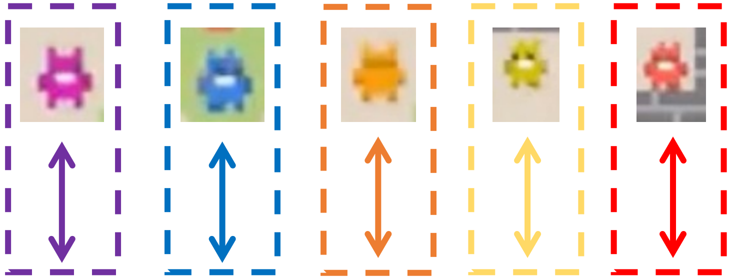
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

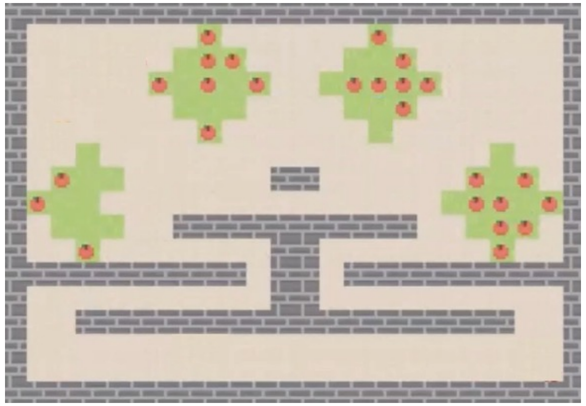
Towards Local Interaction & Global Coordination

Dan Qiao
danqiao@link.cuhk.edu.cn
Supervisor: Baoxiang Wang, Hongyuan Zha
SDS @ CUHKSZ
August 10, 2022

MARL Frameworks

Independent Learning





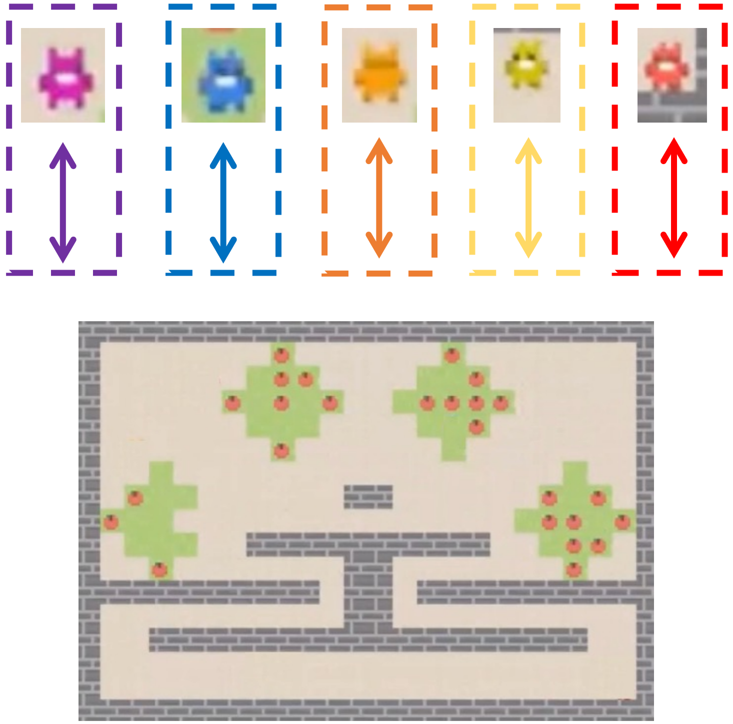
Training: $o_i, a_i \rightarrow \pi_i(a_i|o_i)$

