Multiagent Q-learning with Sub-Team Coordination*

Extended Abstract

Wenhan Huang[1], Kai Li[1], Kun Shao[2], Tianze Zhou[3], Jun Luo[2], Dongge Wang[4], Hangyu Mao[2], Jianye Hao[2], Jun Wang[5] and Xiaotie Deng[6†]
[1]Shanghai Jiao Tong University, [2]Huawei Noah's Ark Lab, [3]Beijing Institute of Technology, [4]EPFL, [5]University College London, [6]Peking University

ABSTRACT

For cooperative mutliagent reinforcement learning tasks, we propose a novel value factorization framework in the popular centralized training with decentralized execution paradigm, called *multiagent Q-learning with sub-team coordination* (QSCAN). This framework could flexibly exploit local coordination within sub-teams for effective factorization while honoring the individual-global-max (IGM) condition. QSCAN encompasses the full spectrum of sub-team coordination according to sub-team size, ranging from the monotonic value function class to the entire IGM function class, with familiar methods such as QMIX and QPLEX located at the respective extremes of the spectrum. Empirical results show that QSCAN's performance dominates state-of-the-art methods in predator-prey tasks and the Switch challenge in MA-Gym.

KEYWORDS

Cooperative multiagent reinforcement learning; Centralized training with decentralized execution; Multiagent Q-learning; Value factorization framework; Sub-team coordination

ACM Reference Format:
W. Huang et al. 2022. Multiagent Q-learning with Sub-Team Coordination: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

In large cooperative multiagent systems, agents must coordinate for success, but full coordination of all agents is not always required. Typically, the task can be divided into several sub-tasks, and each sub-task requires exactly a sub-team of agents [3]. The sub-team structure has been widely used for planning tasks in robotics and unmanned aerial vehicles [9]. However, exploiting sub-team structures for effective coordination in multiagent reinforcement learning (MARL) has not been sufficiently explored.

In this work, we explore value-based cooperative MARL in the centralized training with decentralized execution (CTDE) paradigm [5]. We propose a novel value factorization framework, called *multi-agent Q-learning with Sub-team CoordinAtioN* (QSCAN), to strike a trade-off between the representational capability and the complexity of the network architecture. QSCAN flexibly handles sub-team formation and guarantees the individual-global-max (IGM) [8] condition. The IGM condition is important for value-based methods, whereby the individuals' greedy action corresponds to the optimal joint action for the team. We establish a coordination hierarchy based on QSCAN where the monotonic and the IGM function classes are located at the respective extremes. Two specific implementations, `QPAIR` and `QSCAN`, are proposed for our framework. To characterize coordination among agents, `QPAIR` simply enumerates all agent pairs, while `QSCAN` employs self-attention mechanisms. We empirically evaluate these implementations in predator-prey tasks [6] and the Switch challenge [4]. Comparing our implementations with QMIX [7] and QPLEX [11], we illustrate that the sub-team coordination pattern improves the results in these tasks. `QSCAN` significantly outperforms the two baselines in these settings, while `QPAIR` achieves comparable performance. These results show that our method can provide significant benefits to the CTDE paradigm. Our work suggests a way forward for more flexible sub-team coordination and learning in multiple settings beyond the CTDE paradigm.

*This work is partially supported by Science and Technology Innovation 2030 - "New Generation Artificial Intelligence" Major Project (No. 2018AAA0100901). It was done when WH and KL were interns at Huawei Noah's Ark Lab.
†Correspondence to xiaotie@pku.edu.cn (Xiaotie Deng)

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

2 MULTIAGENT Q-LEARNING WITH SUB-TEAM COORDINATION

In this section, we present the architecture of QSCAN. We would first describe our base architecture. We then discuss QSCAN's coordination module in detail and propose a general representation of the relation among sub-team coordination classes, called *coordination hierarchy*. Finally, we present two different implementations based on this hierarchy, `QPAIR` and `QSCAN`.

