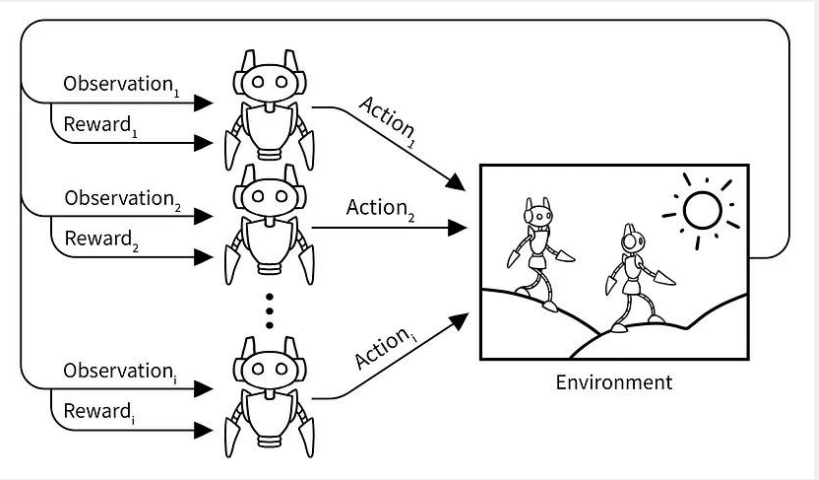
Fictitious Play in Product Markov Games with Kullback-Leibler Control Cost

Khaled Nakhleh

Texas A&M University, College Station
IEEE Asilomar conference on signals, systems & computers
October 28, 2025

A multi-agent system (MAS)

Multi-agent system: a system where separate entities take decisions autonomously and interact within a shared environment.

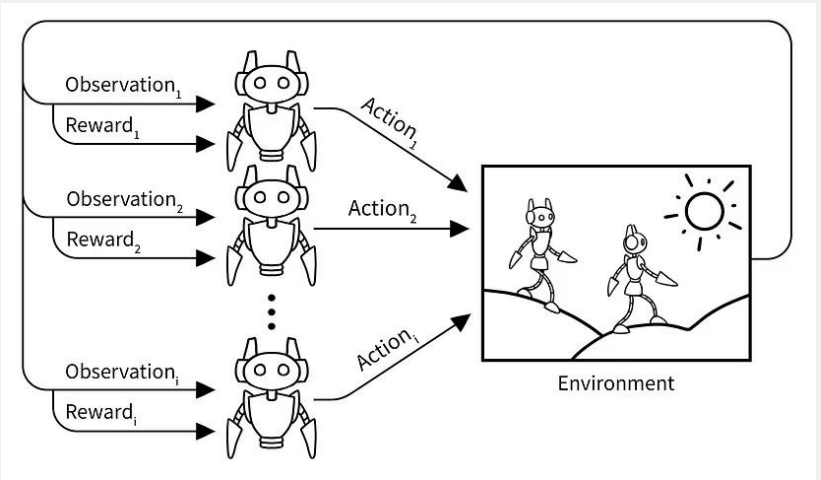


MAS illustration. Copyright Justin Terry 2021.

Sven Gronauer and Klaus Diepold. In: *Artificial Intelligence Review* (2022)

A multi-agent system (MAS)

Autonomous vehicles

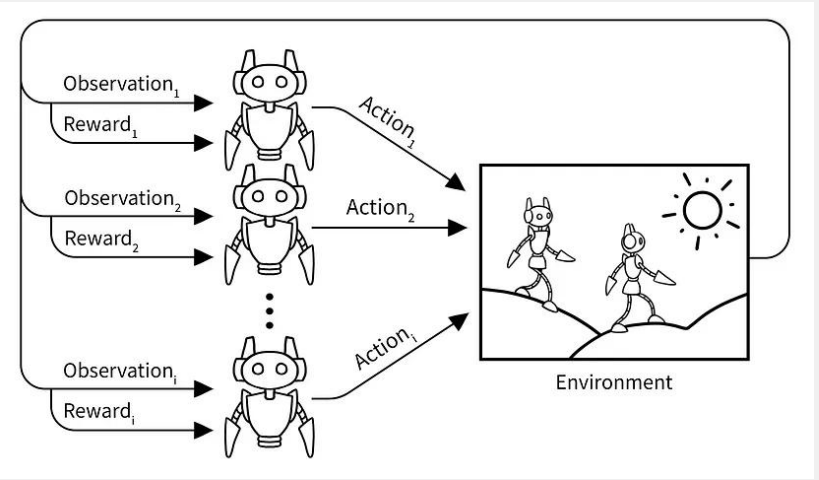


MAS illustration. Copyright Justin Terry 2021.

