Multiagent Reinforcement Learning

Marc Lanctot

RLSS @ Lille, July 11th 2019

Joint work with many great collaborators!

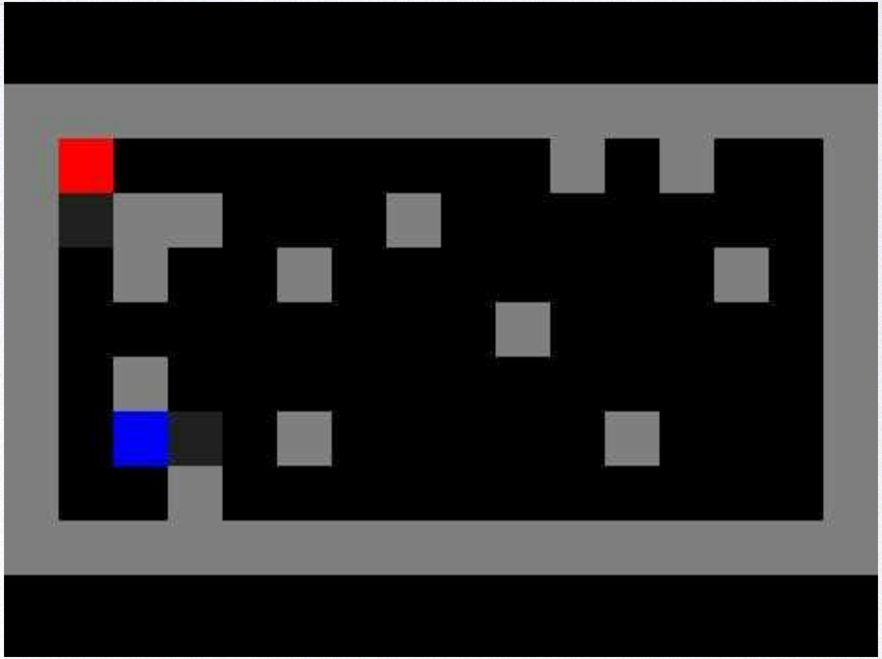
Talk plan

1. What is Multiagent Reinforcement Learning (MARL)?
2. Foundations & Background
3. Basic Formalisms & Algorithms
4. Advanced Topics

Part 1: What is MARL?

Multiagent Reinforcement Learning

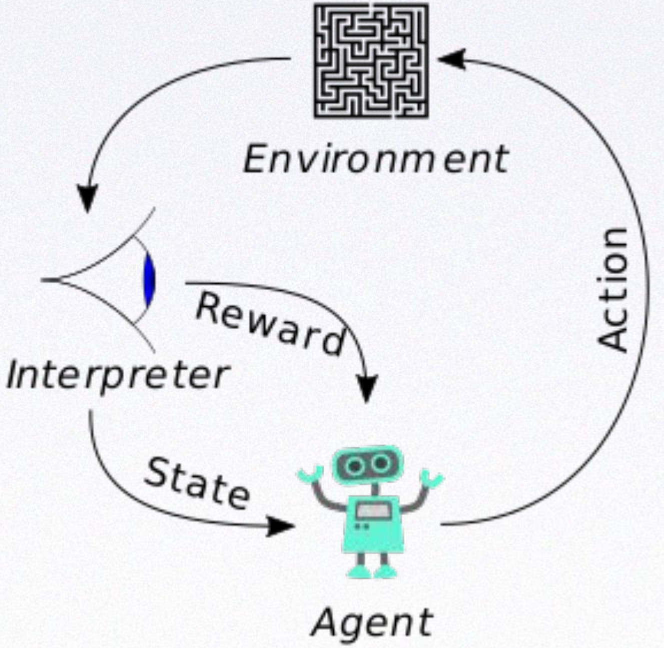
pommerman.com



Laser Tag

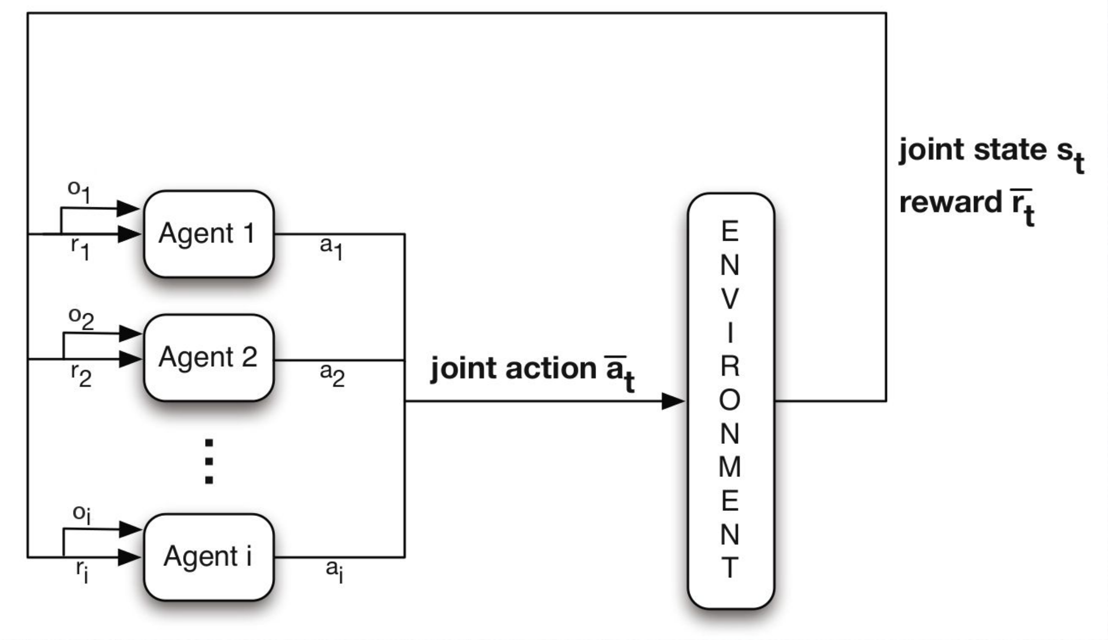
Multiagent Reinforcement Learning

Traditional (Single-Agent) RL



Source: Wikipedia

Multiagent Reinforcement Learning



Source: Nowe, Vrancx & De Hauwere 2012

Motivations: Research in Multiagent RL

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Approximate Solution Methods	Approximate Solution Methods
Small Problems	Tabular Solution Methods	Tabular Solution Methods

Motivations: Research in Multiagent RL

Sutton & Barto '98, '18

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Approximate Solution Methods	Approximate Solution Methods
Small Problems	Tabular Solution Methods	Tabular Solution Methods

Motivations: Research in Multiagent RL

First era of multiagent RL

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Approximate Solution Methods	Approximate Solution Methods
Small Problems	Tabular Solution Methods	Tabular Solution Methods

Motivations: Research in Multiagent RL

Multiagent Deep RL era ('16 - now)

Large Problems	Approximate Solution Methods	Approximate Solution Methods
Small Problems	Tabular Solution Methods	Tabular Solution Methods
	Single Agent	Multiple (e.g. 2) Agents

Motivations: Research in Multiagent RL

Talk focus

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Approximate Solution Methods	Approximate Solution Methods
Small Problems	Tabular Solution Methods	Tabular Solution Methods

Motivations: Research in Multiagent RL

My 10-year mission

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Approximate Solution Methods	Approximate Solution Methods
Small Problems	Tabular Solution Methods	Tabular Solution Methods

Important Historical Note

If multi-agent learning is the answer,
what is the question?

Yoav Shoham, Rob Powers, and Trond Grenager
Stanford University
{shoham,powers,grenager}@cs.stanford.edu

February 15, 2006

Artificial Intelligence, Volume 171, Issue 7

Foundations of multi-agent learning: Introduction to the special issue

Rakesh V. Vohra, Michael P. Wellman

Pages 363-364

An economist's perspective on multi-agent learning

Drew Fudenberg, David K. Levine

Pages 378-381

Perspectives on multiagent learning

Tuomas Sandholm

Pages 382-391

Artificial Intelligence, Volume 171, Issue 7

Agendas for multi-agent learning

Geoffrey J. Gordon

Pages 392-401

Multiagent learning is not the answer. It is the question

Peter Stone

Pages 402-405

What evolutionary game theory tells us about multiagent learning

Karl Tuyls, Simon Parsons

Pages 406-416

Artificial Intelligence, Volume 171, Issue 7

Multi-agent learning and the descriptive value of simple models

Ido Erev, Alvin E. Roth

Pages 423-428

The possible and the impossible in multi-agent learning

H. Peyton Young

Pages 429-433

No regrets about no-regret

Yu-Han Chang

Pages 434-439

Artificial Intelligence, Volume 171, Issue 7

A hierarchy of prescriptive goals for multiagent learning

Martin Zinkevich, Amy Greenwald, Michael L. Littman

Pages 440-447

Learning equilibrium as a generalization of learning to optimize

Dov Monderer, Moshe Tennenholtz

Pages 448-452

Some Specific Axes of MARL

Centralized:

- One brain / algorithm deployed across many agents

Decentralized:

- All agents learn individually
- Communication limitations defined by environment

Some Specific Axes of MARL

Prescriptive:

- Suggests how agents *should* behave

Descriptive:

- Forecast how agent *will* behave

Some Specific Axes of MARL

Cooperative: Agents cooperate to achieve a goal

Competitive: Agents compete against each other

Neither: Agents maximize their utility which may require cooperating and/or competing

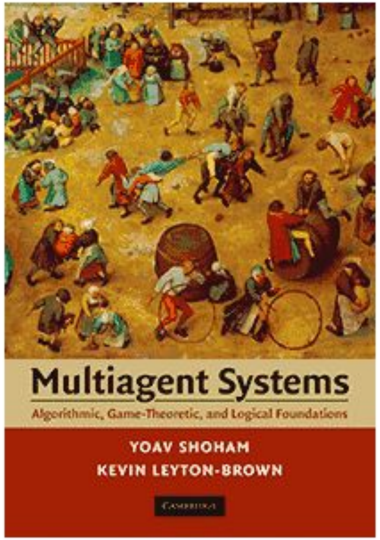
Our Focus Today

1. Centralized training for decentralized execution (very common)
2. Mostly prescriptive
3. Mostly competitive; sprinkle of cooperative and neither

Part 2: Foundations & Background

Shoham & Leyton-Brown '09

Main Page Table of Contents Instructional Resources Errata eBook Download new!



Multiagent Systems
Algorithmic, Game-Theoretic, and Logical Foundations

Yoav Shoham
Stanford University
Kevin Leyton-Brown
University of British Columbia

Cambridge University Press, 2009
Order online: amazon.com

masfoundations.org

Foundations of (MA)RL

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Reinforcement Learning	Multiagent Reinforcement Learning
Small Problems	Approximate Dynamic Programming	Game Theory

Foundations of Multiagent RL

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Reinforcement Learning	Multiagent Reinforcement Learning
Small Problems	Approximate Dynamic Programming	**Game Theory**

Biscuits vs Cookies

A Note on Terminology

Player	Agent
Game	Environment
Strategy	Policy
Best Response	Greedy Policy
Utility	Reward
State	(Information) State

Normal-form "One-Shot" Games

- Set of **players** $i \in \mathcal{N} = \{1, 2, \cdots, n\}$

Normal-form “One-Shot” Games

- Set of **players** $i \in \mathcal{N} = \{1, 2, \cdots, n\}$
- Each player has set of **actions** $\mathcal{A}_i \in \{a_1, a_2, \ldots\}$

Normal-form "One-Shot" Games

- Set of **players** $i \in \mathcal{N} = \{1, 2, \cdots, n\}$
- Each player has set of **actions** $\mathcal{A}_i \in \{a_1, a_2, \ldots\}$
- Set of **joint** actions $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$

Normal-form "One-Shot" Games

- Set of **players** $i \in \mathcal{N} = \{1, 2, \cdots, n\}$
- Each player has set of **actions** $\mathcal{A}_i \in \{a_1, a_2, \ldots\}$
- Set of **joint** actions $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$
- A **utility** function $u : \mathcal{N} \times \mathcal{A} \rightarrow U \subseteq \Re$

Example: (Bi-)Matrix Games (n = 2)

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

Example: (Bi-)Matrix Games (n = 2)

column player

actions

row player

	A	B
a	0 , 0	1 , -1
b	-1 , 1	0, 0

Example: (Bi-)Matrix Games (n = 2)

			column player	
			A	B
row player	a		0 , 0	1 , -1
	b		-1 , 1	0, 0

utility to row player

Example: (Bi-)Matrix Games (n = 2)

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

utility to column player

utility to row player

Example: (Bi-)Matrix Games (n = 2)

column player

utility to column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

for joint action (a,B)

utility to row player

Normal-form "One-Shot" Games

- Set of **players** $i \in \mathcal{N} = \{1, 2, \cdots, n\}$
- Each player has set of **actions** $\mathcal{A}_i \in \{a_1, a_2, \ldots\}$
- Set of **joint** actions $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$
- A **utility** function $u : \mathcal{N} \times \mathcal{A} \rightarrow U \subseteq \Re$

Each player: $\pi_i \in \Delta(\mathcal{A}_i)$, maximize $\mathbb{E}_{a \sim \pi}[u_i(a)]$

Normal-form "One-Shot" Games

- Set of **players** $i \in \mathcal{N} = \{1, 2, \cdots, n\}$
- Each player has set of **actions** $\mathcal{A}_i \in \{a_1, a_2, \ldots\}$
- Set of **joint** actions $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$
- A **utility** function $u : \mathcal{N} \times \mathcal{A} \rightarrow U \subseteq \Re$

Each player: $\pi_i \in \Delta(\mathcal{A}_i), \;\; \text{maximize } \mathbb{E}_{a \sim \pi}[u_i(a)]$

Problem! This is a *joint* policy

Best Response

Suppose we are player i and we fix policies of other players

Best Response

Suppose we are player i and we fix policies of other players ($-i = \mathcal{N} - \{i\}$)

Best Response

Suppose we are player i and we fix policies of other players $(-i = \mathcal{N} - \{i\})$

$$\pi_i \in \Delta(\mathcal{A}_i), \ \ \text{maximize } \mathbb{E}_{a \sim \pi}[u_i(a)]$$

Best Response

Suppose we are player i and we fix policies of other players $(-i = \mathcal{N} - \{i\})$

$$\pi_i \in \Delta(\mathcal{A}_i), \;\; \text{maximize } \mathbb{E}_{a \sim \pi}[u_i(a)]$$

$$\pi_i \in BR(\pi_{-i}) \Leftrightarrow u_i(\pi_i, \pi_{-i}) = \max_{\pi'_i} \mathbb{E}_{a \sim (\pi'_i, \pi_{-i})}[u_i(a)]$$

Best Response

Suppose we are player i and we fix policies of other players ($-i = \mathcal{N} - \{i\}$)

$$\pi_i \in \Delta(\mathcal{A}_i), \;\; \text{maximize} \; \mathbb{E}_{a \sim \pi}[u_i(a)]$$

$$\pi_i \in BR(\pi_{-i}) \Leftrightarrow u_i(\pi_i, \pi_{-i}) = \max_{\pi'_i} \mathbb{E}_{a \sim (\pi'_i, \pi_{-i})}[u_i(a)]$$

π_i is a **best response** to π_{-i}

Solving a Matrix Game

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

Solving a Matrix Game

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

Let’s start here

Solving a Matrix Game

		column player A	column player B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

Both players have *incentive to deviate*
(assuming the opponent stays fixed)

Solving a Matrix Game

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

Solving a Matrix Game

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

Solving a Matrix Game

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

(a,A) is a *fixed point* of this process

Solving a Matrix Game

column player

		A	B
row player	a	0 , 0	1 , -1
	b	-1 , 1	0, 0

(a,A) is a *fixed point* of this process

$$\pi_i \in \Delta(\mathcal{A}_i), \ \text{maximize } \mathbb{E}_{a\sim\pi}[u_i(a)]$$

Let's Try Another....

column player

		A	B
row player	a	1 , -1	-1 , 1
	b	-1 , 1	1, -1

Let's Try Another….

column player

row player

	A	B
a	1 , -1	-1 , 1
b	-1 , 1	1, -1

Nash equilibrium

A Nash equilibrium is a **joint policy** π such that no player has incentive to deviate *unilaterally*.

Nash equilibrium: A Solution Concept

A Nash equilibrium is a **joint policy** π such that no player has incentive to deviate *unilaterally*.

$$\forall i \in \mathcal{N}, \pi_i \in BR(\pi_{-i})$$

Some Facts

- Nash equilibrium always exists in finite games
- Computing a Nash eq. is PPAD-Complete
  - One solution is to focus on tractable subproblems
  - Another is to compute approximations
- Assumes players are (unbounded) rational
- Assumes knowledge:
  - Utility / value functions
  - Rationality assumption is common knowledge

Two-Player Zero-Sum Games

Matching Pennies: $u_1(\cdot) = -u_2(\cdot)$

column player

row player

	A	B
a	1 , -1	-1 , 1
b	-1 , 1	1, -1

Two-Player Zero-Sum Games

Matching Pennies: $u_1(\cdot) = -u_2(\cdot)$

$$\max V$$

column player

		A	B
row player	**a**	**1** , -1	**-1** , 1
	b	**-1** , 1	**1**, -1

Two-Player Zero-Sum Games

Matching Pennies: $u_1(\cdot) = -u_2(\cdot)$

column player

		A	B
row player	**a**	**1** , -1	**-1** , 1
	b	**-1** , 1	**1**, -1

$$\max V$$

$$\pi(a) - \pi(b) \geq V \quad \text{(vs. A)}$$

Two-Player Zero-Sum Games

Matching Pennies: $u_1(\cdot) = -u_2(\cdot)$

column player

		A	B
row player	**a**	**1** , -1	**-1** , 1
	b	**-1** , 1	**1**, -1

$$\max V$$

$$\pi(a) - \pi(b) \geq V \quad (\text{vs. A})$$
$$-\pi(a) + \pi(b) \geq V \quad (\text{vs. B})$$

Two-Player Zero-Sum Games

Matching Pennies: $u_1(\cdot) = -u_2(\cdot)$

column player

		A	B
row player	**a**	**1** , -1	**-1** , 1
	b	**-1** , 1	**1**, -1

$$\max V$$

$$\pi(a) - \pi(b) \geq V \quad \text{(vs. A)}$$
$$-\pi(a) + \pi(b) \geq V \quad \text{(vs. B)}$$

$$\pi(a) + \pi(b) = 1$$

$$0 \leq \pi(a), \pi(b) \leq 1$$

Best Response Condition

For any (possibly stochastic) joint policy π_{-i} ,

There exists a **deterministic** best response:

$$\pi_i^b \in BR(\pi_{-i})$$

Best Response Condition

For any (possibly stochastic) joint policy π_{-i} ,

There exists a **deterministic** best response:

$$\pi_i^b \in BR(\pi_{-i})$$

Proof: Assume otherwise. The values of each deterministic policy (action) must be the same, by def. of BR. Then we can put full weight on any of them.

Two-Player Zero-Sum Games

Matching Pennies: $u_1(\cdot) = -u_2(\cdot)$

column player

row player

	A	B
a	**1** , -1	**-1** , 1
b	**-1** , 1	**1**, -1

$$\max V$$

$$\pi(a) - \pi(b) \geq V \quad \text{(vs. A)}$$
$$-\pi(a) + \pi(b) \geq V \quad \text{(vs. B)}$$

$$\pi(a) + \pi(b) = 1$$

$$0 \leq \pi(a), \pi(b) \leq 1$$

This is a Linear Program!

- Solvable in polynomial time (!)
  - Easy to apply off-the-shelf solvers
- Will find one solution
- Matching Pennies: $\pi(a) = \pi(b) = \frac{1}{2}, V = 0$

Minimax

John von Neumann 1928

Max-min: P1 looks for a π_1 such that

$$v_1 = \max_{\pi_1} \min_{\pi_2} u_1(\pi_1, \pi_2)$$

Min-max: P1 looks for a π_1 such that

$$v_1 = \min_{\pi_2} \max_{\pi_1} u_1(\pi_1, \pi_2)$$

In two-player, zero-sum these are the same!

