Hierarchical Mean-Field Deep Reinforcement Learning for Large-Scale Multiagent Systems

Chao Yu[1, 2]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] Pengcheng Laboratory, Shenzhen, China
yuchao3@mail.sysu.edu.cn

Abstract

Learning for efficient coordination in large-scale multiagent systems suffers from the problem of the curse of dimensionality due to the exponential growth of agent interactions. *Mean-Field* (MF)-based methods address this issue by transforming the interactions within the whole system into a single agent played with the average effect of its neighbors. However, considering the neighbors merely by their average may ignore the varying influences of each neighbor, and learning with this kind of local average effect would likely lead to inferior system performance due to lack of an efficient coordination mechanism in the whole population level. In this work, we propose a *Hierarchical Mean-Field* (HMF) learning framework to further improve the performance of existing MF methods. The basic idea is to approximate the average effect for a sub-group of agents by considering their different influences within the sub-group, and realize population-level coordination through the interactions among different sub-groups. Empirical studies show that HMF significantly outperforms existing baselines on both challenging cooperative and mixed cooperative-competitive tasks with different scales of agent populations.

Introduction

In many real-world *Multi-Agent System* (MAS) applications, such as city-level traffic light control (Wang et al. 2020; Chen et al. 2020), fleet management (Yu et al. 2019), multi-robot systems (Zhou and Xu 2020; Xia, Yu, and Wu 2021), and epidemics control (Libin et al. 2021; Dong, Yu, and Xia 2020), a large number of agents interact with each other to achieve a certain goal in the same environment. Learning for an optimal performance in this kind of large-scale systems is fundamentally challenging: each agent learning individually would normally perform poorly due to the high non-stationarity caused by concurrent interactions and policy updates of the agents, while learning jointly over the whole interactions would rapidly fail due to the huge joint state-action space which grows exponentially in the scale of agent population (Nguyen, Nguyen, and Nahavandi 2020).

Mean Field Reinforcement Learning (MFRL) methods (Luo et al. 2020; Yang et al. 2018; Guo et al. 2019; Carmona, Laurière, and Tan 2019; Subramanian et al. 2020) approximate the interactions within the population of agents by pair-wise interactions between each central agent and the average of its neighbors (or the overall population in the extreme case). While significantly reducing the complexity of interactions among agents, the existing MFRL methods still face the following downside issues. First, since all the existing methods approximate an agent's interactions with its neighbors merely by their average, the different influences of each individual neighbor cannot be distinguished. This can be problematic in some scenarios either when the agents are not necessarily fully homogeneous, or when they are homogeneous in terms of observations, action choices and rewarding mechanisms *etc.*, but possess different roles and thus distinct importance in influencing their neighbors. For example, in the prey and predator game, the importance of each predator in the group can differ to a large extent during the hunting process due to the dynamic change of their positions related to the prey. As such, it would be more efficient for the predators to distinguish their dynamic influences during their collective decision-making process. Second, since the MF approximation in the existing methods involves local Taylor expansion with respect to each central agent and ignores the high-order terms, in principle it is only applicable to modeling local interactions such that the approximation error can be maintained in an acceptable range. As a result, there lacks efficient coordination among those agents that are not within the neighborhood of each other, leading to potentially inferior performance in the whole population level. Last but not the least, the existing MFRL methods require the availability of the global state or the actions of neighbors for action selection in the execution process (Zhang et al. 2021). This premise significantly limits their applicability to scenarios with prefect communication or global observability of the agents.

In this paper, we propose a *Hierarchical Mean-Field* (HMF) deep reinforcement learning framework to alleviate the overall large-scale MAS coordination problem by decomposing the whole population into a number of groups and enabling two levels of coordinated learning processes among the agents: the bottom level for intra-group coordination and the top level for inter-group coordination. In specific, the value function of each agent is factorized into the combination of a local part and a group-interaction MF part. The agents take actions solely based on their local

value functions using locally accessible information, making our method applicable to wider scenarios when agents only have partial observability or their communication capabilities are constrained. Unlike all the existing methods that each agent maintains a specific neighbor-interaction MF function and learns directly over this function, we only model a *Group MF* (GMF) function for each group, thus capturing the overall learning characteristics for the agents in this group. Then, the top level in HMF enables coordinated learning among different groups over their GMF functions, in order to achieve system-level coordination of the whole population. The highlight of HMF lies in its capability of modeling multi-scale coordination in an agent system, ranging from individuals, to groups and the whole population, thus ensuring more efficient learning performance as well as better scalability.

Based on HMF, we then propose a dynamic grouping mechanism, *Information and Policy Consistency* (IPC), to align the new agents' policies with the group mean policy such that the overall approximation errors during the dynamic grouping process can be reduced. We assess the performance of our method by comparing it against some benchmark algorithms on cooperative tasks including the particle environments (Mordatch and Abbeel 2018) and the MAgent environment (Zheng et al. 2018), as well as a mixed cooperative-competitive task in the social dilemma *Bar Game* (Arthur 1994). Empirical studies show that HMF significantly outperforms existing baselines in different scales of agent populations.

Background

Stochastic Game

An MAS coordination problem can be modelled as an N-agent (or, N-player) stochastic game, defined as a tuple $\langle S, A^1, ..., A^N, r^1, .., r^N, p, \gamma \rangle$, where S is the state space, and A^k is the action space of agent k, and $\gamma \in [0, 1)$ is a discount factor. The reward function $r^k : S \times A^1 \times ... \times A^N \to \mathbb{R}$ provides real-valued rewards at each time step, and the transition function $p : S \times A^1 \times ... \times A^N \to \Omega(S)$ indicates the probability distribution over the next state $\Omega(S)$ when the system transits from state s given joint actions $\boldsymbol{a} = (a^1, ..., a^N)$ for all agents. At time step t, the agents choose actions according to their policies $\boldsymbol{\pi}$, where $\boldsymbol{\pi} \triangleq [\pi^1, ..., \pi^N]$ denotes the joint policy of all agents. The goal of the agent is to learn an optimal strategy to obtain the maximum expected discounted returns. Given an initial state s, the value function of agent k is expressed as the expected cumulative discounted rewards: $v^k_{\boldsymbol{\pi}}(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{p}}[r^k_t | s_0 = s]$. Based on the Bellman equation, the Q-function can then be formulated as: $Q^k_{\boldsymbol{\pi}}(s, \boldsymbol{a}) = r^k(s, \boldsymbol{a}) + \gamma \mathbb{E}_{s' \sim p}[v^k_{\pi}(s')]$, where s' represents the next state. The value function v^k_{π} can be expressed as $v^k_{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}}[Q^k_{\boldsymbol{\pi}}(s, \boldsymbol{a})]$. Since taking all of the global information as the input of each local Q-function is not scalable, a more direct way is to take solely the local information as state input, transforming the model to the Partially Observable Stochastic Games (Yang and Wang 2020).

