Episodic Multi-agent Reinforcement Learning with Curiosity-driven Exploration

Lulu Zheng*[1], Jiarui Chen*[2,3], Jianhao Wang[1], Jiamin He[4], Yujing Hu[3], Yingfeng Chen[3], Changjie Fan[3], Yang Gao[2], Chongjie Zhang[1]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University
[2] Department of Computer Science and Technology, Nanjing University
[3] Fuxi AI Lab, NetEase
[4] Department of Computing Science, University of Alberta

Abstract

Efficient exploration in deep cooperative multi-agent reinforcement learning (MARL) still remains challenging in complex coordination problems. In this paper, we introduce a novel Episodic Multi-agent reinforcement learning with Curiosity-driven exploration, called EMC. We use prediction errors of individual Q-values as intrinsic rewards for coordinated exploration and utilize episodic memory to exploit explored informative experience to boost policy training.

Contributions:

(i) We present a novel multi-agent curiosity-driven exploration framework which can be adopted in many value-based MARL algorithms.
(ii) We are the first to utilize the mechanism of episodic control in deep multi-agent reinforcement learning.
(iii) Our method achieves state-of-the-art on the challenging tasks in the StarCraft II micromanagement benchmark.

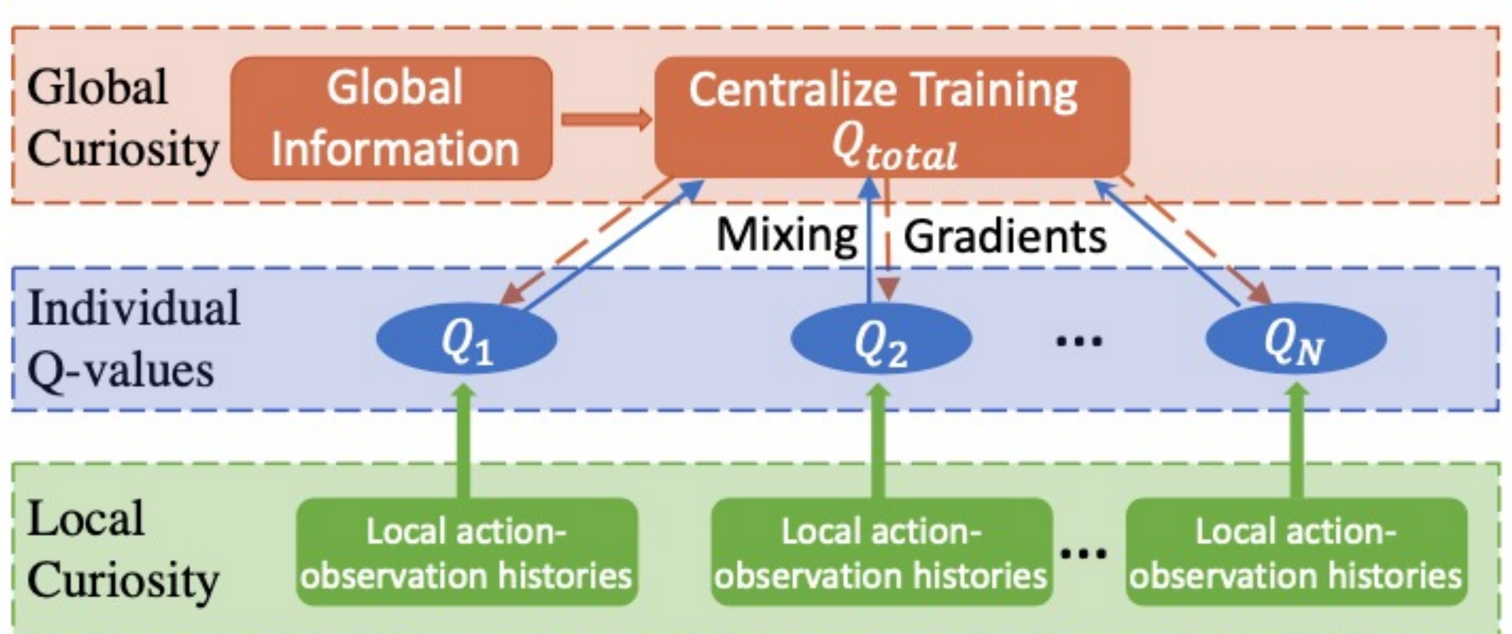
Motivation

Background:

Curiosity is a type of intrinsic motivation for exploration, which usually uses prediction errors on different spaces. However, due to the exponentially growing state space and partial observability in MARL, curiosity-driven exploration methods cannot be adopted into MARL directly.

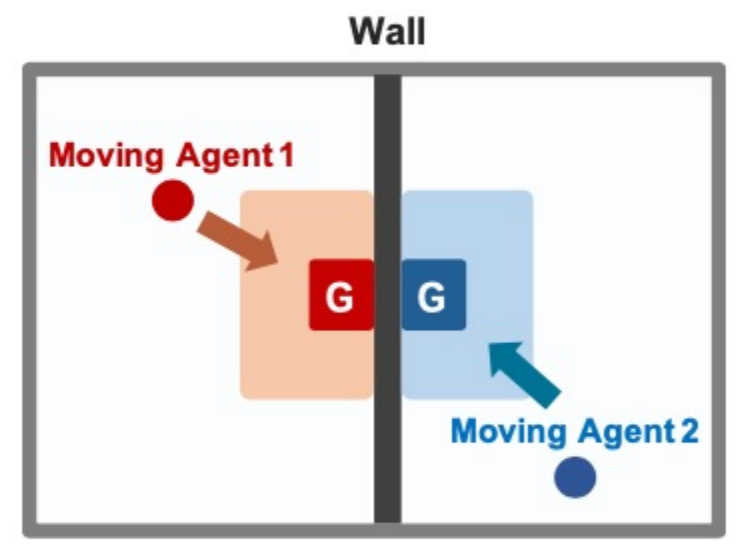
Problem: In which space to define curiosity in MARL?

Centralized (Global) Space: It is inefficient to find structured but sparse interactions between agents in the exponentially growing state space.
Decentralized (Local) Space: it will fail to guide agents to coordinate due to partial observability in the MARL setting.
Middle Point (Individual Q-values Space): (1) provides a novelty measure of joint observation histories with scalability; (2) captures the influence from other agents due to the implicit credit during centralized training.



