LEARNING META REPRESENTATIONS FOR AGENTS IN MULTI-AGENT REINFORCEMENT LEARNING

Shenao Zhang
Georgia Institute of Technology
shenao@gatech.edu

Li Shen & Lei Han
Tencent AI Lab & Tencent Robtics X
{lshen.lsh,leihan.cs}@gmail.com

Li Shen
JD Explore Academy
mathshenli@gmail.com

ABSTRACT

In multi-agent reinforcement learning, the behaviors that agents learn in a single Markov Game (MG) are typically confined to the given agent number. Every single MG induced by varying population sizes may possess distinct optimal joint strategies and game-specific knowledge, which are modeled independently in modern multi-agent algorithms. In this work, we focus on creating agents that generalize across population-varying MGs. Instead of learning a unimodal policy, each agent learns a policy set that is formed by effective strategies across a variety of games. We propose *Meta Representations for Agents* (MRA) that explicitly models the game-common and game-specific strategic knowledge. By representing the policy sets with multi-modal latent policies, the common strategic knowledge and diverse strategic modes are discovered with an iterative optimization procedure. We prove that as an approximation to a constrained mutual information maximization objective, the learned policies can reach Nash Equilibrium in every evaluation MG under the assumption of Lipschitz game on a sufficiently large latent space. When deploying it at practical latent models with limited size, fast adaptation can be achieved by leveraging the first-order gradient information. Extensive experiments show the effectiveness of MRA on both training performance and generalization ability in hard and unseen games.

1 INTRODUCTION

Behaviors of agents learned in a single Markov Game (MG) highly depend on the environmental settings, especially the number of agents, *i.e.*, population size Suarez et al. (2019); Long* et al. (2020). Many multi-agent reinforcement learning (MARL) algorithms Sukhbaatar et al. (2016); Foerster et al. (2016); Lowe et al. (2017) are developed in games with fixed population sizes. However, the algorithms may suffer from generalization issues, *i.e.*, the policies learned in a single MG are brittle to the change of game setting Suarez et al. (2019). Recent works have experimentally shown the benefit of knowledge transfer between MGs with different population sizes Agarwal et al. (2019); Long* et al. (2020), which is required to perform between successive games. Unfortunately, the resulting agents are still confined to particular training games, with less ability for extrapolation.

In this work, we are concerned with learning multi-agent policies that generalize across Markov Games constructed by varying the population from the same underlying environment. The created agents are expected to behave well in both training and novel (or unseen) evaluation MGs. However, optimizing one unimodal policy even for maximizing the performance of the entire *training* set is still challenging Teh et al. (2017). Effective policies in population-varying games, *e.g.*, the ones that achieve Nash Equilibrium in each game, may behave dramatically different due to the game-specific strategic knowledge of themselves. Such discrepancy will hamper the performance in individual games Brunskill & Li (2013). In this regard, it is desirable to learn *sets of policies* that contain the optimal strategies for each training MG, while transferring knowledge to *unseen* MGs is still challenging nevertheless.

To cope with this challenge, we explicitly model the *MG-specific* and *MG-common* strategic knowledge. In unseen games, although the optimal game-specific knowledge that leads to optimal policies is unobtainable, the common strategic knowledge and various strategic modes can still be captured during training by imposing *knowledge variations*, *i.e.*, the *suboptimal* game-specific knowledge. By learning to make the smartest decisions under multiple *imagined* variations instead of only fitting the best response, the strategic knowledge is learned in an unsupervised manner and agents can effectively generalize to novel MGs.

Since games induced by varying populations possess distinct optimal policy behaviors characterized by different (egocentric) strategic relationships, we model the game-specific knowledge as such relationship. For example, the optimal game-specific knowledge for PacMan agents in a ghost-dominant game is to focus on ghost agents for survival. However, this leads to PacMan agents that are unaware of eating more food if evaluated in other games, *e.g.*, a PacMan-dominant game. By additionally impelling the PacMan to pay more attention to other competing PacMan, although the resulting policy may be suboptimal in the training ghost-dominant game, the common knowledge can be learned, which involves eluding ghosts while moving towards food that is with less competition.

We propose *Meta Representations for Agents* (MRA) to discover the underlying strategic structures in the games. By meta-representing the policy sets with multi-modal latent policies, the game-common strategic knowledge and diverse policy modes are captured with an iterative optimization procedure. We prove that as an approximation to a constrained mutual information maximization objective, the latent policies can reach Nash Equilibrium in every evaluation MG under Lipschitz game assumption and on a large latent space. When with practical limited-size latent models, fast adaptation is achieved by leveraging first-order gradient information. We further empirically validate the benefits of MRA, which is capable of boosting training performance and extrapolating over a variety of unseen MGs.

2 PRELIMINARIES

Game: An N-agent Markov Game is defined by state set $\mathcal{S}$, action sets $\{\mathcal{A}_1, \dots, \mathcal{A}_N\}$, and observation sets $\{\mathcal{O}_1, \dots, \mathcal{O}_N\}$. For every agent i, $o^i \in \mathcal{O}_i$ is an observation of the global state $s \in \mathcal{S}$. State transition and per-agent reward function are defined as $\mathcal{P}(\mathcal{S}, \mathcal{A}_1, \dots, \mathcal{A}_N, \mathcal{S}')$ and $\mathcal{R}_i : \mathcal{S} \times \mathcal{A}_i \to [0, 1]$, respectively. The joint strategy is denoted as $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N) = (\pi^i, \boldsymbol{\pi}^{\text{-}\boldsymbol{i}})$, where π^i is the strategy of agent i and $\boldsymbol{\pi}^{\text{-}\boldsymbol{i}}$ is the joint strategy excluding it.

In this work, we consider role-symmetric MGs Suarez et al. (2019); Muller et al. (2020), where homogeneous agents are with the same reward function and action space. The type number of homogeneous agents is denoted as h, *e.g.*, $h = 2$ for Pac-Man and ghosts in any Pac-Man game.

Relational Representation is an opponent modeling framework to capture the strategic relationship between agents and output embedding e for further policy or value function learning Long* et al. (2020); Agarwal et al. (2019); Iqbal & Sha (2018). Consider the observation o^i of agent i with entities $o^i = \left[o^i_s, o^i_1, \dots, o^i_j, \dots, o^i_N\right]$, where o^i_s is agent i's self properties (*e.g.*, its speed), o^i_j is agent i's observation on agent j (*e.g.*, distance from agent j), and the observed environment information (*e.g.*, landmark locations) is concatenated to these entities. Then with self-attention Vaswani et al. (2017) generating the pair-wise relation $g^{i,j}$, *i.e.*, the j-th entity of agent i's egocentric relational graph g^i, the representation embedding e^i for agent i is formulated as:

$$e^i = \sum_{j \neq i} g^{i,j} V(o^i_j), \text{where } g^{i,j} = \frac{\exp(Q(o^i_s)^\top K(o^i_j))}{\sum_{j \neq i} \exp(Q(o^i_s)^\top K(o^i_j))}. \tag{1}$$

Here, $V(\cdot)$, $Q(\cdot)$ and $K(\cdot)$ denote linear transformations. The observation embedding with an arbitrary number of agents can thus be represented with a fixed length.

Nash Equilibrium: A core concept in game theory is Nash Equilibrium (NE). When every agent in game m acts according to the joint strategy $\boldsymbol{\pi}$ at state s, the value $v^{i,m}_{\boldsymbol{\pi}}(s)$ of agent i is the expectation of i's γ-discounted cumulative reward:

$$v^{i,m}_{\boldsymbol{\pi}}(s) = \mathop{\mathbb{E}}_{\substack{\boldsymbol{a} \sim \boldsymbol{\pi}, s_0 = s \\ s_t \sim \mathcal{P}_m}} \left[\sum_t \gamma^t r^i_m(s_t, \boldsymbol{a_t})\right]. \tag{2}$$

In this work, the bold symbol is joint over all agents, and variables with superscript i are of agent i. Denote the value of the best response for agent i as $v^{*i,m}_{\boldsymbol{\pi}^{-i}}$, which is the best policy of agent i when $\boldsymbol{\pi}^{-i}$ is executed, *i.e.*, $v^{*i,m}_{\boldsymbol{\pi}^{-i}} = \max_{\pi^i} v^{i,m}_{\pi^i,\boldsymbol{\pi}^{-i}}$. Then $\boldsymbol{\pi}$ reaches NE if $\forall i \in \{1, ..., N\}, v^{i,m}_{\boldsymbol{\pi}} = v^{*i,m}_{\boldsymbol{\pi}^{-i}}$.

A common metric to measure the distance to a Nash Equilibrium is NASHCONV, which represents how much each player gains by deviating to their best response (unilaterally) in total. And it can be approximately calculated in small games Johanson et al. (2011); Lanctot et al. (2017). We denote the NASHCONV of $\boldsymbol{\pi}$ in game m as $\mathcal{D}_m(\boldsymbol{\pi})$. Then the joint strategy $\boldsymbol{\pi}$ reaches NE in m if $\mathcal{D}_m(\boldsymbol{\pi}) = 0$.

$$\mathcal{D}_m(\boldsymbol{\pi}) = \mathcal{D}_m(\pi^i, \boldsymbol{\pi}^{-i}) = \left\| \left\| v^{*i,m}_{\boldsymbol{\pi}^{-i}} - v^{i,m}_{\boldsymbol{\pi}} \right\|_{s,\infty} \right\|_{i,1} \tag{3}$$

3 LEARNING META REPRESENTATIONS FOR AGENTS

3.1 PROBLEM STATEMENT

In a single stochastic game, achieving Nash Equilibrium gives reasonable solutions and is of great importance Hu & Wellman (2003); Yang et al. (2018); Pérolat et al. (2017). To enable generalization in different MGs, the most straightforward way is to learn a joint strategy set $\boldsymbol{\Pi}$ that contains effective joint strategies for every MG, *e.g.*, the ones that achieve NE. Denote the set of all training MGs as $\mathcal{M}$ and the set of evaluation MGs as $\mathcal{M}'$. The goal is then to obtain an optimal discrete (finite) or continuous (infinite) joint strategy set $\boldsymbol{\Pi}^*$ such that:

$$\forall m' \in \mathcal{M}', \exists \boldsymbol{\pi} \in \boldsymbol{\Pi}^*, \text{ s.t. } \mathcal{D}_{m'}(\boldsymbol{\pi}) = 0. \tag{4}$$

For a satisfiable $\boldsymbol{\Pi}$, we first need its size $|\boldsymbol{\Pi}|$ to be sufficiently large to contain at least one effective strategy for every $m' \in \mathcal{M}'$. Then $\boldsymbol{\Pi}$ should be improved with respect to the worst-performing m', *i.e.*, the game with no effective strategy contained in $\boldsymbol{\Pi}$, to achieve low regret $\mathcal{D}_{m'}$ for all m'. In other words, $\boldsymbol{\Pi}$ is updated to include the joint strategy $\boldsymbol{\pi}$ that minimizes $\mathcal{D}_{m'}(\boldsymbol{\pi})$. Formally,

$$\boldsymbol{\Pi}^* = \arg\min_{\boldsymbol{\Pi}} \mathcal{L}(\boldsymbol{\Pi}), \text{ where } \mathcal{L}(\boldsymbol{\Pi}) = \min_{\boldsymbol{\pi} \sim \boldsymbol{\Pi}} \max_{m' \in \mathcal{M}'} \mathcal{D}_{m'}(\boldsymbol{\pi}). \tag{5}$$

However, minimizing $\mathcal{L}(\boldsymbol{\Pi})$ over the unseen games in $\mathcal{M}'$ is impractical in general. We cope with such intractability by introducing a heuristic algorithm and showing that the resulting objective is indeed equivalent to equation 5 and can thus lead to the optimal $\boldsymbol{\Pi}^*$ in equation 4.

