EPSILON-NASH EQUILIBRIA IN HISTORY-DEPENDENT SA-MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study state-adversarial Markov decision processes (SA-MDP) as a game of observation-space attacks: at each step, an agent selects an action from a received observation while an adversary—who knows the true state the agent is in—chooses a perturbed observation within a state-dependent proximity set. While existing work focuses on Markovian policies, we develop a solution concept and computational approach for SA-MDPs under history dependence. History dependence can materially change equilibrium outcomes and can force both the agent and the adversary to adapt their strategies. First, we prove the non-existence of universal (agnostic of the initial state distribution) history-dependent equilibrium policies. Our main result presents the first algorithmic route to computing ϵ-approximations of initial-state dependent equilibria. We do so by reducing SA-MDPs to a strategically equivalent constrained zero-sum one-sided partially observable stochastic game. We test our algorithm on small analytically verifiable games and showing it scales to larger, more realistic benchmarks, including Atari Freeway rollouts with a 12-period ahead horizon.

1 INTRODUCTION

Reinforcement learning under adversarially manipulated observations poses fundamental challenges that are not captured by standard Markovian robustness models. In this work, we study state-adversarial Markov decision processes with history-dependent policies, characterize their equilibrium structure, and provide the first algorithmic approach for computing approximate history-dependent equilibria.

We consider a state-adversarial Markov decision process (SA-MDP) modeling an observation-space attack game. In each time step t, an *agent* (player 1) takes an action a_1^t based on its observation, while an *adversary* (player 2) simultaneously selects an observation o^t to send to the agent. The adversary knows the true state s^t and must choose o^t from an allowed set $B(s^t)$ (a *proximity constraint* neighborhood of s^t), which limits how far the observation can deviate from reality. The agent receives o^t (but not s^t) and the environment then transitions to a new state s^{t+1} with probability $p(s^{t+1} \mid s^t, a_1^t)$, and a reward $R(s^t, a_1^t)$ is obtained by the agent (the adversary's payoff is the negation). This repeats indefinitely with discount factor $\gamma \in (0, 1)$. The adversary's goal is to minimize the agent's total expected reward by cleverly manipulating observations. Such SA-MDP models have been studied in robust RL and planning settings, usually under the restriction that both players use *Markovian* (memoryless) policies.

Most prior work on SA-MDPs assumes stationary optimal policies depending only on the current state (for the adversary) or observation (for the agent). Under this assumption, the interaction can be formulated as a two-player zero-sum game with a stationary equilibrium policy pair (π_1^*, π_2^*) (for agent and adversary respectively). However, as we argue below, an agent that can condition its strategy on past observations (i.e., a *history-dependent* policy) can potentially achieve higher rewards than any memoryless policy by detecting and exploiting patterns in the observation sequence. We therefore seek a solution concept allowing **history-dependent policies** in SA-MDPs, and develop the first algorithm to compute such an equilibrium.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 Related Work

This paper is inspired by the foundational work of cooperative multi-agent learning, aiming to build adversarially-robust, collaborative autonomous agents (Panait & Luke, 2005; Wang et al., 2017; Cai et al., 2022; Wang et al., 2022a). Early work studied adversarial attacks on reinforcement-learning policies mainly as empirical attack constructions and vulnerability demonstrations: adversarial perturbations to observations can substantially degrade learned policies, and attack timing can matter over trajectories (Huang et al., 2017; Lin et al., 2017); policy-induction attacks during learning were also considered early on (Behzadan & Munir, 2017). The SA-MDP line then formalized worst-case attacks on state observations and robust learning in this setting (Zhang et al., 2020). Follow-up work studied stronger attacks and robust training in closely related fixed-agent, alternating-training, or restricted-policy-class settings, including learned optimal adversaries with alternating training (Zhang et al., 2021), efficient strongest-attack methods (Wang et al., 2022c), and robust/adversarial-training methods such as those of Oikarinen et al. (2021) and Liang et al. (2022).

Several nearby game-theoretic extensions are also relevant (Pendharkar, 2012). Illusory attacks impose information-theoretic detectability constraints on observation attacks (Franzmeyer et al., 2023). Broader online attack/defense formulations consider manipulation of states, observations, actions, and rewards (McMahan et al., 2024; Sarkadi et al., 2019; Motwani et al., 2024; Wang et al., 2022b; Tu et al., 2021; Birmpas et al., 2021). Temporally coupled perturbations have been modeled as a partially observable two-player zero-sum game and solved approximately via equilibrium computation (Liang et al., 2024; Khakpour & Parker, 2025). In multi-agent settings, state-adversarial Markov games can fail to admit standard optimal-policy or robust-Nash solution concepts (Han et al., 2022).

Our algorithmic approach also builds directly on the planning and game-solving literature: partially observable Markov decision processes (POMDPs) provide the underlying partial-observability formalism on the agent side (Kaelbling et al., 1998), Heuristic search value iteration (HSVI) supplies the core search/value-iteration template (Smith & Simmons, 2004), and Horák et al. (2023) gives the immediate zero-sum one-sided POSG framework and solver that our reduction and algorithm build on. Relative to this literature, our contribution is to study observation-space SA-MDPs with history-dependent policies, prove that universal history-dependent state-robust policies need not exist, and give a reduction-based route to computing ϵ-approximate equilibria for a given initial state distribution.

1.2 Background and Notation

For any set X, let $\Delta(X)$ denote the set of probability distributions over X. Notation of variables is that x_i^t indicates the ith player, ith entry, or ith index of variable x at time or game-stage t. Any absent notation is to be understand as the vector of the variable for all i or t in the relevant set of potential outcomes. This split notation is to prevent confusion between a time subscript of $2t$ and "$2, t$", which both appear in the paper.

Zhang et al. (2020) introduce the state-adversarial Markov decision process (SA-MDP) to model an RL agent (player 1) interacting with an environment under adversarial observation perturbations (player 2). An SA-MDP can be defined as a tuple $M = (S, A, R, p, \gamma, B)$, where S is the true state space, A the agent's action space, $R : S \times A \to \mathbb{R}$ the reward function, $p : S \times A \to \Delta(S)$ the state transition probabilities, $\gamma \in (0, 1)$ the discount factor, and B the function that specifies the allowable perturbations of the adversary. In particular, for each true state $s \in S$, $B(s) \subseteq S$ is the set of states that the *adversary* can present to our *agent* as a perturbed observation. We denote the adversary's (possibly stochastic) policy by $\nu : S \to \Delta(S)$, where $\nu(s) \in \Delta(B(s))$. At each time step, the agent observes $\tilde{s} \sim \nu(s)$, and then selects an action $a \in A$ according to its policy $\pi : S \to \Delta(A)$ (which is applied to the observation $\tilde{s}$). The environment then transitions to the next true state s' drawn from $p(s'|s, a)$, and a reward $R(s, a)$ is emitted. Notably, the adversary influences the agent's perception but *does not alter* the underlying state transition; s' depends on the actual s and a taken, not directly on $\tilde{s}$. The adversary's objective is assumed to be worst-case in that it aims to minimize the agent's total return (e.g., by choosing perturbations that lead the agent to make suboptimal decisions). This worst-case observation model cannot be captured by a standard partially-observable MDP, since a POMDP uses fixed observation probabilities (Kaelbling et al., 1998), whereas here the observation distribution observed by the agent depends on the adversary's policy.

Formally, we can define the value of a given stationary agent policy π under a specific adversary ν starting from state s as

$$V^{\pi,\nu}(s) = \mathbb{E}_{\pi,\nu}\Big[\sum_{t=0}^{\infty} \gamma^t R(s^t, a^t) \,\Big|\, s^0 = s\Big],$$

where $a^t \sim \pi(\cdot \mid \tilde{s}^t)$, $\tilde{s}^t \sim \nu(s^t)$, and $s^{t+1} \sim p(\cdot \mid s^t, a^t)$. Of particular interest is the optimal adversary for a given π, ν^*_π. We define the optimal adversarial value function as:

$$V^{\pi,\nu^*_\pi}(s) = \min_\nu V^{\pi,\nu}(s).$$

In an SA-MDP, the agent faces a two-player zero-sum game against the adversary, who chooses the worst observable state at each step. Zhang et al. (2020) show that for a fixed π, the value function under the optimal adversary ν^*_π satisfies a robust Bellman equation. Specifically, letting $\bar{V}_\pi(s) = V^{\pi,\nu^*_\pi}(s)$ denote the value under the strongest adversary, we have:

$$\bar{V}_\pi(s) = \min_{\tilde{s}\in B(s)} \sum_{a\in A} \pi(a \mid \tilde{s}) \cdot \sum_{s'\in S} p(s' \mid s, a)\Big[R(s,a) + \gamma \bar{V}_\pi(s')\Big]. \tag{1}$$

which defines a contraction operator on value functions. Intuitively, at state s the adversary chooses a perturbation $\tilde{s} \in B(s)$ that minimizes the expected return and expected distribution over future values, knowing the agent will act according to π on $\tilde{s}$. This worst-case Bellman operator allows "policy evaluation" under the optimal adversary.

However, in order to consider history dependent policies, we introduce another game, the Zero-Sum One-Sided Partially Observable Stochastic Game, formalized by Horák et al. (2023).

Definition 1.1 (Zero-Sum One-Sided Partially Observable Stochastic Game)**.** A Zero-Sum One-Sided Partially Observable Stochastic Game (or zero-sum OS-POSG) is a tuple $G = (S, A_1, A_2, O, T, R, \gamma)$ where:

- S is a finite set of game states,
- A_1 and A_2 are finite sets of actions of player 1 and player 2, respectively,
- O is a finite set of observations
- $T(\cdot|s, a_1, a_2) \in \Delta(O \times S)$ represents probabilistic transition function for every $(s, a_1, a_2) \in S \times A_1 \times A_2$,
- $R : S \times A_1 \times A_2 \to \mathbb{R}$ is player 1's reward function,
- $\gamma \in (0, 1)$ is a discount factor.

The play unfolds as follows. First, an initial state $s^0 \sim b^{\text{init}}$ is drawn from the probability distribution b^{init}. Then, for each stage, the current state s^t is revealed to player 2 but remains hidden from player 1. Simultaneously, player 1 chooses $a^t_1 \in A_1$ and player 2 chooses $a^t_2 \in A_2$. An unobserved reward $R\big(s^t, a^t_1, a^t_2\big)$ is awarded to player 1, while player 2's payoff is $-R\big(s^t, a^t_1, a^t_2\big)$. The system transitions to a new state s^{t+1} and emits an observation o^t according to $T\big(o^t, s^{t+1} \mid s^t, a^t_1, a^t_2\big)$. Player 2 then observes the full outcome of the stage: $\big(s^t,\, a^t_1,\, a^t_2,\, o^t\big)$. Player 1, by contrast, sees only $\big(a^t_1,\, o^t\big)$, without learning a^t_2, s^t, or s^{t+1}.

2 Theoretical results

In Section 2.1, we show the non-existence of universal Nash equilibria, i.e. policies independent of the initial state distribution. Assuming an initial state distribution is given, in Section 2.2, we show the importance of history-dependent policies, arguing that agents are incentivized to use them, and adversaries must then modify their policies accordingly. We then show in Section 2.3 how to reduce the SA-MDP problem to a modified version of the zero-sum OS-POSG, which allows us to find an epsilon-approximate Nash equilibrium for the original SA-MDP. This equilibrium allows us to find policies that guarantee near optimal rewards.

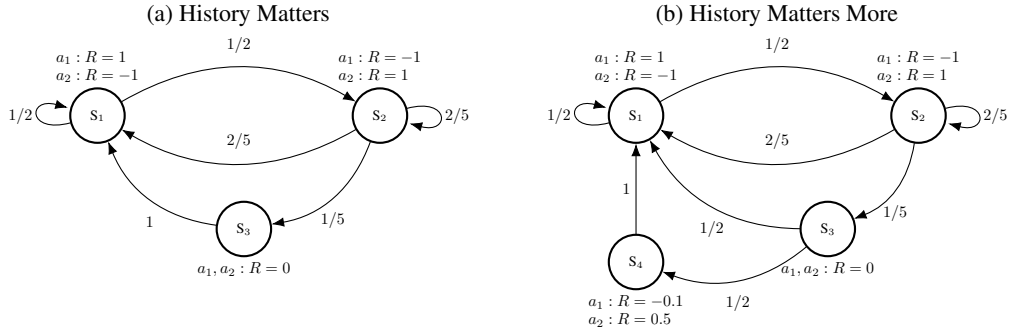


Figure 1: Examples illustrating the impact of history-dependent policies.

2.1 Non-Existence of Universal History Dependent Nash Equilibrium Policies

In a zero-sum game, the agent's optimal strategy can be defined in two equivalent ways: as a robust policy that maximizes value against the worst-case adversary, or as part of a Nash equilibrium (NE) of the game. We show that these notions coincide for SA-MDPs.

Lemma 2.1 (Robust Optimality $\iff$ Nash Equilibrium)**.** *Consider a zero-sum SA-MDP with finite state, action, and observation sets, nonempty perturbation sets $B(s)$, bounded rewards, a fixed initial distribution $b \in \Delta(S)$, and unrestricted full-history behavioral strategy spaces Σ_1 and Σ_2 for the agent and the adversary, respectively. We assume perfect recall in which each player's private history records all previously observed information and all of that player's previous actions. For a strategy profile $(\sigma, \tau) \in \Sigma_1 \times \Sigma_2$, let $V(b; \sigma, \tau)$ denote the agent's expected return for the horizon under consideration when $s^0 \sim b$; when the horizon is $N < \infty$, we also write $V^N(b; \sigma, \tau)$ to make the horizon explicit. For a finite horizon, we allow $\gamma \in (0, 1]$, with $\gamma = 1$ corresponding to the undiscounted return; for an infinite discounted horizon, we require $\gamma \in (0, 1)$. In either case, the following equality holds for every initial distribution $b \in \Delta(S)$:*

$$\arg\max_{\sigma \in \Sigma_1} \min_{\tau \in \Sigma_2} V(b; \sigma, \tau) = \left\{ \sigma \in \Sigma_1 : \exists \tau \in \Sigma_2 \text{ with } (\sigma, \tau) \text{ a Nash equilibrium when } s^0 \sim b \right\}.$$

Both sets are nonempty.

See proof in Appendix A.

We use this lemma to demonstrate the impossibility of finding an algorithm that always returns policies Nash equilibrium agnostic of the initial state distribution. Zhang et al. (2020) showed that there exist SA-MDPs where there is no Markovian optimal agent policy that is independent of the initial state distribution b. They leave open the question of whether the same holds for history dependent policies. We revisit their concrete example and show that no history dependent policy can achieve the minimax value against all adversarial responses. According to the previous lemma, this implies that no initial-state-distribution-agnostic Nash equilibrium exists.

Theorem 2.2. *There exists a SA-MDP such that no history-dependent policy is state-robust optimal.*

See proof in Appendix A.

2.2 Why History Matters

Most models assume Markovian (memoryless) policies for both the agent and the adversary, meaning both the adversary's and agent's actions depend only on the current state or observation and not on past history. This yields a tractable two-player zero-sum framework with stationary, optimal policies. However, a purely Markovian agent might miss opportunities that arise from factoring in the history of play. Consider Figure 1a where the transition probabilities do not depend on the actions, the rewards are inverted for state s_1 and s_2, state s_3 deterministically transitions to state s_1 without emitting a reward, and $B(s_1) = B(s_2) = \{s_1, s_2\},\ B(s_3) = \{s_3\}$.

