## A Related Work

In the sequel, we discuss the related works.

**RL for solving NE in Markov games** Our work adds to the vast body of existing literature on RL for finding Nash equilibria in Markov games. In particular, there is a line of works that generalizes single-agent RL algorithms to Markov games under either the generative model (Azar et al., 2013) or offline settings with well-explored datasets (Littman, 2001; Greenwald et al., 2003; Hu & Wellman, 2003; Lagoudakis & Parr, 2012; Hansen et al., 2013; Perolat et al., 2015; Jia et al., 2019; Sidford et al., 2020; Cui & Yang, 2020; Fan et al., 2020; Daskalakis et al., 2021; Zhao et al., 2021). These works all aim to find the Nash equilibrium and their algorithms are generalizations of single-agent RL algorithms. In particular, Littman (2001; 1994); Greenwald et al. (2003); Hu & Wellman (2003) generalize Q-learning (Watkins & Dayan, 1992) to Markov games and establish asymptotic convergence guarantees. Jia et al. (2019); Sidford et al. (2020); Zhang et al. (2020a); Cui & Yang (2020) propose variants of Q-learning or value iteration (Shapley, 1953) algorithms under the generative model setting. Moreover, Perolat et al. (2015); Fan et al. (2020) study the sample efficiency of fitted value iteration (Munos & Szepesvári, 2008) for zero-sum Markov games under the offline setting. They assume the behavior policy is explorative in the sense that the concentrability coefficients (Munos & Szepesvári, 2008) are uniformly bounded. Under similar assumptions, Daskalakis et al. (2021); Zhao et al. (2021) study the sample complexity of policy gradient (Sutton et al., 1999) under the well-explored offline setting. Moreover, under the online setting, there is a recent line of research that proposes provably efficient RL algorithms for zero-sum Markov games. See, e.g., Wei et al. (2017); Bai et al. (2020); Bai & Jin (2020); Liu et al. (2020a); Tian et al. (2020); Xie et al. (2020); Chen et al. (2021b) and the references therein. These works propose optimism-based algorithms and establish sublinear regret guarantees for finding NE. Among these works, our work is particularly related to Xie et al. (2020); Chen et al. (2021b), whose algorithms also incorporate the linear function approximation. Compared with these aforementioned works, we focus on solving the Stackelberg-Nash equilibrium, which involves a bilevel structure and is fundamentally different from the Nash equilibrium. Thus, our work is not directly comparable.

**Related single-agent RL methods** Broadly speaking, our work is also related to the recent line of works that achieve sample efficiency in single-agent RL under the online setting. See, e.g., (Azar et al., 2017; Jin et al., 2018; Yang & Wang, 2019; Zanette & Brunskill, 2019; Jin et al., 2020b; Zhou et al., 2020; Ayoub et al., 2020; Yang & Wang, 2020; Zanette et al., 2020b;a; Zhang et al., 2020c;b; Agarwal et al., 2020) and the references therein. In particular, following the *optimism in the face of uncertainty* principle, these works achieve near-optimal regret under either tabular or function approximation settings. Meanwhile, for offline RL with an arbitrary dataset, various recent works propose to utilize pessimism for achieving robustness. See, e.g., (Yu et al., 2020; Kidambi et al., 2020; Kumar et al., 2020; Jin et al., 2020c; Liu et al., 2020b; Buckman et al., 2020; Rashidinejad et al., 2021) and the references therein. These aforementioned works all focus on the single-agent setting and we prove that optimism and pessimism also play an indispensable role in achieving sample efficiency in finding SNE.

## B Notation

We denote by $\|\cdot\|_2$ the $\ell_2$-norm of a vector or the spectral norm of a matrix. We also let $\|\cdot\|_{\text{op}}$ denote the matrix operator norm. Furthermore, for a positive definite matrix $A$, we denote by $\|x\|_A$ the weighted norm $\sqrt{x^\top A x}$ of a vector $x$. Also, we denote by $\Delta(\mathcal{A})$ the set of probability distributions on a set $\mathcal{A}$. For some positive integer $K$, $[K]$ denotes the index set $\{1, 2, \cdots, K\}$.

## C Proof of Theorem 3.3

*Proof of Theorem 3.3.* Recall the regret defined in (3.1) takes the following form

$$\text{Regret}(K) = \underbrace{\sum_{k=1}^{K} V_{l,1}^{\pi^*,\nu^*}(x_1^k) - V_{l,1}^{\pi^k,\nu^*(\pi^k)}(x_1^k)}_{\text{Regret}_l(K)} + \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} V_{f_i,1}^{\pi^k,\nu^*(\pi^k)}(x_1^k) - V_{f_i,1}^{\pi^k,\nu_{f_i}^k,\nu_{f_{-i}}^*(\pi^k)}(x_1^k)}_{\text{Regret}_f(K)}. \tag{C.1}$$

By the leader-controller assumption (Assumption 2.1), we have the following lemma.

**Lemma C.1.** For any $k \in [K]$, we have $\nu^k = \nu^*(\pi^k)$. Here $\nu^*(\cdot)$ is defined in (2.6).

*Proof.* See §C.1 for a detailed proof. □

Combining Lemma C.1 and (C.1), we have

$$\text{Regret}(K) = \sum_{k=1}^{K} [V_{l,1}^{\pi^*,\nu^*}(x_1^k) - V_{l,1}^{\pi^k,\nu^k}(x_1^k)]. \tag{C.2}$$

Then we establish an upper bound for this term. To facilitate our analysis, for any $(k,h) \in [K]\times[H]$ we define the model prediction error by

$$\delta_h^k = r_{l,h} + \mathbb{P}_h V_{h+1}^k - Q_h^k. \tag{C.3}$$

Moreover, for any $(k,h) \in [K] \times [H]$, we define $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ as

$$\begin{aligned}
\zeta_{k,h}^1 &= [V_h^k(x_h^k) - V_{l,h}^{\pi^k,\nu^k}(x_h^k)] - [Q_h^k(x_h^k,a_h^k,b_h^k) - Q_{l,h}^{\pi^k,\nu^k}(x_h^k,a_h^k,b_h^k)],\\
\zeta_{k,h}^2 &= [(\mathbb{P}_h V_{h+1}^k)(x_h^k,a_h^k,b_h^k) - (\mathbb{P}_h V_{l,h+1}^{\pi^k,\nu^k})(x_h^k,a_h^k,b_h^k)] - [V_{h+1}^k(x_{h+1}^k) - V_{l,h+1}^{\pi^k,\nu^k}(x_{h+1}^k)].
\end{aligned} \tag{C.4}$$

Recall that $(\pi^k, \nu^k = \{\nu_{f_i}^k\}_{i\in[N]})$ are the policies executed by the leader and the followers in the $k$-th episode, which generate a trajectory $\{x_h^k, a_h^k, b_h^k = \{b_{i,h}^k\}_{i\in[N]}\}_{h\in[H]}$. Thus, we know that $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ characterize the randomness of choosing actions $a_h^k \sim \pi_h^k(\cdot\,|\,x_h^k)$ and $b_h^k \sim \nu_h^k(\cdot\,|\,x_h^k)$ and the randomness of drawing the next state $x_{h+1}^k \sim \mathcal{P}_h(\cdot\,|\,x_h^k,a_h^k,b_h^k)$, respectively.

To establish an upper bound for (C.2), we introduce the following lemma, which decomposes this term into three parts using the notations defined above.

**Lemma C.2** (Regret Decomposition)**.** We can decompose (C.2) as follows.

$$\begin{aligned}
\text{Regret}(K) = &\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu^*}[\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k)\times\nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k)\times\nu_h^k(\cdot\,|\,x_h^k)\rangle]}_{(l.1)\text{: Computational Error}}\\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^*,\nu^*}[\delta_h^k(x_h,a_h,b_h)] - \delta_h^k(x_h^k,a_h^k,b_h^k)\big)}_{(l.2)\text{: Statistical Error}} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}(\zeta_{k,h}^1+\zeta_{k,h}^2)}_{(l.3)\text{: Randomness}},
\end{aligned}$$

where $\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k)\times\nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k)\times\nu_h^k(\cdot\,|\,x_h^k)\rangle = \langle Q_h^k(x_h^k,\cdot,\cdot,\cdots,\cdot), \pi_h^*(\cdot\,|\,x_h^k)\times\nu_{f_1,h}^*(\cdot\,|\,x_h^k)\times\cdots\nu_{f_N,h}^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k)\times\nu_{f_1,h}^k(\cdot\,|\,x_h^k)\times\cdots\nu_{f_N,h}^k(\cdot\,|\,x_h^k)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$.

*Proof.* See §C.2 for a detailed proof. □

**Remark C.3.** Similar regret decomposition results also appear in the single-agent RL literature (Cai et al., 2020; Efroni et al., 2020; Yang et al., 2020), and they can be regarded as the special case of Lemma C.2. Moreover, our regret decomposition lemma is independent of the leader-controller linear setting in Assumption 2.1, and thus, can be applied to more general settings.

Lemma C.2 states that the regret has three sources: (i) computational error, which represents the convergence of the algorithm with the known model, (ii) statistical error, that is, the error caused by the inaccurate estimation of the model, and (iii) randomness, as aforementioned, which comes from executing random policies and interaction with random environment.

Returning to the main proof, we only need to characterize these three types of errors, respectively. We first characterize the computational error by the following lemma.

**Lemma C.4** (Optimization Error)**.** It holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\nu^*}[\langle Q_h^k(x_h^k,\cdot,\cdot),\pi_h^*(\cdot\,|\,x_h^k)\times\nu_h^*(\cdot\,|\,x_h^k)-\pi_h^k(\cdot\,|\,x_h^k)\times\nu_h^k(\cdot\,|\,x_h^k)\rangle]\le\epsilon KH.$$

*Proof.* See §C.3 for a detailed proof. □

Then, we establish an upper bound for the statistical error. Due to the uncertainty that arises from only observing limited data, the model prediction errors can be possibly large for the triple $(x,a,b)$ that are less visited or even unseen. Fortunately, however, we have the following lemma which characterizes the model prediction errors defined in (C.3).

**Lemma C.5** (Optimism)**.** It holds with probability at least $1-p/2$ that

$$-2\min\{H,\Gamma_h^k(x,a)\}\le\delta_h^k(x,a,b)\le 0$$

for any $(k,h)\in[K]\times[H]$ and $(x,a,b)\in\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$.

*Proof.* See §C.4 a detailed proof. □

Lemma C.5 states that $\delta_h^k(x,a,b)\le 0$ for any $(x,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{A}$. Combining the definition of model prediction error in (C.3), we obtain

$$Q_h^k(x,a,b)\ge r_{l,h}(x,a,b)+(\mathbb{P}_hV_{h+1}^k)(x,a,b),$$

which further implies that the estimated Q-function $Q_{\star,h}^k$ is "optimistic in the face of uncertainty". Moreover, Lemma C.5 implies that $-\delta_h^k(x,a,b)\le 2\min\{H,\Gamma_h^k(x,a)\}$. Thus we only need to establish an upper bound for $2\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{H,\Gamma_h^k(x_h^k,a_h^k)\}$, which is the total price paid for the optimism. As shown in the following lemma, we can derive an upper bound for this term by the elliptical potential lemma (Abbasi-Yadkori et al., 2011).

**Lemma C.6.** For the bonus function $\Gamma_h^k$ defined in Line 7 of Algorithm 1, it holds that

$$2\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{H,\Gamma_h^k(x_h^k,a_h^k)\}\le\mathcal{O}(\sqrt{d^2H^3T\iota^2}).$$

Here $p\in(0,1)$ and $\iota=\log(2dT/p)$ are defined in Theorem 3.3.

*Proof.* See §C.5 for a detailed proof. □

It remains to analyze the randomness, which is the purpose of the following lemma.

