CONSERVATIVE POLICY ITERATION FOR ZERO-SUM POSGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Occupancy-state reductions make dynamic programming possible for zero-sum partially observable stochastic games (zs-POSGs). Existing point-based solvers usually optimise a central value function. This restores Bellman recursion, but it requires point-based updates over high-dimensional occupancy spaces. We introduce an approximate policy-iteration algorithm for finite-horizon zs-POSGs. The method maintains one policy per player and updates both jointly along occupancy trajectories induced by conservative policy updates. Instead of approximating a full central value function, each update uses a single supporting hyperplane induced by the current joint policy. The algorithm alternates local policy improvements, conservative updates, and persistent exploration. Computation is therefore concentrated on occupancy regions where the current joint policy is exploitable. On finite-horizon zero-sum POSG benchmarks, the method is competitive at small horizons and often improves the runtime–exploitability trade-off at larger horizons compared with PBVI, HSVI, and CFR+.

1 INTRODUCTION

Planning under imperfect information requires agents to act from histories rather than from the true state. In partially observable Markov decision processes (POMDPs), belief states restore a Markovian dynamic program and support exact and point-based solvers (Aström, 1965; Smallwood & Sondik, 1973; Sondik, 1978; Pineau et al., 2003; Smith & Simmons, 2004). In partially observable stochastic games (POSGs), partial observability is coupled with strategic uncertainty. Each player must reason about the hidden state, its own information, and the possible information and actions of the other players (Hansen et al., 2004; Kovařík et al., 2022). The two-player zero-sum case, called a zero-sum partially observable stochastic game (zs-POSG), is the adversarial form of this problem. Dynamic programming is classical for fully observable stochastic games (Shapley, 1953; Filar & Vrieze, 1997), but partial observability requires a sufficient statistic for planning. Common-information, plan-time, and occupancy-state reductions provide such statistics for decentralised control and Dec-POMDPs (Nayyar et al., 2013; Dibangoye et al., 2013; 2016a; Oliehoek, 2013). Recent work has extended this programme to adversarial settings, including value-structure results, one-sided partially observable stochastic games, public-observation games, and general two-player zs-POSGs (Wiggers et al., 2016; Horák et al., 2017; Horák & Bošanský, 2019; Kartik & Nayyar, 2019; Buffet et al., 2020; Delage et al., 2023; Sokota et al., 2023; Escudie et al., 2025). In these reductions, a zs-POSG is lifted to an occupancy stochastic game. The lifted state is a distribution over hidden states and histories. The lifted actions are decision rules. The reduced game is fully observed by the planner. It preserves the value of the original game and restores Bellman recursion.

This progress leaves an algorithmic gap. Existing dynamic-programming solvers for zs-POSGs are mostly point-based value-iteration methods (Pineau et al., 2003; Smith & Simmons, 2004; Delage et al., 2023; Escudie et al., 2025). They approximate the optimal central value on sampled occupancy states and apply Bellman backups, for all samples in sync. This is principled, but it creates a coverage problem. The approximation must support backups over a large occupancy space. Player-specific policy improvement needs accurate value information only where the current policy can fail to be secure. In zero-sum games, these regions are not determined by the current policy alone. They are exposed by unilateral deviations and worst-case responses. The difficulty is therefore not only to extract executable policies from a central planner. It is the mismatch between global value approximation and local policy improvement.

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

Policy-update methods address another side of the same problem. Fictitious play, counterfactual regret minimisation, and CFR+ update policies directly in imperfect-information games (Robinson, 1951; Zinkevich et al., 2007; Neller & Lanctot, 2013; Tammelin, 2014; Moravčík et al., 2017; Brown et al., 2020). Classical fictitious play, regret minimisation, and CFR-style methods show that repeated local improvement can solve large zero-sum imperfect-information games without explicitly representing a global dynamic program (Robinson, 1951; Zinkevich et al., 2007; Neller & Lanctot, 2013; Tammelin, 2014). Sequence-form methods and information-set decompositions provide compact representations of extensive-form games (Koller et al., 1996; Moravčík et al., 2017; Brown et al., 2020; Kovařík et al., 2022). Policy-improvement and trust-region methods provide another route to executable updates in reinforcement learning and multi-agent reinforcement learning (Bertsekas & Tsitsiklis, 1996; Munos, 2003; Kakade & Langford, 2002; Schulman et al., 2015; 2017; Yu et al., 2022; Zhong et al., 2024). Search and policy improvement have also been combined in cooperative or partially observable multi-agent settings (Lerer et al., 2020; Van der Pol & Oliehoek, 2016; Wei et al., 2019; Pieroth et al., 2024). These methods are attractive because they update policies at histories or information states that are executable by construction.

Existing methods provide either a principled value equation or executable policy updates. They do not yet provide both in a form suited to focused approximation. The central occupancy reduction provides a Bellman equation, but its state is broad and not directly executable. Focal planning makes policy extraction player-specific, but its state is set-valued and lacks a tractable value representation. Policy-space methods update executable policies, but they do not provide a Bellman representation for focused point-based backups. The missing object is therefore a player-specific value representation that keeps executable decision rules while supporting local Bellman backups.

To address this gap, we introduce Conservative Policy Iteration (CPI) for finite-horizon zs-POSGs. CPI maintains one policy per player and improves both policies along occupancy trajectories induced by conservative policy updates. For each player, the current policy induces a value function in the corresponding player-specific planning process. CPI uses this policy-induced value function for local Bellman backups, instead of constructing a global central value function. Each backup computes an improved decision rule for one player from its current value function. The two policies are updated together, so the opposing policy used during rollouts is supplied by the other player-specific planning process. The improved decision rules are averaged into the current policies, and GLIE pertubations maintain exploration. Thus, updates are performed where the current joint policy is tested by adversarial interaction, rather than by uniformly covering the central occupancy space. The convergence guarantees are conditional on exact local backups, sufficient coverage, and the stated regularity assumptions. We evaluate CPI on standard finite-horizon zero-sum POSG benchmarks. The method is competitive at small horizons. At larger horizons, it often improves the runtime-exploitability trade-off compared with PBVI (Escudie et al., 2025), HSVI (Delage et al., 2023), and CFR+ (Tammelin, 2014).

2 PRELIMINARIES

Finite-horizon zero-sum POSGs are stochastic games with hidden states and imperfect information.

Definition 1. A finite-horizon zero-sum partially observable stochastic game (zs-POSG) is a tuple $\mathcal{M} = (\mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{Z}^0, p, (r^i)_{i\in\mathcal{I}}, T, b_0)$, where $\mathcal{I} = \{1, 2\}$, $\mathcal{S}$ is a finite state space, $\mathcal{A} = \times_{i\in\mathcal{I}}\mathcal{A}^i$ is the joint action space, $\mathcal{Z} = \times_{i\in\mathcal{I}}\mathcal{Z}^i$ is the joint private observation space, $\mathcal{Z}^0$ is the public observation space, $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{Z}^0 \times \mathcal{Z})$ is the transition kernel, $r^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is player i's reward, $r^2 = -r^1$, $b_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and T is the horizon.

The game unfolds over stages $t = 0, \dots, T-1$. At stage t, the players choose a_t, each player i receives $r^i(s_t, a_t)$, and the game samples $(s_{t+1}, z^0_{t+1}, z_{t+1})$ according to $p(\cdot \mid s_t, a_t)$. Because the state is hidden, i acts from its private history h^i_t and the public history h^0_t. These histories are empty at $t = 0$. For $t \geq 1$, $h^i_t \doteq a^i_0 z^i_1 \cdots a^i_{t-1} z^i_t$ records i's past actions and private observations, and $h^0_t \doteq z^0_1 \cdots z^0_t$ records the public observations. We write $h_t \doteq (h^1_t, h^2_t)$ for a joint private history, with reachable sets $\mathcal{H}^i_t$, $\mathcal{H}_t$, and $\mathcal{H}^0_t$.

A decision rule for player i is a map $\pi^i_t : \mathcal{H}^i_t \times \mathcal{H}^0_t \to \Delta(\mathcal{A}_i)$. A joint decision rule is $\pi_t = (\pi^1_t, \pi^2_t) \in \Pi_t \doteq \Pi^1_t \times \Pi^2_t$. A behavioural policy profile is $\pi = (\pi_0, \dots, \pi_{T-1}) \in \Pi$. We write $\pi^i \in \Pi^i$ for a full behavioural policy of player i. We write $\pi^i_{t:T} \in \Pi_{t:T}$ for player i's policy suffix

and $\pi_{t:T} \in \Pi_{t:T}$ for a policy-profile suffix from stage t onward. We write $\pi^i_{0:t} \in \Pi^i_{0:t}$ for player i's policy prefix and $\pi_{0:t} \in \Pi_{0:t}$ for a policy-profile prefix.

Policies are evaluated by expected cumulative payoff. Given b_0, i's payoff under π is $J^\pi_i \doteq \mathbb{E}_{b_0,p,\pi}[\sum_{k=0}^{T-1} r^i(S_k, A_k)]$. Since $r^2 = -r^1$, we write the zero-sum game from player 1's perspective. The finite-horizon zs-POSG induces a finite zero-sum extensive-form game with perfect recall, so the classical minimax and behavioural-policy foundations apply (Neumann, 1928; Kuhn, 1953). Thus,

$$v_0^{1,*}(b_0) = \max_{\pi^1 \in \Pi^1} \min_{\pi^2 \in \Pi^2} J_1^{(\pi^1,\pi^2)} = \min_{\pi^2 \in \Pi^2} \max_{\pi^1 \in \Pi^1} J_1^{(\pi^1,\pi^2)}.$$

A policy profile attaining this value is a Nash equilibrium of the induced zero-sum game. Equivalently, player i's component is maximin optimal for i's payoff for any $i \in \mathcal{I}$. This player-specific view is central below. A value-based method must ultimately select decision rules that can be implemented from each player's private and public information.

2.1 Occupancy-state dynamic programming

Bellman recursion relies on a state that is sufficient for the continuation problem. For a policy-profile prefix $\pi_{0:t}$, the central occupancy state is $o_t(\pi_{0:t}) = \mathbb{P}_{b_0,p,\pi_{0:t}}(S_t, H_t, H^0_t)$. We write o_t when the prefix is clear. It belongs to $\mathcal{O}_t \subseteq \Delta(\mathcal{S} \times \mathcal{H}_t \times \mathcal{H}^0_t)$. Although o_t is not observed by either player, it is Markovian for a central planner. Given o_t and π_t, the next occupancy is $\tau_t(o_t, \pi_t) = \mathbb{P}_{o_t,\pi_t}(S_{t+1}, H_{t+1}, H^0_{t+1})$. Player i's one-step expected reward is $\rho_{i,t}(o_t, \pi_t) = \mathbb{E}_{o_t,\pi_t}[r^i(S_t, A_t)]$. The central occupancy state restores Bellman's principle for the zero-sum value. With $v_T^{1,*}(\cdot) \doteq 0$, the central dynamic program is

$$v_t^{1,*}(o_t) = NE_{(\pi_t^1,\pi_t^2)}\{\rho_{1,t}(o_t, \pi_t) + v_{t+1}^{1,*}(\tau_t(o_t, \pi_t))\}$$

where NE denotes the operator that returns the value associated to the Nash equilibrium of the corresponding subgame.

This recursion is the basis of exact and approximate occupancy-based planning methods for Dec-POMDPs and zs-POSGs (Dibangoye et al., 2012; 2013; 2014; 2016b;a; Dibangoye & Buffet, 2018; Delage et al., 2023; Peralez et al., 2025; 2024; Escudie et al., 2025). The difficulty is not the recursion itself, but the space on which it acts. The number of coordinates needed to represent o_t grows exponentially with the horizon. Point-based value iteration is a natural approximation paradigm for such continuous-state dynamic programs, but its zs-POSG instantiations provide limited guidance on where to place samples (Pineau et al., 2003; Smith & Simmons, 2004; Shani et al., 2013; Delage et al., 2023; Escudie et al., 2025).

The central planner game provides the correct value equation and, in principle, produces the optimal decision rules. The difficulty is instead computational: the central occupancy state induces a very large decision space. In particular, solving the corresponding subgame requires optimizing over decision rules defined on the possible central states, which can be viewed as a large matrix whose rows correspond to decision rules. This representation quickly becomes intractable as the horizon and information spaces grow.

The challenge lies in exploiting the decision rules produced by the central formulation without having to explicitly enumerate or sample the entire central occupancy space. In particular, we seek a representation that retains the correct Bellman recursion while allowing the relevant parts of the decision-rule space to be explored efficiently.

3 Focal Belief Games

We rewrite the central state by conditioning on the information relevant to the focal player i and on the policies that determine the other players' behaviour. More precisely, we condition on the public history, the opponent's private history, and the focal player's policy prefix. This representation allows us to separate two steps of the solution procedure. First, during policy evaluation, we fix the focal player's current policy and compute the value it can guarantee against an opponent that best responds to it. This gives a pessimistic continuation value for the focal player's policy. Second,

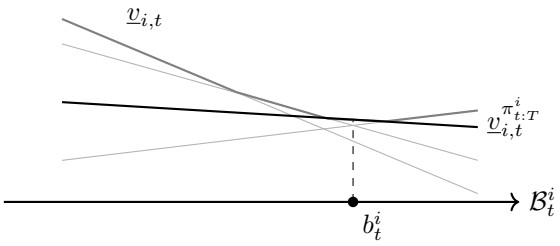

Figure 1: The optimal value function of the game $\mathcal{M}^i$ is convex in the belief space. While value-iteration methods such as PBVI sample this space and operate on a set of vectors to represent $\underline{v}_{i,t}$, our method works with a single vector $\underline{v}_{i,t}^{\pi^i_{t:T}}$ at a time, reducing the complexity of the associated algorithms.

during policy improvement, this continuation value is used to determine a better decision rule for the focal player at the current stage, in the same spirit as a Bellman backup.

3.1 Conditional occupancy states

Fix focal player i. At stage t, fix the public history h_t^0, the opponent private history h_t^{-i}, and the focal policy prefix $\pi^i_{0:t}$. Conditioning the central occupancy state on these variables gives *the conditional occupancy state*

$$c_t^i(h_t^0, h_t^{-i}, \pi^i_{0:t}) \doteq \mathbb{P}_{b_0,p}\big(S_t, H_t^i \mid H_t^0 = h_t^0, H_t^{-i} = h_t^{-i}, \pi^i_{0:t}\big)\,.$$

We write c_t^i when the histories and prefix are clear. It is a distribution over the hidden state and the focal player's private history. If the conditioning event has probability zero, the conditional update is fixed arbitrarily.

Let $\mathcal{C}_t^i$ be the set of reachable conditional occupancy states at stage t. In this conditional representation, the opponent's plan-time decision rule is written as $\pi_t^{-i}\colon \mathcal{C}_t^i \to \mathcal{A}^{-i}$, with space Π_t^{-i}. It does not denote an execution-time policy of the original zs-POSG. We use this representation for fixed-policy evaluation and improvement. For fixed-policy evaluation, pure opponent actions are sufficient: after fixing the focal player's policy suffix, the objective is affine in the opponent's mixed action, i.e., in its distribution over $\mathcal{A}^{-i}$. Hence, the minimum over mixed actions is attained by a pure action.

Given c_t^i, π_t^i, and a_t^{-i}, the model induces a probability kernel $p^{\mathsf{c}}_{i,t}(\cdot \mid c_t^i, \pi_t^i, a_t^{-i})$ on $\mathcal{C}_{t+1}^i$. This kernel is obtained by conditioning on the next public observation and the next opponent private observation. The conditional Bayes update and its observation likelihood are denoted by $\tau^{\mathsf{c}}_{i,t}$ and $w^{\mathsf{c}}_{i,t}$, and are given in Appendix A.1. The one-step conditional payoff is $g^{\mathsf{c}}_{i,t}(c_t^i, \pi_t^i, a_t^{-i})$.

