

## A Algorithms

---

### Algorithm 1 Multi-agent Soft Q-learning

---

**Result:** policy  $\pi^i$ , opponent model  $\rho^i$

**Initialization:**

Initialize replay buffer  $\mathcal{M}$  to capacity  $M$ .

Initialize  $Q_{\omega^i}(s, a^i, a^{-i})$  with random parameters  $\omega^i$ ,  $P(a^{-i}|s)$  arbitrarily, set  $\gamma$  as the discount factor.

Initialize target  $Q_{\bar{\omega}^i}(s, a^i, a^{-i})$  with random parameters  $\bar{\omega}^i$ , set  $C$  the target parameters update interval.

**while** not converge **do**

**Collect experience**

For the current state  $s_t$  compute the opponent model  $\rho^i(a_t^{-i}|s_t)$  and conditional policy  $\pi^i(a_t^i|s_t, a_t^{-i})$  respectively from:

$$\rho^i(a_t^{-i}|s_t) \propto P(a_t^{-i}|s_t) \left( \sum_{a_t^i} \exp\left(\frac{1}{\alpha} Q_{\omega^i}(s_t, a_t^i, a_t^{-i})\right) \right)^\alpha,$$

$$\pi^i(a_t^i|s_t, \hat{a}_t^{-i}) \propto \exp\left(\frac{1}{\alpha} Q_{\omega^i}(s_t, a_t^i, \hat{a}_t^{-i})\right).$$

Compute the marginal policy  $\pi^i(a_t^i|s_t)$  and sample an action from it:

$$a_t^i \sim \pi^i(a_t^i|s_t) = \sum_{a_t^{-i}} \pi^i(a_t^i|s_t, a_t^{-i}) \rho(a_t^{-i}|s_t).$$

Observe next state  $s_{t+1}$ , opponent action  $a_t^{-i}$  and reward  $r_t^i$ , save the new experience in the replay buffer:

$$\mathcal{M} \leftarrow \mathcal{M} \cup \{(s_t, a_t^i, a_t^{-i}, s_{t+1}, r_t^i)\}.$$

Update the prior from the replay buffer:

$$P(a_t^{-i}|s_t) = \frac{\sum_{m=1}^{|\mathcal{M}|} \mathbb{I}(s = s_t, a^{-i} = a_t^{-i})}{\sum_{m=1}^{|\mathcal{M}|} \mathbb{I}(s = s_t)} \forall s_t, a_t^{-i} \in \mathcal{M}.$$

**Sample a mini-batch from the replay buffer:**

$$\{s_t^{(n)}, a_t^{i,(n)}, a_t^{-i,(n)}, s_{t+1}^{(n)}, r_t^{(n)}\}_{n=1}^N \sim \mathcal{M}.$$

**Update**  $Q_{\omega^i}(s, a^i, a^{-i})$ :

**for** each tuple  $(s_t^{(n)}, a_t^{i,(n)}, a_t^{-i,(n)}, s_{t+1}^{(n)}, r_t^{(n)})$  **do**

Sample  $\{a^{-i,(n,k)}\}_{k=1}^K \sim \rho$ ,  $\{a^{i,(n,k)}\}_{k=1}^K \sim \pi$ .

Compute empirical  $\bar{V}^i(s_{t+1}^{(n)})$  as:

$$\bar{V}^i(s_{t+1}^{(n)}) = \log \left( \frac{1}{K} \sum_{k=1}^K \frac{\left( P_{\alpha}^{\frac{1}{\alpha}}(a^{-i,(n,k)}|s_{t+1}^{(n)}) \exp\left(\frac{1}{\alpha} Q_{\bar{\omega}^i}(s_{t+1}^{(n)}, a^{i,(n,k)}, a^{-i,(n,k)})\right) \right)^\alpha}{\pi(a^{i,(n,k)}|s_{t+1}^{(n)}, a^{-i,(n,k)}) \rho(a^{-i,(n,k)}|s_{t+1}^{(n)})} \right).$$

Set

$$y^{(n)} = \begin{cases} r_t^{(n)} & \text{for terminal } s_{t+1}^{(n)} \\ r_t^{(n)} + \gamma \bar{V}^i(s_{t+1}^{(n)}) & \text{for non-terminal } s_{t+1}^{(n)} \end{cases}$$