Ming Zhou et al. In: *Conference on robot learning*. PMLR. 2021

A multi-agent system (MAS)

Optimizing communications networks

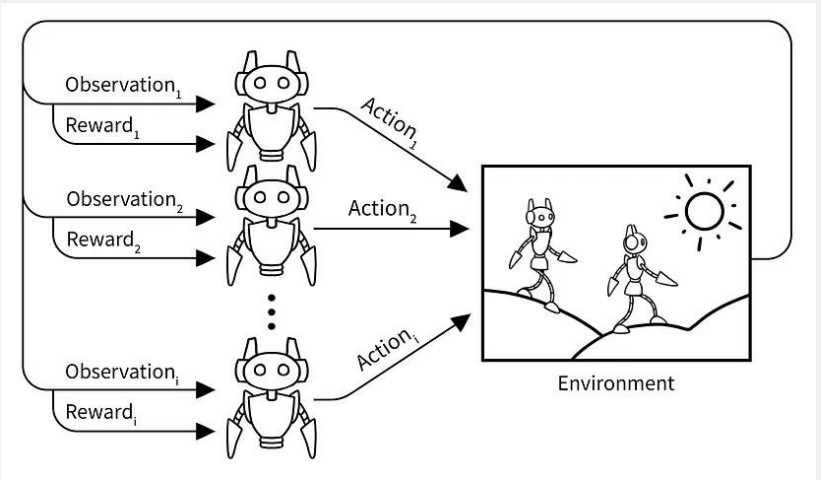


MAS illustration. Copyright Justin Terry 2021.

Nguyen Cong Luong et al. In: *IEEE communications surveys & tutorials* (2019)

A multi-agent system (MAS)

Internet marketing

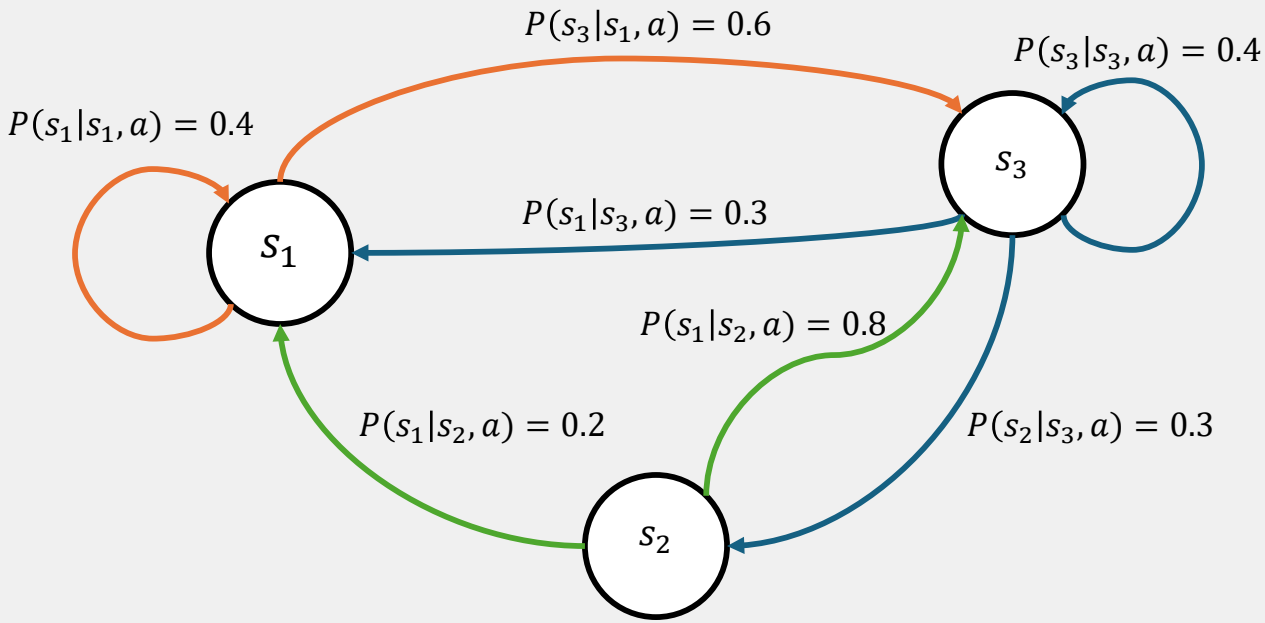


MAS illustration. Copyright Justin Terry 2021.

