On Dynamic Programming Theory for Leader–Follower Stochastic Games

Jilles Steeve Dibangoye[1], **Thibaut Le Marre**[2], **Ocan Sankur**[3], **François Schwarzentruber**[4]

[1]University of Groningen, j.s.dibangoye@rug.nl
[2]ENS de Lyon, CNRS, University Claude Bernard Lyon 1, Inria, LIP, UMR 5668, France, t.v.d.le.marre@rug.nl
[3]ENS de Lyon, CNRS, University Claude Bernard Lyon 1, Inria, LIP, UMR 5668, France, francois.schwarzentruber@ens-lyon.fr
[4] Univ Rennes, Inria, CNRS, IRISA, France, ocan.sankur@cnrs.fr

Abstract

Leader–follower general-sum stochastic games (LF-GSSGs) model sequential decision-making under asymmetric commitment, where a leader commits to a policy and a follower best responds, yielding a strong Stackelberg equilibrium (SSE) with leader-favourable tie-breaking. This paper introduces a dynamic programming (DP) framework that applies Bellman recursion over credible sets—state abstractions formally representing all rational follower best responses under partial leader commitments—to compute SSEs. We first prove that any LF-GSSG admits a lossless reduction to a Markov decision process (MDP) over credible sets. We further establish that synthesising an optimal memoryless deterministic leader policy is NP-hard, motivating the development of ε-optimal DP algorithms with provable guarantees on leader exploitability. Experiments on standard mixed-motive benchmarks—including security games, resource allocation, and adversarial planning—demonstrate empirical gains in leader value and runtime scalability over state-of-the-art methods.

1 Introduction

Leader–follower general-sum stochastic games (LF-GSSGs) model sequential planning under asymmetric commitment, where a leader commits to a policy and a follower best responds. The solution concept of interest is the strong Stackelberg equilibrium (SSE), which assumes that the follower resolves ties in favour of the leader. LF-GSSGs arise in numerous applications requiring robust decision-making under strategic uncertainty, including adversarial patrolling, network security, infrastructure protection, and cyber-physical planning (Yin et al. 2012; An et al. 2012; Xin Jiang et al. 2013; Basilico, Nittis, and Gatti 2016; Basilico, Coniglio, and Gatti 2016; Chung, Hollinger, and Isler 2011).

Dynamic programming (DP) has proven foundational in planning under uncertainty for Markov decision processes (MDPs) (Bellman 1957; Puterman 1994) and zero-sum stochastic games (zs-SGs) (Shapley 1953; Horák and Bošanskỳ 2019; Zheng, Jung, and Lin 2022; Horák et al. 2023), where recursive decompositions and Markovian policies enable tractable value computation. However, in LF-GSSGs, the asymmetry of commitment and general-sum structure fundamentally alter the nature of planning: the follower's response may depend on the entire interaction history, and the leader must anticipate all such best responses. Subgame decomposability fails, standard state-based recursions no longer hold, and Markov policies are not sufficient for optimality (Vorobeychik and Singh 2012; López et al. 2022).

These limitations manifest even in simple deterministic environments. In the centipede game of Figure 1, backward induction predicts that rational agents will terminate the game immediately, despite the existence of more rewarding cooperative trajectories (Rosenthal 1981; Binmore 1987; McKelvey and Palfrey 1992; Megiddo 1986; Reny 1988; Kreps 2020). This outcome illustrates the failure of Bellman-style reasoning in the presence of social dilemmas,[1] a limitation that only intensifies under stochastic transitions and asymmetric commitment.

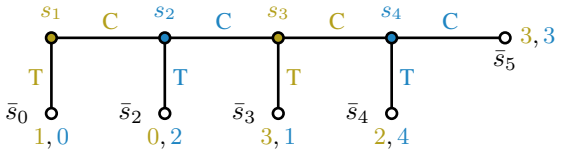

Figure 1: Centipede game. The game begins in state s_1. The leader (Player 1) acts in s_1 and s_3; the follower (Player 2) acts in s_2 and s_4. Terminal states $\bar{s}_1, \ldots, \bar{s}_5$ yield payoffs shown as (leader, follower).

The computational landscape reflects these structural difficulties. While SSEs can be computed in polynomial time for normal-form games (Conitzer and Sandholm 2006), the problem becomes NP-hard with succinct action representations (Korzhyk, Conitzer, and Parr 2010), PSPACE-complete in STRIPS-like planning domains (Behnke and Steinmetz 2024), and even NEXPTIME-complete in multi-objective arenas (Bruyère et al. 2024). Existing methods either encode the problem as large mixed-integer programs (Vorobeychik and Singh 2012; Letchford and Conitzer 2010; Vorobeychik et al. 2014) or rely on extensive linear programs (Conitzer and Sandholm 2006), both of which scale poorly with horizon or state space. Simplifying assumptions—such as myopic or omniscient followers (Denardo 1967; Whitt 1980), or stationary Markov policies (López et al. 2022)—limit practical applicability and fail to preserve generality.

This work introduces a new value-based dynamic programming framework for LF-GSSGs, grounded in a structural

[1]Situations where mutual cooperation yields the highest joint payoff, yet each agent has incentives to defect unilaterally.

reduction to what we call a *credible Markov decision process* (credible MDP). In this reformulation, states correspond to *credible sets*—finite collections of occupancy states induced by a fixed leader policy and all rational follower responses. Each occupancy state is a distribution over joint histories of environment states and actions, induced by a follower response to a partial leader policy. Transitions between credible sets are deterministic and governed by the leader's decision rules, while accounting for all follower responses that are *admissible* under SSE semantics—that is, responses that could be extended into an optimal policy with leader-favourable tie-breaking, under an extension of the leader policy consistent with the current prefix.

Contributions. The reduction is shown to be lossless: optimal leader value is preserved without requiring explicit follower enumeration. To characterise computational hardness, it is further proven that synthesising an optimal memoryless deterministic leader policy in an LF-GSSG is *NP*-hard. Nonetheless, the value function over credible sets exhibits *uniform continuity*, enabling Bellman-style recursion and point-based approximation with provable guarantees on leader exploitability. The resulting dynamic programming algorithms achieve ε-optimality without full policy enumeration and scale to large-horizon settings. Empirical evaluations on mixed-motive benchmarks—including security games, resource allocation, and adversarial planning—demonstrate improvements in leader value and runtime over MILP-based and DP-inspired baselines.

2 Background

This section formalises leader–follower general-sum stochastic games (LF-GSSGs), outlines the interaction model, and recalls key concepts from dynamic programming relevant to our framework. A summary of the notation used throughout the paper is provided in Appendix A.

Definition 1. *A* leader–follower general-sum stochastic game *(LF-GSSG) is a tuple* $M \doteq (I, S, A_L, A_F, p, r_L, r_F, s_0, \gamma)$, *where* $I = \{L, F\}$ *is the set of players, comprising the leader* L *and the follower* F*;* S *is a finite set of environment states;* A_L *and* A_F *are the finite action spaces of the leader and the follower, respectively. The transition function* $p\colon S \times A_L \times A_F \times S \to [0,1]$ *defines the probability of moving to state* s' *from* s *under joint action* (a_L, a_F)*. The one-step reward functions* $r_L, r_F\colon S \times A_L \times A_F \times S \to \mathbb{R}$ *determine the immediate rewards received by each player. The game begins in a known initial state* $s_0 \in S$ *and proceeds over an infinite number of stages, with rewards geometrically discounted by a factor* $\gamma \in [0,1)$.

Since exact infinite-horizon planning is computationally prohibitive, we approximate it using a finite planning horizon $\ell < \infty$. This is theoretically justified: for any $\varepsilon > 0$, the total expected return under a policy truncated at stage ℓ differs from its infinite-horizon value by at most ε, provided that

$$\ell \geq \left\lceil \log_\gamma \left(\frac{(1-\gamma)\varepsilon}{m} \right) \right\rceil,$$

where $m > 0$ is a uniform upper bound on one-step reward magnitudes. This follows from geometric tail decay and enables tractable approximation to arbitrary precision.

Information structure and example. The interaction unfolds over discrete time steps $t \in \{0, 1, \ldots, \ell - 1\}$. At each stage t, both players observe the current state s_t, but neither observes the other's action unless explicitly encoded in the state. Each player is informed only of the current state and their own past actions. We use the centipede game as a running example throughout the paper

Example 1. *The centipede game, shown in Figure 1 (Rosenthal 1981), involves two players who alternately choose whether to continue or terminate the game by taking a larger share of an increasing pot. If a player chooses to take, the game ends immediately with that player receiving the larger share. Despite its simplicity, the centipede game illustrates a breakdown of backward induction due to strategic interdependence: classical dynamic programming terminates too early, failing to exploit deeper cooperative potential. This failure motivates our development of a more robust planning framework for LF-GSSGs.*

Histories and decision rules. Each player makes decisions based on their *private history*. For player $i \in \{L, F\}$, the history at time t is $h_{i,t} \doteq (s_0, a_{i,0}, s_1, \ldots, a_{i,t-1}, s_t)$, capturing the sequence of observed states and the player's own actions. Let $H_{i,t}$ denote the set of such histories. A *decision rule* is a mapping $\delta_{i,t}\colon H_{i,t} \to \Delta(A_i)$; the leader may use stochastic rules, while the follower is restricted to deterministic ones (Conitzer and Sandholm 2006). Let $\Delta_{i,t}$ be the set of all such rules. A *policy* is a sequence $\pi_i = (\delta_{i,0}, \ldots, \delta_{i,\ell-1})$, and the joint policy is $\pi = (\pi_L, \pi_F)$.

Value functions under a joint policy. Fix a joint policy $\pi = (\pi_L, \pi_F)$. For player i, the action-value function at stage t is defined for environment state s, joint history $h = (h_L, h_F)$, and joint action $a = (a_L, a_F)$ as

$$q^\pi_{i,t}(s, h, a) \doteq \mathbb{E}_\pi \left[\textstyle\sum_{k=t}^{\ell-1} \gamma^{k-t} r_i(s_k, a_{L,k}, a_{F,k}, s_{k+1}) \right],$$

where the expectation is taken over future trajectories consistent with policy π, given that history (h_L, h_F) is observed at stage t and joint action (a_L, a_F) is executed. The environment state s is included explicitly (though determined by the histories) to simplify belief updates in later constructions.

The value function at stage t is then

$$v^\pi_{i,t}(s, h_L, h_F) \doteq \mathbb{E}\left[q^\pi_{i,t}(s, h_L, h_F, a_L, a_F) \right],$$

with the expectation taken over $a_L \sim \delta_{L,t}(h_L)$, $a_F = \delta_{F,t}(h_F)$, and $s_{t+1} \sim p(s_t, a_L, a_F, \cdot)$. The terminal value is $v^\pi_{i,\ell} \equiv 0$ (Puterman 1994).

Follower best-response set. Given a fixed leader policy π_L, the set of follower best responses is

$$\mathtt{BR}(\pi_L) \doteq \operatorname{argmax}_{\pi_F \in \Pi_F} v^{(\pi_L, \pi_F)}_{F,0}(s_0, s_0, s_0).$$

To evaluate leader performance against optimal responses, define the leader-evaluated action-value function as $q^{\pi_L}_{i,t}(s, h, a) \doteq \max_{\pi_F \in \mathtt{BR}(\pi_L)} q^{(\pi_L, \pi_F)}_{i,t}(s, h, a)$, and the corresponding value function:

$$v^{\pi_L}_{i,t}(s, h_L, h_F) \doteq \max_{\pi_F \in \mathtt{BR}(\pi_L)} \mathbb{E}[q^{(\pi_L, \pi_F)}_{i,t}(s, h_L, h_F, a_L, a_F)].$$

Strong Stackelberg equilibrium. A leader policy $\pi_L^\star$ induces a *strong Stackelberg equilibrium* (SSE) if

$$\pi_L^\star \in \operatorname{argmax}_{\pi_L \in \Pi_L} \max_{\pi_F \in \mathtt{BR}(\pi_L)} v_{L,0}^{(\pi_L,\pi_F)}(s_0, s_0, s_0).$$

Limitations of state-of-the-art approaches. State-of-the-art methods for computing SSEs in LF-GSSGs fall into two main categories: solving large-scale MILPs (Vorobeychik and Singh 2012; Letchford and Conitzer 2010; Vorobeychik et al. 2014) and extensive linear programs (Conitzer and Sandholm 2006). These approaches rely on normal-form representations and scale poorly, while offering limited structural insight. Dynamic programming, highly effective in zero-sum and common-payoff settings, fails to extend to general-sum games. A natural but flawed idea is to compute, at each state and stage, a leader action that maximises expected payoff given a best-responding follower. The following proposition shows that replacing a single-stage leader decision rule using backward induction—even under the full follower value function—can invalidate equilibrium guarantees.

Proposition 2.1 (Proof in Appendix B.1)**.** *Let π_L be a Markov leader policy. Define a new policy π'_L that differs from π_L only at stage $t < \ell$, where π'_L selects a decision rule $\delta'_{L,t}(s)$ solving:*

$$\max_{\delta_L(s)\in\Delta(A_L)} \max_{a_F \in A_F} \mathbb{E}_{a_L\sim\delta_L(s)}\left[q_{L,t}^{\pi_L}(s,s,s,a_L,a_F)\right]$$
$$\text{s.t.} \quad \mathbb{E}_{a_L\sim\delta_L(s)}\left[q_{F,t}^{\pi_L}(s,s,s,a_L,a_F)\right] = v_{F,t}^{\pi_L}(s,s,s).$$

Then π'_L may fail to induce a strong Stackelberg equilibrium, even if $\delta'_{L,t}(s) = \delta_{L,t}(s)$ for all s and all other stages.

This result motivates the *myopic follower* approximation, which replaces the full follower value function with immediate reward (Fujiwara 2006; López et al. 2022; Zhong et al. 2023). Proposition 2.1 shows that even when using the correct long-term value function, backward induction may yield policies that violate SSE optimality. The myopic simplification is thus even less principled and less reliable in general. Such failure arises in domains like CENTIPEDE, where locally rational leader deviations may trigger early termination and suboptimal outcomes (e.g., $(1, 0)$ instead of $(2, 4)$). This breakdown stems from two structural barriers: (i) the leader faces a *multi-objective* optimisation problem, in which actions optimal against one best response may be dominated under another; and (ii) the problem is *non-Markovian* from the follower's perspective, making correct interpretation of leader actions dependent on history.

How can a leader-centric perspective support a reliable and scalable dynamic programming framework for computing strong Stackelberg equilibria in LF-GSSGs?

3 Credible Markov Decision Processes

We present a lossless reduction of any leader–follower general-sum stochastic game (LF-GSSG) into a Markov decision process (MDP) from the leader's perspective. The core idea is to model the leader's planning task as a sequential decision process in which each state summarises the game's evolution under all rational follower responses. We call the resulting model the *credible MDP*.

The credible MDP is constructed step by step. We first define *joint decision-rule histories* and the *occupancy states* they induce. These occupancy states are grouped into *credible sets*, each indexed by a partial leader policy. We then define the deterministic transitions and cumulative rewards associated with these sets. To reduce redundancy, we introduce a filtering mechanism based on value-dominance within belief classes. This enables forward planning while preserving all information relevant to equilibrium computation. A formal proof that the credible MDP constitutes a *lossless reduction* is provided in Section 4.

Occupancy States and Decision-Rule Histories

At each stage t, the decision process is indexed by joint decision-rule histories $\theta_t = (\theta_{L,t}, \theta_{F,t})$, where $\theta_{i,t} = (\delta_{i,0}, \ldots, \delta_{i,t-1})$ denotes the sequence of local decision rules applied by agent $i \in I$ up to time step $t-1$. Each joint history θ_t induces an *occupancy state* o_{θ_t}, defined as the distribution over joint local environment histories given θ_t:

$$o_{\theta_t} : (s_t, h_{L,t}, h_{F,t}) \mapsto \Pr(s_t, h_{L,t}, h_{F,t} \mid \theta_t).$$

Although s_t is is already given in $h_{L,t}$ and $h_{F,t}$, we explicitly mention it in o_{θ_t} to support belief-based filtering over states. This distribution is induced by the initial state, the system dynamics, and the policy history θ_t. Given any occupancy state o_θ and a joint decision rule $\delta = (\delta_L, \delta_F)$, we define a deterministic transition operator:

$$\tau : (o_\theta, \delta) \mapsto o_{(\theta,\delta)},$$

where $o_{(\theta,\delta)}$ is the next occupancy state resulting from applying δ to o_θ and propagating through the dynamics. Formally, for any state s', joint history (h_L, h_F), and joint action (a_L, a_F) : $o_{(\theta,\delta)}(s', (h_L, a_L, s'), (h_F, a_F, s')) = \sum_s o_\theta(s, h_L, h_F)\delta_L(a_L|h_L)\delta_F(a_F|h_F)p(s, a_L, a_F, s')$. In addition, we define the *value-so-far function* $\rho_i : o_t \mapsto \mathbb{R}$ for each agent $i \in I$, mapping each occupancy state o_t to the expected cumulative discounted reward accrued up to stage t:

$$\rho_i(o_t) \doteq \mathbb{E}_{o_t}\left[\sum_{t'=0}^{t-1} \gamma^{t'}\, r_i(s_{t'}, a_{L,t'}, a_{F,t'}, s_{t'+1})\right]$$

where the expectation is taken with respect to the joint distribution over histories encoded in o_t. Note that o_t encodes sufficient information to evaluate this cumulative reward.

Credible Sets under Epistemic Pessimism

We now define the credible state space for the leader's planning problem. At stage t, the leader's local decision-rule history $\theta_{L,t}$ is fully specified, while the follower's decision-rule history $\theta_{F,t}$ is unknown and must be reasoned over. To account for all possible rational follower behaviours consistent with the game structure, we adopt a pessimistic closure over follower responses. Specifically, we define the *credible set* at stage t as the set of occupancy states reachable from the known leader decision-rule history and all follower decision-rule histories:

$$c_{\theta_{L,t}} \doteq \left\{ o_{(\theta_{L,t},\theta_{F,t})} \;\middle|\; \theta_{F,t} \in \Theta_{F,t} \right\},$$

where $\Theta_{F,t}$ is the set of all follower decision-rule histories of length t. When the leader selects a new decision rule $\delta_{L,t}$, her decision-rule history updates to $\theta_{L,t+1} = \theta_{L,t}\delta_{L,t}$. The corresponding successor credible set is defined by applying the transition operator τ to each element of $c_{\theta_{L,t}}$ under every follower decision rule:

$$c_{\theta_{L,t+1}} \doteq \left\{\tau(o_t, (\delta_{L,t}, \delta_{F,t})) \mid o_t \in c_{\theta_{L,t}}, \delta_{F,t} \in \Delta_{F,t}\right\},$$

where $c_{\theta_{L,t+1}} \doteq T(c_{\theta_{L,t}}, \delta_{L,t})$ denotes the resulting set. This transition is deterministic from the leader's perspective and compactly encodes all possible follower responses compatible with epistemic pessimism.

Marginal Beliefs and Their Role in Filtering

Forward search over credible sets leads to a rapid growth in the number of reachable occupancy states. However, many of these states may be strategically redundant *under Markov policies*: they differ in the distribution over local histories but induce the same belief over the current environment state s_t. In order to capture this redundancy, we define the *marginal belief* induced by a joint decision-rule history θ as

$$b_{o_\theta}(s_t) \doteq \sum_{(h_{L,t}, h_{F,t})} o_\theta(s_t, h_{L,t}, h_{F,t}).$$

Since the game is fully observable and Markovian, the marginal belief b_{o_θ} summarises all information relevant for predicting future transitions and rewards. However, it is important to emphasise that *belief equivalence does not imply full equivalence of occupancy states*. Two occupancy states may induce the same marginal belief but differ in their distributions over local histories, and thus in the cumulative reward they have already accrued. As a result, belief equivalence is insufficient for reducing the occupancy-state space directly. Nevertheless, when comparing future continuations from belief-equivalent states, the belief is sufficient to reason about the *value-to-go*. We exploit this fact to define a pruning procedure over credible sets that removes dominated occupancy states sharing the same marginal belief.

