Learning to Steer Markovian Agents under Model Uncertainty

Jiawei Huang   Vinzenz Thoma   Zebang Shen

Heinrich H. Nax   Niao He

Motivation

- Agents following some ***typical*** learning dynamics may not always converge to the desired policy

Motivation

- Agents following some ***typical*** learning dynamics may not always converge to the desired policy

For example: the Nash with highest total utility

Motivation

- Agents following some ***typical*** learning dynamics may not always converge to the desired policy

- Two-Player Stag Hunt Game
  - Two actions: H (Hunt) and G (Gather)
  - Pay-off Matrix

	H	G
H	(5, 5)	(0, 4)
G	(4, 0)	(2, 2)

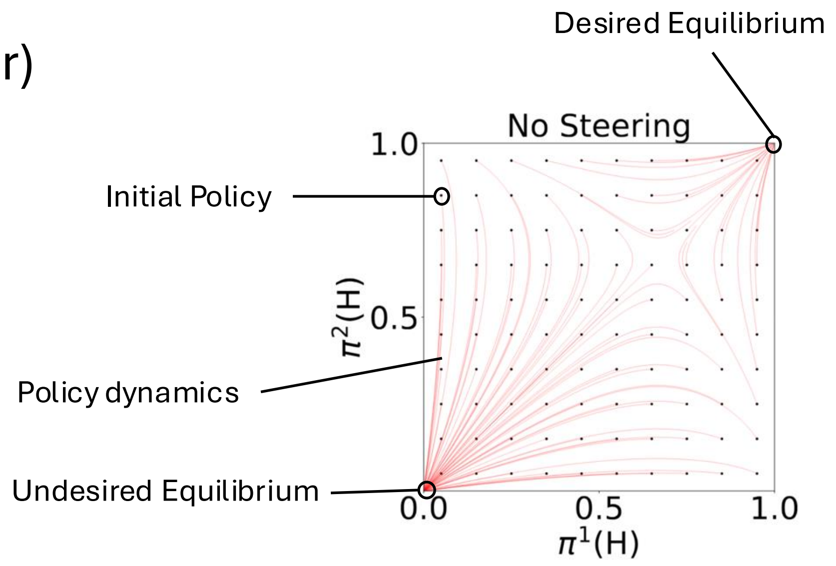
Motivation

- Agents following some ***typical*** learning dynamics may not always converge to the desired policy

- Two-Player Stag Hunt Game
  - Two actions: H (Hunt) and G (Gather)
  - Pay-off Matrix

	H	G
H	(5, 5)	(0, 4)
G	(4, 0)	(2, 2)

  - Replicator Dynamics
    - $\forall t \in [T], i \in \{1,2\}, \ \ \pi_{t+1}^{i}(\cdot) \ \propto \pi_t^{i}(\cdot) \exp(\alpha \, r^{i}(\cdot, \pi_t^{-i}))$



