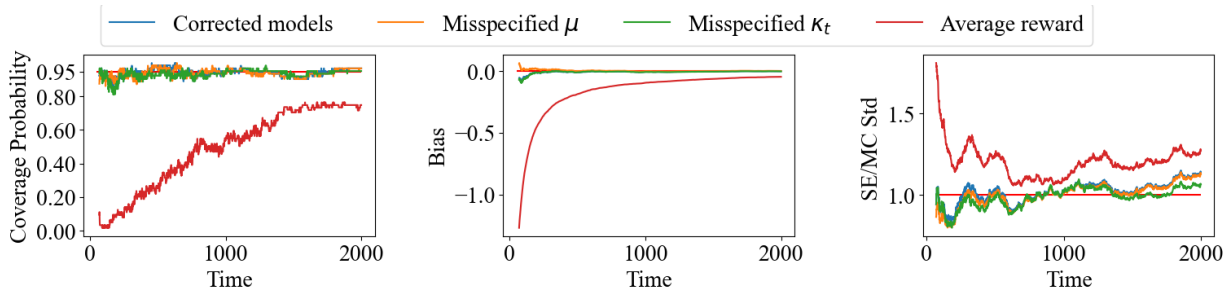
Supplementary to 'Doubly Robust Interval Estimation for Optimal Policy Evaluation in Online Learning'

Ye Shen [*], Hengrui Cai [†], Rui Song [‡]

This supplementary article provides sensitivity analyses and all the technical proofs for the established theorems for policy evaluation in online learning under the contextual bandits. Note that the theoretical results in Section 7 can be proven in a similar manner by arguments in Section B. Thus we omit the details here.

A Sensitivity Test for the Choice of p_t

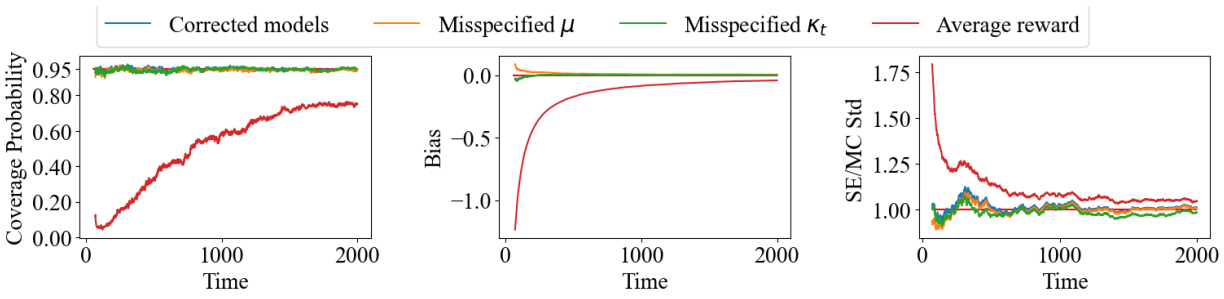
We conduct a sensitivity test for the choice of p_t in this section. We run all the simulations with $p_t = 0.01, 0.05, 0.1$ and find that Algorithm 1 is not sensitive to the choice of p_t.



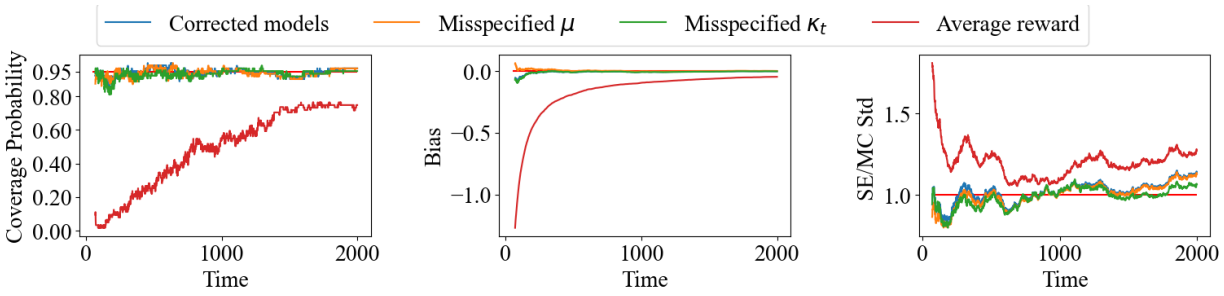
(a) $p_t = 0.01$

Figure A.1: Results by DREAM under UCB with different model specifications in comparison to the averaged reward. Left panel: the coverage probabilities of the 95% two-sided Wald-type CI, with the red line representing the nominal level at 95%. Middle panel: the bias between the estimated value and the true value. Right panel: the ratio between the standard error and the Monte Carlo standard deviation, with the red line representing the nominal level at 1.

[*]Equal Contribution, Department of Statistics, North Carolina State University
[†]Equal Contribution, Department of Statistics, University of California Irvine
[‡]Amazon.com, Inc

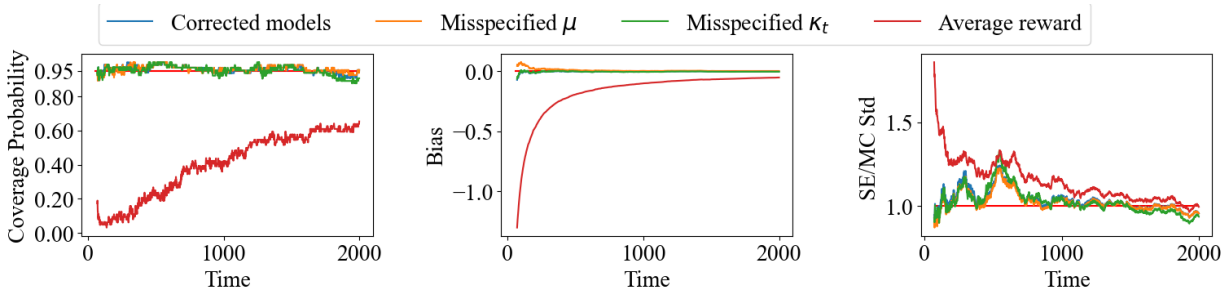


(a) $p_t = 0.05$

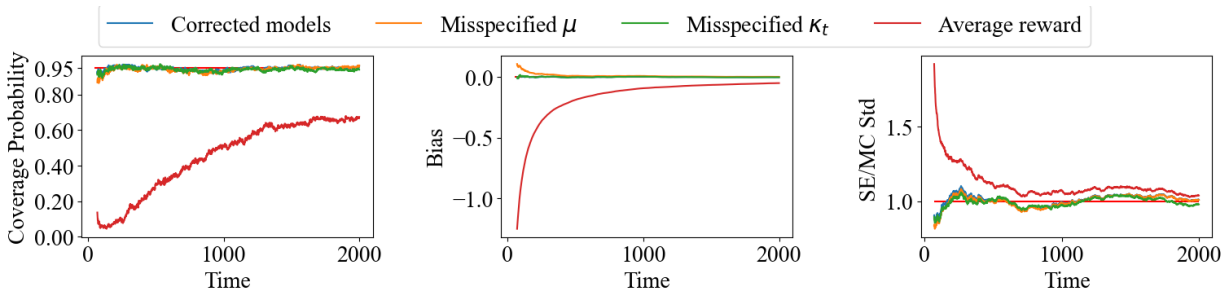


(b) $p_t = 0.1$

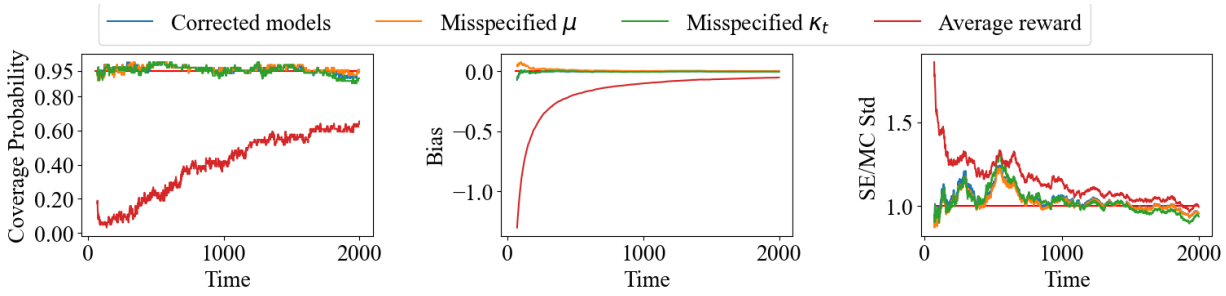
Figure A.1: Results by DREAM under UCB with different model specifications in comparison to the averaged reward. Left panel: the coverage probabilities of the 95% two-sided Wald-type CI, with the red line representing the nominal level at 95%. Middle panel: the bias between the estimated value and the true value. Right panel: the ratio between the standard error and the Monte Carlo standard deviation, with the red line representing the nominal level at 1.



(c) $p_t = 0.01$

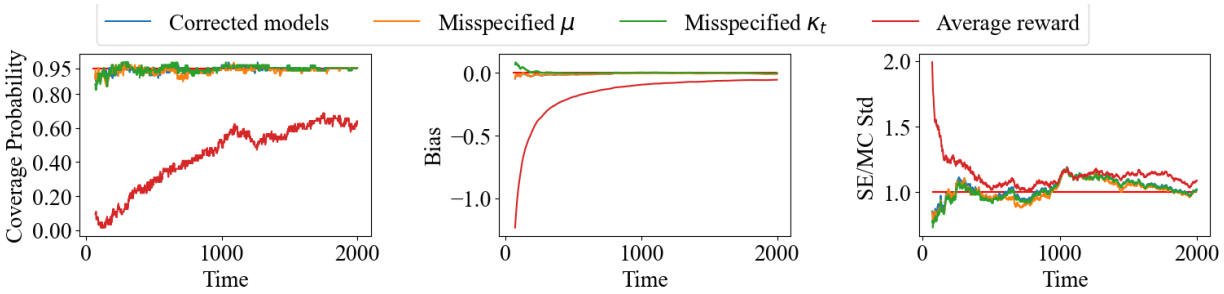


(d) $p_t = 0.05$

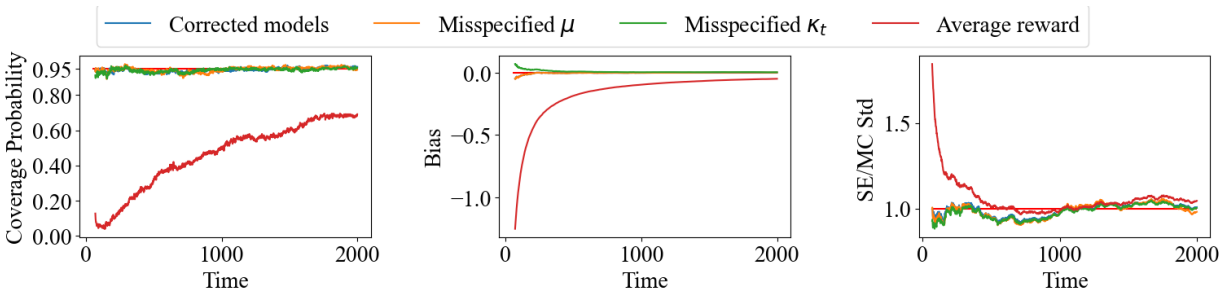


(e) $p_t = 0.1$

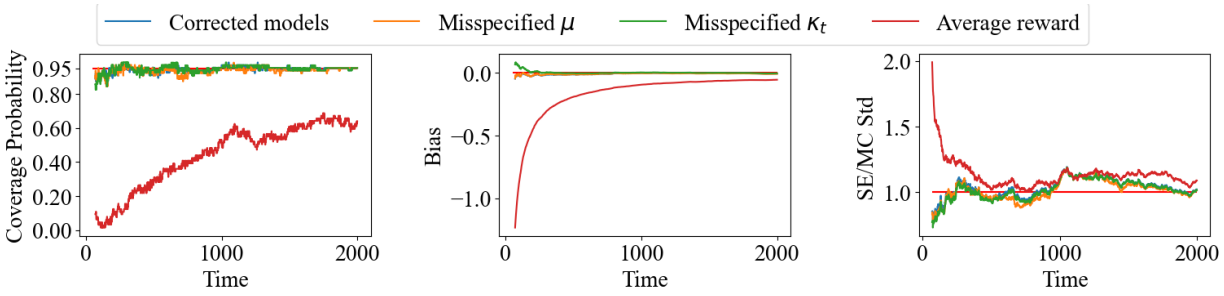
Figure A.2: Results by DREAM under TS with different model specifications in comparison to the averaged reward. Left panel: the coverage probabilities of the 95% two-sided Wald-type CI, with the red line representing the nominal level at 95%. Middle panel: the bias between the estimated value and the true value. Right panel: the ratio between the standard error and the Monte Carlo standard deviation, with the red line representing the nominal level at 1.



(a) $p_t = 0.01$



(b) $p_t = 0.05$



(c) $p_t = 0.1$

Figure A.3: Results by DREAM under EG with different model specifications in comparison to the averaged reward. Left panel: the coverage probabilities of the 95% two-sided Wald-type CI, with the red line representing the nominal level at 95%. Middle panel: the bias between the estimated value and the true value. Right panel: the ratio between the standard error and the Monte Carlo standard deviation, with the red line representing the nominal level at 1.

B Technical Proofs for Main Results

This section provides all the technical proofs for the established theorems for policy evaluation in online learning under the contextual bandits.

B.1 Proof of Lemma 4.1

The proof of Lemma 4.1 consists of three main steps. To be specific, we first reconstruct the target difference $\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)$ and decompose it into two parts. Then, we establish the bound for each part and derive its lower bound $\Pr(\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\|_1 \leq h)$.

Step 1: Recall Equation (1) in the main paper with $\boldsymbol{D}_{t-1}(a)$ being a $\boldsymbol{N}_{t-1}(a) \times d$ design matrix at time $t-1$ with $\boldsymbol{N}_{t-1}(a)$ as the number of pulls for action a, we have

$$\widehat{\boldsymbol{\beta}}_t(a) = \left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{1}{t} \omega \boldsymbol{I}_d \right\}^{-1} \left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i r_i \right\}.$$

We are interested in the quantity

$$\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a) = \left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{1}{t} \omega \boldsymbol{I}_d \right\}^{-1} \left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i r_i \right\} - \boldsymbol{\beta}(a). \quad \text{(B.1)}$$

Note that $\boldsymbol{\beta}(a)$ can be written as

$$\left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{1}{t} \omega \boldsymbol{I}_d \right\}^{-1} \left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{1}{t} \omega \boldsymbol{I}_d \right\} \boldsymbol{\beta}(a),$$

and since $r_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}(a) + e_i$, we can write (B.1) as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a) = \quad & \underbrace{\left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{1}{t} \omega \boldsymbol{I}_d \right\}^{-1} \left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i e_i \right\}}_{\eta_3} \quad \text{(B.2)} \\ & - \underbrace{\left\{ \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(a_i = a) \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{1}{t} \omega \boldsymbol{I}_d \right\}^{-1} \frac{\omega}{t} \boldsymbol{\beta}(a)}_{\eta_4}. \end{aligned}$$

Our goal is to find a lower bound of $\Pr(\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\|_1 \leq h)$ for any $h > 0$. Notice that by the triangle inequality we have $\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\|_1 \leq \|\eta_3\|_1 + \|\eta_4\|_1$, thus we can find the lower bound using the inequality as

$$\Pr\left(\left\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\|_1 \leq h\right) \geq \Pr\left(\|\eta_3\|_1 + \|\eta_4\|_1 \leq h\right). \tag{B.3}$$

Step 2: We focus on bounding η_4 first. By the relationship between eigenvalues and the L_2 norm of symmetric matrix, we have $\|\boldsymbol{M}^{-1}\|_2 = \lambda_{\max}(\boldsymbol{M}^{-1}) = \{\lambda_{\min}(\boldsymbol{M})\}^{-1}$ for any invertible matrix $\boldsymbol{M}$. Thus we can obtain that

$$\begin{aligned}
\left\|\left\{\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega\boldsymbol{I}_d\right\}^{-1}\right\|_2 &= \left\{\lambda_{\min}\left(\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega\boldsymbol{I}_d\right)\right\}^{-1} \\
&= \frac{1}{\lambda_{\min}\left(\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right) + \frac{1}{t}\omega} \\
&\leq \frac{1}{p_t\lambda_{\min}(\boldsymbol{\Sigma}) + \frac{1}{t}\omega} \leq \frac{1}{p_t\lambda + \frac{1}{t}\omega},
\end{aligned}$$

which leads to

$$\|\eta_4\|_2 \leq \left\|\left\{\frac{1}{t}\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega\boldsymbol{I}_d\right\}^{-1}\right\|_2 \left\|\frac{\omega}{t}\boldsymbol{\beta}(a)\right\|_2 \leq \frac{\omega}{tp_t\lambda + \omega}\|\boldsymbol{\beta}(a)\|_2. \tag{B.4}$$

By Cauchy-Schwartz inequality, we further have bound of the L_1 norm of η_4 as

$$\|\eta_4\|_1 \leq \sqrt{d}\|\eta_4\|_2 \leq \frac{\omega\sqrt{d}}{tp_t\lambda + \omega}\|\boldsymbol{\beta}(a)\|_2 \leq \frac{\sqrt{d}}{1 + tp_t\lambda/\omega}\|\boldsymbol{\beta}(a)\|_2 \leq \sqrt{d}\|\boldsymbol{\beta}(a)\|_2. \tag{B.5}$$

Step 3: Lastly, using the results in (B.5), we have

$$\Pr\left(\|\eta_3\|_1 + \|\eta_4\|_1 \leq h\right) \geq \Pr\left(\|\eta_3\|_1 \leq h - \sqrt{d}\|\boldsymbol{\beta}(a)\|_2\right). \tag{B.6}$$

By the definition of η_3 and Lemma 2 in Chen et al. (2020), for any constant $c > 0$, we have

$$\Pr\left(\|\eta_3\|_1 > c\right) \leq 2d\exp\left\{-\frac{t\left(\frac{p_t\lambda}{2}\right)^2 c^2}{2d^2\sigma^2 L_{\boldsymbol{x}}^2}\right\} = 2d\exp\left\{-\frac{tp_t^2\lambda^2c^2}{8d^2\sigma^2L_{\boldsymbol{x}}^2}\right\}.$$

Therefore, from (B.3) and (B.6), taking $c = h - \sqrt{d}\left\|\boldsymbol{\beta}(a)\right\|_2$, we have that under event E_t,

$$\Pr\left(\left\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\|_1 > h\right) \leq 2d\exp\left\{-\frac{tp_t^2\lambda^2\left(h - \sqrt{d}\left\|\boldsymbol{\beta}(a)\right\|_2\right)^2}{8d^2\sigma^2 L_{\boldsymbol{x}}^2}\right\}. \tag{B.7}$$

Based on the above results, it is immediate that the online ridge estimator $\widehat{\boldsymbol{\beta}}_t(a)$ is consistent to $\boldsymbol{\beta}(a)$ if $tp_t^2 \to \infty$ as $t \to \infty$. The proof is hence completed.