For a fixed focal policy suffix $\pi^i_{t:T}$, these objects define the ingredients for a fixed-policy evaluation. The policy suffix induces a decision rule π_k^i at every stage $k \geq t$. With terminal value $v_{i,T}^{\cdot} \equiv 0$, the fixed-policy value function satisfies

$$v_{i,k}^{\pi^i_{t:T}}(c_k^i) = \min_{a_k^{-i}} \left[g^{\mathsf{c}}_{i,k}(c_k^i, \pi_k^i, a_k^{-i}) + \textstyle\sum_{c^i_{k+1} \in \mathcal{C}^i_{k+1}} p^{\mathsf{c}}_{i,k}(c^i_{k+1} \mid c_k^i, \pi_k^i, a_k^{-i}) v_{i,k+1}^{\pi^i_{k+1:T}}(c^i_{k+1}) \right].$$

This recursion evaluates a fixed focal policy suffix by accumulating one-step conditional payoffs. It is the value function used later as the continuation term in local policy improvement.

3.2 The focal belief game

We now lift the conditional occupancy states to distributions over such states. This gives the player-specific planning state used below.

Fix a focal player i. The distribution over conditional occupancy state induced at stage t *the focal belief*. A focal belief $b_t^i \in \mathcal{B}_t^i$ is a probability distribution over $\mathcal{C}_t^i$. Given b_t^i, the focal player chooses

π_t^i. The opponent's rule is represented in the conditional coordinates by π_t^{-i}. The transition τ_t^i is the law of the next focal belief under $(b_t^i, \pi_t^i, \pi_t^{-i})$. The reward g_t^i is the expected conditional payoff under the same belief and decision rules. Both are given in Appendix A.2

Definition 2. $\mathcal{M}^i \doteq (\mathcal{I}, (\mathcal{B}_t^i)_{t=0}^T, (\Pi_t^i)_{t=0}^{T-1}, (\Pi_t^{-i})_{t=0}^{T-1}, (\tau_t^i)_{t=0}^{T-1}, (g_t^i)_{t=0}^{T-1}, b_0^i, T)$ is the focal belief game associated with the original game $\mathcal{M}$.

The focal belief game is fully observed at the planning level. It is not a strict lossless reduction of the original zs-POSG $\mathcal{M}$. The opponent's decision problem is represented in conditional coordinates and can therefore be viewed as a virtual player. Fixed-policy value function have the value in both representations, and the focal decision-rule space is unchanged. The problem is inherently one-sided: we fix the policy of a focal player and evaluate how well this player performs against an opponent that minimizes the remaining payoff. For a fixed focal policy suffix $\pi_{t:T}^i$, let $\underline{v}_{i,t}^{\pi_{t:T}^i}(b_t^i)$ denote the value obtained from b_t^i when player i follows $\pi_{t:T}^i$ and the opponent chooses actions to minimize the remaining payoff. This one-sided fixed-policy value function is the value function used by CPI.

3.3 Structure of the focal belief value

The focal belief game reveals structural properties of the value function that are not explicit in the set-valued focal state representation, as the theorem below states. In particular, the fixed-policy value function $\underline{v}_{i,t}^{\pi_{t:T}^i}$ defines a linear function over the belief space $\mathcal{B}_t^i$.

Theorem 1. *For any fixed focal policy suffix $\pi_{t:T}^i$, the function $\underline{v}_{i,t}^{\pi_{t:T}^i}$ is linear on $\mathcal{B}_t^i$ and satisfies*

$$\underline{v}_{i,t}^{\pi_{t:T}^i}(b_t^i) = \sum_{c_t^i \in \mathtt{supp}(b_t^i)} b_t^i(c_t^i) v_{i,t}^{\pi_{t:T}^i}(c_t^i).$$

Although beliefs b_t^i typically have sparse support, their representation remains intractable due to the continuous nature of the conditional state space $\mathcal{C}_t^i$. Nevertheless, this complexity can be mitigated: the value function induced by a policy $\pi_{t:T}^i$ can be exactly characterized using a reduced representation, relying only on a subset of relevant components.

Proposition 1. *For any fixed focal policy suffix $\pi_{t:T}^i$, $v_{i,t}^{\pi_{t:T}^i}$ is piecewise linear and concave on $\mathcal{B}_t^i$. That is, there exists a finite collection $\Lambda_{i,t}^{\pi_{t:T}^i}$ of α-vectors $\alpha_{i,t}^{\pi_{t:T}^i} \in \mathbb{R}^{S_t \times H_t^i}$, each defining a linear function over the ConditionalSpace space $\mathcal{C}_t^i$, such that:*

$$v_{i,t}^{\pi_{t:T}^i}(c_t^i) = \min_{\alpha_{i,t}^{\pi_{t:T}^i} \in \Lambda_{i,t}^{\pi_{t:T}^i}} \alpha_{i,t}^{\pi_{t:T}^i} \cdot c_t^i.$$

4 Conservative Policy Iteration

We describe Conservative Policy Iteration (CPI) for finite-horizon zs-POSGs. CPI maintains one behavioural policy per player. At each iteration, these policies generate an occupancy trajectory after conservative averaging and GLIE exploration. For each player, it evaluates the current policy through the fixed-policy value function of Section 3. The resulting value function is used as the continuation term in a local Bellman backup. The backup returns an improved decision rule for the focal player. This rule is averaged into the current policy instead of being installed directly. CPI therefore performs local policy improvement along sampled occupancy trajectories. It does not construct the optimal value function of the focal belief game. It also does not attempt to cover the full central occupancy space.

4.1 Local policy improvement

Fix a stage t, a focal belief b_t^i, and a current focal policy suffix $\pi_{t+1:T}^i$. The fixed-policy value function $\underline{v}_{i,t+1}^{\pi_{t+1:T}^i}$ is used as the continuation value function after stage t. We write $\hat{q}_{i,t}$ for the corresponding local Q-value, defined by $\hat{q}_{i,t}(b_t^i, \pi_t) \doteq g_t^i(b_t^i, \pi_t) + \underline{v}_{i,t+1}^{\pi_{t+1:T}^i}(\tau_t^i(b_t^i, \pi_t))$.

Algorithm 1 Conservative policy iteration.

1: Initialise $(\Lambda_{i,t})_{t\in\{0,\dots,T\},i\in\mathcal{I}}$
2: Initialise $(\alpha_k)_{k\geq 0}$, $(\epsilon_k)_{k\geq 0}$, $k \leftarrow 0$
3: Initialise $(\bar{\pi}_{i,t}, u_{i,t})_{t\in\{0,\dots,T-1\},i\in\mathcal{I}}$
4: **while** not converged **do**
5: $\quad$ `Recurse`(o_0, k)
6: $\quad \pi^i_t \leftarrow \tilde{\pi}_{i,t}, \forall t \in \{0, ..., T-1\}$
7: $\quad k \leftarrow k + 1$
8: **end while**
9:
10: **Function** `CGLIE`(π^i_t, k):
11: $\quad \bar{\pi}_{i,t} \leftarrow (1-\alpha_k)\bar{\pi}_{i,t} + \alpha_k \pi^i_t$
12: $\quad$ **return** $(1-\epsilon_k)\bar{\pi}_{i,t} + \epsilon_k u_{i,t}$

1: **Function** `Recurse`(o_t, k):
2: **if** $t < T$ **then**
3: $\quad$ `Recurse`$(\tau_t(o_t, (\pi^1_t, \pi^2_t)), k)$
4: $\quad \pi^i_t \leftarrow$ `LP`$(\text{BELIEF}_i(o_t), \Lambda_{i,t+1}), \forall i \in \mathcal{I}.$
5: $\quad \tilde{\pi}_{i,t} \leftarrow$ `CGLIE`$(\pi^i_t, k), \forall i \in \mathcal{I}$
6: $\quad \Lambda_{i,t} \leftarrow$ `VECTOR`$(\tilde{\pi}_{i,t}), \forall i \in \mathcal{I}$
7: **end if**
8:
9: **Function** $\text{BELIEF}_i(o_t)$:
10: $\quad$ **return** $\mathbb{P}_{o_t}(C^i_t)$

The Bellman backup at (t, b^i_t) induces a local zero-sum continuation game, *i.e.*, the finite zero-sum game rooted at (t, b^i_t) with a fixed-policy value function. The focal player chooses a decision rule π^i_t. The opponent chooses a decision rule π^{-i}_t on conditional states. The payoff is $\hat{q}_{i,t}(b^i_t, (\pi^i_t, \pi^{-i}_t))$. We denote this local continuation game by $\mathcal{M}^i_t(\hat{q}_{i,t}(b^i_t, \cdot))$. We write $\mathcal{M}^i_t$ when the continuation value is clear from context.

Definition 3. Tuple $\mathcal{M}^i_t \doteq (\mathcal{I}, \Pi^i_t, \Pi^{-i}_t, \hat{q}_{i,t}(b^i_t, \cdot))$ is the finite zero-sum game with players $\mathcal{I}$, action spaces Π^i_t and Π^{-i}_t, and payoff $\hat{q}_{i,t}(b^i_t, \cdot)$.

Revisiting stage t at another belief therefore defines another local continuation game. It does not perturb the previous one. For fixed $\hat{q}_{i,t}$, the local backup at b^i_t is $\max_{\pi^i_t} \min_{\pi^{-i}_t} \hat{q}_{i,t}(b^i_t, (\pi^i_t, \pi^{-i}_t))$. This is the local Bellman optimisation problem represented by $\mathcal{M}^i_t(\hat{q}_{i,t}(b^i_t, \cdot))$. A local solution is a max–min optimal focal decision rule together with an opponent minimising decision rule (see Theorem 6). Since histories, observations, and actions are finite, deterministic local decision rules are finite. The focal planner may commit to mixed decision rules. The opponent admits deterministic minimising decision rules because the local payoff is affine in its mixed rule.

4.2 Conservative Policy Iteration

We now introduce conservative policy iteration (CPI). The algorithm alternes central occupancy expansion with local policy improvement in the two player-specific conditional representations. Algorithm 1 maintains a fixed-policy value representation, step sizes, and averaged decision rules for both players; see Lines 1–3. Each iteration starts from the initial central occupancy state o_0 and calls `Recurse`; see Line 5. The recursion propagates the central occupancy state because the joint policy acts on the original zs-POSG. Line 3 expands then central occupancy state using the current joint decision rule. At each reached o_t, Line 10 constructs, for each player i, the focal belief induced by conditioning o_t in player i's basis Line 4 applies the LP backup (see Theorem 6) at this focal belief and returns an improved decision rule for each player. Line 11–12 average the improved decision rule into the current decision rule and add GLIE exploration. GLIE is used to ensure that all components of the belief remain represented. Without exploration, the policy may assign zero probability to some actions, causing the induced belief to become degenerate and eliminating support from parts of the distribution. Maintaining exploration therefore prevents components of the belief from vanishing entirely, similarly to the distribution restart mechanism used in Kakade & Langford (2002). Line 6 computes the vector associated with the decision rule just computed, as established by Proposition 1.

Linear program. For a reached belief b^i_t, the local LP is written at the conditional level. The input $\Lambda_{i,t+1}$ denotes the stored lower envelop of α-vectors representing the fixed-policy value function used for the continuation term. The LP implements the local backup in $\mathcal{M}^i_t$, *cf.* Appendix A.6.1. In Algorithm 1, `LP`$(b^i_t, \lambda_{i,t+1})$ denotes this local policy-improvement routine.

Theorem 2. *Let the policy updates of player i be given by*

$$\pi^{i,k+1} = (1-\alpha_k)\pi^{i,k} + \alpha_k \tilde{F}^i\left(\pi^{i,k}, \pi^{-i,k}\right) + U_{i,k+1},$$

where $\tilde{F}^i$ represents all the stage-wise policy optimization operator $\tilde{F}^i_t$ obtained from the corresponding LP (Line 4 in Algorithm 1), and $U_{i,k+1}$ is a noise term. Assume that the step sizes satisfy the Robbins-Monro conditions

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Under the standard stochastic approximation conditions on $U_{i,k}$ and the approximation error between the discrete operator $\tilde{F}^i$ and the continuous dynamics F^i (see Definition 4), the algorithm converges almost surely to the set of Nash equilibria.

Theorem 2 establishes the convergence of Algorithm 1, since its discrete update operator coincides with the operator defined in the algorithm. The GLIE noise introduced during exploration does not affect this convergence result, as it satisfies the conditions required by the convergence theorem.

5 Empirical Evaluation

We evaluate CPI on standard finite-horizon zero-sum POSG benchmarks: *Adversarial Tiger*, *Recycling*, *Mabc*, and *Matching Pennies*. Benchmark descriptions are available at `http://masplan.org/`. Some instances originate from common-payoff problems and are converted to the zero-sum setting by negating player 1's payoff for player 2. We compare CPI with PBVI (Escudie et al., 2025), HSVI (Delage et al., 2023) and CFR+(Zinkevich et al., 2007; Tammelin, 2014). The CFR+ baseline adapts the extensive-form game CFR+ directly to the occupancy-state reduction of zs-POSGs; implementation details are given in Appendix D.1. In contrast, our CFR+ formulation naturally leverages the objects introduced in this paper, further strengthening the theoretical contribution of the paper. In particular, the POSG setting provides a compact representation of histories and occupancy states that can be exploited to speed up CFR. This allows CFR+ to reason directly over clusters of probabilistically equivalent private histories, rather than operating on the significantly larger extensive-form representation. All methods are evaluated on the same finite-horizon POSG instances. Table 1 reports runtime, value, and exploitability. Results for extensive-form games are discussed solely in Appendix C.

CPI vs PBVI. CPI is competitive with PBVI at small horizons and improves the runtime–exploitability trade-off at larger horizons. Indeed, PBVI grows exponentially with the horizon, as it adds an exponentially increasing number of occupancy states by construction. This difference is particularly visible in games such as Matching Pennies ($T = 25, 50$ in Table 1 and Figure 2), where PBVI is sensitive to the sampled belief set. For instance, insufficient or poorly distributed samples may slow convergence or degrade the quality of its central value function approximation. This sensitivity is particularly relevant because exploitability is approximated in our experiments: PBVI must converge to sufficiently low exploitability for this approximation to provide a reliable indication of safety, so a temporary increase in measured exploitability can indicate that the policy remains exploitable. In contrast, CPI updates policies along trajectories induced by the current conservative joint policy, which can guide updates toward regions where the policies are still exploitable. From a computational perspective, PBVI applies a relatively expensive backup operator over a growing collection of sampled belief points. As the horizon increases, maintaining and updating this point set becomes prohibitive. CPI instead performs local policy updates along sampled trajectories, avoiding the explicit construction of large point sets. In our experiments, this generally resulted in more stable scaling behaviour (Section A.7).

Finally, although CPI and CFR+ outperform PBVI for higher horizon, PBVI achieves stronger results on Adversarial Tiger and Recycling (see Figure 2). Adversarial Tiger is particularly well suited to methods that cover a broader occupancy space such as point-based value iteration algorithms. Notably, computation time increases proportionally with the horizon for CPI, whereas PBVI's computation time may vary depending on the sampling, as observed for Recycling in Table 1.

