Electronic Companion

In this supplement, we include experimental details for Figure 3 in Section EC. 1, extend our framework to MDPs in Section EC. 2, provide additional results in Section EC. 3, briefly describe the datasets and the pretrained model in Section EC. 4, and provide detailed proofs of the theoretical results, including Lemma 1, Theorem 1, Theorem 2, Theorem 3 and Theorem 4 in Section EC. 5. Supporting lemmas are included in Section EC. 6.

EC. 1. Experimental Details for Figure 3

We consider 4 actions with $\phi(x,a_1) = (0.2,0,0.1)^\top, \phi(x,a_2) = (0.1,-0.9,0.1)^\top, \phi(x,a_3) = (0.2,0.1,-0.1)^\top$ and $\phi(x,a_4) = (0,0.1,0)^\top$. The true reward parameter is $\theta_* = (-1,0.1,1)^\top$. The optimal action is $a_4 = \arg\max_{a\in\{a_1,a_2,a_3,a_4\}} \theta_*^\top \phi(x,a)$. For the experiment, T actions are randomly selected from a_1, a_2 and a_3 with probabilities $0.45, 0.45, 0.1$ to estimate θ_* as $\hat{\theta}$, intentionally excluding the optimal action a_4 from selection. Based on $\hat{\theta}$, we apply both the greedy policy and the pessimistic policy. The simulation results are derived from 100 independent runs.

EC. 2. Extension to Markov Decision Processes

Now, we extend our framework to MDPs. We consider a finite-horizon MDP characterized by the tuple $(\mathcal{S}, \mathcal{A}, N, \{P_i\}_{i=1}^N, \{r_i\}_{i=1}^N, \rho)$. Here, $\mathcal{S}$ represents the state space, $\mathcal{A}$ is the action space, N denotes the horizon length, $P_i : \mathcal{S}\times\mathcal{A} \mapsto \Delta(\mathcal{S})$ is the probability transition at step i, $r_i : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$ is the reward function, and ρ is the initial state distribution. At each step i, after taking action a in state s, the system transitions to a new state s' with probability $P_i(s'|s,a)$, and a reward $r_i(s,a)$ is received.

We assume the availability of two trajectories starting from the same initial state for comparison. Initially, we sample the starting state s_0 from a fixed distribution ρ, followed by two trajectories $\tau^{(0)} = (s_0^{(0)}, a_0^{(0)}, s_1^{(0)}, a_1^{(0)}, \cdots, s_N^{(0)}, a_N^{(0)})$ and $\tau^{(1)} = (s_0^{(1)}, a_0^{(1)}, s_1^{(1)}, a_1^{(1)}, \cdots, s_N^{(1)}, a_N^{(1)})$, where both start from s_0, i.e., $s_0^{(0)} = s_0^{(1)} = s_0$. The preference of a teacher with rationality parameter β over the two trajectories $\tau^{(0)}$ and $\tau^{(1)}$ is given by

$$\mathbb{P}(Y=1|s_0,\tau^{(0)},\tau^{(1)},\beta,\theta_*) = \frac{e^{\beta\theta_*^\top \sum_{i=0}^N \phi(s_i^{(1)},a_i^{(1)})}}{e^{\beta\theta_*^\top \sum_{i=0}^N \phi(s_i^{(0)},a_i^{(0)})} + e^{\beta\theta_*^\top \sum_{i=0}^N \phi(s_i^{(1)},a_i^{(1)})}}. \tag{EC.1}$$

We have a dataset $\{(o^{(i)}, \tau_i^{(0)}, \tau_i^{(1)})\}_{i=1}^n$, where $o^{(i)}$ denotes the type of the trajectory, and define $z^{(i)} = \sum_{i=0}^N [\phi(s_i^{(1)}, a_i^{(1)}) - \phi(s_i^{(0)}, a_i^{(0)})]$ for reward learning with a sample budget constraint. To

estimate θ_*, we select T samples from $(z^{(1)}, \cdots, z^{(n)})$ and T teachers from $\{\beta_1^{(k)}, \cdots, \beta_m^{(k)}\}_{k=1}^g$ using Algorithm 1, by modifying only the calculation of z. The conclusions regarding the MLE $\hat{\theta}_T$ derived from the contextual bandit setting using Algorithm 1 are applicable here.

A deterministic policy $\pi_i : \mathcal{S} \mapsto \mathcal{A}$ is a function that maps a state to an action at step i. We use π to denote the collection of policies $\{\pi_i\}_{i=1}^N$. The associated value function $V^\pi(s) = \mathbb{E}[\sum_{i=0}^N r_i(s_i, a_i)|s_0 = s, a_i = \pi_i(s_i)]$ represents the expected cumulative reward from starting in state s and adhering to π_i at each step i. We define the state occupancy measure $d^\pi(s) = \sum_{i=1}^N \mathbb{P}_i(s_i = s|\pi)$ and the state-action occupancy measure $d^\pi(s, a) = \sum_{i=1}^N \mathbb{P}_i(s_i = s, a_i = a|\pi)$, where $\mathbb{P}_i(s_i = s|\pi)$ denotes the probability of visiting state $s_i = s$ (similarly for $s_i = s, a_i = a$) at step i after executing policy π and starting from $s_0 \sim \rho$.

For analyzing sub-optimality, we employ a pessimistic estimate of the rewards. When the transition distribution P is known, the occupancy measure d^π can be directly computed. If P is unknown, it can be estimated by collecting state-action trajectories through interactions with the environment, as outlined in the method proposed by Zhan et al. (2024b). Given the definition of d^π, one has $\mathbb{E}_{s\sim\rho}[V^\pi(s)] = \mathbb{E}_{s,a\sim d^\pi}[r(s, a)]$. The pessimistic expected value function is formulated as

$$\hat{J}_T(\pi) = \min_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}_{s\sim d^\pi}\theta^\top\phi(s, \pi(s)) = \hat{\theta}_T^\top \mathbb{E}_{s\sim d^\pi}\phi(s, \pi(s)) - \|\mathbb{E}_{s\sim d^\pi}\phi(s, \pi(s))\|_{\bar{H}_T^{-1}(\hat{\theta}_T)}\gamma(T, d, \delta).$$

Then, the pessimistic policy is obtained as $\hat{\pi}_T = \arg\max_\pi \hat{J}_T(\pi)$.

THEOREM 4. *Under Assumptions 1 and 2, for any $0 < \delta < 1$, with probability at least $1 - \delta$, when $T > T_0$ for some positive constant T_0, the sub-optimality of the pessimistic policy $\hat{\pi}_T$ for the offline MDPs is bounded by*

$$\mathit{SubOpt}(\hat{\pi}_T) \le 2\sqrt{\frac{C_3}{T}\left[d\log\left(e + \frac{C_4 T}{d}\right) + \log\frac{2}{\delta}\right]}\|M^{-1/2}(\xi_*, \theta_*)\mathbb{E}_{s\sim d^{\pi^*}}\phi(s, \pi^*(s))\|_2,$$

where C_3 and C_4 are some positive constants.

EC. 3. Additional Experiments

In this section, we implement our policy within large language models, utilizing the public datasets `Anthropic` (Bai et al. 2022) and `UltraFeedback` (Cui et al. 2024). We collect all the prompts with single-turn dialogues from the dataset and process them into a pairwise training format, where each question is paired with two answers. Here, each question serves as context x, and the two answers serve as $a^{(0)}$ and $a^{(1)}$. Each answer is given a rating score[2], and the answer with the higher

score is treated as the chosen one. We randomly select 40000 samples and divide them into a training subset and a test subset with a 4:1 ratio.

The pretrained model employed is `Gemma-7b-it`[3] (Team et al. 2024). The feature map ϕ in (1) is derived by removing the last layer of the pretrained language model, yielding a d-dimensional vector, where the dimension $d = 3072$ is determined by the `Gemma-7b-it` model. More details of the pretrained model and the real data are in Appendix EC. 4.

Our goal in this experiment is to learn the reward function specified in (1) within a sample budget of $T = 5000$, i.e., selecting 5000 samples from $n = 32000$ training samples. We briefly describe the process of the experiments. The question and two corresponding answers $(x, a^{(0)}, a^{(1)})$ are first input into the `Gemma-7b-it` model. After being processed through the last layer of the `Gemma-7b-it` model, the triple $(x, a^{(0)}, a^{(1)})$ is transformed into a 3072-dimensional vector $\phi(x, a^{(0)}, a^{(1)})$. The preference y denoting the preference between $a^{(0)}$ and $a^{(1)}$ follows the Bernoulli distribution as described in (3). Our objective is to estimate the parameter θ_* in (3) using MLE.

Since no information about the rationality of teachers is available in the dataset, we engage various LLMs as synthetic teachers. The LLMs considered are `Qwen2.5-7B-Instruct` (Team 2024), `Yi-1.5-34B-Chat`[4] and `glm-4-9b-chat` (GLM et al. 2024). These LLMs provide their preference y on two answer options for a single question. Questions are categorized into $g = 5$ groups using k-means clustering (Buitinck et al. 2013). We first derive the MLE $\hat{\theta}$ from the original data. Using $\hat{\theta}$, the information of the type of question, and the preference of each LLM, we estimate the rationality of each LLM $\beta_j^{(k)}$ under the constraints $\sum_{k=1}^{5} \beta_j^{(k)} = 10$ and $\beta_j^{(k)} \in [0, 10]$ using MLE. We implement different policies to select conversations and query the synthetic teacher for the preference between the two answers, and estimate θ_*.

Since the true reward parameter in (3) is unknown, we evaluate the effectiveness of different policies using the reward accuracy, which is widely used in assessing reward estimation in large language models (Yao et al. 2023, Das et al. 2025). Using the estimator $\hat{\theta}$, we can obtain the estimated reward $\hat{\theta}^\top \phi(x, a)$. The reward accuracy is defined as the percentage of instances where the estimated reward of the chosen response exceeds that of the rejected one. A higher reward accuracy signifies a better policy.