**Lemma C.7.** For the $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ defined in Lemma C.2 and any $p\in(0,1)$, it holds with probability at least $1-p/2$ that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(\zeta_{k,h}^1+\zeta_{k,h}^2)\le\sqrt{16KH^3\cdot\log(4/p)}.$$

*Proof.* See §C.6 for a detailed proof. □

Putting above lemmas together, we obtain

$$\text{Regret}(K)\le\mathcal{O}(\sqrt{d^2H^3T\iota^2}) \tag{C.5}$$

with probability at least $1-p$, which concludes the proof of Theorem 3.3. □

### C.1 Proof of Lemma C.1

*Proof of Lemma C.1.* Fix $k \in [K]$, by the definition of the best response in (2.5), we have

$$\begin{aligned}\mathrm{BR}(\pi^k) &= \{\nu = \{\nu_{f_i}\}_{i\in[N]} \,|\, \nu \text{ is the NE of the followers given the leader policy } \pi^k\}\\ &= \{\nu = \{\nu_{f_i}\}_{i\in[N]} \,|\, \nu \text{ is the NE of } \{V_{f_i,h}^{\pi^k,\nu}(x)\}_{i\in[N]},\ \forall h \in [H] \text{ and } x \in \mathcal{S}\}\\ &= \{\nu = \{\nu_{f_i}\}_{i\in[N]} \,|\, \nu \text{ is the NE of } \{r_{f_i,h}^{\pi^k,\nu}(x)\}_{i\in[N]},\ \forall h \in [H] \text{ and } x \in \mathcal{S}\}, \qquad \text{(C.6)}\end{aligned}$$

where $r_{f_i,h}^{\pi^k,\nu}(x) = \langle r_{f_i,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x) \times \nu_{f_1,h}(\cdot\,|\,x) \times \cdots \times \nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here the last inequality uses Bellman equality (2.2) and Assumption 2.1, which assumes the leader-controller game. Moreover, by the definition of $\nu^*(\pi^k)$ defined in (2.6), we have that

$$\nu_h^*(\pi^k) = \{\nu_{f_i,h}^*(\pi^k)\}_{i\in[N]} \in \operatorname*{argmin}_{\nu\in\mathrm{BR}(\pi^k)} V_{l,h}^{\pi^k,\nu}(x) = \operatorname*{argmin}_{\nu\in\mathrm{BR}(\pi^k)} r_{l,h}^{\pi^k,\nu}(x), \qquad \text{(C.7)}$$

where $r_{l,h}^{\pi^k,\nu}(x) = \langle r_{l,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x) \times \nu_{f_1,h}(\cdot\,|\,x) \times \cdots \times \nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here the last equality uses Assumption 2.1.

Recall that, in the subroutine $\epsilon$-SNE (Algorithm 2), we pick the function $\widetilde{Q} \in \mathcal{Q}_{h,\epsilon}^k$ such that $\|Q_h^k - \widetilde{Q}\|_\infty \le \epsilon$ and solve the matrix game defined in (3.6). Here $\mathcal{Q}_{h,\epsilon}^k$ is the class of functions $Q : \mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f \to \mathbb{R}$ that takes form

$$Q(\cdot,\cdot,\cdot) = r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\big\{\phi(\cdot,\cdot)^\top w + \beta\cdot\big(\phi(\cdot,\cdot)^\top\Lambda^{-1}\phi(\cdot,\cdot)\big)^{1/2}\big\}, \qquad \text{(C.8)}$$

where $\|w\|_2 \le H\sqrt{dk}$ and $\lambda_{\min}(\Lambda) \ge 1$. Thus, given the leader policy $\pi^k$, the best response of the followers for the matrix game defined in (3.6) takes the form

$$\begin{aligned}\mathrm{BR}'(\pi^k) &= \{\nu \,|\, \nu \text{ is the NE of } \{\langle r_{f_i,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle\}_{i\in[N]}, \forall h\in[H] \text{ and } x\in\mathcal{S}\}\\ &= \mathrm{BR}(\pi^k) \qquad \text{(C.9)}\end{aligned}$$

where $\langle r_{f_i,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle$ is the shorthand of $\langle r_{f_i,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_{f_1,h}(\cdot\,|\,x)\times\cdots\times\nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here the last equality uses (C.6). Similarly, by the definition of $\mathcal{Q}_{h,\epsilon}^k$ in (C.8), we can obtain that

$$\operatorname*{argmin}_{\nu_h}\langle\widetilde{Q}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle = \operatorname*{argmin}_{\nu_h}\langle r_{l,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle, \qquad \text{(C.10)}$$

where $\langle r_{l,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle$ is the abbreviation of $\langle r_{f_i,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_{f_1,h}(\cdot\,|\,x)\times\cdots\times\nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$ Together with (C.7) and (C.9), we have that, for the matrix game with payoff matrices $(\widetilde{Q}(x_h^k,\cdot,\cdot), \{r_{f_i,h}^k(x_h^k,\cdot,\cdot)\}_{i\in[N]})$, the policy $\nu_h^k(\cdot\,|\,x_h^k) = \{\nu_{f_i,h}^k(\cdot\,|\,x_h^k)\}_{i\in[N]}$ is also the best response of $\pi_h^k(\cdot\,|\,x_h^k)$ and breaks ties against favor of the leader. Therefore, we have $\nu^k = \nu^*(\pi^k)$ for any $k \in [K]$, which concludes the proof of Lemma C.1. □

### C.2 Proof of Lemma C.2

First, we establish a more general regret decomposition lemma, which immediately implies Lemma C.2.

**Lemma C.8** (General Decomposition for One Episode)**.** Fix $k \in [K]$. Suppose $(\pi^k, \nu^k = \{\nu_{f_i}^k\}_{i\in[N]})$ are the policies executed by the leader $l$ and the followers $\{f_i\}_{i\in[N]}$ in the $k$-th episode. Moreover, suppose that $Q_{\star,h}^k$ and $V_{\star,h}^k = \langle Q_{\star,h}^k, \pi_h^k\times\nu_h^k\rangle$ are the estimated Q-function and value function for any $\star \in \{l, f_1, \cdots, f_N\}$ at $h$-th step of $k$-th episode. Then, for any policies

$(\pi, \nu = \{\nu_{f_i}\}_{i\in[N]})$ and $\star \in \{l, f_1, \cdots, f_N\}$, we have

$$
\begin{aligned}
&V_{\star,1}^{\pi,\nu}(x_1^k) - V_{\star,1}^{\pi^k,\nu^k}(x_1^k) \\
&\quad = \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi,\nu}[\langle Q_{\star,h}^k(x_h^k,\cdot,\cdot), \pi_h(\cdot\,|\,x_h^k)\times\nu_h(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k)\times\nu_h^k(\cdot\,|\,x_h^k)\rangle]}_{\text{Computational Error}} \\
&\qquad + \underbrace{\sum_{h=1}^{H}\big(\mathbb{E}_{\pi,\nu}[\delta_{\star,h}^k(x_h,a_h,b_h)] - \delta_{\star,h}^k(x_h^k,a_h^k,b_h^k)\big)}_{\text{Statistical Error}} + \underbrace{\sum_{h=1}^{H}(\zeta_{\star,k,h}^1 + \zeta_{\star,k,h}^2)}_{\text{Randomness}},
\end{aligned}
$$

where $\langle Q_{\star,h}^k, \pi_h^k\times\nu_h^k\rangle = \langle Q_{\star,h}^k, \pi_h^k\times\nu_{f_1,h}^k\times\cdots\times\nu_{f_N,h}^k\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$ and $\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k)\times\nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k)\times\nu_h^k(\cdot\,|\,x_h^k)\rangle = \langle Q_h^k(x_h^k,\cdot,\cdot,\cdots,\cdot), \pi_h^*(\cdot\,|\,x_h^k)\times\nu_{f_1,h}^*(\cdot\,|\,x_h^k)\times\cdots\nu_{f_N,h}^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k)\times\nu_{f_1,h}^k(\cdot\,|\,x_h^k)\times\cdots\nu_{f_N,h}^k(\cdot\,|\,x_h^k)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here $\delta_{\star,h}^k$ is the model prediction error defined by

$$
\delta_{\star,h}^k = r_{\star,h} + \mathbb{P}_h V_{\star,h+1}^k - Q_{\star,h}^k, \tag{C.11}
$$

and $\zeta_{\star,k,h}^1$ and $\zeta_{\star,k,h}^2$ are defined by

$$
\begin{aligned}
\zeta_{\star,k,h}^1 &= [V_{\star,h}^k(x_h^k) - V_{\star,h}^{\pi^k,\nu^k}(x_h^k)] - [Q_{\star,h}^k(x_h^k,a_h^k,b_h^k) - Q_{\star,h}^{\pi^k,\nu^k}(x_h^k,a_h^k,b_h^k)], \\
\zeta_{\star,k,h}^2 &= [(\mathbb{P}_h V_{\star,h+1}^k)(x_h^k,a_h^k,b_h^k) - (\mathbb{P}_h V_{\star,h+1}^{\pi^k,\nu^k})(x_h^k,a_h^k,b_h^k)] - [V_{\star,h+1}^k(x_{h+1}^k) - V_{\star,h+1}^{\pi^k,\nu^k}(x_{h+1}^k)].
\end{aligned} \tag{C.12}
$$

*Proof of Lemma C.8.* To facilitate our analysis, for any $\nu = \{\nu_{f_i}\}_{i\in[N]}$ and $(h,x)\in[H]\times\mathcal{S}$, we denote $\nu_{f_1,h}(\cdot\,|\,x)\times\cdots\nu_{f_N,h}(\cdot\,|\,x)$ by $\nu_h(\cdot\,|\,x)$. Moreover, we define two operators $\mathbb{J}_h$ and $\mathbb{J}_{k,h}$ respectively by

$$
\begin{aligned}
(\mathbb{J}_h f)(x) &= \langle f(x,\cdot,\cdot), \pi_h(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle, \\
(\mathbb{J}_{k,h} f)(x) &= \langle f(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h^k(\cdot\,|\,x)\rangle
\end{aligned} \tag{C.13}
$$

for any $h\in[H]$ and any function $f : \mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f\to\mathbb{R}$. Also, we define

$$
\begin{aligned}
\xi_{\star,h}^k(x) &= (\mathbb{J}_h Q_{\star,f}^k)(x) - (\mathbb{J}_{k,h} Q_{\star,f}^k)(x) \\
&= \langle Q_{\star,h}^k(x,\cdot,\cdot), \pi_h(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x) - \pi_h^k(\cdot\,|\,x)\times\nu_h^k(\cdot\,|\,x)\rangle
\end{aligned} \tag{C.14}
$$

for any $(h,x)\in[H]\times\mathcal{S}$ and $\star\in\{l, f_1, \cdots, f_N\}$.

Under the above notations, we decompose the regret at the $k$-th episode into the following two terms,

$$
V_{\star,1}^{\pi,\nu}(x_1^k) - V_1^{\pi^k,\nu^k}(x_1^k) = \underbrace{V_{\star,1}^{\pi,\nu}(x_1^k) - V_{\star,1}^k(x_1^k)}_{\text{(i)}} + \underbrace{V_{\star,1}^k(x_1^k) - V_1^{\pi^k,\nu^k}(x_1^k)}_{\text{(ii)}}. \tag{C.15}
$$

Then we characterize these two terms respectively.