Mean Field Theory

In large-scale stochastic games, learning the standard Q-function ahead is infeasible due to the exponential growth of agent interactions. To solve this problem, a natural way is to decompose the standard Q-function of each agent into a set of local Q-functions that capture pairwise interactions:

$$Q^k(s, \boldsymbol{a}) = \frac{1}{N^k} \sum_{j \in N(k)} Q^k(s, a^k, a^j), \tag{1}$$

where N^k is the number of neighbors of agent k and $N(k)$ is the index set of neighboring agents. Then, the Q-function can be approximated as the mean field Q-function $Q^k(s, \boldsymbol{a}) \approx Q^k_{\mathrm{MF}}(s, a^k, \bar{a}^k)$ according to the mean-field theory (Domb 2000), which approximates the interactions within the population of agents into a single agent played with the average effect from the overall (local) population. The mean action $\bar{a}^k = \frac{1}{N^k} \sum_{j \in N(k)} a^j$ represents the neighbor action distribution, where a^j is the action of each neighbor j. The mean field Q-function can be updated in a recurrent manner as follows:

$$Q^k_{t+1}(s, a^k, \bar{a}^k) = (1 - \alpha) Q^k_t(s, a^k, \bar{a}^k) + \alpha [r^k_t + \gamma v^k_t(s')], \tag{2}$$

where α is the learning rate. The value function $v^k_t(s')$ for agent k at time t is given by:

$$v^k_t(s') = \sum_{a^k} \pi^k_t(a^k | s', \bar{a}^k) \mathbb{E}_{\boldsymbol{a}^{-k} \sim \pi^{-k}_t}[Q^k_t(s', a^k, \bar{a}^k)]. \tag{3}$$

Then, the policy for each agent k is calculated as follows:

$$\pi^k_t(a^k | s, \bar{a}^k) = \frac{\exp(-\beta Q^k_t(s, a^k, \bar{a}^k))}{\sum_{a^{k'} \in A^k} \exp(-\beta Q^k_t(s, a^{k'}, \bar{a}^k))}. \tag{4}$$

However, this method depends on the global state and cannot be applied to environments with local observations. To address this problem, Zhang *et al.,* (Zhang et al. 2021) decomposed the individual joint Q-function into a local Q-function and a neighbor MF Q-function:

$$Q^k(s, \boldsymbol{a}) \approx Q^k_{\mathrm{LOC}}(o^k, a^k) + Q^k_{\mathrm{NB}}(\mu_o(\boldsymbol{o}^{-k}), \mu_a(\boldsymbol{a}^{-k})), \tag{5}$$

where $Q^k_{\mathrm{LOC}}(o^k, a^k)$ represents the agent k's efforts conditioned on local information, and $Q^k_{\mathrm{NB}}(\mu_o(\boldsymbol{o}^{-k}), \mu_a(\boldsymbol{a}^{-k}))$ represents its neighbors' influence conditioned on the average effect of neighboring agents.

The HMF Learning Framework

The MF Approximation

In our formulation, the individual joint Q-function for each agent is decomposed as follows:

$$Q^k(s, \boldsymbol{a}) = Q^k_{\mathrm{LOC}}(x_k) + Q^k_{\mathrm{IMF}}(\mu_o(\boldsymbol{x}), x_k), \tag{6}$$

where $x_k = (o_k, a_k)$ and $\mu(\boldsymbol{x})$ is the group's observation-action distribution. The local Q-function $Q^k_{\mathrm{LOC}}(x_k)$ is conditioned on its own information while the *Individual MF* (IMF) function Q^k_{IMF} is conditioned on the average information in agent k's neighborhood (including its neighbors

and itself). To bridge the gap between this neighborhood interaction and the global interaction, the individual joint Q-functions of all the agents within a group are added to form the group Q-function Q_{GRP}, weighted by the different influences of agents in the group:

$$Q_{\mathrm{GRP}}\left(\boldsymbol{x}\right)=\sum_{k\in N}\lambda_k\left(Q_{\mathrm{LOC}}^k(x_k)+Q_{\mathrm{IMF}}^k\left(\mu(\boldsymbol{x}),x_k\right)\right), \tag{7}$$

where $\lambda_k \in [0,1]$ is the weight representing the importance of agent k in the group and $\sum_{k\in N}\lambda_k = 1$. Denote $r(\mu(\boldsymbol{x}),x_k)$ as a virtual interaction reward between agent k and its neighbors such that the expected discounted sum of $r(\mu(\boldsymbol{x}),x_k)$ amounts to Q_{IMF}^k, Then, assuming this virtual reward can be simply decomposed into two rewards $r(\mu(\boldsymbol{x}_t))$ and $r(x_{k,t}))$, which represent the effect from the neighborhood and the agent itself, respectively, the weighted sum of the IMF Q-function can be transformed as follows:

$$\begin{aligned}
&\sum_{k\in N}\lambda_k Q_{\mathrm{IMF}}^k\left(\mu(\boldsymbol{x}),x_k\right)=\sum_{k\in N}\lambda_k\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r(\mu\left(\boldsymbol{x}_t\right),x_{k,t})\right]\\
&=\sum_{k\in N}\lambda_k\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}\left(r\left(\mu\left(\boldsymbol{x}_t\right)\right)+r\left(x_{k,t}\right)\right)\right]\\
&=\sum_{k\in N}\lambda_k\left\{\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r\left(\mu\left(\boldsymbol{x}_{t=1}\right)\right)\right]+\mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r\left(x_{k,t}\right)\right]\right\}\\
&:=\sum_{k\in N}\lambda_k\left[\bar{Q}\left(\mu\left(\boldsymbol{x}\right)\right)+\bar{Q}_k\left(x_k\right)\right]\\
&=\bar{Q}\left(\mu\left(\boldsymbol{x}\right)\right)+\sum_{k\in N}\lambda_k\bar{Q}_k\left(x_k\right),
\end{aligned} \tag{8}$$

