A Algorithms

Algorithm 1 Multi-agent Soft Q-learning

Result: policy π^i, opponent model ρ^i

Initialization:
Initialize replay buffer $\mathcal{M}$ to capacity M.
Initialize $Q_{\omega^i}(s, a^i, a^{-i})$ with random parameters ω^i, $P(a^{-i}|s)$ arbitrarily, set γ 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:

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

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

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

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

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

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

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 ω^i
Every C gradient descent steps, reset target parameters:

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

end for
end while
Compute converged π^i **and** ρ^i

Algorithm 2 Multi-agent Variational Actor Critic

Result: policy π_{θ^i}, opponent model ρ_{ϕ^i}
Initialization:
Initialize parameters θ^i, ϕ^i, ω^i, ψ^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 γ as the discount factor.
for Each episode $d = (1, \ldots, D)$ **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:
$\hat{a}_t^{-i} \leftarrow g_{\phi^{-i}}(\epsilon^{-i}; s_t)$, where $\epsilon_t^{-i} \sim \mathcal{N}$,
$a_t^i \leftarrow f_{\theta^i}(\epsilon^i; s_t, \hat{a}_t^{-i})$, 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:
$\hat{a}_{t+1}^{-i,(n)} \leftarrow g_{\phi^{-i}}(\epsilon^{-i}; s_{t+1}^{(n)})$, where $\epsilon_{t+1}^{-i} \sim \mathcal{N}$,
$a_{t+1}^{i,(n)} \leftarrow f_{\bar{\theta}^i}(\epsilon^i; s_{t+1}^{(n)}, \hat{a}_{t+1}^{-i,(n)})$, 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

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

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

$$\begin{aligned} &\nabla_{\theta^i} \mathcal{J}_\pi(\theta^i) = \nabla_{\theta^i} \alpha \log \pi_{\theta^i}(a_t^{i,(n)}|s_t^{(n)}, \hat{a}_t^{-i,(n)}) \\ &+ (\nabla_{a_t^{i,(n)}} \alpha \log \pi_{\theta^i}(a_t^{i,(n)}|s_t^{(n)}, \hat{a}_t^{-i,(n)}) - \nabla_{a_t^{i,(n)}} Q_\omega(s_t^{(n)}, a_t^{i,(n)}, \hat{a}_t^{-i,(n)})) \nabla_\theta f_{\theta^i}(\epsilon_t^i; s_t^{(n)}, \hat{a}_t^{-i,(n)}) \end{aligned}$$

$$\begin{aligned} &\nabla_{\phi^i} \mathcal{J}_\rho(\phi^i) = \nabla_{\phi^i} \log \rho_{\phi^i}(\hat{a}_t^{-i,(n)}|s_t^{(n)}) \\ &+ (\nabla_{\hat{a}_t^{-i,(n)}} \log \rho_{\phi^i}(\hat{a}_t^{-i,(n)}|s_t^{(n)}) - \nabla_{\hat{a}_t^{-i,(n)}} \log P(\hat{a}_t^{-i,(n)}|s_t^{(n)}) - \nabla_{\hat{a}_t^{-i,(n)}} Q_{\omega^i}(s_t^{(n)}, a_t^{i,(n)}, \hat{a}_t^{-i,(n)}) \\ &+ \nabla_{\hat{a}_t^{-i,(n)}} \alpha \log \pi_{\theta^i}(a^{i,(n)}|s_t^{(n)}, \hat{a}_t^{-i,(n)})) \nabla_{\phi^i} g_{\phi^i}(\epsilon_t^{-i}; s_t^{(n)}) \end{aligned}$$

Update parameters:
$\omega^i = \omega^i - \lambda_Q \nabla_{\omega^i} \mathcal{J}_Q(\omega^i)$
$\theta^i = \theta^i - \lambda_\pi \nabla_{\theta^i} \mathcal{J}_\pi(\theta^i)$
$\phi^i = \phi^i - \lambda_{\phi^i} \nabla_{\phi^i} \mathcal{J}_\rho(\phi^i)$
end for
Every C gradient descent steps, reset target parameters:

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

.
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 $P(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^{-i}_{1:T})$ as :

$$P(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^{-i}_{1:T}) = P(s_1)\prod_t P(s_{t+1}|s_t, a_t)P(a^i_t|a^{-i}_t, s_t, o^{-i}_t)P(a^{-i}_t|s_t, o^{-i}_t), \tag{29}$$

where $P(a^i_t|a^{-i}_t, s_t, o^{-i}_t)$ 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 $P(a^i_t|a^{-i}_t, s_t, o^{-i}_t) \propto 1$.

