{\mathbb{1}[A_t = a](Y_t - f(a, X_t))}{\pi^{\mathrm{b}}(A_t \mid X_t)} + f(a, X_t).$$

Then, the sequence $\{Z_t\}_{t=1}^{T}$ is an MDS, i.e.,

$$\begin{aligned}
&\mathbb{E}\big[Z_t \mid \mathcal{S}_{t-1}\big] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \Gamma^{\mathrm{AIPW}}_t(a; \pi^{\mathrm{b}}_{0:-T'}, \hat{f}_{t-1}) - R(\pi^{\mathrm{e}}_t) \mid \mathcal{S}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi^{\mathrm{e}}_t) \mid \mathcal{S}_{t-1}\right] + \mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi^{\mathrm{e}}_t(a \mid X_t, \mathcal{S}_{t-1}) \mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}_{0:-T'}(A_t \mid X_t)} \mid \mathcal{S}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi^{\mathrm{e}}_t) \mid \mathcal{S}_0\right] + \mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi^{\mathrm{e}}_t(a \mid X_t, \mathcal{S}_{t-1}) \mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}_{0:-T'}(A_t \mid X_t)} \mid \mathcal{S}_0\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi^{\mathrm{e}}_t) \mid \mathcal{S}_0\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}_{0:-T'}(a \mid X_t)} \mid X_t, \mathcal{S}_0\right] \mid \mathcal{S}_0\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi^{\mathrm{e}}_t) + \sum_{a=1}^{K} \pi^{\mathrm{e}}_t(a \mid X_t, \mathcal{S}_{t-1}) f^*(a, X_t) - \sum_{a=1}^{K} \pi^{\mathrm{e}}_{0:-T'}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) \mid \mathcal{S}_0\right] \\
&= 0.
\end{aligned}$$

Therefore, to derive the asymptotic distribution, we consider applying the CLT for an MDS introduced in Proposition 4. There are following three conditions in the statement.

(a) $\mathbb{E}\big[Z^2_{t,0:-T'}\big] = \nu^2_t > 0$ with $\big(1/T\big)\sum_{t=1}^{T} \nu^2_t \to \nu^2 > 0$;

(b) $\mathbb{E}\big[|Z_{t,0:-T'}|^r\big] < \infty$ for some $r > 2$;

(c) $\big(1/T\big)\sum_{t=1}^{T} Z^2_{t,0:-T'} \xrightarrow{\mathrm{p}} \nu^2$.

The proof procedure is almost same as that of Theorem 2. Therefore, we omit showing the above conditions here.

F Efficient Experimental Design

In the case (b) ITOPE, we have an evaluation probability at a period $t = 1$. Therefore, a naive idea for estimating the policy value is to conduct an RCT (A/B testing) using a dataset $\mathcal{S}$. However, as van der Laan (2008), Hahn et al. (2011), and Kato et al. (2020a) proposed, we can optimize the behavior probability π^{b} for minimizing the asymptotic variance of the OPE estimator, which is more efficient than a plain RCT. In this section, following the existing studies of van der Laan (2008), Hahn et al. (2011), and Kato et al. (2020a), which propose an experimental design for one evaluation probability, we introduce an efficient experimental design for multiple evaluation probabilities.

Let us consider a situation where there is a set of M evaluation probabilities, $\Pi_t = \left\{\pi^{\mathrm{e},(1)}, \pi^{\mathrm{e},(2)}, \ldots, \pi^{\mathrm{e},(M)}\right\}$ at period $t = 1$. Let us assume that we know the variance $\nu^*(a, x)$ before experiment. In Remark 9, we introduce a method that weakens this assumption. In this case, our recommendation for designing an efficient experiment via optimizing a behavior probability is to use the behavior probability defined as

$$\pi^{\mathrm{b}*} = \arg\min_{\pi^{\mathrm{b}} \in \mathcal{H}} \max_{\pi^{\mathrm{e}} \in \Pi_t} \sigma^2_{\mathrm{OPT}}(\pi^{\mathrm{b}}, \pi^{\mathrm{e}}), \tag{5}$$

where $\mathcal{H}$ is a class of possible behavior probabilities. note that σ^2_{OPT} is the semi-parametric lower bound given a behavior probability π^{b} and evaluation probability π^{e}. This minimax formulation gives us a behavior probability that minimizes the worst semi-parametric lower bound among M-candidate probabilities.

For independent M hypothesis testings on the null hypothesis $R(\pi^{\mathrm{e},(m)}) = 0$ for $m = 1, 2, \ldots, M$, we also show the sample size need for conducting hypothesis testing with power β when using AIPW estimator under regularity conditions.

Corollary 1 (Sample size for the experiment, Kato et al. (2020a))**.** *Suppose that the effect size Δ and AIPW estimator has the asymptotic normality with the asymptotic variance achieving the semi-parametric lower bound. Then, the sample size of power β test with controlling Type I error at α is*

$$T_\beta(\Delta) = \frac{\max_{\pi^{\mathrm{e}} \in \Pi_t} \sigma^2_{\mathrm{AIPW}}(\pi^{\mathrm{b}*}, \pi^{\mathrm{e}})}{\Delta^2} \left(z_{1-\alpha/2} - z_\beta\right)^2.$$

Remark 9 (Sequential Estimation)**.** When the variance ν^* is unknown, we can sequentially estimate the variance and update an estimator of the efficient behavior probability $\pi^{\mathrm{b}*}(a \mid x)$. For this method in ATE estimation, Kato et al. (2020a) showed that an A2IPW estimator has the same asymptotic distribution as AIPW estimator with the known variance.

Remark 10 (Parallel RCTs)**.** We consider the case where we cannot use OPE for estimating the policy value, and need to conduct parallel RCTs for each evaluation probability by separating the dataset without adjusting the behavior probability, i.e., the evaluation probability is equal to the evaluation probability. In this case, we can still optimize the sample size of each RCT for equalizing the power of each test.

