CONFIDENCE INTERVAL FOR OFF-POLICY EVALUATION FROM DEPENDENT SAMPLES VIA BANDIT ALGORITHM: APPROACH FROM STANDARDIZED MARTINGALES

A PREPRINT

Masahiro Kato
CyberAgent Inc.
Tokyo, Japan
masahiro_kato@cyberagent.co.jp

June 15, 2020

ABSTRACT

This study addresses the problem of off-policy evaluation (OPE) from dependent samples obtained via the bandit algorithm. The goal of OPE is to evaluate a new policy using historical data obtained from behavior policies generated by the bandit algorithm. Because the bandit algorithm updates the policy based on past observations, the samples are not independent and identically distributed (i.i.d.). However, several existing methods for OPE do not take this issue into account and are based on the assumption that samples are i.i.d. In this study, we address this problem by constructing an estimator from a standardized martingale difference sequence. To standardize the sequence, we consider using evaluation data or sample splitting with two-step estimation. This technique produces an estimator with asymptotic normality without restricting a class of behavior policies. In an experiment, the proposed estimator performs better than existing methods, which assume that the behavior policy converges to a time-invariant policy.

1 Introduction

The multi-armed bandit (MAB) problem is one of the sequential decision-making problems, which is applied in various applications, such as ad-optimization, personalized medicine, search engines, and recommendation systems. In the MAB problem, we choose an action at a period and observe the outcome. To reduce the cost of searching for better policies, the evaluation of a new policy using historical data obtained from past trials has gathered great attention (Beygelzimer & Langford, 2009; Li et al., 2010; Athey & Wager, 2017). This framework is called off-policy evaluation (OPE) (Dudík et al., 2011; Wang et al., 2017; Narita et al., 2019; Bibaut et al., 2019; Kallus & Uehara, 2019a; Oberst & Sontag, 2019). Although several methods for OPE have been proposed, existing studies often presume that the samples are independent and identically distributed (i.i.d.). However, in the MAB problem algorithm, the policy is usually updated based on past observations, and the samples are not i.i.d. due to the policy updating process. In this case, the consistency and asymptotic normality of the existing methods are not guaranteed. In particular, the asymptotic normality is critical because it guarantees the $\sqrt{n}$-consistency and is needed for the confidence interval in hypothesis testing to determine whether the new policy is better than the existing policy. Thus, the motivation for establishing a novel method for performing OPE from dependent samples is strong.

Several previous studies consider OPE from dependent samples (der Laan & Mark, 2008; Hadad et al., 2019; Kallus & Uehara, 2019b; Kato et al., 2020a). The strategies for deriving asymptotic normality can be classified into three approaches. In the first approach, under the assumption that the policy used in past trials converges to the time-invariant policy in probability, the asymptotic normality is derived using the theories related to martingales (der Laan & Mark, 2008; Hadad et al., 2019; Kato et al., 2020a). In the second approach, the policy is assumed to be batch updated, where, although the policy is updated using past observations, the sample size under a fixed policy is sufficient (Hahn et al., 2011). In the third approach, both the stationarity and the conditions of mixingales (Kosorok, 2008), which requires the independence of time-separated samples, are assumed (Kallus & Uehara, 2019b). However,
existing methods have several drawbacks. With the first approach, the assumption that the policy converges to a time-invariant policy sometimes does not hold. For example, if the algorithms of the MAB problem change during past trials, it is difficult to justify this assumption. The second and third approaches cannot be applied in cases where the policy is sequentially updated, and the conditions of mixingales do not hold, respectively.

To overcome these drawbacks, in this study, we propose a novel estimator with the asymptotic normality, which does not assume policy convergence, the existence of batches, and the independence of time-separated samples. To accomplish this goal, we refocus on the first approach. In the first approach, we use a martingale difference sequence (MDS) for constructing an estimator. In this approach, the convergence of the policy used in past trials to the time-invariant policy in probability is assumed for using central limit theorem (CLT) of an MDS, which requires that the variance of an MDS becomes asymptotically constant. In this paper, instead of assuming the convergence of the policy, we consider constructing an MDS with asymptotically constant variance by standardization using an estimator of the variance. To implement the standardization, we assume access to evaluation (test) data or conduct sample splitting. Through this technique, we can guarantee the asymptotic normality of the proposed estimator with fewer assumptions.

Contributions and Organization of this Paper: This paper has three main contributions. First, this paper provides a theoretical solution for causal inference from dependent samples obtained via bandit feedback, which is more realistic than the situation that has been often considered in the existing work. Second, the proposed estimator achieves the asymptotic normality with fewer assumptions compared with existing solutions. Third, the estimator also achieves a lower mean squared error (MSE) in experiments using benchmark datasets.

Additional Related Work: There are various studies for OPE under the assumption that samples are i.i.d. (Dudík et al., 2011; Wang et al., 2017; Narita et al., 2019; Bibaut et al., 2019; Kallus & Uehara, 2019; Oberst & Sontag, 2019), but there are fewer studies focus on the case in which samples are not i.i.d. When the policy converges, der Laan & Mark (2008), Hadad et al. (2019), and Kato et al. (2020a) proposed using an estimator based on an MDS. When including the covariates for OPE, der Laan & Mark (2008) mainly suggested using targeted maximum likelihood estimation. On the other hand, Hadad et al. (2019) and Kato et al. (2020a) proposed using an estimator based on the doubly robust or augmented IPW estimator. Kallus & Uehara (2019) dealt with the application of reinforcement learning and proposed an estimator constructed from dependent samples. In their study, they proposed time cross-fitting, which is a variant of cross-fitting of double/debiased machine learning (Chernozhukov et al., 2018). The theoretical guarantee is based on mixingale theory. For obtaining the asymptotic normality, conducting the standardization is also proposed by Luedtke & van der Laan (2016). Luedtke & van der Laan (2016) tries to solve a similar problem in the batch policy updating. In addition to this difference of policy updating, there are the following two points different from us. First, in their solution, they require estimators of $f^*$ and $e^*$ to satisfy some conditions like the Donsker condition. However, in our estimator, we do not impose such restrictions on estimators of $f^*$ and $e^*$ except for the boundedness. This difference is critical because, in practice, as Chernozhukov et al. (2018) pointed out, the estimators of $f^*$ and $e^*$ usually do not satisfy the Donsker condition when we use modern methods such as the Lasso and Random forest regression. Second, for obtaining an estimator with the asymptotic normality under sequential policy updating, we propose using evaluation data or sample splitting.

2 Problem Setting

In this section, we introduce our problem setting.

2.1 Data Generating Process

Let $A_t$ be the action taking variable in $A = \{1, 2, \ldots, K\}$, $X_t$ be the covariate observed by the decision maker when choosing an action, and $\mathcal{X}$ be the domain of covariate. Let us denote a random variable of a reward at a period $t$, and a historical data $\mathcal{S}_T = \{(X_t, A_t, Y_t)\}_{t=1}^T$, with the following data generating process (DGP):

$$\{(X_t, A_t, Y_t)\}_{t=1}^T \sim p(x)p_t(a | x, \Omega_{t-1})p(y | a, x),$$

where $Y_t = \sum_{a=1}^K \mathbb{1}[A_t = a]Y_t(a), p(x)$ denote the density of the covariate $X_t$, $p_t(a | x, \Omega_{t-1})$ denote the probability of assigning an action $A_t$ conditioned on a covariate $X_t$, $p(y | a, x)$ denote the density of an outcome $Y_t$ conditioned on $A_t$ and $X_t$, and $\Omega_{t-1} \in \mathcal{M}_{t-1}$ denotes the history defined as $\Omega_{t-1} = \{X_{t-1}, A_{t-1}, Y_{t-1}, \ldots, X_1, A_1, Y_1\}$ with the domain $\mathcal{M}_{t-1}$. We assume that $p(x)$ and $p(y | a, x)$ are invariant across periods, but $p_t(a | x, \Omega_{t-1})$ can take different value across periods. Besides, we allow the decision maker to change $p_t(a | x, \Omega_{t-1})$ based on past observations. In this case, samples $\{(X_t, A_t, Y_t)\}_{t=1}^T$ are correlated across periods, i.e., samples are not i.i.d. Here, we introduce
behavior policies, which determine the probability \( p_t(a \mid x, \Omega_{t-1}) \). Let a behavior policy \( \pi_t : \mathcal{A} \times \mathcal{X} \times \mathcal{M}_{t-1} \to (0, 1) \) in the \( t \)-th period be a function of a covariate \( X_t \), an action \( A_t \), and a history \( \Omega_{t-1} \). In this study, we assume that 
\[
\pi_t(a \mid x, \Omega_{t-1}) = p_t(a \mid x, \Omega_{t-1}).
\]

**Remark 1 (Evaluation Data).** In Section 4.3 we also assume that there is access to evaluation data, \( \mathcal{E}_N = \{X_t\}_{t=1}^N \). In Remark 7 we relax this assumption by introducing the sample splitting.

### 2.2 Off-Policy Evaluation

Under the DGP defined in the previous subsection, we consider estimating the value of an evaluation policy using samples obtained under the behavior policies. Let an evaluation policy \( \pi^* : \mathcal{X} \to (0, 1) \) be a function of a covariate \( X_t \) and an action \( A_t \); this evaluation policy can be considered as the probability of taking the action \( A_t \) conditional on the covariate \( X_t \). We are interested in estimating the expected reward from a given pre-specified evaluation policy \( \pi^*(a \mid x) \). Then, we define the expected reward under \( \pi^* \) as 
\[
R(\pi^*) := \mathbb{E} \left[ \sum_{t=1}^K \pi^*(a, X_t)Y_t(a) \right].
\]
We also denote \( R(\pi^*) \) as \( \theta_0 \). The goal of this study is to estimate \( R(\pi^*) \) using dependent samples under a sequentially-updated policy. To identify \( R(\pi^*) \), we make the following assumptions on the policies and the outcomes.

**Assumption 1.** The behavior policies \( \pi_t(a \mid \cdot, \Omega_{t-1}) \) and the evaluation policy \( \pi^*(a \mid \cdot) \) are known. Additionally, the evaluation policy is deterministic.

**Assumption 2.** There exist \( C_1 \) and \( C_2 \) such that
\[
\frac{\pi^*(a \mid x)}{\pi_t(a \mid x, \Omega_{t-1})} \leq C_1 \quad \text{and} \quad |Y_t| \leq C_2.
\]

The deterministic evaluation policy looks restrictive. However, the optimal policy that maximizes the expected reward is a deterministic policy when the policy is not restricted. Therefore, when we find the optimal policy from a set of policies, such as off-policy learning (Zhao et al. 2012; Kitagawa & Tetenov 2018; Zhou et al. 2018; Chernozhukov et al. 2019), this assumption can be accommodated. Thus, restricting the policy class to the deterministic policy class would be reasonable in practice. In fact, most of the existing methods follow this way.

