Multi-agent Reinforcement Learning (2)

Prof. Jun Wang

Computer Science, UCL

Oct 15 2018, SJTU

Recap on MARL (1)

- Stochastic Games
  - Policy Iteration/Value Iteration (model based)
- Equilibrium Learners (model free)
  - Nash-Q
  - Minimax-Q
  - Friend-Foe-Q
- Best-Response Learners (model free)
  - JAL and Opponent Modelling
  - Iterated Gradient Ascent
  - Wolf-IGA

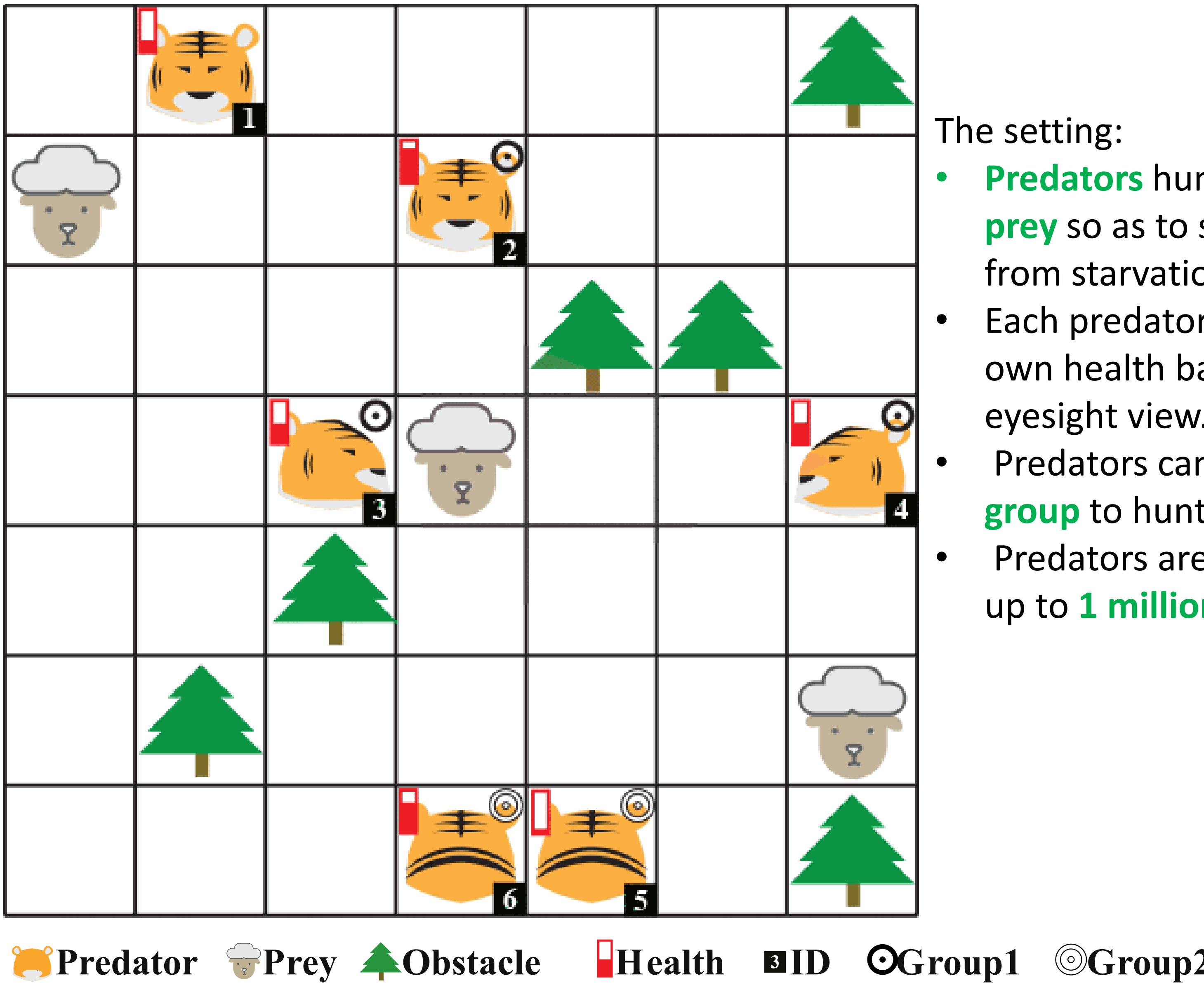
Tables of Content

- Many agents
  - Population dynamics
  - CLEAN rewards
  - Mean-field MARL
- Multi-agent Communications
  - CommNet
  - DIAL
  - BiCNet

Tables of Content

- Many agents
  - Population dynamics
  - CLEAN rewards
  - Mean-field MARL
- Multi-agent Communications
  - CommNet
  - DIAL
  - BiCNet

Artificial Population: Large-scale predator-prey world



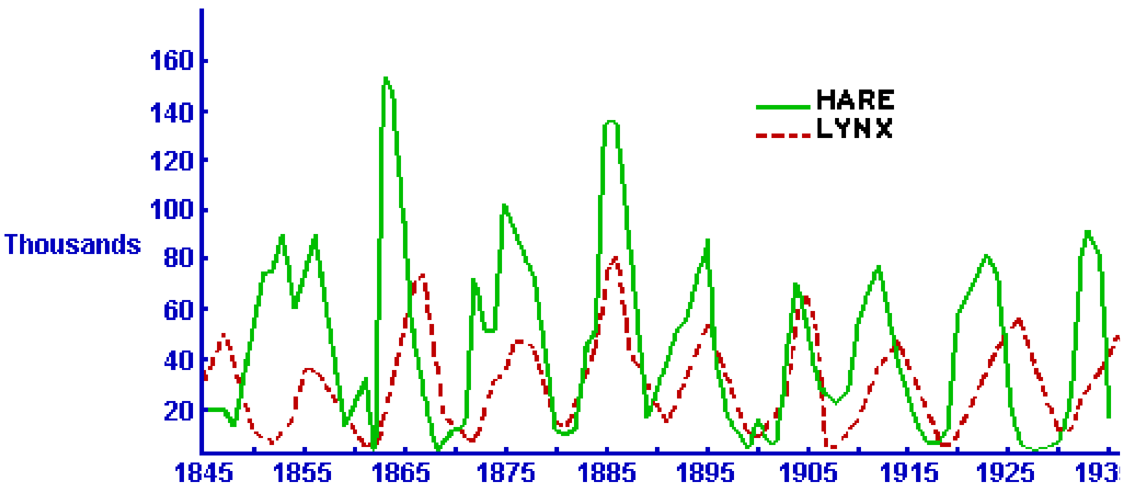
The setting:

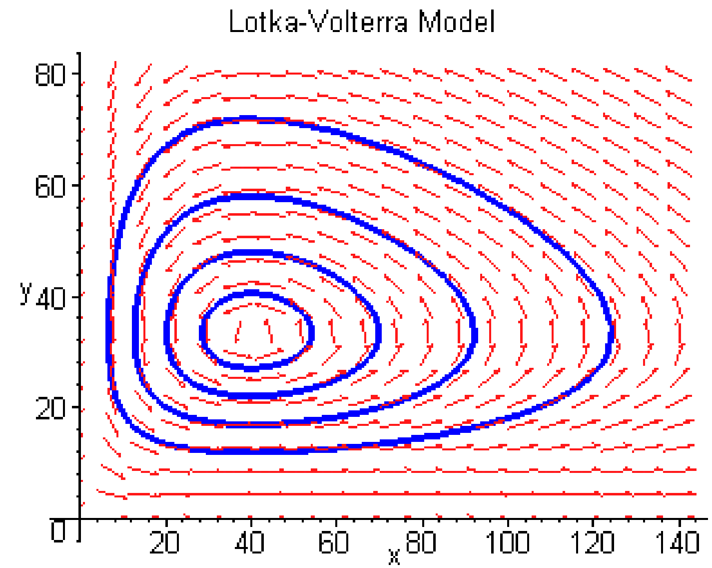
- **Predators** hunt the **prey** so as to survive from starvation.
- Each predator has its own health bar and eyesight view.
- Predators can form **a group** to hunt the prey
- Predators are scaled up to **1 million**

Yaodong Yang , Lantao Yu , Yiwei Bai , Jun Wang , Weinan Zhang , Ying Wen , Yong Yu, , Dynamics of Artificial Populations by Million-agent Reinforcement Learning, 2017