Junqi Jin et al. In: *Proceedings of the 27th ACM international conference on information and knowledge management*. 2018

Markov Decision Processes (MDPs) and Markov Games

Model decision-making as **Multi-agent MDPs** or **Markov games**



Underlying Markov chain example

Markov Decision Processes (MDPs) and Markov Games

For multi-agent (and single-agent) MDPs:

- **Goal:** solve for the optimal V^* and an optimal π^*
- **Solvers:**
  - value iteration, policy iteration [Zhang et al. 2023]
  - Learning methods such as Q-learning

Yizhou Zhang et al. In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* (2023)

Markov Decision Processes (MDPs) and Markov Games

For Markov games:

- **Goal:** solve for a Nash equilibrium point π^*
- **Solvers:** depends on the game setting
  - **Identical-interest Markov games:**
    monolithic MDP then standard MDP solvers [Unlu et al. 2023]
  - **Zero-sum two-agent Markov games:**
    Minimax value using a contraction operator [Shapley 1953]

Onur Unlu and Muhammed O Sayin. In: *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE. 2023

Lloyd S Shapley. In: *Proceedings of the national academy of sciences* (1953)

Markov Decision Processes (MDPs) and Markov Games

- **General-sum Markov games:**
  Case when $\text{reward}_i + \text{reward}_j \neq 0$ for any two agents $i, j \in \mathcal{N}$?
- Our approach:
  exploit structure in state transitions and the cost function

Contribution

- Product Markov games: agent controls their own transitions
- KL control provides a tractable convergence proof
- New learning dynamics for Markov games variant
  - Addresses different rewards or cost functions per agent

Fictitious Play in Markov games
with Kullback-Leibler Control Cost

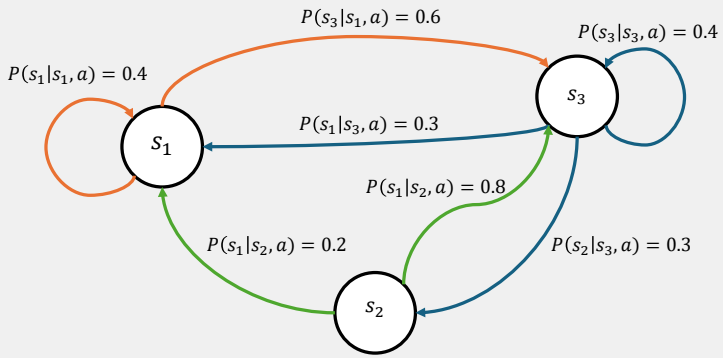
Joint work with:

Sarper Aydin Ceyhun Eksin Sabit Ekin

Problem Setting

A discounted multi-agent MDP or a Markov game is defined as

- $\mathcal{N}$: agents' set
- $\mathcal{S}_i$: agent i state space
- $\mathcal{A}_i$: agent i action space
- q_i : agent i one-step cost (reward) function
- $\gamma \in [0, 1)$: discount factor
- P : probability transition function
- ρ : states' initial distribution



MDP example.

Problem Setting

- Agent i base policy $P_{i,0}$ of the underlying Markov chain
- Similar to Markov product games [Flesch 2008]

$$P_0(s'|s) = \prod_{i=1}^{n} P_{i,0}(s_i'|s)$$

- Agent i KL control cost

$$D_{KL}[\pi_i(\cdot|s)||P_{i,0}(\cdot|s)] := \sum_{s_i' \in \mathcal{S}} \pi_i(s_i'|s) \ln\left(\frac{\pi_i(s_i'|s)}{P_{i,0}(s_i'|s)}\right)$$

János Flesch, Gijs Schoenmakers, and Koos Vrieze. In: *Mathematics of Operations Research* (2008)

Problem Setting

- V_i is different from V_j for any $i, j \in \mathcal{N}$

$$V_i^{\pi}(s) := \mathop{\mathbb{E}}_{s\sim\pi(\cdot|s)}\Big[\sum_{t=0}^{\infty}\gamma^t\Big[C_i(s_t) + D_{KL}\Big(\pi_i(\cdot|s_t)||P_{i,0}(\cdot|s_t)\Big)\Big]|s_{t=0} = s\Big].$$

$$V_j^{\pi}(s) := \mathop{\mathbb{E}}_{s\sim\pi(\cdot|s)}\Big[\sum_{t=0}^{\infty}\gamma^t\Big[C_j(s_t) + D_{KL}\Big(\pi_j(\cdot|s_t)||P_{j,0}(\cdot|s_t)\Big)\Big]|s_{t=0} = s\Big].$$

