多智能体深度强化学习及（量子）博弈分析

Ying Wen / 温颖
John Hopcroft Center for Computer Science
Shanghai Jiao Tong University
ying.wen@sjtu.edu.cn

21-22/Nov/2020

2020 量子物理与智能计算交叉研讨会

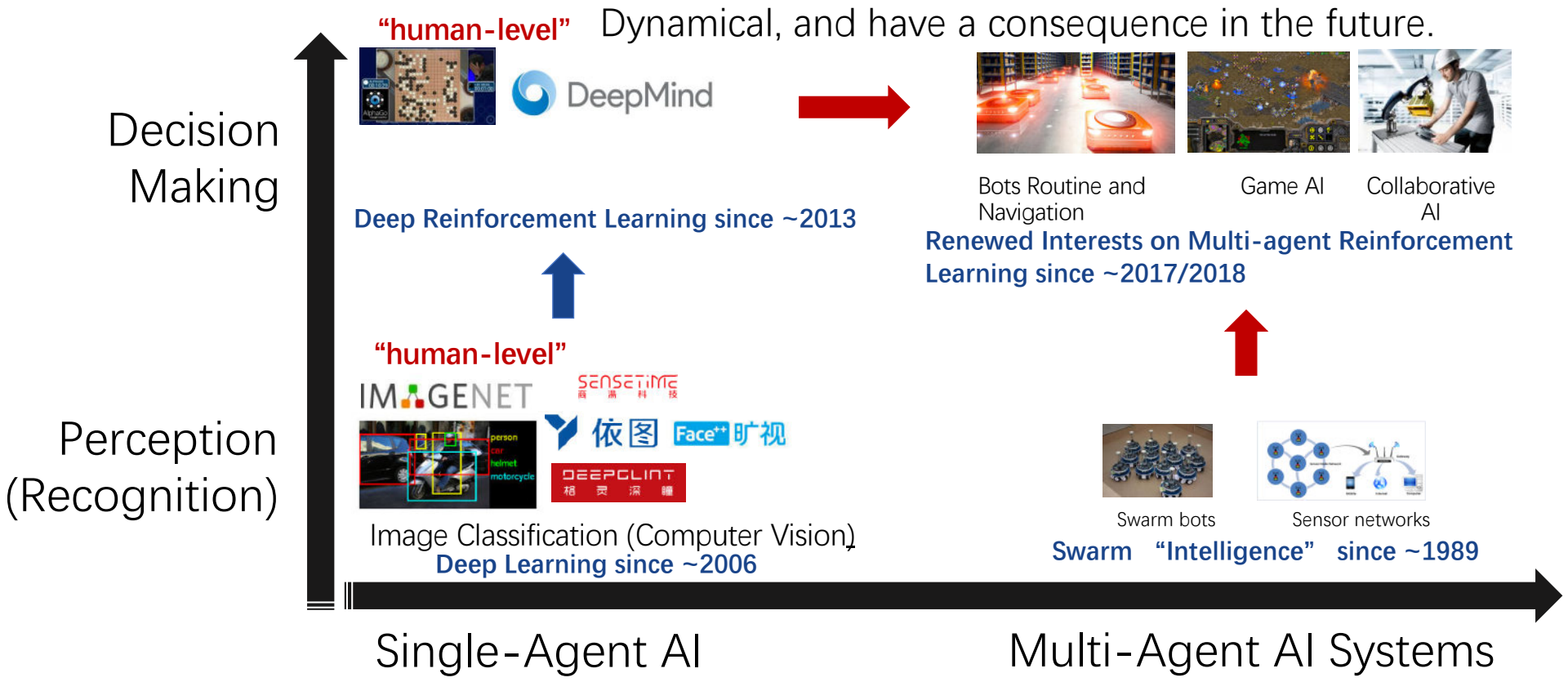
Agenda

1. Introduction
2. Direct Interaction Modelling - Emergent Communication
3. Indirect Interaction Modelling
    a. Agent Modelling
    b. Implicit Interaction Modelling via Centralized Critic
    c. Explicit Policy Interaction Modelling via Policy-Space Analysis
4. Challenges and Perspectives X Quantum Theory

Agenda

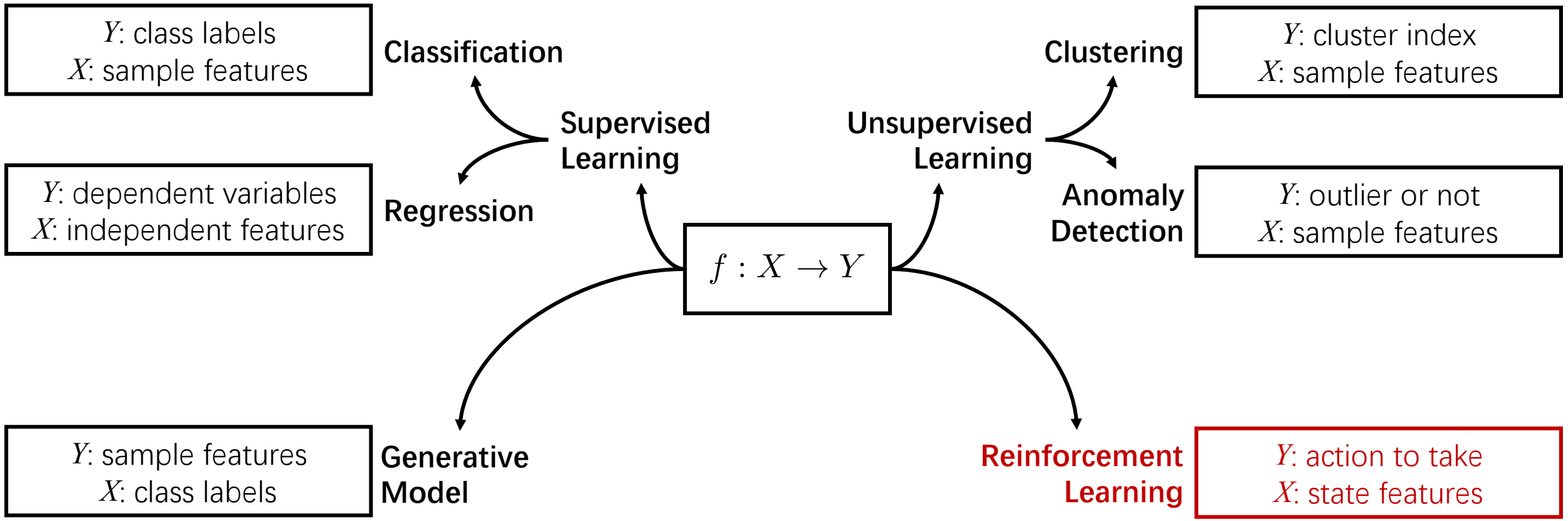
1. Introduction
2. Direct Interaction Modelling - Emergent Communication
3. Indirect Interaction Modelling
    a. Agent Modelling
    b. Implicit Interaction Modelling via Centralized Critic
    c. Explicit Policy Interaction Modelling via Policy-Space Analysis
4. Challenges and Perspectives X Quantum Theory

Some Observations on AI Progress



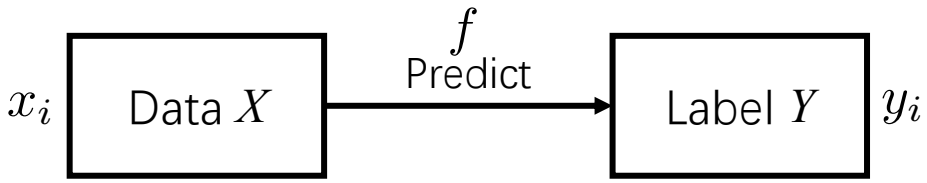
Machine Learning as Function Mapping

- Machine learning tasks are different in the experience available, the objective function, and the specific learning algorithms.



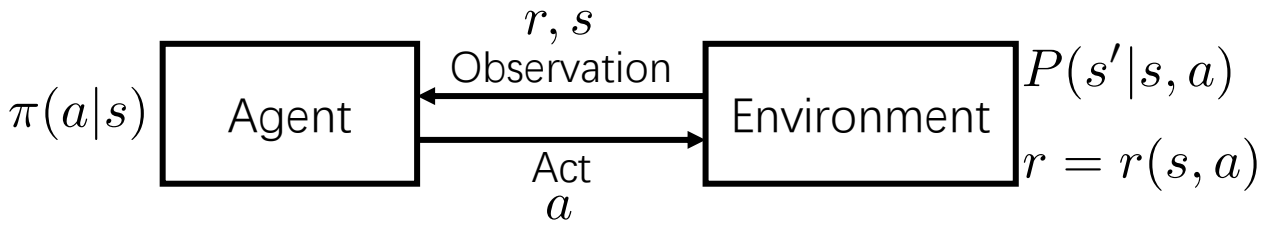
Typical Machine learning

Supervised Learning



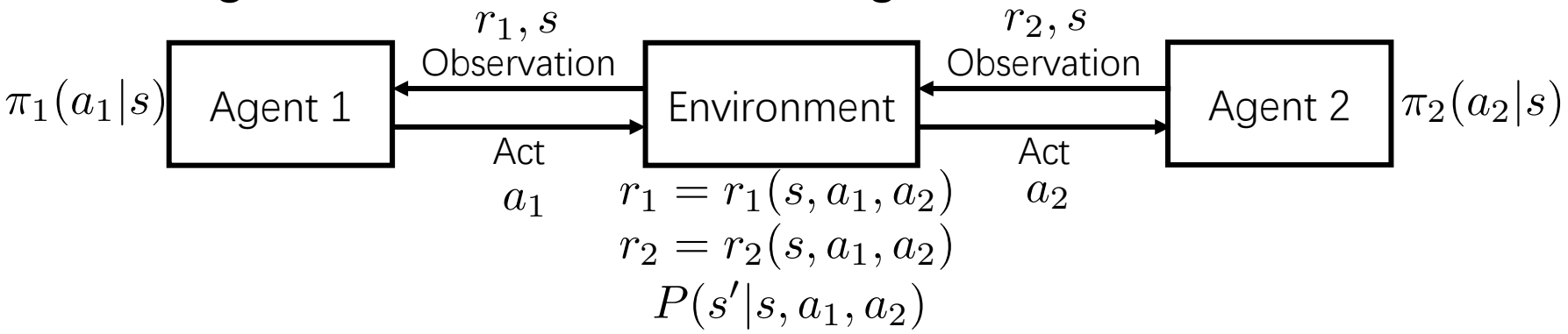
$$\arg\min_{\theta} \sum_{i=1}^{n} \text{Loss}\left(f_\theta\left(x_i\right), y_i\right)$$

Reinforcement Learning



$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=1}^{\infty} \gamma^t r_t\right]$$

Multi-Agent Reinforcement Learning



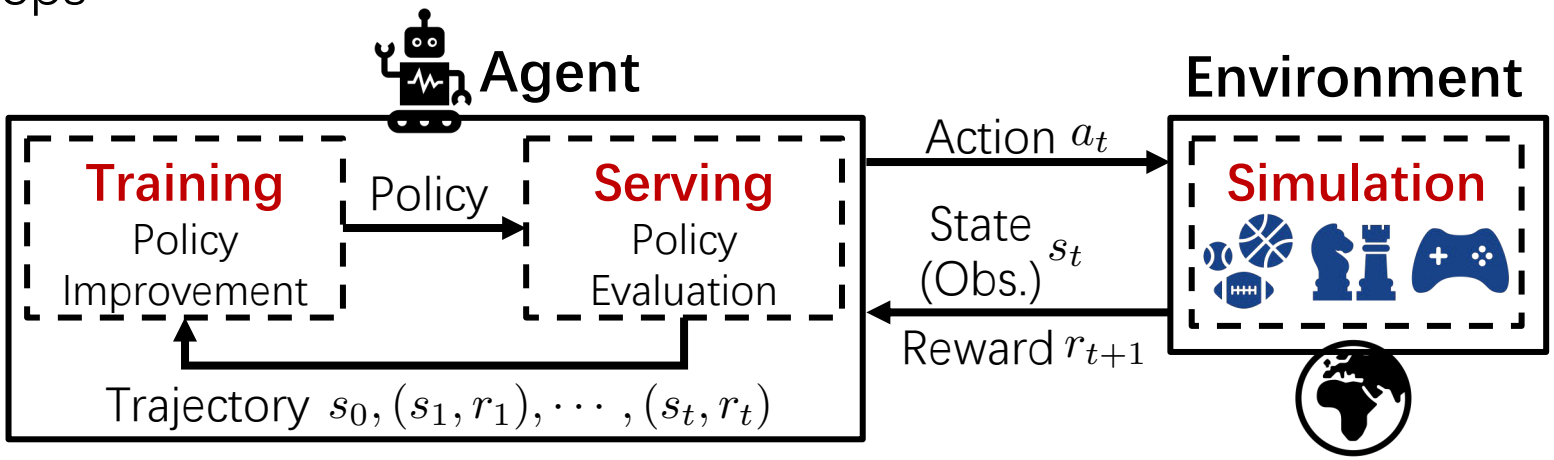
$$\max_{\pi_1} \mathbb{E}_{\pi_1,\pi_2}\left[\sum_{t=1}^{\infty} \gamma^t r_{1,t} \middle| P, \pi_2\right]$$

- Machine learning tasks are different in the experience available, the objective function, and the specific learning algorithms

Reinforcement Learning Settings

Agent interacts at discrete time steps

- ▶Observes state
- ▶Selects action.
- ▶Obtains immediate reward.
- ▶Observes resulting state.



- ▶Learning from interactions.
- ▶Learning how to map observations to actions.
- ▶Aim at maximize long-term discounted return.