Ecology: the Lotka-Volterra (LV) model

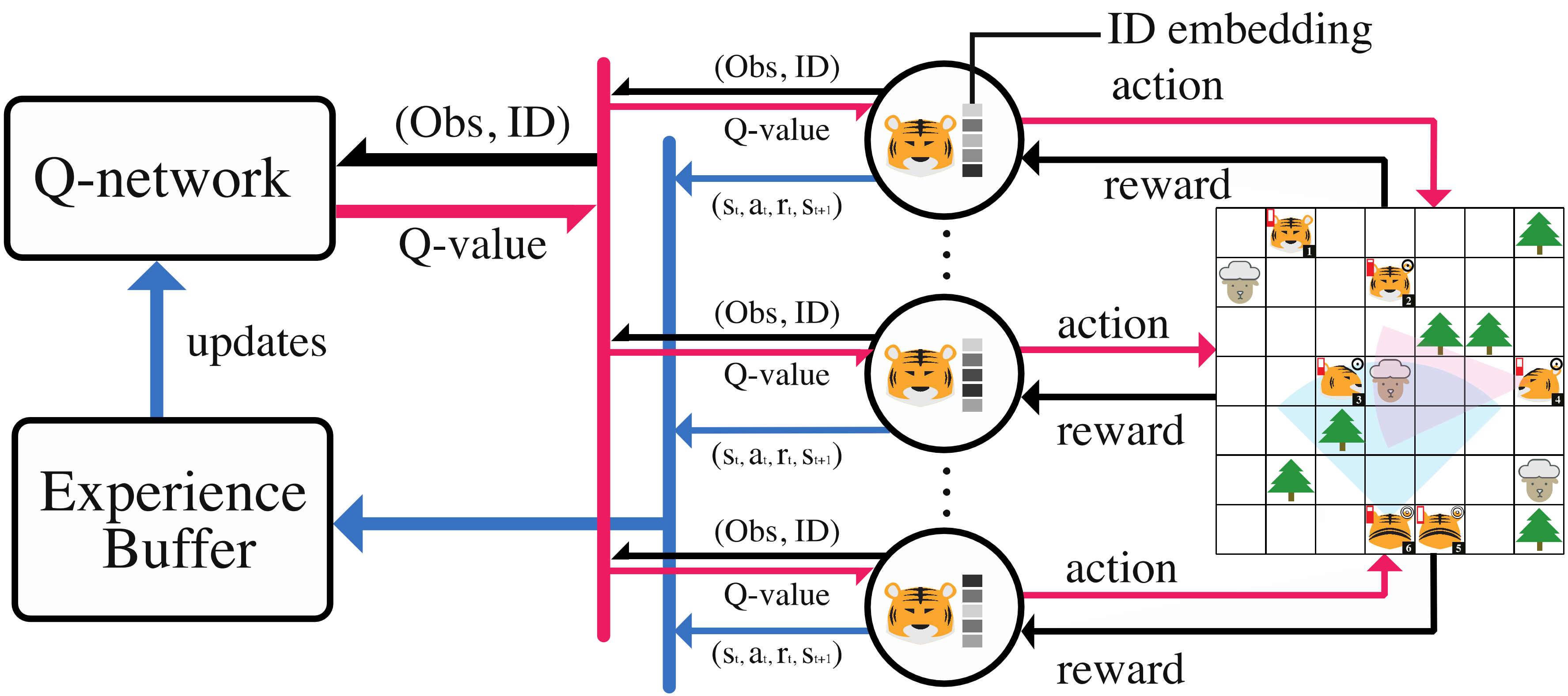
- A major topic of population dynamics is the cycling of predator and prey populations
- The *Lotka-Volterra* model is used to model this
- lynx (wild cat) and hare





Lotka, A. J. (1910). "Contribution to the Theory of Periodic Reaction". *J. Phys. Chem.* **14** (3): 271–274.

Reinforcement Learning with 1 millions agents

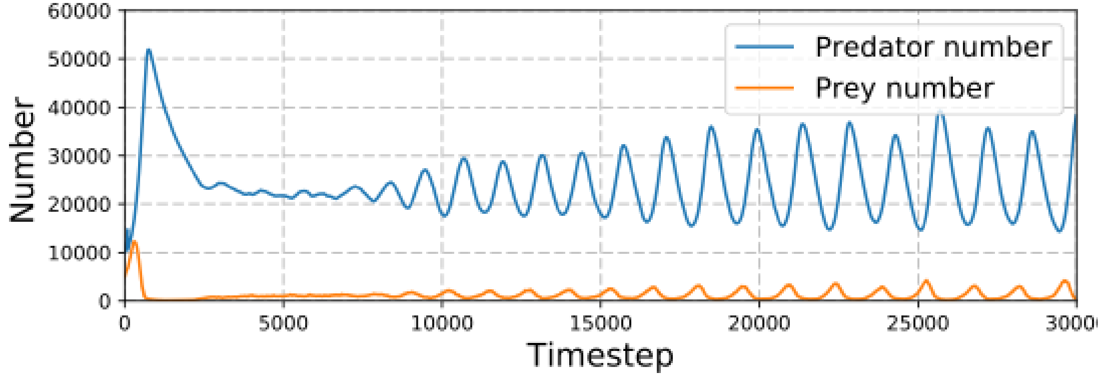


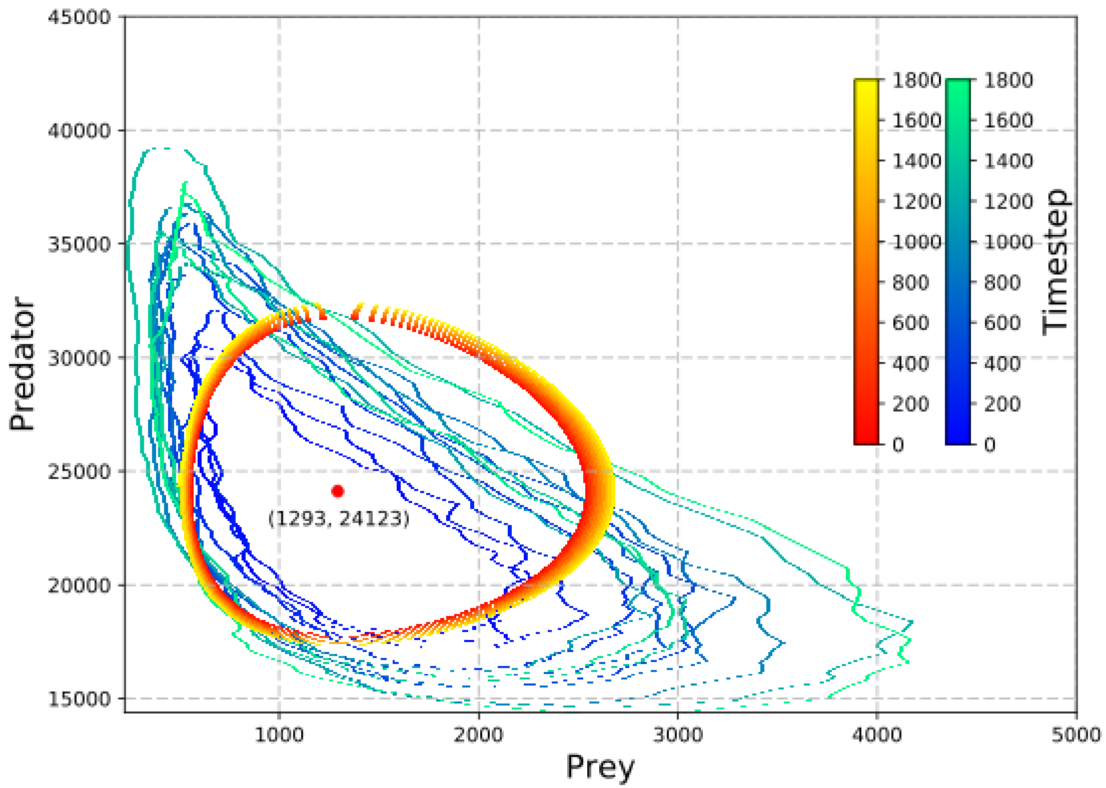
$$Q(s_t^i, a_t^i) \leftarrow Q(s_t^i, a_t^i) + \alpha[r_t^i + \gamma \max_{a' \in A} Q(s_{t+1}^i, a') - Q(s_t^i, a_t^i)].$$

The action space A: {move forward, backward, left, right, rotate left, rotate right, stand still, join a group, and leave a group}.