B.2 Proof of Corollary 1

Since $\widehat{\mu}_t(\boldsymbol{x}_t, a) - \mu(\boldsymbol{x}_t, a) = \boldsymbol{x}_t^\top\left(\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right)$, by Holder's inequality, we have

$$|\widehat{\mu}_t(\boldsymbol{x}_t, a) - \mu(\boldsymbol{x}_t, a)| \leq \|\boldsymbol{x}_t\|_\infty \left\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\|_1 \leq L_{\boldsymbol{x}} \left\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\|_1,$$

which follows

$$\Pr\left\{|\widehat{\mu}_t(\boldsymbol{x}_t, a) - \mu(\boldsymbol{x}_t, a)| > \xi\right\} \leq \Pr\left\{L_{\boldsymbol{x}}\left\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\|_1 > \xi\right\} = \Pr\left\{\left\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\|_1 > \xi/L_{\boldsymbol{x}}\right\}.$$

By Lemma 4.1, we further have

$$\begin{aligned}
\Pr\left\{|\widehat{\mu}_t(\boldsymbol{x}_t, a) - \mu(\boldsymbol{x}_t, a)| > \xi\right\} &\leq \Pr\left\{\left\|\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\|_1 > \frac{\xi}{L_{\boldsymbol{x}}}\right\} \\
&\leq 2d\exp\left\{-\frac{tp_t^2\lambda^2\left(\frac{\xi}{L_{\boldsymbol{x}}} - \sqrt{d}\left\|\boldsymbol{\beta}(a)\right\|_2\right)^2}{8d^2\sigma^2 L_{\boldsymbol{x}}^2}\right\} \\
&= 2d\exp\left\{-\frac{tp_t^2\lambda^2\left(\xi - \sqrt{d}L_{\boldsymbol{x}}\left\|\boldsymbol{\beta}(a)\right\|_2\right)^2}{8d^2\sigma^2 L_{\boldsymbol{x}}^4}\right\}.
\end{aligned}$$

Note that by the Triangle Inequality,

$$\begin{aligned}
|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}| &= |\{\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 1)\} - \{\widehat{\mu}_t(\boldsymbol{x}_t, 0) - \mu(\boldsymbol{x}_t, 0)\}| \\
&\leq |\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 1)| + |\widehat{\mu}_t(\boldsymbol{x}_t, 0) - \mu(\boldsymbol{x}_t, 0)|,
\end{aligned}$$

thus for $|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}|$, we have

$$
\begin{aligned}
&\Pr\left\{|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}| > \xi\right\} \\
&\leq \Pr\left\{|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 1)| + |\widehat{\mu}_t(\boldsymbol{x}_t, 0) - \mu(\boldsymbol{x}_t, 0)| > \xi\right\} \\
&\leq \Pr\left\{|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \mu(\boldsymbol{x}_t, 1)| > \xi/2\right\} + \Pr\left\{|\widehat{\mu}_t(\boldsymbol{x}_t, 0) - \mu(\boldsymbol{x}_t, 0)| > \xi/2\right\} \\
&\leq 2d \exp\left\{-\frac{t p_t^2 \lambda^2 \left(\xi/2 - \sqrt{d} L_{\boldsymbol{x}} \left\|\boldsymbol{\beta}(1)\right\|_2\right)^2}{8 d^2 \sigma^2 L_{\boldsymbol{x}}^4}\right\} + 2d \exp\left\{-\frac{t p_t^2 \lambda^2 \left(\xi/2 - \sqrt{d} L_{\boldsymbol{x}} \left\|\boldsymbol{\beta}(0)\right\|_2\right)^2}{8 d^2 \sigma^2 L_{\boldsymbol{x}}^4}\right\} \\
&\leq 4d \exp\left\{-(t-1) p_{t-1}^2 c_\xi\right\},
\end{aligned}
$$

with

$$
c_\xi = \frac{\lambda^2 \left[\min\left\{\left(\xi/2 - \sqrt{d} L_{\boldsymbol{x}} \left\|\boldsymbol{\beta}(1)\right\|_2\right)^2, \left(\xi/2 - \sqrt{d} L_{\boldsymbol{x}} \left\|\boldsymbol{\beta}(0)\right\|_2\right)^2\right\}\right]}{8 d^2 \sigma^2 L_{\boldsymbol{x}}^4},
$$

consistent with time t.

B.3 Proof of Theorem 1

The proof of Theorem 1 consists of two main parts to show the probability of exploration under UCB and TS, respectively, by noting the probability of exploration under EG is given by its definition.

B.3.1 Proof for UCB

We first show the probability of exploration under UCB. This proof consists of three main steps stated as following:

1. We first rewrite the target probability by its definition and express it as

$$
\Pr\left\{|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)| < c_t \left|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)\right|\right\}.
$$

2. Then, we establish the bound for the variance estimation such that

$$
c_t \left|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)\right| \leq \frac{2 c_t L_{\boldsymbol{x}}}{\sqrt{(t-1) p_{t-1} \lambda}}.
$$

3. Lastly, we bound $\Pr\left\{|\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)|<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\right\}$ using the result in Corollary 1.

Step 1: We rewrite the target probability by definition and decompose it into two parts.

Let $\Delta_{\boldsymbol{x}_t}\equiv\mu(\boldsymbol{x}_t,1)-\mu(\boldsymbol{x}_t,0)$. Based on the definition of the probability of exploration and the form of the estimated optimal policy $\widehat{\pi}_t(\boldsymbol{x}_t)$, we have

$$\begin{aligned}&\kappa_t(\boldsymbol{x}_t)=\Pr\{a_t\neq\widehat{\pi}_t(\boldsymbol{x}_t)\}=\mathbb{E}[\mathbb{I}\{a_t\neq\widehat{\pi}_t(\boldsymbol{x}_t)\}] \qquad \text{(B.8)}\\ =\;& \underbrace{\mathbb{E}[\mathbb{I}(a_t=0)|\widehat{\pi}_t(\boldsymbol{x}_t)=1]\Pr\{\widehat{\pi}_t(\boldsymbol{x}_t)=1\}}_{\eta_0}+\underbrace{\mathbb{E}[\mathbb{I}(a_t=1)|\widehat{\pi}_t(\boldsymbol{x}_t)=0]\Pr\{\widehat{\pi}_t(\boldsymbol{x}_t)=0\}}_{\eta_1},\end{aligned}$$

where the expectation is taken with respect to the history $\mathcal{H}_{t-1}$ before time point t.

Next, we rewrite η_0 and η_1 using the estimated mean and variance components $\widehat{\mu}_{t-1}(\boldsymbol{x}_t,a)$ and $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,a)$, where $a=0,1$. We focus on η_0 first.

Given $\widehat{\pi}_t(\boldsymbol{x}_t)=1$, i.e., $\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)>0$, based on the definition of the taken action in Lin-UCB that

$$a_t=\mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)+c_t\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,1)>\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)+c_t\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,0)\right\},$$

the probability of choosing action 0 rather than action 1 is

$$\begin{aligned}&\mathbb{E}[\mathbb{I}(a_t=0)|\widehat{\pi}_t(\boldsymbol{x}_t)=1]\\ =&\Pr\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)+c_t\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,1)<\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)+c_t\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,0)|\widehat{\pi}_t(\boldsymbol{x}_t)=1\right\}\\ =&\Pr\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)<c_t\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,0)-c_t\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,1)|\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)>0\right\}\\ =&\Pr\left[0<\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)<c_t\left\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,0)-\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,1)\right\}\right]/\Pr\{\widehat{\pi}_t(\boldsymbol{x}_t)=1\},\end{aligned}$$

where the second equality is to rearrange the estimated mean and variance components, and the last equality comes from the definition of the conditional probability. Combining this with (B.8), we have

$$\eta_0=\Pr\left[0<\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)<c_t\left\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,0)-\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,1)\right\}\right].$$

Similarly, we have

$$\eta_1=\Pr\left[c_t\left\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,0)-\widehat{\sigma}_{t-1}(\boldsymbol{x}_t,1)\right\}<\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)<0\right].$$

Thus combined with Equation (B.8), we have

$$\begin{aligned}\kappa_t(\boldsymbol{x}_t) = \eta_0 + \eta_1 =& \Pr\left[0 < \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0) < c_t\left\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)\right\}\right] \\ &+ \Pr\left[c_t\left\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)\right\} < \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0) < 0\right] \\ &= \Pr\left\{|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)| < c_t\,|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)|\right\}.\end{aligned} \tag{B.9}$$

The rest of the proof is aims to bound the probability

$$\Pr\left\{|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)| < c_t\,|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)|\right\}.$$

Step 2: Secondly, we bound the variance $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)$ and $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)$.

We consider the quantity $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) = \sqrt{\{\boldsymbol{x}_t^\top\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\}^{-1}\boldsymbol{x}_t\}}$ first. Let $\mathbf{v}$ be any $d \times 1$ vector, then the sample variance under action 0 is given by

$$\begin{aligned}\boldsymbol{x}_t^\top\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\}^{-1}\boldsymbol{x}_t &= \|\boldsymbol{x}_t\|_2^2(\frac{\boldsymbol{x}_t}{\|\boldsymbol{x}_t\|_2})^\top\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\}^{-1}(\frac{\boldsymbol{x}_t}{\|\boldsymbol{x}_t\|_2}) \\ &\le \|\boldsymbol{x}_t\|_2^2 \max_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\}^{-1}\mathbf{v} \\ &\le \|\boldsymbol{x}_t\|_2^2\lambda_{\max}\{(\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d)^{-1}\},\end{aligned} \tag{B.10}$$

where the first inequality is to replace $(\boldsymbol{x}_t/\|\boldsymbol{x}_t\|_2)^\top$ with any normalized vector, and the second inequality is due to the definition. According to (B.10), combined with Assumption 4.1, we can further bound $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)$ by

$$\|\boldsymbol{x}_t\|_2\sqrt{\lambda_{\max}\{(\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d)^{-1}\}} \le \frac{L_{\boldsymbol{x}}}{\sqrt{\lambda_{\min}\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\}}}. \tag{B.11}$$

It is immediate from (B.10) and (B.11) that

$$0 < \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) \le \frac{L_{\boldsymbol{x}}}{\sqrt{\lambda_{\min}\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\}}}. \tag{B.12}$$

Note that

$$\lambda_{\min}\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\} = \lambda_{\min}\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0)\} + \omega,$$

combined with the fact that $\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) = \sum_{i=1}^{t-1}(1-a_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top$, then $\lambda_{\min}\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\}$ can be further expressed as

$$\begin{aligned}\lambda_{\min}\{\boldsymbol{D}_{t-1}(0)^\top\boldsymbol{D}_{t-1}(0) + \omega\boldsymbol{I}_d\} &= (t-1)\lambda_{\min}\left\{\frac{1}{t-1}\sum_{i=1}^{t-1}(1-a_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right\} + \omega \\ &> (t-1)p_{t-1}\lambda_{\min}(\boldsymbol{\Sigma}) + \omega > (t-1)p_{t-1}\lambda + \omega,\end{aligned}$$

where the first inequality is owing to Assumption 4.2 , and the second inequality is owing to Assumption 4.1. This together with (B.12) gives the lower and upper bounds of $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)$ as

$$0 < \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) \leq \frac{L_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda + \omega}} < \frac{L_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}. \tag{B.13}$$

Similarly we have

$$0 < \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1) \leq \frac{L_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}, \tag{B.14}$$

which follows that

$$c_t \left|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)\right| \leq c_t \left(\left|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)\right| + \left|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)\right|\right) \leq \frac{2c_t L_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}.$$

Combining (B.9) and the above equation, we get the conclusion that

$$\begin{aligned} \kappa_t(\boldsymbol{x}_t) &\leq \Pr\left\{\left|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\right| < c_t \left|\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) - \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)\right|\right\} \\ &\leq \Pr\left\{\left|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\right| < \frac{2c_t L_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\right\}. \end{aligned} \tag{B.15}$$

Step 3: Lastly, we aim to bound $\Pr\left\{\left|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\right| < \frac{2c_t L_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\right\}$ using the result in Corollary 1.

For any $\xi > 0$, define $E := \{|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}| \leq \xi\}$, which satisfies $\Pr\{E\} \geq 1 - 4d\exp\{-tp_t^2 c_\xi\}$ by Corollary 1. Then on the Event E, we have

$$\begin{aligned} \left|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\right| &= \left|\Delta_{\boldsymbol{x}_t} + \{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}\}\right| \\ &\geq \left|\Delta_{\boldsymbol{x}_t}\right| - \left|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}\right| \geq \left|\Delta_{\boldsymbol{x}_t}\right| - \xi. \end{aligned}$$

Thus for the probability $\Pr\left\{|\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)|<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\right\}$, we have

$$\begin{aligned}
\kappa_t(\boldsymbol{x}_t) &\leq \Pr\left\{|\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)|<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\right\} \\
&\leq \Pr\left\{|\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)|<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\;\middle|\; E\right\}+\Pr\{E^c\} \\
&\leq \Pr\left\{|\Delta_{\boldsymbol{x}_t}|-\xi<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\right\}+4d\exp\left\{-(t-1)p_{t-1}^2c_\xi\right\} \\
&= \Pr\left\{|\Delta_{\boldsymbol{x}_t}|<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}+\xi\right\}+4d\exp\left\{-(t-1)p_{t-1}^2c_\xi\right\}.
\end{aligned} \tag{B.16}$$

Sine $tp_t\to\infty$ as $t\to\infty$, for any constant $\delta>\xi$, there exist large enough t satisfying $\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}\leq\delta-\xi$. Then by Assumption 4.3, there exists some constant γ such that

$$\Pr\left\{|\Delta_{\boldsymbol{x}_t}|<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}+\xi\right\}=\mathcal{O}\left\{\left(\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}+\xi\right)^\gamma\right\},$$

i.e., there exists some constant C such that

$$\Pr\left\{|\Delta_{\boldsymbol{x}_t}|<\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}+\xi\right\}=C\left(\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}+\xi\right)^\gamma.$$

Therefore, combined with the Equation (B.16), we have

$$\kappa_t(\boldsymbol{x}_t)\leq C\left(\frac{2c_tL_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}+\xi\right)^\gamma+4d\exp\left\{-(t-1)p_{t-1}^2c_\xi\right\}.$$

The proof is hence completed.

B.3.2 Proof for TS

We next show the probability of exploration under TS consisting of three main steps:

1. We firstly define an event $E := \{|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}| \leq \xi\}$ for any $0 < \xi < |\Delta_{\boldsymbol{x}_t}|/2$, where the estimated difference between mean functions is close to the true difference. And we have $\Pr\{E\} \geq 1 - 4d\exp\{-tp_t^2 c_\xi\}$ by Corollary 1.

2. Next, we bound the probability of exploration on the event E.

3. Lastly, we combine the results in the previous two steps to get the unconditioned probability of exploration .

Step 1: For any $0 < \xi < |\Delta_{\boldsymbol{x}_t}|/2$, define $E := \{|\widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) - \Delta_{\boldsymbol{x}_t}| \leq \xi\}$, which satisfies $\Pr\{E\} \geq 1 - 4d\exp\left\{-(t-1)p_{t-1}^2 c_\xi\right\}$ by Corollary 1. Then for the probability $\Pr\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\}$, we have

$$\Pr\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\} \leq \Pr\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)|E\} + \Pr\{E^c\} \leq \Pr\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)|E\} + 4d\exp\left\{-(t-1)p_{t-1}^2 c_\xi\right\}. \tag{B.17}$$

Without loss of generality, we assume $\Delta_{\boldsymbol{x}_t} > 0$, then $E := \{0 < \Delta_{\boldsymbol{x}_t} - \xi \leq \widehat{\mu}_t(\boldsymbol{x}_t, 1) - \widehat{\mu}_t(\boldsymbol{x}_t, 0) \leq \Delta_{\boldsymbol{x}_t} + \xi\}$, which implies $\widehat{\pi}_t(\boldsymbol{x}_t) = 1$.

