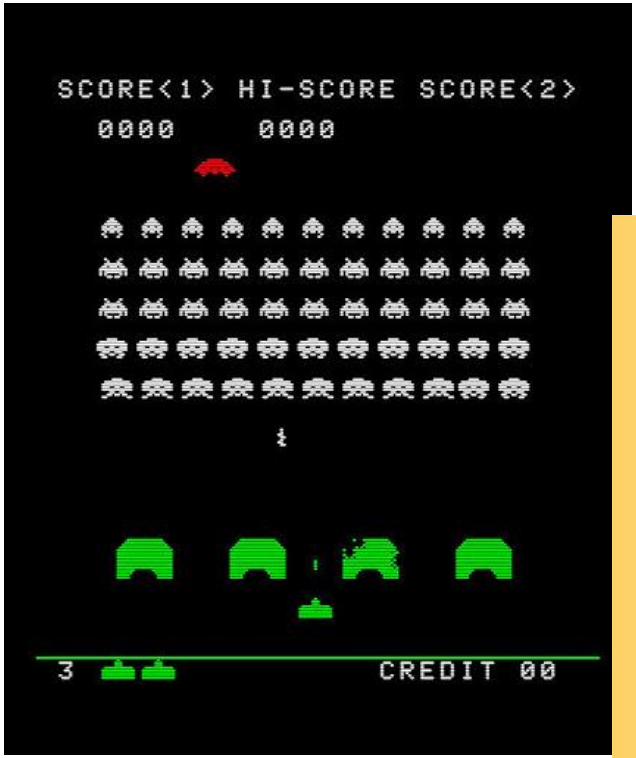
Self-Play or Group Practice: Learning to Play Alternating Markov Game in Multi-Agent System

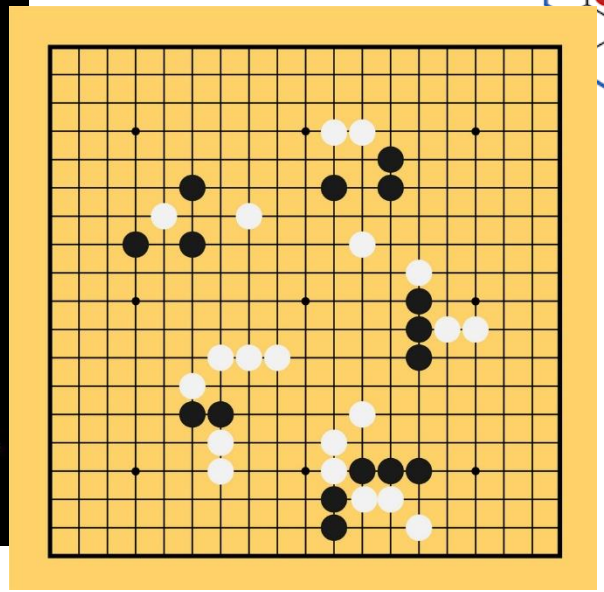
CHIN-WING LEUNG, SHUYUE HU AND HO-FUNG LEUNG

