A PROOF OF PROPOSITION 4.1

Proof. For all $(s, a^i) \in \mathcal{S} \times \mathcal{A}^i$, taking the derivative of the Lagrangian of the optimization problem with simplex constraints over π^i,

$$\alpha_{t,(i)} \cdot \big\langle \widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, \cdot) - \lambda_i \cdot \log \pi^i_{\theta_t}(\cdot \,|\, s), \pi^i(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s) \big\rangle \\ - \mathrm{KL}\big(\pi^i(\cdot \,|\, s) \,\big\|\, \pi^i_{\theta_t}(\cdot \,|\, s)\big) + C \cdot \bigg(\sum_{a^i \in \mathcal{A}^i} \pi^i(a^i \,|\, s) - 1 \bigg) + C' \cdot \sum_{a^i \in \mathcal{A}^i} \pi^i(a^i \,|\, s)$$

with respect to $\pi^i(a^i \,|\, s)$, we obtain

$$\alpha_{t,(i)} \cdot \big(\widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, a^i) - \lambda_i \cdot \log \pi^i_{\theta_t}(a^i \,|\, s) \big) - \log \pi^i(a^i \,|\, s) + \log \pi^i_{\theta_t}(a^i \,|\, s) + C + C' - 1, \tag{A.1}$$

where C is a constant. Setting (A.1) to zero we further have

$$\pi^i_{t+1}(a^i \,|\, s) = \big(\pi^i_{\theta_t}(a^i \,|\, s)\big)^{1-\lambda_i \alpha_{t,(i)}} \cdot \exp\big\{ \alpha_{t,(i)} \cdot \widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, a^i) + C + C' - 1 \big\} \\ = \exp\big\{ \alpha'_{t,(i)} \cdot \mathcal{E}_{\theta_t,(i)}(s, a^i) + \alpha_{t,(i)} \cdot \widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, a^i) + C + C' - 1 \big\},$$

where the second equality follows from $\lambda_i = (1 - \alpha'_{t,(i)})/\alpha_{t,(i)}$. Thus, we know that (4.2) is $\pi^i_{t+1}(\cdot \,|\, s) \propto \exp\{\alpha'_{t,(i)} \cdot \mathcal{E}_{\theta_t,(i)}(s, \cdot) + \alpha_{t,(i)} \cdot \widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, \cdot)\}$, which coincides with (3.3). Therefore, we finish the proof. □

B DISCUSSION ON ASSUMPTION 4.4

In this section, we make a detailed discussion on Assumption 4.4. First, we give the following condition on the mean squared error of the estimations $\widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}$ and $\mathcal{E}_{\theta_{t+1},(i)}$.

Condition B.1 (Approximation Error). For all $0 \le t \le T-1$, the estimator $\widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, a^i)$ of the marginalized $Q_{(i)}$-function $\widetilde{Q}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, a^i)$ satisfies

$$\mathbb{E}_{\sigma_t}\Big[\big(\widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, a^i) - \widetilde{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}(s, a^i) \big)^2 \Big] \le \varepsilon_t, \tag{B.1}$$

and the energy function update $\mathcal{E}_{\theta_{t+1},(i)}(s, a^i)$ satisfy

$$\mathbb{E}_{\sigma_t}\Big[\big(\mathcal{E}_{\theta_{t+1},(i)}(s, a^i) - \widehat{\mathcal{E}}_{\theta_{t+1},(i)}(s, a^i) \big)^2 \Big] \le \varepsilon'_t. \tag{B.2}$$

Condition B.1 can be satisfied for arbitrary small errors ε_t and ε'_t if the following conditions are satisfied: (I) the representation powers of parametrization $\mathcal{E}_{\theta_{t+1},(i)}(s, a^i)$ and the paramerization of $\widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}$ are strong enough. (II) the algorithms for learning $\widehat{\mathcal{E}}_{t+1,(i)}(s, a^i)$ and $\widetilde{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}$ attain stationary points after sufficiently many iterations. In this section, as an example, we take the most commonly used neural network parameterization for $\mathcal{E}_{\theta_{t+1},(i)}(s, a^i)$ and $\widehat{Q}_{(i)}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}$.

Representation Power. There is a line of literature discussing the representation power of neural networks (see, e.g., Daniely et al. (2016); Khrulkov et al. (2017)) showing that the overparameterized neural networks possess strong representation power. Specifically, the representation power of neural networks can be approximated as a subset of the reproducing kernel Hilbert space (RKHS) with neural tangent kernel (Jacot et al., 2018; Chizat & Bach, 2018; Allen-Zhu et al., 2018; Lee et al., 2019; Arora et al., 2019), which is a sufficiently rich function class. Moreover, we present the following lemma justifying the choice of the truncation parameter $\mathcal{E}_{(i)}^{\max}$ in Theorem 4.5.

Lemma B.2 (Bounded Energy Function)**.** Let the regularization parameter λ_i and the stepsizes $\alpha_{t,(i)}, \alpha'_{t,(i)}$ be chosen as in Theorem 4.5. In Algorithm 1, setting $\mathcal{E}_{(i)}^{\max} = Q_{(i)}^{\max}/(\lambda_i - M_i)$ makes the function class $\mathcal{F}_{\mathcal{E}_{(i)}^{\max}}$ always cover the range of the estimated energy function update $\widehat{\mathcal{E}}_{t,(i)}$ obtained in Line 6 of Algorithm 1.

Proof. See Appendix H for a detailed proof. □

Lemma B.2 states that, truncating the energy functions $\mathcal{E}_{\theta_t,(i)}$ within the function class $\mathcal{F}_{Q_{(i)}^{\max}/(\lambda_i - M_i)}$ does not compromise the MSE ε_t' defined in (B.2).

Learning Algorithms. There are some recent advances (Cai et al., 2019; Liu et al., 2019) showing that, when equipped with neural network parameterization, temporal-difference (TD) learning converges to the stationary point. Also, as discussed in Liu et al. (2019), under neural network parameterization, stochastic gradient descent also converges to the optima at a sublinear rate.

Next, by Lemmas 4.7-4.8 of Liu et al. (2019), we lay out error bounds in the form of those in Assumption 4.4.

Lemma B.3. ϵ_t and ϵ_t' in Assumption 4.4 take the forms of

$$\epsilon_t = \max_{i\in\{1,2\}} \left\{|\mathcal{A}^i|\right\} \cdot (\varepsilon_t')^2, \quad \epsilon_t' = \varepsilon_t' \cdot \max_{i\in\{1,2\}} \{\phi^*_{\pi^i_{\theta_t}}\} + \varepsilon_t \cdot \psi_t^*,$$

where

$$\psi_t^* = \mathbb{E}_{\sigma_t}\left[\left|\frac{\mathrm{d}\sigma^*}{\mathrm{d}\sigma_t} - \frac{\mathrm{d}\nu^*}{\mathrm{d}\nu_t}\right|\right], \quad \phi^*_{\pi^i_{\theta_t}} = \mathbb{E}_{\sigma_t}\left[\left|\frac{\mathrm{d}\pi^i_*}{\mathrm{d}\pi^i_0} - \frac{\mathrm{d}\pi^i_{\theta_t}}{\mathrm{d}\pi^i_0}\right|\right],$$

Here the density ratios are Radon-Nikodym derivatives.

Thus, as long as the learning algorithms run sufficiently many iterations such that the errors ε_t and ε_t' are sufficiently small, for σ defined in (4.13), $\sigma = \widetilde{O}(1)$ can be achieved. Finally, we remark that, with some recent advances in the variance reduced techniques for policy optimization (Papini et al., 2018; Xu et al., 2020; Shen et al., 2019; Xu et al., 2019; Huang et al., 2020), we expect the MSEs in (B.1) and (B.2) being further reduced, which could possibly allow us to use relatively smaller neural networks for $\widehat{Q}_{(i)}^{\pi^i_{\theta_t},\pi^{-i}_{\theta_t}}$ and $\mathcal{E}_{\theta_{t+1},(i)}$. This can help to boost practicality of smooth FSP due to the reduced computational cost associated with the reduced network size. We leave this direction to our future research.

C Proof of Theorem 4.5

In this section, we lay out the proof of Theorem 4.5. We have the following proposition on the regularization bias of $\mathcal{J}_{(i)}(\pi^i, \pi^{-i})$ defined in (4.3).

Proposition C.1 (Regularization Bias)**.** The regularized performance function $\mathcal{J}_{(i)}(\pi^i, \pi^{-i})$ satisfies

$$\mathcal{J}(\pi^1,\pi^2) < \mathcal{J}_{(1)}(\pi^1,\pi^2) \le \mathcal{J}(\pi^1,\pi^2) + \frac{\lambda_1 \cdot \log|\mathcal{A}^1|}{1-\gamma}, \tag{C.1}$$

$$-\mathcal{J}(\pi^1,\pi^2) < \mathcal{J}_{(2)}(\pi^2,\pi^1) \le -\mathcal{J}(\pi^1,\pi^2) + \frac{\lambda_2 \cdot \log|\mathcal{A}^2|}{1-\gamma} \tag{C.2}$$

for all policy pairs $[\pi^1, \pi^2]$.

Proof. See Appendix D for a detailed proof. □

Based on Assumptions 4.2 and 4.3, we have the following lemma on the Lipschitz continuity of the marginalized $Q_{(i)}$-function. Recall that ζ is the concentrability coefficient in Assumption 4.2 and ι_i is the Lipschitz coefficient in Assumption 4.3.

Lemma C.2 (Lipschitz Marginalized $Q_{(i)}$-Function)**.** Suppose that Assumption 4.2 holds. We choose the regularization parameter λ_i, the truncation parameter $\mathcal{E}_{(i)}^{\max}$, and the stepsizes $\alpha_{t,(i)}, \alpha'_{t,(i)}$ as in Theorem 4.5. We have for all $[\pi^i; \pi^{-i}] = [\pi^i_{\theta_t}; \pi^{-i}_{\theta_t}]$ generated by the policy update in Line 4 of

Algorithm 1 that,

$$\begin{aligned}
&\Big|\mathbb{E}_{\nu^*}\Big[\big\langle \widetilde{Q}_{(i)}^{\pi^i,\pi^{-i}}(s,\,\cdot\,) - \widetilde{Q}_{(i)}^{\pi^i,\pi_*^{-i}}(s,\,\cdot\,), \pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\big\rangle\Big]\Big| \\
&\quad \le \sqrt{2\iota_i\cdot\iota_{-i}}\cdot\Big\{\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)^{1/2}\cdot\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)^{1/2}\Big] \\
&\qquad + \frac{V_{(i)}^{\max}+Q_{(i)}^{\max}\cdot\zeta}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)\Big]^{1/2}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}\Big\}.
\end{aligned}$$

Proof. See Appendix E for a detailed proof. □

Recall that $\mathcal{J}_{(i)}(\pi^i,\pi^{-i})$ is defined in (4.3). We present the following extended performance difference lemma, which extends the performance difference lemma of Kakade & Langford (2002) to the two-agent setting with entropy regularization.