Using the law of iterated expectations, based on the definition of the probability of exploration and the form of the estimated optimal policy $\widehat{\pi}_t(\boldsymbol{x}_t)$, on the event E, we have

$$\Pr\{a_t \neq 1|E\} = \mathbb{E}[\mathbb{I}\{a_t \neq 1\}|E] = \mathbb{E}\left(\mathbb{E}[\mathbb{I}\{a_t = 0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1), \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)]|E\right). \tag{B.18}$$

Step 2: Next, we focus on deriving the bound of $\mathbb{E}[\mathbb{I}\{a_t = 0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1), \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)]$ on the Event E.

Recalling the bandit mechanism of TS, we have $a_t = \mathbb{I}\left\{\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(1) > \boldsymbol{x}_t^\top \boldsymbol{\beta}_t(0)\right\}$, where $\boldsymbol{\beta}_t(a)$ is drawn from the posterior distribution of $\boldsymbol{\beta}(a)$ given by

$$\mathcal{N}_d[\widehat{\boldsymbol{\beta}}_{t-1}(a), \rho^2\{\boldsymbol{D}_{t-1}(a)^\top \boldsymbol{D}_{t-1}(a) + \omega \boldsymbol{I}_d\}^{-1}].$$

From the posterior distributions and the definitions of $\widehat{\mu}_{t-1}(\boldsymbol{x}_t, a)$ and $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, a)$, we have

$$\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(a) \sim \mathcal{N}[\boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_{t-1}(a), \rho^2 \boldsymbol{x}_t^\top \{\boldsymbol{D}_{t-1}(a)^\top \boldsymbol{D}_{t-1}(a) + \omega \boldsymbol{I}_d\}^{-1} \boldsymbol{x}_t],$$

that is,

$$\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(a) \sim \mathcal{N}[\widehat{\mu}_{t-1}(\boldsymbol{x}_t, a), \rho^2 \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, a)^2].$$

Notice that $\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(1)$ and $\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(0)$ are drawn independently, thus,

$$\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(1) - \boldsymbol{x}_t^\top \boldsymbol{\beta}_t(0) \sim \mathcal{N}[\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0), \rho^2\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)^2 + \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)^2\}]. \tag{B.19}$$

Recall $a_t = \mathbb{I}\left\{\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(1) > \boldsymbol{x}_t^\top \boldsymbol{\beta}_t(0)\right\}$ in TS, based on the posterior distribution in (B.19). Therefore, on the Event E we have

$$\begin{aligned}
&\mathbb{E}[\mathbb{I}\{a_t = 0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1), \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)] \\
&= \Pr\left\{\boldsymbol{x}_t^\top \boldsymbol{\beta}_t(1) - \boldsymbol{x}_t^\top \boldsymbol{\beta}_t(0) < 0|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1), \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\right\} \\
=&\Phi[-\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\}/\sqrt{\rho^2\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)^2 + \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)^2\}}] \\
=&1 - \Phi[\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\}/\sqrt{\rho^2\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)^2 + \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)^2\}}],
\end{aligned} \tag{B.20}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Denote $\widehat{z}_t \equiv \{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\}/\sqrt{\rho^2\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)^2 + \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)^2\}} > 0$, since $\widehat{\pi}_t(\boldsymbol{x}_t) = \mathbb{I}\{\widehat{\mu}_{t-1}(\boldsymbol{x}, 1) > \widehat{\mu}_{t-1}(\boldsymbol{x}, 0)\} = 1$, i.e., $\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0) > 0$. By applying the tail bound established for the normal distribution in Section 7.1 of Feller (2008), we have (B.20) can be bounded as

$$\mathbb{E}[\mathbb{I}\{a_t = 0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1), \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)] \leq \exp(-\widehat{z}_t^2/2).$$

This yields that on the Event E,

$$\begin{aligned}
&\mathbb{E}[\mathbb{I}\{a_t = 0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1), \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)] \\
&\leq \mathbb{E}\left\{\exp(-\widehat{z}_t^2/2)\right\} = \mathbb{E}\left\{\exp\left(-\frac{\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\}^2}{2\rho^2\{\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)^2 + \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)^2\}}\right)\right\}.
\end{aligned}$$

Using similar arguments in proving (B.13)that $\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0) \leq \frac{L_{\boldsymbol{x}}}{\sqrt{(t-1)p_{t-1}\lambda}}$, we have

$$\widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 1)^2 + \widehat{\sigma}_{t-1}(\boldsymbol{x}_t, 0)^2 \leq \frac{2L_{\boldsymbol{x}}^2}{(t-1)p_{t-1}\lambda}.$$

Therefore, combining the above two equations leads to

$$\mathbb{E}[\mathbb{I}\{a_t = 0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1), \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)] \leq \mathbb{E}\left\{\exp\left(-\frac{\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\}^2(t-1)p_{t-1}\lambda}{4\rho^2 L_{\boldsymbol{x}}^2}\right)\right\},$$

where the expectation is taken with respect to history $\mathcal{H}_{t-1}$.

Note that on the Event E, we have

$$\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, 0)\}^2 \geq \left(|\Delta_{\boldsymbol{x}_t}| - \xi\right)^2,$$

which follows that on the Event E,

$$\begin{aligned}\mathbb{E}[\mathbb{I}\{a_t=0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1),\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)] &\leq \mathbb{E}\left\{\exp\left(-\frac{\left(|\Delta_{\boldsymbol{x}_t}|-\xi\right)^2(t-1)p_{t-1}\lambda}{4\rho^2L_{\boldsymbol{x}}^2}\right)\right\}\\ &\leq \exp\left(-\frac{\left(|\Delta_{\boldsymbol{x}_t}|-\xi\right)^2(t-1)p_{t-1}\lambda}{4\rho^2L_{\boldsymbol{x}}^2}\right).\end{aligned} \tag{B.21}$$

Step 3: Combined with Equation (B.17), Equation (B.18) and Equation (B.21), we have

$$\begin{aligned}\Pr\{a_t\neq\widehat{\pi}_t(\boldsymbol{x}_t)\} &\overset{\text{(B.17)}}{\leq} \Pr\{a_t\neq\widehat{\pi}_t(\boldsymbol{x}_t)|E\}+4d\exp\left\{-(t-1)p_{t-1}^2c_\xi\right\}\\ &\overset{\text{(B.18)}}{\leq} \mathbb{E}\left(\mathbb{E}[\mathbb{I}\{a_t=0\}|\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1),\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)]|E\right)+4d\exp\left\{-(t-1)p_{t-1}^2c_\xi\right\}\\ &\overset{\text{(B.21)}}{\leq} \exp\left(-\frac{\left(|\Delta_{\boldsymbol{x}_t}|-\xi\right)^2(t-1)p_{t-1}\lambda}{4\rho^2L_{\boldsymbol{x}}^2}\right)+4d\exp\left\{-(t-1)p_{t-1}^2c_\xi\right\}.\end{aligned}$$

The proof is hence completed.

B.4 Proof of Theorem 2

We detail the proof of Theorem 2 in this section. Using the similar arguments in (B.2) in the proof of Lemma 4.1, we can rewrite $\sqrt{t}\{\widehat{\boldsymbol{\beta}}_t(a)-\boldsymbol{\beta}(a)\}$ as

$$\begin{aligned}\sqrt{t}\{\widehat{\boldsymbol{\beta}}_t(a)-\boldsymbol{\beta}(a)\} = &\underbrace{\left\{\frac{1}{t}\sum_{i=1}^t\mathbb{I}(a_i=a)\boldsymbol{x}_i\boldsymbol{x}_i^\top+\frac{1}{t}\omega\boldsymbol{I}_d\right\}^{-1}}_{\boldsymbol{\xi}}\underbrace{\left\{\frac{1}{\sqrt{t}}\sum_{i=1}^t\mathbb{I}(a_i=a)\boldsymbol{x}_ie_i\right\}}_{\boldsymbol{\eta}_1}\\ &\underbrace{-\left\{\frac{1}{t}\sum_{i=1}^t\mathbb{I}(a_i=a)\boldsymbol{x}_i\boldsymbol{x}_i^\top+\frac{1}{t}\omega\boldsymbol{I}_d\right\}^{-1}\frac{1}{\sqrt{t}}\boldsymbol{\beta}(a)}_{\boldsymbol{\eta}_2}.\end{aligned}$$

Our goal is to prove that $\sqrt{t}\{\widehat{\boldsymbol{\beta}}_t(a)-\boldsymbol{\beta}(a)\}$ is asymptotically normal. The proof is to generalize Theorem 3.1 in Chen et al. (2020) by considering commonly used bandit algorithms, including UCB, TS, and EG here. We complete the proof in the following four steps:

- Step 1: Prove that $\boldsymbol{\eta}_1 = (1/\sqrt{t})\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i e_i \xrightarrow{D} \mathcal{N}_d\left(\mathbf{0}_d, G_a\right)$, where G_a is the variance matrix to be spesified shortly.
- Step 2: Prove that $\boldsymbol{\xi} = \left\{(1/t)\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + (\omega/t)\boldsymbol{I}_d\right\}^{-1} \xrightarrow{p} \sigma_a^2 G_a^{-1}$, where $\sigma_a^2 = \mathbb{E}(e_t^2|a_t = a)$ for $a = 0, 1$.
- Step 3: Prove that $\boldsymbol{\eta}_2 = \left\{(1/t)\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + (\omega/t)\boldsymbol{I}_d\right\}^{-1}(1/\sqrt{t})\boldsymbol{\beta}(a) \xrightarrow{p} \mathbf{0}_d$.
- Step 4: Combine above results in steps 1-3 using Slutsky's theorem.

Step 1: We first focus on proving that $\boldsymbol{\eta}_1 = (1/\sqrt{t})\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{x}_i e_i \xrightarrow{D} \mathcal{N}_d\left(\mathbf{0}_d, G_a\right)$. Using Cramer-Wold device, it suffices to show that for any $\boldsymbol{v} \in \mathbb{R}^d$,

$$\boldsymbol{\eta}_1(\boldsymbol{v}) \equiv \frac{1}{\sqrt{t}}\sum_{i=1}^{t}\mathbb{I}(a_i = a)\boldsymbol{v}^\top\boldsymbol{x}_i e_i \xrightarrow{D} \mathcal{N}_d\left(0, \boldsymbol{v}^\top G_a\boldsymbol{v}\right).$$

Note that $\mathbb{E}\left(\mathbb{I}(a_i = a)\boldsymbol{v}^\top\boldsymbol{x}_i e_i \mid \mathcal{H}_{i-1}\right) = \mathbb{E}\left(\mathbb{I}(a_i = a)\boldsymbol{v}^\top\boldsymbol{x}_i \mid \mathcal{H}_{i-1}\right)\mathbb{E}\left(e_i \mid \mathcal{H}_{i-1}, a_i = a\right) = 0$, we have that $\mathbb{I}(a_i = a)\boldsymbol{v}^\top\boldsymbol{x}_i e_i$ is a Martingale difference sequence. We next show the asymptotic normality of $\boldsymbol{\eta}_1(\boldsymbol{v})$ using Martingale central limit theorem, by the following two parts: i) check the conditional Lindeberg condition; ii) derivative the limit of the conditional variance.

Firstly, we check the conditional Lindeberg condition. For any $\delta > 0$, denote

$$\psi = \sum_{i=1}^{t}\mathbb{E}\left[\frac{1}{t}\mathbb{I}(a_i = a)\left(\boldsymbol{v}^\top\boldsymbol{x}_i\right)^2 e_i^2\mathbb{I}\left\{\left|\frac{1}{\sqrt{t}}\mathbb{I}(a_i = a)\boldsymbol{v}^\top\boldsymbol{x}_i e_i\right| > \delta\right\} \mid \mathcal{H}_{i-1}\right]. \tag{B.22}$$

Notice that $\left(\boldsymbol{v}^\top\boldsymbol{x}_i\right)^2 \leq \|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d$, we have

$$\mathbb{I}\left\{\left|\frac{1}{\sqrt{t}}\mathbb{I}(a_i = a)\boldsymbol{v}^\top\boldsymbol{x}_i e_i\right| > \delta\right\} \leq \mathbb{I}\left\{\mathbb{I}(a_i = a)e_i^2\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d > t\delta^2\right\} = \mathbb{I}\left\{\mathbb{I}(a_i = a)e_i^2 > \frac{t\delta^2}{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}\right\}.$$

Combining this with (B.22), we obtain that

$$\psi \leq \frac{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}{t}\sum_{i=1}^{t}\mathbb{E}\left(\mathbb{I}(a_i = a)e_i^2\mathbb{I}\left\{\mathbb{I}(a_i = a)e_i^2 > \frac{t\delta^2}{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}\right\} \mid \mathcal{H}_{i-1}\right), \tag{B.23}$$

where the right hand side equals

$$\frac{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}{t} \sum_{i=1}^{t} \mathbb{E}\left(\mathbb{I}(a_i = a) \mid \mathcal{H}_{i-1}\right) \mathbb{E}\left(e_i^2 \mathbb{I}\left\{\mathbb{I}(a_i = a) e_i^2 > \frac{\delta^2 t}{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}\right\} \mid \mathcal{H}_{i-1}\right).$$

Then, we can further write (B.23) as

$$\psi \leq \frac{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}{t} \sum_{i=1}^{t} \mathbb{E}\left(e_{i(a)}^2 \mathbb{I}\left\{e_{i(a)}^2 > \frac{\delta^2 t}{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}\right\} \mid \mathcal{H}_{i-1}\right). \tag{B.24}$$

where $e_{i(a)} = e_i$ when $a_i = a$ and 0 otherwise. Since e_i conditioned on a_i are *i.i.d.*, $\forall i$, we have the right hand side of the above inequality equals

$$\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d \mathbb{E}\left(e^2 \mathbb{I}\left\{e^2 > \frac{t\delta^2}{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}\right\}\right),$$

where e is the random variable given by $e_i | \mathcal{H}_{i-1}$. Note that $e^2 \mathbb{I}\left\{e^2 > t\delta^2/(\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d)\right\}$ is dominated by e^2 with $\mathbb{E}e^2 < \infty$ and converges to 0, as $t \to \infty$. Then, by Dominated Convergence Theorem, the results in (B.24) can be further bounded by

$$\psi \leq \frac{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}{t} \sum_{i=1}^{t} \mathbb{E}\left(e^2 \mathbb{I}\left\{e^2 > \frac{t\delta^2}{\|\boldsymbol{v}\|_2^2 L_{\boldsymbol{x}}^2 d}\right\}\right) \to 0, \text{ as } t \to \infty.$$

Therefore, conditional Lindeberg condition holds.

Secondly, we derive the limit of the conditional variance. Notice that

$$\begin{aligned}
\frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\mathbb{I}(a_i = a)\left(\boldsymbol{v}^\top \boldsymbol{x}_i\right)^2 e_i^2 \mid \mathcal{H}_{i-1}\right] &= \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left\{\mathbb{E}\left[\mathbb{I}(a_i = a)\left(\boldsymbol{v}^\top \boldsymbol{x}_i\right)^2 e_i^2 \mid a_i, \boldsymbol{x}_i\right] \mid \mathcal{H}_{i-1}\right\} \\
&= \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left\{\mathbb{I}(a_i = a)\left(\boldsymbol{v}^\top \boldsymbol{x}_i\right)^2 \mathbb{E}\left[e_i^2 \mid a_i = a, \boldsymbol{x}_i\right] \mid \mathcal{H}_{i-1}\right\}.
\end{aligned}$$

Since e_t is independent of $\mathcal{H}_{i-1}$ and $\boldsymbol{x}_i$ given a_t, and $\mathbb{E}\left[e_i^2 \mid a_i = a\right] = \sigma_a^2$, we have

$$\begin{aligned}
\frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\mathbb{I}(a_i = a)\left(\boldsymbol{v}^\top \boldsymbol{x}_i\right)^2 e_i^2 \mid \mathcal{H}_{i-1}\right] &= \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\mathbb{I}(a_i = a)\left(\boldsymbol{v}^\top \boldsymbol{x}_i\right)^2 \mathbb{E}\left[e_i^2 \mid a_i = a\right] \mid \mathcal{H}_{i-1}\right] \\
&= \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\mathbb{I}(a_i = a) \boldsymbol{v}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{v} \sigma_a^2 \mid \mathcal{H}_{i-1}\right] \\
&= \frac{1}{t} \sum_{i=1}^{t} \sigma_a^2 \mathbb{E}\left[\mathbb{I}(a_i = a) \boldsymbol{v}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{v} \mid \mathcal{H}_{i-1}\right],
\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}(a_i=a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\mid\mathcal{H}_{i-1}\right] = &\quad \mathbb{E}\{\mathbb{E}\left[\mathbb{I}(a_i\neq\pi^*(\boldsymbol{x}_i))\mathbb{I}(\pi^*(\boldsymbol{x}_i)\neq a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\mid\boldsymbol{x}_i\right]\mid\mathcal{H}_{i-1}\}\\ &+\quad \mathbb{E}\{\mathbb{E}\left[\mathbb{I}(a_i=\pi^*(\boldsymbol{x}_i))\mathbb{I}(\pi^*(\boldsymbol{x}_i)=a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\mid\boldsymbol{x}_i\right]\mid\mathcal{H}_{i-1}\}\\ &=\quad \mathbb{E}\{\mathbb{E}\left[\mathbb{I}(a_i\neq\pi^*(\boldsymbol{x}_i))\mid\boldsymbol{x}_i,\mathcal{H}_{i-1}\right]\mathbb{I}(\pi^*(\boldsymbol{x}_i)\neq a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\}\\ &+\quad \mathbb{E}\{\mathbb{E}\left[\mathbb{I}(a_i=\pi^*(\boldsymbol{x}_i))\mid\boldsymbol{x}_i,\mathcal{H}_{i-1}\right]\mathbb{I}(\pi^*(\boldsymbol{x}_i)=a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\}.\end{aligned}$$