Perform gradient descent step on  $(y^{(n)} - Q_{\omega^i}(s_{t+1}^{(n)}, a^{i,(n)}, a^{-i,(n)}))^2$  with respect to parameters  $\omega^i$

Every  $C$  gradient descent steps, reset target parameters:

$$\bar{\omega}^i \leftarrow \omega$$

**end for**

**end while**

**Compute converged**  $\pi^i$  **and**  $\rho^i$

---

---

**Algorithm 2** Multi-agent Variational Actor Critic

---

**Result:** policy  $\pi_{\theta^i}$ , opponent model  $\rho_{\phi^i}$

**Initialization:**

Initialize parameters  $\theta^i, \phi^i, \omega^i, \psi^i$  for each agent  $i$  and the random process  $\mathcal{N}$  for action exploration.

Assign target parameters of joint action Q-function:  $\bar{\omega}^i \leftarrow \omega$ .

Initialize learning rates  $\lambda_V, \lambda_Q, \lambda_\pi, \lambda_\phi, \alpha$ , and set  $\gamma$  as the discount factor.

**for** Each episode  $d = (1, \dots, D)$  **do**

Initialize random process  $\mathcal{N}$  for action exploration.

**for** each time step  $t$  **do**

For the current state  $s_t$ , sample an action and opponent's action using:

$\hat{a}_t^{-i} \leftarrow g_{\phi^{-i}}(\epsilon_t^{-i}; s_t)$ , where  $\epsilon_t^{-i} \sim \mathcal{N}$ ,

$a_t^i \leftarrow f_{\theta^i}(\epsilon_t^i; s_t, \hat{a}_t^{-i})$ , where  $\epsilon_t^i \sim \mathcal{N}$ .

Observe next state  $s_{t+1}$ , opponent action  $a_t^{-i}$  and reward  $r_t^i$ , save the new experience in the replay buffer:

$$\mathcal{D}^i \leftarrow \mathcal{D}^i \cup \{(s_t, a_t^i, a_t^{-i}, \hat{a}_t^{-i}, s_{t+1}, r_t^i)\}.$$

Update the prior from the replay buffer:

$$\psi^i = \arg \max \mathbb{E}_{\mathcal{D}^i}[-P(a^{-i}|s) \log P_{\psi^i}(a^{-i}|s)]$$

Sample a mini-batch from the reply buffer:

$$\{s_t^{(n)}, a_t^{i,(n)}, a_t^{-i,(n)}, \hat{a}_t^{-i,(n)}, s_{t+1}^{(n)}, r_t^{(n)}\}_{n=1}^N \sim \mathcal{M}.$$

For the state  $s_{t+1}^{(n)}$ , sample an action and opponent's action using:

$\hat{a}_{t+1}^{-i,(n)} \leftarrow g_{\phi^{-i}}(\epsilon_{t+1}^{-i}; s_{t+1}^{(n)})$ , where  $\epsilon_{t+1}^{-i} \sim \mathcal{N}$ ,

$a_{t+1}^{i,(n)} \leftarrow f_{\theta^i}(\epsilon_{t+1}^i; s_{t+1}^{(n)}, \hat{a}_{t+1}^{-i,(n)})$ , where  $\epsilon_{t+1}^i \sim \mathcal{N}$ .

$\bar{V}^i(s_{t+1}^{(n)}) = Q_{\bar{\omega}}(s_{t+1}^{(n)}, a_{t+1}^{i,(n)}, \hat{a}_{t+1}^{-i,(n)}) - \alpha \log \pi_{\theta^i}(a_{t+1}^{i,(n)} | s_{t+1}^{(n)}, \hat{a}_{t+1}^{-i,(n)}) - \log \rho_{\phi^i}(\hat{a}_{t+1}^{-i,(n)} | s_{t+1}^{(n)}) + \log P_{\psi^i}(\hat{a}_{t+1}^{-i,(n)} | s_{t+1}^{(n)})$ .