Lemma C.3 (Extended Performance Difference)**.** We have for all $[\pi^i;\pi^{-i}]$ that,

$$\begin{aligned}
&\big[\mathcal{J}_{(i)}(\pi_*^i,\pi_*^{-i}) - \mathcal{J}_{(i)}(\pi^i,\pi_*^{-i})\big] + \frac{\lambda_i}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)\Big] \\
&\qquad = \frac{1}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\big\langle \widetilde{Q}_{(i)}^{\pi^i,\pi_*^{-i}}(s,\,\cdot\,) - \lambda_i\cdot\log\pi^i(\cdot\,|\,s), \pi_*^i(\cdot\,|\,s) - \pi^i(\cdot\,|\,s)\big\rangle\Big].
\end{aligned}$$

Proof. See Appendix F for a detailed proof. □

The following lemma establishes the one-step descent of the KL-divergence between a Nash equilibrium $[\pi_*^1;\pi_*^2]$ and the policy sequence $\{[\pi_{\theta_t}^1;\pi_{\theta_t}^2]\}_{0\le t\le T-1}$ generated by Line 4 of Algorithm 1 in the infinite-dimensional policy space, which extends the analysis of mirror descent (Nemirovski & Yudin, 1983; Nesterov, 2013). Recall that the energy function $\mathcal{E}_{\theta_t,(i)}$ is obtained in Line 7 of Algorithm 1 and the ideal energy function update $\overline{\mathcal{E}}_{t+1,(i)}$ is defined in (4.10).

Lemma C.4 (One-Step Descent)**.** Suppose that the stepsizes satisfy $\alpha'_{t,(i)} = 1-\lambda_i\alpha_{t,(i)}$. For the policy sequence $\{\pi_{\theta_t}^i\}_{0\le t\le T-1}$ generated by the policy update in Line 4 of Algorithm 1, we have for all $s\in\mathcal{S}$ that,

$$\begin{aligned}
&\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi_{\theta_{t+1}}^i(\cdot\,|\,s)\big) - \mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi_{\theta_t}^i(\cdot\,|\,s)\big) \qquad (C.3)\\
&\quad \ge \big\langle \mathcal{E}_{\theta_{t+1},(i)}(s,\cdot) - \overline{\mathcal{E}}_{t+1,(i)}(s,\cdot), \pi_*^i(\cdot\,|\,s) - \pi_{\theta_t}^i(\cdot\,|\,s)\big\rangle - 1/2\cdot\big\|\mathcal{E}_{\theta_{t+1},(i)}(s,\cdot) - \mathcal{E}_{\theta_t,(i)}(s,\cdot)\big\|_\infty^2 \\
&\qquad + \alpha_{t,(i)}\cdot\big\langle -\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_{\theta_t}^{-i}}(s,\,\cdot\,) + \lambda_i\cdot\log\pi_{\theta_t}^i(\cdot\,|\,s), \pi_*^i(\cdot\,|\,s) - \pi_{\theta_t}^i(\cdot\,|\,s)\big\rangle.
\end{aligned}$$

Proof. See Appendix G for a detailed proof. □

Proof of Theorem 4.5. For notational simplicity, we write $\mathrm{KL}(\pi_*^i(\cdot\,|\,s)\,\|\,\pi_{\theta_{t+1}}^i(\cdot\,|\,s))$ as $\mathrm{KL}_{t,(i)}(s)$ throughout this proof. By the choices of the stepsizes in (4.15) of Theorem 4.5, we have $\alpha'_{t,(i)} = 1-\lambda_i\alpha_{t,(i)}$. By Lemma C.4, we have under (4.12) of Assumption 4.4 that,

$$\begin{aligned}
&\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t+1,(i)}(s)\big] - \mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big] \qquad (C.4)\\
&\quad \le \epsilon'_t + \alpha_{t,(i)}\cdot\underbrace{\mathbb{E}_{\nu^*}\Big[\big\langle -\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_*^{-i}}(s,\,\cdot\,) + \lambda_i\cdot\log\pi_{\theta_t}^i(\cdot\,|\,s), \pi_*^i(\cdot\,|\,s) - \pi_{\theta_t}^i(\cdot\,|\,s)\big\rangle\Big]}_{\text{(I)}} \\
&\qquad + \alpha_{t,(i)}\cdot\underbrace{\mathbb{E}_{\nu^*}\Big[\big\langle -\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_{\theta_t}^{-i}}(s,\,\cdot\,) + \widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_*^{-i}}(s,\,\cdot\,), \pi_*^i(\cdot\,|\,s) - \pi_{\theta_t}^i(\cdot\,|\,s)\big\rangle\Big]}_{\text{(II)}} \\
&\qquad + \underbrace{\mathbb{E}_{\nu^*}\Big[1/2\cdot\big\|\mathcal{E}_{\theta_{t+1},(i)}(s,\cdot) - \mathcal{E}_{\theta_t,(i)}(s,\cdot)\big\|_\infty^2\Big]}_{\text{(III)}}.
\end{aligned}$$

For (I), by Lemma C.3, we have

$$\text{(I)} = (1-\gamma)\cdot\big[\mathcal{J}_{(i)}(\pi^i_{\theta_t},\pi^{-i}_*) - \mathcal{J}_{(i)}(\pi^i_*,\pi^{-i}_*)\big] - \lambda_i\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big]. \tag{C.5}$$

For (II), by Lemma C.2, we have

$$\begin{aligned}\text{(II)} &\le \sqrt{2\iota_i\cdot\iota_{-i}}\cdot\Big\{\mathbb{E}_{\nu^*}\big[\mathrm{KL}^{1/2}_{t,(i)}(s)\cdot\mathrm{KL}^{1/2}_{t,(i)}(s)\big] \\ &\qquad + \frac{V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big]^{1/2}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(-i)}(s)\big]^{1/2}\Big\} \\ &\le \bigg(1+\frac{V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta}{1-\gamma}\bigg)\cdot\Big(\iota_i\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big]+\iota_{-i}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(-i)}(s)\big]\Big).\end{aligned} \tag{C.6}$$

For (III), by (4.11) of Assumption 4.4 and the definition of $\widehat{\mathcal{E}}_{t+1,(i)}$ in (3.8), we have

$$\begin{aligned}\text{(III)} &\le \mathbb{E}_{\nu^*}\Big[\big\|\mathcal{E}_{\theta_{t+1},(i)}(s,\cdot)-\widehat{\mathcal{E}}_{t+1,(i)}(s,\cdot)\big\|^2_\infty + \big\|\widehat{\mathcal{E}}_{t+1,(i)}(s,\cdot)-\mathcal{E}_{\theta_t,(i)}(s,\cdot)\big\|^2_\infty\Big] \\ &\le \epsilon_t + \alpha^2_{t,(i)}\cdot\mathbb{E}_{\nu^*}\Big[\big\|-\lambda_i\cdot\mathcal{E}_{\theta_t,(i)}(s,\,\cdot\,)+\widehat{Q}^{\pi^i_{\theta_t},\pi^{-i}_{\theta_t}}_{(i)}(s,\,\cdot\,)\big\|^2_\infty\Big] \\ &\le \epsilon_t + \alpha^2_{t,(i)}\cdot\mathbb{E}_{\nu^*}\Big[2\lambda_i^2\cdot\big\|\mathcal{E}_{\theta_t,(i)}(s,\,\cdot\,)\big\|^2_\infty + 2\cdot\big\|\widehat{Q}^{\pi^i_{\theta_t},\pi^{-i}_{\theta_t}}_{(i)}(s,\,\cdot\,)\big\|^2_\infty\Big] \\ &\le \epsilon_t + \big[2+2\lambda_i^2/(\lambda_i-M_i)^2\big]\cdot(Q^{\max}_{(i)})^2\cdot\alpha^2_{t,(i)}.\end{aligned} \tag{C.7}$$

Here the last inequality follows from the truncations $\widehat{Q}^{\pi^i_{\theta_t},\pi^{-i}_{\theta_t}}_{(i)} \in \mathcal{F}_{Q^{\max}_{(i)}}$ and $\mathcal{E}_{\theta_t,(i)}\in\mathcal{F}_{\mathcal{E}^{\max}_{(i)}}$, where $\mathcal{E}^{\max}_{(i)} = Q^{\max}_{(i)}/(\lambda_i-M_i)$.

