A Mathematical preliminaries

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

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

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

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

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

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

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

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

Remark 6. *The weak law of large numbers for martingale holds when the random variable is bounded by a constant.*

B Proof of Theorem 1

Proof of Theorem 1. We show asymptotic normality of

$$\widehat{R}_T^{\mathrm{ADR}}(\pi^{\mathrm{e}}) = \frac{1}{T}\sum_{t=1}^T \left\{\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) + \phi_2(X_t; \hat{f}_{t-1})\right\},$$

where

$$\phi_1(X_t, A_t, Y_t; g, f) = \sum_{a=1}^K \frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t = a]\left(Y_t - f(a, X_t)\right)}{g(a|X_t)}$$

$$\phi_2(X_t; f) = \sum_{a=1}^K \pi^{\mathrm{e}}(a|X_t) f(a, X_t).$$

Let us define an AIPW estimator with $\hat{f} = f^*$ as

$$\widehat{R}^*(\pi^{\mathrm{e}}) = \frac{1}{T}\sum_{t=1}^T \left\{\phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) + \phi_2(X_t; f^*)\right\}.$$

We decompose $\sqrt{T}\left(R^{\mathrm{ADRE}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right)$ as

$$\sqrt{T}\left(R^{\mathrm{ADRE}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right) = \sqrt{T}\left(\widehat{R}_T^{\mathrm{ADR}}(\pi^{\mathrm{e}}) - \widehat{R}^*(\pi^{\mathrm{e}}) + \widehat{R}^*(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right).$$

From Proposition 1 of Kato et al. (2020) and Assumption 1 and 3, because $\sqrt{T}\left(\widehat{R}^*(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right)$ follows asymptotic normal distribution, we want to show

$$\widehat{R}_T^{\mathrm{ADR}}(\pi^{\mathrm{e}}) - \widehat{R}^*(\pi^{\mathrm{e}}) = \mathrm{o}_p(1/\sqrt{T}).$$

Here, we have

$$
\begin{aligned}
&\widehat{R}_T^{\mathrm{ADR}}(\pi^{\mathrm{e}}) - \widehat{R}^*(\pi^{\mathrm{e}}) \\
&= \frac{1}{T}\sum_{t=1}^{T}\Bigg\{\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
&\quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right]\Bigg\} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1})|\Omega_{t-1}\right] + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1})|\Omega_{t-1}\right] \\
&\quad - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_2(X_t; f^*)|\Omega_{t-1}\right].
\end{aligned}
$$

In the following parts, we separately show that

$$
\begin{aligned}
&\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Bigg\{\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
&\quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right]\Bigg\} \\
&= \mathrm{o}_p(1);
\end{aligned} \tag{1}
$$

and

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1})|\Omega_{t-1}\right] + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1})|\Omega_{t-1}\right] \\
&- \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_2(X_t; f^*)|\Omega_{t-1}\right] = \mathrm{o}_p(1/\sqrt{T}).
\end{aligned} \tag{2}
$$

Proof of (1)**.** For any $\varepsilon > 0$, to show that

$$
\begin{aligned}
&\mathbb{P}\Bigg(\Bigg|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Bigg\{\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
&\quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right]\Bigg\}\Bigg| > \varepsilon\Bigg) \\
&\to 0,
\end{aligned}
$$

we show that the mean is 0 and the variance of the component converges to 0. Then, from the Chebyshev's inequality, this result yields the statement.

The mean is calculated as

$$
\begin{aligned}
&\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Bigg[\Bigg\{\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)\\
&\quad-\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]\\
&\quad+\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)-\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right]\Bigg\}\Bigg]\\
&=\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg\{\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)\\
&\quad-\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]\\
&\quad+\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)-\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right]\Bigg\}|\Omega_{t-1}\Bigg]\Bigg]\\
&=0
\end{aligned}
$$

Because the mean is 0, the variance is

$$
\begin{aligned}
&\mathrm{Var}\Bigg(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Bigg\{\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)\\
&\quad-\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]\\
&\quad+\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)-\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right]\Bigg\}\Bigg)\\
&=\mathbb{E}\Bigg[\Bigg(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Bigg\{\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)\\
&\quad-\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]\\
&\quad+\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)-\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right]\Bigg\}\Bigg)^2\Bigg]\\
&=\frac{1}{T}\mathbb{E}\Bigg[\Bigg(\sum_{t=1}^{T}\Bigg\{\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)\\
&\quad-\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]\\
&\quad+\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)-\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right]\Bigg\}\Bigg)^2\Bigg].
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Bigg[\Bigg(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
&\quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right]\Bigg)^2\Bigg] \\
&\quad + \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}\Bigg[\Bigg(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
&\quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right]\Bigg) \\
&\quad \times \Bigg(\phi_1(X_s, A_s, Y_s; \hat{g}_{s-1}, \hat{f}_{s-1}) - \phi_1(X_s, A_s, Y_s; \pi_{s-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_s, A_s, Y_s; \hat{g}_{s-1}, \hat{f}_{s-1}) - \phi_1(X_s, A_s, Y_s; \pi_{s-1}, f^*)|\Omega_{s-1}\right] \\
&\quad + \phi_2(X_s; \hat{f}_{s-1}) - \phi_2(X_s; f^*) - \mathbb{E}\left[\phi_2(X_s; \hat{f}_{s-1}) - \phi_2(X_s; f^*)|\Omega_{s-1}\right]\Bigg)\Bigg].
\end{aligned}
$$

For $s > t$, we can vanish the covariance terms as

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Bigg(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
&\quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right]\Bigg) \\
&\quad \times \Bigg(\phi_1(X_s, A_s, Y_s; \hat{g}_{s-1}, \hat{f}_{s-1}) - \phi_1(X_s, A_s, Y_s; \pi_{s-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_s, A_s, Y_s; \hat{g}_{s-1}, \hat{f}_{s-1}) - \phi_1(X_s, A_s, Y_s; \pi_{s-1}, f^*)|\Omega_{s-1}\right] \\
&\quad + \phi_2(X_s; \hat{f}_{s-1}) - \phi_2(X_s; f^*) - \mathbb{E}\left[\phi_2(X_s; \hat{f}_{s-1}) - \phi_2(X_s; f^*)|\Omega_{s-1}\right]\Bigg)\Bigg] \\
&= \mathbb{E}\Bigg[U\mathbb{E}\Bigg[\Bigg(\phi_1(X_s, A_s, Y_s; \hat{g}_{s-1}, \hat{f}_{s-1}) - \phi_1(X_s, A_s, Y_s; \pi_{s-1}, f^*) \\
&\quad - \mathbb{E}\left[\phi_1(X_s, A_s, Y_s; \hat{g}_{s-1}, \hat{f}_{s-1}) - \phi_1(X_s, A_s, Y_s; \pi_{s-1}, f^*)|\Omega_{s-1}\right] \\
&\quad + \phi_2(X_s; \hat{f}_{s-1}) - \phi_2(X_s; f^*) - \mathbb{E}\left[\phi_2(X_s; \hat{f}_{s-1}) - \phi_2(X_s; f^*)|\Omega_{s-1}\right]\Bigg)|\Omega_{s-1}\Bigg]\Bigg] \\
&= 0,
\end{aligned}
$$

where

$$
\begin{aligned}
U = \Bigg(& \phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
& - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
& + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) \\
& - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right] \Bigg).
\end{aligned}
$$

Therefore, the variance is calculated as

$$
\begin{aligned}
& \mathrm{Var}\Bigg(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Bigg\{ \phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
& \quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
& \quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) \\
& \quad - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right] \Bigg\}\Bigg) \\
& = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Bigg[\Bigg(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
& \quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
& \quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) \\
& \quad - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right] \Bigg)^2\Bigg] \\
& = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Bigg[\mathbb{E}\Bigg[\Bigg(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
& \quad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right] \\
& \quad + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) \\
& \quad - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right] \Bigg)^2 |\Omega_{t-1}\Bigg]\Bigg] \\
& = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Var}\left(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) + \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right)\right] \\
& = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Var}\left(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right)\right] \\
& \quad + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Var}\left(\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right)\right] \\
& \quad + \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Cov}\left(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*), \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right)\right].
\end{aligned}
$$

Then, we want to show that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Var}\left(\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right)\right]\to 0, \tag{3}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Var}\left(\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right)\right]\to 0, \tag{4}$$

$$\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Cov}\left(\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*),\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right)\right]\to 0 \tag{5}$$

For showing (3)–(5), we consider showing

$$\mathrm{Var}\left(\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right)=\mathrm{o}_p(1), \tag{6}$$

$$\mathrm{Var}\left(\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right)=\mathrm{o}_p(1) \tag{7}$$

$$\mathrm{Cov}\left(\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*),\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right)=\mathrm{o}_p(1), \tag{8}$$

