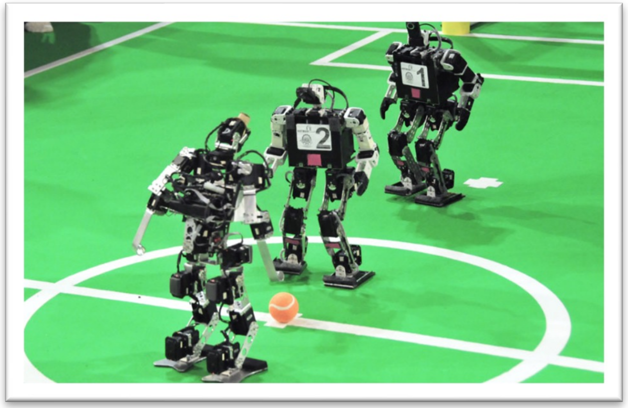
6.S890: Topics in Multiagent Learning

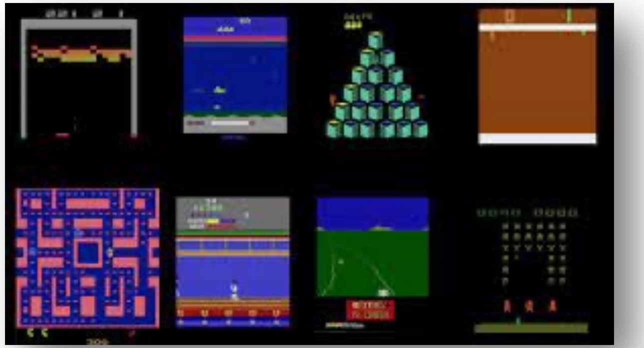
Lecture 9 – Prof. Daskalakis

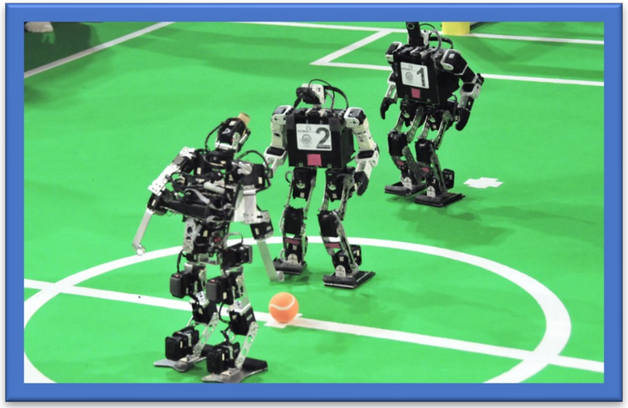
Fall 2023

Reinforcement Learning: breakthroughs & frontiers



Reinforcement Learning: breakthroughs & frontiers

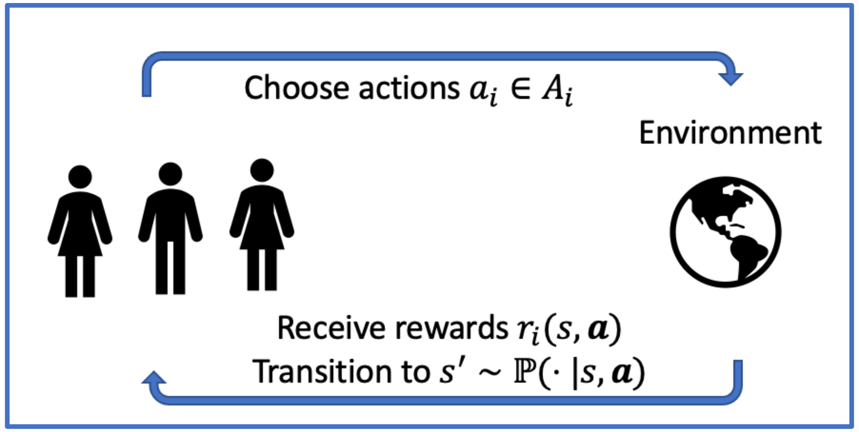




many involve
multiple players!

Lectures 9-11: investigate questions regarding equilibrium *existence*, *computation* and *learning* in multi-player RL and its underlying game-theoretic models