**Term (i).** By the Bellman equation in (2.2) and the definition of the operator $\mathbb{J}_h$ in (C.13), we have $V_{\star,h}^{\pi,\nu} = \mathbb{J}_h Q_{\star,h}^{\pi,\nu}$. Similar, by the definition of $V_{\star,h}^k$ and the definition of the operator $\mathbb{J}_{k,h}$ in (C.13), we have $V_{\star,h}^k = \mathbb{J}_{k,h} Q_{\star,h}^k$. Hence, for any $h\in[H]$, we have

$$
\begin{aligned}
V_{\star,h}^{\pi,\nu} - V_{\star,h}^k &= \mathbb{J}_h Q_{\star,h}^{\pi,\nu} - \mathbb{J}_{k,h} Q_{\star,h}^k = (\mathbb{J}_h Q_{\star,h}^{\pi,\nu} - \mathbb{J}_h Q_{\star,h}^k) + (\mathbb{J}_h Q_{\star,h}^k - \mathbb{J}_{k,h} Q_{\star,h}^k) \\
&= \mathbb{J}_h(Q_{\star,h}^{\pi,\nu} - Q_{\star,h}^k) + \xi_{\star,h}^k,
\end{aligned} \tag{C.16}
$$

where the last inequality is obtained by the fact that $\mathbb{J}_h$ is a linear operator and the definition of $\xi_{\star,h}^k$ in (C.14). Meanwhile, by the Bellman equation in (2.2) and the definition of the prediction error $\delta_{\star,h}^k$ in (C.3), we obtain

$$
\begin{aligned}
Q_{\star,h}^{\pi,\nu} - Q_{\star,h}^k &= (r_{\star,h} + \mathbb{P}_h V_{\star,h+1}^{\pi,\nu}) - (r_{\star,h} + \mathbb{P}_h V_{\star,h+1}^k - \delta_{\star,h}^k) \\
&= \mathbb{P}_h(V_{\star,h+1}^{\pi,\nu} - V_{\star,h+1}^k) + \delta_{\star,h}^k.
\end{aligned} \tag{C.17}
$$

Putting (C.16) and (C.17) together, we further obtain

$$V^{\pi,\nu}_{\star,h} - V^k_{\star,h} = \mathbb{J}_h\mathbb{P}_h(V^{\pi,\nu}_{\star,h+1} - V^k_{\star,h+1}) + \mathbb{J}_h\delta^k_{\star,h} + \xi^k_{\star,h} \tag{C.18}$$

for any $h \in [H]$ and $\star \in \{l, f_1, \cdots, f_N\}$. By recursively applying (C.18) for all $h \in [H]$, we have

$$\begin{aligned} V^{\pi,\nu}_{\star,1} - V^k_{\star,1} &= \Big(\prod_{h=1}^H \mathbb{J}_h\mathbb{P}_h\Big)(V^{\pi,\nu}_{\star,H+1} - V^{\pi^k,\nu^k}_{\star,H+1}) + \sum_{h=1}^H\Big(\sum_{i=1}^h \mathbb{J}_i\mathbb{P}_i\Big)\mathbb{J}_h\delta^k_{\star,h} + \sum_{h=1}^H\Big(\sum_{i=1}^h \mathbb{J}_i\mathbb{P}_i\Big)\xi^k_{\star,h} \\ &= \sum_{h=1}^H\Big(\sum_{i=1}^h \mathbb{J}_i\mathbb{P}_i\Big)\mathbb{J}_h\delta^k_{\star,h} + \sum_{h=1}^H\Big(\sum_{i=1}^h \mathbb{J}_i\mathbb{P}_i\Big)\xi^k_{\star,h}, \end{aligned} \tag{C.19}$$

where the last equality follows from the fact that $V^{\pi,\nu}_{\star,H+1} = V^{\pi^k,\nu^k}_{\star,H+1} = 0$. Thus, by utilizing the definition of $\xi^k_{\star,h}$ in (C.14), we further obtain

$$\begin{aligned} V^{\pi,\nu}_{\star,1}(x^k_1) - V^k_{\star,1}(x^k_1) = \mathbb{E}_{\pi,\nu}\Bigg[\sum_{h=1}^H \langle Q^k_{\star,h}(x^k_h,\cdot,\cdot), \pi_h(\cdot\,|\,x^k_h)\times\nu_h(\cdot\,|\,x^k_h) - \pi^k_h(\cdot\,|\,x^k_h)\times\nu^k_h(\cdot\,|\,x^k_h)\rangle\Bigg] \\ + \mathbb{E}_{\pi,\nu}\Bigg[\sum_{h=1}^H \delta^k_{\star,h}(x_h,a_h,b_h)\Bigg] \end{aligned} \tag{C.20}$$

for any $k \in [K]$ and $\star \in \{l, f_1, \cdots, f_N\}$.

**Term (ii).** Recall that we denote $\{b^k_{f_i,h}\}_{i\in[N]}$ by $b^k_h$ for any $h \in [H]$. Then, for any $h \in [H]$ and $\star \in \{l, f_1, \cdots, f_N\}$, by the definition of model prediction error in (C.11), we have

$$\begin{aligned} \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h) &= r^k_{\star,h}(x^k_h,a^k_h,b^k_h) + (\mathbb{P}_hV^k_{\star,h+1})(x^k_h,a^k_h,b^k_h) - Q^k_{\star,h}(x^k_h,a^k_h,b^k_h) \\ &= [r^k_{\star,h}(x^k_h,a^k_h,b^k_h) + (\mathbb{P}_hV^k_{\star,h+1})(x^k_h,a^k_h,b^k_h) - Q^{\pi^k,\nu^k}_{\star,h}(x^k_h,a^k_h,b^k_h)] \\ &\quad + [Q^{\pi^k,\nu^k}_{\star,h}(x^k_h,a^k_h,b^k_h) - Q^k_{\star,h}(x^k_h,a^k_h,b^k_h)] \\ &= (\mathbb{P}_hV^k_{\star,h+1} - \mathbb{P}_hV^{\pi^k,\nu^k}_{\star,h+1})(x^k_h,a^k_h,b^k_h) + (Q^{\pi^k,\nu^k}_{\star,h} - Q^k_{\star,h})(x^k_h,a^k_h,b^k_h) \end{aligned} \tag{C.21}$$

where the last equation is obtained by the Bellman equation in (2.2). Thus, by (C.21), we have

$$\begin{aligned} &V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) \\ &\quad= V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) + (Q^{\pi^k,\nu^k}_{\star,h} - Q^k_{\star,h})(x^k_h,a^k_h,b^k_h) \\ &\qquad + (\mathbb{P}_hV^k_{\star,h+1} - \mathbb{P}_hV^{\pi^k,\nu^k}_{\star,h+1})(x^k_h,a^k_h,b^k_h) - \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h) \\ &\quad= V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) - (Q^k_{\star,h} - Q^{\pi^k,\nu^k}_{\star,h})(x^k_h,a^k_h,b^k_h) \\ &\qquad + \big(\mathbb{P}_h(V^k_{\star,h+1} - V^{\pi^k,\nu^k}_{\star,h+1})\big)(x^k_h,a^k_h,b^k_h) - (V^k_{\star,h+1} - V^{\pi^k,\nu^k}_{\star,h+1})(x^k_h) \\ &\qquad + (V^k_{\star,h+1} - V^{\pi^k,\nu^k}_{\star,h+1})(x^k_h) - \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h) \end{aligned} \tag{C.22}$$

for any $h \in [H]$ and $\star \in \{l, f_1, \cdots, f_N\}$. By the definitions of $\zeta^1_{\star,k,h}$ and $\zeta^2_{\star,k,h}$ in (C.12), (C.22) can be written as

$$V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) = [V^k_{\star,h+1}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h+1}(x^k_h)] + \zeta^1_{\star,k,h} + \zeta^2_{\star,k,h} - \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h). \tag{C.23}$$

For any $\star \in \{l, f_1, \cdots, f_N\}$, recursively expanding (C.23) across $h \in [H]$ yields

$$\begin{aligned} &V^k_{\star,1}(x^k_1) - V^{\pi^k,\nu^k}_{\star,1}(x^k_1) \\ &\quad= V^k_{\star,H+1}(x^k_{H+1}) - V^{\pi^k,\nu^k}_{\star,H+1}(x^k_{H+1}) + \sum_{h=1}^H(\zeta^1_{\star,k,h} + \zeta^2_{\star,k,h}) - \sum_{h=1}^H \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h) \\ &\quad= \sum_{h=1}^H(\zeta^1_{\star,k,h} + \zeta^2_{\star,k,h}) - \sum_{h=1}^H \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h), \end{aligned} \tag{C.24}$$

where the last equality follows from the fact that $V_{\star,H+1}^{k}(x_{H+1}^{k}) = V_{\star,H+1}^{\pi^k,\nu^k}(x_{H+1}^{k}) = 0$.

Plugging (C.20) and (C.24) into (C.15), we obtain

$$
\begin{aligned}
&V_{\star,1}^{\pi,\nu}(x_1^k) - V_{\star,1}^{\pi^k,\nu^k}(x_1^k) \\
&\quad = \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi,\nu}[\langle Q_{\star,h}^{k}(x_h^k,\cdot,\cdot), \pi_h(\cdot\,|\,x_h^k) \times \nu_h(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle]}_{\text{Computational Error}} \\
&\qquad + \underbrace{\sum_{h=1}^{H} \big(\mathbb{E}_{\pi,\nu}[\delta_{\star,h}^{k}(x_h,a_h,b_h)] - \delta_{\star,h}^{k}(x_h^k,a_h^k,b_h^k)\big)}_{\text{Statistical Error}} + \underbrace{\sum_{h=1}^{H} (\zeta_{\star,k,h}^{1} + \zeta_{\star,k,h}^{2})}_{\text{Randomness}}
\end{aligned}
$$

for any $(\pi,\nu)$ and $\star \in \{l, f_1, \cdots, f_N\}$. Therefore, we conclude the proof of Lemma C.2. □

*Proof of Lemma C.2.* For any $k \in [K]$, applying Lemma C.8 with $(\pi,\nu) = (\pi^*,\nu^*)$, we obtain

$$
\begin{aligned}
&V_{l,1}^{\pi^*,\nu^*}(x_1^k) - V_{l,1}^{\pi^k,\nu^k}(x_1^k) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu^*}[\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle] \\
&\qquad + \sum_{h=1}^{H} \big(\mathbb{E}_{\pi^*,\nu^*}[\delta_h^k(x_h,a_h,b_h)] - \delta_h^k(x_h^k,a_h^k,b_h^k)\big) + \sum_{h=1}^{H} (\zeta_{k,h}^{1} + \zeta_{k,h}^{2}).
\end{aligned}
$$

Taking summation over $k \in [K]$, we decompose (C.2) as desired, which concludes the proof of Lemma C.2.

□

### C.3 Proof of Lemma C.4

*Proof of Lemma C.4.* By the same argument in §C.1 (replacing $\pi^k$ by $\pi^*$), we have that, for the matrix game with payoff matrices $(\widetilde{Q}(x_h^k,\cdot,\cdot), \{r_{f_i,h}^k(x_h^k,\cdot,\cdot)\}_{i\in[N]})$, $\nu_h^*(\cdot\,|\,x_h^k)$ is also the best response of $\pi_h^*(\cdot\,|\,x_h^k)$ and breaks ties against favor of the leader.

Recall that $(\pi_h^k(\cdot\,|\,x_h^k), \nu_h^k(\cdot\,|\,x_h^k) = \{\nu_{f_i,h}^k(\cdot\,|\,x_h^k)\}_{i\in[N]})$ is the Stackelberg-Nash equilibrium of the matrix game with payoff matrices $(\widetilde{Q}(x_h^k,\cdot,\cdot,\cdot), \{r_{f_i,h}^k(x_h^k,\cdot,\cdot,\cdot)\}_{i\in[N]})$, which implies that $\pi_h^k(\cdot\,|\,x_h^k)$ is the "best response to the best response", which further implies that

$$
\langle \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \le 0 \tag{C.25}
$$

for any $(k,h) \in [K] \times [H]$. Thus, for any $(k,h) \in [K] \times [H]$, we have

$$
\begin{aligned}
&\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \\
&\quad = \langle \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \\
&\qquad + \langle Q_h^k(x_h^k,\cdot,\cdot) - \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \\
&\quad \le \epsilon,
\end{aligned} \tag{C.26}
$$

where the last inequality uses (C.25) and the fact that $\|Q_h^k - \widetilde{Q}\|_\infty \le \epsilon$. By taking summation over $(k,h) \in [K] \times [H]$, we conclude the proof of Lemma C.4.

