A. Multiplayer General-sum Markov Games

In this section, we extend both our model-based algorithms (Algorithm 1 and Algorithm 2) to the setting of multiplayer general-sum Markov games, and present corresponding theoretical guarantees.

A.1. Problem formulation

A general-sum Markov game (general-sum MG) with m players is a tuple $\mathrm{MG}(H, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, \mathbb{P}, \{r_i\}_{i=1}^m)$, where $H, \mathcal{S}$ denote the length of each episode and the state space. Different from the two-player zero-sum setting, we now have m different action spaces, where $\mathcal{A}_i$ is the action space for the i^{th} player and $|\mathcal{A}_i| = A_i$. We let $\boldsymbol{a} := (a_1, \cdots, a_m)$ denote the (tuple of) joint actions by all m players. $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ is a collection of transition matrices, so that $\mathbb{P}_h(\cdot|s, \boldsymbol{a})$ gives the distribution of the next state if actions $\boldsymbol{a}$ are taken at state s at step h, and $r_i = \{r_{h,i}\}_{h\in[H]}$ is a collection of reward functions for the i^{th} player, so that $r_{h,i}(s, \boldsymbol{a})$ gives the reward received by the i^{th} player if actions $\boldsymbol{a}$ are taken at state s at step h.

In this section, we consider three versions of equlibrium for general-sum MGs: Nash equilibrium (NE), correlated equilibrium (CE), and coarse correlated equilibrium (CCE), all being standard solution notions in games (Nisan et al., 2007). These three notions coincide on two-player zero-sum games, but are not equivalent to each other on multi-player general-sum games; any one of them could be desired depending on the application at hand. Below we introduce their definitions.

(Approximate) Nash equilibrium in general-sum MGs. The policy of the i^{th} player is denoted as $\pi_i := \big\{\pi_{h,i} : \mathcal{S} \to \Delta_{\mathcal{A}_i}\big\}_{h\in[H]}$. We denote the product policy of all the players as $\pi := \pi_1 \times \cdots \times \pi_M$, and denote the policy of all the players except the i^{th} player as π_{-i}. We define $V_{h,i}^{\pi}(s)$ as the expected cumulative reward that will be received by the i^{th} player if starting at state s at step h and all players follow policy π. For any strategy π_{-i}, there also exists a *best response* of the i^{th} player, which is a policy $\mu^\dagger(\pi_{-i})$ satisfying $V_{h,i}^{\mu^\dagger(\pi_{-i}),\pi_{-i}}(s) = \sup_{\pi_i} V_{h,i}^{\pi_i,\pi_{-i}}(s)$ for any $(s,h) \in \mathcal{S}\times[H]$. We denote $V_{h,i}^{\dagger,\pi_{-i}} := V_{h,i}^{\mu^\dagger(\pi_{-i}),\pi_{-i}}$. The Q-functions of the best response can be defined similarly.

Our first objective is to find an approximate Nash equilibrium of Markov games.

Definition 7 (ϵ-approximate Nash equilibrium in general-sum MGs)**.** A product policy π is an **ϵ-approximate Nash equilibrium** if $\max_{i\in[m]} (V_{1,i}^{\dagger,\pi_{-i}} - V_{1,i}^{\pi})(s_1) \le \epsilon$.

The above definition requires the suboptimality gap $(V_{1,i}^{\dagger,\pi_{-i}} - V_{1,i}^{\pi})(s_1)$ to be less than ϵ for all player i. This is consistent with the two-player case (Definition 1) up to a constant of 2, since in the two-player zero-sum setting, we have $V_{1,1}^{\pi}(s_1) = -V_{1,2}^{\pi}(s_1)$ for any product policy $\pi = (\mu, \nu)$, and therefore $(V_{1,1}^{\dagger,\nu} - V_{1,1}^{\mu,\dagger})(s_1) \le 2\max_{i\in[2]} (V_{1,i}^{\dagger,\pi_{-i}} - V_{1,i}^{\pi})(s_1) \le 2(V_{1,1}^{\dagger,\nu} - V_{1,1}^{\mu,\dagger})(s_1)$.We can similarly define the regret.

Definition 8 (Nash-regret in general-sum MGs)**.** Let π^k denote the (product) policy deployed by the algorithm in the k^{th} episode. After a total of K episodes, the regret is defined as

$$\mathrm{Regret}_{\mathsf{Nash}}(K) = \sum_{k=1}^{K} \max_{i\in[m]} (V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k})(s_1).$$

(Approximate) CCE in general-sum MGs. The coarse correlated equilibrium (CCE) is a relaxed version of Nash equilibrium in which we consider general correlated policies instead of product policies. Let $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ denote the joint action space.

Definition 9 (CCE in general-sum MGs)**.** A (correlated) policy $\pi := \{\pi_h(s) \in \Delta_{\mathcal{A}} : (h,s) \in [H]\times\mathcal{S}\}$ is a **CCE** if $\max_{i\in[m]} V_{h,i}^{\dagger,\pi_{-i}}(s) \le V_{h,i}^{\pi}(s)$ for all $(s,h) \in \mathcal{S}\times[H]$.

Compared with a Nash equilibrium, a CEE is not necessarily a product policy, that is, we may not have $\pi_h(s) \in \Delta_{\mathcal{A}_1} \times \cdots \times \Delta_{\mathcal{A}_m}$. Similarly, we also define ϵ-approximate CCE and CCE-regret below.

Definition 10 (ϵ-approximate CCE in general-sum MGs)**.** A policy $\pi := \{\pi_h(s) \in \Delta_{\mathcal{A}} : (h,s) \in [H]\times\mathcal{S}\}$ is an **ϵ-approximate CCE** if $\max_{i\in[m]} (V_{1,i}^{\dagger,\pi_{-i}} - V_{1,i}^{\pi})(s_1) \le \epsilon$.

Definition 11 (CCE-regret in general-sum MGs)**.** Let policy π^k denote the (correlated) policy deployed by the algorithm in the k^{th} episode. After a total of K episodes, the regret is defined as

$$\mathrm{Regret}_{\mathsf{CCE}}(K) = \sum_{k=1}^{K} \max_{i\in[m]} (V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k})(s_1).$$

(Approximate) CE in general-sum MGs. The correlated equilibrium (CE) is another relaxation of the Nash equilibrium. To define CE, we first introduce the concept of strategy modification: A strategy modification $\phi := \{\phi_{h,s}\}_{(h,s)\in[H]\times\mathcal{S}}$ for player i is a set of $S \times H$ functions from $\mathcal{A}_i$ to itself. Let Φ_i denote the set of all possible strategy modifications for player i.

One can compose a strategy modification ϕ with any Markov policy π and obtain a new policy $\phi \diamond \pi$ such that when policy π chooses to play $\boldsymbol{a} := (a_1, \ldots, a_m)$ at state s and step h, policy $\phi \diamond \pi$will play $(a_1, \ldots, a_{i-1}, \phi_{h,s}(a_i), a_{i+1}, \ldots, a_m)$ instead.

Definition 12 (CE in general-sum MGs)**.** A policy $\pi := \{\pi_h(s) \in \Delta_{\mathcal{A}} : (h,s) \in [H] \times \mathcal{S}\}$ is a **CE** if $\max_{i\in[m]} \max_{\phi\in\Phi_i} V_{h,i}^{\phi\diamond\pi}(s) \le V_{h,i}^{\pi}(s)$ holds for all $(s,h) \in \mathcal{S} \times [H]$.

Similarly, we have an approximate version of CE and CE-regret.

Definition 13 (ϵ-approximate CE in Markov games)**.** A policy $\pi := \{\pi_h(s) \in \Delta_{\mathcal{A}} : (h,s) \in [H]\times\mathcal{S}\}$ is an **ϵ-approximate CE** if $\max_{i\in[m]} \max_{\phi\in\Phi_i} (V_{1,i}^{\phi\diamond\pi} - V_{1,i}^{\pi})(s_1) \le \epsilon$.

Definition 14 (CE-regret in multiplayer Markov games)**.** Let policy π^k denote the policy deployed by the algorithm in the k^{th} episode. After a total of K episodes, the regret is defined as

$$\mathrm{Regret}_{\mathsf{CE}}(K) = \sum_{k=1}^{K} \max_{i\in[m]} \max_{\phi\in\Phi_i} (V_{1,i}^{\phi\diamond\pi^k} - V_{1,i}^{\pi^k})(s_1).$$

Relationship between Nash, CE, and CCE For general-sum MGs, we have $\{\text{Nash}\} \subseteq \{\text{CE}\} \subseteq \{\text{CCE}\}$, so that they form a nested set of notions of equilibria (Nisan et al., 2007). Indeed, one can easily verify that if we restrict the choice of strategy modification ϕ to those consisting of only constant functions, i.e., $\phi_{h,s}(a)$ being independent of a, Definition 12 will reduce to the definition of CCE policy. In addition, any Nash equilibrium is a CE by definition. Finally, since a Nash equilibrium always exists, so does CE and CCE.

A.2. Multiplayer optimistic Nash value iteration

Here we present the Multi-Nash-VI algorithm, which is an extension of Algorithm 1 for multi-player general-sum Markov games.

The Equilibrium Subroutine. Our Equilibrium subroutine in Line 11 could be taken from either one of the $\{\textsc{Nash}, \text{CE}, \text{CCE}\}$ subroutines for *one-step* games. When using Nash, we compute the Nash equilibrium of a one-step multi-player game (see, e.g., Berg & Sandholm (2016) for an overview of the available algorithms); the worst-case computational complexity of such a subroutine will be PPAD-hard (Daskalakis, 2013). When using CE or CCE, we find CEs or CCEs of the one-step games respectively, which can be solved in polynomial time using linear programming. However, the policies found are not guaranteed to be a product policy. We remark that in Algorithm 1 we used the CCE subroutine for finding Nash in two-player zero-sum games, which seemingly contrasts the principle of using the right subroutine for finding the right equilibrium, but nevertheless works as the Nash equilibrium and CCE are equivalent in zero-sum games.

Now we are ready to present the theoretical guarantees for Algorithm 3. We let π^k denote the policy computed in line 11 of Algorithm 3 in the k^{th} episode.

Theorem 15 (Multi-Nash-VI)**.** *There exists an absolute constant c, for any $p \in (0,1]$, let $\iota = \log(SABT/p)$, then with probability at least $1-p$, Algorithm 3 with bonus $\beta_t = c\sqrt{SH^2\iota/t}$ and* Equilibrium *being one of* $\{\textsc{Nash}, \text{CE}, \text{CCE}\}$ *satisfies (repsectively):*

- π^{out} *is an ϵ-approximate* $\{\textsc{Nash},\text{CE},\text{CCE}\}$*, if the number of episodes $K \ge \Omega(H^4S^2(\prod_{i=1}^m A_i)\iota/\epsilon^2)$.*

Algorithm 3 Multiplayer Optimistic Nash Value Iteration (Multi-Nash-VI)

```
Initialize: for any $(s, \boldsymbol{a}, h, i)$, $\overline{Q}_{h,i}(s, \boldsymbol{a}) \leftarrow H$, $\underline{Q}_{h,i}(s, \boldsymbol{a}) \leftarrow 0$, $\Delta \leftarrow H$, $N_h(s, \boldsymbol{a}) \leftarrow 0$.
for episode $k = 1, \dots, K$ do
    for step $h = H, H-1, \dots, 1$ do
        for $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ do
            $t \leftarrow N_h(s, \boldsymbol{a})$;
            if $t > 0$ then
                for player $i = 1, 2, \dots, m$ do
                    $\overline{Q}_{h,i}(s, \boldsymbol{a}) \leftarrow \min\{(r_{h,i} + \widehat{\mathbb{P}}_h \overline{V}_{h+1,i})(s, \boldsymbol{a}) + \beta_t, H\}$.
                    $\underline{Q}_{h,i}(s, \boldsymbol{a}) \leftarrow \max\{(r_{h,i} + \widehat{\mathbb{P}}_h \underline{V}_{h+1,i})(s, \boldsymbol{a}) - \beta_t, 0\}$.
        for $s \in \mathcal{S}$ do
            $\pi_h(\cdot|s) \leftarrow$ EQUILIBRIUM$(\overline{Q}_{h,1}(s, \cdot), \overline{Q}_{h,2}(s, \cdot), \cdots, \overline{Q}_{h,M}(s, \cdot))$.
            for player $i = 1, 2, \dots, m$ do
                $\overline{V}_{h,i}(s) \leftarrow (\mathbb{D}_{\pi_h} \overline{Q}_{h,i})(s)$;   $\underline{V}_{h,i}(s) \leftarrow (\mathbb{D}_{\pi_h} \underline{Q}_{h,i})(s)$.
    if $\max_{i \in [m]} (\overline{V}_{1,i} - \underline{V}_{1,i})(s_1) < \Delta$ then
        $\Delta \leftarrow \max_{i \in [m]} (\overline{V}_{1,i} - \underline{V}_{1,i})(s_1)$ and $\pi^{\text{out}} \leftarrow \pi$.
    for step $h = 1, \dots, H$ do
        take action $\boldsymbol{a}_h \sim \pi_h(\cdot|s_h)$, observe reward $r_h$ and next state $s_{h+1}$.
        add 1 to $N_h(s_h, \boldsymbol{a}_h)$ and $N_h(s_h, \boldsymbol{a}_h, s_{h+1})$.
        $\widehat{\mathbb{P}}_h(\cdot|s_h, \boldsymbol{a}_h) \leftarrow N_h(s_h, \boldsymbol{a}_h, \cdot)/N_h(s_h, \boldsymbol{a}_h)$.
Output $\pi^{\text{out}}$.
```

- $\text{Regret}_{\{\mathsf{Nash,CE,CCE}\}}(K) \le \mathcal{O}(\sqrt{H^3 S^2 (\prod_{i=1}^m A_i) T\iota})$.

In the situation where the EQUILIBRIUM subroutine is taken as NASH, Theorem 15 provides the sample complexity bound of Multi-Nash-VI algorithm to find an ϵ-approximate Nash equilibrium and its regret bound. Compared with our earlier result in two-player zero-sum games (Theorem 3), here the sample complexity scales as S^2H^4 instead of SH^3. This is because the auxiliary bonus and Bernstein concentration technique do not apply here. Furthermore, the sample complexity is proportional to $\prod_{i=1}^m A_i$, which increases exponentially as the number of players increases.

Runtime of Algorithm 3 We remark that while the Nash guarantee is the strongest among the three guarantees presented in Theorem 15, the runtime of Algorithm 3 in the Nash case is not guaranteed to be polynomial and in the worst case PPAD-hard (due to the hardness of the NASH subroutine). In contrast, the CE and CCE guarantees are weaker, but the corresponding algorithms are guaranteed to finish in polynomial time.

A.3. Multiplayer reward-free learning

We can also generalize VI-Zero to the multiplayer setting and obtain Algorithm 4, Multi-VI-Zero, which is almost the same as VI-Zero except that its exploration bonus β_t is larger than that of VI-Zero by a $\sqrt{S}$ factor.

Similar to Theorem 5, we have the following theoretical guarantee claiming that any $\{$NASH,CCE,CE$\}$ of the $\mathcal{M}(\widehat{\mathbb{P}}, \widehat{r}^i)$ $(i \in [N])$ is also an approximate $\{$NASH,CCE,CE$\}$ of the true Markov game $\mathcal{M}(\mathbb{P}, r^i)$, where $\widehat{\mathbb{P}}^{\text{out}}$ is the empirical transition outputted by Algorithm 4 and $\widehat{r}^i$ is the empirical estimate of r^i.

Theorem 16 (Multi-VI-Zero)**.** *There exists an absolute constant c, for any $p \in (0, 1]$, $\epsilon \in (0, H]$, $N \in \mathbb{N}$, if we choose bonus $\beta_t = c\sqrt{H^2 S\iota/t}$ with $\iota = \log(NSABT/p)$ and $K \ge c(H^4 S^2 (\prod_{i=1}^m A_i)\iota/\epsilon^2)$, then with probability at least $1 - p$, the output $\widehat{\mathbb{P}}^{out}$ of Algorithm 4 has the following property: for any N fixed reward functions $r^1, \dots, r^N$, any $\{$NASH,CCE,CE$\}$ of Markov game $\mathcal{M}(\widehat{\mathbb{P}}^{out}, \widehat{r}^i)$ is also an ϵ-approximate $\{$NASH,CCE,CE$\}$ of the true Markov game $\mathcal{M}(\mathbb{P}, r^i)$ for all $i \in [N]$.*

The proof of Theorem 16 can be found in Appendix F.2. It is worth mentioning that the empirical Markov game $\mathcal{M}(\widehat{\mathbb{P}}^{\text{out}}, \widehat{r}^i)$ may have multiple $\{$Nash equilibria,CCEs,CEs$\}$ and Theorem 16 ensures that all of them are ϵ-approximate $\{$Nash equilibria,CCEs,CEs$\}$ of the true Markov game. Also, note that the sample complexity here is quadratic in the number of states because we are using the exploration bonus $\beta_t = \sqrt{H^2 S\iota/t}$ that is larger than usual by a $\sqrt{S}$ factor.

