Efficient Adaptive Experimental Design for Average Treatment Effect Estimation

A Preprint

Masahiro Kato[1,2], Takuya Ishihara[2], Junya Honda[2,3], and Yusuke Narita[4]

[1]CyberAgent Inc.
[2]The University of Tokyo
[3]RIKEN
[4]Yale University

October 27, 2020

Abstract

The goal of many scientific experiments including A/B testing is to estimate the *average treatment effect* (ATE), which is defined as the difference between the expected outcomes of two or more treatments. In this paper, we consider a situation where an experimenter can assign a treatment to research subjects sequentially. In *adaptive experimental design*, the experimenter is allowed to change the probability of assigning a treatment using past observations for estimating the ATE efficiently. However, with this approach, it is difficult to apply a standard statistical method to construct an estimator because the observations are not independent and identically distributed. We thus propose an algorithm for efficient experiments with estimators constructed from dependent samples. We also introduce a *sequential testing* framework using the proposed estimator. To justify our proposed approach, we provide finite and infinite sample analyses. Finally, we experimentally show that the proposed algorithm exhibits preferable performance.

1 Introduction

Discovering causality from observations is a fundamental task in statistics and machine learning. In this paper, we follow Rubin (1974) to define a causal effect as the difference between the average outcomes resulting from two different actions, i.e., the *average treatment effect* (ATE). One of these actions corresponds to the *treatment* and the other corresponds to the *control* (Imbens & Rubin, 2015). One naive method for estimating the ATE using scientific experiments is the *randomized control trial* (RCT). In an RCT, we randomly assign one of the two actions to each research subject (Kendall, 2003) to obtain an unbiased estimator of the ATE (Imbens & Rubin, 2015).

However, while an RCT is a reliable method for scientific experiments, it often requires a large sample size for estimating the ATE precisely enough. To mitigate this problem, *adaptive experimental designs* have garnered increasing attention in various fields such as medicine and social science (Chow SC, 2005; van der Laan, 2008; Komiyama et al., 2009; Hahn et al., 2011; Chow & Chang, 2011; Villar et al., 2015; FDA, 2019). Compared to usual non-adaptive designs, adaptive designs often allow experimenters to detect the true causal effect while exposing fewer subjects to potentially harmful treatment. This motivates the US Food and Drug Administration (FDA) to recommend adaptive designs (FDA, 2019).

This paper proposes an adaptive experimental design that sequentially estimates a treatment assignment probability that minimizes the asymptotic variance of an estimator of the ATE and assigns a treatment according to the estimated probability. The proposed method is inspired by van der Laan (2008) and Hahn et al. (2011), which conduct hypothesis testing based on the asymptotic distribution for samples gather to minimize the asymptotic variance. In this paper, we show two directions of hypothesis testing based on the asymptotic distribution and the concentration inequality. The asymptotic variance based hypothesis testing is an extension of the methods proposed by van der Laan

(2008) and Hahn et al. (2011). The concentration inequality is a recently developed framework for hypothesis testing (Balsubramani & Ramdas, 2016), which is potentially useful in the real-world applications, such as ad-optimization.

One of the theoretical difficulties comes from the dependency among samples. Because we update the assignment probability using past observations, samples are not *independent and identically distributed* (i.i.d.). Therefore, instead of using existing results under the i.i.d. assumption for deriving the theoretical properties, we use the theoretical results of *martingales*. The main contributions of this paper are as follows: (i) We establish a framework of causal inference from samples obtained from a time-dependent algorithm with theoretical properties. (ii) We propose an algorithm for scientific experiments that achieves the lower bound of the asymptotic variance with several statistical hypothesis testing methods. This paper thus contributes to the literature and practice of RCTs and A/B testing by proposing an efficient experimental design with theoretical guarantees.

Organization of this Paper: In the following sections, we introduce the proposed algorithm with its theoretical analysis and experimental results. First, in Section 2, we define the problem setting. In Section 3, we present a new estimator constructed from samples with dependency. In Section 4, we introduce sequential hypothesis testing, which has the potential to reduce the sample size compared with conventional hypothesis testing. Then, we propose an algorithm for constructing an efficient estimator of the treatment effect in Section 5. Finally, in Section 6, we elucidate the empirical performance of the proposed algorithm using synthetic and semi-synthetic datasets.

2 Problem Setting

In the problem setting, a research subject arrives in a certain period and an experimenter assigns a treatment to the research subject. For simplicity, we assume the immediate observation of the outcome of a treatment. After several trials, we decide whether the treatment has an effect.

2.1 Data Generating Process

We define the data generating process (DGP) as follows. In period $t \in \mathbb{N}$, a research subject visits an experimenter, and the experimenter assigns an action $A_t \in \mathcal{A} = \{0, 1\}$ based on the *covariate* $X_t \in \mathcal{X}$, where $\mathcal{X}$ denotes the space of the covariate. After assigning the action, the experimenter observes a reward $Y_t \in \mathbb{R}$ immediately, which has a potential outcome denoted by a random variable, $Y_t : \mathcal{A} \to \mathbb{R}$. We have access to a set $\mathcal{S}_T = \{(X_t, A_t, Y_t)\}_{t=1}^T$ with the following DGP:

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

where $Y_t = \mathbb{1}[A_t = 0]Y_t(0) + \mathbb{1}[A_t = 1]Y_t(1)$ for an indicator function $\mathbb{1}[\cdot]$, $p(x)$ denotes the density of the covariate X_t, $p_t(a \mid x, \Omega_{t-1})$ denotes the probability of assigning an action A_t conditioned on a covariate X_t, $p(y \mid a, x)$ denotes 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 $\Omega_{t-1} = \{X_{t-1}, A_{t-1}, Y_{t-1}, \dots, X_1, A_1, Y_1\}$ with the space $\mathcal{M}_{t-1}$. We assume that $p(x)$ and $p(y \mid a, x)$ are invariant over time, but $p_t(a \mid x)$ can take different values. Further, we allow the decision maker to change $p_t(a \mid x, \Omega_{t-1})$ based on past observations. In this case, the samples $\left\{(X_t, A_t, Y_t)\right\}_{t=1}^T$ are correlated over time (i.e., the samples are not i.i.d.). The probability $p_t(a \mid x, \Omega_{t-1})$ is determined by a *policy* $\pi_t : \mathcal{A} \times \mathcal{X} \times \mathcal{M}_{t-1} \to (0, 1)$, which is a function of a covariate X_t, an action A_t, and a history Ω_{t-1}. For the policy $\pi_t(a \mid x, \Omega_{t-1})$, we consider the following process. First, we draw a random variable ξ_t following the uniform distribution on $[0, 1]$ in period t. Then, in each period t, we select an action A_t such that $A_t = \mathbb{1}[\xi_t \leq \pi_t(X_t, \otimes_{t-1})]$. Under this process, we regard the policy as the probability (i.e., $p_t(a \mid x, \Omega_{t-1}) = \pi_t(a \mid x, \Omega_{t-1})$).

Remark 1 (Observation of a Reward)**.** We assume that an outcome can be observed immediately after assigning an action. This setting is also referred to as *bandit feedback*. The case in which we observe a reward after some time can be considered as a special case of bandit feedback.

2.2 Average Treatment Effect Estimation

Our goal is to estimate the treatment effect, which is a *counterfactual* value because we can only observe an outcome of an action when assigning the action. Therefore, following the causality formulated by Rubin (1974), we consider estimating the ATE between $d = 1$ and $d = 0$ as $\theta_0 = \mathbb{E}[Y_t(1) - Y_t(0)]$ (Imbens & Rubin, 2015). For identification of θ_0, we put the following assumption.

Assumption 1 (Boundedness)**.** There exist C_1 and C_2 such that $\frac{1}{p_t(a|x)} \leq C_1$ and $|Y_t| \leq C_2$.

Remark 2 (Stable Unit Treatment Value Assumption)**.** In the DGP, we assume that the *Stable Unit Treatment Value Assumption*, namely, $p(y \mid a, x)$, is invariant no matter what mechanism is used to assign an action (Rubin, 1986).

Remark 3 (Unconfoundedness)**.** Existing methods often make an assumption called unconfoundedness: the outcomes $(Y_t(1), Y_t(0))$ and the action A_t are conditionally independent on X_t. In the DGP, this assumption is satisfied because we choose an action based on the observed outcome.

Notations: Let us denote $\mathbb{E}[Y_t(k) \mid x]$, $\mathbb{E}[Y_t^2(k) \mid x]$, $\mathrm{Var}(Y_t(k) \mid x)$, and $\mathbb{E}[Y_t(1) - Y_t(0) \mid x]$ as $f^*(k, x)$, $e^*(k, x)$, $v^*(k, x)$, and $\theta_0(x)$, respectively. Let $\hat{f}_t(k, x)$ be an estimator of $f^*(k, x)$, which is constructed from Ω_t. Let $\mathcal{N}(\mu, \mathrm{var})$ be the normal distribution with the mean μ and the variance var.

2.3 Existing Estimators

We review three types of standard estimators of the ATE in the case in which we know the probability of assigning an action and the samples are i.i.d., that is, the probability of assigning an action is invariant as $p(a \mid x) = p_1(a \mid x, \Omega_0) = p_2(a \mid x, \Omega_1) = \cdots$. The first estimator is an *inverse probability weighting* (IPW) estimator given by $\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{1}[A_t=1]Y_t}{p(1|X_t)} - \frac{\mathbb{1}[A_t=0]Y_t}{p(0|X_t)}\right)$ (Horvitz & Thompson, 1952; Rubin, 1987; Hirano et al., 2003; Swaminathan & Joachims, 2015). Although this estimator is unbiased when the behavior policy is known, it suffers from high variance. The second estimator is a direct method (DM) estimator $\frac{1}{T}\sum_{t=1}^{T}\left(\hat{f}_t(1, X_t) - \hat{f}_t(0, X_t)\right)$ (Hahn, 1998). This estimator is known to be weak against model misspecification for $\mathbb{E}[Y_t(k) \mid X_t]$. The third estimator is an augmented IPW (AIPW) estimator (Robins et al., 1994) defined as

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{1}[A_t=1]\big(Y_t - \hat{f}_T(1, X_t)\big)}{p(1 \mid X_t)} + \hat{f}_T(1, X_t) - \frac{\mathbb{1}[A_t=0]\big(Y_t - \hat{f}_T(0, X_t)\big)}{p(0 \mid X_t)} + \hat{f}_T(0, X_t)\right).$$

For the unbiasedness of the IPW and AIPW estimators, we can calculate the variance explicitly. The variance of the IPW estimator is $\left(\mathbb{E}\left[\frac{e^*(1,X_t)}{p(1|X_t)}\right] + \mathbb{E}\left[\frac{e^*(0,X_t)}{p(0|X_t)}\right] - \theta_0^2\right)/T$. The variance of the AIPW estimator is

$$\left(\mathbb{E}\left[\frac{v^*(1, X_t)}{p(1 \mid X_t)}\right] + \mathbb{E}\left[\frac{v^*(0, X_t)}{p(0 \mid X_t)}\right] + \mathbb{E}\big[(f^*(1, X_t) - f^*(0, X_t) - \theta_0)^2\big]\right)/T,$$

when $\hat{f}_T = f^*$. The asymptotic variances of the IPW and AIPW estimators are the same as their respective variances. Further, the variance and asymptotic variance are equal to the mean squared error (MSE) and asymptotic MSE (AMSE), respectively. As an important property, the (asymptotic) variance of the AIPW estimator achieves the lower bound of the asymptotic variance among regular $\sqrt{T}$-consistent estimators (van der Vaart, 1998, Theorem 25.20).

2.4 Semiparametric Efficiency

The lower bound of the variance is defined for an estimator under some posited models of the DGP. If this posited model is parametric, it is equal to the Cramér–Rao lower bound. When this posited model is a non- or semiparametric, we can still define the corresponding lower bound Bickel et al. (1998). As Narita (2018) showed, the semiparametric lower bound of (1) under $p_1(a \mid x) = p_2(a \mid x) = \cdots = p_T(a \mid x) = p(a \mid x)$ is given as $\mathbb{E}\left[\left\{\sum_{k=0}^{1}\frac{v\left(k, X_t\right)}{p(k|X_t)} + \left(\theta_0(X_t) - \theta_0\right)^2\right\}\right]$.

2.5 Efficient Policy

We consider minimizing the variance by appropriately optimizing the policy. Following Hahn et al. (2011), the efficient policies for IPW and AIPW estimators are given in the following proposition.

Proposition 1 (Efficient Probability of Assigning an Action)**.** For an IPW estimator, a probability minimizing the variance is given as

$$\pi^{\mathrm{IPW}}(1 \mid X_t) = \frac{\sqrt{e^*(1, X_t)}}{\sqrt{e^*(1, X_t)} + \sqrt{e^*(0, X_t)}}.$$

For an AIPW estimator, a probability minimizing the variance is given as

$$\pi^{\mathrm{AIPW}}(1 \mid X_t) = \frac{\sqrt{v^*(1, X_t)}}{\sqrt{v^*(1, X_t)} + \sqrt{v^*(0, X_t)}}.$$

The derivation of an AIPW estimator is shown in Hahn et al. (2011). For an IPW estimator, we show the proof in Appendix B. In the following sections, we show that using the probability in Proposition 1, which minimizes the variance, we can also minimize the asymptotic variance and upper bound of the concentration inequality of appropriately defined estimators. Because the variance is equal to the MSE, a policy minimizing the variance also minimizes the MSE.

2.6 Related Work

In this section, we review existing studies related to our problem setting.

Adaptive Experimental Design Among various methods for the adaptive experimental design, we share the motivation with Hahn et al. (2011) and van der Laan (2008). To best our knowledge, other existing studies have not pursued this direction of experimental design. Hahn et al. (2011) proposed the two-stage adaptive experimental design. Using the samples in the first stage, they estimated the conditional variance of outcomes to construct the optimal policy that minimizes the asymptotic variance of an estimator of the ATE (Proposition 1). In the second stage, they assigned the treatments to samples following the policy constructed in the first stage. There are three essential differences between them. First, because our method enables us to simultaneously construct the optimal policy and assign a treatment, we do not have to decide the sample size of the first stage in advance. Second, because of this property, our method and sequential testing introduced in Section 4 are compatible. Third, by applying martingale theory as van der Laan & Lendle (2014), we do not require Donsker's condition for the nuisance estimator. Our proposed method is also closely related to the method of van der Laan (2008). Compared with their method, our contributions are a generalization of the estimator using the martingale-based construction of the nuisance estimator and a proposition of nonparametric sequential testing based on concentration inequality. In addition, following Klaassen (1987), van der Laan & Lendle (2014), and Chernozhukov et al. (2018), we point out that Donkser's condition is unnecessary to the nuisance estimator.

ATE Estimation from Dependent Samples Several existing studies offer statistical inference from samples with dependency (van der Laan, 2008; van der Laan & Lendle, 2014; Luedtke & van der Laan, 2016; Hadad et al., 2019). The asymptotic normality of our proposed estimator is a special and simplified version of the estimator proposed by van der Laan & Lendle (2014), but the concentration inequality of the estimator is a new result. Although Hadad et al. (2019) recommended using adaptive weights for stabilization, we consider that we can control the behavior of the assignment probability by ourselves in our setting.

Best Arm Identification: While the common goal of the MAB problem is to maximize the profit obtained from treatments, another framework called the *best arm identification* (BAI) aims to find actions with better rewards, whose motivation is similar to ours. For example, Yang et al. (2017) and Jamieson & Jain (2018) proposed a method to conduct a statistical test to find better actions using as a small sample size as possible. However, the approach is different. In the BAI without covariates, we typically compare the sample average of the rewards of each arm and tries to find an arm whose expected reward is the best among those arms with a high probability. On the contrary, in the best arm identification with covariates, we aim to find an arm whose expected reward conditional on the covariates is the best among the arms with a high probability. The problem setting in this paper shares the same goal as the best arm identification without covariates; however, we can also use the covariate information. In conclusion, the problem setting in this paper can be regarded as a new approach to the BAI. This setting can be called *semiparametric best arm identification*. Similarly, Deshpande et al. (2018) and Yao et al. (2020) consider controlling the power of the test in the MAB problem, but have different motivations with different problem setting.

Ethics and Fairness: While an RCT is a reliable framework for scientific experiments, it has some ethical problems (Nardini, 2014). For example, in clinical trials, the use of placebos concerns the problem of deception. A researcher sometimes prescribes placebos to patients, and patients must be made to believe they are receiving a working treatment, even though they are not, for the placebo effect to play any role at all (Nardini, 2014). Thus, clinical trials are not only costly but also unethical. In addition, simple randomization sometimes obtains unfair results. On the contrary, compared with adaptive randomization based on past observations, an RCT with a completely random assignment might require more samples. Thus, ethics and fairness in RCTs and adaptive experimental design are critical problems. In the proposed algorithm, we allocate the treatment based on the standard deviations. If this seems unfair, we can incorporate some fairness criteria as a constraint into the minimization in Proposition 1. For example, if we place a constraint on the overall treatment probability as $\mathbb{E}[\pi_t(1, X_t)] = p$ for a constant $p > 0$, we can add this constraint when solving the minimization problem of Proposition 1. For another approach, Narita (2018) proposed using mechanism design for designing an RCT.

Other Related Work: Balsubramani & Ramdas (2016) proposed nonparametric sequential testing based on the law of the iterated logarithm (LIL). We apply their results to adaptive ATE estimation with the AIPW estimator. Abhishek & Mannor (2017) consider sequential testing with a complex metric, and the motivation is different from us. Some of the adversarial bandit algorithms also use IPW to obtain an unbiased estimator (Auer et al., 2003), but we have a different motivation.

3 Efficient ATE Estimation

As shown in the previous section, by setting

$$\pi_t(1 \mid x, \Omega_{t-1}) = \pi^{\mathrm{AIPW}}(1 \mid x) = \frac{\sqrt{v^*(1,x)}}{\sqrt{v^*(1,x)} + \sqrt{v^*(0,x)}},$$

we can minimize the variance of the estimators. However, how to conduct statistical inference from the policy is unclear. There are two problems. First, we do not know $v^*(k,x) = \sigma^2(k,x)$. The second problem is how to conduct statistical inference from samples with dependency, which comes from the construction of $\pi_t(1 \mid x, \Omega_{t-1})$ (i.e., the estimation of $v^*(k,x)$). We solve the first problem by estimating $v^*(k,x)$ sequentially. For example, we can estimate $v^*(k,x) = e^*(k,x) - (f^*(k,x))^2$ by estimating $f^*(k,x)$ and $e^*(k,x)$. In this section, for solving the second problem, we propose estimators for samples with dependency and analyze the behavior of the estimators for infinite and finite samples.

3.1 Adaptive Estimators from Samples with Dependency

Here, we define the estimators constructed from samples with dependency. First, we define the adaptive IPW (AdaIPW) estimator as $\hat{\theta}^{\mathrm{AdaIPW}}_T = \frac{1}{T}\sum^T_{t=1}\left(\frac{\mathbb{1}[A_t=1]Y_t}{\pi_t(1|X_t,\Omega_{t-1})} - \frac{\mathbb{1}[A_t=0]Y_t}{\pi_t(0|X_t,\Omega_{t-1})}\right)$. Second, we define the adaptive AIPW (A2IPW) estimator as $\hat{\theta}^{\mathrm{A2IPW}}_T = \frac{1}{T}\sum^T_{t=1} h_t$, where

$$h_t = \left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t)\right).$$

For $z_t = h_t - \theta_0$, $\{z_t\}^T_{t=1}$ is a martingale difference sequence (MDS), that is, $\mathbb{E}\big[z_t \mid \Omega_{t-1}\big] = \theta_0$. Using this property, we derive the theoretical results of $\hat{\theta}^{\mathrm{A2IPW}}_T$ in the following section. We omit the discussion for $\hat{\theta}^{\mathrm{AdaIPW}}_T$, but can derive the theoretical properties as well as $\hat{\theta}^{\mathrm{A2IPW}}_T$.

3.2 Asymptotic Distribution of A2IPW

For the A2IPW estimator $\hat{\theta}^{\mathrm{A2IPW}}_T$, we derive the asymptotic distribution.

Theorem 1 (Asymptotic Distribution of A2IPW). *Suppose that (i) point convergence in probability of $\hat{f}_{t-1}$ and π_t, i.e., for all $x \in \mathcal{X}$ and $k \in \{0, 1\}$, $\hat{f}_{t-1}(k, x) - f^*(k, x) \xrightarrow{\mathrm{p}} 0$ and $\pi_t(k \mid x, \Omega_{t-1}) - \tilde{\pi}(k \mid x) \xrightarrow{\mathrm{p}} 0$, where $\tilde{\pi} : \mathcal{A}\times\mathcal{X} \to (0, 1)$ and (ii) there exits a constant C_3 such that $|\hat{f}_{t-1}| \leq C_3$. TheThen, under Assumption 1, for the A2IPW estimator, we have $\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$, where $\sigma^2 = \mathbb{E}\left[\sum^1_{k=0}\frac{\nu^*\left(k, X_t\right)}{\tilde{\pi}(k|X_t)} + \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2\right]$.*

The proof is shown in Appendix C. This result is a special case of van der Laan & Lendle (2014). Unlike van der Laan & Lendle (2014), we do not impose the convergence rate of $\hat{f}_{t-1}$ owing to the unbiasedness. As well as cross-fitting (Klaassen, 1987; Zheng & van der Laan, 2011; Chernozhukov et al., 2018), we do not have to impose Donsker condition as pointed by van der Laan & Lendle (2014). The asymptotic variance is semiparametric efficient under the policy $\tilde{\pi}$. It can also be regarded as the AMSE defined between θ_0 and $\hat{\theta}^{\mathrm{A2IPW}}_T$. As a corollary, in Appendix E, we show the corresponding estimator and its asymptotic variance for the off-policy evaluation, which is a generalization of the ATE estimation. Finally, we also show the consistency by using the weak law of large numbers for an MDS (Proposition 4 in Appendix A). We omit the proof because we can easily show it from the boundedness of z_t.

