Can Reinforcement Learning Solve Asymmetric Combinatorial-Continuous Zero-Sum Games?

ICLR 2025

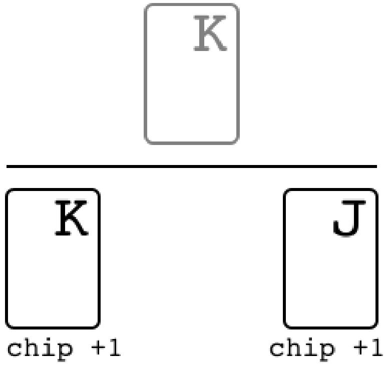
Yuheng Li, Panpan Wang, Haipeng Chen

Data-Driven Decision Intelligence Lab
College of William & Mary

William & Mary | DATA SCIENCE

Motivation

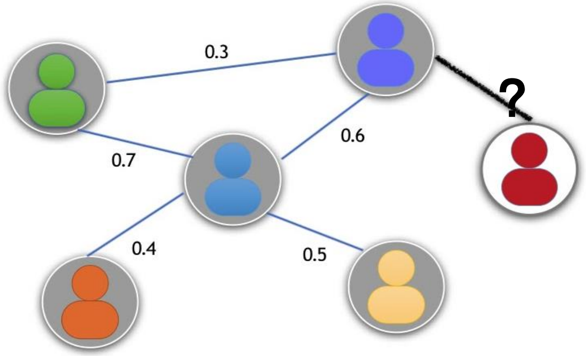
Board Game



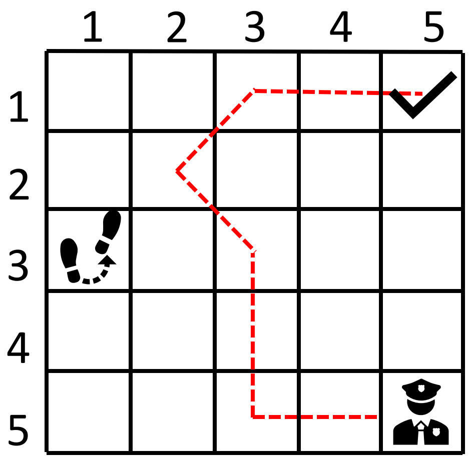
Leduc Poker

		WOMAN	
		Boxing	Shopping
MAN	Boxing	2,1	0,0
	Shopping	0,0	1,2

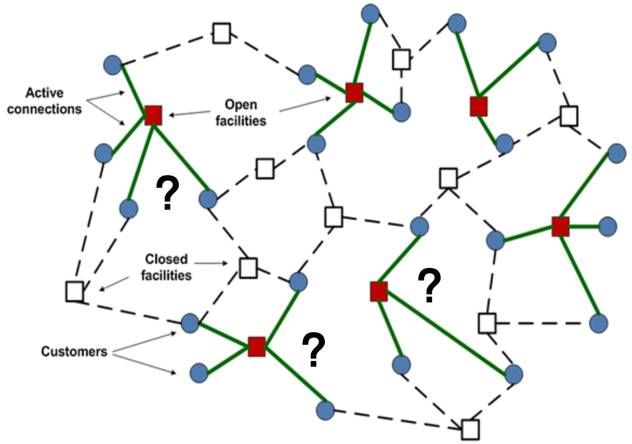
Battle of the Sexes game



Influence Maximization with uncertain weights



Patrolling Game



Facility location under unknown effect

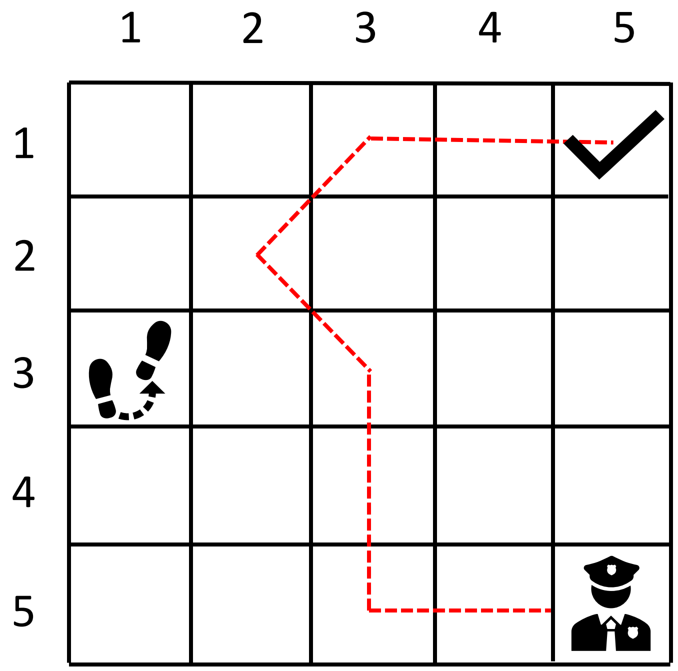
No Game Definition on these games!

