A Algorithm and Theory for Unobserved Confounder

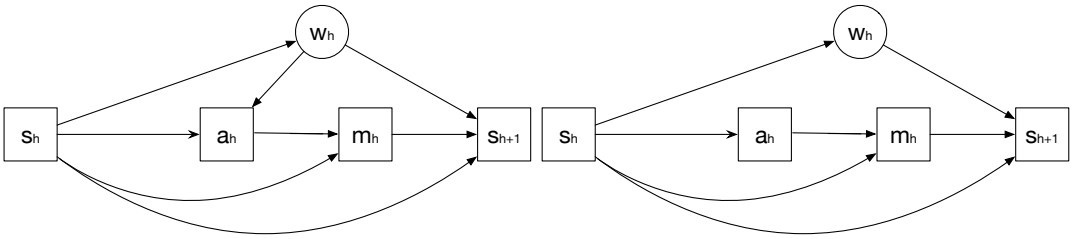
In this section, we extend DOVI to handle the case where the confounders are unobserved in both the online setting and the offline setting. We then characterize the regret of such an extension of DOVI, namely DOVI$^+$. In comparison with DOVI, DOVI$^+$ additionally incorporates an intermediate state at each step, which extends the length of each episode from H to $2H$.

A.1 Algorithm

Frontdoor Adjustment. Since the confounders $\{w_h\}_{h\in[H]}$ are unobserved in the offline setting, the confounded observational data $\{(s_h^i, a_h^i, r_h^i)\}_{(i,h)\in[n]\times[H]}$ are insufficient for the identification of the causal effect $\mathbb{P}(s_{h+1} \,|\, s_h, \mathrm{do}(a_h))$ [32, 33]. However, such a causal effect is identifiable if we observe the intermediate states $\{m_h\}_{h\in[H]}$ that satisfy the following frontdoor criterion.

Assumption A.1 (Frontdoor Criterion [32, 33])**.** In the SCM defined in §2, for all $h \in [H]$, there additionally exists an observed intermediate state m_h that satisfies the frontdoor criterion, that is,

- m_h intercepts every directed path from a_h to s_{h+1},
- conditioning on s_h, no path between a_h and m_h has an incoming arrow into a_h, and
- conditioning on s_h, a_h d-separates every path between m_h and s_{h+1} that has an incoming arrow into m_h.



(a) Offline Setting (b) Online Setting

Figure 3: Causal diagrams of the h-th step of the confounded MDP with the intermediate state (a) in the offline setting and (b) in the online setting, respectively.

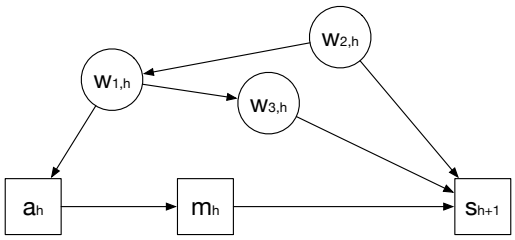


Figure 4: An illustration of the frontdoor criterion. The causal diagram corresponds to the h-th step of the confounded MDP conditioning on s_h. Here $w_h = \{w_{1,h}, w_{2,h}, w_{3,h}\}$ is the confounder and the intermediate state m_h satisfies the frontdoor criterion.

See Figure 3 for the causal diagram that describes such an SCM and Figure 4 for an example that satisfies the frontdoor criterion. Intuitively, Assumption A.1 ensures that, conditioning on s_h, (i) the intermediate state m_h is caused by the action a_h and the causal effect of the action a_h on the next state s_{h+1} is summarized by m_h, while (ii) the action a_h and the intermediate state m_h are not confounded. In the sequel, we denote by $\mathcal{M}$ the space of intermediate states and $\breve{\mathcal{P}}_h(\cdot \,|\, \cdot, \cdot)$ the transition kernel that determines m_h given s_h and a_h. The causal effect $\mathbb{P}(s_{h+1} \,|\, s_h, \mathrm{do}(a_h))$ is identified as follows.

Proposition A.2 (Frontdoor Adjustment [32])**.** Under Assumption A.1, it holds that

$$\mathbb{P}\big(s_{h+1} \,\big|\, s_h, \mathrm{do}(a_h)\big) = \mathbb{E}_{m_h, a_h'}\big[\mathbb{P}(s_{h+1} \,|\, s_h, a_h', m_h)\big],$$

where the expectation $\mathbb{E}_{m_h, a_h'}$ is taken with respect to $m_h \sim \breve{\mathcal{P}}_h(\cdot \,|\, s_h, a_h)$ and $a_h' \sim \mathbb{E}_{w_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}[\nu_h(\cdot \,|\, s_h, w_h)]$. Here (s_{h+1}, s_h, a_h, m_h) follows the SCM define in §2 with the intermediate states $\{m_h\}_{h \in [H]}$ in the offline setting.

Frontdoor-Adjusted Bellman Equation. In the sequel, we assume without loss of generality that the reward r_h is deterministic and only depends on the state s_h and the action a_h. In parallel to (3.3), we have

$$Q_h^\pi(s_h, a_h) = r_h(s_h, a_h) + \mathbb{E}_{s_{h+1}}\big[V_{h+1}^\pi(s_{h+1})\big], \tag{A.1}$$

where the expectation $\mathbb{E}_{s_{h+1}}$ is taken with respect to $s_{h+1} \sim \mathbb{P}(\cdot \,|\, s_h, \mathrm{do}(a_h))$. We define the the following transition operators,

$$(\mathbb{P}_{h+1/2} V)(s_h, m_h) = \mathbb{E}_{s_{h+1} \sim \mathbb{P}(\cdot \,|\, s_h, \mathrm{do}(m_h))}\big[V(s_{h+1})\big], \quad \forall V : \mathcal{S} \mapsto \mathbb{R},\ (s_h, m_h) \in \mathcal{S} \times \mathcal{M},$$
$$(\mathbb{P}_h \widetilde{V})(s_h, a_h) = \mathbb{E}_{m_h \sim \mathbb{P}(\cdot \,|\, s_h, \mathrm{do}(a_h))}\big[\widetilde{V}(s_h, m_h)\big], \quad \forall \widetilde{V} : \mathcal{S} \times \mathcal{M} \mapsto \mathbb{R},\ (s_h, a_h) \in \mathcal{S} \times \mathcal{A}.$$

We highlight that, under Assumption A.1, the causal effect $\mathbb{P}(m_h \,|\, s_h, \mathrm{do}(a_h))$ coincides with the conditional probability $\mathbb{P}(m_h \,|\, s_h, a_h)$, since a_h and m_h are not confounded given s_h. In the sequel, we define the value function at the intermediate state by $V_{h+1/2}^\pi(s_h, m_h) = (\mathbb{P}_{h+1/2} V_{h+1}^\pi)(s_h, m_h)$. We have the following Bellman equation,

$$\begin{aligned} Q_h^\pi(s_h, a_h) &= r_h(s_h, a_h) + \big(\mathbb{P}_h(\mathbb{P}_{h+1/2} V_{h+1}^\pi)\big)(s_h, a_h) \\ &= r_h(s_h, a_h) + (\mathbb{P}_h V_{h+1/2}^\pi)(s_h, a_h). \end{aligned} \tag{A.2}$$

Correspondingly, the Bellman optimality equation takes the following form,

$$\begin{aligned} &Q_h^*(s_h, a_h) = r_h(s_h, a_h) + (\mathbb{P}_h V_{h+1/2}^*)(s_h, a_h), \\ &V_{h+1/2}^*(s_h, m_h) = (\mathbb{P}_{h+1/2} V_{h+1}^*)(s_h, m_h), \quad V_h^*(s_h) = \max_{a_h \in \mathcal{A}} Q_h^*(s_h, a_h). \end{aligned} \tag{A.3}$$

Linear Function Approximation. In parallel to Assumption 3.3, we focus on the following setting with linear transition kernels and reward functions [7, 16, 42, 43], which corresponds to a linear SCM [33].

Assumption A.3 (Linear Confounded MDP)**.** We assume that

$$\mathcal{P}_h(s_{h+1} \,|\, s_h, m_h, w_h) = \langle \rho_h(s_h, m_h, w_h), \mu_h(s_{h+1}) \rangle, \quad \forall h \in [H],\ (s_h, m_h, w_h) \in \mathcal{S} \times \mathcal{M} \times \mathcal{W},$$
$$\breve{\mathcal{P}}_h(m_h \,|\, s_h, a_h) = \langle \gamma_h(s_h, a_h), \overline{\mu}_h(m_h) \rangle, \quad \forall h \in [H],\ (m_h, s_h, a_h) \in \mathcal{M} \times \mathcal{S} \times \mathcal{A}.$$

where $\rho_h(\cdot, \cdot, \cdot)$, $\gamma_h(\cdot, \cdot)$, $\mu_h(\cdot) = (\mu_{1,h}(\cdot), \ldots, \mu_{d,h}(\cdot))^\top$, and $\overline{\mu}_h(\cdot) = (\overline{\mu}_{1,h}(\cdot), \ldots, \overline{\mu}_{d,h}(\cdot))^\top$ are $\mathbb{R}^d$-valued functions. We assume that $\|\rho_h(s_h, m_h, w_h)\|_2 \le 1$, $\|\gamma_h(s_h, a_h)\|_2 \le 1$, $\sum_{i=1}^d \|\mu_{i,h}\|_1^2 \le d$, and $\sum_{i=1}^d \|\overline{\mu}_{i,h}\|_1^2 \le d$ for all $h \in [H]$ and $(s_h, a_h, m_h, w_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{M} \times \mathcal{W}$. Meanwhile, we assume that

$$r_h(s_h, a_h) = \gamma_h(s_h, a_h)^\top \theta_h, \quad \forall (h, k) \in [H] \times [K],$$

where $\theta_h \in \mathbb{R}^d$ and $\|\theta_h\|_2 \le \sqrt{d}$ for all $h \in [H]$.

Proposition A.4. We define $\widetilde{\nu}_h(a_h \,|\, s_h) = \mathbb{E}_{w_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}[\nu_h(a_h \,|\, s_h, w_h)]$, where $\nu = \{\nu_h\}_{h \in [H]}$ is the behavior policy. With a slight abuse of notation, we define the frontdoor-adjusted feature as follows,

$$\phi_h(s_h, a_h, m_h) = \frac{\mathbb{E}_{w_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}\big[\rho_h(s_h, m_h, w_h) \cdot \nu_h(a_h \,|\, s_h, w_h)\big]}{\widetilde{\nu}_h(a_h \,|\, s_h)}, \quad \forall h \in [H]. \tag{A.4}$$

Under Assumption A.3, it holds that

$$\mathbb{P}(s_{h+1} \,|\, s_h, a_h, m_h) = \langle \phi_h(s_h, a_h, m_h), \mu_h(s_{h+1}) \rangle. \tag{A.5}$$

Proof. See §F.2 for a detailed proof. □

Algorithm 2 DOVI$^+$ for Confounded MDP.

Require: Observational data $\{(s_h^i, a_h^i, m_h^i, r_h^i)\}_{i\in[n],h\in[H]}$, tuning parameters $\lambda, \beta > 0$, features $\{\phi_h\}_{h\in[H]}$ and $\{\psi_h\}_{h\in[H]}$, which are defined in (A.4) and (A.6), respectively.

1: **Initialization:** Set $\{Q_h^0, V_{h+1/2}^0, V_h^0\}_{h\in[H]}$ as zero functions and V_{H+1}^k as a zero function for $k \in [K]$.
2: **for** $k = 1, \ldots, K$ **do**
3: **for** $h = H, \ldots, 1$ **do**
4: **Update** $V_{h+1/2}^k$**:**
5: Set $\omega_{1,h}^k \leftarrow \operatorname{argmin}_{\omega\in\mathbb{R}^d} \sum_{\tau=1}^{k-1}(V_{h+1}^\tau(s_{h+1}^\tau) - \omega^\top\psi_h(s_h^\tau, m_h^\tau))^2 + \lambda\|\omega\|_2^2 + L_{1,h}^k(\omega)$, where $L_{1,h}^k$ is defined in (A.9).
6: Set $V_{h+1/2}^k(s_h, m_h) \leftarrow \min\{\psi_h(s_h, m_h)^\top\omega_{1,h}^k + \Gamma_{h+1/2}^k(s_h, m_h), H - h\}$ for all $(s_h, m_h) \in \mathcal{S}\times\mathcal{M}$, where $\Gamma_{h+1/2}^k$ is defined in (A.12).
7: **Update** Q_h^k**:**
8: Set $\omega_{2,h}^k \leftarrow \operatorname{argmin}_{\omega\in\mathbb{R}^d} \sum_{\tau=1}^{k-1}(r_h^k + V_{h+1/2}^k(s_h^\tau, m_h^\tau) - \omega^\top\gamma_h(s_h^\tau, a_h^\tau))^2 + \lambda\|\omega\|_2^2 + L_{2,h}^k(\omega)$, where $L_{2,h}^k$ is defined in (A.14).
9: Set $Q_h^k(s_h, a_h) \leftarrow \min\{\gamma_h(s_h, a_h)^\top\omega_{2,h}^k + \Gamma_h^k(s_h, a_h), H - h\}$ for all $(s_h, a_h) \in \mathcal{S}\times\mathcal{A}$, where Γ_h^k is defined in (A.15).
10: **Update** π_h^k **and** V_h^k**:**
11: Set $\pi_h^k(\cdot \mid s_h) \leftarrow \operatorname{argmax}_{a_h\in\mathcal{A}} Q_h^k(s_h, a_h)$ for all $s_h \in \mathcal{S}$.
12: Set $V_h^k(\cdot) \leftarrow \langle\pi_h^k(\cdot \mid \cdot), Q_h^k(\cdot, \cdot)\rangle_{\mathcal{A}}$.
13: **end for**
14: Obtain s_1^k from the environment.
15: **for** $h = 1, \ldots, H$ **do**
16: Take $a_h^k \sim \pi_h^k(\cdot \mid s_h^k)$. Obtain $r_h^k = r_h(s_h^k, a_h^k)$, m_h^k, and s_{h+1}^k.
17: **end for**
18: **end for**

DOVI$^+$: Update of $V_{h+1/2}^k$. With a slight abuse of notation, we define the following feature,

$$\psi_h(s_h, m_h) = \mathbb{E}_{w_h\sim\widetilde{\mathcal{P}}_h(\cdot \mid s_h)}\big[\rho_h(s_h, m_h, w_h)\big]. \tag{A.6}$$