Analogously, we factorize $q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T}, o^{-i}_{1:T})$ as:

$$\begin{aligned}q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T}, o^{-i}_{1:T}) &= P(s_1)\prod_t P(s_{t+1}|s_t, a_t)q(a^i_t|a^{-i}_t, s_t, o^i_t, o^{-i}_t)q(a^{-i}_t|s_t, o^i_t, o^{-i}_t) &(30)\\ &= P(s_1)\prod_t P(s_{t+1}|s_t, a_t)\pi(a^i_t|s_t, a^{-i}_t)\rho(a^{-i}_t|s_t), &(31)\end{aligned}$$

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

With the above factorization, we have:

$$\begin{aligned}&\log P(o^i_{1:T}|o^{-i}_{1:T})\\ &= \log \sum_{a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}} P(o^i_{1:T}, a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^{-i}_{1:T}) &(32)\\ &\geq \sum q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T}, o^{-i}_{1:T}) \log \frac{P(o^i_{1:T}, a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^{-i}_{1:T})}{q(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}|o^i_{1:T}, o^{-i}_{1:T})} &(33)\\ &= \mathbb{E}_{(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}\sim q)}[\sum_{t=1}^{T} \log P(o^i_t|s_t, a^i_t, a^{-i}_t) + \cancel{\log P(s_1)} + \cancel{\sum_{t=1}^{T}\log P(s_{t+1}|s_t, a^i_t, a^{-i}_t)} &(34)\\ &\cancel{-\log P(s_1)} - \cancel{\sum_{t=1}^{T}\log P(s_{t+1}|s_t, a^i_t, a^{-i}_t)} &(35)\\ &- \sum_{t=1}^{T}\log \pi(a^i_t|s_t, a^{-i}_t) - \sum_{t=1}^{T}\log \frac{\rho(a^{-i}_t|s_t)}{P(a^{-i}_t|s_t, o^{-i}_t)} + \sum_{t=1}^{T}\log P(a^{-i}_t|s_t, a^{-i}_t, o^{-i}_t)] &(36)\\ &= \mathbb{E}_{(a^i_{1:T}, a^{-i}_{1:T}, s_{1:T}\sim q)}[\sum_{t=1}^{T} R^i(s_t, a^i_t, a^{-i}_t) - \log \pi(a^i_t|s_t, a^{-i}_t) - \log \frac{\rho(a^{-i}_t|s_t)}{P(a^{-i}_t|s_t, o^{-i}_t)} + 1] &(37)\\ &= \sum_t \mathbb{E}_{(s_t, a^i_t, a^{-i}_t)\sim q}[R^i(s_t, a^i_t, a^{-i}_t) + H(\pi(a^i_t|s_t, a^{-i}_t)) - D_{\mathrm{KL}}(\rho(a^{-i}_t|s_t)||P(a^{-i}_t|s_t, o^{-i}_t))], &(38)\end{aligned}$$

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

$$P(o^i|s, a^i, a^{-i}, o^{-i}) = P(o^i|s, a^i, a^{-i}). \tag{39}$$

B.2 The Lower Bound on Opponent model

From Eq. 11, we have:

$$\begin{aligned}\log P(\rho = \pi^{-i*}|s) &= \log P(B^i(\rho) = \pi^{i*})\\ &= \log \mathbb{E}_{a^{-i}\sim\rho}[P(B^i(a^{-i}) = \pi^{i*}|s, a^{-i})]\\ &= \log \mathbb{E}_{a^i\sim B^i(a^{-i}), a^{-i}\sim\rho}[P(a^i \sim \pi^{i*}|s, a^i, a^{-i})]\\ &= \log \mathbb{E}_{a^i\sim B^i(a^{-i}), a^{-i}\sim\rho}[P(o^i|s, a^i, a^{-i})]\\ &\geq \mathbb{E}_{a^i\sim\pi^i(a^{-i}), a^{-i}\sim\rho}[\log P(o^i|s, a^i, a^{-i}) + H(\pi(a^i|s, a^{-i}))], &(40)\end{aligned}$$

where we set the best response function to an arbitrary opponent policies as: $B^i(\rho) \propto 1$.

