ROTATE: Regret-driven Open-ended Training for Ad Hoc Teamwork

Caroline Wang*, Arrasy Rahman*, Jiaxun Cui,
Yoonchang Sung, Peter Stone

LARG
Learning Agents Research Group
The University of Texas at Austin

Sony AI

The Ad Hoc Teamwork Problem

Goal[1] : to design an AHT/ego agent to coordinate with a set of **unknown** but "good faith" teammates

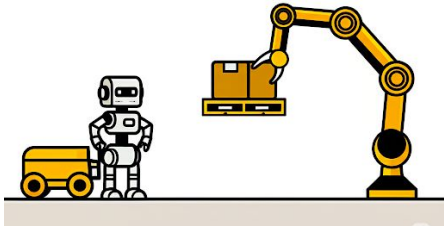
Dec-POMDP $\langle N, S, \{\mathcal{A}^i\}_{i=1}^{|N|}, P, p_0, R, \{\Omega^i\}_{i=1}^{|N|}, O, \gamma \rangle$

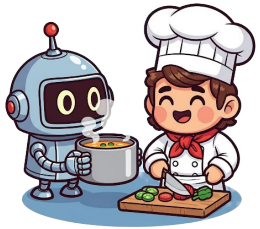
AHT/ego agent π^{ego}; teammates π^{-i}; teammate set Π

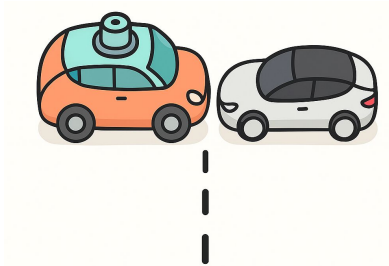
Objective: $\max_{\pi^{\text{ego}}} \mathbb{E}_{\pi^{-i} \sim \psi^{\text{eval}}(\Pi^{\text{eval}}), s_0 \sim p_0} [V(s_0 | \pi^{-i}, \pi^{\text{ego}})]$

Key Challenge: Π^{eval} is unknown

*Problem formulation: N agents; algorithm/experiments: 2 agents







[1] Stone et al., Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination, AAAI 2018.

Ad Hoc Teamwork Methods

Training an AHT Agent [1, 2, 3, 4, 5]:

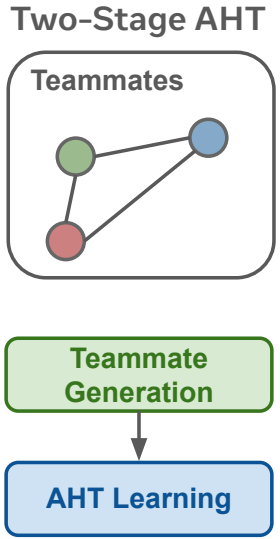
- Train a π^{ego} to coordinate with a pre-specified Π^{train}
  - Generalization is inherently limited by Π^{train}
  - Challenging to construct a good training set

Teammate Generation/Zero-Shot Coordination [6, 7, 8, 9]:

- Goal is to generate diverse Π^{train} such that π^{ego} can generalize to Π^{eval}
  - Unclear how diversity maximization heuristics connect to AHT objective

Two-stage Framework for AHT?

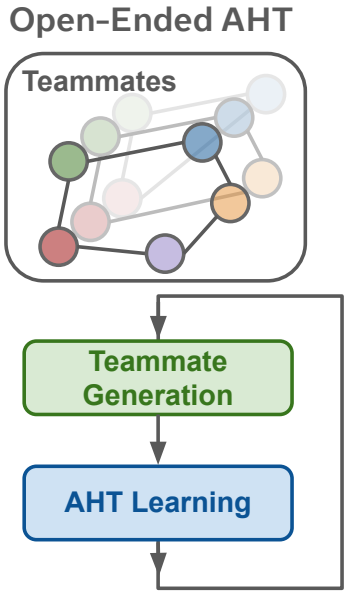
- Challenging to specify how many teammates to generate
- Limits consideration of teammates should influence ego agent learning