Then plugging (C.5), (C.6), and (C.7) into (C.4), we obtain

$$\begin{aligned}&\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t+1,(i)}(s)\big] - \bigg\{1-\bigg[\lambda_i-\bigg(1+\frac{V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta}{1-\gamma}\bigg)\cdot\iota_i\bigg]\cdot\alpha_{t,(i)}\bigg\}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big] \\ &\quad\le (\epsilon_t+\epsilon'_t) + \bigg(1+\frac{V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta}{1-\gamma}\bigg)\cdot\iota_{-i}\cdot\alpha_{t,(i)}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(-i)}(s)\big]^{1/2} \\ &\qquad + (1-\gamma)\alpha_{t,(i)}\cdot\big[\mathcal{J}_{(i)}(\pi^i_{\theta_t},\pi^{-i}_*) - \mathcal{J}_{(i)}(\pi^i_*,\pi^{-i}_*)\big] + \big[2+2\lambda_i^2/(\lambda_i-M_i)^2\big]\cdot(Q^{\max}_{(i)})^2\cdot\alpha^2_{t,(i)}.\end{aligned} \tag{C.8}$$

Summing (C.8) for $i\in\{1,2\}$ and setting $\alpha_{t,(1)}=\alpha_{t,(2)}=\eta_t$, we obtain

$$\begin{aligned}&(1-\gamma)\eta_t\cdot\sum_{i\in\{1,2\}}\big[\mathcal{J}_{(i)}(\pi^i_*,\pi^{-i}_*) - \mathcal{J}_{(i)}(\pi^i_{\theta_t},\pi^{-i}_*)\big] \\ &\quad\le \sum_{i\in\{1,2\}}\bigg\{1-\bigg\{\lambda_i-\bigg[2+\frac{\sum_{i\in\{1,2\}}(V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta)}{1-\gamma}\bigg]\cdot\iota_i\bigg\}\cdot\eta_t\bigg\}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big] \\ &\qquad - \sum_{i\in\{1,2\}}\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t+1,(i)}(s)\big] + \bigg[2(\epsilon_t+\epsilon'_t)+\sum_{i\in\{1,2\}}\big[2+2\lambda_i^2/(\lambda_i-M_i)^2\big]\cdot(Q^{\max}_{(i)})^2\cdot\eta_t^2\bigg] \\ &\quad= \sum_{i\in\{1,2\}}\big[1-(\lambda_i-M_i)\cdot\eta_t\big]\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big] - \sum_{i\in\{1,2\}}\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t+1,(i)}(s)\big] \\ &\qquad + \bigg\{2(\epsilon_t+\epsilon'_t)+\sum_{i\in\{1,2\}}\big[2+2\lambda_i^2/(\lambda_i-M_i)^2\big]\cdot(Q^{\max}_{(i)})^2\cdot\eta_t^2\bigg\}.\end{aligned} \tag{C.9}$$

Multiplying the both sides of (C.9) by $t+1$ and setting $\eta_t = 1/[(t+1)\cdot \min_{i\in\{1,2\}}\{\lambda_i - M_{(i)}\}]$, we obtain for all $0 \le t \le T-1$,

$$\begin{aligned}
&\frac{1-\gamma}{\min_{i\in\{1,2\}}\{\lambda_i - M_i\}} \cdot \sum_{i\in\{1,2\}} \big[\mathcal{J}_{(i)}(\pi_*^i, \pi_*^{-i}) - \mathcal{J}_{(i)}(\pi_{\theta_t}^i, \pi_*^{-i})\big] \\
&\quad \le t \cdot \sum_{i\in\{1,2\}} \mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big] - (t+1)\cdot \sum_{i\in\{1,2\}} \mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t+1,(i)}(s)\big] \\
&\qquad + \bigg\{2(t+1)\cdot(\epsilon_t + \epsilon_t') + \frac{\sum_{i\in\{1,2\}}\big[2 + 2\lambda_i^2/(\lambda_i - M_i)^2\big]\cdot (Q_{(i)}^{\max})^2}{(t+1)\cdot \min_{i\in\{1,2\}}\{\lambda_i - M_i\}^2}\bigg\}. \qquad \text{(C.10)}
\end{aligned}$$

Telescoping (C.10) over $0 \le t \le T-1$, by $\sum_{t=1}^T 1/t \le \log T$ and the nonnegativity of the KL-divergence, we obtain

$$\begin{aligned}
&\frac{1-\gamma}{T\cdot \min_{i\in\{1,2\}}\{\lambda_i - M_i\}} \cdot \sum_{t=0}^{T-1}\sum_{i\in\{1,2\}} \big[\mathcal{J}_{(i)}(\pi_*^i, \pi_*^{-i}) - \mathcal{J}_{(i)}(\pi_{\theta_t}^i, \pi_*^{-i})\big] \\
&\quad \le \frac{1}{T}\cdot \bigg\{2\cdot \sum_{t=0}^{T-1}(t+1)\cdot(\epsilon_t + \epsilon_t') + \frac{\sum_{i\in\{1,2\}}\big[2 + 2\lambda_i^2/(\lambda_i - M_i)^2\big]\cdot (Q_{(i)}^{\max})^2}{\min_{i\in\{1,2\}}\{\lambda_i - M_i\}^2}\cdot \sum_{t=1}^T \frac{1}{t}\bigg\} \\
&\qquad - \sum_{i\in\{1,2\}} \mathbb{E}_{\nu^*}\big[\mathrm{KL}_{T,(i)}(s)\big] \\
&\quad \le \frac{2\sigma}{T} + \frac{\sum_{i\in\{1,2\}}\big[2 + 2\lambda_i^2/(\lambda_i - M_i)^2\big]\cdot (Q_{(i)}^{\max})^2}{\min_{i\in\{1,2\}}\{\lambda_i - M_i\}^2}\cdot \frac{\log T}{T}, \qquad \text{(C.11)}
\end{aligned}$$

where σ is defined in (4.13). Finally, by Proposition C.1, we have

$$\sum_{i\in\{1,2\}} \big[\mathcal{J}_{(i)}(\pi_*^i, \pi_*^{-i}) - \mathcal{J}_{(i)}(\pi_{\theta_t}^i, \pi_*^{-i})\big] \ge \mathcal{J}(\pi_*^1, \pi_{\theta_t}^2) - \mathcal{J}(\pi_{\theta_t}^1, \pi_*^2) - \sum_{i\in\{1,2\}} \lambda_i \cdot \log|\mathcal{A}^i|,$$

combining which with (C.11), we obtain

$$\begin{aligned}
&\frac{1}{T}\cdot \sum_{t=0}^{T-1}\big[\mathcal{J}(\pi_*^1, \pi_{\theta_t}^2) - \mathcal{J}(\pi_{\theta_t}^1, \pi_*^2)\big] \\
&\quad \le \frac{2\sigma\cdot \min_{i\in\{1,2\}}\{\lambda_i - M_i\}}{(1-\gamma)\cdot T} + \frac{\sum_{i\in\{1,2\}}\big[2 + 2\lambda_i^2/(\lambda_i - M_i)^2\big]\cdot (Q_{(i)}^{\max})^2}{(1-\gamma)\cdot \min_{i\in\{1,2\}}\{\lambda_i - M_i\}}\cdot \frac{\log T}{T} + \sum_{i\in\{1,2\}} \lambda_i \cdot \log|\mathcal{A}^i|.
\end{aligned}$$

Thus, we conclude the proof of Theorem 4.5. □

D Proof of Proposition C.1

Proof. By the definition of $\mathcal{J}_{(1)}(\pi^1, \pi^2)$ and $\mathcal{J}(\pi^1, \pi^2)$, we have

$$\begin{aligned}
\mathcal{J}_{(1)}(\pi^1, \pi^2) - \mathcal{J}(\pi^1, \pi^2) &= \mathbb{E}_{\nu^*}\big[V_{(1)}^{\pi^1,\pi^2}(s) - V_1^{\pi^1,\pi^2}(s)\big] \\
&= \frac{1}{1-\gamma}\cdot \mathbb{E}_{s'\sim\rho_{\nu^*}^{\pi^1,\pi^2}}\big[r_{(1)}^{\pi^1,\pi^2}(s') - r_1^{\pi^1,\pi^2}(s')\big] \\
&= \frac{1}{1-\gamma}\cdot \mathbb{E}_{s'\sim\rho_{\nu^*}^{\pi^1,\pi^2}}\Big[\lambda_1 \cdot H\big(\pi^1(\cdot\,|\,s)\big)\Big].
\end{aligned}$$

Since $0 < H(\pi^1(\cdot\,|\,s)) \le \log|\mathcal{A}^1|$, we further obtain

$$\mathcal{J}(\pi^1, \pi^2) < \mathcal{J}_{(1)}(\pi^1, \pi^2) \le \mathcal{J}(\pi^1, \pi^2) + \frac{\lambda_1 \cdot \log|\mathcal{A}^1|}{1-\gamma},$$

which finishes the proof of (C.1). Using the same argument, we can also prove (C.2).Thus, we conclude the proof of Proposition C.1. □

E PROOF OF LEMMA C.2

In the subsequent analysis, using the notion of the state-transition operator $\mathcal{P}^{\pi^i,\pi^{-i}}$ in (4.7), we write $\rho^{\pi^i,\pi^{-i}}_{s,a^i,\pi^{-i}}$, which is defined in (4.5), as

$$\rho^{\pi^i,\pi^{-i}}_{s,a^i,\pi^{-i}} = \bigg[(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\cdot(\mathcal{P}^{\pi^i,\pi^{-i}})^t\bigg]\circ\mathcal{P}^{a^i,\pi^{-i}}\circ\delta_s, \tag{E.1}$$

where δ_s is the Dirac delta function.

Lemma E.1. Under Assumption 4.3, we have

$$\bigg\|\sum_{t=0}^{\infty}\gamma^t\cdot\big[(\mathcal{P}^{\pi^i,\pi^{-i}})^t-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\big]\bigg\|_{\mathrm{op}}\le\frac{\gamma\iota_i}{(1-\gamma)^2}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}.$$

Proof. Since $\|\mathcal{P}^{\pi^1,\pi^2}\|_{\mathrm{op}}\le 1$ (Lasota & Mackey, 2013), we have

$$\begin{aligned}
&\big\|(\mathcal{P}^{\pi^i,\pi^{-i}})^t-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\big\|_{\mathrm{op}} \qquad\qquad (\text{E.2})\\
&=\Big\|(\mathcal{P}^{\pi^i,\pi^{-i}})^{t-1}\circ(\mathcal{P}^{\pi^i,\pi^{-i}}-\mathcal{P}^{\pi^i,\pi_*^{-i}})+\big[(\mathcal{P}^{\pi^i,\pi^{-i}})^{t-1}-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^{t-1}\big]\circ\mathcal{P}^{\pi^i,\pi_*^{-i}}\Big\|_{\mathrm{op}}\\
&\le\big\|(\mathcal{P}^{\pi^i,\pi^{-i}})^{t-1}\circ(\mathcal{P}^{\pi^i,\pi^{-i}}-\mathcal{P}^{\pi^i,\pi_*^{-i}})\big\|_{\mathrm{op}}+\Big\|\big[(\mathcal{P}^{\pi^i,\pi^{-i}})^{t-1}-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^{t-1}\big]\circ\mathcal{P}^{\pi^i,\pi_*^{-i}}\Big\|_{\mathrm{op}}\\
&\le\|\mathcal{P}^{\pi^i,\pi^{-i}}-\mathcal{P}^{\pi^i,\pi_*^{-i}}\|_{\mathrm{op}}+\big\|(\mathcal{P}^{\pi^i,\pi^{-i}})^{t-1}-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^{t-1}\big\|_{\mathrm{op}}.
\end{aligned}$$

Recursively applying (E.2) gives

$$\begin{aligned}
\big\|(\mathcal{P}^{\pi^i,\pi^{-i}})^t-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\big\|_{\mathrm{op}}&\le t\cdot\|\mathcal{P}^{\pi^i,\pi^{-i}}-\mathcal{P}^{\pi^i,\pi_*^{-i}}\|_{\mathrm{op}}\\
&\le t\cdot\iota_i\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)^{1/2}\Big],
\end{aligned}$$

where the second inequality follows from Assumption 4.3. Thus, we have

$$\begin{aligned}
\bigg\|\sum_{t=0}^{\infty}\gamma^t\cdot\big[(\mathcal{P}^{\pi^i,\pi^{-i}})^t-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\big]\bigg\|_{\mathrm{op}}&\le\sum_{t=0}^{\infty}\gamma^t\cdot\big\|(\mathcal{P}^{\pi^i,\pi^{-i}})^t-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\big\|_{\mathrm{op}}\\
&\le\bigg(\sum_{t=0}^{\infty}t\gamma^t\bigg)\cdot\iota_i\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}\\
&=\frac{\gamma\iota_i}{(1-\gamma)^2}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2},
\end{aligned}$$

which concludes the proof of Lemma E.1. □

Now we are ready to prove Lemma C.2.