Theorem 2 (Consistency of A2IPW). *Suppose that there exits a constant C_3 such that $|\hat{f}_{t-1}| \leq C_3$. Then, under Assumption 1, $\hat{\theta}^{\mathrm{A2IPW}}_T \xrightarrow{\mathrm{p}} \theta_0$.*

3.3 Regret Bound of A2IPW

For the finite sample analysis, instead of asymptotic theory, we introduce the *regret analysis* framework often used in the literature on the MAB problem. In this paper, we define regret based on the MSE. We define the optimal policy Π^{OPT} as a policy that chooses a treatment with the probability π^{AIPW} defined in (1), and an estimator $\hat{\theta}^{\mathrm{OPT}}_T$ with oracle f^* as

$$\hat{\theta}^{\mathrm{OPT}}_T = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{1}[A_t=1]\big(Y_t - f^*(1,X_t)\big)}{\pi^{\mathrm{AIPW}}(1\mid X_t)} - \frac{\mathbb{1}[A_t=0]\big(Y_t - f^*(0,X_t)\big)}{1-\pi^{\mathrm{AIPW}}(1\mid X_t)} + f^*(1,X_t) - f^*(0,X_t)\right).$$

Then, for a policy Π adapted by the experimenter, we define the regret of between Π and Π^{OPT} as $\texttt{regret} = \mathbb{E}_{\Pi}\left[\left(\theta_0 - \hat{\theta}^{\mathrm{A2IPW}}_T\right)^2\right] - \mathbb{E}_{\Pi^{\mathrm{OPT}}}\left[\left(\theta_0 - \hat{\theta}^{\mathrm{OPT}}_T\right)^2\right]$, where the expectations are taken over each policy. The upper bound is in the following theorem.

Theorem 3 (Regret Bound of A2IPW)**.** *Suppose that there exits a constant C_3 such that $|\hat{f}_{t-1}| \le C_3$. Then, under Assumption 1, the* `regret` *is bounded by*

$$\frac{1}{T^2}\sum_{t=1}^{T}\sum_{k=0}^{1}\left\{\mathrm{O}\left(\mathbb{E}\left[\left|\sqrt{\pi^{\mathrm{AIPW}}(k\mid X_t)} - \sqrt{\pi_t(k\mid X_t,\Omega_{t-1})}\right|\right]\right) + \mathrm{O}\left(\mathbb{E}\left[\left|f^*(k,X_t) - \hat{f}_{t-1}(k,X_t)\right|\right]\right)\right\},$$

where the expectation is taken over the random variables including Ω_{t-1}.

The proof is shown in Appendix D. Then, by substituting the finite sample bounds of $\mathbb{E}\left[\left|\sqrt{\pi^{\mathrm{AIPW}}(k\mid X_t)} - \sqrt{\pi_t(k\mid X_t,\Omega_{t-1})}\right|\right]$ and $\mathbb{E}\left[\left|f^*(k,X_t) - \hat{f}_{t-1}(k,X_t)\right|\right]$, the regret bound for finite samples can be obtained. We can bound $\hat{f}_{t-1}(k,X_t)$ and $\sqrt{\pi_t(k\mid X_t,\Omega_{t-1})}$ by the same argument as existing work on the MAB problem such as Yang & Zhu (2002).

Remark 4. This result tells us that regret is bounded by $\mathrm{o}(1/T)$ under the appropriate convergence rates of π_t and $\hat{f}_t$. By contrast, if we use a constant value for π_t, regret is $\mathrm{O}(1/T)$.

4 Sequential Hypothesis Testing with A2IPW Estimator

The goal of various applications including A/B testing is to conduct decision making between *null* ($\mathcal{H}_0$) and an *alternative* ($\mathcal{H}_1$) hypothesis while controlling both *false positives* (*Type I error*) and *false negatives* (*Type II error*). *Standard hypothesis testing* generates a confidence interval based on a fixed sample size T. In this case, we can use the asymptotic distribution derived in Theorem 1. On the contrary, for the case in which samples arrive in a stream, there is interest in conducting decision making without waiting for the sample size to reach T. Under this motivation, we discuss *sequential hypothesis testing*, which decides to accept or reject the null hypothesis at any time $t = 1, 2, \dots, T$. The preliminaries of the hypothesis testing are in Appendix F.

4.1 Sequential Testing and ∂ ç Type I error

In sequential testing, we sequentially conduct decision making and stop whenever we want (Wald, 1945). However, if we sequentially conduct standard hypothesis testing based on the p-value defined for a fixed sample size, the probability of the Type I error increases (Balsubramani & Ramdas, 2016). Therefore, the main issue of sequential testing is to control the Type I error, and various approaches have been proposed (Wald, 1945). One classical method is to correct the p-value based on multiple testing correction, such as the Bonferroni (BF) and Benjamini–Hochberg procedures. For example, when we conduct standard hypothesis testing at $t = 100, 200, 300, 400, 500$ by constructing the corresponding p-values of p_{100}, p_{200}, p_{300}, p_{400}, and p_{500}, the BF procedure corrects the p-values to p_{100}, $p_{200}/2$, $p_{300}/3$, $p_{400}/4$, and $p_{500}/5$. Although this correction enables us to control the Type I error, it is also known to be exceedingly conservative and tends to produce suboptimal results (Balsubramani & Ramdas, 2016; Jamieson & Jain, 2018). Further, owing to this conservativeness, we cannot conduct decision making in each period. For example, in the case in which we conduct standard hypothesis testing in period $t = 1, 2, 3, \dots, t, \dots$, the corresponding p-values become too small $(p_1, p_2/2, p_3/3, p_4/4, \dots, p_t/t, \dots)$. Therefore, when conducting sequential testing based on multiple testing, we need to split the stream of samples into several batches (Balsubramani & Ramdas, 2016). To avoid the drawback of multiple testing, recent work has proposed using *adaptive concentration inequalities* for an adaptively chosen number of samples (i.e., the inequality holds at any randomly chosen $t = 1, 2, \dots$) (Balsubramani, 2014;

Jamieson et al., 2014; Johari et al., 2015; Balsubramani & Ramdas, 2016; Zhao et al., 2016; Jamieson & Jain, 2018). This concentration inequality enables us to conduct sequential testing without separating samples into batches while controlling the Type I error under appropriate conditions.

There are two approaches for introducing such concentration inequalities into sequential testing: *confidence sequence* (Darling & Robbins, 1967; Lai, 1984; Zhao et al., 2016) and *always valid p-values* (Johari et al., 2015; Jamieson & Jain, 2018). These two approaches are equivalent, as shown by Ramdas (2018), and we adapt the former herein. For simplicity, let us define the null and alternative hypotheses as $\mathcal{H}_0 : \theta_0 = \mu$ and $\mathcal{H}_1 : \theta_0 \neq \mu$, respectively, where μ is a constant, and consider controlling the Type I error at α. Then, for the A2IPW estimator $\hat{\theta}_t^{\mathrm{A2IPW}}$ of θ_0, we define a sequence of positive values $\{q_t\}_{t=1}^{T}$, which satisfies $\mathbb{P}(\exists t \in \mathbb{N} : t\hat{\theta}_t^{\mathrm{A2IPW}} - t\mu > q_t) \leq \alpha$ when the null hypothesis is true. Using $\{q_t\}_{t=1}^{T}$, we consider the following process: if $t\hat{\theta}_t^{\mathrm{A2IPW}} - t\mu > q_t$, we reject the null hypothesis $\mathcal{H}_0$; otherwise, we temporally accept the null hypothesis $\mathcal{H}_0$. Because $\{q_t\}_{t\in\mathbb{N}}$ satisfies $\mathbb{P}\big(\text{reject } \mathcal{H}_0\big) = \mathbb{P}\left(\exists t \in \mathbb{N} : |t\hat{\theta}_t^{\mathrm{A2IPW}} - t\mu| > q_t\right) \leq \alpha$ when the null hypothesis is true, we can control the Type I error at α. This procedure of hypothesis testing has some desirable properties. First, it controls the Type I error with α in any period t. Second, the Type II error of the hypothesis testing with this procedure is less than or equal to that under standard hypothesis testing (Balsubramani & Ramdas, 2016). Third, it enables us to stop the experiment whenever we obtain sufficient samples for decision making.

4.2 Sequential Testing with LIL

Next, we consider constructing $\{q_t\}_{t\in\mathbb{N}}$ with the Type I error α using the proposed A2IPW estimator. Among the various candidates, concentration inequalities based on the LIL have garnered attention recently. The LIL was originally derived as a asymptotic property of independent random variables by Khintchine (1924) and Kolmogoroff (1929). Following their methods, several works have derived an asymptotic LIL for an MDS under some regularity conditions (Stout, 1970; Fisher, 1992), and Balsubramani & Ramdas (2016) derived a nonasymptotic LIL-based concentration inequality for hypothesis testing. The reason for using the LIL-based concentration inequality is that sequential testing with the LIL-based confidence sequence $\{q_t\}_{t\in\mathbb{N}}$ requires the smallest sample size needed to identify the parameter of interest (Jamieson et al., 2014; Balsubramani & Ramdas, 2016). For this tightness of the inequality, LIL-based concentration inequalities have been widely accepted in sequential testing (Balsubramani & Ramdas, 2016) and in the best arm identification in the MAB problem (Jamieson et al., 2014; Jamieson & Jain, 2018). Therefore, we also construct the confidence sequence $\{q_t\}_{t\in\mathbb{N}}$ based on the LIL-based concentration inequality for the A2IPW estimator as follows.

Theorem 4 (Concentration Inequality of A2IPW)**.** *Suppose that there exists C such that $|z_t| \leq C$. Suppose that there exists C_4 such that $|(z_t - z_{t-1})^2 - \mathbb{E}[(z_t - z_{t-1})^2 \mid \Omega_{t-1}]| \leq C_4$. For any δ, with probability $\geq 1 - \delta$, for all $t \geq \tau_0$ simultaneously,*

$$\left|\sum_{i=1}^{t} z_i\right| = \left|t\hat{\theta}_t^{\mathrm{A2IPW}} - t\theta_0\right| \leq \frac{2C}{e^2}\left(C_0(\delta) + \sqrt{2C_1\hat{V}_t^{*}\left(\log\log\hat{V}_t^{*} + \log\left(\frac{4}{\delta}\right)\right)}\right),$$

where $\hat{V}_t^{*} = C_3\left(\frac{e^4}{4C^2}\sum_{i=1}^{t} z_i^2 + \frac{2C_0(\delta)C_4}{e^2}\right)$, $C_0(\delta) = 3(e-2) + 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right)$, $C_1 = 6(e-2)$ and C_3 is an absolute constant.

We can obtain this result by applying the result of Balsubramani (2014). The proof is in Appendix D.1. Then, we obtain confidence sequences, $\{q_t\}_{t=1}^{T}$, with the Type I error at α from the results of Theorem 4 and Balsubramani & Ramdas (2016) as $q_t \propto \log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum_{i=1}^{t} z_i^2\left(\log\frac{\log\sum_{i=1}^{t} z_i^2}{\alpha}\right)}$. Balsubramani & Ramdas (2016) proposed using the constant 1.1 to specify q_t, namely, $q_t = 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum_{i=1}^{t} z_i^2\left(\log\frac{\log\sum_{i=1}^{t} z_i^2}{\alpha}\right)}\right)$. This choice is motivated by the asymptotic property of the LIL such that $\limsup_{t\to\infty}\frac{\left|t\hat{\theta}_t^{\mathrm{A2IPW}} - t\theta_0\right|}{\sqrt{2\tilde{V}_t^{*}\left(\log\log\tilde{V}_t^{*}\right)}} = 1$ with probability 1 for sufficiently large samples (Stout, 1970; Balsubramani & Ramdas, 2016), where $\tilde{V}_t^2 = \sum_{i=1}^{t}\mathbb{E}[z_i^2 \mid \Omega_{i-1}]$, and the empirical results of Balsubramani & Ramdas (2016).

Table 1: Experimental results using Datasets 1–2. The best performing method is in bold.

	Dataset 1: $\mathbb{E}[Y(1)] = 0.8$, $\mathbb{E}[Y(0)] = 0.3$, $\theta_0 \neq 0$								Dataset 2: $\mathbb{E}[Y(1)] = 0.5$, $\mathbb{E}[Y(0)] = 0.5$, $\theta_0 = 0$							
	$T = 150$			$T = 300$			ST		$T = 150$			$T = 300$			ST	
	MSE	STD	Testing	MSE	STD	Testing	LIL	BF	MSE	STD	Testing	MSE	STD	Testing	LIL	BF
RCT	0.145	0.178	25.0%	0.073	0.100	46.0%	455.4	370.4	0.084	0.129	4.7%	0.044	0.062	4.9%	497.2	481.8
A2IPW (K-nn)	0.085	0.116	38.4%	0.038	0.054	67.9%	389.5	302.8	0.050	0.071	5.6%	0.026	0.037	5.6%	497.2	477.3
A2IPW (NW)	0.064	0.092	51.4%	0.025	0.035	88.1%	303.8	239.8	**0.029**	0.045	**4.4%**	**0.012**	0.018	4.7%	496.2	480.6
MA2IPW (K-nn)	0.092	0.126	38.5%	0.044	0.058	66.2%	387.5	303.4	0.052	0.073	5.4%	0.025	0.034	4.7%	497.9	477.0
MA2IPW (NW)	**0.062**	0.085	52.7%	**0.023**	0.033	90.2%	303.3	236.6	0.032	0.047	6.3%	**0.012**	0.018	4.4%	496.6	475.3
AdaIPW (K-nn)	0.151	0.208	26.1%	0.075	0.103	43.6%	446.3	367.0	0.088	0.126	5.6%	0.043	0.062	5.2%	495.8	478.1
AdaIPW (NW)	0.161	0.232	23.4%	0.081	0.115	41.1%	446.6	375.0	0.094	0.140	5.8%	0.045	0.064	5.3%	495.6	471.6
DM (K-nn)	0.175	0.252	**88.7%**	0.086	0.126	**96.1%**	**59.9**	**164.6**	0.096	0.129	85.3%	0.046	0.063	89.5%	97.3	188.3
DM (NW)	0.111	0.167	82.1%	0.045	0.066	95.6%	119.6	176.2	0.054	0.075	53.7%	0.023	0.032	55.4%	312.8	305.3
Hahn 50 (K-nn)	0.109	0.149	35.2%	0.046	0.064	63.3%	398.5	316.0	0.060	0.089	5.4%	0.029	0.041	6.6%	493.8	473.4
Hahn 50 (NW)	0.085	0.128	45.7%	0.033	0.046	82.8%	313.1	257.0	0.040	0.057	5.6%	0.016	0.025	6.9%	493.7	477.7
Hahn 100 (K-nn)	0.141	0.200	29.6%	0.057	0.081	60.%	408.2	332.6	0.071	0.104	6.3%	0.029	0.044	5.2%	495.2	475.6
Hahn 100 (NW)	0.107	0.146	32.1%	0.036	0.050	75.2%	365.3	294.6	0.043	0.063	4.8%	0.014	0.019	**3.7%**	**498.2**	**483.5**
OPT	0.008	0.011	100.0%	0.004	0.005	100.0%	63.9	150.0	0.005	0.007	4.4%	0.002	0.003	4.4%	498.4	483.0

5 Main Algorithm: AERATE

In this section, we define our main algorithm, referred to as *Adaptive ExpeRiments for efficient ATE estimation* (AERATE). The details are in Appendix G.

First, we consider estimating $f^*(a, x) = \mathbb{E}\big[Y_t(a) \mid x\big]$ and $e^*(a, x) = \mathbb{E}\big[Y_t^2(a) \mid x\big]$. When estimating $f^*(a, x)$ and $e^*(a, x)$, we need to construct consistent estimators from dependent samples obtained from an adaptive policy. In a MAB problem, several nonparametric estimators are consistent, such as the K-nearest neighbor regression estimator and Nadaraya–Watson kernel regression estimator (Yang & Zhu, 2002; Qian & Yang, 2016).

For simplicity, we only show the algorithm using A2IPW, and we can derive the procedure when using the AdaIPW estimator similarly. The proposed algorithm consists of three main steps: in period t, (i) estimate $\nu(k, x)$ using nonparametric estimators in the MAB problem (Yang & Zhu, 2002; Qian & Yang, 2016); (ii) assign an action with an estimator of the optimal policy, which is defined as $\pi^{\mathrm{A2IPW}}(1 \mid x) = \frac{\sqrt{\nu^*(1,x)}}{\sqrt{\nu^*(1,x)}+\sqrt{\nu^*(0,x)}}$; and (iii) conduct testing when sequential testing is chosen as the hypothesis testing method. Moreover, to stabilize the algorithm, we introduce the following three elements: (a) the estimator $\hat{\nu}_{t-1}(k, x)$ of $\nu^*(k, x)$ is constructed as $\max\big(\underline{\nu}, \hat{e}_{t-1}(k, x) - \hat{f}^2_{t-1}(k, x)\big)$, where $\underline{\nu}$ is the lower bound of ν^*, and $\hat{f}_{t-1}$ and $\hat{e}_{t-1}$ are the estimators of f^* and e^* only using Ω_{t-1}, respectively; (b) let a policy be $\pi_t(1 \mid x, \Omega_{t-1}) = \gamma\frac{1}{2} + (1-\gamma)\frac{\sqrt{\hat{\nu}_{t-1}(1,x)}}{\sqrt{\hat{\nu}_{t-1}(1,x)}+\sqrt{\hat{\nu}_{t-1}(0,x)}}$, where $\gamma = \mathrm{O}(1/\sqrt{T})$; and (c) as a candidate of the estimators, we also propose the mixed A2IPW (MA2IPW) estimator defined as $\hat{\theta}^{\mathrm{MA2IPW}}_t = \zeta\hat{\theta}^{\mathrm{AdaIPW}}_t + (1-\zeta)\hat{\theta}^{\mathrm{A2IPW}}_t$, where $\zeta = \mathrm{o}(1/\sqrt{t})$. The motivation of (a) is to prevent $\hat{\nu}_{t-1}$ from taking a negative value or zero technically, and we do not require accurate knowledge of the lower bound. The motivation of (b) is to stabilize the probability of assigning an action. The motivation of (c) is to control the behavior of an estimator by avoiding the situation in which $\hat{f}_{t-1}$ takes an unpredicted value in the early stage. Because the nonparametric convergence rate is lower bounded by $\mathrm{O}(1/\sqrt{t})$ in general, the convergence rate of π_t to π^{AIPW} is also upper bounded by $\mathrm{O}(1/\sqrt{t})$. Therefore, $\gamma = \mathrm{O}(1/\sqrt{t})$ does not affect the convergence rate of the policy. Similarly, the asymptotic distribution of $\hat{\theta}^{\mathrm{MA2IPW}}_T$ is the same as $\hat{\theta}^{\mathrm{A2IPW}}_T$. The pseudo code is in Appendix G.

6 Experiments

In this section, we show the effectiveness of the proposed algorithm experimentally. We compare the proposed AdaIPW, A2IPW, and MA2IPW estimators in AERATE with an RCT with $p(A_t = 1|X_t) = 0.5$, the method of Hahn et al. (2011), the estimator $\hat{\theta}^{\mathrm{OPT}}_T$ under the optimal policy, and the standard DM estimators. To best our knowledge, there is no recent method proposed in this problem setting. In AERATE, we set $\gamma = 1/\sqrt{t}$. For the MA2IPW estimator, we set $\zeta = t^{-1/1.5}$. When estimating f^* and e^*, we use K-nearest neighbor regression and Nadaraya–Watson regression. In the method of Hahn et al. (2011), we first use 50 and 100 samples to estimate the optimal policy. In this experiment, we use synthetic and semi-synthetic datasets. In each dataset, we conduct the following three patterns of hypothesis testing. For all the settings, the null and alternative hypotheses are $\mathcal{H}_0 : \theta_0 = 0$ and $\mathcal{H}_1 : \theta_0 \neq 0$, respectively. We conduct standard hypothesis testing with T-statistics when the sample sizes are 250 and 500, sequential testing based on multiple testing with the BF correction when the sample sizes are 150, 250, 350, and 450, and sequential testing with the LIL based on the concentration inequality shown in Theorem 4.