We evaluate reward accuracy on test samples across various T values and batch sizes ($K = 50, 100$) using different methods. All experiments were conducted on a single Nvidia A100 GPU. Due to the prohibitive computational cost demonstrated in simulations, the case with $K = 1$ is excluded. Figures EC.1 and EC.2 showcase the results, highlighting the superior performance of **DualAL**

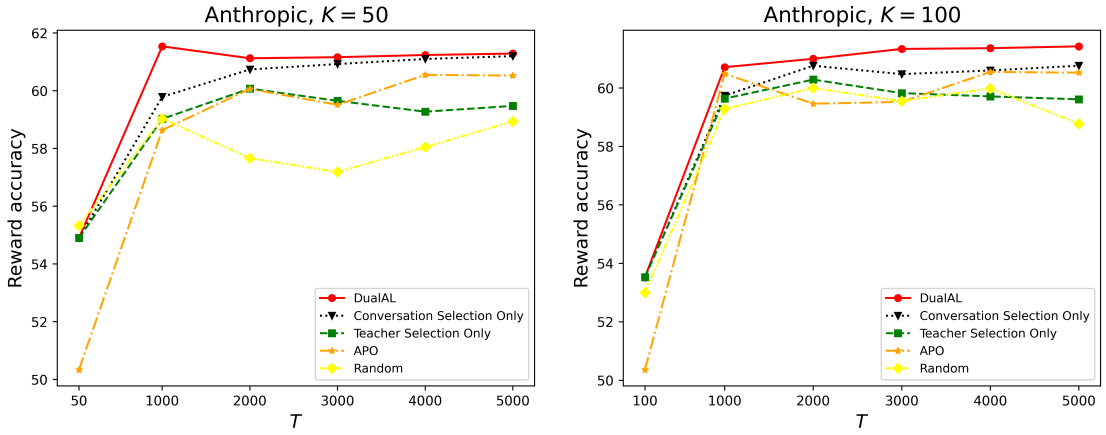


Figure EC.1 Reward accuracy with different methods using the `Anthropic` dataset.

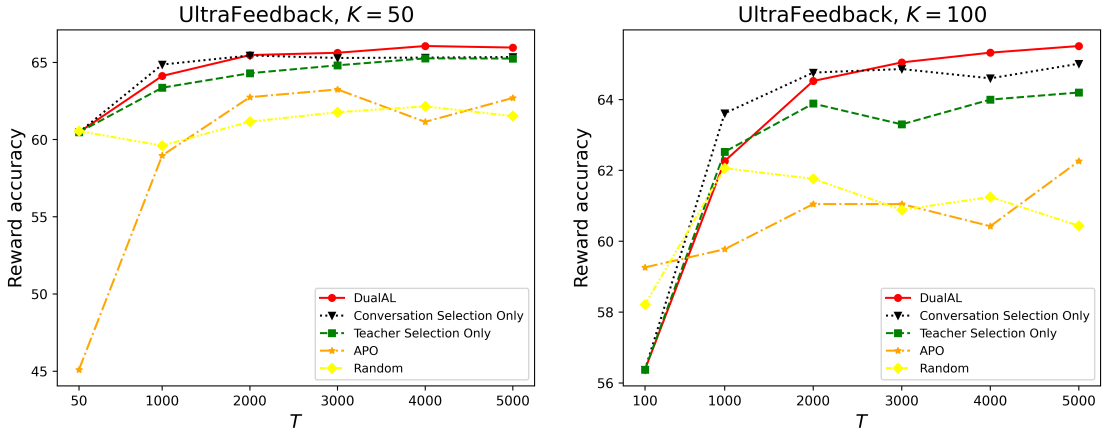


Figure EC.2 Reward accuracy with different methods using the `UltraFeedback` dataset.

compared to benchmark policies. Additionally, we explore the computational efficiency of the batch version of **DualAL**, observing marked reductions in computation time with increased batch sizes: for `Anthropic`, times are 6.06 hours ($K = 50$), and 2.70 hours ($K = 100$); for `UltraFeedback`, times are 6.05 hours ($K = 50$), and 2.68 hours ($K = 100$).

EC. 4. The Datasets and the Pretrained Model

In this section, we give a brief description of the datasets and the pretrained model used in the experiments. All the descriptions are adapted from `Hugging Face`[5]. The dataset `prism-alignment`[6] (Kirk et al. 2024) is a diverse human feedback dataset for preference and value alignment in LLMs. It maps the characteristics and stated preferences of humans from a detailed survey onto their real-time interactions with LLMs and contextual preference ratings. The dataset `Anthropic`[7] (Bai et al. 2022) is about helpfulness and harmlessness, and is meant to

train preference (or reward) models for subsequent RLHF training. These data are not meant for supervised training of dialogue agents. For helpfulness, the data are grouped into train/test splits in three tranches: from our base models (context-distilled 52B language models), via rejection sampling (mostly with best-of-16 sampling) against an early preference model, and a dataset sampled during our iterated "online" process. For harmlessness, the data are only collected for our base models, but otherwise formatted in the same way. The dataset `UltraFeedback`[8] (Cui et al. 2024) is a large-scale, fine-grained, diverse preference dataset, used for training powerful reward models and critic models. About 64k prompts are collected from diverse resources (including UltraChat, ShareGPT, Evol-Instruct, TruthfulQA, FalseQA, and FLAN). These prompts are then used to query multiple LLMs and generate 4 different responses for each prompt, resulting in a total of 256k samples. The `Gemma-7b-it`[9] model is among the Gemma (Team et al. 2024) family, which is a collection of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models. They are text-to-text, decoder-only large language models, available in English, with open weights, pretrained variants, and instruction-tuned variants. Gemma models are well-suited for a variety of text generation tasks, including question answering, summarization, and reasoning.

EC. 5. Proofs

EC. 5.1. Proof of Theorem 1 Recall the definition of $H_t(\hat{\theta}_{t-1})$ in (7). When $H_{t-1}(\hat{\theta}_{t-1})$ is nonsingular, we have

$$\det[H_{t-1}(\hat{\theta}_{t-1}) + \dot{\mu}(\beta\hat{\theta}_{t-1}^\top z)\beta^2 zz^\top] = \det H_t(\hat{\theta}_{t-1})\left[1 + \dot{\mu}(\beta z^\top \hat{\theta}_{t-1})\beta^2 z^\top H_{t-1}^{-1}(\hat{\theta}_{t-1})z\right]. \quad \text{(EC.2)}$$

The above equation follows from Lemma 5 with $R = H_{t-1}(\hat{\theta}_{t-1})$, $\tilde{T} = 1$, $S = \dot{\mu}(\beta\hat{\theta}_{t-1}^\top z)\beta^2 z$, $U = z^\top$. At step t, $H_{t-1}(\hat{\theta}_{t-1})$ is fixed. From (EC.2), the maximization of $\det[H_t(\hat{\theta}_{t-1}) + \dot{\mu}(\beta\hat{\theta}_{t-1}^\top z)\beta^2 zz^\top]$ is equivalent to the maximization of $\dot{\mu}(\beta z^\top \hat{\theta}_{t-1})\beta^2 z^\top H_{t-1}^{-1}(\hat{\theta}_{t-1})z$. For ease of presentation, we denote $h(\beta|z, \hat{\theta}_{t-1}) = \dot{\mu}(\beta z^\top \hat{\theta}_{t-1})\beta^2 z^\top H_{t-1}^{-1}(\hat{\theta}_{t-1})z$. The rationality parameter β influences $h(\beta|z, \hat{\theta}_{t-1})$ through two aspects: $\dot{\mu}(\beta\hat{\theta}_{t-1}^\top z)$ and β^2. Recall that $\beta > 0$. On the one hand, a large β leads to a larger β^2, which contributes to the increase of $h(\beta|z, \hat{\theta}_{t-1})$. On the other hand, β affects $h(\beta|z, \hat{\theta}_{t-1})$ through $\dot{\mu}(\beta\hat{\theta}_{t-1}^\top z)$. By simple calculation, we have $\dot{\mu}(\beta\hat{\theta}_{t-1}^\top z) = \mu(\beta\hat{\theta}_{t-1}^\top z)[1 - \mu(\beta\hat{\theta}_{t-1}^\top z)]$. Clearly, an increase in β does not always lead to an increase in $\dot{\mu}(\beta\hat{\theta}_{t-1}^\top z)$. Thus, a more rational teacher is not always the most informative.

EC. 5.2. Proof of Lemma 1 We denote $\mathcal{L}_T(\theta) = -TL_T(\theta)$, where $L_T(\theta)$ is defined in (4). Then, the MLE is $\hat{\theta}_T = \arg\min_{\theta\in\Theta} \mathcal{L}_T(\theta)$. By the Taylor expansion (Lee et al. 2024), we have

$$\mathcal{L}_T(\theta) = \mathcal{L}_T(\theta_*) + \nabla\mathcal{L}_T(\theta_*)^\top(\theta-\theta_*) + \|\theta-\theta_*\|^2_{G_T(\theta_*,\theta)}, \tag{EC.3}$$

where

$$G_T(\theta_*,\theta) = \sum_{t=1}^T \left[\int_0^1 (1-v)\dot{\mu}(\beta_t z_t^\top(\theta_* + v(\theta-\theta_*)))dv\right]\beta_t^2 z_t z_t^\top.$$

By the definition of $H_T(\theta)$ in (7), we have

$$\begin{aligned}
H_T(\theta) &= \sum_{t=1}^T \dot{\mu}(\beta_t\theta^\top z_t)\beta_t^2 z_t z_t^\top \\
&\preceq \sum_{t=1}^T \left[C(2+|\beta_t z_t^\top(\theta-\theta_*)|)^2\int_0^1 (1-v)\dot{\mu}(\beta_t z_t^\top(\theta_* + v(\theta-\theta_*)))dv\right]\beta_t^2 z_t z_t^\top \\
&\preceq \sum_{t=1}^T \left[C(2+2C_\beta C_\theta C_z)^2\int_0^1 (1-v)\dot{\mu}(\beta_t z_t^\top(\theta_* + v(\theta-\theta_*)))dv\right]\beta_t^2 z_t z_t^\top \\
&= C(2+2C_\beta C_\theta C_z)^2 G_T(\theta_*,\theta),
\end{aligned}$$

where the first inequality follows from Lemma 12 with some constant $C > 1$, and the second inequality is due to Assumption 2. Then $H_T(\hat{\theta}_T) \preceq C(2+2C_\beta C_\theta C_z)^2 G_T(\theta_*,\hat{\theta}_T)$ for some $C > 1$. Together with (EC.3), we have

$$\begin{aligned}
\|\hat{\theta}_T-\theta_*\|^2_{H_T(\hat{\theta}_T)} &\le C(2+2C_\beta C_\theta C_z)^2\|\hat{\theta}_T-\theta_*\|^2_{G_T(\theta_*,\hat{\theta}_T)} \\
&= C(2+2C_\beta C_\theta C_z)^2[\mathcal{L}_T(\hat{\theta}_T) - \mathcal{L}_T(\theta_*) + \nabla\mathcal{L}_T(\theta_*)^\top(\theta_*-\hat{\theta}_T)] \\
&\le C(2+2C_\beta C_\theta C_z)^2\nabla\mathcal{L}_T(\theta_*)^\top(\theta_*-\hat{\theta}_T),
\end{aligned} \tag{EC.4}$$

where the last inequality is from $\mathcal{L}_T(\hat{\theta}_T) \le \mathcal{L}_T(\theta_*)$. Now, we bound $\nabla\mathcal{L}_T(\theta_*)^\top(\theta_*-\hat{\theta}_T)$. We define $\xi_t = \mu(\beta_t z_t^\top\theta_*) - y_t$. Then,

$$\nabla\mathcal{L}_T(\theta_*)^\top(\theta_*-\theta) = \sum_{t=1}^T[\mu(\beta_t z_t^\top\theta_*) - y_t]\beta_t z_t^\top(\theta_*-\theta) = \sum_{t=1}^T \xi_t\beta_t z_t^\top(\theta_*-\theta). \tag{EC.5}$$

