Report on Final Project

Phu Sakulwongtana

November 2019

Abstract

In this report, I will go over the main content of the final project. Starting from the most plausible one and then move toward more interesting projects ideas. The main idea is to try, firstly, solve the problem of inconsistency between ROMMEO and Balancing-Q. After solving this "main challenge" we will move on to other applications. This serves as the brainstorm area. Finally, on each section, we will also include the background on the contexts and inspirations. We will minimally cite the works and save the citation space for the more obscure papers.

1 Solving the inconsistencies

We will start by showing how to marry ROMMEO (which only works on cooperative game, and based on systematic probabilistic derivation) and Balacing-Q learning (which works on both types of games – competitive and cooperation – however lacking the derivation, since the loss is created out of no-where)

1.1 Deriving the algorithms

Before we start let's review how each algorithm are derived, while for the theoretical studies, we will refer to the next section. In the big picture, ROMMEO uses variational inferences, while Balancing-Q are based on constrained optimization that turns out to have very interesting properties.

1.1.1 ROMMEO

ROMMEO assumes that the joint probability between 2 agents (agent and its opponents, which can be further factorized) as

$$\pi(a, a^{-i}|s) = \pi(a|a^{-i}, s)\rho(a^{-i}|s) \tag{1}$$

factorizing in other ways yields PR2. We will start by drawing the graphical model of joint probabilities that we want to approximate is depicted in Figure 1. Given this, we can show that the prior joint probability is equal to the following

$$\begin{aligned} &P(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}, \mathcal{O}^i_{1:T} = 1|\mathcal{O}^{-i}_{1:T} = 1) \\ &\quad = P(s_0)\prod_{t=0}^{T} P_{\text{prior}}(a^i_t|s_t, a^{-i}_t)P_{\text{prior}}(a^{-i}_t|s_t)P(s_{t+1}|s_t, a^i_t, a^{-i}_t)P(\mathcal{O}^i_t = 1|s_t, a^i_t, a^{-i}_t, \mathcal{O}^{-i}_t = 1) \end{aligned} \tag{2}$$

where the optimality variable is defined as

$$P(\mathcal{O}^i_t = 1|s_t, a^i_t, a^{-i}_t, \mathcal{O}^{-i}_t = 1) = \exp\left(\beta R(s_t, a^i_t, a^{-i}_t)\right) \tag{3}$$

Since this is the cooperative setting, the given the fact that the optimal opponent model, the agent should simply increase its reward. *I believe that it is abit problematic although this make sense.* Furthermore, we call β the temperature variable, in which it will play a crucial role and be the main source of the inconsistency. Now for the variations joint probabilities that we want to optimize on the graphical model is depicted in Figure 2. The joint probability of variational distribution can be calculated as following

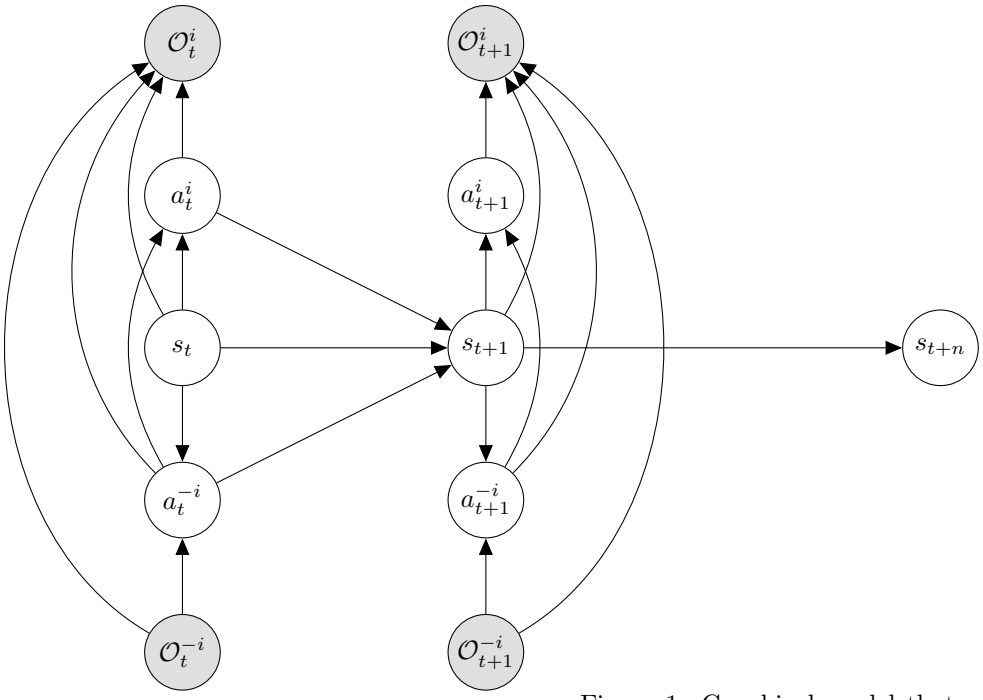

Figure 1: Graphical model that we want to approximate. This is based on the joint probability provided in ROMMEO paper. The author assume that the opponent we are playing against is optimal

$$q(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}) = P(s_0) \prod_{t=0}^{T} \pi_\theta(a^i_t|s_t, a^{-i}_t) \rho_\phi(a^{-i}_t|s_t) P(s_{t+1}|s_t, a^i_t, a^{-i}_t) \tag{4}$$

For now, all we have to do is to perform variational approximation, so that we can optimize both opponent model and agent's model *together* (this can be problematic as we will show in the next section). We would like to solve the following optimization problem

$$\arg\min_{\pi,\phi \in \Pi \times \Phi} D_{\mathrm{KL}} \left(q(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}) \middle\| P(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T} | \mathcal{O}^i_{1:T} = 1 \mathcal{O}^{-i}_{1:T} = 1) \right) \tag{5}$$

This lead to optimizing the following ELBO. The derivation is shown in the appendix A.1.1, which translates the problem into a maximization problem of the following

$$\arg\max_{\pi,\phi \in \Pi \times \Phi} \mathbb{E}_q \left[\sum_{t=0}^{T} \gamma^t \left(\beta R(s_t, a^i_t, a^{-i}_t) - \log \frac{\pi_\theta(a^i_t|s_t, a^{-i}_t)}{P_{\text{prior}}(a^i_t|s_t, a^{-i}_t)} - \log \frac{\rho_\phi(a^{-i}_t|s_t)}{P_{\text{prior}}(a^{-i}_t|s_t)} \right) \right] \tag{6}$$

We can see that it is almost equal to normal reinforcement learning problem with regularization (don't forget that we also train the opponent model together with the agent) toward the common reward. As we try to optimize, we want the agent to be close to the designated prior too. Now, we will try to solve this optimization problem in closed form. We will starting off by consider the last time step, as we will progress backward in time. We will starting by consider the last time step, which is equivalent to

$$\mathbb{E}_q \left[\beta R(s_T, a^i_T, a^{-i}_T) - \log \frac{\pi_\theta(a^i_T|s_T, a^{-i}_T)}{P_{\text{prior}}(a^i_T|s_T, a^{-i}_T)} - \log \frac{\rho_\phi(a^{-i}_T|s_T)}{P_{\text{prior}}(a^{-i}_T|s_T)} \right] \tag{7}$$

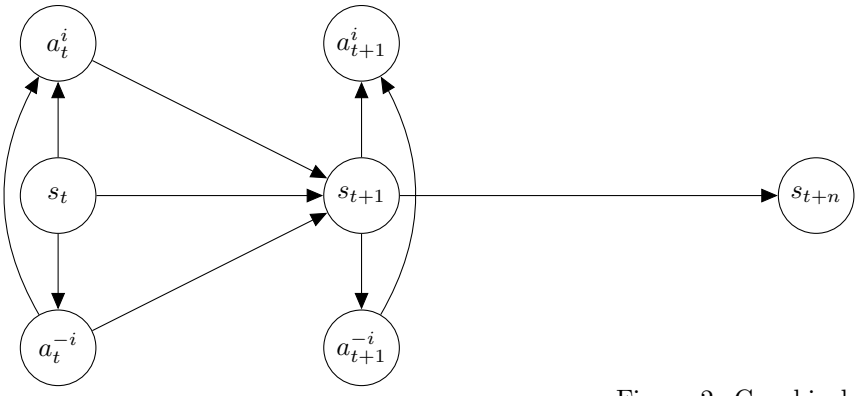

Figure 2: Graphical model that we are going to optimize our policies on. This has almost the same structure as the version that we want to approximate. However, we doesn't care about the optimality of the agent itself, as we want to approximate its policy (denote as π) and the its opponent model (denote as ρ)

We can show that the optimal policy is equal to

$$\begin{aligned} \pi_\theta(a_T^i|s_T, a_T^{-i}) &= \frac{\exp\left(\beta R(s_T, a_T^i, a_T^{-i})\right) P_{\text{prior}}(a_T^i|s_T, a_T^{-i})}{\exp\left(Q^*(s_T, a_T^{-i})\right)} \\ \text{where } Q^*(s_T, a_T^{-i}) &= \log \int \exp\left(\beta R(s_T, a_T^i, a_T^{-i}\right) P_{\text{prior}}(a_T^i|s_T, a_T^{-i})\ \mathrm{d}a_T^i \end{aligned} \tag{8}$$

The proof will be presented in the appendix A.1.2. Since we now have the optimal agent's policy, we can find the optimal opponent model's policy by plugging agent's policy back into the equation 7, with the similar process (see appendix A.1.3) for more details) we have the following optimal opponent's model:

$$\begin{aligned} \rho_\phi(a_T^{-i}|s_T) &= \frac{\exp\left(Q^*(s_T, a_T^{-i})\right) P_{\text{prior}}(a_T^{-i}|s_T)}{\exp\left(V^*(s_T)\right)} \\ \text{where } V^*(s_T) &= \log \int \exp\left(Q^*(s_T, a_T^{-i})\right) P_{\text{prior}}(a_T^{-i}|s_T)\ \mathrm{d}a_T^{-i} \end{aligned} \tag{9}$$

Note that, instead of directly plug into what we have left from the derivation of optimal policy, we plug the optimal agent policy to the original objective, as this will be what true opponent model reacts to. This step will be important since we allows to decouple the training process, as this will be useful in our solution for fixing the inconsistency. Furthermore, the the differences between our formulation and the original isn't signification as the authors decides to explicitly move the weighting parameter outward, while setting the objective to be as:

$$\mathbb{E}_q\left[R(s_T, a_T^i, a_T^{-i}) - \frac{1}{\beta}\log \underbrace{\frac{\pi_\theta(a_T^i|s_T, a_T^{-i})}{\mathcal{U}(a_T^i)}}_{(1)} - \log \frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right] \tag{10}$$

where $\mathcal{U}(a_T^i)$ is uniform distribution, in which we can reduce part (1) to be entropy maximization. This objective slightly violates the implementation because we can't freely setting the weighting on each log (ironically, this is why ROMMEO is incompatible with balancing-Q in the first place). Before, we move on to the arbitrary time-step t. We want to consider what quantity we get from the time step T that will be passed to time-step before. From the derivation in A.1.4 by plugging the policy and opponent model to the objective, we can see that we are left with the

$$\gamma \mathbb{E}_{P(s_T)}\left[V(s_T)\right] \tag{11}$$

This quantity will be passed down to the later time. Now for time t, the objective that we are optimizing becomes

$$\mathbb{E}_q\Bigg[\underbrace{\beta R(s_t, a_T^i, a_T^{-i}) + \gamma \mathbb{E}_{P(s_{t+1}|s_t,a_t^i,a_t^{-i})}\left[V(s_{t+1})\right]}_{Q^*(s_t,a_t^i,a_t^{-i})} - \log\frac{\pi_\theta(a_T^i|s_t,a_T^{-i})}{P_{\text{prior}}(a_T^i|s_t,a_T^{-i})} - \log\frac{\rho_\phi(a_T^{-i}|s_t)}{P_{\text{prior}}(a_T^{-i}|s_t)}\Bigg] \tag{12}$$

We can see that this leads to very similar problems as the one we have before, therefore, by using the same proving process, we can derive the optimal agent's policy and optimal opponent model policy to be

$$\pi_\theta^*(a_t^i|s_t,a_t^{-i}) = \frac{\exp\left(Q^*(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^i|s_t,a_t^{-i})}{\exp\left(Q^*(s_t,a_t^{-i})\right)} \qquad \rho_\phi^*(a_t^{-i}|s_t) = \frac{\exp\left(Q^*(s_t,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)}{\exp\left(V^*(s_t)\right)} \tag{13}$$

where the analogous "Bellman equation" is the following (with the value and action value functions being)

$$\begin{aligned} Q^*(s_t,a_t^i,a_t^{-i}) &= \beta R(s_t,a_T^i,a_T^{-i}) + \gamma\mathbb{E}_{P(s_{t+1}|s_t,a_t^i,a_t^{-i})}\left[V^*(s_{t+1})\right] \\ \text{where } Q^*(s_t,a_t^{-i}) &= \log\int \exp\left(Q(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^i|s_t,a_t^{-i})\ \mathrm{d}a_t^i \\ V^*(s_t) &= \log\int \exp\left(Q(s_t,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a_t^{-i} \end{aligned} \tag{14}$$

This finishes the derivation for ROMMEO in its most general form. As we will move to Balancing-Q.

1.1.2 Balancing-Q

Balancing-Q is the multi-agent extension to the well-established bounded rationality reinforcement framework, in which the objective for the agent to optimize is defined as :

$$\begin{aligned} \pi^* &= \arg\max_{\pi\in\Pi}\ \mathbb{E}\left[\sum_{t=0}^{T}\gamma^t R(s_t,a_t)\right] \\ &\text{Such That } D_{\text{KL}}\left(\pi(a_t|s_t)\middle\|P_{\text{prior}}(a_t|s_t)\right) \le C \end{aligned} \tag{15}$$

Given the this, we can transform it to unconstaint optimization as

$$\arg\max_{\pi\in\Pi}\ \mathbb{E}\left[\sum_{t=0}^{T}\gamma^t\left(R(s_t,a_t) - \beta D_{\text{KL}}\left(\pi(a_t|s_t)\middle\|P_{\text{prior}}(a_t|s_t)\right)\right)\right] \tag{16}$$

We can see that bounded rationality can be seen another interpretation to control-as-inference framework. The authors of Balancing-Q proposed the following optimization objective for both agents (no opponent model and it is semi-centralized training as we will see in the derivation of optimal agent policies):

$$\mathbb{E}\left[\sum_{t=0}^{T}\gamma\left(R(s_t,a_t,a_t^{-i}) - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_t|s_t)}{P_{\text{prior}}(a_t^i|s_t)} - \frac{1}{\beta^{-i}}\log\frac{\rho_\phi(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)}\right)\right] \tag{17}$$

As discussed in [18], Balancing-Q joint policy can be viewed as the following factorization

$$\pi(a,a^{-i}|s) = \pi(a|s)\rho(a^{-i}|s) \tag{18}$$

However, there is no probability graphical model that can represent this. Recall that Balancing-Q learning allows agents to learn both in minimax game and cooperative game based on the temperature parameter β^i and β^{-i}. Normally, β denotes how close we want our policy to be to the prior, the higher the closer, which means the value should be positive. In this case, we can also use the temperature variable to control the behavior of the agents by control how each agent "perceives" the joint reward if the temperature varaibles are in opposite sign, then the game becomes minimax (one want to maximize it other want to minimize it)

and vice versa. Now let's fully derive the optimal policy and optimal opponent model from this objective. Using the same method as ROMMEO, we will start with the last time step T, then derive the message, and find the objective for general time step t. The objective at time step T is

$$\mathbb{E}_{P(s_T)P(a_T,a_T^{-i}|s)}\left[R(s_T,a_T,a_T^{-i}) - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_T|s_T)}{P_{\text{prior}}(a_T^i|s_T)} - \frac{1}{\beta^{-i}}\log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right] \tag{19}$$

The optimal agent's policy is then equal to (see proof in appendix A.2.1)

$$\begin{aligned}
\pi_\theta(a_T^i|s_T) &= \frac{\exp\left(Q^i(s_T,a_T^i)\right)P_{\text{prior}}(a_T^i|s_T)}{\exp\left(V^i(s_T)\right)} \\
\text{where } Q^i(s_T,a_T^i) &= \frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}R(s_T,a_T^i,a_T^{-i})\right)P_{\text{prior}}(a_T^{-i}|s_T)\ \mathrm{d}a^{-i} \\
\text{and } V^i(s_T) &= \log\int\exp\left(Q^i(s_T,a_T^i)\right)P_{\text{prior}}(a_T^i|s_T)\ \mathrm{d}a^i
\end{aligned} \tag{20}$$

Similarly the the optimal opponent policy is (the proof is almost the same)

$$\begin{aligned}
\rho_\phi(a_T^{-i}|s_T) &= \frac{\exp\left(Q^{-i}(s_T,a_T^{-i})\right)P_{\text{prior}}(a_T^{-i}|s_T)}{\exp\left(V^{-i}(s_T)\right)} \\
\text{where } Q^{-i}(s_T,a_T^{-i}) &= \frac{\beta^{-i}}{\beta^i}\log\int\exp\left(\beta^{i}R(s_T,a_T^i,a_T^{-i})\right)P_{\text{prior}}(a_T^{i}|s_T)\ \mathrm{d}a^{i} \\
\text{and } V^{-i}(s_T) &= \log\int\exp\left(Q^{-i}(s_T,a_T^{-i})\right)P_{\text{prior}}(a_T^{-i}|s_T)\ \mathrm{d}a^{-i}
\end{aligned} \tag{21}$$

Now, as we want to move on to the next time-step, we will have to consider the message passed by the time step before. Interestingly, the agent's backward message (LHS), for agent, is calculated by plugging back the "conditional" opponent's policy and the optimal policy (we infer from the result where the full calculate is shown in appendix), A.2.2) Interesting, the backward message for opponent (RHS) is shown to be equal to the agents (We will collectively call it $V(s)$)

$$\gamma\mathbb{E}_{P(s_t)}\left[V(s)\right] = \gamma\mathbb{E}_{P(s_t)}\left[\frac{1}{\beta^i}V^i(s_t)\right] = \gamma\mathbb{E}_{P(s_t)}\left[\frac{1}{\beta^{-i}}V^{-i}(s_t)\right] \tag{22}$$

Now, we can consider the objective on arbitrary time step t as

$$\mathbb{E}_{P(s_t,a_t,a_t^{-i})}\left[\underbrace{R(s_t,a_t,a_t^{-i}) + \gamma\mathbb{E}_{P(s_{t+1}|s_t,a_t,a_t^{-i})}\left[V(s)\right]}_{Q(s_t,a_t^i,a_t^{-i})} - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_t|s_t)}{P_{\text{prior}}(a_t^i|s_t)} - \frac{1}{\beta^{-i}}\log\frac{\rho_\phi(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)}\right] \tag{23}$$

Given the same method, we can derive the final policy for agent and its opponent as

$$\pi_\theta^*(a_t^i|s_t) = \frac{\exp\left(Q^i(s_t,a_t^i)\right)P_{\text{prior}}(a_t^i|s_t)}{\exp\left(V^i(s_t)\right)} \qquad \rho_\phi^*(a_t^{-i}|s_t) = \frac{\exp\left(Q^{-i}(s_t,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)}{\exp\left(V^{-i}(s_t)\right)} \tag{24}$$

Where the "Bellman" equation are listed as

$$\begin{aligned}
Q^*(s_t,a_t^i,a_t^{-i}) &= R(s_t,a_t,a_t^{-i}) + \mathbb{E}_{P(s_{t+1}|s_t,a_t,a_t^{-i})}\left[V^*(s)\right] \\
\text{where } Q^{i*}(s_t,a_t^i) &= \frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i} \\
V^{i*}(s_t) &= \log\int\exp\left(Q^i(s_t,a_t^i)\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i \\
Q^{-i*}(s_t,a_t^{-i}) &= \frac{\beta^{-i}}{\beta^i}\log\int\exp\left(\beta^{i}Q(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{i}|s_t)\ \mathrm{d}a^{i} \\
V^{-i*}(s_t) &= \log\int\exp\left(Q^{-i}(s_t,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i} \\
V^*(s_t) &= \frac{1}{\beta^i}V^i(s_t) = \frac{1}{\beta^{-i}}V^{-i}(s_t)
\end{aligned} \tag{25}$$

Thus finish the derivation of Balancing-Q policy. We will move on to the theoretical properties of both algorithms, while leaving the observation on the results and implementation details to the sections afterward.

1.2 Theoretical Properties

Now we shall consider theoretical properties for both algorithms, which is mostly guarantee the results from the algorithm will be useful, and the algorithm itself will converges to optimal value.

1.2.1 ROMMEO

We will starting with the contraction mapping of the Bellman-like equation (14), and therefore, we can show that the action value function will converge to optimal value. Defining the Bellman-Operator as

$$\begin{aligned} \mathcal{T}Q(s_t, a_t^i, a_t^{-i}) &= \beta R(s_t, a_T^i, a_T^{-i}) + \mathbb{E}_{P(s_{t+1}|s_t, a_t^i, a_t^{-i})}\left[V(s_{t+1})\right] \\ \text{where } V(s_{t+1}) &= \log \int \left(\int \exp\left(Q(s_t, a_t^i, a_t^{-i})\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \, \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \, \mathrm{d}a_t^{-i} \end{aligned} \tag{26}$$

We can show that it is contraction mapping i.e

$$\left\| \mathcal{T}Q^1(s_t, a_t^i, a_t^{-i}) - \mathcal{T}Q^2(s_t, a_t^i, a_t^{-i}) \right\|_\infty \leq \gamma \left\| Q^1(s_t, a_t^i, a_t^{-i}) - Q^2(s_t, a_t^i, a_t^{-i}) \right\|_\infty \tag{27}$$

For $\gamma \in [0, 1]$. Then by Banach fixed-point theorem, repeatedly applying this contraction mapping will lead to optimal action value function. The proof is show in appendix A.3.1. Furthermore, we would like to consider non-recursive (not necessary optimal) definition of action value function $Q_n^{\pi,\rho}(s_n, a_n^i, a_n^{-i})$ (short-handed as $Q^{\pi,\rho}$) as follows

$$\begin{aligned} Q_n^{\pi,\rho} =& \beta R(s_n, a_n^i, a_n^{-i}) + \\ & \mathbb{E}_{s_t, a_t^i, a_t^{-i} \sim \pi, \rho, P} \left[\sum_{t=n+1}^{T} \gamma^{t-n+1} \left(\beta R(s_t, a_t^i, a_t^{-i}) - \log \frac{\pi(a_t^i|s_t, a_t^{-i})}{P_{\text{prior}}(a_t^i|s_t, a_t^{-i})} - \log \frac{\rho(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)} \right) \right] \end{aligned} \tag{28}$$

We can show that $\mathbb{E}_{s_t, a_t^i, a_t^{-i} \sim \pi, \rho, P}\left[Q^*(s_n, a_n^i, a_n^{-i})\right] = Q_n^{\pi^*, \rho^*}(s_n, a_n^i, a_n^{-i})$ since after all $Q_n^{\pi,\rho}$ is the objective for the optimization that we have solved in earlier section, furthermore, one can expand the equation and arrived at the same equation, see appendix A.3.2. Now, we want to see given arbitrary opponent model ρ_s, we want to show that the improved policy $\tilde{\pi}$ defined as

$$\tilde{\pi}(a_t^i|s_t, a_t^{-i}) = \frac{\exp\left(Q^{\pi,\rho_s}(s_t, a_t^i, a_t^{-i})\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i})}{\exp\left(Q^{\pi,\rho_s}(s_t, a_t^{-i})\right)} \tag{29}$$

where the normalizing factor is equal to

$$Q^{\pi,\rho_s}(s_t, a_t^{-i}) = \log \int \exp\left(Q^{\pi,\rho_s}(s_t, a_t^i, a_t^{-i})\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \, \mathrm{d}a_t^i \tag{30}$$

improves the action value function i.e $Q^{\pi,\rho_s}(s_t, a_t^i, a_t^{-i}) \leq Q^{\tilde{\pi},\rho_s}(s_t, a_t^i, a_t^{-i})$ The proofs will be presented in appendix A.3.3. Similarly, for the opponent model, given arbitrary policy π_s, we define improvement of the opponent model ρ as

$$\tilde{\rho}_\phi(a_t^{-i}|s_t) = \frac{\exp\left(Q^{\pi_s,\rho}(s_t, a_t^{-i})\right) P_{\text{prior}}(a_t^{-i}|s_t)}{\exp\left(V^{\pi_s,\rho}(s_t)\right)} \tag{31}$$

Once can show that $Q^{\pi_s,\rho}(s_t, a_t^{-i}) \leq Q^{\pi_s,\tilde{\rho}}(s_t, a_t^{-i})$. The proofs will be presented in appendix A.3.4 Given the results on the improvement it is clear that

$$Q^{\pi,\rho}(s_t, a_t^{-i}) \leq Q^{\tilde{\pi},\tilde{\rho}}(s_t, a_t^{-i}) \tag{32}$$

Although the results is satisfactory, we can see that ROMMEO *truely* works only in cooperate setting and by making the opponent model to update toward to increase the agent's reward will not be useful in non-cooperating setting.

1.2.2 Balancing-Q

For Balancing-Q we will consider, first, the contraction mapping of the defined Bellman-equation in equation 25. Note that all the algorithms will be based on the point of view of the agent.

$$\begin{aligned}&\mathcal{T}Q^*(s_t, a_t^i, a_t^{-i}) = R(s_t, a_t, a_t^{-i}) + \mathbb{E}_{P(s_{t+1}|s_t, a_t, a_t^{-i})}\left[V^*(s_{t+1})\right] \\ &\text{where } V^*(s_{t+1}) = \log \int \exp\left(\frac{\beta^i}{\beta^{-i}} \log \int \exp\left(\beta^{-i} Q(s_t, a_t^i, a_t^{-i})\right) P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right) P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\end{aligned} \tag{33}$$

We can show that it is contraction mapping i.e

$$\left\|\mathcal{T}Q^1(s_t, a_t^i, a_t^{-i}) - \mathcal{T}Q^2(s_t, a_t^i, a_t^{-i})\right\|_\infty \leq \gamma \left\|Q^1(s_t, a_t^i, a_t^{-i}) - Q^2(s_t, a_t^i, a_t^{-i})\right\|_\infty \tag{34}$$

The proof (see appendix A.4.1) is almost the same as ROMMEO, however, we presented this as an alternative version to the main paper. Now, we would like to explore the ideas of how the β^i and β^{-i} are affecting the agents' behaviors via policy improvement theorem. Now, we have the action-value function for both agent's and opponent's

$$\begin{aligned}Q_n^{\pi,\rho} =& \beta R(s_n, a_n^i, a_n^{-i}) + \\ &\mathbb{E}_{s_t, a_t^i, a_t^{-i} \sim \pi, \rho, P}\left[\sum_{t=n+1}^{T} \gamma^{t-n+1}\left(R(s_t, a_t^i, a_t^{-i}) - \frac{1}{\beta^i}\log\frac{\pi(a_t^i|s_t)}{P_{\text{prior}}(a_t^i|s_t)} - \frac{1}{\beta^{-i}}\log\frac{\rho(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)}\right)\right]\end{aligned} \tag{35}$$

Without lose of generality, we will assume that β^i is always positive and β^{-i} can either be positive or negative relative to the agent's β^i. We want to show that, given the policy improvement on the opponent

$$\tilde{\rho}(a_t^{-i}|s_t) = \frac{\exp\left(Q^{-i,\pi,\rho}(s_t, a_t^{-i})\right) P_{\text{prior}}(a_s^{-i})}{\exp(V^{-i,\pi,\rho}(s_t))} \tag{36}$$

The action-value function is affected in difference way according to the sign of β^{-i}

$$\begin{cases} Q^{\pi,\rho}(s_n, a_n^i, a_n^{-i}) \geq Q^{\pi,\tilde{\rho}}(s_n, a_n^i, a_n^{-i}) & \text{if } \beta^{-i} < 0 \\ Q^{\pi,\rho}(s_n, a_n^i, a_n^{-i}) \leq Q^{\pi,\tilde{\rho}}(s_n, a_n^i, a_n^{-i}) & \text{if } \beta^{-i} > 0 \end{cases} \tag{37}$$

We shall show the proof in appendix A.4.2, given the assumption that

$$\begin{cases} D_{\text{KL}}\left(\pi(a_{t+1}^i|s_{t+1}) \middle\| \mathbb{E}_{a_{t+1}^{-i} \sim \rho}\left[\pi(a_{t+1}^i|a_{t+1}^{-i}, s_{t+1})\right]\right) \\ \quad \leq D_{\text{KL}}\left(\pi(a_{t+1}^i|s_{t+1}) \middle\| \mathbb{E}_{a_{t+1}^{-i} \sim \tilde{\rho}}\left[\pi(a_{t+1}^i|a_{t+1}^{-i}, s_{t+1})\right]\right) & \text{for } \beta^{-i} < 0 \\ D_{\text{KL}}\left(\pi(a_{t+1}^i|s_{t+1}) \middle\| \mathbb{E}_{a_{t+1}^{-i} \sim \rho}\left[\pi(a_{t+1}^i|a_{t+1}^{-i}, s_{t+1})\right]\right) \\ \quad \geq D_{\text{KL}}\left(\pi(a_{t+1}^i|s_{t+1}) \middle\| \mathbb{E}_{a_{t+1}^{-i} \sim \tilde{\rho}}\left[\pi(a_{t+1}^i|a_{t+1}^{-i}, s_{t+1})\right]\right) & \text{for } \beta^{-i} > 0 \end{cases} \tag{38}$$

Although for the case of $\beta^{-i} < 0$ might not seem to be intuitive at first glance, we will provide an interpretation of this in the next section.

1.3 Probabilistic Fictitious Play

In this next section, we will alsoto define ROMMEO like algorithms, which resulted in Fictitious Play like algorithms. As we can see in the derivation of optimal policy of Balancing-Q, we can in fact decouple the graphical models so that we could optimize both agent's policy and its opponent model separately. We will now present ROMMEO like agent that is able to handle competitive game, which is inspired by Balancing-Q algorithm.