Objective: $\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^{\infty} \gamma^t r_t \right]$

- ▶Key Features:
  - Learner is not told which action to take.
  - Delayed reward.
  - Exploration and exploitation.
  - Possibility of delayed reward
  - In between supervised and unsupervised learning

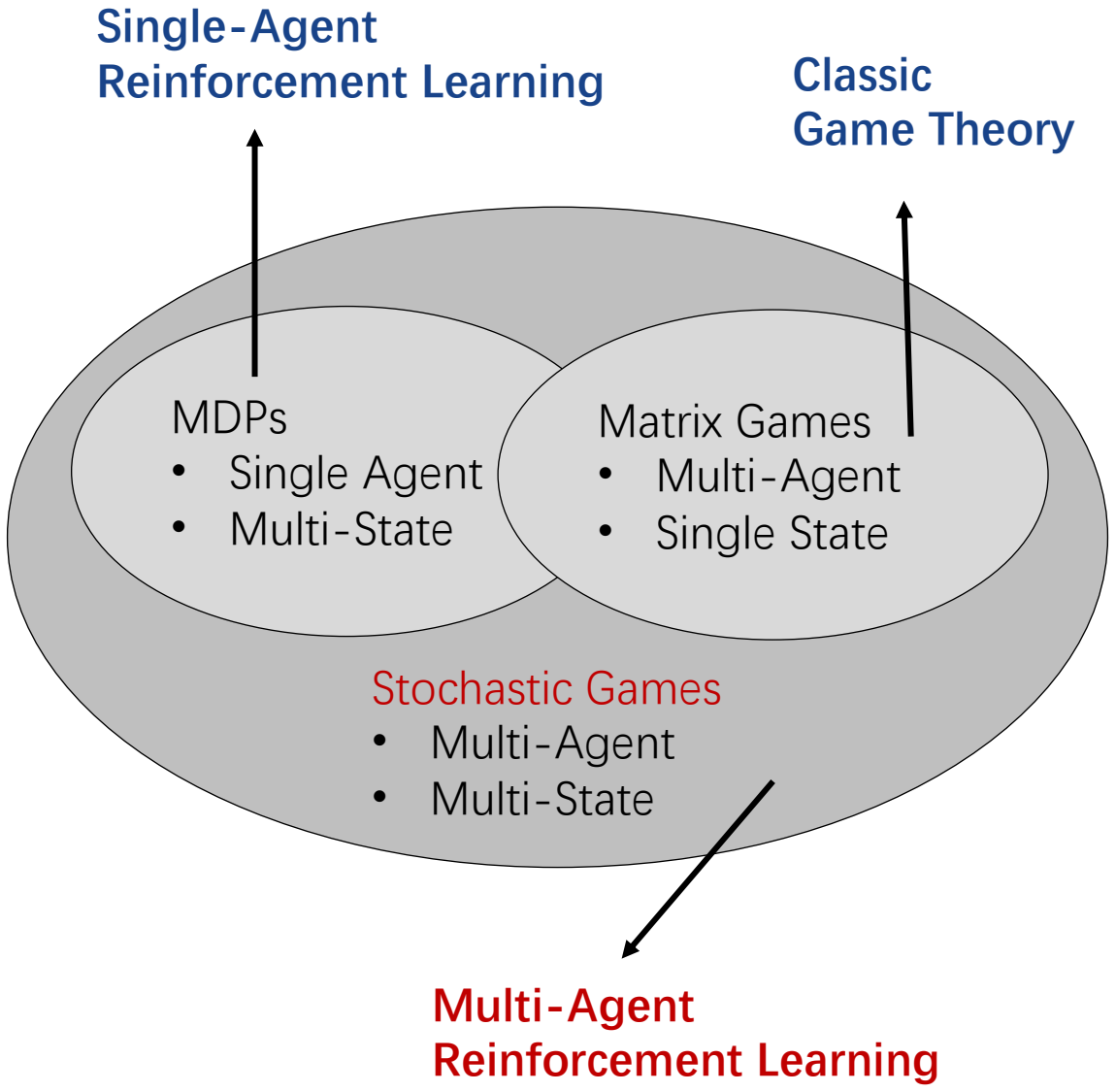
Multi-Agent in the Wild

- ►Games (Poker, Go, Shogi, StarCraft)
- ►Auctions (ad auctions, spectrum auctions)
- ►Online platforms (sharing economy, crowdsourcing)
- ►Cooperative interaction (language, advice, decision support)
- ►Resource allocation (packet routing, server allocation)

Multi-Agent System: **Agents** acting towards its **objectives** while all interacting in a shared **environment**.

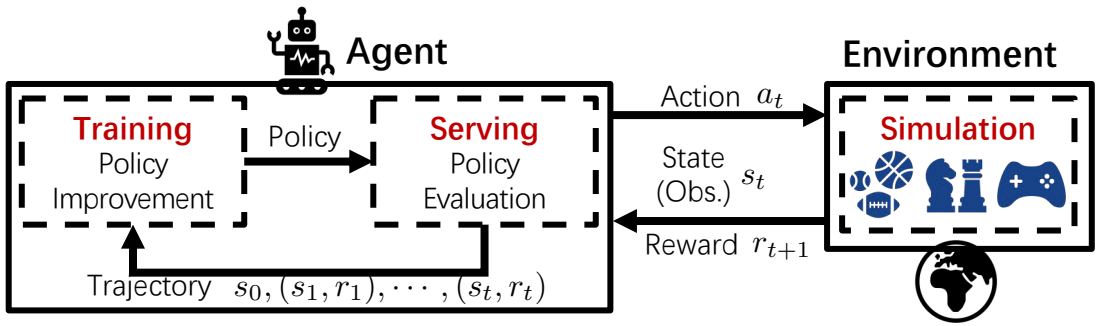
What Are We Studying?

- ►Markov decision processes
  - one decision maker
  - multiple states
- ►Matrix games / Repeated games
  - multiple decision makers
  - one state (e.g., one normal form game)
- ►Stochastic games (Markov games)
  - multiple decision makers
  - multiple states (e.g., multiple normal form games)



From Single-Agent to Multi-Agent RL

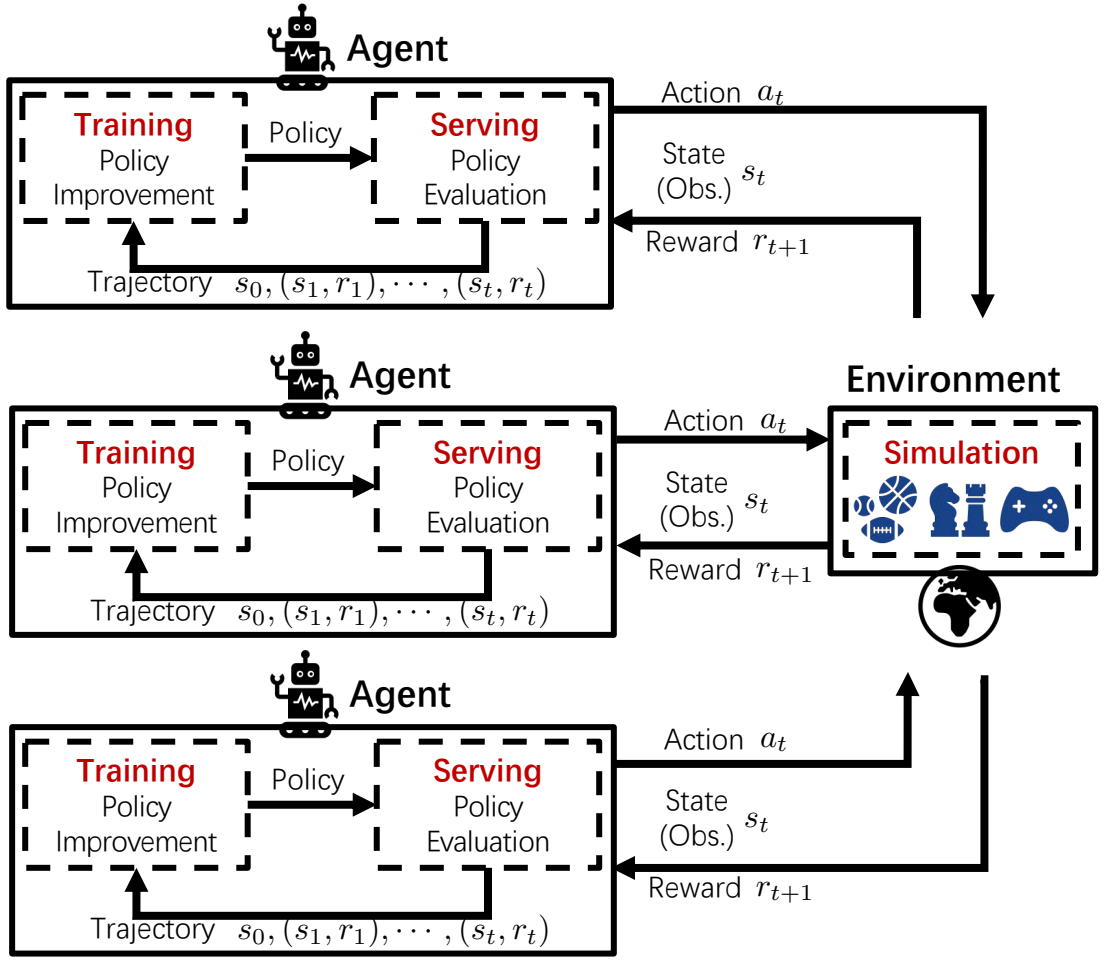
- Combination of state and joint action space: curse of dimensionality => Scalability
- Agents **interact** with the environment and each other
  - => Multi-objectives, Game Theory, Communication
  - Learning is simultaneous => Non-stationary



$$s' \sim P(s'|s, a)$$

Single-Agent

Usually formalized as Stochastic Game (SG), where $s' \sim P(s'|s, a_1, a_2, a_3 \ldots)$.

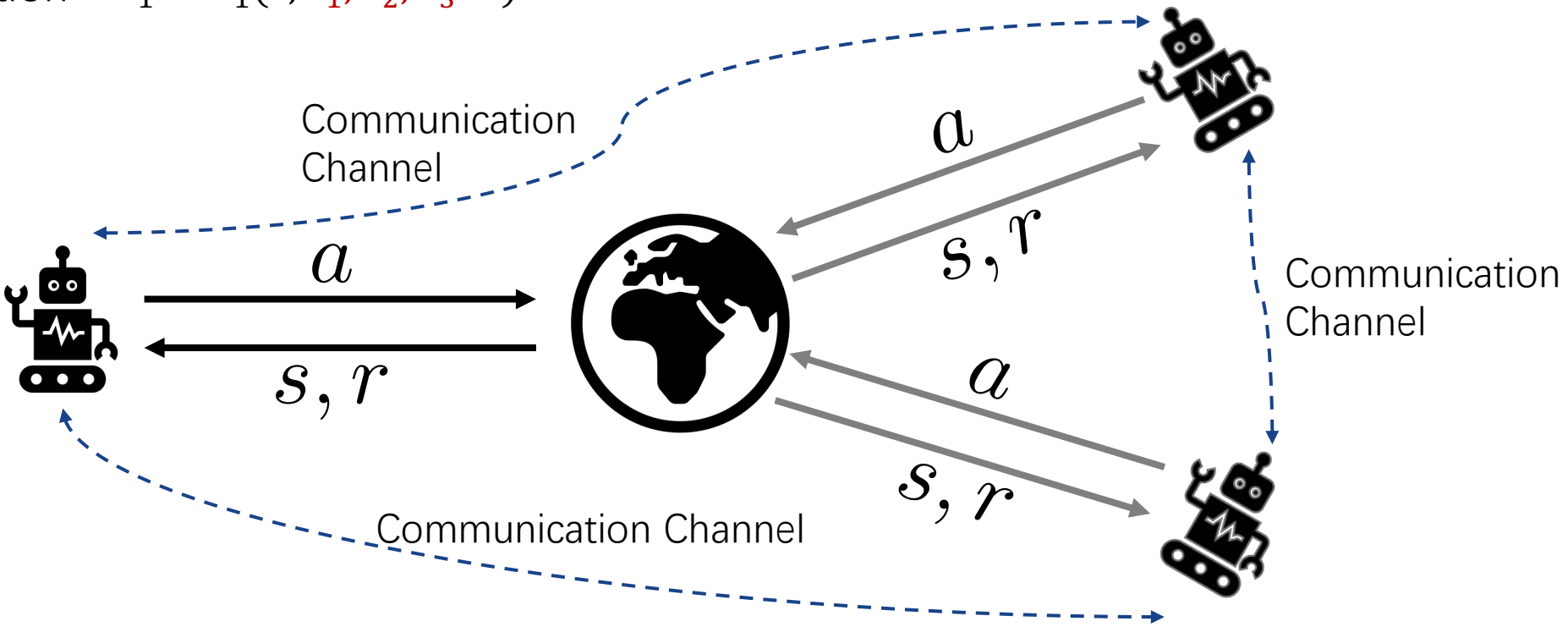


Multi-Agent

Direct and Indirect Interactions

► Direct Interaction
  - Communication

► Indirect Interaction via Environment
  - State joint action transition $s' \sim P(s'|s, a_1, a_2, a_3 \ldots)$
  - Reward function $r_1 = R_1(s, a_1, a_2, a_3 \ldots)$

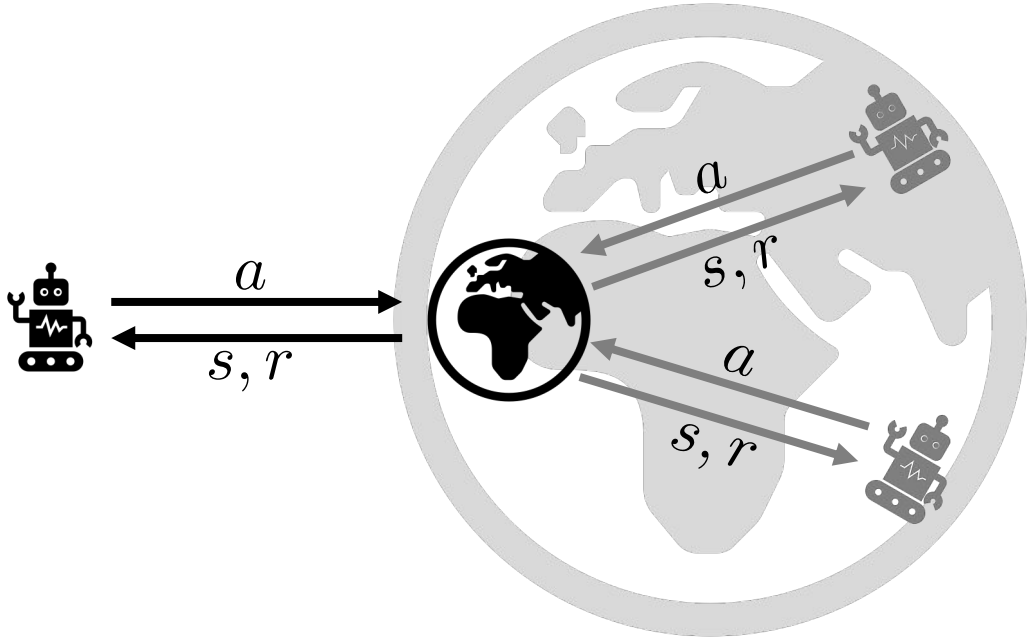


Interactions in a multi-agent system

Interaction Example: Agents in Traffic

Naïve Solution: Independent Learner

- ►Naive extension to multi-agent setting.
- ►Agents mutually ignore each other.
- ►Treat the interactions with other agents as noise.
- ►Downgrade to a **MDP**.



Why Multi-Agent Learning is Hard?

The agents are in a shared environment, they are mutually affected during the interactions, cannot only optimize an agent independently.