Set

$$y^{(n)} = \begin{cases} r_t^{(n)} & \text{for terminal } s_{t+1}^{(n)} \\ r_t^{(n)} + \gamma \bar{V}^i(s_{t+1}^{(n)}) & \text{for non-terminal } s_{t+1}^{(n)} \end{cases}$$

$$\nabla_{\omega^i} \mathcal{J}_Q(\omega^i) = \nabla_{\omega^i} Q_{\omega^i}(s_t^{(n)}, a_t^{i,(n)}, a_t^{-i,(n)})(Q_{\omega^i}(s_t^{(n)}, a_t^{i,(n)}, a_t^{-i,(n)}) - y^{(n)})$$

$$\nabla_{\theta^i} \mathcal{J}_\pi(\theta^i) = \nabla_{\theta^i} \alpha \log \pi_{\theta^i}(a_t^{i,(n)} | s_t^{(n)}, \hat{a}_t^{-i,(n)})$$

$$+ (\nabla_{a_t^{i,(n)}} \alpha \log \pi_{\theta^i}(a_t^{i,(n)} | s_t^{(n)}, \hat{a}_t^{-i,(n)}) - \nabla_{a_t^{i,(n)}} Q_{\omega^i}(s_t^{(n)}, a_t^{i,(n)}, \hat{a}_t^{-i,(n)})) \nabla_{\theta^i} f_{\theta^i}(\epsilon_t^i; s_t^{(n)}, \hat{a}_t^{-i,(n)})$$

$$\nabla_{\phi^i} \mathcal{J}_\rho(\phi^i) = \nabla_{\phi^i} \log \rho_{\phi^i}(\hat{a}_t^{-i,(n)} | s_t^{(n)})$$

$$+ (\nabla_{\hat{a}_t^{-i,(n)}} \log \rho_{\phi^i}(\hat{a}_t^{-i,(n)} | s_t^{(n)}) - \nabla_{\hat{a}_t^{-i,(n)}} \log P(\hat{a}_t^{-i,(n)} | s_t^{(n)}) - \nabla_{\hat{a}_t^{-i,(n)}} Q_{\omega^i}(s_t^{(n)}, a_t^{i,(n)}, \hat{a}_t^{-i,(n)}))$$

$$+ \nabla_{\hat{a}_t^{-i,(n)}} \alpha \log \pi_{\theta^i}(a_t^{i,(n)} | s_t^{(n)}, \hat{a}_t^{-i,(n)}) \nabla_{\phi^i} g_{\phi^i}(\epsilon_t^{-i}; s_t^{(n)})$$

Update parameters:

$$\omega^i = \omega^i - \lambda_Q \nabla_{\omega^i} \mathcal{J}_Q(\omega^i)$$

$$\theta^i = \theta^i - \lambda_\pi \nabla_{\theta^i} \mathcal{J}_\pi(\theta^i)$$

$$\phi^i = \phi^i - \lambda_\phi \nabla_{\phi^i} \mathcal{J}_\rho(\phi^i)$$

**end for**

Every  $C$  gradient descent steps, reset target parameters:

$$\bar{\omega}^i = \beta \omega^i + (1 - \beta) \bar{\omega}^i$$

**end for**

---

## B Variational Lower Bounds in Multi-agent Reinforcement Learning

### B.1 The Lower Bound of The Log Likelihood of Optimality

We can factorize  $P(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^{-i})$  as :

$$P(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^{-i}) = P(s_1) \prod_t P(s_{t+1} | s_t, a_t) P(a_t^i | a_t^{-i}, s_t, o_t^{-i}) P(a_t^{-i} | s_t, o_t^{-i}), \quad (24)$$

where  $P(a_t^i | a_t^{-i}, s_t, o_t^{-i})$  is the conditional policy of agent  $i$  when other agents  $-i$  achieve optimality. As agent  $i$  has no knowledge about rewards of other agents, we set  $P(a_t^i | a_t^{-i}, s_t, o_t^{-i}) \propto 1$ .