Yaodong Yang , Lantao Yu , Yiwei Bai , Jun Wang , Weinan Zhang , Ying Wen , Yong Yu, , Dynamics of Artificial Populations by Million-agent Reinforcement Learning, 2017

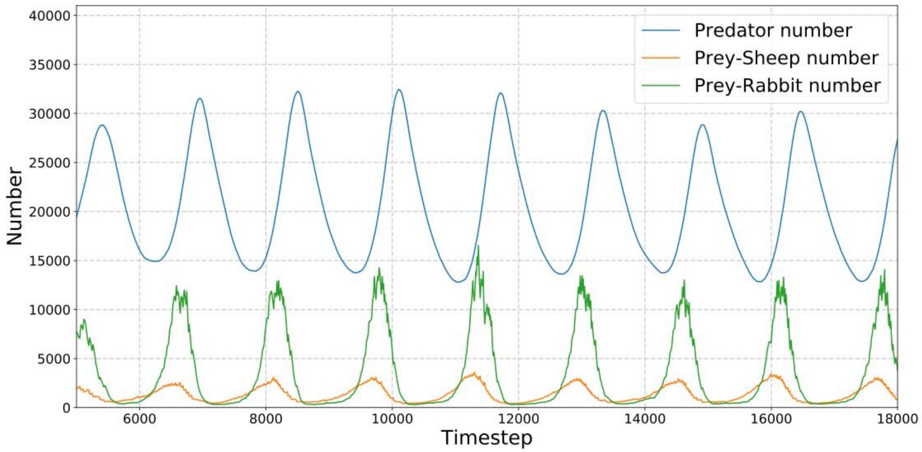
The Dynamics of the Artificial Population

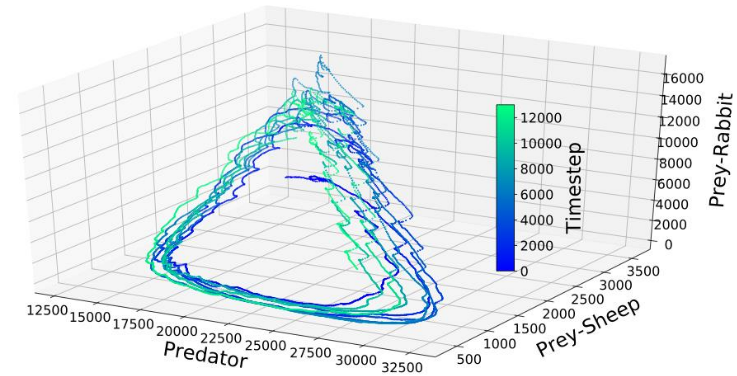


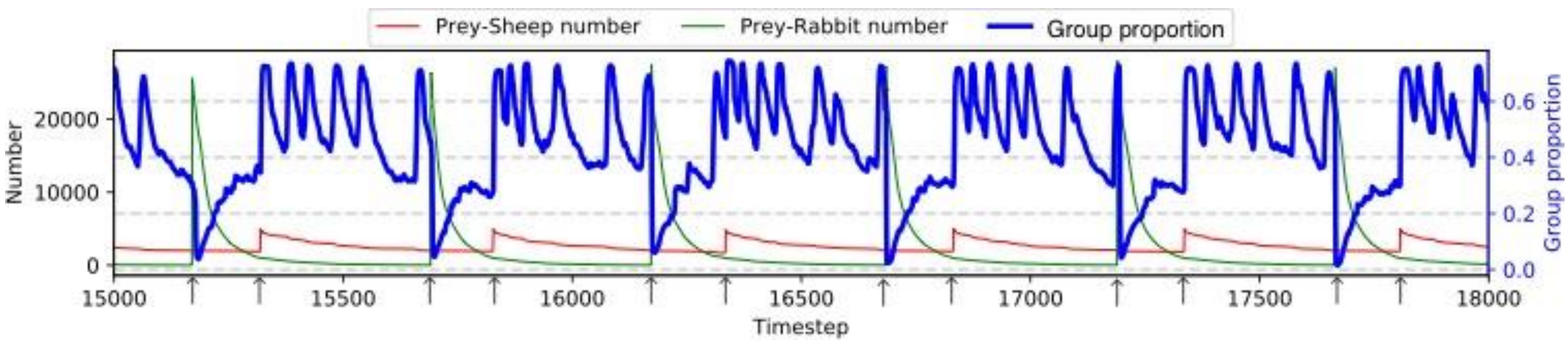


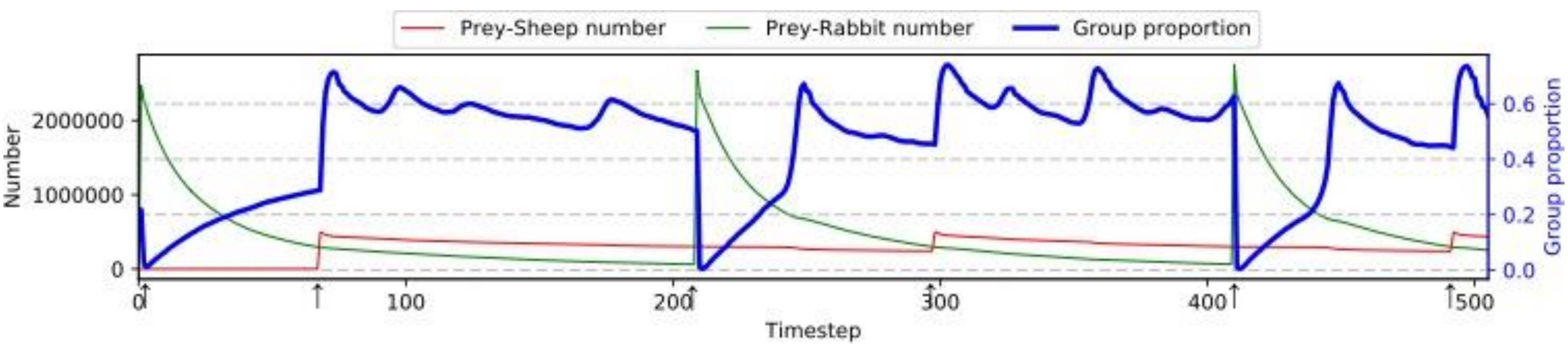
Yaodong Yang , Lantao Yu , Yiwei Bai , Jun Wang , Weinan Zhang , Ying Wen , Yong Yu, , Dynamics of Artificial Populations by Million-agent Reinforcement Learning, 2017

Tiger-sheep-rabbit: Grouping









Yaodong Yang , Lantao Yu , Yiwei Bai , Jun Wang , Weinan Zhang , Ying Wen , Yong Yu, , Dynamics of Artificial Populations by Million-agent Reinforcement Learning, 2017

Exploratory Action Noise

- Agents in the system provide a constantly changing background in which each agent needs to learn its task
  - As a consequence, agents need to extract the underlying reward signal from the noise of other agents acting within the environment
- This learning noise can have a significant impact on the resultant system performance

$$Q_i(s, ai, a_j)$$

Condition on other agent actions: a_j but they are also exploring – the actual a_j contains some element of exploration and not their intended actions

C. Holmesparker, M. E. Taylor, A. K. Agogino, and K. Tumer. Clean rewards to improve coordination by removing exploratory action noise. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03, pages 127–134, 2014.

CLEAN rewards

- **Coordinated** Learning without **Exploratory Action Noise** (CLEAN) aims to remove exploratory noise present in the global reward
  - This is achieved by private exploration
- Specifically, at each learning episode, each agent executes an action by following its greedy policy (i.e. without exploration);
- then all the agents receive a global re-ward.
- Each agent then privately computes the (global) reward it would have received had it executed an exploratory action, while the rest of the agents followed their greedy policies.

$$Q_i(s, \overbrace{ai, a_j}^{\text{Result in greedy actions}})$$

a_j =argmax Q_j

a_i =argmax Q_i

$$Q_j(s, aj, a_i)$$

C. Holmesparker, M. E. Taylor, A. K. Agogino, and K. Tumer. Clean rewards to improve coordination by removing exploratory action noise. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03, pages 127–134, 2014.

CLEAN rewards

- CLEAN rewards were defined:

$$D_i = \widehat{R}_i(s, a_i^c, aj) - Ri(s, a_i\ \ , aj)$$

  - where $(a_i\ \ , aj)$ is the joint action executed when all agents followed their greedy policies,
  - a_i^c is the counterfactual (offline) action taken by agent i following ε-greedy,
  - R_i is the reward of agent i received when all agents executed their greedy policies and
  - $\widehat{R}_i(s, a_i^c, aj)$ is the counterfactual (offline) reward agent i would have received, had it executed the counterfactual action a_i^c, instead of action a_i, while the rest of the agents followed their greedy policies.
- Each agent then uses the following formula to update its Q-values:

$$Q_i(s, a_i^c, aj) \leftarrow Q_i(s, a_i^c, aj) + \alpha(D_i - Q_i(s, a_i^c, aj))$$

- which removes the exploratory noise caused by other agents and
- allow each agent to effectively determine which actions are beneficial or not

C. Holmesparker, M. E. Taylor, A. K. Agogino, and K. Tumer. Clean rewards to improve coordination by removing exploratory action noise. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03, pages 127–134, 2014.