G Proof of Theorem 2

The procedure of this proof follows Kato et al. (2020a).

Proof. For $Z_t = \sum_{a=1}^K \pi^{\mathrm{e}}_t(a \mid X_t, \mathcal{S}_{t-1}) \Gamma^{\mathrm{AIPW}}_t(a; \pi^{\mathrm{b}}, \hat{f}_{t-1}) - R(\pi^{\mathrm{e}}_t)$, we show

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^T Z_t\right) \to \mathcal{N}\left(0, \widetilde{\sigma}^2\right),$$

where note that

$$\Gamma^{\mathrm{AIPW}}_t(a; \pi^{\mathrm{b}}, f) = \frac{\mathbb{1}[A_t = a](Y_t - f(a, X_t))}{\pi^{\mathrm{b}}(A_t \mid X_t)} + f(a, X_t).$$

Then, the sequence $\{Z_t\}_{t=1}^T$ is an MDS, i.e.,

$$
\begin{aligned}
&\mathbb{E}\big[Z_t \mid \mathcal{S}_{t-1}\big] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \Gamma_t^{\mathrm{AIPW}}(a; \pi, \hat{f}_{t-1}) - R(\pi_t^{\mathrm{e}}) \mid \mathcal{S}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}}) \mid \mathcal{S}_{t-1}\right] + \mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}(A_t \mid X_t)} \mid \mathcal{S}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}}) \mid \mathcal{S}_{t-1}\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}(a \mid X_t)} \mid X_t, \mathcal{S}_{t-1}\right] \mid \mathcal{S}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}}) + \sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) f^*(a, X_t) - \sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) \mid \mathcal{S}_{t-1}\right] \\
&= 0.
\end{aligned}
$$

Therefore, to derive the asymptotic distribution, we consider applying the CLT for an MDS introduced in Proposition 4. There are following three conditions in the statement.

(a) $\mathbb{E}\big[Z_t^2\big] = \nu_t^2 > 0$ with $\big(1/T\big) \sum_{t=1}^{T} \nu_t^2 \to \nu^2 > 0$;

(b) $\mathbb{E}\big[|Z_t|^r\big] < \infty$ for some $r > 2$;

(c) $\big(1/T\big) \sum_{t=1}^{T} Z_t^2 \xrightarrow{\mathrm{p}} \nu^2$.

Because we assumed the boundedness of z_t by assuming the boundedness of Y_t, $\hat{f}_{t-1}$, and $1/\pi_t$, the condition (b) holds. Therefore, the remaining task is to show the conditions (a) and (c) hold.

Step 1: Check of Condition (a)

We can rewrite $\mathbb{E}\big[Z_t^2\big]$ as

$$
\begin{aligned}
\mathbb{E}\big[Z_t^2\big] =& \mathbb{E}\left[\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \left(\frac{\mathbb{1}[A_t = a](Y_t - f(a, X_t))}{\pi^{\mathrm{b}}(a \mid X_t)} + f(a, X_t)\right) - R(\pi_t^{\mathrm{e}})\right)^2\right] \\
&- \mathbb{E}\left[\sum_{a=1}^{K} \frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \nu^*\big(a, X_t\big)}{\pi^{\mathrm{b}}(a \mid X_t)} + \Big(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\right] \\
&+ \mathbb{E}\left[\sum_{a=1}^{K} \frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \nu^*\big(a, X_t\big)}{\pi^{\mathrm{b}}(a \mid X_t)} + \Big(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\right].
\end{aligned}
$$

We will prove that the RHS of the following equation varnishes asymptotically to show that the condition (a) holds.

$$
\begin{aligned}
&\mathbb{E}\big[Z_t^2\big] - \mathbb{E}\left[\sum_{a=1}^{K} \frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \nu^*\big(a, X_t\big)}{\pi^{\mathrm{b}}(a \mid X_t)} + \Big(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \left(\frac{\mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}(a \mid X_t)} + \hat{f}_{t-1}(a, X_t)\right) - R(\pi_t^{\mathrm{e}})\right)^2\right] \\
&\quad - \mathbb{E}\left[\sum_{a=1}^{K} \frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \nu^*\big(a, X_t\big)}{\pi^{\mathrm{b}}(a \mid X_t)} + \Big(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\right]. \qquad (6)
\end{aligned}
$$

First, for the first term of the RHS of (6),

$$
\begin{aligned}
&\mathbb{E}\left[\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\left(\frac{\mathbb{1}[A_t=a](Y_t-\hat{f}_{t-1}(a,X_t))}{\pi^{\mathrm{b}}(a\mid X_t)}+\hat{f}_{t-1}(a,X_t)\right)-R(\pi_t^{\mathrm{e}})\right)^2\right]\\
&=\sum_{a=1}^{K}\mathbb{E}\left[\left(\frac{\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\mathbb{1}[A_t=a]\big(Y_t-\hat{f}_{t-1}(a,X_t)\big)}{\pi^{\mathrm{b}}(a\mid X_t)}\right)^2\right]\\
&\quad+\mathbb{E}\left[\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2\right]\\
&\quad+2\sum_{a=1}^{K-1}\sum_{a'=a+1}^{K}\mathbb{E}\left[\left(\frac{\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\mathbb{1}[A_t=a]\big(Y_t-\hat{f}_{t-1}(a,X_t)\big)}{\pi^{\mathrm{b}}(a\mid X_t)}\right)\left(\frac{\pi_t^{\mathrm{e}}(a'\mid X_t,\mathcal{S}_{t-1})\mathbb{1}[A_t=a']\big(Y_t-\hat{f}_{t-1}(a',X_t)\big)}{\pi^{\mathrm{b}}(a'\mid X_t)}\right)\right]\\
&\quad+2\sum_{a=1}^{K}\mathbb{E}\left[\left(\frac{\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\mathbb{1}[A_t=a]\big(Y_t-\hat{f}_{t-1}(a,X_t)\big)}{\pi^{\mathrm{b}}(a\mid X_t)}\right)\left(\sum_{a'=1}^{K}\pi_t^{\mathrm{e}}(a'\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a',X_t)-R(\pi_t^{\mathrm{e}})\right)\right].
\end{aligned}
$$