Analogously, we factorize  $q(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^i, o_{1:T}^{-i})$  as:

$$q(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^i, o_{1:T}^{-i}) = P(s_1) \prod_t P(s_{t+1} | s_t, a_t) q(a_t^i | a_t^{-i}, s_t, o_t^i, o_t^{-i}) q(a_t^{-i} | s_t, o_t^i, o_t^{-i}) \quad (25)$$

$$= P(s_1) \prod_t P(s_{t+1} | s_t, a_t) \pi(a_t^i | s_t, a_t^{-i}) \rho(a_t^{-i} | s_t), \quad (26)$$

where  $\pi(a_t^i | a_t^{-i}, s_t)$  is agent 1's conditional policy at optimum and  $\rho(a_t^{-i} | s_t)$  is agent 1's model about opponents' optimal policies.

With the above factorization, we have:

$$\begin{aligned} & \log P(o_{1:T}^i | o_{1:T}^{-i}) \\ &= \log \sum_{a_{1:T}^i, a_{1:T}^{-i}, s_{1:T}} P(o_{1:T}^i, a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^{-i}) \end{aligned} \quad (27)$$

$$\geq \sum q(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^i, o_{1:T}^{-i}) \log \frac{P(o_{1:T}^i, a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^{-i})}{q(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} | o_{1:T}^i, o_{1:T}^{-i})} \quad (28)$$

$$= \mathbb{E}_{(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} \sim q)} [\sum_{t=1}^T \log P(o_t^i | s_t, a_t^i, a_t^{-i}) + \log P(s_1) + \sum_{t=1}^T \log P(s_{t+1} | s_t, a_t^i, a_t^{-i})] \quad (29)$$

$$- \log P(s_1) - \sum_{t=1}^T \log P(s_{t+1} | s_t, a_t^i, a_t^{-i}) \quad (30)$$

$$- \sum_{t=1}^T \log \pi(a_t^i | s_t, a_t^{-i}) - \sum_{t=1}^T \log \frac{\rho(a_t^{-i} | s_t)}{P(a_t^{-i} | s_t, o_t^{-i})} + \sum_{t=1}^T \log P(a_t^{-i} | s_t, a_t^{-i}, o_t^{-i})] \quad (31)$$

$$= \mathbb{E}_{(a_{1:T}^i, a_{1:T}^{-i}, s_{1:T} \sim q)} [\sum_{t=1}^T R^i(s_t, a_t^i, a_t^{-i}) - \log \pi(a_t^i | s_t, a_t^{-i}) - \log \frac{\rho(a_t^{-i} | s_t)}{P(a_t^{-i} | s_t, o_t^{-i})} + 1] \quad (32)$$

$$= \sum_t \mathbb{E}_{(s_t, a_t^i, a_t^{-i}) \sim q} [R^i(s_t, a_t^i, a_t^{-i}) + H(\pi(a_t^i | s_t, a_t^{-i})) - D_{\text{KL}}(\rho(a_t^{-i} | s_t) || P(a_t^{-i} | s_t, o_t^{-i}))], \quad (33)$$

where we assume that given joint actions  $(a^i, a^{-i})$  and state  $s$ , the optimality of agent  $i = 1$  is independent of other agents' optimalities:

$$P(o^i | s, a^i, a^{-i}, o^{-i}) = P(o^i | s, a^i, a^{-i}). \quad (34)$$

## C Multi-Agent Soft-Q Learning

### C.1 Soft Q-Function

We define the soft state-action value function  $Q_{soft}^{\pi, \rho}(s, a, a^{-i})$  of agent  $i$  in a stochastic game as:

$$Q_{soft}^{\pi, \rho}(s_t, a_t^i, a_t^{-i}) = r_t + \mathbb{E}_{(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i}, \dots) \sim q} \left[ \sum_{l=1}^{\infty} \gamma^l (r_{t+l} + \alpha H(\pi(a_{t+l}^i | s_{t+1}, a_{t+l}^{-i})) - D_{KL}(\rho(a_{t+l}^{-i} | s_{t+1}) || P(a_{t+l}^{-i} | s_{t+1}))) \right] \quad (35)$$

$$= \mathbb{E}_{(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})} [r_t + \gamma (\alpha H(\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\rho(a_{t+1}^{-i} | s_{t+1}) || P(a_{t+1}^{-i} | s_{t+1}))) + Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})] \quad (36)$$

$$= \mathbb{E}_{(s_{t+1}, a_{t+1}^{-i})} [r_t + \gamma (\alpha H(\pi(\cdot | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\rho(a_{t+1}^{-i} | s_{t+1}) || P(a_{t+1}^{-i} | s_{t+1}))) + \mathbb{E}_{a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})]] \quad (37)$$

$$= \mathbb{E}_{(s_{t+1})} [r_t + \gamma (\mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [\alpha H(\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i}))] - D_{KL}(\rho(\cdot | s_{t+1}) || P(\cdot | s_{t+1}))) + \mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})]] \quad (38)$$

