清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Conservative Offline Policy Adaptation in Multi-Agent Games

Chengjie Wu[1], Pingzhong Tang[12], Jun Yang[3], Yujing Hu[4], Tangjie Lv[4], Changjie Fan[4], Chongjie Zhang[5]

[1]*Institute for Interdisciplinary Information Sciences, Tsinghua University*

[2]*Turingsense* [3]*Department of Automation, Tsinghua University*

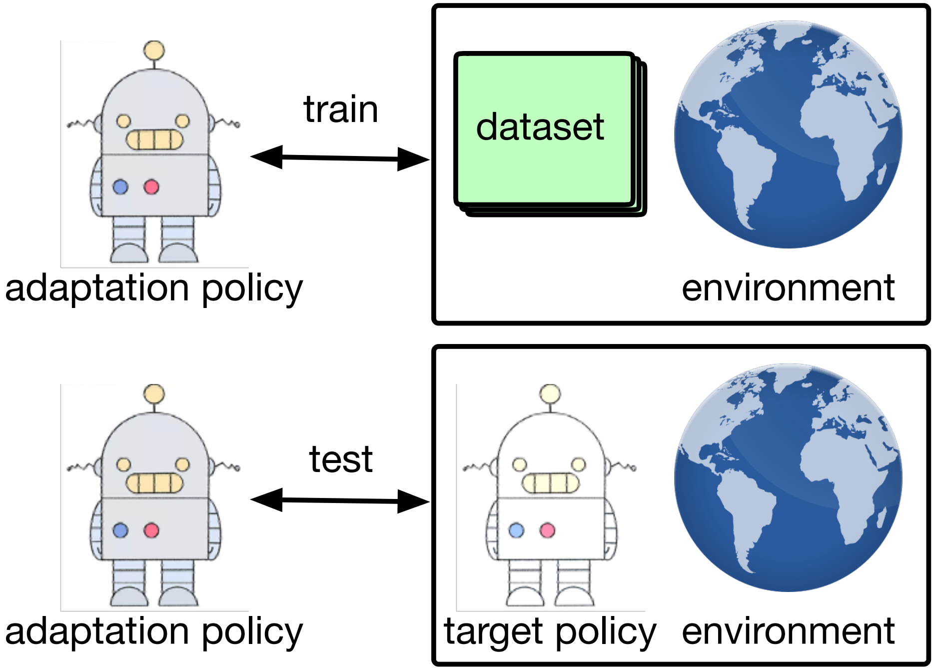
[4]*Fuxi AI Lab, NetEase*

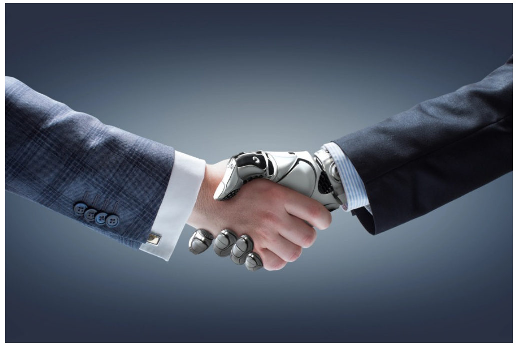
[5]*Department of Computer Science & Engineering, Washington University in St. Louis*

wucj19@mails.tsinghua.edu.cn

Introduction

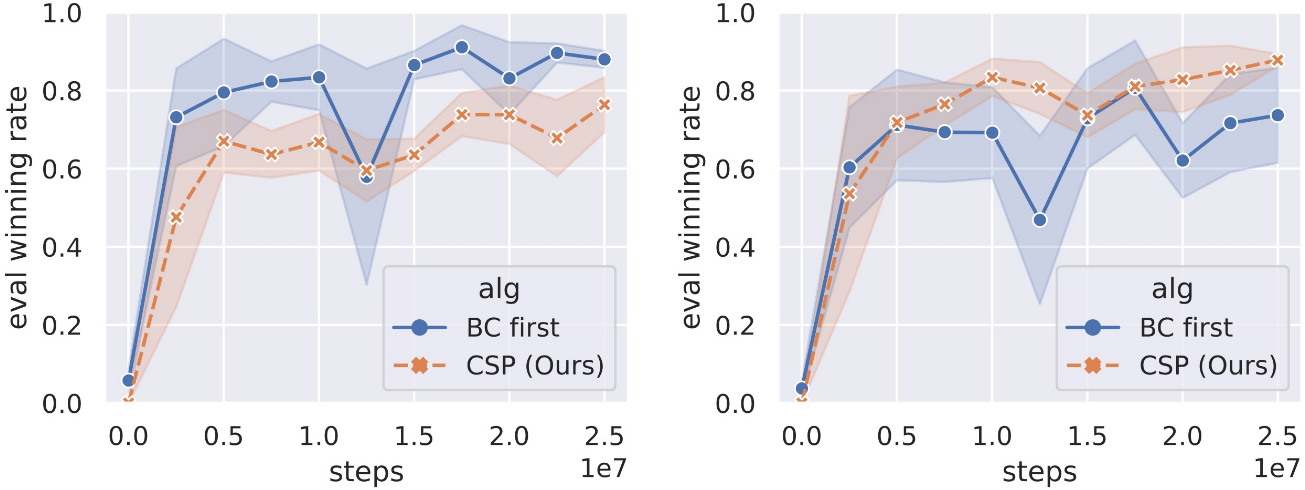
- Problem Formulation: Offline Policy Adaptation in Multi-Agent Games





Introduction

- Challenges
  - ***Distributional shift:*** our estimation about the target agent can differ arbitrarily from the real target on out-of-distribution (OOD) states
  - ***Risk free deviation:*** "conservatism" for addressing distributional shift may not be sufficient in this setting. It is possible to benefit from deviation in multi-agent games



Method

- Conservative Offline Adaptation (COA)

Definition 4.5. ***Conservative offline adaptation (COA).*** Given an unknown policy π_B, and a dataset D of its behavior data, conservative offline adaptation optimizes the worst-case performance against any possible dataset-consistent policy:

$$\max_{\pi} \min_{\mu} J(\pi, \mu), \ \ s.t. \ \mu \in \mathcal{C}_D. \tag{2}$$

The adaptation policy π^* is an optimal risk-free adaptation policy if it is a solution to Objective (2).