Algorithm 4 Multiplayer Optimistic Value Iteration with Zero Reward (Multi-VI-Zero)

```
Initialize: for any $(s, \boldsymbol{a}, h)$, $\widetilde{V}_h(s, \boldsymbol{a}) \leftarrow H$, $\Delta \leftarrow H$, $N_h(s, \boldsymbol{a}) \leftarrow 0$.
for episode $k = 1, \ldots, K$ do
    for step $h = H, H-1, \ldots, 1$ do
        for $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ do
            $t \leftarrow N_h(s, \boldsymbol{a})$.
            if $t > 0$ then
                $\widetilde{Q}_h(s, \boldsymbol{a}) \leftarrow \min\{(\widehat{\mathbb{P}}_h \widetilde{V}_{h+1})(s, \boldsymbol{a}) + \beta_t, H\}$.
        for $s \in \mathcal{S}$ do
            $\pi_h(s) \leftarrow \arg\max_{\boldsymbol{a} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_m} \widetilde{Q}_h(s, \boldsymbol{a})$.
            $\widetilde{V}_h(s) \leftarrow (\mathbb{D}_{\pi_h} \widetilde{Q}_h)(s)$.
    if $\widetilde{V}_1(s_1) < \Delta$ then
        $\Delta \leftarrow \widetilde{V}_1(s_1)$ and $\widehat{\mathbb{P}}^{\mathrm{out}} \leftarrow \widehat{\mathbb{P}}$.
    for step $h = 1, \ldots, H$ do
        take action $\boldsymbol{a}_h \sim \pi_h(\cdot, \cdot | s_h)$, observe next state $s_{h+1}$.
        add 1 to $N_h(s_h, \boldsymbol{a}_h)$ and $N_h(s_h, \boldsymbol{a}_h, s_{h+1})$.
        $\widehat{\mathbb{P}}_h(\cdot | s_h, \boldsymbol{a}_h) \leftarrow N_h(s_h, \boldsymbol{a}_h, \cdot) / N_h(s_h, \boldsymbol{a}_h)$.
Output $\widehat{\mathbb{P}}^{\mathrm{out}}$.
```

B. Bellman Equations for Markov Games

In this section, we present the Bellman equations for different types of values in Markov games.

Fixed policies. For any pair of Markov policy (μ, ν), by definition of their values in (1) (2), we have the following Bellman equations:

$$Q_h^{\mu,\nu}(s,a,b) = (r_h + \mathbb{P}_h V_{h+1}^{\mu,\nu})(s,a,b), \qquad V_h^{\mu,\nu}(s) = (\mathbb{D}_{\mu_h \times \nu_h} Q_h^{\mu,\nu})(s)$$

for all $(s,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$, where $V_{H+1}^{\mu,\nu}(s) = 0$ for all $s \in \mathcal{S}$.

Best responses. For any Markov policy μ of the max-player, by definition, we have the following Bellman equations for values of its best response:

$$Q_h^{\mu,\dagger}(s,a,b) = (r_h + \mathbb{P}_h V_{h+1}^{\mu,\dagger})(s,a,b), \qquad V_h^{\mu,\dagger}(s) = \inf_{\nu \in \Delta_{\mathcal{B}}} (\mathbb{D}_{\mu_h \times \nu} Q_h^{\mu,\dagger})(s),$$

for all $(s,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$, where $V_{H+1}^{\mu,\dagger}(s) = 0$ for all $s \in \mathcal{S}$.

Similarly, for any Markov policy ν of the min-player, we also have the following symmetric version of Bellman equations for values of its best response:

$$Q_h^{\dagger,\nu}(s,a,b) = (r_h + \mathbb{P}_h V_{h+1}^{\dagger,\nu})(s,a,b), \qquad V_h^{\dagger,\nu}(s) = \sup_{\mu \in \Delta_{\mathcal{A}}} (\mathbb{D}_{\mu \times \nu_h} Q_h^{\dagger,\nu})(s).$$

for all $(s,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$, where $V_{H+1}^{\dagger,\nu}(s) = 0$ for all $s \in \mathcal{S}$.

Nash equilibria. Finally, by definition of Nash equilibria in Markov games, we have the following Bellman optimality equations:

$$\begin{aligned} Q_h^{\star}(s,a,b) =& (r_h + \mathbb{P}_h V_{h+1}^{\star})(s,a,b) \\ V_h^{\star}(s) = \sup_{\mu \in \Delta_{\mathcal{A}}} \inf_{\nu \in \Delta_{\mathcal{B}}} & (\mathbb{D}_{\mu \times \nu} Q_h^{\star})(s) = \inf_{\nu \in \Delta_{\mathcal{B}}} \sup_{\mu \in \Delta_{\mathcal{A}}} (\mathbb{D}_{\mu \times \nu} Q_h^{\star})(s) \end{aligned}$$

for all $(s,a,b,h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times [H]$, where $V_{H+1}^{\star}(s) = 0$ for all $s \in \mathcal{S}$.

C. Properties of Coarse Correlated Equilibrium

Recall the definition for CCE in our main paper (4), we restate it here after rescaling. For any pair of matrices $P, Q \in [0,1]^{n\times m}$, the subroutine $\mathrm{CCE}(P, Q)$ returns a distribution $\pi \in \Delta_{n\times m}$ that satisfies:

$$\begin{aligned} \mathbb{E}_{(a,b)\sim\pi} P(a,b) &\ge \max_{a^\star} \mathbb{E}_{(a,b)\sim\pi} P(a^\star, b) \\ \mathbb{E}_{(a,b)\sim\pi} Q(a,b) &\le \min_{b^\star} \mathbb{E}_{(a,b)\sim\pi} Q(a, b^\star) \end{aligned} \tag{5}$$

We make three remarks on CCE. First, a CCE always exists since a Nash equilibrium for a general-sum game with payoff matrices (P, Q) is also a CCE defined by (P, Q), and a Nash equilibrium always exists. Second, a CCE can be efficiently computed, since above constraints (5) for CCE can be rewritten as $n + m$ linear constraints on $\pi \in \Delta_{n\times m}$, which can be efficiently resolved by standard linear programming algorithm. Third, a CCE in general-sum games needs not to be a Nash equilibrium. However, a CCE in zero-sum games is guaranteed to be a Nash equalibrium.

Proposition 17. *Let $\pi = \mathrm{CCE}(Q, Q)$, and (μ, ν) be the marginal distribution over both players' actions induced by π. Then (μ, ν) is a Nash equilibrium for payoff matrix Q.*

Proof of Proposition 17. Let $N^\star$ be the value of Nash equilibrium for Q. Since $\pi = \mathrm{CCE}(Q, Q)$, by definition, we have:

$$\begin{aligned} \mathbb{E}_{(a,b)\sim\pi} Q(a,b) &\ge \max_{a^\star} \mathbb{E}_{(a,b)\sim\pi} Q(a^\star, b) = \max_{a^\star} \mathbb{E}_{b\sim\nu} Q(a^\star, b) \ge N^\star \\ \mathbb{E}_{(a,b)\sim\pi} Q(a,b) &\le \min_{b^\star} \mathbb{E}_{(a,b)\sim\pi} Q(a, b^\star) = \min_{b^\star} \mathbb{E}_{a\sim\mu} Q(a, b^\star) \le N^\star \end{aligned}$$

This gives:

$$\max_{a^\star} \mathbb{E}_{b\sim\nu} Q(a^\star, b) = \min_{b^\star} \mathbb{E}_{a\sim\mu} Q(a, b^\star) = N^\star$$

which finishes the proof. □

Intuitively, a CCE procedure can be used in Nash Q-learning for finding an approximate Nash equilibrium, because the values of upper confidence and lower confidence ($\overline{Q}$ and $\underline{Q}$) will be eventually very close, so that the preconditions of Proposition 17 becomes approximately satisfied.

D. Proof for Section 3 – Optimistic Nash Value Iteration

D.1. Proof of Theorem 3

We denote V^k, Q^k, π^k, μ^k and ν^k [4] for values and policies at the *beginning* of the k-th episode. In particular, $N_h^k(s,a,b)$ is the number we have visited the state-action tuple (s,a,b) at the h-th step before the k-th episode. $N_h^k(s,a,b,s')$ is defined by the same token. Using this notation, we can further define the empirical transition by $\widehat{\mathbb{P}}_h^k(s'|s,a,b) := N_h^k(s,a,b,s')/N_h^k(s,a,b)$. If $N_h^k(s,a,b) = 0$, we set $\widehat{\mathbb{P}}_h^k(s'|s,a,b) = 1/S$.

As a result, the bonus terms can be written as

$$\beta_h^k(s,a,b) := C\left(\sqrt{\frac{\iota H^2}{\max\{N_h^k(s,a,b), 1\}}} + \frac{H^2 S\iota}{\max\{N_h^k(s,a,b), 1\}}\right) \tag{6}$$

$$\gamma_h^k(s,a,b) := \frac{C}{H}\widehat{\mathbb{P}}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a,b) \tag{7}$$

for some large absolute constant $C > 0$.

Lemma 18. *Let c_1 be some large absolute constant. Define event E_0 to be: for all h, s, a, b, s' and $k \in [K]$,*

$$\begin{cases} |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^\star](s,a,b)| \le c_1\sqrt{\dfrac{H^2\iota}{\max\{N_h^k(s,a,b), 1\}}}, \\ |(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(s' \mid s,a,b)| \le c_1\left(\sqrt{\dfrac{\min\{\mathbb{P}_h(s' \mid s,a,b), \widehat{\mathbb{P}}_h^k(s' \mid s,a,b)\}\iota}{\max\{N_h^k(s,a,b), 1\}}} + \dfrac{\iota}{\max\{N_h^k(s,a,b), 1\}}\right). \end{cases}$$

[4] recall that (μ_h^k, ν_h^k) are the marginal distributions of π_h^k.

We have $\mathbb{P}(E_1) \geq 1 - p$*.*

Proof. The proof is standard and folklore: apply standard concentration inequalities and then take a union bound. For completeness, we provide the proof of the second one here.

Consider a fixed (s, a, b, h) tuple.

Let's consider the following equivalent random process: (a) before the agent starts, the environment samples $\{s^{(1)}, s^{(2)}, \dots, s^{(K)}\}$ independently from $\mathbb{P}_h(\cdot \mid s, a, b)$; (b) during the interaction between the agent and environment, the i^{th} time the agent reaches (s, a, b, h), the environment will make the agent transit to $s^{(i)}$. Note that the randomness induced by this interaction procedure is exactly the same as the original one, which means the probability of any event in this context is the same as in the original problem. Therefore, it suffices to prove the target concentration inequality in this 'easy' context. Denote by $\widehat{\mathbb{P}}_h^{(t)}(\cdot \mid s, a, b)$ the empirical estimate of $\mathbb{P}_h(\cdot \mid s, a, b)$ calculated using $\{s^{(1)}, s^{(2)}, \dots, s^{(t)}\}$. For a fixed t and s', by applying the Bernstein inequality and its empirical version, we have with probability at least $1 - p/S^2ABT$,

$$|(\mathbb{P}_h - \widehat{\mathbb{P}}_h^{(t)})(s' \mid s, a, b)| \leq \mathcal{O}\left(\sqrt{\frac{\min\{\mathbb{P}_h(s' \mid s, a, b), \widehat{\mathbb{P}}_h^{(t)}(s' \mid s, a, b)\}\iota}{t}} + \frac{\iota}{t}\right).$$

Now we can take a union bound over all s, a, b, h, s' and $t \in [K]$, and obtain that with probability at least $1 - p$, for all s, a, b, h, s' and $t \in [K]$,

$$|(\mathbb{P}_h - \widehat{\mathbb{P}}_h^{(t)})(s' \mid s, a, b)| \leq \mathcal{O}\left(\sqrt{\frac{\min\{\mathbb{P}_h(s' \mid s, a, b), \widehat{\mathbb{P}}_h^{(t)}(s' \mid s, a, b)\}\iota}{t}} + \frac{\iota}{t}\right).$$

Note that the agent can reach each (s, a, b, h) for at most K times, this directly implies that the third inequality also holds with probability at least $1 - p$. □

We begin with an auxiliary lemma bounding the lower-order term.

Lemma 19. *Suppose event* E_0 *holds, then there exists absolute constant* c_2 *such that: if function* $g(s)$ *satisfies* $|g|(s) \leq (\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s)$ *for all* s*, then*

$$\begin{aligned}
&|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)g(s, a, b)| \\
\leq& c_2\left(\frac{1}{H}\min\{\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s, a, b), \mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s, a, b)\} + \frac{H^2S\iota}{\max\{N_h^k(s, a, b), 1\}}\right).
\end{aligned}$$

Proof. By triangle inequality,

$$\begin{aligned}
|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)g(s, a, b)| &\leq \sum_{s'} |(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(s'|s, a, b)||g|(s') \\
&\leq \sum_{s'} |(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(s'|s, a, b)|(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s') \\
&\overset{(i)}{\leq} \mathcal{O}\left(\sum_{s'}(\sqrt{\frac{\iota\widehat{\mathbb{P}}_h^k(s'|s, a, b)}{\max\{N_h^k(s, a, b), 1\}}} + \frac{\iota}{\max\{N_h^k(s, a, b), 1\}})(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s')\right) \\
&\overset{(ii)}{\leq} \mathcal{O}\left(\sum_{s'}(\frac{\widehat{\mathbb{P}}_h^k(s'|s, a, b)}{H} + \frac{H\iota}{\max\{N_h^k(s, a, b), 1\}})(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s')\right) \\
&\leq \mathcal{O}\left(\frac{\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s, a, b)}{H} + \frac{H^2S\iota}{\max\{N_h^k(s, a, b), 1\}}\right),
\end{aligned}$$

where (i) is by the second inequality in event E_0 and (ii) is by AM-GM inequality. This proves the empirical version. Similarly, we can show

$$|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)g(s,a,b)| \le \mathcal{O}\left(\frac{\mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a,b)}{H} + \frac{H^2 S\iota}{\max\{N_h^k(s,a,b),1\}}\right),$$

Combining the two bounds completes the proof. □

Now we can prove the upper and lower bounds are indeed upper and lower bounds of the best reponses.