Conditioning on the state s_h, the confounder w_h satisfies the backdoor criterion for identifying the causal effect $\mathbb{P}(s_{h+1} \mid s_h, \mathrm{do}(m_h))$, although it is unobserved. In the sequel, we assume that either the density of $\{\widetilde{\mathcal{P}}_h(\cdot \mid s_h)\}_{h\in[H]}$ is known to us or the features $\{\phi_h\}_{h\in[H]}$ and $\{\psi_h\}_{h\in[H]}$ are known to us. Following from (A.6), Proposition 3.2, and Assumption A.3, it holds for all $h \in [H]$ and $(s_{h+1}, s_h, m_h) \in \mathcal{S}\times\mathcal{S}\times\mathcal{M}$ that

$$\mathbb{P}\big(s_{h+1} \,\big|\, s_h, \mathrm{do}(m_h)\big) = \langle\psi_h(s_h, m_h), \mu_h(s_{h+1})\rangle. \tag{A.7}$$

Hence, by the Bellman equation and the Bellman optimality equation in (A.2) and (A.3), respectively, the value functions at the intermediate state $V_{h+1/2}^\pi$ and $V_{h+1/2}^*$ are linear in the feature ψ_h for all π. To solve for $V_{h+1/2}^*$ in the Bellman optimality equation in (A.3), we minimize the following empirical mean-squared Bellman error as follows at each step,

$$\omega_{1,h}^k \leftarrow \operatorname*{argmin}_{\omega\in\mathbb{R}^d} \sum_{\tau=1}^{k-1}\big(V_{h+1}^\tau(s_{h+1}^\tau) - \omega^\top\psi_h(s_h^\tau, m_h^\tau)\big)^2 + \lambda\|\omega\|_2^2 + L_{1,h}^k(\omega), \quad h = H, \ldots, 1, \tag{A.8}$$

where we set $V_{H+1}^k = 0$ for all $k \in [K]$ and V_{h+1}^τ is defined in Line 12 of Algorithm 2 for all $(\tau, h) \in [K]\times[H-1]$. Here k is the index of episode, $\lambda > 0$ is a tuning parameter, and $L_{1,h}^k$ is a regularizer, which is constructed based on the confounded observational data. More specifically, we define

$$L_{1,h}^k(\omega) = \sum_{i=1}^n\big(V_{h+1}^\tau(s_{h+1}^i) - \omega^\top\phi_h(s_h^i, a_h^i, m_h^i)\big)^2, \quad \forall(k, h) \in [K]\times[H], \tag{A.9}$$

which corresponds to the least-squares loss for regressing $V_{h+1}^{\tau}(s_{h+1}^i)$ against $\phi_h(s_h^i, a_h^i, m_h^i)$ for all $i \in [n]$. Here $\{(s_h^i, a_h^i, m_h^i, r_h^i)\}_{(i,h)\in[n]\times[H]}$ are the confounded observational data, where $s_{h+1}^i \sim \mathcal{P}_h(\cdot \mid s_h^i, a_h^i, w_h^i)$, $m_h^i \sim \check{\mathcal{P}}_h(\cdot \mid s_h^i, a_h^i)$, and $a_h^i \sim \nu_h(\cdot \mid s_h^i, w_h^i)$ with $\nu = \{\nu_h\}_{h\in[H]}$ being the behavior policy.

The update in (A.8) takes the following explicit form,

$$\omega_{1,h}^k \leftarrow (\Lambda_{1,h}^k)^{-1}\bigg(\sum_{\tau=1}^{k-1}\psi_h(s_h^\tau, m_h^\tau)\cdot V_{h+1}^k(s_{h+1}^\tau) + \sum_{i=1}^{n}\phi_h(s_h^i, a_h^i, m_h^i)\cdot V_{h+1}^k(s_{h+1}^i)\bigg), \quad \text{(A.10)}$$

where

$$\Lambda_{1,h}^k = \sum_{\tau=1}^{k-1}\psi_h(s_h^\tau, m_h^\tau)\psi_h(s_h^\tau, m_h^\tau)^\top + \sum_{i=1}^{n}\phi_h(s_h^i, a_h^i, m_h^i)\phi_h(s_h^i, a_h^i, m_h^i)^\top + \lambda I. \quad \text{(A.11)}$$

Meanwhile, we employ the following UCB of $\psi_h(s_h^k, m_h^k)^\top\omega_{1,h}^k$ for all $(s_h^k, m_h^k) \in \mathcal{S}\times\mathcal{M}$,

$$\Gamma_{h+1/2}^k(s_h^k, m_h^k) = \beta\cdot\Big(\log\det\big(\Lambda_{1,h}^k + \psi_h(s_h^k, m_h^k)\psi_h(s_h^k, m_h^k)^\top\big) - \log\det(\Lambda_{1,h}^k)\Big)^{1/2}. \quad \text{(A.12)}$$

The update of $V_{h+1/2}^k$ is defined in Line 6 of Algorithm 2.

DOVI⁺: Update of Q_h^k. Upon obtaining $V_{h+1/2}^k$, we solve for Q_h^k by minimizing the following empirical mean-squared Bellman error as follows at each step,

$$\begin{aligned}\omega_{2,h}^k \leftarrow \operatorname*{argmin}_{\omega\in\mathbb{R}^d}\sum_{\tau=1}^{k-1}\big(r_h^k + V_{h+1/2}^k(s_h^\tau, m_h^\tau) - \omega^\top\gamma_h(s_h^\tau, a_h^\tau)\big)^2 \\ + \lambda\|\omega\|_2^2 + L_{2,h}^k(\omega), \quad h = H, \ldots, 1. \end{aligned} \quad \text{(A.13)}$$

Here $L_{2,h}^k$ is a regularizer, which is defined as follows,

$$L_{2,h}^k(\omega) = \sum_{i=1}^{n}\big(r_h^i + V_{h+1/2}^k(s_h^i, m_h^i) - \omega^\top\gamma_h(s_h^i, a_h^i)\big)^2, \quad \forall(k,h)\in[K]\times[H]. \quad \text{(A.14)}$$

The update in (A.13) takes the following explicit form,

$$\omega_{2,h}^k \leftarrow (\Lambda_{2,h}^k)^{-1}\bigg(\sum_{\tau=1}^{k-1}\gamma_h(s_h^\tau, a_h^\tau)\cdot\big(V_{h+1/2}^k(s_h^\tau, m_h^\tau) + r_h^\tau\big) + \sum_{i=1}^{n}\gamma_h(s_h^i, a_h^i)\cdot\big(V_{h+1/2}^k(s_h^i, m_h^i) + r_h^i\big)\bigg),$$

where

$$\Lambda_{2,h}^k = \sum_{\tau=1}^{k-1}\gamma_h(s_h^\tau, a_h^\tau)\gamma_h(s_h^\tau, a_h^\tau)^\top + \sum_{i=1}^{n}\gamma_h(s_h^i, a_h^i)\gamma_h(s_h^i, a_h^i)^\top + \lambda I.$$

We employ the following UCB of $\gamma_h(s_h^k, a_h^k)^\top\omega_{2,h}^k$ for all $(s_h^k, a_h^k) \in \mathcal{S}\times\mathcal{A}$,

$$\Gamma_h^k(s_h^k, a_h^k) = \beta\cdot\Big(\log\det\big(\Lambda_{2,h}^k + \gamma_h(s_h^k, a_h^k)\gamma_h(s_h^k, a_h^k)^\top\big) - \log\det(\Lambda_{2,h}^k)\Big)^{1/2}. \quad \text{(A.15)}$$

The update of Q_h^k is defined in Line 9 of Algorithm 2.

A.2 Theory

In parallel to Theorem 3.5, the following theorem characterizes the regret of DOVI⁺, which is defined in (2.3)

Theorem A.5 (Regret of DOVI⁺)**.** Let $\beta = CdH\sqrt{\log(d(T+nH)/\zeta)}$ and $\lambda = 1$, where $C > 0$ and $\zeta \in (0, 1]$ are absolute constants. Under Assumptions A.1 and A.3, it holds with probability at least $1 - 5\zeta$ that

$$\text{Regret}(T) \le C'\cdot(\Delta_{1,H} + \Delta_{2,H})\cdot\sqrt{d^3H^3T}\cdot\sqrt{\log\big(d(T+nH)/\zeta\big)},$$

where $C' > 0$ is an absolute constant and

$$\Delta_{1,H} = \frac{1}{\sqrt{dH^2}} \sum_{h=1}^{H} \big(\log\det(\Lambda_{1,h}^{K+1}) - \log\det(\Lambda_{1,h}^{1})\big)^{1/2},$$
$$\Delta_{2,H} = \frac{1}{\sqrt{dH^2}} \sum_{h=1}^{H} \big(\log\det(\Lambda_{2,h}^{K+1}) - \log\det(\Lambda_{2,h}^{1})\big)^{1/2}.$$

Proof. See §F.4 for a detailed proof. □

See the discussion of Theorem 3.5 in §3, where Δ_H corresponds to $\Delta_{1,H}$ and $\Delta_{2,H}$ in Theorem A.5. In particular, $\Delta_{1,H}$ and $\Delta_{2,H}$ admit the same information-theoretic interpretation.

B Literature Review on Causal Bandit

In this section, we present literature review on causal bandit that are closely related to our work. [26] propose the causal upper confidence bound (C-UCB) and causal Thompson Sampling (C-TS) algorithms, which attain the $\sqrt{T}$-regret. [34] propose an algorithm based on importance sampling in policy evaluation. In the pure offline setting, [17, 18] propose algorithms for contextual bandit with confounders in the observational data. Their algorithms are based on the analysis of sensitivity [3, 27, 38, 44], which characterizes the worst-case difference between the causal effect and the conditional density obtained from the confounded observational data. In a combination of the online setting and the offline setting, [11] study multi-armed bandit with both the interventional data and the confounded observational data. In contrast to this line of work, we study causal RL in a combination of the online setting and the offline setting. Causal RL is more challenging than causal bandit, which corresponds to $H = 1$, as it involves the transition dynamics and is more challenging in exploration.

C Connection Between Confounded MDP and Other Extensions of MDP

In what follows, we discuss the connection between confounded MDP and other extensions of MDP and SCM.

- **Dynamic Treatment Regimes (DTR).** In a DTR [45], all the states $\{s_h\}_{h\in[H]}$ are confounded by a global confounder w, whereas in a confounded MDP, each state s_h depends on an individual confounder w_{h-1}, which further depends on the previous state s_{h-1}. If w_{h-1} does not depend on s_{h-1}, the confounded MDP reduces to a DTR by summarizing the confounders into $w = (w_1, \ldots, w_H)$. In addition, we remark that our proposed DOVI and DOVI$^+$ can handle global confounders as long as the backdoor and frontdoor criterion holds, respectively.
- **Contextual MDP (CMDP).** A confounded MDP is similar to a CMDP [12] if we cast the confounders $\{w_h\}_{h\in[H]}$ as the context therein. In a CMDP, which focuses on the online setting, the context is fixed throughout an episode, whereas in a confounded MDP, the confounders $\{w_h\}_{h\in[H]}$ vary across the H steps. Moreover, in a CMDP, the goal is to minimize the regret against the globally optimal policy that depends on the context, which is a stronger benchmark than π^* in (2.3), since π^* does not depend on the confounders $\{w_h\}_{h\in[H]}$.
- **Partially Observable MDP (POMDP).** A confounded MDP is a simplified POMDP [39] if we cast the confounders $\{w_h\}_{h\in[H]}$ as the hidden states therein (assuming that the confounders are unobserved in the offline setting as in §A). A POMDP is more challenging to solve, since marginalizing over the hidden states does not yield an MDP, which is the case in a confounded MDP.

D Mechanism of Utilizing Confounded Observational Data

In this section, we discuss the mechanism of incorporating the confounded observational data.

D.1 Partially Observed Confounder

Corresponding to Line 4 of Algorithm 1, DOVI effectively estimates the causal effect $\mathbb{P}(\cdot \,|\, s_h, \mathrm{do}(a_h))$ using

$$\psi_h(s_h, a_h)^\top (\Lambda_h^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \cdot \delta_{s_{h+1}^\tau}(\cdot) + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \cdot \delta_{s_{h+1}^i}(\cdot) \bigg), \tag{D.1}$$

where we denote by $\delta_s(\cdot)$ the Dirac measure at s. To see why it works, let the tuning parameter λ be sufficiently small. By the definition of Λ_h^k in (3.10), we have

$$\begin{aligned} \mathbb{P}\big(\cdot \,\big|\, s_h, \mathrm{do}(a_h)\big) &= \langle \psi_h(s_h, a_h), \mu_h(\cdot) \rangle \\ &\approx \psi_h(s_h, a_h)^\top (\Lambda_h^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \cdot \langle \psi_h(s_h^\tau, a_h^\tau), \mu_h(\cdot) \rangle \\ &\qquad + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \cdot \langle \phi_h(s_h^i, a_h^i, u_h^i), \mu_h(\cdot) \rangle \bigg). \end{aligned} \tag{D.2}$$

Meanwhile, Assumption 3.3 and Proposition 3.4 imply

$$\begin{aligned} \mathbb{P}\big(\cdot \,\big|\, s_h, \mathrm{do}(a_h)\big) &= \langle \psi_h(s_h, a_h), \mu_h(\cdot) \rangle, \\ \mathcal{P}_h(\cdot \,|\, s_h, a_h, u_h) &= \langle \phi_h(s_h, a_h, u_h), \mu_h(\cdot) \rangle, \end{aligned}$$

which rely on the backdoor adjustment. Since s_{h+1}^τ and s_{h+1}^i in (D.1) are sampled following $\mathbb{P}(\cdot \,|\, s_h^\tau, \mathrm{do}(a_h^\tau))$ and $\mathcal{P}_h(\cdot \,|\, s_h^i, a_h^i, u_h^i)$, respectively, (D.1) approximates the right-hand side of (D.2) as its empirical version. As $k, n \to +\infty$, (D.1) converges to the right-hand side of (D.2) as well as the causal effect $\mathbb{P}(\cdot \,|\, s_h, \mathrm{do}(a_h))$.