Solution concept

Solution concept: Nash Equilibrium (NE)

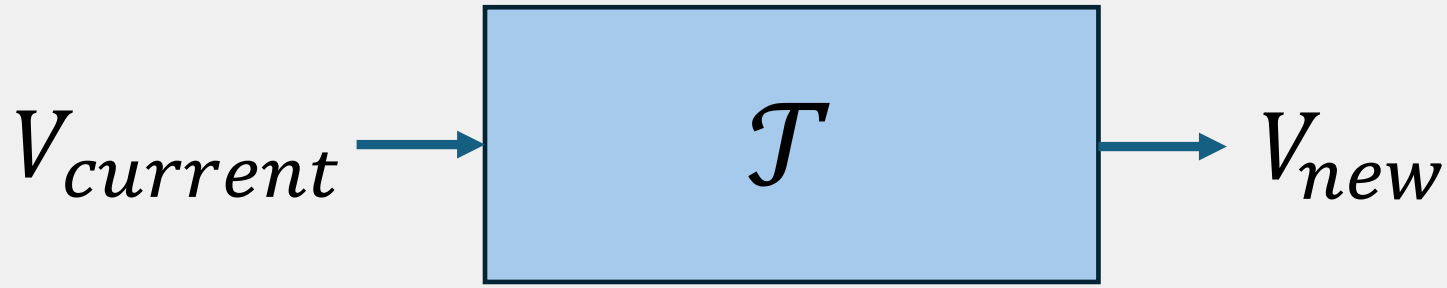
Joint control policy $\pi = (\pi_i, \pi_{-i})$ is a ϵ-NE if and only if

$$V_i^{(\pi_i, \pi_{-i})} \leq V_i^{(\pi_i', \pi_{-i})} + \epsilon,$$

for all $i \in \mathcal{N}$ and $\pi_i' \in \Pi_i$.

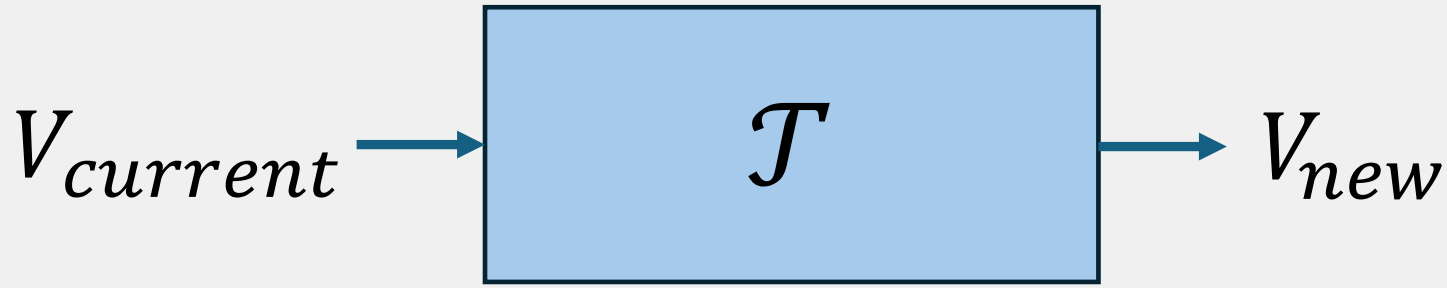
- Suitable for Markov games with different value functions
- Also suitable for product policies

Operator $\mathcal{T}$ for Updating the Value Function



- $\mathcal{T} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ maps value functions into a new estimate
- $V_{new} = \mathcal{T}^{(\pi_i, \pi_{-i})} V_{current}$
- Maps to a different V_{new} given the policies (π_i, π_{-i})
- Mappings happen over iterations $k = 0, 1, 2, \ldots$

Operator $\mathcal{T}$ for Updating the Value Function



- $\mathcal{T} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ maps value functions into a new estimate
- $V_{new} = \mathcal{T}^{(\pi_i, \pi_{-i})} V_{current}$
- Maps to a different V_{new} given the policies (π_i, π_{-i})
- Mappings happen over iterations $k = 0, 1, 2, \ldots$

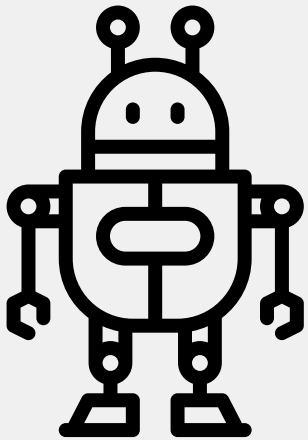
Challenges when $V_i \neq V_j$

- Agent i does not know the policies $\pi_{j,k}$ of $j \in \mathcal{N} \setminus \{i\}$
- Agent i cannot perform $V_{i,k+1} = \mathcal{T}^{(\pi_{i,k}, \pi_{-i,k})} V_{i,k}$
- Our approach: agent i construct a belief $\pi_{j,k}^{(i)}$ on agent j policy