Because $\mathbb{1}[A_t=a]\mathbb{1}[A_t=a']=0$ for $a\neq a'$, $\mathbb{1}[A_t=a]\mathbb{1}[A_t=a]=\mathbb{1}[A_t=a]$, and $\mathbb{1}[A_t=a]Y_t=Y_t(a)$ for all $a\in\mathcal{A}$ and $a'\neq a$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\mathbb{1}[A_t=a]\big(Y_t-\hat{f}_{t-1}(a,X_t)\big)}{\pi^{\mathrm{b}}(a\mid X_t)}\right)^2\right]=\mathbb{E}\left[\frac{\left(\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\right)^2\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}\right],\\
&\mathbb{E}\left[\left(\frac{\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\mathbb{1}[A_t=a]\big(Y_t-\hat{f}_{t-1}(a,X_t)\big)}{\pi^{\mathrm{b}}(a\mid X_t)}\right)\left(\frac{\pi_t^{\mathrm{e}}(a'\mid X_t,\mathcal{S}_{t-1})\mathbb{1}[A_t=a']\big(Y_t-\hat{f}_{t-1}(a',X_t)\big)}{\pi^{\mathrm{b}}(a'\mid X_t)}\right)\right]=0,\\
&\mathbb{E}\left[\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\frac{\mathbb{1}[A_t=a](Y_t-\hat{f}_{t-1}(a,X_t))}{\pi^{\mathrm{b}}(a\mid X_t)}\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\frac{\mathbb{1}[A_t=a](Y_t-\hat{f}_{t-1}(a,X_t))}{\pi^{\mathrm{b}}(a\mid X_t)}\mid X_t,\mathcal{S}_{t-1}\right]\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right]\\
&=\mathbb{E}\left[\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right].
\end{aligned}
$$

Therefore, for the first term of the RHS of (6),

$$
\begin{aligned}
&\mathbb{E}\left[\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\left(\frac{\mathbb{1}[A_t=a](Y_t-\hat{f}_{t-1}(a,X_t))}{\pi^{\mathrm{b}}(a\mid X_t)}+\hat{f}_{t-1}(a,X_t)\right)-R(\pi_t^{\mathrm{e}})\right)^2\right]\\
&=\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\right)^2\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}+\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2+\right.\\
&\quad\left.2\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right].
\end{aligned}
$$

and, for the second term of the RHS of (6),

$$\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a\mid X_t)\right)^2\nu^*\left(a,X_t\right)}{\pi^{\mathrm{b}}(a\mid X_t)}+\Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\Big)^2\right]$$
$$=\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a\mid X_t)\right)^2\left(Y_t(a)-f^*(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}+\Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\Big)^2\right].$$

Then, using these equations, the RHS of (6) can be calculated as

$$\mathbb{E}\left[\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\left(\frac{\mathbb{1}[A_t=a](Y_t-\hat{f}_{t-1}(a,X_t))}{\pi^{\mathrm{b}}(a\mid X_t)}+\hat{f}_{t-1}(a,X_t)\right)-R(\pi_t^{\mathrm{e}})\right)^2\right]$$
$$-\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a\mid X_t)\right)^2\nu^*\left(a,X_t\right)}{\pi^{\mathrm{b}}(a\mid X_t)}+\Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\Big)^2\right]$$
$$=\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\right)^2\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}+\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2+\right.$$
$$\left.2\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right]$$
$$-\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a\mid X_t)\right)^2\left(Y_t(a)-f^*(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}+\Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\Big)^2\right].$$

Then, from the triangle inequality, we have

$$\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\right)^2\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}+\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2+\right.$$
$$\left.2\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right]$$
$$-\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a\mid X_t)\right)^2\left(Y_t(a)-f^*(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}+\Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\Big)^2\right]$$
$$\leq\sum_{a=1}^{K}\mathbb{E}\left[\left|\frac{\left(\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\right)^2\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}-\frac{\left(\pi^{\mathrm{e}}(a\mid X_t)\right)^2\left(Y_t(a)-f^*(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}\right|\right]$$
$$+\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2-\Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\Big)^2\right|\right]$$
$$+2\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right|\right].$$

Because all elements are assumed to be bounded and $b_1^2-b_2^2=(b_1+b_2)(b_1-b_2)$ for variables b_1 and b_2, there exist constants $\tilde{C}_0$, $\tilde{C}_1$, $\tilde{C}_2$, and $\tilde{C}_3$ such that

$$
\begin{aligned}
&\sum_{a=1}^{K}\mathbb{E}\left[\left|\frac{(\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1}))^2\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}-\frac{\pi^{\mathrm{e}}(a\mid X_t)\left(Y_t(a)-f^*(a,X_t)\right)^2}{\pi^{\mathrm{b}}(a\mid X_t)}\right|\right]\\
&\quad+\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2-\left(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\right)^2\right|\right]\\
&\quad+2\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right|\right]\\
&\leq\tilde{C}_0\sum_{a=1}^{K}\mathbb{E}\left[\left|\frac{\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)}{\sqrt{\pi^{\mathrm{b}}(a\mid X_t)}}-\frac{\pi^{\mathrm{e}}(a\mid X_t)\left(Y_t(a)-f^*(a,X_t)\right)}{\sqrt{\pi^{\mathrm{b}}(a\mid X_t)}}\right|\right]\\
&\quad+\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2-\left(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\right)^2\right|\right]\\
&\quad+2\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right|\right]\\
&\leq\tilde{C}_1\sum_{a=1}^{K}\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\left(Y_t(a)-\hat{f}_{t-1}(a,X_t)\right)-\pi^{\mathrm{e}}(a\mid X_t)\left(Y_t(a)-f^*(a,X_t)\right)\right|\right]\\
&\quad+\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)^2-\left(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a\mid X_t)f^*(a,X_t)-R(\pi^{\mathrm{e}})\right)^2\right|\right]\\
&\quad+2\mathbb{E}\left[\left|\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})f^*(a,X_t)-\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-R(\pi_t^{\mathrm{e}})\right)\right|\right]\\
&\leq\tilde{C}_1\sum_{a=1}^{K}\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-\pi^{\mathrm{e}}(a\mid X_t)f^*(k,X_t)\right|\right]\\
&\quad+\tilde{C}_2\sum_{a=1}^{K}\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})-\pi^{\mathrm{e}}(a\mid X_t)\right|\right]+\tilde{C}_3\sum_{a=1}^{K}\mathbb{E}\left[\left|\hat{f}_{t-1}(a,X_t)-f^*(a,X_t)\right|\right].
\end{aligned}
$$

