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:

$$\hat{\theta}_T^{\mathrm{A2IPW}} =$$

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

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.,

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

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),

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

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

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

Then, from the triangle inequality, we have

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

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

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

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

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

Then, as $t \to \infty$,

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

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

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

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

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

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

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

Next, we show that

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

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

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

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

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

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

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

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

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

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.

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

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

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

We use the following result:

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

Therefore,

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

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

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

On the other hand, we have

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

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

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

Therefore,

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

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

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

Therefore, we have

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

□

D.1 Proof of Theorem 4

The procedure of this proof mainly follows Balsubramani & Ramdas (2016). For a martingale M_t, let $V_t=\sum_{i=1}^{t}\mathbb{E}\big[(M_i-M_{i-1})^2\mid\Omega_{i-1}\big]$. 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 $t\geq 1$, $\big|M_t-M_{t-1}\big|\leq e^2/2$ with probability 1. Fix any $\delta>0$, and define $\tau_0=\min\big\{s:2(e-2)V_s\geq$*

$173 \log\left(\frac{4}{\delta}\right) \big\}$*, 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,*

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

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*

$$V_t \leq C_3 \left(\sum_{i=1}^{t}(M_i - M_{i-1})^2 + \frac{2C_4 C_0(\delta)}{e^2}\right),$$

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

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

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 $\mathbb{P}(\tau_1 < \tau_0) \le \frac{\delta}{2} < \delta$. □

Proof of Lemma 2

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

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

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

$$\left|M_t\right| \le \max\left(3(e-2), \sqrt{2C_1V_t\log\log V_t + C_1V_t\log\left(\frac{4}{\delta}\right)}\right), \tag{3}$$

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

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

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

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

$$\left|M_t\right| \le \begin{cases} 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right) & \text{if } t \le \tau_0 \\ \frac{2(e-2)}{e^2(1+\sqrt{1/3})}V_t \text{ and } \max\left(3(e-2), \sqrt{2C_1V_t\log\log V_t + C_1V_t\log\left(\frac{4}{\delta}\right)}\right) & \text{if } t \ge \tau_0. \end{cases}$$

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

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

□

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 $\left|(M_t - M_{t-1})^2 - \mathbb{E}\left[(M_i - M_{i-1})^2 \mid \Omega_{i-1}\right]\right| \le C_4$ with probability 1 in which the existence is guaranteed by the boundedness of $M_i - M_{i-1}$, i.e., $|M_i - M_{i-1}| \le e^2/2$ 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 $\ge 1-\delta$

$$\left|\tilde{M}_t\right| \le \frac{2C_4}{e^2}\left(C_0(\delta) + \sqrt{C_1B_t\left(2\log\log B_t + \log\left(\frac{4}{\delta}\right)\right)}\right),$$

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

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

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

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

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 $2C_4^2 C_1 \log\log\left(e^2 C_4 C_0(\delta) + e^4 \sum_{i=1}^{t}(M_i - M_{i-1})^2/2\right) \geq C_4^2 C_1 \log\left(\frac{4}{\delta}\right)$, by using a constant C_5, the RHS is bounded as

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

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):

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

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 $\pi^{\mathrm{e}} : \mathcal{A} \times \mathcal{X} \to (0, 1)$ 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 $R(\pi^{\mathrm{e}}) := \mathbb{E}\left[\sum_{k=1}^{K} \pi^{\mathrm{e}}(k, X_t)Y_t(k)\right]$. We also denote $R(\pi^{\mathrm{e}})$ 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 $\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$, where $\sigma^2 = \mathbb{E}\left[\sum_{k=1}^{K} \frac{\left(\pi^{\mathrm{e}}\left(k, X_t\right)\right)^2 \nu^*\left(k, X_t\right)}{\tilde{\pi}(k \mid X_t)} + \left(\sum_{k=1}^{K} \pi^{\mathrm{e}}(k, X_t) f^*(k, X_t) - \theta_0\right)^2\right]$.

F Further Discussion of Related Work

In this section, we review the details of related work.

Two-Stage Adaptive 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. In this paper, we consider an extension of the method of Hahn et al. (2011) to by introducing sequential policy updating. However, there are three essential differences between the methods of Hahn et al. (2011) and those in this paper. 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, we can derive the finite sample analysis for the proposed estimator, but we cannot obtain the finite sample results for the method of Hahn et al. (2011) because this depends on the asymptotic property of the first stage. Thus, our method is an extension of the method of Hahn et al. (2011), but is quite different. Moreover, we can regard the method of Hahn et al. (2011) as a special case of our method. We can also apply the method of Hahn et al. (2011) and our proposed method simultaneously, i.e., even after estimating the optimal policy in the first stage of the experiment, we can continue to update the estimated policy in the second stage without loss of statistical property required for hypothesis testing.

