Robust PAC Learning of Concurrent Stochastic Games

Angel Y. He
Department of Engineering Science
University of Oxford
angelyhe@robots.ox.ac.uk

David Parker
Department of Computer Science
University of Oxford
david.parker@cs.ox.ac.uk

Abstract

We introduce the first Probably Approximately Correct (PAC) learning framework for general-sum concurrent stochastic games (CSGs) with transition uncertainty, while addressing the challenge of Nash equilibrium (NE) existence. Our algorithm maintains data-driven L^1 confidence sets over transition kernels and solves a robust CSG to compute a social-welfare optimal ε-NE, using a robust MDP-based exploration mechanism to drive joint state–action coverage. Crucially, we introduce a Nash margin characterisation that enables principled reasoning about equilibrium existence: the framework either returns an ε-approximate NE whose social-welfare value is ε-close to optimal, or provides a sound certificate that no exact NE exists. Under a minimum reachability condition $p_{\mathrm{reach}} > 0$ over relevant state–action pairs, the algorithm terminates after a polynomial number of trajectory samples, with sample complexity $\widetilde{\mathcal{O}}\big(R_{\max}^2 H^4 |S|^2 |A| / (p_{\mathrm{reach}} \varepsilon^2)\big)$. Empirical results on benchmark CSGs demonstrate near-optimal performance, correct handling of equilibrium (non-)existence, and sample complexity consistent with theory.

1 Introduction

Learning to act optimally in multi-agent stochastic environments is fundamentally harder when the transition dynamics are unknown: even small estimation errors can destabilise equilibria or make a nominally near-optimal strategy far from equilibrium under the true model. Concurrent stochastic games (CSGs) [39], where players choose actions simultaneously without observing one another, are the canonical model for such settings — yet most existing approaches to equilibrium computation assume that transition probabilities are precisely known. This paper asks: *how many environment interactions suffice to compute an approximately optimal equilibrium with high probability, and what happens when no exact equilibrium exists?*

Uncertainty in stochastic decision-making has been studied from two complementary angles. On the robust-planning side, robust MDPs (RMDPs) [21, 35, 48] including interval MDPs [17], and their stochastic-game extensions [4, 9] optimise worst-case performance over a prescribed uncertainty set. Recent work has extended this line to robust CSGs (RCSGs) [18] and distributionally robust Markov games (DRMGs) [30, 38, 40, 51]. On the learning side, *Probably Approximately Correct* (PAC) [44] algorithms provide sample-complexity guarantees for MDPs [2, 7, 22, 24, 42] and turn-based stochastic games [3, 32, 50]. However, these two lines of work have remained largely separate in the concurrent general-sum setting: robust methods typically assume availability of the uncertainty set or a generative model [21, 30, 35, 40, 51], while PAC methods do not address robust equilibrium computation or equilibrium existence under unknown dynamics. Work at their intersection on *robust PAC learning* remains largely restricted to the single-agent setting [36, 37, 46, 49, 52].

Combining robust equilibrium computation with PAC learning in concurrent games poses two challenges absent from prior work: **1) Equilibrium existence under uncertainty.** While ε-NE exist for any $\varepsilon > 0$ [13], exact stationary (i.e., memoryless and time-independent) NE may fail to exist in CSGs [6]. The issue is not merely pathological: in multi-agent systems, an NE represents a stable

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

operating point in which no agent has unilateral incentive to deviate. When no such equilibrium exists, any deployed strategy may be inherently unstable, potentially inducing unpredictable or undesirable system behaviour, as observed in domains such as traffic coordination, network routing, and competitive markets [19, 23]. Moreover, even when an exact NE exists in the true game, an uncertainty set around the true transition kernel may not have a *robust* equilibrium [18]. A PAC framework for CSGs should therefore support both approximate equilibrium computation and principled handling of non-existence. **2) Global coverage from stochastic trajectories.** Obtaining sufficient joint state–action coverage from online trajectories is challenging in our setting because no individual player controls the joint action and the transition kernel is uncertain. Under centralised exploration, we address this by coordinating joint actions and constructing an auxiliary *exploration RMDP* on which we compute joint profiles to robustly reach under-visited state–action pairs. This yields a per-episode lower bound on visiting insufficiently explored pairs in the true game. Because the set of under-visited pairs and the exploration policy adapt to the observed history, these episode-level events are dependent; we therefore use a martingale concentration argument based on Freedman's inequality [16] to aggregate the conditional lower bound into a high-probability global coverage and sample-complexity guarantee.

Contributions. We introduce *PAC-CSG*, the first PAC-learning framework for two-player general-sum CSGs under transition uncertainty with *sound* detection of stationary equilibrium non-existence. When an exact NE exists, the algorithm returns an ε-NE (i.e., stable up to ε under unilateral deviations) whose social-welfare value is ε-close to optimal; otherwise, it may return a sound certificate of non-existence. We consider finite- and infinite-horizon probabilistic and reward reachability objectives, which are standard in verification and planning [27, 28], unlike the discounted reward objectives common in reinforcement learning (RL) [2, 42]. To realise this framework, we design an algorithm that maintains L^1 confidence sets over transition kernels and combines robust equilibrium computation with an auxiliary RMDP-based exploration mechanism that formulates joint state–action coverage as a robust reachability problem. Under a graph reachability condition, the algorithm terminates after $\widetilde{\mathcal{O}}\big(R_{\max}^2 H^4 |S|^2 |A| / (p_{\mathrm{reach}} \varepsilon^2)\big)$ trajectory samples and outputs either an ε-optimal ε-NE or a sound non-existence certificate, where p_{reach} is the minimum reachability probability over all relevant state–action pairs. Finally, we evaluate the approach on six CSG benchmarks, demonstrating near-optimality, correct handling of equilibrium (non-)existence, and sample complexity scaling consistent with the theoretical bounds.

Related work. *Robust planning* in stochastic games addresses decision-making under a prescribed model uncertainty set. In the single-agent setting, this includes robust MDPs [21, 35, 48] and interval MDPs [17], which optimise worst-case performance over admissible transition kernels. Multi-agent extensions are more recent: robust turn-based stochastic games [9] and qualitative objectives [4] have been studied, and He and Parker [18]'s robust CSG framework introduces robust equilibrium notions for general-sum concurrent games. These approaches assume that the uncertainty set is specified a priori and focus on computing robust policies or equilibria, rather than learning them from data.

PAC learning methods are well developed for MDPs via E^3 [24], R-MAX [7], and related PAC-MDP analyses [42]. Closely related exploration guarantees include regret-minimisation methods such as UCRL2 [22] and UCBVI [2]. These results have been extended to turn-based stochastic games in both zero-sum self-play [3, 32] and general-sum settings [50]. However, they do not address the concurrent setting, where no single agent controls state–action visitation, nor do they incorporate robustness to transition uncertainty or equilibrium existence considerations.

Robust RL considers learning under model uncertainty from data. In the single-agent setting, both model-based and model-free approaches provide sample complexity guarantees [36, 37, 46, 49, 52]. In multi-agent settings, recent work studies distributionally robust Markov games (DRMGs), primarily in zero-sum settings [30, 38, 40, 51], including sample-efficient algorithms under generative-model assumptions that largely remove the need for exploration. Other work considers alternative solution concepts, e.g., robust correlated equilibria [33]. Closest to our setting is Farhat et al. [15], who study online robust multi-agent RL under similar uncertainty sets. However, they optimise regret rather than providing PAC guarantees, and do not address equilibrium existence in general-sum concurrent games. In contrast, we provide a high-confidence output guarantee: either an ε-NE when one exists, or a sound certificate of non-existence. This distinction is fundamental in our concurrent general-sum setting: PAC learning targets the quality of a single output equilibrium, whereas regret minimisation controls performance over the entire learning trajectory.

2 Preliminaries

We denote by $\mathcal{D}(X)$ the set of discrete probability distributions over a finite set X, and let $\mathbb{1}[A]$ be the indicator function that equals 1 if A holds and 0 otherwise.

Definition 1 (Concurrent stochastic game)**.** *A (reward-augmented)* concurrent stochastic game *(CSG) is a tuple $\mathcal{G} = (N, S, \bar{s}, A, \Gamma, P, r)$ where: $N = \{1, \dots, n\}$ is a finite set of players; S is a finite set of states with initial state $\bar{s} \in S$; $A = \bigtimes_{i \in N}(A_i \cup \{\bot\})$ is the set of joint actions, where each A_i is a finite action set and $\bot$ is a distinguished idle action; $\Gamma : S \to 2^{\cup_i A_i}$ assigns available actions; $P : S \times A \rightharpoonup \mathcal{D}(S)$ is a transition kernel; and $r = (r_1, \dots, r_n)$ with rewards $r_i : S \times A \to \mathbb{R}$.*

At each state s, each player $i \in N$ simultaneously selects an action $a_i \in A_i(s)$, where $A_i(s) = \Gamma(s) \cap A_i$ if $\Gamma(s) \cap A_i \neq \emptyset$ and $A_i(s) = \{\bot\}$ otherwise. The joint action $a = (a_1, \dots, a_n) \in A(s) := \bigtimes_{i \in N} A_i(s)$ induces a transition to state s' according to $P_{sa} := P(s, a)$. Henceforth let $\mathcal{X} := \{(s, a) : s \in S, a \in A(s)\}$ denote the set of all admissible state–action pairs (or "slots").

A *strategy* for player i is a function σ_i mapping finite histories to distributions over actions. A *stationary* (or *memoryless*) strategy depends only on the current state, i.e., $\sigma_i(\pi) = \sigma_i(s)$ for all histories π ending in state s. We focus on this class of strategies throughout this work. A *strategy profile* (or just *profile*) is a tuple of strategies, one for each player, denoted $\sigma = (\sigma_1, \dots, \sigma_n) \in \Sigma := \bigtimes_{i \in N} \Sigma_i$. For a profile σ and a unilateral deviation $\sigma'_i \in \Sigma_i$, we write $\sigma_{-i}[\sigma'_i]$ for the profile obtained by replacing player i's strategy σ_i with σ'_i.

Each player i is associated with an objective X_i mapping infinite paths to $\mathbb{R}$. For a profile σ, kernel P, and state s, we define players' total expected utility, i.e., the *social welfare*, as $u(\sigma, P \mid s) := \sum_{i \in N} u_i(\sigma, P \mid s)$, where $u_i(\sigma, P \mid s) := \mathbb{E}_s^{\sigma, P}[X_i]$ is player i's expected utility. For the value in the initial state $\bar{s}$, we write $u(\sigma, P) := u(\sigma, P \mid \bar{s})$.

Robust CSGs (RCSGs) [18] add transition uncertainty to CSGs. Formally, an RCSG is a tuple $\mathcal{G} = (N, S, \bar{s}, A, \Gamma, \mathcal{P}, r)$, where $\mathcal{P} : S \times A \rightharpoonup 2^{\mathcal{D}(S)}$ is an uncertain transition kernel, and all other components are as in Definition 1. Fixing a $P \in \mathcal{P}$ in an RCSG induces a CSG, which we denote $\mathcal{G}_P$. We focus on (s, a)-rectangular uncertainty, where the global uncertainty set factorises as $\mathcal{P} = \prod_{(s,a) \in \mathcal{X}} \mathcal{P}(s, a)$, so transition uncertainty at each slot can be resolved independently. For a profile σ, we denote its *robust* and *optimistic* values over $\mathcal{P}$, respectively, as

$$\underline{u}_{\mathcal{P}}(\sigma) := \inf_{P \in \mathcal{P}} u(\sigma, P), \qquad \bar{u}_{\mathcal{P}}(\sigma) := \sup_{P \in \mathcal{P}} u(\sigma, P).$$

Objectives. Following He and Parker [18], Kwiatkowska et al. [27], we focus on four objectives (two finite-horizon, two infinite-horizon). Let $S_T \subseteq S$ be a set of target states, and $\mathcal{X}_{\text{nt}} := \{(s, a) \in \mathcal{X} : s \notin S_T\}$ the set of non-target slots. For finite-horizon objectives fix $H \in \mathbb{N}$: (1) *Bounded probabilistic reachability*: $X(\pi) = \mathbb{1}[\exists j \leq H.\ \pi(j) \in S_T]$; (2) *Bounded cumulative reward*: $X(\pi) = \sum_{i=0}^{H-1} r(\pi(i), \pi[i])$; (3) *Probabilistic reachability*: $X(\pi) = \mathbb{1}[\exists j \in \mathbb{N}.\ \pi(j) \in S_T]$; and (4) *Reachability reward (or stochastic shortest paths)*: $X(\pi) = \sum_{i=0}^{\tau_T - 1} r(\pi(i), \pi[i])$ if $\exists j \in \mathbb{N}.\ \pi(j) \in S_T$ and $X(\pi) = \infty$ otherwise, where $\tau_T = \min\{j \in \mathbb{N} \mid \pi(j) \in S_T\}$.

For the reachability objectives, transitions after entering S_T do not affect the objective, so we only need to resolve uncertainty over $\mathcal{X}_{\text{nt}}$. Thus, henceforth we consider $\mathcal{P} = \prod_{(s,a) \in \mathcal{X}_{\text{nt}}} \mathcal{P}(s, a)$. For the bounded cumulative reward objective, $S_T = \emptyset$ so $\mathcal{X}_{\text{nt}} = \mathcal{X}$.

Solution concepts. A *Nash equilibrium* (NE) in a CSG is a profile σ such that, for all players i, states s and deviations σ'_i, $u_i(\sigma, P \mid s) \geq u_i(\sigma_{-i}[\sigma'_i], P \mid s)$. A *social-welfare optimal NE* (SWNE) is an NE maximising $u(\sigma, P \mid s)$. We require these equilibrium conditions to hold from every state, yielding a *subgame-perfect* equilibrium.

For an RCSG, we further require the same equilibrium requirement to hold robustly, i.e., across every $P \in \mathcal{P}$. In addition, since exact stationary NE may not exist [6] but ε-NE exists for any $\varepsilon > 0$, we use *approximate* robust equilibria:

Definition 2 (ε-RNE)**.** *A profile σ is a* robust ε-Nash equilibrium *(ε-RNE) iff for all players i, states s, and deviations σ'_i, $\inf_{P \in \mathcal{P}} [u_i(\sigma, P \mid s) - u_i(\sigma_{-i}[\sigma'_i], P \mid s)] \geq -\varepsilon$. The corresponding ε-RNE value is defined as $\underline{u}_{\mathcal{P}}(\sigma) := \inf_{P \in \mathcal{P}} u(\sigma, P)$.*

The case $\varepsilon = 0$ yields a *robust NE* (RNE). Following this definition, a *robust social-welfare optimal ε-NE* (ε-RSWNE) is just an ε-RNE σ that maximises $\underline{u}_{\mathcal{P}}(\sigma)$, the robust total utility of the players. Henceforth we write $\Sigma_{\varepsilon\text{-RNE}}(\mathcal{P})$ for the set of all ε-RNE under $\mathcal{P}$, $\Sigma_{\text{RNE}}(\mathcal{P})$ for the set of RNE, and similarly $\Sigma_{\varepsilon\text{-NE}}(P)$ and $\Sigma_{\text{NE}}(P)$ for the set of ε-NE and NE, respectively, under kernel P.

3 PAC-CSG

Game setting. We model the true environment as a finite two-player general-sum CSG $\mathcal{G}^\star = (N = \{1, 2\}, S, \bar{s}, A, \Gamma, P^\star, r)$, where $P^\star$ is unknown but the state space, action availability, rewards, and transition support of $\mathcal{G}^\star$ are known. When it exists, denote the SWNE value of $\mathcal{G}^\star$ by $V^\star := \sup_{\sigma \in \Sigma_{\text{NE}}(P^\star)} u(\sigma, P^\star) = u(\sigma^\star, P^\star)$, with $\sigma^\star$ being the corresponding profile. We assume bounded rewards: $|r_i(s, a)| \leq R_{\max}$. We also enforce the standard *graph preservation* constraint [10, 34], requiring all $P \in \mathcal{P}$ to share the same support. This ensures the tractability of robust dynamic programming [21, 35] over (s, a)-rectangular uncertainty models.

Learning. Throughout, we assume a *centralised* learner that jointly controls both players. Learning proceeds episodically: quantities derived from data up to episode t are indexed by t, and T denotes the total number of episodes before termination. At episode t, the learner maintains an uncertainty set $\mathcal{P}_t$ around an empirical kernel $\widehat{P}_t$. This set is defined via per-slot L^1 confidence balls: for each $(s, a) \in \mathcal{X}_{\text{nt}}$, $\|P_{sa} - \widehat{P}_{sa,t}\|_1 \leq \alpha_t(s, a) := \sqrt{\frac{2}{n} \ln\left((2^{|S|} - 2)/\delta^{\text{cont}}_{sa,n}\right)}$, where $n := N_t(s, a)$ is the visit count of (s, a) up to episode t and $\delta^{\text{cont}}_{sa,n}$ is a suitable failure probability defined in Corollary 10, App. B. The radius $\alpha_t(s, a)$ follows from Weissman's L^1 concentration inequality [47] and is derived in Lemma 9, App. B. $\mathcal{P}_t$ contains all transition kernels consistent with the observed data so far with high probability, and induces an empirical L^1-CSG $\mathcal{G}_t = (N, S, \bar{s}, A, \Gamma, \mathcal{P}_t, r)$. Further, the learner selects an exploration profile σ^e_t by solving an exploration RMDP $\mathcal{G}^e_t$ (see Definition 5) induced by the same uncertainty set $\mathcal{P}_t$, executes it in $\mathcal{G}^\star$, and uses the resulting trajectories to update $\mathcal{G}_t$.

3.1 PAC-CSG

Unlike single-agent or zero-sum settings, exact stationary NE need not exist in our setting.

Remark 1 (NE existence)**.** *For any finite CSG, $\Sigma_{\varepsilon\text{-NE}}(P) \neq \emptyset$ for all $\varepsilon > 0$ [6, 13]. However, exact stationary NE ($\varepsilon = 0$) may fail to exist in CSGs. For finite-horizon CSGs, backward induction generally yields time-dependent equilibrium strategies. In the infinite-horizon setting, Bouyer et al. [6] illustrate non-existence via two-player concurrent terminal-reward games with no NE at all (and hence no stationary NE), including the Hide-or-run game used in our experiments; see Sec. 5. Finally, for an RCSG $\mathcal{G}$, even if every induced CSG $\mathcal{G}_P$ admits an (ε-)NE, an (ε-)RNE need not exist in $\mathcal{G}$ [18], as it must satisfy equilibrium conditions simultaneously for all admissible kernels P.*

To reason about equilibrium existence quantitatively, we introduce the notion of *Nash margin*.

Definition 3 (Nash margin)**.** *For a CSG with kernel P and profile space Σ, define the* Nash margin *of a profile $\sigma \in \Sigma$ as $\mu(\sigma, P) := \min_{i \in N} \min_{s \in S} \inf_{\sigma'_i \in \Sigma_i} [u_i(\sigma, P \mid s) - u_i(\sigma_{-i}[\sigma'_i], P \mid s)]$. Further define the* global Nash margin $\bar{\mu}$ *in $\mathcal{G}^\star$ as $\bar{\mu} := \max_{\sigma \in \Sigma} \mu(\sigma, P^\star)$; and, if $\bar{\mu} \geq 0$, define the* social-welfare optimal Nash margin *as $\mu^\star := \mu(\sigma^\star, P^\star)$.*

Thus, $\mu(\sigma, P)$ measures how close a profile is to being an NE, while $\bar{\mu}$ captures how close $\mathcal{G}^\star$ is to admitting an equilibrium overall. In particular, $\mu(\sigma, P) \geq 0$ iff $\sigma \in \Sigma_{\text{NE}}(P)$, and $\bar{\mu} \geq 0$ iff $\Sigma_{\text{NE}}(P^\star) \neq \emptyset$. Since an ε-NE exists in any finite CSG for all $\varepsilon > 0$, the global margin satisfies $\bar{\mu} \geq -\varepsilon$ for arbitrarily small ε. The use of $\max$ (rather than $\sup$) in the definition of $\bar{\mu}$ ensures that this margin is well-defined only when the maximum is attained by some profile. Finally, $V^\star$ is well-defined iff $\mu^\star \geq 0$, which holds precisely when $\bar{\mu} \geq 0$, i.e., $\mathcal{G}^\star$ admits an exact NE. We formalise this relationship in Proposition 8, App. A.

