Tacit Commitments Emergence in Multi-agent Reinforcement Learning

Boyin Liu[1,2], Zhiqiang Pu[1,2], Junlong Gao[3], Jianqiang Yi[1,2], and Zhenyu Guo[3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3] Alibaba Group, Hangzhou, China
{liuboyin2019,zhiqiang.pu,jianqiang.yi}@ia.ac.cn
{Junlong Gao,Zhenyu Guo}@alibaba-inc.com

Abstract. Tacit commitments have been widely seen as a crucial underpinning for real-world cooperation. Similarly, it could also be a key to multi-agent cooperation. This paper proposes a novel tacit commitment emergence multi-agent reinforcement learning (MARL) framework (TCEM). In MARL, we define commitment as the unique state that the agent will exhibit through its action. TCEM first equips each agent with a commitment inference module (CIM) to infer its neighbor's commitments. Then, TCEM proposes that commitments influence intrinsic motivation (CIR) to encourage agents to have casual influence on others' actions. Finally, commitment acceptance intrinsic (CAI) motivation is constructed to guide the agent in behaving considering neighbors' commitments. CIR and CAI calculate intrinsic reward using counterfactual reasoning deriving from causal inference. Empirical results show that our method can effectively improve learning performance and deliver better cooperation among agents, which helps our method show superior performance on the Google Research Football benchmark.

Keywords: Casual Inference · Counterfactual Reasoning · Intrinsic Reward · Multi-agent Systems · Reinforcement Learning.

1 Introduction

In recent years, cooperative multi-agent reinforcement learning (MARL) has achieved meaningful progress, and many deep approaches have been proposed [3, 15, 12, 8]. However, learning complicated and effective coordination policies among agents is still a challenge in the MARL field.

The human ability to make and stick to commitments is crucial to human social cooperation [9, 1]. Tacit commitment induces effective cooperation behavior, which improves the productivity and efficiency of human society [10]. Analogically, the emergence of tacit commitment should also be essential for multi-agent cooperation. Tacit commitments mean that the transmission of intention among agents is implicit without explicit communication. For the commitments receiving agent, it is necessary to understand other agents‘ commitments through

historical observation. Then, its behaviors must be influenced by others' commitments and also make corresponding commitments.

In this paper, we propose a novel tacit commitment emergence MARL framework (TCEM) to develop effective cooperation among agents. To equip agents with the ability to know others' commitments, TCEM first learns a commitment inference module (CIM) to predict neighbors' commitments. Humans usually infer others' commitments according to their historical behavior and current observation. CIM is trained using neighbors' past behavior data and outputs their commitments according to current observation. To further encourage tacit commitment emergence, we propose two terms of intrinsic motivation, i.e., commitments influence intrinsic motivation (CIR) and commitments acceptance intrinsic motivation (CAI). CIR gives an agent an additional reward for having a casual commitment influence towards others. With this intrinsic reward, we hope the agent makes meaningful commitments. CAI encourages the agent to be influenced by its inferring neighbors' commitments. CIR and CAI form a closed circle, one for making commitments and another for accepting commitments.

We benchmark our approach on Google Research Football (GRF) benchmark. The superior performance of our approach on challenging benchmarking tasks shows that our approach achieves significantly higher coordination capacity than baselines while using tacit commitment as a catalyst for more robust talent policies.

2 Related Work

A fully cooperative multi-agent task can be formulated as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [11]. There are n agents in the environment, where each agent i receives a local observation o_i^t and then executes its action a_i^t and gets a shared reward r^t. All agents aim to maximize their expected return $E[\sum_{t=0}^{T} \gamma^t r^t]$, where γ denotes the discount factor and T the time horizon. Many methods have been proposed for Dec-POMDP, most of which follow the paradigm of centralized training and decentralized execution (CTDE). One of the promising ways to implement the CTDE framework is value function factorization [12, 14, 13, 19]. QMIX [12] proposes monotonicity for factorization structures. QPLEX [15] uses a duplex dueling network architecture for factorization. Except for value decomposition methods, policy gradient method [17, 7, 3, 4, 18] is also popular in MARL field. COMA [3] proposes a counterfactual baseline for multi-agent credit assignment.