**Remark 2 (Existing Methods for OPE).** We review three types of standard estimators of \( R(\pi^*) \) under the case where \( p_t(a \mid x) = p_2(a \mid x) = \cdots = p_T(a \mid x) = p(a \mid x) \) in the DGP defined in (1). The first estimator is an inverse probability weighting (IPW) estimator given by
\[
\frac{1}{T} \sum_{t=1}^{T} \sum_{a=1}^{K} \frac{\pi(a | X_t)}{p(a | X_t)} Y_t(a).
\]
This estimator is unbiased when the behavior policy is known, it has a high variance. The second estimator is a direct method (DM) estimator
\[
\frac{1}{T} \sum_{t=1}^{T} \sum_{a=1}^{K} \frac{\pi(a | X_t)}{p(a | X_t)} Y_t(a) \cdot f_t(a, X_t),
\]
where \( f_t(a, X_t) \) is an estimator of \( f^*(a, X_t) \) (Hainz 1998). This estimator is known to be weak against model misspecification for \( f^*(a, X_t) \). The third estimator is a doubly robust estimator defined as
\[
\frac{1}{T} \sum_{t=1}^{T} \sum_{a=1}^{K} \left( \frac{\pi(a | X_t)}{p(a | X_t)} Y_t - \hat{f}_t(a, X_t) \right) + \pi^*(a \mid X_t) \hat{f}_t(a, X_t)
\]
(Robins et al. 1994; Chernozhukov et al. 2018). Under certain conditions, it is known that this estimator achieves the lower 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).

**Remark 3 (Unconfoundedness).** Existing methods often make an assumption called unconfoundedness, that is, they assume that the outcomes \( Y_t(0), Y_t(1), \ldots, Y_t(K) \) and the action \( A_t \) are conditionally independent on \( X_t \) for identification. In the DGP, we can obtain a similar result from the assumption that an action \( A_t \) is chosen according to probability \( \pi_t(a \mid x, \Omega_{t-1}) \).

**Remark 4 (Semiparametric Lower Bound).** The lower bound of the variance can be defined for an estimator under some posited models of the DGP. If this posited model is a parametric model, then the lower bound is equal to the Cramér-Rao lower bound. When this posited model is a non- or semiparametric model, the corresponding lower bound can still be defined (Bickel et al. 1998). Narita et al. (2019) shows that the semiparametric lower bound of the DGP (1) under \( p_t(a \mid x, \Omega_0) = p_2(a \mid x, \Omega_1) = \cdots = p_T(a \mid x, \Omega_{T-1}) = p(a \mid x) \) is
\[
\mathbb{E} \left[ \sum_{a=1}^{K} \left( \frac{\pi(a | X_t)}{p(a | X_t)} f^*(a, X_t) - \theta_0 \right)^2 \right].
\]

**Notations:** Let \( a \) be an action in \( \mathcal{A} \). We denote \( \mathbb{E}[Y_t(a) \mid x], \mathbb{E}[Y_t^2(a) \mid x], \) and \( \text{Var}(Y_t(a) \mid x) \) as \( f^*(a, x), e^*(a, x), \) and \( v^*(a, x) \), respectively. Let \( F \) be the domain of \( f^*(a, x) \). Let \( f_t(a, x), e_t(a, x), \) and \( \theta_t \) be estimators of \( f^*(a, x), e^*(a, x), \) and \( \theta_0 \) constructed from \( \Omega_t \), respectively. Let \( \mathcal{N}(\mu, \text{var}) \) be the normal distribution with mean \( \mu \) and variance \( \text{var} \).

1The proposed method requires the function of the polices, not only \( \pi_t(a \mid X_t, \Omega_{t-1}) \) for a specific \( X_t \).
2.3 Policy Updating

In the DGP defined in Section 2.1, we allow the behavior policy \( \pi_t(a \mid x, \Omega_{t-1}) \) to be updated based on past observations. For example, when applying some algorithms proposed in the MAB problem, we usually optimize the probability to maximize the cumulative reward using past observations. When the behavior policy \( \pi_t(a \mid x, \Omega_{t-1}) \) depends on past observations, the samples are not i.i.d. In this case, we cannot apply the standard methods for OPE, which assume that the samples are i.i.d. To clarify this problem, we classify the patterns of policy updating into two scenarios: sequential policy updating and batch policy updating. In sequential policy updating, the policy can be updated each period. In batch policy updating, the policy is updated after observing some samples (Hahn et al., 2011; Narita et al., 2019). In this study, we consider a problem using sequential policy updating.

3 OPE under a Converging Policy

In this section, we review OPE when a policy converges to the time-invariant policy in probability. As der Laan & Mark (2008), Hadad et al. (2019) and Kato et al. (2020a) proposed, in such a situation, we can construct an estimator with the asymptotic normality by using an MDS (Hall et al., 2014). To conduct OPE when the behavior policy is known, we can use the following score function \( \psi : \Theta \times X \times A \times R \times \Pi_t \times \Pi_t \times F \to R \) (Chernozhukov et al., 2018):

\[
\psi_t(\theta; x, k, y, \pi_t, \pi^e, f) = \sum_{a=1}^{K} \left\{ \pi^e(a \mid x) \mathbb{I}[A_t = a] \left\{ y - f(a, x) \right\} + \pi^e(a \mid f(a, x)) \right\} - \theta,
\]

where \( \Theta, \Pi_t, \Pi \) are spaces of a parameter \( \theta \), a behavior policies \( \pi_t \), and an evaluation policy \( \pi^e \). Then, for the true policy value \( \theta_0 \), the score function with \( \theta = \theta_0 \), \( \psi_t(\theta_0; X_t, A_t, Y_t, \pi_t, \pi^e, f_{t-1}) \), satisfies \( \mathbb{E}[\psi_t(\theta_0; X_t, A_t, Y_t, \pi_t, \pi^e, f_{t-1})] = 0 \). As we explain in Remark 2, there are several other candidate estimators, including the IPW estimator. However, the estimator based on the score function defined in (2) achieves the lowest asymptotic variance when \( \pi_t \) is fixed among periods (Bickel et al., 1998; Scharfstein et al., 1999). Using \( \{\psi_t(\theta_0; X_t, A_t, Y_t, \pi_t, \pi^e, f_{t-1})\}_{t=1}^{T} \), we consider constructing an MDS \( \{z_t\}_{t=1}^{T} \) that satisfies the conditions of the following CLT for an MDS.

**Proposition 1.** [CLT for an MDS, Hamilton (1994), Proposition 7.9, p. 194] Let \( \{z_t\}_{t=1}^{\infty} \) be an MDS. Suppose that (a) \( \mathbb{E}[z_t^2] = \nu^2 > 0 \) with \( \left(1/T\right) \sum_{t=1}^{T} \nu_t^2 \to \nu^2 > 0 \); (b) \( \mathbb{E}[|z_t|^r] < \infty \) for some \( r > 2 \); (c) \( \left(1/T\right) \sum_{t=1}^{T} z_t^2 \overset{p}{\to} \nu^2 \).

Then \( \sqrt{T} \sum_{t=1}^{T} z_t \overset{d}{\to} N(0, \nu^2) \).

When \( Y_t, \pi^e/\pi_t, \) and \( f_{t-1} \) are bounded, we can easily show that the condition (b) holds (Assumption 2). The remaining task is to show that conditions (a) and (c) hold.

Kato et al. (2020a) constructed \( \hat{z}_t = \psi_t(\theta_0; X_t, A_t, Y_t, \pi_t, \pi^e, f_{t-1}) \), where \( \{\hat{z}_t\}_{t=1}^{T} \) is an MDS. Then, by assuming \( \pi_t(a \mid x, \Omega_{t-1}) \overset{p}{\to} \alpha(a \mid x) \) and some regularity conditions, Kato et al. (2020a) proved that the sequence \( \{\hat{z}_t\}_{t=1}^{T} \) satisfies the conditions (a) and (c) with \( \nu^2 = \mathbb{E}\left[\left(\sum_{a=1}^{K} \frac{\pi^e(a \mid x) \mathbb{I}[A_t = a] Y_t f_{t-1}(a, X_t)}{\alpha(a \mid x)} + \pi^e(a \mid f_{t-1}(a, X_t)) - \theta_0\right)^2\right] \). Thus, by the assumption \( \pi_t(a \mid x, \Omega_{t-1}) \overset{p}{\to} \alpha(a \mid x) \), the variance of \( h_t \) converges to time-invariant value, and this property is the motivation of the assumption proposed by Kato et al. (2020a). Then, under \( \pi_t(a \mid x, \Omega_{t-1}) \overset{p}{\to} \alpha(a \mid x) \), the estimator proposed defined as \( \hat{\theta}_T^{AIPW} = \frac{1}{T} \sum_{t=1}^{T} \sum_{a=1}^{K} \frac{\pi^e(a \mid x) \mathbb{I}[A_t = a] (Y_t - f_{t-1}(a, X_t))}{\pi(a \mid x, \Omega_{t-1})} + \pi^e(a \mid x) f_{t-1}(a, X_t) \) can be proved to be consistent and has the asymptotic normality (Hadad et al., 2019; Kato et al., 2020a). We refer to this estimator as Adaptive AIPW (A2IPW). We can also define the corresponding IPW estimator and refer to it as adaptive IPW (AdaIPW) estimator.

**Remark 5** (Efficiency of A2IPW). The asymptotic variance of the A2IPW estimator matches the semiparametric lower bound under a time-invariant policy \( \alpha \) (Kato et al., 2020a).
4 OPE from Dependent Samples under Sequential Policy Updating

In this section, we discuss a strategy for constructing an estimator without assuming the convergence of the policy. Then, we show a method for constructing an estimator with the asymptotic normality.

4.1 Strategy for OPE from Dependent Samples under Sequential Policy Updating

However, there are various applications in which we cannot assume that \( \pi_t(a \mid x, \Omega_{t-1}) \overset{D}{\rightarrow} \alpha(a \mid x) \). Therefore, we have a strong motivation for constructing an estimator with asymptotic normality without assuming that \( \pi_t(a \mid x, \Omega_{t-1}) \overset{D}{\rightarrow} \alpha(a \mid x) \). Our strategy is to construct an MDS \( \{\hat{z}_t\}_{t=1}^T \) satisfying conditions (a) and (c) by standardization of \( \hat{z}_t \) as \( \hat{z}_t = \left( \frac{\text{Var}(\hat{z}_t \mid \Omega_{t-1})}{\text{Var}(\hat{z}_t \mid \Omega_{t-1})} \right)^{-1} \hat{z}_t \), where \( \text{Var}(\hat{z}_t \mid \Omega_{t-1}) \) is the conditional variance of \( \hat{z}_t \). The standardization means that the variance of \( \hat{z}_t \) becomes 1. Then, we can easily show that the conditions (a) and (b) hold. Thus, if we know the variance \( \text{Var}(\hat{z}_t \mid \Omega_{t-1}) \), we can construct an MDS with constant variance by the standardization.