3.2 RELATIONAL REPRESENTATION WITH LATENT VARIABLE POLICIES

Instead of learning independent policies to form $\boldsymbol{\Pi}$, we adopt hierarchical latent variable policies to represent the multimodality. In this way, the MG-common and MG-specific strategic knowledge can be explicitly modeled. Specifically, in the relational representation framework, the specific strategic knowledge for population-varying MGs is captured by the egocentric relational graph g since agents optimally behave in each game by learning *per-game optimal* relationship. Besides, agents take different actions when incorporating different strategic relationships, *i.e.*, multiple policy modes are obtained with varied g. Therefore, we treat g as a higher-level latent variable that is dynamically generated by $g = \phi(o, z)$. Here, z is a lower-level latent sampled from a learned distribution $p(z|m; \psi)$. Then the common knowledge that how agents optimally behave under different g can be distilled into the policy parameter θ, which includes the transform V in equation 1 and the successive policy network parameters. An agent takes action $a \sim \pi(\cdot|o, g; \theta)$, where $g = \phi(o, z)$ and $z \sim p(z|m; \psi)$.

Although the instantiation can be applied to MARL multi-task learning setups, the design of latent policies by itself does *not* immediately suggest an algorithm that generalizes to novel MGs. As such, we present the key ingredient of the Meta Representations for Agents (MRA) algorithm as follows.

3.3 GENERALIZATION BY STRATEGIC KNOWLEDGE DISCOVERY

Our core idea to enable generalization is to discover the underlying strategic structures in the underlying games. Although the effective policies in the evaluation games are never known during training, agents can still learn the common strategic knowledge and different behavioral modes solely in the *training* games in an unsupervised manner with the imposed *suboptimal* game-specific knowledge. In particular for population varying games, the agents are assigned different strategic relationships, *i.e.*, each agent pays additional attention to some agents while ignoring others. Instead of learning only one optimal joint policy, training with multiple strategic relationships enables the unsupervised discovery of behavior modes, some of which offer appreciable returns in evaluation MGs. Thus, when evaluating in novel games, the desired policy behaviors can be quickly generated by adaptation. In the extreme case that sufficiently many strategic modes are captured with an extremely large latent space, the desired policy for evaluation games can be directly found.

Specifically, agents optimally behave in each game $m \in \mathcal{M}$ with the optimal policy parameter θ^* and the (per-game) optimal relational graph g^*. By imposing *knowledge* (or *relation*) *variations* in m, *i.e.*, *multiple suboptimal* g at a certain observation, agents learn how the best decisions to accomplish the task are made, *i.e.*, learn θ^* that achieves the highest average return. With the discovery of distinct strategic modes, the MG-common knowledge contained in θ^* is obtained. Thus, when agents are in the novel MGs $m' \in \mathcal{M}'$, their policies can effectively adapt by learning the optimal relational graph in m', or achieve zero-shot transfer (without adaptation) if the latent space is large. This gives the objective of θ^i that maximizes the average return of all knowledge variations and all training MGs:

$$\max_{\theta^i} \mathcal{L}(\theta^i) = \max_{\theta^i} \mathop{\mathbb{E}}_{m\sim\mathcal{M}, g^i} \left[v^{i,m}_{\pi^i(\cdot|\cdot,g^i;\theta^i),\boldsymbol{\pi}^{\text{-}i}} \right] \tag{6}$$

Notably, equation 6 differs from the objective of multi-task learning where the average training return is maximized by learning the optimal θ^* and a *single optimal* relational graph in each game.

In order to perform well in all $m' \in \mathcal{M}'$, the strategic modes captured during training should cover as many behaviors as possible. This requires both a large latent space size $|Z|$ and *diverse* actions. For fixed $|Z|$, we introduce a diversity-driven objective that encourages high mutual information between g and a for behavior diversity, as well as between m and g to encourage game-specific knowledge learning. With high dependence between g and a, distinct variations g can generate diverse actions a.

$$\max_{\psi^i,\phi^i} \mathcal{L}(\psi^i, \phi^i) = \max_{\psi^i,\phi^i} \mathcal{I}(g^i; a^i|o^i) + \mathcal{I}(m; g^i|o^i), \tag{7}$$

where $\mathcal{I}(g; a|o) = \mathcal{H}(a|o) - \mathcal{H}(a|o, g)$ is the mutual information between g and a conditioned on observation o. With a slight abuse of notation, m denotes the basic information of game m, *e.g.*, the populations of each role. Iterative optimization of equation 6 and equation 7 is then performed.

3.4 Fast Adaptation with Limited Latent Space Size

Despite the diversity-inducing objective, we also need a large latent space $|Z|$. For $\mathbf{\Pi}$ parameterized by $\Theta = \{\psi, \phi, \theta\}$, denote $\mathbf{\Pi}_\Theta = \mathbf{\Pi}$. Specially, achieving zero-shot transfer requires $|Z| = |\mathbf{\Pi}_\Theta| \geq |\mathbf{\Pi}^*|$. However, if the settings of games in $\mathcal{M}'$ are not restricted, the size $|\mathbf{\Pi}^*|$ is unbounded. For practical limited-size latent models and unrestricted $\mathcal{M}'$, fast adaptation ability is thus desired.

The generated graph g, together with observations, actions are stored in the replay buffer. Then various knowledge variations, *i.e.*, (o, g) pair, are sampled to update θ. Compared with only $|Z|$ variations generated in an on-policy manner, this leads to a much larger size of variations. Then we use similar techniques from Reptile Nichol et al. (2018) to achieve fast adaptation.

The optimization of θ is now to perform K policy gradient steps on each individual MG, instead of on the average return (over m) in a joint training way. In game m, the objective for θ in the k-th mini-batch changes from equation 6 to $\mathcal{L}^k_m(\theta^i) = \mathbb{E}_{g^i}\left[v^{i,m}_{\pi^i(\cdot|\cdot,g^i;\theta^i),\boldsymbol{\pi}^{\text{-}i}}\right]$. Let $U^K_m(\theta)$ denote the policy parameter after K gradient steps with learning rate β. Then θ is updated by $\theta \leftarrow \theta + \alpha\Delta\theta$, where α is a hyperparameter and $\Delta\theta = U^K_m(\theta) - \theta$. By doing so, the first-order gradient information can be leveraged to update θ towards the instance-specific adapted policy parameter. Specifically, the expected policy parameter update $\mathbb{E}_k[\Delta\theta]$ over mini-batches in game m is

$$\mathbb{E}_k[\Delta\theta] = (K-1)\mathbb{E}_k\left[\mathcal{L}^k_m(\theta)\right] + \frac{(K-1)(K-2)\beta}{2}\mathbb{E}_{j,k}\left[\nabla\left(\nabla\mathcal{L}^k_m(\theta)\nabla\mathcal{L}^j_m(\theta)\right)\right], \tag{8}$$

where $\nabla\mathcal{L}^k_m(\theta)$ is the gradient at the initial θ. The derivation is in Appendix B. The second term of the RHS in equation 8 is with direction that increases the inner product between gradients of different mini-batches j, k. That is, θ is optimized not only to maximize the return under all relation variations, but also towards the place that gradients of different variations point to the same direction, *i.e.*, the place that is easy to optimize from. With this property, when the optimal strategic modes of evaluation MGs are not discovered during training, θ can still fast adapt to effective policies.

4 Analysis

In this section, we provide a theoretical analysis of MRA. We show how the tractable objective in equation 6 and equation 7 can be derived out of the primary optimization problem in equation 4.

To begin with, we introduce the Markov state transition operator $\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}$ in MG m, defined as

$$\left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}} x\right)(s) = \int_{s'\sim\mathcal{S}} x(s') \mathop{\mathbb{E}}_{\pi^i,\boldsymbol{\pi}^{-i}} \left[\mathcal{P}_m(ds'|s,a^i,\boldsymbol{a}^{-i})\right].$$

Here, $x\colon \mathcal{S} \to \mathbb{R}$ is an L_1 Lebesgue integrable function. The norm of the operator is defined as $\|\Lambda\|_{op} = \sup\{\|\Lambda x\|_{L_1(\mathcal{S})} : \|x\|_{L_1(\mathcal{S})} \le 1\}$, where $\|\cdot\|_{L_1(\mathcal{S})}$ is the L_1-norm over the state space $\mathcal{S}$.

Then we make the assumption of Lipschitz Game.

Assumption 1. *(Lipschitz Game). For any Markov Game $m \in (\mathcal{M} \cup \mathcal{M}')$, there exists a Lipschitz coefficient $\iota_m > 0$ such that for all agent in m and $s \in \mathcal{S}$:*

$$\left\|\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op} \le \iota_m \left\|\left\|\pi^{*i}(a|s) - \pi^i(a|s)\right\|_{a,1}\right\|_{s,\infty}. \tag{9}$$

Similar assumptions also appear in many previous works Liu et al. (2021); Zhang et al. (2019). We note that Assumption 1 is reasonable since the Lipschitz coefficient ι_m can be interpreted as the *influence* of agents Radanovic et al. (2019); Dimitrakakis et al. (2019), which measures how much the policy changing of an agent can affect the game environment.

Then we define a distance metric that measures the discrepancy between $\mathcal{M}$ and $\mathcal{M}'$ by comparing and computing the distance to NE in the games of the two sets. Let N_m denote the total number of agents in game m, and $h_{i,m}$ denote the homogeneous agent set of agent i in m.

Definition 1. *For two sets of MGs $\mathcal{M}$ and $\mathcal{M}'$, define the distance ς between $\mathcal{M}$ and $\mathcal{M}'$ by*

$$\varsigma = \max_{\substack{m'\in\mathcal{M}' \\ i\in\{1,\dots,N_{m'}\}}} \min_{\substack{m\in\mathcal{M}, i'\in h_{i,m} \\ \boldsymbol{\pi}\in\{\boldsymbol{\pi}|\mathcal{D}_m(\boldsymbol{\pi})=0\} \\ \boldsymbol{\pi}'\in\{\boldsymbol{\pi}'|\mathcal{D}_{m'}(\boldsymbol{\pi}')=0\}}} \mathcal{D}_{m'}(\pi^{i'}, \boldsymbol{\pi}'^{-i}). \tag{10}$$

We also define the ϵ-range joint strategy set $\hat{\boldsymbol{\Pi}}$ to guide the policy learning of agents during training.

Definition 2. *For the training MG set $\mathcal{M}$ and $\epsilon > 0$, the ϵ-range joint strategy set $\hat{\boldsymbol{\Pi}}$ is defined as:*

$$\hat{\boldsymbol{\Pi}} = \bigcup_{m\in\mathcal{M}} \hat{\boldsymbol{\Pi}}_{\boldsymbol{m}}, \text{ where } \hat{\boldsymbol{\Pi}}_{\boldsymbol{m}} = \{\boldsymbol{\pi}|\mathcal{D}_m(\boldsymbol{\pi}) \le \epsilon\}. \tag{11}$$

By bounding ϵ that characterizes a large set $\hat{\boldsymbol{\Pi}}$, equation 5 can be shown to be equivalent to a constrained mutual information maximization objective. Formally, we provide the following theorem. The variables and parameters are per agent, *e.g.*, $\boldsymbol{\pi}_\theta$ is joint over π_{θ^i}. The superscript is omitted for clarity.

Theorem 1. *If $|\boldsymbol{\Pi}_\Theta| \ge |\hat{\boldsymbol{\Pi}}|$ and ϵ satisfies $\epsilon \ge \varsigma - \min_{\iota_m,\iota_{m'}} \frac{\varsigma\gamma(\iota_{m'}-\iota_m)}{\gamma\iota_{m'}+1-\gamma}$, then with the optimal parameters $\Theta^* = \{\psi^*, \phi^*, \theta^*\}$ given by*

$$\psi^*, \phi^* = \arg\max_{\psi,\phi} \mathcal{I}(g;a|o) + \mathcal{I}(m;g|o) \text{ s.t. } \boldsymbol{\pi}_{\theta^*} \in \hat{\boldsymbol{\Pi}}, \tag{12}$$

for every evaluation Markov Game $m' \in \mathcal{M}'$, there exists a joint strategy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\Theta^*}$ that reaches Nash Equilibrium (i.e., $\boldsymbol{\Pi}_{\Theta^*} = \boldsymbol{\Pi}^*$ satisfies equation 4).

