Supplementary Material for "Learning While Playing in Mean-Field Games: Convergence and Optimality"

A. Proof of Lemma 1

Proof. By the definition of Λ, we have

$$\begin{aligned}
&\left\|\Lambda^\lambda(\mu)-\Lambda^\lambda(\mu')\right\|_{\mathcal{H}}\\
=&\left\|\Gamma_2\left(\Gamma_1^\lambda(\mu),\mu\right)-\Gamma_2\left(\Gamma_1^\lambda(\mu'),\mu'\right)\right\|_{\mathcal{H}}\\
\le&\left\|\Gamma_2\left(\Gamma_1^\lambda(\mu),\mu\right)-\Gamma_2\left(\Gamma_1^\lambda(\mu'),\mu\right)\right\|_{\mathcal{H}}+\left\|\Gamma_2\left(\Gamma_1^\lambda(\mu'),\mu\right)-\Gamma_2\left(\Gamma_1^\lambda(\mu'),\mu'\right)\right\|_{\mathcal{H}} && \text{triangle inequality}\\
\le&d_2 D\left(\Gamma_1^\lambda(\mu),\Gamma_1^\lambda(\mu')\right)+d_3\left\|\mu-\mu'\right\|_{\mathcal{H}} && \text{Assumption 3}\\
\le&d_1d_2\left\|\mu-\mu'\right\|_{\mathcal{H}}+d_3\left\|\mu-\mu'\right\|_{\mathcal{H}}, && \text{Assumption 2}
\end{aligned}$$

which proves the lemma. □

B. Technical Lemmas

The proofs of our main Theorems 1 and 2 involve several common steps. We summarize these steps as several lemmas, which are proved below.

B.1. Properties of KL-Divergence

We start with two lemmas about boundedness and Lipschitzness of KL-divergence.

Lemma 2. *Let p^* and $p\in\Delta(\mathcal{A})$ and $\widehat{p}=(1-\eta)p+\eta\frac{\mathbf{1}_{|\mathcal{A}|}}{|\mathcal{A}|}$. Then*

$$\begin{aligned}
D_{\mathrm{KL}}\left(p^*\|\widehat{p}\right)&\le\log\frac{|\mathcal{A}|}{\eta},\\
D_{\mathrm{KL}}\left(p^*\|\widehat{p}\right)-D_{\mathrm{KL}}\left(p^*\|p\right)&\le 2\eta.
\end{aligned}$$

Proof. By definition we have

$$\begin{aligned}
D_{\mathrm{KL}}\left(p^*\|\widehat{p}\right)&=\sum_{a\in\mathcal{A}}p^*(a)\log\frac{p^*(a)}{\widehat{p}(a)}\\
&=\sum_{a\in\mathcal{A}}p^*(a)\log\frac{p^*(a)}{(1-\eta)p(a)+\frac{\eta}{|\mathcal{A}|}}\\
&\le\sum_{a\in\mathcal{A}}p^*(a)\log\frac{1}{0+\frac{\eta}{|\mathcal{A}|}}\\
&=\log\frac{|\mathcal{A}|}{\eta},
\end{aligned}$$

thereby proving the first inequality.

Note that

$$D_{\mathrm{KL}}\left(p^*\|\widehat{p}\right)-D_{\mathrm{KL}}\left(p^*\|p\right)=\sum_{a\in\mathcal{A}}p^*(a)\log\left(\frac{p(a)}{\widehat{p}(a)}\right). \tag{14}$$

For each a such that $p(a) \le \widehat{p}(a)$, we have

$$\log\left(\frac{p(a)}{\widehat{p}(a)}\right) \le 0 \le 2\eta;$$

for each a' such that $p(a') \ge \widehat{p}(a')$, we have

$$\begin{aligned}\log\left(\frac{p(a')}{\widehat{p}(a')}\right) &= \log\left(\frac{p(a')}{(1-\eta)p(a') + \eta/|\mathcal{A}|}\right)\\ &\le \log\left(\frac{p(a')}{(1-\eta)p(a')}\right)\\ &\le \frac{\eta}{1-\eta} \le 2\eta,\end{aligned}$$

where the third step follows from the fact that $\log(z) \le z-1$ for all $z>0$ and the last step holds as $\eta \in [0, \frac{1}{2}]$. Applying the above two inequalities to (14) completes the proof. □

The following Lemma states that the KL-divergence is Lipschitz w.r.t. $\|\cdot\|_1$ under certain conditions.

Lemma 3. *Let x, y and $z \in \Delta(\mathcal{A})$. If $x(a) \ge \alpha_1$, $y(a) \ge \alpha_1$ and $z(a) \ge \alpha_2$ for all $a \in \mathcal{A}$, then*

$$D_{\mathrm{KL}}(x\|z) - D_{\mathrm{KL}}(y\|z) \le \left(1 + \log\frac{1}{\min\{\alpha_1, \alpha_2\}}\right) \cdot \|x-y\|_1 .$$

Proof. Under the lower bound assumption of the lemma, we have

$$\frac{\mathrm{d}D_{\mathrm{KL}}(x\|z)}{\mathrm{d}x(a)} = 1 + \log\frac{x(a)}{z(a)} \le 1 + \log\frac{1}{\alpha_2}$$

and

$$-\frac{\mathrm{d}D_{\mathrm{KL}}(x\|z)}{\mathrm{d}x(a)} \le -1 - \log\alpha_1.$$

It follows that

$$\left\|\frac{\mathrm{d}D_{\mathrm{KL}}(x\|z)}{\mathrm{d}x(a)}\right\|_\infty \le \max\left\{1 + \log\frac{1}{\alpha_2}, -1 - \log\alpha_1\right\} \le 1 + \log\frac{1}{\min\{\alpha_1,\alpha_2\}}.$$

Hence the function $x \mapsto D_{\mathrm{KL}}(x\|z)$ is Lipschitz w.r.t. $\|\cdot\|_1$, the dual norm of $\|\cdot\|_\infty$. □

B.2. Policy Improvement

To analyze the convergence of policy sequence, we need the following lemma, which characterizes the policy improvement step.

Lemma 4. *For any distributions $p^*, p \in \Delta(\mathcal{A})$,state $s \in \mathcal{S}$ and function $G : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, it holds for $p' \in \Delta(\mathcal{A})$ with $p'(\cdot) \propto p(\cdot)\cdot\exp\left[\alpha G(s,\cdot)\right]$ that*

$$D_{\mathrm{KL}}\left(p^*\|p'\right) \le D_{\mathrm{KL}}\left(p^*\|p\right) - \alpha\left\langle G(s,\cdot), p^* - p\right\rangle + \alpha^2\left\|G(s,\cdot)\right\|_\infty^2/2.$$

Proof. For any function $g : \mathcal{A}\to\mathbb{R}$ and distribution $p \in \Delta(\mathcal{A})$, let $z : \mathcal{A}\to\mathbb{R}$ be a constant function defined by

$$z(a) = \log\left(\sum_{a'\in\mathcal{A}} p(a')\cdot\exp\left(\alpha g(a')\right)\right).$$

Note that for any distributions $p^*, p' \in \Delta(\mathcal{A})$,$\langle z, p^* - p'\rangle = 0$. Since

$$p'(\cdot) \propto p(\cdot)\cdot\exp\left(\alpha g(\cdot)\right),$$

we have $\alpha g(\cdot) = z(\cdot) + \log(p'(\cdot)/p(\cdot))$. Hence

$$\begin{aligned}\alpha \langle g, p^* - p' \rangle &= \langle z + \log(p'/p), p^* - p' \rangle \\ &= \langle z, p^* - p' \rangle + \langle \log(p^*/p), p^* \rangle + \langle \log(p'/p^*), p^* \rangle + \langle \log(p'/p), -p' \rangle \\ &= D_{\mathrm{KL}}(p^* \| p) - D_{\mathrm{KL}}(p^* \| p') - D_{\mathrm{KL}}(p' \| p).\end{aligned}$$

Therefore, for each state $s \in \mathcal{S}$, we have

$$\begin{aligned}\alpha \langle G(s,\cdot), p^* - p \rangle &= \alpha \langle G(s,\cdot), p^* - p' \rangle + \alpha \langle G(s,\cdot), p' - p \rangle \\ &= D_{\mathrm{KL}}(p^* \| p) - D_{\mathrm{KL}}(p^* \| p') - D_{\mathrm{KL}}(p' \| p) + \alpha \langle G(s,\cdot), p' - p \rangle \\ &\le D_{\mathrm{KL}}(p^* \| p) - D_{\mathrm{KL}}(p^* \| p') - D_{\mathrm{KL}}(p' \| p) + \alpha \|G(s,\cdot)\|_\infty \cdot \|p - p'\|_1.\end{aligned}$$