C Multi-Agent Soft-Q Learning

C.1 Soft Q-Function

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

$$Q^{\pi,\rho}_{soft}(s_t, a^i_t, a^{-i}_t)$$

$$= r_t + \mathbb{E}_{(s_{t+l}, a^i_{t+l}, a^{-i}_{t+l}, \ldots) \sim q}[\sum_{l=1}^{\infty} \gamma^l (r_{t+l} + \alpha H(\pi(a^i_{t+l}|a^{-i}_{t+l}, s_{t+l})) - D_{\mathrm{KL}}(\rho(a^{-i}_{t+l}|s_{t+l})||P(a^{-i}_{t+l}|s_{t+l}))] \tag{41}$$

$$= \mathbb{E}_{(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})}[r_t + \gamma(\alpha H(\pi(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1})) - D_{\mathrm{KL}}(\rho(a^{-i}|s_{t+1})||P(a^{-i}|s_{t+1})) + Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1}))] \tag{42}$$

$$= \mathbb{E}_{(s_{t+1}, a^{-i}_{t+1})}[r_t + \gamma(\alpha H(\pi(\cdot|s_{t+1}, a^{-i}_{t+1})) - D_{\mathrm{KL}}(\rho(a^{-i}|s_{t+1})||P(a^{-i}|s_{t+1})) + \mathbb{E}_{a^i_{t+1} \sim \pi}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})])] \tag{43}$$

$$= \mathbb{E}_{(s_{t+1})}[r_t + \gamma(\mathbb{E}_{a^{-i}_{t+1} \sim \rho, a^i_{t+1} \sim \pi}[\alpha H(\pi(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1}))] - D_{\mathrm{KL}}(\rho(\cdot|s_{t+1})||P(\cdot|s_{t+1}))]$$
$$+ \mathbb{E}_{a^{-i}_{t+1} \sim \rho, a^i_{t+1} \sim \pi}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})])], \tag{44}$$

Then we can easily see that the objective in Eq. 9 can be rewritten as:

$$\mathcal{J}(\pi, \phi) = \sum_t \mathbb{E}_{(s_t, a^i_t, a^{-i}_t) \sim (p_s, \pi, \rho)}[Q^{\pi,\rho}_{soft}(s_t, a^i_t, a^{-i}_t) + \alpha H(\pi(a^i_t|s_t, a^{-i}_t)) - D_{\mathrm{KL}}(\rho(a^{-i}_t|s_t)||P(a^{-i}_t|s_t))], \tag{45}$$

by setting $\alpha = 1$.

C.2 Policy Improvement and Opponent Model Improvement

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

$$\tilde{\pi}(\cdot|s, a^{-i}) \propto \exp(\frac{1}{\alpha} Q^{\pi,\rho}_{soft}(s, \cdot, a^{-i})), \forall s, a^{-i}. \tag{46}$$

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 π and $\tilde{\pi}$). Then $Q^{\tilde{\pi},\rho}_{soft}(s, a^i, a^{-i}) \geq Q^{\pi,\rho}_{soft}(s, a^i, a^{-i}) \forall s, a$.

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

$$\tilde{\rho}(\cdot|s) \propto \exp(\sum_{a^i} Q^{\pi,\rho}_{soft}(s, a^i, \cdot)\pi(a^i|\cdot, s) + \alpha H(\pi(s)) + \log P(\cdot|s)), \forall s, a^i. \tag{47}$$

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 ρ and $\tilde{\rho}$). Then $Q^{\pi,\tilde{\rho}}_{soft}(s, a^i, a^{-i}) \geq Q^{\pi,\rho}_{soft}(s, a^i, a^{-i}) \forall s, a$.

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

$$\alpha H(\pi(\cdot|s, a^{-i})) + \mathbb{E}_{a^i \sim \pi}[Q^{\pi,\rho}_{soft}(s, a^i, a^{-i})] \leq \alpha H(\tilde{\pi}(\cdot|s, a^{-i})) + \mathbb{E}_{a^i \sim \tilde{\pi}}[Q^{\pi,\rho}_{soft}(s, a^i, a^{-i})], \tag{48}$$

and

$$\mathbb{E}_{a^{-i}_{t+1} \sim \rho, a^i_{t+1} \sim \pi}[\alpha H(\pi(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1}))] - D_{\mathrm{KL}}(\rho(\cdot|s_{t+1})||P(\cdot|s_{t+1}))] + \mathbb{E}_{a^{-i}_{t+1} \sim \rho, a^i_{t+1} \sim \pi}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})] \tag{49}$$

