Heuristic Search Value Iteration can solve zero-sum Partially Observable Stochastic Games

Aurélien Delage, Olivier Buffet, Jilles Dibangoye, Abdallah Saffidine

Heuristic Search Value Iteration can solve zero-sum Partially Observable Stochastic Games

Aurélien Delage
Univ Lyon, INSA Lyon, Inria, CITI
69621 Villeurbanne, France
aurelien.delage@insa-lyon.fr

Olivier Buffet
Univ. Lorraine, INRIA, CNRS, LORIA
F-54000, Nancy, France
olivier.buffet@inria.fr

Jilles-Steeve Dibangoye
Univ Lyon, INSA Lyon, Inria, CITI
69621 Villeurbanne, France
jilles-steeve.dibangoye@insa-lyon.fr

Abdallah Saffidine
University of New South Wales
Sidney, Australia
abdallahs@cse.unsw.edu.au

ABSTRACT

State-of-the-art methods for solving 2-player zero-sum imperfect information games rely on linear programming or regret minimization, though not on dynamic programming (DP) or heuristic search (HS), while the latter are often at the core of state-of-the-art solvers for other sequential decision-making problems. In partially observable or collaborative settings (*e.g.*, POMDPs and Dec-POMDPs), DP and HS require introducing an appropriate statistic that induces a fully observable problem as well as bounding (convex) approximators of the optimal value function. This approach has succeeded in some subclasses of 2-player zero-sum partially observable stochastic games (zs-POSGs) as well, but how to apply it in the general case still remains an open question. We answer it by (i) rigorously defining an equivalent game to work with, (ii) proving mathematical properties of the optimal value function that allow deriving bounds that come with solution strategies, (iii) proposing for the first time an HSVI-like solver that provably converges to an ϵ-optimal solution in finite time, and (iv) empirically analyzing it. This opens the door to a novel family of promising approaches complementing those relying on linear programming or iterative methods.

ACM Reference Format:
Aurélien Delage, Olivier Buffet, Jilles-Steeve Dibangoye, and Abdallah Saffidine. 2024. Heuristic Search Value Iteration can solve zero-sum Partially Observable Stochastic Games. In *Appears at the 11th Multiagent Sequential Decision Making under Uncertainty Workshop (MSDM 2023). Held as part of the Workshops at the 21st International Conference on Autonomous Agents and Multiagent Systems., London, England, May-June 2023*, IFAAMAS, 9 pages.

Note : The research report containing proofs and details can be found in [11].

Appears at the 11th Multiagent Sequential Decision Making under Uncertainty Workshop (MSDM 2023). Held as part of the Workshops at the 21st International Conference on Autonomous Agents and Multiagent Systems., Yifeng Zeng, Yuchen Xiao, Yinghui Pan, Prashant Doshi (Chairs), May-June 2023, London, England. © 2024 Copyright held by the owner/author(s). …$ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

1 INTRODUCTION

Solving imperfect information sequential games is a challenging field with many applications from playing Poker [21] to security games [2]. We focus on finite-horizon 2-player zero-sum partially observable stochastic games (zs-POSGs), an important class of games that is equivalent to that of zero-sum extensive-form games (zs-EFGs) [24][1]. From the viewpoint of (maximizing) player 1, we aim at finding a strategy with a worst-case expected return (*i.e.*, whatever player 2's strategy) within ϵ of the Nash equilibrium value (NEV).

A first approach to solving a zs-POSG is to turn it into a zs-EFG addressed as a *sequence form* linear program (SFLP) [5, 17, 31], giving rise to an exact algorithm. A second approach is to use an iterative game solver, *i.e.*, either a counterfactual-regret-based method (CFR) [6, 35], or a first-order method [14, 20], both coming with asymptotic convergence properties. CFR-based approaches now incorporate deep reinforcement learning and search, some of them winning against top human players at heads-up no limit hold'em poker [6, 7, 23]. A third approach, proposed by Wiggers [32], is to use two parallel searches in strategy space, one per player, so that the gap between both strategies' security levels (*i.e.*, the values of their opponent's best responses) bounds the distance to the NEV.

In contrast, dynamic programming and heuristic search have not been applied to general zs-POSGs, while often at the core of state-of-the-art solvers in other problem classes that involve Markovian dynamics, partial observability and multiple agents (POMDP [1, 26], Dec-POMDP [12, 28], or subclasses of zs-POSGs with simplifying observability assumptions [3, 9, 10, 13, 15, 16]). They all rely on some statistic that induces a fully observable problem whose value function (V^*) exhibits continuity properties that allow deriving bounding approximations. Wiggers et al. [33, 34] progress in this direction for zs-POSGs by demonstrating an important continuity property of the optimal value function, and proposing a reformulation as a particular equivalent game. We work in a similar direction, (1) using a game with different observability hypotheses, (2) proving theoretical results they implicitly rely on, and (3) building on some of their results to derive an HSVI-like algorithm solving the zs-POSG.

[1]Note: POSGs are equivalent to the large class of "well-behaved" EFGs as defined by Kovařík et al. [19].

Section 2 presents some necessary background, including the concept of *occupancy state* [12, 33] (*i.e.*, the probability distribution over the players' past action-observation histories), and properties that rely on it. Then, Section 3 describes theoretical contributions. First, we rigorously reformulate the problem as a non-observable game, and demonstrate that the Nash equilibrium value can be expressed with a recursive formula, which is a required tool for DP and HS (Sec. 3.1). Second, we exhibit novel continuity properties of optimal value functions and derive bounding approximators, a second tool made necessary due to the continuous state space of the new game, before showing that these approximators come with valid solution strategies for the zs-POSG (Sec. 3.2). Third, we adapt Smith and Simmons' [27] HSVI's algorithmic scheme to ϵ-optimally solve the problem in finitely many iterations (Sec. 3.3). Section 4 presents an empirical analysis of the approach.

2 BACKGROUND

Here, we first give basic definitions about zs-POSGs, including the solution concept at hand. Then we introduce an equivalent game where a state corresponds to a statistic summarizing past behaviors, which leads to some important properties of the game's optimal value.

2.1 zs-POSGs

Definition 2.1 (zs-POSGs)**.** As illustrated through a dynamic influence diagram in Figure 1, a (2-player) zero-sum partially observable stochastic game (zs-POSG) is defined by a tuple $\langle \mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, \mathcal{Z}^1, \mathcal{Z}^2, P, r, H, \gamma, b_0 \rangle$, where

- $\mathcal{S}$ is a finite set of states;
- $\mathcal{A}^i$ is (player) i's finite set of actions;
- $\mathcal{Z}^i$ is i's finite set of observations;
- $P^{z^1,z^2}_{a^1,a^2}(s'|s)$ is the probability to transition to state s' and receive observations z^1 and z^2 when actions a^1 and a^2 are performed while in state s;
- $r(s, a^1, a^2)$ is a (scalar) reward function;
- $H \in \mathbb{N}$ is a (finite) temporal horizon;
- $\gamma \in [0, 1]$ is a discount factor; and
- b_0 is the initial belief state, *i.e.*, a probability distribution over states at $t = 0$.

From the Dec-POMDP, POSG and EFG literature, we use the following concepts and definitions:

$\theta^i_\tau = (a^i_0, z^i_1, \ldots, a^i_{\tau-1}, z^i_\tau)$ is a length-τ *action-observation history* (AOH) for i. We note Θ^i_τ the set of all AOHs for player i at horizon τ, so that any AOH θ^i_τ is in $\cup_{t=0}^{H-1}\Theta^i_t$.

β^i_τ is a *(behavioral) decision rule* (DR) at τ for i, *i.e.*, a mapping from private AOHs in Θ^i_τ to *distributions* over private actions. $\beta^i_\tau(\theta^i_\tau, a^i)$ is the probability to pick a^i when facing θ^i_τ.

$\beta^i_{\tau:\tau'} = (\beta^i_\tau, \ldots, \beta^i_{\tau'})$ is a *behavioral strategy* for i from time step τ to τ' (included).