This setting is already complex enough for it to become difficult to analytically determine the best Markovian policy, not to mention the best history dependent policy. However, we can assume

that in a Markovian equilibrium, the adversary would focus its manipulations on s_1 and s_2, as an incorrect action in s_3 or s_4 simply has less of a negative effect on the agent. In a history dependent setting, however, the presence of s_4 allows the adversary to reduce the certainty the agent had in the previous example of transitioning from s_3 to s_1. Thus, we would expect the adversary to increase the misreports between s_1 and s_4 in a history dependent environment. The inherent complexity of even this small MDP and the interplay between agent and attacker policy motivate the following study of an algorithm to find approximate history dependent Nash equilibria.

Allowing for history dependent strategies vastly increases the strategy space and brings new analytical challenges. As shown in Theorem 2.2, history dependent state-robust Nash equilibrium policies (i.e. policies that are optimal for every state) do not always exist. Therefore, we seek a Nash equilibrium strategy profile for a given initial state distribution. To compute such an equilibrium under history dependence, we modify an algorithm developed by Horák et al. (2023). The authors introduce an HSVI-based solver for one-sided partially observable stochastic games, where one player (the attacker) has full state information and the other (the agent) has only partial observability. This algorithm can efficiently approximate equilibrium strategies even in long-horizon problems. Our attacker–agent interaction fits the one-sided paradigm: the attacker knows the true state, while the agent only sees perturbed observations. By leveraging Horak's algorithm, we can obtain a history dependent equilibrium policy pair for our game given the initial state distribution.

2.3 Modified Zero-Sum OS-POSG to Find Epsilon-Equilibria

We now state one of our main theorems: that any SA-MDP can be transformed into a strategically equivalent zero-sum one-sided POSG.

Theorem 2.3. *Any SA-MDP G_{seq} can be transformed into a strategically equivalent constrained zero-sum one-sided POSG $\widehat{G}$ in which both players act simultaneously at every stage. The solution strategy to $\widehat{G}$ can be mapped back to a unique solution strategy of G_{seq}.*

Proof. For space reasons, the formal mathematical objects and proof are in the appendix. The key underlying idea in the transformation is to encode each sequential SA-MDP as two simultaneous micro-stages while removing all non-decisive degrees of freedom. We introduce alternating dummy actions for both the agent and the adversary. Thus, each player will still have to choose an action at every stage, but will only need to make meaningful decisions every other stage, maintaining the nature of sequential decisions within a simultaneous move game.

In order to accommodate the dummy actions, we also introduce dummy states, which act as placeholders while the agent takes the meaningful action that will actually inform the transition probabilities.

The remaining parts of the proof consist of showing that this altered representation does not affect the set of feasible actions, that the mapping between the two strategy spaces corresponding to the original and new sequential games is a bijection, that payoffs are preserved, and that an epsilon NE of the original game directly maps via our bijection on strategies to an epsilon NE of the new game.

See the full proof on page 16. □

2.4 Constrained Zero-sum OS-POSG

Definition 2.4 (Information-compatible constrained zero-sum OS-POSG)**.** An information-compatible constrained zero-sum one-sided partially observable stochastic game is a tuple

$$G = (S, A_1, A_2, O, T, R, \gamma, L_1, L_2),$$

where $(S, A_1, A_2, O, T, R, \gamma)$ is a finite discounted zero-sum one-sided POSG, player 1 is the uninformed maximizing player, and player 2 is the fully informed minimizing player.

For player 2,

$$L_2 : S \to 2^{A_2} \setminus \{\emptyset\}$$

may depend on the true state. For player 1, the feasible action set is required to be measurable with respect to player 1's information: whenever two states s, s' may occur at the same player-1 information history,

$$L_1(s) = L_1(s').$$

Equivalently, at every reachable player-1 information history h_1, there is a nonempty set $L_1(h_1) \subseteq A_1$ that is common to all states compatible with h_1.

A behavioral strategy is feasible if its support is contained in the corresponding feasible action set at every information history.

The game is played as before, with the additional requirement that, at state s, player 1 chooses $a_1 \in L_1(s)$, and player 2 chooses $a_2 \in L_2(s)$.

Theorem 2.5. *The HSVI algorithm of Horák et al. (2023) extends to the constrained zero-sum one-sided POSG in Definition 2.4 by restricting each local stage game to the players' feasible strategy sets. The algorithm maintains a lower-bound value function V_{LB} and an upper-bound value function V_{UB}, written as V_{LB}^{Γ} and V_{UB}^{Υ}, respectively, in Algorithm 2, satisfying*

$$V_{LB}(b) \leq V^*(b) \leq V_{UB}(b)$$

for every reachable belief b, where V^* is the optimal value function of the constrained game.

Let

$$\delta = \frac{U_0 - L_0}{2},$$

where L_0 and U_0 are the discounted minimum and maximum feasible reward bounds. If $\delta = 0$, the initial bounds coincide. If $\delta > 0$ and

$$0 < D < \frac{(1-\gamma)\epsilon}{2\delta},$$

then the algorithm terminates at the initial belief b_0 with

$$V_{UB}(b_0) - V_{LB}(b_0) \leq \epsilon.$$

The strategy obtained from the lower-bound value function for player 1 and the strategy obtained from the upper-bound value function for player 2 are feasible and form an ϵ-Nash equilibrium. In particular, the result applies to the constrained OS-POSG $\widehat{G}$ constructed in Theorem 2.3.

See proof in section A.

The termination condition deteriorates as $\gamma \to 1$. Under the uniform reward bounds used here,

$$\delta = \frac{R_{\max} - R_{\min}}{2(1-\gamma)},$$

so the sufficient condition may be written as

$$D < \frac{(1-\gamma)^2\epsilon}{R_{\max} - R_{\min}}.$$

This is a worst-case sufficient bound and becomes conservative for discount factors close to one; we discuss its practical implications and our finite-horizon use of $\gamma = 0.999$ when we discuss the limitations of our algorithm in Section 3.3.

2.5 Algorithm

Using our representation of a history-dependent SA-MDP as a constrained zero-sum OS-POSG, we devise an algorithm to solve our game. The algorithm first converts a given SA-MDP game (as encoded in a custom text file format) into an equivalent constrained OS-POSG with phase-dependent singleton feasible action sets and adversarial proximity constraints. Given this constrained OS-POSG, we then leverage Horák et al. (2023) with HSVI, modified to respect state-dependent feasibility constraints for both players. Further details of the algorithm are provided in Appendix D.

3 Experiments

3.1 Example SA-MDP Where History Matters

We construct an example history-dependent SA-MDP with analytically verifiable optimal strategies to explicitly find how much worse off a player facing a history-dependent adversary is and to provide a benchmark setting to test our algorithm.

The game consists of states $S = \{s_0, s_1, s_2, s_3\}$, actions $A_1 = \{A, B, C\}$, and a time horizon $H = 2$. Adversaries can misreport only in restricted ways, where the restriction on the support of misreporting strategies is given by $B(s) \in 2^S \setminus \emptyset$ for state $s \in S$. For our problem $B(s_1) = \{s_1, s_2\}$, $B(s_2) = \{s_2, s_3\}$ and $B(s_3) = \{s_3, s_1\}$. Rewards are sparse and are only realized at the end of the game. The full game is shown in Figure 2.

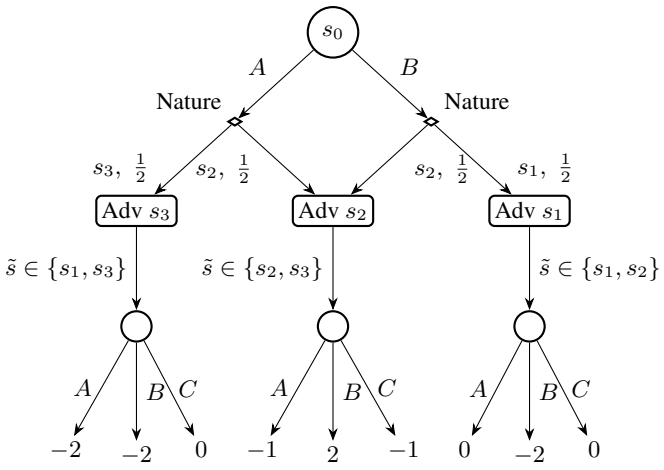


Figure 2: Finite horizon version of SA-MDP game tree (action C at s^0 yields -10 and so omitted as never chosen): the agent chooses A or B at s^0, Nature draws the second-stage state, the adversary reports a neighbor state $\tilde{s} \in B(s)$, and the agent chooses A, B, or C for the terminal payoff.

An analysis of the game in Appendix C demonstrates that the value associated with Markovian strategies for both the agent and the adversary $V_{\text{Markov}} = 0.25$, while for a history-dependent agent and adversary, we get $V_{\text{Hist}} = 0$. This means that the agent is strictly worse off facing a history-dependent adversary, even after being allowed history-dependent strategies.

There are multiple equilibrium strategies for each player. Our code successfully recovers one of these equilibrium strategies that defines the boundary of the region containing optimal strategies under a barrier-point linear program solver for each stage game. Full recovered strategies are shown visually in Figure 3 and written out explicitly in Appendix B.

Our algorithm returns an upper and lower bound that contain the true value for the history-dependent strategies that are less than tolerance ϵ apart on convergence. The midpoint of these upper and lower bounds for history-dependent strategies is given by $\hat{V}_{\text{Hist}} = 0 \pm O(10^{-8})$, compared to a known analytic value of 0, for a convergence tolerance $\epsilon = 0.01$. Because we use $\gamma \sim 1$ ($\gamma = 1$ breaks HSVI), even with exact precision we would have small numerical error left over for an arbitrary problem. In practice, over moderate horizons, the difference is negligible. In comparison, our Markovian game solver finds a solution with a numerical tolerance of $O(10^{-6})$, based on the numerical precision of the linear programming solver. The final solution found via the linear solver is $\hat{V}_{\text{Markov}} = 0.2496 \pm O(10^{-6})$, compared to a true analytical value for $\gamma = 0.999 \sim 1$ of 0.2496 and 0.25 for the exact $\gamma = 1$.

3.2 Larger Problems

We test our algorithm up to depth 10 with 4 actions per node on randomly generated trees with 3.1 million states and 1.4 million information sets. Solving takes 2200s on a single MacBook M4 Pro CPU core and involves 18,000 explorations of the tree. We also tested our algorithm with 5 different seeds for games of sizes 2, 4, 6, and 8. We get a roughly log-linear plot in the amount of solve time per additional layer of depth. Roughly every 2 additional horizons increase solve time by 20–30×. See Figure 6 and Table 1 for detailed scaling information and Appendix D for further discussion.

Finally, we practically test our algorithm on Atari Freeway at depth 12 with a frame-skip of 30. Here, following the procedure in BRIDGE (Laidlaw et al., 2023) we generate an exact enumeration of state-actions reachable within the given horizon and merge behaviorally equivalent states. We found that on the default Atari settings with frameskip 30, agents were able to find an optimal

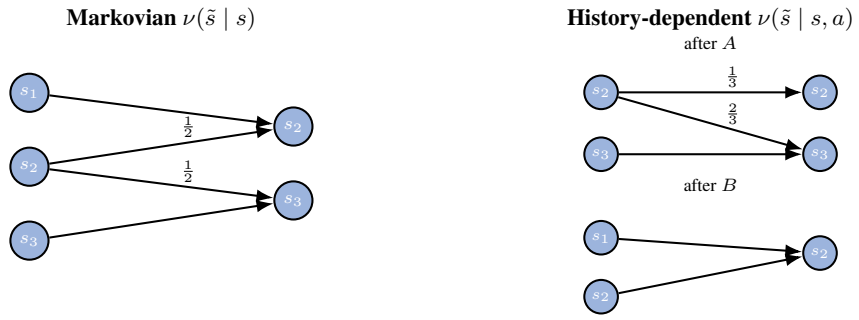


Figure 3: Adversary equilibrium strategies for Markovian vs. history-dependent adversary. Left: Markovian. Right: history-dependent kernels conditioned on player action in state s^0 (A vs. after B).

Table 1: Runtime and exploration scaling by depth over 5 seeds. Explorations are the number required to reach an excess value-function gap below 10^{-2}. Times are rounded to hundredths of a second.

Depth	States	Mean time (s)	Std. dev. (s)	Explorations
2	23	0.00	0.00	6
4	743	0.08	0.02	28
6	12,263	1.48	0.30	257
8	196,583	52.49	21.49	2,564

deterministic strategy independent of actual observations on the initial reset, and thus immune to adversary perturbations. Therefore, we made the Atari initial state "random" by hidden warm-up uncertainty and an initial adversary maximal perturbation of the radius in L2 distance as 500 (about 2.73 grayscale levels per pixel if spread uniformly), which permits multiple reports. The two initial states were generated by "warming-up" the Freeway environment and letting it run initially for a random length of time. We then searched across these possible random initial states until we found two whose distance was less than 500 apart from each other (their initial screen distance is 486.7018). This ensured that these two initial states are confusable. We then placed an equal probability of the agent starting in either of the two initial states in the Atari simulation. Here, $V_{\text{HD,HD}}$ lies below the no-adversary value and exceeds the approximate Markov FP value, showing history-dependent strategies matter and help the agent here. Atari Freeway is depth 12 but takes just 2.07s to solve for optimal history-dependent adversary and agent policies. This is because the number of info partitions in the OS-POSG is small after this behavioral consolidation (5,242 partitions versus 87,375 for depth 8, action 4 game).

In this Atari Freeway setting, history matters, and we find a history-dependent value of 0.8167225 (0.816584–0.816861 lower and upper value bounds from HSVI after converging under 0.01 excess gap; a reported excess gap of $\sim 2 \times 10^{-4}$) for both the agent and the adversary exhibiting history-dependence, compared to 0.856748 for no adversary and 0.758971 for both agents being Markovian. Here, history dependence helps the agent detect manipulations by the adversary, leading to improved performance. Theoretically, however, the direction could go either way.

Larger settings are feasible to scale to, but are limited by the use of HSVI as our algorithm for exact solving; the primary constraint is memory as our depth 10 setting uses peak RAM of 8.6GB.

3.3 Limitations of Algorithm

The primary limitation of our algorithm is that, for general games, the solver wall time scales in a roughly exponential fashion in the maximum length of history-dependence we solve for in the original SA-MDP problem. This is unsurprising, given that the number of histories themselves scale in this fashion and we are aiming to exactly solve the solution. In principle, however, any solution algorithm for general zero-sum OS-POSG games can now be applied. This means that approximate solutions involving depth-limited search, public belief-space methods, flexible neural network-based value function or policy function estimators learned via self-play could be applied to our algorithm, at the cost of our strong guarantees on convergence that we get from utilizing HSVI in exchange for even further scalability. Already, however, we have shown we can tackle SA-MDP

problems of real-world relevance and size, while still retaining the use of HSVI with guarantees on the validity of the solution. A further limitation arises from using a discount factor close to one to represent finite-horizon objectives within the infinite-horizon HSVI framework. In our finite-horizon experiments, we use $\gamma = 0.999$ and absorbing zero-reward states after the terminal depth. This introduces a small difference between the intended undiscounted finite-horizon return and the discounted return optimized by the solver. Thus, although $\gamma = 0.999$ provides a close approximation to the undiscounted short-horizon objective, the corresponding worst-case termination bound can be highly conservative. Our experiments indicate that practical convergence is substantially better than this bound, but the approach remains most suitable for short- to moderate-horizon problems.