Here, the first equation comes from iteration expectation over $\boldsymbol{x}_i$ and the fact that $\mathbb{I}(\pi^*(\boldsymbol{x}_i)\neq a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}$ is a constant given $\boldsymbol{x}_i$ and $\mathcal{H}_{i-1}$.

$$\begin{aligned}\mathbb{I}(a_i=a) &= \mathbb{I}(a_i=a\neq\pi^*(\boldsymbol{x}_i))+\mathbb{I}(a_i=a=\pi^*(\boldsymbol{x}_i))\\ &= \mathbb{I}(a_i\neq\pi^*(\boldsymbol{x}_i))\mathbb{I}(\pi^*(\boldsymbol{x}_i)\neq a)+\mathbb{I}(a_i=\pi^*(\boldsymbol{x}_i))\mathbb{I}(\pi^*(\boldsymbol{x}_i)=a),\end{aligned}$$

and the second equation is owing to the fact that $\mathbb{I}(\pi^*(\boldsymbol{x}_i)=a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}$ is a constant given $\boldsymbol{x}_i$ and independent of $\mathcal{H}_{i-1}$. Define

$$\nu_i\left(\boldsymbol{x}_i,\mathcal{H}_{i-1}\right)\equiv\Pr\left\{a_i\neq\pi^*(\boldsymbol{x}_i)\middle|\,\boldsymbol{x}_i,\mathcal{H}_{i-1}\right\}=\mathbb{E}\left[\mathbb{I}\left\{a_i\neq\pi^*(\boldsymbol{x}_i)\right\}|\boldsymbol{x}_i,\mathcal{H}_{i-1}\right],\tag{B.25}$$

then the conditional variance can be expressed as

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^t\mathbb{E}\left[\mathbb{I}(a_i=a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\mid\mathcal{H}_{i-1}\right] = &\frac{1}{t}\sum_{i=1}^t\mathbb{E}\{\nu_i\left(\boldsymbol{x}_i,\mathcal{H}_{i-1}\right)\mathbb{I}(\pi^*(\boldsymbol{x}_i)\neq a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\}\\ &+\frac{1}{t}\sum_{i=1}^t\mathbb{E}\left[\{1-\nu_i\left(\boldsymbol{x}_i,\mathcal{H}_{i-1}\right)\}\mathbb{I}(\pi^*(\boldsymbol{x}_i)=a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v}\right],\end{aligned}$$

which can be expressed as

$$\begin{aligned}&\frac{1}{t}\sum_{i=1}^t\int\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right)\mathbb{I}\{\boldsymbol{x}^\top\boldsymbol{\beta}(a)<\boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}}\\ &+\frac{1}{t}\sum_{i=1}^t\int\{1-\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right)\}\mathbb{I}\left\{\boldsymbol{x}^\top\boldsymbol{\beta}(a)\geq\boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\right\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}}\\ &=\int\frac{1}{t}\sum_{i=1}^t\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right)\mathbb{I}\{\boldsymbol{x}^\top\boldsymbol{\beta}(a)<\boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}}\\ &+\int\{1-\frac{1}{t}\sum_{i=1}^t\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right)\}\mathbb{I}\left\{\boldsymbol{x}^\top\boldsymbol{\beta}(a)\geq\boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\right\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}}.\end{aligned}$$

Since $\lim_{i\to\infty} \Pr\{a_i \neq \pi^*(\boldsymbol{x})\} = \kappa_\infty(\boldsymbol{x})$, we have for any $\epsilon > 0$, there exist a constant $t_0 > 0$ such that $|\Pr\{a_i \neq \pi^*(\boldsymbol{x})\} - \kappa_\infty(\boldsymbol{x})| < \epsilon$ for all $i \geq t_0$. Therefore, for the expectation of $\frac{1}{t}\sum_{i=1}^{t} \nu_i(\boldsymbol{x}, \mathcal{H}_{i-1})$ over the history, we have

$$\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} \nu_i(\boldsymbol{x}, \mathcal{H}_{i-1})\right] = \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}\left[\nu_i(\boldsymbol{x}, \mathcal{H}_{i-1})\right] = \frac{1}{t}\sum_{i=1}^{t} \Pr\{a_i \neq \pi^*(\boldsymbol{x})\}.$$

It follows immediately that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} \nu_i(\boldsymbol{x}, \mathcal{H}_{i-1})\right] - \kappa_\infty(\boldsymbol{x}) \\
&= \frac{1}{t}\sum_{i=1}^{t_0} \left[\Pr\{a_i \neq \pi^*(\boldsymbol{x})\} - \kappa_\infty(\boldsymbol{x})\right] + \frac{1}{t}\sum_{i=t_0}^{t} \left[\Pr\{a_i \neq \pi^*(\boldsymbol{x})\} - \kappa_\infty(\boldsymbol{x})\right].
\end{aligned}$$

Therefore, by the triangle inequality, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} \nu_i(\boldsymbol{x}, \mathcal{H}_{i-1})\right] - \kappa_\infty(\boldsymbol{x})\right| \\
&= \frac{1}{t}\sum_{i=1}^{t_0} \left|\Pr\{a_i \neq \pi^*(\boldsymbol{x})\} - \kappa_\infty(\boldsymbol{x})\right| + \frac{1}{t}\sum_{i=t_0}^{t} \left|\Pr\{a_i \neq \pi^*(\boldsymbol{x})\} - \kappa_\infty(\boldsymbol{x})\right| \\
&< \frac{1}{t}\sum_{i=1}^{t_0} \left[\left|\Pr\{a_i \neq \pi^*(\boldsymbol{x})\}\right| + \left|\kappa_\infty(\boldsymbol{x})\right|\right] + \frac{1}{t}\sum_{i=t_0}^{t} \epsilon \\
&\leq \frac{1}{t}\sum_{i=1}^{t_0} 2 + \frac{1}{t}\sum_{i=t_0}^{t} \epsilon = \frac{2t_0}{t} + \frac{t - t_0}{t}\epsilon.
\end{aligned}$$

Since the above equation holds for any $\epsilon > 0$ and $0 < \frac{t-t_0}{t} < 1$, we have

$$\left|\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} \nu_i(\boldsymbol{x}, \mathcal{H}_{i-1})\right] - \kappa_\infty(\boldsymbol{x})\right| \leq \frac{2t_0}{t},$$

which goes to zero as $t \to \infty$. Thus,

$$\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t} \nu_i(\boldsymbol{x}, \mathcal{H}_{i-1})\right] = \kappa_\infty(\boldsymbol{x}) + o_p(1). \tag{B.26}$$

Next, we consider the variance of $\frac{1}{t}\sum_{i=1}^{t}\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right)$. Denote $\mathbb{E}\left[\frac{1}{t}\sum_{i=1}^{t}\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right)\right] = \mu_\nu(\boldsymbol{x})$, we have $\mu_\nu(\boldsymbol{x}) = \kappa_\infty(\boldsymbol{x})+o_p(1)$. Notice that $\nu_i\left(\boldsymbol{x}_i,\mathcal{H}_{i-1}\right) \equiv \Pr\left\{a_i \neq \pi^*(\boldsymbol{x}_i)|\, \boldsymbol{x}_i,\mathcal{H}_{i-1}\right\} \in [0,1]$, by Lemma B.1, we have

$$\operatorname{Var}\left[\frac{1}{t}\sum_{i=1}^{t}\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right)\right] \leq \mu_\nu(\boldsymbol{x}) - \mu_\nu(\boldsymbol{x})^2 = \kappa_\infty(\boldsymbol{x})\left\{1-\kappa_\infty(\boldsymbol{x})\right\} + o_p(1),$$

which goes to zero as $t \to \infty$. Combined with Equation (B.26), it follows immediately that as t goes to ∞, we have

$$\frac{1}{t}\sum_{i=1}^{t}\nu_i\left(\boldsymbol{x},\mathcal{H}_{i-1}\right) \to \kappa_\infty(\boldsymbol{x}).$$

Therefore, as t goes to ∞, we have $\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\mathbb{I}(a_i=a)\boldsymbol{v}^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{v} \mid \mathcal{H}_{i-1}\right]$ converges to

$$\begin{aligned}&\int \kappa_\infty(\boldsymbol{x})\mathbb{I}\{\boldsymbol{x}^\top\boldsymbol{\beta}(a) < \boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}} \\ &+\int\left\{1-\kappa_\infty(\boldsymbol{x})\right\}\mathbb{I}\left\{\boldsymbol{x}^\top\boldsymbol{\beta}(a) \geq \boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\right\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}}.\end{aligned} \tag{B.27}$$

Thus, following the similar arguments in S1.2 in Chen et al. (2020), we have

$$\boldsymbol{\eta}_1(\boldsymbol{v}) = \frac{1}{\sqrt{t}}\sum_{i=1}^{t}\mathbb{I}(a_i=a)\boldsymbol{v}^\top\boldsymbol{x}_i e_i \xrightarrow{D} \mathcal{N}_d\left(0,\boldsymbol{v}^\top G_a\boldsymbol{v}\right),$$

where

$$\begin{aligned}\boldsymbol{v}^\top G_a\boldsymbol{v} = \sigma_a^2\Bigg\{&\int \kappa_\infty(\boldsymbol{x})\mathbb{I}\{\boldsymbol{x}^\top\boldsymbol{\beta}(a) < \boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}} \\ &+\int\left\{1-\kappa_\infty(\boldsymbol{x})\right\}\mathbb{I}\left\{\boldsymbol{x}^\top\boldsymbol{\beta}(a) \geq \boldsymbol{x}^\top\boldsymbol{\beta}(1-a)\right\}\boldsymbol{v}^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{v}dP_{\mathcal{X}}\Bigg\}.\end{aligned}$$

Finally, by Martingale Central Limit Theorem, we have

$$\boldsymbol{\eta}_1 = \frac{1}{\sqrt{t}}\sum_{i=1}^{t}\mathbb{I}(a_i=a)\boldsymbol{x}_i e_i \xrightarrow{D} \mathcal{N}_d\left(0,G_a\right),$$

where

$$G_a = \sigma_a^2 \left\{ \int \kappa_\infty(\boldsymbol{x}) \mathbb{I}\{\boldsymbol{x}^\top \boldsymbol{\beta}(a) < \boldsymbol{x}^\top \boldsymbol{\beta}(1-a)\} \boldsymbol{x}\boldsymbol{x}^\top dP_{\mathcal{X}} \right. \\ \left. + \int \{1-\kappa_\infty(\boldsymbol{x})\} \mathbb{I}\left\{\boldsymbol{x}^\top \boldsymbol{\beta}(a) \geq \boldsymbol{x}^\top \boldsymbol{\beta}(1-a)\right\} \boldsymbol{x}\boldsymbol{x}^\top dP_{\mathcal{X}} \right\}. \tag{B.28}$$

The first part is thus completed.

Step 2: We next show that $\boldsymbol{\xi} = \left\{(1/t)\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega \boldsymbol{I}_d\right\}^{-1} \xrightarrow{p} \sigma_a^2 G_a^{-1}$, which is sufficient to find the limit of $\left\{(1/t)\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega \boldsymbol{I}_d\right\}$. By Lemma 6 in Chen et al. (2020), it suffices to show the limit of $\frac{1}{t}\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{v}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{v}$ for any $\boldsymbol{v} \in \mathbb{R}^d$.

Since $\Pr(|\mathbb{I}(a_i = a)\boldsymbol{v}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{v}| > h) \leq \Pr(|\boldsymbol{v}^\top \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{v}| > h)$ for each $h > 0$ and $i \leq 1$, by Theorem 2.19 in Hall and Heyde (2014), we have

$$\frac{1}{t}\sum_{i=1}^t \left[\mathbb{I}(a_i = a)\boldsymbol{v}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{v} - \mathbb{E}\left\{\mathbb{I}(a_i = a)\boldsymbol{v}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{v} \mid \mathcal{H}_{i-1}\right\}\right] \xrightarrow{p} 0, \text{ as } t \to \infty. \tag{B.29}$$

Recall the results in (B.26) and (B.28), we have $\mathbb{E}\left\{\mathbb{I}(a_i = a)\boldsymbol{v}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{v} \mid \mathcal{H}_{i-1}\right\} = G_a/\sigma_a^2$. Combining this with (B.29), we have

$$\frac{1}{t}\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{v}^\top \boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{v} \xrightarrow{p} \frac{G_a}{\sigma_a^2}, \text{ as } t \to \infty.$$

By Lemma 6 in Chen et al. (2020) and Continuous Mapping Theorem, we further have

$$\boldsymbol{\xi} = \left\{\frac{1}{t}\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega \boldsymbol{I}_d\right\}^{-1} \xrightarrow{p} \sigma_a^2 G_a^{-1}.$$

Step 3: We focus on proving $\boldsymbol{\eta}_2 = \left\{(1/t)\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + (\omega/t)\boldsymbol{I}_d\right\}^{-1}(1/\sqrt{t})\boldsymbol{\beta}(a) \xrightarrow{p} \mathbf{0}_d$ next. This suffices to show that $\boldsymbol{b}_i^\top \boldsymbol{\eta}_2 \xrightarrow{p} 0$ holds for any standard basis $\boldsymbol{b}_i \in \mathbb{R}^d$. Since

$$\boldsymbol{b}_i^\top \boldsymbol{\eta}_2 = \left\| \boldsymbol{b}_i^\top \left\{\frac{1}{t}\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega \boldsymbol{I}_d\right\}^{-1} \frac{1}{\sqrt{t}}\boldsymbol{\beta}(a) \right\|_2 \\ \leq \frac{1}{\sqrt{t}} \left\|\boldsymbol{b}_i^\top\right\|_2 \left\| \left\{\frac{1}{t}\sum_{i=1}^t \mathbb{I}(a_i = a)\boldsymbol{x}_i\boldsymbol{x}_i^\top + \frac{1}{t}\omega \boldsymbol{I}_d\right\}^{-1} \right\|_2 \|\boldsymbol{\beta}(a)\|_2,$$

and by (B.4), we have

$$\boldsymbol{b}_i^\top \boldsymbol{\eta}_2 \leq \frac{\|\boldsymbol{\beta}(a)\|_2}{\sqrt{tp_t^2}\lambda + \frac{1}{\sqrt{t}}\omega}.$$

Thus, we have $\boldsymbol{b}_i^\top \boldsymbol{\eta}_2 \xrightarrow{p} 0$, as $tp_t^2 \to \infty$.

Step 4: Finally, we combine the above results using Slutsky's theorem, and conclude that

$$\sqrt{t}\{\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\} = \boldsymbol{\xi}\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2 \xrightarrow{D} \mathcal{N}_d\left(\mathbf{0}_d, \sigma_a^4 G_a^{-1}\right),$$

where G_a is defined in (B.28). Denote the variance term as

$$\sigma^2_{\boldsymbol{\beta}(a)} = \sigma_a^2 \left[\int \kappa_\infty(\boldsymbol{x}) \mathbb{I}\{\boldsymbol{x}^\top \boldsymbol{\beta}(a) < \boldsymbol{x}^\top \boldsymbol{\beta}(1-a)\} \boldsymbol{x}\boldsymbol{x}^\top dP_{\mathcal{X}} \right.$$
$$\left. + \int \{1 - \kappa_\infty(\boldsymbol{x})\} \mathbb{I}\left\{\boldsymbol{x}^\top \boldsymbol{\beta}(a) \geq \boldsymbol{x}^\top \boldsymbol{\beta}(1-a)\right\} \boldsymbol{x}\boldsymbol{x}^\top dP_{\mathcal{X}} \right]^{-1},$$

with $\sigma_a^2 = \mathbb{E}(e_t^2 | a_t = a)$ denoting the conditional variance of e_t given $a_t = a$, for $a = 0, 1$, we have

$$\sqrt{t}\left\{\widehat{\boldsymbol{\beta}}_t(a) - \boldsymbol{\beta}(a)\right\} \xrightarrow{D} \mathcal{N}_d\{\mathbf{0}_d, \sigma^2_{\boldsymbol{\beta}(a)}\}.$$

The proof is hence completed.

B.5 Proof of Theorem 3

Finally, we prove the asymptotic normality of the proposed value estimator under DREAM in Theorem 3 in this section. The proof consists of four steps. In step 1, we aim to show

$$\widehat{V}_T = \widetilde{V}_T + o_p(T^{-1/2}),$$

where

$$\widetilde{V}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1 - \kappa_t(\boldsymbol{x}_t)} \left[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right] + \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}.$$

Next, in Step 2, we establish

$$\widetilde{V}_T = \overline{V}_T + o_p(T^{-1/2}),$$

where

$$\overline{V}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \left[r_t - \mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\}\right] + \mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\}.$$

The above two steps yields that

$$\widehat{V}_T = \overline{V}_T + o_p(T^{-1/2}). \tag{B.30}$$

Then, in Step 3, based on (B.30) and Martingale Central Limit Theorem, we show

$$\sqrt{T}(\widehat{V}_T - V^*) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2_{DR}\right),$$

with

$$\sigma^2_{DR} = \int_{\boldsymbol{x}} \frac{\pi^*(\boldsymbol{x})\sigma_1^2 + \{1-\pi^*(\boldsymbol{x})\}\sigma_0^2}{1-\kappa_\infty(\boldsymbol{x})} dP_{\mathcal{X}} + \mathrm{Var}\left[\mu\{\boldsymbol{x}, \pi^*(\boldsymbol{x})\}\right],$$

where $\sigma_a^2 = \mathbb{E}(e_t^2 | a_t = a)$ for $a = 0, 1$, and $\kappa_t \to \kappa_\infty$ as $t \to \infty$.

