Universität Zürich[UZH]

ALPI

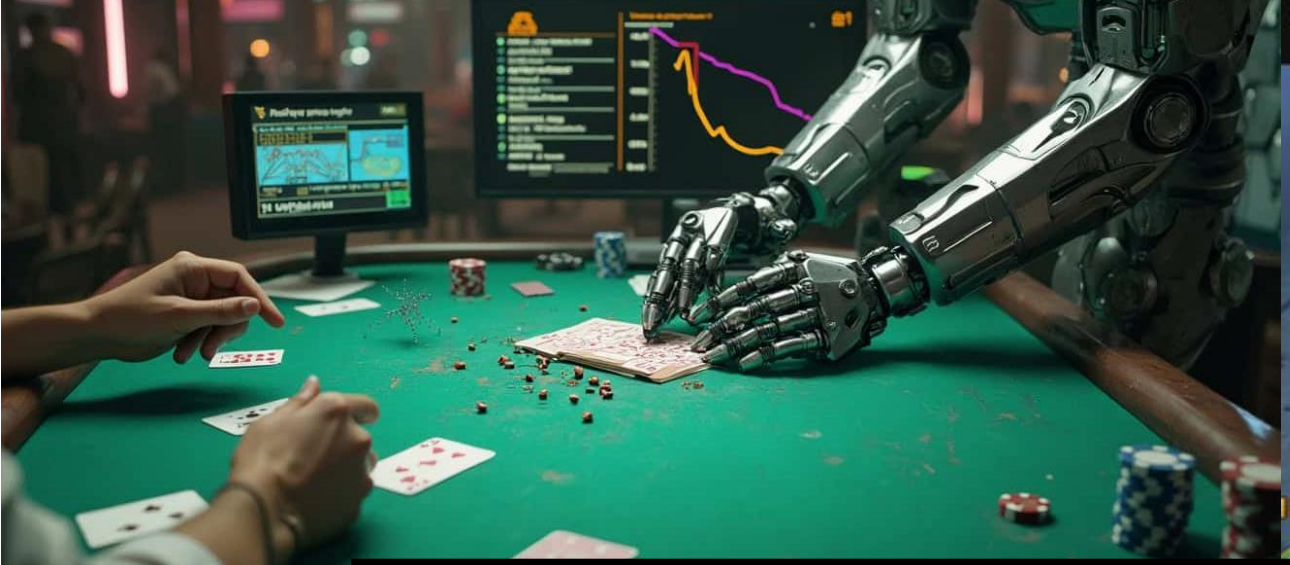
Multi-agent Imitation Learning

Prof. Dr. Giorgia Ramponi

ETH AI CENTER

Swiss National Science Foundation

UZH .ai



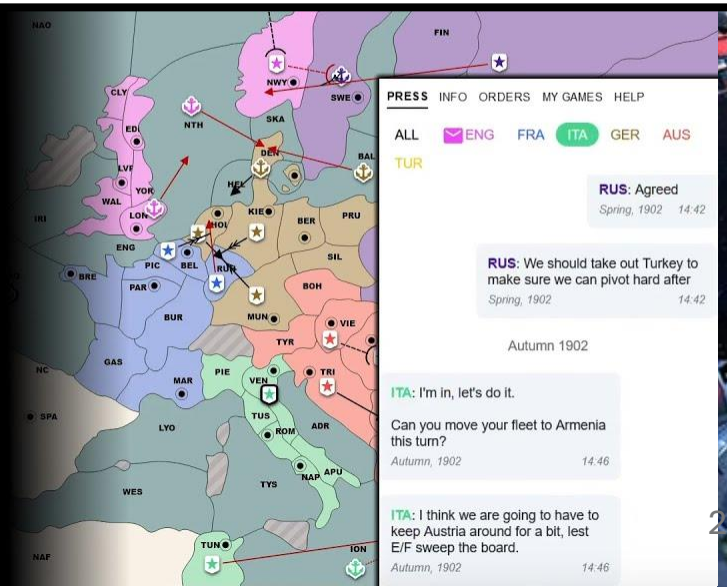
Multi-Agent Reinforcement Learning has achieved super-human play

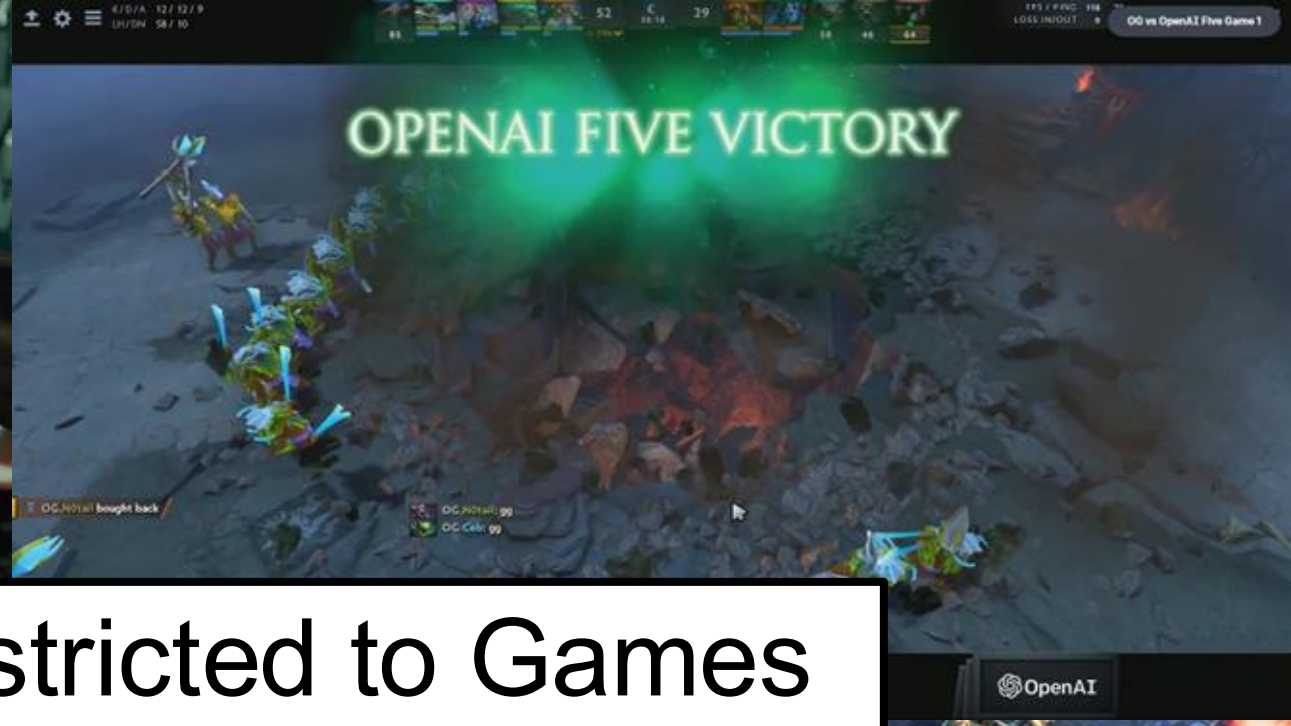
Expert H

VS

CICERO AI (x6)