□

### C.4 Proof of Lemma C.5

*Proof of Lemma C.5.* Recall that the estimated Q-function $Q_h^k$ defined in Line 8 of Algorithm 1 takes the following form:

$$\begin{aligned} &Q_h^k(\cdot,\cdot,\cdot) \leftarrow r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot)^\top w_h^k + \Gamma_h^k(\cdot,\cdot)\}, \\ &\text{where } w_h^k = (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(x_h^\tau, a_h^\tau)\cdot V_{h+1}^k(x_{h+1}^\tau)\Big). \end{aligned} \tag{C.27}$$

Here $\Lambda_h^k$ and $\Gamma_h^k$ are defined in Lines 5 and 7 of Algorithm 1, respectively. Meanwhile, by Assumption 2.1, we have

$$\begin{aligned} (\mathbb{P}_h V_{h+1}^k)(x,a,b) &= \phi(x,a)^\top \langle \mu_h, V_{h+1}^k\rangle \\ &= \phi(x,a)^\top (\Lambda_h^k)^{-1}\Lambda_h^k \langle \mu_h, V_{h+1}^k\rangle \end{aligned} \tag{C.28}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$. Here $\langle \mu_h, V_{h+1}^k\rangle = \int_{\mathcal{S}} V_{h+1}^k(x')\mathrm{d}\mu_h(x')$. Together with the definition of $\Lambda_h^k$ in Line 5 of Algorithm 1, we further obtain

$$\begin{aligned} (\mathbb{P}_h V_{h+1}^k)(x,a,b) &= \phi(x,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\phi(x_h^\tau,a_h^\tau)^\top \langle \mu_h, V_{h+1}^k\rangle + \langle \mu_h, V_{h+1}^k\rangle\Big) \\ &= \phi(x,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\cdot (\mathbb{P}_h V_{h+1}^k)(x_h^\tau,a_h^\tau,b_h^\tau) + \langle \mu_h, V_{h+1}^k\rangle\Big), \end{aligned} \tag{C.29}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$. Here the last equality uses (C.28). Putting (C.27) and (C.29) together, we have

$$\begin{aligned} &\phi(x,a)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a,b) \\ &\quad= \underbrace{\phi(x,a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\cdot\big(V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau,a_h^\tau,b_h^\tau)\big)\Big)}_{\text{(i)}} \\ &\qquad - \underbrace{\phi(x,a)^\top (\Lambda_h^k)^{-1}\langle \mu_h, V_{h+1}^k\rangle}_{\text{(ii)}} \end{aligned} \tag{C.30}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$. Then we upper bound these two terms respectively.

**Term (i).** By Cauchy-Schwarz inequality, we have

$$|\text{(i)}| \le \|\phi\|_{(\Lambda_h^k)^{-1}} \cdot \Big\|\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\cdot\big(V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau,a_h^\tau,b_h^\tau)\big)\Big\|_{(\Lambda_h^k)^{-1}} \tag{C.31}$$

for any $(k,h,x,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l$. Under the event $\mathcal{E}$ defined in Lemma C.9, we further have

$$|\text{(i)}| \le C' dH\sqrt{\log(2dT/p)}\cdot\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} \tag{C.32}$$

for any $(k,h,x,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l$.

**Term (ii).** Similarly, by Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} |\text{(ii)}| &\le \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\cdot\|\langle \mu_h, V_{h+1}^k\rangle\|_{(\Lambda_h^k)^{-1}} \\ &\le \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}}\cdot\|\langle \mu_h, V_{h+1}^k\rangle\|_2 \le \sqrt{d}\cdot\|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} \end{aligned} \tag{C.33}$$

for any $(k,h,x,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l$. Here the second inequality follows from the fact that $\Lambda_h^k \succeq I$ and the last inequality is obtained by

$$\|\langle \mu_h, V_{h+1}^k\rangle\|_2 \le \|V_{h+1}^k\|_\infty\cdot\|\mu_h(\mathcal{S})\|_2 \le H\sqrt{d}.$$

Here we use the fact that $\|V_{h+1}^k\|_\infty \le H$ and Assumption 2.1, which assumes $\|\mu_h(\mathcal{S})\|_2 \le \sqrt{d}$. Plugging (C.32) and (C.33) into (C.30), we obtain that

$$|\phi(x,a)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a,b)| \le CdH\sqrt{\log(2dT/p)} \cdot \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} \tag{C.34}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$ under the event $\mathcal{E}$. Here $C>0$ is a constant. By setting

$$\beta = CdH\sqrt{\log(2dT/p)} \tag{C.35}$$

in Line 7 of Algorithm 1, (C.34) gives

$$|\phi(x,a)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a,b)| \le \Gamma_h^k(x,a) \tag{C.36}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$ under the event $\mathcal{E}$. Meanwhile, by the truncation in Line 8 of Algorithm 1 and the fact that $r_{l,h} \in [-1,1]$, we have $Q_h^k \in [-(H-h+1), H-h+1]$, which further implies that

$$V_h^k \in [-(H-h+1), H-h+1] \tag{C.37}$$

for any $(k,h) \in [K]\times[H]$. Hence, by (C.36), we have

$$\phi(x,a)^\top w_h^k + \Gamma_h^k(x,a) \ge (\mathbb{P}_h V_{h+1}^k)(x,a,b) \ge H-h \tag{C.38}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$ under the event $\mathcal{E}$, where the last inequality is obtained by (C.37). Thus, for the model prediction error defined in (C.3), we have

$$\begin{aligned} -\delta_h^k(x,a,b) &= Q_h^k(x,a,b) - r_{l,h}(x,a,b) - \mathbb{P}_h V_{h+1}^k(x,a,b) \\ &\le \phi(x,a)^\top w_h^k + \Gamma_h^k(x,a) - \mathbb{P}_h V_{h+1}^k(x,a,b) \\ &\le 2\Gamma_h^k(x,a) \end{aligned} \tag{C.39}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$ under the event $\mathcal{E}$. Moreover, by the definition of the model prediction error, we have $-\delta_h^k(\cdot,\cdot,\cdot) \le 2H$. Together with (C.39), we have

$$-\delta_h^k(x,a,b) \le 2\min\{H, \Gamma_h^k(x,a)\} \tag{C.40}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$ under the event $\mathcal{E}$. On the other hand, by (3.4), we have

$$\begin{aligned} \delta_h^k(x,a,b) &= r_{l,h}(x,a,b) + \mathbb{P}_h V_{h+1}^k(x,a,b) - Q_h^k(x,a,b) \\ &\le \mathbb{P}_h V_{h+1}^k(x,a,b) - \min\{\phi(x,a)^\top w_h^k + \Gamma_h^k(x,a), H-h\} \\ &= \max\{\mathbb{P}_h V_{h+1}^k(x,a,b) - \phi(x,a)^\top w_h^k - \Gamma_h^k(x,a), \mathbb{P}_h V_{h+1}^k(x,a,b) - (H-h)\} \\ &\le 0 \end{aligned} \tag{C.41}$$

for any $(k,h,x,a,b) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$ under the event $\mathcal{E}$. Here the last inequality follows from (C.36) and the fact that $V_{h+1}^k \le H-h$. Combining (C.40) and (C.41), we conclude the proof of Lemma C.5. □

**Lemma C.9.** For any $p \in (0,1]$, the event $\mathcal{E}$ that, for any $(k,h) \in [K]\times[H]$,

$$\Big\| \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau, b_h^\tau)\big) \Big\|_{(\Lambda_h^k)^{-1}} \le C' dH\sqrt{\log(2dT/p)}$$

happens with probability at least $1-p/2$, where $C'>0$ is an absolute constant.

*Proof of Lemma C.9.* Fix $(k,h) \in [K]\times[H]$. By Lemma C.10, we have $w_{h+1}^k \le H\sqrt{dk}$, which implies that $Q_{h+1}^k \in \mathcal{Q}_{h+1,\epsilon}^k$. Here $\mathcal{Q}_{h+1,\epsilon}^k$ is defined in (3.5). Moreover, as shown in Algorithm 2, we find a $\widetilde{Q}$ in the $\epsilon$-net $\mathcal{Q}_{h+1,\epsilon}^k$ such that $\|Q_{h+1}^k - \widetilde{Q}\|_\infty \le \epsilon$. For any $x \in \mathcal{S}$, let

$(\widetilde{\pi}(\cdot\,|\,x), \widetilde{\nu} = \{\nu_{f_i}\}_{i=1}^N)$ be the Stackelberg-Nash equilibrium of the matrix game with payoff matrices $(\widetilde{Q}(x,\cdot,\cdot), \{r_{f_i,h+1}(x,\cdot,\cdot)\}_{i=1}^N)$. Moreover, we define $\widetilde{V}(x) = \mathbb{E}_{a\sim\widehat{\pi}(\cdot\,|\,x), b\sim\widehat{\nu}(\cdot\,|\,x)}[\widetilde{Q}(x,a,b)]$ for any $x \in \mathcal{S}$. Then, we have

$$
\begin{aligned}
&\Big\|\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\cdot\big(V_{h+1}^k(x_{h+1}^\tau)-(\mathbb{P}_hV_{h+1}^k)(x_h^\tau,a_h^\tau,b_h^\tau)\big)\Big\|_{(\Lambda_h^k)^{-1}}\\
&\quad\le\underbrace{\Big\|\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\cdot\big(\widetilde{V}(x_{h+1}^\tau)-(\mathbb{P}_h\widetilde{V})(x_h^\tau,a_h^\tau,b_h^\tau)\big)\Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(i)}} \qquad \text{(C.42)}\\
&\qquad+\underbrace{\Big\|\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\cdot\big([V_{h+1}^k(x_{h+1}^\tau)-\widetilde{V}(x_{h+1}^\tau)]-\big(\mathbb{P}_h(V_{h+1}^k-\widetilde{V})\big)(x_h^\tau,a_h^\tau,b_h^\tau)\big)\Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(ii)}}.
\end{aligned}
$$

By Lemma H.2 and a union bound argument, it holds for any $\widetilde{Q} \in \mathcal{Q}_{h+1,\epsilon}^k$ with probability at least $1-p/2$ that

$$
|(\mathrm{i})| \le 4H^2\Big(\frac{d}{2}\log(k+1)+\log\frac{2\mathcal{N}_\epsilon}{p}\Big), \qquad \text{(C.43)}
$$

where $\mathcal{N}_\epsilon$ is the covering number of $Q_{h+1,\epsilon}$. Meanwhile, by applying Lemma H.4 with $L = H\sqrt{dk}$ and $\lambda = 1$, (C.43) gives that

$$
|(\mathrm{i})| \le C'dH\sqrt{\log(dT/p)}, \qquad \text{(C.44)}
$$

with probability at least $1-p/2$. Here $C'$ is a constant. Meanwhile, by the definition of $V_{h+1}^k$ in Line 10 of Algorithm 1, we have $V_{h+1}^k(x) = \mathbb{E}_{a\sim\widehat{\pi}(\cdot\,|\,x), b\sim\widehat{\nu}(\cdot\,|\,x)}[Q_{h+1}^k(x,a,b)]$, which yields that

$$
\begin{aligned}
|V_{h+1}^k(x)-\widetilde{V}(x)| &= \big|\mathbb{E}_{a\sim\widehat{\pi}(\cdot\,|\,x), b\sim\widehat{\nu}(\cdot\,|\,x)}[Q_{h+1}^k(x,a,b)-\widetilde{Q}(x,a,b)]\big|\\
&\le \mathbb{E}_{a\sim\widehat{\pi}(\cdot\,|\,x), b\sim\widehat{\nu}(\cdot\,|\,x)}|Q_{h+1}^k(x,a,b)-\widetilde{Q}(x,a,b)| \le \epsilon
\end{aligned}
$$

for any $x \in \mathcal{S}$, which further implies that

$$
|(\mathrm{ii})| \le \epsilon\cdot\sum_{\tau=1}^{k-1}\|\phi(x_h^\tau,a_h^\tau)\|_{(\Lambda_h^k)^{-1}} \le \epsilon k, \qquad \text{(C.45)}
$$

where the last inequality follows from the fact that $\|\phi(\cdot,\cdot)\|_{(\Lambda_h^k)^{-1}} \le \|\phi(\cdot,\cdot)\|_2 \le 1$ for any $(k,h) \in [K]\times[H]$. Plugging (C.44) and (C.45) into (C.42), together with the fact that $\epsilon = 1/KH$, we conclude the proof of Lemma C.9. □

**Lemma C.10** (Bounded Weight of Value Functions)**.** For all $(k,h) \in [K]\times[H]$, the linear coefficient $w_h^k$ defined in (3.3) satisfies $\|w_h^k\| \le H\sqrt{kd}$.