[1] Barrett et al., Making friends on the fly: Cooperating with new teammates. Artificial Intelligence, 242:143-171, 2017
[2] Ravula et al., Ad Hoc Teamwork With Behavior Switching Agents. IJCAI 2019.
[3] Macke et al., Expected Value of Communication for Planning in Ad Hoc Teamwork. AAAI 2021.
[4] Papoudakis et al., Agent Modelling under Partial Observability for Deep Reinforcement Learning. NeurIPS 2021.
[5] Rahman et al., Towards Open Ad Hoc Teamwork Using Graph-based Policy Learning. ICML 2021.
[6] Strouse et al., Collaborating with Humans without Human Data. NeurIPS 2021.
[7] Lupu et al., Trajectory Diversity for Zero-Shot Coordination. ICML 2021.
[8] Zhao et al., Maximum Entropy Population-Based Training for Zero-Shot Human-AI Coordination. AAAI 2023.
[9] Sarkar et al., Diverse Conventions in Human-AI Collaboration. NeurIPS 2023.

Open-Ended Ad Hoc Teamwork

Can we frame AHT as an open-ended[1], interactive learning process between a teammate generator and an ego agent?



[1] Dennis et al., Emergent Complexity and Zero-Shot Transfer via Unsupervised Environment Design. NeurIPS 2020.

Problem Formulation: Open-Ended AHT

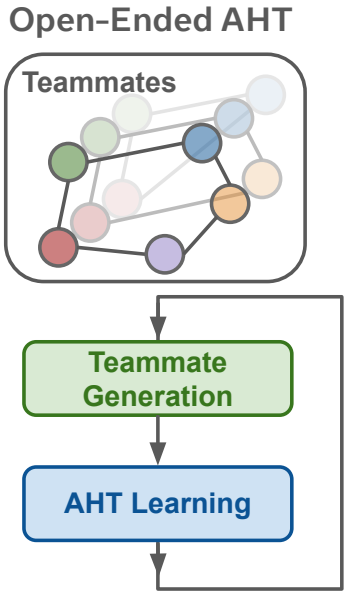
Minimax Regret Objective: $\min_{\pi^{\text{ego}}} \max_{\pi^{-i} \in \Pi^{-i}} \mathbb{E}_{s_0 \sim p_0} \left[\text{CR}(\pi^{\text{ego}}, \pi^{-i}, s_0) \right]$

Ego Learner

Teammate generator

Cooperative Regret[1]: $\text{CR}(\pi^{\text{ego}}, \pi^{-i}, s) = V\left(s | \pi^{-i}, \text{BR}(\pi^{-i})\right) - V\left(s | \pi^{-i}, \pi^{\text{ego}}\right)$

where $\text{BR}(\pi^{-i}) := \max_{\pi} \mathbb{E}_{s \sim p_0} \left[V(s | \pi, \pi^{-i}) \right]$



[1] Erlebach and Cook, RACCOON: Regret-based Adaptive Curricula for Cooperation, CoCoMARL@RLC 2024.

From the AHT Objective to Minimax Coop. Regret

Cooperative Regret[1]: $\mathrm{CR}(\pi^{\mathrm{ego}}, \pi^{-i}, s) = V\left(s|\pi^{-i},\ \mathrm{BR}(\pi^{-i})\right) - V\left(s|\pi^{-i}, \pi^{\mathrm{ego}}\right)$

AHT Objective: $\max_{\pi^{\mathrm{ego}}} \mathbb{E}_{\pi^{-i}\sim\psi^{\mathrm{eval}}(\Pi^{\mathrm{eval}}), s_0\sim p_0}\left[V(s_0|\pi^{-i}, \pi^{\mathrm{ego}})\right]$