---> The Minimax Theorem

Consequences of Minimax

The optima $\pi^* = (\pi_1^*, \pi_2^*)$

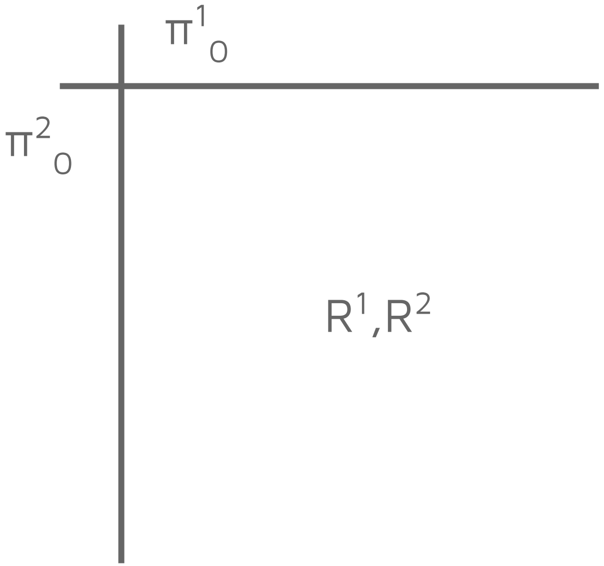
- These exist! (They sometimes might be stochastic.)
- Calles a minimax-optimal joint policy. Also, a Nash equilibrium.
- They are **interchangeable:**

$$\forall \pi^*, \pi^{*\prime} \Rightarrow (\pi_1^*, \pi_2^{*\prime}), (\pi_1^{*\prime}, \pi_2^*) \quad \text{also minimax-optimal}$$

- Each policy is a best response to the other.

Normal Form Games: Algorithms

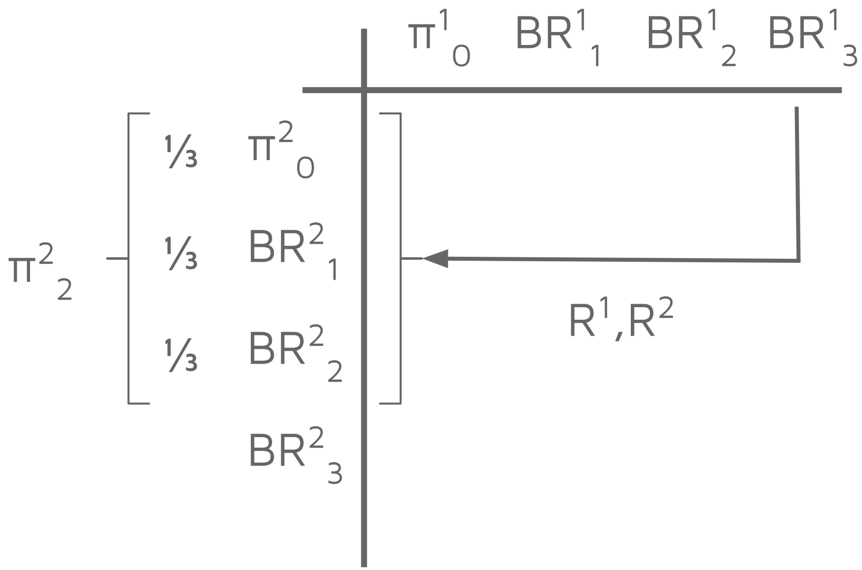
- Fictitious Play:



- Start with an arbitrary policy per player (π^1_0, π^2_0),

Normal Form Games: Algorithms

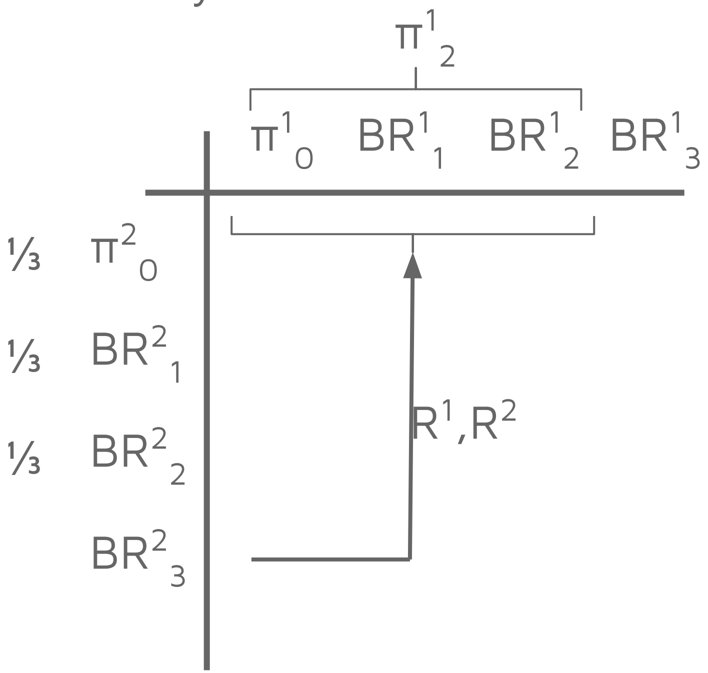
- Fictitious Play:



- Start with an arbitrary policy per player (π^1_0,π^2_0),
    - Then, play best response against a uniform distribution over the past policy of the opponent (BR^1_n,BR^2_n).

Normal Form Games: Algorithms

- Fictitious Play:



- Start with an arbitrary policy per player (π^1_0, π^2_0),
  - Then, play best response against a uniform distribution over the past policy of the opponent (BR^1_n, BR^2_n).

Normal Form Games: Algorithms

- Fictitious Play:
- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)

	R
R	0

Normal Form Games: Algorithms

- Fictitious Play:

	R	P
R	0	1
P	-1	0

- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)
- Iteration 1:
  - BR^1_1, BR^2_1 = P, P
  - (½, ½, 0), (½, ½, 0)

Normal Form Games: Algorithms

- Fictitious Play:

	R	P	P
R	0	1	1
P	-1	0	0
P	-1	0	0

- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)
- Iteration 1:
    - $BR^1{}_1, BR^2{}_1$ = P, P
    - (½, ½, 0), (½, ½, 0)
- Iteration 2:
    - $BR^1{}_2, BR^2{}_2$ = P, P
    - (⅓, ⅔, 0), (⅓, ⅔, 0)

Normal Form Games: Algorithms

- Fictitious Play:

	R	P	P	S
R	0	1	1	-1
P	-1	0	0	1
P	-1	0	0	1
S	1	-1	-1	0

- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)
- Iteration 1:
  - BR^1_1, BR^2_1 = P, P
  - (½, ½, 0), (½, ½, 0)
- Iteration 2:
  - BR^1_2, BR^2_2 = P, P
  - (⅓, ⅔, 0), (⅓, ⅔, 0)
- Iteration 3:
  - BR^1_3, BR^2_3 = S, S
  - (¼,½,¼), (¼,½,¼)

Normal Form Games: Algorithms

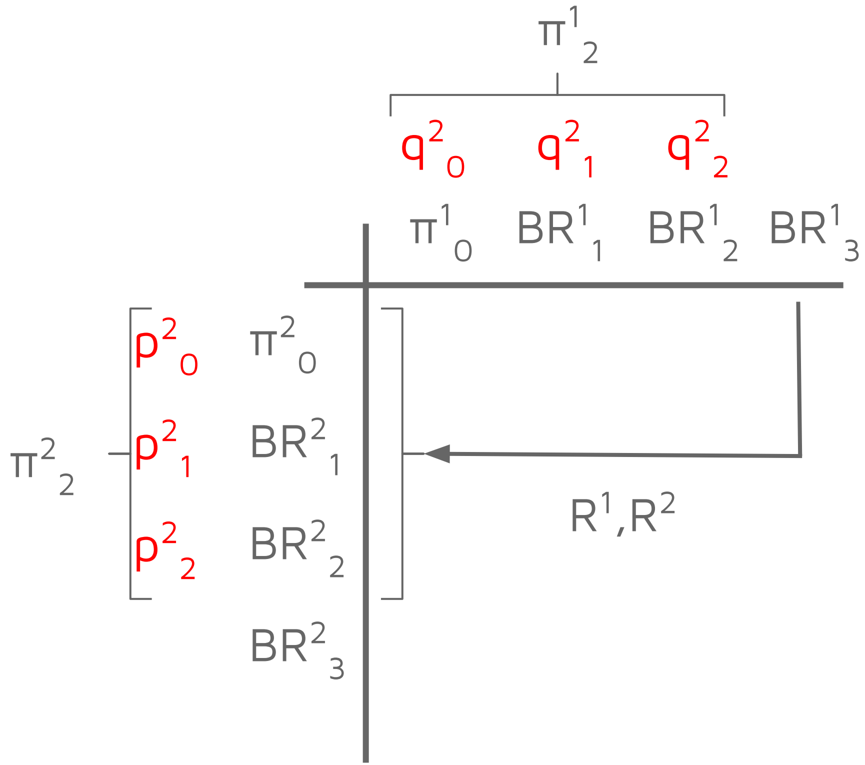
- Fictitious Play:

	R	P	P	S	S
R	0	1	1	-1	-1
P	-1	0	0	1	1
P	-1	0	0	1	1
S	1	-1	-1	0	0
S	1	-1	-1	0	0

- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)
- Iteration 1:
  - BR^1_1, BR^2_1 = P, P
  - (½, ½, 0), (½, ½, 0)
- Iteration 2:
  - BR^1_2, BR^2_2 = P, P
  - (⅓, ⅔, 0), (⅓, ⅔, 0)
- Iteration 3:
  - BR^1_3, BR^2_3 = S, S
  - (¼,½,¼), (¼,½,¼)

Normal Form Games: Algorithms

- double oracle [HB McMahan 2003]:



- Start with an arbitrary policy per player (π^1_0,π^2_0),
  - Compute (p^n,q^n) by solving the game at iteration n
  - Then, best response against (p^n,q^n) and get a new best response (BR^1_n,BR^1_n).

Normal Form Games: Algorithms

- double oracle:
- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)

	R
R	0

Normal Form Games: Algorithms

- double oracle:

	R	P
R	0	1
P	-1	0

- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)
- Iteration 1:
  - BR^1_1, BR^2_1 = P, P
  - Solve the game : (0, 1, 0), (0, 1, 0)

Normal Form Games: Algorithms

- double oracle:

	R	P	S
R	0	1	-1
P	-1	0	1
S	1	-1	0

- Start with (R, P, S)= (1, 0, 0), (1, 0, 0)
- Iteration 1:
  - BR^1_1, BR^2_1 = P, P
  - Solve the game : (0, 1, 0), (0, 1, 0)
- Iteration 2:
  - BR^1_2, BR^2_2 = S, S
  - (⅓, ⅓, ⅓), (⅓, ⅓, ⅓)

Cooperative Games

$$u_i(\cdot) = u_j(\cdot)$$

column player

row player

	A	B	C
a	1, 1	0, 0	0, 0
b	0, 0	2, 2	0, 0
c	0, 0	0, 0	5, 5

Cooperative Games

$$u_i(\cdot) = u_j(\cdot)$$

column player

row player

	A	B	C
a	1, 1	0, 0	0, 0
b	0, 0	2, 2	0, 0
c	0, 0	0, 0	5, 5

These are all Nash equilibria!

General-Sum Games

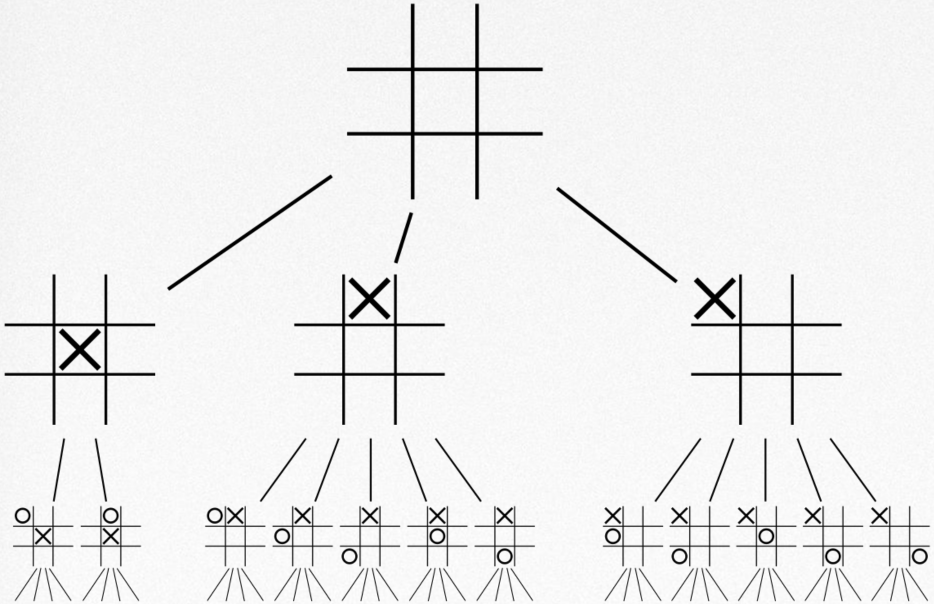
No constraints on utilities!

		column player A	column player B
row player	a	3, 2	0, 0
row player	b	0, 0	2, 3

The Sequential Setting: Extensive-Form Games

What about sequential games…?

Perfect Information Games



(Finite) Perfect Information Games: Model

- Start with an *episodic* MDP

(Finite) Perfect Information Games: Model

- Start with an *episodic* MDP
- Add a **player identity** function:

$$\tau(s) \in \mathcal{N} \cup \{s\}$$

Simultaneous move node (many players play simultaneously)

(Finite) Perfect Information Games: Model

- Start with an *episodic* MDP
- Add a **player identity** function:

$$\tau(s) \in \mathcal{N} \cup \{s\}$$

- Define rewards *per player*:

$$r_i(s, a, s') \text{ for } i \in \mathcal{N}$$

(Finite) Perfect Information Games: Model

- Start with an *episodic* MDP
- Add a **player identity** function:

$$\tau(s) \in \mathcal{N} \cup \{s\}$$

- Define rewards *per player*:

$$r_i(s, a, s') \text{ for } i \in \mathcal{N}$$

- (Similarly for returns: $G_{t,i}$ is the return *to* player i from s_t)

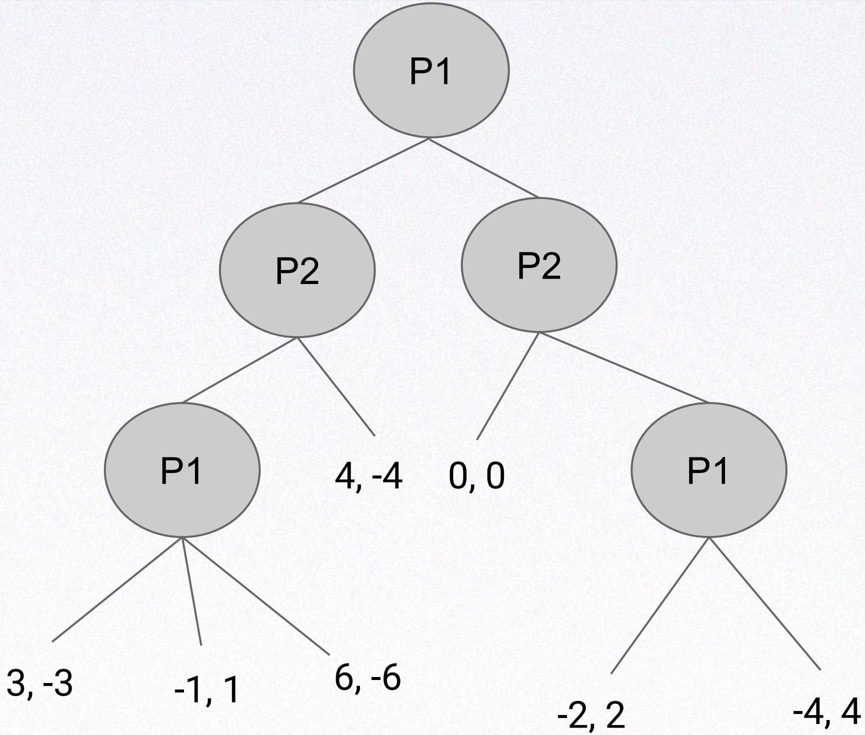
Part 3: Basic Formalisms & Algorithms

Foundations of RL

	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Reinforcement Learning	Multiagent Reinforcement Learning
Small Problems	Approximate Dynamic Programming	Game Theory

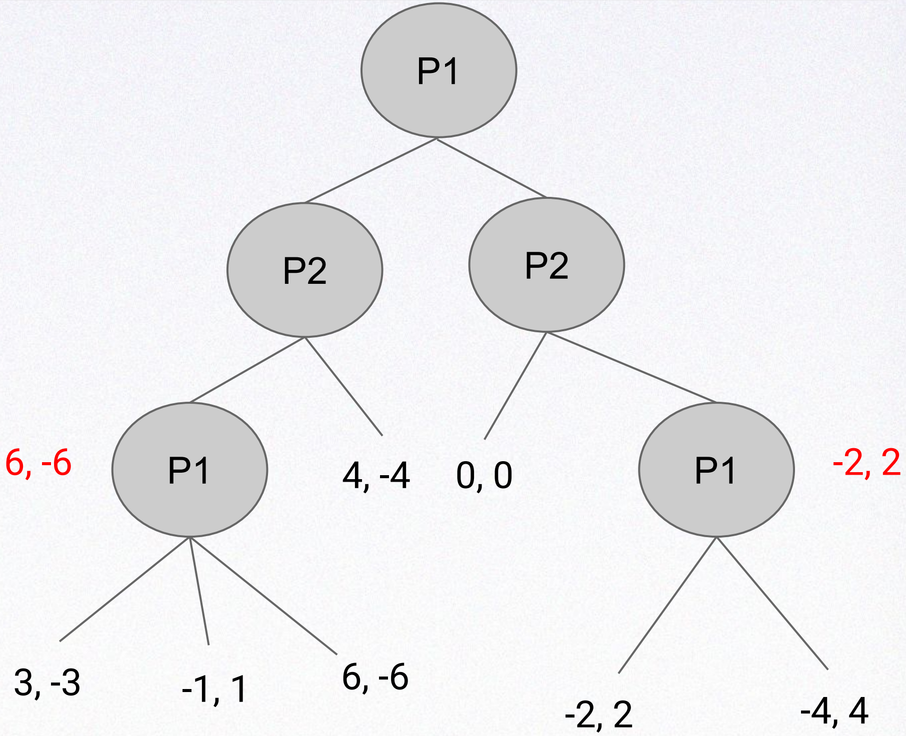
Backward Induction

Solving a *turn-taking* perfect information game



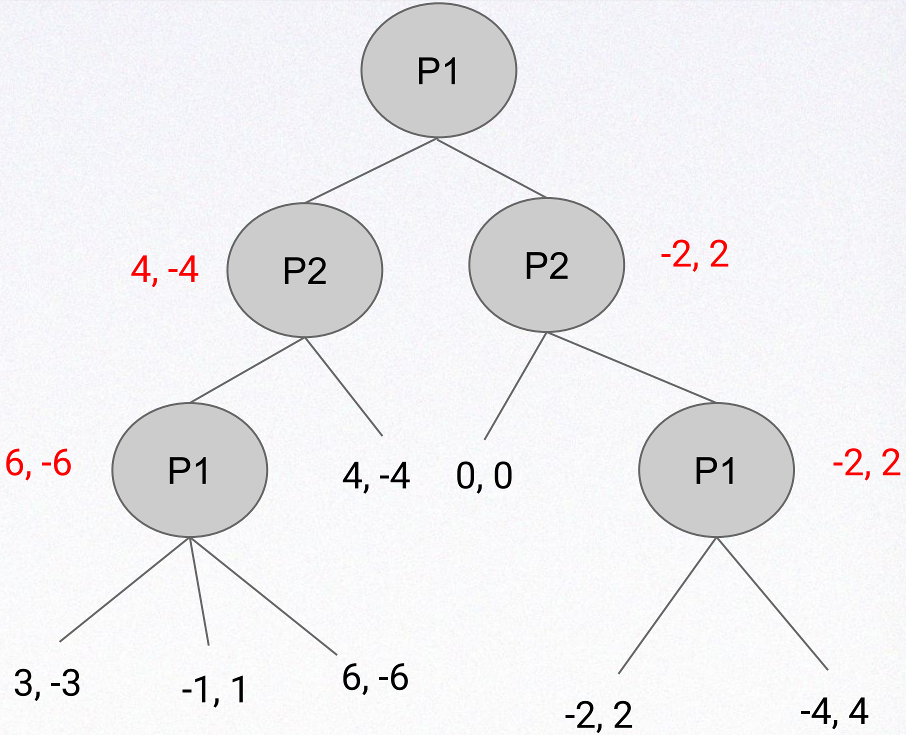
Backward Induction

Solving a *turn-taking* perfect information game



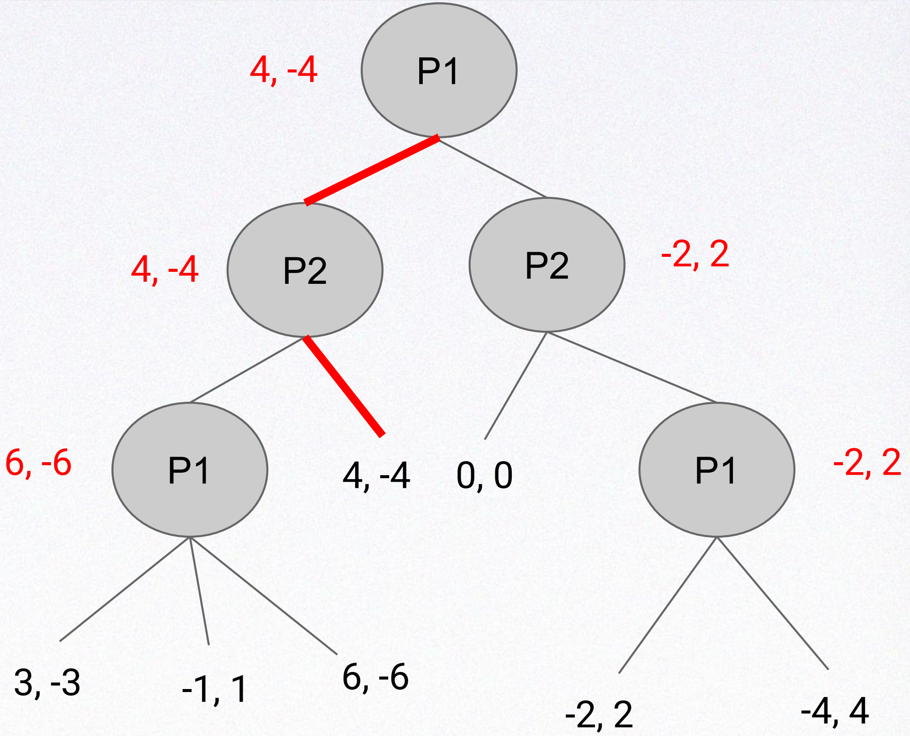
Backward Induction

Solving a *turn-taking* perfect information game



Backward Induction

Solving a *turn-taking* perfect information game



Intro to RL: Tabular Approximate Dyn. Prog.

