

Robust PAC Learning of Concurrent Stochastic Games

Angel Y. He¹ David Parker²

¹Department of Engineering Science

²Department of Computer Science

University of Oxford

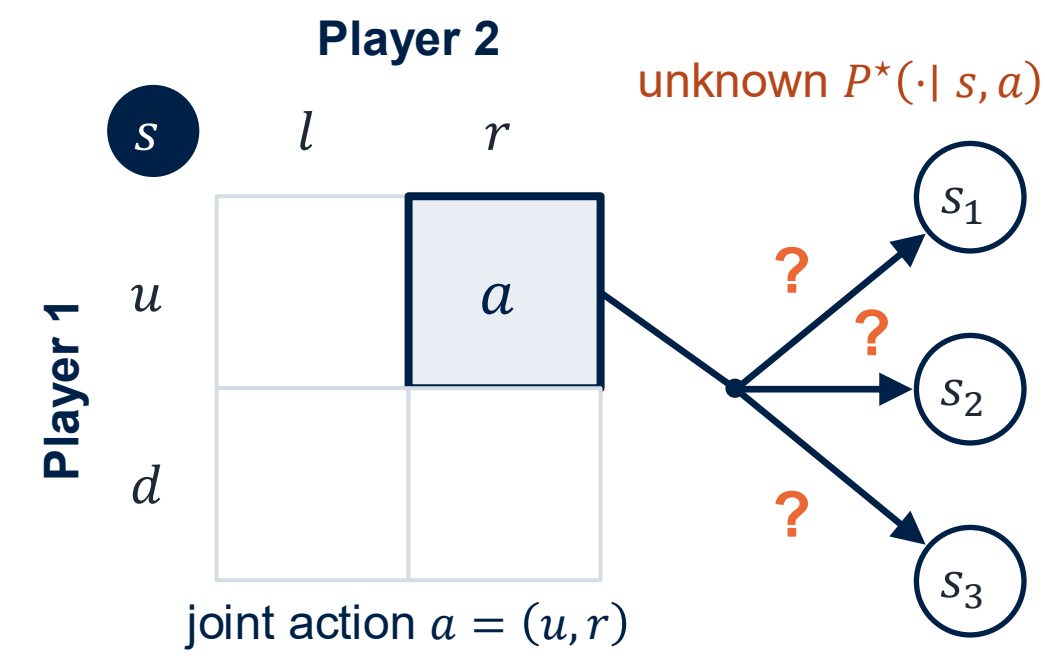


TL;DR

The first PAC-learning framework for two-player general-sum concurrent stochastic games (CSGs) with uncertain transitions, returning with probability $\geq 1 - \delta$ either an ε -optimal ε -Nash equilibrium (ε -NE) if an exact NE exists, or a **sound certificate** of stationary NE non-existence.

$$\tilde{O}\left(R_{\max}^2 H^4 |S|^2 |A| / (\varepsilon^2 p_{\text{reach}})\right) \text{ trajectory samples (Theorem 7)}$$

Learning equilibria from samples



- **2-player CSG**: players act *simultaneously*; their joint action draws the next state from the true kernel P^* .
- P^* is **unknown**.
- **Objectives**: probabilistic and reward reachability, given bounded rewards $|r_t| \leq R_{\max}$.

How many interactions suffice to compute an approximately optimal equilibrium with high probability — and what happens when no exact equilibrium exists?

Two challenges



1 Equilibrium existence under uncertainty

Exact stationary NE may not exist [1]; even if one does, an uncertainty set around P^* may admit no *robust* NE [2].



2 Global coverage from stochastic trajectories

Sufficient joint state–action coverage must come from online trajectories, under joint control and uncertain transitions.

