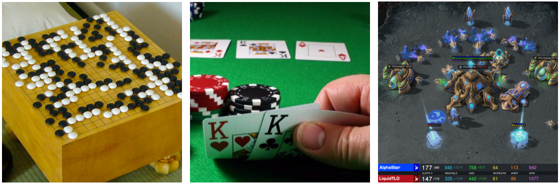
On Improving Model-Free Algorithms for Decentralized Multi-Agent Reinforcement Learning

Weichao Mao[1], Lin F. Yang[2], Kaiqing Zhang[3], Tamer Başar[1]

[1] ECE & CSL, University of Illinois Urbana-Champaign
[2] ECE, University of California, Los Angeles
[3] LIDS, Massachusetts Institute of Technology.

Background

- Multi-agent reinforcement learning (MARL)
  - Sequential decision-making problem where a group of agents strategically interact with each other in a shared environment
  - Cooperative or competitive: Each agent takes actions to maximize its own benefits
  - Leads to breakthroughs in AI applications: Go, Poker, and real-time strategy games.

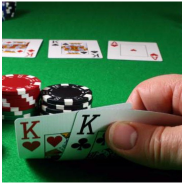


(a) Go (b) Texas hold'em (c) StarCraft II

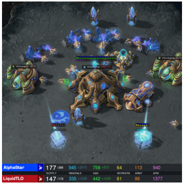
Background

- Multi-agent reinforcement learning (MARL)
  - Sequential decision-making problem where a group of agents strategically interact with each other in a shared environment
  - Cooperative or competitive: Each agent takes actions to maximize its own benefits
  - Leads to breakthroughs in AI applications: Go, Poker, and real-time strategy games.

(a) Go



(b) Texas hold'em



(c) StarCraft II

- Sample-efficient solutions are lacking
  - Often suffer from an exponential sample complexity dependence on the number of agents. "The curse of multiagents"

Decentralized Learning

- We consider a more practical setting: decentralized learning
  - Each agent makes decisions based on only its local information, i.e., local action and reward history
  - Need not communicate with other agents, nor be coordinated by any central controller during training
  - In fact, can be completely oblivious to the presence of other agents

Decentralized Learning

- We consider a more practical setting: decentralized learning
  - Each agent makes decisions based on only its local information, i.e., local action and reward history
  - Need not communicate with other agents, nor be coordinated by any central controller during training
  - In fact, can be completely oblivious to the presence of other agents
- Advantages:
  - Does not suffer exponential sample & computation complexity
  - More practical even if communication is expensive or unreliable
  - Naturally model-free: higher space efficiency, and compatible with deep RL architectures

Contributions

A series of sample-efficient decentralized MARL algorithms:

1. For general-sum Markov games, we propose a stage-based V-learning algorithm that learns an ε-approximate coarse correlated equilibrium (CCE) in $\widetilde{O}(H^5 S A_{\max}/\varepsilon^2)$ episodes
   - Stage-based Q-learning for exploration, and adversarial bandit subroutine for policy update
   - Circumvents a rather complicated no-weighted-regret bandit subroutine in existing works

Contributions

A series of sample-efficient decentralized MARL algorithms:

1. For general-sum Markov games, we propose a stage-based V-learning algorithm that learns an ε-approximate coarse correlated equilibrium (CCE) in $\widetilde{O}(H^5 S A_{\max}/\varepsilon^2)$ episodes
   - Stage-based Q-learning for exploration, and adversarial bandit subroutine for policy update
   - Circumvents a rather complicated no-weighted-regret bandit subroutine in existing works
2. For general-sum Markov games, learns an ε-approximate correlated equilibrium (CE) in $\widetilde{O}(H^5 S A_{\max}^2/\varepsilon^2)$ episodes

Contributions

A series of sample-efficient decentralized MARL algorithms:

1. For general-sum Markov games, we propose a stage-based V-learning algorithm that learns an ε-approximate coarse correlated equilibrium (CCE) in $\widetilde{O}(H^5 S A_{\max}/\varepsilon^2)$ episodes
   - Stage-based Q-learning for exploration, and adversarial bandit subroutine for policy update
   - Circumvents a rather complicated no-weighted-regret bandit subroutine in existing works
2. For general-sum Markov games, learns an ε-approximate correlated equilibrium (CE) in $\widetilde{O}(H^5 S A_{\max}^2/\varepsilon^2)$ episodes
3. For Markov potential games, an independent policy gradient algorithm with a decentralized momentum-based variance reduction technique for learning Nash equilibrium (NE).

Contributions

A series of sample-efficient decentralized MARL algorithms:

1. For general-sum Markov games, we propose a stage-based V-learning algorithm that learns an ε-approximate coarse correlated equilibrium (CCE) in $\widetilde{O}(H^5 S A_{\max}/\varepsilon^2)$ episodes
   - Stage-based Q-learning for exploration, and adversarial bandit subroutine for policy update
   - Circumvents a rather complicated no-weighted-regret bandit subroutine in existing works
2. For general-sum Markov games, learns an ε-approximate correlated equilibrium (CE) in $\widetilde{O}(H^5 S A_{\max}^2/\varepsilon^2)$ episodes
3. For Markov potential games, an independent policy gradient algorithm with a decentralized momentum-based variance reduction technique for learning Nash equilibrium (NE).
4. Numerical simulations corroborate our theoretical findings