Theorem 1 suggests a general paradigm of diversity-driven learning that is effective when $\hat{\boldsymbol{\Pi}}$ satisfies certain properties. In practical MGs, however, the unknown Lipschitz coefficient and the hardness of calculating ς pose challenges to compute the satisfying ϵ. An approximation to the optimal parameters in equation 12 is to perform iterative optimization following equation 6 and equation 7.

With fixed ϕ and ψ, the objective of θ in equation 6 (greedily) maximizes the expected value over variations in order to minimize the distances to Nash of different policy modes. In other words, the distance $\mathcal{D}_m(\boldsymbol{\pi})$ of the corresponding joint strategies is minimized to satisfy $\mathcal{D}_m(\boldsymbol{\pi}) \leq \epsilon$ in the long run. Then the optimization of ϕ and ψ follows to maximize $\mathcal{I}(g;a|o) + \mathcal{I}(m;g|o)$. By iteratively improving the mutual information and updating θ towards the ϵ-range $\hat{\mathbf{\Pi}}$, the obtained solutions are close to the optimal parameters in equation 12. Besides, the condition $|\mathbf{\Pi}_\Theta| \geq |\hat{\mathbf{\Pi}}|$ in the theorem supports the intuition that a sufficiently large policy set (or latent space) is required for zero-shot transfer. However, as an approximation to the theorem, MRA also has some limitations, which we discuss and provide potential improvements in Section 7.

5 Practical Algorithm

We have shown that the iterative optimization of MRA can arise from a theoretically justified objective. In this section, we present practical implementations of the two optimization procedures. We provide the pseudocode in Appendix D. Implementation details can be found in Appendix E.

5.1 Maximization of Expected Value over Variations

The policy parameter θ^i in objective equation 6 is optimized by introducing a *centralized* critic Q_{ζ^i} for each agent i Lowe et al. (2017). Denote the target network with delayed policy and critic parameters as $\bar{\theta}$, $\bar{\zeta}$, and replay buffer as D. The parameterized critic Q_{ζ^i} is optimized to minimize:

$$\mathcal{L}(\zeta^i) = \mathop{\mathbb{E}}_{(\boldsymbol{o},\boldsymbol{a},\boldsymbol{o'},\boldsymbol{r})\sim D}\left[\left(Q_{\zeta^i}(\boldsymbol{o},\boldsymbol{a}) - y^i\right)^2\right], \text{where } y^i = r^i + \gamma \mathop{\mathbb{E}}_{\boldsymbol{a'}\sim\boldsymbol{\pi}_{\bar{\theta}}}\left[Q_{\bar{\zeta}^i}(\boldsymbol{o'},\boldsymbol{a'})\right] \tag{13}$$

Then the gradient of the policy parameter θ^i of agent i during training is given by:

$$\nabla_{\theta^i}\mathcal{L}(\boldsymbol{\pi}) = \mathop{\mathbb{E}}_{\substack{(\boldsymbol{o},\boldsymbol{g})\sim D\\ \boldsymbol{a}\sim\boldsymbol{\pi}}}\left[\nabla_{\theta^i}\log\pi_{\theta^i}(a^i|o^i,g^i)Q_{\zeta^i}(\boldsymbol{o},\boldsymbol{a})\right]. \tag{14}$$

When evaluation in a novel MG, θ^i and ϕ^i is fine-tuned to greedily maximize agents' individual rewards. Denote $\omega^i = \{\theta^i, \phi^i\}$. Then the gradient of ω^i is given by:

$$\nabla_{\omega^i}\mathcal{L}(\boldsymbol{\pi}) = \mathop{\mathbb{E}}_{\boldsymbol{o}\sim D,\boldsymbol{a}\sim\boldsymbol{\pi}}\left[\nabla_{\omega^i}\log\pi_{\theta^i}(a^i|o^i,\phi^i(o^i,z^i))Q_{\zeta^i}(\boldsymbol{o},\boldsymbol{a})\right]. \tag{15}$$

5.2 Mutual Information Maximization

In an iteration, several update steps of actor and critic are followed by mutual information $\mathcal{I}(g^i;a^i|o^i) + \mathcal{I}(m;g^i|o^i)$ maximization. According to the definition of mutual information and the non-negativeness of KL divergence, the following bound holds. And ϕ is optimized to optimize equation 16 by gradient ascent. All the derivations below are provided in Appendix C.

$$\mathcal{I}(g^i;a^i|o^i) \geq \mathop{\mathbb{E}}_{\substack{o^i\sim D,g^i\\ a^i\sim\pi_{\theta^i}(\cdot|o^i,g^i)}}\left[\log\frac{\pi_{\bar{\theta}^i}(a^i|o^i,g^i)}{p(a^i|o^i)}\right], \text{where } p(a^i|o^i) = \mathop{\mathbb{E}}_{\substack{z'\sim p(\cdot|m)\\ g'=\phi^i(o^i,z')}}\left[\pi_{\bar{\theta}}(a^i|o^i,g')\right]. \tag{16}$$

For $|\mathcal{M}|$ training games, the $\mathcal{I}(m;g^i|o^i)$ term can be simplified as:

$$\mathcal{I}(m;g^i|o^i) = \mathbb{E}_{m,o^i\sim D}\left[\log p(m|o^i,g^i)\right] + \log|\mathcal{M}|, \tag{17}$$

To calculate the RHS of equation 17, we introduce an auxiliary inference network ξ. Denote the game's one-hot index as x. Then the auxiliary network outputs the prediction $\hat{x}$, ***i.e.***, $p(\hat{x}|o,g;\xi)$. By minimizing the cross-entropy loss in equation 18, ψ and ξ are simultaneously optimized. Gumbel-softmax trick Jang et al. (2016) is used for discrete z.

$$\min_{\psi,\xi}\mathbb{E}_{\substack{z\sim p(\cdot|m;\psi)\\ o^i\sim D}}\left[-x\log\left(p\left(\hat{x}|o^i,\phi^i(o^i,z);\xi\right)\right)\right]. \tag{18}$$

6 Experiments

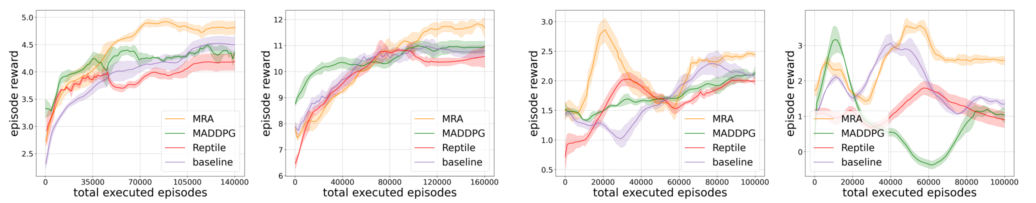
Experiments are conducted in three environments which cover both competitive and mixed games and are built based on the particle-world framework Lowe et al. (2017). Models are trained with 4 random seeds. Details of environment settings and hyperparameters are listed in Appendix E.

6.1 Benefits of Meta Representations

We first conduct experiments to show the benefits of the proposed meta-representations. We aim to answer the question that if the MG-common strategic knowledge in various training MGs can be extracted by MRA, and if it can benefit individual training games.

In Fig. 1, we compare MRA and the following methods: (1) the MADDPG algorithm Lowe et al. (2017) with relational representations (**MADDPG**); (2) the Reptile algorithm Nichol et al. (2018) (**Reptile**); (3) baseline with the same network architectures as MRA, but agents learn their policies only in a *single* MG (**baseline**).

For MRA and Reptile, the size of training MG set $|\mathcal{M}|$ for the three environments is $4, 4, 3$, respectively. Further specifications are in Appendix E. We note that the *actual* executed episodes of MRA in one MG are $|\mathcal{M}|$ times *smaller* than that for baseline and MADDPG, which reveals the efficiency of the proposed meta-representation. Although the Pacman-like world is not a zero-sum game, we still provide cross-comparison results in Appendix E for completeness.



(a) Treasure collection. (b) Resource occupying. (c) Pacman-like world: **Left:** Pac-Man. **Right:** Ghost.

Figure 1: Benefits of meta-representations in the three environments. **(a):** 6 collectors and 20 treasures; **(b):** 12 agents with 6 resources; **(c):** 8 Pac-Man, 4 ghosts and 20 food dots.

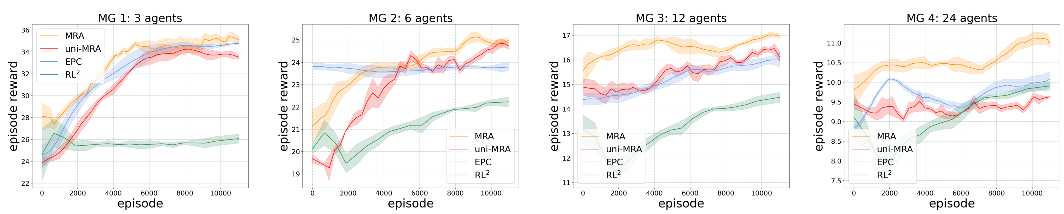


Figure 2: Multi-MG training curves in resource occupying environment. Total number of training MGs are 4, with population $3, 6, 12, 24$.

6.2 Performance Comparison in Multiple Games

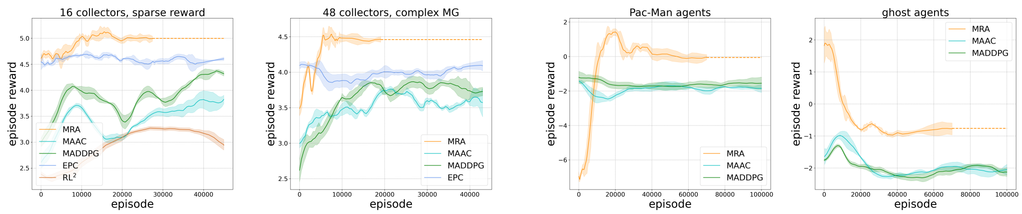
In this part, we compare the performance of MRA with multi-task and meta-learning methods, including **EPC** Long* et al. (2020) and **RL**^{2} Duan et al. (2016).

The curriculum learning EPC is implemented by initializing 3 parallel sets of agents and mix-and-match the top 2 sets to the successive MG. For RL^2, each trial contains a cycle of all the $|\mathcal{M}|$ MGs. We also compare another MRA variant, **uni-MRA**, that samples z from a uniform distribution. In resource occupying environment with 8 resources, the results are shown in Fig. 2.

Due to the discrepancy between effective policies in different MGs, the game-common strategic knowledge is not well exploited by EPC. RL^2 agents are also observed to perform poorly in some MGs, which verifies the benefits of *explicitly* modeling the MG-common knowledge and MG-specific knowledge when population varies. Since no game-specific information is conditioned in uni-MRA, we observe that some MGs dominate the others. Detailed settings and curves in other environments are provided in Appendix E.

6.3 Generalization Evaluation

If the learned policies both (1) adapt better and faster; and (2) perform well in novel games with no additional training (*i.e.*, zero-shot transfer), then the algorithm is considered to generalize well. Now we evaluate the generalization ability of MRA and other methods based on these two metrics.



(a) MG with sparse reward. (b) Complex MG. (c) Imbalanced competitive agents.

Figure 3: Adaptation performance comparison. **(a):** Sparse reward treasure collection task; **(b):** Complex treasure collection MG with large population size; **(c):** Imbalanced Pac-Man and ghost agents: 2 Pac-Man vs 8 ghosts, where random exploration bottleneck exists.

Better Adaptation: We first show that MRA have better adaptation ability benefited from the distilled common knowledge and first-order gradient information. Comparisons are conducted between **MRA**, **EPC**, **RL**^{2}, **MADDPG** and **MAAC** Iqbal & Sha (2018). MADDPG and MAAC are trained from scratch, while the others are fine-tuned from the parameters trained in multiple MGs.

The common knowledge can provide agents a good policy initialization and overcome the random exploration bottleneck. For example, the random exploration often results in Pac-Man agents being killed if ghost agents dominate the game. And all agents end up with almost random behaviors. However, with common knowledge guiding the Pac-Man to take reasonable actions, agents will get useful information and learn to accomplish the task. In the Pac-Man game where 8 ghost agents chase 2 Pac-Man agents, the benefits of the common knowledge are reflected in Fig. 3(c). The random exploration leads to worse performance of MAAC and MADDPG agents.