What’s the Asymmetric Combinatorial-Continuous zEro-Sum (ACCES) Game?

- Player 1: **Combinatorial** strategy space
- Player 2: **Infinite and compact** strategy space with a continuous utility function

Ep. Patrolling Game,

- Player 1: defender, choosing **a feasible constrained route** to patrol.
- Player 2: attacker, deciding the **attack probability** for targets.
- Utility function: the expectation of successfully protected target values



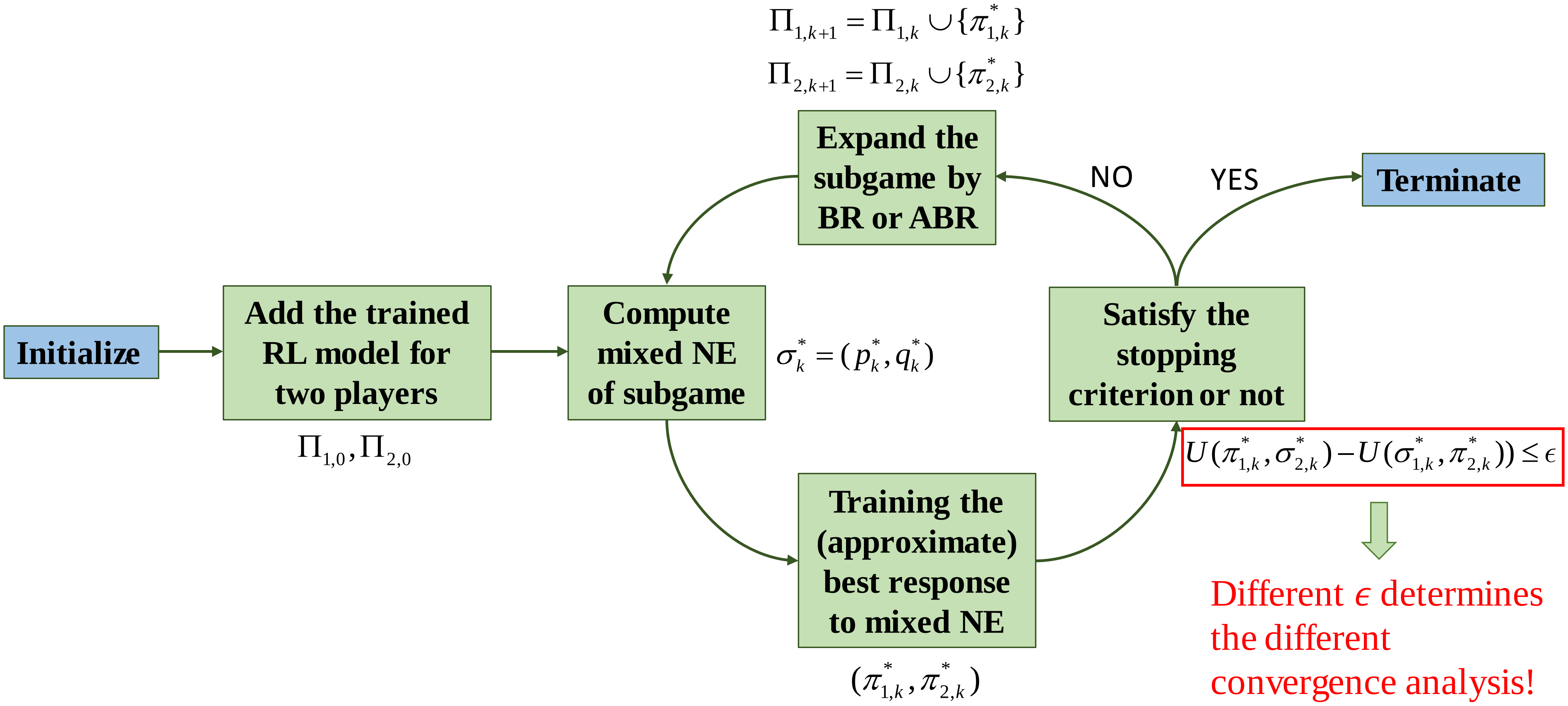
Patrolling Game

Contributions

1. Summarize and define the ACCES game
2. The **existence of mixed NE** in ACCES games
3. **CCDO & CCDO-RL Framework**
    - Novel Convergence Guarantee
    - First practical algorithm to solve ACCES games
4. **Empirical evaluations** on three instances

W&M DATA SCIENCE

CCDO & CCDO-RL Framework



CCDO & CCDO-RL Convergence Analysis

- **Existence of NE (Theorem 1):**

The ACCES game has a mixed strategy Nash Equilibrium.

- **CCDO Convergence Analysis (Theorem 2):**

1. When the stopping criterion $\boldsymbol{\epsilon = 0}$, CCDO possibly iterates in an infinite number of iterations. However, every weakly convergent subsequence in the **subgame equilibrium sequence** $\{p_k^*, q_k^*\}$ converges to **the equilibrium of the whole game**.
2. When the stopping criterion $\boldsymbol{\epsilon > 0}$, CCDO converges to an **ϵ-equilibrium** in a finite number of epochs.

CCDO & CCDO-RL Convergence Analysis

- **CCDO-RL Convergence Analysis (Theorem 3):**

1. When the stopping criterion $\boldsymbol{\epsilon = 0}$, if the approximate response oracle for Player 2 has a **uniform lower bound** for every mixed strategy, then CCDO-RL must converge to an $(\boldsymbol{\epsilon + \epsilon_1 + \epsilon_2})$**-equilibrium** in a finite iterations.