Execution: $\pi_i(a_i|o_i)$

IQL suffers Nonstationary Issues

MARL Frameworks

Independent Learning

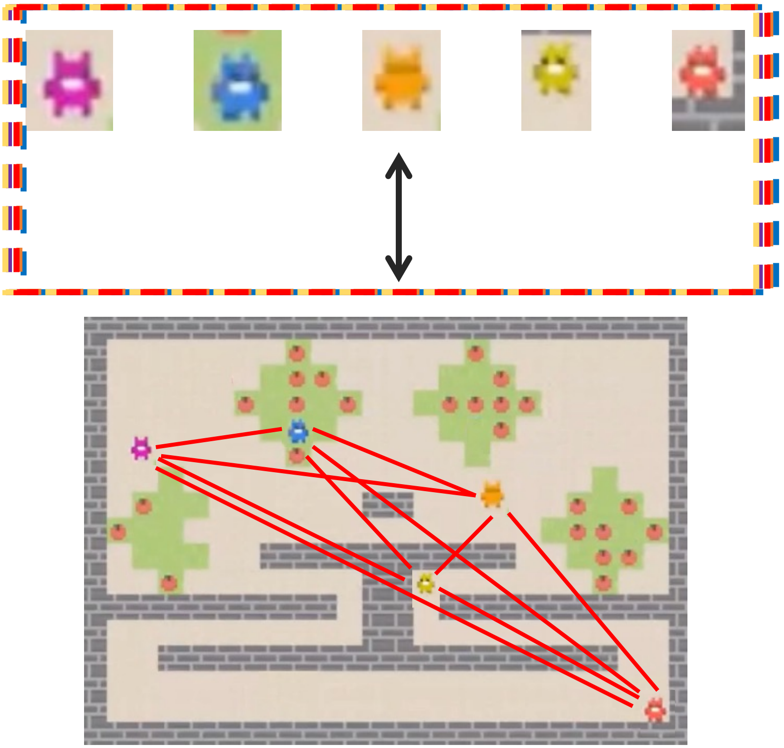


Training: $o_i, a_i \rightarrow \pi_i(a_i|o_i)$

Execution: $\pi_i(a_i|o_i)$

IQL suffers Nonstationary Issues

CTDE/Joint-action Learning



Training: $o_1, o_2, \dots, o_5, a_1, a_2, \dots, a_5 \rightarrow \pi_i(a_i|o_i)$

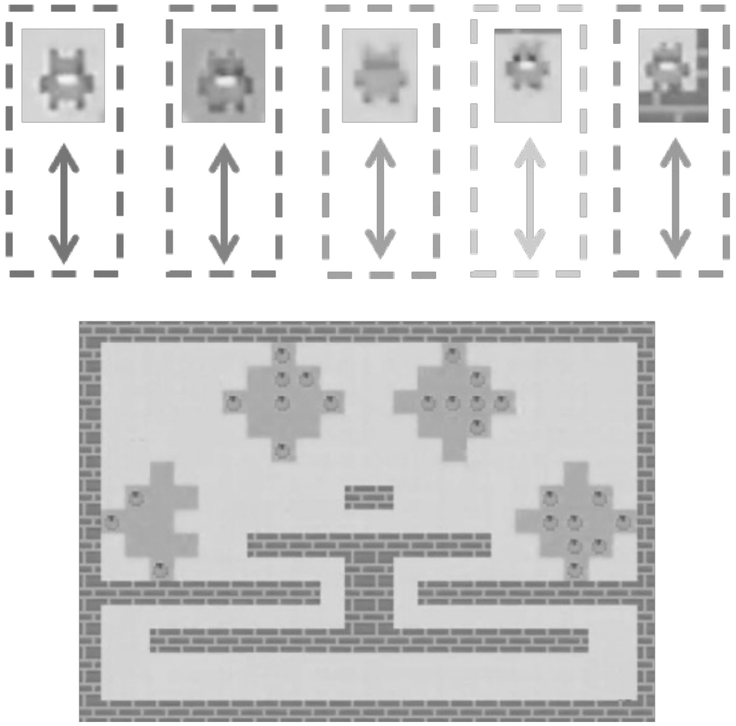
Execution: $\pi_i(a_i|o_i)$

CTDE suffers Overgeneralization Issues

"Employed value functions cannot estimate well because agents sometimes choose uncoordinated actions, and thus the optimal policy cannot be learned" -Yi et al., 2022

