On Bellman's Optimality Principle for zs-POSGs

Olivier Buffet, Jilles Dibangoye, Aurélien Delage, Abdallah Saffidine, Vincent Thomas

On Bellman's Optimality Principle for zs-POSGs

Olivier Buffet[1] Jilles Dibangoye[2] Aurélien Delage[1,2] Abdallah Saffidine[3] Vincent Thomas[1]

[1] Université de Lorraine, CNRS, INRIA, LORIA, F-54000 Nancy, France
[2] Univ. Lyon, INSA Lyon, INRIA, CITI, F-69621 Villeurbanne, France
[3] The University of New South Wales, Sydney, Australia

(firstname.lastname@inria.fr|abdallahs[@]cse.unsw.edu.au)

Abstract

Many non-trivial sequential decision-making problems are efficiently solved by relying on Bellman's optimality principle, i.e.*, exploiting the fact that sub-problems are nested recursively within the original problem. Here we show how it can apply to (infinite horizon) 2-player zero-sum partially observable stochastic games (zs-POSGs) by (i) taking a central planner's viewpoint, which can only reason on a sufficient statistic called occupancy state, and (ii) turning such problems into zero-sum occupancy Markov games (zs-OMGs). Then, exploiting the Lipschitz-continuity of the value function in occupancy space, one can derive a version of the HSVI algorithm (Heuristic Search Value Iteration) that provably finds an ϵ-Nash equilibrium in finite time.*

Keywords

POSG; partially observable stochastic game; Bellman's optimality principle; Heuristic Search Value Iteration.

1 Introduction

Bellman's optimality principle (Bellman's OP) [3] led to state-of-the-art solvers in many non-trivial sequential decision-making problems, assuming partial observability [25], multi-objective criteria [29, 21], collaborating agents, *e.g.*, modeled as decentralized partially observable Markov decision processes (Dec-POMDPs) [13, 30, 9], or some non-collaborative perfect information games (from Shapley's seminal work [26] to [6]). In all these settings this principle exploits the fact that sub-problems are nested recursively within the original problem. An open question is whether—and how—it could be applied to imperfect information games, which are encountered in diverse applications such as Poker [18] or security games [1]. This paper answers this question in the setting of 2-player zero-sum partially observable stochastic games (zs-POSGs), *i.e.*, imperfect information games with simultaneous moves, perfect recall, discounted rewards and a possibly infinite time horizon.

As general POSGs and Dec-POMDPs, infinite-horizon zs-POSGs are undecidable, and their finite-horizon approximations are in NEXP [22, 4]. As further discussed in Section 2, solution techniques for finite-horizon POSGs, or other impartial information games that can be formulated as extensive-form games (EFGs), typically solve an equivalent normal-form game [27] or use a dedicated regret-minimization mechanism [34, 5]. They thus do not rely on Bellman's optimality principle, except (i) a dynamic programming approach that only constructs sets of non-dominated solutions [13], (ii) in collaborative problems (Decentralized POMDPs), adopting the viewpoint of a (blind) central planner [30, 9], and (iii) for (mostly 2-player zero-sum) settings with observability assumptions such that one can reason on player beliefs [12, 7, 2, 15, 8, 14]. Here, we do not make any assumption beyond the game being 2-player zero-sum, in particular regarding observability of the state and actions.

As for a number of Dec-POMDP solvers, our approach adopts the viewpoint not of a player, but of a central (offline) planner that prescribes individual strategies to the players [30], which allows turning a zs-POSG into a non-observable game for which Bellman's optimality principle applies. This is achieved in Section 4 (after background Section 3) while reasoning not on a player's belief over the game state (as feasible in POMDPs or some particular games), but on the central planner's (blind) belief, a statistic called occupancy state and that we prove to be sufficient for optimal planning, as Dibangoye et al. did for Dec-POMDPs [9]. In Section 5, our Bellman/Shapley operator is proved to induce an optimal game value function that is Lipschitz-continuous in occupancy space, which leads to deriving value function approximators, including upper- and lower-bounding ones, and discussing their initialization. Finally, Section 6 describes a variant of HSVI for zs-POSGs, and demonstrates its finite-time convergence to an ϵ-optimal solution despite the continuous (occupancy) state and action spaces.

2 Related Work

Infinite horizon POSGs are undecidable [22], which justifies searching for near-optimal solutions, *e.g.*, through finite horizon solutions, as we will do. There is little work on solving POSGs, in particular through exploiting Bellman's optimality principle. One exception is Hansen and Zilberstein's work on finite horizon POSGs [13] , where dynamic programming (DP) incrementally constructs non-dominated policy trees for each player, which allows then deriving a solver for common-payoff POSGs, *i.e.*, decentralized partially observable Markov decision processes (Dec-POMDPs). Here, Bellman's OP thus serves as a preprocessing phase, while we aim at employing it in the core of algorithms.

Dec-POMDPs Bellman's OP appears as the core component of a Dec-POMDP solver when Szer et al. [30] adopt a planner-centric viewpoint whereby the planner aims at providing the players with their private policies without knowing which action-observation histories they have experienced. The planner's information state at t thus contains the initial belief and the joint policy up to t. This leads to turning a Dec-POMDP into an information-state MDP, and obtaining a deterministic shortest path problem that can be solved using an A* search called MAA* (multi-agent A*). Then, another important step is when Dibangoye et al. [9] show that (i) the *occupancy state*, a statistic used to compute expected rewards in MAA*, is in fact sufficient for planning, and (ii) the optimal value function is piecewise linear and convex (PWLC) in occupancy space, which allows adapting point-based POMDP solvers using approximators of V^*.

Subclasses of POSGs Recent works addressed particular cases of discounted partially observable stochastic games (POSGs), 2-player and zero-sum if not specified otherwise, exploiting the structure of the problem to turn it into an equivalent problem for which Bellman's principle applies. Ghosh et al. [12] considered POSGs with public actions and shared observations, which can be turned into stochastic games defined over the common belief space, similarly to POMDPs turned into belief MDPs. Chatterjee and Doyen [7], Basu and Stettner [2], and Horák et al. [15] considered One-Sided POSGs, *i.e.*, scenarios where (player) 2 (w.l.o.g.) only partially observes the system state, and 1 has access to the system state, plus the action and observation of 2. Cole and Kocherlakota [8] considered (n-player) POSGs with independent private states, partially shared observability, and i's utility function depending on his private state and on the shared observation. Horák and Bošanský [14] considered zs-POSGs with independent private states and public observations, *i.e.*, scenarios where (i) each player i has a private state s_i he fully observes, and (ii) both players receive the same public observations of each player's private state. Any player's belief over the other player's private state is thus common knowledge.

Focusing on the work of Horák et al. [15, 14], in both cases convexity or concavity properties of the optimal value function are obtained, which allow deriving upper- and lower-bounding approximators. These approximators are then employed in HSVI-based algorithms. Yet, moving from MDPs and POMDPs (as in Smith's work) to these settings induces a tree of possible futures with an infinite branching factor, which requires changes to the algorithm, and thus to the theoretical analysis of the finite-time convergence. As we shall see, the present work adopts similar changes.
Wiggers et al. [32] prove that, using appropriate representations, the value function associated to a zs-POSG is convex for (maximizing) player 1 and concave for (minimizing) player 2. Yet, this did not allow deriving a solver based on approximating the value function. Here, we exploit no convexity or concavity property of the optimal value function, as they may not hold, but its Lipschitz continuity.

Imperfect Information Games Finite horizon (general-sum) POSGs can be written as extensive-form games with imperfect information and perfect recall (EFGs, often referred to as *imperfect information games*) [24], which makes solution techniques for EFGs relevant even for infinite-horizon POSGs. A first approach to solving EFGs is to turn them into a normal-form game before looking for a Nash equilibrium, thus ignoring the temporal aspect of the problem [27] and inducing a combinatorial explosion. For (2-player) zs-EFGs, this leads to solving two linear programs (one for each player). Koller and Megiddo [16] propose a different linear programming approach for zs-EFGs that exploits the temporal aspect through the choice of decision variables, but still does not apply Bellman's OP (see also [31, 17]).
More recently, Counterfactual Regret minimization (CFR) [34] has been introduced, allowing to solve large imperfect-information games with bounded regret such as heads-up no limit hold'em poker, now winning against top human players [5]. While some CFR-based algorithms use heuristic-search techniques, thus somehow exploit the sequentiality of the game, they do not rely on Bellman's OP either.

3 Background

For the sake of clarity, the concepts and results of the EFG literature used in this work will be recast in the POSG setting. We shall employ the terminology of pure/mixed/behavioral strategies and strategy profiles—more convenient in our non-collaborative setting—instead of deterministic or stochastic policies (private or joint ones)—common in the collaborative setting of Dec-POMDPs.
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 in state s;
- $r(s, a^1, a^2)$ is a (scalar) reward function;
- $H \in \mathbb{N} \cup \{\infty\}$ is a temporal horizon;
- $\gamma \in [0, 1)$ is a discount factor; and
- b_0 is the (public/common) initial belief state.

1 would like to maximize the expected return, defined as the discounted sum of future rewards, while 2 would like to minimize it, what we formalize next.
From the Dec-POMDP, POSG and EFG literature, we use the following concepts and definitions, where $i \in \{1, 2\}$:

$-i$ is i's opponent.

$\theta^i_\tau = (a^i_1, z^i_1, \ldots, a^i_\tau, z^i_\tau)$ is a length-τ *action-observation history* for i. The set of histories is $\Theta^i = \Theta^i_0 \cup \Theta^i_1 \cup \Theta^i_2 \cup \ldots$, with one subset per time step.