Remark 6 (Lindeberg Condition). The proposed strategy can be interpreted using the Lindeberg condition. For all \( \epsilon > 0 \) and an MDS \( \{z_t\}_{t=1}^T \) such that \( \sum_{t=1}^T E[z_i^2] / \Omega_{t-1} / \sum_{t=1}^T E[z_i^2] \overset{D}{\rightarrow} 1 \), the Lindeberg condition is\[
\left( \sum_{t=1}^T E[z_i^2] / \Omega_{t-1} / \sum_{t=1}^T E[z_i^2] \right)^{-1} \sum_{t=1}^T E[z_i^2 \mathbb{1}_{|z_i| \geq \epsilon \sqrt{\sum_{t=1}^T E[z_i^2]}}] \overset{D}{\rightarrow} 0 \text{ as } T \rightarrow \infty \text{ (Brown 1971)}.
\]The first condition \( \sum_{t=1}^T E[z_i^2] / \Omega_{t-1} / \sum_{t=1}^T E[z_i^2] \overset{D}{\rightarrow} 1 \) restricts the class of an MDS to an MDS with asymptotically time-invariant variance. Here, we note that \( \sigma^2 = E[z_i^2] / \Omega_{t-1} \). The strategy of this paper can be intuitively interpreted as a transformation of the original sequence \( z_t \) into such an MDS with asymptotically time-invariant variance.

4.2 OPE with a Weighted Average Estimator: Case with Known Variance

Then, we propose an estimator based on the MDS \( \{\hat{z}_t\}_{t=1}^T \) and refer to the estimator as a weighted Average Adaptive Augmented IPW (A3IPW) estimator. For \( \{\hat{f}_t\}_{t=0}^T \) and \( \sigma_t^2 = \text{Var}(\hat{z}_t \mid \Omega_{t-1}) \), we define the A3IPW estimator as
\[
\hat{\theta}_{T}^{A3IPW} = \left( \sum_{t=1}^T \frac{1}{\sigma_t^2} \right)^{-1} \sum_{t=1}^T \frac{1}{\sigma_t^2} \sum_{a=1}^K \left\{ \frac{\pi^*(a \mid X_t) \mathbb{1}_{[A_t=a]} \{Y_t - f_{t-1}(a, X_t)\}}{\pi_t(a \mid X_t, \Theta_{t-1})} + \pi^*(a \mid x) \hat{f}_{t-1}(a, X_t) \right\}.
\]
Then, we can show that the estimator has the asymptotic normality as follows.

Theorem 1. Suppose that there exists \( C_2 \) such that \( |\hat{f}_t| \leq C_2 \). Then, under Assumption \( \hat{\theta}_{T}^{A3IPW} - \theta_0 \overset{D}{\rightarrow} N(0, 1) \) as \( T \rightarrow \infty \).

Proof. For \( \hat{z}_t = \frac{1}{\sigma_t^2} \left( \sum_{a=1}^K \pi^*(a \mid X_t) \mathbb{1}_{[A_t=a]} \{Y_t - f_{t-1}(a, X_t)\} / \pi_t(a \mid X_t, \Theta_{t-1}) + \pi^*(a \mid x) \hat{f}_{t-1}(a, X_t) - \theta_0 \right) \), \( \{\hat{z}_t\}_{t=1}^T \) is an MDS because \( E[\hat{z}_t \mid \Omega_{t-1}] = \frac{1}{\sigma_t^2} E[\hat{z}_t \mid \Omega_{t-1}] = 0 \). Because \( \text{Var}(\hat{z}_t / \sigma_t^2) = E[(\hat{z}_t - E[\hat{z}_t])^2 \mid \Omega_{t-1} / \sigma_t^2] = E[E[\hat{z}_t^2 \mid \Omega_{t-1}] / \sigma_t^2] = E[\text{Var}(\hat{z}_t \mid \Omega_{t-1}) / \sigma_t^2] = 1 \), we have \( (1/T) \sum_{t=1}^T \text{Var}(\hat{z}_t) = 1 \). Here, we used \( E[\hat{z}_t \mid \Omega_{t-1}] = E[\hat{z}_t] = 0 \). Besides, we can show that \( \hat{z}_t \) is bounded from assumptions, and \( (1/T) \sum_{t=1}^T \hat{z}_t^2 \overset{P}{\rightarrow} 1 \) from the weak law of large numbers of Proposition 5 in Appendix A, where we apply CLT for an MDS.

4.3 OPE with a Weighted Average Estimator: Case with Unknown Variance

In general, the variance \( \sigma_t^2 \) is unknown and needed to be replaced with an estimator. However, we cannot use the samples used for constructing \( \pi_t \) and \( f_{t-1} \) to estimate \( \sigma_t^2 \) due to the dependency. For this reason, we assume access to a dataset only with covariates, \( \mathcal{E}_N = \{X_i\}_{i=1}^N \), which is independent from the historical data \( \mathcal{S}_T \). Because \( \mathcal{E}_N \) is independent from \( \mathcal{S}_T = \{ (X_t, A_t, Y_t) \}_{t=1}^T \), we can approximate \( \sigma_t^2 \) by applying the law of large numbers to the sample average \( \hat{g}_{t,N}^{\epsilon} \), defined as
\[
\frac{1}{N} \sum_{i=1}^N \sum_{a=1}^K \left\{ \frac{\pi^*(a \mid X_i)^2 (\hat{z}_{t-1} - f_{t-1}(a, X_i))}{\pi_t(a \mid X_i, \Omega_{t-1})} + \left( \frac{\pi^*(a \mid X_i) \hat{f}_{t-1}(a, X_i) - \theta_{t-1}}{\sigma_t^2} \right)^2 \right\}.
\]

The consistency of \( \hat{g}_{t,N}^{\epsilon} \) to the variance of \( \psi_t(\theta_0; X_t, A_t, Y_t, \pi_t, \pi^*, f^*) \) is shown as follows.
Theorem 2. Suppose that, for all \( x \in \mathcal{X} \) and \( a \in \mathcal{A} \), \( \hat{f}_{t-1}(a, x) - f^*(a, x) \xrightarrow{p} 0 \), \( \hat{e}_{t-1}(a, x) - e^*(a, x) \xrightarrow{p} 0 \) and \( \theta_t \xrightarrow{p} \theta_0 \) as \( t \to \infty \). Then, for \( \pi_t \in \Pi \), we have \( g'_{t,(1-r)T} \xrightarrow{p} 0 \) as \( t \to \infty \) and \( N \to \infty \), where

\[
\sigma_t^2 = \mathbb{E} \left[ \sum_{a=1}^{K} \left\{ \pi'(a|X_t) (\hat{e}_{t-1}(a, X_t) - e^*(a, X_t))^2 \right\} \right].
\]

In practice, we need to avoid having the estimator of \( \sigma_t^2 \) be less than or equal to 0. Thus, instead of \( g'_{t,(1-r)T} \), we use the estimator \( g_{t,N} \), defined as \( g_{t,(1-r)T} = \max \{ g_{t,N}, \epsilon \} \), where \( \epsilon > 0 \) is the lower bound of the variance \( \sigma_t^2 \). The estimator \( g_{t,N} \) is consistent, as is \( g_{t,N} \). The value \( \epsilon \) is assumed to be known, and we can use any sufficiently small value for \( \epsilon \) because it is simply a technical term. The proof is shown in Appendix [B]. Then, using \( g_{t,N} \), we define the Feasible A3IPW (FA3IPW) estimator \( \hat{\theta}_T^{FA3IPW} \) as

\[
\left( \sum_{t=1}^{T} \frac{1}{\sqrt{g_{t,N}}} \right)^{-1} \sum_{t=1}^{T} \sum_{a=1}^{K} \frac{1}{\sqrt{g_{t,N}}} \left\{ \frac{\pi'(a|X_t)(\hat{e}_{t-1}(a, X_t) - f_{t-1}(a, X_t))^2}{\pi(a|X_t, \Omega_{t-1})} + \pi'(a|X_t)f_{t-1}(a, X_t) - \theta_0 \right\}.
\]

Similarly, we can also define the corresponding IPW estimator and refer to it as Feasible weighted Average Adaptive IPW estimator (FA2dIPW).

In some applications, we can obtain the dataset \( \mathcal{E}_N = \{ X_t \}_{t=1}^{N} \) as evaluation (test) data, which is the target of a new policy (Kato et al., 2020b). On the other hand, by introducing a sample splitting, we can relax the assumption as the following remark.

Remark 7 (Sample Splitting). To relax the access to \( \mathcal{E}_N \), we consider splitting \( \{ (X_t, A_t, Y_t) \}_{t=1}^{T} \) into \( \{ (X_t, A_t, Y_t) \}_{t=1}^{[rT]} \) and \( \{ (X_t, A_t, Y_t) \}_{t=1}^{[rT]+1} \), where \( 0 < r < 1 \). In the following discussion, for simplicity, we assume that \( rT \) is an integer (i.e., \( [rT] = rT \)). The samples \( \{ X_t \}_{t=1}^{[rT]} \) and \( \{ X_t \}_{t=1}^{[rT]+1} \) are independent. Thus, to approximate \( \sigma_t^2 \), we can apply the law of large numbers to the sample average \( g_{t,(1-r)T} = \max \{ g'_{t,(1-r)T}, \epsilon \} \), where \( g_{t,(1-r)T} \) is defined as

\[
\frac{1}{(1-r)T} \sum_{s=rT}^{T} \sum_{a=1}^{K} \left\{ \frac{\pi'(a|X_s)(\hat{e}_{t-1}(a, X_s) - f_{t-1}(a, X_s))^2}{\pi(a|X_s, \Omega_{t-1})} + \pi'(a|X_s)f_{t-1}(a, X_s) - \theta_0 \right\}.
\]

More details on sample splitting are provided in Appendix [D].

4.4 Asymptotic Properties of a FA3IPW Estimator

Let \( \hat{\epsilon}_t = \sum_{t=1}^{T} \frac{1}{\sqrt{g_{t,N}}} \left( \sum_{a=1}^{K} \frac{\pi'(a|X_t)(\hat{e}_{t-1}(a, X_t) - f_{t-1}(a, X_t))^2}{\pi(a|X_t, \Omega_{t-1})} + \pi'(a|X_t)f_{t-1}(a, X_t) - \theta_0 \right) \). We can show that \( \hat{\epsilon}_t \) is an MDS as \( \mathbb{E} [\hat{\epsilon}_t | \Omega_{t-1}] = \frac{1}{\sqrt{g_{t,N}}} \mathbb{E} [\epsilon_t | \Omega_{t-1}] = 0 \). Based on this property, the following theorems give us the consistency and asymptotic normality of \( \hat{\theta}_T^{FA3IPW} \). First, we show the consistency of \( \hat{\theta}_T^{FA3IPW} \), i.e., \( \hat{\theta}_T^{FA3IPW} \xrightarrow{d} \theta_0 \).