Lemma 20. *Suppose event E_0 holds. Then for all h, s, a, b and $k \in [K]$, we have*

$$\begin{cases} \overline{Q}_h^k(s,a,b) \ge Q_h^{\dagger,\nu^k}(s,a,b) \ge Q_h^{\mu^k,\dagger}(s,a,b) \ge \underline{Q}_h^k(s,a,b), \\ \overline{V}_h^k(s) \ge V_h^{\dagger,\nu^k}(s) \ge V_h^{\mu^k,\dagger}(s) \ge \underline{V}_h^k(s). \end{cases} \tag{8}$$

Proof. The proof is by backward induction. Suppose the bounds hold for the Q-values in the $(h+1)^{\mathrm{th}}$ step, we now establish the bounds for the V-values in the $(h+1)^{\mathrm{th}}$ step and Q-values in the h^{th}-step. For any state s:

$$\begin{aligned} \overline{V}_{h+1}^k(s) &= \mathbb{D}_{\pi_{h+1}^k}\overline{Q}_{h+1}^k(s) \\ &\ge \max_\mu \mathbb{D}_{\mu\times\nu_{h+1}^k}\overline{Q}_{h+1}^k(s) \\ &\ge \max_\mu \mathbb{D}_{\mu\times\nu_{h+1}^k} Q_{h+1}^{\dagger,\nu^k}(s) = V_{h+1}^{\dagger,\nu^k}(s). \end{aligned} \tag{9}$$

Similarly, we can show $\underline{V}_{h+1}^k(s) \le V_{h+1}^{\mu^k,\dagger}(s)$. Therefore, we have: for all s,

$$\overline{V}_{h+1}^k(s) \ge V_{h+1}^{\dagger,\nu^k}(s) \ge V_{h+1}^\star(s) \ge V_{h+1}^{\mu^k,\dagger}(s) \ge \underline{V}_{h+1}^k(s).$$

Now consider an arbitrary triple (s,a,b) in the h^{th} step. We have

$$\begin{aligned} &(\overline{Q}_h^k - Q_h^{\dagger,\nu^k})(s,a,b) \\ \ge &\min\left\{(\widehat{\mathbb{P}}_h^k\overline{V}_{h+1}^k - \mathbb{P}_h V_{h+1}^{\dagger,\nu^k} + \beta_h^k + \gamma_h^k)(s,a,b), 0\right\} \\ \ge &\min\left\{(\widehat{\mathbb{P}}_h^k V_{h+1}^{\dagger,\nu^k} - \mathbb{P}_h V_{h+1}^{\dagger,\nu^k} + \beta_h^k + \gamma_h^k)(s,a,b), 0\right\} \\ = &\min\Bigg\{\underbrace{(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(V_{h+1}^{\dagger,\nu^k} - V_{h+1}^\star)(s,a,b)}_{(A)} + \underbrace{(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^\star(s,a,b)}_{(B)} + (\beta_h^k + \gamma_h^k)(s,a,b), 0\Bigg\}. \end{aligned} \tag{10}$$

Invoking Lemma 19 with $g = V_{h+1}^{\dagger,\nu^k} - V_{h+1}^\star$,

$$|(A)| \le \mathcal{O}\left(\frac{\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a,b)}{H} + \frac{H^2 S\iota}{\max\{N_h^k(s,a,b),1\}}\right).$$

By the first inequality in event E_0,

$$|(B)| \le \mathcal{O}\left(\sqrt{\frac{H^2\iota}{\max\{N_h^k(s,a,b),1\}}}\right).$$

Plugging the two inequalities above back into (10) and recalling the definition of β_h^k and γ_h^k, we obtain $\overline{Q}_h^k(s,a,b) \ge Q_h^{\dagger,\nu^k}(s,a,b)$. Similarly, we can show $\underline{Q}_h^k(s,a,b) \le Q_h^{\mu^k,\dagger}(s,a,b)$. □

Finally we come to the proof of Theorem 3.

Proof of Theorem 3. Suppose event E_0 holds. We first upper bound the regret. By Lemma 20, the regret can be upper bounded by

$$\sum_k (V_1^{\dagger,\nu^k}(s_1^k) - V_1^{\mu^k,\dagger}(s_1^k)) \le \sum_k (\overline{V}_1^k(s_1^k) - \underline{V}_1^k(s_1^k)).$$

For brevity's sake, we define the following notations:

$$\begin{cases} \Delta_h^k := (\overline{V}_h^k - \underline{V}_h^k)(s_h^k), \\ \zeta_h^k := \Delta_h^k - (\overline{Q}_h^k - \underline{Q}_h^k)(s_h^k, a_h^k, b_h^k), \\ \xi_h^k := \mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k, b_h^k) - \Delta_{h+1}^k. \end{cases} \tag{11}$$

Let $\mathcal{F}_h^k$ be the σ-field generated by the following random variables:

$$\{(s_i^j, a_i^j, b_i^j, r_i^j)\}_{(i,j)\in[H]\times[k-1]} \bigcup \{(s_i^k, a_i^k, b_i^k, r_i^k)\}_{i\in[h-1]} \bigcup \{s_h^k\}.$$

It's easy to check ζ_h^k and ξ_h^k are martingale differences with respect to $\mathcal{F}_h^k$. With a slight abuse of notation, we use β_h^k to refer to $\beta_h^k(s_h^k, a_h^k, b_h^k)$ and N_h^k to refer to $N_h^k(s_h^k, a_h^k, b_h^k)$ in the following proof.

We have

$$\begin{aligned}
\Delta_h^k =& \zeta_h^k + \left(\overline{Q}_h^k - \underline{Q}_h^k\right)(s_h^k, a_h^k, b_h^k) \\
\le& \zeta_h^k + 2\beta_h^k + 2\gamma_h^k + \widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)\left(s_h^k, a_h^k, b_h^k\right) \\
\overset{(i)}{\le}& \zeta_h^k + 2\beta_h^k + 2\gamma_h^k + \mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)\left(s_h^k, a_h^k, b_h^k\right) + c_2\left(\frac{\mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k, b_h^k)}{H} + \frac{H^2 S\iota}{\max\{N_h^k, 1\}}\right) \\
\overset{(ii)}{\le}& \zeta_h^k + 2\beta_h^k + \mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)\left(s_h^k, a_h^k, b_h^k\right) + 2c_2 C\left(\frac{\mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h^k, a_h^k, b_h^k)}{H} + \frac{H^2 S\iota}{\max\{N_h^k, 1\}}\right) \\
\le& \zeta_h^k + \left(1 + \frac{2c_2 C}{H}\right)\mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)\left(s_h^k, a_h^k, b_h^k\right) + 4c_2 C\left(\sqrt{\frac{\iota H^2}{\max\{N_h^k, 1\}}} + \frac{H^2 S\iota}{\max\{N_h^k, 1\}}\right) \\
=& \zeta_h^k + \left(1 + \frac{2c_2 C}{H}\right)\xi_h^k + \left(1 + \frac{2c_2 C}{H}\right)\Delta_{h+1}^k + 4c_2 C\left(\sqrt{\frac{\iota H^2}{\max\{N_h^k, 1\}}} + \frac{H^2 S\iota}{\max\{N_h^k, 1\}}\right)
\end{aligned}$$

where (i) and (ii) follow from Lemma 19.

Define $c_3 := 1 + 2c_2 C$ and $\kappa := 1 + c_3/H$. Recursing this argument for $h \in [H]$ and summing over k,

$$\sum_{k=1}^K \Delta_1^k \le \sum_{k=1}^K \sum_{h=1}^H \left[\kappa^{h-1}\zeta_h^k + \kappa^h \xi_h^k + \mathcal{O}\left(\sqrt{\frac{\iota H^2}{\max\{N_h^k, 1\}}} + \frac{H^2 S\iota}{\max\{N_h^k, 1\}}\right)\right].$$

By Azuma-Hoeffding inequality, with probability at least $1 - p$,

$$\begin{cases} \sum_{k=1}^K \sum_{h=1}^H \kappa^{h-1}\zeta_h^k \le \mathcal{O}\Big(H\sqrt{HK\iota}\Big) = \mathcal{O}\Big(\sqrt{H^2 T\iota}\Big), \\ \sum_{k=1}^K \sum_{h=1}^H \kappa^h \xi_h^k \le \mathcal{O}\Big(H\sqrt{HK\iota}\Big) = \mathcal{O}\Big(\sqrt{H^2 T\iota}\Big). \end{cases} \tag{12}$$

By pigeon-hole argument,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\sqrt{\max\{N_h^k,1\}}} \le \sum_{s,a,b,h:\ N_h^K(s,a,b)>0}\ \sum_{n=1}^{N_h^K(s,a,b)}\frac{1}{\sqrt{n}} + HSAB \le \mathcal{O}\Big(\sqrt{HSABT} + HSAB\Big),$$

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\max\{N_h^k,1\}} \le \sum_{s,a,b,h:\ N_h^K(s,a,b)>0}\ \sum_{n=1}^{N_h^K(s,a,b)}\frac{1}{n} + HSAB \le \mathcal{O}(HSAB\iota).$$

Put everything together, with probability at least $1-2p$ (one p comes from $\mathbb{P}(E_0)\ge 1-p$ and the other is for equation (12)),

$$\sum_{k=1}^{K}(V_1^{\dagger,\nu^k}(s_1^k) - V_1^{\mu^k,\dagger}(s_1^k)) \le \mathcal{O}\Big(\sqrt{H^3SABT\iota} + H^3S^2AB\iota^2\Big)$$

For the PAC guarantee, recall that we choose $\pi^{\text{out}} = \pi^{k^\star}$ such that $k^\star = \text{argmin}_k\left(\overline{V}_1^k - \underline{V}_1^k\right)(s_1)$. As a result,

$$(V_1^{\dagger,\nu^{k^\star}} - V_1^{\mu^{k^\star},\dagger})(s_1) \le (\overline{V}_1^{k^\star} - \underline{V}_1^{k^\star})(s_1) \le \frac{1}{K}\mathcal{O}\Big(\sqrt{H^3SABT\iota} + H^3S^2AB\iota^2\Big),$$

which concludes the proof. □

D.2. Proof of Theorem 4

We use the same notation as in Appendix D.1 except the form of bonus. Besides, we define the empirical variance operator

$$\widehat{\mathbb{V}}_h^k V(s,a,b) := \text{Var}_{s'\sim\widehat{\mathbb{P}}_h^k(\cdot|s,a,b)}V(s')$$

and the true (population) variance operator

$$\mathbb{V}_h V(s,a,b) := \text{Var}_{s'\sim\mathbb{P}_h(\cdot|s,a,b)}V(s')$$

for any function $V\in\Delta^S$. If $N_h^k(s,a,b)=0$, we simply set $\widehat{\mathbb{V}}_h^k V(s,a,b) := H^2$ regardless of the choice of V.

As a result, the bonus terms can be written as

$$\beta_h^k(s,a,b) := C\left(\sqrt{\frac{\iota\widehat{\mathbb{V}}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a,b)}{\max\{N_h^k(s,a,b),1\}}} + \frac{H^2S\iota}{\max\{N_h^k(s,a,b),1\}}\right) \tag{13}$$

for some absolute constant $C>0$.

Lemma 21. *Let c_1 be some large absolute constant. Define event E_1 to be: for all h,s,a,b,s' and $k\in[K]$,*

$$\begin{cases} |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^\star](s,a,b)| \le c_1\left(\sqrt{\dfrac{\widehat{\mathbb{V}}_h^k V_{h+1}^\star(s,a,b)\iota}{\max\{N_h^k(s,a,b),1\}}} + \dfrac{H\iota}{\max\{N_h^k(s,a,b),1\}}\right), \\ |(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(s'\mid s,a,b)| \le c_1\left(\sqrt{\dfrac{\min\{\mathbb{P}_h(s'\mid s,a,b),\widehat{\mathbb{P}}_h^k(s'\mid s,a,b)\}\iota}{\max\{N_h^k(s,a,b),1\}}} + \dfrac{\iota}{\max\{N_h^k(s,a,b),1\}}\right), \\ \|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(\cdot\mid s,a,b)\|_1 \le c_1\sqrt{\dfrac{S\iota}{\max\{N_h^k(s,a,b),1\}}}. \end{cases}$$

We have $\mathbb{P}(E_1)\ge 1-p$.

The proof of Lemma 21 is highly similar to that of Lemma 18. Specifically, the first two can be proved by following basically the same argument in Lemma 18; the third one is standard (e.g., equation (12) in (Azar et al., 2017)). We omit the proof here.

Since the proof of Lemma 19 does not depend on the form of the bonus, it can also be applied in this section. As in Appendix D.1, we will prove the upper and lower bounds are indeed upper and lower bounds of the best reponses.

Lemma 22. *Suppose event E_1 holds. Then for all h, s, a, b and $k \in [K]$, we have*

$$\begin{cases} \overline{Q}_h^k(s,a,b) \geq Q_h^{\dagger,\nu^k}(s,a,b) \geq Q_h^{\mu^k,\dagger}(s,a,b) \geq \underline{Q}_h^k(s,a,b), \\ \overline{V}_h^k(s) \geq V_h^{\dagger,\nu^k}(s) \geq V_h^{\mu^k,\dagger}(s) \geq \underline{V}_h^k(s). \end{cases} \tag{14}$$

Proof. The proof is by backward induction and very similar to that of Lemma 20. Suppose the bounds hold for the Q-values in the $(h+1)^{\mathrm{th}}$ step, we now establish the bounds for the V-values in the $(h+1)^{\mathrm{th}}$ step and Q-values in the h^{th}-step.

The proof for the V-values is the same as (9).

For the Q-values, the decomposition (10) still holds and (A) is bounded using Lemma 19 as before. The only difference is that we need to bound (B) more carefully.

First, by the first inequality in event E_1,

$$|(B)| \leq \mathcal{O}\left(\sqrt{\frac{\widehat{\mathbb{V}}_h^k V_{h+1}^\star(s,a,b)\iota}{\max\{N_h^k(s,a,b),1\}}} + \frac{H\iota}{\max\{N_h^k(s,a,b),1\}}\right).$$

By the relation of V-values in the $(h+1)^{\mathrm{th}}$ step,

$$\begin{aligned} &|[\widehat{\mathbb{V}}_h^k(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2] - \widehat{\mathbb{V}}_h^k V_{h+1}^\star|(s,a,b) \\ \leq &|[\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2 - (\widehat{\mathbb{P}}_h^k V_{h+1}^\star)^2|(s,a,b) + |\widehat{\mathbb{P}}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2]^2 - \widehat{\mathbb{P}}_h^k(V_{h+1}^\star)^2|(s,a,b) \\ \leq &4H\widehat{\mathbb{P}}_h^k|(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2 - V_{h+1}^\star|(s,a,b) \\ \leq &4H\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s,a,b), \end{aligned} \tag{15}$$

which implies

$$\begin{aligned} &\sqrt{\frac{\iota\widehat{\mathbb{V}}_h^k V_{h+1}^\star(s,a,b)}{\max\{N_h^k(s,a,b),1\}}} \\ \leq &\sqrt{\frac{\iota[\widehat{\mathbb{V}}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2] + 4H\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s,a,b)}{\max\{N_h^k(s,a,b),1\}}} \\ \leq &\sqrt{\frac{\iota\widehat{\mathbb{V}}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a,b)}{\max\{N_h^k(s,a,b),1\}}} + \sqrt{\frac{4\iota H\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)](s,a,b)}{\max\{N_h^k(s,a,b),1\}}} \\ \overset{(i)}{\leq} &\sqrt{\frac{\iota\widehat{\mathbb{V}}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2](s,a,b)}{\max\{N_h^k(s,a,b),1\}}} + \frac{\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)}{H} + \frac{4H^2\iota}{\max\{N_h^k(s,a,b),1\}}, \end{aligned} \tag{16}$$

where (i) is by AM-GM inequality.