MARL Frameworks

Independent Learning

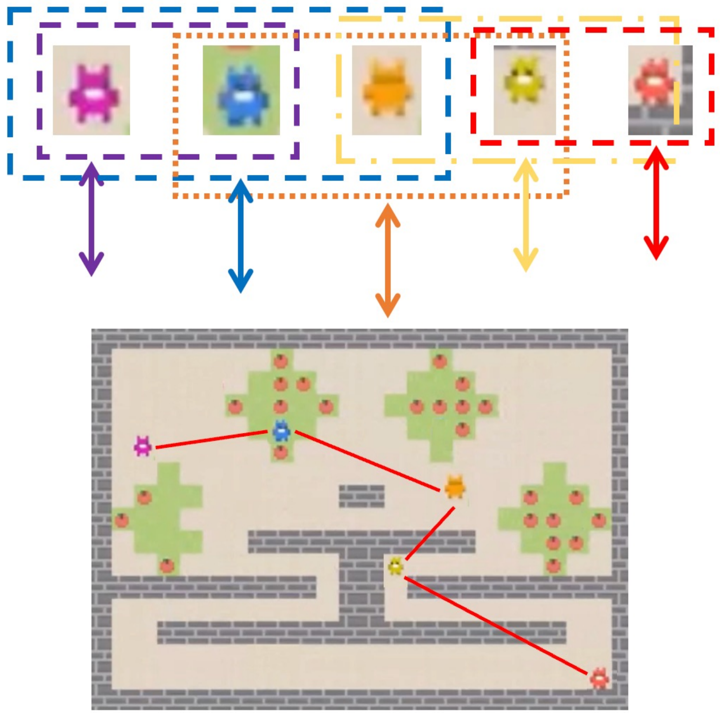


Training: $o_i, a_i \rightarrow \pi_i(a_i|o_i)$

Execution: $\pi_i(a_i|o_i)$

IQL suffers Nonstationary Issues

DTDE/Neighbor-action Learning

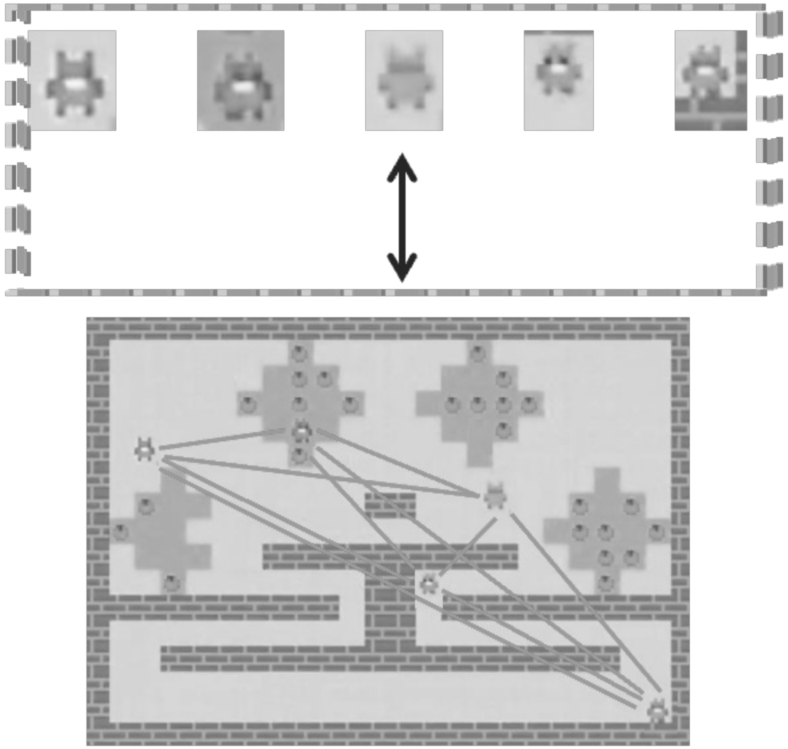


Training: $o_i, o_{j\in N_i}, a_i, a_{j\in N_i} \rightarrow \pi_i\left(a_i|o_i, o_{j\in N_i}\right)$

Execution: $\pi_i\left(a_i|o_i, o_{j\in N_i}\right)$

DTDE agents exchange local information to limited neighbors over communication topology G without a center, which leverages networks to enable distributed cooperation and less overgeneralization.