Stochastic Games [Shapley'53]
infinite horizon, finite states/actions

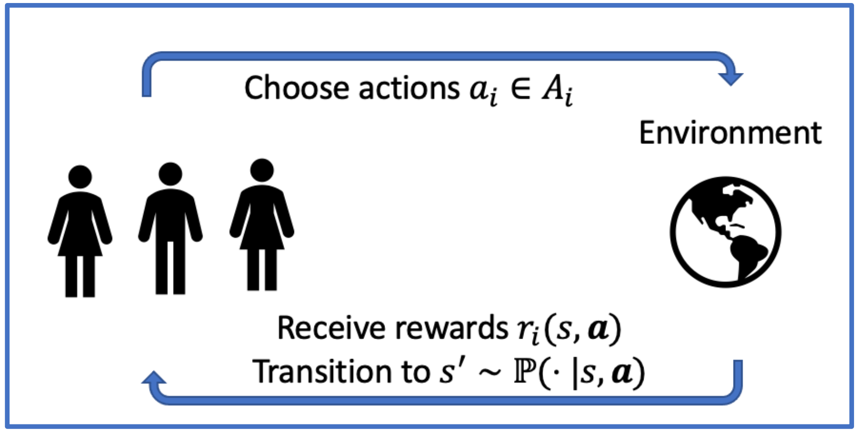


- An m-player, infinite-horizon, finite state/action space, stochastic (or Markov) game $G = (S, A, \mathbb{P}, r, \gamma, \mu)$ is specified via the following ingredients:
  - S : a finite set of **states**
  - $A = A_1 \times \cdots \times A_m$: a **joint action set**, where A_i is the finite **action set** of agent $i \in [m]$
  - $\mathbb{P}(s'|s, \boldsymbol{a})$, for $s, s' \in S$, $\boldsymbol{a} \in A$: the **transition matrix** of the environment
  - $r = (r_1, \dots, r_m)$: the **reward functions** of the environment where $r_i(s, \boldsymbol{a})$ is the **reward function** of agent i
  - $\gamma \in (0,1)$: the **discount factor**
  - $\mu \in \Delta(S)$: the **initial state distribution**
- Given an infinite state-action sequence $(s_t, \boldsymbol{a}_t)_t$ players derive discounted utilities: $u_i((s_t, \boldsymbol{a}_t)_t) = \sum_{t \geq 0} \gamma^t \cdot r_i(s_t, \boldsymbol{a}_t)$
- A randomized strategy, or policy, of player i is a function $\pi_i: S \times (S \times A)^* \to \Delta(A_i)$, mapping histories to action distributions
- Given policies $\pi_1, \dots, \pi_m$ the discounted expected utility of agent i is:

$$u_i(\pi_1, \dots, \pi_m) = \mathbb{E}_{\substack{s_0 \sim \mu \\ \boldsymbol{a}_{t,i} \sim \pi_i(\cdot | s_t, (s_\tau, \boldsymbol{a}_\tau)_{\tau < t}) \\ s_{t+1} \sim \mathbb{P}(\cdot | s_t, \boldsymbol{a}_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r_i(s_t, \boldsymbol{a}_t)\right]$$

Stochastic Games [Shapley'53]

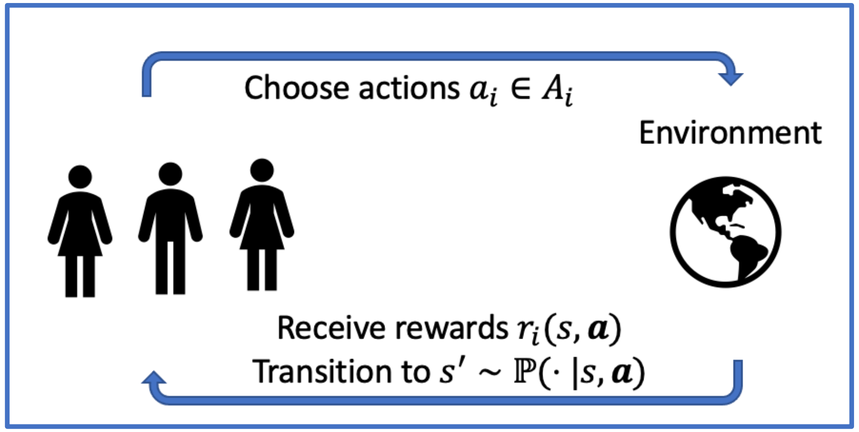
infinite horizon, finite states/actions



- In general, a policy $\pi_i: S \times (S \times A)^* \to \Delta(A_i)$ can be *history dependent*
- A policy is *history-independent* or *Markovian* if it only depends on the current state and time
  - i.e. for all $t, s, (s_\tau, \boldsymbol{a}_\tau)_{\tau=1}^{t-1}, (s'_\tau, \boldsymbol{a}'_\tau)_{\tau=1}^{t-1}$: $\pi_i(s, (s_\tau, \boldsymbol{a}_\tau)_{\tau=1}^{t-1}) = \pi_i(s, (s_\tau{}', \boldsymbol{a}_\tau{}')_{\tau=1}^{t-1})$
  - such policy can be also represented as a function $\pi_i: S \times \mathbb{N} \to \Delta(A_i)$
- A policy is stationary and Markovian if it only depends on the current state
  - such policy can be also represented as a function $\pi_i: S \to \Delta(A_i)$
- Given stationary, Markovian policies $\pi_1, \dots, \pi_m$: $u_i(\pi_1, \dots, \pi_m) = \mathbb{E}_{\substack{s_0 \sim \mu \\ \boldsymbol{a}_{t,i} \sim \pi_i(\cdot|s_t) \\ s_{t+1} \sim \mathbb{P}(\cdot|s_t, \boldsymbol{a}_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r_i(s_t, \boldsymbol{a}_t)\right]$
- **[Takahashi'64, Fink'64]:** There exists a Nash equilibrium in stationary, Markovian policies, i.e. a collection of stationary and Markovian policies $\pi_1, \dots, \pi_m$ s.t. for all i, for all (possibly history-dependent) π'_i: $u_i(\pi_i, \pi_{-i}) \geq u_i(\pi'_i, \pi_{-i})$.
- **[Shapley'53]:** In two-player zero-sum stochastic games: $\max_{\pi_1} \min_{\pi_2} u_1(\pi_1, \pi_2) = \min_{\pi_2} \max_{\pi_1} u_1(\pi_1, \pi_2)$.
- **Costis's comment:** pretty cool because $u_i(\pi_i; \pi_{-i})$ is non-concave in π_i

