Appendix

Table of Contents

A Policy-Induced Joint Measure ................................................. 14
B Properties of Modification ..................................................... 15
C Proof of Theorem 1 .................................................................. 17
   C.1 Proof of item 1 for unconstrained Markov game ................... 17
   C.2 Proof of item 2 for constrained Markov game ..................... 18
D Proof of Theorem 2 .................................................................. 19
E The Range of the Optimal Dual Variable ................................... 20
F Proof of Theorem 3 .................................................................. 20
G Proof of Corollary 5.1 .............................................................. 23

A Policy-Induced Joint Measure

We introduce policy-induced joint measure as a useful tool to help prove Theorems 1 and 2. Specifically, for any \( h, h' \), we define \( p_\pi(s_{1:h}, a_{1:h'}) \) as the joint distribution of \( (s_{1:h}, a_{1:h'}) \) induced by policy \( \pi \). Based on the underlying Markov decision process, it can be easily verified that \( p_\pi \) takes the following forms.

\[
p_\pi(s_1) = \rho(s_1),
\]

\[
p_\pi(s_{1:h}, a_{1:h-1}) = p_\pi(s_{1:h-1}, a_{1:h-1}) P(s_h|s_{h-1}, a_{h-1}),
\]

\[
p_\pi(s_{1:h}, a_{1:h}) = \rho(s_1) \pi_1(a_1|s_1) \prod_{h'=2}^h P(s_{h'}|s_{h'-1}, a_{h'-1}) \pi_{h'}(a_{h'}|s_{1:h'}, a_{1:h'-1}).
\]

Eqs. (17) and (18) further imply that

\[
p_\pi(s_{1:h}, a_{1:h}) = p_\pi(s_{1:h}, a_{1:h-1}) \pi(a_h|s_{1:h}, a_{1:h-1}).
\]

Conversely, given a proper joint probability measure \( p \), we can infer its corresponding inducing policy \( \pi \) as follows.

**Lemma A.1.** Consider any joint probability measure \( p \) that satisfies eqs. (16) and (17) (replace all \( p_\pi \) with \( p \)). Then, the following policy \( \pi \) induces \( p \).

\[
\pi_h(a_h|s_{1:h}, a_{1:h-1}) = \begin{cases} 
p(h_{1:h}, a_{1:h}) & \text{if } p(s_{1:h}, a_{1:h-1}) > 0 \\
\frac{p(s_{1:h}, a_{1:h-1})}{p(s_{1:h}, a_{1:h-1})} & \text{if } p(s_{1:h}, a_{1:h-1}) = 0 
\end{cases}
\]

**Proof.** It suffices to prove that \( p_\pi(s_{1:h}, a_{1:h}) = p(s_{1:h}, a_{1:h}) \) for any \( s_{1:h}, a_{1:h} \), i.e., \( p \) is exactly the joint measure \( p_\pi \) induced by \( \pi \). We consider the following two cases.

**Case 1:** \( p(s_{1:h}, a_{1:h-1}) > 0 \). In this case, we must have that \( p(s_{1:h'}, a_{1:h'-1}) > 0 \) for any \( 1 \leq h' \leq h \). Therefore, by eq. (20) we have \( \pi_{h'}(a_{h'}|s_{1:h'}, a_{1:h'-1}) = \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'}, a_{1:h'-1})} \) for any \( 1 \leq h' \leq h \).

Substitute this policy \( \pi \) into eq. (18) and note that \( \rho(s_1) = p(s_1) \), we obtain that

\[
p_\pi(s_{1:h}, a_{1:h}) = p(s_1) \frac{h}{p(s_1)} \prod_{h'=2}^h P(s_{h'}|s_{h'-1}, a_{h'-1}) \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'}, a_{1:h'-1})}
\]
We note that where (i) uses eq. (17). Thus we conclude that (Case 2.2) If (Case 2.1) If 

Thus, both the value function viewed as the policy following linear combination of these two joint measures.

Next, let as follows: we first generate joint action policy in this section, we present some useful properties of the modification operator. Recall that for any , respectively, for any , as well. We further consider the following two subcases.

(Case 2.1) If , then , 0 by substituting , 0 into eq. (18).

(Case 2.2) If , 0 because Case 2 assumes that , 0, then there must exist such that , 0 and , 0. On the other hand, note that eq. (18) implies that contains the following multiplicative factor

\[
p_{\pi}(s_{1:h}, a_{1:h}) \propto \pi(h_{1:h} | s_{1:h}) \prod_{h'=2}^{h} \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'-1}, a_{1:h'-1})} = p(s_{1:h}, a_{1:h}),
\]

where (i) follows from eq. (17) (replace all with ).

\[
\text{Case 2: } p(s_{1:h}, a_{1:h-1}) = 0. \text{ In this case, we have that } p(s_{1:h}, a_{1:h}) = 0. \text{ Hence, it suffices to prove that } p_{\pi}(s_{1:h}, a_{1:h}) = 0 \text{ as well. We further consider the following two subcases.}
\]

\[
\text{(Case 2.1) If } p(s_1) = \rho(s_1) = 0, \text{ then } p_{\pi}(s_{1:h}, a_{1:h}) = 0 \text{ by substituting } \rho(s_1) = 0 \text{ into eq. (18).}
\]

\[
\text{(Case 2.2) If } p(s_1) = \rho(s_1) > 0 \text{ and because Case 2 assumes that } p(s_{1:h}, a_{1:h-1}) = 0, \text{ then there must exist } 1 < h' < h - 1 \text{ such that } p(s_{1:h'}, a_{1:h'-1}) > 0 \text{ and } p(s_{1:h'+1}, a_{1:h'}) = 0. \text{ On the other hand, note that eq. (18) implies that } p_{\pi}(s_{1:h}, a_{1:h}) \propto \pi(h_{1:h} | s_{1:h}) \prod_{h'=2}^{h} \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'-1}, a_{1:h'-1})} = 0
\]

\[\pi(h_{1:h} | s_{1:h}) \propto \pi(h_{1:h} | s_{1:h}) \prod_{h'=2}^{h} \frac{p(s_{1:h'}, a_{1:h'})}{p(s_{1:h'-1}, a_{1:h'-1})} = 0 \]

where (i) uses eq. (17). Thus we conclude that .