Value iteration

Initialize array V arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)

Repeat
  $\Delta \leftarrow 0$
  For each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \max_a \sum_{s',r} p(s', r | s, a)\big[r + \gamma V(s')\big]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, $\pi \approx \pi_*$, such that
  $\pi(s) = \arg\max_a \sum_{s',r} p(s', r | s, a)\big[r + \gamma V(s')\big]$

Turn-Taking 2P Zero-sum Perfect Info. Games

- Player to play at s: $\tau(s)$
- Reward to player i: r_i
- Subset of legal actions $\textsc{LegalActions}(s)$
- Often assume episodic and $\gamma = 1$

Values of a state **to player i**: $V_i(s)$

Identities:

$$\forall s, a, s' : r_1 = -r_2, \quad V_1(s) = -V_2(s)$$

2P Zero-Sum Perfect Info. Value Iteration

Value iteration

Initialize array V_i arbitrarily (e.g., $V_i(s) = 0$ for all $s \in \mathcal{S}^+$)

Repeat

$\Delta \leftarrow 0$

For each $s \in \mathcal{S}$: *Let $i = t(s)$*

$v \leftarrow V_i(s)$

$V_i(s) \leftarrow \max_a \sum_{s', r_i} p(s', r_i | s, a) [r_i + \gamma V_i(s')]$

$\Delta \leftarrow \max(\Delta, |v - V_i(s)|)$

until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, $\pi \approx \pi_*$, such that *$i = t(s)$*

$\pi(s) = \operatorname{argmax}_a \sum_{s', r_i} p(s', r_i | s, a) [r_i + \gamma V_i(s')]$

Minimax

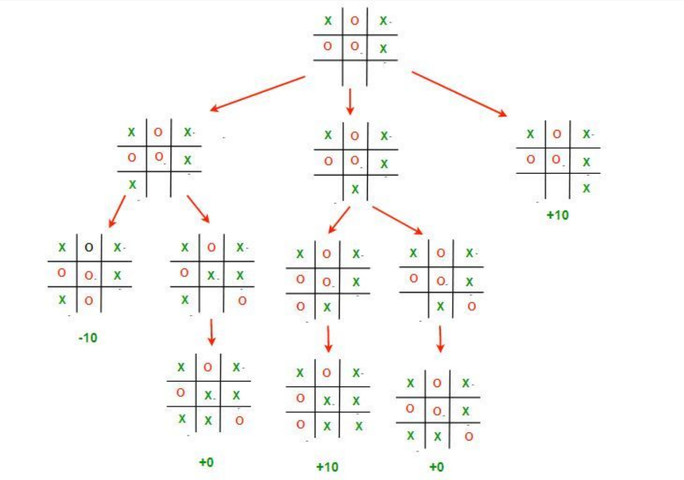
A.K.A. Alpha-Beta, Backward Induction, Retrograde Analysis, etc...

Start from search state s,

Compute a depth-limited approximation:

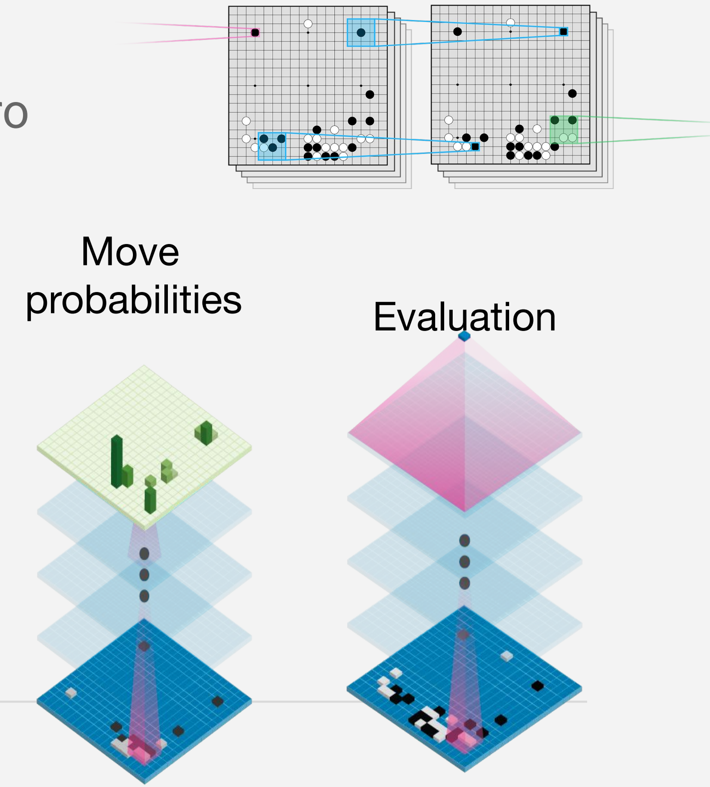
$$V_{i,d}(s) = \begin{cases} u_i(s) & \text{if } s \text{ is terminal,} \\ h_i(s) & \text{if } d = 0, \\ \sum_{s'} p(s,a,s')V_{i,d-1}(s') & \text{otherwise.} \end{cases}$$

---> Minimax Search



Two-Player Zero-Sum Policy Iteration

- Analogous to adaptation of value iteration
- Foundation of AlphaGo, AlphaGo Zero, AlphaZero
  - Better policy improvement via MCTS
  - Deep network func. approximation
    - Policy prior cuts down *breadth*
    - Value network cuts the *depth*



2P Zero-Sum Games with Simultaneous Moves

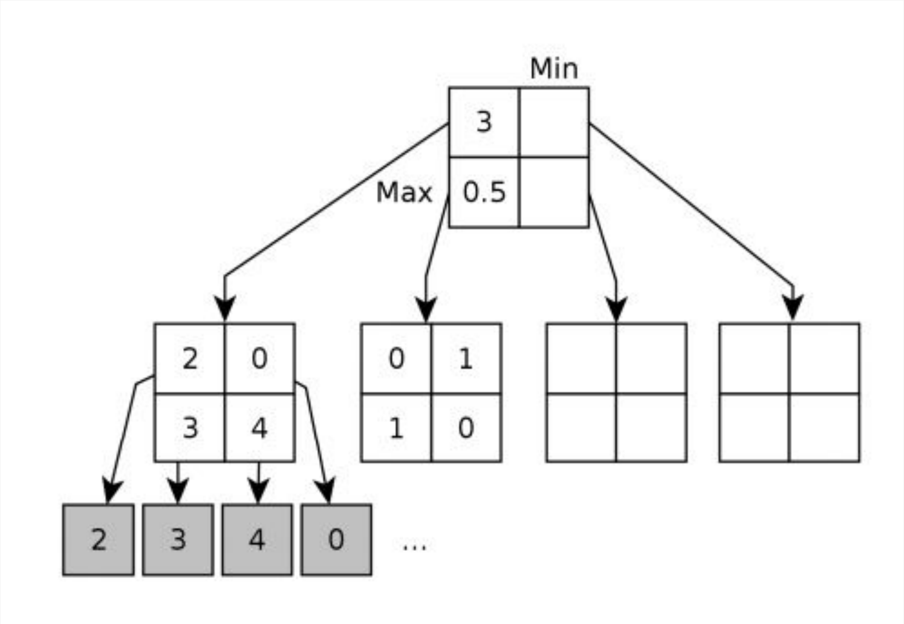


Image from Bozansky et al. 2016

Markov Games

“Markov Soccer”

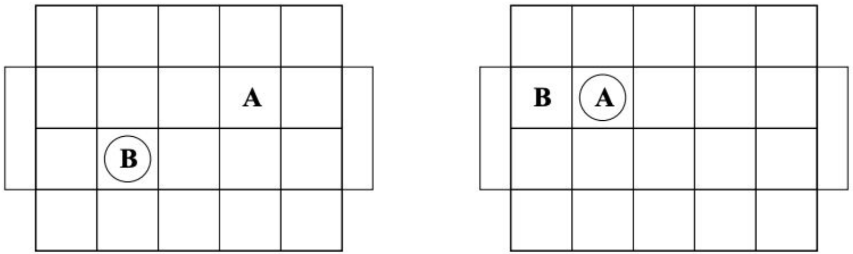
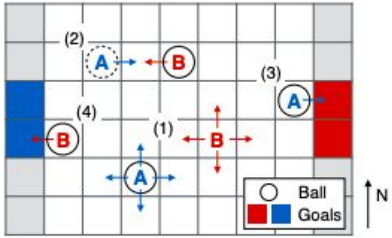


Figure 2: An initial board (left) and a situation requiring a probabilistic choice for A (right).



	Defensive	Offensive
w/ ball	Avoid opponent	Advance to goal
w/o ball	Defend goal	Intercept the ball

Figure 3. Left: Illustration of the soccer game. *Right:* Strategies of the hand-crafted rule-based agent.

Littman ‘94

He et al. ‘16

Also: Lagoudakis & Parr ‘02, Uther & Veloso ‘03, Collins ‘07

Value Iteration for Zero-Sum Markov Games

Value iteration

Initialize array V arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)

Repeat
  $\Delta \leftarrow 0$
  For each $s \in \mathcal{S}$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \cancel{\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]}$ $\min_{\pi_2(s)} \max_{\pi_1(s)} \mathbb{E}_{a \sim \pi(s), s'}[r_1(s,a,s') + \gamma V_1(s')]$
    $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)

Output a ~~deterministic~~ policy, $\pi \approx \pi_*$, ~~such that~~ computed above
  $\cancel{\pi(s) = \arg\max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]}$

First MARL Algorithm: Minimax-Q (Littman '94)

1. Start with arbitrary joint value functions $q(s, a, o)$

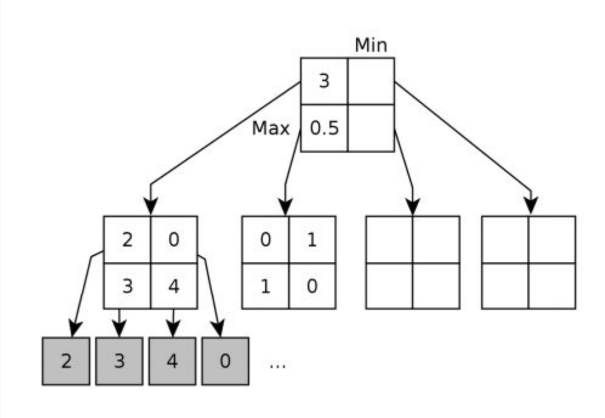
my action

opponent action

DeepMind

First MARL Algorithm: Minimax-Q (Littman '94)

1. Start with arbitrary joint value functions $q(s, a, o)$



my action

opponent action

Induces a matrix of values

First MARL Algorithm: Minimax-Q (Littman '94)

1. Start with arbitrary joint value functions $q(s, a, o)$
2. Define policy π as in value iteration (by solving an LP)

First MARL Algorithm: Minimax-Q (Littman '94)

1. Start with arbitrary joint value functions $q(s, a, o)$
2. Define policy π as in value iteration (by solving an LP)
3. Generate trajectories of tuple (s, a, o, s') using behavior policy $\pi' = \epsilon \textsc{Unif}(\mathcal{A}) + (1 - \epsilon)\pi$

First MARL Algorithm: Minimax-Q (Littman '94)

1. Start with arbitrary joint value functions $q(s, a, o)$
2. Define policy π as in value iteration (by solving an LP)
3. Generate trajectories of tuple (s, a, o, s') using behavior policy $\pi' = \epsilon \textsc{Unif}(\mathcal{A}) + (1-\epsilon)\pi$
4. Update $q(s, a, o) = (1-\alpha)q(s, a, o) + \alpha(r(s, a, o, s') + \gamma v(s'))$

First Era of MARL

Follow-ups to Minimax Q:

- Friend-or-Foe Q-Learning (Littman ‘01)
- Correlated Q-learning (Greenwald & Hall ‘03)
- Nash Q-learning (Hu & Wellman ‘03)
- Coco-Q (Sodomka et al. ‘13)

Function approximation:

- LSPI for Markov Games (Lagoudakis & Parr ‘02)

First Era of MARL

Nash Convergence of Gradient Dynamics in General-Sum Games

Satinder Singh
AT&T Labs
Florham Park, NJ 07932
baveja@research.att.com

Michael Kearns
AT&T Labs
Florham Park, NJ 07932
mkearns@research.att.com

Yishay Mansour
Tel Aviv University
Tel Aviv, Israel
mansour@math.tau.ac.il

Singh, Kearns & Mansour ‘03, [Infinitesimal Gradient Ascent (IGA)]

First Era of MARL

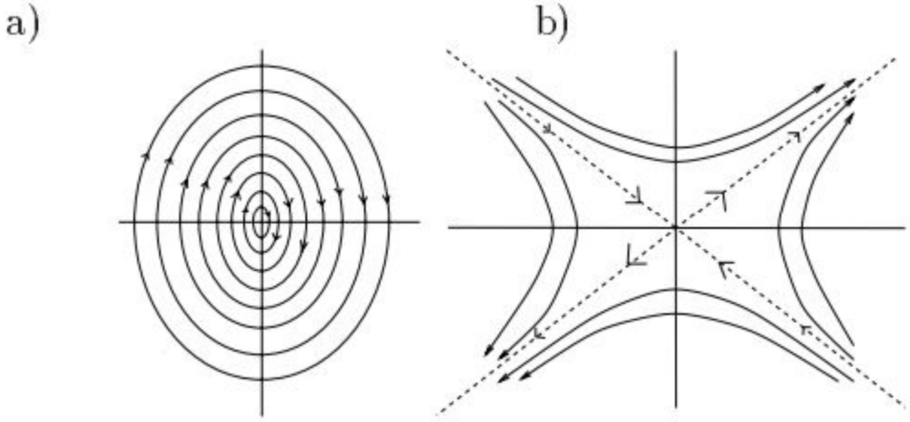


Figure 1: The general form of the dynamics: a) when U has imaginary eigenvalues and b) when U has real eigenvalues.

Image from Singh, Kearns, & Mansour ‘03

Formalize optimization as a dynamical system:

policy gradients

Analyze using well-established techniques

First Era of MARL

→ Evolutionary Game Theory: **replicator dynamics**

$$\dot{\pi}_t(a) = \pi_t(a)\left[u(a, \boldsymbol{\pi}_t) - \bar{u}(\boldsymbol{\pi}_t)\right]$$

time derivative

First Era of MARL

→ Evolutionary Game Theory: **replicator dynamics**

$$\dot{\pi}_t(a) = \pi_t(a)\left[u(a, \boldsymbol{\pi}_t) - \bar{u}(\boldsymbol{\pi}_t)\right]$$

time derivative

utility of action a against
the joint policy / population
of other players

First Era of MARL

→ Evolutionary Game Theory: **replicator dynamics**

$$\dot{\pi}_t(a) = \pi_t(a)\left[u(a, \boldsymbol{\pi}_t) - \bar{u}(\boldsymbol{\pi}_t)\right]$$

time derivative

utility of action a against the joint policy / population of other players

Expected / average utility of the joint policy / population

First Era of MARL

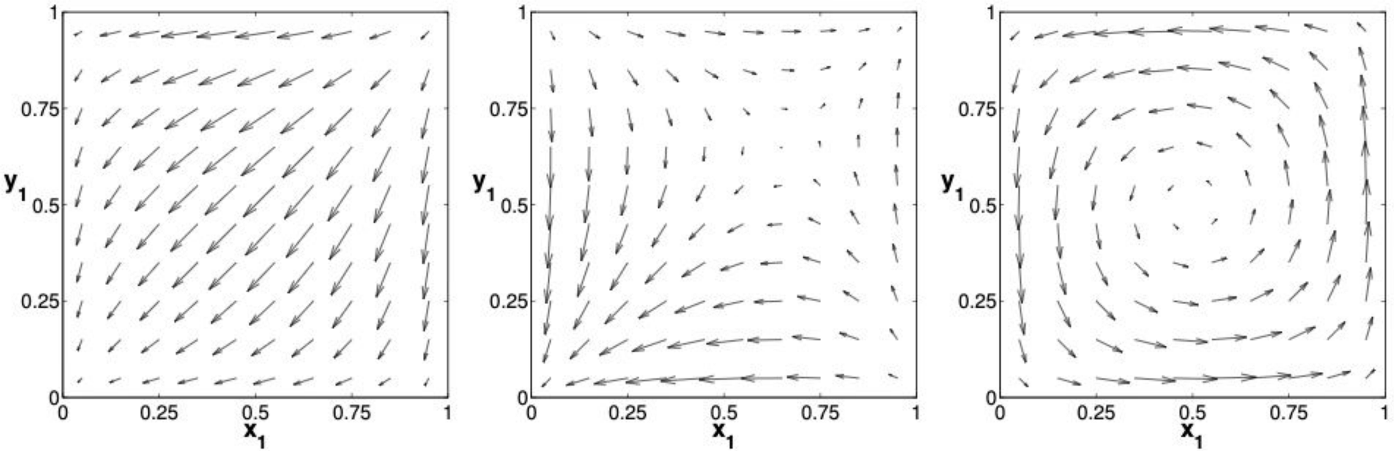


Figure 4: The replicator dynamics, plotted in the unit simplex, for the prisoner's dilemma (left), the stag hunt (center), and matching pennies (right).

[Bloembergen et al. 2015]

First Era of MARL

WoLF: Win or Learn Fast. (Bowling & Veloso '01).

IGA is **rational** but not **convergent**!

- *Rational*: opponents converge to a fixed joint policy

  → learning agent converges to a best response of joint policy

- *Convergent*: learner necessarily converges to a fixed policy

Use specific *variable learning rate* to ensure convergence (in 2x2 games)

First Era of MARL

Follow-ups to policy gradient and replicator dynamics:

- WoLF-IGA, WoLF-PHC
- WoLF-GIGA (Bowling ‘05)
- Weighted Policy Learner (Abdallah & Lesser ‘08)
- Infinitesimal Q-learning (Wunder et al. ‘10)
- Frequency-Adjusted Q-Learning (Kaisers et al. ‘10, Bloembergen et al. ‘11)
- Policy Gradient Ascent with Policy Prediction (Zhang & Lesser ‘10)
- Evolutionary Dynamics of Multiagent Learning (Bloembergen et al. ‘15)

So……

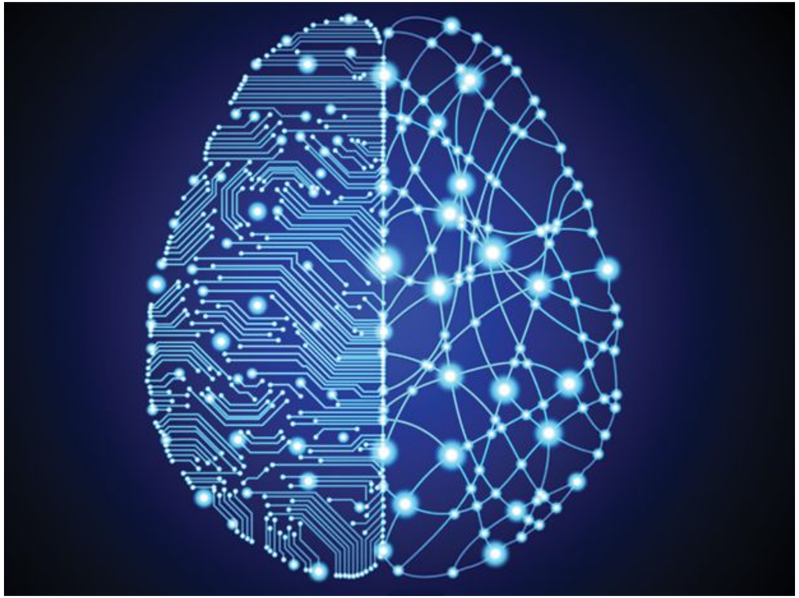
Why call it “the first era”?