Method

- Constrained Self-Play (CSP)

$$\max_{\pi}\min_{\mu} J(\pi, \mu)\ s.t.\ \max_{s\in D} D_{\mathrm{KL}}(\pi_B(\cdot|s)\|\mu(\cdot|s)) \leq \delta$$

Theorem 4.8. *Let π^* be the optimal risk-free offline adaptation policy at the convergence of the optimization of objective 2, and let $\tilde{\pi}$ be the policy at the convergence of objective 5. Then the worst-case adaptation performance of $\tilde{\pi}$ is near-optimal:*

$$\min_{\mu\in\mathcal{C}_D} J(\pi^*, \mu) \geq \min_{\mu\in\mathcal{C}_D} J(\tilde{\pi}, \mu) \geq \min_{\mu\in\mathcal{C}_D} J(\pi^*, \mu) - R_M\sqrt{2\delta}\left(1 + \frac{2\gamma\delta}{(1-\gamma)^2}\right). \tag{6}$$

Experiment

- Constrained Self-Play (CSP)

Table 4: Winning rates of offline adaptation policy in 4 scenarios of Google Football: 3vs1 (defender), 3vs1 (attacker), RPS (defender), and Counterattack Easy (defender). For each scenario, we experiment with 5 independent target opponents.

Scenario	Method	1	2	3	4	5
3vs1 defender	**CSP(Ours)**	**0.9 ± 0.06**	**0.6 ± 0.12**	**0.45 ± 0.04**	**0.32 ± 0.11**	**0.76 ± 0.16**
	BC-First	0.64 ± 0.04	0.46 ± 0.09	0.2 ± 0.01	0.16 ± 0.04	0.56 ± 0.09
	Self-Play	0.29 ± 0.09	0.34 ± 0.1	0.26 ± 0.16	0.29 ± 0.18	0.3 ± 0.13
3vs1 attacker	**CSP(Ours)**	**0.81 ± 0.14**	**0.88 ± 0.02**	**0.83 ± 0.07**	**0.84 ± 0.05**	**0.78 ± 0.08**
	BC-First	**0.83 ± 0.03**	0.74 ± 0.21	0.68 ± 0.07	0.79 ± 0.11	0.71 ± 0.15
	Self-Play	0.75 ± 0.15	0.76 ± 0.08	0.73 ± 0.13	0.7 ± 0.21	0.74 ± 0.08
RPS defender	**CSP(Ours)**	0.51 ± 0.33	**0.71 ± 0.07**	**0.56 ± 0.13**	**0.38 ± 0.07**	**0.79 ± 0.09**
	BC-First	0.51 ± 0.24	0.41 ± 0.07	0.34 ± 0.18	**0.38 ± 0.04**	0.71 ± 0.02
	Self-Play	**0.57 ± 0.14**	0.36 ± 0.04	0.25 ± 0.04	**0.37 ± 0.1**	0.76 ± 0.03
Counter-attack defender	**CSP(Ours)**	**0.93 ± 0.02**	**0.88 ± 0.04**	**0.81 ± 0.25**	**0.81 ± 0.02**	**0.75 ± 0.06**
	BC-First	0.7 ± 0.17	0.52 ± 0.07	0.53 ± 0.09	0.69 ± 0.14	0.5 ± 0.09
	Self-Play	0.55 ± 0.13	0.41 ± 0.1	0.46 ± 0.09	0.36 ± 0.08	0.36 ± 0.07

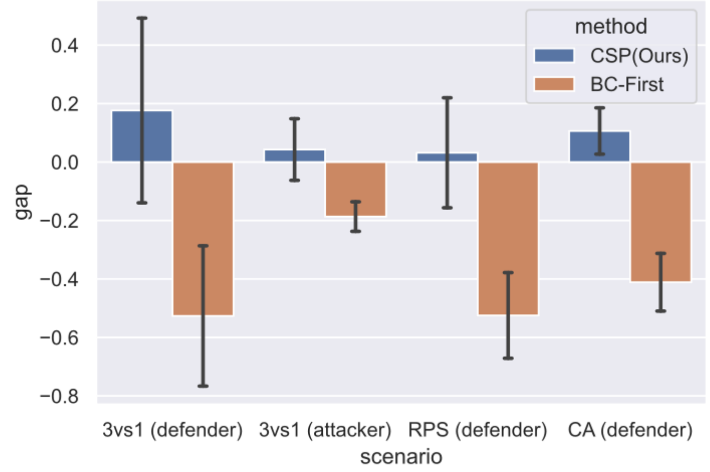


Figure 3: The average test-train performance gap in four scenarios of Google Football. A negative gap indicates the occurrence of unsafe exploitation.

清華大学 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Conservative Offline Policy Adaptation in Multi-Agent Games

Thank you!