This characterises our PAC formulation in Definition 4, which conditions on whether $\bar{\mu} \geq 0$. We present a concrete realisation of this framework in Alg. 1.

Definition 4 (PAC-CSG)**.** *An algorithm for a finite CSG is* PAC *if, for any $\varepsilon, \delta \in (0, 1)$, with probability at least $1 - \delta$, after a number of interactions polynomial in $|S|, |A|, 1/\varepsilon, \log(1/\delta)$ and problem constants, it outputs one of the following:*

(a) (Approximate equilibrium) *An ε-NE profile $\hat{\sigma}$ such that, if $\bar{\mu} \geq 0$, then $V^\star - u(\hat{\sigma}, P^\star) \leq \varepsilon$, i.e., $\hat{\sigma}$ is ε-optimal; in this case, the algorithm always terminates in this branch.*

(b) (Near-non-existence certificate) *A sound declaration that $\bar{\mu}$ is not well-defined or $\bar{\mu} < 0$, i.e., no exact NE exists in $\mathcal{G}^\star$.*

Importantly, here ε-NE and ε-optimality are distinct requirements: the former bounds unilateral incentives to deviate, while the latter bounds the social-welfare loss relative to the optimal NE.

The existence branch (a) certifies that the true game admits a well-separated (SW)NE and returns a corresponding profile, while the non-existence branch (b) certifies that no exact NE exists based on the (well-definedness) and value of $\bar{\mu}$. Branch (b), however, provides a *sound* but *incomplete* certificate: if the algorithm outputs $\bar{\mu} < 0$, then $\mathcal{G}^\star$ admits no exact NE. The converse need not hold: even when $\bar{\mu} < 0$, the algorithm may still terminate via branch (a), since ε-NE always exist. Intuitively, non-existence is detected only with sufficiently strong evidence (i.e., enough samples). We develop this condition in the next section.

Algorithm 1 PAC learning of 2-player CSGs

Input: $\delta, \varepsilon, H, R_{\max}, (S, A, \Gamma, r), \mathrm{Supp}(P^\star)$

$\delta^{\mathrm{cont}}, \delta^{\mathrm{cov}} \leftarrow \delta/2$; $t \leftarrow 1$

$N, \mathcal{K}, \widehat{P}, n_{\min}, p_{\mathrm{reach}} \leftarrow$ INITIALISE$(S, A, \Gamma, R_{\max}, H, \mathrm{Supp}(P^\star))$ ▷ see Alg. 2, App. J

while true **do**

  $\mathcal{P} \leftarrow$ CONSTRUCTUNCERTAINKERNEL$(N, \delta^{\mathrm{cont}}, S, A, \Gamma, \mathrm{Supp}(P^\star))$ ▷ see Alg. 3, App. J

  $\mathcal{G} \leftarrow$ BUILDL^1CSG$(S, A, \Gamma, \mathcal{P}, r)$

  $\Delta \leftarrow \frac{1}{2} R_{\max} H^2 \max_{(s,a) \in \mathcal{X}_{\mathrm{nt}}} \sup_{p \in \mathcal{P}(s,a)} \|p - \widehat{P}_{sa}\|_1$

  if $\Delta \leq \varepsilon/4$ **then**

    $\hat{\sigma},\ \hat{V},\ \mathrm{found} \leftarrow$ SOLVEL^1CSG$(\mathcal{G}, \varepsilon)$

    if found $=$ false **then return** "NOT_FOUND: $\bar{\mu}$ undefined or $\bar{\mu} < 0$" **else return** $\hat{\sigma},\ \hat{V}$

    end if

  end if

  $\mathcal{G}^e \leftarrow$ BUILDEXPLORATIONRMDP$(S, A, \Gamma, \mathcal{P}, r, \mathcal{K})$ ▷ following Def. 5

  $\sigma^e \leftarrow$ SOLVEL^1MDP$(\mathcal{G}^e)$

  $\delta_t^{\mathrm{cov}} \leftarrow \delta^{\mathrm{cov}}/(t(t+1))$; $N^\pi \leftarrow \lceil -(1/p_{\mathrm{reach}}) \log \delta_t^{\mathrm{cov}} \rceil$

  $\pi \leftarrow$ SAMPLETRAJECTORIES(σ^e, H, N^π) ▷ execute σ^e in $\mathcal{G}^\star$

  $N, \mathcal{K} \leftarrow$ UPDATE$(N, \mathcal{K}, \pi, n_{\min}, S, A, \Gamma)$; $t \leftarrow t+1$ ▷ see Alg. 4, App. J

end while

4 PAC-CSG guarantees

We begin with finite-horizon objectives and extend the analysis to the infinite-horizon setting in Sec. 4.3. Our approach follows a three-step pipeline: 1) In Sec. 4.1, we relate statistical uncertainty in the transition kernel to value estimation error via a sensitivity lemma (Lemma 2), showing that solving a robust equilibrium in $\mathcal{P}_t$ yields an approximate NE in the true game once $\Delta_t \leq \varepsilon/4$. 2) In Sec. 4.2, we construct an exploration RMDP and show that, under Assumption 1, every relevant (s, a)-pair is reachable, ensuring that targeted exploration drives Δ_t below the required threshold. 3) Finally, in Sec. 4.3, we combine these results and use a Freedman martingale argument to lift stochastic per-episode visitation into a high-probability bound on the number of episodes, yielding the overall sample complexity. Throughout, we condition on the high-probability event $\mathcal{E}^{\mathrm{cont}}$ that $P^\star \in \mathcal{P}_t$ for all episodes $t \geq 1$, as established in Corollary 10, App. B.

4.1 Estimation Error and Value Stability

Our analysis begins by relating uncertainty in the transition kernel to errors in value estimates. Suppose $\mathcal{E}^{\mathrm{cont}}$ holds, i.e., $P^\star \in \mathcal{P}_t$. We show that this uncertainty induces a controlled deviation in the value of any fixed strategy profile.[1]

[1] In Alg. 1, Δ_t is computed conservatively by maximising the transition error over the confidence set $\mathcal{P}_t$; on $\mathcal{E}^{\mathrm{cont}}$, this computed quantity upper-bounds the true Δ_t, and hence subsequent claims on Δ_t apply to the computed quantity too.

Lemma 2 (Sensitivity). *Define* $\Delta_t := \frac{1}{2} R_{\max} H^2 \Delta_t^{L^1}$, *where* $\Delta_t^{L^1} := \max_{(s,a)\in\mathcal{X}_{\mathrm{nt}}} \|P^\star_{sa} - \widehat{P}_{sa}\|_1$. *Then for any* $P \in \mathcal{P}_t$, *profile* σ, *player* $i \in N$, *and state* $s_0 \in S$, $|u_i(\sigma, P \mid s_0) - u_i(\sigma, \widehat{P} \mid s_0)| \le \Delta_t$.

Proof sketch. We track how a per-(s,a) L^1 perturbation of size $\Delta_t^{L^1}$ propagates through the trajectory distribution. At step h, the deviation between state marginals satisfies $\|\mu_h - \hat{\mu}_h\|_1 \le h\,\Delta_t^{L^1}$ (Proposition 11, App. B). Summing the induced reward difference $R_{\max}\|\mu_h - \hat{\mu}_h\|_1$ over $h = 0, \dots, H-1$ and using $\sum_{h=0}^{H-1} h \le H^2/2$ yields the bound. Full proof in App. C. □

Thus, on $\mathcal{E}^{\mathrm{cont}}$, for any fixed profile $\hat{\sigma}$, its value in the true game $\mathcal{G}^\star$ is close to its robust value over $\mathcal{P}_t$, up to an error of order Δ_t, which vanishes as the confidence radius shrinks with additional data. The following lemma further lets us reason about NE existence in $\mathcal{G}^\star$ by solving for a robust equilibrium over $\mathcal{P}_t$ (using a black-box RCSG solver; see Assumption 2, App. D): it shows that any true NE induces an approximate RNE in $\mathcal{P}_t$, with quality degrading proportionally in Δ_t. Consequently, solver failure to find an ε-RNE implies that no NE in $\mathcal{G}^\star$ has sufficiently large margin, i.e., $\ge 4\Delta_t - \varepsilon$.

Lemma 3 (NE($P^\star$) $\Rightarrow$ approximate RNE($\mathcal{P}_t$)). *If* σ *is an (exact) NE under* $P^\star$ *achieving Nash margin* $\mu := \mu(\sigma, P^\star) \ge 0$, *then it is also an* $(4\Delta_t - \mu)$*-RNE under* $\mathcal{P}_t$.

Proof sketch. For any player $i \in \{1, 2\}$ and unilateral deviation σ_i', Lemma 2 bounds the difference in deviation gains under $P^\star$ and any $P \in \mathcal{P}_t$ by $4\Delta_t$. Since σ has Nash margin μ under $P^\star$, its deviation gain there is at most $-\mu$. Thus, under every $P \in \mathcal{P}_t$, the deviation gain is at most $4\Delta_t - \mu$, implying that σ is an $(4\Delta_t - \mu)$-RNE. Full proof in App. D. □

Proposition 4 (Solver guarantees for equilibrium computation). *On the event* $\mathcal{E}^{cont}$*:*

*(1) (*Equilibrium transfer*) If the solver returns an* ε*-RSWNE* $\hat{\sigma}$ *of* $\mathcal{P}_t$, *then* $\hat{\sigma}$ *is an* ε*-NE of* $\mathcal{G}^\star$.

*(2) (*Failure implies near non-existence*) If the solver returns* `NOT_FOUND`, *then either* $\mu^\star$ *is undefined or* $\mu^\star < 4\Delta_t - \varepsilon$.

Proof sketch. (1) Any ε-RSWNE under $\mathcal{P}_t$ is, by definition, an ε-RNE under $\mathcal{P}_t$, and hence an ε-NE for every $P \in \mathcal{P}_t$, including $P^\star$ (on $\mathcal{E}^{\mathrm{cont}}$). (2) If the solver returns `NOT_FOUND`, then by Assumption 2 no ε-RNE exists under $\mathcal{P}_t$. We then distinguish two cases. If no NE exists under $P^\star$, then the SWNE does not exist and $\mu^\star$ is undefined. Otherwise, the SWNE exists, and if $\mu^\star \ge 4\Delta_t - \varepsilon$, the NE–RNE transfer lemma (Lemma 3) would imply that the SWNE is an ε-RNE under $\mathcal{P}_t$, contradicting the solver's failure. Full proof in App. D. □

Together, these results link equilibrium computation in the empirical model to the true game: when $\mu^\star \ge 4\Delta_t - \varepsilon$, the solver returns an approximate NE; when $\mu^\star < 4\Delta_t - \varepsilon$, solver failure yields a sound certificate of non-existence once uncertainty is sufficiently small. Combined with the value stability results, this yields the following per-episode guarantee.

Corollary 5 (Per-episode value gap). *If* $\mu^\star \ge 4\Delta_t - \varepsilon$, *then* $V^\star - u(\hat{\sigma}_t, P^\star) \le 4\Delta_t$. [Proof in App. E]

It therefore remains to ensure that Δ_t falls below $\varepsilon/4$ in a finite number of samples. Since Δ_t scales with the maximum per-(s,a) confidence radius, this requires sufficient coverage of all relevant slots. We tackle this using a targeted, RMDP-based exploration scheme presented in the next section.

4.2 Exploration

A key challenge in PAC learning of concurrent games is obtaining sufficient joint state–action coverage from stochastic trajectories under transition uncertainty. To isolate this statistical challenge, we assume centralised exploration, allowing the learner to coordinate the players' joint actions. To this end, we construct an auxiliary *exploration RMDP* that casts coverage of under-visited (s,a)-pairs as a robust reachability problem. For each (s,a), we maintain a boolean flag indicating whether the slot is *known*, i.e., whether it has been visited sufficiently many times; this threshold is precisely defined in Lemma 20, App. F. Let $\mathcal{U}_t$ and $\mathcal{K}_t$ denote the sets of unknown and known (non-target) slots at the beginning of episode t, respectively (so $\mathcal{U}_1 = \mathcal{X}_{\mathrm{nt}}$; see formal definition in App. F).

Exploration RMDP. At each episode t, we construct an *exploration RMDP* $\mathcal{G}_t^e$ that coincides with the empirical RCSG $\mathcal{G}_t$ on all known slots. On each unknown slot $(s,a) \in \mathcal{U}_t$, the model instead transitions deterministically to an auxiliary absorbing state z, with a nonzero reward accrued upon this transition. Note that this absorbing construction is used for analysis only (see App. F). The learner's objective is to maximise expected reward, which is equivalent to maximising the probability of reaching unknown slots under this construction.

Definition 5 (Exploration RMDP)**.** *The* exploration RMDP *at episode t is an L^1-MDP $\mathcal{G}_t^e = (S^e, s, A^e, \mathcal{P}_t^e, r_t^e)$ where $S^e := S \cup \{z\}$ and $A^e := A_1 \times A_2$. For each $(s,a) \in S^e \times A^e$: if $(s,a) \in \mathcal{U}_t \vee s = z \vee s \in S_T$ then $\mathcal{P}_t^e(s,a) = \{\delta_z\}$, where δ_z is the Dirac distribution that assigns probability 1 to z and 0 to all others; else $\mathcal{P}_t^e(s,a) = \mathcal{P}_t(s,a)$; r_t^e is such that $r_t^e(s,a) := \mathbb{1}[(s,a) \in \mathcal{U}_t]$.*

Next, let $p_t^e(P) := \mathbb{P}^{\sigma_t^e, P}[\text{reach } \mathcal{U}_t]$, i.e., the probability of reaching any unknown (s,a)-pair during episode t. Define the exploration profile as

$$\sigma_t^e \in \arg\max_{\sigma \in \Sigma} \underline{u}_{\mathcal{P}_t^e}(\sigma) = \arg\max_{\sigma \in \Sigma} \inf_{P \in \mathcal{P}_t^e} p_t^e(P). \tag{1}$$

This pessimistic (robust) objective is essential for the high-probability global coverage guarantee (Lemma 19, App. G) used to prove Theorem 7: a best-case visitation probability under $P \in \mathcal{P}_t$ need not lower-bound the visitation probability under the true kernel $P^\star$.

Assumption 1 (Graph reachability)**.** *Let $\mathcal{P}_{\text{Supp}} := \{P : \text{Supp}(P) = \text{Supp}(P^\star)\}$. The game graph of $\mathcal{G}^\star$ is such that every $(s,a) \in \mathcal{X}_{\text{nt}}$ is reachable from $\bar{s}$ under some profile, i.e.,*

$$p_{\text{reach}} := \min_{(s,a) \in \mathcal{X}_{\text{nt}}} \max_{\sigma \in \Sigma} \inf_{P \in \mathcal{P}_{\text{Supp}}} \mathbb{P}^{\sigma, P}[\text{reach } (s,a) \text{ before entering } S_T] > 0. \tag{2}$$

Reachability conditions like Assumption 1 are standard in prior work [1, 7, 22]. Since the transition support of $P^\star$ is known as input to Alg. 1, p_{reach} can be precomputed by solving a corresponding reachability RMDP over the known game graph (see App. J for details). Together with the exploration RMDP construction, this yields the following result; the full proof is in App. F.

Lemma 6 (Exploration probability)**.** *For all episodes $t \geq 1$ with $\mathcal{U}_t \neq \emptyset$, $p_t^e(P^\star) \geq p_{\text{reach}}$.*

We execute σ_t^e in $\mathcal{G}^\star$ to sample trajectories. By collecting multiple samples per episode (with the number specified in Proposition 17, App. F), each episode visits an unknown slot with nontrivial probability. Repeating this across episodes yields high-probability coverage of all relevant slots, ensuring that each is visited sufficiently often. The precise sample and episode counts are derived in Lemma 19 and Corollary 21 (App. F), and are used to establish the sample complexity bound in the next section.

4.3 PAC-CSG

We now combine the previous results to obtain the PAC guarantee, in line with Definition 4.

Theorem 7 (PAC guarantee for finite-horizon objectives)**.** *Set $n_{\min} = \widetilde{\mathcal{O}}\big(R_{\max}^2 H^4 |S| / \varepsilon^2\big)$. Then with probability at least $1 - \delta$, Alg. 1 terminates after at most*

$$\widetilde{\mathcal{O}}\left(\frac{R_{\max}^2 H^4 |S|^2 |A|}{\varepsilon^2 p_{\text{reach}}}\right) \tag{3}$$

trajectory samples, where $\widetilde{\mathcal{O}}(\cdot)$ suppresses logarithmic factors in the problem parameters. Upon termination, the algorithm's output satisfies the PAC-CSG guarantee of Definition 4.

Proof sketch. We split the failure budget as $\delta^{\text{cont}} = \delta^{\text{cov}} = \delta/2$ and condition on $\mathcal{E} := \mathcal{E}^{\text{cont}} \cap \mathcal{E}^{\text{cov}}$, which holds with probability $\geq 1 - \delta$. On $\mathcal{E}^{\text{cont}}$, Corollary 5 reduces near-optimality to ensuring $\Delta_t \leq \varepsilon/4$, which holds once every $(s,a) \in \mathcal{X}_{\text{nt}}$ is visited at least $n_{\min} = \widetilde{\mathcal{O}}\big(R_{\max}^2 H^4 |S| / \varepsilon^2\big)$ times (Lemma 20). Under Assumption 1, Lemma 19 shows that this coverage is achieved after $\widetilde{\mathcal{O}}(n_{\min} |S||A|)$ episodes: for each episode, let X_t indicate whether at least one sampled trajectory reaches an unknown slot. Since these indicators are history-dependent, we apply Freedman's inequality to their martingale differences, using the lower bound p_{reach} from Lemma 6, to obtain a high-probability bound on the number of episodes needed until all relevant slots are sufficiently visited. Moreover, Proposition 17 yields $\widetilde{\mathcal{O}}(\log t / p_{\text{reach}})$ trajectories per episode t, giving the stated sample complexity. Finally, Proposition 4 yields the output branching. Full details are in App. G. □

Infinite-horizon extension. We further assume almost-sure reachability of S_T, i.e., S_T is reached with probability 1 from every state under every $P \in \mathcal{P}_{\mathrm{Supp}}$ by some profile (Assumption 3), following prior work [5, 18, 27]. This ensures value iteration convergence but not the existence of stationary NE. Under this assumption, the process admits an *effective horizon* H_{eff} which upper bounds the stopping time, with $H_{\mathrm{eff}} \leq |S|/p_T$ (see Lemma 24). Here, p_T is the minimum worst-case probability of reaching S_T within $|S|$ steps from any state. Replacing H with H_{eff} yields the corresponding infinite-horizon guarantees, with constants adjusted accordingly. Full details are in App. H.

5 Experimentation

We evaluate our method on a suite of benchmark CSGs to validate our theoretical guarantees.

Experimental setup. We implement Alg. 1 in Java and evaluate it in the PRISM-games model checker [26]. Unless otherwise stated, we set $\delta = 0.05$, $\varepsilon = 0.2$ and $R_{\max} = 1$. For RQ1, we instead use $\varepsilon = 0.1$. Results are averaged over 10 random seeds for trajectory sampling. For the learning-dynamics experiments in Appendix K.2, we additionally evaluate the empirical model and strategy against the known ground-truth CSG; these quantities are used only for evaluation and are not available to the learner. Experiments ran on an Apple M4 Pro with 24 GB RAM. Peak memory usage never exceeded 0.5 GB; the benchmarks are compute-bound rather than memory-bound, with the largest configuration requiring approximately 5 hours per seed. Our implementation, including the models and property specifications, is available at `https://github.com/7angel4/prism-games/tree/csg-learning`. We compare our learners' output against the PRISM-games CSG solver as ground truth.

Benchmarks. We evaluate on six small, interpretable CSGs, chosen to enable repeated exact NE computation and comparison against exact solutions. These are illustrated in Fig. 2, App. K. Each benchmark is paired with a set of finite- and infinite-horizon properties expressed in *probabilistic alternating-time temporal logic with rewards* (rPATL) [11, 27], designed to isolate a distinct challenge: 1) *Cyclic Preferences* has no exact stationary NE for infinite-horizon reachability reward, testing detection of non-existence; 2) *Delayed Coordination* isolates hard exploration under sparse, delayed rewards; 3) *Hide-or-Run* adapted from De Alfaro et al. [13] has no stationary NE profiles; its value is only achieved in the limit via mixed strategies converging to "hide w.p. 1" for the zero-sum objective of player 1 eventually reaching s_1; 4) *Mixed NE* requires a mixed equilibrium; 5) *Safe vs. Risky* contrasts actions with differing sensitivity to transition probabilities; and 6) *Traffic Merge* evaluates safety-critical multi-step coordination. Together, they cover all objective types in our framework.