where $\bar{Q}(\mu(\boldsymbol{x})) = \mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r\left(\mu\left(\boldsymbol{x}_t\right)\right)\right]$ and $\bar{Q}_k = \mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r\left(x_{k,t}\right)\right]$. Each agent's local observation-action can be expressed as the sum of the group's observation-action distribution $\mu(\boldsymbol{x})$ and a small fluctuation $\delta\bar{x}_k$:

$$x_k=\mu(\boldsymbol{x})+\delta\bar{x}_k,\quad \text{where}\quad \mu(\boldsymbol{x})=\sum_{k\in N}\lambda_k x_k. \tag{9}$$

Assuming sufficient differentiability of $\bar{Q}_k$, Equation (8) can be further decomposed based on the Taylor's theorem:

$$\begin{aligned}
&\bar{Q}\left(\mu\left(\boldsymbol{x}\right)\right)+\sum_{k\in N}\lambda_k\bar{Q}_k\left(x_k\right)\\
&=\bar{Q}\left(\mu(\boldsymbol{x})\right)+\sum_{k\in N}\lambda_k\Big[\bar{Q}_k(\mu(\boldsymbol{x}))+\\
&\quad\delta x_k^{\mu(\boldsymbol{x})}\nabla_{\mu(\boldsymbol{x})}\bar{Q}_k(\mu(\boldsymbol{x}))+R_{2,k}\Big]\\
&=\bar{Q}(\mu(\boldsymbol{x}))+\sum_{k\in N}\left[\lambda_k\bar{Q}_k(\mu(\boldsymbol{x}))+\lambda_k R_{2,k}\right]\\
&\approx\bar{Q}(\mu(\boldsymbol{x}))+\sum_{k\in N}\lambda_k\bar{Q}_k(\mu(\boldsymbol{x}))\\
&=2\bar{Q}(\mu(\boldsymbol{x})),
\end{aligned} \tag{10}$$

where the remainder of Taylor polynomial (*i.e.,* $R_{2,k}$) can be seen as a small fluctuation under some mild condition (the proof refers to (Yang et al. 2018)), and in the second equation in the above derivation, the first-order term is dropped since $\sum_{k\in N}\lambda_k\delta x_k^{\mu(\boldsymbol{x})} = 0$ by Equation (9). In addition, $\mu(\boldsymbol{x}) = \sum_{k\in N}\lambda_k x_k = N\times\boldsymbol{\lambda}\times\bar{x}$ where $\bar{x}=\frac{1}{N}\sum_{k\in N}x_k$ can be seen as the empirical distribution of the population action-observation, thus $\bar{Q}$ conditioned on $\mu(\boldsymbol{x})$ can capture the overall learning characteristics for the agents in the group. Therefore, to distinguish from the IMF function used in previous algorithms, we denote $\bar{Q}$ as Q_{GMF}, the *Group MF* (GMF) function, and thus Equation (7) can then be reformulated as follows:

$$Q_{\mathrm{GRP}}(\boldsymbol{x})=\sum_{k\in N}\lambda_k Q_{\mathrm{LOC}}^k(x_k)+2Q_{\mathrm{GMF}}(\mu(\boldsymbol{x})) \tag{11}$$

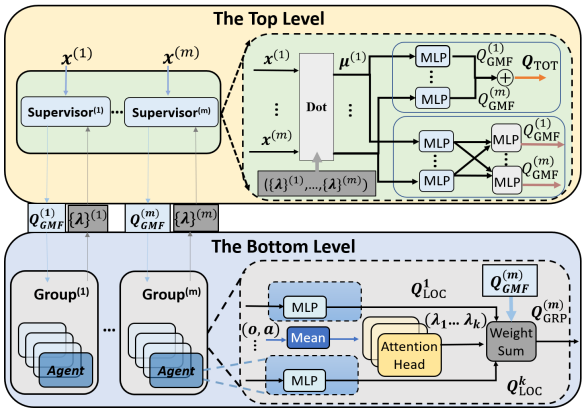


Figure 1: The HMF learning framework. Each group is associated with a virtual supervisor that governs the learning processes of multiple agents. At the bottom level, the average effect for a group is approximated by considering the different influences of the agents within the group based on the mean-field theory and attention mechanism, while the top level realizes system-level coordination through multiple supervisors learning over their group MF information.

The Hierarchical Learning Processes

The HMF learning framework consists of two levels: the bottom-level for intra-group coordination and the top-level for inter-group coordination. The overall HMF framework is illustrated in Figure 1.

The Bottom-Level Learning This level mainly focuses on the coordinated learning of agents within each group. As shown in Equation (6), the individual joint Q-function of each agent is decomposed to a local part that depends solely on locally observable information, and an IMF part that considers the interactions between this agent and its neighbors. Summing up all the individual joint Q-functions of a group of agents indicates the group Q-function as given in Equation (7). As shown in Equation (11), this group Q-function can be divided into two parts: the local Q-function of each agent and the group mean field function. To make HMF suitable for scenarios when agents are partially observable or their communication is restricted, the local Q-function takes locally observable information (*i.e.,* x_k) as in-