CPI vs CFR. Both CPI and CFR-based methods improve policies iteratively, but through different update mechanisms. In our experiments, CFR iterations are computationally cheaper than CPI updates. However, exploitability often evolves non-monotonically across iterations (see Figure 2 for Recycling and Adversarial Tiger). This finding is consistent with observations reported in (Delage et al., 2023). CPI instead performs more expensive local updates based on explicit continuation

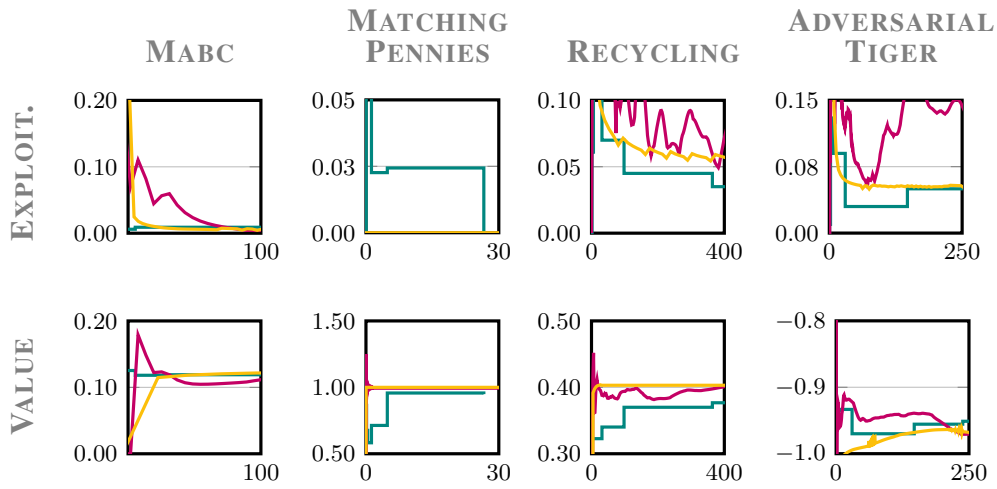

Figure 2: Evolution of value and exploitability across different games under a fixed horizon $\ell = 5$. The results compare three methods: CFR+, PBVI, and CPI, highlighting their respective performance profiles in terms of convergence behaviour and solution quality across the considered benchmarks.

Table 1: Snapshot of the empirical results, including the computational cost of exploitability evaluation. For each game, horizon and algorithm, we report time, value and exploitability ϵ associated with the computed ϵ-Nash equilibrium. OOT occurs when the target exploitability threshold cannot be reached within the prescribed time limit, while OOM means that the algorithms ran out of memory. **magenta** = lowest exploitability/value per benchmark

Problem	T	CPI (Ours)			PBVI			HSVI			CFR+		
		t	val.	ϵ	t	val.	ϵ	t	val.	ϵ	t	val.	ϵ
MABC	2	0.3	**0.078**	**0.**	**0.2**	**0.078**	**0.**	0.4	**0.078**	**0.**	0.3	**0.078**	**0.**
	3	**11**	**0.097**	**0.**	24	0.096	**0.**	70	**0.097**	**0.**	12	0.095	0.
	4	**23**	**0.11**	**0.**	65	**0.11**	0.01	OOT	OOT	OOT	49	**0.11**	0.01
	5	**29**	**0.12**	**0.**	31	**0.12**	**0.**	OOT	OOT	OOT	82	**0.12**	0.01
	7	**97**	**0.14**	0.01	190	**0.14**	**0.005**	OOT	OOT	OOT	292	**0.14**	0.02
	10	**100**	**0.18**	**0.01**	637	0.17	**0.01**	OOT	OOT	OOT	380	**0.18**	0.02
	100	**1920**	**1.06**	**0.03**	OOT	OOT	OOT	OOT	OOT	OOT	10042	**1.06**	**0.2**
MATCHING PENNIES	2	**0.01**	**0.4**	**0.**	2	0.39	**0.**	0.1	**0.4**	**0.**	0.06	**0.4**	0.005
	3	**0.02**	**0.6**	**0.**	18	**0.6**	**0.**	1	**0.6**	**0.**	0.1	**0.6**	0.001
	4	**0.04**	**0.8**	**0.**	34	0.77	0.	4	**0.8**	**0.**	0.2	**0.8**	**0.**
	5	**0.04**	**1.**	**0.**	52	0.95	**0.**	60	**1.**	**0.**	0.25	**1.**	**0.**
	7	**13**	**1.4**	**0.**	180	1.31	0.03	OOT	OOT	OOT	15	1.39	0.01
	10	**92**	**2.**	**0.**	OOT	OOT	OOT	OOT	OOT	OOT	103	1.99	0.02
	25	**15**	**5.**	**0.**	OOT	OOT	OOT	OOT	OOT	OOT	132	**5.**	0.02
	50	**217**	**10.**	**0.**	OOT	OOT	OOT	OOT	OOT	OOT	718	10.	0.03
RECYCLING	2	**0.4**	**0.26**	**0.**	**0.4**	**0.26**	**0.**	1.2	**0.26**	**0.**	2.8	**0.26**	0.004
	3	**17**	**0.32**	0.01	78	**0.32**	**0.**	120	**0.32**	**0.**	75	**0.32**	0.01
	4	**42**	**0.36**	0.03	56	0.35	**0.01**	OOT	OOT	OOT	139	0.35	0.04
	5	213	**0.4**	**0.05**	**98**	0.38	**0.05**	OOT	OOT	OOT	239	0.38	0.06
	7	**271**	**0.5**	0.06	1481	0.48	**0.04**	OOT	OOT	OOT	624	**0.46**	0.12
	10	**505**	**0.63**	0.07	3232	0.6	**0.06**	OOT	OOT	OOT	1249	**0.63**	0.15
	100	**4180**	4.9	**0.68**	OOT	OOT	OOT	OOT	OOT	OOT	13087	**4.96**	2
ADVERSARIAL TIGER	2	**0.1**	**-0.4**	**0.**	**0.1**	**-0.4**	**0.**	0.2	**-0.4**	**0.**	**0.1**	**-0.4**	**0.0**
	3	**0.2**	**-0.56**	**0.**	0.3	**-0.56**	**0.**	500	**-0.56**	**0.**	5	**-0.56**	**0.**
	4	**16**	**-0.75**	**0.01**	46	-0.76	0.02	OOT	OOT	OOT	76	**-0.75**	0.03
	5	**82**	-0.96	**0.05**	141	-0.96	**0.05**	OOT	OOT	OOT	91	**-0.95**	**0.05**
	7	**248**	-1.36	0.13	866	-1.36	**0.1**	OOT	OOT	OOT	459	**-1.35**	0.13
	10	**824**	-1.97	0.24	956	-1.98	**0.18**	OOT	OOT	OOT	3901	**-1.95**	0.3

evaluations. This generally produces smoother exploitability profiles in the reported experiments (see Figure 2). A practical difference concerns exploitability evaluation. For CFR, exploitability must be estimated through an additional evaluation procedure whose cost increases with horizon. For fairness, these evaluation costs are included in the reported runtimes (see Table 1). Runtimes excluding the measurement cost are reported in Table 4. In CPI, the quantities required to estimate deviations are maintained during policy improvement steps via the fixed-policy value function. So exploitability estimates are obtained directly from the algorithm. For larger horizons, exploitability estimates in CFR also become increasingly noisy, which can make convergence trends more difficult to interpret empirically. CFR+ remains attractive because of its inexpensive iterations, whereas CPI performs fewer but more costly updates. The two approaches therefore reflect different computational trade-offs between update cost and exploitability estimation.

Takehome message. CFR+ benefits from inexpensive iterations, but typically requires more iterations and repeated exploitability evaluations to reach solution quality comparable to CPI. In our experiments, this often resulted in higher overall runtime at larger horizons. CPI instead performs fewer but more expensive updates, leading to a different trade-off between update cost, exploitability estimation, and empirical stability.

6 Conclusion

We introduced an approximate policy-iteration algorithm for finite-horizon zs-POSGs. The method combines elements from point-based value iteration and policy-search approaches. It maintains one policy per player and updates both policies along occupancy trajectories induced by conservative policy updates. Instead of approximating a global central value function, each update relies on a fixed-policy continuation value function associated with the current joint policy. The algorithm combines local policy improvements, conservative updates, and persistent exploration. The empirical results show that this approach is competitive on standard finite-horizon zs-POSG benchmarks. In several instances and horizons, it improves the runtime–exploitability trade-off compared with PBVI, HSVI, and CFR+.

Several limitations remain. First, the algorithm requires solving local optimisation problems for both players throughout training. Second, convergence can remain slow and the sample complexity high, which may limit scalability at larger horizons. Third, the current implementation still relies on linear programming subroutines, which may become computationally demanding in larger games. Future work will investigate iterative approximations of these optimisation steps, improved exploration and scheduling mechanisms, and neural approximations of the continuation value functions.

AI use statement

In this work, we used generative AI tools to assist with translation, provide critical ingredients for proving mathematical claims, assist in the development and refinement of mathematical proofs, edit the manuscript to improve readability, create and modify scientific figures, create and edit software code, summarize and analyze existing literature, and assist with sourcing and searching for relevant information.

We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design research methodology or experiments, provide feedback on research methodology or experiments, implement the methods underlying the research, or interpret the results. The remaining categories of required disclosure are not applicable to this work.

All AI-assisted content and outputs were reviewed by the authors. The authors take full responsibility for the final content of this work, including all text, mathematical claims, proofs, figures, code, and other artifacts produced with the assistance of generative AI.

References

Karl J. Aström. Optimal control of Markov decision processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005. doi: 10.1137/S0363012904439301. URL https://doi.org/10.1137/S0363012904439301.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-dynamic programming.*, volume 3 of *Optimization and neural computation series*. Athena Scientific, 1996. ISBN 1886529108.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Olivier Buffet, Jilles Steeve Dibangoye, Abdallah Safidine, and Vincent Thomas. ϵ-optimally solving zs-POSGs using Bellman's optimality principle. Technical report, Inria Nancy Grand-Est, 2020.

Rafael F. Cunha, Jacopo Castellini, Johan Peralez, and Jilles Steeve Dibangoye. On convex optimal value functions for POSGs. *arXiv preprint arXiv:2311.09459*, 2023.

Aurélien Delage, Olivier Buffet, Jilles Steeve Dibangoye, and Abdallah Saffidine. HSVI can solve zero-sum partially observable stochastic games. *Dynamic Games and Applications*, pp. 1–55, 2023.

Jilles Steeve Dibangoye and Olivier Buffet. Learning to act in decentralized partially observable MDPs. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1233–1242, 2018.

Jilles Steeve Dibangoye, Christopher Amato, and Arnaud Doniec. Scaling up decentralized MDPs through heuristic search. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 217–226, 2012.

Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 90–96, 2013.

Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Exploiting separability in multiagent planning with continuous-state MDPs. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. ACM, 2014.

Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. *Journal of Artificial Intelligence Research*, 55: 443–497, 2016a.

Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. *Journal of Artificial Intelligence Research*, 2016b.

Erwan C. Escudie, Matthia Sabatelli, Olivier Buffet, and Jilles Steeve Dibangoye. ϵ-optimally solving two-player zero-sum posgs. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer, 1997.

Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI)*, 2004.

Karel Horák and Branislav Bošanský. Solving partially observable stochastic games with public observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Karel Horák, Branislav Bošanský, and Michal Pěchouček. Heuristic search value iteration for one-sided partially observable stochastic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Dhruva Kartik and Ashutosh Nayyar. Zero-sum stochastic games with asymmetric information. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 4061–4066, 2019. doi: 10.1109/CDC40024.2019.9030177.

Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and Economic Behavior*, 14(2):247–259, June 1996.

Vojtěch Kovařík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisý. Rethinking formal models of partially observable multiagent decision making. *Artificial Intelligence*, 303:103645, 2022. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103645.

H. W. Kuhn. *Extensive Games and the Problem of Information*. Princeton University Press, Princeton, NJ, 1953.

Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7187–7194, 2020.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael B. Johanson, and Michael H. Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356:508–513, 2017.

Rémi Munos. Error bounds for approximate policy iteration. In Tom Fawcett and Nina Mishra (eds.), *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 560–567. AAAI Press, 2003.

Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013.

T. Neller and Marc Lanctot. An introduction to counterfactual regret minimization. 2013.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

Frans Adriaan Oliehoek. Sufficient plan-time statistics for decentralized POMDPs. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

Johan Peralez, Aurélien Delage, Olivier Buffet, and Jilles Steeve Dibangoye. Solving hierarchical information-sharing dec-POMDPs: An extensive-form game approach. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Johan Peralez, Aurélien Delage, Jacopo Castellini, Rafael F. Cunha, and Jilles Steeve Dibangoye. Optimally solving simultaneous-move dec-POMDPs: The sequential central planning approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):9411–9419, 2025.

Fabian Raoul Pieroth, Katherine Fitch, and Lenz Belzner. Detecting influence structures in multi-agent reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 40740–40761. PMLR, 2024.

Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. Point-based value iteration: An anytime algorithm for POMDPs. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 3, pp. 1025–1032, 2003.

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, 1951. ISSN 0003486X, 19398980.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1889–1897. JMLR.org, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51, 2013.

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

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov decision processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

Trey Smith and Reid G. Simmons. Heuristic search value iteration for POMDPs. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI-2004)*, pp. 520–527. AUAI Press, 2004.

Samuel Sokota, Ryan D'Orazio, Chun Kai Ling, David J. Wu, J. Zico Kolter, and Noam Brown. Abstracting imperfect information away from two-player zero-sum games. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32169–32193, 2023.

Edward J. Sondik. The optimal control of partially observable Markov decision processes over the infinite horizon: Discounted cost. *Operations Research*, 12:282–304, 1978.

Oskari Tammelin. Solving large imperfect information games using CFR+. *CoRR*, 2014.

Elise Van der Pol and Frans A. Oliehoek. Coordinated deep reinforcement learners for traffic light control. In *NeurIPS Workshop on Learning, Inference and Control of Multi-Agent Systems*, 2016.

Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui Li. Presslight: Learning max pressure control to coordinate traffic signals in arterial network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1290–1298, 2019.

Auke J. Wiggers, Frans A. Oliehoek, and Diederik M. Roijers. Structure in the value function of two-player zero-sum games of incomplete information. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI)*, 2016.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.

Martin A. Zinkevich, Michael Bradley Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Neural Information Processing Systems*, 2007.

A PROOFS

Lemma 1. *In a finite-horizon finite model, all behavioural-policy spaces are compact, and all payoff maps are continuous.*

Proof. Each Π_t^i is a finite product of finite-dimensional simplexes. Each Π^i and each $\Pi_{t:T}$ is a finite product of decision-rule spaces. Finite products of compact sets are compact. Every finite-horizon payoff is a finite sum of multilinear functions of behavioural-policy coordinates. Hence every finite-horizon payoff map is continuous. The extreme-value theorem gives attainment for every optimisation over these spaces. □

A.1 CONDITIONAL KERNELS

If a conditioning event has zero probability, $\tau_{i,t}^{\mathsf{c}} \colon \mathcal{C}_t^i \times \Pi_t^i \times \mathcal{A}^{-i} \times \mathcal{Z}^0 \times \mathcal{Z}^{-i} \to \mathcal{C}_{t+1}^i$ is assigned an arbitrary fixed element of $\mathcal{C}_{t+1}^i$. All such terms are multiplied by their zero likelihood.

Lemma 2. *$\tau_{i,t}^{\mathsf{c}}$ is the Bayes update of (S_{t+1}, H_{t+1}^i) given $(Z_{t+1}^0, Z_{t+1}^{-i})$.*

Proof. Fix $(c_t^i, \pi_t^i, a_t^{-i}, z_{t+1}^0, z_{t+1}^{-i})$.

For (s_{t+1}, h_{t+1}^i), define

$$m(s_{t+1}, h_{t+1}^i) \doteq \sum_{s_t, h_t^i, a_t^i, z_{t+1}^i} c_t^i(s_t, h_t^i)\pi_t^i(a_t^i \mid h_t^i) \cdot p(s_{t+1}, z_{t+1}^0, z_{t+1}^i, z_{t+1}^{-i} \mid s_t, a_t^i, a_t^{-i}) \cdot \mathbf{1}\{h_{t+1}^i = (h_t^i, a_t^i, z_{t+1}^i)\}.$$

Let $W \doteq \sum_{s_{t+1}, h_{t+1}^i} m(s_{t+1}, h_{t+1}^i)$.