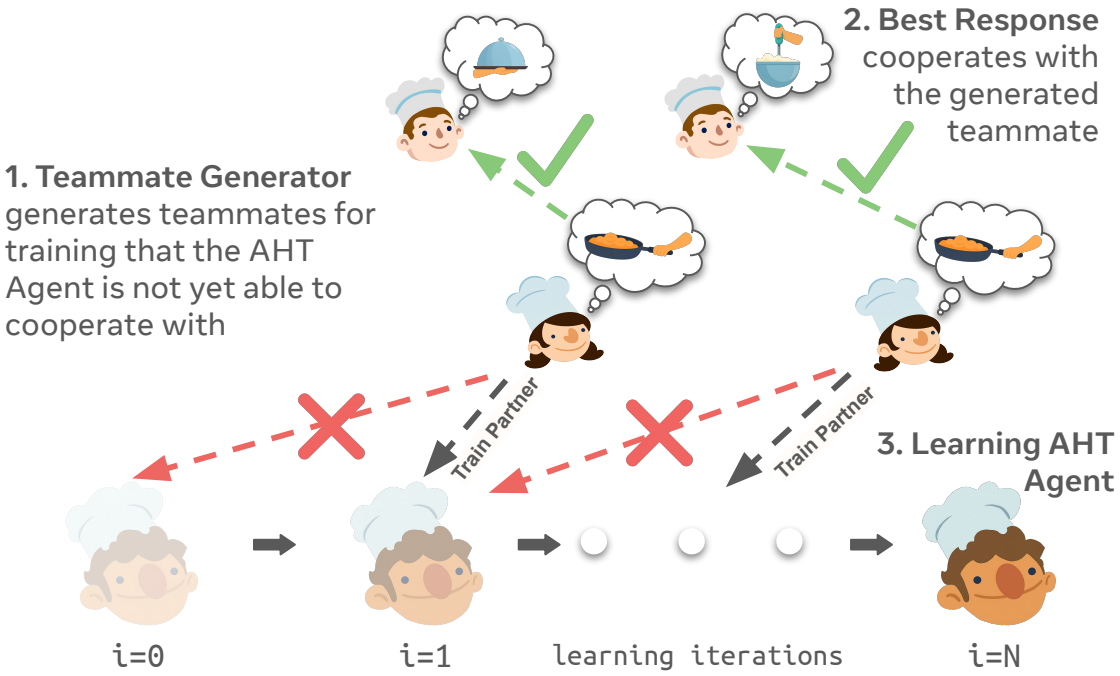
$\iff \min_{\pi^{\mathrm{ego}}} \mathbb{E}_{\pi^{-i}\sim\psi^{\mathrm{eval}}(\Pi^{\mathrm{eval}}), s_0\sim p_0}\left[\mathrm{CR}(\pi^{\mathrm{ego}}, \pi^{-i}, s_0)\right]$

worst case assumption

Minimax Regret Objective: $\min_{\pi^{\mathrm{ego}}} \max_{\pi^{-i}\in\Pi^{-i}} \mathbb{E}_{s_0\sim p_0}\left[\mathrm{CR}(\pi^{\mathrm{ego}}, \pi^{-i}, s_0)\right]$

[1] Erlebach and Cook, RACCOON: Regret-based Adaptive Curricula for Cooperation, CoCoMARL@RLC 2024.

ROTATE: Regret-driven Open-ended Training for AHT



ROTATE: Teammate Generation Objective

How should we generate regret-maximizing teammates?