Belief-Based Filtering over Credible Sets

To mitigate the combinatorial blow-up during planning, we introduce a *value-dominance filtering* operation under Markov policies. It retains only those occupancy states that are not strictly Pareto-dominated and not exact duplicates within the same marginal belief class. Let $\rho_L(o)$ and $\rho_F(o)$ denote the cumulative value-so-far for the leader and follower, respectively, under occupancy state o. Given a credible set c, we define $B_c \doteq \{b_o \mid o \in c\}$ and, for each $b \in B_c$, define the belief class

$$c_b \doteq \{o \in c \mid b_o = b\}.$$

For all c_b, for any occupancy state $o \in c_b$, define c_b^o as:

$$\{o' \in c_b \mid \forall i \in I, \rho_i(o') \geq \rho_i(o),\ \exists j \in I, \rho_j(o') > \rho_j(o)\}.$$

The set c_b^o contains the occupancy states whose the margin belief is b and where the values-so-far Pareto-dominates o. We then define the filtered set of c as

$$\mathbb{F}(c) \doteq \cup_{b \in B_c} \texttt{Unique}(\{o \in c_b \mid c_b^o = \emptyset\}),$$

where $\texttt{Unique}(\cdot)$ retains exactly one representative[2] per unique cumulative value-so-far vector (ρ_L, ρ_F). That is, for each marginal belief profile b, we retain all Pareto-undominated occupancy states and eliminate any redundant duplicates with identical accumulated returns. Applying this operator to any credible set c yields the filtered set $\tilde{c} \doteq \mathbb{F}(c)$. This pruning step guarantees that all strategically relevant occupancy states are preserved: since the marginal belief captures the information required for computing continuation values, and ρ_i accounts for the accumulated return, the filtered set $\tilde{c}$ supports sound and efficient forward planning within the credible MDP. A formal proof of the losslessness of the filtering operator is provided in Appendix D.

Reward Model and Terminal Sets

We define a reward function R over credible sets that captures the leader's cumulative return at terminal stages. Intuitively, at intermediate stages $t < \ell$, the leader's policy is still under construction, and uncertainty about the follower's best response prevents assigning a meaningful reward. In contrast, at stage ℓ, the leader's policy is fixed, and all follower responses consistent with epistemic pessimism are known. This enables a reliable assessment of the total reward. Let $\tilde{F} \doteq \{\tilde{c}_{\theta_{L,\ell}} \in \tilde{C} \mid |\theta_{L,\ell}| = \ell\}$ denote the terminal filtered credible sets. The reward function R is defined by $R(\tilde{c}) \doteq 0$ if $\tilde{c} \notin \tilde{F}$, and for all $\tilde{c} \in \tilde{F}$,

$$R(\tilde{c}) \doteq \max\left\{\rho_L(o) \,\middle|\, o \in \operatorname*{argmax}_{o' \in \tilde{c}} \rho_F(o')\right\}.$$

The reward is thus zero during transient planning and evaluated only once the leader's policy is complete, ensuring strategic soundness at termination.

Formal Definition

Definition 2. *Given a leader–follower general-sum stochastic game M with horizon ℓ, the* credible MDP *is defined as $M' \doteq (\tilde{C}, \Delta_L, \tilde{T}, R, \tilde{F}, c_0, \ell)$ where:*

- $\tilde{C} \doteq \{\tilde{c}_{\theta_{L,t}} \mid \theta_{L,t} \in \Theta_{L,t}, t \in \{0, 1, \ldots, \ell\}\}$ *is the set of filtered credible sets;*
- Δ_L *is the set of admissible leader decision rules;*
- $\tilde{T}\colon (\tilde{c}, \delta_L) \mapsto \mathbb{F}(T(\tilde{c}, \delta_L))$ *is the deterministic transition function;*
- $R\colon \tilde{C} \to \mathbb{R}$ *is the reward function defined above;*
- $\tilde{F} \subseteq \tilde{C}$ *is the set of terminal filtered credible sets at stage ℓ;*
- $c_0 \in \tilde{C}$ *is the initial (filtered) credible set induced by $s_0 \in M$;*
- $\ell \in \mathbb{N}$ *is the finite planning horizon.*

[2] To define $\texttt{Unique}(\{o \in c_b \mid c_b^o = \emptyset\})$, we first partition $\Omega := \{o \in c_b \mid c_b^o = \emptyset\}$ in sets $\Omega^{\nu_L,\nu_F} = \{o \in \Omega \mid \rho_L(o) = \nu_L, \rho_F(o) = \nu_F\}$. Then we consider the function $\texttt{Choose}$ that picks an element in a set: $\texttt{Choose}(X)$ is an element in X for all non-empty sets X. Then $\texttt{Unique}(\{o \in c_b \mid c_b^o = \emptyset\}) = \{\texttt{Choose}(\Omega^{\nu_L,\nu_F}) \mid \Omega^{\nu_L,\nu_F} \neq \emptyset \text{ and } \nu_L, \nu_F \in \mathbb{R}\}$.

Example 2. *Figure 2 illustrates the structure of the credible MDP prior to belief-based filtering, using the centipede game under a deterministic leader. Nodes represent occupancy states summarised by their realised state and values-so-far.*

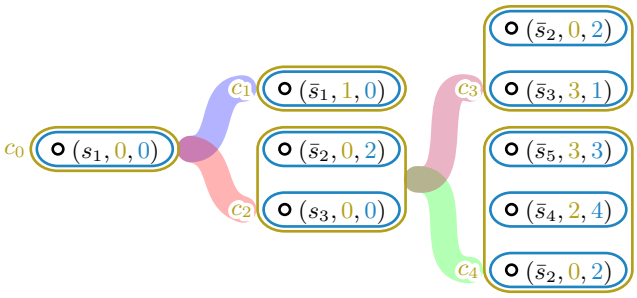

Figure 2: Reduction of the centipede game (Figure 1) into the credible MDP prior to filtering. In this example, the filtering step has no effect. Each node is annotated with a tuple (s_t, ρ_L, ρ_F), where s_t denotes the current state, and ρ_L, ρ_F denote the cumulative discounted rewards for the leader and follower, respectively. Determinism ensures that each such tuple corresponds to a unique occupancy state. Brown contours indicate credible sets, each indexed by the leader's decision-rule history. River arcs depict epistemically pessimistic transitions, capturing the leader's choice under all admissible follower responses.

The construction of M' equips the leader with a forward-searchable state space that faithfully captures all rational follower responses through credible sets. This reduction preserves the strategic essence of the original game while enabling algorithmic tractability through dynamic programming techniques. Importantly, it avoids the pitfalls of earlier formulations that relied on game states or belief states lacking temporal and strategic coherence. Having completed the definition of the reduced model, we now turn to the formal analysis of its computational and structural properties.

Lossless Reduction via the Credible MDP

The concept of a *lossless reduction* formalises when a surrogate model preserves all decision-relevant structure of an LF-GSSG, including planning semantics and equilibrium outcomes, without introducing spurious policies or additional informational power (Sanjari, Başar, and Yüksel 2023).

Definition 3 (Lossless reduction)**.** *A reduction from an LF-GSSG M to a surrogate model M' is said to be* lossless *if the following three conditions hold: (i)* value preservation *— for every leader policy $\pi_L \in \Pi_L$, there exists a policy $\pi'_L \in \Pi'_L$ such that $v^{\pi_L}_{M,L,0}(s_0) = v^{\pi'_L}_{M',L,0}(s_0)$; (ii)* equilibrium correspondence *— there exists a* surjective *map from the set of SSEs in M onto the set of SSEs in M', preserving leader and follower payoffs; (iii)* information compatibility *— the surrogate model M' introduces no additional observability or decision flexibility beyond what is available in M.*

The credible MDP M', defined in Section 3, satisfies all three conditions above and thus serves as a faithful surrogate model for planning in LF-GSSGs.

Theorem 3.1 (Proof in Appendix C)**.** *The credible MDP M' associated with the LF-GSSG M constitutes a lossless reduction in the sense of Definition 3.*

Proof sketch. Every leader policy $\pi_L = (\delta_{L,0}, \ldots, \delta_{L,\ell-1})$ in M induces a unique sequence of decision rules, which can be executed identically in M' to obtain the same trajectory of credible sets and cumulative value. Conversely, each policy in M' is defined by a sequence of decision rules that correspond to a well-defined leader policy in M, ensuring value preservation. Because the follower reactions embedded in M' are precisely the admissible best responses under SSE semantics, all equilibrium solutions in M' correspond to equilibria in M, establishing surjectivity. Finally, both models rely exclusively on private histories and decision rules defined in M, satisfying information compatibility. □

4 Theoretical Results for Credible MDPs

This section establishes the theoretical foundation for solving the original leader–follower stochastic game M through its lossless reduction to the credible MDP M'. We present three main results, each addressing a fundamental aspect: computational complexity, dynamic programming structure, and geometric properties of the value function. Together, they provide both theoretical guarantees and structural insights that enable the design of scalable solution algorithms.

We first establish that computing an optimal leader policy is *NP*-hard, even when restricted to Markov policies. This result highlights the intrinsic computational intractability of synthesising equilibrium strategies, motivating the need for approximate approaches.

Theorem 4.1 (Proof in Appendix D.4)**.** *The deterministic memoryless SSE decision problem is NP-hard both in the finite-horizon and infinite-horizon cases.*

To characterise the recursive structure of leader decision-making over credible sets, we first formalise the optimal state- and action-value functions, which together form the basis of Bellman's optimality equations.

Definition 4 (Optimal state-value function)**.** *Let $\pi_L \doteq (\delta_{L,0}, \ldots, \delta_{L,\ell-1})$ be a leader policy. The leader value function $v^{\pi_L}_{L,t} \colon \tilde{C} \to \mathbb{R}$ at stage t under π_L is given as follows: for any terminal credible set $\tilde{c} \in \tilde{F}$, we have $v^{\pi_L}_L(\tilde{c}) = R(\tilde{c})$ and for any transient credible set $\tilde{c} \notin \tilde{F}$ at stage t,*

$$v^{\pi_L}_{L,t} \colon \tilde{c} \mapsto R\big(\tilde{T}\big(\tilde{T}(\cdots \tilde{T}(\tilde{c}, \delta_{L,t}) \cdots, \delta_{L,\ell-2}), \delta_{L,\ell-1}\big)\big).$$

The optimal (leader) state-value function $v^*_{L,t} \colon \tilde{C} \to \mathbb{R}$ at stage t of surrogate model M' is given:

$$v^*_{L,t} \colon \tilde{c} \mapsto \max_{\pi_L \in \Pi_L} v^{\pi_L}_{L,t}(\tilde{c}).$$

Definition 5 (Optimal action-value function)**.** *Let $\pi_L \doteq (\delta_{L,0}, \ldots, \delta_{L,\ell-1})$ be a leader policy. The optimal leader action-value function $q^{\pi_L}_{L,t} \colon \tilde{C} \times \Delta_L \to \mathbb{R}$ at stage t under π_L is defined, for any transient credible set $\tilde{c} \in \tilde{C}$ and leader decision rule δ_L, as*

$$q^{\pi_L}_{L,t} \colon (\tilde{c}, \delta_L) \mapsto v^{\pi_L}_{L,t}(\tilde{T}(\tilde{c}, \delta_L)),$$

where $\tilde{T}(\tilde{c}, \delta_L)$ is the successor credible set induced by δ_L starting in credible set $\tilde{c}$. The optimal (leader) action-value function $q^*_{L,t} \colon \tilde{C} \times \Delta_L \to \mathbb{R}$ at stage t of surrogate model M' is given:

$$q^*_{L,t} \colon (\tilde{c}, \delta_L) \mapsto \max_{\pi_L \in \Pi_L} q^{\pi_L}_{L,t}(\tilde{c}, \delta_L).$$

We now show that these value functions satisfy Bellman optimality equations, formalising how optimal values propagate across decision stages.

Theorem 4.2 (Proof in Appendix D.7)**.** *The optimal leader state-value function $v^*_{L,t} \colon \tilde{C} \to \mathbb{R}$ at stage $t < \ell$ satisfies:*

$$v^*_{L,t} \colon \tilde{c} \mapsto \max_{\delta_L \in \Delta_L} v^*_{L,t+1}\big(\tilde{T}(\tilde{c}, \delta_L)\big).$$

We then establish a key regularity result, showing that the optimal value function is piecewise linear and convex over marginal belief states. This property underpins the design of point-based approximation methods by guaranteeing that value representations can be compactly expressed.

Define, for any $i \in \{L, F\}$, the value function at stage t induced by the vector set Γ at an occupancy state o as

$$v^{\Gamma}_{i,t}(o) \doteq \rho_i(o) + \gamma^t \sum_{h_F} \Pr(h_F \mid o) \max_{\alpha \in \bar{\Gamma}(o_{h_F})} \alpha_i(o_{h_F}),$$

where $\bar{\Gamma}(o_{h_F}) \doteq \operatorname{argmax}_{\alpha \in \Gamma} \alpha_F(o_{h_F})$ is a filtered vector set induced by Γ and $o_{h_F} \colon (s, h_L) \mapsto \frac{o(s,h_L,h_F)}{\Pr(h_F \mid o)}$ is a conditional occupancy state.

Theorem 4.3 (Proof in Appendix D.11)**.** *The optimal leader value function $v^*_{L,t} \colon \tilde{C} \to \mathbb{R}$ at stage t, solution of Bellman's optimality equations (Theorem 4.2), is uniformly continuous across credible sets. There exists a collection Λ_t of finite sets Γ of value vectors (α_L, α_F) linear across conditional occupancy states at stage t:*

$$v^*_{L,t} \colon \tilde{c} \mapsto \max_{\Gamma \in \Lambda_t} \left\{ v^{\Gamma}_{L,t}(o) \;\middle|\; o \in \operatorname*{argmax}_{o' \in \tilde{c}} v^{\Gamma}_{F,t}(o') \right\}.$$

Uniform continuity guarantees that small perturbations in credible sets lead to small changes in value, a prerequisite for approximating $v^*_{L,t}$ using a finite set of representative credible sets. To approximate optimal value function $v^*_{L,t}$, we introduce a point-based value iteration (PBVI) scheme (Pineau et al. 2003; Peralez et al. 2024, 2025). Building on the uniform continuity property, we now formalise how an approximate greedy leader decision rule can be computed efficiently for any given credible set, showing that it reduces to solving a mixed-integer linear program (MILP).

Theorem 4.4 (Proof in Appendix D.14)**.** *Let Λ_{t+1} be a finite collection of finite sets Γ_{t+1} of linear functions approximating the optimal leader value function v^*_L at stage $t{+}1$. Then, for any credible set $\tilde{c}$ at stage t, the greedy leader decision rule $\delta^{\tilde{c}}_L$ maximising the leader action-value is the solution of a MILP (see Appendix D, Table 3).*

Intuitively, the MILP exploits the uniform continuity structure of the action-value functions to jointly select leader decisions and consistent follower best responses, using integer variables to encode max-selection constraints. Next, we leverage these insights to design practical algorithms for scalable equilibrium computation.

5 Point-Based Value Iteration

For computing an SSE in M, we adapt point-based value iteration (PBVI, *see* Algorithm 1) Pineau et al. (2003) to solving credible MDP M' induced by LF-GSSG M. The algorithm alternates between three phases: expansion of credible sets alongside with (conditional) occupancy states, improvement of leader value function, and pruning to control complexity. The pseudo-code is given in Section E.

The expansion phase (*see* Algorithm 2) samples a finite set $\tilde{C}' \subset \tilde{C}$ of representative credible sets through forward simulations. Starting with the initial state s_0, each credible set $\tilde{c} \in \tilde{C}'$ is expanded by selecting leader decisions $\delta_L \in \Delta_L$, and populating occupancy states in $\tilde{T}(\tilde{c}, \delta_L)$ by sampling the corresponding follower response $\delta_F \in \Delta_F$ starting in an occupancy state $o \in \tilde{c}$, and applying the transition $\tau(o, \delta_L, \delta_F)$ to generate new occupancy states. Newly reached occupancy states are retained only if their ℓ_1-distance from the current credible set exceeds a fixed threshold, ensuring coverage. The improvement phase (*see* Algorithm 3) computes, through point-based backups and mixed-integer linear programming, a finite set $\Gamma^{\tilde{c}}$ of value vectors (α_L, α_F) linear over conditional occupancy states for every credible sample set $\tilde{c} \in \tilde{C}'$. The collection $\Lambda \doteq \{\Gamma^{\tilde{c}} \mid \tilde{c} \in \tilde{C}'\}$ induces leader value function v_L which approximates optimal leader value function v^*_L. This point-based update procedure iterates until convergence or computational resources are exhausted. To further improve scalability, we incorporate two pruning strategies. The first (Algorithm 4) eliminates dominated sets of value vectors. The second (Algorithm 5) removes noncredible sets. Both procedures introduce approximation error, but yield significant computational savings.

To analyse the approximation quality of point-based value iteration (PBVI) in the credible MDP M', we define a sample-based approximation of the optimal leader value function over a finite set $\tilde{C}' \subset \tilde{C}$ of sampled credible sets. Let $\Lambda_0, \ldots, \Lambda_\ell$ be the collections of approximate value vectors constructed via Theorem 4.4, inducing an approximate leader value function $v_{L,t} \colon \tilde{C} \to \mathbb{R}$ at stage t, generalised to arbitrary credible sets $\tilde{c} \in \tilde{C}$ via nearest-neighbour evaluation:

$$v_{L,t}(\tilde{c}) = \min \left\{ v_{L,t}(\tilde{c}^*) \;\middle|\; \tilde{c}^* \in \arg\min_{\tilde{c}' \in \tilde{C}'} d_H(\tilde{c}, \tilde{c}') \right\}.$$

The distance $d_H(\tilde{c}, \tilde{c}')$ denotes the *Hausdorff distance* between finite subsets $\tilde{c}, \tilde{c}' \subseteq O$ of occupancy states:

$$d_H(\tilde{c}, \tilde{c}') \doteq \max \left\{ \sup_{o \in \tilde{c}} \inf_{\bar{o} \in \tilde{c}'} \|o - \bar{o}\|_1,\ \sup_{\bar{o} \in \tilde{c}'} \inf_{o \in \tilde{c}} \|\bar{o} - o\|_1 \right\}.$$

Let $v^*_{L,t} \colon \tilde{C} \to \mathbb{R}$ be the optimal leader value function defined in Theorem 4.2, and let $m \doteq \max\{\|r_L\|_\infty, \|r_F\|_\infty\}$ denote a uniform bound on instantaneous payoffs. By Theorem 4.3, the mapping $\tilde{c} \mapsto v^*_{L,t}(\tilde{c})$ is uniformly continuous with respect to d_H. We now express the worst-case approximation error incurred by nearest-neighbour PBVI in terms of the sampling resolution:

Theorem 5.1. *Let $\sigma \doteq \sup_{\tilde{c} \in \tilde{C}} \min_{\tilde{c}' \in \tilde{C}'} d_H(\tilde{c}, \tilde{c}')$ be the Hausdorff covering radius of the PBVI sample set $\tilde{C}'$. Then*

the exploitability ε of the leader policy returned by PBVI satisfies: $\varepsilon \leq \frac{2m\sigma}{(1-\gamma)^2}\left[1+\ell\gamma^{\ell+1}-(\ell+1)\gamma^{\ell}\right]$.

6 Experimental Evaluation

We evaluate the empirical performance of our dynamic programming framework for computing strong Stackelberg equilibria (SSEs) in leader–follower general-sum stochastic games (LF-GSSGs). The main objective is to quantify the impact of history-dependent planning on leader value and runtime across diverse strategic settings. Experiments were run on a machine with 64 GB RAM and a 4.9 GHz CPU, with a timeout of 10 minutes per run. Linear programs and mixed-integer programs were solved using ILOG CPLEX. Source code will be released publicly.