1.3.1 Defining the model

We will start with the graphical model of the agent's policy with its prior on its prior on the opponent's policy that we want to optimize toward, which is represented in figure 3, whereby we defined variatioanl joint probability distribution that is the same as figure 2, with the opponet still representing by the prior and not the quatity that we want to optimize. With these, we can formalize the joint probability that we

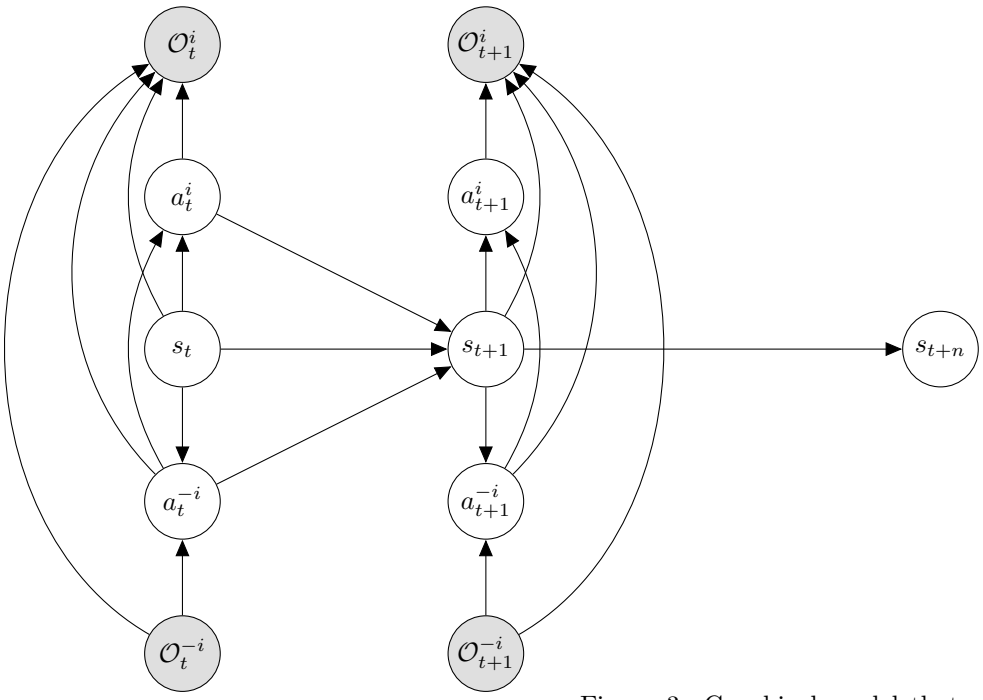

Figure 3: Graphical model that we want to approximate. For Unified Probabilistic Model, we doesn't assume that the opponent is optimal

want to approximate as

$$P(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}, \mathcal{O}^i_{1:T} = 1, \mathcal{O}^{-i}_{1:T} = 1)$$
$$= P(s_0) \prod_{t=0}^{T} P_{\text{prior}}(a_t^i | s_t, a_t^{-i}) P_{\text{prior}}(a_t^{-i} | s_t, \mathcal{O}_t^{-i} = 1) P(s_{t+1} | s_t, a_t^i, a_t^{-i}) P(\mathcal{O}_t^i = 1 | s_t, a_t^i, a_t^{-i}) P(\mathcal{O}_t^{-i} = 1) \tag{39}$$

While the variation probability is equal to

$$q(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}) = P(s_0) \prod_{t=0}^{T} \pi_\theta(a_t^i | s_t, a_t^{-i}) P_{\text{prior}}(a_t^{-i} | s_t, \mathcal{O}_t^{-i} = 1) P(s_{t+1} | s_t, a_t^i, a_t^{-i}) P(\mathcal{O}_t^{-i} = 1) \tag{40}$$

Please note in mind that the opponent we are playing against is the update given prior of the opponent as we assume "optimality" of the opponent (we assume our best knowledge of our opponent). Furthermore, we define agent's optimality random variable to be

$$P(\mathcal{O}_t^i = 1 | s_t, a_t^i, a_t^{-i}, \mathcal{O}_t^{-i} = 1) = \exp\left(\beta^i R(s_t, a_t^i, a_t^{-i})\right) \tag{41}$$

With this, we can do the variational inference and deriving the agent's policy to be

$$\begin{aligned}
\pi(a_t^i|a_t^{-i}, s_t) &= \frac{\exp\left(Q^i(s_t, a_t^i, a_t^{-i}))\right) P_{\text{prior}}(a_t^i|a_t^{-i}, s_t)}{\exp(Q^i(s_t, a_t^{-i}))} \\
\text{where } Q^i(s_t, a_t^i, a_t^{-i}) &= \beta^i R(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}, a_{t+1}^{-i} \sim P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1}, \mathcal{O}_{t+1}^{-i}=1)} \left[Q^i(s_{t+1}, a_{t+1}^{-i})\right] \\
Q^i(s_t, a_t^{-i}) &= \log \int \exp\left(Q^i(s_t, a_t^i, a_t^{-i}))\right) P_{\text{prior}}(a_t^i|a_t^{-i}, s_t)\ \mathrm{d}a_t^i
\end{aligned} \tag{42}$$

From now on, we will write $P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})$ instead of $P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1}, \mathcal{O}_{t+1}^{-i} = 1)$ to save space, which we "approximate" it to be $\rho_\phi(a_t^{-i}|s_t)$. All the proof will be presented in appendix A.5.1. Now, we define the joint probability of opponent's model represented in figure 4 We defined the optimal random variable to be

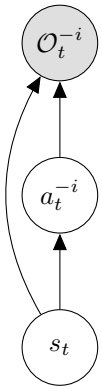

Figure 4: Graphical model that we want to approximate. This follows the VIREL type of graphical model rather than traditional probabilistic.

$$P(\mathcal{O}_t^{-i} = 1|s_t, a_t^{-i}) = \exp\left(\frac{\beta^{-i}}{\beta^i} Q^i(s_t, a_t^{-i}))\right) \tag{43}$$

The joint probability that we want to optimize is

$$P(s_t, a_t^{-i}, \mathcal{O}_t^{-i} = 1) = P(\mathcal{O}_t^{-i} = 1|s_t, a_t^{-i}) P(s_t) P_{\text{prior}}(a_t^{-i}|s_t) \tag{44}$$

And the variational probability is simply

$$q(s_t, a_t^{-i}) = P(s_t) \rho_\phi(a_t^{-i}|s_t) \tag{45}$$

Therefore, the optimal opponent model is defined to be (see appendix A.5.2 for the proof)

$$\begin{aligned}
\rho_\phi(a_t^{-i}|s_t) &= \frac{\exp\left(\frac{\beta^{-i}}{\beta^i} Q^i(s_t, a_t^{-i})\right) P_{\text{prior}}(a_t^{-i}|s_t)}{\exp\left(V^{-i}(s_t)\right)} \\
\text{where } V^{-i}(s_t) &= \log \int \exp\left(\frac{\beta^{-i}}{\beta^i} Q^i(s_t, a_t^{-i})\right) P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a_t^{-i}
\end{aligned} \tag{46}$$

With this we can train the agent by alternating between training the agent and opponent model. (using opponent model as next iteration prior for training the agent). Given this, we can derived the equality between the $V^{-i}(s_t)$ and $Q^i(s_t, a_t^{-i})$ as follows

$$Q^i(s_t, a_t^{-i}) = \frac{\beta^i}{\beta^{-i}} \left[\log \frac{\rho_\phi(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)} + V^{-i}(s_t)\right] \tag{47}$$

Let's trace-back into the the definition of $Q^i(s_t, a_t^i, a_t^{-i})$, we have the following equality

$$\frac{\beta^i}{\beta^{-i}}\left[\log\frac{\rho_\phi(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)} + V^{-i}(s_t)\right] = Q^i(s_t, a_t^{-i}) = -\log\frac{\pi(a_t^i|s_t, a_t^{-i})}{P_{\text{prior}}(a_t^i|s_t, a_t^{-i})} + Q^i(s_t, a_t^i, a_t^{-i}) \tag{48}$$

We see that the "connection" between value function of the opponent is

$$V^{-i}(s_t) = \frac{\beta^{-i}}{\beta^i}Q^i(s_t, a_t^i, a_t^{-i}) - \frac{\beta^{-i}}{\beta^i}\log\frac{\pi(a_t^i|a_t^{-i}, s_t)}{P_{\text{prior}}(a_t^i|a_t^{-i}, s_t)} - \log\frac{\rho_\phi(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)} \tag{49}$$

We can see that the addition in equation 47 is the way we "remove" the regularization of the opponent model, instead of doing this, let's directly inserted "corrected" $V^{-i}(s_t)$ in to the Bellman equation derived from the inverse of a factor in front of $Q^i(s_t, a_t^i, a_t^{-i})$ giving

$$Q^i(s_t, a_t^i, a_t^{-i}) = \beta^i R(s_t, a_t^i, a_t^{-i}) + \gamma\mathbb{E}_{s_{t+1}}\left[\frac{\beta^i}{\beta^{-i}}V^{-i}(s_{t+1})\right] \tag{50}$$

Now, what is $P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})$? According to our formulation and parallel connection to Balance-Q, we can use $P_{\text{prior}}(a_t^{-i}|s_t)$ to be our latest opponent model (our "best" model we can found). The will finally have the recursive formula that rollouts to be almost the same as Balance-Q.

1.4 Minimax-PR2

1.5 Perfect-Model

Now, we are left with the problem of totally difference reward as the reward from agent's and its opponent aren't proportional anymore. Although, this would be the most general case, we sacrifice the convergence property. We will starting with the definition of the optimality, which is simply

$$P(\mathcal{O}_t^i = 1|s_t, a_t^i, a_t^{-i}, \mathcal{O}_t^{-i} = 1) = \exp\left(R^i(s_t, a_t^i, a_t^{-i})\right) \tag{51}$$

Now, assume the same graphical model/joint distribution as the 1.3, we have similar solution to equation 42. However, the interesting part comes when we consider the optimality of the opponent model, which can be defined as

$$P(\mathcal{O}^{-i} = 1|s_t, a_t^{-i}) = \exp\left(Q^{-i}(s_t, a_t^{-i})\right) \tag{52}$$

where we usually have $Q^{-i}(s_t, a_t^{-i})$ to be the soft-max of the $Q^{-i}(s_t, a_t^i, a_t^{-i})$, which is $Q^{-i}(s_t, a_t^{-i}) = \log\int P_{\text{prior}}(a_t^i|a_t^{-i}, s_t)\exp\left(Q^{-i}(s_t, a_t^i, a_t^{-i})\right)\ \mathrm{d}a_t^i$, now the problem for us is how to arrived at such a Q-function, which should be defined as

$$Q^{-i}(s_t, a_t^i, a_t^{-i}) = r^{-i}(s_t, a_t^i, a_t^{-i}) + \gamma\mathbb{E}_{s_{t+1}}\left[V^{-i}(s_{t+1})\right] \tag{53}$$

Now, we can see that the opponent model can be equal to

$$\begin{aligned}\rho_\phi(a_t^{-i}|s_t) &= \frac{\exp\left(Q^{-i}(s_t, a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)}{\exp\left(V^{-i}(s_t)\right)} \\ \text{where } V^{-i}(s_t) &= \log\int\exp\left(Q^{-i}(s_t, a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a_t^{-i}\end{aligned} \tag{54}$$

Now, looking back at the property $Q^i(s_t, a_t^i, a_t^{-i})$, as we have the following Bellman equation

$$Q^i(s_t, a_t^i, a_t^{-i}) = \beta^i R(s_t, a_t^i, a_t^{-i}) + \gamma\mathbb{E}_{s_{t+1}, a_{t+1}^{-i}\sim P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1}, \mathcal{O}_{t+1}^{-i}=1)}\left[Q^i(s_{t+1}, a_{t+1}^{-i})\right] \tag{55}$$

Note that now, we also lose the guarantee that $V^i(s_t) \propto V^{-i}(s_t)$ as in the Balancing-Q case. After investigating the role of $V(s_t)$ from the section before, we can see that all we have to do left is to apply the entropy regularization to the next step calculation i.e

$$Q^i(s_t, a_t^i, a_t^{-i}) = \beta^i R(s_t, a_t^i, a_t^{-i}) + \gamma\mathbb{E}_{s_{t+1}, a_{t+1}^{-i}}\left[Q^i(s_{t+1}, a_{t+1}^{-i}) - \log\frac{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1}, \mathcal{O}_{t+1}^{-i}=1)}{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})}\right] \tag{56}$$

1.6 Intuitive Interpretation and Implementation

Before we consider how can we merge 2 difference models together let's consider conceptually how they works and how can we implement the algorithms.

1.6.1 Interpretation

Connection between algorithms: Probabilistic multi-agent reinforcement learning can be very confusing especially when we move to non-cooperative games as there is no more single objective that all agent should follows. However, before we move on to the conceptual overview of algorithms, we would like to gives a quick walkthrough of PR2. Starting with its factorization of joint policy probability

$$P(a_t^i, a_t^{-i}|s_t) = \pi(a_t^i|s_t)\rho(a_t^i|s_t, a_t^{-i}) \tag{57}$$

Now, we want to find agent policy that understand how to act given the opponent model that reacts to agent's action. Now it is very clear that this very similar to Balancing-Q learning in the case that each agent's policy is based on its opponent policy that conditional on the agent's action (nonetheless PR2 only works on cooperative cases). On the other hand, from our calculation, ROMMEO is PR2 on the case of opponent models, and Balancing Q (cooperative case) is simply ROMMEO where its opponent models are used (obviously opponent model for agent i is playing against opponent model for agent $-i$). We believes that ROMMEO makes more sense in the real world setting. That is because the "acting" policy is condition on other's action. As raised in Balancing-Q paper, since both agent doesn't know what the other's going to do, it has to use its prior of what other agent is, in order to calculate $Q^{i*}(s_t, a_t^i)$ and $Q^{-i*}(s_t, a_t^i)$. This also applies in the case of ROMMEO and PR2.

Why ROMMEO and PR2 doesn't work in minimax game ? As shown briefly in the opponent model improvement of ROMMEO, since the temperature variable doesn't contributes to opponent model as in Balancing-Q, the opponent model will be updated according to main objective of the agent, which perfectly make sense in coopeartive game, while being disastrous in competitive game. This is reflected in the assumption in equation 38, in the case that $\beta^{-i} > 0$ the opponent model should be updated in such a way that its agent conditional policy reflects the "real" agent's policy. On the other hand, in the case that $\beta^{-i} < 0$, we want the opponent to play in such a way that the agent's policy doesn't follows its own objective (minimax behavior). Thus further verify the hypothesis that β^{-i} reflected how opponent perceives its reward.

1.6.2 Implementation

There are 2 main approaches that we can use for single agent probabilistic reinforcement learning, mainly Soft actor critic and SVGD. We will start with Soft actor critic first, that is because it is easier to implement and understand. Suppose, we have the agent policy that is defined as

$$\begin{aligned} \pi(a|s) &= \frac{\exp\left(Q(s,a)\right) P_{\text{prior}}(a|s)}{\exp\left(V(s)\right)} \\ \text{where } Q(s,a) &= \beta R(s,a) + \mathbb{E}_{s' \sim P(s'|s,a)}\left[V(s')\right] \\ V(s) &= \log \int \exp\left(Q(s,a)\right) P_{\text{prior}}(a|s)\ \mathrm{d}a \end{aligned} \tag{58}$$

Soft-Actor Critic: There are mainly 3 quantities that we can parameterized on, with the following objectives:

- Value function: $V_\psi(s_t)$

$$\mathcal{L}_V(\psi) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\frac{1}{2}\left(V_\psi(s_t) - \mathbb{E}_{a_t \pi_\phi}\left[Q_\theta(s,a) - \log \frac{\pi_\phi(a_t|s_t)}{P_{\text{prior}}(a_t|s_t)}\right]\right)^2\right] \tag{59}$$

- Action value function: $Q_\theta(s_t, a_t)$

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s_t,a_t,s_{t+1})}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t) - \left[R(s_t,a_t) + \gamma \bar{V}_\psi(s_{t+1})\right]\right)^2\right] \tag{60}$$

- Policy: $\pi_\phi(\cdot|s_t) = f_\phi(\varepsilon; s_t)$ where $\varepsilon \sim \mathcal{N}(0, I)$, we shall use the KL-minimization, furthermore, $V(s)$ doesn't depends on ϕ, therefore, it can be safely removed from the objective.

$$\mathcal{L}_\pi(\phi) = \mathbb{E}_{s_t\sim\mathcal{D},\varepsilon_t\sim\mathcal{N}}\left[\log f_\phi(\varepsilon_t; s_t) - Q_\theta(s_t, f_\phi(\varepsilon_t; s_t)) - \log P_{\text{prior}}(a_t^i|s_t^i)\right] \tag{61}$$

where $\mathcal{D}$ is the replay buffer, $\bar{V}_\psi$ is the target value.

Soft-Q Learning : In this case, we represent the value function as an expectation of the action value function instead, while optimizing the policy based on SVGD.

- Value function $V_\theta(s_t)$: calculated as the importance sampling from arbitrary distribution q

$$V_\theta(s_t) = \log \mathbb{E}_{a'\sim q}\left[\frac{\exp\left(Q(s_t,a')\right) P_{\text{prior}}(a'|s_t)}{q(a')}\right] \tag{62}$$

- Action value function: $Q_\theta(s_t, a_t)$ trained to minimizes the following objective

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s_t,a_t,s_{t+1})}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t) - \left[R(s_t,a_t) + \gamma V_\theta(s_{t+1})\right]\right)^2\right] \tag{63}$$

- Policy: $\pi_\phi(\cdot|s_t) = f_\phi(\varepsilon; s_t)$ where we relies on SVGD, where

$$\begin{aligned}
&\frac{\partial \mathcal{L}_\pi(\phi)}{\partial \phi} \propto \mathbb{E}_\varepsilon\left[\Delta f_\phi(\varepsilon; s_t)\frac{\partial f_\phi(\varepsilon; s_t)}{\partial \phi}\right] \\
&\text{where } \Delta f_\phi(\cdot, s_t) = \mathbb{E}_{a_t\sim f_\phi}\Big[\kappa(a_t, f_\phi(\cdot, s_t))\nabla_{a'} Q_\theta(s_t, a')\log P_{\text{prior}}(a'|s_t)\big|_{a'=a_t} \\
&\qquad\qquad + \nabla_{a'}\kappa(a', f_\phi(\cdot, s_t))\big|_{a'=a_t}\Big]
\end{aligned} \tag{64}$$

  where κ is kernel function defined as $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\|x-x'\|}{\sigma^2})$

Balancing-Q: Now, we will present the algorithm that optimizes the objective given in equation 24 and 25, for Balancing-Q, as the authors doesn't extends to continuous setting from the agent's perspective (however value function can be shared), and we required to have some prior of the opponent on us and its β^{-i}

- Value function $V_\theta(s_t)$ can only be calculated by importance sampling from arbirary distributions: $q(a^{-i})$ and $q(a^i)$, as we doesn't have intermediate opponent policy.

$$\begin{aligned}
V_\theta(s_t) &= \frac{1}{\beta^{-i}}\log \mathbb{E}_{a^{-i}\sim q(a^{-i})}\left[\frac{\exp\left(Q_\theta^{-i}(s_t, a^{-i})\right) P_{\text{prior}}(a^{-i}|s)}{q(a^{-i})}\right] \\
\text{where } Q_\theta^{-i}(s_t, a^{-i}) &= \frac{\beta^{-i}}{\beta^i}\log \mathbb{E}_{a^i\sim q(a^i)}\left[\frac{\exp(\beta^i Q(s_t, a^i, a_t^{-i}))P_{\text{prior}}(a^i|s_t)}{q(a^i)}\right]
\end{aligned} \tag{65}$$

- Action value function $Q_\theta(s_t, a_t^i, a_t^{-i})$ is trained by minimizing the following objective

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s_t,a_t^i,a_t^{-i}s_{t+1})\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t^i,a_t^{-i}) - \left[R(s_t,a_t^i,a_t^{-i}) - \gamma V_\theta(s_{t+1})\right]\right)^2\right] \tag{66}$$

- Partial Action value function $Q_\theta^i(s_t, a_t^i)$ is calculated via importance sampling

$$Q_\theta^i(s_t, a_t^i) = \frac{\beta^i}{\beta^{-i}}\log \mathbb{E}_{a^{-i}\sim q(a^{-i})}\left[\frac{\exp(\beta^{-i} Q(s_t, a^i, a_t^{-i}))P_{\text{prior}}(a^{-i}|s_t)}{q(a^{-i})}\right] \tag{67}$$

- Policy: $\pi_\phi(\cdot|s_t) = f_\phi(\varepsilon; s_t)$ where $\varepsilon \sim \mathcal{N}(0, I)$ can be trained by minimizing KL-divergence

$$\mathcal{L}_\pi(\phi) = \mathbb{E}_{s_t\sim\mathcal{D},\varepsilon_t\sim\mathcal{N}}\left[\log f_\phi(\varepsilon_t; s_t) - Q_\theta^i(s_t, a_t^i) - \log P_{\text{prior}}(a_t^i|s_t)\right] \tag{68}$$

2 Hierarchical MARL

We will now consider the notion of master policy, which controls the agent on higher-level of temporal abstraction. This could be done via a non-trivial extension of graphical model (which we have describe in earlier section). Furthermore, this could be seen as one step toward public belief agent (See section on Representing Belief). Given this, we could also extends multi-tasking perspective on the agent, which could serves as the way to handle difference kinds of opponents.

2.1 Single Agent Soft-Hierachical Reinforcement Learning

Now, before we move on to the problem of multi-agent reinforcement learning, we will have to consider the graphical model of single agent problem. This would turn out to be non-trivial to solve using our regular technique in the section before. We will associate our model with option learning paradigm, and extends the result in [7] by cooperating Soft-Actor Critic type solution as it is standard method for solving probabilistic reinforcement learning. Furthermore, our approach is similar to [1], in terms of using similar probabilistic technique, however, in the paper, the authors concerning more on inferring policies from expert data, while didn't explicitly solve for each policy given the optimally condition (what we are trying to do now). In the nutshell, we have a master policy π^H that suggesting the "mode of execution" or option to lower level policy π based on termination policy π^T, which telling what the next option should be. The joint distribution of actions i.e $P(s_{1:T}, a_{1:T}, \mathcal{O}_{1:T}, z_{1:T}, b_{1:T})$ is depicted in figure 5. The joint distribution is then

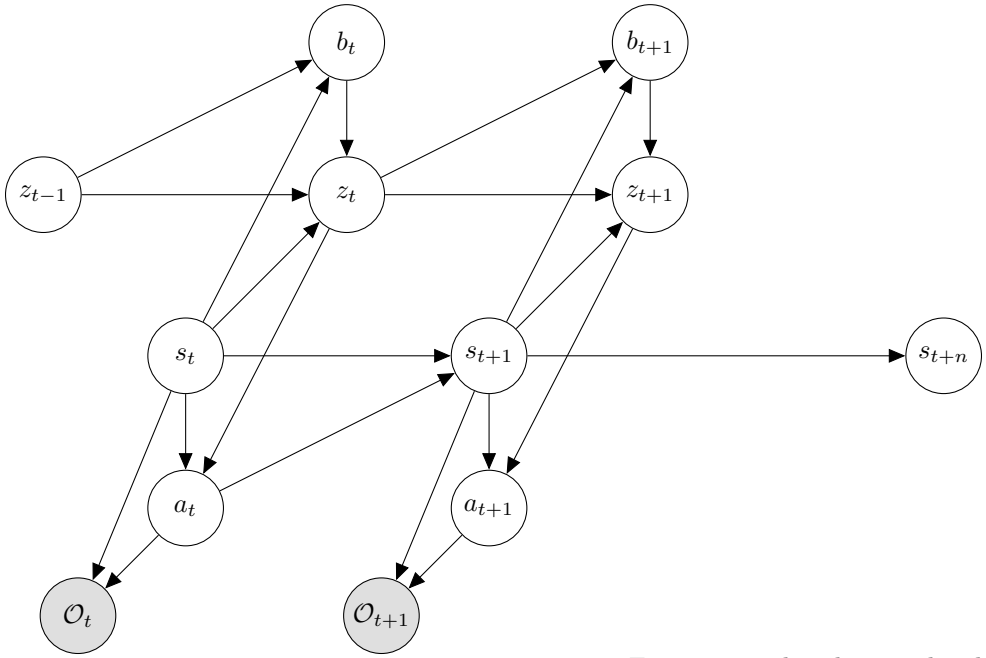

Figure 5: The the graphical model of hierarchical agent. We have the master policy which produces z_t while having switch goal policy producing policy b_t

$$P(s_{1:T}, a_{1:T}, z_{1:T}, b_{1:T}) = P(s_0)P(z_0)\prod_{t=0}^{T} P(s_{t+1}|s_t, a_t)P_{\text{prior}}(a_t|s_t, z_t) \\ P^H_{\text{prior}}(z_t|s_t, z_{t-1}, b_t)P^T_{\text{prior}}(b_t|s_t, z_{t-1})P(\mathcal{O}_t = 1|s_t, a_t) \tag{69}$$

The variations joint distribution is defined as

$$q(s_{1:T}, a_{1:T}, z_{1:T}, b_{1:T}) = P(s_0)P(z_0)\prod_{t=0}^{T} P(s_{t+1}|s_t, a_t)\pi_\theta(a_t|s_t, z_t) \\ \pi_\phi^H(z_t|s_t, z_{t-1}, b_t)\pi_\phi^T(b_t|s_t, z_{t-1}) \tag{70}$$

This is the most general form of the joint distribution. Usually, the master policy and its prior are defined as, which is straightforward.

$$\pi_\phi^H(z_t|s_t, z_{t-1}, b_t) = (1-b_t)\delta(z_t - z_{t-1}) + b_t q_\phi^H(z_t|s_t) \\ P_{\text{prior}}^H(z_t|s_t, z_{t-1}, b_t) = (1-b_t)\delta(z_t - z_{t-1}) + b_t \frac{1}{m} \tag{71}$$

Now, the ELBO is equal to

$$\mathbb{E}\left[\sum_{t=1}^{T} \beta r(s_t, a_t) - \log \frac{\pi_\theta(a_t|s_t, z_t)}{P_{\text{prior}}(a_t|s_t, z_t)} - \log \frac{\pi_\phi^H(z_t|s_t, z_{t-1}, b_t)}{P_{\text{prior}}^H(z_t|s_t, z_{t-1}, b_t)} - \log \frac{\pi_\phi^T(b_t|s_t, z_{t-1})}{P_{\text{prior}}^T(b_t|s_t, z_{t-1})}\right] \tag{72}$$

Now solving this can be non-trivial, since there are more than 2 policies, which are not directly influence the reward, unlike multi-agent case. We will solve the problem in closed form in appendix B.1.1, where we starting from solving the master policy first (as most of the option framework does). However, we can verify that solving by deriving the lower-level and then use the partition function to solve the master policy also yields the same solution (with simpler steps) and clarify on the differences between method in [12] (which contain a crucial mistake) and ours (that intuitively corrects the solution). All of the additional finding is presented in appendix B.1.2. Now, we have the following closed form solution for Soft-Hierarchical Reinforcement Learning: For the low-level policy $\pi_\theta(a_t|s_t, z_t)$, we have

$$\pi_\theta(a_t|s_t, z_t) = \frac{P_{\text{prior}}(a_t|s_t, z_t)\exp\left(Q(s_t, a_t, z_t)\right)}{\exp V(s_t, z_t)} \\ \text{where } V(s_t, z_t) = \log \int P_{\text{prior}}(a_t|s_t, z_t)\exp\left(Q(s_t, a_t, z_t)\right)\ \mathrm{d}a_t \tag{73}$$

For master policy $\pi_\phi^H(z_t|s_t)$ we have the following

$$\pi^H(z_t|s_t) = \frac{\frac{1}{m}\exp\left(Q^H(s_t, z_t)\right)}{\exp\left(V^H(s_t)\right)} \\ \text{where } Q^H(s_t, z_t) = \mathbb{E}_{a_t\sim\pi(a_t|s_t, z_t)}\left[Q(s_t, a_t, z_t)\right] \\ V^H(s_t) = \frac{1}{m}\log \int \exp\left(Q(s_t, z_t)\right)\ \mathrm{d}z_T \tag{74}$$

Finally, the termination policy $\pi^T(b_t|s_t, z_{t-1})$ is

$$\pi^T(b_t|s_t, z_{t-1}) = \frac{P_{\text{prior}}^T(b_t|s_t, z_{t-1})\exp U(s_t, z_{t-1}, b_t)}{\exp U(s_t, z_{t-1})} \\ \text{where } U(s_t, z_{t-1}, b_t) = b_t\left[V^H(s_t)\right] + (1-b_t)V(s_t, z_{t-1}) \\ U(s_t, z_{t-1}) = \log \int P_{\text{prior}}^T(b_t|s_t, z_{t-1})\exp\left(b_t\left[V^H(s_t)\right] + (1-b_t)V(s_t, z_{t-1})\right)\ \mathrm{d}b_t \tag{75}$$

Finally, the Bellman equation for soft-hierarchical single-agent reinforcement learning is

$$Q(s_t, a_t, z_t) = \beta r(s_t, a_t) + \mathbb{E}_{s_{t+1}}\left[\mathbb{E}_{a_{t+1}, z_{t+1}, b_{t+1}}\left[U(s_{t+1}, z_t)\right]\right] \tag{76}$$

Given this, we can derive the soft-actor critic objective for all components.