Policy under Replicator Dynamics

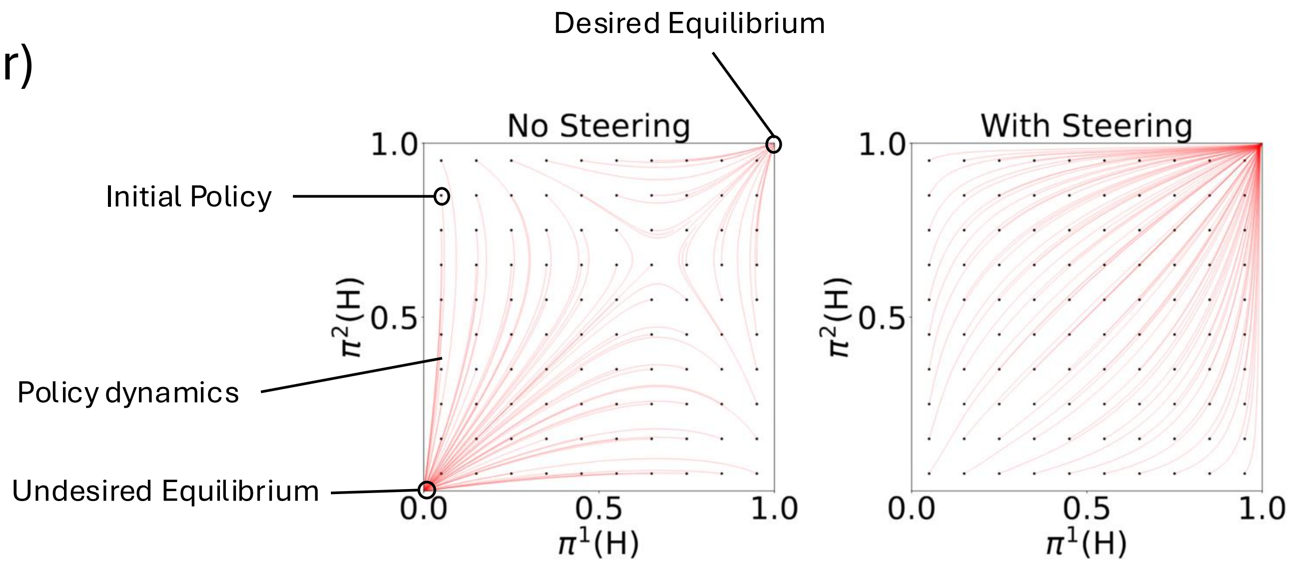
Motivation

A "mediator" may exist, ***steering*** the agents' behaviors by ***providing additional rewards.***
e.g. Financial subsidy by governments to companies.

- Agents following some ***typical*** learning dynamics may not always converge to the desired policy

- Two-Player Stag Hunt Game
  - Two actions: H (Hunt) and G (Gather)
  - Pay-off Matrix

	H	G
H	(5, 5)	(0, 4)
G	(4, 0)	(2, 2)



Policy under Replicator Dynamics

  - Replicator Dynamics
    - $\forall t \in [T], i \in \{1,2\},\ \ \pi_{t+1}^i(\cdot) \propto \pi_t^i(\cdot)\,\exp(\alpha\, r^i(\cdot, \pi_t^{-i}))$

Motivation

A "mediator" may exist, ***steering*** the agents' behaviors by ***providing additional rewards.***
e.g. Financial subsidy by governments to companies.

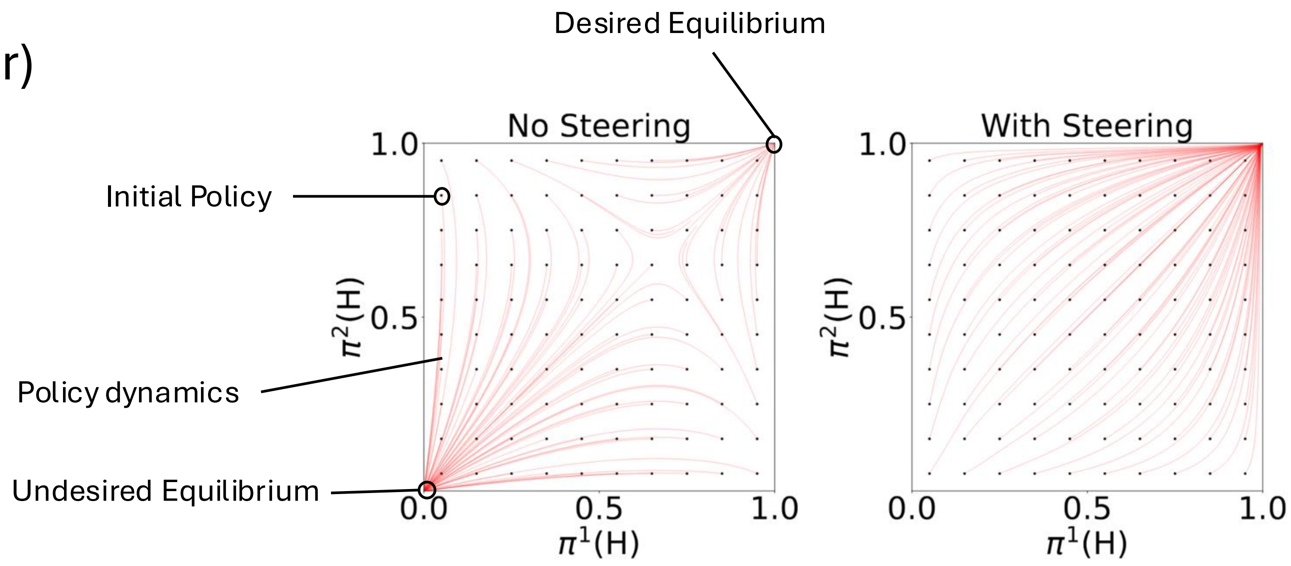
- Agents following some ***typical*** learning dynamics may not always converge to the desired policy