CTDE/Joint-action Learning



Training: $o_1, o_2, \ldots, o_5, a_1, a_2, \ldots, a_5 \rightarrow \pi_i(a_i|o_i)$

Execution: $\pi_i(a_i|o_i)$

CTDE suffers Overgeneralization Issues

"Employed value functions cannot estimate well because agents sometimes choose uncoordinated actions, and thus the optimal policy cannot be learned" -Yi et al., 2022

DTDE networked MARL

Policy Evaluation

- Multi-agent reinforcement learning via double averaging primal-dual optimization. Hoi-To Wai. 2018 NIPS
- Finite-Time Analysis of Distributed TD(0) with Linear Function Approximation for Multi-Agent Reinforcement Learning. Doan. 2019 ICML
- Multiagent fully decentralized value function learning with linear convergence rates. Ali Sayed. 2020 TAC
- Finite-Sample Analysis For Decentralized Batch Multi-Agent Reinforcement Learning With Networked Agents. Zhang Kaiqing. 2021 TAC
- Decentralized Deterministic Multi-Agent Reinforcement Learning. Grosnit. 2021 arxiv
- Finite time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward. Liu Jia. 2022 ICLR

Learning Optimal Policy

- A Distributed Actor-Critic Algorithm and Applications to Mobile Sensor Network Coordination Problems. Pennesi. 2010 TAC
- QD-Learning: A Collaborative Distributed Strategy for Multi-agent Reinforcement Learning Through Consensus_ Innovations. Kar. 2013 TSP
- Fully Decentralized Multi-agent Reinforcement Learning with Networked Agents. Zhang Kaiqing. 2018 ICML
- Diff-DAC: Distributed actor-critic for average multitask deep reinforcement learning. Sergio. 2018 ALA
- Value Propagation for Decentralized Networked Deep Multi-agent Reinforcement Learning. Qu Chao 2019 NIPS
- F2A2: Flexible fully-decentralized approximate actor-critic for cooperative multi-agent reinforcement learning. Li Wenhao. 2021 arxiv
- Learning to Share in Multi-Agent Reinforcement Learning. Yi. 2022 ICLR LToS

Communication Efficient

- Coordinating multi-agent reinforcement learning with limited communication. Zhang Chongjie. 2013 AAMAS
- Communication-efficient distributed reinforcement learning. Chen 2018 arxiv
- A communication efficient hierarchical distributed optimization algorithm for multi-agent reinforcement learning. Ren J. 2019 ICMLworkshop
- A communication-efficient multi-agent actor-critic algorithm for distributed reinforcement learning. Lin, Basar 2019 CDC
- Event-Triggered Multi-agent Reinforcement Learning with Communication under Limited-bandwidth Constraint. Zhao dongbin. 2021 TNNLS
- Communication-Efficient Policy Gradient Methods for Distributed Reinforcement Learning. Chen Tianyi. 2021 TCNS

Related Works

- Multi-agent reinforcement learning in time-varying networked systems. Lin Yiheng. 2020 arxiv
- Decentralized Multi-agent Reinforcement : Learning with Multi-time Scale of Decision Epochs. Jia Qingshan 2020
- Neighborhood Cognition Consistent Multi-Agent Reinforcement Learning. Mao Hangyu. 2020 AAAI
- Multi-agent Reinforcement Learning for Networked System Control. Chu Tianshu 2020 ICLR NeurComm
- Scalable Reinforcement Learning of Localized Policies for Multi-Agent Networked Systems. Qu Guannan. 2020 NIPS

Presented with XMind

- Consider a networked MARL system consisting of N-agents

$$\{\mathcal{S}, \{\mathcal{A}^i\}, P, \{r_i\}_{i\in\mathcal{V}}, \{\mathcal{G}_t\}_{t\geq 0}, \{\mathcal{M}_{ij}\}_{ij\in\mathcal{E}}\}$$

Following the standard setting in QD-Learning (Kar. 13), each agent updates the estimation of global Q-value with local information of neighbor agents. The information m is Q-value.

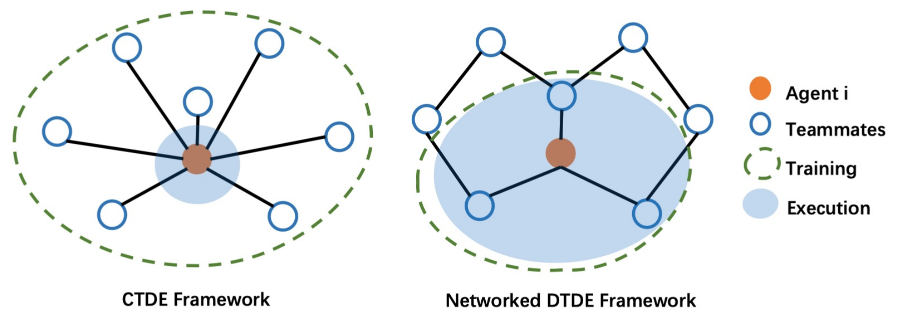


Fig. 1: In CTDE (left), agent i uses all agents' observation in centralized training and self observation in decentralized excution. In networked DTDE (right), agent i uses local neighbor observation in both of training and execution.