2.2 Hierachical Multi-Agent Reinforcement Learning

Now, we will consider the problem in which the master policy have to assign the right option given the opponent's action. This would be a direct extension to the single agent HRL problem. We should start with the graphical model of the problem, see figure 6. Given this, we can see that the joint probability distribution becomes **[[Phu: Fix the Time step]]**

$$\begin{aligned}
&P(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}, z^i_{1:T}, z^{-i}_{1:T}, b^i_{1:T}, b^{-i}_{1:T}, \mathcal{O}^i_{1:T} = 1, \mathcal{O}^{-i}_{1:T} = 1) \\
&\quad = P(s_0)P(z^i_0)P(z^{-i}_0)P(b^i_0)P(b^{-i}_0)\prod_{t=0}^{T} P(s_{t+1}|s_t, a^i_t, a^{-i}_t)P_{\text{prior}}(a^i_t|s_t, z^i_t)P^{H,i}_{\text{prior}}(z^i_t|s_t, z^i_{t-1}, b^i_t, a^{-i}_t) \\
&\quad P^{T,i}_{\text{prior}}(b^i_t|s_t, z^i_{t-1})P_{\text{prior}}(a^{-i}_t|s_t, z^{-i}_t)P^{H,-i}_{\text{prior}}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t)P^{T,-i}_{\text{prior}}(b^{-i}_t|s_t, z^{-i}_{t-1}) \\
&\quad P(\mathcal{O}^i_t = 1|\mathcal{O}^{-i}_{1:T} = 1)P(\mathcal{O}^{-i}_{1:T} = 1)
\end{aligned} \tag{77}$$

Where we have the optimal random variable to be

$$P(\mathcal{O}^i_t = 1|\mathcal{O}^{-i}_{1:T} = 1) \propto \exp\left(\beta R^i(s_t, a^i_t, a^{-i}_t)\right) \tag{78}$$

The variational joint distribution defined as

$$\begin{aligned}
&q(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}, z^i_{1:T}, z^{-i}_{1:T}, b^i_{1:T}, b^{-i}_{1:T}, \mathcal{O}^{-i}_{1:T} = 1) \\
&\quad = P(s_0)P(z^i_0)P(z^{-i}_0)P(b^i_0)P(b^{-i}_0)\prod_{t=0}^{T} P(s_{t+1}|s_t, a^i_t, a^{-i}_t)\pi_\theta(a^i_t|s_t, z^i_t)q^{H,i}(z^i_t|s_t, z^i_{t-1}, b^i_t, a^{-i}_t) \\
&\quad q^{T,i}(b^i_t|s_t, z^i_{t-1})\rho_\theta(a^{-i}|s_t, z^{-i}_t)q^{H,-i}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t)q^{T,-i}(b^{-i}_t|s_t, z^{-i}_{t-1})P(\mathcal{O}^{-i}_{1:T} = 1)
\end{aligned} \tag{79}$$

Given both of them, we can derive the ELBO by finding KL-divergence, which is equal to

$$\begin{aligned}
\mathbb{E}_q\Bigg[\sum_{t=0}^{T} \beta R(s_t, a^i_t, a^{-i}_t) &- \log\frac{\pi_\theta(a^i_t|s_t, z^i_t)}{P_{\text{prior}}(a^i_t|s_t, z^i_t)} - \log\frac{\rho_\theta(a^{-i}_t|s_t, z^{-i}_t)}{P_{\text{prior}}(a^{-i}_t|s_t, z^{-i}_t)} - \log\frac{q^{H,i}(z^i_t|s_t, z^i_{t-1}, b^i_t, a^{-i}_t)}{P^{H,i}_{\text{prior}}(z^i_t|s_t, z^i_{t-1}, b^i_t, a^{-i}_t)} \\
&- \log\frac{q^{H,-i}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t)}{P^{H,-i}_{\text{prior}}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t)} - \log\frac{q^{T,i}(b^i_t|s_t, z^i_{t-1})}{P^{T,i}_{\text{prior}}(b^i_t|s_t, z^i_{t-1})} - \log\frac{q^{T,-i}(b^{-i}_t|s_t, z^{-i}_{t-1})}{P^{T,-i}_{\text{prior}}(b^{-i}_t|s_t, z^{-i}_{t-1})}\Bigg]
\end{aligned} \tag{80}$$

Before we move on, let's specifically define the master policy's conditional probability and its prior:

$$\begin{aligned}
q^{H,i}(z^i_t|s_t, z^i_{t-1}, b^i_t, a^{-i}_t) &= (1 - b^i_t)\delta(z^i_t - z^i_{t-1}) + b^i_t q^{H,i}_\phi(z^i_t|s_t, a^{-i}_t) \\
P_{\text{prior}}(z^i_t|s_t, z^i_{t-1}, b^i_t, a^{-i}_t) &= (1 - b^i_t)\delta(z^i_t - z^i_{t-1}) + b^i_t\frac{1}{m} \\
q^{H,-i}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t) &= (1 - b^{-i}_t)\delta(z^{-i}_t - z^{-i}_{t-1}) + b^{-i}_t q^{H,-i}_\phi(z^{-i}_t|s_t) \\
P_{\text{prior}}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t) &= (1 - b^{-i}_t)\delta(z^{-i}_t - z^{-i}_{t-1}) + b^{-i}_t\frac{1}{m}
\end{aligned} \tag{81}$$

Now, the solution will be the following:

$$\begin{aligned}
\pi_\theta(a^i_t|s_t, z^i_t, a^{-i}_t, z^{-i}_t) &= \frac{P_{\text{prior}}(a^i_t|s_t, z^i_t, a^{-i}_t, z^{-i}_T)\exp\left(Q(a^i_t, s_t, z^i_t, a^{-i}_t, z^{-i}_t)\right)}{\exp Q^H(s_t, z^i_t, a^{-i}_t, z^{-i}_t)} \\
\text{where } Q^H(s_t, z^i_t, a^{-i}_t, z^{-i}_t) &= \int P_{\text{prior}}(a^i_t|s_t, z^i_t, a^{-i}_t, z^{-i}_T)\exp\left(Q(a^i_t, s_t, z^i_t, a^{-i}_t, z^{-i}_t)\right)\,\mathrm{d}a^i_t
\end{aligned}$$

Now for the

$$\begin{aligned}
q^{H,i}(z^i_t|s_t, a^{-i}_t, z^{-i}_t) &= \frac{P^{H,i}_{\text{prior}}(z^i_t|s_t, a^{-i}_t, z^{-i}_t)\exp(Q^H(s_t, z^i_t, a^{-i}_t, z^{-i}_t))}{\exp Q(s_t, a^{-i}_t, z^{-i}_t)} \\
\text{where } Q(s_t, a^{-i}_t, z^{-i}_t) &= \log\int P^{H,i}_{\text{prior}}(z^i_t|s_t, a^{-i}_t, z^{-i}_t)\exp(Q^H(s_t, z^i_t, a^{-i}_t, z^{-i}_t))\,\mathrm{d}z^i_t
\end{aligned} \tag{82}$$

Now we have

$$q^{T,i}(b_t^i|s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i}) = \frac{P_{\text{prior}}(b_t^i|s_t, z_{t-1}^i, z_t^{-i}) \exp Q(b_t^i, s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i})}{\exp Q(s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i})}$$
$$\text{where } Q(b_t^i, s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i}) = b_t^i \left[Q(s_t, a_t^{-i}, z_t^{-i})\right] + (1 - b_t^i) \left[Q(s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i})\right]$$
$$Q(s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i}) = \log \int P_{\text{prior}}(b_t^i|s_t, z_{t-1}^i, z_t^{-i}) \exp Q(b_t^i, s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i}) \, \mathrm{d}b_t^i$$

And

$$\rho_\theta(a_t^{-i}|s_t, z_t^{-i}, z_{t-1}^i) = \frac{P_{\text{prior}}(a_t^{-i}|s_t, z_t^{-i}, z_{t-1}^i) \exp Q(s_t, z_{t-1}^i, a_t^{-i}, z_t^{-i})}{\exp Q^H(s_t, z_{t-1}^i, z_t^{-i})} \tag{83}$$
$$\text{where } Q^H(s_t, z_{t-1}^i, z_t^{-i}) = \log \int P_{\text{prior}}(a_t^{-i}|s_t, z_t^{-i}, z_{t-1}^i) \exp Q^H(s_t, z_{t-1}^i, z_t^{-i}) \, \mathrm{d}a_t^i$$

And

$$q^{H,-i}(z_t^{-i}|s_t, z_{t-1}^i) = \frac{P_{\text{prior}}(z_t^{-i}|s_t, z_{t-1}^i) \exp\left(Q^H(s_t, z_{t-1}^i, z_t^{-i})\right)}{\exp Q(s_t, z_{t-1}^i)} \tag{84}$$
$$\text{where } Q(s_t, z_{t-1}^i) = \log \int P_{\text{prior}}(z_t^{-i}|s_t, z_{t-1}^i) \exp\left(Q^H(s_t, z_{t-1}^i, z_t^{-i})\right) \, \mathrm{d}z_t^{-i}$$

For the changing stuff.

$$q^{T,-i}(b_t^{-i}|s_t, z_{t-1}^{-i}, z_{t-1}^i) = \frac{P_{\text{prior}}(b_t^{-i}|s_t, z_{t-1}^{-i}, z_{t-1}^i) \exp\left(Q(b_t^{-i}, s_t, z_{t-1}^i, z_t^{-i})\right)}{\exp Q(s_t, z_{t-1}^i, z_t^{-i})}$$
$$\text{where } Q(b_t, s_t, z_{t-1}^i, z_{t-1}^{-i}) = b_t^i \left[Q(s_t, z_{t-1}^i)\right] + (1 - b_t^i) \left[Q^H(s_t, z_{t-1}^i, z_{t-1}^{-i})\right] \tag{85}$$
$$Q(s_t, z_{t-1}^i, z_{t-1}^{-i}) = \log \int P_{\text{prior}}(b_t^{-i}|s_t, z_{t-1}^{-i}, z_{t-1}^i) \exp\left(Q(b_t^{-i}, s_t, z_{t-1}^i, z_{t-1}^{-i})\right) \, \mathrm{d}b_t^{-i}$$

Finally, for the Bellman equation, we got

$$Q(s_t, a_t^i, a_t^{-i}, z_t^i, z_t^{-i}) = \beta R(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}} \left[Q(s_{t+1}, z_t^i, z_t^{-i})\right] \tag{86}$$

3 EM MERL

We are going back to the basic of probabilistic multi-agent reinforcement learning, and we are going to solve the problem via expectation-maximization (variational EM to be exact). Hopefully this would lead to more exciting results and intuitions.

3.1 ROMMEO based EM-Algorithm

Now, let's consider the following graphical model that model the interaction between agents, similar to one proposed in VIREL. This is depicted in figure We want to find $\pi(a^i|s)$ that maximizes the probability of optimality, which is defined as

$$P(\mathcal{O}^i = 1|s, a^i, a^{-i}) \propto \exp Q_\omega^i(s, a^i, a^{-i}) \tag{87}$$

Where $Q_\omega^i(s, a^i, a^{-i})$ is the joint-action value function for agent i. Note that this is entirely agent i's perspective, which includes its opponent model. Following the Bayes' rule, we want to do the following inference + we will follow the MPO's direction of defining the prior over actions and state where $P(s, a^i, a^{-i}) = \mathcal{U}(s) P_{\text{prior}}(a^i|s, a^{-i}) P_{\text{prior}}(a^{-i}|s)$, which is equal to

$$P(s, a^i, a^{-i}|\mathcal{O}^i = 1) = \frac{\exp Q_\omega^i(s, a^i, a^{-i}) \mathcal{U}(s) P_{\text{prior}}(a^i|s, a^{-i}) P_{\text{prior}}(a^{-i}|s)}{\int \exp Q_\omega^i(s, a^i, a^{-i}) \mathcal{U}(s) P_{\text{prior}}(a^i|s, a^{-i}) P_{\text{prior}}(a^{-i}|s) \, \mathrm{d}s \, \mathrm{d}a^i \, \mathrm{d}a^{-i}} \tag{88}$$

See the proof in appendix C.1.1. Given this let's consider the ROMMEO case where we want to infer the $P(a^i|s, a^{-i}, \mathcal{O}^i = 1)$:

$$P(a^i|s, a^{-i}, \mathcal{O}^i = 1) = \frac{\exp\left(Q^i_\omega(s, a^i, a^{-i})\right) P_{\text{prior}}(a^i|s, a^{-i})}{\int \exp\left(Q^i_\omega(s, a^i, a^{-i})\right) P_{\text{prior}}(a^i|s, a^{-i})\ \mathrm{d}a^i} \tag{89}$$

The proof will be appear in appendix C.1.1. This would be our form of the agent. Now, let's derived the variational EM. Let's consider maximizing the probability of being optimal, which lead us to following factorization, following our usual EM procedure:

$$\begin{aligned}\log P(\mathcal{O}^i = 1) = &\int q_\theta(s, a^i, a^{-i}) \log\left(\frac{P_\omega(\mathcal{O}^i = 1, s, a^i, a^{-i})}{q_\theta(s, a^i, a^{-i})}\right) \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i} \\ &+ \int q_\theta(s, a^i, a^{-i}) \log\left(\frac{q_\theta(s, a^i, a^{-i})}{P_\omega(s, a^i, a^{-i}|\mathcal{O}^i = 1)}\right) \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}\end{aligned} \tag{90}$$

Where the joint variational distribution is defined as $\mathcal{U}(s)\pi_\theta(a^i|s, a^{-i})\rho_\theta(a^i|s, a^{-i})$ Now, as we shown before, maximizing ELBO is the same thing as variational inference, therefore, we consider only the ELBO term, which we can expand it to be:

$$\begin{aligned}\mathcal{L}(\theta, \omega) = \mathbb{E}_q\left[Q^i_\omega(s, a^i, a^{-i})\right] &- D_{\text{KL}}\left(\pi_\theta(a^i|s, a^{-i}) \middle\| P_{\text{prior}}(a^i|s, a^{-i})\right) \\ &- D_{\text{KL}}\left(\rho_\theta(a^i|s) \middle\| P_{\text{prior}}(a^{-i}|s)\right)\end{aligned} \tag{91}$$

The proof will be in appedix C.1.2 Now the training scheme for Variational EM is the following

$$\theta_{t+1} \leftarrow \arg\max_{\theta_t} \mathcal{L}(\theta_t, \omega_t) \qquad \omega_{t+1} \leftarrow \arg\max_{\omega_{t+1}} \mathcal{L}(\theta_{t+1}, \omega_t) \tag{92}$$

It is similar to ROMMEO model, in which we that with 2 distinct differences

- We directly use the Q-function with isn't usually the soft Q, and derived from other centralized methods based on our opponent action. Note that this isn't that hard assuming fully observable state with ability to keep track of opponent models. For more integrated definition for Q-function training, we shall consider the inclusion of error within optimally condition.
- Instead of using our predefined prior (usually uniform distribution) we use our last step policy/opponent model to be our prior. This is similar usage to [14], which shows that method similar to this (optimized KL-divergence regularized reward from both agent and opponent converges to Nash equilibrium in special zero-sum game). Looking more deeply, we can see that the value $Q^i(s_t, a^i_t)$ is based on agent's old policy rather than its prior i.e

$$Q^*\left(s_t, a^{-i}_t\right) = \log \int \exp Q\left(s_t, a^i_t, a^{-i}_t\right) \pi_{t-1}\left(a^i_t|s_t, a^{-i}_t\right) \mathrm{d}a^i_t \tag{93}$$

  This would assume that the opponent/opponent model have a direct access to agent at last time step, and anticipate the agent's policy, which isn't the case for models that we have developed till now, as we always assume static prior policy over any given policies. We can therefore, see the training as nested "recursive reasoning"

3.2 Generalized EM-MARL

Given the first point, we have now reach the problem posted in the first part, in which, the opponent model update according to the agent's reward rather than its own reward. Luckily, we have solved the problem posted before hand, and therefore, we are able to quickly applied what we have found to this model. Drawing from the insight that Balancing-Q can be made into decentralized version by consider the following condition the agent, in which, the optimally of $\mathcal{O}^i$ is based on the soft-max of a^{-i}. Starting with the graphical model of agent's (figure 8), given known model Now the joint probability becomes

$$P(s, a^i, a^{-i}, \mathcal{O}^i = 1, \mathcal{O}^{-i} = 1) = \mathcal{U}(s)P(a^i|s, a^{-i})P(a^{-i}|s_t, \mathcal{O}_t^{-i} = 1) \\ P(\mathcal{O}^{-i})P(\mathcal{O}^i = 1|\mathcal{O}^{-i} = 1, a^i, a^{-i}, s) \tag{94}$$

where the optimality condition for the agent being:

$$P(\mathcal{O}^i = 1|\mathcal{O}^{-i} = 1, a^i, a^{-i}, s) \propto \exp Q_\omega^i(s, a^i, a^{-i}) \tag{95}$$

Please note that $Q_\omega^i(s, a^i, a^{-i})$ embedded the value function in which, we have to access the optimal opponent's policy. Now, following the same procedure, we have the following posterior on the agent, given the prior $P(s, a^i, a^{-i}) = \mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P^*(a^{-i}|s, \mathcal{O}^{-i} = 1)P(\mathcal{O}^{-i} = 1)$

$$P(a^i|s, a^{-i}, \mathcal{O}^i = 1, \mathcal{O}^{-i} = 1) = \frac{\exp\left(Q_\omega^i(s, a^i, a^{-i})\right) P_{\text{prior}}(a^i|s, a^{-i})}{\int \exp\left(Q_\omega^i(s, a^i, a^{-i})\right) P_{\text{prior}}(a^i|s, a^{-i}) \ \mathrm{d}a^i} \tag{96}$$

See appendix C.2.1 for proof. Following the same procedure we have the following ELBO, where we assign the variational distribution to be $q(s, a^i, a^{-i}, \mathcal{O}^{-i} = 0) = \mathcal{U}(s)\pi_\theta(a^i|s, a^{-i})P^*(a^{-i}|s, \mathcal{O}^{-i} = 1)P(\mathcal{O}^{-i} = 1)$, as we assume to have the knowledge of optimal opponent model.

$$\mathcal{L}^i(\theta, \omega) = \mathbb{E}_q\left[Q_\omega^i(s, a^i, a^{-i})\right] - D_{\text{KL}}\left(\pi_\theta(a^i|s, a^{-i}) \middle\| P_{\text{prior}}(a^i|s, a^{-i})\right) \tag{97}$$

which we can use the variational-EM. Now, before, we show the finaly training loop, let's consider how can we calculate/infer the "optimal" agent model. We shall use the graphical model similar to one presented in first sector. We have the following graphical model (figure 9), which is similar to one on presented Now, we can see that the joint probability is equal to

$$P(s, a^i a^{-i}, \mathcal{O}^{-i} = 1, \mathcal{O}^i = 1) = \mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i}, s, \mathcal{O}^i = 1) \\ P(\mathcal{O}^i = 1)P(\mathcal{O}^{-i} = 1|\mathcal{O}^i = 1, a^{-i}, s) \tag{98}$$

Where the optimality condition is defined to be

$$P(\mathcal{O}^{-i} = 1|\mathcal{O}^i = 1, a^{-i}, s) = \int \pi_\theta(a^i|a^{-i}, s) \exp\left(Q_\omega^{-i}(s, a^i, a^{-i})\right) \ \mathrm{d}a^i = \exp Q_{\psi,\theta}^{-i}(s, a^{-i}) \tag{99}$$

where $\pi_\theta(a^i|a^{-i}, s)$ comes from the solution of $\mathcal{L}^i(\theta, \omega)$. This is seem as the exponential "soft-max" with prior given by the current best agent policy, and $Q^{-i}(s, a^i, a^{-i})$ can be action value function of the opponent, which is arbitrary. We can show that the posterior of the policy is

$$P(a^{-i}|s, \mathcal{O}^{-i} = 1, \mathcal{O}^i = 1) = \frac{Q^{-i}(s, a^{-i})P(a^{-i}|s)}{\int \exp\left(Q^{-i}(s, a^{-i})\right) P(a^{-i}|s) \ \mathrm{d}a^{-i}} \tag{100}$$

The proof will be in appendix C.2.2. Now, let's consider variational distribution, which will be as follows (Assuming we are aware of agent's optimality), which is $q(s, a^i, a^{-i}, \mathcal{O}^i = 1) = \mathcal{U}(s)\rho_\phi(a^{-i}|s)P^*(a^i|a^{-i}, s, \mathcal{O}^i = 1)P(\mathcal{O}^i = 1)$ Now the ELBO associated with this should be

$$\mathcal{L}_\theta^{-i}(\phi, \psi) = \mathbb{E}_q\left[Q_{\psi,\theta}^{-i}(s, a^{-i})\right] - D_{\text{KL}}\left(\rho_\phi(a^{-i}|s) \middle\| P_{\text{prior}}(a^{-i}|s)\right) \tag{101}$$

Which, we can use variational EM to do coordinate descent on this objective. Now, it should be the time to consider the training loop for 2 of the ELBO (equation 97 and 101), we simply starting with the agents and then its opponent model, similar to our training on ROMMEO EM (the process starts from left to right and then next line).

$$\theta_{t+1} \leftarrow \arg\max_{\theta_t} \mathcal{L}^i(\theta_t, \omega_t) \qquad \omega_{t+1} \leftarrow \arg\max_{\omega_t} \mathcal{L}^i(\theta_{t+1}, \omega_t) \\ \phi_{t+1} \leftarrow \arg\max_{\phi_t} \mathcal{L}_{\theta_{t+1}}^{-i}(\phi_t, \psi_t) \qquad \psi_{t+1} \leftarrow \arg\max_{\psi_t} \mathcal{L}_{\theta_{t+1}}^{-i}(\phi_{t+1}, \psi_t) \tag{102}$$

4 Representing Belief

Represent belief is crucial for unknown environment, which is inherently part of multi-agent reinforcement learning, as we don't know what other agents are up to, all we can do is speculate and update our belief based on the observation. We will explore mainly 2 path: common belief, and I-POMDP. Common belief is the way to remove the need for recursive reasoning and replace with consensus update of public knowledge (more formal treatment of the subject will be explored in the text), while I-POMDP approach embrace the recursive reasoning, which is usually infinity, by finitely update the belief of others and itself anticipating on the outcome of the opponent model.

4.1 Solving Probabilistic Recursive Reasoning

4.1.1 Understanding Cooperative

Now, we will consider the model as proposed in [18]. We shall start by consider how recursive reasoning can be represented. The authors in [18] uses the the fact that we can utilized our understanding of other opponent as the prior. Instead of solving ELBO in terms of the hierarchy of policies, we can just simply include the other agent's lower-level policy as our prior belief. This lead to the following objective: For the agent's or opponent's policy at reasoning step k.

Now, let's consider what does it means for recursive reasoning. We will following the chain of reasoning as below

$$\rho_0(a_t^{-i}|s_t) \to \pi_1(a_t^i|a_t^{-i}, s_t) \to \rho_2(a_t^{-i}|s_t, a_t^i) \tag{103}$$

Now consider the first part, note that this is visually inverse from the one we proposed earlier (although the $\rho_0(a_t^{-i}|s_t)$ would likely best response from even lower-level policy). Given this, we can see that $Q_0(s_t, a_t^i, a_t^{-i})$ will be defined as

$$Q_1(s_t, a_t^i, a_t^{-i}) = R(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}}\left[V_1(s_t)\right] \tag{104}$$

where $V({}_0s_t)$ is defined by the soft-max of both and we are left with the following marginalized policies

$$\begin{aligned} V_1(s_t) &= \log \int \rho_0(a_t^{-i}|s_t) \int \pi_0(a_t^i|a_t^{-i}, s_t) \exp Q_1(s_t, a_t^i, a_t^{-i}) \ \mathrm{d}a_t^i, a_t^{-i} \\ \rho_1(a_t^i|s_t) &= \frac{\rho_0(a_t^{-i}|s_t) \exp Q_0(s_t, a_t^{-i})}{\exp V_0(s_{t+1})} \end{aligned} \tag{105}$$

Please note that the marginalization in the case of $\pi(a_t^i|s_t)$ would be based on ρ_1 that is because the Q-function $Q(s_t, a_t^i, a_t^{-i})$ is based on the fact that the next state is executed by $\rho_1(a_t^i|s_t)$ therefore it doesn't make sense to use ρ_0 as at the end the Bellman equation will calculate the expected reward given the fact that ρ_1 is playing along and not ρ_0. Now, the associated value-function for $\rho_2(a_t^{-i}|s_t, a_t^i)$ would be

$$Q_1(s_t, a_t^i, a_t^{-i}) = R(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}}\left[V_1(s_{t+1})\right] \tag{106}$$

Then we have the following value function and marginalized policies

$$\begin{aligned} V_1(s_t) &= \log \int \pi_1(a_t^i|s_t) \int \rho_0(a_t^{-i}|s_t, a_t^i) \exp Q_1(s_t, a_t^i, a_t^{-i}) \ \mathrm{d}a_t^i, a_t^{-i} \\ \pi_2(a_t^i|s_t) &= \frac{\pi_0(a_t^i|s_t) \exp Q_1(s_t, a_t^i)}{\exp V_1(s_t)} \\ \pi_0(a_t^i|s_t) &= \int \rho_0(a_t^{-i}|s_t, a_t^i) \pi_0(a_t^i|s_t) \ \mathrm{d}a_t^{-i} \end{aligned} \tag{107}$$

The level-2 is the only special case for us that we have to marginalize the policy and using the opponent prior as if it is independent of the agent's action. Given both cases, we are able to state the generalized

version for k level policy $\pi_k(a_t^i|a_t^{-i}, s_t)$, which is

$$\begin{aligned}
\pi_k(a_t^i|a_t^{-i}, s_t) &= \frac{\pi_{k-1}(a_t^i|s_t)\exp\left(Q_k(s_t, a_t^i, a_t^{-i})\right)}{\exp Q_k(s_t, a_t^{-i})} \\
\rho_k(a_t^{-i}|s_t) &= \frac{\rho_{k-1}(a_t^{-i}|s_t)\exp\left(Q_k(s_t, a_t^{-i})\right)}{\exp V_k(s_t)} \\
\pi_k(a_t^i|s_t) &= \int \pi_k(a_t^i|a_t^{-i}, s_t)\rho_k(a_t^{-i}|s_t)\ \mathrm{d}a_t^{-i} \\
Q_k(s_t, a_t^i, a_t^{-i}) &= R(s_t, a_t^i, a_t^{-i}) + \gamma\mathbb{E}_{s_{t+1}}\left[V_k(s_{t+1})\right]
\end{aligned} \tag{108}$$

Similarly, the generalized version for k level opponent model would be

$$\begin{aligned}
\rho_k(a_t^{-i}|a_t^i, s_t) &= \frac{\rho_{k-1}(a_t^{-i}|s_t)\exp\left(Q_k(s_t, a_t^i, a_t^{-i})\right)}{\exp Q_k(s_t, a_t^{-i})} \\
\pi_k(a_t^i|s_t) &= \frac{\pi_{k-1}(a_t^i|s_t)\exp\left(Q_k(s_t, a_t^{-i})\right)}{\exp V_k(s_t)} \\
\rho_k(a_t^{-i}|s_t) &= \int \rho_k(a_t^{-i}|a_t^i, s_t)\pi_k(a_t^i|s_t)\ \mathrm{d}a_t^i \\
Q_k(s_t, a_t^i, a_t^{-i}) &= R(s_t, a_t^i, a_t^{-i}) + \gamma\mathbb{E}_{s_{t+1}}\left[V_k(s_{t+1})\right]
\end{aligned} \tag{109}$$

Note that there are 3 observations that we made with our current model

1. The use of marginalized agent between our approach to the one proposed in [18]. As explain earlier, our choice of Q-function (the already optimized agent) already uses the assumption that the a so-called marginalized opponent's policy is "optimal" according its prior on agent i.e $\pi_{k-2}(a_t^i|a_t^{-i}, s_t)$, which reinforce the fact that $\pi_k(a_t^i|a_t^{-i}, s_t)$ should be superior as it also taking account the opponent's best response to its lower-level policy. We believe that our current policy/optimal model will already distill the knowledge of all the level before. Finally, we might need the lower level agent since level-k opponent model is model against

2. The training will be drastic difference from the one presented [18] that is because we will train level by level rather than fully trained all levels together. Please note that for all training till now, we only try to approximate the optimal agent. Furthermore, there will be multiple number of values functions that are used for each level of agent, which wouldn't be shared, as in our case, we taking the notion of "best response" to the truest sense. It can be confusing of how the action value function represents in [18] as the authors intended to force it to works across entire level without noticing that the the final value function implicitly consider the level-k policy and its opponent model actions rollout. Therefore, it is impossible to construct such a value function.

3. This is similar to EM-algorithm in the case that we reuse our old policy as prior. There are 2 training differences we have (we will ignore the objective):

   - In the case of EM we change the policy every time step of training (although if we delay the prior update we will have almost the same training behavior).
   - EM always based the training on one kind of agent i.e ROMMEO or PR2 but not doing it alternatively as described here

The process of the training is depicted in the following figure:

4.2 Probabilistic Public Belief

4.2.1 Public Belief MDP

As we know, the recursive reasoning can be very expensive. We haven't start our examination into how can we model recursive *belief* as we only consider how agent participates its opponent actions. We will increase

the complexity of the task a bit by considering the coopeartive setting, which enable us to use a public agent. By using public agent, we can reduce the problem into single agent POMDP task, which would be easier to manage. This technique is proposed in [13] and extended to scale over by approximation and neural network in [3].

Now, let's start on the formulation of public Belief MDP. The public belief MDP base on the fact that every agent can some how infer the underlying state (and other informations) given a public observation (can also be an action too). After forming such a belief, each agent can executes and action that condition on its private observation and the belief over the state. Given a decoupled process, we can have a centralized agent that only infer the underlying state and the distribute the common belief to each lower level agent to execute its action. The process is as follows:

- The higher-level agent observers a public observation that all the agents should see and construct a message $m_t \sim \pi_t^{\text{pub}}(\cdot|o_t^{\text{pub}}, a_t^{\text{pub}})$
- The lower-level agent after receieves the message from ther higher-level agent will execute its action based on its private observation and the message given $a_t^i \sim \pi^i(\cdot|m_t, o_t^{\text{pri}})$.

Please note that by having a centralized public belief agent makes the coordination problem much easier, and we shouldn't treat this as an decentralized learning (even if we have a common seed). In addition, in [3], the authors argue for a deterministic execution of lower-level agent, which should improve the coordination. Furthermore, we purposefully use the term "message" because it isn't entirely state but more than that as the higher-level agent have to carefully choose the message and not merely spill out the belief over state. Both of the agent and its higher-level counter part will try to optimize the same reward.

4.2.2 Solving Single agent POMDP via Variational Monte Carlo

Our work will be based on [8] and the interpretation via graphical model [16], which is similar to [6] but instead of having one state representation the authors consider the case where we try to improve the ELBO bound to be tighter by particle filtering. We will revisit the derivation and proposed minor extension by trying to the problem in closed form so that we are able to perform Soft-Actor critic like update.

4.2.3 Public Belief ELBO

Now, let's formulize the problem of public Belief into graphical model setting. We will follow more closely for [16] as we assume that the observation of the next step is depending on the action of the agents. The graphical model is depicted in figure 10. Let's start by defining our joint distribution for this:

$$P(s_0)\prod_{t=0}^{T} P(s_{t+1}|s_t, a_t^i, a_t^{-i})P(o_{t+1}^{\text{pri},i}|s_{t+1}, a_t^i)P(o_{t+1}^{\text{pri},-i}|s_{t+1}, a_t^{-i})P(o_{t+1}^{\text{pub}}|s_{t+1}, a_t^i, a_t^{-i})$$
$$P(a_t^i|s_t, o_t^{\text{pri},i}, m_t, o_t^{\text{pub}})P(a_t^{-i}|s_t, o_t^{\text{pri},-i}, m_t, o_t^{\text{pub}})P(\mathcal{O}_t|r_t)P(r_t|s_t, a_t^i, a_t^{-i})P(m_t|s_t) \tag{110}$$

There should be 2 steps for this, first the public agent see the public observation only, it updates its belief on the state, after this, the private agent uses its local information (and possibly) to further refine the state representation and execute the action. It is very clear from this representation that the if we simply doing single agent POMDP as we can only observe the public observation and execute the action by sending right "message" to the agents. Our plan of attack is as follows:

- We have to realized that the beauty of public agent lies in the fact that by having one message that coordinates every agents, we can ignore other agents' actions and even its observation as noted in [3] that the transition of PuB-MDP "depends not just on the executed action, but on the counterfactual actions". Now, we can simply have the understanding of other agents by simply looking at m_t.
- There are multiple cases that we might want to follows ranging from most restrictive to most relaxed:
  - We are given non-related public observation that is derived from the state. We will have to ignore how the actions are calculation and consider only $Q(s_t, m_t)$ and infer the hidden state given only

the public observation. This is typical single agent POMDP framework in which we are not aware of how the underlying dynamics works and we only aware of the reward given out.