*Proof of Lemma C.10.* By the definition of $w_h^k$ in (3.3) and triangle inequality, we have

$$
\begin{aligned}
\|w_h^k\| &= \Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(x_h^\tau,a_h^\tau)\cdot V_{h+1}^k(x_{h+1}^\tau)\Big)\Big\|\\
&\le \sum_{\tau=1}^{k-1}\|(\Lambda_h^k)^{-1}\phi(x_h^\tau,a_h^\tau)\cdot V_{h+1}^k(x_{h+1}^\tau)\|. \qquad \text{(C.46)}
\end{aligned}
$$

Together with the fact that $|V_h^k(\cdot)| \le H$ for any $(k,h) \in [K] \times [H]$, (C.46) gives

$$\begin{aligned}\|w_h^k\| &\le H \cdot \sum_{\tau=1}^{k-1} \|(\Lambda_h^k)^{-1}\phi(x_h^\tau, a_h^\tau)\| \\ &\le H \cdot \sum_{\tau=1}^{k-1} \|(\Lambda_h^k)^{-1/2}\| \cdot \|\phi(x_h^\tau, a_h^\tau)\|_{(\Lambda_h^k)^{-1}} \\ &\le H \cdot \sum_{\tau=1}^{k-1} \|\phi(x_h^\tau, a_h^\tau)\|_{(\Lambda_h^k)^{-1}}, \end{aligned} \tag{C.47}$$

where the second inequality uses Cauchy-Schwarz inequality and the last inequality follows from the fact that $\Lambda_h^k \succeq I$ for any $(k,h) \in [K] \times [H]$. Then, by Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\sum_{\tau=1}^{k-1} \|\phi(x_h^\tau, a_h^\tau)\|_{(\Lambda_h^k)^{-1}} &\le \sqrt{k} \cdot \Big(\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)^\top (\Lambda_h^k)^{-1} \phi(x_h^\tau, a_h^\tau)\Big)^{1/2} \\ &= \sqrt{k} \cdot \Big(\sum_{\tau=1}^{k-1} \mathrm{Tr}\big(\phi(x_h^\tau, a_h^\tau)^\top (\Lambda_h^k)^{-1} \phi(x_h^\tau, a_h^\tau)\big)\Big)^{1/2} \\ &= \sqrt{k} \cdot \Big(\mathrm{Tr}\big((\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top\big)\Big)^{1/2}. \end{aligned} \tag{C.48}$$

Meanwhile, recall that $\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + I$, we have

$$\mathrm{Tr}\Big((\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top\Big) \le \mathrm{Tr}(I) = d. \tag{C.49}$$

Plugging (C.48) and (C.49) into (C.47), we conclude the proof of Lemma C.10. □

### C.5 Proof of Lemma C.6

*Proof of Lemma C.6.* Recall the definition of $\Gamma_h^k$ in Line 7 of Algorithm 1, we have

$$\begin{aligned}2\sum_{k=1}^{K}\sum_{h=1}^{H} \min\{H, \Gamma_h^k(x_h^k, a_h^k)\} &= 2\beta \cdot \sum_{k=1}^{K}\sum_{h=1}^{H} \min\{H/\beta, \|\phi(x_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}\} \\ &\le 2\beta \cdot \sum_{k=1}^{K}\sum_{h=1}^{H} \min\{1, \|\phi(x_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}\}. \end{aligned} \tag{C.50}$$

Here the last inequality uses the fact that $\beta = CdH\sqrt{\log(2dT/p)}$, where $C > 1$ is a constant. By Cauchy-Schwarz inequality, we further obtain that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H} \min\{1, \|\phi(x_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}\} &\le \sum_{h=1}^{H}\Big(K \cdot \sum_{k=1}^{K} \min\{1, \|\phi(x_h^k, a_h^k)\|^2_{(\Lambda_h^k)^{-1}}\}\Big) \\ &\le \sum_{h=1}^{H} \sqrt{K} \cdot \Big(2\log\Big(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\Big)\Big)^{1/2}, \end{aligned} \tag{C.51}$$

where the last inequality follows from Lemma H.1. Moreover, Assumption 2.1 gives that

$$\|\phi(x,a)\|_2 \le 1$$

for any $(k,h,x,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, which further implies that

$$\Lambda_h^{K+1} = \sum_{k=1}^{K} \phi(x_h^k, a_h^k)\phi(x_h^k, a_h^k)^\top + I \preceq (K+1) \cdot I \tag{C.52}$$

for any $h \in [H]$. Combining (C.52) and the fact that $\Lambda_h^1 = I$, we obtain

$$2\log\Big(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\Big) \le 2d\cdot\log(K+1) \le 4d\cdot\log(K). \tag{C.53}$$

Combining (C.50), (C.51), (C.52) and (C.53), it holds that

$$2\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{H, \Gamma_h^k(x_h^k, a_h^k)\} \le 2\beta\sqrt{dHT\cdot\log(K)} \le \mathcal{O}(\sqrt{d^3H^3T\iota^2}),$$

where $\iota = \log(2dT/p)$. Therefore, we conclude the proof of Lemma C.6. □

### C.6 Proof of Lemma C.7

*Proof of Lemma C.7.* First, we show that $\{\zeta_{k,h}^1, \zeta_{k,h}^2\}_{(k,h)\in[K]\times[H]}$ can be written as a bounded martingale difference with respect to a filtration. Similar to Cai et al. (2020), we construct the following filtration. For any $(k,h) \in [K]\times[H]$, we define $\sigma$-algebras $\mathcal{F}_{k,h}^1$ and $\mathcal{F}_{k,h}^2$ as follows:

$$\begin{aligned}
\mathcal{F}_{k,h}^2 &= \sigma\big(\{x_i^\tau, a_i^\tau, b_{1,i}^\tau, \cdots, b_{N,i}^\tau\}_{(\tau,i)\in[k-1]\times[h]} \cup \{x_i^k, a_i^k, b_{1,i}^k, \cdots, b_{N,i}^k\}_{i\in[h]}\big),\\
\mathcal{F}_{k,h}^2 &= \sigma\big(\{x_i^\tau, a_i^\tau, b_{1,i}^\tau, \cdots, b_{N,i}^\tau\}_{(\tau,i)\in[k-1]\times[h]} \cup \{x_i^k, a_i^k, b_{1,i}^k, \cdots, b_{N,i}^k\}_{i\in[h]} \cup \{x_{h+1}^k\}\big),
\end{aligned} \tag{C.54}$$

where $x_{H+1}$ is a null state for any $k \in [K]$. Here $\sigma(\cdot)$ denotes the $\sigma$-algebra generated by a finite set. Moreover, for any $(k,h,m) \in [K]\times[H]\times[2]$, we define the timestep index $t(k,h,m)$ as

$$t(k,h,m) = (k-1)\cdot 2H + (h-1)\cdot 2 + m. \tag{C.55}$$

By the definitions of $\sigma$-algebras in (C.54), we have $\mathcal{F}_{k,h}^m \subset \mathcal{F}_{k',h'}^{m'}$ for any $t(k,h,m) \le t(k',m',h')$, which implies that the $\sigma$-algebra sequence $\{\mathcal{F}_{k,h}^m\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a filtration. Moreover, by the definitions of $\{\zeta_{k,h}^1, \zeta_{k,h}^2\}_{(k,h)\in[K]\times[H]}$ in (C.4), we have

$$\zeta_{k,h}^1 \in \mathcal{F}_{k,h}^1, \quad \zeta_{k,h}^2 \in \mathcal{F}_{k,h}^2, \quad \mathbb{E}[\zeta_{k,h}^1 \,|\, \mathcal{F}_{k,h-1}^2] = 0, \quad \mathbb{E}[\zeta_{k,h}^2 \,|\, \mathcal{F}_{k,h}^1] = 0 \tag{C.56}$$

for any $(k,h) \in [K]\times[H]$. Here we identify $\mathcal{F}_{k,0}^2$ with $\mathcal{F}_{k-1,H}^2$ for any $k \ge 2$ and define $\mathcal{F}_{1,0,2}$ be the empty set. Hence, we can define the martingale

$$\mathcal{M}_{k,h}^m = \Big\{\sum_{k',h',m'} \zeta_{k',h'}^{m'} : t(k',h',m') \le t(k,h,m)\Big\}. \tag{C.57}$$

Such a martingale is adaptive to the filtration $\{\mathcal{F}_{k,h}^m\}_{(k,h,m)\in[K]\times[H]\times[2]}$. In particular, we have

$$\mathcal{M}_{K,H}^2 = \sum_{k=1}^{K}\sum_{h=1}^{H}(\zeta_{k,h}^1 + \zeta_{k,h}^2). \tag{C.58}$$

Moreover, note the fact that $V_h^k, Q_h^k, V_{l,h}^{\pi^k,\nu^k}, Q_{l,h}^{\pi^k,\nu^k} \in [-H, H]$, we further obtain $|\zeta_{k,h}^m| \le 2H$, for any $(k,h,m) \in [K]\times[H]\times[2]$. Finally, by applying the Azuma-Hoeffding inequality to $\mathcal{M}_{K,H}^2$ defined in (C.58), we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(\zeta_{k,h}^1 + \zeta_{k,h}^2) \le \sqrt{16KH^3\cdot\log(4/p)}$$

with probability at least $1 - p/2$, which concludes the proof of Lemma C.7. □

## D PSEUDOCODE OF REWARD-FREE EXPLORE

**Algorithm 4** Reward-Free Explore

**Input:** iteration number $K_0$ and $K$.
Let policy class $\Phi = \varnothing$.
**for** $(x, h) \in \mathcal{S} \times [H]$ **do**
  $r_{h'}(x', a') \leftarrow \mathbf{1}[x' = x \text{ and } h' = h]$ for all $(x', a', h') \in \mathcal{S} \times \mathcal{A} \times [H]$.
  $\Phi^{(x,h)} \leftarrow \text{EULER}(r, K_0)$. [2]
  $\pi_h(\cdot \,|\, x) \leftarrow \text{Uniform}(\mathcal{A})$ for all $\pi \in \Phi^{(x,h)}$.
  $\Psi \leftarrow \Psi \cup \Phi^{(x,h)}$.
**end for**
**for** $k = 1, \cdots, K$ **do**
  Sample policy $\pi \sim \text{Uniform}(\Psi)$.
  Play the game $\mathcal{M}$ using policy $\pi$ and uniform policy $\nu_{uni}$, and observe the trajectory $\{x_h^k, a_h^k, b_h^k\}_{h\in[H]}$ and rewards $\{r_{\star,h}(x_h^k, a_h^k, b_h^k)\}_{h\in[H]}$.
**end for**
Calculate the empirical reward as

$$\widehat{r}_{\star,h}(x, a, b) = \frac{\sum_{k=1}^K r_{\star,h}(x, a, b) \cdot \mathbf{1}[x_h^k = x, a_h^k = a, x_{h+1}^k = x']}{\sum_{k=1}^K \mathbf{1}[x_h^k = x, a_h^k = a, x_{h+1}^k = x']}.$$

## E UNKNOWN REWARD SETTING

To relax the assumption that the reward is known, in this subsection, we consider the case where the reward functions are unknown. We focus on the tabular case for simplicity, and the extension to linear case is left as future work. We assume that $S = |\mathcal{S}|$, $A_l = |\mathcal{A}_l|$ and $A_f = |\mathcal{A}_f| = |\mathcal{A}_{f_1} \times \cdots \times \mathcal{A}_{f_N}|$. For simplicity, we use the shorthand $V_\star^{\pi,\nu} = V_{\star,1}^{\pi,\nu}(x_1)$, where $x_1 \in \mathcal{S}$ is the fixed initial state.

At a high level, we first conduct a reward-free exploration algorithm (Algorithm 4 in §D), a variant of Reward-Free RL-Explore algorithm in Jin et al. (2020a), to obtain estimated reward functions $\{\widehat{r}_l, \widehat{r}_{f_1}, \cdots \widehat{r}_{f_N}\}$. As asserted before, we can use Algorithm 1, to find the SNE with respect to the *known* estimated reward functions $\{\widehat{r}_l, \widehat{r}_{f_1}, \cdots \widehat{r}_{f_N}\}$. Hence, we can obtain the approximate SNE if the value functions of estimated value functions are good approximation of the true value functions. Fortunately, we have the following lemma to guarantee it.

