Self-Play or Group Practice: Learning to Play Alternating Markov Game in Multi-Agent System

Chin-wing Leung, Ho-fung Leung
The Chinese University of Hong Kong
Hong Kong, China

Shuyue Hu
National University of Singapore
Singapore

Summary

- We consider a population of agents each independently learns to play an alternating Markov game (AMG)
- We propose a new training framework ---group practice--- for a population of decentralized RL agents
- The convergence result to the optimal value function and the Nash equilibrium are proved under the GP framework
- Experiments verify that GP is the more efficient training scheme than self-play (SP) given the same amount of training

Group Practice Framework

Algorithm 1 GP framework for standard Q-learning

Input: $N, NIter, Nep, scheme, \alpha, \gamma, \epsilon$

create and initialize $Agents[0]$ to $Agents[N-1]$
for $iteration = 1$ to $NIter$ **do**
  $groups = generateGroups(scheme)$
  for $episode = 1$ to Nep **do**
    for all $group$ in $groups$ **parallel do**
      $pairs = pairUp(group)$
      for all $pair$ in $pairs$ **parallel do**
        $playGame(Agents[pair[0]], Agents[pair[1]])$
      end for
      for all $agent$ in $Agents$ **parallel do**
        $trainAgent(agent)$
      end for
    end for
  end for
end for

Experiment settings

Environments and RL Models

- Connect Four: 4x5 → standard Q-learning
- Connect Four: 6x7 → Deep reinforcement learning with MCTS
- Hex: 7x7 → Deep reinforcement learning with MCTS

Training Schemes

- SP-0.0: agents are trained under SP;
- SP-0.2: agents are trained under SP with an additional exploration probability of 0.2;
- GP-RGS: agents are trained under GP with random grouping scheme;
- GP-LGS-6: agents are trained under GP with local grouping scheme with group size 6;
- GP-RGS-12: agents are trained under GP with local grouping scheme with group size 12.

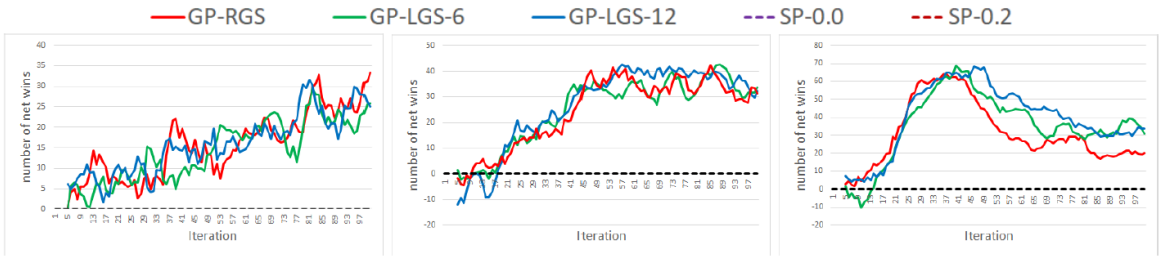
Proof of Convergence

Assumption 1. *For every agent i, the state-action pair (i, s, a) is visited infinitely often during training.*

Assumption 2. *The learning rate is decayed such that $0 \leq \alpha_t \leq 1$, $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.*

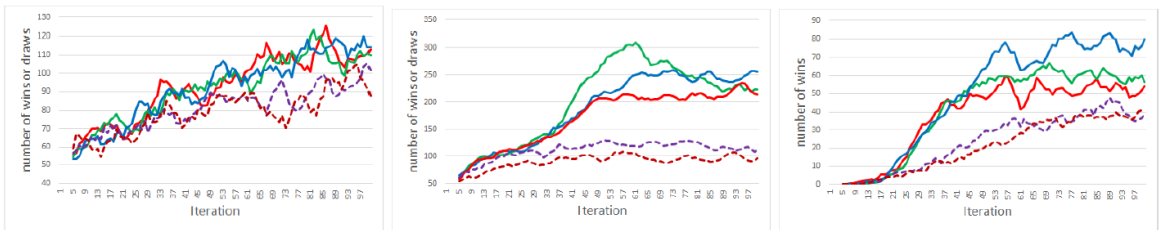
Theorem 1. *With Assumptions 1 and 2, the Q-values for all agents will converge to the fixed point Q^* in alternating Markov game under the GP framework.*

Experiment results



(a) Connect 4 (4x5), Q-learning (b) Connect 4 (6x7), deep Q-learning (c) Hex (7x7), deep Q-learning

Fig. 2: average number of net wins by iterations: matching against SP-0.0 agents



(a) Connect 4 (4x5), Q-learning (b) Connect 4 (6x7), deep Q-learning (c) Hex (7x7), deep Q-learning

Fig. 3: average number of wins or draws by iterations: matching against 90% perfect agent