The first equation (6) is shown as

$$\begin{aligned}
&\mathrm{Var}\left(\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right)\\
&\le \mathbb{E}\left[\left\{\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}-\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\pi_{t-1}(a|X_t)}\right\}^2|\Omega_{t-1}\right]\\
&=\mathbb{E}\left[\left\{\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}-\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}\right.\right.\\
&\quad\left.\left.+\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}-\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\pi_{t-1}(a|X_t)}\right\}^2|\Omega_{t-1}\right]\\
&\le 2\mathbb{E}\left[\left\{\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}-\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}w\right\}^2 z|\Omega_{t-1}\right]\\
&\quad+2\mathbb{E}\left[\left\{\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}-\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\pi_{t-1}(a|X_t)}\right\}^2|\Omega_{t-1}\right]\\
&\le 2C\|f^*-\hat{f}_{t-1}\|_2^2+2\times 4C\|\hat{g}_{t-1}-\pi_{t-1}\|_2^2=\mathrm{o}_p(1),
\end{aligned}$$

where $C>0$ is a constant. Here, we have used a parallelogram law from the second line to the third line. We have used $|\hat{f}_{t-1}|<C_3$, and $0<\frac{\pi^{\mathrm{e}}}{\pi_{t-1}}<C_4$, convergence of π_{t-1} and convergence rate conditions from the third line to the fourth line. Then, from the L^r convergence theorem (Proposition 2) and the boundedness of the random variables, we can show that as $t\to\infty$,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Var}\left(\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right)\right]\\
&\le\mathbb{E}\left[\left|\mathrm{Var}\left(\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right)\right|\right]\to 0.
\end{aligned}$$

Therefore, for any $\epsilon > 0$, there exists a constant $C > 0$ such that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Var}\left(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right)\right] \leq C/T + \epsilon.$$

The second equation (7) is derived by Jensen's inequality, and we show (4) as well as (3) by using L^r convergence theorem.

Next, we show the third equation (8) as

$$\begin{aligned}
&\mathrm{Cov}\left(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*), \phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)|\Omega_{t-1}\right) \\
&\leq \Bigg|\mathbb{E}\Bigg[\Big(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) - \mathbb{E}\Big[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_t \\
&\qquad\qquad \times \Big(\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)\right]\Big)|\Omega_{t-1}\Bigg]\Bigg| \\
&\leq \mathbb{E}\Bigg[\Bigg|\Big(\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) - \mathbb{E}\Big[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_t \\
&\qquad\qquad \times \Big(\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*) - \mathbb{E}\left[\phi_2(X_t; \hat{f}_{t-1}) - \phi_2(X_t; f^*)\right]\Big)\Bigg||\Omega_{t-1}\Bigg] \\
&\leq C\mathbb{E}\Bigg[\Bigg|\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&\qquad\qquad - \mathbb{E}\left[\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*)|\Omega_{t-1}\right]\Bigg||\Omega_{t-1}\Bigg] \\
&= \mathrm{o}_p(1),
\end{aligned}$$

where $C > 0$ is a constant. From the second to third line, we used Jensen's inequality. From the fourth to fifth line, we used consistencies of $\hat{f}_{t-1}$ and $\hat{g}_{t-1}$, which imply that for all $X_t \in \mathcal{X}$,

$$\begin{aligned}
&\phi_1(X_t, A_t, Y_t; \hat{g}_{t-1}, \hat{f}_{t-1}) - \phi_1(X_t, A_t, Y_t; \pi_{t-1}, f^*) \\
&= \sum_{a=1}^{K}\left(\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t = a]\left(Y_t - \hat{f}_{t-1}(a, X_t)\right)}{\hat{g}_{t-1}(a|X_t)} - \frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t = a]\left(Y_t - f^*(a, X_t)\right)}{\pi_{t-1}(a|X_t)}\right) \\
&\leq \sum_{a=1}^{K}\left|\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t = a]\left(Y_t - \hat{f}_{t-1}(a, X_t)\right)}{\hat{g}_{t-1}(a|X_t)} - \frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t = a]\left(Y_t - f^*(a, X_t)\right)}{\pi_{t-1}(a|X_t)}\right| \\
&\leq C\sum_{a=1}^{K}\left|\pi_{t-1}(a|X_t)\left(Y_t - \hat{f}_{t-1}(a, X_t)\right) - \hat{g}_{t-1}(a|X_t)\left(Y_t - f^*(a, X_t)\right)\right| \\
&\leq C\sum_{a=1}^{K}\left|\pi_{t-1}(a|X_t) - \hat{g}_{t-1}(a|X_t)\right| \\
&\quad - C\sum_{a=1}^{K}\left|\pi_{t-1}(a|X_t)\hat{f}_{t-1}(a, X_t) - \hat{g}_{t-1}(a|X_t)\hat{f}_{t-1}(a, X_t) + \hat{g}_{t-1}(a|X_t)\hat{f}_{t-1}(a, X_t) - \hat{g}_{t-1}(a|X_t)f^*(a, X_t)\right| \\
&\leq C\sum_{a=1}^{K}\left|\pi_{t-1}(a|X_t) - \hat{g}_{t-1}(a|X_t)\right| - C\sum_{a=1}^{K}\left|\hat{f}_{t-1}(a, X_t) - f^*(a, X_t)\right| = \mathrm{o}_p(1),
\end{aligned}$$

where $C > 0$ is a constant.

Thus, from (3)–(5), the variance of the bias term converges to 0. Then, from Chebyshev's inequality,

$$
\begin{aligned}
&\mathbb{P}\Bigg(\Bigg|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Bigg\{\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)\\
&\quad-\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]\\
&\quad+\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)-\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right]\Bigg\}\Bigg|>\varepsilon\Bigg)\\
&\leq \mathrm{Var}\Bigg(\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\Bigg\{\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)\\
&\quad-\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})-\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]\\
&\quad+\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)-\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})-\phi_2(X_t;f^*)|\Omega_{t-1}\right]\Bigg\}\Bigg)/\varepsilon^2\\
&\to 0.
\end{aligned}
$$

Proof of (2)**.**

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\hat{g}_{t-1},\hat{f}_{t-1})|\Omega_{t-1}\right]+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_2(X_t;\hat{f}_{t-1})|\Omega_{t-1}\right]\\
&\quad-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_1(X_t,A_t,Y_t;\pi_{t-1},f^*)|\Omega_{t-1}\right]-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\phi_2(X_t;f^*)|\Omega_{t-1}\right]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}|\Omega_{t-1}\right]\\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\pi^{\mathrm{e}}(a,X_t)\hat{f}_{t-1}(a,X_t)|\Omega_{t-1}\right]\\
&\quad-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-f^*(a,X_t)\right)}{\pi_{t-1}(a|X_t,\Omega_{t-1})}|\Omega_{t-1}\right]\\
&\quad-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\pi^{\mathrm{e}}(a,X_t)f^*(a,X_t)|\Omega_{t-1}\right].
\end{aligned}
\tag{9}
$$

Because (9) is 0,

$$
\begin{aligned}
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}|\Omega_{t-1}\right] \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\pi^{\mathrm{e}}(a,X_t)\hat{f}_{t-1}(a,X_t)|\Omega_{t-1}\right] - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\pi^{\mathrm{e}}(a,X_t)f^*(a,X_t)|\Omega_{t-1}\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\frac{\pi^{\mathrm{e}}(a|X_t)\mathbb{1}[A_t=a]\left(Y_t-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}|\Omega_{t-1}\right] \\
&\quad - \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K}\pi^{\mathrm{e}}(a,X_t)\Big(f^*(a,X_t)-\hat{f}_{t-1}(a,X_t))\Big)|\Omega_{t-1}\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\mathbb{E}\Bigg[\mathbb{E}\Bigg[\frac{\pi^{\mathrm{e}}(a|X_t)\pi_{t-1}(a|X_t,\Omega_{t-1})\left(f^*(a,X_t)-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)} \\
&\qquad\qquad - \pi^{\mathrm{e}}(a,X_t)\Big(f^*(a,X_t)-\hat{f}_{t-1}(a,X_t))\Big)|X_t,\Omega_{t-1}\Bigg]|\Omega_{t-1}\Bigg] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{K}\left|\mathbb{E}\left[\frac{\pi^{\mathrm{e}}(a|X_t)\Big(\pi_{t-1}(a|X_t)-\hat{g}_{t-1}(a|X_t)\Big)\left(f^*(a,X_t)-\hat{f}_{t-1}(a,X_t)\right)}{\hat{g}_{t-1}(a|X_t)}|\Omega_{t-1}\right]\right|.
\end{aligned}
$$

By using Hölder's inequality $\|fg\|_1 \leq \|f\|_2\|g\|_2$, for a constant $C>0$, we have

$$
\begin{aligned}
&\leq \frac{C}{T}\sum_{t=1}^{T}\left\|\pi_{t-1}(a|X_t,\Omega_{t-1})-\hat{g}_{t-1}(a|X_t)\right\|_2\left\|f^*(a,X_t)-\hat{f}_{t-1}(a,X_t)\right\|_2 \\
&= \frac{C}{T}\sum_{t=1}^{T}\mathrm{o}_p(t^{-p})\mathrm{o}_p(t^{-q}) \\
&= \frac{C}{T}\sum_{t=1}^{T}\mathrm{o}_p(t^{-1/2}).
\end{aligned}
$$

□

C Adaptive-fitting and batched samples

Section C.1 supplements the description of adaptive-fitting. Next, in Section C.2, we introduce the AIPW estimator when the samples are given in batch form, and the true logging policy is given π_t. This is essentially the same as the generalized method of moment (GMM), which gives an asymptotically normal estimator for martigaale difference sequences (MDS), and we do not use adaptive-fitting. Based on this estimator, in Section C.3, we introduce the ADR estimator when the data is given in batch form, but the true logging policy π_t is not given.