D.2 Unobserved Confounder

If the confounders $\{w_h\}_{h \in [H]}$ are unobserved in the offline setting, the backdoor adjustment in §3 is not applicable. Alternatively, the intermediate states $\{m_h\}_{h \in [H]}$ allow us to estimate the causal effect without observing the confounders. The key is that the frontdoor criterion in Assumption A.1 implies

$$\mathbb{P}\big(s_{h+1} \,\big|\, s_h, \mathrm{do}(a_h)\big) = \int_{\mathcal{M}} \mathbb{P}\big(s_{h+1} \,\big|\, s_h, \mathrm{do}(m_h)\big) \cdot \mathbb{P}\big(m_h \,\big|\, s_h, \mathrm{do}(a_h)\big) \mathrm{d}m_h. \tag{D.3}$$

It remains to estimate $\mathbb{P}(s_{h+1} \,|\, s_h, \mathrm{do}(m_h))$ and $\mathbb{P}(m_h \,|\, s_h, \mathrm{do}(a_h))$ on the right-hand side of (D.3). Since a_h and m_h are not confounded given s_h, the causal effect $\mathbb{P}(m_h \,|\, s_h, \mathrm{do}(a_h))$ coincides with the conditional distribution $\mathbb{P}(m_h \,|\, s_h, a_h)$, which can be estimated based on the observational data. To estimate the causal effect $\mathbb{P}(s_{h+1} \,|\, s_h, \mathrm{do}(m_h))$, we utilize the backdoor adjustment in Proposition 3.2 with u_h replaced by a_h, which is enabled by Assumption A.1. More specifically, it holds that

$$\mathbb{P}\big(s_{h+1} \,\big|\, s_h, \mathrm{do}(m_h)\big) = \mathbb{E}_{a_h' \sim \mathbb{P}(\cdot \,|\, s_h)} \big[\mathcal{P}_h(s_{h+1} \,\big|\, s_h, a_h', m_h) \big]. \tag{D.4}$$

Correspondingly, we construct the value function at the intermediate state $V_{h+1/2}$ and adapt the value iteration following the Bellman optimality equation in (A.3). To estimate the value functions $\{V_{h+1/2}^k\}_{h \in [H]}$ based on the confounded observational data, we utilize the adjustment in (D.4). Corresponding to Line 5 of Algorithm 2, DOVI$^+$ effectively estimates the causal effect $\mathbb{P}(\cdot \,|\, s_h, \mathrm{do}(m_h))$ using

$$\psi_h(s_h, m_h)^\top (\Lambda_{1,h}^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, m_h^\tau) \cdot \delta_{s_{h+1}^\tau}(\cdot) + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, m_h^i) \cdot \delta_{s_{h+1}^i}(\cdot) \bigg), \tag{D.5}$$

To see why it works, let the tuning parameter λ be sufficiently small. By the definition of $\Lambda_{1,h}^k$ in (A.11), we have

$$\begin{aligned}\mathbb{P}\big(\cdot \,\big|\, s_h, \mathrm{do}(m_h)\big) &= \langle \psi_h(s_h, m_h), \mu_h(\cdot)\rangle \\ &\approx \psi_h(s_h, m_h)^\top (\Lambda_{1,h}^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, m_h^\tau) \cdot \langle \psi_h(s_h^\tau, m_h^\tau), \mu_h(\cdot)\rangle \\ &\qquad + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, m_h^i) \cdot \langle \phi_h(s_h^i, a_h^i, m_h^i), \mu_h(\cdot)\rangle \bigg). \end{aligned} \tag{D.6}$$

Meanwhile, Assumption A.3 and Proposition A.4 imply

$$\begin{aligned}\mathbb{P}\big(\cdot \,\big|\, s_h, \mathrm{do}(m_h)\big) &= \langle \psi_h(s_h, m_h), \mu_h(\cdot)\rangle, \\ \mathbb{P}(\cdot \,|\, s_h, a_h, m_h) &= \langle \phi_h(s_h, a_h, m_h), \mu_h(\cdot)\rangle.\end{aligned}$$

Since s_{h+1}^τ and s_{h+1}^i in (D.6) are sampled following $\mathbb{P}(\cdot \,|\, s_h^\tau, \mathrm{do}(m_h^\tau))$ and $\mathbb{P}(\cdot \,|\, s_h^i, a_h^i, m_h^i)$, respectively, (D.5) approximates the right-hand side of (D.6) as its empirical version. As $k, n \to +\infty$, (D.5) converges to the right-hand side of (D.6) as well as the causal effect $\mathbb{P}(\cdot \,|\, s_h, \mathrm{do}(m_h))$.

E Limitation and Future Study

In this paper, we propose confounded MDP, which captures the data generating processes in both the offline setting and the online setting as well as their mismatch due to the confounding issue. We propose DOVI and DOVI$^+$, which handles the confounding issue if backdoor or frontdoor criteria hold, respectively. Nevertheless, our work requires knowing the linear features in the transition dynamics. Moreover, our work requires taking expectations over the feature embeddings with respect to the variable for adjustment. In reality, such feature and expectation are in general unavailable. It remains unknown if efficient reinforcement learning is possible without knowning the features a priori, which we left as our future study. Moreover, our study is restricted to two types of adjustment, namely, the backdoor and frontdoor adjustment, respectively. The design of DOVI and DOVI$^+$ is tightly related to the estimation equation corresponding to the backdoor and frontdoor adjustments, respectively, which estimates the counterfactual effect of actions on the cumulative rewards. In our future study, we also want to generalize our work for general adjustment with estimation equations given.

F Proof of Main Result

F.1 Proof of Proposition 3.4

Proof. Following from Assumption 3.3 and Proposition 3.2, it holds for all $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned}\mathbb{P}\big(s_{h+1} \,|\, s_h, \mathrm{do}(a_h)\big) &= \mathbb{E}_{u_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}\big[\mathcal{P}_h(\cdot \,|\, s_h, a_h, u_h)\big] = \mathbb{E}_{u_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}\big[\langle \phi_h(s_h, a_h, u_h), \mu_h(s_{h+1})\rangle\big] \\ &= \langle \psi_h(s_h, a_h), \mu_h(s_{h+1})\rangle,\end{aligned}$$

where

$$\psi_h(s_h, a_h) = \mathbb{E}_{u_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}\big[\phi_h(s_h, a_h, u_h)\big], \quad \forall (s_h, a_H) \in \mathcal{S} \times \mathcal{A}.$$

Similarly, following from Assumption 3.3 and Proposition 3.2, it holds for all $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$ that

$$R_h(s_h, a_h) = \mathbb{E}\big[r_h \,\big|\, s_h, \mathrm{do}(a_h)\big] = \mathbb{E}_{u_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}\big[\phi_h(s_h, a_h, u_h)^\top \theta_h\big] = \psi_h(s_h, a_h)^\top \theta_h.$$

Hence, following from the Bellman equations in (3.3) and (3.4), the action-value functions Q_h^π and Q_h^* are linear in the backdoor-adjusted feature ψ_h for all π. Thus, we complete the proof of Proposition 3.4. □

F.2 Proof of Proposition A.4

Proof. It holds for all $h \in [H]$ and $(s_{h+1}, s_h, a_h, m_h) \in \mathcal{S} \times \mathcal{S} \times \mathcal{A} \times \mathcal{M}$ that

$$\begin{aligned}
&\mathbb{P}(s_{h+1}, s_h, a_h, m_h) \\
&\quad = \int_{\mathcal{W}} \mathcal{P}_h(s_{h+1} \,|\, s_h, a_u, w_h) \cdot \nu_h(a_h \,|\, s_h, w_h) \cdot \widetilde{\mathcal{P}}_h(w_h \,|\, s_h) \cdot \breve{\mathcal{P}}_h(m_h \,|\, s_h, a_h) \cdot \mathbb{P}(s_h) \mathrm{d}w_h.
\end{aligned}$$

Meanwhile, it holds for all $h \in [H]$ and $(s_h, a_h, m_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{M}$ that

$$\mathbb{P}(s_h, a_h, m_h) = \int_{\mathcal{W}} \nu_h(a_h \,|\, s_h, w_h) \cdot \widetilde{\mathcal{P}}_h(w_h \,|\, s_h) \cdot \breve{\mathcal{P}}_h(m_h \,|\, s_h, a_h) \cdot \mathbb{P}(s_h) \mathrm{d}w_h.$$

Hence, we have

$$\begin{aligned}
\mathbb{P}(s_{h+1} \,|\, s_h, a_h, m_h) &= \frac{\mathbb{P}(s_{h+1}, s_h, a_h, m_h)}{\mathbb{P}(s_h, a_h, m_h)} \\
&= \frac{\int_{\mathcal{W}} \mathcal{P}_h(s_{h+1} \,|\, s_h, a_u, w_h) \cdot \nu_h(a_h \,|\, s_h, w_h) \cdot \widetilde{\mathcal{P}}_h(w_h \,|\, s_h) \mathrm{d}w_h}{\int_{\mathcal{W}} \nu_h(a_h \,|\, s_h, w_h) \cdot \widetilde{\mathcal{P}}_h(w_h \,|\, s_h) \mathrm{d}w_h}. \qquad \text{(F.1)}
\end{aligned}$$

Meanwhile, following from Assumption A.3, we have

$$\mathcal{P}_h(s_{h+1} \,|\, s_h, a_h, w_h) = \langle \rho_h(s_h, a_h, w_h), \mu_h(s_{h+1}) \rangle. \qquad \text{(F.2)}$$

Recall that we define $\widetilde{\nu}_h(a_h \,|\, s_h) = \mathbb{E}_{w_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)}[\pi(a_h \,|\, s_h, u_h)]$. Hence, by plugging (F.2) into (F.1), we obtain that

$$\mathbb{P}(s_{h+1} \,|\, s_h, a_h, m_h) = \langle \phi_h(s_h, a_h, m_h), \mu_h(s_{h+1}) \rangle,$$

where we define for all $h \in [H]$ and $(s_h, a_h, m_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{M}$ that

$$\begin{aligned}
\phi_h(s_h, a_h, m_h) &= \frac{\int_{\mathcal{W}} \rho_h(s_h, a_u, w_h) \cdot \nu_h(a_h \,|\, s_h, w_h) \cdot \widetilde{\mathcal{P}}_h(w_h \,|\, s_h) \mathrm{d}w_h}{\int_{\mathcal{W}} \nu_h(a_h \,|\, s_h, w_h) \cdot \widetilde{\mathcal{P}}_h(w_h \,|\, s_h) \mathrm{d}w_h} \\
&= \frac{\mathbb{E}_{w_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)} \big[\rho_h(s_h, m_h, w_h) \cdot \nu_h(a_h \,|\, s_h, w_h)\big]}{\widetilde{\nu}_h(a_h \,|\, s_h)}.
\end{aligned}$$

Thus, we complete the proof of Proposition A.4. □

F.3 Proof of Theorem 3.5

Proof. We first define for all $(k, h) \in [K] \times [H]$ the model prediction error ι_h^k as follows,

$$\iota_h^k(s_h, a_h) = -Q_h^k(s_h, a_h) + R_h(s_h, a_h) + (\mathbb{P}_h V_{h+1}^k)(s_h, a_h), \quad \forall (s_h, a_h) \in \mathcal{S} \times \mathcal{A}. \qquad \text{(F.3)}$$

We define the filtrations associated with Algorithm 1 as follows.

Definition F.1 (Filtration)**.** For all $(k, h) \in [K] \times [H]$, we define $\mathcal{F}_{k,h,1}$ the σ-algebra generated by the following set,

$$\begin{aligned}
B_{k,h,1} = &\big\{(s_h^i, a_h^i, u_h^i, r_h^i)\big\}_{(i,h) \in [n] \times [H]} \cup \big\{(s_j^\tau, a_j^\tau, r_j^\tau)\big\}_{(\tau, j) \in [k-1] \times [H]} \\
&\cup \big\{(s_j^k, a_j^k, r_j^k)\big\}_{j \in [h-1]} \cup \big\{(s_h^k, a_h^k)\big\}. \qquad \text{(F.4)}
\end{aligned}$$

Similarly, we define $\mathcal{F}_{k,h,2}$ the σ-algebra generated by the following set,

$$B_{k,h,2} = B_{k,h,1} \cup \{s_{h+1}^k\} \cup \{r_h^k\}. \qquad \text{(F.5)}$$

Moreover, we define $\mathcal{F}_{0,h,2}$ the σ-algebra generated by the set $\{(s_h^i, a_h^i, u_h^i, r_h^i)\}_{(i,h) \in [n] \times [H]}$ for all $h \in [H]$. We define the timestep index as follows,

$$t(k, h, m) = 2H \cdot k + 2(h-1) + m. \qquad \text{(F.6)}$$

It then holds for $t(k, h, m) \le t(k', h', m')$ that $\mathcal{F}_{k,h,m} \subseteq \mathcal{F}_{k',h',m'}$. Hence, the set of σ-algebra $\{\mathcal{F}_{k,h,m}\}_{(k,h,m) \in [K] \times [H] \times [2]}$ is a filtration with the timestep index $t(\cdot, \cdot, \cdot)$ defined in (F.6).

The following lemma characterizes the model prediction errors defined in (F.3).

Lemma F.2. Let $\beta = CdH\sqrt{\log(d(T+nH)/\zeta)}$ and $\zeta \in (0,1]$. Under Assumption 3.3, it holds with probability at least $1-2\zeta$ that

$$-2\Gamma_h^k(s_h,a_h) \le \iota_h^k(s_h,a_h) \le 0, \quad \forall (k,h) \in [K]\times[H],\ (s_h,a_h) \in \mathcal{S}\times\mathcal{A}.$$

Proof. See §G.1 for a detailed proof. □

In the sequel, we define the following operators,

$$(\mathbb{J}_h f)(s) = \langle f(s,\cdot), \pi_h^*(\cdot\,|\,s)\rangle_{\mathcal{A}}, \qquad (\mathbb{J}_{k,h} f)(s) = \langle f(s,\cdot), \pi_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}}, \quad \forall s \in \mathcal{S}.$$

Meanwhile, recall that we define

$$(\mathbb{P}_h V)(s_h,a_h) = \mathbb{E}_{s_{h+1}\sim\mathbb{P}(\cdot\,|\,s_h,\mathrm{do}(a_h))}\big[V(s_{h+1})\big], \quad \forall (s_h,a_h) \in \mathcal{S}\times\mathcal{A}.$$

We define the following martingale adapted to the filtration $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$,

$$M_{k,h,m} = \sum_{\substack{(\tau,i,\ell)\in[K]\times[H]\times[2]\\ t(\tau,i,\ell)\le t(k,h,m)}} D_{\tau,i,\ell},$$

where

$$D_{k,h,1} = \big(\mathbb{J}_{k,h}(Q_h^k - Q_h^{\pi^k,k})\big)(s_h^k) - (Q_h^k - Q_h^{\pi^k,k})(s_h^k,a_h^k), \quad \forall (k,h)\in[K]\times[H],$$
$$D_{k,h,2} = \big(\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(s_h^k,a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k,k})(s_{h+1}^k), \quad \forall (k,h)\in[K]\times[H].$$

The following lemma is adapted from [7].