Optimal Control Policy to the Beliefs' Illustration

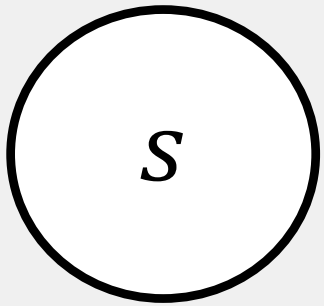
Consider two agents and a joint state $s = [s_i, s_j]$

Agent i

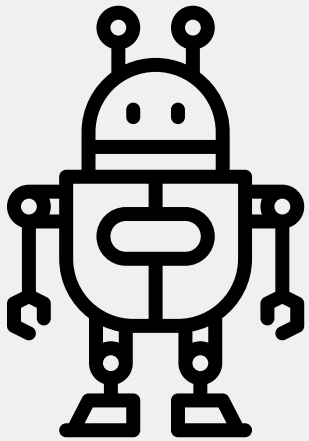


choices

s_1, s_3, s_6



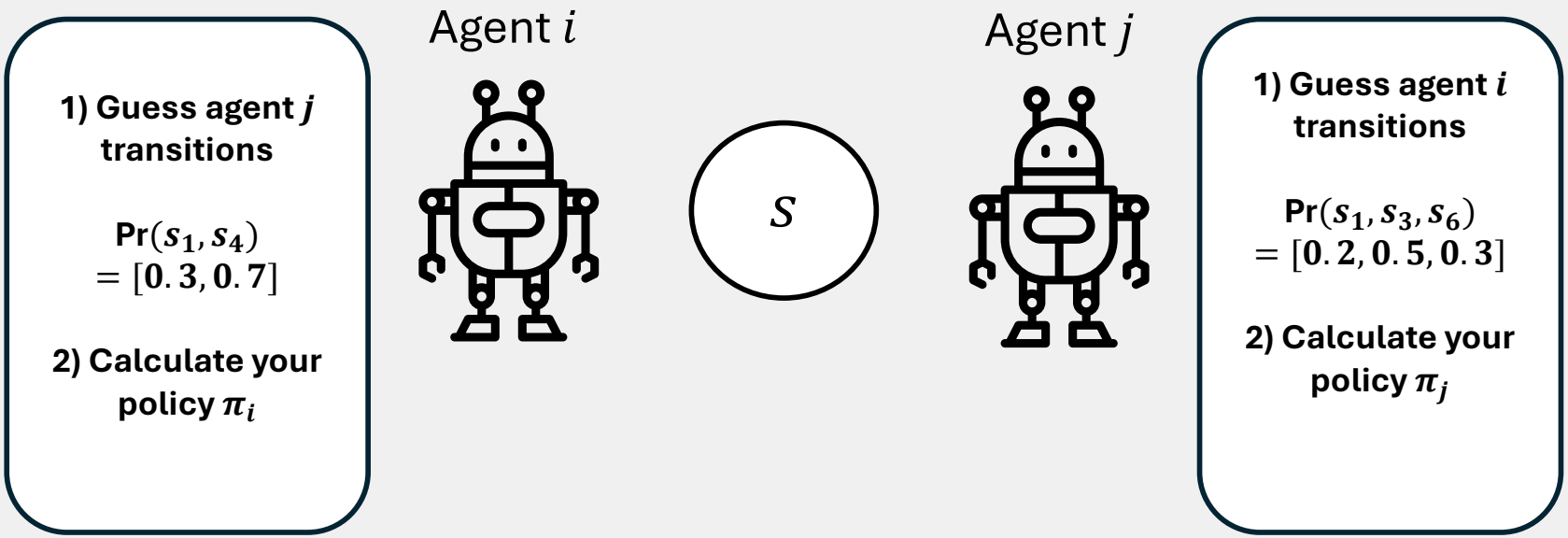
Agent j



choices

s_1, s_4

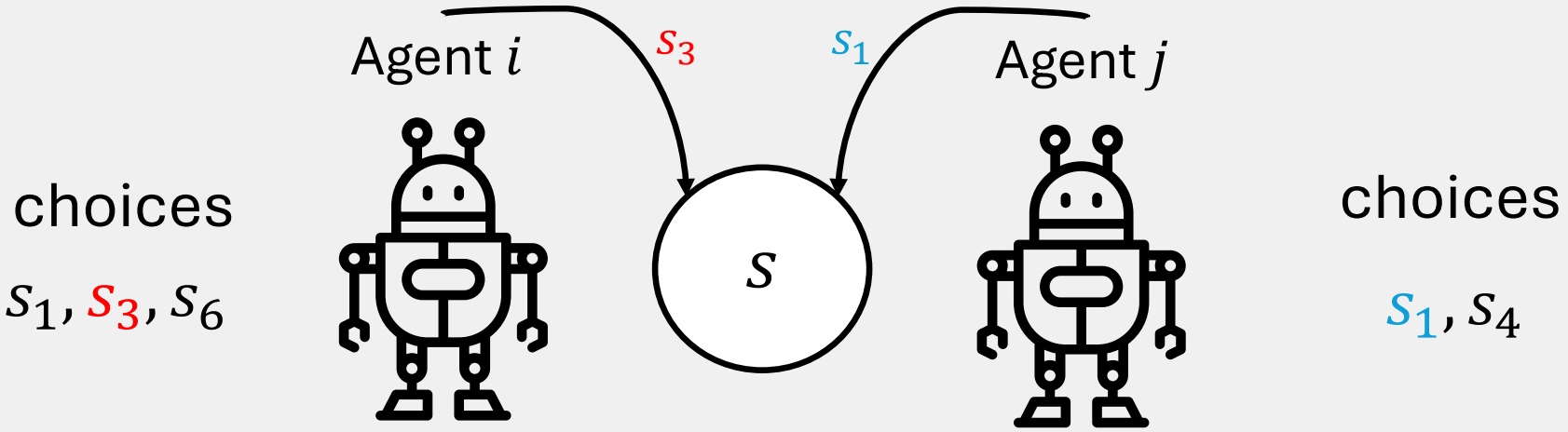
Optimal Control Policy to the Beliefs' Illustration



Each agent only knows their value function V and control policy π

Optimal Control Policy to the Beliefs' Illustration

Agents reveal their choices and update their beliefs



Agent i belief goes from $[0.3, 0.7]$ to $[0.4, 0.6]$

Agent j belief goes from $[0.2, 0.5, 0.3]$ to $[0.15, 0.6, 0.25]$

Fictitious Play Learning Dynamics

- Two timescales: learning rate β_k and learning rate α_k
- **First step:** update the value function

$$V_{i,new}(s) = V_{i,current}(s) + \beta_k \Big([\mathcal{T} V_{i,current}](s) - V_{i,current}(s) \Big)$$

- **Second step:** Update the belief

$$\pi^{(i)}_{j,new}(\cdot|s) = \pi^{(i)}_{j,current}(\cdot|s) + \alpha_k \Big(\text{realized sub-state} - \pi^{(i)}_{j,current}(\cdot|s) \Big)$$