CLEAN rewards: Experiment

- The Gaussian Squeeze Domain (GSD):
  - There is a set of agents in which each agent contributes to a system objective

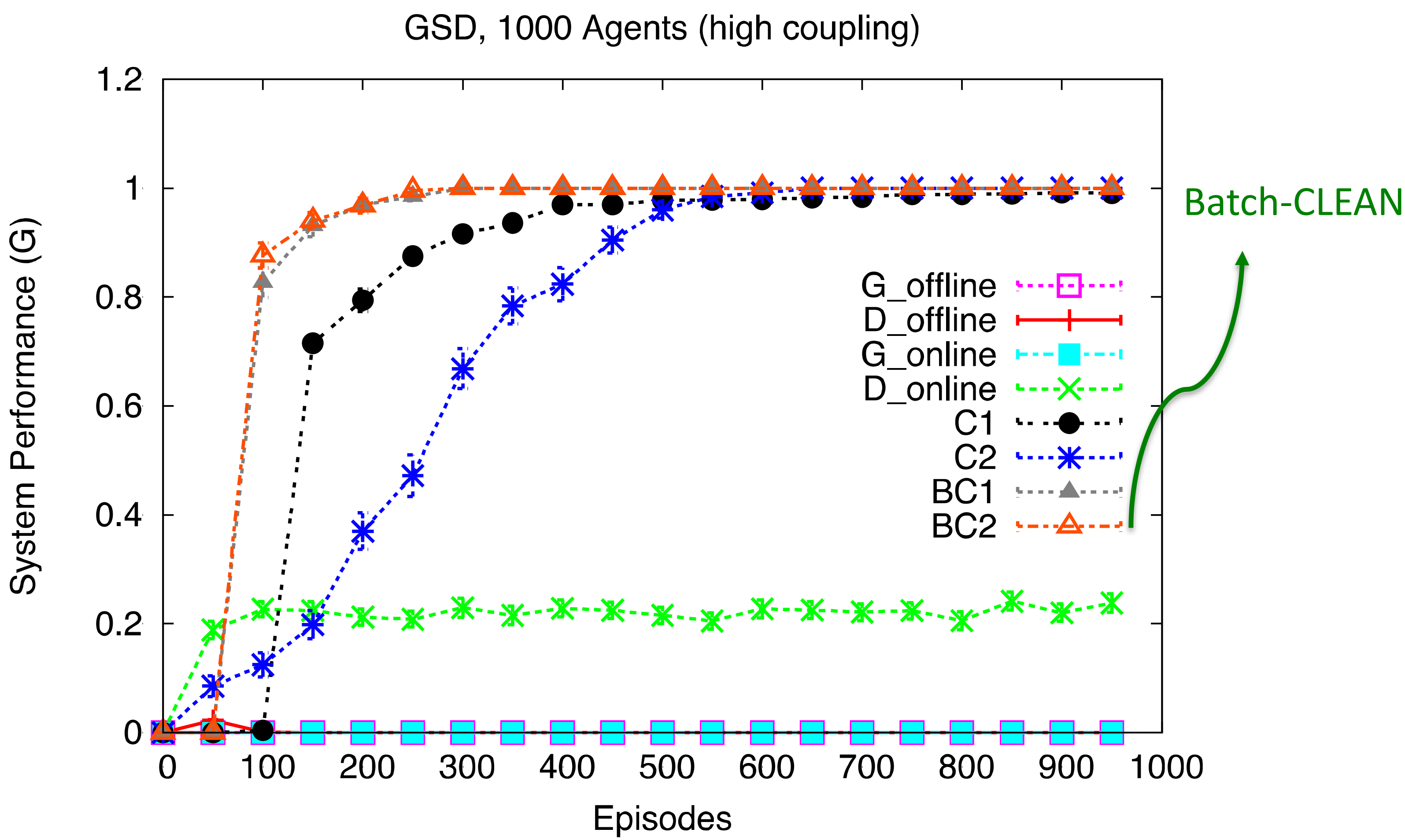
$$G(x) = xe^{\frac{-(x-\mu)^2}{\sigma^2}} \qquad x = \sum_{i=0}^{n} a_i$$

μ and σ are parameters

  - The goal of the agents is to choose their individual actions a_i in such a way that the sum of their individual actions optimize the objective

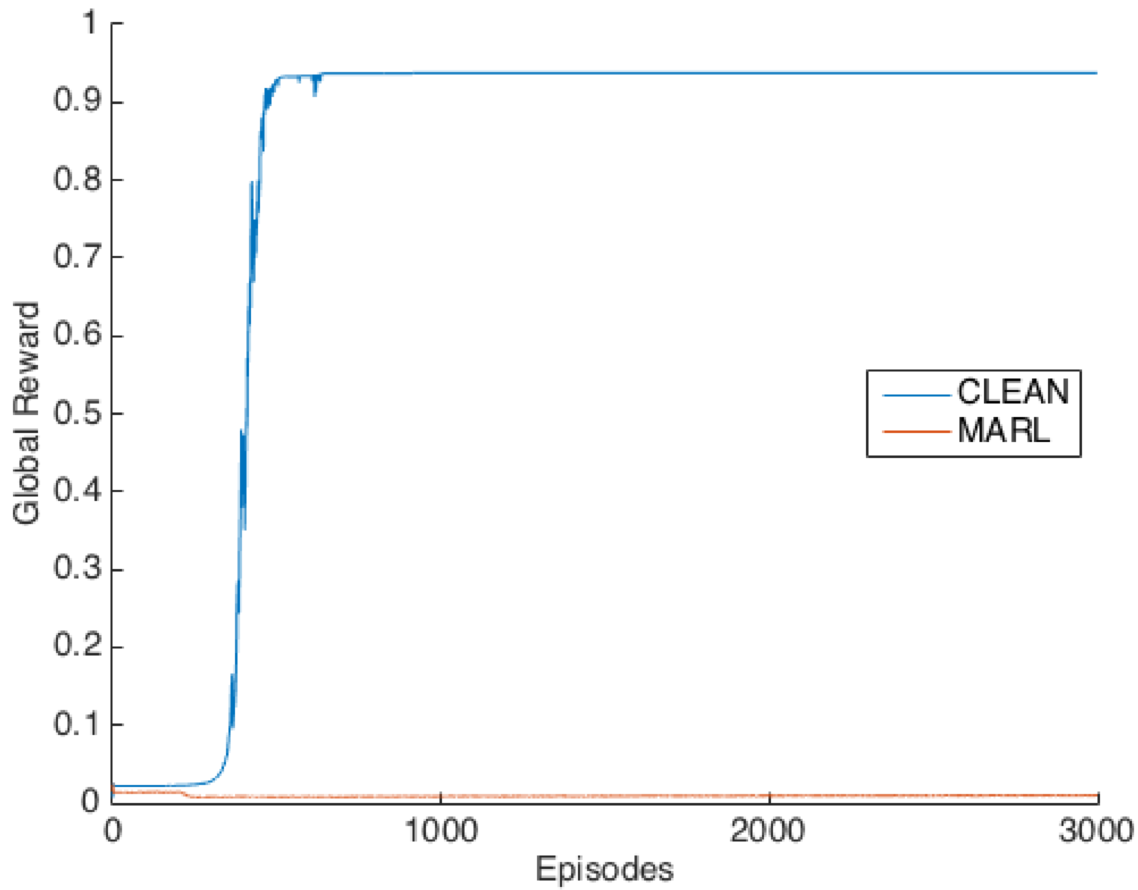
C. Holmesparker, M. E. Taylor, A. K. Agogino, and K. Tumer. Clean rewards to improve coordination by removing exploratory action noise. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03, pages 127–134, 2014.

CLEAN rewards: Experiment



C. Holmesparker, M. E. Taylor, A. K. Agogino, and K. Tumer. Clean rewards to improve coordination by removing exploratory action noise. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03, pages 127–134, 2014.

CLEAN rewards: Feature selection Experiment



Malialis, Kleanthis, et al. "Feature Selection as a Multiagent Coordination Problem." arXiv preprint arXiv:1603.05152(2016).