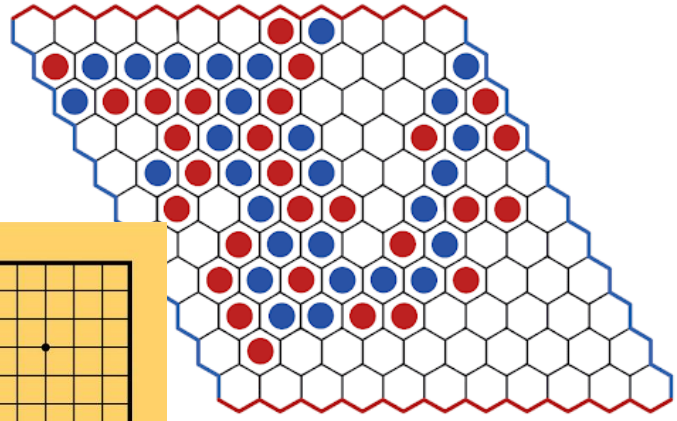
Abstract

- 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

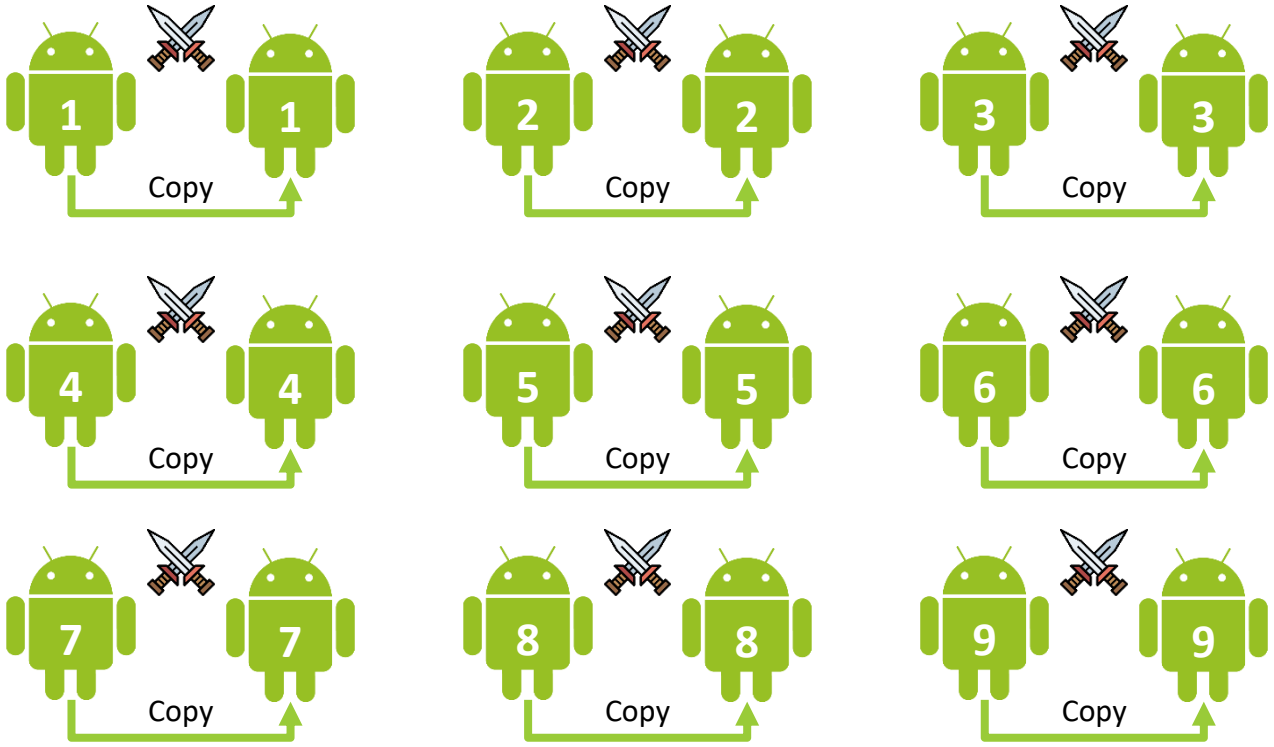
RL in competitive multi-player games



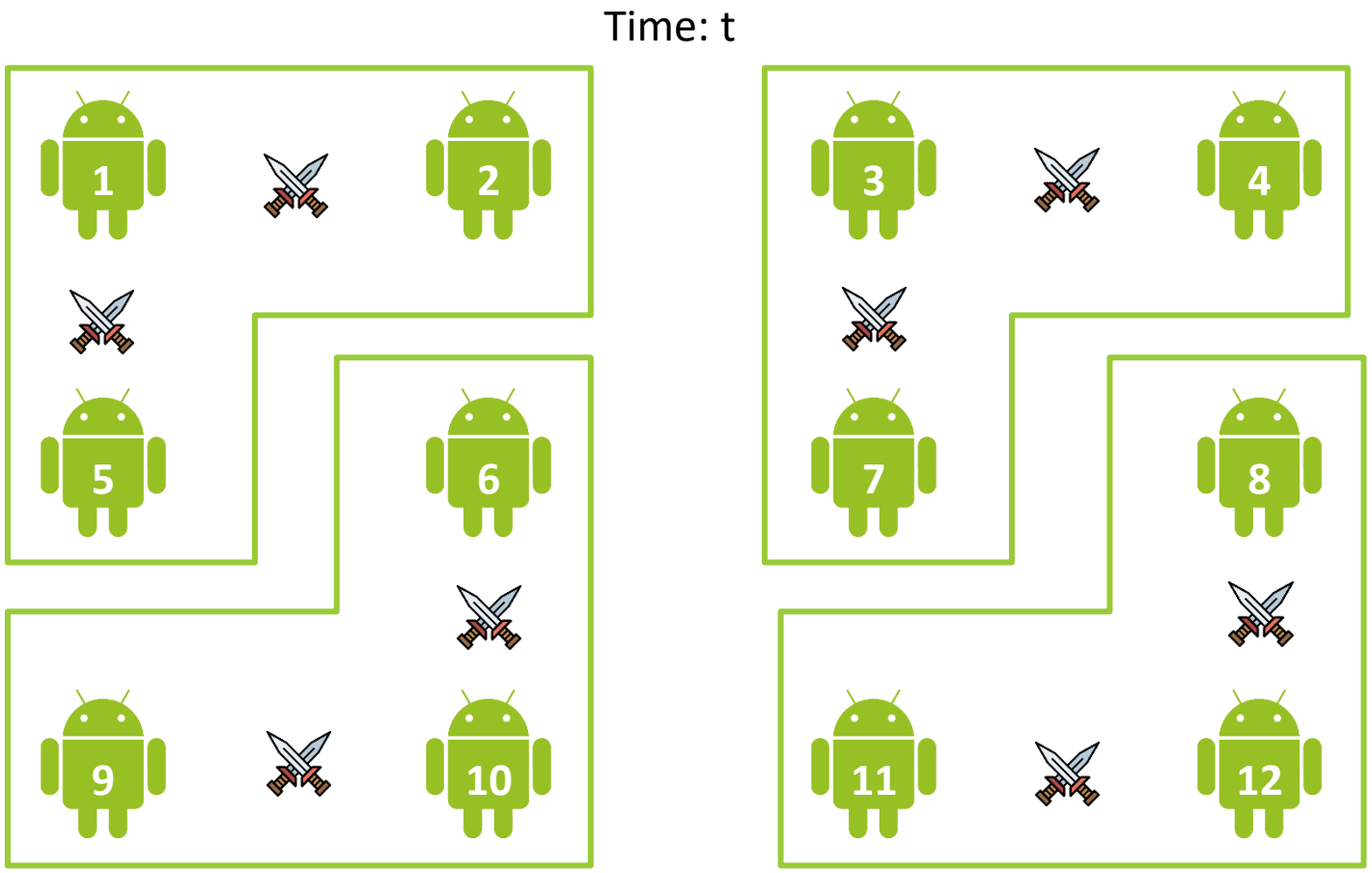




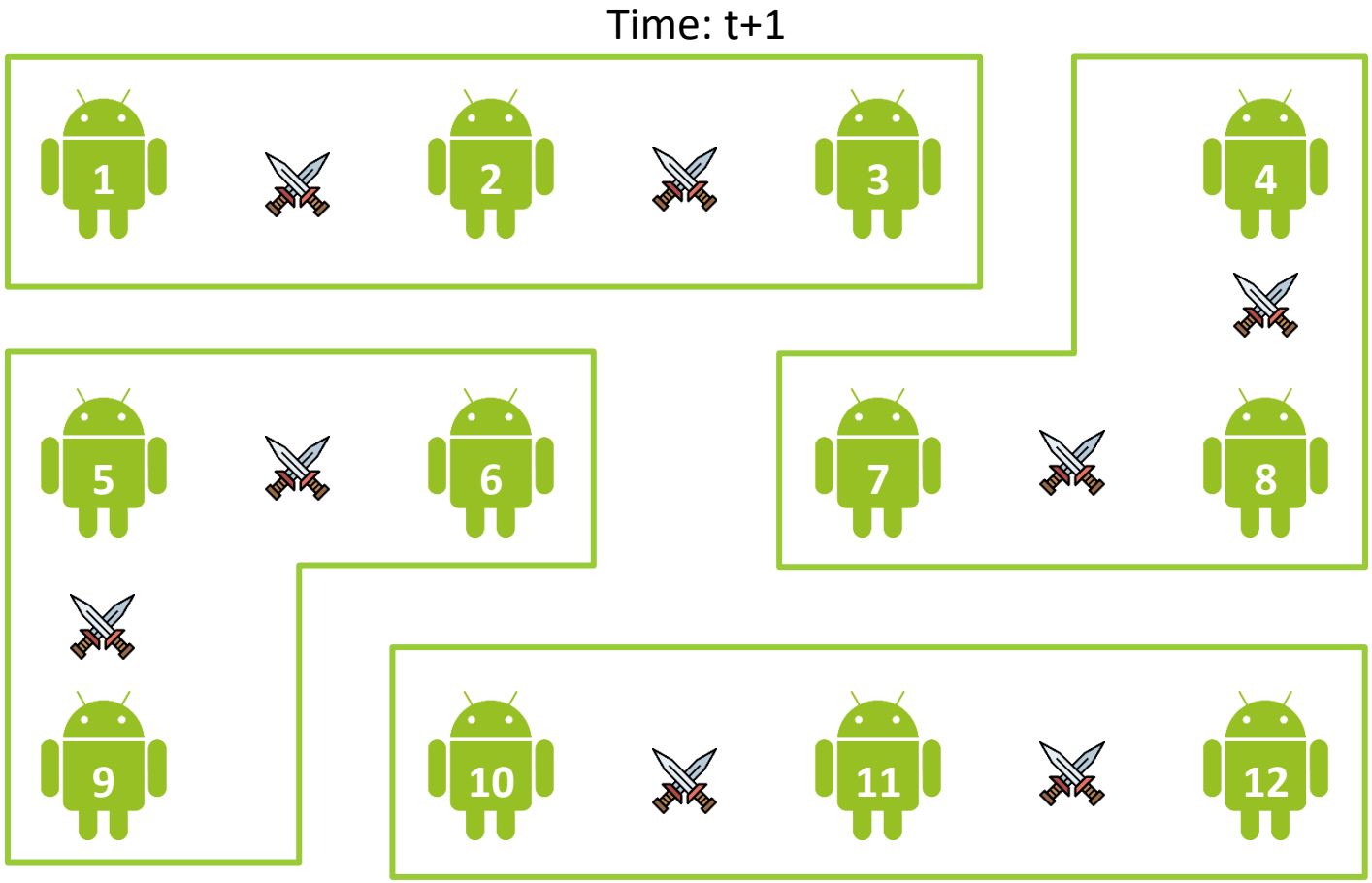
Self-Play


1
1
Copy
2
2
Copy
3
3
Copy
4
4
Copy
5
5
Copy
6
6
Copy
7
7
Copy
8
8
Copy
9
9
Copy

Group Practice



Group Practice



Collaborative Learning in Educational Psychology