Standard QD-Learning

- Consider a networked MARL system consisting of N-agents

$$\{\mathcal{S}, \{\mathcal{A}^i\}, P, \{r_i\}_{i\in\mathcal{V}}, \{\mathcal{G}_t\}_{t\geq 0}, \{\mathcal{M}_{ij}\}_{ij\in\mathcal{E}}\}$$

Following the standard setting in QD-Learning (Kar. 13), each agent updates the estimation of global Q-value with local information of neighbor agents. The information m is Q-value.

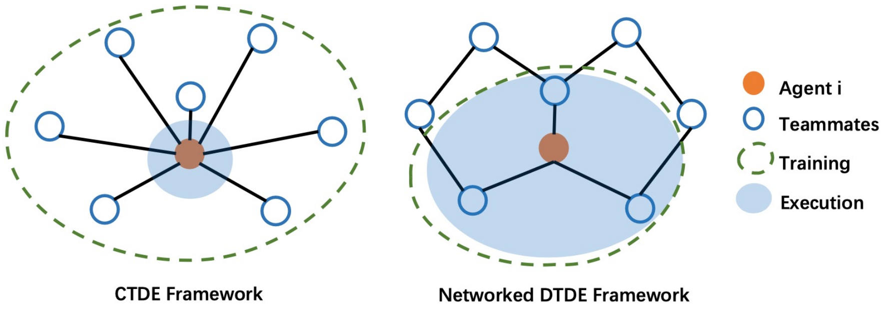


Fig. 1: In CTDE (left), agent i uses all agents' observation in centralized training and self observation in decentralized excution. In networked DTDE (right), agent i uses local neighbor observation in both of training and execution.

- The Q-value of each agent n for each pair (s, a) evolves in the form of *consensus* + *innovation*

$$Q^n_{i,u}(t+1) = Q^n_{i,u}(t) - \beta_{i,u}(t)\underbrace{\boxed{\sum_{l\in\Omega_n(t)}\left(Q^n_{i,u}(t) - Q^l_{i,u}(t)\right)}}_{\textit{Consensus term}} + \alpha_{i,u}(t)\underbrace{\boxed{\left(c_n(\mathbf{x}_t, \mathbf{u}_t) + \gamma \min_{v\in\mathcal{U}} Q^n_{\mathbf{x}_{t+1},v}(t) - Q^n_{i,u}(t)\right)}}_{\textit{Bellman innovation term}}$$

Standard QD-Learning

- Consider a networked MARL system consisting of N-agents

$$\{\mathcal{S}, \{\mathcal{A}^i\}, P, \{r_i\}_{i\in\mathcal{V}}, \{\mathcal{G}_t\}_{t\geq 0}, \{\mathcal{M}_{ij}\}_{ij\in\mathcal{E}}\}$$

Following the standard setting in QD-Learning (Kar. 13), each agent updates the estimation of global Q-value with local information of neighbor agents. The information m is Q-value.

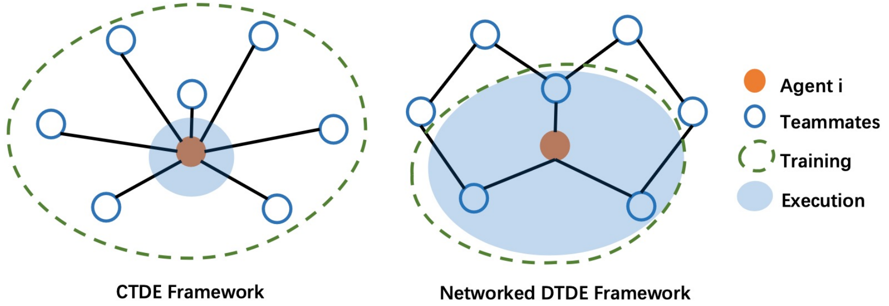


Fig. 1: In CTDE (left), agent i uses all agents' observation in centralized training and self observation in decentralized excution. In networked DTDE (right), agent i uses local neighbor observation in both of training and execution.