  - If action is visible, then we have to consider how each agent's will response to our message. This information would be very important as we can use it to "understand" how other agent will work. Instead of modeling $Q(s_t, m_t)$ we will have to model the each agent's observation function, while assuming the same agent model. By doing this, we might be able to have better quality message that acts directly on agents' action.

- After getting message m_t, we can learn the representation of other agents by considering the following value function: $Q(a_t^i, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}})$ for any agent i given the reward (ignoring all other opponent action)

$$\pi(a_t^i | o_t^{\text{pri},i}, o_t^{\text{pub}}, m_t) = \frac{P_{\text{prior}}(a_t^i | o_t^{\text{pri},i}, o_t^{\text{pub}}, m_t) \exp\left(Q(a_t^i, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}})\right)}{\exp Q(m_t, o_t^{\text{pri},i}, o_t^{\text{pub}})} \tag{111}$$

  Or we can even trying to infer the underlying state, refining the original one done by public agent. We can imagine this as "knowing the location of one self" and given the message m_t we might be able to infer the location of others by inferring from our position, while the message m_t has the property of how the other agent's will acts. Summing up we have $Q(s_t, m_t, a_t)$ where we have

$$\pi(a_t^i | s_t, m_t) = \frac{P_{\text{prior}}(s_t, m_t, a_t) \exp\left(Q(s_t, m_t, a_t^i)\right)}{\exp Q(s_t, m_t)} \tag{112}$$

Now, let's walk-through 2 kinds of games: a card game, and the messenger game (where all the agent are in a grid world and heard the message coming by). Given our plan, let's consider how can we solve this:

- **Card Game**: In card games, usually the public belief consists of action of agents and possibly a public card that all agents can observe. The action of agents can also affects the public card. The state is the configuration of the cards for player and the private observation is simple a mask to the card removing opponent private card and allows us to see only our card. The public observation is a mask over the private observations with past actions.

- **Communication Game**: Now the state of the game is a whole map representation. We will treat message from the coordinator as part of the state, which will be masked to a public observation. Similarly, the private observation is simply a masked version of the state.

The problem becomes *would it be necessary for the agent to do further refine its state belief ? especially in the case of fully non-observable.* The effect we want for the agent (overall agent) is the ability to infer others action purely from the partial observability. For example, suppose one agent is in the room and then notice a change in its observation. We would like that agent to infer that the other agents might came to this room before. This leads to 2 questions:

- would the state representation on the public agent *enough* for our agent to use as a real state representation ?

- would the message m_t enough to convey other agent's action (which is a product of its private observation, m_t, and state s_t)

Now, we can see that if we assume the other agent's being like us (us in this case means the lower level agent) then it's a matter of perspective given the fully observable state. The method of archiving this is clear once we see the action as part of public knowledge as it forces us to infer the private observation given the state, which in the case of executing agent we can model this as "our own observation" (if our observation is a pixel world then we shall assume the others to be the same as us, although the others might be given a text. However the representation power should be the same or less than). The difference behavior comes purely from the augmentation of such observation, since the opponent's inputs are the same except its observation. One can partly see this problem as multi-view [10], which has been shown to works effectively in single agent reinforcement learning. However, in multi-view case we are also aware of the other's observation. In our

case, the only thing that are difference is the private observation. By this, we would *safely* assume/reassure that m_t is selected in such a way that will guide the opponent's action to achieve the common goal, and all of the counter-factual values calculation is done via difference particles of s_t^k when selecting the right m_t. *The only problem that remains what to do with the inferred observation ?* There are at least 2 properties that we have to consider:

- The quality of the private observation: We want to make sure that the agent action will be impacted the most by the private information, as it is essentially the only quantity that distinguish the action of agents.
- The interpretability of the observation. Since the public agent will have almost no idea about the observation of agents (there is some because we have a "model" of the agents, in this case, we only aware of the "size" of the vector corresponding to the agent but not the domain it is in). The observation can be a noisy image and still works. This problem is almost the same problem as "suppose the action is up, given the state what is the observation of the agent". Due to this nature, it is clear that the problem arises because agent itself isn't injective.

We can fix the first part by using mutual information regularization, where we maximize the mutual information between the representation of the observation and the output of the agent. This would force the hidden observation to be more useful to the agent.**[[Phu: Is the related to the MI regularization ?]]**

For the interpretability, we can see some problem with this. Since we want the public agent to have *absolutely zero* idea about the private observation of *all the agents*, while maintaining the same m_t between each agents. The best we can do is to trying to map this latent description of the private observation to the real private observation of the agent. The best we can do is to simply independently train a decoder that decodes the latent description to the real agent observation, which we can now choose to continue using latent description or to use the decoded observation. similar to the opponent model. By this way, we have achieve "model other as ones own". However, it is crucial to see that the way public agent is trained should strictly be the same and we can't reinforce a gradient of the decoder back to the public agent.

Public Agent (No knowledge of action): In this case, we consider a more simpler model of agents, which is almost the same with normal POMDP framework in the case where the optimality of public agent is

$$\begin{aligned} P(\mathcal{O}_t, r_t | s_t, m_t) &= \exp\left(\beta R(s_t, m_t)\right) \\ \text{where } R(s_t, m_t) &= \mathbb{E}_{a_t^i \sim \pi(a_t^i | m_t, s_t, o_t^{\text{pri},i}, o_t^{\text{pub},i})}\left[R(s_t, a_t^i, a_t^{-i})\right] \end{aligned} \tag{113}$$

The reward can be calculated by simply rollout where we assume not knowing the underlying dynamics as normal reinforcement learning case. There would be no harm to it as it is a single agent case. However, this might not be interesting enough.

Public Agent (Action as part of Public Knowledge): Now, we have move the the part where we can consider the action to be communicable. Now consider the joint probability, we had, which is in equation 110. Now, consider the variational distribution:

$$\begin{aligned} &P(s_0)\left(\prod_{t=0}^{T} q(s_{t+1}|s_t, a_t^i, a_t^{-i}, o_{t+1}^{\text{pub}}, m_t, \mathcal{O}_t, r_t)\right)\left(\prod_{t=0}^{T} q(o_{t+1}^{\text{pri},i}|s_t, a_t^i, o_{t+1}^{\text{pub}}, m_t, \mathcal{O}_t, r_t)\right) \\ &\left(\prod_{t=0}^{T} q(o_{t+1}^{\text{pri},-i}|s_t, a_t^{-i}, o_{t+1}^{\text{pub}}, m_t, \mathcal{O}_t, r_t)\right)\left(\prod_{t=0}^{T} q(m_t|o_{1:t}^{\text{pub}}, a_{1:t}^i, a_{1:t}^{-i}, \mathcal{O}_{1:T}, r_{1:T})\right) \end{aligned} \tag{114}$$

Let's consider the ELBO for the problem. We have the following:

$$\begin{aligned} \mathbb{E}\Bigg[\sum_{t=1}^{T} \log &\frac{P(s_{t+1}|s_t, a_t^i, a_t^{-i}) P(o_{t+1}^{\text{pri},i}|s_{t+1}, a_t^i) P(o_{t+1}^{\text{pri},-i}|s_{t+1}, a_t^{-i}) P(o_{t+1}^{\text{pub}}|s_{t+1}, a_t^i, a_t^{-i}) P(a_t^i|s_t, o_t^{\text{pri},i}, m_t, o_t^{\text{pub}})}{q(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}|s_t, a_t^i, a_t^{-i}, o_{t+1}^{\text{pub}}, m_t, \mathcal{O}_t, r_t)} \\ &\cdot \frac{P(m_t|s_t) P(\mathcal{O}_t, r_t|r_t) P(r_t|s_t, a_t^i, a_t^{-i}) P(a_t^{-i}|s_t, o_t^{\text{pri},-i}, m_t, o_t^{\text{pub}})}{q(m_t|o_{1:t}^{\text{pub}}, a_{1:t}^i, a_{1:t}^{-i}, \mathcal{O}_{1:T}, r_{1:T})}\Bigg] \end{aligned} \tag{115}$$

Now, we want to infer the unknown given what we can observe:

$$P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t | a_{1:t}^i, a_{1:t}^{-i}, \mathcal{O}_{1:T}, r_{1:T}, o_{1:t+1}^{\text{pub}}) \\ = \frac{P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t, a_t^i, a_t^{-i}, \mathcal{O}_t, r_t, o_{t+1}^{\text{pub}} | a_{1:t-1}^i, a_{1:t-1}^{-i}, \mathcal{O}_{1:t-1}, r_{1:t-1}, o_{1:t}^{\text{pub}})}{P(a_t^i, a_t^{-i}, \mathcal{O}_t, r_t, o_{t+1}^{\text{pub}} | a_{1:t-1}^i, a_{1:t-1}^{-i}, \mathcal{O}_{1:t-1}, r_{1:t-1}, o_{1:t}^{\text{pub}})} \tag{116}$$

This quantity can be infer recursively, as follows fron the pattern, we shall consider the numerator as follows, since we can see that the current is conditional independent given the current state.

$$P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t, a_t^i, a_t^{-i}, \mathcal{O}_t, r_t, o_{t+1}^{\text{pub}} | a_{1:t-1}^i, a_{1:t-1}^{-i}, \mathcal{O}_{1:t-1}, r_{1:t-1}, o_{1:t}^{\text{pub}}) \\ = \int P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t, a_t^i, a_t^{-i}, \mathcal{O}_t, r_t, o_{t+1}^{\text{pub}} | s_t) P(s_t | a_{1:t-1}^i, a_{1:t-1}^{-i}, \mathcal{O}_{1:t-1}, r_{1:t-1}, o_{1:t}^{\text{pub}}) \, \mathrm{d}s_t \tag{117}$$

Let's perform the importance sampling using the proposal distribution or the variational distribution:

$$\int P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t, a_t^i, a_t^{-i}, \mathcal{O}_t, r_t, o_{t+1}^{\text{pub}} | s_t) P(s_t | a_{1:t-1}^i, a_{1:t-1}^{-i}, \mathcal{O}_{1:t-1}, r_{1:t-1}, o_{1:t}^{\text{pub}}) \, \mathrm{d}s_t \\ = \int \frac{P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t, a_t^i, a_t^{-i}, \mathcal{O}_t, r_t, o_{t+1}^{\text{pub}} | s_t)}{q(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t)} q(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t) \\ P(s_t | a_{1:t-1}^i, a_{1:t-1}^{-i}, \mathcal{O}_{1:t-1}, r_{1:t-1}, o_{1:t}^{\text{pub}}) \, \mathrm{d}s_t \tag{118}$$

We will denote the importance weight $w_{t+1}(s_t, s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t)$ as

$$w_{t+1}(s_t, s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t) = \frac{P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t, a_t^i, a_t^{-i}, \mathcal{O}_t, r_t, o_{t+1}^{\text{pub}} | s_t)}{q(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t)} \tag{119}$$

Now given the following importance sampling, we can see that by initially sampling the state s_t by ancestor index u_t^k. Let's approximate this using ancestor index type sampling:

$$\frac{1}{K} \sum_{k=1}^{K} w_{t+1}(s_t^{u_t^k}, s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t) \cdot q(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t | s_t^{u_t^k}, a_t^i, a_t^{-i}, o_{t+1}^{\text{pub}}) \\ = \frac{1}{K} \sum_{k=1}^{K} w_{t+1}(s_t^{u_t^k}, s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t) q(s_{t+1} | s_t^{u_t^k}, a_t^i, a_t^{-i}, o_{t+1}^{\text{pub}}) q(o_{t+1}^{\text{pri},i} | s_t^{u_t^k}, a_t^i, m_t, \mathcal{O}_{1:T}, r_{1:T}, o_{t+1}^{\text{pub}}) \\ q(o_{t+1}^{\text{pri},-i} | s_t^{u_t^k}, a_t^{-i}, m_t, \mathcal{O}_{1:T}, r_{1:T}, o_{t+1}^{\text{pub}}) q(m_t | o_{1:t}^{\text{pub}}, a_{1:t}^{-i}, a_{1:t}^i, \mathcal{O}_{1:T}, r_{1:T}) \\ \approx \frac{1}{K} \sum_{k=1}^{K} w_{t+1}(s_t^{u_t^k}, s_{t+1}^k, o_{t+1}^{\text{pri},i,k}, o_{t+1}^{\text{pri},-i,k}, m_t^1) \delta(s_{t+1} - s_{t+1}^k) \delta(o_{t+1}^{\text{pri},i} - o_{t+1}^{\text{pri},i,k}) \\ \delta(o_{t+1}^{\text{pri},-i} - o_{t+1}^{\text{pri},-i,k}) \delta(m_t - m_t^1) \tag{120}$$

We can see above that we can further approximate our equation using sampling from our proposal/variational distribution. Now, we can fine the denominator by marginalization. Given this we have the final posterior being weighted mixture of particles ($P(s_{t+1}, o_{t+1}^{\text{pri},i}, o_{t+1}^{\text{pri},-i}, m_t | a_{1:t}^i, a_{1:t}^{-i}, \mathcal{O}_{1:T}, r_{1:T}, o_{1:t+1}^{\text{pub}})$):

$$\frac{1}{K} \sum_{k=1}^{K} W_{t+1}^k \delta(s_{t+1} - s_{t+1}^k) \delta(o_{t+1}^{\text{pri},i} - o_{t+1}^{\text{pri},i,k}) \delta(o_{t+1}^{\text{pri},-i} - o_{t+1}^{\text{pri},-i,k}) \delta(m_t - m_t^1) \tag{121}$$

where the W_{t+1}^k is the re-normalized weight. Now, the probability over the state is simply:

$$P(s_{t+1} | a_{1:t}^i, a_{1:t}^{-i}, \mathcal{O}_{1:T}, r_{1:T}, o_{1:t+1}^{\text{pub}}) = \frac{1}{K} \sum_{k=1}^{K} W_{t+1}^k \delta(s_{t+1} - s_{t+1}^k) \tag{122}$$

which we can further sample like the one above. Now, since the probability of the observable variable is:

$$P(a^i_{1:\tau}, a^{-i}_{1:\tau}, \mathcal{O}_{1:\tau}, r_{1:\tau} o^{\text{pub}}_{1:\tau}) = \prod_{t=0}^{\tau} \frac{1}{K} \sum_{k=1}^{K} w_{t+1}(s_t^{u_t^k}, s^k_{t+1}, o^{\text{pri},i,k}_{t+1}, o^{\text{pri},-i,k}_{t+1}, m^1_t) \tag{123}$$

which is a tighter variational lower bound. Now, the objective becomes:

$$\mathbb{E}\left[\sum_{t=0}^{\tau} \log \frac{1}{K} \sum_{k=1}^{K} w_{t+1}(s_t^{u_t^k}, s^k_{t+1}, o^{\text{pri},i,k}_{t+1}, o^{\text{pri},-i,k}_{t+1}, m^1_t)\right] \tag{124}$$

Let's recall the definition of the importance weight with suitable indexes:

$$\begin{aligned}
&\frac{P(s^k_{t+1}|s_t^{u_t^k}, a^i_t, a^{-i}_t) P(o^{\text{pri},i,k}_{t+1}|s^k_{t+1}, a^i_t) P(o^{\text{pri},-i,k}_{t+1}|s^k_{t+1}, a^{-i}_t) P(o^{\text{pub}}_{t+1}|s^k_{t+1}, a^i_t, a^{-i}_t)}{q(s^k_{t+1}|s_t^{u_t^k}, a^i_t, a^{-i}_t, o^{\text{pub}}_{t+1}, m_t, \mathcal{O}_t, r_t)} \\
\times &\frac{P(a^i_t, a^{-i}_t|s_t^{u_t^k}, o^{\text{pri},i}_t, o^{\text{pri},-i}_t, m_t, o^{\text{pub}}_t) P(m_t|s_t^{u_t^k}) P(\mathcal{O}_t|r_t)}{q(o^{\text{pri},i,k}_{t+1}, o^{\text{pri},-i,k}_{t+1}|s_t^{u_t^k}, a^i_t, a^{-i}_t, m_t, o^{\text{pub}}_{t+1}, \mathcal{O}_t, r_t)} \\
\times &\frac{P(\mathcal{O}_t|r_t)}{q(m_t|o^{\text{pub}}_{1:t}, a^i_{1:t}, a^{-i}_{1:t}, \mathcal{O}_{1:T}, r_{1:T})}
\end{aligned} \tag{125}$$

Since the second terms doesn't correspond to the indexing of the particle (hence we separate the reward from the optimality), we have the following objective that follows maximum entropy framework:

$$\begin{aligned}
\mathbb{E}\Bigg[&\sum_{t=0}^{\tau} \log P(\mathcal{O}_t|r_t) + \mathbb{H}\left[q(m_t|o^{\text{pub}}_{1:t}, a^i_{1:t}, a^{-i}_{1:t}, \mathcal{O}_{1:T}, r_{1:T})\right] \\
&+ \log \frac{1}{K} \sum_{k=1}^{K} \Bigg(\frac{P(s^k_{t+1}|s_t^{u_t^k}, a^i_t, a^{-i}_t) P(o^{\text{pri},i,k}_{t+1}|s^k_{t+1}, a^i_t) P(o^{\text{pri},-i,k}_{t+1}|s^k_{t+1}, a^{-i}_t) P(o^{\text{pub}}_{t+1}|s^k_{t+1}, a^i_t, a^{-i}_t)}{q(s^k_{t+1}|s_t^{u_t^k}, a^i_t, a^{-i}_t, o^{\text{pub}}_{t+1}, m_t, \mathcal{O}_t, r_t)} \\
&\cdot \frac{P(a^i_t, a^{-i}_t|s_t^{u_t^k}, o^{\text{pri},i,k}_t, o^{\text{pri},-i,k}_t, m_t, o^{\text{pub}}_t) P(m_t|s_t^{u_t^k}) P(\mathcal{O}_t|r_t)}{q(o^{\text{pri},i,k}_{t+1}, o^{\text{pri},-i,k}_{t+1}|s_t^{u_t^k}, a^i_t, a^{-i}_t, m_t, o^{\text{pub}}_{t+1}, \mathcal{O}_t, r_t)} \Bigg)\Bigg]
\end{aligned} \tag{126}$$

Now, in practice the public agent shall be a weighted sum of underlying state, public observation, and agent actions i.e

$$\sum_{k=1}^{K} W^k_t \pi(m^k_t|s^k_t, o^{\text{pub}}_t, a^i_t, a^{-i}_t) \tag{127}$$

There are 3 points that we want to consider (that we have mentioned above):

- The interpretability of $o^{\text{pri},i,k}_t$ and $o^{\text{pri},-i,k}_t$. There isn't a lot we can do with this apart from constructing a network that would translate the public agent's private observation into the the agent's private observation. Or we can pretrain a network that output the agent's private observation. However, we have to make sure for the sake of decentralized training to not change the main public agent's network by backpropagate the local feature.

- The joint action $P(a^i_t, a^{-i}_t|s_t^{u_t^k}, o^{\text{pri},i}_t, o^{\text{pri},-i}_t, m_t, o^{\text{pub}}_t)$. After having all the private observation and public observation, we can trying to do the reconstruction loss on the agents' actions. However, we can further assume that the agents' policy are based on boltzmann distribution of action value function. Given this assumption, the maximum-log-likelihood estimation should follows the usual adversarial imitation learning that we have extensively discussed in chapter before (in the special case of partially difference goal). For a normal scenario, we simply able to train its own Q function. As noted before, we shouldn't include any additional information from the agent to help us with the public agent (including agent's own action value function).

- The relation between the agent action and its private observation. As mention before, one of our concern is on the quality of the private observation of the agent, which might lead to a total ignoring of this important factor. We, therefore, need to have a regularization that forces the agent model to takes more attention to the private observation and improves the quality of private observation representation. We can achieve this by consider mutual information regularization via contrastive loss [17] by trying to distinguish between the positive pair of $\{a_t^i, o_t^{\text{pri},i}\}$ and the negative pair $\{a_t^i, o_t^{\text{pri},-i}\}$

$$\mathcal{L}_{\text{contrast}} = -\mathbb{E}_{(a_1^i, o_1^{\text{pri},i}),\cdots,(a_N^i, o_N^{\text{pri},-i})}\left[\log \frac{h_\theta(a_1^i, o_1^{\text{pri},i})}{\sum_{n=2}^N h_\theta(a_n^i, o_n^{\text{pri},-i})}\right] \tag{128}$$

  We didn't decide to use contrastive loss between each observation since there exists some cases where the private observation are the same and we still wants them to be almost the same. We only want to emphasize that vastly difference private observation leads agent to difference actions. **[[Phu: Check on MI again.]]**

Finally, we would like to say that all the problem above can be lifted if we assume that each agent owns its own public agent and trying to model the others using oneself or other methods. By ignoring the common knowledge of messages m_t, we lose a guarantee that the agents will be in "sync" of each other thus returning to a problem of equilibrium selection.

Now, we have lifted the heavy part of the problem. Now, we shall consider the individual agent that corresponds on the message and its own private observation. We will define the following value function $Q(s_t, a_t^i, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}})$, which the agent is simply a weighted sum:

$$\begin{aligned}
\pi(a_t^i | s_t, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}}) &= \sum_{k=1}^K W_t^k \frac{P_{\text{prior}}(a_t^i | s_t^k, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}}) \exp Q(s_t^k, a_t^i, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}})}{Q(s_t^k, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}})} \\
\text{where } \sum_{k=1}^K W_t^k Q(s_t^k, a_t^i, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}}) &= r_t + \gamma\left(\sum_{k=1}^K W_t^k Q(s_{t+1}^k, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}})\right) \\
V(s_t, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}}) &= \int P_{\text{prior}}(a_t^i | s_t^k, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}}) \exp Q(s_t, a_t^i, m_t, o_t^{\text{pri},i}, o_t^{\text{pub}}) \, \mathrm{d}a_t^i
\end{aligned} \tag{129}$$

4.3 Variational Sequential Monte-Carlo I-POMDP

5 Muti-Agent imitation Learning

It would be impossible for any agent to act without having any knowledge of how others objectives are. There has been lots of studies on imitation learning and inverse reinforcement learning both in single agent and multi-agent. We will start with interpolating some of the single agent techniques (such as f-divergence minimization and Bayesian adversarial imitation learning) to multi-agent domain. But these concepts will not be utilized to their fullest potential if we didn't include cooperate this into training and execution.

We will start by understand how to achieve such a thing. The result of this comes from the very fact that in order to derive the agent, we have to find the solution for another - a common theme in this work. Knowing this fact, we start deriving basic context aware agent, in which we try to infer the other agent's "mode of behavior" using VAE like objective. Given this, we extends to adversarial setting, which leads to the most basic form. In the second section, we will survey possible extensions to this adversarial approach including works in GAN literature and imitation learning, inviting researchers to consider more on the contribution of one field to another.

5.1 Imitation learning for action execution

5.1.1 Propose of Inference Model

In this section, we will consider the use of initiation learning of the opponent model in order to inform how the agent acts. As we observe in earlier sections, we can see that we should consider the interaction of agents

to be correlated together as proposed in [11]. However, the authors doesn't consider how the training of the agent or its opponent takes place, while only use supervised learning to "approximate" the policies. Our proposal to consider normal opponent adversarial learning, while as a by product able to arrive at the agent's policy that "best response" to the model. This comes from the observation that we can't calculated $\rho(a^{-i}|s)$ without implicitly derive $\pi(a^i|s, a^{-i})$. Furthermore, in order for us to fully utilized the model, we shall assume context variable z that characterizes the opponent's reward (Note that the agent's reward/action is known + we shall assume that the opponent's data is collected while playing the game) as shown in single agent case [19, 15]. Now, we would like to consider the variatioanl auto-encoder like architecture with context variable m as the opponent's reward representation. Let's consider the variational auto-encoding see figure 11. We can see that the by doing Variational EM leads to an VAE objective, where we have a variational distribution $q_\phi(m|\tau)$ i.e

$$\mathcal{L}(\theta, \psi) = \mathbb{E}_{q_\psi(m|\tau)} \left[\log P_\theta(\tau|m)\right] - D_{\mathrm{KL}} \left(q_\psi(m|\tau) \middle\| P(m)\right) \tag{130}$$

Note that we can represent $P_\theta(\tau|m)$ to be as the following:

$$\begin{aligned}
\rho_\phi(a_t^{-i}|s_t; m) &= \frac{P_{\text{prior}}(a_t^{-i}|s_t) \exp\left(Q_\theta^{-i}(s_t, a_t^{-i}; m)\right)}{\exp V_\theta^{-i}(s_t; m)} \\
\text{where } Q_\theta^{-i}(s_t, a_t^{-i}; m) &= \log \int P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \exp\left(Q_\theta^{-i}(s_t, a_t^i, a_t^{-i}; m)\right) \, \mathrm{d}a_t^i \\
Q_\theta^{-i}(s_t, a_t^i, a_t^{-i}; m) &= r^{-i}(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s'} \left[V_\theta^{-i}(s'; m)\right] \\
V_\theta^{-i}(s_t; m) &= \log \int P_{\text{prior}}(a_t^{-i}|s_t) \exp\left(Q_\theta^{-i}(s_t, a_t^{-i}; m)\right) \, \mathrm{d}a_t^{-i}
\end{aligned} \tag{131}$$

We can replace the $\pi(a_t^i|s_t, a_t^{-i})$ if we would like to assume that the opponent has almost perfect knowledge about the agent (which isn't the usual case). Furthermore, the Bellman function should actually include the real agent's action as it is how the environment progress. We will call this model opponent-context imitation learning. On the other hand, if we made assumption that the opponent's reward is proportional to the agent, similar to Balancing Q case, which will make m to be reduced to 1-dimension. We will have more interesting case, there are direct connection between agent's action value function and oppoent's action value function, or we are solving,

$$\begin{aligned}
\rho_\phi(a_t^{-i}|s_t) &= \frac{P_{\text{prior}}(a_t^{-i}|s_t) \exp\left(\frac{\beta^{-i}}{\beta^i} Q_\theta^i(s_t, a_t^{-i})\right)}{\exp V_\theta^{-i}(s_t)} \\
\text{where } Q_\theta^i(s_t, a_t^{-i}) &= \log \int P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \exp\left(Q_\theta^i(s_t, a_t^i, a_t^{-i})\right) \, \mathrm{d}a_t^i \\
Q_\theta^i(s_t, a_t^i, a_t^{-i}) &= \beta^i(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s'} \left[\frac{\beta^i}{\beta^{-i}} V_\theta^{-i}(s')\right] \\
V_\theta^{-i}(s_t) &= \log \int P_{\text{prior}}(a_t^{-i}|s_t) \exp\left(\frac{\beta^{-i}}{\beta^i} Q_\theta^i(s_t, a_t^{-i})\right) \, \mathrm{d}a_t^{-i}
\end{aligned} \tag{132}$$

where the By-product of the calculation is agent's policy

$$\pi(a_t^i|s_t, a_t^{-i}) = \frac{P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \exp\left(Q_\theta^i(s_t, a_t^i, a_t^{-i})\right)}{\exp Q_\theta^i(s_t, a_t^{-i})} \tag{133}$$

As now, we have the method to infer either full context variable or β^{-i}. We will call this model: Balance-Imitation Learning (from Balancing-Model).