Mean-field MARL

- *Mean Field Reinforcement Learning*
  - interactions within the population of agents are approximated by those between a single agent and the average effect from neighbouring agents;
  - the interplay between the two entities is mutually reinforced:
    - the learning of the individual agent's optimal policy depends on the dynamics of the population,
    - while the dynamics of the population change according to the collective patterns of the individual policies.

$$Q^j(s, \boldsymbol{a}) \equiv \frac{1}{N^j} \sum_{k \in \mathcal{K}^j} Q^j(s, a^j, a^k),$$

Joint action is replaced by pairwise interactions

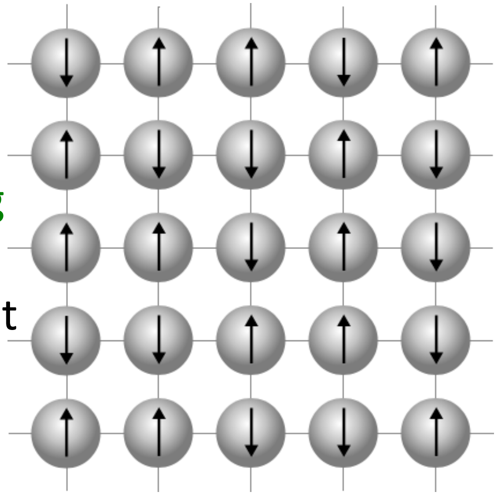
$$Q^j_{t+1}(s, a^j, \bar{a}) = (1 - \alpha_t) Q^j_t(s, a^j, \bar{a}) + \alpha_t \left[r^j_t + \gamma v^j_t(s') \right]$$

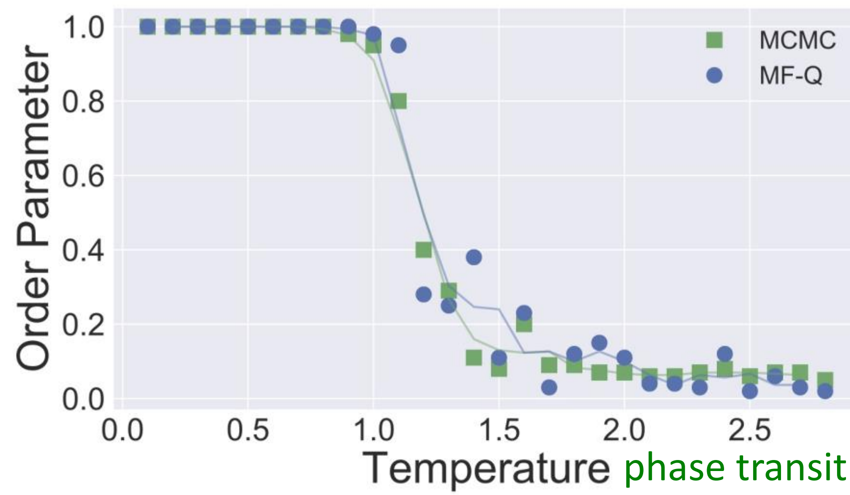
Interplayed with a mean agent

Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J. Mean Field Multi-Agent Reinforcement Learning. arXiv preprint arXiv:1802.05438. 2018 Feb 15.

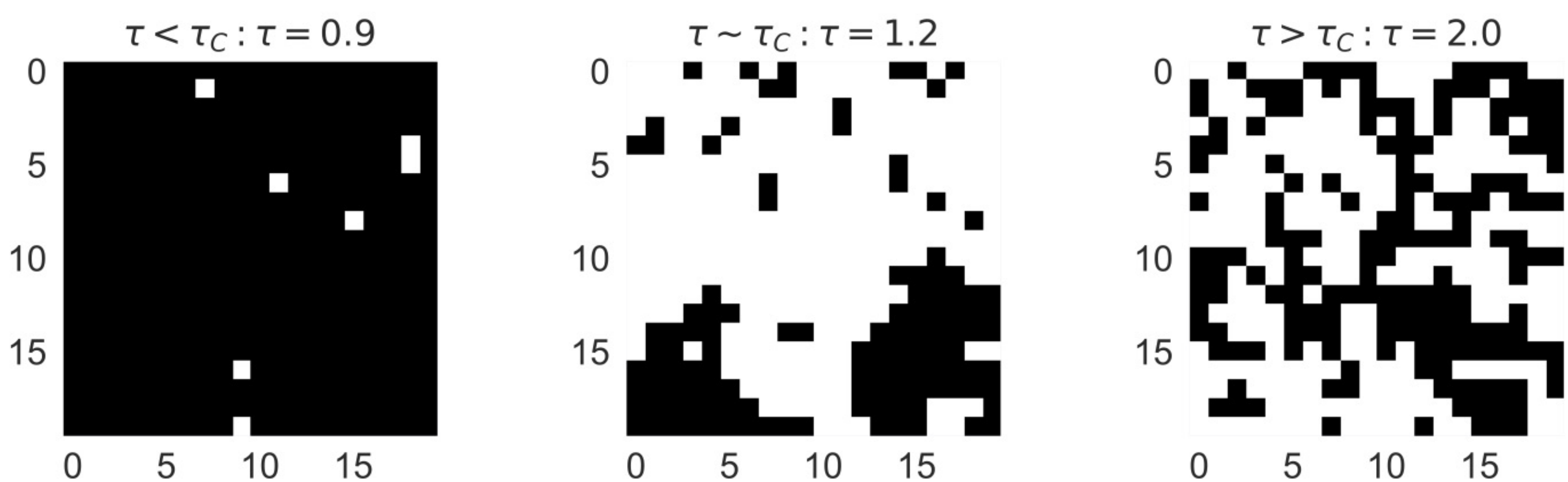
Mean-field MARL: experiments

A model-free method to learning the Ising model (atomic spins that can be in one of two states)





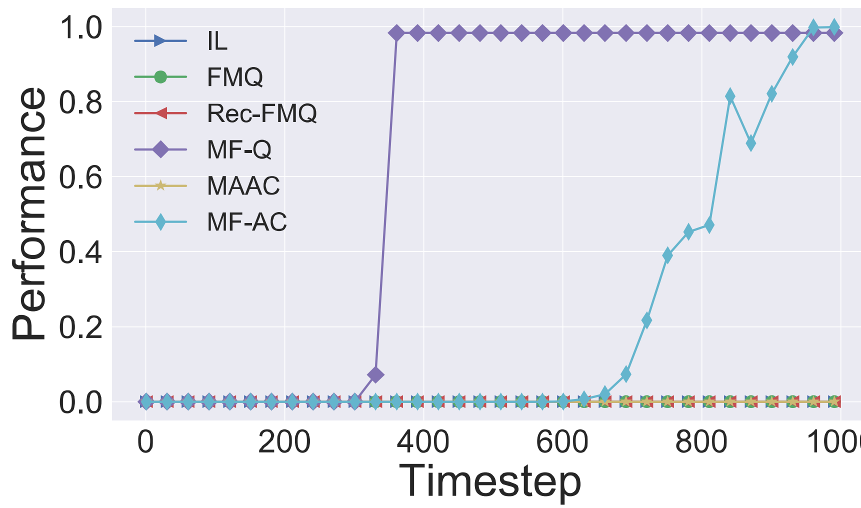
phase transitions



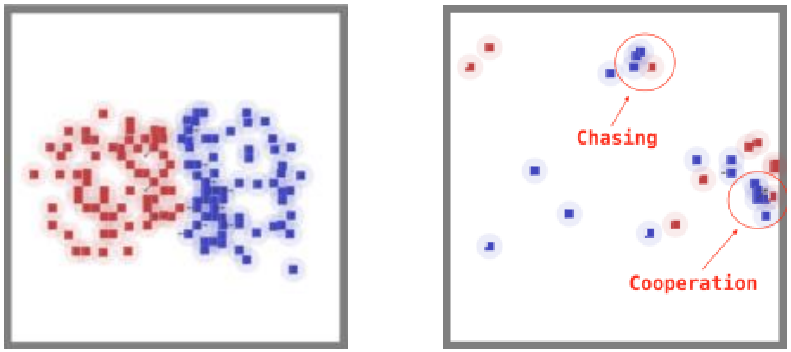
Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J. Mean Field Multi-Agent Reinforcement Learning. arXiv preprint arXiv:1802.05438. 2018 Feb 15.

Mean-field MARL: experiments

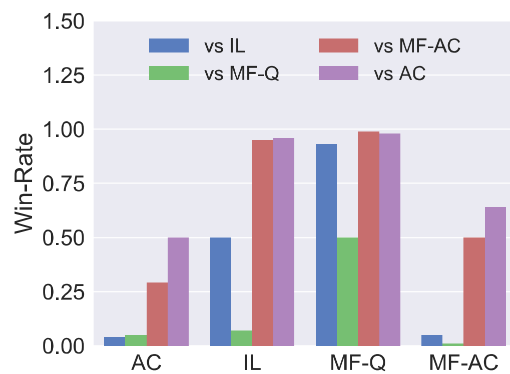
- The Gaussian Squeeze Domain (GSD):
  - each agent contributes to a system objective
- Battle games:



1000 agents



(a) two groups agents (b) chasing and cooperation



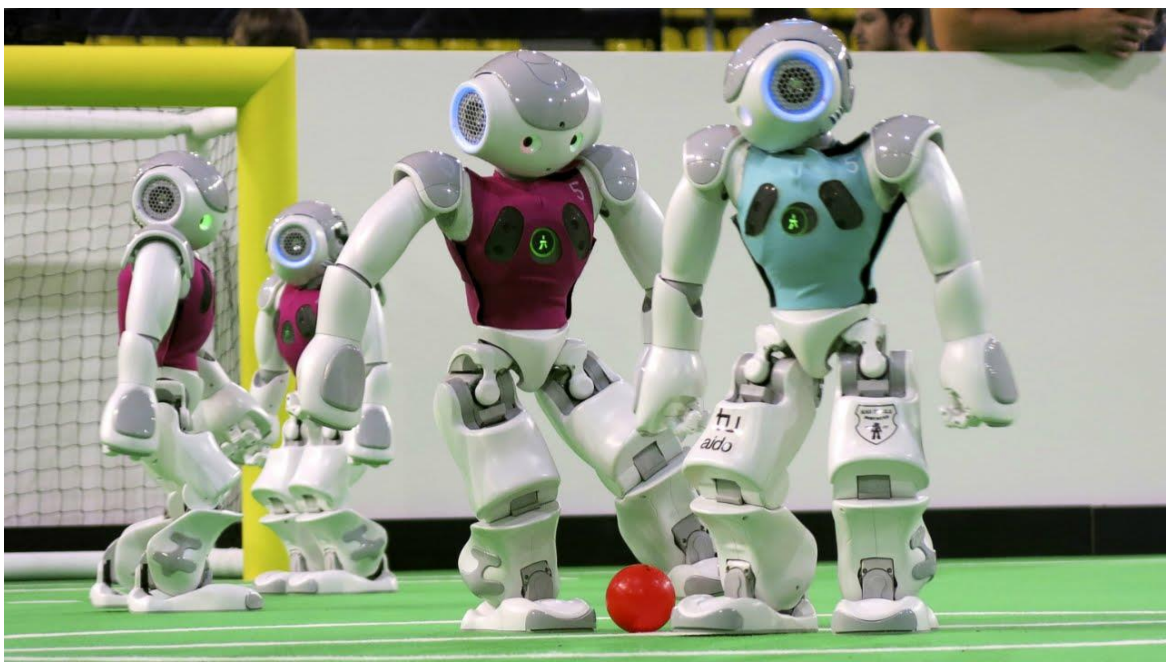
Yang Y, Luo R, Li M, Zhou M, Zhang W, Wang J. Mean Field Multi-Agent Reinforcement Learning. arXiv preprint arXiv:1802.05438. 2018 Feb 15.

Tables of Content

- Many agents
  - Population dynamics
  - CLEAN rewards
  - Mean-field MARL
- Multi-agent Communications
  - CommNet
  - DIAL
  - BiCNet

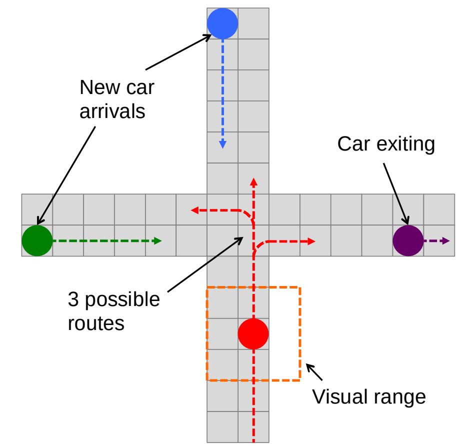
Communications among agents

- AI require the collaboration of multiple agents
- the communication between agents is vital to coordinate the behaviour of each individual

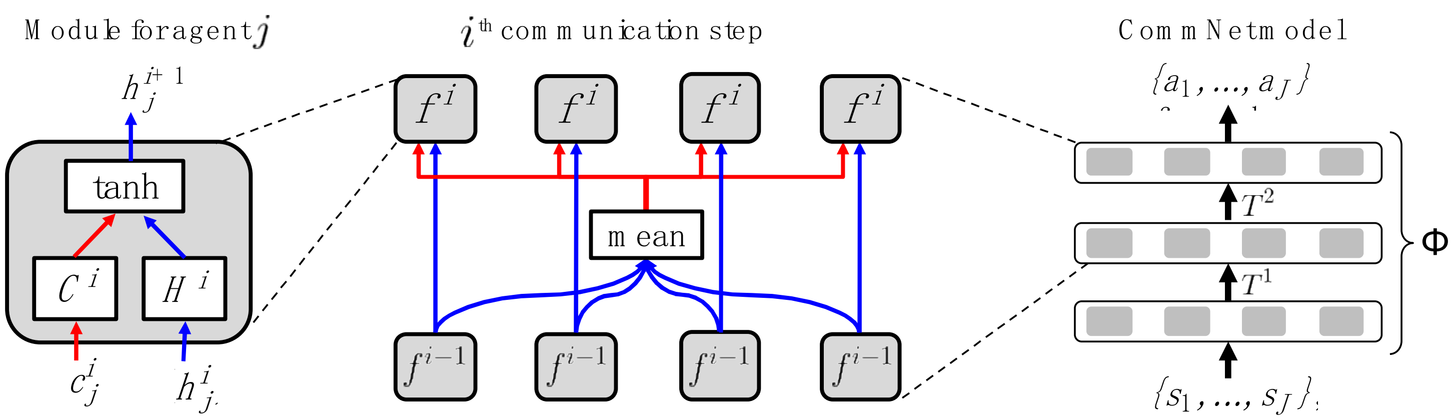


CommNets

- Full cooperation between agents
- The model consists of multiple agents and the communication between them is learned alongside their policy.



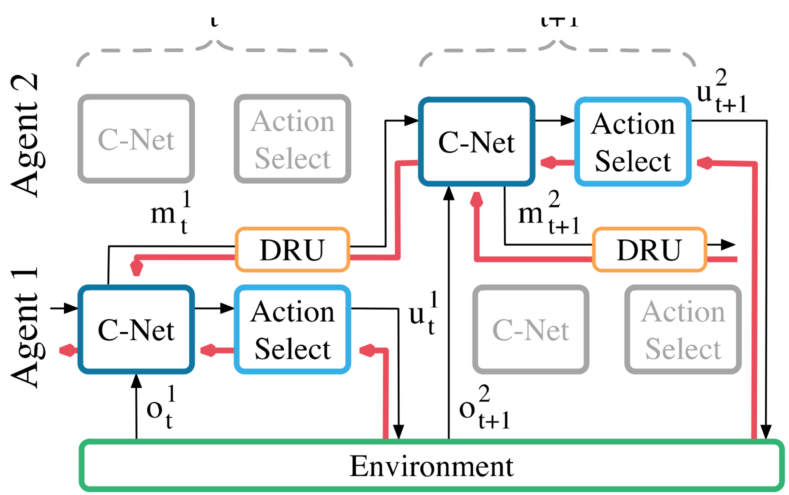
pass the through the junction without colliding

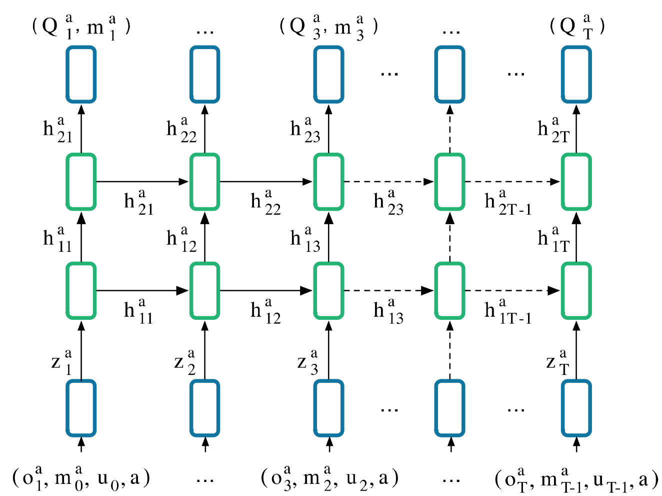


Sukhbaatar, Sainbayar, and Rob Fergus. "Learning multiagent communication with backpropagation." *Advances in Neural Information Processing Systems*. 2016.

Differentiable Inter-Agent Learning (DIAL)

- Uses centralised learning but decentralised execution
  - during learning, agents can backpropagate error derivatives through (noisy) communication channels