Plugging the above inequalities back into (10) and recalling the definition of β_h^k and γ_h^k completes the proof. □

We need one more lemma to control the error of the empirical variance estimator:

Lemma 23. *Suppose event E_1 holds. Then for all h, s, a, b and $k \in [K]$, we have*

$$|\widehat{\mathbb{V}}_h^k[(\overline{V}_{h+1}^k + \underline{V}_{h+1}^k)/2] - \mathbb{V}_h V_{h+1}^{\pi^k}|(s,a,b)$$

$$\leq 4H\mathbb{P}_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s,a,b) + \mathcal{O}\bigg(1 + \frac{H^4 S\iota}{\max\{N^k_h(s,a,b),1\}}\bigg).$$

Proof. By Lemma 22, we have $\overline{V}^k_h(s) \geq V^{\pi^k}_h(s) \geq \underline{V}^k_h(s)$. As a result,

$$\begin{aligned}
&|\widehat{\mathbb{V}}^k_h[(\overline{V}^k_{h+1} + \underline{V}^k_{h+1})/2] - \mathbb{V}_h V^{\pi^k}_{h+1}|(s,a,b) \\
=&|[\widehat{\mathbb{P}}^k_h(\overline{V}^k_{h+1} + \underline{V}^k_{h+1})^2/4 - \mathbb{P}_h(V^{\pi^k}_{h+1})^2](s,a,b) - [(\widehat{\mathbb{P}}^k_h(\overline{V}^k_{h+1} + \underline{V}^k_{h+1}))^2/4 - (\mathbb{P}_h V^{\pi^k}_{h+1})^2](s,a,b)| \\
\leq&[\widehat{\mathbb{P}}^k_h(\overline{V}^k_{h+1})^2 - \mathbb{P}_h(\underline{V}^k_{h+1})^2 - (\widehat{\mathbb{P}}^k_h\underline{V}^k_{h+1})^2 + (\mathbb{P}_h\overline{V}^k_{h+1})^2](s,a,b) \\
\leq&[|(\widehat{\mathbb{P}}^k_h - \mathbb{P}_h)(\overline{V}^k_{h+1})^2| + |\mathbb{P}_h[(\overline{V}^k_{h+1})^2 - (\underline{V}^k_{h+1})^2]| \\
&+ |(\widehat{\mathbb{P}}^k_h\underline{V}^k_{h+1})^2 - (\mathbb{P}_h\underline{V}^k_{h+1})^2| + |(\mathbb{P}_h\underline{V}^k_{h+1})^2 - (\mathbb{P}_h\overline{V}^k_{h+1})^2|](s,a,b)
\end{aligned}$$

These terms can be bounded separately by using event E_1:

$$\begin{aligned}
&|(\widehat{\mathbb{P}}^k_h - \mathbb{P}_h)(\overline{V}^k_{h+1})^2|(s,a,b) \leq H^2\|(\widehat{\mathbb{P}}^k_h - \mathbb{P}_h)(\cdot \mid s,a,b)\|_1 \leq \mathcal{O}(H^2\sqrt{\frac{S\iota}{\max\{N^k_h(s,a,b),1\}}}), \\
&|\mathbb{P}_h[(\overline{V}^k_{h+1})^2 - (\underline{V}^k_{h+1})^2]|(s,a,b) \leq 2H[\mathbb{P}_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})](s,a,b), \\
&|(\widehat{\mathbb{P}}^k_h\underline{V}^k_{h+1})^2 - (\mathbb{P}_h\underline{V}^k_{h+1})^2|(s,a,b) \leq 2H[(\widehat{\mathbb{P}}^k_h - \mathbb{P}_h)\underline{V}^k_{h+1}](s,a,b) \leq \mathcal{O}(H^2\sqrt{\frac{S\iota}{\max\{N^k_h(s,a,b),1\}}}), \\
&|(\mathbb{P}_h\underline{V}^k_{h+1})^2 - (\mathbb{P}_h\overline{V}^k_{h+1})^2|(s,a,b) \leq 2H[\mathbb{P}_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})](s,a,b).
\end{aligned}$$

Combining with $H^2\sqrt{\frac{S\iota}{\max\{N^k_h(s,a,b),1\}}} \leq 1 + \frac{H^4 S\iota}{\max\{N^k_h(s,a,b),1\}}$ completes the proof. □

Finally we come to the proof of Theorem 4.

Proof of Theorem 4. Suppose event E_1 holds. We define Δ^k_h, ζ^k_h abd ξ^k_h as in the proof of Theorem 3. As before we have

$$\begin{aligned}
\Delta^k_h \leq& \zeta^k_h + \left(1 + \frac{c_3}{H}\right)\mathbb{P}_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})\left(s^k_h, a^k_h, b^k_h\right) \\
&+ 4c_2 C\left(\sqrt{\frac{\iota\widehat{\mathbb{V}}^k_h[(\overline{V}^k_{h+1} + \underline{V}^k_{h+1})/2](s^k_h, a^k_h, b^k_h)}{\max\{N^k_h(s^k_h, a^k_h, b^k_h),1\}}} + \frac{H^2 S\iota}{\max\{N^k_h(s^k_h, a^k_h, b^k_h),1\}}\right).
\end{aligned} \tag{17}$$

By Lemma 23,

$$\begin{aligned}
&\sqrt{\frac{\iota\widehat{\mathbb{V}}^k_h[(\overline{V}^k_{h+1} + \underline{V}^k_{h+1})/2](s,a,b)}{\max\{N^k_h(s,a,b),1\}}} \\
\leq&\mathcal{O}\left(\sqrt{\frac{\iota\mathbb{V}_h V^{\pi^k}_{h+1}(s,a,b) + \iota}{\max\{N^k_h(s,a,b),1\}}} + \sqrt{\frac{H\iota\mathbb{P}_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s,a,b)}{\max\{N^k_h(s,a,b),1\}}} + \frac{H^2\sqrt{S}\iota}{\max\{N^k_h(s,a,b),1\}}\right) \\
\leq&c_4\left(\sqrt{\frac{\iota\mathbb{V}_h V^{\pi^k}_{h+1}(s,a,b) + \iota}{\max\{N^k_h(s,a,b),1\}}} + \frac{\mathbb{P}_h(\overline{V}^k_{h+1} - \underline{V}^k_{h+1})(s,a,b)}{H} + \frac{H^2\sqrt{S}\iota}{\max\{N^k_h(s,a,b),1\}}\right),
\end{aligned} \tag{18}$$

where c_4 is some absolute constant. Define $c_5 := 4c_2c_4C + c_3$ and $\kappa := 1 + c_5/H$. Plugging (18) back into (17), we have

$$\Delta^k_h \leq \kappa\Delta^k_{h+1} + \kappa\xi^k_h + \zeta^k_h + \mathcal{O}\bigg(\sqrt{\frac{\iota\mathbb{V}_h V^{\pi^k}_{h+1}(s^k_h, a^k_h, b^k_h)}{N^k_h(s^k_h, a^k_h, b^k_h)}} + \sqrt{\frac{\iota}{N^k_h(s^k_h, a^k_h, b^k_h)}} + \frac{H^2 S\iota}{N^k_h(s^k_h, a^k_h, b^k_h)}\bigg)\bigg\}. \tag{19}$$

Recursing this argument for $h \in [H]$ and summing over k,

$$\sum_{k=1}^{K} \Delta_1^k \leq \sum_{k=1}^{K}\sum_{h=1}^{H}\left[\kappa^{h-1}\zeta_h^k + \kappa^h \xi_h^k + \mathcal{O}\left(\sqrt{\frac{\iota \mathbb{V}_h V_{h+1}^{\pi^k}(s_h^k, a_h^k, b_h^k)}{\max\{N_h^k, 1\}}} + \sqrt{\frac{\iota}{\max\{N_h^k, 1\}}} + \frac{H^2 S\iota}{\max\{N_h^k, 1\}}\right)\right].$$

The remaining steps are the same as that in the proof of Theorem 3 except that we need to bound the sum of variance term.

By Cauchy-Schwarz,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\mathbb{V}_h V_{h+1}^{\pi^k}(s_h^k, a_h^k, b_h^k)}{\max\{N_h^k(s_h^k, a_h^k, b_h^k), 1\}}} \leq \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_h V_{h+1}^{\pi^k}(s_h^k, a_h^k, b_h^k) \cdot \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{\max\{N_h^k(s_h^k, a_h^k, b_h^k), 1\}}}.$$

By the Law of total variation and standard martingale concentration (see Lemma C.5 in Jin et al. (2018) for a formal proof), with probability at least $1-p$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_h V_{h+1}^{\pi^k}(s_h^k, a_h^k, b_h^k) \leq \mathcal{O}\big(HT + H^3\iota\big).$$

Putting all relations together, we obtain that with probability at least $1-2p$ (one p comes from $\mathbb{P}(E_1) \geq 1-p$ and the other comes from the inequality for bounding the variance term),

$$\text{Regret}(K) = \sum_{k=1}^{K}(V_1^{\dagger,\nu^k} - V_1^{\mu^k,\dagger})(s_1) \leq \mathcal{O}(\sqrt{H^2 SABT\iota} + H^3 S^2 AB\iota^2).$$

Rescaling p completes the proof. □

E. Proof for Section 4 – Reward-Free Learning

E.1. Proof of Theorem 5

In this section, we prove Theorem 5 for the single reward function case, i.e., $N = 1$. The proof for multiple reward functions ($N > 1$) simply follows from taking a union bound, that is, replacing the failure probability p by Np.

Let (μ^k, ν^k) be an arbitrary Nash-equilibrium policy of $\widehat{\mathcal{M}}^k := (\widehat{\mathbb{P}}^k, \widehat{r}^k)$, where $\widehat{\mathbb{P}}^k$ and $\widehat{r}^k$ are our empirical estimate of the transition and the reward at the beginning of the k'th episode in Algorithm 2, respectively. We use $N_h^k(s,a,b)$ to denote the number we have visited the state-action tuple (s,a,b) at the h'=th step before the k'th episode. And the bonus used in the k'th episode can be written as

$$\beta_h^k(s,a,b) := C\left(\sqrt{\frac{H^2\iota}{\max\{N_h^k(s,a,b), 1\}}} + \frac{H^2 S\iota}{\max\{N_h^k(s,a,b), 1\}}\right), \tag{20}$$

where $\iota = \log(SABT/p)$ and C is some large absolute constant.

We use $\widehat{Q}^k$ and $\widehat{V}^k$ to denote the empirical optimal value functions of $\widehat{\mathcal{M}}^k$ as following.

$$\begin{cases} \widehat{Q}_h^k(s,a,b) = (\widehat{\mathbb{P}}_h^k \widehat{V}_{h+1})(s,a,b) + \widehat{r}_h^k(s,a,b), \\ \widehat{V}_h^k(s) = \max_{\mu}\min_{\nu} \mathbb{D}_{\mu\times\nu}\widehat{Q}_h^k(s). \end{cases} \tag{21}$$

Since (μ^k, ν^k) is a Nash-equilibrium policy of $\widehat{\mathcal{M}}^k$, we also have $\widehat{V}_h^k(s) = \mathbb{D}_{\mu^k\times\nu^k}\widehat{Q}_h^k(s)$.

We begin with stating a useful property of matrix game that will be frequently used in our analysis. Since its proof is quite simple, we omit it here.

Lemma 24. *Let* $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^{A\times B}$ *and* Δ_d *be the* d*-dimensional simplex. Suppose* $|\mathbf{X}-\mathbf{Y}| \le \mathbf{Z}$*, where the inequality is entry-wise. Then*

$$\left| \max_{\mu\in\triangle_A} \min_{\nu\in\triangle_B} \mu^\top \mathbf{X}\nu - \max_{\mu\in\triangle_A} \min_{\nu\in\triangle_B} \mu^\top \mathbf{Y}\nu \right| \le \max_{i,j} \mathbf{Z}_{ij}. \tag{22}$$

Lemma 25. *Let* c_1 *be some large absolute constant such that* $c_1^2 + c_1 \le C$*. Define event* E_1 *to be: for all* h, s, a, b, s' *and* $k \in [K]$*,*

$$\begin{cases} |[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^\star](s,a,b)| \le \dfrac{c_1}{10}\sqrt{\dfrac{H^2\iota}{\max\{N_h^k(s,a,b),1\}}}, \\ |(\widehat{r}_h^k - r_h)(s,a,b)| \le \dfrac{c_1}{10}\sqrt{\dfrac{H^2\iota}{\max\{N_h^k(s,a,b),1\}}}, \\ |(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(s' \mid s,a,b)| \le \dfrac{c_1}{10}\left(\sqrt{\dfrac{\widehat{\mathbb{P}}_h^k(s' \mid s,a,b)\iota}{\max\{N_h^k(s,a,b),1\}}} + \dfrac{\iota}{\max\{N_h^k(s,a,b),1\}}\right). \end{cases} \tag{23}$$

We have $\mathbb{P}(E_1) \ge 1-p$*.*

Proof. The proof is standard: apply concentration inequalities and then take a union bound. For completeness, we provide the proof of the third one here.

Consider a fixed (s,a,b,h) tuple.

Let's consider the following equivalent random process: (a) before the agent starts, the environment samples $\{s^{(1)}, s^{(2)}, \ldots, s^{(K)}\}$ independently from $\mathbb{P}_h(\cdot \mid s,a,b)$; (b) during the interaction between the agent and the environment, the i^{th} time the agent reaches (s,a,b,h), the environment will make the agent transit to $s^{(i)}$. Note that the randomness induced by this interaction procedure is exactly the same as the original one, which means the probability of any event in this context is the same as in the original problem. Therefore, it suffices to prove the target concentration inequality in this 'easy' context. Denote by $\widehat{\mathbb{P}}_h^{(t)}(\cdot \mid s,a,b)$ the empirical estimate of $\mathbb{P}_h(\cdot \mid s,a,b)$ calculated using $\{s^{(1)}, s^{(2)}, \ldots, s^{(t)}\}$. For a fixed t and s', by the empirical Bernstein inequality, we have with probability at least $1 - p/S^2ABT$,

$$|(\mathbb{P}_h - \widehat{\mathbb{P}}_h^{(t)})(s' \mid s,a,b)| \le \mathcal{O}\left(\sqrt{\frac{\widehat{\mathbb{P}}_h^{(t)}(s' \mid s,a,b)\iota}{t}} + \frac{\iota}{t}\right).$$

Now we can take a union bound over all s, a, b, h, s' and $t \in [K]$, and obtain that with probability at least $1-p$, for all s, a, b, h, s' and $t \in [K]$,

$$|(\mathbb{P}_h - \widehat{\mathbb{P}}_h^{(t)})(s' \mid s,a,b)| \le \mathcal{O}\left(\sqrt{\frac{\widehat{\mathbb{P}}_h^{(t)}(s' \mid s,a,b)\iota}{t}} + \frac{\iota}{t}\right).$$

Note that the agent can reach each (s,a,b,h) for at most K times, so we conclude the third inequality also holds with probability at least $1-p$. □

The following lemma states that the empirical optimal value functions are close to the true optimal ones, and their difference is controlled by the exploration value functions calculated in Algorithm 2.

Lemma 26. *Suppose event* E_1 *(defined in Lemma 25) holds. Then for all* h, s, a, b *and* $k \in [K]$*, we have,*

$$\begin{cases} \left|\widehat{Q}_h^k(s,a,b) - Q_h^\star(s,a,b)\right| \le \widetilde{Q}_h^k(s,a,b), \\ \left|\widehat{V}_h^k(s) - V_h^\star(s)\right| \le \widetilde{V}_h^k(s). \end{cases} \tag{24}$$

Proof. Let's prove by backward induction on h. The case of $h = H+1$ holds trivially.

Assume the conclusion hold for $(h+1)$'th step. For h'th step,

$$\begin{aligned}
&\left|\widehat{Q}_h^k(s,a,b) - Q_h^\star(s,a,b)\right| \\
\le{}& \min\left\{\left|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^\star](s,a,b)\right| + |(\widehat{r}_h^k - r_h)(s,a,b)| + \left|[\widehat{\mathbb{P}}_h^k(\widehat{V}_{h+1}^k - V_{h+1}^\star)](s,a,b)\right|, H\right\} \\
\overset{(i)}{\le}{}& \min\left\{\beta_h^k(s,a,b) + (\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s,a,b), H\right\} \overset{(ii)}{=} \widetilde{Q}_h^k(s,a,b),
\end{aligned} \tag{25}$$

where (i) follows from the induction hypothesis and event E_1, and (ii) follows from the definition of $\widetilde{Q}_h^k$. By Lemma 24, we immediately obtain $|\widehat{V}_h^k(s) - V_h^\star(s)| \le \widetilde{V}_h^k(s)$. □

Now, we are ready to establish the key lemma in our analysis using Lemma 26.

Lemma 27. *Suppose event E_1 (defined in Lemma 25) holds. Then for all h, s, a, b and $k \in [K]$, we have*

$$\begin{cases} |\widehat{Q}_h^k(s,a,b) - Q_h^{\dagger,\nu^k}(s,a,b)| \le \alpha_h \widetilde{Q}_h^k(s,a,b), \\ |\widehat{V}_h^k(s) - V_h^{\dagger,\nu^k}(s)| \le \alpha_h \widetilde{V}_h^k(s), \end{cases} \tag{26}$$

and

$$\begin{cases} |\widehat{Q}_h^k(s,a,b) - Q_h^{\mu^k,\dagger}(s,a,b)| \le \alpha_h \widetilde{Q}_h^k(s,a,b), \\ |\widehat{V}_h^k(s) - V_h^{\mu^k,\dagger}(s)| \le \alpha_h \widetilde{V}_h^k(s), \end{cases} \tag{27}$$

where $\alpha_{H+1} = 0$ and $\alpha_h = [(1+\frac{1}{H})\alpha_{h+1} + \frac{1}{H}] \le 4$.