**Lemma E.1.** Fix $\varepsilon, p > 0$. If we set $K_0 \geq \Omega(H^7 S^4 A_l/\varepsilon)$ and $K \geq \Omega(H^3 S^2 A_l A_f/\varepsilon^2)$ in Algorithm 4 (cf. §D), then we have the empirical rewards $\{\widehat{r}_l, \widehat{r}_{f_1}, \cdots \widehat{r}_{f_N}\}$ and corresponding value functions $(\widehat{V}_l, \widehat{V}_{f_1}, \cdots, \widehat{V}_{f_N})$ satisfying that

$$\sup_{\pi,\nu} |\widehat{V}_\star^{\pi,\nu} - V_\star^{\pi,\nu}| \leq \varepsilon$$

with probability at least $1 - p$. Here $\Omega(\cdot)$ hides some logarithmic factors.

*Proof.* This lemma is a simple extension of Lemma D.1 in Bai et al. (2021). They focus on the MDP setting and we consider the more complex leader-controller Markov games. The detailed proof is given in §E.1. □

Lemma E.1 states that we can obtain estimated reward functions and the associated value functions is an $\varepsilon$-approximation of the true value functions, which further implies that the SNE with respect to the estimated reward functions is a good approximation of the SNE in the original problem. We also remark that if we consider the Markov games with only one follower and aim to find the Stackelberg equilibria, we can provide a more refined analysis. See §E.2 for more details.

[2] Here EULER is a single-agent RL algorithm proposed in Zanette & Brunskill (2019).

### E.1 PROOF OF LEMMA E.1

Before our proof, we present a useful lemma.

**Lemma E.2.** We define the set of $\delta$-significant states as

$$\mathcal{S}_h^\delta = \{s : \max_\pi \mathbb{P}_h^\pi(x) \geq \delta\}, \tag{E.1}$$

where $\mathbb{P}_h^\pi(x)$ is the probability of visiting $x$ at $h$-th step under policy $\pi$. Then, We have

$$\max_\pi \frac{\mathbb{P}_h^\pi(x)}{\frac{1}{K_0}\sum_{\pi\in\Phi^{(x,h)}} \mathbb{P}_h^\pi(x,a)} \leq 2$$

for any $s \in \mathcal{S}_h^\delta$. Here $\mathbb{P}_h^\pi(x,a)$ is the probability of visiting $(x,a)$ at $h$-th step under policy $\pi$.

*Proof.* See the proof of Theorem 3.3 in Jin et al. (2020a) for more details. □

Now, we are ready to proof Lemma E.1.

*Proof of Lemma E.1.* For any $(\pi,\nu)$, we denote $\mathbb{P}_h^{\pi,\nu}(x,a,b)$ as the probability of visiting $(x,a,b)$ at $h$-th step under policies $(\pi,\nu)$. Under this notion, by Lemma E.2 and the fact that all policies in $\Phi^{(x,h)}$ are uniform at $(x,h)$, we have

$$\max_{\pi,a} \frac{\mathbb{P}_h^\pi(x,a)}{\frac{1}{K_0}\sum_{\pi\in\Phi^{(x,h)}} \mathbb{P}_h^\pi(x,a)} \leq 2A_l,$$

where $|\mathcal{A}_l| = A_l$. Together with the fact that we use the uniform policy $\nu_{uni}$ in Algorithm 4 to gather data, we further obtain that

$$\max_{\pi,\nu,a,b} \frac{\mathbb{P}_h^{\pi,\nu}(x,a,b)}{\frac{1}{K_0}\sum_{\pi\in\Phi^{(x,h)}} \mathbb{P}_h^{\pi,\nu_{uni}}(x,a,b)} \leq 2A_lA_f,$$

where $A_f = |\mathcal{A}_f| = |\mathcal{A}_{f_1} \times \cdots \times \mathcal{A}_{f_N}|$. Thus, for any $\delta$-significant $(x,h)$, we have

$$\max_{\pi,\nu,a,b} \frac{\mathbb{P}_h^{\pi,\nu}(x,a,b)}{\frac{1}{K_0SH}\sum_{\pi\in\cup\{\Phi^{(x,h)}\}_{(x,h)}} \mathbb{P}_h^{\pi,\nu_{uni}}(x,a,b)} \leq 2SA_lA_fH.$$

Then the data obtained from Algorithm 4 is sampled i.i.d. from some distribution $\zeta_h$, such that

$$\max_{\pi,\nu,a,b} \frac{\mathbb{P}_h^{\pi,\nu}(x,a,b)}{\zeta_h(x,a,b)} \leq 2SA_lA_fH. \tag{E.2}$$

for any $s \in \mathcal{S}_h^\delta$. Back to our proof, we have

$$\begin{aligned}
|\widehat{V}_\star^{\pi,\nu} - V_\star^{\pi,\nu}| &= \left|\sum_{h=1}^H \sum_{x,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\left(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\right)\right| \\
&= \left|\sum_{h=1}^H \sum_{x,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\left(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\right)\right| \\
&\leq \underbrace{\left|\sum_{h=1}^H \sum_{x\notin\mathcal{S}_h^\delta,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\left(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\right)\right|}_{\text{(i)}} \\
&\quad + \underbrace{\left|\sum_{h=1}^H \sum_{x\in\mathcal{S}_h^\delta,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\left(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\right)\right|}_{\text{(ii)}}.
\end{aligned} \tag{E.3}$$

Clearly,

$$\text{(i)} \le \sum_{h=1}^{H} \sum_{x \notin \mathcal{S}_h^\delta, a, b} \mathbb{P}_h^{\pi,\nu}(s,a,b) = \sum_{h=1}^{H} \sum_{x \notin \mathcal{S}_h^\delta} \mathbb{P}_h^{\pi}(x) \le HS\delta \le \varepsilon/2, \tag{E.4}$$

where the second inequality uses the definition of $\delta$-significant set in (E.1) and the last inequality is implied by the fact that $\delta = \varepsilon/2H^2S$. Meanwhile, we have

$$\begin{aligned}\text{(ii)} &\le \sum_{h=1}^{H} \Big| \sum_{x \in \mathcal{S}_h^\delta, a, b} \mathbb{P}_h^{\pi,\nu}(x,a,b) \cdot \big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big) \Big| \\ &\le \sum_{h=1}^{H} \underbrace{\Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \mathbb{P}_h^{\pi,\nu}(x,a,b) \cdot \big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)^2 \Big)^{1/2}}_{\Delta_h}. \end{aligned} \tag{E.5}$$

Note that $\mathbb{P}_h^{\pi,\nu}(x,a,b) = \mathbb{P}_h^{\pi,\nu}(x) \cdot \pi_h(a \,|\, x) \cdot \nu_h(b \,|\, x)$, together with Cauchy-Schwarz inequality, we further have

$$\begin{aligned}\Delta_h &\le \max_{\pi': \mathcal{S} \to \mathcal{A}_l, \nu': \mathcal{S} \to \mathcal{A}_f} \Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \mathbb{P}_h^{\pi}(x) \cdot \big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)^2 \mathbf{1}[a = \pi'(s), b = \nu'(s)] \Big)^{1/2} \\ &\le \max_{\pi': \mathcal{S} \to \mathcal{A}_l, \nu': \mathcal{S} \to \mathcal{A}_f} \Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \mathbb{P}_h^{\pi}(x) \cdot \big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)^2 \mathbf{1}[a = \pi'(s), b = \nu'(s)] \Big)^{1/2} \\ &\le \max_{\pi': \mathcal{S} \to \mathcal{A}_l, \nu': \mathcal{S} \to \mathcal{A}_f} (2SA_lA_fH)^{1/2} \\ &\quad \times \Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \zeta_h(x,a,b) \cdot \big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)^2 \mathbf{1}[a = \pi'(s), b = \nu'(s)] \Big)^{1/2}, \end{aligned} \tag{E.6}$$

where the last inequality follows from (E.2). Meanwhile, by Hoeffding inequality and a union bound for the reward estimations we have

$$\begin{aligned}&\Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \zeta_h(x,a,b) \cdot \big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)^2 \mathbf{1}[a = \pi'(s), b = \nu'(s)] \Big)^{1/2} \\ &\quad \le \Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \zeta_h(x,a,b) \cdot \widetilde{\mathcal{O}}\big(\frac{1}{N_h(s,a,b)}\big) \mathbf{1}[a = \pi'(s), b = \nu'(s)] \Big)^{1/2}. \end{aligned} \tag{E.7}$$

Choose $\delta = \varepsilon/2H^2S$. Together with (E.2), we have $\zeta_h(s,a,b) \ge \varepsilon/4H^3S^2A_lA_f$ for any $s \in \mathcal{S}_h^\delta$. Hence, we have $K \ge \Omega(H^3S^2A_lA_f/\varepsilon) \ge \Omega(1/\min_{s,a,b} \zeta_h(s,a,b))$. Applying multiplicative Chernoff bound for the counter $N_h(s,a,b) \sim \mathrm{Bin}(K, \zeta_h(s,a,b))$, we have

$$\begin{aligned}&\Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \zeta_h(x,a,b) \cdot \widetilde{\mathcal{O}}\big(\frac{1}{N_h(s,a,b)}\big) \mathbf{1}[a = \pi'(s), b = \nu'(s)] \Big)^{1/2} \\ &\quad \le \Big( \sum_{x \in \mathcal{S}_h^\delta, a, b} \zeta_h(x,a,b) \cdot \widetilde{\mathcal{O}}\big(\frac{1}{K\zeta_h(s,a,b)}\big) \mathbf{1}[a = \pi'(s), b = \nu'(s)] \Big)^{1/2} \\ &\quad = \widetilde{\mathcal{O}}\Big(\sqrt{\frac{S}{K}}\Big). \end{aligned} \tag{E.8}$$

Plugging (E.6), (E.7), and (E.7) into (E.5), we have

$$\text{(ii)} \le \widetilde{\mathcal{O}}\Big(\sqrt{\frac{H^3S^2A_lA_f}{K}}\Big) \le \varepsilon/2, \tag{E.9}$$

where the last inequality follows form our choice that $K \ge \Omega(H^3S^2A_lA_f/\varepsilon^2)$. Combining (E.3), (E.4) and (E.9), we have $|\widehat{V}_\star^{\pi,\nu} - V_\star^{\pi,\nu}| \le \varepsilon$ for any $(\pi, \nu)$, which concludes the proof of Lemma E.1. □

### E.2 LEARNING STACKELBERG EQUILIBRIA

In this section, we analyze the sample-efficiency of learning Stackelberg equilibria in two-player tabular Markov games without the known reward assumption.

