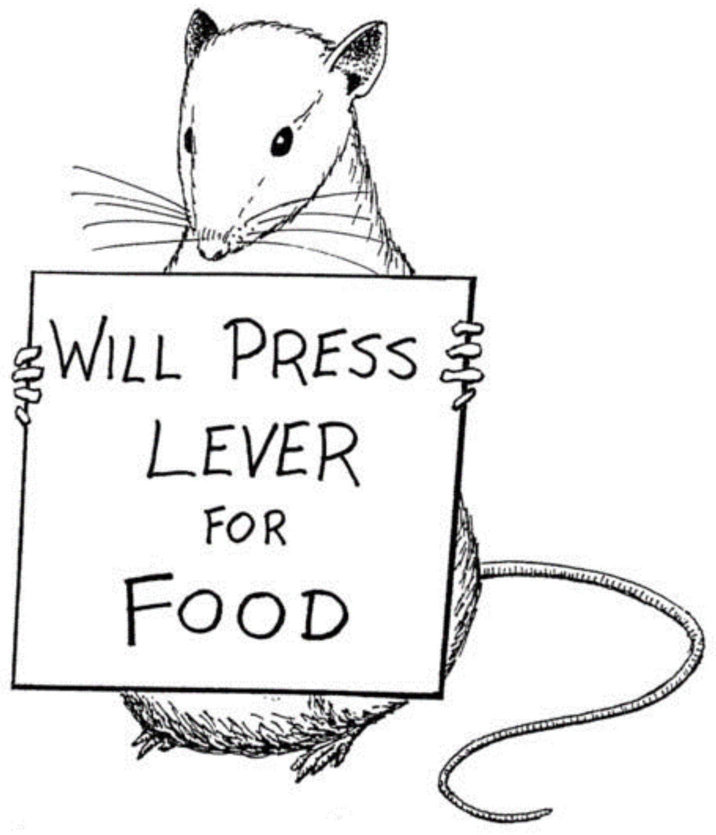
CS 277: Control and Reinforcement Learning

Winter 2024

Lecture 19: Multi-Agent RL

Roy Fox
Department of Computer Science
School of Information and Computer Sciences
University of California, Irvine



Logistics

assignments

- Quiz 8 due Wednesday
- Exercise 5 due next Monday
- Exercise 1–3, Quiz 4–7 grades to be published soon