$\boldsymbol{\theta}_\tau = (\theta^1_\tau, \theta^2_\tau)$ is a *joint history* at time τ. The set of joint histories is $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1 \cup \boldsymbol{\Theta}_2 \cup \ldots$, with one subset per time step.

[o_τ] An *occupancy state* o_τ at time τ is a probability distribution over state–joint-history pairs $(s, \boldsymbol{\theta}_\tau)$. ($o_0$ is completely specified by b_0.) The set of occupancy states is $O = O_0 \cup O_1 \cup O_2 \cup \ldots$, with one subset per time step. Note that this notion applies to POSGs despite the use of stochastic actions.

[$\pi^i_{0:\tau}$] A *pure strategy* for i is a mapping $\pi^i_{0:\tau}$ from private histories in Θ^i_t ($\forall t \in \{0\,..\,\tau\}$) to **single** private actions. By default, $\pi^i \stackrel{\text{def}}{=} \pi^i_{0:H-1}$.

$\boldsymbol{\pi}_{0:\tau} = \langle \pi^1_{0:\tau}, \pi^2_{0:\tau} \rangle$ is a *pure strategy profile*.

[$\mu^i_{0:\tau}$] A *mixed strategy* $\mu^i_{0:\tau}$ for i is a probability distribution over pure strategies. It is used by first sampling one of the pure strategies (at $t = 0$), and then executing that strategy until $t = \tau$.

$\boldsymbol{\mu}_{0:\tau} = \langle \mu^1_{0:\tau}, \mu^2_{0:\tau} \rangle$ is a *mixed strategy profile*.

[β^i_τ] A *(behavioral) decision rule* at time τ for i is a mapping β^i_τ from private histories in Θ^i_τ to **distributions** over private actions. For convenience, we will note $\beta^i_\tau(\theta^i_\tau, a^i)$ the probability to pick action a^i when facing history θ^i_τ.

$\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$ is a *decision rule profile* ($\in \mathcal{B}_\tau$, and noting $\mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1 \cup \ldots$).

$\beta^i_{\tau:\tau'} = (\beta^i_\tau, \ldots, \beta^i_{\tau'})$ is a *behavioral strategy* for i from time step τ to τ' (included). By default, $\beta^i \stackrel{\text{def}}{=} \beta^i_{0:H-1}$.

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

[$V_0(o_0, \boldsymbol{\beta})$] The *value* of a behavioral strategy profile $\boldsymbol{\beta}$ in occupancy state o_0 (from time step 0 on) is:

$$V_0(o_0, \boldsymbol{\beta}) = E[\sum_{t=0}^{\infty} \gamma^t R_t | O_0 = o_0, \boldsymbol{\beta}],$$

where R_t is the random variable associated to the instant reward at time step t. [Note: This definition extends naturally to pure and mixed strategy profiles.]

The primary objective is here to find a Nash equilibrium strategy (NES), *i.e.*, a mixed strategy profile $\boldsymbol{\mu}^* = \langle \mu^{1*}, \mu^{2*} \rangle$ such that no player has an incentive to deviate, which can be written:

$$\forall \mu^1, V_0(o_0, \mu^{1*}, \mu^{2*}) \geq V_0(o_0, \mu^1, \mu^{2*}),$$
$$\forall \mu^2, V_0(o_0, \mu^{1*}, \mu^{2*}) \leq V_0(o_0, \mu^{1*}, \mu^2).$$

In such a 2-player zero-sum game, all NESs have the same Nash-equilibrium value (NEV) $V^*_0(o_0) \stackrel{\text{def}}{=} V_0(o_0, \mu^{1*}, \mu^{2*})$. Finite horizon POSGs being equivalent to EFGs with imperfect information and perfect recall, the following key result for EFGs applies to (finite H) POSGs:

Theorem 1. *[19, 11] In a game of perfect recall, mixed and behavioral strategies are equivalent. (More precisely: Every mixed strategy is equivalent to the unique behavioral strategy it generates, and each behavioral strategy is equivalent to every mixed strategy that generates it.)*

4 Solving POSGs as Occupancy MGs

In this section, unless stated otherwise, we assume finite horizons and exact solutions (no ϵ error).
Here, we show (i) how a zs-POSG can be reformulated as a different zero-sum Markov game, and (ii) that Bellman's optimality principle applies in this game.

4.1 From zs-POSGs to zs-OMGs

To solve a zs-POSG, we take the viewpoint of a central planner that searches *offline* for the best behavioral strategy profile before providing it to the players. This contrasts with Dec-POMDPs where deterministic strategy profiles suffice, and means exploring a (bounded) continuous space rather than a (finite) discrete one as for Dec-POMDPs. Such a planner grows a partial strategy $\boldsymbol{\beta}_{0:\tau-1}$ by appending a decision rule profile $\boldsymbol{\beta}_\tau$.
Note that any partial strategy $\boldsymbol{\beta}_{0:\tau}$ is in one-to-one correspondence with an occupancy state o_τ. So, the controlled process induced in occupancy space, where actions are decision rule profiles, is both deterministic and Markovian (see formal details about the dynamics below): applying $\boldsymbol{\beta}_\tau$ in o_τ (*i.e.*, appending it to $\boldsymbol{\beta}_{0:\tau}$) leads to a unique $o_{\tau+1}$.

Also, the expected reward at time τ is linear in occupancy space (more precisely in the corresponding distribution over states). All this allows reasoning not on partial behavioral strategy profiles, but on occupancy states. The central planner will thus (i) infer occupancy states seen as "beliefs" over the possible situations ("situation" here meaning the current state s and the players' joint action-observation history $\boldsymbol{\theta}_\tau$) which may have been reached, although without knowing what actually happened, and (ii) map each occupancy state to a decision rule profile telling the players how to act depending on their actual action-observation histories.[1] Each zs-POSG is thus turned into an equivalent game, called a *zero-sum occupancy Markov game* (zs-OMG)[2] formally defined by the tuple $\langle O, \mathcal{B}, T, r, H, \gamma, b_0 \rangle$, where:

- O is the set of occupancy states induced by the zs-POSG;
- $\mathcal{B}$ is the set of decision rule profiles of the zs-POSG;
- T is a deterministic transition function that maps each pair $(o_\tau, \boldsymbol{\beta}_\tau)$ to the (only) possible next occupancy state $o_{\tau+1}$; formally (see Lemma 7 in App. A.1), $\forall s', \theta^1_\tau, a^1, z^1, \theta^2_\tau, a^2, z^2$,

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

- r is a reward function naturally induced from the zs-POSG as the expected reward for the current occupancy state and decision rule profile:

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

  we use the same notation r for zs-POSGs as the context shall indicate which one is discussed;

- H, γ, and b_0 are as in the zs-POSG.

Note first that, for convenience, we directly consider behavioral decision rules, which correspond to mixed strategies. Of course, at τ, i's possible actions should be decision rules defined over histories that have non-zero probability in current o_τ. The dynamics being deterministic and the actions public, both players of that new game (also denoted 1 and 2 while these are different players) know the next state after each transition. But this is no standard zs Markov game also since (i) the mixture of two actions is equivalent to another action already in the (continuous) action space at hand, and (ii) at each time step, the state (occupancy) space O_τ is continuous.

We shall study the subgames of a zs-OMG, *i.e.*, situations where some occupancy state o_τ has somehow been reached at time step τ, and the central solver is looking for rational strategies ($\beta^1_{\tau:H-1}$ and $\beta^2_{\tau:H-1}$) to provide to the players. o_τ tells which action-observation histories each player could be facing with non-zero probability, and thus which are relevant for planning. We can then extend the definition of value function from time step 0 only to any time step τ as follows (using behavioral strategies):

$$
\begin{aligned}
&V_\tau(o_\tau, \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}) \\
&= E[\sum_{t=\tau}^{\infty} \gamma^{t-\tau} R_t | O_\tau = o_\tau, \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}].
\end{aligned}
$$

Note that o_τ is in one-to-one relationship with a strategy profile $\boldsymbol{\beta}_{0:\tau-1|o_\tau\rangle}$, so that we can denote $\boldsymbol{\beta}_{0:H-1|o_\tau\rangle}$ its concatenation with a $\boldsymbol{\beta}_{\tau:H-1}$.

4.2 Back to Mixed Strategies

We now re-introduce, and generalize, mixed strategies as a mathematical tool to handle subgames of a zs-OMG as normal-form games, and give some preliminary results.

For a given o_τ, let $\boldsymbol{\mu}_{0:\tau-1|o_\tau\rangle}$ be an arbitrarily chosen mixed strategy profile that leads to (/is compatible with) o_τ, thus defined over time interval $0 : \tau - 1$. To complete this mixed (prefix) strategy, the central planner should provide each player with a different (suffix) strategy to execute for each θ^i_τ it could be facing. We now detail how to build an equivalent set of mixed (full) strategies for i. Each of the pure (prefix) strategies $\pi^i_{0:\tau-1}$ used in $\mu^i_{0:\tau-1|o_\tau\rangle}$ (belonging to a set denoted $\Pi^i_{0:\tau-1|o_\tau\rangle}$) can be extended by appending a different pure (suffix) strategy $\pi^i_{\tau:H-1}$ at each of its leaf nodes, which leads to a large set of pure strategies $\Pi^i_{0:H-1}(\pi^i_{0:\tau-1})$. Then, let $M^i_{0:H-1|o_\tau\rangle}$ be the set of mixed (full) strategies $\mu^i_{0:H-1|o_\tau\rangle}$ obtained by considering the distributions over $\bigcup_{\pi^i_{0:\tau-1} \in \Pi^i_{0:\tau-1|o_\tau\rangle}} \Pi^i_{0:H-1}(\pi^i_{0:\tau-1})$ that verify, $\forall \pi^i_{0:\tau-1}$,

$$
\sum_{\substack{\pi^i_{0:H-1} \in \\ \Pi^i_{0:H-1}(\pi^i_{0:\tau-1})}} \mu^i_{0:H-1|o_\tau\rangle}(\pi^i_{0:H-1}) = \mu^i_{0:\tau-1|o_\tau\rangle}(\pi^i_{0:\tau-1}). \quad (1)
$$

This is the set of mixed strategies compatible with o_τ.