C.1 Details of adaptive-fitting

As Section 5, let us define the parameter of interest θ_0 that satisfies $\mathbb{E}[\psi(W_t;\theta_0,\eta_0)]=0$, where $\{W_t\}_{t=1}^T$ are observations, η_0 is a nuisance parameter, and ψ is a score function. Let us define two estimators $\hat{\theta}_T$ and $\check{\theta}_T$ as $\frac{1}{T}\sum_{t=1}^T\psi(W_t,\hat{\theta}_T,\eta_{t-1})=0$ and $\frac{1}{T}\sum_{t=1}^T\psi(W_t,\check{\theta}_T,\eta_0)=0$. Suppose that $\check{\theta}_T$ is an asymptotically normal estimator of θ_0. Then, if $\check{\theta}_T-\hat{\theta}_T$ converges to 0 with convergence rate $\mathrm{o}_p(1/\sqrt{T})$, $\hat{\theta}_T$ is also an asymptotically normal estimator. In general, we cannot obtain such a fast convergence rate. However, by using double robustness, we can obtain the convergence rate. In ADR estimator, this conditions appears as $\|\hat{g}_{t-1}(a|X_t)-\pi_{t-1}(a|X_t,\Omega_{t-1})\|_2=\mathrm{o}_p(t^{-p})$, and

$\|\hat{f}_{t-1}(a, X_t) - f^*(a, X_t)\|_2 = \mathrm{o}_p(t^{-q})$, where $p, q > 0$ such that $p + q = 1/2$, and the expectation of the norm is taken over X_t. This allows us to obtain $\mathrm{o}_p(1/\sqrt{T})$ of the asympttoic bias. The image of the vanishing asymptotic bias is shown in Figure 3.

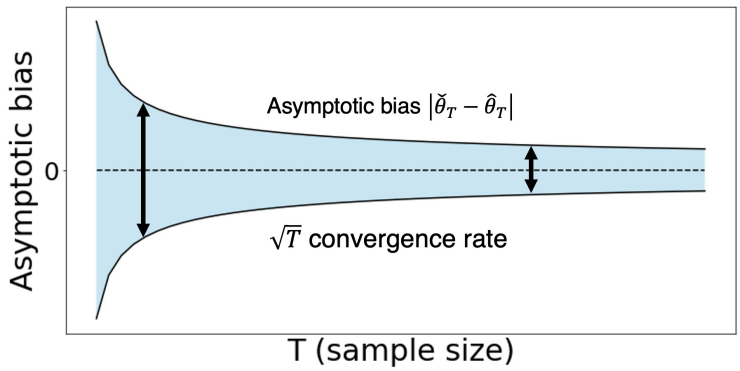

Figure 3: Convergence of asymptotic bias $|\check{\theta}_T - \hat{\theta}_T|$.

C.2 AIPW estimator with batched samples when the true logging policy is known

The proposed adaptive-fitting can be applied when batched samples are given. Let M denote the number of batch updates and $\tau \in I = \{1, 2, \ldots, M\}$ denotes the batch index. For $\tau \in I$, the probability is updated at a period t_τ, where $t_\tau - t_{\tau-1} = Tr_\tau$, using samples $\{(X_t, Y_t, A_t)\}_{t=t_{\tau-1}}^{t_\tau}$, where $r_1 + r_2 + \cdots + r_M = 1$ and $t_0 = 0$. Thus, in addition to the DGP, we assume that

$$\big\{(X_t, A_t, Y_t)\big\}_{t=t_{\tau-1}}^{t_\tau} \overset{i.i.d.}{\sim} p(x)\pi_\tau(a|x, \Omega_{t_{\tau-1}})p(y_a|x),$$

where $\pi_\tau(a|x, \Omega_{t_{\tau-1}})$ denotes the probability of choosing an action updated based on samples until the period $t_{\tau-1}$.

We consider asymptotic properties based on the assumption of $t_\tau - t_{\tau-1} \to \infty$ as $T \to \infty$ for fixed τ. This strategy is the same as Zhang et al. (2020). Because $\big\{(X_t, A_t, Y_t)\big\}_{t=t_{\tau-1}}^{t_\tau}$ is i.i.d., we can use the standard limit theorems for the partial sum of the samples to obtain an asymptotically normal estimator of $\theta_0 = R(\pi^{\mathrm{e}})$. However, we also have the motivation to use all samples together to increase the efficiency of the estimator. Therefore, by using the GMM, we propose an estimator of θ_0 considering the sample averages of each block as an empirical moment conditions. Although we cannot use standard CLT, we can can apply the CLT for the MDS by appropriately constructing an estimator. Note that the original GMM is proved when the MDS is applicable, which also includes the method of Zhang et al. (2020) as a special case.

For an index of batch $\tau \in I$, a function $f \in \mathcal{F}$ such that $f : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ and an evaluation policy $\pi^{\mathrm{e}} \in \Pi$, we define a function h_t as

$$h_t(x, k, y; \tau, R, f, \pi_\tau, \pi^{\mathrm{e}}) = \frac{1}{r_\tau}\xi_t(x, k, y; \tau, R, f, \pi_\tau, \pi^{\mathrm{e}})\mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big],$$

where $R \in \mathbb{R}$, $\xi_t(x, k, y; \tau, R, f, \pi_\tau, \pi^{\mathrm{e}}) := \phi_t(x, k, y; \tau, f, \pi_\tau, \pi^{\mathrm{e}}) - R$ and

$$\phi_t(x, k, y; \tau, f, \pi_\tau, \pi^{\mathrm{e}}) := \sum_{a=1}^{K} \pi^{\mathrm{e}}(a|x)\left\{\frac{\mathbb{1}[k = a]\big\{y - f(a, x)\big\}}{\pi_\tau(a|x, \Omega_{t_{\tau-1}})} + f(a, x)\right\}.$$

Let us note that the sequence $\left\{h_t(X_t, A_t, Y_t; \tau, R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}})\right\}_{t=1}^T$ is an MDS: for $h_t(X_t, A_t, Y_t; \tau, R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}})$, by using $\mathbb{E}[\mathbb{1}[A_t = a]|\mathbb{H}_{t-1}] = \pi_\tau(a|X_t, \Omega_{t_{\tau-1}})$, we have

$$\begin{aligned}
&\mathbb{E}\left[h_t(X_t, A_t, Y_t; \tau, R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}})|\Omega_{t-1}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{1}\left[t_{\tau-1} < t \leq t_\tau\right]}{r_\tau}\xi_t(x, k, y; \tau, R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}})|\Omega_{t-1}\right] \\
&= \frac{\mathbb{1}\left[t_{\tau-1} < t \leq t_\tau\right]}{r_\tau}\mathbb{E}\left[\xi_t(x, k, y; \tau, R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}})|\Omega_{t-1}\right] \\
&= \frac{\mathbb{1}\left[t_{\tau-1} < t \leq t_\tau\right]}{r_\tau} \times 0 = 0.
\end{aligned}$$

Let us also define

$$\boldsymbol{h}_t\left(X_t, A_t, Y_t; R, \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right) := \begin{pmatrix} h_t(X_t, A_t, Y_t; 1, R, \hat{f}_{t-1}, \pi_1, \pi^{\mathrm{e}}) \\ h_t(X_t, A_t, Y_t; 2, R, \hat{f}_{t-1}, \pi_2, \pi^{\mathrm{e}}) \\ \vdots \\ h_t(X_t, A_t, Y_t; M, R, \hat{f}_{t-1}, \pi_M, \pi^{\mathrm{e}}) \end{pmatrix}.$$

Then, the sequence $\left\{\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right\}_{t=1}^T$ is an MDS with respect to $\left\{\Omega_t\right\}_{t=0}^{T-1}$, i.e.,

$$\mathbb{E}\left[\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)|\Omega_{t-1}\right] = \mathbf{0}.$$

Using the sequence $\left\{\boldsymbol{h}_t\left(X_t, A_t, Y_t; R, \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right\}_{t=1}^T$, we define an estimator of $R(\pi^{\mathrm{e}})$ as

$$\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) = \arg\min_{R \in \mathbb{R}} \left(\hat{\boldsymbol{q}}_T(R)\right)^\top \hat{W}_T \left(\hat{\boldsymbol{q}}_T(R)\right), \tag{10}$$

where $\hat{\boldsymbol{q}}_T(R) = \frac{1}{T}\sum_{t=1}^T \boldsymbol{h}_t\left(X_t, A_t, Y_t; R, \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)$ and $\hat{W}_T$ is a data-dependent $(M \times M)$-dimensional positive semi-definite matrix. Let us note that the estimator defined in Eq. (10) is an application of GMM with the moment condition

$$\boldsymbol{q}(R(\pi^{\mathrm{e}})) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), \hat{f}_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right] = 0.$$