2. When the stopping criterion $\boldsymbol{\epsilon = 0}$ and CCDO-RL iterates **infinite** rounds, every weakly convergent subsequence converges to an $\boldsymbol{\epsilon_1}$**- equilibrium**.
3. When the stopping criterion $\boldsymbol{\epsilon > 0}$, CCDO-RL converges to an $(\boldsymbol{\epsilon + \epsilon_1 + \epsilon_2})$ **-equilibrium** in a finite number of epochs.

ϵ_1 and ϵ_2 are the approximate error bound of approximate best responses for Player 1 and 2 respectively.

Experiments

In three instances under two types of adversary, CCDO-RL and stochastic adversary, CCDO-RL has

- **Better average reward on seen graphs.**
- **Greater generalizability on unseen graphs.**

Table 1: Average reward against CCDO-RL's adversary (on seen graphs)

method	ACSP (Mean±Std)		ACVRP (Mean±Std)		PG (Mean±Std)	
	20 nodes	50 nodes	20 nodes	50 nodes	20 nodes	50 nodes
Heuristic	6.13±1.20	7.55±1.42	7.65±1.23	13.38±1.70	2.64±1.03	4.53±1.84
RL against Stoc	3.50±0.47	4.55±0.62	7.55±1.16	13.90±1.63	2.71±0.90	4.80±2.18
CCDO-RL	**3.25**±0.42	**4.31**±0.51	**7.42**±1.21	**13.28**±1.52	**2.75**±0.87	**5.01**±1.91

Table 2: Generalizability against CCDO-RL's adversary (on unseen graphs)

method	ACSP (Mean±Std)		ACVRP (Mean±Std)		PG (Mean±Std)	
	20 nodes	50 nodes	20 nodes	50 nodes	20 nodes	50 nodes
Heuristic	6.20±1.33	7.60±1.37	7.64±1.30	13.27±1.87	2.43±0.98	4.19±1.69
RL against Stoc	3.56±0.37	4.57±0.58	7.67±1.30	13.85±1.53	2.50±0.95	4.26±2.17
CCDO-RL	**3.31**±0.35	**4.39**±0.52	**7.55**±1.28	**13.15**±1.59	**2.56**±0.92	**4.70**±1.94

[1] For the average reward of ACSP and ACVRP, smaller is better while for that of PG larger is better.