We aim to address the following research questions:

RQ1: Correctness. (a) **Near-optimality.** When $\mathcal{G}^\star$ admits an NE, does the algorithm output an ε-optimal ε-NE? (b) **Equilibrium existence detection.** When no exact stationary NE exists, when does the algorithm return a sound non-existence certificate, and when does incompleteness arise?

RQ2: Exploration strategy. How does the robust exploration RMDP compare with simpler or optimistic exploration rules under an otherwise identical PAC loop and stopping condition? We compare against: (i) uniform-random joint-action exploration; (ii) round-robin exploration targeting the least-visited unknown slot under the point-estimate model; and (iii) optimistic exploration, which maximises *best-case* rather than worst-case reachability (see Eq. (1)) over the current uncertainty set. We define each explorer in detail in Appendix K.1.

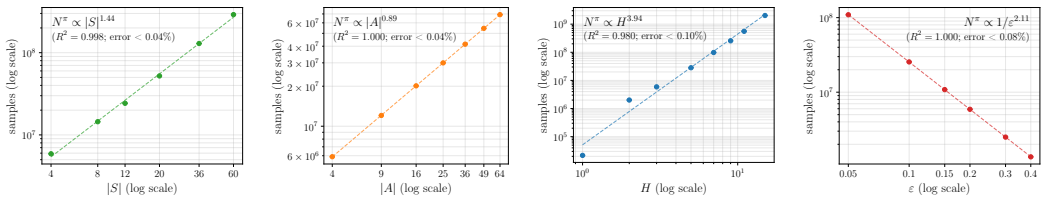
RQ3: Empirical sample-complexity scaling. How does the number of sampled trajectories until convergence scale with the number of states $|S|$, joint actions $|A|$, horizon H and precision ε? We use parametrised variants of *Safe vs. Risky* that vary one problem dimension at a time while preserving the underlying equilibrium structure. Specifically, we test $|S| \in \{4, 8, 12, 20, 36, 60\}$, $|A| \in \{4, 9, 16, 25, 36, 49, 64\}$, $H \in \{1, 2, 3, 5, 7, 9, 11, 15\}$ and $\varepsilon \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4\}$. For each scaling experiment, all parameters other than the varied dimension are held at the default values $|S| = 4, |A| = 4$ (as in Fig. 2e), with $H = 3, \varepsilon = 0.2$, and the same property used throughout.

Table 1: Comparison of our algorithm's output with the CSG-solver oracle across benchmarks. All cases use $\varepsilon = 0.1$. The returned profile attains the oracle's true-game value whenever a stationary NE exists. For *Cyclic Preferences* and the first *Hide-or-Run* case, no stationary NE exists; the algorithm returns `NOT_FOUND` in the former and an ε-NE approaching the limiting equilibrium value in the latter. For the second *Safe vs. Risky* case, the NE value is ∞, which our algorithm also recovers. The final column reports the robust-value estimate gap $V^\star - \hat{V}$. Full results are in Table 3, App. K.3.

Case study	**Property**	**ε-NE found?**	**NE exists?**	$\boldsymbol{V^\star - \hat{V}}$ $(\times 10^{-3})$
Cyclic Prefs.	$\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{R}_{r_1}[\mathtt{F}\, s_3] + \mathtt{R}_{r_2}[\mathtt{F}\, s_3])$	✗	✗	–
Delayed Coord.	$\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{R}_{r_1}[\mathtt{C}^{\leq 5}] + \mathtt{R}_{r_2}[\mathtt{C}^{\leq 5}])$	✓	✓	1.573 ± 0.069
Hide-or-run	$\langle\!\langle p_1 \rangle\!\rangle \mathtt{P}_{\max=?}[\mathtt{F}\, s_1]$	✓	✗	0.000 ± 0.000
Hide-or-run	$\langle\!\langle p_1 \rangle\!\rangle \mathtt{P}_{\max=?}[\mathtt{F}^{\leq 5} s_1]$	✓	✓	0.000 ± 0.000
Mixed NE	$\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{P}[\mathtt{F}\, s_1] + \mathtt{P}[\mathtt{F}\, s_2])$	✓	✓	3.083 ± 0.000
Safe vs. Risky	$\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{P}[\mathtt{F}\, s_1] + \mathtt{P}[\mathtt{F}\, s_2])$	✓	✓	1.743 ± 0.000
Safe vs. Risky	$\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{R}_{r_1}[\mathtt{F}\, s_1] + \mathtt{R}_{r_2}[\mathtt{F}\, s_2])$	✓	✓(∞)	–
Traffic Merge	$\langle\!\langle p_1 \rangle\!\rangle \mathtt{P}_{\max=?}[\mathtt{F}^{\leq 5} s_4]$	✓	✓	0.000 ± 0.000

Table 2: Samples to convergence (millions) and value gap $V^\star - \hat{V}$ for our robust exploration RMDP and the optimistic, round-robin, and uniform-random baselines. Values are reported as mean $\pm$ standard deviation; the lowest value gap and sample count for each benchmark are shown in bold.

	Safe vs. Risky $\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{P}[\mathtt{F}^{\leq 3} s_1] + \mathtt{P}[\mathtt{F}^{\leq 3} s_2])$		**Traffic Merge** $\langle\!\langle p_1 \rangle\!\rangle \mathtt{P}_{\max=?}[\mathtt{F}^{\leq 5} s_4]$		**Cyclic Prefs.** $\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{R}_{r_1}[\mathtt{F}\, s_3] + \mathtt{R}_{r_2}[\mathtt{F}\, s_3])$		**Delayed Coord.** $\langle\!\langle p_1 : p_2 \rangle\!\rangle_{\max=?} (\mathtt{P}[\mathtt{F}\, s_3] + \mathtt{P}[\mathtt{F}\, s_3])$	
Explorer	**Samples**	$V^\star - \hat{V}$	**Samples**	$V^\star - \hat{V}$	**Samples**	$V^\star - \hat{V}$	**Samples**	$V^\star - \hat{V}$
Robust (ours)	$\mathbf{5.90 \pm 0.00}$	$\mathbf{0.005 \pm 0.000}$	$\mathbf{6.90 \pm 0.00}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{18.41 \pm 0.01}$	–	$\mathbf{42.95 \pm 0.00}$	$\mathbf{0.000 \pm 0.000}$
Optimistic	$\mathbf{5.90 \pm 0.00}$	$\mathbf{0.005 \pm 0.000}$	$\mathbf{6.90 \pm 0.00}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{18.41 \pm 0.01}$	–	43.15 ± 0.87	0.060 ± 0.097
Round-robin	6.09 ± 0.09	0.005 ± 0.000	16.22 ± 0.00	$\mathbf{0.000 \pm 0.000}$	19.07 ± 0.01	–	44.92 ± 0.83	0.060 ± 0.097
Uniform	11.61 ± 0.01	0.005 ± 0.000	7.02 ± 0.00	$\mathbf{0.000 \pm 0.000}$	46.84 ± 0.02	–	730.09 ± 0.28	0.180 ± 0.063



(a) $|S|$-scaling (b) $|A|$-scaling (c) H-scaling (d) ε-scaling

Figure 1: Scaling of sample complexity (log–log scale) in the *Safe vs. Risky* game, following the setup in RQ3, with respect to (a) $|S|$, (b) $|A|$, (c) H, and (d) ε. Error bars show standard deviation over 10 runs; all are plotted but negligible ($< 0.10\%$ of the mean) and thus not visible.

5.1 Results and Discussion

RQ1(a) Near-optimality. For all case studies in Table 1 that admit an exact NE (i.e., $\bar{\mu} \geq 0$), the learned profile $\hat{\sigma}$ attains the same true-game value as the oracle up to ties between indifferent actions. For the finite-valued cases, this yields a value gap $V^\star - u(\hat{\sigma}, P^\star) = 0 \ll \varepsilon$. The robust value estimate $\hat{V}$ can, however, differ slightly from $V^\star$. Nevertheless, the estimation gap $V^\star - \hat{V}$ consistently remains below $4\Delta_t \leq \varepsilon$, as guaranteed by Lemma 14 (App. E). This empirically validates the stopping condition $\Delta_t \leq \varepsilon/4$, which ensures near-optimality of the output profile via Corollary 5: once Δ_t falls below this threshold, the robust solution computed on $\mathcal{P}_t$ transfers to the true model with error at most $4\Delta_t \leq \varepsilon$. Further, the second *Safe vs. Risky* property exhibits a degenerate case in which the SWNE value is unbounded: since both s_1 and s_2 are absorbing self-loops, once either is reached the game loops indefinitely, allowing the corresponding player to accumulate unbounded reward (while the reachability objective is effectively zero-sum); see Fig. 2, App. K. The algorithm correctly returns $V^\star = \infty$ together with a profile that attains it, demonstrating that our implementation handles unbounded total reward despite assuming bounded per-step rewards in the analysis.

RQ1(b) Equilibrium existence detection. Two benchmarks specifically consider the NE existence problem. 1) *Cyclic Preferences* provides a *clean non-existence* case. The game does not exhibit a

fixed-point solution and possesses no exact NE. Our solver correctly returns `NOT_FOUND` in 10/10 runs, yielding a sound certificate that no exact stationary NE exists. This is consistent with Proposition 4 which stipulates that solver failure implies $\mu^\star < 4\Delta_t - \varepsilon$, which is ≤ 0 at termination; and thus $\bar{\mu} \leq 0$ by Proposition 8, App. A. 2) *Hide-or-Run* illustrates *incompleteness* of the non-existence certificate. The game has no optimal stationary profile, but only an NE value in the limit (i.e., approached by a sequence of mixed strategies). In all runs, our finite-sample learner returns an ε-NE approaching the limiting equilibrium (hide w.p. 0.999 at s_0) rather than `NOT_FOUND`. This does not violate soundness, however: the non-existence certificate is required to be correct when emitted, but need not be emitted for every game lacking an exact stationary NE.

RQ2 Exploration strategy. Table 2 compares all explorers under the same PAC loop and stopping criterion. Uniform exploration is consistently less sample-efficient, requiring $2.0\times$ more trajectories on *Safe vs. Risky*, $2.5\times$ on *Cyclic Preferences*, and $17\times$ on *Delayed Coordination*. Round-robin performs similarly to the robust explorer on the shallow benchmarks but requires $2.3\times$ more trajectories on *Traffic Merge*. Optimistic exploration has similar sample consumption to robust exploration on most benchmarks, but uses slightly more samples on *Delayed Coordination*, where it also gives less stable and less accurate robust-value estimates: its estimated value differs from the true value by 0.06, whereas our robust explorer gives a near-perfect estimate. These results suggest that targeted exploration matters most when useful state-action pairs are difficult to reach, while pessimistic or robust planning provides more reliable coverage under transition uncertainty. We further analyse the learning dynamics of our robust explorer in Appendix K.2, including the evolution of uncertainty, known-slot coverage, confidence-radius imbalance, and per-episode sample consumption.

RQ3 Sample complexity scaling. Fig. 1 shows the dependence of total sample consumption on different problem parameters. Log–log fits give empirical exponents of 1.44 for $|S|$, 0.89 for $|A|$, 3.94 for H and 2.11 for $1/\varepsilon$. These are consistent with the worst-case bound of $|S|^2|A|H^4/\varepsilon^2$ in Theorem 7. In particular, the horizon exponent closely tracks the theoretical dependence over the tested range, while state and action space growth is milder than the upper bound. These experiments do not establish tight asymptotic rates, but show that the observed scaling is consistent with the theorem over the tested instances.

6 Conclusion

We introduced PAC-CSG, the first PAC learning framework for general-sum concurrent stochastic games under transition uncertainty that handles stationary NE (non-)existence. The algorithm combines L^1 uncertainty sets with a new RMDP formulation that casts joint state–action coverage as a robust reachability problem, and uses a black-box RCSG solver to compute ε-RSWNEs. Under a graph reachability condition, it terminates after $\widetilde{\mathcal{O}}\big(R_{\max}^2 H^4 |S|^2 |A| / (\varepsilon^2 p_{\text{reach}})\big)$ trajectory samples and outputs either an ε-social-welfare-optimal ε-NE or a sound certificate of NE non-existence. Beyond these guarantees, our results highlight two key insights: 1) robust equilibrium computation in the empirical model transfers to the true game; and 2) equilibrium non-existence can be detected in a sound, data-driven manner. Empirically, our framework achieves near-optimality, sound non-existence detection, and sample complexity consistent with theory. Future work includes relaxing the standard assumptions of centralised exploration, known transition support, and reachability, as well as improving scalability in $|S|$ and $|A|$, e.g., through value-aware exploration; see App. L for further discussion.

Acknowledgments and Disclosure of Funding

This work is supported by the EPSRC Centre for Doctoral Training no. EP/Y035070/1 and EPSRC grant EP/Y028872/1, Mathematical Foundations of Intelligence: An "Erlangen Programme" for AI.

References

[1] A. Al-Marjani, A. Tirinzoni, and E. Kaufmann. Active coverage for pac reinforcement learning. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5044–5109. PMLR, 2023. URL `https://proceedings.mlr.press/v195/al-marjani23a.html`.

[2] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 263–272. JMLR.org, 2017.

[3] Y. Bai and C. Jin. Provable self-play algorithms for competitive reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[4] R. Berthon, J.-P. Katoen, M. Mittelmann, and A. Murano. Robust strategies for stochastic multi-agent systems. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 2437–2439, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

[5] D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

[6] P. Bouyer, N. Markey, and D. Stan. Mixed nash equilibria in concurrent terminal-reward games. In *FSTTCS 2014*, 2014.

[7] R. I. Brafman and M. Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[8] T. Brihaye, V. Bruyère, A. Goeminne, J.-F. Raskin, and M. Van den Bogaard. The complexity of subgame perfect equilibria in quantitative reachability games. *Logical Methods in Computer Science*, 16, 2020.

[9] P. F. Castro and P. D'Argenio. Polytopal stochastic games. In *Principles of Verification: Cycling the Probabilistic Landscape: Essays Dedicated to Joost-Pieter Katoen on the Occasion of His 60th Birthday*, pages 99—-117. Springer, 2024.

[10] K. Chatterjee, K. Sen, and T. A. Henzinger. Model-checking ω-regular properties of interval Markov chains. In *International Conference on Foundations of Software Science and Computational Structures*, pages 302–317. Springer, 2008.

[11] T. Chen, V. Forejt, M. Kwiatkowska, D. Parker, and A. Simaitis. Automatic verification of competitive stochastic systems. In C. Flanagan and B. König, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, pages 315–330, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-28756-5.

[12] L. De Alfaro and T. A. Henzinger. Concurrent omega-regular games. In *Proceedings Fifteenth Annual IEEE Symposium on Logic in Computer Science (Cat. No. 99CB36332)*, pages 141–154. IEEE, 2000.

[13] L. De Alfaro, T. A. Henzinger, and O. Kupferman. Concurrent reachability games. *Theoretical computer science*, 386(3):188–217, 2007.

[14] L. Euler. De serie lambertina plurimisque eius insignibus proprietatibus. *Acta Academiae scientiarum imperialis petropolitanae*, pages 29–51, 1783.

[15] Z. U. Farhat, D. Ghosh, G. K. Atia, and Y. Wang. Sample-efficient distributionally robust multi-agent reinforcement learning via online interaction, 2026. URL `https://arxiv.org/abs/2508.02948`.

[16] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[17] R. Givan, S. Leach, and T. Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122(1-2):71–109, 2000.

[18] A. Y. He and D. Parker. Robust verification of concurrent stochastic games. In *Proc. 32nd International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'26)*, LNCS. Springer, 2026.

[19] C. A. Holt and A. E. Roth. The nash equilibrium: A perspective. *Proceedings of the National Academy of Sciences*, 101(12):3999–4002, 2004. doi: 10.1073/pnas.0308738101. URL https://www.pnas.org/doi/abs/10.1073/pnas.0308738101.

[20] J. Hu and M. P. Wellman. Nash q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4(null):1039–1069, Dec. 2003. ISSN 1532-4435.

[21] G. N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

[22] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, Aug. 2010. ISSN 1532-4435.

[23] X. Jin, M. Dan, N. Zhang, W. Yu, X. Fu, and S. K. Das. Chapter 2 - game theory for infrastructure security: The power of intent-based adversary models11part of the results included in this chapter was presented in [1], [2]. In *Handbook on Securing Cyber-Physical Critical Infrastructure*, pages 31–53. Morgan Kaufmann, Boston, 2012. ISBN 978-0-12-415815-3. doi: https://doi.org/10.1016/B978-0-12-415815-3.00002-9. URL https://www.sciencedirect.com/science/article/pii/B9780124158153000029.

[24] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2):209–232, 2002. ISSN 1573-0565. doi: 10.1023/A:1017984413808. URL https://doi.org/10.1023/A:1017984413808.

[25] M. Kwiatkowska, G. Norman, and D. Parker. PRISM model checker. https://www.prismmodelchecker.org/, 2013–2026.

[26] M. Kwiatkowska, G. Norman, D. Parker, and G. Santos. PRISM-games 3.0: Stochastic game verification with concurrency, equilibria and time. In *Proc. 32nd International Conference on Computer Aided Verification (CAV'20)*, volume 12225 of *LNCS*, pages 475–487. Springer, 2020.

[27] M. Kwiatkowska, G. Norman, D. Parker, and G. Santos. Automatic verification of concurrent stochastic systems. *Formal Methods in System Design*, 58(1):188–250, 2021.

[28] M. Kwiatkowska, G. Norman, D. Parker, and G. Santos. Correlated equilibria and fairness in concurrent stochastic games. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 60–78. Springer, 2022.

[29] J. H. Lambert. Observationes variae in mathesin puram. *Acta Helvetica*, 3(1):128–168, 1758.

[30] G. Li, Y. Chi, Y. Wei, and Y. Chen. Minimax-optimal multi-agent rl in markov games with a generative model. *Advances in Neural Information Processing Systems*, 35:15353–15367, 2022.

[31] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In W. W. Cohen and H. Hirsh, editors, *Machine Learning Proceedings 1994*, pages 157–163. Morgan Kaufmann, San Francisco (CA), 1994. ISBN 978-1-55860-335-6. doi: https://doi.org/10.1016/B978-1-55860-335-6.50027-1. URL https://www.sciencedirect.com/science/article/pii/B9781558603356500271.

[32] Q. Liu, T. Yu, Y. Bai, and C. Jin. A sharp analysis of model-based reinforcement learning with self-play. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7001–7010. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21z.html.

[33] S. Ma, Z. Chen, S. Zou, and Y. Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *J. Mach. Learn. Res.*, 24(1), Jan. 2023. ISSN 1532-4435.

[34] T. Meggendorfer, M. Weininger, and P. Wienhöft. Solving robust Markov decision processes: Generic, reliable, efficient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(25), pages 26631–26641, 2025.

[35] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[36] K. Panaganti and D. Kalathil. Sample complexity of robust reinforcement learning with a generative model. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9582–9602. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

[37] K. Panaganti, Z. Xu, D. Kalathil, and M. Ghavamzadeh. Robust reinforcement learning using offline data. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[38] Z. Roch and Y. Wang. Distributionally robust Markov games with average reward, 2025. URL https://arxiv.org/abs/2508.03136.

[39] L. S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

[40] L. Shi, E. Mazumdar, Y. Chi, and A. Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[41] J. Stirling. *Methodus differentialis: sive tractatus de summatione et interpolatione serierum infinitarum*. G. Strahan, 1730. Typis Gul. Bowyer.

[42] A. L. Strehl, L. Li, and M. L. Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(84):2413–2444, 2009. URL http://jmlr.org/papers/v10/strehl09a.html.

[43] M. Suilen, T. Badings, E. M. Bovy, D. Parker, and N. Jansen. Robust Markov decision processes: A place where AI and formal methods meet. In *Principles of Verification: Cycling the Probabilistic Landscape: Essays Dedicated to Joost-Pieter Katoen on the Occasion of His 60th Birthday*, pages 126–154. Springer, 2024.

[44] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[45] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior, princeton, 1944.

[46] S. Wang, N. Si, J. Blanchet, and Z. Zhou. A finite sample complexity bound for distributionally robust q-learning. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3370–3398. PMLR, 2023. URL https://proceedings.mlr.press/v206/wang23b.html.