PAC-CSG: an equilibrium or a certificate

Nash margin — how far a stationary profile σ is from being an NE:

$$\mu(\sigma, P) := \min_{i \in N} \min_{s \in S} \inf_{\sigma'_i \in \Sigma_i} [u_i(\sigma, P | s) - u_i(\sigma_{-i}[\sigma'_i], P | s)]$$

$$\mu(\sigma, P) \geq 0 \Leftrightarrow \sigma \text{ is an NE}$$

$$\bar{\mu} := \max_{\sigma} \mu(\sigma, P^*) \geq 0 \Leftrightarrow \text{true game } \mathcal{G}^* \text{ has an exact NE}$$

PAC guarantee (Def. 4) — with probability $\geq 1 - \delta$, after polynomially many interactions, the output is:

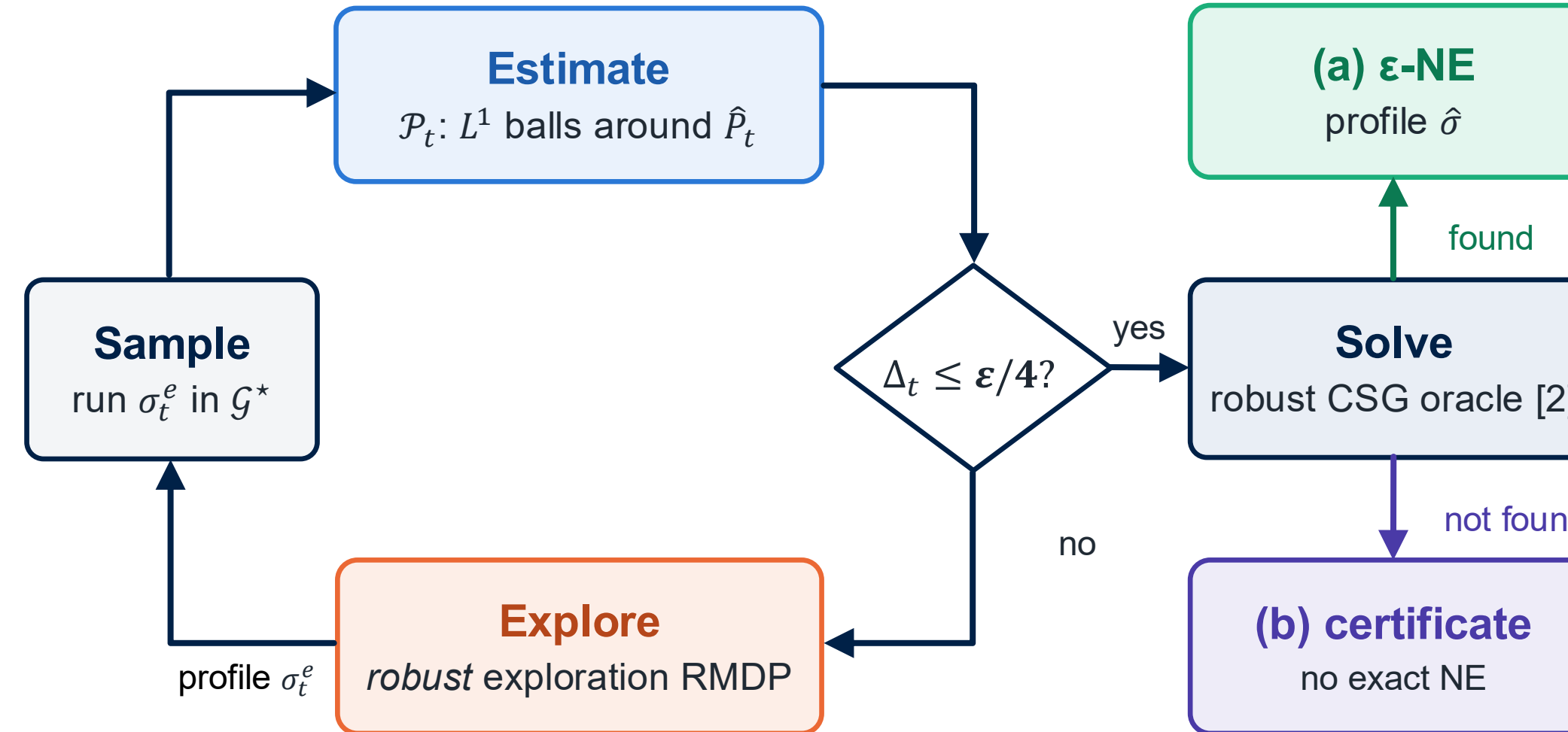


(a) if an exact stationary NE exists: an **ε -NE** $\hat{\sigma}$ that is also **ε -optimal**, i.e., $V^* - u(\hat{\sigma}, P^*) \leq \varepsilon$



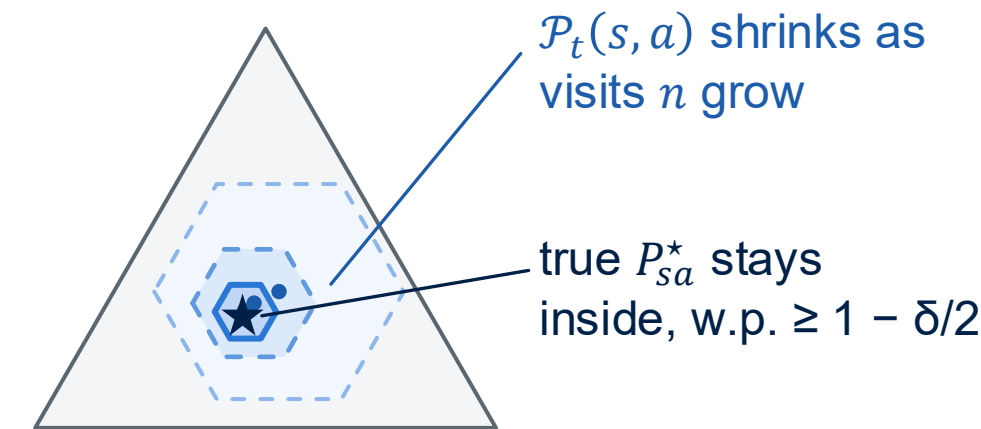
(b) a **sound certificate** that $\mathcal{G}^*$ has **no exact stationary NE**

Algorithm: estimate, test, explore



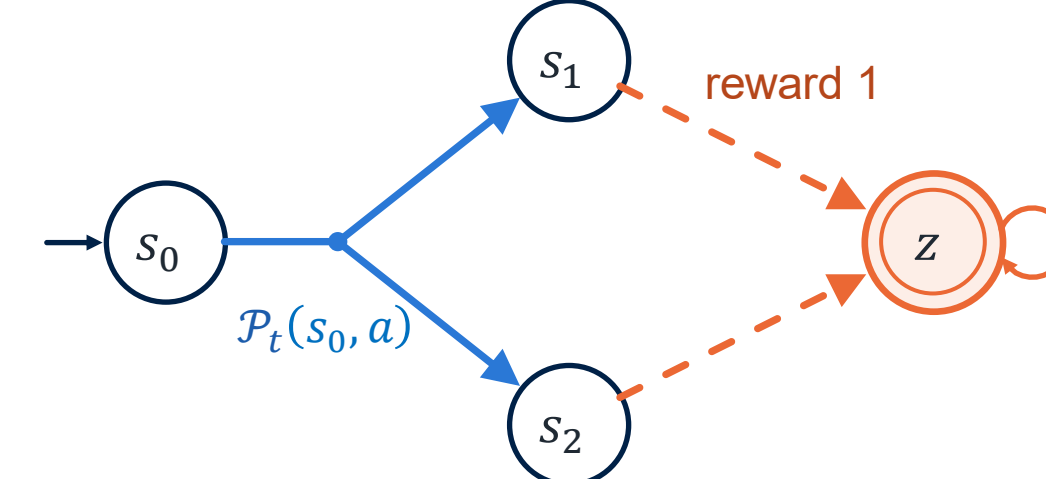
Exploration

L^1 confidence sets



$$\alpha_t(s, a) = \sqrt{\frac{2}{n} \ln \frac{2(2^{|S|} - 2) |S| |A| n(n+1)}{\delta}}$$