With the policy-induced joint measure , we can rewrite the value function

\[
V_{j}^{(m)}(\pi) := \mathbb{E}_{\pi} \left[ \sum_{h=1}^{H} \gamma_{i,j}^{(m)} | s_1 \sim \rho \right] \text{ and the Lagrangian function eq. 6} \text{ as follows.}
\]

\[
V_{j}^{(m)}(\pi) = \tilde{V}_{j}^{(m)}(\pi_{\pi}) := \sum_{s_{1:H}, a_{1:H}} p_{\pi}(s_{1:H}, a_{1:H}) \sum_{h=1}^{H} \gamma_{j,h}^{(m)} (s_{h}, a_{h}),
\]

\[
L^{(m)}(\pi, \lambda^{(m)}) = \tilde{L}^{(m)}(\pi_{\pi}, \lambda^{(m)}) := \tilde{V}_{0}^{(m)}(\pi_{\pi}) + \sum_{j=1}^{d_{m}} \lambda_{j} (\tilde{V}_{j}^{(m)}(\pi_{\pi}) - c_{j}^{(m)}).
\]

Thus, both the value function and the Lagrangian function can be rewritten as linear functions of . Such a linear form helps simplify the problem and prove the key Theorems 1 and 2.

**B Properties of Modification**

In this section, we present some useful properties of the modification operator. Recall that for any policy and any modification operator , the modified policy is defined as follows: we first generate joint action from . Then, randomly modifies to . To summarize, the modified policy takes the following form.

\[
(\phi_{h}^{(m)} \circ \pi_{h})(a_{h}^{(m)} | s_{1:h}, a_{1:h-1}) = \sum_{a_{h}^{(m)}} \phi_{h}^{(m)}(a_{h}^{(m)} | s_{1:h}, a_{1:h-1}, a_{h}^{(m)}) \pi_{h}(a_{h}^{(m)} | s_{1:h}, a_{1:h-1}).
\]

Next, let be the joint measures induced by the modified policies and , respectively, for any , . In the proof of Theorems 1 and 2, we introduce the following linear combination of these two joint measures.

\[
p_{\lambda} := \lambda p_{\phi^{(m)\circ\pi}} + (1 - \lambda)p_{\bar{\phi}^{(m)\circ\pi}}, \lambda \in \mathbb{R}.
\]

We note that for any , . Hence, is a proper joint measure.

We note that for any , . This case, since the joint measures and satisfy eqs. (16) and (17) by definition, it is easy to verify that also satisfies eqs. (16) and (17) and hence is a proper joint measure. Therefore, by Lemma A.1 we can find its inducing policy using eq. (20). Next, we show that such an inducing policy can actually be viewed as the policy modified by a certain stochastic modification.
Lemma B.1. Regarding the $p_\lambda$ defined in eq. \ref{eq:25}, if $\lambda \in \mathbb{R}$ is selected such that the following two conditions hold for any $s_{1:h}, a_{1:h-1}, a_{h}^{(m)}$, $\overline{a}_h^{(m)}$:

\begin{equation}
 p_\lambda(s_{1:h}, a_{1:h-1}) \geq 0
\end{equation}

\begin{equation}
 \lambda p_\phi^{(m)}(s_{1:h}, a_{1:h-1}) \phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})
 + (1 - \lambda) p_{\overline{\phi}^{(m)}}(s_{1:h}, a_{1:h-1}) \overline{\phi}_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \geq 0,
\end{equation}

then its inducing policy (defined by eq. \ref{eq:20}) can be written as $\pi_\lambda = \phi_h^{(m)}(\cdot|s_{1:h}, a_{1:h-1}, a_h^{(m)})$, where the stochastic modification $\phi_h^{(m)}$ takes the following form.

\begin{equation}
 \phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) = \begin{cases} 
 \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})} & \text{if } p_\lambda(s_{1:h}, a_{1:h-1}) > 0 \\
 (1 - \lambda) p_{\overline{\phi}^{(m)}}(s_{1:h}, a_{1:h-1}) \overline{\phi}_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) 
 + (1 - \lambda) p_\phi^{(m)}(s_{1:h}, a_{1:h-1}) \phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) 
 & \text{if } p_\lambda(s_{1:h}, a_{1:h-1}) = 0
\end{cases}
\end{equation}

Proof. We first show that $\phi_h^{(m)}(\cdot|s_{1:h}, a_{1:h-1}, a_h^{(m)})$ is a proper stochastic modification. By eq. \ref{eq:28}, we only need to consider the case $p_\lambda(s_{1:h}, a_{1:h-1}) > 0$. In this case, based on the condition in eq. \ref{eq:27}, we conclude that $\phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \geq 0$. In addition,

\begin{align*}
 \sum_{\overline{a}_h^{(m)} \in A^{(m)}} \phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) 
 = & \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})} \left( \lambda p_\phi^{(m)}(s_{1:h}, a_{1:h-1}) \sum_{\overline{a}_h^{(m)} \in A^{(m)}} \phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) 
 + (1 - \lambda) p_{\overline{\phi}^{(m)}}(s_{1:h}, a_{1:h-1}) \sum_{\overline{a}_h^{(m)} \in A^{(m)}} \overline{\phi}_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \right) \\
 = & \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})} \left( \lambda p_\phi^{(m)}(s_{1:h}, a_{1:h-1}) + (1 - \lambda) p_{\overline{\phi}^{(m)}}(s_{1:h}, a_{1:h-1}) \right) = 1.
\end{align*}

Therefore, $\phi_h^{(m)}$ is a proper stochastic modification.

Next, we prove that the policy that induces $p_\lambda$ takes the form $\pi_\lambda = \phi_h^{(m)}(\cdot|s_{1:h}, a_{1:h-1}, a_h^{(m)})$. We consider two cases.

Case 1: $p_\lambda(s_{1:h}, a_{1:h-1}) > 0$. In this case, we obtain that

\begin{align*}
 \pi_\lambda(\overline{a}_h|s_{1:h}, a_{1:h-1}) 
 = & \left( i \right) \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})} \left( \lambda p_\phi^{(m)}(s_{1:h}, a_{1:h-1}) + (1 - \lambda) p_{\overline{\phi}^{(m)}}(s_{1:h}, a_{1:h-1}) \right) \\
 = & \left( i \right) \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})} \left( \lambda p_\phi^{(m)}(s_{1:h}, a_{1:h-1}) \phi_h^{(m)}(\overline{a}_h|s_{1:h}, a_{1:h-1}) 
 + (1 - \lambda) p_{\overline{\phi}^{(m)}}(s_{1:h}, a_{1:h-1}) \overline{\phi}_h^{(m)}(\overline{a}_h|s_{1:h}, a_{1:h-1}) \right) \\
 = & \left( i \right) \frac{1}{p_\lambda(s_{1:h}, a_{1:h-1})} \sum_{\overline{a}_h^{(m)} \in A^{(m)}} \left( \lambda p_\phi^{(m)}(s_{1:h}, a_{1:h-1}) \phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) 
 + (1 - \lambda) p_{\overline{\phi}^{(m)}}(s_{1:h}, a_{1:h-1}) \overline{\phi}_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) \right) \\
 = & \left( i \right) \sum_{\overline{a}_h^{(m)} \in A^{(m)}} \pi_\lambda(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}) \phi_h^{(m)}(\overline{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)})
\end{align*}
where (i) uses eqs. (20) and (25), (ii) uses eq. (19), (iii) uses eq. (24), and (iv) uses eq. (28). This proves the claim due to eq. (24).

