# Minimum Coverage Sets for Training Robust Ad Hoc Teamwork Agents

Arrasy Rahman[1], Jiaxun Cui[1], Peter Stone[1,2]
[1]The University of Texas at Austin, [2]Sony AI

## Ad Hoc Teamwork & its Challenges

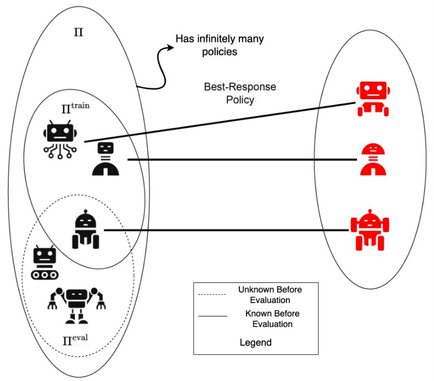
### Ad Hoc Teamwork

- Create an adaptive agent (**learner**) that can collaborate with others without prior coordination mechanisms
- Train learner policy ($\pi^{*,i}$) with training set of teammate policies ($\Pi^{\text{train}}$)
- Given a holdout set of teammate policies, $\Pi^{\text{eval}}$, evaluate the expected returns of $\pi^{*,i}$ defined below:

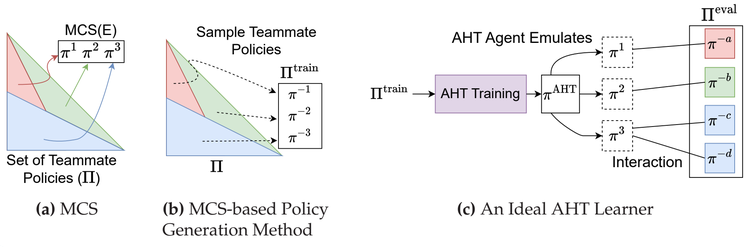
$$\mathbb{E}_{\substack{\boldsymbol{\pi}^{-i}\sim\mathbb{U}(\Pi^{\text{eval}}),a_t^i\sim\pi^{*,i},\\ a_t^{-i}\sim\boldsymbol{\pi}^{-i},P,O}}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\right] \quad (1)$$

### Challenges

- $\Pi^{\text{eval}}$ is unknown when training the learner
- Cannot train with the infinite set of all teammate policies ($\Pi$)
- $\Pi^{\text{train}}$ not necessarily representative of $\Pi^{\text{eval}}$




## Ideal Teammate Policy Generation Method




**(a)** MCS **(b)** MCS-based Policy Generation Method **(c)** An Ideal AHT Learner

1. Find the minimum coverage set (MCS).
2. For each $\pi^i \in$ MCS, sample $\pi^{-i}$ that has $\pi^i$ as its best-response (BR) policy and include as part of $\Pi^{\text{train}}$.

**Goal:**

- Enable learner to emulate BR policy to any $\pi^{-i} \in \Pi$.

## Lagrangian Best Response Diversity (L-BRDiv)

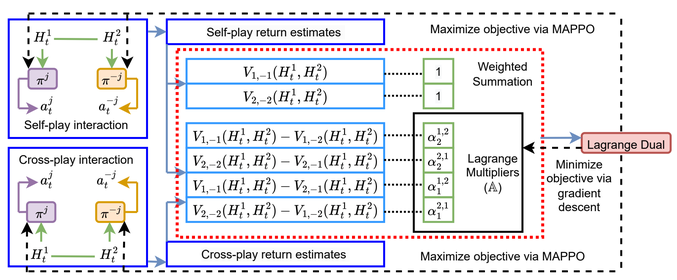
Generate $\Pi^{\text{train}} = \{\pi^{-i}\}_{i=1}^K$ and their set of BR policies $\{\pi^i\}_{i=1}^K$ by optimizing the following objective:

$$\max_{\substack{\{\pi^i\}_{i=1}^K\subseteq\Pi,\\ \{\pi^{-i}\}_{i=1}^K\subseteq\Pi}} \sum_{i\in\{1,2,\ldots,K\}} \mathbb{E}_{s\sim p_0}\left[\mathbf{R}_{i,-i}(H_t)\right], \quad (2)$$

with the following constraints that must be fulfilled for all $i,j \in \{1,2,\ldots,K\}$ and $i \neq j$:

$$\mathbb{E}_{s\sim p_0}\left[\mathbf{R}_{j,-i}(H_t)\right] + \tau \leq \mathbb{E}_{s\sim p_0}\left[\mathbf{R}_{i,-i}(H_t)\right], \quad (3)$$

$$\mathbb{E}_{s\sim p_0}\left[\mathbf{R}_{i,-j}(H_t)\right] + \tau \leq \mathbb{E}_{s\sim p_0}\left[\mathbf{R}_{i,-i}(H_t)\right]. \quad (4)$$

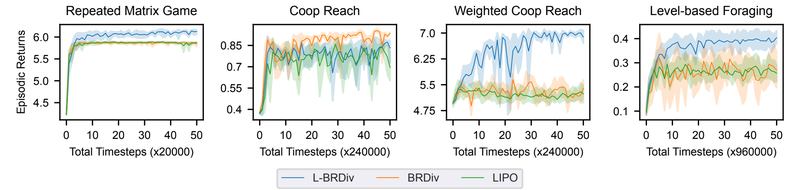



## Experiment Results

**Generalization Experiments**

**(Baselines)** Teammate policy generation methods maximizing adversarial diversity

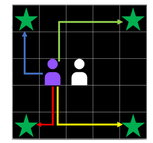
**(Experiment Protocol)** With $\Pi^{\text{train}}$ generated by each compared method, train the learner and evaluate its returns when dealing with holdout policies in $\Pi^{\text{eval}}$



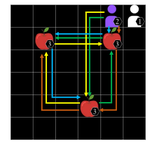

### Analysis of Generated Policies

| | $\pi$(A) | $\pi$(B) | $\pi$(C) |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |

**(a) Repeated Matrix Game**



**(b) (Weighted) Coop Reaching**

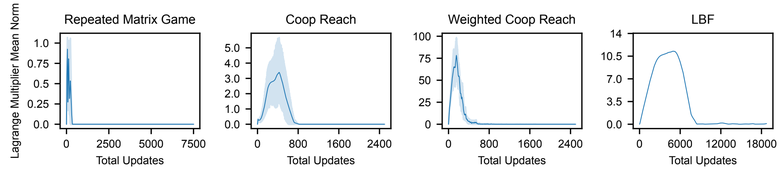


**(c) LBF**

- L-BRDiv facilitates the discovery of $\Pi^{\text{train}}$ having more members with different BR policies
- The BR policies of $\Pi^{\text{train}}$ encompass all members of the MCS

### Lagrange Multiplier Analysis

- Lagrange multipliers keep increasing while constraints are violated
- Eventually, the Lagrange multipliers converge to zero once constraints are fulfilled




## Summary & Future Work

### Our Contributions

- Important concept (i.e. minimum coverage sets) for generating $\Pi^{\text{train}}$ that enable training robust AHT agents
- A Lagrangian multiplier-based teammate generation method (i.e. L-BRDiv) that outperforms existing state-of-the-art baselines

### Future Work

- Extend agent generation method to general-sum games
- Generalizing to N-Player games

## External Links

**(a) Paper**

**(b) Code Repository**