We leave it to future work to explore the application of these methods to solve larger SA-MDPs with our approach.

4 Conclusions

We develop theory for history-dependent strategies in SA-MDPs, as well as an algorithm for computing their solutions using a novel reduction combined with HSVI. We show how this can be extended to study constrained adversaries. Finally, we verify in numerical examples that history-dependence quantitatively and qualitatively matters for optimal policies in SA-MDPs and how much an adversary can punish an agent and show that our algorithm successfully recovers Nash equilibrium strategies. This opens the door to scaling our approach to substantially larger state spaces by using compressed belief representations, using approximate solvers for POSG in a public belief space representation of the setting, and horizon truncation at the cost of guarantees and uniform bounds for HSVI.

AI use statement

In this work, we used generative AI tools for assistance with coding, executing experiments, debugging, further discussion about other literature, checking notational consistency, and grammar and the clarification and rewriting of paragraphs. We also used AI tools with repeated rounds of discussion and heavy guidance from the authors specifically in the construction of or expansion of some of the theorems. General proof structure was developed mostly by humans, but some exact proof argument steps were created by AI.

In terms of required disclosures, we used AI at various points: to formulate (some) mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, and interpret results.

For theoretical results our summary is: Lemma 2.1 was extended from a prior human-written argument that was for finite horizons to one for infinite horizons with extensive human guidance, the lemma statement was correspondingly expanded; Theorem 2.2 was human written; Theorem 2.3 was mostly human written but in later revisions some checks on the information structure and ensuring the mapping is valid were performed with the help of AI; Theorem 2.5 was mostly human written, but some of the exact bounds given an ϵ were run with AI and the Horák et al. (2023) portions that substantially reproduce parts of his argument were mostly adapted from the paper locally to our setting and written explicitly by AI (prior versions of the paper used looser formulations like “from Horák et al. (2023) we find” to skip highly repetitive portions of the logic, but earlier feedback asked for a more explicit argument.) The exact values for the example of a SA-MDP Game where history matters were found by a combination of human and automated search. The actual argument and multiple cases to check were performed by AI, with significant human supervision. For theorems that used AI more heavily, we had AI simultaneously develop mostly self-contained Lean 4 programs to check their validity. Lean has no game theory library in Mathlib at time of writing, so some game structures had just the relevant properties defined in Lean rather than explicitly define the whole structure and some prior standard theorem’s were formalized in Lean 4, but did not have their own proof formalized in Lean 4. These Lean 4 programs cleanly compiled and ran without "sorrys" or similar. These Lean files were checked with “Axle MCP” by AxiomMath.

If requested by the reviewers, we are happy to provide the Lean 4 programs used for validation. Therefore, we disclosed that AI was used to polish writing and for proving mathematical claims.

We have not used generative AI tools for literature search, idea generation, significant paper structure, or theorem writing or proving without the human oversight and guidance described above, as well as originally human written theorems. Therefore, tasks that include using AI to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, assist with translation, clean and reformat datasets, and support qualitative and thematic data analysis are not applicable to this work

Additionally, we used generative AI tools to perform the following recommended disclosures to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, and draft parts of a research paper, sourcing/searching for information, edit a research paper to improve readability, and format references. We have reviewed all AI-assisted work to check its accuracy, both the theoretical work and its final generated content whenever sentences were rewritten or similar. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work advances the theoretical and algorithmic foundations of robust reinforcement learning under adversarial observation manipulation by formalizing and computing history-dependent equilibria in state-adversarial MDPs. By showing when history dependence fundamentally alters equilibrium behavior - and providing the first practical method to compute such equilibria - our results improve understanding of decision-making in safety-critical and adversarial environments.

Potential positive impacts include more reliable autonomous systems and stronger robustness guarantees in sequential decision problems. As with many advances in adversarial modeling, these techniques could also be misused to design more effective attacks; however, we believe that making such vulnerabilities explicit is necessary for developing defenses and improving system resilience overall. Finally, the real-world applicability of our current approach is constrained by its bounded scalability.

Reproducibility statement

The proofs of the theoretical results are given in Appendix A. The analytically verifiable example and its recovered strategies are detailed in Appendices B and C. Appendix D provides the algorithm pseudocode and implementation details; experimental settings and runtime measurements are reported in Section 3.

References

Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *International conference on machine learning and data mining in pattern recognition*, pp. 262–275. Springer, 2017.

Georgios Birmpas, Jiarui Gan, Alexandros Hollender, Francisco J. Marmolejo Cossío, Ninad Rajgopal, and Alexandros A. Voudouris. Optimally deceiving a learning leader in stackelberg games. *J. Artif. Intell. Res.*, 72:507–531, 2021.

He Cai, Youfeng Su, and Jie Huang. Cooperative control of multi-agent systems. *Cham: Springer-Verlag*, 2022.

Tim Franzmeyer, Stephen Marcus McAleer, Joao F Henriques, Jakob Nicolaus Foerster, Philip Torr, Adel Bibi, and Christian Schroeder de Witt. Illusory attacks: Detectability matters in adversarial attacks on sequential decision-makers. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, and Fei Miao. What is the solution for state-adversarial multi-agent reinforcement learning? *Trans. Mach. Learn. Res.*, 2022. arXiv:2212.02705.

Karel Horák, Branislav Bošanskỳ, Vojtěch Kovařík, and Christopher Kiekintveld. Solving zero-sum one-sided partially observable stochastic games. *Artificial Intelligence*, 316:103838, 2023.

Sandy H. Huang, Nicolas Papernot, I. Goodfellow, Yan Duan, and P. Abbeel. Adversarial attacks on neural network policies. In *International Conference on Learning Representations, ICLR*, 2017. arXiv:1702.02284.

IBM. User's manual for cplex, 2024. URL `https://www.ibm.com/docs/en/icos/22.1.2?topic=optimizers-users-manual-cplex`. IBM ILOG CPLEX Optimization Studio documentation.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Narges Khakpour and David Parker. Partially-observable security games for attack-defence analysis in software systems. In Alexandre Madeira and Alexander Knapp (eds.), *Software Engineering and Formal Methods*, pp. 144–161, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-77382-2.

Harold W. Kuhn. Extensive games and the problem of information. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 193–216. Princeton University Press, Princeton, NJ, 1953.

Cassidy Laidlaw, Stuart J Russell, and Anca Dragan. Bridging rl theory and practice with the effective horizon. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 58953–59007. Curran Associates, Inc., 2023.

Yongyuan Liang, Yanchao Sun, Ruijie Zheng, and Furong Huang. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yongyuan Liang, Yanchao Sun, Ruijie Zheng, Xiangyu Liu, Benjamin Eysenbach, Tuomas Sandholm, Furong Huang, and Stephen McAleer. Game-theoretic robust reinforcement learning handles temporally-coupled perturbations. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.12062.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 3756–3762. AAAI Press, 2017. ISBN 9780999241103.

Jeremy McMahan, Young Wu, Xiaojin Zhu, and Qiaomin Xie. Optimal attack and defense for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. arXiv:2312.00198.

Sumeet R Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H Torr, Lewis Hammond, and Christian S de Witt. Secret collusion among ai agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems*, 37:73439–73486, 2024.

James R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, 2nd edition, 2000. ISBN 978-0131816299.

Tuomas Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. *Advances in Neural Information Processing Systems*, 34:26156–26167, 2021.

Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, 11(3):387–434, 2005.

Parag C Pendharkar. Game theoretical applications for multi-agent systems. *Expert Systems with Applications*, 39(1):273–279, 2012.

Ştefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302, 2019.

Trey Smith and Reid Simmons. Heuristic search value iteration for pomdps. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pp. 520–527, Arlington, Virginia, USA, 2004. AUAI Press. ISBN 0974903906.

James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7768–7777, 2021.

Jianrui Wang, Yitian Hong, Jiali Wang, Jiapeng Xu, Yang Tang, Qing-Long Han, and Jürgen Kurths. Cooperative and competitive multi-agent systems: From optimization to games. *IEEE/CAA Journal of Automatica Sinica*, 9(5):763–783, 2022a.

Xinjun Wang, Ye Cao, Ben Niu, and Yongduan Song. A novel bipartite consensus tracking control for multiagent systems under sensor deception attacks. *IEEE Transactions on Cybernetics*, 53(9): 5984–5993, 2022b.

Yifan Wang, Lukas Rahmann, and Olga Sorkine-Hornung. Geometry-consistent neural shape representation with implicit displacement fields. In *International Conference on Learning Representations, ICLR*, 2022c. arXiv:2106.05187.

Yue Wang, Eloy Garcia, David Casbeer, and Fumin Zhang. *Cooperative control of multi-agent systems: Theory and applications*. John Wiley & Sons, 2017.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020.

Huan Zhang, Hongge Chen, Duane S. Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations, ICLR*, 2021. arXiv:2101.08452.

A PROOFS

Lemma 2.1 (Robust Optimality $\iff$ Nash Equilibrium)**.** *Consider a zero-sum SA-MDP with finite state, action, and observation sets, nonempty perturbation sets $B(s)$, bounded rewards, a fixed initial distribution $b \in \Delta(S)$, and unrestricted full-history behavioral strategy spaces Σ_1 and Σ_2 for the agent and the adversary, respectively. We assume perfect recall in which each player's private history records all previously observed information and all of that player's previous actions. For a strategy profile $(\sigma, \tau) \in \Sigma_1 \times \Sigma_2$, let $V(b; \sigma, \tau)$ denote the agent's expected return for the horizon under consideration when $s^0 \sim b$; when the horizon is $N < \infty$, we also write $V^N(b; \sigma, \tau)$ to make the horizon explicit. For a finite horizon, we allow $\gamma \in (0, 1]$, with $\gamma = 1$ corresponding to the undiscounted return; for an infinite discounted horizon, we require $\gamma \in (0, 1)$. In either case, the following equality holds for every initial distribution $b \in \Delta(S)$:*

$$\arg\max_{\sigma \in \Sigma_1} \min_{\tau \in \Sigma_2} V(b; \sigma, \tau) = \left\{\sigma \in \Sigma_1 : \exists \tau \in \Sigma_2 \text{ with } (\sigma, \tau) \text{ a Nash equilibrium when } s^0 \sim b\right\}.$$

Both sets are nonempty.

Proof of Lemma 2.1. Throughout, player 1 is the maximizing player, and player 2 is the minimizing player.

Finite horizon. Fix a horizon $N < \infty$. Since all underlying alphabets are finite, the induced N-stage extensive-form zero-sum game has finitely many histories and information sets. By assumption, it has perfect recall.

Let P_i^N be the finite set of player i's pure complete contingent plans, where a plan specifies an action at every information set of that player, including information sets that the plan itself prevents from being reached. Let

$$M_i^N := \Delta(P_i^N)$$

be the corresponding mixed-strategy simplex. If $u_N(\mu, \nu)$ denotes the expected payoff under $(\mu, \nu) \in M_1^N \times M_2^N$, then u_N is bilinear. Von Neumann's minimax theorem, therefore, gives $(\mu_N^*, \nu_N^*) \in M_1^N \times M_2^N$ such that

$$u_N(\mu, \nu_N^*) \le u_N(\mu_N^*, \nu_N^*) \le u_N(\mu_N^*, \nu) \qquad \forall (\mu, \nu) \in M_1^N \times M_2^N.$$

Therefore, by Kuhn's realization-equivalence theorem for finite perfect-recall games, there are behavioral strategies σ_N^* and τ_N^* that are realization-equivalent to μ_N^* and ν_N^*, respectively (Kuhn, 1953). Here, realization equivalence means equality of the induced distribution over terminal histories against every strategy of the opponent. Conversely, every behavioral strategy in this finite game has a realization-equivalent mixed strategy over complete contingent plans.

Consequently, for arbitrary behavioral deviations σ and τ, choose realization-equivalent mixed strategies μ_σ and ν_τ. The preceding normal-form saddle inequalities imply

$$V^N(b; \sigma, \tau_N^*) = u_N(\mu_\sigma, \nu_N^*) \le u_N(\mu_N^*, \nu_N^*) = V^N(b; \sigma_N^*, \tau_N^*)$$

and

$$V^N(b; \sigma_N^*, \tau_N^*) = u_N(\mu_N^*, \nu_N^*) \le u_N(\mu_N^*, \nu_\tau) = V^N(b; \sigma_N^*, \tau).$$

Thus (σ_N^*, τ_N^*) is a behavioral saddle point of the N-stage game.

Infinite discounted horizon. We obtain an infinite-horizon saddle point as a limit of finite-horizon saddle points.

For $n \ge 0$, let $\mathcal{H}_i^n$ denote the finite set of player i's private histories immediately before the stage-n action. Under the information structure above, generic elements of $\mathcal{H}_1^n$ and $\mathcal{H}_2^n$, respectively, have the forms

$$h_1^n = \left(a_1^0, o_0, \ldots, a_1^{n-1}, o^{n-1}\right)$$

and

$$h_2^n = \left(s^0, a_1^0, a_2^0, o^{(0)}, s_1, \ldots, a_1^{n-1}, a_2^{n-1}, o^{n-1}, s^n\right),$$

with $h_1^0 = \varnothing$ and $h_2^0 = (s^0)$. Let

$$\mathcal{H}_i := \bigcup_{n \ge 0} \mathcal{H}_i^n.$$

Because all underlying alphabets are finite, each $\mathcal{H}_i^n$ is finite, so $\mathcal{H}_i$ is countable. For $h \in \mathcal{H}_i$, let $A_i(h)$ be the nonempty finite set of actions available to player i at h. Thus, $A_1(h) = A_1$ for the agent, while $A_2(h) = B(s)$ at an adversary history whose current true state is s. Define

$$\Sigma_i \coloneqq \prod_{h \in \mathcal{H}_i} \Delta(A_i(h))$$

and endow it with the product topology.

Each factor $\Delta(A_i(h))$ is a compact metrizable simplex in a finite-dimensional Euclidean space. Tychonoff's theorem therefore implies that Σ_i is compact in the product topology (Munkres, 2000, Theorem 37.3). Moreover, the probability coordinates (h, a), where $h \in \mathcal{H}_i$ and $a \in A_i(h)$, form a countable set. Identifying each behavioral choice with its action probabilities therefore realizes Σ_i, with the product topology above, as a subspace of a countable product of copies of $[0, 1]$, which is metrizable in the product topology (Munkres, 2000, Theorem 20.5). Thus, each Σ_i, and hence the finite product $\Sigma_1 \times \Sigma_2$, is compact and metrizable. It is therefore sequentially compact (Munkres, 2000, Theorem 28.2).

For $N \geq 1$, define the truncated discounted payoff

$$V^N(b; \sigma, \tau) \coloneqq \mathbb{E}_{s^0 \sim b, \sigma, \tau}\left[\sum_{t=0}^{N-1} \gamma^t R_t\right].$$

For each player i, $V^N(b; \cdot, \cdot)$ depends only on the behavioral coordinates indexed by the finite set $\bigcup_{n=0}^{N-1} \mathcal{H}_i^n$, namely the private histories immediately preceding actions in the first N stages. Expanding the expectation as a sum over the finitely many length-N paths shows that it is a polynomial in those coordinates and therefore is continuous in the product topology.