$$\leq \mathbb{E}_{a^{-i}_{t+1} \sim \tilde{\rho}, a^i_{t+1} \sim \pi}[\alpha H(\pi(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1}))] - D_{\mathrm{KL}}(\tilde{\rho}(a^{-i}_{t+1}|s_{t+1})||P(\cdot|s_{t+1})) + \mathbb{E}_{a^{-i}_{t+1} \sim \tilde{\rho}, a^i_{t+1} \sim \pi}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})]. \tag{50}$$

First, we notice that

$$\alpha H(\pi(\cdot|s, a^{-i})) + \mathbb{E}_{a^i \sim \pi}[Q^{\pi,\rho}_{soft}(s, a^i, a^{-i})] = -\alpha D_{\mathrm{KL}}(\pi(\cdot|s, a^{-i})||\tilde{\pi}(\cdot|s, a^{-i})) + \alpha \log \sum_{a^i} \exp(\frac{1}{\alpha} Q^{\pi,\rho}_{soft}(s, a^i, a^{-i})). \tag{51}$$

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 48.

Similarly, we can have

$$\mathbb{E}_{a^{-i} \sim \rho, a^i \sim \pi}[\alpha H(\pi(a^i|s, a^{-i}))] - D_{\mathrm{KL}}(\rho(\cdot|s)||P(\cdot|s))]) + \mathbb{E}_{a^{-i} \sim \rho, a^i \sim \pi}[Q^{\pi,\rho}_{soft}(s, a^i, a^{-i})]$$
$$= -D_{\mathrm{KL}}(\rho(\cdot|s)||\tilde{\rho}(\cdot|s)) + \log \sum_{a^{-i}} \exp(\sum_{a^i} Q^{\pi,\rho}(s, a^i, a^{-i})\pi(a^i|s, a^{-i}) + \alpha H(\pi(\cdot|s, a^i)) + \log P(a^{-i|s})), \tag{52}$$

which proves the Equation 50.