Targeted Adaptive Design The *targeted adaptive design* is proposed by van der Laan (2008), which tries to minimize the asymptotic variance by sequentially optimizing the assignment probabilities. For overcoming the problem of dependency, they also constructed an estimator from an MDS. Thus, our study and the method proposed by van der Laan (2008) are quite similar, but there are several differences. First, we propose using the A2IPW estimator, which allows us to use a wide class of models for f^*. In contrast, in the method of van der Laan (2008), for deriving the asymptotic normality, we need to restrict the models. In addition, we also show the examples of the consistency of f^* based on the arguments of Yang & Zhu (2002), but van der Laan (2008) did not. Second, in this paper, we also proposed the method for sequential testing based on non-asymptotic concentration inequality. On the other hand, van der Laan (2008) only considers the hypothesis testing based on the asymptotic normality of an estimator.

Best Arm Identification: Our method also has a close relationship with the best arm identification in the MAB problem. The best arm identification is a *pure exploration* problem over multiple bandits, and the goal is to detect the best arm with high probability. The best arm identification with covariates also has recently garnered attention (Soare et al., 2014). In the best arm identification 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, if our interest is in hypothesis testing, the problem setting in this paper can be regarded as a novel case of the best arm identification in the MAB problem. This setting can be called *semiparametric best arm identification*.

When we have an interest in the estimation of the ATE itself, our problem setting can be considered as a novel setting of pure exploration over two-armed bandits, which is different from the best arm identification.

Causal Inference from Dependent Samples: When estimating the ATE from samples obtained via an adaptive policy, we cannot use standard methods of statistical inference. In this context, there are three approaches. In the first approach, the policy determined from past observations converges to a time-invariant policy in probability, as in this paper and (Hadad et al., 2019). In the second approach, the batch updating of a policy is assumed; that is, although the policy is updated using past observations, there are sufficient samples under a fixed policy. In the third approach, in addition to the stationarity of samples, we assume the independence of time-separated samples (Kallus & Uehara, 2019). Theoretically, the first and second approaches use martingale theory (Hall et al., 2014), whereas the third approach uses mixingale theory (Kosorok, 2008). Independently of this paper, Hadad et al. (2019) also derived a similar estimator based on martingale theory. However, in their work, several points need to be fixed. First, they assumed that the adaptive policy π_t and a function $\hat{f}_{t-1}$ converge almost surely to the time-invariant function, but this assumption is superfluous. As we showed, we can prove asymptotic normality only assuming point convergence in the probability of π_t and a function $\hat{f}_{t-1}$. Second, they stated that their estimator has asymptotic normality if either estimators $\hat{f}_{t-1}$ or the treatment assignment probabilities π_t are consistent, i.e., $\hat{f}_{t-1}(k,x) \xrightarrow{\mathrm{p}} \tilde{f}(k,x)$, where $\tilde{f}$ is a time-invariance function, or $\pi_t \xrightarrow{\mathrm{p}} \tilde{\pi}$. However, to show asymptotic normality, as we showed in Theorem 1, we need the point convergences of both $\hat{f}_{t-1}$ and π_t for the asymptotic normality of an estimator of the ATE. Third, although we can derive the asymptotic variance explicitly as σ^2 in this paper, they did not. As a result, they also did not discuss the semiparametric lower bound. While we cannot define the usual semiparametric lower bound in the problem setting, we can consider the semiparametric lower bound under a time-invariant policy, as we discussed in Section 3. Fourth, they did not introduce the covariate X_t and only mentioned that the derivation of the asymptotic distribution with covariate X_t is straightforward. However, as we showed, the derivation is not so trivial. Fifth, they proposed stabilizing their proposed estimator using adaptive weights, whereas we proposed using a combination of the A2IPW and AdaIPW estimators. The proposition of weight matrix is one of the main contribution of Hadad et al. (2019) for stabilization of the initial periods, but we suggest a solution to the same problem by adjusting π_t itself and proposition of MA2IPW estimator.

Ethics and Fairness: While RCT is a reliable framework for scientific experiments, it has some ethical problems (Nardini, 2014,?). In addition, simple randomization sometimes obtains unfair results. On the contrary, compared with adaptive randomization based on past observations such as the algorithms of the MAB problem, an RCT with completely random assignment might be fairer because we do not manipulate the assignment based on the covariates of the research subjects. Thus, ethics and fairness in RCTs and adaptive experimental design is a critical problem.

In the proposed algorithm, we allocate the treatment based on the standard deviation of samples. If this seems unfair, we can incorporate some fairness criteria as a constraint into the minimization that appeared in Proposition 1, which determines the optimal policy. 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 in Theorem 1. This idea is also suggested by Hahn et al. (2011).

For another approach, Narita (2018) proposed using mechanism design for designing the RCT. Based on the preferences of research subjects, the method randomly assigns each treatment Pareto optimally and is also asymptotically incentive compatible for preference elicitation. As a future direction, we could incorporate the method of Narita (2018) into the method proposed in this paper.

G 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 is 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 $\beta(\theta) = \mathbb{P}_\theta(\hat{\theta}^{\mathrm{A2IPW}}_t \in R)$, where R is a rejection region, where, if $\hat{\theta}^{\mathrm{A2IPW}}_t \in R$, 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 type 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.