[47] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, page 125, 2003.

[48] W. Wiesemann, D. Kuhn, and B. Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[49] W. Yang, L. Zhang, and Z. Zhang. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

[50] A. Zehfroosh and H. G. Tanner. Pac reinforcement learning algorithm for general-sum markov games. *IEEE Transactions on Automatic Control*, 68(5):2821–2831, 2023. doi: 10.1109/TAC.2022.3219340.

[51] K. Zhang, S. M. Kakade, T. Başar, and L. F. Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *J. Mach. Learn. Res.*, 24(1), Jan. 2023. ISSN 1532-4435.

[52] Z. Zhou, Z. Zhou, Q. Bai, L. Qiu, J. Blanchet, and P. Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3331–3339. PMLR, 2021. URL https://proceedings.mlr.press/v130/zhou21d.html.

A Proof for Nash margin property (Definition 3)

Proposition 8. $\bar{\mu} \geq 0$ *iff* $\mu^\star \geq 0$.

Proof. We prove both directions.

- ($\Rightarrow$) Suppose $\bar{\mu} \geq 0$. By definition of $\bar{\mu}$, there exists a profile σ with Nash margin at least 0, i.e., an NE profile. Therefore, the set of NE profiles is non-empty. By definition, an SWNE is an NE that maximises social welfare among all NE profiles. Since the set of NE profiles is non-empty, such a profile exists. As every SWNE is an NE, it must also have non-negative Nash margin, and hence $\mu^\star \geq 0$.
- ($\Leftarrow$) Since $\bar{\mu}$ is the maximum achievable Nash margin, by definition $\bar{\mu} \geq \mu^\star \geq 0$.

□

B Concentration bounds and confidence radii

In this section, we establish the Weissman concentration result (Lemma 9) and the containment event $\mathcal{E}^{\text{cont}}$ (via Corollary 10) used in Sec. 4.1.

We begin by analysing the concentration of empirical transition probabilities, which are used to construct the per-slot L^1 uncertainty sets $\mathcal{P}_t(s,a)$ for each episode $t \geq 1$ in Alg. 1.

For a state–action pair $(s,a) \in \mathcal{X}_{\text{nt}}$, let $N_{sa} \geq 0$ denote the number of independent samples drawn from the true transition distribution $P^\star_{sa}(\cdot)$, and let $N_{sas'}$ denote the number of observed transitions to state s'.

Further let

$$\text{Succ}(s,a) := \{s' \mid (s,a,s') \in \text{Supp}(P^\star)\},$$

and define the support-preserving simplex

$$\Delta_+(\text{Succ}(s,a)) := \{p \in \Delta(\text{Succ}(s,a)) : \text{Supp}(p) = \text{Succ}(s,a)\}.$$

Equivalently, $\Delta_+(\text{Succ}(s,a))$ contains exactly those transition distributions whose support equals the known successor support of that slot. In the theoretical analysis, we use this support-preserving class; in implementation, we replace it with the corresponding closed simplex by imposing a 10^{-5} lower bound on probabilities within the support. This ensures graph preservation.

If $N_{sa} = 0$, we define the empirical uncertainty set $\mathcal{P}_t(s,a) = \Delta_+(\text{Succ}(s,a))$ and set the point estimate $\widehat{P}_{sa}$ to be the uniform distribution over this support.

If $N_{sa} > 0$, we define the empirical estimate as $\widehat{P}_{sa}(s') = N_{sas'}/N_{sa}$, and construct the corresponding L^1 uncertainty set as described below.

Lemma 9 (Per-slot concentration)**.** *Consider a given state–action pair (s,a) with $n \geq 1$ i.i.d. draws from $P^\star_{sa}(\cdot)$, and fix a failure probability $\delta^{sa} \in (0,1)$. Define the Weissman L^1 confidence radius*

$$\alpha_{sa}(n;\delta^{sa}) := \sqrt{\frac{2}{n}\ln\frac{2^{|S|}-2}{\delta^{sa}}}.$$

Then

$$\Pr\left(\|P^\star_{sa}(\cdot) - \widehat{P}_{sa}(\cdot)\|_1 \leq \alpha_{sa}(n)\right) \geq 1 - \delta^{sa}.$$

[↑ main]

Proof. We apply the Weissman inequality [47], which states that for any $\varepsilon > 0$,

$$\Pr\left(\|\widehat{P}_{sa} - P^{\star}_{sa}\|_1 \geq \varepsilon\right) \leq (2^{|S|} - 2)\exp\left(-\frac{n\varepsilon^2}{2}\right).$$

Setting the right-hand side equal to δ^{sa} and solving for ε:

$$(2^{|S|} - 2)\exp\left(-\frac{n\varepsilon^2}{2}\right) = \delta^{sa} \implies \varepsilon = \sqrt{\frac{2}{n}\ln\frac{2^{|S|} - 2}{\delta^{sa}}} =: \alpha_{sa}(n; \delta_{sa}).$$

□

Corollary 10 (Containment of $P^{\star}$ in $\mathcal{P}_t$)**.** *Fix $\delta^{cont} \in (0,1)$. For brevity, write $n := N_t(s,a)$ and define $\delta^{cont}_{sa,n} := \delta^{cont}/(|S||A|n(n+1))$. For each episode $t \geq 1$, define the slot-wise uncertainty set*

$$\mathcal{P}_t(s,a) := \left\{p \in \Delta_+(\mathrm{Succ}(s,a)) : \|p - \widehat{P}_{sa,n}\|_1 \leq \alpha_{sa}\big(n; \delta^{cont}_{sa,n}\big)\right\},$$

and the global uncertainty set

$$\mathcal{P}_t := \prod_{(s,a)\in\mathcal{X}_{\mathrm{nt}}} \mathcal{P}_t(s,a),$$

where $\alpha_{sa}\big(n; \delta^{cont}_{sa,n}\big)$ is the Weissman radius from Lemma 9, instantiated at $\delta^{sa} := \delta^{cont}_{sa,n}$. Then, with probability at least $1 - \delta^{cont}$,

$$P^{\star} \in \mathcal{P}_t \qquad \textit{for all } t \geq 1.$$

Proof. For each (s,a) and $n \geq 1$, define the event

$$\mathcal{E}^{\mathrm{cont}}_{sa,n} := \left\{\|P^{\star}_{sa} - \widehat{P}_{sa,n}\|_1 > \alpha_{sa}\big(n; \delta^{\mathrm{cont}}_{sa,n}\big)\,.\right\}$$

By Lemma 9,

$$\Pr(\mathcal{E}^{\mathrm{cont}}_{sa,n}) \leq \delta^{\mathrm{cont}}_{sa,n} = \frac{\delta^{\mathrm{cont}}}{|S||A|n(n+1)}.$$

Taking a union bound over all $(s,a) \in \mathcal{X}_{\mathrm{nt}}$ and all possible sample counts $n \geq 1$,

$$\Pr\left(\bigcup_{(s,a)}\bigcup_{n\geq 1}\mathcal{E}^{\mathrm{cont}}_{sa,n}\right) \leq \sum_{(s,a)}\sum_{n\geq 1}\Pr(\mathcal{E}^{\mathrm{cont}}_{sa,n}) = \delta^{\mathrm{cont}}\sum_{n=1}^{\infty}\frac{1}{n(n+1)} = \delta^{\mathrm{cont}},$$

Hence with probability at least $1 - \delta^{\mathrm{cont}}$, no such failure occurs for any (s,a) and any n, which implies $P^{\star} \in \mathcal{P}_t$ for all episodes $t \geq 1$. □

Conditioning on containment of the true kernel. Define the per-episode and all-episode containment event:

$$\mathcal{E}^{\mathrm{cont}}_t := \{P^{\star} \in \mathcal{P}_t\}, \qquad \mathcal{E}^{\mathrm{cont}} := \bigwedge_{t\geq 1}\mathcal{E}^{\mathrm{cont}}_t.$$

Following Corollary 10, we use the summable failure schedule $\delta^{\mathrm{cont}}_n = \delta^{\mathrm{cont}}/(n(n+1))$ to ensure that

$$\Pr(\mathcal{E}^{\mathrm{cont}}) \geq 1 - \delta^{\mathrm{cont}}.$$

In the remainder of the analysis we condition on $\mathcal{E}^{\mathrm{cont}}$.

C Proofs for value sensitivity (Lemma 2)

We continue to derive the results in Sec. 4.1 of the main paper, focusing on the sensitivity lemma (Lemma 2) and presenting several useful corollaries. In the following, consider a fixed episode $t \geq 1$ and a given strategy profile $\sigma \in \Sigma$.

Proposition 11 (One-step error propagation)**.** *Let $\mu_h, \widehat{\mu}_h$ denote the stopped state marginals at time-step h under P and $\widehat{P}$ respectively, following σ, started from the same initial state $s_0 \in S$. Define*

$$\Delta_t^{L^1} := \max_{(s,a)\in\mathcal{X}_{\mathrm{nt}}} \alpha_t(s,a).$$

Then for all $h \geq 1$,

$$\|\mu_h - \widehat{\mu}_h\|_1 \leq h\Delta_t^{L^1}.$$

Proof. We prove the recursive bound

$$\|\mu_{h+1} - \widehat{\mu}_{h+1}\|_1 \leq \|\mu_h - \widehat{\mu}_h\|_1 + \Delta_t^{L^1}, \tag{4}$$

from which the result follows by induction using $\mu_0 = \widehat{\mu}_0 = \delta_{s_0}$, i.e., the Dirac distribution that assigns probability 1 to state s_0 and 0 to all others.

For $s \notin S_T$, define the mixed one-step kernels

$$K_s := \sum_a \sigma(a|s)P(\cdot|s,a), \qquad \widehat{K}_s := \sum_a \sigma(a|s)\widehat{P}(\cdot|s,a).$$

Then

$$\mu_{h+1} = \sum_{s\notin S_T} \mu_h(s)K_s, \qquad \widehat{\mu}_{h+1} = \sum_{s\notin S_T} \widehat{\mu}_h(s)\widehat{K}_s,$$

and hence

$$\mu_{h+1} - \widehat{\mu}_{h+1} = \sum_{s\notin S_T} (\mu_h(s) - \widehat{\mu}_h(s))K_s + \sum_{s\notin S_T} \widehat{\mu}_h(s)(K_s - \widehat{K}_s).$$

Taking L^1 norms and applying the triangle inequality:

$$\|\mu_{h+1} - \widehat{\mu}_{h+1}\|_1 \leq \Big\| \sum_{s\notin S_T} (\mu_h(s) - \widehat{\mu}_h(s))K_s \Big\|_1 + \Big\| \sum_{s\notin S_T} \widehat{\mu}_h(s)(K_s - \widehat{K}_s) \Big\|_1. \tag{5}$$

First term. Let $c_s := \mu_h(s) - \widehat{\mu}_h(s)$. Since each K_s is a probability distribution,

$$\sum_{s'} \left| \sum_{s\notin S_T} c_s K_s(s') \right| \leq \sum_{s\notin S_T} |c_s| \sum_{s'} K_s(s') = \sum_{s\notin S_T} |c_s|,$$

hence

$$\Big\| \sum_{s\notin S_T} (\mu_h(s) - \widehat{\mu}_h(s))K_s \Big\|_1 \leq \|\mu_h - \widehat{\mu}_h\|_1.$$

Second term. By convexity,

$$\left\| \sum_{s\notin S_T} \widehat{\mu}_h(s)(K_s - \widehat{K}_s) \right\|_1 \leq \sum_{s\notin S_T} \widehat{\mu}_h(s)\|K_s - \widehat{K}_s\|_1 \leq \max_{s\notin S_T} \|K_s - \widehat{K}_s\|_1.$$

For each $s \notin S_T$,

$$\begin{aligned}\|K_s - \widehat{K}_s\|_1 &\leq \sum_a \sigma(a|s)\|P_{sa} - \widehat{P}_{sa}\|_1 \leq \max_a \|P_{sa} - \widehat{P}_{sa}\|_1 \\ &= \max_a \sum_{s'} |P_{sa}(s') - \widehat{P}_{sa}(s')| \leq \max_a \alpha_t(s,a) = \Delta_t^{L^1}.\end{aligned}$$

where the last inequality follows from the construction of $\mathcal{P}_t$ (see Corollary 10).

Substituting into (5) gives

$$\|\mu_{h+1} - \widehat{\mu}_{h+1}\|_1 \leq \|\mu_h - \widehat{\mu}_h\|_1 + \Delta_t^{L^1}.$$

Unrolling the recursion yields

$$\|\mu_h - \widehat{\mu}_h\|_1 \leq h\,\Delta_t^{L^1}.$$

□

Lemma 2 (Sensitivity)**.** *Define* $\Delta_t := \frac{1}{2} R_{\max} H^2 \Delta_t^{L^1}$*, where* $\Delta_t^{L^1} := \max_{(s,a)\in\mathcal{X}_{\text{nt}}} \|P^\star_{sa} - \widehat{P}_{sa}\|_1$*. Then for any* $P \in \mathcal{P}_t$*, profile* σ*, player* $i \in N$*, and state* $s_0 \in S$*,* $|u_i(\sigma, P \mid s_0) - u_i(\sigma, \widehat{P} \mid s_0)| \le \Delta_t$*.* [↑ main]

Proof. The left-hand side could be rewritten as:

$$|u_i(\sigma, P \mid s_0) - u_i(\sigma, \widehat{P} \mid s_0)| = \left| \mathbb{E}^{\sigma,P}\left[\sum_{h=0}^{H-1} r_i(s_h, a_h)\right] - \mathbb{E}^{\sigma,\widehat{P}}\left[\sum_{h=0}^{H-1} r_i(s_h, a_h)\right] \right|$$
$$= \left| \sum_{h=0}^{H-1} \Big[\mathbb{E}^{\sigma,P}[r_i(s_h, a_h)] - \mathbb{E}^{\sigma,\widehat{P}}[r_i(s_h, a_h)] \Big] \right|.$$

Note that for bounded probabilistic reachability, we use the equivalent finite-horizon terminal-reward encoding, so the same argument applies. Moreover,

$$\mathbb{E}^{\sigma,P}[r_i(s_h, a_h)] = \sum_s \sum_a \mathbb{P}^{\sigma,P}(s_h = s)\sigma(a \mid s) \cdot r_i(s, a)$$
$$= \sum_s \mu_h(s) \sum_a \sigma(a \mid s) r_i(s, a) = \sum_s \mu_h(s) r_i^\sigma(s),$$

where $r_i^\sigma(s) := \mathbb{E}_{a\sim\sigma(s)}[r_i(s, a)]$. By assumption, $|r_i^\sigma(s)| \le R_{\max}$. Thus,

$$LHS = \left| \sum_{h=0}^{H-1} \sum_s r_i^\sigma(s) \big(\mu_h(s) - \widehat{\mu}_h(s)\big) \right|$$
$$\le \sum_{h=0}^{H-1} \max_s |r_i^\sigma(s)| \cdot \sum_s |\mu_h(s) - \widehat{\mu}_h(s)| \le R_{\max} \sum_{h=0}^{H-1} \|\mu_h - \widehat{\mu}_h\|_1$$
$$\le R_{\max} \Delta_t^{L^1} \sum_{h=0}^{H-1} h = R_{\max} \Delta_t^{L^1} \frac{H(H-1)}{2} \qquad \text{(by Proposition 11)}$$
$$\le \frac{1}{2} H^2 R_{\max} \Delta_t^{L^1} =: \Delta_t.$$

□

Henceforth, we define

$$\Delta_t := \tfrac{1}{2} R_{\max} H^2 \Delta_t^{L^1}, \quad \text{where} \quad \Delta_t^{L^1} := \max_{(s,a)\in\mathcal{X}_{\text{nt}}} \|P^\star_{sa} - \widehat{P}_{sa}\|_1.$$

Corollary 12. *For any* $P, P' \in \mathcal{P}_t$*, any strategy profile* σ*, and each player* $i \in N$*:*

$$|u_i(\sigma, P) - u_i(\sigma, P')| \le 2\Delta_t.$$

Proof. By the triangle inequality and Lemma 2 (applied twice),

$$|u_i(\sigma, P) - u_i(\sigma, P')| \le \Big|u_i(\sigma, P) - u_i(\sigma, \widehat{P})\Big| + \Big|u_i(\sigma, \widehat{P}) - u_i(\sigma, P')\Big| = 2\Delta_t.$$

□

Corollary 13 (Social-welfare sensitivity)**.** *For any* $P \in \mathcal{P}_t$ *and any strategy profile* σ*:*

$$\Big|u(\sigma, P) - u(\sigma, \widehat{P})\Big| \le 2\Delta_t.$$

Consequently, for any $P, P' \in \mathcal{P}_t$*:*

$$|u(\sigma, P) - u(\sigma, P')| \le 4\Delta_t.$$

Proof. Apply Lemma 2 to each player $i \in N$ and sum up:

$$\Big|u(\sigma, P) - u(\sigma, \widehat{P})\Big| \le \sum_{i\in N} \Big|u_i(\sigma, P) - u_i(\sigma, \widehat{P})\Big| \le \sum_{i\in N} \Delta_t = 2\Delta_t.$$

The pairwise bound follows identically from Corollary 12. □

D Proofs for equilibrium existence (Proposition 4)

This section establishes the key properties of the solver oracle (Proposition 4) that link robust equilibrium computation in the empirical model to equilibrium existence in the true game. We first formalise the solver oracle (Assumption 2), then show that any exact NE of the true game induces an approximate RNE in the empirical model (Lemma 3), and use this to derive the solver guarantees (Proposition 4).

Henceforth, we assume access to a black-box (L^1-)RCSG solver SOLVEL^1CSG with tolerance $\varepsilon \geq 0$, which computes an ε-RSWNE of the empirical RCSG $\mathcal{G}_t$. Below we use the notation of Alg. 1.

Assumption 2 (SOLVEL^1CSG oracle)**.** *Consider an (L^1-)RCSG $\mathcal{G}$ with uncertainty set $\mathcal{P}$. Given tolerance $\varepsilon \geq 0$, the solver returns $(\hat{\sigma}, \hat{V}, \text{found}) = \text{SOLVE}L^1\text{CSG}(\mathcal{G}, \varepsilon)$ such that $\text{found} = \text{true}$ iff $\hat{\sigma}$ is an ε-RSWNE of $\mathcal{G}$, in which case $\hat{V} = \underline{u}_{\mathcal{P}}(\hat{\sigma})$. Equivalently, $\text{found} = \text{false}$ iff $\Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}) = \emptyset$.*

Note that $\hat{\sigma} \in \Sigma_{\varepsilon\text{-RNE}}(\mathcal{P})$ implies $\Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}) \neq \emptyset$, so the failure condition is equivalently stated as $\Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}) = \emptyset$.

Assumption 2 abstracts equilibrium computation as an oracle, yielding an information-theoretic PAC guarantee that is decoupled from the solver's computational complexity. Practical solvers for general-sum L^1-RCSGs under our objectives do exist (e.g., in PRISM-games [18]), although polynomial-time implementations are currently known only in certain settings (e.g., the zero-sum case [18]). We defer a detailed discussion of computational aspects to App. L.

At each episode t, the solver either succeeds or fails to find an ε-RSWNE of $\mathcal{P}_t$. The following result characterises how each outcome relates to equilibrium properties of the true game $\mathcal{G}^\star$.

Lemma 3 (NE($P^\star$) $\Rightarrow$ approximate RNE($\mathcal{P}_t$))**.** *If σ is an (exact) NE under $P^\star$ achieving Nash margin $\mu := \mu(\sigma, P^\star) \geq 0$, then it is also an $(4\Delta_t - \mu)$-RNE under $\mathcal{P}_t$.* [↑ main]

Proof. Let σ be an NE under $P^\star$ with Nash margin $\mu \geq 0$. By the triangle inequality, for any $P \in \mathcal{P}_t$ and player $i \in N$, we have:

$$\begin{aligned} u_i(\sigma_{-i}[\sigma_i'], P) - u_i(\sigma, P) &\leq [u_i(\sigma_{-i}[\sigma_i'], P) - u_i(\sigma_{-i}[\sigma_i'], P^\star)] + [u_i(\sigma_{-i}[\sigma_i'], P^\star) - u_i(\sigma, P^\star)] \\ &\quad + [u_i(\sigma, P^\star) - u_i(\sigma, P)] \\ &\leq 2\Delta_t - \mu + 2\Delta_t = 4\Delta_t - \mu \end{aligned}$$

where the last inequality follows from Corollary 12 and the definition of μ (Definition 3). Since the bound holds for every $P \in \mathcal{P}_t$, taking $\inf$ preserves the inequality, giving

$$\inf_{P \in \mathcal{P}_t} [u_i(\sigma_{-i}[\sigma_i'], P) - u_i(\sigma, P)] \leq 4\Delta_t - \mu \qquad \forall i \in N.$$

Hence σ is an $(4\Delta_t - \mu)$-RNE under $\mathcal{P}_t$. □

Proposition 4 (Solver guarantees for equilibrium computation)**.** *On the event $\mathcal{E}^{cont}$:*

*(1) (*Equilibrium transfer*) If the solver returns an ε-RSWNE $\hat{\sigma}$ of $\mathcal{P}_t$, then $\hat{\sigma}$ is an ε-NE of $\mathcal{G}^\star$.*

*(2) (*Failure implies near non-existence*) If the solver returns `NOT_FOUND`, then either $\mu^\star$ is undefined or $\mu^\star < 4\Delta_t - \varepsilon$.*

[↑ main]

Proof. We prove each part individually.