Question: How to design steering rewards?

- Two-Player Stag Hunt Game
  - Two actions: H (Hunt) and G (Gather)
  - Pay-off Matrix

	H	G
H	(5, 5)	(0, 4)
G	(4, 0)	(2, 2)

  - Replicator Dynamics
    - $\forall t \in [T], i \in \{1,2\},\ \ \pi_{t+1}^{i}(\cdot) \propto \pi_{t}^{i}(\cdot)\exp(\alpha\, r^{i}(\cdot,\pi_{t}^{-i}))$



Policy under Replicator Dynamics

Steering Problem Setup

- Finite-Horizon N-Player Markov Games $G := (N, \mathcal{S}, \mathcal{A}, s_1, H, \mathbb{P}, r)$
  - State space $\mathcal{S}$; Action space $\mathcal{A} := \mathcal{A}^1 \times \cdots \mathcal{A}^N$;
  - Transition $\mathbb{P}$; Reward $\boldsymbol{r} := \{r^n\}_{n\in[N]}$,
  - Policy $\boldsymbol{\pi} := (\pi^1, \dots, \pi^N)$
  - Total return $J(\boldsymbol{\pi}|\boldsymbol{r}) := \mathbb{E}_{\boldsymbol{\pi}}[\sum_{n\in[N],h\in[H]} r_h^n(s_h, a_h^1, \dots, a_h^N)]$

Steering Problem Setup

- Markovian learning dynamics

$$\forall t \in [T], \quad \underbrace{\boxed{\boldsymbol{\pi}_{t+1}}}_{\text{Agents' policy}} \sim \underbrace{\boxed{f}}_{\text{Agents' learning dynamics}}(\cdot|\boldsymbol{\pi}_t, \underbrace{\boxed{\boldsymbol{r}}}_{\text{(original) reward}}).$$

Steering Problem Setup

- Markovian learning dynamics

Agents' policy (original) reward

$$\forall t \in [T], \quad \boxed{\boldsymbol{\pi}_{t+1}} \sim \boxed{f}(\cdot|\boldsymbol{\pi}_t, \boxed{\boldsymbol{r}}).$$

Agents' learning dynamics

  - Subsume a broad class of policy-based methods
    - Replicator dynamics, gradient descent, etc.
  - Complementary to no-regret dynamics studied before (Zhang et. al., 2023)
  - Considered in a concurrent work (Canyakmaz et al., 2024)

Steering Problem Setup

- Steering Markovian Agents for T steps

$$\forall t \in [T], \quad \underbrace{\boxed{\boldsymbol{u}_t}}_{\text{steering reward}} \sim \underbrace{\boxed{\psi_t}}_{\text{steering strategy}}(\cdot|\boldsymbol{\pi}_1, \boldsymbol{u}_1, ..., \boldsymbol{\pi}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{\pi}_t), \quad \boldsymbol{\pi}_{t+1} \sim f(\cdot|\boldsymbol{\pi}_t, \boldsymbol{r} + \boldsymbol{u}_t),$$

Steering Problem Setup

- Steering Markovian Agents for T steps

$$\forall t \in [T], \quad \boxed{\boldsymbol{u}_t} \sim \boxed{\psi_t}(\cdot|\boldsymbol{\pi}_1, \boldsymbol{u}_1, ..., \boldsymbol{\pi}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{\pi}_t), \quad \boldsymbol{\pi}_{t+1} \sim f(\cdot|\boldsymbol{\pi}_t, \boldsymbol{r} + \boldsymbol{u}_t),$$

steering reward   steering strategy

- Our goal
  - [**Primary**] Agents' Behavior
    - $\eta^{\text{goal}}(\boldsymbol{\pi}_{T+1}) \approx \max_{\pi} \eta^{\text{goal}}(\boldsymbol{\pi})$, for some measure η^{goal}