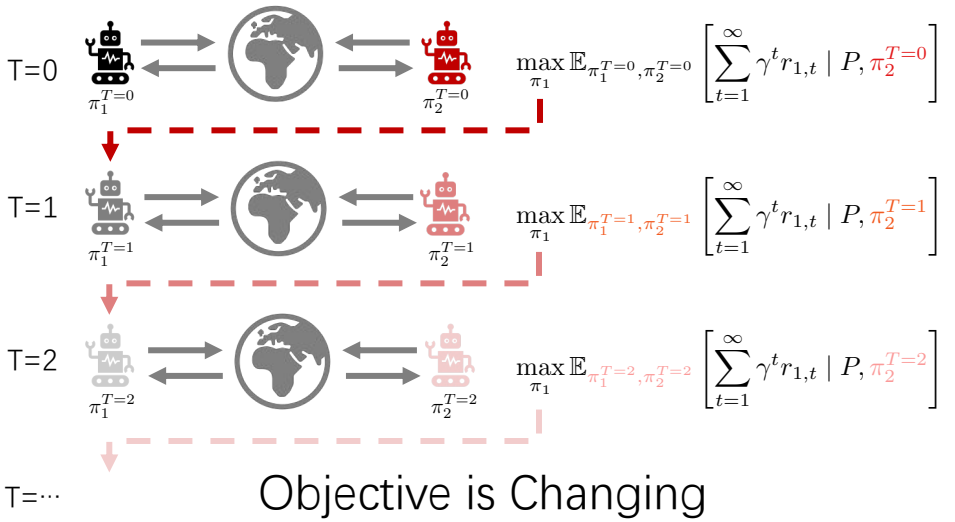
- The system is a moving target: during the learning, agents' policies are changing simultaneously.
- Need for coordination/compromise: agents behaviors have impacts on other agents (like merge).

Can we take the mutual influence or interactions into account to shape the learning?

The other agents are learning and updating its policy over time: $P^t(a_{-i}|s) \neq P^{t+\delta t}(a_{-i}|s)$

Therefore, for a single naive agent i, the environment it perceives is non-stationary:

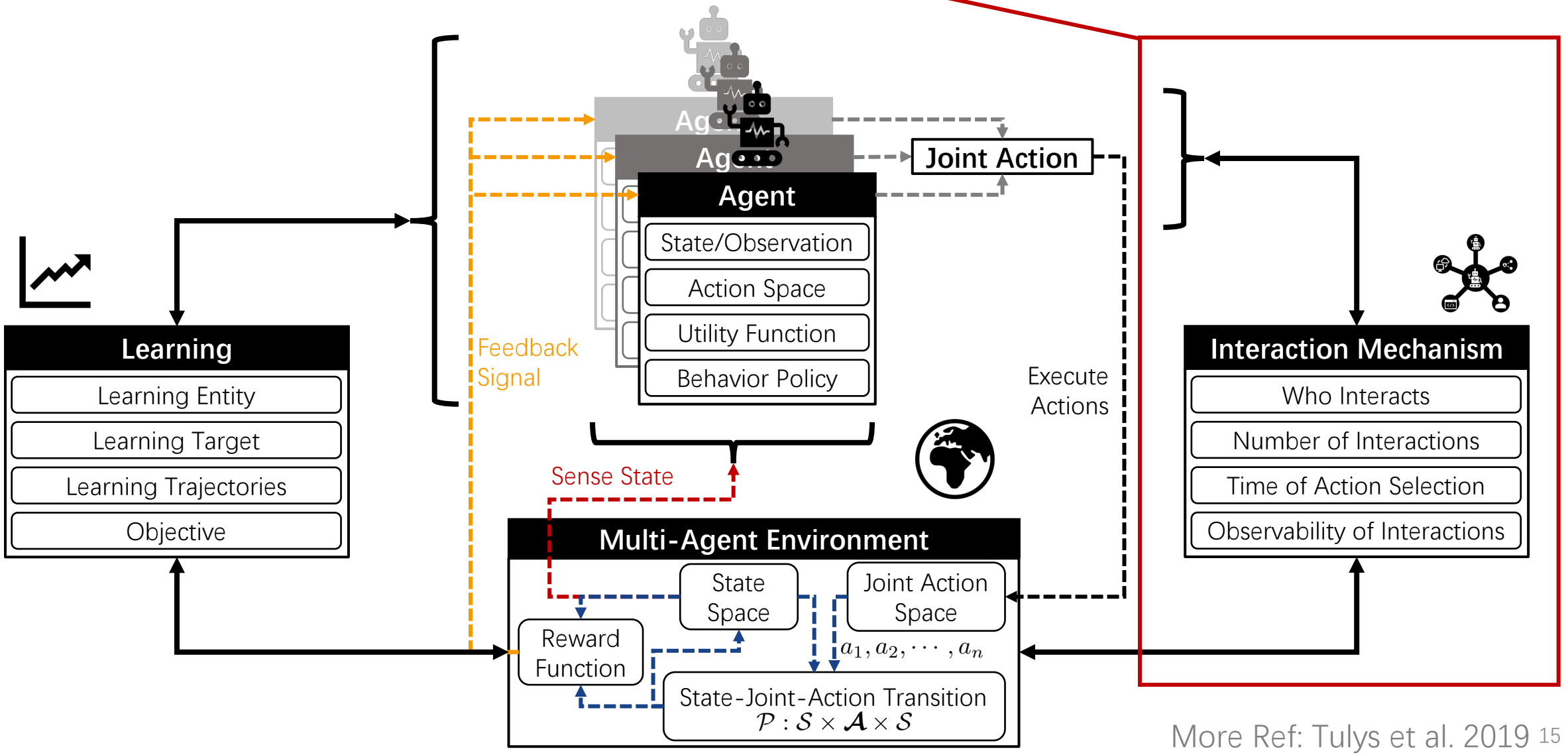
$$
\begin{aligned}
P_t(s'|s,a_i) &= \sum_{a_{-i}} P(s'|s,a_i,a_{-i}) P_t(a_{-i}|s) \\
&\neq \sum_{a_{-i}} P(s'|s,a_i,a_{-i}) P^{t+\delta t}(a_{-i}|s) \\
&= P^{t+\delta t}(s'|s,a_i).
\end{aligned}
$$



Objective is Changing

Why Multi-Agent Learning is Hard?

►How to manage additional interactions in MARL?



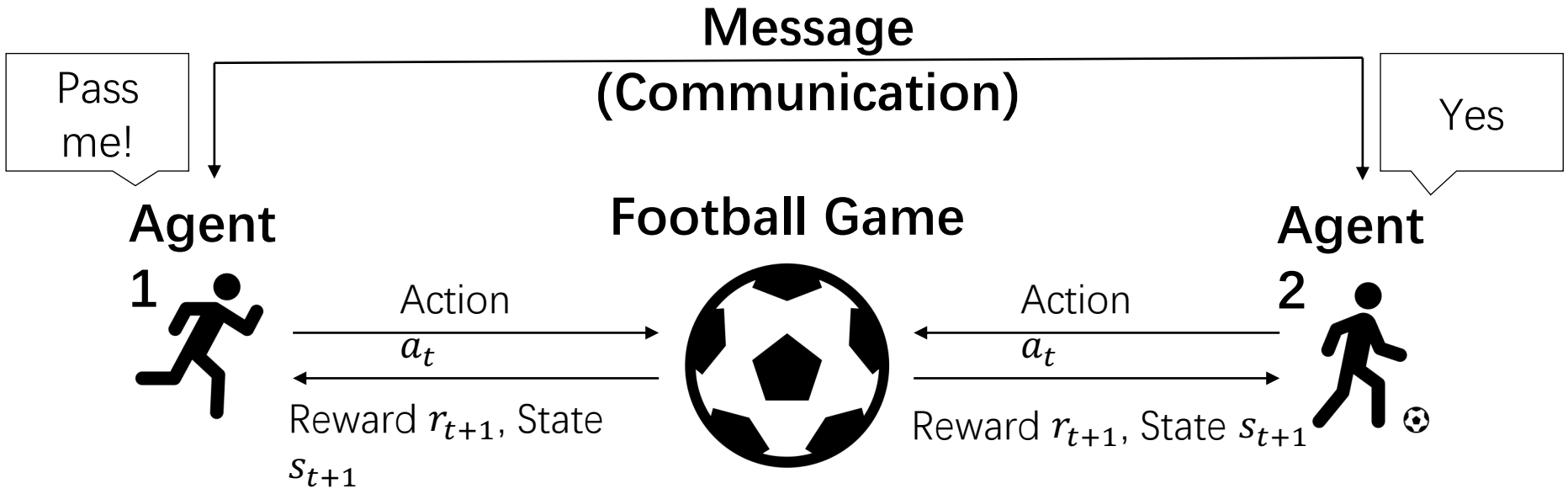
More Ref: Tulys et al. 2019

Agenda

1. Introduction
2. Direct Interaction Modelling - Emergent Communication
3. Indirect Interaction Modelling
    a. Agent Modelling
    b. Implicit Interaction Modelling via Centralized Critic
    c. Explicit Policy Interaction Modelling via Policy-Space Analysis
4. Challenges and Perspectives X Quantum Theory

Direct Interaction Modelling – Emergent Communication

- AI requires the collaboration of multiple agents
- The communication between agents is vital to coordinate the behavior of each individual



Cooperate with Communication

Bidirectional-Coordinated Nets (BiCNet)

- ►Bi-directional recurrent networks
  - Means of communication
  - Connect each individual agent's policy and and Q networks.
- ►Multi-agent deterministic actor-critic.

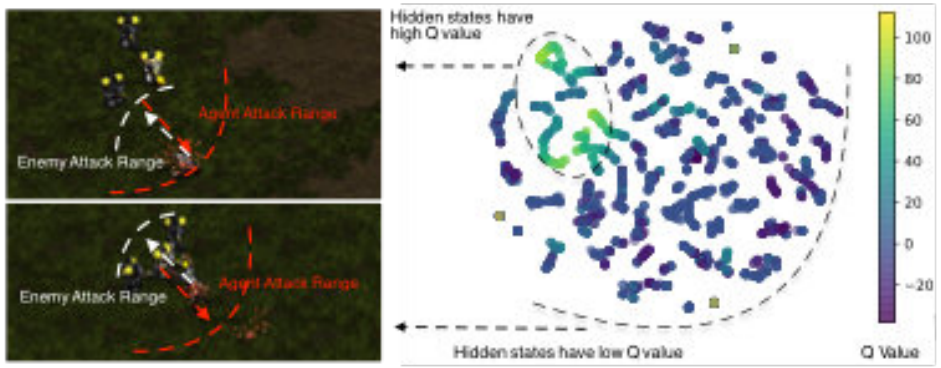
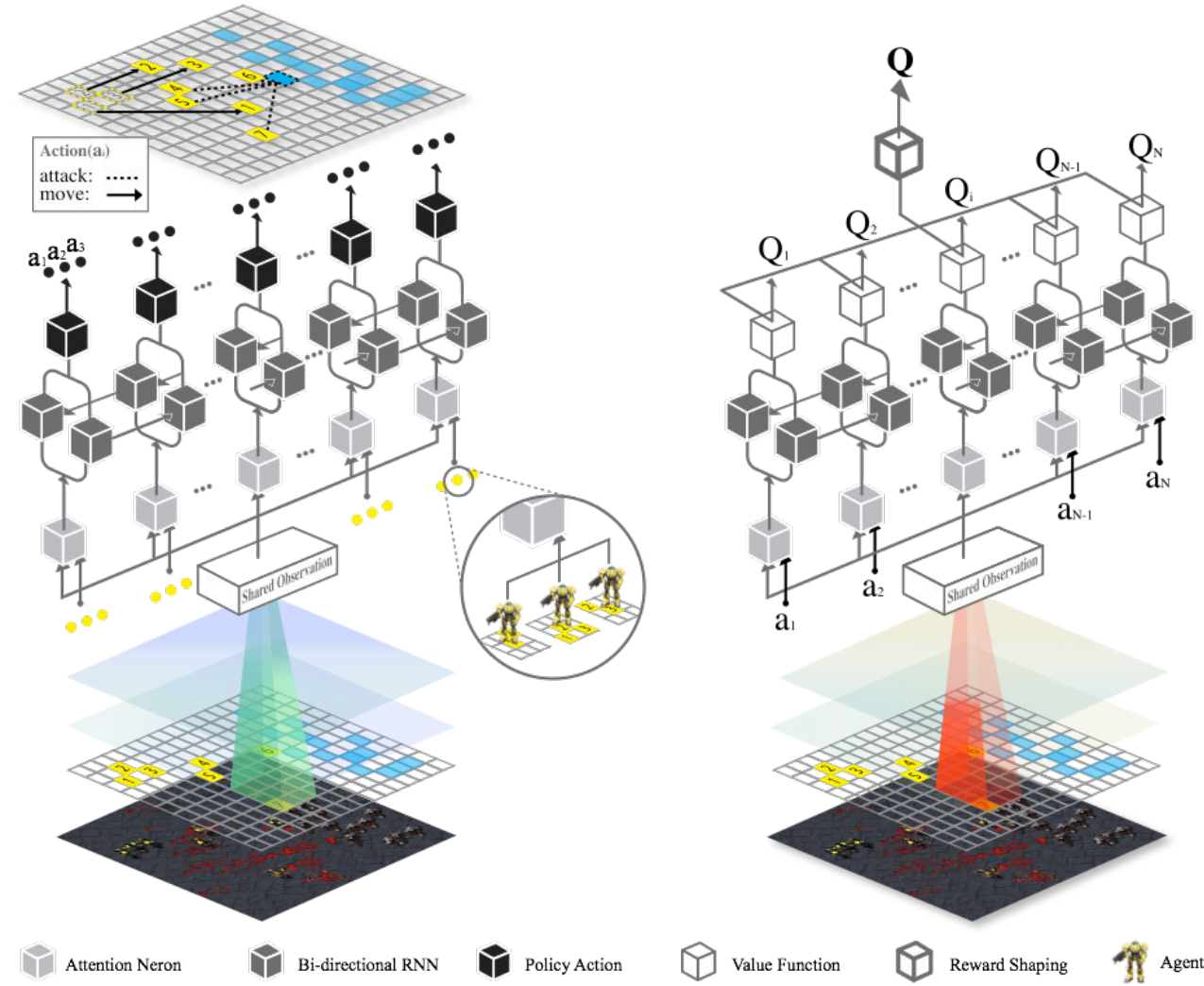


Figure 4: Visualisation for 3 Marines vs. 1 Super Zergling combat. **Upper Left**: State with high Q value; **Lower Left**: State with low Q value; **Right**: Visualisation of hidden layer outputs for each step using TSNE, coloured by Q values.

High Q-value steps are aggregated in the same area.



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

[Peng, Peng*, and Wen, Ying*. 2017.]