Here ξ_t is a martingale difference sequence w.r.t. $\mathcal{F}_{t-1} = \sigma(z_1,\beta_1,y_1,\cdots,z_{t-1},\beta_{t-1},y_{t-1},z_t,\beta_t)$. Then $\xi_t\beta_t z_t^\top(\theta_*-\theta)$ is a martingale difference sequence. Since $|\xi_t\beta_t z_t^\top(\theta_*-\theta)| \le 2C_\beta C_z C_\theta$ and

$\mathbb{E}\{[\xi_t\beta_t z_t^\top(\theta_*-\theta)]^2|\mathcal{F}_{t-1}\}=\dot{\mu}(\beta_t z_t^\top\theta_*)[\beta_t z_t^\top(\theta_*-\theta)]^2$, by Lemma 11, for any $\eta\in(0,\frac{1}{2C_\beta C_z C_\theta}]$, with probability at least $1-\frac{\delta}{2}$, we have

$$\begin{aligned}\sum_{t=1}^{T}\xi_t\beta_t z_t^\top(\theta_*-\theta)&\le(e-2)\eta\sum_{t=1}^{T}\dot{\mu}(\beta_t z_t^\top\theta_*)[\beta_t z_t^\top(\theta_*-\theta)]^2+\frac{1}{\eta}\log\frac{2}{\delta}\\&=(e-2)\eta\|\theta_*-\theta\|^2_{H_T(\theta_*)}+\frac{1}{\eta}\log\frac{2}{\delta}.\end{aligned}\tag{EC.6}$$

By (EC.5) and (EC.6), replacing θ with $\hat{\theta}_T$, with probability at least $1-\frac{\delta}{2}$, we have

$$\nabla\mathcal{L}_T(\theta_*)^\top(\theta_*-\hat{\theta}_T)\le(e-2)\eta\|\theta_*-\hat{\theta}_T\|^2_{H_T(\theta_*)}+\frac{1}{\eta}\log\frac{2}{\delta}.\tag{EC.7}$$

By setting $\eta=\frac{1}{(e-2)(4+4C_\beta C_z C_\theta)}$, similar to the arguments in Lemma 6 of Lee et al. (2024), with probability at least $1-\frac{\delta}{2}$, we can obtain

$$\|\theta_*-\hat{\theta}_T\|^2_{H_T(\theta_*)}\le C'(C_\beta C_z C_\theta)^2\left[d\log\left(e+\frac{C_\beta C_z C_\theta T}{d}\right)+\log\frac{2}{\delta}\right]\tag{EC.8}$$

for some positive constant C'. By (EC.7) and (EC.8), with probability at least $1-\delta$, we have

$$\nabla\mathcal{L}_T(\theta_*)^\top(\theta_*-\hat{\theta}_T)\le\frac{C'(C_\beta C_z C_\theta)^2}{4+4C_\beta C_z C_\theta}\left[d\log\left(e+\frac{C_\beta C_z C_\theta T}{d}\right)+\log\frac{2}{\delta}\right]+(e-2)(4+4C_\beta C_z C_\theta)\log\frac{2}{\delta}.\tag{EC.9}$$

We define $C_1=\frac{CC'(C_\beta C_z C_\theta)^2(2+2C_\beta C_z C_\theta)}{2}+2C(e-2)(2+2C_\beta C_z C_\theta)^3$ and $C_2=C_\beta C_z C_\theta$. By (EC.4) and (EC.9), with probability at least $1-\delta$, we have

$$\|\hat{\theta}_T-\theta_*\|_{H_T(\hat{\theta}_T)}\le\sqrt{C_1\left[d\log\left(e+\frac{C_2T}{d}\right)+\log\frac{2}{\delta}\right]}.$$

We define $\bar{H}_T(\hat{\theta}_T)=\frac{1}{T}H_T(\hat{\theta}_T)$. Then, with probability at least $1-\delta$, it follows that

$$\|\hat{\theta}_T-\theta_*\|_{\bar{H}_T(\hat{\theta}_T)}\le\sqrt{\frac{C_1}{T}\left[d\log\left(e+\frac{C_2T}{d}\right)+\log\frac{2}{\delta}\right]}.$$

EC. 5.3. Proof of Theorem 2 Now we prove the asymptotic property of $\hat{\theta}_T$. The convergence of this adaptively generated information matrix $M(\xi_T,\hat{\theta}_T)$ is established in the following theorem.

THEOREM 5. *Assuming that Assumptions 1 and 2 are satisfied and $\hat{\theta}_T$ is the estimator derived from Algorithm 1, let $M(\xi_T,\hat{\theta}_T)$ be as defined in (5) and $M(\xi_*,\theta_*)$ as in (6). It follows that*

$$M(\xi_T,\hat{\theta}_T)\xrightarrow{a.s.}M(\xi_*,\theta_*),\ as\ T\to\infty,$$

where $\xrightarrow{a.s.}$ denotes convergence almost surely.

Theorem 5 asserts that the information matrix $M(\xi_T, \hat{\theta}_T)$ converges almost surely to $M(\xi_*, \theta_*)$, which maximizes $\det M(\xi, \theta_*)$ over the set of all designs. The proof of Theorem 5 is deferred to Section EC. 5.4.

We take the gradient of $L_T(\theta)$ with respect to θ as follows,

$$\frac{\partial L_T(\theta)}{\partial \theta} = \frac{1}{T}\sum_{t=1}^{T} \dot{\mu}(\beta_t \theta^\top z_t)\left[\frac{y_t \beta_t z_t}{\mu(\beta_t \theta^\top z_t)} - \frac{(1-y_t)\beta_t z_t}{1-\mu(\beta_t \theta^\top z_t)}\right] = \frac{1}{T}\sum_{t=1}^{T}[y_t - \mu(\beta_t \theta^\top z_t)]\beta_t z_t. \quad \text{(EC.10)}$$

We denote $S_T(\theta) = \frac{\partial L_T(\theta)}{\partial \theta}$ as the score function. Since $S_T(\hat{\theta}_T) = 0$, by the Taylor expansion, we have

$$\begin{aligned}
S_T(\theta^*) &= S_T(\theta^*) - S_T(\hat{\theta}_T) \\
&= \frac{1}{T}\sum_{t=1}^{T}[\mu(\beta_t \hat{\theta}_T^\top z_t) - \mu(\beta_t z_t^\top \theta_*)]\beta_t z_t \\
&= \frac{1}{T}\sum_{t=1}^{T}\dot{\mu}(\beta_t \tilde{\theta}_t^\top z_t)\beta_t^2 z_t z_t^\top (\hat{\theta}_T - \theta_*) \\
&= \frac{1}{T}\left[\sum_{t=1}^{T}\dot{\mu}(\beta_t \tilde{\theta}_t^\top z_t)\beta_t^2 z_t z_t^\top - \sum_{t=1}^{T}\dot{\mu}(\beta_t z_t^\top \theta_*)\beta_t^2 z_t z_t^\top + \sum_{t=1}^{T}\dot{\mu}(\beta_t z_t^\top \theta_*)\beta_t^2 z_t z_t^\top\right](\hat{\theta}_T - \theta_*) \\
&= \left\{\frac{1}{T}\sum_{t=1}^{T}[\dot{\mu}(\beta_t \tilde{\theta}_t^\top z_t) - \dot{\mu}(\beta_t z_t^\top \theta_*)]\beta_t^2 z_t z_t^\top + M(\xi_T, \theta_*)\right\}(\hat{\theta}_T - \theta_*),
\end{aligned}$$

where $\tilde{\theta}_t$ is on the line segment joining θ_* and $\hat{\theta}_T$. We denote $M_* = M(\xi_*, \theta_*)$. Therefore,

$$\sqrt{T} M_*^{-1/2} S_T(\theta_*) = M_*^{-1/2}\left\{\frac{1}{T}\sum_{t=1}^{T}[\dot{\mu}(\beta_t \tilde{\theta}_t^\top z_t) - \dot{\mu}(\beta_t z_t^\top \theta_*)]\beta_t^2 z_t z_t^\top + M(\xi_T, \theta_*)\right\}\sqrt{T}(\hat{\theta}_T - \theta_*). \quad \text{(EC.11)}$$

We propose a lemma to show that the left side of (EC.11) converges to a multivariate normal distribution.

LEMMA 2. *Let $M_* = M(\xi_*, \theta_*)$ be defined in (6) and $S_T(\theta_*)$ be the score function defined in (EC.10). Under Assumption 2, we have*

$$\sqrt{T} M_*^{-1/2} S_T(\theta_*) \xrightarrow{d} N(0, I_d).$$

Proof. Let $\tilde{v} \in \mathbb{R}^d$ and $v = \tilde{v}/\|\tilde{v}\|$. Then, $\|v\| = 1$. Recall that $e_i = y_i - \mu(\beta_i z_i^\top \theta^*)$ is defined in (EC.22). By (EC.10), we have

$$\sqrt{T} v^\top M_*^{-1/2} S_T(\theta_*) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} e_t \beta_t v^\top M_*^{-1/2} z_t. \quad \text{(EC.12)}$$

We define the σ-field generated by the historical data as follows,

$$\mathcal{F}_t = \sigma(z_1, \ldots, z_t; \beta_1, \cdots, \beta_t; y_1, \ldots, y_t) \tag{EC.13}$$

Under Assumption 2, $|\sum_{t=1}^{T} e_t \beta_t v^\top M_*^{-1/2} z_t|$ is bounded. Since z_t and β_t in Algorithm 1 are determined by $\mathcal{F}_{t-1}$, z_t and β_t are measurable with respect to $\mathcal{F}_{t-1}$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{t=1}^{T} e_t \beta_t v^\top M_*^{-1/2} z_t | \mathcal{F}_{T-1}\right) &= \mathbb{E}\left(\sum_{t=1}^{T-1} e_t \beta_t v^\top M_*^{-1/2} z_t | \mathcal{F}_{T-1}\right) + \mathbb{E}(e_T \beta_T v^\top M_*^{-1/2} z_T | \mathcal{F}_{T-1}) \\
&= \sum_{t=1}^{T-1} e_t \beta_t v^\top M_*^{-1/2} z_t + \mathbb{E}(e_T | \mathcal{F}_{T-1}) \beta_T v^\top M_*^{-1/2} z_T \\
&= \sum_{t=1}^{T-1} e_t \beta_t v^\top M_*^{-1/2} z_t,
\end{aligned}$$