Steering Problem Setup

- Steering Markovian Agents for T steps

$$\forall t \in [T], \quad \boxed{\boldsymbol{u}_t} \sim \boxed{\psi_t}(\cdot|\boldsymbol{\pi}_1, \boldsymbol{u}_1, ..., \boldsymbol{\pi}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{\pi}_t), \quad \boldsymbol{\pi}_{t+1} \sim f(\cdot|\boldsymbol{\pi}_t, \boldsymbol{r} + \boldsymbol{u}_t),$$

steering reward

steering strategy

- Our goal
  - [**Primary**] Agents' Behavior
    - $\eta^{\text{goal}}(\boldsymbol{\pi}_{T+1}) \approx \max_{\pi} \eta^{\text{goal}}(\boldsymbol{\pi})$, for some measure $\boxed{\eta^{\text{goal}}}$

Example 1: $-||\boldsymbol{\pi} - \boldsymbol{\pi}^*||$

Example 2: $J(\boldsymbol{\pi}|r)$

Steering Problem Setup

- Steering Markovian Agents for T steps

$$\forall t \in [T], \quad \underbrace{\boxed{\boldsymbol{u}_t}}_{\text{steering reward}} \sim \underbrace{\boxed{\psi_t}}_{\text{steering strategy}}(\cdot|\boldsymbol{\pi}_1, \boldsymbol{u}_1, \ldots, \boldsymbol{\pi}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{\pi}_t), \quad \boldsymbol{\pi}_{t+1} \sim f(\cdot|\boldsymbol{\pi}_t, \boldsymbol{r} + \boldsymbol{u}_t),$$

- Our goal
  - **[Primary]** Agents' Behavior
    - $\eta^{\text{goal}}(\boldsymbol{\pi}_{T+1}) \approx \max_{\pi} \eta^{\text{goal}}(\boldsymbol{\pi})$, for some measure $\boxed{\eta^{\text{goal}}}$

Example 1: $-||\boldsymbol{\pi} - \boldsymbol{\pi}^*||$

Example 2: $J(\boldsymbol{\pi}|r)$

  - **[Secondary]** The steering cost
    - $\eta^{\text{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t) \coloneqq J(\boldsymbol{\pi}_t|\boldsymbol{u}_t)$
    - We expect $\sum_{t\in[T]} \eta^{\text{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t)$ is "reasonable"

Steering Problem Setup

- Steering Markovian Agents for T steps

$$\forall t \in [T],\ \underbrace{\boxed{\color{red}{\boldsymbol{u}_t}}}_{\text{steering reward}} \sim \underbrace{\boxed{\psi_t}}_{\text{steering strategy}}(\cdot|\boldsymbol{\pi}_1, \boldsymbol{u}_1, ..., \boldsymbol{\pi}_{t-1}, \boldsymbol{u}_{t-1}, \boldsymbol{\pi}_t), \quad \boldsymbol{\pi}_{t+1} \sim f(\cdot|\boldsymbol{\pi}_t, \boldsymbol{r} + \color{red}{\boldsymbol{u}_t}),$$

- Our goal
  - **[Primary]** Agents' Behavior
    - $\eta^{\text{goal}}(\boldsymbol{\pi}_{T+1}) \approx \max_{\pi} \eta^{\text{goal}}(\boldsymbol{\pi})$, for some measure $\boxed{\eta^{\text{goal}}}$

Example 1: $-||\boldsymbol{\pi} - \boldsymbol{\pi}^*||$

Example 2: $J(\boldsymbol{\pi}|r)$

  - **[Secondary]** The steering cost
    - $\eta^{\text{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t) \coloneqq J(\boldsymbol{\pi}_t|\boldsymbol{u}_t)$
    - We expect $\sum_{t\in[T]} \eta^{\text{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t)$ is "reasonable"
- Steering dynamics as an MDP
  - State $\boldsymbol{\pi}_t$; Action $\boldsymbol{u}_t$
  - Transition f; Reward function η^{goal} and η^{cost}
  - We can use Reinforcement Learning (RL) to learn ψ_t