Rearranging terms yields

$$D_{\mathrm{KL}}(p^* \| p') \le D_{\mathrm{KL}}(p^* \| p) - \alpha \langle G(s,\cdot), p^* - p \rangle - D_{\mathrm{KL}}(p' \| p) + \alpha \|G(s,\cdot)\|_\infty \cdot \|p - p'\|_1. \tag{15}$$

Meanwhile, by Pinsker's inequality, it holds that

$$D_{\mathrm{KL}}(p' \| p) \ge \|p - p'\|_1^2 / 2. \tag{16}$$

By combining (15) and (16), we obtain

$$\begin{aligned}D_{\mathrm{KL}}(p^* \| p') &\le D_{\mathrm{KL}}(p^* \| p) - \alpha \langle G(s,\cdot), p^* - p \rangle - \|p - p'\|_1^2 / 2 + \alpha \|G(s,\cdot)\|_\infty \cdot \|p - p'\|_1 \\ &\le D_{\mathrm{KL}}(p^* \| p) - \alpha \langle G(s,\cdot), p^* - p \rangle + \alpha^2 \|G(s,\cdot)\|_\infty^2 / 2,\end{aligned}$$

which concludes the proof. □

C. Proof of Theorem 1

In order to obtain an upper bound on the optimality gap

$$\sigma_\mu^t := \|\mu_t - \mu^*\|_{\mathcal{H}}, \tag{17}$$

where μ^* is the embedded mean-field state of the entropy regularized NE, we also need to estimate the gap between π_t and the optimal solution π_t^* to the entropy regularized MDP_{μ_t}. We define

$$\sigma_\pi^t := \mathbb{E}_{s \sim \rho_t^*} \left[D_{\mathrm{KL}}(\pi_t^*(\cdot|s) \| \pi_t(\cdot|s)) \right] \tag{18}$$

to quantify the convergence of policy sequence.

Before proceeding, we establish the following properties of entropy regularized MDPs, which are central to the convergence analysis.

Properties of Regularized MDP. The following lemma quantifies the performance difference between two policies for a regularized MDP — measured in terms of the expected total reward — through the Q-function and their KL-divergence. The proof is provided in Appendix C.1.

Lemma 5 (Performance Difference)**.** *For each $\mu \in \mathcal{M}$ and policies $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, it holds that*

$$\begin{aligned}&J_\mu^\lambda(\pi') - J_\mu^\lambda(\pi) + \frac{\lambda}{1-\gamma} \mathbb{E}_{s \sim \rho_\mu^{\pi'}} \left[D_{\mathrm{KL}}(\pi'(\cdot|s) \| \pi(\cdot|s)) \right] \\ =&\frac{1}{1-\gamma} \mathbb{E}_{s \sim \rho_\mu^{\pi'}} \left[\left\langle Q_\mu^{\lambda,\pi}(s,\cdot) - \lambda \log \pi(\cdot|s), \pi'(\cdot|s) - \pi(\cdot|s) \right\rangle \right],\end{aligned} \tag{19}$$

where $\rho_\mu^{\pi'}$ is the discounted state visitation distribution induced by the policy π' on MDP_μ.

We can characterize the optimal policy $\pi_\mu^{\lambda,*}$ in terms of the optimal Q-function $Q_\mu^{\lambda,*}$ as a Boltzmann distribution of the form (Cen et al., 2020; Nachum et al., 2017)

$$\pi_\mu^{\lambda,*}(a|s) \propto \exp\left(\frac{Q_\mu^{\lambda,*}(s,a)}{\lambda}\right). \tag{20}$$

For the setting where the reward function is bounded, we then can obtain a lower bound on $\pi_\mu^{\lambda,*}$, as stated in the following lemma. The proof is provided in Appendix C.2

Lemma 6. *Suppose that there exists a constant $R_{\max} > 0$ such that $0 \le \sup_{(s,a,\mu)\in\mathcal{S}\times\mathcal{A}\times\mathcal{M}} r(s,a,\mu) \le R_{\max}$. For each $\mu \in \mathcal{M}$, and each policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, we have*

$$\left\|Q_\mu^{\lambda,\pi}\right\|_\infty \le Q_{\max} := \frac{R_{\max} + \gamma\lambda \log|\mathcal{A}|}{1-\gamma}.$$

Also, the optimal policy $\pi_\mu^{\lambda,*}$ for the regularized MDP_μ *satisfies*

$$\pi_\mu^{\lambda,*}(a|s) \ge \frac{1}{e^{Q_{\max}/\lambda}|\mathcal{A}|}, \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

Convergence Analysis. We now move to the convergence analysis. For clarity of exposition, we use $\mathbb{E}_\rho\left[\|\pi - \pi'\|_1\right]$ as shorthand for $\mathbb{E}_{s\sim\rho}\left[\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1\right]$, where $\rho \in \Delta(\mathcal{S})$; we also use $\mathbb{E}_\rho\left[D_{\mathrm{KL}}\left(\pi\|\pi'\right)\right]$ as shorthand for $\mathbb{E}_{s\sim\rho}\left[D_{\mathrm{KL}}\left(\pi(\cdot|s)\|\pi'(\cdot|s)\right)\right]$. We recall that the step sizes are chosen as

$$\alpha_t \equiv \alpha = c_\alpha T^{-2/5}, \qquad \beta_t \equiv \beta = c_\beta T^{-4/5},$$

where the parameters c_α and c_β satisfy that:

$$c_\alpha T^{-2/5}\lambda < 1, \qquad c_\beta T^{-4/5}\overline{d} < 1. \tag{21}$$

Here $\overline{d} := 1 - d_1 d_2 - d_3 > 0$, where d_1 appears in Assumption 2, and d_2, d_3 appear in Assumption 3.

Step 1: Convergence of Policy. We first characterize the convergence behavior of the policy sequence $\{\pi_t\}_{t\ge0}$. Recall that $\sigma_\pi^t = \mathbb{E}_{s\sim\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_t(\cdot|s)\right)\right]$. We start with establishing a recursive relationship between σ_π^{t+1} and σ_π^t, as stated in the following lemma. The proof is provided in Section C.3.

Lemma 7. *Under the setting of Theorem 1, for each $t \ge 0$, we have*

$$\sigma_\pi^{t+1} \le (1-\lambda\alpha_t)\sigma_\pi^t + \left(d_0 \log\frac{|\mathcal{A}|}{\eta} + \kappa C_\rho d_1\right)\|\mu_{t+1} - \mu_t\|_{\mathcal{H}} + 2\varepsilon_Q\alpha_t + \frac{Q_{\max}^2}{2}\alpha_t^2 + 2\eta, \tag{22}$$

where $\kappa = \frac{4}{1-\gamma}\log\frac{|\mathcal{A}|}{\eta} + \frac{2R_{\max}}{\lambda(1-\gamma)}$.

Recall that $\mu_{t+1} = (1-\beta_t)\mu_t + \beta_t \cdot \Gamma_2(\pi_t, \mu_t)$. Under Assumption 1, we have

$$\|\mu_{t+1} - \mu_t\|_{\mathcal{H}} = \beta_t\|\mu_t - \Gamma_2(\pi_t,\mu_t)\|_{\mathcal{H}} \le 2\beta_t. \tag{23}$$

Lemma 7 implies that

$$\sigma_\pi^{t+1} \le (1-\lambda\alpha_t)\sigma_\pi^t + \overline{C}_1\beta_t + 2\varepsilon_Q\alpha_t + \frac{Q_{\max}^2}{2}\alpha_t^2 + 2\eta, \tag{24}$$

where we define

$$\overline{C}_1 := 2\left(d_0\log\frac{|\mathcal{A}|}{\eta} + \kappa C_\rho d_1\right).$$

With $\alpha_t \equiv \alpha$, $\beta_t \equiv \beta$, from Equation (24) we have that

$$\sigma_\pi^t \le \frac{1}{\lambda\alpha}\left(\sigma_\pi^t - \sigma_\pi^{t+1}\right) + \frac{\overline{C}_1\beta}{\lambda\alpha} + \frac{2\varepsilon_Q}{\lambda} + \frac{Q_{\max}^2}{2\lambda}\alpha + \frac{2\eta}{\lambda\alpha}. \tag{25}$$