evaluations

- Course evaluations due this weekend

Today's lecture

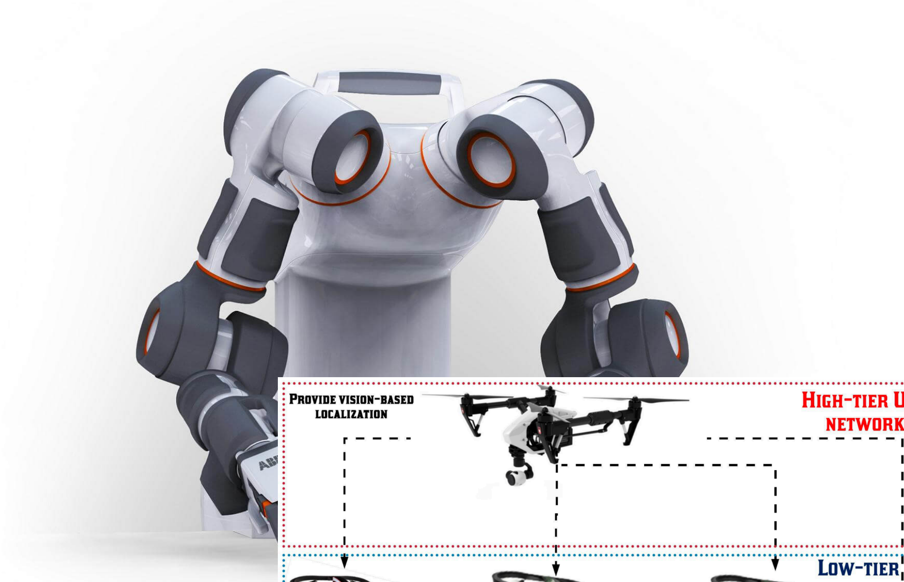
Centralized vs. decentralized RL

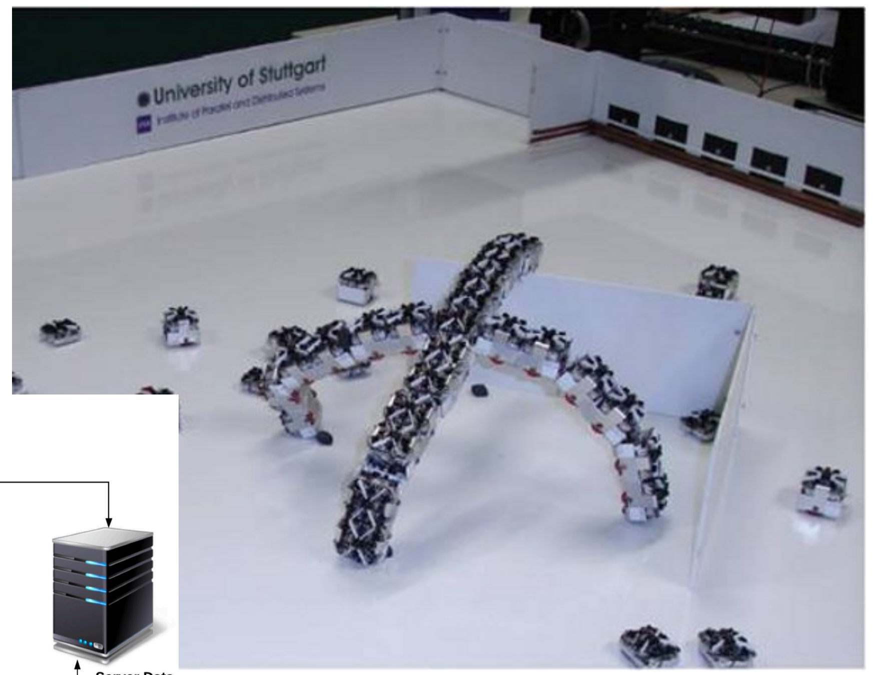
(Fictitious) Self Play

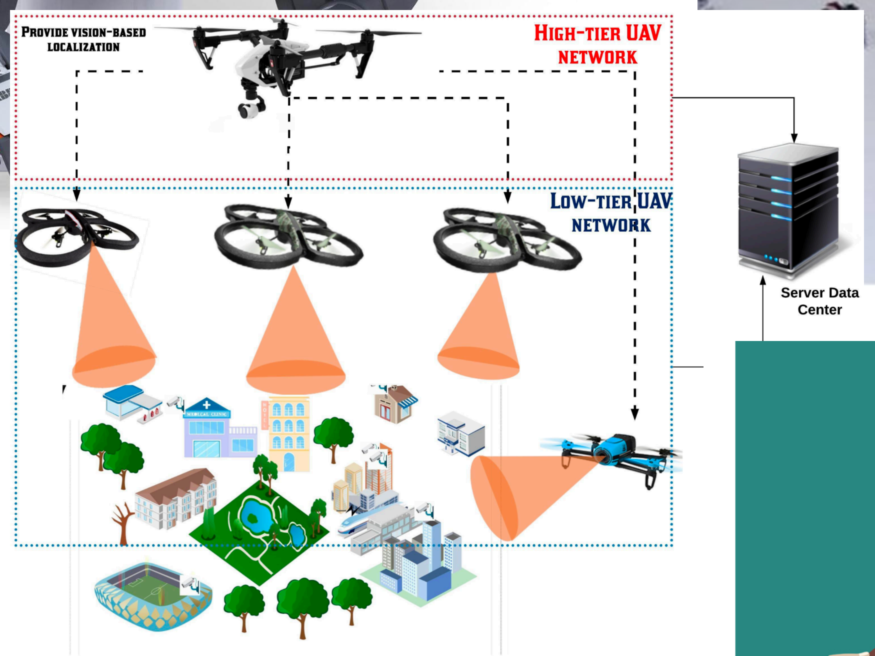
Double Oracle

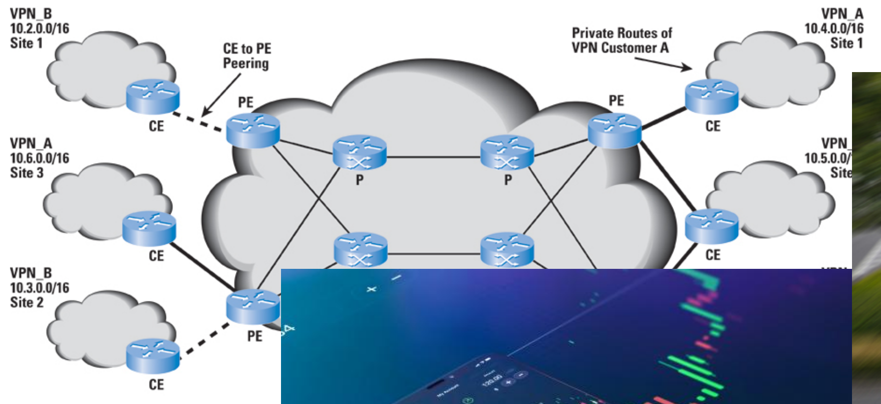
Multi-agent systems

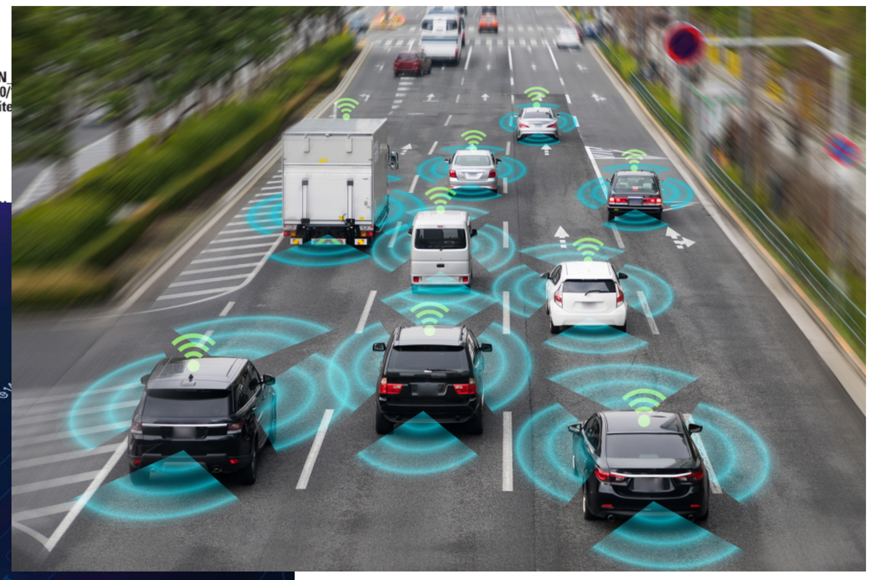
- Agent = actuator + sensor + self-interest (reward function) + optimizer
- Multi-agent system:
  - Distributed actuation
  - Distributed sensing / information hiding
  - Distinct interests (cooperative / competitive / indifferent / mix)
  - Distributed optimization
  - ⇒ distributed memory state ⇒ Theory of Mind