For simplicity, we use the shorthands $f = f_1$ and $V_\star^{\pi,\nu} = V_{\star,1}^{\pi,\nu}(x_1)$, where $x_1 \in \mathcal{S}$ is the fixed initial state. Meanwhile, for any $\varepsilon > 0$, we define the $\varepsilon$-approximate value of worst-case best response by

$$V_\varepsilon^\pi = \operatorname*{argmin}_{\nu \in \mathrm{BR}_\varepsilon(\pi)} V_l^{\pi,\nu},$$
$$\mathrm{BR}_\varepsilon(\pi) = \{\nu : V_f^{\pi,\nu} \geq \max_{\nu'} V_f^{\pi,\nu'} - \varepsilon\}.$$

By the above definitions, we can immediately obtain that $\mathrm{BR}(\pi) \subseteq \mathrm{BR}_\varepsilon(\pi)$, which further implies $V_\varepsilon^\pi \leq V_l^{\pi,\nu^*(\pi)}$. Then we can define the gap

$$\mathrm{gap}_\varepsilon = \max_\pi V_l^{\pi,\nu^*(\pi)} - \max_\pi V_\varepsilon^\pi = V_l^{\pi^*,\nu^*} - \max_\pi V_\varepsilon^\pi. \tag{E.10}$$

As stated before, we first conduct a Reward-Free Explore algorithm (Algorithm 4) to obtain the estimated rewards $(\widehat{r}_l, \widehat{r}_f)$. We also define $(\widehat{V}_l, \widehat{V}_f)$ as the corresponding value functions. Then we use Algorithm 1 to solve the SNE with respect to the *known* reward functions $(\widehat{r}_l, \widehat{r}_f)$. Specifically, we consider the following optimization problem of finding approximation Stackelberg equilibria with respect to the empirical rewards $(\widehat{r}_l, \widehat{r}_f)$.

$$\begin{aligned}
&\operatorname*{argmax}_\pi \widehat{V}_{3\varepsilon/4}(\pi) = \operatorname*{argmax}_\pi \widehat{V}_l^{\pi,\nu(\pi)}, \\
&\nu(\pi) = \operatorname*{argmin}_{\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)} \widehat{V}_l^{\pi,\nu}, \\
&\widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) = \big\{\nu : \widehat{V}_f^{\pi,\nu} \geq \max_\nu \widehat{V}_f^{\pi,\nu} - 3\varepsilon/4\big\}.
\end{aligned} \tag{E.11}$$

Since $(\widehat{r}_l, \widehat{r}_f)$ are known to us, we can use Algorithm 1 to obtain the solution $(\widehat{\pi}, \widehat{\nu} = \nu(\widehat{\pi}))$, which is our approximate solution. See Algorithm 5 for more details.

**Algorithm 5** Reward-Free Explore then Commit

1: **Input:** Accuracy coefficient $\varepsilon > 0$.
2: Run the Reward-Free Explore algorithm (Algorithm 4) with $K_0 \geq \Omega(H^7 S^4 A_l/\varepsilon)$ and $K \geq \Omega(H^3 S^2 A_l A_f/\varepsilon^2)$, and obtain empirical rewards $(\widehat{r}_l, \widehat{r}_f)$.
3: Use Algorithm 1 as an oracle to solve the problem defined in (E.11) and obtain the solution $(\widehat{\pi}, \widehat{\nu} = \nu(\pi))$.
4: **Output:** $(\widehat{\pi}, \widehat{\nu})$.

### E.3 THEORETICAL RESULTS

The performance of Algorithm 5 is guaranteed by the following theorem.

**Theorem E.3.** Suppose Algorithm 5 outputs $(\widehat{\pi}, \widehat{\nu})$. Then it holds with probability at least $1 - p$ that

$$V_l^{\widehat{\pi},\nu^*(\widehat{\pi})} \geq V_l^{\pi^*,\nu^*} - \mathrm{gap}_\varepsilon - \varepsilon, \qquad V_{f_i}^{\widehat{\pi},\widehat{\nu}} \geq V_f^{\widehat{\pi},\nu^*(\widehat{\pi})} - \varepsilon.$$

*Proof.* Similar analysis also appears in Bai et al. (2021). As stated before, however, their setting is different with ours. For completeness, we provide a detailed proof here. First, we show that

$$\mathrm{BR}_{\varepsilon/2}(\pi) \subseteq \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) \subseteq \mathrm{BR}_\varepsilon(\pi). \tag{E.12}$$

By choosing a large absolute constant in $K$, together with Lemma E.1, it holds for any $\star \in \{l, f\}$ that

$$\sup_{\pi,\nu} |\widehat{V}_\star^{\pi,\nu} - V_\star^{\pi,\nu}| \leq \varepsilon/8. \tag{E.13}$$

Meanwhile, for the empirical rewards $(\widehat{r}_l, \widehat{r}_f)$, we define the best response of leader's policy $\pi$ as $\widehat{\nu^*(\pi)}$. Under this notation, for any $\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)$, we have

$$\begin{aligned} & V_f^{\pi,\nu^*(\pi)} - V_f^{\pi,\nu} \\ &= \underbrace{(V_f^{\pi,\nu^*(\pi)} - \widehat{V}_f^{\pi,\nu^*(\pi)})}_{\text{(i)}} + \underbrace{(\widehat{V}_f^{\pi,\nu^*(\pi)} - \widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}})}_{\text{(ii)}} + \underbrace{(\widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}} - \widehat{V}_f^{\pi,\nu})}_{\text{(iii)}} + \underbrace{(\widehat{V}_{f_i}^{\pi,\nu} - V_f^{\pi,\nu})}_{\text{(iv)}} \\ &\le \varepsilon/8 + 0 + 3\varepsilon/4 + \varepsilon/8 \le \varepsilon. \end{aligned} \tag{E.14}$$

where (i) $\le \varepsilon/8$ and (iv) $\le \varepsilon/8$ is implied by the uniform convergence in (E.13), (ii) $\le 0$ uses the definition of $\widehat{\nu^*(\pi)}$, and (iii) $\le 0$ follows from the fact that $\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)$.

Similarly, for any $\nu \in \mathrm{BR}_{\varepsilon/2}(\pi)$, we can show that

$$\begin{aligned} & \widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}} - \widehat{V}_f^{\pi,\nu} \\ &= (\widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}} - V_f^{\pi,\widehat{\nu^*(\pi)}}) + (V_f^{\pi,\widehat{\nu^*(\pi)}} - V_f^{\pi,\nu^*(\pi)}) + (V_f^{\pi,\nu^*(\pi)} - V_f^{\pi,\nu}) + (V_f^{\pi,\nu} - \widehat{V}_f^{\pi,\nu}) \\ &\le \varepsilon/8 + 0 + \varepsilon/2 + \varepsilon/8 = 3\varepsilon/4. \end{aligned} \tag{E.15}$$

Combining (E.14) and (E.15), we obtain $\mathrm{BR}_{\varepsilon/2}(\pi) \subseteq \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) \subseteq \mathrm{BR}_{\varepsilon}(\pi)$ as desired.

Back to our proof, by the fact that $\widehat{\pi}$ maximizes $\widehat{V}_{3\varepsilon/4}^{\pi} = \min_{\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)} \widehat{V}_l(\pi, \nu)$, we have

$$\min_{\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})} \widehat{V}_l^{\widehat{\pi},\nu} = \widehat{V}_{3\varepsilon/4}^{\widehat{\pi}} \ge \widehat{V}_{3\varepsilon/4}^{\pi} = \min_{\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)} V_l^{\pi,\nu} \ge \min_{\nu \in \mathrm{BR}_{\varepsilon}(\pi)} \widehat{V}_l^{\pi,\nu}, \tag{E.16}$$

for any $\pi$. Here the last inequality uses the fact $\widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) \subseteq \mathrm{BR}_{\varepsilon}(\pi)$ in (E.12). Together with the uniform convergence in (E.13), (E.16) yields

$$\min_{\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})} V_l^{\widehat{\pi},\nu} \ge \min_{\nu \in \mathrm{BR}_{\varepsilon}(\pi)} V_l^{\pi,\nu} - \varepsilon/8 = V_{\varepsilon}^{\pi} - \varepsilon/8 \tag{E.17}$$

Meanwhile, by the fact $\mathrm{BR}_{\varepsilon/2}(\pi) \subseteq \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)$ in (E.13), we have

$$V_{\varepsilon/2}^{\widehat{\pi}} = \min_{\nu \in \mathrm{BR}_{\varepsilon/2}(\widehat{\pi})} V_l^{\widehat{\pi},\nu} \ge \min_{\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})} V_l^{\widehat{\pi},\nu}. \tag{E.18}$$

Combining (E.17) and (E.18), we have

$$V_{\varepsilon/2}^{\widehat{\pi}} \ge \max_{\pi} V_{\varepsilon}^{\pi} - \varepsilon/8 = \max_{\pi} V_l^{\pi,\nu^*(\pi)} - \mathrm{gap}_{\varepsilon} - \varepsilon/8 \ge V_l^{\pi^*,\nu^*} - \mathrm{gap}_{\varepsilon} - \varepsilon, \tag{E.19}$$

where the equality uses the definition of $\mathrm{gap}_{\varepsilon}$ in (E.10). Clearly, we also have $V_{\varepsilon/2}^{\widehat{\pi}} \le V_l^{\widehat{\pi},\nu^*(\widehat{\pi})}$. Plugging this inequality into (E.19), we obtain

$$V_l^{\widehat{\pi},\nu^*(\widehat{\pi})} \ge V_l^{\pi^*,\nu^*} - \mathrm{gap}_{\varepsilon} - \varepsilon$$

as desired. Meanwhile, by the facts that $\widehat{\nu} \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})$ and $\widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi}) \subseteq \mathrm{BR}_{\varepsilon}(\widehat{\pi})$, we have

$$V_{f_i}^{\widehat{\pi},\widehat{\nu}} \ge V_f^{\widehat{\pi},\nu^*(\widehat{\pi})} - \varepsilon.$$

Therefore, we conclude the proof of Theorem E.3. □

## F Proof of Theorem 4.2

To facilitate our analysis, we first define the prediction error

$$\delta_h = r_{l,h} + \widehat{Q}_h - \mathbb{P}_h \widehat{V}_h \tag{F.1}$$

for any $h \in [H]$. Then we show the proof of Theorem 4.2.

*Proof of Theorem 4.2.* Recall that the definition of optimality gap defined in (4.1) takes the following form

$$\text{SubOpt}(\widehat{\pi}, \widehat{\nu}, x) = V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\nu^*(\widehat{\pi})}(x) + \sum_{i=1}^{N}[V_{f_i,1}^{\widehat{\pi},\nu^*(\widehat{\pi})}(x) - V_{f_i,1}^{\widehat{\pi},\widehat{\nu}}(x)]. \tag{F.2}$$

Similar to Lemma C.1, we have the following lemma.

**Lemma F.1.** It holds that $\widehat{\nu} = \nu^*(\widehat{\pi})$. Here $\nu(\cdot)$ is defined in (2.6).

*Proof.* This proof is similar to the proof of Lemma C.1, and we omit it to avoid repetition. □

By Lemma F.1, we have $\nu^*(\widehat{\pi}) = \widehat{\nu}$, which implies that the suboptimality of followers decays to zero. Then we only need to characterize the quantity $V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x)$, which can be decomposed by the following lemma.

**Lemma F.2.** For the $\widehat{V}_1$ defined in Line 9 of Algorithm 3 and any $(\pi, \nu)$, it holds that

$$V_{l,1}^{\pi,\nu}(x) - \widehat{V}_1(x) = \mathbb{E}_{\pi,\nu}\bigg[\sum_{h=1}^{H}\langle\widehat{Q}_h(x_h,\cdot,\cdot), \pi_h(\cdot\,|\,x_h)\times\nu_h(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h)\times\widehat{\nu}_h(\cdot\,|\,x_h)\rangle\bigg] + \mathbb{E}_{\pi,\nu}\bigg[\sum_{h=1}^{H}\delta_h(x_h,a_h,b_h)\bigg].$$

*Proof.* This proof is the same as the proof of (C.20), and we omit it to avoid repetition. □

Applying Lemma F.2 with $(\pi, \nu) = (\pi^*, \nu^*)$, we have

$$V_{l,1}^{\pi^*,\nu^*}(x) - \widehat{V}_1(x) = \mathbb{E}_{\pi^*,\nu^*}\bigg[\sum_{h=1}^{H}\langle\widehat{Q}_h(x_h,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h)\times\nu_h^*(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h)\times\widehat{\nu}_h(\cdot\,|\,x_h)\rangle\bigg] + \mathbb{E}_{\pi^*,\nu^*}\bigg[\sum_{h=1}^{H}\delta_h(x_h,a_h,b_h)\bigg]. \tag{F.3}$$

Similarly, applying Lemma F.2 with $(\pi, \nu) = (\widehat{\pi}, \widehat{\nu})$ gives that

$$\widehat{V}_1(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x) = -\mathbb{E}_{\widehat{\pi},\widehat{\nu}}\bigg[\sum_{h=1}^{H}\delta_h(x_h,a_h,b_h)\bigg]. \tag{F.4}$$

Combining (F.3) and (F.4), we obtain

$$V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x) = \mathbb{E}_{\pi^*,\nu^*}\bigg[\sum_{h=1}^{H}\langle\widehat{Q}_h(x_h,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h)\times\nu_h^*(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h)\times\widehat{\nu}_h(\cdot\,|\,x_h)\rangle\bigg] + \mathbb{E}_{\pi^*,\nu^*}\bigg[\sum_{h=1}^{H}\delta_h(x_h,a_h,b_h)\bigg] - \mathbb{E}_{\widehat{\pi},\widehat{\nu}}\bigg[\sum_{h=1}^{H}\delta_h(x_h,a_h,b_h)\bigg]. \tag{F.5}$$

As stated in §C, these two terms characterize the optimization error and the statistical error, respectively. Similar to Lemmas C.4 and C.5, we introduce the following two lemmas to analyze these two errors.