Then we can easily see that the objective in Eq. 7 can be rewritten as:

$$\mathcal{J}(\pi, \phi) = \sum_t \mathbb{E}_{(s_t, a_t^i, a_t^{-i}) \sim (p_s, \pi, \rho)} [Q_{soft}^{\pi, \rho}(s_t, a_t^i, a_t^{-i}) + \alpha H(\pi(a_t^i | s_t, a_t^{-i})) - D_{KL}(\rho(a_t^{-i} | s_t) || P(a_t^{-i} | s_t))], \quad (39)$$

by setting  $\alpha = 1$ .

### C.2 Policy Improvement and Opponent Model Improvement

**Theorem 4.** (Policy improvement theorem) Given a conditional policy  $\pi$  and opponent model  $\rho$ , define a new conditional policy  $\tilde{\pi}$  as

$$\tilde{\pi}(\cdot | s, a^{-i}) \propto \exp\left(\frac{1}{\alpha} Q_{soft}^{\pi, \rho}(s, \cdot, a^{-i})\right), \forall s, a^{-i}. \quad (40)$$

Assume that throughout our computation,  $Q$  is bounded and  $\sum_{a^i} Q(s, a^i, a^{-i})$  is bounded for any  $s$  and  $a^{-i}$  (for both  $\pi$  and  $\tilde{\pi}$ ). Then  $Q_{soft}^{\tilde{\pi}, \rho}(s, a^i, a^{-i}) \geq Q_{soft}^{\pi, \rho}(s, a^i, a^{-i}) \forall s, a$ .

**Theorem 5.** (Opponent model improvement theorem) Given a conditional policy  $\pi$  and opponent model  $\rho$ , define a new opponent model  $\tilde{\rho}$  as

$$\tilde{\rho}(\cdot | s) \propto \exp\left(\sum_{a^i} Q_{soft}^{\pi, \rho}(s, a^i, \cdot) \pi(a^i | \cdot, s) + \alpha H(\pi(s)) + \log P(\cdot | s)\right), \forall s, a^i. \quad (41)$$

Assume that throughout our computation,  $Q$  is bounded and  $\sum_{a^{-i}} \exp(\sum_{a^i} Q(s, a^i, a^{-i}) \pi(a^i | s, a^{-i}))$  is bounded for any  $s$  and  $a^i$  (for both  $\rho$  and  $\tilde{\rho}$ ). Then  $Q_{soft}^{\pi, \tilde{\rho}}(s, a^i, a^{-i}) \geq Q_{soft}^{\pi, \rho}(s, a^i, a^{-i}) \forall s, a$ .

The proof of Theorem 4 and 5 is based on two observations that:

$$\alpha H(\pi(\cdot | s, a^{-i})) + \mathbb{E}_{a^i \sim \pi} [Q_{soft}^{\pi, \rho}(s, a^i, a^{-i})] \leq \alpha H(\tilde{\pi}(\cdot | s, a^{-i})) + \mathbb{E}_{a^i \sim \tilde{\pi}} [Q_{soft}^{\pi, \rho}(s, a^i, a^{-i})], \quad (42)$$

and

$$\mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [\alpha H(\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\rho(\cdot | s_{t+1}) || P(\cdot | s_{t+1}))] + \mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})] \quad (43)$$

$$\leq \mathbb{E}_{a_{t+1}^{-i} \sim \tilde{\rho}, a_{t+1}^i \sim \pi} [\alpha H(\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\tilde{\rho}(a_{t+1}^{-i} | s_{t+1}) || P(\cdot | s_{t+1}))] + \mathbb{E}_{a_{t+1}^{-i} \sim \tilde{\rho}, a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})]. \quad (44)$$