5.1.2 From VAE to Discriminator for imitation learning

Given 2 models, we shall consider scalable implementation of this, as we will borrow method from [2]. We will consider the fact that VAE model is simply a log-likelihood with KL-regularization, as the KL-regularization

can be changed to be mutual information regularization also as proposed in [19]; we will deal with is in later part of this section. We will focus on the terms $\mathbb{E}_{q_\psi(m|\tau)}\left[\log P_\theta(\tau|m)\right]$. Let's see how it rollouts. Furthermore, we shall assume that

$$\begin{aligned}
\log P_\theta(\tau|m) &= \sum_{n=1}^{T} \log P(s_n, a_n^{-i}|m) = \sum_{n=1}^{T} \log P(s_n) + \log \rho_\phi(a_n^{-i}|s_m; m) \\
&= \sum_{n=1}^{T} \log P(s_n) + \log \frac{P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)}{\int P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)\ \mathrm{d}a_n^{-i}} \\
&= \text{const} + \sum_{n=1}^{T} \log \left[P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)\right] - \log \left[\mathbb{E}_{P_{\text{prior}}(a_n^{-i}|s_n)}\left[Q_\theta^{-i}(s_n, a_n^{-i}; m)\right]\right] \\
&= \sum_{n=1}^{T} \log \left[P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)\right] + \log \mathbb{E}_\mu \left[\frac{P_{\text{prior}}(a_n^{-i}|s_n) Q_\theta^{-i}(s_n, a_n^{-i}; m)}{\frac{1}{2}\tilde{p}(a_n^{-i}|s_n) + \frac{1}{2}\rho_\phi(a_n^{-i}|s_n; m)}\right]
\end{aligned}$$

The last equation will ignore the the constant and apply importance sampling shuffling between approximate distribution $\tilde{p}(a_n^{-i}|s_n)$ and variational distribution where $\mu = \frac{1}{2}\tilde{p}(a_n^{-i}|s_n) + \frac{1}{2}\rho_\phi(a_n^{-i}|s_t; m)$. Please consider that $\rho_\phi(a_n^{-i}|s_n; m)$ might not always be the same as our closed form. Now, given this let's move to the definition of optimal discriminator:

$$D_\theta(s_n, a_n^{-i}; m) = \frac{P_\theta(a_n^{-i}|s_n; n)}{P_\theta(a_n^{-i}|s_n; m) + \rho_\phi(a_n^{-i}|s_n)} = \frac{\frac{1}{Z} P_{\text{prior}}(a_t^{-i}|s_t) \exp\left(Q_\theta^{-i}(s_t, a_t^{-i}; m)\right)}{\frac{1}{Z} P_{\text{prior}}(a_t^{-i}|s_t) \exp\left(Q_\theta^{-i}(s_t, a_t^{-i}; m)\right) + \rho_\phi(a_n^{-i}|s_n)} \tag{134}$$

Now, let's see how the generator loss (that we have to maximize) be as we can show that it would be the same as the KL-regularized objective for opponent model (we want to max the discriminator when it is agreeing with our result and want to min the discriminator when it isn't):

$$\begin{aligned}
\mathcal{L}_{\text{gen}}(\phi) &= \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)}\left[\log D_\theta(s_n, a_n^{-i}; m)\right] - \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)}\left[\log(1 - D_\theta(s_n, a_n^{-i}; m))\right] \\
&= \quad \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)}\left[\log \frac{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)}{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) + \rho_\phi(a_n^{-i}|s_n)}\right] \\
&\quad -\mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)}\left[\log \frac{\rho_\phi(a_n^{-i}|s_n)}{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) + \rho_\phi(a_n^{-i}|s_n)}\right] \\
&= \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)}\left[\log \frac{1}{Z} + \log P_{\text{prior}}(a_n^{-i}|s_n) + Q_\theta^{-i}(s_n, a_n^{-i}; m) - \log \rho_\phi(a_n^{-i}|s_n)\right] \\
&= \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)}\left[Q_\theta^{-i}(s_n, a_n^{-i}; m) - \log \frac{\rho_\phi(a_n^{-i}|s_n)}{P_{\text{prior}}(a_n^{-i}|s_n)}\right] - \mathbb{E}\left[Z\right]
\end{aligned} \tag{135}$$

Now, let's consider the discriminate loss, which we can show that it is equivalent to the log-likelihood approach with respective to derivative over variable θ. Starting with the definition of discriminator loss

$$\mathcal{L}_{\text{dis}}(\theta) = \mathbb{E}_{P(s_n)P(a_n^{-i}|s_n;m)}\left[\log D_\theta(s_n, a_n^{-i}; m)\right] + \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n;m)}\left[\log\left(1 - D_\theta(s_n, a_n^{-i}; m)\right)\right] \tag{136}$$

where the log likelihood is equal to, furthermore we have the distribution $\tilde{p}(a_n^{-i}|s_n)$ to be the real sampling distribution i.e

$$\begin{aligned}
\mathcal{L}_{\text{cost}}(\theta) =& \log \left[P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)\right] \\
&+ \log \mathbb{E}_\mu \left[\frac{P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)}{\frac{1}{2Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) + \frac{1}{2}\rho_\phi(a_n^{-i}|s_n; m)}\right]
\end{aligned} \tag{137}$$

Let's consider the derivation of the cost function, we can see that

$$\begin{aligned}\partial_\theta \mathcal{L}_{\text{cost}}(\theta) &= \partial_\theta \left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) + \\ &\quad + \mathbb{E}_\mu \left[\frac{P_{\text{prior}}(a_n^{-i}|s_n)\partial_\theta \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)}{\tilde{\mu}_\theta(s_n, a_n)}\right] \cdot \mathbb{E}_\mu \left[\frac{\tilde{\mu}_\theta(s_n, a_n)}{P_{\text{prior}}(a_n^{-i}|s_n) Q_\theta^{-i}(s_n, a_n^{-i}; m)}\right] \\ &= \partial_\theta \left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) + \mathbb{E}_\mu \left[\frac{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) \partial_\theta Q_\theta^{-i}(s_n, a_n^{-i}; m)}{\tilde{\mu}_\theta(s_n, a_n)}\right]\end{aligned} \tag{138}$$

We denote $\tilde{\mu}_\theta(s_n, a_n) = \frac{1}{2Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) + \frac{1}{2}\rho_\phi(a_n^{-i}|s_n; m)$. The author shown that we have to treat $\tilde{\mu}_\theta(s_n, a_n)$ to be constant because we don't want to update gradient based on the importance weight. Now, we shall show that the derivation is equal to discriminator:

$$\begin{aligned}\mathcal{L}_{\text{dis}}(\theta) &= \mathbb{E}_{P(s_n)P(a_n^{-i}|s_n)} \left[\log D_\theta(s_n, a_n^{-i}; m)\right] + \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)} \left[\log\left(1 - D_\theta(s_n, a_n^{-i}; m)\right)\right] \\ &= \mathbb{E}_{P(s_n)P(a_n^{-i}|s_n)} \left[\log \frac{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)}{2\tilde{\mu}_\theta(s_n, a_n)}\right] + \mathbb{E}_{P(s_n)\rho_\phi(a_n^{-i}|s_n)} \left[\log \frac{\rho_\phi(a_n^{-i}|s_n)}{2\tilde{\mu}_\theta(s_n, a_n)}\right] \\ &= -\mathbb{E}_P \left[\log Z\right] + \mathbb{E}_P \left[P_{\text{prior}}(a_n^{-i}|s_n)\right] + \mathbb{E}_P \left[Q_\theta^{-i}(s_n, a_n^{-i}; m)\right] - 4\mathbb{E}_\mu \left[\log \tilde{\mu}_\theta(s_n, a_n)\right] + \mathbb{E}_\rho \left[\log \rho_\phi(a_n^{-i}|s_n)\right]\end{aligned} \tag{139}$$

Consider the derivative of the the first and third term (now we will have to differentiate $\tilde{\mu}_\theta(s_n, a_n)$)

$$\partial_\theta \mathcal{L}_{\text{dis}}(\theta) = \mathbb{E}_P \left[\partial_\theta Q_\theta^{-i}(s_n, a_n^{-i}; m)\right] - 4\partial_\theta \mathbb{E}_\mu \left[\log \tilde{\mu}_\theta(s_n, a_n)\right] \tag{140}$$

We can see that this is similar to the derivation cost function with minor difference in the multiplication in the second term. Now, we have establish the connection between the MLE to Adversarial training. Now, given what we know let's consider the training algorithm, as we shown that the log-likelihood is the same as discriminator, we have

$$\begin{aligned}\min_\phi \max_{\theta,\psi} \; &\mathbb{E}_{P_\psi(m|\tau)P(s_n)P(a_n^{-i}|s_n;m)} \left[\sum_{n=1}^N \log D_\theta(s_n, a_n^{-i}; m)\right] \\ &+ \mathbb{E}_{P_\psi(m|\tau)P(s_n)\rho_\phi(a_n^{-i}|s_n;m)} \left[\sum_{n=1}^N \log\left(1 - D_\theta(s_n, a_n^{-i}; m)\right)\right] \\ &- D_{\text{KL}}\left(q_\psi(m|\tau) \middle\| P(m)\right) \\ \text{where } D_\theta(s_n, a_n^{-i}; m) &= \frac{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right)}{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^{-i}; m)\right) + \rho_\phi(a_n^{-i}|s_n; m)}\end{aligned} \tag{141}$$

The first max is the VAE objective and the second min corresponds to the adversarial objective. Note that this also works for Balancing model with minor changes. Note that everything is based on the fact that we are aware of the agent itself. Now, we can model the agent by the following:

$$\begin{aligned}\pi(a_t^i|a_t^{-i}, s_t; m) &= \frac{\exp\left(Q^i(s_t, a_t^i, a_t^{-i}; m))\right) P_{\text{prior}}(a_t^i|a_t^{-i}, s_t)}{\exp(Q^i(s_t, a_t^{-i}; m))} \\ \text{where } Q^i(s_t, a_t^i, a_t^{-i}; m) &= R^i(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}, a_{t+1}^{-i} \sim P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1}, \mathcal{O}_{t+1}^{-i}=1; m)} \left[Q^i(s_{t+1}, a_{t+1}^{-i})\right] \\ Q^i(s_t, a_t^{-i}) &= \log \int \exp\left(Q^i(s_t, a_t^i, a_t^{-i}; m))\right) P_{\text{prior}}(a_t^i|a_t^{-i}, s_t) \; \mathrm{d}a_t^i\end{aligned} \tag{142}$$

where we have $P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1}, \mathcal{O}_{t+1}^{-i} = 1; m)$ to be our opponent model $\rho(a_{t+1}^{-i}|s_{t+1}; m)$, which we can't find it without having access to the model. This is what we want because now, we can infer the context variable m and then able to transform the agent without even considering the opponent model (since it is intrinsically embedded) or we can have the model to "predict" the opponent action that is also fine too in practice. On

the other hand, one can have PR2 counter part, which we assume that the opponent model reacts based on agent's action, which lead to the following discriminator (where we can think that a_t^i, s_t is one special state)

$$D_\theta(s_n, a_n^i, a_n^{-i}; m) = \frac{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|a_n^i, s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^i, a_n^{-i}; m)\right)}{\frac{1}{Z} P_{\text{prior}}(a_n^{-i}|a_n^i, s_n) \exp\left(Q_\theta^{-i}(s_n, a_n^i, a_n^{-i}; m)\right) + \rho_\phi(a_n^{-i}|a_n^i, s_n; m)} \tag{143}$$

Now, we can see that the opponent model can be equal to

$$\begin{aligned}
\pi_\phi(a_t^i|s_t; m) &= \frac{\exp\left(Q^i(s_t, a_t^i; m)\right) P_{\text{prior}}(a_t^i|s_t)}{\exp\left(V^i(s_t; m)\right)} \\
\text{where } V^i(s_t; m) &= \log \int \exp\left(Q^i(s_t, a_t^i; m)\right) P_{\text{prior}}(a_t^i|s_t) \, \mathrm{d}a_t^i \\
Q^i(s_t, a_t^i; m) &= \log \int \exp\left(Q^i(s_t, a_t^i, a_t^{-i}; m)\right) \rho_\phi(a_n^{-i}|a_n^i, s_n; m) \, \mathrm{d}a_t^{-i} \\
Q^i(s_t, a_t^i, a_t^{-i}; m) &= r^i(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}}\left[V^i(s_{t+1}; m)\right]
\end{aligned} \tag{144}$$

Note that we could have take the prior we have over the opponent, however, there should not be a problem if we want to gain and advantage over the opponent as we want to "win" over the current opponent not to train to get consensus with the opponent.

Finally, Note that there are significant differences between our proposal and one in [11]. We are more concerned on how to correctly training a policy that is able to adapt given the unknown environment via variational inference, while [11] substitute unknown component with supervised learning. We will treat the objective for opponent model as a basis for extensions, where we can improve this in 2 ways: Improving the VAE objective (by using mutual information or adversarial on latent variable) or improving the adversarial training for opponent model discriminator (f-GAN, IWGAN, Baysian imitation learning), which we will explore in later sections. This come from the fact that since we are aware of the agent, we effectively reduce our problem into single agent problem, which we can use the works done in single agent domain to solve our problem. We will survey some of the interesting results in the sections below.

5.2 Extension to Discriminator

Now we have establish the practical method of inference for understanding the opponent's behavior drawn from the current development of adversarial imitations learning and assumption on how joint action is factorized. We are now presenting multiple extensions to our model based on well established work on single-agent adversarial imitation learning and GANs as we will focus on the adversarial term i.e

$$\begin{aligned}
\min_\phi \max_{\theta,\psi} \mathbb{E}_{P_\psi(m|\tau)P(s_n)P(a_n^{-i}|s_n;m)} &\left[\sum_{n=1}^N \log D_\theta(s_n, a_n^{-i}; m)\right] \\
+ \mathbb{E}_{P_\psi(m|\tau)P(s_n)\rho_\phi(a_n^{-i}|s_n;m)} &\left[\sum_{n=1}^N \log\left(1 - D_\theta(s_n, a_n^{-i}; m)\right)\right]
\end{aligned} \tag{145}$$

where we have "known" random variable m, where the generator is defined by our policy and opponent model.

5.2.1 Bayesian Discriminator

Now, viewing this from another perspective, following from [5, 9], we trying to approximate the $q_\theta(y|s_n, a_n^{-i}; m)$, which represents a discriminator, while the generator is represented by $P_\phi(s_n, a_n^{-i}|y; m)$ which is defined as

$$P_\phi(s_n, a_n^{-i}|y; m) = \begin{cases} P_\phi(a_n^{-i}|s_n; m) P_{\text{obs}}(s_n) & \text{if } y = 0 \\ P_{\text{obs}}(s_n, a_n) & \text{if } y = 1 \end{cases} \tag{146}$$

Given a know discriminator θ. The authors proposed to approximate the posterior of the discriminator via SVGD using the following objective

$$\max_\theta \; \log \mathbb{E}_{P(y)P_\phi(s_n, a_n^{-i}|y;m)}\left[q_\theta(y|s_n, a_n^{-i}; m)\right] \tag{147}$$

in which we can consider the lower bound from Jensen's inequality

$$\begin{aligned}
&\log \mathbb{E}_{P(y)P_\phi(s_n,a_n^{-i}|y;m)}\left[q_\theta(y|s_n,a_n^{-i};m)\right] \\
&\quad = \log \mathbb{E}_{P(y)}\left[q_\theta(y=1|s_{n,\text{obs}},a_{n,\text{obs}}^{-i};m)\cdot q_\theta(y=0|s_n,a_n^{-i};m)\right] + \text{const.} \\
&\quad \geq \mathbb{E}_{P(y)}\left[\log\left(D_\theta(s_{n,\text{obs}},a_{n,\text{obs}}^{-i};m)\right) + \log\left(1-D_\theta(s_n,a_n^{-i};m)\right)\right]
\end{aligned} \tag{148}$$

Now looking back, if we want to infer θ i.e $P(\theta|y,s,a^{-i},\phi)$ as we can see that this is proportion to the following

$$P(\theta,s,a^{-i}|y,\phi) \propto P(\theta)P(\phi)P(y|s,a^{-i},\phi)P(s,a^{-i}|\phi) \tag{149}$$

where s,a^{-i} can represent either the opponent's action or our generated action. Given this, it is appropriate to apply SVGD to infer the posterior of the discriminator. We have now estimate ranges of plausible opponent reward, which gives rise to similar mode of behavior represented by m.

5.2.2 IWGAN

This will be based on a model proposed by [5], since it might be dissatisfied for the us to simply average the result of discriminator as there are some prediction of discriminator that we should pay for attention. We would like to maximize the log-likelihood of $q^{(r)}(y)$ for the generator, where $q^{(r)}(y|s,a^{-i}) = q(1-y|s,a^{-i})$ is the reverse prediction. Given this, we can find the lower bound of the problem, which is

$$\log q_\theta(y) \geq \mathbb{E}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{q_\theta^{(r)}(y|s_k,a_k^{-i};m)P_\phi(s_k,a_k^{-i};m)}{P_\phi(s_k,a_k^{-i}|y;m)}\right] = \mathcal{L}(\phi) \tag{150}$$

We will denote the weight w_k as $\frac{q_\theta^{(r)}(y|s_k,a_k^{-i};m)P_\phi(s_k,a_k^{-i};m)}{P_\phi(s_k,a_k^{-i}|y;m)}$. Now, let's consider the derivative of the lower bound with respect to the generator, which we can show that it is equal to

$$\nabla_\phi \mathcal{L}(\phi) = \mathbb{E}_{(s_1,a_1^{-i}),\cdots,(s_K,a_K^{-i})}\left[\sum_{k=1}^{K}\frac{w_k}{\sum_{j=1}^{J}w_j}\nabla_\phi \log q_\theta^{(r)}(y|s_k,a_k^{-i};m)\right] \tag{151}$$

Gradient of some part of the weight is removed due to the fact that it is equivalent to the gradient of JSD which is removed from GAN training in general see [5] for more details.

5.2.3 A Divergence Minimization Perspective on Multi-Agent Imitation Learning

We shall consider the final extension to imitation learning framework, which is to consider the family of GANs via the len of f-GAN. In the nutshell, we can consider the adversarial optimization as follows:

$$\begin{aligned}
\min_\phi \max_{\theta,\psi} &\mathbb{E}_{P_\psi(m|\tau)P(s_n)P(a_n^{-i}|s_n;m)}\left[\sum_{n=1}^{N} f^*\left(T_\omega(s_n,a_n)\right)\right] \\
&+ \mathbb{E}_{P_\psi(m|\tau)P(s_n)\rho_\phi(a_n^{-i}|s_n;m)}\left[\sum_{n=1}^{N} T_\omega(s_n,a_n)\right]
\end{aligned} \tag{152}$$

Where f^* is the convex conjugate of f. We can choose f and T_ω so that it suited for our task. Given this, we can generalize to suitable objective. We refer to [4] for more example and extension.

5.2.4 InfoGAN

We now consider the use of information theory as in [19] in which the authors constrains the latent variable based on the mutual information between latent variable and environment rollout, which is represented in the following objective

$$\min -I(m;\tau) + \mathbb{E}_{P(\tau)}\left[D_{\text{KL}}\left(P_\theta(m|\tau)\middle\|q_\psi(m|\tau)\right)\right] \tag{153}$$

Instead of using the VAE objective the authors decided to consider the following variational inference problem. Now, we can show that this is equivalent to

$$\mathbb{E}_{P_\phi(\tau,m)}\left[\log q_\psi(m|\tau)\right] + \mathbb{E}_{p(m)}\left[\log p(m)\right] \tag{154}$$

By removing the last term due to intractability, we have the the regularization for mutual information, which is used instead of KL-regularization in [19]. Please note that this information theory bound is very similar to one used in InfoGAN without the removal of entropy term.

References

[1] Christian Daniel, Herke Van Hoof, Jan Peters, and Gerhard Neumann. Probabilistic inference for determining options in reinforcement learning. *Machine Learning*, 104(2-3):337–357, 2016.

[2] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.

[3] Jakob N Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1811.01458*, 2018.

[4] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. *arXiv preprint arXiv:1911.02256*, 2019.

[5] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.

[6] Shiyu Huang, Hang Su, Jun Zhu, and Ting Chen. Svqn: Sequential variational soft q-learning networks.

[7] Maximilian Igl, Andrew Gambardella, Jinke He, Nantas Nardelli, N Siddharth, Wendelin Böhmer, and Shimon Whiteson. Multitask soft option learning. *arXiv preprint arXiv:1904.01033*, 2019.

[8] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. *arXiv preprint arXiv:1806.02426*, 2018.

[9] Wonseok Jeon, Seokin Seo, and Kee-Eung Kim. A bayesian approach to generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 7429–7439, 2018.

[10] Minne Li, Lisheng Wu, WANG Jun, and Haitham Bou Ammar. Multi-view reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1418–1429, 2019.

[11] Minghuan Liu, Ming Zhou, Weinan Zhang, Yuzheng Zhuang, Jun Wang, Wulong Liu, and Yong Yu. Multi-agent interactions modeling with correlated policies. *arXiv preprint arXiv:2001.03415*, 2020.

[12] Elita Lobo and Scott Jordan. Soft options critic. *arXiv preprint arXiv:1905.11222*, 2019.

[13] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013.

[14] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincar\'e recurrence to convergence in imperfect information games: Finding equilibrium via regularization. *arXiv preprint arXiv:2002.08456*, 2020.

[15] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. *arXiv preprint arXiv:1903.08254*, 2019.

[16] Pavel Shvechikov, Alexander Grishin, Arsenii Kuznetsov, Alexander Fritzler, and Dmitry Vetrov. Joint belief tracking and reward optimization through approximate inference.

[17] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[18] Ying Wen, Yaodong Yang, Rui Lu, and Jun Wang. Multi-agent generalized recursive reasoning. *arXiv preprint arXiv:1901.09216*, 2019.

[19] Lantao Yu, Tianhe Yu, Chelsea Finn, and Stefano Ermon. Meta-inverse reinforcement learning with probabilistic context variables. In *Advances in Neural Information Processing Systems*, pages 11749–11760, 2019.

A Part 1: Appendix

A.1 ROMMEO

This will be a collection of proofs related to ROMMEO not exclusive to any section in the paper, which will make finding much easier.

A.1.1 ROMMEO ELBO derivation

Starting off with the KL-divergence (we have shown that this is equivalent to ELBO derived from Jensen's inequality) and we expand by doing a algebraic manipulation to get the following:

$$
\begin{aligned}
& D_{\mathrm{KL}}\left(q(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}) \middle\| P(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T} | \mathcal{O}^i_{1:T} = 1, \mathcal{O}^{-i}_{1:T} = 1)\right) \\
=& \mathbb{E}_q\left[\log \frac{q(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}) P(\mathcal{O}^i_{1:T} = 1)}{P(s_{1:T}, a^i_{1:T}, a^{-i}_{1:T}, \mathcal{O}^i_{1:T} = 1 | \mathcal{O}^{-i}_{1:T} = 1)}\right] \\
=& \mathbb{E}_q\left[\log \frac{P(\mathcal{O}^i_{1:T} = 1) \cancel{P(s_0)} \prod_{t=0}^T \pi_\theta(a^i_t | s_t, a^{-i}_t) \rho_\phi(a^{-i}_t | s_t) \cancel{P(s_{t+1} | s_t, a^i_t, a^{-i}_t)}}{\cancel{P(s_0)} \prod_{t=0}^T P_{\text{prior}}(a^i_t | s_t, a^{-i}_t) P_{\text{prior}}(a^{-i}_t | s_t) \cancel{P(s_{t+1} | s_t, a^i_t, a^{-i}_t)} P(\mathcal{O}^i_t = 1 | s_t, a^i_t, a^{-i}_t, \mathcal{O}^{-i}_t = 1)}\right] \\
=& \mathbb{E}_q\left[\log P(\mathcal{O}^i_{1:T} = 1) + \sum_{t=0}^T -\beta R(s_t, a^i_t, a^{-i}_t) + \log \frac{\pi_\theta(a^i_t | s_t, a^{-i}_t)}{P_{\text{prior}}(a^i_t | s_t, a^{-i}_t)} + \log \frac{\rho_\phi(a^{-i}_t | s_t)}{P_{\text{prior}}(a^{-i}_t | s_t)}\right] \\
=& -\mathbb{E}_q\left[\sum_{t=0}^T \beta R(s_t, a^i_t, a^{-i}_t) - \log \frac{\pi_\theta(a^i_t | s_t, a^{-i}_t)}{P_{\text{prior}}(a^i_t | s_t, a^{-i}_t)} - \log \frac{\rho_\phi(a^{-i}_t | s_t)}{P_{\text{prior}}(a^{-i}_t | s_t)}\right] - \text{const.}
\end{aligned}
$$

A.1.2 ROMMEO Optimal Agent's Policy

We starting by expanding the equation and separate the opponent's model. We then going to see that the objective is equivalent to minimizing the KL-divergence, which by Gibbs' inequality, we can imply that the optimal agent's policy can be found by setting the agent's policy to be equal to the quantity in KL-divergence.

$$
\begin{aligned}
& \mathbb{E}_q\left[\beta R(s_T, a^i_T, a^{-i}_T) - \log \frac{\pi_\theta(a^i_T | s_T, a^{-i}_T)}{P_{\text{prior}}(a^i_T | s_T, a^{-i}_T)} - \log \frac{\rho_\phi(a^{-i}_T | s_T)}{P_{\text{prior}}(a^{-i}_T | s_T)} + Q(s_T, a^{-i}_T) - Q(s_T, a^{-i}_T)\right] \\
=& \mathbb{E}_{q(s_T, a^{-i}_T)}\Bigg[\underbrace{\mathbb{E}_{\pi_\theta}\left[\beta R(s_T, a^i_T, a^{-i}_T) - \log \frac{\pi_\theta(a^i_T | s_T, a^{-i}_T)}{P_{\text{prior}}(a^i_T | s_T, a^{-i}_T)} - Q(s_T, a^{-i}_T)\right]}_{①} - \log \frac{\rho_\phi(a^{-i}_T | s_T)}{P_{\text{prior}}(a^{-i}_T | s_T)} + Q(s_T, a^{-i}_T)\Bigg]
\end{aligned}
$$

We will consider the only part (1) as it depends only on the optimal policy.

$$
\begin{aligned}
&\mathbb{E}_{\pi_\theta(a_T^i|s_T,a_T^{-i})}\left[\beta R(s_T,a_T^i,a_T^{-i}) - \log\frac{\pi_\theta(a_T^i|s_T,a_T^{-i})}{P_{\text{prior}}(a_T^i|s_T,a_T^{-i})} - Q(s_T,a_T^{-i})\right]\\
=&\mathbb{E}_{\pi_\theta(a_T^i|s_T,a_T^{-i})}\left[\log\frac{\exp\left(\beta R(s_T,a_T^i,a_T^{-i})\right)P_{\text{prior}}(a_T^i|s_T,a_T^{-i})}{\pi_\theta(a_T^i|s_T,a_T^{-i})\exp\left(Q(s_T,a_T^{-i})\right)}\right]\\
=&-D_{\text{KL}}\left(\pi_\theta(a_T^i|s_T,a_T^{-i})\middle\|\frac{\exp\left(\beta R(s_T,a_T^i,a_T^{-i})P_{\text{prior}}(a_T^i|s_T,a_T^{-i})\right)}{\exp\left(Q(s_T,a_T^{-i})\right)}\right)
\end{aligned}
$$

We can maximizing this value by minimizing the KL-divergence. And we can see that we can set $Q(s_T,a_T^{-i})$ to be normalizing constant. Thus finish the proof. Furthermore, we can see that this would works with arbitary function if we replace $R(s,a_T^i,a_T^{-i})$ by $N(s,a_T^i,a_t^{-i})$, the result would be in the same form.

A.1.3 ROMMEO Optimal Opponent's Policy

Similar proving process to the Optimal agent's policy. We will start by plugging the optimal agent's policy into the equation first, then we can use the KL-divergence trick.

$$
\begin{aligned}
&\mathbb{E}_q\left[\beta R(s_T,a_T^i,a_T^{-i}) - \log\frac{\pi_\theta(a_T^i|s_T,a_T^{-i})}{P_{\text{prior}}(a_T^i|s_T,a_T^{-i})} - \log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right]\\
=&\mathbb{E}_q\left[\beta R(s_T,a_T^i,a_T^{-i}) - \log\frac{\exp\left(\beta R(s_T,a_T^i,a_T^{-i})\right)\cancel{P_{\text{prior}}(a_T^i|s_T,a_T^{-i})}}{\cancel{P_{\text{prior}}(a_T^i|s_T,a_T^{-i})}\exp\left(Q(s_T,a_T^{-i})\right)} - \log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right]\\
=&\mathbb{E}_q\left[\log\left(\exp\left(Q(s_T,a_T^{-i})\right)\right) + \cancel{\beta R(s_T,a_T^i,a_T^{-i})} - \cancel{\log\left(\exp\left(\beta R(s_T,a_T^i,a_T^{-i})\right)\right)} - \log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right]\\
=&\mathbb{E}_{\rho_\phi(a_T^{-i}|s_T)P(s_T)}\left[Q(s_T,a_T^{-i}) - \log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right]\\
=&\mathbb{E}_{\rho_\phi(a_T^{-i}|s_T)P(s_T)}\left[Q(s_T,a_T^{-i}) - \log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)} + V(s_T) - V(s_T)\right]\\
=&\mathbb{E}_{\rho_\phi(a_T^{-i}|s_T)P(s_T)}\left[\log\frac{\exp\left(Q(s_T,a_T^{-i})\right)P_{\text{prior}}(a_T^{-i}|s_T)}{\rho_\phi(a_T^{-i}|s_T)\exp\left(V(s_T)\right)}\right] + \mathbb{E}_{P(s_T)}\left[V(s_T)\right]
\end{aligned}
$$

We will consider only the LHS of the equation, which is equal to the KL-divergence (technically expected KL-divergence since we also find expectation over state s_T)

$$
-D_{\text{KL}}\left(\rho_\phi(a_T^{-i}|s_T)\middle\|\frac{\exp\left(Q(s_T,a_T^{-i})\right)P_{\text{prior}}(a_T^{-i}|s_T)}{\exp\left(V(s_T)\right)}\right)
$$

This finishes the proof (where $V(s_T)$ is the normalizing factor). Similarly, we can replace $Q(s,a_T^{-i})$ with other functions and the result will be in the similar form.

A.1.4 ROMMEO Message Passing

Consider the objective, we simply going to plugin the values and then we are left with the message.

$$
\begin{aligned}
&\mathbb{E}_q\left[\beta R(s_T, a_T^i, a_T^{-i}) - \log\frac{\pi_\theta(a_T^i|s_T, a_T^{-i})}{P_{\text{prior}}(a_T^i|s_T, a_T^{-i})} - \log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right] \\
=&\mathbb{E}_q\left[\beta R(s_T, a_T^i, a_T^{-i}) - \log\frac{\exp\left(\beta R(s_T, a_T^i, a_T^{-i})\right)\cancel{P_{\text{prior}}(a_T^i|s_T, a_T^{-i})}}{\cancel{P_{\text{prior}}(a_T^i|s_T, a_T^{-i})}\exp\left(Q(s_t, a_t^{-i})\right)} - \log\frac{\exp\left(Q(s_T, a_T^{-i})\right)\cancel{P_{\text{prior}}(a_T^{-i}|s_T)}}{\cancel{P_{\text{prior}}(a_T^{-i}|s_T)}\exp\left(V(s_T)\right)}\right] \\
=&\mathbb{E}_q\Bigg[\beta R(s_T, a_T^i, a_T^{-i}) - \log\exp\left(\beta R(s_T, a_T^i, a_T^{-i})\right) + \log\exp\left(Q(s_t, a_t^{-i})\right) \\
&\qquad\qquad - \log\exp\left(Q(s_t, a_t^{-i})\right) + \log\exp\left(V(s_T)\right)\Bigg] \\
=&\mathbb{E}_q\left[V(s_T)\right]
\end{aligned}
$$

A.2 Balancing-Q

Interestingly Balancing-Q proof requires more efforts than ROMMEO, however, the methods are exactly the same. Furthermore, some of the proof will not be totally rigorous, although we tried to be as rigorous as possible, while trying to replicate the result so that it matches what is described in paper.