When considering both players, the last 3 concepts become:

$\boldsymbol{\theta}_\tau = (\theta^1_\tau, \theta^2_\tau)$ ($\in \boldsymbol{\Theta} = \cup_{t=0}^{H-1}\boldsymbol{\Theta}_t$), a *joint* AOH at τ,

$\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$ ($\in \mathcal{B} = \cup_{t=0}^{H-1}\mathcal{B}_t$), a *decision rule profile*, and

$\boldsymbol{\beta}_{\tau:\tau'} = \langle \beta^1_{\tau:\tau'}, \beta^2_{\tau:\tau'} \rangle$, a *behavioral strategy profile.*

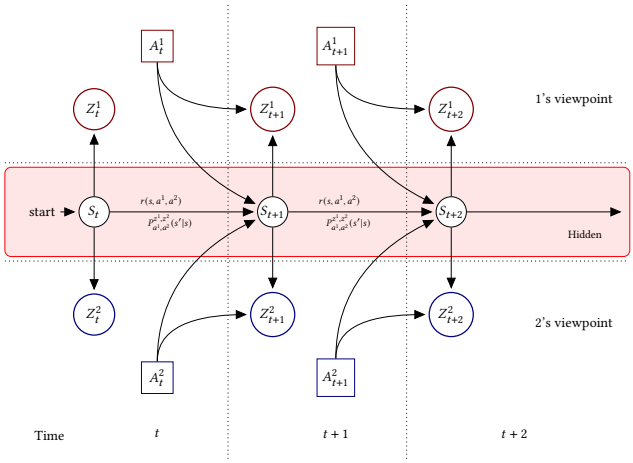

Figure 1: Dynamic influence diagram representing the evolution of a zs-POSG

Nash Equilibria. Here, player 1 (respectively 2) wants to maximize (resp. minimize) the expected return, or *value*, of strategy profile $\boldsymbol{\beta}_{0:H-1}$, defined as the discounted sum of future rewards, *i.e.*,

$$V_0(\boldsymbol{\beta}_{0:H-1}) = E\left[\sum_{t=0}^{H-1} \gamma^t R_t \mid \boldsymbol{\beta}_{0:H-1}\right],$$

where R_t is the random variable associated to the instant reward at t. This leads to the solution concept of Nash equilibrium strategy (NES).

Definition 2.2 (Nash Equilibrium)**.** The strategy profile $\boldsymbol{\beta}^*_{0:H-1} = \langle \beta^{1*}_{0:H-1}, \beta^{2*}_{0:H-1} \rangle$ is a NES if no player has an incentive to deviate, which can be written:

$$\forall \beta^1_{0:H-1},\ V_0(\beta^{1*}_{0:H-1}, \beta^{2*}_{0:H-1}) \geq V_0(\beta^1_{0:H-1}, \beta^{2*}_{0:H-1}) \text{ and}$$
$$\forall \beta^2_{0:H-1},\ V_0(\beta^{1*}_{0:H-1}, \beta^{2*}_{0:H-1}) \leq V_0(\beta^{1*}_{0:H-1}, \beta^2_{0:H-1}).$$

In such a game, all NESs have the same Nash-equilibrium value (NEV), $V^*_0 \stackrel{\text{def}}{=} V_0(\beta^{1*}_{0:H-1}, \beta^{2*}_{0:H-1})$. Our specific objective is to find an ϵ-NES, *i.e.*, a behavioral strategy profile such that any player would gain at most ϵ by deviating.

Why writing a Bellman Optimality Equation is Hard. Our approach requires writing Bellman optimality equations. The main obstacle to achieve this is to find an appropriate characterization of a *subproblem* that allows

(1) *predicting* both the immediate reward and the next possible subproblems given an immediate decision;
(2) *connecting* a subproblem's solution with solutions of its own (lower-level) subproblems; and
(3) *prescripting* a solution strategy for the subproblem built on solutions of lower-level subproblems.

In our setting, a player's AOH does not characterize a subproblem since her opponent's strategy is also required to predict the expected reward and the next AOHs. For their part, joint AOHs allow predicting next joint AOHs given both player's immediate decision rules, but would not be appropriate either, since player i cannot decide how

to act when facing some individual AOH θ^i_τ without considering all possible AOHs of his opponent $\neg i$.

If reasoning on all possible executions rather than one execution at a time, partial behavioral strategy profiles (sequences of behavioral decision rule profiles from $t = 0$ to some τ) contain enough information to completely describe the situation at τ, and are thus necessarily *predictive*. We still need to demonstrate that they are *connected*, despite decision rules not being public, and *prescriptive*, despite the need to address global-consistency issues illustrated in the following example.

Example 1. Matching pennies is a well-known 2-player zero-sum game in which each player has a penny and secretly chooses one side (head or tail). Then, both pennies sides are revealed, and player 1 wins (payoff of +1) if both chosen sides match and looses (payoff of −1) if not.

We here formalize this game as a zs-POSG (as illustrated in Figure 2) where player 1 actually picks his action at $t = 0$, and player 2 at $t = 1$. Hence the tuple $\langle \mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, \mathcal{Z}^1, \mathcal{Z}^2, P, r, H, \gamma, b_0 \rangle$ where:

- $\mathcal{S} = \{s_i, s_h, s_t\}$, where s_i is the initial state, and s_h and s_t represent a memory of 1's last move: respectively "head" or "tail";
- $\mathcal{A}^1 = \mathcal{A}^2 = \{a_h, a_t\}$ for playing "head" (a_h) or "tail" (a_t);
- $\mathcal{Z}^1 = \mathcal{Z}^2 = \{z_n\}$ a "none" trivial observation;
- $P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) = T(s, \boldsymbol{a}, s') \cdot O(\boldsymbol{a}, s', \boldsymbol{z})$, using the next two definitions;
- T is deterministic and such that ($\cdot$ is used to denote "for all")
  - $T(\cdot, \cdot, a_h) = s_h$,
  - $T(\cdot, \cdot, a_t) = s_t$;
- O is deterministic and always returns "z_n";
- r is such that
  - $r(s_i, \cdot, \cdot) = 0$,
  - $r(s_t, \cdot, a_t) = r(s_h, \cdot, a_h) = +1$,
  - $r(s_t, \cdot, a_h) = r(s_h, \cdot, a_t) = -1$;
- $H = 2$;
- $\gamma = 1$;
- b_0 is such that the initial state is s_i with probability 1.

Let us then assume that both players' DRs at $t = 0$ are fixed, with β^1_0 randomly picking a_t or a_h (*i.e.*, it induces a NES whatever his DR at $t = 1$), so that, at $t = 1$, we face a "subgame" where any strategy profile $\langle \beta^1_{1:1}, \beta^2_{1:1} \rangle$ is a NES profile with Nash equilibrium value 0. In particular, 2 can pick a deterministic strategy $\beta^2_{1:1}$, which will be said to be *locally consistent*. Yet, for 2, such a NES in the subgame at $\tau = 1$ is not necessarily *globally consistent*, *i.e.*, it may not be part of a NES for the original game (*i.e.*, starting from $\tau = 0$). Intuitively, in such global-consistency issues [18, 25] (also called *safety issues* [8]), the choices made at latter time steps do not account for possible deviations from the opponent at earlier time steps.

As detailed in the next section, we will characterize a subproblem not with the raw data of partial strategy profiles, but with a sufficient statistic, and this characterization will be used as the state of a new dynamic game equivalent to the zs-POSG.

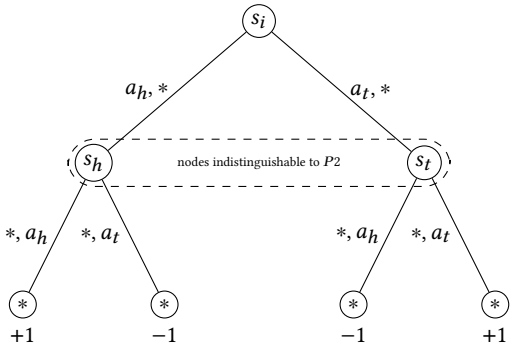

Figure 2: Simplified tree representation of the sequentialized Matching Pennies game. Irrelevant actions, noted $*$, allow merging edges with the same action for (i) player 2 at $t = 0$, and (ii) player 1 at $t = 1$. Notes: (a) Due to irrelevant actions, this game can be seen as an Extensive Form Game, despite players acting simultaneously. (b) Players only know about their past action history (in this observation-free game).

2.2 Occupancy State and Occupancy Markov Game

We now introduce an equivalent game, in which trajectories correspond to behavioral strategy profiles, and which we will be able to decompose temporally (and recursively), a first key tool for DP and HS.

To cope with the necessarily continuous nature of its state space, we will set this game in occupancy space, *i.e.*, a statistic that sums up past DR profiles. This will let us derive continuity properties on which to build point-based approximators.