Thus, the sequence of partial sums $\sum_{t=1}^{T} e_t \beta_t v^\top M_*^{-1/2} z_t$ is a martingale with respect to $\mathcal{F}_T$. Since

$$\begin{aligned}
\mathbb{E}(e_t^2 | \mathcal{F}_{t-1}) &= \mathbb{E}\{[y_t - \mu(\beta_t \theta_*^\top z_t)]^2 | \mathcal{F}_{t-1}\} \\
&= \mathbb{E}(y_t^2 | \mathcal{F}_{t-1}) + \mu^2(\beta_t \theta_*^\top z_t) - 2\mathbb{E}(y_t | \mathcal{F}_{t-1}) \mu(\beta_t \theta_*^\top z_t) \\
&= \mu(\beta_t \theta_*^\top z_t) - \mu^2(\beta_t \theta_*^\top z_t) \\
&= \dot{\mu}(\beta_t \theta_*^\top z_t),
\end{aligned}$$

we have

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[(e_t \beta_t v^\top M_*^{-1/2} z_t)^2 | \mathcal{F}_{t-1}] &= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[e_t^2 | \mathcal{F}_{t-1}] (\beta_t v^\top M_*^{-1/2} z_t)^2 \\
&= \frac{1}{T} \sum_{t=1}^{T} v^\top M_*^{-1/2} \dot{\mu}(\beta_t \theta_*^\top z_t) \beta_t^2 z_t z_t^\top M_*^{-1/2} v \\
&= v^\top M_*^{-1/2} M(\xi_T, \theta_*) M_*^{-1/2} v \\
&\xrightarrow{a.s.} 1,
\end{aligned} \tag{EC.14}$$

where the convergence follows from Theorem 5. For all $\epsilon > 0$, we have

$$\begin{aligned}
&\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[(e_t \beta_t v^\top M_*^{-1/2} z_t)^2 \mathbb{I}(|e_t \beta_t v^\top M_*^{-1/2} z_t| > \sqrt{T}\epsilon) | \mathcal{F}_{t-1}] \\
\leq\ &\frac{1}{\epsilon^2 T^2} \sum_{t=1}^{T} \mathbb{E}[(e_t \beta_t v^\top M_*^{-1/2} z_t)^4 | \mathcal{F}_{t-1}] \\
=\ &\frac{1}{\epsilon^2 T^2} \sum_{t=1}^{T} (\beta_t v^\top M_*^{-1/2} z_t)^4 \mathbb{E}(e_t^4 | \mathcal{F}_{t-1}) \\
\xrightarrow{a.s.}\ &0,
\end{aligned} \tag{EC.15}$$

where the first inequality follows from

$$(e_t\beta_t v^\top M_*^{-1/2}z_t)^2\mathbb{I}(|e_t\beta_t v^\top M_*^{-1/2}z_t|>\sqrt{T}\epsilon)\le\frac{(e_t\beta_t v^\top M_*^{-1/2}z_t)^4}{\epsilon^2 T},$$

and the convergence is from the fact that $(\beta_t v^\top M_*^{-1/2}z_t)^4\mathbb{E}(e_t^4|\mathcal{F}_{t-1})$ is bounded under Assumption 2. By (EC.14), (EC.15) and Lemma 9, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}e_t\beta_t v^\top M_*^{-1/2}z_t\xrightarrow{d}N(0,1).$$

Using (EC.12), we have

$$\sqrt{T}v^\top M_*^{-1/2}S_T(\theta^*)\xrightarrow{d}N(0,1).$$

Let $\tilde{Z}$ be a normal vector with $\tilde{Z}\sim N(0,I_d)$. Then, $v^\top\tilde{Z}\sim N(0,1)$ because of $\|v\|=1$. Therefore, for any $\tilde{v}\in\mathbb{R}^d$, we have

$$\sqrt{T}\frac{\tilde{v}^\top}{\|\tilde{v}\|}M_*^{-1/2}S_T(\theta^*)\xrightarrow{d}\frac{\tilde{v}^\top}{\|\tilde{v}\|}\tilde{Z}.$$

Thus,

$$\sqrt{T}\tilde{v}^\top M_*^{-1/2}S_T(\theta^*)\xrightarrow{d}\tilde{v}^\top\tilde{Z}.$$

By Lemma 10, we have

$$\sqrt{T}M_*^{-1/2}S_T(\theta^*)\xrightarrow{d}N(0,I_d).$$

□

We now return to the proof of Theorem 2. By (EC.11) and Lemma 2, we have

$$M_*^{-1/2}\left\{\frac{1}{T}\sum_{t=1}^{T}[\dot{\mu}(\beta_t\tilde{\theta}_t^\top z_t)-\dot{\mu}(\beta_t z_t^\top\theta_*)]\beta_t^2 z_t z_t^\top+M(\xi_T,\theta_*)\right\}\sqrt{T}(\hat{\theta}_T-\theta_*)\xrightarrow{d}N(0,I_d). \tag{EC.16}$$

Under Assumption 2, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}[\dot{\mu}(\beta_t\tilde{\theta}_t^\top z_t)-\dot{\mu}(\beta_t z_t^\top\theta_*)]\beta_t^2 z_t z_t^\top&\le\max_{1\le t\le T}|\dot{\mu}(\beta_t\tilde{\theta}_t^\top z_t)-\dot{\mu}(\beta_t\theta_*^\top z_t)|\frac{1}{T}\sum_{t=1}^{T}\beta_t^2\|z_t\|^2\\&\le\max_{1\le t\le T}|\dot{\mu}(\beta_t\tilde{\theta}_t^\top z_t)-\dot{\mu}(\beta_t\theta_*^\top z_t)|C_\beta^2C_z^2.\end{aligned} \tag{EC.17}$$

We denote $\ddot{\mu}(w)=\frac{d\dot{\mu}(w)}{dw}=\frac{d\mu(w)[1-\mu(w)]}{dw}=\mu(w)[1-\mu(w)][1-2\mu(w)]$ for $w\in\mathbb{R}$. Under Assumption 2, there exists a positive constant C_μ such that $\ddot{\mu}(\beta\theta^\top z)\le C_\mu$ for any $\beta\in\mathcal{B}$, $z\in\mathcal{Z}$ and $\theta\in\Theta$. By the Taylor expansion, we have

$$\max_{1\le t\le T}|\dot{\mu}(\beta_t\tilde{\theta}_t^\top z_t)-\dot{\mu}(\beta_t\theta_*^\top z_t)|=\max_{1\le t\le T}|\ddot{\mu}(\beta_t\bar{\theta}_t^\top z_t)\beta_t z_t^\top(\tilde{\theta}_t-\theta_*)|\quad\le C_\mu C_\beta C_z\max_{1\le t\le T}\|\tilde{\theta}_t-\theta_*\|^2. \tag{EC.18}$$

Recall that $\tilde{\theta}_t$ is between $\hat{\theta}_t$ and θ_*. We have

$$\max_{1\le t\le T} \|\tilde{\theta}_t - \theta_*\|^2 \le \max_{1\le t\le T} \|\hat{\theta}_t - \theta_*\|^2 \xrightarrow{a.s.} 0, \tag{EC.19}$$

where the convergence follows from Lemma 3. By (EC.17), (EC.18) and (EC.19), we have

$$\frac{1}{T}\sum_{t=1}^{T}[\dot{\mu}(\beta_t\tilde{\theta}_t^\top z_t) - \dot{\mu}(\beta_t z_t^\top \theta_*)]\beta_t^2 z_t z_t^\top \xrightarrow{a.s.} 0.$$

Together with Theorem 5, we have

$$M_*^{-1/2}\left\{\frac{1}{T}\sum_{t=1}^{T}[\dot{\mu}(\beta_t\tilde{\theta}_t^\top z_t) - \dot{\mu}(\beta_t z_t^\top \theta_*)]\beta_t^2 z_t z_t^\top + M(\xi_T, \theta_*)\right\} \xrightarrow{a.s.} M_*^{1/2}.$$

By (EC.11) and Lemma 2, we have

$$M_*^{1/2}\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{d} N(0, I_d).$$

It follows that

$$\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{d} N(0, M_*^{-1}).$$

The proofs of the asymptotic properties of $\hat{\theta}_T^c$ and $\hat{\theta}_T^t$ follow steps similar to those for $\hat{\theta}_T$, and are therefore omitted.

EC. 5.4. Proof of Theorem 5 In this section, we first propose Lemma 3 to show the strong consistency of the adaptive MLE $\hat{\theta}_T$. This lemma plays a pivotal role as a fundamental component in the proof of Theorem 5. Under Assumption 1, the initial information matrix $M(\xi_{t_0}, \theta)$ is constructed as positive definite for any $\theta \in \Theta$ for a theoretical requirement.

LEMMA 3. *Denote $\hat{\theta}_T$ as the estimator from Algorithm 1. We have*

$$\hat{\theta}_T \xrightarrow{a.s.} \theta_*.$$

Proof. According to (4), we calculate the log-likelihood difference between θ_* and $\theta \in \Theta$ as

$$\begin{aligned} L_T(\theta_*) - L_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\left\{y_t \log\frac{\mu(\beta_t z_t^\top \theta_*)}{\mu(\beta_t z_t^\top \theta)} + (1-y_t)\log\frac{1-\mu(\beta z_t^\top \theta_*)}{1-\mu(\beta_t z_t^\top \theta)}\right\} \\ &= \frac{1}{T}\sum_{t=1}^{T}\left\{y_t\left[\log\frac{\mu(\beta_t z_t^\top \theta_*)}{1-\mu(\beta z_t^\top \theta_*)} - \log\frac{\mu(\beta_t z_t^\top \theta)}{1-\mu(\beta_t z_t^\top \theta)}\right] + \log\frac{1-\mu(\beta z_t^\top \theta_*)}{1-\mu(\beta_t z_t^\top \theta)}\right\} \\ &= \frac{1}{T}\sum_{t=1}^{T}\{y_t\beta_t z_t^\top(\theta^* - \theta) + \log[1-\mu(\beta z_t^\top \theta^*)] - \log[1-\mu(\beta z_t^\top \theta)]\}, \end{aligned} \tag{EC.20}$$

where the last equality is from

$$\log \frac{\mu(\beta_t z_t^\top \theta_*)}{1-\mu(\beta z_t^\top \theta_*)} - \log \frac{\mu(\beta_t z_t^\top \theta)}{1-\mu(\beta_t z_t^\top \theta)} = \log e^{\beta_t z_t^\top \theta_*} - \log e^{\beta_t z_t^\top \theta} = \beta_t z_t^\top (\theta_* - \theta).$$