So……

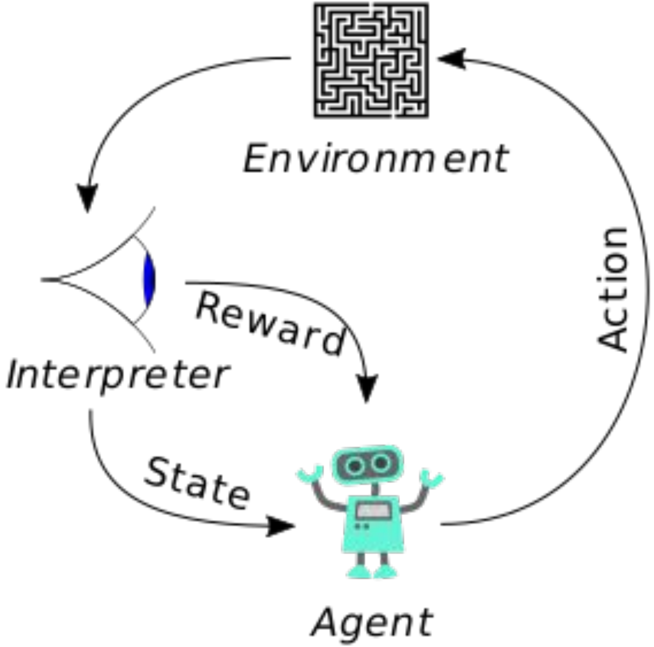
Why call it "the first era"?

Scalability was a major problem.

Second Era: Deep Learning meets Multiagent RL



Source: spectrum.ieee.org

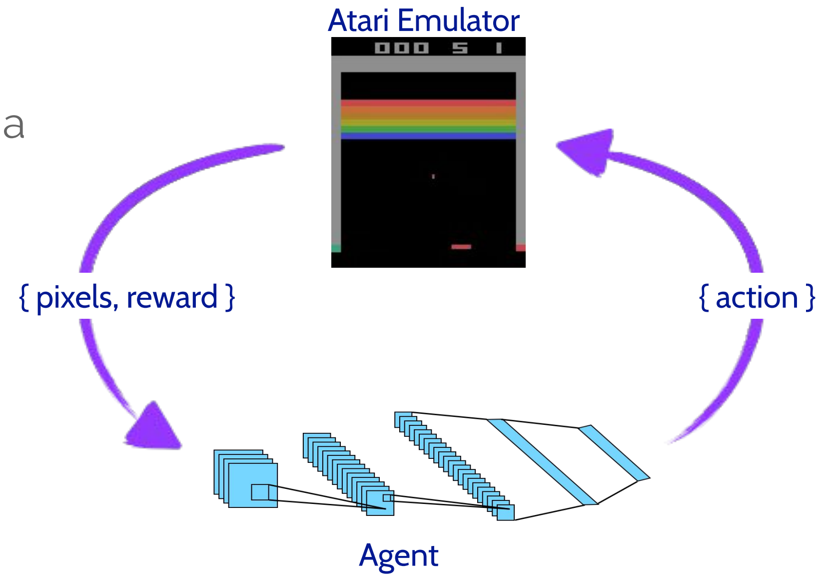


Source: wikipedia.org

Deep Q-Networks (DQN) Mnih et al. 2015

"Human-level control through deep reinforcement learning"

- Represent the action value (Q) function using a convolutional neural network.
- Train using end-to-end Q-learning.
- Can we do this in a stable way?

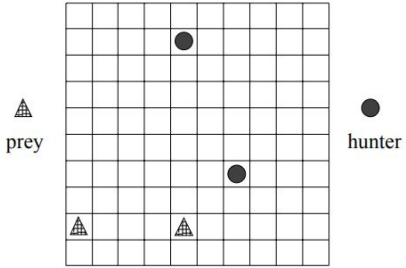


Independent Q-Learning Approaches

Independent Q-learning [Tan, 1993]

$$Q(x,a) \leftarrow Q(x,a) + \beta(r + \gamma V(y) - Q(x,a))$$

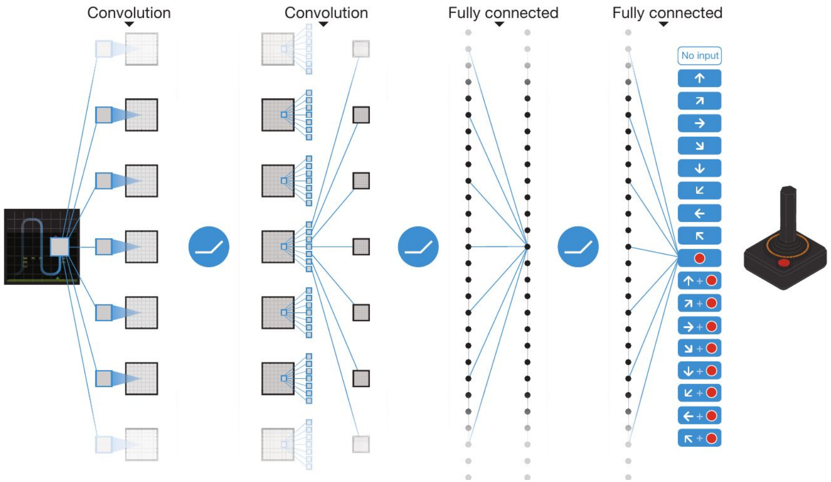
$$V(x) = \max_{b \in actions} Q(x,b)$$

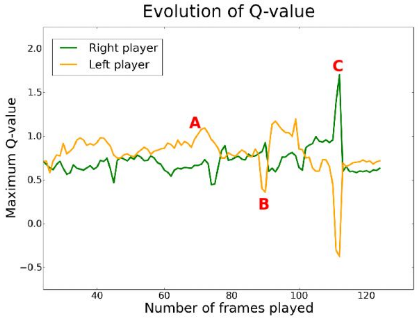


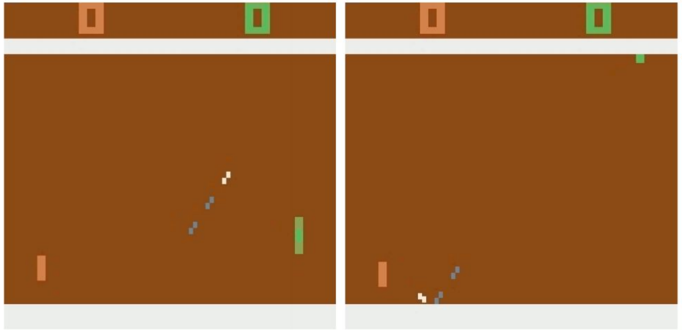
N-of-prey/N-of-hunters	1/1	1/2
Random hunters	123.08	56.47
Learning hunters	25.32	12.21

Table 1: Average Number of Steps to Capture a Prey

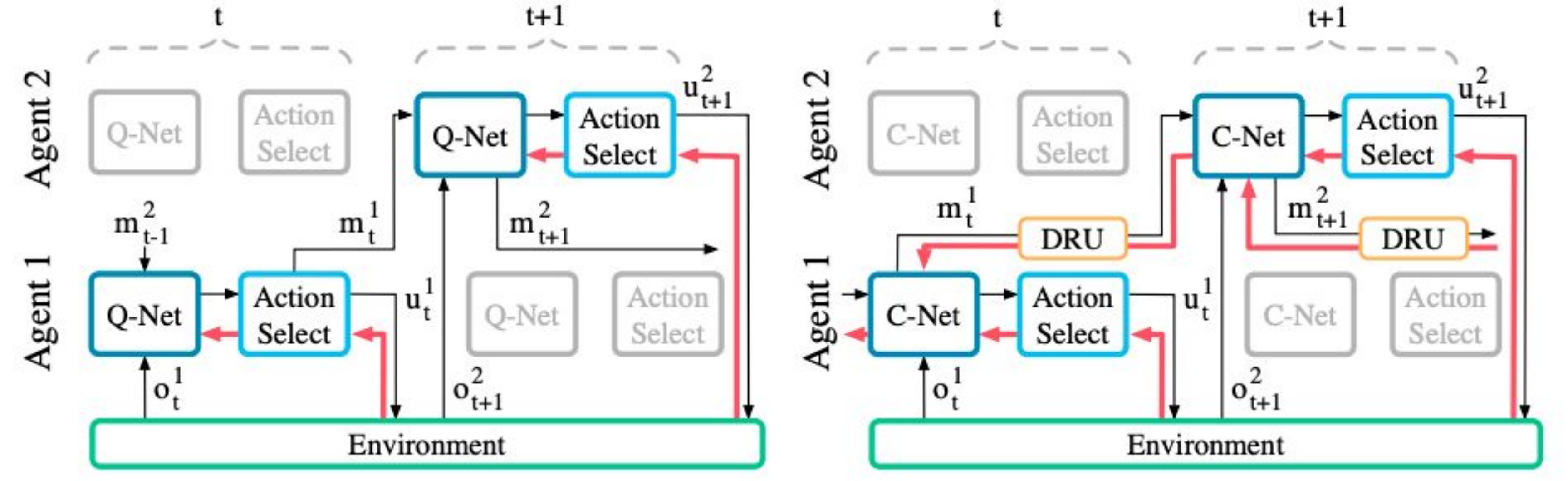
Independent Deep Q-Networks [Tampuu et al., 2015]







Learning to Communicate

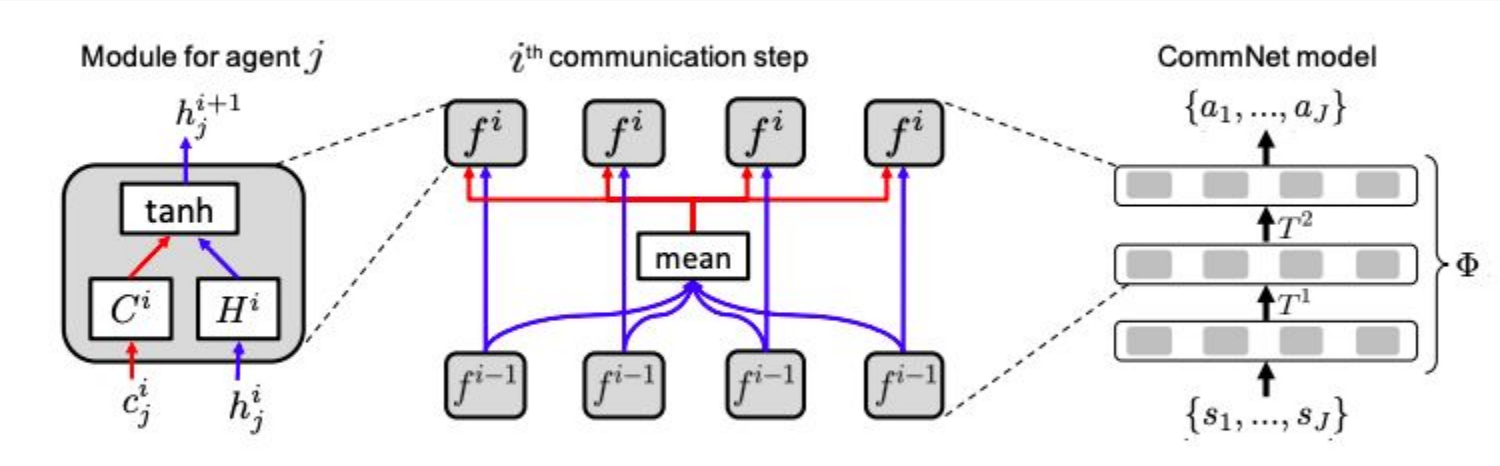


(a) RIAL - RL based communication

(b) DIAL - Differentiable communication

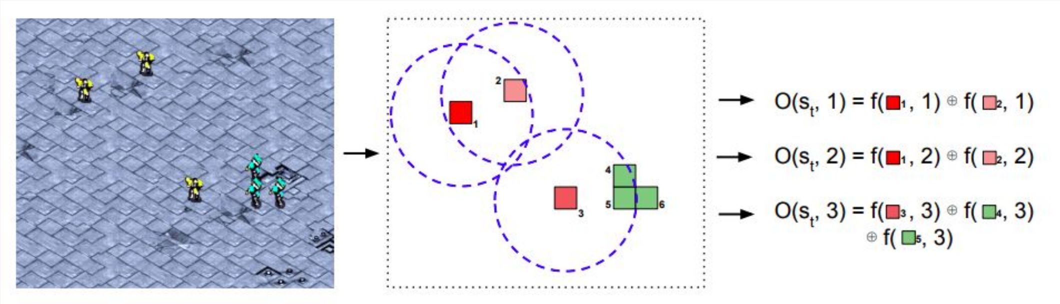
Foerster et al. '16

Learning to Communicate



Sukhbaatar et al. '16

Cooperative Multiagent Tasks

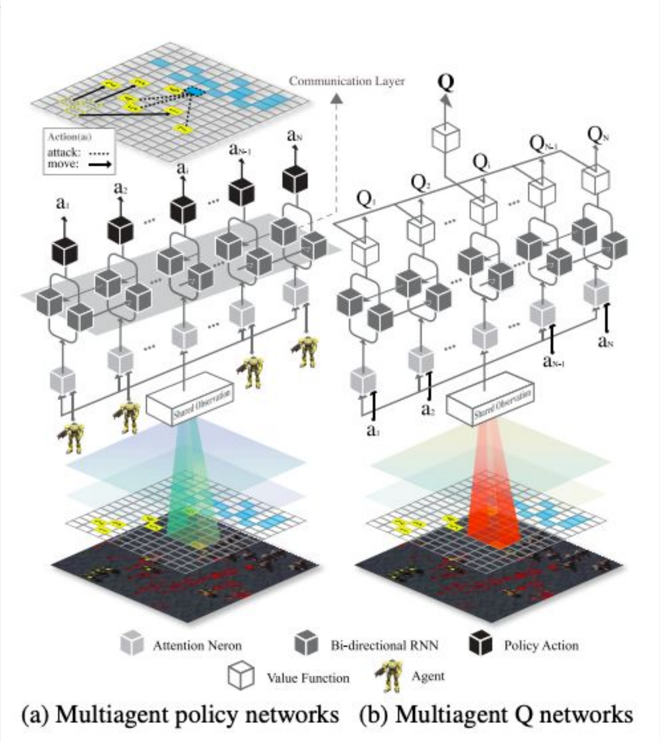


Foerster et al. ‘18

Episodic Exploration for Deep Deterministic Policies: An Application to StarCraft Micromanagement Tasks

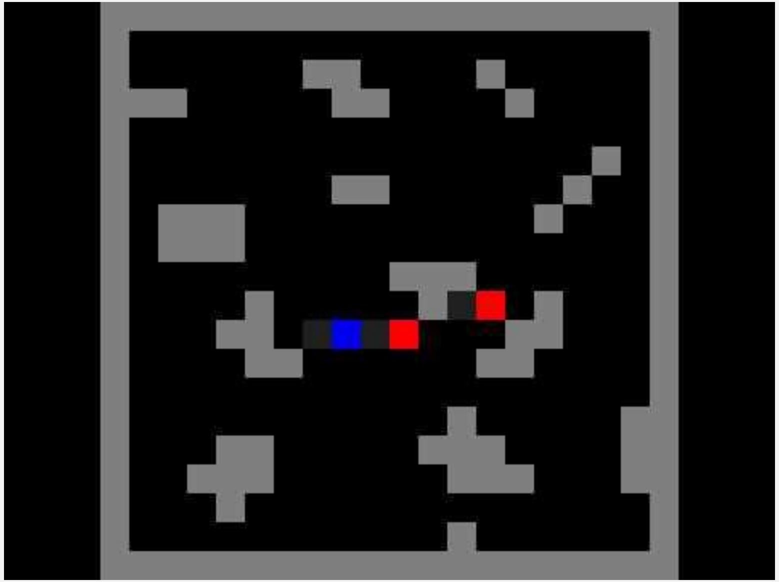
Nicolas Usunier*, Gabriel Synnaeve*, Zeming Lin, Soumith Chintala
Facebook AI Research
usunier,gab,zlin,soumith@fb.com

November 29, 2016



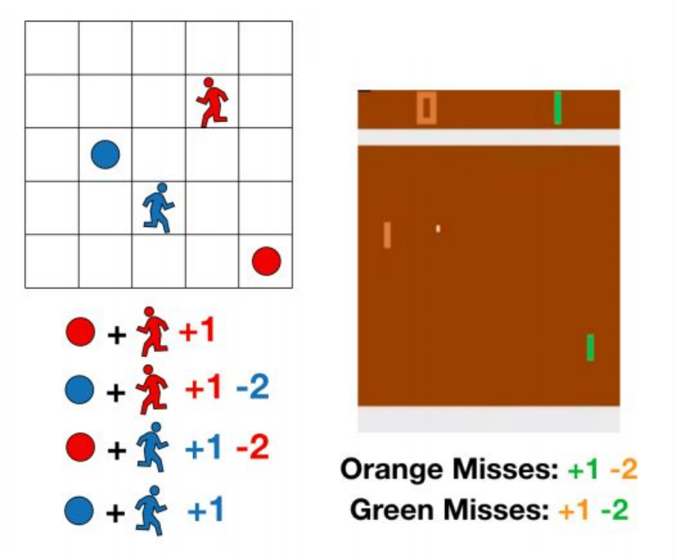
BIC-Net (Peng et al.’17)

Sequential Social Dilemmas



https://www.youtube.com/watch?v=0kaIqz6AvwE

Leibo et al. ‘17



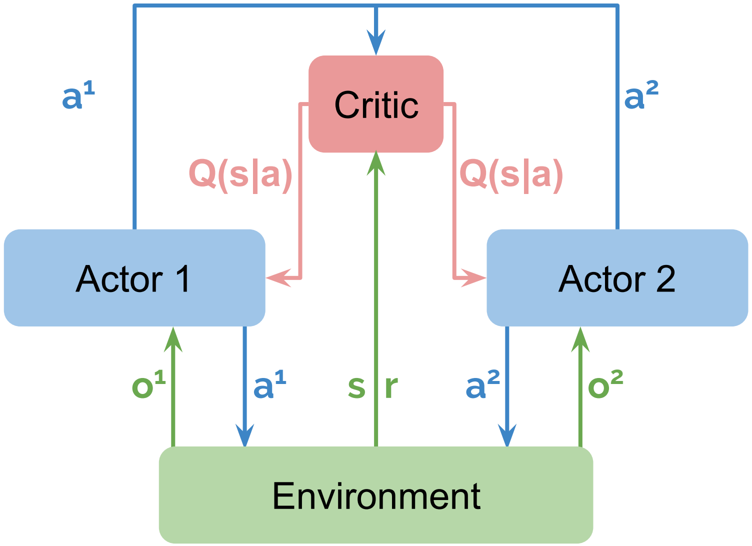
(a) Coins

(b) PPD

Lerer & Peyskavich ‘18

Centralized Critic Decentralized Actor Approaches

- **Idea:** reduce nonstationarity & credit assignment issues using a central critic
- **Examples:** MADDPG [Lowe et al., 2017] & COMA [Foerster et al., 2017]
- Apply to both cooperative and competitive games



Centralized critic trained to minimize loss:

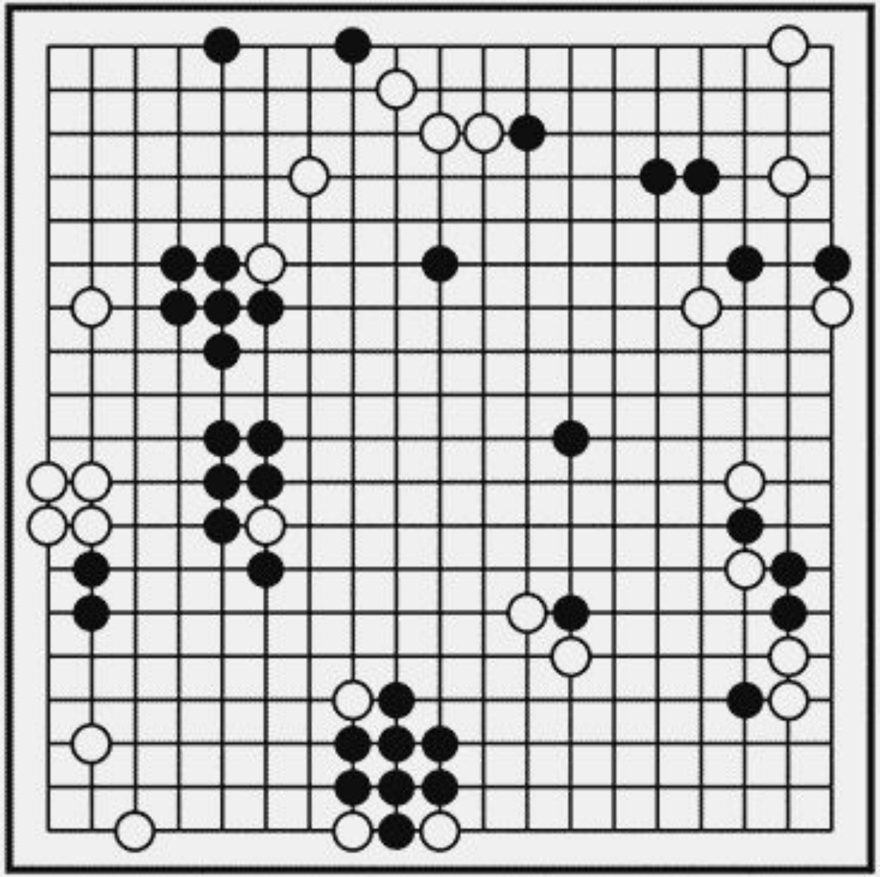
$$\mathcal{L}(\theta_i) = \mathbb{E}_{\mathbf{x},a,r,\mathbf{x}'}[(Q_i^{\boldsymbol{\pi}}(\mathbf{x}, a_1, \ldots, a_N) - y)^2],$$

$$y = r_i + \gamma Q_i^{\boldsymbol{\pi}'}(\mathbf{x}', a'_1, \ldots, a'_N)\big|_{a'_j = \boldsymbol{\pi}'_j(o_j)}$$