Practical Considerations

- **Model uncertainty**: true dynamics f^* is unknown

Practical Considerations

- **Model uncertainty**: true dynamics f^* is unknown
  - A model-based learning setup
  - A model class $\mathcal{F}$, $|\mathcal{F}| < +\infty$ available
  - [**Realizability Assumption**] $f^* \in \mathcal{F}$

Practical Considerations

- **Model uncertainty**: true dynamics f^* is unknown
  - A model-based learning setup
  - A model class $\mathcal{F}$, $|\mathcal{F}| < +\infty$ available
  - [**Realizability Assumption**] $f^* \in \mathcal{F}$

- **Non-episodic setup** ("You can only steer once")
  - Can not reset agents to "initial policy" again.
  - Learn a ***history-dependent*** steering strategy ψ

Practical Considerations

- **Model uncertainty**: true dynamics f^* is unknown
  - A model-based learning setup
  - A model class $\mathcal{F}$, $|\mathcal{F}| < +\infty$ available
  - [**Realizability Assumption**] $f^* \in \mathcal{F}$

- **Non-episodic setup** ("You can only steer once")
  - Can not reset agents to "initial policy" again.
  - Learn a ***history-dependent*** steering strategy ψ

Key Question: How can we learn a good history-dependent steering strategy under model uncertainty?

Learning Objective

- Denote Ψ as the collection of all history-dependent strategies

$$\psi^* \leftarrow \mathrm{argmax}_{\psi \in \Psi} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{E}_{\psi, f}[\beta \cdot \eta^{\mathrm{goal}}(\boldsymbol{\pi}_{T+1}) - \sum_{t \in [T]} \eta^{\mathrm{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t)]$$

- Proposition 3.3 (Under some assumptions)
  1. $\boldsymbol{\pi}_{T+1}$ under ψ^* approximately maximizes η^{goal}
  2. ψ^* is "pareto-optimal" for η^{goal} and η^{cost}.

Learning Objective

- Denote Ψ as the collection of all history-dependent strategies

$$\psi^* \leftarrow \mathrm{argmax}_{\psi \in \Psi} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{E}_{\psi, f}[\beta \cdot \eta^{\mathrm{goal}}(\boldsymbol{\pi}_{T+1}) - \sum_{t \in [T]} \eta^{\mathrm{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t)]$$

- Proposition 3.3 (Under some assumptions)
  1. $\boldsymbol{\pi}_{T+1}$ under ψ^* approximately maximizes η^{goal}
  2. ψ^* is "pareto-optimal" for η^{goal} and η^{cost}.

Example (Section 4)
$\mathcal{F}$ is a class of "distinguishable" policy mirror descent dynamics.

Learning Objective

- Denote Ψ as the collection of all history-dependent strategies

$$\psi^* \leftarrow \text{argmax}_{\psi \in \Psi} \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{E}_{\psi, f}[\beta \cdot \eta^{\text{goal}}(\boldsymbol{\pi}_{T+1}) - \sum_{t \in [T]} \eta^{\text{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t)]$$

- Proposition 3.3 (Under some assumptions)
  1. $\boldsymbol{\pi}_{T+1}$ under ψ^* approximately maximizes η^{goal}
  2. ψ^* is "pareto-optimal" for η^{goal} and η^{cost}.

Example (Section 4)
$\mathcal{F}$ is a class of "distinguishable" policy mirror descent dynamics.