Base architecture. We employ the duplex dueling structure [11] to achieve a high representational capability. This structure factorizes each agent i's action-value function Q_i into its individual value and an advantage function as $V_i(\boldsymbol{\tau}) = \max_{a_i'} Q_i(\boldsymbol{\tau}, a_i')$, $A_i(\boldsymbol{\tau}, a_i) = Q_i(\boldsymbol{\tau}, a_i) - V_i(\boldsymbol{\tau})$. Meanwhile, the global action-value function $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a})$ is factorized into the global value function $V_{tot}(\boldsymbol{\tau})$ and the global advantage function $A_{tot}(\boldsymbol{\tau}, \boldsymbol{a})$, i.e. $Q_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) = V_{tot}(\boldsymbol{\tau}) + A_{tot}(\boldsymbol{\tau}, \boldsymbol{a})$. For a task with n agents, $V_{tot}(\boldsymbol{\tau}) = \sum_{i=1}^{n} V_i(\tau)$ and $A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) \approx \sum_{i=1}^{n} \lambda_i(\boldsymbol{\tau}, \boldsymbol{a}) A_i(\boldsymbol{\tau}, a_i)$, where λ_i is a non-negative importance weight for agent i. We extend the approximation of A_{tot} with monotonic functions that $A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) \approx f([\lambda_i(\boldsymbol{\tau}, \boldsymbol{a}) A_i(\boldsymbol{\tau}, a_i)]_{i=1}^{n}, s)$, where $f(x_1, \ldots, x_n, s)$ is a monotonic function with respect to each $x_i \leq 0$ (i.e., $\frac{\partial f}{\partial x_i} \geq 0$), and maintains a maximum value 0 ($f(\mathbf{0}_n, s) = 0$). Following previous work [2, 7, 11], we would use the global state s as the centralized information, if applicable, or the joint history $\boldsymbol{\tau}$. Since the base architecture adopts only monotonic functions, it guarantees the IGM condition.

Coordination hierarchy. In cooperative MARL tasks, agents could participate in different sub-teams for different sub-objectives.

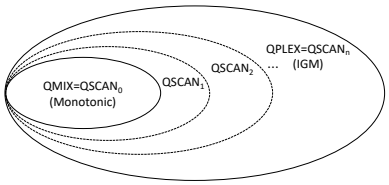
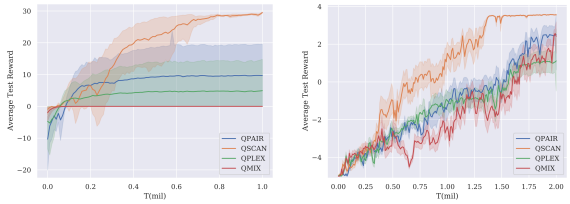


Figure 1: Coordination hierarchy: From monotonic function class to the entire IGM function class.



(a) Predator-Prey (6 versus 6). **(b) Switch4 challenge.**

Figure 2: Learning curves of QPAIR, QSCAN, QPLEX, and QMIX in two different tasks. The average reward with 95% confidence intervals is shown.

Intuitively, the team reward can be credited to each sub-team and then to each individual. Specifically, consider a sub-team $C \subseteq \mathcal{N}$ containing k agents, where $\mathcal{N} = \{1, \ldots, n\}$ is the agent set of the task. Let $\boldsymbol{a}_C$ denote the joint action of C. The global credit could be assigned to each agent as follows:

$$A_{tot}(\boldsymbol{\tau}, \boldsymbol{a}) \approx \sum_{C:C\subseteq\mathcal{N},|C|=k} \left[\sum_{i\in C} \left(g_i(\boldsymbol{\tau}, \boldsymbol{a}_C) \cdot A_i(\boldsymbol{\tau}, a_i)\right)\right] = \sum_{i=1}^{n} \left(\sum_{C:i\in C\subseteq\mathcal{N},|C|=k} g_i(\boldsymbol{\tau}, \boldsymbol{a}_C)\right) A_i(\boldsymbol{\tau}, a_i).$$

Since each agent i's preference over actions is characterized by $A_i(\boldsymbol{\tau}, a_i)$, we employ a function $g_i(\boldsymbol{\tau}, \boldsymbol{a}_C)$ as the importance weight of A_i and use $g_i(\boldsymbol{\tau}, \boldsymbol{a}_C) \cdot A_i(\boldsymbol{\tau}, a_i)$ to evaluate i's contribution to the sub-team C. The credit of the whole sub-team is thus a simple sum of all $g_i \cdot A_i$. Based on this insight, we propose QSCAN$_k$.

Definition 1 (QSCAN$_k$). *QSCAN$_k$ is a branch of QSCAN which concerns coordination within sub-teams containing only k agents. Specifically, QSCAN$_k$ adopts the following weights yielding from the sub-team coordination module $\lambda_i(\boldsymbol{\tau}, \boldsymbol{a}) = h(\sum_{C:i\in C\subseteq\mathcal{N},|C|=k} g_i(\boldsymbol{\tau}, \boldsymbol{a}_C))$, where h is a non-negative activation function.*

$\lambda_i(\boldsymbol{\tau}, \boldsymbol{a})$ corresponds to the total contribution of agent i to all sub-teams containing k members. Since the advantage as the disparity from the optimal action should keep non-positive, we use function h to ensure the positivity of each λ.