Lemma F.3 (Lemma 4.2 of [7])**.** It holds that

$$\begin{aligned}\mathrm{Regret}(T) &= \sum_{k=1}^K V_1^{\pi^*}(x_1^k) - V_1^{\pi_k}(x_1^k)\\ &= Y + \mathcal{M}_{K,H,2} + \sum_{k=1}^K\sum_{h=1}^H \Big(\mathbb{E}_{\pi^*}\big[\iota_h^k(s_h,a_h)\,\big|\,s_1 = s_1^k\big] - \iota_h^k(s_h^k,a_h^k)\Big),\end{aligned} \tag{F.7}$$

where

$$Y = \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot\,|\,s_h) - \pi_h^k(\cdot\,|\,s_h)\rangle\,\big|\,s_1 = s_1^k\big]. \tag{F.8}$$

Proof. See [7] for a detailed proof. □

In what follows, we upper bound the right-hand side of (F.7) in Lemma F.3. By Algorithm 1, it holds that π_h^k is the greedy policy with respect to the action-value function Q_h^k. Hence, for Y defined in (F.8) of Lemma F.3, we have

$$Y = \sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot\,|\,s_h) - \pi_h^k(\cdot\,|\,s_h)\rangle\,\big|\,s_1 = s_1^k\big] \le 0. \tag{F.9}$$

Meanwhile, following from the proof of Theorem 3.1 in [7], it holds with probability at least $1-\zeta/2$ that

$$M_{K,H,2} \le C_0\cdot\sqrt{d^3H^3T}\cdot\sqrt{\log(1/\zeta)}, \tag{F.10}$$

where $C_0 > 0$ is an absolute constant. In addition, following from Lemma F.2, it holds with probability at least $1-2\zeta$ that

$$\sum_{k=1}^K\sum_{h=1}^H \Big(\mathbb{E}_{\pi^*}\big[\iota_h^k(s_h,a_h)\,\big|\,s_1 = s_1^k\big] - \iota_h^k(s_h^k,a_h^k)\Big) \le 2\sum_{k=1}^K\sum_{h=1}^H \Gamma_h^k(s_h^k,a_h^k). \tag{F.11}$$

Recall that for all $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$, we define

$$\Gamma_h^k(s_h, a_h) = \beta \cdot \Big(\log\det\big(\Lambda_h^k + \psi_h(s_h, a_h)\psi_h(s_h, a_h)^\top\big) - \log\det(\Lambda_h^k)\Big)^{1/2}. \tag{F.12}$$

Hence, by the Cauchy-Schwartz inequality, we obtain that

$$\begin{aligned}
\sum_{k=1}^K \sum_{h=1}^H \Gamma_h^k(s_h^k, a_h^k) &= \beta \sum_{k=1}^K \sum_{h=1}^H \Big(\log\det\big(\Lambda_h^k + \psi_h(s_h^k, a_h^k)\psi_h(s_h^k, a_h^k)^\top\big) - \log\det(\Lambda_h^k)\Big)^{1/2} \\
&\le \beta \sum_{h=1}^H \Bigg(K \sum_{k=1}^K \big(\log\det(\Lambda_h^{k+1}) - \log\det(\Lambda_h^k)\big)\Bigg)^{1/2} \\
&= \beta\sqrt{K} \sum_{h=1}^H \big(\log\det(\Lambda_h^{K+1}) - \log\det(\Lambda_h^1)\big)^{1/2}.
\end{aligned} \tag{F.13}$$

In what follows, we define

$$\Delta_H = \frac{1}{\sqrt{dH^2}} \sum_{h=1}^H \big(\log\det(\Lambda_h^{K+1}) - \log\det(\Lambda_h^1)\big)^{1/2}. \tag{F.14}$$

Thus, by plugging (F.14) and $\beta = CdH \cdot \sqrt{\log(d(T+nH)/\zeta)}$ into (F.13), it holds with probability at least $1 - 2\zeta$ that,

$$\sum_{k=1}^K \sum_{h=1}^H \Gamma_h^k(s_h^k, a_h^k) \le C \cdot \Delta_H \cdot \sqrt{d^3H^3T} \cdot \sqrt{\log\big(d(T+nH)/\zeta\big)}, \tag{F.15}$$

where recall that we define $T = HK$. By further plugging (F.15) into (F.11), it holds with probability at least $1 - 2\zeta$ that,

$$\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H \Big(\mathbb{E}_{\pi^*}\big[\iota_h^k(s_h, a_h) \,\big|\, s_1 = s_1^k\big] - \iota_h^k(s_h^k, a_h^k)\Big) \\
&\qquad \le 2C \cdot \Delta_H \cdot \sqrt{d^3H^3T} \cdot \sqrt{\log\big(d(T+nH)/\zeta\big)}.
\end{aligned} \tag{F.16}$$

Finally, combining Lemma F.3, (F.9), (F.10), and (F.16), it holds with probability at least $1 - 5\zeta/2$ that

$$\mathrm{Regret}(T) \le C' \cdot \Delta_H \cdot \sqrt{d^3H^3T} \cdot \sqrt{\log\big(d(T+nH)/\zeta\big)},$$

where $C' > 0$ is an absolute constant and

$$\Delta_H = \frac{1}{\sqrt{dH^2}} \sum_{h=1}^H \big(\log\det(\Lambda_h^{K+1}) - \log\det(\Lambda_h^1)\big)^{1/2}.$$

Thus, we complete the proof of Theorem 3.5. □

F.4 Proof of Theorem A.5

Proof. In the sequel, we define the following operators,

$$(\mathbb{J}_h f)(s) = \langle f(s, \cdot), \pi_h^*(\cdot \,|\, s)\rangle_{\mathcal{A}}, \qquad (\mathbb{J}_{k,h} f)(s) = \langle f(s, \cdot), \pi_h^k(\cdot \,|\, s)\rangle_{\mathcal{A}}. \tag{F.17}$$

Meanwhile, recall that we define the following transition operators,

$$\mathbb{P}_{h+1/2} V(s_h, m_h) = \mathbb{E}\Big[V(s_{h+1}) \,\Big|\, s_{h+1} \sim \mathbb{P}\big(\cdot \,\big|\, s_h, \mathrm{do}(m_h)\big)\Big], \quad \forall V : \mathcal{S} \mapsto \mathbb{R},\ (s_h, m_h) \in \mathcal{S} \times \mathcal{M}.$$

$$\mathbb{P}_h V'(s_h, a_h) = \mathbb{E}\big[V'(s_h, m_h) \,\big|\, m_h \sim \check{\mathcal{P}}_h(\cdot \,|\, s, a)\big], \quad \forall V' : \mathcal{S} \times \mathcal{M} \mapsto \mathbb{R},\ (s_h, a_h) \in \mathcal{S} \times \mathcal{A}.$$

We further define for all $(k, h) \in [K] \times [H]$ the following transition operator,

$$\widetilde{\mathbb{P}}_{h+1/2} V(s_h, a_h, m_h) = \mathbb{E}\big[V(s_{h+1}) \,\big|\, s_{h+1} \sim \mathbb{P}(\cdot \,|\, s_h, a_h, m_h)\big], \quad \forall V : \mathcal{S} \mapsto \mathbb{R},\ (s_h, a_h, m_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{M}.$$

We define the following model prediction errors,

$$\iota_h^k(s_h, a_h) = -Q_h^k(s_h, a_h) + r_h(s_h, a_h) + (\mathbb{P}_h V_{h+1/2}^k)(s_h, a_h), \quad \forall (s_h, a_h) \in \mathcal{S} \times \mathcal{A},$$
$$\iota_{h+1/2}^k(s_h, m_h) = -V_{h+1/2}^k(s_h, m_h) + (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h, m_h), \quad \forall (s_h, m_h) \in \mathcal{S} \times \mathcal{M}. \tag{F.18}$$

In parallel to Definition F.1, we define the following filtrations that correspond to Algorithm 2.

Definition F.4 (Filtration)**.** For $(k, h) \in [K] \times [H]$, we define $\mathcal{F}'_{k,h,1}$ the σ-algebra generated by the following set,

$$B'_{k,h,1} = \big\{(s_h^i, a_h^i, m_h^i, r_h^i)\big\}_{(i,h)\in[n]\times[H]} \cup \big\{(s_j^\tau, a_j^\tau, m_j^\tau, r_j^\tau)\big\}_{(\tau,j)\in[k-1]\times[H]} \cup \big\{(s_j^k, a_j^k, m_j^k, r_j^k)\big\}_{j\in[h-1]} \cup \big\{(s_h^k, a_h^k)\big\}. \tag{F.19}$$

Similarly, we define $\mathcal{F}'_{k,h,2}$ the σ-algebra generated by the following set,

$$B'_{k,h,2} = B'_{k,h,1} \cup \{m_h^k\} \cup \{r_h^k\}, \tag{F.20}$$

and we define $\mathcal{F}'_{k,h,3}$ the σ-algebra generated by the following set,

$$B'_{k,h,3} = B'_{k,h,2} \cup \{s_{h+1}^k\}, \tag{F.21}$$

Moreover, we define $\mathcal{F}'_{0,h,3}$ the σ-algebra generated by the set $\{(s_h^i, a_h^i, m_h^i, r_h^i)\}_{(i,h)\in[n]\times[H]}$ for all $h \in [H]$. We define the timestep index as follows,

$$t'(k, h, m) = 3H \cdot k + 3(h-1) + m. \tag{F.22}$$

It then holds for $t'(k, h, m) \le t'(k', h', m')$ that $\mathcal{F}'_{k,h,m} \subseteq \mathcal{F}'_{k',h',m'}$. Hence, the set of σ-algebra $\{\mathcal{F}'_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[3]}$ is a filtration with the timestep index $t'(\cdot,\cdot,\cdot)$ defined in (F.22).

The following lemma characterizes the model prediction errors defined in (F.18).

Lemma F.5. Let $\beta = CdH\sqrt{\log(d(T+nH)/\zeta)}$ and $\zeta \in (0, 1]$. Under Assumption A.3, it holds with probability at least $1 - 4\zeta$ that

$$-2\Gamma_{h+1/2}^k(s_h, m_h) \le \iota_{h+1/2}^k(s_h, m_h) \le 0, \ \forall (k, h) \in [K] \times [H], \ (s_h, m_h) \in \mathcal{S} \times \mathcal{M}, \tag{F.23}$$
$$-2\Gamma_h^k(s_h, a_h) \le \iota_h^k(s_h, a_h) \le 0, \quad \forall (k, h) \in [K] \times [H], \ (s_h, a_h) \in \mathcal{S} \times \mathcal{A}. \tag{F.24}$$

Proof. See §G.2 for a detailed proof. □

Our goal is to upper bound the regret, which takes the following form,

$$\begin{aligned}\text{Regret}(T) &= \sum_{k=1}^K V_1^{\pi^*}(s_1^k) - V_1^{\pi^k}(s_1^k) \\ &= \underbrace{\sum_{k=1}^K \big(V_1^{\pi^*}(s_1^k) - V_1^k(x_1^k)\big)}_{\text{(i)}} + \underbrace{\sum_{k=1}^K \big(V_1^k(s_1^k) - V_1^{\pi^k}(x_1^k)\big)}_{\text{(ii)}},\end{aligned} \tag{F.25}$$

where $\{V_h^k\}_{(k,h)\in[K]\times[H]}$ is the output of Algorithm 2. In what follows, we calculate terms (i) and (ii) on the right-hand side of (F.25) separately.

Term (i). We now calculate term (i) on the right-hand side of (F.25). By (F.17), for all $h \in [H]$, it holds that

$$V_h^{\pi^*} - V_h^k = \mathbb{J}_h Q_h^{\pi^*} + \mathbb{J}_{k,h} Q_h^k = \mathbb{J}_h (Q_h^{\pi^*} - Q_h^k) + (\mathbb{J}_h - \mathbb{J}_{k,h}) Q_h^k. \tag{F.26}$$

We first calculate the term $Q_h^{\pi^*} - Q_h^k$ on the right-hand side of (F.26). Recall that we define

$$\iota_h^k = -Q_h^k + r_h + \mathbb{P}_h V_{h+1/2}^k, \qquad \iota_{h+1/2}^k = -V_{h+1/2}^k + \mathbb{P}_{h+1/2} V_{h+1}^k.$$

Meanwhile, following from the Bellman equation in (A.2), we obtain that

$$Q_h^{\pi^*} = r_h + \mathbb{P}_h V_{h+1/2}^{\pi^*}, \qquad V_{h+1/2}^{\pi^*} = \mathbb{P}_{h+1/2} V_{h+1}^{\pi^*}.$$

Thus, it holds that

$$Q_h^{\pi^*} - Q_h^k = \iota_h^k + \mathbb{P}_h(V_{h+1/2}^{\pi^*} - V_{h+1/2}^k) = \iota_h^k + \mathbb{P}_h \iota_{h+1/2}^k + \mathbb{P}_h \mathbb{P}_{h+1/2}(V_{h+1}^{\pi^*} - V_{h+1}^k). \quad \text{(F.27)}$$

Recall that we set $V_{H+1}^{\pi^*} = V_{H+1}^k = 0$. Hence, upon recursion, we obtain from (F.26) and (F.27) that

$$\begin{aligned} V_1^{\pi^*} - V_1^k &= \Bigg(\prod_{h=1}^{H} \mathbb{J}_h \mathbb{P}_h \mathbb{P}_{h+1/2}\Bigg)(V_{H+1}^{\pi^*} - V_{H+1}^k) + \sum_{h=1}^{H}\Bigg(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i \mathbb{P}_{i+1/2}\Bigg)\mathbb{J}_h \iota_h^k \\ &\quad + \sum_{h=1}^{H}\Bigg(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i \mathbb{P}_{i+1/2}\Bigg)\mathbb{J}_h \mathbb{P}_h \iota_{h+1/2}^k + \sum_{h=1}^{H}\Bigg(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i \mathbb{P}_{i+1/2}\Bigg)(\mathbb{J}_h - \mathbb{J}_{k,h}) Q_h^k \\ &= \sum_{h=1}^{H}\Bigg(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i \mathbb{P}_{i+1/2}\Bigg)(\mathbb{J}_h \iota_h^k + \mathbb{J}_h \mathbb{P}_h \iota_{h+1/2}^k) + \sum_{h=1}^{H}\Bigg(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i \mathbb{P}_{i+1/2}\Bigg)(\mathbb{J}_h - \mathbb{J}_{k,h}) Q_h^k. \end{aligned} \quad \text{(F.28)}$$

By the definition of $\mathbb{J}_h$ and $\mathbb{J}_{k,h}$ in (F.17), we further obtain from (F.28) that

$$\begin{aligned} \sum_{k=1}^{K}\big(V_1^{\pi^*}(s_1^k) - V_1^k(s_1^k)\big) &= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\big[\iota_h^k(s_h, a_h) + \iota_{h+1/2}^k(s_h, m_h) \,\big|\, s_1 = s_1^k\big] \\ &\quad + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\big[\langle Q_h^k(s_h, \cdot), \pi_h^*(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle \,\big|\, s_1 = s_1^k\big], \end{aligned} \quad \text{(F.29)}$$

which completes the calculation of term (i) on the right-hand side of (F.25).