Besides, in Fig. 3(a) when reward shaping is removed, MRA agents can still adapt with fewer episodes and has better asymptotic performance. We also show in Fig. 3(b) that the complexity brought by the large populations, *e.g.*, 48, can be successfully handled by MRA. The results verify the benefits of common knowledge compared with the transfer knowledge in EPC.

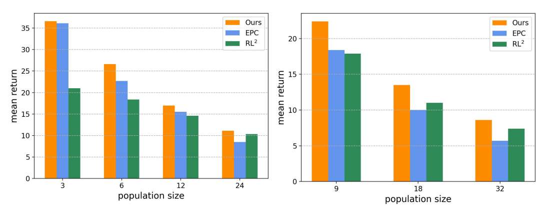


Figure 4: **Left:** Evaluation in the training MGs. **Right:** Zero-shot transfer to novel MGs.

Zero-shot Transfer: MRA also has better zero-shot transferability than EPC and RL^2. The results in resource occupying are reported in Fig. 4. Performance of MRA is calculated by taking expectations over the latent. However, its return will get higher if enumerated trial-and-error is taken, *i.e.*, choose the best policy mode by trying every z.

7 Conclusion & Discussions

In this paper, we propose meta representations for agents (MRA) that can generalize in Markov Games with varying populations. With latent variable policies and relational representations, the diverse strategic modes are captured. As an approximation to a theoretically justified objective, MRA effectively discovers the underlying strategic structures in the games that facilities generalizable knowledge learning. Experimental results also verify the benefits of MRA.

Our work also opens some new problems. Theorem 1 requires the computation of ς and ϵ as well as an extremely large latent space, both of which are impractical. Although approximations that MRA makes are reasonable, obtaining optimal $\mathbf{\Pi}^*$ will not always be guaranteed. Possible improvements include: bound ς by imposing restrictions on the evaluation MG set, or enlarge the latent space size by *e.g.*, adopting continuous latent variables, which we would like to explore as future work.

With role-symmetric game settings, MRA has benefits in many research problems, including dealing with population complexity, overcoming the multi-agent random exploration bottleneck, and adapting faster with the meta-represented agents. A fruitful avenue for future work is to augment MRA by *e.g.*, adapting roles Wang et al. (2020), to apply to other game settings. Besides, achieving NE may not indicate the global optimality in general-sum MGs, and metrics such as social optimum can be investigated.

For population-varying MGs, we model game-specific knowledge as strategic relationship. Although it may lose the universality in broader scopes compared with general meta-RL algorithms, we hope the idea of explicit strategic knowledge modeling can inspire algorithms that adjust with the task of interest.

REFERENCES

Akshat Agarwal, Sumit Kumar, and Katia Sycara. Learning transferable cooperative behavior in multi-agent teams. *arXiv preprint arXiv:1906.01202*, 2019.

Emma Brunskill and Lihong Li. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

Christos Dimitrakakis, David Parkes, Goran Radanovic, and Paul Tylkin. Multi-view decision processes: the helper-ai problem. In *31st Conference on Neural Information Processing Systems (NIPS)*. Neural Information Processing Systems Foundation, 2019.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Carlos Florensa, Yan Duan, and Pieter Abbeel. Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint arXiv:1704.03012*, 2017.

Jakob N Foerster, Yannis M Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *arXiv preprint arXiv:1605.06676*, 2016.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. *arXiv preprint arXiv:1810.02912*, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Michael Johanson, Kevin Waugh, Michael Bowling, and Martin Zinkevich. Accelerating best response calculation in large extensive games. In *IJCAI*, volume 11, pp. 258–265, 2011.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4190–4203, 2017.

Boyi Liu, Zhuoran Yang, and Zhaoran Wang. Policy optimization in zero-sum markov games: Fictitious self-play provably attains nash equilibria, 2021. URL `https://openreview.net/forum?id=c3MWGN_cTf`.

Qian Long*, Zihan Zhou*, Abhinav Gupta, Fei Fang, Yi Wu†, and Xiaolong Wang†. Evolutionary population curriculum for scaling multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJxbHkrKDH`.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pp. 6379–6390, 2017.

Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Remi Munos. A generalized training approach for multiagent learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=Bkl5kxrKDr`.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning nash equilibrium for general-sum markov games from batch data. In *Artificial Intelligence and Statistics*, pp. 232–241. PMLR, 2017.

Goran Radanovic, Rati Devidze, David C Parkes, and Adish Singla. Learning to collaborate in markov decision processes. *arXiv preprint arXiv:1901.08029*, 2019.

Joseph Suarez, Yilun Du, Phillip Isola, and Igor Mordatch. Neural mmo: A massively multiagent game environment for training and evaluating intelligent agents. *arXiv preprint arXiv:1903.00784*, 2019.

Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. In *Advances in neural information processing systems*, pp. 2244–2252, 2016.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Yee Teh, Victor Bapst, Wojciech M Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 4496–4506, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. Multi-agent reinforcement learning with emergent roles. *arXiv preprint arXiv:2003.08039*, 2020.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. *arXiv preprint arXiv:1802.05438*, 2018.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *arXiv preprint arXiv:1906.00729*, 2019.

A Proofs

In this section, we provide three propositions in A.1, A.2 and A.3. The proof of Theorem 1 in A.4 is built upon these propositions.

A.1 Proposition 1

Before giving Proposition 1, we provide a useful lemma of the Markov state transition operator $\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{\text{-}i}}$. The following Lemma 1 is given in (Liu et al., 2021), which is concerned with Fictitious Self-Play. We modify it to be suitable for multi-agent general-sum Markov Games with our notations.

Lemma 1. *(Liu et al., 2021). The Markov state transition operator satisfies:*

$$\left\| \sum_t \gamma^t \left[\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{\text{-}i}}\right)^t - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{\text{-}i}}\right)^t \right] \right\|_{op} \leq \frac{\gamma}{(1-\gamma)^2} \left\| \mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{\text{-}i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{\text{-}i}} \right\|_{op}$$

Proof. To begin with, we have the following inequality of the operator $\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}$:

$$\begin{aligned}
&\left\|\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^t - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^t\right\|_{op} \\
&= \left\|\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^{t-1}\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right) + \left(\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^{t-1} - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^{t-1}\right)\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op} \\
&\leq \left\|\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^{t-1}\left(\mathcal{P}^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)\right\|_{op} + \left\|\left(\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^{t-1} - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^{t-1}\right)\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op} \\
&\leq \left\|\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op} + \left\|\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^{t-1} - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^{t-1}\right\|_{op} \\
&\leq \left\|\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op} + \left\|\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op} + \left\|\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^{t-2} - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^{t-2}\right\|_{op} \\
&\leq t \cdot \left\|\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op}
\end{aligned} \tag{19}$$

Then we have:

$$\begin{aligned}
&\left\|\sum_t \gamma^t\left[\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^t - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^t\right]\right\|_{op} \\
&\leq \sum_t \gamma^t\left\|\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^t - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^t\right\|_{op} \\
&\leq \left(\sum_t t\gamma^t\right)\left\|\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op}
\end{aligned} \tag{20}$$

Thus,:

$$\left\|\sum_t \gamma^t\left[\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}}\right)^t - \left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^t\right]\right\|_{op} \leq \frac{\gamma}{(1-\gamma)^2}\left\|\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right\|_{op}$$

□

Now we state our first proposition.

Let the distance $\kappa(\pi^i)$ between policy π^i and the best response π^{*i} in the action space be defined as:

$$\kappa(\pi^i) = \left\|\left\|\pi^{*i}(a|s) - \pi^i(a|s)\right\|_{a,1}\right\|_{s,\infty}.$$

Proposition 1. *For the joint strategy $\boldsymbol{\pi}$, $\mathcal{D}_m(\boldsymbol{\pi})$ is bounded by:*

$$\mathcal{D}_m(\boldsymbol{\pi}) \leq \left(\frac{\gamma L_m}{(1-\gamma)^2} + \frac{1}{1-\gamma}\right)\left\|\kappa(\pi^i)\right\|_{i,1}. \tag{21}$$

Proof. First, a state visitation measure $\rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}}$ of the joint strategy $(\pi^{*i},\boldsymbol{\pi}^{-i})$ in MG m is defined as follows:

$$\begin{aligned}
\rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}} &= \left(\mathcal{I} - \gamma\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^{-1}\delta_s \\
&= \left(\sum_t \gamma^t\left(\mathcal{P}_m^{\pi^i,\boldsymbol{\pi}^{-i}}\right)^t\right)\delta_s,
\end{aligned} \tag{22}$$

where δ_s is a Dirac delta function.

By converting the value of strategy to the integration of reward over state measure, we have:

$$
\begin{aligned}
& v_{\boldsymbol{\pi}^{-i}}^{*i,m}(s) - v_{\boldsymbol{\pi}}^{i,m}(s) \\
&= \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}}} \left[\mathbb{E}_{\pi^{*i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) \right] - \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{i},\boldsymbol{\pi}^{-i}}} \left[\mathbb{E}_{\pi^{i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) \right] \\
&= \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}}} \left[\mathbb{E}_{\pi^{i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) \right] - \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{i},\boldsymbol{\pi}^{-i}}} \left[\mathbb{E}_{\pi^{i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) \right] \\
&\quad + \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}}} \left[\mathbb{E}_{\pi^{*i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) - \mathbb{E}_{\pi^{i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) \right] \\
&\leq \left\| \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \rho_{s,m}^{\pi^{i},\boldsymbol{\pi}^{-i}} \right\|_{L_1(\mathcal{S})} + \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}}} \left[\mathbb{E}_{\pi^{*i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) - \mathbb{E}_{\pi^{i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) \right]
\end{aligned} \tag{23}
$$

We then bound the two terms separately. For the first term, we have from Lemma 1 that:

$$
\left\| \sum_t \gamma^t \left[\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} \right)^t - \left(\mathcal{P}_m^{\pi^{i},\boldsymbol{\pi}^{-i}} \right)^t \right] \right\|_{op} \leq \frac{\gamma}{(1-\gamma)^2} \left\| \mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^{i},\boldsymbol{\pi}^{-i}} \right\|_{op}
$$

Then by the definition of L_m, we have:

$$
\begin{aligned}
& \left\| \sum_t \gamma^t \left[\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} \right)^t - \left(\mathcal{P}_m^{\pi^{i},\boldsymbol{\pi}^{-i}} \right)^t \right] \right\|_{op} \\
&\leq \frac{\gamma}{(1-\gamma)^2} \left\| \mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \mathcal{P}_m^{\pi^{i},\boldsymbol{\pi}^{-i}} \right\|_{op} \\
&\leq \frac{\gamma L_m}{(1-\gamma)^2} \kappa(\pi^i)
\end{aligned} \tag{24}
$$

Thus, the first term satisfies:

$$
\begin{aligned}
\left\| \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}} - \rho_{s,m}^{\pi^{i},\boldsymbol{\pi}^{-i}} \right\|_{L_1(\mathcal{S})} &= \left\| \left(\sum_t \gamma^t \left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} \right)^t \right) \delta_s - \left(\sum_t \gamma^t \left(\mathcal{P}_m^{\pi^{i},\boldsymbol{\pi}^{-i}} \right)^t \right) \delta_s \right\|_{L_1(\mathcal{S})} \\
&\leq \left\| \sum_t \gamma^t \left[\left(\mathcal{P}_m^{\pi^{*i},\boldsymbol{\pi}^{-i}} \right)^t - \left(\mathcal{P}_m^{\pi^{i},\boldsymbol{\pi}^{-i}} \right)^t \right] \right\|_{op} \cdot \|\delta_s\|_{L_1(\mathcal{S})} \\
&\leq \frac{\gamma L_m}{(1-\gamma)^2} \kappa(\pi^i)
\end{aligned} \tag{25}
$$