We compare six planning variants reported in Appendix F: **H** denotes point-based value iteration (PBVI) over credible sets induced by history-dependent policies; **S** restricts the leader to Markov (state-dependent) policies, which limit the expressiveness of the resulting credible sets compared to history-dependent planning; **BI** applies backward induction using the full follower value function; **MY** is a myopic variant that considers only stage-wise follower rewards; **LP** solves a linear program per follower policy (Conitzer and Sandholm 2006); and **MILP** encodes the full strategy space in a single mixed-integer program (Vorobeychik et al. 2014). The benchmark suite includes five domains: CENTIPEDE, MATCH, PATROLLING, DEC-TIGER, and MABC. These domains span cooperative, competitive, and mixed-motive scenarios, with varying reliance on history. Formal game definitions are deferred to Appendix G. Table 1 summarises, for each method and horizon length ℓ, the resulting SSE leader value (V) and runtime (T in seconds). Additional sensitivity analyses appear in Appendix H.

Results and insights. In history-sensitive games such as CENTIPEDE and MATCH, the history-aware PBVI variant **H** consistently achieves the highest leader value, highlighting the benefit of planning over credible sets. By contrast, **S** performs similarly in domains where history plays a limited strategic role (*e.g.,* DEC-TIGER, MABC, PATROLLING). Both **BI** and **MY** often fail to induce valid SSEs because backward induction may ignore how a local policy deviation retroactively alters the follower's interpretation of prior commitment, as formalised in Proposition 2.1. Both **LP** and **MILP** become computationally intractable as the horizon increases due to exponential growth in the space of deterministic history-dependent strategies. Finally, in fully cooperative or adversarial domains (*e.g.,* MABC, DEC-TIGER, PATROLLING), all methods converge to the same SSE value under cooperative or adversarial symmetry, consistent with DP solutions for MDPs (Puterman 1994) and zero-sum SGs (Shapley 1953).

	H		**S**		**BI**		**MY**		**LP**		**MILP**	
ℓ	V	T	V	T	V	T	V	T	V	T	V	T
CENTIPEDE (Rosenthal 1981)												
1	**1**	0.01	**1**	0.01	**1**	0.01	**1**	0.01	**1**	0.25	**1**	0.14
2	**1**	0.01	**1**	0.01	**1**	0.01	**1**	0.01	**1**	20.21	**1**	7.29
3	**2**	0.01	**2**	0.01	1	0.01	1	0.01	—	—	—	—
6	**4.67**	3.68	**4.67**	0.11	1	0.17	1	0.13	—	—	—	—
MATCH (Vorobeychik and Singh 2012)												
1	**-1000**	0.01	**-1000**	0.01	**-1000**	0.01	**-1000**	0.01	**-1000**	0.2	**-1000**	0.3
2	**-1000**	0.01	**-1000**	0.01	**-1000**	0.01	**-1000**	0.01	—	—	—	—
3	**0**	0.01	-1000	0.01	-1000	0.04	-1000	0.02	—	—	—	—
6	**0**	0.03	-2000	0.03	-2000	0.03	-2000	0.03	—	—	—	—
DEC-TIGER (Nair et al. 2003)												
1	**20**	0.01	**20**	0.01	**20**	0.01	**20**	0.01	**20**	0.19	**20**	0.22
2	**40**	0.01	**40**	0.01	**40**	0.01	**40**	0.01	**40**	1.07	**40**	0.23
3	**60**	0.04	**60**	0.03	**60**	0.01	**60**	0.01	—	—	—	—
6	**120**	0.45	**120**	0.30	**120**	0.20	**120**	0.10	—	—	—	—
MABC (Hansen, Bernstein, and Zilberstein 2004)												
1	**1**	0.01	**1**	0.01	**1**	0.06	**1**	0.03	**1**	0.20	**1**	0.19
2	**2**	0.11	**2**	0.05	**2**	0.09	**2**	0.08	—	—	—	—
3	**2.99**	0.73	**2.99**	0.49	**2.99**	0.05	**2.99**	0.06	—	—	—	—
6	**5.93**	208.6	**5.93**	0.59	**5.93**	0.09	**5.93**	0.08	—	—	—	—
PATROLLING (Vorobeychik and Singh 2012)												
1	**0**	0.01	**0**	0.01	**0**	0.01	**0**	0.01	**0**	0.1	**0**	0.09
2	**0**	0.04	**0**	0.02	**0**	0.02	**0**	0.02	—	—	—	—
3	**0**	3.41	**0**	1.1	**0**	0.10	**0**	0.07	—	—	—	—
6	**0**	66.23	**0**	4.0	**0**	0.14	**0**	0.26	—	—	—	—

Table 1: SSE leader value (V) and runtime (T in seconds). **Bold leader values** indicate the highest return across planning variants for each configuration. "—" indicates timeout.

7 Conclusion

We presented a dynamic programming framework for computing strong Stackelberg equilibria (SSEs) in leader–follower general-sum stochastic games (LF-GSSGs). Our central contribution is a lossless reduction of the original game M, which is multi-objective and non-Markovian, into a credible Markov decision process M' defined over an infinite state space of credible sets. Each credible set captures all follower occupancy states consistent with a fixed leader decision-rule history. This construction preserves optimal leader values, admits recursive decomposition via Bellman's principle, and supports the transfer of point-based value iteration (PBVI) techniques. We formally proved that optimal value functions over credible sets are uniformly continuous, which supports greedy decision-rule extraction through mixed integer linear programming and enables bounded-error approximation via credible-set sampling. Our PBVI algorithms (**H** and **S**) provably yield ε-optimal leader policies with respect to the exact SSE value, and outperform existing planning baselines on five LF-GSSG benchmarks, especially in mixed-motive settings requiring history-aware strategies.

Beyond these contributions, our reduction offers a foundation for extending exact dynamic programming to more expressive settings such as leader–follower general-sum partially observable stochastic games (Chang and III 2021). By explicitly modelling belief updates and credible best responses through occupancy-state transitions, our framework addresses a longstanding open question on the applicability of Bellman-style recursion to asymmetric general-sum stochastic games (Shapley 1953). As such, it paves the way for unified methods that integrate strategic planning, information asymmetry, and sample-efficient learning.

References

An, B.; Kempe, D.; Kiekintveld, C.; Shieh, E.; Singh, S.; Tambe, M.; and Vorobeychik, Y. 2012. Security Games with Limited Surveillance. In Hoffmann, J.; and Selman, B., eds., *AAAI*, 1241–1248.

Basilico, N.; Coniglio, S.; and Gatti, N. 2016. Methods for Finding Leader-Follower Equilibria with Multiple Followers: (Extended Abstract). In Jonker, C. M.; Marsella, S.; Thangarajah, J.; and Tuyls, K., eds., *AAMAS*, 1363–1364. ACM.

Basilico, N.; Nittis, G. D.; and Gatti, N. 2016. A Security Game Combining Patrolling and Alarm-Triggered Responses Under Spatial and Detection Uncertainties. In Schuurmans, D.; and Wellman, M. P., eds., *AAAI*, 404–410. AAAI Press.

Behnke, G.; and Steinmetz, M. 2024. On the Computational Complexity of Stackelberg Planning and Meta-Operator Verification. In Bernardini, S.; and Muise, C., eds., *ICAPS*, 20–24. AAAI.

Bellman, R. E. 1957. *Dynamic Programming*. Dover Publications, Incorporated.

Binmore, K. 1987. Modeling rational players: Part I. *Economics & Philosophy*, 3(2): 179–214.

Bruyère, V.; Fievet, B.; Raskin, J.; and Tamines, C. 2024. Stackelberg-Pareto Synthesis. *ACM*, 25(2): 14:1–14:49.

Chang, Y.; and III, C. C. W. 2021. Worst-case analysis for a leader–follower partially observable stochastic game. *IISE Transactions*, 54(4): 376–389.

Chung, T. H.; Hollinger, G. A.; and Isler, V. 2011. Search and pursuit-evasion in mobile robotics: A survey. *Autonomous Robots*, 31: 299–316.

Conitzer, V.; and Sandholm, T. 2006. Computing the Optimal Strategy to Commit to. In *ACM*, 82–90. ACM.

Denardo, E. W. 1967. Contraction Mappings in the Theory Underlying Dynamic Programming. *SIAM Review*, 9(2): 165–177.

Feinberg, E. A. 2000. Constrained Discounted Markov Decision Processes and Hamiltonian Cycles. *Mathematics of Operations Research*, 25(1): 130–140.

Fujiwara, K. 2006. A Stackelberg Model of Duopoly with a Myopic Follower. *Economics Bulletin*, 4(18): 1–9.

Hansen, E. A.; Bernstein, D. S.; and Zilberstein, S. 2004. Dynamic programming for partially observable stochastic games. In *AAAI*, 709–715.

Horák, K.; and Bošanskỳ, B. 2019. Solving Partially Observable Stochastic Games with Public Observations. In *AAAI*.

Horák, K.; Bosanský, B.; Kovarík, V.; and Kiekintveld, C. 2023. Solving zero-sum one-sided partially observable stochastic games. *Artif. Intell.*, 316: 103838.

Korzhyk, D.; Conitzer, V.; and Parr, R. 2010. Complexity of Computing Optimal Stackelberg Strategies in Security Resource Allocation Games. In Fox, M.; and Poole, D., eds., *AAAI*, 805–810. AAAI.

Kreps, D. M. 2020. *A course in microeconomic theory*. Princeton university press.

Letchford, J.; and Conitzer, V. 2010. Computing Optimal Strategies to Commit to in Extensive-Form Games. In *AAMAS*, 985–992. IFAAMAS.

López, V. B.; Della Vecchia, E.; Jean-Marie, A.; and Ordoñez, F. 2022. Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games. *IEEE Transactions on Automatic Control*.

McKelvey, R. D.; and Palfrey, T. R. 1992. An experimental study of the centipede game. *Econometrica: Journal of the Econometric Society*, 803–836.

Megiddo, N. 1986. *Remarks on bounded rationality*. IBM Corporation.

Nair, R.; Tambe, M.; Yokoo, M.; Pynadath, D. V.; and Marsella, S. 2003. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *IJCAI*, 705–711.

Peralez, J.; Delage, A.; Buffet, O.; and Dibangoye, J. S. 2024. Solving Hierarchical Information-Sharing dec-POMDPs: An Extensive-Form Game Approach. In *ICML*.

Peralez, J.; Delage, A.; Castellini, J.; Cunha, R. F.; and Dibangoye, J. S. 2025. Optimally Solving Simultaneous-Move dec-POMDPs: The Sequential Central Planning Approach. *AAAI*, 38(8): 9411–9419.

Pineau, J.; Gordon, G.; Thrun, S.; et al. 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI*, volume 3, 1025–1032.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons. ISBN 978-0-471-61977-2.

Reny, P. 1988. Rationality, common knowledge and the theory of games. *Princeton: Princeton University*.

Rosenthal, R. W. 1981. Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic theory*, 25(1): 92–100.

Sanjari, S.; Başar, T.; and Yüksel, S. 2023. Isomorphism Properties of Optimality and Equilibrium Solutions under Equivalent Information Structure Transformations: Stochastic Dynamic Games and Teams. *SIAM Journal on Control and Optimization*, 61(5): 3102–3130.

Shapley, L. S. 1953. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100.

Vorobeychik, Y.; An, B.; Tambe, M.; and Singh, S. 2014. Computing solutions in infinite-horizon discounted adversarial patrolling games. In *AAMAS*, 314–322. AAAI Press.

Vorobeychik, Y.; and Singh, S. 2012. Computing Stackelberg equilibria in discounted stochastic games. In *AAAI*, 1478–1484. AAAI Press.

Whitt, W. 1980. Representation and Approximation of Noncooperative Sequential Games. *SIAM Journal on Control and Optimization*, 18(1): 33–48.

Xin Jiang, A.; Yin, Z.; Zhang, C.; Tambe, M.; and Kraus, S. 2013. Game-theoretic Randomization for Security Patrolling with Dynamic Execution Uncertainty. In *AAMAS*.

Yin, Z.; Jiang, A. X.; Tambe, M.; Kiekintveld, C.; Leyton-Brown, K.; Sandholm, T.; and Sullivan, J. P. 2012. TRUSTS:

Scheduling Randomized Patrols for Fare Inspection in Transit Systems Using Game Theory. *AI Magazine*, 33(4): 59.

Zheng, W.; Jung, T.; and Lin, H. 2022. The Stackelberg equilibrium for one-sided zero-sum partially observable stochastic games. *Automatica*, 140: 110231.

Zhong, H.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2023. Can Reinforcement Learning Find Stackelberg-Nash Equilibria in General-Sum Markov Games with Myopically Rational Followers? *Journal of Machine Learning Research*, 24(35): 1–52.

A Table of Notations

Leader–Follower General-Sum Stochastic Game (M)	
Leader and Follower	
Leader's / Follower's action set	A_L / A_F
Leader's / Follower's action	a_L / a_F
Joint action	$a = (a_L, a_F)$
Leader's / Follower's reward function	r_L / r_F
Leader's / Follower's policy	π_L / π_F
Joint policy	$\pi = (\pi_L, \pi_F)$
Leader's / Follower's decision rule at stage t	$\delta_{L,t}$, $\delta_{F,t}$
Joint decision rule	$\delta = (\delta_L, \delta_F)$
Game Structure	
LF-GSSG model	M
Number of stages (horizon)	ℓ
Stage index	$t \in \{0, 1, \ldots, \ell\}$
State space	S
Environment state	$s \in S$
Transition function	$p(s' \mid s, a_L, a_F)$
Discount factor	$\gamma \in [0, 1)$
Set of best responses to π_L	$\mathtt{BR}(\pi_L)$
Credible MDP (M')	
Leader's / Follower's private environment history	$h_{L,t}$, $h_{F,t}$
Joint private history	$h = (h_L, h_F)$
Decision-rule history up to stage t	$\theta_{L,t}$, $\theta_{F,t}$
Joint decision-rule history	$\theta = (\theta_L, \theta_F)$
Occupancy state (distribution over joint private histories)	$o = o_\theta$
Credible set	$c_{\theta_L} = \{o_{(\theta_L, \theta_F)}\}$
Filtered credible set	$\tilde{c} = \mathbb{F}(c_{\theta_L})$
Set of filtered credible sets	$\tilde{C}$
Initial credible set	c_0
Filtered transition function	$\tilde{T}(\tilde{c}, \delta_L)$
Deterministic transition on occupancy states	$\tau(o, \delta_L, \delta_F)$
Expected cumulative reward (leader / follower)	$\rho_L(o)$, $\rho_F(o)$
Marginal belief (over states)	$b_o \in \Delta(S)$
Reward function on $\tilde{C}$	$R(\tilde{c})$
Value Functions	
Leader / Follower value function at stage t under policy π	$v^\pi_{L,t}$, $v^\pi_{F,t}$
Leader / Follower q-value at stage t	$q^\pi_{L,t}$, $q^\pi_{F,t}$
Planning Objects	
Set of α-vector pairs	$\Gamma = \{(\alpha_L^{(k)}, \alpha_F^{(k)})\}_{k=1}^K$
Collection of such sets	$\Lambda = \{\Gamma^{(j)}\}_{j=1}^J$

Table 2: Summary of notations used throughout the paper. All terms are formally defined in the main text.

B Naive Dynamic Programming: Why It Fails

A naive dynamic programming approach may fail to compute strong Stackelberg equilibria (SSEs), even when applying seemingly rational local improvement steps. The following proposition illustrates this failure mode.

Proposition B.1. *Let π_L be a Markov leader policy. Construct a new policy π'_L by selecting, at each state $s \in S$ and stage $t < \ell$,*

a decision rule $\delta'_{L,t}(s)$ solving:

$$\max_{\delta_L \in \Delta(A_L)} \max_{a_F \in A_F} \mathbb{E}_{a_L \sim \delta_L}\left[q^{\pi_L}_{L,t}(s,s,s,a_L,a_F)\right]$$
$$\text{s.t.} \quad \mathbb{E}_{a_L \sim \delta_L}\left[q^{\pi_L}_{F,t}(s,s,s,a_L,a_F)\right] = v^{\pi_L}_{F,t}(s,s,s).$$

Even if the solution satisfies $\pi'_L = \pi_L$, the resulting policy π'_L may fail to induce a strong Stackelberg equilibrium.

Proof. We construct a counterexample based on the finite-horizon centipede game (Figure 1). Consider a Markov leader policy π_L that selects action `take` in both states s_1 and s_3, inducing a follower best response π_F in which the follower plays `take` in both s_2 and s_4. The resulting joint policy $\pi = (\pi_L, \pi_F)$ yields the following value vectors:

$$v^{\pi}_0(s_1) = (1,0),$$
$$v^{\pi}_1(s_2) = (0,2),$$
$$v^{\pi}_2(s_3) = (3,1),$$
$$v^{\pi}_3(s_4) = (2,4).$$

Suppose we now apply the local improvement rule defined in the proposition to π_L. At both s_1 and s_3, the decision rule $\delta_{L,t}(s)$ corresponding to `take` already satisfies the local optimisation objective, and hence the improvement step returns the same policy: $\pi'_L = \pi_L$.

However, consider an alternative leader policy π''_L that plays `continue` in both s_1 and s_3. The follower best response under π''_L is to play `continue` in s_2 and `take` in s_4, leading to the joint policy $\pi'' = (\pi''_L, \pi''_F)$ with:

$$v^{\pi''}_0(s_1) = (2,4).$$

This dominates the outcome under π'_L, which yielded only $(1,0)$, thereby contradicting the assumption that π'_L induces an SSE.

Hence, the local improvement rule may yield a fixed point that fails to satisfy the global optimality required by the strong Stackelberg equilibrium. □

C Credible Markov Decision Processes

Lossless Reduction

Lemma C.1. *For every Markov leader policy $\pi_L \in \Pi_L$ in LF-GSSG M, let c_{π_L} be the credible set induced by π_L starting from the initial state s_0. Then, $R(c_{\pi_L}) = R\big(\mathbb{F}(c_{\pi_L})\big)$, where $\mathbb{F}(\cdot)$ denotes the filter operator.*

Proof. Let c be any credible set. Define $B_c \doteq \{b_o \mid o \in c\}$ and, for each $b \in B_c$, define the belief class

$$c_b \doteq \{o \in c \mid b_o = b\}.$$

Within each c_b, for any occupancy state $o \in c_b$, define

$$c^o_b \doteq \{o' \in c_b \mid \forall i \in I, \rho_i(o') \geq \rho_i(o),\ \exists j \in I, \rho_j(o') > \rho_j(o)\}.$$

We then define the filtered set as

$$\mathbb{F}(c) \doteq \cup_{b \in B_c} \texttt{Unique}(\{o \in c_b \mid c^o_b = \emptyset\}),$$

where $\texttt{Unique}(\cdot)$ retains exactly one representative per unique cumulative value-so-far vector (ρ_L, ρ_F). We proceed by contradiction. Assume the statement does not hold, that is,

$$R(c_{\pi_L}) \neq R\big(\mathbb{F}(c_{\pi_L})\big),$$

which implies that there exists an occupancy state $o \in c_{\pi_L}$ whose removal by the filter operator $\mathbb{F}$ changes the value:

$$R(c_{\pi_L}) \neq R\big(c_{\pi_L} \setminus \{o\}\big).$$

Let b be the marginal belief state associated with o, and let $c_b \subseteq c_{\pi_L}$ be the subset of occupancy states in c_{π_L} sharing the marginal belief b. For the removal of o to affect the value, it must satisfy:

$$\rho_F(o) > \max_{o' \in \mathbb{F}(c_{\pi_L})} \rho_F(o') \quad \text{or} \quad \rho_F(o) = \max_{o' \in \mathbb{F}(c_{\pi_L})} \rho_F(o') \wedge \rho_L(o) > R\big(\mathbb{F}(c_{\pi_L})\big).$$

where $\rho_L(o)$ and $\rho_F(o)$ denote the expected cumulative rewards for the leader and follower, respectively, under occupancy state o. Since o was removed by $\mathbb{F}$, there exists $o' \in c^o_b$ such that

$$\forall i \in I, \rho_i(o') \geq \rho_i(o),\ \exists j \in I, \rho_j(o') > \rho_j(o).$$

Consequently, we have

$$R\big(\mathbb{F}(c_{\pi_L})\big) > \rho_L(o),$$

which contradicts the earlier assumption that

$$\rho_F(o) > \max_{o' \in \mathbb{F}(c_{\pi_L})} \rho_F(o') \quad \text{or} \quad \rho_F(o) = \max_{o' \in \mathbb{F}(c_{\pi_L})} \rho_F(o') \wedge \rho_L(o) > R\big(\mathbb{F}(c_{\pi_L})\big).$$

Therefore, the assumption is false, and we conclude that

$$R(c_{\pi_L}) = R\big(\mathbb{F}(c_{\pi_L})\big).$$

□

Theorem C.2. *The credible MDP M' associated with the LF-GSSG M constitutes a lossless reduction in the sense of Definition 3.*

Proof. We verify each of the three criteria in Definition 3.