Lastly, in Step 4, we show the variance estimator in Equation (5) in the main paper is a consistent estimator of σ^2_{DR}. The proof for Theorem 3 is thus completed.

Step 1: We first show $\widehat{V}_T = \widetilde{V}_T + o_p(T^{-1/2})$. To this end, define a middle term as

$$\widetilde{\phi}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\kappa_t(\boldsymbol{x}_t)}\Big[r_t - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big] + \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}.$$

Thus, it suffices to show $\widehat{V}_T = \widetilde{\phi}_T + o_p(T^{-1/2})$ and $\widetilde{\phi}_T = \widetilde{V}_T + o_p(T^{-1/2})$.

Firstly, we have

$$\begin{aligned}
\widehat{V}_T - \widetilde{\phi}_T &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\kappa_t(\boldsymbol{x}_t)}\right]\Big[r_t - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big] \qquad \text{(B.31)} \\
&= \frac{1}{T}\sum_{t=1}^{T}\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right].
\end{aligned}$$

We can further decompose (B.31) by

$$\begin{aligned}
\widehat{V}_T - \widetilde{\phi}_T = \frac{1}{T}\sum_{t=1}^{T} \quad & \{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\} \\
& \left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} + \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1 - \kappa_t(\boldsymbol{x}_t)\}}\right] \\
= \frac{1}{T}\sum_{t=1}^{T} \quad & \{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\} \left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1 - \kappa_t(\boldsymbol{x}_t)\}}\right] \qquad \text{(B.32)} \\
+ \frac{1}{T}\sum_{t=1}^{T} \quad & \{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\} \left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1 - \kappa_t(\boldsymbol{x}_t)\}}\right].
\end{aligned}$$

We first show that the first term in (B.32) is $o_p(T^{-1/2})$. Define a class of function

$$\mathcal{F}_\kappa(\boldsymbol{x}, a, r) = \left\{ \{\widehat{\kappa}(\boldsymbol{x}) - \kappa(\boldsymbol{x})\} \left[\frac{\mathbb{I}\{a = \pi(\boldsymbol{x})\}\Big[r - \mu\{\boldsymbol{x}, \pi(\boldsymbol{x})\}\Big]}{\{1 - \widehat{\kappa}(\boldsymbol{x})\}\{1 - \kappa(\boldsymbol{x})\}}\right] : \widehat{\kappa}(\cdot), \kappa(\cdot) \in \Lambda, \pi(\cdot) \in \Pi \right\},$$

where Π and Λ are two classes of functions that maps context $\boldsymbol{x} \in \mathcal{X}$ to a probability. Define the supremum of the empirical process indexed by $\mathcal{F}_\kappa$ as

$$||\mathbb{G}_n||_{\mathcal{F}} \equiv \quad \sup_{\pi \in \Pi} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) - \mathbb{E}\{\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t)|\mathcal{H}_{t-1}\}\right]. \qquad \text{(B.33)}$$

Notice that

$$\begin{aligned}
& \mathbb{E}\{\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t)|\mathcal{H}_{t-1}\} \\
=& \mathbb{E}\left(\{\widehat{\kappa}(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\} \left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1 - \widehat{\kappa}(\boldsymbol{x}_t)\}\{1 - \kappa_t(\boldsymbol{x}_t)\}}\right] \mid \mathcal{H}_{t-1}\right) \\
=& \mathbb{E}\left(\{\widehat{\kappa}(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\} \left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}e_t}{\{1 - \widehat{\kappa}(\boldsymbol{x}_t)\}\{1 - \kappa_t(\boldsymbol{x}_t)\}}\right] \mid \mathcal{H}_{t-1}\right),
\end{aligned}$$

by the definitions and thus, using the iteration of expectation, we have

$$\begin{aligned}
&\mathbb{E}\{\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t)|\mathcal{H}_{t-1}\} \\
=&\mathbb{E}\left\{\mathbb{E}\left(\{\widehat{\kappa}(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}e_t}{\{1-\widehat{\kappa}(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right] \mid a_t, \boldsymbol{x}_t\right) \mid \mathcal{H}_{t-1}\right\} \\
=&\mathbb{E}\left\{\{\widehat{\kappa}(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}\left[\frac{\mathbb{E}\{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\} \mid \boldsymbol{x}_t\}}{\{1-\widehat{\kappa}(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]\mathbb{E}\left(e_t \mid a_t, \boldsymbol{x}_t\right) \mid \mathcal{H}_{t-1}\right\} = 0,
\end{aligned}$$

where the last equation is due to the definition of the noise e_t. Therefore, Equation (B.33) can be further written as

$$||\mathbb{G}_n||_{\mathcal{F}} \equiv \sup_{\pi\in\Pi}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t).$$

Next, we show the second moment is bounded by

$$\begin{aligned}
&\mathbb{E}\left\{\left(\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]\right)^2 \mid \mathcal{H}_{t-1}\right\} \\
=&\mathbb{E}\left\{\left[\frac{\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]^2 \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]^2 \mid \mathcal{H}_{t-1}\right\}, \\
=&\mathbb{E}\left\{\left[\frac{\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]^2 \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}e_t^2 \mid \mathcal{H}_{t-1}\right\}
\end{aligned}$$

by the definitions and thus, using the iteration of expectation, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left(\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]\right)^2 \mid \mathcal{H}_{t-1}\right\} \\
=&\mathbb{E}\left(\mathbb{E}\left\{\left[\frac{\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]^2 \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}e_t^2 \mid a_t, \boldsymbol{x}_t\right\} \mid \mathcal{H}_{t-1}\right) \\
=&\mathbb{E}\left(\left[\frac{\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]^2 \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\mathbb{E}\left\{e_t^2 \mid a_t, \boldsymbol{x}_t\right\} \mid \mathcal{H}_{t-1}\right) \\
=&\mathbb{E}\left(\left[\frac{1}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)} - \frac{1}{1-\kappa_t(\boldsymbol{x}_t)}\right]^2 \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\sigma_{a_t}^2 \mid \mathcal{H}_{t-1}\right).
\end{aligned}$$

Notice that $\widehat{\kappa}_t(\boldsymbol{x}_t) \leq C_1 < 1$ and $\kappa_t(\boldsymbol{x}_t) \leq C_1 < 1$ for sure for some constant $C_1 < 1$ (by definition of a valid bandit algorithm and results of Theorem 1), we have

$$\left[\frac{1}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)} - \frac{1}{1-\kappa_t(\boldsymbol{x}_t)}\right]^2 \leq \left\{\left|\frac{1}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)}\right| + \left|\frac{1}{1-\kappa_t(\boldsymbol{x}_t)}\right|\right\}^2 \leq \left(\frac{2}{1-C_1}\right)^2 \equiv C_2,$$

where C_2 is a bounded constant. Thus we have

$$\begin{aligned}
&\mathbb{E}\left(\left[\frac{1}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)} - \frac{1}{1-\kappa_t(\boldsymbol{x}_t)}\right]^2 \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\sigma^2_{a_t} \mid \mathcal{H}_{t-1}\right) \\
&\leq \mathbb{E}\left(C_2 \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\sigma^2_{a_t} \mid \mathcal{H}_{t-1}\right) \\
&\leq C_2 \max\{\sigma_0^2, \sigma_1^2\}.
\end{aligned}$$

Therefore, for the second moment of the inner term of the first term, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left\{\left(\frac{1}{\sqrt{T}}\left\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\right\}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\left[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right]}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right]\right)^2 \mid \mathcal{H}_{t-1}\right\} \\
&\leq \sum_{t=1}^{T} \frac{C_2 \max\{\sigma_0^2, \sigma_1^2\}}{T} = C_2 \max\{\sigma_0^2, \sigma_1^2\} < \infty.
\end{aligned}$$

Therefore, we have

$$d_1(f) \equiv \left\|\mathbb{E}\left(|\ \mathcal{F}_\kappa(\boldsymbol{x}_1, a_1, r_1)\|\mathcal{H}_0\right)\right\|_\infty < \infty,$$

and

$$d_2(f) \equiv \left\|\mathbb{E}\left((\mathcal{F}_\kappa(\boldsymbol{x}_1, a_1, r_1))^2 \mid \mathcal{H}_0\right)\right\|_\infty^{1/2} < \infty.$$

It follows from the maximal inequality developed in Section 4.2 of Dedecker and Louhichi (2002) that there exist some constant $K \geq 1$ such that

$$\mathbb{E}\Big[||\mathbb{G}_n||_{\mathcal{F}}\Big] \lesssim K\left(\sqrt{p}d_2(f) + \frac{1}{\sqrt{T}}\left\|\max_{1\leq t\leq T}|\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) - \mathbb{E}\left(\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) \mid \mathcal{H}_{t-1}\right)|\right\|\right).$$

The above right-hand-side is upper bounded by

$$O(1)\sqrt{T^{-1/2}},$$

where $O(1)$ denotes some universal constant. Hence, we have

$$\mathbb{E}\Big[||\mathbb{G}_n||_{\mathcal{F}}\Big] = \mathcal{O}_p(T^{-1/2}). \tag{B.34}$$

Combined with Equation (B.33), we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) = \mathcal{O}_p(T^{-1/2}).$$

Therefore, for the first term in (B.32), we have

$$\frac{1}{T}\sum_{t=1}^{T}\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right] = \mathcal{O}_p(T^{-1}) = o_p(T^{-1/2}).$$

Then we consider the second term in (B.32), where

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\{\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big[\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]}{\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}}\right] \\
\leq\quad &\frac{1}{T}\sum_{t=1}^{T} B_\kappa \left|\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\right| \left|\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right|,
\end{aligned}$$

for some B_κ as the bound of $\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}/[\{1-\widehat{\kappa}_t(\boldsymbol{x}_t)\}\{1-\kappa_t(\boldsymbol{x}_t)\}]$. By Cauchy-Schwartz inequality, we have the above term further bounded by

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} B_\kappa \left|\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\right| \left|\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right| \\
\leq\quad &B_\kappa\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left|\widehat{\kappa}_t(\boldsymbol{x}_t) - \kappa_t(\boldsymbol{x}_t)\right|^2 \frac{1}{T}\sum_{t=1}^{T}\left|\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right|^2}.
\end{aligned}$$

Given Assumption 4.4, we have the above bounded by $o_p(T^{-1/2})$, and thus the second term in (B.32) is $o_p(T^{-1/2})$.

Therefore, we have $\widehat{V}_T = \widetilde{\phi}_T + o_p(T^{-1/2})$ hold.

Then, we focus on proving

$$\widetilde{\phi}_T - \widetilde{V}_T = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\kappa_t(\boldsymbol{x}_t)} - 1\right]\left[\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right], \tag{B.35}$$

is $o_p(T^{-1/2})$. Specifically, since

$$\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} = \mathcal{O}_p(t^{-1/2}),$$

by Theorem 2, and notice that

$$-1 \leq \frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\kappa_t(\boldsymbol{x}_t)} - 1 \leq \frac{1}{1-\kappa_t(\boldsymbol{x}_t)} - -1 \leq \frac{1}{1-C_1} - 1,$$

which is bounded, we have the inner part of (B.35) satisfies

$$\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\kappa_t(\boldsymbol{x}_t)} - 1\right]\left[\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \widehat{\mu}_{t-1}\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right] = \mathcal{O}_p(t^{-1/2}) = o_p(t^{-1/2-\alpha}),$$

for $0 < \alpha < 1/2$. Therefore, by Lemma 6 in Luedtke and Van Der Laan (2016), we have

$$\widetilde{\phi}_T - \widetilde{V}_T = \frac{1}{T}\sum_{t=1}^{T} o_p(t^{-1/2-\alpha}) = o_p(T^{-1/2-\alpha}) = o_p(T^{-1/2}).$$

Hence, $\widetilde{\phi}_T = \widetilde{V}_T + o_p(T^{-1/2})$ holds, and thus $\widehat{V}_T = \widetilde{V}_T + o_p(T^{-1/2})$ holds. The first step is thus completed.

Step 2: We next focus on proving $\widetilde{V}_T = \overline{V}_T + o_p(T^{-1/2})$. By definition of $\widetilde{V}_T$ and $\overline{V}_T$, we have

$$\begin{aligned}\sqrt{T}(\widetilde{V}_T - \overline{V}_T) = \quad & \underbrace{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\kappa_t(\boldsymbol{x}_t)} - 1\right]\left[\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right]}_{\eta_5} \\ & + \underbrace{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1-\kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}\right]\left[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right]}_{\eta_6}.\end{aligned} \tag{B.36}$$

We first show $\eta_5 = o_p(1)$. Since $\kappa_t(\boldsymbol{x}_t) \leq C_2 < 1$ for sure for some constant $0 < C_2 < 1$ (by definition of a valid bandit algorithm as results for Theorem 1), it suffices to show

$$\psi_5 = T^{-1/2} \sum_{t=1}^{T} |\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \pi^*\}| = o_p(1),$$

which is the direct conclusion of Lemma B.3.

Next, we show $\eta_6 = o_p(1)$. Firstly we can express η_6 as

$$\begin{aligned}
\eta_6 =& \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\frac{\mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}{1 - \kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right] \Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} \Big] \\
=& \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\frac{1}{1 - \kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right] \Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} \Big] \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\} \\
& - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \Big[r_t - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} \Big] \mathbb{I}\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\} \\
=& \underbrace{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\frac{1}{1 - \kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right] e_t \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\}}_{\psi_6} \\
& \underbrace{- \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{1}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \Big[\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} \Big] \mathbb{I}\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\} \mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}_{\psi_7}.
\end{aligned}$$

Note that

$$\begin{aligned}
\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\} &\geq \Pr\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t), \widehat{\pi}_t(\boldsymbol{x}_t) = \pi^*(\boldsymbol{x}_t)\} \\
&\geq \Pr\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\} + \Pr\{\widehat{\pi}_t(\boldsymbol{x}_t) = \pi^*(\boldsymbol{x}_t)\} - 1 \\
&= \Pr\{\widehat{\pi}_t(\boldsymbol{x}_t) = \pi^*(\boldsymbol{x}_t)\} - \kappa_t(\boldsymbol{x}_t),
\end{aligned}$$

where $\kappa_t(\boldsymbol{x}_t) = o_p(1)$ by Theorem 1 and $\Pr\{\widehat{\pi}_t(\boldsymbol{x}_t) = \pi^*(\boldsymbol{x}_t)\} \geq 1 - ct^{-\alpha\gamma}$ by Lemma B.2 as $t \to \infty$. Thus we have for large enough t,

$$\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\} > C_1, \tag{B.37}$$

for some constant $C_1 > 0$.

Then we focus on proving $\psi_6 = o_p(1)$ here. Define a class of function

$$\mathcal{F}_\kappa(\boldsymbol{x}, a, r) = \left\{ \left[\frac{1}{1-\kappa(\boldsymbol{x})} - \frac{\mathbb{I}\{a = \pi^*(\boldsymbol{x})\}}{\Pr\{a = \pi^*(\boldsymbol{x})\}} \right] e_t \mathbb{I}\{a = \pi(\boldsymbol{x})\} : \kappa_t \in \Lambda, \pi(\cdot) \in \Pi \right\},$$

where Π and Λ are two classes of functions that maps context $\boldsymbol{x} \in \mathcal{X}$ to a probability. Define the supremum of the empirical process indexed by $\mathcal{F}_\kappa$ as

$$||\mathbb{G}_n||_{\mathcal{F}} \equiv \quad \sup_{\pi \in \Pi} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) - \mathbb{E}\{\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) | \mathcal{H}_{t-1}\} \right]. \tag{B.38}$$

Firstly we notice that

$$\mathbb{E}\{\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) | \mathcal{H}_{t-1}\} = \mathbb{E}\left(\left[\frac{1}{1-\kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right] e_t \mathbb{I}\{a = \pi_t(\boldsymbol{x})\} \mid \mathcal{H}_{t-1} \right) = 0,$$

since $\mathbb{E}\left(e_t | \{a_t, \mathcal{H}_{t-1}\right) = 0$.
Secondly since $\kappa_t(\boldsymbol{x}_t) \leq C_2 < 1$ for sure for some constant $0 < C_2 < 1$ (by definition of a valid bandit algorithm an results for Theorem 1), by the Triangle inequality, we have

$$\left| \frac{1}{1-\kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right| \leq \left| \frac{1}{1-\kappa_t(\boldsymbol{x}_t)} - \right| + \left| \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right| \leq \frac{1}{1-C_2} + \frac{1}{C_1} \triangleq C.$$

Therefore, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left\{ \left(\frac{1}{\sqrt{T}} \left[\frac{1}{1-\kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right] e_t \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\} \right)^2 \mid \mathcal{H}_{t-1} \right\} \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left\{ \left(\frac{1}{\sqrt{T}} \left[\frac{1}{1-\kappa_t(\boldsymbol{x}_t)} - \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \right] e_t \mathbb{I}\{a_t = \widehat{\pi}_t(\boldsymbol{x}_t)\} \right)^2 \mid \mathcal{H}_{t-1} \right\} \\
&\leq \sum_{t=1}^{T} \frac{C^2}{T} \mathbb{E}\left\{ e_t^2 \mid \mathcal{H}_{t-1} \right\} \leq \sum_{t=1}^{T} \frac{C^2}{T} \max\{\sigma_0^2, \sigma_1^2\} = C^2 \max\{\sigma_0^2, \sigma_1^2\} < \infty.
\end{aligned}$$

Therefore, we have

$$d_1(f) \equiv \left\| \mathbb{E}\left(| \mathcal{F}_\kappa(\boldsymbol{x}_1, a_1, r_1) \| \mathcal{H}_0 \right) \right\|_\infty < \infty,$$

and

$$d_2(f) \equiv \left\| \mathbb{E}\left(\left(\mathcal{F}_\kappa(\boldsymbol{x}_1, a_1, r_1) \right)^2 \mid \mathcal{H}_0 \right) \right\|_\infty^{1/2} < \infty.$$

It follows from the maximal inequality developed in Section 4.2 of Dedecker and Louhichi (2002) that there exist some constant $K \geq 1$ such that

$$\mathbb{E}\Big[||\mathbb{G}_n||_{\mathcal{F}}\Big] \lesssim K\left(\sqrt{p}d_2(f) + \frac{1}{\sqrt{T}}\left\|\max_{1\leq t\leq T}|\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) - \mathbb{E}\left(\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) \mid \mathcal{H}_{t-1}\right)|\right\|\right)$$

The above right-hand-side is upper bounded by

$$\mathcal{O}(1)\sqrt{T^{-1/2}},$$

where $\mathcal{O}(1)$ denotes some universal constant. Hence, we have

$$\mathbb{E}\Big[||\mathbb{G}_n||_{\mathcal{F}}\Big] = \mathcal{O}_p(T^{-1/2}). \tag{B.39}$$

Combined with Equation (B.38), we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathcal{F}_\kappa(\boldsymbol{x}_t, a_t, r_t) = \mathcal{O}_p(T^{-1/2}).$$

and $\psi_6 = o_p(1)$.