Diplomacy Commentary





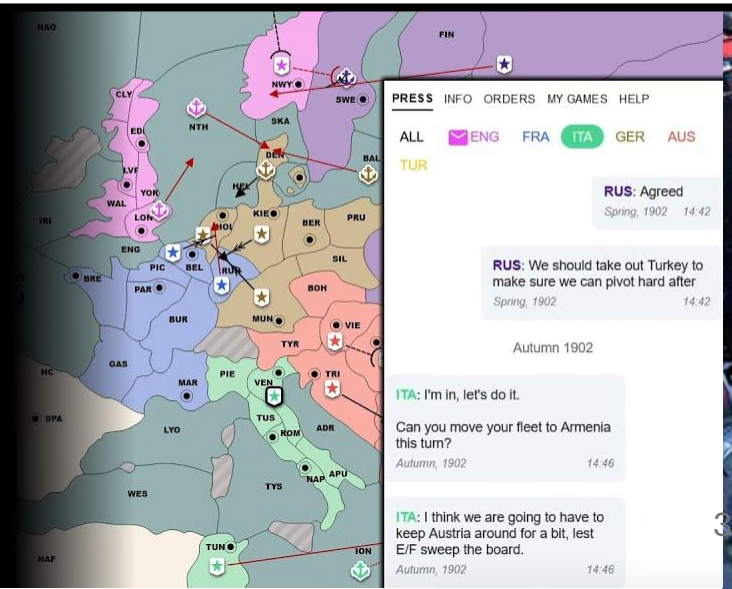
Successes are restricted to Games with well-specified reward functions

Expert H

VS

CICERO AI (x6)

Diplomacy Commentary

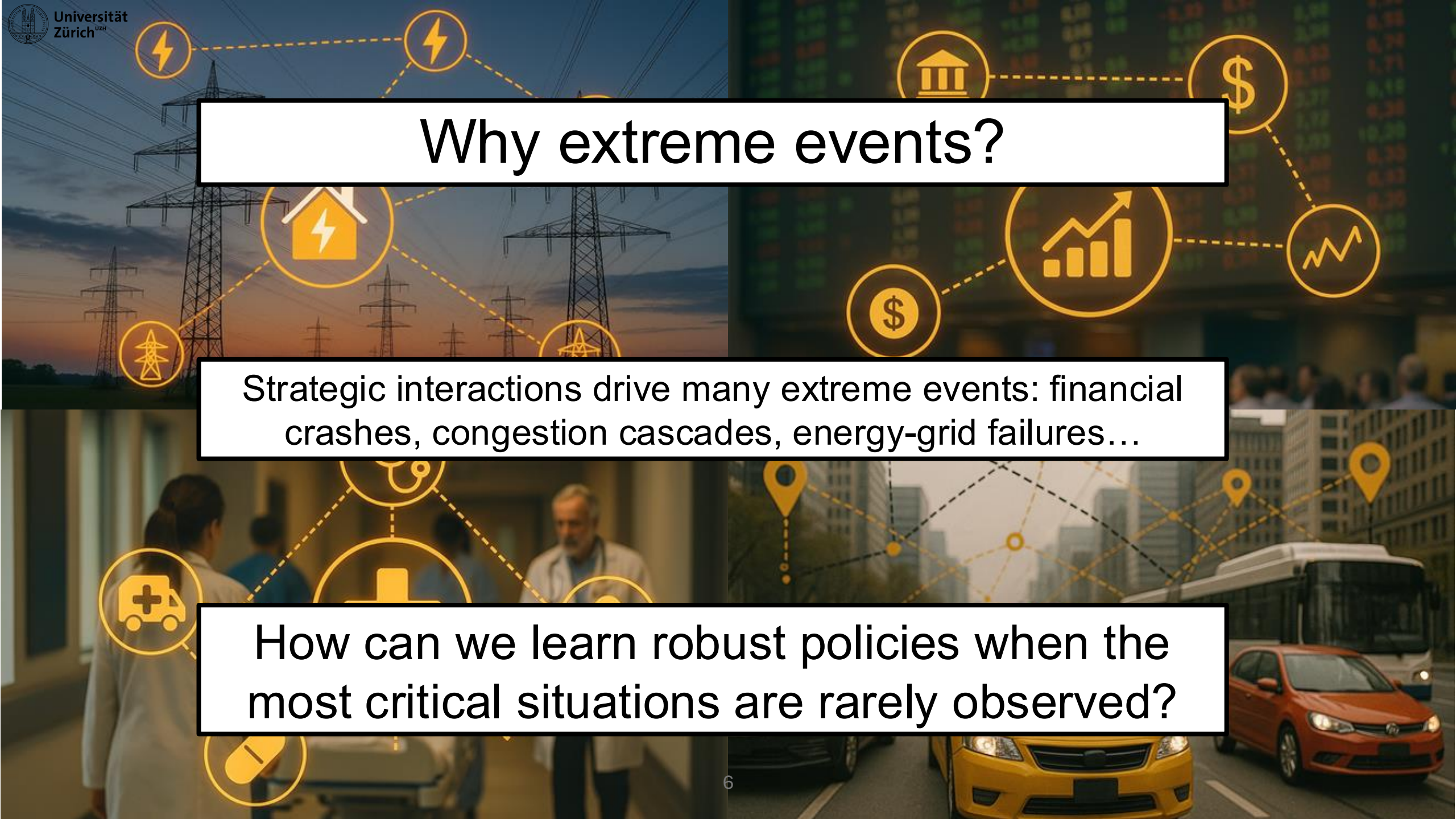


Can we scale to real-world scenarios with unknown reward functions?

Idea: Leverage expert demonstrations

Why extreme events?

Strategic interactions drive many extreme events: financial crashes, congestion cascades, energy-grid failures…

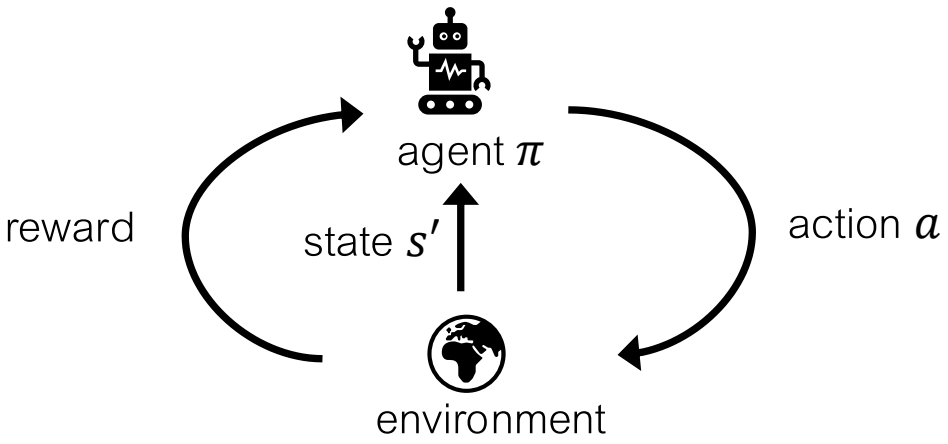


Why extreme events?