1. Conditioning on $\mathcal{E}^{\text{cont}}$ we have $P^\star \in \mathcal{P}_t$. By definition of an ε-RSWNE (see after Definition 2), every ε-RSWNE under $\mathcal{P}_t$ is also an ε-RNE under $\mathcal{P}_t$, and every ε-RNE under $\mathcal{P}_t$ is an ε-NE of every $P \in \mathcal{P}_t$ including $P^\star$. Therefore, if the solver returns an ε-RSWNE $\hat{\sigma}$ under $\mathcal{P}_t$, then $\hat{\sigma}$ is an ε-NE of $P^\star$ (the transition kernel of $\mathcal{G}^\star$).

2. Suppose the solver returns `NOT_FOUND`. By Assumption 2, this means that $\Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}_t) = \emptyset$, i.e., no ε-RNE exists under $\mathcal{P}_t$. We now distinguish two cases according to the sign of $\bar{\mu} := \max_{\sigma \in \Sigma} \mu(\sigma, P^\star)$ (defined in Definition 3).

(a) If $\bar{\mu} < 0$, then no exact NE, and hence no SWNE (i.e., the NE that achieves the highest social welfare) exists under $P^\star$. However, the social-welfare optimal Nash margin $\mu^\star := \mu(\sigma^\star, P^\star)$ (Definition 3) is well-defined only when the SWNE $\sigma^\star$ exists, therefore trivially $\mu^\star$ is undefined.

(b) If $\bar{\mu} \geq 0$, then $\Sigma_{\text{NE}}(P^\star) \neq \emptyset$, so the SWNE $\sigma^\star$ exists and $\mu^\star = \mu(\sigma^\star, P^\star)$ is well-defined. Now suppose, for contradiction, that $\mu^\star \geq 4\Delta_t - \varepsilon$. Then Lemma 3 implies that $\sigma^\star$ is also a $(4\Delta_t - \mu^\star)$-RNE under $\mathcal{P}_t$. Since

$$4\Delta_t - \mu^\star \leq 4\Delta_t - (4\Delta_t - \varepsilon) = \varepsilon,$$

$\sigma^\star$ is also an ε-RNE under $\mathcal{P}_t$. Hence $\Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}_t) \neq \emptyset$, and Assumption 2 guarantees that the solver would return `FOUND` with an ε-RSWNE under $\mathcal{P}_t$, contradicting the assumption that the solver returned `NOT_FOUND`.

□

E Proofs for anytime guarantees (Corollary 5)

We now derive *anytime* guarantees on the robust estimation error $V^\star - \underline{u}_{\mathcal{P}_t}(\hat{\sigma})$ (Lemma 14) and the value gap $V^\star - u(\hat{\sigma}, P^\star)$ (Corollary 5, Sec. 4.1). In this section we assume $\bar{\mu} \geq 0$ so that $\Sigma_{\text{NE}}(P^\star) \neq \emptyset$ and hence $V^\star$ is well-defined. Also recall that $\hat{\sigma}$ is an ε-RSWNE under $\mathcal{P}_t$ by Assumption 2.

Lemma 14 (Anytime upper bound for $V^\star$)**.** *If $\mu^\star \geq 4\Delta_t - \varepsilon$, then $V^\star - \underline{u}_{\mathcal{P}_t}(\hat{\sigma}_t) \leq 4\Delta_t$.*

Proof. Since $\mu^\star \geq 4\Delta_t - \varepsilon$, Lemma 3 implies that $\sigma^\star \in \Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}_t) \neq \emptyset$. By Corollary 13 applied to $\sigma^\star$, for every $P \in \mathcal{P}_t$:

$$u(\sigma^\star, P) \geq u(\sigma^\star, P^\star) - 4\Delta_t = V^\star - 4\Delta_t.$$

Taking $\inf$ over $P \in \mathcal{P}_t$:

$$\inf_{P \in \mathcal{P}_t} u(\sigma^\star, P) = \underline{u}_{\mathcal{P}_t}(\sigma^\star) \geq V^\star - 4\Delta_t.$$

Since $\hat{\sigma}_t$ maximises the robust value over $\Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}_t)$ and $\sigma^\star \in \Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}_t)$:

$$\underline{u}_{\mathcal{P}_t}(\hat{\sigma}_t) = \sup_{\sigma \in \Sigma_{\varepsilon\text{-RNE}}(\mathcal{P}_t)} \underline{u}_{\mathcal{P}_t}(\sigma) \geq \underline{u}_{\mathcal{P}_t}(\sigma^\star) \geq V^\star - 4\Delta_t \quad \Longrightarrow \quad V^\star - \underline{u}_{\mathcal{P}_t}(\hat{\sigma}_t) \leq 4\Delta_t.$$

□

Corollary 5 (Per-episode value gap)**.** *If $\mu^\star \geq 4\Delta_t - \varepsilon$, then $V^\star - u(\hat{\sigma}_t, P^\star) \leq 4\Delta_t$.* [↑ main]

Proof. Immediate from Lemma 14 since $u(\hat{\sigma}_t, P^\star) \geq \underline{u}_{\mathcal{P}_t}(\hat{\sigma}_t)$ conditioning on $\mathcal{E}^{\text{cont}}$. □

F Supplementary details and proofs for exploration principles (Sec. 4.2)

This section provides the technical results underlying Sec. 4.2. We first show that the confidence allocation δ_t^{cov} sums to δ^{cov} (Proposition 18), then bound the number of trajectories per episode required to discover an unknown slot with probability at least $1 - \delta_t^{\text{cov}}$ (Proposition 17), and finally establish the exploration probability lower bound (Lemma 6).

Recall the following notation for the exploration analysis. A non-target slot (s, a) is *known* if it has been visited at least $n_{\min}$ times (defined in Lemma 20, App. G), and *unknown* otherwise. Let

$$\mathcal{U}_t = \{(s, a) \in \mathcal{X}_{\text{nt}} : N_t(s, a) < n_{\min}\}, \qquad \mathcal{K}_t = \mathcal{X}_{\text{nt}} \setminus \mathcal{U}_t$$

denote the sets of unknown and known slots, respectively, at the beginning of episode t. Here, $N_t(s, a)$ is the visit count of (s, a) by episode t. Further define $N_{\text{tot}} := |\mathcal{X}_{\text{nt}}| n_{\min} \leq |S||A| n_{\min}$, the total number of visits required across all relevant slots.

Interpreting the exploration RMDP construction. Consider Definition 5 of the *exploration RMDP*. The absorbing construction is used purely for analysis: in the implementation, trajectories are not terminated upon visiting $\mathcal{U}_t$ or S_T. This does not affect the guarantee, since visiting some $(s, a) \in \mathcal{U}_t$ depends only on the trajectory prefix up to the first such visit. Continuing the trajectory can only increase the probability of this event, so the lower bound in Lemma 6 remains valid. The state z ensures that expected reward coincides with visitation probability (each trajectory contributes at most once); without it, the objective would instead reflect expected visit counts. Finally, the exploration RMDP remains unchanged as long as $\mathcal{U}_t$ does not change. This allows an implementation optimisation: $\mathcal{G}_t^e$ need only be re-solved when new (s, a) pairs become known, instead of doing so at every episode.

We now show that the reachability constant p_{reach} lower bounds the exploration probability $p_t^e(P^\star)$.

Lemma 6 (Exploration probability)**.** *For all episodes $t \geq 1$ with $\mathcal{U}_t \neq \emptyset$, $p_t^e(P^\star) \geq p_{\text{reach}}$.* [↑ main]

Proof. Fix episode t with $\mathcal{U}_t \neq \emptyset$ and any $(s^\dagger, a^\dagger) \in \mathcal{U}_t$.

By Assumption 1, there exists a policy $\sigma_{\text{reach}} \in \Sigma$ such that

$$\inf_{P:\,\text{Supp}(P)=\text{Supp}(P^\star)} \mathbb{P}^{\sigma_{\text{reach}},P}[\text{reach } (s^\dagger, a^\dagger) \text{ before } S_T] \geq p_{\text{reach}}.$$

By definition of σ_t^e (1) as a maximiser of the robust value under r_t^e, for any comparator $\sigma' \in \Sigma$,

$$\inf_{P\in\mathcal{P}_t^e} \mathbb{E}^{\sigma_t^e,P}\left[\sum_{h=0}^{H-1} r_t^e(s_h, a_h)\right] = \underline{u}_{\mathcal{P}_t^e}(\sigma_t^e) \geq \underline{u}_{\mathcal{P}_t^e}(\sigma') = \inf_{P\in\mathcal{P}_t^e} \mathbb{E}^{\sigma',P}\left[\sum_{h=0}^{H-1} r_t^e(s_h, a_h)\right].$$

Due to the absorbing construction of $\mathcal{G}_t^e$, any trajectory visits $\mathcal{U}_t$ at most once. Hence for any policy σ and kernel P,

$$\sum_{h=0}^{H-1} r_t^e(s_h, a_h) = \mathbb{1}[\text{reach } \mathcal{U}_t],$$

and hence maximising expected cumulative reward is equivalent to maximising $\mathbb{P}[\text{reach } \mathcal{U}_t]$.

Therefore, by optimality of σ_t^e and taking $\sigma' = \sigma_{\text{reach}}$,

$$\begin{aligned}\inf_{P\in\mathcal{P}_t^e} \mathbb{P}^{\sigma_t^e,P}[\text{reach } \mathcal{U}_t] &\geq \inf_{P\in\mathcal{P}_t^e} \mathbb{P}^{\sigma_{\text{reach}},P}[\text{reach } \mathcal{U}_t] \\ &\geq \inf_{P\in\mathcal{P}_t^e} \mathbb{P}^{\sigma_{\text{reach}},P}[\text{reach } (s^\dagger, a^\dagger) \text{ before } S_T] \geq p_{\text{reach}} \qquad (6)\end{aligned}$$

where the last inequality follows from the definition of p_{reach} (2).

Now, since $(s^\dagger, a^\dagger) \in \mathcal{U}_t$, the event $\{\text{reach } (s^\dagger, a^\dagger)\}$ depends only on the trajectory prefix before the first visit to $\mathcal{U}_t$. By construction of $\mathcal{G}_t^e$ (Definition 5), any unknown slot $(s, a) \in \mathcal{U}_t$ transitions deterministically to the absorbing state z, so a trajectory reaching $(s^\dagger, a^\dagger)$ must avoid all other unknown slots beforehand. Consequently, the trajectory up to the hitting time of $\mathcal{U}_t$ depends only on transitions over $\mathcal{K}_t$.

Finally, to relate this bound to the true kernel, define $P^{e,\star} \in \mathcal{P}_t^e$ to coincide with $P^\star \in \mathcal{P}_t$ (conditioning on $\mathcal{E}^{\text{cont}}$) on $\mathcal{K}_t$ and use the absorbing transitions on $\mathcal{U}_t \cup S_T$. Since the event $\{\text{reach } \mathcal{U}_t\}$ depends only on transitions prior to the first visit to $\mathcal{U}_t$, and these transitions lie entirely within $\mathcal{K}_t$, we have

$$p_t^e(P^\star) = \mathbb{P}^{\sigma_t^e,P^\star}[\text{reach } \mathcal{U}_t] \geq \mathbb{P}^{\sigma_t^e,P^{e,\star}}[\text{reach } \mathcal{U}_t] \geq \inf_{P\in\mathcal{P}_t^e} \mathbb{P}^{\sigma_t^e,P}[\text{reach } \mathcal{U}_t] \overset{(6)}{\geq} p_{\text{reach}}.$$

□

Multi-trajectory exploration. Let $(\mathcal{F}_t)_{t\geq 0}$ be the filtration generated by the interaction history up to the end of episode t, and let $\mathcal{F}_{t-1}$ denote the history available at the start of episode t. At each episode t, we draw N_t^π (conditionally) independent trajectories under σ_t^e.

For $i \in \{1, \ldots, N_t^\pi\}$, define

$$X_{t,i} := \mathbb{1}[\text{trajectory } i \text{ visits some } (s, a) \in \mathcal{U}_t],$$

and let

$$X_t := \mathbb{1}[\exists\, i \leq N_t^\pi : X_{t,i} = 1].$$

Proposition 15. *Conditioned on $\mathcal{F}_{t-1}$, the variables $X_{t,1}, \ldots, X_{t,N_t^\pi}$ are independent and satisfy*

$$\Pr(X_{t,i} = 1 \mid \mathcal{F}_{t-1}) \geq p_t^e(P^\star) \qquad \textit{for all } i \in \{1, \ldots, N_t^\pi\}.$$

Consequently,

$$\Pr(X_t = 0 \mid \mathcal{F}_{t-1}) \leq \left(1 - p_t^e(P^\star)\right)^{N_t^\pi}.$$

Proof. Immediate from conditional independence of trajectories given $\mathcal{F}_{t-1}$ and the definition of $p_t^e(P^\star)$ (see Sec. 4.2). □

Corollary 16 (Per-episode exploration probability)**.**

$$\mathbb{P}(X_t = 0 \mid \mathcal{F}_{t-1}) \leq \exp(-p_{\text{reach}} N_t^\pi),$$

Proof. Combining Lemma 6 and Proposition 15 we have

$$\mathbb{P}(X_t = 0 \mid \mathcal{F}_{t-1}) = \left(1 - p_t^e(P^\star)\right)^{N_t^\pi} \leq (1 - p_{\text{reach}})^{N_t^\pi} \leq \exp(-p_{\text{reach}} N_t^\pi),$$

where the last inequality follows from the fact that $1 - x \leq e^{-x}$ for $x \in (0, 1)$. □

Proposition 17 (Number of samples per-episode)**.** *Consider a given failure probability $\delta_t^{cov} > 0$. If*

$$N_t^\pi := \left\lceil \frac{1}{p_{\text{reach}}} \log \frac{1}{\delta_t^{cov}} \right\rceil$$

then

$$\mathbb{P}(X_t = 0 \mid \mathcal{F}_{t-1}) \leq \delta_t^{cov}.$$

Proof. Substituting the given definition of N_t^π into the upper bound from Corollary 16 yields

$$\exp(-p_{\text{reach}} N_t^\pi) \leq \exp\left(-p_{\text{reach}} \cdot \frac{1}{p_{\text{reach}}} \log \frac{1}{\delta_t^{\text{cov}}}\right) = \exp\left(-\log \frac{1}{\delta_t^{\text{cov}}}\right) = \delta_t^{\text{cov}}.$$

Hence $\mathbb{P}(X_t = 0 \mid \mathcal{F}_{t-1}) \leq \delta_t^{\text{cov}}$ as required. □

Proposition 18 (Coverage confidence allocation)**.** *Let*

$$\delta_t^{cov} := \frac{\delta^{cov}}{t(t+1)}. \tag{7}$$

Then, for every episode $t \geq 1$ with $\mathcal{U}_t \neq \emptyset$, the probability of visiting some unknown (s, a) is at least $1 - \delta^{cov}$.

Proof. Using the telescoping identity

$$\sum_{t \geq 1} \frac{1}{t(t+1)} = \sum_{t \geq 1} \left(\frac{1}{t} - \frac{1}{t+1}\right) = 1,$$

we obtain

$$\sum_{t \geq 1} \delta_t^{\text{cov}} = \delta^{\text{cov}}.$$

By Proposition 17, each episode satisfies

$$\mathbb{P}(X_t = 0 \mid \mathcal{F}_{t-1}) \leq \delta_t^{\text{cov}}.$$

Applying a union bound over all episodes (with $\mathcal{U}_t \neq \emptyset$) yields

$$\mathbb{P}(\exists t \geq 1 : X_t = 0) \leq \sum_{t \geq 1} \delta_t^{\text{cov}} = \delta^{\text{cov}},$$

so with probability at least $1 - \delta^{\text{cov}}$, every episode visits some unknown (s, a). □

Together, Propositions 17 and 18 give the multi-trajectory exploration strategy used in Alg. 1: drawing $N_t^\pi = \lceil (1/p_{\text{reach}}) \log(1/\delta_t^{\text{cov}}) \rceil$ trajectories per episode guarantees that each episode is exploratory with probability at least $1 - \delta_t^{\text{cov}}$.

G Proofs for PAC-CSG (Theorem 7)

In this section we prove the main PAC guarantee, Theorem 7, by combining the three components developed in the previous Appendix items (B–F). The proof proceeds in four steps:

(1) *Confidence split*: we set $\delta^{\text{cont}} = \delta^{\text{cov}} = \delta/2$ and condition on $\mathcal{E} := \mathcal{E}^{\text{cont}} \cap \mathcal{E}^{\text{cov}}$, which holds with probability $\geq 1 - \delta$.

(2) *Episode count*: Lemma 19 bounds the number of episodes until all relevant slots are known at $T^\star = \widetilde{\mathcal{O}}\big(R_{\max}^2 H^4 |S|^2 |A|/\varepsilon^2\big)$.

(3) *Sample count*: at each episode t, Proposition 17 (App. F) requires $N_t^\pi = \mathcal{O}(\log t/p_{\text{reach}})$ sample trajectories, so the total number of samples required is $\sum_{t \leq T^\star} N_t^\pi = \widetilde{\mathcal{O}}(T^\star/p_{\text{reach}})$.

(4) *Correctness*: at termination, Corollary 22 applies, yielding either an ε-approximate equilibrium in $\mathcal{G}^\star$ or the non-existence certificate.

Henceforth we fix the target accuracy $\varepsilon \in (0, 1)$ and confidence $\delta \in (0, 1)$. Following Corollary 5, to attain the near-optimality guarantee, we aim for

$$4\Delta_t \leq \varepsilon \quad \Longrightarrow \quad \Delta_t \leq \frac{\varepsilon}{4},$$

and under this scheme the non-existence margin threshold in Proposition 4(2) becomes $4\Delta_t - \varepsilon = 0$.

For the confidence parameter δ, we distribute it as $\delta^{\text{cont}} = \delta^{\text{cov}} := \delta/2$.

Lemma 19 (Counting episodes)**.** *Fix a confidence parameter $\delta^{cov} \in (0, 1)$. With probability at least $1 - \delta^{cov}$, the total number of episodes executed by the algorithm until all relevant slots are known is at most*

$$T^\star := 2N_{\text{tot}} + \frac{8}{3} \ln(1/\delta^{cov}). \tag{8}$$

Proof. Let T be a fixed number of episodes before all relevant slots are known. For $t = 1, \ldots, T$, recall the definitions of $X_{t,i}$, X_t, and $\mathcal{F}_{t-1}$ from App. F. Further define

$$\mu_t := \mathbb{E}[X_t \mid \mathcal{F}_{t-1}], \qquad N^e := \sum_{t=1}^{T} X_t \leq T.$$

So N^e is the number of exploratory episodes.

Define martingale differences $Y_t := X_t - \mu_t$ and $Z_t := -Y_t$. Then Z_t are martingale differences with respect to the filtration $\mathcal{F}_{t-1}$. Thus

$$M_T := \sum_{t=1}^{T} Z_t = \sum_{t=1}^{T} \mu_t - N^e.$$

We have $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ and $|Z_t| \leq 1$. The predictable variance is

$$v := \sum_{t=1}^{T} \mathbb{E}[Z_t^2 \mid \mathcal{F}_{t-1}] = \sum_{t=1}^{T} \mu_t(1 - \mu_t) \leq \sum_{t=1}^{T} \mu_t \leq T, \tag{9}$$

where the last inequality uses $\mu_t \leq 1$.