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

$$Q^{\pi,\rho}_{soft}(s_t, a^i_t, a^{-i}_t)$$
$$= \mathbb{E}_{(s_{t+1},a^i_{t+1},a^{-i}_{t+1})}[r_t + \gamma(\alpha H(\pi(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1})) - D_{\mathrm{KL}}(\rho(a^{-i}_{t+1}|s_{t+1})||P(a^{-i}_{t+1}|s_{t+1})) + Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1}))] \tag{53}$$
$$= \mathbb{E}_{(s_{t+1},a^{-i}_{t+1})}[r_t + \gamma(\alpha H(\pi(\cdot|s_{t+1}, a^{-i}_{t+1})) - D_{\mathrm{KL}}(\rho(a^{-i}_{t+1}|s_{t+1})||P(a^{-i}_{t+1}|s_{t+1})) + \mathbb{E}_{a^i_{t+1}\sim\pi}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})])] \tag{54}$$
$$\leq \mathbb{E}_{(s_{t+1},a^{-i}_{t+1})}[r_t + \gamma(\alpha H(\tilde{\pi}(\cdot|s_{t+1}, a^{-i}_{t+1})) - D_{\mathrm{KL}}(\rho(a^{-i}_{t+1}|s_{t+1})||P(a^{-i}_{t+1}|s_{t+1})) + \mathbb{E}_{a^i_{t+1}\sim\tilde{\pi}}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})])] \tag{55}$$
$$= \mathbb{E}_{(s_{t+1})}[r_t + \gamma(\mathbb{E}_{a^{-i}_{t+1}\sim\rho,a^i_{t+1}\sim\pi}[\alpha H(\tilde{\pi}(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1}))] - D_{\mathrm{KL}}(\rho(\cdot|s_{t+1})||P(\cdot|s_{t+1}))$$
$$+ \mathbb{E}_{a^{-i}_{t+1}\sim\rho,a^i_{t+1}\sim\pi}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})])] \tag{56}$$
$$\leq \mathbb{E}_{(s_{t+1})}[r_t + \gamma(\mathbb{E}_{a^{-i}_{t+1}\sim\tilde{\rho},a^i_{t+1}\sim\pi}[\alpha H(\tilde{\pi}(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1}))] - D_{\mathrm{KL}}(\tilde{\rho}(\cdot|s_{t+1})||P(\cdot|s_{t+1}))$$
$$+ \mathbb{E}_{a^{-i}_{t+1}\sim\tilde{\rho},a^i_{t+1}\sim\pi}[Q^{\pi,\rho}_{soft}(s_{t+1}, a^i_{t+1}, a^{-i}_{t+1})])] \tag{57}$$
$$= \mathbb{E}_{(s_{t+1},a^i_{t+1},a^{-i}_{t+1})\sim\tilde{q}}[r_t + \gamma(\alpha H(\tilde{\pi}(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1})) - D_{\mathrm{KL}}(\tilde{\rho}(a^{-i}|s_{t+1})||P(a^{-i}|s_{t+1})) + r_{t+1})$$
$$+ \gamma^2\mathbb{E}_{(s_{t+2},a^{-i}_{t+2})}[\alpha H(\pi(\cdot|s_{t+2}, a^{-i}_{t+2})) - D_{\mathrm{KL}}(\rho(a^{-i}_{t+2}|s_{t+2})||P(a^{-i}_{t+2}|s_{t+2})) + \mathbb{E}_{a^i_{t+2}\sim\pi}[Q^{\pi,\rho}_{soft}(s_{t+2}, a^i_{t+2}, a^{-i}_{t+2})]]] \tag{58}$$
$$\leq \mathbb{E}_{(s_{t+1},a^i_{t+1},a^{-i}_{t+1})}[r_t + \gamma(\alpha H(\tilde{\pi}(a^i_{t+1}|s_{t+1}, a^{-i}_{t+1})) - D_{\mathrm{KL}}(\tilde{\rho}(a^{-i}|s_{t+1})||P(a^{-i}|s_{t+1})) + r_{t+1})$$
$$+ \gamma^2\mathbb{E}_{(s_{t+2},a^{-i}_{t+2})}[\alpha H(\pi(\cdot|s_{t+2}, a^{-i}_{t+2})) - D_{\mathrm{KL}}(\rho(a^{-i}_{t+2}|s_{t+2})||P(a^{-i}_{t+2}|s_{t+2})) + \mathbb{E}_{a^i_{t+2}\sim\tilde{\pi}}[Q^{\pi,\rho}_{soft}(s_{t+2}, a^i_{t+2}, a^{-i}_{t+2})]]] \tag{59}$$
$$\vdots$$
$$\leq r_t + \mathbb{E}_{(s_{t+l},a^i_{t+l},a^{-i}_{t+l},\ldots)\sim\tilde{q}}[\sum_{l=1}^{\infty}\gamma^l(r_{t+l} + \alpha H(\tilde{\pi}(a^i_{t+l}|a^{-i}_{t+l}, s_{t+l})) - D_{\mathrm{KL}}(\tilde{\rho}(a^{-i}_{t+l}|s_{t+l})||P(a^{-i}_{t+l}|s_{t+l}))] \tag{60}$$
$$= Q^{\tilde{\pi},\tilde{\rho}}_{soft}(s_t, a^i_t, a^{-i}_t). \tag{61}$$

With Theorem 4 and 5 and the above inequalities, we can see that, if we start from an arbitrary conditional policy π_0 and an arbitrary opponent model ρ_0 and we iterate between policy improvement as

$$\pi_{i+1}(\cdot|s, a^{-i}) \propto \exp(\frac{1}{\alpha}Q^{\pi_t,\rho_t}_{soft}(s, \cdot, a^{-i})), \tag{62}$$

and opponent model improvement as

$$\rho_{t+1}(\cdot|s) \propto \exp(\sum_{a^i} Q^{\pi_{t+1},\rho_t}_{soft}(s, a^i, \cdot)\pi_{t+1}(a^i|\cdot, s) + \alpha H(\pi_{t+1}(s)) + \log P(\cdot|s)), \tag{63}$$

then $Q^{\pi_t,\rho_t}_{soft}(s, a^i, a^{-i})$ 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:

$$\pi^*(a^i|s, a^{-i}) = \frac{\frac{1}{\alpha}\exp(Q^*(s, a^i, a^{-i}))}{\sum_{a^i}\exp(\frac{1}{\alpha}Q^*(s, a^i, a^{-i}))}, \tag{64}$$