Summing over $t = 0, 1, \dots T-1$ on both sides of (25) and dividing by t gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\pi^t &\le \frac{1}{T\lambda\alpha}\left(\sigma_\pi^0 - \sigma_\pi^T\right) + \frac{\overline{C}_1\beta}{\lambda\alpha} + \frac{2\varepsilon_Q}{\lambda} + \frac{Q_{\max}^2}{2\lambda}\alpha + \frac{2\eta}{\lambda\alpha} \\ &\le \frac{1}{T\lambda\alpha}\sigma_\pi^0 + \frac{\overline{C}_1\beta}{\lambda\alpha} + \frac{2\varepsilon_Q}{\lambda} + \frac{Q_{\max}^2}{2\lambda}\alpha + \frac{2\eta}{\lambda\alpha}. \end{aligned} \tag{26}$$

When choosing $\alpha = \mathcal{O}(T^{-2/5})$, $\beta = \mathcal{O}(T^{-4/5})$ and $\eta = \mathcal{O}(T^{-1})$, we have $\overline{C}_1 = \mathcal{O}(\log T)$. Therefore, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\pi^t \lesssim \frac{\log T}{\lambda T^{2/5}} + \frac{2\varepsilon_Q}{\lambda}. \tag{27}$$

If we let T be a random number sampled uniformly from $\{0, \dots, T-1\}$, then the above equation can be written equivalently as

$$\mathbb{E}_{\mathsf{T}}\left[\sigma_\pi^{\mathsf{T}}\right] \lesssim \frac{\log T}{\lambda T^{2/5}} + \frac{2\varepsilon_Q}{\lambda}. \tag{28}$$

Step 2: Convergence of Mean-field Embedding. We now proceed to characterize the optimality gap for the embedded mean-field state. We obtain the following upper bound on the optimality gap $\sigma_\mu^t = \|\mu_t - \mu^*\|_{\mathcal{H}}$. The proof is provided in Section C.4.

Lemma 8. *Under the setting of Theorem 1, we have*

$$\sigma_\mu^{t+1} \le \left(1 - \beta_t\overline{d}\right)\sigma_\mu^t + d_2\overline{C}_\rho\beta_t\sqrt{\sigma_\pi^t}, \qquad \forall t \in [T],$$

where $\overline{d} = 1 - d_1d_2 - d_3 > 0$.

Lemma 8 implies that

$$\sigma_\mu^t \le \frac{1}{\overline{d}\beta_t}\left(\sigma_\mu^t - \sigma_\mu^{t+1}\right) + \frac{d_2\overline{C}_\rho}{\overline{d}}\sqrt{\sigma_\pi^t}. \tag{29}$$

With $\beta_t \equiv \beta = \mathcal{O}(T^{-4/5})$, averaging equation (29) over iteration $t = 0, \dots, T-1$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\mu^t &\le \frac{1}{\overline{d}\beta T}\left(\sigma_\mu^0 - \sigma_\mu^T\right) + \frac{d_2\overline{C}_\rho}{\overline{d}T}\sum_{t=0}^{T-1}\sqrt{\sigma_\pi^t} \\ &\le \frac{\sigma_\mu^0}{\overline{d}\beta T} + \frac{d_2\overline{C}_\rho}{\overline{d}T}\sum_{t=0}^{T-1}\sqrt{\sigma_\pi^t} \\ &\le \frac{\sigma_\mu^0}{\overline{d}\beta T} + \frac{d_2\overline{C}_\rho}{\overline{d}}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\pi^t}, \end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality.

From Eq. (27), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\mu^t &\lesssim \frac{\sigma_\mu^0}{\overline{d}}T^{-1/5} + \frac{d_2\overline{C}_\rho}{\overline{d}}\sqrt{\frac{\log T}{\lambda T^{2/5}} + \frac{2\varepsilon_Q}{\lambda}} \\ &\lesssim \sqrt{\frac{\log T}{\lambda T^{2/5}} + \frac{2\varepsilon_Q}{\lambda}} \\ &\lesssim \frac{1}{\sqrt{\lambda}}\left(\frac{\sqrt{\log T}}{T^{1/5}} + \sqrt{\varepsilon_Q}\right). \end{aligned}$$

This equation, together with Jensen's inequality, proves equation (13) in Theorem 1.

Turning to equation (12) in Theorem 1, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} D\left(\pi_t, \pi_t^*\right) &= \mathbb{E}_{\mathsf{T}}\left[D\left(\pi_{\mathsf{T}}, \pi_{\mathsf{T}}^*\right)\right] \\
&= \mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho^*}\left[\left\|\pi_{\mathsf{T}}^*(\cdot|s) - \pi_{\mathsf{T}}(\cdot|s)\right\|_1\right] \\
&= \mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[\frac{\rho^*(s)}{\rho_{\mathsf{T}}^*(s)}\left\|\pi_{\mathsf{T}}^*(\cdot|s) - \pi_{\mathsf{T}}(\cdot|s)\right\|_1\right] \\
&\overset{(i)}{\leq} \sqrt{\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[\left|\frac{\rho^*(s)}{\rho_{\mathsf{T}}^*(s)}\right|^2\right]\cdot\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[\left\|\pi_{\mathsf{T}}^*(\cdot|s) - \pi_{\mathsf{T}}(\cdot|s)\right\|_1^2\right]} \\
&\overset{(ii)}{\leq} \sqrt{\overline{C}_\rho^2\cdot\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[2D_{\mathrm{KL}}\left(\pi_{\mathsf{T}}^*(\cdot|s)\|\pi_{\mathsf{T}}(\cdot|s)\right)\right]} \\
&= \sqrt{\overline{C}_\rho^2\cdot 2\mathbb{E}_{\mathsf{T}}\left[\sigma_\pi^{\mathsf{T}}\right]} \\
&\overset{(iii)}{\lesssim} \frac{1}{\sqrt{\lambda}}\left(\frac{\sqrt{\log T}}{T^{1/5}} + \sqrt{\varepsilon_Q}\right),
\end{aligned}$$

where step (i) follows from Cauchy-Schwarz inequality, step (ii) follows from Assumption 4 and Pinsker's inequality, and step (iii) follows from the bound in equation (28). The above equation, together with Jensen's inequality, proves equation (12). We have completed the proof of Theorem 1.

C.1. Proof of Lemma 5

Proof. The proof follows similar argument as that of Lemma 1 in (Cen et al., 2020). We provide the full proof for completeness. By the definition of $V_\mu^{\lambda,\pi}$ in (3), we have

$$\begin{aligned}
&V_\mu^{\lambda,\pi'}(s) \\
=&\mathbb{E}_{a_t\sim\pi'(s_t), s_{t+1}\sim\mathrm{P}(\cdot|s_t,a_t,\mu)}\left[\sum_{t=0}^{\infty}\gamma^t\left[r_\mu^{\lambda,\pi'}(s,a) + V_\mu^{\lambda,\pi}(s_t) - V_\mu^{\lambda,\pi}(s_t)\right] \mid s_0 = s\right]. \\
=&\mathbb{E}_{a_t\sim\pi'(s_t), s_{t+1}\sim\mathrm{P}(\cdot|s_t,a_t,\mu)}\left[\sum_{t=0}^{\infty}\gamma^t\left[r_\mu^{\lambda,\pi'}(s,a) + \gamma V_\mu^{\lambda,\pi}(s_{t+1}) - V_\mu^{\lambda,\pi}(s_t)\right] \mid s_0 = s\right] + V_\mu^{\lambda,\pi}(s). \qquad (30)
\end{aligned}$$

Recall that the Q-function $Q_\mu^{\lambda,\pi}$ of a policy π for the regularized MDP_μ is related to $V_\mu^{\lambda,\pi}$ as

$$\begin{aligned}
V_\mu^{\lambda,\pi}(s) &= \mathbb{E}_{a\sim\pi(s)}\left[Q_\mu^{\lambda,\pi}(s,a) - \lambda\log\pi(a|s)\right] = \left\langle Q_\mu^{\lambda,\pi}(s,\cdot), \pi(\cdot|s)\right\rangle + \lambda\mathbb{H}\left(\pi(\cdot|s)\right), \qquad \forall s\in\mathcal{S}, \\
Q_\mu^{\lambda,\pi}(s,a) &= r(s,a,\mu) + \gamma\mathbb{E}_{s_1\sim\mathrm{P}(\cdot|s,a,\mu)}\left[V_\mu^{\lambda,\pi}(s_1)\right], \qquad \forall (s,a)\in\mathcal{S}\times\mathcal{A}.
\end{aligned}$$