Proof of Lemma C.2. By the definition of the $Q_{(i)}$-function in (2.5), we have

$$Q^{\pi^i,\pi^{-i}}_{(i)}(s,a^i,a^{-i})=r_i(s,a^i,a^{-i})+\gamma\cdot\mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a^i,a^{-i})}\big[V^{\pi^i,\pi^{-i}}_{(i)}(s')\big],$$

which gives

$$\begin{aligned}
&\Big|\mathbb{E}_{\nu^*}\Big[\big\langle \widetilde{Q}_{(i)}^{\pi^i,\pi^{-i}}(s,\,\cdot\,) - \widetilde{Q}_{(i)}^{\pi^i,\pi_*^{-i}}(s,\,\cdot\,), \pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\big\rangle\Big]\Big| \qquad \text{(E.3)}\\
&\quad= \Big|\mathbb{E}_{\nu^*}\Big[\big\langle r_i^{\cdot,\pi^{-i}}(s) - r_i^{\cdot,\pi_*^{-i}}(s), \pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\big\rangle\\
&\qquad\quad+ \big\langle [\mathcal{P}^{\cdot,\pi^{-i}} \circ V_{(i)}^{\pi^i,\pi^{-i}}](s) - [\mathcal{P}^{\cdot,\pi_*^{-i}} \circ V_{(i)}^{\pi^i,\pi_*^{-i}}](s), \pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\big\rangle\Big]\Big|\\
&\quad= \underbrace{\Big|\mathbb{E}_{\nu^*}\Big[\big\langle r_i^{\cdot,\pi^{-i}}(s) - r_i^{\cdot,\pi_*^{-i}}(s), \pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\big\rangle\Big]\Big|}_{\text{(I)}}\\
&\qquad+ \underbrace{\Big|\mathbb{E}_{\nu^*}\Big[\big\langle [\mathcal{P}^{\cdot,\pi^{-i}} \circ V_{(i)}^{\pi^i,\pi^{-i}}](s) - [\mathcal{P}^{\cdot,\pi_*^{-i}} \circ V_{(i)}^{\pi^i,\pi_*^{-i}}](s), \pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\big\rangle\Big]\Big|}_{\text{(II)}}.
\end{aligned}$$

Upper Bounding (I): By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\text{(I)} &\le \mathbb{E}_{\nu^*}\Big[\big\|r_i^{\cdot,\pi^{-i}}(s) - r_i^{\cdot,\pi_*^{-i}}(s)\big\|_\infty \cdot \big\|\pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\big\|_1\Big] \qquad \text{(E.4)}\\
&\le \sqrt{2}\cdot\iota_i\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)^{1/2}\cdot\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)^{1/2}\Big],
\end{aligned}$$

where the second inequality follows from Assumption 4.3 and the Pinsker's inequality. Meanwhile, we have

$$\begin{aligned}
\text{(I)} &= \Big|\mathbb{E}_{\nu^*}\Big[\big\langle r_i^{\pi_*^i,\cdot}(s) - r_i^{\pi_*^i,\cdot}(s), \pi^{-i}(\cdot\,|\,s) - \pi_*^{-i}(\cdot\,|\,s)\big\rangle\Big]\Big| \qquad \text{(E.5)}\\
&\le \sqrt{2}\cdot\iota_{-i}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{i}(\cdot\,|\,s)\,\big\|\,\pi^{i}(\cdot\,|\,s)\big)^{1/2}\cdot\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)^{1/2}\Big],
\end{aligned}$$

where the inequality follows from the same argument as (E.4). Combining (E.4) and (E.5), we obtain

$$\text{(I)} \le \sqrt{2\iota_i\cdot\iota_{-i}}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{i}(\cdot\,|\,s)\,\big\|\,\pi^{i}(\cdot\,|\,s)\big)^{1/2}\cdot\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)^{1/2}\Big]. \qquad \text{(E.6)}$$

Upper Bounding (II): Recall that we have

$$\begin{aligned}
V_{(i)}^{\pi^i,\pi^{-i}}(s) &= \frac{1}{1-\gamma}\cdot\mathbb{E}_{s'\sim\rho_s^{\pi^i,\pi^{-i}}}\big[r_{(i)}^{\pi^i,\pi^{-i}}(s')\big]\\
&= \frac{1}{1-\gamma}\cdot\mathbb{E}_{s'\sim\rho_s^{\pi^i,\pi^{-i}}}\Big[r_i^{\pi^i,\pi^{-i}}(s') + \lambda_i\cdot H\big(\pi^i(\cdot\,|\,s')\big)\Big],
\end{aligned}$$

which gives

$$\begin{aligned}
&(1-\gamma)\cdot\big|[\mathcal{P}^{a^i,\pi^{-i}}\circ V_{(i)}^{\pi^i,\pi^{-i}}](s) - [\mathcal{P}^{a^i,\pi_*^{-i}}\circ V_{(i)}^{\pi^i,\pi_*^{-i}}](s)\big| \qquad \text{(E.7)}\\
&\quad= \Big|\mathbb{E}_{s'\sim\rho_{s,a^i,\pi^{-i}}^{\pi^i,\pi^{-i}}}\Big[r_i^{\pi^i,\pi^{-i}}(s') + \lambda_i\cdot H\big(\pi^i(\cdot\,|\,s')\big)\Big] - \mathbb{E}_{s'\sim\rho_{s,a^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}\Big[r_i^{\pi^i,\pi_*^{-i}}(s') + \lambda_i\cdot H\big(\pi^i(\cdot\,|\,s')\big)\Big]\Big|\\
&\quad\le \Big|\mathbb{E}_{s'\sim\rho_{s,a^i,\pi^{-i}}^{\pi^i,\pi^{-i}}}\Big[r_i^{\pi^i,\pi^{-i}}(s') + \lambda_i\cdot H\big(\pi^i(\cdot\,|\,s')\big)\Big] - \mathbb{E}_{s'\sim\rho_{s,a^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}\Big[r_i^{\pi^i,\pi^{-i}}(s') + \lambda_i\cdot H\big(\pi^i(\cdot\,|\,s')\big)\Big]\Big|\\
&\qquad+ \iota_i\cdot\mathbb{E}_{s'\sim\rho_{s,a^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}\Big[\big\|\pi^{-i}(\cdot\,|\,s') - \pi_*^{-i}(\cdot\,|\,s')\big\|_1\Big]\\
&\quad\le \big(1+\lambda_i\cdot\log|\mathcal{A}^i|\big)\cdot\underbrace{\|\rho_{s,a^i,\pi^{-i}}^{\pi^i,\pi^{-i}} - \rho_{s,a^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}\|_{L_1(\mathcal{S})}}_{\text{(III)}} + \iota_i\cdot\underbrace{\mathbb{E}_{s'\sim\rho_{s,a^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s')\,\big\|\,\pi^{-i}(\cdot\,|\,s')\big)^{1/2}\Big]}_{\text{(IV)}},
\end{aligned}$$

where the first inequality follows from (4.9) of Assumption 4.3, and $\rho_{s,a^i,\pi^{-i}}^{\pi^i,\pi^{-i}}$ is defined in (4.5).

By (E.1), we have

$$\begin{aligned}
\text{(III)} &= (1-\gamma)\cdot\bigg\|\bigg[\sum_{t=0}^{\infty}\gamma^t\cdot(\mathcal{P}^{\pi^i,\pi^{-i}})^t\bigg]\circ\mathcal{P}^{a^i,\pi^{-i}}\circ\delta_s-\bigg[\sum_{t=0}^{\infty}\gamma^t\cdot(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\bigg]\circ\mathcal{P}^{a^i,\pi_*^{-i}}\circ\delta_s\bigg\|_{L_1(\mathcal{S})}\\
&\le(1-\gamma)\cdot\bigg\|\bigg[\sum_{t=0}^{\infty}\gamma^t\cdot\big[(\mathcal{P}^{\pi^i,\pi^{-i}})^t-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\big]\bigg]\circ\mathcal{P}^{a^i,\pi^{-i}}\circ\delta_s\bigg\|_{L_1(\mathcal{S})}\\
&\quad+\bigg\|\bigg[(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\cdot(\mathcal{P}^{\pi^i,\pi^{-i}})^t\bigg]\circ(\mathcal{P}^{a^i,\pi^{-i}}-\mathcal{P}^{a^i,\pi_*^{-i}})\circ\delta_s\bigg\|_{L_1(\mathcal{S})}. \qquad \text{(E.8)}
\end{aligned}$$

Since $\|\mathcal{P}^{\pi^i,\pi^{-i}}\|_{\mathrm{op}}\le 1$, we have

$$\bigg\|(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\cdot(\mathcal{P}^{\pi^i,\pi^{-i}})^t\bigg\|_{\mathrm{op}}\le(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\cdot\|\mathcal{P}^{\pi^i,\pi^{-i}}\|_{\mathrm{op}}^t\le 1,$$

plugging which and Lemma E.1 into (E.8) gives

$$\begin{aligned}
\text{(III)} &\le(1-\gamma)\cdot\bigg\|\sum_{t=0}^{\infty}\gamma^t\cdot\big[(\mathcal{P}^{\pi^i,\pi^{-i}})^t-(\mathcal{P}^{\pi^i,\pi_*^{-i}})^t\big]\bigg\|_{\mathrm{op}}\cdot\|\mathcal{P}^{a^i,\pi^{-i}}\|_{\mathrm{op}}\cdot\|\delta_s\|_{L_1(\mathcal{S})}\\
&\quad+\bigg\|(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t\cdot(\mathcal{P}^{\pi^i,\pi^{-i}})^t\bigg\|\cdot\|\mathcal{P}^{a^i,\pi^{-i}}-\mathcal{P}^{a^i,\pi_*^{-i}}\|_{\mathrm{op}}\cdot\|\delta_s\|_{L_1(\mathcal{S})}\\
&\le\frac{\iota_i}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}. \qquad \text{(E.9)}
\end{aligned}$$