Strategic interactions drive many extreme events: financial crashes, congestion cascades, energy-grid failures…

How can we learn robust policies when the most critical situations are rarely observed?

Reinforcement Learning



Objective: find the optimal policy

$$\max_{\pi} V(\pi, r) = \max_{\pi} \mathbb{E}\left[\sum_{h=0}^{H} r(s_h, a_h) | s_0 \sim \mu, \pi, P\right]$$

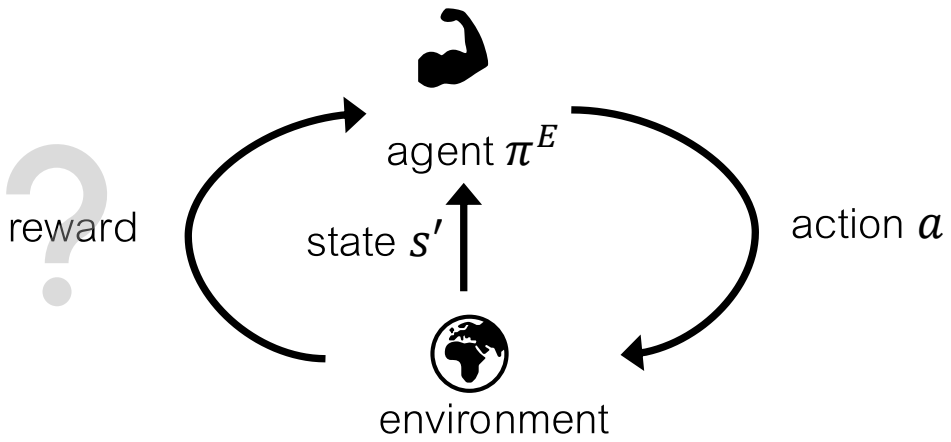
$$\max_{\pi} V(\pi, r) = \max_{\pi} \sum_{s,a} \mu^{\pi}(s, a) r(s, a)$$

$$\mu^{\pi}(s, a) = \mathbb{E}\left[\sum_{h=0}^{H} P(s_h = s, a_h = a) | s_0 \sim \mu, \pi, P\right]$$

- In many domains design the reward function is hard
- Hard to express numerically
- Inherently multi-objective problems

Learning from experts

Learning from “expert” demonstrations [1,2] (human demonstrations, powerful teacher models, etc.)



We have access to **expert’s demonstrations:**

$$D = \{\tau_1, \dots, \tau_m\},$$
$$\tau_i = \{s_1, a_1, s_2, a_2, \dots, s_n, a_n\}$$

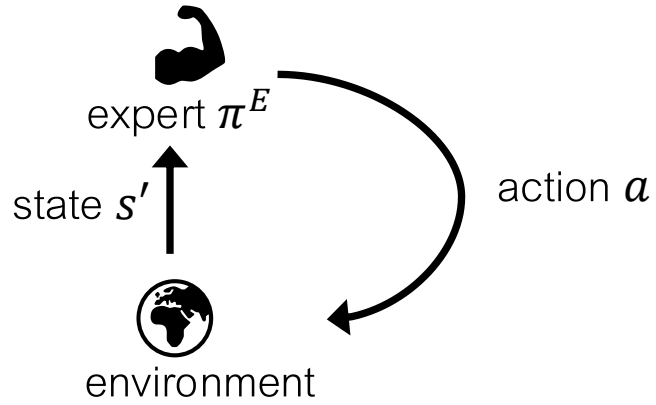
OBJECTIVE (VALUE GAP)

$$\text{ValueGap}(\pi) = \text{SubOpt}(\pi) = \min_{\pi} V(\pi, r) - V(\pi^E, r)$$

⇒ same optimal policy as the underlying RL problem

Imitation Learning

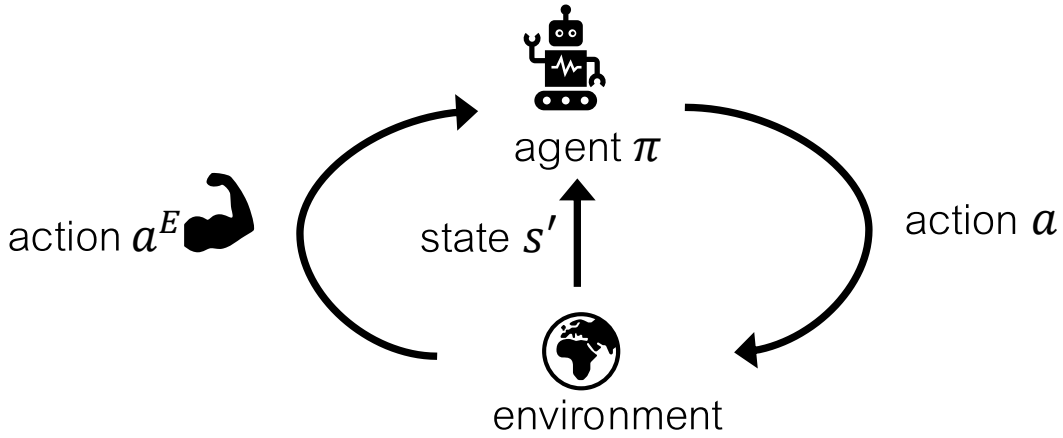
NON-INTERACTIVE



Observing $\mathcal{D} = \{\tau_k\}_{k=1}^N \sim \pi^E$

$\pi^E \in \text{argmax}_\pi V(\pi, r)$ of an unknown r

INTERACTIVE



The IL learning agent interact with the environment

Asking the expert advice on the visited states

Imitation Learning

OBJECTIVE (VALUE GAP)

$$\text{ValueGap}(\pi) = \text{SubOpt}(\pi) = \min_{\pi} V(\pi, r) - V(\pi^E, r)$$

We do not know r!

(informal) non-interactive imitation learning (BC) is optimal in a worst case sense [2]

PROXY OBJEC

Behavior Cloning

Regression task by estimating action distributions.

$$\min_{\pi} \sum_{i=1} loss(\pi(s_i), a_i)$$

Adversarial Imitation Learning [3,4] (GAIL)

Objective is to match the expert state-action occupancy measure

$$\min_{\pi} \sum_{i=1}^{nT} loss\left(\mu^{\pi}(s_i, a_i), \mu^{\pi^E}(s_i, a_i)\right)$$

[1] Pomerleau (1991). *Efficient Training of Artificial Neural Networks for Autonomous Navigation*
[2] Foster et al. (2024). *Is Behavior Cloning All You Need? Understanding Horizon in Imitation Learning*
[3] Ho and Ermon (2016). *Generative Adversarial Imitation Learning*
[4] Garg, D., Chakraborty, S., Cundy, C., Song, J., & Ermon, S. (2021). Iq-learn: Inverse soft-q learning for imitation. Advances in Neural Information Processing Systems, 34, 4028-4039.