As Wiggers et al. [33], let us formally define an *occupancy state* (OS) $\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$ as the probability distribution over joint AOHs $\boldsymbol{\theta}_\tau$ given partial strategy profile $\boldsymbol{\beta}_{0:\tau-1}$. This statistic exhibits the usual Markov and sufficiency properties:

Proposition 2.3 (Adapted from Dibangoye et al. [12, Thm. 1] – Proof in [11] (App. B.1)). *$\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$, together with $\boldsymbol{\beta}_\tau$, is a sufficient statistic to compute (i) the next* OS, $T(\sigma_{\boldsymbol{\beta}_{0:\tau-1}}, \boldsymbol{\beta}_\tau) \overset{\text{def}}{=} \sigma_{\boldsymbol{\beta}_{0:\tau}}$, *and (ii) the expected reward at τ: $r(\sigma_{\boldsymbol{\beta}_{0:\tau-1}}, \boldsymbol{\beta}_\tau) \overset{\text{def}}{=} \mathbb{E}\left[R_\tau \mid \boldsymbol{\beta}_{0:\tau-1} \oplus \boldsymbol{\beta}_\tau\right]$, where $\oplus$ denotes a concatenation.*

Writing from now on σ_τ, as short for $\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$, the OS associated with some prefix strategy profile $\boldsymbol{\beta}_{0:\tau-1}$, the proof essentially relies on deriving the following formulas: $\forall \theta^1_\tau, a^1, z^1, \theta^2_\tau, a^2, z^2$,

$$\begin{aligned} T(\sigma_\tau, \boldsymbol{\beta}_\tau)((\theta^1_\tau, a^1, z^1), (\theta^2_\tau, a^2, z^2)) \\ &\overset{\text{def}}{=} Pr((\theta^1_\tau, a^1, z^1), (\theta^2_\tau, a^2, z^2) | \sigma_\tau, \boldsymbol{\beta}_\tau) \\ &= \beta^1_\tau(\theta^1_\tau, a^1)\beta^2_\tau(\theta^2_\tau, a^2)\sigma_\tau(\boldsymbol{\theta}_\tau) \sum_{s,s'} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) b(s|\boldsymbol{\theta}_\tau), \end{aligned} \tag{1}$$

where $b(s|\boldsymbol{\theta}_\tau)$ is a *belief state* obtained by Hidden Markov Model filtering; and

$$\begin{aligned} r(\sigma_\tau, \boldsymbol{\beta}_\tau) &\overset{\text{def}}{=} E[r(S, A^1, A^2) | \sigma_\tau, \boldsymbol{\beta}_\tau] \\ &= \sum_{s, \boldsymbol{\theta}_\tau, \boldsymbol{a}} \sigma_\tau(\boldsymbol{\theta}_\tau) b(s|\boldsymbol{\theta}_\tau) \beta^1_\tau(a^1|\theta^1_\tau) \beta^2_\tau(a^2|\theta^2_\tau) r(s, \boldsymbol{a}). \end{aligned} \tag{2}$$

We can then derive, from a zs-POSG, a non-observable zero-sum game similar to Wiggers et al.'s *plan-time NOSG* [33, Definition 4], but without assuming that the players' past strategies are public.

Definition 2.4 (zero-sum occupancy Markov Game (zs-oMG))**.** A *zero-sum occupancy Markov game* (zs-oMG)[2] is defined by the tuple $\langle \mathcal{O}^\sigma, \mathcal{B}, T, r, H, \gamma \rangle$, where:

- $\mathcal{O}^\sigma (= \cup_{t=0}^{H-1} \mathcal{O}_t^\sigma)$ is the set of oss induced by the zs-POSG;
- $\mathcal{B}$ is the set of DR profiles of the zs-POSG;
- T is the deterministic transition function in Eq. (1);
- r is the reward function in Eq. (2); and
- H and γ are as in the zs-POSG

(b_0 is not in the tuple but serves to define T and r).

In this game, as in the zs-POSG, a player's solution is a behavioral strategy. Besides, the value of a strategy profile $\boldsymbol{\beta}_{0:H-1}$ is the same for both zs-oMG and zs-POSG, so that they share the same ϵ-NEV and ϵ-NESs. We can thus work with zs-oMGs as a means to solve zs-POSGs.

The following aims at deriving a recursive expression of V_0^*, as well as continuity properties.

Bellman Optimality Equation. Despite the OS at $\tau > 0$ not being accessible to any player, let us define a fictious *subgame* at σ_τ as the restriction starting from time step τ under this particular occupancy state, meaning that we are seeking strategies $\beta^1_{\tau:H-1}$ and $\beta^2_{\tau:H-1}$. σ_τ tells us which AOHs each player could be facing with non-zero probability, and are thus relevant for planning. We can then define the value function in any OS σ_τ for any strategy profile $\boldsymbol{\beta}_{\tau:H-1}$ as follows:

$$V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:H-1}) \stackrel{\text{def}}{=} E[\sum_{t=\tau}^{\infty} \gamma^{t-\tau} r(S_t, A_t) | \sigma_\tau, \boldsymbol{\beta}_{\tau:H-1}]. \tag{3}$$

The optimal value of a subgame rooted at σ_τ, $V^*(\sigma_\tau)$, is thus the unique NEV for the previous criterion[3]. Wiggers et al. then proved key continuity properties of V^* discussed next.

Concavity and Convexity Results. As a preliminary step, Wiggers et al. decompose an occupancy state σ_τ into a *marginal term* $\sigma_\tau^{m,1}$ and a *conditional term* $\sigma_\tau^{c,1}$, where

- $\sigma_\tau^{m,1}(\theta_\tau^1) = \sum_{\theta_\tau^2} \sigma_\tau(\theta_\tau^1, \theta_\tau^2)$ is the probability of 1 facing θ_τ^1 under σ_τ, and
- $\sigma_\tau^{c,1}(\theta_\tau^2 | \theta_\tau^1) = \frac{\sigma_\tau(\theta_\tau^1, \theta_\tau^2)}{\sigma_\tau^{m,1}(\theta_\tau^1)}$ is the probability of 2 facing θ_τ^2 under σ_τ given that 1 faces θ_τ^1,

so that $\sigma_\tau(\theta_\tau^1, \theta_\tau^2) = \sigma_\tau^{m,1}(\theta_\tau^1) \cdot \sigma_\tau^{c,1}(\theta_\tau^2 | \theta_\tau^1)$. (Symmetric definitions apply by swapping players 1 and 2.) In addition, let us denote $T_m^1(\sigma_\tau, \boldsymbol{\beta}_\tau)$ and $T_c^1(\sigma_\tau, \boldsymbol{\beta}_\tau)$ the marginal and conditional terms associated to $T(\sigma_\tau, \boldsymbol{\beta}_\tau)$.

Now, if 1 faces AOH θ_τ^1, knows 2's future strategy $\beta^2_{\tau:H-1}$, and has access to $\sigma_\tau^{c,1}(\theta_\tau^2 | \theta_\tau^1)$ for any θ_τ^2, then she faces a POMDP whose optimal value we denote $v^2_{[\sigma_\tau^{c,1}, \beta^2_{\tau:H-1}]}(\theta_\tau^1)$. This leads to defining the best-response value vector $v^2_{[\sigma_\tau^{c,1}, \beta^2_{\tau:H-1}]}$, which contains one component per AOH θ_τ^1, and writing the value of 1's best response against $\beta^2_{\tau:H-1}$ under σ_τ as $\sigma_\tau^{m,1} \cdot v^2_{[\sigma_\tau^{c,1}, \beta^2_{\tau:H-1}]}$. But then, because 2 also knows σ_τ, she can in fact pick $\beta^2_{\tau:H-1}$ to minimize this value, so that we get the following theorem.