First, we notice that

$$\alpha H(\pi(\cdot | s, a^{-i})) + \mathbb{E}_{a^i \sim \pi} [Q_{soft}^{\pi, \rho}(s, a^i, a^{-i})] = -\alpha D_{KL}(\pi(\cdot | s, a^{-i}) || \tilde{\pi}(\cdot | s, a^{-i})) + \alpha \log \sum_{a^i} \exp\left(\frac{1}{\alpha} Q_{soft}^{\pi, \rho}(s, a^i, a^{-i})\right). \quad (45)$$

Therefore, the LHS is only maximized if the KL-Divergence on the RHS is minimized. This KL-Divergence is minimized only when  $\pi = \tilde{\pi}$ , which proves the Equation 42.

Similarly, we can have

$$\begin{aligned} & \mathbb{E}_{a^{-i} \sim \rho, a^i \sim \pi} [\alpha H(\pi(a^i | s, a^{-i})) - D_{KL}(\rho(\cdot | s) || P(\cdot | s))] + \mathbb{E}_{a^{-i} \sim \rho, a^i \sim \pi} [Q_{soft}^{\pi, \rho}(s, a^i, a^{-i})] \\ &= -D_{KL}(\rho(\cdot | s) || \tilde{\rho}(\cdot | s)) + \log \sum_{a^{-i}} \exp\left(\sum_{a^i} Q_{soft}^{\pi, \rho}(s, a^i, a^{-i}) \pi(a^i | s, a^{-i}) + \alpha H(\pi(\cdot | s, a^{-i})) + \log P(a^{-i} | s)\right), \end{aligned} \quad (46)$$

which proves the Equation 44.

With the above observations, the proof of Theorem 4 and 5 is completed by as follows:

$$\begin{aligned} & Q_{soft}^{\pi, \rho}(s_t, a_t^i, a_{t+1}^{-i}) \\ &= \mathbb{E}_{(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})} [r_t + \gamma(\alpha H(\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\rho(a_{t+1}^{-i} | s_{t+1}) || P(a_{t+1}^{-i} | s_{t+1})) + Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i}))] \end{aligned} \quad (47)$$

$$= \mathbb{E}_{(s_{t+1}, a_{t+1}^{-i})} [r_t + \gamma(\alpha H(\pi(\cdot | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\rho(a_{t+1}^{-i} | s_{t+1}) || P(a_{t+1}^{-i} | s_{t+1})) + \mathbb{E}_{a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})])] \quad (48)$$

$$\leq \mathbb{E}_{(s_{t+1}, a_{t+1}^{-i})} [r_t + \gamma(\alpha H(\tilde{\pi}(\cdot | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\rho(a_{t+1}^{-i} | s_{t+1}) || P(a_{t+1}^{-i} | s_{t+1})) + \mathbb{E}_{a_{t+1}^i \sim \tilde{\pi}} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})])] \quad (49)$$

$$\begin{aligned} &= \mathbb{E}_{(s_{t+1})} [r_t + \gamma(\mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [\alpha H(\tilde{\pi}(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i}))] - D_{KL}(\rho(\cdot | s_{t+1}) || P(\cdot | s_{t+1})) \\ &+ \mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})])] \end{aligned} \quad (50)$$

$$\begin{aligned} &\leq \mathbb{E}_{(s_{t+1})} [r_t + \gamma(\mathbb{E}_{a_{t+1}^{-i} \sim \tilde{\rho}, a_{t+1}^i \sim \pi} [\alpha H(\tilde{\pi}(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i}))] - D_{KL}(\tilde{\rho}(\cdot | s_{t+1}) || P(\cdot | s_{t+1})) \\ &+ \mathbb{E}_{a_{t+1}^{-i} \sim \tilde{\rho}, a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})])] \end{aligned} \quad (51)$$