Decentralized actors trained via policy gradient:

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}_{s\sim p^{\boldsymbol{\mu}}, a_i \sim \boldsymbol{\pi}_i}[\nabla_{\theta_i} \log \underbrace{\boldsymbol{\pi}_i(a_i|o_i)}_{\text{Actor}} \underbrace{Q_i^{\boldsymbol{\pi}}(\mathbf{x}, a_1, \ldots, a_N)}_{\text{Critic}}]$$


AlphaGo

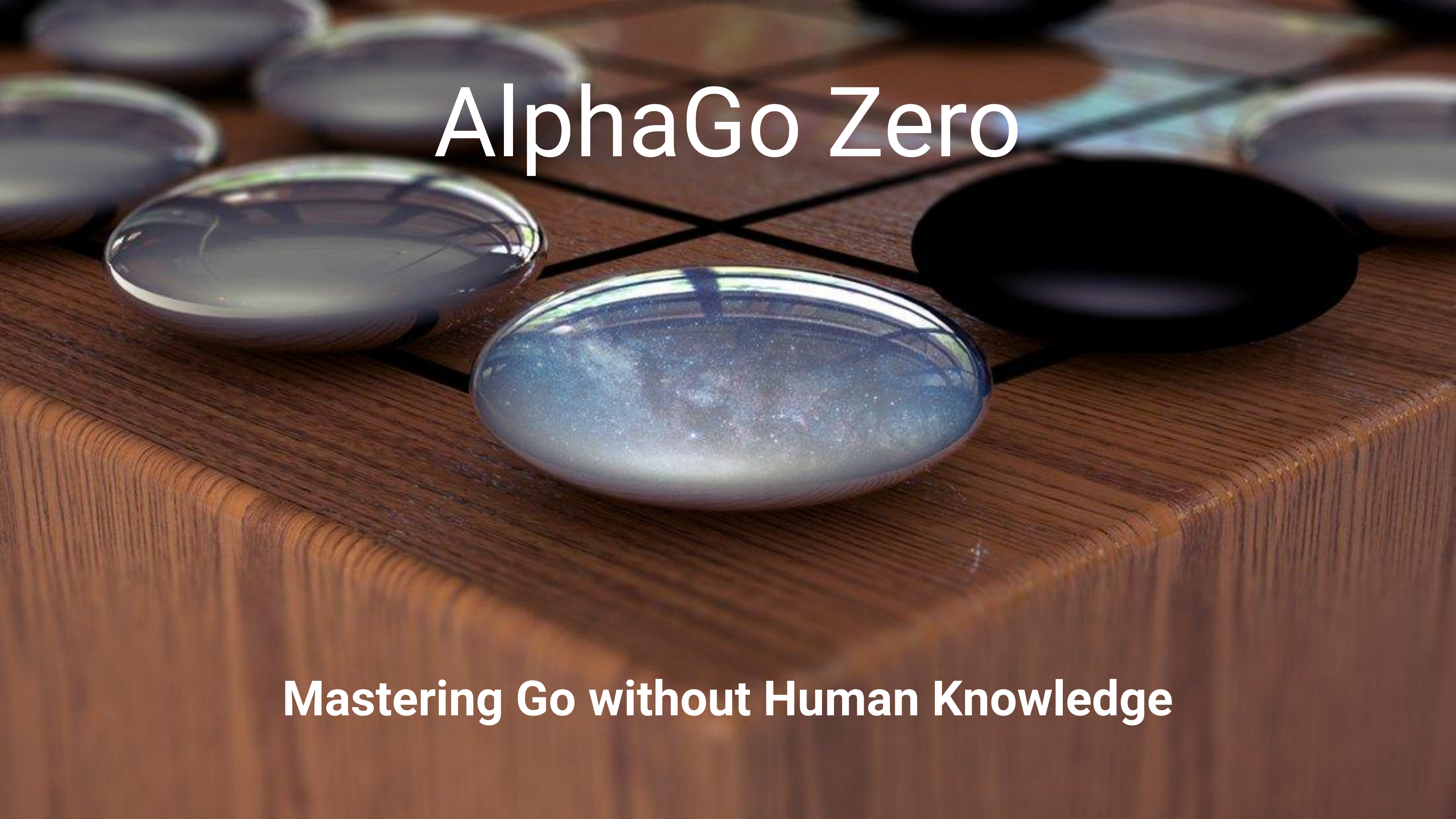


AlphaGo vs. Lee Sedol

Lee Sedol (9p): winner of 18 world titles

Match was played in Seoul, March 2016

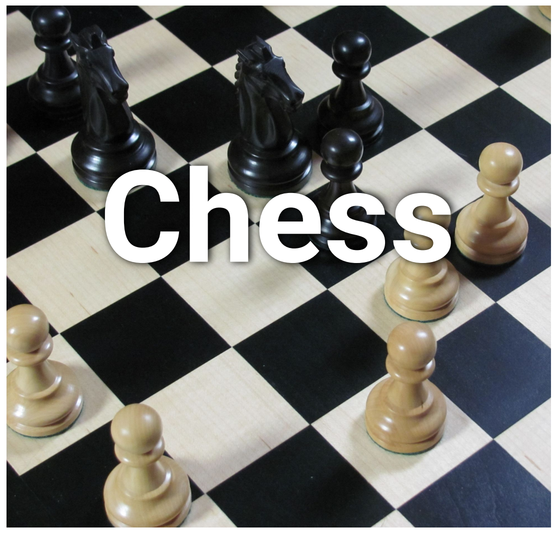
AlphaGo won the match 4-1



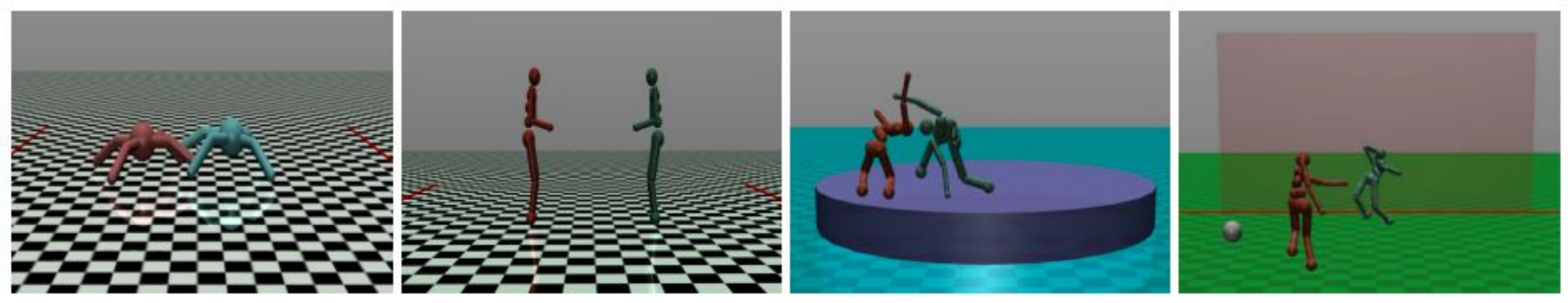
AlphaGo Zero

Mastering Go without Human Knowledge

AlphaZero: One Algorithm, Three Games



3D Worlds



Bansal et al. '18

Meta-Learning in RoboSumo

Al-Shedivat et al. '17

Emergent Coordination Through Competition

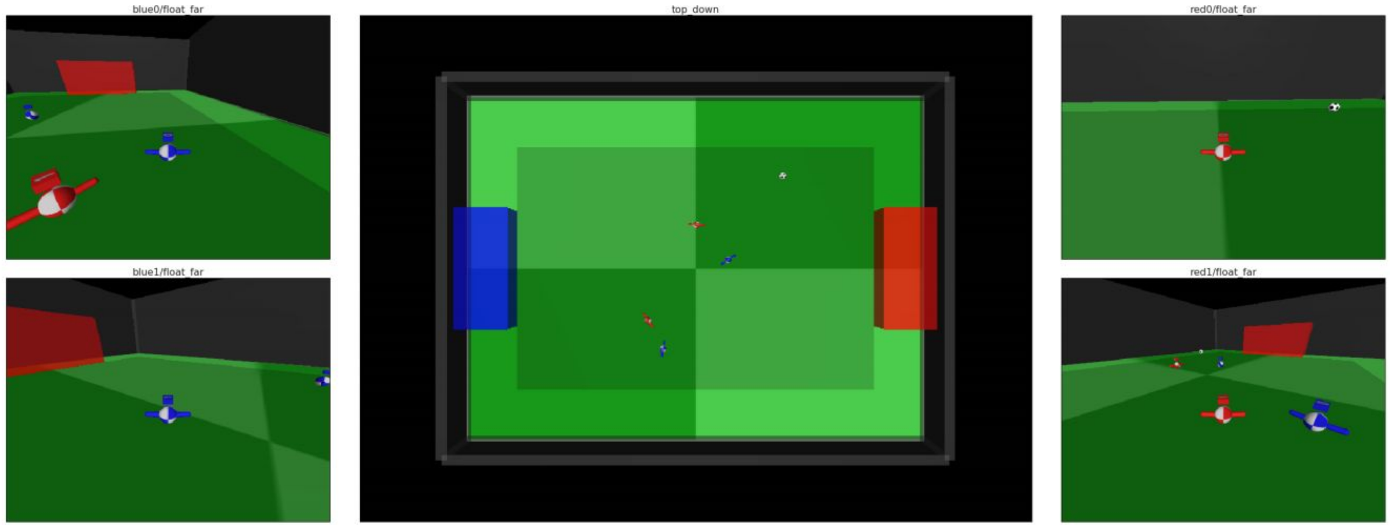
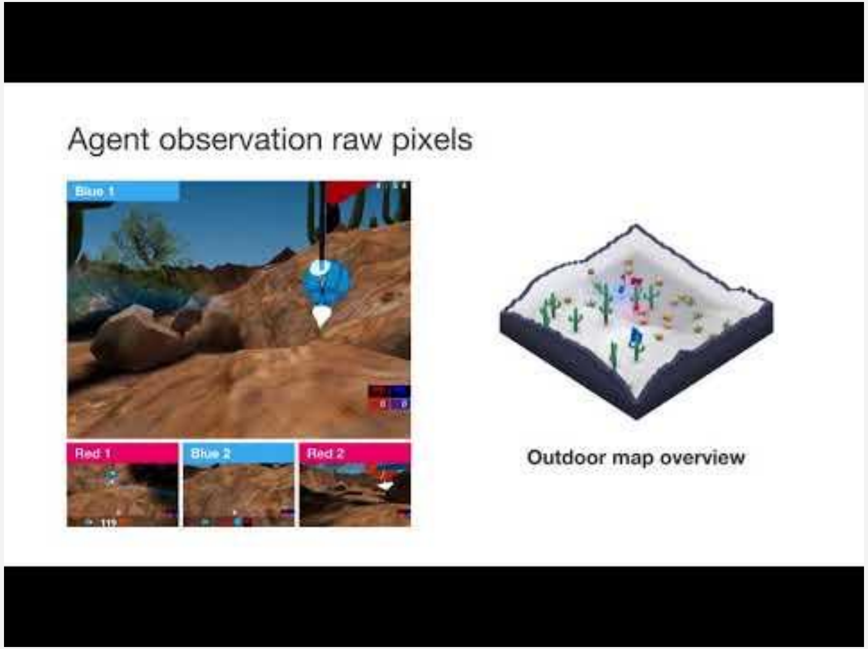
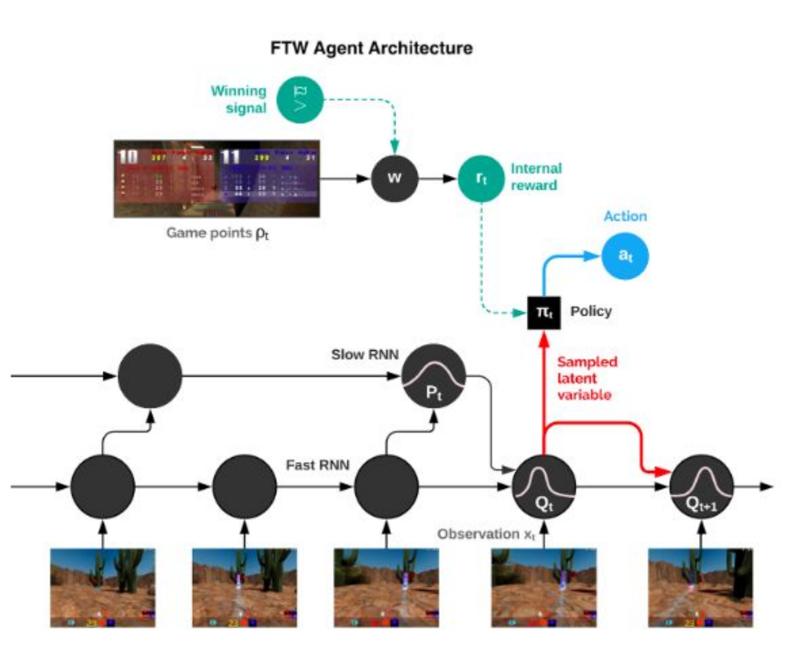


Figure 1: Top-down view with individual camera views of 2v2 multi-agent soccer environment.

Liu et al. ‘19   and   http://git.io/dm_soccer

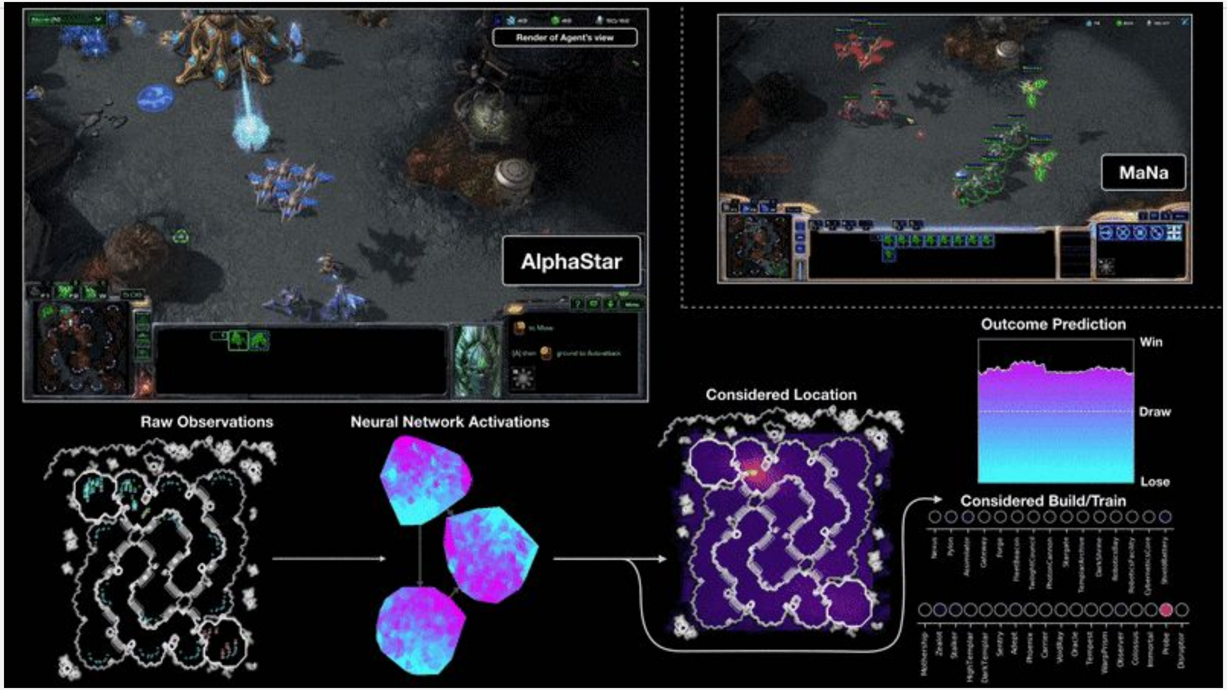
Capture-the-Flag (Jaderberg et al. ‘19)





https://deepmind.com/blog/capture-the-flag-science/

AlphaStar (Vinyals et al. '19)



https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

Dota 2: OpenAI Five

https://openai.com/blog/openai-five-finals/

Deep Multiagent RL Survey

Is multiagent deep reinforcement learning the answer or the question?
A brief survey

Pablo Hernandez-Leal , Bilal Kartal and Matthew E. Taylor
{pablo.hernandez,bilal.kartal,matthew.taylor}@borealisai.com

Borealis AI
University of Alberta CCIS 3-232
Edmonton, Canada

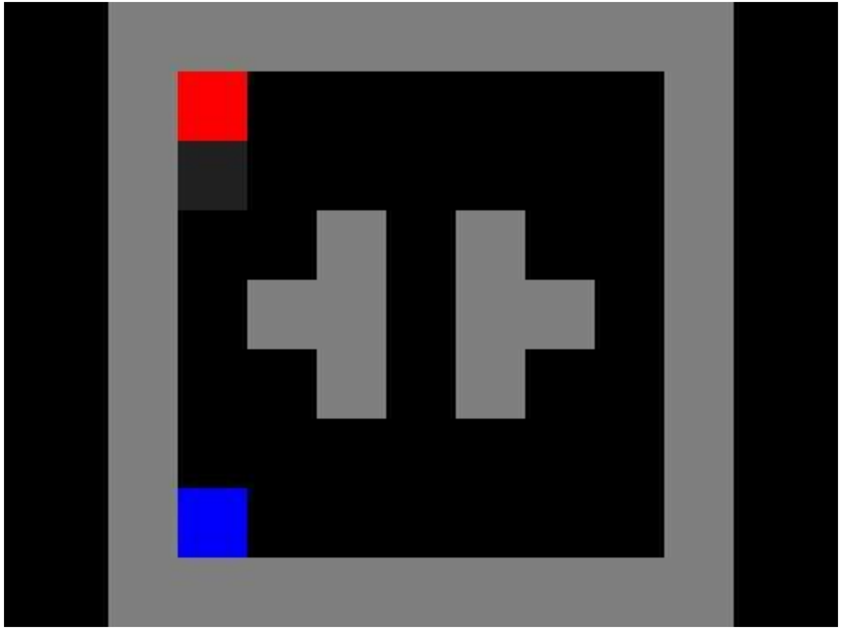
https://arxiv.org/abs/1810.05587

Part 4: Partial Observability

Foundations of Multiagent RL

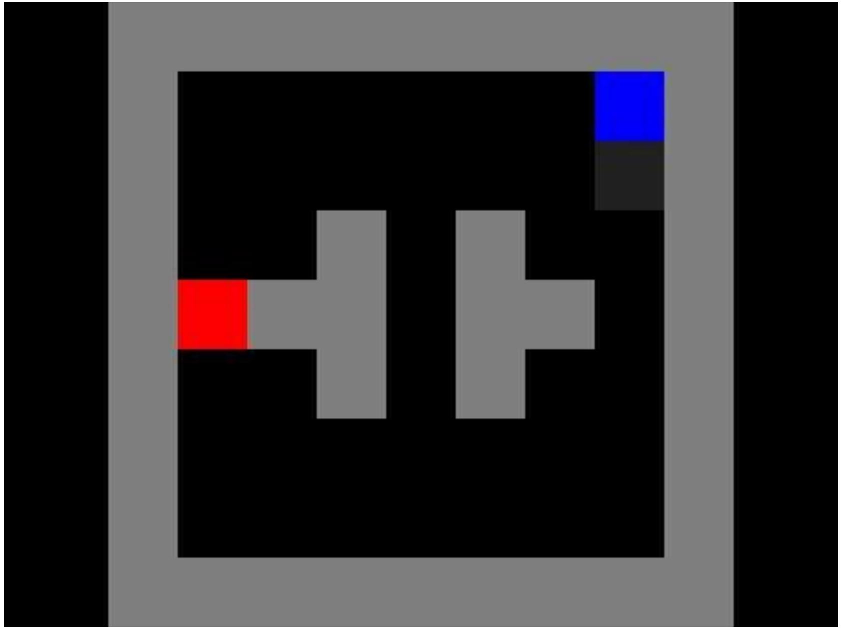
	Single Agent	Multiple (e.g. 2) Agents
Large Problems	Reinforcement Learning	**Multiagent Reinforcement Learning**
Small Problems	Approximate Dynamic Programming	Game Theory