➔ Key challenge of naive regret maximization (per-trajectory regret): *self-sabotage*

- Previously observed by teammate generation methods optimizing regret-like objectives [1, 2, 3]

Terminology

XP (cross-play): Interactions between $\pi^{\text{ego}}, \pi^{-i}$

SP (self-play): Interactions between $\pi^{-i}, \text{BR}(\pi^{-i})$

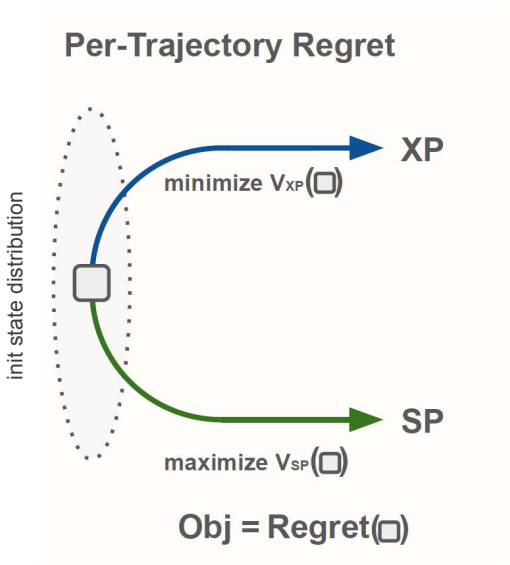
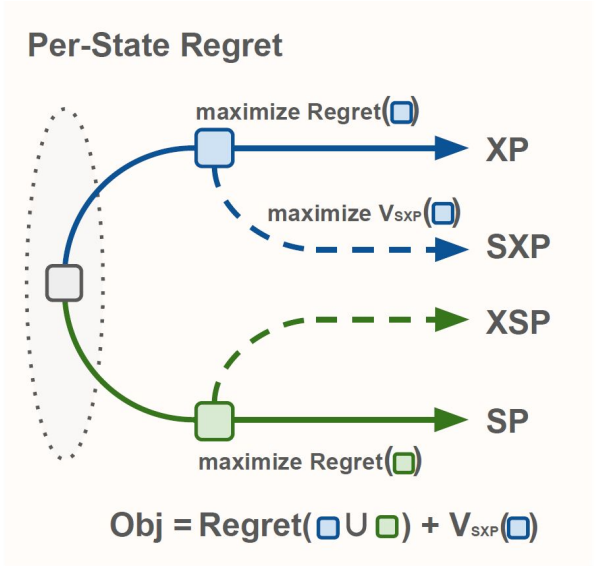


Fig 1. Regret as “SP - XP”.

[1] Sarkar et al., Diverse Conventions in Human-AI Collaboration. NeurIPS 2023.
[2] Rahman et al., Generating Teammates for Training Robust Ad Hoc Teamwork Agents via Best Response Diversity. TMLR 2023.
[3] Sarkar et al., Diverse Conventions in Human-AI Collaboration. NeurIPS 2023.

ROTATE: Teammate Generation Objective

- ➔ We propose a per-state regret objective to mitigate sabotage
- ➔ Key idea: maximize regret from all states encountered in SP and XP
  - From XP states: discourages teammate from irreversibly harming coordination
  - From SP states: discourages teammate from identifying BR from state alone (signalling)



Terminology

XP (cross-play): Interactions between $\pi^{\text{ego}}, \pi^{-i}$

SP (self-play): Interactions between π^{-i}, $\text{BR}(\pi^{-i})$

SXP: self-play from cross-play states

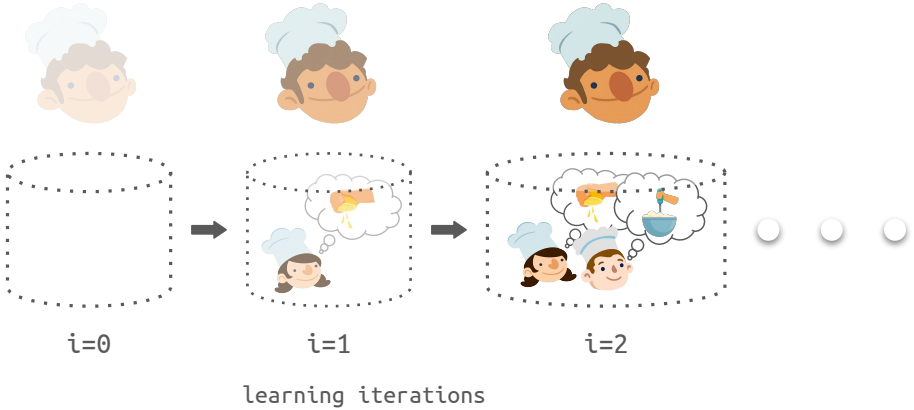
XSP: cross-play from self-play states

ROTATE: Ego Agent Training

How can we train a regret-minimizing ego agent?