(i) Value preservation. Let $\pi_L = (\delta_{L,0}, \ldots, \delta_{L,\ell-1}) \in \Pi_L$ be a leader policy in M. We define the corresponding policy $\pi'_L \in \Pi'_L$ in M' as the sequence of decision rules $\pi'_L \doteq (\delta_{L,0}, \ldots, \delta_{L,\ell-1})$, applied deterministically at each stage. Since the transitions in M' are deterministic and coincide with the application of decision rules over credible sets, this correspondence preserves execution paths.

The value in M' starting from initial credible set $c_0 = \{s_0\}$ is

$$v^{\pi'_L}_{M',L,0}(c_0) = R\big(\mathbb{F}(c_{\pi'_L})\big) = R(c_{\pi'_L}),$$

where the second equality follows from Lemma C.1.

By definition of the terminal reward function, we then have:

$$v^{\pi'_L}_{M',L,0}(c_0) = \max \left\{ \rho_L(o) \;\middle|\; o \in \operatorname*{argmax}_{o' \in c_{\pi'_L}} \rho_F(o') \right\}.$$

Each $o \in c_{\pi'_L}$ corresponds to an occupancy state induced by some follower response $\pi_F \in \mathtt{BR}(\pi_L)$. Since $\rho_i(o) = v^{\pi_L,\pi_F}_{M,i,0}(s_0)$ by construction of occupancy states, this implies

$$v^{\pi'_L}_{M',L,0}(c_0) = v^{\pi_L}_{M,L,0}(s_0).$$

Conversely, given a policy $\pi'_L \in \Pi'_L$, the sequence of decision rules defines a leader policy $\pi_L \in \Pi_L$ in M that induces the same sequence of occupancy sets and the same terminal credible set $c_{\pi_L} = c_{\pi'_L}$. Hence,

$$v^{\pi_L}_{M,L,0}(s_0) = v^{\pi'_L}_{M',L,0}(c_0).$$

This establishes value preservation in both directions.

(ii) Equilibrium correspondence. Let $\pi = (\pi_L, \pi_F)$ be a strong Stackelberg equilibrium (SSE) in M. Then $\pi_F \in \mathtt{BR}(\pi_L)$, and all best responses induce occupancy states $o \in c_{\pi_L}$. As above, the corresponding leader policy π'_L in M' induces the same credible set $c_{\pi'_L} = c_{\pi_L}$, and the follower best response in M' coincides with selecting the occupancy state $o \in c_{\pi_L}$ maximising ρ_F. Hence, the value of the SSE is preserved, and the mapping $\pi \mapsto \pi'_L$ defines a surjection from SSEs in M to optimal leader policies in M'. Multiple SSEs in M can map to the same policy in M' if they induce the same decision-rule sequence.

(iii) Information compatibility. The leader policies in both M and M' are defined over the same private state–action histories, and decision rules are chosen with respect to credible sets indexed by these histories. The surrogate model M' does not introduce any additional observability or decision-making capability. All transitions and outcomes are grounded in the environment and policy structure of the original model M.

Therefore, M' satisfies all three criteria of Definition 3 and is a lossless reduction of M. □

D Theoretical Results for Credible MDPs

Complexity Results

We consider the following problem. INFINITE-HORIZON EXACT VALUE: Given a MDP M, and a rational number v, decide whether there is a memoryless deterministic policy $\pi : S \to A$ whose value V^π (from the initial state) is exactly v. We also define FINITE HORIZON EXACT VALUE which further accepts a horizon ℓ.

Theorem D.1. *The* FINITE HORIZON EXACT VALUE *is NP-complete.*

Proof. By reduction from Hamiltonian path. Given directed graph $G = (V, E)$, we build an MDP M whose states are the vertices of G, and whose transitions are the edges of G. The initial state s_0 of M is chosen to be an arbitrary vertex of G. Define $B = |V| + 1$. We associate an index $0 \leq i \leq |V| - 1$ to each state of M. The reward when leaving the state with index i is B^{i-1}. The discount factor is $\gamma = 1$. We set $v = 1 + B + ...B^{|V|-1}$.

The graph G admits a Hamiltonian path iff $(M, v, \ell = |V|)$ is a positive instance of the finite horizon exact value problem.

$\boxed{\Rightarrow}$ If there is a Hamiltonian path, consider policy π that follows that path. Its value is $v = 1 + B + ...B^{|V|-1}$ because we reach exactly each state once.

$\boxed{\Leftarrow}$ Consider a policy π whose value is $v = 1 + B + ...B^{|V|-1}$. Let x_i be the number of times the state with index i is visited. The value can also be written $v = x_1 + x_2 B + \times x_{|V|} B^{|V|-1}$. As $x_i < B$, the two writings of number v in base B are unique, thus $x_i = 1$ for all i. So π follows a Hamiltonian path. □

The infinite-horizon case is also NP-hard but for this proof, we restrict to memoryless deterministic policies.

Theorem D.2. INFINITE HORIZON EXACT VALUE *is NP-complete.*

Proof. The membership is proven as follows. Guess π. Compute the exact value in the induced Markov chain in polynomial time. Check that it is equal to v.

We now focus on the NP-hardness, proven by reduction from Hamiltonian cycle. Let $G = (V, E)$ be a directed graph. We construct an MDP M as follows. The state set is V, and the transitions in M are given by the edges of G. Fix any vertex s_0 in V to be the initial state of M. The reward is 1 when going out s_0 and 0 everywhere else. We fix the discount factor $\gamma = 1/2$. Let $v = \frac{1}{1-\gamma^{|V|}}$.

The instance (M, v) is computable in poly-time from G. It remains to prove that G has a Hamiltonian cycle iff (M, v) is a positive instance.

$\boxed{\Rightarrow}$ Suppose G has a Hamiltonian cycle. Consider π to be the strategy that goes through the cycle forever. Every $|V|$ steps, we reach s_0 and the reward is 1. The gain is

$$V^{\pi}(s_0) = 1 + \gamma^{|V|} + \gamma^{2|V|} + ... = v.$$

So M has a strategy whose value is exactly v. So (M, v) is a positive instance.

$\boxed{\Leftarrow}$ Suppose that (M, v) is a positive instance, that is, M has a strategy π whose value is exactly v. Because M is deterministic, a single infinite path in G is generated under π.

Suppose that under π, s_0 is never reached again after the first step. Then $V^{\pi}(s_0) = 1$, which contradicts $V^{\pi}(s_0) = v$. So s_0 must be reached again. Because π is deterministic and memoryless, it induces a simple cycle that contains s_0, which is repeated indefinitely. Let n be the number of states before reaching s_0 again, that is, the length of the simple cycle. We have

$$V^{\pi}(s_0) = 1 + \gamma^n + \gamma^{2n} + ... = \frac{1}{1 - \gamma^n}.$$

As $V^{\pi}(s_0) = v$, we have $n = |V|$. Thus π induces a Hamiltonian cycle. □

A *constrained MDP (CMDP)* (we consider the single-cost version of CMDP given in (Feinberg 2000)) is a tuple

$$(S, A, p, r, c, \nu, \gamma, s_0)$$

where:

- S: Finite state space, $s_0 \in S$ initial state
- A: Finite action space
- $p(s' \mid s, a)$: Transition probabilities
- $r(s, a) \in \mathbb{Q}$: Reward function
- $c(s, a) \in \mathbb{Q}_{\geq 0}$: Cost function
- $\gamma \in (0, 1)$: Discount factor
- $\nu \in \mathbb{Q}$: Cost threshold

The *infinite-horizon CMDP feasibility problem* is the the decision problem that, given a CMDP, and rational R, asks whether there exists a policy π satisfying:

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \geq R, \tag{1}$$

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \leq \nu. \tag{2}$$

In words, we ask for the existence of a policy π whose expected gain is greater than or equal to R, while expected cost is bounded by ν.

We similarly define the *finite-horizon CMDP feasibility problem* by considering a horizon ℓ. In the sequel, we ask for the existence of a *deterministic* policy.

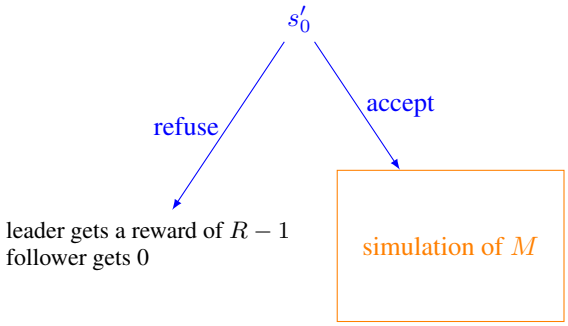

Figure 3: The LF-GSSG M' constructed from the CMDP feasibility instance (M, R, ν) works as follows. In the initial state s'_0 the follower starts by either refusing or accepting. If the follower refuses, the leader gets a reward of $R-1$ while the follower gets 0. If the follower accepts, a simulation of M starts where the leader plays. The leader receives the reward in M while the reward of F handles the constraint.

Theorem D.3. *The infinite-horizon and finite horizon CMDP feasability problems are NP-hard.*

Proof. Reduction from INFINITE HORIZON EXACT VALUE POLICY of Theorem D.2. Consider an instance (M, v) of INFINITE HORIZON EXACT VALUE POLICY. The CMDP M' is M, and the costs are identical to rewards: $c(s,a) = r(s,a)$.

Then, there is a policy in M whose value is v iff there is a policy in M' whose value is $\geq v$ and whose cost is $\leq v$.

For the finite horizon case, we similarly reduce from the FINITE HORIZON EXACT VALUE POLICY using Theorem D.1. □

The *"two-bound" SSE decision problem* is the following: Given rationals R, ν', and LF-GSSG M, determine if there is a joint policy $\pi = (\pi_L, \pi_F)$ such that $v^\pi_L(s_0) \geq R$ and $v^\pi_F(s_0) \geq \nu'$. The *memoryless deterministic* SSE decision problem is the restriction of this problem to leader policies that are memoryless and deterministic.

Theorem D.4. *The memoryless deterministic "two-bound" SSE decision problem is NP-hard both in the finite horizon and infinite horizon cases.*

Proof. We prove NP-hardness via a polynomial-time reduction from the CMDP feasibility problems (Theorem D.3)

From a CMDP feasibility instance M, and bounds R, ν, we construct an infinite-horizon LF-GSSG M' with bounds $R, \nu' = -\nu$ on the same state space as M. The leader's actions are identical to actions available in M, and the follower has a single dummy action which is ignored in all transitions. Furthermore, we have the following rewards: $r_L(s,a) = r(s,a)$, and $r_F(s,a) = -c(s,a)$ for all s, a.

Assume that (1)-(2) hold for policy π. Let $\pi_L = \pi$, and π_F be an arbitrary policy. Then $v^\pi_L(s_0) \geq R$ since the value for the leader is identical to the value in M, and $v^\pi_F(s_0) \geq \nu'$ since the follower's value is minus the expected cost in M.

Conversely, if $v^\pi_L(s_0) \geq R$ and $v^\pi_F(s_0) \geq \nu'$, then by letting $\pi = \pi_L$, we satisfy (1)-(2).

The proof for the finite horizon case is analogous using Theorem D.3. □

The *infinite-horizon SSE decision problem* is the following: Given rational R, and LF-GSSG M, determine if there is a joint policy $\pi = (\pi_L, \pi_F)$ such that $\pi_F \in \mathtt{BR}(\pi_L)$ and $v^\pi_L(s_0) \geq R$. We also define the *finite-horizon SSE decision problem* which further takes the horizon ℓ as input.

Theorem D.5. *The deterministic memoryless SSE decision problem is NP-hard both in the finite-horizon and infinite-horizon cases.*

Proof. We prove NP-hardness via a polynomial-time reduction from the finite-horizon CMDP feasibility problem (Theorem D.3).

Let us start with the finite horizon case. From a finite-horizon deterministic CMDP instance (M, R, ν, ℓ) we construct the following finite-horizon deterministic SSE-instance (M', R, ℓ). The LF-GSSG M' depicted in Figure 3 is constructed as follows:

- We introduce a fresh initial state s'_0 (distinct from the states of M);
- In s'_0, the follower decides either to "accept" to simulate a play in M, or to "refuse".
- If the follower refuses, the leader obtains a reward of $R-1$ and the follower a reward of 0 and the LF-GSSG stops.
- If the follower accepts, we enter the initial state of a copy of M, in which only the leader determines the actions to be taken (actions played by the follower are not relevant), and where the reward of the leader mimics the reward in M, while the reward of F takes the constraint into account as follows. At state s, when the leader plays a, the follower receives the reward:

$$r_F(s,a) := \frac{\nu}{\ell} - c(s,a)$$

where $c(s,a)$ is the cost of executing a from state s in M.

The parameter R and ℓ are the same than in (M, R, ν, ℓ). The construction of (M', R, ℓ) from (M, R, ν, ℓ) can be performed in poly-time.

It remains to prove that there is a policy π in the CMDP M such that the gain is $\geq R$ and the cost is $\leq \nu$ iff there is π' a SSE in M' with the leader-value $\geq R$.

$\boxed{\Rightarrow}$ Suppose there is such a π. We set π' so that π_L mimics π in the M-part of M', and π_F accepts in s_0. The leader-value for π' is $\geq R$. Meanwhile, as the cost is $\leq \nu$ in M with π, the follower value is ≥ 0. So π_F is a best response, and π' is indeed a SSE.

$\boxed{\Leftarrow}$ Suppose there is a SSE π' in M' such that the leader-value is $\geq R$ with π'. Let π the policy that mimics the leader-policy π_L in the M-part. As the leader-value is $\geq R$, it means that the follower has accepted in s_0. Thus, as the follower plays a best response, the follower-value is as good as the "refuse" choice, *i.e.* greater than or equal to 0. It follows that the total cost under π in M is less than or equal to ν.

For the infinite-horizon case, we consider the same reduction with the following differences. In the initial state s'_0, if the follower refuses, the players receive $R - 1$ and 0 respectively as before, and we extend the game with a self-loop with reward 0 for both players. In the copy of M, given state-action pair (s, a), the leader receives the reward $r_L(s, a) = r(s, a)/\gamma$, while the follower receives $r_F(s, a) = \nu(1 - \gamma)/\gamma - c(s, a)/\gamma$. Note that we divide the rewards and costs by γ to account for the additional step we introduced due to s'_0.

Consider a joint policy π. If the follower refuses, then the leader receives $R - 1$ and the follower 0. If the follower accepts, then the leader's policy is played in the copy of M, and the leader receives

$$\mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)/\gamma\right] = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right].$$

The follower receives

$$\begin{aligned}
&\mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty} \gamma^t (\nu(1 - \gamma)/\gamma - c(s_t, a_t)/\gamma)\right] \\
&= \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t (\nu(1 - \gamma) - c(s_t, a_t))\right] \\
&= \nu - \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right].
\end{aligned}$$

The correctness of the reduction is then identical to the proof in the finite-horizon case. □

Proof of Losslessness of the Filtering Operator

Theorem D.6 (Losslessness of the Filtering Operator)**.** *Let M be a finite-horizon leader–follower general-sum stochastic game (LF-GSSG) with horizon ℓ, and let $\mathbb{F}$ denote the belief-based filtering operator. Fix any Markov leader policy $\pi_L \in \Pi_L$, and let $(c_0, \ldots, c_\ell)$ denote the sequence of credible sets generated by forward simulation under π_L. Define the filtered sequence $(\tilde{c}_0, \ldots, \tilde{c}_\ell)$ recursively by:*

- $\tilde{c}_0 \doteq \mathbb{F}(c_0)$,
- $\tilde{c}_{t+1} \doteq \mathbb{F}(T(\tilde{c}_t, \delta_{L,t}))$ *for all* $t < \ell$,

where $\delta_{L,t}$ is the decision rule applied by π_L at stage t, and T denotes the transition function of the credible MDP. Then, the final cumulative leader reward is invariant under recursive filtering:

$$R(c_\ell) = R(\tilde{c}_\ell).$$

Proof. We proceed by induction on $t \in \{0, \ldots, \ell\}$, maintaining the following invariant.

> **Filtered Reachability Preservation (FRP).** At every stage t, the filtered set $\tilde{c}_t$ contains, for each marginal belief b, at least one occupancy state per value-so-far vector (ρ_L, ρ_F) that is reachable from c_0 under π_L, and that may lead to a final occupancy state attaining the optimal cumulative reward $R(c_\ell)$.

Base case ($t = 0$). The initial credible set c_0 is a singleton, and the filtering operator does not remove it: $\tilde{c}_0 = \mathbb{F}(c_0) = c_0$. The invariant holds trivially.

Inductive step. Assume the invariant holds at stage t. We must show it holds at $t + 1$. Let $o \in \tilde{c}_t$ be any occupancy state preserved by filtering. By the inductive hypothesis, every marginal belief and value-so-far vector relevant for optimal return is represented in $\tilde{c}_t$. Since the leader policy is Markov and the environment is fully observable and Markovian, the successor

occupancy states resulting from applying $T(o, \delta_{L,t})$ depend only on the marginal belief b and the joint action, not on local histories.

Moreover, the transition function T deterministically propagates marginal beliefs and extends cumulative rewards additively. Thus, the set $c_{t+1} \doteq T(\tilde{c}_t, \delta_{L,t})$ includes all marginal belief and cumulative reward pairs relevant for reaching optimal terminal values. Applying $\mathbb{F}$ to c_{t+1} preserves one representative per such pair, hence the invariant holds at $t+1$.

Terminal step. At stage $t = \ell$, the credible MDP accumulates the total leader reward. By the inductive invariant, the final filtered set $\tilde{c}_\ell$ contains all marginal belief and reward pairs relevant for achieving $R(c_\ell)$. By Lemma C.1, which guarantees that filtering does not reduce the maximal cumulative reward,

$$R(\tilde{c}_\ell) = R(c_\ell).$$

Conclusion. Filtering preserves all value-relevant paths under Markov leader policies. Thus, filtering may be applied at each stage without loss of optimality. □

Bellman's Optimality Equations

Theorem D.7. *The optimal leader state-value function $v^*_{L,t} \colon \tilde{C} \to \mathbb{R}$ at stage $t < \ell$ satisfies:*

$$v^*_{L,t} \colon \tilde{c} \mapsto \max_{\delta_L \in \Delta_L} v^*_{L,t+1}\big(\tilde{T}(\tilde{c}, \delta_L)\big).$$

Proof. The proof proceeds from backward induction for finite-horizon discrete-time Markov decision process (Puterman 1994). We begin with the definition: for any transient credible set $\tilde{c}$ at stage $t < \ell$,

$$v^*_{L,t}(\tilde{c}) = \max_{\pi_L \in \Pi_L} v^{\pi_L}_{L,t}(\tilde{c}).$$

By decomposing leader policy $\pi_L \doteq (\delta_{L,t}, \ldots, \delta_{L,\ell-1})$, we obtain the following expression

$$v^*_{L,t}(\tilde{c}) = \max_{\delta_{L,t} \in \Delta_{L,t}} \cdots \max_{\delta_{L,\ell-1} \in \Delta_{L,\ell-1}} v^{\pi_L}_{L,t}(\tilde{c}).$$

Since the policy tail $\pi'_L \doteq (\delta_{L,t+1}, \ldots, \delta_{L,\ell-1})$ applies to the next credible set $\tilde{T}(\tilde{c}, \delta_{L,t})$ reached by executing decision rule $\delta_{L,t}$ at $\tilde{c}$, we obtain:

$$v^*_{L,t}(\tilde{c}) = \max_{\delta_{L,t} \in \Delta_{L,t}} \max_{\pi'_L \in \Delta_{L,t+1:\ell-1}} v^{\pi'_L}_{L,t+1}(\tilde{T}(\tilde{c}, \delta_{L,t})).$$

Injecting the definition of the optimal value function yields:

$$\begin{aligned} v^*_{L,t}(\tilde{c}) &= \max_{\delta_{L,t} \in \Delta_{L,t}} v^*_{L,t+1}(\tilde{T}(\tilde{c}, \delta_{L,t})) \\ &= \max_{\delta_{L,t} \in \Delta_{L,t}} q^*_{L,t}(\tilde{c}, \delta_{L,t}). \end{aligned}$$

This concludes the proof. □

Uniform Continuity Properties

We establish the uniform continuity of the leader's value function over finite credible sets. For clarity, we decompose the proof into three lemmas leading to the main proposition. Before proceeding any further we start with the definition of uniform continuity and the Hausdorff distance across credible sets useful for establishing the uniform continuity properties.