Networked updates: centralized agent updates and transmits over network

Asymptotic Convergence

Theorem: Convergence to an NE

For agent i, and given

- $\alpha_k \geq \beta_k \Rightarrow$ two-timescale updates
- $\sum_{k=0}^{K} \alpha_k = \infty, \sum_{k=0}^{K} \beta_k = \infty$ (non-summable)
- $\sum_{k=0}^{K} (\alpha_k)^2 < \infty, \sum_{k=0}^{K} (\beta_k)^2 < \infty$ (square-summable)
- $\alpha_k \to 0$, $\beta_k \to 0$ as $k \to \infty$

Example: $\alpha_k = \frac{\ln(k)}{k}, \beta_k = \frac{1}{k}$

Asymptotic Convergence

Theorem: Convergence to an NE

Then value function $V_{i,k}$ converges to a fixed point state-wise associated with an NE π^*

- First convergence result for product Markov games
- Extension from single controller setting [Sayin 2022] to product policies setting

Muhammed O. Sayin, Kaiqing Zhang, and Asuman Ozdaglar. In: ACM EC, 2022

Belief convergence result

Result: Value function improvement

For agent i and given the learning rates' condition, we have

$$\lim \inf_{k\to\infty} \sum_{k}^{k+1} y_{i,k} = \lim \inf_{k\to\infty} \sum_{k}^{k+1} \beta_k \left((\mathcal{T}^{(\pi_{i,k}, \pi^{(i)}_{-i,k})} V_{i,k})(s) - V_{i,k}(s) \right) \leq 0$$