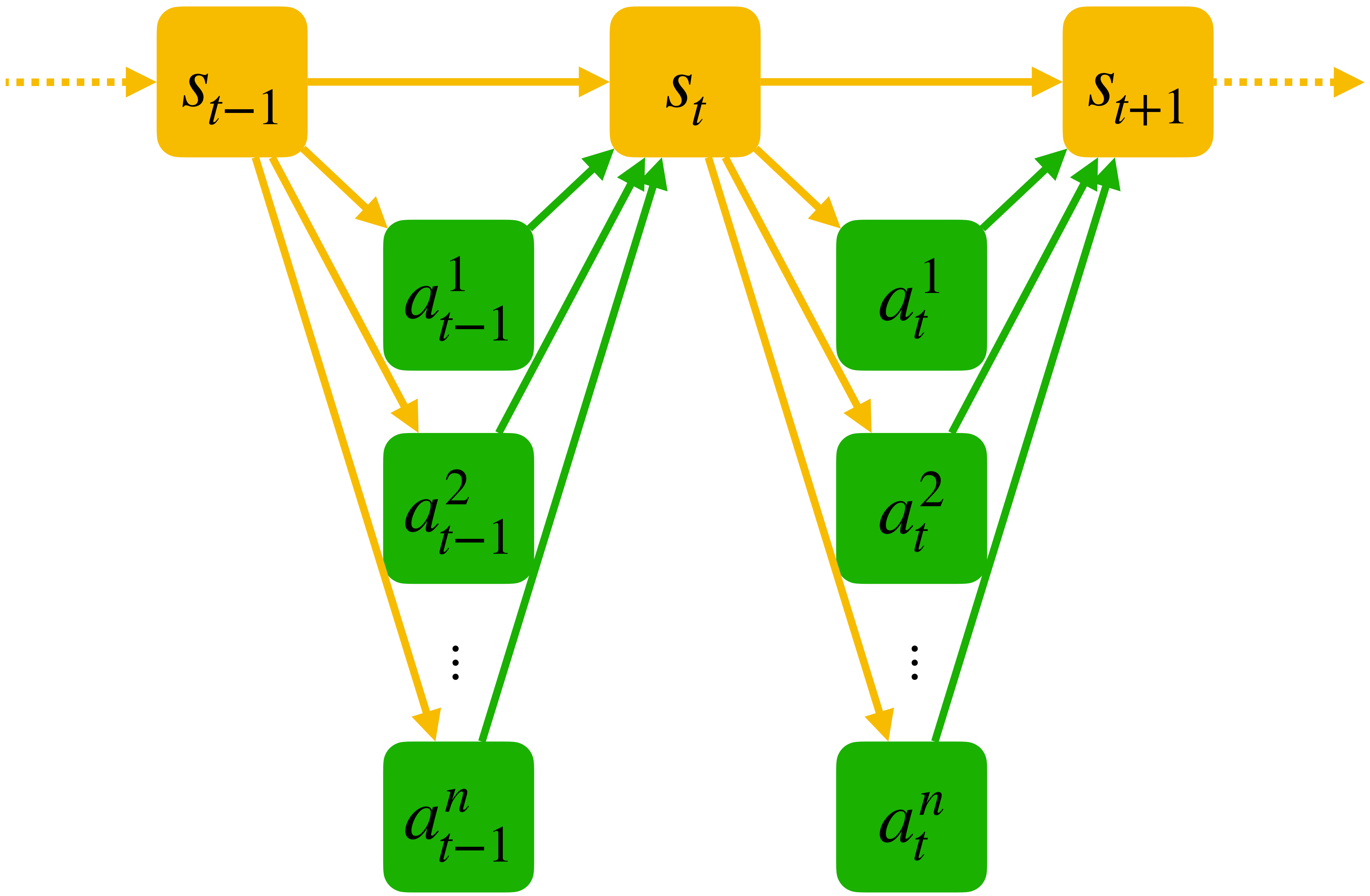




Centralized cooperative RL

- n agents = players; joint action = $a = (a^1, \ldots, a^n) \in \mathscr{A} = \mathscr{A}^1 \times \cdots \times \mathscr{A}^n$

- State transition = $p(s' \mid s, a)$; policy profile = $\pi = (\pi^1, \ldots, \pi^n)$



Centralized cooperative RL

- n agents = players; joint action = $a = (a^1, \ldots, a^n) \in \mathscr{A} = \mathscr{A}^1 \times \cdots \times \mathscr{A}^n$

- State transition = $p(s' | s, a)$; policy profile = $\pi = (\pi^1, \ldots, \pi^n)$

- Cooperative RL = all agents share the same rewards (payoffs) $r^1 = \cdots = r^n$

- Assume each agent gets observation o^i with probability $p(o^i | s)$
  - $\Rightarrow$ policy structure: $\pi(a | o) = \prod_i \pi^i(a^i | o^i)$

    agent i's action a^i can only depend on o^i

    action distributions are independent
  - Can jointly optimize π under this independence structure
  - E.g. PG: $\nabla_\theta \mathscr{L}_\theta = \nabla_\theta \log \pi_\theta(a | o) R = \sum_i \nabla_{\theta_i} \log \pi_{\theta_i}(a^i | o) R$

Independent RL

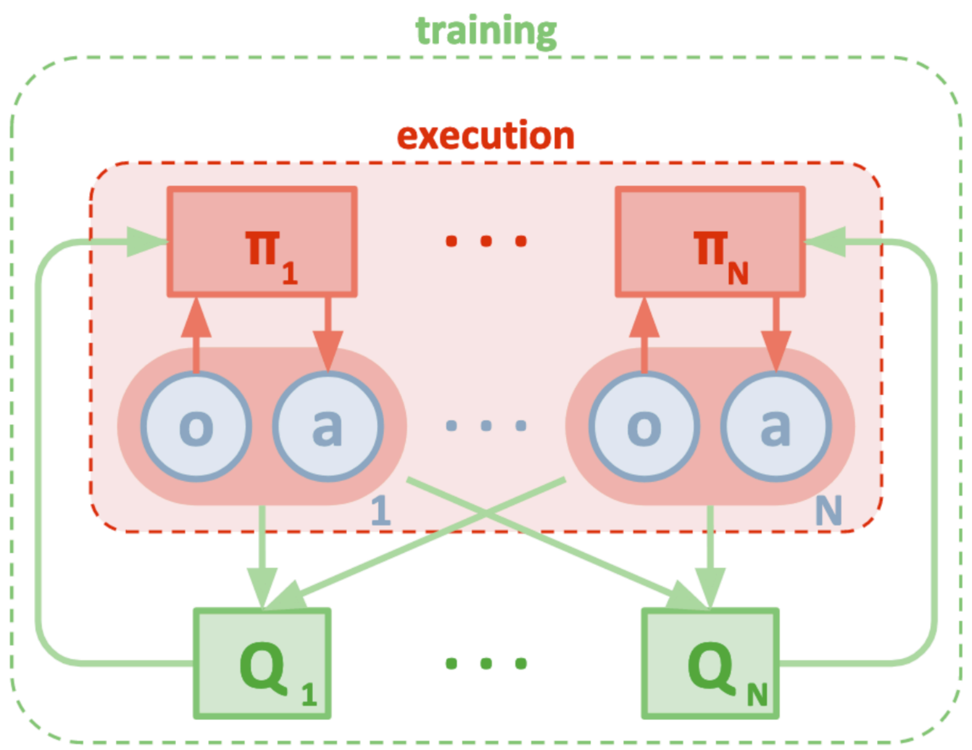
- Return R (or $R_{\geq t}$) is shared by all agents, but has high variance
  - Can we use some TD learning? Q-learning, AC, etc. ⇒ need Q^i

all agents except i

- Independent RL = train each agent i in MDP induced by others $-i$
  - $p(s' \mid s, a^i) = \mathbb{E}_{a^{-i}|o \sim \pi^{-i}}[p(s' \mid s, a)]$
  - Can train $Q^i(o^i, a^i)$ from experience $(o^i_t, a^i_t, r_t, o^i_{t+1})$
- Problem: the MDP keeps changing with π^{-i} ⇒ instability
  - May still work well in practice