Theorem 3 (Consistency of the Proposed Estimator). Suppose that there exist \( C_2 \) and \( C_3 \) such that \( |\hat{f}_t| \leq C_2 \) and \( 0 < 1/\sqrt{g_{t,N}} \leq C_3 \). Then, under Assumptions \( \hat{\theta}_T^{FA3IPW} \xrightarrow{p} \theta_0 \).

Proof. From the boundedness, we can apply the law of large numbers for an MDS (Proposition 5 in Appendix [A]). Therefore, \( \hat{\epsilon}_t \xrightarrow{p} 0 \).

The following theorem shows the asymptotic normality of \( \hat{\theta}_T^{FA3IPW} \). The proof is in Appendix [C].

Theorem 4 (Asymptotic Distribution of the Proposed Estimator). Suppose that

(i) there exist \( C_2 \) and \( C_3 \) such that \( |\hat{f}_t| \leq C_2 \) and \( 0 < 1/\sqrt{g_{t,N}} \leq C_3 \);

(ii) Pointwise convergence in probability of \( \hat{f}_t \), i.e., \( \forall x \in \mathcal{X}, \hat{f}_t(a, x) - f^*(a, x) \xrightarrow{p} 0 \) as \( t \to \infty \);

(iii) \( g_{t,N} - \sigma_t^2 \xrightarrow{p} 0 \) as \( t \to \infty \) and \( N \to \infty \).

Then, under Assumptions \( \hat{\theta}_T^{FA3IPW} \xrightarrow{d} \mathcal{N} (0, 1) \) as \( T, N \to \infty \).
Algorithm 1 TSFA3IPW Estimation

For $t = 1, 2, \ldots, T$, let $q_t = \sum_{a=1}^{K} \left\{ \pi^*(a|X_t)1(A_t=a) \left\{ Y_t-f_{t-1}(a,X_t) \right\} + \pi^*(a|X_t)f_{t-1}(a,X_t) \right\}$.

**Initialization:** Let $\tilde{g}_t$ be 1 or some other positive value for $t = 1, 2, \ldots, T - 1$.

Obtain $\{\tilde{\theta}_t\}_{t=1}^{T-1}$ using $\{\tilde{g}_t\}_{t=1}^{T-1}$ by $\tilde{\theta}_t = \left( \sum_{s=1}^{t} \frac{1}{\sqrt{g_s}} \right)^{-1} \sum_{s=1}^{t} \frac{1}{\sqrt{g_s}} \tilde{g}_s$.

Obtain $\{g_{t,N}\}_{t=1}^{T}$ using $\{\tilde{\theta}_t\}_{t=1}^{T-1}$, an appropriate value $\text{\hat{g}}^{(1)}$ such as $\text{\hat{g}}^{(1)} = 0$, and $\bar{E}_N$.

Obtain $\bar{\theta}_T^{TSFA3IPW}$ using $\{g_{t,N}\}_{t=1}^{T}$ by $\bar{\theta}_T^{TSFA3IPW} = \left( \sum_{s=1}^{T} \frac{1}{\sqrt{g_{s,N}}} \right)^{-1} \sum_{s=1}^{T} \frac{1}{\sqrt{g_{s,N}}} q_s$.

**Remark 8** (Heuristic Stabilization). The proposed estimator requires a sequence of $\{f_t\}_{t=0}^{T-1}$. Because the early estimator of $f_t$ may be unstable due to the insufficient amount of training data, some heuristic stabilizations may improve its performance. Hadad et al. (2019) proposed a similar estimator and suggested using a weight sequence for stabilization. In the setting of Kato et al. (2020a), they can change $p_t(a|X_t)$ and proposed adjusting the probability to stabilize their proposed estimator. In this study, we take a simple approach by suggesting discarding some samples at first and constructing the proposed estimator without them. We call this estimator the stabilized FA3IPW (SFA3IPW) estimator. In some experiments, this approach improves the estimator’s performance.

**Remark 9** (Difference of Assumption between the Consistency and Asymptotic Normality). The asymptotic normality requires $g_{t,N} - \sigma_t^2 \xrightarrow{\mathbb{P}} 0$. However, we can obtain the consistency $\bar{\theta}_T^{A3IPW} \xrightarrow{\mathbb{P}} \theta_0$ without assuming $g_{t,N} - \sigma_t^2 \xrightarrow{\mathbb{P}} 0$. From this property, we can obtain the consistent estimator of the variance, $\sigma_t^2$, which requires the consistent estimator of the policy value, $\theta_0$.

4.5 Efficiency of A3IPW Estimator

In the problem setting, it is not easy to discuss an efficiency of an estimator. For simplicity, we consider the lower bound of the variance of an estimator within a class of estimators such that $\bar{\theta}_T = \sum_{t=1}^{T} w_t \bar{\psi}_t(X_t, A_t, Y_t, \pi_t)$, where $\bar{\psi}$ is a score function and $w_t$ is a weight satisfying $\sum_{t=1}^{T} w_t = 1$. Let us note that we usually use $w_t = 1/T$ in the case where samples are i.i.d. First, we consider the variance of $\bar{\psi}_t$. In OPE, the lowest variance is given as $\mathbb{E} \left[ \sum_{t=1}^{K} \left\{ \frac{\pi^*(a|X_t)}{\pi_t(a|X_t)} w^*(a,X_t) \right\}^2 \right]$. In the proposed estimator, each element of $\{\tilde{\psi}_t\}_{t=1}^{T-1}$ achieves this variance. In addition, it is well known that for a sequence of positive values $\phi_t = \mathbb{E}[\bar{\psi}_t^2 | \Omega_{t-1}]$, the weight $w_t = \frac{1}{\sqrt{\phi_t}} / \sum_{t=1}^{T} \frac{1}{\sqrt{\phi_t}}$ minimizes the variance of $\bar{\theta}_T$ when the covariances among the score functions are 0. When $\{\phi_t\}_{t=1}^{T}$ is an MDS, the covariance is 0. Thus, the proposed estimator uses the weights and scores that minimize the variance for an estimator with a form that satisfies $\sum_{t=1}^{T} w_t \bar{\psi}_t(X_t, A_t, Y_t, \pi_t)$.

5 Main Algorithm: Two-step FA3IPW Estimation

Under the assumption of $g_{t,N} - \sigma_t^2 \xrightarrow{\mathbb{P}} 0$, the proposed estimator $\bar{\theta}_T^{A3IPW}$ has the asymptotic normality. However, to estimate $\sigma_t^2 = \text{Var} (\bar{\theta}_t | \Omega_{t-1})$, a consistent estimator of $\sigma_t^2$ is required, which is why we use $g_{t,N}$. However, we can obtain a consistent estimator of $\theta_0$ without imposing the assumption that $g_{t,N} - \sigma_t^2 \xrightarrow{\mathbb{P}} 0$. Based on these properties, we propose two-step estimation. First, using arbitrary appropriate values for $g_{t,N} = \bar{g}_t$ for $t = 1, 2, \ldots, T - 1$, such as $\bar{g}_t = 1$, we obtain the initial estimates, $\{\tilde{\theta}_t\}_{t=1}^{T-1}$. Then, using $\tilde{\theta}_t \xrightarrow{\mathbb{P}} \theta_0$, we construct $g_{t,N} - \sigma_t^2 \xrightarrow{\mathbb{P}} 0$. Finally, we can estimate the next estimator of $\theta_0$, $\bar{\theta}_T^{TSFA3IPW}$, with asymptotic normality from Theorem 4. We summarize the process of two-step FA3IPW (TSFA3IPW) estimation in Algorithm 1.

**Remark 10** (Estimation of Functions $f_t^*$ and $e_t^*$). Several nonparametric regression estimators can be proved to be consistent for dependent samples via certain bandit algorithms (Yang & Zhu, 2002).
Table 1: Experimental results under the RW policy. The method with the lowest MSE is in bold.

| Datasets | satimage | pendigits | mnist | dna | letter | sensorless | connect-4 | covtype |
|----------|----------|-----------|-------|-----|--------|------------|-----------|---------|
| Metrics  | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD |
| FA3IPW   | 0.037 | 0.193 | 0.064 | 0.254 | 0.048 | 0.219 | 0.012 | 0.178 | 0.128 | 0.358 | 0.101 | 0.313 | 0.039 | 0.163 | 0.045 | 0.210 |
| SFA3IPW  | 0.037 | 0.192 | 0.061 | 0.243 | 0.051 | 0.225 | 0.029 | 0.169 | 0.142 | 0.376 | 0.096 | 0.308 | 0.033 | 0.178 | 0.042 | 0.206 |
| DM       | 0.101 | 0.168 | 0.056 | 0.171 | 0.028 | 0.178 | 0.020 | 0.122 | 0.089 | 0.143 | 0.043 | 0.153 | 0.100 | 0.158 | 0.109 | 0.165 |
| A2IPW    | 0.068 | 0.261 | 0.089 | 0.297 | 0.069 | 0.263 | 0.063 | 0.251 | 0.129 | 0.359 | 0.133 | 0.360 | 0.089 | 0.242 | 0.070 | 0.252 |
| FA2aIPW  | 0.038 | 0.193 | 0.063 | 0.250 | 0.049 | 0.220 | 0.035 | 0.185 | 0.119 | 0.345 | 0.100 | 0.314 | 0.032 | 0.164 | 0.052 | 0.226 |
|          | 0.068 | 0.261 | 0.090 | 0.301 | 0.074 | 0.273 | 0.057 | 0.238 | 0.146 | 0.373 | 0.133 | 0.358 | 0.075 | 0.237 | 0.059 | 0.236 |

Table 2: Experimental results under the UCB policy. The method with the lowest MSE is in bold.