First, we conducted an experiment using the following synthetic datasets. We generated a covariate $X_t \in \mathbb{R}^5$ at each round as $X_t = (X_{t1}, X_{t2}, X_{t3}, X_{t4}, X_{t5})^\top$, where $X_{tk} \sim \mathcal{N}(0, 1)$ for $k = 1, 2, 3, 4, 5$. In this experiment, we used $Y_t(d) = \mu_d + \sum_{k=1}^{5} X_{tk} + e_{td}$ as a model of a potential outcome, where μ_d is a constant, e_{td} is the error term, and $\mathbb{E}[Y_t(d)] = \mu_d$ The error term e_{td} follows the normal distribution, and we denote the standard deviation as std_d. We made two datasets with different μ_d and std_d, Datasets 1–2, with 500 periods (samples). For Datasets 1, we set $\mu_1 = 0.8$ and $\mu_0 = 0.3$ with $\mathrm{std}_1 = 0.8$ and $\mathrm{std}_1 = 0.3$. For Datasets 1, we set $\mu_1 = \mu_0 = 0.5$ with $\mathrm{std}_1 = 0.8$ and $\mathrm{std}_1 = 0.3$. We ran 1000 independent trials for each setting. The results of experiment are shown in Table 1. We show the MSE between θ and $\hat{\theta}$, the standard deviation of MSE (STD), and percentages of rejections of hypothesis testing using T-statistics at the 150th (mid) round and the 300th (final) periods. Besides, we also showed the stopping time of the LIL based algorithm (LIL) and multiple testing with BF correction. When using BF correction, we conducted testing at $t = 150, 250, 350, 450$. In sequential testing, if we do not reject the hypothesis, we return the stopping time as 500. In many datasets, the proposed algorithm achieves the lower MSE than an the other methods. The DM estimators rejects the null hypothesis with small samples in Dataset 1, but also often reject the null hypothesis in Dataset II, i.e, the Type II error is large. The details are shown in Appendix H.

Appendix H shows the additional experimental results. In Appendix H, we investigate the performance of the proposed algorithm for other synthetic and semi-synthetic datasets constructed from the Infant Health and Development Program (IHDP). The IHDP dataset consists of simulated outcomes and covariate data from a real study following the simulation proposed by Hill (2011). In the IHDP data, we can reduce the sample size by $1/5$ compared with the RCT by using the proposed methods.

7 Conclusion

In this paper, we proposed an algorithm of the MAB problem that yields an efficient estimator of the treatment effect. Using martingale theory, we derived the theoretical properties of the proposed algorithm for cases with both infinite and finite samples for the framework of sequential testing. As the experimental results imply, the proposed methods potentially decrease the sample size for scientific experiments. Because each standard hypothesis testing and sequential testing have different properties, we should choose an appropriate one for each application.

References

Abhishek, V. and Mannor, S. A nonparametric sequential test for online randomized experiments. In *WWW*, 2017.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Math. J. (2)*, 19(3):357–367, 1967.

Balsubramani, A. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

Balsubramani, A. and Ramdas, A. Sequential nonparametric testing with the law of the iterated logarithm. In *UAI*, pp. 42–51. AUAI Press, 2016.

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

Casella, G. *Statistical inference*. Duxbury advanced series. Duxbury/Thomson Learning, Australia ; Pacific Grove, Calif., 2nd ed. edition, 2002.

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*. Chapman and Hall/CRC, 2 edition, 2011.

Chow SC, Chang M, P. A. Statistical consideration of adaptive methods in clinical development. *J Biopharm Stat*, 2005.

Darling, D. A. and Robbins, H. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66–68, 1967.

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. Accurate inference for adaptive linear models. In *ICML*, pp. 1194–1203, 2018.

Durrett, R. *Probability: Theory and Examples*. Cambridge University Press, USA, 4th edition, 2010.

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.

Fisher, E. On the law of the iterated logarithm for martingales. *The Annals of Probability*, 20(2):675–680, 1992.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*, 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. and Hayde, C. *Martingale Limit Theory and Its Application*. Probability and mathematical statistics. Academic Press, 1980.

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, Princeton, NJ, 1994.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

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.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

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.

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

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *COLT*, volume 35, 2014.

Jamieson, K. G. and Jain, L. A bandit approach to sequential experimental design with false discovery control. In *NeurIPS*, pp. 3664–3674. Curran Associates, Inc., 2018.

Johari, R., Pekelis, L., and Walsh, D. J. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

Kendall, J. M. Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal*, 20(2):164–168, 2003.

Khintchine, A. Über einen satz der wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae*, 6(1):9–20, 1924.

Klaassen, C. A. J. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, 1987.

Kolmogoroff, A. Über das gesetz des iterierten logarithmus. *Mathematische Annalen*, 101:126–135, 1929.

Komiyama, O., Koshimizu, T., Suganami, H., Sakai, H., Orhihashi, Y., and Tomiyama, H. Adaptive designs in clinical drug development: An executive summary of the phrma working group. *Rinsho yakuri/Japanese Journal of Clinical Pharmacology and Therapeutics*, 40(6):303–310, 2009.

Lai, T. Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach. *Communications in Statistics - Theory and Methods*, 13(19):2355–2368, 1984.

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*, 2016.

Nardini, C. The ethics of clinical trials. *Ecancermedicalscience*, 8:387, 2014.

Narita, Y. Experiment-as-Market: Incorporating Welfare into Randomized Controlled Trials. Cowles Foundation Discussion Papers 2127, Cowles Foundation for Research in Economics, Yale University, 2018.

Qian, W. and Yang, Y. Kernel estimation and model combination in a bandit problem with covariates. *Journal of Machine Learning Research*, 17(149):1–37, 2016.

Ramdas, A. Sequential testing, always valid p-values. *Martingales 2 : Sequential Analysis*, 2018.

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. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Rubin, D. B. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

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

Stout, W. F. A martingale analogue of kolmogorov's law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 15(4):279–290, 1970.

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

Tomking, R. J. Some iterated logarithm results related to the central limit theorem. *Transactions of the American Mathematical Society*, 156, 1971.

van der Laan, M. J. The construction and analysis of adaptive group sequential designs. 2008.

van der Laan, M. J. and Lendle, S. D. Online targeted learning. 2014.

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

Villar, S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30:199–215, 2015.

Wald, A. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 1945.

Yang, F., Ramdas, A., Jamieson, K. G., and Wainwright, M. J. A framework for multi-a(rmed)/b(andit) testing with online fdr control. In *NeurIPS*, pp. 5957–5966. Curran Associates, Inc., 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, 2002.

Yao, J., Brunskill, E., Pan, W., Murphy, S., and Doshi-Velez, F. Power-constrained bandits, 2020.

Zhao, S., Zhou, E., Sabharwal, A., and Ermon, S. Adaptive concentration inequalities for sequential decision problems. In *NeurIPS*, pp. 1343–1351. Curran Associates, Inc., 2016.

Zheng, W. and van der Laan, M. J. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics. 2011.

A Preliminaries

A.1 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 2. [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 3 (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.

A.2 Martingale Limit Theorems

Proposition 4. [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 5. The weak law of large numbers for martingale holds when the random variable is bounded by a constant.

Proposition 5. [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 Proof of Proposition 1

Proof. Let $\mathcal{P}$ be a function class of $p : \mathcal{X} \to (0, 1)$, and let us define the following function $b : \mathcal{P} \to \mathbb{R}$:

$$b(p) = \mathbb{E}\left[\frac{e(1, X_t)}{b(X_t)}\right] + \mathbb{E}\left[\frac{e(0, X_t)}{1 - b(X_t)}\right].$$

Here, we rewrite $b(p)$ as follows:

$$b(p) = \mathbb{E}\left[\mathbb{E}\left[\frac{e(1, X_t)}{p(X_t)} + \frac{e(0, X_t)}{1 - p(X_t)}\middle| X_t\right]\right].$$

We consider minimizing $b(p)$ by minimizing $\tilde{b}(q) = \mathbb{E}\left[\frac{e(1,X_t)}{q} + \frac{e(0,X_t)}{1-q}\middle| X_t\right]$ for $q \in [\varepsilon, 1-\varepsilon]$. The first derivative of $\tilde{b}(q)$ with respect to q is given as follows:

$$\tilde{b}'(q) = -\frac{e(1,X_t)}{q^2} + \frac{e(0,X_t)}{(1-q)^2}.$$

The second derivative of f is given as follows:

$$\tilde{b}''(q) = 2\frac{e(1,X_t)}{q^3} + 2\frac{e(0,X_t)}{(1-q)^3}.$$

For $\varepsilon < q < 1-\varepsilon$, because $\tilde{b}''(q) > 0$, the minimizer q^* of $\tilde{b}$ satisfies the following equation:

$$-\frac{e(1,X_t)}{(q^*)^2} + \frac{e(0,X_t)}{(1-q^*)^2} = 0.$$

This equation is equivalent to

$$\begin{aligned}
&-(q^*)^2 e(0,X_t) + (1-q^*)^2 e(1,X_t) = 0\\
&\Leftrightarrow\ q^*\sqrt{e(0,X_t)} = (1-q^*)\sqrt{e(1,X_t)}\\
&\Leftrightarrow\ q^* = \frac{\sqrt{e(1,X_t)}}{\sqrt{e(1,X_t)} + \sqrt{e(0,X_t)}}.
\end{aligned}$$

Therefore,

$$b^{\mathrm{OPT}}(D=1|X_t) = \frac{\sqrt{e(1,X_t)}}{\sqrt{e(1,X_t)} + \sqrt{e(0,X_t)}}.$$

□

C Proof of Theorem 1

Proof. Note that the estimator is given as follows:

$$\hat{\theta}^{\mathrm{A2IPW}}_T = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{1}[A_t=1]\big(Y_t - \hat{f}_{t-1}(1,X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t=0]\big(Y_t - \hat{f}_{t-1}(0,X_t)\big)}{\pi_t(0 \mid x, \Omega_{t-1})} + \hat{f}_{t-1}(1,X_t) - \hat{f}_{t-1}(0,X_t)\right).$$

Let us note that z_t is defined as

$$\frac{\mathbb{1}[A_t=1]\big(Y_t - \hat{f}_{t-1}(1,X_t)\big)}{\pi_t(1 \mid x, \Omega_{t-1})} - \frac{\mathbb{1}[A_t=0]\big(Y_t - \hat{f}_{t-1}(0,X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1,X_t) - \hat{f}_{t-1}(0,X_t) - \theta_0.$$

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

$$\begin{aligned}
&\mathbb{E}\big[z_t \mid \Omega_{t-1}\big]\\
&= \mathbb{E}\left[\frac{\mathbb{1}[A_t=1]\big(Y_t - \hat{f}_{t-1}(1,X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t=k]\big(Y_t - \hat{f}_{t-1}(0,X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(0,X_t) - \hat{f}_{t-1}(0,X_t) - \theta_0 \mid \Omega_{t-1}\right]\\
&= \mathbb{E}\Bigg[\hat{f}_{t-1}(1,X_t) - \hat{f}_{t-1}(0,X_t) - \theta_0\\
&\qquad + \mathbb{E}\left[\frac{\mathbb{1}[A_t=1]\big(Y_t - \hat{f}_{t-1}(1,X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t=0]\big(Y_t - \hat{f}_{t-1}(0,X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} \mid X_t, \Omega_{t-1}\right] \mid \Omega_{t-1}\Bigg]\\
&= \mathbb{E}\Big[\hat{f}_{t-1}(1,X_t) - \hat{f}_{t-1}(0,X_t) - \theta_0 + f^*(1,X_t) - f^*(0,X_t) - \hat{f}_{t-1}(1,X_t) + \hat{f}_{t-1}(0,X_t) \mid \Omega_{t-1}\Big] = 0.
\end{aligned}$$

Therefore, to derive the asymptotic distribution, we consider applying the CLT for an MDS introduced in Proposition 5. There are the 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

$$\begin{aligned}
&\mathbb{E}\big[z_t^2\big]\\
&= \mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\right]\\
&= \mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\right]\\
&\quad - \mathbb{E}\left[\sum_{k=0}^{1} \frac{v\big(k, X_t\big)}{\tilde{\pi}(k \mid X_t)} + \Big(\theta_0(X_t) - \theta_0\Big)^2\right] + \mathbb{E}\left[\sum_{k=0}^{1} \frac{v\big(k, X_t\big)}{\tilde{\pi}(k \mid X_t)} + \Big(\theta_0(X_t) - \theta_0\Big)^2\right].
\end{aligned}$$

Therefore, we prove that the RHS of the following equation varnishes asymptotically to show that the condition (a) holds.

$$\begin{aligned}
&\mathbb{E}\big[z_t^2\big] - \mathbb{E}\left[\sum_{k=0}^{1} \frac{v\big(k, X_t\big)}{\tilde{\pi}(k \mid X_t)} + \Big(\theta_0(X_t) - \theta_0\Big)^2\right]\\
&= \mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\right]\\
&\quad - \mathbb{E}\left[\sum_{k=0}^{1} \frac{v\big(k, X_t\big)}{\tilde{\pi}(k \mid X_t)} + \Big(\theta_0(X_t) - \theta_0\Big)^2\right]. \qquad (2)
\end{aligned}$$

First, for the first term of the RHS of (2),

$$\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\right]$$

$$= \mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})}\right)^2\right]$$

$$+ \mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})}\right)^2\right]$$

$$+ \mathbb{E}\left[\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\right]$$

$$- 2\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})}\right)\left(\frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})}\right)\right]$$

$$+ 2\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})}\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\right]$$

$$- 2\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})}\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\right].$$

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

$$\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = k]\big(Y_t - \hat{f}_{t-1}(k, X_t)\big)}{\pi_t(k \mid X_t, \Omega_{t-1})}\right)^2\right] = \mathbb{E}\left[\frac{\big(Y_t(k) - \hat{f}_{t-1}(k, X_t)\big)^2}{\pi_t(k \mid X_t, \Omega_{t-1})}\right],$$

$$\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})}\right)\left(\frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})}\right)\right] = 0,$$

$$\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})}\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} \mid X_t, \Omega_{t-1}\right]\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\right]$$

$$= \mathbb{E}\left[\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\right].$$

Therefore, for the first term of the RHS of (2),

$$\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\right]$$

$$= \mathbb{E}\Bigg[\frac{\big(Y_t(1) - \hat{f}_{t-1}(1, X_t)\big)^2}{\pi_t(1 \mid X_t, \Omega_{t-1})} + \frac{\big(Y_t(0) - \hat{f}_{t-1}(0, X_t)\big)^2}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2$$

$$+ 2\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\Bigg].$$

and, for the second term of the RHS of (2),

$$\mathbb{E}\left[\sum_{k=0}^{1}\frac{v(k, X_t)}{\tilde{\pi}(k \mid X_t)} + \left(\theta_0(X_t) - \theta_0\right)^2\right]$$

$$= \mathbb{E}\left[\frac{\big(Y_t(1) - f^*(1, X_t)\big)^2}{\tilde{\pi}(1 \mid X_t)} + \frac{\big(Y_t(0) - f^*(0, X_t)\big)^2}{\tilde{\pi}(0 \mid X_t)} + \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2\right].$$

Then, using these equations, the RHS of (2) can be calculated as

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{\mathbb{1}[A_t=1]\big(Y_t-\hat{f}_{t-1}(1,X_t)\big)}{\pi_t(1\mid X_t,\Omega_{t-1})}-\frac{\mathbb{1}[A_t=0]\big(Y_t-\hat{f}_{t-1}(0,X_t)\big)}{\pi_t(0\mid X_t,\Omega_{t-1})}+\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\right)^2\right]\\
&\qquad-\mathbb{E}\left[\sum_{k=0}^{1}\frac{v\big(k,X_t\big)}{\tilde{\pi}(k\mid X_t)}+\Big(\theta_0(X_t)-\theta_0\Big)^2\right]\\
&=\mathbb{E}\Bigg[\frac{\big(Y_t(1)-\hat{f}_{t-1}(1,X_t)\big)^2}{\pi_t(1\mid X_t,\Omega_{t-1})}+\frac{\big(Y_t(0)-\hat{f}_{t-1}(0,X_t)\big)^2}{\pi_t(0\mid X_t,\Omega_{t-1})}+\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)^2\\
&\quad+2\Big(f^*(1,X_t)-f^*(0,X_t)-\hat{f}_{t-1}(1,X_t)+\hat{f}_{t-1}(0,X_t)\Big)\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)\Bigg]\\
&\quad-\mathbb{E}\Bigg[\frac{\big(Y_t(1)-f^*(1,X_t)\big)^2}{\tilde{\pi}(1\mid X_t)}+\frac{\big(Y_t(0)-f^*(0,X_t)\big)^2}{\tilde{\pi}(0\mid X_t)}+\Big(f^*(1,X_t)-f^*(0,X_t)-\theta_0\Big)^2\Bigg].
\end{aligned}
$$

By taking the absolute value, we can bound the RHS as

$$
\begin{aligned}
&\mathbb{E}\Bigg[\frac{\big(Y_t(1)-\hat{f}_{t-1}(1,X_t)\big)^2}{\pi_t(1\mid X_t,\Omega_{t-1})}+\frac{\big(Y_t(0)-\hat{f}_{t-1}(0,X_t)\big)^2}{\pi_t(0\mid X_t,\Omega_{t-1})}+\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)^2\\
&\quad+2\Big(f^*(1,X_t)-f^*(0,X_t)-\hat{f}_{t-1}(1,X_t)+\hat{f}_{t-1}(0,X_t)\Big)\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)\Bigg]\\
&\quad-\mathbb{E}\Bigg[\frac{\big(Y_t(1)-f^*(1,X_t)\big)^2}{\tilde{\pi}(1\mid X_t)}+\frac{\big(Y_t(0)-f^*(0,X_t)\big)^2}{\tilde{\pi}(0\mid X_t)}+\Big(f^*(1,X_t)-f^*(0,X_t)-\theta_0\Big)^2\Bigg]\\
&\le\mathbb{E}\Bigg[\Bigg|\Bigg\{\frac{\big(Y_t(1)-\hat{f}_{t-1}(1,X_t)\big)^2}{\pi_t(1\mid X_t,\Omega_{t-1})}+\frac{\big(Y_t(0)-\hat{f}_{t-1}(0,X_t)\big)^2}{\pi_t(0\mid X_t,\Omega_{t-1})}+\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)^2\\
&\quad+2\Big(f^*(1,X_t)-f^*(0,X_t)-\hat{f}_{t-1}(1,X_t)+\hat{f}_{t-1}(0,X_t)\Big)\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)\Bigg\}\\
&\quad-\Bigg\{\frac{\big(Y_t(1)-f^*(1,X_t)\big)^2}{\tilde{\pi}(1\mid X_t)}+\frac{\big(Y_t(0)-f^*(0,X_t)\big)^2}{\tilde{\pi}(0\mid X_t)}+\Big(f^*(1,X_t)-f^*(0,X_t)-\theta_0\Big)^2\Bigg\}\Bigg|\Bigg].
\end{aligned}
$$

Then, from the triangle inequality, we have

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Bigg|\Bigg\{\frac{\big(Y_t(1)-\hat{f}_{t-1}(1,X_t)\big)^2}{\pi_t(1\mid X_t,\Omega_{t-1})}+\frac{\big(Y_t(0)-\hat{f}_{t-1}(0,X_t)\big)^2}{\pi_t(0\mid X_t,\Omega_{t-1})}+\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)^2\\
&\quad+2\Big(f^*(1,X_t)-f^*(0,X_t)-\hat{f}_{t-1}(1,X_t)+\hat{f}_{t-1}(0,X_t)\Big)\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)\Bigg\}\\
&\quad-\Bigg\{\frac{\big(Y_t(1)-f^*(1,X_t)\big)^2}{\tilde{\pi}(1\mid X_t)}+\frac{\big(Y_t(0)-f^*(0,X_t)\big)^2}{\tilde{\pi}(0\mid X_t)}+\Big(f^*(1,X_t)-f^*(0,X_t)-\theta_0\Big)^2\Bigg\}\Bigg|\Bigg]\\
&\le\sum_{k=0}^{1}\mathbb{E}\left[\left|\frac{\big(Y_t(k)-\hat{f}_{t-1}(k,X_t)\big)^2}{\pi_t(k\mid X_t,\Omega_{t-1})}-\frac{\big(Y_t(k)-f^*(k,X_t)\big)^2}{\tilde{\pi}(k\mid X_t)}\right|\right]\\
&\quad+\mathbb{E}\left[\left|\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)^2-\Big(f^*(1,X_t)-f^*(0,X_t)-\theta_0\Big)^2\right|\right]\\
&\quad+2\mathbb{E}\left[\left|\Big(f^*(1,X_t)-f^*(0,X_t)-\hat{f}_{t-1}(1,X_t)+\hat{f}_{t-1}(0,X_t)\Big)\Big(\hat{f}_{t-1}(1,X_t)-\hat{f}_{t-1}(0,X_t)-\theta_0\Big)\right|\right].
\end{aligned}
$$