Centralized critic / decentralized actors

- Actor–Critic presents opportunity:
  - No critic in test time ⇒ critic may be unrealizable



- Multi-Agent Deep Deterministic Policy Gradient (MADDPG):
  - Train critic $Q(o,a)$ for joint observation + action from experience (o_t, a_t, r_t, o_{t+1})

    or $Q^i(o, a^i)$
  - Use critic to train actors $\pi^i(a^i \mid o^i)$
- Stochastic actors: $\nabla_{\theta_i} \mathcal{L}_i = \nabla_{\theta_i} \log \pi_{\theta_i}(a^i \mid o^i) Q(o,a)$ (like AC)
- Deterministic actors: $\nabla_{\theta_i} \mathcal{L}_i = \nabla_{\theta_i} \mu_{\theta_i}(o^i) \nabla_{a^i} Q(o,a) \Big|_{a_i = \mu_{\theta_i}(o^i)}$ (like DDPG)

Today's lecture

Centralized vs. decentralized RL

(Fictitious) Self Play

Double Oracle

Solution concept: Nash equilibrium

- Best response of player i to π^{-i}: $b^i(\pi^{-i}) = \arg\max_{\pi^i} \mathbb{E}_{\pi^i,\pi^{-i}}[R^i]$
- Nash equilibrium π = each π^i is best response to π^{-i}
  - $\Rightarrow$ player i has no incentive to deviate
- Example 1: Prisoner's Dilemma

	Cooperate	Defect
Cooperate	-1 \ -1	-3 \ 0
Defect	0 \ -3	-2 \ -2

- Example 2: Matching Pennies
  - Generally, stochastic policies needed

mixed equilibrium

	Heads	Tails
Heads	1 \ -1	-1 \ 1
Tails	-1 \ 1	1 \ -1

Nash equilibrium: challenges

- Problem 1: is finding a Nash equilibrium all we need?
  - Example: Coordination Game

	action 1	action 2
action 1	1 \ 1	0 \ 0
action 2	0 \ 0	2 \ 2

  - Nash equilibrium is a pretty weak (but simple) solution concept
- Problem 2: how to find a Nash equilibrium?
  - Iteratively switch to each player's best response?
  - Counter-example: Rock–Paper–Scissors

	Rock	Paper	Scissors
Rock	0 \ 0	-1 \ 1	1 \ -1
Paper	1 \ -1	0 \ 0	-1 \ 1
Scissors	-1 \ 1	1 \ -1	0 \ 0

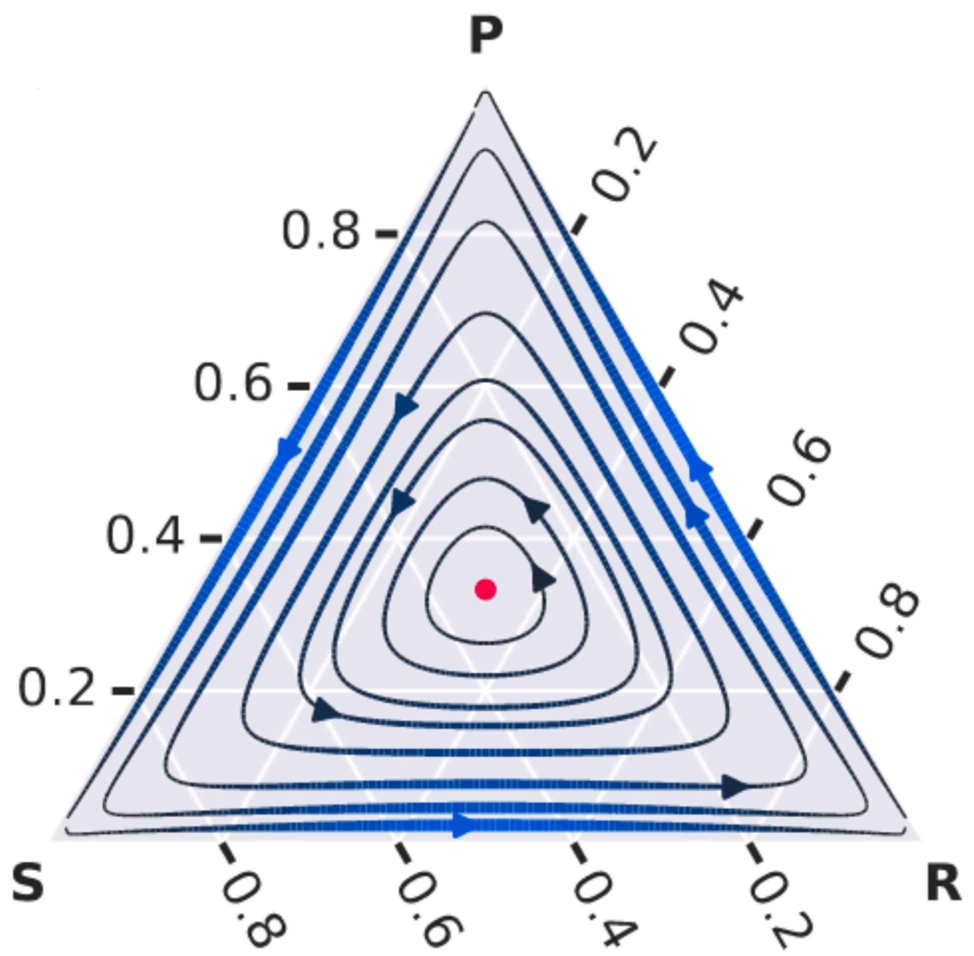
    - Best response can be deterministic; equilibrium may require stochastic