Theorem 2.5 ([33, Thm. 2])**.** *For any $\tau \in \{0 \,..\, H-1\}$, V_τ^* is (i) concave w.r.t. $\sigma_\tau^{m,1}$ for a fixed $\sigma_\tau^{c,1}$, and (ii) convex w.r.t. $\sigma_\tau^{m,2}$ for a fixed $\sigma_\tau^{c,2}$. More precisely,*

$$V_\tau^*(\sigma_\tau) = \min_{\beta^2_{\tau:H-1}} \left[\sigma_\tau^{m,1} \cdot v^2_{[\sigma_\tau^{c,1}, \beta^2_{\tau:H-1}]} \right] = \max_{\beta^1_{\tau:H-1}} \left[\sigma_\tau^{m,2} \cdot v^1_{[\sigma_\tau^{c,2}, \beta^1_{\tau:H-1}]} \right],$$

where

$$v^2_{[\sigma_\tau^{c,1}, \beta^2_{\tau:H-1}]}(\theta_\tau^1) \stackrel{\text{def}}{=} \max_{\beta^1_{\tau:H-1}} \mathbb{E}_{\theta_\tau^2 \sim \sigma_\tau^{c,1}(\theta_\tau^1)} \left\{ \sum_{t=\tau}^{H-1} \gamma^{t-\tau} r(S_t, A_t^1, A_t^2) \mid \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1} \right\}. \tag{4}$$

An important observation that ensues from this theorem is that V_τ^* is concave in $\sigma_\tau^{m,1}$ and convex in $\sigma_\tau^{m,2}$. In practice, however, such continuity properties alone only allow upper-bounding V_τ^* for finitely many conditional terms $\sigma_\tau^{c,i}$, thus *not* for the whole occupancy space, as required to enable DP and HS in our game.

In the following, we complement Wiggers et al.'s results with properties of V^* in subgames, plus continuity properties that help designing bounding approximators, which will lead us to an HSVI-like solver.

For convenience, we may replace in the following: (i) subscript "$\tau : H-1$" with "$\tau :$", (ii) any function $f(\boldsymbol{x})$ linear in vector $\boldsymbol{x}$ with either $f(\cdot) \cdot \boldsymbol{x}$ or $\boldsymbol{x}^\top \cdot f(\cdot)$, (iii) a full tuple with its few elements of interest, and (iv) an element (a "field") x of a specific tuple t by $x[t]$.

3 THEORETICAL CONTRIBUTIONS

In this section, we demonstrate how to implement dynamic programming and heuristic search by (1) rigorously showing that Bellman optimality equation (Sec. 3.1) holds, (2) deriving bounding approximators of two novel optimal value functions, which come with solution strategies (Sec. 3.2), and (3) proposing a variant of HSVI that computes (in finite time) a player's strategy whose value is within ϵ of the zs-POSG's NEV (Sec. 3.3).

3.1 The Optimal Value Function V^* and its Recursive Expression

Let us first recall that, contrary to Wiggers et al. [33, Section 5, Lemma 4], we do not make the strong assumption that past decision rules can be considered as public (and, thus, we do not assume that any player knows σ_τ). Indeed, while it is valid in Dec-POMDPs because the players are willing to coordinate their behaviors, it is *a priori* not valid in zs-POSGs, since players are, in the contrary, willing to deceive one another. Safety issues as presented in Example 1 illustrate the possible flaws of such an assumption.

We now discuss the existence of an optimal value function V_τ^* and its properties. These results are implicitly used by Wiggers et al., but it seems important to state and demonstrate them. A first step is to demonstrate that von Neumann's minimax theorem [30] applies when in σ_τ, thus justifying the definition of the optimal (Nash equilibrium) value of a subgame.

[2] We use (i) "Markov game" instead of "stochastic game" because the dynamics are not stochastic, and (ii) "partially observable stochastic game" to stick with the literature.
[3] We will come back to the validity of this point in Section 3.1.

Theorem 3.1 (Minimax theorem – Proof in [11](App. C.1.2)). *The subgame defined in Eq.* (3) *admits a unique NEV*

$$V^*_\tau(\sigma_\tau) \overset{\text{def}}{=} \max_{\beta^1_{\tau:H-1}} \min_{\beta^2_{\tau:H-1}} V_\tau(\sigma_\tau, \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}). \tag{5}$$

$V(\sigma_\tau, \cdot, \cdot)$ not being bilinear in the space of behavioral strategies ([11] App. C.1.1), the proof requires reasoning with mixed strategies (for which the bilinearity holds), *i.e.*, distributions over pure (deterministic) strategies defined over *all* time steps. Yet, when in a subgame, we have to reason only on mixed strategies *compatible* with the associated occupancy state σ_τ (*i.e.*, which ensure that the os at τ is σ_τ), one step being to extend Kuhn's equivalence results between behavioral and mixed strategies [21] to the subgames.

Then, defining the optimal action-value function:

$$Q^*_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau) \overset{\text{def}}{=} r(\sigma_\tau, \boldsymbol{\beta}_\tau) + \gamma V^*_{\tau+1}(T(\sigma_\tau, \boldsymbol{\beta}_\tau)), \tag{6}$$

we can now prove that a Bellman optimality equation exists, which justifies reasoning on subgames despite the non-observability.

Theorem 3.2 (Bellman optimality equation – Proof in [11](App. C.1.2)). *$V^*_\tau(\sigma_\tau)$ satisfies the following functional equation:*

$$V^*_\tau(\sigma_\tau) = \max_{\beta^1_\tau} \min_{\beta^2_\tau} r(\sigma_\tau, \boldsymbol{\beta}_\tau) + \gamma V^*_{\tau+1}(T(\sigma_\tau, \boldsymbol{\beta}_\tau)) = \max_{\beta^1_\tau} \min_{\beta^2_\tau} Q^*_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau).$$

The proof requires decomposing min and max operators over different time steps before swapping them appropriately to end up recognizing the optimal value function at the next time step ($V^*_{\tau+1}$).

Theorems 3.1 and 3.2 together show that Theorem 2.5 holds even without player's strategies being public, so that we can now build on the convex-concave property to solve zs-oMGs.

3.2 Towards Solving zs-OMGs

This section aims at providing the second tool for DP and HS with continuous state spaces, *i.e.*, bounding approximators of optimal value functions which will allow generalization across occupancy space. Their update and selection operators are written as linear programs, and they turn out to come with solution strategies.

3.2.1 Bounding value functions. So far, several issues prevented to apply the HSVI scheme to zs-POSGs, starting with the continuous spaces of 1. occupancy states (zs-OMG states) and 2. decision rules (zs-OMG actions). One can address (1) by introducing the bounding functions $\overline{V}_\tau(\sigma_\tau)$ and $\underline{V}_\tau(\sigma_\tau)$ of $V^*_\tau(\sigma_\tau)$ (*cf.* [11], App. D.2), for instance:

$$\overline{V}_\tau(\sigma_\tau) = \min_{\langle \tilde{\sigma}^{c,1}_\tau, \langle \overline{v}^2_\tau, \beta^2_{\tau:}\rangle\rangle \in \overline{\mathcal{J}}_\tau} \left[\sigma^{m,1}_\tau \cdot \overline{v}^2_\tau + \lambda_\tau \|\sigma_\tau - \sigma^{m,1}_\tau \tilde{\sigma}^{c,1}_\tau\|_1 \right],$$

where $\overline{v}^2_\tau$ component-wise upper-bounds $v^2_{[\tilde{\sigma}^{c,1}_\tau, \beta^2_{\tau:}]}$ for some $\beta^2_{\tau:}$. They allow generalizing knowledge from the subgame rooted at σ_τ to any other one rooted at $\tilde{\sigma}_\tau$. To do so, we use V^*'s Lipschitz-Continuity proven below.

Theorem 3.3 (Lipschitz-Continuity of V^* - Proof in [11](App. D.1.3)). *Let $h_\tau \overset{\text{def}}{=} \frac{1-\gamma^{H-\tau}}{1-\gamma}$ (or $h_\tau \overset{\text{def}}{=} H-\tau$ if $\gamma = 1$). Then $V^*_\tau(\sigma_\tau)$ is λ_τ-Lipschitz continuous in σ_τ at any depth $\tau \in \{0\,..\,H-1\}$, where $\lambda_\tau = \frac{1}{2} h_\tau (r_{\max} - r_{\min})$.*

Yet, this yields (generally non-convex) Lipschitz-continuous functions whose max-min optimization would be intractable, so that (2) remains an issue. Also, we do not know how to retrieve valid solution strategies. In particular, and as illustrated in Example 1, simply concatenating decision rules backwards from $\tau = H-1$ to 0 would not guarantee globally-consistent solutions, and could result in exploitable strategies.