Stochastic Games [Shapley'53]

finite-horizon variant



- An m-player, **finite-horizon**, finite state/action space, stochastic (or Markov) game $G = (S, A, \mathbb{P}, r, H, \mu, \gamma)$ is specified via the following ingredients:
  - S : a finite set of **states**
  - $A = A_1 \times \cdots \times A_m$: a **joint action set**, where A_i is the finite **action set** of agent $i \in [m]$
  - $\mathbb{P}(s'|s, \boldsymbol{a})$, for $s, s' \in S$, $\boldsymbol{a} \in A$: the **transition matrix** of the environment
  - $r = (r_1, \dots, r_m)$: the **reward functions** of the environment where $r_i(s, \boldsymbol{a})$ is the **reward function** of agent i
  - $\boldsymbol{H} \in \mathbb{N}_+$: the **number of interaction steps**
  - $\mu \in \Delta(S)$: the **initial state distribution**
  - $\gamma \in (0,1]$: the discount factor γ; not that in contrast to the infinite-horizon setting, γ can be chosen to be 1
- Given a **finite** state-action sequence $(s_t, \boldsymbol{a}_t)_{t=1}^{\boldsymbol{H}}$ players derive discounted utilities: $u_i((s_t, \boldsymbol{a}_t)_t) = \sum_{t=0}^{\boldsymbol{H-1}} \gamma^t r_i(s_t, \boldsymbol{a}_t)$
- A randomized strategy, or policy, of player i is a function $\pi_i : S \times (S \times A)^{<\boldsymbol{H}} \to \Delta(A_i)$, mapping histories to action distributions
- Given policies $\pi_1, \dots, \pi_m$ the discounted expected utility of agent i is:

$$u_i(\pi_1, \dots, \pi_m) = \mathbb{E}_{\substack{s_0 \sim \mu \\ \boldsymbol{a}_{t,i} \sim \pi_i(\cdot|s_t, (s_\tau, \boldsymbol{a}_\tau)_{\tau<t}) \\ s_{t+1} \sim \mathbb{P}(\cdot|s_t, \boldsymbol{a}_t)}} \left[\sum_{t<H} \gamma^t \cdot r_i(s_t, \boldsymbol{a}_t)\right]$$

changes compared to the infinite horizon case in light blue

Stochastic Games: Single- vs Multi-Agent Case

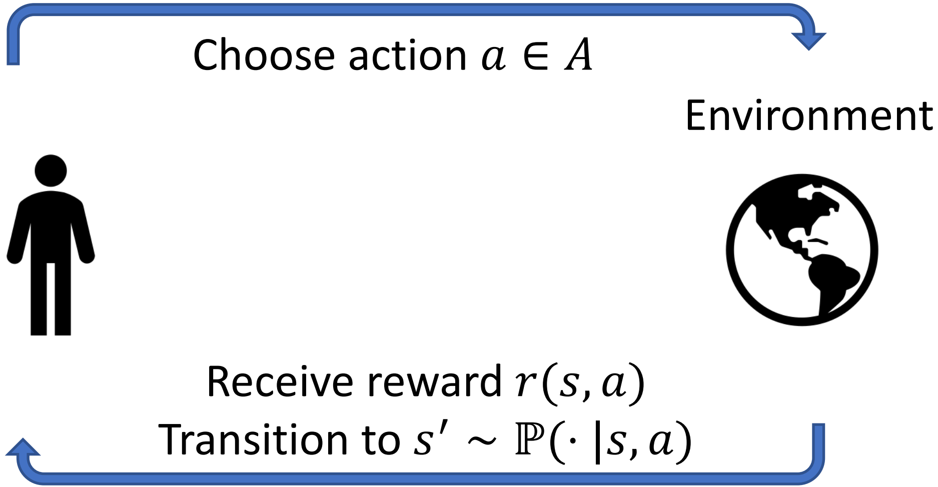
Markov Decision Process (n=1)

$$s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$$
$$r(s_t, a_t)$$

Agent's policy $\pi: S \times (S \times A)^* \to \Delta(A)$

Agent's objective:

$$u(\pi) = \mathbb{E}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot \mid s_t, (s_\tau, a_\tau)_\tau) \\ s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r(s_t, a_t)\right]$$