and

$$\begin{aligned}\rho^*(a^{-i}|s) &= \frac{\exp(\sum_{a^i} Q^*(s, a^i, a^{-i})\pi^*(a^i|s, a^{-i}) + \alpha H(\pi^*(a^i|s, a^{-i})) + \log P(a^{-i}|s))}{\sum_{a^{-i}}\exp(\sum_{a^i} Q^*(s, a^i, a^{-i})\pi^*(a^i|s, a^{-i}) + \alpha H(\pi^*(a^i|s, a^{-i})) + \log P(a^{-i}|s))} \\ &= \frac{P(a^{-i}|s)\left(\sum_{a^i}\exp(Q^*_{soft}(s, a^i, a^{-i}))\right)^{\alpha}}{\exp(V^*(s))},\end{aligned} \tag{65}$$

where the equality in Eq. 65 comes from substituting π^* with Eq. 64 and we define the soft sate value function $V_{soft}^{\pi,\rho}(s)$ of agent i as:

$$V_{soft}^{\pi,\rho}(s_t) = \log \sum_{a_t^{-i}} P(a_t^{-i}|s_t) \left(\sum_{a_t^i} \exp \left(\frac{1}{\alpha} Q_{soft}^{\pi,\rho}(s_t, a_t^i, a_t^{-i}) \right) \right)^\alpha . \tag{66}$$

Then we can show that

$$\begin{aligned}
&Q_{soft}^{\pi^*,\rho^*}(s, a^i, a^{-i}) \\
&= r_t + \gamma \mathbb{E}_{s' \sim p_s}[(\mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi}[\alpha H(\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i}))] - D_{\mathrm{KL}}(\rho(\cdot|s_{t+1})||P(\cdot|s_{t+1}))] \\
&+ \mathbb{E}_{a_{t+1}^{-i} \sim \rho, a_{t+1}^i \sim \pi}[Q_{soft}^{\pi,\rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})])] \\
&= r_t + \gamma \mathbb{E}_{s' \sim p_s}[V^*(s')].
\end{aligned} \tag{67}$$

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

$$\mathcal{T} Q(s, a^i, a^{-i}) = R(s, a^i, a^{-i}) + \gamma \mathbb{E}_{s' \sim p_s} \left[\log \sum_{a^{-i\prime}} P(a^{-i\prime}|s') \left(\sum_{a^{i\prime}} \exp \left(\frac{1}{\alpha} Q(s', a^{i\prime}, a^{-i\prime}) \right) \right)^\alpha \right]. \tag{68}$$

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 $\left\| Q_1^i - Q_2^i \right\| \triangleq \max_{s,a^i,a^{-i}} \left| Q_1^i\left(s, a^i, a^{-i}\right) - Q_2^i\left(s, a^i, a^{-i}\right) \right|$. Let $\varepsilon = \left\| Q_1^i - Q_2^i \right\|$, then we have:

$$\begin{aligned}
\log \sum_{a^{-i\prime}} P(a^{-i\prime}|s') \left(\sum_{a^{i\prime}} \exp \left(\frac{1}{\alpha} Q_1(s', a^{i\prime}, a^{-i\prime}) \right) \right)^\alpha &\leq \log \sum_{a^{-i\prime}} P(a^{-i\prime}|s') \left(\sum_{a^{i\prime}} \exp \left(\frac{1}{\alpha} Q_2(s', a^{i\prime}, a^{-i\prime}) + \varepsilon \right) \right)^\alpha \\
&= \log \sum_{a^{-i\prime}} P(a^{-i\prime}|s') \left(\sum_{a^{i\prime}} \exp \left(\frac{1}{\alpha} Q_2(s', a^{i\prime}, a^{-i\prime}) \right) \exp(\varepsilon) \right)^\alpha \\
&= \log \sum_{a^{-i\prime}} P(a^{-i\prime}|s') \exp(\varepsilon)^\alpha \left(\sum_{a^{i\prime}} \exp \left(\frac{1}{\alpha} Q_2(s', a^{i\prime}, a^{-i\prime}) \right) \right)^\alpha \\
&= \alpha\varepsilon + \log \sum_{a^{-i\prime}} P(a^{-i\prime}|s') \left(\sum_{a^{i\prime}} \exp \left(\frac{1}{\alpha} Q_2(s', a^{i\prime}, a^{-i\prime}) \right) \right)^\alpha .
\end{aligned} \tag{69}$$

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