We aim to ensure with high probability that $N^e \geq N_{\text{tot}}$, the total number of visits required across all relevant slots. That is,

$$\Pr\left(N^e \leq N_{\text{tot}}\right) \leq \delta^{\text{cov}}. \tag{10}$$

By Proposition 17, $\mathbb{P}(X_t = 0 \mid \mathcal{F}_{t-1}) \leq \delta_t^{\text{cov}}$. Therefore $\mu_t = \mathbb{P}(X_t = 1 \mid \mathcal{F}_{t-1}) \geq 1 - \delta_t^{\text{cov}}$ and

$$\sum_{t=1}^{T} \mu_t \geq \sum_{t=1}^{T} (1 - \delta_t^{\text{cov}}) = T - \sum_{t=1}^{T} \delta_t^{\text{cov}} \overset{(7)}{\geq} T - \delta^{\text{cov}}.$$

Therefore, on the event $\{N^e \le N_{\mathrm{tot}}\}$ we have

$$M_T = \sum_{t=1}^{T} \mu_t - N^e \ge T - \delta^{\mathrm{cov}} - N_{\mathrm{tot}} =: a. \tag{11}$$

Thus,

$$\Pr(N^e \le N_{\mathrm{tot}}) \le \Pr(M_T \ge a)\,. \tag{12}$$

Applying Freedman's inequality [16], and using the bound on the variance $v \le T$ (9), we obtain

$$\Pr(M_T \ge a) \le \exp\left(-\frac{a^2}{2(v + a/3)}\right) \le \exp\left(-\frac{a^2}{2(T + a/3)}\right) \tag{13}$$

We now choose

$$T = 2N_{\mathrm{tot}} + CL, \quad L := \ln(1/\delta^{\mathrm{cov}}), \tag{14}$$

for a constant $C > 0$ to be determined. Then

$$a = T - N_{\mathrm{tot}} - \delta^{\mathrm{cov}} = N_{\mathrm{tot}} + CL - \delta^{\mathrm{cov}}.$$

Substituting into the exponent in (13), to achieve (10) it suffices to ensure

$$\frac{(N_{\mathrm{tot}} + CL)^2}{2\big(T + a/3\big)} \ge \frac{(N_{\mathrm{tot}} + CL - \delta^{\mathrm{cov}})^2}{2\big(T + a/3\big)} \ge L,$$

since $\delta^{\mathrm{cov}} > 0$. Further,

$$T + \frac{a}{3} = 2N_{\mathrm{tot}} + CL + \frac{1}{3}\big(N_{\mathrm{tot}} + CL - \delta^{\mathrm{cov}}\big) \le 2N_{\mathrm{tot}} + CL + \frac{1}{3}\big(N_{\mathrm{tot}} + CL\big) = \frac{1}{3}(7N_{\mathrm{tot}} + 4CL)$$

Therefore, it suffices to ensure

$$\frac{3(N_{\mathrm{tot}} + CL)^2}{2\,(7N_{\mathrm{tot}} + 4CL)} \ge L. \tag{15}$$

We find that $C = 8/3$ is the minimal constant ensuring that (15) holds; we defer the case analysis establishing this to the end (see **Solving for** C).

With this choice of C, (13) yields

$$\Pr(M_T \ge a) \le \exp(-L) = \delta^{\mathrm{cov}}.$$

By (12), this implies

$$\Pr(N^e \le N_{\mathrm{tot}}) \le \Pr(M_T \ge a) \le \delta^{\mathrm{cov}}.$$

Thus, with probability at least $1 - \delta^{\mathrm{cov}}$, we have $N^e \ge N_{\mathrm{tot}}$.

Since each exploratory episode discovers at least one previously unknown slot, a standard counting argument implies that after at most N_{tot} such episodes all relevant slots are known. Hence, with probability at least $1 - \delta^{\mathrm{cov}}$, all relevant slots become known within T episodes.

Finally, substituting $C = 8/3$ into (14) yields

$$T^\star = 2N_{\mathrm{tot}} + \tfrac{8}{3}\ln(1/\delta^{\mathrm{cov}}).$$

Solving for C. We now return to (15) and solve for the smallest constant C that ensures the inequality holds. Introduce $x := N_{\mathrm{tot}}/L \ge 0$ (so $N_{\mathrm{tot}} = xL$). Then

$$\begin{aligned}
&\frac{3L^2(x + C)^2}{2L(7x + 4C)} = \frac{3L(x + C)^2}{2(7x + 4C)} \ge L \\
\iff &\frac{3(x + C)^2}{2(7x + 4C)} \ge 1 \\
\iff &3(x + C)^2 \ge 14x + 8C \\
\iff &h(x) := 3(x + C)^2 - 14x - 8C = 3x^2 + (6C - 14)x + (3C^2 - 8C) \ge 0.
\end{aligned} \tag{16}$$

We require $h(x) \ge 0$ for all $x \ge 0$. Since h is convex, its minimum occurs at

$$x^\star = \frac{7}{3} - C.$$

We distinguish two cases.

Case A: $x^\star \geq 0$ **(i.e.,** $C \leq 7/3$**).** The minimum of h on $[0, \infty)$ occurs at $x^\star$, yielding

$$h(x^\star) = 6C - \frac{49}{3}.$$

Requiring $h(x^\star) \geq 0$ gives $C \geq 49/18$, contradicting $C \leq 7/3$. Hence no feasible C exists in this regime.

Case B: $x^\star < 0$ **(i.e.,** $C > 7/3$**).** In this case, the minimum of h on $[0, \infty)$ occurs at the boundary $x = 0$, giving

$$h(0) = C(3C - 8).$$

Requiring $h(0) \geq 0$ yields $C \geq 8/3$.

Therefore, the smallest constant C such that (16) holds for all $x \geq 0$ is

$$\boxed{C = \tfrac{8}{3}}.$$

□

Lemma 20 (Counting per-slot visits)**.** *Define* $\delta^{cont}_{sa,n} := \delta^{cont}/(|S||A|n(n+1))$*. If a slot* $(s,a) \in \mathcal{X}_{\mathrm{nt}}$ *is visited*

$$n_{\min} := \left\lceil -\frac{16R_{\max}^2 H^4}{\varepsilon^2} W_{-1}\left(-\frac{\varepsilon^2}{16R_{\max}^2 H^4}\sqrt{\frac{\delta^{cont}}{2(2^{|S|}-2)|S||A|}}\right)\right\rceil = \widetilde{\mathcal{O}}\left(\frac{R_{\max}^2 H^4 |S|}{\varepsilon^2}\right)$$

times, then

$$\alpha_{sa}(n; \delta^{cont}_{sa,n}) \leq \varepsilon/(2R_{\max}H^2).$$

Proof. Applying Lemma 9, we require

$$\sqrt{\frac{2}{n}\ln\frac{(2^{|S|}-2)}{\delta^{\text{cont}}_{sa,n}}} = \sqrt{\frac{2}{n}\ln\frac{(2^{|S|}-2)|S||A|n(n+1)}{\delta^{\text{cont}}}} \leq \frac{\varepsilon}{2R_{\max}H^2}.$$

Solving for n**.** Squaring both sides and rearranging gives

$$\frac{2}{n}\ln\frac{(2^{|S|}-2)|S||A|n(n+1)}{\delta^{\text{cont}}} \leq \frac{\varepsilon^2}{4R_{\max}^2H^4} \quad \Longrightarrow \quad \ln\frac{(2^{|S|}-2)|S||A|n(n+1)}{\delta^{\text{cont}}} \leq \frac{\varepsilon^2}{8R_{\max}^2H^4}n.$$

Let

$$C := \frac{(2^{|S|}-2)|S||A|}{\delta^{\text{cont}}}, \qquad \beta := \frac{\varepsilon^2}{8R_{\max}^2H^4}.$$

Then we need

$$\ln(Cn(n+1)) \leq \beta n.$$

Since $n(n+1) \leq 2n^2$ for all $n \geq 1$, it is enough to require

$$\ln(2Cn^2) \leq \beta n.$$

Solving the corresponding equality gives

$$n = -\frac{2}{\beta}W_{-1}\left(-\frac{\beta}{2\sqrt{2C}}\right), \tag{17}$$

where W is the Lambert W function [14, 29] defined by $W(x)e^{W(x)} = x$, and W_{-1} is the lower real branch of W mapping $x \in [-1/e, 0)$[2] to $W_{-1}(x) \leq -1$.

Therefore a sufficient condition is

$$n \geq \left\lceil -\frac{16R_{\max}^2H^4}{\varepsilon^2}W_{-1}\left(-\frac{\varepsilon^2}{16R_{\max}^2H^4}\sqrt{\frac{\delta^{\text{cont}}}{2(2^{|S|}-2)|S||A|}}\right)\right\rceil =: n_{\min}$$

Asymptotically,

$$n_{\min} = \mathcal{O}\left(\frac{R_{\max}^2H^4}{\varepsilon^2}\log\frac{(2^{|S|}-2)|S||A|}{\delta^{\text{cont}}}\right) = \widetilde{\mathcal{O}}\left(\frac{R_{\max}^2H^4|S|}{\varepsilon^2}\right). \tag{18}$$

□

[2]The argument of W_{-1} lies in $[-1/e, 0)$ under the standard PAC parameter regimes (e.g., ($\delta^{\text{cont}} \in (0,1)$).

Henceforth we define $n_{\min}$ as in Lemma 20.

Corollary 21. *If every $(s,a) \in \mathcal{X}_{\mathrm{nt}}$ is visited at least $n_{\min}$ times, then $\Delta_t \le \varepsilon/4$.*

Proof. By Lemma 20, if every $(s,a) \in \mathcal{X}_{\mathrm{nt}}$ is visited at least $n_{\min}$ times, then on $\mathcal{E}^{\mathrm{cont}}$ we have:

$$\|\widehat{P}_{sa} - P^\star_{sa}\|_1 \le \alpha_{sa}(n_{\min}; \delta^{\mathrm{cont}}_{sa,n_{\min}}) \le \frac{\varepsilon}{2R_{\max}H^2} \qquad \forall (s,a) \in \mathcal{X}_{\mathrm{nt}}.$$

Therefore, by definition of Δ_t:

$$\Delta_t = \frac{1}{2}R_{\max}H^2 \max_{(s,a)\in\mathcal{X}_{\mathrm{nt}}} \|\widehat{P}_{sa} - P^\star_{sa}\|_1 \le \frac{1}{2}R_{\max}H^2 \cdot \frac{\varepsilon}{2R_{\max}H^2} = \frac{\varepsilon}{4}.$$

□

Corollary 22. *Consider the* SOLVEL^1CSG *oracle from Assumption 2 with tolerance ε. Suppose $\Delta_t \le \varepsilon/4$. Then:*

(a) If $\bar{\mu} \ge 0$, the solver returns `FOUND`*, and the output profile $\hat{\sigma}$ is an ε-NE under $P^\star$, satisfying*

$$V^\star - u(\hat{\sigma}, P^\star) \le \varepsilon.$$

(b) If $\bar{\mu} < 0$, then:

(i) if the solver returns `FOUND`*, the output profile $\hat{\sigma}$ is still an ε-NE under $P^\star$;*
(ii) if the solver returns `NOT_FOUND`*, this provides a sound certificate that $\bar{\mu}$ is not well-defined or $\bar{\mu} < 0$.*

Proof. Suppose $\Delta_t \le \varepsilon/4$. Then $4\Delta_t - \varepsilon \le 0$. We now proceed by a case analysis on the sign of $\bar{\mu}$.

(a) Suppose $\bar{\mu} \ge 0$. By Proposition 8, this is equivalent to $\mu^\star \ge 0$. Since $\mu^\star \ge 0 \ge 4\Delta_t - \varepsilon$, the contrapositive of Proposition 4(2) implies that the solver returns `FOUND`. Now by Assumption 2, returning `FOUND` means that the output profile $\hat{\sigma}$ is an ε-RSWNE under $\mathcal{P}_t$. Then condition on $\mathcal{E}^{\mathrm{cont}}$, Proposition 4(1) implies that $\hat{\sigma}$ is also an ε-NE of $P^\star \in \mathcal{P}_t$.

Furthermore, since $\mu^\star \ge 4\Delta_t - \varepsilon$, we may apply Corollary 5, which gives

$$V^\star - u(\hat{\sigma}, P^\star) \le 4\Delta_t \le \varepsilon.$$

(b) Now suppose $\bar{\mu} < 0$.

(i) If the solver returns `FOUND`, outputting an ε-RSWNE $\hat{\sigma}$ of $\mathcal{P}_t$, then Proposition 4(1) implies that $\hat{\sigma}$ is also an ε-NE of $P^\star$.
(ii) If the solver returns `NOT_FOUND`, then by Proposition 4(2), either $\mu^\star$ is undefined or $\mu^\star < 4\Delta_t - \varepsilon \le 0$. Therefore, by Proposition 8, $\bar{\mu}$ is either undefined or satisfies $\bar{\mu} < 0$.

□

Theorem 7 (PAC guarantee for finite-horizon objectives)**.** *Set $n_{\min} = \widetilde{\mathcal{O}}\big(R^2_{\max}H^4|S|/\varepsilon^2\big)$. Then with probability at least $1-\delta$, Alg. 1 terminates after at most*

$$\widetilde{\mathcal{O}}\left(\frac{R^2_{\max}H^4|S|^2|A|}{\varepsilon^2 p_{\mathrm{reach}}}\right) \tag{3}$$

trajectory samples, where $\widetilde{\mathcal{O}}(\cdot)$ suppresses logarithmic factors in the problem parameters. Upon termination, the algorithm's output satisfies the PAC-CSG guarantee of Definition 4. [↑ main]

Proof. Recall the containment and coverage events

$$\mathcal{E}^{\mathrm{cont}} := \bigcap_{t\ge 1}\{P^\star \in \mathcal{P}_t\}, \qquad \mathcal{E}^{\mathrm{cov}} := \{\text{all } (s,a) \in \mathcal{X}_{\mathrm{nt}} \text{ are visited at least } n_{\min} \text{ times}\}.$$

By Corollary 10 and Lemma 19,

$$\Pr(\mathcal{E}^{\mathrm{cont}}) \ge 1 - \delta^{\mathrm{cont}}, \qquad \Pr(\mathcal{E}^{\mathrm{cov}}) \ge 1 - \delta^{\mathrm{cov}}.$$

Setting $\delta^{\mathrm{cont}} = \delta^{\mathrm{cov}} = \delta/2$ and defining $\mathcal{E} := \mathcal{E}^{\mathrm{cont}} \cap \mathcal{E}^{\mathrm{cov}}$, we have $\Pr(\mathcal{E}) \ge 1 - \delta$.

Sample complexity. Let

$$T := \inf\{t : \Delta_t \le \varepsilon/4\}.$$

For any $t < T$, the stopping condition is false, in particular $\Delta_t > \varepsilon/4$.

Substituting $n_{\min}$ from Corollary 21 into the episode bound $T^\star$ from Lemma 19 yields

$$T^\star = 2|S||A|n_{\min} + \frac{8}{3}\ln\left(\frac{1}{\delta^{\text{cov}}}\right) = \mathcal{O}\left(\frac{R_{\max}^2 H^4 |S|^2 |A|}{\varepsilon^2}\right).$$

By Proposition 17, each episode requires

$$\left\lceil \frac{1}{p_{\text{reach}}} \log \frac{1}{\delta_t^{\text{cov}}} \right\rceil = \mathcal{O}\left(\frac{1}{p_{\text{reach}}} \log \frac{t}{\delta^{\text{cov}}}\right)$$

trajectories. Summing over episodes,

$$\begin{aligned}
\sum_{t=1}^{T^\star} \frac{1}{p_{\text{reach}}} \log \frac{t}{\delta^{\text{cov}}} &= \frac{1}{p_{\text{reach}}} \left(\sum_{t=1}^{T^\star} \log t - T^\star \log \delta^{\text{cov}} \right) \\
&= \frac{1}{p_{\text{reach}}} \left(\log(T^\star!) - T^\star \log \delta^{\text{cov}} \right) \\
&= \mathcal{O}\left(\frac{T^\star}{p_{\text{reach}}} \log \frac{T^\star}{\delta^{\text{cov}}} \right) = \widetilde{\mathcal{O}}\left(\frac{T^\star}{p_{\text{reach}}} \right),
\end{aligned}$$

where we use Stirling's approximation [41] $\log(T!) = \mathcal{O}(T \log T)$.

Termination. Let T^{cov} denote the first episode at which $\mathcal{E}^{\text{cov}}$ holds. On $\mathcal{E}^{\text{cov}}$, Lemma 19 gives

$$T^{\text{cov}} \le T^\star.$$

By Corollary 21, on $\mathcal{E}^{\text{cont}}$, once $\mathcal{E}^{\text{cov}}$ holds we have

$$\Delta_{T^{\text{cov}}} \le \varepsilon/4.$$

Hence, on $\mathcal{E}$, the stopping condition is satisfied by episode T^{cov}, and the algorithm terminates no later than $T^\star$.

Correctness of the output. On $\mathcal{E}$, we have both $P^\star \in \mathcal{P}_t$ for all t and $\Delta_T \le \varepsilon/4$. Therefore, Corollary 22 applies at termination, yielding the required output guarantee and completing the proof. □

H Supplementary details and proofs for infinite-horizon objectives

In this section, we extend the PAC-CSG framework to infinite-horizon objectives (unbounded probabilistic reachability and reachability reward / SSP). Following standard SSP formulations [5, 18, 27], we assume the existence of at least one proper profile and restrict attention to this class throughout the analysis. Let

$$\tau_T := \inf\{h \ge 0 : s_h \in S_T\}$$

denote the hitting time of the target set S_T. Further let $\Sigma^{\text{prop}} \subseteq \Sigma$ be the set of proper profiles, under which $\tau_T < \infty$ almost surely and all values are well-defined and that the underlying SOLVEL^1CSG oracle converges. The assumption can be verified via standard graph-based reachability analyses under the known support structure induced by graph preservation [12].

Assumption 3 (Existence of a proper profile)**.** *There exists a proper profile $\sigma \in \Sigma^{\text{prop}}$, i.e., one that reaches the target set S_T almost surely from all states under all support-preserving kernels:*

$$\mathbb{P}_s^{\sigma,P}(\tau_T < \infty) = 1 \quad \forall s \in S,\ \forall P : \text{Supp}(P) = \text{Supp}(P^\star).$$

Additionally, the stopping probability is uniformly bounded away from zero:

$$p_T := \min_{s \in S} \inf_{\sigma \in \Sigma^{\text{prop}}} \inf_{P : \text{Supp}(P) = \text{Supp}(P^\star)} \mathbb{P}^{\sigma,P}[\tau_T \le |S| \mid s_0 = s] > 0.$$

This condition strengthens properness by requiring a uniform lower bound on finite-time reachability. It is analogous to standard reachability and diameter assumptions in PAC and reinforcement learning analyses [2, 7, 22, 24], and to the properness conditions used in SSP formulations [5]. It is required to obtain the geometric tail bound in Proposition 23 and the finite effective horizon in Lemma 24. We discuss this assumption and its practical implications further in App. L.

Next, we establish a geometric tail bound on the stopping time in terms of p_T.

Proposition 23. *Under Assumption 3, for all $\sigma \in \Sigma^{\mathrm{prop}}$ and all support-preserving P:*

$$\mathbb{P}^{\sigma,P}[\tau_T > k|S|] \leq (1 - p_T)^k \qquad \forall k \geq 0.$$

Proof. We proceed by induction on k. The base case $k = 0$ is immediate.

For the inductive step, suppose $\mathbb{P}^{\sigma,P}[\tau_T > k|S|] \leq (1 - p_T)^k$. Conditioned on $\{\tau_T > k|S|\}$ we have $s_{k|S|} \notin S_T$.