For the minimization problem defined in Eq. (10), we can analytically calculate the minimizer as

$$\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) = w_T^\top D_T(\pi^{\mathrm{e}}),$$

where $w_T = (w_{T,1} \ \cdots \ w_{T,M})^\top$ is an M-dimensional vector such that $\sum_{\tau=1}^M w_{T,\tau} = 1$, and

$$D_T(\pi^{\mathrm{e}}) = \begin{pmatrix} \frac{1}{t_1}\sum_{t=1}^{t_1} \phi_t(X_t, A_t, Y_t; 1, f_{t-1}, \pi_1, \pi^{\mathrm{e}}) \\ \frac{1}{t_2 - t_1}\sum_{t=t_1+1}^{t_2} \phi_t(X_t, A_t, Y_t; 2, f_{t-1}, \pi_2, \pi^{\mathrm{e}}) \\ \vdots \\ \frac{1}{T - t_{M-1}}\sum_{t=t_{M-1}+1}^{T} \phi_t(X_t, A_t, Y_t; M, f_{t-1}, \pi_M, \pi^{\mathrm{e}}) \end{pmatrix}.$$

Here, we show the asymptotic normality of the proposed estimator $\hat{\theta}_T^{\mathrm{OPE}}$.

Theorem 3 (Asymptotic distribution the proposed estimator)**.** *Suppose that*

(i) $w_T = (w_{T,1} \ \cdots \ w_{T,M})^\top \xrightarrow{\mathrm{p}} w = (w_1 \ \cdots \ w_M)^\top$*;*

(ii) $w_{T,\tau} > 0$ *and* $\sum_{\tau=1}^M w_{T,\tau} = 1$.

Then, under Assumptions 1, 3, 5, 7,

$$\sqrt{T}\big(\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\big) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \sigma^2\big),$$

where $\sigma^2 = \sum_{\tau=1}^{M} w_\tau \Psi(\pi^{\mathrm{e}}, \pi_\tau)$.

In the proposed method, we construct a moment condition using martingale difference sequences. On the other hand, for some readers, using martingale difference sequences may look unnecessary because samples are i.i.d in each block between $t_{\tau-1}$ and t_τ. Therefore, such readers also might feel that we can use $f_T(a, x)$, which is an estimator of $\mathbb{E}[Y(a) \mid x]$ using samples until T-th period, without going through constructing several estimators $\{f_{t_\tau}\}_{\tau=0}^{M-1}$. However, in that case, it is difficult to guarantee the asymptotic normality of the proposed estimator. For example, we can consider Cramér-Wold theorem to consider this problem. This motivations shares with the GMM and the method of Zhang et al. (2020).

The proof of Theorem 3 is shown as follows.

Choice of weight w_T. We discuss the choice of weight w_T. A naive choice is weighting the moment conditions equality; that is, $w_{\tau,T} = \frac{1}{M}$. Next, we consider an efficient weight w_T that minimizes the asymptotic variance of $\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}})$. In the GMM, the τ-th element of the efficient weight is given as $w_\tau^* = \frac{1}{\Psi(\pi^{\mathrm{e}}, \pi_\tau)} / \sum_{\tau'=1}^{M} \frac{1}{\Psi(\pi^{\mathrm{e}}, \pi_{\tau'})}$ (Hamilton, 1994). Here, we use the orthogonality among moment conditions; that is, zero covariance. In this case, the asymptotic variance becomes $1/\sum_{\tau'=1}^{M} \frac{1}{\sigma_{\tau'}^2}$. Therefore, for gaining efficiency, we use a weight $\hat{w}_{T,\tau} = \frac{1}{\hat{\Psi}_T(\pi^{\mathrm{e}}, \pi_\tau)} / \sum_{\tau'=1}^{M} \frac{1}{\hat{\Psi}_T(\pi^{\mathrm{e}}, \pi_{\tau'})}$, where $\hat{\Psi}_T(\pi^{\mathrm{e}}, \pi_\tau)$ is an estimator of $\Psi(\pi^{\mathrm{e}}, \pi_\tau)$.

Proof of Theorem 3. Instead of $\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) = w_T D_T(\pi^{\mathrm{e}})$, from the original formulation Eq. (10), we consider an estimator $\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) = \big(I^\top \hat{W}_T I\big)^{-1} I^\top \hat{W}_T D_T(\pi^{\mathrm{e}})$, where W_T is a $(M \times M)$-dimensional positive-definite matrix. Let us note that $w_T = \big(I^\top \hat{W}_T I\big)^{-1} I^\top \hat{W}_T$. We prove the following theorem, which is a generalized statement of Theorem 3.

Theorem 4 (Asymptotic distribution the AIPW estimator under batch update)**.** *Suppose that*

(i) $\hat{W}_T \xrightarrow{\mathrm{p}} W$*;*

(ii) W *is a positive definite;*

Then, under Assumptions 1, 3, 5, 7,

$$\sqrt{T}\big(\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\big) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \sigma^2\big),$$

where $\sigma^2 = \big(I^\top W I\big)^{-1} I^\top W \Sigma W^\top I \big(I^\top W I\big)^{-1}$ *and* Σ *is a* $(M \times M)$ *diagonal matrix such that the* $(\tau \times \tau)$*-element is* $\Psi(\pi^{\mathrm{e}}, \pi_\tau)$.

Proof. Let us define $\hat{\boldsymbol{q}}_T(R(\pi^{\mathrm{e}})) = (\hat{q}_{1,T}(R(\pi^{\mathrm{e}}))\ \hat{q}_{2,T}(R(\pi^{\mathrm{e}}))\ \cdots\ \hat{q}_{M,T}(R(\pi^{\mathrm{e}})))$, where

$$\hat{q}_{\tau,T}(R(\pi^{\mathrm{e}})) = \frac{1}{T}\frac{1}{r_1}\sum_{t=1}^{T}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\{Y_t - \hat{f}_{t-1}(a, X_t)\}}{\pi_1(a \mid x, \Omega_0)} + \pi^{\mathrm{e}}(a \mid X_t)\hat{f}_{t-1}(a, X_t)\right\} - R(\pi^{\mathrm{e}})\right) \times \mathbb{1}\big[t_0 = 0 < t \le t_1\big].$$

For $\sqrt{T}\big(\hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\big) = \big(I^\top \hat{W}_T I\big)^{-1} I^\top \hat{W}_T \sqrt{T}\hat{\boldsymbol{q}}_T(R(\pi^{\mathrm{e}}))$, we show that

$$\sqrt{T}\hat{\boldsymbol{q}}_T(R(\pi^{\mathrm{e}})) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \Sigma\big),$$

where Σ is a diagonal matrix such that the (τ, τ)-element is

$$\frac{1}{r_\tau}\mathbb{E}\left[\sum_{a=1}^{K}\frac{\big(\pi^{\mathrm{e}}(a \mid X)\big)^2 \mathrm{Var}(Y(a) \mid X)}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \left(\sum_{a=1}^{K}\pi^{\mathrm{e}}(a \mid X)\mathbb{E}\big[Y(a) \mid X\big] - R(\pi^{\mathrm{e}})\right)^2\right].$$

Then, from Slutsky Theorem, we can show that

$$\big(I^\top \hat{W}_T I\big)^{-1} I^\top \hat{W}_T \sqrt{T}\hat{\boldsymbol{q}}_T(R(\pi^{\mathrm{e}})) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \big(I^\top W I\big)^{-1} I^\top W \Sigma W^\top I \big(I^\top W I\big)^{-1}\big).$$

To show this result, we use the CLT for MDS by checking the following conditions:

(a) $(1/T)\sum_{t=1}^T \Sigma_t \to \Sigma$, where

$$\Sigma_t = \mathbb{E}\left[\left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)\left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)^\top\right];$$

(b) $\mathbb{E}\big[\tilde{h}_t(i, R(\pi^{\mathrm{e}}), f_{t-1}, \pi^{\mathrm{e}})\tilde{h}_t(j, R(\pi^{\mathrm{e}}), f_{t-1}, \pi^{\mathrm{e}})\tilde{h}_t(a, R(\pi^{\mathrm{e}}), f_{t-1}, \pi^{\mathrm{e}})\tilde{h}_t(l, R(\pi^{\mathrm{e}}), f_{t-1}, \pi^{\mathrm{e}})\big] < \infty$ for $i, j, k, l \in I$, where $\tilde{h}_t(a, R(\pi^{\mathrm{e}}), f_{t-1}, \pi^{\mathrm{e}}) = h_t^{\mathrm{OPE}}(X_t, A_t, Y_t; k, R(\pi^{\mathrm{e}}), f_k, \pi^{\mathrm{e}})$ for $k \in I$;

(c) $\frac{1}{T}\sum_{t=1}^T \left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)\left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)^\top \xrightarrow{p} \Sigma$,

Note that the GMM shows the asymptotic normality for more general cases.