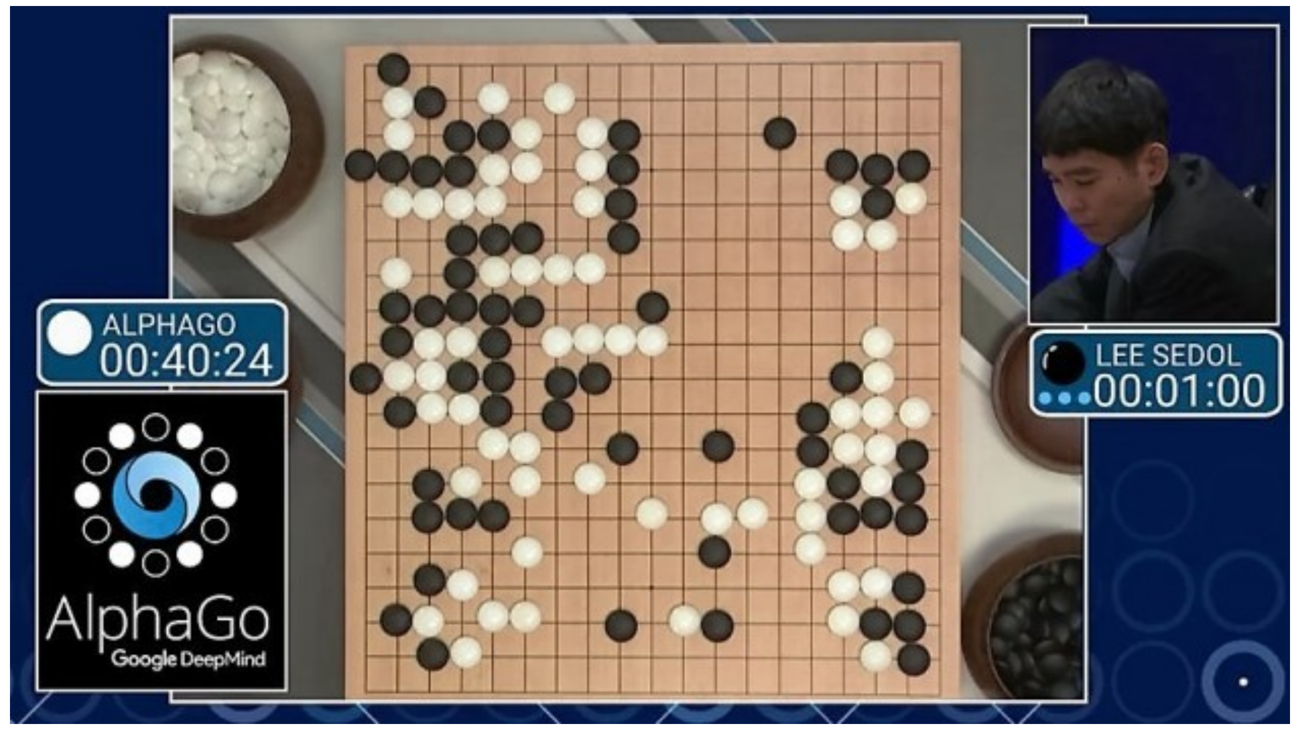
Two-player zero-sum games

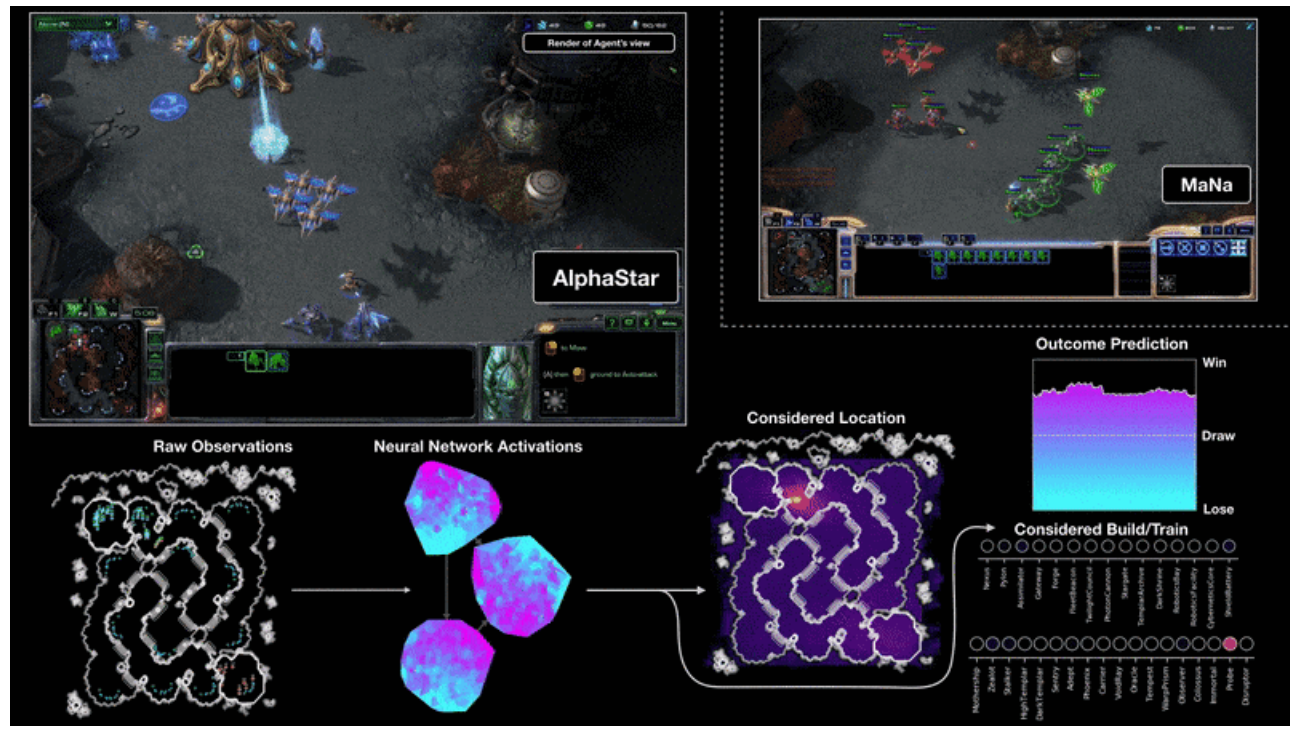
- Zero-sum: $r^1 = -r^2 = r$
- Optimization problem: $\max_{\pi^1} \min_{\pi^2} \mathbb{E}_{\pi^1,\pi^2}[R]$
  - Under mild conditions: max-min = min-max (no duality gap)
  - All Nash equilibria have the same value
- Very hard optimization problem
  - Gradient-based algorithms usually try to avoid a saddle-point
  - Here we're seeking a saddle-point