Lemma 1 (Proof in App. A.2)**.** *$M^i_{0:H-1|o_\tau\rangle}$ is convex and equivalent to the set of behavioral strategies $\beta^i_{0:H-1|o_\tau\rangle}$, thus sufficient to search for a Nash equilibrium in o_τ.*

While only future rewards are relevant when making a decision at τ, reasoning with mixed strategies defined from $t = 0$ will be convenient because $V_\tau(o_\tau, \cdot, \cdot)$ is linear in $\mu^i_{0:H-1|o_\tau\rangle}$, which allows coming back to a standard normal-form game and applying known results.

In the remaining, we simply note μ^i (without index) the mixed strategies in $M^i_{0:H-1|o_\tau\rangle}$, set which we now

[1] In contrast, in a POMDP, the belief state depends on the agent's action-observation history, and is mapped to a single action.

[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.

note $M^i_{|o_\tau\rangle}$. Also, since we shall work with local game $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$, let us define: $M^i_{|o_\tau,\beta^j_\tau\rangle}$ the set of i's mixed strategies compatible with occupancy states reachable given o_τ and β^j_τ (with either $j = i$ or $j = -i$). Then, $M^i_{|T(o_\tau,\boldsymbol{\beta}_\tau)\rangle} \subseteq M^i_{|o_\tau,\beta^j_\tau\rangle} \subseteq M^i_{|o_\tau\rangle}$ (inclusion due to the latter sets being less constrained in their definition). As a consequence, if maximizing some function f over i's mixed strategies compatible with a given o_τ:

$$\max_{\mu^i \in M^i_{|o_\tau\rangle}} f(o_\tau, \mu^i, \dots) \geq \max_{\mu^i \in M^i_{|o_\tau,\beta^j_\tau\rangle}} f(o_\tau, \mu^i, \dots) \geq \max_{\mu^i \in M^i_{|o_\tau,\boldsymbol{\beta}_\tau\rangle}} f(o_\tau, \mu^i, \dots).$$

As can be easily demonstrated (cf. Lemma 8 in App. A.3), any Nash equilibrium solution of our original game $V_0(o_0, \cdot, \cdot)$ induces a Nash equilibrium in any of its reachable subgames.[3] But this does not tell whether Bellman's optimality principle applies, what we discuss next.

4.3 Bellman's Optimality Principle

For any τ and o_τ, let us define (i) $\boldsymbol{\beta}^*_{\tau:H-1}(o_\tau)$ a NE profile for the subgame at o_τ, (ii) $V^*_\tau(o_\tau)$ the NE value of the subgame at any o_τ, and (iii) the *local* subgame at o_τ

$$Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau) \stackrel{\text{def}}{=} r(o_\tau, \boldsymbol{\beta}_\tau) + \gamma V^*_{\tau+1}(T(o_\tau, \boldsymbol{\beta}_\tau)).$$

Then, given Nash equilibrium solutions for any $o_{\tau+1}$, the applicability of Bellman's optimality principle shall be proved if a Nash equilibrium of $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ can be found by (i) solving the local subgame $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$ to get a decision rule profile $\boldsymbol{\beta}^*_\tau$ and (ii) appending it to $\boldsymbol{\beta}^*_{\tau+1:H-1}(T(o_\tau, \boldsymbol{\beta}^*_\tau))$.

An Abnormal-Form Game? A first question is whether this game $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$ is in fact a normal-form game, *i.e.*, whether it could be defined by a payoff matrix over pure decision rules, payoffs for behavioral decision rules being obtained through linear mixtures.
$V_\tau(o_\tau, \cdot, \cdot)$ is linear in each player's decision rule space at each time step (*i.e.*, in $\beta^i_{\tau'}$ for any i and $\tau' \in \{\tau \,..\, H-1\}$), but multilinear in each player's behavioral strategy space (see Lemma 2 App. A.4.1), which suggests that $Q^*_\tau(o_\tau, \cdot, \cdot)$ may not be convex-concave (and thus not (bi)linear) in the space of decision rules at τ. As a consequence, we are possibly facing an **ab**normal-form game and cannot use von Neumann's Minimax theorem.

Properties of the Maximin and Minimax Values Rather than digging the convexity-concavity property further, we now show that computing the maximin and minimax values of $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$ induces finding a NE of $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ given NEs for any $o_{\tau+1}$.

Theorem 2 (Proof in App. A.4.2)**.** *In the 2p zs abnormal-form game $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$, the maximin and minimax values are both equal to $V^*_\tau(o_\tau)$*—i.e.*, as previously defined, the NEV for game $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$—and correspond to a NES.*

[3] In contrast, a *subgame perfect equilibrium* requires a Nash equilibrium in any subgame reachable by *some* strategy profile, which is more constraining.

Proof. (sketch) The proof relies on first developing the maximin of $Q^*_\tau(o_\tau, \beta^1_\tau, \beta^2_\tau)$, then using (i) the equivalence of maximin and minimax for mixed strategies (when von Neumann's minimax theorem applies), and (ii) the equivalence of mixed and behavioral strategies. □

Theorem 3 (Proof in App. A.4.2)**.** *As in 2p zs normal-form games, game $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$ has at least one NES; all its NESs are all value-equivalent; and solving for maximin and minimax values allows finding one NES.*

Maximin and Minimax Computation The last results tell us that we can exploit knowledge of the optimal value function at $\tau + 1$ (for all $o_{\tau+1}$) to find optimal decision rules at τ for any given o_τ by computing the maximin and minimax values of the local (abnormal-form) game at hand. Yet, we cannot use an LP as for normal-form games. To find an appropriate solution method, let us now look at properties of this game, noting that we lack any convexity/concavity property, and presenting a preliminary result.

Lemma 2 (Proof in App. A.4.3)**.** *At depth τ, $T(o_\tau, \boldsymbol{\beta}_\tau)$ is linear in β^1_τ, β^2_τ, and o_τ, where $\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$. It is more precisely 1-Lipschitz-continuous in o_τ (in 1-norm),* i.e.*, for any o_τ, o'_τ:*

$$\|T(o'_\tau, \boldsymbol{\beta}_\tau) - T(o_\tau, \boldsymbol{\beta}_\tau)\|_1 \leq 1 \cdot \|o'_\tau - o_\tau\|_1.$$

The Lipschitz continuity (LC) property would also hold in 2-norm or ∞-norm, due to the equivalence between norms, but with different constants.

Lemma 3 (Proof in App. A.4.3)**.** *For any τ and o_τ, $Q^*_\tau(o_\tau, \beta^1_\tau, \beta^2_\tau)$ is Lipschitz continuous in both β^1_τ and β^2_τ.*

The payoff function of our game is thus LC in each private decision-rule space, which suggests using error-bounded global optimization techniques, as Munos's DOO (Deterministic Optimistic Optimization) [23]. Here, searching for a maximin (resp. minimax) value suggests using two nested optimization processes: an "outer" one for the $\max$ (resp. $\min$) operator, and an "inner" one for the $\min$ (resp. $\max$). To ensure being within ϵ of the maximin value, each process could, for example, use an $\frac{\epsilon}{2}$ tolerance threshold. Yet, in such a nested optimization process, the inner process may stop, at each call, before reaching $\frac{\epsilon}{2}$-optimality if it leads the outer process to explore a different point.
Due to the continuous state space of zs-OMGs, V^* cannot be computed exactly. We shall now see how to approximate it, before exploiting the resulting approximators in a specific version of HSVI in Sec. 6.

5 Properties of V^*

In this section, we again assume finite horizon problems (unless stated otherwise). The main objective here is to propose upper- and lower-bounding approximators that exploit V^*'s

Lipschitz continuity (rather than PWLC) property, as Fehr et al. [10] did in the setting of (single agent) information-oriented control, but here with simpler derivations.

5.1 Finite-Horizon Lipschitz Continuity of V^*

The following lemma proves that the expected instant reward at any τ is linear in o_τ, and thus so is the expected value of a finite-horizon strategy profile from τ onwards (trivial proof by induction).