We have

$$\begin{aligned}
\left\langle Q_\mu^{\lambda,\pi}(s,\cdot), \pi'(\cdot|s)\right\rangle &= \mathbb{E}_{a\sim\pi'(s)}\left[Q_\mu^{\lambda,\pi}(s,a)\right], \\
&= \mathbb{E}_{a\sim\pi'(s)}\left[r(s,a,\mu) + \gamma\mathbb{E}_{s_1\sim\mathrm{P}(\cdot|s,a,\mu)}\left[V_\mu^{\lambda,\pi}(s_1)\right]\right] \\
&= \mathbb{E}_{a\sim\pi'(s), s_1\sim\mathrm{P}(\cdot|s,a,\mu)}\left[r_\mu^{\lambda,\pi'}(s,a) + \gamma V_\mu^{\lambda,\pi}(s_1) + \lambda\log\pi'(a|s)\right] \\
&= \mathbb{E}_{a\sim\pi'(s), s_1\sim\mathrm{P}(\cdot|s,a,\mu)}\left[r_\mu^{\lambda,\pi'}(s,a) + \gamma V_\mu^{\lambda,\pi}(s_1)\right] - \lambda\mathbb{H}\left(\pi'(\cdot|s)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\langle Q_\mu^{\lambda,\pi}(s,\cdot), \pi'(\cdot|s) - \pi(\cdot|s)\right\rangle \\
=&\mathbb{E}_{a\sim\pi'(s), s_1\sim\mathrm{P}(\cdot|s,a,\mu)}\left[r^{\lambda,\pi'}(s,a,\mu) + \gamma V_\mu^{\lambda,\pi}(s_1)\right] - \lambda\mathbb{H}\left(\pi'(\cdot|s)\right) - V_\mu^{\lambda,\pi}(s) + \lambda\mathbb{H}\left(\pi(\cdot|s)\right) \\
=&\mathbb{E}_{a\sim\pi'(s), s_1\sim\mathrm{P}(\cdot|s,a,\mu)}\left[r^{\lambda,\pi'}(s,a,\mu) + \gamma V_\mu^{\lambda,\pi}(s_1) - V_\mu^{\lambda,\pi}(s)\right] - \lambda\left[\mathbb{H}\left(\pi'(\cdot|s)\right) - \mathbb{H}\left(\pi(\cdot|s)\right)\right]. \qquad (31)
\end{aligned}$$

Plugging (31) into (30), we have

$$\begin{aligned}
&V_\mu^{\lambda,\pi'}(s) - V_\mu^{\lambda,\pi'}(s) \\
=&\mathbb{E}_{a_t\sim\pi'(s_t),s_{t+1}\sim \mathrm{P}(\cdot|s_t,a_t,\mu)}\left[\sum_{t=0}^{\infty}\gamma^t\left\langle Q_\mu^{\lambda,\pi}(s_t,\cdot),\pi'(\cdot|s_t)-\pi(\cdot|s_t)\right\rangle \mid s_0=s\right] \\
&+\mathbb{E}_{a_t\sim\pi'(s_t),s_{t+1}\sim \mathrm{P}(\cdot|s_t,a_t,\mu)}\left[\sum_{t=0}^{\infty}\gamma^t\lambda\left(\mathbb{H}\left(\pi'(\cdot|s_t)\right)-\mathbb{H}\left(\pi(\cdot|s_t)\right)\right) \mid s_0=s\right]. \qquad (32)
\end{aligned}$$

Recall the definition of $J_\mu^\lambda(\pi)$ in (4). Taking expectation with respect to $s\sim\nu_0$ on both sides of (32) yields □

$$\begin{aligned}
&J_\mu^\lambda(\pi') - J_\mu^\lambda(\pi) \\
=&\mathbb{E}_{s_0\sim\nu_0,a_t\sim\pi'(s_t),s_{t+1}\sim \mathrm{P}(\cdot|s_t,a_t,\mu)}\left[\sum_{t=0}^{\infty}\gamma^t\left\langle Q_\mu^{\lambda,\pi}(s_t,\cdot),\pi'(\cdot|s_t)-\pi(\cdot|s_t)\right\rangle\right] \\
&+\mathbb{E}_{s_0\sim\nu_0,a_t\sim\pi'(s_t),s_{t+1}\sim \mathrm{P}(\cdot|s_t,a_t,\mu)}\left[\sum_{t=0}^{\infty}\gamma^t\lambda\left(\mathbb{H}\left(\pi'(\cdot|s_t)\right)-\mathbb{H}\left(\pi(\cdot|s_t)\right)\right)\right] \\
=&\frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho_\mu^{\pi'}}\left[\left\langle Q_\mu^{\lambda,\pi}(s,\cdot),\pi'(\cdot|s)-\pi(\cdot|s)\right\rangle+\lambda\left(\mathbb{H}\left(\pi'(\cdot|s)\right)-\mathbb{H}\left(\pi(\cdot|s)\right)\right)\right]. \qquad (33)
\end{aligned}$$

For the entropy term in (33), we have

$$\begin{aligned}
&\mathbb{E}_{s\sim\rho_\mu^{\pi'}}\left[\mathbb{H}\left(\pi'(\cdot|s)\right)-\mathbb{H}\left(\pi(\cdot|s)\right)\right] \\
=&\mathbb{E}_{s\sim\rho_\mu^{\pi'}}\left[\left\langle\log\frac{1}{\pi'(\cdot|s)},\pi'(\cdot|s)\right\rangle-\left\langle\log\frac{1}{\pi(\cdot|s)},\pi(\cdot|s)\right\rangle\right] \\
=&\mathbb{E}_{s\sim\rho_\mu^{\pi'}}\left[\left\langle\log\frac{1}{\pi(\cdot|s)}-\log\frac{\pi'(\cdot|s)}{\pi(\cdot|s)},\pi'(\cdot|s)\right\rangle-\left\langle\log\frac{1}{\pi(\cdot|s)},\pi(\cdot|s)\right\rangle\right] \\
=&\mathbb{E}_{s\sim\rho_\mu^{\pi'}}\left[\left\langle\log\frac{1}{\pi(\cdot|s)},\pi'(\cdot|s)-\pi(\cdot|s)\right\rangle-D_{\mathrm{KL}}\left(\pi'(\cdot|s)\|\pi(\cdot|s)\right)\right]. \qquad (34)
\end{aligned}$$

Taking (34) into (33) yields the desired equation in Lemma 5.

C.2. Proof of Lemma 6

Proof. Note that the value function $V_\mu^{\lambda,\pi}$ can be written as

$$V_\mu^{\lambda,\pi}(s)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_\mu^{\lambda,\pi}(s_t,a_t)|s_0=s\right].$$

By the definition of $r_\mu^{\lambda,\pi}$ in (1), we have $0\le\mathbb{E}_\pi\left[r_\mu^{\lambda,\pi}(s_t,a_t)\right]\le R_{\max}+\lambda\log|\mathcal{A}|$. Therefore,

$$0\le V_\mu^{\lambda,\pi}(s)\le\frac{R_{\max}+\lambda\log|\mathcal{A}|}{1-\gamma},\qquad\forall s\in\mathcal{S},$$

and

$$0\le Q_\mu^{\lambda,\pi}(s,a)\le R_{\max}+\gamma\frac{R_{\max}+\lambda\log|\mathcal{A}|}{1-\gamma}=\frac{R_{\max}+\gamma\lambda\log|\mathcal{A}|}{1-\gamma},\qquad\forall s\in\mathcal{S},a\in\mathcal{A}.$$

For the second inequality, we have

$$\begin{aligned}
\pi_\mu^{\lambda,*}(a|s)&=\frac{\exp\left(Q_\mu^{\lambda,*}(s,a)/\lambda\right)}{\sum_{b\in\mathcal{A}}\exp\left(Q_\mu^{\lambda,*}(s,b)/\lambda\right)} \\
&\ge\frac{1}{\sum_{b\in\mathcal{A}}\exp\left(Q_{\max}/\lambda\right)}=\frac{1}{e^{Q_{\max}/\lambda}|\mathcal{A}|}
\end{aligned}$$

as claimed. □

C.3. Proof of Lemma 7

Recall that at the t-th iteration, the policy is improved as follows:

$$\widehat{\pi}_{t+1}(\cdot|s) \propto \pi_t(\cdot|s) \cdot \exp\left[\alpha_t \left(\widehat{Q}_t^\lambda(s,\cdot) - \lambda \log \pi_t(\cdot|s)\right)\right].$$

Applying the Lemma 4 of policy improvement, we have for each $s \in \mathcal{S}$,