By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\text{(IV)} &=\mathbb{E}_{s'\sim\nu^*}\bigg[\frac{\mathrm{d}\rho_{s,a^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}{\mathrm{d}\nu^*}\cdot\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)^{1/2}\bigg]\\
&\le\mathbb{E}_{\nu^*}\bigg[\bigg|\frac{\mathrm{d}\rho_{s,a^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}{\mathrm{d}\nu^*}\bigg|^2\bigg]^{1/2}\cdot\mathbb{E}_{s'\sim\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s')\,\big\|\,\pi^{-i}(\cdot\,|\,s')\big)\Big]^{1/2}\\
&\le\zeta\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}, \qquad \text{(E.10)}
\end{aligned}$$

where the last inequality follows from Assumption 4.2 and the Pinsker's inequality. Plugging (E.9) and (E.10) into (E.7), we obtain for all $s\in\mathcal{S}$ and $a^i\in\mathcal{A}^i$ that,

$$\begin{aligned}
&\big|[\mathcal{P}^{a^i,\pi^{-i}}\circ V_{(i)}^{\pi^i,\pi^{-i}}](s)-[\mathcal{P}^{a^i,\pi_*^{-i}}\circ V_{(i)}^{\pi^i,\pi_*^{-i}}](s)\big|\\
&\quad\le\frac{V_{(i)}^{\max}+\zeta}{1-\gamma}\cdot\iota_i\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2},
\end{aligned}$$

which further gives

$$\begin{aligned}
\text{(II)} &\le\mathbb{E}_{\nu^*}\Big[\big\|[\mathcal{P}^{\cdot,\pi^{-i}}\circ V_{(i)}^{\pi^i,\pi^{-i}}](s)-[\mathcal{P}^{\cdot,\pi_*^{-i}}\circ V_{(i)}^{\pi^i,\pi_*^{-i}}](s)\big\|_\infty\cdot\big\|\pi^i(\cdot\,|\,s)-\pi_*^i(\cdot\,|\,s)\big\|_1\Big] \qquad \text{(E.11)}\\
&\le\frac{V_{(i)}^{\max}+\zeta}{1-\gamma}\cdot\iota_i\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}\cdot\mathbb{E}_{\nu^*}\Big[\big\|\pi^i(\cdot\,|\,s)-\pi_*^i(\cdot\,|\,s)\big\|_1\Big]\\
&\le\frac{\sqrt{2}\cdot(V_{(i)}^{\max}+\zeta)\cdot\iota_i}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{i}(\cdot\,|\,s)\,\big\|\,\pi^{i}(\cdot\,|\,s)\big)\Big]^{1/2},
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the last inequality follows from the Pinsker's inequality. Alternatively, we can also write

$$\begin{aligned}
\text{(II)} &= \Big|\mathbb{E}_{\nu^*}\Big[\Big\langle\big[(\mathcal{P}^{\pi^i,\cdot} - \mathcal{P}^{\pi_*^i,\cdot}) \circ V_{(i)}^{\pi^i,\pi^{-i}}\big](s), \pi^{-i}(\cdot\,|\,s) - \pi_*^{-i}(\cdot\,|\,s)\Big\rangle\Big] \\
&\qquad + \mathbb{E}_{\nu^*}\Big[\Big\langle\big[\mathcal{P}^{\cdot,\pi_*^{-i}} \circ (V_{(i)}^{\pi^i,\pi^{-i}} - V_{(i)}^{\pi^i,\pi_*^{-i}})\big](s), \pi^i(\cdot\,|\,s) - \pi_*^i(\cdot\,|\,s)\Big\rangle\Big]\Big| \\
&\le \underbrace{\Big|\mathbb{E}_{\nu^*}\Big[\Big\langle\big[(\mathcal{P}^{\pi^i,\cdot} - \mathcal{P}^{\pi_*^i,\cdot}) \circ V_{(i)}^{\pi^i,\pi^{-i}}\big](s), \pi^{-i}(\cdot\,|\,s) - \pi_*^{-i}(\cdot\,|\,s)\Big\rangle\Big]\Big|}_{\text{(V)}} \\
&\qquad + \underbrace{\Big|\mathbb{E}_{\nu^*}\Big[\big[(\mathcal{P}^{\pi^i,\pi_*^{-i}} - \mathcal{P}^{\pi_*^i,\pi_*^{-i}}) \circ (V_{(i)}^{\pi^i,\pi^{-i}} - V_{(i)}^{\pi^i,\pi_*^{-i}})\big](s)\Big]\Big|}_{\text{(VI)}}. \qquad \text{(E.12)}
\end{aligned}$$

By (4.8) of Assumption 4.3, we have for all $s \in \mathcal{S}$ and $a^{-i} \in \mathcal{A}^{-i}$ that,

$$\begin{aligned}
\Big|\big[(\mathcal{P}^{\pi^i,a^{-i}} - \mathcal{P}^{\pi_*^i,a^{-i}}) \circ V_{(i)}^{\pi^i,\pi^{-i}}\big](s)\Big| &\le V_{(i)}^{\max} \cdot \|\mathcal{P}^{\pi^i,a^{-i}} - \mathcal{P}^{\pi_*^i,a^{-i}}\|_{\mathrm{op}} \\
&\le V_{(i)}^{\max} \cdot \iota_{-i} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s) \,\big\|\, \pi^i(\cdot\,|\,s)\big)\Big]^{1/2},
\end{aligned}$$

which gives

$$\begin{aligned}
\text{(V)} &\le \mathbb{E}_{\nu^*}\Big[\Big\|\big[(\mathcal{P}^{\pi^i,\cdot} - \mathcal{P}^{\pi_*^i,\cdot}) \circ V_{(i)}^{\pi^i,\pi^{-i}}\big](s)\Big\|_\infty \cdot \big\|\pi^{-i}(\cdot\,|\,s) - \pi_*^{-i}(\cdot\,|\,s)\big\|_1\Big] \qquad \text{(E.13)} \\
&\le \sqrt{2} \cdot V_{(i)}^{\max} \cdot \iota_{-i} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s) \,\big\|\, \pi^i(\cdot\,|\,s)\big)\Big]^{1/2} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s) \,\big\|\, \pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}. \\
&\le \frac{\sqrt{2} \cdot V_{(i)}^{\max}}{1-\gamma} \cdot \iota_{-i} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s) \,\big\|\, \pi^i(\cdot\,|\,s)\big)\Big]^{1/2} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s) \,\big\|\, \pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}.
\end{aligned}$$

Here the second inequality follows from the Pinsker's inequality.

Meanwhile, we have

$$\begin{aligned}
\text{(VI)} &\le \mathbb{E}_{\nu^*}\Big[\Big|\big[(\mathcal{P}^{\pi^i,\pi_*^{-i}} - \mathcal{P}^{\pi_*^i,\pi_*^{-i}}) \circ (V_{(i)}^{\pi^i,\pi^{-i}} - V_{(i)}^{\pi^i,\pi_*^{-i}})\big](s)\Big|\Big] \qquad \text{(E.14)} \\
&\le \|\mathcal{P}^{\pi^i,\pi_*^{-i}} - \mathcal{P}^{\pi_*^i,\pi_*^{-i}}\|_{\mathrm{op}} \cdot \mathbb{E}_{\nu^*}\Big[\big|V_{(i)}^{\pi^i,\pi^{-i}}(s) - V_{(i)}^{\pi^i,\pi_*^{-i}}(s)\big|\Big] \\
&\le \iota_{-i} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s) \,\big\|\, \pi^i(\cdot\,|\,s)\big)\Big]^{1/2} \cdot \mathbb{E}_{\nu^*}\Big[\big|V_{(i)}^{\pi^i,\pi^{-i}}(s) - V_{(i)}^{\pi^i,\pi_*^{-i}}(s)\big|\Big].
\end{aligned}$$

By the same argument in the proof of Lemma C.3, we have

$$V_{(i)}^{\pi^i,\pi^{-i}}(s) - V_{(i)}^{\pi^i,\pi_*^{-i}}(s) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{s' \sim \rho_s^{\pi^i,\pi_*^{-i}}}\Big[\big\langle \widetilde{Q}_{(i)}^{\pi^i,\pi^{-i}}(s',\cdot), \pi_*^{-i}(\cdot\,|\,s') - \pi^{-i}(\cdot\,|\,s')\big\rangle\Big],$$

which gives

$$\begin{aligned}
\mathbb{E}_{\nu^*}\Big[\big|V_{(i)}^{\pi^i,\pi^{-i}}(s) - V_{(i)}^{\pi^i,\pi_*^{-i}}(s)\big|\Big] &\le \frac{1}{1-\gamma} \cdot \mathbb{E}_{s' \sim \rho_{\nu^*,\pi_*^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}\Big[\big\|\widetilde{Q}_{(i)}^{\pi^i,\pi^{-i}}(s',\cdot)\big\|_\infty \cdot \big\|\pi_*^{-i}(\cdot\,|\,s') - \pi^{-i}(\cdot\,|\,s')\big\|_1\Big] \\
&\le \frac{Q_{(i)}^{\max}}{1-\gamma} \cdot \mathbb{E}_{s' \sim \rho_{\nu^*,\pi_*^i,\pi_*^{-i}}^{\pi^i,\pi_*^{-i}}}\Big[\big\|\pi_*^{-i}(\cdot\,|\,s') - \pi^{-i}(\cdot\,|\,s')\big\|_1\Big] \\
&\le \frac{\sqrt{2} \cdot Q_{(i)}^{\max} \cdot \zeta}{1-\gamma} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s) \,\big\|\, \pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}, \qquad \text{(E.15)}
\end{aligned}$$

where the last inequality follows from the same arguments in (E.10). Taking (E.15) into (E.14), we obtain

$$\text{(VI)} \le \frac{\sqrt{2} \cdot Q_{(i)}^{\max} \cdot \zeta}{1-\gamma} \cdot \iota_{-i} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s) \,\big\|\, \pi^i(\cdot\,|\,s)\big)\Big]^{1/2} \cdot \mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s) \,\big\|\, \pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2},$$

pugging which and (E.13) into (E.12) gives

$$(\mathrm{II}) \leq \frac{\sqrt{2}\cdot(V_{(i)}^{\max}+Q_{(i)}^{\max}\cdot\zeta)}{1-\gamma}\cdot\iota_{-i}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)\Big]^{1/2}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}. \tag{E.16}$$