Choose $\bar{R} < \infty$ such that $|R_t| \leq \bar{R}$ almost surely. Then, uniformly over $(\sigma, \tau) \in \Sigma_1 \times \Sigma_2$,

$$\left|V(b; \sigma, \tau) - V^N(b; \sigma, \tau)\right| \leq \sum_{t=N}^{\infty} \gamma^t \bar{R} = \frac{\bar{R}\gamma^N}{1-\gamma} =: \varepsilon_N \longrightarrow 0. \tag{2}$$

Thus $V(b; \cdot, \cdot)$ is the uniform limit of the continuous functions $V^N(b; \cdot, \cdot)$, so $V(b; \cdot, \cdot)$ is jointly continuous on $\Sigma_1 \times \Sigma_2$.

For every N, let (σ^N, τ^N) be a behavioral saddle point of the N-stage discounted game, whose existence follows from the finite-horizon argument, and extend these strategies arbitrarily at histories occurring after stage N. Sequential compactness of $\Sigma_1 \times \Sigma_2$ gives a subsequence $N_k \to \infty$ and a profile (σ^*, τ^*) such that

$$(\sigma^{N_k}, \tau^{N_k}) \longrightarrow (\sigma^*, \tau^*).$$

Fix arbitrary infinite-horizon strategies $(\sigma, \tau) \in \Sigma_1 \times \Sigma_2$. Their restrictions to the first N_k stages are admissible behavioral deviations in the N_k stage game. Therefore

$$V^{N_k}(b; \sigma, \tau_{N_k}) \leq V^{N_k}(b; \sigma^{N_k}, \tau^{N_k}) \leq V^{N_k}(b; \sigma^{N_k}, \tau).$$

Using (2) gives

$$V(b; \sigma, \tau^{N_k}) \leq V(b; \sigma^{N_k}, \tau^{N_k}) + 2\varepsilon_{N_k}$$

and

$$V(b; \sigma^{N_k}, \tau^{N_k}) \leq V(b; \sigma^{N_k}, \tau) + 2\varepsilon_{N_k}.$$

Joint continuity of $V(b; \cdot, \cdot)$, convergence of the strategy profiles, and $\varepsilon_{N_k} \to 0$ now imply

$$V(b; \sigma, \tau^*) \leq V(b; \sigma^*, \tau^*) \leq V(b; \sigma^*, \tau) \qquad \forall (\sigma, \tau) \in \Sigma_1 \times \Sigma_2.$$

Hence, (σ^*, τ^*) is an infinite-horizon behavioral saddle point.

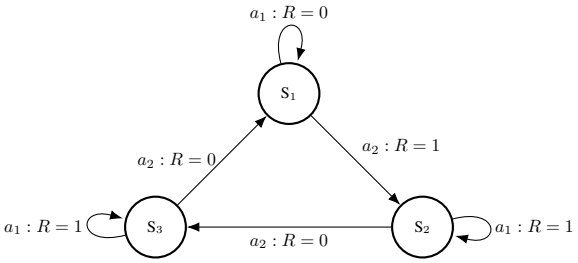
Equality of the two policy sets. In either horizon model, let (σ^0, τ^0) be a saddle point and set

$$v \coloneqq V(b; \sigma^0, \tau^0).$$

The saddle inequalities give

$$v = \max_{\sigma \in \Sigma_1} \min_{\tau \in \Sigma_2} V(b; \sigma, \tau) = \min_{\tau \in \Sigma_2} \max_{\sigma \in \Sigma_1} V(b; \sigma, \tau).$$

Figure 4: MDP without History-Dependent State-Robust Policy



Suppose that $\widehat{\sigma}$ is robust-optimal. Then

$$\min_{\tau} V(b; \widehat{\sigma}, \tau) = v,$$

and hence

$$V(b; \widehat{\sigma}, \tau) \geq v \qquad \forall \tau \in \Sigma_2.$$

The saddle property of (σ^0, τ^0) also gives

$$V(b; \sigma, \tau^0) \leq v \qquad \forall \sigma \in \Sigma_1.$$

In particular, $V(b; \widehat{\sigma}, \tau^0) = v$, and consequently

$$V(b; \sigma, \tau^0) \leq V(b; \widehat{\sigma}, \tau^0) \leq V(b; \widehat{\sigma}, \tau) \qquad \forall (\sigma, \tau).$$

Thus $(\widehat{\sigma}, \tau^0)$ is a saddle point, equivalently a Nash equilibrium.

Conversely, if $(\bar{\sigma}, \bar{\tau})$ is any Nash equilibrium, then, because the game is zero-sum, it is a saddle point, and its payoff equals the game value v. Hence

$$\min_{\tau \in \Sigma_2} V(b; \bar{\sigma}, \tau) = V(b; \bar{\sigma}, \bar{\tau}) = v,$$

so $\bar{\sigma}$ is robust-optimal. The existence of a saddle point also proves that both sets in the statement are nonempty. □

Theorem 2.2. *There exists a SA-MDP such that no history-dependent policy is state-robust optimal.*

Proof of Theorem 2.2. We begin with a high-level overview of the proof and then proceed through the details.

Let σ_v^A be the policy that always plays a_1. In the SA-MDP of Figure 4, σ_v^A achieves robust value 0 at s_1 and $\frac{1}{1-\gamma}$ at s_2, s_3, while uniformly mixing a_1, a_2 achieves $\frac{1}{2(1-\gamma)}$ everywhere. Any state-robust optimal σ_v must therefore match σ_v^A at s_2, s_3 and strictly improve on it at s_1. To improve at s_1 against the identity adversary, σ_v must, after some first observed (i.e. reported by adversary) history $h^* = (s_1, a_1, \ldots, s_1)$, place positive probability on a_2. But the agent conditions only on reported histories: an adversary starting from true state s_3 can fabricate the same h^* by reporting s_1 at every step, since σ_v plays a_1 surely along the prefix and the true state remains s_3. Triggering a_2 at true s_3 strictly reduces the return, so $J_{\sigma_v}(s_3) < \frac{1}{1-\gamma}$, contradicting state-robust optimality at s_3. For a history-dependent agent policy σ_v and state $s \in S$, define the worst-case value

$$J_{\sigma_v}(s) \;:=\; \min_{\sigma_a} V^{\sigma_v, \sigma_a}(s).$$

A policy σ_v is *state-robust optimal* if $J_{\sigma_v}(s) \geq J_{\sigma_v'}(s)$ for every state s and every alternative policy σ_v'.

Consider the SA-MDP in Figure 4. Let σ_v^A denote the policy that always plays a_1, regardless of the reported history. Direct computation gives

$$J_{\sigma_v^A}(s_1) = 0, \qquad J_{\sigma_v^A}(s_2) = J_{\sigma_v^A}(s_3) = \frac{1}{1-\gamma}.$$

For $\gamma = 0.99$, these values are 0, 100, and 100, respectively.

Now let σ_v^B be the policy that plays each of $\{a_1, a_2\}$ with probability $\frac{1}{2}$ at every history. Under σ_v^B, the expected one-step reward is $\frac{1}{2}$ in every state, so

$$J_{\sigma_v^B}(s) \;=\; \frac{1}{2(1-\gamma)} \quad \text{for all } s \in \{s_1, s_2, s_3\},$$

which equals 50 when $\gamma = 0.99$.

Therefore, any state-robust optimal policy σ_v must satisfy

$$J_{\sigma_v}(s_1) \geq 50, \qquad J_{\sigma_v}(s_2) \geq 100, \qquad J_{\sigma_v}(s_3) \geq 100.$$

Since no policy can achieve more than $\frac{1}{1-\gamma} = 100$ from s_2 or s_3, any state-robust optimal policy must in fact satisfy

$$J_{\sigma_v}(s_2) = J_{\sigma_v}(s_3) = \frac{1}{1-\gamma} \quad \text{and} \quad J_{\sigma_v}(s_1) \geq \frac{1}{2(1-\gamma)} > 0. \tag{3}$$

Contradiction argument. Assume for contradiction that there exists a history-dependent policy σ_v satisfying (3).

Let σ_a^I denote the *identity adversary* that always reports the true state. Since $J_{\sigma_v}(s_1) > 0$, we have in particular

$$V^{\sigma_v, \sigma_a^I}(s_1) \;\geq\; J_{\sigma_v}(s_1) \;>\; 0.$$

Hence, starting from true state s_1 against σ_a^I, there must exist some first time step n at which σ_v assigns positive probability to action a_2. Otherwise, σ_v would play a_1 surely at every step in s_1, yielding value 0, which contradicts $V^{\sigma_v, \sigma_a^I}(s_1) > 0$.

By minimality of n, along the identity-adversary play starting from s_1, the reported history up to time n must be

$$h^* \;=\; (s_1, a_1, s_1, a_1, \ldots, s_1, a_1, s_1),$$

i.e., n repeated reports of s_1 with action a_1 taken with probability 1 at all earlier time steps. After this history, the policy plays a_2 with positive probability:

$$\Pr_{\sigma_v}\big[a_2 \,\big|\, h^*\big] \;>\; 0.$$

Mimicking the history from s_3. Now consider instead starting from true state s_3, and let σ_a' be the adversary that reports s_1 for the first n observations (recall $s_1 \in B(s_3)$, so this is feasible). Since σ_v conditions only on the *reported* history, it plays a_1 with probability 1 on the first $n-1$ steps—exactly as it does along h^* from true state s_1. Because s_3 is absorbing under action a_1, the true state remains s_3 throughout these steps.

At time n, the reported history is again h^*, so the policy plays a_2 with positive probability. However, at true state s_3, action a_2 yields reward 0 and transitions out of s_3, while a_1 yields reward 1 and keeps the agent in s_3. Conditioning on the event $\{a_2 \text{ played at step } n\}$, which occurs with positive probability under σ_v given h^*, the discounted return is strictly less than $\frac{1}{1-\gamma}$. Therefore

$$V^{\sigma_v, \sigma_a'}(s_3) \;<\; \frac{1}{1-\gamma},$$

which implies

$$J_{\sigma_v}(s_3) \;\leq\; V^{\sigma_v, \sigma_a'}(s_3) \;<\; \frac{1}{1-\gamma}.$$

This contradicts (3), which required $J_{\sigma_v}(s_3) = \frac{1}{1-\gamma}$. Hence no such history-dependent policy σ_v can exist, and no history-dependent policy is state-robust optimal in this SA-MDP. □

Theorem 2.3. *Any SA-MDP G_{seq} can be transformed into a strategically equivalent constrained zero-sum one-sided POSG $\widehat{G}$ in which both players act simultaneously at every stage. The solution strategy to $\widehat{G}$ can be mapped back to a unique solution strategy of G_{seq}.*

Proof of Theorem 2.3. We begin the proof by defining the relevant objects for a transformation and then establish that this transformation induces a game that is strategically equivalent to the original observation space attack game G_{seq}

DEFINING THE TRANSFORMATION

We begin with a sequential observation space attack game $G_{\text{seq}} = (S, A_v, R, p, \gamma, B)$ and construct a constrained zero-sum OS-POSG $\widehat{G} = (\widehat{S}, \widehat{A}_1, \widehat{A}_2, \widehat{O}, \widehat{T}, \widehat{R}, \sqrt{\gamma}, L_1, L_2)$ that is strategically equivalent to G_{seq} but meets the *simultaneous* action requirements of Horák et al. (2023). We define the components of $\widehat{G}$ with the components of G_{seq} and necessary dummy actions and states.

State space. For every $s \in S$ create two states, an *original* state s and a *dummy* state $\bar{s}$. Hence, $\widehat{S} = S \cup \bar{S}, \bar{S} = \{\bar{s} : s \in S\}$.

Action sets. Introduce distinguished dummy actions $\perp_1 \notin A_v$ for the agent and $\perp_2 \notin S$ for the adversary, and define $\widehat{A}_1 = A_v \cup \{\perp_1\}, \widehat{A}_2 = S \cup \{\perp_2\}$.

Agent 1 represents the agent, so its meaningful actions are the original agent actions in A_v. Agent 2 represents the attacker, whose meaningful actions encode the observation reported to the agent and therefore range over S. The dummy actions $\perp_1, \perp_2$ are used only on non-decisive micro-stages. To eliminate spurious signaling opportunities, we impose state-dependent feasible action sets: $L_1(s) = \{\perp_1\}, L_2(s) = B(s)$ for each original state $s \in S$, and $L_1(\bar{s}) = A_v, L_2(\bar{s}) = \{\perp_2\}$ for each dummy state $\bar{s} \in \bar{S}$.

Stage decomposition. One stage of G_{seq} is encoded as *two* micro-stages in $\widehat{G}$:

1. Suppose the system is in an original state s. The attacker chooses a feasible action $a_2 \in L_2(s) = B(s)$, which corresponds to the observation it wants the agent to receive, while the agent is forced to play $\perp_1$. The system deterministically transitions to $\bar{s}$ and emits observation $o = a_2$.

2. The system is now in the dummy state $\bar{s}$. The agent chooses $a_1 \in L_1(\bar{s}) = A_v$ given the received observation $o = a_2$, while the attacker is forced to play $\perp_2$. The system then transitions according to the original probability $p(\cdot \mid s, a_1)$ to the next original state s' and emits a special dummy observation $\bar{o}$.

Transition Kernel $\widehat{T}$. For every original state $s \in S$ and feasible attacker action $a_2 \in B(s)$, $\widehat{T}(a_2, \bar{s} \mid s, \perp_1, a_2) = 1$. For every dummy state $\bar{s} \in \bar{S}$ and agent action $a_1 \in A_v$, $\widehat{T}(\bar{o}, s' \mid \bar{s}, a_1, \perp_2) = p(s' \mid s, a_1)$. All other transitions have probability zero (equivalently, they correspond to infeasible action tuples under the state-dependent constraints above).

Note that the first micro-stage depends only on the attacker's meaningful action, while the second micro-stage depends only on the agent's meaningful action. The dummy actions never affect transitions, rewards, or observations.

Reward. Define $\widehat{R}(s, a_1, a_2) = 0, \widehat{R}(\bar{s}, a_1, a_2) = \frac{1}{\sqrt{\gamma}} R(s, a_1)$ where the $\frac{1}{\sqrt{\gamma}}$ accounts for the rewards being applied at odd indices.

Observations. The agent's observation alphabet is $\widehat{O} = O \cup \{\bar{o}\}$. Player 1 receives the attacker-selected observation a_2 at the first micro-stage and the dummy observation $\bar{o}$ at the second micro-stage. The observation $\bar{o}$ signals that the next micro-stage is attacker-decisive and the agent is constrained to play $\perp_1$. Player 2 always observes the full state.

Discount Factor. The transformed game has discount factor $\sqrt{\gamma}$. Since the only non-zero rewards occur at the second micro-stage of each encoded step, we scale rewards on dummy states by $1/\sqrt{\gamma}$. Hence $(\sqrt{\gamma})^{2t+1} \widehat{R}(\bar{s}, a_1, \perp_2) = (\sqrt{\gamma})^{2t+1} \frac{1}{\sqrt{\gamma}} R(s, a_1) = \gamma^t R(s, a_1)$, so the transformed discounted return matches the original discounted return exactly.

Consider the following visualization (Figure 5a) of how one stage in the sequential observation space attack game can be simulated by a simultaneous move game. We will consider the first stage of the game, transitioning from s^0 to either s_1 or s_2.