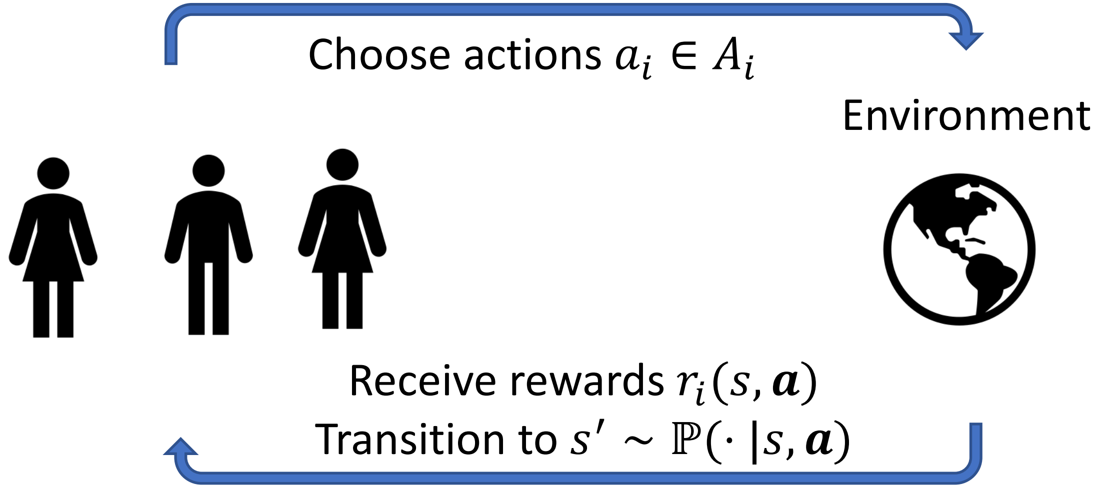
Stochastic Game (n>1)

$$s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_{t,1}, \dots, a_{t,m})$$
$$r_i(s_t, a_{t,1}, \dots, a_{t,m})$$

Agent i's policy $\pi_i: S \times (S \times A)^* \to \Delta(A_i)$

Agent i's objective:

$$u_i(\pi) = \mathbb{E}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot \mid s_t, (s_\tau, \boldsymbol{a}_\tau)_\tau) \\ s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, \boldsymbol{a}_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r_i(s_t, \boldsymbol{a}_t)\right]$$



Stochastic Games: Single- vs Multi-Agent Case

Markov Decision Process (n=1)

$$s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$$
$$r(s_t, a_t)$$

Agent's policy $\pi: S \times (S \times A)^* \to \Delta(A)$

Agent's objective:

$$u(\pi) = \mathbb{E}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot \mid s_t, (s_\tau, a_\tau)_\tau) \\ s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r(s_t, a_t)\right]$$

Folklore Result: exists optimal policy that is stationary and Markovian

- optimal policy can be found using Linear Programming
- also using policy iteration/value iteration methods

Stochastic Game (n>1)

$$s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_{t,1}, \dots, a_{t,m})$$
$$r_i(s_t, a_{t,1}, \dots, a_{t,m})$$

Agent i's policy $\pi_i: S \times (S \times A)^* \to \Delta(A_i)$

Agent i's objective:

$$u_i(\pi) = \mathbb{E}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot \mid s_t, (s_\tau, \boldsymbol{a}_\tau)_\tau) \\ s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, \boldsymbol{a}_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r_i(s_t, \boldsymbol{a}_t)\right]$$

Corresponding Result: ∃ Nash eq in stationary Markovian policies

- computing Nash equilibrium: PPAD-hard
- in zero-sum games: open in general; tractable if discount factor bounded away from 1 and goal is approximate min-max
- correlated equilibria: open in general; some hardness results, depending on type
- more tractable when game is finite horizon

Stochastic Games: Planning vs Learning

Markov Decision Process (n=1)

$$s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)$$
$$r(s_t, a_t)$$

Agent's policy $\pi: S \times (S \times A)^* \to \Delta(A)$

Agent's objective:

$$u(\pi) = \mathbb{E}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot | s_t, (s_\tau, a_\tau)_\tau) \\ s_{t+1} \sim \mathbb{P}(\cdot | s_t, a_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r(s_t, a_t) \right]$$

Planning: find a good policy with knowledge of environment i.e. dynamics & rewards

Reinforcement Learning: find a good policy without a priori knowledge (or at least not complete knowledge) of the environment

- by interacting with environment
- or with simulator access to the environment
- or with enough offline data

RL through Q-learning, policy gradient methods,…

Stochastic Game (n>1)

$$s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_{t,1}, \dots, a_{t,m})$$
$$r_i(s_t, a_{t,1}, \dots, a_{t,m})$$

Agent i's policy $\pi_i: S \times (S \times A)^* \to \Delta(A_i)$

Agent i's objective:

$$u_i(\pi) = \mathbb{E}_{\substack{s_0 \sim \mu \\ a_t \sim \pi(\cdot | s_t, (s_\tau, \boldsymbol{a}_\tau)_\tau) \\ s_{t+1} \sim \mathbb{P}(\cdot | s_t, \boldsymbol{a}_t)}} \left[\sum_{t \geq 0} \gamma^t \cdot r_i(s_t, \boldsymbol{a}_t) \right]$$

Distinction between planning and learning similar

extra complication: do agents observe each other's actions? can agents communicate?