Combining (E.11) and (E.16), we obtain

$$(\mathrm{II}) \leq \frac{V_{(i)}^{\max}+Q_{(i)}^{\max}\cdot\zeta}{1-\gamma}\cdot\sqrt{2\iota_i\cdot\iota_{-i}}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)\Big]^{1/2}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}. \tag{E.17}$$

Finally, taking (E.6) and (E.17) in to (E.3), we obtain

$$\begin{aligned}
&\Big|\mathbb{E}_{\nu^*}\Big[\big\langle \widetilde{Q}_{(i)}^{\pi^i,\pi^{-i}}(s,\cdot\,)-\widetilde{Q}_{(i)}^{\pi^i,\pi_*^{-i}}(s,\cdot\,),\pi^i(\cdot\,|\,s)-\pi_*^i(\cdot\,|\,s)\big\rangle\Big]\Big| \\
&\quad\leq \sqrt{2\iota_i\cdot\iota_{-i}}\cdot\bigg\{\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)^{1/2}\cdot\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)^{1/2}\Big] \\
&\qquad+\frac{V_{(i)}^{\max}+Q_{(i)}^{\max}\cdot\zeta}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)\Big]^{1/2}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^{-i}(\cdot\,|\,s)\,\big\|\,\pi^{-i}(\cdot\,|\,s)\big)\Big]^{1/2}\bigg\},
\end{aligned}$$

which concludes the proof of Lemma C.2. □

F Proof of Lemma C.3

Proof. The proof extends that of Lemma 6.1 in Kakade & Langford (2002) to zero-sum Markov games. By the definition of $V_{(i)}^{\pi^i,\pi^{-i}}(s)$ in (2.1), we have

$$\begin{aligned}
&V_{(i)}^{\pi_*^i,\pi_*^{-i}}(s) \qquad\qquad (\text{F.1})\\
&=\mathbb{E}_{\pi_*^i,\pi_*^{-i}}\bigg[\sum_{t=0}^{\infty}\gamma^t\cdot\big(r_{(i)}^{\pi_*^i,\pi_*^{-i}}(s_t,a^i,a^{-i})+V_{(i)}^{\pi^i,\pi_*^{-i}}(s_t)-V_{(i)}^{\pi^i,\pi_*^{-i}}(s_t)\big)\,\bigg|\,s_0=s\bigg]\\
&=\mathbb{E}_{\pi_*^i,\pi_*^{-i}}\bigg[\sum_{t=0}^{\infty}\gamma^t\cdot\big(r_{(i)}^{\pi_*^i,\pi_*^{-i}}(s_t,a^i,a^{-i})+\gamma\cdot V_{(i)}^{\pi^i,\pi_*^{-i}}(s_{t+1})-V_{(i)}^{\pi^i,\pi_*^{-i}}(s_t)\big)\,\bigg|\,s_0=s\bigg]+V_{(i)}^{\pi^i,\pi_*^{-i}}(s).
\end{aligned}$$

By the definition of $Q_{(i)}$-function, we have

$$Q_{(i)}^{\pi^i,\pi_*^{-i}}(s,a^i,a^{-i})=r(s,a^i,a^{-i})+\gamma\cdot\mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a^i,a^{-i})}\big[V_{(i)}^{\pi^i,\pi_*^{-i}}(s')\big],$$

which gives

$$\begin{aligned}
&\mathbb{E}_{\pi_*^i,\pi_*^{-i}}\Big[r_{(i)}^{\pi_*^i,\pi_*^{-i}}(s)+\gamma\cdot\mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a^i,a^{-i})}\big[V_{(i)}^{\pi^i,\pi_*^{-i}}(s')\big]-V_{(i)}^{\pi^i,\pi_*^{-i}}(s)\Big] \qquad (\text{F.2})\\
&\quad=\mathbb{E}_{\pi_*^i,\pi_*^{-i}}\Big[r(s,a^1,a^{-i})-\lambda_i\cdot\log\pi_*^i(a^i\,|\,s)+\gamma\cdot\mathbb{E}_{s'\sim\mathcal{P}(\cdot\,|\,s,a^i,a^{-i})}\big[V_{(i)}^{\pi^i,\pi_*^{-i}}(s')\big]-V_{(i)}^{\pi^i,\pi_*^{-i}}(s)\Big]\\
&\quad=\mathbb{E}_{\pi_*^i,\pi_*^{-i}}\big[Q_{(i)}^{\pi^i,\pi_*^{-i}}(s,a^i,a^{-i})-V_{(i)}^{\pi^i,\pi_*^{-i}}(s)\big]+\lambda_i\cdot H\big(\pi_*^i(\cdot\,|\,s)\big)\\
&\quad=\big\langle\widetilde{Q}_{(i)}^{\pi^i,\pi_*^{-i}}(s,\cdot\,),\pi_*^i(\cdot\,|\,s)-\pi^i(\cdot\,|\,s)\big\rangle+\lambda_i\cdot H\big(\pi_*^i(\cdot\,|\,s)\big)-\lambda_i\cdot H\big(\pi^i(\cdot\,|\,s)\big).
\end{aligned}$$

Here the last equality follows from

$$\begin{aligned}
V_{(i)}^{\pi^i,\pi_*^{-i}}(s)&=\mathbb{E}_{\pi^i,\pi_*^{-i}}\big[Q_{(i)}^{\pi^i,\pi_*^{-i}}(s,a^i,a^{-i})-\lambda_i\cdot\log\pi^i(a^i\,|\,s)\big]\\
&=\big\langle\widetilde{Q}_{(i)}^{\pi^i,\pi_*^{-i}}(s,\cdot\,),\pi^i(\cdot\,|\,s)\big\rangle+\lambda_i\cdot H\big(\pi^i(\cdot\,|\,s)\big).
\end{aligned}$$

For the entropy terms in (F.2), we have

$$\begin{aligned}
H\big(\pi_*^i(\cdot\,|\,s)\big)&=\big\langle-\log\pi_*^i(\cdot\,|\,s),\pi_*^i(\cdot\,|\,s)\big\rangle\\
&=\big\langle-\log\pi^i(\cdot\,|\,s),\pi_*^i(\cdot\,|\,s)\big\rangle-\Big\langle\log\big(\pi_*^i(\cdot\,|\,s)/\pi^i(\cdot\,|\,s)\big),\pi_*^i(\cdot\,|\,s)\Big\rangle\\
&=\big\langle-\log\pi^i(\cdot\,|\,s),\pi_*^i(\cdot\,|\,s)\big\rangle-\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big),
\end{aligned}$$

which gives

$$H\big(\pi_*^i(\cdot\,|\,s)\big) - H\big(\pi^i(\cdot\,|\,s)\big) = \big\langle -\log\pi^i(\cdot\,|\,s), \pi_*^i(\cdot\,|\,s) - \pi^i(\cdot\,|\,s)\big\rangle - \mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big). \tag{F.3}$$

Taking (F.2) and (F.3) into (F.1), we have

$$\begin{aligned} &V_{(i)}^{\pi_*^i,\pi_*^{-i}}(s) - V_{(i)}^{\pi^i,\pi_*^{-i}}(s) \\ &= \mathbb{E}_{\pi_*^i,\pi_*^{-i}}\bigg[\sum_{t=0}^{\infty}\gamma^t\cdot\Big(\big\langle\widetilde{Q}_{(i)}^{\pi^1,\pi_*^{-i}}(s_t,\,\cdot\,) - \lambda_i\cdot\log\pi^i(\cdot\,|\,s_t), \pi_*^i(\cdot\,|\,s_t) - \pi^i(\cdot\,|\,s_t)\big\rangle \\ &\quad - \lambda_i\cdot\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s_t)\,\big\|\,\pi^i(\cdot\,|\,s_t)\big)\Big)\,\bigg|\,s_0 = s\bigg] \\ &= \frac{1}{1-\gamma}\cdot\mathbb{E}_{s'\sim\rho_s^{\pi_*^i,\pi_*^{-i}}}\Big[\big\langle\widetilde{Q}_{(i)}^{\pi^1,\pi_*^{-i}}(s',\,\cdot\,) - \lambda_i\cdot\log\pi^i(\cdot\,|\,s'), \pi_*^i(\cdot\,|\,s') - \pi^i(\cdot\,|\,s')\big\rangle \\ &\quad - \lambda_i\cdot\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s')\,\big\|\,\pi^i(\cdot\,|\,s')\big)\Big], \end{aligned} \tag{F.4}$$

where $\rho_s^{\pi_*^i,\pi_*^{-i}}$ is defined in (4.4) as the visitation measure of the policy pair $[\pi_*^i;\pi_*^{-i}]$ starting from state s. Taking $\mathbb{E}_{\nu^*}[\,\cdot\,]$ on both sides of (F.4) and recalling the definition of $\mathcal{J}_{(i)}(\pi^i,\pi^{-i})$ in (4.3), we have

$$\begin{aligned} &\mathcal{J}_{(i)}(\pi_*^i,\pi_*^{-i}) - \mathcal{J}_{(i)}(\pi^i,\pi_*^{-i}) + \frac{\lambda_i}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi^i(\cdot\,|\,s)\big)\Big] \\ &\quad = \frac{1}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\Big[\big\langle\widetilde{Q}_{(i)}^{\pi^i,\pi_*^{-i}}(s,\,\cdot\,) - \lambda_i\cdot\log\pi^i(\cdot\,|\,s), \pi_*^i(\cdot\,|\,s) - \pi^i(\cdot\,|\,s)\big\rangle\Big]. \end{aligned}$$

Here we use the fact that $\mathbb{E}_{s'\sim\nu^*}[\,\cdot\,] = \mathbb{E}_{s'\sim\rho_s^{\pi_*^i,\pi_*^{-i}},s\sim\nu^*}[\,\cdot\,]$. Thus,we finish the proof of Lemma C.3. □