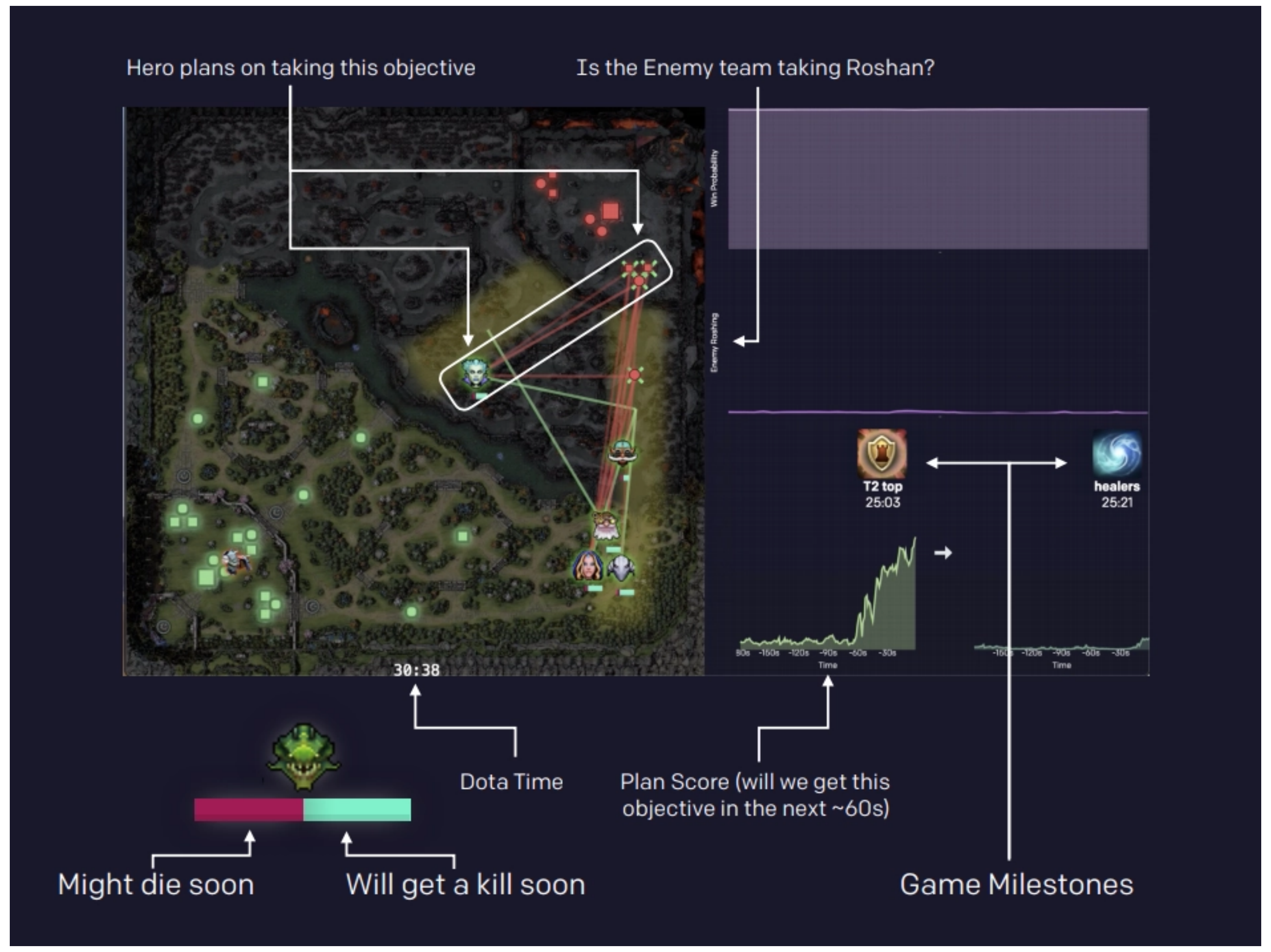
Self Play

- Self Play (= independent RL) = train each agent in MDP induced by others
- Problem: no guarantees of convergence to Nash equilibrium
  - E.g., not clear how to keep policies sufficiently stochastic
- But may work well in practice, particularly in games of skill









Fictitious Play (FP)

- Self Play has the right idea: if $b^i(\pi^{-i})$ is better than π^i $\Rightarrow$ update toward it
  - But by how much?
- Fictitious Play
  - Add $b^i(\pi^{-i})$ to a population Π
  - $\pi^i \leftarrow$ average of population
- π guaranteed to converge to Nash equilibrium

	Rock	Paper	Scissors
Pop. avg.	1	0	0
BR	0	1	0
Pop. avg.	0.5	0.5	0
BR	0	1	0
Pop. avg.	0.33	0.67	0
BR	0	0	1
Pop. avg.	0.25	0.5	0.25
BR	0	0	1
Pop. avg.	0.2	0.4	0.4
BR	1	0	0
Pop. avg.	0.33	0.33	0.33
BR	1	0	0

⋮

- How to implement this with (Deep) RL?

Neural Fictitious Self Play (NFSP)

- Representation: "best-response" values Q^i + "average" policies π^i
  - Use DQN to train Q^i against π^{-i}
    - Roll out episodes using $(\epsilon\text{-greedy}(Q^i), \pi^{-i}) \to$ replay buffer
    - Sample $(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$ from replay buffer $\to$ descend on square Bellman error
  - Use policy distillation (supervised learning) to average greedy(Q^i) into π^i

    when not exploring

    - Sample (s^i, a^i) from replay buffer $\to$ descend on NLL loss $-\log \pi^i(a^i \mid s^i)$
- Unlike FP, Q^i isn't immediately best response $\Rightarrow$ NFSP can be unstable

Today's lecture

Centralized vs. decentralized RL

(Fictitious) Self Play

Double Oracle

Double Oracle (DO)

- Unweighted population average guaranteed asymptotic convergence
  - Some policies are better than others (e.g. late vs. early in training) ⇒ weights?

