Dynamic Conservative Degree Allocation for Offline Multi-Agent Reinforcement Learning

Extended Abstract

Haosheng Chen
East China Normal University
Shanghai, China
hschen@stu.ecnu.edu.cn

Yun Hua†
East China Normal University
Shanghai, China
yunhua@stu.ecnu.edu.cn

Junjie Sheng
East China Normal University
Shanghai, China
jarvis@stu.ecnu.edu.cn

Wenhao Li
Tongji University
Shanghai, China
whli@tongji.edu.cn

Bo Jin
Tongji University
Shanghai, China
bjin@tongji.edu.cn

Xiangfeng Wang
East China Normal University &
Shanghai Formal-Tech Information
Technology Co., Lt
Shanghai, China
xfwang@sei.ecnu.edu.cn

ABSTRACT

Offline Multi-agent Reinforcement Learning (MARL) has been designed to learn policies from pre-collected datasets without real-time interaction in multi-agent systems. A primary concern in offline MARL is the conservative degree allocation, which involves assigning different conservatism levels to agents based on their varying influence on the system. Current approaches frequently neglect this crucial aspect, resulting in suboptimal performance, particularly when agents have differing impacts on the environment. In this paper, we propose OMCDA, a novel offline MARL algorithm that addresses the issue of conservative degree allocation by assigning dynamic conservatism levels to each agent based on their individual influence on system performance. OMCDA decomposes the Q-function into two components: one for computing the return and another for capturing deviations from the behavior policy. Additionally, OMCDA employs a dynamic allocation mechanism that adjusts conservatism levels for agents based on varying impacts, while maintaining coherent credit assignment and ensuring robust system performance throughout learning. We evaluate OMCDA on MuJoCo and SMAC, showing it outperforms existing offline MARL methods in challenging tasks by effectively addressing conservative degree allocation.

KEYWORDS

Multi-agent reinforcement learning; Offline reinforcement learning

ACM Reference Format:
Haosheng Chen, Yun Hua†, Junjie Sheng, Wenhao Li, Bo Jin, and Xiangfeng Wang. 2025. Dynamic Conservative Degree Allocation for Offline Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

† Corresponding author (yunhua@stu.ecnu.edu.cn).

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has garnered significant attention in domains like autonomous driving [2], collaborative robotics[12], and multi-player games[1], where agents must cooperate or compete to achieve specific goals. However, the majority of these approaches presume that agents are able to interact with the environment during training, which poses challenges when direct interaction is costly, hazardous, or otherwise impractical[16]. To address these limitations, offline reinforcement learning has been proposed[5, 7, 9, 10, 18].

Offline RL trains policies using pre-collected datasets, with the main challenge being sensitivity to the training data distribution [8], especially during bootstrapping. Evaluating Q-functions on out-of-distribution actions introduces errors that cannot be corrected without new data, leading to instability. Conservative strategies [9] are often used to limit policy updates and prevent significant deviation from the dataset behavior [18]. While single-agent offline RL has seen progress, applying these methods to multi-agent systems presents additional challenges, as agent interactions increase non-stationarity and uncertainty, complicating credit assignment and stable policy learning.

Recent works address multi-agent challenges by proposing offline MARL [17][19] using the Centralized Training with Decentralized Execution (CTDE) framework [11]. They apply global-to-local value regularization to bridge multi-agent value decomposition with offline learning and maintain credit assignment consistency. However, these methods overlook the issue of conservative degree allocation, applying the same level of conservatism to all agents, regardless of their varying impacts on the system. Agents in multi-agent settings often have unequal influence on outcomes due to their roles [15] or environmental interactions [4]. In offline MARL, the degree of deviation from behavior policies should depend on each agent's influence, as a uniform strategy may overly constrain some agents while allowing others to become too aggressive, risking policy failure.

In this paper, we introduce a novel Offline MARL approach, **Offline MARL with Conservative Degree Allocation (OMCDA)**, which addresses the challenge of distributing conservatism