A.2.1 Derivation of Optimal Policy on last step

Since there is not opponent model, the derivation is symmetric, as we will only show the proof for agent's policy. We start with the objective and the derived "conditional" optimal policy, which is the same case as ROMMEO's agent's policy $P(a^i|s, a^{-i})$ or PR2's opponent's model $P(a^{-i}|s, a^i)$ and then the agent can the react to that:

$$
\begin{aligned}
&\mathbb{E}_{P(s_T)P(a_T, a_T^{-i}|s_T)}\left[R(s_T, a_T, a_T^{-i}) - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_T^i|s_T)}{P_{\text{prior}}(a_T^i|s_T)} - \frac{1}{\beta^{-i}}\log\frac{\rho_\phi(a_T^{-i}|s_T)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right] \\
=&\mathbb{E}_{P(s_T, a_T^i)}\Bigg[\underbrace{\mathbb{E}_{\rho(a^{-i}|s_T, a_T^i)}\left[R(s_T, a_T, a_T^{-i}) - \frac{1}{\beta^{-i}}\log\frac{\rho(a^{-i}|s_T, a_T^i)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right]}_{①} - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_T^i|s_T)}{P_{\text{prior}}(a_T^i|s_T)}\Bigg]
\end{aligned}
$$

We will consider the quality ①, which is equivalent to KL-divergence as we introduce the normalizing factor $Q^{-i}(s_T, a_T^i)$ and move β^{-i} to the reward:

$$
\begin{aligned}
&\mathbb{E}_{\rho(a^{-i}|s_T, a_T^i)}\left[R(s_T, a_T, a_T^{-i}) - \frac{1}{\beta^{-i}}\log\frac{\rho(a_T^{-i}|s_T, a_T^i)}{P_{\text{prior}}(a_T^{-i}|s_T)}\right] \\
=&\mathbb{E}_{\rho(a_T^{-i}|s_T, a_T^i)}\left[\beta^{-i}R(s_T, a_T, a_T^{-i}) - \log\frac{\rho(a_T^{-i}|s_T, a_T^i)}{P_{\text{prior}}(a_T^{-i}|s_T)} - Q^{-i}(s_T, a_T^i) + Q^{-i}(s_T, a_T^i)\right] \\
=&\mathbb{E}_{\rho(a^{-i}|s_T, a_T^i)}\left[\log\frac{\exp\left(\beta^{-i}R(s_T, a_T, a_T^{-i})\right)P_{\text{prior}}(a_T^{-i}|s_T)}{\rho(a_T^{-i}|s_T, a_T^i)\exp\left(Q^{-i}(s, a^i)\right)}\right] + \mathbb{E}_{\rho(a_T^{-i}|s_T, a_T^i)}\left[Q^{-i}(s_T, a_T^i)\right] \\
=&-D_{\text{KL}}\left(\rho(a_T^{-i}|s_T, a_T^i)\middle\|\frac{\exp\left(\beta^{-i}R(s_T, a_T, a_T^{-i})\right)P_{\text{prior}}(a_T^{-i}|s_T)}{\exp\left(Q^{-i}(s_T, a_T^i)\right)}\right) + Q^{-i}(s_T, a_T^i)
\end{aligned}
$$

With this we can set the optional "conditional" opponent policy as (the most obvious way to)

$$\rho(a_T^{-i}|s_T, a_T^i) = \frac{\exp\left(\beta^{-i} R(s_T, a_T, a_T^{-i})\right) P_{\text{prior}}(a_T^{-i}|s_T)}{\exp\left(Q^{-i}(s_T, a_T^i)\right)}$$
$$\text{where } Q^{-i}(s_T, a_T^i) = \log \int \exp\left(\beta^{-i} R(s_T, a_T, a_T^{-i})\right) P_{\text{prior}}(a_T^{-i}|s_T)\ \mathrm{d}a^{-i}$$

Let's plug this into the original objective to find the optimal policy

$$\begin{aligned}
&\mathbb{E}_{P(s_T, a_T, a_T^{-i})}\left[R(s_T, a_T, a_T^{-i}) - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_T^i|s_T)}{P_{\text{prior}}(a_T^i|s_T)} - \frac{1}{\beta^{-i}}\log\frac{\exp\left(\beta^{-i} R(s_T, a_T, a_T^{-i})\right)\cancel{P_{\text{prior}}(a_T^{-i}|s_T)}}{\cancel{P_{\text{prior}}(a_T^{-i}|s_T)}\exp\left(Q^{-i}(s_T, a_T^i)\right)}\right] \\
=&\mathbb{E}_{P(s_T, a_T, a_T^{-i})}\left[\frac{1}{\beta^{-i}}Q^{-i}(s_T, a_T^i) - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_T^i|s_T)}{P_{\text{prior}}(a_T^i|s_T)}\right] \\
=&\mathbb{E}_{P(s_T, a_T, a_T^{-i})}\Bigg[\underbrace{\frac{\beta^i}{\beta^{-i}}Q^{-i}(s_T, a_T^i)}_{Q^i(s_T, a_T^i)} - \log\frac{\pi_\theta(a_T^i|s_T)}{P_{\text{prior}}(a_T^i|s_T)}\Bigg]
\end{aligned}$$

Using the same method, we can we get the final agent's policy by minimizing the KL-divergence

$$\begin{aligned}
&\mathbb{E}_{P(s_T, a_T)}\left[Q^i(s_T, a_T^i) - \log\frac{\pi_\theta(a_T^i|s_T)}{P_{\text{prior}}(a_T^i|s_T)}\right] \\
=&\mathbb{E}_{P(s_T, a_T)}\left[Q^i(s_T, a_T^i) - \log\frac{\pi_\theta(a_T^i|s_T)}{P_{\text{prior}}(a_T^i|s_T)} + V^i(s_T) - V^i(s_T)\right] \\
=&\mathbb{E}_{P(s_T, a_T)}\left[\log\frac{\exp\left(Q^i(s_T, a_T^i)\right) P_{\text{prior}}(a_T^i|s_T)}{\pi_\theta(a_T^i|s_T)\exp\left(V^i(s_T)\right)}\right] + \mathbb{E}_{P(s_T)}\left[V^i(s_T)\right] \\
=&-D_{\text{KL}}\left(\pi_\theta(a_T^i|s_T)\middle\|\frac{\exp\left(Q^i(s_T, a_T^i)\right) P_{\text{prior}}(a_T^i|s_T)}{\exp\left(V^i(s_T)\right)}\right) + \mathbb{E}_{P(s_T)}\left[V^i(s_T)\right]
\end{aligned}$$

This can be maximize by setting agent policy to appropriate value.

A.2.2 Backward message

To calculate this, plugging back optimal opponent's policy and optimal agent's policy to the main objective:

$$\begin{aligned}
&\mathbb{E}_{P(s_t, a_t, a_t^{-i})}\left[R(s_t, a_t, a_t^{-i}) - \frac{1}{\beta^i}\log\frac{\pi_\theta(a_t|s_t)}{P_{\text{prior}}(a_t^i|s_t)} - \frac{1}{\beta^{-i}}\log\frac{\rho_\phi(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)}\right] \\
=&\mathbb{E}_{P(s_t, a_t, a_t^{-i})}\Bigg[R(s_t, a_t, a_t^{-i}) - \frac{1}{\beta^i}\log\frac{\exp\left(Q^i(s_t, a_t^i)\right)\cancel{P_{\text{prior}}(a_t^i|s_t)}}{\exp\left(V^i(s_t)\right)\cancel{P_{\text{prior}}(a_t^i|s_t)}} \\
&\qquad - \frac{1}{\beta^{-i}}\log\frac{\exp\left(\beta^{-i} R(s_t, a_t, a_t^{-i})\right)\cancel{P_{\text{prior}}(a_t^{-i}|s_t)}}{\cancel{P_{\text{prior}}(a_t^{-i}|s_t)}\exp\left(Q^{-i}(s_t, a_t^i)\right)}\Bigg] \\
=&\mathbb{E}_{P(s_t, a_t, a_t^{-i})}\Bigg[R(s_t, a_t, a_t^{-i}) - \frac{1}{\beta^i}\exp Q^i(s_t, a_t^i) + \frac{1}{\beta^i}V^i(s_t) \\
&\qquad - R(s_t, a_t, a_t^{-i}) + \frac{1}{\beta^{-i}}Q^{-i}(s_t, a_t^i)\Bigg] \\
=&\mathbb{E}_{P(s_t, a_t, a_t^{-i})}\left[\frac{1}{\beta^i}V^i(s_t)\right]
\end{aligned}$$

A.3 ROMMEO (Theory)

A.3.1 ROMMEO Contraction Mapping

The proof will be very similar to one shown in soft-Q learning, apart from the fact that there are 2 integrals we have to pass through. Starting by recalling the mapping .

$$\mathcal{T}Q(s_t, a_t^i, a_t^{-i}) = \beta R(s_t, a_T^i, a_T^{-i}) + \mathbb{E}_{P(s_{t+1}|s_t, a_t^i, a_t^{-i})}\left[V(s_{t+1})\right]$$
$$\text{where } V(s_{t+1}) = \log \int \left(\int \exp\left(Q(s_t, a_t^i, a_t^{-i}\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \ \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \ \mathrm{d}a_t^{-i}$$

and infinite norm on the Q-values is defined as

$$\left\|Q^1(s_t, a_t^i, a_t^{-i}) - Q^2(s_t, a_t^i, a_t^{-i})\right\|_\infty = \max_{s_t, a_t^i, a_t^{-i}} \left|Q^1(s_t, a_t^i, a_t^{-i}) - Q^2(s_t, a_t^i, a_t^{-i})\right|$$

Let's first plug in the definition of the mapping

$$\begin{aligned}
&\left\|\mathcal{T}Q^1(s_t, a_t^i, a_t^{-i}) - \mathcal{T}Q^2(s_t, a_t^i, a_t^{-i})\right\|_\infty \\
=&\left\|\beta R(s_t, a_T^i, a_T^{-i}) + \gamma \mathbb{E}_{P(s_{t+1}|s_t, a_t^i, a_t^{-i})}\left[V^1(s_{t+1})\right] - \beta R(s_t, a_T^i, a_T^{-i}) - \gamma \mathbb{E}_{P(s_{t+1}|s_t, a_t^i, a_t^{-i})}\left[V^2(s_{t+1})\right]\right\| \\
=&\gamma \max_{s_t, a_t, a_t^{-i}} \left|\mathbb{E}_{P(s_{t+1}|s_t, a_t^i, a_t^{-i})}\left[V^1(s_{t+1}) - V^2(s_{t+1})\right]\right| \\
\leq&\gamma \max_{s_t} \left|V^1(s_t) - V^2(s_t)\right|
\end{aligned}$$

The last inequality is based on the fact that the expectation of constant (with respect to random variable) is constant itself. Now, we want to shows that

$$\max_{s_t} \left|V^1(s_t) - V^2(s_t)\right| \leq \|Q^1(s_t, a_t^i, a_t^{-i}) - Q^2(s_t, a_t^i, a_t^{-i})\|_\infty = \varepsilon$$

Or we have to show that $V^1(s_t) - V^2(s_t) \leq \varepsilon$ and $V^1(s_t) - V^2(s_t) \geq -\varepsilon$. Starting with the first case, where it is clear that the inequality holds because $\varepsilon = \max_{s_t, a_t, a_t^{-i}} |Q^1(s_t, a_t^i, a_t^{-i} - Q^2(s_t, a_t^i, a_t^{-i})|_\infty$

$$\begin{aligned}
&\log \int \left(\int \exp\left(Q^1(s_t, a_t^i, a_t^{-i}\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \ \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \ \mathrm{d}a_t^{-i} \\
&\quad \leq \log \int \left(\int \exp\left(Q^2(s_t, a_t^i, a_t^{-i} + \varepsilon\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \ \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \ \mathrm{d}a_t^{-i} \\
&\quad = \varepsilon + \log \int \left(\int \exp\left(Q^2(s_t, a_t^i, a_t^{-i}\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \ \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \ \mathrm{d}a_t^{-i}
\end{aligned}$$

We can move the ε up the integration, since it is a constant. In the second case, we use the similar trick (with opposite signs) as follows:

$$\begin{aligned}
&\log \int \left(\int \exp\left(Q^1(s_t, a_t^i, a_t^{-i}\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \ \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \ \mathrm{d}a_t^{-i} \\
&\quad \geq \log \int \left(\int \exp\left(Q^2(s_t, a_t^i, a_t^{-i} - \varepsilon\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \ \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \ \mathrm{d}a_t^{-i} \\
&\quad = -\varepsilon + \log \int \left(\int \exp\left(Q^2(s_t, a_t^i, a_t^{-i}\right) P_{\text{prior}}(a_t^i|s_t, a_t^{-i}) \ \mathrm{d}a_t^i \right) P_{\text{prior}}(a_t^{-i}|s_t) \ \mathrm{d}a_t^{-i}
\end{aligned}$$

With 2 of these we can see that the condition holds, where, now we can conclude that

$$\left\|\mathcal{T}Q^1(s_t, a_t^i, a_t^{-i}) - \mathcal{T}Q^2(s_t, a_t^i, a_t^{-i})\right\|_\infty \leq \gamma \|Q^1(s_t, a_t^i, a_t^{-i}) - Q^2(s_t, a_t^i, a_t^{-i})\|_\infty \tag{155}$$

Thus proving the fact that Bellman equation from ROMMEO derivation is contraction mapping.

A.3.2 ROMMEO Expanding Q-value

Now, let's consider the optimal policies of both agents, then we can work backward to see what is the value of $V(s_t)$ is (under the expectation $\mathbb{E}_{s_t,a_t^i,a_t^{-i}\sim\pi,\rho,P}$) starting with the agent's optimal policy

$$\pi_\theta^*(a_t^i|s_t,a_t^{-i}) = \frac{\exp\left(Q^*(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^i|s_t,a_t^{-i})}{\exp\left(Q^*(s_t,a_t^{-i})\right)} \iff Q^*(s_t,a_t^{-i}) = Q^*(s_t,a_t^i,a_t^{-i}) - \log\left(\frac{\pi_\theta^*(a_t^i|s_t,a_t^{-i})}{P_{\text{prior}}(a_t^i|s_t,a_t^{-i})}\right)$$

We can see that this is done by rearranging the equation. Let's move to the definition of $Q^*(s_t,a_t^{-i})$, which can be recovered from optimal opponent model.

$$\rho_\phi^*(a_t^{-i}|s_t) = \frac{\exp\left(Q^*(s_t,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)}{\exp\left(V^*(s_t)\right)} \iff V^*(s_t) = Q^*(s_t,a_t^{-i}) - \log\left(\frac{\rho_\phi^*(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)}\right)$$

Combining this, we can see that we have

$$V^*(s_t) = Q^*(s_t,a_t^i,a_t^{-i}) - \log\left(\frac{\pi_\theta^*(a_t^i|s_t,a_t^{-i})}{P_{\text{prior}}(a_t^i|s_t,a_t^{-i})}\right) - \log\left(\frac{\rho_\phi^*(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)}\right)$$

Plugging this back into the Bellman equation, we can see that the results will repeatedly layout as the definition of $Q_n^{\pi,\rho}(s_n,a_n^i,a_n^{-i})$. (Note that the $V^*(s)$ in Bellman equation is the value function of the next time step, and we therefore have to take the expectation with respect to policy acting on the next state).

A.3.3 ROMMEO Improved Policy improves value function

We will show this partially, starting with the definition of the action value function

$$\begin{aligned}Q_n^{\pi,\rho} =& \beta R(s_n,a_n^i,a_n^{-i}) + \\ &\mathbb{E}_{s_t,a_t^i,a_t^{-i}\sim\pi,\rho,P}\left[\sum_{t=n+1}^{T}\gamma^{t-n+1}\left(\beta R(s_t,a_t^i,a_t^{-i}) - \log\frac{\pi(a_t^i|s_t,a_t^{-i})}{P_{\text{prior}}(a_t^i|s_t,a_t^{-i})} - \log\frac{\rho(a_t^{-i}|s_t)}{P_{\text{prior}}(a_t^{-i}|s_t)}\right)\right]\end{aligned}$$

the Bellman equation would be

$$\begin{aligned}Q_n^{\pi,\rho}(s_n,a_n^i,a_n^{-i}) &= \beta R(s_n,a_n^i,a_n^{-i}) \\ &+ \gamma\mathbb{E}_{s_{n+1},a_{n+1}^i,a_{n+1}^{-i}}\left[-\log\frac{\pi(a_{n+1}^i|s_{n+1},a_{n+1}^{-i})}{P_{\text{prior}}(a_{n+1}^i|s_{n+1},a_{n+1}^{-i})} - \log\frac{\rho(a_{n+1}^{-i}|s_{n+1})}{P_{\text{prior}}(a_{n+1}^{-i}|s_{n+1})} + Q_{n+1}^{\pi,\rho}(s_{n+1},a_{n+1}^i,a_{n+1}^{-i})\right]\end{aligned}$$

Now, we can use the following inequality, given the improvement step:

$$\tilde{\pi}(a_t^i|s_t,a_t^{-i}) = \frac{\exp\left(Q^{\pi,\rho_s}(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^i|s_t,a_t^{-i})}{\exp\left(Q^{\pi,\rho_s}(s_t,a_t^{-i})\right)}$$

We can show that:

$$\begin{aligned}&\mathbb{E}_{a_b^i\sim\pi}\left[-\log\frac{\pi(a_b^i|s_b,a_b^{-i})}{P_{\text{prior}}(a_b^i|s_b,a_b^{-i})} + Q_b^{\pi,\rho_s}(s_b,a_b^i,a_b^{-i})\right] \\ &\qquad\leq \mathbb{E}_{a_b^i\sim\tilde{\pi}}\left[-\log\frac{\tilde{\pi}(a_b^i|s_b,a_b^{-i})}{P_{\text{prior}}(a_b^i|s_b,a_b^{-i})} + Q_b^{\pi,\rho_s}(s_b,a_b^i,a_b^{-i})\right]\end{aligned}$$

For any time step b. This is obvious since in A.1.2, we have made extensive efforts on trying to minimizing the KL-value and left with the $Q(s_t,a_t^{-i})$, which is illustrated in the following equation

$$\begin{aligned}&\mathbb{E}_{a_{n+1}^i\sim\pi}\left[-\log\frac{\pi(a_{n+1}^i|s_{n+1},a_{n+1}^{-i})}{P_{\text{prior}}(a_{n+1}^i|s_{n+1},a_{n+1}^{-i})} + Q_{n+1}^{\pi,\rho_s}(s_{n+1},a_{n+1}^i,a_{n+1}^{-i})\right] \\ &\qquad= -D_{\text{KL}}\left(\pi(a_{n+1}^i|s_{n+1},a_{n+1}^{-i})\middle\|\tilde{\pi}(a_{n+1}^i|s_{n+1},a_{n+1}^{-i})\right) + Q^{\pi,\rho_s}(s_{n+1},a_{n+1}^{-i})\end{aligned}$$

Now consider the Bellman equation and its expansion:

$$
\begin{aligned}
\bar{Q}_n^{\pi,\rho}(s_n, a_n^i, a_n^{-i}) &= \beta R(s_n, a_n^i, a_n^{-i}) + \gamma \mathbb{E}_{s_{n+1}, a_{n+1}^i, a_{n+1}^{-i}} \Bigg[-\log \frac{\pi(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})}{P_{\text{prior}}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})} \\
&\quad - \log \frac{\rho(a_{n+1}^{-i} | s_{n+1})}{P_{\text{prior}}(a_{n+1}^{-i} | s_{n+1})} + Q_{n+1}^{\pi,\rho_s}(s_{n+1}, a_{n+1}^i, a_{n+1}^{-i}) \Bigg] \\
&\leq \beta R(s_n, a_n^i, a_n^{-i}) + \gamma \mathbb{E}_{s_{n+1}, a_{n+1}^i \sim \tilde{\pi}, a_{n+1}^{-i}} \Bigg[-\log \frac{\tilde{\pi}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})}{P_{\text{prior}}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})} \\
&\quad - \log \frac{\rho(a_{n+1}^{-i} | s_{n+1})}{P_{\text{prior}}(a_{n+1}^{-i} | s_{n+1})} + Q_{n+1}^{\pi,\rho_s}(s_{n+1}, a_{n+1}^i, a_{n+1}^{-i}) \Bigg] \\
&= \beta R(s_n, a_n^i, a_n^{-i}) + \gamma \mathbb{E}_{s_{n+1}, a_{n+1}^i \sim \tilde{\pi}, a_{n+1}^{-i}} \Bigg[-\log \frac{\tilde{\pi}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})}{P_{\text{prior}}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})} \\
&\quad - \log \frac{\rho(a_{n+1}^{-i} | s_{n+1})}{P_{\text{prior}}(a_{n+1}^{-i} | s_{n+1})} + \beta R(s_{n+1}, a_{n+1}^i, a_{n+1}^{-i}) + \\
&\quad \gamma \mathbb{E}_{s_{n+2}, a_{n+2}^i, a_{n+2}^{-i}} \Bigg[-\log \frac{\pi(a_{n+2}^i | s_{n+2}, a_{n+2}^{-i})}{P_{\text{prior}}(a_{n+2}^i | s_{n+2}, a_{n+2}^{-i})} \\
&\quad - \log \frac{\rho(a_{n+2}^{-i} | s_{n+2})}{P_{\text{prior}}(a_{n+2}^{-i} | s_{n+2})} + Q_{n+2}^{\pi,\rho_s}(s_{n+2}, a_{n+2}^i, a_{n+2}^{-i}) \Bigg] \Bigg] \\
&\leq \beta R(s_n, a_n^i, a_n^{-i}) + \gamma \mathbb{E}_{s_{n+1}, a_{n+1}^i \sim \tilde{\pi}, a_{n+1}^{-i}} \Bigg[-\log \frac{\tilde{\pi}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})}{P_{\text{prior}}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})} \\
&\quad - \log \frac{\rho(a_{n+1}^{-i} | s_{n+1})}{P_{\text{prior}}(a_{n+1}^{-i} | s_{n+1})} + \beta R(s_{n+1}, a_{n+1}^i, a_{n+1}^{-i}) + \\
&\quad \gamma \mathbb{E}_{s_{n+2}, a_{n+2}^i \sim \tilde{\pi}, a_{n+2}^{-i}} \Bigg[-\log \frac{\tilde{\pi}(a_{n+2}^i | s_{n+2}, a_{n+2}^{-i})}{P_{\text{prior}}(a_{n+2}^i | s_{n+2}, a_{n+2}^{-i})} \\
&\quad - \log \frac{\rho(a_{n+2}^{-i} | s_{n+2})}{P_{\text{prior}}(a_{n+2}^{-i} | s_{n+2})} + Q_{n+2}^{\pi,\rho_s}(s_{n+2}, a_{n+2}^i, a_{n+2}^{-i}) \Bigg] \Bigg] \\
&\vdots \\
&\leq \bar{Q}^{\tilde{\pi},\rho_s}(s_n, a_n^i, a_n^{-i})
\end{aligned}
$$

This is the application of repeatedly using the inequality that we derived before. Therefore, we can see that

$$
Q_n^{\pi,\rho}(s_n, a_n^i, a_n^{-i}) \leq Q^{\tilde{\pi},\rho_s}(s_n, a_n^i, a_n^{-i}) \tag{156}
$$

A.3.4 ROMMEO Improved opponent model improves value function

We will get the definition of $Q_n^{\pi,\rho}$ from appendix A.3.3, which is defined by Bellman equation as

$$
\begin{aligned}
Q_n^{\pi,\rho}(s_n, a_n^i, a_n^{-i}) &= \beta R(s_n, a_n^i, a_n^{-i}) \\
&+ \gamma \mathbb{E}_{s_{n+1}, a_{n+1}^i, a_{n+1}^{-i}} \left[-\log \frac{\pi(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})}{P_{\text{prior}}(a_{n+1}^i | s_{n+1}, a_{n+1}^{-i})} - \log \frac{\rho(a_{n+1}^{-i} | s_{n+1})}{P_{\text{prior}}(a_{n+1}^{-i} | s_{n+1})} + Q_{n+1}^{\pi,\rho}(s_{n+1}, a_{n+1}^i, a_{n+1}^{-i}) \right]
\end{aligned}
$$

Given the opponent model improvement step:

$$
\tilde{\rho}_\phi(a_t^{-i} | s_t) = \frac{\exp\left(Q^{\pi_s,\rho}(s_t, a_t^{-i})\right) P_{\text{prior}}(a_t^{-i} | s_t)}{\exp\left(V^{\pi_s,\rho}(s_t)\right)}
$$

We can show that

$$\mathbb{E}_{a_b^i\sim\pi,a_b^{-i}\sim\rho}\left[-\log\frac{\pi(a_b^i|s_b,a_b^{-i})}{P_{\text{prior}}(a_b^i|s_b,a_b^{-i})}-\log\frac{\rho(a_b^{-i}|s_b)}{P_{\text{prior}}(a_b^{-i}|s_b)}+Q_b^{\pi,\rho_s}(s_b,a_b^i,a_b^{-i})\right]$$
$$\leq\mathbb{E}_{a_b^i\sim\pi,a_b^{-i}\sim\tilde{\rho}}\left[-\log\frac{\pi(a_b^i|s_b,a_b^{-i})}{P_{\text{prior}}(a_b^i|s_b,a_b^{-i})}-\log\frac{\tilde{\rho}(a_b^{-i}|s_b)}{P_{\text{prior}}(a_b^{-i}|s_b)}+Q_b^{\pi,\rho_s}(s_b,a_b^i,a_b^{-i})\right]$$

By showing the following equality:

$$\mathbb{E}_{a_b^i\sim\pi,a_b^{-i}\sim\rho}\left[-\log\frac{\pi(a_b^i|s_b,a_b^{-i})}{P_{\text{prior}}(a_b^i|s_b,a_b^{-i})}-\log\frac{\rho(a_b^{-i}|s_b)}{P_{\text{prior}}(a_b^{-i}|s_b)}+Q_b^{\pi,\rho_s}(s_b,a_b^i,a_b^{-i})\right]$$
$$=-D_{\text{KL}}\left(\rho(a_b^{-i}|s_b)\middle\|\tilde{\rho}(a_b^{-i}|s_b)\right)-\mathbb{E}_{a_b^i,a_b^{-i}\sim\pi,\rho}\left[-\log\frac{\pi(a_b^i|s_b,a_b^{-i})}{P_{\text{prior}}(a_b^i|s_b,a_b^{-i})}\right]+V^{\pi_s,\rho}(s_b)$$

Using the similar expansion to the agent improvement policy proof, we can thus proving the fact that updating opponent model implies increases in policy's value.

A.4 Balancing-Q (Theory)

A.4.1 Balancing-Q value Contraction Mapping

Given the following Bellma-equation,

$$\begin{aligned}&\mathcal{T}Q^*(s_t,a_t^i,a_t^{-i})=R(s_t,a_t,a_t^{-i})+\mathbb{E}_{P(s_{t+1}|s_t,a_t,a_t^{-i})}\left[V^*(s_{t+1})\right]\\&\text{where } V^*(s_{t+1})=\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\end{aligned}\tag{157}$$

similar to the proof of ROMMEO (appendix A.3.1), we want to show that

$$\max_{s_t}\left|V^1(s_t)-V^2(s_t)\right|\leq\|Q^1(s_t,a_t^i,a_t^{-i})-Q^2(s_t,a_t^i,a_t^{-i})\|_\infty=\varepsilon$$

Or we have to show that $V^1(s_t)-V^2(s_t)\leq\varepsilon$ and $V^1(s_t)-V^2(s_t)\geq-\varepsilon$. Starting with the first case

$$\begin{aligned}&\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^1(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\\&\leq\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^2(s_t,a_t^i,a_t^{-i})+\varepsilon\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\\&=\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^2(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}+\varepsilon\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\\&=\varepsilon+\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^2(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\end{aligned}$$

Now consider the second case, where

$$\begin{aligned}&\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^1(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\\&\geq\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^2(s_t,a_t^i,a_t^{-i})-\varepsilon\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\\&=\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^2(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}-\varepsilon\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\\&=-\varepsilon+\log\int\exp\left(\frac{\beta^i}{\beta^{-i}}\log\int\exp\left(\beta^{-i}Q^2(s_t,a_t^i,a_t^{-i})\right)P_{\text{prior}}(a_t^{-i}|s_t)\ \mathrm{d}a^{-i}\right)P_{\text{prior}}(a_t^i|s_t)\ \mathrm{d}a^i\end{aligned}$$

And we finish the proof.

A.4.2 Balancing-Q value Opponent improvement

The Bellman equation for $\bar{Q}^{\pi,\rho}(s_t, a_t^i, a_t^{-i})$ is equal to

$$Q^{\pi,\rho}(s_t, a_t^i, a_t^{-i}) = R(s_t, a_t^i, a_t^{-i}) + \mathbb{E}_{s_{t+1}, a_{t+1}^i, a_{t+1}^{-i}}\Bigg[- \frac{1}{\beta^i} \log \frac{\pi(a_{t+1}^i|s_{t+1})}{P_{\text{prior}}(a_{t+1}^i|s_{t+1})} - \frac{1}{\beta^{-i}} \log \frac{\rho(a_{t+1}^{-i}|s_{t+1})}{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})} + Q^{\pi,\rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})\Bigg]$$

With this we have the following factorization

$$Q^{\pi,\rho}(s_t, a_t^i, a_t^{-i}) = R(s_t, a_t^i, a_t^{-i}) + \mathbb{E}_{s_{t+1}}\Bigg[\mathbb{E}_{a_{t+1}^i, a_{t+1}^{-i}}\Bigg[- \frac{1}{\beta^i} \log \frac{\pi(a_{t+1}^i|s_{t+1})}{P_{\text{prior}}(a_{t+1}^i|s_{t+1})} - \frac{1}{\beta^{-i}} \log \frac{\rho(a_{t+1}^{-i}|s_{t+1})}{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})} + Q^{\pi,\rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})\Bigg]\Bigg]$$

We want to show that, for $\beta^{-i} < 0$ and

$$\begin{aligned}
&\mathbb{E}_{a_{t+1}^i, a_{t+1}^{-i} \sim \rho}\Bigg[- \frac{1}{\beta^i} \log \frac{\pi(a_{t+1}^i|s_{t+1})}{P_{\text{prior}}(a_{t+1}^i|s_{t+1})} - \frac{1}{\beta^{-i}} \log \frac{\rho(a_{t+1}^{-i}|s_{t+1})}{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})} + Q^{\pi,\rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})\Bigg] \\
&\quad \geq \mathbb{E}_{a_{t+1}^i, a_{t+1}^{-i} \sim \tilde{\rho}}\Bigg[- \frac{1}{\beta^i} \log \frac{\pi(a_{t+1}^i|s_{t+1})}{P_{\text{prior}}(a_{t+1}^i|s_{t+1})} - \frac{1}{\beta^{-i}} \log \frac{\tilde{\rho}(a_{t+1}^{-i}|s_{t+1})}{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})} + Q^{\pi,\rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})\Bigg]
\end{aligned}$$

This can be proved by expanding the KL-divergence as follows, starting with only KL-divergence term

$$\begin{aligned}
&-\frac{1}{\beta^{-i}} D_{\text{KL}}\left(\rho(a_{t+1}^{-i}|s_{t+1}) \middle\| \tilde{\rho}(a_{t+1}^{-i}|s_{t+1})\right) \\
&\quad = -\frac{1}{\beta^{-i}} \int \rho(a_{t+1}^{-i}|s_{t+1}) \log \frac{\rho(a_{t+1}^{-i}|s_{t+1})}{\tilde{\rho}(a_{t+1}^{-i}|s_{t+1})} \, \mathrm{d}a_t^{-i} \\
&\quad = -\frac{1}{\beta^{-i}} \int \rho(a_{t+1}^{-i}|s_{t+1}) \log \frac{\rho(a_{t+1}^{-i}|s_{t+1}) \exp(V^{-i,\pi,\rho}(s_{t+1}))}{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1}) \exp\left(Q^{-i,\pi,\rho}(s_{t+1}, a_{t+1}^{-i})\right)} \, \mathrm{d}a_{t+1}^{-i}
\end{aligned}$$

We can see that once we expand we have the following terms

$$\begin{aligned}
&-\frac{1}{\beta^{-i}} \int \rho(a_{t+1}^{-i}|s_{t+1}) \log \frac{\rho(a_{t+1}^{-i}|s_{t+1})}{P_{\text{prior}}(a_{t+1}^{-i}|s_{t+1})} \, \mathrm{d}a_{t+1}^{-i} - \frac{1}{\beta^{-i}} \exp(V^{-i,\pi,\rho}(s_t)) \\
&\quad - \frac{1}{\beta^{-i}} \int \rho(a_{t+1}^{-i}|s_{t+1}) Q^{-i,\pi,\rho}(s_{t+1}, a_{t+1}^{-i}) \, \mathrm{d}a_{t+1}^i
\end{aligned}$$

As we can see, if we plug the value of KL-divergence inside, all the terms match except this term:

$$\mathbb{E}_{a_{t+1}^i \sim \pi}\left[\frac{1}{\beta^{-i}} \int \rho(a_{t+1}^{-i}|s_{t+1}) Q^{-i,\pi,\rho}(s_{t+1}, a_{t+1}^{-i}) \, \mathrm{d}a_{t+1}^{-i}\right]$$

We will have to relax a bit by assuming that π is based on opponent's imagination of conditional agent policy (see appendix A.2.1) that is used for deriving the optimal objective. Given the fact that

$$\pi(a_t^i|a_t^{-i}, s_t) = \frac{\exp(\beta^i Q^{\pi,\rho}(s_t, a_t^i, a_t^{-i})) P_{\text{prior}}(a_t^i|s_t)}{\exp\left(Q^{i,\pi,\rho}(s_t, a_t^{-i})\right)}$$

We can see that (in the expectation of π) notice the action value function of opponent now.

$$Q^{-i,\pi,\rho}(s_t, a_t^{-i}) = \frac{\beta^{-i}}{\beta^i}\left(\beta^i Q(s_t, a_t^i, a_t^{-i}) - \log \frac{\pi(a_t^i|a_t^{-i}, s_t)}{P_{\text{prior}}(a_t^i|s_t)}\right)$$

Then, we can plug in the equation into the problematic terms, yielding the following

$$
\begin{aligned}
&\mathbb{E}_{a^i_{t+1}\sim\pi}\left[\frac{1}{\beta^{-i}}\int \rho(a^{-i}_{t+1}|s_{t+1})Q^{-i,\pi,\rho}(s_{t+1},a^{-i}_{t+1})\ \mathrm{d}a^{-i}_{t+1}\right]\\
=&\mathbb{E}_{a^i_{t+1}\sim\pi}\left[\frac{1}{\beta^{-i}}\int \rho(a^{-i}_{t+1}|s_{t+1})\left(\beta^{-i}Q(s_{t+1},a^i_{t+1},a^{-i}_{t+1})-\frac{\beta^{-i}}{\beta^i}\log\frac{\pi(a^i_{t+1}|a^{-i}_{t+1},s_{t+1})}{P_{\text{prior}}(a^i_{t+1}|s_{t+1})}\right)\ \mathrm{d}a^{-i}_{t+1}\right]\\
=&\mathbb{E}_{a^i_{t+1},a^{-i}_{t+1}\sim\pi,\rho}\left[Q(s_{t+1},a^i_{t+1},a^{-i}_{t+1})\right]-\frac{1}{\beta^i}\mathbb{E}_{a^i_{t+1}}\left[\log\frac{\mathbb{E}_{a^{-i}_{t+1}}\left[\pi(a^i_{t+1}|a^{-i}_{t+1},s_{t+1})\right]}{P_{\text{prior}}(a^i_{t+1}|s_{t+1})}\right]
\end{aligned}
$$

As this is difference cases from the ROMMEO in which the value of $\pi(a^i_{t+1}|a^{-i}_{t+1},s_{t+1})$ is the same, the representation of agent's policy in opponent view might not always equivalent. Now, we have the following equality

$$
\begin{aligned}
&\mathbb{E}_{a^i_{t+1},a^{-i}_{t+1}\sim\rho}\Bigg[-\frac{1}{\beta^i}\log\frac{\pi(a^i_{t+1}|s_{t+1})}{P_{\text{prior}}(a^i_{t+1}|s_{t+1})}-\frac{1}{\beta^{-i}}\log\frac{\rho(a^{-i}_{t+1}|s_{t+1})}{P_{\text{prior}}(a^{-i}_{t+1}|s_{t+1})}+Q^{\pi,\rho}(s_{t+1},a^i_{t+1},a^{-i}_{t+1})\Bigg]\\
&\qquad=\mathbb{E}_{a^i_{t+1}}\left[-\frac{1}{\beta^{-i}}D_{\text{KL}}\left(\rho(a^{-i}_{t+1}|s_{t+1})\middle\|\tilde{\rho}(a^{-i}_{t+1}|s_{t+1})\right)\right]+\frac{V^{-i,\pi,\rho}(s_t)}{\beta^{-i}}\\
&\qquad-\frac{1}{\beta^i}D_{\text{KL}}\left(\pi(a^i_{t+1}|s_{t+1})\middle\|\mathbb{E}_{a^{-i}_{t+1}}\left[\pi(a^i_{t+1}|a^{-i}_{t+1},s_{t+1})\right]\right)
\end{aligned}
$$

Finally, we can see that this inequality will be applicable by assumption in equation 38. Given all the facts, we will consider the effect of arbitrary ρ to its update.