**Case 2:** \( p_\lambda(s_{1:h}, a_{1:h-1}) = 0 \). In this case, both \( \phi^{(m)}_{\lambda,h} \) and \( \pi_\lambda \) can be arbitrarily defined. Hence, we can simply define \( \pi_\lambda \) by eq. (24).

**C. Proof of Theorem 1**

**C.1 Proof of item 1 for unconstrained Markov game**

Throughout the proof, for any policy \( \pi \), we denote \( \tilde{\phi}^{(m)} \) as the *optimal* stochastic modification associated with \( \pi \), i.e., \( V_0^{(m)}(\tilde{\phi}^{(m)} \circ \pi) \) achieves the maximum value over all stochastic modifications. In order for \( \pi \) to be a CE, it must satisfy \( V_0^{(m)}(\pi) \geq V_0^{(m)}(\tilde{\phi}^{(m)} \circ \pi) \).

If for any optimal stochastic modification \( \tilde{\phi}^{(m)} \) associated with \( \pi \), we can construct a corresponding deterministic modification \( \phi^{(m)} \) such that \( V_0^{(m)}(\phi^{(m)} \circ \pi) = V_0^{(m)}(\tilde{\phi}^{(m)} \circ \pi) \), then the condition of item 1 guarantees that \( \pi \) is a CE and then item 1 is proved.

Next, for any policy \( \pi \) and any associated optimal stochastic modification \( \tilde{\phi}^{(m)} \), we construct a deterministic modification \( \phi^{(m)} \) as follows: for any \( s_{1:h}, a_{1:h-1}, a_h \), select an arbitrary \( \tilde{a}_h^{(m)} \) such that \( \tilde{\phi}^{(m)}(\tilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) > 0 \) (this always exists) and then simply define \( \phi^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, a_h^{(m)}) = 1 \), and 0 otherwise. It suffices to prove that \( V_0^{(m)}(\phi^{(m)} \circ \pi) = V_0^{(m)}(\tilde{\phi}^{(m)} \circ \pi) \) for any \( \pi \) satisfying the condition of item 1.

To proceed, we claim that one can find \( \lambda < 0 \) such that the joint measure \( p_\lambda := \lambda p_\phi^{(m)} \circ \pi + (1 - \lambda)p_{\tilde{\phi}^{(m)}} \circ \pi \) satisfies eqs. (26) and (27). We will prove the validity of this claim later. Suppose this claim holds. Then based on Lemma B.1, the inducing policy of \( p_\lambda \) takes the form \( \pi_\lambda = \phi^{(m)}_\lambda \circ \pi \), where \( \phi^{(m)}_\lambda \) is defined by eq. (28). Then, we obtain that

\[
V_0^{(m)}(\phi^{(m)} \circ \pi) \overset{(i)}{=} V_0^{(m)}(\tilde{\phi}^{(m)} \circ \pi) \overset{(ii)}{=} V_0^{(m)}(\pi) \overset{(iii)}{=} V_0^{(m)}(\lambda p_\phi^{(m)} \circ \pi + (1 - \lambda)\tilde{\phi}^{(m)} \circ \pi) \overset{(iv)}{=} V_0^{(m)}(\phi^{(m)} \circ \pi) + (1 - \lambda)\tilde{V}_0^{(m)}(p_{\tilde{\phi}^{(m)}} \circ \pi),
\]

where (i) uses the optimality of \( \tilde{\phi}^{(m)} \), (ii)-(iv) use the linear form of \( \tilde{V}_0^{(m)}(p_{\tilde{\phi}^{(m)}} \circ \pi) \) defined in eq. (22).

The above inequality along with \( \lambda < 0 \) implies that \( \lambda V_0^{(m)}(\phi^{(m)} \circ \pi) \geq \tilde{V}_0^{(m)}(\phi^{(m)} \circ \pi) \). On the other hand, \( V_0^{(m)}(\phi^{(m)} \circ \pi) \geq V_0^{(m)}(\tilde{\phi}^{(m)} \circ \pi) \) based on the optimality of the stochastic modification \( \tilde{\phi}^{(m)} \). Hence, \( V_0^{(m)}(\phi^{(m)} \circ \pi) = V_0^{(m)}(\tilde{\phi}^{(m)} \circ \pi) \) as desired. All left is to find \( \lambda < 0 \) such that the joint measure \( p_\lambda \) satisfies eqs. (26) and (27). We prove them as follows.

**Proof of eq. (26):** Recall that \( p_\lambda = \lambda p_\phi^{(m)} + (1 - \lambda)p_{\tilde{\phi}^{(m)}} \circ \pi \) and \( \lambda < 0 \). If \( p_{\phi^{(m)}}(s_{1:h}, a_{1:h-1}) \neq 0 \), it is clear that eq. (26) holds. So we just need to consider the other case where \( p_{\phi^{(m)}}(s_{1:h}, a_{1:h-1}) = 0 \). In this case and by eq. (18), we must have that \( \rho(s_1), \phi^{(m)}(\pi_1)(a_1|s_1), \mathcal{P}(s_{1:h}, a_{1:h-1}, \phi^{(m)}(\pi_1)|a_h|s_{1:h}, a_{1:h-1}) > 0 \) for any \( h = 2, \ldots, H \). Then, eq. (24) implies that for any \( h = 2, \ldots, H \),

\[
0 < \phi^{(m)}(a_h|m)|s_{1:h}, a_{1:h-1}) \mathcal{P}(a_h|m)|s_{1:h}, a_{1:h-1}) = \sum_{\tilde{a}_h^{(m)}} \phi^{(m)}(a_h|m)|s_{1:h}, a_{1:h-1}) \mathcal{P}(a_h|m)|s_{1:h}, a_{1:h-1}) > 0.
\]