Lemma 4 (Proof in App. A.5.1)**.** *At depth τ, $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ is linear w.r.t. o_τ.*

Corollary 1 (Proof in App. A.5.1)**.** *$V^*_\tau(o_\tau)$ is Lipschitz continuous in o_τ at any depth $\tau \in \{0\,..\,H-1\}$.*

Refining the Lipschitz constant(s) We have just discussed the LC of V^* based on the LC of finite-horizon strategies, reasoning on worst case Lipschitz constants (one per time step) that hold for all strategies. Now, (i) could we refine those constants based on knowledge regarding V^*, in particular upper and lower bounds U and L (see next sections)? And (ii) could we make use of those refined constants in the planning process?
Regarding question (i), U and L tell us that any strategy profile from time τ on (and thus with remaining horizon H) has values within $L^{\min}_\tau \stackrel{\text{def}}{=} \min_{o_\tau} L_\tau(o_\tau)$ and $U^{\max}_\tau \stackrel{\text{def}}{=} \max_{o_\tau} U_\tau(o_\tau)$, hence the refined Lipschitz constant:

$$\lambda^{LU}_\tau = \frac{U^{\max}_\tau - L^{\min}_\tau}{2}.$$

Regarding question (ii), as L and U are refined during the planning process, these refined depth-dependent constants would progressively shrink, thus speeding up planning! This phenomenon could encourage improving the value function bounds where they seem high (for U) or low (for L).

5.2 Approximating V^*

Note: For the sake of readability, the depth index τ may be omitted when it can be inferred from the occupancy state.

Approximators An HSVI-like algorithm requires maintaining both an upper and a lower approximator of V^*. We denote them U and L, and $\hat{V} = (L, U)$.
The LC of V^* suggests employing LC function approximators for U at depth τ (U_τ) in the form of a lower envelope of (i) an initial upper-bound $U^{(0)}(o)$ and (ii) downward-pointing L1-cones, where an upper-bounding cone $c^U_\omega = \langle \omega, u \rangle$—located at ω, with "summit" value u, and slope $\lambda_{(H-\tau)}$—induces a function $U^{(\omega)}(o) \stackrel{\text{def}}{=} u + \lambda_{(H-\tau)} \|\omega - o\|_1$. The upper-bound is thus defined as the lower envelope of $U^{(0)}$ and the set of cones $C^U_\tau = \{c^U_\omega\}_{\omega \in \Omega^U_\tau}$, *i.e.*,

$$U(o) = \min\{U^{(0)}(o), \min_{\omega \in \Omega^U_\tau} U^{(\omega)}(o)\}.$$

Respectively, for the lower-bounding approximator at depth τ: a lower-bounding (upward-pointing) cone $c^L_\omega = \langle \omega, l \rangle$ induces a function $L_\omega(o) = l - \lambda_{(H-\tau)} \|\omega - o\|_1$; and the lower bound is defined as the upper envelope of an initial lower bound $L^{(0)}$ and the set of cones $C^L_\tau = \{c^L_\omega\}_{\omega \in \Omega^L_\tau}$, *i.e.*,

$$L(o) = \max\{L^{(0)}(o), \max_{\omega \in \Omega^L_\tau} L^{(\omega)}(o)\}.$$

(Point-based) Operator and Value Updates One cannot apply an operator (noted $\mathcal{H}$) to update a value function approximator uniformly. Instead, when visiting some occupancy state o (at depth $\tau \in \{0\,..\,H-1\}$), we perform a *point-based* update of the upper-bound U by (i) finding the NEV of the following game (which relies on U at $\tau+1$):

$$\begin{aligned} &U(o, \boldsymbol{\beta}_\tau) \\ &= \sum_{s, a^1, a^2} \left(\sum_{\boldsymbol{\theta}} o(s, \boldsymbol{\theta}) \beta^1(\theta^1, a^1) \beta^2(\theta^2, a^2) \right) r(s, a^1, a^2) \\ &\quad + \gamma U\left(T(o, \boldsymbol{\beta}_\tau)\right) \end{aligned}$$

then (ii) adding a downward-pointing cone to C^U_τ. We note $\mathcal{K}_o U$ the upper bound after this update at point o. The same applies to L with upward-pointing cones instead, and using notation $\mathcal{K}_o L$.

5.3 Initializations

Due to the symmetry between players in a zs-POSG, without loss of generality, let us look for an upper bound of the optimal value function V^*, *i.e.*, an optimistic bound (an admissible heuristic) for (maximizing) player 1. A usual approach to look for optimistic bounds is to relax the problem for the player at hand. To that end, one can here envision manipulating the players' knowledge, their control over the system, the action ordering, or the opponent's objective, *e.g.*:

1. providing more (*e.g.* full) state observability to 1;
2. providing less (*e.g.* no) state observability to 2;
3. letting 1 know what 2 observes;
4. letting 1 control chance (2's choice would then only restrict the set of reachable states), but this would require that 1 has full observability;
5. letting 2 act first, and telling 1 about 2's selected action (exactly or through a partial observation);
6. turning 2 into a collaborator by making him maximize, rather than minimize, the expected return.

Accounting for related Markov models for sequential decision-making, this suggests turning the zs-POSG at hand for example into:

- a Dec-POMDP by turning the opponent into a collaborator (or even into a POMDP or an MDP); or
- a One-Sided POSGs [15] by combining (i) full state observability, (ii) observability of 2's observation, and (iii) observability of 2's action.

Note that making both players' actions or observations public (as in PO-POSGs [14]) would not be a viable solution as this would imply providing more knowledge to both players at the same time, which may prevent the resulting optimal value function from being an upper bound for our problem.

6 HSVI for zs-POSGs

In this section, we assume infinite horizon problems (unless stated otherwise) and ϵ-optimal solutions.

6.1 Algorithm

As we shall see, ϵ-optimally solving an ∞-horizon zs-POSG amounts, as often, to solving a problem with finite horizon $H_{\max}$, which allows exploiting the results derived up to now. For convenience, we assume $H_{\max}$ already known and use horizon-dependent constants (*e.g.*, Lipschitz constants).

HSVI for zs-OMGs is detailed in Algorithm 1. As vanilla HSVI, it relies on (i) generating trajectories while acting optimistically (lines 9–11), *i.e.*, player 1 (resp. 2) acting "greedily" w.r.t. U (resp. L), and (ii) locally updating the upper- and lower-bounding approximators (lines 8 and 13). Here, computations of value updates and strategies rely on solving our local zero-sum abnormal form games (possibly a maximin/minimax optimization exploiting the Lipschitz continuity as discussed in Sec. 4.3). A key difference lies in the criterion for stopping trajectories. In vanilla HSVI (for POMDPs), the finite branching factor allows looking at the convergence of U and L at each point reachable under an optimal strategy. To ensure ϵ-convergence at b_0, trajectories just need to be interrupted when the current width at b_τ ($width(\hat{V}(b_\tau))$, where $width(x, y) \stackrel{\text{def}}{=} y - x$) is smaller than a threshold $\gamma^{-\tau}\epsilon$. Here, dealing with an infinite branching factor, one may converge towards an optimal solution while always visiting new points of the occupancy space. Yet, as the sequence of generated (deterministic) trajectories converges to an optimal trajectory, the density of visited points around it increases, so that the Lipschitz approximation error tends to zero. One can thus bound the width within balls around visited points by exploiting the Lipschitz continuity of the optimal value function. As proposed by Horák et al. [15], this is achieved by adding a term $-\sum_{i=1}^{\tau} 2\rho\lambda_\tau\gamma^{-i}$ to ensure that the width is below $\gamma^\tau\epsilon$ within a ball of radius ρ around the current point (here the occupancy state o_τ). Hence the threshold

$$thr(\tau) \stackrel{\text{def}}{=} \gamma^{-\tau}\epsilon - \sum_{i=1}^{\tau} 2\rho\lambda_\tau\gamma^{-i}. \tag{2}$$

Algorithm 1: zs-OMG-HSVI (in red: differences with HSVI)

```
1  Fct HSVI (ε)
2      Initialize L and U
3      while width(o_0) > ε do
4          RecursivelyTry (o_0, τ = 0)
5      return L, U
6  Fct RecursivelyTry (o, τ)
7      if width(o) > thr(τ) then
8          Update (o)
9          β_τ^U ← NES(Γ^s(U))
10         β_τ^L ← NES(Γ^s(L))
11         o' ← T(o, β_τ^{U,1}, β_τ^{L,2})
12         RecursivelyTry (o', τ + 1)
13         Update (o)
14     return
15 Fct Update (o)
16     L ← Update (L, o) /* uses NEV(Γ^o(L))   */
17     U ← Update (U, o) /* uses NEV(Γ^o(U))   */
```

Setting ρ As can be observed, this threshold function should always return positive values, which requires a small enough ρ. For a given problem, the maximum possible value ρ shall depend on the Lipschitz constants at each time step, which themselves depend on the upper and lower bounds of the optimal value function (and thus may evolve during the planning process). For the sake of simplicity, let us consider a single Lipschitz constant λ common to all time steps, which always exists.

Lemma 5 (Proof in App. A.6)**.** *Assuming a single depth-independent Lipschitz constant λ, and noting that*

$$thr(\tau) = \gamma^{-\tau}\epsilon - 2\rho\lambda\frac{\gamma^{-\tau} - 1}{1 - \gamma}, \tag{3}$$

one can ensure positivity of the threshold at any $\tau \geq 1$ by enforcing $\rho < \frac{1-\gamma}{2\lambda}\epsilon$.

We shall thus pick ρ in $(0, \frac{1-\gamma}{2\lambda}\epsilon)$. But what is the effect of setting ρ to small or large values?

- The smaller ρ, the larger $thr(\tau)$, the shorter the trajectories, but the smaller the balls and the higher the required density of points around the optimal trajectory, thus the more trajectories needed to converge.
- The larger ρ, the smaller $thr(\tau)$, the longer the trajectories, but the larger the balls and the lower the required density of points around the optimal trajectory, thus the less trajectories needed to converge.