Step 1: Condition (a)

From

$$\Sigma_t = \mathbb{E}\left[\left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)\left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)^\top\right],$$

the matrix $(1/T)\sum_{t=1}^T \Omega_t$ becomes a diagonal matrix such that the (τ, τ)-element is

$$\frac{1}{r_\tau^2 T}\sum_{t=1}^T \mathbb{E}\Bigg[\left(\sum_{a=1}^K \left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - R(\pi^{\mathrm{e}})\right)^2 \times \mathbb{1}\big[t_{\tau-1} < t \le t_\tau\big]\Bigg].$$

For $\tau \in I$ and t such that $t_{\tau-1} < t \le t_\tau$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{a=1}^K \left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - R(\pi^{\mathrm{e}})\right)^2\right] \\
&- \mathbb{E}\left[\left(\sum_{a=1}^K \left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - \mathbb{E}\big[Y_t(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - R(\pi^{\mathrm{e}})\right)^2\right] \\
&\le \mathbb{E}\Bigg[\Bigg|\left(\sum_{a=1}^K \left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - R(\pi^{\mathrm{e}})\right)^2 \\
&\quad - \left(\sum_{a=1}^K \left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - \mathbb{E}\big[Y_t(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - R(\pi^{\mathrm{e}})\right)^2\Bigg|\Bigg]
\end{aligned}$$

Because $\alpha^2 - \beta^2 = (\alpha + \beta)(\alpha - \beta)$, there exists a constant $\gamma_0 > 0$ such that

$$\begin{aligned}
&\le \gamma_0 \mathbb{E}\Bigg[\Bigg|\sum_{a=1}^K \left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right. \\
&\qquad \left. - \frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - \mathbb{E}\big[Y_t(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} - \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\}\Bigg|\Bigg]
\end{aligned}$$

Then, there exist constants $\gamma_1 > 0$ such that

$$\leq \gamma_1 \mathbb{E}\left[\sum_{a=1}^{K}\left|f_{t-1}(a, X_t) - \mathbb{E}\big[Y(a) \mid X_t\big]\right|\right].$$

Here, from the assumption that $f_{t-1}(a,x) - \mathbb{E}\big[Y(a) \mid X\big] \xrightarrow{\mathrm{p}} 0$ for $\tau = 2, 3, \ldots, M$, and $f_{t_{\tau-1}}(a,x)$ is bounded for $\tau \in I$, we can use L^r convergence theorem. First, to use L^r convergence theorem, we use boundedness of f_{t_m} to derive the uniform integrability of f_{t_m} for $m = 0, 1, \ldots, \tau - 1$. Then, from L^r convergence theorem, we have $\mathbb{E}\big[|f_{t_m}(a, X) - \mathbb{E}[Y(a) \mid X]|\big] \to 0$ as $t_m \to \infty$. Using this results, we can show that, as $t_{\tau-1} \to \infty$ (this also means $T \to \infty$),

$$\gamma_1 \sum_{a=1}^{K} \mathbb{E}\left[\left|f_{t-1}(a, X_t) - \mathbb{E}\big[Y(a) \mid X_t\big]\right|\right] \to 0.$$

Therefore, as $t_{\tau-1} \to \infty$ ($T \to \infty$),

$$\mathbb{E}\left[\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - R(\pi^{\mathrm{e}})\right)^2\right]$$
$$\to \mathbb{E}\left[\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - \mathbb{E}\big[Y_t(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - R(\pi^{\mathrm{e}})\right)^2\right].$$

Then, by using $\mathbb{1}[A_t = a]\mathbb{1}[A_t = l] = 0$, $\mathbb{E}\left[\frac{\mathbb{1}[A_t = a]Y_t^2}{\left(\pi_\tau(a|X_t, \Omega_{t_{\tau-1}})\right)^2}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[Y_t^2(a)|X_t\right]}{\pi_\tau(a|X_t, \Omega_{t_{\tau-1}})}\right]$, and $\frac{1}{r_\tau T}\sum_{t=1}^{T}\mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big] = 1$,

$$\mathbb{E}\left[\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - \mathbb{E}\big[Y_t(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - R(\pi^{\mathrm{e}})\right)^2\right]$$
$$= \mathbb{E}\left[\sum_{a=1}^{K}\left\{\frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \mathrm{Var}\big(Y_t(a) \mid X_t\big)}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \Big(\pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y_t(a) \mid X_t\big] - R(\pi^{\mathrm{e}})\Big)^2\right\}\right].$$

In addition, the variance does not depend on t. We represent the independence by omitting the subscript t, i.e.,

$$\mathbb{E}\left[\sum_{a=1}^{K}\left\{\frac{\big(\pi^{\mathrm{e}}(a \mid X_t)\big)^2 \mathrm{Var}\big(Y_t(a) \mid X_t\big)}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \Big(\pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y_t(a) \mid X_t\big] - R(\pi^{\mathrm{e}})\Big)^2\right\}\right]$$
$$= \mathbb{E}\left[\sum_{a=1}^{K}\left\{\frac{\big(\pi^{\mathrm{e}}(a \mid X)\big)^2 \mathrm{Var}\big(Y(a) \mid X\big)}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \Big(\pi^{\mathrm{e}}(a \mid X)\mathbb{E}\big[Y_t(a) \mid X\big] - R(\pi^{\mathrm{e}})\Big)^2\right\}\right].$$

Therefore, we have

$$\frac{1}{r_\tau^2 T}\sum_{t=1}^{T}\mathbb{E}\Bigg[\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t = a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X_t, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - R(\pi^{\mathrm{e}})\right)^2$$
$$\times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big]\Bigg]$$
$$\to \frac{1}{r_\tau}\mathbb{E}\left[\sum_{a=1}^{K}\left\{\frac{\big(\pi^{\mathrm{e}}(a \mid X)\big)^2 \mathrm{Var}\big(Y(a) \mid X\big)}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \Big(\pi^{\mathrm{e}}(a \mid X)\mathbb{E}\big[Y(a) \mid X\big] - R(\pi^{\mathrm{e}})\Big)^2\right\}\right].$$

Thus, the matrix $(1/T)\sum_{t=1}^{T}\Sigma_t$ converges to a diagonal matrix Σ as $T \to \infty$, where the (τ, τ)-element of Σ is

$$\frac{1}{r_\tau}\mathbb{E}\left[\sum_{a=1}^{K}\left\{\frac{\big(\pi^{\mathrm{e}}(a \mid X)\big)^2 \mathrm{Var}\big(Y(a) \mid X\big)}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \Big(\pi^{\mathrm{e}}(a \mid X)\mathbb{E}\big[Y(a) \mid X\big] - R(\pi^{\mathrm{e}})\Big)^2\right\}\right].$$

Step 2: Condition (b)

Because we assume that all variables are bounded, this condition holds.

Step 3: Condition (c)

Here, we check that $(1/T)\sum_{t=1}^{T}\left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)\left(\boldsymbol{h}_t\left(X_t, A_t, Y_t; R(\pi^{\mathrm{e}}), f_{t-1}, \pi_\tau, \pi^{\mathrm{e}}\right)\right)^\top \xrightarrow{p} \Sigma$. The (τ,τ)-element of the matrix is

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{r_\tau^2}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t=a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - \theta\right)^2 \\
&\qquad\qquad \times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big] \\
&= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{r_\tau^2}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t=a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - \theta\right)^2 \\
&\qquad\qquad \times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big] \\
&- \frac{1}{T}\sum_{t=1}^{T}\frac{1}{r_\tau^2}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t=a]\big\{Y_t - \mathbb{E}\big[Y(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - \theta\right)^2 \\
&\qquad\qquad \times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big] \\
&+ \frac{1}{T}\sum_{t=1}^{T}\frac{1}{r_\tau^2}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t=a]\big\{Y_t - \mathbb{E}\big[Y(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - \theta\right)^2 \\
&\qquad\qquad \times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big].
\end{aligned}
$$

The part

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{r_\tau^2}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t=a]\big\{Y_t - f_{t-1}(a, X_t)\big\}}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t) f_{t-1}(a, X_t)\right\} - \theta\right)^2 \\
&\qquad\qquad \times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big] \\
&- \frac{1}{T}\sum_{t=1}^{T}\frac{1}{r_\tau^2}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t=a]\big\{Y_t - \mathbb{E}\big[Y(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - \theta\right)^2 \\
&\qquad\qquad \times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big]
\end{aligned}
$$

converges in probability to 0 because $f_{t-1}(a, X_t) \xrightarrow{\mathrm{p}} \mathbb{E}\big[Y(a) \mid X_t\big]$. The term