Then, from $b_1b_2-b_3b_4=(b_1-b_3)b_4-(b_4-b_2)b_1$ for variables b_1, b_2, b_3, and b_4, there exist $\tilde{C}_4$ and $\tilde{C}_5$ such that

$$
\begin{aligned}
&\tilde{C}_1\sum_{a=1}^{K}\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})\hat{f}_{t-1}(a,X_t)-\pi^{\mathrm{e}}(a\mid X_t)f^*(k,X_t)\right|\right]\\
&\quad+\tilde{C}_2\sum_{a=1}^{K}\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})-\pi^{\mathrm{e}}(a\mid X_t)\right|\right]+\tilde{C}_3\sum_{a=1}^{K}\mathbb{E}\left[\left|\hat{f}_{t-1}(a,X_t)-f^*(a,X_t)\right|\right]\\
&\leq\tilde{C}_4\sum_{a=1}^{K}\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})-\pi^{\mathrm{e}}(a\mid X_t)\right|\right]+\tilde{C}_5\sum_{a=1}^{K}\mathbb{E}\left[\left|\hat{f}_{t-1}(a,X_t)-f^*(a,X_t)\right|\right].
\end{aligned}
$$

From the assumption that the pointwise convergences in probability, i.e., for all $x\in\mathcal{X}$ and $k\in\mathcal{A}$, $\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})-\pi^{\mathrm{e}}(a\mid X_t)\xrightarrow{\mathrm{p}}0$ and $\hat{f}_{t-1}(a,x)-f^*(a,x)\xrightarrow{\mathrm{p}}0$ as $t\to\infty$, if $\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})$, and $\hat{f}_{t-1}(a,x)$ are uniformly integrable, for fixed $x\in\mathcal{X}$, we can prove that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid X_t,\mathcal{S}_{t-1})-\pi^{\mathrm{e}}(a\mid X_t)\right|\mid X_t=x\right]=\mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a\mid x,\mathcal{S}_{t-1})-\pi^{\mathrm{e}}(a\mid x)\right|\right]\to0,\\
&\mathbb{E}\left[\left|\hat{f}_{t-1}(a,X_t)-f^*(a,X_t)\right|\mid X_t=x\right]=\mathbb{E}\left[\left|\hat{f}_{t-1}(a,x)-f^*(a,x)\right|\right]\to0,
\end{aligned}
$$

as $t \to \infty$ using L^r-convergence theorem (Proposition 2). Here, we used the fact that $\hat{f}_{t-1}(a,x)$ and $\pi_t^{\mathrm{e}}(a \mid x, \mathcal{S}_{t-1})$ are independent from X_t. For fixed $x \in \mathcal{X}$, we can show that $\pi_t^{\mathrm{e}}(a \mid x, \mathcal{S}_{t-1})$ and $\hat{f}_{t-1}(a,x)$ are uniformly integrable from the boundedness of $\pi_t^{\mathrm{e}}(a \mid x, \mathcal{S}_{t-1})$, and $\hat{f}_{t-1}(a,x)$ (Proposition 1). From the pointwise convergence of $\mathbb{E}[|\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) - \pi^{\mathrm{e}}(a \mid X_t)| \mid X_t = x]$ and $\mathbb{E}[|\hat{f}_{t-1}(a, X_t) - f^*(a, X_t)| \mid X_t = x]$, by using the Lebesgue's dominated convergence theorem, we can show that

$$\mathbb{E}_{X_t}\big[\mathbb{E}\big[|\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) - \pi^{\mathrm{e}}(a \mid X_t)| \mid X_t\big]\big] \to 0,$$
$$\mathbb{E}_{X_t}\big[\mathbb{E}\big[|\hat{f}_{t-1}(a, X_t) - f^*(a, X_t)| \mid X_t\big]\big] \to 0.$$

Then, as $t \to \infty$,

$$\mathbb{E}\big[Z_t^2\big] - \mathbb{E}\left[\sum_{a=1}^K \frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \nu^*\big(a, X_t\big)}{\pi^{\mathrm{b}}(a \mid X_t)} + \Big(\sum_{a=1}^K \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\right] \to 0.$$

Therefore, for any $\epsilon > 0$, there exists $\tilde{t} > 0$ such that

$$\frac{1}{T}\sum_{t=1}^T \left(\mathbb{E}\big[Z_t^2\big] - \mathbb{E}\left[\sum_{a=1}^K \frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \nu^*\big(a, X_t\big)}{\pi^{\mathrm{b}}(a \mid X_t)} + \Big(\sum_{a=1}^K \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\right]\right) \leq \frac{\tilde{t}}{T} + \epsilon.$$