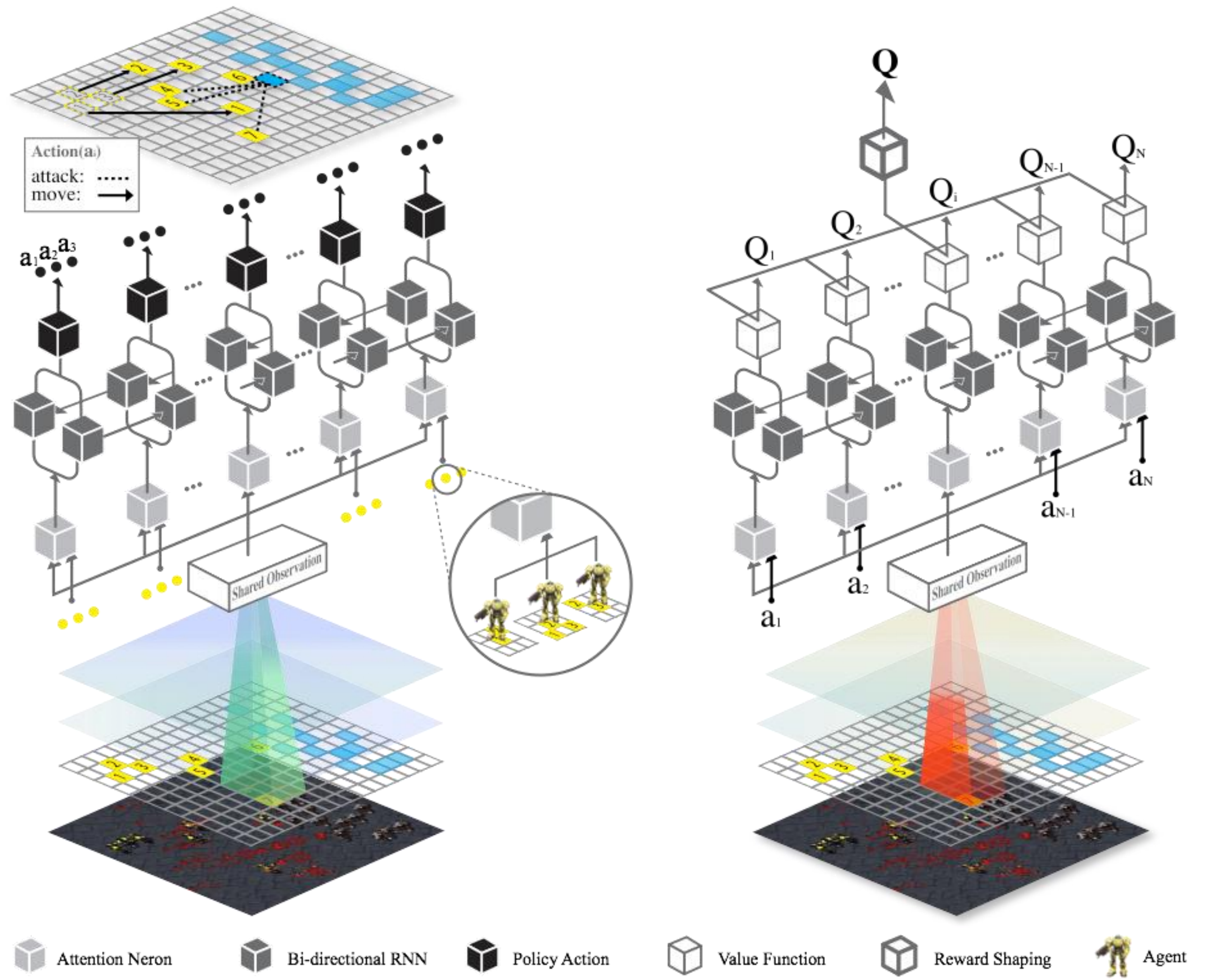


Foerster J, Assael IA, de Freitas N, Whiteson S. Learning to communicate with deep multi-agent reinforcement learning. InAdvances in Neural Information Processing Systems 2016 (pp. 2137-2145).

AI plays StarCraft

- One of the most difficult games for computers
- At least 10^{1685} possible states (for reference, the game of Go has about 10^{170} states)!
- Multiagent reinforcement learning: how large-scale multiple AI agents could learn human-level collaborations, or competitions, from their experiences?

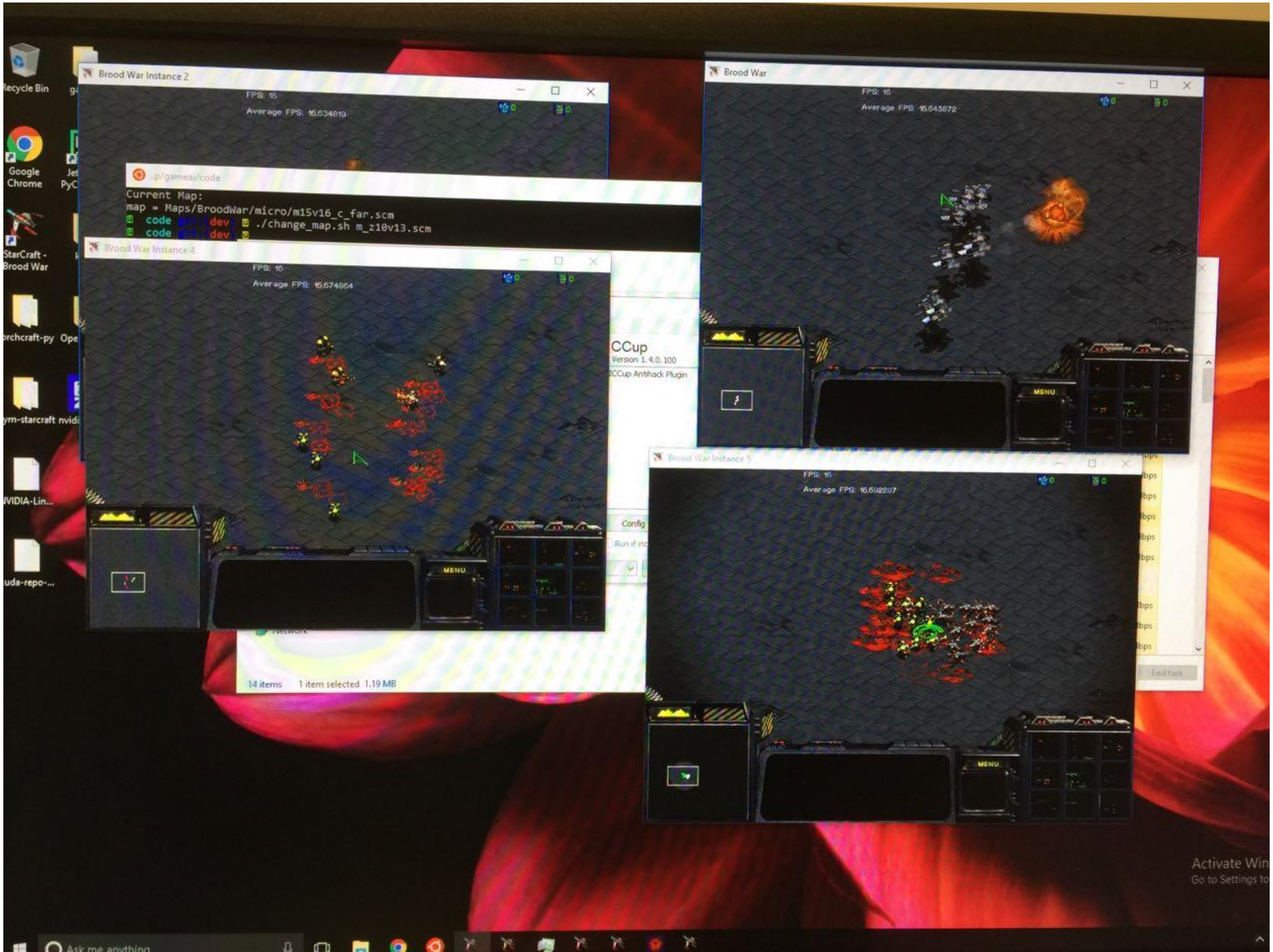
Bidirectional-Coordinated nets (BiCNet)



(a) Multiagent policy networks with grouping (b) Multiagent Q networks with reward shaping

Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, Jun Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games, 2017

Unsupervised training without human demonstration and labelled data



Coordinated moves without collision

Combat 3 Marines (ours) vs. 1 Super Zergling (enemy)

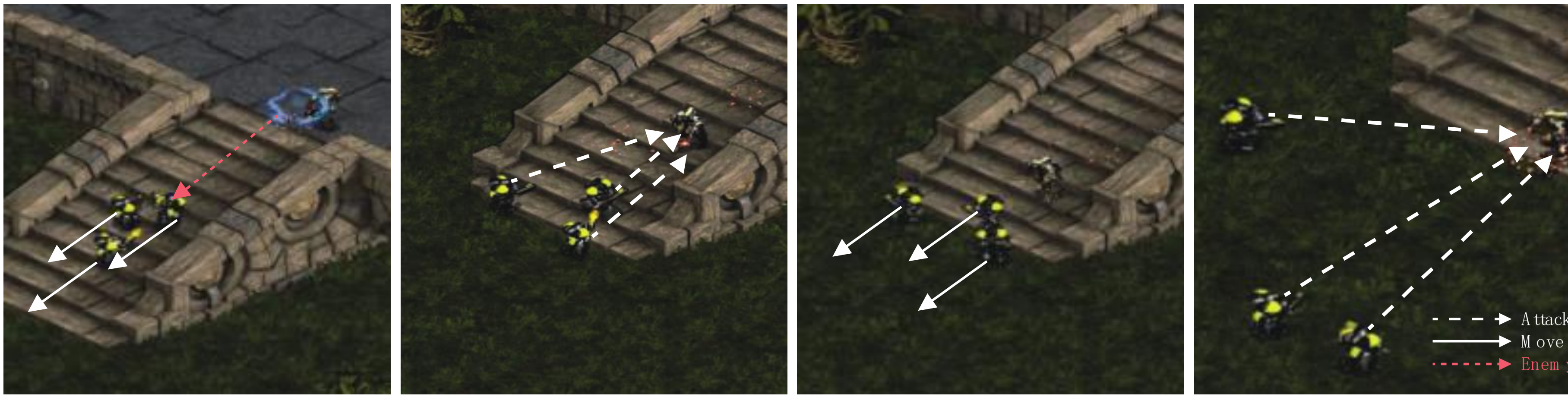
(a) Early stage of training (b) Early stage of training (c) Well-trained (d) Well-trained