To develop MARL methods qualified for complex tasks, many concepts in human collaboration are introduced into MARL, such as role [16], diversity [2, 8], individuality [5], etc. ROMA [16] constructs a stochastic role embedding space by introducing two novel regularizers and conditioning individual policies on roles. MAVEN [8] learns a diverse ensemble of monotonic approximations with the help of a latent space to explore. In this paper, we introduce tacit commitments into MARL to help agents construct more effective cooperation.

3 Method

In this section, we propose a novel tacit commitments emergence MARL framework (TCEM) that guides agents to learn high-level cooperation. TCEM adopts the CTDE paradigm. As shown in Fig.1, to enable tacit commitments emergence among agents, TCEM firstly equips each agent with a commitments inference module (CIM). CIM allows agents to speculate on other agents' commitments based on the observed historical trajectory. Then, the inferred commitments combined with the agent's hidden states are fed into a multi-layer perception network to output the Q function of the agent. More importantly, TCEM proposes two terms of intrinsic reward, one for encouraging the agent to exert its commitment to others and another for guiding the agent to learn to be influenced by others' commitments.

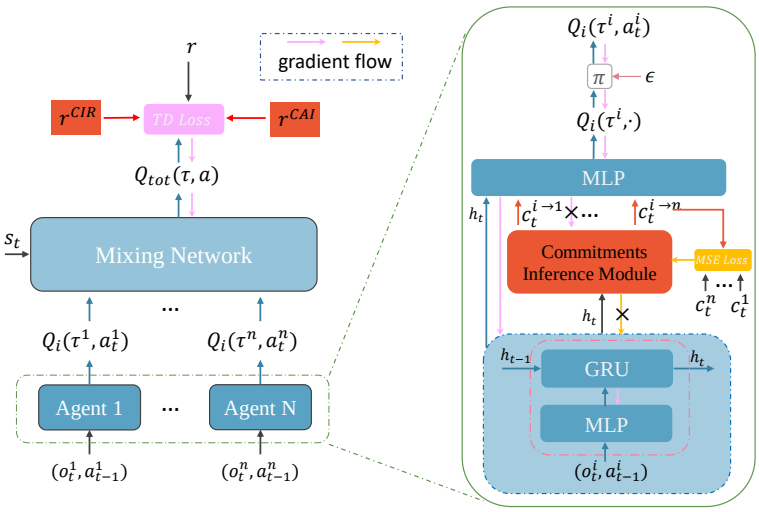

Fig. 1. Schematics of our approach. $c_t^{i\to n}$ denotes the inferred n's commitments by i. c_t^n represents n's real commitments, observed by other agents after several steps. Additionally, the TD Loss gradient does not influence the parameters in CIM , which updates its parameters isolately.

3.1 Commitments Inference Module

In high-level human cooperation, such as professional football, people usually make decisions based on their teammates' unspoken commitments. Superior coordination is shown when the decision-maker infers the teammates accurately. To learn tacit commitments among agents, we first need to define the agent's commitments.

Definition 1 *The **controllable states** of an agent is the states that can be directly and completely achieved by agent itself through corresponding actions.*

Remark 1. For example, in football game, the controllable states refer to the states such as position and direction of the agent itself but not the positions of the ball, neighbors or opponents.

Definition 2 *The **commitments** c_i of the agent i is its expected performing controllable states at next step.*

To construct tacit commitments among agents, each agent must learn to infer the possible commitments of its neighbors. As agents constantly learn and improve their policies, they must also constantly learn to infer their neighbors' commitments. Therefore, commitments need to be learned by neural networks. We adopt a MLP to construct the CIM, with extracted hidden states as input and inferred neighbors' commitments as output. CIM can be updated when the agent sees the real controllable states of its neighbors. It is noted that the gradients of TD Loss do not influence the parameters of CIM.