Group Practice

Algorithm 1 GP framework for standard Q-learning

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

```
1: create and initialize Agents[0] to Agents[N − 1]
2: for iteration = 1 to NIter do
3:    groups = generateGroups(scheme)
4:    for episode = 1 to Nep do
5:       for all group in groups parallel do
6:          pairs = pairUp(group)
7:          for all pair in pairs parallel do
8:             playGame(Agents[pair[0]], Agents[pair[1]])
9:          end for
10:         for all agent in Agents parallel do
11:            trainAgent(agent)
12:         end for
13:      end for
14:   end for
15: end for
```

Proof of convergence

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

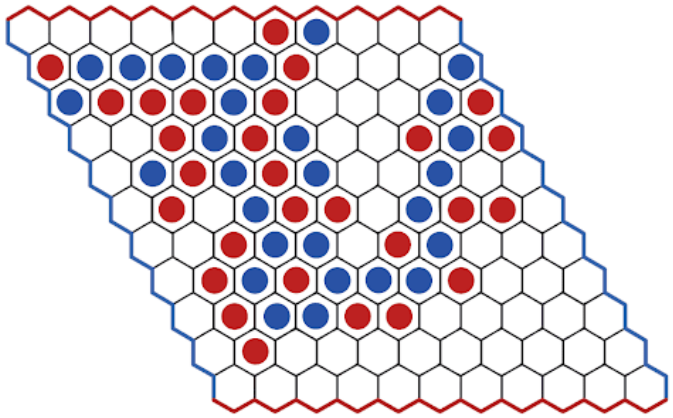
Assumption 2. *The learning rate is decayed such that $0 \leq \alpha_t \leq 1$, $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.*