normal form

- Assume payoffs / utilities given by matrix U_{π^1,π^2} for all $\pi^1 \in \Pi^1$, $\pi^2 \in \Pi^2$

- Idea: weight by mixed Nash equilibrium on population
  - $\sigma \leftarrow$ find Nash equilibrium restricted to population policies Π^i
  - Add best response to population: $\Pi^i \leftarrow \Pi^i \cup \{b^i(\sigma^{-i})\}$

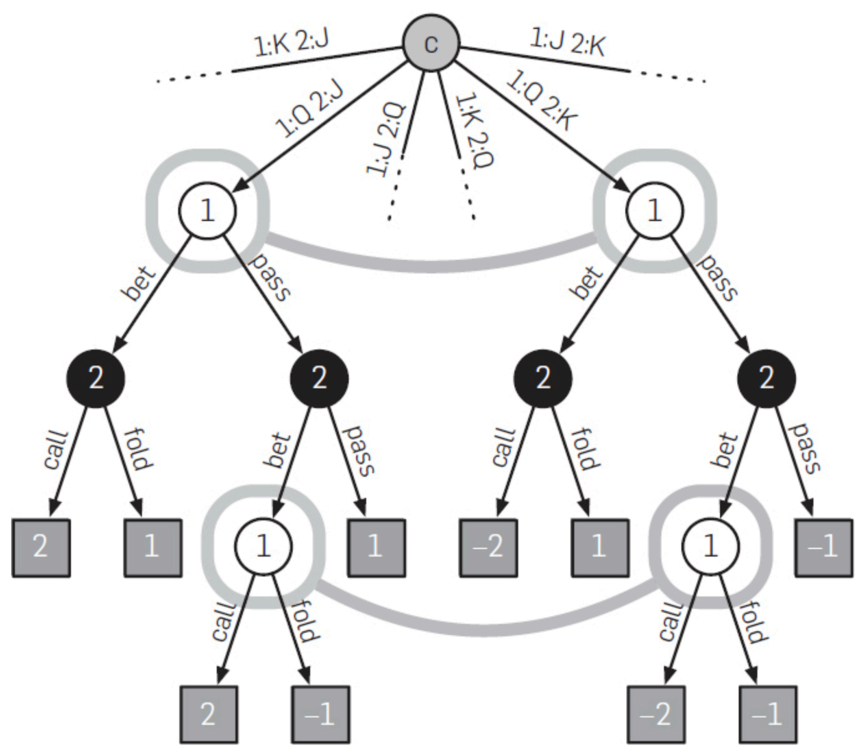
	R	P	S
Pop.NE	1	0	0
BR	0	1	0
Pop.NE	0	1	0
BR	0	0	1
Pop.NE	0.33	0.33	0.33

- Guarantee: $\sigma \rightarrow$ Nash equilibrium; hopefully before all policies added

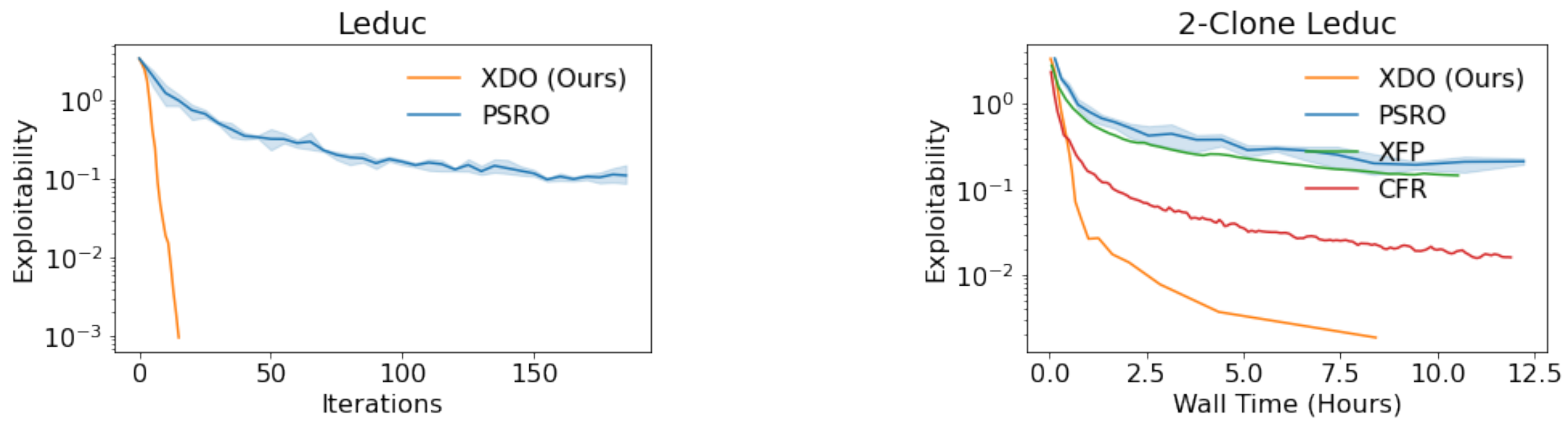
PSRO

- Problem: computing and storing entire utility matrix is infeasible in RL
  - Policy-space size is exponential in belief-space size $|\mathscr{A}|^{|\mathscr{B}|}$
- Idea: Policy-Space Response Oracles (PSRO)
  - Match pairs of population policies $\Rightarrow$ estimate $U_{\pi^1,\pi^2} = \mathbb{E}_{\pi^1,\pi^2}[R]$
  - Find meta-Nash equilibrium over population policies Π^i
  - $\Rightarrow$ meta-policy σ^i = mixture over Π^i
  - Use Deep RL to train best response to σ^{-i}, add to Π^i
- Guarantee: $\sigma \to$ Nash equilibrium; hopefully before all (many!!) policies added

Extensive-form Double Oracle (XDO)