Hence, there must exist \( \tilde{a}_h^{(m)} \) such that \( \phi^{(m)}(a_h|m)|s_{1:h}, a_{1:h-1}) \mathcal{P}(a_h|m)|s_{1:h}, a_{1:h-1}) > 0 \). As \( \phi^{(m)}_\lambda \) is the deterministic modification constructed at the beginning of this proof, we must have
\( \phi_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \phi_h^{(m)}) = 1 \) and therefore the corresponding stochastic modification satisfies \( \widetilde{\phi}_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \phi_h^{(m)}) > 0 \). Then, eq. (24) implies that
\[
(\phi_h^{(m)} \circ \pi_h)(a_h|s_{1:h}, a_{1:h-1}) = \sum_{\widetilde{a}_h^{(m)}} \widetilde{\phi}_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \pi_h([a_h^{(m)}, a_h^{(\text{\top})}])|s_{1:h}, a_{1:h-1}) \geq \widetilde{\phi}_h^{(m)}(a_h^{(m)}|s_{1:h}, a_{1:h-1}, \pi_h([a_h^{(m)}, a_h^{(\text{\top})}])|s_{1:h}, a_{1:h-1}) > 0.
\]
Similarly, we can prove that \( (\phi_1^{(m)} \circ \pi_1)(a_1|s_1) > 0 \) from \( (\phi_1^{(m)} \circ \pi_1)(a_1|s_1) > 0 \). Therefore, based on eq. (18), it is proved that whenever \( p_{\phi(m) \circ \pi}^{(m)}(s_{1:h}, a_{1:h-1}) > 0 \), we have
\[
p_{\widetilde{\phi}(m) \circ \pi}^{(m)}(s_1, a_1) = \rho(s_1)(\phi_1^{(m)} \circ \pi_1)(a_1|s_1) \prod_{h'=2}^h P(s_{h'}|s_{h'-1}, a_{h'-1})(\widetilde{\phi}_h^{(m)} \circ \pi_h')(a_{h'}|s_{1:h'}, a_{1:h'-1}) > 0. \tag{29}
\]
Therefore, eq. (26) holds for
\[
0 > \lambda \geq \frac{p_{\widetilde{\phi}(m) \circ \pi}^{(m)}(s_1, a_1)}{p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1)} - p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1) := w(s_1, a_1),
\]
for any \( s_1, a_1 \); whenever \( p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1) > p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1) \), which implies that \( p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1) = 0 \) and therefore \( p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1) = 0 \) based on eq. (29). Thus, we conclude that \( w(s_1, a_1) < 0 \). Consider the finite (and possibly empty) set \( A_1 := \{w(s_1, a_1) : 1 \leq h \leq H, p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1) > p_{\phi(m) \circ \pi}^{(m)}(s_1, a_1)\} \). If it is non-empty, eq. (26) holds for all \( 0 > \lambda \geq \max A_1 \); for constant \( \lambda \); otherwise, eq. (26) holds for all \( \lambda < 0 \).

**Proof of eq. (27):** If \( p_{\phi(m) \circ \pi}^{(m)}(s_{1:h}, a_{1:h-1}) \phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, \phi_h^{(m)}) = 0 \) and \( \lambda < 0 \), then eq. (27) holds. Consider the case \( \phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, \phi_h^{(m)}) > 0 \) and \( \lambda < 0 \), and in this case, we have \( p_{\phi(m) \circ \pi}^{(m)}(s_{1:h}, a_{1:h-1}) > 0 \) as proved in the proof of eq. (26), and we also have \( \phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, \phi_h^{(m)}) > 0 \) and thus \( \phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, \phi_h^{(m)}) > 0 \) based on the construction of \( \phi_h^{(m)}(\widetilde{a}_h^{(m)}|s_{1:h}, a_{1:h-1}, \phi_h^{(m)}) > 0 \). Following the same proof logic as that of eq. (26), we can find a constant \( A_2 < 0 \) such that eq. (27) holds for all \( \lambda > -A_2 \).

In summary, we have proved that there exists \( \lambda < 0 \) that guarantees both eqs. (26) and (27).

### C.2 Proof of item 2 for constrained Markov game

Here we construct a counter example to prove item 2. Consider a constrained Markov game with only one state \( S = \{s\} \), two agents with action spaces \( A_1 = A_2 = \{0, 1\} \) and horizon \( H = 1 \). For simplicity, we drop the time step index \( h \) and state \( s \) in all notations throughout this example. Specifically, we denote \( \pi(a^{(1)}, a^{(2)}), \pi^{(m)}(a^{(m)}(\pi)), \phi_h^{(m)}(\pi), m = 1, 2 \), as the joint policy, marginal policy and stochastic modification, respectively.

For both agents \( m = 1, 2 \), we define rewards \( r_0^{(m)} = a^{(m)}, r_1^{(m)} = a^{(0)} + a^{(1)}, r_2^{(m)} = 2 - a^{(0)} - a^{(1)} \) and constraint thresholds \( c_1^{(m)} = c_2^{(m)} = 0.6 \). Therefore, \( V_0^{(m)}(\pi) = \pi r_0^{(m)}(\pi) = \pi^{(m)}(1) \), \( V_1^{(m)}(\pi) = \pi r_1^{(m)}(\pi) = \pi^{(1)}(1) + \pi^{(2)}(1) \) and \( V_2^{(m)}(\pi) = \pi r_2^{(m)}(\pi) = 2 - \pi^{(1)}(1) - \pi^{(2)}(1) = \pi^{(1)}(0) + \pi^{(2)}(0) \). Therefore, for both agents \( m = 1, 2 \), their value function constraints \( V_1^{(m)}(\pi) \geq 0.6, V_2^{(m)}(\pi) \geq 0.6 \) are equivalent to the following condition
\[
0.6 \leq \pi^{(1)}(1) + \pi^{(2)}(1) \leq 1.4. \tag{30}
\]

Now consider a uniform policy \( \pi \) where \( \pi(a^{(1)}, a^{(2)}) = 0.25 \) for all \( a^{(1)}, a^{(2)} \in \{0, 1\} \). This is a product policy which generates independent uniformly distributed actions \( a^{(1)}, a^{(2)} \) with \( \pi^{(1)}(1) = \pi^{(2)}(1) = 0.5 \) that satisfy the constraints in eq. (30). Note that \( A_1 \) only includes two actions. Hence, the set of all possible deterministic modifications \( \phi^{(1)} \) includes the following three cases.
(i) \(\phi^{(1)} \circ \pi = \pi\): either \(\phi^{(1)}\) modifies any \(a^{(1)}\) to \(a^{(1)}\) or modifies any \(a^{(1)}\) to \(1 - a^{(1)}\); 
(ii) \(\phi^{(1)} \circ \pi = \pi'\) that always generates \(a^{(1)} = 0\) and generates \(a^{(2)}\) uniformly at random: \(\phi^{(1)}\) modifies any \(a^{(1)}\) to \(0\); 
(iii) \(\phi^{(1)} \circ \pi = \pi''\) that always generates \(a^{(1)} = 1\) and generates \(a^{(2)}\) uniformly at random: \(\phi^{(1)}\) modifies any \(a^{(1)}\) to \(1\).

However, \(\pi'\) and \(\pi''\) do not satisfy the constraint (30) since \(\pi'^{(1)}(1) + \pi''^{(2)}(1) = 0.5\) and \(\pi''^{(1)}(1) + \pi''^{(2)}(1) = 1.5\). Hence, the only feasible deterministic modifications \(\phi^{(1)}\) are the two ones in (i) with \(\phi^{(1)} \circ \pi = \pi\), which implies that \(V^{(1)}_{0}(\phi^{(1)} \circ \pi) = V^{(1)}_{0}(\pi) = \pi^{(1)}(1) = 0.5\). Therefore, such a \(\pi\) satisfies the assumption of item 2.