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_{k=0}^{1}\mathbb{E}\left[\left|\frac{\big(Y_t(k) - \hat{f}_{t-1}(k, X_t)\big)^2}{\pi_t(k \mid X_t, \Omega_{t-1})} - \frac{\big(Y_t(k) - f^*(k, X_t)\big)^2}{\tilde{\pi}(k \mid X_t)}\right|\right]\\
&\quad + \mathbb{E}\left[\left|\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)^2 - \Big(f^*(1, X_t) - f^*(0, X_t) - \theta_0\Big)^2\right|\right]\\
&\quad + 2\mathbb{E}\left[\left|\Big(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\Big)\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)\right|\right]\\
&\leq \tilde{C}_0\sum_{k=0}^{1}\mathbb{E}\left[\left|\frac{\big(Y_t(k) - \hat{f}_{t-1}(k, X_t)\big)}{\sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}} - \frac{\big(Y_t(k) - f^*(k, X_t)\big)}{\sqrt{\tilde{\pi}(k \mid X_t)}}\right|\right]\\
&\quad + \mathbb{E}\left[\left|\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)^2 - \Big(f^*(1, X_t) - f^*(0, X_t) - \theta_0\Big)^2\right|\right]\\
&\quad + 2\mathbb{E}\left[\left|\Big(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\Big)\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)\right|\right]\\
&\leq \tilde{C}_1\sum_{k=0}^{1}\mathbb{E}\left[\left|\sqrt{\tilde{\pi}(k \mid X_t)}\big(Y_t - \hat{f}_{t-1}(k, X_t)\big) - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}\big(Y_t - f^*(k, X_t)\big)\right|\right]\\
&\quad + \mathbb{E}\left[\left|\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)^2 - \Big(f^*(1, X_t) - f^*(0, X_t) - \theta_0\Big)^2\right|\right]\\
&\quad + 2\mathbb{E}\left[\left|\Big(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\Big)\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)\right|\right]\\
&\leq \tilde{C}_1\sum_{k=0}^{1}\mathbb{E}\left[\left|\sqrt{\tilde{\pi}(k \mid X_t)}\hat{f}_{t-1}(k, X_t) - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}f^*(k, X_t)\right|\right]\\
&\quad + \tilde{C}_2\sum_{k=0}^{1}\mathbb{E}\left[\left|\sqrt{\tilde{\pi}(k \mid X_t)} - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}\right|\right] + \tilde{C}_3\sum_{k=0}^{1}\mathbb{E}\left[\left|\hat{f}_{t-1}(k, X_t) - f^*(k, 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

$$\begin{aligned}
&\tilde{C}_1\sum_{k=0}^{1}\mathbb{E}\left[\left|\sqrt{\tilde{\pi}(k \mid X_t)}\hat{f}_{t-1}(k, X_t) - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}f^*(k, X_t)\right|\right]\\
&\quad + \tilde{C}_2\sum_{k=0}^{1}\mathbb{E}\left[\left|\sqrt{\tilde{\pi}(k \mid X_t)} - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}\right|\right] + \tilde{C}_3\sum_{k=0}^{1}\mathbb{E}\left[\left|\hat{f}_{t-1}(k, X_t) - f^*(k, X_t)\right|\right]\\
&\leq \tilde{C}_4\sum_{k=0}^{1}\mathbb{E}\left[\left|\sqrt{\tilde{\pi}(k \mid X_t)} - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}\right|\right] + \tilde{C}_5\sum_{k=0}^{1}\mathbb{E}\left[\left|\hat{f}_{t-1}(k, X_t) - f^*(k, X_t)\right|\right].
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\big[|\sqrt{\pi_t(k \mid X_t, \Omega_{t-1})} - \sqrt{\tilde{\pi}(k \mid X_t)}| \mid X_t = x, \Omega_{t-1}\big] = \mathbb{E}\big[|\sqrt{\pi_t(k \mid x, \Omega_{t-1})} - \sqrt{\tilde{\pi}(k \mid x)}|\big] \to 0,\\
&\mathbb{E}\big[|\hat{f}_{t-1}(k, X_t) - f^*(k, X_t)| \mid X_t = x, \Omega_{t-1}\big] = \mathbb{E}\big[|\hat{f}_{t-1}(k, x) - f^*(k, x)|\big] \to 0,
\end{aligned}$$

as $t \to \infty$ using L^r-convergence theorem (Proposition 3). Here, we used the fact that $\hat{f}_{t-1}(k, x)$ and $\sqrt{\pi_t(k \mid x, \Omega_{t-1})}$ are independent from X_t. For fixed $x \in \mathcal{X}$, we can show that $\sqrt{\pi_t(k \mid x, \Omega_{t-1})}$, and $\hat{f}_{t-1}(k, x)$ are uniformly integrable from the boundedness of $\sqrt{\pi_t(k \mid x, \Omega_{t-1})}$, and $\hat{f}_{t-1}(k, x)$ (Proposition 2). From the point convergence of $\mathbb{E}[|\sqrt{\pi_t(k \mid X_t, \Omega_{t-1})} - \sqrt{\tilde{\pi}(k \mid X_t)}| \mid X_t = x]$ and $\mathbb{E}[|\hat{f}_{t-1}(k, X_t) - f^*(k, X_t)| \mid X_t = x]$, by using the Lebesgue's dominated convergence theorem, we can show that

$$\mathbb{E}_{X_t, \Omega_{t-1}}\left[\mathbb{E}[|\sqrt{\pi_t(k \mid X_t, \Omega_{t-1})} - \sqrt{\tilde{\pi}(k \mid X_t)}| \mid X_t, \Omega_{t-1}]\right] \to 0,$$
$$\mathbb{E}_{X_t, \Omega_{t-1}}\left[\mathbb{E}[|\hat{f}_{t-1}(k, X_t) - f^*(k, X_t)| \mid X_, \Omega_{t-1} t]\right] \to 0.$$

Then, as $t \to \infty$,

$$\mathbb{E}\left[z_t^2\right] - \mathbb{E}\left[\sum_{k=0}^{1} \frac{v\big(k, X_t\big)}{\tilde{\pi}(k \mid X_t)} + \Big(\theta_0(X_t) - \theta_0\Big)^2\right] \to 0.$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[z_t^2\right] - \mathbb{E}\left[\sum_{k=0}^{1} \frac{v\big(k, X_t\big)}{\tilde{\pi}(k \mid X_t)} + \Big(\theta_0(X_t) - \theta_0\Big)^2\right]\right) \leq \tilde{t}/T + \epsilon.$$

Here, $\mathbb{E}\left[\sum_{k=0}^{1} \frac{v(k, X_t)}{\tilde{\pi}(k|X_t)} + \Big(\theta_0(X_t) - \theta_0\Big)^2\right] = \mathbb{E}\left[\sum_{k=0}^{1} \frac{v(k, X)}{\tilde{\pi}(k|X)} + \Big(\theta_0(X) - \theta_0\Big)^2\right]$ 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

$$\sigma^2 = \mathbb{E}\left[\sum_{k=0}^{1} \frac{v\big(k, X\big)}{\tilde{\pi}(k \mid X)} + \Big(\theta_0(X) - \theta_0\Big)^2\right].$$

Step 2: Check of Condition (b)

From the boundedness of each variable in z_t, we can easily show that the condition (b) holds.

Step 3: Check of Condition (c)

Let u_t be an MDS such that

$$u_t = z_t^2 - \mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right]$$
$$= \left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2$$
$$- \mathbb{E}\left[\left(\frac{\mathbb{1}[A_t = 1]\big(Y_t - \hat{f}_{t-1}(1, X_t)\big)}{\pi_t(1 \mid X_t, \Omega_{t-1})} - \frac{\mathbb{1}[A_t = 0]\big(Y_t - \hat{f}_{t-1}(0, X_t)\big)}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2 \mid \Omega_{t-1}\right].$$

From the boundedness of each variable in z_t, we can apply weak law of large numbers for an MDS (Proposition 4 in Appendix A), and obtain

$$\frac{1}{T}\sum_{t=1}^{T} u_t = \frac{1}{T}\sum_{t=1}^{T}\left(z_t^2 - \mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right]\right) \xrightarrow{\mathrm{p}} 0.$$

Next, we show that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] - \sigma^2 \xrightarrow{\mathrm{p}} 0.$$

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

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

Then, we consider showing $\mathbb{E}\left[\left|\mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] - \sigma^2\right|\right] \to 0$. Here, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] - \sigma^2\right|\right]\\
&= \mathbb{E}\Bigg[\Bigg|\mathbb{E}\Bigg[\frac{\left(Y_t(1) - \hat{f}_{t-1}(1, X_t)\right)^2}{\pi_t(1 \mid X_t, \Omega_{t-1})} + \frac{\left(Y_t(0) - \hat{f}_{t-1}(0, X_t)\right)^2}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\\
&\quad + 2\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\\
&\quad - \frac{\left(Y_t(1) - f^*(1, X_t)\right)^2}{\tilde{\pi}(1 \mid X_t)} - \frac{\left(Y_t(0) - f^*(0, X_t)\right)^2}{\tilde{\pi}(0 \mid X_t)} - \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2 \mid \Omega_{t-1}\Bigg]\Bigg|\Bigg]\\
&= \mathbb{E}\Bigg[\Bigg|\mathbb{E}\Bigg[\mathbb{E}\Bigg[\frac{\left(Y_t(1) - \hat{f}_{t-1}(1, X_t)\right)^2}{\pi_t(1 \mid X_t, \Omega_{t-1})} + \frac{\left(Y_t(0) - \hat{f}_{t-1}(0, X_t)\right)^2}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\\
&\quad + 2\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\\
&\quad - \frac{\left(Y_t(1) - f^*(1, X_t)\right)^2}{\tilde{\pi}(1 \mid X_t)} - \frac{\left(Y_t(0) - f^*(0, X_t)\right)^2}{\tilde{\pi}(0 \mid X_t)} - \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2 \mid X_t, \Omega_{t-1}\Bigg] \mid \Omega_{t-1}\Bigg]\Bigg|\Bigg].
\end{aligned}$$

Then, by using Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] - \sigma^2\right|\right]\\
&\leq \mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg|\mathbb{E}\Bigg[\frac{\left(Y_t(1) - \hat{f}_{t-1}(1, X_t)\right)^2}{\pi_t(1 \mid X_t, \Omega_{t-1})} + \frac{\left(Y_t(0) - \hat{f}_{t-1}(0, X_t)\right)^2}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\\
&\quad + 2\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\\
&\quad - \frac{\left(Y_t(1) - f^*(1, X_t)\right)^2}{\tilde{\pi}(1 \mid X_t)} - \frac{\left(Y_t(0) - f^*(0, X_t)\right)^2}{\tilde{\pi}(0 \mid X_t)} - \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2 \mid X_t, \Omega_{t-1}\Bigg]\Bigg| \mid \Omega_{t-1}\Bigg]\Bigg]\\
&= \mathbb{E}\Bigg[\Bigg|\mathbb{E}\Bigg[\frac{\left(Y_t(1) - \hat{f}_{t-1}(1, X_t)\right)^2}{\pi_t(1 \mid X_t, \Omega_{t-1})} + \frac{\left(Y_t(0) - \hat{f}_{t-1}(0, X_t)\right)^2}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\\
&\quad + 2\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\\
&\quad - \frac{\left(Y_t(1) - f^*(1, X_t)\right)^2}{\tilde{\pi}(1 \mid X_t)} - \frac{\left(Y_t(0) - f^*(0, X_t)\right)^2}{\tilde{\pi}(0 \mid X_t)} - \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2 \mid X_t, \Omega_{t-1}\Bigg]\Bigg|\Bigg].
\end{aligned}$$

Because $\hat{f}_{t-1}$ and π_t are constructed from Ω_{t-1},

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] - \sigma^2\right|\right]\\
&\leq \mathbb{E}\Bigg[\Bigg|\mathbb{E}\Bigg[\frac{\left(Y_t(1) - \hat{f}_{t-1}(1, X_t)\right)^2}{\pi_t(1 \mid X_t, \Omega_{t-1})} + \frac{\left(Y_t(0) - \hat{f}_{t-1}(0, X_t)\right)^2}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\\
&\quad + 2\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\\
&\quad - \frac{\left(Y_t(1) - f^*(1, X_t)\right)^2}{\tilde{\pi}(1 \mid X_t)} - \frac{\left(Y_t(0) - f^*(0, X_t)\right)^2}{\tilde{\pi}(0 \mid X_t)} - \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2 \mid X_t, \hat{f}_{t-1}, \pi_t\Bigg]\Bigg|\Bigg].
\end{aligned}$$

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

$$\mathbb{E}\left[\left|\mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] - \sigma^2\right|\right]$$
$$\leq \mathbb{E}\left[\left|\mathbb{E}\left[\frac{\left(Y_t(1) - \hat{f}_{t-1}(1, X_t)\right)^2}{\pi_t(1 \mid X_t, \Omega_{t-1})} + \frac{\left(Y_t(0) - \hat{f}_{t-1}(0, X_t)\right)^2}{\pi_t(0 \mid X_t, \Omega_{t-1})} + \left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)^2\right.\right.\right.$$
$$\left. + 2\left(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\right)\left(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\right)\right\}$$
$$\left.\left.\left. - \frac{\left(Y_t(1) - f^*(1, X_t)\right)^2}{\tilde{\pi}(1 \mid X_t)} + \frac{\left(Y_t(0) - f^*(0, X_t)\right)^2}{\tilde{\pi}(0 \mid X_t)} - \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2 \mid X_t, \hat{f}_{t-1}, \pi_t\right]\right|\right]$$
$$\leq \tilde{C}_4 \sum_{k=0}^{1} \mathbb{E}\left[\left|\sqrt{\tilde{\pi}(k \mid X_t)} - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}\right|\right] + \tilde{C}_5 \sum_{k=0}^{1} \mathbb{E}\left[\left|\hat{f}_{t-1}(k, X_t) - f^*(k, X_t)\right|\right].$$

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

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

As a conclusion,

$$\frac{1}{T}\sum_{t=1}^{T} z_t^2 - \sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\left(z_t^2 - \mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] + \mathbb{E}\left[z_t^2 \mid \Omega_{t-1}\right] - \sigma^2\right) \xrightarrow{\mathrm{p}} 0.$$

Conclusion

From Steps 1–3, we can use CLT for an MDS. Hence, we have

$$\sqrt{T}\left(\hat{\theta}_T^{\mathrm{A2IPW}} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right),$$

where $\sigma^2 = \mathbb{E}\left[\sum_{k=0}^{1} \frac{\nu\left(k, X_t\right)}{\tilde{\pi}(k|X_t)} + \left(\theta_0(X_t) - \theta_0\right)^2\right]$. □

D Proof of Theorem 3

Proof.

$$\left(\theta_0 - \hat{\theta}_T^{\mathrm{A2IPW}}\right)^2 = \left(\frac{1}{T}\theta - \frac{1}{T}h_1 + \cdots + \frac{1}{T}\theta - \frac{1}{T}h_T\right)^2 = \frac{1}{T^2}\left(\theta - h_1 + \cdots + \theta - h_T\right)^2.$$

Let z_t be $\theta_0 - h_t$. Then,

$$\mathbb{E}_{\Pi}\left[(\theta - \hat{\theta}_T^{\mathrm{A2IPW}})^2\right] = \frac{1}{T^2}\mathbb{E}_{\Pi}\left[\left(\sum_{t=1}^{T} z_t\right)^2\right] = \frac{1}{T^2}\mathbb{E}_{\Pi}\left[\sum_{t=1}^{T} z_t^2 + 2\sum_{t=1}^{T}\sum_{s=1}^{t-1} z_t z_s\right].$$

We use the following result:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{t-1} z_t z_s\right]$$
$$= \sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{E}_{\Omega_{t-1}}\left[\mathbb{E}_{\Pi|\Omega_{t-1}}\left[z_t z_s \mid \Omega_{t-1}\right]\right]$$
$$= \sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{E}_{\Omega_{t-1}}\left[\mathbb{E}_{\Pi|\Omega_{t-1}}\left[z_t \mid \Omega_{t-1}\right] z_s\right]$$
$$= \sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{E}_{\Omega_{t-1}}\left[0 \times z_s\right] = 0.$$

Therefore,

$$\mathbb{E}_{\Pi}\left[(\theta_0 - \hat{\theta}^{\mathrm{A2IPW}}_T)^2\right] = \frac{1}{T^2}\mathbb{E}_{\Pi}\left[\sum^T_{t=1} z^2_t\right] = \frac{1}{T^2}\sum^T_{t=1}\mathbb{E}_{\Pi}\left[z^2_t\right].$$

As we showed in Step 1 of the proof of Theorem 1, we have

$$\begin{aligned}
&\mathbb{E}_{\Pi}\left[(\theta_0 - \hat{\theta}^{\mathrm{A2IPW}}_T)^2\right]\\
&= \frac{1}{T^2}\sum^T_{t=1}\mathbb{E}_{\Pi}\Bigg[\frac{\big(Y_t(1) - \hat{f}_{t-1}(1, X_t)\big)^2}{\pi_t(1\mid X_t, \Omega_{t-1})} + \frac{\big(Y_t(0) - \hat{f}_{t-1}(0, X_t)\big)^2}{\pi_t(0\mid X_t, \Omega_{t-1})} + \Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)^2\\
&\quad + 2\Big(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\Big)\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)\Bigg].
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
&\mathbb{E}_{\Pi^{\mathrm{OPT}}}\left[\Big(\theta_0 - \hat{\theta}^{\mathrm{OPT}}_T\Big)^2\right]\\
&= \frac{1}{T^2}\sum^T_{t=1}\mathbb{E}_{\Pi^{\mathrm{OPT}}}\left[\left(\frac{\mathbb{1}[\tilde{A}_t = 1]\big(Y_t - f^*(1, X_t)\big)}{\pi^{\mathrm{AIPW}}(1\mid X_t)} - \frac{\mathbb{1}[\tilde{A}_t = 0]\big(Y_t - f^*(0, X_t)\big)}{\pi^{\mathrm{AIPW}}(0\mid X_t)} + f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2\right],
\end{aligned}$$

where $\tilde{A}_t$ denotes the stochastic variable of an action under a policy π^{AIPW}. Then, we have

$$\begin{aligned}
&\frac{1}{T^2}\sum^T_{t=1}\mathbb{E}_{\Pi^{\mathrm{OPT}}}\left[\left(\frac{\mathbb{1}[\tilde{A}_t = 1]\big(Y_t - f^*(1, X_t)\big)}{\pi^{\mathrm{AIPW}}(1\mid X_t)} - \frac{\mathbb{1}[\tilde{A}_t = 0]\big(Y_t - f^*(0, X_t)\big)}{\pi^{\mathrm{AIPW}}(0\mid X_t)} + f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2\right]\\
&= \frac{1}{T^2}\sum^T_{t=1}\mathbb{E}\left[\frac{\big(Y_t(1) - f^*(1, X_t)\big)^2}{\pi^{\mathrm{AIPW}}(1\mid X_t)} + \frac{\big(Y_t(0) - f^*(0, X_t)\big)^2}{\pi^{\mathrm{AIPW}}(0\mid X_t)} + \Big(f^*(1, X_t) - f^*(0, X_t) - \theta_0\Big)^2\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\Pi}\left[\Big(\theta_0 - \hat{\theta}^{\mathrm{A2IPW}}_T\Big)^2\right] - \mathbb{E}_{\Pi^{\mathrm{OPT}}}\left[\Big(\theta_0 - \hat{\theta}^{\mathrm{OPT}}_T\Big)^2\right]\\
&= \frac{1}{T^2}\sum^T_{t=1}\mathbb{E}\Bigg[\frac{\big(Y_t(1) - \hat{f}_{t-1}(1, X_t)\big)^2}{\pi_t(1\mid X_t, \Omega_{t-1})} + \frac{\big(Y_t(0) - \hat{f}_{t-1}(0, X_t)\big)^2}{\pi_t(0\mid X_t, \Omega_{t-1})} + \Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)^2\\
&\quad + 2\Big(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\Big)\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)\Bigg]\\
&\quad - \frac{1}{T^2}\sum^T_{t=1}\mathbb{E}_{\Pi}\left[\frac{\big(Y_t(1) - f^*(1, X_t)\big)^2}{\pi^{\mathrm{AIPW}}(1\mid X_t)} + \frac{\big(Y_t(0) - f^*(0, X_t)\big)^2}{\pi^{\mathrm{AIPW}}(0\mid X_t)} + \Big(f^*(1, X_t) - f^*(0, X_t) - \theta_0\Big)^2\right]\\
&\leq \frac{1}{T^2}\sum^T_{t=1}\mathbb{E}\Bigg[\Bigg|\Bigg\{\frac{\big(Y_t(1) - \hat{f}_{t-1}(1, X_t)\big)^2}{\pi_t(1\mid X_t, \Omega_{t-1})} + \frac{\big(Y_t(0) - \hat{f}_{t-1}(0, X_t)\big)^2}{\pi_t(0\mid X_t, \Omega_{t-1})} + \Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)^2\\
&\quad + 2\Big(f^*(1, X_t) - f^*(0, X_t) - \hat{f}_{t-1}(1, X_t) + \hat{f}_{t-1}(0, X_t)\Big)\Big(\hat{f}_{t-1}(1, X_t) - \hat{f}_{t-1}(0, X_t) - \theta_0\Big)\Bigg\}\\
&\quad - \left\{\frac{\big(Y_t(1) - f^*(1, X_t)\big)^2}{\pi^{\mathrm{AIPW}}(1\mid X_t)} + \frac{\big(Y_t(0) - f^*(0, X_t)\big)^2}{\pi^{\mathrm{AIPW}}(0\mid X_t)} + \Big(f^*(1, X_t) - f^*(0, X_t) - \theta_0\Big)^2\right\}\Bigg|\Bigg],
\end{aligned}$$

where the expectation of the last equation is taken over random variables including Ω_{t-1}.