Imitation Learning

Key insight : the imitation dataset induces the right state distribution for the Value gap

$$V(\pi^E) - V(\pi) = H \sum_h \mathbb{E}_{s,a\sim\pi^E}\left[Q^{\pi^E}(s,a) - Q^{\pi}(s,\pi(\cdot\,|s))\right]$$

$$\leq H^2 \sum_h \mathbb{E}_{s,a\sim\pi_h^E}\left[||\pi_h^E(\cdot\,|s) - \pi_h(\cdot\,|s)||_1\right]$$

This can be optimized by a single regression over the dataset

More data = Better results

Imitation Learning

Key insight : the imitation dataset induces the right state distribution for the Value gap

Can we conclude the same for Multi-Agent Imitation Learning?

More data = Better results

Multi-agent Imitation Learning

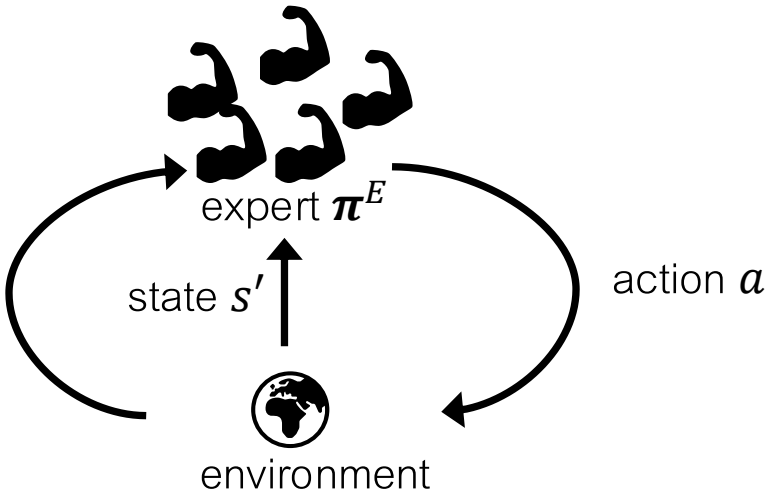
On imitation in mean-field games, NeurIPS 2023
G Ramponi, P Kolev, O Pietquin, N He, M Laurière, M Geist

Learning Equilibria from Data: Provably Efficient Multi-Agent Imitation Learning, NeurIPS 2025
T Freihaut, L Viano, V Cevher, M Geist, G Ramponi

Rate optimal learning of equilibria from data, AISTATS 2026
T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026
L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Multi-Agent Imitation Learning



π^E is a joint policy maximizing what?

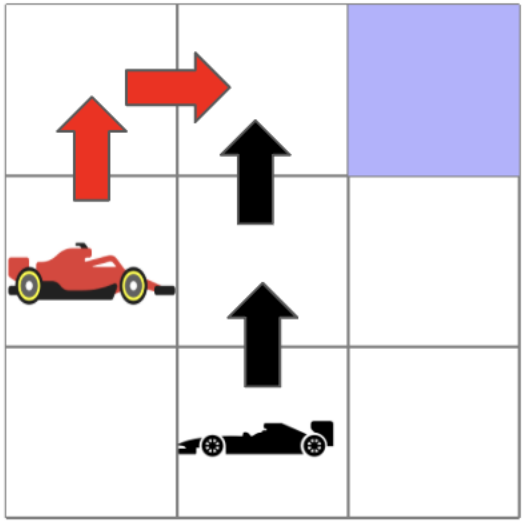
Maximizing Nash Equilibria, i.e.

$$V(\pi^E, r_i) - \max_{\pi_i} V\left((\pi_i, \pi^E_{-i}\right), r_i) \geq 0 \quad \forall i$$

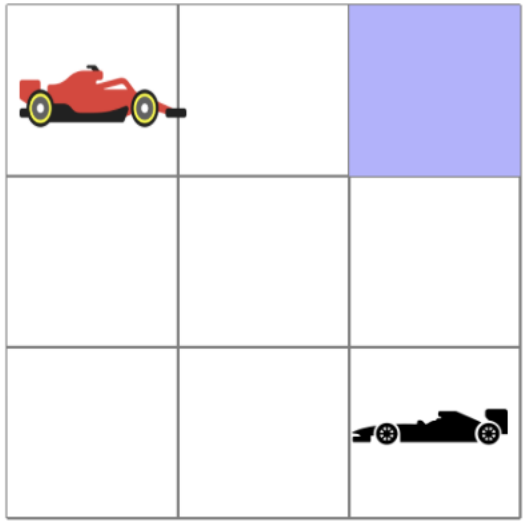
OBJECTIVE (NASH IMITATION GAP)

$$\text{NashGap}(\boldsymbol{\pi}) = \sum_i \max_{\pi'} V_i(\pi', \boldsymbol{\pi}_{-i}) - V_i(\boldsymbol{\pi})$$

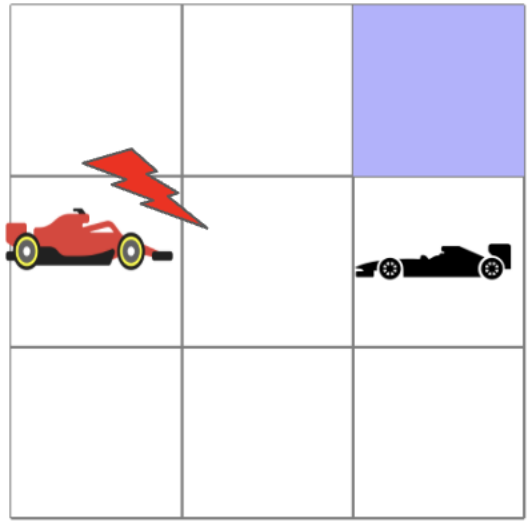
Is Behavior Cloning all you need in MAIL?