Didactic Example

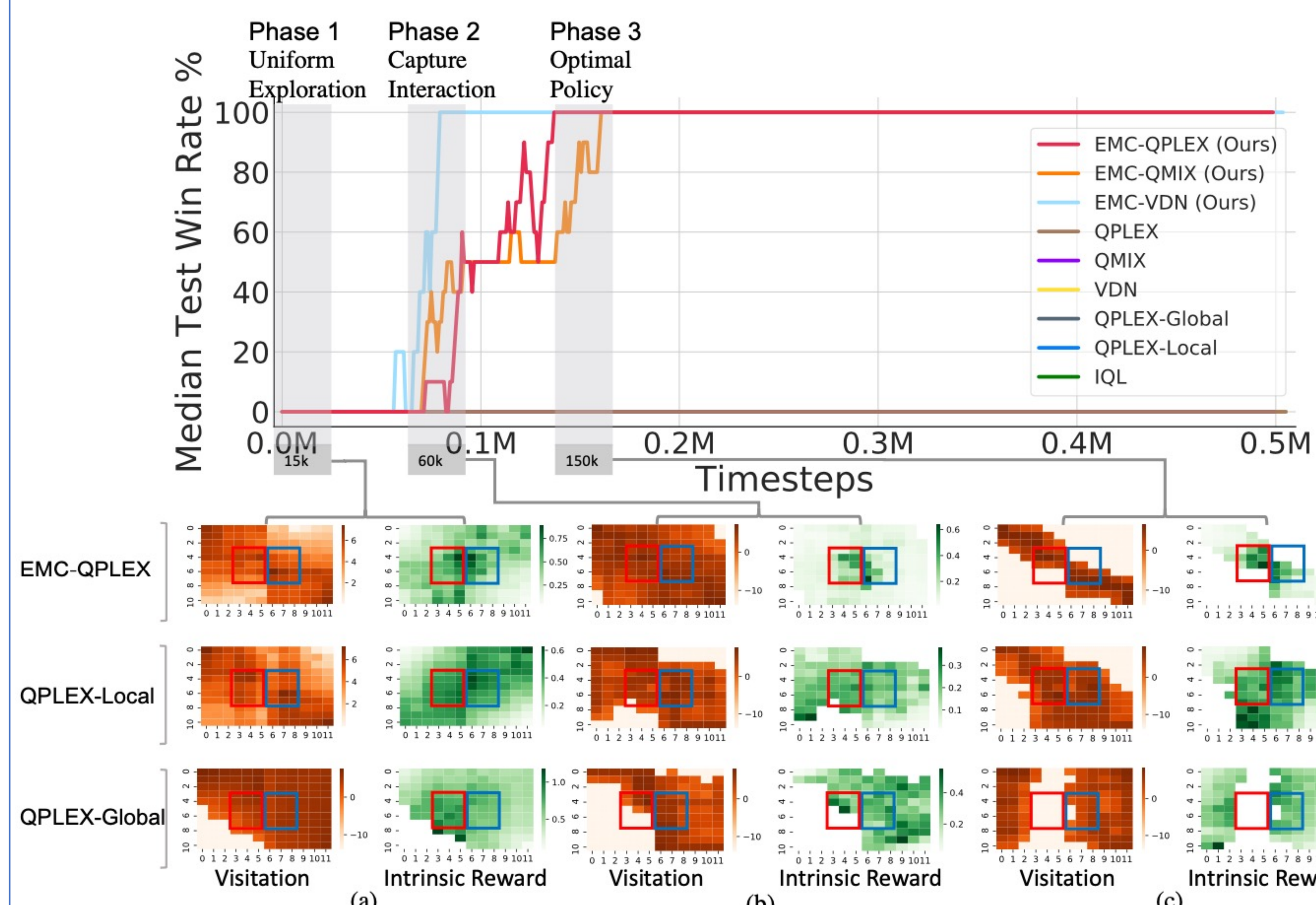
An illustrative gridworld game requiring coordinated exploration



Challenges:

(i) **Partial observability**: one agent cannot be observed by the other until it gets into the shaded area.
(ii) **Sparse Reward**: positive reward if and only if the two agents arrive at the goal grid at the same time. Otherwise, they will get incoordinate punishment.