Next, for ψ_7, by triangle inequality, we have

$$\begin{aligned}
|\psi_7| &= \left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}\Big[\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]\mathbb{I}\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\}\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}\right| \\
&\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left|\frac{1}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}\Big[\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\Big]\mathbb{I}\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\}\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}\right| \\
&\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left|\frac{1}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}\right|\left|\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right|\left|\mathbb{I}\{a_t \neq \widehat{\pi}_t(\boldsymbol{x}_t)\}\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}\right| \\
&\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{C_1}\left|\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right|.
\end{aligned}$$

Notice that $\psi_5 = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left|\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\}\right| = o_p(1)$, we have

$$|\psi_7| \leq= \frac{1}{C_1}\psi_5 = o_p(1).$$

Therefore, $\eta_6 = \psi_6 + \psi_7 = o_p(1)$. Hence, $\widetilde{V}_T = \overline{V}_T + o_p(T^{-1/2})$ hold.

Step 3: Then, to show the asymptotic normality of the proposed value estimator under DREAM, based on the above two steps, it is sufficient to show

$$\sqrt{T}(\widehat{V}_T - V^*) = \sqrt{T}(\overline{V}_T - V^*) + o_p(1) \xrightarrow{D} \mathcal{N}\left(0, \sigma_{DR}^2\right), \tag{B.40}$$

as $T \to \infty$, using Martingale Central Limit Theorem. By $r_t = \mu\{\boldsymbol{x}_t, a_t) + e_t$, we define

$$\begin{aligned} \xi_t &= \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}\Big[r_t - \mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\}\Big] + \mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} \\ &= \underbrace{\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}e_t}_{Z_t} + \underbrace{\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - V^*}_{W_t}. \end{aligned} \tag{B.41}$$

By (B.41), we have

$$\begin{aligned} \mathbb{E}\{Z_t \mid \mathcal{H}_{t-1}\} &= \mathbb{E}\{\mathbb{E}\left[Z_t \mid a_t\right] \mid \mathcal{H}_{t-1}\} = \mathbb{E}\left\{\mathbb{E}\left[\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}e_t \mid a_t, \boldsymbol{x}_t\right] \mid \mathcal{H}_{t-1}\right\} \\ &= \mathbb{E}\left\{\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}\mathbb{E}\left[e_t \mid a_t, \boldsymbol{x}_t\right] \mid \mathcal{H}_{t-1}\right\}. \end{aligned}$$

Since e_t is independent of $\mathcal{H}_{i-1}$ and $\boldsymbol{x}_i$ given a_t, and $\mathbb{E}\left\{e_t \mid a_t\right\} = 0$, we have

$$\mathbb{E}\{Z_t \mid \mathcal{H}_{t-1}\} = \mathbb{E}\left\{\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}\mathbb{E}\left\{e_t \mid a_t\right\} \mid \mathcal{H}_{t-1}\right\} = 0.$$

Notice that $\mathbb{E}[\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\}] = V^*$ by the definition, we have

$$\mathbb{E}\{W_t \mid \mathcal{H}_{t-1}\} = \mathbb{E}\left\{\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - V^*\right\} = 0.$$

Thus, we have $\mathbb{E}\{\xi_t \mid \mathcal{H}_{t-1}\} = 0$, which implies that $\{Z_t\}_{t=1}^T$, $\{W_t\}_{t=1}^T$ and $\{\xi_t\}_{t=1}^T$ are Martingale difference sequences. To prove Equation (B.40), it suffices to prove that $(1/\sqrt{T})\sum_{t=1}^T \xi_t \xrightarrow{D} \mathcal{N}(0, \sigma_{DR}^2)$, as $T \to \infty$, using Martingale Central Limit Theorem.

Firstly we calculate the conditional variance of ξ_t given $\mathcal{H}_{t-1}$. Note that

$$\mathbb{E}\left(Z_t^2 \mid \mathcal{H}_{t-1}\right) = \mathbb{E}\left\{\left(\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}}e_t\right)^2 \mid \mathcal{H}_{t-1}\right\} = \mathbb{E}\left(\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{[\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}]^2}e_t^2 \mid \mathcal{H}_{t-1}\right),$$

and

$$\begin{aligned}\mathbb{E}\left(Z_t^2 \mid \mathcal{H}_{t-1}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{[\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}]^2} e_t^2 \mid a_t, \boldsymbol{x}_t\right) \mid \mathcal{H}_{t-1}\right] \\ &= \mathbb{E}\left[\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{[\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}]^2} \mathbb{E}\left(e_t^2 \mid a_t, \boldsymbol{x}_t\right) \mid \mathcal{H}_{t-1}\right].\end{aligned}$$

Since e_t is independent of $\mathcal{H}_{i-1}$ and $\boldsymbol{x}_i$ given a_t, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(Z_t^2 \mid \mathcal{H}_{t-1}\right) &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{[\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}]^2}\mathbb{E}\left(e_t^2 \mid a_t\right) \mid \mathcal{H}_{t-1}\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{[\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}]^2}\sigma_{a_t}^2 \mid \mathcal{H}_{t-1}\right). \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{[\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}]^2}\sigma_{\pi^*(\boldsymbol{x}_t)}^2 \mathbb{E}\left(\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\} \mid \boldsymbol{x}_t, \mathcal{H}_{t-1}\right)\right].\end{aligned}$$

By the definition of Equation (B.25),

$$\nu_i\left(\boldsymbol{x}_i, \mathcal{H}_{i-1}\right) \equiv \Pr\left\{a_i \neq \pi^*(\boldsymbol{x}_i) \middle| \boldsymbol{x}_i, \mathcal{H}_{i-1}\right\} = \mathbb{E}\left[\mathbb{I}\left\{a_i \neq \pi^*(\boldsymbol{x}_i)\right\} | \boldsymbol{x}_i, \mathcal{H}_{i-1}\right],$$

we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(Z_t^2 \mid \mathcal{H}_{t-1}\right) &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{[\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}]^2}\sigma_{\pi^*(\boldsymbol{x}_t)}^2\left\{1 - \nu_t\left(\boldsymbol{x}, \mathcal{H}_{t-1}\right)\right\}\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\int\frac{\left\{1 - \nu_t\left(\boldsymbol{x}, \mathcal{H}_{t-1}\right)\right\}}{[\Pr\{a_t = \pi^*(\boldsymbol{x})\}]^2}\sigma_{\pi^*(\boldsymbol{x})}^2 dP_{\mathcal{X}} \\ &= \int\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\left\{1 - \nu_t\left(\boldsymbol{x}, \mathcal{H}_{t-1}\right)\right\}}{[\Pr\{a_t = \pi^*(\boldsymbol{x})\}]^2}\right]\sigma_{\pi^*(\boldsymbol{x})}^2 dP_{\mathcal{X}}.\end{aligned}$$

Similar as before, since $\lim_{i\to\infty}\Pr\{a_i \neq \pi^*(\boldsymbol{x})\} = \kappa_\infty(\boldsymbol{x})$, we have for any $\epsilon > 0$, there exist a constant $T_0 > 0$ such that $|\Pr\{a_i \neq \pi^*(\boldsymbol{x})\} - \kappa_\infty(\boldsymbol{x})| < \epsilon$ for all $i \geq T_0$.
We firstly consider the expectation of $(1/T)\sum_{t=1}^{T}\left\{1 - \nu_t\left(\boldsymbol{x}, \mathcal{H}_{t-1}\right)\right\}/[\Pr\{a_t = \pi^*(\boldsymbol{x})\}]^2$.

Note that $\Pr\{a_t = \pi^*(\boldsymbol{x})\}$ is not conditional on $\mathcal{H}_{t-1}$, thus we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right] = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2} = \frac{1}{T}\sum_{t=1}^{T}\frac{\Pr\{a_t=\pi^*(\boldsymbol{x})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}$$
$$= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}}.$$

Therefore by the triangle inequality we have

$$\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right] - \frac{1}{1-\kappa_\infty(\boldsymbol{x})}\right| = \left|\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}} - \frac{1}{1-\kappa_\infty(\boldsymbol{x})}\right|$$
$$= \left|\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}} - \frac{1}{1-\kappa_\infty(\boldsymbol{x})}\right]\right| = \left|\frac{1}{T}\sum_{t=1}^{T}\frac{1-\kappa_\infty(\boldsymbol{x})-\Pr\{a_t=\pi^*(\boldsymbol{x})\}}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}}\right|$$
$$= \left|\frac{1}{T}\sum_{t=1}^{T}\frac{\Pr\{a_t\neq\pi^*(\boldsymbol{x})-\kappa_\infty(\boldsymbol{x})\}}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}}\right| \leq \frac{1}{T}\sum_{t=1}^{T}\frac{|\Pr\{a_t\neq\pi^*(\boldsymbol{x})-\kappa_\infty(\boldsymbol{x})\}|}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}}$$
$$= \frac{1}{T}\sum_{t=1}^{T_0}\frac{|\Pr\{a_t\neq\pi^*(\boldsymbol{x})-\kappa_\infty(\boldsymbol{x})\}|}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}} + \frac{1}{T}\sum_{t=T_0}^{T}\frac{|\Pr\{a_t\neq\pi^*(\boldsymbol{x})-\kappa_\infty(\boldsymbol{x})\}|}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}}$$
$$< \frac{1}{T}\sum_{t=1}^{T_0}\frac{|\Pr\{a_t\neq\pi^*(\boldsymbol{x})\}|+|\kappa_\infty(\boldsymbol{x})|}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}} + \frac{1}{T}\sum_{t=T_0}^{T}\frac{\epsilon}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}}.$$

Since the above equation holds for any $\epsilon > 0$, we have

$$\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right] - \frac{1}{1-\kappa_\infty(\boldsymbol{x})}\right| \leq \frac{1}{T}\sum_{t=1}^{T_0}\frac{|\Pr\{a_t=\pi^*(\boldsymbol{x})\}|+|\kappa_\infty(\boldsymbol{x})|}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}}$$
$$\leq \frac{1}{T}\sum_{t=1}^{T_0}\frac{2}{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\{1-\kappa_\infty(\boldsymbol{x})\}}.$$

Since $\kappa_\infty(\boldsymbol{x}) \leq C_2 < 1$ for sure for some constant $0 < C_2 < 1$ (by definition of a valid bandit algorithm an results for Theorem 1), and by Equation (B.37), we have $\Pr\{a_t = \pi^*(\boldsymbol{x})\} > C_1$ for some constant $C_1 > 0$, therefore we have

$$\left|\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right] - \frac{1}{1-\kappa_\infty(\boldsymbol{x})}\right| \leq \frac{1}{T}\sum_{t=1}^{T_0}\frac{2}{C_1(1-C_2)} = \frac{2T_0}{C_1(1-C_2)T} \to 0,$$

as $T \to \infty$.
Then we consider the variance of $(1/T)\sum_{t=1}^{T}\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}/[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2$ over different histories. By Lemma B.1, we have

$$\mathrm{Var}\left[\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\right] \le \Pr\{a_t=\pi^*(\boldsymbol{x})\}\left[1-\Pr\{a_t=\pi^*(\boldsymbol{x})\}\right].$$

Therefore, we have

$$\begin{aligned}
&\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right] = \frac{1}{T^2}\mathrm{Var}\left[\sum_{t=1}^{T}\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right]\\
\le&\frac{1}{T}\sum_{t=1}^{T}\mathrm{Var}\left[\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right] = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathrm{Var}\left[\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}\right]}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^4} = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathrm{Var}\left[\{\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}\right]}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^4}\\
\le&\frac{1}{T}\sum_{t=1}^{T}\frac{\Pr\{a_t=\pi^*(\boldsymbol{x})\}\left[1-\Pr\{a_t=\pi^*(\boldsymbol{x})\}\right]}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^4} = \frac{1}{T}\sum_{t=1}^{T}\frac{1-\Pr\{a_t=\pi^*(\boldsymbol{x})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^3}\\
\le&\frac{1}{T}\sum_{t=1}^{T}\frac{1-\Pr\{a_t=\pi^*(\boldsymbol{x})\}}{[1-C_2]^3} = \frac{1}{[1-C_2]^3}\frac{1}{T}\sum_{t=1}^{T}\Pr\{a_t\neq\pi^*(\boldsymbol{x})\}.
\end{aligned}$$

Similarly as before, we could proof

$$\frac{1}{T}\sum_{t=1}^{T}\Pr\{a_t\neq\pi^*(\boldsymbol{x})\}\to\kappa_\infty(\boldsymbol{x}),$$

which follows

$$\mathrm{Var}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\right]\to\frac{1}{[1-C_2]^3}\kappa_\infty(\boldsymbol{x}).$$

Therefore, as T goes to infinity, we have

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\{1-\nu_t(\boldsymbol{x},\mathcal{H}_{t-1})\}}{[\Pr\{a_t=\pi^*(\boldsymbol{x})\}]^2}\to\frac{1}{1-\kappa_\infty(\boldsymbol{x})},$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(Z_t^2\mid\mathcal{H}_{t-1}\right)\to\int\frac{1}{1-\kappa_\infty(\boldsymbol{x})}\sigma^2_{\pi^*(\boldsymbol{x})}dP_{\mathcal{X}}.$$

Using the same technique of conditioning on a_t and $\boldsymbol{x}_t$, we have

$$\begin{aligned}\mathbb{E}\left(W_t Z_t \mid \mathcal{H}_{t-1}\right) &= \mathbb{E}\left\{\left(\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - V^*\right) \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} e_t \mid \mathcal{H}_{t-1}\right\} \\ &= \mathbb{E}\left\{\mathbb{E}\left[\left(\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - V^*\right) \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} e_t \mid a_t, \boldsymbol{x}_t\right] \mid \mathcal{H}_{t-1}\right\} \\ &= \mathbb{E}\left\{\left(\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} - V^*\right) \frac{\mathbb{I}\{a_t = \pi^*(\boldsymbol{x}_t)\}}{\Pr\{a_t = \pi^*(\boldsymbol{x}_t)\}} \mathbb{E}\left(e_t \mid a_t\right) \mid \mathcal{H}_{t-1}\right\} = 0.\end{aligned}$$

Thus, we further have

$$\mathbb{E}\left(\xi_t^2 \mid \mathcal{H}_{t-1}\right) = \mathbb{E}\left\{\left(Z_t + W_t\right)^2 \mid \mathcal{H}_{t-1}\right\} = \mathbb{E}\left(Z_t^2 \mid \mathcal{H}_{t-1}\right) + \mathbb{E}\left(W_t^2 \mid \mathcal{H}_{t-1}\right),$$

and

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left(\xi_t^2 \mid \mathcal{H}_{t-1}\right) = \int \frac{1}{1 - \kappa_\infty(\boldsymbol{x})} \sigma^2_{\pi^*(\boldsymbol{x})} dP_{\mathcal{X}} + \frac{1}{T}\sum_{t=1}^{T} \mathrm{Var}\left[\mu\{\boldsymbol{x}_t, \pi^*(\boldsymbol{x}_t)\} \mid \mathcal{H}_{t-1}\right].$$

Therefore as T goes to infinity, we have

$$\sum_{t=1}^{T} \mathbb{E}\left\{\left(\frac{1}{\sqrt{T}}\xi_t\right)^2 \mid \mathcal{H}_{t-1}\right\} \longrightarrow \sigma^2_{DR},$$

where

$$\sigma^2_{DR} = \int_{\boldsymbol{x}} \frac{\sigma_1^2 \mathbb{I}\{\mu(\boldsymbol{x}, 1) > \mu(\boldsymbol{x}, 0)\} + \sigma_0^2 \mathbb{I}\{\mu(\boldsymbol{x}, 1) < \mu(\boldsymbol{x}, 0)\}}{1 - \kappa_\infty(\boldsymbol{x})} dP_{\mathcal{X}} + \mathrm{Var}\left[\mu\{\boldsymbol{x}, \pi^*(\boldsymbol{x})\}\right]. \tag{B.42}$$

Then we check the conditional Lindeberg condition. For any $h > 0$, we have

$$\sum_{t=1}^{T} \mathbb{E}\left\{\left(\frac{1}{\sqrt{T}}\xi_t\right)^2 \mathbb{I}\left\{\left|\frac{1}{\sqrt{T}}\xi_t\right| > h\right\} \mid \mathcal{H}_{t-1}\right\} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left\{\xi_t^2 \mathbb{I}\left\{\xi_t^2 > Th^2\right\} \mid \mathcal{H}_{t-1}\right\}.$$

Since $\xi_t^2 \mathbb{I}\left\{\xi_t^2 > Th^2\right\}$ converges to zero as T goes to infinity and is dominated by ξ_t^2 given $\mathcal{H}_{t-1}$. Therefore, by Dominated Convergence Theorem, we conclude that

$$\sum_{t=1}^{T} \mathbb{E}\left\{\left(\frac{1}{\sqrt{T}}\xi_t\right)^2 \mathbb{I}\left\{\left|\frac{1}{\sqrt{T}}\xi_t\right| > h\right\} \mid \mathcal{H}_{t-1}\right\} \to 0, \text{ as } t \to \infty.$$

Thus the conditional Lindeberg condition is checked.