Agenda

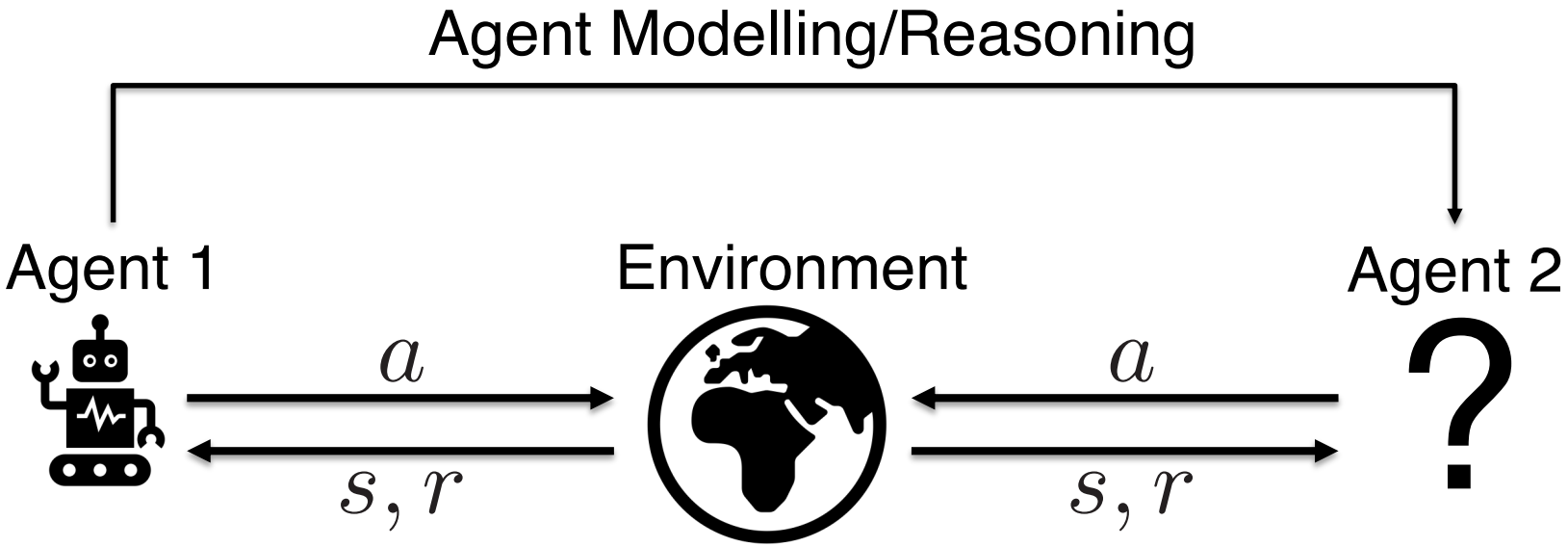
1. Introduction
2. Direct Interaction Modelling - Emergent Communication
3. Indirect Interaction Modelling
    a. **Agent Modelling**
    b. Implicit Interaction Modelling via Centralized Critic
    c. Explicit Policy Interaction Modelling via Policy-Space Analysis
4. Challenges and Perspectives X Quantum Theory

Indirect Interaction Modelling - Agent Modelling

Act based on the other agents' movement?

- But the problem is the other agents wouldn't do everything as you wish!
- Besides, the agent i's objective $\eta_i(\pi) = \mathbb{E}_{a_i, a_{-i} \sim \pi_i, \pi_{-i}}[r_i]$ doesn't explicitly include the other agents's policies:

Build a model for the other agents and include it in the learning objective!



Opponent-Aware Models

- Maintain a belief over environment state and the other agents' models (e.g., learning algorithms, observation, actions, their beliefs over other agents, etc.)

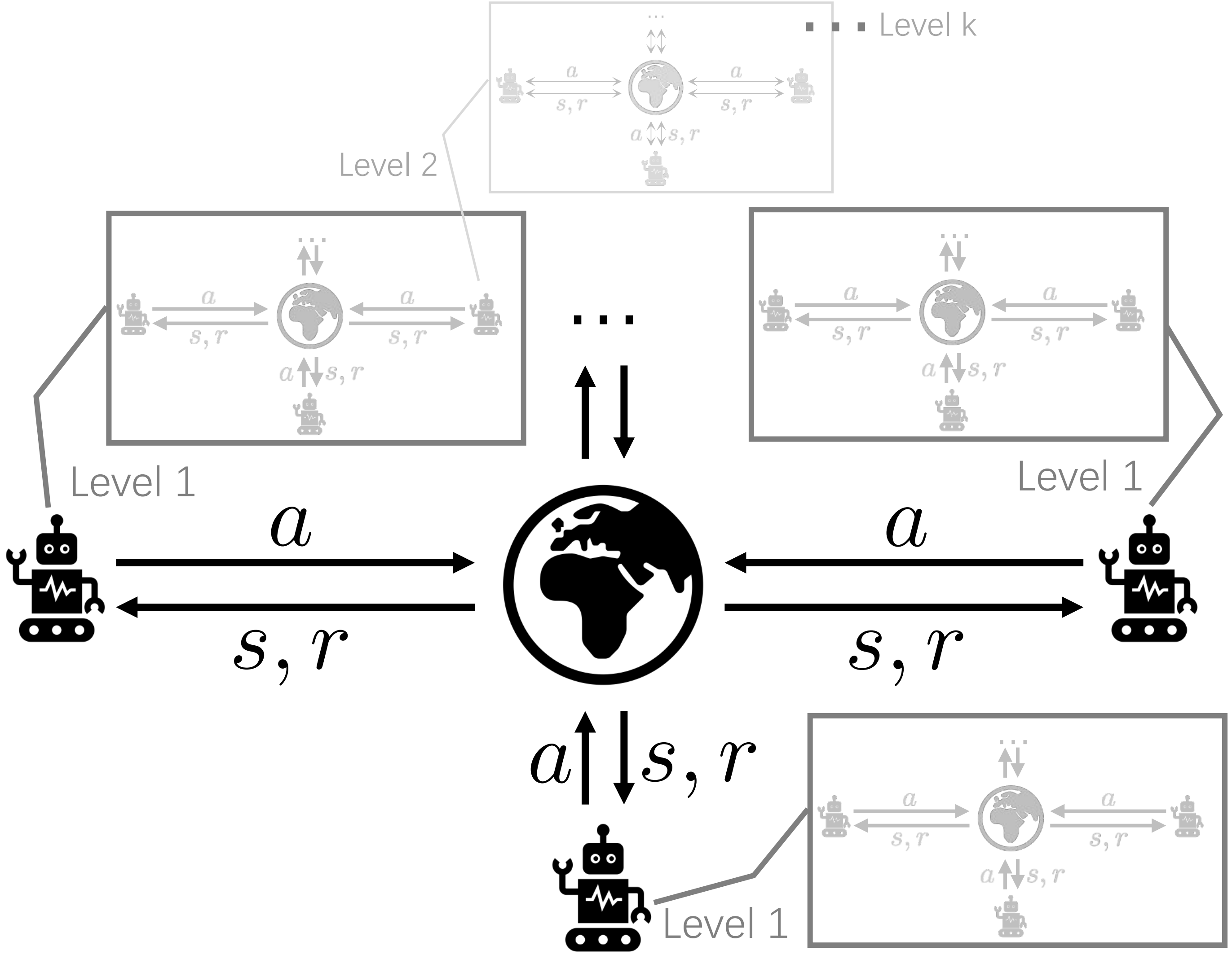
Observed interaction history (past actions, states, ...)

↓

Agent / Opponent Model

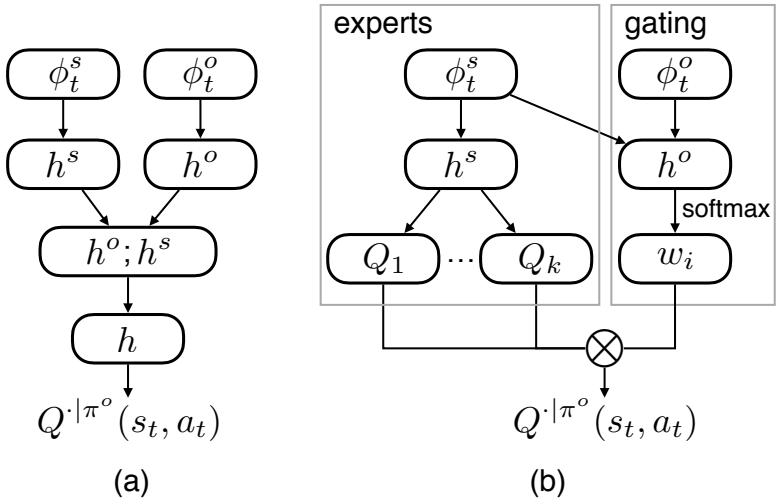
↓

Predicted property of interest (actions, class, goal, ...)



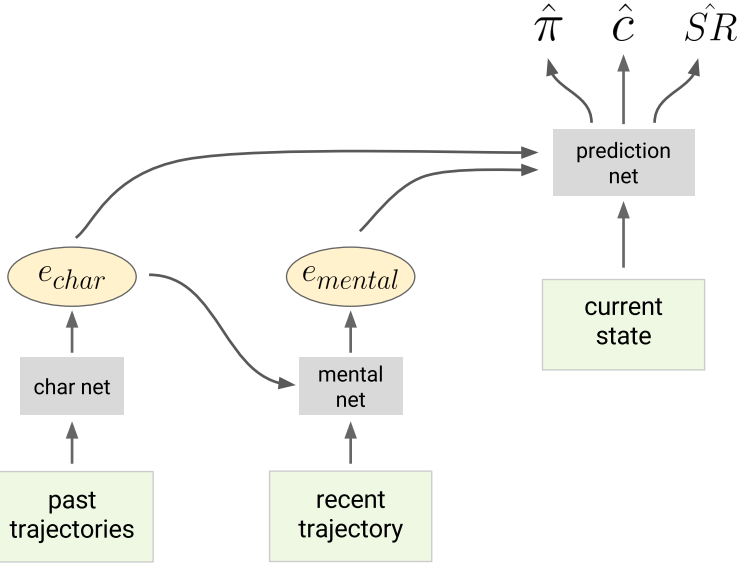
Idea: account for beliefs, models, and/or learning algorithms of other agents

Deep Reinforcement Opponent Networks



(a) DRON-concat: opponent representation is concatenated with the state representation.

(b) (b) DRON-MoE: Q-values predicted by K experts are combined linearly by weights from the gating network.

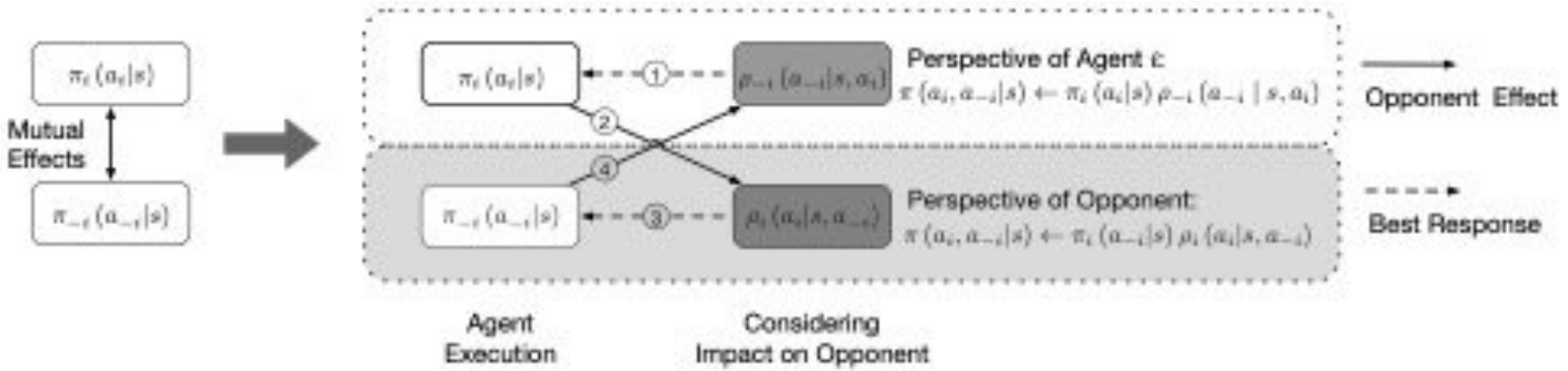


Theory of Mind neural network (*ToMnet*)

- The character net: parses trajectories to a character embedding, $\mathbf{e}_{\text{char}}$.
- The mental state net parses the trajectory on the current episode, to mental state, $\mathbf{e}_{\text{mental}}$.
- These embeddings are fed into the prediction net and outputs
  - next-step action probabilities,
  - probabilities of whether certain objects will be consumed
  - predicted successor representations

[He, He, et al. 2016, Raileanu, Roberta, et al. 2018. Rabinowitz, Neil C., et al. 2018,]

Mutual Influence as Recursive Reasoning



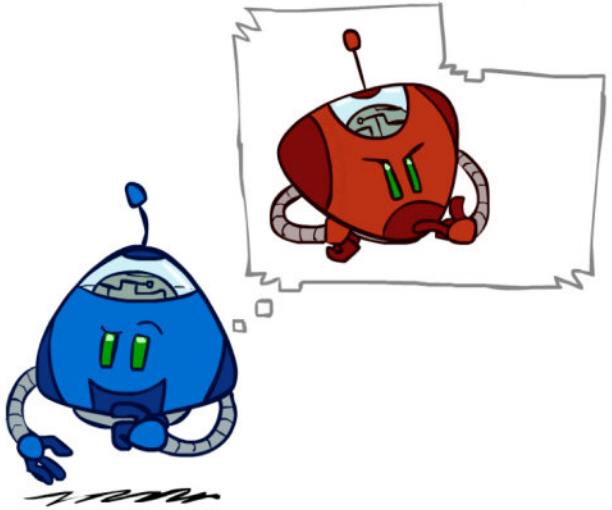
Probabilistic recursive reasoning framework. Recursive reasoning decouples the connections between agents. ①: agent i takes the best response after considering all the potential consequences of opponents' actions given its own action a_i. ②: how agent i behaves in the environment serves as the prior for the opponents to learn how their actions would affect a_i. ③: similar to ①, opponents take the best response to agent i. ④: similar to ②, opponents' actions are the prior knowledge to agent i on estimating how a_i will affect the opponents. Looping from step 1 to 4 forms recursive reasoning.

$$\pi(a_i, a_{-i}|s) = \underbrace{\pi_{-i}(a_{-i}|s, a_i)\pi_i(a_i|s)}_{\text{Perspective of Agent}i} = \underbrace{\pi_i(a_i|s, a_{-i})\pi_{-i}(a_{-i}|s)}_{\text{Perspective of Agent}-i}.$$

$$\begin{aligned}
\pi\left(a_i, a_{-i} \mid s\right) &= \pi_{-i}\left(a_{-i} \mid s, a_i\right) \pi_i\left(a_i \mid s\right) \rightarrow \text{ Correlated, Perspective of Agent } i \\
&= \pi_i\left(a_i \mid s, a_{-i}\right) \pi_{-i}\left(a_{-i} \mid s\right) \rightarrow \text{ Correlated, Perspective of Agent } -i \\
&\approx \pi_i\left(a_i \mid s\right) \pi_{-i}\left(a_{-i} \mid s\right) \rightarrow \text{ Non-correlated}
\end{aligned}$$

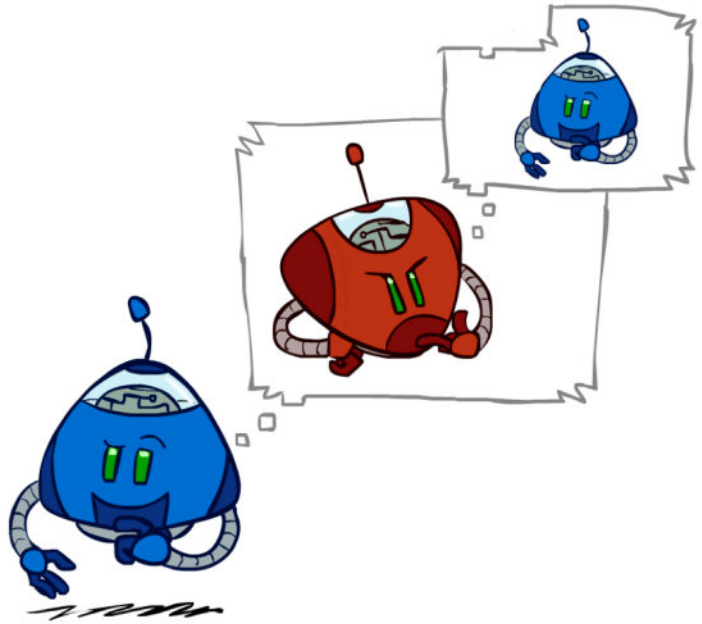
[Wen, Ying*, Yang, Yaodong*. et al. 2019.]