Proof. We only prove the first set of inequalities. The second one follows exactly the same. Again, the proof is by performing backward induction on h. It is trivial to see the conclusion holds for $(H+1)$'th step with $\alpha_{H+1} = 0$. Now, assume the conclusion holds for $(h+1)$'th step. For h'th step,

$$\begin{aligned}
&|\widehat{Q}_h^k(s,a,b) - Q_h^{\dagger,\nu^k}(s,a,b)| \\
\le&\min\left\{|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(V_{h+1}^{\dagger,\nu^k} - V_{h+1}^\star)](s,a,b)| + |(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V_{h+1}^\star(s,a,b)|\right. \\
&\left.+ |(\widehat{r}_h^k - r_h)(s,a,b)| + |[\widehat{\mathbb{P}}_h(\widehat{V}_{h+1}^k - V_{h+1}^{\dagger,\nu^k})](s,a,b)|, H\right\} \\
\le&\min\left\{\underbrace{|[(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(V_{h+1}^{\dagger,\nu^k} - V_{h+1}^\star)](s,a,b)|}_{(T_1)} + c_1\sqrt{\frac{H^2\iota}{\max\{N_h^k(s,a,b),1\}}} + \underbrace{|[\widehat{\mathbb{P}}_h(\widehat{V}_{h+1}^k - V_{h+1}^{\dagger,\nu^k})](s,a,b)|}_{(T_2)}, H\right\},
\end{aligned} \tag{28}$$

where the second inequality follows from the definition of event E_1.

We can control the term (T_1) by combining Lemma 26 and the induction hypothesis to bound $|V_{h+1}^{\dagger,\nu^k} - V_{h+1}^\star|$, and then applying the third inequality in event E_1:

$$\begin{aligned}
(T_1) \le& \sum_{s'} |\widehat{\mathbb{P}}_h^k(s' \mid s,a,b) - \mathbb{P}_h(s' \mid s,a,b)||V_{h+1}^{\dagger,\nu^k} - V_{h+1}^\star(s')| \\
\le& \sum_{s'} |\widehat{\mathbb{P}}_h^k(s' \mid s,a,b) - \mathbb{P}_h(s' \mid s,a,b)|\Big(|V_{h+1}^{\dagger,\nu^k} - \widehat{V}_{h+1}^k(s')| + |\widehat{V}_{h+1}^k - V_{h+1}^\star(s')|\Big) \\
\le& \sum_{s'} |\widehat{\mathbb{P}}_h^k(s' \mid s,a,b) - \mathbb{P}_h(s' \mid s,a,b)|(\alpha_{h+1}+1)\widetilde{V}_{h+1}^k \\
\le& \frac{(\alpha_{h+1}+1)}{H}(\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s,a,b) + \frac{c_1^2(\alpha_{h+1}+1)H^2 S\iota}{\max\{N_h^k(s,a,b),1\}}.
\end{aligned} \tag{29}$$

The term (T_2) is bounded by directly applying the induction hypothesis

$$|[\widehat{\mathbb{P}}_h(\widehat{V}_{h+1}^k - V_{h+1}^{\dagger,\nu^k})](s,a,b)| \le \alpha_{h+1}[\widehat{\mathbb{P}}_h \widetilde{V}_{h+1}^k](s,a,b). \tag{30}$$

Plugging (29) and (30) into (28), we obtain

$$\begin{aligned}
&\left|\widehat{Q}_h^k(s,a,b) - Q_h^{\dagger,\nu^k}(s,a,b)\right| \\
\le& \min\left\{(1+\frac{1}{H})\alpha_{h+1} + \frac{1}{H}[\widehat{\mathbb{P}}_h^k\widetilde{V}_{h+1}^k](s,a,b) + c_1\sqrt{\frac{H^2\iota}{\max\{N_h^k(s,a,b),1\}}} + \frac{c_1^2(\alpha_{h+1}+1)H^2S\iota}{\max\{N_h^k(s,a,b),1\}}, H\right\} \\
\overset{(i)}{\le}& \min\left\{\left((1+\frac{1}{H})\alpha_{h+1} + \frac{1}{H}\right)[\widehat{\mathbb{P}}_h^k\widetilde{V}_{h+1}^k](s,a,b) + \beta_h^k(s,a,b), H\right\} \\
\overset{(ii)}{\le}& \left((1+\frac{1}{H})\alpha_{h+1} + \frac{1}{H}\right)\widetilde{Q}_h^k(s,a,b),
\end{aligned} \tag{31}$$

where (i) follows from the definition of β_h^k, and (ii) follows from the definition of $\widetilde{Q}_h^k$. Therefore, by (31), choosing $\alpha_h = [(1+\frac{1}{H})\alpha_{h+1} + \frac{1}{H}]$ suffices for the purpose of induction.

Now, let's prove the inequality for V functions.

$$\begin{aligned}
|(\widehat{V}_h^k - V_h^{\dagger,\nu^k})(s)| &\overset{(i)}{=} |\max_{\mu\in\triangle_A}(\mathbb{D}_{\mu,\nu^k}\widehat{Q}_h^k)(s) - \max_{\mu\in\triangle_A}(\mathbb{D}_{\mu,\nu^k}Q_h^{\dagger,\nu^k})(s)| \\
&\overset{(ii)}{\le} \max_{a,b}\left[\alpha_h\widetilde{Q}_h^k(s,a,b)\right] = \alpha_h\widetilde{V}_h^k(s),
\end{aligned} \tag{32}$$

where (i) follows from the definition of $\widehat{V}_h^k$ and $V_h^{\dagger,\nu^k}$, and (ii) uses (31) and Lemma 24. □

Theorem 28 (Guarantee for UCB-VI from Azar et al. (2017))**.** *For any $p \in (0,1]$, choose the exploration bonus β_t in Algrothm 2 as (20). Then, with probability at least $1-p$,*

$$\sum_{k=1}^K \widetilde{V}_1^k(s_1) \le \mathcal{O}(\sqrt{H^4SAK\iota} + H^3S^2A\iota^2).$$

Proof of Theorem 5. Recall that out $= \arg\min_{k\in[K]}\widetilde{V}_h^k(s)$. By Lemma 27 and Theorem 28, with probability at least $1-2p$,

$$\begin{aligned}
V_h^{\dagger,\nu^{\text{out}}}(s) - V_h^{\mu^{\text{out}},\dagger}(s) \le& |V_h^{\dagger,\nu^{\text{out}}}(s) - \widehat{V}_h^{\text{out}}(s)| + |\widehat{V}_h^{\text{out}}(s) - V_h^{\mu^{\text{out}},\dagger}(s)| \\
\le& 8\widetilde{V}_h^{\text{out}}(s) \le \mathcal{O}(\sqrt{\frac{H^4SA\iota}{K}} + \frac{H^3S^2A\iota^2}{K}).
\end{aligned} \tag{33}$$

Rescaling p completes the proof. □

E.2. Vanilla Nash Value Iteration

Here, we provide one optional algorithm, Vanilla Nash VI, for computing the Nash equilibrium policy for a *known* model. Its only difference from the value iteration algorithm for MDPs is that the maximum operator is replaced by the minimax operator in Line 7. We remark that the Nash equilibrium for a two-player zero-sum game can be computed in polynomial time.

By recalling the definition of best responses in Appendix B, one can directly see that the output policy $(\hat{\mu},\hat{\nu})$ is a Nash equilibrium for $\widehat{\mathcal{M}}$.

E.3. Proof of Theorem 6

In this section, we first prove a $\Theta(AB/\epsilon^2)$ lower bound for reward-free learning of matrix games, i.e., $S = H = 1$, and then generalize it to $\Theta(SABH^2/\epsilon^2)$ for reward-free learning of Markov games.

Algorithm 5 Vanilla Nash Value Iteration

```
Input: model $\widehat{\mathcal{M}} = (\widehat{\mathbb{P}}, \widehat{r})$.
Initialize: for all $(s, a, b)$, $V_{H+1}(s, a, b) \leftarrow 0$.
for step $h = H, H-1, \dots, 1$ do
    for $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ do
        $Q_h(s, a, b) \leftarrow [\widehat{\mathbb{P}}_h V_{h+1}](s, a, b) + \widehat{r}_h(s, a, b)$.
    for $s \in \mathcal{S}$ do
        $(\hat{\mu}_h(\cdot \mid s), \hat{\nu}_h(\cdot \mid s)) \leftarrow$ NASH-ZERO-SUM$(Q_h(s, \cdot, \cdot))$.
        $V_h(s) \leftarrow \hat{\mu}_h(\cdot \mid s)^\top Q_h(s, \cdot, \cdot) \hat{\nu}_h(\cdot \mid s)$.
Output $(\hat{\mu}, \hat{\nu}) \leftarrow \{(\hat{\mu}_h(\cdot \mid s), \hat{\nu}_h(\cdot \mid s))\}_{(h,s) \in [H] \times \mathcal{S}}$.
```

E.3.1. Reward-free learning of matrix games

In the matrix game, let the max-player pick row and the min-player pick column. We consider the following family of Bernoulli matrix games:

$$\mathfrak{M}(\epsilon) = \left\{ \mathcal{M}^{a^\star b^\star} \in \mathbb{R}^{A \times B} \text{ with } \mathcal{M}^{a^\star b^\star}_{ab} = \frac{1}{2} + (1 - 2 \cdot \mathbf{1}\{a \neq a^\star \& b = b^\star\})\epsilon : \ (a^\star, b^\star) \in [A] \times [B] \right\}, \tag{34}$$

where in matrix game $\mathcal{M}^{a^\star b^\star}$, the reward is sampled from $\mathrm{Bernoulli}(\mathcal{M}^{a^\star b^\star}_{ab})$ if the max-player picks the a'th row and the min-player picks the b'th column.

$$\begin{array}{ccccccccc}
 & & & & \text{Min-player} & & & & \\
 & \text{action} & 1 & \dots & b^\star - 1 & b^\star & b^\star + 1 & \dots & B \\
 & 1 & + & \dots & + & - & + & \dots & + \\
 & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 & a^\star - 1 & + & \dots & + & - & + & \dots & + \\
\text{Max-player} & a^\star & + & \dots & + & + & + & \dots & + \\
 & a^\star + 1 & + & \dots & + & - & + & \dots & + \\
 & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
 & A & + & \dots & + & - & + & \dots & +
\end{array} \tag{35}$$

Above, we visualize $\mathcal{M}^{a^\star b^\star}$ by using $+$ and $-$ to represent $1/2 + \epsilon$ and $1/2 - \epsilon$, respectively. It is direct to see that the optimal (Nash equilibrium) policy for the max-player is always picking the $a^\star$'th row. If the max-player picks the $a^\star$'th row with probability smaller than $2/3$, it is at least $\epsilon/10$ suboptimal.

Lemma 29. *Consider an arbitrary **fixed** matrix game $\mathcal{M}^{a^\star b^\star}$ from $\mathfrak{M}(\epsilon)$ and $N \in \mathbb{N}$. If there exists an algorithm $\mathcal{A}$ such that when running on $\mathcal{M}^{a^\star b^\star}$, it uses at most N samples and outputs an $\epsilon/10$-optimal policy with probability at least p, then there exists an algorithm $\hat{\mathcal{A}}$ that can identify $a^\star$ with probability at least p using at most N samples.*

Proof. We simply define $\hat{\mathcal{A}}$ as running algorithm $\mathcal{A}$ and choosing the most played row by its output policy as the guess for $a^\star$. Because any $\epsilon/10$-optimal policy must play $a^\star$ with probability at least $2/3$, we obtain $\hat{\mathcal{A}}$ will correctly identify $a^\star$ with probability at least p. □

Lemma 29 directly implies that in order to prove the desired lower bound for reward-free matrix games:

Claim 30. *for **any** reward-free algorithm $\mathcal{A}$ using at most $N = AB/(10^3\epsilon^2)$ samples, there exists a matrix game $\mathcal{M}^{a^\star b^\star}$ in $\mathfrak{M}(\epsilon)$ such that when running $\mathcal{A}$ on $\mathcal{M}^{a^\star b^\star}$, it will output a policy that is **at least** $\epsilon/10$ suboptimal for the max-player with probability at least $1/4$,*

it suffices to prove the following claim:

Claim 31. *for **any** reward-free algorithm $\hat{\mathcal{A}}$ using at most $N = AB/(10^3\epsilon^2)$ samples, there exists a matrix game $\mathcal{M}^{a^\star b^\star}$ in $\mathfrak{M}(\epsilon)$ such that when running $\hat{\mathcal{A}}$ on $\mathcal{M}$, it will fail to identify the optimal row with probability at least $1/4$.*

Remark 32. *By Lemma 29, the existence of such 'ideal' $\mathcal{A}$ implies the existence of an 'ideal' $\hat{\mathcal{A}}$, so to prove such 'ideal' $\mathcal{A}$ does not exist (Claim 30), it suffices to show such 'ideal' $\hat{\mathcal{A}}$ does not exist (Claim 31).*

Proof of Claim 31. WLOG, we assume $\hat{\mathcal{A}}$ is deterministic. Since $\hat{\mathcal{A}}$ is *reward-free*, being deterministic means that in the exploration phase algorithm $\hat{\mathcal{A}}$ always pulls each arm (a, b) for some *fixed* $n(a, b)$ times (because there is no information revealed in this phase), and in the planning phase it outputs a guess for $a^\star$, which is a deterministic function of the reward information revealed.

We define the following notations:

- L: the stochastic reward information revealed after algorithm $\hat{\mathcal{A}}$'s pulling.
- $\mathbb{P}_\star$: the probability measure induced by picking $\mathcal{M}^{a^\star b^\star}$ uniformly at random from $\mathfrak{M}(\epsilon)$ and then running $\hat{\mathcal{A}}$ on $\mathcal{M}$.
- $\mathbb{P}_{ab}$: the probability measure induced by running $\hat{\mathcal{A}}$ on $\mathcal{M}^{ab}$.
- $\mathbb{P}_{0b}$: the probability measure induced by running $\mathcal{A}$ on matrix $\mathcal{M}^{0b}$, whose b'th column are all $(1/2 - \epsilon)$'s and other columns are all $(1/2 + \epsilon)$'s.[5]
- $f(L)$: the output of $\hat{\mathcal{A}}$ based on the stochastic reward information L revealed. More precisely, f is function mapping from $[0, 1]^N$ to $[A]$.

We have

$$\begin{aligned}
\mathbb{P}_\star(f(L) \neq a^\star) &\geq \frac{1}{AB}\sum_{a,b}\mathbb{P}_{0b}(f(L) \neq a) - \frac{1}{AB}\sum_{a,b}\|\mathbb{P}_{ab}(L = \cdot) - \mathbb{P}_{0b}(L = \cdot)\|_1 \\
&\geq 1 - \frac{1}{A} - \frac{1}{AB}\sum_{a,b}\sqrt{2\mathrm{KL}(\mathbb{P}_{0b}\|\mathbb{P}_{ab})} \\
&= 1 - \frac{1}{A} - \frac{1}{AB}\sum_{a,b}\sqrt{2n(a,b)[(\frac{1}{2} - \epsilon)\log\frac{\frac{1}{2}-\epsilon}{\frac{1}{2}+\epsilon} + (\frac{1}{2} + \epsilon)\log\frac{\frac{1}{2}+\epsilon}{\frac{1}{2}-\epsilon}]} \qquad (36) \\
&\geq 1 - \frac{1}{A} - \frac{10}{AB}\sum_{a,b}\sqrt{n(a,b)\epsilon^2} \\
&\geq 1 - \frac{1}{A} - \sqrt{\frac{100N\epsilon^2}{AB}},
\end{aligned}$$

where the second inequality follows from $\sum_{a,b}\mathbb{P}_{0b}(f(L) \neq a) = \sum_{a,b}[1 - \mathbb{P}_{0b}(f(L) = a)] = B(A-1)$ and Pinsker's inequality. Finally, plugging in $N = AB/(10^3\epsilon^2)$ concludes the proof. □

Remark 33. *The arguments in proving Claim 31 basically follows the same line in proving lower bounds for multi-arm bandits (e.g., see Lattimore & Szepesvári, 2018).*

E.3.2. Reward-free learning of Markov games

Now let's generalize the $\Theta(AB/\epsilon^2)$ lower bound for reward-free learning of matrix games to $\Theta(SABH^2/\epsilon^2)$ for reward-free learning of Markov games. We can follow the same way of generalizing a lower bound for multi-arm bandits to a lower bound for MDPs (see e.g., Dann & Brunskill, 2015; Lattimore & Szepesvári, 2018; Zhang et al., 2020b).