If $W > 0$, then

$$\tau_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, a_t^{-i}, z_{t+1}^0, z_{t+1}^{-i})(s_{t+1}, h_{t+1}^i) = \frac{m(s_{t+1}, h_{t+1}^i)}{W}.$$

If $W = 0$, every weighted term involving this update is zero. □

Lemma 3. *$w_{i,t}^{\mathsf{c}} \colon \mathcal{C}_t^i \times \Pi_t^i \times \mathcal{A}^{-i} \to \Delta(\mathcal{Z}^0 \times \mathcal{Z}^{-i})$ is a probability law over $\mathcal{Z}^0 \times \mathcal{Z}^{-i}$.*

Proof. Fix $(c_t^i, \pi_t^i, a_t^{-i})$.

With m as in Lemma 2, define

$$w_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, a_t^{-i}, z_{t+1}^0, z_{t+1}^{-i}) \doteq \sum_{s_{t+1}, h_{t+1}^i} m(s_{t+1}, h_{t+1}^i).$$

Then $w_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, a_t^{-i}, z_{t+1}^0, z_{t+1}^{-i}) \geq 0$, and

$$\begin{aligned}
&\sum_{z_{t+1}^0, z_{t+1}^{-i}} w_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, a_t^{-i}, z_{t+1}^0, z_{t+1}^{-i}) \\
&= \sum_{s_t, h_t^i} c_t^i(s_t, h_t^i) \sum_{a_t^i} \pi_t^i(a_t^i \mid h_t^i) \cdot \sum_{s_{t+1}, z_{t+1}^0, z_{t+1}^i, z_{t+1}^{-i}} p(s_{t+1}, z_{t+1}^0, z_{t+1}^i, z_{t+1}^{-i} \mid s_t, a_t^i, a_t^{-i}) \\
&= \sum_{s_t, h_t^i} c_t^i(s_t, h_t^i) \sum_{a_t^i} \pi_t^i(a_t^i \mid h_t^i) = 1.
\end{aligned}$$

□

Lemma 4. *$g_{i,t}^{\mathsf{c}} \colon \mathcal{C}_t^i \times \Pi_t^i \times \mathcal{A}^{-i} \to \mathbb{R}$ is the conditional expectation of $r^i(S_t, A_t^i, a_t^{-i})$.*

Proof. Fix $(c_t^i, \pi_t^i, a_t^{-i})$.

$$\begin{aligned}\mathbb{E}[r^i(S_t, A_t^i, a_t^{-i}) \mid c_t^i, \pi_t^i, a_t^{-i}] &= \sum_{s_t, h_t^i} c_t^i(s_t, h_t^i) \sum_{a_t^i} \pi_t^i(a_t^i \mid h_t^i) r^i(s_t, a_t^i, a_t^{-i}) \\ &= g_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, a_t^{-i}).\end{aligned}$$

□

Lemma 5. *$p_{i,t}^{\mathsf{c}} \colon \mathcal{C}_t^i \times \Pi_t^i \times \mathcal{A}^{-i} \to \Delta(\mathcal{C}_{t+1}^i)$ is the pushforward of $w_{i,t}^{\mathsf{c}}$ by $\tau_{i,t}^{\mathsf{c}}$.*

Proof. Fix $(c_t^i, \pi_t^i, a_t^{-i})$. For every $c_{t+1}^i \in \mathcal{C}_{t+1}^i$,

$$\begin{aligned}&\mathbb{P}(C_{t+1}^i = c_{t+1}^i \mid c_t^i, \pi_t^i, a_t^{-i}) \\ &= \sum_{z_{t+1}^0, z_{t+1}^{-i}} w_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, a_t^{-i}, z_{t+1}^0, z_{t+1}^{-i}) \cdot \mathbf{1}\{c_{t+1}^i = \tau_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, a_t^{-i}, z_{t+1}^0, z_{t+1}^{-i})\} \\ &= p_{i,t}^{\mathsf{c}}(c_{t+1}^i \mid c_t^i, \pi_t^i, a_t^{-i}).\end{aligned}$$

□

For fixed $\pi_{t:T}^i$, the recursion below defines $v_{i,k}^{\pi_{t:T}^i} \colon \mathcal{C}_k^i \to \mathbb{R}$ on $\mathcal{C}_k^i$ for every $k \in \{t, \dots, T\}$. Define

$$v_{i,k}^{\pi_{t:T}^i} \colon c_k^i \mapsto \min_{\pi_{t:T}^{-i}} \mathbb{E}_{c_k^i, \pi_{t:T}^i}\left[\textstyle\sum_{\kappa=k}^{T-1} r^i(S_\kappa, A_\kappa^i, A_\kappa^{-i})\right].$$

Theorem 3. *For fixed $\pi_{t:T}^i$, $v_{i,k}^{\pi_{t:T}^i}$ satisfies the conditional Bellman recursion for every $k \geq t$.*

Proof. At $k = T$, set $v_{i,T}^{\pi_{t:T}^i} \equiv 0$.

Fix $k < T$, $c_k^i \in \mathcal{C}_k^i$, and the rule π_k^i induced by $\pi_{t:T}^i$.

For every $a_k^{-i} \in \mathcal{A}^{-i}$, Lemmas 4 and 5 give

$$\begin{aligned}&\mathbb{E}[r^i(S_k, A_k^i, a_k^{-i}) + v_{i,k+1}^{\pi_{t:T}^i}(C_{k+1}^i) \mid c_k^i, \pi_k^i, a_k^{-i}] \\ &= g_{i,k}^{\mathsf{c}}(c_k^i, \pi_k^i, a_k^{-i}) + \sum_{c_{k+1}^i \in \mathsf{supp}(p_{i,k}^{\mathsf{c}}(\cdot \mid c_k^i, \pi_k^i, a_k^{-i}))} p_{i,k}^{\mathsf{c}}(c_{k+1}^i \mid c_k^i, \pi_k^i, a_k^{-i}) v_{i,k+1}^{\pi_{t:T}^i}(c_{k+1}^i).\end{aligned}$$

Taking $\min_{a_k^{-i} \in \mathcal{A}^{-i}}$ gives the recursion. □

A.2 Belief dynamics

Define $g_t^i \colon (b_t^i, \pi_t) \mapsto \mathbb{E}_{b_t^i, \pi_t}[r^i(S_t, A_t)]$.

Lemma 6. *g_t^i is the expectation of $g_{i,t}^{\mathsf{c}}$ under b.*

Proof. Fix $b \in \mathcal{B}_t^i$ and $\pi_t = (\pi_t^i, \pi_t^{-i})$.

By definition,

$$\begin{aligned}g_t^i(b, \pi_t) &\doteq \mathbb{E}_{b_t^i, \pi_t}[r^i(S_t, A_t)] \\ &= \mathbb{E}_{b_t^i, \pi_t}[\mathbb{E}_{C_t^i, \pi_t^i, \pi_t^{-i}(C_t^i)}[r^i(S_t, A_t)]] \\ &= \sum_{c_t^i \in \mathcal{C}_t^i} b(c_t^i) g_{i,t}^{\mathsf{c}}(c_t^i, \pi_t^i, \pi_t^{-i}(c_t^i)).\end{aligned}$$

□

Lemma 7. *τ_t^i is the law of C_{t+1}^i under (b, π_t).*

Proof. Fix $b \in \mathcal{B}_t^i$ and $\pi_t = (\pi_t^i, \pi_t^{-i})$.

For every $c_{t+1}^i \in \mathcal{C}_{t+1}^i$,

$$\begin{aligned}&\mathbb{P}_{b,\pi_t}(C_{t+1}^i = c_{t+1}^i)\\ &= \sum_{c_t^i \in \mathcal{C}_t^i} b(c_t^i) p_{i,t}^{\mathsf{c}}(c_{t+1}^i \mid c_t^i, \pi_t^i, \pi_t^{-i}(c_t^i))\\ &= \tau_t^i(b, \pi_t)(c_{t+1}^i).\end{aligned}$$

□

Lemma 8. *For separable payoffs, the minimisation over π_t^{-i} decomposes pointwise.*

Proof. Let $F \colon \mathcal{C}_t^i \times \mathcal{A}^{-i} \to \mathbb{R}$.

For every π_t^{-i},

$$\sum_{c_t^i} b(c_t^i) F(c_t^i, \pi_t^{-i}(c_t^i)) \geq \sum_{c_t^i} b(c_t^i) \min_{a_t^{-i}} F(c_t^i, a_t^{-i}).$$

Choose

$$\pi_{t,*}^{-i}(c_t^i) \in \arg\min_{a_t^{-i} \in \mathcal{A}^{-i}} F(c_t^i, a_t^{-i}).$$

Then

$$\sum_{c_t^i} b(c_t^i) F(c_t^i, \pi_{t,*}^{-i}(c_t^i)) = \sum_{c_t^i} b(c_t^i) \min_{a_t^{-i}} F(c_t^i, a_t^{-i}).$$

□

A.3 Continuation values

Define

$$v_{i,t}^{\mathrm{cont}}(b) \doteq \max_{\pi_{t:T}^i} \min_{\pi_{t:T}^{-i}} \mathbb{E}_{b,\pi_{t:T}^i,\pi_{t:T}^{-i}} \left[\sum_{k=t}^{T-1} r^i(S_k, A_k) \right].$$

Lemma 9. *$v_{i,t}^{\mathrm{cont}}$ satisfies the focal belief Bellman recursion.*

Proof. At $t = T$, $v_{i,T}^{\mathrm{cont}} = 0$.

Fix $t < T$. Rearranging $\max$ and $\min$ operators require replacing opponent deterministic policy suffix π_t^{-i} by a mixed one. For any (π_t^i, π_t^{-i}), Lemmas 6 and 7 give

$$\begin{aligned}&\mathbb{E}_{b,\pi_t,\pi_{t+1:T}^i,\pi_{t+1:T}^{-i}} \left[\sum_{k=t}^{T-1} r^i(S_k, A_k) \right]\\ &= g_t^i(b, \pi_t) + \mathbb{E}_{\tau_t^i(b,\pi_t),\pi_{t+1:T}^i,\pi_{t+1:T}^{-i}} \left[\sum_{k=t+1}^{T-1} r^i(S_k, A_k) \right].\end{aligned}$$

Decomposing the policy space and optimising gives

$$v_{i,t}^{\mathrm{cont}}(b) = \max_{\pi_t^i} \min_{\pi_t^{-i}} \left[g_t^i(b, \pi_t) + v_{i,t+1}^{\mathrm{cont}}(\tau_t^i(b, \pi_t)) \right].$$

where the interchange of $\max$ and $\min$ is established using arguments similar to those in Delage et al. (2023) and Escudie et al. (2025). □

Theorem 4. *The optimal focal belief value function $\underline{v}_t^{i,*}$ satisfies the focal belief Bellman equation and equals $v_{i,t}^{\mathrm{cont}}$.*

Proof. Both functions have terminal value zero.

Assume $\underline{v}^{i,*}_{t+1} = v^{\mathrm{cont}}_{i,t+1}$. By Lemma 9,

$$v^{\mathrm{cont}}_{i,t}(b) = \max_{\pi^i_t} \min_{\pi^{-i}_t} \left[g^i_t(b, \pi_t) + \underline{v}^{i,*}_{t+1}(\tau^i_t(b, \pi_t)) \right].$$

The right-hand side is the Bellman definition of $\underline{v}^{i,*}_t(b)$. Backward induction gives the claim. □

A.4 Focal decomposition

Lemma 10. *A prefix pair $(\pi^i_{0:t}, \pi^{-i}_{0:t})$ induces $b^i_{k+1} = \tau^i_k(b^i_k, \pi_k)$.*

Proof. At $k = 0$, b^i_0 is induced by b_0. If b^i_k is the law of C^i_k, then Lemma 7 gives

$$b^i_{k+1} = \tau^i_k(b^i_k, \pi_k).$$

□

A.5 Proof of Theorem 1

For fixed $\pi^i_{t:T}$, define

$$\underline{v}^{\pi^i_{t:T}}_{i,k}(b) \doteq \min_{\pi^{-i}_{k:T}} \mathbb{E}_{b, \pi^i_{k:T}, \pi^{-i}_{k:T}} \left[\sum_{\ell=k}^{T-1} r^i(S_\ell, A_\ell) \right],$$

where $\pi^i_{k:T}$ is the tail of $\pi^i_{t:T}$.

Lemma 11. *For fixed $\pi^i_{t:T}$ and $k \geq t$, $\underline{v}^{\pi^i_{t:T}}_{i,k}$ is linear on $\mathcal{B}^i_k$.*

Proof. For $c \in \mathcal{C}^i_k$, set

$$w^{\pi^i_{t:T}}_{i,k}(c) \doteq v^{\pi^i_{t:T}}_{i,k}(c).$$

We prove by backward induction that

$$\underline{v}^{\pi^i_{t:T}}_{i,k}(b) = \sum_{c \in \mathcal{C}^i_k} b(c) w^{\pi^i_{t:T}}_{i,k}(c).$$

The claim is zero at $k = T$.

Assume it holds at $k+1$. The stage-k focal rule π^i_k is fixed. By Lemmas 6, 7, and the induction hypothesis,

$$\underline{v}^{\pi^i_{t:T}}_{i,k}(b) = \min_{\pi^{-i}_k} \sum_{c \in \mathcal{C}^i_k} b(c) \Bigg[g^{\mathsf{c}}_{i,k}(c, \pi^i_k, \pi^{-i}_k(c)) + \sum_{c' \in \mathcal{C}^i_{k+1}} p^{\mathsf{c}}_{i,k}(c' \mid c, \pi^i_k, \pi^{-i}_k(c)) w^{\pi^i_{t:T}}_{i,k+1}(c') \Bigg].$$

By Lemma 8,

$$\underline{v}^{\pi^i_{t:T}}_{i,k}(b) = \sum_{c \in \mathcal{C}^i_k} b(c) \min_{a^{-i}_k} \Bigg[g^{\mathsf{c}}_{i,k}(c, \pi^i_k, a^{-i}_k) + \sum_{c' \in \mathcal{C}^i_{k+1}} p^{\mathsf{c}}_{i,k}(c' \mid c, \pi^i_k, a^{-i}_k) w^{\pi^i_{t:T}}_{i,k+1}(c') \Bigg].$$

By Theorem 3, the inner minimum equals $w^{\pi^i_{t:T}}_{i,k}(c)$. Thus

$$\underline{v}^{\pi^i_{t:T}}_{i,k}(b) = \sum_{c \in \mathcal{C}^i_k} b(c) w^{\pi^i_{t:T}}_{i,k}(c).$$

Linearity follows. □

Theorem 5. *$\underline{v}_t^{i,*}$ is convex, and $\underline{v}_t^{i,*}(b) = \max_{\underline{v}\in\Gamma_{i,t}} \underline{v}(b)$ for all $b \in \mathcal{B}_t^i$.*

Proof. Let

$$\Gamma_{i,t} \doteq \{\underline{v}_{i,t}^{\pi_{t:T}^i} : \pi_{t:T}^i \in \Pi_{t:T}\}.$$

By Lemma 1, $\Gamma_{i,t}$ is compact in the finite-dimensional space of linear functionals.