| Datasets | satimage | pendigits | mnist | dna | letter | sensorless | connect-4 | covtype |
|----------|----------|-----------|-------|-----|--------|------------|-----------|---------|
| Metrics  | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD |
| FA3IPW   | 0.041 | 0.201 | 0.095 | 0.308 | 0.098 | 0.308 | 0.026 | 0.160 | 0.042 | 0.213 | 0.077 | 0.244 | 5.484 | 1.795 | 0.066 | 0.257 |
| SFA3IPW  | 0.042 | 0.198 | 0.071 | 0.263 | 0.126 | 0.353 | 0.029 | 0.167 | 0.488 | 0.679 | 0.070 | 0.262 | 5.248 | 2.240 | 0.077 | 0.275 |
| DM       | 0.022 | 0.089 | 0.218 | 0.114 | 0.245 | 0.168 | 0.195 | 0.130 | 0.489 | 0.144 | 0.401 | 0.141 | 0.112 | 0.201 | 0.107 | 0.172 |
| A2IPW    | 0.078 | 0.275 | 0.099 | 0.314 | 0.188 | 0.421 | 0.051 | 0.223 | 0.891 | 0.929 | 0.101 | 0.315 | 11.424 | 2.560 | 0.072 | 0.267 |
| FA2aIPW  | 0.142 | 0.360 | 0.386 | 0.563 | 0.097 | 0.304 | 0.025 | 0.158 | 0.854 | 0.909 | 0.083 | 0.286 | 2.363 | 1.400 | 0.060 | 0.245 |
|          | 0.072 | 0.263 | 0.124 | 0.351 | 0.173 | 0.398 | 0.052 | 0.225 | 0.481 | 0.680 | 0.094 | 0.304 | 11.031 | 2.628 | 0.079 | 0.281 |

6 Experiments

In this section, using benchmark datasets, we demonstrate the effectiveness of the proposed FA3IPW estimator and the FA3IPW estimator with discarding some early samples for stabilization (SFA3IPW estimator in Remark 9). Following Dudik 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 satImage, pendigits, mnist, dna, letter, sensorless, connect-4, and covtype datasets. First, we create an adaptive policy $\pi^t$ by following a random walk and perform upper confidence bound (UCB) algorithms in the MAB problem. Then, we construct a behavior policy as a mixture of $\pi_t^u$ and the uniform random policy $\pi_t^u$, defined as $\pi_t = 0.7\pi_t^u + 0.3\pi_t^u$ (Kallus & Uehara, 2019a). When we use the RW policy, the policy $\pi_t^u$ does not converge. In contrast, when we use the UCB policy to generate $\pi_t^u$, the policy converges to a time-invariant policy. To construct an evaluation policy, we create a deterministic policy $\pi_d$ by training a logistic regression classifier on historical data. Then, we construct the evaluation policy $\pi^e$ as a mixture of $\pi_d$ and the uniform random policy $\pi^u$, defined as $\pi^e = 0.7\pi_d + 0.3\pi^u$. Through experiments, the behavior policy $\pi_t$ is assumed to be known. More details are in Appendix E.

We compare the MSES of six estimators, the FA3IPW, SFA3IPW, DM, A2aIPW, A2IPW, and FA2aIPW estimators. For the FA3IPW, SFA3IPW, and FA2aIPW estimators, we applied the two-step estimation. In each experiment, we have access to historical data with $T = 1000$ and evaluation data with $N = 1000$. With the SFA3IPW estimator, we discard the first 1000 samples of the historical data. When estimating $f^*$ and $e^*$, we use the Nadaraya-Watson regression (NW) estimator, which is used in Yang & Zhu (2002). The resulting MSES and their standard deviations (SDs) over 20 replications of each experiment are shown in Tables 1 and 2. In the experimental results, when the RW policy is used, the FA3IPW and SFA3IPW estimators show preferable performances compared with the A2IPW estimator. This is because the performance of the A2IPW estimator is guaranteed only when the policy converges to a time-invariant policy. On the other hand, when the UCB policy is used, the A2IPW estimator also works well as we can expect. From the results, we recommend using the FA3IPW estimator when the policy does not converge. More importantly, in the case, we can construct confidence intervals from the FA3IPW estimator, but cannot construct it from the A2IPW estimator. In Appendix E, we show the experimental results with other settings.

7 Conclusion

This study presented solutions for causal inference from dependent samples obtained via bandit feedback. By utilizing evaluation data or sample splitting, we can standardize the variance of an MDS. Then, we can apply the CLT for an MDS to obtain the asymptotic normality of the proposed estimator. In experiments, the proposed estimators showed theoretically expected performances.

https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
Broader Impact

In various applications, such as A/B testing and clinical trials, the problem of causal inference from dependent samples arises. Because this study promotes the use of OPE for those applications by presenting an effective solution to this problem, the ethical issues related to such applications will become more critical. For example, in clinical trials, the proposed method also should take the criteria proposed by the US Food and Drug Administration (FDA, 2019) into account.

Besides, although this study solved an essential problem with OPE, the limitations of methods for OPE should still be recognized. The existing methods for OPE try to find causality only from data without requiring domain-specific knowledge, but they sometimes lead to inappropriate decision making by producing incorrect results. When finding causality in the real world, it might be desirable to utilize domain-specific knowledge combined with the results obtained from machine learning methods.

References

Athey, S. and Wager, S. Efficient policy learning. arXiv preprint arXiv:1702.02896, 2017.
Beygelzimer, A. and Langford, J. The offset tree for learning with partial labels. In KDD, 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 ICML, volume 97, pp. 654–663. PMLR, 2019.
Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. Efficient and Adaptive Estimation for Semiparametric Models. Springer, 1998.
Brown, B. M. Martingale central limit theorems. Ann. Math. Statist., 42(1):59–66, 02 1971.
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.
Chernozhukov, V., Demirer, M., Lewis, G., and Syrgkanis, V. Semi-parametric efficient policy learning with continuous actions. In NeurIPS, 2019.
Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In AISTATS, volume 15 of Proceedings of Machine Learning Research, pp. 208–214, 2011.
der Laan, V. and Mark, J. I. The construction and analysis of adaptive group sequential designs. 2008.
Dudík, M., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Learning. In ICML, pp. 1097–1104, 2011.
Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. ICML, pp. 1447–1456, 2018.
FDA. Adaptive designs for clinical trials of drugs and biologics: Guidance for industry. Technical report, U.S. Department of Health and Human Services Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), 2019.
Greene, W. Econometric Analysis. Pearson Education, 2003.
Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments, 2019.
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.
Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260):663–685, 1952.
Kallus, N. and Uehara, M. Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning. In NeurIPS, pp. 3320–3329. 2019a.
Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019b.

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, 2020b.

Kitagawa, T. and Tetenov, A. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86:591–616, 2018.

Kosorok, M. R. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, New York, NY, 2008.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pp. 661–670, 2010.

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

Luedtke, A. R. and van der Laan, M. J. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713–742, Apr 2016.

Narita, Y., Yasui, S., and Yata, K. Efficient counterfactual learning from bandit feedback. *AAAI*, 2019.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *ICML*, volume 97, pp. 4881–4890, 2019.

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*, 89:846–866, 1994.

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

Scharfstein, D., Rotnizky, A., and Robins, J. M. Adjusting for nonignorable dropout using semi-parametric models. *Journal of the American Statistical Association*, 94:1096–1146, 1999.

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

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. *Ann. Statist.*, 30(1):100–121, 02 2002. doi: 10.1214/aos/1015362186.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107:1106–1118, 2012.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *arxiv preprint arxiv:1810.04778*, 2018.
A Preliminaries

A.1 Mathematical Tools

**Proposition 2.** [Slutsky Theorem, Greene (2003), Theorem D. 16, p. 1117] If \( a_n \xrightarrow{d} a \) and \( b_n \xrightarrow{p} b \), then
\[
a_n b_n \xrightarrow{d} ba.
\]

**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
\[
\mathbb{E}[|A_t| \cdot 1(|A_t| \geq c)] < \epsilon
\]
for all \( t \).

**Proposition 3.** [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 4** (\( L^r \) Convergence Theorem, Loève (1977)). Let \( 0 < r < \infty \), suppose that \( \mathbb{E}[|a_n|^r] < \infty \) for all \( n \) and that \( a_n \xrightarrow{p} a \) as \( n \to \infty \). The following are equivalent:

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

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

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

A.2 Martingale Limit Theorems

**Proposition 5.** [Weak Law of Large Numbers for Martingale, Hall et al. (2014)] Let \( \{S_n = \sum_{i=1}^{n} X_i, \Omega_t, t \geq 1\} \) be a martingale and \( \{b_n\} \) a sequence of positive constants with \( b_n \to \infty \) as \( n \to \infty \). Then, writing \( X_{ni} = X_i \cdot 1(|X_i| \leq b_n) \), \( 1 \leq i \leq n \), we have that \( b_n^{-1} S_n \xrightarrow{p} 0 \) as \( n \to \infty \) if

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

(ii) \( b_n^{-1} \sum_{i=1}^{n} \mathbb{E}[X_{ni} | \Omega_{t-1}] \xrightarrow{p} 0 \), and;

(iii) \( b_n^{2} \sum_{i=1}^{n} \{ \mathbb{E}[X_{ni}^2] - \mathbb{E}[X_{ni} | \Omega_{t-1}]^2 \} \to 0 \).

**Remark 11.** The weak law of large numbers for martingale holds when the random variable is bounded by a constant.

B Proof of Theorem [2]

**Proof.** The variance \( \sigma_t^2 \) can be calculated as follows:
\[
\sigma_t^2 = \mathbb{E} \left[ \left( \sum_{a=1}^{K} \left( \frac{\pi^e(a | X_t) \mathbb{1}[A_t = a] \{ Y_t - f^e(a, X_t) \}}{\pi_t(a | X_t, \Omega_{t-1})} + \pi^e(a | X_t) f^e(a, X_t) \right) - \theta_0 \right)^2 \right]
\]
\[
= \mathbb{E} \left[ \left( \sum_{a=1}^{K} \left( \frac{\pi^e(a | X_t) \mathbb{1}[A_t = a] \{ Y_t - f^e(a, X_t) \}}{\pi_t(a | X_t, \Omega_{t-1})} + \pi^e(a | X_t) f^e(a, X_t) \right) - \theta_0 \right)^2 \right]
\]
\[
= \mathbb{E} \left[ \sum_{a=1}^{K} \left( \frac{(\pi^e(a | X_t))^2 (e^e(a, X_t) - (f^e(a, X_t))^2)}{\pi_t(a | X_t, \Omega_{t-1})} + \left( \pi^e(a | X_t) f^e(a, X_t) - \theta_0 \right)^2 \right) \right].
\]
Therefore, what we want to show is
\[
\frac{1}{N} \sum_{i=1}^{N} \sum_{a=1}^{K} \left\{ \left( \pi^e(a \mid X_i) \right)^2 \left( \hat{\epsilon}_{t-1}(a \mid X_i) - \hat{f}_{t-1}(a \mid X_i) \right) \right\} + \left( \pi^e(a \mid X_i) \hat{f}_{t-1}(a, X_i) - \hat{\theta}_{t-1} \right)^2
\]
\[= - \sum_{a=1}^{K} \left\{ \left( \pi^e(a \mid X_i) \right)^2 \left( e^*(a, X_i) - f^*(a, X_i) \right) \right\} + \left( \pi^e(a \mid X_i) f^*(a, X_i) - \theta_0 \right)^2 \]
\[\overset{P}{\to} 0.\]