Algorithm 1: The HMF Learning Algorithm

Initial $Q_{\mathrm{LOC}}^{\theta_k}, \hat{Q}_{\mathrm{LOC}}^{\hat{\theta}_k}$ for all agents, and $Q_{\mathrm{GMF}}^{\phi^{(i)}}, \hat{Q}_{\mathrm{GMF}}^{\hat{\phi}^{(i)}}$ for all groups.
while training not finished **do**
  for each agent k **do**
    Sample action a_k from $Q_{\mathrm{LOC}}^{\theta_k}$ with ϵ-greedy policy.
  end for
  Take joint observations $\boldsymbol{o} = [o_k]_{k=1}^{N\times M}$, joint action $\boldsymbol{a} = [a_k]_{k=1}^{N\times M}$, joint reward $\boldsymbol{r} = [r_k]_{k=1}^{N\times M}$ and joint next observation $\boldsymbol{o'} = [o'_k]_{k=1}^{N\times M}$.
  Store $\langle \boldsymbol{o}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{o'} \rangle$ in the replay buffer D.
  Sample a mini-batch of K experience $\langle \boldsymbol{o}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{o'} \rangle$ from D.
  Get next action $\boldsymbol{a'} = [a']_k^{N\times M}$ from $[\hat{Q}_{\mathrm{LOC}}^{\hat{\theta}_k}]_k^{N\times M}$.
  for each group i **do**
    Set $Q_{\mathrm{GRP}}^{(i)}$ based on Equation (11).
    Update the Q-networks by minimizing the loss: $L_{\mathrm{GRP}}^{(i)} = \mathbb{E}_{\boldsymbol{o}_t^{(i)}, \boldsymbol{o}_{t+1}^{(i)}, \boldsymbol{a}_t^{(i)}, \boldsymbol{a}_{t+1}^{(i)}} (Q_{\mathrm{GRP}}^{(i)} - y^{(i)})^2$.
  end for
  Update the Q-network by minimizing the loss: $\mathcal{L}_{\mathrm{TOT}} = \mathbb{E}_{\boldsymbol{o}_t^{(i)}, \boldsymbol{o}_{t+1}^{(i)}, \boldsymbol{a}_t^{(i)}, \boldsymbol{a}_{t+1}^{(i)}} (Q_{\mathrm{TOT}} - y)^2$.
  Update the parameters of the target network for each agent and each group with learning rate τ:
    $\hat{\theta}_k \leftarrow \tau\theta_k + (1-\tau)\hat{\theta}_k$
    $\hat{\phi^{(i)}} \leftarrow \tau\phi^{(i)} + (1-\tau)\hat{\phi}^{(i)}$
end while

put, while the group mean field function takes the empirical group observation-action distribution, $\mu(\boldsymbol{x})$, as input. $\mu(\boldsymbol{x})$ is implemented as the weighted sum of $\{x_k\}_{k\in N}$, where the weight parameter (*i.e.,* λ_k) is computed using the attention mechanism (Vaswani et al. 2017), thus capturing different influences of the agents in the group.

The group Q-function $Q_{\mathrm{GRP}}^{(i)}$ for group i is updated to minimize the regression loss:

$$\mathcal{L}_{\mathrm{GRP}}^{(i)} = \mathbb{E}_{\boldsymbol{x}_t^{(i)}, \boldsymbol{x}_{t+1}^{(i)}} \left[Q_{\mathrm{GRP}}^{(i)} \left(\boldsymbol{x}_t^{(i)} \right) - y_t^{(i)} \right]^2, \tag{12}$$

where $y_t^{(i)} = r_t^{(i)} + \gamma \sum_{k\in N(i)} \lambda_{k,t+1} \hat{Q}_{\mathrm{LOC}}^k(x_{k,t+1}^{(i)}) + 2\hat{Q}_{\mathrm{GMF}}^{(i)}(\mu(\boldsymbol{x}_{k,t+1}^{(i)}))$ is the TD target, and $r_t^{(i)}$ is the sum of agents' rewards within the group at time step t, $\hat{Q}_{\mathrm{LOC}}$ and $\hat{Q}_{\mathrm{GMF}}$ denote the target Q-functions w.r.t. Q_{LOC} and Q_{GMF}, respectively.

The Top-Level Learning The top level realizes coordinated learning among different groups over their GMF functions, in order to achieve system-level coordination in the whole population. For each group i, we denote the group observation-action $(\{o\}^{(i)}, \{a\}^{(i)})$ as $\boldsymbol{x}^{(i)}$. Based on the weight value $\{\lambda\}^{(i)}$ from the bottom level, $\boldsymbol{x}^{(i)}$ is transformed into $\mu(\boldsymbol{x}^{(i)})$ of each group. The mean-field action-value network takes group $i's$ information $\mu(\boldsymbol{x}^{(i)})$ as input and calculates the GMF Q value $Q_{\mathrm{GMF}}(\mu(\boldsymbol{x}^{(i)}))$, and all the GMF Q values are summed up to derive the global Q value Q_{TOT} as follows:

$$Q_{\mathrm{TOT}}(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}) = \sum_{i\in m} Q_{\mathrm{GMF}}^{(i)}(\mu(\boldsymbol{x}^{(i)})), \tag{13}$$

where m is the number of groups, and Q_{TOT} is optimized by minimizing the loss function $\mathcal{L}_{\mathrm{TOT}}$:

$$\mathcal{L}_{\mathrm{TOT}} = \mathbb{E}_{\{\boldsymbol{x}_t^{(j)}, \boldsymbol{x}_{t+1}^{(j)}\}_{j=1}^m} \left(Q_{\mathrm{TOT}} \left(\left\{ \boldsymbol{x}_t^{(j)} \right\}_{j=1}^m \right) - y_t \right)^2, \tag{14}$$

where $y_t = r_t + \gamma \sum_{i\in m} \hat{Q}_{\mathrm{GMF}}^{(i)}(\mu(\boldsymbol{x}_{t+1}^{(i)}))$, and r_t is the sum of all the agents' rewards at time step t. In order to address the coordination problem in competitive tasks, we expand the top level learning with a central critic which evaluates the joint mean observation-action of multiple groups. Then, Q_{GMF} can be trained to minimize:

$$\mathcal{L}_{\mathrm{TOT}} = \mathbb{E}_{\mu(\boldsymbol{x}_t^{(i)}), \mu(\boldsymbol{x}_{t+1}^{(i)})} [Q_{\mathrm{GMF}}(\mu(\boldsymbol{x}_t^{(1)}), ..., \mu(\boldsymbol{x}_t^{(m)})) - y_t], \tag{15}$$

where y_t is given as follows:

$$y_t = r_t + \gamma \hat{Q}_{\mathrm{GMF}}(\mu(\boldsymbol{x}_{t+1}^{(1)}), ..., \mu(\boldsymbol{x}_{t+1}^{(m)})). \tag{16}$$