3.2 Commitments Influence Intrinsic Motivation

Based on CIM, we then propose commitments influence intrinsic reward (CIR) to motivate the agent having a casual influence on another agent's actions. Suppose there are two agents i and j, agent i infers $c_t^{i\to j}$ the commitments c_t^j of agent j using CIM. The prediction of $c_t^{i\to j}$ is based on the history observation of agent j. Thus, the agent i computes the probability of its next action as $p(a_t^i|o_t^i, c_t^{i\to j}, c_t^{i\to\sim(i,j)})$ where $\sim(i,j)$ denotes agents $1, 2, ..., n$ except agents i and j. We then replace $c_t^{i\to j}$ with counterfactual commitments $\tilde{c}_t^{i\to j}$, and calculating the counterfactual policy distribution $p(a_t^i|o_t^i, \tilde{c}_t^{i\to j}, c_t^{i\to\sim(i,j)})$. The agent j's commitments influence towards i is then computed by asking a question: How would agent i's policy distribution change if i had only inferred j's commitments different? If agent i's policy change as the inferred agent j's commitments change, we can say that agent j has commitments influence towards agent i.

The marginal distribution of $p(a_t^i|o_t^i, \tilde{c}_t^{i\to j}, c_t^{i\to\sim(i,j)})$ can not be directly computed. We achieve it by sampling several counterfactual commitments of agent j:

$$p\left(a_t^i \mid c_t^{i\to\sim(i,j)}, o_t^i\right) = \sum_{\tilde{c}_t^{i\to j}} p\left(a_t^i \mid \tilde{c}_t^{i\to j}, c_t^{i\to\sim(i,j)}, o_t^i\right) p\left(\tilde{c}_t^{i\to j} \mid c_t^{i\to\sim(i,j)}, o_t^i\right). \quad (1)$$

Thus, the causal influence reward c_{CIR} for agent j is defined as:

$$c_{CIR}^j = \sum_{i=0, i\neq j}^{N} \left[D_{KL}\left[p\left(a_t^i \mid c_t^{i\to j}, c_t^{i\to\sim(i,j)}, o_t^i\right) \| p\left(a_t^i \mid c_t^{i\to\sim(i,j)}, o_t^i\right)\right]\right]. \quad (2)$$

3.3 Commitments Acceptance Intrinsic Motivation

CIR encourages agents to be influencers. To show tacit commitments emergence among agents, the agent that is influenced by inferred other agents' commitments, could also be rewarded. Therefore, we propose commitments acceptance intrinsic motivation (CAI) to give an agent an additional reward for being influenced by inferred other agents' commitments. For agent j, it calculates CAI by asking a retrospective question: How would agent j's policy distribution change if i had inferred neighbors' commitments different?

Thus, the causal influence reward c_{CAI} for agent j is defined as:

$$c_{CAI}^{j} = \left[D_{KL} \left[p\left(a_t^j \mid c_t^{j \to \sim j,}, o_t^j \right) \| p\left(a_t^j \mid o_t^j \right) \right] \right]. \tag{3}$$

3.4 Overall Learning Objective

In this paper, the proposed learning framework TCEM adopts QMIX [12] style mixing network. With CIR and CAI, the final intrinsic reward function is calculated as:

$$r_I = \beta_1 r_{CIR} + \beta_2 r_{CAI}, \tag{4}$$

where β_1 and β_2 are weight coefficient of CIR and CAI respectively.

we add r_I to extrinsic rewards r and use the following TD loss:

$$\mathcal{L}_{TD}(\theta) = \left[r + \beta r_I + \gamma \max_{a'} Q_{tot}\left(s', a'; \theta^{-}\right) - Q_{tot}(s, a; \theta) \right]^2, \tag{5}$$

where θ is the parameters of the whole framework, θ^{-} is periodically fixed parameters copied from θ for a stable update, and β is the weight of intrinsic rewards.

4 Experiments Setup

To clearly interpret the mechanism and show the effectiveness of TCEM, we evaluate our method in the scenarios in Google Research Football (GRF) [6]. In this section, we describe the environment and experimental setup.

4.1 Environments

Football is a game that needs high-level cooperation among agents, and players essentially require tacit commitment to score goals. As shown in Fig.2, we chose three challenging football tasks i.e., *academy 3vs1 with keeper*, *academy 3vs3*, and *academy counterattack hard* to test the learning performance of TCEM and other baselines. In these tasks, agents need to choose an action from 19 actions at each step, including move, pass, shot, etc. In our experiments, we control left-side players (in yellow) except the goalkeeper. The right-side players are rule-based

Fig. 2. Initial snapshot of three GRF tasks.

bots controlled by the game engine. All the agents must coordinate well, and then it is possible to overcome the opponent and score a goal. Except for the goal reward of +80, a reward +5 is given for the ball being first controlled by agents.