Exploration RMDP



— known (s, a) : keeps $\mathcal{P}_t(s, a)$
 - - - unknown (s, a) : to z , reward 1

Why “robust” exploration? Conditioning on $P^* \in \mathcal{P}_t$, the worst-case probability of reaching an unknown slot lower-bounds the true one:

$$p_t^e(P^*) \geq \inf_{P \in \mathcal{P}_t} p_t^e(P) \geq p_{\text{reach}} > 0$$

Why it works

1 Sensitivity Lem. 2

$|u_i(\sigma, P | s) - u_i(\sigma, \hat{P}_t | s)| \leq \Delta_t := 1/2 R_{\max} H^2 \Delta_t^{L^1}$ for all $P \in \mathcal{P}_t$, where $\Delta_t^{L^1}$ is the largest L^1 radius of $\mathcal{P}_t$

2 NE transfer Lem. 3, Prop. 4

NE of $\mathcal{G}^*$ with margin $\mu \Rightarrow (4\Delta_t - \mu)$ -NE under every $P \in \mathcal{P}_t$.
 $\therefore$ Found $\Rightarrow \varepsilon$ -NE of $\mathcal{G}^*$; not found $\Rightarrow$ no NE has margin $\geq 4\Delta_t - \varepsilon$.

3 Coverage Lem. 6, 19, 20; Cor. 21

$p_t^e(P^*) \geq p_{\text{reach}} + \text{Freedman's inequality} \Rightarrow \text{w.p. } \geq 1 - \delta/2$ every (s, a) gets at least $n_{\min}$ visits

Main result

Theorem 7. W.p. $\geq 1 - \delta$, Alg. 1 terminates after at most

$$n_{\min} \text{ visits per slot } \tilde{O}\left(\frac{R_{\max}^2 H^4 |S|^2 |A|}{\varepsilon^2 p_{\text{reach}}}\right) \begin{matrix} \text{Slots} \\ \text{Trajectories per episode} \end{matrix}$$

trajectory samples, and its output satisfies Def. 4, where

$$p_{\text{reach}} := \min_{(s,a): s \notin S_T} \max_{\sigma} \inf_{P: \text{Supp}(P)=\text{Supp}(P^*)} \mathbb{P}^{\sigma, P}[\text{reach}(s, a) \text{ before } S_T]$$

Infinite horizon: $\tilde{O}\left(R_{\max}^2 |S|^6 |A| / (\varepsilon^2 p_T^4 p_{\text{reach}})\right)$, where $p_T > 0$ is the minimum worst-case probability of reaching S_T within $|S|$ steps from any state.