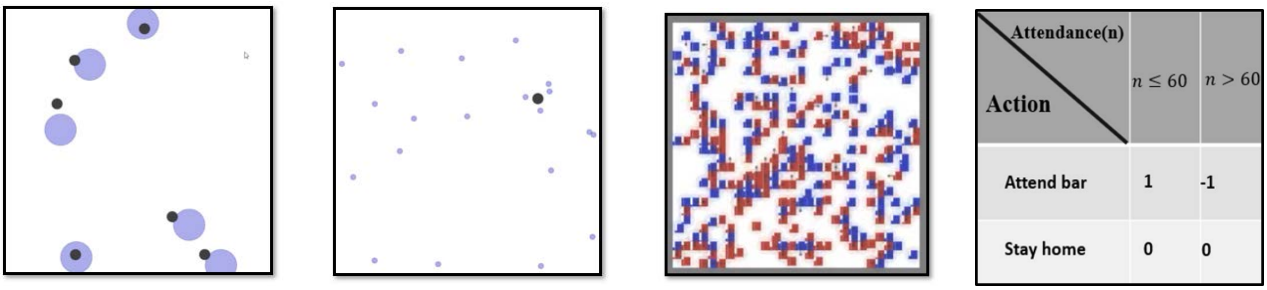
The pseudo code of HMF is given by Algorithm 1.

Error Reduction in the MF Approximation Note that $R_{2,k}$, *i.e.,* the remainder of Taylor expansion in Equation (10), increases w.r.t. $\delta x_k^{\mu(\boldsymbol{x})}$, *i.e.,* the distance between x_k and $\mu(\boldsymbol{x})$, since $R_{2,k}$ can be further expanded assuming sufficient differentiability: $R_{2,k} = \left(\delta x_k^{\mu(\boldsymbol{x})}\right)^2 \nabla^2_{\mu(\boldsymbol{x})} \bar{Q}_k(\mu(\boldsymbol{x})) + o\left(\delta x_k^{\mu(\boldsymbol{x})}\right)^2$. This implies that the error of the mean field approximation incurred by ignoring $\sum_{k\in N} \lambda_k R_{2,k}$ can be enormous especially in large-scale scenarios. To this end, we propose an error reduction technique to constrain the learning of each agent. This technique is expressed as a regularization term: $\alpha \sum_{k\in N} \left(x_k - \mu(\boldsymbol{x})\right)^2$, keeping the distance of agents to the center point within an acceptable range, in order to limit the overall error of mean field approximation.

The Dynamic Grouping Mechanism

The existing MF-based methods are all based on an assumption that the mean field approximation can accurately capture the average interactions between an agent and its neighbors. However, in some cases, particularly in large-scale dynamic systems when agents' behaviors are less likely similar even if they are close to each other, ignoring the higher-order remainder of the Taylor polynomial of the Q-function in these scenarios will lead to poor performance due to the errors in the mean field approximation process. To mitigate this issue, we propose the *Information and Policy Consistency* (IPC) mechanism in the HMF method in order to realize dynamic grouping in the learning process.

In specific, in the bottom-level learning, agents are grouped using the $K - Means$ method. At the beginning of each episode, a randomly chosen agent is treated as the central node μ of the first group. We then calculate the



(a) Spread (b) Pursuit Evasion (c) Battle (d) Bar Game

Figure 2: Spread: N agents, N landmarks. Agents are rewarded based on how far any agent is from each landmark. So, agents have to learn to cover all the landmarks. Pursuit Evasion: N pursuers(agents), 1 evader. Agents are rewarded based on the distance from evader to the closest pursuer. Battle: N agents, M enemies. The goal of each army is to get more rewards by collaborating with teammates to destroy all the opponents.

shortest distance to the central node μ for each agent k by $D_k = \arg min ||x_k - \mu_r||_2^2$, where r is the index of the central node. After each iteration, the node with the maximum distance in D_k is chosen as the new central node. Then, the mean information of each group i can be calculated according to the observation o_k and action a_k of agent k:

$$\bar{x}^{(i)} = \left(\frac{1}{N^{(i)}} \sum_{k \in N(i)} (o_k), \frac{1}{N^{(i)}} \sum_{k \in N^{(i)}} (a_k) \right). \quad (17)$$

Then, we can get the Euclidean distance between agent k and each of these groups: $d(k,i) = |x_k - \bar{x}^{(i)}|_2^2$, and the agent will be assigned to the group with smallest $d(k,i)$. The central node of the group will be updated according to the new added agents. In this way, the similar agents can be clustered in a group after several iterations.

However, the above dynamic grouping method will lead to diverse policies of agents in the group, making it hard for the group MF policy to adapt to new added agents. To resolve this issue, each group i will produce a *mean policy* $\bar{\pi}^{(i)}_{GMF}$, which reflects the average historical policy in this group, and the policy update of each agent is adjusted by the difference between the mean policy and each agent's individual policy, in order to lower the the new agents' influence on the group policy.

Specifically, we minimize the KL divergence between the policy distribution $\pi_k(o_k;\theta)$ of each agent k and the mean policy $\bar{\pi}^{(i)}_{GMF}$ as follows:

$$D_{KL}(\pi_k || \bar{\pi}^{(i)}_{GMF}) = \mathbb{E}\left[\log \pi_k - \log \bar{\pi}^{(i)}_{GMF}\right]. \quad (18)$$

The overall IPC framework and the pseudo code are provided in the Appendix.

Experiments

In this section, we first assess HMF on the *Multi-Agent Particle Environments* (MPE) (Mordatch and Abbeel 2018), *i.e.*, the *Spread* and the *Pursuit Evasion*, in order to evaluate the performance in relatively small-scale agent systems. Then, we evaluate the performance of HMF on the *Battle* task in the larger-scale *MAgent* (Zheng et al. 2018) environment involving hundreds of agents. We also conduct an experiment in the social dilemma *Bar* game (Arthur 1994) to show that HMF can also solve more complex coordination problems involving a large number of agents with mixed cooperative-competitive interests. Tables 1 and 2 in the Appendix respectively show the hyper-parameter setting of our methods and the other baselines.