So, setting ρ means making a compromise between the number of generated trajectories and their length.

6.2 Finite-Time Convergence

First, the following result bounds the length of HSVI's trajectories using the bounded width of $\hat{V}$ and the exponential growth of $thr(\tau)$.

Lemma 6 (Proof in App. A.6)**.** *Assuming a depth-independent Lipschitz constant λ, and with $W \stackrel{\text{def}}{=} \|U^{(0)} - L^{(0)}\|_\infty$, the length of trajectories is upper-bounded by*

$$T_{\max} \stackrel{\text{def}}{=} \left\lceil \log_\gamma \frac{\epsilon - \frac{2\rho\lambda}{1-\gamma}}{W - \frac{2\rho\lambda}{1-\gamma}} \right\rceil .$$

Note that (i) the classical upper-bound is retrieved when $\rho = 0$ (Eq. (6.7) in [28]), and (ii) this gives us the maximum horizon $H_{\max}$ needed to solve the problem. Now, knowing that any trial terminates in bounded time allows deriving the following results, in order.

Theorem 4 (Proof in App. A.6)**.** *Consider a trial $(o_0, \ldots, o_\tau)$ of length τ and consider that the backward updates of $U_{\tau-1}$ and $L_{\tau-1}$ have* **not yet** *been performed. Then*

1. *$width(\mathcal{K}_{o_{\tau-1}} \hat{V}(o_{\tau-1})) \leq thr(\tau - 1) - 2\rho\lambda_\tau$, and*
2. *for every $o'_{\tau-1}$ satisfying $\|o'_{\tau-1} - o_{\tau-1}\|_1 \leq \rho$, it holds: $width(\mathcal{K}_{o_{\tau-1}} \hat{V}(o'_{\tau-1})) \leq thr(\tau - 1)$.*

Theorem 5. *Algorithm 1 terminates with an ϵ-approximation of $V_0^*(o_0)$.*

Proof. (Adapted from [14]) Assume for the sake of contradiction that the algorithm does not terminate and generates an infinite number of explore trials. Since the length of a trial is bounded by a finite number $T_{\max}$, the number of trials of length T (for some $0 \leq T \leq T_{\max}$) must be infinite. It is impossible to fit an infinite number of occupancy points o_T satisfying $\|o_T - o'_T\|_1 > \rho$ within O_T. Hence there must be two trials of length T, $\{o_{\tau,1}\}_{\tau=0}^T$ and $\{o_{\tau,2}\}_{\tau=0}^T$, such that $\|o_{T-1,1} - o_{T-1,2}\|_1 \leq \rho$. Without loss of generality, assume that $o_{T-1,1}$ was visited the first. According to Lemma 4, the point-based update in $o_{T-1,1}$ resulted in $width(\hat{V}(o_{T-1,2})) \leq thr(T-1)$—which contradicts that the condition on line 11 of Algorithm 1 has not been satisfied for $o_{T-1,2}$ (and hence that$\{o_{t,2}\}_{t=0}^T$ was a trial of length T). □

Note that the number of trials could be (tediously) upper-bounded by determining how many balls of radius ρ are required to cover occupancy simplexes at each depth.

7 Discussion

Inspired by techniques solving POMDPs as belief MDPs or Dec-POMDPs as occupancy MDPs, we have demonstrated that zs-POSGs could be turned into a new type of sequential game, namely zs-OMGs, allowing to apply Bellman's optimality principle. Value function approximators (with heuristic initializations) can be used thanks to the Lipschitz continuity of V^*, and despite V^* possibly **not** being concave or convex in any relevant statistic. A variant of HSVI has been derived which provably converges in finite time to an ϵ-optimal solution.

This approach was motivated by the fact that the corresponding techniques for POMDPs and Dec-POMDPs provide state-of-the-art solvers. The time complexity of the algorithm shall depend, among other things, on that of the maximin/minimax optimization technique in use, and on how many trials are required before convergence. We also currently lack empirical comparisons of the resulting algorithm with existing zs-POSG solution techniques.

Several implementation details could be further discussed as the maximin/minimax error-bounded optimization algorithm, the need to regularly prune dominated cones in U and L, and the possible use of compression techniques to reduce the dimensionality of the occupancy subspaces, as in FB-HSVI [9].

Regarding execution, as in single-agent or collaborative multi-agent settings, while exploration is guided by optimistic decisions (greediness w.r.t. U for 1 and L for 2), actual decisions should be pessimistic, *i.e.*, 1 should act "greedily" w.r.t. L, and 2 w.r.t. U.

Handling finite-horizon settings requires little changes. The maximum length of trials shall be the minimum between this horizon and the bound that depends on ϵ and ρ. Additionally considering $\gamma = 1$ shall require revising the Lipschitz constants and some other formulas.

As often with Dec-POMDPs [30, 9], each player's strategy is here history-dependent, because one could not come up with private belief states, which is feasible under certain assumptions [15, 14]. One could possibly address this issue as MacDermed and Isbell [20] did by assuming that a bounded number of beliefs is sufficient to solve the problem.

Public actions and observations, as in Poker, could be exploited by turning the non-observable sequential decision problem faced by the central planner into a partially observable one, and thus the deterministic OMG into a probabilistic one.

References

[1] N. Basilico, G. De Nittis, and N. 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.

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

[3] R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Science*, 38:716–719, 1952.

[4] D. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

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

[6] O. Buffet, J. Dibangoye, A. Saffidine, and V. Thomas. Heuristic search value iteration for zero-sum stochastic games. *IEEE Transactions on Games*, 2020.

[7] K. Chatterjee and L. Doyen. Partial-observation stochastic games: How to win when belief fails. volume 15, page 16, 2014.

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

[9] J. Dibangoye, C. Amato, O. Buffet, and F. Charpillet. Optimally solving Dec-POMDPs as continuous-state MDPs. *Journal of Artificial Intelligence Research*, 55:443–497, 2016.

[10] M. Fehr, O. Buffet, V. Thomas, and J. Dibangoye. ρ-POMDPs have Lipschitz-continuous ϵ-optimal value functions. In *Advances in Neural Information Processing Systems 31*, pages 6933–6943, 2018.

[11] D. Fudenberg and J. Tirole. *Game Theory*. The MIT Press, 1991.

[12] M. K. Ghosh, D. McDonald, and S. Sinha. Zero-sum stochastic games with partial information. *Journal of Optimization Theory and Applications*, 121(1):99–118, Apr. 2004.

[13] E. A. Hansen, D. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 2004.

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

[15] K. Horák, B. Bošanský, and M. 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.

[16] D. Koller and N. Megiddo. The complexity of two-person zero-sum games in extensive form. *Games and Economic Behavior*, 4(4):528–552, 1992.

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

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

[19] H. W. Kuhn. *Extensive Games and the Problem of Information*, volume Contributions to the Theory of Games II of *Annals of Mathematics (AM-28)*, pages 193–216. Princeton University Press, 1953.

[20] L. C. MacDermed and C. Isbell. Point based value iteration with optimal belief compression for Dec-POMDPs. In *Advances in Neural Information Processing Systems 26*, 2013.

[21] E. Machuca. An analysis of multiobjective search algorithms and heuristics. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[22] O. Madani, S. Hanks, and A. Condon. On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 1999.

[23] R. Munos. From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–130, 2014.

[24] F. Oliehoek and N. 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] K. Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174 – 205, 1965.

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

[27] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

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

[29] B. S. Stewart and C. C. White, III. Multiobjective A*. *Journal of the ACM*, 38(4):775–814, Oct. 1991.

[30] D. Szer, F. Charpillet, and S. 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.

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

[32] A. Wiggers, F. Oliehoek, and D. 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.

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

[34] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems 20*, 2007.

A Appendix

This appendix mainly provides proofs of several theoretical claims of the paper.

A.1 From zs-POSGs to zs-OMGs

The following result shows that the occupancy state is Markovian, *i.e.*, its value at τ only depends on its previous value ($o_{\tau-1}$), the system dynamics ($P^{z^1,z^2}_{a^1,a^2}$), and the last behavioral decision rules ($\beta^1_{\tau-1}$ and $\beta^2_{\tau-1}$).