But then, combining Theorems 2.5 and 3.2 leads to introducing a novel value function (denoted $W^{1,*}_\tau$) through writing, for any os σ_τ:

$$V^*_\tau(\sigma_\tau) = \max_{\beta^1_\tau} \underbrace{\min_{\beta^2_{\tau:H-1} \in \mathcal{B}^2_\tau} \left[r(\sigma_\tau, \boldsymbol{\beta}_\tau) + \gamma \sigma^{m,1}_{\tau+1} \cdot v^2_{[\sigma^{c,1}_{\tau+1}, \beta^2_{\tau+1:H-1}]} \right]}_{\overset{\text{def}}{=} W^{1,*}_\tau(\sigma_\tau, \beta^1_\tau)}.$$

Assuming that player 2 can only respond with one of finitely many stored strategies, the concavity and λ_τ-Lipschitz-continuity of $W^{1,*}_\tau$ allow upper-bounding it with finitely many tuples $w = \langle \tilde{\sigma}_\tau, \beta^2_\tau, \langle \overline{v}^2_{\tau+1}, \beta^2_{\tau+1:}\rangle\rangle$ stored in sets $\overline{\mathcal{I}}_\tau$, and where $\overline{v}^2_{\tau+1}$ upper-bounds $v^2_{[\tilde{\sigma}^{c,1}_{\tau+1}, \beta^2_{\tau+1:}]}$.

Proposition 3.4 (Proof in [11](App. D.2.2)). *Let $\overline{\mathcal{I}}_\tau$ be a set of tuples $w = \langle \tilde{\sigma}_\tau, \beta^2_\tau, \langle \overline{v}^2_{\tau+1}, \beta^2_{\tau+1:}\rangle\rangle$. Then,*

$$\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \overset{\text{def}}{=} \min_{\langle \tilde{\sigma}_\tau, \beta^2_\tau, \langle \overline{v}^2_{\tau+1}, \beta^2_{\tau+1:}\rangle\rangle \in \overline{\mathcal{I}}_\tau} \Big[r(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) + \gamma T^1_m(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) \cdot \overline{v}^2_{\tau+1} + \lambda_{\tau+1} \|T(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) - T^1_m(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) T^1_c(\tilde{\sigma}^{c,1}_\tau, \beta^2_\tau)\|_1 \Big] \tag{7}$$

upper-bounds $W^{1,*}_\tau$ over the whole space $\mathcal{O}^\sigma_\tau \times \mathcal{B}^1_\tau$.

Symmetrically, we define $\underline{W}_\tau$ as the lower bound of the symmetrically defined $W^{2,*}_\tau$. As explained in the next two sections, $\overline{W}_\tau$ will be easier to deal with compared to $\overline{V}_\tau$, allowing 1 to seek for decision rules optimistically, and providing valid solution strategies for 2 for the subgame at τ, *i.e.*, ignoring consistency with higher-level subgames.

3.2.2 Action Selection and Backup Operators. We now detail the decision rule selection for 1 using $\overline{W}_\tau$ to optimistically guide a trajectory in occupancy space, and how to update $\overline{W}_\tau$ by providing backup operations.

To that end, first note that linearities in β^1_τ within Eq. (7) allow writing $\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) = \min_{w \in \overline{\mathcal{I}}_\tau} \beta^{1\top}_\tau \cdot M^{\sigma_\tau}_{(\cdot,w)}$, where β^1_τ and $M^{\sigma_\tau}_{(\cdot,w)}$ (for each w) are column vectors of dimension $|\Theta^1 \times \mathcal{A}^1|$. M^{σ_τ} (see developed formula in [11], App. D.3.1) is thus a $|\Theta^1_\tau \times \mathcal{A}^1| \times |\overline{\mathcal{I}}_\tau|$ matrix. Then, $\overline{W}_\tau$ being a lower envelope of hyperplanes leads to a convenient way of computing $\max_{\beta^1_\tau} \overline{W}_\tau(\sigma_\tau, \beta^1_\tau)$.

Proposition 3.5 (Proof in [11](App. D.3.1)). *For any given σ_τ and any set $\overline{\mathcal{I}}_\tau$ of tuples $w = \langle \tilde{\sigma}_\tau, \beta^2_\tau, \langle \overline{v}^2_{\tau+1}, \beta^2_{\tau+1:}\rangle\rangle$, $\max_{\beta^1_\tau} \overline{W}_\tau(\sigma_\tau, \beta^1_\tau)$ is*

equivalent to the LP and dual LP:

$$
\begin{array}{lllllll}
\text{LP}\overline{W}_\tau(\sigma_\tau) : & \max\limits_{\beta^1_\tau, v} v & \textit{s.t.} & \textit{(i)} & \forall w \in \overline{\mathcal{I}}_\tau, & v \le \beta^{1\top}_\tau \cdot M^{\sigma_\tau}_{(\cdot,w)} \\
& \textit{and} & & \textit{(ii)} & \forall \theta^1_\tau \in \Theta^1_\tau, & \sum\limits_{a^1} \beta^1_\tau(a^1|\theta^1_\tau) = 1, \\
\text{DLP}\overline{W}_\tau(\sigma_\tau) : & \min\limits_{\psi^2_\tau, v} v & \textit{s.t.} & \textit{(i)} & \forall (\theta^1_\tau, a^1), & v \ge M^{\sigma_\tau}_{((\theta^1_\tau,a^1),\cdot)} \cdot \psi^2_\tau \\
& \textit{and} & & \textit{(ii)} & & \sum\limits_{w \in \overline{\mathcal{I}}_\tau} \psi^2_\tau(w) = 1.
\end{array} \tag{8}
$$

Remark 3.6 (Outcomes of this game). Since $\overline{W}_\tau$ upper-bounds $W^{1,*}_\tau$, solving this LP provides 1 with an *optimistically* selected immediate decision rule β^1_τ. For 2, ψ^2_τ is a probability distribution over tuples containing strategies $\beta^2_\tau \oplus \beta^2_{\tau+1:H-1}$, thus recursively induces a strategy, which can be turned into a behavioral strategy $\beta^2_{\tau:H-1}$ (more details in [11], App. D.3.3) whose value is *at worst* (from 2's viewpoint) the LP's value, *i.e.*, against 1's best response to it.

Then, the following properties allow performing backups, *i.e.*, filling up the set $\overline{\mathcal{I}}_{\tau-1}$ with new tuples w containing, in particular, vectors $\overline{v}^2_\tau$.

Lemma 1 (Proof in [11](App. D.3.2))**.** *For any $\psi^2_\tau = \text{DLP}\overline{W}_\tau(\sigma_\tau)$, the vector $v^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]}$ is component-wise upper-bounded by*

$$\overline{v}^2_\tau \overset{\text{def}}{=} \frac{1}{\sigma^{m,1}_\tau} M^{\sigma_\tau}_{((\theta^1_\tau,a^1),\cdot)} \cdot \psi^2_\tau.$$

Proposition 3.7 (update)**.** *Let us assume that*

- *a transition $\sigma_{\tau-1} \to \sigma_\tau$ has been performed through playing $\langle \beta^1_{\tau-1}, \beta^2_{\tau-1} \rangle$, and*
- *solving $\text{DLP}\overline{W}_\tau(\sigma_\tau)$ provides both*
  - *a tree strategy ψ^2_τ (as the main solution of the DLP), and*
  - *a vector $\overline{v}^2_\tau = \frac{1}{\sigma^{m,1}_\tau} M^{\sigma_\tau}_{((\theta^1_\tau,a^1),\cdot)} \cdot \psi^2_\tau$ (as a by-product).*

Then,

(1) $\overline{\mathcal{I}}_{\tau-1} \leftarrow \overline{\mathcal{I}}_{\tau-1} \cup \{\langle \sigma^{c,1}_{\tau-1}, \beta^2_{\tau-1}, \langle \overline{v}^2_\tau, \psi^2_\tau \rangle \rangle\}$ is a valid update operator in the sense that it preserves $\overline{W}_\tau$'s upper-bounding property, and

(2) similarly, $\overline{\mathcal{J}}_\tau \leftarrow \overline{\mathcal{J}}_\tau \cup \{\langle \sigma^{c,1}_\tau, \langle \overline{v}^2_\tau, \psi^2_\tau \rangle \rangle\}$ is a valid update operator for $\overline{V}_\tau$.

3.2.3 Initialization. To initialize the bounds $\overline{W}_\tau$ and $\overline{V}_\tau$ for any time step, we begin by generating a trajectory in a forward phase. At each time step, a uniform decision rule is picked for both players to derive a sequence of occupancy states $\sigma_0, \dots, \sigma_{H-1}$. Then, during a backward phase, for each time step $\tau = H-1, \dots, 1$, we create a tuple $w_{\tau-1,init} = \langle \sigma^{c,1}_{\tau-1}, \beta^2_{\tau-1}, \langle \overline{v}^2_\tau, \psi^2_\tau \rangle \rangle$, where

- $\sigma^{c,1}_{\tau-1}$ is the conditional term associated to $\sigma_{\tau-1}$;
- $\beta^2_{\tau-1}$ is a uniform decision rule;
- ψ^2_τ is
  - a degenerate distribution over the only next tuple $w_{\tau+1}$ if $\tau < H-1$ (which induces a concatenation of uniform decision rules for all future time steps);
  - undefined if $\tau = H-1$;

  and
- $\overline{v}^2_\tau(\theta^1_\tau) = r_{max} \cdot (H - \tau)$ for any history θ^1_τ that player 1 could face.