By Theorem 4,

$$\underline{v}_t^{i,*}(b) = v_{i,t}^{\mathrm{cont}}(b).$$

By definition of $v_{i,t}^{\mathrm{cont}}$,

$$v_{i,t}^{\mathrm{cont}}(b) = \max_{\pi_{t:T}^i\in\Pi_{t:T}} \underline{v}_{i,t}^{\pi_{t:T}^i}(b).$$

Hence

$$\underline{v}_t^{i,*}(b) = \max_{\underline{v}\in\Gamma_{i,t}} \underline{v}(b).$$

For $b, b' \in \mathcal{B}_t^i$ and $\lambda \in [0,1]$,

$$\begin{aligned}\underline{v}_t^{i,*}(\lambda b + (1-\lambda)b') &= \max_{\underline{v}\in\Gamma_{i,t}} \underline{v}(\lambda b + (1-\lambda)b') \\ &= \max_{\underline{v}\in\Gamma_{i,t}} [\lambda \underline{v}(b) + (1-\lambda)\underline{v}(b')] \\ &\le \lambda \underline{v}_t^{i,*}(b) + (1-\lambda)\underline{v}_t^{i,*}(b').\end{aligned}$$

Thus $\underline{v}_t^{i,*}$ is convex. □

A.6 Local Bellman backups

The theorem is conditional on finite coverage, Bellman closure, globally valid inserted vectors, infinite exact updates, and covered-backup consistency.

A.6.1 Local point-based backup

We introduce the following linear program to compute a greedy decision rule with respect to a fixed continuation policy $\pi_{t+1:T}^i$ and focal belief b_t^i.

The program uses the following variables:

- $\mathbf{v}(\mathbf{c_t^i})$, representing the value associated with conditional state c_t^i,
- $\eta(\mathbf{h_{t+1}^{-i}}, \mathbf{h_{t+1}^0})$, representing auxiliary continuation values,
- $\pi_\mathbf{t}^\mathbf{i} \in \mathbf{\Pi_t^i}$, the focal decision rule to optimize.

Theorem 6. *For any fixed suffix policy $\pi_{t+1:T}^i$ (associated with its value function $\Lambda_{i,t+1}$), and focal belief $b_t^i \in \mathcal{B}_t^i$, solving the linear program*

$$\begin{aligned}&\textit{Max} \quad \textstyle\sum_{c_t^i} b_t^i(c_t^i)\, \mathbf{v}(\mathbf{c_t^i}) \\ &\textit{Subject to} \quad \mathbf{v}(\mathbf{c_t^i}) \le \sum_{z_t^{-i}, z_t^0} \eta(\mathbf{h_{t+1}^{-i}}, \mathbf{h_{t+1}^0}), \quad \forall c_t^i, a_t^{-i} \\ &\qquad \eta(\mathbf{h_{t+1}^{-i}}, \mathbf{h_{t+1}^0}) \le \sum_{h_t^i, a_t^i} \pi_\mathbf{t}^\mathbf{i}(\mathbf{a_t^i} \mid \mathbf{h_t^i}, \mathbf{h_t^0})\, Q_{(h_{t+1}^{-i}, h_{t+1}^i, \alpha)}, \\ &\qquad\qquad\qquad\qquad \forall h_{t+1}^{-i}, h_{t+1}^0, \alpha \in \Lambda_{i,t+1}, \\ &\qquad \pi_\mathbf{t}^\mathbf{i} \in \Pi_t^i\end{aligned}$$

where

$$(h_{t+1}^{-i}, h_{t+1}^0) = (c_t^i, a_t^{-i}, z_t^{-i}, z_t^0),$$

yields a decision rule π_t^i that greedily maximizes

$$\underline{q}_t^{\pi_{t+1:T}^i}(b_t^i, \pi_t^i) = \min_{\pi_t^{-i}} g_t^i(b_t^i, \pi_t^i, \pi_t^{-i}) + \underline{v}_{i,t+1}^{\pi_{t+1:T}^i}\big(\tau_t^i(b_t^i, \pi_t^i, \pi_t^{-i})\big),$$

corresponding to the Q-value induced by the fixed-policy continuation $(\pi_t^i, \pi_{t+1:T}^i)$.

Proof. By construction, $\underline{q}_t^{\pi^i_{t+1:T}}$ is consistent with $\underline{v}_{i,t}^{\pi^i_{t+1:T}}$, and an optimal π_t^{-i} can be taken deterministic (since the suffix policy $\pi^i_{t+1:T}$ is fixed).

A greedy decision rule maximizes $\underline{q}_t^{\pi^i_{t+1:T}}$ over reachable beliefs:

$$\begin{aligned}\max_{\pi_t^i} \underline{q}_t^{\pi^i_{t+1:T}}(b_t^i, \pi_t^i) &= \max_{\pi_t^i} \min_{\pi_t^{-i}} \left[g_t^i(b_t^i, \pi_t^i, \pi_t^{-i}) + \underline{v}_{i,t}^{\pi^i_{t+1:T}}\left(\tau_t^i(b_t^i, \pi_t^i, \pi_t^{-i})\right) \right] \\ &= \max_{\pi_t^i} \min_{\pi_t^{-i}} \left[g_t^i(b_t^i, \pi_t^i, \pi_t^{-i}) + \sum_{c_{t+1}^i} b_{t+1}^i(c_{t+1}^i) \min_{\alpha^{\pi^i_{t+1:T}}} \alpha^{\pi^i_{t+1:T}} \cdot c_{t+1}^i \right],\end{aligned}$$

where $b_{t+1}^i = \tau_t^i(b_t^i, \pi_t^i, \pi_t^{-i})$.

Expanding the belief update and rearranging terms, we obtain:

$$\begin{aligned}\max_{\pi_t^i} \underline{q}_t^{\pi^i_{t+1:T}}(b_t^i, \pi_t^i) = \max_{\pi_t^i} \min_{\pi_t^{-i}} & \sum_{h_t^{-i}, a_t^{-i}, z_t^{-i}, z_t^0} \pi_t^{-i}(a_t^{-i} \mid h_t^{-i}, h_t^0) \\ & \cdot \min_{\alpha^{\pi^i_{t+1:T}}} \sum_{h_t^i, a_t^i} \pi_t^i(a_t^i \mid h_t^i, h_t^0) \sum_{s_t} b_t^i(s_t, h_t) \\ & \cdot \left[g_t^i(s_t, a_t^i, a_t^{-i}) + \sum_{s_{t+1}, z_t} p(s_{t+1}, z_t, z_t^0 \mid s_t, a_t) \alpha^{\pi^i_{t+1:T}}(c_{t+1}^i) \right].\end{aligned}$$

This expression makes explicit the max–min structure together with the inner minimization over α, where α is taken after the expectation over next-stage variables. This placement is crucial: it preserves the piecewise-linear concave structure of the value function and yields directly the desired linear programming formulation.

If we denote $(h_{t+1}^{-i}, h_{t+1}^0) = (c_t^i, a_t^{-i}, z_t^{-i}, z_t^0)$, using the bijection between c_t^i and the pair of histories (h_t^{-i}, h_t^0) , we can define

$$\begin{aligned}& Q_{(h_{t+1}^{-i}, h_{t+1}^i, \alpha)} \\ &= \sum_{s_t} b_t^i(s_t, h_t) \left[g_t^i(s_t, a_t^i, a_t^{-i}) + \sum_{s_{t+1}, z_t} p(s_{t+1}, z_t, z_t^0 \mid s_t, a_t)\, \alpha^{\pi^i_{t+1:T}}(c_{t+1}^i) \right],\end{aligned}$$

then the above quantity corresponds to the q-value of the best response where player $-i$ plays a_t^{-i} in c_t^i and follows $\alpha^{\pi^i_{t+1:T}}$ at (h_{t+1}^{-i}, h_t^0). This yields the coefficients of the linear program described in the theorem.

We also impose an additional constraint on π_t^i to ensure that it is a valid decision rule, i.e., that each local strategy belongs to the simplex:

$$\pi_t^i(\cdot \mid h_t^i, h_t^0) \in \Delta(\mathcal{A}^i), \quad \forall h_t^i, h_t^0$$

ensuring that π_t^i defines a proper probability distribution over actions at every local history. □

A.7 Comparative Analysis of Backup Complexity

Comparative analysis: Delage et al. (2023) The OMG greedy-update formulation Delage et al. (2023) enforces pessimism against *all* deterministic opponent decision rules $\pi_t^2 : H_t^2 \to A_t^2$. Because there are $|A_t^2|^{|H_t^2|}$ such mappings, the LP description contains $O(|A_t^2|^{|H_t^2|})$ constraints (Table 2), which is exponential in the size of the opponent's private-history space. Our backup removes this bottleneck by never indexing constraints by opponent decision rules: our LP constrains only against finitely many vectors in a cached envelope Λ. Hence our LP admits polynomial-size descriptions in the explicit parameters $|H_t^1|, |H_t^2|, |A_t^1|, |A_t^2|, |Z_t^2|, |\Lambda|$.

Update step	# variables	# constraints
Ours: LP	$\lvert H_t^1\rvert\lvert A_t^1\rvert + \lvert H_t^2\rvert\lvert A_t^2\rvert\lvert Z_t^2\rvert + \lvert H_t^2\rvert$	$O(\lvert H_t^2\rvert\lvert A_t^2\rvert\lvert Z_t^2\rvert\lvert\Lambda\rvert)$
(Delage et al., 2023)	(rule variables + epigraph)	**Exponential:** $O\big(\lvert A_t^2\rvert^{\lvert H_t^2\rvert}\big)$
(Escudie et al., 2025)	$\lvert H_t^1\rvert\lvert A_t^1\rvert\lvert\Gamma_1\rvert + \lvert H_t^2\rvert\lvert A_t^2\rvert\lvert Z_t^2\rvert + \lvert H_t^2\rvert$	$O\big(\lvert\Gamma_1\rvert \cdot \lvert\Gamma_2^*\rvert \cdot \lvert H_t^2\rvert \cdot \lvert A_t^2\rvert \cdot \lvert Z_t^2\rvert\big)$

Table 2: LP complexity comparison (input-size view). The OMG greedy-update LP has constraints indexed by opponent decision rules, yielding an exponential dependence on $|H_t^2|$ (Delage et al., 2023). Both the cached-envelope greedy LP (Escudie et al., 2025) and our backup scale polynomially in the same explicit representation objects, but expose different optimization structures. In our formulation, both the number of variables and constraints are reduced by a factor of $|\Gamma_1|$. Here, $|\Gamma_1|$ denotes the number of piecewise-linear concave functions maintained, while $|\Lambda|$ is the number of piecewise-linear vectors associated with each element of Γ_1.

Comparative analysis: Escudie et al. (2025) In (Escudie et al., 2025), the LP explicitly incorporates all candidate next policies of i through the sets Γ_1 of piecewise linear concave function Λ, which grow over iterations. Specifically, a single monolithic greedy LP jointly couples the outer cache index Γ_1, the inner value index $\alpha \in \Lambda$, and the opponent action-observation indices (a_t^2, z_t^2). This full coupling within a single program leads to a constraint count of $O\big(|\Gamma_1|\cdot|\Lambda|\cdot|H_t^2|\cdot|A_t^2|\cdot|Z_t^2|\big)$, as reported in Table 2.

Although the reduction in $\mid \Gamma_1 \mid$ has little impact for short horizons, it becomes critical for longer horizons. Indeed, in PBVI (Escudie et al., 2025), the size of Γ_1 grows with the number of sampled points, which itself increases with the number of stages required to cover an ever-expanding continuous state space.

B ALGORITHM

Value iteration in (Escudie et al., 2025) amounts to solving a focal belief game using sampled opponent behaviors. However, this unguided sampling process induces an exponential growth in the number of sampled points, resulting in a rapidly increasing computational complexity. To address this limitation, we introduce a policy iteration approach that exploits the structural properties of the game together with the reduced formulations defined previously. Leveraging this structure yields a more efficient iterative procedure while preserving optimality guarantees.

We prove that the dynamics induced by the game can be viewed as a discretization of a continuous-time dynamical system. Along the resulting continuous trajectory, the dual gap is non-increasing and converges to zero. The proof relies on three key properties: the convexity of the dual gap in the policy space, the differentiability of the value function almost everywhere, and the uniform continuity of the value function. Together, these properties allow us to characterize the evolution of the dual gap along the continuous dynamics and establish its convergence.

This section introduces and formally defines the main objects used throughout the analysis, including the dual gap and the dynamics guiding the update F_t^i, and related quantities.

The continuous dynamics describe the infinitesimal policy updates underlying the algorithm. At each stage t, the update acts on the policy π_t^i while the policies at other stages are kept fixed. The resulting direction therefore belongs to the local policy space Π_t^i, while depending on the complete policy of player i and on the opponent's policies up to stage $t-1$ through the induced belief state. This defines a vector field over the policy space, whose integral curves describe the continuous version of the algorithm.

Definition 4 (Continuous Dynamics)**.** For player i and stage t, the continuous dynamics are defined as the vector field

$$F_t^i : \Pi_{0:T-1}^i \times \Pi_{0:t-1}^{-i} \to \Pi_t^i,$$

given by

$$F_t^i\left(\pi_{0:T-1}^i, \pi_{0:t-1}^{-i}\right) = \nabla_{\pi_t^i} \underline{q}^{\pi_{t+1:T}^i}\left(b_t^i, \pi_t^i\right),$$

where b_t^i is the focal belief induced by $\pi_{0:t-1}^i$ and $\pi_{0:t-1}^{-i}$ and $\underline{q}^{\pi_{t+1:T-1}^i}$ is defined in Theorem 6. Thus, F_t^i gives the safest most improving partial direction of the pessimistic Q-value with respect to π_t^i, while keeping the belief fixed.

We may ask whether having the agents follow their respective induced flows improves their value $\underline{v}_t^i$. The direction induced by F_t^i is safely improving for timestep t to T, as all future best-responses are considered in $\underline{q}^{\pi_{t+1:T-1}^i}$. However, it remains unclear whether this improvement is preserved globally across all stages. Indeed, F_t^i is defined under the assumption that the opponent's past policy remains fixed, as reflected by its dependence on $\pi_{0:t-1}^{-i}$. When the policies are updated simultaneously across iterations, this assumption may no longer hold: changes to the opponent's policy at earlier stages can modify the induced belief b_t^i and, consequently, the continuation value at later stages. This creates a situation similar to fictitious play, where both players update their policies simultaneously and each player's update can therefore affect the optimization problem faced by the other.

Definition 5 (Dual gap)**.** Given a joint policy $\pi_{t:T-1}$ and the belief state b_t^i induced by the joint policy up to time $t-1$, the dual gap at time t is defined as

$$d_t^{\pi_{t:T-1}}\left(b_t^i\right) = \underline{q}_t^{\pi_{t+1:T-1}^{-i}}\left(b_t^i, \pi_t^{-i}\right) - \underline{q}_t^{\pi_{t+1:T-1}^i}\left(b_t^i, \pi_t^i\right).$$

Lemma 12 (Local stability of the best response)**.** *At any regular policy profile, the best response of the opponent in the constrained subgame rooted at the current policy is locally constant. Consequently, the pessimistic Q-value*

$$\underline{q}_t^i\left(b_t^i, \pi_t^i\right) \doteq \underline{q}_t^{i,\pi_{t+1:T-1}^i}\left(b_t^i, \pi_t^i\right)$$

defined in Theorem 6 is differentiable with respect to π_t^i in a sufficiently small neighborhood of the current policy.

Proof. At a regular policy profile, the opponent's best response is unique, since the game is concave-convex and admits a finite set of deterministic policies. Consequently, there exists a neighborhood of the current policy in which this best response remains unchanged. Within this neighborhood, the pessimistic continuation value is therefore evaluated against a fixed best response. The remaining terms in the Bellman recursion are differentiable with respect to π_t^i. Hence,

$$\underline{q}_t^i\left(b_t^i, \pi_t^i\right)$$

is differentiable with respect to π_t^i. □

The Q-value function is not differentiable everywhere, but this issue is not specific to our setting: the same non-differentiability arises for the dual gap in normal-form games. Since our method uses the same type of local improvement mechanism in policy space, it can handle these points in the same way. Moreover, our discretization does not encounter pathological cases in practice, and the value function explicitly accounts for possible deviations from the current policy. This provides additional protection against numerical instabilities caused by nondifferentiable points and ensures that the improvement step remains robust to such cases.