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{r_\tau^2}\left(\sum_{a=1}^{K}\left\{\frac{\pi^{\mathrm{e}}(a \mid X_t)\mathbb{1}[A_t=a]\big\{Y_t - \mathbb{E}\big[Y(a) \mid X_t\big]\big\}}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \pi^{\mathrm{e}}(a \mid X_t)\mathbb{E}\big[Y(a) \mid X_t\big]\right\} - R(\pi^{\mathrm{e}})\right)^2 \\
&\qquad\qquad \times \mathbb{1}\big[t_{\tau-1} < t \leq t_\tau\big].
\end{aligned}
$$

converges in probability to

$$
\frac{1}{r_\tau}\mathbb{E}\left[\sum_{a=1}^{K}\left\{\frac{\big(\pi^{\mathrm{e}}(a \mid X)\big)^2 \mathrm{Var}\big(Y(a) \mid X\big)}{\pi_\tau(a \mid X, \Omega_{t_{\tau-1}})} + \Big(\pi^{\mathrm{e}}(a \mid X)\mathbb{E}\big[Y(a) \mid X\big] - R(\pi^{\mathrm{e}})\Big)^2\right\}\right].
$$

from the weak law of large numbers for i.i.d. samples as $t_{\tau-1} - t_\tau \to \infty$ because the samples are i.i.d. between $t_{\tau-1}$ and t_τ. □

C.3 OPE estimator when the true logging policy is unknown

Then, we consider estimating the policy value without using the true logging policy π_t. We use adaptive-fitting for obtaining an asymptotically normal estimator. As the ADR estimator under

Assumption 2, we estimate π_t and f^* only using Ω_{t-1} and denote them as g_{t-1} and $\hat{f}_{t-1}$, respectively. Then, we define an estimator of $R(\pi^{\mathrm{e}})$ as

$$\tilde{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) = \arg\min_{R \in \mathbb{R}} \left(\tilde{\boldsymbol{q}}_T(R)\right)^\top \hat{W}_T \left(\tilde{\boldsymbol{q}}_T(R)\right),$$

where $\tilde{\boldsymbol{q}}_T(R) = \frac{1}{T}\sum_{t=1}^T \boldsymbol{h}_t\left(X_t, A_t, Y_t; R, \hat{f}_{t-1}, \hat{g}_{t-1}\pi^{\mathrm{e}}\right)$ and $\hat{W}_T$ is a data-dependent $(M \times M)$-dimensional positive semi-definite matrix. This estimator is the ADR estimator under batch update. As well as the proof of Theorem 1, it is hold that

$$\left|\tilde{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) - \hat{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}})\right| = \mathrm{o}_p(1/\sqrt{T})$$

if $\|\hat{g}_{t-1}(a|X_t) - \pi_{t-1}(a|X_t, \Omega_{t-1})\|_2 = \mathrm{o}_p(t^{-p})$, and $\|\hat{f}_{t-1}(a, X_t) - f^*(a, X_t)\|_2 = \mathrm{o}_p(t^{-q})$, where $p, q > 0$ such that $p + q = 1/2$. Therefore, we can obtain the following theorem.

Theorem 5 (Asymptotic distribution the ADR estimator under batch update)**.** *Suppose that*

(i) $\hat{W}_T \xrightarrow{\mathrm{p}} W$*;*

(ii) W *is a positive definite.*

Then, under Assumptions 1, 3, 4–6,

$$\sqrt{T}\left(\tilde{R}_T^{\mathrm{batch}}(\pi^{\mathrm{e}}) - R(\pi^{\mathrm{e}})\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2\right),$$

where $\sigma^2 = \left(I^\top W I\right)^{-1} I^\top W \Sigma W^\top I \left(I^\top W I\right)^{-1}$ *and* Σ *is a* $(M \times M)$ *diagonal matrix such that the* $(\tau \times \tau)$*-element is* $\Psi(\pi^{\mathrm{e}}, \pi_\tau)$*.*

D Details of experiments

The description of the dataset is shown in Table 3. We use LinUCB and LinTS policies. We add uniform sampling to make overlap between policies. We can relax this requirement by considering different DGPs, such as batched sampling. However, for brevity, we adopt this setting. Additional results are shown as follows.

For numerical experiments in Section 6.3, we show the result with sample sizes $T = 100, 1,000$ in Table 4. In addition, we show the error distribution with sample size $T = 100$ in Figure 4; we show the error distribution with sample size $T = 500$ in Figure 5; we show the error distribution with sample size $T = 500$ in Figure 6; we show the error distribution with sample size $T = 1,000$ in Figure 7.

For experiments with dependent samples in Section 7, we show the additional results with different settings in Tables 5–10. In Table 5, we show the results using the benchmark datasets with 800 samples generated from the LinUCB algorithm. In Table 6, we show the results using the benchmark datasets with $1,000$ samples generated from the LinUCB algorithm. In Table 7, we show the results using the the benchmark datasets with $1,200$ samples generated from the LinUCB algorithm. In Table 8, we show the results using the the benchmark datasets with 800 samples generated from the LinTS algorithm. In Table 9, we show the results using the the benchmark datasets with $1,000$ samples generated from the LinTS algorithm. In Table 10, we show the results using the the benchmark datasets with $1,200$ samples generated from the LinTS algorithm.

Next, we compare the estimators using the benchmark datasets generated from the logistic regression as well as the evaluation weight; that is, the samples are i.i.d. For $\alpha \in \{0.7, 0.4, 0.1\}$ and the sample sizes 800, $1,000$, and $1,200$, we calculate the RMSEs and the SDs over 10 trials. We show additional results with different settings in Tables 11–13. In Table 11, we show the results using the benchmark datasets with 800 samples. In Table 12, we show the results using the benchmark datasets with $1,000$ samples. In Table 13, we show the results using the benchmark datasets with $1,200$ samples. In these experiments, AIPW estimators show better performances than the ADR estimator. We conjecture that this is because the logging policy is not unstable unlike the case with dependent samples; that is, the paradox is specific to the case where samples are dependent. Note this case (i.i.d. samples) is not in the scope of the proposed method. In such cases, it is common to use other methods, such as

the cross-fitting proposed by Chernozhukov et al. (2018), instead of the AIPW and ADR estimators discussed in this paper. We show this result to clarify the cause of the paradox. Since the ADR and AIPW estimators have the same asymptotic variance, and since the AIPW estimator uses more information, it is a natural result that the AIPW estimator performs better. However, when the samples are dependent, the ADR estimator paradoxically outperforms the AIPW estimator because of the unstable behavior of the logging policy π_t, as pointed out by Hadad et al. (2021).

As a surprising discovery, our proposed ADR estimator shows better results than the AIPW estimator, although the AIPW estimator uses more information (true logging policy π_t) than the ADR estimator and their asymptotic properties are the same. As discussed above, we consider that this result is due to the logging policy's unstable behavior. Even when knowing the true logging policy π_t, we can stabilize estimation by reestimating the logging policy from (A_t, X_t). This paradox is similar to the well-known property that the IPW estimator using an estimated propensity score shows a smaller asymptotic variance than the IPW using the true propensity score (Hirano et al., 2003; Henmi & Eguchi, 2004; Henmi et al., 2007). However, we consider that they are different phenomena. In previous studies, the paradox is mainly explained by differences in the asymptotic variance between the IPW estimators with the true and an estimated propensity score. On the other hand, for our case, the AIPW and ADR estimators have the same asymptotic variance, unlike IPW-type estimators; therefore, we cannot elucidate the paradox by traditional explanation.

Table 3: Specification of datasets

Dataset	the number of samples	Dimension	the number of classes
`mnist`	60,000	780	10
`satimage`	4,435	35	6
`sensorless`	58,509	48	11
`connect-4`	67,557	126	3

Table 4: The results of Section 6.1 with sample sizes $T = 100, 1,000$. We show the RMSE, SD, and coverage ratio of the confidence interval (CR). We highlight in red bold two estimators with the lowest RMSE. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

LinUCB policy																		
T	ADR †$*$			IPW †			AIPW †			AW-AIPW †			DM $*$			EIPW $*$		
	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR
100	**0.106**	0.011	0.86	0.158	0.036	0.8	0.139	0.024	0.87	0.161	0.035	0.34	0.116	0.016	0.29	0.144	0.031	0.73
1,000	**0.030**	0.001	0.93	0.083	0.035	0.88	0.061	0.011	0.87	0.189	0.016	0.0	0.033	0.001	0.22	0.133	0.012	0.07

LinTS policy																		
T	ADR †$*$			IPW †			AIPW †			AW-AIPW †			DM $*$			EIPW $*$		
	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR	RMSE	SD	CR
100	**0.087**	0.008	0.92	0.181	0.062	0.88	0.150	0.032	0.92	0.124	0.02	0.42	0.100	0.011	0.25	0.140	0.026	0.82
1,000	**0.026**	0.001	0.98	0.065	0.008	0.89	0.053	0.005	0.94	0.155	0.011	0.0	0.033	0.001	0.15	0.114	0.009	0.09