Recursive Reasoning Levels



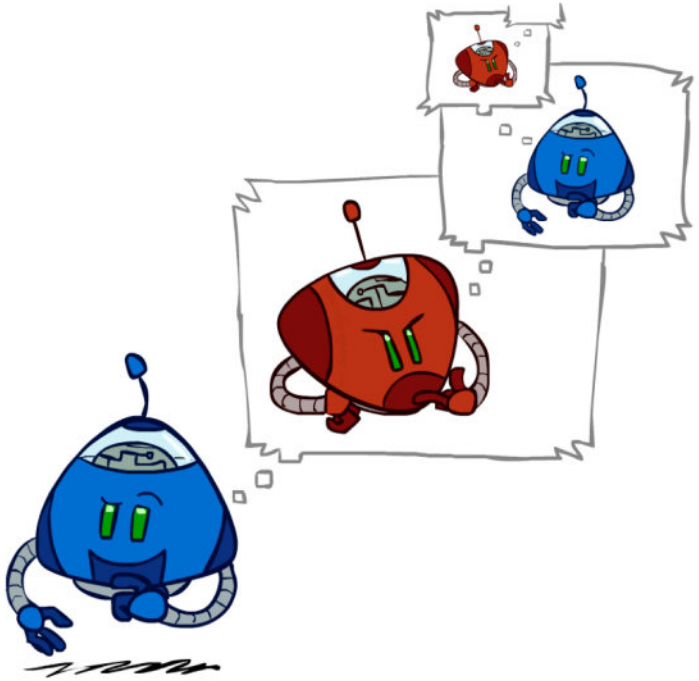
Source: CS 188, Berkeley.

Level 1



Source: CS 188, Berkeley.

Level 2

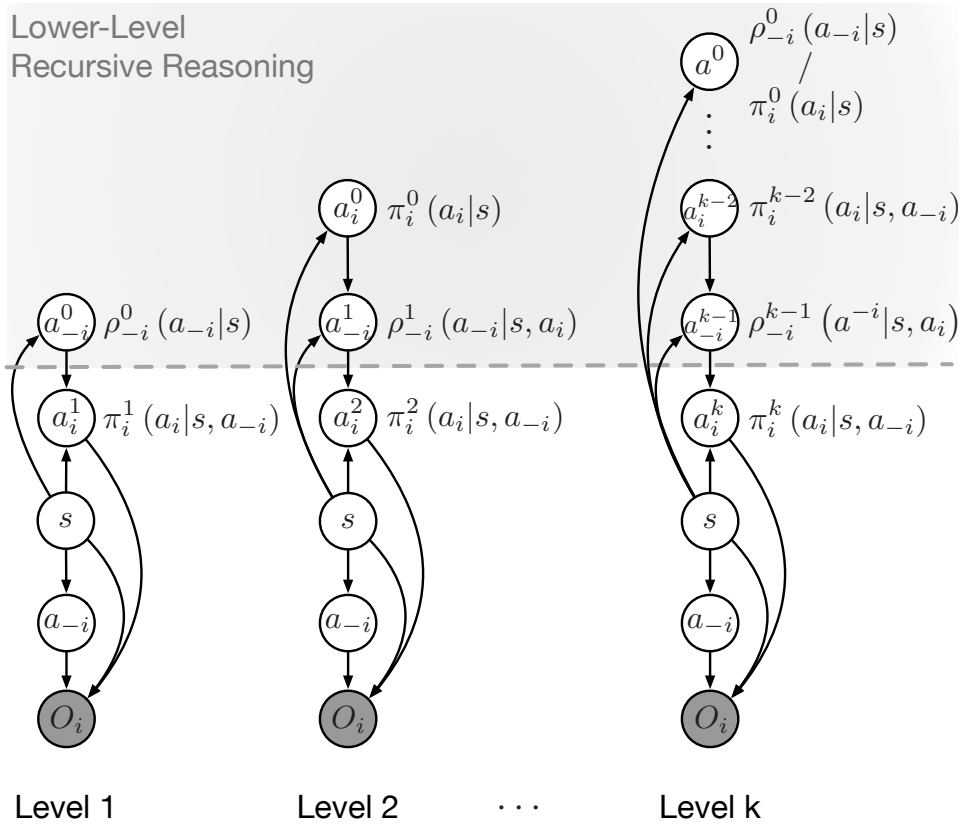


Source: CS 188, Berkeley.

Level 3

Level-k Recursive Reasoning

Agent i rolls out the recursive reasoning about opponents in its mind (grey area). In the recursion, agents with higher-level beliefs take the best response to the lower-level thinkers' actions. Higher-level models would conduct all the computations that the lower-level models have done, e.g. level-2 model contains level-1 model by integrating out $\pi_{i,0}(a_i|s)$.



Graphical model of the level-k reasoning model.

[Wen, Ying*, Yang, Yaodong*. et al. 2020.]

More Agent Modeling Work

►More comprehensive survey about opponent modeling:

- Albrecht, Stefano V., and Peter Stone. "**Autonomous agents modelling other agents: A comprehensive survey and open problems**." Artificial Intelligence 258 (2018): 66-95.

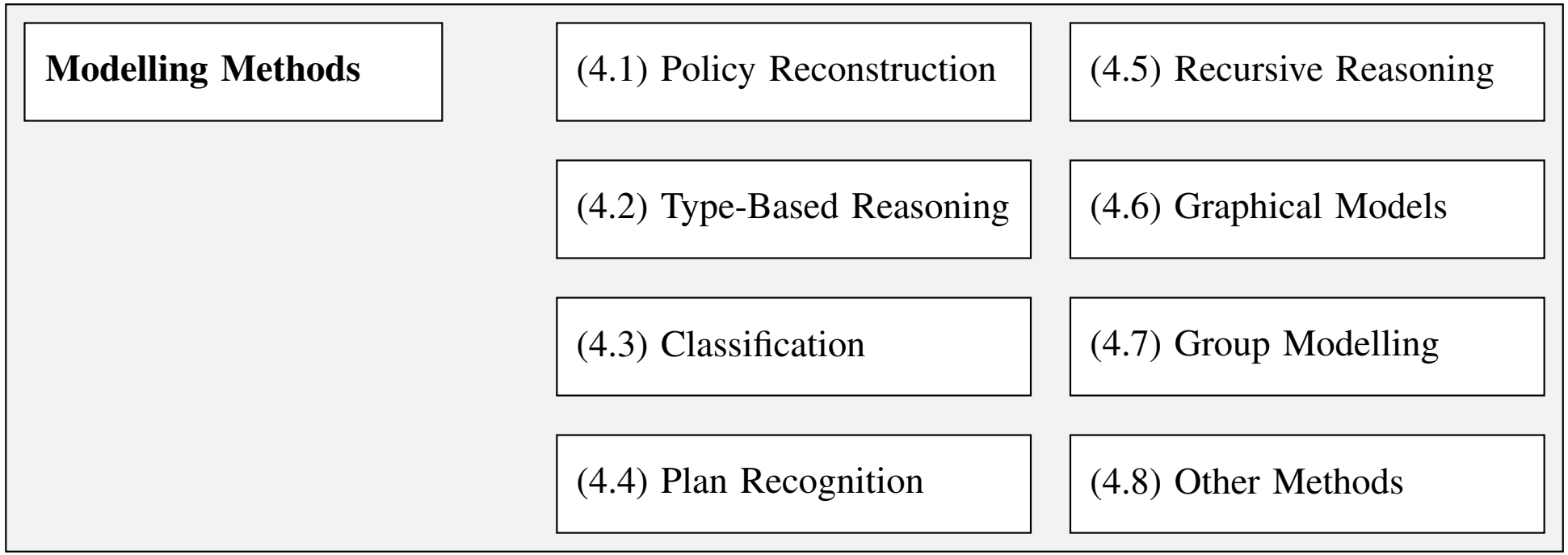
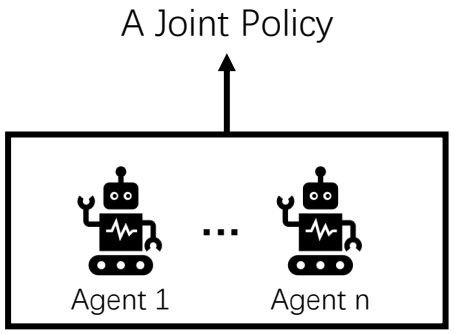


Figure 2: Surveyed modelling methods. Brackets show linked section numbers.

Agenda

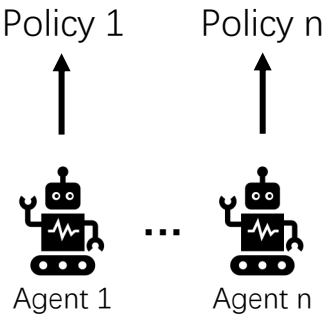
1. Introduction
2. Direct Interaction Modelling - Emergent Communication
3. Indirect Interaction Modelling
   a. Agent Modelling
   b. Implicit Interaction Modelling via Centralized Critic
   c. Explicit Policy Interaction Modelling via Policy-Space Analysis
4. Challenges and Perspectives X Quantum Theory

Implicit Interaction Modelling via Centralized Critic



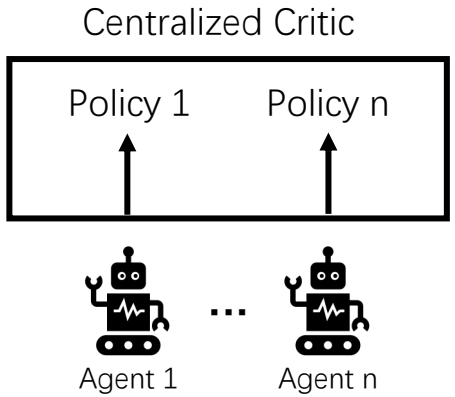
Fully Centralized Controller

Scalability ✗



Fully Decentralized Controllers

Non-stationarity
Credit assignment ✗



Centralized Critic with Decentralized Policy

Centralized Training with Decentralized Execution ✓

How to learn joint policy $\boldsymbol{\pi}$ or value function Q ?

Multi-Agent Deep Deterministic Policy Gradient

- If consider N continuous policies μ_i w.r.t. parameters θ_i, the gradient can be written as

$$\nabla_{\theta_i} J(\boldsymbol{\mu}_i) = \mathbb{E}_{\mathbf{x},a\sim\mathcal{D}}[\nabla_{\theta_i}\boldsymbol{\mu}_i(a_i|o_i)\nabla_{a_i} Q_i^{\boldsymbol{\mu}}(\mathbf{x}, a_1, ..., a_N)|_{a_i=\boldsymbol{\mu}_i(o_i)}]$$

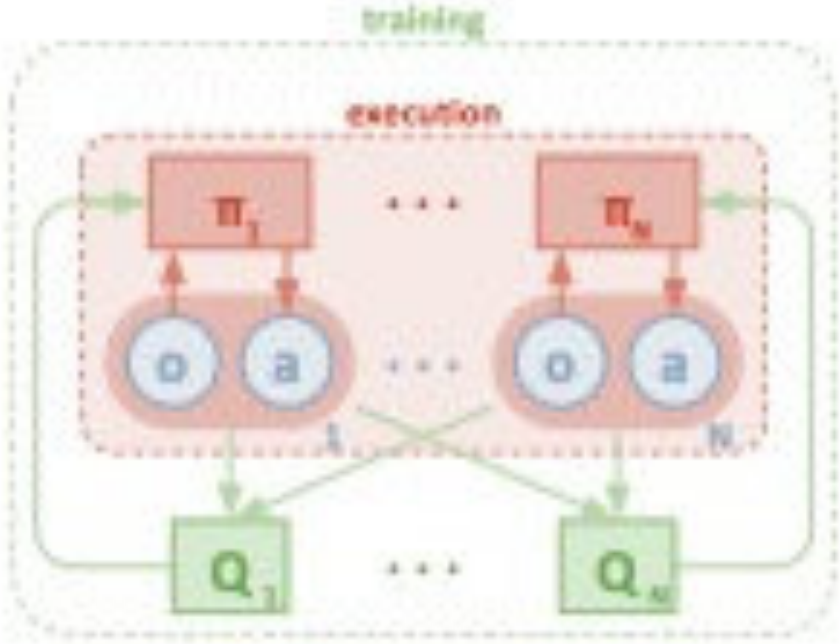
- The Q-function can be optimized by loss function:

$$\mathcal{L}(\varphi_i) = \mathbb{E}_{\mathbf{x},a,r,\mathbf{x}'}[(Q_i^{\boldsymbol{\mu}}(\mathbf{x}, a_1, \ldots, a_N) - y)^2], \quad y = r_i + \gamma\, Q_i^{\boldsymbol{\mu}'}(\mathbf{x}', a'$$

- Decentralized actors (policies, π) to keep the self-play framework.
- Centralized critic (baseline, Q) to ease learning and reduce variance

- Robust Extension : MiniMax Multi-agent Deep Deterministic Policy Gradient (M3DDPG)
  - Minimax objective: introduce minimax concept into deep MARL
  - Key idea: replace the inner-loop minimization by a one-step gradient descent

$$\nabla_{\theta_i} J(\theta_i) = \left[\begin{array}{c} \nabla_{\theta_i}\mu_i(o_i)\,\nabla_{a_i} Q^{\mu}_{\mathrm{M},i}(\mathrm{x}, a_1^\star, \ldots, a_i, \ldots a_N^\star) \mid \\ a_i = \mu_i(o_i) \\ a_j^\star = a_j + \hat{\epsilon}_j, \quad \forall j \neq i \\ \hat{\epsilon}_j = -\alpha_j \nabla_{a_j} Q^{\mu}_{\mathrm{M},i}(\mathrm{x}, a_1, \ldots, a_N) \end{array}\right]$$



Centralized training and decentralized execution

[Lowe, Ryan, et al. 2017. Li, Shihui, et al. 2019.]