Proof sketch. Given the family of Bernoulli matrix games $\mathfrak{M}(\cdot)$ defined in (34), we simply construct a Markov game to consist of SH Bernoulli matrix games $\{M^{s,h}\}_{(s,h)\in[S]\times[H]}$ where $M^{s,h}$'s are sampled independently and identically from the uniform distribution over $\mathfrak{M}(\epsilon/H)$. We will define the transition measure to be totally 'uniform at random' so that in each episode the agent will always reach each $M^{s,h}$ with probability $1/S$ (it is not $1/(SH)$ because in each episode the agent can visit H matrix games). As a result, to guarantee ϵ-optimality, the output policy must be at least $2\epsilon/H$-optimal for at least $SH/2$ different $M^{s,h}$'s. Recall Claim 30 shows learning a $2\epsilon/H$-optimal policy for a single $M^{s,h}$ requires $\Omega(H^2AB/\epsilon^2)$ samples. Therefore, we need $\Omega(H^3AB/\epsilon^2)$ samples in total for learning $SH/2$ different $M^{s,h}$'s.

[5] We comment that matrix $\mathcal{M}^{0b}$ does not belong to $\mathfrak{M}(\epsilon)$.

Below, we provide a rigorous proof where the constants may be slightly different from those in our sketch. We remark that although the notations we will use are involved, they are only introduced for rigorousness and there is no real technical difficulty or new informative idea in the following proof.

Construction We define the following family of Markov games:

$$\mathfrak{J}(\epsilon) := \left\{ \mathcal{J}(\boldsymbol{a}^\star, \boldsymbol{b}^\star) : \ (\boldsymbol{a}^\star, \boldsymbol{b}^\star) \in [A]^{H\times S} \times [B]^{H\times S} \right\}, \tag{37}$$

where MG $\mathcal{J}(\boldsymbol{a}^\star, \boldsymbol{b}^\star)$ is defined as

- **States and actions:** $\mathcal{J}(\boldsymbol{a}^\star, \boldsymbol{b}^\star)$ is a finite-horizon MG with $S+1$ states and of length $H+1$. There is a fixed initial state s_0 in the first step, S states $\{s_1, \ldots, s_S\}$ in the remaining steps. The two players have A and B actions, respectively.
- **Rewards:** there is no reward in the first step. For the remaining steps $h \in \{2, \ldots, H+1\}$, if the agent takes action (a, b) at state s_i in the h^{th} step, it will receive a binary reward sampled from

$$\text{Bernoulli}\Big(\frac{1}{2} + (1 - 2\cdot \mathbf{1}\{a \neq \boldsymbol{a}^\star_{h-1,i} \& b = \boldsymbol{b}^\star_{h-1,i}\})\frac{\epsilon}{H}\Big)$$

- **Transitions:** The agent always starts at a fixed initial state s_0 in the first step Regardless of the current state, actions and index of steps, the agent will always transit to one of $s_1, \ldots, s_S$ uniformly at random.

It is direct to see that $\mathcal{J}(\boldsymbol{a}^\star, \boldsymbol{b}^\star)$ is a collection of SH matrix games from $\mathfrak{M}(\epsilon/H)$. Therefore, the optimal policy for the max-player is to always pick action $\boldsymbol{a}^\star_{h-1,i}$ whenever it reaches state s_i at step h ($h \geq 2$).

Formal proof of Theorem 6. Now, let's use $\mathfrak{J}(\epsilon)$ to prove the $\Theta(SABH^2/\epsilon^2)$ lower bound (in terms of number of episodes) for reward-free learning of Markov games. We start by proving an analogue of Lemma 29.

Lemma 34. *Consider an arbitrary fixed matrix game $\mathcal{J}(\boldsymbol{a}^\star, \boldsymbol{b}^\star)$ from $\mathfrak{J}(\epsilon)$ and $N \in \mathbb{N}$. If an algorithm $\mathcal{A}$ can output a policy that is at most $\epsilon/10^3$ suboptimal with probability at least p using at most N samples, then there exists an algorithm $\hat{\mathcal{A}}$ that can correctly identify at least $SH - \lfloor SH/500 \rfloor$ entries of $\boldsymbol{a}^\star$ with probability at least p using at most N samples.*

Proof. Denote by π the output policy for the max player. Denote by Z the collection of (h, i)'s in $[H] \times [S]$ such that $\pi_{h+1}(\boldsymbol{a}^\star_{h,i} \mid s_i) \leq 2/3$.

Observe that each time the max player picks a suboptimal action, it will incur an $2\epsilon/H$ suboptimality in expectation. As a result, if π is at most $\epsilon/10^3$-suboptimal, we must have

$$\frac{1}{S} \sum_{(h,i)\in Z} (1 - \pi_{h+1}(\boldsymbol{a}^\star_{h,i} \mid s_i)) \times \frac{2\epsilon}{H} \leq \frac{\epsilon}{10^3},$$

which implies $|Z| \leq SH/500$, that is, for at most $\lfloor SH/500 \rfloor$ different (h, i)'s, $\pi_{h+1}(\boldsymbol{a}^\star_{h,i} \mid s_i) \leq 2/3$. Therefore, we can simply pick $\text{argmax}_a\, \pi_{h+1}(a \mid s_i)$ as the guess for $\boldsymbol{a}^\star_{h,i}$. Since policy π is at most $\epsilon/10^3$ suboptimal with probability at least p, our guess will be correct for at least $SH - \lfloor SH/500 \rfloor$ different (s, h) pairs also with probability no smaller than p. □

Similar to the funtion of Lemma 29, Lemma 34 directly implies that in order to prove the desired lower bound for reward-free learning of Markov games:

Claim 35. *for any reward-free algorithm $\mathcal{A}$ that interacts with the environment for at most $K = SABH^2/(10^4\epsilon^2)$ episodes, there exists $\mathcal{J} \in \mathfrak{J}(\epsilon)$ such that when running $\mathcal{A}$ on $\mathcal{J}$, it will output a policy that is at least $\epsilon/10^3$ suboptimal for the max-player with probability at least $1/4$,*

it suffices to prove the following claim:

Claim 36. *for any reward-free learning algorithm $\hat{\mathcal{A}}$ that interacts with the environment for at most $K = ABSH^2/(10^4\epsilon^2)$ episodes, there exists $\mathcal{J} \in \mathfrak{J}(\epsilon)$ such that when running $\hat{\mathcal{A}}$ on $\mathcal{J}$, it will fail to correctly identify $\boldsymbol{a}^\star_{h,i}$ for at least $\lfloor SH/500 \rfloor + 1$ different (h, i) pairs with probability at least $1/4$.*

Proof of Claim 36. Denote by $\mathbb{P}_\star$ ($\mathbb{E}_\star$) the probability measure (expectation) induced by picking $\mathcal{J}$ uniformly at random from $\mathfrak{J}(\epsilon)$ and then running $\hat{\mathcal{A}}$ on $\mathcal{J}$. Denote by $n_{\rm wrong}$ the number of (s,h) pairs for which $\hat{\mathcal{A}}$ fails to identify the optimal actions. Denote by $\mathrm{error}_{h,i}$ the indicator function of the event that $\hat{\mathcal{A}}$ fails to identify the optimal action for $(h+1,i)$.

We prove by contradiction. Suppose for any $\mathcal{J}\in\mathfrak{J}(\epsilon)$, $\hat{\mathcal{A}}$ can identify the optimal actions for at least $SH-\lfloor SH/500\rfloor$ different (s,h) pairs with probability larger than $3/4$. Then we have

$$\mathbb{E}_\star[n_{\rm wrong}]\le\frac{1}{4}\times SH+\frac{3}{4}\times\left\lfloor\frac{SH}{500}\right\rfloor\le\frac{101SH}{400}.$$

Since $\sum_{(h,i)\in[H]\times[S]}\mathbb{E}_\star[\mathrm{error}_{h,i}]=\mathbb{E}_\star[n_{\rm wrong}]$, there must exists $(h',i')\in[H]\times[S]$ such that $\mathbb{E}_\star[\mathrm{error}_{h',i'}]\le 101/400$. However, in the following, we show that for every $(h,i)\in[H]\times[S]$, $\mathbb{E}_\star[\mathrm{error}_{h,i}]\ge 1/3$. As a result, we obtain a contraction and Claim 36 holds. In the remainder of this section, we will prove for every $(h,i)\in[H]\times[S]$, $\mathbb{E}_\star[\mathrm{error}_{h,i}]\ge 1/3$.

WLOG, we assume $\hat{\mathcal{A}}$ is deterministic. It suffices to consider an arbitrary *fixed* (h',i') pair and prove $\mathbb{E}_\star[\mathrm{error}_{h',i'}]\ge 1/3$.

For technical reason, we introduce a new MG $\mathcal{J}_{-(h',i')}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$ as below:

- **States, actions and transitions:** same as $\mathcal{J}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$.
- **Rewards:** there is no reward in the first step. For the remaining steps $h\in\{2,\ldots,H+1\}$, if the agent takes action (a,b) at state s_i in the $h^{\rm th}$ step such that $(h-1,i)\neq(h',i')$, it will receive a binary reward sampled from

$$\mathrm{Bernoulli}\Big(\frac{1}{2}+(1-2\cdot\mathbf{1}\{a\neq\boldsymbol{a}^\star_{h-1,i}\&b=\boldsymbol{b}^\star_{h-1,i}\})\frac{\epsilon}{H}\Big),$$

  otherwise it will receive a binary reward sampled from

$$\mathrm{Bernoulli}\Big(\frac{1}{2}+(1-2\cdot\mathbf{1}\{b=\boldsymbol{b}^\star_{h-1,i}\})\frac{\epsilon}{H}\Big).$$

Remark 37. *Briefly speaking, $\mathcal{J}_{-(h',i')}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$ is the same as $\mathcal{J}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$ except the matrix game embedded at state $s_{i'}$ at step $h'+1$, where for the max player all its actions are equivalently bad* [6]. *Finally, we remark that $\mathcal{J}_{-(h',i')}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$ is* ***independent*** *of $\boldsymbol{a}^\star_{h',i'}$.*

To proceed, we introduce (and recall) the following notations:

- $n(a,b)$: the number of times $\hat{\mathcal{A}}$ picks action (a,b) at state $s_{i'}$ at step $(h'+1)$ within K episode.
- $\mathbb{P}_{\boldsymbol{a}^\star\boldsymbol{b}^\star}$ ($\mathbb{E}_{\boldsymbol{a}^\star\boldsymbol{b}^\star}$): the probability measure (expectation) induced by running algorithm $\hat{\mathcal{A}}$ on $\mathcal{J}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$.
- $\mathbb{P}^-_{\boldsymbol{a}^\star\boldsymbol{b}^\star}$ ($\mathbb{E}^-_{\boldsymbol{a}^\star\boldsymbol{b}^\star}$): the probability measure (expectation) induced by running algorithm $\hat{\mathcal{A}}$ on $\mathcal{J}_{-(h',i')}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$.
- $\mathbb{P}_\star$ ($\mathbb{E}_\star$) the probability measure (expectation) induced by picking $\mathcal{J}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$ uniformly at random from $\mathfrak{J}(\epsilon)$ and running $\hat{\mathcal{A}}$ on $\mathcal{J}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$.
- L: the whole interaction trajectory of states, actions and rewards produced by algorithm $\hat{\mathcal{A}}$ within K episodes.
- $f(L)$: the guess of $\hat{\mathcal{A}}$ for $\boldsymbol{a}^\star_{h',i'}$ based on L.

The key observation here is that for any $(a,b)\in[A]\times[B]$ and $(\boldsymbol{a}^\star,\boldsymbol{b}^\star)\in[A]^{H\times S}\times[B]^{H\times S}$, the expectation $\mathbb{E}^-_{\boldsymbol{a}^\star\boldsymbol{b}^\star}[n(a,b)]$ is independent of $(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$ because the agent does not receive any reward information when interacting with the environment and the transition dynamics of different $\mathcal{J}_{-(h',i')}(\boldsymbol{a}^\star,\boldsymbol{b}^\star)$'s are exactly the same. For simplicity of notation, we denote this expectation by $m(a,b)$. Moreover, note that $\sum_{a,b}m(a,b)=K/S$ because the agent always reach state $s_{i'}$ in step $(h'+1)$ with probability $1/S$ regardless of the actions taken.

[6] A graphic illustration based on (35) would be replacing the column $[-,\ldots,-,+,-,\ldots,-]^\top$ with a column of all $-$'s in the matrix game embedded at state $s_{i'}$ at step $h'+1$.

By mimicking the arguments in (36), we have

$$
\begin{aligned}
&\mathbb{E}_\star[\mathrm{error}_{h',i'}] = \mathbb{P}_\star(f(L) \neq \boldsymbol{a}^\star_{h',i'}) \\
&= \frac{1}{(AB)^{SH}} \sum_{(\boldsymbol{a}^\star,\boldsymbol{b}^\star)\in[A]^{H\times S}\times[B]^{H\times S}} \mathbb{P}_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(f(L) \neq \boldsymbol{a}^\star_{h',i'}) \\
&\geq \frac{1}{(AB)^{SH}} \sum_{\boldsymbol{a}^\star,\boldsymbol{b}^\star} \left(\mathbb{P}^-_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(f(L) \neq \boldsymbol{a}^\star_{h',i'}) - \left\| \mathbb{P}^-_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(L=\cdot) - \mathbb{P}_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(L=\cdot) \right\|_1 \right) \\
&= 1 - \frac{1}{A} - \frac{1}{(AB)^{SH}} \sum_{\boldsymbol{a}^\star,\boldsymbol{b}^\star} \left\| \mathbb{P}^-_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(L=\cdot) - \mathbb{P}_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(L=\cdot) \right\| \\
&\geq 1 - \frac{1}{A} - \frac{1}{(AB)^{SH}} \sum_{\boldsymbol{a}^\star,\boldsymbol{b}^\star} \sqrt{2\mathrm{KL}(\mathbb{P}^-_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(L=\cdot), \mathbb{P}_{\boldsymbol{a}^\star\boldsymbol{b}^\star}(L=\cdot))} \\
&= 1 - \frac{1}{A} - \frac{1}{(AB)^{SH}} \sum_{\boldsymbol{a}^\star,\boldsymbol{b}^\star} \sqrt{2m(\boldsymbol{a}^\star_{h',i'}, \boldsymbol{b}^\star_{h',i'})\big[(\frac{1}{2} - \frac{\epsilon}{H}) \log \frac{\frac{1}{2} - \frac{\epsilon}{H}}{\frac{1}{2} + \frac{\epsilon}{H}} + (\frac{1}{2} + \frac{\epsilon}{H}) \log \frac{\frac{1}{2} + \frac{\epsilon}{H}}{\frac{1}{2} - \frac{\epsilon}{H}}\big]} \\
&= 1 - \frac{1}{A} - \frac{1}{AB} \sum_{(a,b)\in[A]\times[B]} \sqrt{2m(a,b)\big[(\frac{1}{2} - \frac{\epsilon}{H}) \log \frac{\frac{1}{2} - \frac{\epsilon}{H}}{\frac{1}{2} + \frac{\epsilon}{H}} + (\frac{1}{2} + \frac{\epsilon}{H}) \log \frac{\frac{1}{2} + \frac{\epsilon}{H}}{\frac{1}{2} - \frac{\epsilon}{H}}\big]} \\
&\geq 1 - \frac{1}{A} - \frac{10}{AB} \sum_{a,b} \sqrt{m(a,b)\frac{\epsilon^2}{H^2}} \\
&\geq 1 - \frac{1}{A} - \frac{10\epsilon}{ABH} \sqrt{AB \sum_{a,b} m(a,b)} = 1 - \frac{1}{A} - \sqrt{\frac{100K\epsilon^2}{SABH^2}}.
\end{aligned}
\tag{38}
$$

Plugging in $K = SABH^2/(10^4\epsilon^2)$ completes the proof. □

F. Proof for Appendix A – Multi-player General-sum Markov Games

F.1. Proof of Theorem 15

F.1.1. NE Version

In this section, we prove Theorem 15 (NE version). As before, we begin with proving the optimistic estimations are indeed upper bounds of the corresponding V-value and Q-value functions.