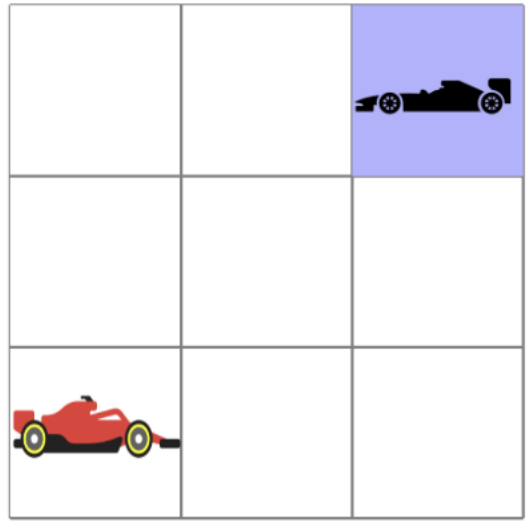
2 out of 72 states visited



Best responding car moves to unseen state



Car takes suboptimal action



Black car wins

→ Best responding agent visits a state not within the expert dataset
→ BC policy does not know how to act
→ The policy of the individual agent becomes exploitable

Analysis idea

$$\text{NashGap}_1(\hat{\pi}_1, \hat{\pi}_2) = \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\hat{\pi}_1, \hat{\pi}_2)$$

Analysis idea

$$\text{NashGap}_1(\hat{\pi}_1, \hat{\pi}_2) = \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\hat{\pi}_1, \hat{\pi}_2)$$

$$= \underbrace{\max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1^E, \pi_2^E)}_{\text{Exploit-gap}} + \underbrace{V_1(\pi_1^E, \pi_2^E) - V_1(\hat{\pi}_1, \hat{\pi}_2)}_{\text{Value-gap}}$$

easy to analyze with
single-agent techniques

Analysis idea

$$\text{Exploit} - \text{gap}_1(\hat{\pi}_1, \hat{\pi}_2) = \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1^E, \pi_2^E)$$

$$\leq \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1, \pi_2^E)$$

Analysis idea

$$\text{Exploit} - \text{gap}_1(\hat{\pi}_1, \hat{\pi}_2) = \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1^E, \pi_2^E)$$

$$\leq \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1, \pi_2^E)$$

$$\leq H \max_{\pi_1} \sum_h \mathbb{E}_{(\pi_1, \pi_2^E)}[||\hat{\pi}_{h,2}(\cdot|s) - \pi_2^E(\cdot|s)||_1]$$

Analysis idea

$$\text{Exploit} - \text{gap}_1(\hat{\pi}_1, \hat{\pi}_2) = \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1^E, \pi_2^E)$$

$$\leq \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1, \pi_2^E)$$

$$\leq H \max_{\pi_1} \sum_h \mathbb{E}_{(\pi_1, \pi_2^E)}[||\hat{\pi}_{h,2}(\cdot|s) - \pi_2^E(\cdot|s)||_1]$$

$$\leq H \max_{\pi_1} \left\| \frac{\mu^{\pi_1, \pi_2^E}}{\mu^{\pi_1^E, \pi_2^E}} \right\| \sum_h \mathbb{E}_{(\pi_1^E, \pi_2^E)}[||\hat{\pi}_{h,2}(\cdot|s) - \pi_2^E(\cdot|s)||_1]$$

Analysis idea

New quantity we need to control:
deviation of the best response of the recovered policy and the equilibrium one

$$C_{\max} = \max_{\pi_1,\pi_2} \left\{ \max_{br(\pi_2)} \left\| \frac{\mu^{br(\pi_2),\pi_2^E}}{\mu^{\pi_1^E,\pi_2^E}} \right\|, \max_{br(\pi_1)} \left\| \frac{\mu^{\pi_1^E,br(\pi_1)}}{\mu^{\pi_1^E,\pi_2^E}} \right\| \right\}$$

But is this the best we can do?

Analysis idea

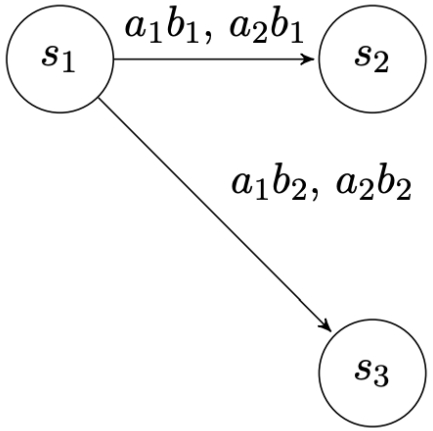
New quantity we need to control:
deviation of the best response of the recovered policy and the equilibrium one

$$C_{\max} = \max_{\pi_1,\pi_2} \left\{ \max_{br(\pi_2)} \left\| \frac{\mu^{br(\pi_2),\pi_2^E}}{\mu^{\pi_1^E,\pi_2^E}} \right\|, \max_{br(\pi_1)} \left\| \frac{\mu^{\pi_1^E,br(\pi_1)}}{\mu^{\pi_1^E,\pi_2^E}} \right\| \right\}$$

But is this the best we can do?

Not only an upper bound…

(Informal) For any non-interactive algorithm, to achieve a ϵ −NashGap there exists a Markov Game that requires an expert dataset of size $N = \Omega(\frac{C_{\max}}{\epsilon^2})$.



1. The Nash Imitation Gap cannot be reduced with non-interactive IL techniques!

2. Nash equilibrium only guarantees robustness against unilateral deviations

[1] Rate optimal learning of equilibria from data, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

Other hardness results: J. Tang, G. Swamy, F. Fang, and Z. S. Wu. Multi-agent imitation learning: Value is easy, regret is hard. In NIPS '24

More data = better results?

C_{max} is required in any non-interactive MAIL algorithm [1]

If $C_{max} = \infty \rightarrow$ any non-interactive IL algorithm suffers from a Nash gap that is linear in the order of the horizon (or discount factor)!

Observed equilibria are not enough

- Strategic deviations
  - Rare failures
- Extreme outcomes

$\rightarrow$ ever

$\rightarrow$ if the learner knows the transition dynamics [1]

[1] Rate optimal learning of equilibria from data, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

Other hardness results: J. Tang, G. Swamy, F. Fang, and Z. S. Wu. Multi-agent imitation learning: Value is easy, regret is hard. In NIPS '24

More data = better results?

C_{max} is required in any non-interactive MAIL algorithm [1]

If $C_{max} = \infty \rightarrow$ any non-interactive IL algorithm suffers from a Nash gap that is linear in the order of the horizon (or discount factor)!

We need interactive imitation learning!

$\rightarrow$ if the learner knows the transition dynamics [1]

[1] Rate optimal learning of equilibria from data, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

Other hardness results: J. Tang, G. Swamy, F. Fang, and Z. S. Wu. Multi-agent imitation learning: Value is easy, regret is hard. In NIPS '24

Interactive Multi-Agent Imitation Learning

- Observing the best-response [1]
- Having access only to the equilibrium policy [2]
- Using the structure [3]

[1] Learning Equilibria from Data: Provably Efficient Multi-Agent Imitation Learning, NeurIPS 2025T Freihaut, L Viano, V Cevher, M Geist, G Ramponi

[2] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

[3] Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Interactive Multi-Agent Imitation Learning

- Observing the best-response [1]
- Having access only to the equilibrium policy [2]
- Using the structure [3]

[1] Learning Equilibria from Data: Provably Efficient Multi-Agent Imitation Learning, NeurIPS 2025, T Freihaut, L Viano, V Cevher, M Geist, G Ramponi

[2] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

[3] Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026, L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Multi-Agent Imitation Learning with Best Response Oracle (MAIL-BRO)