Taking the first-order derivative of $\log[1-\mu(w)]$ with respect to w, we obtain

$$\frac{d \log[1-\mu(w)]}{dw} = -\frac{\dot{\mu}(w)}{1-\mu(w)} = -\frac{\mu(w)[1-\mu(w)]}{1-\mu(w)} = -\mu(w),$$

and the second-order derivative is

$$\frac{d^2 \log[1-\mu(w)]}{dw^2} = -\dot{\mu}(w).$$

Therefore, by the second-order Taylor expansion of $\log[1-\mu(\beta z^\top \theta)]$ at $\beta z^\top \theta_*$, we have

$$\log[1-\mu(\beta z^\top \theta)] = \log[1-\mu(\beta z^\top \theta_*)] - \mu(\beta z^\top \theta_*)\beta z^\top (\theta - \theta_*) - \frac{1}{2}\dot{\mu}(\beta z^\top \tilde{\theta})[\beta z^\top (\theta - \theta_*)]^2,$$

where $\tilde{\theta}$ is between θ and θ_*. Therefore,

$$\log[1-\mu(\beta z^\top \theta_*)] - \log[1-\mu(\beta z^\top \theta)] = \mu(\beta z^\top \theta_*)\beta z^\top (\theta - \theta_*) + \frac{1}{2}\dot{\mu}(\beta z^\top \tilde{\theta})[\beta z^\top (\theta - \theta_*)]^2. \tag{EC.21}$$

Now, we define the error terms as

$$e_t = y_t - \mu(\beta_t z_t^\top \theta_*). \tag{EC.22}$$

Combining (EC.20) and (EC.21), we obtain

$$\begin{aligned} L_T(\theta_*) - L_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\{[\mu(\beta_t z_t^\top \theta_*) + e_t]\beta_t z_t^\top (\theta_* - \theta) + \log[1-\mu(\beta z_t^\top \theta_*)] - \log[1-\mu(\beta z_t^\top \theta)]\} \\ &= \frac{1}{T}\sum_{t=1}^{T} e_t \beta_t z_t^\top (\theta_* - \theta) + \frac{1}{2T}\sum_{t=1}^{T}\dot{\mu}(\beta_t z_t^\top \tilde{\theta}_t)[\beta_t z_t^\top (\theta - \theta_*)]^2 \\ &\geq \frac{1}{T}\sum_{t=1}^{T} e_t \beta_t z_t^\top (\theta_* - \theta) + \frac{\kappa}{2T}\sum_{t=1}^{T}[\beta_t z_t^\top (\theta - \theta_*)]^2. \end{aligned} \tag{EC.23}$$

For any $\delta > 0$, we define the parameter subset $C(\theta_*, \delta) = \{\theta \in \Theta : \|\theta - \theta_*\| \geq \delta\}$. Then, for any $\delta > 0$, by (EC.23), we have

$$L_T(\theta^*) - \sup_{\theta \in C(\theta_*, \delta)} L_T(\theta) \geq -\frac{1}{T}\sup_{\theta \in \Theta}\left|\sum_{t=1}^{T} e_t \beta_t z_t^\top (\theta^* - \theta)\right| + \frac{\kappa}{2T}\inf_{\theta \in C(\theta_*, \delta)}\sum_{t=1}^{T}[\beta_t z_t^\top (\theta - \theta_*)]^2. \tag{EC.24}$$

Letting $n_{i,j}^{(k)}$ be the number of observations taken at $(z^{(i)}, \beta_j^{(k)})$ under the generated design ξ_T, we have

$$\begin{aligned}\frac{1}{T}\inf_{\theta\in C(\theta_*,\delta)}\sum_{t=1}^{T}[\beta_t z_t^\top(\theta-\theta_*)]^2 &= \frac{1}{T}\inf_{\theta\in C(\theta_*,\delta)}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{g} n_{i,j}^{(k)}[\beta_j^{(k)}(\theta-\theta_*)^\top z^{(i)}]^2 \\ &= \inf_{\theta\in C(\theta_*,\delta)}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{g}\xi_T(\beta_j^{(k)}, z^{(i)})[\beta_j^{(k)}(\theta-\theta_*)^\top z^{(i)}]^2.\end{aligned} \tag{EC.25}$$

For any $\theta\in\Theta$ and $\theta\neq\theta_*$, we define $c_\theta = (\theta-\theta_*)/\|\theta-\theta_*\|$. By Theorem 2.6 in Freise et al. (2021), there exist $t_0>0, \epsilon>0$ and $\alpha\in(0,1)$ such that for all $T\geq t_0$ and $\theta\neq\theta_*$,

$$\sum_{\beta\in\mathcal{B},z\in\mathcal{Z}}\xi_T(\beta,z)\mathbb{I}(|\sqrt{\dot\mu(\beta z^\top\theta)}\beta c_\theta^\top z|\leq\epsilon)\leq\alpha.$$

Noting that $\dot\mu(\beta z^\top\theta)\leq 1/4$, we have $|\sqrt{\dot\mu(\beta z^\top\theta)}\beta c_\theta^\top z|\leq|\beta c_\theta^\top z|/2$. Therefore, $|\beta c_\theta^\top z|\leq 2\epsilon$ implies $|\sqrt{\dot\mu(\beta z^\top\theta)}\beta c_\theta^\top z|\leq\epsilon$. Then, $\mathbb{I}(|\beta c_\theta^\top z|\leq 2\epsilon)\leq\mathbb{I}(|\sqrt{\dot\mu(\beta z^\top\theta)}\beta c_\theta^\top z|\leq\epsilon)$. Thus, there exist $t_0>0, \epsilon>0$ and $\alpha\in(0,1)$ such that for all $T\geq t_0$ and $\theta\neq\theta_*$,

$$\sum_{\beta\in\mathcal{B},z\in\mathcal{Z}}\xi_T(\beta,z)\mathbb{I}(|\beta c_\theta^\top z|\leq 2\epsilon)\leq\alpha.$$

Because $\sum_{\beta\in\mathcal{B},z\in\mathcal{Z}}\xi_T(\beta,z)=1$, we have

$$\sum_{\beta\in\mathcal{B},z\in\mathcal{Z}}\xi_T(\beta,z)\mathbb{I}(|\beta c_\theta^\top z|>2\epsilon)\geq 1-\alpha. \tag{EC.26}$$

Then,

$$\begin{aligned}&\inf_{\theta\in C(\theta_*,\delta)}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{g}\xi_T(\beta_j^{(k)},z^{(i)})[\beta_{jk}(\theta-\theta_*)^\top z^{(i)}]^2 \\ &\geq \inf_{\theta\in C(\theta_*,\delta)}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{g}\xi_T(\beta_j^{(k)},z^{(i)})[\beta_{jk}(\theta-\theta_*)^\top z^{(i)}]^2\mathbb{I}(|\beta_{jk}(\theta-\theta^*)^\top z^{(i)}|>2\epsilon\|\theta-\theta^*\|) \\ &\geq 4\inf_{\theta\in C(\theta_*,\delta)}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{g}\xi_T(\beta_j^{(k)},z^{(i)})\epsilon^2\|\theta-\theta^*\|^2\mathbb{I}(|\beta_{jk}(\theta-\theta^*)^\top z^{(i)}|>2\epsilon\|\theta-\theta^*\|) \\ &\geq 4\epsilon^2\delta^2(1-\alpha),\end{aligned} \tag{EC.27}$$

where the last inequality is from (EC.26) and the fact that $|\beta c_\theta^\top z|>2\epsilon$ is equivalent to $|\beta(\theta-\theta_*)^\top z|>2\epsilon\|\theta-\theta_*\|$. By (EC.25) and (EC.27) we have

$$\inf_{\theta\in C(\theta_*,\delta)}\frac{1}{T}\sum_{t=1}^{T}[\beta_t z_t^\top(\theta-\theta_*)]^2\geq 4\epsilon^2\delta^2(1-\alpha). \tag{EC.28}$$

By Lemma A.1 in Freise et al. (2021), we have

$$\frac{1}{T}\sup_{\theta\in\Theta}\left|\sum_{t=1}^{T} e_t\beta_t z_t^\top(\theta_*-\theta)\right| \xrightarrow{a.s.} 0. \tag{EC.29}$$

By (EC.24), (EC.28) and (EC.29), we have

$$L_T(\theta_*) - \sup_{\theta\in C(\theta_*,\delta)} L_T(\theta) \geq 2\kappa\epsilon^2\delta^2(1-\alpha)\ a.s.$$

It follows

$$\liminf_{T\to\infty}[L_T(\theta_*) - \sup_{\theta\in C(\theta_*,\delta)} L_T(\theta)] > 0\ a.s.$$

By Lemma 7, we have $\hat{\theta}_T \xrightarrow{a.s.} \theta_*$. □

Now we proceed with the proof of Theorem 5. By the definition of $M(\xi_T,\theta)$ as shown in (5), we calculate the difference of the Fisher matrices between $\hat{\theta}_T$ and θ_* at the design ξ_T as follows,

$$\begin{aligned}
\|M(\xi_T,\hat{\theta}_T) - M(\xi_T,\theta_*)\| &= \left\|\frac{1}{T}\sum_{t=1}^{T}\dot{\mu}(\beta_t\hat{\theta}_T^\top z_t)\beta_t^2 z_t z_t^\top - \frac{1}{T}\sum_{t=1}^{T}\dot{\mu}(\beta_t\theta_*^\top z_t)\beta_t^2 z_t z_t^\top\right\| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\|\dot{\mu}(\beta_t\hat{\theta}_T^\top z_t)\beta_t^2 z_t z_t^\top - \dot{\mu}(\beta_t\theta_*^\top z_t)\beta_t^2 z_t z_t^\top\| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\|[\dot{\mu}(\beta_t\hat{\theta}_T^\top z_t) - \dot{\mu}(\beta_t\theta_*^\top z_t)]\beta_t^2 z_t z_t^\top\| \\
&\leq \frac{C_\beta^2 C_z^2}{T}\sum_{t=1}^{T}\|\dot{\mu}(\beta_t\hat{\theta}_T^\top z_t) - \dot{\mu}(\beta_t\theta_*^\top z_t)\| \\
&\leq C_\beta^2 C_z^2 \max_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}}\|\dot{\mu}(\beta\hat{\theta}_T^\top z) - \dot{\mu}(\beta\theta_*^\top z)\|
\end{aligned} \tag{EC.30}$$

By Lemma 3, we know $\hat{\theta}_T \xrightarrow{a.s.} \theta_*$. Since the real-valued function $(z,\beta,\theta)\mapsto\dot{\mu}(\beta\theta^\top z)$ is uniformly continuous on its compact domain $\mathcal{Z}\times\mathcal{B}\times\Theta$, we have

$$\|\dot{\mu}(\beta\hat{\theta}_T^\top z) - \dot{\mu}(\beta\theta_*^\top z)\| \xrightarrow{a.s.} 0.$$

Using (EC.30), we have

$$\|M(\xi_T,\hat{\theta}_T) - M(\xi_T,\theta_*)\| \xrightarrow{a.s.} 0. \tag{EC.31}$$