Here,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{a=1}^K \frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \nu^*\big(a, X_t\big)}{\pi^{\mathrm{b}}(a \mid X_t)} + \Big(\sum_{a=1}^K \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2\right] \\
&= \mathbb{E}\left[\sum_{a=1}^K \frac{\big(\pi^{\mathrm{e}}(a \mid X)\big)^2 \nu^*\big(a, X\big)}{\pi^{\mathrm{b}}(a \mid X)} + \Big(\sum_{a=1}^K \pi^{\mathrm{e}}(a \mid X) f^*(a, X) - R(\pi^{\mathrm{e}})\Big)^2\right]
\end{aligned}$$

does not depend on periods. Therefore, $\big(1/T\big)\sum_{t=1}^T \sigma_t^2 - \sigma^2 \leq \tilde{t}/T + \epsilon \to 0$ as $T \to \infty$, where

$$\sigma_{\mathrm{AP}}^2 = \mathbb{E}\left[\sum_{a=1}^K \frac{\big(\pi^{\mathrm{e}}(a \mid X)\big)^2 \nu^*\big(a, X\big)}{\pi^{\mathrm{b}}(a \mid X)} + \Big(\sum_{a=1}^K \pi^{\mathrm{e}}(a \mid X) f^*(a, X) - R(\pi^{\mathrm{e}})\Big)^2\right].$$

Step 2: Check of Condition (c)

Let U_t be an MDS such that

$$\begin{aligned}
U_t &= Z_t^2 - \mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big] \\
&= \left(\sum_{a=1}^K \frac{\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}(A_t \mid X_t)} + \sum_{a=1}^K \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 \\
&- \mathbb{E}\left[\left(\sum_{a=1}^K \frac{\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \mathbb{1}[A_t = a](Y_t - \hat{f}_{t-1}(a, X_t))}{\pi^{\mathrm{b}}(A_t \mid X_t)} + \sum_{a=1}^K \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 \mid \mathcal{S}_{t-1}\right].
\end{aligned}$$

From the boundedness of each variable in Z_t, we can apply weak law of large numbers for an MDS (Proposition 3 in Appendix A), and obtain

$$\frac{1}{T}\sum_{t=1}^T U_t = \frac{1}{T}\sum_{t=1}^T \big(Z_t^2 - \mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big]\big) \xrightarrow{\mathrm{p}} 0.$$

Next, we show that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big] - \sigma_{\mathrm{AP}}^2 \xrightarrow{\mathrm{p}} 0.$$

From Markov's inequality, for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big] - \sigma_{\mathrm{AP}}^2\right| \geq \varepsilon\right) \\
&\leq \frac{\mathbb{E}\left[\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big] - \sigma_{\mathrm{AP}}^2\right|\right]}{\varepsilon} \\
&\leq \frac{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left|\mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big] - \sigma_{\mathrm{AP}}^2\right|\right]}{\varepsilon}.
\end{aligned}$$

Then, we consider showing $\mathbb{E}\left[\left|\mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big] - \sigma_{\mathrm{AP}}^2\right|\right] \to 0$. Here, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\big[Z_t^2 \mid \mathcal{S}_{t-1}\big] - \sigma_{\mathrm{AP}}^2\right|\right] \\
&= \mathbb{E}\left[\left|\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\right)^2\left(Y_t(a) - \hat{f}_{t-1}(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} + \left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 + \right.\right. \\
&\quad 2\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})f^*(a, X_t) - \sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\hat{f}_{t-1}(a, X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right) - \\
&\quad \left.\left.\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a \mid X_t)\right)^2\left(Y_t(a) - f^*(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} - \Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X_t)f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2 \mid \mathcal{S}_{t-1}\right]\right|\right] \\
&= \mathbb{E}\left[\left|\mathbb{E}\left[\mathbb{E}\left[\sum_{a=1}^{K}\frac{\left(\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\right)^2\left(Y_t(a) - \hat{f}_{t-1}(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} + \left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 + \right.\right.\right. \\
&\quad 2\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})f^*(a, X_t) - \sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\hat{f}_{t-1}(a, X_t)\right)\left(\sum_{a=1}^{K}\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right) - \\
&\quad \left.\left.\left.\sum_{a=1}^{K}\frac{\left(\pi^{\mathrm{e}}(a \mid X_t)\right)^2\left(Y_t(a) - f^*(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} - \Big(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X_t)f^*(a, X_t) - R(\pi^{\mathrm{e}})\Big)^2 \mid X_t, \mathcal{S}_{t-1}\right] \mid \mathcal{S}_{t-1}\right]\right|\right].
\end{aligned}$$

Then, by using Jensen's inequality,

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] - \sigma^2\right|\right] \\
&\leq \mathbb{E}\left[\left|\mathbb{E}\left[\left|\mathbb{E}\left[\sum_{a=1}^{K} \frac{\left(\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\right)^2 \left(Y_t(a) - \hat{f}_{t-1}(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} + \left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 + \right.\right.\right. \\
&\qquad 2\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) f^*(a, X_t) - \sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t)\right)\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right) - \\
&\qquad \left.\left.\left. \sum_{a=1}^{K} \frac{\left(\pi^{\mathrm{e}}(a \mid X_t)\right)^2 \left(Y_t(a) - f^*(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} - \left(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\right)^2 \mid X_t, \mathcal{S}_{t-1}\right] \right| \mid \mathcal{S}_{t-1}\right]\right] \\
&= \mathbb{E}\left[\left|\mathbb{E}\left[\sum_{a=1}^{K} \frac{\left(\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\right)^2 \left(Y_t(a) - \hat{f}_{t-1}(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} + \left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 + \right.\right. \\
&\qquad 2\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) f^*(a, X_t) - \sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t)\right)\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right) - \\
&\qquad \left.\left. \sum_{a=1}^{K} \frac{\left(\pi^{\mathrm{e}}(a \mid X_t)\right)^2 \left(Y_t(a) - f^*(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} - \left(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\right)^2 \mid X_t, \mathcal{S}_{t-1}\right] \right|\right].
\end{aligned}
$$