Term (ii). We now calculate term (ii) on the right-hand side of (F.25). By (F.17), for all $h \in [H]$, we have

$$V_h^k(s_h^k) - V_h^{\pi^k}(s_h^k) = \big(\mathbb{J}_{k,h}(Q_h^k - Q_h^{\pi^k})\big)(s_h^k). \quad \text{(F.30)}$$

Meanwhile, by (F.18) it holds that

$$\begin{aligned} \iota_h^k(s_h^k, a_h^k) &= r_h(s_h^k, a_h^k) + (\mathbb{P}_h V_{h+1/2}^k)(s_h^k, a_h^k) - Q_h^k(s_h^k, a_h^k) \\ &= r_h(s_h^k, a_h^k) - Q_h^{\pi^k}(s_h^k, a_h^k) + \mathbb{P}_h V_{h+1/2}^k(s_h^k, a_h^k) + (Q_h^{\pi^k} - Q_h^k)(s_h^k, a_h^k)(s_h^k, a_h^k) \\ &= \big(\mathbb{P}_h(V_{h+1/2}^k - V_{h+1/2}^{\pi^k})\big)(s_h^k, a_h^k) - (Q_h^k - Q_h^{\pi^k})(s_h^k, a_h^k), \end{aligned} \quad \text{(F.31)}$$

where the second equality follows from the Bellman equation $Q_h^{\pi^k}(s_h, a_h) = r_h(s_h, a_h) + (\mathbb{P}_h V_{h+1/2}^{\pi^k})(s_h, a_h)$. Similarly, we have

$$\iota_{h+1/2}^k(s_h^k, m_h^k) = \big(\mathbb{P}_{h+1/2}(V_{h+1}^k - V_{h+1}^{\pi^k})\big)(s_h^k, m_h^k) - (V_{h+1/2}^k - V_{h+1/2}^{\pi^k})(s_h^k, m_h^k). \quad \text{(F.32)}$$

Thus, by combining (F.30), (F.31), and (F.32), we have

$$\begin{aligned} &(V_h^k - V_h^{\pi^k})(s_h^k) + \iota_h^k(s_h^k, a_h^k) + \iota_{h+1/2}^k(s_h^k, m_h^k) \\ &\quad = (V_{h+1}^k - V_{h+1}^{\pi^k})(s_{h+1}^k) + \underbrace{\big(\mathbb{J}_{k,h}(Q_h^k - Q_h^{\pi^k})\big)(s_h^k) - (Q_h^k - Q_h^{\pi^k})(s_h^k, a_h^k)}_{D_{k,h,1}} \\ &\qquad + \underbrace{\big(\mathbb{P}_h(V_{h+1/2}^k - V_{h+1/2}^{\pi^k})\big)(s_h^k, a_h^k) - (V_{h+1/2}^k - V_{h+1/2}^{\pi^k})(s_h^k, m_h^k)}_{D_{k,h,2}} \\ &\qquad + \underbrace{\big(\mathbb{P}_{h+1/2}(V_{h+1}^k - V_{h+1}^{\pi^k})\big)(s_h^k, m_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k})(s_{h+1}^k)}_{D_{k,h,3}}. \end{aligned} \quad \text{(F.33)}$$

Meanwhile, note that $V_{H+1}^{\pi^k} = V_{H+1}^k = 0$. Hence, by recursively applying (F.33), we obtain that

$$\begin{aligned}&(V_1^k - V_1^{\pi^k})(s_1^k)\\ &\quad= \sum_{h=1}^H (D_{k,h,1} + D_{k,h,2} + D_{k,h,3}) - \sum_{h=1}^H \big(\iota_h^k(s_h^k, a_h^k) + \iota_{h+1/2}^k(s_h^k, m_h^k)\big).\end{aligned} \tag{F.34}$$

By the definition of filtration in (F.4), for the terms $D_{k,h,1}$, $D_{k,h,2}$ and $D_{k,h,3}$ on the right-hand side of (F.33), it holds for all $(k,h) \in [K] \times [H]$ that

$$D_{k,h,1} \in \mathcal{F}_{k,h,1}, \quad D_{k,h,2} \in \mathcal{F}_{k,h,2}, \quad D_{k,h,3} \in \mathcal{F}_{k,h,3}.$$

Moreover, it holds that

$$\mathbb{E}[D_{k,h,1} \,|\, \mathcal{F}_{k,h-1,3}] = \mathbb{E}[D_{k,h,2} \,|\, \mathcal{F}_{k,h,1}] = \mathbb{E}[D_{k,h,3} \,|\, \mathcal{F}_{k,h,2}] = 0.$$

Hence, the terms $D_{k,h,1}$, $D_{k,h,2}$ and $D_{k,h,3}$ defines a martingale $M'_{k,h,m}$ with respect to the timestep index $t'(\cdot,\cdot,\cdot)$ as follows,

$$M'_{k,h,m} = \sum_{\substack{(\tau,i,\ell) \in [K]\times[H]\times[3] \\ t'(\tau,i,\ell) \le t'(k,h,m)}} D_{\tau,i,\ell}, \tag{F.35}$$

where $t'(\cdot,\cdot,\cdot)$ is defined in (F.22) of Definition F.4. In specific, we have

$$M'_{K,H,3} = \sum_{k=1}^K \sum_{h=1}^H (D_{k,h,1} + D_{k,h,2} + D_{k,h,3}). \tag{F.36}$$

By further taking sum of (F.34) over $k \in [K]$, we obtain from (F.36) that

$$\sum_{k=1}^K (V_1^k - V_1^{\pi^k})(s_1^k) = M'_{K,H,3} - \sum_{k=1}^K \sum_{h=1}^H \big(\iota_h^k(s_h^k, a_h^k) + \iota_{h+1/2}^k(s_h^k, m_h^k)\big), \tag{F.37}$$

which completes the calculation of term (ii) on the right-hand side of (F.25).

Finally, by plugging (F.29) and (F.37) into (F.25), we conclude that

$$\begin{aligned}\mathrm{Regret}(T) = &\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle \,\big|\, s_1 = s_1^k\big] + M'_{K,H,3}\\ &+ \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\iota_h^k(s_h,a_h) + \iota_{h+1/2}^k(s_h,m_h) \,\big|\, s_1 = s_1^k\big]\\ &- \sum_{k=1}^K \sum_{h=1}^H \big(\iota_h^k(s_h^k,a_h^k) + \iota_{h+1/2}^k(s_h^k,m_h^k)\big),\end{aligned} \tag{F.38}$$

where $M'_{K,H,3}$ is defined in (F.36).

We now upper bound the right-hand side of (F.38). The following proof is similar to that of Theorem 3.5 in §F.3. In the sequel, we define

$$Y' = \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle Q_h^k(s_h,\cdot), \pi_h^*(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle \,\big|\, s_1 = s_1^k\big],$$

$$Z' = \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\iota_h^k(s_h,a_h) + \iota_{h+1/2}^k(s_h,m_h) \,\big|\, s_1 = s_1^k\big] - \sum_{k=1}^K \sum_{h=1}^H \big(\iota_h^k(s_h^k,a_h^k) + \iota_{h+1/2}^k(s_h^k,m_h^k)\big).$$

It then follows from (F.38) that

$$\mathrm{Regret}(T) = Y' + M'_{K,H,3} + Z'. \tag{F.39}$$

Recall that we set π_h^k to be the greedy policy with respect to the action-value function Q_h^k. Thus, it holds that

$$Y' = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\big[\langle Q_h^k(s_h,\cdot),\pi_h^*(\cdot\,|\,s_h)-\pi_h^k(\cdot\,|\,s_h)\rangle\,\big|\,s_1=s_1^k\big]\le 0. \tag{F.40}$$

Meanwhile, following from the truncation of Q_h^k in Algorithm 2 and the assumption that $r_h\in[0,1]$, for terms $D_{k,h,i}$ defined in (F.33), we have

$$|D_{k,h,i}|\le 2H,\quad \forall(k,h,i)\in[K]\times[H]\times[3].$$

Hence, by the Azumas-Hoeffding lemma, it holds with probability at least $1-\zeta$ that

$$M'_{K,H,3}\le C_1\cdot\sqrt{d^3H^3T}\cdot\sqrt{\log(dT/\zeta)}, \tag{F.41}$$

where $M'_{K,H,3}$ is the martingale defined in (F.35), $C_1>0$ is an absolute constant, and $T=HK$. Following from Lemma F.5, it holds with probability at least $1-4\zeta$ that

$$Z'\le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_{h+1/2}^k(s_h^k,m_h^k)+2\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(s_h^k,a_h^k). \tag{F.42}$$

Following from the definition of $\Gamma_{h+1/2}^k$ in (A.12), we obtain that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_{h+1/2}^k(s_h^k,m_h^k)&=2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\log\det\big(\Lambda_{1,h}^k+\psi_h(s_h^k,m_h^k)\psi_h(s_h,m_h)^\top\big)-\log\det(\Lambda_{1,h}^k)\Big)^{1/2}\\&=2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\log\det(\Lambda_{1,h}^{k+1})-\log\det(\Lambda_{1,h}^k)\big)^{1/2}.\end{aligned} \tag{F.43}$$

Thus, by the Cauchy-Schwartz inequality, we obtain from (F.43) that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_{h+1/2}^k(s_h^k,m_h^k)&\le\beta\sum_{h=1}^{H}\Bigg(K\cdot\sum_{k=1}^{K}\big(\log\det(\Lambda_{1,h}^{k+1})-\log\det(\Lambda_{1,h}^1)\big)\Bigg)^{1/2}\\&\le\beta\cdot\sqrt{K}\sum_{h=1}^{H}\big(\log\det(\Lambda_{1,h}^{K+1})-\log\det(\Lambda_{1,h}^1)\big)^{1/2}.\end{aligned} \tag{F.44}$$

Similarly, we obtain that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(s_h^k,a_h^k)\le\beta\cdot\sqrt{K}\sum_{h=1}^{H}\big(\log\det(\Lambda_{2,h}^{k+1})-\log\det(\Lambda_{2,h}^1)\big)^{1/2}. \tag{F.45}$$

In what follows, we define

$$\begin{aligned}\Delta_{1,H}&=\frac{1}{\sqrt{dH^2}}\sum_{h=1}^{H}\big(\log\det(\Lambda_{1,h}^{K+1})-\log\det(\Lambda_{1,h}^1)\big)^{1/2},\\\Delta_{2,H}&=\frac{1}{\sqrt{dH^2}}\sum_{h=1}^{H}\big(\log\det(\Lambda_{2,h}^{k+1})-\log\det(\Lambda_{2,h}^1)\big)^{1/2}.\end{aligned}$$

By plugging (F.44), (F.45), and $\beta=CdH\cdot\sqrt{\log(d(T+nH)/\zeta)}$ into (F.42), we obtain that

$$Z'\le 2C\cdot(\Delta_{1,H}+\Delta_{2,H})\cdot\sqrt{d^3H^3T}\cdot\sqrt{\log\big(d(T+nH)/\zeta\big)}, \tag{F.46}$$

which holds with probability at least $1-4\zeta$. Here recall that we define $T=HK$. Finally, by plugging (F.40), (F.41), and (F.46) into (F.39), it holds with probability at least $1-5\zeta$ that

$$\mathrm{Regret}(T)\le C'\cdot(\Delta_{1,H}+\Delta_{2,H})\cdot\sqrt{d^3H^3T}\cdot\sqrt{\log\big(d(T+nH)/\zeta\big)},$$

where $C'>0$ is an absolute constant. Thus, we complete the proof of Theorem A.5. □

G Proof of Auxiliary Result

G.1 Proof of Lemma F.2

Proof. Recall that we define

$$\begin{aligned}(\mathbb{P}_h V)(s_h, a_h) &= \mathbb{E}\Big[V(s_{h+1}) \,\Big|\, s_{h+1} \sim \mathbb{P}\big(\cdot \,|\, s_h, \mathrm{do}(a_h)\big)\Big] \\ &= \mathbb{E}\big[V(s_{h+1}) \,\big|\, s_{h+1} \sim \mathcal{P}_h(\cdot \,|\, s_h, a_h, u_h), u_h \sim \widetilde{\mathcal{P}}_h(\cdot \,|\, s_h)\big],\end{aligned}$$

where the second equality follows from Proposition 3.2. In the sequel, we define

$$(\widetilde{\mathbb{P}}_h V)(s_h, a_h, u_h) = \mathbb{E}\Big[V(s_{h+1}) \,\Big|\, s_{h+1} \sim \mathcal{P}_h\big(\cdot \,|\, s_h, a_h, u_h\big)\Big].$$

By Assumption 3.3, we obtain that

$$\mathbb{P}_h V_{h+1}^k = \psi_h^\top \langle \mu_h, V_{h+1}^k \rangle = \psi_h^\top (\Lambda_h^k)^{-1} \Lambda_h^k \langle \mu_h, V_{h+1}^k \rangle, \quad \widetilde{\mathbb{P}}_h V_{h+1}^k = \phi_h^\top \langle \mu_h, V_{h+1}^k \rangle. \tag{G.1}$$

Recall that

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \psi_h(s_h^\tau, a_h^\tau)^\top + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \phi_h(s_h^i, a_h^i, u_h^i)^\top + \lambda I.$$