Lemma 13 (Local improving direction)**.** *At a regular policy profile, the direction F_t^i is an improving direction for player i's pessimistic value at stage t. In particular,*

$$\frac{d}{d\alpha}\underline{q}_t^i\left(b_t^i, \pi_t^i + \alpha F_t^i\right)\Big|_{\alpha=0} > 0$$

whenever π_t^i is not locally optimal.

Proof. By Lemma 12, the local pessimistic Q-value is differentiable around the current policy. By construction of F_t^i,

$$F_t^i = \nabla_{\pi_t^i} \underline{q}_t^i \left(b_t^i, \pi_t^i\right).$$

Therefore,

$$\frac{d}{d\alpha} \underline{q}_t^i \left(b_t^i, \pi_t^i + \alpha F_t^i\right)\Big|_{\alpha=0} = \left\|F_t^i\right\|^2 > 0$$

whenever $F_t^i \neq 0$. □

Lemma 14 (Safe improvement for $T = 2$)**.** *Suppose that $T = 2$. For sufficiently small step sizes, the update of player i following the dynamics F_t^i strictly improves her pessimistic value whenever the current policy is not locally optimal.*

Proof. Consider the joint policy $(\pi_{0:1}^{1,k}, \pi_{0:1}^{2,k})$ at iteration k. The decision rule $\pi_0^{i,k}$ is optimal for the pessimistic continuation problem induced by $\pi_1^{i,k}$.

We first update the last-stage decision rule according to the above dynamics. Since there is no subsequent decision rule, the associated one-stage pessimistic problem is solved directly. Thus, considering any step size $\alpha_1^k > 0$, the update

$$\pi_1^{i,k+1} = \pi_1^{i,k} + \alpha_1^k F_1^i$$

improves the corresponding local pessimistic value, and hence decreases the corresponding local dual gap.

We then update the first-stage decision rule while keeping the newly obtained $\pi_1^{i,k+1}$ fixed. Define

$$v^k = \underline{v}_0^{i,[\pi_0^{i,k}, \pi_1^{i,k}]},$$

$$\widetilde{v}^k = \underline{v}_0^{i,[\pi_0^{i,k}, \pi_1^{i,k+1}]},$$

and

$$v^{k+1} = \underline{v}_0^{i,[\pi_0^{i,k+1}, \pi_1^{i,k+1}]}.$$

By construction of the first-stage update, F_0^i is an improving direction for the local function. Therefore, the newly obtained policy $\pi_0^{i,k+1}$ strictly improve the overall pessimistic value

$$\pi_0^i \mapsto \underline{v}_0^{i,[\pi_0^i, \pi_1^{i,k+1}]},$$

and thus,

$$v^{k+1} > \widetilde{v}^k.$$

It remains to ensure that the improvement at stage $t = 0$ is not cancelled by the preceding update at stage $t = 1$. To this end, we regard the stage $t = 1$ update as a homotopy

$$\pi_1^{i,k+1}(\alpha) = \pi_1^{i,k} + \alpha F_1^i, \qquad \alpha \in [0, \bar{\alpha}].$$

Let

$$\Delta(\alpha) = \underline{v}_0^{i,[\pi_0^{i,k+1}(\alpha), \pi_1^{i,k+1}(\alpha)]} - \underline{v}_0^{i,[\pi_0^{i,k}, \pi_1^{i,k}]},$$

where $\pi_0^{i,k+1}(\alpha)$ denotes the stage $t = 0$ update computed from the continuation policy $\pi_1^{i,k+1}(\alpha)$.

At $\alpha = 0$, the continuation policy coincides with the policy at iteration k. Since $\pi_0^{i,k+1}$ is an improving direction for the corresponding local problem, we have $\Delta(0) > 0$ at every non-degenerate point where the stage 0 direction is strictly improving. By continuity of the pessimistic value in the policy and, locally, of the stage 0 update, there exists $\bar{\alpha} > 0$ such that

$$\Delta(\alpha) > 0 \qquad \text{for all } 0 < \alpha < \bar{\alpha}.$$

Hence, the stage 1 step can be chosen sufficiently small so that the improvement obtained at stage 0 is not cancelled by the preceding continuation update. Consequently,

$$v^{k+1} > v^k.$$

□

Lemma 15. *For a horizon T, if the stage wise updates from stage 1 to stage $T-1$ can be chosen to strictly improve the pessimistic value, then the update at stage $t = 0$ can also be chosen so that the complete iteration strictly improves the pessimistic value.*

Proof. Assume that the result holds for a horizon T. Consider a horizon $T + 1$. Starting from the last stage, we apply the construction from Lemma 14 to stages

$$1, \dots, T-1.$$

By the induction hypothesis, the continuation problem from stage 1 onwards can be updated with sufficiently small step sizes so that its pessimistic value is improved.

We then consider stage $t = 0$ while keeping the resulting continuation policies fixed. The same local argument applies: the continuation pessimistic value induces a differentiable local Q-value around the current stage $t = 0$ policy, and the update follows an improving direction of this function.

Finally, the step sizes at stages $1, \dots, T-1$ can be chosen sufficiently small so that the improvement produced at stage 0 dominates the change induced by the continuation updates. By continuity, there therefore exist positive step sizes

$$\alpha_0^k, \dots, \alpha_T^k$$

for which the complete iteration yields a strict improvement of the pessimistic value. □

Lemma 16. *For any starting joint policy π, there exists a discretized trajectory*

$$\phi = (\pi^{1,k}, \pi^{2,k})_{k \geq 0}$$

starting from any $(\pi^{1,0}, \pi^{2,0})$ and following the flow induced by $(F^i_{0:T-1}, F^{-i}_{0:T-1})$ such that the dual gap strictly decreases at every non-degenerate iteration.

Proof. We construct the trajectory iteratively. At iteration k, the policy at each stage is updated according to

$$\pi_t^{i,k+1} = \pi_t^{i,k} + \alpha_t^k F_t^i \left(\pi_{0:T-1}^{i,k}, \pi_{0:T-1}^{-i,k} \right),$$

with $\alpha_t^k > 0$.

By Lemma 15, whenever an improvement at an earlier stage could be affected by the updates of the continuation stages, the corresponding continuation step sizes can be chosen sufficiently small so that this improvement is preserved.

Consequently, at every non-degenerate iteration, the step sizes

$$\alpha_0^k, \dots, \alpha_{T-1}^k$$

can be reduced arbitrarily while remaining strictly positive and can be chosen so that the complete iteration strictly improves the pessimistic value. Equivalently, the associated dual gap satisfies

$$d_0(b_0, (\pi^{1,k+1}, \pi^{2,k+1})) < d_0(b_0, (\pi^{1,k}, \pi^{2,k})).$$

Starting from any π, choosing such step sizes at each iteration defines a discretized trajectory ϕ following the induced flow. Along this trajectory, the dual gap therefore strictly decreases at every non-degenerate iteration. □

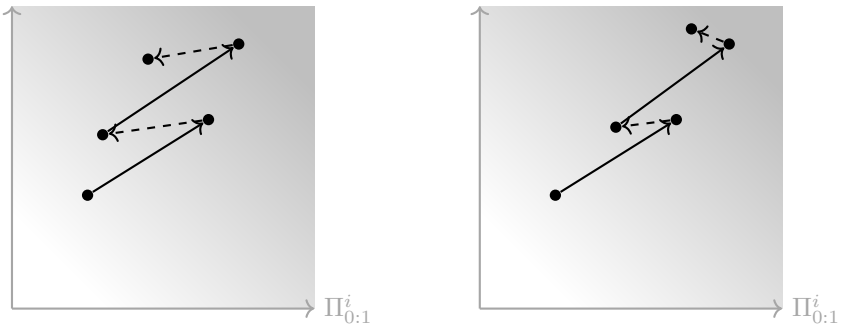

Figure 3: **Example of a situation at** $T = 2$. The background represents the gradient of the dual gap at $t = 0$. *The black arrows* represent updates of the policy where only π^i_0 is changing. At $t = 0$, since i has an optimal view of the pessimistic value function, the move strictly improves the global dual gap. *The dashed arrows* represent updates at $t = 1$; since the update is also influenced by $-i$ past, nothing guarantees that the update will follow the global dual gap d_0 gradient direction. Here, we present an extreme case where the update at $t = 1$ almost cancels the one at $t = 0$. **On the left**, we do not adapt the step sizes, **on the right**, we apply the proof such that destructive updates are avoided and we are sure the updates go to the right direction.

We have established the existence of a trajectory following the flow induced by F^i_t. Since the step size can be chosen arbitrarily small, we can construct a sequence of discrete trajectories with vanishing step sizes that increasingly closely approximates the continuous flow. Along this trajectory, the dual gap d_0 is non-increasing. It remains to show that the trajectory converges to the set of Nash equilibria.

Lemma 17. *Suppose that, at a joint policy π, a safe local improving direction exists for the overall objective. Then there exists a player i and a stage t such that the dynamics F^i_t identify a strictly improving direction at the corresponding induced belief state b^i_t.*

Proof. We prove the result by contradiction. Consider first the case $T = 2$. Suppose that a safe improving direction exists (considering all stages at once), but that no improving direction can be identified inductively at either stage.

In particular, suppose that no improving direction can be identified at $t = 0$. Then the improvement cannot be obtained from the current decision alone and must arise from the continuation decision at $t = 1$. Hence, there exists an induced belief state b^i_1, arbitrarily close to the current one, at which a local change of the policy at $t = 1$ strictly improves the continuation objective.

By the local improvement property at $t = 1$, this strict improvement implies the existence of an improving direction at that belief state. Since the continuation value is continuous with respect to the induced belief state, the same improving direction remains improving in a neighborhood of the current belief state. It can therefore be detected at the current belief state, contradicting the assumption that no improving direction can be identified inductively.

Thus, for $T = 2$, every safe local improvement is detected at one of the two stages.

The argument extends inductively to an arbitrary finite horizon. Suppose that no improving direction can be identified at any stage. Then, after ruling out an improvement at the current stage, any remaining improvement must arise from the continuation problem. Applying the same argument

to this continuation problem yields a later stage at which a strict local improvement exists. By continuity, this improvement must already be detectable at the corresponding current induced belief state, contradicting the assumption.

Therefore, whenever a safe global improving direction exists, the dynamics identify a strictly improving direction at some stage t. □

The next theorem characterizes the points toward which the continuous dynamics converges.

Theorem 7. *The continuous dynamics ϕ introduced in Lemma 16 converges to the set of Nash equilibria of the game.*

Proof. Consider a joint policy

$$\pi \notin \mathcal{N},$$

outside of the set of the Nash Equilibria. Since the dual gap vanishes if and only if the joint policy is a Nash equilibrium,

$$d_0(\pi) > 0.$$

$$D_{F_t^i} d_0(\pi) < 0.$$

Therefore, at every joint policy outside $\mathcal{N}$, the dynamics admit a direction that strictly decreases the dual gap. By Lemma 16, the updates can be chosen sufficiently small so that the simultaneous evolution of the two players does not cancel this strict improvement. Consequently, along the continuous dynamics,

$$\frac{\mathrm{d}}{\mathrm{d}\tau} d_0(\phi(\tau)) < 0 \qquad \text{whenever } \phi(\tau) \notin \mathcal{N}.$$

Thus, the dual gap is a strict Lyapunov function outside the set $\mathcal{N}$.

Because the joint policy space is compact and the dual gap is non-negative, the function

$$g(\tau) = d_0(\phi(\tau))$$

is monotonically decreasing and bounded from below. Hence, it converges to some $g_\infty \geq 0$.

Suppose, for contradiction, that

$$g_\infty > 0.$$

Then $\phi(\tau)$ remains outside $\mathcal{N}$ for all sufficiently large τ. At every such point, the argument above provides a strictly improving direction for the dynamics, and therefore the dual gap must continue to decrease. This contradicts the convergence of $g(\tau)$ to the strictly positive limit g_∞.

Hence,

$$g_\infty = 0,$$

and therefore

$$\lim_{\tau \to \infty} d_0(\phi(\tau)) = 0.$$

Since the dual gap is zero if and only if the joint policy belongs to $\mathcal{N}$, the trajectory $\phi(\tau)$ converges to the set of Nash equilibria. □

The previous result establishes that the continuous dynamics converge to the set of Nash equilibria. We now relate these continuous trajectories to the discrete-time policy updates considered in practice. In particular, we show that, under the appropriate conditions, the discrete dynamics track the continuous dynamics sufficiently closely to inherit the same convergence property. This connection is formalized in the following theorem.

Proof of Theorem 2. The discrete algorithm can be written as a stochastic approximation of the continuous dynamics introduced in Definition 4. In particular, for each player i,

$$\pi^{i,k+1} = (1 - \alpha_k)\pi^{i,k} + \alpha_k \tilde{F}_t^i \left(\pi^{i,k}, \pi^{-i,k}\right) + U_{i,k+1}.$$

The operator $\tilde{F}_t^i$ is the discrete counterpart of the continuous improving direction F_t^i. Both operators perform the same local pessimistic optimization, but they differ in how this optimization is carried out. While F_t^i describes the continuous direction of improvement, $\tilde{F}_t^i$ is obtained by solving the stage-wise LPs used by the algorithm. These LPs exploit the uniform continuity of the value function to construct a local approximation of the pessimistic Q-value. In particular, $\tilde{F}_t^i$ maximizes this local approximation while matching the exact pessimistic value at the current point (see Theorem 6). Thus, the LP provides a discrete approximation of the same local optimization characterized by F_t^i.

Consequently, the interpolated sequence of policies can be viewed as a stochastic approximation of the continuous system

$$\dot{\pi^i} = F^i \left(\pi^i, \pi^{-i}\right).$$

Under the Robbins-Monro conditions and the standard assumptions on the noise, the cumulative effect of the stochastic perturbations remains controlled, while the updates continue to explore the policy space. Together with the assumed consistency of the LP operator with the local continuous dynamics, this ensures that the interpolated trajectory is an asymptotic pseudo-trajectory of the flow induced by $\left(F^1, F^2\right)$.

Proposition 1.3 of Benaïm et al. (2005) provides precisely the required connection. It states, under the corresponding stochastic approximation conditions, that the asymptotic behavior of the algorithm is preserved when its iterates are viewed as perturbed solutions of the limiting differential inclusion. Since we have established that the trajectories of this continuous dynamics converge to a Nash equilibrium, the proposition allows us to transfer this convergence result to the discrete algorithm.

By Lemma 7, the limiting continuous dynamics converges to the set $\mathcal{N}$ of Nash equilibria. Hence, the only internally chain-transitive invariant set of the limiting dynamics is contained in $\mathcal{N}$. Benaïm et al. (2005)'s stochastic approximation result then implies that the limit set of the discrete iterates is almost surely contained in $\mathcal{N}$. Therefore,

$$\operatorname{dist}\left(\pi^k, \mathcal{N}\right) \xrightarrow[k\to\infty]{\text{a.s.}} 0.$$

Thus, the algorithm converges almost surely to the set of Nash equilibria. □

C Extensive Form Game

C.1 A Hierarchy of Planners

Escudie et al. (2025) introduce a hierarchy of planners for zero-sum partially observable stochastic games (zs-POSGs), based on the observation that different planning procedures arise from progressively stronger assumptions about how future decisions are treated during planning. In particular, the hierarchy distinguishes planners according to the information they have access to and the extent to which future decisions—both their own and those of the opponent—are treated as fixed or optimized during planning.

At the two levels relevant here, the hierarchy can be summarized as follows:

- **Focal planner:** a player-specific planning process in which only the focal player's policy is known. The planner therefore reasons about its own decisions without access to the opponent's policy.