Because $\hat{f}_{t-1}$ and π_t^{e} are constructed from $\mathcal{S}_{t-1}$,

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] - \sigma^2\right|\right] \\
&\leq \mathbb{E}\left[\left|\sum_{a=1}^{K} \frac{\left(\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\right)^2 \left(Y_t(a) - \hat{f}_{t-1}(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} + \left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 + \right. \\
&\qquad 2\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) f^*(a, X_t) - \sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t)\right)\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right) - \\
&\qquad \left.\left. \sum_{a=1}^{K} \frac{\left(\pi^{\mathrm{e}}(a \mid X_t)\right)^2 \left(Y_t(a) - f^*(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} - \left(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\right)^2 \mid X_t, \hat{f}_{t-1}, \pi_t^{\mathrm{e}}\right] \right|\right].
\end{aligned}
$$

From the results of Step 1, there exist $\tilde{C}_4$ and $\tilde{C}_5$ such that

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] - \sigma^2\right|\right] \\
&\leq \mathbb{E}\left[\left|\mathbb{E}\left[\sum_{a=1}^{K} \frac{\left(\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1})\right)^2 \left(Y_t(a) - \hat{f}_{t-1}(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} + \left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right)^2 + \right.\right. \\
&\qquad 2\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) f^*(a, X_t) - \sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t)\right)\left(\sum_{a=1}^{K} \pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) \hat{f}_{t-1}(a, X_t) - R(\pi_t^{\mathrm{e}})\right) - \\
&\qquad \left.\left. \sum_{a=1}^{K} \frac{\left(\pi^{\mathrm{e}}(a \mid X_t)\right)^2 \left(Y_t(a) - f^*(a, X_t)\right)^2}{\pi^{\mathrm{b}}(a \mid X_t)} - \left(\sum_{a=1}^{K} \pi^{\mathrm{e}}(a \mid X_t) f^*(a, X_t) - R(\pi^{\mathrm{e}})\right)^2 \mid X_t, \hat{f}_{t-1}, \pi_t^{\mathrm{e}}\right] \right|\right] \\
&\leq \tilde{C}_4 \sum_{k=0}^{1} \mathbb{E}\left[\left|\pi_t^{\mathrm{e}}(a \mid X_t, \mathcal{S}_{t-1}) - \pi^{\mathrm{e}}(a \mid X_t)\right|\right] + \tilde{C}_5 \sum_{a=1}^{K} \mathbb{E}\left[\left|\hat{f}_{t-1}(a, X_t) - f^*(a, X_t)\right|\right].
\end{aligned}
$$

Then, from L^r convergence theorem, by using pointwise convergence of π_t^{e} and $\hat{f}_{t-1}$ and the boundedness of z_t, we have $\mathbb{E}\left[\left|\mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] - \sigma^2\right|\right] \to 0$. Therefore,

$$
\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] - \sigma^2\right| \geq \varepsilon\right) \leq \frac{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left|\mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] - \sigma^2\right|\right]}{\varepsilon} \to 0.
$$

As a conclusion,

$$\frac{1}{T}\sum_{t=1}^{T} Z_t^2 - \sigma_{\mathrm{AP}}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(Z_t^2 - \mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] + \mathbb{E}\left[Z_t^2 \mid \mathcal{S}_{t-1}\right] - \sigma_{\mathrm{AP}}^2\right) \xrightarrow{\mathrm{p}} 0.$$

Conclusion

We can use CLT for an MDS. Hence, we have

$$\sqrt{T}\left(\widehat{R}^{\mathrm{AP}}\left(\{\pi_t^{\mathrm{e}}\}_{t=1}^T\right) - \frac{1}{T}\sum_{t=1}^{T} R(\pi_t^{\mathrm{e}})\right) \to \mathcal{N}\left(0, \sigma_{\mathrm{AP}}^2\right),$$

where $\sigma_{\mathrm{AP}}^2 = \mathbb{E}\left[\sum_{a=1}^K \frac{\left(\pi^{\mathrm{e}}(a|X)\right)^2 \nu^*\left(a,X\right)}{\pi^{\mathrm{b}}(a|X)} + \left(\sum_{a=1}^K \pi^{\mathrm{e}}(a \mid X) f^*(a,X) - R(\pi^{\mathrm{e}})\right)^2\right]$. □

H OPE with Hypothesis Testing

Here, we consider constructing evaluation probabilities by using $\mathcal{S}_{T^{(1)}}^{(1)}$ and evaluate them by using $\mathcal{S}_{T^{(2)}}^{(2)}$. Let us assume the number of evaluation probabilities is $M = 2$, and the L OPE estimators jointly follows the asymptotic distribution with the covariance matrix Σ^L, i.e., for a L dimensional vector $I_L = (1\ 1\ \cdots\ 1)$ $\sqrt{T}\left(\widehat{\boldsymbol{R}}(\pi^{\mathrm{e}}) - I_L R(\pi^{\mathrm{e}})\right) \xrightarrow{\mathrm{d}}$ $\mathcal{N}(0,\Sigma)$, where $\widehat{\boldsymbol{R}}(\pi^{\mathrm{e}}) = \left(\widehat{R}^1(\pi^{\mathrm{e}})\ \cdots\ \widehat{R}^L(\pi^{\mathrm{e}})\right)^\top$. Then, if Σ is known, we can obtain an efficient estimator as $\widehat{R}^{\mathrm{efficient}}(\pi^{\mathrm{e}}) = \left(I^\top \Sigma^{-1} I\right)^{-1} I^\top \Sigma^{-1} \widehat{\boldsymbol{R}}(\pi^{\mathrm{e}})$, which has the variance $\sigma^2_{\mathrm{efficient}} = \left(I^\top \Sigma^{-1} I\right)^{-1}$ (Hamilton, 1994; Greene, 2003). Under these settings, we consider test the null hypothesis $R(\pi^{\mathrm{e},(1)}) = R(\pi^{\mathrm{e},(2)})$. By using an efficient OPE estimator $\widehat{R}^{\mathrm{efficient}}(\pi^{\mathrm{e},(1)}) - \widehat{R}^{\mathrm{efficient}}(\pi^{\mathrm{e},(2)})$, which has the asymptotic normality, we can conduct hypothesis testing. In general, we can estimate $R^{\mathrm{DVE}}(\pi^{\mathrm{e},(1)}, \pi^{\mathrm{e},(2)})$ by $\widehat{R}^{\mathrm{efficient}}(\pi^{\mathrm{e},(1)} - \pi^{\mathrm{e},(2)})$.