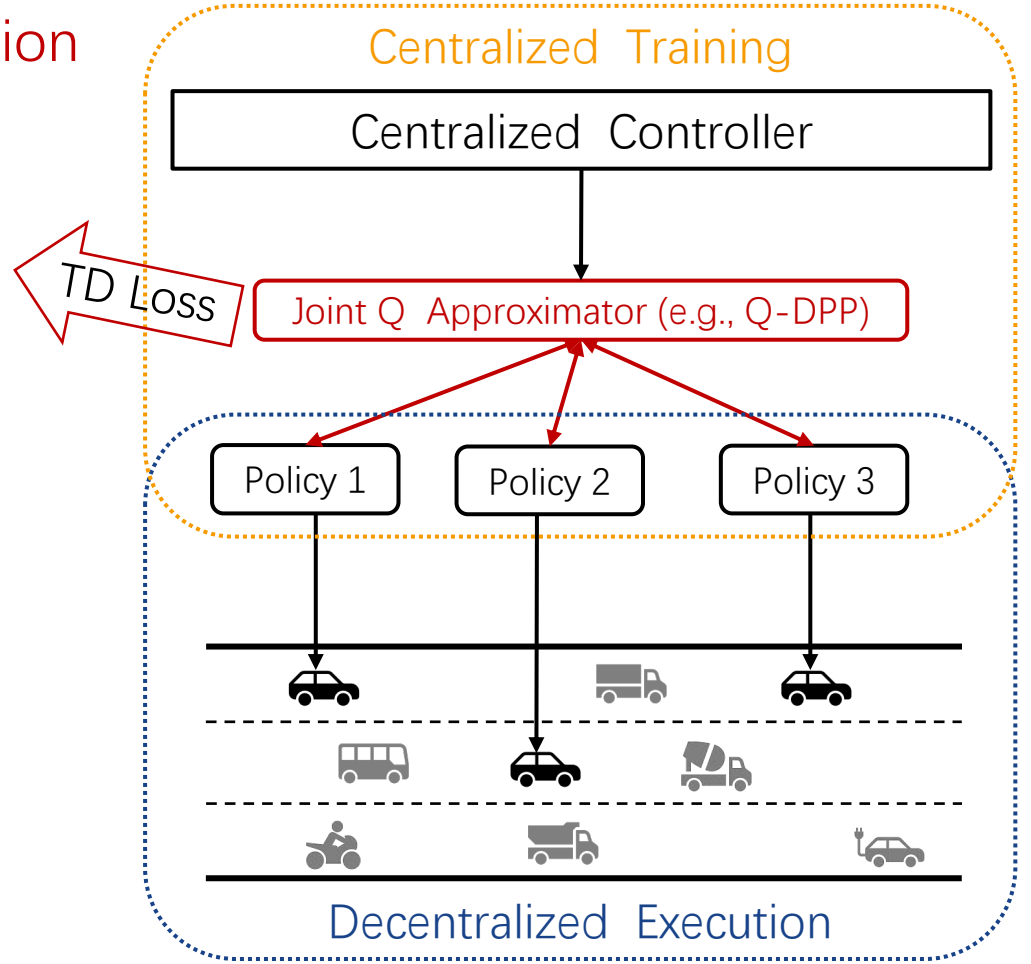
Centralized Critic Decomposition Methods

- Setting: Dec-POMDP
- Paradigm: Centralized training with decentralized execution

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{\tau},\boldsymbol{a},r,\boldsymbol{\tau}')\in D}\left[\left(r + \gamma V\left(\tau';\boldsymbol{\theta}^{-}\right) - Q(\boldsymbol{\tau},\boldsymbol{a};\boldsymbol{\theta})\right)^2\right]$$

$$V\left(\tau';\theta^{-}\right) = \max_{a'} Q\left(\tau',a';\theta^{-}\right)$$

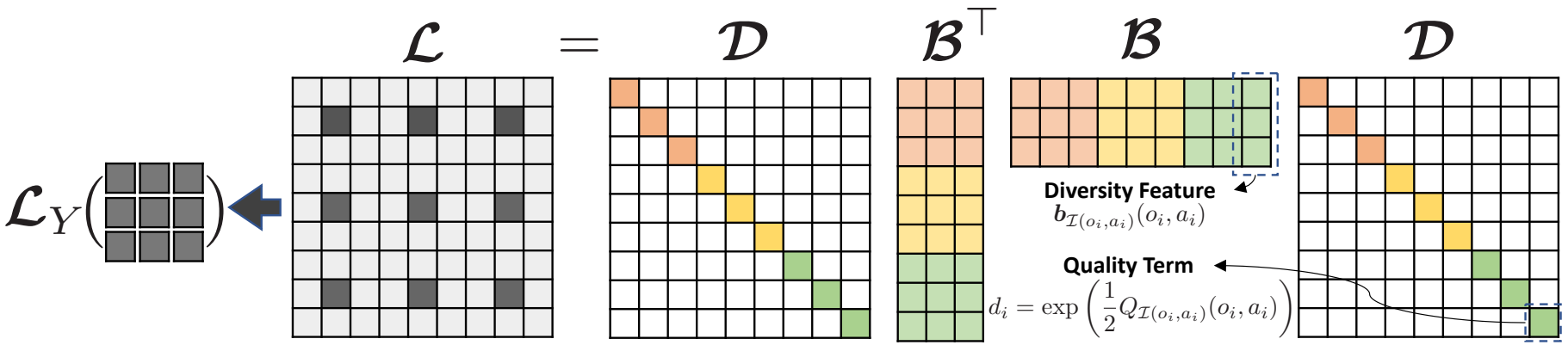
- Individual-Global Maximization (IGM) Principle
  - Consistent action selection between joint and individuals
  - $$\arg\max_{\boldsymbol{a}} Q^{\pi}(s,\boldsymbol{a}) = \left\{\arg\max_{a_i} Q_i\left(s,a_i\right)\right\}_{i\in\mathcal{N}}$$
- Cooperative multi-agent RL shares a single joint reward signal
  - Need to learn to decompose the team value function into agent-wise value functions.
- Although learning requires some centralization, the learned agents can be deployed independently.



Q-DPP: A Function Approximator for Cooperative MARL

The DPP kernel $\boldsymbol{\mathcal{L}}$ can represent the joint Q-function $Q^{\boldsymbol{\pi}}(\boldsymbol{o}, \boldsymbol{a})$ with both quality (maximizing reward) & diversity (different behaviors).

$$
\begin{aligned}
Q^{\boldsymbol{\pi}}(\boldsymbol{o}, \boldsymbol{a}) :&= \operatorname{logdet}\left(\boldsymbol{\mathcal{L}}_{Y=\left\{\left(o_1^t, a_1\right), \ldots, \left(o_N^t, a_N\right)\right\} \in \mathcal{C}\left(\boldsymbol{o}^t\right)} = \boldsymbol{\mathcal{W}}_Y \boldsymbol{\mathcal{W}}_Y^{\top}\right) \\
&= \log\left(\operatorname{tr}\left(\boldsymbol{\mathcal{D}}_Y^{\top} \boldsymbol{\mathcal{D}}_Y\right) \det\left(\boldsymbol{\mathcal{B}}_Y^{\top} \boldsymbol{\mathcal{B}}_Y\right)\right) \qquad \text{Quality-Diversity Decomposition} \\
&= \sum_{i=1}^{N} \underbrace{Q_{\mathcal{I}\left(o_i^t, a_i\right)}\left(o_i, a_i\right)}_{\text{Quality}} + \underbrace{\operatorname{logdet}\left(\boldsymbol{\mathcal{B}}_Y^{\top} \boldsymbol{\mathcal{B}}_Y\right)}_{\text{Diversity}}
\end{aligned}
\tag{3}
$$



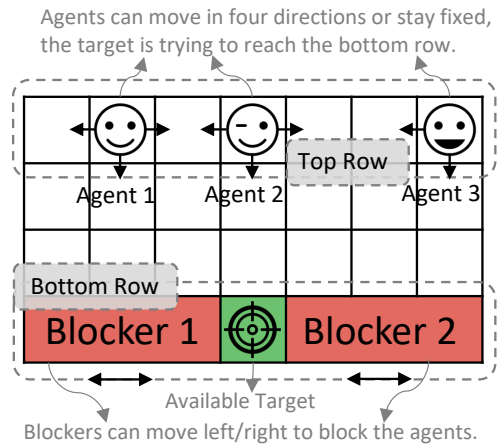
Example of Q-DPP with quality-diversity kernel decomposition in a single-state three-player learning task.

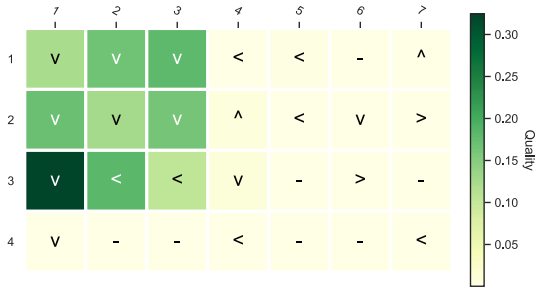
[Yang, Yaodong*; Wen, Ying*, et al. 2020.]

Q-DPP Performance

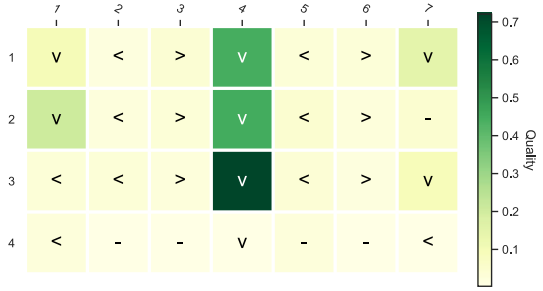
The acquired diverse behaviors on Blocker Game.

- The optimal action of one agent doesn't depend on the others.
- The behaviors acquired by different agents are orthogonal.
- Q-DPP meets the decentralizability assumption naturally.

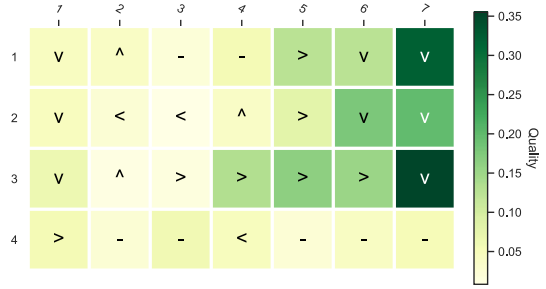




Agent 1



Agent 2

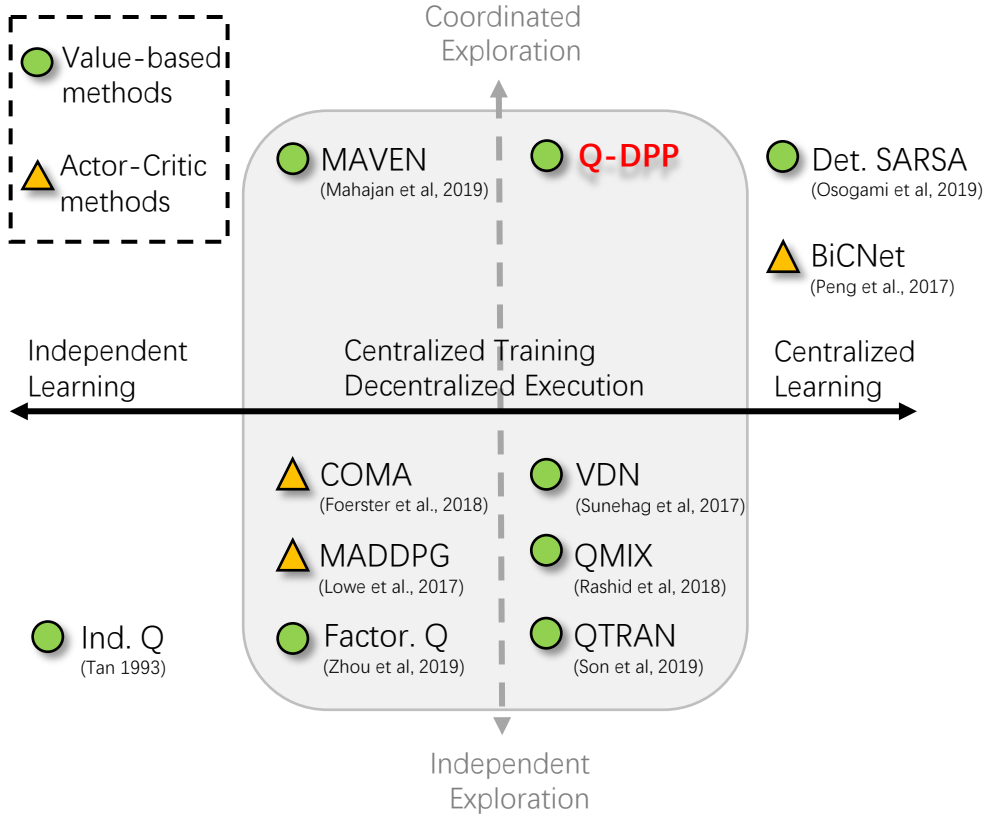


Agent 3

Learned state values ($\text{argmax}_{a_i} Q(s, a_i)$) for each agent.

Centralized Critic Decomposition Methods

- ►Value-Based Methods
  - Paradigm: centralized training with decentralized execution.
  - Methods: VDN, QMIX, QTRAN, MAVEN, QPLEX, NDQ, ROMA...
- ►Continuous actions extensions: COVDN and COMIX. VDN and QMIX respectively.
- ►Actor-Critic Methods
  - Paradigm: centralized critic and decentralized actors.
  - DOP: off-policy decomposed multi-agent policy gradient.
- ►Question:
  - Can we exploit the implicit structures in the centralized critic to boost the learning?