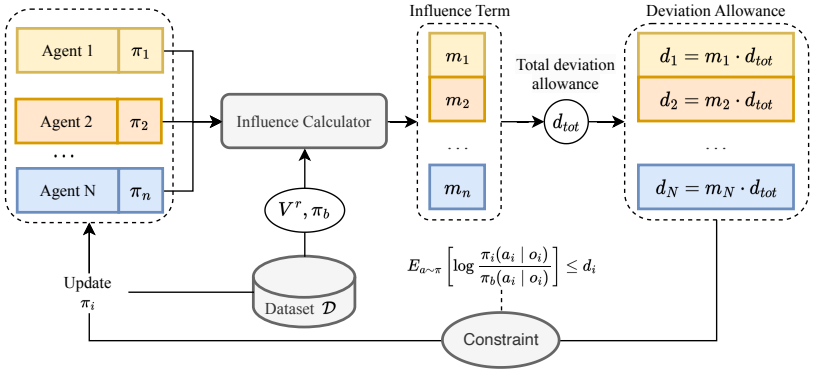

Figure 1: The overview of OMCDA

Table 1: MUJOCO and SMAC Tasks

Task	Dataset	BCQ-MA	CQL-MA	ICQ	OMAR	OMIGA	OMCDA
Hopper	expert	77.85±58.04	159.14±313.83	754.74±806.28	2.36±1.46	859.63±709.47	**1214.25±830.72**
Hopper	medium	44.58±20.62	401.27±199.88	501.79±14.03	21.34±24.90	1189.26±544.30	**1441.53±488.91**
Hopper	medium-replay	26.53±24.04	31.37±15.16	195.39±103.61	3.30±3.22	774.18±494.27	**1733.27±379.71**
Hopper	medium-expert	54.31±23.66	64.82±123.31	355.44±373.86	1.44±0.86	709.00±595.66	**1047.13±523.67**
5m_vs_6m	good	7.76±0.15	8.08±0.21	7.87±0.30	7.40±0.63	8.25±0.37	**8.46±0.12**
5m_vs_6m	medium	7.58±0.10	7.78±0.10	7.77±0.30	7.08±0.51	7.92±0.57	**8.10±0.18**
5m_vs_6m	poor	7.61±0.36	7.43±0.10	7.26±0.19	7.27±0.42	7.52±0.21	**7.67±0.10**

among agents based on their deviations from behavior policies and their impact on system performance. OMCDA decomposes the Q-function in offline MARL with regulation into two components for both return and deviation. The conservative degree of each agent is dynamically adjusted based on the effect of their deviations on the overall return. This dynamic allocation is integrated into the OMCDA framework, ensuring a balance between conservatism and flexibility, and consistent credit assignment during training.

The key contributions of this paper are as follows: 1) A comprehensive analysis of conservative degree allocation in offline MARL, exploring how varying conservative degrees affect individual agent returns and overall system performance. 2) The introduction of OMCDA, a novel offline MARL algorithm that dynamically adjusts each agent's conservative degree based on its impact on system performance, balancing conservatism and flexibility while ensuring consistent credit assignment. 3) Extensive experiments on diverse datasets, including multi-agent MuJoCo [3] and the StarCraft Multi-Agent Challenge (SMAC) [14], showing that OMCDA consistently outperforms existing methods.

2 OMCDA FRAMEWORK

In this section, we present the overview of the dynamic conservatism degree allocation framework in OMCDA, as is in Figure. 1. The influence calculator first takes the policy from each agent, along with the return-based state-value function V^r and the behavior policy π_b derived from the data in dataset, as input to generate the influence term for each agent on the system. Then each agent's deviation allowance is then allocated from the total deviation allowance based on the influence term. Finally, the deviation allowance is incorporated into the constraint to update the conservatism level, which controls the range of policy updates for each agent, while achieving dynamic conservative degree allocation and ensuring consistent credit assignment.

3 EXPERIMENT

In this section, we evaluate OMCDA using Multi-Agent MuJoCo [3] and the StarCraft Multi-Agent Challenge (SMAC) [14]. We compare our approach with five recent offline MARL algorithms: the multi-agent versions of BCQ [6] and CQL [7] (BCQ-MA and CQL-MA), as well as ICQ [19], OMAR [13], and OMIGA [16]. Representative results are presented in Table 1.

4 CONCLUSION

In conclusion, a novel offline MARL framework OMCDA is introduced to tackle the challenge of conservative degree allocation. OMCDA decomposes the Q-function in offline MARL with regulation into two components: one for computing the return and another for capturing deviations from the behavior policy. It dynamically adjusts each agent's conservative degree based on their influence on the overall system's performance, ensuring coherent credit assignment and robust performance throughout the learning process. Meanwhile, extensive experiments demonstrate that OMCDA consistently outperforms existing offline MARL methods across various environments.

ACKNOWLEDGMENTS

This work was supported STCSM (No.22QB1402100).

REFERENCES

[1] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

[2] Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. 2012. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics* 9, 1 (2012), 427–438.

[3] Christian Schroeder de Witt, Bei Peng, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. 2020. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709* 19 (2020).

[4] 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.

[5] Scott Fujimoto and Shixiang Shane Gu. 2021. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems* 34 (2021), 20132–20145.

[6] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*. PMLR, 2052–2062.

[7] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2021. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169* (2021).

[8] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems* 32 (2019).

[9] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 1179–1191.

[10] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* (2020).

[11] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).

[12] James Orr and Ayan Dutta. 2023. Multi-agent deep reinforcement learning for multi-robot applications: A survey. *Sensors* 23, 7 (2023), 3625.

[13] Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. 2022. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification. In *International conference on machine learning*. PMLR, 17221–17237.

[14] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043* (2019).

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

[16] Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. 2024. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. *Advances in Neural Information Processing Systems* 36 (2024).

[17] Xiangsen Wang and Xianyuan Zhan. 2023. Offline multi-agent reinforcement learning with coupled value factorization. *arXiv preprint arXiv:2306.08900* (2023).

[18] Yifan Wu, George Tucker, and Ofir Nachum. 2019. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361* (2019).

[19] Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. 2021. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 10299–10312.