Definition 6 (Uniform Continuity)**.** *A function $v \colon \tilde{C} \to \mathbb{R}$ is uniformly continuous on $\tilde{C}$ if:*

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ such that } \forall \tilde{c}, \tilde{c}' \in \tilde{C}, |\tilde{c} - \tilde{c}'| < \delta \implies |v(\tilde{c}) - v(\tilde{c}')| < \varepsilon.$$

Next, we introduce a distance between two credible sets, $\tilde{c}$ and $\tilde{c}'$, drawn from the occupancy-state space O, using the Hausdorff distance with the ℓ_1-norm. The Hausdorff distance between two credible sets measures how far apart they are by capturing the worst-case minimal ℓ_1-distance needed to match each occupancy state in one set to some occupancy state in the other. If every occupancy state in both credible sets is close to at least one in the other, the sets are considered behaviorally close.

Definition 7 (Hausdorff Distance)**.** *Let $\tilde{c}, \tilde{c}' \subset O$ be finite credible sets, and equip the space of finite subsets of O with the Hausdorff distance d_H induced by the ℓ_1-norm:*

$$d_H(\tilde{c}, \tilde{c}') \doteq \max\left\{\sup_{o \in \tilde{c}} \inf_{o' \in \tilde{c}'} \|o - o'\|_1,\ \sup_{o' \in \tilde{c}'} \inf_{o \in \tilde{c}} \|o' - o\|_1\right\}.$$

[2]The filtering operator $\mathbb{F}$ retains exactly one representative per cumulative reward vector within each marginal belief class. Since transitions and rewards in the credible MDP depend only on these quantities under Markov policies, the tie-breaking rule used to select representatives has no impact on value.

Lemma D.8 (Stability of filtered maxima under Hausdorff perturbations)**.** *Let $v\colon O \to \mathbb{R}$ be uniformly continuous, and let $x, \tilde{c}' \subset O$ be credible sets. Then for every $\varepsilon > 0$, there exists $\delta > 0$ such that if $d_H(\tilde{c}, \tilde{c}') < \delta$, then*

$$\left| \max_{o \in x} v(o) - \max_{\bar{o} \in \tilde{c}'} v(\bar{o}) \right| < \varepsilon.$$

Proof. Fix $\varepsilon > 0$. By uniform continuity of $v\colon O \to \mathbb{R}$, there exists $\delta > 0$ such that:

$$\|o - \bar{o}\|_1 < \delta \quad \Rightarrow \quad |v(o) - v(\bar{o})| < \varepsilon.$$

Assume $d_H(\tilde{c}, \tilde{c}') < \delta$. Then:

1. For every $o \in \tilde{c}$, there exists $\bar{o} \in \tilde{c}'$ with $\|o - \bar{o}\|_1 < \delta$, hence $|v(o) - v(\bar{o})| < \varepsilon$. Taking maxima over $o \in \tilde{c}$, we obtain:

$$\max_{o \in \tilde{c}} v(o) < \max_{\bar{o} \in \tilde{c}'} v(\bar{o}) + \varepsilon.$$

2. Similarly, for every $\bar{o} \in \tilde{c}'$, there exists $o \in \tilde{c}$ with $\|\bar{o} - o\|_1 < \delta$, hence $|v(\bar{o}) - v(o)| < \varepsilon$, and thus:

$$\max_{\bar{o} \in \tilde{c}'} v(\bar{o}) < \max_{o \in \tilde{c}} v(o) + \varepsilon.$$

Combining both bounds: $|\max_{o \in \tilde{c}} v(o) - \max_{\bar{o} \in \tilde{c}'} v(\bar{o})| < \varepsilon$. □

Lemma D.9 (Leader value function uniform continuity)**.** *Let K and N be finite index sets, and let O be a compact metric space with distance $\|\cdot\|$. For each $k \in K$ and $n \in N$, let $v_{kn}, u_{kn}\colon O \to \mathbb{R}$ be uniformly continuous functions. For any credible set $\tilde{c} \subset O$, define the filtered set*

$$\tilde{c}^*_{kn} \doteq \left\{ o \in \tilde{c} \mid u_{kn}(o) = \max_{o' \in \tilde{c}} \max_{n' \in N} u_{kn'}(o') \right\},$$

and define the composite value function

$$v^*(\tilde{c}) \doteq \max_{k \in K} \max_{n \in N} \max_{o \in \tilde{c}^*_{kn}} v_{kn}(o).$$

Then v^* is uniformly continuous over credible sets equipped with the Hausdorff distance d_H induced by $\|\cdot\|$.

Proof. Let $\varepsilon > 0$ be arbitrary.

By uniform continuity of each u_{kn} and by Lemma D.8, there exists $\delta^u_{kn} > 0$ such that if $d_H(\tilde{c}, \tilde{c}') < \delta^u_{kn}$, then the filtered sets satisfy:

$$d_H(\tilde{c}^*_{kn}, \tilde{c}'^*_{kn}) < \delta^v_{kn},$$

where δ^v_{kn} will be chosen next.

By uniform continuity of each v_{kn} and again by Lemma D.8, there exists $\delta^v_{kn} > 0$ such that if $d_H(\tilde{c}^*_{kn}, \tilde{c}'^*_{kn}) < \delta^v_{kn}$, then:

$$\left| \max_{o \in \tilde{c}^*_{kn}} v_{kn}(o) - \max_{o' \in \tilde{c}'^*_{kn}} v_{kn}(o') \right| < \varepsilon.$$

Define:

$$\delta \doteq \min_{k,n} \{\delta^u_{kn}, \delta^v_{kn}\}.$$

Then, under $d_H(\tilde{c}, \tilde{c}') < \delta$, each index pair satisfies:

$$\left| \max_{o \in \tilde{c}^*_{kn}} v_{kn}(o) - \max_{o' \in \tilde{c}'^*_{kn}} v_{kn}(o') \right| < \varepsilon,$$

and the global maximum over finitely many (k, n) stays controlled:

$$|v^*(\tilde{c}) - v^*(\tilde{c}')| < \varepsilon.$$

Therefore, v^* is uniformly continuous. □

In demonstrating the uniform continuity of $v^*_L\colon \tilde{C} \to \mathbb{R}$ across credible sets, it will prove useful to show the underlying structure of value function v^*_L across credible sets.

Lemma D.10. *The optimal leader state-value function $v^*_{L,t}\colon \tilde{C} \to \mathbb{R}$ at stage t is such that: there exists a collection Λ_t of finite sets Γ_t of linear function pairs (α_L, α_F), each defined over occupancy states, such that:*

$$v^*_{L,t}(\tilde{c}) = \max_{\Gamma \in \Lambda_t} \left\{ \rho_L(o) + \gamma^t \sum_{h_F} \Pr(h_F \mid o) \cdot \alpha_L^{o_{h_F}}(o_{h_F}) \mid (o, \alpha^{\cdot}) \in \operatorname*{argmax}_{o' \in \tilde{c}, \alpha' \in \Gamma^{|\mathbb{C}_F(\tilde{c})|}} \left(\rho_F(o') + \gamma^t \sum_{h_F} \Pr(h_F \mid o) \cdot \alpha_F'^{o_{h_F}}(o'_{h_F}) \right) \right\}.$$

Proof. The proof follows directly from the definition of the optimal leader state-value function at stage t: for any arbitrary credible set $\tilde{c} \in \tilde{C}$, we have that:

$$v^*_{L,t}(\tilde{c}) \doteq \max_{\pi_L \in \Pi_L} v^{\pi_L}_{L,t}(\tilde{c}). \tag{3}$$

Let $\tilde{c}_{\pi_L} \doteq \tilde{T}(\tilde{T}(\dots(\tilde{T}(\tilde{c}, \delta_{L,t}), \delta_{L,t+1})\dots), \delta_{L,\ell-1})$ denote a terminal credible set induced by $\tilde{c}$ and leader policy $\pi_L \doteq (\delta_{L,t}, \delta_{L,t+1}, \dots, \delta_{L,\ell-1})$ from stage t onward. Then, it follows from (3) that:

$$v^*_{L,t}(\tilde{c}) = \max_{\pi_L \in \Pi_L} R(\tilde{c}_{\pi_L}). \tag{4}$$

Let o_{π_L,π_F} denote a terminal occupancy state induced by occupancy state $o \in \tilde{c}$ and leader policy $\pi_L \doteq (\delta_{L,t}, \delta_{L,t+1}, \dots, \delta_{L,\ell-1})$ and follower policy $\pi_F \doteq (\delta_{F,t}, \delta_{F,t+1}, \dots, \delta_{F,\ell-1})$ from stage t onward. The application of the definition of the reward in a credible Markov decision process into (4) leads us to the following expression:

$$v^*_{L,t}(\tilde{c}) = \max_{\pi_L \in \Pi_L} \max_{o_{\pi_L,\pi_F} \in \tilde{c}_{\pi_L}} \left\{ \rho_L(o_{\pi_L,\pi_F}) \mid \rho_F(o_{\pi_L,\pi_F}) = \max_{o'_{\pi_L,\pi'_F} \in \tilde{c}_{\pi_L}} \rho_F(o'_{\pi_L,\pi'_F}) \right\}. \tag{5}$$

By the application of the definition of the state-value function $v^{\pi_L,\pi_F}_{i,t} : S \times H_L \times H_F \to \mathbb{R}$ under leader–follower policies (π_L, π_F) from stage t onward into (5) we obtain:

$$v^*_{L,t}(\tilde{c}) = \max_{\pi_L \in \Pi_L} \max_{o \in \tilde{c}} \left\{ \rho_L(o) + \gamma^t v^{\pi_L,\pi_F}_{L,t}(o) \mid \rho_F(o) + \gamma^t v^{\pi_L,\pi_F}_{F,t}(o) = \max_{o' \in \tilde{c}} \left(\rho_F(o') + \gamma^t v^{\pi_L,\pi'_F}_{F,t}(o') \right) \right\}. \tag{6}$$

If we let

$$\Gamma_{\pi_L,t:\ell-1} \doteq \{(v^{\pi_L,t:\ell-1,\pi_{F,t:\ell-1}(h_{F,t})}_{L,t}, v^{\pi_L,t:\ell-1,\pi_{F,t:\ell-1}(h_{F,t})}_{F,t}) \mid \pi_{F,t:\ell-1} \in \Pi_{F,t:\ell-1}, h_{F,t} \in H_{F,t}\}$$

where $\pi_{F,t:\ell-1}(h_{F,t})$ is a follower policy tree from stage t onward, and $\Lambda_t \doteq \{\Gamma_{\pi_L,t:\ell-1} \mid \pi_{L,t:\ell-1} \in \Pi_{L,t:\ell-1}\}$, then (6) becomes:

$$v^*_{L,t}(\tilde{c}) = \max_{\Gamma \in \Lambda_t} \left\{ \rho_L(o) + \gamma^t \sum_{h_F} \Pr(h_F \mid o) \cdot \alpha^{o_{h_F}}_L(o_{h_F}) \mid (o, \alpha^{\cdot}) \in \operatorname*{argmax}_{o' \in \tilde{c}, \alpha' \in \Gamma^{|\mathbb{C}_F(\tilde{c})|}} \left(\rho_F(o') + \gamma^t \sum_{h_F} \Pr(h_F \mid o) \cdot \alpha'^{o_{h_F}}_F(o'_{h_F}) \right) \right\}.$$

Which ends the proof. □

Theorem D.11. *The optimal leader state-value function $v^*_{L,t} : \tilde{C} \to \mathbb{R}$ at stage t is uniformly continuous across credible sets.*

Proof. Let $v^{\Gamma}_{\alpha^{\cdot}} : o \mapsto \rho_L(o) + \gamma^t \sum_{h_F} \Pr(h_F \mid o) \cdot \alpha^{o_{h_F}}_L(o_{h_F})$ and $u^{\Gamma}_{\alpha^{\cdot}} : o \mapsto \rho_F(o) + \gamma^t \sum_{h_F} \Pr(h_F \mid o) \cdot \alpha^{o_{h_F}}_F(o_{h_F})$. Both are uniformly continuous over occupancy states O linearity of ρ_i and $\alpha^{\cdot}_i$ for all $i \in \{L, F\}$. The result then follows directly from Lemma D.10 and Lemma D.9, which ensures the uniform continuity of

$$v^* : \tilde{c} \mapsto \max_{\Gamma \in \Lambda} \{v^{\Gamma}_{\alpha^{\cdot}}(o) \mid (o, \alpha^{\cdot}) \in \operatorname*{argmax}_{o' \in \tilde{c}, \alpha'^{\cdot} \in \Gamma^{|\mathbb{C}_F(\tilde{c})|}} u^{\Gamma}_{\alpha'^{\cdot}}(o')\}.$$

□

Theorem D.12. *Let $\sigma \doteq \sup_{\tilde{c} \in \tilde{C}} \min_{\tilde{c}' \in \tilde{C}'} d_H(\tilde{c}, \tilde{c}')$ be the Hausdorff covering radius of the PBVI sample set $\tilde{C}'$. Then the exploitability ε of the leader policy returned by PBVI satisfies:*

$$\varepsilon \leq \tfrac{2m\sigma}{(1-\gamma)^2} \left[1 + \ell\gamma^{\ell+1} - (\ell+1)\gamma^{\ell} \right].$$

Proof. Let $\pi_L \doteq \delta_{L,0:\ell-1}$ be an optimal leader policy with value $v^*_L(\tilde{c}_0)$. Let $(\tilde{c}_0, \dots, \tilde{c}_\ell)$ denote the sequence of credible sets induced by π_L and the follower (best-response) policy $\pi_F \doteq \delta_{F,0:\ell-1}$, for which PBVI yields the worst estimate. Let $(\tilde{c}'_0, \dots, \tilde{c}'_\ell)$ be the closest sequence of credible sets to $(\tilde{c}_0, \dots, \tilde{c}_\ell)$ in Hausdorff distance $d_H(\cdot,\cdot)$, induced by the sampled credible sets $\tilde{C}'_{0:\ell}$. As a consequence, the following inequality holds $d_H(\tilde{c}_t, \tilde{c}'_t) \leq \sigma$ for any stage t. Let v_L be the approximate value function, $\pi'_L \doteq \delta'_{L,0:\ell-1}$ the induced leader policy computed by PBVI over $\tilde{C}'_{0:\ell}$ and the follower (best-response) policy $\pi'_F \doteq \delta'_{F,0:\ell-1}$. Let $v^{\pi_L,\pi_F}_{L,0:\ell}$ and $v^{\pi'_L,\pi'_F}_{L,0:\ell}$ be the value functions induced by pairs of behavioural strategies, each uniformly continuous over credible sets, such that $v^*_{L,0}(\tilde{c}_0) = v^{\pi_L,\pi_F}_{L,0}(s_0)$ and $v_{L,0}(\tilde{c}_0) = v^{\pi'_L,\pi'_F}_{L,0}(s_0)$, respectively. Then,

$$\varepsilon \doteq v^*_{L,0}(\tilde{c}_0) - v^{\pi'_L,\pi'_F}_{L,0}(s_0).$$

Let $\nu_L(o_t, \delta_{L,t}, \delta_{F,t}) \doteq \mathbb{E}[r(s_t, a_{L,t}, a_{F,t}, s_{t+1}) \mid o_t, \delta_{L,t}, \delta_{F,t}]$ denotes the expected immediate leader payoff upon taking decision rules $(\delta_{L,t}, \delta_{F,t})$ in occupancy state $o_t \doteq o_{\delta_{L,0:t-1}\delta_{F,0:t-1}} \in \tilde{c}_t$. Given $v^*_{L,0}(\tilde{c}_0) = v^{\pi_L,\pi_F}_{L,0}(s_0)$ and $v_{L,0}(\tilde{c}_0) = v^{\pi'_L,\pi'_F}_{L,0}(s_0)$, it follows that:

$$\begin{aligned}
v^*_{L,0}(\tilde{c}_0) - v^{\pi'_L,\pi'_F}_{L,0}(s_0) &= v^{\pi_L,\pi_F}_{L,0}(s_0) - v^{\pi'_L,\pi'_F}_{L,0}(s_0) \\
&= \left(\textstyle\sum_{t=0}^{\ell-1} \gamma^t \cdot \nu_L(o_t, \delta_{L,t}, \delta_{F,t})\right) - v^{\pi'_L,\pi'_F}_{L,0}(s_0) \quad \text{(by definition)} \\
&= \left(\textstyle\sum_{t=0}^{\ell-1} \gamma^t \cdot \nu_L(o_t, \delta_{L,t}, \delta_{F,t})\right) - \textstyle\sum_{t=1}^{\ell} \gamma^t (v^{\pi'_L,\pi'_F}_{L,t}(o_t) - v^{\pi'_L,\pi'_F}_{L,t}(o_t)) - v^{\pi'_L,\pi'_F}_{L,0}(s_0) \quad \text{(adding zero).}
\end{aligned}$$

Using the convention $v^{\pi'_L,\pi'_F}_{L,\ell}(\cdot) \doteq 0$, we rearrange terms:

$$\begin{aligned}
&= \textstyle\sum_{t=0}^{\ell-1} \gamma^t \cdot \nu_L(o_t, \delta_{L,t}, \delta_{F,t}) + \left(\gamma^\ell \cdot v^{\pi'_L,\pi'_F}_{L,\ell}(o_\ell) + \sum_{t=1}^{\ell-1} \gamma^t \cdot v^{\pi'_L,\pi'_F}_{L,t}(o_t)\right) - \left(\gamma^0 \cdot v^{\pi'_L,\pi'_F}_{L,0}(s_0) + \sum_{t=1}^{\ell-1} \gamma^t \cdot v^{\pi'_L,\pi'_F}_{L,t}(o_t)\right) \\
&= \textstyle\sum_{t=0}^{\ell-1} \gamma^t \cdot \nu_L(o_t, \delta_{L,t}, \delta_{F,t}) + \sum_{t=0}^{\ell-1} \gamma^{t+1} \cdot v^{\pi'_L,\pi'_F}_{L,t}(o_{t+1}) - \sum_{t=0}^{\ell-1} \gamma^t \cdot v^{\pi'_L,\pi'_F}_{L,t}(o_t) \\
&= \textstyle\sum_{t=0}^{\ell-1} \gamma^t \left(\nu_L(o_t, \delta_{L,t}, \delta_{F,t}) + \gamma \cdot v^{\pi'_L,\pi'_F}_{L,t}(o_{t+1}) - v^{\pi'_L,\pi'_F}_{L,t}(o_t)\right) \\
&= \textstyle\sum_{t=0}^{\ell-1} \gamma^t \left(q^{\pi'_L,\pi'_F}_{L,t}(o_t, \delta'_{L,t}, \delta'_{F,t}) - v^{\pi'_L,\pi'_F}_{L,t}(o_t)\right).
\end{aligned}$$