Best response oracle

For any $(\hat{\mu}, \hat{\nu})$ we can query the best response policies $\mathrm{br}(\hat{\mu})$ and $\mathrm{br}(\hat{\nu})$

$$\text{Exploit} - \text{gap}_1(\hat{\pi}_1, \hat{\pi}_2) = \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1^E, \pi_2^E)$$

$$\leq H \max_{\pi_1} \sum_h \mathbb{E}_{(\pi_1, \pi_2^E)}[||\hat{\pi}_{h,2}(\cdot|s) - \pi_2^E(\cdot|s)||_1]$$

Multi-Agent Imitation Learning with Best Response Oracle (MAIL-BRO)

$$\min_{\hat{\pi}_2} \; H \max_{\pi_1} \sum_h \mathbb{E}_{(\pi_1, \pi_2^E)}[||\hat{\pi}_{h,2}(\cdot\,|s) - \pi_2^E(\cdot\,|s)\,||_1]$$

1. This optimization problem is non-convex
2. We need to collect data from the occupancy measure of π_2^E, π_1 which depends on the minimizer itself

To solve this we showed that it is enough to minimize:

$$\mathbb{E}_{s\sim \mathrm{br}(\pi_{k,2}),\pi_{k,2}}\left[\left|\left|\hat{\pi}_{h,2}(\cdot\,|s) - \pi_2^E(\cdot\,|s)\right|\right|^2\right]$$

We can apply a no-regret algorithm to these loss sequences

Sample-complexity

$$\mathcal{O}\left(\frac{H^4|S||\mathrm{A}_{\max}|^2}{\epsilon^4}\right)$$

- Compared to single-agent the sample complexity is ϵ^4 instead of ϵ^2
- We can avoid the dependence on the single policy deviation concentrability coefficient

How to remove the best-response oracle assumption?

Interactive Multi-Agent Imitation Learning

- Observing the best-response [1]
- Having access only to the equilibrium policy [2]
- Using the structure [3]

P.s.: I will use the finite-horizon setting for the rest of the presentation but the results can be extended to infinite-horizon

[1] Learning Equilibria from Data: Provably Efficient Multi-Agent Imitation Learning, NeurIPS 2025, T Freihaut, L Viano, V Cevher, M Geist, G Ramponi

[2] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

[3] Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

We need to control the $C_{\max}$

$$\text{Exploit} - \text{gap}_1(\hat{\pi}_1, \hat{\pi}_2) = \max_{\pi_1} V_1(\pi_1, \hat{\pi}_2) - V_1(\pi_1^E, \pi_2^E)$$

$$\leq H \max_{\pi_1} \left\| \frac{\mu^{\pi_1, \pi_2^E}}{\mu^{\pi_1^E, \pi_2^E}} \right\| \sum_h \mathbb{E}_{(\pi_1^E, \pi_2^E)} [\, || \hat{\pi}_{h,2}(\cdot \,|s) - \pi_2^E(\cdot \,|s) \,||_1]$$

[1] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

Multi-Agent Imitation Learning with reward-free warm-up (MAIL-WARM) [1]

Remember that what we need to control is C_{max} and we can query the experts π^E

Step 1: Reward-free warm-up phases

- Construct datasets for each player i $D^{\pi^E_{-i}}$

Policy Generation

- For each state–time pair (s, h), define a reward that is positive only if state s is reached at time h
- Fix the π_{-i}-players to the Nash queryable policy π^E_{-i}
- Solve the resulting RL problem and add the policy to $\Psi^{\pi^E_{-i}}$

Policy Set Construction

- For N times: fix the π_{-i}-players to π^E_{-i} ;sample π_i-player's policy uniformly from $\Psi^{\pi^E_{-i}}$
- Collect one trajectory for sampled policy

[1] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

Multi-Agent Imitation Learning with reward-free warm-up (MAIL-WARM) [1]

Remember that what we need to control is C_{max} and we can query the experts π^E

Step 1: Reward-free warm-up phases

- Construct datasets for each player i $D^{\pi^E_{-i}}$

Step 2: BC over the datasets generated in Phase 1

- For each agent i, use the dataset $D^{\pi^E_{-i}}$ to compute

$$\hat{\pi}_i = \text{argmin}_\nu \sum_{s,b \in D^{\pi^E_{-i}}} -\log \nu(b|s)$$

Intuition: we are constructing a dataset with a good coverage of the environment

[1] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

Sample complexity

(Informal) For any $\epsilon > 0$ and $\delta_{\text{fail}} \in (0,1)$,if we run MAIL-WARM with

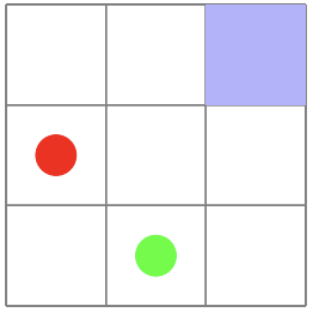
$$N = \mathcal{O}\left(\frac{H^6 S^3 A_{\max}^3 \log(S/\delta_{\text{fail}})}{\epsilon^2}\right)$$

then with probability at least $1 - \delta_{\text{fail}}$,the output policies $(\hat{\boldsymbol{\pi}})$ satisfy:

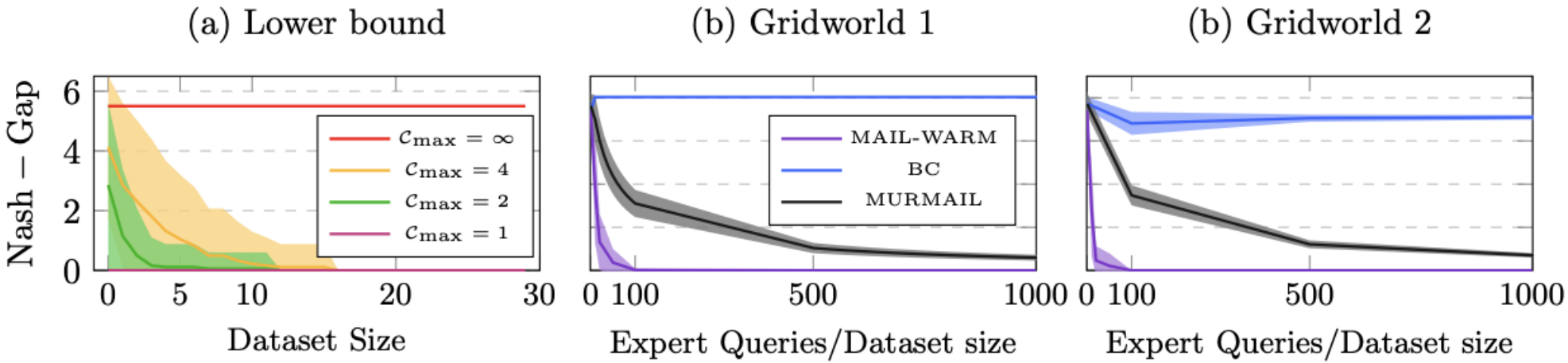
$$\text{Nash-Gap}(\hat{\boldsymbol{\pi}}) \leq \mathcal{O}(\epsilon).$$