Lemma 38. *With probability* $1 - p$, *for any* $(s, \boldsymbol{a}, h, i)$ *and* $k \in [K]$:

$$\overline{Q}^k_{h,i}(s, \boldsymbol{a}) \geq Q^{\dagger,\pi^k_{-i}}_{h,i}(s, \boldsymbol{a}), \quad \underline{Q}^k_{h,i}(s, \boldsymbol{a}) \leq Q^{\pi^k}_{h,i}(s, \boldsymbol{a}), \tag{39}$$

$$\overline{V}^k_{h,i}(s) \geq V^{\dagger,\pi^k_{-i}}_{h,i}(s), \quad \underline{V}^k_{h,i}(s) \leq V^{\pi^k}_{h,i}(s). \tag{40}$$

Proof. For each fixed k, we prove this by induction from $h = H+1$ to $h = 1$. For the base case, we know at the $(H+1)$-th step,$\overline{V}^k_{H+1,i}(s) = V^{\dagger,\pi^k_{-i}}_{H+1,i}(s) = 0$. Now, assume the inequality (40) holds for the $(h+1)$-th step, for the h-th step, by the definition of Q-functions,

$$
\begin{aligned}
\overline{Q}^k_{h,i}(s, \boldsymbol{a}) - Q^{\dagger,\pi^k_{-i}}_{h,i}(s, \boldsymbol{a}) &= \left[\widehat{\mathbb{P}}^k_h \overline{V}^k_{h+1,i}\right](s, \boldsymbol{a}) - \left[\mathbb{P}_h V^{\dagger,\pi^k_{-i}}_{h+1,i}\right](s, \boldsymbol{a}) + \beta_t \\
&= \underbrace{\widehat{\mathbb{P}}^k_h \left(\overline{V}^k_{h+1,i} - V^{\dagger,\pi^k_{-i}}_{h+1,i}\right)(s, \boldsymbol{a})}_{(A)} + \underbrace{\left(\widehat{\mathbb{P}}^k_h - \mathbb{P}_h\right) V^{\dagger,\pi^k_{-i}}_{h+1,i}(s, \boldsymbol{a})}_{(B)} + \beta_t.
\end{aligned}
$$

By induction hypothesis, for any s', $\left(\overline{V}_{h+1,i}^{k} - V_{h+1,i}^{\dagger,\pi_{-i}^{k}}\right)(s') \geq 0$, and thus $(A) \geq 0$. By uniform concentration (e.g., Lemma 12 in Bai & Jin, 2020), $(B) \leq C\sqrt{SH^2\iota/N_h^k(s,\boldsymbol{a})} = \beta_t$. Putting everything together we have $\overline{Q}_{h,i}^{k}(s,\boldsymbol{a}) - Q_{h,i}^{\dagger,\pi_{-i}^{k}}(s,\boldsymbol{a}) \geq 0$. The second inequality can be proved similarly.

Now assume inequality (39) holds for the h-th step, by the definition of V-functions and Nash equilibrium,

$$\overline{V}_{h,i}^{k}(s) = \mathbb{D}_{\pi^k}\overline{Q}_{h,i}^{k}(s) = \max_{\mu}\mathbb{D}_{\mu\times\pi_{-i}^{k}}\overline{Q}_{h,i}^{k}(s).$$

By Bellman equation,

$$V_{h,i}^{\dagger,\pi_{-i}^{k}}(s) = \max_{\mu}\mathbb{D}_{\mu\times\pi_{-i}^{k}}Q_{h,i}^{\dagger,\pi_{-i}^{k}}(s).$$

Since by induction hypothesis, for any $(s,\boldsymbol{a})$, $\overline{Q}_{h,i}^{k}(s,\boldsymbol{a}) \geq Q_{h,i}^{\dagger,\pi_{-i}^{k}}(s,\boldsymbol{a})$. As a result, we also have $\overline{V}_{h,i}^{k}(s) \geq V_{h,i}^{\dagger,\pi_{-i}^{k}}(s)$, which is exactly inequality (40) for the h-th step. The second inequality can be proved similarly. □

Proof of Theorem 15. Let us focus on the i-th player and ignore the subscript when there is no confusion. To bound

$$\max_i\left(V_{1,i}^{\dagger,\pi_{-i}^{k}} - V_{1,i}^{\pi^k}\right)(s_h^k) \leq \max_i\left(\overline{V}_{1,i}^{k} - \underline{V}_{1,i}^{k}\right)(s_h^k),$$

we notice the following propogation:

$$\begin{cases}(\overline{Q}_{h,i}^{k} - \underline{Q}_{h,i}^{k})(s,\boldsymbol{a}) \leq \widehat{\mathbb{P}}_h^k(\overline{V}_{h+1,i}^{k} - \underline{V}_{h+1,i}^{k})(s,\boldsymbol{a}) + 2\beta_h^k(s,\boldsymbol{a}),\\ (\overline{V}_{h,i} - \underline{V}_{h,i})(s) = [\mathbb{D}_{\pi_h}(\overline{Q}_{h,i}^{k} - \underline{Q}_{h,i}^{k})](s).\end{cases} \tag{41}$$

We can define $\widetilde{Q}_h^k$ and $\widetilde{V}_h^k$ recursively by $\widetilde{V}_{H+1}^k = 0$ and

$$\begin{cases}\widetilde{Q}_h^k(s,\boldsymbol{a}) = \widehat{\mathbb{P}}_h^k\widetilde{V}_{h+1}^k(s,\boldsymbol{a}) + 2\beta_h^k(s,\boldsymbol{a}),\\ \widetilde{V}_h^k(s) = [\mathbb{D}_{\pi_h}\widetilde{Q}_h^k](s).\end{cases} \tag{42}$$

Then we can prove inductively that for any k, h, s and $\boldsymbol{a}$ we have

$$\begin{cases}\max_i(\overline{Q}_{h,i}^{k} - \underline{Q}_{h,i}^{k})(s,\boldsymbol{a}) \leq \widetilde{Q}_h^k(s,\boldsymbol{a}),\\ \max_i(\overline{V}_{h,i} - \underline{V}_{h,i})(s) \leq \widetilde{V}_h^k(s).\end{cases} \tag{43}$$

Thus we only need to bound $\sum_{k=1}^K \widetilde{V}_1^k(s)$. Define the shorthand notation

$$\begin{cases}\beta_h^k := \beta_h^k(s_h^k,\boldsymbol{a}_h^k),\\ \Delta_h^k := \widetilde{V}_h^k(s_h^k),\\ \zeta_h^k := [\mathbb{D}_{\pi^k}\widetilde{Q}_h^k]\left(s_h^k\right) - \widetilde{Q}_h^k(s_h^k,\boldsymbol{a}_h^k),\\ \xi_h^k := [\mathbb{P}_h\widetilde{V}_{h+1}^k](s_h^k,\boldsymbol{a}_h^k) - \Delta_{h+1}^k.\end{cases} \tag{44}$$

We can check ζ_h^k and ξ_h^k are martingale difference sequences. As a result,

$$\begin{aligned}\Delta_h^k =&\mathbb{D}_{\pi^k}\widetilde{Q}_h^k\left(s_h^k\right)\\ =&\zeta_h^k + \widetilde{Q}_h^k\left(s_h^k,\boldsymbol{a}_h^k\right)\\ =&\zeta_h^k + 2\beta_h^k + [\widehat{\mathbb{P}}_h^k\widetilde{V}_{h+1}^k]\left(s_h^k,\boldsymbol{a}_h^k\right)\end{aligned}$$

$$\begin{aligned}&\le \zeta_h^k + 3\beta_h^k + \left[\mathbb{P}_h \widetilde{V}_{h+1}^k\right]\left(s_h^k, \boldsymbol{a}_h^k\right)\\ &= \zeta_h^k + 3\beta_h^k + \xi_h^k + \Delta_{h+1}^k.\end{aligned}$$

Recursing this argument for $h \in [H]$ and taking the sum,

$$\sum_{k=1}^{K} \Delta_1^k \le \sum_{k=1}^{K} \left(\zeta_h^k + 3\beta_h^k + \xi_h^k\right) \le O\left(S\sqrt{H^3 T\iota \prod_{i=1}^{M} A_i}\right).$$

□

F.1.2. CCE Version

The proof is very similar to the NE version. Specifically, the only part that uses the properties of NE there is Lemma 38. We prove a counterpart here.

Lemma 39. *With probability* $1-p$*, for any* $(s, \boldsymbol{a}, h, i)$ *and* $k \in [K]$*:*

$$\overline{Q}_{h,i}^k(s, \boldsymbol{a}) \ge Q_{h,i}^{\dagger, \pi_{-i}^k}(s, \boldsymbol{a}), \quad \underline{Q}_{h,i}^k(s, \boldsymbol{a}) \le Q_{h,i}^{\pi^k}(s, \boldsymbol{a}), \tag{45}$$

$$\overline{V}_{h,i}^k(s) \ge V_{h,i}^{\dagger, \pi_{-i}^k}(s), \quad \underline{V}_{h,i}^k(s) \le V_{h,i}^{\pi^k}(s). \tag{46}$$

Proof. For each fixed k, we prove this by induction from $h = H+1$ to $h = 1$. For the base case, we know at the $(H+1)$-th step, $\overline{V}_{H+1,i}^k(s) = V_{H+1,i}^{\dagger, \pi_{-i}^k}(s) = 0$. Now, assume the inequality (40) holds for the $(h+1)$-th step, for the h-th step, by the definition of Q-functions,

$$\begin{aligned}\overline{Q}_{h,i}^k(s, \boldsymbol{a}) - Q_{h,i}^{\dagger, \pi_{-i}^k}(s, \boldsymbol{a}) &= \left[\widehat{\mathbb{P}}_h^k \overline{V}_{h+1,i}^k\right](s, \boldsymbol{a}) - \left[\mathbb{P}_h V_{h+1,i}^{\dagger, \pi_{-i}^k}\right](s, \boldsymbol{a}) + \beta_t\\ &= \underbrace{\widehat{\mathbb{P}}_h^k \left(\overline{V}_{h+1,i}^k - V_{h+1,i}^{\dagger, \pi_{-i}^k}\right)(s, \boldsymbol{a})}_{(A)} + \underbrace{\left(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h\right) V_{h+1,i}^{\dagger, \pi_{-i}^k}(s, \boldsymbol{a})}_{(B)} + \beta_t.\end{aligned}$$

By induction hypothesis, for any s', $\left(\overline{V}_{h+1,i}^k - V_{h+1,i}^{\dagger, \pi_{-i}^k}\right)(s') \ge 0$, and thus $(A) \ge 0$. By uniform concentration, $(B) \le C\sqrt{SH^2\iota / N_h^k(s, \boldsymbol{a})} = \beta_t$. Putting everything together we have $\overline{Q}_{h,i}^k(s, \boldsymbol{a}) - Q_{h,i}^{\dagger, \pi_{-i}^k}(s, \boldsymbol{a}) \ge 0$. The second inequality can be proved similarly.

Now assume inequality (45) holds for the h-th step, by the definition of V-functions and CCE,

$$\overline{V}_{h,i}^k(s) = \mathbb{D}_{\pi^k} \overline{Q}_{h,i}^k(s) \ge \max_{\mu} \mathbb{D}_{\mu \times \pi_{-i}^k} \overline{Q}_{h,i}^k(s).$$

By Bellman equation,

$$V_{h,i}^{\dagger, \pi_{-i}^k}(s) = \max_{\mu} \mathbb{D}_{\mu \times \pi_{-i}^k} Q_{h,i}^{\dagger, \pi_{-i}^k}(s).$$

Since by induction hypothesis, for any $(s, \boldsymbol{a})$, $\overline{Q}_{h,i}^k(s, \boldsymbol{a}) \ge Q_{h,i}^{\dagger, \pi_{-i}^k}(s, \boldsymbol{a})$. As a result, we also have $\overline{V}_{h,i}^k(s) \ge V_{h,i}^{\dagger, \pi_{-i}^k}(s)$, which is exactly inequality (40) for the h-th step. The second inequality can be proved similarly. □

F.1.3. CE Version

The proof is very similar to the NE version. Specifically, the only part that uses the properties of NE there is Lemma 38. We prove a counterpart here.

Lemma 40. *With probability* $1-p$*, for any* $(s, \boldsymbol{a}, h, i)$ *and* $k \in [K]$*:*

$$\overline{Q}^{k}_{h,i}\left(s,\boldsymbol{a}\right) \geq \max_{\phi \in \Phi_i} Q^{\phi \diamond \pi^k}_{h,i}\left(s,\boldsymbol{a}\right), \quad \underline{Q}^{k}_{h,i}\left(s,\boldsymbol{a}\right) \leq Q^{\pi^k}_{h,i}\left(s,\boldsymbol{a}\right), \tag{47}$$

$$\overline{V}^{k}_{h,i}\left(s\right) \geq \max_{\phi \in \Phi_i} V^{\phi \diamond \pi^k}_{h,i}\left(s\right), \quad \underline{V}^{k}_{h,i}\left(s\right) \leq V^{\pi^k}_{h,i}\left(s\right). \tag{48}$$

Proof. For each fixed k, we prove this by induction from $h = H+1$ to $h = 1$. For the base case, we know at the $(H+1)$-th step, $\overline{V}^{k}_{H+1,i}\left(s\right) = \max_{\phi} V^{\phi \diamond \pi^k}_{H+1,i}\left(s\right) = 0$. Now, assume the inequality (40) holds for the $(h+1)$-th step, for the h-th step, by definition of Q-functions,

$$\begin{aligned}
&\overline{Q}^{k}_{h,i}\left(s,\boldsymbol{a}\right) - \max_{\phi} Q^{\phi \diamond \pi^k}_{h,i}\left(s,\boldsymbol{a}\right) \\
=& \left[\widehat{\mathbb{P}}^{k}_{h} \overline{V}^{k}_{h+1,i}\right]\left(s,\boldsymbol{a}\right) - \left[\mathbb{P}_h \max_{\phi} V^{\phi \diamond \pi^k}_{h+1,i}\right]\left(s,\boldsymbol{a}\right) + \beta_t \\
=& \underbrace{\widehat{\mathbb{P}}^{k}_{h}\left(\overline{V}^{k}_{h+1,i} - \max_{\phi} V^{\phi \diamond \pi^k}_{h+1,i}\right)\left(s,\boldsymbol{a}\right)}_{(A)} + \underbrace{\left(\widehat{\mathbb{P}}^{k}_{h} - \mathbb{P}_h\right) \max_{\phi} V^{\phi \diamond \pi^k}_{h+1,i}\left(s,\boldsymbol{a}\right)}_{(B)} + \beta_t.
\end{aligned}$$

By induction hypothesis, for any s', $\left(\overline{V}^{k}_{h+1,i} - \max_{\phi} V^{\phi \diamond \pi^k}_{h+1,i}\right)\left(s'\right) \geq 0$, and thus $(A) \geq 0$. By uniform concentration, $(B) \leq C\sqrt{SH^2\iota / N^k_h(s,\boldsymbol{a})} = \beta_t$. Putting everything together we have $\overline{Q}^{k}_{h,i}\left(s,\boldsymbol{a}\right) - \max_{\phi} Q^{\phi \diamond \pi^k}_{h,i}\left(s,\boldsymbol{a}\right) \geq 0$. The second inequality can be proved similarly.

Now assume inequality (47) holds for the h-th step, by the definition of V-functions and CE,

$$\overline{V}^{k}_{h,i}\left(s\right) = \mathbb{D}_{\pi^k} \overline{Q}^{k}_{h,i}\left(s\right) = \max_{\phi} \mathbb{D}_{\phi \diamond \pi^k} \overline{Q}^{k}_{h,i}\left(s\right).$$

By Bellman equation,

$$\max_{\phi} V^{\phi \diamond \pi^k}_{h,i}\left(s\right) = \max_{\phi} \mathbb{D}_{\phi \diamond \pi^k} \max_{\phi'} Q^{\phi' \diamond \pi^k}_{h,i}\left(s\right).$$

Since by induction hypothesis, for any $(s,\boldsymbol{a})$, $\overline{Q}^{k}_{h,i}\left(s,\boldsymbol{a}\right) \geq \max_{\phi} Q^{\phi \diamond \pi^k}_{h,i}\left(s,\boldsymbol{a}\right)$. As a result, we also have $\overline{V}^{k}_{h,i}\left(s\right) \geq \max_{\phi} V^{\phi \diamond \pi^k}_{h,i}\left(s\right)$, which is exactly inequality (40) for the h-th step. The second inequality can be proved similarly. □

F.2. Proof of Theorem 16

In this section, we prove each theorem for the single reward function case, i.e., $N = 1$. The proof for the case of multiple reward functions ($N > 1$) simply follows from taking a union bound, that is, replacing the failure probability p by Np.