Therefore, by (G.1), we obtain that

$$\begin{aligned}(\mathbb{P}_h V_{h+1}^k)(\cdot,\cdot) &= \psi_h(\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \psi_h(s_h^\tau, a_h^\tau)^\top \langle \mu_h, V_{h+1}^k \rangle + \lambda \cdot \langle \mu_h, V_{h+1}^k \rangle \\ &\qquad + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \phi_h(s_h^i, a_h^i, u_h^i)^\top \langle \mu_h, V_{h+1}^k \rangle \bigg) \\ &= \psi_h(\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \cdot (\mathbb{P}_h V_{h+1}^k)(s_h^\tau, a_h^\tau) + \lambda \cdot \langle \mu_h, V_{h+1}^k \rangle \\ &\qquad + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \cdot (\widetilde{\mathbb{P}}_h V_{h+1}^k)(s_h^i, a_h^i, u_h^i) \bigg).\end{aligned} \tag{G.2}$$

Recall that we define the counterfactual reward as follows,

$$R_h(s_h, a_h) = \mathbb{E}_{u_h}\big[r(s_h, a_h, u_h) \,\big|\, S_h = s_h\big], \quad \forall (s_h, a_h) \in \mathcal{S} \times \mathcal{A}. \tag{G.3}$$

It then follows from Assumption 3.3 and Proposition 3.4 that $R_h(\cdot,\cdot) = \psi_h(\cdot,\cdot)^\top \theta_h$. Hence, it holds for all $h \in [H]$ that

$$\begin{aligned}r_h(\cdot,\cdot,\cdot) &= \phi_h(\cdot,\cdot,\cdot)^\top \theta_h = \phi_h(\cdot,\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \Lambda_h^k \theta_h \\ &= \phi_h(\cdot,\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \psi_h(s_h^\tau, a_h^\tau)^\top \theta_h + \lambda \cdot \langle \mu_h, V_{h+1}^k \rangle \\ &\qquad + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \phi_h(s_h^i, a_h^i, u_h^i)^\top \theta_h \bigg) \\ &= \phi_h(\cdot,\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \cdot R_h(s_h^\tau, a_h^\tau) + \lambda \cdot \theta_h \\ &\qquad + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \cdot \mathbb{E}[r_h \,|\, s_h^i, a_h^i, u_h^i] \bigg).\end{aligned} \tag{G.4}$$

Meanwhile, following from the explicit update of ω_h^k in (3.9), we obtain that

$$\begin{aligned}\psi_h(\cdot,\cdot)^\top \omega_h^k = \psi_h(\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \bigg(&\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^k(s_{h+1}^\tau) + r_h^\tau\big) \\ &+ \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \cdot \big(V_{h+1}^k(s_{h+1}^i) + r_h^i\big) \bigg).\end{aligned} \tag{G.5}$$

Hence, combining (G.2), (G.4), and (G.5), we obtain that

$$\begin{aligned}&\psi_h(\cdot,\cdot)^\top\omega_h^k - R_h(\cdot,\cdot) - (\mathbb{P}_hV_{h+1}^k)(\cdot,\cdot)\\&\quad= \psi_h(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}(S_{1,h}+S_{2,h}+S_{3,h}+S_{4,h}) - \psi_h(\cdot,\cdot)^\top\lambda\cdot\big(\langle\mu_h,V_{h+1}^k\rangle+\theta_h\big),\end{aligned}\tag{G.6}$$

where we define

$$S_{1,h} = \sum_{\tau=1}^{k-1}\psi_h(s_h^\tau,a_h^\tau)\cdot\big(V_{h+1}^k(s_{h+1}^\tau) - (\mathbb{P}_hV_{h+1}^k)(s_h^\tau,a_h^\tau)\big),\tag{G.7}$$

$$S_{2,h} = \sum_{i=1}^{n}\phi_h(s_h^i,a_h^i,u_h^i)\cdot\big(V_{h+1}^k(s_{h+1}^i) - (\widetilde{\mathbb{P}}_hV_{h+1}^k)(s_h^i,a_h^i,u_h^i)\big),$$

$$S_{3,h} = \sum_{\tau=1}^{k-1}\psi_h(s_h^\tau,a_h^\tau)\cdot\big(r_h^\tau - R(s_h^\tau,a_h^\tau)\big),\quad\text{and}\quad S_{4,h} = \sum_{i=1}^{n}\phi_h(s_h^i,a_h^i,u_h^i)\cdot\big(r_h^i - \mathbb{E}[r_h\,|\,s_h^i,a_h^i,u_h^i]\big).$$

In what follows, we upper bound the right-hand side of (G.6). By the Cauchy-Schwartz inequality, we obtain that

$$\begin{aligned}&|\psi_h(\cdot,\cdot)^\top\omega_h^k - R_h(\cdot,\cdot) - (\mathbb{P}_hV_{h+1}^k)(\cdot,\cdot)|\\&\le\big(\psi_h(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}\psi_h(\cdot,\cdot)\big)^{1/2}\cdot\Big(\Big\|\sum_{\ell=1}^{4}S_{\ell,h}\Big\|_{(\Lambda_h^k)^{-1}} + \lambda\cdot\big(\|\langle\mu_h,V_{h+1}^k\rangle\|_{(\Lambda_h^k)^{-1}} + \|\theta_h\|_{(\Lambda_h^k)^{-1}}\big)\Big),\end{aligned}\tag{G.8}$$

where $S_{1,h}$, $S_{2,h}$, $S_{3,h}$, and $S_{4,h}$ are defined in (G.7). By Lemma H.6, for $\lambda = 1$, it holds with probability at least $1-2\zeta$ that

$$\Big\|\sum_{\ell=1}^{4}S_{\ell,h}\Big\|_{(\Lambda_h^k)^{-1}} \le C'dH\sqrt{\log\big(2(C+1)d(T+nH)/\zeta\big)},\tag{G.9}$$

where $C>0$ and $C'>0$ are absolute constants. Meanwhile, by Assumption 3.3, it holds that

$$\begin{aligned}\|\langle\mu_h,V_{h+1}^k\rangle\|_{(\Lambda_h^k)^{-1}} &\le \|\langle\mu_h,V_{h+1}^k\rangle\|_2/\sqrt{\lambda}\\&\le\Big(\sum_{\ell=1}^{d}\|\mu_{\ell,h}\|_1^2\Big)^{1/2}\cdot\|V_{h+1}^h\|_\infty/\sqrt{\lambda}\le H\sqrt{d/\lambda},\end{aligned}\tag{G.10}$$

where the first inequality follows from the fact that $\Lambda_h^k\succeq\lambda I$, the second inequality follows from the Hölder's inequality, and the third inequality follows from Assumption 3.3 and the fact that $V_{k+1}^h\le H$. Similarly, it holds from Assumption 3.3 that

$$\|\theta_h\|_{(\Lambda_h^k)^{-1}}\le\|\theta_h\|_2/\sqrt{\lambda}\le\sqrt{d/\lambda}.\tag{G.11}$$

Finally, by plugging (G.9), (G.10), and (G.11) into (G.8) with $\lambda=1$, it holds with probability at least $1-2\zeta$ that

$$|\psi_h(\cdot,\cdot)^\top\omega_h^k - R_h(\cdot,\cdot) - (\mathbb{P}_hV_{h+1}^k)(\cdot,\cdot)|\le\beta/\sqrt{2}\cdot\big(\psi_h(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}\psi_h(\cdot,\cdot)\big)^{1/2},\tag{G.12}$$

where we set $\beta = C''dH\sqrt{\log(d(T+nH)/\zeta)}$ for a sufficiently large absolute constant $C''>0$. By further applying Lemma H.7 to (G.12), for $\lambda=1$, it holds with probability at least $1-2\zeta$ that

$$\begin{aligned}&|\psi_h(\cdot,\cdot)^\top\omega_h^k - R_h(\cdot,\cdot) - (\mathbb{P}_hV_{h+1}^k)(\cdot,\cdot)|\\&\le\beta\cdot\Big(\log\det\big(\Lambda_h^k+\psi_h(\cdot,\cdot)\psi_h(\cdot,\cdot)^\top\big) - \log\det(\Lambda_h^k)\Big)^{1/2} = \Gamma_h^k(\cdot,\cdot).\end{aligned}\tag{G.13}$$

Recall that we set

$$Q_h^k(\cdot,\cdot) = \min\big\{\psi_h(\cdot,\cdot)^\top\omega_h^k + \Gamma_h^k(\cdot,\cdot), H-h\big\}.$$

Hence, by (G.13), it holds with probability at least $1-2\zeta$ that

$$\begin{aligned}-\iota_h^k(\cdot,\cdot) &= Q_h^k(\cdot,\cdot) - R_h(\cdot,\cdot) - (\mathbb{P}_hV_{h+1}^k)(\cdot,\cdot)\\&\le\psi_h(\cdot,\cdot)^\top\omega_h^k + \Gamma_h^k(\cdot,\cdot) - R_h(\cdot,\cdot) - (\mathbb{P}_hV_{h+1}^k)(\cdot,\cdot)\le 2\Gamma_h^k(\cdot,\cdot),\end{aligned}$$

and

$$\iota_h^k(\cdot,\cdot) = -Q_h^k(\cdot,\cdot) + R_h(\cdot,\cdot) + (\mathbb{P}_h V_{h+1}^k)(\cdot,\cdot)$$
$$\leq \max\big\{(\mathbb{P}_h V_{h+1}^k)(\cdot,\cdot) + R_h(\cdot,\cdot) - \psi_h(\cdot,\cdot)^\top \omega_h^k - \Gamma_h^k, R_h(\cdot,\cdot) + (\mathbb{P}_h V_{h+1}^k)(\cdot,\cdot) - H + h\big\} \leq 0,$$

where the second inequality follows from (G.13) the facts that $V_{h+1}^k \leq H - h - 1$ and $R_h \leq 1$. In conclusion, it holds with probability at least $1 - 2\zeta$ that

$$-2\Gamma_h^k(\cdot,\cdot) \leq \iota_h^k(\cdot,\cdot) \leq 0,$$

which concludes the proof of Lemma F.2. □

G.2 Proof of Lemma F.5

Proof. Recall that we define the following transition operators,

$$\mathbb{P}_{h+1/2} V(s_h, m_h) = \mathbb{E}\Big[V(s_{h+1}) \,\Big|\, s_{h+1} \sim \mathbb{P}\big(\cdot \,\big|\, s_h, \mathrm{do}(m_h)\big)\Big]$$
$$\widetilde{\mathbb{P}}_{h+1/2} V(s_h, a_h, m_h) = \mathbb{E}\big[V(s_{h+1}) \,\big|\, s_{h+1} \sim \mathbb{P}(\cdot \,|\, s_h, a_h, m_h)\big]. \tag{G.14}$$

Following from Assumption A.3 and (A.7), we have

$$\mathbb{P}_{h+1/2} V_{h+1}^k = \psi_h^\top \langle \mu_h, V_{h+1}^k \rangle = \psi_h^\top (\Lambda_{1,h}^k)^{-1} \Lambda_{1,h}^k \langle \mu_h, V_{h+1}^k \rangle, \tag{G.15}$$
$$\widetilde{\mathbb{P}}_{h+1/2} V_{h+1}^k = \phi_h^\top \langle \mu_h, V_{h+1}^k \rangle, \tag{G.16}$$

where we define

$$\Lambda_{1,h}^k = \sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, m_h^\tau) \psi(s_h^\tau, m_h^\tau)^\top + \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, m_h^i) \phi_h(s_h^i, a_h^i, m_h^i)^\top + \lambda I. \tag{G.17}$$

Hence, following from (G.15), it holds for all $(s_h, m_h) \in \mathcal{S} \times \mathcal{M}$ that

$$\mathbb{P}_{h+1/2} V_{h+1}^k(s_h, m_h)$$
$$= \psi_h(s_h, m_h)^\top (\Lambda_{1,h}^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, m_h^\tau) \psi(s_h^\tau, m_h^\tau)^\top \langle \mu_h, V_{h+1}^k \rangle + \lambda \cdot \langle \mu_h, V_{h+1}^k \rangle \tag{G.18}$$
$$+ \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, m_h^i) \phi_h(s_h^i, a_h^i, m_h^i)^\top \langle \mu_h, V_{h+1}^k \rangle \bigg).$$

By plugging (G.15) and (G.16) into (G.18), we further obtain that

$$\mathbb{P}_{h+1/2} V_{h+1}^k(s_h, m_h)$$
$$= \psi_h(s_h, m_h)^\top (\Lambda_{1,h}^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, m_h^\tau) \cdot (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h^\tau, m_h^\tau) + \lambda \cdot \langle \mu_h, V_{h+1}^k \rangle \tag{G.19}$$
$$+ \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, m_h^i) \cdot (\widetilde{\mathbb{P}}_{h+1/2} V_{h+1}^k)(s_h^i, a_h^i, m_h^i) \bigg).$$

Following from the update of $\omega_{1,h}^k$ in (A.10), it holds for all $h \in [H]$ and $(s_h, m_h) \in \mathcal{S} \times \mathcal{M}$ that

$$\psi_h(s_h, m_h)^\top \omega_{1,h}^k = \psi_h(s_h, m_h)^\top (\Lambda_{1,h}^k)^{-1} \bigg(\sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, m_h^\tau) \cdot V_{h+1}^k(s_{h+1}^\tau) \tag{G.20}$$
$$+ \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, m_h^i) \cdot V_{h+1}^k(s_{h+1}^i) \bigg).$$

Hence, combining (G.19) and (G.20), we obtain for all $h \in [H]$ and $(s_h, m_h) \in \mathcal{S} \times \mathcal{M}$ that

$$\psi_h(s_h, m_h)^\top \omega_{1,h}^k - \mathbb{P}_{h+1/2} V_{h+1}^k(s_h, m_h)$$
$$= \psi_h(s_h, m_h)^\top (\Lambda_{1,h}^k)^{-1} (S'_{1,h} + S'_{2,h}) + \lambda \cdot \psi_h(s, m)^\top \langle \mu_h, V_{h+1}^k \rangle, \tag{G.21}$$

where we define

$$S'_{1,h} = \sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, m_h^\tau) \cdot \big(V_{h+1}^k(s_{h+1}^\tau) - (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h^\tau, m_h^\tau)\big),$$
$$S'_{2,h} = \phi_h(s_h^i, a_h^i, m_h^i) \cdot \big(V_{h+1}^k(s_{h+1}^i) - (\widetilde{\mathbb{P}}_{h+1/2} V_{h+1}^k)(s_h^i, a_h^i, m_h^i)\big).$$