- The first two (a) and (b) illustrate that the collision happens when the agents are close by during the early stage of the training;
- the last two (c) and (d) illustrate coordinated moves over the well-trained agents

Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, Jun Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games, 2017

"Hit and Run" tactics

combat 3 Marines (ours) vs. 1 Zealot (enemy)



(a) time step 1: run when attacked

(b) time step 2: fight back when safe

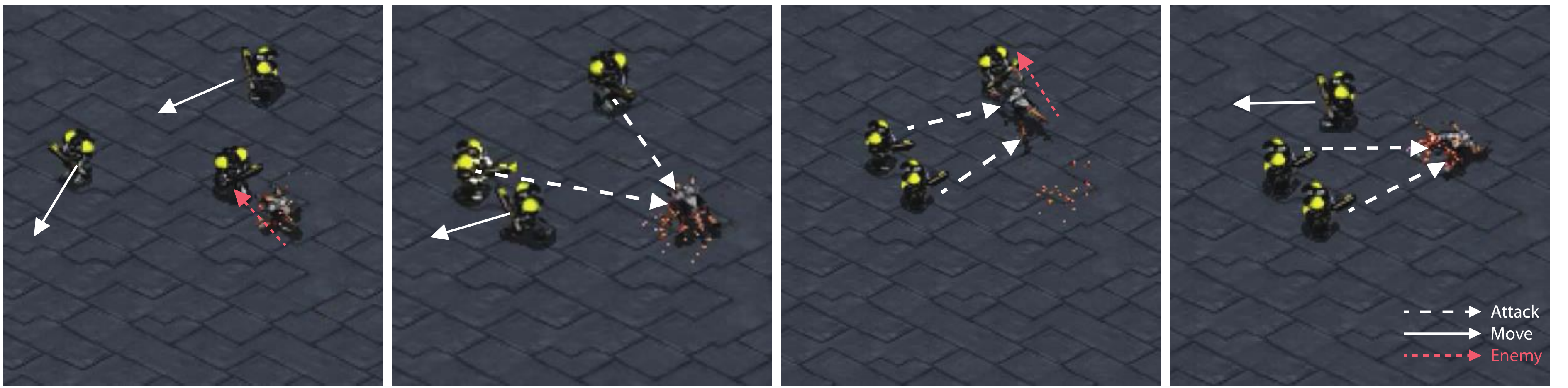
(c) time step 3: run again

(d) time step 4: fight back again

Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, Jun Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games, 2017

Coordinated moves without collision

Combat 3 Marines (ours) vs. 1 Zergling (enemy)

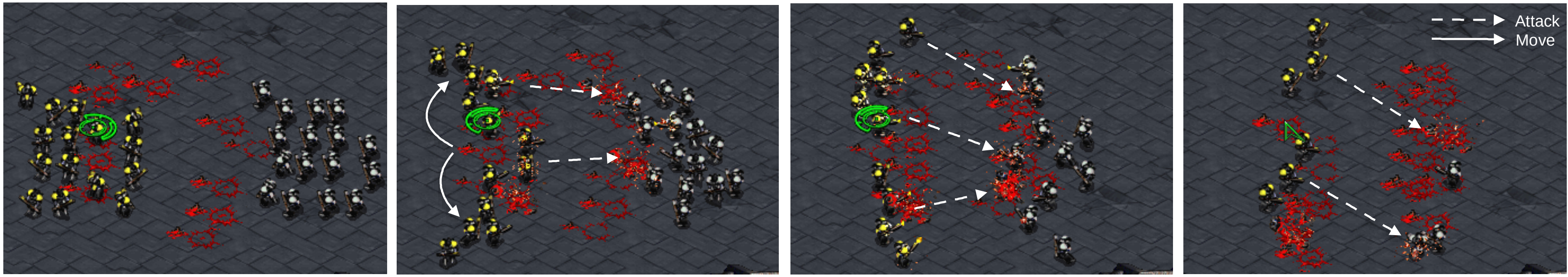


(a) time step 1 (b) time step 2 (c) time step 3 (d) time step 4

Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, Jun Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games, 2017

Focus fire

combat 15 Marines (ours) vs. 16 Marines (enemy)

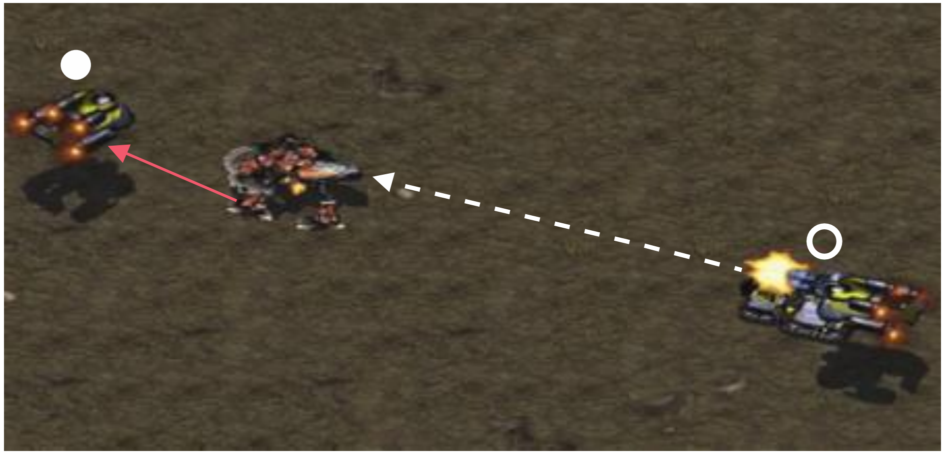


(a) time step 1 (b) time step 2 (c) time step 3 (d) time step 4

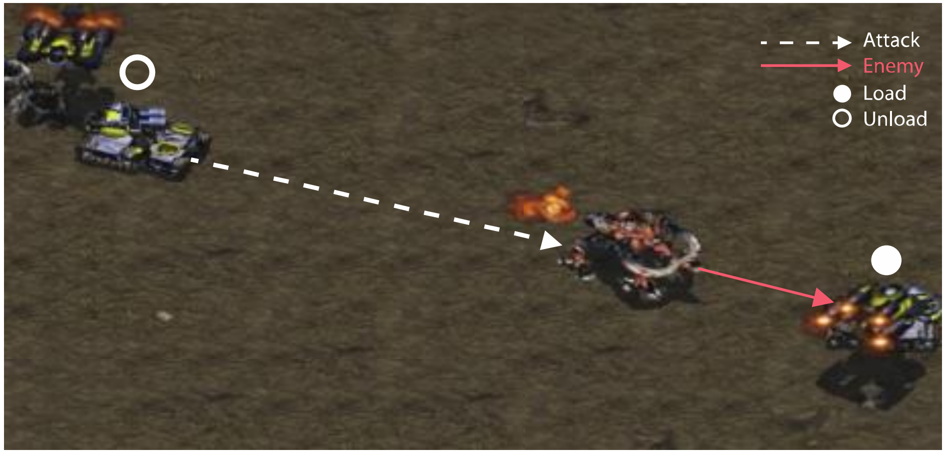
Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, Jun Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games, 2017

Coordinated heterogeneous agents

combat 2 Dropships and 2 tanks vs. 1 Ultralisk



(a) time step 1



(b) time step 2

Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, Jun Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games, 2017

References

- Peng Peng, Quan Yuan, Ying Wen, Yaodong Yang, Zhenkun Tang, Haitao Long, Jun Wang, Multiagent Bidirectionally-Coordinated Nets for Learning to Play StarCraft Combat Games, 2017
- Yaodong Yang , Lantao Yu , Yiwei Bai , Jun Wang , Weinan Zhang , Ying Wen , Yong Yu, , Dynamics of Artificial Populations by Million-agent Reinforcement Learning, 2017
- C. Holmesparker, M. E. Taylor, A. K. Agogino, and K. Tumer. Clean rewards to improve coordination by removing exploratory action noise. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03, pages 127–134, 2014.
- Malialis, Kleanthis, et al. "Feature Selection as a Multiagent Coordination Problem." arXiv preprint arXiv:1603.05152(2016).
- Sukhbaatar, Sainbayar, and Rob Fergus. "Learning multiagent communication with backpropagation." *Advances in Neural Information Processing Systems*. 2016.
- Foerster J, Assael IA, de Freitas N, Whiteson S. Learning to communicate with deep multi-agent reinforcement learning. InAdvances in Neural Information Processing Systems 2016 (pp. 2137-2145).