Now consider a stochastic modification \(\phi^{(1)}\) defined by \(\phi^{(1)}(1|a_1) = 0.9\) and \(\phi^{(1)}(0|a_1) = 0.1\) for \(a_1 \in \{0, 1\}\). Then \(\phi^{(1)} \circ \pi\) independently generates Bernoulli distributed actions \(a^{(1)} \sim \text{Bern}(0.9)\) and \(a^{(2)} \sim \text{Bern}(0.5)\). Hence, \((\phi^{(1)} \circ \pi)^{(1)}(1) + (\phi^{(1)} \circ \pi)^{(2)}(1) = 1.4\), which means \(\phi^{(1)}\) is feasible based on eq. (30). In addition, \(V^{(1)}_{0}(\phi^{(1)} \circ \pi) = (\phi^{(1)} \circ \pi)^{(1)}(1) = 0.9\), which is strictly larger than \(V^{(1)}_{0}(\pi) = 0.5\). Therefore, \(\pi\) is not a CE as defined in Definition 3.2.

D Proof of Theorem 2

For any policy \(\pi\) and its associated joint measure \(p_{\pi}\), recall the following equivalent Lagrangian functions defined in eq. (23).

\[
L^{(m)}(\pi, \lambda^{(m)}) = \widetilde{L}^{(m)}(p_{\pi}, \lambda^{(m)}).
\]

Then, the desired strong duality result shown in eq. (7) is equivalent to the following equation.

\[
\max_{p \in \mathcal{X}} \min_{\lambda^{(m)} \in \mathbb{R}^{d_{m}^+}} \widetilde{L}^{(m)}(p, \lambda^{(m)}) = \min_{\lambda^{(m)} \in \mathbb{R}^{d_{m}^+}} \max_{p \in \mathcal{X}} \widetilde{L}^{(m)}(p, \lambda^{(m)}),
\]

where the set \(\mathcal{X} := \{p_{\phi^{(m)} \circ \pi} : \phi^{(m)}\) is a stochastic modification\} is defined for the fixed \(\pi\). The nice property of the Lagrangian function \(\widetilde{L}^{(m)}(p, \lambda^{(m)})\) is that it is a linear function in \(p\), which has an advantage toward establishing strong duality.

Based on the minimax theorem (Lemma 9.2 of [2]), it suffices to prove the following properties:

(I). \(\widetilde{L}^{(m)}(p, \cdot)\) is convex and lower semi-continuous, and \(\widetilde{L}^{(m)}(\cdot, p)\) is concave. These properties directly follow from the definition of \(\widetilde{L}\) in eq. (23).

(II). \(\mathbb{R}^{d_{m}^+}\) is a convex set, which holds obviously.

(III). \(\mathcal{X}\) is a convex set, which follows from Lemma B.1 since eqs. (26) and (27) always hold for \(\lambda \in [0, 1]\).

(IV). \(\mathcal{X}\) is a compact set.

Hence, it remains to prove (IV).

As the state space \(S\), action space \(A\) and the horizon \(H\) are finite, we can represent \(p_{\pi}\) as a vector with entries \(p_{\pi}(s^{1:1}, a^{1:H})\) for every \(s^{1:1}, a^{1:H} \in S^H \times A^H\). Hence, the set \(\mathcal{X} \subset [0, 1]^{(|S||A|^H)}\) is bounded. Then, it suffices to prove that \(\mathcal{X}\) is a closed set, i.e., \(p \in \mathcal{X}\) if \(p_{\phi^{(m)} \circ \pi}(s^{1:1}, a^{1:H})\) \(\xrightarrow{k}\) \(p(s^{1:1}, a^{1:H})\), \(\forall s^{1:1}, a^{1:H}\), for some \(p_{\phi^{(m)} \circ \pi} \in \mathcal{X}\) (Note that the notation \(\phi^{(m)}_{[k]}\) indexed by \(k\) differs from \(\phi^{(m)}_{[h]}\) where \(h\) denotes time step).

Similar to \(\mathcal{X}\), any stochastic modification \(\phi^{(m)}\) can also be seen as a bounded finite-dimensional vector with entries \(\phi^{(m)}(\alpha^{(m)}_{s^{1:1}, a^{1:H-1}, a^{m}_{h}}) \in [0, 1]\). Hence, \(\{\phi^{(m)}_{[k]} : k \in \mathbb{N}^+\}\) has a convergent subsequence \(\{\phi^{(m)}_{[k]} : k \in \mathbb{N}^+\}\) such that \(\phi^{(m)}_{[k]}(\alpha^{(m)}_{s^{1:1}, a^{1:H-1}, a^{m}_{h}}) \xrightarrow{i} \phi^*(\alpha^{(m)}_{h})\) for any \(s^{1:1}, a^{1:H-1}, a^{m}_{h}\), which implies that \(\phi^*(\alpha^{(m)}_{h}) \geq 0\) and \(\sum_{a^{(m)}_{h}} \phi^*(\alpha^{(m)}_{h}) \geq 0\) for any \(s^{1:1}, a^{1:H-1}, a^{m}_{h}\). Therefore, \(\phi^*\) is a proper stochastic modification.
Then based on eq. (24), it holds for any \( s_{1:h}, a_{1:h} \) that
\[
(\phi^{(m)}_{[k,j],h} \circ \pi_h)(a_h | s_{1:h}, a_{1:h} - 1) = \sum_{\hat{a}_h} \phi^{(m)}_{[k,j],h}(a^{(m)}_h | s_{1:h}, a_{1:h} - 1, \hat{a}_h) \pi_h(\hat{a}_h | s_{1:h}, a_{1:h} - 1)
\]
\[
= \sum_{\hat{a}_h} \phi^{(m)}_{[k,j],h}(a^{(m)}_h | s_{1:h}, a_{1:h} - 1, \hat{a}_h) \pi_h\left(\left[\begin{array}{c} \hat{a}_h \\ a^{(m)}_h \end{array}\right] | s_{1:h}, a_{1:h} - 1\right)
\]
\[
\Rightarrow \sum_{\hat{a}_h} \phi^{(m)}_{[k,j],h}(a^{(m)}_h | s_{1:h}, a_{1:h} - 1, \hat{a}_h) \pi_h\left(\left[\begin{array}{c} \hat{a}_h \\ a^{(m)}_h \end{array}\right] | s_{1:h}, a_{1:h} - 1\right) = (\phi^{(m)}_{[k,j],h} \circ \pi_h)(a_h | s_{1:h}, a_{1:h} - 1).
\] (31)