Remark 11 (Cross hypothesis testing)**.** As the cross-validation, after the above process, we can construct evaluation probabilities by using $\mathcal{S}_{T^{(1)}}^{(1)}$ and test them by using $\mathcal{S}_{T^{(2)}}^{(2)}$. However, this additional hypothesis testing cases multiple hypothesis testing and should be avoided if possible. When conducting such a process, we can correct the confidence interval by multiple testing methods such as Bonferroni correction.

I Other Discussions

Variance-regret trade-off

There is a trade-off between regret and adaptability of OPE. Let us consider the setting where there is no covariate, i.e., we can only observe (A_t, Y_t). When it is possible to perform OPE for general evaluation policies, the behavior policy cannot achieve the best order of the expected regret. Let Θ be a set of all such problems satisfying Assumption 2. For each $\theta \in \Theta$, let us define the expected regret as follows:

$$\mathrm{regret}_\theta \coloneqq \sum_{a \neq a^*} \Delta_a \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[A_t = a]\right],$$

where $\Delta_a \coloneqq \mathbb{E}\left[Y_t(a^*) - Y_t(a)\right]$ and $a^* \in \arg\max_{a\in[K]} \mathbb{E}\left[Y_t(a)\right]$. The goal of the MAB problem is to minimize the expected regret. We call an algorithm is *consistent* if for all $\theta \in \Theta$, $\mathrm{regret}_\theta = o(T^\alpha)$ for all $\alpha \in (0,1)$. Applying the OPE algorithm requires the additional assumption that, for all t and $a \in [K]$, $0 \leq \frac{\pi^{\mathrm{e}}(a)}{\pi_t^{\mathrm{b}}(a)} \leq C_1$. (Assumption 1). Therefore, if we want to conduct OPE using the samples generated from MAB algorithms, the MAB algorithms need to perform uniform exploration with some small constant $\gamma \in (0,1)$. This will cause additional γT regret. The theoretical regret lower bound for any consistent algorithm for most of the Stochastic MAB problem is known to be $\Omega(\log T)$ and there exists an algorithm to achieve $\mathcal{O}(\log T)$ regret, asymptotically Lai & Robbins (1985). It implies

Table 5: Specification of datasets

Dataset	the number of samples	Dimension	the number of classes
mnist	60,000	780	10
pendigits	7,496	16	10
sensorless	58,509	48	11
connect-4	67,557	126	3

when it is possible to perform OPE for general evaluation policies, the behavior policy cannot achieve the best order of the expected regret. Thus, there is a trade-off between the regret minimization of behavior MAB policies and OPE evaluation.

Importance of Asymptotic Normality: The asymptotic normality is an important criterion for obtaining a confidence interval and convergence rate. For obtaining the asymptotic normality, we need several conditions for OPE estimators. For example, Chernozhukov et al. (2018) devised DDM, which guarantees the asymptotic normality when using nonparametric models for estimating f^* under a mild condition. Although such techniques are incorporated in various OPE methods, it is also reported that such methods worsen the estimator in the sense of MSE between the true value and OPE estimators. Therefore, by giving up the asymptotic normality, we may increase the performance empirically.

J Details of Experiments

In this section, we describe the details of experiments. The dataset description is shown in Table 5. We show the results using sensorless and connect-4 datasets in Tables 6–8. The other information is described as follows.

Evaluation policy: All algorithms are implemented by scikit-learn, which is one of the most famous machine learning libraries in Python[4]. We solve classification problems and regard the output with softmax function as the evaluation probability. For all methods, we use cross-validation for the hyper-parameter tuning. For the SVM with RBF kernel, we select the Kernel coefficient from a set $\{0.01, 0.1, 1\}$. For the random forest, we select the max depth from a set $\{5, 10, 15, 20\}$ and the number of estimators from a set $\{10, 50, 100\}$. For all algorithms, we applied L2 regularization with a parameter chosen from $\{0.01, 0.1, 1\}$. The cross-validation is 2-fold.

Estimator of f^*: In DM LR, we use a naive linear regression. For the other methods, we estimate f^* by using the kernel ridge regression with the Kernel coefficient from a set $\{0.01, 0.1, 1\}$ and regularization parameter chosen from a set $\{0.01, 0.1, 1\}$.

OPE estimators: For AIPW, we apply 2-fold DDM for guaranteeing the asymptotic normality.

K Proof of Theorem 3

Proof. First, p^* is given as the follows:

$$p^* = \mathop{\arg\max}_{p \in \Delta^{M-1}} \min_{w \in \Delta^{L-1}} \sum_{j=1}^{L} w_j \hat{R}^j(\pi_p^{\mathrm{e}}),$$

where $\pi_p^{\mathrm{e}} = \sum_{i=1}^{M} p_i \pi^{\mathrm{e},(i)}$. Since $\hat{R}^j(\pi) = \frac{1}{T}\sum_{t=1}^{T} \langle \pi(X_t), \Gamma_{j,t} \rangle$, we have:

[4]https://scikit-learn.org/stable/

Table 6: Experimental results of OPE2D with various OPE estimators using mnist and pendigits datasets. The best and worst methods are highlighted in red and blue, respectively.