The $\mathcal{L}_{+\infty}$-norm over s of the second term is bounded by:

$$
\begin{aligned}
& \left\| \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}}} \left[\mathbb{E}_{\pi^{*i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) - \mathbb{E}_{\pi^{i},\boldsymbol{\pi}^{-i}} r_m^i(s', \boldsymbol{a}) \right] \right\|_{s,\infty} \\
\leq & \left\| \mathbb{E}_{s' \sim \rho_{s,m}^{\pi^{*i},\boldsymbol{\pi}^{-i}}} \left[\left\| \pi^{*i}(\cdot|s) - \pi^i(\cdot|s) \right\|_1 \right] \right\|_{s,\infty} \\
\leq & \kappa(\pi^i) \left(\sum_t \gamma^t \right) \\
= & \frac{\kappa(\pi^i)}{1-\gamma}
\end{aligned} \tag{26}
$$

Finally, from equation 25 and equation 26, we have the following bound for $\mathcal{D}_m(\boldsymbol{\pi})$:

$$\begin{aligned}
\mathcal{D}_m(\boldsymbol{\pi}) &= \left\| \left\| v^{*i,m}_{\boldsymbol{\pi}^{\text{-}i}}(s) - v^{i,m}_{\boldsymbol{\pi}}(s) \right\|_{s,\infty} \right\|_{i,1} \\
&\leq \left\| \left\| \left\| \rho^{\pi^{*i},\boldsymbol{\pi}^{\text{-}i}}_{s,m} - \rho^{\pi^{i},\boldsymbol{\pi}^{\text{-}i}}_{s,m} \right\|_{L_1(\mathcal{S})} + \mathbb{E}_{s' \sim \rho^{\pi^{*i},\boldsymbol{\pi}^{\text{-}i}}_{s,m}} \left[\mathbb{E}_{\pi^{*i},\boldsymbol{\pi}^{\text{-}i}} r^i_m(s', \boldsymbol{a}) - \mathbb{E}_{\pi^{i},\boldsymbol{\pi}^{\text{-}i}} r^i_m(s', \boldsymbol{a}) \right] \right\|_{s,\infty} \right\|_{i,1} \\
&\leq \left\| \frac{\gamma L_m}{(1-\gamma)^2} \kappa(\pi^i) + \frac{\kappa(\pi^i)}{1-\gamma} \right\|_{i,1} \\
&= \left(\frac{\gamma L_m}{(1-\gamma)^2} + \frac{1}{1-\gamma} \right) \left\| \kappa(\pi^i) \right\|_{i,1}
\end{aligned} \tag{27}$$

□

A.2 Proposition 2

Proposition 2. *For the training MG set $\mathcal{M}$ and the evaluation MG set $\mathcal{M}'$, if ϵ satisfies*

$$\epsilon \geq \varsigma - \min_{\iota_m, \iota_{m'}} \frac{\varsigma\gamma\left(\iota_{m'} - \iota_m\right)}{\gamma\iota_{m'} + 1 - \gamma}, \tag{28}$$

then for every evaluation Markov Game $m' \in \mathcal{M}'$, the joint strategy that achieves Nash Equilibrium in m' is guaranteed to be contained in the ϵ-range joint strategy set $\hat{\boldsymbol{\Pi}}$.

Proof. For Definition 1 of the distance ς between the training MG set $\mathcal{M}$ and the evaluation MG set $\mathcal{M}'$, we have the following equivalent logic statement:

$$\forall m' \in \mathcal{M}', i \in \{1, \ldots, N_{m'}\}, \exists m \in \mathcal{M}, \boldsymbol{\pi}, \boldsymbol{\pi'}, i' \in h_{i,m},$$

$$\text{s.t. } \mathcal{D}_m(\boldsymbol{\pi}) = 0, \mathcal{D}_{m'}(\boldsymbol{\pi'}) = 0, \forall \pi^i \in \boldsymbol{\pi}, \mathcal{D}_{m'}(\pi^{i'}, \boldsymbol{\pi'}^{\text{-}i}) \leq \varsigma.$$

In an evaluation MG $\tilde{m}' \in \mathcal{M}'$, let i and i' be the agent index defined as follows:

$$i = \operatorname*{arg\,max}_{i \in \{1,\ldots,N_{\tilde{m}'}\}} \left[\min_{i' \in h_{i,m}} \mathcal{D}_{\tilde{m}'}(\pi^{i'}, \boldsymbol{\pi'}^{\text{-}i}) \right], \text{ s.t. } \mathcal{D}_m(\boldsymbol{\pi}) = 0, \mathcal{D}_{\tilde{m}'}(\boldsymbol{\pi'}) = 0 \tag{29}$$

Intuitively, the above agent i is the agent in $\tilde{m}'$ that being replaced by the trained policy $\pi^{i'}$ in the policy set leads to the largest distance to the Nash Equilibrium. And agent i' is the corresponding agent that $\pi^{i'}$ achieves an NE in a particular training MG.

With i and i' denied in equation 29, the bound in Proposition 1 can be specified as follows:

$$\begin{aligned}
\mathcal{D}_{\tilde{m}'}(\pi^{i'}, \boldsymbol{\pi'}^{\text{-}i}) &= \left\| \left\| v^{*i,\tilde{m}'}_{\boldsymbol{\pi'}^{\text{-}i}} - v^{i,\tilde{m}'}_{\pi^{i'},\boldsymbol{\pi'}^{\text{-}i}} \right\|_{s,\infty} \right\|_{i,1} \\
&= \left\| v^{i,\tilde{m}'}_{\boldsymbol{\pi'}} - v^{i,\tilde{m}'}_{\pi^{i'},\boldsymbol{\pi'}^{\text{-}i}} \right\|_{s,\infty} \\
&= \left\| v^{i,\tilde{m}'}_{\pi'^{i},\boldsymbol{\pi'}^{\text{-}i}} - v^{i,\tilde{m}'}_{\pi^{i'},\boldsymbol{\pi'}^{\text{-}i}} \right\|_{s,\infty} \\
&\leq \frac{\gamma\iota_{\tilde{m}'}}{(1-\gamma)^2} \kappa(\pi^{i'}) + \frac{\kappa(\pi^{i'})}{1-\gamma}
\end{aligned} \tag{30}$$

The second equality holds since $\mathcal{D}_{\tilde{m}'}(\boldsymbol{\pi'}) = 0$, and the distance from the joint strategy $(\pi^{i'}, \boldsymbol{\pi'}^{\text{-}i})$ to a Nash Equilibrium is equal to the distance to $\boldsymbol{\pi'}$. The last inequality holds due to the bound in Proposition 1 and the definition of i and i'.

This implies that for any MG $\tilde{m}'$ we have:

$$\max_{i\in\{1,\dots,N_{\tilde{m}'}\}} \min_{\substack{m\in\mathcal{M},i'\in h_i\\ \boldsymbol{\pi}\in\{\boldsymbol{\pi}|\mathcal{D}_m(\boldsymbol{\pi})=0\}\\ \boldsymbol{\pi}'\in\{\boldsymbol{\pi}'|\mathcal{D}_{\tilde{m}'}(\boldsymbol{\pi}')=0\}}} \mathcal{D}_{\tilde{m}'}(\pi^{i'}, \boldsymbol{\pi}'^{\text{-}i}) \leq \frac{\gamma \iota_{\tilde{m}'}}{(1-\gamma)^2}\kappa(\pi^{i'}) + \frac{\kappa(\pi^{i'})}{1-\gamma} \tag{31}$$

With the maximum influence $\iota_{m'}$ over the evaluation MG $m' \in \mathcal{M}'$, we obtain:

$$\varsigma = \max_{\iota_{m'}} \frac{\gamma \iota_{m'}}{(1-\gamma)^2}\kappa(\pi^{i'}) + \frac{\kappa(\pi^{i'})}{1-\gamma} \tag{32}$$

Since $\mathcal{D}_m(\boldsymbol{\pi}) = 0$ and $\mathcal{D}_{m'}(\boldsymbol{\pi}') = 0$, the best response in MG m with other agent's strategies fixed as $\boldsymbol{\pi}^{\text{-}i'}$ is $\pi^{i'}$. And in MG m', the best response with other agent's strategies fixed as $\boldsymbol{\pi}'^{\text{-}i}$ is π'^{i}. Thus we get:

$$\kappa(\pi^{i'}) = \kappa(\pi'^{i}) = \left\| \left\| \pi^{i'}(a) - \pi'^{i}(a) \right\|_{a,1} \right\|_{s,\infty} \tag{33}$$

For ϵ that satisfies equation 28, we obtain:

$$\begin{aligned} \epsilon &\geq \max_{\iota_{m'},\iota_m} \varsigma - \frac{\varsigma\gamma\left(\iota_{m'} - \iota_m\right)}{\gamma\iota_{m'} + 1 - \gamma} \\ &\geq \max_{\iota_{m'},\iota_m} \left(\frac{\gamma\iota_m + 1 - \gamma}{\gamma\iota_{m'} + 1 - \gamma}\right) \cdot \left(\frac{\gamma\iota_{m'}}{(1-\gamma)^2}\kappa(\pi^{i'}) + \frac{\kappa(\pi^{i'})}{1-\gamma}\right) \\ &\geq \max_{\iota_m} \left(\frac{\gamma\iota_m + 1 - \gamma}{\gamma\iota_{\tilde{m}'} + 1 - \gamma}\right) \cdot \left(\frac{\gamma\iota_{\tilde{m}'}}{(1-\gamma)^2}\kappa(\pi'^{i}) + \frac{\kappa(\pi'^{i})}{1-\gamma}\right) \\ &\geq \max_{\iota_m} \frac{\gamma\iota_m}{(1-\gamma)^2}\kappa(\pi'^{i}) + \frac{\kappa(\pi'^{i})}{1-\gamma} \end{aligned} \tag{34}$$

For i and i' defined in equation 29, we have the following inequality by noticing that $\mathcal{D}_m(\boldsymbol{\pi}) = 0$:

$$\begin{aligned} \mathcal{D}_m\left(\left(\pi^{i'}, \boldsymbol{\pi}^{\text{-}i'}\right)\right) &= \left\| \left\| v^{*i,m}_{\boldsymbol{\pi}^{\text{-}i'}} - v^{i,m}_{\pi'^{i},\boldsymbol{\pi}^{\text{-}i'}} \right\|_{s,\infty} \right\|_{i,1} \\ &= \left\| v^{i,m}_{\boldsymbol{\pi}} - v^{i,m}_{\pi'^{i},\boldsymbol{\pi}^{\text{-}i'}} \right\|_{s,\infty} \\ &= \left\| v^{i,m}_{\pi^{i'},\boldsymbol{\pi}^{\text{-}i'}} - v^{i,m}_{\pi'^{i},\boldsymbol{\pi}^{\text{-}i'}} \right\|_{s,\infty} \\ &\leq \frac{\gamma\iota_m}{(1-\gamma)^2}\kappa(\pi'^{i}) + \frac{\kappa(\pi'^{i})}{1-\gamma} \\ &\leq \max_{\iota_m} \frac{\gamma\iota_m}{(1-\gamma)^2}\kappa(\pi'^{i}) + \frac{\kappa(\pi'^{i})}{1-\gamma} \\ &\leq \epsilon \end{aligned} \tag{35}$$

This indicates that if we choose ϵ to satisfy equation 28, then the policy π'^{i} that achieves Nash Equilibrium in MG $\tilde{m}'$ is guaranteed to be included in the policy set.

Since all the above inequalities hold for any $\tilde{m}' \in \mathcal{M}'$, the policy that achieves NE in all MGs in the evaluation MG set $\mathcal{M}'$ is guaranteed to be included in the policy set. This completes the proof. □

A.3 PROPOSITION 3

Proposition 3. *If* $|\boldsymbol{\Pi}_\Theta| \geq |\hat{\boldsymbol{\Pi}}|$ *and* ϵ *satisfies* $\epsilon \geq \varsigma - \min_{\iota_m,\iota_{m'}} \frac{\varsigma\gamma(\iota_{m'}-\iota_m)}{\gamma\iota_{m'}+1-\gamma}$, *then the optimal* $\boldsymbol{\Pi}$ *that maximizes the objective:*

$$\mathcal{L}(\boldsymbol{\Pi}) = \max_{\boldsymbol{\pi}\sim\boldsymbol{\Pi}} \min_{\hat{\boldsymbol{\pi}}\sim\hat{\boldsymbol{\Pi}}} \mathbb{E}_{\boldsymbol{a}\sim\hat{\boldsymbol{\pi}},\boldsymbol{o}} \left[\log \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{o})\right], \tag{36}$$

for every evaluation Markov Game $m' \in \mathcal{M}'$, there exists a joint strategy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\Theta^*}$ that reaches Nash Equilibrium (i.e., $\boldsymbol{\Pi}_{\Theta^*} = \boldsymbol{\Pi}^*$ satisfies equation 4).