Here, from the weak law of large numbers, as \( N \to \infty, \)
\[
\frac{1}{N} \sum_{i=1}^{N} \sum_{a=1}^{K} \left\{ \left( \pi^e(a \mid X_i) \right)^2 \left( e^*(a, X_i) - f^*(a, X_i) \right) \right\} + \left( \pi^e(a \mid X_i) f^*(a, X_i) - \theta_0 \right)^2
\]
\[= \sum_{a=1}^{K} \left\{ \left( \pi^e(a \mid X_i) \right)^2 \left( e^*(a, X_i) - f^*(a, X_i) \right) \right\} + \left( \pi^e(a \mid X_i) f^*(a, X_i) - \theta_0 \right)^2 \]
\[\overset{P}{\to} 0 \text{ as } t \to \infty.\]

Therefore, as \( t \to \infty \) and \( N \to \infty, \)
\[
\frac{1}{N} \sum_{i=1}^{N} \sum_{a=1}^{K} \left\{ \left( \pi^e(a \mid X_i) \right)^2 \left( \hat{\epsilon}_{t-1}(a \mid X_i) - \hat{f}_{t-1}(a \mid X_i) \right) \right\} + \left( \pi^e(a \mid X_i) \hat{f}_{t-1}(a, X_i) - \hat{\theta}_{t-1} \right)^2
\]
\[- \sum_{a=1}^{K} \left\{ \left( \pi^e(a \mid X_i) \right)^2 \left( e^*(a, X_i) - f^*(a, X_i) \right) \right\} + \left( \pi^e(a \mid X_i) f^*(a, X_i) - \theta_0 \right)^2 \]
\[\overset{P}{\to} 0.\]

\[\square\]

### C Proof of Theorem 4

**Proof.** We have
\[
\sqrt{T}(\hat{\theta}_T - \theta_0) = \left( \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\sqrt{g_{t,N}}} \right)^{-1} \sqrt{T} \sum_{t=1}^{T} \hat{z}_t,
\]
where note that
\[
\hat{z}_t = \frac{1}{\sqrt{g_{t,N}}} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^e(a \mid X_i) 1[A_t = a] \{Y_t - \hat{f}_{t-1}(a, X_t)\}}{\pi^e(a \mid X_i, \Omega_{t-1})} + \pi^e(a \mid x) \hat{f}_{t-1}(a, X_t) \right\} - \theta_0 \right).
\]
If we show
\[
\sqrt{T} \sum_{t=1}^{T} \hat{z}_t \overset{d}{\to} N(0, 1) \text{ as } T \to \infty,
\]
as in (3).

12
then we can show that
\[
\left( \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\sqrt{d_{i,N}}} \right) \sqrt{T} (\hat{\theta}_T - \theta_0) = \sqrt{T} \frac{1}{T} \sum_{t=1}^{T} \hat{z}_t \overset{d}{\to} N(0, 1).
\]

To show (a), we use the central limit theorem for an MDS (Proposition 1 in Appendix A). Because \( \{\hat{z}_t\}_{t=1}^{T} \) is an MDS, we check the following three conditions of Proposition 1.

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

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

(c) \((1/T) \sum_{t=1}^{T} \hat{z}_t^2 \overset{p}{\to} \nu^2.\)

**Step 1: Check of Condition (a)**

We have
\[
\mathbb{E}[\hat{z}_t^2] = \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{d_{i,N}}} \sum_{a=1}^{K} \left\{ \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - \hat{f}_{t-1}(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right)^2 \right]
\]
\[
= \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{d_{i,N}}} \sum_{a=1}^{K} \left\{ \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - \hat{f}_{t-1}(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right)^2 \right]
\]
\[
- \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{\sigma_t^2}} \sum_{a=1}^{K} \left\{ \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - f^*(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right)^2 \right] + \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{\sigma_t^2}} \sum_{a=1}^{K} \left\{ \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - f^*(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right)^2 \right].
\]

Because the last term is 1, We have
\[
\mathbb{E}[\hat{z}_t^2] - 1 = \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{d_{i,N}}} \sum_{a=1}^{K} \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - \hat{f}_{t-1}(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right]^2
\]
\[
- \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{\sigma_t^2}} \sum_{a=1}^{K} \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - f^*(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right]^2.
\]

First, we show that the first and second terms vanishes asymptotically. We have
\[
\mathbb{E} \left[ \left\{ \frac{1}{\sqrt{d_{i,N}}} \sum_{a=1}^{K} \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - \hat{f}_{t-1}(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right]^2
\]
\[
- \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{\sigma_t^2}} \sum_{a=1}^{K} \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - f^*(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right]^2
\]
\[
\leq \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{d_{i,N}}} \sum_{a=1}^{K} \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - \hat{f}_{t-1}(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right]^2
\]
\[
- \left\{ \frac{1}{\sqrt{\sigma_t^2}} \sum_{a=1}^{K} \frac{\pi^a(a | X_i) \mathbb{I}[A_t = a] \{ Y_i - f^*(a, X_i) \}}{\pi_t(a | X_t, \Omega) - \theta_0} \right\} \right]^2.
\]
We can show that $1$ is uniformly integrable, we can prove

$$
\begin{align*}
\mathbb{E} & \left[ \left\{ \frac{1}{\sqrt{g_{t,N}}} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ y - \hat{f}_{t-1}(a, X_t) \} + \pi^\beta(a \mid x) \hat{f}_{t-1}(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \\
& - \mathbb{E} \left[ \left\{ \frac{1}{\sigma_i^2} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ Y_t - f^*(a, X_t) \} + \pi^\beta(a \mid x) f^*(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \\
& \leq \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{g_{t,N}}} - \frac{1}{\sqrt{\sigma_i^2}} \right\} \hat{C}_1 + \mathbb{E} \left[ \left\{ \hat{f}_{t-1}(a, X_t) - f^*(a, X_t) \right\} \right] \hat{C}_2. 
\end{align*}
$$

From the boundedness of each variable in $\hat{z}_t$, there exist absolute constants $\tilde{C}_1$ and $\tilde{C}_2$ such that

$$
\begin{align*}
& \leq \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{g_{t,N}}} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ y - \hat{f}_{t-1}(a, X_t) \} + \pi^\beta(a \mid x) \hat{f}_{t-1}(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \\
& - \mathbb{E} \left[ \left\{ \frac{1}{\sigma_i^2} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ Y_t - f^*(a, X_t) \} + \pi^\beta(a \mid x) f^*(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \\
& \leq \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{g_{t,N}}} - \frac{1}{\sqrt{\sigma_i^2}} \right\} \hat{C}_1 + \mathbb{E} \left[ \left\{ \hat{f}_{t-1}(a, X_t) - f^*(a, X_t) \right\} \right] \hat{C}_2. 
\end{align*}
$$

From the continuous mapping theorem and assumption that $g_{t,N} - \sigma_i^2 \overset{P}{\to} 0$ as $t \to \infty$ and $N \to \infty$, we have $1/\sqrt{g_{t,N}} - 1/\sigma_i^2 \overset{P}{\to} 0$ as $t \to \infty$ and $N \to \infty$ from the continuous mapping theorem. Then, if $1/\sqrt{g_{t,N}}$ is uniformly integrable, we can prove $\mathbb{E}[|1/\sqrt{g_{t,N}} - 1/\sigma_i^2|] \to 0$ as $t \to \infty$ using $L^*$-convergence theorem (Proposition 4 in Appendix A).

Because $\hat{f}_{t-1}(a, x)$ is independent from $X_t$, we have $\mathbb{E}[\hat{f}_{t-1}(a, X_t) \mid X_t = x] = \mathbb{E}[\hat{f}_{t-1}(a, x)]$. For $\mathbb{E}[\hat{f}_{t-1}(a, x)]$ and fixed $x$, from the point convergence of $\hat{f}_{t-1}(a, x)$, we have

$$
\mathbb{E} \left[ \left| \hat{f}_{t-1}(a, X_t) - \mathbb{E} f^*(a, X_t) \right| \mid X_t = x \right] = \mathbb{E} \left[ \left| \hat{f}_{t-1}(a, x) - \mathbb{E} f^*(a, x) \right| \right] \to 0.
$$

Then, from the Lebesgue’s dominated convergence theorem, for $\mathbb{E}[\hat{f}_{t-1}(a, X_t) \mid X_t = x]$, which pointwisely converges to $\mathbb{E}[f^*(a, X_t) \mid X_t = x]$, we can show that

$$
\mathbb{E}_X \left[ \left| \mathbb{E}[\hat{f}_{t-1}(a, X_t) \mid X_t = X_t] - \mathbb{E}[f^*(a, X_t) \mid X_t] \right| \right] \to 0.
$$

We can show that $1/\sqrt{g_{t,N}}$ and $\hat{f}_{t-1}$ are uniformly integrable from the boundedness of $1/\sqrt{g_{t,N}}$ and $\hat{f}_{t-1}$ (Proposition 4 in Appendix A). Then, as $t \to \infty$ and $T \to \infty$,

$$
\begin{align*}
& \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{g_{t,N}}} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ y - \hat{f}_{t-1}(a, X_t) \} + \pi^\beta(a \mid x) \hat{f}_{t-1}(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \\
& - \mathbb{E} \left[ \left\{ \frac{1}{\sigma_i^2} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ Y_t - f^*(a, X_t) \} + \pi^\beta(a \mid x) f^*(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \\
& \leq \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{g_{t,N}}} - \frac{1}{\sqrt{\sigma_i^2}} \right\} \hat{C}_1 + \mathbb{E} \left[ \left\{ \hat{f}_{t-1}(a, X_t) - f^*(a, X_t) \right\} \right] \hat{C}_2 \right] \to 0.
\end{align*}
$$

Therefore, for any $\epsilon > 0$, there exists $\tilde{t} > 0$ and $\tilde{T} > 0$ such that, for $T > \max \{ \tilde{T}, \tilde{t}/r \}$,

$$
\begin{align*}
& \mathbb{E} \left[ \tilde{z}_t^2 \right] - 1 \\
& = \frac{1}{T} \sum_{t=1}^{T} \left( \mathbb{E} \left[ \left\{ \frac{1}{\sqrt{g_{t,N}}} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ y - \hat{f}_{t-1}(a, X_t) \} + \pi^\beta(a \mid x) \hat{f}_{t-1}(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \\
& - \mathbb{E} \left[ \left\{ \frac{1}{\sigma_i^2} \left( \sum_{a=1}^{K} \left\{ \frac{\pi^\alpha(a \mid X_t) \mathbb{I}[A_t = a] \{ Y_t - f^*(a, X_t) \} + \pi^\beta(a \mid x) f^*(a, X_t) \} - \theta_0 \right\} \right)^2 \right\} \right] \right) \\
& \leq \tilde{t}/T + \epsilon.
\end{align*}
$$

Thus, we have $(1/T) \sum_{t=1}^{T} \nu_t^2 - 1 \leq \tilde{t}/T + \epsilon \to 0$ as $T \to \infty$.