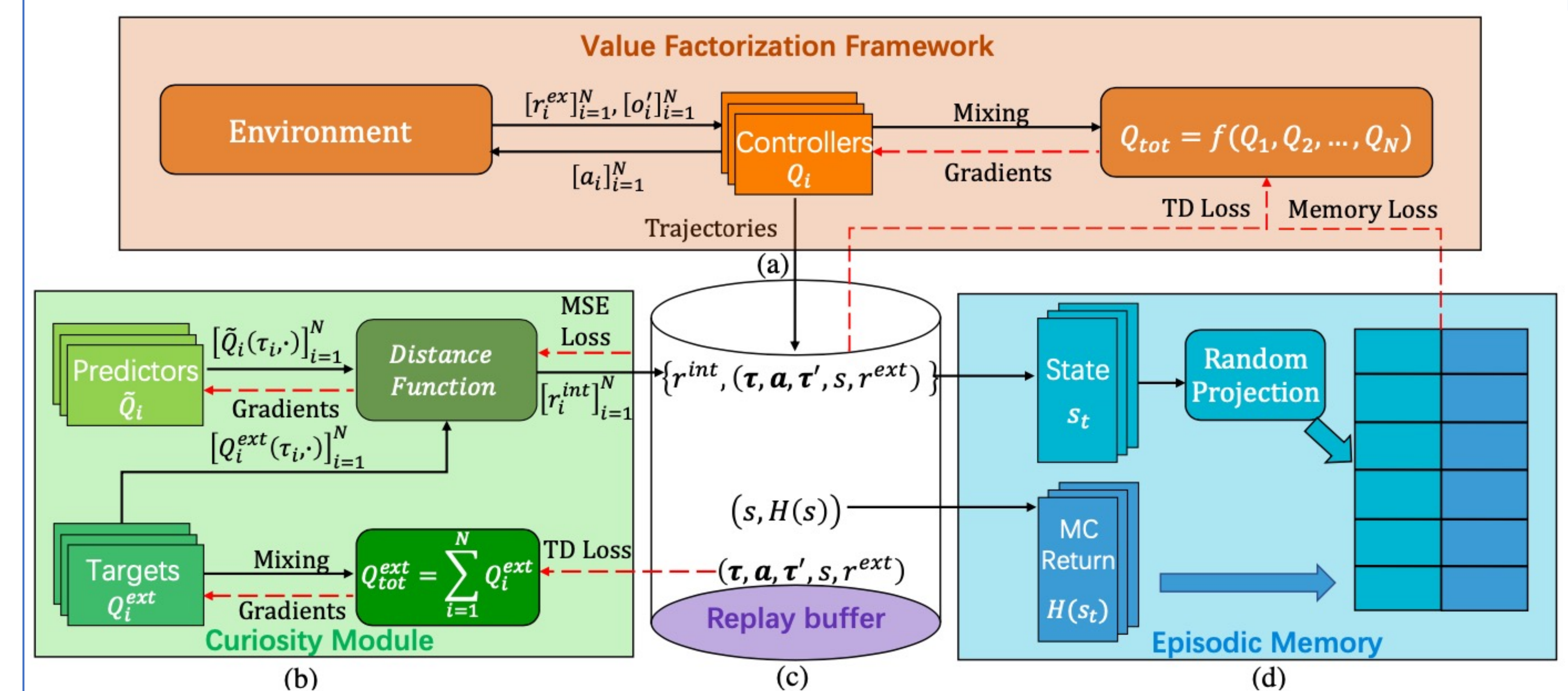
Visualization



Analysis:

Centralized (Global) Space: encourages agents to visit all configurations without bias which is inefficient and not scalable.
Decentralized (Local) Space: cannot encourage agents to coordinate due to the partial observability in decentralized execution.
Our method: can capture valuable and spare interactions among agents and bias exploration into new or promising states.
Results: Therefore, only our methods can win the game while other methods failed.

Framework



Curiosity Module: use the prediction error of local Q-values as intrinsic rewards