**Lemma F.3.** It holds that

$$\mathbb{E}_{\pi^*,\nu^*}\bigg[\sum_{h=1}^{H}\langle\widehat{Q}_h(x_h,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h)\times\nu_h^*(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h)\times\widehat{\nu}_h(\cdot\,|\,x_h)\rangle\bigg] \le \epsilon H.$$

*Proof.* This proof is similar to the proof of Lemma C.4, and we omit it to avoid repetition. □

**Lemma F.4.** It holds with probability at least $1 - p/2$ that

$$0 \le \delta_h(x, a, b) \le 2\Gamma_h(x, a)$$

for any $h \in [H]$ and $(x, a, b) \in \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$.

*Proof.* See §F.1 for a detailed proof. □

Combining (F.5) and Lemmas F.3 and F.4, we further obtain that

$$\begin{aligned} V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x) &\le \epsilon H + 2\mathbb{E}_{\pi^*}\Big[\sum_{h=1}^{H} \Gamma_h(x_h, a_h)\Big] \\ &\le 3\beta' \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\big[\big(\phi(s_h, a_h)^\top (\Lambda_h)^{-1} \phi(s_h, a_h)\big)^{1/2}\big], \end{aligned} \tag{F.6}$$

where the last inequality is obtained by the definition of $\Gamma_h$ in Line 6 of Algorithm 3 and the fact that $\epsilon = d/KH$. Therefore, we conclude the proof of Theorem 4.2. □

### F.1 Proof of Lemma F.4

*Proof of Lemma F.4.* Similar to (C.36), it holds with probability at least $1 - p/2$ that

$$|\phi(x, a)^\top w_h - (\mathbb{P}_h \widehat{V}_{h+1})(x, a, b)| \le \Gamma_h(x, a) \tag{F.7}$$

for any $h \in [H]$. The only exception is that we use Lemma H.3 instead of the classical concentration lemma (Lemma H.2) for the self-normalized process. Here we omit the detailed proof to avoid repetition.

By (F.7) and the fact that $\widehat{V}_{h+1}(\cdot) \le H - h$, we obtain

$$\phi(x, a)^\top w_h - \Gamma_h(x, a) \le (\mathbb{P}_h \widehat{V}_{h+1})(x, a, b) \le H - h. \tag{F.8}$$

Thus, we have $\widehat{Q}_h \ge \phi^\top w_h - \Gamma_h$, which further implies that

$$\begin{aligned} \delta_h(x, a, b) &= r_{l,h}(x, a, b) + \mathbb{P}_h \widehat{V}_{h+1}(x, a, b) - \widehat{Q}_h(x, a, b) \\ &\le \mathbb{P}_h \widehat{V}_{h+1}(x, a, b) - \phi(x, a)^\top w_h + \Gamma_h(x, a) \\ &\le 2\Gamma_h(x, a), \end{aligned} \tag{F.9}$$

where the last inequality uses (F.7). Meanwhile, it holds that

$$\begin{aligned} \delta_h(x, a, b) &= r_{l,h}(x, a, b) + \mathbb{P}_h \widehat{V}_{h+1}(x, a, b) - \widehat{Q}_h(x, a, b) \\ &\ge \mathbb{P}_h \widehat{V}_{h+1}(x, a, b) - \max\{\phi(x, a)^\top w_h - \Gamma_h^k(x, a), -(H - h)\} \\ &= \min\{\mathbb{P}_h V_{h+1}^k(x, a, b) - \phi(x, a)^\top w_h^k + \Gamma_h^k(x, a), \mathbb{P}_h V_{h+1}^k(x, a, b) + (H - h)\} \\ &\ge 0, \end{aligned} \tag{F.10}$$

where the last inequality follows from (F.7). Combining (F.9) and (F.10), we conclude the proof of Lemma F.4. □

## G Proof of Corollary 4.3

*Proof of Corollary 4.3.* The proof is similar to the proof of Corollary 4.5 in Jin et al. (2020c). For completeness, we present the detailed proof here. For notational simplicity, we define

$$\Sigma_h(x) = \mathbb{E}_{\pi^*,x}[\phi(s_h, a_h)\phi(s_h, a_h)^\top]$$

for all $x \in \mathcal{S}$ and $h \in [H]$. With this notation and Cauchy-Schwarz inequality, we have

$$\begin{aligned} \mathbb{E}_{\pi^*,x}\big[\sqrt{\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)}\big] &= \mathbb{E}_{\pi^*,x}\big[\sqrt{\mathrm{Tr}\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)}\big] \\ &= \mathbb{E}_{\pi^*,x}\big[\sqrt{\mathrm{Tr}\big(\phi(s_h, a_h)\phi(s_h, a_h)^\top \Lambda_h^{-1}\big)}\big] \\ &= \mathbb{E}_{\pi^*,x}\big[\sqrt{\mathrm{Tr}\big(\Sigma_h(x)\Lambda_h^{-1}\big)}\big]. \end{aligned} \tag{G.1}$$

Plugging (G.1) into Theorem 4.2, together with the assumption that $\Lambda_h \succeq I + c \cdot K \cdot \mathbb{E}_{\pi^*,x}[\phi(s_h,a_h)\phi(s_h,a_h)^\top]$ with probability at least $1-p/2$ and a union bound argument, we further with probability at least $1-p$ have

$$\begin{aligned}\text{SubOpt}(\widehat{\pi},\widehat{\nu},x) &\le 3\beta' \sum_{h=1}^{H} \mathbb{E}_{\pi^*,x}\Big[\sqrt{\text{Tr}\Big(\Sigma_h(x)\big(I + c\cdot K\cdot \Sigma_h(x)\big)^{-1}\Big)}\Big] \\ &= 3\beta' \sum_{h=1}^{H} \sqrt{\sum_{j=1}^{d} \frac{\lambda_{h,j}(x)}{1+cK\lambda_{h,j}(x)}} \end{aligned} \tag{G.2}$$

for all $x \in \mathcal{S}$. Here $\{\lambda_{h,j}(x)\}_{j=1}^d$ are the eigenvalues of $\Sigma_h(x)$. Meanwhile, by Jensen's inequality, we obtain

$$\|\Sigma_h(x)\|_{\text{op}} \le \mathbb{E}_{\pi^*,x}[\|\phi(s_h,a_h)\phi(s_h,a_h)^\top\|_{\text{op}}] \le 1, \tag{G.3}$$

where the last inequality follows from the fact that $\|\phi(\cdot,\cdot)\|_2 \le 1$. Combining (G.2) and (G.3), it holds with probability at least $1-p$ that

$$\begin{aligned}\text{SubOpt}(\widehat{\pi},\widehat{\nu},x) &\le 3\beta' \sum_{h=1}^{H} \sqrt{\sum_{j=1}^{d} \frac{1}{1+cK}} \\ &\le \bar{C}\cdot d^{3/2}H^2\sqrt{\log(4dHK/p)/K},\end{aligned}$$

where $\bar{C} = 3C/\sqrt{c}$, which concludes the proof of Corollary 4.3. □

## H Supporting Lemmas

**Lemma H.1** (Elliptical Potential Lemma (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Jin et al., 2020b; Cai et al., 2020))**.** Let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued sequence. Meanwhile, let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive-definite matrix and $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t-1}\phi_j\phi_j^\top$. It holds for any $t \in \mathbb{Z}_+$ that

$$\sum_{j=1}^{t} \min\{1, \|\phi_j\|^2_{\Lambda_j^{-1}}\} \le 2\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg).$$

*Proof.* See Lemma 11 of Abbasi-Yadkori et al. (2011) for a detailed proof. □

**Lemma H.2** (Concentration of Self-Normalized Process (Abbasi-Yadkori et al., 2011))**.** Let $\{\widetilde{\mathcal{F}}_t\}_{t=0}^\infty$ be a filtration and $\{\eta_t\}_{t=1}^\infty$ be an $\mathbb{R}$-valued stochastic process such that $\eta_t$ is $\widetilde{\mathcal{F}}_t$-measurable for any $t \ge 0$. We also assume that, for any $t \ge 0$, conditioning on $\widetilde{\mathcal{F}}_t$, $\eta_t$ is a zero-mean and $\sigma$-sub-Gaussian random variable, that is,

$$\mathbb{E}[\eta_t \,|\, \widetilde{\mathcal{F}}_t] = 0, \qquad \mathbb{E}[e^{\lambda\eta_t} \,|\, \widetilde{\mathcal{F}}_t] \le e^{\lambda^2\sigma^2/2} \tag{H.1}$$

for any $\lambda \in \mathbb{R}$. Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that $X_t$ is $\widetilde{\mathcal{F}}_t$-measurable for any $t \ge 0$. Also, let $Y \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix. For any $t \ge 0$, we define

$$\overline{Y}_t = Y + \sum_{s=1}^{t} X_s X_s^\top, \quad S_t = \sum_{s=1}^{t} \eta_s \cdot X_s.$$

For any $\delta > 0$ and $t \ge 0$., it holds with probability at least $1-\delta$ that

$$\|S_t\|^2_{\overline{Y}_t^{-1}} \le 2\sigma^2 \cdot \log\bigg(\frac{\det(\overline{Y}_t)^{1/2}\det(Y)^{-1/2}}{\delta}\bigg).$$

*Proof.* See Theorem 1 of Abbasi-Yadkori et al. (2011) for a detailed proof. □

**Lemma H.3.** For any fixed $h \in [H]$, let $V : \mathcal{S} \to [0, H]$ be any fixed value function. Under Assumption 4.1, for any fixed $\delta > 0$, we have

$$P_{\mathcal{D}}\bigg(\Big\|\sum_{k=1}^{K}\phi(x_h^\tau, a_h^\tau)\cdot\big(V(x_{h+1}^\tau) - \mathbb{P}_h V(x_h^\tau, a_h^\tau, b_h^\tau)\big)\Big\|_{\Lambda_h^{-1}} > H^2\cdot\big(2\log(1/\delta) + d\cdot\log(1+K)\big)\bigg) \le \delta.$$

*Proof.* See Lemma B.2 of Jin et al. (2020c) for a detailed proof. □

**Lemma H.4** (Covering)**.** Let $\mathcal{Q}_h$ be the class of value functions $Q : \mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f \to \mathbb{R}$ that takes the form

$$Q(\cdot,\cdot,\cdot) = r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{(\phi(\cdot,\cdot)^\top w + \beta\cdot\big(\phi(\cdot,\cdot)^\top\Lambda^{-1}\phi(\cdot,\cdot)\big)^{1/2}\},$$

which are parameterized by $(w, \Lambda) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ such that $\|w\| \le L$ and $\lambda_{\min}(\Lambda) \ge \lambda$. We assume that $\beta$ is fixed and satisfy that $\beta \in [0, B]$, and the feature map $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ satisfies that $\|\phi(\cdot,\cdot)\|_2 \le 1$. We have that, for any $L, B, \epsilon > 0$, there exists an $\epsilon$-covering of $\mathcal{Q}_h$ with respect to the $\ell_\infty$ norm such that the covering number $\mathcal{N}_\epsilon$ satisfies

$$\log\mathcal{N}_\epsilon \le d\cdot\log(1 + 4L/\epsilon) + d^2\cdot\log\big(1 + 8B^2\sqrt{d}/(\epsilon^2\lambda)\big).$$

*Proof.* See Jin et al. (2020b) for a detailed proof. □