➔ Generated teammate policies are added to a population buffer

➔ Recurrent S5 actor-critic architecture[1] , optimized using PPO[2]



[1] Lu et al., Structured State Space Models for In-Context Reinforcement Learning. NeurIPS 2023.
[2] Schulman et al., Proximal Policy Optimization Algorithms. arXiv:1707.06347 2017.

Experiments

- Tasks:
  - 5 Overcooked layouts[1, 2]
  - 7x7 LBF[3]
  - Illustrative matrix game
- Evaluation teammates:
  - 8-14 manually designed agents per task
  - Reward shaping+PPO, heuristic agents; BRDIV
- Metric: normalized return
- Baselines:
  - UED: PAIRED[4], Minimax Return[5]
  - AHT: FCP[6]; BRDiv[7]; CoMeDi[8]

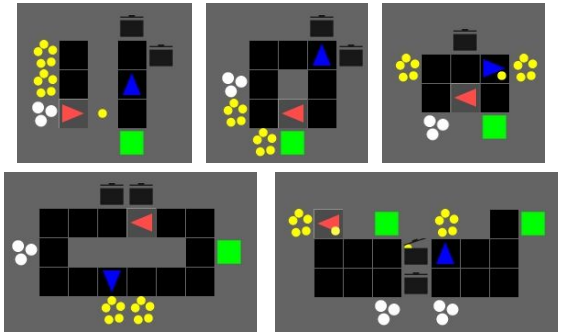


Figure 1. Overcooked layouts.

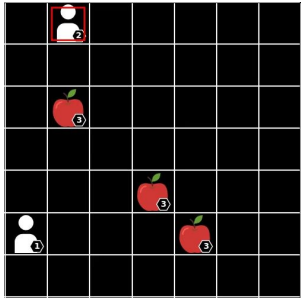
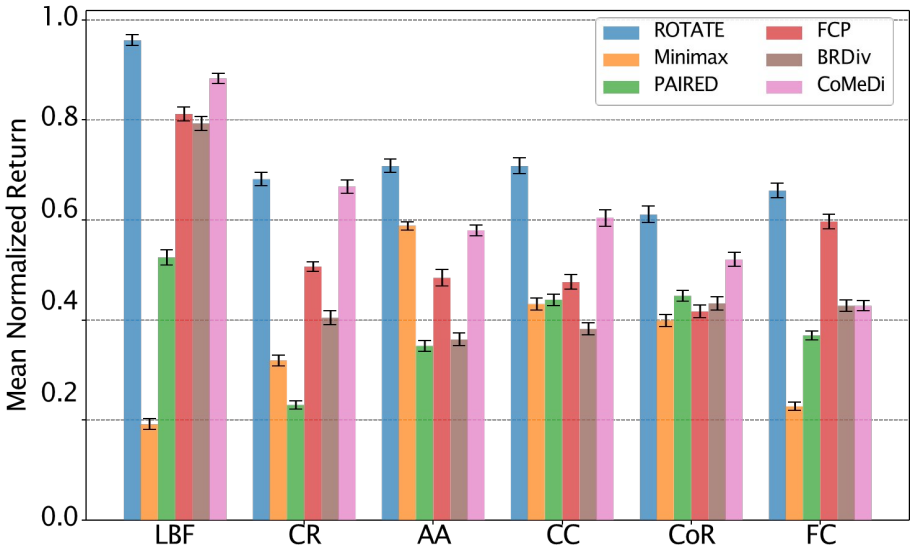
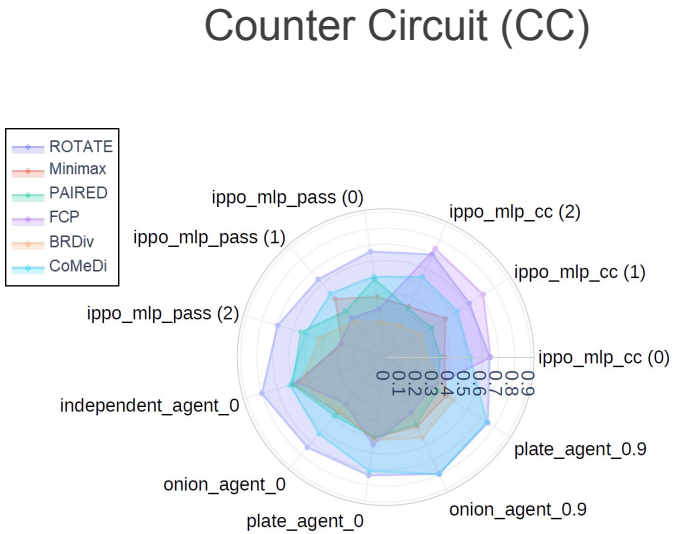