Now substitute $o'_t \in \tilde{c}'_t$ such that $\|o_t - o'_t\|_1 \leq \sigma$ in place of o_t. Notice that o'_t must exist since $d_H(\tilde{c}_t, \tilde{c}'_t) \leq \sigma$. Thus,

$$= \textstyle\sum_{t=0}^{\ell-1} \gamma^t \left(q^{\pi'_L,\pi'_F}_{L,t}(o_t, \delta'_{L,t}, \delta'_{F,t}) - q^{\pi'_L,\pi'_F}_{L,t}(o'_t, \delta'_{L,t}, \delta'_{F,t}) + q^{\pi'_L,\pi'_F}_{L,t}(o'_t, \delta'_{L,t}, \delta'_{F,t}) - v^{\pi'_L,\pi'_F}_{L,t}(o_t)\right).$$

Because the greedy rule for $q^{\pi'_L,\pi'_F}_{L,t}(o'_t, \cdot, \cdot)$ achieves value $v^{\pi'_L,\pi'_F}_{L,t}(o'_t)$, we have:

$$\begin{aligned}
&\leq \textstyle\sum_{t=0}^{\ell-1} \gamma^t \left(q^{\pi'_L,\pi'_F}_{L,t}(o_t, \delta'_{L,t}, \delta'_{F,t}) - q^{\pi'_L,\pi'_F}_{L,t}(o'_t, \delta'_{L,t}, \delta'_{F,t}) + v^{\pi'_L,\pi'_F}_{L,t}(o'_t) - v^{\pi'_L,\pi'_F}_{L,t}(o_t)\right) \\
&= \textstyle\sum_{t=0}^{\ell-1} \gamma^t \left[(q^{\pi'_L,\pi'_F}_{L,t}(o_t, \delta'_{L,t}, \delta'_{F,t}) - v^{\pi'_L,\pi'_F}_{L,t}(o_t)) - (q^{\pi'_L,\pi'_F}_{L,t}(o'_t, \delta'_{L,t}, \delta'_{F,t}) - v^{\pi'_L,\pi'_F}_{L,t}(o'_t))\right] \\
&= \textstyle\sum_{t=0}^{\ell-1} \gamma^t \left[q^{\pi'_L,\pi'_F}_{L,t}(\cdot, \delta'_{L,t}, \delta'_{F,t}) - v^{\pi'_L,\pi'_F}_{L,t}(\cdot)\right] \cdot [o_t - o'_t], \text{ (by linearity over occupancy states).}
\end{aligned}$$

Applying Hölder's inequality, using the definition of σ, and the fact that payoffs are bounded:

$$\begin{aligned}
v^*_{L,0}(\tilde{c}_0) - v^{\pi'_L,\pi'_F}_{L,0}(s_0) &\leq \textstyle\sum_{t=0}^{\ell-1} \gamma^t \cdot \left\| q^{\pi'_L,\pi'_F}_{L,t}(\cdot, \delta'_{L,t}, \delta'_{F,t}) - v^{\pi'_L,\pi'_F}_{L,t}(\cdot)\right\|_\infty \cdot \|o_t - o'_t\|_1 \\
&\leq \sigma \textstyle\sum_{t=0}^{\ell-1} \gamma^t \cdot \left\| q^{\pi'_L,\pi'_F}_{L,t}(\cdot, \delta'_{L,t}, \delta'_{F,t}) - v^{\pi'_L,\pi'_F}_{L,t}(\cdot)\right\|_\infty \\
&\leq 2m\sigma \textstyle\sum_{t_0=0}^{\ell-1} \gamma^{t_0} \sum_{t_1=t_0}^{\ell-1} \gamma^{t_1-t_0}, \text{ (where } q^{\pi'_L,\pi'_F}_{L,t} \text{ and } v^{\pi'_L,\pi'_F}_{L,t} \text{ linear over occupancy states)} \\
&= 2m\sigma \textstyle\sum_{t_0=0}^{\ell-1} \sum_{t_1=t_0}^{\ell-1} \gamma^{t_1} \\
&= 2m\sigma \textstyle\sum_{t_0=0}^{\ell-1} \frac{\gamma^{t_0} - \gamma^\ell}{1-\gamma} \\
&= \frac{2m\sigma}{1-\gamma} \textstyle\sum_{t_0=0}^{\ell-1} (\gamma^{t_0} - \gamma^\ell) \\
&= \frac{2m\sigma}{(1-\gamma)^2} \left[1 + \ell\gamma^{\ell+1} - (\ell+1)\gamma^\ell\right].
\end{aligned}$$

Which ends the proof. □

Greedy Leader Decision Rules

Definition 8. *Let o be an occupancy state. A conditional occupancy state o_{h_F} induced by occupancy state o and follower history h_F is given by: for any state $s \in S$ and leader history $h_L \in H_L$,*

$$o_{h_F}(s, h_L) \doteq \frac{o(s, h_L, h_F)}{\sum_{s,h_L} o(s, h_L, h_F)}.$$

Each occupancy state o admits a convex decomposition of the form

$$o \;=\; \textstyle\sum_{h_F} \Pr(h_F \mid o) \cdot (o_{h_F} \otimes \boldsymbol{e}_{h_F}),$$

where $\boldsymbol{e}_{h_F}$ is the one-hot vector over follower histories and $\otimes$ denotes the Kronecker product.

Lemma D.13. *Let o be an occupancy state, o_{h_F} be a conditional occupancy state induced by occupancy state o and follower history h_F. Let δ_L be a leader decision rule, and a_F be a follower action. The next conditional occupancy state $o_{h_F a_F s'} \doteq \tau_{s'}(o_{h_F}, \delta_L, a_F)$ upon receiving next state s' after taking leader-follower decisions (δ_L, a_F) starting in conditional occupancy state o_{h_F} is given by: for any state s' and leader history $h'_L \doteq h_L a_L s'$,*

$$o_{h_F a_F s'}(s', h'_L) \propto \textstyle\sum_s o_{h_F}(s, h_L) \cdot \delta_L(a_L | h_L) \cdot p(s' | s, a_L, a_F),$$

with normalising constant $\eta_{s'}(o_{h_F}, \delta_L, a_F)$.

$$\delta_L^{\Gamma,o} \in \operatorname{argmax}_{\delta_L \in \Delta_L} q_{L,t}^{\Gamma}(o, \delta_L) \tag{7}$$

$$\text{subject to: } q_{F,t}^{\Gamma}(o, \delta_L) \geq q_{F,t}^{\Gamma}(o\prime, \delta_L),\ \forall o\prime \in \tilde{c} \tag{8}$$

$$q_{L,t}^{\Gamma}(o, \delta_L) = \rho_L(o) + \gamma^t \textstyle\sum_{h_F} \Pr(h_F \mid o) \cdot g_{L,t}^{\Gamma}(o_{h_F}, \delta_L) \tag{9}$$

$$q_{F,t}^{\Gamma}(o\prime, \delta_L) = \rho_F(o\prime) + \gamma^t \textstyle\sum_{h_F} \Pr(h_F \mid o\prime) \cdot g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L), \forall o\prime \in \tilde{c} \tag{10}$$

$$g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L) \geq \sum_{s' \in S} \beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s'),\ \forall a_F \in A_F, \forall o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}) \tag{11}$$

$$g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L) \leq \sum_{s' \in S} \beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s') + \mathtt{M} \cdot (1 - \kappa_F^{\Gamma,o}(o\prime_{h_F}, a_F)),\ \forall a_F \in A_F, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}) \tag{12}$$

$$\sum_{a_F \in A_F} \kappa_F^{\Gamma,o}(o\prime_{h_F}, a_F) = 1,\ \forall o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}) \tag{13}$$

$$\beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s') \geq \nu_F(o\prime_{h_F}, \delta_L, a_F, s') + \gamma \alpha_F(\tau_{s'}(o\prime_{h_F}, \delta_L, a_F)),\ \forall s' \in S, a_F \in A_F, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}), \alpha \in \Gamma \tag{14}$$

$$\beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s') \leq \nu_F(o\prime_{h_F}, \delta_L, a_F, s') + \gamma \alpha_F(\tau_{s'}(o\prime_{h_F}, \delta_L, a_F)) + \mathtt{M} \cdot (1 - w_F^{\Gamma,o}(o\prime_{h_F}, s', \alpha)) \tag{15}$$
$$\forall s' \in S, a_F \in A_F, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}), \alpha \in \Gamma$$

$$\sum_{\alpha \in \Gamma} w_F^{\Gamma,o}(o\prime_{h_F}, s', \alpha) = 1,\ \forall s' \in S, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}) \tag{16}$$

$$g_{L,t}^{\Gamma}(o_{h_F}, \delta_L) \geq \sum_{s'} \beta_{L,t}(o_{h_F}, \delta_L, a_F, s') - \mathtt{M} \cdot (1 - \kappa_F^{\Gamma,o}(o_{h_F}, a_F)),\ \forall a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o) \tag{17}$$

$$g_{L,t}^{\Gamma}(o_{h_F}, \delta_L) \leq \sum_{s'} \beta_{L,t}(o_{h_F}, \delta_L, a_F, s') + \mathtt{M} \cdot (1 - \kappa_F^{\Gamma,o}(o_{h_F}, a_F)),\ \forall a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o) \tag{18}$$

$$\beta_{L,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') \geq \nu_L(o_{h_F}, \delta_L, a_F, s') + \gamma \alpha_L(\tau_{s'}(o_{h_F}, \delta_L, a_F)) - \mathtt{M} \cdot (1 - w_F^{\Gamma,o}(o_{h_F}, s', \alpha)), \tag{19}$$
$$\forall s' \in S, a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o), \alpha \in \Gamma.$$

$$\beta_{L,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') \leq \nu_L(o_{h_F}, \delta_L, a_F, s') + \gamma \alpha_L(\tau_{s'}(o_{h_F}, \delta_L, a_F)) + \mathtt{M} \cdot (1 - w_F^{\Gamma,o}(o_{h_F}, s', \alpha)), \tag{20}$$
$$\forall s' \in S, a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o), \alpha \in \Gamma.$$

Table 3: This mixed-integer linear program computes the greedy leader decision rule $\delta_L^{\tilde{c}}$ maximising the stage-t leader action-value as described in Theorem D.14. $\mathtt{M}$ is a sufficiently large constant for big-M constraints, $w_F^{\Gamma}(o\prime_{h_F}, s', \alpha)$ and $\kappa_F^{\Gamma,o}(o\prime_{h_F}, a_F)$ are binary selection variables, and $\delta_L(\cdot|\cdot)$ are free variables denoting altogether the leader decision rule, and $q_{i,t}^{\Gamma}(o, \delta_L)$, $g_{i,t}^{\Gamma}(o_{h_F, \delta_L})$, $\beta_{i,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s\prime)$ are free variables. Note that $\mathbb{C}_F(o) \doteq \{o_{h_F} \mid h_F \in H_F, \sum_{s,h_L} o(s, h_L, h_F) > 0\}$ and $\mathbb{C}_F(\tilde{c}) \doteq \cup_{o' \in \tilde{c}} \mathbb{C}_F(o')$.

Interpretation of the MILP for Greedy Backup

This section explains the mixed-integer linear program (MILP) in Table 3, which performs a point-based greedy backup for the leader in a dynamic programming framework for solving leader–follower general-sum stochastic games (LF-GSSGs).

Problem setting. Given:

- a filtered credible set $\tilde{c} \subseteq \Delta(S \times H_L \times H_F)$,
- a specific occupancy state $o \in \tilde{c}$,
- and a finite set Γ of α-vector pairs (α_L, α_F),

the goal is to compute a leader decision rule $\delta_L^{\Gamma,o}$ that maximises the stage-t expected return of the leader at o, under the constraint that the follower best-responds in each conditional occupancy state $o_{h_F} \in \mathbb{C}_F(\tilde{c})$, and that the value approximations respect the α-vector representation.

High-level structure. The MILP maximises the leader's action-value function while enforcing:

1. follower best response at state o (Stackelberg consistency),
2. consistent value approximation via binary selection of α-vectors,
3. alignment of leader and follower values under these selections.

This is encoded using continuous variables (for probabilities and expected values) and binary variables (for action and vector selection).

Decision variables. The MILP defines the following key variables:

- **Leader decision rule:** $\delta_L(a_L \mid h_L)$, defined for all $a_L \in A_L$ and $h_L \in \bigcup_{o' \in \tilde{c}} H_L(o')$, where

$$H_L(o') \doteq \left\{ h_L \in H_L \;\middle|\; \sum_{s,h_F} o'(s, h_L, h_F) > 0 \right\}.$$

- **Expected values:**
  - $g^{\Gamma}_{L,t}(o_{h_F}, \delta_L)$: expected leader return at $o_{h_F} \in \mathbb{C}_F(o)$,
  - $g^{\Gamma}_{F,t}(o'_{h_F}, \delta_L)$: expected follower return at $o'_{h_F} \in \mathbb{C}_F(\tilde{c})$,
  - $\beta^{\Gamma}_{i,t}(o_{h_F}, \delta_L, a_F, s')$: expected return of player $i \in \{L, F\}$ when executing a_F and transitioning to state $s' \in S$.
- **Binary selectors:**
  - $\kappa^{\Gamma,o}_F(o_{h_F}, a_F) \in \{0, 1\}$: indicates selected follower action in each $o_{h_F} \in \mathbb{C}_F(\tilde{c})$,
  - $w^{\Gamma,o}_F(o_{h_F}, s', \alpha) \in \{0, 1\}$: selects the α-vector used to approximate value continuation at state s'.

Constraint semantics. We now explain how the constraints enforce the structure of the backup:

- **Leader value maximisation (Constraint 7):**

$$\max_{\delta_L} q^{\Gamma}_{L,t}(o, \delta_L).$$

  This defines the objective of the MILP.
- **Follower best response (Constraint 8):**

$$q^{\Gamma}_{F,t}(o, \delta_L) \geq q^{\Gamma}_{F,t}(o', \delta_L) \quad \forall o' \in \tilde{c}.$$

  The follower must prefer the policy induced at o, enforcing Stackelberg equilibrium consistency.
- **Follower value representation (Constraints 11–13):** These jointly enforce:

$$g^{\Gamma}_{F,t}(o_{h_F}, \delta_L) = \max_{a_F \in A_F} \sum_{s'} \beta^{\Gamma}_{F,t}(o_{h_F}, \delta_L, a_F, s').$$

  This is implemented via:
  - Constraint 12 (lower bound, always active),
  - Constraint 13 (upper bound, only active for selected action),
  - Constraint 14 (exactly one action selected).
- **Follower value via α_F-vectors (Constraints 15–16):** These define:

$$\beta^{\Gamma}_{F,t}(o_{h_F}, \delta_L, a_F, s') = \max_{\alpha \in \Gamma} \{\nu_F + \gamma \alpha_F(\tau)\},$$

  where $\tau = \tau_{s'}(o_{h_F}, \delta_L, a_F)$. Constraint 16 enforces that exactly one α-vector is used per state.
- **Leader value under consistent follower response (Constraints 17–18):** These encode:

$$g^{\Gamma}_{L,t}(o_{h_F}, \delta_L) = \max_{a_F} \left\{ \sum_{s'} \beta^{\Gamma}_{L,t}(o_{h_F}, \delta_L, a_F, s') \;\middle|\; a_F \text{ selected for } g^{\Gamma}_{F,t} \right\}.$$

- **Leader value via matched α_L-vectors (Constraints 19–20):** These enforce:

$$\beta^{\Gamma}_{L,t}(o_{h_F}, \delta_L, a_F, s') = \max_{\alpha \in \Gamma} \{\nu_L + \gamma \alpha_L(\tau) \mid \text{matching } \alpha_F(\tau) \text{ in follower value}\}.$$

Conclusion. This MILP operationalises the greedy dynamic programming step for the leader in an LF-GSSG. It combines leader optimisation, follower rationality, and joint value approximation via α-vector selection. The binary variables encode deterministic follower responses and consistent value decompositions, enabling scalable and interpretable value backups over filtered credible sets.

Theorem D.14. *Let Λ_{t+1} be a finite collection of sets of functions Γ approximating the optimal leader value function $v^*_{L,t+1}\colon \tilde{C} \to \mathbb{R}$ at stage $t{+}1$. Then, for any credible set $\tilde{c}$ at stage t, the greedy leader decision rule $\delta^{\tilde{c}}_L$ maximising the stage-t leader action-value is the solution of a mixed-integer linear program (see Table 3).*

Proof. The proof starts with the definition of the greedy leader decision-rule $\delta^{\tilde{c}}_L$ selection at credible set $\tilde{c}$, assuming we have access to collection Λ_{t+1} of sets Γ approximating v^*_L as leader value function v_L. That is,

$$\begin{aligned}\delta^{\tilde{c}}_L &\in \operatorname*{argmax}_{\delta_L \in \Delta_L} v_{L,t+1}(\tilde{T}(\tilde{c}, \delta_L)) \\ &\in \operatorname*{argmax}_{\delta_L \in \Delta_L} q_{L,t}(\tilde{c}, \delta_L).\end{aligned}$$

By the application of uniform continuity property, see Lemma D.10, the following holds:

$$\begin{aligned}&\delta^{\tilde{c}}_L \in \operatorname{argmax}\{q^{\Gamma}_{L,t}(o, \delta_L) \mid \{q^{\Gamma}_{F,t}(o, \delta_L) \geq q^{\Gamma}_{F,t}(o\prime, \delta_L) \mid \forall o\prime \in \tilde{c}\}, \forall o \in \tilde{c}, \delta_L \in \Delta_L, \Gamma \in \Lambda_{t+1}\} \\ &\quad q^{\Gamma}_{i,t}(o, \delta_L) \doteq \rho_i(o) + \gamma^t \textstyle\sum_{h_F} \Pr(h_F \mid o) \cdot g^{\Gamma}_{i,t}(o_{h_F}, \delta_L),\ \forall i \in I \\ &\quad g^{\Gamma}_{i,t}(o_{h_F}, \delta_L) \doteq \max_{a_F \in \bar{A}_F(o_{h_F}, \delta_L)} \sum_{s' \in S} \beta^{\Gamma}_{i,t}(o_{h_F}, \delta_L, a_F, s'),\ \forall i \in I. \\ &\quad \beta^{\Gamma}_{i,t}(o_{h_F}, \delta_L, a_F, s') \doteq \mathbb{E}[r_i(s, a_L, a_F, s') \mid o_{h_F}, \delta_L, a_F, s'] + \gamma \max_{\alpha \in \bar{\Gamma}(\tau_{s'}(o_{h_F}, \delta_L, a_F))} \alpha_i(\tau_{s'}(o_{h_F}, \delta_L, a_F)),\ \forall i \in I.\end{aligned}$$

where $\bar{\Gamma}(o_{h_F}) \doteq \operatorname{argmax}_{\alpha' \in \Gamma} \alpha'_F(o_{h_F})$ and $\bar{A}_F(o_{h_F}, \delta_L) \doteq \operatorname{argmax}_{a'_F \in A_F} \sum_{s'} \beta^{\Gamma}_{F,t}(o_{h_F}, \delta_L, a'_F, s')$. Define the expected immediate payoff for agent i, conditional occupancy state o_{h_F}, leader–follower decisions (δ_L, a_F) as follows:

$$\begin{aligned}\nu_i(o_{h_F}, \delta_L, a_F, s') &\doteq \mathbb{E}[r_i(s, a_L, a_F, s') \mid o_{h_F}, \delta_L, a_F, s'] \\ &= \sum_{s \in S} \sum_{h_L \in H_L(o_{h_F})} o_{h_F}(s, h_L) \sum_{a_L \in A_L} \delta_L(a_L \mid h_L) \cdot p(s' \mid s, a_L, a_F) \cdot r_i(s, a_L, a_F, s').\end{aligned}$$

Define the expected payoff for agent i, "value function" α, current conditional occupancy o_{h_F}, leader-follower decisions (δ_L, a_F), next state s' as follows:

$$\begin{aligned}&\alpha_i(\tau_{s'}(o_{h_F}, \delta_L, a_F)) \doteq \sum_{h_L \in H_L(o_{h_F})} \sum_{a_L \in a_L} \tau_{s'}(o_{h_F}, \delta_L, a_F)(s', (h_L, a_L, s')) \cdot \alpha_i(s', (h_L, a_L, s')), \\ &\forall \alpha \in \Gamma, \forall i \in I, \forall s' \in S, \forall \tau_{s'}(o_{h_F}, \delta_L, a_F) \in O.\end{aligned} \tag{21}$$