- Lemma goal: estimates either improve or remain the same
- KL control allows to define the tractable difference $\hat{y}_{i,k}$
- $\hat{y}_{i,k}$ upper bounds the updates $y_{i,k}$

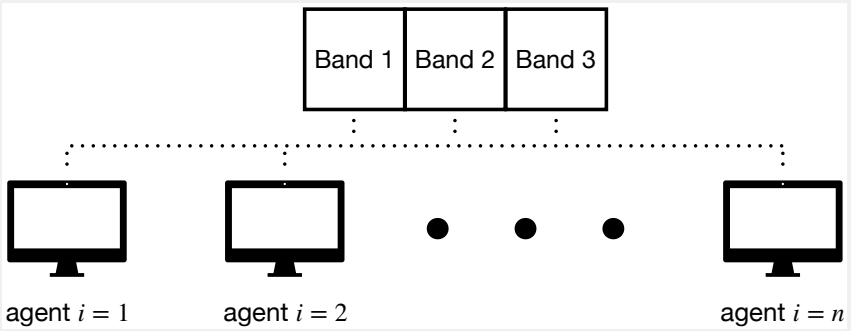
Belief convergence result

- Main goal: beliefs converge to agents' policies asymptotically
- approach: beliefs lie in the limit set of the policies
  $\dot{\pi}^{(i)}_{i,k}(\cdot|s) + \pi^{(i)}_{i,k}(\cdot|s) \in \mathtt{BR(s)}$ [Benaim 2005; Theorem 4]

Michel Benaim, Josef Hofbauer, and Sylvain Sorin. In: *SIAM Journal on Control and Optimization* (2005)

Cloud Radio Access Network Markov Game

- Agents obtain a minimum cost of -4 and -3 if they reach their goal frequency bands
- Incur a cost of 1 if agents occupy a non-target frequency band



Cloud Radio Access Network Markov Game

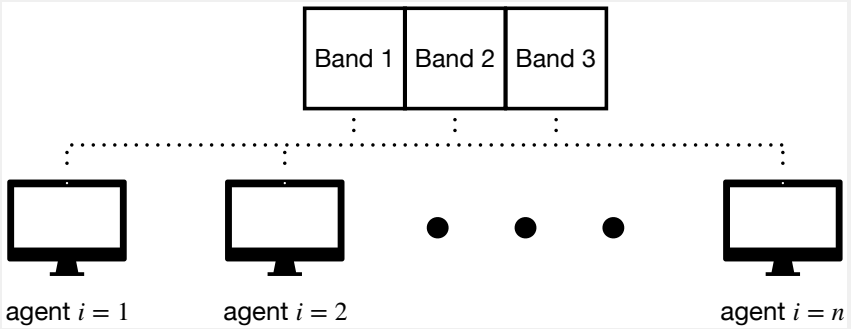
- Base policy $P_{i,0}$ is obtained through a Score Function

$$\phi(s_i', s_i) = \max\{0, \min\{1, A_i(s_i) \cdot s_i' + D_i(s_i) \cdot w_i\}\} + \epsilon$$

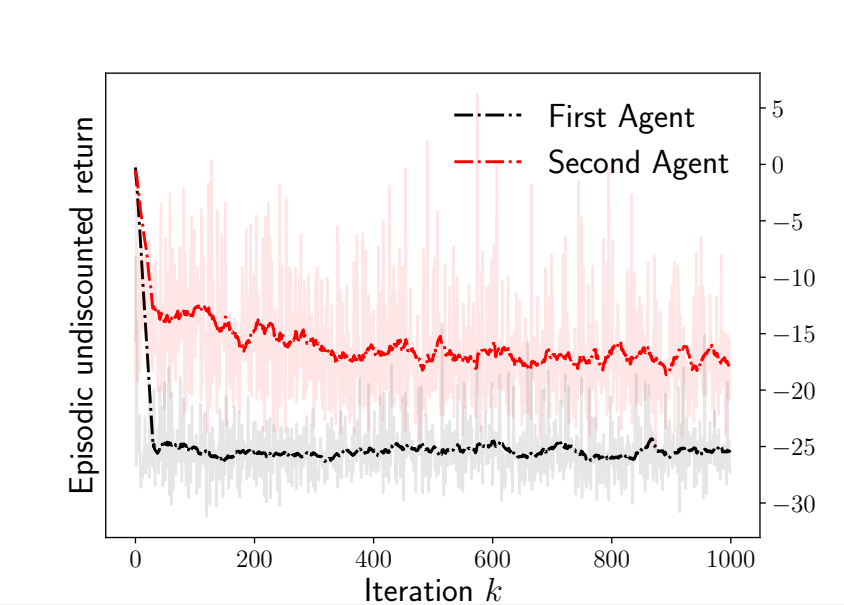
- $A_i(s_i)$: linear function controlling deterministic transitions
- $D_i(s_i)$: noise scaling function for stochastic channel variations
- $w_i \sim \mathcal{N}(0, 1)$: Gaussian noise
- $\epsilon > 0$: small constant for numerical stability