To speed up training and reduce useless exploration, the episodes are terminated either when some events happen (including score, ball possession loss, and game stops) or the time steps exceeding 400).

4.2 Baseline and Ablation Methods Setup

In this section, we compare our methods with QMIX [12], MAVEN [8] and COMA [3]. In addition, we carry out the following ablation studies: (1) QMIX-CIM. QMIX with CIM module to infer other agents' commitments. (2)QMIX-CIR (QMIX-CIM-CIR). Based on QMIX-CIM, we add commitments influence intrinsic motivation to encourage agents to behave with influence. (3) QMIX-CAI (QMIX-CIM-CAI). Based on QMIX-CIM, we add commitments acceptance and intrinsic motivation to encourage agents to be influenced by others' commitments.

The content of controllable states are essential to effectively tacit commitment emergence. In GRF task, the agent's controllable state is its position and direction. For agent i, the agent j's commitments for him is its relative position and direction after m steps towards current i and ball's position and direction. Football is a game about the ball. Therefore, we also add the ball's information into the commitments design. Agent j's commitments about the ball are its expected position or direction after m steps relative to the current ball, so it is still controllable for agent j. We choose $m = 2$ in GRF tasks.

For all experiments, the optimization is conducted using RMSprop with a learning rate of 5×10^4, α of 0.99, and with no momentum or weight decay. For exploration, we use ϵ-greedy, with ϵ annealed linearly from 1.0 to 0.05 over 500K time steps and kept constant for the rest of the training for both TCME and all the baselines and ablations. We introduce three important hyperparameters: β, β_1, β_2. The $(\beta, \beta_1, \beta_2)$ of our methods is shown in Table 4.2. The intrinsic motivation stops after 600 million steps of training. In addition, to reduce randomness, we show the average and variance of the performance for our method, baselines, and ablations tested with three random seeds.

Table 1. Parameters setting of our methods for all tasks.

Methods	Parameters Setting $(\beta, \beta_1, \beta_2)$
TCEM	$(0.08, 1, 2)$
QMIX-CIM	$(0, 0, 0)$
QMIX-CIR	$(0.08, 1, 0)$
QMIX-CAI	$(0.08, 0, 2)$

5 Experiments Results

5.1 Comparison Results of Baseline Methods

As illustrated by Fig. 3, TCEM outperforms all other methods for each task with acceptable variance across random seeds. With tacit commitments introduced, TCEM evidently learns better strategy and forms effective coordination faster. The baseline QMIX can achieve satisfactory performance compared with MAVEN and COMA. However, at *academy 3vs1 with keeper* and *academy 3vs3* tasks, QMIX converges to the strategy with a lower wining rate compared with TCEM. In addition, TCEM always rises faster than QMIX without falling into the local optimum. At *academy 3vs1 with keeper* task, although TCEM's learning curve rises later than QMIX, it rises faster and achieves better performance finally. Among these baselines, MAVEN and COMA fail to show meaningful strategy learning.

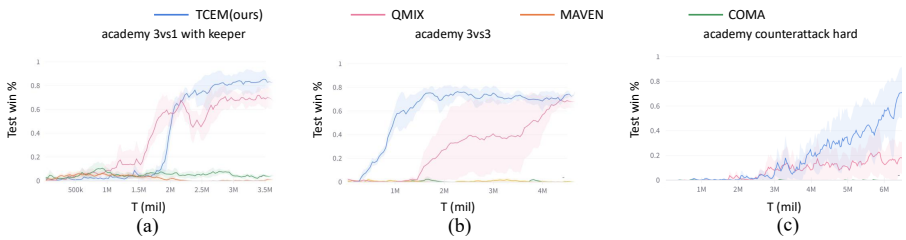

Fig. 3. Comparison of our method against baseline algorithms.

5.2 Comparison Results of Ablation Methods

In this section, we conduct ablation studies, comparing with the ablations explained in Section 4.2 at three GRF tasks. As shown in 4, TCEM offers the best performance among ablation methods. The ablation of each part of our method will induce an evident decrease in learning performance.