Multi-Agent Reinforcement Learning

less well explored

Algorithms/Learning/Complexity: next week (guest: Noah Golowich)

Equilibrium Existence Results: this week

Equilibrium Existence:
Finite Horizon Stochastic Games

Proposition: Exists Nash equilibrium in Markovian policies

Proof: via “backwards induction”

- Construct Nash equilibrium policies inductively, starting at $t = H - 1$ (last interaction round) and proceeding backwards
  - I.e. for all i’s together compute $\pi_i(\cdot \mid s, t)$ from $t = H - 1$ down to 0
  - Auxiliary variables constructed inductively $V_{i,t}(s)$: continuation value that player i expects to receive if they were to start at state s at time t under selected Nash equilibrium policies at times t, t+1,…

Base Case:

$V_{i,H}(s) \leftarrow 0$ for all s, i

Inductive step ($\boldsymbol{t = H - 1, \ldots, 0}$)

1. Assume given $V_{i,t+1}: S \to \mathbb{R}$
2. For each $s \in S$, construct a game where i’s utility $F_{is}: A \to \mathbb{R}$ is as shown at right
3. Compute a Nash equilibrium of the game $(F_{1s}, \ldots, F_{ms})$, and let that be $\pi(\cdot \mid s, t) \in \Delta(A)$
4. Let $V_{i,t}(s) \coloneqq \mathbb{E}_{\boldsymbol{a} \sim \pi(\cdot|s,t)}[F_{is}(\boldsymbol{a})]$.

$a_2 \in A_2$

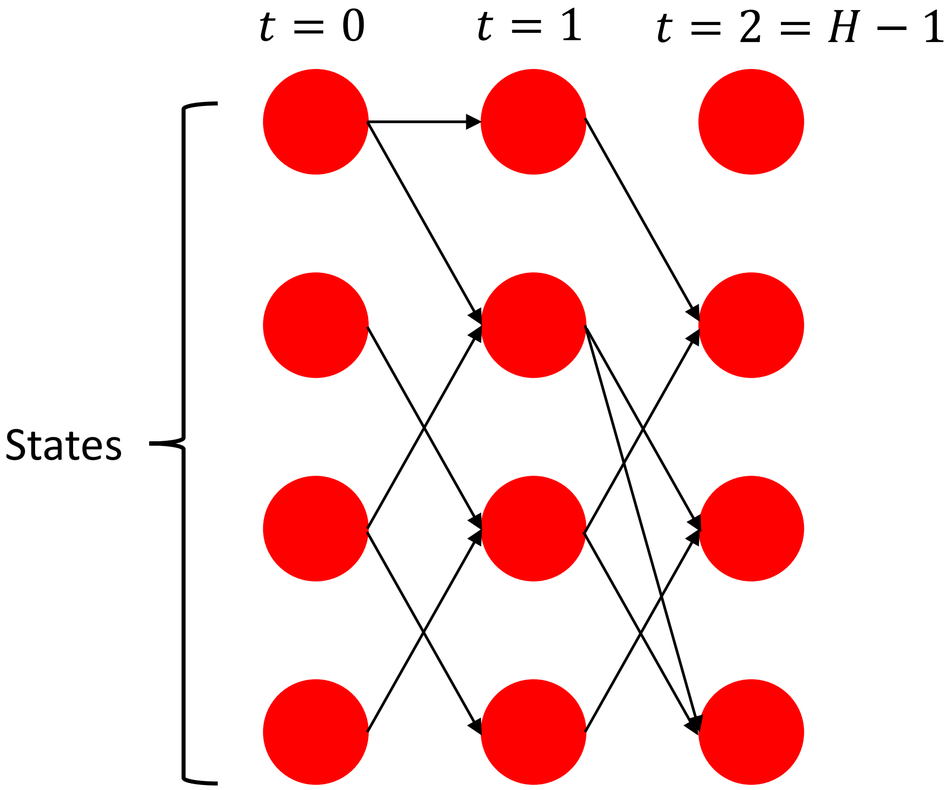
$a_1 \in A_1$					
			$F_{is}(a_1, a_2)$		

$F_{is}(\boldsymbol{a}) \coloneqq r_i(s, \boldsymbol{a}) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,\boldsymbol{a})}[V_{i,t+1}(s')]$

Equilibrium Existence:
Finite-Horizon Stochastic Games

Construct Nash equilibrium policies inductively, starting at $t = H - 1$ (last interaction round) and proceeding backwards