As we proved in Step 1 of proof of Theorem 1, there exist constants $\tilde{C}_0$ and $\tilde{C}_1$ such that

$$\mathbb{E}\left[\left(\theta_0 - \hat{\theta}_T^{\mathrm{A2IPW}}\right)^2\right] - \mathbb{E}\left[\left(\theta_0 - \hat{\theta}_T^{\mathrm{OPT}}\right)^2\right] \\ \leq \frac{\tilde{C}_0}{T^2}\sum_{t=1}^{T}\sum_{k=0}^{1}\mathbb{E}\left[\left|\sqrt{\pi^{\mathrm{AIPW}}(k \mid X_t)} - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}\right|\right] + \frac{\tilde{C}_1}{T^2}\sum_{t=1}^{T}\sum_{k=0}^{1}\mathbb{E}\left[\left|\hat{f}_{t-1}(k, X_t) - f^*(k, X_t)\right|\right].$$

Therefore, we have

$$\mathbb{E}\left[\left(\theta_0 - \hat{\theta}_T^{\mathrm{A2IPW}}\right)^2\right] - \mathbb{E}\left[\left(\theta_0 - \hat{\theta}_T^{\mathrm{OPT}}\right)^2\right] \\ = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{k=0}^{1}\left\{\mathrm{O}\left(\mathbb{E}\left[\left|\sqrt{\pi^{\mathrm{AIPW}}(k \mid X_t)} - \sqrt{\pi_t(k \mid X_t, \Omega_{t-1})}\right|\right]\right) + \mathrm{O}\left(\mathbb{E}\left[\left|f^*(k, X_t) - \hat{f}_{t-1}(k, X_t)\right|\right]\right)\right\}.$$

□

D.1 Proof of Theorem 4

The procedure of this proof mainly follows Balsubramani & Ramdas (2016). For a martingale M_t, let $V_t = \sum_{i=1}^{t}\mathbb{E}\left[(M_i - M_{i-1})^2 \mid \Omega_{i-1}\right]$. Before proving Theorem 4, we prove the following three lemmas.

Lemma 1 (Small Sample Bound for an MDS)**.** *Let M_t be a martingale such that for all $t \geq 1$, $\left|M_t - M_{t-1}\right| \leq e^2/2$ with probability 1. Fix any $\delta > 0$, and define $\tau_0 = \min\left\{s : 2(e-2)V_s \geq 173\log\left(\frac{4}{\delta}\right)\right\}$, Then, with probability $\geq 1-\delta$, for all $t \leq \tau_0$,*

$$|M_t| \leq 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right)$$

Lemma 2 (Uniform Bernstein Bound for Martingales at Any Time)**.** *Let M_t be a martingale such that for all $t \geq 1$, $\left|M_t - M_{t-1}\right| \leq e^2/2$ with probability 1. Then, with probability $\geq 1-\delta$, for all t simultaneously,*

$$|M_t| \leq C_0(\delta) + \sqrt{2C_1V_t\left(\log\log V_t + \log\left(\frac{4}{\delta}\right)\right)},$$

where $C_0(\delta) = 3(e-2) + 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right)$ and $C_1 = 6(e-2)$.

Remark 6. For the Napier's constant e, $e^2/2 \approx 3.694$.

Lemma 3 (Upper Bound of the Variance)**.** *Let M_t be a martingale such that for all $t \geq 1$, $\left|M_t - M_{t-1}\right| \leq e^2/2$ with probability 1. Suppose that there exists C_4 such that $\left|(M_t - M_{t-1})^2 - \mathbb{E}\left[(M_i - M_{i-1})^2 \mid \Omega_{i-1}\right]\right| \leq C_4$. With probability $\geq 1-\delta$, for all t, for sufficiently large V_t and $\sum_{i=1}^{t}(M_i - M_{i-1})^2$, there is an absolute constant C_3 such that*

$$V_t \leq C_3\left(\sum_{i=1}^{t}(M_i - M_{i-1})^2 + \frac{2C_4C_0(\delta)}{e^2}\right),$$

where $C_0(\delta) = 3(e-2) + 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right)$.

In this section, we use the following three propositions.

Proposition 6 (Balsubramani (2014), Lemma 23.)**.** Suppose that, for all $\ell \geq 3$ and t, $\mathbb{E}[(M_t - M_{t-1})^\ell \mid \Omega_{t-1}] \leq \frac{1}{2}\ell!\left(e/\sqrt{2}\right)^{2(\ell-2)}\mathbb{E}[(M_t - M_{t-1})^2 \mid \Omega_{t-1}]$. Then, for any $\lambda \in \left(-\frac{1}{e^2}, \frac{1}{e^2}\right)$, the process $U_t^\lambda := \exp(\lambda M_t - \lambda^2 V_t)$ is a super martingale.

Remark 7. The condition that, for all $\ell \geq 3$ and all t, $\mathbb{E}[(M_t - M_{t-1})^\ell \mid \Omega_{t-1}] \leq \frac{1}{2}\ell!\left(e/\sqrt{2}\right)^{2(\ell-2)}\mathbb{E}[(M_t - M_{t-1})^2 \mid \Omega_{t-1}]$ is satisfied when $|M_t - M_{t-1}| \leq \frac{e^2}{2}$ for all t with probability 1.

Proposition 7 (Uniform Bernstein Bound for Martingales, Balsubramani (2014), Theorem 5.)**.** Let M_t be a martingale such that for all $t \geq 1$, $\left|M_t - M_{t-1}\right| \leq e^2$ with probability 1. Fix any $\delta < 1$ and define $\tau_0 = \min\left\{s : 2(e-2)V_s \geq 173\log\left(\frac{4}{\delta}\right)\right\}$. Then, with probability $\geq 1-\delta$, for all $t \geq \tau_0$ simultaneously, $\left|M_t\right| \leq \frac{2(e-2)}{e^2(1+\sqrt{1/3})}V_t$ and

$$\left|M_t\right| \leq \sqrt{6(e-2)V_t\left(2\log\log\left(\frac{3(e-2)V_t}{\left|M_t\right|}\right) + \log\left(\frac{2}{\delta}\right)\right)}.$$

Proposition 8. Suppose b_1, b_2, c are positive constants, $r \geq 8\max\left(e^4 b_1 \log\log(e^4 r/4), e^4 b_2\right)$, and $r - \sqrt{b_1 e^4 r \log\log\left(e^4 r/4\right) + b_2 e^4 r} - c \leq 0$. Then,

$$\sqrt{r} \leq \sqrt{b_1 e^4 \log\log(e^4 c/2) + b_2 e^4} + \sqrt{c}.$$

This proposition is almost same as Lemma 9 of Balsubramani (2014), but we changed the statement a little. We show the proof as follows.

Proof of Lemma 8. Since $r \geq 8e^4 b_2$,

$$0 \leq \frac{r}{8} - e^4 b_2 = \frac{r}{4} - \frac{r}{8} - e^4 b_2 = \frac{r}{4} - b_1\frac{r}{8b_1} - e^4 b_2 \to 0 \leq \frac{r^2}{4} - b_1 r\frac{r}{8b_1} - b_2 e^4 r.$$

Substituting the assumption $\frac{r}{8b_1} \geq e^4 \log\log(e^4 r/4)$ gives

$$\begin{aligned}
0 &\leq \frac{r^2}{4} - b_1 r\frac{r}{8b_1} - b_2 e^4 r \leq \frac{r^2}{4} - b_1 r e^4 \log\log\left(e^4 r/4\right) - b_2 e^4 r \\
&\to \sqrt{b_1 r e^4 \log\log\left(e^4 r/4\right) + b_2 e^4 r} \leq \frac{r}{2}.
\end{aligned}$$

Then, by substituting this into $r - \sqrt{b_1 e^4 r \log\log\left(e^4 r/4\right) + b_2 e^4 r} - c \leq 0$, we have $r \leq 2c$. Therefore, again using $r - \sqrt{b_1 e^4 r/4 \log\log\left(e^4 r/4\right) + b_2 e^4 r} - c \leq 0$,

$$\begin{aligned}
0 &\geq r - \sqrt{b_1 e^4 r \log\log\left(e^4 r/4\right) + b_2 e^4 r} - c \\
&\geq r - \sqrt{b_1 e^4 r \log\log\left(e^4 c/2\right) + b_2 e^4 r} - c.
\end{aligned}$$

This is a quadratic in $\sqrt{r}$. By solving it, we have

$$\begin{aligned}
\sqrt{r} &\leq \frac{1}{2}\left(\sqrt{b_1 e^4 \log\log\left(e^4 c/2\right) + b_2 e^4} + \sqrt{b_1 e^4 \log\log\left(e^4 c/2\right) + b_2 e^4 + 4c}\right) \\
&\leq \sqrt{b_1 e^4 \log\log\left(e^4 c/2\right) + b_2 e^4} + \sqrt{c}
\end{aligned}$$

□

Then, we prove Lemmas 1–3 and Theorem 4 as follows.

Proof of Lemma 1

Proof. This proof mostly follows the proof of Theorem 24 of Balsubramani (2014).

First, by using Proposition 6, we show that $2 \geq \mathbb{E}\left[\exp\left(\lambda_0|M_\tau| - \lambda_0^2 V_\tau\right)\right]$ for any stopping time τ and $\lambda \in \left(-\frac{1}{e^2}, \frac{1}{e^2}\right)$. From Proposition 6, $U_t^\lambda := \exp(\lambda M_t - \lambda^2 V_t)$ is a super martingale. The condition that, for all $\ell \geq 3$, $\mathbb{E}[(M_t - M_{t-1})^\ell \mid \Omega_{t-1}] \leq \frac{1}{2}\ell!\left(e/\sqrt{2}\right)^{2(\ell-2)}\mathbb{E}[(M_t - M_{t-1})^2 \mid \Omega_{t-1}]$ holds from the assumption that $|M_t - M_{t-1}| \leq e^2/2$ for all t with probability 1. For $\lambda_0 \in \left(-\frac{1}{e^2}, \frac{1}{e^2}\right)$, let us consider a situation where $\lambda \in \{-\lambda_0, \lambda_0\}$ with probability $1/2$ each. After marginalizing over λ, the resulting process is

$$\begin{aligned}
\tilde{U}_t &= \frac{1}{2}\exp(\lambda_0 M_t - \lambda_0^2 V_t) + \frac{1}{2}\exp(-\lambda_0 M_t - \lambda_0^2 V_t) \\
&\geq \frac{1}{2}\exp(\lambda_0 M_t - \lambda_0^2 V_t).
\end{aligned}$$

On the other hand, for any stopping time τ, from the optimal stopping theorem for a super martingale (Durrett, 2010), we have

$$\mathbb{E}\left[\exp(\lambda_0 M_\tau - \lambda_0^2 V_\tau)\right] \leq \mathbb{E}\left[\exp(\lambda_0 M_0 - \lambda_0^2 V_0)\right] = 1.$$

Similarly,

$$\mathbb{E}\left[\exp(-\lambda_0 M_\tau - \lambda_0^2 V_\tau)\right] \leq \mathbb{E}\left[\exp(-\lambda_0 M_0 - \lambda_0^2 V_0)\right] = 1.$$

Combining these results, we have

$$\mathbb{E}\left[\tilde{U}_t\right] = \mathbb{E}\left[\frac{1}{2}\exp(\lambda_0 M_t - \lambda_0^2 V_t) + \frac{1}{2}\exp(-\lambda_0 M_t - \lambda_0^2 V_t)\right] \leq 1,$$

and $1 \geq \mathbb{E}\left[\frac{1}{2}\exp(\lambda_0 M_t - \lambda_0^2 V_t)\right]$. Thus, we proved $2 \geq \mathbb{E}\left[\exp\left(\lambda_0 |M_\tau| - \lambda_0^2 V_\tau\right)\right]$.

Next, note that $\tau_0 = \min\left\{s : V_s \geq \frac{173}{2(e-2)}\log\left(\frac{4}{\delta}\right)\right\}$. Therefore, by defining the stopping time $\tau_1 = \min\left\{s : |M_t| \geq 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right)\right\}$ and using $\lambda_0 = \sqrt{\frac{2(e-2)}{173}} \approx 0.091 \leq \frac{1}{e^2} \approx 0.135$,

$$\begin{aligned}
2 &\geq \mathbb{E}\left[\exp\left(\lambda_0 |M_{\tau_1}| - \lambda_0^2 V_{\tau_1}\right)\right] \\
&\geq \mathbb{E}\left[\exp\left(\lambda_0 |M_{\tau_1}| - \lambda_0^2 V_{\tau_1}\right) \mid \tau_1 < \tau_0\right] \mathbb{P}\left(\tau_1 < \tau_0\right) \\
&\geq \mathbb{E}\left[\exp\left(2\lambda_0 \sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right) - \lambda_0^2 \frac{173}{2(e-2)}\log\left(\frac{4}{\delta}\right)\right) \mid \tau_1 < \tau_0\right] \mathbb{P}\left(\tau_1 < \tau_0\right) \\
&\geq \mathbb{E}\left[\exp\left(\log\left(\frac{4}{\delta}\right)\right) \mid \tau_1 < \tau_0\right] \mathbb{P}\left(\tau_1 < \tau_0\right) = \frac{4}{\delta}\mathbb{P}\left(\tau_1 < \tau_0\right).
\end{aligned}$$

Thus, we obtain $\mathbb{P}\left(\tau_1 < \tau_0\right) \leq \frac{\delta}{2} < \delta$. □

Proof of Lemma 2

Proof. From Proposition 7, with probability $\geq 1 - \delta/2$, for all $t \geq \tau_0$ simultaneously, $\left|M_t\right| \leq \frac{2(e-2)}{e^2(1+\sqrt{1/3})} V_t$ and

$$\left|M_t\right| \leq \sqrt{6(e-2)V_t\left(2\log\log\left(\frac{3(e-2)V_t}{\left|M_t\right|}\right) + \log\left(\frac{4}{\delta}\right)\right)}.$$

Therefore we have that, with probability $\geq 1 - \delta/2$, for all $t \geq \tau_0$, simultaneously, $\left|M_t\right| \leq \frac{2(e-2)}{e^2(1+\sqrt{1/3})} V_t$ and

$$\left|M_t\right| \leq \max\left(3(e-2), \sqrt{2C_1 V_t \log\log V_t + C_1 V_t \log\left(\frac{4}{\delta}\right)}\right), \tag{3}$$

where note that $C_1 = 6(e-2)$.

Next, from Lemma 1, with probability $\geq 1 - \delta/4$, for all $t \leq \tau_0$ simultaneously,

$$|M_t| \leq 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right)$$

By taking a union bound of (3), with probability $\geq 1 - \delta$, the following inequality holds for all t simultaneously:

$$\left|M_t\right| \leq \begin{cases} 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right) & \text{if } t \leq \tau_0 \\ \frac{2(e-2)}{e^2(1+\sqrt{1/3})} V_t \text{ and } \max\left(3(e-2), \sqrt{2C_1 V_t \log\log V_t + C_1 V_t \log\left(\frac{4}{\delta}\right)}\right) & \text{if } t \geq \tau_0. \end{cases}$$

Then, with probability $\geq 1 - \delta$, the following relationship holds for all t simultaneously:

$$|M_t| \leq C_0(\delta) + \sqrt{C_1 V_t \left(2\log\log V_t + \log\left(\frac{4}{\delta}\right)\right)}.$$

□

Proof of Lemma 3

Proof. Let $\tilde{M}_t$ be $\sum_{i=1}^{t}(M_i - M_{i-1})^2 - V_t$, where note that $V_t = \sum_{i=1}^{t}\mathbb{E}\left[(M_i - M_{i-1})^2 \mid \Omega_{i-1}\right]$ Suppose that there exists C_4 such that $\left|(M_t - M_{t-1})^2 - \mathbb{E}\left[(M_i - M_{i-1})^2 \mid \Omega_{i-1}\right]\right| \leq C_4$ with probability 1 in which the existence is guaranteed by the boundedness of $M_i - M_{i-1}$, i.e., $|M_i - M_{i-1}| \leq e^2/2$ for all t with probability 1. Because $\tilde{M}_t$ is a martingale, we can apply Proposition 2, i.e., for all t, with probability $\geq 1-\delta$

$$\left|\tilde{M}_t\right| \leq \frac{2C_4}{e^2}\left(C_0(\delta) + \sqrt{C_1 B_t\left(2\log\log B_t + \log\left(\frac{4}{\delta}\right)\right)}\right),$$

where $B_t = \mathbb{E}\left[\left(\sum_{i=1}^{t}(M_i - M_{i-1})^2 - V_t\right)^2 \mid \Omega_{i-1}\right]$. For B_t, we have

$$\begin{aligned}
B_t &= \sum_{i=1}^{t}\left(\mathbb{E}\left[(M_i - M_{i-1})^4 \mid \Omega_{i-1}\right] - \left(\mathbb{E}\left[(M_i - M_{i-1})^2 \mid \Omega_{i-1}\right]\right)^2\right)\\
&\leq \sum_{i=1}^{t}\mathbb{E}\left[(M_i - M_{i-1})^4 \mid \Omega_{i-1}\right] \leq \left(e^8/2^4\right)\sum_{i=1}^{t}\mathbb{E}\left[(M_i - M_{i-1})^4/(e^8/2^4) \mid \Omega_{i-1}\right]
\end{aligned}$$

Because $M_i - M_{i-1} \leq e^2/2 \to \frac{(M_i - M_{i-1})^2}{e^4/2^2} \leq 1$, we have $(M_i - M_{i-1})^2/(e^4/2^2) \geq (M_i - M_{i-1})^4/(e^8/2^4)$, and

$$\sum_{i=1}^{t}\mathbb{E}\left[(M_i - M_{i-1})^4 \mid \Omega_{i-1}\right] \leq e^8/2^4\sum_{i=1}^{t}\mathbb{E}\left[(M_i - M_{i-1})^2/(e^4/2^2) \mid \Omega_{i-1}\right] = e^4 V_t/4. \tag{4}$$

Therefore,

$$\begin{aligned}
\left|\tilde{M}_t\right| &\leq \frac{2C_4}{e^2}\left(C_0(\delta) + \sqrt{C_1 B_t\left(2\log\log B_t + \log\left(\frac{4}{\delta}\right)\right)}\right)\\
&\leq \frac{2C_4}{e^2}\left(C_0(\delta) + \sqrt{C_1 e^4 V_t/4\left(2\log\log\left(e^4 V_t/4\right) + \log\left(\frac{4}{\delta}\right)\right)}\right).
\end{aligned}$$

This can be relaxed to

$$\begin{aligned}
&-\sum_{i=1}^{t}(M_i - M_{i-1})^2 + V_t - \frac{2C_4}{e^2}\left(C_0(\delta) + \sqrt{C_1 e^4 V_t/4\left(2\log\log\left(e^4 V_t/4\right) + \log\left(\frac{4}{\delta}\right)\right)}\right)\\
&= -\sum_{i=1}^{t}(M_i - M_{i-1})^2 + V_t - \left(\frac{2C_4 C_0(\delta)}{e^2} + \sqrt{\frac{C_4^2 C_1}{e^4}e^4 V_t\left(2\log\log\left(e^4 V_t/4\right) + \log\left(\frac{4}{\delta}\right)\right)}\right) \leq 0.
\end{aligned}$$

We consider two cases for V_t. First, we consider a case where $V_t \geq 8\max\left(e^4\frac{C_4^2 C_1}{e^4}2\log\log\left(e^4 V_t\right), e^4\frac{C_4^2 C_1}{e^4}\log\left(\frac{4}{\delta}\right)\right)$. Then, from Proposition 8, we have

$$\begin{aligned}
\sqrt{V_t} \leq &\sqrt{\frac{C_4^2 C_1}{e^4}2e^4\log\log\left(e^2 C_4 C_0(\delta) + e^4\sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right) + e^4\frac{C_4^2 C_1}{e^4}\log\left(\frac{4}{\delta}\right)}\\
&+ \sqrt{\frac{2C_4 C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2}\\
= &\sqrt{2C_4^2 C_1\log\log\left(e^2 C_4 C_0(\delta) + e^4\sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right) + C_4^2 C_1\log\left(\frac{4}{\delta}\right)}\\
&+ \sqrt{\frac{2C_4 C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2}.
\end{aligned}$$

For sufficiently high $\sum_{i=1}^{t}(M_i - M_{i-1})^2$ such that $2C_4^2C_1 \log\log\left(e^2C_4C_0(\delta) + e^4\sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right) \geq C_4^2C_1\log\left(\frac{4}{\delta}\right)$, by using a constant C_5, the RHS is bounded as

$$\sqrt{2C_4^2C_1 \log\log\left(e^2C_4C_0(\delta) + e^4\sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right) + C_4^2C_1\log\left(\frac{4}{\delta}\right)} + \sqrt{\frac{2C_4C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2}$$

$$\leq \sqrt{4C_4^2C_1 \log\log\left(e^2C_4C_0(\delta) + e^4\sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right)} + \sqrt{\frac{2C_4C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2}$$

$$\leq \sqrt{4C_4^2C_1 \left(e^2C_4C_0(\delta) + e^4\sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right)} + \sqrt{\frac{2C_4C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2}.$$

Then, by squaring both sides of

$$\sqrt{V_t} \leq \sqrt{4C_4^2C_1 \left(e^2C_4C_0(\delta) + e^4\sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right)} + \sqrt{\frac{2C_4C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2}$$

$$= \sqrt{2e^4C_4^2C_1 \left(\frac{2C_4C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2\right)} + \sqrt{\frac{2C_4C_0(\delta)}{e^2} + \sum_{i=1}^{t}(M_i - M_{i-1})^2},,$$

we obtain

$$V_t \leq C_3\left(\sum_{i=1}^{t}(M_i - M_{i-1})^2 + \frac{2C_4C_0(\delta)}{e^2}\right),$$

where C_3 is a constant. When $V_t < 8\max\left(e^4\frac{C_4^2C_1}{e^4}2\log\log\left(e^4V_t\right), e^4\frac{C_4^2C_1}{e^4}\log\left(\frac{4}{\delta}\right)\right)$, the statement clearly holds for sufficiently high V_t such that $V_t < e^4\frac{C_4^2C_1}{e^4}2\log\log\left(e^4V_t\right)$. □

Proof of Theorem 4

Finally, combining the above results, we show Theorem 4 as follows.