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.*

Experiments

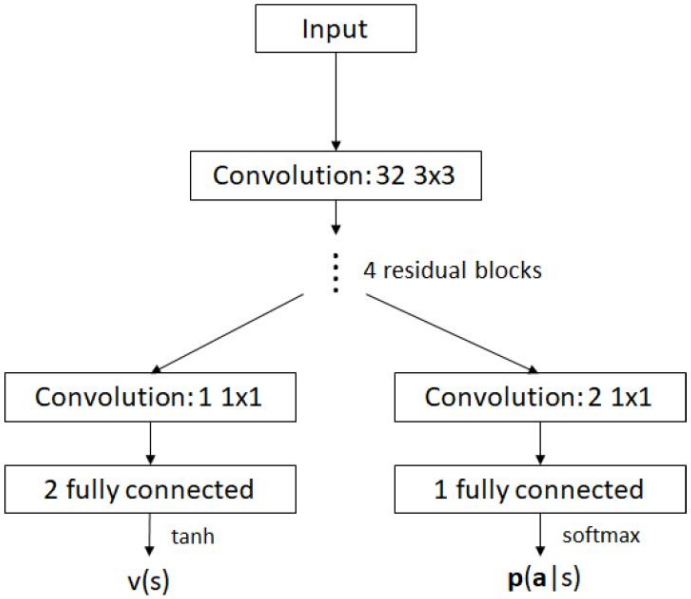
Environments

- Connect Four: 4x5
- Connect Four: 6x7
- Hex: 7x7



Models

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

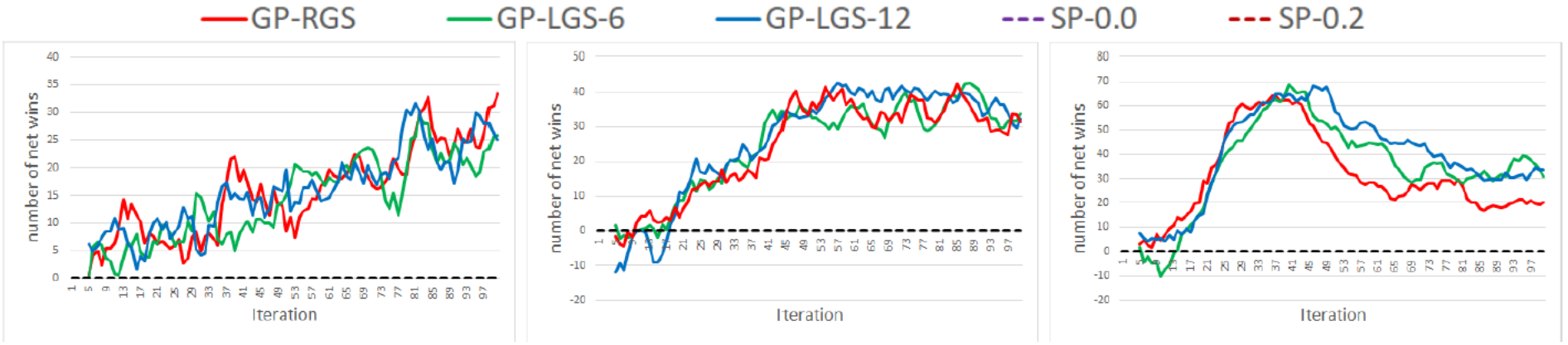


Training Schemes

- Number of agents per scheme = 48

- **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-LGS-12**: agents are trained under GP with local grouping scheme with group size 12.

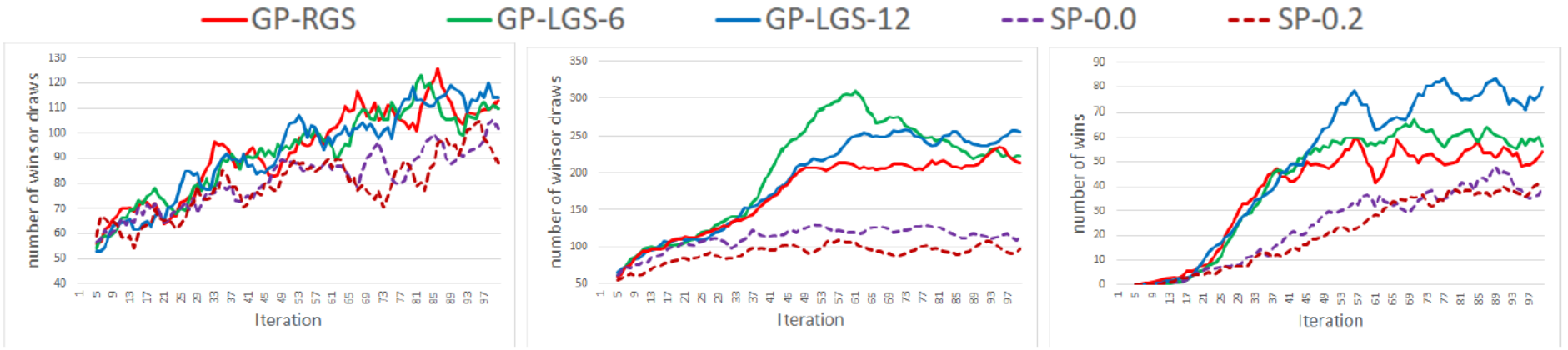
GP agents matching against SP agents



(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

SP, GP agents matching against 90-percent perfect player



(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

Conclusion

- We propose the new group practice (GP) training framework for a population of decentralized RL agents
- We prove that for a population of Q-learning agents, training via group practice will naturally result in the convergence to the optimal value function and the Nash equilibrium
- We show that given the same amount of training, agents trained via group practice generally defeat those trained via self-play across diverse settings
- We also show that the learning effectiveness can even be improved when applying local grouping to agents

End