- if $\beta^{-i}>0$ (in a cooperative case) we can see that given the original ρ will make the value *less* than when plugging $\tilde{\rho}$ into the equation by making the first KL-divergence more than the
- if $\beta^{-i}<0$ (in a competitive case) we can see that given the original ρ will make the value *more* than when plugging $\tilde{\rho}$ into the equation.

Thus, we have finishes the proof for the improvement, which will lead to more interesting implication in section 1.6.1

A.5 Unified View

A.5.1 Agent's policy Variational Inference

All the derivation is almost the same as ROMMEO cases. We have the posterior as

$$
\begin{aligned}
&P(s_{1:T},a^i_{1:T},a^{-i}_{1:T},\mathcal{O}^i_{1:T}=1)\\
&\qquad=P(s_0)\prod_{t=0}^{T}P_{\text{prior}}(a^i_t|s_t,a^{-i}_t)P_{\text{prior}}(a^{-i}_t|s_t)P(s_{t+1}|s_t,a^i_t,a^{-i}_t)P(\mathcal{O}^i_t=1|s_t,a^i_t,a^{-i}_t)
\end{aligned}
$$

The variational distribution is defined as

$$
q(s_{1:T},a^i_{1:T},a^{-i}_{1:T})=P(s_0)\prod_{t=0}^{T}\pi_\theta(a^i_t|s_t,a^{-i}_t)P_{\text{prior}}(a^{-i}_t|s_t)P(s_{t+1}|s_t,a^i_t,a^{-i}_t)
$$

Then we have to find the KL-divergence and its corresponding ELBO

$$
\begin{aligned}
&D_{\text{KL}}\left(q(s_{1:T},a^i_{1:T},a^{-i}_{1:T})\middle\|\frac{P(s_{1:T},a^i_{1:T},a^{-i}_{1:T},\mathcal{O}^i_{1:T}=1)}{P(\mathcal{O}^i_{1:T}=1))}\right)\\
&\qquad=\ -\text{const.}+-\mathbb{E}_q\left[\sum_{t=0}^{T}\beta^iR(s_t,a^i_t,a^{-i}_t)-\log\frac{\pi(a^i_t|a^{-i}_t,s_t)}{P_{\text{prior}}(a^i_t|a^{-i}_t,s_t)}\right]
\end{aligned}
$$

With this, we can follows the solving technique from ROMMEO to arrived at the optimal policy.

A.5.2 Agent's optimal opponent model Variational Inference

Starting with the posterior

$$P(s_t, a_t^{-i}, \mathcal{O}_t^{-i} = 1) = P(\mathcal{O}_t^{-i} = 1 | s_t, a_t^{-i}) P(s_t) P_{\text{prior}}(a_t^{-i} | s_t)$$

where the optimal random variable defined as

$$P(\mathcal{O}_t^{-i} = 1 | s_t, a_t^{-i}) = \exp\left(\frac{\beta^{-i}}{\beta^i} Q^i(s_t, a_t^{-i}))\right)$$

Given the variational distribution to be

$$q(s_t, a_t^{-i}) = P(s_t) \rho_\phi(a_t^{-i} | s_t)$$

The we want to find the KL-divergence of these 2 and trying to optimize

$$\begin{aligned} &D_{\text{KL}}\left(q(s_t, a_t^{-i}) \middle\| \frac{P(s_t, a_t^{-i}, \mathcal{O}_t^{-i} = 1)}{P(\mathcal{O}_t^{-i} = 1)}\right) \\ &\quad = D_{\text{KL}}\left(\rho_\phi(a_t^{-i} | s_t) \middle\| \frac{\exp\left(\frac{\beta^{-i}}{\beta^i} Q^i(s_t, a_t^{-i})) P_{\text{prior}}(a_t^{-i} | s_t)\right)}{V^{-i}(s_t)}\right) \end{aligned}$$

Therefore, we can set $\rho_\phi(a_t^{-i} | s_t)$ to be equal to

$$\frac{\exp\left(\frac{\beta^{-i}}{\beta^i} Q^i(s_t, a_t^{-i})) P_{\text{prior}}(a_t^{-i} | s_t)\right)}{V^{-i}(s_t)}$$

Thus finish the proof. Note that $V^{-i}(s_t)$ is simply the normalizing factor.

A.5.3 Contraction Mapping of Agent's Action value function Evaluation

We will carry out the proof that is similar to

$$\begin{aligned} &\mathcal{T} Q^i(s_t, a_t^i, a_t^{-i}) \\ &\quad = \beta^i R(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}, a_t^{-i} \sim P_{\text{prior}}(a_t^{-i} | s_t)} \left[\log \int \exp\left(Q^i(s_t, a_t^i, a_t^{-i}))\right) P_{\text{prior}}(a_t^i | a_t^{-i}, s_t) \, \mathrm{d}a_t^i\right] \end{aligned} \tag{158}$$

We can see that is is equivalence to soft-Q learning, when we simpily have to augment the transition probability with opponent's policy and the agent's $Q^i(s_t, a_t^{-i})$ function.

A.5.4 Agent policy improvement

Start with the definition of action value function

$$Q^{i,\pi,\rho}(s_t, a_t^i, a_t^{-i}) = \beta^i R(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}, a_{t+1}^{-i} \sim \rho(a_{t+1}^{-i} | s_{t+1})} \left[Q^{i,\pi,\rho}(s_{t+1}, a_{t+1}^{-i})\right]$$

Consider expanding the Bellman-equation, we can see that

$$\begin{aligned} Q^{i,\pi,\rho}(s_t, a_t^i, a_t^{-i}) = \beta^i R(s_t, a_t^i, a_t^{-i}) + \gamma \mathbb{E}_{s_{t+1}, a_{t+1}^{-i} \sim \rho(a_{t+1}^{-i} | s_{t+1}), a_{t+1}^i \sim \pi} \Bigg[&- \log \frac{\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})}{P_{\text{prior}}(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})} \\ &+ Q^{i,\pi,\rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})\Bigg] \end{aligned}$$

We can show that

$$\begin{aligned} &\mathbb{E}_{s_{t+1}, a_{t+1}^{-i} \sim \rho(a_{t+1}^{-i} | s_{t+1}), a_{t+1}^i \sim \pi} \left[- \log \frac{\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})}{P_{\text{prior}}(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})} + Q^{i,\pi,\rho}(s_{t+1}, a_{t+1}^i, a_{t+1}^{-i})\right] \\ &\qquad = - D_{\text{KL}}\left(\pi(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i}) \middle\| \tilde{\pi}(a_{t+1}^i | s_{t+1}, a_{t+1}^{-i})\right) + Q^{i,\pi,\rho}(s_{t+1}, a_{t+1}^{-i}) \end{aligned}$$

Therefore, we can prove that

$$\mathbb{E}_{s_{t+1},a^{-i}_{t+1}\sim\rho(a^{-i}_{t+1}|s_{t+1}),a^{i}_{t+1}\sim\pi}\left[-\log\frac{\pi(a^{i}_{t+1}|s_{t+1},a^{-i}_{t+1})}{P_{\text{prior}}(a^{i}_{t+1}|s_{t+1},a^{-i}_{t+1})}+Q^{i,\pi,\rho}(s_{t+1},a^{i}_{t+1},a^{-i}_{t+1})\right]$$
$$\leq\mathbb{E}_{s_{t+1},a^{-i}_{t+1}\sim\rho(a^{-i}_{t+1}|s_{t+1}),a^{i}_{t+1}\sim\tilde{\pi}}\left[-\log\frac{\tilde{\pi}(a^{i}_{t+1}|s_{t+1},a^{-i}_{t+1})}{P_{\text{prior}}(a^{i}_{t+1}|s_{t+1},a^{-i}_{t+1})}+Q^{i,\pi,\rho}(s_{t+1},a^{i}_{t+1},a^{-i}_{t+1})\right]$$

Which we can recursively apply this inequality and arrived at the result we need.

B Part 2: Appendix

B.1 Single Agent Case

B.1.1 On Solving Hierarchical RL

We will follow the usual hierarchical literature, where we mostly consider the the optimal master policy first $\pi^H(z_T|s_T)$. Consider the objective that we want to maximize, we have the following results

$$\begin{aligned}&b_T\mathbb{E}_{\pi(z_T|s_T)}\left[\mathbb{E}_{a_T\sim\pi(a_T|s_T,z_T)}\left[\beta r(s_T,a_T)\right]-\log\frac{\pi^M_\phi(z_T|s_T)}{P_{\text{prior}}(z_T|s_T)}\right]\\&+(1-b_T)\mathbb{E}_{a_T\sim\pi(a_T|s_T,z_{T-1})}\left[\beta r(s_T,a_T)\right]+\text{ other quantities}\end{aligned}\tag{159}$$

which would lead to the following solution i.e

$$\begin{aligned}\pi^H(z_T|s_T)&=\frac{\frac{1}{m}\exp\left(Q(s_T,z_T)\right)}{\exp\left(V^H(s_T)\right)}\\\text{where }Q^H(s_T,z_T)&=\mathbb{E}_{a_T\sim\pi(a_T|s_T,z_T)}\left[\beta r(s_T,a_T)-\log\frac{\pi_\theta(a_T|s_T,z_T)}{P_{\text{prior}}(a_T|s_T,z_T)}\right]\\V^H(s_T)&=\frac{1}{m}\log\int\exp\left(Q^H(s_T,z_T)\right)\,\mathrm{d}z_T\end{aligned}\tag{160}$$

Before, we move on, let's consider the objective in the fullest form, so that we won't be missing any variables. We using the fact that the terminating condition is discrete therefore, we are able to arrive at the following expectation

$$\begin{aligned}\mathbb{E}_{s_T,b_T\sim\pi^T(b_T|s_T,z_{T-1})}\Bigg[b_T\mathbb{E}\Bigg[&\beta r(s_T,a_T)-\log\frac{\pi_\theta(a_T|s_T,z_T)}{P_{\text{prior}}(a_T|s_T,z_T)}-\log\frac{\pi^T(b_T|s_T,z_{T-1})}{P^T_{\text{prior}}(b_T|s_T,z_{T-1})}-\log\frac{\pi^H_\phi(z_T|s_T)}{P^H_{\text{prior}}(z_T|s_T)}\Bigg]\\&+(1-b_T)\mathbb{E}\left[\beta r(s_T,a_T)-\log\frac{\pi_\theta(a_T|s_T,z_{T-1})}{P_{\text{prior}}(a_T|s_T,z_{T-1})}-\log\frac{\pi^T(1-b_T|s_T,z_{T-1})}{P^T_{\text{prior}}(1-b_T|s_T,z_{T-1})}\right]\Bigg]\end{aligned}\tag{161}$$

We can see that in 2 possible cases the agent acts differently according to z_T assigned by the master policy $\pi^H(z_T|s_T,z_{T-1},b_T)$. Now let's replace the back the optimal policy to the, which gives us

$$\begin{aligned}\mathbb{E}_{s_T,b_T\sim\pi^T(b_T|s_T,z_{T-1})}\Bigg[b_T\mathbb{E}\Bigg[&V^H(s_T)-\log\frac{\pi^T(b_T|s_T,z_{T-1})}{P^T_{\text{prior}}(b_T|s_T,z_{T-1})}\Bigg]\\&+(1-b_T)\mathbb{E}\left[\beta r(s_T,a_T)-\log\frac{\pi_\theta(a_T|s_T,z_{T-1})}{P_{\text{prior}}(a_T|s_T,z_{T-1})}-\log\frac{\pi^T(1-b_T|s_T,z_{T-1})}{P^T_{\text{prior}}(1-b_T|s_T,z_{T-1})}\right]\Bigg]\end{aligned}\tag{162}$$

Let's proceed to maximize the objective via the low-level policy. We can see that the optimization objective (removing irrelevant terms)

$$\mathbb{E}_{s_T,b_T\sim\pi^T(b_T|s_T,z_{T-1})}\left[b_T\mathbb{E}\left[V^H(s_T)\right]+(1-b_T)\mathbb{E}\left[\beta r(s_T,a_T)-\log\frac{\pi_\theta(a_T|s_T,z_{T-1})}{P_{\text{prior}}(a_T|s_T,z_{T-1})}\right]\right]\tag{163}$$

we shall consider the fact that the π^H is already maximized, meaning that (by rearranging the optimal policy term):

$$V^H(s_T) = \mathbb{E}_{z_T \sim \pi^*} \left[\int \pi(a_T|s_T, z_T) \left(\beta r(s_T, a_T) - \log \frac{\pi_\theta(a_T|s_T, z_T)}{P_{\text{prior}}(a_T|s_T, z_T)} \right) \, \mathrm{d}a_T - \log \frac{\pi^H(z_T|s_T)}{P_{\text{prior}}(z_T|s_T)} \right] \tag{164}$$

Now, we are back with the following optimization objective after removing irrelevant terms (this seem to be a walkabout just to make sure that every component is correct). Furthermore, note that we have also merge the terminating condition and master policy:

$$\mathbb{E}_{s_T, b_T, z_T \sim \pi^{H,*}, a_T} \left[\beta r(s_T, a_T) - \log \frac{\pi_\theta(a_T|s_T, z_T)}{P_{\text{prior}}(a_T|s_T, z_T)} \right] \tag{165}$$

which lead to the following optimization problem

$$\begin{aligned} \pi_\theta(a_T|s_T, z_T) &= \frac{P_{\text{prior}}(a_T|s_T, z_T) \exp\left(Q(s_T, a_T, z_T)\right)}{\exp V(s_T, z_T)} \\ \text{where } Q(s_T, a_T, z_T) &= \beta r(s_T, a_T) \\ V(s_T, z_T) &= \log \int P_{\text{prior}}(a_T|s_T, z_T) \exp\left(Q(s_T, a_T, z_T)\right) \, \mathrm{d}a_T \end{aligned} \tag{166}$$

The following objective is reduced to

$$\mathbb{E}_{s_T, z_T, z_{T-1}, a_T, b_T} \left[b_T \left[V^H(s_T) \right] + (1 - b_T) V(s_T, z_{T-1}) - \log \frac{\pi^T(b_T|s_T, z_{T-1})}{P^T_{\text{prior}}(b_T|s_T, z_{T-1})} \right] \tag{167}$$

Now let's solve the equation to get optimal termination condition. Using the same method, we arrived at

$$\begin{aligned} \pi^T(b_T|s_T, z_{T-1}) &= \frac{P^T_{\text{prior}}(b_T|s_T, z_{T-1}) \exp U(s_T, z_{T-1}, b_T)}{\exp U(s_T, z_{T-1})} \\ \text{where } U(s_T, z_{T-1}, b_T) &= b_T \left[V^H(s_T) \right] + (1 - b_T) V(s_T, z_{T-1}) \\ U(s_T, z_{T-1}) &= \log \int P^T_{\text{prior}}(b_T|s_T, z_{T-1}) \exp\left(b_T \left[V^H(s_T) \right] + (1 - b_T) V(s_T, z_{T-1}) \right) \, \mathrm{d}b_T \end{aligned} \tag{168}$$

Now, let's plug back the result of optimal termination condition, we are left with the following "message", $U(s_T, z_{T-1})$. However, let's unroll $U(s_T, z_{T-1})$ the whole message, we can see that this value fits perfectly.

$$\begin{aligned} U(s_T, z_{T-1}) &= U(s_T, z_{T-1}, b_T) - \log \frac{\pi^T(b_T|s_T, z_{T-1})}{P^T_{\text{prior}}(b_T|s_T, z_{T-1})} \\ &= b_T \left[V^H(s_T) \right] + (1 - b_T) V(s_T, z_{T-1}) - \log \frac{\pi^T(b_T|s_T, z_{T-1})}{P^T_{\text{prior}}(b_T|s_T, z_{T-1})} \\ &= b_T \left[Q^H(s_T, a_T) - \log \frac{\pi^H(z_T|s_T)}{P_{\text{prior}}(z_T|s_T)} \right] + (1 - b_T) \left[\beta r(s_T, a_T) - \log \frac{\pi_\theta(a_T|s_T, z_T)}{P_{\text{prior}}(a_T|s_T, z_T)} \right] \\ &\quad - \log \frac{\pi^T(b_T|s_T, z_{T-1})}{P^T_{\text{prior}}(b_T|s_T, z_{T-1})} \\ &= \beta r(s_T, a_T) - \log \frac{\pi_\theta(a_T|s_T, z_T)}{P_{\text{prior}}(a_T|s_T, z_T)} - \frac{\pi^T(b_T|s_T, z_{T-1})}{P^T_{\text{prior}}(b_T|s_T, z_{T-1})} - b_T \left[\log \frac{\pi^H(z_T|s_T)}{P_{\text{prior}}(z_T|s_T)} \right] \end{aligned}$$

The regularization for π^H only applies when we actually use the master agent. Now, we have a message passing consider arbitrary time step t, we have the following incoming messages

$$\begin{aligned}\mathbb{E}_{s_t,b_t\sim\pi^T(b_t|s_t,z_{T-1})}\Bigg[b_t\mathbb{E}\left[-\log\frac{\pi_\theta(a_t|s_t,z_t)}{P_{\text{prior}}(a_t|s_t,z_t)}-\log\frac{\pi^T(b_t|s_t,z_{T-1})}{P^T_{\text{prior}}(b_t|s_t,z_{T-1})}-\log\frac{\pi^H_\phi(z_t|s_t)}{P^H_{\text{prior}}(z_t|s_t)}\right]\\ +(1-b_t)\mathbb{E}\left[-\log\frac{\pi_\theta(a_t|s_t,z_{T-1})}{P_{\text{prior}}(a_t|s_t,z_{T-1})}-\log\frac{\pi^T(1-b_t|s_t,z_{T-1})}{P^T_{\text{prior}}(1-b_t|s_t,z_{T-1})}\right]\\ +\beta r(s_t,a_t)+\mathbb{E}_{s_{t+1}}\left[U(s_{t+1},z_t)\right]\Bigg]\end{aligned} \tag{169}$$

We will define a Bellman equation, which is a complete analogous to Option-framework

$$Q(s_t,a_t,z_t)=\beta r(s_t,a_t)+\mathbb{E}_{s_{t+1}}\left[\mathbb{E}_{a_{t+1},z_{t+1},b_{t+1}}\left[U(s_{t+1},z_t)\right]\right] \tag{170}$$

Given this we now have the following optimization problem

$$\begin{aligned}\mathbb{E}_{s_t,b_t\sim\pi^T(b_t|s_t,z_{t-1})}\Bigg[b_t\mathbb{E}\left[Q(s_t,a_t,z_t)-\log\frac{\pi_\theta(a_t|s_t,z_t)}{P_{\text{prior}}(a_t|s_t,z_t)}-\log\frac{\pi^T(b_t|s_t,z_{t-1})}{P^T_{\text{prior}}(b_t|s_t,z_{t-1})}-\log\frac{\pi^H_\phi(z_t|s_t)}{P^H_{\text{prior}}(z_t|s_t)}\right]\\ +(1-b_t)\mathbb{E}\left[Q(s_t,a_t,z_t)-\log\frac{\pi_\theta(a_t|s_t,z_{t-1})}{P_{\text{prior}}(a_t|s_t,z_{t-1})}-\log\frac{\pi^T(1-b_t|s_t,z_{t-1})}{P^T_{\text{prior}}(1-b_t|s_t,z_{t-1})}\right]\Bigg]\end{aligned} \tag{171}$$

Now, we shall repeat the process from the above, we are arriving at the following solutions for each policies: For the low-level policy, we have

$$\begin{aligned}\pi_\theta(a_t|s_t,z_t)&=\frac{P_{\text{prior}}(a_t|s_t,z_t)\exp\left(Q(s_t,a_t,z_t)\right)}{\exp V(s_t,z_t)}\\ \text{where } V(s_t,z_t)&=\log\int P_{\text{prior}}(a_t|s_t,z_t)\exp\left(Q(s_t,a_t,z_t)\right)\,\mathrm{d}a_t\end{aligned} \tag{172}$$

For the high-level policy, we have

$$\begin{aligned}\pi^H(z_t|s_t)&=\frac{\frac{1}{m}\exp\left(Q^H(s_t,z_t)\right)}{\exp\left(V^H(s_t)\right)}\\ \text{where } Q^H(s_t,z_t)&=V(s_t,z_t)\\ V^H(s_t)&=\frac{1}{m}\log\int\exp\left(Q(s_t,z_t)\right)\,\mathrm{d}z_T\end{aligned} \tag{173}$$

Finally, the termination policy is

$$\begin{aligned}\pi^T(b_t|s_t,z_{t-1})&=\frac{P^T_{\text{prior}}(b_t|s_t,z_{t-1})\exp U(s_t,z_{t-1},b_t)}{\exp U(s_t,z_{t-1})}\\ \text{where } U(s_t,z_{t-1},b_t)&=b_t\left[V^H(s_t)\right]+(1-b_t)V(s_t,z_{t-1})\\ U(s_t,z_{t-1})&=\log\int P^T_{\text{prior}}(b_t|s_t,z_{t-1})\exp\left(b_t\left[V^H(s_t)\right]+(1-b_t)V(s_t,z_{t-1})\right)\,\mathrm{d}b_t\end{aligned} \tag{174}$$

Thus finishing the derivation of soft-hierarchical agents.

B.1.2 Solving Lower-Level Policy first

We can solve the lower-level policy first (similar to multi-agent where we optimize the conditional polity first i.e $\pi(a^i|a^{-i}, s)$). We have the following equation

$$\begin{aligned} \pi_\theta(a_T|s_T, z_T) &= \frac{P_{\text{prior}}(a_T|s_T, z_T)\exp\left(Q(s_T, a_T, z_T)\right)}{\exp V(s_T, z_T)} \\ \text{where } Q(s_T, a_T, z_T) &= \beta r(s_T, a_T) \\ V(s_T, z_T) &= \log \int P_{\text{prior}}(a_T|s_T, z_T)\exp\left(Q(s_T, a_T, z_T)\right)\ \mathrm{d}a_T \end{aligned} \tag{175}$$

Let's consider arbitrary time step (where this is the solution given briefly without explicit proof in [12])

$$\begin{aligned} \pi_\theta(a_t|s_t, z_t) &= \frac{P_{\text{prior}}(a_t|s_t, z_t)\exp\left(Q(s_t, a_t, z_t)\right)}{\exp V^{\text{lower}}(s_t, z_t)} \\ \text{where } Q(s_t, a_t, z_t) &= \beta r(s_t, a_t) + \mathbb{E}_{s_{t+1}}\Big[\mathbb{E}_{b_{t+1}\sim P(b_{t+1}|s_{t+1}, z_t), z'_{t+1}\sim P(z_{t+1}|s_{t+1})}\Big[(1-b_{t+1})V(s_{t+1}, z_t) \\ &\qquad + b_{t+1}V(s_{t+1}, z'_t)\Big]\Big] \\ V^{\text{lower}}(s_t, z_t) &= \log \int P_{\text{prior}}(a_t|s_t, z_t)\exp\left(Q(s_t, a_t, z_t)\right)\ \mathrm{d}a_t \end{aligned} \tag{176}$$

The mistake made in this equation is when we doesn't consider the KL-regularization of the master policy $\pi^H_\phi(z_t|s_t)$ where

$$\begin{aligned} V^H(s_t) &= \mathbb{E}_{z_t^*}\left[\mathbb{E}_{a_t^*\sim\pi^*(a_t|s_t, z_t)}\left[Q(s_t, a_t, z_t)\right] - \log\frac{\pi^H(z_t|s_t)}{P_{\text{prior}}(z_t|s_t)}\right] \\ &= \mathbb{E}_{z_t^*}\left[\mathbb{E}_{a_t}\left[V(s_t, z_t) + \log\frac{\pi_\theta(a_t|s_t, z_t)}{P_{\text{prior}}(a_t|s_t, z_t)}\right] - \log\frac{\pi^H(z_t|s_t)}{P_{\text{prior}}(z_t|s_t)}\right] \end{aligned} \tag{177}$$

while the equation 176 gives the impression that

$$V^{\text{lower}}(s_t, z_t) = \mathbb{E}_{a_t^*\sim\pi^*(a_t|s_t, z_t)}\left[Q(s_t, a_t, z_t)\right] - \log\frac{\pi_\theta(a_t|s_t, z_t)}{P_{\text{prior}}(a_t|s_t, z_t)} \tag{178}$$

Solving further the equation 176, we have the following equation for the master policy:

$$\begin{aligned} \pi^H(z_t|s_t) &= \frac{\frac{1}{m}\exp\left(V^{\text{lower}}(s_t, z_t)\right)}{\exp V(s_t)} \\ \text{where } V(s_t) &= \frac{1}{m}\log\int \exp\left(V^{\text{lower}}(s_t, z_t)\right)\ \mathrm{d}z_t \end{aligned} \tag{179}$$

Given this, we can see that the expectation of $\mathbb{E}_{z_t}\left[V^{\text{lower}}(s_t, z_t)\right]$ isn't the same as the regularized reward, since we are missing such a term, as we can show that

$$V(s_t) = \mathbb{E}_{z_t}\left[V^{\text{lower}}(s_t, z_t) - \log\frac{\pi^H(z_t|s_t)}{P_{\text{prior}}(z_t|s_t)}\right] \tag{180}$$

In conclusion, if we replace $V(s_{t+1}, z_t)$ with $V(s_{t+1})$ recover the solution in section before, while having easier/more intuitive way of solving, as we can solve the equation backward starting with lower-level policy.