Tuples $w_{\tau-1,init}$ are added to sets $\overline{\mathcal{I}}_{\tau-1}$. For any time step $\tau \ge 0$, we similarly create tuples $\langle \sigma^{c,1}_\tau, \langle \overline{v}^2_\tau, \psi^2_{\tau,:} \rangle \rangle$ and add them to sets $\overline{\mathcal{J}}_\tau$. The lower bounds are initialized symmetrically.

We now show that occupancy states can also be prescriptive, allowing one to retrieve an ϵ-NES for the subgame at occupancy state σ_τ once the bounds are within ϵ from each other, in particular at $\tau = 0$.

3.2.4 Extracting a NES. Vectors $\overline{v}^2_0$ upper bounding the value of their associated strategies, the following result tells when and how to extract an ϵ-optimal solution strategy for this player.

Theorem 3.8. *If sets $\overline{\mathcal{J}}_0$ and $\underline{\mathcal{J}}_0$ are such that $\overline{V}_0(\sigma_0) - \underline{V}_0(\sigma_0) \le \epsilon$, then $\arg\max_{\underline{w} \in \underline{\mathcal{J}}_0} \underline{v}^2_0$ and $\arg\min_{\overline{w} \in \overline{\mathcal{J}}_0} \overline{v}^2_0$ respectively provide strategies ψ^1_0 and ψ^2_0 that form an ϵ-NES of the zs-POSG.*

Proof. First, let us notice that, at $\tau = 0$, the occupancy-state space is reduced to a singleton, $\{\sigma_0 = \langle 1 \rangle\}$, because of the single (empty) joint AOH. The value vectors v are thus one-dimensional, and here considered as scalar numbers.

Let us assume that sets $\overline{\mathcal{J}}_0$ and $\underline{\mathcal{J}}_0$ are such that

$$\overline{V}_0(\sigma_0) - \underline{V}_0(\sigma_0) \le \epsilon,$$

and let $\underline{w}^* = \langle \sigma^{c,1}_0, \langle \underline{v}^*_0, \psi^{1,*}_0 \rangle \rangle$ and $\overline{w}^* = \langle \sigma^{c,1}_0 \langle \overline{v}^*_0, \psi^{2,*}_0 \rangle \rangle$ be the tuples returned by $\arg\max_{\underline{w} \in \underline{\mathcal{J}}_0} \underline{v}^2_0$ and $\arg\min_{\overline{w} \in \overline{\mathcal{J}}_0} \overline{v}^1_0$. Then, noting that $\sigma_0 = \langle 1 \rangle$,

$$
\begin{aligned}
v^1_{[\sigma^{c,2}_0, \psi^{1,*}_0]} - v^2_{[\sigma^{c,1}_0, \psi^{2,*}_0]} &\le \overline{v}^*_0 - \underline{v}^*_0 \\
&= \max_{\underline{w} \in \underline{\mathcal{J}}_0} \underline{v}^2_0 - \min_{\overline{w} \in \overline{\mathcal{J}}_0} \overline{v}^1_0 \\
&= \overline{V}_0(\sigma_0) - \underline{V}_0(\sigma_0) \\
&\le \epsilon.
\end{aligned}
$$

Thus, ψ^1_0 and ψ^2_0 are two strategies whose security levels (values against best-responding opponents) are ϵ-close, and thus form an ϵ-NES of the zs-POSG. □

Note: This result can be generalized to any σ_τ at later time steps, but this generalization is not used in practice.

Distributions ψ^2_0 are stored and could be executed as is. [11] (App. D.3.3) still presents a conversion process to retrieve a behavioral strategy $\beta^2_{0:H-1}$ from a distribution ψ^2_0 over tuples $w \in \overline{\mathcal{I}}_0$. Next, we see how to design a practical HSVI-based algorithm that provably returns sets $\overline{\mathcal{J}}_0$ and $\underline{\mathcal{J}}_0$ satisfying Theorem 3.8 after finitely many iterations.

3.3 HSVI for zs-POSGs

This section details our adaptation of the general HSVI scheme for ϵ-optimally solving zs-POSGs, and presents a theoretical finite-time convergence property.

3.3.1 Algorithm. HSVI for zs-POSGs is described in Algorithm 1. As vanilla HSVI, it relies on (i) generating trajectories while acting optimistically (lines 10+11), *i.e.*, player 1 (resp. 2) acting "greedily" w.r.t. $\overline{W}_\tau$ (resp. $\underline{W}_\tau$), and (ii) locally updating the upper and lower

bounds (lines 17+18). Both phases rely on solving the same games described by LP (8). At $\tau = H-1$, line 14 selects DRs by solving an exact game, and line 20 returns a distribution reduced to the single element just added in line 15.

A key difference with Smith and Simmons's HSVI algorithm [27] lies in the criterion for stopping trajectories. The branching factor for zs-oMGs being infinite, we make use of V^*'s Lipschitz-continuity to implement the same adaptations as [16] used for zs-OS-POSGs. The Lipschitz-continuity allows controlling the variations of the value function within small balls of radius ρ around a previously visited occupancy-state. A finite number of such balls is sufficient to cover the whole space. Then, Theorem 3.9 (below) ensures ϵ-optimality in finite time if stopping trajectories when $\overline{V}_\tau(\sigma_\tau) - \underline{V}_\tau(\sigma_\tau) \leq thr(\tau)$, with the threshold function $thr(\tau) \stackrel{\text{def}}{=} \gamma^{-\tau}\epsilon - \sum_{i=1}^{\tau} 2\rho\lambda_{\tau-i}\gamma^{-i}$.

Algorithm 1: zs-oMG-HSVI(b_0, $[\epsilon, \rho]$)
[here returning a tuple w_0 containing a solution strategy $\underline{\psi}_0^1$ for player 1]

1 **Fct** `zs-oMG-HSVI`($b_0 \simeq \sigma_0$)
2 **foreach** $\tau \in 0\,..\,H-1$ **do**
3 Initialize $\overline{V}_\tau$, $\underline{V}_\tau$, $\overline{W}_\tau$, & $\underline{W}_\tau$
4 **while** $\left[\overline{V}_0(\sigma_0) - \underline{V}_0(\sigma_0) > thr(0)\right]$ **do**
5 `Explore`($\sigma_0, 0, -, -$)
6 **return** $\arg\max_{w_0 \in \underline{\mathcal{J}}_0} \underline{v}_0^1$
7 **Fct** `Explore`($\sigma_\tau, \tau, \sigma_{\tau-1}, \boldsymbol{\beta}_{\tau-1}$)
8 **if** $\left[\overline{V}_\tau(\sigma_\tau) - \underline{V}_\tau(\sigma_\tau) > thr(\tau)\right]$ **then**
9 **if** $\tau < H-1$ **then**
10 $\overline{\beta}_\tau^1 \leftarrow \text{LP}\overline{W}_\tau(\sigma)$
11 $\underline{\beta}_\tau^2 \leftarrow \text{LP}\underline{W}_\tau(\sigma)$
12 `Explore`$(T(\sigma_\tau, \overline{\beta}_\tau^1, \underline{\beta}_\tau^2), \tau+1, \sigma_\tau, \langle \overline{\beta}_\tau^1, \underline{\beta}_\tau^2 \rangle)$
13 **else** ($\tau = H-1$)
14 $(\overline{\beta}_\tau^1, \underline{\beta}_\tau^2) \leftarrow NES\left(r(\sigma, \beta_\tau^1, \beta_\tau^2)\right)$
15 $\overline{\mathcal{I}}_\tau^1 \leftarrow \overline{\mathcal{I}}_\tau^1 \cup \{\langle \sigma_\tau^{c,1}, \underline{\beta}_\tau^2, - \rangle\}$
16 $\underline{\mathcal{I}}_\tau^2 \leftarrow \underline{\mathcal{I}}_\tau^2 \cup \{\langle \sigma_\tau^{c,2}, \overline{\beta}_\tau^1, - \rangle\}$
17 $\overline{\textbf{Update}}(\overline{W}_{\tau-1}, \langle \sigma_\tau, \sigma_{\tau-1}^{c,1}, \underline{\beta}_{\tau-1}^2 \rangle)$
18 $\underline{\textbf{Update}}(\underline{W}_{\tau-1}, \langle \sigma_\tau, \sigma_{\tau-1}^{c,2}, \overline{\beta}_{\tau-1}^1 \rangle)$
19 **Fct** $\overline{\textbf{Update}}(\overline{W}_{\tau-1}, \langle \sigma_\tau, \sigma_{\tau-1}^{c,1}, \overline{\beta}_{\tau-1}^2 \rangle)$
20 $\langle \overline{v}_\tau^2, \psi_\tau^2 \rangle \leftarrow \text{DLP}\overline{W}_\tau(\sigma_\tau,)$
21 $\overline{\mathcal{I}}_{\tau-1} \leftarrow \overline{\mathcal{I}}_{\tau-1} \cup \{\langle \sigma_{\tau-1}^{c,1}, \underline{\beta}_{\tau-1}^2, \langle \overline{v}_\tau^2, \psi_\tau^2 \rangle \rangle\}$
22 $\overline{\mathcal{J}}_\tau \leftarrow \overline{\mathcal{J}}_\tau \cup \{\langle \sigma_\tau^{c,1}, \langle \overline{v}_\tau^2, \psi_\tau^2 \rangle \rangle\}$