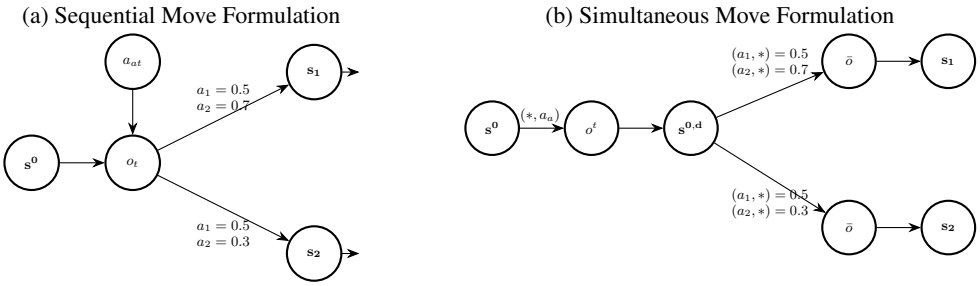


Figure 5: Strategically equivalent forms of an OS-POSG.

Note that in the sequential setting, the observation emitted by the system is modified according to the attacker's action. This observation informs the agent's action, while the true state dictates the actual transition probabilities.

Strategic Equivalence

We now show that the transformed simultaneous zero-sum OS-POSG $\widehat{G}$ is strategically equivalent to the original sequential observation-space attack game G_{seq}. We establish the following three properties:

1. *Strategy correspondence:* after restricting each non-decisive micro-stage to its singleton dummy action, there is a bijection
$$\Phi_i : \Sigma_i^{\text{seq}} \longrightarrow \widehat{\Sigma}_i^{\text{feas}}, \qquad i \in \{1, 2\},$$
between each player's behavioral strategies in G_{seq} and feasible behavioral strategies in $\widehat{G}$.

2. *Pairwise payoff preservation:* for every initial belief b, its canonical embedding $\widehat{b}$ in the transformed state space, and every strategy pair (σ_1, σ_2),
$$V^{\text{seq}}(b; \sigma_1, \sigma_2) = V^{\text{sim}}\big(\widehat{b}; \Phi_1(\sigma_1), \Phi_2(\sigma_2)\big).$$

3. *ϵ-equilibrium correspondence:* a strategy pair is an ϵ-Nash equilibrium, equivalently an ϵ-saddle point, of G_{seq} if and only if its image under (Φ_1, Φ_2) is an ϵ-Nash equilibrium of $\widehat{G}$ over feasible strategies.

These three results prove Theorem 2.3. Result 1 guarantees that a strategy found in $\widehat{G}$ will have a corresponding unique strategy in G_{seq}. Result 2 guarantees that the values of the Nash equilibria in both games are equal. Finally, Result 3 guarantees that the strategy in G_{seq} still achieves an ϵ-approximate Nash equilibrium.

Preliminaries and Definitions

Let σ_1 (resp. σ_2) denote the agent (resp. adversary) strategy in the sequential game G_{seq}, with value
$$V^{\text{seq}}(b; \sigma_1, \sigma_2) = \mathbb{E}_{s^0 \sim b}\Big[\sum_{t=0}^{\infty} \gamma^t R\big(s_t, a_1^t\big)\Big],$$
and let $\widehat{\sigma}_1, \widehat{\sigma}_2$ be analogous strategies in the simultaneous game $\widehat{G}$ with value
$$V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1, \widehat{\sigma}_2) = \mathbb{E}_{s^0 \sim b}\Big[\sum_{n=0}^{\infty} (\sqrt{\gamma})^n \widehat{R}\big(x_n, a_1^n, a_2^n\big)\Big],$$
where x_n, a_i^n denote states/actions at micro–step.

Define a bijection $\Phi : (a_2^0, o^0, a_1^0, a_2^1, o^1, a_1^1, \dots) \to (\perp_1, a_2^0, o^0, a_1^0, \perp_2, \bar{o}, \perp_1, a_2^1, o^1, a_1^1, \perp_2, \bar{o}, ...)$ that interleaves each stage of G_{seq} into two constrained simultaneous micro-stages of $\widehat{G}$: at step $2t$ the attacker chooses a_2^t while the agent is forced to play $\perp_1$, and at step $2t+1$ the agent chooses a_1^t while the adversary is forced to play $\perp_2$. Because the non-decisive actions are fixed and observations are emitted in the same order as in the sequential game, the mapping Φ is bijective and preserves each player's information set.

Given a sequential strategy pair (σ_1, σ_2), define the feasible simultaneous strategy pair $(\widehat{\sigma}_1, \widehat{\sigma}_2)$ by

$$\widehat{\sigma}_1(\cdot \mid \Phi(h^{\text{seq}}), o^{(2t+1)}) = \sigma_1(\cdot \mid h^{\text{seq}}, o^t), \quad \widehat{\sigma}_2(\cdot \mid \Phi(h^{\text{seq}}), x^{2t}) = \sigma_2(\cdot \mid h^{\text{seq}}),$$

with

$$\widehat{\sigma}_1(\cdot \mid \Phi(h^{\text{seq}}), o_2^t) = \delta_{\perp_1}, \qquad \widehat{\sigma}_2(\cdot \mid \Phi(h^{\text{seq}}), x_{2,t+1}) = \delta_{\perp_2}.$$

Here h^{seq} denotes the relevant player's private history. For player 1, Φ inserts the forced dummy components into its action-observation history, while for player 2, it inserts the dummy states and actions into its full state-action history.

Conversely, every feasible simultaneous strategy pair induces a unique sequential strategy pair by deleting the forced dummy moves.

Definition A.1 (ϵ–equilibrium concepts)**.** Recall that player 1 is the maximizing player and player 2 is the minimizing player.

A feasible strategy pair $(\widehat{\sigma}_1^*, \widehat{\sigma}_2^*) \in \widehat{\Sigma}_1^{\text{feas}} \times \widehat{\Sigma}_2^{\text{feas}}$ is an ϵ–Nash equilibrium, equivalently an ϵ–saddle point, of $\widehat{G}$ if

$$V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1, \widehat{\sigma}_2^*) \le V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1^*, \widehat{\sigma}_2^*) + \epsilon \qquad \forall\, \widehat{\sigma}_1 \in \widehat{\Sigma}_1^{\text{feas}}, \tag{4}$$

$$V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1^*, \widehat{\sigma}_2) \ge V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1^*, \widehat{\sigma}_2^*) - \epsilon \qquad \forall\, \widehat{\sigma}_2 \in \widehat{\Sigma}_2^{\text{feas}}. \tag{5}$$

A strategy pair $(\sigma_1^*, \sigma_2^*) \in \Sigma_1^{\text{seq}} \times \Sigma_2^{\text{seq}}$ is an ϵ–Nash equilibrium, equivalently an ϵ–saddle point, of G_{seq} if

$$V^{\text{seq}}(b; \sigma_1, \sigma_2^*) \le V^{\text{seq}}(b; \sigma_1^*, \sigma_2^*) + \epsilon \qquad \forall\, \sigma_1 \in \Sigma_1^{\text{seq}}, \tag{6}$$

$$V^{\text{seq}}(b; \sigma_1^*, \sigma_2) \ge V^{\text{seq}}(b; \sigma_1^*, \sigma_2^*) - \epsilon \qquad \forall\, \sigma_2 \in \Sigma_2^{\text{seq}}. \tag{7}$$

THEOREM STATEMENTS AND PROOFS

Lemma A.2 (Exact value preservation)**.** *Under optimal play,*

$$V^{\text{seq}}(b) = \max_{\sigma_1} \min_{\sigma_2} V^{\text{seq}}(b; \sigma_1, \sigma_2) = \max_{\widehat{\sigma}_1} \min_{\widehat{\sigma}_2} V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1, \widehat{\sigma}_2) = V^{\text{sim}}(\widehat{b}).$$

Proof. We use the bijection Φ to translate any sequential strategy pair (σ_1, σ_2) into a feasible simultaneous pair $(\widehat{\sigma}_1, \widehat{\sigma}_2)$ and vice-versa, without altering the induced probability measure over outcome streams. Observe:

1. *Reward alignment:* by construction, the only non-zero reward in the transformed game occurs at the second micro-stage x^{2t+1}, and
$$\widehat{R}(x^{2t+1}, a_1^{2t+1}, a_2^{2t+1}) = \frac{1}{\sqrt{\gamma}} R(s_t, a_1^t).$$
2. *Discount consistency:* therefore
$$(\sqrt{\gamma})^{2t+1} \widehat{R}(x^{2t+1}, a_1^{2t+1}, a_2^{2t+1}) = (\sqrt{\gamma})^{2t+1} \frac{1}{\sqrt{\gamma}} R(s_t, a_1^t) = \gamma^t R(s_t, a_1^t).$$
3. *Measure preservation:* at each original stage, corresponding strategies assign the same probabilities to the attacker's report and the agent's action. The inserted moves are deterministic, and the second micro-stage uses the same transition probability $p(s^{t+1} \mid s_t, a_1^t)$. Hence, after deleting the dummy components, corresponding trajectories have the same probability.

4. *Feasibility preservation:* the singleton constraints ensure that the transformed game introduces no additional strategic choices beyond those already present in G_{seq}.

Therefore, for any corresponding strategy pair,

$$V^{\text{seq}}(b; \sigma_1, \sigma_2) = V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1, \widehat{\sigma}_2).$$

Taking max–min over admissible strategies yields the desired equality. □

Lemma A.3 (ϵ-equilibrium correspondence)**.** *Fix an initial belief b. For each player $i \in \{1, 2\}$, let*

$$\Phi_i : \Sigma_i^{\text{seq}} \longrightarrow \widehat{\Sigma}_i^{\text{feas}}$$

be a bijection between the behavioral strategies in G_{seq} and the feasible behavioral strategies in $\widehat{G}$. Suppose that, for every (σ_1, σ_2),

$$V^{\text{seq}}(b; \sigma_1, \sigma_2) = V^{\text{sim}}\big(\widehat{b}; \Phi_1(\sigma_1), \Phi_2(\sigma_2)\big).$$

Then (σ_1^*, σ_2^*) is an ϵ-Nash equilibrium of G_{seq} if and only if $\big(\Phi_1(\sigma_1^*), \Phi_2(\sigma_2^*)\big)$ is an ϵ-Nash equilibrium of $\widehat{G}$.

Proof. Write

$$\widehat{\sigma}_i^* = \Phi_i(\sigma_i^*), \qquad i \in \{1, 2\}.$$

Assume first that (σ_1^*, σ_2^*) is an ϵ-Nash equilibrium of G_{seq}. For any feasible deviation $\widehat{\sigma}_1 \in \widehat{\Sigma}_1^{\text{feas}}$, bijectivity gives $\sigma_1 = \Phi_1^{-1}(\widehat{\sigma}_1)$. Hence,

$$\begin{aligned} V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1, \widehat{\sigma}_2^*) &= V^{\text{seq}}(b; \sigma_1, \sigma_2^*) \\ &\le V^{\text{seq}}(b; \sigma_1^*, \sigma_2^*) + \epsilon \\ &= V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1^*, \widehat{\sigma}_2^*) + \epsilon. \end{aligned}$$

Similarly, for any feasible deviation $\widehat{\sigma}_2 \in \widehat{\Sigma}_2^{\text{feas}}$, setting $\sigma_2 = \Phi_2^{-1}(\widehat{\sigma}_2)$ yields

$$\begin{aligned} V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1^*, \widehat{\sigma}_2) &= V^{\text{seq}}(b; \sigma_1^*, \sigma_2) \\ &\ge V^{\text{seq}}(b; \sigma_1^*, \sigma_2^*) - \epsilon \\ &= V^{\text{sim}}(\widehat{b}; \widehat{\sigma}_1^*, \widehat{\sigma}_2^*) - \epsilon. \end{aligned}$$

Thus, $(\widehat{\sigma}_1^*, \widehat{\sigma}_2^*)$ is an ϵ-Nash equilibrium of $\widehat{G}$. The converse follows by the same argument, applying the inverse bijections Φ_i^{-1} to arbitrary deviations in G_{seq}. Finally, the feasible strategy spaces of $\widehat{G}$ admit no additional deviations at dummy micro-stages, since the acting player has a singleton feasible action set there. Therefore, every feasible deviation in $\widehat{G}$ corresponds uniquely to a deviation in G_{seq}, as required. □

Lemma A.4 (Well-defined strategy correspondence)**.** *The bijection Φ ensures a one-to-one correspondence between strategies in $\widehat{G}$ and strategies in G_{seq}.*

Proof. By construction of Φ. □

This concludes the proof of Theorem 2.3. □

Theorem 2.5. *The HSVI algorithm of Horák et al. (2023) extends to the constrained zero-sum one-sided POSG in Definition 2.4 by restricting each local stage game to the players' feasible strategy sets. The algorithm maintains a lower-bound value function V_{LB} and an upper-bound value function V_{UB}, written as V_{LB}^{Γ} and V_{UB}^{Υ}, respectively, in Algorithm 2, satisfying*

$$V_{LB}(b) \le V^*(b) \le V_{UB}(b)$$

for every reachable belief b, where V^* is the optimal value function of the constrained game.

Let

$$\delta = \frac{U_0 - L_0}{2},$$

where L_0 and U_0 are the discounted minimum and maximum feasible reward bounds. If $\delta = 0$, the initial bounds coincide. If $\delta > 0$ and

$$0 < D < \frac{(1-\gamma)\epsilon}{2\delta},$$

then the algorithm terminates at the initial belief b_0 with

$$V_{UB}(b_0) - V_{LB}(b_0) \le \epsilon.$$

The strategy obtained from the lower-bound value function for player 1 and the strategy obtained from the upper-bound value function for player 2 are feasible and form an ϵ-Nash equilibrium. In particular, the result applies to the constrained OS-POSG $\widehat{G}$ constructed in Theorem 2.3.

Proof of Theorem 2.5. We adapt the proof of Horák et al. (2023) by restricting the action sets in each local stage game. We state the parts of their argument that are unchanged and make explicit where the feasibility restrictions enter.

Model and feasible stage strategies. At a player-1 information history h_1, let $S(h_1)$ be the states that may be current at that history. By the condition in Definition 2.4, all states in $S(h_1)$ have the common feasible action set $L_1(h_1)$. Thus, at a belief $b \in \Delta(S(h_1))$, the stage strategies are

$$\pi_1 \in \Pi_1(h_1) \coloneqq \Delta(L_1(h_1)), \qquad \pi_2 \in \Pi_2(h_1) \coloneqq \prod_{s \in S(h_1)} \Delta(L_2(s)).$$

Player 1 therefore uses one distribution over $L_1(h_1)$; only player 2 may condition its action on the current state. Both strategy sets are nonempty, compact, and convex.