Proof. Since $|\boldsymbol{\Pi}| \geq |\hat{\boldsymbol{\Pi}}|$, the optimal $|\boldsymbol{\Pi}|$ that maximizes $\mathcal{L}(\boldsymbol{\Pi})$ in equation 36 must satisfy:

$$\min_{\hat{\boldsymbol{\pi}} \sim \hat{\boldsymbol{\Pi}}} \mathbb{E}_{\boldsymbol{a} \sim \hat{\boldsymbol{\pi}}, \boldsymbol{o}} \left[\log \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{o})\right] = 1$$

.

In other words, for every policy $\hat{\boldsymbol{\pi}}$ in the joint strategy set $\hat{\boldsymbol{\Pi}}$, *i.e.*, $\hat{\boldsymbol{\pi}} \in \hat{\boldsymbol{\Pi}}$, there exists a learned policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, such that $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$. Note that the above statement is true only when $|\boldsymbol{\Pi}| \geq |\hat{\boldsymbol{\Pi}}|$.

For ϵ that satisfies equation 28, from Proposition 2 we know that for every evaluation MG $m' \in \mathcal{M}'$, the strategy that achieves Nash Equilibrium are guaranteed to be contained in $\hat{\boldsymbol{\Pi}}$. Thus, the optimal policy set $\boldsymbol{\Pi}$ that results from optimizing equation 36 also contains the strategies that are NE in every MG $m' \in \mathcal{M}'$.

So the optimal policy set $\boldsymbol{\Pi}$ satisfies equation 4, which completes the proof.

□

A.4 Proof of Theorem 1

Proof. Since $|\boldsymbol{\Pi}_\Theta| \geq |\hat{\boldsymbol{\Pi}}|$, we can simplify the objective in equation 36 by updating the joint strategy $\boldsymbol{\pi}$ in a fixed-size set $\boldsymbol{\Pi}$. That is, maximizing equation 36 is equivalent to:

$$\max_{\boldsymbol{\Pi}} \min_{\hat{\boldsymbol{\pi}} \sim \hat{\boldsymbol{\Pi}}} \max_{\boldsymbol{\pi} \sim \boldsymbol{\Pi}} \mathbb{E}_{\boldsymbol{a} \sim \hat{\boldsymbol{\pi}}, \boldsymbol{o}} \left[\log \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{o})\right] = \max_{\boldsymbol{\pi} \sim \boldsymbol{\Pi}} \min_{\hat{\boldsymbol{\pi}} \sim \hat{\boldsymbol{\Pi}}} \mathbb{E}_{\boldsymbol{a} \sim \hat{\boldsymbol{\pi}}, \boldsymbol{o}} \left[\log \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{o})\right] \tag{37}$$

From the non-negativeness of KL divergence, we have:

$$\mathcal{D}_{KL}\left(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}\right) = \mathbb{E}_{\hat{\boldsymbol{\pi}}} \left[\log \frac{\hat{\boldsymbol{\pi}}}{\boldsymbol{\pi}}\right] \geq 0$$

The equality holds when $\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}$.

Thus, we have:

$$\max_{\boldsymbol{\pi}} \mathbb{E}_{\hat{\boldsymbol{\pi}}} \left[\log \boldsymbol{\pi}\right] \leq -\mathcal{H}(\hat{\boldsymbol{\pi}}) \tag{38}$$

If $\boldsymbol{\pi} \in \hat{\boldsymbol{\Pi}}$ is constrained, then the equality holds and $\max_{\boldsymbol{\pi}} \mathbb{E}_{\hat{\boldsymbol{\pi}}} \left[\log \boldsymbol{\pi}\right] = -\mathcal{H}(\hat{\boldsymbol{\pi}})$

This leads to:

$$\begin{aligned} \max_{\boldsymbol{\pi} \sim \boldsymbol{\Pi}} \min_{\hat{\boldsymbol{\pi}} \sim \hat{\boldsymbol{\Pi}}} \mathbb{E}_{\hat{\boldsymbol{\pi}}} \left[\log \boldsymbol{\pi}\right] &= \min_{\hat{\boldsymbol{\pi}} \sim \hat{\boldsymbol{\Pi}}} -\mathcal{H}(\hat{\boldsymbol{\pi}}) \\ &= \max_{\hat{\boldsymbol{\pi}} \sim \hat{\boldsymbol{\Pi}}} \mathcal{H}(\hat{\boldsymbol{\pi}}) \\ &= \max_{\boldsymbol{\pi} \sim \boldsymbol{\Pi}} \mathcal{H}(\boldsymbol{\pi}), \ \text{s.t.}\ \boldsymbol{\pi} \in \hat{\boldsymbol{\Pi}} \end{aligned} \tag{39}$$

The above equation states that the strategy $\boldsymbol{\pi}$ is learned to fit $\hat{\boldsymbol{\pi}}$. And to cover all the $\hat{\boldsymbol{\pi}} \in \hat{\boldsymbol{\Pi}}$, the entropy of strategies in $\boldsymbol{\Pi}$ should also be maximized.

Then we get the following equality:

$$\begin{aligned} \max_{\boldsymbol{\pi} \sim \boldsymbol{\Pi}} \mathcal{H}(\boldsymbol{\pi}) &= \max_{\Omega} \mathcal{I}(m; a|o) + \mathcal{H}(a|m, o) \\ &= \max_{\Omega} \mathcal{I}(m; a|o) + \mathcal{I}(g; a|m, o) + \mathcal{H}(a|m, o, g) \\ &= \max_{\Omega} \mathcal{I}(m; a|o) + \mathcal{I}(g; a|o) + \mathcal{H}(a|m, o, g) \\ &= \max_{\Omega} \mathcal{I}(m; a|o) + \mathcal{I}(g; a|o), \end{aligned} \tag{40}$$

where the first equality holds following the definition of mutual information and by noticing that policy $\boldsymbol{\pi}$ is meta-represented by Ω. The last equality holds since the size of the meta-represented policy set $|\boldsymbol{\Pi}_\Theta|$ is sufficiently large and satisfying $|\boldsymbol{\Pi}_\Theta| \geq |\hat{\boldsymbol{\Pi}}|$.

Combining equation 39 and equation 40. we have:

$$\min_{\hat{\boldsymbol{\pi}} \sim \hat{\boldsymbol{\Pi}}} \max_{\boldsymbol{\pi} \sim \boldsymbol{\Pi}} \mathbb{E}_{\hat{\boldsymbol{\pi}}} [\log \boldsymbol{\pi}] = \max_{\Omega} \mathcal{I}(m; a|o) + \mathcal{I}(g; a|o), \text{ s.t. } \boldsymbol{\pi} \in \hat{\boldsymbol{\Pi}} \tag{41}$$

Then by equation 37 and equation 41, we have that the solution of the following objective is equivalent to the solution of equation 36:

$$\psi^*, \phi^* = \arg\max_{\psi,\phi} \mathcal{I}(g; a|o) + \mathcal{I}(m; g|o) \text{ s.t. } \boldsymbol{\pi}_{\theta^*} \in \hat{\boldsymbol{\Pi}}$$

Thus, for every evaluation Markov Game $m' \in \mathcal{M}'$, there exists a joint strategy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_\Theta$ that reaches Nash Equilibrium (*i.e.*, $\boldsymbol{\Pi}_{\Theta^*} = \boldsymbol{\Pi}^*$ satisfies equation 4). □

B Fast Adaptation with First-Order Gradient

Reptile Nichol et al. (2018) is a meta-learning algorithm that uses first-order gradient information for fast adaptation.

For parameter θ that maximizes objective $\mathcal{L}_m^k(\theta)$ in the k-th mini-batch of game m, θ is updated by $\theta \leftarrow \theta + \alpha\Delta\theta$, where $\Delta\theta = U_m^K(\theta) - \theta$ and $U_m^K(\theta)$ denotes the updated θ after K gradient steps with learning rate β, and α is a hyperparameter.

Denote the k-th step parameter as θ_k, then the update $\Delta\theta$ of K gradient steps is as follows:

$$\begin{aligned}
\Delta\theta &= \theta_K - \theta_1 \\
&= \beta \sum_{k=1}^{K-1} \nabla \mathcal{L}_m^k(\theta_k) \\
&= \beta \sum_{k=1}^{K-1} \left(\nabla \mathcal{L}_m^k(\theta_1) + \nabla^2 \mathcal{L}_m^k(\theta_1) (\theta_k - \theta_1) + \mathcal{O}\left(\|\theta_k - \theta_1\|^2 \right) \right) \\
&= \beta \sum_{k=1}^{K-1} \left(\nabla \mathcal{L}_m^k(\theta_1) + \beta \nabla^2 \mathcal{L}_m^k(\theta_1) \sum_{j=1}^{k-1} \nabla \mathcal{L}_m^k(\theta_j) + \mathcal{O}\left(\beta^2\right) \right) \\
&= \beta \left[\sum_{k=1}^{K-1} \left(\nabla \mathcal{L}_m^k(\theta_1) + \beta \sum_{j=1}^{k-1} \left(\nabla^2 \mathcal{L}_m^k(\theta_1) \nabla \mathcal{L}_m^j(\theta_1) \right) \right) + \mathcal{O}\left(\beta^2\right) \right],
\end{aligned} \tag{42}$$

where the last equation holds since $\nabla \mathcal{L}_m^k(\theta_j) = \nabla \mathcal{L}_m^k(\theta_1) + \mathcal{O}(\beta)$.

For the initial parameter $\theta = \theta_1$, the term $\sum_{k=1}^{K-1} \nabla \mathcal{L}_m^k(\theta_1)$ maximizes the overall performance at θ in all the K mini-batches in an MG m. The key difference from the joint training objective is the term $\nabla^2 \mathcal{L}_m^k(\theta_1) \nabla \mathcal{L}_m^j(\theta_1)$. When the expectation are taken under mini-batch sampling in m, denote $\mathbb{E}_k$ as the expectation over the mini-batch defined by J^k. Omitting the higher-order term $\mathcal{O}\left(\beta^2\right)$, we have:

$$\begin{aligned}
\mathbb{E}[\Delta\theta] &= (K-1)\mathbb{E}_k \left[\mathcal{L}_m^k(\theta) \right] + (K-1)(K-2)\beta \mathbb{E}_{j,k} \left[\nabla^2 \mathcal{L}_m^k(\theta) \nabla \mathcal{L}_m^j(\theta) \right] \\
&= (K-1)\mathbb{E}_k \left[\mathcal{L}_m^k(\theta) \right] + \frac{(K-1)(K-2)\beta}{2} \mathbb{E}_{j,k} \left[\nabla^2 \mathcal{L}_m^k(\theta) \nabla \mathcal{L}_m^j(\theta) + \nabla^2 \mathcal{L}_m^j(\theta) \nabla \mathcal{L}_m^k(\theta) \right] \\
&= (K-1)\mathbb{E}_k \left[\mathcal{L}_m^k(\theta) \right] + \frac{(K-1)(K-2)\beta}{2} \mathbb{E}_{j,k} \left[\nabla \left(\nabla \mathcal{L}_m^k(\theta) \nabla \mathcal{L}_m^j(\theta) \right) \right]
\end{aligned} \tag{43}$$

Thus, updating θ by $\theta \leftarrow \theta + \alpha\Delta\theta$ not only maximizes the average performance in K mini-batches of all MGs, but also maximizes the inner product between gradients of different mini-batches, *i.e.*, $\nabla \mathcal{L}_m^k(\theta) \nabla \mathcal{L}_m^j(\theta)$. Thus, the generalization ability is improved and fast adaptation is achieved.