- **Main Challenge**: Learning history-dependent policy

Solving Main Objective

- Scenario 1: $|\mathcal{F}|$ is small

$$\begin{aligned}\psi^* &\leftarrow \mathrm{argmax}_{\psi\in\Psi}\frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}\mathbb{E}_{\psi,f}[\beta\cdot\eta^{\mathrm{goal}}(\boldsymbol{\pi}_{T+1})-\sum_{t\in[T]}\eta^{\mathrm{cost}}(\boldsymbol{\pi}_t,\boldsymbol{u}_t)]\\ &= \mathrm{argmax}_{\psi\in\Psi}\mathbb{E}_{\psi,f\sim\mathrm{Uniform}(\mathcal{F})}[\beta\cdot\eta^{\mathrm{goal}}(\boldsymbol{\pi}_{T+1})-\sum_{t\in[T]}\eta^{\mathrm{cost}}(\boldsymbol{\pi}_t,\boldsymbol{u}_t)]\end{aligned}$$

  - A POMDP perspective
    - Hidden state is $x_t=(f,\boldsymbol{\pi}_t)$, but only $o_t=\pi_t$ is revealed.

Solving Main Objective

- Scenario 1: $|\mathcal{F}|$ is small

$$\psi^* \leftarrow \mathrm{argmax}_{\psi\in\Psi}\frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}\mathbb{E}_{\psi,f}\left[\beta\cdot\eta^{\mathrm{goal}}(\boldsymbol{\pi}_{T+1})-\sum_{t\in[T]}\eta^{\mathrm{cost}}(\boldsymbol{\pi}_t,\boldsymbol{u}_t)\right]$$
$$=\mathrm{argmax}_{\psi\in\Psi}\mathbb{E}_{\psi,f\sim\mathrm{Uniform}(\mathcal{F})}\left[\beta\cdot\eta^{\mathrm{goal}}(\boldsymbol{\pi}_{T+1})-\sum_{t\in[T]}\eta^{\mathrm{cost}}(\boldsymbol{\pi}_t,\boldsymbol{u}_t)\right]$$

  - A POMDP perspective
    - Hidden state is $x_t=(f,\boldsymbol{\pi}_t)$, but only $o_t=\pi_t$ is revealed.
  - Learn a belief-state based ψ instead
    - Belief states is posterior distribution of f
    - Easy to compute when $|\mathcal{F}|$ is small

Solving Main Objective

- Scenario 2: $|\mathcal{F}|$ is large

$$\boxed{\psi^* \leftarrow \operatorname{argmax}_{\psi\in\Psi} \frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} \mathbb{E}_{\psi,f}[\beta \cdot \eta^{\text{goal}}(\boldsymbol{\pi}_{T+1}) - \sum_{t\in[T]} \eta^{\text{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t)]}$$

  - Exact solution is intractable in general;

Solving Main Objective

$$\psi^* \leftarrow \mathrm{argmax}_{\psi\in\Psi} \frac{1}{|\mathcal{F}|} \sum_{f\in\mathcal{F}} \mathbb{E}_{\psi,f}\left[\beta \cdot \eta^{\mathrm{goal}}(\boldsymbol{\pi}_{T+1}) - \sum_{t\in[T]} \eta^{\mathrm{cost}}(\boldsymbol{\pi}_t, \boldsymbol{u}_t)\right]$$

- Scenario 2: $|\mathcal{F}|$ is large
  - Exact solution is intractable in general;
  - Trade-off tractability and optimality
  - A First-Explore-Then-Exploit Framework
    - Explore and estimate $\hat{f}^*$ in the first T_0 steps
    - Deploy optimal strategy in $\hat{f}^*$ for the rest $T - T_0$ steps
    - Only learn history-dependent strategy for T_0 steps

Experiments

- Empirical verification of proposed methods for two scenarios

- See Section 6 for more details

Summary

- We study steering Markovian agents under model uncertainty

Summary

- We study steering Markovian agents under model uncertainty

- Take Aways
  - Formulation for steering Markovian agents
  - A learning objective with guarantees
  - Algorithms overcoming challenges in learning history-dependent strategies (with empirical evaluation)

Summary

- We study steering Markovian agents under model uncertainty

- Take Aways
  - Formulation for steering Markovian agents
  - A learning objective with guarantees
  - Algorithms overcoming challenges in learning history-dependent strategies (with empirical evaluation)

- Future works
  - Better objective function?
  - Non-Markovian agents?

Thank You!

Paper Link