Belief update. Fix $b \in \Delta(S(h_1))$, $a_1 \in L_1(h_1)$, and $\pi_2 \in \Pi_2(h_1)$. Define

$$p(o, s' \mid b, a_1, \pi_2) = \sum_{s \in S(h_1)} \sum_{a_2 \in L_2(s)} b(s)\pi_2(a_2 \mid s) T(o, s' \mid s, a_1, a_2),$$

$$P(o \mid b, a_1, \pi_2) = \sum_{s' \in S} p(o, s' \mid b, a_1, \pi_2).$$

For the successor history $h_1' = (h_1, a_1, o)$, the states that may be current are

$$S(h_1') = \{s' \in S : T(o, s' \mid s, a_1, a_2) > 0 \text{ for some } s \in S(h_1),\ a_2 \in L_2(s)\}.$$

Action-observation pairs for which $S(h_1') = \emptyset$ are omitted from all continuation sums and lower-bound constraints, since their probability is zero under every feasible stage strategy. Thus $S(h_1')$ depends only on $S(h_1)$, a_1, and o, and every posterior generated by a feasible π_2 on a remaining branch is supported on $S(h_1')$. When $P(o \mid b, a_1, \pi_2) > 0$, player 1's posterior after observing its action a_1 and observation o is

$$\tau(b, a_1, \pi_2, o)(s') = \frac{p(o, s' \mid b, a_1, \pi_2)}{P(o \mid b, a_1, \pi_2)}. \tag{8}$$

On a zero-probability branch, the posterior may be chosen arbitrarily in $\Delta(S(h_1'))$. If player 1 uses π_1, the probability of the branch (a_1, o) is

$$\Pr_{b, \pi_1, \pi_2}[a_1, o] = \pi_1(a_1) P(o \mid b, a_1, \pi_2).$$

The factor $\pi_1(a_1)$ does not appear in (8) because player 1 conditions on the action it took. The condition in Definition 2.4 also ensures that every positive-probability successor belief is supported on a set of states with one common feasible action set for player 1.

Values and structure. For a feasible strategy σ_1 of player 1, define

$$V_{\sigma_1}(b) = \inf_{\sigma_2} \mathbb{E}_{b, \sigma_1, \sigma_2}\left[\sum_{t \ge 1} \gamma^{t-1} R(s_t, a_t^1, a_t^2)\right], \qquad V^*(b) = \sup_{\sigma_1} V_{\sigma_1}(b),$$

where both optimizations are over feasible behavioral strategies. Since player 2 observes the state, a best response to a fixed σ_1 can be chosen separately for each initial state. Hence V_{σ_1} is affine in b on each $\Delta(S(h_1))$, and V^* is convex there.

Let

$$R_{\min} = \min_{s\in S,\, a_1\in L_1(s),\, a_2\in L_2(s)} R(s,a_1,a_2),$$
$$R_{\max} = \max_{s\in S,\, a_1\in L_1(s),\, a_2\in L_2(s)} R(s,a_1,a_2),$$
$$L_0 = \frac{R_{\min}}{1-\gamma}, \qquad U_0 = \frac{R_{\max}}{1-\gamma}, \qquad \delta = \frac{U_0 - L_0}{2}.$$

Then $L_0 \le V_{\sigma_1}(b) \le U_0$. As in Horák et al. (2023), every V_{σ_1}, and therefore V^*, is δ-Lipschitz in $\|\cdot\|_1$ on each $\Delta(S(h_1))$.

Bellman operator. Let V be a bounded convex continuous continuation value on the reachable beliefs. At $b \in \Delta(S(h_1))$, define

$$u_{V,b}(\pi_1,\pi_2) = \mathbb{E}_{b,\pi_1,\pi_2}[R(s,a_1,a_2)] + \gamma \sum_{a_1\in L_1(h_1)} \sum_{o\in O} \Pr_{b,\pi_1,\pi_2}[a_1,o] V\big(\tau(b,a_1,\pi_2,o)\big),$$
$$[HV](b) = \max_{\pi_1\in\Pi_1(h_1)} \min_{\pi_2\in\Pi_2(h_1)} u_{V,b}(\pi_1,\pi_2),$$

where zero-probability branches contribute zero. For fixed π_2, this payoff is affine in π_1. For fixed π_1, each continuation term is the perspective of the convex function V and is therefore continuous and convex in π_2. Sion's minimax theorem gives

$$\max_{\pi_1\in\Pi_1(h_1)} \min_{\pi_2\in\Pi_2(h_1)} u_{V,b}(\pi_1,\pi_2) = \min_{\pi_2\in\Pi_2(h_1)} \max_{\pi_1\in\Pi_1(h_1)} u_{V,b}(\pi_1,\pi_2).$$

Thus the local stage game has a value and feasible optimal strategies. The value-composition argument of Horák et al. (2023) applies on each $\Delta(S(h_1))$. It follows that H maps convex δ-Lipschitz continuation values to convex δ-Lipschitz values on the current simplex.

For any bounded V and W,

$$\|HV - HW\|_\infty \le \gamma \|V - W\|_\infty,$$

where the supremum is over all reachable beliefs. The proof is unchanged: for fixed stage strategies, replacing V by W changes the expected continuation value by at most $\gamma\|V-W\|_\infty$, and taking the maximum and minimum preserves the inequality. Moreover, every feasible strategy of player 1 after h_1 consists of a feasible distribution in $\Delta(L_1(h_1))$ and feasible continuation strategies after each (a_1,o), and every such choice defines a feasible behavioral strategy. The strategy-decomposition argument of Horák et al. (2023) therefore gives $HV^* = V^*$. Hence H has a unique fixed point, which is V^*.

Lower-bound stage program. The lower bound is represented by sets of α-vectors, one for each distinct set $S(h_1)$. Consider a point update at $b \in \Delta(S(h_1))$. For each (a_1,o), write $h_1' = (h_1,a_1,o)$ for the successor information history, and let $\Gamma_{a_1,o} = \{\alpha_1^{a_1,o}, \ldots, \alpha_{k_{a_1,o}}^{a_1,o}\}$ be the current lower-bound

vectors on that successor set $S(h_1')$. The lower-bound program of Horák et al. (2023) becomes

$$\begin{aligned}
&\max_{\pi_1,\widehat{\lambda},\widehat{\alpha},V_s} \quad \sum_{s\in S(h_1)} b(s)V_s \\
&\text{s.t.} \quad V_s \le \sum_{a_1\in L_1(h_1)} \pi_1(a_1)R(s,a_1,a_2) \\
&\qquad\qquad + \gamma \sum_{a_1\in L_1(h_1)} \sum_{o\in O} \sum_{s'\in S(h_1')} T(o,s' \mid s,a_1,a_2)\widehat{\alpha}_{a_1,o}(s') \qquad \begin{smallmatrix}\forall s\in S(h_1),\\ a_2\in L_2(s),\end{smallmatrix} \\
&\widehat{\alpha}_{a_1,o}(s') = \sum_{i=1}^{k_{a_1,o}} \widehat{\lambda}_i^{a_1,o}\alpha_i^{a_1,o}(s') \qquad \begin{smallmatrix}\forall a_1\in L_1(h_1),\ o\in O,\\ s'\in S(h_1'),\end{smallmatrix} \\
&\sum_{i=1}^{k_{a_1,o}} \widehat{\lambda}_i^{a_1,o} = \pi_1(a_1) \qquad \begin{smallmatrix}\forall a_1\in L_1(h_1),\\ o\in O,\end{smallmatrix} \\
&\sum_{a_1\in L_1(h_1)} \pi_1(a_1) = 1, \qquad \pi_1 \ge 0, \qquad \widehat{\lambda} \ge 0.
\end{aligned}$$

For every a_1 with $\pi_1(a_1) > 0$, set

$$\alpha_{a_1,o} = \frac{\widehat{\alpha}_{a_1,o}}{\pi_1(a_1)} \in \mathrm{conv}(\Gamma_{a_1,o}).$$

When $\pi_1(a_1) = 0$, choose $\alpha_{a_1,o}$ arbitrarily from $\mathrm{conv}(\Gamma_{a_1,o})$. The vector added to the lower bound is the one-step vector obtained from π_1 and these continuation vectors. Equivalently, it is defined on every $s \in S(h_1)$ by the minimum of the right-hand side of the first constraint over $a_2 \in L_2(s)$. This fixes every coordinate of the new vector, including states to which b assigns zero probability, and its value at b equals the optimal objective of the program.

The only changes are the restrictions $a_1 \in L_1(h_1)$, $a_2 \in L_2(s)$, and the use of the lower-bound vectors for the successor belief. The vector $\widehat{\alpha}_{a_1,o}$ is already multiplied by the probability $\pi_1(a_1)$, since its mixture weights sum to $\pi_1(a_1)$; it is therefore not multiplied by $\pi_1(a_1)$ again. The dual is the dual of Horák et al. (2023) with one realization weight $b(s)\pi_2(a_2 \mid s)$ for every $a_2 \in L_2(s)$. Both programs are finite, feasible, and bounded, so strong duality holds.

Bounds and point updates. For each distinct $S(h_1)$, initialize the lower bound with the constant vector L_0 and the upper bound with the points (e_s, U_0), $s \in S(h_1)$. The first is a valid lower bound because no feasible strategy can receive less than L_0. The lower δ-Lipschitz envelope of the upper points is the constant U_0 and is a valid upper bound.

Every vector produced by the lower-bound program combines a feasible local strategy of player 1 with continuation vectors that already lower-bound the values of feasible continuation strategies. Following those strategies after the corresponding (a_1, o) therefore gives a feasible behavioral strategy whose value is at least the new vector. Convex combinations of stored vectors remain valid because player 1 can randomize privately among the corresponding strategies. It follows by induction that the lower bound remains valid after every point update. Moreover, every coordinate of every lower-bound vector remains in $[L_0, U_0]$. This is true initially, and the one-step construction preserves these bounds because $R_{\min} + \gamma L_0 = L_0$ and $R_{\max} + \gamma U_0 = U_0$. Each vector therefore defines a δ-Lipschitz linear function, and their pointwise maximum is also δ-Lipschitz.

For the upper bound, add the point $(b, [HV_{UB}](b))$ and take the same lower δ-Lipschitz envelope as Horák et al. (2023), using only points on the current $\Delta(S(h_1))$. If the stored points upper-bound V^*, then monotonicity of H gives

$$[HV_{UB}](b) \ge [HV^*](b) = V^*(b),$$

so the new point is also valid. Convexity and the δ-Lipschitz property then show that the updated envelope upper-bounds V^* throughout that simplex. No interpolation is performed between sets with different feasible actions. Thus, throughout the algorithm,

$$V_{LB}(b) \le V^*(b) \le V_{UB}(b).$$

The same updates also preserve the two inequalities used by the strategy construction of Horák et al. (2023). For the initial lower vector, any feasible π_1 together with the same constant continuation gives a one-step value of at least L_0. Every subsequently added vector is the one-step vector obtained from the feasible π_1 returned by the lower-bound program and continuation vectors in the corresponding sets $\mathrm{conv}(\Gamma_{a_1,o})$. The same property holds for convex combinations. Indeed, if $\alpha = \sum_j \lambda_j \alpha^j$, use $\pi_1 = \sum_j \lambda_j \pi_1^j$ and, after an action a_1 with $\pi_1(a_1) > 0$, mix the corresponding continuation vectors with weights $\lambda_j \pi_1^j(a_1)/\pi_1(a_1)$. This gives a feasible one-step vector that dominates α coordinatewise. For the upper bound,

$$[HV_{UB}](b) \le V_{UB}(b)$$

holds initially because $HU_0 \le R_{\max} + \gamma U_0 = U_0$, and it continues to hold after every update. To see the latter, let V'_{UB} be the envelope after adding a point. Since $V'_{UB} \le V_{UB}$, every old point (b_i, y_i) satisfies

$$y_i \ge V_{UB}(b_i) \ge [HV_{UB}](b_i) \ge [HV'_{UB}](b_i),$$

and the new point satisfies

$$[HV_{UB}](b) \ge [HV'_{UB}](b).$$

The convexity and δ-Lipschitz continuity of HV'_{UB} then imply that its lower δ-Lipschitz envelope, V'_{UB}, also lies above HV'_{UB}. This is the same argument as in Horák et al. (2023), applied separately on each $\Delta(S(h_1))$.

Termination. If $\delta = 0$, the initial bounds coincide. Suppose $\delta > 0$ and define

$$\rho(0) = \epsilon, \qquad \rho(t+1) = \frac{\rho(t) - 2\delta D}{\gamma}, \qquad \mathrm{excess}(b,t) = V_{UB}(b) - V_{LB}(b) - \rho(t).$$

Let π_1^{UB} be optimal in the upper-bound stage game and π_2^{LB} be optimal in the lower-bound stage game. The saddle-point inequalities give

$$[HV_{UB}](b) - [HV_{LB}](b) \le \gamma \sum_{a_1,o} \Pr_{b,\pi_1^{UB},\pi_2^{LB}}[a_1, o] \times \big(V_{UB} - V_{LB}\big)\big(\tau(b, a_1, \pi_2^{LB}, o)\big).$$

Consequently, the forward rule of Horák et al. (2023) is unchanged: it selects the branch maximizing its probability multiplied by the successor excess, using the action-conditioned posterior in (8).

If the selected quantity is nonpositive, every positive-probability successor has gap at most $\rho(t+1)$. After the point update, the preceding inequality gives

$$V_{UB}(b) - V_{LB}(b) \le \gamma\rho(t+1) = \rho(t) - 2\delta D.$$

Both bounds are δ-Lipschitz on the current simplex, so their difference is 2δ-Lipschitz. Hence every belief in that simplex within distance D of b has nonpositive excess at depth t. If

$$0 < D < \frac{(1-\gamma)\epsilon}{2\delta},$$

then $\rho(t)$ eventually exceeds the largest possible bound gap, so the recursion depth is finite. Fix a depth t and a structural state set $S(h_1)$, and consider the last belief reached by each trial that terminates there. After the point update at such a belief b, the excess is nonpositive throughout the D-neighborhood of b in $\Delta(S(h_1))$. Later updates can only reduce the gap, so a later trial can terminate at that depth and state set only at a belief more than D from b. These terminal beliefs are therefore D-separated. Compactness of each $\Delta(S(h_1))$, together with the finite recursion depth and the finitely many distinct subsets $S(h_1) \subseteq S$, implies that only finitely many trials occur. Hence the algorithm terminates with

$$V_{UB}(b_0) - V_{LB}(b_0) \le \epsilon.$$

Strategy extraction. The strategy construction of Horák et al. (2023) uses exactly the two inequalities established above. For player 1, begin with a lower-bound vector α attaining $V_{LB}(b_0)$. At each information history, play the feasible distribution associated with α and, after observing (a_1, o), replace α by the associated continuation vector in $\mathrm{conv}(\Gamma_{a_1,o})$. This construction depends only on player 1's action-observation history and the retained continuation vector; it does not require player 1

to know the current belief. If it is followed for K stages and an arbitrary feasible strategy is used thereafter, the finite-horizon induction of Horák et al. (2023) gives a value of at least

$$V_{LB}(b_0) - \gamma^K (U_0 - L_0).$$

Letting $K \to \infty$ gives a feasible strategy σ_1^{LB} such that

$$\inf_{\sigma_2} V(b_0; \sigma_1^{LB}, \sigma_2) \geq V_{LB}(b_0).$$

For player 2, choose at each stage the minimizing strategy in $[HV_{UB}](b)$. Player 2 observes the current state and knows the stage strategies it has used, so it can update b by (8). The inequality $HV_{UB} \leq V_{UB}$ gives, after K stages and an arbitrary feasible continuation, an upper bound of $V_{UB}(b_0) + \gamma^K (U_0 - L_0)$. Letting $K \to \infty$ gives a feasible strategy σ_2^{UB} such that

$$\sup_{\sigma_1} V(b_0; \sigma_1, \sigma_2^{UB}) \leq V_{UB}(b_0).$$

Since the two bounds differ by at most ϵ, neither player can gain more than ϵ by deviating. Thus $(\sigma_1^{LB}, \sigma_2^{UB})$ is an ϵ-Nash equilibrium, equivalently an ϵ-saddle point.