By the Markov property, the probability of reaching S_T within the next $|S|$ steps depends only on the current state $s_{k|S|}$. By Assumption 3, from any such state the probability of reaching S_T within the next $|S|$ steps is at least p_T, uniformly over the realised state $s_{k|S|}$:

$$\mathbb{P}^{\sigma,P}[\tau_T \leq (k+1)|S| \mid \tau_T > k|S|,\ s_{k|S|}] \geq p_T.$$

Hence,

$$\mathbb{P}^{\sigma,P}[\tau_T > (k+1)|S| \mid \tau_T > k|S|] \leq 1 - p_T.$$

Therefore,

$$\mathbb{P}^{\sigma,P}[\tau_T > (k+1)|S|] = \mathbb{P}^{\sigma,P}[\tau_T > (k+1)|S| \mid \tau_T > k|S|] \cdot \mathbb{P}^{\sigma,P}[\tau_T > k|S|] \leq (1 - p_T)^{k+1},$$

where the final inequality uses the inductive hypothesis. □

Under Assumption 3, we have $\tau_T < \infty$ almost surely under all proper policies and support-preserving kernels, so the reachability reward is finite. Moreover, the assumption provides a uniform lower bound $p_T > 0$ on the probability of reaching S_T within $|S|$ steps from any state; this quantity can be computed by solving a reachability RMDP for S_T. This induces an *effective horizon*.

Effective horizon. Define the effective horizon as

$$H_{\mathrm{eff}} := \sup_{\sigma \in \Sigma^{\mathrm{prop}}} \sup_{P:\, \mathrm{Supp}(P) = \mathrm{Supp}(P^\star)} \mathbb{E}^{\sigma,P}[\tau_T].$$

The following result shows that H_{eff} is finite and admits the bound $H_{\mathrm{eff}} \leq |S|/p_T$.

Lemma 24 (Effective horizon)**.**

$$H_{\mathrm{eff}} \leq \frac{|S|}{p_T} < \infty.$$

Proof. For any $\sigma \in \Sigma^{\mathrm{prop}}$ and support-preserving P,

$$\begin{aligned}
\mathbb{E}^{\sigma,P}[\tau_T] &= \sum_{n=0}^{\infty} \mathbb{P}^{\sigma,P}[\tau_T > n] = \sum_{k=0}^{\infty} \sum_{j=0}^{|S|-1} \mathbb{P}^{\sigma,P}[\tau_T > k|S| + j] \\
&\leq \sum_{k=0}^{\infty} \sum_{j=0}^{|S|-1} \mathbb{P}^{\sigma,P}[\tau_T > k|S|] \leq |S| \sum_{k=0}^{\infty} \mathbb{P}^{\sigma,P}[\tau_T > k|S|] \\
&\leq |S| \sum_{k=0}^{\infty} (1 - p_T)^k = \frac{|S|}{p_T},
\end{aligned}$$

where the last inequality follows from Proposition 23. □

Hence, H_{eff} can be conservatively bounded by $|S|/p_T$ when unknown. In the following, we extend the sensitivity lemma (Lemma 2) to the infinite-horizon setting by replacing H with H_{eff}, which induces only a constant-factor change in Δ_t.

We first prove the stopped analogue of Proposition 11 from the finite-horizon setting (App. C), following the same proof structure.

Proposition 25 (Stopped one-step error propagation)**.** *Let $\mu_h, \widehat{\mu}_h$ denote the stopped marginals at time-step h under P and $\widehat{P}$ respectively, following a proper profile σ from initial state $s_0 \in S$:*

$$\mu_h(s) := \mathbb{P}^{\sigma,P}[s_h = s,\ h < \tau_T], \quad \widehat{\mu}_h(s) := \mathbb{P}^{\sigma,\widehat{P}}[s_h = s,\ h < \tau_T].$$

Define

$$\Delta_t^{L^1} := \max_{(s,a)\in\mathcal{X}_{\mathrm{nt}}} \|P_{sa} - \widehat{P}_{sa}\|_1.$$

Then for all $h \geq 1$,

$$\|\mu_h - \widehat{\mu}_h\|_1 \leq h\, \Delta_t^{L^1}.$$

Proof. We prove the recursive bound

$$\|\mu_{h+1} - \widehat{\mu}_{h+1}\|_1 \leq \|\mu_h - \widehat{\mu}_h\|_1 + \Delta_t^{L^1},$$

from which the result follows by induction using $\mu_0 = \widehat{\mu}_0 = \delta_{s_0}$.

Define the stopped one-step kernels, for $s \notin S_T$,

$$K_s := \mathbb{1}[\cdot \notin S_T] \sum_a \sigma(a|s) P(\cdot|s,a), \qquad \widehat{K}_s := \mathbb{1}[\cdot \notin S_T] \sum_a \sigma(a|s) \widehat{P}(\cdot|s,a).$$

These are substochastic, i.e., $\sum_{s'} K_s(s') \leq 1$.

Then

$$\mu_{h+1} = \sum_s \mu_h(s) K_s, \qquad \widehat{\mu}_{h+1} = \sum_s \widehat{\mu}_h(s) \widehat{K}_s,$$

so

$$\mu_{h+1} - \widehat{\mu}_{h+1} = \sum_s (\mu_h(s) - \widehat{\mu}_h(s)) K_s + \sum_s \widehat{\mu}_h(s)(K_s - \widehat{K}_s).$$

Taking L^1 norms and applying the triangle inequality:

$$\|\mu_{h+1} - \widehat{\mu}_{h+1}\|_1 \leq \Big\| \sum_s (\mu_h - \widehat{\mu}_h) K_s \Big\|_1 + \Big\| \sum_s \widehat{\mu}_h(s)(K_s - \widehat{K}_s) \Big\|_1.$$

First term. Substochasticity implies

$$\Big\| \sum_s (\mu_h - \widehat{\mu}_h) K_s \Big\|_1 \leq \|\mu_h - \widehat{\mu}_h\|_1.$$

Second term. By convexity,

$$\Big\| \sum_s \widehat{\mu}_h(s)(K_s - \widehat{K}_s) \Big\|_1 \leq \max_{s \notin S_T} \|K_s - \widehat{K}_s\|_1.$$

For each $s \notin S_T$,

$$\|K_s - \widehat{K}_s\|_1 \leq \sum_a \sigma(a|s) \|P_{sa} - \widehat{P}_{sa}\|_1 \leq \Delta_t^{L^1}.$$

Substituting gives

$$\|\mu_{h+1} - \widehat{\mu}_{h+1}\|_1 \leq \|\mu_h - \widehat{\mu}_h\|_1 + \Delta_t^{L^1}.$$

Unrolling yields

$$\|\mu_h - \widehat{\mu}_h\|_1 \leq h\, \Delta_t^{L^1}.$$

□

We now lift Proposition 25 to derive the infinite-horizon counterpart of the sensitivity lemma (Lemma 2).

Lemma 26 (Infinite-horizon sensitivity)**.** *Under Assumption 3, define*

$$\Delta_t := R_{\max} H_{\text{eff}}^2 \Delta_t^{L^1}, \qquad H_{\text{eff}} := \left\lceil \frac{|S|}{p_T} \right\rceil.$$

Then for any $P \in \mathcal{P}_t$, profile $\sigma \in \Sigma^{\text{prop}}$, player $i \in N$ and state $s_0 \in S$,

$$|u_i(\sigma, P \mid s_0) - u_i(\sigma, \widehat{P} \mid s_0)| \le \Delta_t.$$

Proof. Let $V_P(s) := u_i(\sigma, P \mid s)$ and $\widehat{V}(s) := u_i(\sigma, \widehat{P} \mid s)$, with $V_P(s) = \widehat{V}(s) = 0$ for $s \in S_T$. For $s \notin S_T$, let

$$P_s^\sigma := \sum_a \sigma(a \mid s) P(\cdot \mid s, a), \qquad \widehat{P}_s^\sigma := \sum_a \sigma(a \mid s) \widehat{P}(\cdot \mid s, a).$$

Then

$$V_P(s) = r_i^\sigma(s) + \sum_{s'} P_s^\sigma(s') V_P(s'),$$

and analogously for $\widehat{V}$. Since $|\tau_T|$ has expectation at most H_{eff} under Assumption 3,

$$\|V_P\|_\infty, \|\widehat{V}\|_\infty \le R_{\max} H_{\text{eff}}.$$

Moreover, for every $s \notin S_T$,

$$\|P_s^\sigma - \widehat{P}_s^\sigma\|_1 \le \sum_a \sigma(a \mid s) \|P_{sa} - \widehat{P}_{sa}\|_1 \le \Delta_t^{L^1}.$$

Hence, for $s \notin S_T$,

$$|V_P(s) - \widehat{V}(s)| \le R_{\max} H_{\text{eff}} \Delta_t^{L^1} + \sum_{s'} P_s^\sigma(s') |V_P(s') - \widehat{V}(s')|.$$

Unrolling this inequality until τ_T gives

$$|V_P(s_0) - \widehat{V}(s_0)| \le R_{\max} H_{\text{eff}} \Delta_t^{L^1} \, \mathbb{E}_{s_0}^{\sigma, P}[\tau_T] \le R_{\max} H_{\text{eff}}^2 \Delta_t^{L^1} = \Delta_t.$$

□

Lemma 26 recovers Lemma 2 (up to constant factors) with the substitution $H \leftarrow H_{\text{eff}}$. All intermediate results in the finite-horizon case carry over with $H \leftarrow H_{\text{eff}}$ and $n_{\min}$ as stated in Theorem 27 below.

Theorem 27 (PAC guarantee for infinite-horizon objectives)**.** *Alg. 1 with*

$$n_{\min} = \left\lceil -\frac{64 R_{\max}^2 |S|^4}{\varepsilon^2 p_T^4} W_{-1}\left(-\frac{\varepsilon^2 p_T^4}{64 R_{\max}^2 |S|^4} \sqrt{\frac{\delta^{cont}}{2(2^{|S|} - 2)|S||A|}} \right) \right\rceil = \widetilde{\mathcal{O}}\left(\frac{R_{\max}^2 |S|^5}{\varepsilon^2 p_T^4} \right)$$

satisfies the PAC-CSG guarantee of Definition 4 under the unbounded probabilistic reachability and reachability reward objectives, with sample complexity

$$T^\star = \widetilde{\mathcal{O}}\left(\frac{R_{\max}^2 |S|^6 |A|}{\varepsilon^2 p_T^4 p_{\text{reach}}} \right).$$

I Extension: Learning zero-sum CSGs

We now specialise the framework to two-player zero-sum CSGs. Since the players have directly opposing objectives, i.e., $X_1 = -X_2$, we write $X := X_1$, so that $X_2 = -X$. Accordingly, we consider a single utility and reward function, writing $u := u_1 = -u_2$ and $r := r_1 = -r_2$.

In a zero-sum CSG $\mathcal{G}$, the (minimax) *value* of $\mathcal{G}$ with respect to X is defined whenever the game is *determined*, i.e., when

$$\sup_{\sigma_1 \in \Sigma_1} \inf_{\sigma_2 \in \Sigma_2} u(\sigma_1, \sigma_2) = \inf_{\sigma_2 \in \Sigma_2} \sup_{\sigma_1 \in \Sigma_1} u(\sigma_1, \sigma_2).$$

In this case, the value is unique, and the corresponding optimal strategies form the set of Nash equilibria of $\mathcal{G}$ [45].

Extending to robust CSGs, let $\mathcal{G}$ be an RCSG with uncertainty set $\mathcal{P}$. We say the game is *robustly determined* if determinacy holds under worst-case transition uncertainty, i.e., with respect to $\underline{u}_{\mathcal{P}}(\sigma) = \inf_{P\in\mathcal{P}} u(\sigma, P)$. The resulting *robust value* of $\mathcal{G}$ is

$$\underline{V} := \sup_{\sigma_1} \inf_{\sigma_2} \inf_{P\in\mathcal{P}} u(\sigma, P),$$

which coincides with the value of the induced CSG under an adversarial kernel (cf. He and Parker [18, Definition 6]). The corresponding *robust optimal strategies* guarantee a value of at least $\underline{V}$ for all $P \in \mathcal{P}$.

For our objectives of interest, zero-sum CSGs are always determined in the finite-horizon setting, and optimal strategies exist due to the finite game tree. In contrast, in the infinite-horizon setting, a value exists (e.g., in the limit-sure sense for reachability objectives), but optimal strategies need not exist [13].

Simplifications relative to the general-sum case. Since only a single utility function u is involved, the sensitivity (Lemma 2) and per-episode value gap (Corollary 5) from the general-sum case apply without the factor-of-2 loss incurred when summing over players. Consequently, the stopping threshold on Δ_t improves from $\varepsilon/4$ to $\varepsilon/2$ (line 9 in Alg. 1), and the constant in $n_{\min}$ improves accordingly:

$$n_{\min} = \left\lceil -\frac{4R_{\max}^2 H^4}{\varepsilon^2} W_{-1}\left(-\frac{\varepsilon^2}{4R_{\max}^2 H^4} \sqrt{\frac{\delta^{\text{cont}}}{2(2^{|S|}-2)|S||A|}} \right) \right\rceil = \widetilde{\mathcal{O}}\left(\frac{R_{\max}^2 H^4 |S|}{\varepsilon^2} \right).$$

All other bounds remain unchanged. Moreover, the RCSG solver SOLVEL^1CSG reduces to a minimax L^1-CSG solver, which is equivalent to solving an RMDP and admits a polynomial-time solution [35, 37, 43, 48].

J Helper functions for Algorithm 1

The helper routines, Algs. 2 to 4, are called by Alg. 1 but omitted from the main paper for brevity. Note that in Alg. 2, p_{reach} is computed by solving a reachability problem on an RMDP with the same support as $P^\star$, full-simplex transition uncertainty over that support, and S_T as the terminal set; this does not require knowing the transition probabilities. In implementation, we instantiate the support-preserving simplex with a 10^{-5} lower bound on probabilities within the support, which ensures graph preservation and yields a strictly positive p_{reach} consistent with Assumption 1.

Algorithm 2 Initialise

Input: $S, A, \Gamma, R_{\max}, H, \text{Supp}(P^\star)$

1: $p_{\text{reach}} \leftarrow \min_{(s,a)\in\mathcal{X}_{\text{nt}}} \max_{\sigma\in\Sigma} \inf_{P\in\mathcal{P}_{\text{Supp}}} \mathbb{P}^{\sigma,P}[\text{reach } (s,a) \text{ before entering } S_T]$ ▷ Eq. (2)
2: **for** $(s,a) \in \mathcal{X}_{\text{nt}}$ **do**
3: $\quad N(s,a) \leftarrow 0,\ N(s,a,s') \leftarrow 0\ \forall s'$
4: $\quad \mathcal{K}(s,a) \leftarrow false$
5: $\quad \widehat{P}(\cdot \mid s,a) \leftarrow$ Uniform
6: **end for**
7: $c \leftarrow -16R_{\max}^2 H^4/\varepsilon^2$; $\quad n_{\min} \leftarrow \left\lceil c \cdot W_{-1}\left(\sqrt{\delta^{\text{cont}}/(2(2^{|S|}-2)|S||A|)}/c\right)\right\rceil$ ▷ by Cor. 21
8: **return** $N, \mathcal{K}, \widehat{P}, p_{\text{reach}}$

K Additional experimental details

K.1 Exploration Baselines

We give precise definitions of the four exploration strategies compared in RQ2 (Table 2). In each case, the resulting sampling policy replaces σ_t^e in Alg. 1; only the choice of exploration policy differs, with the PAC loop, stopping condition $\Delta_t \le \varepsilon/4$, and all other components held fixed.

Algorithm 3 Construct uncertain transition set $\mathcal{P}(s,a)$

Input: $N, \delta^{\text{cont}}, S, A, \Gamma, \text{Supp}(P^\star)$
for $(s,a) \in \mathcal{X}_{\text{nt}}$ **do**
  if $N(s,a) = 0$ **then**
    $\widehat{P}_{sa} \leftarrow \text{Uniform}(\text{Succ}(s,a))$
    $\mathcal{P}(s,a) \leftarrow \Delta_+(\text{Succ}(s,a))$
  else
    $\widehat{P}_{sa}(s') \leftarrow N(s,a,s')/N(s,a) \quad \forall s' \in \text{Succ}(s,a)$
    $\delta^{\text{cont}}_{sa} \leftarrow \delta^{\text{cont}} / \left[|S||A| \cdot N(s,a)(N(s,a)+1)\right]$
    $\alpha \leftarrow \sqrt{(2/N(s,a)) \ln\left[(2^{|S|}-2)/\delta^{\text{cont}}_{sa}\right]}$ ▷ by Lem. 9
    $\mathcal{P}(s,a) \leftarrow \left\{p \in \Delta_+(\text{Succ}(s,a)) : \|p - \widehat{P}_{sa}\|_1 \le \alpha\right\}$
  end if
end for
return $\mathcal{P}$

Algorithm 4 Update

Input: $N, \mathcal{K}, \pi = (\pi_1, \ldots, \pi_{N^\pi}), n_{\min}, S, A, \Gamma$
for $(s,a) \in \mathcal{X}_{\text{nt}}$ **do**
  for $i \leftarrow 1$ **to** N^π **do**
    $N(s,a) \leftarrow N(s,a) + \#(s,a)$ in π_i
    for $s' \in S$ **do**
      $N(s,a,s') \leftarrow N(s,a,s') + \#(s,a,s')$ in π_i
    end for
  end for
  $\mathcal{K}(s,a) \leftarrow (N(s,a) \ge n_{\min})$
end for
return $N, \mathcal{K}$

(i) **Robust (ours).** Solves the exploration RMDP $\mathcal{G}^e_t$ (Definition 5) pessimistically, i.e., $\sigma^e_t \in \arg\max_{\sigma\in\Sigma} \inf_{P\in\mathcal{P}^e_t} \mathbb{E}^{\sigma^e,P}[\cdot]$, maximising worst-case reachability of unknown (s,a)-pairs over the current uncertainty set (see Eq. (1)). This is the exploration rule used by Alg. 1 and underlies our sample-complexity guarantee (Theorem 7).

(ii) **Optimistic.** Solves the same exploration RMDP $\mathcal{G}^e_t$, but with the inner minimisation over $\mathcal{P}^e_t$ replaced by a maximisation, i.e. $\sigma^e_t \in \arg\max_{\sigma\in\Sigma} \sup_{P\in\mathcal{P}^e_t} \mathbb{E}^{\sigma^e,P}[\cdot]$, targeting best-case rather than worst-case reachability. This corresponds to the optimism-in-the-face-of-uncertainty principle common in PAC-MDP exploration, but is not covered by our analysis (Sec. 4.2): an optimistic σ^e_t need not lower-bound the true visitation probability under $P^\star$, so the coverage guarantee of Lemma 6 need not hold.

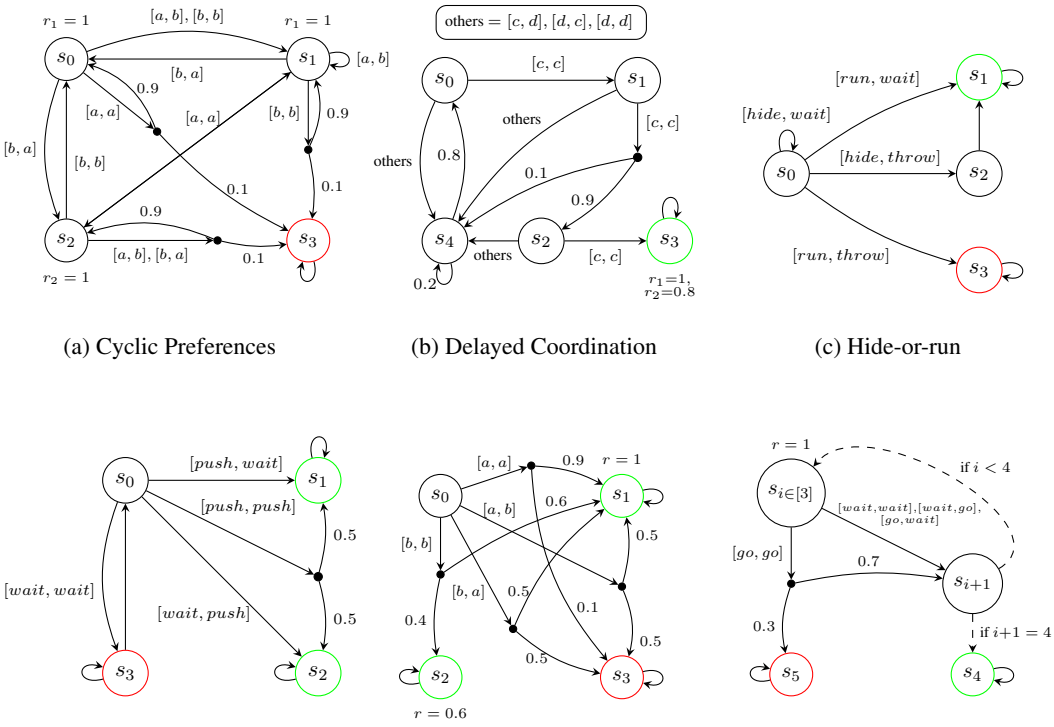
(iii) **Round-robin.** Maintains a single target slot: the unknown (s,a)-pair with the fewest samples so far, ties broken by index. Rather than solving the exploration RMDP, it computes a policy on the *point-estimate* MDP $\widehat{P}_t$ (i.e., the empirical kernel treated as exact, with zero confidence radius) that maximises the probability of reaching the target slot within the horizon, and re-solves only when the target slot changes (i.e., becomes known).