Algorithm 1 MRA: Training in MG set $\mathcal{M}$

```
while not converged do
    for MG m = 1 ... |M| do
        sample lower-level latent z ~ p_ψ(·|m);
        execute action a ~ π_θ(·|o, g), where g = φ(o, z);
        push (o, a, o', g, r) to replay buffer;
        for k = 1 ... K do
            update critic ζ by minimizing equation 13;
            update policy at the k-th step θ_k by equation 14;
        end for
        update θ by θ ← θ + α(θ_K − θ);
        update φ to maximize the RHS of equation 16;
        update ψ and auxiliary network ξ by equation 18;
        update delayed parameters θ̄ and ζ̄;
    end for
end while
```

Algorithm 2 MRA: Adaptation in a novel evaluation MG m'

```
while not converged do
    sample lower-level latent z ~ p_ψ(·|m');
    execute action a ~ π_θ(·|o, g), where g = φ(o, z);
    push (o, a, o', r) to replay buffer;
    update critic ζ by minimizing equation 13;

    update θ and φ by equation 15;
    update delayed parameters θ̄ and ζ̄;
end while
```

C Derivation of Mutual Information Calculation

The two mutual information terms in equation 12, *i.e.*, $\mathcal{I}(g;a|o)$ and $\mathcal{I}(m;g|o)$ can be calculated as follows:

$$
\begin{aligned}
\mathcal{I}(g;a|o) &= \int p(a,o,g)\log\frac{p(a|o,g)}{p(a|o)}da\,do\,dg \\
&= \mathbb{E}_{a,o,g}[\log\frac{\pi_{\bar{\theta}}(a|o,g)}{p(a|o)}] + \mathbb{E}_{a,o,g}[\mathcal{D}_{KL}(p(a|o,g)||\pi_{\bar{\theta}}(a|o,g))] \\
&\geq \mathbb{E}_{a,o,g}[\log\frac{\pi_{\bar{\theta}}(a|o,g)}{p(a|o)}] \\
&\approx \mathbb{E}_{\substack{o\sim D\\ z\sim p(\cdot|m)\\ a\sim\pi_\theta(\cdot|o,\phi(o,z))}}\left[\log\frac{\pi_{\bar{\theta}}(a|o,\phi(o,z))}{\mathbb{E}_{\substack{z'\sim p(\cdot|m)\\ g'=\phi(o,z')}}[\pi_{\bar{\theta}}(a|o,g')]}\right]
\end{aligned} \tag{44}
$$

Similarly, we have:

$$
\begin{aligned}
\mathcal{I}(m;g|o) &= \int p(m,o,g)\log\frac{p(m|o,g)}{p(m|o)}dm\,do\,dg \\
&= \mathbb{E}_{m,o,g}[\log\frac{p(m|o,g)}{p(m)}] \\
&= \mathbb{E}_{m,o\sim D}\left[\log p(m|o,g)\right] + \log|\mathcal{M}|
\end{aligned} \tag{45}
$$

D Pseudocode

The high level algorithmic frameworks of the training process and evaluation process are shown in Algorithm 3 and Algorithm 2, respectively.

The complete pseudocode for training MRA is provided in Algorithm 3, and the pseudocode of the MRA adaptation process is given in Algorithm 4.

Algorithm 3 Training Procedure of MRA in Set $\mathcal{M}$ of Population-Varying Markov Games

Input: Markov Games $m \in \mathcal{M}$ with index y.
Output: Parameter ψ and θ^i, ϕ^i, ζ^i for each agent i.
Initialize P threads of games;
Initialize $T_{\text{update}} \leftarrow 0$;
Initialize replay buffer $\mathcal{D}_m$ for each Markov Game m;
while total episode number not reach **do**
  for Markov Game m in $\mathcal{M}$ **do**
    Reset game, each agent i samples lower-level latent code $z^i \sim p_\psi(z|m)$;
    for time steps in an episode **do**
      Each agent i executes action $a^i \sim \pi\big(\cdot|o^i, \phi_i(o^i, z^i); \theta^i\big)$ simultaneously and get reward r^i, next observation o'^i;
      Push $(\boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o'}, \boldsymbol{g}, \boldsymbol{r})$ to replay buffer $\mathcal{D}_m$;
      $\boldsymbol{o} \leftarrow \boldsymbol{o'}$
      $T_{\text{update}} \leftarrow T_{\text{update}} + P$;
      if $T_{\text{update}}\%(\text{min steps per update}) \leq P$ **then**
        A mini-batch of B samples of $(\boldsymbol{o}_b, \boldsymbol{a}_b, \boldsymbol{o'_b}, \boldsymbol{g}_b, \boldsymbol{r}_b)$ is sampled from $\mathcal{D}_m$;
        for $k = 1$ to K **do**
          Update all agents' critic parameter ζ^i by minimizing:
            $\mathcal{L}\left(\zeta^i\right) = \frac{1}{B}\sum_b \left(\mathcal{B}^i_{\boldsymbol{\pi}} Q - Q\left(\boldsymbol{o}_b, \boldsymbol{a}_b, g^i_b; \zeta^i\right)\right)^2$, where $\mathcal{B}^i_{\boldsymbol{\pi}} Q = r^i_b + \gamma \mathbb{E}_{\boldsymbol{a'} \sim \bar{\boldsymbol{\pi}}}\left[Q\left(\boldsymbol{o'_b}, \boldsymbol{a'}, g^i_b; \bar{\zeta}^i\right)\right]$;
          The k-th step of policy parameter θ^i_k is updated by gradient ascent:
            $\nabla_{\theta^i} J = \frac{1}{B}\sum_b \nabla_{\theta^i} \log \pi\big(a^i|o^i_b, g^i_b; \theta^i_k\big)\, Q\left(\boldsymbol{o}_b, \boldsymbol{a}_b, g^i_b; \zeta^i\right)$;
        end for
        Update all agents' θ^i by $\theta^i \leftarrow \theta^i + \alpha(\theta^i_K - \theta^i)$;
        Sample n latent code z'^i and approximate $p(a|o^i_b)$ by $\frac{1}{n}\sum \pi\big(a^i|o^i_b, \phi^i\left(o^i_b, z'^i\right); \bar{\theta}^i\big)$;
        Update all agents' ϕ_i by maximizing:
          $\mathcal{L}\left(\phi^i\right) = \frac{1}{B}\sum_b \mathbb{E}_{a^i \sim \pi(\cdot|o^i_b, \phi^i(o^i_b, z^i); \theta^i)}\left[\log \frac{\pi(a^i|o^i_b, \phi^i(o^i_b, z^i); \bar{\theta}^i)}{p(a|o)}\right]$;
        Update ψ and auxiliary network ξ simultaneously by minimizing:
          $\mathcal{L}\left(\psi, \xi\right) = \mathbb{E}_{z'^i \sim p_\psi(\cdot|m)}\left[-y \log\left(p\left(\hat{y}|o^i_b, \phi(o^i_b, z'^i); \xi\right)\right)\right]$;
        Update all agents' delayed parameters $\bar{\theta}^i$ and $\bar{\zeta}^i$;
      end if
    end for
  end for
end while

Algorithm 4 Adaptation Procedure of MRA in a Novel Markov Game m'

Input: Trained parameters from Alg. 3.
Output: Adapted parameters.
Initialize replay buffer $\mathcal{D}$;
Each agent i samples lower-level latent code $z^i \sim p_\psi(z|m')$;
while total episode number not reach **do**
  Reset game and receive initial observation $\boldsymbol{o}$;
  for time steps in an episode **do**
    Each agent i executes action $a^i \sim \pi\big(\cdot|o^i, \phi_i(o^i, z^i); \theta^i\big)$ simultaneously and get reward r^i, next observation o'^i;
    Push $(\boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o'}, \boldsymbol{r})$ to replay buffer $\mathcal{D}_{m'}$;
    $\boldsymbol{o} \leftarrow \boldsymbol{o'}$
    A mini-batch of B samples of $(\boldsymbol{o}_b, \boldsymbol{a}_b, \boldsymbol{o'_b}, \boldsymbol{r}_b)$ is sampled from $\mathcal{D}_m$;
    Calculate the detached relational graph $g^i = \phi_i(o_b^i, z_b^i)$
    Update all agents' critic parameter ζ^i by minimizing:
      $\mathcal{L}\left(\zeta^i\right) = \frac{1}{B}\sum_b \left(\mathcal{B}^i_{\boldsymbol{\pi}} Q - Q\left(\boldsymbol{o}_b, \boldsymbol{a}_b, g^i; \zeta^i\right)\right)^2$, where $\mathcal{B}^i_{\boldsymbol{\pi}} Q = r_b^i + \gamma \mathbb{E}_{\boldsymbol{a'}\sim\bar{\boldsymbol{\pi}}}\left[Q\left(\boldsymbol{o'_b}, \boldsymbol{a'}, g^i; \bar{\zeta}^i\right)\right]$;
    Update all agents' parameter $\omega^i = (\theta^i, \phi^i)$ by gradient ascent:
      $\nabla_{\omega^i} J = \frac{1}{B}\sum_b \nabla_{\omega^i} \log \pi\big(a^i|o_b^i, \phi^i(o_b^i, z_b^i); \theta_k^i\big)\, Q\left(\boldsymbol{o}_b, \boldsymbol{a}_b, g^i; \zeta^i\right)$;
    Update all agents' delayed parameters $\bar{\theta}^i$ and $\bar{\zeta}^i$;
  end for
end while

E Additional Experiments

E.1 Training Setups and Hyperparameters

Treasure Collection: Each agent is with the goal to collect more treasures in an episode. Treasures disappear and re-generate at a random location when touched by agents.

Resource Occupying: Agents receive rewards for occupying varisized resource landmarks: higher reward if one agent is occupying a larger resource with fewer other agents in it.

Pacman-like World: It is similar to predatory-prey games, but with additional food dots. Pac-Man agents are with goals to collect food and elude ghosts, and ghosts are with goals to touch Pac-Man.

The treasures in the treasure collection environment and the food dots in the Pacman-like world are randomly initialized in the position range of $[-1, +1]$, and regenerated when touched by collector agents and PacMan agents, respectively. The sizes of the resource landmarks are pre-defined as $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ and fixed in each episode.

When evaluating the benefits of MRA in Section 6.1 of the main paper, the training Markov Games in the three environments are as follows. In treasure collection environment, there are 4 training MGs in total. The agent populations in the 4 MGs are $3, 6, 12, 24$, respectively, which we denote as $\{3, 6, 12, 24\}$. The 4 training MGs in the resource occupying environment is $\{6, 9, 12, 15\}$. In the PacMan-like world, there are 3 training MGs in total: $\{(4, 2), (6, 3), (8, 4)\}$, where $(4, 2)$ in the first MG denotes that there are 4 PacMan agents and 2 ghost agents.

We adopt the same set of hyperparameter for experiments. 12 rollouts are executed in parallel when training. The maximum length of the replay buffer is $1e6$. Episode length is set to 20. The dimension of the latent code z is 6. The critic also adopts a self-attention network in a similar way with MAAC Iqbal & Sha (2018). And the number of gradient steps of policy and critic parameters in each update, *i.e.*, K, is set as 10. And $\alpha = 1$ works well in experiments. Batch size is set to 1024 and Adam is used as the optimizer. The initial learning rate is set to 0.0003. In all experiments, we use one NVIDIA Tesla P40 GPU.

E.2 CROSS-COMPARISON RESULTS

We provide the cross-comparison results in the PacMan-like world. The comparisons are conducted between the MRA agents trained in multiple MGs and the agents trained in a single MG. The score is summed in each episode, averaged across homogeneous agents on 40 runs and normalized.

Table 1: PacMan scores.

Ghosts \ PacMan	Single	MRA
Single	0.78	**1.00**
MRA	0.54	0.89

Table 2: Ghost scores.