Proof. Let us note that we can construct an MDS from $z_t = q_t - \theta_0$ as $\{z_t\}_{t=1}^{T}$. Let us suppose that there exists a constant C such that $|z_t| \leq C$. Let $\tilde{z}_t$ and $\tilde{V}_t$ be $z_te^2/(2C)$ and $\sum_{i=1}^{t}\mathbb{E}\left[\tilde{z}_i^2 \mid \Omega_{i-1}\right]$, respectively. From this boundedness of z_t, there exists a constant C_4 such that $|z_t^2 - \mathbb{E}[z_t^2 \mid \Omega_{t-1}]| \leq C_4$. Then, for fixed δ, from Proposition 2, with probability $\geq 1-\delta$, the following true for all t simultaneously:

$$\left|t\hat{\theta}_t^{\mathrm{A2IPW}} - t\theta_0\right| \leq \frac{2C}{e^2}\left(C_0(\delta) + \sqrt{2C_1\tilde{V}_t^*\left(\log\log\tilde{V}_t^* + \log\left(\frac{4}{\delta}\right)\right)}\right).$$

Here, by using Proposition 3, we have

$$\tilde{V}_t \leq C_3\left(\sum_{i=1}^{t}\tilde{z}_i^2 + \frac{2C_4C_0(\delta)}{e^2}\right),$$

Then,

$$\left|t\hat{\theta}_t^{\mathrm{A2IPW}} - t\theta_0\right|$$

$$\leq \frac{2C}{e^2}\left(C_0(\delta) + \sqrt{2C_1C_3\left(\frac{e^4}{4C^2}\sum_{i=1}^{t}z_t^2 + \frac{2C_4C_0(\delta)}{e^2}\right)\left(\log\log C_3\left(\frac{e^4}{4C^2}\sum_{i=1}^{t}z_t^2 + \frac{2C_4C_0(\delta)}{e^2}\right) + \log\left(\frac{4}{\delta}\right)\right)}\right).$$

□

E A2IPW Estimator for Off-policy Evaluation

In *off-policy evaluation* (OPE), we consider the following problem setting. Let A_t be the action taking variable 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 domain of covariate. Let us denote a random variable of a reward at period t as a function $Y_t : \mathcal{A} \to \mathbb{R}$. In this paper, we have access to a set $\mathcal{S}_T = \{(X_t, A_t, Y_t)\}_{t=1}^T$ with the following data generating process (DGP):

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

where $p(x)$ denote the density of the covariate X_t, $\pi_t(a \mid x, \Omega_{t-1})$ denote the probability of assigning an action A_t conditioned on a covariate X_t, which is also called *behavior policy*, $p(y \mid a, x)$ denote the density of an outcome Y_t conditioned on A_t and X_t, and $\Omega_{t-1} \in \mathcal{M}_{t-1}$

Under the DGP defined above, we consider estimating the value of an *evaluation policy* using samples obtained under the behavior policy. Let an evaluation policy $\boldsymbol{x} : \mathcal{A} \times \mathcal{X} \to (0, 1)$ be a function of a covariate X_t and an action A_t, which can be considered as the probability of taking an action A_t conditioned on a covariate X_t. We are interested in estimating the expected reward from any given pre-specified evaluation policy $\pi^{\mathrm{e}}(a \mid x)$. Then, we define the expected reward under an evaluation policy as $R(\boldsymbol{x}) := \mathbb{E}\left[\sum_{k=1}^K \boldsymbol{x}(k, X_t)Y_t(k)\right]$. We also denote $R(\boldsymbol{x})$ as θ_0. The goal of OPE is to estimate $R(\pi^{\mathrm{e}})$ using dependent samples under a batch updated behavior policies.

For OPE, we can obtain the following corollary from Theorem 1.

Corollary 1 (Asymptotic Distribution of A2IPW for OPE)**.** *Suppose that*

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

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

Then, under Assumption 1, for the A2IPW estimator, we have $\sqrt{T}\left(\hat{\theta}_T^{\mathrm{A2IPW}} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$, where $\sigma^2 = \mathbb{E}\left[\sum_{k=1}^K \frac{\left(\boldsymbol{x}\left(k, X_t\right)\right)^2 \nu^*\left(k, X_t\right)}{\tilde{\pi}(k \mid X_t)} + \left(\sum_{k=1}^K \boldsymbol{x}(k, X_t)f^*(k, X_t) - \theta_0\right)^2\right]$.

F Details of Statistical Hypothesis Testing

This section provides the preliminaries of statistical hypothesis testing.

A hypothesis refers to a statement about a population parameter (Casella, 2002). Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be the null and alternative hypotheses, respectively. For simplicity, we only discuss the following hypotheses: $\mathcal{H}_0 : \theta_0 = \mu$ and $\mathcal{H}_1 : \theta_0 \neq \mu$ for $\mu \in \mathbb{R}$. In a hypothesis testing problem, after observing the sample, the experimenter must decide either to accept $\mathcal{H}_0$ as true or to reject $\mathcal{H}_0$ as false (Casella, 2002). In deciding to accept or reject the null hypothesis $\mathcal{H}_0$, an experimenter might be making a mistake, which are classified into Type I and Type II errors. In the Type I error, the hypothesis testing incorrectly decides to reject $\mathcal{H}_0$, but $\theta_0 = \mu$ holds in the population. In the Type II error, the test incorrectly decides to accept $\mathcal{H}_0$, but $\theta_0 \neq \mu$ holds in the population. As criteria for controlling these errors, we consider their probabilities. Let $\mathbb{P}_{\mathcal{H}_0}$ and $\mathbb{P}_{\mathcal{H}_1}$ be the probabilities when the null and alternative hypotheses are correct, respectively. When $\mathbb{P}_{\mathcal{H}_0}\big(\text{reject } \mathcal{H}_0\big) \leq \alpha$, we say that we control the Type I error at α. When $\mathbb{P}_{\mathcal{H}_1}\big(\text{reject } \mathcal{H}_0\big) \leq \beta$, we say that we control the Type II error at $1 - \beta$. To discuss this more generally, let us define $\mathbb{P}_\theta(\text{reject } \mathcal{H}_0)$, where $\mathbb{P}_\theta$ denotes $\mathbb{P}_{\mathcal{H}_0}$ if the null hypothesis is correct; otherwise $\mathbb{P}_\theta$ denotes $\mathbb{P}_{\mathcal{H}_1}$. This probability is also known as the power function $\beta(\theta) = \mathbb{P}_\theta(\hat{\theta}_t^{\mathrm{A2IPW}} \in R)$, where R is a rejection region, where, if $\hat{\theta}_t^{\mathrm{A2IPW}} \in R$, then we reject the null hypothesis.

The methods of hypothesis testing can be classified into two approaches. In the first approach, we assume a fixed sample size, and construct the confidence interval after obtaining a set of samples with the sample size. This types of hypothesis testing is well accepted and a standard of hypothesis testing. In the second approach, we conduct the hypothesis testing sequentially, in which the sample size is regarded as a random variable. This approach recently gathered attention because it is more suitable to the situation with sequential decision making such as the MAB problem.

F.1 Standard Statistical Test with a Fixed Sample Size

First, we consider the standard statistical test with a fixed (predetermined) sample size and the proposed A2IPW estimator under $\pi_t(k \mid x) - \tilde{\pi}(k \mid x) \xrightarrow{\mathrm{p}} 0$ for all $x \in \mathcal{X}$. In this case, we can use the (asymptotic) *Student's t-test* or z-test with the following t-statistic:

$$t\text{-statistic} = \frac{\hat{\theta}^{\mathrm{A2IPW}}_T - \mu}{\sqrt{\hat{\sigma}^2/T}},$$

where $\hat{\sigma}^2$ is an estimator of $\sigma^2 = \mathbb{E}\left[\sum^1_{k=0} \frac{\nu^*\left(k, X_t\right)}{\tilde{\pi}(k|X_t)} + \left(f^*(1, X_t) - f^*(0, X_t) - \theta_0\right)^2\right]$. Then, by considering a situation where there are sufficient samples and $\hat{\sigma}^2 = \sigma^2$, if the null hypothesis is correct (i.e., $\theta_0 = 0$ is true), the T-statistic asymptotically follows the standard normal distribution. By using this results, the test rejects the null hypothesis whenever

$$\left|\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \mu\right)\right| > \sqrt{\hat{\sigma}^2} z_{1-\alpha/2},$$

where z_α is the α quantile of the standard normal distribution. Then, when the sample size T is large, the Type I error is controlled as

$$\mathbb{P}_{\mathcal{H}_0}\left(\left|\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \mu\right)\right| > \sqrt{\sigma^2} z_{\alpha/2}\right) \leq \alpha.$$

F.2 Sequential Testing

For a null $\mathcal{H}_0$ and an alternative $\mathcal{H}_1$ hypothesis, we have an incentive to make our decision using experiments with as small a sample size as possible. In sequential testing, we do not have to decide the sample size in advance. We sequentially conduct decision making and stop whenever we want. However, if we sequentially conduct standard statistical testing, the probability of the Type I error increases (Balsubramani & Ramdas, 2016).

In sequential testing, the probability of the Type II error does not increase (Balsubramani & Ramdas, 2016). However, we can control more precisely the Type II error by introducing certain methods (Jamieson & Jain, 2018).

Sequential Testing with Multiple Testing Correction: As explained in Section 4, one standard method for reducing errors is applying a kind of multiple testing correction such as the BF and Benjamini–Hochberg procedures. Some concepts can be used to control the Type I error in multiple testing, such as the false discovery rate and family-wise error rate (Jamieson & Jain, 2018). However, we do not discuss these concepts in detail because of space limitations.

Sequential Testing with the LIL: However, these corrections are exceedingly conservative and they produce sub-optimal results over a large number of tests (Balsubramani & Ramdas, 2016). To avoid this problem, the concentration inequalities derived from the LIL are useful (Balsubramani, 2014; Jamieson et al., 2014; Johari et al., 2015; Balsubramani & Ramdas, 2016); the properties of the LIL in sequential testing were further investigated by Zhao et al. (2016) and Jamieson & Jain (2018). As we explained in Section 4, the LIL-based sequential testing has been already used in various existing studies (Jamieson & Jain, 2018).

Remark 8 (LIL and an MDS)**.** Khintchine (1924) and Kolmogoroff (1929) derived the LIL for independent random variables. Following their methods, several works have derived other LILs for an MDS under certain regularity conditions (Stout, 1970; Fisher, 1992). Further, a result is related to the CLT under certain rate conditions (Tomking, 1971). On the convergence rate of the CLT for an MDS, see Hall & Hayde (1980). In this paper, we do not introduce the asymptotic LIL for an MDS explicitly.

F.3 Sample Size and Stopping Time

In hypothesis testing, we are interested in the sample size required to reject the null hypothesis with controlling Type II error at β when the alternative hypothesis $\mathcal{H}_1$ is true. To control the Type II error, we introduce a parameter $\Delta > 0$, which is called the *effect size* in the literature on hypothesis testing. Let us redefine the alternative hypothesis as $\mathcal{H}_1(\Delta) : |\theta_0 - \mu| > \Delta$, where $\mathbb{P}_{\mathcal{H}_1(\Delta)}$ is the probability when the alternative hypothesis is correct. Let R_n be a rejection region when controlling the Type II error at β, i.e., when $\hat{\theta}^{\mathrm{A2IPW}}_n \in R_n$ and the alternate hypothesis $\mathcal{H}_1$ is true, the null hypothesis is rejected with the probability of the Type II error at least $1 - \beta$. Then, for Δ and β, the minimum sample size with controlling Type II error at β is defined as $n^*_\beta(\Delta) = \min\left\{n : \mathbb{P}_{\mathcal{H}_1(\Delta)}\left(\hat{\theta}^{\mathrm{A2IPW}}_n \in R_n\right) \geq 1 - \beta\right\}$, which is also referred to as *sample complexity* in the MAB problem. In sequential testing, the sample size corresponds to the stopping time when the algorithm stops by rejecting the null hypothesis. Let τ be the stopping time of sequential testing.

F.4 Minimum Sample Size under the Optimal Policy

For discussing the minimum sample required in hypothesis testing, we derive the minimum sample size under an ideal situation where we know the optimal policy and use it as a policy for choosing an action, i.e., $\pi_t = \pi^{\mathrm{AIPW}}$.

Let us denote the minimum sample size in this case as $n_\beta^{\mathrm{OPT}*}(\Delta)$. For the sufficiently large sample size T, from Theorem 1, we have

$$\sqrt{T}\left(\hat{\theta}_T^{\mathrm{A2IPW}} - \mu\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \tilde{\sigma}^2\right),$$

where

$$\tilde{\sigma}^2 = \mathbb{E}\left[\sum_{k=0}^{1} \frac{\tilde{\nu}^*\big(k, X_t\big)}{\pi^{\mathrm{AIPW}}(k \mid X_t)} + \left(\tilde{f}^*(1, X_t) - \tilde{f}^*(0, X_t) - \mu\right)^2\right],$$

$\tilde{f}(k, X_t) = \mathbb{E}_{\mathcal{H}_0}[Y_t(k) \mid X_t]$, $\tilde{\nu}^*\big(k, X_t\big) = \mathbb{E}_{\mathcal{H}_0}[\big(Y_t(k) - \tilde{f}(k, X_t)\big)^2 \mid X_t]$, $\tilde{\pi} = \frac{\sqrt{\tilde{\nu}^*\left(1, X_t\right)}}{\sqrt{\tilde{\nu}^*\left(1, X_t\right)} + \sqrt{\tilde{\nu}^*\left(0, X_t\right)}}$, and $\mathbb{E}_{\mathcal{H}_0}$ denotes the expectation when the null hypothesis is true. From this result, we have

$$\frac{\sqrt{T}\left(\hat{\theta}_T^{\mathrm{A2IPW}} - \mu\right)}{\sqrt{\tilde{\sigma}^2}} \xrightarrow{\mathrm{d}}_{\mathcal{H}_0} \mathcal{N}(0, 1), \quad \frac{\sqrt{T}\left(\hat{\theta}_T^{\mathrm{A2IPW}} - \mu - \Delta\right)}{\sqrt{\ddot{\sigma}^2}} \xrightarrow{\mathrm{d}}_{\mathcal{H}_1} \mathcal{N}(0, 1),$$

where

$$\ddot{\sigma}^2 = \mathbb{E}\left[\sum_{k=0}^{1} \frac{\ddot{\nu}^*\big(k, X_t\big)}{\ddot{\pi}^{\mathrm{AIPW}}(k \mid X_t)} + \left(\ddot{f}^*(1, X_t) - \ddot{f}^*(0, X_t) - \mu - \Delta\right)^2\right],$$

$\ddot{f}(k, X_t) = \mathbb{E}_{\mathcal{H}_1}[Y_t(k) \mid X_t]$, $\ddot{\nu}^*\big(k, X_t\big) = \mathbb{E}_{\mathcal{H}_1}[\big(Y_t(k) - \ddot{f}(k, X_t)\big)^2 \mid X_t]$, $\ddot{\pi} = \frac{\sqrt{\ddot{\nu}^*\left(1, X_t\right)}}{\sqrt{\ddot{\nu}^*\left(1, X_t\right)} + \sqrt{\ddot{\nu}^*\left(0, X_t\right)}}$, and $\mathbb{E}_{\mathcal{H}_1}$ denotes the expectation when the alternate hypothesis is true.

Based on these results, when we have sufficient samples and know $\tilde{\sigma}^2$, we rejects the null hypothesis whenever

$$\left|\sqrt{T}\left(\hat{\theta}_T^{\mathrm{A2IPW}} - \mu\right)\right| > \sqrt{\tilde{\sigma}^2} z_{1-\alpha/2},$$

where note that $z_{1-\alpha/2}$ is s the $1 - \alpha/2$ quantile of the standard normal distribution.

For ease of discussion, we put the following two assumptions,

Assumption 2. The density of $p(x)$ is the same under the null and alternate hypothesis.

Assumption 3. For the models of conditional outcomes,

$$\ddot{f}(k, X_t) = \Delta + \tilde{f}(k, X_t).$$

Besides, when the null hypothesis is true,

$$Y(k) = \tilde{f}(k, X_t) + \tilde{\varepsilon}_t;$$

when the alternate hypothesis is true,

$$Y(k) = \ddot{f}(k, X_t) + \ddot{\varepsilon}_t,$$

where $\tilde{\varepsilon}_t$ and $\ddot{\varepsilon}_t$ are random variables with mean zero and independent from X_t.

Let us note that, under these assumptions, we have $\ddot{\sigma}^2 = \tilde{\sigma}^2$. As explained in Section F.1, the Type I error is controlled at α. On the other hand, the asymptotic power is given as

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_1}\left(\left|\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \mu\right)\right| \in R_t\right)\\
&= \mathbb{P}_{\mathcal{H}_1}\left(\left|\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \mu\right)\right| > \sqrt{\tilde{\sigma}^2} z_{1-\alpha/2}\right)\\
&= \mathbb{P}_{\mathcal{H}_1}\left(\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \mu\right) > \sqrt{\tilde{\sigma}^2} z_{1-\alpha/2}\right) + \mathbb{P}_{\mathcal{H}_1}\left(\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \mu\right) < -\sqrt{\tilde{\sigma}^2} z_{1-\alpha/2}\right)\\
&= \mathbb{P}_{\mathcal{H}_1}\left(\sqrt{T}\frac{\hat{\theta}^{\mathrm{A2IPW}}_T - \mu - \Delta}{\sqrt{\tilde{\sigma}^2}} > \frac{\sqrt{\tilde{\sigma}^2}}{\sqrt{\tilde{\sigma}^2}} z_{1-\alpha/2} - \frac{\sqrt{T}\Delta}{\sqrt{\tilde{\sigma}^2}}\right) + \mathbb{P}_{\mathcal{H}_1}\left(\sqrt{T}\frac{\hat{\theta}^{\mathrm{A2IPW}}_T - \mu - \Delta}{\sqrt{\tilde{\sigma}^2}} < -\frac{\sqrt{\tilde{\sigma}^2}}{\sqrt{\tilde{\sigma}^2}} z_{1-\alpha/2} - \frac{\sqrt{T}\Delta}{\sqrt{\tilde{\sigma}^2}}\right)\\
&= 1 - \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{T}\Delta}{\sqrt{\tilde{\sigma}^2}}\right) + \Phi\left(-\frac{\sqrt{T}\Delta}{\sqrt{\tilde{\sigma}^2}} - z_{1-\alpha/2}\right)\\
&\geq 1 - \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{T}\Delta}{\sqrt{\tilde{\sigma}^2}}\right).
\end{aligned}$$

Thus, the power is $1 - \Phi\left(z_{1-\alpha/2} - \frac{\sqrt{T}\Delta}{\sqrt{\tilde{\sigma}^2}}\right)$. From this result, it is clear that for $T \geq \frac{\tilde{\sigma}^2}{\Delta^2}\left(z_{1-\alpha/2} - z_\beta\right)^2$, the power becomes at least β. It means that, for achieving the power β, we need $\frac{\tilde{\sigma}^2}{\Delta^2}\left(z_{1-\alpha/2} - z_\beta\right)^2$ samples, i.e.,

$$n^{\mathrm{OPT}*}_\beta(\Delta) = \frac{\tilde{\sigma}^2}{\Delta^2}\left(z_{1-\alpha/2} - z_\beta\right)^2.$$

F.5 Early Stopping under the Optimal Policy

In sequential testing using a LIL-based concentration inequality of this paper, we proposed an algorithm that rejects the null hypothesis when

$$\left|t\hat{\theta}^{\mathrm{A2IPW}}_t - t\mu\right| > 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum^t_{i=1} z^2_i\left(\log\frac{\log\sum^t_{i=1} z^2_i}{\alpha}\right)}\right) = q_t.$$

Let τ be the stopping time of the sequential testing, i.e., $\tau = \min\left\{t : \left|t\hat{\theta}^{\mathrm{A2IPW}}_t - t\mu\right| > q_t\right\}$ When $t = \tau$, it rejects the null hypothesis. In this section, we calculate the upper bound of the expected stopping time τ.