- The Q-value of each agent n for each pair (s, a) evolves in the form of *consensus* + *innovation*

$$Q^n_{i,u}(t+1) = Q^n_{i,u}(t) - \beta_{i,u}(t)\underbrace{\boxed{\sum_{l\in\Omega_n(t)}\left(Q^n_{i,u}(t) - Q^l_{i,u}(t)\right)}}_{\text{Consensus term}} + \alpha_{i,u}(t)\underbrace{\boxed{\left(c_n(\mathbf{x}_t, \mathbf{u}_t) + \gamma \min_{v\in\mathcal{U}} Q^n_{\mathbf{x}_{t+1},v}(t) - Q^n_{i,u}(t)\right)}}_{\text{Bellman innovation term}}$$

Considering that the transmitted information over networks could be eavesdropped or monitored by malicious agents, which is highly-related to agents' privacy, it is still an open problem to consider privacy-protecting in networked MARL systems.

DP-QDL: Protecting the Privacy of Networked MARL

Homomorphic encryption: computationally expensive on mobile devices;

End2End encoding: unexplainable;

Differential Privacy: low cost, provable protections, widely used in the database of Google, Amazon, etc.

DP-QDL: Protecting the Privacy of Networked MARL

Homomorphic encryption: computationally expensive on mobile devices;

End2End encoding: unexplainable;

Differential Privacy: low cost, provable protections, widely used in the database of Google, Amazon, etc.

- Add a random Laplace noise into the transmitted Q-value

$$Q^i_{s,a}(t+1) = Q^i_{s,a}(t) - \beta_{s,a}(t) \sum_{v_j \in \mathcal{N}_i(t)} (Q^i_{s,a}(t) - \hat{Q}^j_{s,a}(t)) + \alpha_{s,a}(t)(r_i(s_t, a_t) + \gamma \max_{a' \in \mathcal{A}} Q^i_{s',a'}(t) - Q^i_{s,a}(t))$$

DP-QDL: Protecting the Privacy of Networked MARL

Homomorphic encryption: computationally expensive on mobile devices;

End2End encoding: unexplainable;

Differential Privacy: low cost, provable protections, widely used in the database of Google, Amazon, etc.

- Add a random Laplace noise into the transmitted Q-value

$$Q^i_{s,a}(t+1) = Q^i_{s,a}(t) - \beta_{s,a}(t) \sum_{v_j \in \mathcal{N}_i(t)} (Q^i_{s,a}(t) - \boxed{\hat{Q}^j_{s,a}(t)}) + \alpha_{s,a}(t)(r_i(s_t, a_t) + \gamma \max_{a' \in \mathcal{A}} Q^i_{s',a'}(t) - Q^i_{s,a}(t))$$

DP-Protected message in the consensus term

$$\hat{Q}^i_{s,a}(t) = Q^i_{s,a}(t) + \eta_i(t), \quad \eta_i(t) \sim Lap(0, \iota_i(t)), and \quad \iota_i(t) = s_i q_i^t, q_i \in (0,1),$$

Theorem 1 (Consensus in expectation a.s.)

The Q-value of each agent in DP-QDL can achieve consensus in expectation almost surely as

$$\lim_{t\to\infty} \mathbb{E}[Q^i_{s,a}(t) - \bar{Q}_{s,a}(t)] = 0, i, j = 1, \ldots, N,$$

$$\bar{Q}_{s,a}(t) = \frac{1}{N}\sum_{i=1}^{N} Q^i_{s,a}(t)$$

Brief Proof. $\mathrm{E}[\hat{Q}] = E[Q + \eta] = E[Q]$

Theorem 2 (Consensus in mean square a.s.)

The Q-value of each agent in DP-QDL can achieve asymptotically consensus in mean square almost surely as

$$\lim_{t\to\infty} \mathbb{E}[(Q^i_{s,a}(t) - Q^j_{s,a}(t))^2] = 0, i, j = 1, \ldots, N.$$

3 key steps in proof:

- Construct an auxiliary process y including Laplace noise.
- Y achieves mean square convergence.
- The error between Q and y converges to zero.