$$\begin{aligned} &= \mathbb{E}_{(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i}) \sim \tilde{q}} [r_t + \gamma(\alpha H(\tilde{\pi}(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\tilde{\rho}(a_{t+1}^{-i} | s_{t+1}) || P(a_{t+1}^{-i} | s_{t+1})) + r_{t+1}) \\ &+ \gamma^2 \mathbb{E}_{(s_{t+2}, a_{t+2}^{-i})} [\alpha H(\pi(\cdot | s_{t+2}, a_{t+2}^{-i})) - D_{KL}(\rho(a_{t+2}^{-i} | s_{t+2}) || P(a_{t+2}^{-i} | s_{t+2})) + \mathbb{E}_{a_{t+2}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+2}, a_{t+2}^i, a_{t+2}^{-i})]]] \end{aligned} \quad (52)$$

$$\begin{aligned} &\leq \mathbb{E}_{(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})} [r_t + \gamma(\alpha H(\tilde{\pi}(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})) - D_{KL}(\tilde{\rho}(a_{t+1}^{-i} | s_{t+1}) || P(a_{t+1}^{-i} | s_{t+1})) + r_{t+1}) \\ &+ \gamma^2 \mathbb{E}_{(s_{t+2}, a_{t+2}^{-i})} [\alpha H(\pi(\cdot | s_{t+2}, a_{t+2}^{-i})) - D_{KL}(\rho(a_{t+2}^{-i} | s_{t+2}) || P(a_{t+2}^{-i} | s_{t+2})) + \mathbb{E}_{a_{t+2}^i \sim \tilde{\pi}} [Q_{soft}^{\pi, \rho}(s_{t+2}, a_{t+2}^i, a_{t+2}^{-i})]]] \end{aligned} \quad (53)$$

$\vdots$

$$\leq r_t + \mathbb{E}_{(s_{t+l}, a_{t+l}^i, a_{t+l}^{-i}, \dots) \sim \tilde{q}} [\sum_{l=1}^{\infty} \gamma^l (r_{t+l} + \alpha H(\tilde{\pi}(a_{t+l}^i | a_{t+l}^{-i}, s_{t+l})) - D_{KL}(\tilde{\rho}(a_{t+l}^{-i} | s_{t+l}) || P(a_{t+l}^{-i} | s_{t+l})))] \quad (54)$$

$$= Q_{soft}^{\tilde{\pi}, \tilde{\rho}}(s_t, a_t^i, a_t^{-i}). \quad (55)$$

With Theorem 4 and 5 and the above inequalities, we can see that, if we start from an arbitrary conditional policy  $\pi_0$  and an arbitrary opponent model  $\rho_0$  and we iterate between policy improvement as

$$\pi_{i+1}(\cdot | s, a^{-i}) \propto \exp\left(\frac{1}{\alpha} Q_{soft}^{\pi_i, \rho_i}(s, \cdot, a^{-i})\right), \quad (56)$$

and opponent model improvement as

$$\rho_{t+1}(\cdot | s) \propto \exp\left(\sum_{a^i} Q_{soft}^{\pi_{t+1}, \rho_t}(s, a^i, \cdot) \pi_{t+1}(a^i | \cdot, s) + \alpha H(\pi_{t+1}(s)) + \log P(\cdot | s)\right), \quad (57)$$

then  $Q_{soft}^{\pi_t, \rho_t}(s, a^i, a^{-i})$  can be shown to increase monotonically. Similar to [Haarnoja *et al.*, 2017], we can show that with certain regularity conditions satisfied, any non optimal policy and opponent model can be improved this way and Theorem 1 is proved.

### C.3 Soft Bellman Equation

As we show in Appendix C.2, when the training converges, we have:

$$\pi^*(a^i | s, a^{-i}) = \frac{\frac{1}{\alpha} \exp(Q^*(s, a^i, a^{-i}))}{\sum_{a^i} \exp(\frac{1}{\alpha} Q^*(s, a^i, a^{-i}))}, \quad (58)$$

and

$$\begin{aligned} \rho^*(a^{-i} | s) &= \frac{\exp(\sum_{a^i} Q^*(s, a^i, a^{-i}) \pi^*(a^i | s, a^{-i}) + \alpha H(\pi^*(a^i | s, a^{-i})) + \log P(a^{-i} | s))}{\sum_{a^{-i}} \exp(\sum_{a^i} Q^*(s, a^i, a^{-i}) \pi^*(a^i | s, a^{-i}) + \alpha H(\pi^*(a^i | s, a^{-i})) + \log P(a^{-i} | s))} \\ &= \frac{P(a^{-i} | s) \left( \sum_{a^i} \exp(Q_{soft}^*(s, a^i, a^{-i})) \right)^\alpha}{\exp(V^*(s))}, \end{aligned} \quad (59)$$