MAIL-WARM achieves the optimal sample complexity in the interactive setting

[1] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi



Experimental simulations



Learning Equilibria from Data: Provably Efficient Multi-Agent Imitation Learning, T Freihaut, L Viano, V Cevher, M Geist, G Ramponi, NeurIPS 2025

Rate optimal learning of equilibria from data, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi, AISTATS 2026

Interactive Multi-Agent Imitation Learning

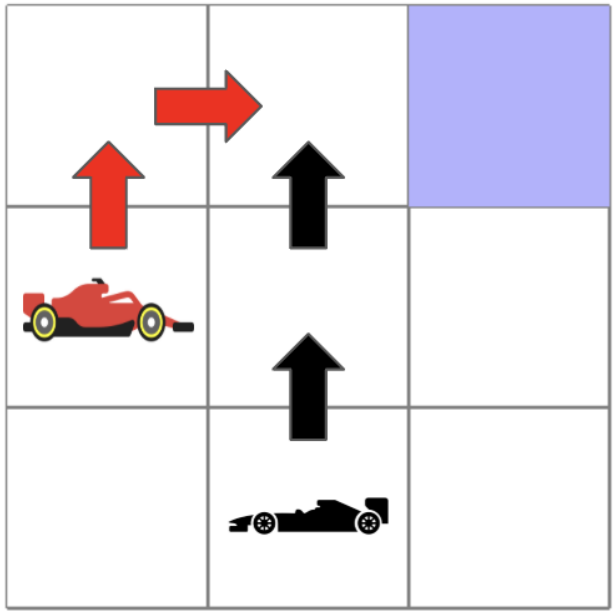
- Observing the best-response [1]
- Having access only to the equilibrium policy [2]
- Using the structure [3]

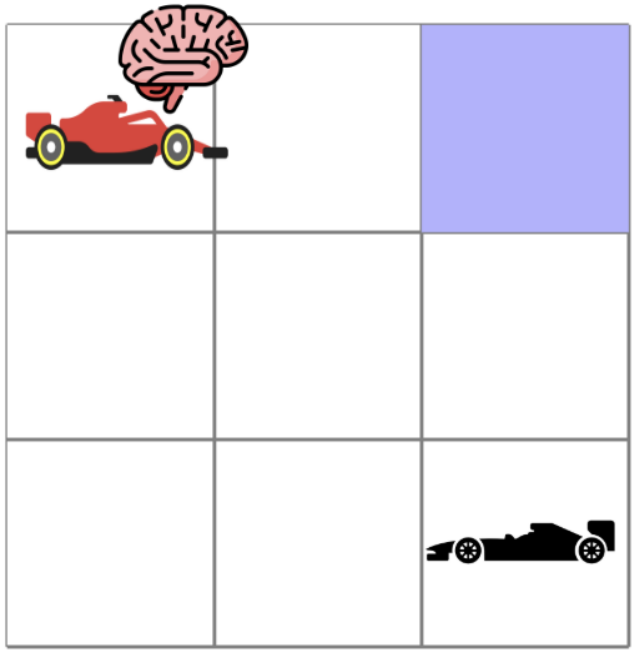
[1] Learning Equilibria from Data: Provably Efficient Multi-Agent Imitation Learning, NeurIPS 2025, T Freihaut, L Viano, V Cevher, M Geist, G Ramponi

[2] Rate optimal learning of equilibria from data, AISTATS 2026, T Freihaut, L Viano, E Nevali, V Cevher, M Geist, G Ramponi

[3] Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026, L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Going beyond the tabular setting





→ Smart agents can generalize to states similar to the seen in the expert dataset
→ Can we formalize this?

Multi-agent imitation learning with function approximation: Linear Markov games and beyond,
ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Going beyond the tabular setting

Linear Markov Games

- Known feature map $\phi(s, a)$
- parametrized Q-function $Q_n^{\pi_1,\pi_2}(s,a) = \phi(s,a)^\top \theta_n^{\pi_1,\pi_2},$

→ agents can generalize across states
→ Coverage is now on the feature level:

$$\mathcal{C}_{\phi,\max_1} = \max_{\pi_2 \in \Pi_2^{\text{softlin}}} \max_{\pi_1^\star \in \text{br}(\pi_2)} \|\phi$$

$\mathcal{C}_{\phi,\max_1}$ can be much smaller than $\mathcal{C}_{\max}$

we can observe generalization across states when the features of two states are not orthogonal

Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Going beyond the tabular setting

Take Away: The best features for BC are the ones with lowest values of $\mathcal{C}_{\phi,\max_1}$ i.e. the ones that allow the best generalization across states, while being expressive enough to realize the Linear MG assumption.

However we are still depending on $\mathcal{C}_{\phi,\max_1}$

Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

We still need interactive learning

Algorithm 1 LSVI-UCB-ZERO-BC

Input: Horizon H, Number of episodes K.
Output: Learned policy pair $\pi_{\text{out}} = (\pi^1_{\text{out}}, \pi^2_{\text{out}})$.
% Loop over players:
for $n \in \{1, 2\}$ **do**
% Exploration Phase:
Define the MDP $\mathcal{M}_n$ where the policy of player n is fixed to π^n_E.
Create the dataset for player n, updating the policy of player $-n$:

$$\mathcal{D}^n \leftarrow \text{LSVI-UCB-ZERO}(\mathcal{M}_n, \texttt{active} = -n).$$

% Imitation Phase:
Run BC on dataset $\mathcal{D}^n$:

$$\pi^n_{\text{out}} \leftarrow \arg \min_{\pi^n \in \Pi^n_{\text{softlin}}} \sum_{(X,A) \in \mathcal{D}^n} -\log \pi^n(A|X)$$

end for

Similar as before we are doing a reward-free warm-up phase for each player

Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

We still need interactive learning

(Informal) The total number of queries LSVI-UCB-ZERO-BC needs to learn an ϵ-optimal policy is

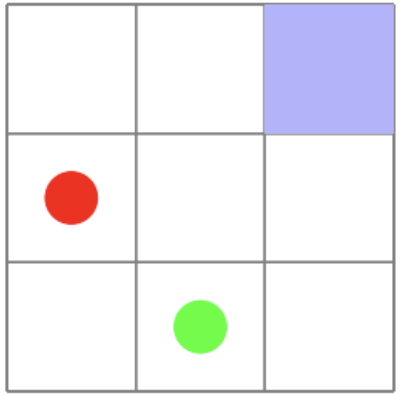
$$KH = \tilde{\mathcal{O}}\left(\frac{H^7 d^4 B^4 \log(B_\theta A_{\max} \delta^{-1})}{\varepsilon^2}\right)$$