Lemma 7. *Given an occupancy state $o_{\tau-1}$ and a behavioral decision rule profile $\beta_{\tau-1}$, next occupancy state o_τ is given by the following formula (for any s', $\theta^1_{\tau-1}$, a^1, z^1, $\theta^2_{\tau-1}$, a^2, z^2):*

$$o_\tau(s', (\theta^1_{\tau-1}, a^1, z^1), (\theta^2_{\tau-1}, a^2, z^2)) = \beta^1_{\tau-1}(\theta^1_{\tau-1}, a^1) \cdot \beta^2_{\tau-1}(\theta^2_{\tau-1}, a^2) \sum_s P^{z^1,z^2}_{a^1,a^2}(s'|s) \cdot o_{\tau-1}(s, \theta^1_{\tau-1}, \theta^2_{\tau-1}).$$

Proof. The proof goes by simply developing the definition:

$$\begin{aligned}
o_\tau(s', (\theta^1_{\tau-1}, a^1, z^1), (\theta^2_{\tau-1}, a^2, z^2)) &\overset{\text{def}}{=} Pr(s', (\theta^1_{\tau-1}, a^1, z^1), (\theta^2_{\tau-1}, a^2, z^2)) \\
&= \sum_s Pr(s, s', (\theta^1_{\tau-1}, a^1, z^1), (\theta^2_{\tau-1}, a^2, z^2)) \\
&= \sum_s Pr(s', z^1, z^2 | s, \theta^1_{\tau-1}, a^1, \theta^2_{\tau-1}, a^2) \cdot Pr(s, \theta^1_{\tau-1}, a^1, \theta^2_{\tau-1}, a^2) \\
&= \sum_s P^{z^1,z^2}_{a^1,a^2}(s'|s) \cdot Pr(a^1, a^2 | s, \theta^1_{\tau-1}, \theta^2_{\tau-1}) \cdot Pr(s, \theta^1_{\tau-1}, \theta^2_{\tau-1}) \\
&= \sum_s P^{z^1,z^2}_{a^1,a^2}(s'|s) \cdot Pr(a^1 | s, \theta^1_{\tau-1}, \theta^2_{\tau-1}) \cdot Pr(a^2 | s, \theta^1_{\tau-1}, \theta^2_{\tau-1}) \cdot Pr(s, \theta^1_{\tau-1}, \theta^2_{\tau-1}) \\
&= \sum_s P^{z^1,z^2}_{a^1,a^2}(s'|s) \cdot \beta^1_{\tau-1}(\theta^1_{\tau-1}, a^1) \cdot \beta^2_{\tau-1}(\theta^2_{\tau-1}, a^2) \cdot o_{\tau-1}(s, \theta^1_{\tau-1}, \theta^2_{\tau-1}) \\
&= \beta^1_{\tau-1}(\theta^1_{\tau-1}, a^1) \cdot \beta^2_{\tau-1}(\theta^2_{\tau-1}, a^2) \sum_s P^{z^1,z^2}_{a^1,a^2}(s'|s) \cdot o_{\tau-1}(s, \theta^1_{\tau-1}, \theta^2_{\tau-1}).
\end{aligned}$$

□

A.2 Back to Mixed Strategies

The following result demonstrate that, instead of reasoning on mixed strategies constrained to be "compatible" with some occupancy state o_τ, one can reason equivalently with behavioral strategies.

Lemma 1. (originally stated on page 4) *$M^i_{0:H-1|o_\tau\rangle}$ is convex and equivalent to the set of behavioral strategies $\beta^i_{0:H-1|o_\tau\rangle}$, thus sufficient to search for a Nash equilibrium in o_τ.*

Proof. Let $\mu^i_{0:H-1|o_\tau\rangle}$ and $\nu^i_{0:H-1|o_\tau\rangle}$ be two mixed strategies in $M^i_{0:H-1|o_\tau\rangle}$, *i.e.*, which are both compatible with occupancy state o_τ at time step τ, and $\alpha \in [0, 1]$. Then, for any $\pi^i_{0:\tau-1}$,

$$\begin{aligned}
&\sum_{\pi^i_{0:H-1} \in \Pi^i_{0:H-1}(\pi^i_{0:\tau-1})} \left[\alpha \cdot \mu^i_{0:H-1|o_\tau\rangle}(\pi^i_{0:H-1}) + (1-\alpha) \cdot \nu^i_{0:H-1|o_\tau\rangle}(\pi^i_{0:H-1})\right] \\
&= \alpha \left[\sum_{\pi^i_{0:H-1} \in \Pi^i_{0:H-1}(\pi^i_{0:\tau-1})} \mu^i_{0:H-1|o_\tau\rangle}(\pi^i_{0:H-1})\right] + (1-\alpha) \left[\sum_{\pi^i_{0:H-1} \in \Pi^i_{0:H-1}(\pi^i_{0:\tau-1})} \nu^i_{0:H-1|o_\tau\rangle}(\pi^i_{0:H-1})\right]
\end{aligned}$$

(because both mixed strategies are compatible with o_τ (eq. 1, p. 4:)

$$\begin{aligned}
&= \alpha \cdot \mu^i_{0:\tau-1|o_\tau\rangle}(\pi^i_{0:\tau-1}) + (1-\alpha) \cdot \mu^i_{0:\tau-1|o_\tau\rangle}(\pi^i_{0:\tau-1}) \\
&= \mu^i_{0:\tau-1|o_\tau\rangle}(\pi^i_{0:\tau-1}).
\end{aligned}$$

Eq. 1 thus also applies to $\alpha \cdot \mu^i_{0:H-1|o_\tau\rangle} + (1-\alpha) \cdot \nu^i_{0:H-1|o_\tau\rangle}$, proving that it belongs to $M^i_{0:H-1|o_\tau\rangle}$ and, as a consequence, that this set is convex.
The equivalence with the set of behavioral strategies simply relies on the fact that all mixed strategies over $\tau : H-1$ can be independently generated at each action-observation history $\theta^i_{0:\tau-1}$. □

A.3 Nash Equilibria in Subgames?

The definition of Nash equilibrium still applies in each of the infinitely many (induced) subgames and, as explained by the following lemma, NEs are "nested".

Lemma 8. *A Nash equilibrium of $V_\tau(o_\tau, \cdot, \cdot)$ induces a Nash equilibrium in any (induced) subgame $V_{\tau'}(o_{\tau'}, \cdot, \cdot)$ for $\tau' \geq \tau$.*

Proof. For any $\tau \in \{0 \,..\, H-1\}$ and any o_τ, (i) $V_\tau(o_\tau, \mu^1, \mu^2)$ is linear in both μ^1 and μ^2, and (ii) the space of mixed strategy profiles constrained by o_τ is convex (Lemma 1). This allows applying von Neumann's minimax theorem, so that this subgame at o_τ induces a 2-player zero-sum normal-form game for which at least one NE exists (and all NEs are equivalent).
Lemma 1 (p. 4) allows reasoning with behavorial strategy profiles instead of mixed ones, and we can focus on what happens from τ on. So, let $\boldsymbol{\beta}^*_{\tau:H-1}$ be a NES of $V_\tau(o_\tau, \cdot, \cdot)$. Then the definition of Nash equilibrium allows writing in particular, $\forall \beta^1_{\tau+1:H-1}$:

$$V_\tau(o_\tau, \langle \beta^{1,*}_\tau \oplus \beta^{1,*}_{\tau+1:H-1} \rangle, \beta^{2,*}_{\tau:H-1}) \geq V_\tau(o_\tau, \langle \beta^{1,*}_\tau \oplus \beta^{1}_{\tau+1:H-1} \rangle, \beta^{2,*}_{\tau:H-1}),$$

thus,

$$\begin{aligned} r(o_\tau, \beta^{1,*}_\tau, \beta^{2,*}_\tau) + \gamma V_{\tau+1}(T(o_\tau, \beta^{1,*}_\tau, \beta^{2,*}_\tau), \beta^{1,*}_{\tau+1:H-1}, \beta^{2,*}_{\tau+1:H-1}) \\ \geq r(o_\tau, \beta^{1,*}_\tau, \beta^{2,*}_\tau) + \gamma V_{\tau+1}(T(o_\tau, \beta^{1,*}_\tau, \beta^{2,*}_\tau), \beta^{1}_{\tau+1:H-1}, \beta^{2,*}_{\tau+1:H-1}) \end{aligned}$$

or, equivalently,

$$V_{\tau+1}(T(o_\tau, \beta^{1,*}_\tau, \beta^{2,*}_\tau), \beta^{1,*}_{\tau+1:H-1}, \beta^{2,*}_{\tau+1:H-1}) \geq V_{\tau+1}(T(o_\tau, \beta^{1,*}_\tau, \beta^{2,*}_\tau), \beta^{1}_{\tau+1:H-1}, \beta^{2,*}_{\tau+1:H-1}).$$

With the symmetric property holding for player 2, this implies that $\boldsymbol{\beta}^*_{\tau+1:H-1} \overset{\text{def}}{=} (\beta^{1,*}_{\tau+1:H-1}, \beta^{2,*}_{\tau+1:H-1})$ is a NES of the *constrained* 2-player zero-sum (normal-form) game $V_{\tau+1}(T(o_\tau, \boldsymbol{\beta}^*_\tau), \cdot, \cdot)$. By induction, a NES is obtained for any subgame $V_{\tau'}(o_{\tau'}, \cdot, \cdot)$ ($\tau' \geq \tau$). □

In particular, as expected, any Nash equilibrium solution of our original game $V_0(o_0, \cdot, \cdot)$ induces a Nash equilibrium in any of its reachable subgames, ensuring a rational behavior at each time step.

A.4 Solving POSGs as Occupancy Markov Games

A.4.1 An Abnormal-Form Game?

The next two lemmas lead to demonstrating that $V_\tau(o_\tau, \langle \cdot, \cdot \rangle)$ is linear in $\beta^i_{\tau'}$ for $i \in \{1, 2\}$ and $\tau' \geq \tau$.