On one hand, the above inequality and eq. (18) imply that for any \( s_{1:h}, a_{1:h} \), \( p\phi^{(m)}_{[k,j],h} \circ \pi_h(s_{1:h}, a_{1:h}) \Rightarrow p\phi^\ast \circ \pi_h(s_{1:h}, a_{1:h}). \) On the other hand, \( p\phi^{(m)}_{[k,j],h} \circ \pi_h(s_{1:h}, a_{1:h}) \Rightarrow p(s_{1:h}, a_{1:h}). \) Therefore, \( p = p\phi^\ast \circ \pi \) for \( \phi^\ast \) being a stochastic modification, and thus \( p \in \mathcal{X}. \)

E The Range of the Optimal Dual Variable

Before proving Theorem 3 and Corollary 5.1 on the non-asymptotic convergence of Algorithm 1, we first consider the optimal dual variable \( \lambda^{(m)}_s \) of the minimax optimization problem in eq. (7) and denote \( \phi^\ast \) as the optimal solution to the constrained optimization problem in eq. (4) and \( \xi_j \) which is important for the selection of the projection set \( \Lambda^{(m)} \) in Algorithm 1.

**Lemma E.1.** The optimal dual variable \( \lambda^{(m)}_s \) satisfies the following range.

\[
\lambda^{(m)}_s \leq \frac{H_{r,0,\max}^{(m)}}{\xi_j}, \quad j = 1, \ldots, d_m.
\] (32)

**Proof.** Given \( \pi \), denote \( \lambda^{(m)}_s \) as the optimal solution to the constrained optimization problem in eq. (4) and denote \( \tilde{\phi}^{(m)} \) as the stochastic modification that satisfies Assumption 2, i.e., \( V^{(m)}_0(\tilde{\phi}^{(m)} \circ \pi) - V^{(m)}(\phi^\ast \circ \pi) \geq \xi_j^{(m)}. \) Then we have

\[
H_{r,0,\max}^{(m)}(i) \geq V^{(m)}_0(\phi^\ast \circ \pi)
\]
\[
= \max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)}_s)
\]
\[
\geq L^{(m)}(\tilde{\phi}^{(m)} \circ \pi, \lambda^{(m)}_s)
\]
\[
= V^{(m)}_0(\tilde{\phi}^{(m)} \circ \pi) + \sum_{j=1}^{d_m} \lambda^{(m)}_{s,j} (V^{(m)}_j(\tilde{\phi}^{(m)} \circ \pi) - \xi_j^{(m)})
\]
\[
\geq \sum_{j=1}^{d_m} \lambda^{(m)}_{s,j} \xi_j^{(m)},
\]

where (i) and (ii) use \( V^{(m)}_0(\pi) \in [0, H_{r,0,\max}^{(m)}], \forall j = 0, 1, \ldots, d_m \) which is directly implied by Assumption 2, (ii) uses Theorem 2 which implies the equivalence between the constrained optimization problem in eq. (4) and the minimax optimization problem in eq. (7), and (iii) also uses \( \lambda^{(m)}_{s,j} \geq 0 \) and \( V^{(m)}_j(\pi) - \xi_j^{(m)} \geq \xi_j^{(m)}. \) Since \( \xi_j^{(m)} > 0 \), the above inequality implies eq. (32). \qed

F Proof of Theorem 3

Assumption 2 and the value functions defined in eq. (4) imply that for any \( m = 1, \ldots, M, j = 0, 1, \ldots, d_m \) and joint policy \( \pi \), we have

\[
0 \leq V^{(m)}_j(\pi) = E_{\pi}\left[\sum_{h=1}^{H} r^{(m)}_{j,h}(s_h, a_h) \bigg| s_1 \sim \rho\right] \leq H_{r,j,\max}^{(m)}.
\] (33)
Hence, for any $m = 1, \ldots, M$ and joint policy $\pi$

$$
\|V^{(m)}(\pi)\| = \sqrt{\sum_{j=1}^{d_m} V_j^{(m)}(\pi)^2} \leq H \sqrt{\sum_{j=1}^{d_m} (r_{j,\max}^{(m)})^2} = HR_{\max}^{(m)}
$$

(34)

Furthermore, Assumption 1 implies that there is a joint policy $\pi'$ such that $0 \leq c^{(m)} \leq V^{(m)}(\pi')$, so

$$
\|c^{(m)}\| \leq \|V^{(m)}(\pi')\| \leq HR_{\max}^{(m)}.
$$

(35)

Then,

$$
0 \leq \left\| \lambda_T^{(m)} \right\|^2 \\
\overset{(i)}{=} \sum_{t=0}^{T-1} \left( \left\| \lambda_{t+1}^{(m)} \right\|^2 - \left\| \lambda_t^{(m)} \right\|^2 \right) \\
\overset{(ii)}{\leq} \sum_{t=0}^{T-1} \left( \|\lambda_t^{(m)} - \eta (V^{(m)}(\pi_t) - c^{(m)})\|^2 - \|\lambda_t^{(m)}\|^2 \right) \\
\overset{(iii)}{\leq} 2\eta \sum_{t=0}^{T-1} \lambda_t^{(m)\top}(c^{(m)} - V^{(m)}(\pi_t)) + \eta^2 \sum_{t=0}^{T-1} \left( \|V^{(m)}(\pi_t)\| + \|c^{(m)}\| \right)^2 \\
\overset{(iv)}{\leq} 2\eta \sum_{t=0}^{T-1} \lambda_t^{(m)\top}(V^{(m)}(\phi^{(m)}_{t^*} \circ \pi_t) - V^{(m)}(\pi_t)) + 4T(\eta HR_{\max}^{(m)})^2,
$$

where (i) uses the initialization $\lambda_0^{(m)} = 0$, (ii) uses eq. (12), and $0 \in \Lambda^{(m)}$, (iii) uses triangular inequality, and (iv) uses eqs. (34) and (35) and the constraint that $V^{(m)}(\phi^{(m)}_{t^*} \circ \pi) \geq c^{(m)}$ satisfied by the optimal modification $\phi^{(m)}_{t^*}$ of the constrained optimization problem in eq. (4) for $\pi = \pi_t$. Rearranging the above inequality yields that

$$
\sum_{t=0}^{T-1} \lambda_t^{(m)\top}(V^{(m)}(\pi_t) - V^{(m)}(\phi^{(m)}_{t^*} \circ \pi_t)) \leq 2\eta T(HR_{\max}^{(m)})^2.
$$

(36)