We show that, when sufficient periods passed, the probability that the sequential testing does not reject the hypothesis testing becomes small. Let us bound $\mathbb{P}_{\mathcal{H}_1}(\tau > \tilde{t})$ for sufficiently large $\tilde{t}$ such that $\tilde{t}\Delta \gg 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2C^2\tilde{t}\left(\log\frac{\log C^2\tilde{t}}{\alpha}\right)}\right)$. First, for a stopping time τ, we consider the probability of $\tau \geq \tilde{t}$. Here, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_1}(\tau > \tilde{t}) = 1 - \mathbb{P}_{\mathcal{H}_1}(\tau \leq \tilde{t})\\
&= 1 - \mathbb{P}_{\mathcal{H}_1}\left(\exists t \leq \tilde{t} : \left|t\hat{\theta}^{\mathrm{A2IPW}}_t - t\mu\right| > q_t\right)\\
&\leq 1 - \mathbb{P}_{\mathcal{H}_1}\left(\left|\tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu\right| > q_{\tilde{t}}\right)\\
&= \mathbb{P}_{\mathcal{H}_1}\left(\left|\tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu\right| \leq q_{\tilde{t}}\right)\\
&= \mathbb{P}_{\mathcal{H}_1}\left(-q_{\tilde{t}} \leq \tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu \leq q_{\tilde{t}}\right)\\
&= \mathbb{P}_{\mathcal{H}_1}\left(-q_{\tilde{t}} - \tilde{t}\Delta \leq \tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu - \tilde{t}\Delta \leq q_{\tilde{t}} - \tilde{t}\Delta\right)\\
&\leq \mathbb{P}_{\mathcal{H}_1}\left(\tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu - \tilde{t}\Delta \leq q_{\tilde{t}} - \tilde{t}\Delta\right).
\end{aligned}$$

Then, by substituting $q_{\tilde{t}} = 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum_{i=1}^{\tilde{t}} z_i^2\left(\log\frac{\log\sum_{i=1}^{\tilde{t}} z_i^2}{\alpha}\right)}\right)$,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_1}(\tau > \tilde{t})\\
&\leq \mathbb{P}_{\mathcal{H}_1}\left(\tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu - \tilde{t}\Delta \leq 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum_{i=1}^{\tilde{t}} z_i^2\left(\log\frac{\log\sum_{i=1}^{\tilde{t}} z_i^2}{\alpha}\right)}\right) - \tilde{t}\Delta\right)\\
&= \mathbb{P}_{\mathcal{H}_1}\left(\frac{\tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu - \tilde{t}\Delta}{\sqrt{\tilde{\sigma}^2}} \leq \frac{1.1}{\sqrt{\tilde{\sigma}^2}}\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum_{i=1}^{\tilde{t}} z_i^2\left(\log\frac{\log\sum_{i=1}^{\tilde{t}} z_i^2}{\alpha}\right)}\right) - \frac{\tilde{t}\Delta}{\sqrt{\tilde{\sigma}^2}}\right)\\
&\leq \mathbb{P}_{\mathcal{H}_1}\left(\frac{\tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu - \tilde{t}\Delta}{\sqrt{\tilde{\sigma}^2}} \leq \frac{1.1}{\sqrt{\tilde{\sigma}^2}}\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2C^2\tilde{t}\left(\log\frac{\log C^2\tilde{t}}{\alpha}\right)}\right) - \frac{\tilde{t}\Delta}{\sqrt{\tilde{\sigma}^2}}\right).
\end{aligned}$$

Here, we used $|z_t| \leq C$ for all t. Let $\preceq$ and $\asymp$ be $\leq$ and $=$ when ignoring constants. Then, by using Azuma-Heoffding inequality for martingales (Hoeffding, 1963; Azuma, 1967), $|z_t - z_{t-1}| \leq 2C$, and $\tilde{t}\Delta \gg 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2C^2\tilde{t}\left(\log\frac{\log C^2\tilde{t}}{\alpha}\right)}\right)$,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{H}_1}(\tau > \tilde{t})\\
&\leq \mathbb{P}_{\mathcal{H}_1}\left(\tilde{t}\hat{\theta}^{\mathrm{A2IPW}}_{\tilde{t}} - \tilde{t}\mu - \tilde{t}\Delta \leq 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2C^2\tilde{t}\left(\log\frac{\log C^2\tilde{t}}{\alpha}\right)}\right) - \frac{\tilde{t}\Delta}{\sqrt{\tilde{\sigma}^2}}\right)\\
&\leq \exp\left(-\frac{\left(\tilde{t}\Delta - 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2C^2\tilde{t}\left(\log\frac{\log C^2\tilde{t}}{\alpha}\right)}\right)\right)^2}{8\tilde{t}C^2}\right)\\
&\asymp \exp\left(-\frac{\tilde{t}\Delta^2}{8C^2}\right).
\end{aligned}$$

For $n^{\mathrm{OPT}*}_\beta(\Delta)$, let us assume $n^{\mathrm{OPT}*}_\beta(\Delta)\Delta \gg 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2C^2 n^{\mathrm{OPT}*}_\beta(\Delta)\left(\log\frac{\log C^2 n^{\mathrm{OPT}*}_\beta(\Delta)}{\alpha}\right)}\right)$. This assumption holds when β is sufficiently close to 0. For $n^{\mathrm{OPT}*}_\beta(\Delta) = \frac{\tilde{\sigma}^2}{\Delta^2}\left(z_{1-\alpha/2} - z_\beta\right)^2$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_1}[\tau] &= \sum_{n\geq 1}\mathbb{P}_{\mathcal{H}_1}(\tau > n)\\
&\leq n^{\mathrm{OPT}*}_\beta(\Delta) + \sum_{t\geq n^{\mathrm{OPT}*}_\beta(\Delta)+1}\mathbb{P}_{\mathcal{H}_1}(\tau > t)\\
&\leq n^{\mathrm{OPT}*}_\beta(\Delta) + \sum_{t\geq n^{\mathrm{OPT}*}_\beta(\Delta)-1}\mathbb{P}_{\mathcal{H}_1}(\tau > t)\\
&\preceq n^{\mathrm{OPT}*}_\beta(\Delta) + \sum_{t\geq n^{\mathrm{OPT}*}_\beta(\Delta)-1}^{\infty}\exp\left(-\frac{t\Delta^2}{8C^2}\right)\\
&= n^{\mathrm{OPT}*}_\beta(\Delta) + \exp\left(-\frac{\left(n^{\mathrm{OPT}*}_\beta(\Delta) - 1\right)\Delta^2}{8C^2}\right) + \exp\left(-\frac{n^{\mathrm{OPT}*}_\beta(\Delta)\Delta^2}{8C^2}\right) + \cdots\\
&= n^{\mathrm{OPT}*}_\beta(\Delta) + \exp\left(-\frac{\left(n^{\mathrm{OPT}*}_\beta(\Delta) - 1\right)\Delta^2}{8C^2}\right)\sum_{s=1}^{\infty}\exp\left(-\frac{(s-1)\Delta^2}{8C^2}\right).
\end{aligned}$$

Then by using the infinite geometric series sum formula,

$$\begin{aligned}
&n_\beta^{\mathrm{OPT}*}(\Delta) + \exp\left(-\frac{\left(n_\beta^{\mathrm{OPT}*}(\Delta) - 1\right)\Delta^2}{8C^2}\right)\sum_{s=1}^{\infty}\exp\left(-\frac{(s-1)\Delta^2}{8C^2}\right)\\
&= n_\beta^{\mathrm{OPT}*}(\Delta) + \exp\left(-\frac{\left(n_\beta^{\mathrm{OPT}*}(\Delta) - 1\right)\Delta^2}{8C^2}\right)\frac{1}{1-\exp\left(-\frac{\Delta^2}{8C^2}\right)}\\
&= n_\beta^{\mathrm{OPT}*}(\Delta) + \exp\left(-\frac{n_\beta^{\mathrm{OPT}*}(\Delta)\Delta^2}{8C^2}\right)\frac{1}{\exp\left(\frac{\Delta^2}{8C^2}\right)-1}.
\end{aligned}$$

By substituting $\exp\left(-\frac{\tilde{t}\Delta^2}{8C^2}\right) \asymp \mathbb{P}_{\mathcal{H}_1}(\tau > \tilde{t})$,

$$\mathbb{E}_{\mathcal{H}_1}[\tau] \preceq n_\beta^{\mathrm{OPT}*}(\Delta) + \frac{\mathbb{P}_{\mathcal{H}_1}(\tau > n_\beta^{\mathrm{OPT}*}(\Delta))}{\exp\left(\frac{\Delta^2}{8C^2}\right) - 1}.$$

Using the inequality, $1 - \exp(-r) \leq r$, and $n_\beta^{\mathrm{OPT}*}(\Delta) = \frac{\tilde{\sigma}^2}{\Delta^2}\left(z_{1-\alpha/2} - z_\beta\right)^2$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{H}_1}[\tau]\\
&\preceq n_\beta^{\mathrm{OPT}*}(\Delta) + \frac{8C^2}{\Delta^2}\mathbb{P}_{\mathcal{H}_1}(\tau > n_\beta^{\mathrm{OPT}*}(\Delta))\\
&= n_\beta^{*}(\Delta) + \frac{8C^2 n_\beta^{\mathrm{OPT}*}(\Delta)}{\tilde{\sigma}^2\left(z_{1-\alpha/2} - z_\beta\right)^2}\mathbb{P}_{\mathcal{H}_1}(\tau > n_\beta^{\mathrm{OPT}*}(\Delta))\\
&= \left(1 + \mathrm{O}(1)\right)n_\beta^{\mathrm{OPT}*}(\Delta).
\end{aligned}$$

Thus, we obtain the following corollary.

Corollary 2. *Suppose that* $n_\beta^{\mathrm{OPT}*}(\Delta)\Delta \gg 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2C^2 n_\beta^{\mathrm{OPT}*}(\Delta)\left(\log\frac{\log C^2 n_\beta^{\mathrm{OPT}*}(\Delta)}{\alpha}\right)}\right)$ *and* $\pi_t = \pi^{\mathrm{AIPW}}$. *Then, under* $\mathcal{H}_1$ *and Assumptions 2 and 3, for sufficiently large sample size, the sequential testing using* q_t *has expected stopping time* $\propto n_\beta^{\mathrm{OPT}*}(\Delta)$.

F.6 Minimum Sample Size and Early Stopping under a User-defined Policy

For a user-defined policy π_t, if $\pi_t \xrightarrow{\mathrm{p}} \pi^{\mathrm{A2IPW}}$, we have the same asymptotic variance as $\tilde{\sigma}^2$ from Theorem 1. Therefore, when we use $\pi_t \xrightarrow{\mathrm{p}} \pi^{\mathrm{A2IPW}}$, the minimum sample size required for hypothesis testing is also $n_\beta^{\mathrm{OPT}*}(\Delta)$. By using the same procedure of the previous section, we can easily confirm that the sequential testing under a user-defined policy π_t using q_t has expected stopping time $\propto n_\beta^{\mathrm{OPT}*}(\Delta)$.

G Details of Main Algorithm: AERATE

We show the details of AERATE in Section 5.

G.1 Estimation of $\mathbb{E}\left[Y_t(a) \mid x\right]$ and $\mathbb{E}\left[Y_t^2(a) \mid x\right]$

First, we consider how to estimate $f^*(a,x) = \mathbb{E}\left[Y_t(a) \mid x\right]$ and $e^*(a,x) = \mathbb{E}\left[Y_t^2(a) \mid x\right]$. When estimating $f^*(a,x)$ and $e^*(a,x)$, we need to construct consistent estimators from dependent samples obtained from a adaptive policy. In a MAB problem, several non-parametric estimators are proved to be consistent, such as K-nearest neighbor regression estimator and Nadaraya-Watson kernel regression estimator (Yang & Zhu, 2002; Qian & Yang, 2016). As a example, we show the theoretical properties of K-nearest neighbor regression estimator when using samples with bandit feed back in the following part.

K-nearest neighbor regression: We introduce nonparametric estimation of f^* based on K-nearest neighbor regression using samples with bandit feedback (Yang & Zhu, 2002).

First, we fix $x^* \in \mathcal{X}$. Let $k_n > 0$ be a value depending on the sample size n. Let $N_{t,k}$ be $\sum_{s=1}^{t} \mathbb{1}[A_s = k]$. At t-th round, we gather $N_{t,k}$ samples from the case of $A_{t'} = k$ and reindex the samples as $\{(X_{t'}, Y_{t'})\}_{t'=1}^{N_{t,k}}$. Then, we construct an estimator using the $k_{N_{t,k}}$-NN regression and $\{(X_{t'}, Y_{t'})\}_{t'=1}^{N_{t,k}}$ as follows:

$$\hat{f}_t(k, x^*) = \frac{1}{k_{N_{t,k}}} \sum_{i=1}^{k_{N_{t,k}}} Y^2_{\pi(x^*,i)},$$

where π is the permutation of $\{1, 2, ..., N_{t,k}\}$ such that

$$\|X_{\pi(x^*,1)} - x^*\| \leq \|X_{\pi(x^*,2)} - x^*\| \leq \cdots \leq \|X_{\pi(x^*,N_{t,k})} - x^*\|.$$

For $\hat{f}_{t-1}(k, x)$, Yang & Zhu (2002) showed the following theoretical results. For simplicity, let us assume $\mathcal{X} = [0, 1]^d$ for an integer $d > 0$. First, they put the following assumption.

Assumption 4 (Yang & Zhu (2002), Eq. (5))**.** The function $f^*(k, x)$ be continuous in $x \in \mathcal{X}$ for all $k \in \mathcal{A}$.

Let $\psi(z; f^*(k, \cdot))$ be a modulus of continuity defined by

$$\psi(z; f^*(k, \cdot)) = \sup\left\{|f^*(k, x') - f^*(k, x'')| : |x' - x''|_\infty \leq z\right\}.$$

The term ψ represents the smoothness of the function ν_d.

Assumption 5 (Yang & Zhu (2002), Assumption 2)**.** The probability $p(x)$ is uniformly bounded above and away from 0 on $[0, 1]^d$, i.e., $\underline{c} \leq p(x) \leq \overline{c}$.

Let us assume $Y_t(k) = f^*(k, X_t) + \epsilon_{t,k}$, where $\epsilon_{t,k}$ is a random variable with mean 0 and a finite variable.

Assumption 6 (Yang & Zhu (2002), Assumption 3)**.** The error term $\epsilon_{t,k}$ also satisfies the moment condition such that there exist positive constants v and w satisfying, for all $m \geq 2$,

$$\mathbb{E}[|\epsilon_{t,k}|^m] \leq \frac{m!}{2} v^2 w^{m-2}.$$

Under these assumptions, we can show the following lemma from the result of Yang & Zhu (2002).

Lemma 4 (Yang & Zhu (2002), Eq. (4))**.** *For $\kappa > 0$, let $\eta_\kappa = \sup\{z : \psi(z; f^*(k, \cdot)) \leq \kappa\}$. There exists a constant $M > 0$ such that, for $\kappa > 0$, $h < \eta_{\kappa/4}$, and $k_{N_{t,k}} \leq \underline{c}th^k/2$,*

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{f}_t(k, x^*) - f^*(k, x^*)\right| \geq \kappa\right) \\
&\leq M \exp\left(-\frac{3k_{N_{t,k}}}{14}\right) + \left(t^{d+2} + 1\right)\left(\exp\left(-\frac{3k_{N_{t,k}}\varepsilon}{28}\right) + \exp\left(-\frac{k_{N_{t,k}}\varepsilon^2\kappa^2}{16(v^2 + w\varepsilon\kappa/4)}\right)\right).
\end{aligned}$$

According to Yang & Zhu (2002), for k_t such that $k_t\varepsilon^2/\log t \to \infty$ and $k_{N_{t,k}} = o(t)$, we can choose $h \to 0$ satisfying $h \geq (2k_{N_{t,k}}/(\underline{c}t))^{1/d}$. From the this discussion and the Borel-Cantelli lemma, we can show the following corollary (Yang & Zhu, 2002).

Corollary 3 (Yang & Zhu (2002))**.** *For k_t such that $k_t\varepsilon^2/\log t \to \infty$ and $k_{N_{t,k}} = o(t)$, with probability* 1,

$$\left|\hat{f}_t(k, x^*)) - f^*(k, x^*)\right| \to 0.$$

Besides, when we use $k_{N_{t,k}} = \mathrm{O}(\sqrt{t})$ in our algorithm, which satisfies $k_{N_{t,k}}\varepsilon^2/\log t \to \infty$ and $k_{N_{t,k}} = o(t)$, the following corollary holds.

Corollary 4. *For $k_t = \sqrt{t}$, there exists a constant $M > 0$ such that, for $t > \left(\frac{2}{\underline{c}\eta^k_{\kappa/4}}\right)^2$,*

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{f}_t(k, x^*) - f^*(k, x^*)\right| \geq \kappa\right) \\
&\leq M \exp\left(-\frac{3k_{N_{t,k}}}{14}\right) + \left(t^{d+2} + 1\right)\left(\exp\left(-\frac{3k_t\varepsilon}{28}\right) + \exp\left(-\frac{k_{N_{t,k}}\varepsilon^2\kappa^2}{16(v^2 + w\varepsilon\kappa/4)}\right)\right).
\end{aligned}$$

Using these results, we can bound $\mathbb{E}\left[\left|\hat{f}_t(k,x^*) - f^*(k,x^*)\right|\right]$ by the following lemma.

Lemma 5. *For $\kappa > 0$, $\eta_\kappa = \sup\{z : \psi(z; v_d) \le \kappa\}$, $k_t = \sqrt{t}$, and $t > \left(\frac{2}{\underline{C}\eta^k_{\kappa/4}}\right)^2$, there exists a constant $M > 0$ such that*

$$\begin{aligned}&\mathbb{E}\left[\left|\hat{f}_t(k,x^*) - f^*(k,x^*)\right|\right]\\&\le \kappa + C_2\left(M\exp\left(-\frac{3k_{N_{t,k}}}{14}\right) + \left(t^{d+2}+1\right)\left(\exp\left(-\frac{3k_{N_{t,k}}\varepsilon}{28}\right) + \exp\left(-\frac{k_{N_{t,k}}\varepsilon^2\kappa^2}{16(v^2+w\varepsilon\kappa/4)}\right)\right)\right).\end{aligned}$$

Proof. For $\kappa > 0$, $\eta_\kappa = \sup\{z : \psi(z; v_d) \le \kappa\}$, and $t > \left(\frac{2}{\underline{C}\eta^m_{\kappa/4}}\right)^2$,

$$\begin{aligned}&\mathbb{E}\left[\left|\hat{f}_t(k,x^*) - f^*(k,x^*)\right|\right]\\&\le \kappa + C_2\mathbb{P}\left(\left|\hat{f}_t(k,x^*) - f^*(k,x^*)\right| \ge \kappa\right)\\&\le \kappa + C_2\left(M\exp\left(-\frac{3k_{N_{t,k}}}{14}\right) + \left(T^{d+2}+1\right)\left(\exp\left(-\frac{3k_{N_{t,k}}\varepsilon}{28}\right) + \exp\left(-\frac{k_{N_{t,k}}\varepsilon^2\kappa^2}{16(v^2+w\varepsilon\kappa/4)}\right)\right)\right).\end{aligned}$$

□

Remark 9. The theoretical results of Yang & Zhu (2002) is based on the assumption that the flexibility of the function is restricted and assignment probabilities are > 0 for all actions. Therefore, we can easily check that their results can apply to our case.

G.2 Main Algorithm

The propose algorithm mainly consists of two steps: at a period t, (i) estimate $\nu(k,x)$ and assign an action with the estimated optimal policy, and (ii) conduct testing when sequential testing. Besides, to stabilize the algorithm, we introduce the following three elements: (a) the estimator $\hat{\nu}_{t-1}(k,x)$ of $\nu^*(k,x)$ is constructed as $\max\left(\underline{\nu}, \hat{e}_{t-1}(k,x) - \hat{f}^2_{t-1}(k,x)\right)$, where $\underline{\nu}$ is the lower bound of ν^*, and $\hat{f}_{t-1}$ and $\hat{e}_{t-1}$ are estimators of f^* and e^* only using Ω_{t-1}, respectively; (b) let a policy be $\pi_t(1 \mid x, \Omega_{t-1}) = \gamma\frac{1}{2} + (1-\gamma)\frac{\hat{\nu}_{t-1}(1,x)}{\hat{\nu}_{t-1}(1,x)+\hat{\nu}_{t-1}(0,x)}$, where $\gamma = \mathrm{o}(1/\sqrt{T})$; (c) as a candidate of estimators, we also propose Mixed A2IPW (MA2IPW) estimator defined as $\hat{\theta}^{\mathrm{MA2IPW}}_t = \zeta\hat{\theta}^{\mathrm{AdaIPW}}_t + (1-\zeta)\hat{\theta}^{\mathrm{A2IPW}}_t$, where $\zeta = \mathrm{o}(1/\sqrt{t})$. The motivation of (a) is to prevent $\hat{\nu}_{t-1}$ from taking a negative value. The motivation of (b) is to stabilize the probability of assigning an action. The motivation of (c) is to to control the behavior of estimator by avoiding the situation where $\hat{f}_{t-1}$ takes an unpredicted value in early stage. Because the nonparametric convergence rate is upper bounded by $\mathrm{O}(1/\sqrt{t})$ in general, the convergence rate of policy is also upper bounded by $\mathrm{O}(1/\sqrt{t})$, and $\gamma = \mathrm{o}(1/\sqrt{t})$ does not affect the convergence rate. Similarly, the asymptotic distribution of $\hat{\theta}^{\mathrm{MA2IPW}}_T$ is the same as $\hat{\theta}^{\mathrm{A2IPW}}_T$ because

$$\begin{aligned}&\sqrt{t}\hat{\theta}^{\mathrm{MA2IPW}}_t\\&= \sqrt{t}\left(\zeta\hat{\theta}^{\mathrm{AdaIPW}}_t + (1-\zeta)\hat{\theta}^{\mathrm{A2IPW}}_t\right)\\&= \sqrt{t}\left(\mathrm{o}(1/\sqrt{t})\hat{\theta}^{\mathrm{AdaIPW}}_t + (1-\mathrm{o}(1/\sqrt{t}))\hat{\theta}^{\mathrm{A2IPW}}_t\right)\\&= \sqrt{t}\hat{\theta}^{\mathrm{A2IPW}}_t + \mathrm{o}(1).\end{aligned}$$

Besides, we additionally introduce a hyperparameter ρ, which is technically introduced for initialization. The pseudo code of AERATE is in Algorithm 1.

H Details of Experiments

In this section, we show the effectiveness the proposed algorithm through experiments. We compare the proposed AdaIPW, A2IPW, MA2IPW estimators with an RCT with $p(D_t = 1|X_t) = 0.5$ and the standard IPW, DM, and AIPW estimators. In A2IPW and AIPW estimators, we estimate f^* by K-nearest neighbor regression and Nadaraya-Watson regression. For DM estimator, we used K-nearest neighbor regression and Nadaraya-Watson regression. For three settings of hypothesis testing, we used two datasets; synthetic and semi-synthetic datasets.