- **Uninformed planner:** a planner that additionally knows the opponent's policy, together with the information available to the focal planner. This setting can be formulated as a transition-independent zero-sum stochastic game, and is closely related to the stochastic-game formulation introduced at the beginning of this paper.

Beyond this operational characterization, we argue that the hierarchy admits a deeper structural interpretation. To make this interpretation explicit, we first consider the example of Sokota et al. (2023), which illustrates why a reduction based solely on public policies can fail. In particular, their discussion of the *Backward Dependence Problem* highlights how future decisions may depend on information and decisions made earlier in the game, even when these dependencies are not captured by the public policy representation alone.

C.2 Equilibrium and Dynamic Programming

The distinction between simultaneous- and sequential-move games provides a deeper motivation for the hierarchy of planners introduced by Escudie et al. (2025). We illustrate this point by revisiting the example of Sokota et al. (2023) and comparing it with the approach developed in Escudie et al. (2025).

In a simultaneous-move POSG, the equilibrium sought by the dynamic-programming formulation is a Subgame Perfect Nash Equilibrium (SPNE). This is reflected directly in the corresponding Bellman recursion. In the two-player zero-sum case, the backward recursion can be written schematically as

$$NE_{(\pi_0^1,\pi_0^2)} NE_{(\pi_1^1,\pi_1^2)} \cdots NE_{(\pi_{T-1}^1,\pi_{T-1}^2)} v^{\pi^1,\pi^2},$$

where NE denotes a Nash equilibrium of the corresponding subgame. Thus, the equilibrium is constructed recursively, with the decision rules at each stage being optimized within the corresponding subgame. Because the recursion solves each subgame, the resulting strategy profile satisfies the subgame-perfectness requirement.

This correspondence between dynamic programming and SPNE does not hold in the same way for sequential POSGs, and consequently for extensive-form games (EFGs). In an EFG, the Nash equilibrium sought by the game formulation need not be subgame perfect, even after a reduction to public policies.

This distinction is particularly apparent in Kuhn Poker, where the optimal decision rule at the final stage is not deterministic. If the Bellman recursion above is nevertheless applied directly to the sequential game after reducing it to a public-policy representation, the final stage is optimized through a $\max$ or $\min$ over the available decision rules. This optimization selects a deterministic decision rule at the final stage and therefore fails to capture the dependence of the final decision on deviations at earlier stages. Sokota et al. (2023) identify the same structural issue in their example of the "Backward Dependence Problem."

This observation motivates changing the equilibrium formulation rather than simply applying the simultaneous-game Bellman recursion to a sequential game. Nevertheless, an extensive-form game can be transformed into a lossless game for which a Subgame Perfect Nash Equilibrium exists. The resulting formulation is closely related to the reduction introduced at the beginning of this paper.

C.3 One-Sided Extensive-Form Games

How should we handle terminal values in a finite-horizon formulation? In an extensive-form game (EFG), terminal nodes may be reached before the last possible time step, whereas the dynamic-programming formulation evaluates the value at the final time step. Without loss of generality, we can postpone the realization of any terminal reward until the final time step by introducing absorbing states. More precisely, whenever a terminal node is reached at time $t < T-1$, we transition to an absorbing state and keep the process there until time $T-1$, at which point the corresponding terminal reward is received. This transformation preserves the value of every policy profile while ensuring that all rewards are realized at the final time step. This is done for readability and to maintain consistency across the different game formulations, thereby simplifying the notation and equations.

Definition 6. Let i be a player.We define the associated **one-sided EFG (oEFG)** as

$$\mathcal{E}^i \doteq (\mathcal{I}, \mathcal{D}, \mathbf{H}, \mathbf{H}^k, \mathcal{D}_t^i, u),$$

where the components are defined as follows:

- $\mathcal{I}$ is the set of n agents in the original game.
- $\mathcal{D}$ is the decision tree. Each node at time t represents a joint policy prefix
$$\pi_{0:t} = \{\pi_0^1, \pi_0^2, \dots, \pi_t^k\},$$
where k denotes the player k. A child node π' is accessible from π if
$$\exists\, \pi_t^k \quad \text{such that} \quad \pi' = [\pi, \pi_t^k].$$
- $\mathbf{H}^k$ is the information partition of player k:
$$\mathbf{H}^k(\pi_{0:t}) \doteq \begin{cases} \pi_{0:t}^k & \text{if } k = i, \\ \pi_{0:t} & \text{otherwise.} \end{cases}$$
Thus, the focal player i observes only her own policy prefix, whereas the opponent $-i$ observes the complete joint policy prefix.
- $\mathcal{D}_t^k = \Pi_t^k$ is the space of decision rules available to player k at time t.
- u is the reward function defined on terminal nodes:
$$u(\pi_{0:T}) \doteq v^{\pi_{0:T}^1, \pi_{0:T}^2}.$$

The players make their decisions sequentially, with i choosing first and $-i$ choosing after observing the complete policy prefix of i. Consequently, the value of the game is characterized by a max-min optimization: i chooses a decision rule to maximize the value, while $-i$ observes this choice and subsequently chooses a decision rule to minimize it.

Under this construction, i has access only to her own policy, whereas $-i$ has access to both players' policies. We then show that any zero-sum EFG admits a lossless reduction to the corresponding zero-sum oEFG.

Theorem 8. *Let $\mathcal{E}$ be a zero-sum extensive-form game and let i be the focal player. The reduction*

$$\mathcal{T}_{\mathcal{E} \to \mathcal{E}^i}$$

that projects $\mathcal{E}$ onto the one-sided EFG $\mathcal{E}^i$ is lossless for player i. In particular, it preserves the value of every joint policy and the focal player's maximin value and optimal policies.

Proof. We establish the result by showing that the reduction preserves (i) the correspondence between joint policies and paths in the decision tree, (ii) the associated values, and (iii) the information available to the focal player.

Policy and path correspondence. Consider a joint policy

$$\pi_{0:T} = \left(\pi_{0:T}^i, \pi_{0:T}^{-i}\right)$$

in the original EFG $\mathcal{E}$. The reduction associates with this joint policy a policy $\tilde{\pi}$ in the one-sided EFG $\mathcal{E}^i$.

At each node corresponding to a policy prefix $\pi_{0:t}$, the next decision of the focal player is determined by her own policy prefix $\pi_{0:t}^i$, whereas the next decision of the opponent is determined by the complete policy prefix $\pi_{0:t}$. Thus, the complete joint policy uniquely determines a path in the decision tree $\mathcal{D}$, and conversely, a path in $\mathcal{D}$ specifies the sequence of decision rules constituting the corresponding joint policy.

Value preservation. By construction, the terminal node reached by the path associated with $\pi_{0:T}$ has reward

$$u(\pi_{0:T}) \doteq v^{\pi^i_{0:T}, \pi^{-i}_{0:T}}.$$

Consequently, the expected value induced by any joint policy is unchanged by the reduction:

$$u(\tilde{\pi}) = v^{\pi^i_{0:T}, \pi^{-i}_{0:T}}.$$

Hence, optimizing over corresponding policies in the two representations yields the same payoff. In particular, the reduction does not introduce or remove any payoff-relevant trajectories.

Information preservation for the focal player. The reduction is one-sided precisely because it preserves the information available to the focal player while explicitly exposing additional information to the opponent. In $\mathcal{E}^i$, player i observes only her own policy prefix through the information partition $\mathbf{H}^i$:

$$\mathbf{H}^i(\pi_{0:t}) = \pi^i_{0:t}.$$

This is exactly the information on which her decision rule is allowed to depend. By contrast, the opponent observes the complete policy prefix,

$$\mathbf{H}^{-i}(\pi_{0:t}) = \pi_{0:t},$$

and can therefore condition its decision on the focal player's previous decisions. The opponent thus acts as an informed best responder. Since the theorem concerns losslessness for player i, granting this additional information to the opponent does not alter the focal player's policy space or information structure; rather, it makes the opponent's minimization explicit.

Preservation of the focal player's maximin problem. Because the game is zero-sum, the value associated with a joint policy can be written as

$$v^{\pi^i, \pi^{-i}}.$$

The focal player's security problem in the original game is therefore

$$\max_{\pi^i} \min_{\pi^{-i}} v^{\pi^i, \pi^{-i}}.$$

The policy and value correspondence established above implies that the same optimization problem is represented in $\mathcal{E}^i$:

$$\max_{\tilde{\pi^i}} \min_{\tilde{\pi^{-i}}} u(\tilde{\pi}).$$

Thus, the two formulations have the same maximin value, and a policy is optimal for player i in the original EFG if and only if its corresponding policy is optimal in the one-sided EFG.

Finally, by the minimax theorem for finite zero-sum games, the maximin value coincides with the value attained at a Nash equilibrium. Hence, the focal player's equilibrium value and the set of focal-player maximin strategies are preserved by the reduction, even though the one-sided representation does not claim to preserve the opponent's information structure or the full set of equilibrium strategies. □

Theorem 9. *Under the convention introduced in Definition 6, the one-sided EFG $\mathcal{E}^i$ admits a unilateral Subgame Perfect Nash Equilibrium. In particular, its value satisfies the following Bellman recursion:*

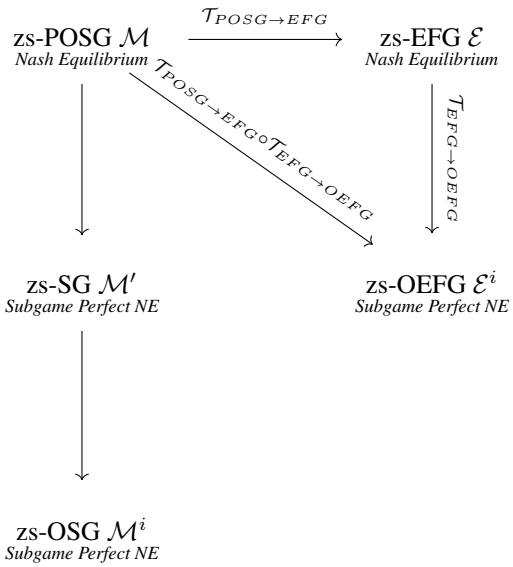

Figure 4: $\mathcal{T}_{\mathcal{G}\to\mathcal{G}'}$ denotes the reduction from game $\mathcal{G}$ to game $\mathcal{G}'$. The reduction $\mathcal{T}_{POSG\to EFG}$ is almost transparent: it simply reformulates the POSG as an extensive-form game (EFG). However, unlike $\mathcal{T}_{POSG\to SG}$, this reduction does not induce an equilibrium concept that can be characterized inductively in the reduced game. The composition $\mathcal{T}_{POSG\to EFG}\circ\mathcal{T}_{EFG\to OEFG}$ corresponds to the focal-planner reduction introduced in Escudie et al. (2025).

$$v_t^{*,i}(\pi^i_{0:t-1}) = \max_{\pi^i_t}\ \min_{\pi^{-i}_{0:T-1}}\ v^{*,i}_{t+1}(\pi^i_{0:t}),$$

where $v_t^*(\pi^i_{0:t-1})$ denotes the optimal continuation value for player i after fixing her policy prefix $\pi^i_{0:t-1}$.

At the final decision stage, the recursion reduces to

$$v^{*,i}_{T-1}(\pi^i_{0:T-2}) = \max_{\pi^i_{T-1}}\ \min_{\pi^{-i}_{0:T-1}}\ u_{T-1}\left(\pi^i_{0:T-1}, \pi^{-i}_{0:T-1}\right).$$

Proof. We prove by backward induction the stronger statement $(*)$,

$$v_t^{*,i}(\pi^i_{0:t-1}) = \max_{\pi^i_{t:T-1}}\ \min_{\pi^{-i}_{0:T-1}}\ u\left([\pi^i_{0:t-1}, \pi^i_{t:T-1}], \pi^{-i}_{0:T-1}\right)$$

for every t.

Base case. At $t = T-1$, there is no future decision of player i after π^i_{T-1}. Therefore, by definition,

$$v^{*,i}_{T-1}(\pi^i_{0:T-2}) = \max_{\pi^i_{T-1}}\ \min_{\pi^{-i}_{0:T-1}}\ u\left([\pi^i_{0:T-2}, \pi^i_{T-1}], \pi^{-i}_{0:T-1}\right),$$

so property $(*)$ holds.

Induction step. Assume that $(*)$ holds at time $t+1$. Then

$$\begin{aligned} v_t^{*,i}(\pi^i_{0:t-1}) &= \max_{\pi^i_t}\ \min_{\pi^{-i}_{0:T-1}}\ v^{*,i}_{t+1}(\pi^i_{0:t}) \\ &= \max_{\pi^i_t}\ \min_{\pi^{-i}_{0:T-1}}\ \max_{\pi^i_{t+1:T-1}}\ \min_{\pi^{-i}_{0:T-1}}\ u\left([\pi^i_{0:t-1}, \pi^i_t, \pi^i_{t+1:T-1}], \pi^{-i}_{0:T-1}\right). \end{aligned}$$

We now justify the exchange of the middle min and max. Fix a prefix $\pi^i_{0:t}$. Player i chooses a complete continuation policy

$$\pi^i_{t+1:T-1},$$

while player $-i$ chooses a complete policy

$$\pi^{-i}_{0:T-1}.$$

For fixed $\pi^i_{0:t}$, let the payoff of the corresponding pure-strategy pair be

$$U\left(\pi^i_{t+1:T-1}, \pi^{-i}_{0:T-1}\right) = u\left([\pi^i_{0:t}, \pi^i_{t+1:T-1}], \pi^{-i}_{0:T-1}\right).$$

When the two players randomize over these complete policies, the expected payoff is bilinear in their mixed strategies. Hence, by the minimax theorem,

$$\min_{\pi^{-i}_{0:T-1}} \max_{\pi^i_{t+1:T-1}} U = \max_{\pi^i_{t+1:T-1}} \min_{\pi^{-i}_{0:T-1}} U.$$

This is the same minimax argument used in Delage et al. (2023). Importantly, the argument is applied only after fixing the focal player's prefix $\pi^i_{0:t}$; the continuation policies of the two players then constitute the strategy spaces of the induced zero-sum normal-form game.

Applying this equality inside the backward recursion gives

$$\begin{aligned} v^{*,i}_t(\pi^i_{0:t-1}) &= \max_{\pi^i_t} \max_{\pi^i_{t+1:T-1}} \min_{\pi^{-i}_{0:T-1}} u\left([\pi^i_{0:t-1}, \pi^i_t, \pi^i_{t+1:T-1}], \pi^{-i}_{0:T-1}\right) \\ &= \max_{\pi^i_{t:T-1}} \min_{\pi^{-i}_{0:T-1}} u\left([\pi^i_{0:t-1}, \pi^i_{t:T-1}], \pi^{-i}_{0:T-1}\right). \end{aligned}$$

This is precisely property $(*)$ at time t, completing the induction.

Finally, applying $(*)$ at $t = 0$ yields

$$v^{*,i}_0 = \max_{\pi^i_{0:T-1}} \min_{\pi^{-i}_{0:T-1}} u\left(\pi^i_{0:T-1}, \pi^{-i}_{0:T-1}\right).$$

Thus, the policy generated by the Bellman recursion is a maximin-optimal policy for player i. Since the opponent observes the complete policy prefix of player i, this sequential maximin construction corresponds to the unilateral Subgame Perfect Nash Equilibrium of the one-sided EFG. Hence, the Bellman recursion computes the equilibrium value and an equilibrium strategy for the focal player. □

A few observations follow from this construction. First, by reducing an EFG to a one-sided EFG, we obtain a well-defined notion of a proper subgame for the sequential game. This allows us to formulate and compute a Subgame Perfect Nash Equilibrium, thereby addressing the backward-dependence issue identified by Sokota et al. (2023). In particular, the one-sided construction makes the information available to the two players explicit and ensures that the continuation problem at each stage is formulated as the appropriate sequential zero-sum game.