Where the probability of being in next state s', leader next history being (h_L, a_L, s') coming from any conditional occupancy state o_{h_F} is the following

$$\begin{aligned}&\tau_{s'}(o_{h_F}, \delta_L, a_F)(s', (h_L, a_L, s')) = \sum_{s \in S} o_{h_F}(s(h_L), h_L) \cdot \delta_L(a_L \mid h_L) \cdot p(s' \mid s(h_L), a_L, a_F) \\ &, \forall h_L \in H_L(o_{h_F}), \forall a_L \in A_L, \forall s' \in S.\end{aligned} \tag{22}$$

Re-arranging terms results in the following :

$$\delta_L^{\tilde{c}} \in \arg\max_{\delta_L \in \Delta_L} \max_{o \in \tilde{c}} \max_{\Gamma \in \Lambda_{t+1}} q_{L,t}^{\Gamma}(o, \delta_L)$$

$$\text{subject to: } q_{F,t}^{\Gamma}(o, \delta_L) = \max_{o\prime \in \tilde{c}} q_{F,t}^{\Gamma}(o\prime, \delta_L),$$

$$q_{L,t}^{\Gamma}(o, \delta_L) = \rho_L(o) + \gamma^t \textstyle\sum_{h_F} \Pr(h_F \mid o) \cdot g_{L,t}^{\Gamma}(o_{h_F}, \delta_L)$$

$$q_{F,t}^{\Gamma}(o\prime, \delta_L) = \rho_F(o\prime) + \gamma^t \textstyle\sum_{h_F} \Pr(h_F \mid o\prime) \cdot g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L)\ \forall o\prime \in \tilde{c}$$

$$g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L) = \max_{a_F \in A_F} \{\sum_{s'} \beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s')\},\ \forall o\prime_{h_F} \in \mathbb{C}_F(\tilde{c})$$

$$\beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s') = \max_{\alpha \in \Gamma} \{\nu_F(o\prime_{h_F}, \delta_L, a_F, s') + \gamma \alpha_F(\tau_{s'}(o\prime_{h_F}, \delta_L, a_F))\},\ \forall s' \in S, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c})$$

$$g_{L,t}^{\Gamma}(o_{h_F}, \delta_L) = \max_{a_F \in A_F} \left\{ \sum_{s'} \beta_{L,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') \mid \sum_{s'} \beta_{F,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') = g_{F,t}^{\Gamma}(o_{h_F}, \delta_L) \right\},\ \forall o_{h_F} \in \mathbb{C}_F(o)$$

$$\begin{aligned}\beta_{L,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') &= \max_{\alpha \in \Gamma} \left\{ \nu_L(o_{h_F}, \delta_L, a_F, s') + \gamma \alpha_L(\tau_{s'}(o_{h_F}, \delta_L, a_F)) \mid \beta_{F,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') \right. \\ &\left. = \nu_F(o_{h_F}, \delta_L, a_F, s') + \gamma \alpha_F(\tau_{s'}(o_{h_F}, \delta_L, a_F)) \right\},\ \forall s' \in S, a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o).\end{aligned}$$

Fix occupancy state $o \in \tilde{c}$ and value-vector set $\Gamma \in \Lambda_{t+1}$, the greedy leader decision rule $\delta_L^{\Gamma,o}$. Let

$$\{\kappa_F^{\Gamma,o}(o\prime_{h_F}, a_F) \mid o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}), a_F \in A_F\}$$

be a set of binary variables $\kappa_F^{\Gamma,o}(o\prime_{h_F}, a_F)$ representing the follower decision rule for occupancy state $o \in \tilde{c}$ and value-vector set $\Gamma \in \Lambda_{t+1}$. Similarly, let

$$\{w_F^{\Gamma,o}(o_{h_F}, s', \alpha) \mid o_{h_F} \in \mathbb{C}_F(\tilde{c}), s' \in S, \alpha \in \Gamma\}$$

be the set of binary variables $w_F^{\Gamma,o}(o_{h_F}, s', \alpha)$ representing the follower policy upon seeing $s' \in S$ after taking an action in conditional occupancy state $o_{h_F} \in \mathbb{C}_F(\tilde{c})$. These binary variables help reformulating maximization as inequality constraints:

$$\delta_L^{\Gamma,o} \in \operatorname{argmax}_{\delta_L \in \Delta_L} q_{L,t}^{\Gamma}(o, \delta_L)$$

$$\text{subject to: } q_{F,t}^{\Gamma}(o, \delta_L) \geq q_{F,t}^{\Gamma}(o\prime, \delta_L),\ \forall o\prime \in \tilde{c}$$

$$q_{L,t}^{\Gamma}(o, \delta_L) = \rho_L(o) + \gamma^t \textstyle\sum_{h_F} \Pr(h_F \mid o) \cdot g_{L,t}^{\Gamma}(o_{h_F}, \delta_L)$$

$$q_{F,t}^{\Gamma}(o\prime, \delta_L) = \rho_F(o\prime) + \gamma^t \textstyle\sum_{h_F} \Pr(h_F \mid o\prime) \cdot g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L), \forall o\prime \in \tilde{c}$$

$$g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L) \geq \sum_{s' \in S} \beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s'),\ \forall a_F \in A_F, \forall o\prime_{h_F} \in \mathbb{C}_F(\tilde{c})$$

$$g_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L) \leq \sum_{s' \in S} \beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s') + \mathtt{M} \cdot (1 - \kappa_F^{\Gamma,o}(o\prime_{h_F}, a_F)),\ \forall a_F \in A_F, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c})$$

$$\sum_{a_F \in A_F} \kappa_F^{\Gamma,o}(o\prime_{h_F}, a_F)) = 1,\ \forall o\prime_{h_F} \in \mathbb{C}_F(\tilde{c})$$

$$\beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s') \geq \nu_F(o\prime_{h_F}, \delta_L, a_F, s') + \gamma \alpha_F(\tau_{s'}(o\prime_{h_F}, \delta_L, a_F)),\ \forall s' \in S, a_F \in A_F, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}), \alpha \in \Gamma$$

$$\beta_{F,t}^{\Gamma}(o\prime_{h_F}, \delta_L, a_F, s') \leq \nu_F(o\prime_{h_F}, \delta_L, a_F, s') + \gamma \alpha_F(\tau_{s'}(o\prime_{h_F}, \delta_L, a_F)) + \mathtt{M} \cdot (1 - w_F^{\Gamma,o}(o\prime_{h_F}, s', \alpha))$$
$$\forall s' \in S, a_F \in A_F, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c}), \alpha \in \Gamma$$

$$\sum_{\alpha \in \Gamma} w_F^{\Gamma,o}(o\prime_{h_F}, s', \alpha) = 1,\ \forall s' \in S, o\prime_{h_F} \in \mathbb{C}_F(\tilde{c})$$

$$g_{L,t}^{\Gamma}(o_{h_F}, \delta_L) \geq \sum_{s'} \beta_{L,t}(o_{h_F}, \delta_L, a_F, s') - \mathtt{M} \cdot (1 - \kappa_F^{\Gamma,o}(o_{h_F}, a_F)),\ \forall a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o)$$

$$g_{L,t}^{\Gamma}(o_{h_F}, \delta_L) \leq \sum_{s'} \beta_{L,t}(o_{h_F}, \delta_L, a_F, s') + \mathtt{M} \cdot (1 - \kappa_F^{\Gamma,o}(o_{h_F}, a_F)),\ \forall a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o)$$

$$\beta_{L,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') \geq \nu_L(o_{h_F}, \delta_L, a_F, s') + \gamma \alpha_L(\tau_{s'}(o_{h_F}, \delta_L, a_F)) - \mathtt{M} \cdot (1 - w_F^{\Gamma,o}(o_{h_F}, s', \alpha)),$$
$$\forall s' \in S, a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o), \alpha \in \Gamma.$$

$$\beta_{L,t}^{\Gamma}(o_{h_F}, \delta_L, a_F, s') \leq \nu_L(o_{h_F}, \delta_L, a_F, s') + \gamma \alpha_L(\tau_{s'}(o_{h_F}, \delta_L, a_F)) + \mathtt{M} \cdot (1 - w_F^{\Gamma,o}(o_{h_F}, s', \alpha)),$$
$$\forall s' \in S, a_F \in A_F, o_{h_F} \in \mathbb{C}_F(o), \alpha \in \Gamma.$$

Which ends the proof. □

For each filtered credible set $\tilde{c} \in \tilde{C}'$ from a sample set of filtered credible sets $\tilde{C}' \subset \tilde{C}$, the greedy leader decision rule is extracted from the solution of the mixed-integer linear program in Theorem D.14 as follows $\delta_L^{\tilde{c}} \in \arg\max_{\delta_L \in \Delta_L} \max_{o \in \tilde{c}, \Gamma \in \Lambda_{t+1}} q_{L,t}^{\Gamma}(o, \delta_L)$. Once $\delta_L^{\tilde{c}}$ has been computed for each filtered credible set $\tilde{c}$, we now can updated the collection Λ_t with the set $\Gamma^{\tilde{c}}$ computing induced by filtered credible set $\tilde{c}$ and leader decision rule $\delta_L^{\tilde{c}}$.

Corollary D.15. *Let $v_{L,t}$ denote the current leader value function at stage t, represented by the collection Λ_t. Let $\tilde{C}' \subset \tilde{C}$ a sample set of filtered credible sets. Let $\tilde{c} \in \tilde{C}'$ be a filtered credible set, and let $(\delta_L^{\tilde{c}}, \Gamma^{\tilde{c}}, w_F^{\Gamma^{\tilde{c}}}, \kappa_F^{\Gamma^{\tilde{c}}})$ be the solution to the mixed-integer linear programme in Theorem D.14 at filtered credible set $\tilde{c}$. Define the updated leader value function $v'_{L,t}$ at stage t by augmenting Λ_t with a new set Γ of value vectors $(\alpha_L^{o_{h_F}}, \alpha_F^{o_{h_F}})$, each linear over conditional occupancy states $o_{h_F} \in \mathbb{C}_F(\tilde{c})$, where: $\forall i \in I, s \in S, h_L \in \cup_{o \in \tilde{c}} H_L(o)$,*

$$\alpha_i^{o_{h_F}}(s, h_L) \doteq \sum_{s' \in S} \sum_{a_F \in A_F} \kappa_F^{\Gamma^{\tilde{c}}}(o_{h_F}, a_F) \sum_{a_L \in A_L} \delta_L^{\tilde{c}}(a_L | h_L) \cdot p(s' \mid s, a_L, a_F) \left(r_i(s, a_L, a_F, s') + \gamma \sum_{\alpha \in \Gamma^{\tilde{c}}} w_F^{\Gamma^{\tilde{c}}}(o_{h_F}, s', \alpha) \cdot \alpha_i(s', h_L a_L s') \right).$$

Then the updated value function satisfies $v'_{L,t}(\tilde{c}) \geq v_{L,t}(\tilde{c})$ for all filtered credible sets $\tilde{c} \in \tilde{C}'$, and $v'_{L,t}(\tilde{c}) > v_{L,t}(\tilde{c})$ for at least one such $\tilde{c}$ whenever the greedy update yields a strict improvement.

E Point-Based Value Iteration for the Credible MDP

This appendix provides the algorithmic details for computing an approximate strong Stackelberg equilibrium (SSE) in the original leader–follower game M using point-based value iteration (PBVI) over the reduced credible MDP $\mathcal{M}'$. This method performs dynamic programming over filtered credible sets, constructing value function approximations using vector pairs that encode leader and follower returns while maintaining Stackelberg-consistent best responses at every step.

Overview of the PBVI Algorithm

Algorithm 1 outlines the PBVI loop. The algorithm maintains:

- a sequence of filtered credible sets $\tilde{C}'_0, \ldots, \tilde{C}'_\ell$, representing reachable belief supports up to horizon ℓ,
- corresponding value approximations $\Lambda_0, \ldots, \Lambda_\ell$, where each Λ_t is a collection of vector sets $\Gamma \subseteq \mathbb{R}^{|\mathbb{C}_F(\tilde{c})|} \times \mathbb{R}^{|\mathbb{C}_F(\tilde{c})|}$ approximating the leader and follower value functions over conditional occupancy states.

Algorithm 1: Computing an approximation of a SSE in M via Point-Based Value Iteration (PBVI) for M'.

function `solveSSE`(M)
$\tilde{C}'_0 \leftarrow \{c_0\}$ (initial set of credible sets)
$\tilde{C}'_1, \ldots, \tilde{C}'_\ell \leftarrow \emptyset$ (sample sets of credible sets)
$\Lambda_0, \ldots, \Lambda_\ell \leftarrow \emptyset$ (collections of sets of vector pairs)
while not converged **do**
  for $t = 0$ to $\ell - 1$ **do**
    $\tilde{C}'_{t+1} \leftarrow$ `expand`$(\tilde{C}'_t, \tilde{C}'_{t+1})$
  for $t = \ell - 1$ to 0 **do**
    $\Lambda_t \leftarrow$ `backup`$(\tilde{C}'_t, \Lambda_{t+1})$
  $(\Lambda_t, \tilde{C}'_t) \leftarrow$ `pruning`$(\Lambda_t, \tilde{C}'_t)$
return `extractSSE`$(\Lambda_0, \ldots, \Lambda_\ell)$.

Each PBVI iteration alternates between three phases:

1. **Expansion** (Algorithm 2): simulate reachable belief sets,
2. **Backup** (Algorithm 3): update value approximations using MILPs,
3. **Pruning** (Algorithm 4, Algorithm 5): reduce redundant vectors and belief points.

The process terminates once value approximations stabilise and an SSE policy can be extracted.

Phase 1: Expansion of Reachable Credible Sets

Algorithm 2 defines the expansion routine `expand`, which constructs the next-stage set $\tilde{C}'_{t+1}$ by sampling possible transitions from each $\tilde{c} \in \tilde{C}'_t$.

Algorithm 2: The `expand` phase for PBVI in M'

function `expand`$(\tilde{C}'_t, \tilde{C}'_{t+1})$
for each $\tilde{c} \in \tilde{C}'_t$ **do**
  Filtered credible set $\tilde{c}' \leftarrow \emptyset$ and set $\mathbb{C}_F(\tilde{c}') \leftarrow \emptyset$
  Sample leader decision rule $\delta_L \in \Delta_L$
  for each occupancy state $o \in \tilde{c}$ **do**
    Initialise $\delta_F \leftarrow \emptyset$
    `isInteresting`$[\delta_F] \leftarrow$ **false**
    for each $h_F \in H_F$ reachable from o **do**
      Sample $\delta_F(h_F) \sim$ `uniform`(A_F)
      for each $s' \in S$ **do**
        $o' \leftarrow \tau_{s'}(o_{h_F}, \delta_L, \delta_F(h_F))$
        if $\mathbb{C}_F(\tilde{c}') = \emptyset$ or $\max_{\bar{o}' \in \mathbb{C}_F(\tilde{c}')} \|\bar{o}' - o'\|_1 > \varepsilon$ **then**
          `isInteresting`$[\delta_F] \leftarrow$ **true**
          $\mathbb{C}_F(\tilde{c}') \leftarrow \mathbb{C}_F(\tilde{c}') \cup \{o'\}$
    if `isInteresting`$[\delta_F]$ **then**
      $\tilde{c}' \leftarrow \tilde{c}' \cup \{\tau(o, \delta_L, \delta_F)\}$
  if $\tilde{C}'_{t+1} = \emptyset$ or $\max_{\tilde{c}'' \in \tilde{C}'_{t+1}} d_H(\tilde{c}'', \tilde{c}') > \varepsilon$ **then**
    $\tilde{C}'_{t+1} \leftarrow \tilde{C}'_{t+1} \cup \{\tilde{c}'\}$
return $\tilde{C}'_{t+1}$

For each $o \in \tilde{c}$:

- A leader decision rule $\delta_L \in \Delta_L$ is sampled.
- Follower policies δ_F are sampled by selecting actions uniformly at random for each reachable follower-private history h_F.
- The resulting conditional occupancy states $o' \in \mathbb{C}_F(\tilde{c}')$ are computed using the environment transition $\tau_{s'}(o_{h_F}, \delta_L, \delta_F(h_F))$ for all $s' \in S$.
- Only sufficiently novel conditional states o' are retained (based on ℓ_1-norm or Hausdorff distance), ensuring diversity of belief support.

This phase efficiently expands the reachable belief space while avoiding redundancy.

Phase 2: Backing Up the Value Function

Algorithm 3 performs value backups at each filtered credible set $\tilde{c} \in \tilde{C}'_t$, using the current approximation Λ_{t+1}.

Algorithm 3: The `backup` phase for PBVI in $\mathcal{M}'$

function `backup`$(\tilde{C}'_t, \Lambda_{t+1})$
for each filtered credible set $\tilde{c} \in \tilde{C}'_t$ **do**
  for each occupancy state $o \in \tilde{c}$ **do**
    for each occupancy state $\Gamma \in \Lambda_{t+1}$ **do**
      $(\delta_L^{\Gamma,o}, w_F^{\Gamma,o}, \kappa_F^{\Gamma,o}, v^{\Gamma,o}) \leftarrow$ `MILP`$(\tilde{c}, o, \Gamma)$, (cf. Table 3)
      $\Phi \leftarrow \Phi \cup \{(\delta_L^{\Gamma,o}, w_F^{\Gamma,o}, \kappa_F^{\Gamma,o}, v^{\Gamma',o})\}$
  $(\delta_L^{\tilde{c}}, w_F^{\tilde{c}}, \kappa_F^{\tilde{c}}, v^{\tilde{c}}) \leftarrow \operatorname{argmax}_{(\delta_L^{\Gamma,o}, w_F^{\Gamma,o}, \kappa_F^{\Gamma,o}, v^{\Gamma,o}) \in \Phi} v^{\Gamma,o}$
  $\Gamma^{\tilde{c}} \leftarrow$ `improve`$(\tilde{c}, \delta_L^{\tilde{c}}, w_F^{\tilde{c}}, \kappa_F^{\tilde{c}})$
  $\Lambda_t \leftarrow \Lambda_t \cup \{\Gamma^{\tilde{c}}\}$
pruning of Λ_t and $\tilde{C}'_t$
return Λ_t

For each $o \in \tilde{c}$ and $\Gamma \in \Lambda_{t+1}$:

- The MILP in Table 3 is solved to obtain:
  - a greedy leader decision rule $\delta_L^{\Gamma,o}$,
  - a consistent follower best response $w_F^{\Gamma,o}$,
  - and the resulting leader value $v^{\Gamma,o}$.
- The best such triple is selected and used to construct a new vector set $\Gamma^{\tilde{c}}$ via `improve`, which is then added to Λ_t.

This step updates the value function while preserving SSE rationality through explicit best-response modelling.

Phase 3: Pruning Redundant Representations

To ensure computational efficiency, two pruning steps are applied:

(a) Value pruning. Algorithm 4 retains only vector sets $\Gamma \in \Lambda$ that are optimal for some filtered credible set $\tilde{c} \in \tilde{C}'$. This ensures that all retained Γ contribute to the current upper envelope of the value function.

Algorithm 4: Bounded Set Pruning

function `BoundedPruning`$(\Lambda, \tilde{C}')$
for each $\Gamma \in \Lambda$ **do**
  `isKept`$[\Gamma] \leftarrow$ **false**
for each $\tilde{c} \in \tilde{C}'$ **do**
  $\Gamma_x \leftarrow \operatorname{argmax}_{\Gamma \in \Lambda} v_L^{\Gamma}(\tilde{c})$
  `isKept`$[\Gamma_{\tilde{c}}] \leftarrow$ **true**
return $\{\Gamma \in \Lambda \mid$ `isKept`$[\Gamma]\}$

(b) Belief support pruning. Algorithm 5 removes filtered credible sets that are well approximated by existing ones. Specifically, a set $\tilde{c}$ is retained only if the leader value induced by its optimal $\Gamma_{\tilde{c}}$ differs from all others by more than a threshold ϵ.