Finally, the game $\widehat{G}$ constructed in Theorem 2.3 satisfies the required condition. In an original state, player 1 has the common feasible action set $\{\perp_1\}$; in a dummy state, it has the common feasible action set A_v. Player 1 observes which phase is being played, while player 2 observes the transformed state and uses $B(s)$ or $\{\perp_2\}$ as appropriate. The result therefore applies to $\widehat{G}$. □

B Numerically Discovered SA-MDP Strategies

We use the standard game theory notation for mixed strategies $p[A] + (1-p)[B]$ to indicate a mixed strategy that plays action A with probability p and B with probability $1-p$, and A to indicate the pure strategy that plays A. We write the adversary's reporting strategies the same way.

B.1 Markovian Strategies

Agent We find Markovian equilibrium strategies, $\pi_1(s^0) = A,\ \pi_1(s_1) = A,\ \pi_1(s_2) = B,\ \pi_1(s_3) = C$, given observations s^0, s_1, s_2, s_3.

Adversary We find equilibrium strategies $\nu(\cdot|s^0) = s^0, \nu(\cdot|s_1) = s_2, \nu(\cdot|s_2) = \frac{1}{2}[s_2] + \frac{1}{2}[s_3]$, $\nu(\cdot|s_3) = s_3$.

B.2 History Dependent Strategies

Agent Let $\pi_1(s^0)$ denote our agent's strategy at state s^0 in the first stage, and $\pi_1(a, \tilde{s})$ denote our agent's optimal strategy after playing action a in the first stage and receiving observation $\tilde{s}$. Our agent's strategy is $\pi_1(s^0) = A$, $\pi_1(A, \tilde{s}_i) = B$ for $\tilde{s}_i \in \{s_2, s_3\}$ for on-policies strategies. Off-policy, our agent best responds $\pi_1(B, \tilde{s}_i) = B$ for $\tilde{s}_i \in \{s_2, s_3\}$. Finally, again off-policy, our agent best responds $\pi_1(C, s^0) = A$.

Adversary Our adversary best responds as: At state s^0: $\nu(\cdot|s^0) = s^0$. After A (on policy): $\nu(\cdot|s_2, A) = \frac{1}{3}[s_2] + \frac{2}{3}[s_3]$, $\nu(\cdot|s_3, A) = 1$. After B, off-policy: $\nu(\cdot|s_i, B) = s_2$, for $s_i \in \{s_1, s_2\}$. After C, off-policy: $\nu(\cdot|s^0, C) = s^0$ (which is forced due to trivial constraint).

C Analysis of Example SA-MDP Game Where History Matters

C.1 Setup

The state space and initial distribution are given by $S = \{s^0, s_1, s_2, s_3, t\},\ \mu_0 = \delta_{s^0}$. The action space is $A = \{A, B, C\}$. Horák et al. (2023) solve for infinite horizon settings. Therefore, we choose $\gamma \sim 1$, and add absorbing states after $t = 2$ that return reward 0 from then on to convert our finite horizon setting to an infinite horizon game. Similarly, t is added to our state space. To simplify the notation of our SA-MDP, we drop this dependence on t, with the understanding that after $t = 2$, rewards are 0 from then on and all transition probabilities from a state to itself are 1 regardless of the action taken.

Transitions Only transitions out of s^0 are nontrivial. Transitions are given by

$$p(s_2 \mid s^0, A) = \tfrac{1}{2}, \qquad p(s_3 \mid s^0, A) = \tfrac{1}{2}, \qquad p(s_1 \mid s^0, A) = 0, \tag{9}$$

$$p(s_1 \mid s^0, B) = \tfrac{1}{2}, \qquad p(s_2 \mid s^0, B) = \tfrac{1}{2}, \qquad p(s_3 \mid s^0, B) = 0, \tag{10}$$

$$p(s^0 \mid s^0, C) = 1. \tag{11}$$

All other states are absorbing.

Reward We take $R(s, a, s') \equiv R(s, a)$ and specify R as

	A	B	C
s^0	0	0	-10
s_1	0	-2	0
s_2	-1	2	-1
s_3	-2	-2	0

Adversary Constraint Sets We restrict the adversary to report only "neighboring" states:

$$\begin{aligned} B(s^0) &= \{s^0\}, \\ B(s_1) &= \{s_1, s_2\}, \\ B(s_2) &= \{s_2, s_3\}, \\ B(s_3) &= \{s_1, s_3\}. \end{aligned} \tag{12}$$

where at s^0 the agent has to report the (trivial) true initial state.

C.2 Markovian vs history-dependent observation adversaries

A *Markovian* adversary is any kernel $\nu(\tilde{s} \mid s)$ such that $\mathrm{supp}(\nu(\cdot \mid s)) \subseteq B(s)$ for each s. Because $B(s^0) = \{s^0\}$, we have $\nu(s^0 \mid s^0) = 1$.

For s_1, s_2, s_3, every such Markovian kernel can be parameterized by $(t_1, t_2, t_3) \in [0, 1]^3$:

$$\begin{aligned} &\nu(s_1 \mid s_1) = t_1, && \nu(s_2 \mid s_1) = 1 - t_1, \\ &\nu(s_2 \mid s_2) = t_2, && \nu(s_3 \mid s_2) = 1 - t_2, \\ &\nu(s_1 \mid s_3) = 1 - t_3, && \nu(s_3 \mid s_3) = t_3. \end{aligned} \tag{13}$$

A *history-dependent* adversary is allowed to choose its reporting kernel as a function of the full history (past states/actions/observations). So the adversary is powerless at s^0 (it must report s^0), and only matters at the second stage.

C.3 Markovian vs history-dependent observation adversaries

A *Markovian* adversary is given by a "Markov misreporting kernel" $\nu(\tilde{s} \mid s)$ such that $\mathrm{supp}(\nu(\cdot \mid s)) \subseteq B(s)$ for each s. Because $B(s^0) = \{s^0\}$, we have $\nu(s^0 \mid s^0) = 1$.

For s_1, s_2, s_3, every such Markovian kernel can be parameterized by $(t_1, t_2, t_3) \in [0, 1]^3$:

$$\begin{aligned} &\nu(s_1 \mid s_1) = t_1, && \nu(s_2 \mid s_1) = 1 - t_1, \\ &\nu(s_2 \mid s_2) = t_2, && \nu(s_3 \mid s_2) = 1 - t_2, \\ &\nu(s_1 \mid s_3) = 1 - t_3, && \nu(s_3 \mid s_3) = t_3. \end{aligned} \tag{14}$$

A *history-dependent* adversary is allowed to choose its reporting kernel as a function of the full history (past states/actions/observations).

C.4 Markovian adversary: best responses and value

The analysis of the Markovian adversary is a fair amount of exhaustion by cases due to the division of the simplex into 4 regions, each satisfying both, either, or none of them.

Setup Fix a Markovian misreporting kernel (14). Let $v_A(t_2, t_3)$ denote the agent's best-response value if it commits to playing action A in state s^0, and similarly $v_B(t_1, t_2)$ the value if it commits to playing action B in s^0. Against the kernel (t_1, t_2, t_3), the agent's best value is

$$F(t_1, t_2, t_3) \coloneqq \max\{v_A(t_2, t_3),\ v_B(t_1, t_2)\}.$$

We compute v_A and v_B, then minimize F over $(t_1, t_2, t_3) \in [0, 1]^3$.

Branch A: value $v_A(t_2, t_3)$ Under action A at s^0, the prior at the second stage is $p^A = (0, \frac{1}{2}, \frac{1}{2})$ over (s_1, s_2, s_3).

From s_2 and s_3, the observation distributions are

- s_2: report s_2 with prob. t_2, report s_3 with prob. $1 - t_2$.
- s_3: report s_1 with prob. $1 - t_3$, report s_3 with prob. t_3.

Total report probabilities under the A-branch:

$$\begin{aligned}
\mathbb{P}(\tilde{s} = s_1 \mid A) &= 0.5(1 - t_3), \\
\mathbb{P}(\tilde{s} = s_2 \mid A) &= 0.5t_2, \\
\mathbb{P}(\tilde{s} = s_3 \mid A) &= 0.5(1 - t_2) + 0.5t_3.
\end{aligned}$$

Report s_1. Only s_3 can produce report s_1, so $P(s_3 \mid \tilde{s} = s_1) = 1$. From (C.1), action C is uniquely optimal at s_3 with value 0.

Report s_2. Only s_2 can produce report s_2, so $P(s_2 \mid \tilde{s} = s_2) = 1$. At s_2, action B is optimal with value 2.

Report s_3. Report s_3 can come from s_2 or s_3. Using Bayes' rule to update beliefs, given

$$\begin{aligned}
w_2 &\coloneqq \mathbb{P}(\tilde{s} = s_3, s_2 \mid A) = 0.5(1 - t_2), \\
w_3 &\coloneqq \mathbb{P}(\tilde{s} = s_3, s_3 \mid A) = 0.5t_3
\end{aligned}$$

we get

$$P(s_2 \mid \tilde{s} = s_3, A) = \frac{w_2}{w_2 + w_3} = \frac{1 - t_2}{1 - t_2 + t_3} =: p.$$

Expected payoffs for the report s_3:

$$\begin{aligned}
\mathbb{E}[R \mid A, \tilde{s} = s_3, A] &= p(-1) + (1 - p)(-2) = -2 + p, \\
\mathbb{E}[R \mid B, \tilde{s} = s_3, A] &= p(2) + (1 - p)(-2) = -2 + 4p, \\
\mathbb{E}[R \mid C, \tilde{s} = s_3, A] &= p(-1) + (1 - p)(0) = -p.
\end{aligned}$$

Comparisons show that

- B beats C iff $p \geq 0.4$.
- B always beats A (for $p \geq 0$).
- C always beats A (for $p \leq 1$).

So, the best action for the report s_3 is B if $p \geq 0.4$ and C if $p < 0.4$. In terms of (t_2, t_3),

$$p \geq 0.4 \iff \frac{1 - t_2}{1 - t_2 + t_3} \geq 0.4 \iff 3 - 3t_2 - 2t_3 \geq 0.$$

Define the region

$$C_A \coloneqq \{(t_2, t_3) : 3 - 3t_2 - 2t_3 \geq 0\}.$$

Branch-A value. *Case A1: $(t_2, t_3) \in C_A$ (so report $s_3 \mapsto B$).*

From s_2 we always play B and get 2. From s_3 we play C with prob. $1 - t_3$ (payoff 0) and B with prob. t_3 (payoff -2), so $\mathbb{E}[R \mid s_3] = -2t_3$. Thus,

$$v_A(t_2, t_3) = 0.5 \cdot 2 + 0.5 \cdot (-2t_3) = 1 - t_3.$$

Case A2: $(t_2, t_3) \notin C_A$ (so report $s_3 \mapsto C$).

From s_2, report s_2 leads to B (payoff 2) and report s_3 leads to C (payoff -1), so

$$\mathbb{E}[R \mid s_2] = t_2 \cdot 2 + (1 - t_2)(-1) = 3t_2 - 1.$$

From s_3 we always play C and get 0. Hence,

$$v_A(t_2, t_3) = 0.5(3t_2 - 1) + 0.5(0) = 1.5t_2 - 0.5.$$

In summary:

$$v_A(t_2, t_3) = \begin{cases} 1 - t_3, & \text{if } 3 - 3t_2 - 2t_3 \geq 0, \\ 1.5t_2 - 0.5, & \text{if } 3 - 3t_2 - 2t_3 < 0. \end{cases} \tag{15}$$

Branch B: value $v_B(t_1, t_2)$ Under action B at s^0, the prior at the second stage is $p^B = (\frac{1}{2}, \frac{1}{2}, 0)$. From s_1 and s_2:

- s_1: report s_1 with prob. t_1, report s_2 with prob. $1 - t_1$.
- s_2: report s_2 with prob. t_2, report s_3 with prob. $1 - t_2$.

Total report probabilities under the B-branch:

$$\begin{aligned} \mathbb{P}(\tilde{s} = s_1 \mid B) &= 0.5t_1, \\ \mathbb{P}(\tilde{s} = s_2 \mid B) &= 0.5(1 - t_1) + 0.5t_2, \\ \mathbb{P}(\tilde{s} = s_3 \mid B) &= 0.5(1 - t_2). \end{aligned}$$

Report s_1. Only s_1 can produce report s_1; best action is A (value 0).

Report s_3. Only s_2 can produce report s_3; best action is B (value 2).

Report s_2. Report s_2 can come from s_1 or s_2. Let

$$\begin{aligned} w_1 &= \mathbb{P}(\tilde{s} = s_2, s_1 \mid B) = 0.5(1 - t_1), \\ w_2 &= \mathbb{P}(\tilde{s} = s_2, s_2 \mid B) = 0.5t_2. \end{aligned}$$

Then

$$\begin{aligned} P(s_1 \mid \tilde{s} = s_2, B) &= \frac{1 - t_1}{1 - t_1 + t_2} \\ P(s_2 \mid \tilde{s} = s_2, B) &= \frac{t_2}{1 - t_1 + t_2}. \end{aligned}$$

Expected payoffs for the report s_2:

$$\begin{aligned} \mathbb{E}[R \mid A, \tilde{s} = s_2, B] &= -\frac{t_2}{1 - t_1 + t_2}, \\ \mathbb{E}[R \mid C, \tilde{s} = s_2, B] &= -\frac{t_2}{1 - t_1 + t_2}, \\ \mathbb{E}[R \mid B, \tilde{s} = s_2, B] &= \frac{-2(1 - t_1) + 2t_2}{1 - t_1 + t_2} = \frac{-2 + 2t_1 + 2t_2}{1 - t_1 + t_2}. \end{aligned}$$

So B beats A/C iff

$$-2 + 2t_1 + 2t_2 \geq -t_2 \iff 2t_1 + 3t_2 \geq 2.$$

Define the region

$$C_B := \{(t_1, t_2) : 2t_1 + 3t_2 \geq 2\}.$$

Branch-B value. *Case B1: $(t_1, t_2) \in C_B$ (so report $s_2 \mapsto B$).*

Mapping: report $s_1 \mapsto A$, report $s_2 \mapsto B$, report $s_3 \mapsto B$.

From s_1 we know that report s_1 gives A (payoff 0) and report s_2 gives B (payoff -2), so $\mathbb{E}[R \mid s_1] = -2(1 - t_1)$. From s_2: both reports lead to B, so $\mathbb{E}[R \mid s_2] = 2$. Thus,

$$v_B(t_1, t_2) = 0.5[-2(1 - t_1)] + 0.5 \cdot 2 = t_1.$$

Case B2: $(t_1, t_2) \notin C_B$ (so report $s_2 \mapsto A$).

Mapping: report $s_1 \mapsto A$, report $s_2 \mapsto A$, report $s_3 \mapsto B$.

