SCIENCE CHINA
Information Sciences

• LETTER •

December 2025, Vol. 68, Iss. 12, 229205:1–229205:2
https://doi.org/10.1007/s11432-024-4597-4

Multiagent reinforcement learning with quantified information-decision content measurement

Ershen WANG[1,2], Xiaotong WU[1], Chen HONG[3,4*], Yanwen WANG[5], Aidong CHEN[3,4], Hongyuan JING[3,4], Song XU[1] & Pingping QU[1]

[1]*School of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, China*
[2]*School of Civil and Aviation, Shenyang Aerospace University, Shenyang 110136, China*
[3]*Multi-Agent Systems Research Centre, Beijing Union University, Beijing 100101, China*
[4]*College of Robotics, Beijing Union University, Beijing 100101, China*
[5]*School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China*

Received 5 August 2024/Revised 3 April 2025/Accepted 25 August 2025/Published online 3 November 2025

Citation Wang E S, Wu X T, Hong C, et al. Multiagent reinforcement learning with quantified information-decision content measurement. Sci China Inf Sci, 2025, 68(12): 229205, https://doi.org/10.1007/s11432-024-4597-4

In large-scale multiagent systems (MASs), partial observability poses a notable challenge. When agents process extensive local information, global information scarcity, and high computation complexity arise [1]. To address the partial observability, researchers have adopted the partially observable Markov decision process (POMDP) as a mathematical framework to simulate and analyze agents' decision-making process.

To address partial observability in multiagent reinforcement learning (MARL), we employ the recursive structure of recurrent neural networks (RNNs) to process sequential information in a decentralized partially observable Markov decision process (Dec-POMDP) [2, 3]. We propose an MARL model with a recurrent Dec-POMDP, in which RNN-generated hidden states replace real states, and prove that the hidden states effectively support rational decision-making under partial observability. The contributions of this study are as follows.

• We propose an MARL model with a recurrent decentralized partially observable Markov decision process (RDec-POMDP) and demonstrate that agents make better decisions using hidden states than real states.

• Within the RDec-POMDP framework, we propose an MARL method with quantified information-decision content measurement (QICM), integrating prioritized experience replay (PER) and temporal-difference learning.

Problem formulation. It is known that RNNs are well-suited for partially observable problems [4]. Inspired by this, we model fully cooperative MARL tasks as RDec-POMDP, leveraging the recursive nature of RNN to obtain hidden states.

RDec-POMDP is represented by a tuple $G = \langle \mathcal{S}, \mathcal{U}, P, \mathcal{R}, \mathcal{X}, \mathcal{Z}, \mathcal{O}, n, \gamma, \mathcal{H} \rangle$. The detailed definitions and explanations of each component are provided in Appendix B.

We outline four main properties of RDec-POMDP. (1) At any time step t, the true state is a function of the true state at time step $t-1$, and new information is updated at time step t. The hidden state is updated recursively, with the history information of the hidden state updated and accumulated by RNNs. (2) The transition probability P to the next real state depends only on the current real state and action, independent of the hidden state. (3) When the information distribution $\mathcal{X}$ is known, the observation o is conditionally independent of action u, such as $p(o \mid x, u) = p(o \mid x)$, with the relationship expressed as $s \to x \to o$. (4) Given the statistic $f(h)$ and action u, the joint probability of reward r and observation o is represented by the information variable x, demonstrating how $f(h)$ effectively captures information from the hidden state h.

RDec-POMDP maximizes the expected return and identifies the optimal policy. The policy is defined as a mapping from histories to probability measures over the action space, defined as $\mathcal{H} \to \Delta(\mathcal{U})$. Given the cyclic recursive nature of the hidden state and the predictability of the next state, the following conclusion can be drawn:

$$\begin{aligned} &f(h') = \theta(f(h), u, o'), \forall h' = (h, u, o'), \\ &p(r, o' \mid h, u) = p(r, o' \mid f(h), u), \forall (h, u, r, o'), \end{aligned} \tag{1}$$

where f represents the statistic, θ is the hidden state update function of RNN. As long as Eq. (1) is satisfied, the statistic f is optimal, indicating that $f(h)$ captures all important information from the hidden state h. Thus, hidden states in the RDec-POMDP framework contain richer information than real states, enabling agents to make optimal decisions. A detailed proof is provided in Appendix C.

Proposed solution. The core principle of QICM is to use environmental information entropy and significantly enhance the decision-making capabilities of agents. The overall QICM architecture is shown in Figure 1(a).