MPE In the *Spread* task in Figure 2a, 6 agents have to cover a set of 6 landmarks while avoiding colliding with each other. The agents can observe the relative positions of other agents and the landmarks, and are rewarded based on the proximity of any agent to each landmark. In the *Pursuit Evasion* task in Figure 2b, 20 pursuers chase an evader using a strategy based on the Voronoi regions (Zhou et al. 2016). In order to encourage a higher level of coordination among the agents, the evader's maximum velocity is set twice as fast as the pursuers' maximum velocity. An episode ends once the evader is caught, *i.e.*, the distance between the closest pursuer and the evader is below a certain threshold. We compare HFM and its IPC version to two classic methods in PE, *i.e.,* MADDPG (Lowe et al. 2017) and VDN (Sunehag et al. 2017), and a latest method MFVFD (Zhang et al. 2021). Figure 3a and Figure 3b show the learning performance averaged over 4 random seeds. We can see that HMF and IPC converge to higher rewards more stably than the other methods. Specifically, MADDPG cannot converge to a reasonable level due to learning over joint state-action space. It is clear that such joint learning scheme would encounter severe dimension issues even in these relatively small domains. MFVFD and VDN perform better than MADDPG but a bit worse than HFM, which proves the benefits of our proposed hierarchical learning and the dynamic grouping mechanism in facilitating higher level of coordination among the agents.

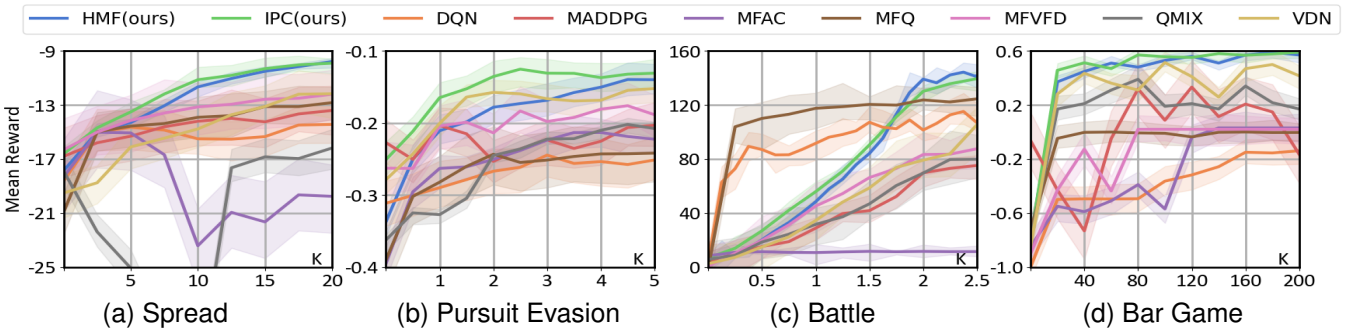


Figure 3: Comparison of different methods in the four different tasks, in terms of the average reward for each learning episode.

MAgent The *Battle* scenario in the MAgent environment shown in Figure 2c involves 200 agents learning to fight against 100 enemies who have superior abilities. The observation range, movement range and attack range of the enemy are twice those of the agents. We choose the pretrained DQN model in MAgent as the enemy. We choose DQN, MFQ (Yang et al. 2018) and MFVFD as our baselines. MADDPG and VDN are excluded as MADDPG cannot deal with such high-dimensional scenarios and VDN requires a team reward. The learning curves in Figure 3c show that HMF and IPC outperform the other baselines, converging to a much higher level of average rewards. Although MFQ and DQN can converge faster during the early period due to their independent learning nature, the final performance flattens at sub-optimal levels. Surprisingly, MFVFD performs rather poorly in this domain. We hypothesize that this is because the high complexity of computing the MF value function for each agent, making it less scalable in large-scale domains. Figure 2 in the Appendix visualizes the dynamic grouping process using the IPC method in the *Battle* scenario. As can be seen, after random initialization, the agents can be clustered quickly within a fixed range of neighborhood.

The Bar Game The *Bar* game in Figure 2d stands for a type of problems to study the emergence of cooperative behaviors in mixed cooperative-competitive environments involving a group of selfish agents. In this game, 100 agents decide whether to attend the bar on a certain night. The only observation available to the agents is the number of agents participating in the bar the night before. The payoff of attending the bar is high (+1) only if the number of attendees on the night is less than or equal to 60. Otherwise, the agents attending the bar would receive the worst payoff (-1) if the number of attendees is greater than 60. Thus, an agent is better off staying home if it believes that the bar would be crowded on that night. The results in Figure 3d show that all the baseline methods fail in this kind of mixed cooperative-competitive environments, except our HMF method that can achieve the near optimal performance of 0.6.

Scalability

We also investigate the scalability of HMF and IPC by comparing them to different algorithms in the MAgent environment with different scales of agent populations. Figure 4 shows that HMF and IPC can scale up to a thousand of agents and outperform existing baselines by a large margin in different population sizes.

Ablation Studies

We examine three aspects of HMF in influencing the final performance: the group size, the regularization constraint, and the availability of top-level learning. To investigate the influence of group sizes, we examine HMF with different numbers of groups in MPE, MAgent and the Bar game environments. As shown in Figure 5a and Figure 5b, in the smaller *Spread* domain, learning with only two groups performs better than with three groups, while in the larger *pursuit evasion* domain, an intermediate number of groups can achieve the best performance. These results indicate that a small number of supervisors are sufficient to efficiently coordinate the whole population if the population size is not too large, but too many supervisors would cause extra coordination burden among them such that the whole learning efficiency would be impaired. As the population size grows further in Figure 5c, more supervisors are required to achieve efficient coordination among the agents, since it is difficult for a supervisor to govern a large number of agents in each group. The distinction, however, is less apparent in the Bar game as shown in Figure 5d, which might be due to the more complex characteristics in this game where agents should behave cooperatively and competitively at the same time in order to achieve a satisfactory outcome. Figures 3 in the Appendix shows that HMF with regularization constraint performs better than the original HMF, and the top level in HMF enables coordinated learning among different groups over their GMF functions, such that the overall learning efficiency can be improved through system-level coordination in the whole population.