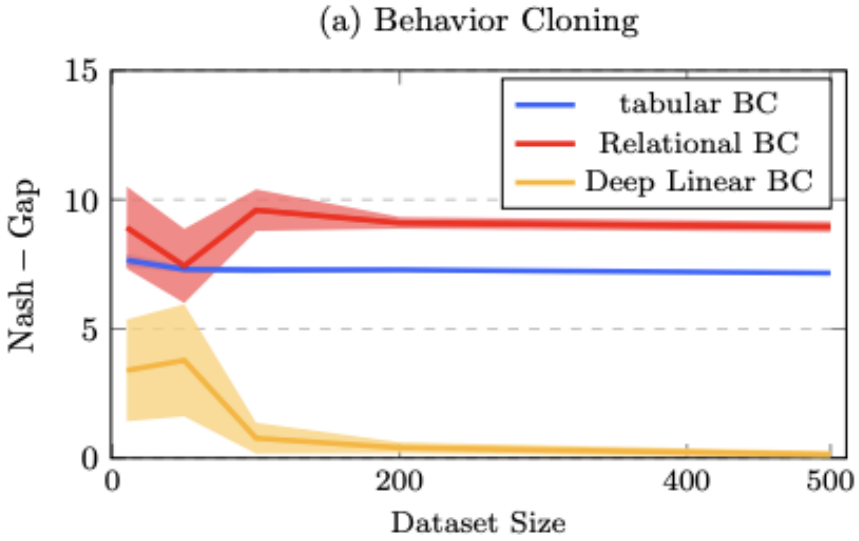
1. Algorithm 1 avoids the dependence on $\mathcal{C}_{\phi,\max_1}$.
2. The sample complexity scales only with the features dimension d and not with the number of states of the game

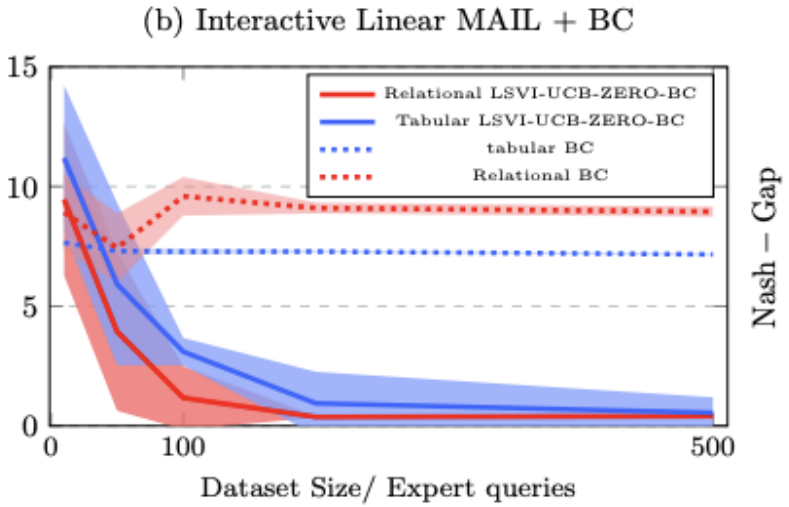
Technical differences with previous results: We have to perform the change of measure at features level in order to avoid dependence on the number of states

Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Experimental evaluation

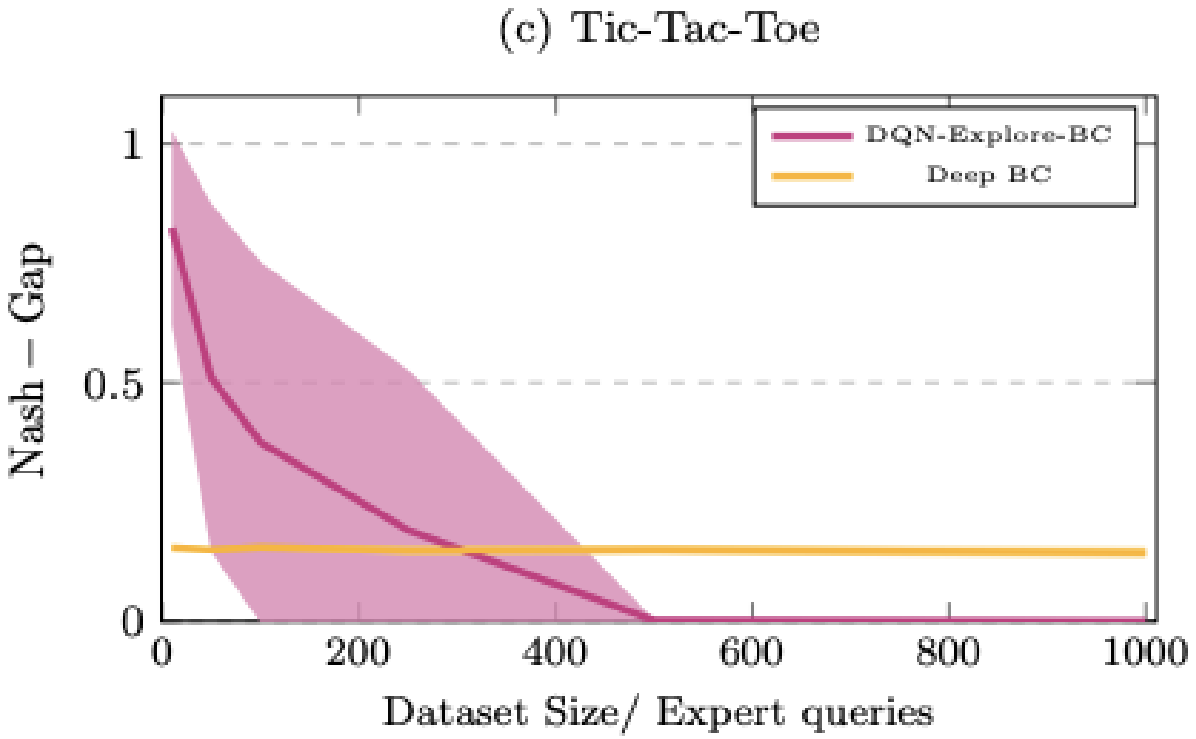






Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Experimental evaluation



Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Experimental evaluation

Opponent	DQN-Explore-BC	BC
Approx. BR	$\mathbf{0.21 \pm 0.04}$	0.00 ± 0.00
Solver noise 1	$\mathbf{0.32 \pm 0.04}$	0.00 ± 0.00
Solver noise 2	$\mathbf{0.60 \pm 0.04}$	0.05 ± 0.01
Solver noise 3	$\mathbf{0.81 \pm 0.01}$	0.21 ± 0.02
Solver noise 4	$\mathbf{0.92 \pm 0.00}$	0.40 ± 0.04
Solver noise 5	$\mathbf{0.97 \pm 0.01}$	0.54 ± 0.05

Multi-agent imitation learning with function approximation: Linear Markov games and beyond, ICML 2026 L Viano, T Freihaut, E Nevali, V Cevher, M Geist, G Ramponi

Conclusion

First rigorous algorithms for **multi-agent imitation learning**

Thank you! Questions?

- Non-interactive IL algorithms cannot learn in MGs
- We described the concentrability coefficient for MAIL
- Learning good features is necessary to use BC in multi-agent games
- We provide two interactive algorithm with rate-optimal sample-complexity
- This framework models and stress-tests rare strategic events beyond historical observations.