Note that

$$
0 \leq \sum_{t=0}^{T-1} \left( \max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi_t, \lambda_t^{(m)}) - L^{(m)}(\phi^{(m)}_{t^*} \circ \pi_t, \lambda_t^{(m)}) \right) \\
\overset{(i)}{=} \sum_{t=0}^{T-1} \left( \max_{\phi^{(m)}} V^{(m)}(\phi^{(m)} \circ \pi_t) - V^{(m)}(\phi^{(m)}_{t^*} \circ \pi_t) \right) \\
\overset{(ii)}{\leq} \sum_{t=0}^{T-1} \left( \epsilon + V^{(m)}(\pi_t) - V^{(m)}(\phi^{(m)}_{t^*} \circ \pi_t) \right) \\
\overset{(iii)}{=} \sum_{t=0}^{T-1} \left( \epsilon + V^{(m)}(\pi_t) - V^{(m)}(\phi^{(m)}_{t^*} \circ \pi_t) + \lambda_t^{(m)\top}(V^{(m)}(\pi_t) - V^{(m)}(\phi^{(m)}_{t^*} \circ \pi_t)) \right) \\
\overset{(iv)}{\leq} \sum_{t=0}^{T-1} \left( \epsilon - D^{(m)}(\pi_t) \right) + 2\eta T(HR_{\max}^{(m)})^2,
$$

(37)

where (i) uses the rewritten Lagrangian function $L^{(m)}(\phi^{(m)} \circ \pi, \lambda^{(m)}) = V^{(m)}(\phi^{(m)} \circ \pi) - \lambda^{(m)\top} c^{(m)}$, (ii) uses eq. (11), (iii) uses $V^{(m)}(\pi) = V^{(m)}(\pi) + \lambda^{(m)\top} V^{(m)}(\pi)$, $\forall \pi$ implies by eqs. (2) and (9), and (iv) uses eqs. (5) and (36). Rearranging the above inequality yields that

$$
\mathbb{E}_t[D^{(m)}(\pi_t)] = \frac{1}{T} \sum_{t=0}^{T-1} D^{(m)}(\pi_t) \leq 2\eta (HR_{\max}^{(m)})^2 + \epsilon,
$$

21
which proves the duality gap in eq. (14) by substituting \( \eta = \frac{1}{\sqrt{L}} \).

Next, we prove the constraint violation in eq. (15).

For any \( \lambda^{(m)} \in \Lambda^{(m)} \), it holds that

\[
\| \lambda_{t+1}^{(m)} - \lambda^{(m)} \|_2^2 \\
\leq (i) \| \lambda_t^{(m)} - \eta (V^{(m)}(\pi_t) - c^{(m)}) - \lambda^{(m)} \|_2^2 \\
\leq (ii) \| \lambda_t^{(m)} - \lambda^{(m)} \|_2^2 - 2\eta (\lambda_t^{(m)} - \lambda^{(m)})^T (V^{(m)}(\pi_t) - c^{(m)}) + \eta^2 (\| V^{(m)}(\pi_t) \|_2^2 + \| c^{(m)} \|_2^2) \\
\leq (iii) \| \lambda_t^{(m)} - \lambda^{(m)} \|_2^2 - 2\eta (\lambda_t^{(m)} - \lambda^{(m)})^T (V^{(m)}(\pi_t) - c^{(m)}) + 4\eta H R_{\text{max}}^{(m)}
\]

where (i) uses eq. (12) and \( \lambda^{(m)} \in \Lambda^{(m)} \), (ii) uses triangular inequality, (iii) uses eqs. (34) and (35). Telescoping the above inequality over \( t = 0, 1, \ldots, T - 1 \) and using \( \lambda_0^{(m)} = 0 \) yields that

\[
\eta \sum_{t=0}^{T-1} \| \lambda_t^{(m)} - \lambda^{(m)} \|_2^2 (V^{(m)}(\pi_t) - c^{(m)}) \leq \frac{1}{2} \| \lambda^{(m)} \|_2^2 + 2T(\eta H R_{\text{max}}^{(m)})^2.
\]

Since \( V^{(m)}(\phi_t^{(m)} \circ \pi_t) \geq c^{(m)} \) and \( \lambda_t^{(m)} \in \mathbb{R}^{d_m}_+ \), eq. (37) implies that

\[
\eta \sum_{t=0}^{T-1} \lambda_t^{(m)}^T (c^{(m)} - V^{(m)}(\pi_t)) \leq \eta \sum_{t=0}^{T-1} (\epsilon + V_0^{(m)}(\pi_t) - V_0^{(m)}(\phi_t^{(m)} \circ \pi_t))
\]

Summing up eqs. (38) and (39) yields that

\[
\eta \sum_{t=0}^{T-1} \lambda_t^{(m)}^T (c^{(m)} - V^{(m)}(\pi_t)) \leq \eta \sum_{t=0}^{T-1} (\epsilon + V_0^{(m)}(\pi_t) - V_0^{(m)}(\phi_t^{(m)} \circ \pi_t)) + \frac{1}{2} \| \lambda^{(m)} \|_2^2 + 2T(\eta H R_{\text{max}}^{(m)})^2.
\]

Denote \( \Phi_t^{(m)} := \{ \phi^{(m)} : V^{(m)}(\phi^{(m)} \circ \pi_t) \geq \min \{ c^{(m)}, V^{(m)}(\pi_t) \} \} \), which is a non-empty set that includes identity modification \( \phi^{(m)} \) such that \( I^{(m)} \circ \pi_t = \pi_t \). Hence,

\[
V_0^{(m)}(\phi_t^{(m)} \circ \pi_t) = \max_{\phi^{(m)} \in \Phi_t^{(m)}} \min_{\lambda^{(m)} \in \mathbb{R}^{d_m}_+} L^{(m)}(\phi^{(m)} \circ \pi_t, \lambda^{(m)}) \leq \min_{\phi^{(m)} \in \Phi_t^{(m)}} \max_{\lambda^{(m)} \in \mathbb{R}^{d_m}_+} L^{(m)}(\phi^{(m)} \circ \pi_t, \lambda^{(m)}) \\
\leq \max_{\phi^{(m)} \in \Phi_t^{(m)}} \{ V_0^{(m)}(\phi^{(m)} \circ \pi_t) + (\lambda_t^{(m)})^T [V^{(m)}(\phi^{(m)} \circ \pi_t) - c^{(m)}] \} \geq \max_{\phi^{(m)} \in \Phi_t^{(m)}} V_0^{(m)}(\phi^{(m)} \circ \pi_t) + (\lambda_t^{(m)})^T \min_{\pi_t} (0, V^{(m)}(\pi_t) - c^{(m)}) \\
\leq V_0^{(m)}(\pi_t) - (\lambda_t^{(m)})^T (c^{(m)} - V^{(m)}(\pi_t)) +
\]

where (i) uses Theorem 2, (ii) uses the fact that \( \Phi_t^{(m)} \) is only a subset of stochastic modifications and denotes that \( \lambda_t^{(m)} = \arg \min_{\lambda^{(m)} \in \mathbb{R}^{d_m}_+} \max_{\phi^{(m)}} L^{(m)}(\phi^{(m)} \circ \pi_t, \lambda^{(m)}) \), (iii) uses \( \lambda_t^{(m)} \in \mathbb{R}^{d_m}_+ \) and the definition of \( \Phi_t^{(m)} \), and (iv) uses the fact that the identity modification \( \phi^{(m)} \in \Phi_t^{(m)} \). Substituting the above inequality into eq. (40) and rearranging it, we obtain that

\[
\eta \sum_{t=0}^{T-1} \lambda_t^{(m)}^T (c^{(m)} - V^{(m)}(\pi_t)) - (\lambda_t^{(m)})^T (c^{(m)} - V^{(m)}(\pi_t)) +
\]