Since QSCAN$_k$ characterizes coordination within k-agent sub-teams, we could establish a hierarchy over QSCAN$_k$ as k varies from 0 to n. Specially, QSCAN$_0$ treats the coeficient λ_i as a fixed positive constant depending only on the joint history $\boldsymbol{\tau}$ and not taking any action as input, i.e. $\lambda_i(\boldsymbol{\tau}, \cdot) \equiv \lambda_i(\boldsymbol{\tau})$. Figure 1 shows a Venn graph of our coordination hierarchy. We can see that the classic methods, QMIX and QPLEX, are typical examples in our hierarchy. In the following, we propose two specific implementations of QSCAN.

Pairwise coordination. QPAIR employs a pairwise coordination module by enumerating all sub-teams with size 2. This module uses a multi-layer perceptron (MLP) to calculate the pairwise coordination coefficients g_i. This structure is an instance of QSCAN$_2$. The coefficient of each $A_i(\tau_i, a_i)$ is based on pairs of agents' actions as well as the global state (s, i, a_i, j, a_j). In this module, $g_i(\boldsymbol{\tau}, \boldsymbol{a}_{\{i,j\}}) = \text{MLP}(s, i, a_i, j, a_j)$ and $\lambda_i = h\left(\sum_{j=1}^{n} g_i(\boldsymbol{\tau}, \boldsymbol{a}_{\{i,j\}})\right)$. In practice, we use the absolute value function for h.

Self-attention. Enumerating all sub-teams will be computationally expensive in large multiagent systems. Inspired by the architecture of Transformer [10], we propose QSCAN which employs self-attention module to characterize coordination among agents hierarchically. The module takes all agent-action pairs and the centralized information as input and directly produces a series of non-negative weights $[\lambda_i]_{i=1}^n$. Specifically, it first receives all agent-action pairs $[(i, a_i)]_{i=1}^n$ and injects the global state s into them through a feedforward network to obtain a series of embedding vectors. These vectors will then be fed into m attention layers to output the weights λ. In each attention layer, we adopt residual learning to provide agents more flexibility to learn different forms of coordination. Since each attention layer aggregates the pairwise information of current inputs, we can expect this module with m attention layers to characterize the coordination in sub-teams with at most size 2^m.

3 EMPIRICAL EVALUATION

We use QMIX and QPLEX as baselines in this work. In experiments, QSCAN uses one self-attention layer. We evaluate QPAIR and QSCAN in predator-prey tasks [1] and the Switch challenge [4]. These two tasks require various styles of coordination to achieve high rewards.

The predator-prey tasks are complicated scenarios with immediate coordination rewards. In these tasks, the agents need to learn spatial-temporal local coordination. We evaluate algorithms in the scenario with 6 predators against 6 prey. The results are shown in Figure 2a. QMIX fails to learn a positive reward due to the relative overgeneralization [1] caused by the miscoordination of "capture" actions. QPLEX could fail to learn the "capture" action, suggesting the difficulty in representing precise coordination patterns. For QPAIR, it performs better than QPLEX because the coordination of "capture" actions only needs pairwise coordination, which QPAIR is forced to learn. For QSCAN, it outperforms all these approaches due to its adaptive balance of the pairwise coordination and each individual's local information. The Switch challenge is a more complicated coordination task due to the sparse and long-term rewards. It is a partially observable task that 4 agents need to reach their corresponding home by passing through the one-agent wide narrow corridor. In this task, the sub-team coordination in the very beginning steps influences deeply over the final rewards. As Figure 2b shows, QSCAN outperforms others in this task while QPLEX performs worst. QMIX and QPAIR achieve comparable results after training but QPAIR achieves better performance during the training phase. Overall, the results indicate that the sub-team coordination benefits in various coordination tasks.

REFERENCES

[1] Wendelin Böhmer, Vitaly Kurin, and Shimon Whiteson. 2020. Deep coordination graphs. In *International Conference on Machine Learning*. PMLR, 980–991.

[2] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[3] Bryan Horling and Victor Lesser. 2004. A survey of multi-agent organizational paradigms. *Knowledge Engineering Review* 19, 4 (2004), 281–316.

[4] Anurag Koul. 2019. ma-gym: Collection of multi-agent environments based on OpenAI gym. https://github.com/koulanurag/ma-gym.

[5] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (2008), 289–353.

[6] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. 2020. Weighted QMIX: Expanding Monotonic Value Function Factorisation. *arXiv preprint arXiv:2006.10800* (2020).

[7] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485* (2018).

[8] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. 2019. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:1905.05408* (2019).

[9] Milind Tambe. 1997. Towards flexible teamwork. *Journal of artificial intelligence research* 7 (1997), 83–124.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[11] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2020. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062* (2020).