Second, this result is closely related to Theorem 3 of Escudie et al. (2025). While our proof is formulated directly in terms of the decision tree and does not explicitly use the occupancy-state representation, the correspondence between the two formulations is immediate: the nodes of the one-sided EFG can equivalently be represented by the corresponding occupancy states. Thus, the construction provides an alternative, more explicit interpretation of the same underlying reduction.

The main distinction is the scope in which this construction is required. The one-sided reduction is necessary when dealing with sequential games, such as EFGs or sequential POSGs, because future decisions may depend on previously revealed decisions and therefore cannot in general be optimized

independently through the standard simultaneous-move Bellman recursion. In contrast, most of the games considered in this paper are simultaneous-move POSGs. In that setting, the standard dynamic-programming formulation already corresponds to the appropriate subgame-perfect solution concept, and the additional one-sided reduction is therefore unnecessary. The relationship between these two cases is illustrated in Figure 4.

We work with the uninformed planner introduced by Escudie et al. (2025) because deviations in the past cannot depend on the public history. If we were to use an informed planner instead, the resulting optimization would allow a different deviation for each public history. This would effectively give the best response access to the public history before the corresponding decision is made, which is not available to the player and would therefore be inconsistent with the information structure of the game.

C.4 Experiments

We extend Algorithm 1 to extensive-form games by adapting the local policy-improvement step to account for deviations from decisions made at earlier stages of the game. In particular, the linear program is modified to retain the relevant information about past decisions when computing the current deviation, allowing the local improvement procedure to operate directly on the extensive-form representation.

To obtain a deviation model consistent with the intended information structure, we include public histories in the support of the belief representation. The deviation is therefore defined with respect to the information available to the player rather than separately for each public history. Without this restriction, the optimization could effectively select a different deviation for every public history, leading to an overly pessimistic evaluation that does not correspond to the intended player model.

We implemented this extension and evaluated it on Kuhn Poker. The resulting algorithm exhibits convergence to the expected equilibrium behavior in our experiments. However, the current implementation is primarily intended as a proof-of-concept: the code requires further cleanup and validation before the experimental results can be presented as part of the main empirical evaluation. We therefore leave a more systematic experimental study of the extensive-form extension to future work.

D Experiments

D.1 Baselines

PBVI in zs-SG In a POMDP, the value function is defined over the belief simplex and is piecewise-linear and convex, making exact value iteration intractable. PBVI addresses this issue by restricting computation to a finite set of representative beliefs. At each iteration, Bellman backups are performed only on this set, yielding an approximate value function that is accurate in the vicinity of the sampled beliefs while remaining tractable.

PBVI is inherently a single-agent method. In adversarial settings, it is therefore applied to a one-sided (focal-planner) surrogate, where one player is optimized while the opponent is treated as part of the environment. Two levels of sampling arise. First, beliefs are expanded via forward simulation to cover reachable regions of the belief space. Second, in the one-sided reduction, the opponent is simulated (e.g., via randomized policies), so that the focal player optimizes against sampled behaviors. This relies on the assumption that relevant opponent responses lie near those induced by sampling.

Each PBVI backup requires a maximization over actions and observations, which in adversarial settings leads to linear programs for worst-case evaluation. The number of α-vectors grows rapidly with iterations, increasing the size of these programs. Consequently, the Bellman operator quickly becomes intractable despite the point-based approximation.

CFR in POSG. Counterfactual Regret Minimization (CFR) is an iterative algorithm for solving extensive-form games (EFGs) that minimizes counterfactual regret at each information set, yielding convergence to a Nash equilibrium in two-player zero-sum settings. At each iteration, CFR eval-

Name in EFG	Symbol	Name in POSG	Symbol
History	h	State and Joint history	(s_t, h_t)
i's information set	I	i's private history	h_i
Strategy profile	σ	Joint policy	π
i's private strategy	σ_i	i's private policy	π_i
Players' contribution except i	$\pi_i^\sigma(I)$	Players' contribution except i	$\Pr(h_i \mid \pi, b_0)$
Players' contribution except i	$\pi_i^\sigma(h)$	Counterfactual state	$c_i(s, h_{-i})$
Counterfactual value of σ at h	$v_i(\sigma, h)$	Value of π at (s, h)	$v^\pi_{\mathcal{M}}(s, h)$
Deviation at I	$\sigma_{I \to a}$	Deviation at h_i	$\pi_{h_i \to a_i}$

Table 3: Translation from EFG to POSG. For more information, we refer the reader to (Kovařík et al., 2022). We connect the history h in EFG to the pair (s, h) in POSG since the reward is not associated to the history but to the state in POSG. The quantity $v^\pi_{\mathcal{M}}(s, h)$ is formally defined in (Cunha et al., 2023), we refer the reader for more information. It is simply the expected cumulated value a player can get at this pair.

uates the current strategy profile, computes counterfactual values, and updates strategies via regret matching.

The connection between POSGs and EFGs has been established in (Kovařík et al., 2022), laying the groundwork for the application of extensive-form techniques such as CFR. Notably, the theoretical objects introduced in our paper naturally embed CFR and yield a refined variant tailored to POSGs. To make this connection explicit, Table 3 presents a correspondence between standard EFG notions and their POSG counterparts, thereby providing a bridge for transferring the algorithm. The relevant notation is summarized in Table 3.

From (Zinkevich et al., 2007; Neller & Lanctot, 2013), we adapt the key ingredients of CFR to the POSG setting. We first define the (instantaneous) regret at (s_t, h_t) for player $-i$ when deviating only at $h_{-i,t}$ to action $a_{-i,t}$:

$$r_{\mathcal{M}}(s_t, h_t, a_{i,t}) = v^{\pi_{h_{i,t} \to a_{i,t}}}_{\mathcal{M}}(s_t, h_t) - v^{\pi}_{\mathcal{M}}(s_t, h_t).$$

In contrast to EFGs, where regret is indexed by histories h, in POSGs it is defined over (s_t, h_t) since rewards depend on the (hidden) state s_t. The history remains necessary to condition the belief over states.

Aggregating over histories that share the same private history $h_{i,t}$ yields a vector-valued regret:

$$\mathbf{r}^{h_{i,t} \to a_{i,t}}_{\mathcal{M}}(s_t, h_{i,t}) = v^{\pi_{h_{-i,t} \to a_{-i,t}}}_{\mathcal{M}}(s_t, h_{i,t}, h_{-i,t}) - v^{\pi}_{\mathcal{M}}(s_t, h_{i,t}, h_{-i,t}),$$

which is a vector over $\mathcal{C}_i$ (one coordinate per state). Using the conditional state c_i notation, we recover a scalar regret:

$$R^{h_{-i,t} \to a_{-i,t}}_{\mathcal{M}} = \mathbf{r}^{h_{-i,t} \to a_{-i,t}} \cdot c_i.$$

This corresponds to the EFG quantity

$$r_{EFG}(I, a) = \sum_{h \in I} \big(v_i(\sigma_{I \to a}, h) - v_i(\sigma, h)\big),$$

where the summation over histories is replaced by expectation under the counterfactual occupancy.

The policy is then updated via regret matching:

$$\pi_{-i,\text{new}}(a_{-i,t} \mid h_{-i,t}) = \frac{\big[R^{h_{-i,t} \to a_{-i,t}}_{\mathcal{M}}\big]_+}{\sum_{a'_{-i,t}} \big[R^{h_{-i,t} \to a'_{-i,t}}_{\mathcal{M}}\big]_+}.$$

The regret of player $-i$ is inherently *vector-valued*, as it is defined conditionally on the hidden state. The conditional state c_i acts as a belief, projecting this vector onto a scalar quantity, exactly as

Table 4: Snapshot of empirical results. For each game, horizon and algorithm, we report time, value and exploitability ϵ associated with the computed ϵ-Nash equilibrium. OOT occurs when the target exploitability threshold cannot be reached within the prescribed time limit, while OOM means that the algorithms ran out of memory. **magenta** = lowest exploitability value per benchmark, while '$-$' means the time budget for computing the exploitability has been exceeded.

Problem	T	CPI (Ours)			PBVI			HSVI			CFR+		
		t	val.	ϵ	t	val.	ϵ	t	val.	ϵ	t	val.	ϵ
MABC	2	0.3	**0.078**	**0.**	**0.2**	**0.078**	**0.**	0.4	**0.078**	**0.**	0.3	**0.078**	**0.**
	3	11	**0.097**	**0.**	**24**	0.096	**0.**	70	**0.097**	**0.**	**7**	0.095	**0.**
	4	**23**	**0.11**	**0.**	65	**0.11**	0.01	OOT	OOT	OOT	25	**0.11**	0.01
	5	**29**	**0.12**	**0.**	31	**0.12**	**0.**	OOT	OOT	OOT	69	0.12	0.01
	7	**97**	**0.14**	0.01	190	**0.14**	**0.0**	OOT	OOT	OOT	92	**0.14**	0.02
	10	**100**	**0.18**	**0.01**	637	0.17	**0.01**	OOT	OOT	OOT	122	**0.18**	0.02
	15	**143**	**0.22**	**0.01**	668	0.22	0.02	OOT	OOT	OOT	222	**0.22**	0.03
	30	**360**	**0.36**	**0.01**	OOT	OOT	OOT	OOT	OOT	OOT	**433**	**0.36**	0.06
MATCHING PENNIES	2	**0.01**	**0.4**	**0.**	2	0.39	**0.**	0.1	**0.4**	**0.**	0.06	**0.4**	0.005
	3	**0.02**	**0.6**	**0.**	18	**0.6**	**0.**	1	**0.6**	**0.**	0.1	**0.6**	0.001
	4	**0.04**	**0.8**	**0.**	34	0.77	0.	4	**0.8**	**0.**	0.2	**0.8**	**0.**
	5	**0.04**	**1.**	**0.**	52	0.95	**0.**	60	**1.**	**0.**	0.25	**1.**	**0.**
	7	**13**	**1.4**	**0.**	180	1.31	0.03	OOT	OOT	OOT	15	1.39	0.01
	10	**92**	**2.**	**0.**	OOT	OOT	OOT	OOT	OOT	OOT	103	1.99	0.02
	25	**15**	**5.**	**0.**	OOT	OOT	OOT	OOT	OOT	OOT	132	**5.**	0.02
	50	**217**	**10.**	**0.**	OOT	OOT	OOT	OOT	OOT	OOT	718	10.	0.03
RECYCLING	2	**0.4**	**0.26**	**0.**	**0.4**	**0.26**	**0.**	1.2	**0.26**	**0.**	2.8	**0.26**	0.004
	3	**17**	**0.32**	**0.01**	78	**0.32**	**0.**	120	**0.32**	**0.**	75	**0.32**	0.01
	4	**42**	**0.36**	**0.03**	56	0.35	0.01	OOT	OOT	OOT	104	0.35	0.04
	5	**213**	**0.4**	**0.05**	98	0.38	0.05	OOT	OOT	OOT	177	0.38	0.05
	7	**271**	**0.5**	**0.06**	1481	0.48	0.04	OOT	OOT	OOT	503	**0.49**	0.12
	10	**505**	0.63	**0.07**	3232	0.6	**0.06**	OOT	OOT	OOT	1151	**0.64**	0.15
	15	**703**	**0.87**	**0.08**	OOT	OOT	OOT	OOT	OOT	OOT	2415	0.86	0.23
ADVERSARIAL TIGER	2	**0.1**	**-0.4**	**0.**	**0.1**	**-0.4**	**0.**	0.2	**-0.4**	**0.**	**0.1**	**-0.4**	**0.0**
	3	**0.2**	**-0.56**	**0.**	0.3	**-0.56**	**0.**	500	**-0.56**	**0.**	5	**-0.56**	**0.**
	4	**16**	**-0.75**	0.01	46	-0.76	**0.02**	OOT	OOT	OOT	76	**-0.75**	0.03
	5	82	**-0.96**	0.05	141	-0.96	**0.05**	OOT	OOT	OOT	**87**	**-0.96**	0.07
	7	**248**	-1.36	0.13	866	-1.36	**0.1**	OOT	OOT	OOT	172	**-1.35**	0.13
	10	824	-1.97	0.24	956	-1.98	0.18	OOT	OOT	OOT	556	-1.95	0.3

counterfactual reach probabilities do in EFGs. This establishes a direct connection with the focal-planner view: regret minimization operates on belief-weighted values, consistent with optimizing a surrogate game for a fixed opponent.

Our framework can be interpreted as performing regret minimization directly in the occupancy space. Instead of explicitly maintaining regrets per history, we operate on regret functions over counterfactual occupancy states, and policy updates correspond to regret matching under these occupancy-induced weights. This yields a compact implementation that avoids explicit enumeration of histories.

D.2 Implementation Details

The results presented in this paper were run on CPU (AMD Ryzen 7 PRO).

Table 5: A snapshot of empirical results with no pruning mechanism applied (see Escudie et al. (2025), Dibangoye et al. (2016a)). For each game, horizon and algorithm, we report time, value and exploitability ϵ associated with the computed ϵ-Nash equilibrium. OOT occurs when the target exploitability threshold cannot be reached within the prescribed time limit, while OOM means that the algorithms ran out of memory. **magenta** = lowest exploitability value per benchmark, while '$-$' means the time budget for computing the exploitability has been exceeded.

Problem	T	CPI (Ours)			PBVI			HSVI			CFR+		
		t	val.	ϵ	t	v	ϵ	t	val.	ϵ	t	val.	ϵ
MABC	2	0.3	**0.078**	**0.**	**0.2**	**0.077**	**0.**	0.3	**0.078**	**0.**	0.3	**0.078**	0.003
	3	**5**	**0.097**	**0.**	19	**0.097**	**0.**	70	**0.097**	**0.**	12	0.096	0.002
	4	**38**	**0.11**	**0.001**	50	0.1	**0.001**	OOT	OOT	OOT	73	**0.11**	0.003
	5	2176	**0.12**	0.01	1365	**0.12**	0.013	OOT	OOT	OOT	1614	**0.12**	**0.005**
MATCHING PENNIES	2	**0.01**	**0.4**	**0.**	0.13	0.38	**0.**	0.1	0.4	0.	**0.01**	**0.4**	**0.**
	3	**0.05**	**0.6**	**0.**	9	**0.6**	**0.**	1	0.6	0.	0.2	**0.6**	**0.**
	4	**0.5**	**0.8**	**0.**	250	0.77	**0.**	4	0.8	0.	2	**0.8**	0.005
	5	11	**1.0**	**0.003**	3207	1.14	0.12	60	1.	0.	**5**	1.	0.006
RECYCLING	2	0.4	**0.26**	**0.**	**0.1**	**0.26**	**0.**	1.2	**0.26**	**0.**	1.5	**0.26**	0.01
	3	**14**	**0.32**	**0.**	286	**0.32**	**0.**	120	**0.32**	**0.**	131	**0.32**	0.01
	4	681	**0.35**	**0.**	**418**	0.34	**0.01**	OOT	OOT	OOT	848	**0.35**	0.03
	5	**5146**	**0.36**	**0.05**	OOT	OOT	OOT	OOT	OOT	OOT	7611	**0.36**	0.1
ADVERSARIAL TIGER	2	**0.**	**-0.4**	**0.**	**0.**	**-0.4**	**0.**	**0.**	**-0.4**	**0.**	**0.**	**-0.4**	0.002
	3	2	**-0.56**	**0.**	**0.17**	**-0.56**	**0.**	500	**-0.56**	**0.**	6	-0.57	0.007
	4	**146**	**-0.76**	**0.01**	213	-0.78	0.04	OOT	OOT	OOT	171	-0.73	0.03
	5	**2082**	**-0.96**	**0.02**	3237	**-0.96**	0.06	OOT	OOT	OOT	2583	**-0.96**	**0.02**