Independent Deep Q-networks (See [Lanctot et al. '17])



https://www.youtube.com/watch?v=8vXpdHuoQH8

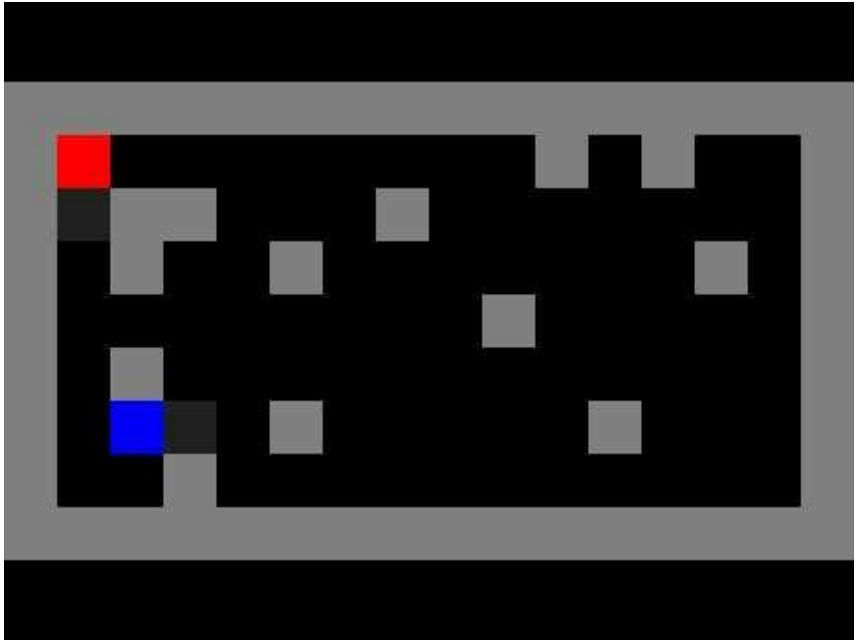
Independent learners who learned together



https://www.youtube.com/watch?v=jOjwOkCM_i8

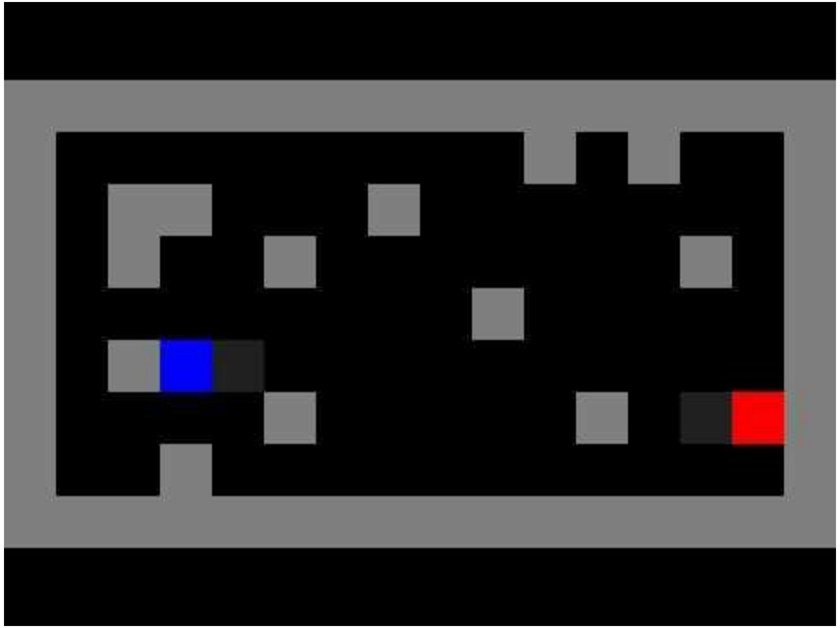
Independent learners who learned using the same algorithm, same architecture, same hyperparameters, but different seed

Independent Deep Q-networks (See [Lanctot et al. '17])



https://www.youtube.com/watch?v=Z5cpIG3GsLw

Independent learners who learned together



https://www.youtube.com/watch?v=zilU0hXvGK4

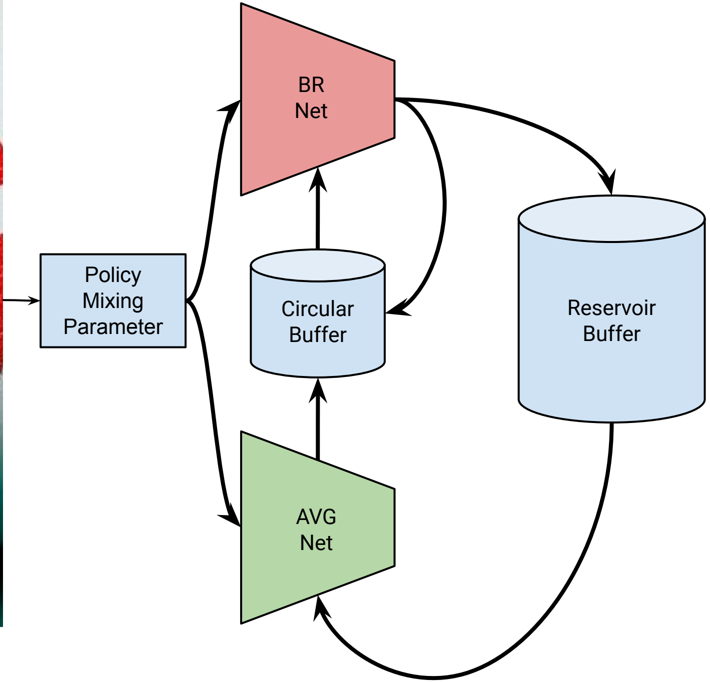
Independent learners who learned using the same algorithm, same architecture, same hyperparameters, but different seed

Fictitious Self-Play [Heinrich et al. '15, Heinrich & Silver 2016]

- **Idea:** Fictitious self-play (FSP) + reinforcement learning
- Update rule in sequential setting *equivalent* to standard fictitious play (matrix game)
- Approximate NE via two neural networks:

1. **Best response net (BR):**
   - Estimate a best response
   - Trained via RL

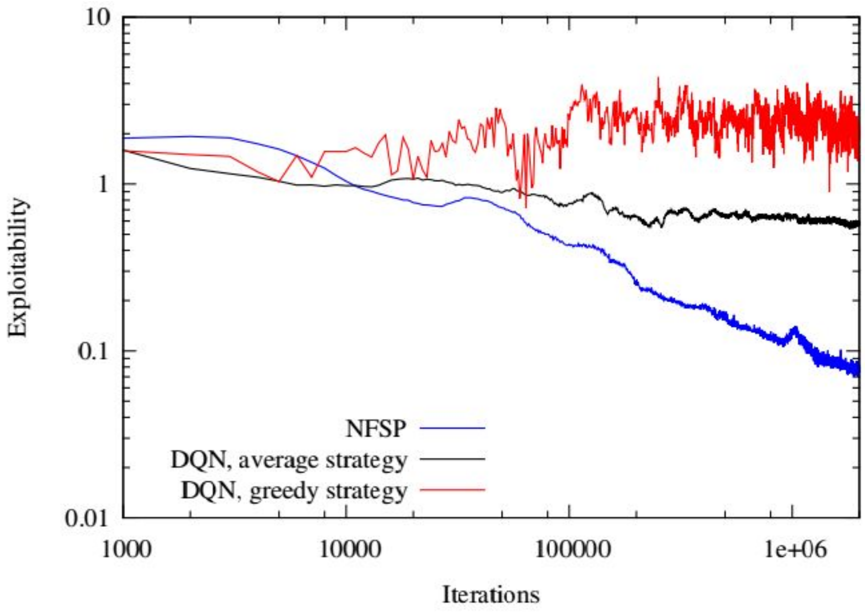
2. **Average policy net (AVG):**
   - Estimate the time-average policy
   - Trained via supervised learning



Neural Fictitious Self-Play [Heinrich & Silver 2016]

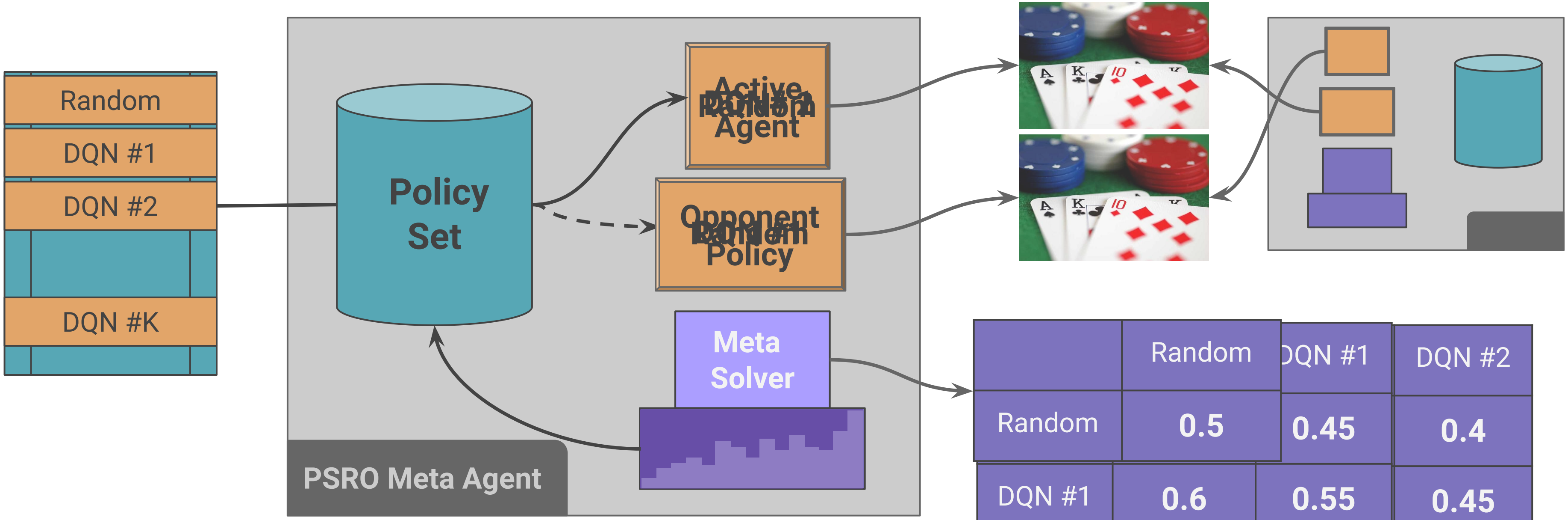
- Leduc Hold’em poker experiments:

“Closeness” to Nash



- 1st scalable end-to-end approach to learn **approximate Nash equilibria w/o prior domain knowledge**
    - Competitive with superhuman computer poker programs when it was released

Policy-Space Response Oracles [(Lanctot et al. ‘17)]



	Random	DQN #1	DQN #2
Random	0.5	0.45	0.4
DQN #1	0.6	0.55	0.45
DQN #2	0.7	0.6	0.56

Quantifying "Joint Policy Correlation"

In RL:

- Each player uses optimizes independently
- After many steps, joint policy (π_1, π_2) co-learned for players 1 & 2

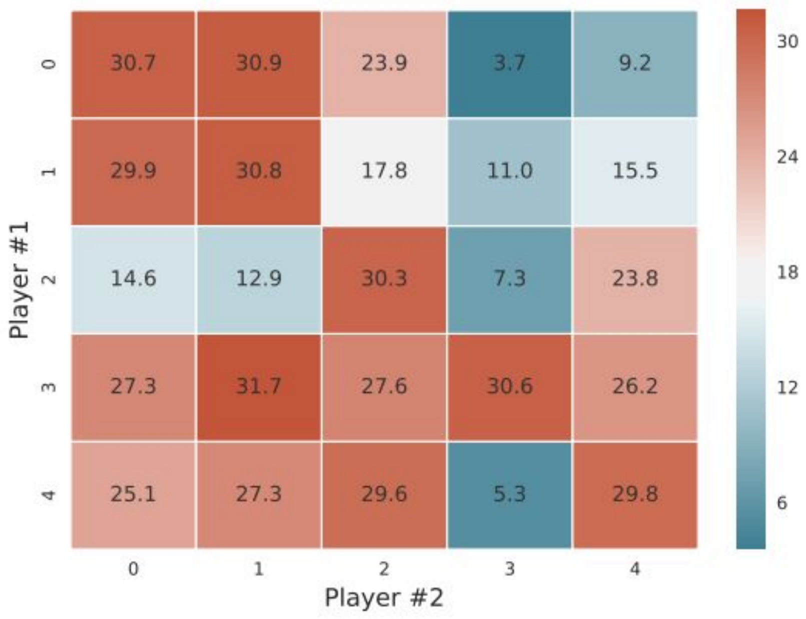
Computing **JPC:** start **5 separate instances of the *same experiment***, with

  - Same hyper-parameter values
  - Differ *only* by seed (!)

- Reload all 25 combinations and play π_1^i with π_2^j for instances i, j

Joint Policy Correlation in Independent RL

InRL in small2 (first) map



InRL in small4 map

JPC Results - Laser Tag

Game	Diag	Off Diag	Exp. Loss
LT small2	30.44	20.03	**34.2 %**
LT small3	23.06	9.06	**62.5 %**
LT small4	20.15	5.71	**71.7 %**
Gathering field	147.34	146.89	none
Pathfind merge	108.73	106.32	none

Exploitability Descent (Lockhart et al. '19)

Algorithm 2: Exploitability Descent (ED)

input : π^0 — initial joint policy

for $t \in \{1, 2, \cdots\}$ **do**

  for $i \in \{1, \cdots, n\}$ **do**

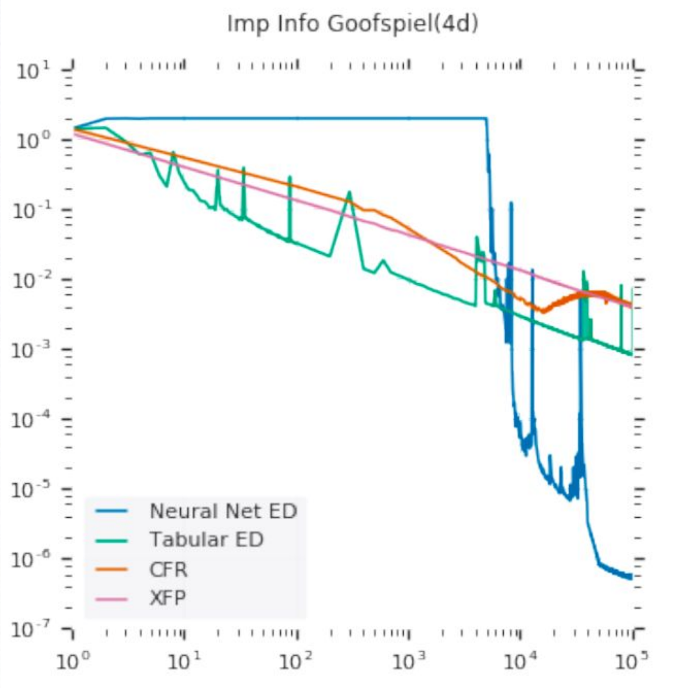
    Compute a best response $\boldsymbol{b}_i^t(\boldsymbol{\pi}_{-i}^{t-1})$

  for $i \in \{1, \cdots, n\}, s \in \mathcal{S}_i$ **do**

    Define $\boldsymbol{b}_{-i}^t = \{\boldsymbol{b}_j^t\}_{j \neq i}$

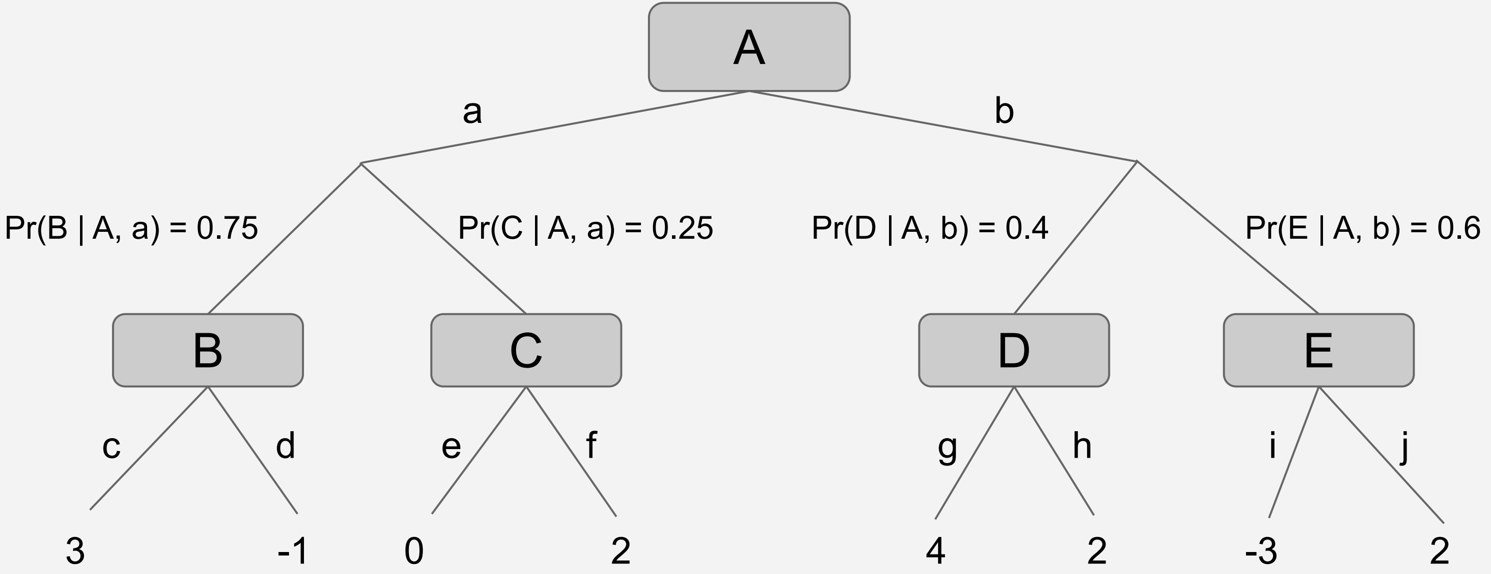
    Let $\mathbf{q}^{\boldsymbol{b}}(s) = \textsc{ValuesVsBRs}(\boldsymbol{\pi}_i^{t-1}(s), \boldsymbol{b}_{-i}^t)$

    $\boldsymbol{\pi}_i^t(s) = \textsc{GradAscent}(\boldsymbol{\pi}_i^{t-1}(s), \alpha^t, \mathbf{q}^{\boldsymbol{b}}(s))$



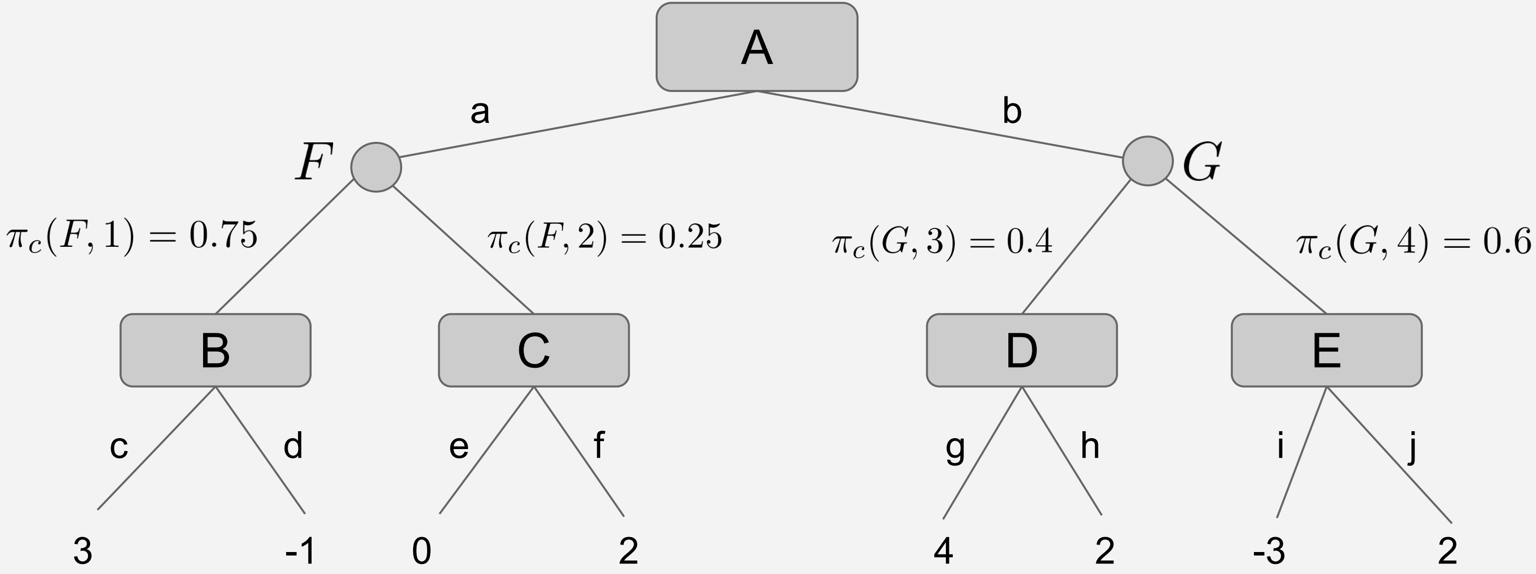
- A FP-like algorithm conv. *without averaging*!
- Amenable to function approximation