We now upper bound the right-hand side of (G.21). By the Cauchy-Schwartz inequality, we obtain from (G.21) that

$$\begin{aligned}
&|\psi_h^\top \omega_{1,h}^k - \mathbb{P}_{h+1/2} V_{h+1}^k| \\
&\quad \le \big(\psi_h^\top (\Lambda_{1,h}^k)^{-1} \psi_h\big)^{1/2} \cdot \big(\|S'_{1,h} + S'_{2,h}\|_{(\Lambda_h^k)^{-1}} + \lambda \cdot \|\langle \mu_h, V_{h+1}^k \rangle\|_{(\Lambda_h^k)^{-1}}\big). \qquad \text{(G.22)}
\end{aligned}$$

Following from similar analysis to the proof of Lemma H.6 in §H, for $\lambda = 1$, it holds with probability at least $1 - 2\zeta$ that

$$\|S'_{1,h} + S'_{2,h}\|_{(\Lambda_h^k)^{-1}} \le C' dH \sqrt{\log\big(2(C+1)d(T+nH)/\zeta\big)}. \qquad \text{(G.23)}$$

Meanwhile, by Assumption A.3, we have

$$\begin{aligned}
\|\langle \mu_h, V_{h+1}^k \rangle\|_{(\Lambda_h^k)^{-1}} &\le \|\langle \mu_h, V_{h+1}^k \rangle\|_2 / \sqrt{\lambda} \\
&\le \bigg(\sum_{\ell=1}^d \|\mu_{\ell,h}\|_1^2\bigg)^{1/2} \cdot \|V_{h+1}^h\|_\infty / \sqrt{\lambda} \le H\sqrt{d/\lambda}, \qquad \text{(G.24)}
\end{aligned}$$

where the first inequality follows from the fact that $\Lambda_{1,h}^k \succeq \lambda I$, the second inequality follows from the Hölder's inequality, and the third inequality follows from Assumption A.3 and the fact that $V_{h+1}^h \le H$. Finally, by plugging (G.23) and (G.24) into (G.22), we obtain for all $(s_h, m_h) \in \mathcal{S} \times \mathcal{M}$ that

$$\begin{aligned}
&|\psi_h(s_h, m_h)^\top \omega_{1,h}^k - (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h, m_h)| \\
&\quad \le \beta/\sqrt{2} \cdot \big(\psi_h(s_h, m_h)^\top (\Lambda_{1,h}^k)^{-1} \psi_h(s_h, m_h)\big)^{1/2} \\
&\quad \le \beta \cdot \Big(\log\det\big(\Lambda_{1,h}^k + \psi_h(s_h, m_h)\psi_h(s_h, m_h)^\top\big) - \log\det(\Lambda_{1,h}^k)\Big)^{1/2} \\
&\quad = \Gamma_{h+1/2}^k(s_h, m_h), \qquad \text{(G.25)}
\end{aligned}$$

where we set $\beta = C'' dH\sqrt{\log(d(T+nH)/\zeta)}$ for a sufficiently large absolute constant $C'' > 0$ and the last inequality follows from Lemma H.7. Here $\Gamma_{h+1/2}^k$ is the UCB defined in (A.12). Recall that for all $(s_h, m_h) \in \mathcal{S} \times \mathcal{M}$, we define

$$V_{h+1/2}^k(s_h, m_h) = \min\big\{\psi_h(s_h, m_h)^\top \omega_{1,h}^k + \Gamma_{h+1/2}^k(s_h, m_h), H - h\big\}.$$

Hence, by (G.25), for all $(s_h, m_h) \in \mathcal{S} \times \mathcal{M}$, it holds with probability at least $1 - 2\zeta$ that

$$\begin{aligned}
-\iota_{h+1/2}^k(s_h, m_h) &= V_{h+1/2}^k(s_h, m_h) - (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h, m_h) \\
&\le \psi_h(s_h, m_h)^\top \omega_h^k + \Gamma_{h+1/2}^k(s_h, m_h) - (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h, m_h) \le 2\Gamma_{h+1/2}^k(s_h, m_h),
\end{aligned}$$

and

$$\begin{aligned}
\iota_{h+1/2}^k(s_h, m_h) &= -V_{h+1/2}^k(s_h, m_h) + (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h, m_h) \\
&\le \max\big\{(\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h, m_h) - \psi_h(s_h, m_h)^\top \omega_{1,h}^k - \Gamma_{h+1/2}^k(s_h, m_h), \\
&\qquad (\mathbb{P}_{h+1/2} V_{h+1}^k)(s_h, m_h) - H + h\big\} \le 0,
\end{aligned}$$

where the second inequality follows from (G.25) and the fact that $V_{h+1}^k \le H - h - 1$. In conclusion, it holds with probability at least $1 - 2\zeta$ that

$$-2\Gamma_{h+1/2}^k(s_h, m_h) \le \iota_{h+1/2}^k(s_h, m_h) \le 0.$$

Similarly, following from the proof of Lemma F.2 with Lemma H.5 in place of Lemma H.4, the reward r_h in place of R_h, and the feature γ_h in place of both ψ_h and ϕ_h, for all $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$, it holds with probability at least $1 - 2\zeta$ that

$$-2\Gamma_h^k(s_h, a_h) \le \iota_h^k(s_h, a_h) \le 0.$$

Thus, we complete the proof of Lemma F.5. □

H Auxiliary Lemma

Lemma H.1 (Concentration of Self-Normalized Process [1, 16])**.** Let $\{\epsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $\epsilon_t \,|\, \mathcal{F}_{t-1}$ be zero-mean and σ-sub-Gaussian. Let $\{\psi_t\}_{t=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process with $\psi_t \in \mathcal{F}_{t-1}$. Let $\overline{\Lambda}_t = \overline{\Lambda}_0 + \sum_{\tau=1}^{t} \psi_\tau \psi_\tau^\top$, where $\overline{\Lambda}_0$ is a positive definite matrix. Let $\delta > 0$ be an absolute constant. It then holds with probability at least $1-\delta$ that

$$\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \epsilon_\tau\Big\|_{\overline{\Lambda}_t^{-1}}^2 \le 2\sigma^2 \cdot \log\Big(\sqrt{\det(\overline{\Lambda}_t)/\det(\overline{\Lambda}_0)} \cdot \delta^{-1}\Big), \quad \forall t \ge 0.$$

Proof. See [1] for a detailed proof. □

Lemma H.2 (Lemma D.4 of [16])**.** Let $\{s_t\}_{t=1}^{\infty}$ and $\{\psi_t\}_{t=1}^{\infty}$ with $\|\psi_t\|_2 \le 1$ be $\mathcal{S}$-valued and $\mathbb{R}^d$-valued stochastic processes adopted to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$, respectively. Let $\overline{\Lambda}_t = \overline{\Lambda}_0 + \sum_{\tau=1}^{t} \psi_\tau \psi_\tau^\top$, where $\overline{\Lambda}_0 \succeq \lambda I$ is a positive definite matrix. Let $\sup_{s\in\mathcal{S}} |V(s)| \le H$ for all $V \in \mathcal{V}$. Let $\delta > 0$ be an absolute constant. It then holds with probability at least $1-\delta$ that

$$\begin{aligned}
&\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \Big(V(s_\tau) - \mathbb{E}\big[V(s_\tau) \,\big|\, \mathcal{F}_{\tau-1}\big]\Big)\Big\|_{\overline{\Lambda}_t^{-1}}^2 \\
&\qquad \le 4H^2 \cdot \Big(d/2 \cdot \log\big(\det(\overline{\Lambda}_t)/\det(\overline{\Lambda}_0)\big) + \log(\mathcal{N}_\epsilon/\delta)\Big) + 8t^2\epsilon^2/\lambda.
\end{aligned}$$

Here $\mathcal{N}_\epsilon$ is the ϵ-covering number of $\mathcal{V}$ with respect to the metric $d(V, V') = \sup_{s\in\mathcal{S}} |V(s) - V'(s)|$ for all $V, V' \in \mathcal{V}$.

Proof. The proof technique is similar to that of Lemma D.4 by [16]. For all $V \in \mathcal{V}$, there exist an element $\widetilde{V}$ in the ϵ-covering of $\mathcal{V}$ satisfying

$$d(V, \widetilde{V}) = \sup_{s\in\mathcal{S}} |V(s) - \widetilde{V}(s)| \le \epsilon. \tag{H.1}$$

In the sequel, we define

$$\Delta_V(\cdot) = V(\cdot) - \widetilde{V}(\cdot). \tag{H.2}$$

It then holds that

$$\begin{aligned}
&\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \Big(V(s_\tau) - \mathbb{E}\big[V(s_\tau) \,\big|\, \mathcal{F}_{\tau-1}\big]\Big)\Big\|_{\overline{\Lambda}_t^{-1}}^2 \\
&\qquad \le 2\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \Big(\widetilde{V}(s_\tau) - \mathbb{E}\big[\widetilde{V}(s_\tau) \,\big|\, \mathcal{F}_{\tau-1}\big]\Big)\Big\|_{\overline{\Lambda}_t^{-1}}^2 \\
&\qquad\quad + 2\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \Big(\Delta_V(s_\tau) - \mathbb{E}\big[\Delta_V(s_\tau) \,\big|\, \mathcal{F}_{\tau-1}\big]\Big)\Big\|_{\overline{\Lambda}_t^{-1}}^2.
\end{aligned} \tag{H.3}$$

Note that $|\widetilde{V}(s)| \le H$ for all $s \in \mathcal{S}$. Hence, following from Lemma H.1 and a union bound argument, it holds with probability at least $1-\delta$ that

$$\begin{aligned}
&2\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \Big(\widetilde{V}(s_\tau) - \mathbb{E}\big[\widetilde{V}(s_\tau) \,\big|\, \mathcal{F}_{\tau-1}\big]\Big)\Big\|_{\overline{\Lambda}_t^{-1}}^2 \\
&\qquad \le 4H^2 \cdot \Big(d/2 \cdot \log\big(\det(\overline{\Lambda}_t)/\det(\overline{\Lambda}_0)\big) + \log(\mathcal{N}_\epsilon/\delta)\Big),
\end{aligned} \tag{H.4}$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering number of $\mathcal{V}$. Meanwhile, it follows from (H.1) and (H.2) that $|\Delta_V(s)| \le \epsilon$ for all $s \in \mathcal{S}$. Hence, we have

$$2\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \Big(\Delta_V(s_\tau) - \mathbb{E}\big[\Delta_V(s_\tau) \,\big|\, \mathcal{F}_{\tau-1}\big]\Big)\Big\|_{\overline{\Lambda}_t^{-1}}^2 \le 8t^2\epsilon^2/\lambda, \tag{H.5}$$

where the inequality follows from the fact that $\overline{\Lambda}_t \succeq \lambda I$. By plugging (H.4) and (H.5) into (H.3), it holds with probability at least $1-\delta$ that

$$\Big\|\sum_{\tau=1}^{t} \psi_\tau \cdot \Big(V(s_\tau) - \mathbb{E}\big[V(s_\tau) \,\big|\, \mathcal{F}_{\tau-1}\big]\Big)\Big\|^2_{\overline{\Lambda}_t^{-1}}$$
$$\leq 4H^2 \cdot \Big(d/2 \cdot \log\big(\det(\overline{\Lambda}_t)/\det(\overline{\Lambda}_0)\big) + \log(\mathcal{N}_\epsilon/\delta)\Big) + 8t^2\epsilon^2/\lambda,$$

which concludes the proof of Lemma H.2. □

Lemma H.3 (Upper Bound of Parameter [16])**.** Under Assumption 3.3, It holds that

$$\|\omega_h^k\|_2 \leq H\big(d(k+n)/\lambda\big)^{1/2}, \quad \forall (k,h) \in [K] \times [H]. \tag{H.6}$$

Proof. See [16] for a detailed proof. □

Lemma H.4 (Covering Number of $\mathcal{V}$ [16])**.** Let $\mathcal{V}$ be a class of functions V satisfying

$$V(\cdot) = \min\big\{\max_{a\in\mathcal{A}} \psi(\cdot,a)^\top \omega + \Gamma(\cdot,a), H-h\big\}, \tag{H.7}$$

where

$$\Gamma(\cdot,\cdot) = \sqrt{2}\beta \cdot \Big(\log\det\big(\Lambda + \psi(\cdot,\cdot)\psi(\cdot,\cdot)^\top\big) - \log\det(\Lambda)\Big)^{1/2}. \tag{H.8}$$

Here the function V is parameterized by (ω, Λ) and the parameter β is fixed. Let $\psi(\cdot,\cdot)$ be an $\mathbb{R}^d$-valued function and $\Lambda \in \mathbb{R}^{d\times d}$. Let $\|\psi(s,a)\|_2 \leq 1$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. For $\|\omega\|_2 \leq L$, $\Lambda \succeq \lambda I$, $\beta \in [0, B]$, and $\epsilon > 0$, there exist an ϵ-covering of $\mathcal{V}$ with respect to the metric $d(V,V') = \sup_{s\in\mathcal{S}} |V(s) - V'(s)|$, such that the covering number $\mathcal{N}_\epsilon$ is upper bounded as follows,

$$\log \mathcal{N}_\epsilon \leq d \cdot \log(1 + 4L/\epsilon) + d^2 \cdot \log\big(1 + 16B^2 d^{1/2}/(\epsilon^2\lambda)\big).$$

Proof. The proof technique is similar to that of Lemma D.6 by [16]. Let V_1 and V_2 be the functions defined in (H.7), which are parameterized by (ω_1, Λ_1) and (ω_2, Λ_2), respectively. Note that

$$\begin{aligned} d(V_1,V_2) &\leq \sup_{s\in\mathcal{S}} \big|\min\big\{\max_{a\in\mathcal{A}} \psi(s,a)^\top\omega_1 + \Gamma_1(s,a), H-h\big\} \\ &\qquad - \min\big\{\max_{a\in\mathcal{A}} \psi(s,a)^\top\omega_2 + \Gamma_2(s,a), H-h\big\}\big| \\ &\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\psi(s,a)^\top(\omega_1-\omega_2) + \Gamma_1(s,a) - \Gamma_2(s,a)|, \end{aligned} \tag{H.9}$$

where the second inequality follows from the fact that $\min\{\cdot, H-h\}$ and $\max_{a\in\mathcal{A}}$ are contraction mappings. Here we define Γ_1 and Γ_2 in (H.8) with $\Lambda = \Lambda_1$ and $\Lambda = \Lambda_2$, respectively. Meanwhile, following from the matrix determinant lemma, we have

$$\begin{aligned} \Gamma_1(s,a) &= \sqrt{2}\beta \cdot \Big(\log\det\big(\Lambda_1 + \psi(s,a)\psi(s,a)^\top\big) - \log\det(\Lambda_1)\Big)^{1/2} \\ &= \sqrt{2}\beta \cdot \Big(\log\big(1 + \psi(s,a)^\top\Lambda_1^{-1}\psi(s,a)\big)\Big)^{1/2}, \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \end{aligned}$$