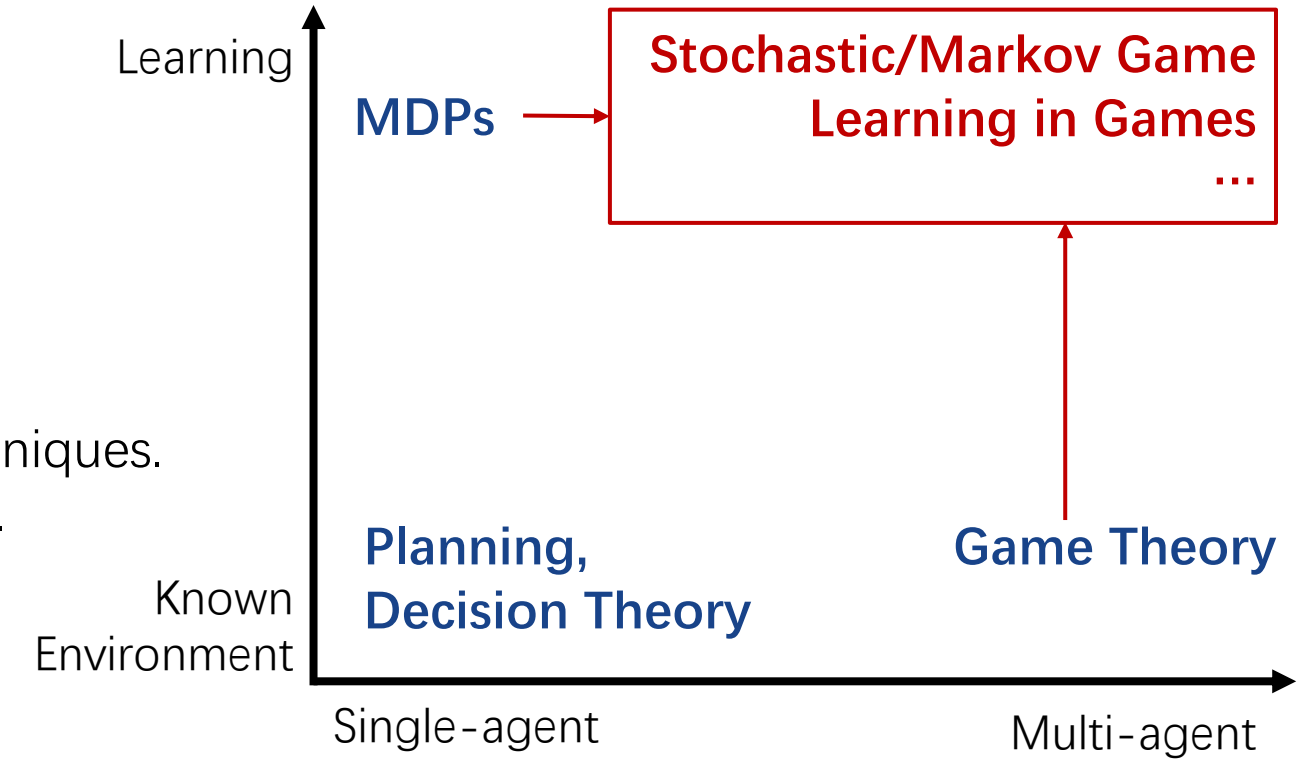


Agenda

1. Introduction
2. Direct Interaction Modelling - Emergent Communication
3. Indirect Interaction Modelling
   a. Agent Modelling
   b. Implicit Interaction Modelling via Centralized Critic
   c. **Explicit Policy Interaction Modelling via Policy-Space Analysis**
4. Challenges and Perspectives X Quantum Theory

Where is Game Theory – Explicit Policy Analysis

- ►Games Theory
  - Modeling the interactions.
- ►Reinforcement Learning
  - Learning decision making.
- ►Some successful combinations:
  - Empirical game theoretic techniques.
  - Policy space response oracles.
  - …



Can we use the game theory to guide
the multi-agent reinforcement learning?

Changing Policies' Relationship

Improvement Against Fixed Opponent Policy

$$\eta_i(\pi_i, \pi_{-i}) \overset{\leq}{\longrightarrow} \eta_i(\pi'_i, \pi_{-i}) \quad \eta_{-i}(\pi_i, \pi'_{-i}) \overset{\geq}{\longleftarrow} \eta_{-i}(\pi_i, \pi_{-i})$$

Unknown Improvement Against Opponent's Simultaneous Learning

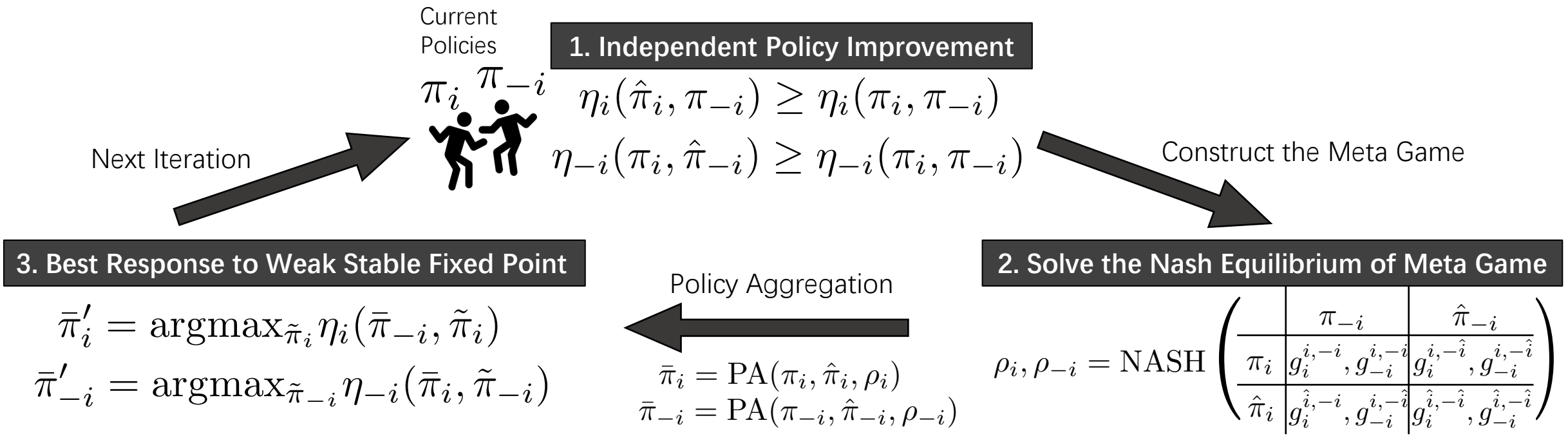
$$? \quad \eta_i(\pi'_i, \pi'_{-i}) \quad \eta_{-i}(\pi'_i, \pi'_{-i}) \quad ?$$

The relationship of discounted returns η_i for an agent i given the different joint policy pairs, where π_i is the current policy, π'_i is the simultaneously updated policy.

Can we allows agents to consider the policy-space mutual influence and shape the learning direction which gives stable improvement guarantee?

Wen, Ying; Chen, Hui; Yang, Yaodong; Tian, Zheng; Li, Minne; Chen, Xu; Wang, Jun, Multi-Agent Trust Region Learning.

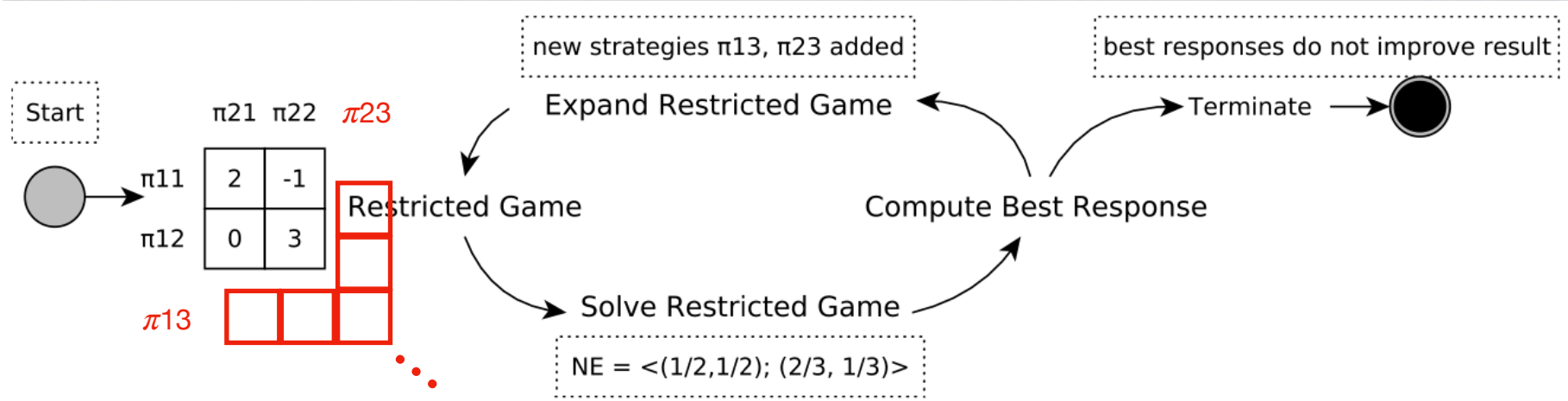
Multi-Agent Trust Region Learning



Overview of the multi-agent trust region learning phases in two-agent games. It can be easily extended to n-agent case by solving the n-agent two-action matrix form meta game.

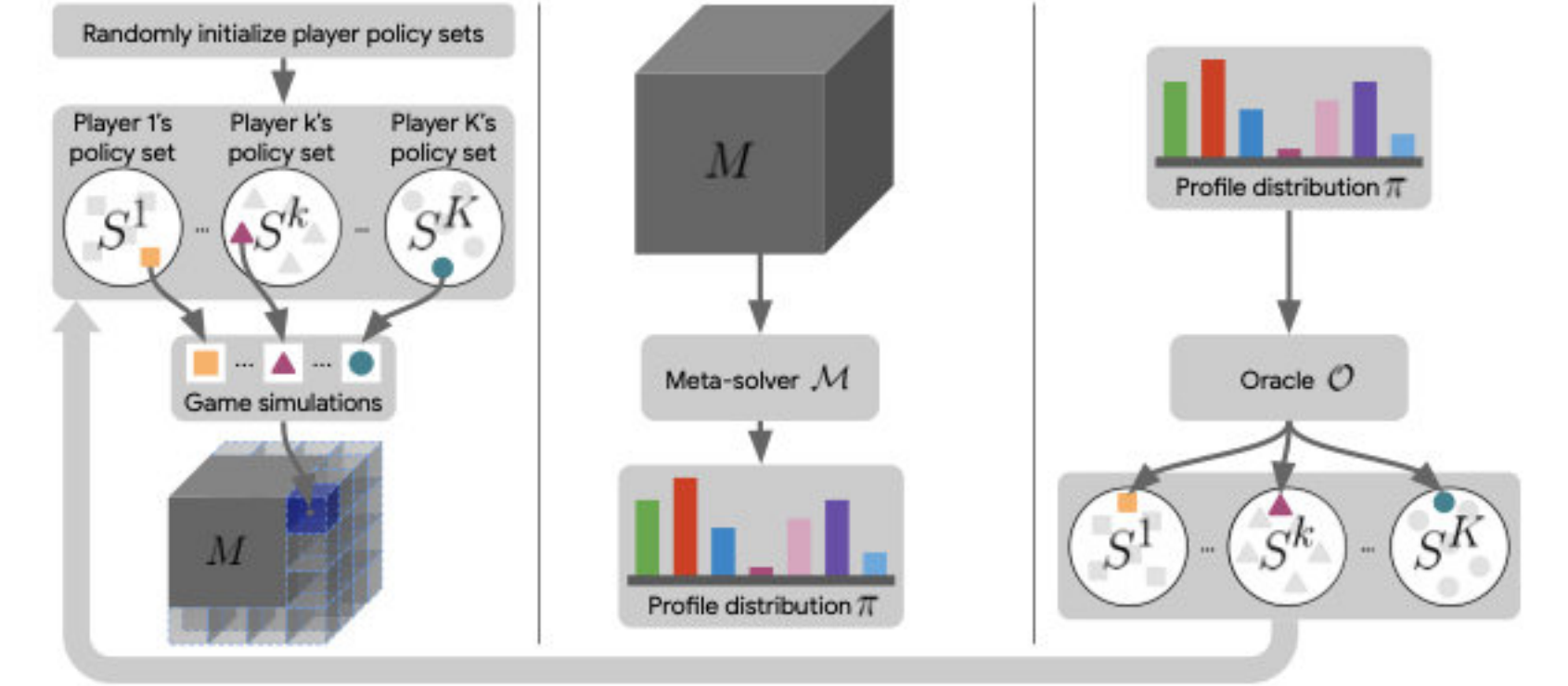
Double Oracle

- Double Oracle is also an iterated best response method, the difference is that, it best responds to the opponent’ s Nash equilibrium at each iteration.
- If the newly added best response is already in the strategy pool, then terminate. Why “oracle” : the search for a best response is guided by an oracle. It guarantees to converge to minimax equilibrium in finite games.



[Lanctot, Marc, et al. 2017.]

PSRO Algorithm Phases



(a) Complete: compute missing payoff tensor M entries via game simulations.

(b) Solve: given the updated payoff tensor M, calculate meta-strategy π via meta-solver $\mathcal{M}$.

(c) Expand: append a new policy to each player's policy space using the oracle $\mathcal{O}$.

[Muller, Paul, et al. 2019.]