B.2 MA-HRL

Starting with the objective, we have the following ELBO to solve at the final time step T. Note that solving this requires standard techniques:

$$\begin{aligned} \mathbb{E}_q\Bigg[\beta R(s_T, a_T^i, a_T^{-i}) &- \log\frac{\pi_\theta(a_T^i|s_T, z_T^i, a_T^{-i}, z_T^{-i})}{P_{\text{prior}}(a_T^i|s_T, z_T^i, a_T^{-i}, z_T^{-i})} - \log\frac{\rho_\theta(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)}{P_{\text{prior}}(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)} - \log\frac{q^{H,i}(z_T^i|s_T, z_{T-1}^i, b_T^i, a_T^{-i}, z_T^{-i})}{P^{H,i}_{\text{prior}}(z_T^i|s_T, z_{T-1}^i, b_T^i, a_T^{-i}, z_T^{-i})} \\ &- \log\frac{q^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^i)}{P^{H,-i}_{\text{prior}}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^i)} - \log\frac{q^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})}{P^{T,i}_{\text{prior}}(b_T^i|s_T, z_{T-1}^i, a_T^{-i})} - \log\frac{q^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})}{P^{T,-i}_{\text{prior}}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})}\Bigg] \end{aligned}$$

Starting with the definition of the policy, we can see that when solving it we have

$$\pi_\theta(a_T^i|s_T, z_T^i, a_T^{-i}, z_T^{-i}) = \frac{P_{\text{prior}}(a_T^i|s_T, z_T^i, a_T^{-i}, z_T^{-i}) \exp\left(Q(a_T^i, s_T, z_T^i, a_T^{-i}, z_T^{-i})\right)}{\exp Q^H(s_T, z_T^i, a_T^{-i}, z_T^{-i})}$$
$$\text{where } Q(a_T^i, s_T, z_T^i, a_T^{-i}, z_T^{-i}) = \beta R(s_T, a_T^i, a_T^{-i})$$
$$Q^H(s_T, z_T^i, a_T^{-i}, z_T^{-i}) = \int P_{\text{prior}}(a_T^i|s_T, z_T^i, a_T^{-i}, z_T^{-i}) \exp\left(Q(a_T^i, s_T, z_T^i, a_T^{-i}, z_T^{-i})\right) \, \mathrm{d}a_T^i$$

Plugging it back and solve for the master policy, we have the following problem

$$\begin{aligned} b_T^i \Bigg[& Q^H(s_T, z_T^i, a_T^{-i}, z_T^{-i}) - \log \frac{\rho_\theta(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)}{P_{\text{prior}}(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)} - \log \frac{q^{H,i}(z_T^i|s_T, a_T^{-i}, z_T^{-i})}{P_{\text{prior}}^{H,i}(z_T^i|s_T, a_T^{-i}, z_T^{-i})} \\ & - \log \frac{q^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i})}{P_{\text{prior}}^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i})} - \log \frac{q^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})}{P_{\text{prior}}^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i})} - \log \frac{q^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})}{P_{\text{prior}}^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})} \Bigg] \\ + (1 - b_T^i) \Bigg[& Q^H(s_T, z_T^i, a_T^{-i}, z_T^{-i}) - \log \frac{\rho_\theta(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)}{P_{\text{prior}}(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)} - \log \frac{q^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i})}{P_{\text{prior}}^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i})} \\ & - \log \frac{q^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})}{P_{\text{prior}}^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i})} - \log \frac{q^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})}{P_{\text{prior}}^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})} \Bigg] \end{aligned} \tag{181}$$

Consider solving for generative model for option, which we have the following

$$\begin{aligned} q^{H,i}(z_T^i|s_T, a_T^{-i}, z_T^{-i}) &= \frac{P_{\text{prior}}^{H,i}(z_T^i|s_T, a_T^{-i}, z_T^{-i}) \exp(Q^H(s_T, z_T^i, a_T^{-i}, z_T^{-i}))}{\exp Q(s_T, a_T^{-i}, z_T^{-i})} \\ \text{where } Q(s_T, a_T^{-i}, z_T^{-i}) &= \log \int P_{\text{prior}}^{H,i}(z_T^i|s_T, a_T^{-i}, z_T^{-i}) \exp(Q^H(s_T, z_T^i, a_T^{-i}, z_T^{-i})) \, \mathrm{d}z_T^i \end{aligned} \tag{182}$$

Plugging this back into the equation, we left with the following problem with the termination policy:

$$\begin{aligned} \mathbb{E}_q \Bigg[& b_T^i \left[Q(s_T, a_T^{-i}, z_T^{-i})\right] + (1 - b_T^i) \left[Q(s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})\right] - \log \frac{q^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})}{P_{\text{prior}}^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})} \\ & - \log \frac{q^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i})}{P_{\text{prior}}^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i})} \log \frac{\rho_\theta(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)}{P_{\text{prior}}(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)} - \log \frac{q^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})}{P_{\text{prior}}^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, a_T^{-i})} \Bigg] \end{aligned}$$

Note that for $q^{T,i}$ we will based our expectation of z_T^i on our optimized $q^{H,i}$, meaning that z_T^i is either based on newly generated or the old one from time-step before. We have the following policy

$$q^{T,i}(b_T^i|s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i}) = \frac{P_{\text{prior}}(b_T^i|s_T, z_{T-1}^i, z_T^{-i}) \exp Q(b_T^i, s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})}{\exp Q(s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})}$$
$$\text{where } Q(b_T^i, s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i}) = b_T^i \left[Q(s_T, a_T^{-i}, z_T^{-i})\right] + (1 - b_T^i) \left[Q(s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i})\right]$$
$$Q(s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i}) = \log \int P_{\text{prior}}(b_T^i|s_T, z_{T-1}^i, z_T^{-i}) \exp Q(b_T^i, s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i}) \, \mathrm{d}b_T^i$$

Given this, we have the following equation left for the opponent's model, this leads to following equation

$$\begin{aligned} \mathbb{E}_q \Bigg[& Q(s_T, z_{T-1}^i, a_T^{-i}, z_T^{-i}) - \log \frac{q^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i}, z_{T-1}^i)}{P_{\text{prior}}^{H,-i}(z_T^{-i}|s_T, z_{T-1}^{-i}, b_T^{-i}, z_{T-1}^i)} - \log \frac{\rho_\theta(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)}{P_{\text{prior}}(a_T^{-i}|s_T, z_T^{-i}, z_{T-1}^i)} \\ & - \log \frac{q^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, z_{T-1}^i)}{P_{\text{prior}}^{T,-i}(b_T^{-i}|s_T, z_{T-1}^{-i}, z_{T-1}^i)} \Bigg] \end{aligned} \tag{183}$$

Starting with the opponent's policy based on $Q(s_T, z^i_{T-1}, a^{-i}_T, z^{-i}_T)$, we have the following

$$\begin{aligned}\rho_\theta(a^{-i}_T|s_T, z^{-i}_T, z^i_{T-1}) &= \frac{P_{\text{prior}}(a^{-i}_T|s_T, z^{-i}_T, z^i_{T-1}) \exp Q(s_T, z^i_{T-1}, a^{-i}_T, z^{-i}_T)}{\exp Q^H(s_T, z^i_{T-1}, z^{-i}_T)} \\ \text{where } Q^H(s_T, z^i_{T-1}, z^{-i}_T) &= \log \int P_{\text{prior}}(a^{-i}_T|s_T, z^{-i}_T, z^i_{T-1}) \exp Q^H(s_T, z^i_{T-1}, z^{-i}_T) \ \mathrm{d}a^i_T\end{aligned} \tag{184}$$

Now, let's split the termination policy, to solve the master policy.

$$\begin{aligned}\mathbb{E}_q \Bigg[b^i_T \Bigg[Q^H(s_T, z^i_{T-1}, z^{-i}_T) - \log \frac{q^{H,-i}(z^{-i}_T|s_T, z^i_{T-1})}{P^{H,-i}_{\text{prior}}(z^{-i}_T|s_T, z^i_{T-1})} - \log \frac{q^{T,-i}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1})}{P^{T,-i}_{\text{prior}}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1})} \Bigg] \\ (1 - b^i_T) \left[Q^H(s_T, z^i_{T-1}, z^{-i}_{T-1}) - \log \frac{q^{T,-i}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1})}{P^{T,-i}_{\text{prior}}(1 - b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1})} \right] \Bigg]\end{aligned} \tag{185}$$

We, therefore, having the following master policy as:

$$\begin{aligned}q^{H,-i}(z^{-i}_T|s_T, z^i_{T-1}) &= \frac{P_{\text{prior}}(z^{-i}_T|s_T, z^i_{T-1}) \exp\left(Q^H(s_T, z^i_{T-1}, z^{-i}_T)\right)}{\exp Q(s_T, z^i_{T-1})} \\ \text{where } Q(s_T, z^i_{T-1}) &= \log \int P_{\text{prior}}(z^{-i}_T|s_T, z^i_{T-1}) \exp\left(Q^H(s_T, z^i_{T-1}, z^{-i}_T)\right) \ \mathrm{d}z^{-i}_T\end{aligned} \tag{186}$$

Finally, we have the following objective left for the opponent condition terminating condition

$$\mathbb{E}_q \left[b^{-i}_T \left[Q(s_T, z^i_{T-1}) \right] + (1 - b^{-i}_T) \left[Q^H(s_T, z^i_{T-1}, z^{-i}_{T-1}) \right] - \log \frac{q^{T,-i}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1})}{P^{T,-i}_{\text{prior}}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1})} \right] \tag{187}$$

Given this, we have the following termination condition

$$\begin{aligned}q^{T,-i}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1}) &= \frac{P_{\text{prior}}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1}) \exp\left(Q(b^{-i}_T, s_T, z^i_{T-1}, z^{-i}_T)\right)}{\exp Q(s_T, z^i_{T-1}, z^{-i}_T)} \\ \text{where } Q(b_T, s_T, z^i_{T-1}, z^{-i}_{T-1}) &= b^i_T \left[Q(s_T, z^i_{T-1}) \right] + (1 - b^i_T) \left[Q^H(s_T, z^i_{T-1}, z^{-i}_{T-1}) \right] \\ Q(s_T, z^i_{T-1}, z^{-i}_{T-1}) &= \log \int P_{\text{prior}}(b^{-i}_T|s_T, z^{-i}_{T-1}, z^i_{T-1}) \exp\left(Q(b^{-i}_T, s_T, z^i_{T-1}, z^{-i}_{T-1})\right) \ \mathrm{d}b^{-i}_T\end{aligned} \tag{188}$$

Finally, we are left with the message of $Q(s_T, z^i_{T-1}, z^{-i}_{T-1})$. Now, we shall consider the update at arbirary time step t, which lead to the following equation:

$$\begin{aligned}\mathbb{E}_q \Bigg[\beta R(s_t, a^i_t, a^{-i}_t) &- \log \frac{\pi_\theta(a^i_T|s_T, z^i_T, a^{-i}_T, z^{-i}_T)}{P_{\text{prior}}(a^i_t|s_t, z^i_t, a^{-i}_t, z^{-i}_T)} - \log \frac{\rho_\theta(a^{-i}_T|s_T, z^{-i}_T, z^i_{T-1})}{P_{\text{prior}}(a^{-i}_T|s_T, z^{-i}_T, z^i_{T-1})} - \log \frac{q^{H,i}(z^i_T|s_T, z^i_{T-1}, b^i_T, a^{-i}_T, z^{-i}_T)}{P^{H,i}_{\text{prior}}(z^i_T|s_T, z^i_{T-1}, b^i_T, a^{-i}_T, z^{-i}_T)} \\ &- \log \frac{q^{H,-i}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t)}{P^{H,-i}_{\text{prior}}(z^{-i}_t|s_t, z^{-i}_{t-1}, b^{-i}_t)} - \log \frac{q^{T,i}(b^i_T|s_T, z^i_{T-1}, a^{-i}_T, z^{-i}_T)}{P^{T,i}_{\text{prior}}(b^i_t|s_t, z^i_{t-1}, a^{-i}_t, z^{-i}_T)} - \log \frac{q^{T,-i}(b^{-i}_t|s_t, z^{-i}_{t-1}, a^{-i}_t)}{P^{T,-i}_{\text{prior}}(b^{-i}_t|s_t, z^{-i}_{t-1}, a^{-i}_t)} \\ &+ \gamma \mathbb{E}_{s_{t+1}} \left[Q(s_{t+1}, z^i_t, z^{-i}_t) \right] \Bigg]\end{aligned} \tag{189}$$

Setting the Bellman equation to be in the following form

$$Q(s_t, a^i_t, a^{-i}_t, z^i_t, z^{-i}_t) = \beta R(s_t, a^i_t, a^{-i}_t) + \gamma \mathbb{E}_{s_{t+1}} \left[Q(s_{t+1}, z^i_t, z^{-i}_t) \right] \tag{190}$$

Given this, we shall re-do all the solving given the definition of $Q(s_t, a^i_t, a^{-i}_t, z^i_t, z^{-i}_t)$, which will lead to the solution that we want.

C Part 3: Appendix

C.1 ROMMEO EM

C.1.1 Action Posterior

Joint Posterior We will start with the inference of s, a^i, a^{-i}, which is simply:

$$\begin{aligned}
P(s, a^i, a^{-i}|\mathcal{O}^i = 1) &= \frac{P(\mathcal{O}^i = 1|s, a^i, a^{-i})P(s, a^i, a^{-i})}{P(\mathcal{O}^i = 1)} \\
&= \frac{P(\mathcal{O}^i = 1|s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)}{\int P(\mathcal{O}^i = 1|s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}} \\
&= \frac{\exp Q^i_\omega(s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)}{\int \exp Q^i_\omega(s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}} \\
&= \frac{\exp Q^i_\omega(s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)}{\int \exp Q^i_\omega(s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}}
\end{aligned}$$

Policy Posterior We follows the product rule of probability to arrive at the final posterior over the agent:

$$\begin{aligned}
P(a^i|s, a^{-i}, \mathcal{O}^i = 1) &= \frac{P(s, a^i, a^{-i}|\mathcal{O}^i = 1)}{\int P(s, a^i, a^{-i}|\mathcal{O}^i = 1)\ \mathrm{d}a^i} \\
&= \frac{P(\mathcal{O}^i = 1|s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)}{\int P(\mathcal{O}^i = 1|s, a^i, a^{-i})\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P_{\text{prior}}(a^{-i}|s)\ \mathrm{d}a^i} \\
&= \frac{P(\mathcal{O}^i = 1|s, a^i, a^{-i})\cancel{\mathcal{U}(s)}P_{\text{prior}}(a^i|s, a^{-i})\cancel{P_{\text{prior}}(a^{-i}|s)}}{\cancel{P_{\text{prior}}(a^{-i}|s)}\cancel{\mathcal{U}(s)}\int P(\mathcal{O}^i = 1|s, a^i, a^{-i})P_{\text{prior}}(a^i|s, a^{-i})\ \mathrm{d}a^i} \\
&= \frac{\exp\left(Q^i_\omega(s, a^i, a^{-i})\right)P_{\text{prior}}(a^i|s, a^{-i})}{\int \exp\left(Q^i_\omega(s, a^i, a^{-i})\right)P_{\text{prior}}(a^i|s, a^{-i})\ \mathrm{d}a^i}
\end{aligned}$$

C.1.2 ELBO Derivation

The ELBO is the same as the

$$\begin{aligned}
\mathcal{L}(\theta, \omega) &= \int q_\theta(s, a^i, a^{-i}) \log\left(\frac{P_\omega(\mathcal{O}^i = 1, s, a^i, a^{-i})}{q_\theta(s, a^i, a^{-i})}\right)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i} \\
&= \mathbb{E}_q\left[\log P_\omega(\mathcal{O}^i = 1|s, a^i, a^{-i})\right] - D_{\text{KL}}\left(q_\theta(s, a^i, a^{-i})\middle\| P_\omega(s, a^i, a^{-i})\right) \\
&= \mathbb{E}_q\left[Q^i_\omega(s, a^i, a^{-i})\right] - D_{\text{KL}}\left(\pi_\theta(a^i|s, a^{-i})\middle\| P_{\text{prior}}(a^i|s, a^{-i})\right) \\
&\qquad - D_{\text{KL}}\left(\rho_\theta(a^i|s)\middle\| P_{\text{prior}}(a^{-i}|s)\right)
\end{aligned} \tag{191}$$

C.2 General Reward EM

C.2.1 Agent Action Posterior

Joint Posterior We will follows the same procedure as the ROMMEO EM derivation starting by finding joint probability. Note that

$$\begin{aligned}
P(a^i, s, a^{-i}, \mathcal{O}^{-i} = 1|\mathcal{O}^i = 1) &= \frac{P(\mathcal{O}^i = 1|a^i, s, a^{-i}, \mathcal{O}^{-i} = 1)P(a^i, s, a^{-i}, \mathcal{O}^{-i} = 1)}{P(\mathcal{O}^i = 1)} \\
&= \frac{P(\mathcal{O}^i = 1|a^i, s, a^{-i}, \mathcal{O}^{-i} = 1)\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P^*(a^{-i}|s, \mathcal{O}^{-i} = 1)P(\mathcal{O}^{-i} = 1)}{\int P(\mathcal{O}^i = 1|a^i, s, a^{-i}, \mathcal{O}^{-i} = 1)\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P^*(a^{-i}|s, \mathcal{O}^{-i} = 1)P(\mathcal{O}^{-i} = 1)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}\ \mathrm{d}\mathcal{O}^{-i}} \\
&= \frac{\left(Q^i_\omega(s, a^i, a^{-i})\right)\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P^*(a^{-i}|s, \mathcal{O}^{-i} = 1)P(\mathcal{O}^{-i} = 1)}{\int \exp\left(Q^i_\omega(s, a^i, a^{-i})\right)\mathcal{U}(s)P_{\text{prior}}(a^i|s, a^{-i})P^*(a^{-i}|s, \mathcal{O}^{-i} = 1)P(\mathcal{O}^{-i} = 1)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}\ \mathrm{d}\mathcal{O}^{-i}}
\end{aligned}$$

Policy Posterior We simply apply the product rule to the joint posterior.

$$
\begin{aligned}
P(a^i|s,a^{-i},\mathcal{O}^{-i}=1,\mathcal{O}^i=1) &= \frac{P(a^i,s,a^{-i},\mathcal{O}^{-i}=1|\mathcal{O}^i=1)}{\int P(a^i,s,a^{-i},\mathcal{O}^{-i}=1|\mathcal{O}^i=1)\ \mathrm{d}a^i} \\
&= \frac{P(\mathcal{O}^i=1|a^i,s,a^{-i},\mathcal{O}^{-i}=1)\mathcal{U}(s)P_{\text{prior}}(a^i|s,a^{-i})P^*(a^{-i}|s,\mathcal{O}^{-i}=1)P(\mathcal{O}^{-i}=1)}{\int P(\mathcal{O}^i=1|a^i,s,a^{-i},\mathcal{O}^{-i}=1)\mathcal{U}(s)P_{\text{prior}}(a^i|s,a^{-i})P^*(a^{-i}|s,\mathcal{O}^{-i}=1)P(\mathcal{O}^{-i}=1)\ \mathrm{d}a^i} \\
&= \frac{P(\mathcal{O}^i=1|a^i,s,a^{-i},\mathcal{O}^{-i}=1)\cancel{\mathcal{U}(s)}P_{\text{prior}}(a^i|s,a^{-i})\cancel{P^*(a^{-i}|s,\mathcal{O}^{-i}=1)}\cancel{P(\mathcal{O}^{-i}=1)}}{\cancel{\mathcal{U}(s)}\cancel{P^*(a^{-i}|s,\mathcal{O}^{-i}=1)}\cancel{P(\mathcal{O}^{-i}=1)}\int P(\mathcal{O}^i=1|a^i,s,a^{-i},\mathcal{O}^{-i}=1)P_{\text{prior}}(a^i|s,a^{-i})\ \mathrm{d}a^i} \\
&= \frac{Q^i_\omega(s,a^i,a^{-i})P_{\text{prior}}(a^i|s,a^{-i})}{\int Q^i_\omega(s,a^i,a^{-i})P_{\text{prior}}(a^i|s,a^{-i})\ \mathrm{d}a^i}
\end{aligned}
\tag{192}
$$

C.2.2 Opponent Action Posterior

Joint Posterior We will follows the same procedure as the agent:

$$
\begin{aligned}
P(a^i,s,a^{-i},\mathcal{O}^i=1|\mathcal{O}^{-i}=1) &= \frac{P(\mathcal{O}^{-i}=1|a^i,s,a^{-i},\mathcal{O}^i=1)P(a^i,s,a^{-i},\mathcal{O}^i=1)}{P(\mathcal{O}^{-i}=1)} \\
&= \frac{P(\mathcal{O}^{-i}=1|a^i,s,a^{-i},\mathcal{O}^i=1)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)}{\int P(\mathcal{O}^{-i}=1|a^i,s,a^{-i},\mathcal{O}^i=1)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}\ \mathrm{d}\mathcal{O}^i} \\
&= \frac{P(\mathcal{O}^{-i}=1|s,a^{-i},\mathcal{O}^i=1)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)}{\int P(\mathcal{O}^{-i}=1|s,a^{-i},\mathcal{O}^i=1)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}\ \mathrm{d}\mathcal{O}^i} \\
&= \frac{\exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)}{\int \exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)\ \mathrm{d}s\ \mathrm{d}a^i\ \mathrm{d}a^{-i}\ \mathrm{d}\mathcal{O}^i} \\
&= \frac{1}{Z}\exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)
\end{aligned}
$$

In the third equality, we can see that the optimality of the opponent model will not be dependent of the agent's action. Now, let's consider the opponent's policy posterior

$$
\begin{aligned}
P(a^{-i}|s,\mathcal{O}^{-i}=1,\mathcal{O}^i=1) &= \frac{\int P(a^i,s,a^{-i},\mathcal{O}^i=1|\mathcal{O}^{-i}=1)\ \mathrm{d}a^i}{\int P(a^i,s,a^{-i},\mathcal{O}^i=1|\mathcal{O}^{-i}=1)\ \mathrm{d}a^i\ \mathrm{d}a^{-i}} \\
&= \frac{\cancel{\frac{1}{Z}}\int \exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)\ \mathrm{d}a^i}{\cancel{\frac{1}{Z}}\int \exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)\mathcal{U}(s)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)P(\mathcal{O}^i=1)\ \mathrm{d}a^i\ \mathrm{d}a^{-i}} \\
&= \frac{\cancel{\mathcal{U}(s)}\cancel{P(\mathcal{O}^i=1)}\int \exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)\ \mathrm{d}a^i}{\cancel{\mathcal{U}(s)}\cancel{P(\mathcal{O}^i=1)}\int \exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)P(a^{-i}|s)P(a^i|a^{-i},s,\mathcal{O}^i=1)\ \mathrm{d}a^i\ \mathrm{d}a^{-i}} \\
&= \frac{\exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)P(a^{-i}|s)\int P(a^i|a^{-i},s,\mathcal{O}^i=1)\ \mathrm{d}a^i}{\int \exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)P(a^{-i}|s)\int P(a^i|a^{-i},s,\mathcal{O}^i=1)\ \mathrm{d}a^i\ \mathrm{d}a^{-i}} \\
&= \frac{Q^{-i}_{\psi,\theta}(s,a^{-i})P(a^{-i}|s)}{\int \exp\left(Q^{-i}_{\psi,\theta}(s,a^{-i})\right)P(a^{-i}|s)\ \mathrm{d}a^{-i}}
\end{aligned}
$$

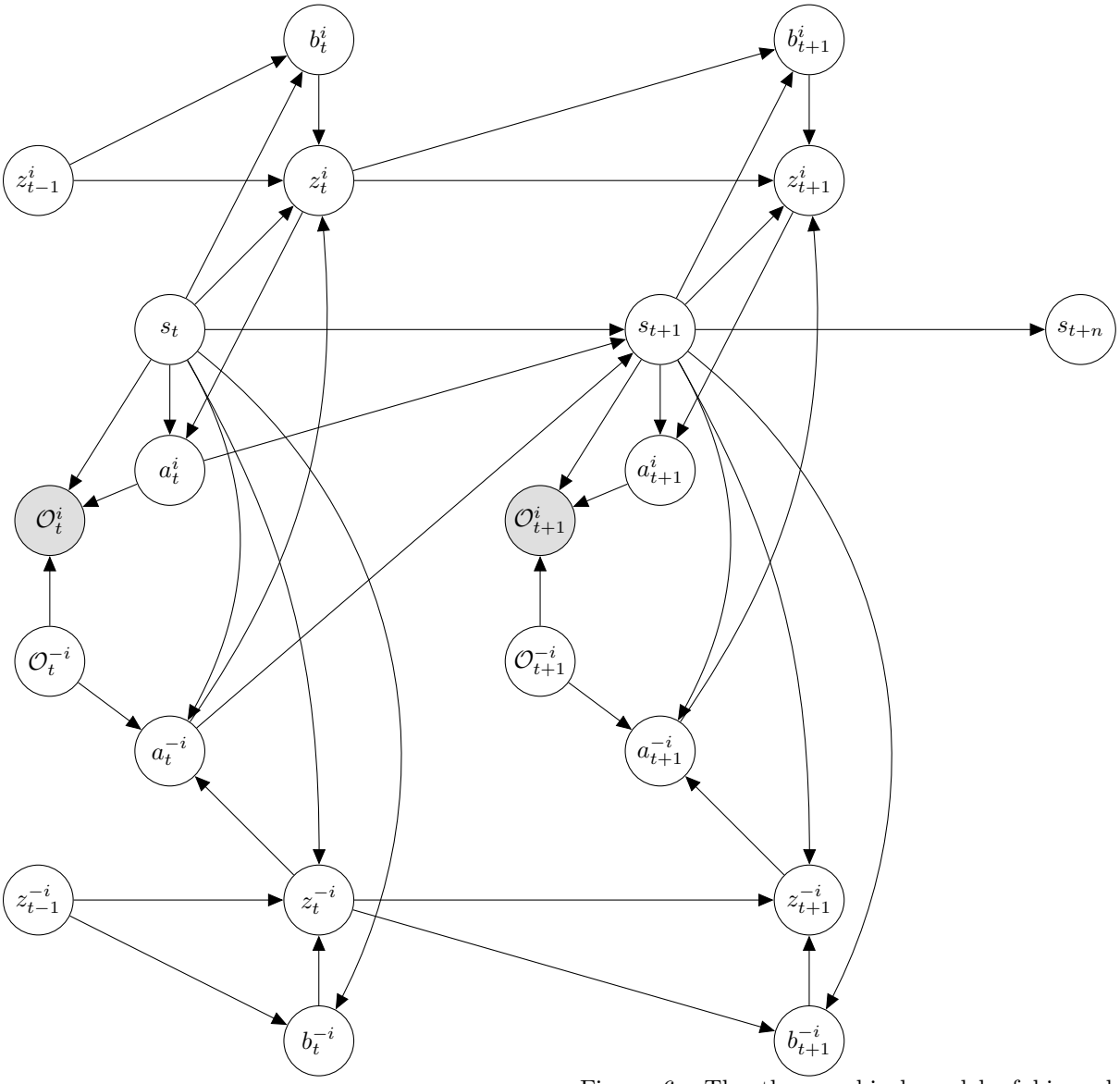

Figure 6: The the graphical model of hierarchical agent. The "main" controller is indeed the policy, while making the lower-level works, rather than also best response to opponent's action

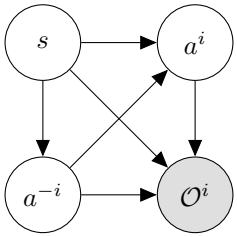

Figure 7: The the graphical model of EM agent. As we will consider the most basic case of ROMMEO where the agent's policy is depending on other agents

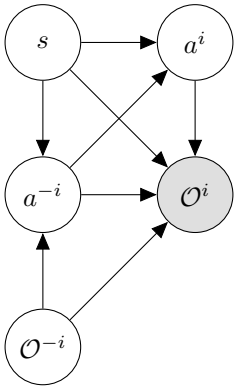

Figure 8: The the graphical model of EM agent. We will consider the case in which we are aware of the opponent's optimal model

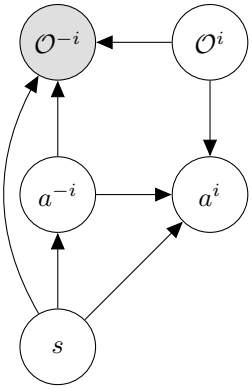

Figure 9: This is the similar from the one proposed before. As it has similar property, we are going to solve them in slightly difference manners, which we also assume that the opponent model aware that the agent is optimal (this structure will fits the training scheme that outlined in EM-training, which repeatedly assign a new prior.)

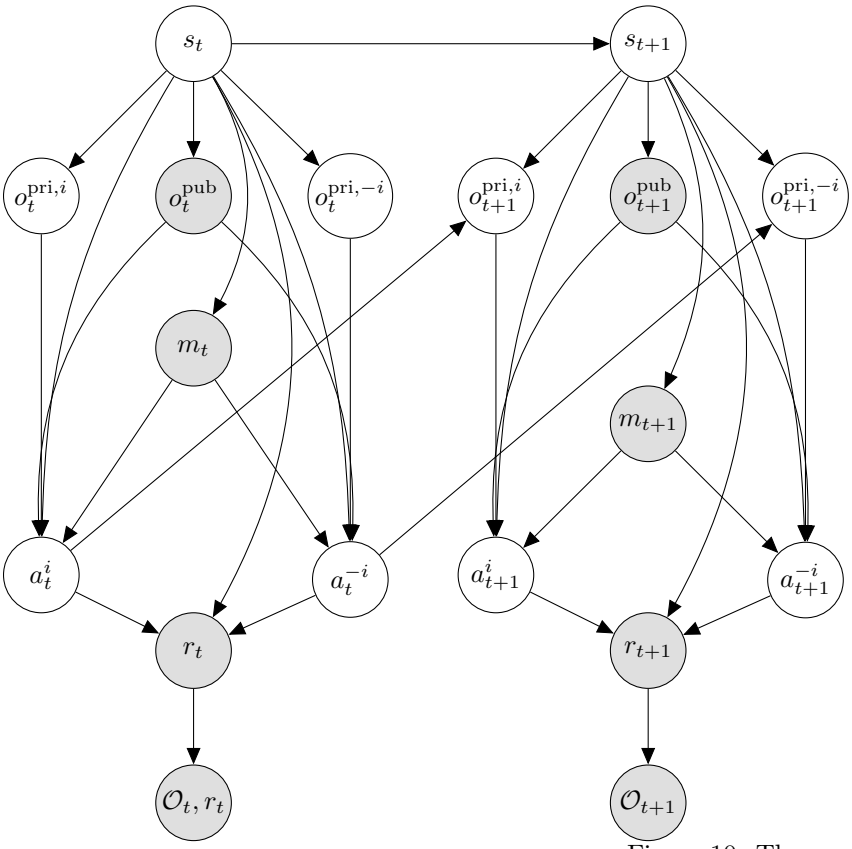

Figure 10: The graphical model representing the public belief POMDP o^{pub} representing the public observation, which depends on the current state and agents' action. We left the state transition function $P(s_{t+1}|s_t, a_t)$ to reduce the clusters. We assume a cooperative setting in which each agents (including the public belief agent) trying to optimize. The reason we split between reward and optimality is because the reward will depends on unknown state hence but when we are training agent we didn't have a knowledge of.

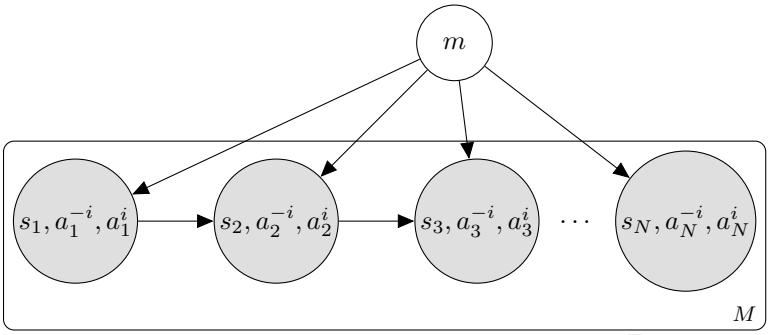

Figure 11: This is the same setting as the one proposed in [19]. As we have a sequence of opponent's trajectories and state, while the opponent is depends on the latent variable z