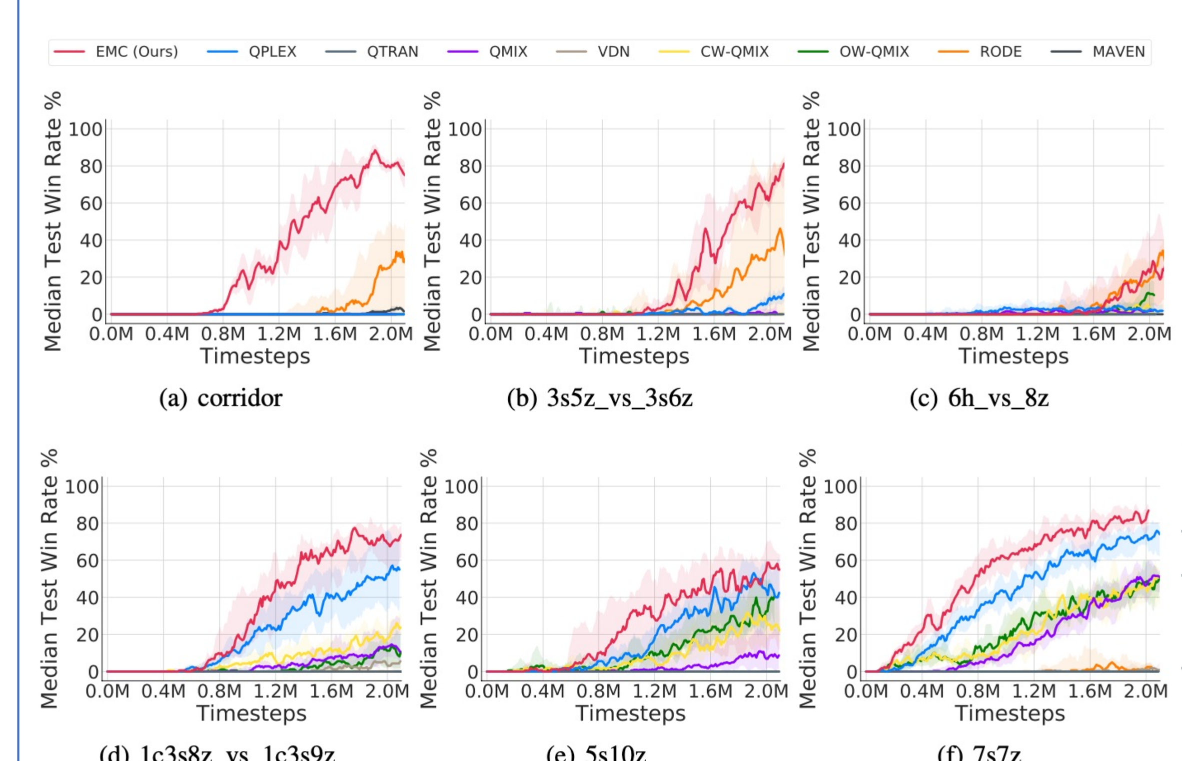
$$r^{int} = \frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{Q}_i(\tau_i,\cdot) - Q_i^{ext}(\tau_i,\cdot)\right\|_2$$

Episodic Memory: record the maximum remembered return of the current state

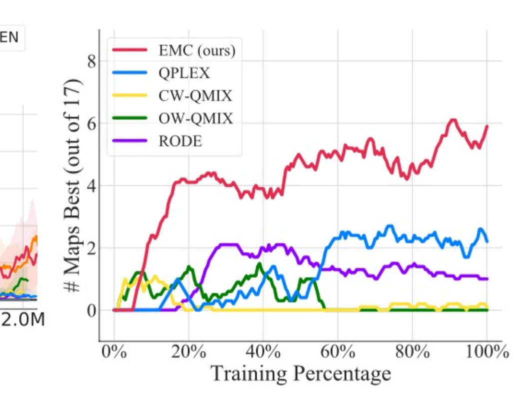
$$H(\phi(s_t)) = \begin{cases} \max\{H(\phi(\hat{s}_t)), R_t(s_t, \boldsymbol{a}_t)\} & \text{if } \|\phi(\hat{s}_t) - \phi(s_t)\|_2 < \delta \\ R_t(s_t, \boldsymbol{a}_t) & \text{otherwise} \end{cases}$$

Experiments

Results of super hard maps in SMAC:



Overall performance:



Our method is the best performer on up to 6 out of 17 tasks, underperforms on just three tasks, and ties for the best performer on the rest tasks.

Conclusion

This paper introduces EMC, a novel episodic multi-agent curiosity-driven exploration framework that allows for efficient coordinated exploration and boosted policy training by exploiting explored informative experiences. EMC achieves state-of-the-art on challenging tasks in the StarCraft II micromanagement benchmark.