Under Assumption 1 and the design in the initialization of Algorithm 1, we have $\lambda_0 := \lambda_{min}(M(\xi_T,\hat{\theta}_T)) > 0$. On the other hand, by Assumption 2, the trace of $M(\xi_T,\hat{\theta}_T)$ is

$$\mathrm{tr}(M(\xi_T,\hat{\theta}_T)) = \frac{1}{T}\sum_{t=1}^{T}\dot{\mu}(\beta_t\hat{\theta}_T^\top z_t)\beta_t^2 z_t^\top z_t \leq \frac{C_\beta^2 C_z^2}{4}.$$

Let $\mathcal{M}$ be the set of all non-negative definite $d \times d$ matrices M such that $\lambda_{min}(M) \geq \lambda_0$ and $\mathrm{tr}(M) \leq C_\beta^2 C_z^2/4$. Obviously, $\mathcal{M}$ is compact. We define a real-valued function G on $\mathcal{Z} \times \mathcal{B} \times \Theta \times \mathcal{M}$ by

$$G(z, \beta, \theta, A) = \dot{\mu}(\beta z^\top \theta)\beta^2 z^\top A^{-1} z,$$

which is uniformly continuous on its compact domain $\mathcal{Z} \times \mathcal{B} \times \Theta \times \mathcal{M}$. Since $M(\xi_T, \hat{\theta}_T) \in \mathcal{M}$ and $M(\xi_T, \theta_*) \in \mathcal{M}$, by (EC.31), we have

$$\max_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} |G(z, \beta, \hat{\theta}_T, M(\xi_T, \hat{\theta}_T)) - G(z, \beta, \theta_*, M(\xi_T, \theta_*))| \xrightarrow{a.s.} 0.$$

Therefore, for a given $\epsilon \in (0, 1)$, there exists t_1 such that for all $(z, \beta) \in \mathcal{Z} \times \mathcal{B}$

$$|\dot{\mu}(\beta \hat{\theta}_T^\top z)\beta^2 z^\top M^{-1}(\xi_T, \hat{\theta}_T) z - \dot{\mu}(\beta \theta_*^\top z)\beta^2 z^\top M^{-1}(\xi_T, \theta_*) z| < \frac{\epsilon}{2} \text{ for all } T \geq t_1. \tag{EC.32}$$

Since $H_{t-1}(\hat{\theta}_{t-1}) = (t-1)M(\xi_{t-1}, \hat{\theta}_{t-1})$, by the generation process of Algorithm 1 and (EC.2), equivalently, we have $z_t, \beta_t = \arg\max_{z\in\mathcal{Z}} \max_{\beta\in\mathcal{B}_k} \det[H_{t-1}(\hat{\theta}_{t-1}) + \dot{\mu}(\beta \hat{\theta}_{t-1}^\top z)\beta^2 z z^\top]$ with k being the type of z.

$$z_{t+1}, \beta_{t+1} = \arg\max_{z\in\mathcal{Z}} \max_{\beta\in\mathcal{B}_k} \dot{\mu}(\beta \hat{\theta}_t^\top z)\beta^2 z^\top M^{-1}(\xi_t, \hat{\theta}_t) z \text{ with } k \text{ being the type of } z. \tag{EC.33}$$

We define

$$z_{t+1}^*, \beta_{t+1}^* = \arg\max_{z\in\mathcal{Z}} \max_{\beta\in\mathcal{B}_k} \dot{\mu}(\beta z^\top \theta_*)\beta^2 z^\top M^{-1}(\xi_t, \theta_*) z \text{ with } k \text{ being the type of } z. \tag{EC.34}$$

Then, for all $t \geq t_1$, we have

$$\begin{aligned}
\dot{\mu}(\beta_{t+1} z_{t+1}^\top \theta_*)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t, \theta_*) z_{t+1} &\geq \dot{\mu}(\beta_{t+1} z_{t+1}^\top \hat{\theta}_t)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t, \hat{\theta}_t) z_{t+1} - \frac{\epsilon}{2} \\
&\geq \dot{\mu}(\beta_{t+1}^* \hat{\theta}_t^\top z_{t+1}^*)\beta_{t+1}^{*2} z_{t+1}^{*\top} M^{-1}(\xi_t, \hat{\theta}_t) z_{t+1}^* - \frac{\epsilon}{2} \\
&\geq \dot{\mu}(\beta_{t+1}^* \theta_*^\top z_{t+1}^*)\beta_{t+1}^{*2} z_{t+1}^{*\top} M^{-1}(\xi_t, \theta_*) z_{t+1}^* - \epsilon \\
&\geq d - \epsilon,
\end{aligned} \tag{EC.35}$$

where the first and third inequalities are from (EC.32), the second inequality is due to (EC.33), and the last inequality is from (EC.34) and the Kiefer–Wolfowitz equivalence theorem (Kiefer and Wolfowitz 1960, White 1973, Freise et al. 2021). By the definition of $M(\xi_t, \theta_*)$, we have

$$(t+1)M(\xi_{t+1}, \theta_*) = tM(\xi_t, \theta_*) + \dot{\mu}(\beta_{t+1} \theta_*^\top z_{t+1})\beta_{t+1}^2 z_{t+1} z_{t+1}^\top.$$

Then, by Lemma 5 with $R = tM(\xi_t, \theta_*)$, $\tilde{T} = 1$, $S = \dot{\mu}(\beta_{t+1}\theta_*^\top z_{t+1})\beta_{t+1}^2 z_{t+1}$, $U = z_{t+1}^\top$, we obtain

$$\det[(t+1)M(\xi_{t+1}, \theta_*)] = \det[tM(\xi_t, \theta_*)]\left[1 + \frac{\dot{\mu}(\beta_{t+1} z_{t+1}^\top \theta_*)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t, \theta_*) z_{t+1}}{t}\right].$$

Therefore,

$$\det M(\xi_{t+1}, \theta_*) = \left(\frac{t}{t+1}\right)^d \det M(\xi_t, \theta_*)\left[1 + \frac{\dot{\mu}(\beta_{t+1} z_{t+1}^\top \theta_*)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t, \theta_*) z_{t+1}}{t}\right].$$

Then,

$$\begin{aligned}
&\log\det(M(\xi_{t+1}, \theta_*)) - \log\det(M(\xi_t, \theta_*)) \\
&= \log\frac{\det(M(\xi_{t+1}, \theta_*))}{\det(M(\xi_t, \theta_*))} \\
&= \log\frac{\left(\frac{t}{t+1}\right)^d \det M(\xi_t, \theta_*)\left[1 + \frac{\dot{\mu}(\beta_{t+1} z_{t+1}^\top \theta_*)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t, \theta_*) z_{t+1}}{t}\right]}{\det(M(\xi_t, \theta_*))} \\
&= \log\left[1 + \frac{\dot{\mu}(\beta_{t+1} z_{t+1}^\top \theta_*)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t, \theta_*) z_{t+1}}{t}\right] - d\log\left(1 + \frac{1}{t}\right).
\end{aligned} \tag{EC.36}$$

By (EC.35) and (EC.36), for all $t \geq t_1$, we have

$$\begin{aligned}
\log\det(M(\xi_{t+1}, \theta^*)) - \log\det(M(\xi_t, \theta^*)) &\geq \log\left(1 + \frac{d-\epsilon}{t}\right) - d\log\left(1 + \frac{1}{t}\right) \\
&= \log\frac{1 + (d-\epsilon)/t}{(1+1/t)^d}.
\end{aligned} \tag{EC.37}$$

On the other hand, we have

$$\log\frac{1 + (d-\epsilon)/t}{(1+1/t)^d} = \log\frac{1 + (d-\epsilon)/t}{1 + (d+c_t)/t}, \tag{EC.38}$$

where we have used that $(1+1/t)^d = 1 + (d+c_t)/t$ with $c_t \geq 0, c_t \to 0$ as $t \to \infty$. We choose $t_2 \geq t_1$ such that $c_t \leq (d-\epsilon)\epsilon$ for all $t \geq t_2$. Then for all $t \geq t_2$, we have

$$\begin{aligned}
\log\frac{1 + (d-\epsilon)/t}{1 + (d+c_t)/t} &\geq \log\frac{1 + (d-\epsilon)/t}{1 + [d + (d-\epsilon)\epsilon]/t} \\
&= -\log\left\{1 + \frac{1 + [d + (d-\epsilon)\epsilon]/t - 1 - (d-\epsilon)/t}{1 + (d-\epsilon)/t}\right\} \\
&= -\log\left[1 + \frac{\epsilon(1 + d - \epsilon)/t}{1 + (d-\epsilon)/t}\right] \\
&\geq -\frac{1}{1 + (d-\epsilon)/t}\frac{\epsilon(1+d-\epsilon)}{t} \\
&= -\frac{\epsilon(1+d-\epsilon)}{t + d - \epsilon} \\
&\geq -\epsilon,
\end{aligned} \tag{EC.39}$$

where the second inequality is due to the fact that $\log(1+x) \le x$ for $x \ge 0$. By (EC.37), (EC.38) and (EC.39), for all $t \ge t_2$, we conclude

$$\log\det M(\xi_{t+1}, \theta_*) - \log\det M(\xi_t, \theta_*) \ge -\epsilon. \tag{EC.40}$$

Now we choose $t_3 \ge t_2$ such that for all $t \ge t_3$,

$$\begin{aligned}
\log\left(1+\frac{d+\epsilon}{t}\right) - d\log\left(1+\frac{1}{t}\right) &= \log\left(1+\frac{d+\epsilon}{t}\right) - \log\left(1+\frac{d+c_t}{t}\right) \\
&\ge \log\left(1+\frac{d+\epsilon}{t}\right) - \log\left[1+\frac{d+\epsilon(1-\frac{t+d+\epsilon}{2t})}{t}\right] \\
&= \log\frac{t+d+\epsilon}{t+d+\epsilon(1-\frac{t+d+\epsilon}{2t})} \\
&= \log\frac{1}{1-\frac{\epsilon}{2t}} \\
&\ge \frac{\epsilon}{2t},
\end{aligned} \tag{EC.41}$$

where the first equality follows from $(1+1/t)^d = 1 + (d+c_t)/t$ with $c_t \ge 0, c_t \to 0$ as $t \to \infty$, and the first inequality is achieved by choosing $t_3 \ge t_2$ such that $c_t \le \epsilon(1-\frac{t+d+\epsilon}{2t})$ for all $t \ge t_3$, and the last inequality is due to the fact that $\log(1-x) \le -x$ for $x < 1$. Now, we propose the following lemma.