Related Work

Although from various perspectives and with distinct problem settings and assumptions, learning in large-scale multiagent systems has long been an interest in the AI community. The norm emergence paradigm (Airiau, Sen, and Villatoro 2014; Sen and Airiau 2007; Yu, Zhang, and Ren 2014; Yu et al. 2015) investigates how a consistent norm can be established in an agent population through learning from local interactions. Works in this area simply focus on toy problems

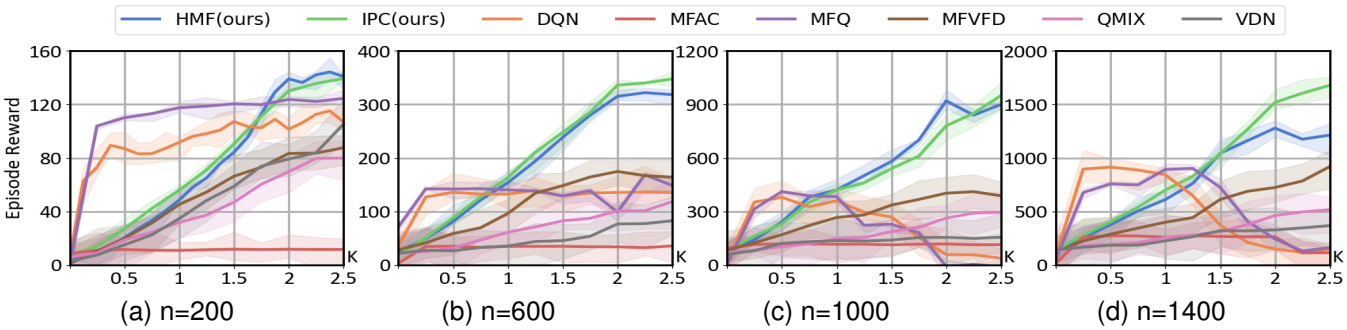


Figure 4: Learning curves in the MAgent environment with different scales of agent populations.

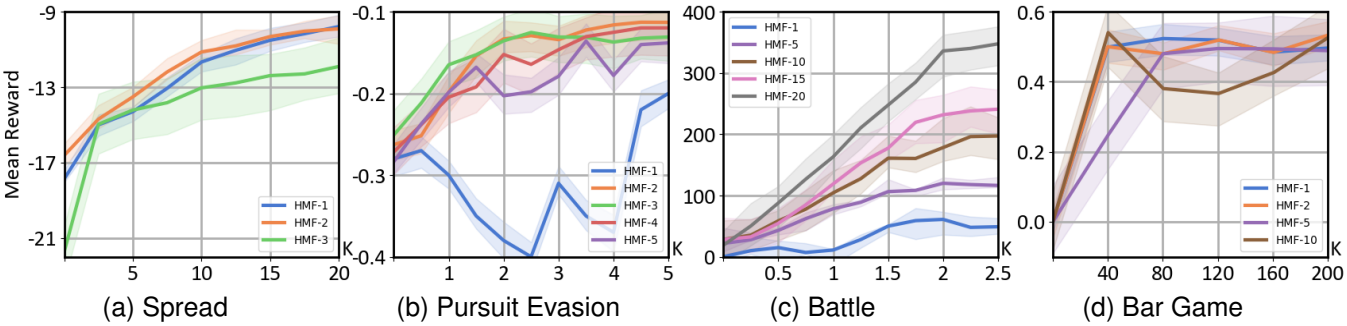


Figure 5: Impact of group size on HMF's performance. Curves of different colors correspond to different size of groups.

where agents' interactions can be modeled as certain matrix games, and thus cannot address more complex problems with high dimensions. Some studies (Qu et al. 2019; Zhang et al. 2018) resorted to fully decentralized learning schemes by transmitting local parameters among the agents, in order to scale the problems to large agent populations. However, a fixed network structure should be provided to enable such local transmissions, thus restricting the applications to limited settings. There are also several works (Ngiuem and Kumar 2018; Khan et al. 2018) that used count-based state representations to address large-scale agent learning problems. However, these methods are purely heuristic and do not enable agent-level control as we did in our work.

MF-based methods have emerged as a promising paradigm for efficient learning in large-scale agent populations. Mguni *et al.*, (2018) utilized fictitious play to achieve *Nash Equilibrium* (NE) in MF games and theoretically proved its convergence. The authors in (Yang et al. 2018; Subramanian and Mahajan 2019) proposed various MF methods such as the MF-Q, MF-AC to approximate the NE in MF games. Carmona *et al.*, (2019) extended the MF theory to continuous settings. Huttenrauch *et al.*, (2019; 2017) combined MF theory with deep reinforcement learning algorithms such as TRPO and DDPG in large-scale robot swarm systems. Unlike all these studies that the MF function approximates the interactions with the neighbors for each agent, HMF features a hierarchical learning scheme that a group MF function bridges the local interactions within each group and inter-group interactions for wider scope of system-level coordination.

Our work also shares some similarity with the hierarchical methods in RL research (Vezhnevets et al. 2017; Pateria et al. 2021). However, unlike these existing works that pay more attention to autonomous decomposition of challenging long-horizon decision-making tasks into simpler subtasks, our focus here is to employ hierarchical control to achieve more efficient learning in large-scale MASs.

Conclusion & Future Work

In this paper, we propose a hierarchical learning framework to equip the MF learning methods with a capability of system-level coordination in large-scale MASs. Through the bottom-level learning for intra-group coordination and the top-level learning for inter-group coordination, HMF is able to model multi-scale coordination in an agent system, ranging from individuals (through modelling different influences of the neighbors), to groups (through modelling a group MF function for each group to capture the overall learning characteristics in this group) and the whole population (through coordinated learning among different groups over their group MF functions), thus ensuring more efficient learning performance as well as better scalability. Empirical studies show that HMF significantly outperforms existing baselines on both cooperative and mixed cooperative-competitive tasks in different scales of agent populations. In the future, we will conduct more evaluation in other large-scale domains such as city-level light or traffic control.

Acknowledgments

We gratefully acknowledge support from the National Natural Science Foundation of China (No. 62076259), the Fundamental and Applicational Research Funds of Guangdong province (No. 2023A1515012946), and the Fundamental Research Funds for the Central Universities-Sun Yat-sen University.