Next, recall the derived conditional variance in (B.42). By Martingale Central Limit Theorem, we have

$$(1/\sqrt{T})\sum_{t=1}^{T}\xi_t \xrightarrow{D} \mathcal{N}(0,\sigma_{DR}^2),$$

as $T\to\infty$. Hence, we complete the proof of Equation (B.40).

Step 4: Finally, to show the variance estimator in Equation (5) in the main paper is a consistent estimator of σ_{DR}^2. Recall that the variance estimator is

$$\begin{aligned}\widehat{\sigma}_T^2 = &\underbrace{\frac{1}{T}\sum_{t=1}^{T}\frac{\widehat{\sigma}_{1,t-1}^2(\boldsymbol{x}_t,1)\mathbb{I}\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)>\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)\}+\widehat{\sigma}_{0,t-1}^2\mathbb{I}\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)<\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)\}}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)}}_{\widehat{\sigma}_{T,1}^2}\\ &+\underbrace{\frac{1}{T}\sum_{t=1}^{T}\left[\widehat{\mu}_T\{\boldsymbol{x}_t,\widehat{\pi}_T(\boldsymbol{x}_t)\}-\frac{1}{T}\sum_{t=1}^{T}\widehat{\mu}_T\{\boldsymbol{x}_t,\widehat{\pi}_T(\boldsymbol{x}_t)\}\right]^2}_{\widehat{\sigma}_{T,2}^2}.\end{aligned} \tag{B.43}$$

Firstly we proof the first line of the above Equation (B.43) is a consistent estimator for

$$\int_{\boldsymbol{x}}\frac{\sigma_1^2\mathbb{I}\{\mu(\boldsymbol{x},1)>\mu(\boldsymbol{x},0)\}+\sigma_0^2\mathbb{I}\{\mu(\boldsymbol{x},1)<\mu(\boldsymbol{x},0)\}}{1-\kappa_\infty(\boldsymbol{x})}dP_{\mathcal{X}}.$$

Recall that we denote $\widehat{\Delta}_{\boldsymbol{x}_t}=\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)-\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)$, thus we can rewrite $\widehat{\sigma}_T^2$ as

$$\begin{aligned}\widehat{\sigma}_{T,1}^2 &= \frac{1}{T}\sum_{t=1}^{T}\frac{\widehat{\sigma}_{1,t-1}^2\mathbb{I}\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)>\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)\}+\widehat{\sigma}_{0,t-1}^2(\boldsymbol{x}_t,0)\mathbb{I}\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t,1)<\widehat{\mu}_{t-1}(\boldsymbol{x}_t,0)\}}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)}\\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{\widehat{\sigma}_{1,t-1}^2\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}+\widehat{\sigma}_{0,t-1}^2\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}<0\}}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)}.\end{aligned}$$

We decompose the proposed variance estimator by

$$\begin{aligned}
\widehat{\sigma}^2_{T,1} = & \frac{1}{T}\sum_{t=1}^{T} \frac{\left\{\widehat{\sigma}^2_{1,t-1} - \sigma_1^2\right\}\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} > 0\} + \left\{\widehat{\sigma}^2_{0,t-1} - \sigma_0^2\right\}\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} < 0\}}{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)} \\
& + \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_1^2\left(\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} > 0\} - \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\}\right) + \sigma_0^2\left(\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} < 0\} - \mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}\right)}{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)} \\
& + \frac{1}{T}\sum_{t=1}^{T} \left(\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} + \sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}\right)\left\{\frac{1}{1 - \widehat{\kappa}_t(\boldsymbol{x}_t)} - \frac{1}{1 - \kappa_t(\boldsymbol{x}_t)}\right\} \\
& + \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} + \sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}}{1 - \kappa_t(\boldsymbol{x}_t)}.
\end{aligned}$$

Our goal is to prove that the first three lines are all $o_p(1)$.

Firstly, recall that

$$\begin{aligned}
\widehat{\sigma}^2_{a,t} &= \{\sum_{i=1}^{t}\mathbb{I}(a_i = a) - d\}^{-1}\sum_{a_i=a}^{1\le i\le t}[\widehat{\mu}_i\{\boldsymbol{x}_i, a_i\} - r_i]^2 \\
&= \{\sum_{i=1}^{t}\mathbb{I}(a_i = a) - d\}^{-1}\sum_{a_i=a}^{1\le i\le t}[\boldsymbol{x}_i^\top\left\{\widehat{\boldsymbol{\beta}}_{i-1}(a) - \boldsymbol{\beta}_{i-1}(a)\right\} - e_i]^2.
\end{aligned}$$

By Lemma 4.1, we have $\|\widehat{\boldsymbol{\beta}}_{i-1}(a) - \boldsymbol{\beta}_{i-1}(a)\|_1 = o_p(1)$. Under Assumption 4.1, we have $\boldsymbol{x}_i^\top\left\{\widehat{\boldsymbol{\beta}}_{i-1}(a) - \boldsymbol{\beta}_{i-1}(a)\right\} = o_p(1)$. Thus by Lemma 6 in Luedtke and Van Der Laan (2016), we have

$$\widehat{\sigma}^2_{a,t} = \{\sum_{i=1}^{t}\mathbb{I}(a_i = a) - d\}^{-1}\sum_{a_i=a}^{1\le i\le t} e_i^2 + o_p(1).$$

Since e_i is i.i.d conditional on a_i, and $\mathbb{E}(e_i^2|a_i = a) = \sigma_a^2$, noting that

$$\lim_{t\to\infty}\frac{\sum_{i=1}^{t}\mathbb{I}(a_i = a)}{\sum_{i=1}^{t}\mathbb{I}(a_i = a) - d} = 1,$$

by Law of Large Numbers we have

$$\widehat{\sigma}^2_{a,t} = \sigma_a^2 + o_p(1).$$

Therefore, the first line is $o_p(1)$.

Secondly, denote the second line as

$$\begin{aligned}\psi_8 &= \frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_1^2\left(\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right)+\sigma_0^2\left(\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}<0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}\right)}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)}\\ &= \frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_1^2\left(\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right)+\sigma_0^2\left(\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}\right)}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)}\\ &= \frac{\sigma_1^2-\sigma_0^2}{T}\sum_{t=1}^{T}\frac{\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}}{1-\widehat{\kappa}_t(\boldsymbol{x}_t)}.\end{aligned}$$

Since $\kappa_t(\boldsymbol{x}_t)\leq C_2<1$ for sure for some constant $0<C_2<1$ (by definition of a valid bandit algorithm an results for Theorem 1), by the triangle inequality, we have

$$\begin{aligned}|\psi_8| &\leq \frac{\sigma_1^2-\sigma_0^2}{T}\sum_{t=1}^{T}\frac{\left|\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right|}{|1-\widehat{\kappa}_t(\boldsymbol{x}_t)|} \leq \frac{\sigma_1^2-\sigma_0^2}{T}\sum_{t=1}^{T}\frac{\left|\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right|}{1-C_2}\\ &\leq \frac{\sigma_1^2-\sigma_0^2}{1-C_2}\frac{1}{T}\sum_{t=1}^{T}\left|\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right|.\end{aligned}$$

Since $\Pr\left(\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}=0\right)=1-ct^{-\alpha\gamma}$ by Lemma B.2, there exists some constant c such that

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left|\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right|\neq 0\right)\leq\sum_{t=1}^{T}\Pr\left(\left|\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right|\neq 0\right)\leq\sum_{t=1}^{T}ct^{-\alpha\gamma}.$$

By Lemma 6 in Luedtke and Van Der Laan (2016), we have $\sum_{t=1}^{T}ct^{-\alpha\gamma}=cT^{-\alpha\gamma}$, thus

$$\Pr\left(|\psi_8|\neq 0\right)=\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left|\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t}>0\}-\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}\right|\neq 0\right)\leq cT^{-\alpha\gamma},$$

which follows $\Pr\left(|\psi_8|\neq 0\right)=o_p(1)$.

Lastly, under the assumption that $\widehat{\kappa}_t(\boldsymbol{x}_t)$ is a consistent estimator for $\kappa_t(\boldsymbol{x}_t)$, we have the third line is $o_p(1)$ by the continuous mapping theorem.

Given the above results, we have

$$\widehat{\sigma}_{T,1}^2 = \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_1^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} + \sigma_0^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}}{1 - \kappa_t(\boldsymbol{x}_t)} + o_p(1) = \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_1^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} + \sigma_0^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}}{1 - \Pr\left\{a_t \neq \widehat{\pi}_t\left(x_t\right)\right\}} + o_p(1).$$

Thus, we can further express $\widehat{\sigma}_{T,1}^2$ as

$$\begin{aligned}\widehat{\sigma}_{T,1}^2 = &\frac{1}{T}\sum_{t=1}^{T} \left(\sigma_1^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} + \sigma_0^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}\right) \left(\frac{1}{1 - \Pr\left\{a_t \neq \widehat{\pi}_t\left(x\right)\right\}} - \frac{1}{1 - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}\right)\right\}}\right) \\ &+ \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_1^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} + \sigma_0^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}}{1 - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}} + o_p(1).\end{aligned} \tag{B.44}$$

Notice that

$$\left(\frac{1}{1 - \Pr\left\{a_t \neq \widehat{\pi}_t\left(x\right)\right\}} - \frac{1}{1 - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}\right)\right\}}\right) = \frac{\Pr\left\{a_t \neq \widehat{\pi}_t\left(x\right)\right\} - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}}{\left(1 - \Pr\left\{a_t \neq \widehat{\pi}_t\left(x\right)\right\}\right)\left(1 - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}\right)\right\}\right)},$$

where

$$\Pr\left\{a_t \neq \widehat{\pi}_t\left(x\right)\right\} - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\} = \mathbb{E}\left(\mathbb{I}\{a_t \neq \widehat{\pi}_t\left(x\right)\} - \mathbb{I}\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\}\right) = \mathbb{E}\left(\mathbb{I}\{\widehat{\pi}_t\left(x\right) \neq \pi^*\left(\boldsymbol{x}_t\right)\}\right).$$

We also note that by Lemma B.2, there exists some constant c and $0 < \alpha < \frac{1}{2}$ such that $\alpha\gamma < \frac{1}{2}$ and

$$\mathbb{E}\left(\mathbb{I}\{\widehat{\pi}_t\left(x\right) = \pi^*\left(\boldsymbol{x}_t\right)\}\right) = \Pr\left(\widehat{\pi}_t\left(x\right) = \pi^*\left(\boldsymbol{x}_t\right)\right\}) \leq ct^{-\alpha\gamma},$$

therefore we have the result that

$$\Pr\left\{a_t \neq \widehat{\pi}_t\left(x\right)\right\} - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\} = \mathbb{E}\left(\mathbb{I}\{\widehat{\pi}_t\left(x\right) \neq \pi^*\left(\boldsymbol{x}_t\right)\}\right) = o_p(1).$$

Thus, Equation (B.44) can be expressed as

$$\begin{aligned}
\widehat{\sigma}^2_{T,1} &= \frac{1}{T}\sum_{t=1}^{T}\left(\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}\right)o_p(1)+\frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}}{1-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}}+o_p(1)\\
&=\frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}}{1-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}}+o_p(1)\\
&=\underbrace{\frac{1}{T}\sum_{t=1}^{T}\left\{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}\right\}\left[\frac{1}{1-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}}-\frac{1}{1-\kappa_\infty(\boldsymbol{x}_t)}\right]}_{\psi_9}\\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\frac{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}}{1-\kappa_\infty(\boldsymbol{x}_t)}+o_p(1).
\end{aligned}$$

Note that

$$\begin{aligned}
\psi_9 &= \frac{1}{T}\sum_{t=1}^{T}\left\{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}\right\}\left[\frac{1}{1-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}}-\frac{1}{1-\kappa_\infty(\boldsymbol{x}_t)}\right]\\
&=\frac{1}{T}\sum_{t=1}^{T}\left\{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}\right\}\frac{\kappa_\infty(\boldsymbol{x}_t)-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}}{\left[1-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}\right]\left[1-\kappa_\infty(\boldsymbol{x}_t)\right]},
\end{aligned}$$

which follows

$$|\psi_9|\leq\frac{1}{T}\sum_{t=1}^{T}\left\{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}\right\}\frac{\left|\kappa_\infty(\boldsymbol{x}_t)-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}\right|}{\left[1-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}\right]\left[1-\kappa_\infty(\boldsymbol{x}_t)\right]}.$$

By Equation (B.37), for large enough t, there exists some constant $C_1>0$ such that

$$\Pr\{a_t\neq\pi^*(\boldsymbol{x}_t)\}<C_1.$$

Since $\kappa_t(\boldsymbol{x}_t)\leq C_2<1$ for sure for some constant $0<C_2<1$ (by the definition of a valid bandit algorithm as results shown in Theorem 1), we also have

$$\kappa_\infty(\boldsymbol{x}_t)<C_2.$$

Therefore,

$$\frac{\sigma_1^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}>0\}+\sigma_0^2\mathbb{I}\{\Delta_{\boldsymbol{x}_t}<0\}}{\left[1-\Pr\left\{a_t\neq\pi^*\left(\boldsymbol{x}_t\right)\right\}\right]\left[1-\kappa_\infty(\boldsymbol{x}_t)\right]}<\frac{\max\{\sigma_0^2,\sigma_1^2\}}{(1-C_1)(1-C_2)}\triangleq C,$$

which follows immediately that

$$|\psi_9| \leq \frac{1}{T}\sum_{t=1}^{T} C\left|\kappa_\infty(\boldsymbol{x}_t) - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}\right|.$$

Since $\lim_{t\to\infty}\Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}\right)\right\} = \kappa_\infty(\boldsymbol{x})$ for any $\boldsymbol{x}$, therefore for any $\epsilon > 0$, there exists some constant T_0, such that $\left|\Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}\right)\right\} - \kappa_t(\boldsymbol{x})\right| < \epsilon$ for any $\boldsymbol{x}$ with $t \geq T_0$, thus we have

$$\begin{aligned}
|\psi_9| &\leq \frac{1}{T}\sum_{t=1}^{T} C\left|\kappa_\infty(\boldsymbol{x}_t) - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}\right| \\
&= \frac{1}{T}\sum_{t=1}^{T_0} C\left|\kappa_\infty(\boldsymbol{x}_t) - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}\right| + \frac{1}{T}\sum_{t=T_0}^{T} C\left|\kappa_\infty(\boldsymbol{x}_t) - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}\right| \\
&< \frac{1}{T}\sum_{t=1}^{T_0} C\left|\kappa_\infty(\boldsymbol{x}_t) - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}\right| + \frac{1}{T}\sum_{t=T_0}^{T} C\epsilon.
\end{aligned}$$

Note that by the triangle inequality,

$$\left|\kappa_\infty(\boldsymbol{x}_t) - \Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}\right| \leq \left|\kappa_\infty(\boldsymbol{x}_t)\right| + \left|\Pr\left\{a_t \neq \pi^*\left(\boldsymbol{x}_t\right)\right\}\right| < C_1 + C_2,$$

thus we have

$$|\psi_9| < \frac{1}{T}\sum_{t=1}^{T_0} C\left(C_1 + C_2\right) + \frac{1}{T}\sum_{t=T_0}^{T} C\epsilon = \frac{T_0 C\left(C_1 + C_2\right)}{T} + \frac{T - T0}{T} C\epsilon.$$

Since the above equation holds for any $\epsilon > 0$, we have

$$|\psi_9| \leq \frac{1}{T}\sum_{t=1}^{T_0} C\left(C_1 + C_2\right) + \frac{1}{T}\sum_{t=T_0}^{T} C\epsilon = \frac{T_0 C\left(C_1 + C_2\right)}{T} = \mathcal{O}(\frac{1}{T}),$$

which follows $\psi_9 = 0_p(1)$. Therefore, we have

$$\widehat{\sigma}^2_{T,1} = \frac{1}{T}\sum_{t=1}^{T} \frac{\sigma_1^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} + \sigma_0^2 \mathbb{I}\{\Delta_{\boldsymbol{x}_t} < 0\}}{1 - \kappa_\infty(\boldsymbol{x}_t)} + o_p(1).$$