Agenda

1. Introduction
2. Direct Interaction Modelling - Emergent Communication
3. Indirect Interaction Modelling
    a. Opponent Modelling
    b. Implicit Interaction Modelling via Centralized Critic
    c. Explicit Policy Interaction Modelling via Policy-Space Analysis
4. Challenges and Perspectives X Quantum Theory

From Fixed-Play to Open-Ended-Play

Policy Space Distribution

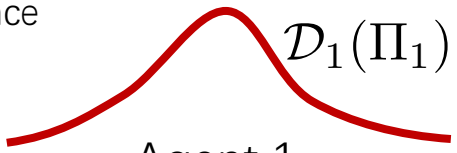
Π_1 Policy Space of Agent i

$\pi_1 \in \Pi_1$ A Policy Instance

$\pi_1 \sim \mathcal{D}_1(\Pi_1)$

Sample a Policy According to the Policy Space Distribution

$\mathcal{D}_i(\Pi_i)$

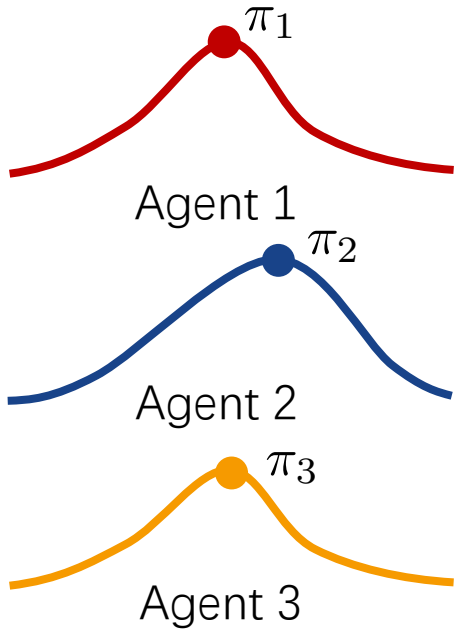


$\pi_2 \in \Pi_2$

$\pi_2 \sim \mathcal{D}_2(\Pi_2)$

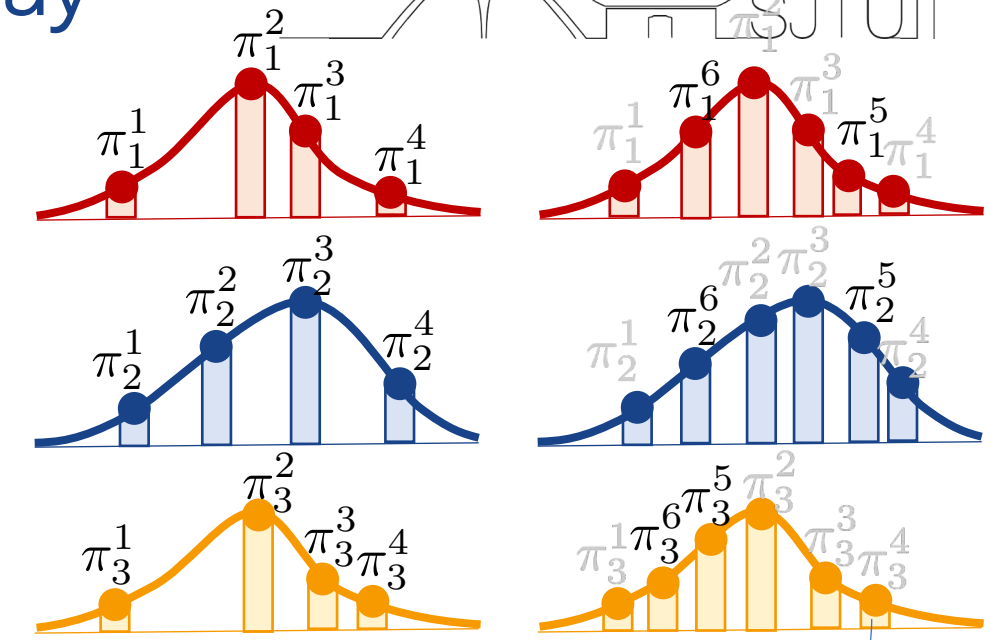
$\pi_3 \in \Pi_3$

$\pi_3 \sim \mathcal{D}_3(\Pi_3)$



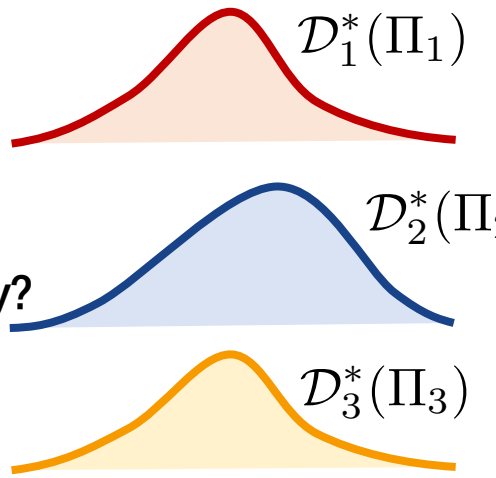
Co-Play/Fixed-Play

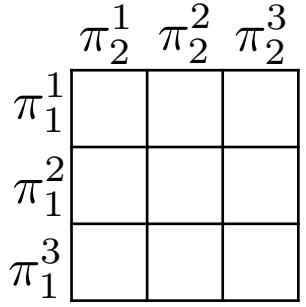
Deterministic distribution



Open-Ended-Play: Eval + Expand

Discrete distribution





Meta-game Analysis

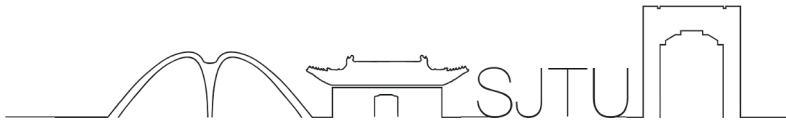
Mixed Policy-Space Strategy

Continuous distribution

Problems:

1. **It is hard to evaluate the policy pool.**
2. **Can we have a continuous distribution to sample the policy?**
3. **Better policy representation?**
4. **How to introduce the prior to select the policy?**

量子力学与多智能体学习



►直接引入，量子博弈，扩张策略集，助力博弈分析

►连续策略空间的建模和优化
- 问题：策略空间过于巨大，无法直接优化
- 方案：通过量子/贝叶斯神经网络构建策略网络，使直接优化策略空间策略的学习变成可能。

►策略空间博弈分析：可以根据经验元博弈构建一个策略交互图的概率转移矩阵
- 问题：一般随机游走求解策略的平稳分布随着策略池大小，求解难度呈指数被增长
- 方案：通过量子游走直接并行线性求解策略交互图的平稳分布

►一般现实博弈中，尤其是竞争博弈，一个纯策略总是有反制策略
- 问题：纯策略在现实中不够鲁棒
- 方案：能否构建量子混合态策略，根据对手的行为进行观测（引入prior？），坍缩成一个纯策略。
- 例如：比如王者荣耀，不同英雄组合，根据阵容及对方行为，采取不同的策略。

Thanks for Listening!

Q&A

Ying Wen
ying.wen@sjtu.edu.cn
https://yingwen.io

This slides is available at: github.com/ying-wen/dai-slides

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

JOHN HOPCROFT
CENTER FOR
COMPUTER SCIENCE

References I

Value Decomposition

[VDN] Sunehag, Peter, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot et al. "Value-decomposition networks for cooperative multi-agent learning." arXiv preprint arXiv:1706.05296 (2017).

[QMIX] Rashid, Tabish, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning." arXiv preprint arXiv:1803.11485 (2018).

[QTRAN] Son, Kyunghwan, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. "Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning." arXiv preprint arXiv:1905.05408 (2019).

[QDPP] Yang, Yaodong*, Ying Wen*, Lihuan Chen, Jun Wang, Kun Shao, David Mguni, and Weinan Zhang. "Multi-Agent Determinantal Q-Learning." arXiv preprint arXiv:2006.01482 (2020).

[QVPD] Yang, Yaodong, Jianye Hao, Guangyong Chen, Hongyao Tang, Yingfeng Chen, Yujing Hu, Changjie Fan, and Zhongyu Wei. "Q-value Path Decomposition for Deep Multiagent Reinforcement Learning." arXiv preprint arXiv:2002.03950 (2020).

[QPLEX] Wang, Jianhao, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. "Qplex: Duplex dueling multi-agent q-learning." arXiv preprint arXiv:2008.01062 (2020).

[ROMA] Wang, Tonghan, Heng Dong, Victor Lesser, and Chongjie Zhang. "Roma: Multi-agent reinforcement learning with emergent roles." In Proceedings of the 37th International Conference on Machine Learning. 2020.

[MAVEN] Mahajan, Anuj, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. "Maven: Multi-agent variational exploration." In Advances in Neural Information Processing Systems, pp. 7613-7624. 2019.

de Witt, Christian Schroeder, Bei Peng, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Böhmer, and Shimon Whiteson. "Deep Multi-Agent Reinforcement Learning for Decentralized Continuous Cooperative Control." arXiv preprint arXiv:2003.06709 (2020).

Centralized Critic Decentralized Actor

[MADDPG] Lowe, Ryan, Yi . Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. "Multi-agent actor-critic for mixed cooperative-competitive environments." In Advances in neural information processing systems, pp. 6379-6390. 2017.

[COMA] Foerster, Jakob, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. "Counterfactual multi-agent policy gradients." arXiv preprint arXiv:1705.08926 (2017).

[DOP] Wang, Yihan, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. "Off-Policy Multi-Agent Decomposed Policy Gradients." arXiv preprint arXiv:2007.12322 (2020).

References II

Learning to Communicate

[BiCNet] Peng, Peng*, Ying Wen*, Quan Yuan, Yaodong Yang, Zhenkun Tang, Haitao Long, and Jun Wang. "Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games." arXiv preprint arXiv:1703.10069 2 (2017): 2.

[DIAL] Foerster, Jakob, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. "Learning to communicate with deep multi-agent reinforcement learning." In Advances in neural information processing systems, pp. 2137-2145. 2016.

[CommNet] Sukhbaatar, Sainbayar, and Rob Fergus. "Learning multiagent communication with backpropagation." In Advances in neural information processing systems, pp. 2244-2252. 2016.

[ATOC] Jiang, Jiechuan, and Zongqing Lu. "Learning attentional communication for multi-agent cooperation." In Advances in neural information processing systems, pp. 7254-7264. 2018.

[VAIN] Hoshen, Yedid. "Vain: Attentional multi-agent predictive modeling." In Advances in Neural Information Processing Systems, pp. 2701-2711. 2017.

Opponent-Aware Learning

[IPOMDP] Gmytrasiewicz, Piotr J., and Prashant Doshi. "A framework for sequential planning in multi-agent settings." Journal of Artificial Intelligence Research 24 (2005): 49-79.

[ROMMEO] Tian, Zheng*, Ying Wen*, Zhichen Gong, Faiz Punakkath, Shihao Zou, and Jun Wang. "A regularized opponent model with maximum entropy objective." arXiv preprint arXiv:1905.08087 (2019).

[PR2] Wen, Ying, Yaodong Yang, Rui Luo, Jun Wang, and Wei Pan. "Probabilistic recursive reasoning for multi-agent reinforcement learning." arXiv preprint arXiv:1901.09207 (2019).

[GR2] Wen, Ying, Yaodong Yang, Rui Luo, and Jun Wang. "Modelling Bounded Rationality in Multi-Agent Interactions by Generalized Recursive Reasoning." arXiv preprint arXiv:1901.09216 (2019).

[LOLA] Foerster, Jakob N., Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. "Learning with opponent-learning awareness." arXiv preprint arXiv:1709.04326 (2017).

[DRON] He, He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. "Opponent modeling in deep reinforcement learning." In International conference on machine learning, pp. 1804-1813. 2016.

[DPIQN] Hong, Zhang-Wei, Shih-Yang Su, Tzu-Yun Shann, Yi-Hsiang Chang, and Chun-Yi Lee. "A deep policy inference q-network for multi-agent systems." arXiv preprint arXiv:1712.07893 (2017).

References III

[MToM] Rabinowitz, Neil C., Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Eslami, and Matthew Botvinick. "Machine theory of mind." arXiv preprint arXiv:1802.07740 (2018).

[SOM] Raileanu, Roberta, Emily Denton, Arthur Szlam, and Rob Fergus. "Modeling others using oneself in multi-agent reinforcement learning." arXiv preprint arXiv:1802.09640 (2018).

[SOS] Letcher, Alistair, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. "Stable opponent shaping in differentiable games." arXiv preprint arXiv:1811.08469 (2018).

Policy-Space Response Oracle

[PSRO] Muller, Paul, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes et al. "A generalized training approach for multiagent learning." arXiv preprint arXiv:1909.12823 (2019).

[Double Oracle] Lanctot, Marc, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. "A unified game-theoretic approach to multiagent reinforcement learning." In Advances in neural information processing systems, pp. 4190-4203. 2017.

[PSROrN] Balduzzi, David, Marta Garnelo, Yoram Bachrach, Wojciech M. Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. "Open-ended learning in symmetric zero-sum games." arXiv preprint arXiv:1901.08106 (2019).

[α-Rank] Omidshafiei, Shayegan, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. "α-rank: Multi-agent evaluation by evolution." Scientific reports 9, no. 1 (2019): 1-29.

[α^{α}-Rank] Yang, Yaodong, Rasul Tutunov, Phu Sakulwongtana, Haitham Bou Ammar, and Jun Wang. "$\alpha^{\alpha} $-Rank: Scalable Multi-agent Evaluation through Evolution." arXiv preprint arXiv:1909.11628 (2019).

Wellman, Michael P. "Methods for empirical game-theoretic analysis." In AAAI, pp. 1552-1556. 2006.

Tuyls, Karl, Julien Perolat, Marc Lanctot, Joel Z. Leibo, and Thore Graepel. "A generalised method for empirical game theoretic analysis." arXiv preprint arXiv:1803.06376 (2018).

Jordan, Patrick R., L. Julian Schvartzman, and Michael P. Wellman. "Strategy exploration in empirical games." In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1, pp. 1131-1138. 2010.

Balduzzi, David, Karl Tuyls, Julien Perolat, and Thore Graepel. "Re-evaluating evaluation." In Advances in Neural Information Processing Systems, pp. 3268-3279. 2018.

References V

Gradient-Based Learning

Mazumdar, Eric, Lillian J. Ratliff, and S. Shankar Sastry. "On Gradient-Based Learning in Continuous Games." SIAM Journal on Mathematics of Data Science 2, no. 1 (2020): 103-131.

Letcher, Alistair. "On the Impossibility of Global Convergence in Multi-Loss Optimization." arXiv preprint arXiv:2005.12649 (2020).

Jin, Chi, Praneeth Netrapalli, and Michael I. Jordan. "What is local optimality in nonconvex-nonconcave minimax optimization?." arXiv preprint arXiv:1902.00618 (2019).

Fiez, Tanner, Benjamin Chasnov, and Lillian Ratliff. "Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study." In International Conference on Machine Learning (ICML). 2020.

[SOS] Letcher, Alistair, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. "Stable opponent shaping in differentiable games." arXiv preprint arXiv:1811.08469 (2018).

[SGA] Balduzzi, David, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. "The mechanics of n-player differentiable games." arXiv preprint arXiv:1802.05642 (2018).

Mazumdar, Eric V., Michael I. Jordan, and S. Shankar Sastry. "On finding local nash equilibria (and only local nash equilibria) in zero-sum games." arXiv preprint arXiv:1901.00838 (2019).

[MATRL] Ying, Wen, Hui, Chen., Yaodong, Yang, Zheng, Tian, Minne, Li, Xu, Chen and Jun, Wang, "Multi-Agent Trust Region Learning."

Others

[NeuRD] Omidshafiei, Shayegan, Daniel Hennes, Dustin Morrill, Remi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, and Karl Tuyls. "Neural replicator dynamics." arXiv preprint arXiv:1906.00190 (2019).

Jaderberg, Max, Wojciech M. Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie et al. "Human-level performance in 3D multiplayer games with population-based reinforcement learning." Science 364, no. 6443 (2019): 859-865.

Leibo, Joel Z., Edward Hughes, Marc Lanctot, and Thore Graepel. "Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research." arXiv preprint arXiv:1903.00742 (2019).

Omidshafiei, Shayegan, Karl Tuyls, Wojciech M. Czarnecki, Francisco C. Santos, Mark Rowland, Jerome Connor, Daniel Hennes et al. "Navigating the Landscape of Games." arXiv preprint arXiv:2005.01642 (2020).

Czarnecki, Wojciech Marian, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. "Real World Games Look Like Spinning Tops." arXiv preprint arXiv:2004.09468 (2020).

Hu, Junling, and Michael P. Wellman. "Nash Q-learning for general-sum stochastic games." Journal of machine learning research 4, no. Nov (2003).