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\widehat{\pi}_{t+1}(\cdot|s)\right)\\
\leq& D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_t(\cdot|s)\right) - \alpha_t\left\langle \widehat{Q}_t^\lambda(s,\cdot) - \lambda\log\pi_t(\cdot|s), \pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\rangle + \left\|\widehat{Q}_t^\lambda\right\|_\infty^2 \alpha_t^2/2\\
=& D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_t(\cdot|s)\right) - \alpha_t\left\langle Q_t^\lambda(s,\cdot) - \lambda\log\pi_t(\cdot|s), \pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\rangle\\
&+ \alpha_t\left\langle Q_t^\lambda(s,\cdot) - \widehat{Q}_t^\lambda(s,\cdot), \pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\rangle + \left\|\widehat{Q}_t^\lambda\right\|_\infty^2 \alpha_t^2/2\\
\leq& D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_t(\cdot|s)\right) - \alpha_t\left\langle Q_t^\lambda(s,\cdot) - \lambda\log\pi_t(\cdot|s), \pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\rangle\\
&+ 2\alpha_t\left\|Q_t^\lambda(s,\cdot) - \widehat{Q}_t^\lambda(s,\cdot)\right\|_\infty + \left\|\widehat{Q}_t^\lambda\right\|_\infty^2 \alpha_t^2/2.
\end{aligned}$$

Recall that $\pi_{t+1}(\cdot|s) = (1-\eta)\widehat{\pi}_{t+1}(\cdot|s) + \frac{\eta}{|\mathcal{A}|}\mathbf{1}_{|\mathcal{A}|}$. Lemma 2 implies that

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_{t+1}(\cdot|s)\right)\\
\leq& D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\widehat{\pi}_{t+1}(\cdot|s)\right) + 2\eta. \qquad (35)\\
\leq& D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_t(\cdot|s)\right) - \alpha_t\left\langle Q_t^\lambda(s,\cdot) - \lambda\log\pi_t(\cdot|s), \pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\rangle\\
&\quad \underbrace{+ 2\alpha_t\left\|Q_t^\lambda(s,\cdot) - \widehat{Q}_t^\lambda(s,\cdot)\right\|_\infty + \left\|\widehat{Q}_t^\lambda\right\|_\infty^2 \alpha_t^2/2 + 2\eta}_{Y_t(s)}. \qquad (36)
\end{aligned}$$

Taking expectation over ρ_t^* on both sides of (36) yields

$$\begin{aligned}
&\mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_{t+1}\right)\right]\\
\leq& \mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_t\right)\right] - \alpha_t\mathbb{E}_{s\sim\rho_t^*}\left[\left\langle Q_t^\lambda(s,\cdot) - \lambda\log\pi_t(\cdot|s), \pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\rangle\right] + \mathbb{E}_{s\sim\rho_t^*}\left[Y_t(s)\right]\\
\overset{(a)}{=}& \mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_t\right)\right] - (1-\gamma)\alpha_t\left[J_{\mu_t}^\lambda(\pi_t^*) - J_{\mu_t}^\lambda(\pi_t)\right] - \alpha_t\lambda\mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_t\right)\right] + \mathbb{E}_{s\sim\rho_t^*}\left[Y_t(s)\right]\\
\overset{(b)}{\leq}& (1-\alpha_t\lambda)\mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_t\right)\right] + \mathbb{E}_{s\sim\rho_t^*}\left[Y_t(s)\right], \qquad (37)
\end{aligned}$$

where step (a) follows from Lemma 5; step (b) follows from the fact that $J_{\mu_t}^\lambda(\pi_t) \leq J_{\mu_t}^\lambda(\pi_t^*)$, as $\pi_t^* = \Gamma_1^\lambda(\mu_t)$ is the optimal policy for the regularized MDP_{μ_t}.

Next we bound the difference between $\mathbb{E}_{\rho_{t+1}^*}\left[D_{\mathrm{KL}}\left(\pi_{t+1}^*\|\pi_{t+1}\right)\right]$ and $\mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_{t+1}\right)\right]$. By triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\rho_{t+1}^*}\left[D_{\mathrm{KL}}\left(\pi_{t+1}^*\|\pi_{t+1}\right)\right]\\
\leq& \mathbb{E}_{\rho_{t+1}^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_{t+1}\right)\right] + \left|\mathbb{E}_{\rho_{t+1}^*}\left[D_{\mathrm{KL}}\left(\pi_{t+1}^*\|\pi_{t+1}\right) - D_{\mathrm{KL}}\left(\pi_t^*\|\pi_{t+1}\right)\right]\right|\\
=& \mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_{t+1}\right)\right] + \underbrace{\left(\mathbb{E}_{\rho_{t+1}^*} - \mathbb{E}_{\rho_t^*}\right)\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_{t+1}\right)\right]}_{B_1} + \underbrace{\left|\mathbb{E}_{\rho_{t+1}^*}\left[D_{\mathrm{KL}}\left(\pi_{t+1}^*\|\pi_{t+1}\right) - D_{\mathrm{KL}}\left(\pi_t^*\|\pi_{t+1}\right)\right]\right|}_{B_2}. \qquad (38)
\end{aligned}$$

We now bound the first and second terms on the RHS of (38) separately.

- For the first term B_1: We have

$$\begin{aligned} B_1 &= \mathbb{E}_{s\sim\rho^*}\left[\frac{\rho^*_{t+1}(s)-\rho^*_t(s)}{\rho^*(s)}\cdot D_{\mathrm{KL}}\left(\pi^*_t\|\pi_{t+1}\right)\right] \\ &\overset{(a)}{\leq} \mathbb{E}_{s\sim\rho^*}\left[\frac{\left|\rho^*_{t+1}(s)-\rho^*_t(s)\right|}{\rho^*(s)}\right]\cdot \mathrm{KL}_{\max}, \\ &\overset{(b)}{\leq} \mathrm{KL}_{\max}\cdot d_0\left\|\mu_t-\mu_{t-1}\right\|_{\mathcal{H}}, \end{aligned} \tag{39}$$

  where step (a) uses the fact that $D_{\mathrm{KL}}\left(\pi^*_t\|\pi_{t+1}\right)\leq \mathrm{KL}_{\max}:=\log\frac{|\mathcal{A}|}{\eta}$ (cf. Lemma 2) and step (b) follows from Assumption 5.

- For the second term B_2: Note that π^*_{t+1} and π^*_t are the optimal policy for the regularized $\mathrm{MDP}_{\mu_{t+1}}$ and MDP_{μ_t}, respectively. Define

$$\tau:=\frac{1}{|\mathcal{A}|}\exp\left(-\frac{R_{\max}+\gamma\lambda\log|\mathcal{A}|}{\lambda(1-\gamma)}\right).$$

  By Lemma 6, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\pi^*_{t+1}(a|s)\geq\tau,\ \text{and}\ \pi^*_t(a|s)\geq\tau.$$

  Applying Lemma 3 yields

$$\begin{aligned} B_2 &\leq \kappa\mathbb{E}_{s\sim\rho^*_{t+1}}\left[\left\|\pi^*_t(\cdot|s)-\pi^*_{t+1}(\cdot|s)\right\|_1\right] \\ &= \kappa\mathbb{E}_{s\sim\rho^*}\left[\frac{\rho^*_{t+1}(s)}{\rho^*(s)}\cdot\left\|\pi^*_t(\cdot|s)-\pi^*_{t+1}(\cdot|s)\right\|_1\right] \\ &\leq \kappa C_\rho\mathbb{E}_{s\sim\rho^*}\left[\left\|\pi^*_t(\cdot|s)-\pi^*_{t+1}(\cdot|s)\right\|_1\right] && \text{Assumption 4} \\ &= \kappa C_\rho D\left(\Gamma^\lambda_1(\mu_t),\Gamma^\lambda_1(\mu_{t+1})\right) \\ &\leq \kappa C_\rho d_1\left\|\mu_t-\mu_{t+1}\right\|_{\mathcal{H}}, && \text{Assumption 2} \end{aligned} \tag{40}$$

  where

$$\begin{aligned} \kappa &:= 1+\log\frac{1}{\min\left\{\tau,\frac{\eta}{|\mathcal{A}|}\right\}} \\ &\leq 2\max\left\{\log\frac{|\mathcal{A}|}{\eta},\frac{2}{1-\gamma}\log|\mathcal{A}|+\frac{R_{\max}}{\lambda(1-\gamma)}\right\} \\ &\leq \frac{4}{1-\gamma}\log\frac{|\mathcal{A}|}{\eta}+\frac{2R_{\max}}{\lambda(1-\gamma)} \\ &= \frac{4}{1-\gamma}\mathrm{KL}_{\max}+\frac{2R_{\max}}{\lambda(1-\gamma)}. \end{aligned}$$