Setting ρ. As can be observed, this threshold function should always return positive values, which requires a small enough (but > 0) ρ. For a given problem (*cf.* [11], Proposition E.1, App. E.1), the maximum possible value $\rho_{\max}$ depends on the Lipschitz constants at each time step, which themselves depend on the initial upper and lower bounds of the optimal value function. Setting $\rho \in (0, \rho_{\max})$ means making a trade-off between generating many trajectories (small ρ) and long ones (large ρ).

3.3.2 *Finite-Time Convergence.*

Theorem 3.9 (Proof in [11](App. E.2.1)). *zs-oMG-HSVI (Alg. 1) terminates in finite time with an ϵ-approximation of $V_0^*(\sigma_0)$ that statisfies Theorem 3.8.*

The finite time complexity suffers from the same combinatorial explosion as for Dec-POMDPs, and is even worse as we have to handle "infinitely branching" trees of possible futures. More precisely, the bound on the number of iterations depends on the number of balls of radius ρ required to cover occupancy simplexes at each depth.

Also, the following proposition allows solving infinite horizon problems as well (when $\gamma < 1$) by bounding the length of HSVI's trajectories using the boundedness of $\overline{V} - \underline{V}$ and the exponential growth of $thr(\tau)$.

Proposition 3.10 (Proof in [11](App. E.2.2)). *When $\gamma < 1$, the length of trajectories is upper bounded by $T_{\max} \stackrel{\text{def}}{=} \left\lceil \log_\gamma \frac{\epsilon - \frac{2\rho\lambda^\infty}{1-\gamma}}{W - \frac{2\rho\lambda^\infty}{1-\gamma}} \right\rceil$, where λ^∞ is a depth-independent Lipschitz constant and $W \stackrel{\text{def}}{=} \|\overline{V}^{(0)} - \underline{V}^{(0)}\|_\infty$ is the maximum width between initializations.*

4 EXPERIMENTS

Experiments presented in this section aim at validating the proposed approach and comparing its behavior to the behavior of some reference algorithms.

4.1 Setup

Benchmark Problems. Five benchmark problems were used. Adversarial Tiger and Competitive Tiger were introduced by Wiggers [32]. Mabc and Recycling Robot are well-known Dec-POMDP benchmark problems (*cf.* http://masplan.org) and were adapted to our competitive setting by making player 2 minimize (rather than maximize) the objective function. The fifth benchmark is the adaptation of the well-known Matching Pennies game detailed in Example 1, with a small difference in that $r(s_h, \cdot, a_h) = +2$ instead of $+1$; this change breaks the symmetry in the optimal strategy, so that HSVI can not find the NES by "chance" by trying uniform strategies. We only consider finite horizons H and $\gamma = 1$.

Algorithms. For conciseness, Algorithm 1 is here denoted HSVI, and compared against *Random* search and *Informed* search [32] (both using Wiggers's implementation (unlicensed and unreleased)), SFLP [17], and CFR+ [29] (both using open_spiel [22] (Apache license)).

All algorithms (but SFLP, which is exact) used a target error $\epsilon = 1\%$ of the initial gap $H \cdot (r_{\max} - r_{\min})$. HSVI ran with $\lambda_\tau = (H - \tau) \cdot (r_{\max} - r_{\min})$, and ρ the middle of its feasible interval. We also use FB-HSVI's LPE lossless compression of probabilistically equivalent action-observation histories in occupancy states, so as to reduce their dimensionality [12]. Experiments ran on an Ubuntu machine with i7-10810U 1.10 GHz Intel processor and 16 GB available RAM, and the code is available under MIT license at https://gitlab.com/aureliendelage1/hsviforzsposgs.

Random and Informed, only ran once, providing fairly representative results.

4.2 Results

Performance Measures. A common performance measure in 2-player zero-sum games is the *exploitability* of a strategy $\beta^i_{0:}$, *i.e.*, the difference between the strategy's *security level* (the value of $\neg i$'s best response to $\beta^i_{0:}$) and the Nash equilibrium value $V^*_0(\sigma_0)$:

$$\begin{aligned}\text{exploitability}(\beta^i_{0:}) &= |V^*(\sigma_0) - \sigma_0^{m,1} \cdot v^i_{[\sigma_0^{c,\neg i}, \beta^i_{0:}]}| \\ &= |V^*(\sigma_0) - v^i_{[\sigma_0^{c,\neg i}, \beta^i_{0:}]}|,\end{aligned}$$

noting that σ_0 is a degenerate distribution over a single element, the pair of empty action-observation histories. In our setting, it will be convenient to look at the (average) *exploitability of a strategy profile* $\langle \beta^1_{0:}, \beta^2_{0:} \rangle$:

$$\begin{aligned}\text{exploitability}(\beta^1_{0:}, \beta^2_{0:}) &= \frac{(V^*(\sigma_0) - v^1_{[\sigma_0^{c,2}, \beta^1_{0:}]}) + (v^2_{[\sigma_0^{c,1}, \beta^2_{0:}]} - V^*(\sigma_0))}{2} \\ &= \frac{v^2_{[\sigma_0^{c,1}, \beta^2_{0:}]} - v^1_{[\sigma_0^{c,2}, \beta^1_{0:}]}}{2}.\end{aligned}$$

This quantity is a more concise statistic than both individual exploitabilities, and can be obtained by solving two POMDPs (fixing one player's strategy or the other) without requiring to know the actual NEV.

This exploitability can also be defined as half of the *gap between security levels* (SL-gap). To analyze the convergence of algorithms with respect to the initial gap, we will look at the *SL-gap percentage*, *i.e.*,

$$\begin{aligned}\text{SL-gap percentage}(\beta^1_{0:}, \beta^2_{0:}) &= \frac{v^2_{[\sigma_0^{c,1}, \beta^2_{0:}]} - v^1_{[\sigma_0^{c,2}, \beta^1_{0:}]}}{H \cdot (R_{\max} - R_{\min})} \\ &= \frac{2 \cdot \text{exploitability}(\beta^1_{0:}, \beta^2_{0:})}{H \cdot (R_{\max} - R_{\min})}.\end{aligned}$$

4.2.1 Comparison with the state of the art. Table 1 gives the convergence time of Wiggers's two heuristic algorithms, CFR+, SFLP, and HSVI on the benchmark problems with various horizons, or the SL-gap percentage when reaching a 1 h time limit. Executions not returning any result (*i.e.*, for Random, Informed and CFR+, not performing a single iteration) are noted out-of-time [OOT].

This table first shows that HSVI always outperforms the heuristic baseline provided by Wiggers's algorithms, thus proving the interest of an HSVI scheme. However, HSVI is outperformed by both SFLP and CFR+, unless they run out of time. As can be noted, HSVI is able to keep improving even when the horizon grows thanks to the LPE compression, taking advantage of underlying structure in some games (*e.g.*, Recycling Robot, a problem with transition+observation independence (TOI), when scaling to larger horizons).

Additional bound and exploitability graphs can be found in [11].