The design of agent networks is crucial for approximating the Q-value of each agent. In our approach, agents utilize

* Corresponding author (email: hchchina@sina.com, xxthongchen@buu.edu.cn)

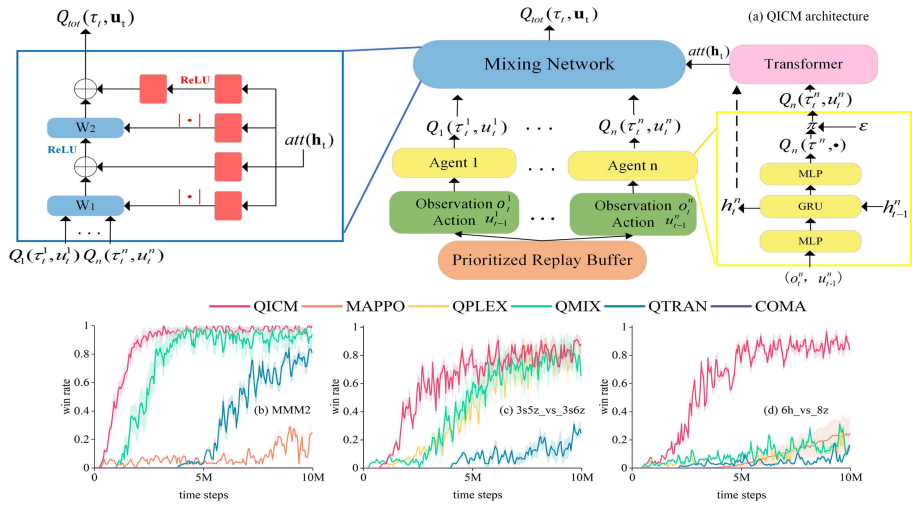

Figure 1 (Color online) (a) QICM framework comprises agent networks (on the right side, yellow), a transformer (on the right side, pink), a mixing network (on the left side), and a PER buffer (at the bottom, in orange). (b)–(d) Win rates for QICM against baseline algorithms on three super-hard maps.

both current partial observations and historical information, integrated via deep recurrent Q-learning (DRQN). Gated recurrent units (GRUs) were used to generate hidden states h_t, which are vital for maintaining the coherence of sequential information. Subsequently, the hidden states h_t were input into the transformer.

The transformer outputs a sequence of hidden states using a multi-head attention mechanism. To obtain the local Q value, each hidden state was transformed using a feed-forward neural network. The local Q-values were fed into the mixing network to generate the global Q value while adhering to the monotonicity constraint. This process ensures that the features extracted by the transformer are effectively integrated through the mixing network while preserving the individual global max (IGM) condition.

The mixing network optimizes the overall collaborative behaviors using a linear combination. The local Q-value outputs from agent networks were combined to produce the total action value Q_{tot}, which reflects the optimal action selected by the agents. The mixing network embodies collective intelligence and efficiency, thereby providing a clear path toward shared goals.

However, the introduction of hidden states increases the uncertainty level. To address this issue, PER was employed to dynamically adjust the probability of experience sampling based on sample uncertainty. In addition, a parameterized TD(λ) was used to estimate the value function of the current state by considering reward signals from multiple time steps within a certain range.

These techniques reduce environmental instability and capture the impact of long-term rewards, thereby ensuring the performance and stability of the QICM. A detailed description of QICM is provided in Appendix D.

Simulation. StarCraft multiagent challenge (SMAC) is a widely used benchmark for evaluating MARL methods. Figures 1(b)–(d) show the win rates of QICM compared to baseline methods on three super-hard SMAC maps. QICM consistently outperforms other methods, demonstrating effective exploration. Ablation experiments confirm that QICM outperforms other baseline methods.

We introduced the friendly survival rate and enemy mortality rate metrics to further verify the effect of QICM. QICM consistently outperforms baseline methods in these metrics. Additional comparison experiments and ablation experiments are presented in Appendix E.

Conclusion. We addressed partial observability in multi-agent settings by introducing RDec-POMDP and proposed an MARL method, QICM, which uses information-rich hidden states to replace real states. QICM incorporates PER and TD(λ) to tackle increased uncertainty. In future work, QICM will be applied to unmanned aerial vehicle (UAV) swarm confrontation game scenes, enhancing its practical applicability and robustness across diverse real-world tasks.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2018AAA0100804), National Natural Science Foundation of China (Grant No. 62173237), Aeronautical Science Foundation of China (Grant No. 20240055054001), Open Fund of Key Laboratory of Technology and Equipment of Tianjin Urban Air Transportation System (Grant No. TJKL-UAM-202305), Joint Fund of Ministry of Natural Resources Key Laboratory of Spatiotemporal Perception and Intelligent Processing (Grant No. 232203), and Applied Basic Research Programs of Liaoning Province (Grant No. 2025JH2/101300011).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Wang Z, Mu C, Hu S, et al. Modeling the dynamics of regret minimization in large agent populations: a master equation approach. In: Proceedings of IJCAI, 2022. 534–540
2 Lambrechts G, Bolland A, Ernst D. Informed POMDP: leveraging additional information in model-based RL. ArXiv:2306.11488
3 Oliehoek F A, Amato C. A Concise Introduction to Decentralized POMDPs. Cham: Springer International Publishing, 2016
4 Phan T, Ritz F, Altmann P, et al. Attention-based recurrence for multiagent reinforcement learning under stochastic partial observability. In: Proceedings of International Conference on Machine Learning, 2023. 27840–27853