Theorem 3 (p, r)-accuracy of the average Q-value

The average Q-value of all agents in DP-QDL can achieve (p, r)-accuracy with the optimal Q* and $r = \frac{\sqrt{2var(\tilde{Q}_{s,a}(t))}}{\sqrt{p}}$.

Step 1: The variance is calculated by using the iterative update of \tilde{Q}.

$$var\left(\tilde{Q}_{s,a}(t)\right) \leq W_0 s_i^2 q_i^{2t-2} \frac{1-\left(\frac{M_t}{q_i^2}\right)^t}{1-\frac{M_t}{q_i^2}},$$

$$M_t = (1 - \alpha_{s,a}(t) + \gamma\alpha_{s,a}(t))^2 \in (0,1) \text{ and } W_0 = \frac{\beta_{s,a}(0)^2}{N^2}\lambda_N(\bar{D}).$$

Step 2: With Chebyshev's inequality and the variance above, we have

$$\begin{aligned} \mathbb{P}(|\breve{Q}_{s,a}(t)| \leq r) &= 1 - \mathbb{P}(|\breve{Q}_{s,a}(t)| > r) \\ &\geq 1 - \frac{2var(\tilde{Q}_{s,a}(t))}{r^2}. \\ &\geq 1 - p, \end{aligned}$$

DP-QDL Experiment Results

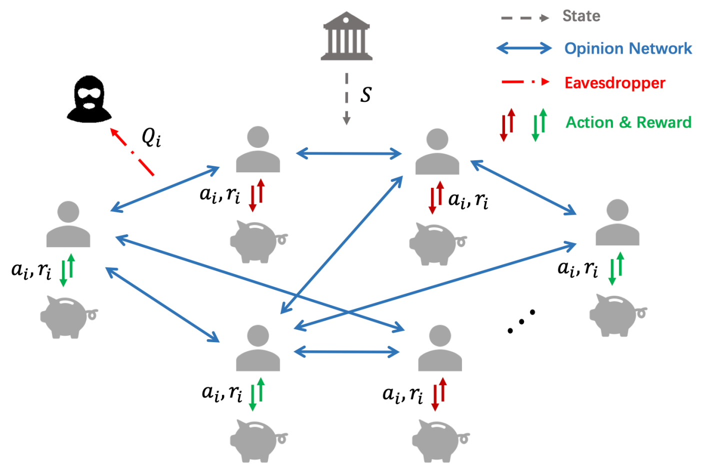


Fig. 1 Center Bank Monetary Policy Environment.

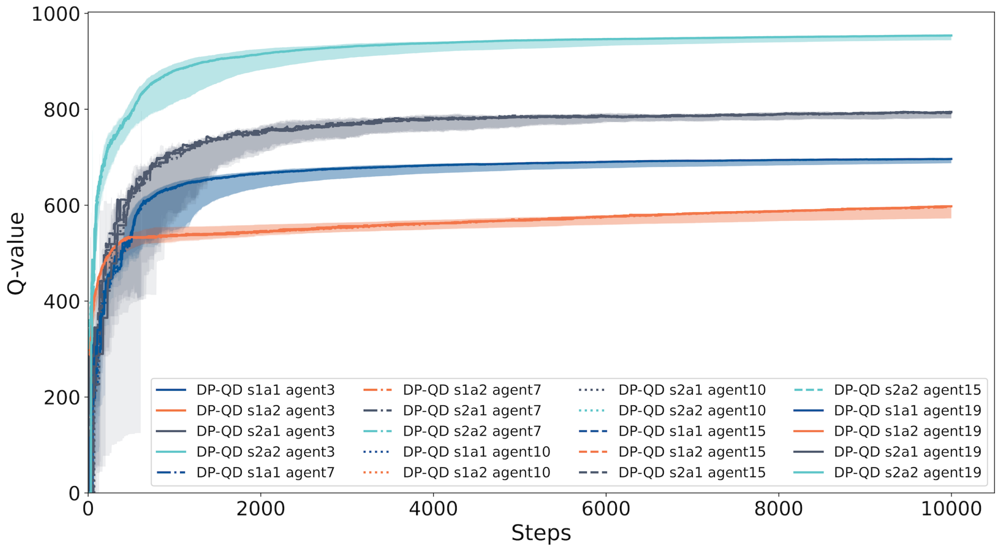


Fig. 2 Convergence in mean square with DP-noise.