Figure 2. Level-based foraging.

[1] Rutherford et al., JaxMARL: Multi-Agent RL Environments and Algorithms in JAX. NeurIPS 2024.
[2] Carroll et al., On the Utility of Learning about Humans for Human-AI Coordination. NeurIPS 2019.
[3] Albrecht and Ramamoorthy, A Game-Theoretic Model and Best-Response Learning Method for Ad Hoc Coordination in Multiagent Systems. AAMAS 2013.
[4] Dennis et al., Emergent Complexity and Zero-Shot Transfer via Unsupervised Environment Design. NeurIPS 2020.
[5] Morimoto and Doya, Robust Reinforcement Learning. Neural Computation, 17(2):335-339, 2005.
[6] Strouse et al., Collaborating with Humans without Human Data. NeurIPS 2021.
[7] Rahman et al., Generating Teammates for Training Robust Ad Hoc Teamwork Agents via Best Response Diversity. TMLR 2023.
[8] Sarkar et al., Diverse Conventions in Human-AI Collaboration. NeurIPS 2023.

Main Result: ROTATE vs Baselines

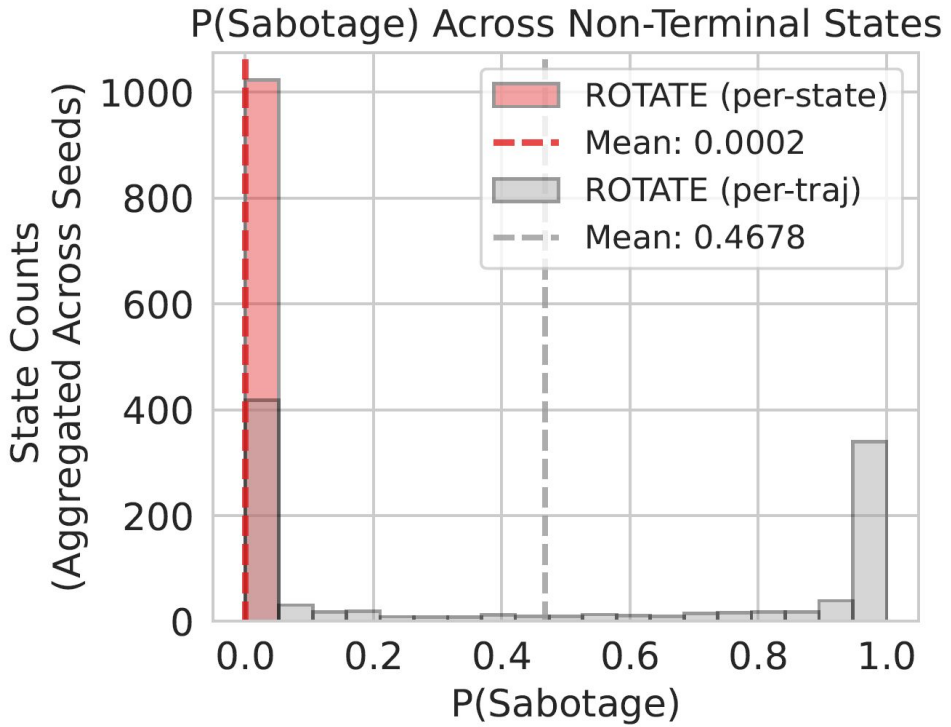




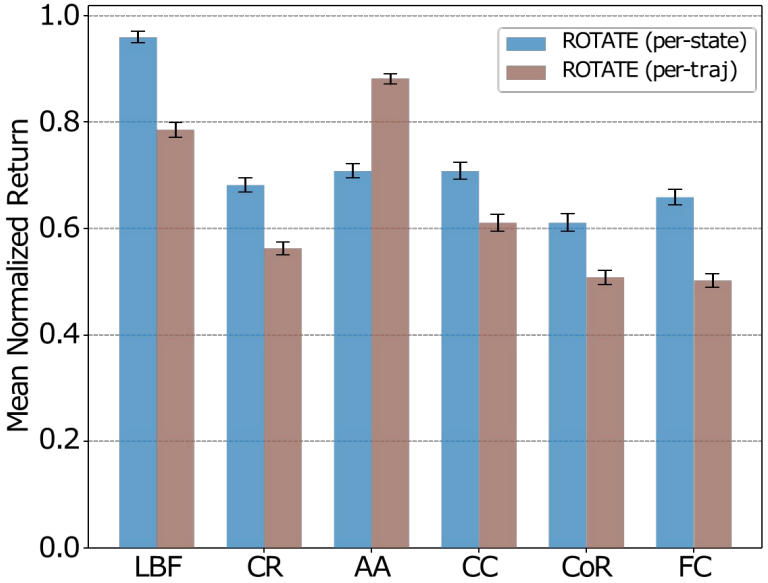
Per-State Regret vs Per-Trajectory Regret

	H	T	S
H	1	0	-1
T	0	1	-1
S	-1	-1	-1

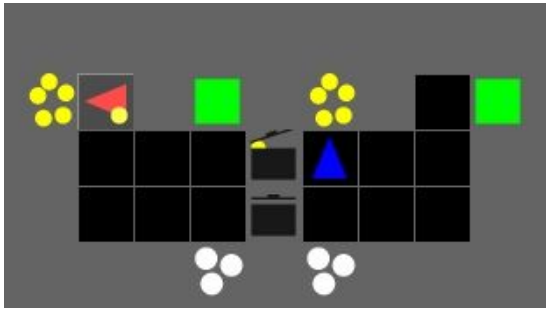
Table 1. Payoff matrix for iterated sabotage game. Game is repeated for 5 iterations.



Per-State Regret vs Per-Trajectory Regret



Asymmetric Advantages



Conclusion

- Contributions:
  - Reformulating AHT as an open-ended learning problem
  - ROTATE, a practical algorithm that substantially outperforms baselines

- Future Work
  - Evaluate ROTATE on broader problem settings:
    - N-agent AHT scenarios
    - Partially observable scenarios
  - Continually learning AHT agents
  - Cooperative regret as an evaluation method for AHT agents

Questions/Comments?

arxiv.org/abs/2505.23686

Caroline Wang

Arrasy Rahman

Jiaxun Cui

Yoonchang Sung

Peter Stone

Effect of Population Buffer