F.2.1. NE VERSION

Let (μ^k, ν^k) be an arbitrary Nash-equilibrium policy of $\widehat{\mathcal{M}}^k := (\widehat{\mathbb{P}}^k, \widehat{r}^k)$, where $\widehat{\mathbb{P}}^k$ and $\widehat{r}^k$ are our empirical estimate of the transition and the reward at the beginning of the k'th episode in Algorithm 4. Given an arbitrary Nash equilibrium π^k of $\widehat{\mathcal{M}}^k$, we use $\widehat{Q}^k_{h,i}$ and $\widehat{V}^k_{h,i}$ to denote its value functions of the i'th player at the h'th step in $\widehat{\mathcal{M}}^k$.

We prove the following two lemmas, which together imply the conclusion about Nash equilibriums in Theorem 16 as in the proof of Theorem 5.

Lemma 41. *With probability* $1-p$*, for any* $(h, s, \boldsymbol{a}, i)$ *and* $k \in [K]$*, we have*

$$\begin{cases}
|\widehat{Q}^k_{h,i}(s,\boldsymbol{a}) - Q^{\pi^k}_{h,i}(s,\boldsymbol{a})| \leq \widetilde{Q}^k_h(s,\boldsymbol{a}), \\
|\widehat{V}^k_{h,i}(s) - V^{\pi^k}_{h,i}(s)| \leq \widetilde{V}^k_h(s).
\end{cases} \tag{49}$$

Proof. For each fixed k, we prove this by induction from $h = H+1$ to $h = 1$. For base case, we know at the $(H+1)$-th step,$\widehat{V}^k_{H+1,i} = V^{\pi^k}_{H+1,i} = \widehat{Q}^k_{H+1,i} = Q^{\pi^k}_{H+1,i} = 0$. Now, assume the conclusion holds for the $(h+1)$'th step, for the h'th step, by definition of Q- functions,

$$\begin{aligned}
&\left|\widehat{Q}^k_{h,i}\left(s,\boldsymbol{a}\right) - Q^{\pi^k}_{h,i}\left(s,\boldsymbol{a}\right)\right| \\
\leq &\left|\left[\widehat{\mathbb{P}}^k_h \widehat{V}^k_{h+1,i}\right]\left(s,\boldsymbol{a}\right) - \left[\mathbb{P}_h V^{\pi^k}_{h+1,i}\right]\left(s,\boldsymbol{a}\right)\right| + \left|r_h(s,a) - \widehat{r}^k_h(s,a)\right| \\
\leq &\underbrace{\left|\widehat{\mathbb{P}}^k_h\left(\widehat{V}^k_{h+1,i} - V^{\pi^k}_{h+1,i}\right)\left(s,\boldsymbol{a}\right)\right|}_{(A)} + \underbrace{\left|\left(\widehat{\mathbb{P}}^k_h - \mathbb{P}_h\right)V^{\pi^k}_{h+1,i}\left(s,\boldsymbol{a}\right)\right| + \left|r_h(s,a) - \widehat{r}^k_h(s,a)\right|}_{(B)}
\end{aligned}$$

By the induction hypothesis,

$$(A) \leq \widehat{\mathbb{P}}^k_h \left|\widehat{V}^k_{h+1,i} - V^{\pi^k}_{h+1,i}\right|\left(s,\boldsymbol{a}\right) \leq (\widehat{\mathbb{P}}^k_h \widetilde{V}^k_{h+1})(s,\boldsymbol{a}).$$

By uniform concentration (e.g., Lemma 12 in Bai & Jin, 2020), $(B) \leq \sqrt{SH^2\iota/N^k_h(s,\boldsymbol{a})} = \beta_t$. Putting everything together we have

$$\left|Q^{\pi^k}_{h,i}\left(s,\boldsymbol{a}\right) - \widehat{Q}^k_{h,i}\left(s,\boldsymbol{a}\right)\right| \leq \min\left\{(\widehat{\mathbb{P}}^k_h \widetilde{V}^k_{h+1})(s,\boldsymbol{a}) + \beta_t, H\right\} = \widetilde{Q}^k_h(s,\boldsymbol{a}),$$

which proves the first inequality in (49). The inequality for V functions follows directly by noting that the value functions are computed using the same policy π^k. □

Lemma 42. *With probability* $1-p$*, for any* $(h, s, \boldsymbol{a}, i, k)$*, we have*

$$\begin{cases}
|\widehat{Q}^k_{h,i}(s,\boldsymbol{a}) - Q^{\pi^k_{-i},\dagger}_{h,i}(s,\boldsymbol{a})| \leq \widetilde{Q}^k_h(s,\boldsymbol{a}), \\
|\widehat{V}^k_{h,i}(s) - V^{\pi^k_{-i},\dagger}_{h,i}(s)| \leq \widetilde{V}^k_h(s).
\end{cases} \tag{50}$$

Proof. For each fixed k, we prove this by induction from $h = H+1$ to $h = 1$. For the base case, we know at the $(H+1)$-th step,$\widehat{V}^k_{H+1,i} = V^{\pi^k_{-i},\dagger}_{H+1,i} = \widehat{Q}^k_{H+1,i} = Q^{\pi^k_{-i},\dagger}_{H+1,i} = 0$. Now, assume the conclusion holds for the $(h+1)$'th step, for the h'th step, by definition of the Q functions,

$$\begin{aligned}
&\left|\widehat{Q}^k_{h,i}\left(s,\boldsymbol{a}\right) - Q^{\pi^k_{-i},\dagger}_{h,i}\left(s,\boldsymbol{a}\right)\right| \\
= &\left|\left[\widehat{\mathbb{P}}^k_h \widehat{V}^k_{h+1,i}\right]\left(s,\boldsymbol{a}\right) - \left[\mathbb{P}_h V^{\pi^k_{-i},\dagger}_{h+1,i}\right]\left(s,\boldsymbol{a}\right)\right| + \left|r_h(s,a) - \widehat{r}^k_h(s,a)\right| \\
\leq &\underbrace{\left|\widehat{\mathbb{P}}^k_h\left(\widehat{V}^k_{h+1,i} - V^{\pi^k_{-i},\dagger}_{h+1,i}\right)\left(s,\boldsymbol{a}\right)\right|}_{(A)} + \underbrace{\left|\left(\widehat{\mathbb{P}}^k_h - \mathbb{P}_h\right)V^{\pi^k_{-i},\dagger}_{h+1,i}\left(s,\boldsymbol{a}\right)\right| + \left|r_h(s,a) - \widehat{r}^k_h(s,a)\right|}_{(B)}
\end{aligned}$$

By the induction hypothesis,

$$(A) \leq \widehat{\mathbb{P}}^k_h \left|\widehat{V}^k_{h+1,i} - V^{\pi^k_{-i},\dagger}_{h+1,i}\right|\left(s,\boldsymbol{a}\right) \leq (\widehat{\mathbb{P}}^k_h \widetilde{V}^k_{h+1})(s,\boldsymbol{a}).$$

By uniform concentration, $(B) \leq \sqrt{SH^2\iota/N^k_h(s,\boldsymbol{a})} = \beta_t$. Putting everything together we have

$$\left|Q^{\pi^k_{-i},\dagger}_{h,i}\left(s,\boldsymbol{a}\right) - \widehat{Q}^k_{h,i}\left(s,\boldsymbol{a}\right)\right| \leq \min\left\{(\widehat{\mathbb{P}}^k_h \widetilde{V}^k_{h+1})(s,\boldsymbol{a}) + \beta_t, H\right\} = \widetilde{Q}^k_h(s,\boldsymbol{a}),$$

which proves the first inequality in (50). It remains to show the inequality for V-functions also hold in the h'th step.

Since π^k is a Nash-equilibrium policy, we have

$$\widehat{V}^k_{h,i}\left(s\right) = \max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} \widehat{Q}^k_{h,i}\left(s\right).$$

By Bellman equation,

$$V^{\pi^k_{-i},\dagger}_{h,i}\left(s\right) = \max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} Q^{\pi^k_{-i},\dagger}_{h,i}\left(s\right).$$

Combining the two equations above, and utilizing the bound we just proved for Q functions, we obtain

$$\left|\widehat{V}^k_{h,i}\left(s\right) - V^{\pi^k_{-i},\dagger}_{h,i}\left(s\right)\right| \leq \left|\max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} \widehat{Q}^k_{h,i}\left(s\right) - \max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} Q^{\pi^k_{-i},\dagger}_{h,i}\left(s\right)\right| \leq \max_{\boldsymbol{a}} \widetilde{Q}^k_h(s,\boldsymbol{a}) = \widetilde{V}^k_h(s),$$

which completes the whole proof. □

F.2.2. CCE VERSION

The proof is almost the same as that for Nash equilibriums. We will reuse Lemma 41 and prove an analogue of Lemma 42. The conclusion for CCEs will follow directly by combining the two lemmas as in the proof of Theorem 5.

Lemma 43. *With probability* $1-p$*, for any* $(h,s,\boldsymbol{a},i)$ *and* $k\in[K]$*, we have*

$$\begin{cases} Q^{\pi^k_{-i},\dagger}_{h,i}(s,\boldsymbol{a}) - \widehat{Q}^k_{h,i}(s,\boldsymbol{a}) \leq \widetilde{Q}^k_h(s,\boldsymbol{a}), \\ V^{\pi^k_{-i},\dagger}_{h,i}(s) - \widehat{V}^k_{h,i}(s) \leq \widetilde{V}^k_h(s). \end{cases} \tag{51}$$

Proof. For each fixed k, we prove this by induction from $h=H+1$ to $h=1$. For base case, we know at the $(H+1)$-th step,$\widehat{V}^k_{H+1,i} = V^{\pi^k_{-i},\dagger}_{H+1,i} = \widehat{Q}^k_{H+1,i} = Q^{\pi^k_{-i},\dagger}_{H+1,i} = 0$. Now, assume the conclusion holds for the $(h+1)$'th step, for the h'th step, by definition of Q -functions,

$$\begin{aligned} & Q^{\pi^k_{-i},\dagger}_{h,i}\left(s,\boldsymbol{a}\right) - \widehat{Q}^k_{h,i}\left(s,\boldsymbol{a}\right) \\ \leq & \left[\mathbb{P}_h V^{\pi^k_{-i},\dagger}_{h+1,i}\right]\left(s,\boldsymbol{a}\right) - \left[\widehat{\mathbb{P}}^k_h \widehat{V}^k_{h+1,i}\right]\left(s,\boldsymbol{a}\right) + \left|r_h(s,a) - \widehat{r}^k_h(s,a)\right| \\ \leq & \underbrace{\widehat{\mathbb{P}}^k_h\left(V^{\pi^k_{-i},\dagger}_{h+1,i} - \widehat{V}^k_{h+1,i}\right)\left(s,\boldsymbol{a}\right)}_{(A)} + \underbrace{\left(\mathbb{P}_h - \widehat{\mathbb{P}}^k_h\right)V^{\pi^k_{-i},\dagger}_{h+1,i}\left(s,\boldsymbol{a}\right) + \left|r_h(s,a) - \widehat{r}^k_h(s,a)\right|}_{(B)}. \end{aligned}$$

By the induction hypothesis, $(A) \leq (\widehat{\mathbb{P}}^k_h \widetilde{V}^k_{h+1})(s,\boldsymbol{a})$.

By uniform concentration, $(B) \leq \sqrt{SH^2\iota/N^k_h(s,\boldsymbol{a})} = \beta_t$. Putting everything together we have

$$Q^{\pi^k_{-i},\dagger}_{h,i}\left(s,\boldsymbol{a}\right) - \widehat{Q}^k_{h,i}\left(s,\boldsymbol{a}\right) \leq \min\left\{(\widehat{\mathbb{P}}^k_h \widetilde{V}^k_{h+1})(s,\boldsymbol{a}) + \beta_t, H\right\} = \widetilde{Q}^k_h(s,\boldsymbol{a}),$$

which proves the first inequality in (51). It remains to show the inequality for V-functions also hold in the h'th step.

Since π^k is a CCE, we have

$$\widehat{V}^k_{h,i}\left(s\right) \geq \max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} \widehat{Q}^k_{h,i}\left(s\right).$$

Observe that $V^{\pi^k_{-i},\dagger}_{h,i}$ obeys the Bellman optimality equation, so we have

$$V^{\pi^k_{-i},\dagger}_{h,i}\left(s\right) = \max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} Q^{\pi^k_{-i},\dagger}_{h,i}\left(s\right).$$

Combining the two equations above, and utilizing the bound we just proved for Q-functions, we obtain

$$V^{\pi^k_{-i},\dagger}_{h,i}\left(s\right) - \widehat{V}^k_{h,i}\left(s\right) \leq \max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} Q^{\pi^k_{-i},\dagger}_{h,i}\left(s\right) - \max_{\mu} \mathbb{D}_{\mu\times\pi^k_{-i}} \widehat{Q}^k_{h,i}\left(s\right) \leq \max_{\boldsymbol{a}} \widetilde{Q}^k_h(s,\boldsymbol{a}) = \widetilde{V}^k_h(s),$$

which completes the whole proof. □

F.2.3. CE VERSION

The proof is almost the same as that for Nash equilibriums. We will reuse Lemma 41 and prove an analogue of Lemma 42. The conclusion for CEs will follow directly by combining the two lemmas as in the proof of Theorem 5.

Lemma 44. *With probability* $1-p$, *for any* $(h, s, \boldsymbol{a}, i)$, $k \in [K]$ *and strategy modification* ϕ *for player* i, *we have*

$$\begin{cases} Q_{h,i}^{\phi\diamond\pi^k}(s,\boldsymbol{a}) - \widehat{Q}_{h,i}^k(s,\boldsymbol{a}) \le \widetilde{Q}_h^k(s,\boldsymbol{a}), \\ V_{h,i}^{\phi\diamond\pi^k}(s) - \widehat{V}_{h,i}^k(s) \le \widetilde{V}_h^k(s). \end{cases} \tag{52}$$

Proof. For each fixed k, we prove this by induction from $h = H+1$ to $h = 1$. For the base case, we know at the $(H+1)$-th step, $\widehat{V}_{H+1,i}^k = V_{H+1,i}^{\phi\diamond\pi^k} = \widehat{Q}_{H+1,i}^k = Q_{H+1,i}^{\phi\diamond\pi^k} = 0$. Now, assume the conclusion holds for the $(h+1)$'th step, for the h'th step, following exactly the same argument as Lemma 43, we can show

$$Q_{h,i}^{\phi\diamond\pi^k}(s,\boldsymbol{a}) - \widehat{Q}_{h,i}^k(s,\boldsymbol{a}) \le \min\left\{(\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s,\boldsymbol{a}) + \beta_t, H\right\} = \widetilde{Q}_h^k(s,\boldsymbol{a}),$$

which proves the first inequality in (52). It remains to show the inequality for V-functions also hold in the h'th step.

Since π^k is a CE, we have

$$\widehat{V}_{h,i}^k(s) = \max_{\tilde{\phi}_{h,s}} \mathbb{D}_{\tilde{\phi}_{h,s}\diamond\pi^k} \widehat{Q}_{h,i}^k(s),$$

where the maximum is take over all possible functions from $\mathcal{A}_i$ to itself.

Observe that $V_{h,i}^{\phi\diamond\pi^k}$ obeys the Bellman optimality equation, so we have

$$V_{h,i}^{\phi\diamond\pi^k}(s) = \max_{\tilde{\phi}_{h,s}} \mathbb{D}_{\tilde{\phi}_{h,s}\diamond\pi^k} Q_{h,i}^{\phi\diamond\pi^k}(s).$$

Combining the two equations above, and utilizing the bound we just proved for Q-functions, we obtain

$$\begin{aligned} V_{h,i}^{\phi\diamond\pi^k}(s) - \widehat{V}_{h,i}^k(s) &= \max_{\tilde{\phi}_{h,s}} \mathbb{D}_{\tilde{\phi}_{h,s}\diamond\pi^k} Q_{h,i}^{\phi\diamond\pi^k}(s) - \max_{\tilde{\phi}_{h,s}} \mathbb{D}_{\tilde{\phi}_{h,s}\diamond\pi^k} \widehat{Q}_{h,i}^k(s) \\ &\le \max_{\boldsymbol{a}} \widetilde{Q}_h^k(s,\boldsymbol{a}) = \widetilde{V}_h^k(s), \end{aligned}$$

which completes the whole proof. □