G PROOF OF LEMMA C.4

Proof. First, we have for any $s\in\mathcal{S}$ that,

$$\begin{aligned} &\mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi_{\theta_t}^i(\cdot\,|\,s)\big) - \mathrm{KL}\big(\pi_*^i(\cdot\,|\,s)\,\big\|\,\pi_{\theta_{t+1}}^i(\cdot\,|\,s)\big) \\ &= \mathrm{KL}\big(\pi_{\theta_{t+1}}^i(\cdot\,|\,s)\,\big\|\,\pi_{\theta_t}^i(\cdot\,|\,s)\big) + \bigg\langle\log\bigg[\frac{\pi_{\theta_{t+1}}^i(\cdot\,|\,s)}{\pi_{\theta_t}^i(\cdot\,|\,s)}\bigg], \pi_{\theta_t}^i(\cdot\,|\,s) - \pi_{\theta_{t+1}}^i(\cdot\,|\,s)\bigg\rangle \\ &\quad + \bigg\langle\log\bigg[\frac{\pi_{\theta_{t+1}}^i(\cdot\,|\,s)}{\pi_{\theta_t}^i(\cdot\,|\,s)}\bigg], \pi_*^i(\cdot\,|\,s) - \pi_{\theta_t}^i(\cdot\,|\,s)\bigg\rangle \\ &= \mathrm{KL}\big(\pi_{\theta_{t+1}}^i(\cdot\,|\,s)\,\big\|\,\pi_{\theta_t}^i(\cdot\,|\,s)\big) + \bigg\langle\log\bigg[\frac{\pi_{\theta_{t+1}}^i(\cdot\,|\,s)}{\pi_{\theta_t}^i(\cdot\,|\,s)}\bigg], \pi_{\theta_t}^i(\cdot\,|\,s) - \pi_{\theta_{t+1}}^i(\cdot\,|\,s)\bigg\rangle \\ &\quad + \bigg\langle\log\bigg[\frac{\pi_{\theta_{t+1}}^i(\cdot\,|\,s)}{\pi_{\theta_t}^i(\cdot\,|\,s)}\bigg] - \alpha_{t,(i)}\cdot\big(\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_{\theta_t}^{-i}}(s,\,\cdot\,) - \lambda_i\cdot\log\pi_{\theta_t}^i(\cdot\,|\,s)\big), \pi_*^i(\cdot\,|\,s) - \pi_{\theta_t}^i(\cdot\,|\,s)\bigg\rangle \\ &\quad + \alpha_{t,(i)}\cdot\big\langle\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_{\theta_t}^{-i}}(s,\,\cdot\,) - \lambda_i\cdot\log\pi_{\theta_t}^i(\cdot\,|\,s), \pi_*^i(\cdot\,|\,s) - \pi_{\theta_t}^i(\cdot\,|\,s)\big\rangle. \end{aligned} \tag{G.1}$$

Recall that

$$\begin{aligned} \overline{\pi}_{t+1}^i(\cdot\,|\,s) &\propto \exp\big\{(1-\lambda_i\alpha_{t,(i)})\cdot\mathcal{E}_{\theta_t,(i)}(s,\cdot) + \alpha_{t,(i)}\cdot\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_{\theta_t}^{-i}}(s,\,\cdot\,)\big\} \\ &\propto \exp\Big\{\mathcal{E}_{\theta_t,(i)}(s,\cdot) + \alpha_{t,(i)}\cdot\big(\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_{\theta_t}^{-i}}(s,\,\cdot\,) - \lambda_i\cdot\log\pi_{\theta_t}^i(\cdot\,|\,s)\big)\Big\}. \end{aligned}$$

Let

$$Z_{t+1,(i)}(s) = \sum_{a^i\in\mathcal{A}^i}\exp\Big\{\mathcal{E}_{\theta_t,(i)}(s,a^i) + \alpha_{t,(i)}\cdot\big(\widetilde{Q}_{(i)}^{\pi_{\theta_t}^i,\pi_{\theta_t}^{-i}}(s,a^i) - \lambda_i\cdot\log\pi_{\theta_t}^i(a^i\,|\,s)\big)\Big\},$$

and

$$Z_{\theta_t,(i)}(s) = \sum_{a^i \in \mathcal{A}^i} \exp\bigl\{\mathcal{E}_{\theta_{t+1},(i)}(s, a^i)\bigr\}.$$

where are only dependent on the state s. It can be verified that $\langle \log Z_{\theta_t,(i)}(s), \pi^i(\cdot \,|\, s) - \pi^{i'}(\cdot \,|\, s)\rangle = \langle \log Z_{t,(i)}(s), \pi^i(\cdot \,|\, s) - \pi^{i'}(\cdot \,|\, s)\rangle = 0$ for all t, π^i, and $\pi^{i'}$, which implies that, on the right-hand-side of (G.1),

$$\begin{aligned}
&\Bigl\langle \log \pi^i_{\theta_t}(\cdot \,|\, s) + \alpha_{t,(i)} \cdot \bigl(\widetilde{Q}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}_{(i)}(s, \,\cdot\,) - \lambda_i \cdot \log \pi^i_{\theta_t}(\cdot \,|\, s)\bigr), \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\Bigr\rangle \\
&\quad = \Bigl\langle \mathcal{E}_{\theta_t,(i)}(s, \cdot) - \log Z_{\theta_t,(i)}(s) + \alpha_{t,(i)} \cdot \bigl(\widetilde{Q}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}_{(i)}(s, \,\cdot\,) - \lambda_i \cdot \log \pi^i_{\theta_t}(\cdot \,|\, s)\bigr), \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\Bigr\rangle \\
&\quad = \Bigl\langle \mathcal{E}_{\theta_t,(i)}(s, \cdot) + \alpha_{t,(i)} \cdot \bigl(\widetilde{Q}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}_{(i)}(s, \,\cdot\,) - \lambda_i \cdot \log \pi^i_{\theta_t}(\cdot \,|\, s)\bigr) - \log Z_{t+1,(i)}(s), \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\Bigr\rangle \\
&\quad = \bigl\langle \log \overline{\pi}^i_{t+1}(\cdot \,|\, s), \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\bigr\rangle, \qquad \text{(G.2)}
\end{aligned}$$

and

$$\begin{aligned}
&\Bigl\langle \log\Bigl[\frac{\pi^i_{\theta_{t+1}}(\cdot \,|\, s)}{\pi^i_{\theta_t}(\cdot \,|\, s)}\Bigr], \pi^i_{\theta_t}(\cdot \,|\, s) - \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\Bigr\rangle \\
&\quad = \bigl\langle \mathcal{E}_{\theta_{t+1},(i)}(s, \cdot) - \mathcal{E}_{\theta_t,(i)}(s, \cdot), \pi^i_{\theta_t}(\cdot \,|\, s) - \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\bigr\rangle \\
&\qquad - \bigl\langle \log Z_{\theta_{t+1},(i)}(s), \pi^i_{\theta_t}(\cdot \,|\, s) - \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\bigr\rangle + \bigl\langle \log Z_{\theta_t,(i)}(s), \pi^i_{\theta_t}(\cdot \,|\, s) - \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\bigr\rangle \\
&\quad = \bigl\langle \mathcal{E}_{\theta_{t+1},(i)}(s, \cdot) - \mathcal{E}_{\theta_t,(i)}(s, \cdot), \pi^i_{\theta_t}(\cdot \,|\, s) - \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\bigr\rangle. \qquad \text{(G.3)}
\end{aligned}$$

Plugging (G.2) and (G.3) into (G.1), we obtain

$$\begin{aligned}
&\mathrm{KL}\bigl(\pi^i_*(\cdot \,|\, s) \,\big\|\, \pi^i_{\theta_t}(\cdot \,|\, s)\bigr) - \mathrm{KL}\bigl(\pi^i_*(\cdot \,|\, s) \,\big\|\, \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\bigr) \qquad \text{(G.4)} \\
&= \Bigl\langle \log\Bigl[\frac{\pi^i_{\theta_{t+1}}(\cdot \,|\, s)}{\overline{\pi}^i_{t+1}(\cdot \,|\, s)}\Bigr], \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\Bigr\rangle + \bigl\langle \mathcal{E}_{\theta_{t+1},(i)}(s, \cdot) - \mathcal{E}_{\theta_t,(i)}(s, \cdot), \pi^i_{\theta_t}(\cdot \,|\, s) - \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\bigr\rangle \\
&\qquad + \alpha_{t,(i)} \cdot \bigl\langle \widetilde{Q}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}_{(i)}(s, \,\cdot\,) - \lambda_i \cdot \log \pi^i_{\theta_t}(\cdot \,|\, s), \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\bigr\rangle + \mathrm{KL}\bigl(\pi^i_{\theta_{t+1}}(\cdot \,|\, s) \,\big\|\, \pi^i_{\theta_t}(\cdot \,|\, s)\bigr) \\
&\geq \Bigl\langle \log\Bigl[\frac{\pi^i_{\theta_{t+1}}(\cdot \,|\, s)}{\overline{\pi}^i_{t+1}(\cdot \,|\, s)}\Bigr], \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\Bigr\rangle - \bigl\|\mathcal{E}_{\theta_{t+1},(i)}(s, \cdot) - \mathcal{E}_{\theta_t,(i)}(s, \cdot)\bigr\|_\infty \cdot \bigl\|\pi^i_{\theta_t}(\cdot \,|\, s) - \pi^i_{\theta_{t+1}}(\cdot \,|\, s)\bigr\|_1 \\
&\qquad + \alpha_{t,(i)} \cdot \bigl\langle \widetilde{Q}^{\pi^i_{\theta_t}, \pi^{-i}_{\theta_t}}_{(i)}(s, \,\cdot\,) - \lambda_i \cdot \log \pi^i_{\theta_t}(\cdot \,|\, s), \pi^i_*(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\bigr\rangle + 1/2 \cdot \bigl\|\pi^i_{\theta_{t+1}}(\cdot \,|\, s) - \pi^i_{\theta_t}(\cdot \,|\, s)\bigr\|_1^2,
\end{aligned}$$

where in the last inequality we use the Cauchy-Schwartz inequality and the Pinsker's inequality. Rearranging the terms in (G.4), we finish the proof of Lemma C.4. □

H PROOF OF LEMMA B.2

Proof. We prove the lemma by induction. First, since $|\widehat{Q}^{\pi^i_{\theta_0}, \pi^{-i}_{\theta_0}}_{(i)}(s, a^i)| \leq Q^{\max}_{(i)}$, we have

$$\begin{aligned}
\bigl|\widehat{\mathcal{E}}_{1,(i)}(s, a^i)\bigr| &= \bigl|(1 - \lambda_i \alpha_{0,(i)}) \cdot \mathcal{E}_{\theta_0,(i)}(s, a^i) + \alpha_{0,(i)} \cdot \widehat{Q}^{\pi^i_{\theta_0}, \pi^{-i}_{\theta_0}}_{(i)}(s, a^i)\bigr| \\
&= \alpha_{0,(i)} \cdot \bigl|\widehat{Q}^{\pi^i_{\theta_0}, \pi^{-i}_{\theta_0}}_{(i)}(s, a^i)\bigr| \leq Q^{\max}_{(i)}/(\lambda_i - M_{(i)}),
\end{aligned}$$

where the last inequality follows from $\alpha_{0,(i)} = 1/(\lambda_i - M_i)$. This means that setting $|\mathcal{E}_{\theta_1,(i)}(s, a^i)| \leq Q^{\max}_{(i)}/(\lambda_i - M_i)$ covers the range of $\widehat{\mathcal{E}}_{1,(i)}(s, a^i)$.