sensorless	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00221	0.00652	0.02250	0.04192	0.00924	0.01374	0.00209	0.00640	0.00620	0.01947	0.00862	0.02834	0.00871	0.02810	0.01831	0.03417
0.4	0.00108	0.00325	0.01807	0.03241	0.00362	0.00828	0.00182	0.00766	0.00655	0.02217	0.00810	0.01867	0.00801	0.01745	0.01084	0.02943
0.0	0.00000	0.00000	0.02865	0.05426	0.00128	0.00902	0.00000	0.00000	0.00302	0.02279	0.01479	0.04042	0.01446	0.03705	0.01194	0.04022

connect-4	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00111	0.00354	0.02257	0.02314	0.00069	0.00193	0.00111	0.00323	0.02257	0.02314	0.01160	0.02007	0.27933	0.31967	0.01213	0.01965
0.4	0.00169	0.00417	0.02054	0.02070	0.00072	0.00202	0.00113	0.00383	0.02054	0.02070	0.01095	0.01711	0.22867	0.29530	0.01162	0.01835
0.0	0.00174	0.00677	0.02161	0.02339	0.00087	0.00309	0.00040	0.00178	0.02161	0.02339	0.00973	0.01743	0.18617	0.26070	0.01269	0.02163

Table 7: Experimental results of ISOPE with various OPE estimators using mnist and pendigits datasets. The best and worst methods are highlighted in red and blue, respectively.

sensorless	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00399	0.01230	0.01564	0.03957	0.00036	0.00134	0.00379	0.01258	0.00550	0.02082	0.00542	0.02509	0.00644	0.02489	0.00957	0.02786
0.4	0.00698	0.01331	0.01688	0.03630	0.00116	0.00380	0.00432	0.00978	0.00527	0.01787	0.00279	0.00922	0.00402	0.01459	0.01592	0.03300
0.0	0.00000	0.00000	0.01921	0.05038	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01320	0.04256	0.01363	0.04230	0.00353	0.02480

connect-4	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.03619	0.02002	0.02581	0.02343	0.00557	0.00884	0.02599	0.02291	0.02581	0.02343	0.01553	0.02071	0.24113	0.33316	0.03176	0.02267
0.4	0.02822	0.02205	0.02189	0.02176	0.00261	0.00655	0.02156	0.02231	0.02189	0.02176	0.01412	0.01900	0.25726	0.31583	0.02365	0.02304
0.0	0.02613	0.02522	0.01922	0.02129	0.00292	0.00619	0.01854	0.02291	0.01922	0.02129	0.00974	0.01611	0.19965	0.26632	0.02596	0.02362

Table 8: Experimental results of OPCV with various OPE estimators using mnist and pendigits datasets. The best and worst methods are highlighted in red and blue, respectively.

sensorless	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.01362	0.02798	0.01814	0.02833	0.01956	0.01624	0.01459	0.03088	0.01632	0.03128	0.01952	0.03225	0.02035	0.02914	0.01386	0.02542
0.4	0.01913	0.02943	0.02450	0.03072	0.01810	0.02203	0.02249	0.03225	0.01587	0.02487	0.01848	0.02390	0.02000	0.02299	0.02133	0.02910
0.0	0.00290	0.02897	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00290	0.02897

connect-4	IPW		DM LR		DM KR		AIPW		MEAN		Minimax		Mix		Maxmax	
α	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.7	0.00270	0.00450	0.01889	0.02271	0.00295	0.00459	0.00360	0.00660	0.01889	0.02271	0.00779	0.01488	0.23013	0.31443	0.01568	0.02167
0.4	0.00837	0.00992	0.02394	0.02142	0.00836	0.01010	0.00768	0.01061	0.02394	0.02142	0.01818	0.02002	0.27391	0.30736	0.01426	0.01659
0.0	0.01630	0.02137	0.02636	0.02269	0.01704	0.02156	0.01567	0.02045	0.02636	0.02269	0.02136	0.02231	0.25140	0.27160	0.01976	0.02244

$$
\begin{aligned}
\sum_{j=1}^{L} w_j \hat{R}^j(\pi_p^{\mathrm{e}}) &= \sum_{j=1}^{L} w_j \left(\frac{1}{T} \sum_{t=1}^{T} \left\langle \sum_{i=1}^{M} p_i \pi^{\mathrm{e},(i)}(X_t), \Gamma_{j,t} \right\rangle \right) \\
&= \sum_{j=1}^{L} w_j \left(\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{M} p_i \left\langle \pi^{\mathrm{e},(i)}(X_t), \Gamma_{j,t} \right\rangle \right) \\
&= \sum_{j=1}^{L} w_j \sum_{i=1}^{M} p_i \left(\frac{1}{T} \sum_{t=1}^{T} \left\langle \pi^{\mathrm{e},(i)}(X_t), \Gamma_{j,t} \right\rangle \right) \\
&= \sum_{j=1}^{L} w_j \sum_{i=1}^{M} p_i \hat{R}^j(\pi^{\mathrm{e},(i)}) \\
&= \sum_{i=1}^{M} \sum_{j=1}^{L} p_i w_j \hat{R}^j(\pi^{\mathrm{e},(i)}).
\end{aligned}
$$

Therefore, we can rewrite p^* as follows:

$$
p^* = \operatorname*{arg\,max}_{p \in \Delta^{M-1}} \min_{w \in \Delta^{L-1}} \sum_{i=1}^{M} \sum_{j=1}^{L} p_i w_j \hat{R}^j(\pi^{\mathrm{e},(i)}).
$$

This equation implies that p^* is a Nash equilibrium strategy in the two-player zero-sum normal-form game with payoff matrix (C_{ij}) where $C_{ij} = \hat{R}^j(\pi^{\mathrm{e},(i)})$. We can rewrite p^* as the solution of the following linear problem:

$$
\begin{aligned}
&\max_{z,p} \ z \\
&\text{s.t. } p^T C \geq z\mathbf{1}^T \\
&\quad p^T\mathbf{1} = 1 \\
&\quad \forall i, p_i \geq 0.
\end{aligned}
$$

□