LEMMA 4. *Let* $t \ge t_3$ *and* $\epsilon \in (0,1)$. *If* $\log\det M(\xi_t, \theta_*) \le \log\det M(\xi_*, \theta_*) - 2\epsilon$, *then,* $\log\det M(\xi_{t+1}, \theta_*) - \log\det M(\xi_t, \theta_*) \ge \frac{\epsilon}{2t}$.

Proof. Since the log-determinant function $\log\det(\cdot)$ is concave on the space of symmetric positive definite matrices (Boyd and Vandenberghe 2004), by the first-order condition for the concave function, we have

$$\begin{aligned}
\log\det M(\xi_*, \theta_*) &\le \log\det M(\xi_t, \theta_*) + \langle M^{-1}(\xi_t, \theta_*), M(\xi_*, \theta_*) - M(\xi_t, \theta_*)\rangle \\
&= \log\det M(\xi_t, \theta_*) + \langle M^{-1}(\xi_t, \theta_*), \sum_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} [\xi_*(z,\beta) - \xi_t(z,\beta)]\dot{\mu}(\beta z^\top \theta_*)\beta^2 z z^\top\rangle \\
&= \log\det M(\xi_t, \theta_*) + \sum_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} [\xi_*(z,\beta) - \xi_t(z,\beta)]\dot{\mu}(\beta z^\top \theta_*)\beta^2 z^\top M^{-1}(\xi_t, \theta_*) z \\
&\le \log\det M(\xi_t, \theta_*) + \max_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} \dot{\mu}(\beta z^\top \theta_*)\beta^2 z^\top M^{-1}(\xi_t, \theta_*) z,
\end{aligned}$$

where the first equality is from the fact that $\frac{\partial \log\det M}{\partial M} = (M^{-1})^\top$ for an invertible matrix M (Harville 1997), and the last inequality is because $\sum_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} \xi(z,\beta) = 1$ with $\xi(z,\beta) \geq 0$. Therefore,

$$\begin{aligned}\log\det M(\xi_*,\theta_*) - \log\det M(\xi_t,\theta_*) &\leq \max_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} \dot{\mu}(\beta z^\top\theta_*)\beta^2 z^\top M^{-1}(\xi_t,\theta_*)z \\ &\leq \dot{\mu}(\beta_{t+1}z_{t+1}^\top\theta_*)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t,\theta_*)z_{t+1} - d + \epsilon,\end{aligned}$$

where the last inequality is from (EC.35). Using the condition $\log\det M(\xi_t,\theta_*) \leq \log\det M(\xi_*,\theta_*) - 2\epsilon$ in Lemma 4, we obtain

$$\dot{\mu}(\beta_{t+1}z_{t+1}^\top\theta^*)\beta_{t+1}^2 z_{t+1}^\top M^{-1}(\xi_t,\theta^*)z_{t+1} \geq d + \epsilon.$$

Together with (EC.36), we have

$$\log\det M(\xi_{t+1},\theta^*) - \log\det M(\xi_t,\theta^*) \geq \log\left(1+\frac{d+\epsilon}{t}\right) - d\log\left(1+\frac{1}{t}\right) \geq \frac{\epsilon}{2t},$$

where the last inequality follows from (EC.41). □

There is some $t_4 \geq t_3$ such that for all $t \geq t_4$

$$\log\det M(\xi_t,\theta_*) > \log\det M(\xi_*,\theta_*) - 2\epsilon \tag{EC.42}$$

since otherwise $\log\det M(\xi_t,\theta^*) \to \infty$ from Lemma 4, which contradicts the fact that $\log\det M(\xi_t,\theta^*)$ is a bounded value, which follows from

$$\begin{aligned}\det M(\xi_t,\theta_*) &\leq \left[\frac{\mathrm{tr}(M(\xi_t,\theta_*))}{d}\right]^d \\ &\leq \left[\frac{\sum_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} \xi_t(z,\beta)\dot{\mu}(\beta z^\top\theta_*)\beta^2\|z\|^2}{d}\right]^d \\ &\leq \left(\frac{C_\beta^2 C_z^2}{4d}\right)^d,\end{aligned}$$

where the last inequality is from Assumption 2 and the facts $0 \leq \dot{\mu}(\cdot) \leq 1/4$ and $\sum_{(z,\beta)\in\mathcal{Z}\times\mathcal{B}} \xi_t(z,\beta) = 1$. Combining (EC.40) and (EC.42), we have

$$\log\det M(\xi_{t_4+1},\theta_*) \geq \log\det M(\xi_{t_4},\theta_*) - \epsilon > \log\det M(\xi_*,\theta^*) - 3\epsilon. \tag{EC.43}$$

If $\log\det M(\xi_{t_4+1},\theta_*) \leq \log\det M(\xi_*,\theta_*) - 2\epsilon$, by Lemma 4 and (EC.43), we have

$$\log\det M(\xi_{t_4+2},\theta^*) \geq \log\det M(\xi_{t_4+1},\theta^*) > \log\det M(\xi_*,\theta^*) - 3\epsilon.$$

If $\log\det M(\xi_{t_4+1},\theta_*) > \log\det M(\xi_*,\theta_*) - 2\epsilon$, by (EC.40) and (EC.43), we have

$$\log\det M(\xi_{t_4+2},\theta_*) \geq \log\det M(\xi_{t_4+1},\theta_*) - \epsilon > \log\det M(\xi_*,\theta_*) - 3\epsilon.$$

Continuing this argument inductively, we obtain, for all $t \geq t_4$,

$$\log\det M(\xi_t,\theta_*) > \log\det M(\xi_*,\theta_*) - 3\epsilon.$$

Therefore,

$$\liminf_{t\to\infty} \log\det M(\xi_t,\theta_*) \geq \log\det M(\xi_*,\theta_*) - 3\epsilon.$$

Since $\epsilon \in (0,1)$ is arbitrary, we have

$$\liminf_{t\to\infty} \log\det M(\xi_t,\theta_*) \geq \log\det M(\xi_*,\theta_*).$$

Since $\xi_* = \arg\max_{\xi\in\mathcal{D}(\mathcal{Z},\mathcal{B})} M(\xi,\theta_*)$, we have

$$\lim_{t\to\infty} \log\det M(\xi_t,\theta_*) = \log\det M(\xi_*,\theta_*).$$

Due to the strict concavity of the criterion $\log\det(\cdot)$, the information matrix at θ_* of a locally D-optimal design at θ_* is unique. Therefore, $\lim_{t\to\infty} M(\xi_t,\theta_*) = M(\xi_*,\theta_*)$. By (EC.31), we have

$$\lim_{T\to\infty} M(\xi_T,\hat{\theta}_T) \xrightarrow{a.s.} M(\xi_*,\theta_*).$$

EC. 5.5. Proof of Theorem 3 By the definition of the sub-optimality (11), we have

$$\mathsf{SubOpt}(\hat{\pi}_T) = J(\pi^*) - J(\hat{\pi}_T) = [J(\pi^*) - \hat{J}(\pi^*)] + [\hat{J}(\pi^*) - \hat{J}(\hat{\pi}_T)] + [\hat{J}(\hat{\pi}_T) - J(\hat{\pi}_T)]. \quad \text{(EC.44)}$$

Since $\hat{\pi}_T$ is the optimal policy under $\hat{J}(\pi)$, we have

$$\hat{J}(\pi^*) - \hat{J}(\hat{\pi}_T) \leq 0. \quad \text{(EC.45)}$$

By the definition of the pessimistic expected value function (9), we obtain

$$\hat{J}(\hat{\pi}_T) - J(\hat{\pi}_T) = \min_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}\theta^\top\phi(x,\hat{\pi}_T(x)) - \mathbb{E}\theta_*^\top\phi(x,\hat{\pi}_T(x)).$$

By Lemma 1, we know that $\theta_* \in C(\hat{\theta}_T,\delta)$ with probability at least $1-\delta$. Therefore, with probability at least $1-\delta$, we have

$$\hat{J}(\hat{\pi}_T) - J(\hat{\pi}_T) \leq 0. \quad \text{(EC.46)}$$

Combining (EC.44), (EC.45) and (EC.46), with probability at least $1-\delta$, we have

$$\begin{aligned}\mathsf{SubOpt}(\hat{\pi}_T) &\le J(\pi^*) - \hat{J}(\pi^*)\\ &= \mathbb{E}\theta_*^\top \phi(x,\pi^*(x)) - \min_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}\theta^\top \phi(x,\pi^*(x))\\ &= \max_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}(\theta_*-\theta)^\top \phi(x,\pi^*(x))\\ &= \max_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}(\theta_*-\hat{\theta}_T+\hat{\theta}_T-\theta)^\top \phi(x,\pi^*(x))\\ &= \mathbb{E}(\theta_*-\hat{\theta}_T)^\top \phi(x,\pi^*(x)) + \max_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}(\hat{\theta}_T-\theta)^\top \phi(x,\pi^*(x)).\end{aligned}$$

By the definition of $C(\hat{\theta}_T,\delta)$ in (8), we obtain

$$\begin{aligned}\max_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}(\hat{\theta}_T-\theta)^\top \phi(x,\pi^*(x)) &\le \max_{\theta\in C(\hat{\theta}_T,\delta)} \mathbb{E}\|\hat{\theta}_T-\theta\|_{\bar{H}_T(\hat{\theta}_T)}\|\phi(x,\pi^*(x))\|_{\bar{H}_T^{-1}(\hat{\theta}_T)}\\ &\le \gamma(T,d,\delta)\mathbb{E}\|\bar{H}_T^{-1/2}(\hat{\theta}_T)\phi(x,\pi^*(x))\|_2\end{aligned}$$

By Lemma 1, we know that $\theta_*\in C(\hat{\theta}_T,\delta)$ with probability at least $1-\delta$. Therefore,

$$\mathsf{SubOpt}(\hat{\pi}_T) \le 2\gamma(T,d,\delta)\mathbb{E}\|\bar{H}_T^{-1/2}(\hat{\theta}_T)\phi(x,\pi^*(x))\|.$$

By Theorem 5, we have $\bar{H}_T(\hat{\theta}_T) = M(\xi_T,\hat{\theta}_T) \xrightarrow{a.s.} M(\xi_*,\theta_*)$. Therefore, there exists a constant T_0 such that $\mathbb{E}\|\bar{H}_T^{-1/2}(\hat{\theta}_T)\phi(x,\pi^*(x))\|_2 \le 2\|M^{-1/2}(\xi_*,\theta_*)\mathbb{E}\phi(x,\pi^*(x))\|_2$ for all $T>T_0$ with probability 1. Thus, when $T>T_0$, with probability at least $1-\delta$, we have

$$\begin{aligned}\mathsf{SubOpt}(\hat{\pi}_T) &\le 2\gamma(T,d,\delta)\|M^{-1/2}(\xi_*,\theta_*)\mathbb{E}\phi(x,\pi^*(x))\|\\ &= 2\sqrt{\frac{C_1}{T}\left[d\log\left(e+\frac{C_2T}{d}\right)+\log\frac{2}{\delta}\right]}\|M^{-1/2}(\xi_*,\theta_*)\mathbb{E}\phi(x,\pi^*(x))\|_2.\end{aligned}$$