A simple MDP

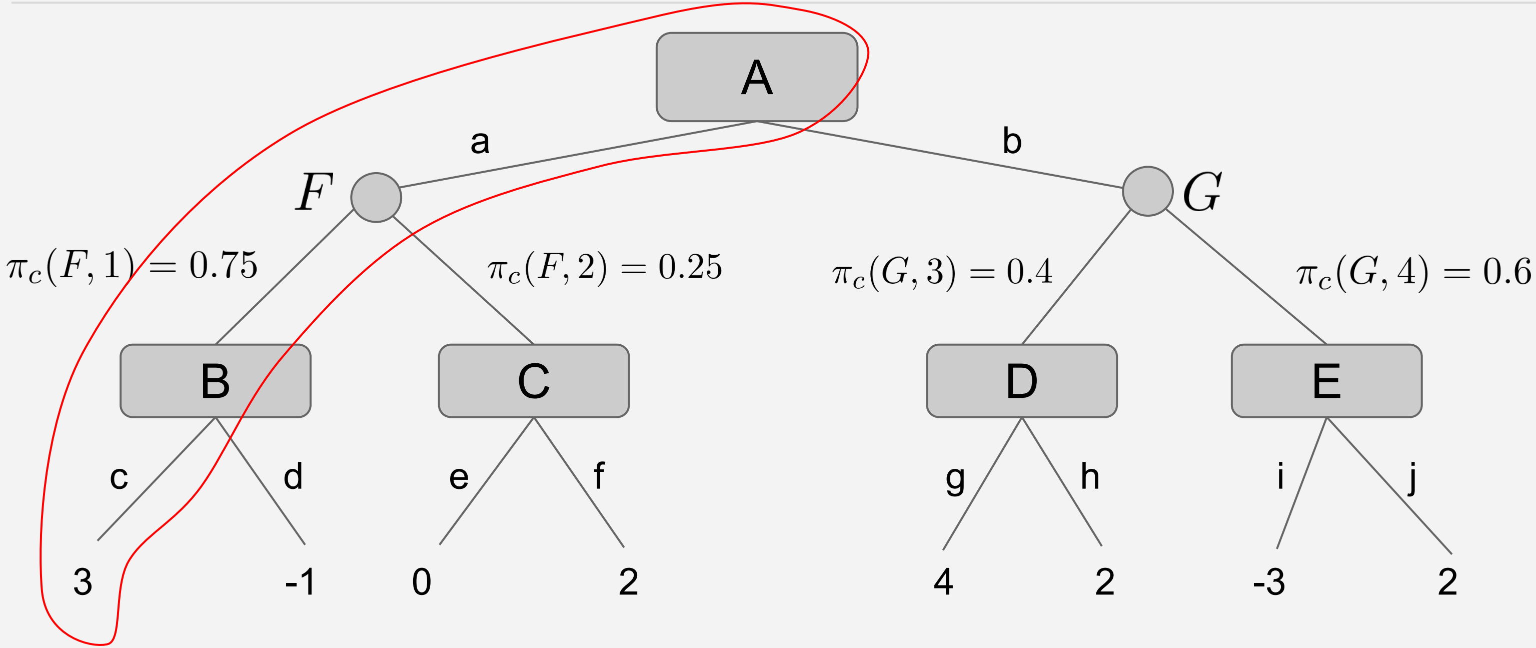


A
a
b
Pr(B | A, a) = 0.75
Pr(C | A, a) = 0.25
Pr(D | A, b) = 0.4
Pr(E | A, b) = 0.6
B
C
D
E
c
d
e
f
g
h
i
j
3
-1
0
2
4
2
-3
2

A simple ~~MDP~~ Multiagent System

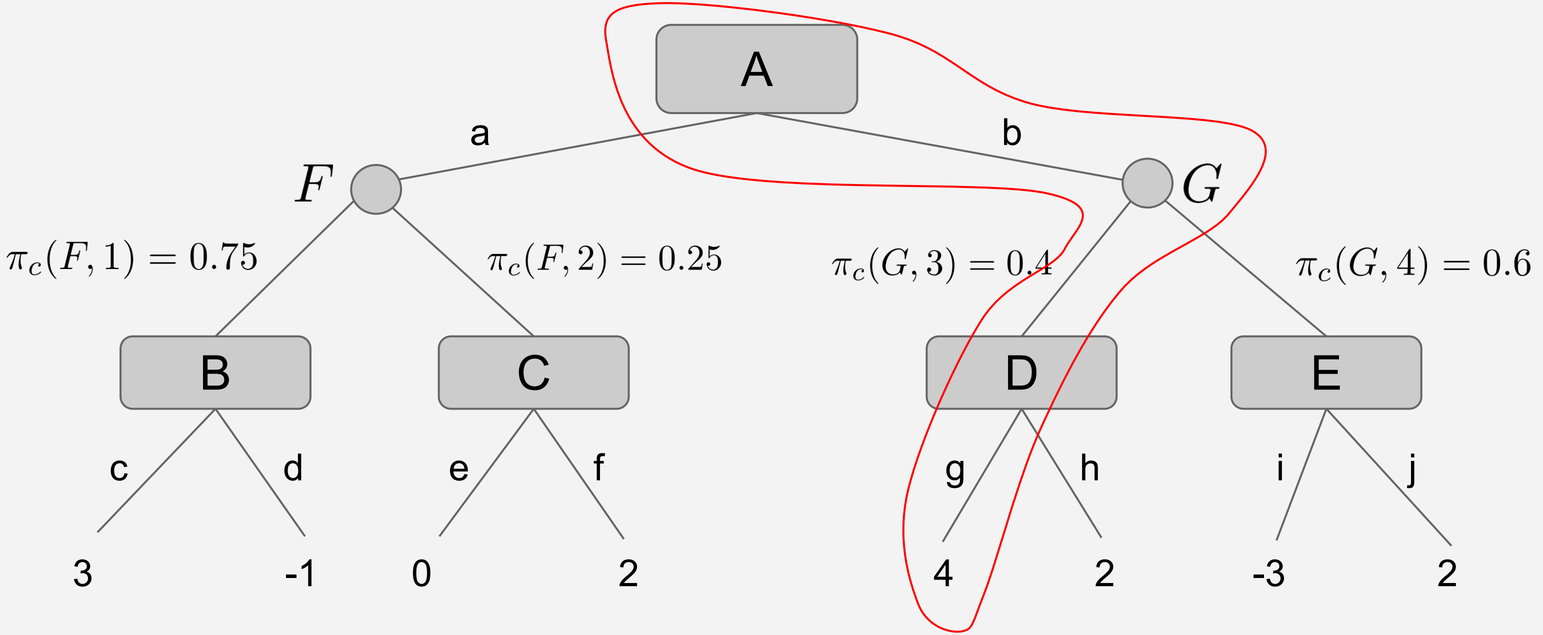


Terminal history A.K.A. Episode



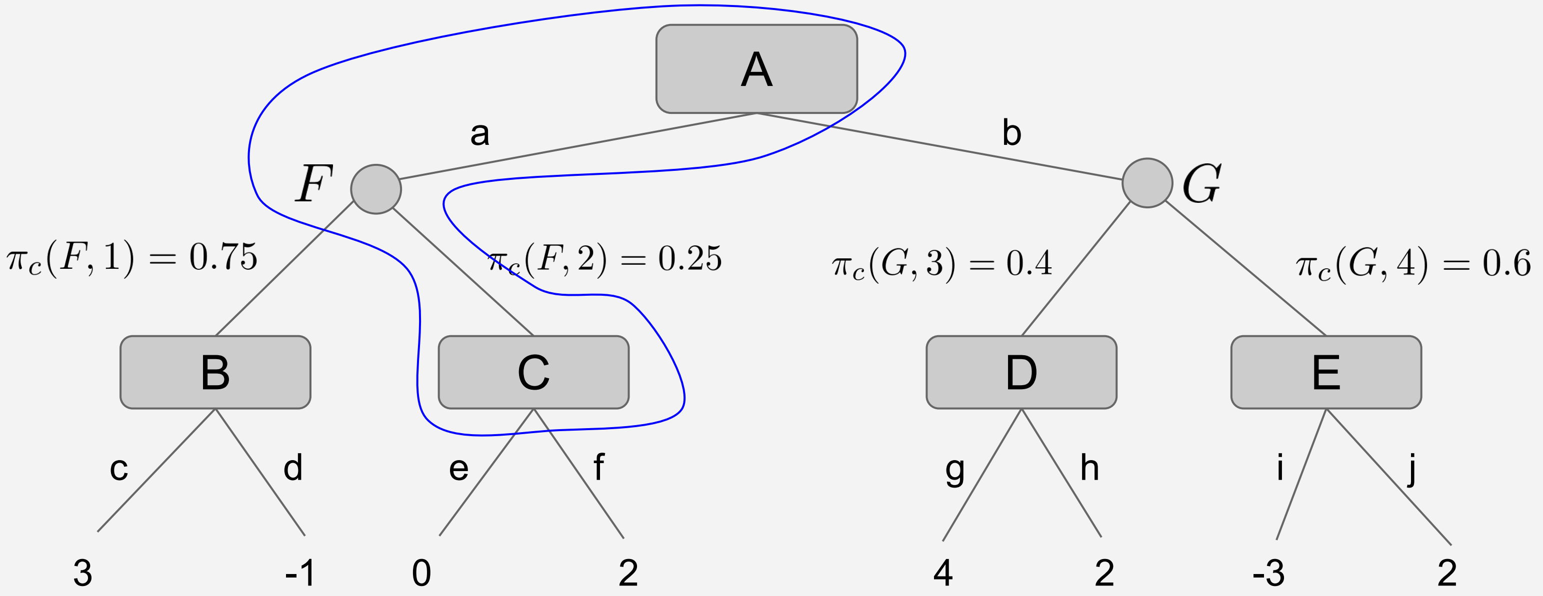
(A, a, F, 1, B, c) is a *terminal* history.

Terminal history A.K.A. Episode



(A, a, F, 1, B, c) is a *terminal* history. (A, b, G, 3, D, g) is a another terminal history.

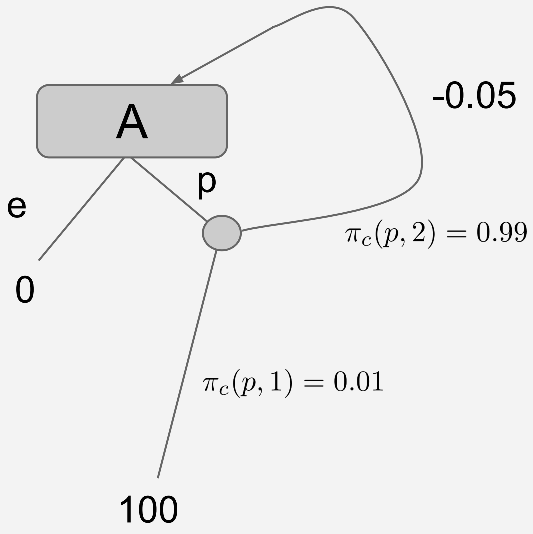
Prefix (non-terminal) Histories



(A, a, F, 2, C) is a history. It is a *prefix* of (A, a, F, 2, C, e) and (A, a, F, 2, C, f).

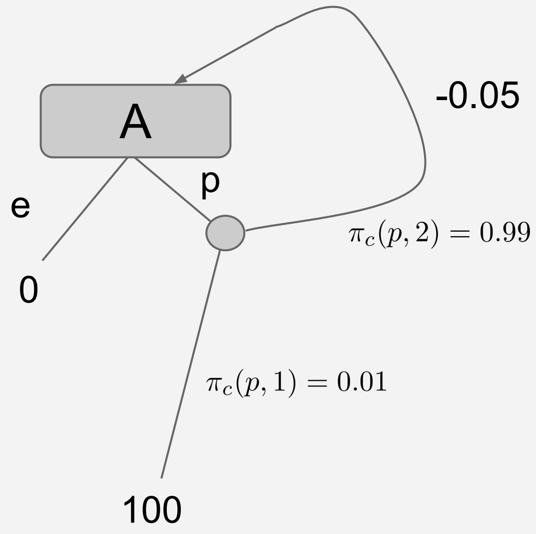
Perfect Recall of Actions and Observations

Another simple MDP:

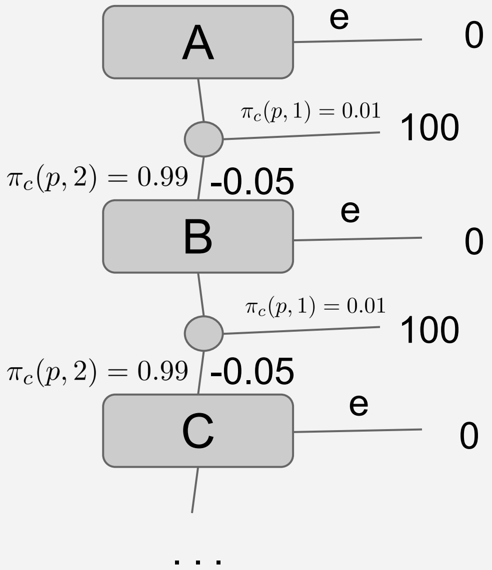


Perfect Recall of Actions and Observations

Another simple MDP:



A different MDP:

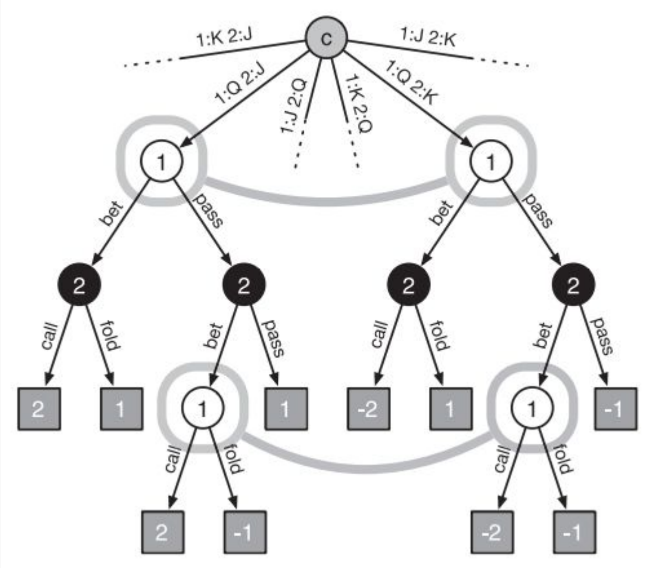


Partially Observable Environment

An **information state** is a set of histories consistent with an agent's observations.

3-card Poker deck:

Jack, Queen, King

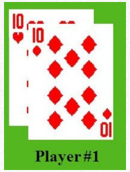


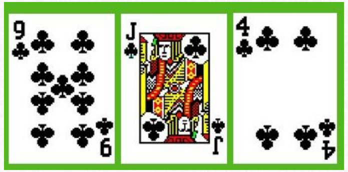
Terminology

What is a "state"?

- An **information state** s corresponds to *sequence of observations*
    - with respect to the player to act at s

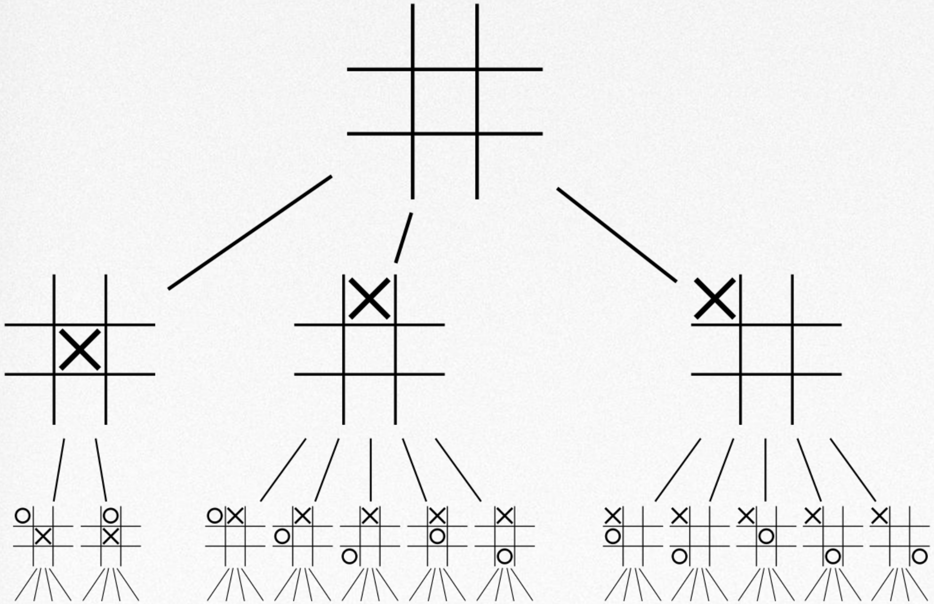
Example information state in Poker:

Ante: 1 chip per player, 

, P1 bets 1 chip, P2 calls, 

Environment is in one of many **world/ground states** $h \in s$

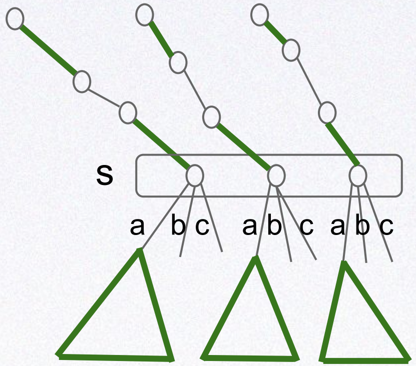
Recall: Turn-Taking Perfect Information Games



→ **Exactly one history per information state!**

{Q,V}-values and Counterfactual Values

What.................. is a counterfactual value?



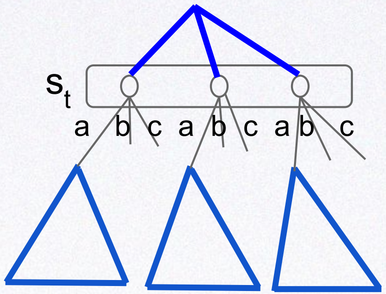
$$v_i^c(\pi, s, a)$$

The portion of the expected return (under s) from the start state, given that:

player i plays to reach information state s (then plays a).

Q-values in Partially Observable Environments

What.................... **is a q-value?**



$$q_{\pi,i}(s_t, a_t) = \mathbb{E}_{\rho\sim\pi}[G_t \mid S_t = s_t, A_t = a_t]$$

Q-values in Partially Observable Environments

$$= \sum_{h,z \in \mathcal{Z}(s_t, a_t)} \Pr(h \mid s_t)\eta^{\pi}(ha, z)u_i(z)$$

All terminal histories z reachable from s, paired with their prefix histories ha, where h is in s

Reach probabilities: product of all policies' state-action probabilities along the portion of the history between ha and z

Return achieved over terminal history z

Q-values in Partially Observable Environments

$$= \sum_{h,z \in \mathcal{Z}(s_t,a_t)} \frac{\Pr(s_t \mid h) \Pr(h)}{\Pr(s_t)} \eta^{\pi}(ha, z) u_i(z)$$

By Bayes rule

Q-values in Partially Observable Environments

$$= \sum_{h,z \in \mathcal{Z}(s_t, a_t)} \frac{\Pr(h)}{\Pr(s_t)} \eta^{\pi}(ha, z) u_i(z)$$

Since h is in s_t and h is unique to s_t

Q-values in Partially Observable Environments

$$= \sum_{h,z \in \mathcal{Z}(s_t, a_t)} \frac{\eta^\pi(h)}{\sum_{h' \in s_t} \eta^\pi(h')} \eta^\pi(ha, z) u_i(z)$$

Q-values in Partially Observable Environments

Only player i's reach probabilities

Player i's opponent's probabilities (*inc. chance*!)

$$\sum_{h,z \in \mathcal{Z}(s_t,a_t)} \frac{\eta_i^\pi(h)\eta_{-i}^\pi(h)}{\sum_{h' \in s_t} \eta_i^\pi(h')\eta_{-i}^\pi(h')} \eta^\pi(ha,z)u_i(z)$$

Similarly here

and here

Q-values in Partially Observable Environments

$$= \sum_{h,z \in \mathcal{Z}(s_t,a_t)} \frac{\eta_i^\pi(h)\eta_{-i}^\pi(h)}{\eta_i^\pi(h)\sum_{h' \in s_t} \eta_{-i}^\pi(h')} \eta^\pi(ha,z)u_i(z)$$

Due to perfect recall (!!)