G.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

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

G.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 suboptimal 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.

G.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 $\hat{\theta}_n^{\mathrm{A2IPW}} \in R_n$ 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 $n^*_\beta(\Delta) = \min\left\{n : \mathbb{P}_{\mathcal{H}_1(\Delta)}\left(\hat{\theta}_n^{\mathrm{A2IPW}} \in R_n\right) \geq 1-\beta\right\}$, 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.

G.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

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

where

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

$\ddot{f}(k, X_t) = \mathbb{E}_{\mathcal{H}_1}[Y_t(k) \mid X_t]$, $\ddot{\nu}^*\big(k, X_t\big) = \mathbb{E}_{\mathcal{H}_1}[\big(Y_t(k) - \ddot{f}(k, X_t)\big)^2 \mid X_t]$, $\ddot{\pi} = \frac{\sqrt{\ddot{\nu}^*\big(1,X_t\big)}}{\sqrt{\ddot{\nu}^*\big(1,X_t\big)}+\sqrt{\ddot{\nu}^*\big(0,X_t\big)}}$, 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

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

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 G.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 $T \geq \frac{\tilde{\sigma}^2}{\Delta^2}\big(z_{1-\alpha/2} - z_\beta\big)^2$, the power becomes at least β. It means that, for achieving the power β, we need $\frac{\tilde{\sigma}^2}{\Delta^2}\big(z_{1-\alpha/2} - z_\beta\big)^2$ samples, i.e.,

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

G.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 $q_{\tilde{t}} = 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum^{\tilde{t}}_{i=1}z^2_i\left(\log\frac{\log\sum^{\tilde{t}}_{i=1}z^2_i}{\alpha}\right)}\right)$,

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

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,

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

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

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

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

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

Thus, we obtain the following corollary.

Corollary 2. *Suppose that* $n^{\mathrm{OPT}*}_{\beta}(\Delta)\Delta\gg 1.1\left(\log\left(\frac{1}{\alpha}\right)+\sqrt{2C^2n^{\mathrm{OPT}*}_{\beta}(\Delta)\left(\log\frac{\log C^2n^{\mathrm{OPT}*}_{\beta}(\Delta)}{\alpha}\right)}\right)$ *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* $\propto n^{\mathrm{OPT}*}_{\beta}(\Delta)$.

G.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 $n^{\mathrm{OPT}*}_{\beta}(\Delta)$. 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 $\propto n^{\mathrm{OPT}*}_{\beta}(\Delta)$.

H Details of Main Algorithm: AERATE

We show the details of AERATE in Section 5.

H.1 Estimation of $\mathbb{E}\big[Y_t(a)\mid x\big]$ and $\mathbb{E}\big[Y^2_t(a)\mid x\big]$

First, we consider how to estimate $f^*(a,x)=\mathbb{E}\big[Y_t(a)\mid x\big]$ and $e^*(a,x)=\mathbb{E}\big[Y^2_t(a)\mid x\big]$. 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 $\sum^t_{s=1}\mathbb{1}[A_s=k]$. At t-th round, we gather $N_{t,k}$ samples from the case of $A_{t'}=k$ and reindex the samples as $\{(X_{t'},Y_{t'})\}^{N_{t,k}}_{t'=1}$. Then, we construct an estimator using the $k_{N_{t,k}}$-NN regression and $\{(X_{t'},Y_{t'})\}^{N_{t,k}}_{t'=1}$ as follows:

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

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

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

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., $\underline{c} \le p(x) \le \overline{c}$.

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 $m \ge 2$,

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

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 $\eta_\kappa = \sup\{z : \psi(z; f^*(k, \cdot)) \le \kappa\}$. There exists a constant $M > 0$ such that, for $\kappa > 0$, $h < \eta_{\kappa/4}$, and $k_{N_{t,k}} \le \underline{c} t h^k / 2$,*

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

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 $h \ge (2k_{N_{t,k}} / (\underline{c}t))^{1/d}$. 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 $t > \left(\frac{2}{\underline{c} \eta^k_{\kappa/4}} \right)^2$,*

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

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

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

□

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.

H.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 lower 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.

I 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.

I.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

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 $\hat{\theta}^{\mathrm{A2IPW}}_t$.
    Construct q_t based on (4.2) with α.
    if $t\hat{\theta}^{\mathrm{A2IPW}}_t > q_t$ **then**
      Reject the null hypothesis.
    end if
  end if
  if Sequential testing based on BF correction **then**
    Construct $\hat{\theta}^{\mathrm{A2IPW}}_t$.
    Construct p-value from $\hat{\theta}^{\mathrm{A2IPW}}_t$ 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 $\hat{\theta}^{\mathrm{A2IPW}}_T$.
  Construct p-value from $\hat{\theta}^{\mathrm{A2IPW}}_T$.
  if If the p-value is less than α **then**
    Reject the null hypothesis.
  end if
end if

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 $Y_t(d) = \mu_d + \sum^5_{k=1} X_{tk} + e_{td}$ 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.

I.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.

I.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