From s_1: always A, so $\mathbb{E}[R \mid s_1] = 0$. From s_2: report s_2 gives A (-1), report s_3 gives B (2), so $\mathbb{E}[R \mid s_2] = t_2(-1) + (1 - t_2)2 = 2 - 3t_2$. Hence,

$$v_B(t_1, t_2) = 0.5(0) + 0.5(2 - 3t_2) = 1 - 1.5t_2.$$

In summary:

$$v_B(t_1, t_2) = \begin{cases} t_1, & \text{if } 2t_1 + 3t_2 \geq 2, \\ 1 - 1.5t_2, & \text{if } 2t_1 + 3t_2 < 2. \end{cases} \tag{16}$$

C.5 Minimizing the Markovian worst-case value

For each kernel (t_1, t_2, t_3), the agent's best-response value is

$$F(t_1, t_2, t_3) = \max\{v_A(t_2, t_3),\ v_B(t_1, t_2)\},$$

with v_A and v_B given by (15)–(16). The Markovian adversary's minimax problem is

$$V_{\text{Markov}} \coloneqq \inf_{t_1, t_2, t_3 \in [0,1]} F(t_1, t_2, t_3).$$

The domain splits into four regions depending on whether the conditions defining v_A and v_B hold:

$$C_A : 3 - 3t_2 - 2t_3 \geq 0,$$
$$C_B : 2t_1 + 3t_2 \geq 2.$$

We consider all four combinations. On each region, v_A and v_B are linear, and F is the maximum of two linear functions.

Region 1: C_A and C_B both hold. Here $v_A = 1 - t_3$ and $v_B = t_1$, so

$$F = \max\{1 - t_3,\ t_1\}.$$

To minimize F we equalize the two arguments

$$t_1 = 1 - t_3.$$

We substitute these into C_B

$$2(1 - t_3) + 3t_2 \geq 2 \iff 3t_2 \geq 2t_3 \iff t_2 \geq \tfrac{2}{3}t_3.$$

From C_A

$$3 - 3t_2 - 2t_3 \geq 0 \iff 3t_2 \leq 3 - 2t_3 \iff t_2 \leq 1 - \tfrac{2}{3}t_3.$$

Together, both constraints imply

$$\tfrac{2}{3}t_3 \leq 1 - \tfrac{2}{3}t_3 \iff t_3 \leq \tfrac{3}{4}.$$

On the equality line Appendix C.5, $F = t_1 = 1 - t_3$, which is minimized by taking t_3 as large as possible, i.e. $t_3 = 3/4$, giving

$$F_{\min,\text{Region 1}} = 1 - \tfrac{3}{4} = \tfrac{1}{4}.$$

Region 2: C_A holds, C_B fails. Here $v_A = 1 - t_3$ and $v_B = 1 - 1.5t_2$, so

$$F = \max\{1 - t_3,\ 1 - 1.5t_2\}.$$

Equalizing gives

$$1 - t_3 = 1 - 1.5t_2 \iff t_3 = 1.5t_2.$$

Constraint C_A with $t_3 = 1.5t_2$:

$$3 - 3t_2 - 2(1.5t_2) \geq 0 \iff 3 - 6t_2 \geq 0 \iff t_2 \leq 0.5.$$

The condition $\neg C_B : 2t_1 + 3t_2 < 2$ only restricts t_1 and can be satisfied for some t_1 at any given $t_2 \leq 0.5$, so it does not constrain t_2 further.

On the equality line $t_3 = 1.5t_2$, the value is

$$F = 1 - 1.5t_2,$$

minimized by taking t_2 as large as possible, i.e. $t_2 = 0.5$, giving

$$F_{\min,\text{Region 2}} = 1 - 1.5 \cdot 0.5 = 0.25.$$

Region 3: C_A fails, C_B holds. Here $v_A = 1.5t_2 - 0.5$ and $v_B = t_1$, so

$$F = \max\{1.5t_2 - 0.5,\ t_1\}.$$

Equalizing gives

$$t_1 = 1.5t_2 - 0.5.$$

Constraint C_B becomes

$$\begin{aligned} 2(1.5t_2 - 0.5) + 3t_2 \geq 2 &\iff 3t_2 - 1 + 3t_2 \geq 2 \\ &\iff 6t_2 \geq 3 \\ &\iff t_2 \geq 0.5. \end{aligned}$$

The condition $\neg C_A : 3 - 3t_2 - 2t_3 < 0$ is

$$3t_2 + 2t_3 > 3.$$

For any $t_2 \geq 0.5$ this can be satisfied (e.g., by choosing t_3 close to 1), so it does not restrict t_2 beyond $t_2 \geq 0.5$.

On the equality line, $F = t_1 = 1.5t_2 - 0.5$, minimized at $t_2 = 0.5$:

$$F_{\min,\text{Region 3}} = 1.5 \cdot 0.5 - 0.5 = 0.25.$$

Region 4: C_A and C_B both fail. Here $v_A = 1.5t_2 - 0.5$ and $v_B = 1 - 1.5t_2$, so

$$F = \max\{1.5t_2 - 0.5,\ 1 - 1.5t_2\},$$

which depends only on t_2.

Equalizing:

$$1.5t_2 - 0.5 = 1 - 1.5t_2 \iff 3t_2 = 1.5 \iff t_2 = 0.5.$$

At $t_2 = 0.5$ we get

$$F = 1.5 \cdot 0.5 - 0.5 = 0.25.$$

Here, the failure of C_A and C_B requires

$$\begin{aligned} 3 - 3t_2 - 2t_3 < 0 &\iff t_3 > 0.75, \\ 2t_1 + 3t_2 < 2 &\iff t_1 < 0.25, \end{aligned}$$

so, for example, $(t_1, t_2, t_3) = (0, 0.5, 0.8)$ lies in Region 4 and has $F = 0.25$.

Conclusion In all four regions, the infimum of $F(t_1, t_2, t_3)$ is $1/4$. We also have explicit triples (for example $(0, 0.5, 0.8)$) that achieve $F = 1/4$. Thus,

$$\begin{aligned} V_{\text{Markov}} &= \inf_{t_1,t_2,t_3 \in [0,1]} F(t_1, t_2, t_3) \\ &= \min_{t_1,t_2,t_3 \in [0,1]} F(t_1, t_2, t_3) \\ &= \frac{1}{4}. \end{aligned}$$

Moreover, for any such minimizing kernel ν^*, the best-response value is exactly $1/4$, so *no (pure) agent strategy can achieve expected reward strictly larger than* $1/4$ against this optimal Markovian adversary.

C.6 History-dependent adversary

Now, allow a history-dependent adversary that can use one kernel on the A-subtree and a different kernel on the B-subtree. Concretely, at the second stage it may choose:

$$\nu^A(\cdot \mid s) \text{ after first-stage action } A$$

and

$$\nu^B(\cdot \mid s) \text{ after first-stage action } B,$$

each respecting the same neighbor constraint (12).

Then, two crucial facts in this new history dependent strategy are:

1. The agent has a baseline strategy with value 0 against *any* adversary: play any mix of A and B at s^0, and at the second stage always play action B regardless of the reported state. Under the A-branch, states $\{s_2, s_3\}$ yield payoffs $+2$ and -2 with equal probability; under the B-branch, states $\{s_1, s_2\}$ yield payoffs -2 and $+2$ with equal probability. In both cases, the expectation is 0, and misreporting does not matter if reports are ignored. Hence, $V_{\text{Hist}} \geq 0$.
2. On each branch separately, the adversary can drive the agent's best-response value down to 0.

The first fact is self-explanatory; however, the second fact requires some additional work. For the second fact, we show that there are branch-specific values that satisfy

$$\inf_{t_2,t_3} v_A(t_2, t_3) = 0, \qquad \inf_{t_1,t_2} v_B(t_1, t_2) = 0.$$

For example,

- On the A-branch, take $(t_2, t_3) = (0, 1)$, which lies in C_A and gives $v_A = 1 - t_3 = 0$.
- On the B-branch, take $(t_1, t_2) = (0, 2/3)$, which lies in $\neg C_B$ and gives $v_B = 1 - 1.5 \cdot (2/3) = 0$.

A history-dependent adversary can therefore use ν^A with $(t_2, t_3) = (0, 1)$ on the A-subtree and ν^B with $(t_1, t_2) = (0, 2/3)$ on the B-subtree. Then any policy that commits to A can achieve at most 0 on that branch, any policy that commits to B can achieve at most 0 on that branch, and mixing yields a convex combination, still at most 0. Therefore, $V_{\text{Hist}} \leq 0$. But by the fact that there is the baseline where the agent just ignores misreports and still obtains 0, we know that $V_{\text{Hist}} \geq 0$

Therefore, combining the baseline strategy with the examples we found, we obtain the exact value

$$V_{\text{Hist}} = 0.$$

C.7 Conclusion of V_{Markov} and V_{Hist} Comparison

Thus, the history-dependent observation adversary is strictly more damaging

$$V_{\text{Hist}} = 0 \; < \; \frac{1}{4} = V_{\text{Markov}}.$$

D ALGORITHM DETAILS

A high-level overview of our algorithm is provided below in Algorithm 1 and Algorithm 2.

In Algorithms 1 and 2, we write V_{LB}^{Γ} and V_{UB}^{Υ} for the lower- and upper-bound value functions denoted above by V_{LB} and V_{UB}. The superscripts emphasize that the lower bound is represented by a set Γ of α-vectors, whereas the upper bound is constructed from a set Υ of belief–value points. Thus, V_{LB}^{Γ} and V_{UB}^{Υ} refer to the same bound functions as V_{LB} and V_{UB}, with their finite representations made explicit.

Algorithm 1 Compute ϵ-equilibrium for history-dependent SA-MDP

Require: SA-MDP $M = (S, A, R, p, \gamma, B)$, initial belief $b_{\text{init}} \in \Delta(S)$, tolerance $\epsilon > 0$, (optional) horizon H

/* Make the model stationary if finite-horizon */
if H is specified **then**
  $S \leftarrow S \cup \{t\}$
  for $s \in S, a \in A$ **do**
    if $t \geq H$ **then**
      $R(s, a) \leftarrow 0 \quad \forall s \in S, a \in A$
      $p(s|s, a) \leftarrow 1$
      $p(s'|s, a) \leftarrow 0, \forall s' \neq s$
    end if
  end for{add time t to state; add absorbing terminal dynamics after $t = H$}
end if
/* Reduce SA-MDP to a constrained zero-sum OS-POSG */
$\widehat{G} \leftarrow$ PARALLELIZETOOSPOSG(M) {dummy states/actions; $\widehat{\gamma} = \sqrt{\gamma}$; attacker action encodes reported observation}
$\widehat{G} \leftarrow$ IMPOSESTATEDEPENDENTCONSTRAINTS$(\widehat{G}, B)$ {set $L_1(s) \leftarrow \{\perp_1\}$ and $L_2(s) \leftarrow B(s)$ on true states;set $L_1(\bar{s}) \leftarrow A_v$ and $L_2(\bar{s}) \leftarrow \{\perp_2\}$ on dummy states}
/* Solve $\widehat{G}$ using HSVI (Horák et al.) */
$(V_{LB}^{\Gamma}, V_{UB}^{\Upsilon}, \Gamma, \Upsilon) \leftarrow$ HSVI$(\widehat{G}, b_{\text{init}}, \epsilon)$
/* Strategy extraction (online policies) */
$\widehat{\sigma}_1 \leftarrow$ CONTINUALRESOLVING$(\widehat{G}, \Gamma, b_{\text{init}})$
$\widehat{\sigma}_2 \leftarrow$ UPPERBOUNDSTAGEPOLICY$(\widehat{G}, V_{UB}^{\Upsilon}, b_{\text{init}})$
/* Map back to SA-MDP by collapsing sub-steps */
$(\pi, \nu) \leftarrow$ COLLAPSEDUMMYSTEPS$(\widehat{\sigma}_1, \widehat{\sigma}_2)$ {drop dummy turns; attacker action $\equiv$ perturbed observation}
return $(\pi, \nu, V_{LB}^{\Gamma}(b_{\text{init}}), V_{UB}^{\Upsilon}(b_{\text{init}}))$

Algorithm 2 HSVI for constrained zero-sum OS-POSGs, blue indicates change relative to Horák et al. (2023)

Require: Constrained OS-POSG $G = (S, A_1, A_2, O, T, R, \gamma, L_1, L_2)$, initial belief b_{init}, tolerance $\epsilon > 0$, neighborhood parameter D

Initialize V_{LB}^{Γ} and V_{UB}^{Υ} (lower/upper bounds)
while $\text{EXCESS}(b_{\text{init}}, 0) > 0$ **do**
  $\text{EXPLORE}(b_{\text{init}}, 0)$
end while
return $(V_{LB}^{\Gamma}, V_{UB}^{\Upsilon}, \Gamma, \Upsilon)$
procedure $\text{EXPLORE}(b_t, t)$
$(\pi_1^{LB}, \pi_2^{LB}) \leftarrow \text{STAGEGAMEEQUILIBRIUM}(b_t, V_{LB}^{\Gamma})$ $\{\pi_1^{LB} \in \Delta(L_1(h_1));\ \pi_2^{LB}(\cdot \mid s) \in \Delta(L_2(s))$ for $s \in S(h_1)\}$
$(\pi_1^{UB}, \pi_2^{UB}) \leftarrow \text{STAGEGAMEEQUILIBRIUM}(b_t, V_{UB}^{\Upsilon})$
$\text{POINTBASEDUPDATE}(b_t, V_{LB}^{\Gamma}, V_{UB}^{\Upsilon})$
$(a_1^{\star}, o^{\star}) \leftarrow \text{FORWARDEXPLOREHEURISTIC}(b_t, t, \pi_1^{UB}, \pi_2^{LB})$
if $\Pr_{b_t, \pi_1^{UB}, \pi_2^{LB}}[a_1^{\star}, o^{\star}] \cdot \text{EXCESS}(\tau(b_t, a_1^{\star}, \pi_2^{LB}, o^{\star}), t{+}1) > 0$ **then**
  $\text{EXPLORE}(\tau(b_t, a_1^{\star}, \pi_2^{LB}, o^{\star}), t{+}1)$
end if
$\text{POINTBASEDUPDATE}(b_t, V_{LB}^{\Gamma}, V_{UB}^{\Upsilon})$
end procedure

We use a linear programming routine to solve for optimal strategies and values at each sub-game node. The code is implemented in C++ with a Python interface and uses linear programming to solve the stage-game for HSVI and the Markovian equilibrium strategies. We use CPLEX with an educational license to solve linear programming routines. (IBM, 2024)

Several implementation-specific modifications are made to our HSVI algorithm to improve computational efficiency. We cache commonly explored state-actions pairs. We also first group histories that lead to the same sets of observations at time t and then subsequently split information sets that overlap only due to the presence of a shared state in the feasibility set of the adversary into separate information sets by duplicating the state in our game tree. This enables solving many smaller convex problems versus solving fewer larger convex problems per stage and is equivalent to retaining separate extensive form (history-dependent) information sets associated with the game. However, computationally, this reduces the number of information sets that we need to keep track of and means that we only "recover the information sets as needed" at each stage.

Our performance is illustrated by Figure 6 and Table 1. In Figure 6, we observe roughly linear scaling in number of states and log-linear scaling in the depth of the game. Min-max range are provided with shaded region, while bars are provided 1 std out. Normality of runs cannot necessarily be assumed, especially with only 5 seeds. Runs were declared to have converged when the excess-gap fell below an epsilon of 0.01. SA-MDP was assumed to have $|A| = 4$, which was kept constant across tree-depth. An actual value of $\gamma = .99$ was used to approximate the $\gamma \sim 1$ case. The adversary's proximity set allows the adversary to perturb any current state to any adjacent state, with first and last states being adjacent. (So, at state 1, the adversary could perturb observations to {4,1,2}, etc.)

Figure 6: HSVI solve time versus tree-depth.