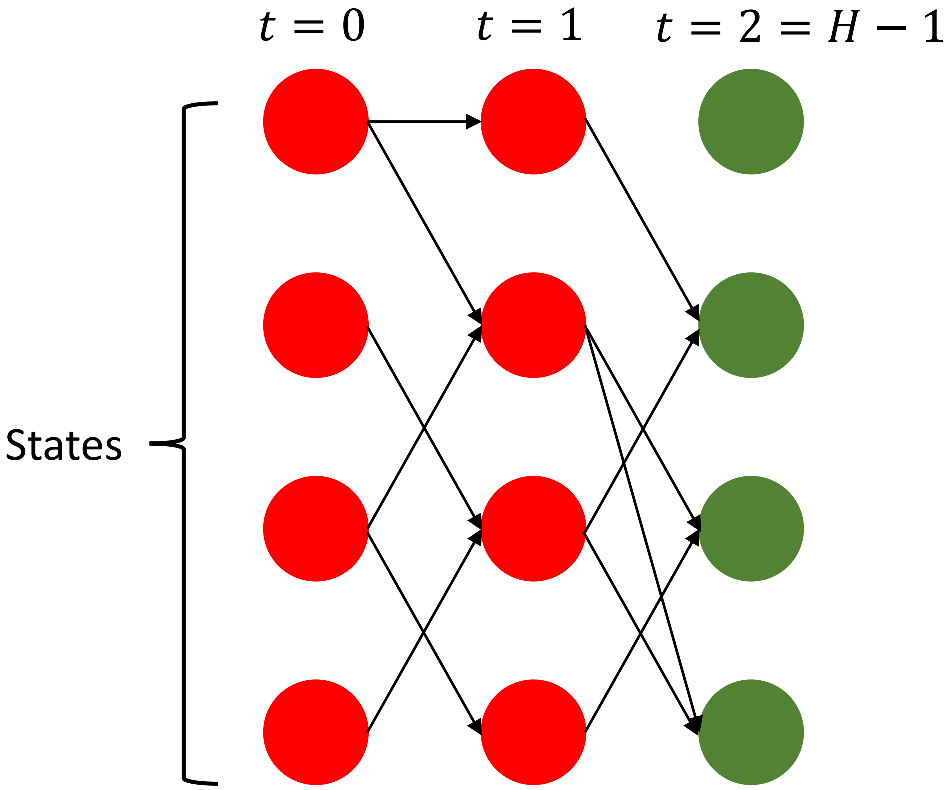
- I.e. for all i, compute $\pi_i(\cdot\,|s, t)$ from $t = H - 1$ down to 0
- Auxiliary variables constructed inductively $V_{i,t}(s)$: continuation value of player i under Nash equilibrium



Equilibrium Existence: Finite-Horizon Stochastic Games

Construct Nash equilibrium policies inductively, starting at $t = H - 1$ (last interaction round) and proceeding backwards

- I.e. for all i, compute $\pi_i(\cdot\,|s,t)$ from $t = H - 1$ down to 0
- Auxiliary variables constructed inductively $V_{i,t}(s)$: continuation value of player i under Nash equilibrium



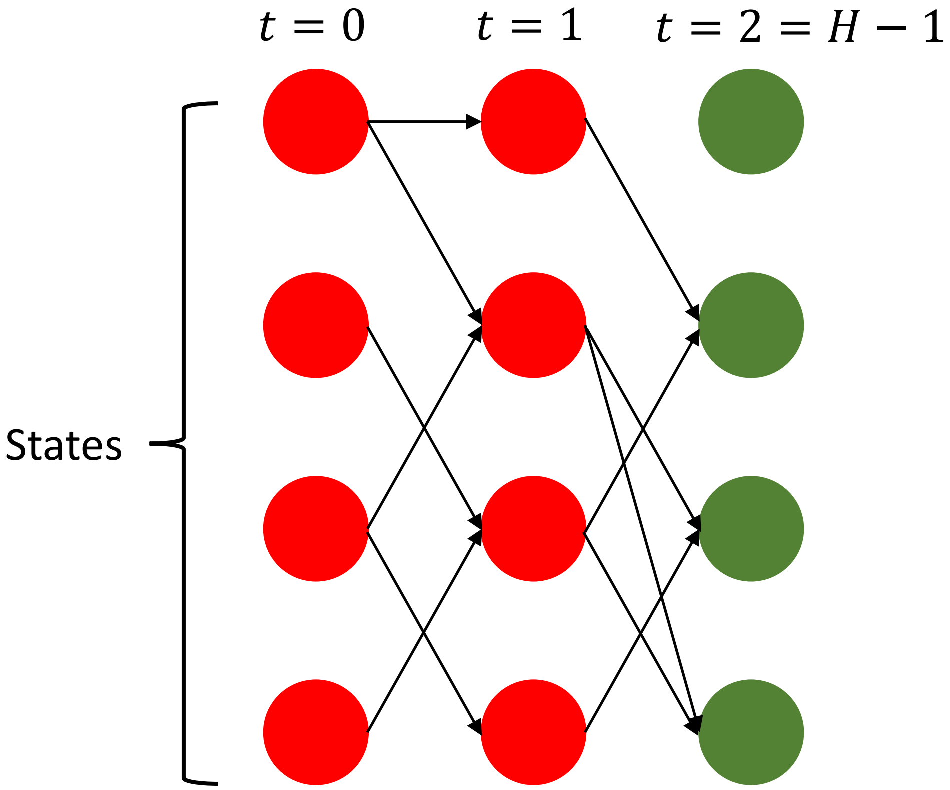
Base Case:

$V_{i,H}(s) \leftarrow 0$ for all s, i

Equilibrium Existence: Finite-Horizon Stochastic Games

Construct Nash equilibrium policies inductively, starting at $t = H - 1$ (last interaction round) and proceeding backwards

- I.e. for all i, compute $\pi_i(\cdot \mid s, t)$ from $t = H - 1$ down to 0
- Auxiliary variables constructed inductively $V_{i,t}(s)$: continuation value of player i under Nash equilibrium



Base Case:

$V_{i,H}(s) \leftarrow 0$ for all s, i

Inductive step:

1. Assume given $V_{i,t+1}: S \to \mathbb{R}$ (e.g., $t = 1$)
2. For each $s \in S$, player $i \in [m]$, define local payoff function $F_{is}: A \to \mathbb{R}$:

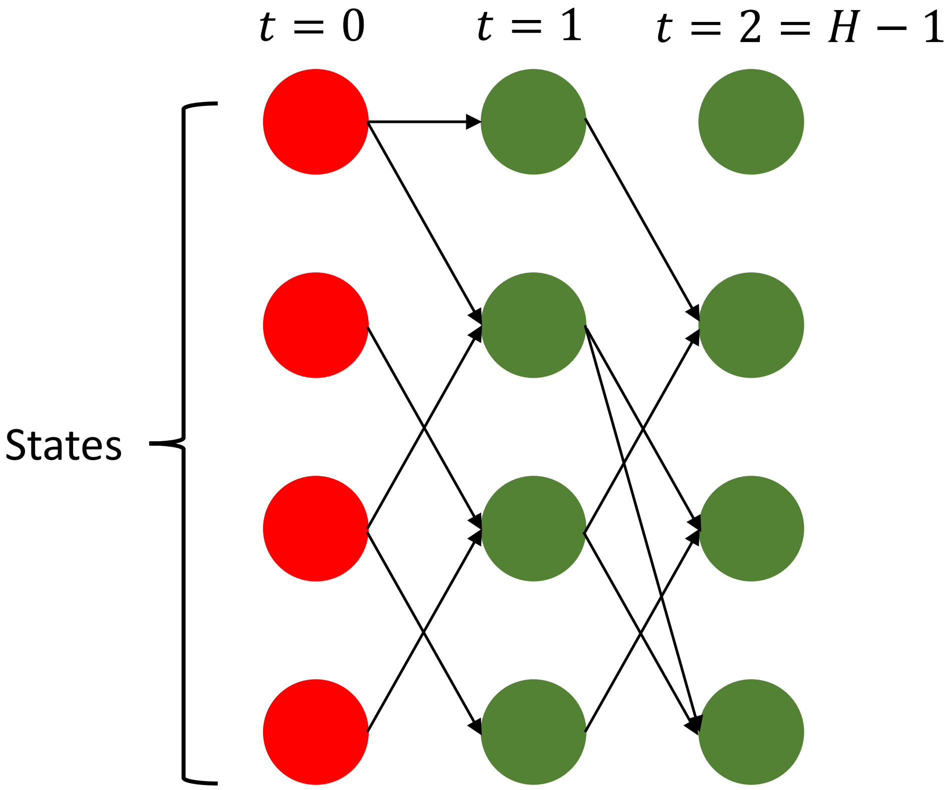
$$F_{is}(\boldsymbol{a}) := r_i(s, \boldsymbol{a}) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s, \boldsymbol{a})}[V_{i,t+1}(s')]$$

3. Compute a Nash equilibrium of game $(F_{1s}, \dots, F_{ms})$ at each state s, and let that be $\pi(\cdot \mid s, t) \in \Delta(A)$

Equilibrium Existence: Finite-Horizon Stochastic Games

Construct Nash equilibrium policies inductively, starting at $t = H - 1$ (last interaction round) and proceeding backwards

- I.e. for all i, compute $\pi_i(\cdot\,|s,t)$ from $t = H - 1$ down to 0
- Auxiliary variables constructed inductively $V_{i,t}(s)$: continuation value of player i under Nash equilibrium



Base Case:

$V_{i,H}(s) \leftarrow 0$ for all s, i

Inductive step:

1. Assume given $V_{i,t+1}: S \to \mathbb{R}$ (e.g., $t = 1$)
2. For each $s \in S$, player $i \in [m]$, define local payoff function $F_{is}: A \to \mathbb{R}$:

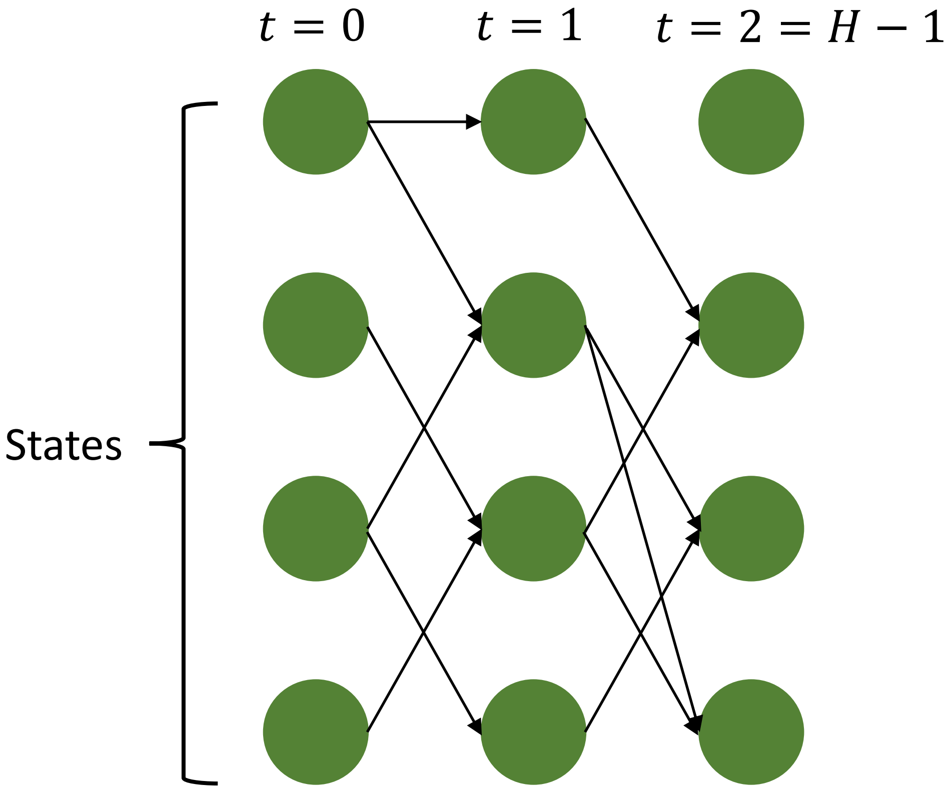
$$F_{is}(\boldsymbol{a}) \coloneqq r_i(s, \boldsymbol{a}) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,\boldsymbol{a})}[V_{i,t+1}(s')]$$

3. Compute a Nash equilibrium of game $(F_{1s}, \dots, F_{ms})$ at each state s, and let that be $\pi(\cdot\,|s,t) \in \Delta(A)$
4. Let $V_{i,h}(s) \coloneqq \mathbb{E}_{\boldsymbol{a} \sim \pi(\cdot|s,t)}[F_{is}(\boldsymbol{a})]$

Equilibrium Existence: Finite-Horizon Stochastic Games

Construct Nash equilibrium policies inductively, starting at $t = H - 1$ (last interaction round) and proceeding backwards

- I.e. for all i, compute $\pi_i(\cdot \mid s, t)$ from $t = H - 1$ down to 0
- Auxiliary variables constructed inductively $V_{i,t}(s)$: continuation value of player i under Nash equilibrium



Base Case:

$V_{i,H}(s) \leftarrow 0$ for all s, i

Inductive step:

1. Assume given $V_{i,t+1}: S \to \mathbb{R}$ (e.g., $t = 1$)
2. For each $s \in S$, player $i \in [m]$, define local payoff function $F_{is}: A \to \mathbb{R}$:

$$F_{is}(\boldsymbol{a}) \coloneqq r_i(s, \boldsymbol{a}) + \mathbb{E}_{s' \sim \mathbb{P}(\cdot \mid s, \boldsymbol{a})}[V_{i,t+1}(s')]$$

3. Compute a Nash equilibrium of game $(F_{1s}, \dots, F_{ms})$ at each state s, and let that be $\pi(\cdot \mid s, t) \in \Delta(A)$
4. Let $V_{i,h}(s) \coloneqq \mathbb{E}_{\boldsymbol{a} \sim \pi(\cdot \mid s, t)}[F_{is}(\boldsymbol{a})]$

Exercise: why are inductively computed policies a Nash equilibrium?

Equilibrium Existence: Infinite-Horizon Stochastic Games

[Takahashi'64, Fink'64]: There exists a Nash equilibrium in stationary, Markovian policies, i.e. a collection of stationary and Markovian policies $\pi_1, \ldots, \pi_m$ s.t. for all i, for all (possibly history-dependent) π_i': $u_i(\pi_i, \pi_{-i}) \geq u_i(\pi_i', \pi_{-i})$.

Proof: on the board