Table 5: The results of benchmark datasets with the LinUCB policy and $T = 800$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.047**	0.003	0.099	0.011	0.132	0.032	0.279	0.025	0.130	0.019
0.4	**0.034**	0.002	0.067	0.005	0.115	0.014	0.281	0.015	0.095	0.009
0.1	**0.053**	0.006	0.079	0.006	0.107	0.013	0.287	0.028	0.119	0.018

satimage	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.015**	0.000	0.097	0.013	0.074	0.007	0.035	0.001	0.065	0.003
0.4	**0.020**	0.000	0.078	0.006	0.088	0.010	0.041	0.001	0.085	0.014
0.1	**0.026**	0.001	0.080	0.006	0.092	0.011	0.059	0.003	0.097	0.012

sensorless	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.050**	0.002	0.253	0.049	0.101	0.016	0.130	0.008	0.121	0.016
0.4	**0.065**	0.006	0.270	0.039	0.104	0.019	0.150	0.013	0.096	0.012
0.1	**0.042**	0.001	0.250	0.043	0.075	0.006	0.127	0.008	0.102	0.010

connect-4	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.024**	0.001	0.121	0.019	0.042	0.002	0.045	0.001	0.140	0.007
0.4	**0.023**	0.001	0.125	0.014	0.033	0.001	0.044	0.002	0.090	0.007
0.1	**0.028**	0.001	0.107	0.012	0.042	0.002	0.068	0.004	0.031	0.001

Table 6: The results of benchmark datasets with the LinUCB policy and $T = 1,000$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.046**	0.002	0.100	0.011	0.162	0.027	0.232	0.013	0.148	0.014
0.4	**0.028**	0.001	0.068	0.005	0.112	0.009	0.249	0.010	0.080	0.004
0.1	0.086	0.006	**0.078**	0.006	0.085	0.008	0.299	0.024	0.091	0.008

satimage	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.013**	0.000	0.098	0.011	0.060	0.004	0.037	0.001	0.056	0.002
0.4	**0.022**	0.000	0.078	0.008	0.019	0.000	0.043	0.001	0.060	0.002
0.1	**0.029**	0.001	0.078	0.005	0.061	0.008	0.041	0.002	0.041	0.002

sensorless	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.043**	0.001	0.270	0.049	0.096	0.014	0.109	0.014	0.068	0.007
0.4	**0.053**	0.004	0.287	0.035	0.070	0.006	0.112	0.008	0.104	0.013
0.1	**0.048**	0.002	0.260	0.031	0.072	0.006	0.154	0.015	0.088	0.012

connect-4	ADR †$*$		IPW †		AIPW †		DM $*$		EIPW $*$	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.030**	0.001	0.135	0.024	0.042	0.002	0.032	0.001	0.173	0.006
0.4	**0.015**	0.000	0.135	0.017	0.037	0.001	0.038	0.001	0.085	0.002
0.1	**0.012**	0.000	0.117	0.013	0.030	0.001	0.051	0.002	0.023	0.001

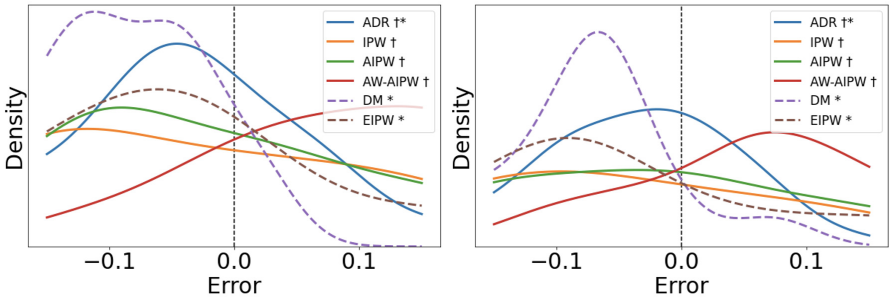

Figure 4: This figure illustrates the error distributions of estimators for OPVE from dependent samples generated with the LinUcB(left) and LinTS(right) with the sample size 100. We smoothed the error distributions using kernel density estimation. Estimators with asymptotic normality are marked with †, and estimators that do not require a true logging policy are marked with $*$.

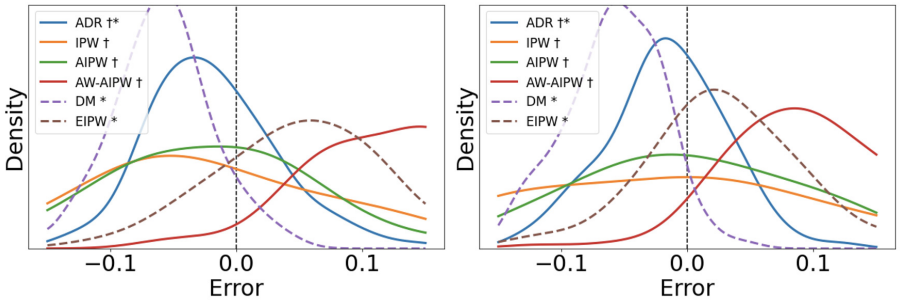

Figure 5: This figure illustrates the error distributions of estimators for OPVE from dependent samples generated with the LinUcB(left) and LinTS(right) with the sample size 250. We smoothed the error distributions using kernel density estimation. Estimators with asymptotic normality are marked with †, and estimators that do not require a true logging policy are marked with $*$.

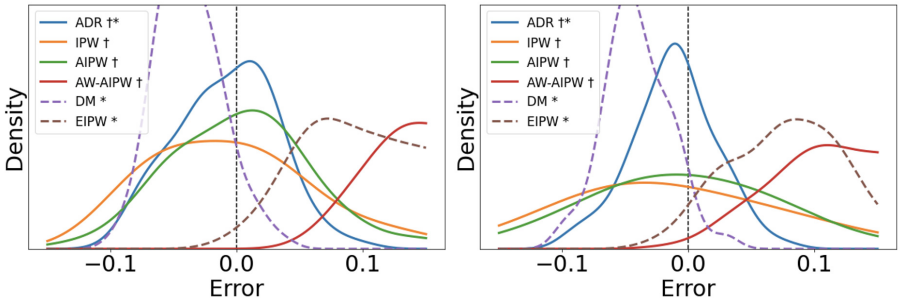

Figure 6: This figure illustrates the error distributions of estimators for OPVE from dependent samples generated with the LinUcB(left) and LinTS(right) with the sample size 500. We smoothed the error distributions using kernel density estimation. Estimators with asymptotic normality are marked with †, and estimators that do not require a true logging policy are marked with $*$.

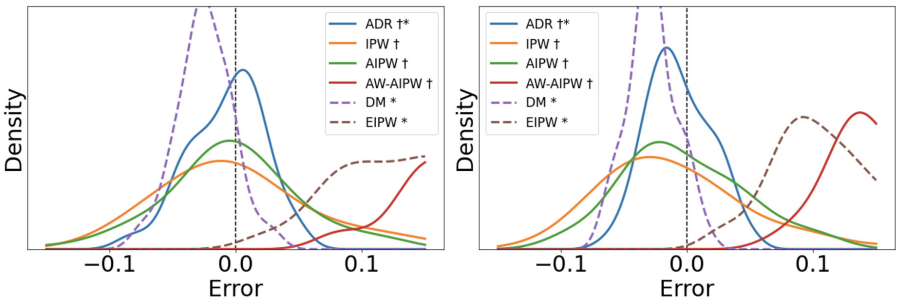

Figure 7: This figure illustrates the error distributions of estimators for OPVE from dependent samples generated with the LinUcB(left) and LinTS(right) with the sample size $1,000$. We smoothed the error distributions using kernel density estimation. Estimators with asymptotic normality are marked with †, and estimators that do not require a true logging policy are marked with $*$.

Table 7: The results of benchmark datasets with the LinUCB policy and $T = 1,200$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.063**	0.003	0.095	0.010	0.114	0.018	0.204	0.007	0.218	0.021
0.4	**0.031**	0.002	0.061	0.004	0.100	0.011	0.227	0.010	0.136	0.015
0.1	0.073	0.005	0.076	0.006	0.093	0.013	0.245	0.013	**0.059**	0.004

satimage	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.015**	0.000	0.091	0.010	0.039	0.002	0.035	0.001	0.039	0.001
0.4	**0.016**	0.000	0.075	0.006	0.041	0.002	0.043	0.001	0.048	0.002
0.1	**0.019**	0.000	0.075	0.005	0.033	0.001	0.045	0.001	0.073	0.005

sensorless	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.049**	0.003	0.272	0.059	0.102	0.018	0.092	0.008	0.078	0.010
0.4	**0.047**	0.004	0.286	0.037	0.071	0.007	0.096	0.006	0.082	0.009
0.1	**0.056**	0.004	0.268	0.047	0.061	0.006	0.116	0.008	0.067	0.005

connect-4	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.033**	0.001	0.138	0.027	0.046	0.002	0.037	0.002	0.184	0.013
0.4	**0.020**	0.000	0.130	0.014	0.052	0.003	0.038	0.001	0.114	0.004
0.1	**0.020**	0.000	0.118	0.013	0.037	0.002	0.034	0.001	0.048	0.003