Now suppose that $|\widehat{\mathcal{E}}_{\theta_t,(i)}(s, a^i)| \leq Q^{\max}_{(i)}/(\lambda_i - M_i)$. By $\lambda_i \geq 2M_i$, we have for all $t \geq 1$ that,

$$\lambda_i \alpha_{t,(i)} = \frac{\lambda_i}{(t+1) \cdot (\lambda_i - M_i)} \leq \frac{\lambda_i}{2(\lambda_i - M_i)} \leq 1.$$

Thus, we have that, for all $t \geq 1$ and $(s, a^i) \in \mathcal{S} \times \mathcal{A}^i$,

$$\begin{aligned}
\big|\widehat{\mathcal{E}}_{t+1,(i)}(s,a^i)\big| &= \big|(1-\lambda_i\alpha_{t,(i)})\cdot\mathcal{E}_{\theta_t,(i)}(s,a^i) + \alpha_{t,(i)}\cdot\widehat{Q}_{(i)}^{\pi^i_{\theta_t},\pi^{-i}_{\theta_t}}(s,a^i)\big| \\
&\leq \left(1-\frac{\lambda_i}{(t+1)\cdot(\lambda_i-M_i)}\right)\cdot\big|\mathcal{E}_{\theta_t,(i)}(s,a^i)\big| + \frac{1}{(t+1)\cdot(\lambda_i-M_i)}\cdot\big|\widehat{Q}_{(i)}^{\pi^i_{\theta_t},\pi^{-i}_{\theta_t}}(s,a^i)\big| \\
&\leq \frac{t\cdot(\lambda_i-M_i)-M_i}{(t+1)\cdot(\lambda_i-M_i)}\cdot\frac{Q^{\max}_{(i)}}{\lambda_i-M_i} + \frac{Q^{\max}_{(i)}}{(t+1)\cdot(\lambda_i-M_i)} \\
&\leq \frac{t}{t+1}\cdot\frac{Q^{\max}_{(i)}}{\lambda_i-M_i} + \frac{Q^{\max}_{(i)}}{(t+1)\cdot(\lambda_i-M_i)} = \frac{Q^{\max}_{(i)}}{\lambda_i-M_i},
\end{aligned}$$

which means that setting $|\mathcal{E}_{\theta_{t+1},(i)}(s,a^i)| \leq Q^{\max}_{(i)}/(\lambda_i - M_i)$ covers the range of $\widehat{\mathcal{E}}_{t+1,(i)}(s,a^i)$. By induction, we conclude the proof of Lemma B.2. □

I IMBALANCED INFLUENCE

When the two players have imbalanced influence to the game (without loss of generality, we assume Player 2 has a dominant influence to the game), we let $\iota_1/\iota_2 = z$. In this case, we can replace (C.6) by

$$\begin{aligned}
\text{(II)} &\leq \sqrt{2z}\cdot\iota_2\cdot\Big\{\mathbb{E}_{\nu^*}\big[\mathrm{KL}^{1/2}_{t,(i)}(s)\cdot\mathrm{KL}^{1/2}_{t,(i)}(s)\big] \\
&\qquad + \frac{V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta}{1-\gamma}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big]^{1/2}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(-i)}(s)\big]^{1/2}\Big\} \\
&\leq \sqrt{2z}\cdot\left(1+\frac{V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta}{1-\gamma}\right)\cdot\Big(\iota_2\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(i)}(s)\big]+\iota_2\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}_{t,(-i)}(s)\big]\Big).
\end{aligned}$$

As a consequence, with M_i in (4.14) replaced by

$$M_i = \sqrt{2z}\cdot\big[2+\textstyle\sum_{i\in\{1,2\}}(V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta)/(1-\gamma)\big]\cdot\iota_2,\quad i\in\{1,2\},$$

the convergence guarantee established in Theorem 4.5 remains valid. Taking $\lambda_1=\lambda_2 = x\cdot\iota_2\sqrt{2z}$ into $\lambda_i \geq 2M_i$, we obtain

$$x \geq 2+\textstyle\sum_{i\in\{1,2\}}(V^{\max}_{(i)}+Q^{\max}_{(i)}\cdot\zeta)/(1-\gamma).$$

For the above inequality to hold, we have a sufficient requirement for the ratio z as

$$z \leq (1-\gamma)^4\Big/\Big[16(1+\gamma)\iota_2\cdot\log\big(|\mathcal{A}^1|\cdot|\mathcal{A}^2|\big)\Big]^2.$$

J STRONGER ASSUMPTION: HANNAN CONSISTENCY

In the proof of Theorem 4.5 in Appendix C, we replace $\pi^2_{\theta_t}$ and π^2_* with $\widetilde{\pi}^2_t$. Also, we replace π^1_* with

$$\pi^1_\dagger = \operatorname*{argmax}_{\pi^1}\left\{\frac{1}{T}\cdot\sum_{t=0}^{T-1}\mathcal{J}(\pi^1,\widetilde{\pi}^2_t)\right\}.$$

With stronger assumptions stated in 4.2, corresponding to (C.8), we have

$$\begin{aligned}
&(1-\gamma)\alpha_{t,(1)}\cdot\big[\mathcal{J}_{(1)}(\pi^1_\dagger,\widetilde{\pi}^2_t)-\mathcal{J}_{(1)}(\pi^1_{\theta_t},\widetilde{\pi}^2_t)\big] \\
&\quad\leq \mathbb{E}_{\nu^*}\big[\mathrm{KL}^\dagger_{t+1,(1)}(s)\big] - \left\{1-\left\{\lambda_1-\left[1+\frac{V^{\max}_{(1)}+Q^{\max}_{(1)}\cdot\zeta}{1-\gamma}\right]\cdot\iota_1\right\}\cdot\alpha_{t,(1)}\right\}\cdot\mathbb{E}_{\nu^*}\big[\mathrm{KL}^\dagger_{t,(1)}(s)\big] \\
&\qquad(\epsilon_t+\epsilon'_t) + \big[2+2\lambda_1^2/(\lambda_1-M_1)^2\big]\cdot(Q^{\max}_{(1)})^2\cdot\alpha^2_{t,(1)},
\end{aligned}\tag{J.1}$$

where $\mathrm{KL}^\dagger_{t+1,(1)}(s) = \mathrm{KL}(\pi^1_\dagger(\cdot\,|\,s)\,\|\,\pi^1_{\theta_t}(\cdot\,|\,s))$. Here we drop the KL-divergence term for Player 2 since π^2_* and $\pi^2_{\theta_t}$ are replaced by the same policy $\widetilde{\pi}^2_t$. Also, the regularization parameter λ_2 is dropped since now Player 2 no longer uses such a regularization parameter to update its policies. Multiplying

both sides of (J.2) and setting the stepsize $\alpha_{t,(1)}$ as the stepsize choice in Theorem 4.5, we obtain

$$\begin{aligned}
&\frac{1-\gamma}{(\lambda_1 - M_1)} \cdot \left[\mathcal{J}_{(1)}(\pi^1_\dagger, \widetilde{\pi}^2_t) - \mathcal{J}_{(1)}(\pi^1_{\theta_t}, \widetilde{\pi}^2_t)\right] \\
&\quad \leq t \cdot \mathbb{E}_{\nu^*}\left[\mathrm{KL}^\dagger_{t,(1)}(s)\right] - (t+1) \cdot \mathbb{E}_{\nu^*}\left[\mathrm{KL}^\dagger_{t+1,(1)}(s)\right] \\
&\qquad + (t+1) \cdot (\epsilon_t + \epsilon'_t) + \frac{\left[2 + 2\lambda_1^2/(\lambda_1 - M_1)^2\right] \cdot (Q^{\max}_{(1)})^2}{(t+1) \cdot (\lambda_1 - M_1)},
\end{aligned} \tag{J.2}$$

applying same argument in the proof of Theorem 4.5, we obtain

$$\begin{aligned}
&\sup_{\pi^1}\left\{\frac{1}{T} \cdot \sum_{t=0}^{T-1}\left[\mathcal{J}(\pi^1, \widetilde{\pi}^2_t) - \mathcal{J}(\pi^1_{\theta_t}, \widetilde{\pi}^2_t)\right]\right\} \\
&= \frac{1}{T} \cdot \sum_{t=0}^{T-1}\left[\mathcal{J}(\pi^1_\dagger, \widetilde{\pi}^2_t) - \mathcal{J}(\pi^1_{\theta_t}, \widetilde{\pi}^2_t)\right] \\
&\leq \frac{\sigma \cdot (\lambda_1 - M_1)}{(1-\gamma) \cdot T} + \frac{\left[2 + 2\lambda_1^2/(\lambda_1 - M_1)^2\right] \cdot (Q^{\max}_{(1)})^2}{(1-\gamma) \cdot (\lambda_1 - M_1)} \cdot \frac{\log T}{T} + \lambda_1 \cdot \log|\mathcal{A}^1|.
\end{aligned}$$

Thus, we conclude the proof of (4.17). Using the same argument as above, we can also prove the same result for Player 2 when Player 1 does not update its policies according to Algorithm 1. Setting $\widetilde{\pi}^1_t = \pi^1_\dagger = \pi^1_*$ and $\widetilde{\pi}^2_t = \pi^2_\dagger = \pi^2_*$, and combining such result for both players, we can replace the left-hand side of the convergence guarantee established in Theorem 4.5 by (4.18). Thus, the proposed Algorithm 1 is Hannan consistent.