**Step 2: Check of Condition (b)**

We can prove this condition using the assumption of the boundedness of the random variables.
Step 3: Check of Condition (c)

To show \((1/T)\sum_{t=1}^{T} \varepsilon_t^2 \xrightarrow{P} \nu^2\), we show (i) \((1/T)\sum_{t=1}^{T} \varepsilon_t^2 - (1/T)\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] \xrightarrow{P} 0\); and (ii) \((1/T)\sum_{t=1}^{T} \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \xrightarrow{P} 0\), where \(\nu^2 = \mathbb{E} \left\{ \frac{1}{\sqrt{\sigma_t^2}} \left( \sum_{a=1}^{K} K \right)^\prime \left\{ \frac{\pi^c(a | X_i) \mathbb{1}[A_t = a]}{\pi_t(a | X_t, \Omega_{t-1})} \right\} + \pi^c(a | x)f^*(a, X_t) - \theta_0 \right\}^2 \right\} = 1.

Proof of (i): We can prove (i) directly applying law of large numbers for an MDS to an MDS \(\phi_t = \varepsilon_t^2 - \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}]\). The conditions for using the law of large numbers are satisfied by the boundedness of components of \(\varepsilon_t\).

Proof of (ii) Next, we show that \(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \xrightarrow{P} 0\).

From Markov’s inequality, for \(\varepsilon > 0\), we have

\[
P\left( \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \geq \varepsilon \right) \leq \frac{\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \right]}{\varepsilon} \leq \frac{\mathbb{E} \left[ \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \right]}{\varepsilon}.
\]

Then, we consider showing \(\mathbb{E} \left[ \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \right] \to 0\). Here, we have

\[
\mathbb{E} \left[ \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \right] = \mathbb{E} \left[ \mathbb{E} \left\{ \left( \sum_{a=1}^{K} \frac{\pi^c(a | X_i) \mathbb{1}[A_t = a]}{\pi_t(a | X_t, \Omega_{t-1})} \right) + \pi^c(a | x)f^*(a, X_t) - \theta_0 \right\}^2 \right].
\]

Then, by using Jensen’s inequality,

\[
\mathbb{E} \left[ \mathbb{E}[\varepsilon_t^2 | \Omega_{t-1}] - \nu^2 \right] \leq \mathbb{E} \left[ \mathbb{E} \left\{ \left( \sum_{a=1}^{K} \frac{\pi^c(a | X_i) \mathbb{1}[A_t = a]}{\pi_t(a | X_t, \Omega_{t-1})} \right) + \pi^c(a | x)f^*(a, X_t) - \theta_0 \right\}^2 \right].
\]

Confidence Interval for Off-Policy Evaluation

from Dependent Samples via Bandit Algorithm:

Approach from Standardized Martingales

A PREPRINT

15
Because \( \hat{f}_{t-1} \) and \( g_{t,N} \) are constructed from \( \Omega_{t-1} \),
\[
\mathbb{E} \left[ \left| \mathbb{E} \left[ z_t^2 \mid \Omega_{t-1} \right] - \nu^2 \right| \right] 
\leq \mathbb{E} \left[ \left( \frac{1}{g_{t,N}} \sum_{a=1}^{K} \frac{\pi^*(a \mid X_t) \mathbf{1}[A_t = a] \left\{ Y_t - \hat{f}_{t-1}(a, X_t) \right\}}{\pi_t(a \mid X_t, \Omega_{t-1})} + \pi^*(a \mid x) f_{t-1}(a, X_t) - \theta_0 \right)^2 \right] \bigg| X_t, \hat{f}_{t-1}, g_{t,N} \bigg].
\]
From the results of Step 1, there exist \( \hat{C}_1 \) and \( \hat{C}_2 \) such that
\[
\mathbb{E} \left[ \left| \mathbb{E} \left[ z_t^2 \mid \Omega_{t-1} \right] - \nu^2 \right| \right] 
\leq \mathbb{E} \left[ \left( \frac{1}{g_{t,N}} \sum_{a=1}^{K} \frac{\pi^*(a \mid X_t) \mathbf{1}[A_t = a] \left\{ Y_t - f^*(a, X_t) \right\}}{\pi_t(a \mid X_t, \Omega_{t-1})} + \pi^*(a \mid x) f^*(a, X_t) - \theta_0 \right)^2 \right] 
\leq \hat{C}_1 + \mathbb{E} \left[ \left| \hat{f}_{t-1}(a, X_t) - f^*(a, X_t) \right| \bigg| \hat{C}_2 \right].
\]
Then, from \( L' \) convergence theorem, by using point convergence and the boundedness of \( g_{t,N} \) and \( \hat{f}_{t-1} \), we have \( \mathbb{E} \left[ \left| \mathbb{E} \left[ z_t^2 \mid \Omega_{t-1} \right] - \nu^2 \right| \right] \rightarrow 0 \) as \( t \rightarrow \infty \) and \( N \rightarrow \infty \). Therefore, as \( T \rightarrow \infty \) and \( N \rightarrow \infty \),
\[
\mathbb{P} \left( \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ z_t^2 \mid \Omega_{t-1} \right] - \nu^2 \bigg| \geq \epsilon \right) 
\leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \left| \mathbb{E} \left[ z_t^2 \mid \Omega_{t-1} \right] - \nu^2 \right| \right] \rightarrow 0.
\]
As a conclusion,
\[
\frac{1}{T} \sum_{t=1}^{T} z_t^2 - \nu^2 = \frac{1}{T} \sum_{t=1}^{T} \left( z_t^2 - \mathbb{E} \left[ z_t^2 \mid \Omega_{t-1} \right] + \mathbb{E} \left[ z_t^2 \mid \Omega_{t-1} \right] - \nu^2 \right) \rightarrow 0.
\]

**Conclusion**

From the results of Steps 1–3, we showed the theorem. \( \square \)

**Remark 12** (Donsker Properties). Readers might feel that functions need to satisfy some assumptions such as Donsker properties (Chernozhukov et al., 2018).

### D Details of Sample Splitting

As well as Theorem\(^2\) we prove the consistency of \( g_{t,(1-r)T} \) in the following theorem.

**Corollary 1.** Suppose that, for all \( x \in \mathcal{X} \) and \( k \in \mathcal{A} \), \( \hat{f}_{t-1}(a, x) - f^*(a, x) \overset{p}{\rightarrow} 0, \hat{e}_{t-1}(a, x) - e^*(a, x) \overset{p}{\rightarrow} 0 \) and \( \hat{\theta}_t \overset{p}{\rightarrow} \theta_0 \) as \( t \rightarrow \infty \). Then, for \( \pi_t \in \Pi \), we have \( g_{t,(1-r)T} - \sigma_t^2 \overset{p}{\rightarrow} 0 \) as \( t \rightarrow \infty \) and \( T \rightarrow \infty \), where
\[
\sigma_t^2 = \mathbb{E} \left[ \sum_{a=1}^{K} \left( \frac{\pi^*(a \mid X_t)^2 \nu^*(a, X_t)}{\pi_t(a \mid X_t, \Omega_{t-1})} + \left( \pi^*(a \mid X_t) f^*(a, X_t) - \theta_0 \right)^2 \right) \right].
\]

Then, using \( g_{t,(1-r)T} \), we define a FA3IPW estimator with sample splitting \( \hat{g}_{r,T}^{\text{FA3IPW-SS}} \) as
\[
\left( \sum_{t=1}^{T} \frac{1}{g_{t,(1-r)T}} \right)^{-1} \sum_{t=1}^{T} \frac{1}{g_{t,(1-r)T}} \left( \frac{\pi^*(a \mid X_t) \mathbf{1}[A_t = a] Y_t - \hat{f}_{t-1}(a, X_t)}{\pi_t(a \mid X_t, \Omega_{t-1})} + \pi^*(a \mid x) f_{t-1}(a, X_t) \right). \]

The asymptotic distribution can be obtained from Theorem\(^1\) as follows.

**Corollary 2** (Asymptotic Distribution of the Proposed Estimator). Suppose that
use the K-nearest neighbor regression. In Table 8, we use a policy
in Table 7, we use a policy
the K-nearest neighbor regression.
the K-nearest neighbor regression.
In Table 6, we use a policy
In Table 4, we use a policy
In Table 3, we use a policy

(i) there exist $C_2$ and $C_3$ such that $|\hat{f}_t| \leq C_2$ and $0 < 1/\sqrt{g_t(1-\gamma)T} \leq C_3$;

(ii) Pointwise convergence in probability of $\hat{f}_t$, i.e., $\forall x \in \mathcal{X}, \hat{f}_t(a,x) - f^*(a,x) \xrightarrow{P} 0$ as $t \to \infty$;

(ii) $g_t(1-\gamma)T - \sigma_t^2 \xrightarrow{P} 0$ as $t \to \infty$ and $T \to \infty$.

Then, under Assumptions $\mathcal{A}$ $\frac{1}{\sqrt{T}} \sqrt{T - \frac{1}{g_t(1-\gamma)T}} \left( \hat{\delta}^{\text{FA3IPW}}_{T} - \theta_0 \right) \xrightarrow{d} \mathcal{N}(0,1)$ as $T \to \infty$.

Thus, we can obtain an estimator with asymptotic normality without assuming the access to $\mathcal{E}_N$.

E Details of Experiments

In this section, we show the details and additional results with different settings of Section 6. First, we show the description of the datasets in Table. All datasets are downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/. For these datasets, in the experiment shown in Section 6 we use a policy $\pi_t = 0.7\pi_t^a + 0.3\pi_t^b$ with 2,000 samples. We use the UCB and RW policy for the adaptive policy. The UCB algorithm follows the LinUCB proposed by Chu et al. (2011). In the random walk policy, we update the policy as follows:

(i) At the first period, the policy $\pi_1^a$ is given from the uniform distribution with a normalization for becoming $\sum_{k=1}^{K} \pi_1^a(k \mid X_1) = 1$.

(ii) At a period $t$, we update the policy as $\pi_{t+1}^a(k \mid X_{t+1}) = \pi_t^a(k \mid X_t) + \eta_{k,t}$, where $\eta_{k,t}$ is a random variable drawn from the normal distribution.

(iii) For satisfying the definition of the probability, we normalize $\pi_{t+1}^a(k \mid X_{t+1})$ to become $\sum_{k=1}^{K} \pi_{t+1}^a(k \mid X_{t+1}) = 1$.

(iv) Iterate (ii) and (iii) until the trial ends.