EC. 5.6. Proof of Theorem 4 The proof of Theorem 4 follows a similar strategy to that of Theorem 3. For completeness, we provide the proof here. By the definition of the sub-optimality (11), we have

$$\mathsf{SubOpt}(\hat{\pi}_T) = J(\pi^*) - J(\hat{\pi}_T) = [J(\pi^*)-\hat{J}(\pi^*)] + [\hat{J}(\pi^*)-\hat{J}(\hat{\pi}_T)] + [\hat{J}(\hat{\pi}_T)-J(\hat{\pi}_T)]. \quad \text{(EC.47)}$$

Since $\hat{\pi}_T$ is the optimal policy under $\hat{J}(\pi)$, we have

$$\hat{J}(\pi^*) - \hat{J}(\hat{\pi}_T) \le 0. \quad \text{(EC.48)}$$

By the definition of the pessimistic expected value function (9), we obtain

$$\hat{J}(\hat{\pi}_T) - J(\hat{\pi}_T) = \min_{\theta \in C(\hat{\theta}_T, \delta)} \mathbb{E}\theta^\top \phi(x, \hat{\pi}_T(x)) - \mathbb{E}\theta_*^\top \phi(x, \hat{\pi}_T(x)).$$

As in the proof of Lemma 1, we can show that $\theta_* \in C(\hat{\theta}_T, \delta)$ with probability at least $1-\delta$. Therefore, with probability at least $1-\delta$, we have

$$\hat{J}(\hat{\pi}_T) - J(\hat{\pi}_T) \le 0. \tag{EC.49}$$

Combining (EC.47), (EC.48) and (EC.49), with probability at least $1-\delta$, we have

$$\begin{aligned}
\mathsf{SubOpt}(\hat{\pi}_T) &\le J(\pi^*) - \hat{J}(\pi^*) \\
&= \mathbb{E}\theta_*^\top \phi(x, \pi^*(x)) - \min_{\theta \in C(\hat{\theta}_T, \delta)} \mathbb{E}\theta^\top \phi(x, \pi^*(x)) \\
&= \max_{\theta \in C(\hat{\theta}_T, \delta)} \mathbb{E}(\theta_* - \theta)^\top \phi(x, \pi^*(x)) \\
&= \max_{\theta \in C(\hat{\theta}_T, \delta)} \mathbb{E}(\theta_* - \hat{\theta}_T + \hat{\theta}_T - \theta)^\top \phi(x, \pi^*(x)) \\
&= \mathbb{E}(\theta_* - \hat{\theta}_T)^\top \phi(x, \pi^*(x)) + \max_{\theta \in C(\hat{\theta}_T, \delta)} \mathbb{E}(\hat{\theta}_T - \theta)^\top \phi(x, \pi^*(x)).
\end{aligned}$$

By the definition of $C(\hat{\theta}_T, \delta)$ in (8), we obtain

$$\begin{aligned}
\max_{\theta \in C(\hat{\theta}_T, \delta)} \mathbb{E}(\hat{\theta}_T - \theta)^\top \phi(x, \pi^*(x)) &\le \max_{\theta \in C(\hat{\theta}_T, \delta)} \mathbb{E}\|\hat{\theta}_T - \theta\|_{\bar{H}_T(\hat{\theta}_T)} \|\phi(x, \pi^*(x))\|_{\bar{H}_T^{-1}(\hat{\theta}_T)} \\
&\le \gamma(T, d, \delta) \mathbb{E}\|\bar{H}_T^{-1/2}(\hat{\theta}_T)\phi(x, \pi^*(x))\|
\end{aligned}$$

As in the proof of Lemma 1, we can show that $\theta_* \in C(\hat{\theta}_T, \delta)$ with probability at least $1-\delta$. Therefore,

$$\mathsf{SubOpt}(\hat{\pi}_T) \le 2\gamma(T, d, \delta) \mathbb{E}\|\bar{H}_T^{-1/2}(\hat{\theta}_T)\phi(x, \pi^*(x))\|.$$

As in the proof of Theorem 5, we can show $\bar{H}_T(\hat{\theta}_T) = M(\xi_T, \hat{\theta}_T) \xrightarrow{a.s.} M(\xi_*, \theta_*)$. Therefore, there exists a constant T_0 such that $\mathbb{E}\|\bar{H}_T^{-1/2}(\hat{\theta}_T)\phi(x, \pi^*(x))\| \le 2\|M^{-1/2}(\xi_*, \theta_*)\mathbb{E}\phi(x, \pi^*(x))\|$ for all $T > T_0$ with probability 1. Thus, when $T > T_0$, with probability at least $1-\delta$, we have

$$\begin{aligned}
\mathsf{SubOpt}(\hat{\pi}_T) &\le 2\gamma(T, d, \delta) \|M^{-1/2}(\xi_*, \theta_*)\mathbb{E}\phi(x, \pi^*(x))\| \\
&= 2\sqrt{\frac{C_3}{T}\left[d \log\left(e + \frac{C_4 T}{d}\right) + \log\frac{2}{\delta}\right]} \|M^{-1/2}(\xi_*, \theta_*)\mathbb{E}\phi(x, \pi^*(x))\|,
\end{aligned}$$

for some positive constants C_3 and C_4.

EC. 6. Supporting Lemmas

LEMMA 5. *(Theorem 18.1.1. (Harville 1997)) Let R represent an $n \times n$ matrix, S an $n \times m$ matrix, $\tilde{T}$ an $m \times m$ matrix, and U an $m \times n$ matrix. If R and $\tilde{T}$ are nonsingular, then*

$$\det(R + S\tilde{T}U) = \det R \det \tilde{T} \det(\tilde{T}^{-1} + UR^{-1}S).$$

LEMMA 6. *(Theorem 1.1 (Tropp 2012)) Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d. Assume that each random matrix satisfies*

$$\mathbf{X}_k \succeq \mathbf{0} \text{ and } \lambda_{max}(\mathbf{X}_k) \le R \text{ almost surely.}$$

Define

$$\mu_{min} := \lambda_{min}\left(\sum_k \mathbb{E}\mathbf{X}_k\right) \text{ and } \mu_{max} := \lambda_{max}\left(\sum_k \mathbb{E}\mathbf{X}_k\right).$$

Then for $\zeta \in [0,1]$,

$$\mathbb{P}\left\{\lambda_{min}\left(\sum_k \mathbf{X}_k\right) \le (1-\delta)\mu_{min}\right\} \le d\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{min}/R} \text{ for } \delta \in [0,1], \text{ and}$$

$$\mathbb{P}\left\{\lambda_{max}\left(\sum_k \mathbf{X}_k\right) \le (1+\delta)\mu_{max}\right\} \le d\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{max}/R} \text{ for } \delta \ge 0.$$

LEMMA 7. *(Lemma 4 (Pronzato 2010)) If for any $\delta > 0$*

$$\liminf_{N\to\infty} \inf_{\|\theta-\theta_*\|\ge\delta} [L_N(\theta_*) - L_N(\theta)] > 0 \text{ almost surely,}$$

then $\hat{\theta}^N_{ML} \xrightarrow{a.s.} \theta_*$.

LEMMA 8. *(Theorem 3.2 (Hall and Heyde 1980)) Let $\{S_{ni}, \mathcal{F}_{n,i}, 1 \le i \le k_n, n \ge 1\}$ be a zero-mean, square-integrable martingale array with differences X_{ni}, and let η^2 be an a.s. finite r.v. Suppose that*

$$\max_i |X_{ni}| \xrightarrow{p} 0, \tag{EC.50}$$

$$\sum_i X^2_{ni} \xrightarrow{p} \eta^2, \tag{EC.51}$$

$$\mathbb{E}\max_i X^2_{ni} \text{ is bounded in } n, \tag{EC.52}$$

and the σ-fields are nested:

$$\mathcal{F}_{n,i} \subseteq \mathcal{F}_{n+1,i} \text{ for } 1 \le i \le k_n, n \ge 1. \tag{EC.53}$$

Then $S_{nk_n} = \sum_i X_{ni} \xrightarrow{d} Z$ (stably), where the r.v. Z has characteristic function $\mathbb{E}e^{-\frac{1}{2}\eta^2 t^2}$.

LEMMA 9. *(Corollary 3.1 (Hall and Heyde 1980)) If* (EC.50) *and* (EC.52) *are replaced by the conditional Lindeberg condition*

$$\text{for all } \epsilon > 0,\ \sum_i \mathbb{E}[X_{ni}^2 \mathbb{I}(|X_{ni}| > \epsilon)|\mathcal{F}_{n,i-1}] \xrightarrow{p} 0,$$

if (EC.51) *is replaced by an analogous condition on the conditional variance:*

$$V_{nk_n}^2 = \sum \mathbb{E}(X_{ni}^2|\mathcal{F}_{n,i-1}) \xrightarrow{p} \eta^2,$$

and if (EC.53) *holds, then the conclusion of Lemma 8 remains true.*

LEMMA 10. *(Theorem 29.4 (Billingsley 1995)) For random vectors* $X_n = (X_{n1}, \cdots, X_{nk})$ *and* $Y = (Y_1, \cdots, Y_k)$*, a necessary and sufficient condition for* $X_n \xrightarrow{d} Y$ *is that* $\sum_{u=1}^k t_u X_{nu} \xrightarrow{d} \sum_{u=1}^k t_u Y_u$ *for each* $(t_1, \cdots, t_k) \in \mathbb{R}^k$.

LEMMA 11. *(Lemma 3 (Lee et al. 2024)) Let* $X_1, \cdots, X_t$ *be a martingale difference sequence satisfying* $\max_s |X_s| \le R$ *a.s., and let* $\mathcal{F}_s$ *be the* σ*-field generated by* $(X_1, \cdots, X_s)$*. Then for any* $\delta \in (0,1)$ *and any* $\eta \in (0, 1/R]$*, the following holds with probability at least* $1-\delta$*:*

$$\sum_{s=1}^t X_s \le (e-2)\eta \sum_{s=1}^t \mathbb{E}(X_s^2|\mathcal{F}_{s-1}) + \frac{1}{\eta}\log\frac{1}{\delta}, \forall t \ge 1.$$

LEMMA 12. *(Lemma 8 (Das et al. 2025)) Let* $z, z' \in \mathbb{R}$ *and* $\tilde{\alpha}(z,z') := \int_0^1 (1-v)\dot{\mu}(z + v(z'-z))dv$*. Then for some* $C > 1$ *(1.01 suffices),*

$$\tilde{\alpha}(z,z') \ge \frac{\dot{\mu}(z')}{C(2+|z-z'|)^2}.$$