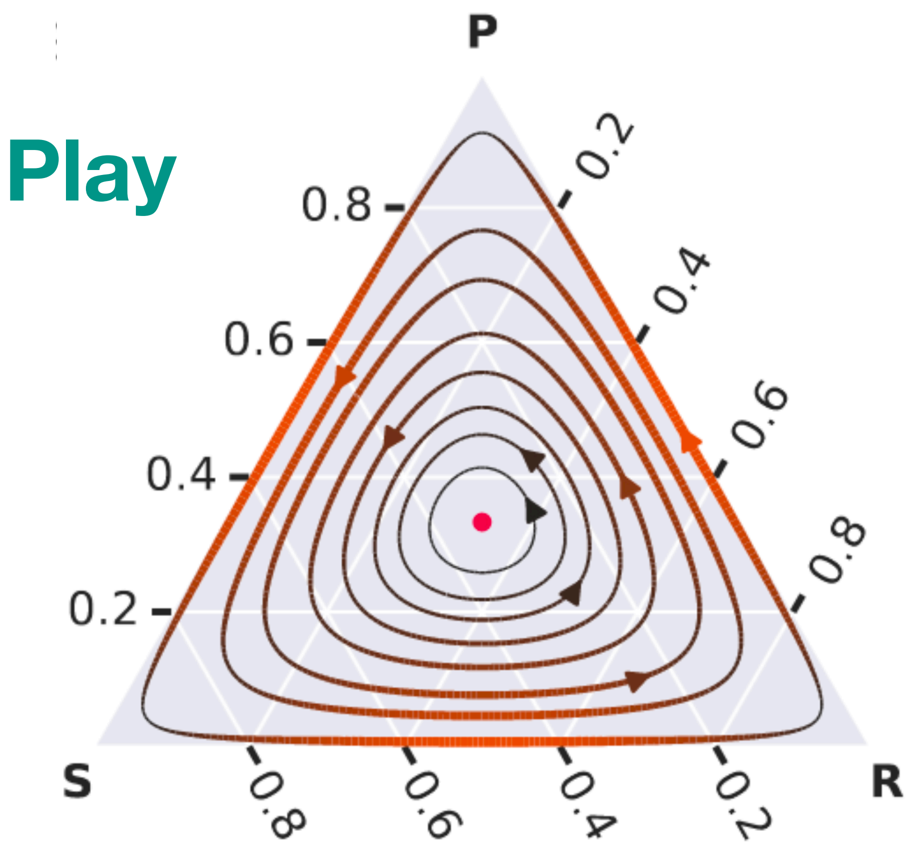
- Extensive form = tree of game histories
    - Information set (infostate) = states with same observable history
- Problem: in long game, mixing over few policies is very exploitable
    - Opponent can identify selected policy ⇒ it becomes deterministic, so exploitable
- Idea: mix over population policies again in every infostate
    - ⇔ extensive-form game restricted to actions by any population policy

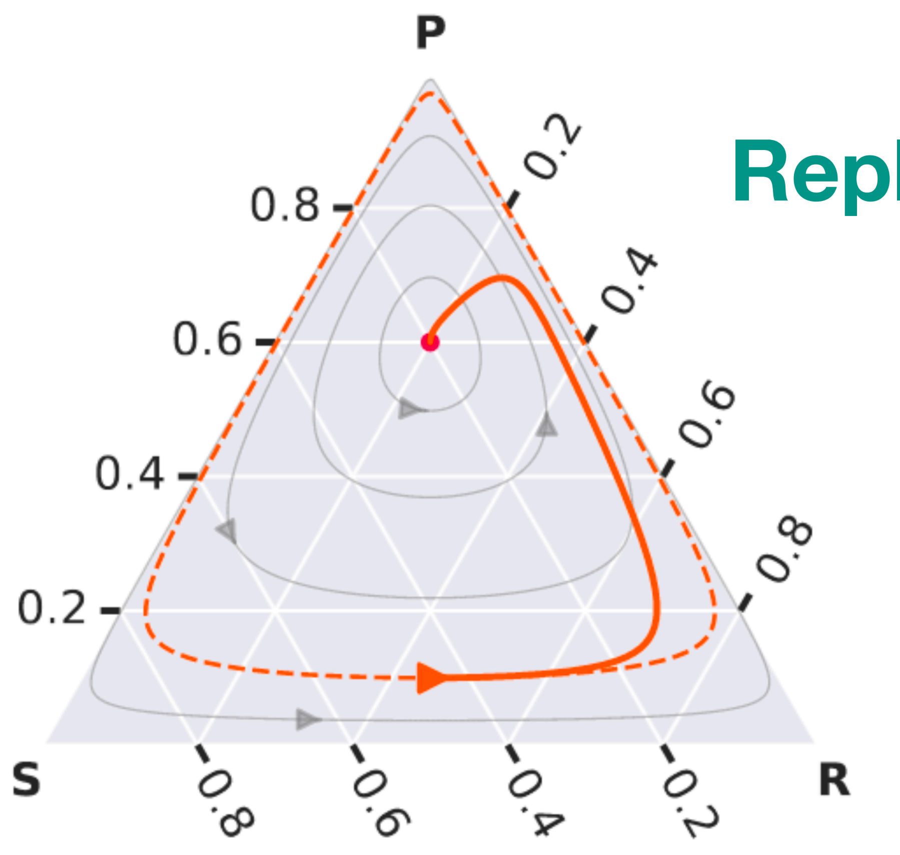


Other methods

- Counterfactual Regret Minimization (CFR)
  - In each episode: $\pi(a \mid h) \propto$ regret of not always taking a in infostate h
- Problem: in RL, we can't really get best responses
  - Idea: policy improvement dynamics that are guaranteed to converge
  - E.g. Replicator Dynamics (RD)

Self Play





Replicator Dynamics

General sum games: challenges

- Between zero-sum and cooperative: competitive + cooperative aspects
- May have multiple Nash equilibria ⇒ which is best? may be ill-defined
  - In one-shot game: which one will my opponent play? ill-defined
- >2 players (nothing special about 0-sum) ⇒ can have coalitions etc.
  - Mixed Nash equilibria exist, but very weak solution concept
  - No really great solution concept is known
- What to do? In practice, Self Play may work well

one critic per player
with shared o and a, but with r^i

  - Can also use MADDPG: $\nabla_{\theta_i}\mathscr{L}_i = \nabla_{\theta_i}\log \pi_{\theta_i}(a^i \mid o^i) Q^i(o, a)$

Recap

- Cooperative / general-sum games
  - ⇒ Self Play (aka independent RL), MADDPG
- Two-player zero-sum games
  - Self Play, MADDPG
  - Fictitious Play (FP), NFSP
  - Double Oracle (DO), PSRO, XDO
  - CFR, DeepCFR
  - Replicator Dynamics (RD), Neural RD (NeuRD)
  - Etc.