Algorithm 5: Pruning Uncredible Sets

function `PruneUncredibleSets`$(\Lambda, \tilde{C}', \epsilon)$
Initialise $\tilde{C}^{\circ} \leftarrow \emptyset$
for each $\tilde{c} \in \tilde{C}'$ **do**
  $\Gamma_{\tilde{c}} \leftarrow \operatorname{argmax}_{\Gamma \in \Lambda} v_L^{\Gamma}(\tilde{c})$
for each $\tilde{c} \in \tilde{C}'$ **do**
  if for all $\tilde{c}' \in \tilde{C}^{\circ}$, $\left| v_L^{\Gamma_{\tilde{c}}}(\tilde{c}) - v_L^{\Gamma_{\tilde{c}'}}(\tilde{c}) \right| > \epsilon$ **then**
    $\tilde{C}^{\circ} \leftarrow \tilde{C}^{\circ} \cup \{\tilde{c}\}$
return $\tilde{C}^{\circ}$

Together, these operations keep the sampled support compact without compromising coverage.

Extracting a Strong Stackelberg Policy

Once PBVI converges, the sequence $\Lambda_0, \ldots, \Lambda_\ell$ contains all the information required to reconstruct an approximate SSE policy.

Recursive representation. Each vector pair $(\alpha_L, \alpha_F) \in \Gamma$ stores:

- the leader decision rule δ_L applied at its parent filtered credible set,
- the follower action(s) selected in each conditional occupancy state,
- pointers to the next-stage vector pairs $(\alpha'_L, \alpha'_F) \in \Gamma' \in \Lambda_{t+1}$, for each possible environment transition $s' \in S$.

Unrolling the policy. Policy extraction proceeds as follows:

1. Identify the initial filtered credible set $\tilde{c}_0$ and the vector pair $(\alpha_L^*, \alpha_F^*) \in \Gamma^* \in \Lambda_0$ maximising the leader value.
2. At each time step t, record the stored δ_L as the leader policy, and the corresponding best-response mapping from conditional occupancy states to a_F.
3. Follow the transition pointers to the next-stage vector pairs (α'_L, α'_F) and repeat until horizon ℓ.

Result. This process reconstructs:

- a full sequence of stage-wise leader decision rules $\delta_{L,0:\ell-1}$,
- best-response mappings for the follower consistent with SSE semantics,
- policies grounded in value approximations stored during PBVI.

The resulting leader–follower policy profile is thus consistent with the approximate value function and satisfies the SSE assumptions by construction. This avoids re-solving any decision problem at execution time.

F Baseline Algorithms

This appendix provides algorithmic details for the baseline methods used in our experiments.

PBVI Baselines

We consider two variations of the Point-Based Value Iteration (PBVI) method introduced in Appendix E:

PBVI-H. This variant computes policy decision rules that map private decision-rule histories to actions. It allows policies to be fully history-dependent.

PBVI-S. This variant restricts the leader's policies to be Markovian, i.e., dependent only on the current state. This constraint enables state-based filtering of credible sets.

Backward Induction Baselines

We evaluate two versions of backward induction (BI), differing in the criterion used to compute follower responses.

BI. This method corresponds to the formulation in Proposition 2.1, where the follower response maximises the complete expected cumulative reward.

MY. In this variant, the follower is assumed to be *myopic*, selecting actions solely to maximise immediate reward. This results in a myopic best response that disregards long-term value.

Normal-Form Game (NFG) Baselines

These baselines reformulate the sequential game as a one-shot normal-form game (NFG), where each row and column of the payoff matrix corresponds to a deterministic policy of the leader and follower, respectively.

Although normal-form representations typically model memoryless (Markovian) interactions, we adapt this formulation to support history-dependent policies. Let Π_L^{det} and $\Pi_F^{\text{det}} \equiv \Pi_F$ denote the sets of deterministic policies for the leader and follower, respectively.

LP Baseline. We use the OpenSpiel implementation of the LP-based SSE solver introduced by Conitzer and Sandholm (2006). For each fixed deterministic follower policy $\pi_F^{\text{det}} \in \Pi_F$, the following linear program identifies a stochastic leader policy $\pi_L \in \Pi_L$ that maximises the leader's expected value under the constraint that π_F^{det} remains a best response:

$$\max_{\pi_L \in \Pi_L} \quad \sum_{\pi_L^{\text{det}}} \pi_L(\pi_L^{\text{det}}) \cdot v_{L,0}^{\pi_L^{\text{det}},\pi_F^{\text{det}}}(s_0, s_0, s_0)$$

$$\text{Subject to:} \quad \sum_{\pi_L^{\text{det}}} \pi_L(\pi_L^{\text{det}}) \cdot v_{F,0}^{\pi_L^{\text{det}},\pi_F^{\text{det}}}(s_0, s_0, s_0) \geq \sum_{\pi_L^{\text{det}}} \pi_L(\pi_L^{\text{det}}) \cdot v_{F,0}^{\pi_L^{\text{det}},\bar{\pi}_F^{\text{det}}}(s_0, s_0, s_0), \quad \forall \bar{\pi}_F^{\text{det}} \in \Pi_F.$$

The optimal policy pair $(\pi_L, \pi_F^{\text{det}})$ is selected as the SSE.

MILP Baseline. We also consider the MILP formulation by Vorobeychik et al. (2014), which encodes follower tie-breaking constraints explicitly. Let $\pi : \Pi_L^{\text{det}} \times \Pi_F \to [0, 1]$ be a joint distribution over deterministic policy pairs. The MILP is defined as:

$$
\begin{aligned}
\max_{\pi} \quad & \sum_{\pi_L^{\det}} \sum_{\pi_F^{\det}} \pi(\pi_L^{\det}, \pi_F^{\det}) \cdot v_{L,0}^{\pi_L^{\det}, \pi_F^{\det}}(s_0, s_0, s_0) \\
\text{s.t.} \quad & \sum_{\pi_F^{\det}} \pi(\pi_L^{\det}, \pi_F^{\det}) = \pi_L(\pi_L^{\det}), \quad \forall \pi_L^{\det} \in \Pi_L^{\det}, \\
& \sum_{\pi_L^{\det}} \pi(\pi_L^{\det}, \pi_F^{\det}) = \pi_F(\pi_F^{\det}), \quad \forall \pi_F^{\det} \in \Pi_F, \\
& \sum_{\pi_L^{\det}} \sum_{\pi_F^{\det}} \pi(\pi_L^{\det}, \pi_F^{\det}) \cdot v_{F,0}^{\pi_L^{\det}, \pi_F^{\det}}(s_0, s_0, s_0) \geq \sum_{\pi_L^{\det}} \pi_L(\pi_L^{\det}) \cdot v_{F,0}^{\pi_L^{\det}, \bar{\pi}_F^{\det}}(s_0, s_0, s_0), \quad \forall \bar{\pi}_F^{\det} \in \Pi_F \\
& \sum_{\pi_L^{\det}} \sum_{\pi_F^{\det}} \pi(\pi_L^{\det}, \pi_F^{\det}) = 1 \\
& \pi(\pi_L^{\det}, \pi_F^{\det}) \geq 0, \pi_L(\pi_L^{\det}) \geq 0, \quad \forall \pi_L^{\det} \in \Pi_L^{\det}, \pi_F^{\det} \in \Pi_F \\
& \pi_F(\pi_F^{\det}) \in \{0, 1\}, \quad \forall \pi_F^{\det} \in \Pi_F.
\end{aligned}
$$

G Benchmark Descriptions

This appendix describes the benchmark environments used in our experiments. Each benchmark is defined by a finite state space S, leader and follower action spaces A_L and A_F, transition function T, and reward functions r_L and r_F. All benchmark files, including state definitions, action sets, transitions, and reward functions, are provided in the supplementary code.

Centipede (Rosenthal 1981). This benchmark adapts the classical Centipede game. Players alternate moves over a finite horizon ℓ: the leader acts on odd stages, the follower on even ones. At each stage, the current player may either *stop*, ending the game and claiming a disproportionate share of the accumulated rewards, or *continue*, increasing the pool for future stages. The benchmark tests commitment and strategic foresight in a setting where early defection yields short-term gain but limits cumulative reward.

Match (Vorobeychik and Singh 2012). The Match benchmark is derived from a counterexample demonstrating the suboptimality of Markov policies in general-sum stochastic games. The follower first chooses between actions that may penalise the leader (e.g., yielding a reward ε to the follower but cost $-\mathbf{M}$ to the leader). The game then transitions deterministically to states encoding the follower's past action. The leader's subsequent decision can reset the game or impose penalties. This benchmark illustrates how history-dependent leader policies can constrain follower behaviour more effectively than Markov ones.

Patrolling (Vorobeychik and Singh 2012). The Patrolling benchmark is a zero-sum security game. The leader (defender) and the follower (attacker) move across a graph of targets. At each timestep, the attacker selects a target to attack, while the defender attempts to intercept. The game is adversarial: the attacker gains if the target is undefended; otherwise, the defender succeeds. Transitions are deterministic based on agent movement. This benchmark tests planning under spatial constraints, imperfect information, and adversarial reasoning.

Dec-Tiger (Nair et al. 2003). This benchmark adapts the fully observable version of the classic Dec-Tiger problem. At each stage, both agents observe the location of a tiger behind one of two doors. They independently choose which door to open. Opening the door without the tiger yields a positive reward; opening the tiger door yields a penalty. The tiger's location is uniformly resampled at each stage. The game tests simultaneous decision-making and partial action coordination under shared state information.

MABC (Hansen, Bernstein, and Zilberstein 2004). The Multi-Agent Broadcast Channel (MABC) benchmark is a cooperative communication problem. Multiple agents share a single broadcast channel and must coordinate their transmissions to avoid collisions. At each timestep, agents independently choose whether to transmit. If exactly one agent transmits, all receive a reward of 1; otherwise, the reward is 0. The current state encodes whether the previous transmission succeeded or failed. This benchmark tests decentralised coordination through implicit signalling under communication constraints.

Scalability Limitations

The number of deterministic history-dependent policies for agent i grows exponentially with the action set A_i, state space S, and time horizon ℓ. The total number of possible deterministic policies is:

$$
|\Pi_i^{\det}| = |A_i|^{\sum_{t=0}^{\ell-1} |A_i \times S|^t}.
$$

As illustrated in Table 4, these baselines are exact but become computationally intractable beyond small horizons due to the combinatorial explosion in policy space.

Benchmark	Horizon 1	Horizon 2	Horizon 3	Horizon 4
Centipede	2^1	2^5	2^{73}	2^{1111}
Match	2^1	2^9	2^{73}	2^{585}
Tiger	2^1	2^5	2^{21}	2^{85}
MABC	2^1	2^9	2^{73}	2^{585}
Patrolling	5^1	5^{31}	5^{931}	5^{27931}

Table 4: Number of deterministic policies per agent per planning horizon, illustrating exponential blow-up.

H Detailed Experimental Results

This appendix provides additional experimental results on the benchmark problems using baseline algorithms over extended horizons, in order to analyse the scalability of the proposed method. We also report the size of the value function for the PBVI baselines, included as an additional metric in the comprehensive result table (Table 5).

	H			S			BI		MY		LP		MILP	
ℓ	V	VF size	T	V	VF size	T	V	T	V	T	V	T	V	T
CENTIPEDE (Rosenthal 1981)														
1	**1**	2	0.01	**1**	2	0.01	**1**	0.01	**1**	0.01	**1**	0.25	**1**	0.14
2	**1**	4	0.01	**1**	3	0.01	**1**	0.01	**1**	0.01	**1**	20.21	**1**	7.29
3	**2**	4	0.01	**2**	4	0.01	1	0.01	1	0.01	—	—	—	—
4	**2.67**	7	1.56	**2.67**	5	0.02	1	0.05	1	0.06	—	—	—	—
5	**4**	9	2.69	**4**	6	0.04	1	0.10	1	0.10	—	—	—	—
6	**4.67**	9	3.68	**4.67**	7	0.11	1	0.17	1	0.13	—	—	—	—
7	**6**	10	5.34	**6**	8	0.20	1	0.18	1	0.15	—	—	—	—
8	**7**	14	8.18	**6.67**	9	0.46	1	0.20	1	0.49	—	—	—	—
9	**8**	13	11.53	**8**	10	0.71	1	0.21	1	0.25	—	—	—	—
10	**9**	15	13.00	**8.67**	11	1.32	1	0.3	1	0.27	—	—	—	—
MATCH (Vorobeychik and Singh 2012)														
1	**-1000**	2	0.01	**-1000**	2	0.01	**-1000**	0.01	**-1000**	0.01	**-1000**	0.2	**-1000**	0.3
2	**-1000**	3	0.01	**-1000**	3	0.01	**-1000**	0.01	**-1000**	0.01	—	—	—	—
3	**0**	4	0.01	-1000	4	0.01	-1000	0.04	-1000	0.02	—	—	—	—
4	**0**	5	0.02	-2000	5	0.02	-2000	0.05	-2000	0.05	—	—	—	—
5	**0**	6	0.03	-2000	6	0.02	-2000	0.05	-2000	0.06	—	—	—	—
6	**0**	7	0.05	-2000	7	0.11	-2000	0.07	-2000	0.08	—	—	—	—
7	**0**	8	0.07	-3000	8	0.15	-3000	0.08	-3000	0.10	—	—	—	—
8	**0**	9	0.11	-3000	9	0.04	-3000	0.10	-3000	0.10	—	—	—	—
9	**0**	10	0.19	-3000	10	0.10	-3000	0.11	-3000	0.12	—	—	—	—
10	**0**	11	0.40	-4000	11	0.22	-4000	0.11	-4000	0.11	—	—	—	—
11	**0**	12	0.96	-4000	12	0.04	-4000	0.12	-4000	0.15	—	—	—	—
12	**0**	13	1.99	-4000	13	0.26	-4000	0.13	-4000	0.15	—	—	—	—
13	**0**	14	5.68	-5000	14	0.23	-5000	0.14	-5000	0.16	—	—	—	—
14	**0**	15	16.5	-5000	15	0.53	-5000	0.17	-5000	0.15	—	—	—	—
15	**0**	16	63.52	-5000	16	0.42	-5000	0.19	-5000	0.17	—	—	—	—
DEC-TIGER (Nair et al. 2003)														
1	**20**	2	0.01	**20**	2	0.01	**20**	0.01	**20**	0.01	**20**	0.19	**20**	0.22
2	**40**	4	0.01	**40**	3	0.01	**40**	0.01	**40**	0.01	**40**	1.07	**40**	0.23
3	**60**	6	0.04	**60**	4	0.03	**60**	0.01	**60**	0.01	—	—	—	—
4	**80**	8	0.06	**80**	5	0.04	**80**	0.05	**80**	0.09	—	—	—	—
5	**100**	10	0.10	**100**	6	0.12	**100**	0.14	**100**	0.11	—	—	—	—
6	**120**	12	0.45	**120**	7	0.30	**120**	0.20	**120**	0.10	—	—	—	—
7	**120**	14	3.94	**120**	8	0.96	**120**	0.22	**120**	0.14	—	—	—	—
8	**160**	16	83.42	**160**	9	3.30	**160**	0.22	**160**	0.11	—	—	—	—
MABC (Hansen, Bernstein, and Zilberstein 2004)														
1	**1**	2	0.01	**1**	2	0.01	**1**	0.06	**1**	0.03	**1**	0.20	**1**	0.19
2	**2**	4	0.11	**2**	3	0.05	**2**	0.09	**2**	0.08	—	—	—	—
3	**2.99**	7	0.73	**2.99**	4	0.49	**2.99**	0.05	**2.99**	0.06	—	—	—	—
4	**3.97**	9	0.96	**3.97**	5	0.53	**3.97**	0.05	**3.97**	0.06	—	—	—	—
5	**4.95**	28	3.60	**4.95**	6	0.57	**4.95**	0.05	**4.95**	0.06	—	—	—	—
6	**5.93**	74	208.6	**5.93**	7	0.59	**5.93**	0.09	**5.93**	0.08	—	—	—	—
PATROLLING (Vorobeychik and Singh 2012)														
1	**0**	2	0.01	**0**	2	0.01	**0**	0.01	**0**	0.01	**0**	0.1	**0**	0.09
2	**0**	6	0.04	**0**	3	0.02	**0**	0.02	**0**	0.02	—	—	—	—
3	**0**	9	3.41	**0**	4	1.1	**0**	0.10	**0**	0.07	—	—	—	—
4	**0**	18	5.07	**0**	5	1.58	**0**	0.01	**0**	0.01	—	—	—	—
5	**0**	26	17.26	**0**	6	3.19	**0**	0.01	**0**	0.01	—	—	—	—
6	**0**	39	66.23	**0**	7	4.0	**0**	0.14	**0**	0.26	—	—	—	—

Table 5: SSE leader value (V), runtime (T in seconds), value function size (VF size in number of hyperplane pairs). **Bold leader values** indicate the highest return across planning variants for each configuration. "—" indicates timeout.

I Sensitivity analysis

This appendix presents a sensitivity analysis of the PBVI baselines. The experimental results for both **H** and **S** are influenced by the configuration of PBVI parameters. The primary source of variability lies in the sampling subprocess, where several interdependent parameters can be tuned. Our analysis focuses on two key parameters: the number of credible sets and the maximum number of occupancy states allowed per credible set.

Sensitivity to the Number of Occupancy States

We begin by examining the sensitivity of methods **H** and **S** to the number of occupancy states. The plots below report the average results over five runs, with the number of credible sets fixed to 1 and the planning horizon set to 4. The number of occupancy states per credible set varies from 1 to 5. The leader's expansion policy is uniform.

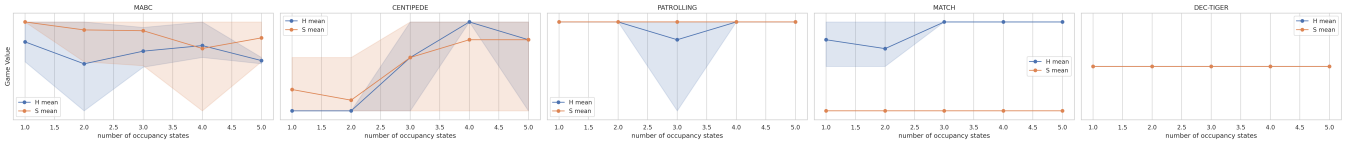

Figure 4: Value sensitivity to the number of occupancy states.

Increasing the number of occupancy states generally improves the value estimates, particularly in environments such as CENTIPEDE and MATCH. However, the magnitude of improvement is not strictly monotonic; for example, in CENTIPEDE, performance plateaus despite additional occupancy states. In DEC-TIGER and PATROLLING, optimal values are often achieved with a single occupancy state. Moreover, in some cases (e.g., PATROLLING with method **H**), adding redundant occupancy states introduces noise, slightly degrading value estimates.

Sensitivity to the Number of Credible Sets

We now analyse the effect of varying the number of credible sets, with the number of occupancy states per set fixed to 5. As before, we fix the planning horizon to 4 and use a uniform leader expansion policy. Results are averaged across five independent runs.

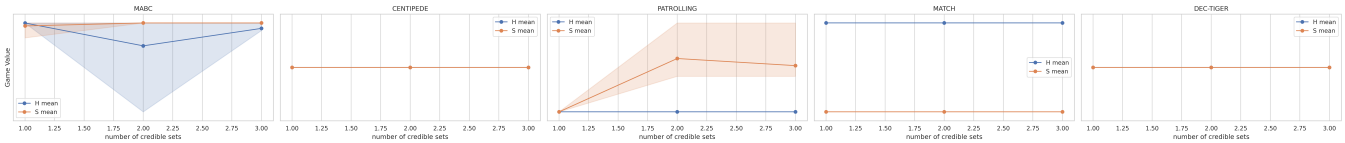

Figure 5: Value sensitivity to the number of credible sets.

In MATCH, CENTIPEDE, and DEC-TIGER, optimal values are typically reached with the first expanded credible set, and additional sets do not yield further improvements. In contrast, in MABC, adding a second credible set causes value variation for method **S**. For PATROLLING, expanding the set of beliefs allows method **H** to exploit artefacts in the follower's best responses, leading to overestimated leader policies.