The superiority of TCEM against QMIX-CIM highlights the contribution of CAI and CIR. By comparing TCEM with QMIX-CAI, we can conclude that the CIM effectively improves learning performance both in speed and quality. In

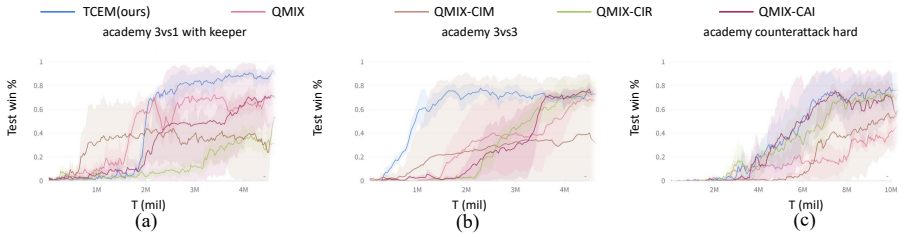

Fig. 4. Comparison of our method against ablation methods.

academy 3vs1 with keeper, there is only two right team players to defend MARL agents. It is noted that QMIX also shows superior learning performance in this easy task, even better than QMIX-CIM, QMIX-CIR, and QMIX-CAI. However, when it comes to *academy 3vs3*, a more challenging task, QMIX-CIR and QMIX-CAI both outperform QMIX. Finally, in *academy counterattack hard*, the most challenging task, even QMIX-CIM, shows better performance than QMIX. These results correspond to the intuition that tacit commitments have broader potential in a complex task. CIM infers commitments and extracts information from observations. The inferred commitments do not provide additional information for agents. Therefore, we observe that QMIX-CIM fails to show better performance than QMIX. Based on QMIX-CIM, adding CIR or CAI will bring noticeable performance promotion for complex tasks. This is because CAI and CIR will guide the agents to notice inferred commitments and form better cooperation. The comparison of the performance of TCEM with QMIX-CIR and QMIX-CAI proves that the commitments influence and acceptance intrinsic motivation loop can effectively speed up training and increase stability.

The performance gap between TCME and ablations is more evident on harder tasks. This observation supports the previous discussion – tacit commitment is more likely to improve labor efficiency in complex tasks.

5.3 Policy Visualization

We further visualize the final learning policy by TCEM at *academy counterattack hard* task, which shows evident tacit commitment emergence between agents.

As shown in Fig. 5, we control the yellow team players. At the start, shown in Fig. 5(a), the yellow player 1 dribbles the ball and runs down to draw the attention of blue player 2. It is noted that the movement of blue player 2 makes space for yellow player 2, who also runs below to bypass the defender. At the same time, yellow player 3 also runs to the backcourt to distract defenders. As shown in Fig. 5(b), from $t = 11$ to $t = 30$, yellow player 1 spends 19 steps dribbling the ball and distracting blue players 1 and 2 with yellow player 3's support. With the help of teammates, yellow player 2 successfully runs to the backfield. At $t = 30$, yellow player 1 suddenly kicks the ball to the penalty arc. The ignored yellow player 2 then runs fast to the penalty arc and successfully receives the ball. Finally, yellow player 2 dribbles the ball to the penalty area and overcomes

the goalkeeper's defense. These mutual commitments among yellow players have led to efficient cooperation and reflect the tacit commitments learned by TCEM.

(a) $t = 11$ (b) $t = 30$

Fig. 5. Visualization of learning policies by TCEM at *academy counterattack hard* task, which achieve complex cooperation with impressive off-the-ball moving strategies.

Throughout the yellow team players' behaviors, there are tacit commitments among agents. First, at the dribbling ball phase, both yellow team players 1 and 3 move to distract two blue players, making commitments that they will draw the blue players' attention. The yellow team player 2 receives the commitments of their teammates and also makes a commitment that he will directly run to the backfield. When it comes to the pass, it is noted that yellow player 1 does not pass the ball straight to yellow player 2, but kicks the ball to the penalty arc. Left team player 2 has made a commitment. Yellow player 1 kicks the ball to receive the commitment from yellow player 2 that he will run to the penalty arc and receive the ball.