Ghosts \ PacMan	Single	MRA
Single	0.82	0.59
MRA	**1.00**	0.85

The cross-comparison results are shown in Table 1 and Table 2. We can see that the agents created by the proposed MRA outperform the single-MG counterparts for both PacMan agents and ghosts agents, validating the effectiveness of the proposed method.

E.3 IMPLEMENTATION VARIANTS

In role-symmetric games, the parameters θ, ϕ, ζ are shared by homogeneous agents. And ϕ can be implemented as the option architecture Sutton et al. (1999), *i.e.*, g corresponds to the z-th option sampled from the categorical distribution. We then conduct experiments on different instantiations of the $\phi(o, z)$ function in MRA.

We denote the default implementation architecture, where different relational graph from different attention heads is controlled by the lower-level latent variable z, as the option Sutton et al. (1999) architecture. Specifically, we apply the multi-head self-attention architecture Vaswani et al. (2017). And the head number is set as the dimension of the categorical distribution $p(z)$, and g corresponds to the graph at the z-th head when sampling z.

The other two variants we consider for implementing $\phi(o, z)$ are the ones that are discussed in Florensa et al. (2017). The first variant is to concatenate z to each entity of the observation decomposition $o^i = \left[o^i_s, o^i_1, \cdots, o^i_j, \cdots, o^i_N\right]$. The same relational representation is also adopted to generate the relational graph g. The second variant is to perform the outer product between each observation entity and z. We refer to the two variants as “concat” and “bilinear”, respectively.

The performance of the three implementations is evaluated in the 6-resource occupying environment. The number of training MGs is set to 3, and the numbers of agents in the 3 games are $\{6, 9, 12\}$. The results in Figure 5 show that all the three implementations can obtain agents that effectively act in all the 3 scenarios. And the default option architecture achieves better performance than the other two variants. The possible reason is that the lower-level latent code z in the option architecture can explicitly control the structural factors and can thus learn the common knowledge more quickly and better.

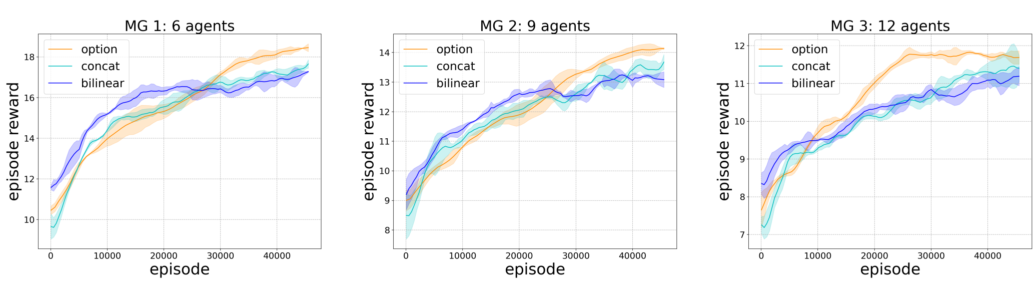


Figure 5: Performances of different implementation variants in the resource occupying environment.

E.4 Ablation Study on the Size of Training MG Set

The information in all the training MGs determines the common knowledge that agents can learn. We provide ablation study on the number of training MGs. In the resource occupying environment, we train the agents in three settings, each of which is with different size of training MG set: $2, 3$ and 4. Specifically, the population size of the three settings are: $\{6, 12\}$, $\{6, 9, 12\}$ and $\{6, 9, 12, 15\}$. The curves in MGs with $\{6, 9, 12\}$ populations are shown in Figure 6.

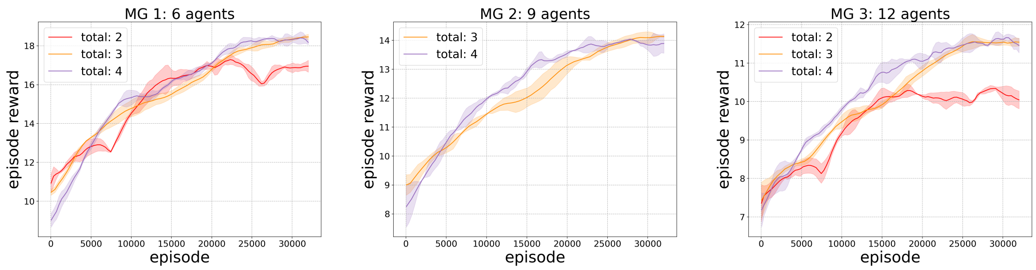


Figure 6: Ablation study on the size of the training MG set in the resource occupying environment.

We observe that when the size of the training MG set is greater than 2, the benefits of the meta-representation are obvious. The knowledge that agents learn from few training MGs, *e.g.*, 1 or 2 training MGs, is limited, and the random exploration bottleneck still exists. However, the performance can be significantly improved by leveraging the information from more training MGs, *e.g.*, 3 or 4 training MGs, where the common knowledge is more likely to be distilled and thus guiding the exploration.

E.5 Additional Curves

When training in multiple treasure collection MGs, the curves of MRA and other related multi-task or meta-learning approaches are shown in Fig. 7.

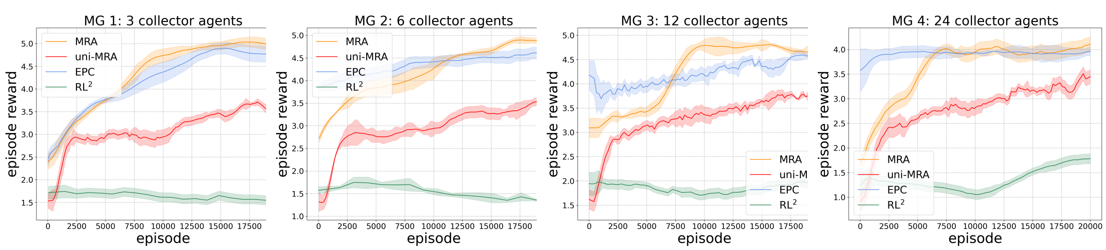


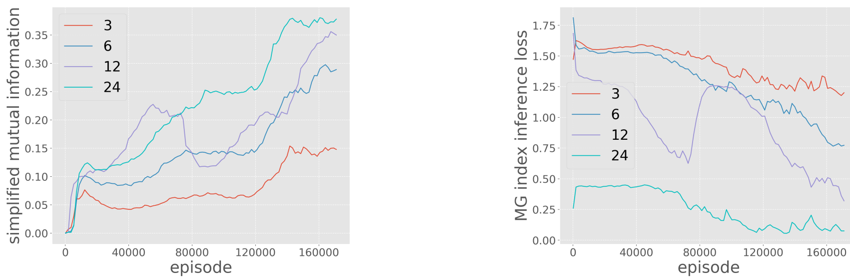
Figure 7: Multi-MG training curves in the treasure collection environment. Total number of training MGs are 4, with population $3, 6, 12, 24$.

We provide the training phase curves of the approximated mutual information $I(g; a|o)$ and the inference loss of MG index output by auxiliary network ξ. The curves are shown in Figure 8.

E.6 Visualizations

The screenshots of the three environments in the experiments are shown in Figure 9.

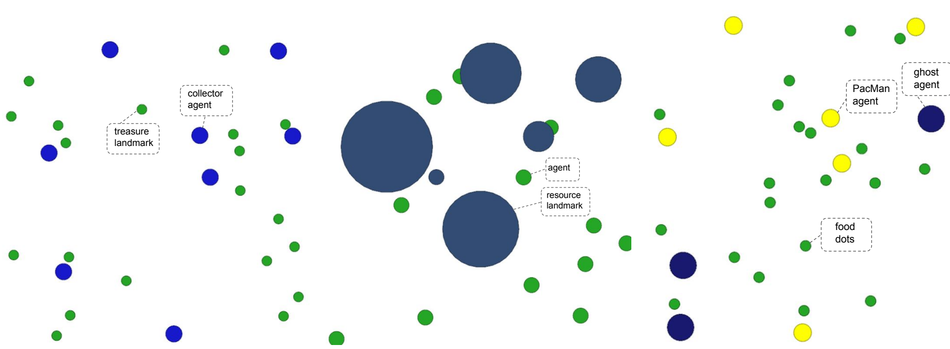
We visualize the trajectories of one agent in a resource occupying MG in Figure 10. Green dots and blue dots are agents and resources, respectively. When agents are only trained in this MG with different random seeds, different behaviors are obtained. This indicates that agents trained in single MGs are confined to environmental settings. Agents only learn the best responses and fit an NE. However, if the agents are only aware of some of the successful behaviors, the generalization



(a) The curve of mutual information $I(g; a|o)$ during training.

(b) The loss curve of the MG index inference.

Figure 8: Additional curves during training. The total number of training MGs is 4, with $\{3, 6, 12, 24\}$ agents in the resource occupying environment.



(a) Treasure collection.

(b) Resource occupying.

(c) Pacman-like world.

Figure 9: The illustration of the three environments that are used in our experiments.

will be constrained. On the contrary, MRA has large capacities to represent multiple strategies by incorporating different relational graph with the distilled common knowledge, which leads to diverse reasonable behaviors.

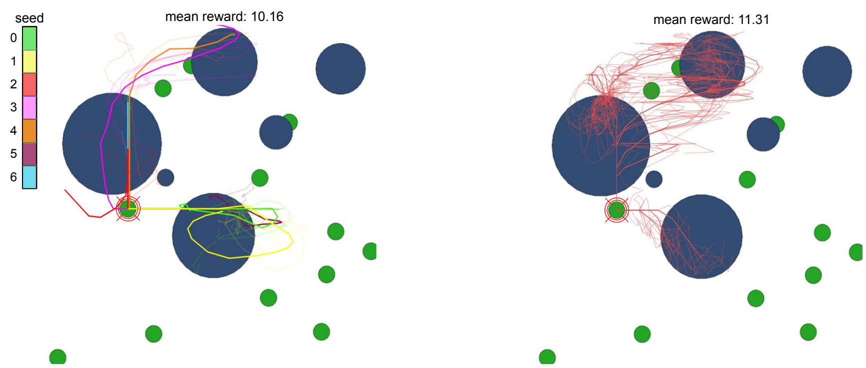
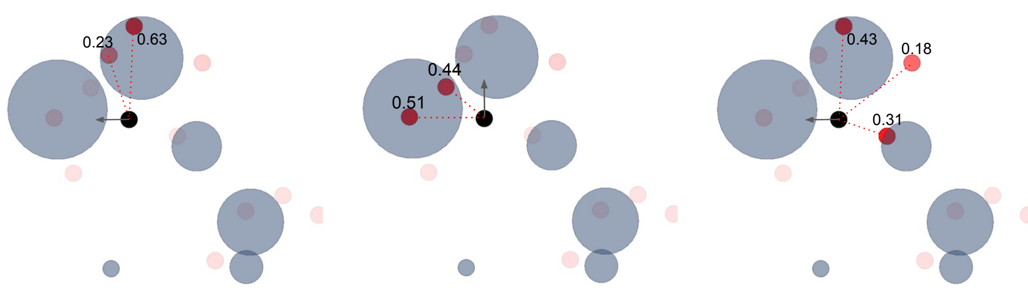


Figure 10: Trajectory visualization in resource occupying environment. **Left:** Trajectories of agents that are trained in a single MG, lighter colors are the exploration trajectories. **Right:** Trajectories of agents that are trained in multiple MGs.

We also visualize some instances of the learned relation variations, *i.e*., different relational graph g under observation o, as well as how agents make the smartest decisions under different variations in Fig. 11. The common knowledge learned by the agent can be interpreted as "moving to less-agent resources". Specifically, in Fig. 11(a) the black agent makes decisions to move left by focusing on the topmost red agents which are occupying a resource. By focusing on the leftmost red agents in Fig. 11(b), the black agent makes decisions to move up. Although such behavior might not be optimal, since the topmost resource is smaller than the leftmost resource, this variation helps agents learn common knowledge and optimally behave in an unseen MG by incorporating the optimal relation mapping in that game.



(a) (b) (c)

Figure 11: Instances of different learned relational graph and the corresponding actions that agents take. We visualize how the black agent makes different reasonable decisions by incorporating relational graph with the common knowledge. The relation scores that are smaller than 0.1 are not shown.