Experiments

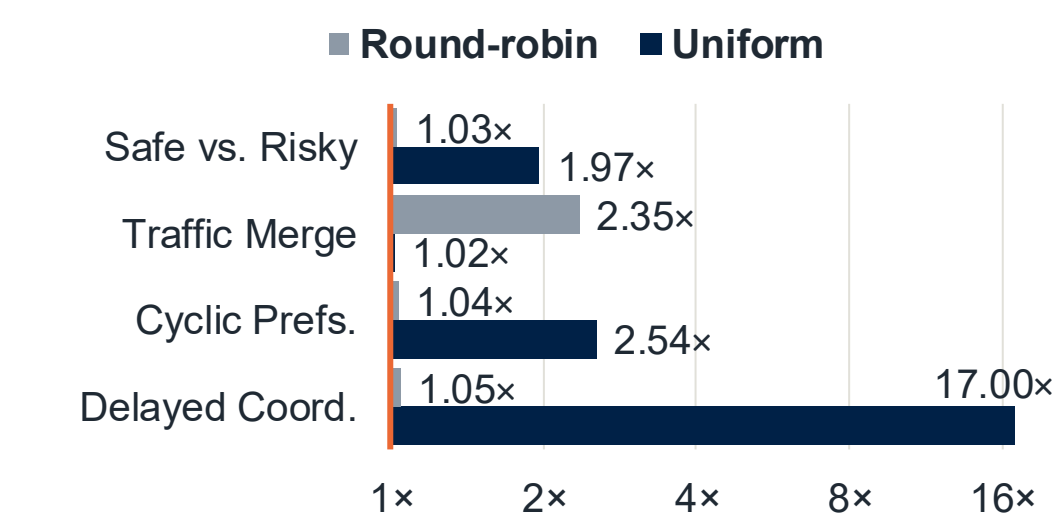


PRISM-games · 6 benchmark CSGs · $\delta = 0.05$

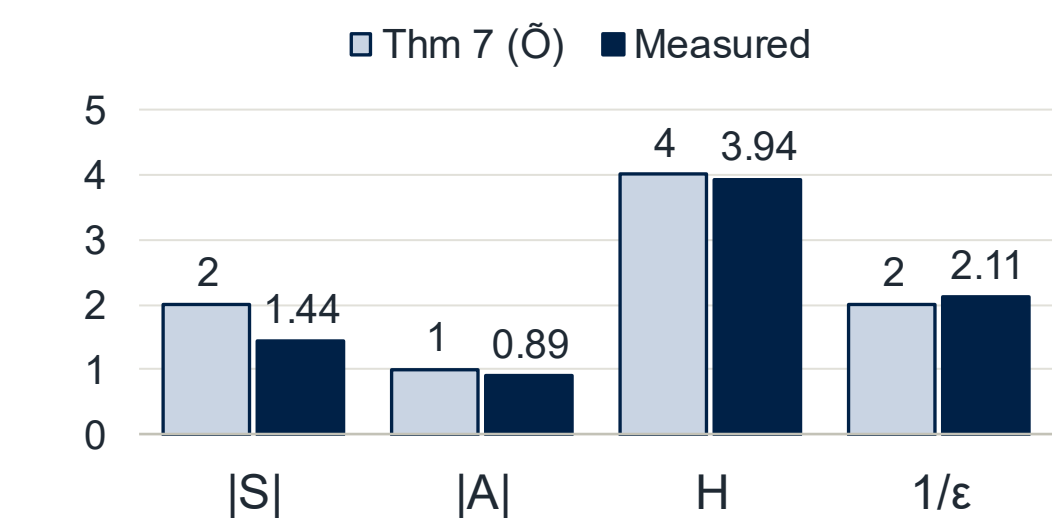
RQ1 · Correct outputs at termination ($\varepsilon = 0.1$)

	returns (a) ε -NE $\hat{\sigma}$	returns (b) NOT_FOUND
exact NE exists ($\bar{\mu} \geq 0$)	✓ oracle's NE value attained all 6 cases	— never (w.p. $\geq 1 - \delta$)
no exact NE	✓ valid ε -NE e.g. <i>Hide-or-run</i> $\Rightarrow$ incomplete	✓ sound certificate e.g. <i>Cyclic Prefs</i> .

RQ2 · Exploration: samples vs. ours (1×)



RQ3 · Scaling (Safe vs. Risky variants)



Optimistic: $\approx$ same samples, but bigger value gap $V^* - \hat{V}$ on Delayed Coord.

Takeaways & limitations

- Robust equilibria computed on the **learned model transfer** to the true game.
- NE **non-existence can be certified soundly** from data.
- **Limitations:** 1) assumes centralised exploration, known transition support, reachability; 2) scalability.



Paper



Project page

References [1] P. Bouyer, N. Markey, D. Stan. Mixed Nash equilibria in concurrent terminal-reward games. FSTTCS 2014.
 [2] A. Y. He, D. Parker. Robust verification of concurrent stochastic games. TACAS 2026.