Combining (37), (38), (40) and (39), we have

$$\begin{aligned} &\mathbb{E}_{\rho^*_{t+1}}\left[D_{\mathrm{KL}}\left(\pi^*_{t+1}\|\pi_{t+1}\right)\right] \\ \leq&\mathbb{E}_{\rho^*_t}\left[D_{\mathrm{KL}}\left(\pi^*_t\|\pi_{t+1}\right)\right]+\mathrm{KL}_{\max}\cdot d_0\left\|\mu_{t+1}-\mu_t\right\|_{\mathcal{H}}+\kappa C_\rho d_1\left\|\mu_t-\mu_{t+1}\right\|_{\mathcal{H}} \\ \leq&(1-\alpha_t\lambda)\mathbb{E}_{\rho^*_t}\left[D_{\mathrm{KL}}\left(\pi^*_t\|\pi_t\right)\right]+\mathbb{E}_{s\sim\rho^*_t}\left[Y_t(s)\right]+\left(d_0\cdot\mathrm{KL}_{\max}+\kappa C_\rho d_1\right)\left\|\mu_{t+1}-\mu_t\right\|_{\mathcal{H}}. \end{aligned} \tag{41}$$

Note that

$$\begin{aligned} \mathbb{E}_{s\sim\rho^*_t}\left[Y_t(s)\right] &= 2\alpha_t\mathbb{E}_{s\sim\rho^*_t}\left[\left\|Q^\lambda_t(s,\cdot)-\widehat{Q}^\lambda_t(s,\cdot)\right\|_\infty\right]+\frac{\left\|\widehat{Q}^\lambda_t\right\|^2_\infty}{2}\alpha_t^2+2\eta \\ &\leq 2\alpha_t\sqrt{\mathbb{E}_{s\sim\rho^*_t}\left[\left\|Q^\lambda_t(s,\cdot)-\widehat{Q}^\lambda_t(s,\cdot)\right\|^2_\infty\right]}+\frac{\left\|\widehat{Q}^\lambda_t\right\|^2_\infty}{2}\alpha_t^2+2\eta \\ &\leq 2\varepsilon_Q\alpha_t+\frac{Q^2_{\max}}{2}\alpha_t^2+2\eta, \end{aligned} \tag{42}$$

where the last step holds by the assumption on the policy evaluation error and the fact that $\widehat{Q}_t^{\lambda} : \mathcal{S} \times \mathcal{A} \to [0, Q_{\max}]$ satisfies $\left\|\widehat{Q}_t^{\lambda}\right\|_{\infty} \leq Q_{\max}$ by definition. Combining (41) and (42) proves the lemma.

C.4. Proof of Lemma 8

Proof. According to the update rule (7) for the embedded mean-field state, we have

$$\begin{aligned}
&\|\mu_{t+1} - \mu^*\|_{\mathcal{H}} \\
=&\|(1-\beta_t)\mu_t + \beta_t \Gamma_2(\pi_t, \mu_t) - \mu^*\|_{\mathcal{H}} \\
=&\left\|(1-\beta_t)\left(\mu_t - \mu^*\right) + \beta_t\left(\Gamma_2\left(\Gamma_1^{\lambda}(\mu_t), \mu_t\right) - \mu^*\right) - \beta_t\left[\Gamma_2\left(\Gamma_1^{\lambda}(\mu_t), \mu_t\right) - \Gamma_2(\pi_t, \mu_t)\right]\right\|_{\mathcal{H}} \\
\leq&(1-\beta_t)\left\|\mu_t - \mu^*\right\|_{\mathcal{H}} + \beta_t\left\|\Gamma_2\left(\Gamma_1^{\lambda}(\mu_t), \mu_t\right) - \mu^*\right\|_{\mathcal{H}} + \beta_t\left\|\Gamma_2\left(\Gamma_1^{\lambda}(\mu_t), \mu_t\right) - \Gamma_2(\pi_t, \mu_t)\right\|_{\mathcal{H}} \\
\overset{(i)}{=}&(1-\beta_t)\left\|\mu_t - \mu^*\right\|_{\mathcal{H}} + \beta_t \underbrace{\left\|\Gamma_2\left(\Gamma_1^{\lambda}(\mu_t), \mu_t\right) - \Gamma_2\left(\Gamma_1^{\lambda}(\mu^*), \mu^*\right)\right\|_{\mathcal{H}}}_{(a)} \\
&+ \beta_t \underbrace{\left\|\Gamma_2\left(\Gamma_1^{\lambda}(\mu_t), \mu_t\right) - \Gamma_2(\pi_t, \mu_t)\right\|_{\mathcal{H}}}_{(b)},
\end{aligned} \tag{43}$$

where the equality (i) follows from the fact that $\mu^* = \Gamma_2\left(\Gamma_1^{\lambda}(\mu^*), \mu^*\right)$.

Lemma 1 implies that $\Lambda(\mu) = \Gamma_2\left(\Gamma_1^{\lambda}(\mu), \mu\right)$ is $d_1 d_2 + d_3$ Lipschitz. It follows that

$$(a) \leq (d_1 d_2 + d_3)\left\|\mu_t - \mu^*\right\|_{\mathcal{H}}. \tag{44}$$

By Assumption 3, we have

$$(b) \leq d_2 D\left(\Gamma_1^{\lambda}(\mu_t), \pi_t\right). \tag{45}$$

Combining Eqs. (43)-(45) yields

$$\|\mu_{t+1} - \mu^*\|_{\mathcal{H}} \leq \left(1 - \beta_t \overline{d}\right)\|\mu_t - \mu^*\|_{\mathcal{H}} + d_2 \beta_t D\left(\Gamma_1^{\lambda}(\mu_t), \pi_t\right) \tag{46}$$

where $\overline{d} = 1 - d_1 d_2 - d_3 > 0$.

Let us bound the second RHS term above. By the definition of policy distance D in equation (11), we have

$$\begin{aligned}
D\left(\Gamma_1^{\lambda}(\mu_t), \pi_t\right) &= \mathbb{E}_{\rho^*}\left[\left\|\Gamma_1^{\lambda}(\mu_t) - \pi_t\right\|_1\right] \\
&= \mathbb{E}_{s\sim\rho^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\|_1\right] \\
&= \mathbb{E}_{s\sim\rho_t^*}\left[\frac{\rho^*(s)}{\rho_t^*(s)}\left\|\pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\|_1\right] \\
&\leq \left\{\mathbb{E}_{s\sim\rho_t^*}\left[\left|\frac{\rho^*(s)}{\rho_t^*(s)}\right|^2\right] \cdot \mathbb{E}_{s\sim\rho_t^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\|_1^2\right]\right\}^{1/2} \\
&\leq \overline{C}_{\rho}\sqrt{\mathbb{E}_{s\sim\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_t(\cdot|s)\right)\right]},
\end{aligned} \tag{47}$$

where the first inequality holds due to Cauchy-Schwartz inequality, the last inequality follows from Assumption 4 and Pinsker's inequality.

Combining (46)-(47) gives

$$\|\mu_{t+1} - \mu^*\|_{\mathcal{H}} \leq \left(1 - \beta_t \overline{d}\right)\|\mu_t - \mu^*\|_{\mathcal{H}} + d_2 \beta_t \overline{C}_{\rho}\sqrt{\mathbb{E}_{\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*\|\pi_t\right)\right]}.$$

This completes the proof. □

D. Proof of Corollary 1

Proof. Note that for each $t \in [T-1]$, we have

$$\begin{aligned} D(\pi_t, \pi^*) &\le D\left(\pi_t, \pi_t^*\right) + D\left(\pi_t^*, \pi^*\right) \\ &= D\left(\pi_t, \pi_t^*\right) + D\left(\Gamma_1^\lambda(\mu_t), \Gamma_1^\lambda(\mu^*)\right) \\ &\le D\left(\pi_t, \pi_t^*\right) + d_1 \left\|\mu_t - \mu^*\right\|_{\mathcal{H}}, \end{aligned}$$

where the last step follows from Assumption 2 on the Lipschitzness of Γ_1^λ. It follows that

$$\begin{aligned} & D\left(\frac{1}{T}\sum_{t=0}^{T-1}\pi_t, \pi^*\right) + \left\|\frac{1}{T}\sum_{t=0}^{T-1}\mu_t - \mu^*\right\|_{\mathcal{H}} \\ \le & \frac{1}{T}\sum_{t=0}^{T-1} D\left(\pi_t, \pi^*\right) + \frac{1}{T}\sum_{t=0}^{T-1}\left\|\mu_t - \mu^*\right\|_{\mathcal{H}} \\ \le & \frac{1}{T}\sum_{t=0}^{T-1}\left(D\left(\pi_t, \pi_t^*\right) + d_1\left\|\mu_t - \mu^*\right\|_{\mathcal{H}}\right) + \frac{1}{T}\sum_{t=0}^{T-1}\left\|\mu_t - \mu^*\right\|_{\mathcal{H}} \\ \lesssim & \frac{1}{\sqrt{\lambda}}\left(\frac{\sqrt{\log T}}{T^{1/5}} + \sqrt{\varepsilon_Q}\right), \end{aligned}$$

where in the last step we apply the bounds (12) and (13) in Theorem 1. □

E. Guarantees under Weaker Assumption On Concentrability

In this section, we show that the ℓ_∞ condition on concentrability coefficient in Assumption 4 can be relaxed to an ℓ_2 condition of the form $\left\{\mathbb{E}\left[\left|\rho_\mu^{\pi_\mu^{\lambda,*}}(s)/\rho^*(s)\right|^2\right]\right\}^{1/2} \le C_\rho$, under which we can establish an $\widetilde{O}(T^{-1/9})$ convergence rate.