6 Conclusion

Many cooperation concepts in human society have shown meaningful potential to improve performance in cooperative MARL. In this paper, we propose TCEM to introduce tacit commitments into MARL to promote cooperation performance among agents. Experimental results demonstrate that our method accelerates the learning and improves the learning performance, which performs best in all GRF tasks. When compared to the baseline methods, the results confirm that TCEM is significantly superior to QMIX, COMA, and MAVEN. Ablation results suggest that each part in TECM makes a meaningful contribution towards ultimate superior performance. Finally, the critical snapshot analysis confirms that TCEM indeed learns tacit commitments among agents.

Acknowledgments. This work was supported by the National Key Research and Development Program of China under Grant 2020AAA0103404, the Na-

tional Natural Science Foundation of China under Grant 62073323 and Alibaba Group through Alibaba Innovative Research (AIR) Program.

References

1. Agranov, M., Potamites, E., Schotter, A., Tergiman, C.: Beliefs and endogenous cognitive levels: An experimental study. Games and Economic Behavior **75**(2), 449–463 (2012)
2. Chenghao, L., Wang, T., Wu, C., Zhao, Q., Yang, J., Zhang, C.: Celebrating diversity in shared multi-agent reinforcement learning. Advances in Neural Information Processing Systems **34** (2021)
3. Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., Whiteson, S.: Counterfactual multi-agent policy gradients. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
4. Iqbal, S., Sha, F.: Actor-attention-critic for multi-agent reinforcement learning. In: International Conference on Machine Learning. pp. 2961–2970. PMLR (2019)
5. Jiang, J., Lu, Z.: The emergence of individuality. In: International Conference on Machine Learning. pp. 4992–5001. PMLR (2021)
6. Kurach, K., Raichuk, A., Stańczyk, P., Zajac, M., Bachem, O., Espeholt, L., Riquelme, C., Vincent, D., Michalski, M., Bousquet, O., et al.: Google research football: A novel reinforcement learning environment. arXiv preprint arXiv:1907.11180 (2019)
7. Lowe, R., Wu, Y.I., Tamar, A., Harb, J., Pieter Abbeel, O., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems **30** (2017)
8. Mahajan, A., Rashid, T., Samvelyan, M., Whiteson, S.: Maven: Multi-agent variational exploration. Advances in Neural Information Processing Systems **32** (2019)
9. Melkonyan, T., Zeitoun, H., Chater, N.: The cognitive foundations of tacit commitments. Available at SSRN 3168669 (2018)
10. Nguyen, N.L.: Tacit knowledge sharing within project teams: an application of social commitments theory. VINE Journal of Information and Knowledge Management Systems (2021)
11. Oliehoek, F.A., Amato, C.: A concise introduction to decentralized POMDPs. Springer (2016)
12. Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., Whiteson, S.: Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: International Conference on Machine Learning. pp. 4295–4304. PMLR (2018)
13. Son, K., Kim, D., Kang, W.J., Hostallero, D.E., Yi, Y.: Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In: International Conference on Machine Learning. pp. 5887–5896. PMLR (2019)
14. Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W.M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J.Z., Tuyls, K., et al.: Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296 (2017)
15. Wang, J., Ren, Z., Liu, T., Yu, Y., Zhang, C.: Qplex: Duplex dueling multi-agent q-learning. arXiv preprint arXiv:2008.01062 (2020)
16. Wang, T., Dong, H., Lesser, V., Zhang, C.: Roma: Multi-agent reinforcement learning with emergent roles. arXiv preprint arXiv:2003.08039 (2020)

17. Wang, Y., Han, B., Wang, T., Dong, H., Zhang, C.: Dop: Off-policy multi-agent decomposed policy gradients. In: International Conference on Learning Representations (2020)
18. Wen, Y., Yang, Y., Luo, R., Wang, J., Pan, W.: Probabilistic recursive reasoning for multi-agent reinforcement learning. arXiv preprint arXiv:1901.09207 (2019)
19. Yang, Y., Hao, J., Liao, B., Shao, K., Chen, G., Liu, W., Tang, H.: Qatten: A general framework for cooperative multiagent reinforcement learning. arXiv preprint arXiv:2002.03939 (2020)