References

Airiau, S.; Sen, S.; and Villatoro, D. 2014. Emergence of conventions through social learning. *Autonomous Agents and Multi-Agent Systems*, 28(5): 779–804.

Arthur, W. 1994. Inductive reasoning and bounded rationarity. *Am. Econ. Rev.*, 84: 488–500.

Carmona, R.; Laurière, M.; and Tan, Z. 2019. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. In *AAMAS*, 251–259.

Chen, C.; Wei, H.; Xu, N.; Zheng, G.; Yang, M.; Xiong, Y.; Xu, K.; and Li, Z. 2020. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3414–3421.

Domb, C. 2000. *Phase transitions and critical phenomena*. Elsevier.

Dong, Y.; Yu, C.; and Xia, L. 2020. Hierarchical reinforcement learning for epidemics intervention. *ACM Knowledge Discovery and Data Mining (KDD) Workshop*.

Guo, X.; Hu, A.; Xu, R.; and Zhang, J. 2019. Learning Mean-Field Games. *NeurIPS*, 32: 4966–4976.

Hüttenrauch, M.; Adrian, S.; Neumann, G.; et al. 2019. Deep reinforcement learning for swarm systems. *Journal of Machine Learning Research*, 20(54): 1–31.

Hüttenrauch, M.; Šošić, A.; and Neumann, G. 2017. Guided deep reinforcement learning for swarm systems. *arXiv preprint arXiv:1709.06011*.

Khan, A.; Zhang, C.; Lee, D. D.; Kumar, V.; and Ribeiro, A. 2018. Scalable centralized deep multi-agent reinforcement learning via policy gradients. *arXiv preprint arXiv:1805.08776*.

Libin, P. J.; Moonens, A.; Verstraeten, T.; Perez-Sanjines, F.; Hens, N.; Lemey, P.; and Nowé, A. 2021. Deep reinforcement learning for large-scale epidemic control. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*, 155–170. Springer.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*.

Luo, G.; Zhang, H.; He, H.; Li, J.; and Wang, F.-Y. 2020. Multiagent adversarial collaborative learning via mean-field theory. *IEEE Transactions on Cybernetics*, 51(10): 4994–5007.

Mguni, D.; Jennings, J.; and de Cote, E. M. 2018. Decentralised learning in systems with many, many strategic agents. In *AAAI*.

Mordatch, I.; and Abbeel, P. 2018. Emergence of grounded compositional language in multi-agent populations. In *AAAI*.

Ngiuem, D. T.; and Kumar, A. 2018. Credit assignment for collective multiagent RL with global rewards. *NeurIPS*, 2–8.

Nguyen, T. T.; Nguyen, N. D.; and Nahavandi, S. 2020. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9): 3826–3839.

Pateria, S.; Subagdja, B.; Tan, A.-h.; and Quek, C. 2021. Hierarchical Reinforcement Learning: A Comprehensive Survey. *ACM Computing Surveys (CSUR)*, 54(5): 1–35.

Qu, C.; Mannor, S.; Xu, H.; Qi, Y.; Song, L.; and Xiong, J. 2019. Value propagation for decentralized networked deep multi-agent reinforcement learning. In *NeurIPS*, 1184–1193.

Sen, S.; and Airiau, S. 2007. Emergence of norms through social learning. In *IJCAI*, volume 1507, 1512.

Subramanian, J.; and Mahajan, A. 2019. Reinforcement learning in stationary mean-field games. In *AAMAS*, 251–259.

Subramanian, S.; Poupart, P.; Taylor, M. E.; and Hegde, N. 2020. Multi Type Mean Field Reinforcement Learning. In *AAMAS*, 411–419.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. In *AAMAS*, 2085–2087.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.

Vezhnevets, A. S.; Osindero, S.; Schaul, T.; Heess, N.; Jaderberg, M.; Silver, D.; and Kavukcuoglu, K. 2017. Feudal networks for hierarchical reinforcement learning. In *ICML*, 3540–3549. PMLR.

Wang, X.; Ke, L.; Qiao, Z.; and Chai, X. 2020. Large-scale traffic signal control using a novel multiagent reinforcement learning. *IEEE transactions on cybernetics*, 51(1): 174–187.

Xia, L.; Yu, C.; and Wu, Z. 2021. Inference-based Hierarchical Reinforcement Learning for Cooperative Multi-agent Navigation. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 57–64. IEEE.

Yang, Y.; Luo, R.; Li, M.; Zhou, M.; Zhang, W.; and Wang, J. 2018. Mean field multi-agent reinforcement learning. In *ICML*, 5571–5580. PMLR.

Yang, Y.; and Wang, J. 2020. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*.

Yu, C.; Wang, X.; Xu, X.; Zhang, M.; Ge, H.; Ren, J.; Sun, L.; Chen, B.; and Tan, G. 2019. Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs. *IEEE Transactions on Intelligent Transportation Systems*, 21(2): 735–748.

Yu, C.; Zhang, M.; and Ren, F. 2014. Collective Learning for the Emergence of Social Norms in Networked Multiagent Systems. *IEEE Transactions on Cybernetics*, 44(12): 2342–2355.

Yu, C.; Zhang, M.; Ren, F.; and Tan, G. 2015. Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE transactions on neural networks and learning systems*, 26(12): 3083–3096.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2018. Fully decentralized multi-agent reinforcement learning with networked agents. In *ICML*, 5872–5881. PMLR.

Zhang, T.; Ye, Q.; Bian, J.; Xie, G.; and Liu, T.-Y. 2021. MFVFD: A Multi-Agent Q-Learning Approach to Cooperative and Non-Cooperative Tasks. In *IJCAI*, 500–506.

Zheng, L.; Yang, J.; Cai, H.; Zhou, M.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In *AAAI*, volume 32, 8222–8223.

Zhou, Z.; and Xu, H. 2020. A novel mean-field-game-type optimal control for very large-scale multiagent systems. *IEEE Transactions on Cybernetics*.

Zhou, Z.; Zhang, W.; Ding, J.; Huang, H.; Stipanović, D. M.; and Tomlin, C. J. 2016. Cooperative pursuit with Voronoi partitions. *Automatica*, 72: 64–72.