Table 1: Comparison of different solvers on various benchmark problems. Reported values are the running times until the algorithm's error gap (based on bounds for HSVI) is lower than 1 %, or, if the timeout limit is reached, the security-level gap percentages (100 % if gap = $H \cdot (R_{\max} - R_{\min})$). Notes: (1) Horizons with a star exponent (H^*) are those for which the security-level computations ran out of time so that, for HSVI, we give the gap between the pessimistic bounds. (2) Even though Random and Informed contain randomness, we ran them only once, getting fairly representative results.

Domain	H	Wiggers		HSVI	SFLP	CFR+
		Rand.	Inf.			
Comp Tiger	2	2.6 %	8.3 %	6 s	**1** s	18 s
	3	7.0 %	6.1 %	3.8 %	**48** s	30 m
	4	12.1 %	7.7 %	4.8 %	**14** m	[OOT]
	5^*	[OOT]	[OOT]	53.3 %	[OOT]	[OOT]
Rec. Robot	2	3.4 %	5.1 %	5 s	**1** s	30 s
	3	9.2 %	15.2 %	4 m	**1** s	13 m
	4	14.1 %	19.6 %	4.9 %	**13** s	1.5 %
	5	[OOT]	[OOT]	**10.7 %**	[OOT]	[OOT]
	6^*	[OOT]	[OOT]	**45.5 %**	[OOT]	[OOT]
Adv Tiger	2	1 s	3.7 %	**1** s	**1** s	**1** s
	3	1.5 %	4.4 %	2 m	**1** s	8 s
	4	2.9 %	5.6 %	2.6 %	**8** s	13 m
MABC	2	45 s	18.8 %	8 s	**3** s	5 s
	3	4.2 %	9.2 %	27 s	**1** s	1 m
	4	18.1 %	36.3 %	4.4 %	**3** s	47 m
MP	4	2 m	46.7 %	5 s	**1** s	2 s
	5	9 m	45.8 %	1 m	**1** s	10 s
	6	2.2 %	44.6 %	8 m	**2** s	1 m

5 DISCUSSION

This paper addresses the problem of ϵ-optimally solving zs-POSGs. In contrast to SFLP or CFR+, we provide the necessary foundational building blocks to apply dynamic programming (in tandem with heuristic search) to solve zs-POSGs. We introduce Bellman optimality equations and uniform-continuity properties of the optimal value function. Next, we exhibit rules for updating value functions while preserving uniform continuity and the ability to extract globally-consistent solutions. Finally, we describe the first effective DP algorithm for zs-POSGs, zs-oMG-HSVI, with finite-time convergence to an ϵ-optimal solution. Experiments support our theoretical findings.

We believe our approach complements existing ones, e.g., SFLP and CFR+, in two dimensions. First, it breaks the original zs-POSG into subgames. Second, it generalizes values from visited subgames to unvisited ones. Our performances are as good as or better than those from SFLP and CFR+ for small-dimensional subgames (*e.g.*, with TOI structure). Unfortunately, the advantage of breaking the original problem into subgames and exploiting uniform continuity properties often fails to fully manifest in the overall computational time.

We hope that this approach will lay the foundation for further work in the area of both exact and approximate DP solutions for zs-POSGs. In the short term, we shall investigate pruning techniques, better Lipschitz constants, and improved initial bounding approximators using solutions from relaxations of zs-POSGs, e.g., zs-OS-POSGs. In the long term, we shall investigate (deep) RL for zs-POSGs, similarly to a recent approach for Dec-POMDPs [4]. The latter shall investigate the trade-off between the update-rule accuracy and the computational efficiency when facing high-dimensional subgames, hence providing competitive solvers. [11] also provides a discussion regarding the connexion between this work and those based on continual re-solving.

REFERENCES

[1] Karl Johan Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174 – 205, 1965. ISSN 0022-247X.

[2] Nicola Basilico, Giuseppe De Nittis, and Nicola Gatti. A security game combining patrolling and alarm–triggered responses under spatial and detection uncertainties. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[3] Arnab Basu and Lukasz Stettner. Finite- and infinite-horizon Shapley games with nonsymmetric partial observation. *SIAM Journal on Control and Optimization*, 53(6):3584–3619, 2015.

[4] Guillaume Bono, Jilles Dibangoye, Laëtitia Matignon, Florian Pereyron, and Olivier Simonin. Cooperative multi-agent policy gradient. In *Proceedings of the Twenty-Eight European Conference on Machine Learning*, 2018.

[5] Branislav Bošanský, Christopher Kiekintveld, Viliam Lisý, and Michal Pěchouček. An exact double-oracle algorithm for zero-sum extensive-form games with imperfect information. *Journal of Artificial Intelligence Research*, 51:829–866, 2014. doi: 10.1613/jair.4477.

[6] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[7] Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17057–17069, 2020.

[8] Neil Burch, Michael Johanson, and Michael Bowling. Solving imperfect information games using decomposition. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[9] Krishnendu Chatterjee and Laurent Doyen. Partial-observation stochastic games: How to win when belief fails. *ACM Transactions on Computational Logic*, 15(2): 16, 2014.

[10] Harold L. Cole and Narayama Kocherlakota. Dynamic games with hidden actions and hidden states. *Journal of Economic Theory*, 98(1):114–126, 2001.

[11] Aurélien Delage, Olivier Buffet, Jilles S Dibangoye, and Abdallah Saffidine. Hsvi can solve zero-sum partially observable stochastic games. *arXiv preprint arXiv:2210.14640*, 2022.

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

[13] Mrinal K. Ghosh, David R. McDonald, and Sagnik Sinha. Zero-sum stochastic games with partial information. *Journal of Optimization Theory and Applications*, 121(1):99–118, April 2004.

[14] Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing Nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512, 2010. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/40801239.

[15] Karel Horák and Branislav Bošanský. Solving partially observable stochastic games with public observations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 2029–2036, 2019.

[16] Karel Horák, Branislav Bošanský, and Michal Pěchouček. Heuristic search value iteration for one-sided partially observable stochastic games. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 558–564, 2017.

[17] Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and Economic Behavior*, 14(51):220–246, 1996.

[18] Vojtěch Kovařík, Dominik Seitz, Viliam Lisý, Jan Rudolf, Shuo Sun, and Karel Ha. Value functions for depth-limited solving in zero-sum imperfect-information games. *Artificial Intelligence*, page 103805, 2022. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2022.103805. URL https://www.sciencedirect.com/science/article/pii/S000437022200145X.

[19] Vojtěch Kovařík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisý. Rethinking formal models of partially observable multiagent decision making. *CoRR*, abs/1906.11110, 2019.

[20] Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 179:385–417, 2020. doi: 10.1007/s10107-018-1336-7.

[21] Harold W. Kuhn. Simplified two-person Poker. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume 1, 1950.

[22] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, abs/1908.09453, 2019.

[23] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit Poker. *Science*, 356(6337):508–513, 2017.

[24] Frans Oliehoek and Nikos Vlassis. Dec-POMDPs and extensive form games: equivalence of models and algorithms. Technical Report IAS-UVA-06-02, Intelligent Systems Laboratory Amsterdam, University of Amsterdam, 2006.

[25] Martin Schmid. *Search in Imperfect Information Games*. PhD thesis, Charles University - Univerzita Karlova, Prague, 2021. URL https://arxiv.org/pdf/2111.05884.pdf.

[26] Trey Smith. *Probabilistic Planning for Robotic Exploration*. PhD thesis, The Robotics Institute, Carnegie Mellon University, 2007.

[27] Trey Smith and R.G. Simmons. Point-based POMDP algorithms: Improved analysis and implementation. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 542–549, 2005.

[28] Daniel Szer, François Charpillet, and Shlomo Zilberstein. MAA*: A heuristic search algorithm for solving decentralized POMDPs. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 576–583, 2005.

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

[30] John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100, 1928. URL https://doi.org/10.1007/BF01448847.

[31] Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(50):220–246, 1996.

[32] Auke Wiggers. Structure in the value function of two-player zero-sum games of incomplete information. Master's thesis, University of Amsterdam, 2015.

[33] Auke Wiggers, Frans Oliehoek, and Diederik Roijers. Structure in the value function of two-player zero-sum games of incomplete information. *Computing Research Repository*, abs/1606.06888, 2016.

[34] Auke Wiggers, Frans Oliehoek, and Diederik Roijers. Structure in the value function of two-player zero-sum games of incomplete information. In *Proceedings of the Twenty-Second European Conference on Artificial Intelligence*, pages 1628–1629, 2016. doi: 10.3233/978-1-61499-672-9-1628.

[35] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20*, 2007.