Lemma 9. *At depth τ, $T(o_\tau, \boldsymbol{\beta}_\tau)$ is linear in β^1_τ and β^2_τ, where $\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$.*

Proof. Let o_τ be an occupancy state at depth τ and $\boldsymbol{\beta}_\tau$ be a decision rule. Then the next occupancy state $\tilde{o} = T(o_\tau, \boldsymbol{\beta}_\tau)$ satisfies, for any $\tilde{s}$ and $(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z})$:

$$\begin{aligned} \tilde{o}(\tilde{s}, (\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z})) &= \sum_{s \in \mathcal{S}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} o_\tau(s, \boldsymbol{\theta}) \boldsymbol{\beta}_\tau(\boldsymbol{\theta}, \boldsymbol{a}) P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) \\ &= \sum_{s \in \mathcal{S}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} o_\tau(s, \boldsymbol{\theta}) \beta^1_\tau(\theta^1, a^1) \beta^2_\tau(\theta^2, a^2) P^{z^1, z^2}_{a^1, a^2}(s'|s) \\ &= \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \beta^1_\tau(\theta^1, a^1) \beta^2_\tau(\theta^2, a^2) \left(\sum_{s \in \mathcal{S}} o_\tau(s, \boldsymbol{\theta}) P^{z^1, z^2}_{a^1, a^2}(s'|s) \right). \end{aligned}$$

The next occupancy state thus also evolves linearly w.r.t. *private* decision rules at τ. □

Lemma 10. *At depth τ, $r_\tau(o_\tau, \boldsymbol{\beta}_\tau)$ is linear in β^1_τ and β^2_τ, where $\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$.*

Proof. When visiting some occupancy state o_τ (at depth $\tau \in \{0, \ldots, \tau_{\max} - 1\}$), for joint behavioral decision rule $\boldsymbol{\beta}_\tau = \langle \beta^1, \beta^2 \rangle$, the expected immediate reward at τ is:

$$r_\tau(o_\tau, \boldsymbol{\beta}_\tau) = \sum_{s, a^1, a^2} \left(\sum_{\boldsymbol{\theta}} o_\tau(s, \boldsymbol{\theta}) \beta^1(\theta^1, a^1) \beta^2(\theta^2, a^2) \right) r(s, a^1, a^2)$$

$$= \sum_{\theta^1,\theta^2} \sum_{a^1,a^2} \beta^1(\theta^1,a^1)\beta^2(\theta^2,a^2) \left(\sum_s o_\tau(s,\boldsymbol{\theta}) r(s,a^1,a^2) \right),$$

i.e., it is linear in β^1 as well as β^2 (hence bilinear). □

Corollary 2. *At depth τ, $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ is linear in $\beta^1_{\tau'}$ as well as $\beta^2_{\tau'}$ for any $\tau' \in \{\tau \,..\, H-1\}$.*

Proof. First, the property trivially holds for $\tau = H-1$.
Let us now assume that it holds for some $\tau+1 \in \{1 \,..\, H-1\}$. Then, we have at τ:

$$\begin{aligned} V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1}) &= r_\tau(o_\tau, \boldsymbol{\beta}_\tau) + \gamma V_{\tau+1}\left(T(o,\boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:H-1}\right) && \{\text{using Lemma 10:}\} \\ &= \left[\sum_{\boldsymbol{\theta},a^1,a^2} \beta^1_\tau(\theta^1,a^1)\beta^2_\tau(\theta^2,a^2) \left(\sum_s o(s,\boldsymbol{\theta}) r(s,a^1,a^2) \right) \right] + \gamma V_{\tau+1}\left(T(o,\boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:H-1}\right). \end{aligned}$$

As

- $T(o_\tau, \boldsymbol{\beta}_\tau)$ is linear in decision rules β^1_τ and β^2_τ (Lemma 9) and
- $V_{\tau+1}(o_{\tau+1}, \boldsymbol{\beta}_{\tau+1:H-1})$ is linear in $o_{\tau+1}$,

then, by composition (and other basic combinations), $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ is linear in decision rules β^1_τ and β^2_τ.
Also, for any $\tau' > \tau$, the first term (expected instant reward) is independent of $\beta_{\tau'}$, and the second term is linear in $\beta^1_{\tau'}$ and $\beta^2_{\tau'}$ (by induction hypothesis), so that $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ is linear in $\beta^1_{\tau'}$ and $\beta^2_{\tau'}$.
Repeating this process, by induction the property holds for all $\tau \in \{0 \,..\, H-1\}$.

□

One issue is that $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ is not linear in $\beta^1_{\tau:H-1}$ but multi-linear in $\beta^1_{\tau'}$ for all $\tau' \geq \tau$ (idem for player 2). As a consequence, this function may not be convex in $\beta^1_{\tau:H-1}$ (or concave in $\beta^2_{\tau:H-1}$).

A.4.2 Properties of the Maximin and Minimax Values

The next two results demonstrate that solving $Q^*(o_\tau, \boldsymbol{\beta}_\tau)$ for maximin and minimax values allows finding one Nash equilibrium strategy profile (NES), so that Bellman's optimality principle can be applied.

Theorem 2. (originally stated on page 5) *In the 2p zs abnormal-form game $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$, the maximin and minimax values are both equal to $V^*_\tau(o_\tau)$*—i.e.*, as previously defined, the NEV for game $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$—and correspond to a NES.*

Proof. Focusing, without loss of generality, on player 1, we have (complementary explanations follow for numbered lines in particular):

$$\begin{aligned} maximin(o_\tau) &\stackrel{\text{def}}{=} \max_{\beta^1_\tau} \min_{\beta^2_\tau} Q^*_\tau(o_\tau, \beta^1_\tau, \beta^2_\tau) \\ &= \max_{\beta^1_\tau} \min_{\beta^2_\tau} \left[r(o_\tau, \beta^1_\tau, \beta^2_\tau) + \gamma V^*_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau)) \right] \end{aligned}$$

($V^*_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau))$ being the Nash equilibrium value of normal-form game $V_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau), \mu^1, \mu^2)$:)

$$\begin{aligned} &= \max_{\beta^1_\tau} \min_{\beta^2_\tau} \left[r(o_\tau, \beta^1_\tau, \beta^2_\tau) + \gamma \max_{\mu^1 \in M^1_{|o_\tau, \boldsymbol{\beta}_\tau\rangle}} \min_{\mu^2 \in M^2_{|o_\tau, \boldsymbol{\beta}_\tau\rangle}} V_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau), \mu^1, \mu^2) \right] \\ &= \max_{\beta^1_\tau} \min_{\beta^2_\tau} \max_{\mu^1 \in M^1_{|o_\tau, \boldsymbol{\beta}_\tau\rangle}} \min_{\mu^2 \in M^2_{|o_\tau, \boldsymbol{\beta}_\tau\rangle}} \left[r(o_\tau, \beta^1_\tau, \beta^2_\tau) + \gamma V_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau), \mu^1, \mu^2) \right] \end{aligned}$$

(using the equivalence between maximin and minimax values for the (constrained normal-form) game at $\tau+1$, the last two max and min operators can be swapped:)

$$= \max_{\beta^1_\tau} \min_{\beta^2_\tau} \min_{\mu^2 \in M^2_{|o_\tau, \boldsymbol{\beta}_\tau\rangle}} \max_{\mu^1 \in M^1_{|o_\tau, \boldsymbol{\beta}_\tau\rangle}} \left[r(o_\tau, \beta^1_\tau, \beta^2_\tau) + \gamma V_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau), \mu^1, \mu^2) \right]$$

(merging both mins (and with explanations thereafter):)

$$= \max_{\beta^1_\tau} \min_{\mu^2 \in M^2_{|o_\tau, \beta^1_\tau\rangle}} \max_{\mu^1 \in M^1_{|o_\tau, \beta^1_\tau, \beta^2_\tau(\mu^2)\rangle}} \left[r(o_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)) + \gamma V_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)), \mu^1, \mu^2) \right] \quad (4)$$

(since ignoring the opponent's decision rule does not influence the expected return:)

$$= \max_{\beta^1_\tau} \min_{\mu^2 \in M^2_{|o_\tau\rangle}} \max_{\mu^1 \in M^1_{|o_\tau,\beta^1_\tau\rangle}} \left[r(o_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)) + \gamma V_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)), \mu^1, \mu^2) \right]$$

(using again the minimax theorem's equivalence between maximin and minimax on an appropriate game:)

$$= \max_{\beta^1_\tau} \max_{\mu^1 \in M^1_{|o_\tau,\beta^1_\tau\rangle}} \min_{\mu^2 \in M^2_{|o_\tau\rangle}} \left[r(o_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)) + \gamma V_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)), \mu^1, \mu^2) \right] \tag{5}$$

(merging both maxs (and with explanations thereafter):)

$$= \max_{\mu^1 \in M^1_{|o_\tau\rangle}} \min_{\mu^2_\tau | o_\tau)} \left[r(o_\tau, \beta^1_\tau(\mu_1), \beta^2_\tau(\mu_2)) + \gamma V_{\tau+1}(T(o_\tau, \beta^1_\tau(\mu^1), \beta^2_\tau(\mu_2)), \mu^1, \mu^2) \right] \tag{6}$$

(again with the equivalence property discussed before the lemma:)

$$\begin{aligned}
&= \max_{\mu^1 \in M^1_{|o_\tau\rangle}} \min_{\mu^2 \in M^2_{|o_\tau\rangle}} V_\tau(o_\tau, \mu^1, \mu^2) \\
&= \max_{\beta^1_{\tau:H-1|o_\tau\rangle}} \min_{\beta^2_{\tau:H-1|o_\tau\rangle}} V_\tau(o_\tau, \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}) \\
&\stackrel{\text{def}}{=} V^*_\tau(o_\tau).
\end{aligned}$$

More precisely, line 4 (and, similarly, line 6) is obtained by observing that

- minimizing over both (i) β^2_τ and (ii) μ^2 constrained by o_τ and $\boldsymbol{\beta}_\tau$ is equivalent to minimizing over μ^2 constrained by o_τ and β^1_τ; and
- in the reminder of the formula, decision rule β^2_τ at time $depth$ can be retrieved as a function of μ^2 (noted $\beta^2_\tau(\mu^2)$).