Cloud Radio Access Network Markov Game

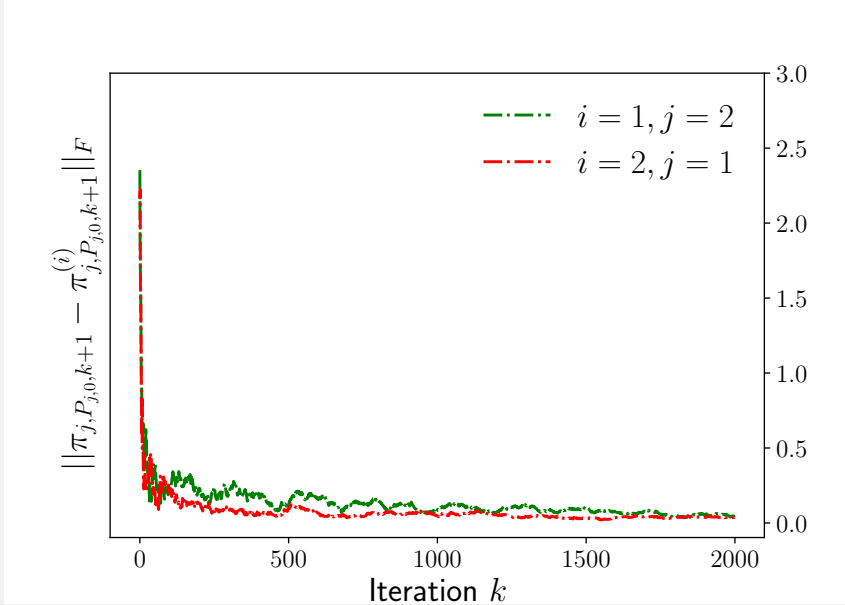
- Discount factor $\gamma = 0.9$
- Beliefs' learning rate $\alpha_k = \frac{\ln(k)}{k}$
- Value function learning rate $\beta_k = \frac{1}{k}$
- Average the results over 10 simulation runs



Results



Iteration k return



Beliefs' convergence

- Return is stable after approximately 180 iterations
- $||\cdot||_F$ value is approximately 0.021 after 2000 iterations

Summary

- General-sum Markov games are largely unexplored
- Proposed new Product Markov games with KL control cost
- Proved the convergence of learning dynamics to a MPE

Additional slides

Additional slides

Additional slides 1

- The convergence relies on showing that the discrete-time value function updates are absolutely summable
- Define the difference value

$$y_{i,k}(s) = [\mathcal{T}^{(\pi^*_{i,P_{i,0},k}, \pi^{(i)}_{-i,P_{-i,0},k})} V_{i,k}](s) - V_{i,k}(s)$$

  between the target update and the current value function estimate
- The goal is to show

$$\liminf_{k \to \infty} \sum_{k}^{k+1} \beta_k y_{i,k}(s) \leq 0$$

Additional slides 2

- To obtain an upper bound on $y_{i,k}(s)$, we introduce a fixed control policy $\hat{\pi}_{i,P_{i,0}}$.
- Define a different mapping $\mathcal{T}^{(\hat{\pi}_{i,P_{i,0}},\pi^{(i)}_{-i,P_{-i,0},k+1})}V_{i,k}(s) = C_i(s) + D_{KL}\left(\hat{\pi}_{i,P_{i,0}}(\cdot|s)||P_{i,0}(\cdot|s)\right) + \gamma \max_{\{s'_i \in \mathcal{S}_i | P_{i,0}(s'_i|s)\neq 0\}} \sum_{s'_{-i}\in\mathcal{S}_{-i}} \pi^{(i)}_{-i,P_{-i,0},k}(s'_{-i}|s) V_{i,k}(s'_i, s'_{-i})$
- Proving the tractable bound $y_{i,k}(s) \leq \hat{y}_{i,k}(s) = \mathcal{T}^{(\hat{\pi}_{i,P_{i,0}},\pi^{(i)}_{-i,P_{-i,0},k+1})}V_{i,k}(s) - V_{i,k}(s) \leq 0$ also proves the original value function updates