Thus, following from the inequalities $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$ and $|\log(1+x) - \log(1+y)| \leq |x-y|$ for all $x, y \geq 0$, we have

$$\begin{aligned} |\Gamma_1(s,a) - \Gamma_2(s,a)| &\leq \sqrt{2}\beta \cdot \Big(\big|\log\big(1 + \psi(s,a)^\top\Lambda_1^{-1}\psi(s,a)\big) - \log\big(1 + \psi(s,a)^\top\Lambda_2^{-1}\psi(s,a)\big)\big|\Big)^{1/2} \\ &\leq \sqrt{2}\beta \cdot \Big(|\psi(s,a)^\top(\Lambda_1^{-1} - \Lambda_2^{-1})\psi(s,a)|\Big)^{1/2}. \end{aligned} \tag{H.10}$$

Combining (H.14) and (H.10), we have

$$\begin{aligned} d(V_1,V_2) &\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\psi(s,a)^\top(\omega_1-\omega_2) + \Gamma_1(s,a) - \Gamma_2(s,a)| \\ &\leq \sup_{\|\psi\|_2\leq 1} |\psi^\top(\omega_1-\omega_2)| + \sqrt{2}\beta \cdot \sup_{\|\psi\|_2\leq 1} \big(|\psi^\top(\Lambda_1^{-1} - \Lambda_2^{-1})\psi|\big)^{1/2} \\ &= \|\omega_1-\omega_2\|_2 + \|2\beta^2\cdot\Lambda_1^{-1} - 2\beta^2\cdot\Lambda_2^{-1}\|_{\mathrm{OP}}^{1/2} \\ &\leq \|\omega_1-\omega_2\|_2 + \|2\beta^2\cdot\Lambda_1^{-1} - 2\beta^2\cdot\Lambda_2^{-1}\|_{\mathrm{F}}^{1/2}, \end{aligned} \tag{H.11}$$

where we denote by $\|\cdot\|_{\mathrm{OP}}$ and $\|\cdot\|_{\mathrm{F}}$ the operator norm and Frobenius norm, respectively. For $\Lambda \succeq \lambda I$ and $\beta \in [0, B]$, it holds that $\|2\beta^2 \cdot \Lambda^{-1}\|_{\mathrm{F}} \leq 2B^2 d^{1/2}\lambda^{-1}$. Meanwhile, let $\mathcal{N}_{\omega,\epsilon}$ be the $\epsilon/2$-covering number of $\{\omega \in \mathbb{R}^d : \|\omega\|_2 \leq L\}$, and $\mathcal{N}_{A,\epsilon}$ be the $\epsilon^2/4$-covering number of $\{A \in \mathbb{R}^{d\times d} : \|A\|_{\mathrm{F}} \leq 2B^2 d^{1/2}\lambda^{-1}\}$. It is known that [41]

$$\mathcal{N}_{\omega,\epsilon} \leq (1 + 4L/\epsilon)^d, \qquad \mathcal{N}_{A,\epsilon} \leq \big(1 + 16B^2 d^{1/2}/(\lambda\epsilon^2)\big)^{d^2}.$$

Hence, by (H.11), we obtain that

$$\log \mathcal{N}_\epsilon \leq \log(\mathcal{N}_{\omega,\epsilon} \cdot \mathcal{N}_{A,\epsilon}) \leq d \cdot \log(1 + 4L/\epsilon) + d^2 \cdot \log\big(1 + 16B^2 d^{1/2}/(\epsilon^2\lambda)\big),$$

which concludes the proof of Lemma H.4. □

Lemma H.5 (Covering Number of Q [16])**.** Let $\mathcal{Q}$ be a class of functions Q satisfying

$$Q(\cdot,\cdot) = \min\big\{\psi(\cdot,\cdot)^\top \omega + \Gamma(\cdot,\cdot), H - h\big\}, \tag{H.12}$$

where

$$\Gamma(\cdot,\cdot) = \sqrt{2}\beta \cdot \Big(\log\det\big(\Lambda + \psi(\cdot,\cdot)\psi(\cdot,\cdot)^\top\big) - \log\det(\Lambda)\Big)^{1/2}. \tag{H.13}$$

Here the function Q is parameterized by (ω, Λ) and the parameter β is fixed. Let $\psi(\cdot,\cdot)$ be an $\mathbb{R}^d$-valued function and $\Lambda \in \mathbb{R}^{d\times d}$. Let $\|\psi(s,m)\|_2 \leq 1$ for all $(s,m) \in \mathcal{S}\times\mathcal{M}$. For $\|\omega\|_2 \leq L$, $\Lambda \succeq \lambda I$, $\beta \in [0, B]$, and $\epsilon > 0$, there exist an ϵ-covering of $\mathcal{Q}$ with respect to the metric $d(V, V') = \sup_{(s,m)\in\mathcal{S}\times\mathcal{M}} |Q(s,m) - Q'(s,m)|$, such that the covering number $\mathcal{N}_\epsilon$ is upper bounded as follows,

$$\log \mathcal{N}_\epsilon \leq d \cdot \log(1 + 4L/\epsilon) + d^2 \cdot \log\big(1 + 16B^2 d^{1/2}/(\epsilon^2\lambda)\big).$$

Proof. The proof is similar to that of Lemma H.4. Let Q_1 and Q_2 be the functions defined in (H.12), which are parameterized by (ω_1, Λ_1) and (ω_2, Λ_2), respectively. Note that

$$\begin{aligned} d(Q_1, Q_2) \leq & \sup_{(s,m)\in\mathcal{S}\times\mathcal{M}} \Big|\min\big\{\psi(s,m)^\top\omega_1 + \Gamma_1(s,m), H - h\big\} \\ & \qquad - \min\big\{\psi(s,m)^\top\omega_2 + \Gamma_2(s,m), H - h\big\}\Big| \\ \leq & \sup_{(s,m)\in\mathcal{S}\times\mathcal{M}} |\psi(s,m)^\top(\omega_1 - \omega_2) + \Gamma_1(s,m) - \Gamma_2(s,m)|, \end{aligned} \tag{H.14}$$

where the second inequality follows from the fact that $\min\{\cdot, H - h\}$ is a contraction mapping. Here we define Γ_1 and Γ_2 in (H.13) with $\Lambda = \Lambda_1$ and $\Lambda = \Lambda_2$, respectively. The rest of the proof is the same as that of Lemma H.4. We omit the proof and refer to the proof of Lemma H.4 for the details. □

Lemma H.6 (Concentration of Self-Normalized Process)**.** Let $\lambda = 1$ and $\beta = CdH\sqrt{\log(d(T + nH)/\zeta)}$. Let $\zeta > 0$ be an absolute constant. It holds with probability at least $1 - 2\zeta$ that

$$\Big\|\sum_{\ell=1}^4 S_{\ell,h}\Big\|_{(\Lambda_h^k)^{-1}} \leq C'dH\sqrt{\log\big(2(C+1)d(T+nH)/\zeta\big)}, \quad \forall (k,h) \in [K]\times[H].$$

where C and C' are positive absolute constants and C' is independent of C.

Proof. Recall that we define

$$S_{1,h} = \sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^k(s_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(s_h^\tau, a_h^\tau)\big),$$

$$S_{2,h} = \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \cdot \big(V_{h+1}^k(s_{h+1}^i) - (\widetilde{\mathbb{P}}_h V_{h+1}^k)(s_h^i, a_h^i, u_h^i)\big),$$

$$S_{3,h} = \sum_{\tau=1}^{k-1} \psi_h(s_h^\tau, a_h^\tau) \cdot \big(r_h^\tau - R(s_h^\tau, a_h^\tau)\big), \quad S_{4,h} = \sum_{i=1}^{n} \phi_h(s_h^i, a_h^i, u_h^i) \cdot \big(r_h^i - \mathbb{E}[r_h \,|\, s_h^i, a_h^i, u_h^i]\big).$$

We define $\mathcal{F}_{-n+i}$ the σ-algebra generated by the set $\{(s_h^\ell, a_h^\ell, u_h^\ell, r_h^\ell)\}_{(\ell,h)\in[i]\times[H]}$ with timestep index $-n+i$. The set of σ-algebra $\{\mathcal{F}_{-n+i}\}_{i\in[n]}$ captures the data generation process in the offline setting. We attach $\{\mathcal{F}_{-n+i}\}_{i\in[n]}$ to the σ-algebra $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K,H,2]}$ with timestep index t defined in Definition F.1 to obtain the complete filtration. By Lemma H.1 with such a complete filtration, it holds with probability at least $1-\zeta$ that

$$\begin{aligned}&\|S_{1,h}+S_{2,h}\|_{(\Lambda_h^k)^{-1}}\\&\quad\le 4H^2\cdot\Big(d/2\cdot\log\big(\det(\Lambda_h^k)/\det(\Lambda_0)\big)+\log(2\mathcal{N}_\epsilon/\zeta)\Big)+8(n+k)^2\epsilon^2/\lambda,\end{aligned}\tag{H.15}$$

where $\Lambda_0=\lambda I$ and

$$\Lambda_h^k=\sum_{\tau=1}^{k-1}\psi_h(s_h^\tau,a_h^\tau)\psi_h(s_h^\tau,a_h^\tau)^\top+\sum_{i=1}^{n}\phi_h(s_h^i,a_h^i,u_h^i)\phi_h(s_h^i,a_h^i,u_h^i)^\top+\lambda I.$$

Similarly, by Lemma H.1, it holds with probability at least $1-\zeta$ that

$$\|S_{3,h}+S_{4,h}\|_{(\Lambda_h^k)^{-1}}\le 4H^2\cdot\Big(d/2\cdot\log\big(\det(\Lambda_h^k)/\det(\Lambda_0)\big)\Big).\tag{H.16}$$

Note that

$$\begin{aligned}\Lambda_h^k&=\sum_{\tau=1}^{k-1}\psi_h(s_h^\tau,a_h^\tau)\psi_h(s_h^\tau,a_h^\tau)^\top+\sum_{i=1}^{n}\phi_h(s_h^i,a_h^i,u_h^i)\phi_h(s_h^i,a_h^i,u_h^i)^\top+\lambda I\\&\preceq(k+n+\lambda)I.\end{aligned}$$

Meanwhile, recall that $\Lambda_0=\lambda I$. Thus, we obtain that

$$\det(\Lambda_h^k)/\det(\Lambda_0)\le(k+n+\lambda)/\lambda.\tag{H.17}$$

On the other hand, we obtain from Lemma H.3 and Lemma H.4 that

$$\log\mathcal{N}_\epsilon\le d\cdot\big(1+4H\sqrt{d(n+k)}/(\epsilon\sqrt{\lambda})\big)+d^2\cdot\log\big(1+16\beta^2\sqrt{d}/(\epsilon^2\lambda)\big),\tag{H.18}$$

where we set $\beta=CdH\sqrt{\log(d(T+nH)/\zeta)}$. Finally, by setting $\epsilon=dH/(n+k)$ in (H.15), plugging (H.17) and (H.18) into (H.15) and (H.16), respectively, and setting $\lambda=1$, we obtain that

$$\begin{aligned}\Big\|\sum_{\ell=1}^{4}S_{\ell,h}\Big\|_{(\Lambda_h^k)^{-1}}&\le\|S_{1,h}+S_{2,h}\|_{(\Lambda_h^k)^{-1}}+\|S_{3,h}+S_{4,h}\|_{(\Lambda_h^k)^{-1}}\\&\le C'dH\sqrt{\log\big(2(C+1)d(T+nH)/\zeta\big)},\end{aligned}$$

which holds with probability at least $1-2\zeta$. Here $T=HK$ and C, C' are absolute constants, where C' is independent of C. Thus, we complete the proof of Lemma H.6. □

Lemma H.7. Let $\Lambda_t\in\mathbb{R}^{d\times d}$ be a positive definite matrix satisfying $\Lambda_t\succeq I$. Let $\psi_t(\cdot,\cdot)$ be a $\mathbb{R}^d$-valued function such that $\|\psi_t(\cdot,\cdot)\|_2\le 1$. Let $\Lambda_{t+1}(\cdot,\cdot)=\Lambda_t+\psi_t(\cdot,\cdot)\psi_t(\cdot,\cdot)^\top$. It then holds that

$$\psi_t(\cdot,\cdot)^\top(\Lambda_t)^{-1}\psi_t(\cdot,\cdot)\le 2\log\det\big(\Lambda_{t+1}(\cdot,\cdot)\big)-2\log\det(\Lambda_t).$$

Proof. Note that $\Lambda_t\succeq I$. Thus, it holds that

$$0\le\psi_t(\cdot,\cdot)^\top(\Lambda_t)^{-1}\psi_t(\cdot,\cdot)\le\|\psi_t(\cdot,\cdot)\|_2^2\le 1.$$

It then follows from the inequality $x\le 2\log(1+x)$ for all $x\in[0,1]$ that

$$\psi_t(\cdot,\cdot)^\top(\Lambda_t)^{-1}\psi_t(\cdot,\cdot)\le 2\log\big(1+\psi_t(\cdot,\cdot)^\top(\Lambda_t)^{-1}\psi_t(\cdot,\cdot)\big).\tag{H.19}$$

Meanwhile, it follows from the matrix determinant lemma that

$$\det\big(\Lambda_{t+1}(\cdot,\cdot)\big)=\det(\Lambda_t)\cdot\big(1+\psi_t(\cdot,\cdot)^\top(\Lambda_t)^{-1}\psi_t(\cdot,\cdot)\big).\tag{H.20}$$

Finally, combining (H.19) and (H.20), we conclude that

$$\psi_t(\cdot,\cdot)^\top(\Lambda_t)^{-1}\psi_t(\cdot,\cdot)\le 2\log\det\big(\Lambda_{t+1}(\cdot,\cdot)\big)-2\log\det(\Lambda_t),$$

which concludes the proof of Lemma H.7. □