where the equality in Eq. 59 comes from substituting  $\pi^*$  with Eq. 58 and we define the soft sate value function  $V_{soft}^{\pi, \rho}(s)$  of agent  $i$  as:

$$V_{soft}^{\pi, \rho}(s_t) = \log \sum_{a_t^{-i}} P(a_t^{-i}|s_t) \left( \sum_{a_t^i} \exp \left( \frac{1}{\alpha} Q_{soft}^{\pi, \rho}(s_t, a_t^i, a_t^{-i}) \right) \right)^\alpha. \quad (60)$$

Then we can show that

$$\begin{aligned} Q_{soft}^{\pi^*, \rho^*}(s, a^i, a^{-i}) &= r_t + \gamma \mathbb{E}_{s' \sim p_s} [(\mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [\alpha H(\pi(a_{t+1}^i|s_{t+1}, a_{t+1}^{-i}))] - D_{KL}(\rho(\cdot|s_{t+1})||P(\cdot|s_{t+1})))] \\ &\quad + \mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi} [Q_{soft}^{\pi, \rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})]] \\ &= r_t + \gamma \mathbb{E}_{s' \sim p_s} [V^*(s')]. \end{aligned} \quad (61)$$

We define the soft value iteration operator  $\mathcal{T}$  as:

$$\mathcal{T}Q(s, a^i, a^{-i}) = R(s, a^i, a^{-i}) + \gamma \mathbb{E}_{s' \sim p_s} \left[ \log \sum_{a^{-i'}} P(a^{-i'}|s') \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q(s', a^{i'}, a^{-i'}) \right) \right)^\alpha \right]. \quad (62)$$

In a symmetric fully cooperative game with only one global optimum, we can show as done in [Wen *et al.*, 2019], the operator defined above is a contraction mapping. We define a norm on Q-values  $\|Q_1^i - Q_2^i\| \triangleq \max_{s, a^i, a^{-i}} |Q_1^i(s, a^i, a^{-i}) - Q_2^i(s, a^i, a^{-i})|$ . Let  $\varepsilon = \|Q_1^i - Q_2^i\|$ , then we have:

$$\begin{aligned} \log \sum_{a^{-i'}} P(a^{-i'}|s') \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q_1(s', a^{i'}, a^{-i'}) \right) \right)^\alpha &\leq \log \sum_{a^{-i'}} P(a^{-i'}|s') \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q_2(s', a^{i'}, a^{-i'}) + \varepsilon \right) \right)^\alpha \\ &= \log \sum_{a^{-i'}} P(a^{-i'}|s') \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q_2(s', a^{i'}, a^{-i'}) \right) \exp(\varepsilon) \right)^\alpha \\ &= \log \sum_{a^{-i'}} P(a^{-i'}|s') \exp(\varepsilon)^\alpha \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q_2(s', a^{i'}, a^{-i'}) \right) \right)^\alpha \\ &= \alpha \varepsilon + \log \sum_{a^{-i'}} P(a^{-i'}|s') \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q_2(s', a^{i'}, a^{-i'}) \right) \right)^\alpha. \end{aligned} \quad (63)$$

Similarly,  $\log \sum_{a^{-i'}} P(a^{-i'}|s') \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q_1(s', a^{i'}, a^{-i'}) \right) \right)^\alpha \geq -\alpha \varepsilon + \log \sum_{a^{-i'}} P(a^{-i'}|s') \left( \sum_{a^{i'}} \exp \left( \frac{1}{\alpha} Q_2(s', a^{i'}, a^{-i'}) \right) \right)^\alpha$ . Therefore  $\|\mathcal{T}Q_1^i - \mathcal{T}Q_2^i\| \leq \gamma \varepsilon = \gamma \|Q_1^i - Q_2^i\|$ , where  $\alpha = 1$ .