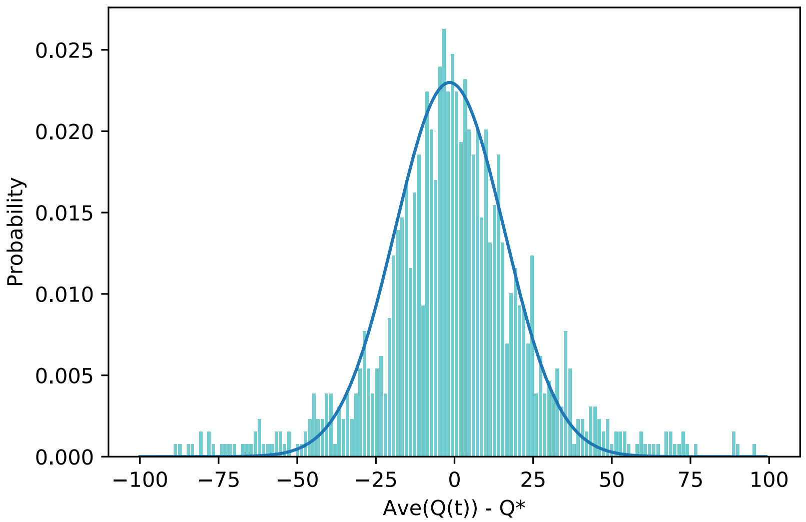


Fig. 3 Average Q-value distribution over 1000 runs.

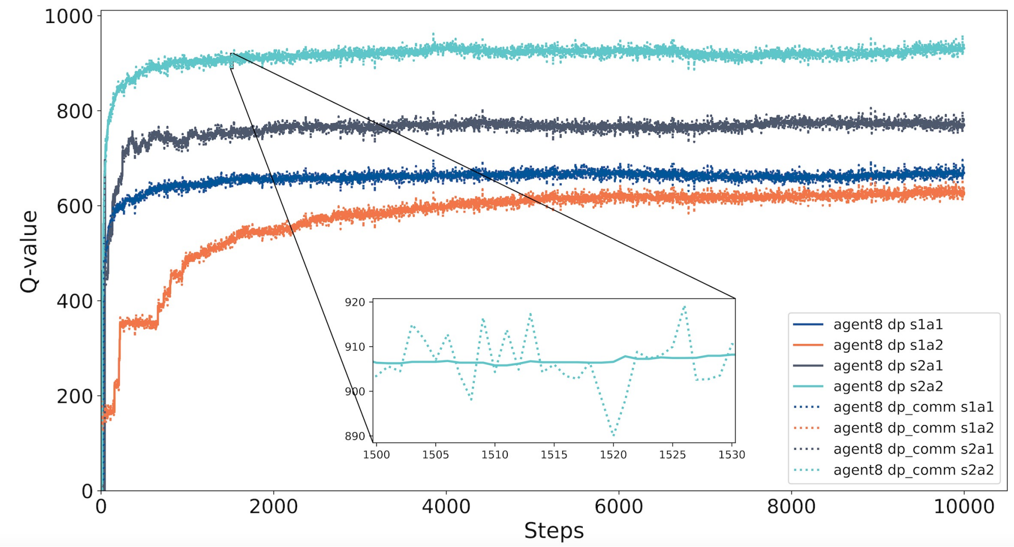


Fig. 4 The Private Q-value and the Real Q-value.

Some Questions

Existing DTDE/networked MARL works focus on the fully cooperative environment without a center. Many works use the consensus of agents' critics to estimate the global critic. The applications of DTDE includes *traffic signals control, grid control, cellular, and multi-robot systems*.

- Considering a networked MARL system with a time-varying communication topology, I'm trying to improve the exploration of networked MARL by maximizing the mutual information between agents and the environment, where the agents can actively change the topology of the information structure. How do we measure this mutual information?
- In addition to the works mentioned above, what are the interesting directions of the DTDE MARL in your opinion?

Thanks!