By Law of Large Numbers, we further have

$$\widehat{\sigma}^2_{T,1} = \int_{\boldsymbol{x}} \frac{\sigma_1^2 \mathbb{I}\{\mu(\boldsymbol{x},1) > \mu(\boldsymbol{x},0)\} + \sigma_0^2 \mathbb{I}\{\mu(\boldsymbol{x},1) < \mu(\boldsymbol{x},0)\}}{1-\kappa_\infty(\boldsymbol{x})} dP_{\mathcal{X}} + o_p(1).$$

Next, we proof the second line of Equation (B.43) is a consistent estimator for $\text{Var}\left[\mu\{\boldsymbol{x}, \pi^*(\boldsymbol{x})\}\right]$. By Central Limit Theorem and Continuous Mapping Theorem, we have

$$\widehat{\sigma}^2_{T,2} = \frac{1}{T}\sum_{t=1}^{T}\left[\widehat{\mu}_T\{\boldsymbol{x}_t, \widehat{\pi}_T(\boldsymbol{x}_t)\} - \frac{1}{T}\sum_{t=1}^{T}\widehat{\mu}_T\{\boldsymbol{x}_t, \widehat{\pi}_T(\boldsymbol{x}_t)\}\right]^2.$$

Since $\boldsymbol{x}_t$ are i.i.d, $\widehat{\mu}_T\{\boldsymbol{x}_t, \widehat{\pi}_T(\boldsymbol{x}_t)\}$ are i.i.d as well. Thus by the Law of Large Numbers, we have

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{\mu}_T\{\boldsymbol{x}_t, \widehat{\pi}_T(\boldsymbol{x}_t)\} = \mathbb{E}\left[\widehat{\mu}_T\{\boldsymbol{x}, \widehat{\pi}_T(\boldsymbol{x})\}\right] + o_p(1),$$

and

$$\widehat{\sigma}^2_{T,2} = \text{Var}\left[\widehat{\mu}_T\{\boldsymbol{x}, \widehat{\pi}_T(\boldsymbol{x})\}\right] + o_p(1).$$

Note that

$$\widehat{\mu}_T\{\boldsymbol{x}, \widehat{\pi}_T(\boldsymbol{x})\} = \mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}} > 0\}\widehat{\mu}_T\{\boldsymbol{x},1\} + \left[1 - \mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}} > 0\}\right]\widehat{\mu}_T\{\boldsymbol{x},0\},$$

since $\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}} > 0\} - \mathbb{I}\{\Delta_{\boldsymbol{x}} > 0\} = o_p(1)$ and the fact that $\widehat{\mu}_T\{\boldsymbol{x},0\}$ and $\widehat{\mu}_T\{\boldsymbol{x},1\}$ are consistent, we have

$$\widehat{\mu}_T\{\boldsymbol{x}, \widehat{\pi}_T(\boldsymbol{x})\} = \mathbb{I}\{\Delta_{\boldsymbol{x}} > 0\}\mu\{\boldsymbol{x},1\} + \left[1 - \mathbb{I}\{\Delta_{\boldsymbol{x}} > 0\}\right]\mu\{\boldsymbol{x},0\} = \mu\{\boldsymbol{x}, \pi(\boldsymbol{x})\} + o_p(1).$$

Therefore

$$\widehat{\sigma}^2_{T,2} = \text{Var}\left[\mu\{\boldsymbol{x}, \pi(\boldsymbol{x})\} + o_p(1)\right] = \text{Var}\left[\mu\{\boldsymbol{x}, \pi(\boldsymbol{x})\}\right] + o_p(1).$$

by Continuous Mapping Theorem. The proof for Theorem 3 is thus completed.

B.6 Results and Proof for Auxiliary Lemmas

Lemma B.1 *Suppose a random variable X is restricted to $[a,b]$ and $\mu = E[X]$, then the variance of X is bounded by $(b-\mu)(\mu-a)$.*

Proof: Firstly consider the case that $a = 0, b = 1$. Notice that we have $E[X^2] \leq E[X]$ since for all $x \in [0,1], x^2 \leq x$. Therefore,

$$\text{Var}[X] = E\left[X^2\right] - \left(E[X]^2\right) = E\left[X^2\right] - \mu^2 \leq \mu - \mu^2 = \mu(1-\mu).$$

Then we consider general interval $[a,b]$. Define $Y = \frac{X-a}{b-a}$, which is restricted in $[0,1]$. Equivalently, $X = (b-a)Y + a$, which follows immediate that

$$\text{Var}[X] = (b-a)^2 \text{Var}[Y] \leq (b-a)^2 \mu_X (1-\mu_Y),$$

where the inequality is based on the first result. Now, by substituting $\mu_Y = \frac{\mu_X - a}{b-a}$, the bound equals

$$(b-a)^2 \frac{\mu_X - a}{b-a}\left(1 - \frac{\mu_X - a}{b-a}\right) = (b-a)^2 \frac{\mu_X - a}{b-a}\frac{b-\mu_X}{b-a} = (\mu_X - a)(b-\mu_X),$$

which is the desired result.

Lemma B.2 *Suppose the conditions in Theorem 2 hold with Assumption 4.3, then there exists some constant c and $0 < \alpha < \frac{1}{2}$ such that $\alpha\gamma < \frac{1}{2}$ and $Pr(\widehat{\pi}_t(x) \neq \pi^*\{\boldsymbol{x}_t)\} \geq 1 - ct^{-\alpha\gamma}$.*

Proof: By Theorem 2, we have $\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t} = \mathcal{O}_p(t^{-\frac{1}{2}})$, thus $\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} > 0\} = \mathbb{I}\{\Delta_{\boldsymbol{x}_t} + \mathcal{O}_p(t^{-\frac{1}{2}}) > 0\}$.
By Assumption 4.3, there exists some constant c and $0 < \alpha < \frac{1}{2}$ such that $\alpha\gamma < \frac{1}{2}$ and

$$\Pr\{0 < |\Delta_{\boldsymbol{x}_t}| < t^{-\alpha}\} \leq ct^{-\alpha\gamma}.$$

Thus, with probability greater than $1 - ct^{-\alpha\gamma}$, we have $|\Delta_{\boldsymbol{x}_t}| > t^{-\alpha}$, which further implies $\mathbb{I}\{\Delta_{\boldsymbol{x}_t} + \mathcal{O}_p(t^{-\frac{1}{2}}) > 0\} = \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\}$. In other words, $\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} > 0\} - \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} = 0$ with probability greater than $1 - ct^{-\alpha\gamma}$, which convergences to 1 as $t \to \infty$. Therefore, as $t \to \infty$, we have

$$\Pr\left(\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} > 0\} - \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > 0\} = 0\right) \geq 1 - ct^{-\alpha\gamma},$$

i.e.,

$$\Pr\left(\widehat{\pi}_t(x) = \pi^*\{\boldsymbol{x}_t)\right\} \geq 1 - ct^{-\alpha\gamma}.$$

Lemma B.3 *Suppose conditions in Lemma 4.1 hold. Assuming Assumptions 4.3 with* $tp_t^2 \to \infty$ *as* $t \to \infty$*, we have*

$$T^{-1/2} \sum_{t=1}^{T} |\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \pi^*\}| = o_p(1).$$

Proof: Without loss of generality, suppose $\widehat{\pi}_t(\boldsymbol{x}_t) = \mathbb{I}\{\boldsymbol{x}_t^\top \boldsymbol{\beta}(1) > \boldsymbol{x}_t^\top \boldsymbol{\beta}(0)\}$. Since $\mu(\boldsymbol{x}_t, a) = \boldsymbol{x}_t^\top \boldsymbol{\beta}(a)$, we have

$$\begin{aligned} \mu(\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t) &= \mathbb{I}\{\boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(1) > \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(0)\} \boldsymbol{x}_t^\top \boldsymbol{\beta}(1) + \mathbb{I}\{\boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(1) \le \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(0)\} \boldsymbol{x}_t^\top \boldsymbol{\beta}(0) \\ &= \mathbb{I}\{\boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(1) > \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(0)\} \boldsymbol{x}_t^\top \{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\} + \boldsymbol{x}_t^\top \boldsymbol{\beta}(0). \end{aligned} \tag{B.45}$$

Similarly to (B.45), we have

$$\mu(\boldsymbol{x}_t, \pi^*) = \mathbb{I}\{\boldsymbol{x}_t^\top \boldsymbol{\beta}(1) > \boldsymbol{x}_t^\top \boldsymbol{\beta}(0)\} \boldsymbol{x}_t^\top \{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\} + \boldsymbol{x}_t^\top \boldsymbol{\beta}(0). \tag{B.46}$$

Combining (B.45) and (B.46), we have

$$\mu(\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t) - \mu(\boldsymbol{x}_t, \pi^*) = \left[\mathbb{I}\{\boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(1) > \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(0)\} - \mathbb{I}\{\boldsymbol{x}_t^\top \boldsymbol{\beta}(1) > \boldsymbol{x}_t^\top \boldsymbol{\beta}(0)\}\right] \boldsymbol{x}_t^\top \{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\}. \tag{B.47}$$

Since $\mathbb{I}\{\boldsymbol{x}_t^\top \boldsymbol{\beta}(1) > \boldsymbol{x}_t^\top \boldsymbol{\beta}(0)\} = 1$ by assumption, (B.47) can be simplified as

$$\mu(\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t) - \mu(\boldsymbol{x}_t, \pi^*) = -\mathbb{I}\{\boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(1) - \boldsymbol{x}_t^\top \widehat{\boldsymbol{\beta}}_t(0) \le 0\} \boldsymbol{x}_t^\top \{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\} = -\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \le 0\} \Delta_{\boldsymbol{x}_t} \le 0,$$

where $\Delta_{\boldsymbol{x}_t} = \boldsymbol{x}_t^\top \{\boldsymbol{\beta}(1) - \boldsymbol{\beta}(0)\}$ and $\widehat{\Delta}_{\boldsymbol{x}_t} = \boldsymbol{x}_t^\top \{\widehat{\boldsymbol{\beta}}(1) - \widehat{\boldsymbol{\beta}}(0)\}$. Thus, we have

$$T^{-1/2} \sum_{t=1}^{T} |\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \pi^*\}| = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \le 0\} \Delta_{\boldsymbol{x}_t}.$$

To show $T^{-1/2} \sum_{t=1}^{T} |\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \pi^*\}| = o_p(1)$, it suffices to show that $(1/\sqrt{T}) \sum_{t=1}^{T} \mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \le 0\} \Delta_{\boldsymbol{x}_t}$ has an upper bound. Since $\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \le 0\} \widehat{\Delta}_{\boldsymbol{x}_t} \le 0$, it suffices to show

$$\zeta = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \le 0\} (\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})$$

has a lower bound. We further notice that for any $\alpha > 0$,

$$\begin{aligned}\zeta = &\underbrace{\Pr\{0 < \Delta_{\boldsymbol{x}_t} < T^{-\alpha}\}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{I}\{0 < \Delta_{\boldsymbol{x}_t} < T^{-\alpha}\}\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})}_{\zeta_1} \\ &+ \underbrace{\Pr\{\Delta_{\boldsymbol{x}_t} > T^{-\alpha}\}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{I}\{\Delta_{\boldsymbol{x}_t} > T^{-\alpha}\}\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})}_{\zeta_2}.\end{aligned} \tag{B.48}$$

To show ζ has a lower bound, it is sufficient to show $\zeta_1 = o_p(1)$ and $\zeta_2 = o_p(1)$ correspondingly.

Firstly, we are going to show $\zeta_1 = o_p(1)$. By Theorem 2, $\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t} = \mathcal{O}_p(t^{-\frac{1}{2}})$, which implies

$$\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t} = o_p\{t^{-(\frac{1}{2}-\alpha\gamma)}\}.$$

Thus we have

$$\begin{aligned}&|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{I}\{0 < \Delta_{\boldsymbol{x}_t} < T^{-\alpha}\}\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})| \\ &\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}|\mathbb{I}\{0 < \Delta_{\boldsymbol{x}_t} < T^{-\alpha}\}||\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}||(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})| \\ &\leq \frac{1}{\sqrt{T}}\sum_{t=1}^{T}|(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})| \leq \frac{\sqrt{T}}{T}\sum_{t=1}^{T}o_p(t^{-(\frac{1}{2}-\alpha\gamma)}) \overset{*}{=} \sqrt{T}o_p(T^{-(\frac{1}{2}-\alpha\gamma)}) = o_p(T^{\alpha\gamma}),\end{aligned}$$

where the equation (*) is derived by Lemma 6 in Luedtke and Van Der Laan (2016). By Assumption 4.3, there exists some constant c and $0 < \alpha < \frac{1}{2}$ such that $\alpha\gamma < \frac{1}{2}$ and

$$\Pr\{0 < \Delta_{\boldsymbol{x}_t} < T^{-\alpha}\} \leq cT^{-\alpha\gamma}.$$

Therefore we have

$$|\zeta_1| \leq \Pr\{0 < \Delta_{\boldsymbol{x}_t} < T^{-\alpha}\}o_p(T^{-(\frac{1}{2}-\alpha\gamma)}) = cT^{-\alpha\gamma}o_p(T^{\alpha\gamma}) = o_p(1). \tag{B.49}$$

Notice that

$$\mathbb{I}\{0 < \Delta_{\boldsymbol{x}_t} < T^{-\alpha}\}\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}) \leq \mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0, \tag{B.50}$$

where the first inequality holds since $\Delta_{\boldsymbol{x}_t} \geq 0$. Combining (B.48), (B.49), and (B.50), we have

$$0 \geq \zeta_1 = o_p(1). \tag{B.51}$$

Next, we consider the second part ζ_2. Note that

$$\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\} = \mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t} \leq -\Delta_{\boldsymbol{x}_t}\} = \mathbb{I}\{|\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}| > \Delta_{\boldsymbol{x}_t}\},$$

we have

$$\left|\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})\right| = \mathbb{I}\{|\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}| > \Delta_{\boldsymbol{x}_t}\}|\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}| \leq \frac{|\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}|^2}{\Delta_{\boldsymbol{x}_t}}, \tag{B.52}$$

where the inequality holds since

$$\mathbb{I}\{|\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}| > \Delta_{\boldsymbol{x}_t}\} \leq \frac{|\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}|}{\Delta_{\boldsymbol{x}_t}}.$$

Thus, by (B.52), we further have

$$\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}) \geq -\frac{(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})^2}{\Delta_{\boldsymbol{x}_t}}. \tag{B.53}$$

Since $\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t} < 0$, based on (B.53), we have

$$0 \geq \zeta_2 \geq \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > T^{-\alpha}\}\mathbb{I}\{\widehat{\Delta}_{\boldsymbol{x}_t} \leq 0\}(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t}) \geq -\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{I}\{\Delta_{\boldsymbol{x}_t} > T^{-\alpha}\} \frac{(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})^2}{\Delta_{\boldsymbol{x}_t}}. \tag{B.54}$$

Notice that $\mathbb{I}\{\Delta_{\boldsymbol{x}_t} > T^{-\alpha}\} \leq \Delta_{\boldsymbol{x}_t} T^{\alpha}$, combining with (B.54), we further have

$$0 \geq \zeta_2 \geq -\frac{1}{\Delta_{\boldsymbol{x}_t}} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\Delta_{\boldsymbol{x}_t}}{T^{-\alpha}} (\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})^2 = -T^{-\frac{1}{2}+\alpha} \sum_{t=1}^{T} (\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})^2 = -T^{\frac{1}{2}+\alpha} \left\{ \frac{1}{T} \sum_{t=1}^{T} (\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})^2 \right\}.$$

By Theorem 2, $\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t} = \mathcal{O}_p(t^{-\frac{1}{2}})$, which implies $(\widehat{\Delta}_{\boldsymbol{x}_t} - \Delta_{\boldsymbol{x}_t})^2 = o_p(T^{-(\frac{1}{2}+\alpha)})$.And by Lemma 6 in Luedtke and Van Der Laan (2016), $T^{-1} \sum_{t=1}^{T} o_p\{t^{-(\frac{1}{2}+\alpha)}\} = o_p\{T^{-(\frac{1}{2}+\alpha)}\}$, we have

$$0 \geq \zeta_2 \geq -T^{\frac{1}{2}+\alpha} o_p(T^{-(\frac{1}{2}+\alpha)}) = o_p(1). \tag{B.55}$$

Therefore, combining Equation (B.51) and Equation (B.55), we have

$$0 \geq \zeta = \zeta_1 + \zeta_2 = o_p(1).$$

Thus, we have

$$T^{-1/2} \sum_{t=1}^{T} |\mu\{\boldsymbol{x}_t, \widehat{\pi}_t(\boldsymbol{x}_t)\} - \mu\{\boldsymbol{x}_t, \pi^*\}| = o_p(1).$$

References

Chen, H., Lu, W. and Song, R. (2020), 'Statistical inference for online decision making: In a contextual bandit setting', *Journal of the American Statistical Association* pp. 1–16.

Dedecker, J. and Louhichi, S. (2002), Maximal inequalities and empirical central limit theorems, *in* 'Empirical process techniques for dependent data', Springer, pp. 137–159.

Feller, W. (2008), *An introduction to probability theory and its applications, vol 2*, John Wiley & Sons.

Hall, P. and Heyde, C. C. (2014), *Martingale limit theory and its application*, Academic press.

Luedtke, A. R. and Van Der Laan, M. J. (2016), 'Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy', *Annals of statistics* **44**(2), 713.