Q-values in Partially Observable Environments

$$= \sum_{h,z \in \mathcal{Z}(s_t,a_t)} \frac{\eta^{\pi}_{-i}(h)}{\sum_{h' \in s_t} \eta^{\pi}_{-i}(h')} \eta^{\pi}(ha,z) u_i(z)$$

Q-values in Partially Observable Environments

$$= \sum_{h,z \in \mathcal{Z}(s_t,a_t)} \frac{\eta^{\pi}_{-i}(h)}{\sum_{h' \in s_t} \eta^{\pi}_{-i}(h')} \eta^{\pi}(ha,z) u_i(z)$$

This is a counterfactual value!

Q-values in Partially Observable Environments

$$= \frac{1}{\sum_{h \in s_t} \eta^{\pi}_{-i}(h)} v_i^c(\pi, s_t, a_t)$$

$$= \frac{1}{\mathcal{B}_{-i}(\pi, s_t)} v_i^c(\pi, s_t, a_t)$$

For full derivation, see Sec 3.2 of [Srinivasan et al. '18]

Yeah.. so.... ?

¯_(ツ)_/¯

Counterfactual Regret Minimization (CFR)

Zinkevich et al. '08

- Algorithm to compute approx Nash eq. In 2P zero-sum games
- Hugely successful in Poker AI
- Size traditionally reduced apriori based on expert knowledge
- Key innovation: counterfactual values: $v_i^c(\pi, s, a)$ $v_i^c(\pi, s)$

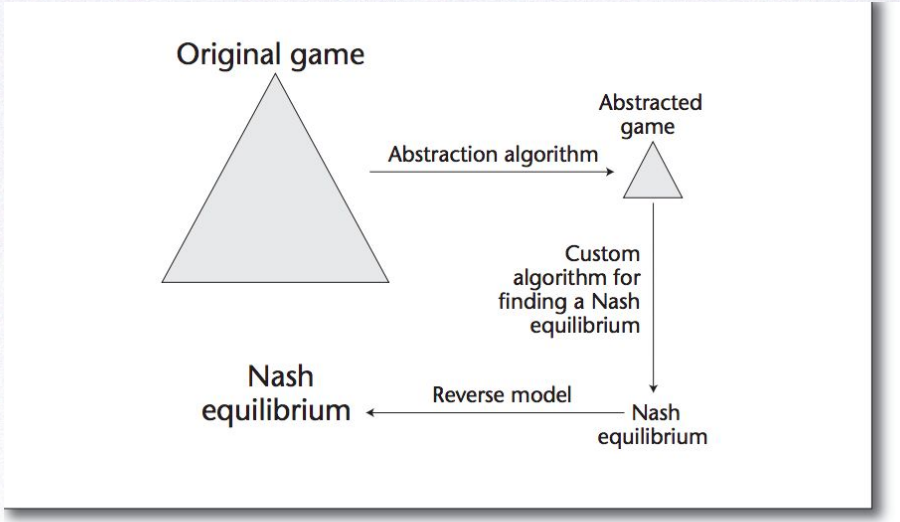


Figure 1. Current Paradigm for Solving Large Incomplete-Information Games.

Image form Sandholm '10

CFR is policy iteration!

- Policy evaluation is analogous
- Policy improvement: use regret minimization algorithms
    - Average strategies converge to Nash in self-play
- Convergence guarantees are on the *average policies*

DeepStack

(Moravcik et al. '17)

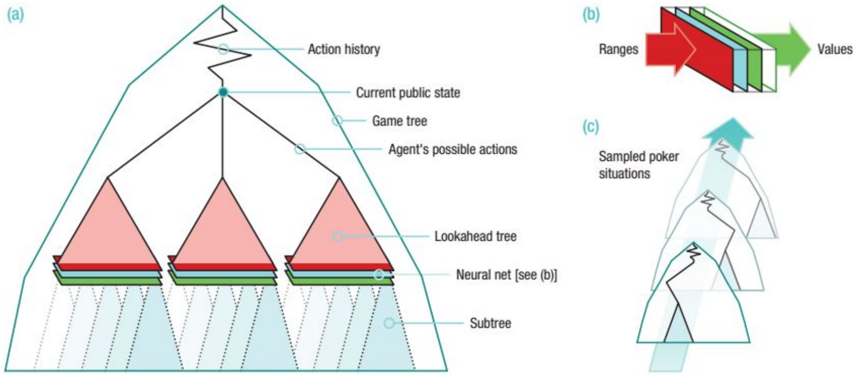
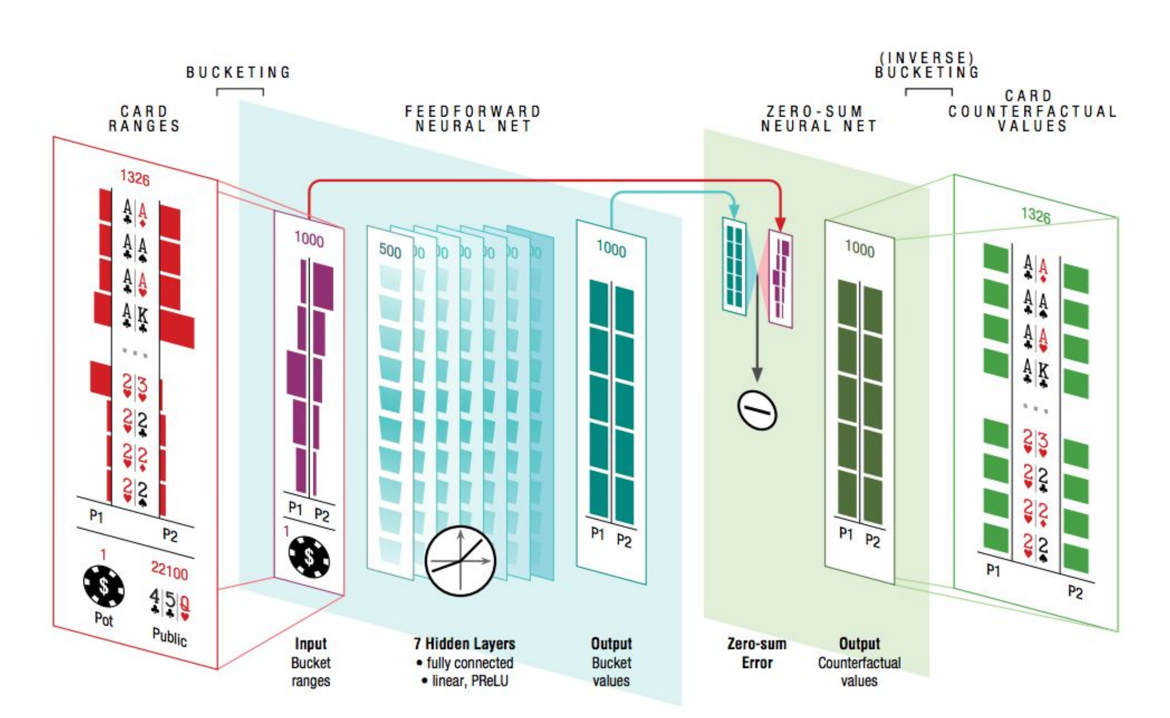


Figure 2: DeepStack overview. (a) DeepStack re-solves for its action at every public state it is to act, using a depth limited lookahead where subtree values are computed using a trained deep neural network (b) trained before play via randomly generated poker situations (c).

DeepStack

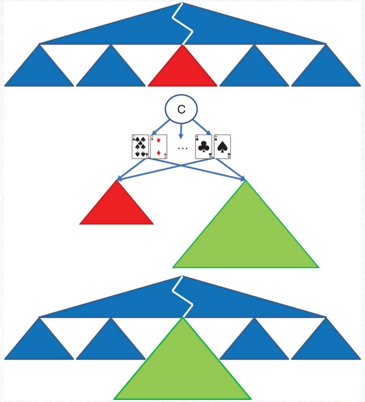
(Moravcik et al. '17)



Libratus [(Brown & Sandholm '18)]

RESEARCH ARTICLE

Superhuman AI for heads-up no-limit poker: Libratus beats top professionals



Policy Gradient Algorithms

Parameterized policy π_θ with parameters θ (e.g. a neural network)

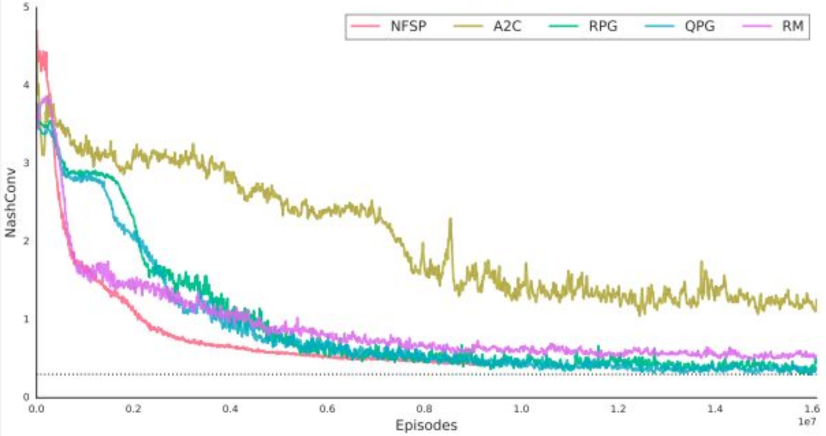
Define a *score function* $J(\pi_\theta) = v_\pi(s_0) = \mathbb{E}_\pi[G_0]$

Main idea: do gradient ascent on J.

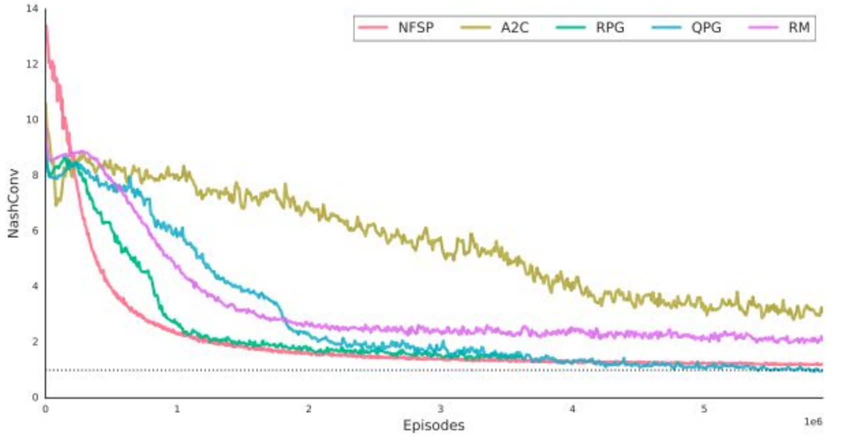
1. **REINFORCE** (Williams '92, see RL book ch. 13) + PG theorem: you can do this via estimates from sample trajectories.
2. **Advantage Actor-Critic (A2C)** (Mnih et al '16): you can use deep networks to estimate the policy *and* baseline value v(s)

Regret Policy Gradients (Srinivasan et al. '18)

- Policy gradient is doing a form of CFR minimization!
- Several new policy gradient variants inspired connection to regret



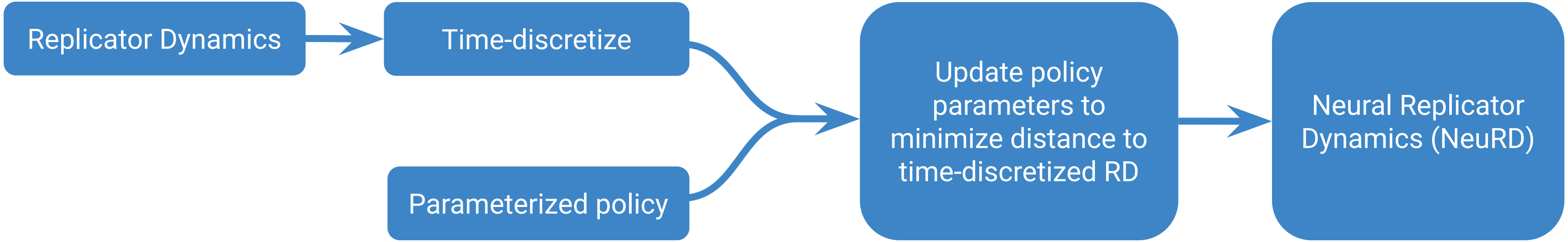
NashConv in 2-player Leduc



NashConv in 3-player Leduc

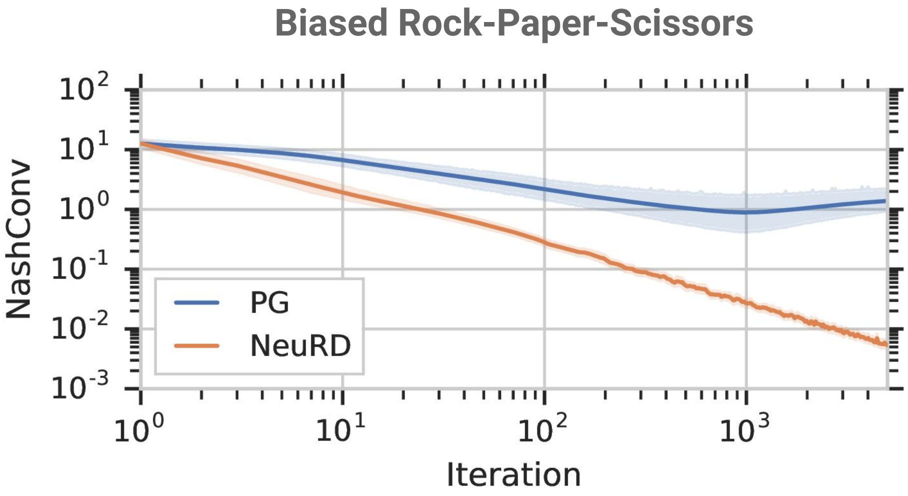
Neural Replicator Dynamics (NeuRD)

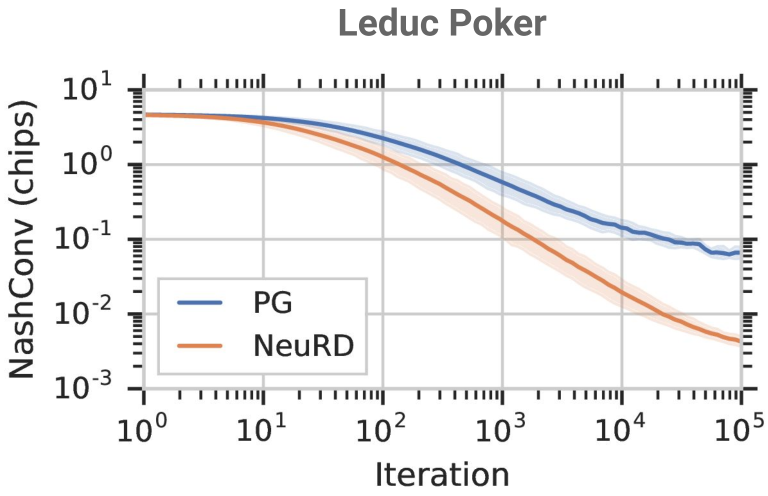
[Omidshafiei, Hennes, Morrill et al. ‘19]



$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t+1} + \eta \sum_{s,a} \nabla_{\boldsymbol{\theta}} \underbrace{y_{t-1}(s_t, a_t; \boldsymbol{\theta})}_{\text{Logits, where policy is } \boldsymbol{\pi} = softmax(\boldsymbol{y})} \underbrace{A(s_t, a_t; \boldsymbol{\theta}, \boldsymbol{w})}_{\text{Advantage } q(s,a)-v(s)}$$

NeuRD: Results



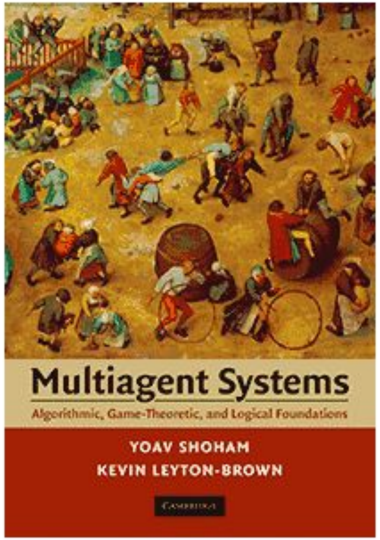


DeepMind

Where to Go From Here?

Shoham & Leyton-Brown ‘09

Main Page Table of Contents Instructional Resources Errata eBook Download*new!*



Multiagent Systems

Algorithmic, Game-Theoretic, and Logical Foundations

Yoav Shoham
Stanford University
Kevin Leyton-Brown
University of British Columbia

Cambridge University Press, 2009
Order online: amazon.com

masfoundations.org

Surveys and Food for Thought

- If multi-agent learning is the answer, what is the question?
  - Shoham et al. '06
  - Hernandez-Leal et al. '19
- A comprehensive survey of MARL (Busoniu et al. '08)
- Game Theory and Multiagent RL (Nowé et al. '12)
- Study of Learning in Multiagent Envs (Hernandez-Leal et al. '17)

The Hanabi Challenge

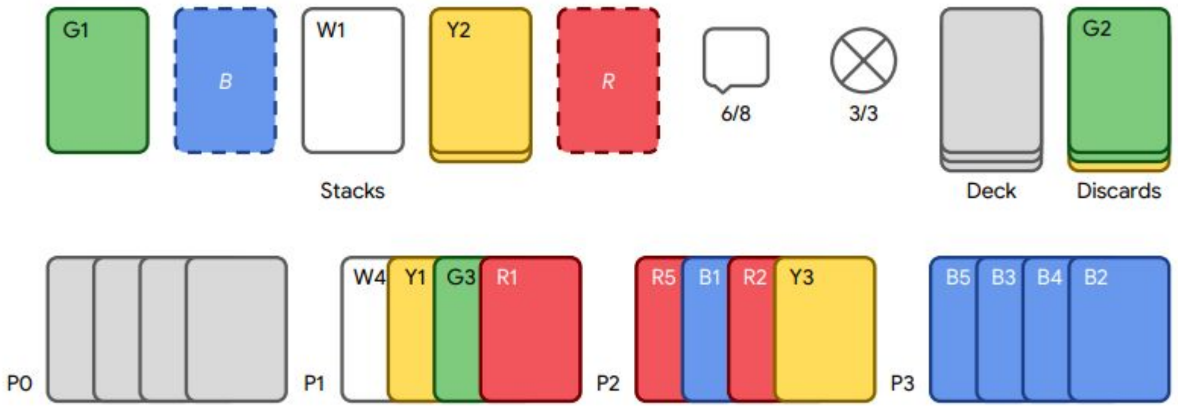
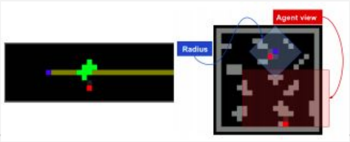


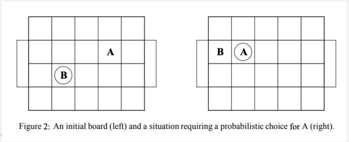
Figure 1: Example of a four player Hanabi game from the point of view of player 0. Player 1 acts after player 0 and so on.

Bard et al. '19

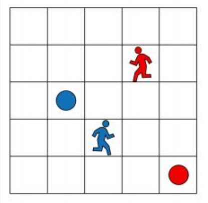
Also Competition at IEEE Cog (ieee-cog.org)

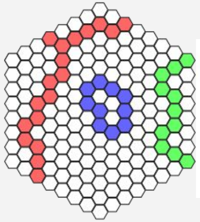
OpenSpiel: Coming Soon!

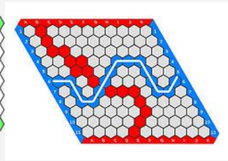




- Open source framework for research in RL & Games
- C++, Python, and Swift impl's
- 25+ games
- 10+ algorithms
- Tell all your friends! (Seriously!)







→ August 2019

AAAI 2020 Workshop on RL in Games?

AAAI19-RLG Summary:

- 39 accepted papers
  - 4 oral presentations
  - 35 posters
- 1 “Mini-Tutorial”
- 3 Invited Talks
- Panel & Discussion

http://aaai-rlg.mlanctot.info/

Questions?

Marc Lanctot

DeepMind

lanctot@google.com

mlanctot.info/

(Please contact me if you have trouble finding any references!)