Table 8: The results of benchmark datasets with the LinTS policy and $T = 800$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.064**	0.006	0.098	0.011	0.150	0.019	0.289	0.026	0.172	0.019
0.4	0.084	0.005	**0.060**	0.003	0.089	0.007	0.311	0.027	0.116	0.016
0.1	0.117	0.011	0.078	0.006	**0.061**	0.004	0.331	0.030	0.065	0.004

satimage	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.032**	0.001	0.102	0.011	0.094	0.007	0.043	0.002	0.095	0.016
0.4	**0.033**	0.001	0.081	0.009	0.073	0.005	0.048	0.002	0.098	0.015
0.1	**0.028**	0.001	0.075	0.004	0.074	0.005	0.043	0.001	0.048	0.003

sensorless	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.055**	0.003	0.252	0.053	0.094	0.011	0.158	0.015	0.112	0.014
0.4	0.084	0.006	0.267	0.033	0.078	0.007	0.177	0.024	**0.059**	0.004
0.1	0.086	0.009	0.247	0.041	0.104	0.017	0.164	0.009	**0.079**	0.006

connect-4	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.021**	0.000	0.129	0.019	0.057	0.003	0.041	0.001	0.115	0.004
0.4	**0.025**	0.001	0.130	0.018	0.055	0.003	0.050	0.003	0.064	0.004
0.1	**0.018**	0.000	0.107	0.011	0.054	0.004	0.054	0.002	0.029	0.001

Table 9: The results of benchmark datasets with the LinTS policy and $T = 1,000$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.053**	0.006	0.097	0.010	0.140	0.023	0.248	0.008	0.184	0.020
0.4	0.077	0.008	**0.070**	0.005	0.098	0.011	0.276	0.028	0.099	0.011
0.1	0.099	0.011	**0.076**	0.006	0.091	0.008	0.296	0.026	0.088	0.007

satimage	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.016**	0.000	0.096	0.011	0.050	0.004	0.044	0.001	0.056	0.003
0.4	**0.021**	0.001	0.077	0.007	0.041	0.002	0.048	0.001	0.063	0.009
0.1	**0.026**	0.001	0.079	0.005	0.087	0.011	0.048	0.002	0.091	0.020

sensorless	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.040**	0.003	0.273	0.057	0.083	0.009	0.114	0.008	0.087	0.008
0.4	**0.034**	0.001	0.291	0.035	0.091	0.010	0.099	0.006	0.070	0.007
0.1	**0.060**	0.002	0.268	0.042	0.075	0.008	0.146	0.011	0.069	0.006

connect-4	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.014**	0.000	0.132	0.020	0.059	0.003	0.045	0.001	0.110	0.006
0.4	**0.019**	0.000	0.136	0.017	0.028	0.001	0.046	0.002	0.066	0.004
0.1	**0.016**	0.000	0.114	0.012	0.023	0.001	0.050	0.002	0.019	0.000

Table 10: The results of benchmark datasets with the LinTS policy and $T = 1,200$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.038**	0.003	0.095	0.010	0.095	0.010	0.234	0.013	0.203	0.028
0.4	**0.048**	0.003	0.066	0.004	0.099	0.012	0.245	0.013	0.117	0.011
0.1	0.058	0.002	0.073	0.006	**0.052**	0.003	0.257	0.021	0.069	0.006

satimage	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.012**	0.000	0.096	0.010	0.058	0.004	0.037	0.001	0.059	0.003
0.4	**0.020**	0.001	0.079	0.007	0.036	0.001	0.043	0.002	0.064	0.010
0.1	**0.027**	0.001	0.079	0.005	0.045	0.002	0.050	0.001	0.050	0.004

sensorless	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.027**	0.001	0.257	0.043	0.063	0.004	0.102	0.005	0.053	0.004
0.4	**0.038**	0.002	0.277	0.036	0.059	0.003	0.106	0.006	0.087	0.010
0.1	0.046	0.002	0.252	0.035	**0.034**	0.002	0.138	0.008	0.045	0.003

connect-4	ADR †∗		IPW †		AIPW †		DM ∗		EIPW ∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	**0.014**	0.000	0.376	0.244	0.044	0.002	0.029	0.001	0.107	0.007
0.4	**0.021**	0.000	0.395	0.283	0.025	0.001	0.033	0.001	0.067	0.004
0.1	**0.215**	0.138	0.385	0.273	**0.215**	0.138	0.217	0.137	**0.215**	0.138

Table 11: The results of benchmark datasets with the i.i.d samples and $T = 800$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.195	0.090	**0.078**	0.061	**0.078**	0.061	**0.078**	0.061	2.085	9.013
0.4	0.192	0.088	**0.078**	0.061	0.079	0.061	0.080	0.061	1.187	2.913
0.1	0.105	0.061	**0.080**	0.061	0.083	0.061	0.108	0.061	0.298	0.189

satimage	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.025	0.001	**0.003**	0.00	0.008	0.000	0.006	0.00	0.076	0.014
0.4	0.019	0.001	**0.006**	0.00	0.017	0.001	0.012	0.00	0.074	0.012
0.1	**0.079**	0.060	**0.079**	0.06	0.083	0.060	0.081	0.06	0.098	0.061

sensorless	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.032	0.003	**0.004**	0.000	0.005	0.000	**0.004**	0.000	0.193	0.099
0.4	0.031	0.003	**0.007**	0.000	0.008	0.000	0.015	0.001	0.154	0.064
0.1	0.052	0.021	**0.046**	0.021	0.048	0.021	0.057	0.021	0.121	0.050

connect-4	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.080	0.047	**0.069**	0.047	**0.069**	0.047	**0.069**	0.047	0.234	0.086
0.4	0.074	0.047	**0.069**	0.047	0.071	0.047	0.071	0.047	0.141	0.052
0.1	**0.070**	0.047	**0.070**	0.047	0.073	0.047	0.074	0.047	0.076	0.047

Table 12: The results of benchmark datasets with the i.i.d samples and $T = 1,000$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.367	0.062	0.006	0.000	0.011	0.000	**0.005**	0.000	4.834	2.294
0.4	0.373	0.037	**0.016**	0.000	0.031	0.001	0.038	0.001	2.672	0.905
0.1	0.165	0.012	**0.029**	0.001	0.049	0.002	0.155	0.005	0.671	0.098

satimage	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.043	0.002	0.006	0.000	0.013	0.000	**0.010**	0.000	0.136	0.010
0.4	0.038	0.001	**0.013**	0.000	0.035	0.001	0.018	0.000	0.148	0.012
0.1	0.028	0.001	**0.026**	0.001	0.034	0.001	0.044	0.002	0.113	0.019

sensorless	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.073	0.005	**0.005**	0.000	0.012	0.000	0.009	0.000	0.454	0.071
0.4	0.077	0.005	**0.009**	0.000	0.031	0.001	0.023	0.001	0.390	0.045
0.1	0.066	0.003	**0.042**	0.003	0.047	0.004	0.083	0.007	0.257	0.066

connect-4	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.057	0.002	**0.009**	0.000	0.014	0.000	0.013	0.000	0.396	0.020
0.4	0.040	0.001	**0.008**	0.000	0.021	0.001	0.019	0.000	0.218	0.010
0.1	0.027	0.001	**0.024**	0.001	0.047	0.003	0.042	0.002	0.056	0.003

Table 13: The results of benchmark datasets with the i.i.d samples and $T = 1,200$. We highlight in red bold the estimator with the lowest RMSE and highlight in under line the estimator with the lowest RMSE among estimators that do not use the true logging policy. Estimators with asymptotic normality are marked with †, and estimators that do not require the true logging policy are marked with $*$.

mnist	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.310	0.036	0.006	0.000	0.013	0.000	**0.005**	0.000	4.709	1.037
0.4	0.346	0.033	0.012	0.000	**0.020**	0.001	0.040	0.001	2.735	0.461
0.1	0.143	0.013	**0.036**	0.001	0.076	0.007	0.161	0.006	0.608	0.137

satimage	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.038	0.001	**0.005**	0.0	0.015	0.000	0.012	0.000	0.114	0.004
0.4	0.029	0.001	**0.011**	0.0	0.019	0.000	0.016	0.000	0.159	0.011
0.1	0.021	0.001	**0.014**	0.0	0.048	0.002	0.044	0.001	0.092	0.007

sensorless	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.096	0.007	**0.007**	0.000	0.012	0.000	0.008	0.000	0.481	0.083
0.4	0.097	0.007	**0.014**	0.000	0.027	0.001	0.025	0.000	0.407	0.058
0.1	0.054	0.002	**0.023**	0.001	0.038	0.002	0.065	0.002	0.208	0.025

connect-4	ADR †∗		IPW †		AIPW †		DM †∗		EIPW †∗	
α	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD	RMSE	SD
0.7	0.060	0.002	**0.005**	0.0	0.012	0.000	0.009	0.000	0.378	0.013
0.4	0.043	0.002	**0.010**	0.0	0.014	0.000	0.023	0.000	0.211	0.012
0.1	**0.017**	0.000	**0.017**	0.0	0.037	0.002	0.035	0.001	0.057	0.003