We now provided the details. Consider the following distance metric between two policies $\pi, \pi' \in \Pi$:

$$W(\pi, \pi') := \sqrt{\mathbb{E}_{s\sim\rho^*}\left[\left\|\pi(\cdot|s) - \pi'(\cdot|s)\right\|_1^2\right]}. \tag{48}$$

Similarly as before, we assume certain Lipschitz properties for the two mappings $\Gamma_1^\lambda : \mathcal{M} \to \Pi$ and $\Gamma_2 : \Pi \times \mathcal{M} \to \mathcal{M}$ defined in Section 2.3. In particular, we impose the following two assumtpions, both stated in terms of the new distance metric $W(\cdot,\cdot)$ defined in (48) above.

Assumption 6. *There exists a constant $d_1 > 0$, such that for any $\mu, \mu' \in \mathcal{M}$, it holds that*

$$W\left(\Gamma_1^\lambda(\mu), \Gamma_1^\lambda(\mu')\right) \le d_1 \left\|\mu - \mu'\right\|_{\mathcal{H}}.$$

Assumption 7. *There exist constants $d_2 > 0, d_3 > 0$ such that for any policies $\pi, \pi' \in \Pi$ and embedded mean-field states $\mu, \mu' \in \mathcal{M}$, it holds that*

$$\begin{aligned} \left\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi', \mu)\right\|_{\mathcal{H}} &\le d_2 W\left(\pi, \pi'\right), \\ \left\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi, \mu')\right\|_{\mathcal{H}} &\le d_3 \left\|\mu - \mu'\right\|_{\mathcal{H}}. \end{aligned}$$

Assumptions 6 and 7 immediately imply Lipschitzness of the composite mapping $\Lambda^\lambda : \mathcal{M} \to \mathcal{M}$, which we recall is defined as $\Lambda^\lambda(\mu) = \Gamma_2\left(\Gamma_1^\lambda(\mu), \mu\right)$.

Lemma 9. *Suppose Assumptions 6 and 7 hold. Then for each $\mu, \mu' \in \mathcal{M}$, it holds that*

$$\left\|\Lambda^\lambda(\mu) - \Lambda^\lambda(\mu')\right\|_{\mathcal{H}} \le \left(d_1 d_2 + d_3\right)\left\|\mu - \mu'\right\|_{\mathcal{H}}.$$

We also consider the following relaxed, ℓ_2-type assumption on the concentrability coefficients.

Assumption 8 (Finite Concentrability Coefficients)**.** *There exist two constants $C_\rho, \overline{C}_\rho > 0$ such that for each $\mu \in \mathcal{M}$, it holds that*

$$\left\{\mathbb{E}_{s\sim\rho_\mu^{\pi_\mu^{\lambda,*}}}\left[\left|\frac{\rho_\mu^{\pi_\mu^{\lambda,*}}(s)}{\rho^*(s)}\right|^2\right]\right\}^{1/2} \le C_\rho \qquad \textit{and} \qquad \left\{\mathbb{E}_{s\sim\rho_\mu^{\pi_\mu^{\lambda,*}}}\left[\left|\frac{\rho^*(s)}{\rho_\mu^{\pi_\mu^{\lambda,*}}(s)}\right|^2\right]\right\}^{1/2} \le \overline{C}_\rho.$$

With the above assumptions and the distance metric W, we can establish the following convergence result for Algorithm 1.

Theorem 2. *Suppose that Assumptions 1, 5, 6, 7, and 8 hold and $d_1 d_2 + d_3 < 1$ and that the error in the policy evaluation step in Algorithm 1 satisfies*

$$\mathbb{E}_{s\sim\rho_t^*}\left[\left\|Q_t^\lambda(s,\cdot) - \widehat{Q}_t^\lambda(s,\cdot)\right\|_\infty^2\right] \le \varepsilon_Q^2, \qquad \forall t \in [T].$$

With the choice of

$$\eta = c_\eta T^{-1}, \qquad \alpha_t \equiv \alpha = c_\alpha T^{-4/9}, \qquad \beta_t \equiv \beta = c_\beta T^{-8/9},$$

for some universal constants $c_\eta > 0$, $c_\alpha > 0$ and $c_\beta > 0$ in Algorithm 1, the resulting policy and embedded mean-field state sequence $\{(\pi_t, \mu_t)\}_{t=0}^T$ satisfy

$$W\left(\frac{1}{T}\sum_{t=0}^{T-1}\pi_t, \frac{1}{T}\sum_{t=0}^{T-1}\pi_t^*\right) \le \frac{1}{T}\sum_{t=0}^{T-1} W(\pi_t, \pi_t^*) \lesssim \frac{1}{\lambda^{1/4}}\left(\frac{(\log T)^{1/4}}{T^{1/9}} + \varepsilon_Q^{1/4}\right), \tag{49}$$

$$\left\|\frac{1}{T}\sum_{t=0}^{T-1}\mu_t - \mu^*\right\|_{\mathcal{H}} \le \frac{1}{T}\sum_{t=0}^{T-1}\|\mu_t - \mu^*\|_{\mathcal{H}} \lesssim \frac{1}{\lambda^{1/4}}\left(\frac{(\log T)^{1/4}}{T^{1/9}} + \varepsilon_Q^{1/4}\right). \tag{50}$$

The following corollary of Theorem 2 shows that after T iterations of our algorithm, the average policy-population pair $\left(\frac{1}{T}\sum_{t=0}^{T-1}\pi_t, \frac{1}{T}\sum_{t=0}^{T-1}\mu_t\right)$ is an $\widetilde{\mathcal{O}}\left(T^{-1/9}\right)$-approximate NE.

Corollary 2. *Under the assumptions of Theorem 2, we have*

$$W\left(\frac{1}{T}\sum_{t=0}^{T-1}\pi_t, \pi^*\right) + \left\|\frac{1}{T}\sum_{t=0}^{T-1}\mu_t - \mu^*\right\|_{\mathcal{H}} \lesssim \frac{1}{\lambda^{1/4}}\left(\frac{(\log T)^{1/4}}{T^{1/9}} + \varepsilon_Q^{1/4}\right).$$

of Theorem 2. The proof follows similar lines as those of Theorem 1 and Corollary 1, with all appearances of the distance D replaced by the new distance W. Below we only point out the modifications needed.

Lemma 7 remains valid as stated. For the proof of this lemma, the only different step is bounding the term B_2 in equation (38). In particular, the bounds in equation (40) should be replaced by the following:

$$\begin{aligned}
B_2 &\le \kappa\mathbb{E}_{s\sim\rho_{t+1}^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_{t+1}^*(\cdot|s)\right\|_1\right] \\
&= \kappa\mathbb{E}_{s\sim\rho^*}\left[\frac{\rho_{t+1}^*(s)}{\rho^*(s)}\cdot\left\|\pi_t^*(\cdot|s) - \pi_{t+1}^*(\cdot|s)\right\|_1\right] \\
&\le \kappa\sqrt{\mathbb{E}_{s\sim\rho^*}\left[\left(\frac{\rho_{t+1}^*(s)}{\rho^*(s)}\right)^2\right]\cdot\mathbb{E}_{s\sim\rho^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_{t+1}^*(\cdot|s)\right\|_1^2\right]} \\
&\le \kappa C_\rho\cdot\sqrt{\mathbb{E}_{s\sim\rho^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_{t+1}^*(\cdot|s)\right\|_1^2\right]} && \text{Assumption 8} \\
&= \kappa C_\rho W\left(\Gamma_1^\lambda(\mu_t), \Gamma_1^\lambda(\mu_{t+1})\right) \\
&\le \kappa C_\rho d_1\|\mu_t - \mu_{t+1}\|_{\mathcal{H}}. && \text{Assumption 6}
\end{aligned} \tag{51}$$

Lemma 8 should be replaced by the following lemma.

Lemma 10. *Under the setting of Theorem 2, for each $t \geq 0$, we have*

$$\sigma_\mu^{t+1} \leq \left(1 - \beta_t \overline{d}\right) \sigma_\mu^t + d_2 \sqrt{\overline{C}_\rho} \beta_t \left(\sigma_\pi^t\right)^{1/4},$$

where $\overline{d} = 1 - d_1 d_2 - d_3 > 0$.