22
where the last inequality uses eq. (32). Substituting the above inequality into eq. (41) yields that

\[
\lambda^{(m)^T}(\epsilon^{(m)} - V^{(m)}(\pi_t)) - (\lambda^{(m)}_{t+})^T(\epsilon^{(m)} - V^{(m)}(\pi_t))_+ \\
\geq \sum_{j=1}^{d_m} \frac{H_r^{(m)}}{\xi_j^{(m)}}(c_j^{(m)} - V_j^{(m)}(\pi_t))_+ ,
\]

where the last inequality uses eq. (32). Substituting the above inequality into eq. (41) yields that

\[
\eta H r^{(m)}_{0,\max} \sum_{t=0}^{T-1} \sum_{j=1}^{d_m} (\xi_j^{(m)})^{-1} (c_j^{(m)} - V_j^{(m)}(\pi_t))_+ \\
\leq \frac{1}{2} \|\lambda^{(m)}\|^2 + 2T(\eta H r^{(m)}_{\max})^2 + \eta T \epsilon \\
\leq 2(H r^{(m)}_{0,\max})^2 \sum_{j=1}^{d_m} (\xi_j^{(m)})^{-2} + 2T(\eta H r^{(m)}_{\max})^2 + \eta T \epsilon ,
\]

where (i) uses \(\|\lambda^{(m)}\| \leq 2H r^{(m)}_{0,\max} \sqrt{\sum_{j=1}^{d_m} (\xi_j^{(m)})^{-2}}\) for our choice \(\lambda_j^{(m)} = 2H r^{(m)}_{0,\max} \sqrt{\sum_{j=1}^{d_m} (\xi_j^{(m)})^{-2}}\).

G Proof of Corollary 5.1

The surrogate rewards defined in eq. (33) has the following bound

\[
0 \leq R_{\lambda,\phi}^{(m)}(s_h, a_h) = r_{0,\lambda}^{(m)}(s_h, a_h) + \lambda_{t,\lambda}^{(m)^T} r_{\lambda}^{(m)}(s_h, a_h) \\
\leq r_{0,\lambda}^{(m)}(s_h, a_h) + \|\lambda_t^{(m)^T}\| r_{\lambda}^{(m)}(s_h, a_h) \\
\leq r_{0,\max}^{(m)} + 2H r_{0,\max}^{(m)} r_{\max}^{(m)} \sum_{j=1}^{d_m} (\xi_j^{(m)})^{-2} := \tilde{R}^{(m)}_{\max} \tag{42}
\]

where (i) uses Assumption 2 and \(\lambda_j^{(m)} \in [0, \frac{2H r_{0,\max}^{(m)}}{\xi_j^{(m)}}]\) (since \(\lambda_j^{(m)} \in \Lambda^{(m)}\) based on eq. (12)). Note that the V-learning in [31] assumes the rewards to range in [0, 1]. To adjust to this assumption, we apply V-learning to the scaled rewards \(\frac{1}{R_{\max}^{(m)}} R_{\lambda,\phi}^{(m)} s_h, a_h \in [0, 1]\) with corresponding value function \(\frac{1}{R_{\max}^{(m)}} V_{\lambda}^{(m)}\). Then based on Theorem 7 of [31], it takes \(\tilde{O}(H^5 S A^2 (\epsilon / R_{\max}^{(m)})^{-2}) = \tilde{O}(H^5 S A^2 \epsilon^{-2})\) samples to reach the \(\epsilon / \tilde{R}_{\max}^{(m)}\)-CE of this scaled Markov game with probability at least \(1 - \delta / T\) for any \(\delta \in (0, 1)\) (we replace \(\delta / T\) which only changes the hidden logarithm factor in \(\tilde{O}\)), that is,

\[
\max_{\phi^{(m)}} \frac{1}{\tilde{R}_{\max}^{(m)}} V_{\lambda}^{(m)}(\phi^{(m)} \circ \pi_t) - \frac{1}{\tilde{R}_{\max}^{(m)}} V_{\lambda}^{(m)}(\pi_t) \leq \frac{\epsilon}{\tilde{R}_{\max}^{(m)}},
\]

which is equivalent to eq. (11). Applying union bound over the \(T\) iterations yields that eq. (11) holds for all iterations \(t = 0, 1, \ldots, T - 1\) with probability at least \(1 - \delta\). In that case, the convergence rates in eqs. (14) and (15) hold. Substituting \(T = \max_m 4\epsilon^{-2}(H r_{\max}^{(m)})^2 (\sum_{j=1}^{d_m} (\xi_j^{(m)})^{-2} + H r_{\max}^{(m)})^2\) and \(r_{0,\max}^{(m)} \geq \frac{1}{n}\) into these convergence rates yields that

\[
\mathbb{E}_{\pi_t}(D^{(m)}(\pi_t)) \leq \frac{2(H r_{\max}^{(m)})^2}{\sqrt{T}} + \epsilon \leq 2\epsilon,
\]

23
\[
\mathbb{E}_t(W^{(m)}(\pi_t)) \leq \frac{2HR^{(m)}_{\text{max}}}{\sqrt{T}} \sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2} + \frac{2H(R^{(m)}_{\text{max}})^2}{Hr^{(m)}_{0,\text{max}}\sqrt{T}} + \frac{\epsilon}{Hr^{(m)}_{0,\text{max}}}
\]
\[
\leq \frac{1}{\sqrt{T}} \left( 2HR^{(m)}_{\text{max}} \sum_{j=1}^{d_m}(\xi_j^{(m)})^{-2} + 2(HR^{(m)}_{\text{max}})^2 \right) + \epsilon \leq 2\epsilon.
\]

The above two inequalities prove that 
\[
\max(\mathbb{E}_tD^{(m)}(\pi_t), \mathbb{E}_tW^{(m)}(\pi_t)) \leq 2\epsilon.
\]

Since each of the \(T = \mathcal{O}(H^4\epsilon^{-2})\) iterations takes \(\tilde{\mathcal{O}}(H^5SA^2\epsilon^{-2})\) samples, the required sample complexity is 
\(T\tilde{\mathcal{O}}(H^5SA^2\epsilon^{-2}) = \tilde{\mathcal{O}}(H^9SA^2\epsilon^{-4})\).