Algorithm 1 AERATE

Parameter: Type I error α. Set $\rho \geq 0$, which is the number of samples that we assign treatments with equal probability. Set $\underline{\nu} > 0$, which is the lower bound of the variance ν.
Initialization:
At $t = 1, 2$, select $A_t = t - 1$. Set $\pi_t(1 \mid X_t, \Omega_{t-1}) = 1/2$.
for $t = 3$ to T **do**
  if $t < \rho$ **then**
    Set $\pi_t(1 \mid X_t, \Omega_{t-1}) = 0.5$.
  else
    Construct estimators $\hat{f}_{t-1}$ and $\hat{e}_{t-1}$ using a nonparametric method.
    Construct $\hat{\nu}_{t-1}$ from $\hat{f}_{t-1}$ and $\hat{e}_{t-1}$.
    Using $\hat{\nu}_{t-1}$, construct an estimator of $\pi^{\mathrm{AIPW}}(k \mid X_t)$ and set it as $\pi_t(k \mid X_t, \Omega_{t-1})$.
  end if
  Draw ξ_t from the uniform distribution on $[0, 1]$.
  $A_t = \mathbb{1}[\xi_t \leq \pi_t(1 \mid X_t, \Omega_{t-1})]$.
  if Sequential testing based on LIL **then**
    Construct $\hat{\theta}_t^{\mathrm{A2IPW}}$.
    Construct q_t based on (4.2) with α.
    if $t\hat{\theta}_t^{\mathrm{A2IPW}} > q_t$ **then**
      Reject the null hypothesis.
    end if
  end if
  if Sequential testing based on BF correction **then**
    Construct $\hat{\theta}_t^{\mathrm{A2IPW}}$.
    Construct p-value from $\hat{\theta}_t^{\mathrm{A2IPW}}$ under BF correction.
    if If the p-value is less than α **then**
      Reject the null hypothesis.
    end if
  end if
end for
if Standard hypothesis testing **then**
  Construct $\hat{\theta}_T^{\mathrm{A2IPW}}$.
  Construct p-value from $\hat{\theta}_T^{\mathrm{A2IPW}}$.
  if If the p-value is less than α **then**
    Reject the null hypothesis.
  end if
end if

H.1 Settings of Testing

In each dataset, we conduct the following three patterns of hypothesis testing, the standard hypothesis testing based on T-test, sequential testing based on multiple testing, and sequential testing based on adaptive confidence sequence based on LIL-based concentration inequality. For all settings, the null and alternate hypothesis are $\mathcal{H}_0 : \theta_0 = 0$ and $\mathcal{H}_1 : \theta_0 \neq 0$. When conducting the standard hypothesis testing, we obtain the confidence intervals obtained from T-statistics constructed from the asymptotic distribution of Theorem 1. When conducting the sequential testing based on multiple testing, we conducting testing at $t = 150, 250, 350, 450$ with BF correction. When conducting the sequential testing based on LIL-based concentration inequality, we construct the confidence intervals from q_t of Section 4.

Experiments with Synthetic Data: In addition to Dataset 1 and 2 in Section 6, we used two synthetic datasets. As Section 6, we generated a covariate $X_t \in \mathbb{R}^5$ at each round as $X_t = (X_{t1}, X_{t2}, X_{t3}, X_{t4}, X_{t5})^\top$, where $X_{tk} \sim \mathcal{N}(0, 1)$ for $k = 1, 2, 3, 4, 5$. In this experiment, we used $Y_t(d) = \mu_d + \sum_{k=1}^{5} X_{tk} + e_{td}$ as a model of a potential outcome, where μ_d is a constant, e_{td} is the error term, and $\mathbb{E}[Y_t(d)] = \mu_d$ The error term e_{td} follows the normal distribution, and we denote the standard deviation as std_d. We made two datasets with different μ_d and std_d, Datasets 3–4, with 500 periods (samples). For Datasets 3, we set $\mu_1 = 0.8$ and $\mu_0 = 0.3$ with $\mathrm{std}_1 = 0.6$ and $\mathrm{std}_1 = 0.4$. For Datasets 4, we set $\mu_1 = \mu_0 = 0.5$ with $\mathrm{std}_1 = 0.6$ and $\mathrm{std}_0 = 0.4$. We ran 1000 independent trials for each setting. The results of experiment are shown in Table 1. We show the MSE between θ and $\hat{\theta}$, the standard deviation of MSE (STD), and percentages of rejections of hypothesis testing using T-statistics at the 150th (mid) round and the 300th (final) periods. Besides, we also showed the stopping time of the LIL based algorithm (LIL) and multiple testing with BF correction. When using BF correction, we conducted testing at $t = 150, 250, 350, 450$. In sequential testing, if we do not reject the hypothesis, we return the stopping time as 500. The results are shown in Tables 2 and 3.

Experiments with Semi-Synthetic Data: In evaluation of algorithms for estimating the treatment effect, it is difficult to find 'real-world' data that can be used for the evaluation. Following previous work, we use semi-synthetic datasets made from the Infant Health and Development Program (IHDP), which consists of simulated outcomes and covariate data from a real study. We follow a setting of simulation proposed by Hill (2011). In the setting of Hill (2011), 747 samples with 6 continuous covariates and 19 binary covariates are used. Hill (2011) generated the outcomes using the covariates artificially. Hill (2011) considered two scenario: response surface A and response surface B. In response surface A, Hill (2011) generated $Y_t(1)$ and $Y_t(0)$ as follows:

$$Y_t(0) \sim \mathcal{N}(X_t\boldsymbol{\beta}_A, 1),$$
$$Y_t(1) \sim \mathcal{N}(X_t\boldsymbol{\beta}_A + 4, 1),$$

where elements of $\boldsymbol{\beta}_A \in \mathbb{R}^{25}$ were randomly sampled from $(0, 1, 2, 3, 4)$ with probabilities $(0.5, 0.2, 0.15, 0.1, 0.05)$. In response surface B, Hill (2011) generated $Y_t(1)$ and $Y_t(0)$ as follows:

$$Y_t(0) \sim \mathcal{N}(\exp\left((X_t + W)\boldsymbol{\beta}_B\right), 1),$$
$$Y_t(1) \sim \mathcal{N}(X_t\boldsymbol{\beta}_B - q, 1)$$

where $\boldsymbol{W}$ was an offset matrix of the same dimension as X_t with every value equal to 0.5, q was a constant to normalize the average treatment effect conditional on $d = 1$ to be 4, and elements of $\boldsymbol{\beta}_B \in \mathbb{R}^{25}$ were randomly sampled values $(0, 0.1, 0.2, 0.3, 0.4)$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$. In the experiments, we randomly chose 500 samples from the datasets. We show the MSE between θ and $\hat{\theta}$, the standard deviation of MSE (STD), and percentages of rejections of hypothesis testing using T-statistics at the 150th (mid) round and the 300th (final) periods. Besides, we also showed the stopping time of the LIL based algorithm (LIL) and multiple testing with BF correction. When using BF correction, we conducted testing at $t = 150, 250, 350, 450$. In sequential testing, if we do not reject the hypothesis, we return the stopping time as 500. The results are shown in Tables 4 and 5.

H.2 Sensitivity Analysis of Hyperparamters

Using Dataset 1 of Section 6, we investigate the sensitivity of the performances against the hyperparameters γ, ζ, and ρ. We compared A2IPW and MA2IPW estimators with Nadara-Watson estimator under various hyperparameters with Hahn 50, Hahn 100, and OPT defined in Section 6. The results are shown in 6. In all cases, the proposed estimators outperforms the existing methods.

H.3 Interpretations

Finally, we discuss the results of each estimator.

DM estimator: First of all, we discuss the results of DM estimator. In almost all experiments, the DM estimator rejects the null hypothesis with smallest samples. However, it also tend to reject the null hypothesis even when the null hypothesis is true. Besides, the MSE of DM estimator is larger than the other methods. Therefore, decision making based on DM estimator might lead us to wrong decision.

Two Step Adaptive Experimental Design: The two step adaptive experimental design proposed by Hahn et al. (2011) also shows preferable performance. However, compared with the proposed method of this paper, the performance seems sub-optimal. We consider that this is because the method cannot reduce the estimation error of the optimal policy after the first stage of the experiment. Therefore, after the first stage of the experiment, the estimation error will remain and it reduces the performance. In experiment using IHDP dataset with surface B, the MSE is less than the proposed method of this paper. However, as shown in Table 5, the sample used in the proposed method in the experiment is 50 samples less than that of the method of Hahn et al. (2011) in LIL and 15 samples less than that of the method of Hahn et al. (2011) in BF. This is because the MSE of the proposed method is smaller than the method of Hahn et al. (2011) in earlier stage than $t = 150$. We show the MSEs of $t = 100, 200, 300, 400$ in Table 7. This is because the proposed method does not require the first stage to estimate the optimal policy and can start assigning treatments following the estimated optimal start from earlier stage.

LIL and BF: In the experiments, the sequential testing based on BF correction seems succeed hypothesis testing using less samples than the sequential testing based on LIL-based concentration inequality. However, BF based sequential testing also tend to reject the null hypothesis even when the null hypothesis is correct (Table 1 and 3). Therefore, because there is a possibility that the BF-based sequential testing just increases the Type I error, it is also difficult to decide which method is better.

Remark 10 (Standard and Sequential Hypothesis Testing)**.** The remaining question is whether to use standard or sequential hypothesis testing. When we want to reject the null hypothesis with a smaller sample size, the sequential hypothesis testing might be better. However, in the case where the null hypothesis is true, the sequential testing may not stop if there are infinite samples. Moreover, unlike the standard hypothesis testing, it is not easy to calculate the sample size. On the other hand, when using the standard hypothesis testing, we can control the test by deciding the sample size. Thus, each of these methods has advantages and disadvantages, and it is necessary to decide which to use for each application.

Table 2: Experimental results using Datasets 3. The best performing method is in bold.

	Dataset 3: $\mathbb{E}[Y(1)] = 0.8$, $\mathbb{E}[Y(0)] = 0.3$, $\mathrm{std1} = 0.6$, $\mathrm{std0} = 0.4$, $\theta_0 \neq 0$							
	$T = 150$			$T = 300$			ST	
	MSE	STD	Testing	MSE	STD	Testing	LIL	BF
RCT	0.139	0.191	24.2%	0.069	0.102	44.8%	450.1	371.7
A2IPW (K-nn)	0.089	0.127	39.0%	0.042	0.064	69.8%	385.8	296.6
A2IPW (NW)	0.061	0.089	53.8%	**0.024**	0.033	90.3%	290.5	230.4
MA2IPW (K-nn)	0.087	0.121	42.6%	0.040	0.054	70.2%	378.1	291.4
MA2IPW (NW)	**0.060**	0.083	53.1%	0.025	0.035	90.8%	292.6	233.6
AdaIPW (K-nn)	0.158	0.214	26.3%	0.076	0.110	46.0%	443.2	365.6
AdaIPW (NW)	0.147	0.202	25.1%	0.080	0.112	46.1%	440.0	367.6
DM (K-nn)	0.167	0.237	**90.3%**	0.084	0.120	96.0%	**57.3**	**162.6**
DM (NW)	0.109	0.156	83.2%	0.044	0.065	**96.8%**	116.8	173.0
Hahn 50 (K-nn)	0.109	0.152	37.1%	0.049	0.064	65.3%	384.3	312.8
Hahn 50 (NW)	0.080	0.110	44.1%	0.029	0.041	85.4%	306.4	255.2
Hahn 100 (K-nn)	0.133	0.179	30.7%	0.050	0.072	59.7%	409.2	330.6
Hahn 100 (NW)	0.101	0.138	30.1%	0.030	0.041	78.0%	362.8	292.6
OPT	0.007	0.010	100.0%	0.003	0.005	100.0%	55.8	150.0

Table 3: Experimental results using Datasets 4. The best performing method is in bold.

	Dataset 3: $\mathbb{E}[Y(1)] = 0.5$, $\mathbb{E}[Y(0)] = 0.5$, $\mathrm{std1} = 0.6$, $\mathrm{std0} = 0.4$, $\theta_0 \neq 0$							
	$T = 150$			$T = 300$			ST	
	MSE	STD	Testing	MSE	STD	Testing	LIL	BF
RCT	0.081	0.117	4.5%	0.041	0.056	**3.5%**	496.3	484.0
A2IPW (K-nn)	0.053	0.073	6.2%	0.024	0.035	5.1%	496.8	474.1
A2IPW (NW)	**0.031**	0.044	5.2%	0.012	0.017	6.1%	495.6	477.0
MA2IPW (K-nn)	0.048	0.065	5.1%	0.024	0.035	4.9%	495.8	477.5
MA2IPW (NW)	0.029	0.042	4.3%	**0.011**	0.015	4.4%	**498.1**	477.6
AdaIPW (K-nn)	0.091	0.120	4.7%	0.048	0.067	6.1%	496.0	475.2
AdaIPW (NW)	0.098	0.132	5.1%	0.049	0.066	5.9%	497.2	474.6
DM (K-nn)	0.101	0.155	84.1%	0.049	0.075	87.2%	102.9	190.4
DM (NW)	0.057	0.086	53.6%	0.023	0.034	57.6%	299.9	306.1
Hahn 50 (K-nn)	0.054	0.076	4.5%	0.025	0.034	5.4%	492.7	474.2
Hahn 50 (NW)	0.033	0.047	4.9%	0.014	0.018	5.4%	495.3	480.2
Hahn 100 (K-nn)	0.065	0.092	5.8%	0.028	0.040	5.4%	495.1	472.6
Hahn 100 (NW)	0.041	0.055	**3.8%**	0.014	0.019	**3.5%**	496.5	**484.8**
OPT	0.004	0.005	4.5%	0.002	0.003	4.5%	497.4	482.3

Table 4: Experimental results using IHDP dataset with surface A. The best performing method is in bold.

	IHDP dataset with surface A, $\theta_0 = 4 \neq 0$							
	$T = 150$			$T = 300$			ST	
	MSE	STD	Testing	MSE	STD	Testing	LIL	BF
RCT	0.674	1.066	60.4%	0.333	0.562	93.4%	355.4	228.0
A2IPW (K-nn)	0.606	0.891	99.6%	0.310	0.500	**100.0%**	86.3	150.5
A2IPW (NW)	0.485	0.740	99.8%	**0.202**	0.311	**100.0%**	76.2	150.2
MA2IPW (K-nn)	0.599	0.961	99.5%	0.275	0.432	**100.0%**	84.6	150.5
MA2IPW (NW)	**0.484**	0.688	99.9%	0.214	0.317	**100.0%**	74.7	**150.1**
AdaIPW (K-nn)	3.287	5.293	63.7%	1.626	2.681	84.8%	293.6	231.8
AdaIPW (NW)	3.694	6.056	61.5%	1.770	2.896	84.7%	302.6	231.1
DM (K-nn)	1.138	1.745	99.9%	0.578	0.892	**100.0%**	15.1	**150.1**
DM (NW)	0.999	1.427	**100.0%**	0.454	0.623	**100.0%**	26.4	150.0
Hahn 50 (K-nn)	0.725	1.164	93.7%	0.320	0.491	**100.0%**	165.9	156.7
Hahn 50 (NW)	0.563	0.872	95.8%	0.277	0.433	**100.0%**	154.5	154.2
Hahn 100 (K-nn)	0.748	1.217	79.4%	0.314	0.494	99.9%	214.6	173.2
Hahn 100 (NW)	0.534	0.775	82.6%	0.238	0.341	**100.0%**	204.6	168.1

Table 5: Experimental results using IHDP dataset with surface B. The best performing method is in bold.

	IHDP dataset with surface B, $\theta_0 \neq 0$							
	$T = 150$			$T = 300$			ST	
	MSE	STD	Testing	MSE	STD	Testing	LIL	BF
RCT	4.522	19.635	53.9%	2.492	9.903	72.7%	355.3	274.4
A2IPW (K-nn)	5.153	33.698	84.5%	2.683	13.545	90.6%	147.7	186.2
A2IPW (NW)	4.379	23.713	84.3%	**2.198**	11.874	91.0%	142.9	185.0
MA2IPW (K-nn)	4.797	21.194	83.9%	2.496	10.330	90.7%	145.5	186.8
MA2IPW (NW)	4.721	18.190	84.3%	2.724	13.127	90.9%	144.%	184.4
AdaIPW (K-nn)	11.376	44.898	55.4%	6.658	29.222	71.5%	308.0	265.6
AdaIPW (NW)	11.674	45.069	56.6%	5.428	15.496	70.9%	311.7	264.4
DM (K-nn)	7.065	23.954	**98.1%**	3.892	14.737	**98.8%**	**18.7**	**152.1**
DM (NW)	7.410	30.313	94.1%	3.821	16.227	96.5%	53.0	162.6
Hahn 50 (K-nn)	4.309	14.939	76.5%	2.190	7.920	89.0%	211.6	200.3
Hahn 50 (NW)	4.650	19.511	75.5%	2.649	12.263	88.1%	209.7	203.4
Hahn 100 (K-nn)	**3.627**	13.561	64.4%	2.985	19.012	85.9%	256.9	224.1
Hahn 100 (NW)	3.858	16.541	66.5%	2.536	16.547	86.8%	251.5	217.7

Table 6: Experimental results of sensitivity analysis using Dataset 1.

				$T = 150$			$T = 300$			ST	
	γ	ζ	ρ	MSE	STD	Testing	MSE	STD	Testing	LIL	BF
A2IPW	$t^{-1/2}$	-	50	0.064	0.092	51.4%	0.025	0.035	88.1%	303.8	239.8
A2IPW	$t^{-1/1.5}$	-	50	0.063	0.091	51.2%	0.025	0.037	89.5%	303.7	240.0
A2IPW	t^{-1}	-	50	0.062	0.090	50.8%	0.024	0.035	88.8%	306.5	239.2
A2IPW	$t^{-1/1.5}$	-	10	0.073	0.106	47.9%	0.027	0.037	84.6%	324.2	254.7
A2IPW	t^{-1}	-	10	0.072	0.098	42.2%	0.028	0.039	83.1%	333.4	265.0
MA2IPW	$t^{-1/2}$	$t^{-1/1.5}$	50	0.062	0.085	**52.7%**	**0.023**	0.033	**90.2%**	303.3	**236.6**
MA2IPW	$t^{-1/1.5}$	t^{-1}	50	0.064	0.094	52.2%	0.025	0.035	88.7%	**301.5**	240.5
MA2IPW	$t^{-1/1.5}$	t^{-2}	50	**0.055**	0.074	49.4%	0.024	0.032	88.4%	311.8	243.4
MA2IPW	t^{-1}	t^{-1}	50	0.064	0.087	49.2%	**0.023**	0.031	86.8%	310.9	245.6
MA2IPW	t^{-1}	t^{-2}	50	0.062	0.093	49.2%	0.024	0.034	88.7%	309.3	245.0
MA2IPW	$t^{-1/1.5}$	t^{-1}	10	0.067	0.096	47.6%	0.025	0.036	86.3%	319.8	250.6
MA2IPW	$t^{-1/1.5}$	t^{-2}	10	0.069	0.092	45.9%	0.028	0.038	84.8%	322.8	254.1
MA2IPW	t^{-1}	t^{-1}	10	0.074	0.105	48.4%	0.027	0.037	84.6%	324.6	253.3
MA2IPW	t^{-1}	t^{-2}	10	0.071	0.103	46.2%	0.026	0.038	84.7%	326.0	254.7
Hahn 50	-	-	50	0.085	0.128	45.7%	0.033	0.046	82.8%	313.1	257.0
Hahn 100	-	-	100	0.107	0.146	32.1%	0.036	0.050	75.2%	365.3	294.6
OPT	-	-	-	0.007	0.011	100.0%	0.004	0.006	100.0%	64.1	150.0

Table 7: Experimental results of MSEs in IHDP dataset with surface B.

	$T = 100$		$T = 200$		$T = 300$		$T = 400$	
	MSE	STD	MSE	STD	MSE	STD	MSE	STD
RCT	8.491	3.605	2.492	9.903	4.522	9.903	4.522	9.903
A2IPW (K-nn)	**7.232**	5.172	2.683	13.545	5.153	13.545	5.153	13.545
A2IPW (NW)	7.256	3.361	2.198	11.874	4.379	11.874	4.379	11.874
MA2IPW (K-nn)	8.917	2.463	2.496	10.330	4.797	10.330	4.797	10.330
MA2IPW (NW)	9.003	3.768	2.724	13.127	4.721	13.127	4.721	13.127
AdaIPW (K-nn)	17.088	10.332	6.658	29.222	11.376	29.222	11.376	29.222
AdaIPW (NW)	16.873	9.245	5.428	15.496	11.674	15.496	11.674	15.496
DM (K-nn)	9.323	9.768	3.892	14.737	7.065	14.737	7.065	14.737
DM (NW)	10.128	10.429	3.821	16.227	7.410	16.227	7.410	16.227
Hahn 50 (K-nn)	8.323	2.632	**2.190**	7.920	**4.309**	7.920	4.309	7.920
Hahn 50 (NW)	9.543	3.889	2.649	12.263	4.650	12.263	4.650	12.263
Hahn 100 (K-nn)	9.249	3.953	2.985	19.012	3.627	19.012	**3.627**	19.012
Hahn 100 (NW)	8.674	5.507	2.536	16.547	3.858	16.547	3.858	16.547