The proof of Lemma 10 is similar to that of Lemma 8. The only different step is the term $D\left(\Gamma_1^\lambda(\mu_t), \pi_t\right)$ in equation (46) should be replaced by $W\left(\Gamma_1^\lambda(\mu_t), \pi_t\right)$, which can be bounded as follows:

$$\begin{aligned}
W\left(\Gamma_1^\lambda(\mu_t), \pi_t\right) &= \sqrt{\mathbb{E}_{s\sim\rho^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\|_1^2\right]} \\
&= \sqrt{\mathbb{E}_{s\sim\rho_t^*}\left[\frac{\rho^*(s)}{\rho_t^*(s)}\left\|\pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\|_1^2\right]} \\
&\leq \left\{\mathbb{E}_{s\sim\rho_t^*}\left[\left|\frac{\rho^*(s)}{\rho_t^*(s)}\right|^2\right] \cdot \mathbb{E}_{s\sim\rho_t^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\|_1^4\right]\right\}^{1/4} \\
&\overset{(i)}{\lesssim} \sqrt{\overline{C}_\rho} \cdot \left\{\mathbb{E}_{s\sim\rho_t^*}\left[\left\|\pi_t^*(\cdot|s) - \pi_t(\cdot|s)\right\|_1^2\right]\right\}^{1/4} \\
&\overset{(ii)}{\lesssim} \sqrt{\overline{C}_\rho} \left\{\mathbb{E}_{s\sim\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s)\|\pi_t(\cdot|s)\right)\right]\right\}^{1/4}.
\end{aligned} \tag{52}$$

where step (i) holds by Assumption 8 and the fact that $\|\nu - \nu'\|_1 \leq 2, \forall \nu, \nu' \in \Delta(\mathcal{A})$, and step (ii) follows Pinsker's inequality.

We now turn to the proof of Theorem 2.

We first establish the convergence for σ_π^t by following the exactly same steps from equation (22) up to equation (26). We restate the bound on $\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\pi^t$ in (26) as follows:

$$\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\pi^t \leq \frac{1}{T\lambda\alpha}\sigma_\pi^0 + \frac{\overline{C}_1\beta}{\lambda\alpha} + \frac{2\varepsilon_Q}{\lambda} + \frac{Q_{\max}^2}{2\lambda}\alpha + \frac{2\eta}{\lambda\alpha}. \tag{53}$$

When choosing $\alpha = \mathcal{O}(T^{-4/9})$, $\beta = \mathcal{O}(T^{-8/9})$ and $\eta = \mathcal{O}(T^{-1})$, we have $\overline{C}_1 = \mathcal{O}(\log T)$. Therefore, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\pi^t \lesssim \frac{\log T}{\lambda T^{4/9}} + \frac{2\varepsilon_Q}{\lambda}. \tag{54}$$

If we let T be a random number sampled uniformly from $\{0, \ldots, T-1\}$, then the above equation can be written equivalently as

$$\mathbb{E}_{\mathsf{T}}\left[\sigma_\pi^{\mathsf{T}}\right] \lesssim \frac{\log T}{\lambda T^{4/9}} + \frac{2\varepsilon_Q}{\lambda}. \tag{55}$$

We now proceed to bound the average embedded mean-field state $\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\mu^t$. Lemma 10 implies

$$\sigma_\mu^t \leq \frac{1}{\overline{d}\beta_t}\left(\sigma_\mu^t - \sigma_\mu^{t+1}\right) + \frac{d_2\sqrt{\overline{C}_\rho}}{\overline{d}}\left(\sigma_\pi^t\right)^{1/4}. \tag{56}$$

With $\beta_t \equiv \beta = \mathcal{O}(T^{-8/9})$, averaging equation (56) over iteration $t = 0, \ldots, T-1$, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\mu^t &\leq \frac{1}{\overline{d}\beta T}\left(\sigma_\mu^0 - \sigma_\mu^T\right) + \frac{d_2\sqrt{\overline{C}_\rho}}{\overline{d}T}\sum_{t=0}^{T-1}\left(\sigma_\pi^t\right)^{1/4} \\
&\leq \frac{\sigma_\mu^0}{\overline{d}\beta T} + \frac{d_2\sqrt{\overline{C}_\rho}}{\overline{d}T}\sum_{t=0}^{T-1}\left(\sigma_\pi^t\right)^{1/4} \\
&\overset{(i)}{\leq} \frac{\sigma_\mu^0}{\overline{d}\beta T} + \frac{d_2\sqrt{\overline{C}_\rho}}{\overline{d}}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sigma_\pi^t}} \\
&\overset{(ii)}{\leq} \frac{\sigma_\mu^0}{\overline{d}\beta T} + \frac{d_2\sqrt{\overline{C}_\rho}}{\overline{d}}\left(\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\pi^t\right)^{1/4}
\end{aligned}$$

where steps (i) and (ii) follow from Cauchy-Schwarz inequality.

From equation (54), we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\sigma_\mu^t &\lesssim \frac{\sigma_\mu^0}{\overline{d}}T^{-1/9} + \frac{d_2\sqrt{\overline{C}_\rho}}{\overline{d}}\left(\frac{\log T}{\lambda T^{4/9}} + \frac{2\varepsilon_Q}{\lambda}\right)^{1/4} \\
&\lesssim \left(\frac{\log T}{\lambda T^{4/9}} + \frac{2\varepsilon_Q}{\lambda}\right)^{1/4} \\
&\lesssim \frac{1}{\lambda^{1/4}}\left(\frac{(\log T)^{1/4}}{T^{1/9}} + \varepsilon_Q^{1/4}\right).
\end{aligned}$$

This equation, together with Jensen's inequality, proves equation (50) in Theorem 2.

Turning to equation (49) in Theorem 2, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} W\left(\pi_t, \pi_t^*\right) &= \mathbb{E}_{\mathsf{T}}\left[W\left(\pi_{\mathsf{T}}, \pi_{\mathsf{T}}^*\right)\right] \\
&= \mathbb{E}_{\mathsf{T}}\sqrt{\mathbb{E}_{s\sim\rho^*}\left[\left\|\pi_{\mathsf{T}}^*(\cdot|s) - \pi_{\mathsf{T}}(\cdot|s)\right\|_1^2\right]} \\
&\overset{(i)}{\leq} \sqrt{\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[\frac{\rho^*(s)}{\rho_{\mathsf{T}}^*(s)}\left\|\pi_{\mathsf{T}}^*(\cdot|s) - \pi_{\mathsf{T}}(\cdot|s)\right\|_1^2\right]} \\
&\overset{(ii)}{\leq} \left\{\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[\left|\frac{\rho^*(s)}{\rho_{\mathsf{T}}^*(s)}\right|^2\right]\cdot\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[\left\|\pi_{\mathsf{T}}^*(\cdot|s) - \pi_{\mathsf{T}}(\cdot|s)\right\|_1^4\right]\right\}^{1/4} \\
&\overset{(iii)}{\lesssim} \left\{\overline{C}_\rho^2\cdot\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[\left\|\pi_{\mathsf{T}}^*(\cdot|s) - \pi_{\mathsf{T}}(\cdot|s)\right\|_1^2\right]\right\}^{1/4} \\
&\overset{(iv)}{\lesssim} \sqrt{\overline{C}_\rho}\cdot\left\{\mathbb{E}_{\mathsf{T}}\mathbb{E}_{s\sim\rho_{\mathsf{T}}^*}\left[D_{\mathrm{KL}}\left(\pi_{\mathsf{T}}^*(\cdot|s)\|\pi_{\mathsf{T}}(\cdot|s)\right)\right]\right\}^{1/4} \\
&= \sqrt{\overline{C}_\rho}\cdot\left\{\mathbb{E}_{\mathsf{T}}\left[\sigma_\pi^{\mathsf{T}}\right]\right\}^{1/4} \\
&\overset{(v)}{\lesssim} \frac{1}{\lambda^{1/4}}\left(\frac{(\log T)^{1/4}}{T^{1/9}} + \varepsilon_Q^{1/4}\right),
\end{aligned}$$

where step (i) holds due to Jensen's inequality, step (ii) follows from Cauchy-Schwarz inequality, step (iii) follows from Assumption 8 and the fact that $\|\nu - \nu'\|_1 \leq 2, \forall \nu, \nu' \in \Delta(\mathcal{A})$, step (iv) comes from Pinsker's inequality, and step (v) follows from the bound in equation (55). The above equation, together with Jensen's inequality, proves equation (49). We have completed the proof of Theorem 2.

□

The proof of Corollary 2 is the same as that of Corollary 1 and is omitted here.