For simplicity, we do not use context in the RW policy. For nonparametric regression, we only use the Nadaraya-Watson regression in Section 6. The evaluation policy is fixed at $\pi^e = 0.7\pi^a + 0.3\pi^b$. The policy $\pi_d$ is a prediction of the logistic regression trained by all samples in $S_T$. If the logistic regression predicts a class $a \in \mathcal{A}$ as the true label given $X_t$, the policy becomes $\pi_d(a, X_t) = 1$ and $\pi_d(a', X_t) = 0$ for $a' \neq a$.

In Table 4 we use a policy $\pi_t = 0.7\pi_t^a + 0.3\pi_t^b$ with $T = 1,000$ and $N = 1,000$. For nonparametric regression, we use the K-nearest neighbor regression.

In Table 5 we use a policy $\pi_t = 0.7\pi_t^a + 0.3\pi_t^b$ with $T = 500$ and $N = 1,000$. For nonparametric regression, we use the Nadaraya-Watson regression.

In Table 6 we use a policy $\pi_t = 0.7\pi_t^a + 0.3\pi_t^b$ with $T = 500$ and $N = 1,000$. For nonparametric regression, we use the K-nearest neighbor regression.

In Table 7 we use a policy $\pi_t = 0.5\pi_t^a + 0.5\pi_t^b$ with $T = 1,000$ and $N = 1,000$. For nonparametric regression, we use the Nadaraya-Watson regression.

In Table 8 we use a policy $\pi_t = 0.5\pi_t^a + 0.5\pi_t^b$ with $T = 1,000$ and $N = 1,000$. For nonparametric regression, we use the K-nearest neighbor regression.
The proposed FA3IPW estimator performs well in many datasets. Especially when we use the RW policy, the performance of the estimator tends to be better than the other methods.

Table 4: Experimental results under the RW policy (the upper table) and UCB policy (the lower table) under a policy $\pi_t = 0.7\pi^0_t + 0.3\pi^1_t$ with $T = 1,000$ and $N = 1,000$, and the K-nearest neighbor regression. The method with the lowest MSE is in bold.

| Datasets  | satimage | pendigits | mnist | dna | letter | sensorsless | connect-4 | covtype |
|----------|----------|-----------|-------|-----|--------|-------------|-----------|---------|
| Metrics  | MSE  | SD  | MSE  | SD  | MSE  | SD  | MSE  | SD  | MSE  | SD  |
| FA2daIPW | 0.196 | 0.204 | 0.099 | 0.314 | 0.039 | 0.196 | 0.203 | 0.434 | 0.109 | 0.328 |
| SF A3IPW | 0.207 | 0.252 | 0.125 | 0.348 | 0.207 | 0.452 | 0.079 | 0.271 | 0.381 | 0.614 |

Table 5: Experimental results under the RW policy (upper table) and UCB policy (lower table) under a policy $\pi_t = 0.7\pi^0_t + 0.3\pi^1_t$ with $T = 500$ and $N = 1,000$, and the Nadaraya-Watson regression. The method with the lowest MSE is in bold.

| Datasets | satimage | pendigits | mnist | dna | letter | sensorsless | connect-4 | covtype |
|----------|----------|-----------|-------|-----|--------|-------------|-----------|---------|
| Metrics  | MSE  | SD  | MSE  | SD  | MSE  | SD  | MSE  | SD  | MSE  | SD  |
| FA2IPW   | 0.035 | 0.104 | 0.116 | 0.338 | 0.081 | 0.176 | 0.076 | 0.394 | 0.098 | 0.173 |
| SF A3IPW | 0.045 | 0.102 | 0.099 | 0.314 | 0.134 | 0.353 | 0.048 | 0.220 | 0.267 | 0.516 |
| DM       | 0.004 | 0.072 | 0.205 | 0.490 | 0.247 | 0.521 | 0.120 | 0.349 | 0.245 | 0.493 |
| AdaIPW   | 0.082 | 0.285 | 0.144 | 0.375 | 0.121 | 0.336 | 0.070 | 0.264 | 0.196 | 0.442 |
| A2IPW    | 0.051 | 0.213 | 0.113 | 0.333 | 0.095 | 0.293 | 0.044 | 0.207 | 0.190 | 0.436 |
| FA2aIPW  | 0.089 | 0.289 | 0.129 | 0.348 | 0.107 | 0.338 | 0.061 | 0.246 | 0.202 | 0.449 |

The proposed FA3IPW estimator performs well in many datasets. Especially when we use the RW policy, the performance of the estimator tends to be better than the other methods.
Table 6: Experimental results under the RW policy (upper table) and UCB policy (lower table) under a policy \( \pi_t = 0.7\pi_t^0 + 0.3\pi_t^p \) with \( T = 500 \) and \( N = 1,000 \), and the K-nearest neighbor regression. The method with the lowest MSE is in bold.

| Datasets | Metrics | satimage | pendigits | mnist | dna | letter | sensorless | connect-4 | covtype |
|----------|---------|----------|----------|-------|-----|--------|------------|-----------|---------|
| \( \pi_t \) | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD |
| FA3IPW | 0.048 | 0.231 | 0.088 | 0.219 | 0.089 | 0.258 | 0.036 | 0.194 | 0.219 | 0.457 | 0.112 | 0.330 | 0.029 | 0.169 | 0.084 | 0.287 |
| SFA3IPW | 0.055 | 0.222 | 0.089 | 0.282 | 0.098 | 0.307 | 0.043 | 0.207 | 0.296 | 0.521 | 0.118 | 0.343 | 0.038 | 0.195 | 0.082 | 0.287 |
| DM | 0.068 | 0.178 | 0.292 | 0.235 | 0.545 | 0.199 | 0.371 | 0.167 | 0.435 | 0.154 | 0.305 | 0.256 | 0.131 | 0.170 | 0.088 | 0.239 |
| AdaIPW | 0.089 | 0.280 | 0.143 | 0.356 | 0.114 | 0.330 | 0.080 | 0.282 | 0.255 | 0.474 | 0.157 | 0.394 | 0.077 | 0.242 | 0.127 | 0.354 |
| A2IPW | 0.048 | 0.210 | 0.086 | 0.277 | 0.080 | 0.280 | 0.042 | 0.206 | 0.239 | 0.469 | 0.113 | 0.332 | 0.029 | 0.170 | 0.077 | 0.277 |
| FA2dIPW | 0.093 | 0.289 | 0.145 | 0.354 | 0.115 | 0.336 | 0.072 | 0.266 | 0.242 | 0.457 | 0.157 | 0.395 | 0.087 | 0.240 | 0.130 | 0.356 |

Table 7: Experimental results under the RW policy (upper table) and UCB policy (lower table) under a policy \( \pi_t = 0.5\pi_t^0 + 0.5\pi_t^p \) with \( T = 1,000 \) and \( N = 1,000 \), and the Nadaraya-Watson regression. The method with the lowest MSE is in bold.

| Datasets | Metrics | satimage | pendigits | mnist | dna | letter | sensorless | connect-4 | covtype |
|----------|---------|----------|----------|-------|-----|--------|------------|-----------|---------|
| \( \pi_t \) | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD |
| FA3IPW | 0.123 | 0.350 | 0.136 | 0.364 | 0.164 | 0.402 | 0.051 | 0.225 | 0.172 | 0.415 | 0.102 | 0.306 | 5.501 | 2.345 | 0.076 | 0.276 |
| SFA3IPW | 0.081 | 0.277 | 0.113 | 0.335 | 0.210 | 0.457 | 0.057 | 0.238 | 0.217 | 0.463 | 0.131 | 0.347 | 5.283 | 1.496 | 0.085 | 0.291 |
| DM | 0.026 | 0.183 | 0.088 | 0.144 | 0.509 | 0.151 | 0.376 | 0.165 | 0.435 | 0.185 | 0.277 | 0.255 | 0.131 | 0.163 | 0.090 | 0.216 |
| AdaIPW | 0.232 | 0.482 | 0.253 | 0.503 | 0.253 | 0.502 | 0.109 | 0.331 | 0.193 | 0.439 | 0.140 | 0.372 | 9.117 | 2.109 | 0.091 | 0.300 |
| A2IPW | 0.130 | 0.359 | 0.132 | 0.359 | 0.168 | 0.410 | 0.051 | 0.225 | 0.161 | 0.399 | 0.100 | 0.305 | 2.805 | 1.661 | 0.070 | 0.265 |
| FA2dIPW | 0.194 | 0.440 | 0.280 | 0.529 | 0.243 | 0.492 | 0.100 | 0.316 | 0.205 | 0.453 | 0.143 | 0.375 | 7.886 | 2.007 | 0.100 | 0.316 |

Table 8: Experimental results under the RW policy (upper table) and UCB policy (lower table) under a policy \( \pi_t = 0.5\pi_t^0 + 0.5\pi_t^p \) with \( T = 1,000 \) and \( N = 1,000 \), and the K-nearest neighbor regression. The method with the lowest MSE is in bold.

| Datasets | Metrics | satimage | pendigits | mnist | dna | letter | sensorless | connect-4 | covtype |
|----------|---------|----------|----------|-------|-----|--------|------------|-----------|---------|
| \( \pi_t \) | MSE | SD | MSE | SD | MSE | SD | MSE | SD | MSE | SD |
| FA3IPW | 0.038 | 0.193 | 0.080 | 0.248 | 0.038 | 0.251 | 0.025 | 0.158 | 0.097 | 0.306 | 0.055 | 0.234 | 0.031 | 0.156 | 0.045 | 0.210 |
| SFA3IPW | 0.040 | 0.199 | 0.056 | 0.237 | 0.071 | 0.261 | 0.029 | 0.172 | 0.088 | 0.313 | 0.080 | 0.281 | 0.031 | 0.156 | 0.045 | 0.210 |
| DM | 0.054 | 0.157 | 0.217 | 0.209 | 0.459 | 0.186 | 0.368 | 0.164 | 0.449 | 0.142 | 0.263 | 0.190 | 0.118 | 0.126 | 0.083 | 0.163 |
| AdaIPW | 0.053 | 0.230 | 0.098 | 0.313 | 0.091 | 0.301 | 0.054 | 0.232 | 0.107 | 0.326 | 0.071 | 0.267 | 0.137 | 0.185 | 0.074 | 0.251 |
| A2IPW | 0.038 | 0.196 | 0.060 | 0.244 | 0.056 | 0.233 | 0.025 | 0.159 | 0.089 | 0.310 | 0.056 | 0.236 | 0.036 | 0.146 | 0.046 | 0.211 |
| FA2dIPW | 0.052 | 0.228 | 0.100 | 0.316 | 0.092 | 0.303 | 0.053 | 0.229 | 0.105 | 0.320 | 0.075 | 0.273 | 0.099 | 0.183 | 0.068 | 0.236 |

A P REPRINT

Confidence Interval for Off-Policy Evaluation from Dependent Samples via Bandit Algorithm: Approach from Standardized Martingales