(iv) **Uniform-random.** No exploration policy is solved; at each state, each player samples independently and uniformly at random from their available actions.

All four share the same per-episode sample count N^π_t (Proposition 17) and stopping condition; only the trajectories collected within each episode differ.

K.2 Learning Dynamics

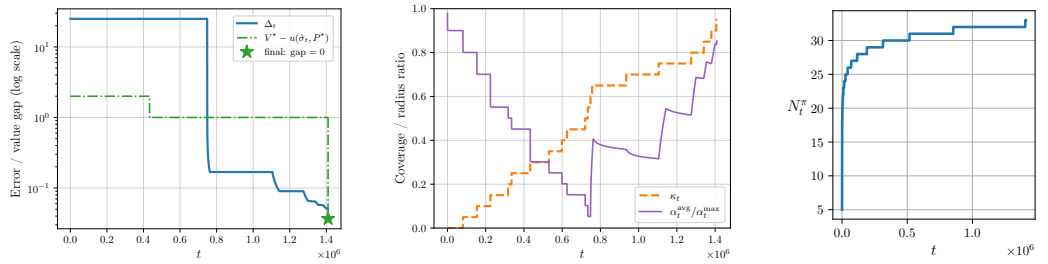
We complement the main experimental evaluation (Sec. 5) with a finer-grained analysis of the learning dynamics induced by our exploration procedure. We ask: *how closely does the model-error proxy Δ_t reflect actual model and decision quality during learning*, and *how does the uncertainty and sampling effort evolve over episodes?*



(a) Cyclic Preferences (b) Delayed Coordination (c) Hide-or-run

(d) Mixed NE (e) Safe vs. Risky (f) Traffic Merge

Figure 2: Six benchmark CSGs ($\bar{s} = s_0$) illustrating distinct challenges for robust PAC learning, including coordination, mixed-strategy equilibria, equilibrium non-existence, and sparse-reward exploration (see Sec. 5). Green states denote target states of the players, while red states denote absorbing trap states.



(a) Δ_t and true value gap (b) κ_t and radius ratio (c) Samples per episode

Figure 3: Learning dynamics over episodes in a representative run of the *Delayed Coordination* game for $\langle\langle p_1 : p_2 \rangle\rangle_{\max=?}(\mathtt{P}[\mathtt{F}\, s_3] + \mathtt{P}[\mathtt{F}\, s_3])$: (a) evolution of Δ_t and the true value gap $V^\star - u(\hat{\sigma}_t, P^\star)$, with $\hat{\sigma}_t$ re-computed every 200 episodes and the final value gap at termination marked by a star; (b) evolution of κ_t and the average-to-maximum confidence-radius ratio; and (c) number of sampled trajectories per episode.

We measure: (i) the model error proxy $\Delta_t := \frac{1}{2} H^2 \max_{(s,a) \in \mathcal{X}_{\text{nt}}} \|\widehat{P}_{t,sa} - P^\star_{sa}\|_1$ ($R_{\max} = 1$); (ii) the fraction $\kappa_t := |\mathcal{K}_t|/|\mathcal{U}_1|$ of known (s, a) pairs by episode t; and (iii) the true value gap $V^\star - u(\hat{\sigma}_t, P^\star)$, where $\hat{\sigma}_t$ is obtained by periodically solving the empirical game every 200 episodes. Together, these quantities allow us to assess how conservative the sufficient stopping condition $\Delta_t \leq \varepsilon/4$ is in practice. We illustrate these dynamics on a representative run of the sparse-reward exploration benchmark, *Delayed Coordination*, for the infinite-horizon property $\langle\langle p_1 : p_2 \rangle\rangle_{\max=?}(\mathtt{P}[\mathtt{F}\, s_3] + \mathtt{P}[\mathtt{F}\, s_3])$: the evolution of Δ_t and the true value gap is shown in Fig. 3a, the evolution of κ_t alongside the

distribution of confidence radii in Fig. 3b, and the number of sampled trajectories per episode in Fig. 3c.

Fig. 3a shows that convergence is not smooth or gradual, especially for the model error Δ_t and the true value gap $V^\star - u(\hat{\sigma}_t, P^\star)$. Rather than decreasing smoothly, Δ_t stays pinned at its worst-case value while a single hard-to-reach (s, a)-pair remains almost entirely unsampled. Since Δ_t is governed by the *largest* per-slot confidence radius rather than the average (Lemma 2), it drops sharply once that slot accumulates enough samples, then follows a staircase pattern as the remaining slots are resolved. Convergence of the extracted strategy is similarly abrupt: $\hat{\sigma}_t$ achieves value 0 (versus the true value $V^\star = 2$) for roughly the first third of training, then rises to and remains at exactly 1 — the value attained by the non-robust point-estimate strategy on most seeds — until the last periodic checkpoint (200 iterations) before Δ_t crosses $\varepsilon/4$. Only in the final handful of episodes does $\hat{\sigma}_t$ resolve to the fully cooperative equilibrium, which *Delayed Coordination* requires at every one of the three consecutive states (s_0, s_1, s_2) to reach the goal s_3 at all (see Fig. 2b). Note that this does not violate Corollary 5's guarantee $V^\star - u(\hat{\sigma}_t, P^\star) \leq 4\Delta_t$: the bound requires $\mu^\star \geq 4\Delta_t - \varepsilon$, and here $\mu^\star = 0$, since all four joint actions at the indifferent state s_4 induce the same transition distribution, so deviating there costs either player nothing and attains the infimum in Definition 3 at 0. Thus, the precondition reduces exactly to $\Delta_t \leq \varepsilon/4$, the algorithm's own stopping condition, so the bound is not expected to hold before the run actually stops, i.e., the regime shown here.

Fig. 3b shows that κ_t increases step-wise but much more smoothly than Δ_t, with no comparably large jump around episode 700–800K, suggesting that the other (s, a)-pairs become known relatively steadily throughout training. κ_t stops at 0.95 (19 of the 20 (s, a)-pairs), indicating that termination is governed by the statistical stopping condition $\Delta_t \leq \varepsilon/4$, which can occur before full coverage (Corollary 22). In contrast, the average-to-maximum confidence-radius ratio $\alpha_t^{\text{avg}}/\alpha_t^{\max}$ decreases for roughly the first half of training as most slots are resolved while the single hardest-to-reach slot keeps $\alpha_t^{\max}$ near its initial value, then rises back towards 1 once that slot is sampled and the remaining radii converge. Together, these trends imply that learning is bottlenecked by a single difficult-to-sample slot: its large confidence radius dominates Δ_t, which reflects the worst-case rather than the average uncertainty across (s, a)-pairs (Lemma 2).

On the other hand, Fig. 3c shows that the number of trajectories sampled per episode, N_t^π, grows logarithmically with the episode index t, matching the adaptive sampling rule $N_t^\omega \propto \log(1/\delta_t^{\text{cov}}) = \mathcal{O}(\log t)$ from Proposition 17 and App. F. This reflects the algorithm's increasingly conservative per-episode coverage requirement. Specifically, as the confidence budget is distributed across an increasing number of episodes, the allowed per-episode failure probability δ_t^{cov} decreases as $\delta_t^{\text{cov}} = \delta^{\text{cov}}/(t(t+1))$ (Proposition 18), so later episodes require more trajectories to maintain their high-probability coverage guarantee, while keeping the total failure probability from coverage at most δ^{cov}.

Overall, these results show that learning is primarily bottlenecked by a small number of hard-to-explore (s, a) pairs: Δ_t and strategy quality improve only once these pairs receive sufficient samples, even when most pairs are already well resolved. The stopping condition $\Delta_t \leq \varepsilon/4$ is therefore conservative in the sense that it is driven by the single worst-case confidence radius rather than decision-relevant uncertainty. To resolve bottlenecks earlier, future improvements could target high-impact, poorly explored regions more directly, e.g., by designing rewards proportional to confidence radii and/or inversely proportional to visitation probabilities.

K.3 Full experimental results

We present the complete set of experimental results in Table 3, in response to RQ1 outlined in Sec. 5 of the main paper.

Licenses for external tools. Our implementation builds on the PRISM-games model checker, which is distributed under the GNU General Public License (GPL), version 2. We adhere to its licensing terms. Additional dependencies and their licenses are documented at Kwiatkowska et al. [25].

Table 3: Full experimental results (mean $\pm$ std over 10 runs). All cases use $\varepsilon = 0.1$, except *Delayed Coordination*'s unbounded-reachability property, which uses $\varepsilon = 0.2$ to keep runtime within a reasonable budget. For every case with a stationary NE returned by the CSG solver, our algorithm achieves the same true-game value, with differences only on action-indifferent states (only s_4 in *Delayed Coordination*). *Hide-or-Run* is a limit-equilibrium case in which the optimal value is approached in the limit but is not attained by any stationary profile. For the third *Safe vs. Risky* property, our learner outputs $\hat{V} = \infty$ in all runs.

Case study	**Property**	**No. of episodes** ($\times 10^6$)	**Total no. of samples** ($\times 10^7$)	$V^\star$	$V^\star - \hat{V}$
Cyclic Prefs.	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{R}_{r_1}[\mathtt{F}\, s_3] + \mathtt{R}_{r_2}[\mathtt{F}\, s_3])$	2.6237 ± 0.0003	8.3317 ± 0.0010	–	–
	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{P}[\mathtt{F}^{\leq 3} s_3] + \mathtt{P}[\mathtt{F}^{\leq 4} s_3])$	2.4997 ± 0.0002	7.9101 ± 0.0007	0.615	0.0021 ± 0.0001
Delayed Coord.	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{P}[\mathtt{F}\, s_3] + \mathtt{P}[\mathtt{F}\, s_3])$	1.4098 ± 0.0000	4.2953 ± 0.0002	2.000	0.0000 ± 0.0000
	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{R}_{r_1}[\mathtt{C}^{\leq 5}] + \mathtt{R}_{r_2}[\mathtt{C}^{\leq 5}])$	5.5241 ± 0.0000	18.3638 ± 0.0002	3.240	0.0016 ± 0.0001
Hide-or-run	$\langle\langle p_1 \rangle\rangle \mathtt{P}_{\max=?} [\mathtt{F}\, s_1]$	0.4671 ± 0.0000	1.3216 ± 0.0000	0.999	0.0000 ± 0.0000
	$\langle\langle p_1 \rangle\rangle \mathtt{P}_{\max=?} [\mathtt{F}^{\leq 5} s_1]$	1.0625 ± 0.0000	3.1834 ± 0.0000	0.800	0.0000 ± 0.0000
Mixed NE	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{P}[\mathtt{F}\, s_1] + \mathtt{P}[\mathtt{F}\, s_2])$	1.4136 ± 0.0000	4.3077 ± 0.0000	1.000	0.0031 ± 0.0000
	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{P}[\mathtt{F}^{\leq 4} s_1] + \mathtt{P}[\mathtt{F}^{\leq 2} s_2])$	1.4136 ± 0.0000	4.3077 ± 0.0000	1.000	0.0031 ± 0.0000
Safe vs. Risky	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{P}[\mathtt{F}\, s_1] + \mathtt{P}[\mathtt{F}\, s_2])$	0.3321 ± 0.0411	5.8710 ± 0.0007	1.000	0.0017 ± 0.0000
	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{P}[\mathtt{F}^{\leq 3} s_1] + \mathtt{P}[\mathtt{F}^{\leq 3} s_2])$	0.1711 ± 0.0066	2.5460 ± 0.0008	1.000	0.0026 ± 0.0000
	$\langle\langle p_1 : p_2 \rangle\rangle_{\max=?} (\mathtt{R}_{r_1}[\mathtt{F}\, s_1] + \mathtt{R}_{r_2}[\mathtt{F}\, s_2])$	0.3321 ± 0.0411	5.8710 ± 0.0007	∞	–
Traffic Merge	$\langle\langle p_1 : p_2 \rangle\rangle_{\min=?} (\mathtt{R}_{t_1}[\mathtt{F}\, s_4] + \mathtt{R}_{t_2}[\mathtt{F}\, s_4])$	0.6712 ± 0.0000	24.3784 ± 0.0003	8.000	0.0000 ± 0.0000
	$\langle\langle p_1 \rangle\rangle \mathtt{P}_{\max=?} [\mathtt{F}^{\leq 5} s_4]$	0.2171 ± 0.0001	3.1098 ± 0.0005	1.000	0.0000 ± 0.0000

L Limitations

We first summarise the assumptions that underpin our analyses, all of which are common in the PAC learning and robust stochastic games literature.

(A1) Centralised exploration. We assume that the players perform *centralised* play during exploration. This assumption is standard in PAC analyses of multi-agent systems, including turn-based stochastic games [7, 42] and Markov games [20, 31], where a centralised learner simplifies the exploration problem. However, it does not capture decentralised or competitive settings where agents act independently or have limited information sharing. We remark that the equilibrium-transfer results in Sec. 4.1 would remain applicable if comparable confidence sets could be obtained under decentralised learning, but the joint-coverage and termination analysis in Secs. 4.2 and 4.3 would require a different exploration mechanism. Developing such a mechanism for decentralised settings remains an important open direction.

(A2) Reachability. Our guarantees rely on a *reachability* condition (Assumption 1) ensuring that every relevant (non-target) (s, a)-pair is reachable under some profile with positive probability at least p_{reach}. This excludes degenerate cases where reachability probabilities vanish and is analogous to standard communicating or reachability assumptions in PAC-MDP analyses [2, 22], adapted to our robust setting. Without such a condition, PAC termination under online trajectories cannot be guaranteed in general, as illustrated by standard lower-bound constructions (e.g., chain-game instances). Relaxing it typically requires additional capabilities such as access to a generative model or structured exploration mechanisms [24], which allow coverage to be enforced explicitly.

For infinite-horizon reachability reward (stochastic shortest path) objectives, we assume almost-sure target reachability (Assumption 3), ensuring $\tau_T < \infty$ almost surely and that value functions are well-defined and finite [5, 27]. We additionally impose a uniform lower bound $p_T > 0$ on the probability of reaching S_T within $|S|$ steps under a proper policy, which yields a geometric tail bound (Proposition 23) and induces a finite effective horizon (Lemma 24). This assumption is consistent with standard reachability and properness conditions in SSP and PAC analyses [2, 7, 22, 24].

In practice, we precompute the set of reachable state–action pairs via standard reachability analysis over the known support graph when solving the game, following established model checking techniques [27]. This allows us to restrict attention to (s, a)-pairs that are reachable under some profile, eliminating vacuous cases where exploration or reaching is impossible. Both probability lower bounds, p_{reach} and p_T, can be computed (or conservatively approximated) by solving associated reachability RMDPs over the known support graph.

(A3) Graph preservation. We assume the standard *graph preservation* condition, whereby all transition kernels in the uncertainty set share the same support. This assumption is common in robust MDP and RCSG formulations [18, 21, 34, 35, 43], and is necessary for tractable robust dynamic programming. However, it excludes settings with support mismatch (e.g., rare or unseen transitions), which may arise in practice when data is sparse.

(A4) Bounded rewards. We assume that the magnitude of rewards is bounded by a known constant $R_{\max}$, which is standard in PAC analyses of reinforcement learning [7, 42]. This assumption ensures well-defined concentration bounds and finite value estimates. In practice, this can typically be satisfied via normalisation.

L.1 Intrinsic limitations

Beyond these assumptions, several intrinsic limitations arise from the problem setting.

Computational complexity. Our approach depends on a *black-box RCSG solver*, instantiated in practice via PRISM-games [18]. The complexity of solving general-sum CSGs is well known to be high: for nonzero-sum reachability objectives, computing subgame-perfect Nash equilibria is PSPACE-complete [8], and practical solvers such as He and Parker [18] can require exponential time in the model size. These complexity results reflect fundamental barriers in equilibrium computation rather than artefacts of our learning framework. Nevertheless, implementation-level optimisations (e.g., parallelised sampling) and increased computational resources can partially mitigate these costs in practice.

Sample complexity dependence. The sample complexity scales polynomially with problem parameters, with a leading H^4 dependence arising from compounding transition uncertainty over trajectories. It may be possible to improve this dependence to H^3 using empirical-Bernstein or stage-wise concentration techniques [15], but such approaches typically require stronger assumptions, such as finer control over variance across timesteps (e.g., martingale-style concentration over trajectories or stage-dependent confidence bounds that exploit variance information at each timestep), and lead to a more technically involved analysis.

Non-existence detection. While our framework supports both equilibrium computation and non-existence detection, the non-existence certificate is *sound but not complete*: when no exact NE exists, the algorithm may still return an ε-NE rather than certifying non-existence. This limitation is inherent to the problem: while ε-NE exist for any $\varepsilon > 0$, this implies that approximate equilibria persist arbitrarily close to the boundary of equilibrium existence, making it fundamentally difficult to distinguish exact equilibrium existence from arbitrarily small Nash margins [6].

Other extensions. Natural extensions include handling more than two players and non-stationary environments. Both introduce additional challenges: the former increases equilibrium complexity and joint-action scaling, while the latter requires adapting to evolving transition kernels and equilibrium structure over time.

M Broader Impacts

This work contributes to the theoretical foundations of learning in multi-agent stochastic systems under uncertainty. Potential positive impacts include improving the robustness and reliability of autonomous systems, such as multi-robot coordination, distributed control, and decision-making in uncertain environments.

At the same time, advances in multi-agent learning and equilibrium computation may have unintended applications in competitive or adversarial settings, such as automated negotiation, financial markets, or strategic resource allocation, where robustness could be exploited to gain advantage. While our work is primarily theoretical and does not target specific deployments, we encourage careful consideration of such downstream uses.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions: a PAC-learning framework for general-sum CSGs with robust equilibrium computation and sound equilibrium non-existence certification, along with theoretical sample complexity guarantees and empirical validation on benchmarks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper provides a consolidated discussion of limitations in App. L, including our assumptions made, complexity and scalability, and the incompleteness of non-existence detection. These are explicitly discussed along with their implications and connections to prior work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are stated under explicit assumptions and supported by complete proofs in the appendix. For results presented in the main text, we include proof sketches there, with full formal statements, detailed proofs, and cross-references provided in Appendix A–H.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the experimental setup (benchmarks, objectives, parameters, randomness sources) in Sec. 5 and provides access to implementation, models, and scripts at `https://github.com/7angel4/prism-games/tree/csg-learning`. Additional details and full results are provided in App. K, enabling reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All models, property specifications, and scripts are provided in a public repository: `https://github.com/7angel4/prism-games/tree/csg-learning`. The implementation builds on the publicly available PRISM-games framework.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all relevant experimental parameters (e.g., ε, δ, $R_{\max}$), benchmark structure, and evaluation protocol, with additional implementation details in App. K.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where appropriate, we report mean results together with standard deviations over 10 runs with different random seeds, shown as error bars in Fig. 1 and as mean $\pm$ standard deviation values in Tables 1 to 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the compute resources in Sec. 5: experiments ran on an Apple M4 Pro with 24 GB RAM, with peak memory usage below 0.5 GB. No specialised hardware was required, and the largest experimental configuration required approximately 5 hours per seed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical and algorithmic, does not involve human subjects or sensitive data, and complies with NeurIPS ethics guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a discussion of broader impacts in App. M, outlining both potential positive applications (e.g., robust multi-agent decision-making systems) and possible risks (e.g., strategic or adversarial use in competitive environments).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets; it proposes theoretical algorithms and evaluation on synthetic benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: The paper's implementation uses the PRISM-games model checker, which is publicly available and open-source. Its licensing terms (GNU GPL v2) are acknowledged and respected, with additional details provided in Appendix K.3.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release an implementation of our PAC-CSG algorithm together with a benchmark suite of concurrent stochastic game models and property specifications at `https://github.com/7angel4/prism-games/tree/csg-learning`. The repository includes documentation (README) describing the models, experimental setup, parameter choices, and instructions for reproducing all results.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No human subjects or crowdsourcing involved.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the methodology or experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.