Also, line 5 results from the observation that, while $M^1_{|o_\tau,\beta^1_\tau\rangle}$ and $M^2_{|o_\tau\rangle}$ allow to actually make decision over different time intervals, we are here minimizing over μ^2 while maximizing over μ^1 over a function that is linear in both input spaces. This amounts to solving some 2-player zero-sum normal-form game, hence the applicability of von Neumann's minimax theorem. The above derivation tells us that the maximin value (the best outcome player 1 can guarantee whatever player 2's strategy) in the one-time-step game is thus the Nash equilibrium value (NEV) for the complete subgame from τ onwards. □

Theorem 3. (originally stated on page 5) *As in 2p zs normal-form games, game $Q^*_\tau(o_\tau, \boldsymbol{\beta}_\tau)$ has at least one NES; all its NESs are all value-equivalent; and solving for maximin and minimax values allows finding one NES.*

Proof. When player 1 (resp. 2) selects a strategy guaranteeing the maximin (resp. minimax) value, the same value is guaranteed for both players, so that none of them can do better by opting for a different strategy. This situation is thus, by definition, a Nash equilibrium. □

A.4.3 Maximin and Minimax Computation

The next two results demonstrate the Lipschitz-continuity of $Q^*_\tau(o_\tau, \beta^1_\tau, \beta^2_\tau)$ in both β^1_τ and β^2_τ, which will allow finding ϵ-optimal solutions of the maximin and minimax problems.

Lemma 2. (originally stated on page 5) *At depth τ, $T(o_\tau, \boldsymbol{\beta}_\tau)$ is linear in β^1_τ, β^2_τ, and o_τ, where $\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$. It is more precisely* 1*-Lipschitz-continuous in o_τ (in* 1*-norm),* i.e.*, for any o_τ, o'_τ:*

$$\|T(o'_\tau, \boldsymbol{\beta}_\tau) - T(o_\tau, \boldsymbol{\beta}_\tau)\|_1 \leq 1 \cdot \|o'_\tau - o_\tau\|_1.$$

Proof. Let o_τ be an occupancy state at time τ and $\boldsymbol{\beta}_\tau$ be a decision rule. Then the next occupancy state $o' = T(o_\tau, \boldsymbol{\beta}_\tau)$ satisfies, for any s' and $(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z})$:

$$\begin{aligned}
o'(s', (\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z})) &\stackrel{\text{def}}{=} Pr(s', \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z} | o, \beta^1_\tau, \beta^2_\tau) \\
&= \sum_{s \in \mathcal{S}} Pr(s', \boldsymbol{z} | s, \boldsymbol{a}) Pr(\boldsymbol{a} | \boldsymbol{\theta}, \beta^1_\tau, \beta^2_\tau) Pr(s, \boldsymbol{\theta} | o) \\
&= \sum_{s \in \mathcal{S}} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s' | s) \boldsymbol{\beta}_\tau(\boldsymbol{\theta}, \boldsymbol{a}) o_\tau(s, \boldsymbol{\theta})
\end{aligned}$$

$$= \beta^1_\tau(\theta^1, a^1)\beta^2_\tau(\theta^2, a^2) \sum_{s \in \mathcal{S}} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) o_\tau(s, \boldsymbol{\theta}).$$

The next occupancy state thus evolves linearly w.r.t. (i) *private* decision rules for a given private history, and (ii) the occupancy state.

The 1-Lipschitz-continuity holds because each component of o_τ is distributed over multiple components of o'. Indeed, let us view two occupancy states as vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and their corresponding next states under $\boldsymbol{\beta}_\tau$ as $M\boldsymbol{x}$ and $M\boldsymbol{y}$, where $M \in \mathbb{R}^{m \times n}$ is the corresponding transition matrix (*i.e.*, which turns o into $o' \overset{\text{def}}{=} T(o_\tau, \boldsymbol{\beta}_\tau)$. Then,

$$\begin{aligned}
\|M\boldsymbol{x} - M\boldsymbol{y}\|_1 &\overset{\text{def}}{=} \sum_{j=1}^{m} |\sum_{i=1}^{n} M_{i,j}(x_i - y_i)| \\
&\leq \sum_{j=1}^{m} \sum_{i=1}^{n} |M_{i,j}(x_i - y_i)| && \text{(convexity of } |\cdot|) \\
&= \sum_{j=1}^{m} \sum_{i=1}^{n} M_{i,j}|x_i - y_i| && (\forall i, j,\ M_{i,j} \geq 0) \\
&= \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{m} M_{i,j}}_{=1} |x_i - y_i| && (M \text{ is a transition matrix}) \\
&\overset{\text{def}}{=} \|\boldsymbol{x} - \boldsymbol{y}\|_1. && \square
\end{aligned}$$

Lemma 3. (originally stated on page 5) *For any τ and o_τ, $Q^*_\tau(o_\tau, \beta^1_\tau, \beta^2_\tau)$ is Lipschitz continuous in both β^1_τ and β^2_τ.*

Proof. As demonstrated in Sec. 5.1, Corollary 1, in finite horizon problems, the optimal value function is LC in occupancy space.

Then, by definition:

$$Q^*_\tau(o_\tau, \beta^1_\tau, \beta^2_\tau) = r(o_\tau, \beta^1_\tau, \beta^2_\tau) + \gamma V^*_{\tau+1}(T(o_\tau, \beta^1_\tau, \beta^2_\tau)),$$

where the first term (reward-based) is λ_r-LC (in each β^i_τ), with $\lambda_r = \frac{\max_{s,a^1,a^2} r(s,a^1,a^2) - \min_{s,a^1,a^2} r(s,a^1,a^2)}{2}$, and the second term is $(\gamma \cdot \lambda_{H-\tau} \cdot 1)$-LC, with $\lambda_{H-\tau} = \frac{V^{\max}_{H-\tau} - V^{\min}_{H-\tau}}{2}$. Q^*_τ is thus $\lambda^{Q^*}_{H-\tau}$-LC with $\lambda^{Q^*}_{H-\tau} = \lambda_r + \gamma \cdot \lambda^V_{H-\tau}$. $\square$

A.5 Properties of V^*

A.5.1 Finite-Horizon Lipschitz-Continuity of V^*

The next two results demonstrate that, in the finite horizon setting, V^* is Lipschitz-continuous (LC) in occupancy space, which allows defining LC upper- and lower-bounding approximators.

Lemma 4. (originally stated on page 6) *At depth τ, $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$ is linear w.r.t. o_τ.*

Proof. This property trivially holds for $\tau = H - 1$ because

$$\begin{aligned}
V_{H-1}(o_{H-1}, \boldsymbol{\beta}_{H-1:H-1}) &= r(o_{H-1}, \boldsymbol{\beta}_{H-1}) \\
&= \sum_{s,a^1,a^2} \left(\sum_{\boldsymbol{\theta}} o_{H-1}(s, \boldsymbol{\theta}) \beta^1_{H-1}(\theta^1, a^1) \beta^2_{H-1}(\theta^2, a^2) \right) r(s, a^1, a^2) \\
&= \sum_{s,\boldsymbol{\theta}} o_{H-1}(s, \boldsymbol{\theta}) \left(\sum_{a^1,a^2} \beta^1_{H-1}(\theta^1, a^1) \beta^2_{H-1}(\theta^2, a^2) r(s, a^1, a^2) \right).
\end{aligned}$$

Now, let us assume that the property holds for $\tau + 1 \in \{1 \,..\, H-1\}$. Then,

$$V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1}) = \left[\sum_{s,a^1,a^2} \left(\sum_{\boldsymbol{\theta}} o(s, \boldsymbol{\theta}) \beta^1_\tau(\theta^1, a^1) \beta^2_\tau(\theta^2, a^2) \right) r(s, a^1, a^2) \right] + \gamma V_{\tau+1}\left(T(o, \boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:H-1}\right)$$

$$= \left[\sum_{s,\boldsymbol{\theta}} o(s,\boldsymbol{\theta}) \left(\sum_{a^1,a^2} \beta^1_\tau(\theta^1,a^1)\beta^2_\tau(\theta^2,a^2) r(s,a^1,a^2) \right) \right] + \gamma V_{\tau+1}\left(T(o,\boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:H-1}\right).$$

As

- $T(o_\tau, \boldsymbol{\beta}_\tau)$ is linear in o_τ (Lemma 2) and
- $V_{\tau+1}(o_{\tau+1}, \boldsymbol{\beta}_{\tau+1:H-1})$ is linear in $o_{\tau+1}$ (induction hypothesis),

their composition, $V_{\tau+1}(T(o_\tau, \boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:H-1})$, is also linear in o_τ, and so is $V_\tau(o_\tau, \boldsymbol{\beta}_{\tau:H-1})$. □

Corollary 1. (originally stated on page 6) *$V^*_\tau(o_\tau)$ is Lipschitz continuous in o_τ at any depth $\tau \in \{0\,..\,H-1\}$.*

Proof. At depth τ, the value of any behavioral strategy $\boldsymbol{\beta}_{\tau:H-1}$ is bounded, independently of o_τ, by

$$V^{\max}_\tau \overset{\text{def}}{=} \frac{1-\gamma^{H-\tau}}{1-\gamma} \max_{s,\boldsymbol{a}} r(s,\boldsymbol{a}) \text{ and}$$
$$V^{\min}_\tau \overset{\text{def}}{=} \frac{1-\gamma^{H-\tau}}{1-\gamma} \min_{s,\boldsymbol{a}} r(s,\boldsymbol{a}).$$

Because the occupancy space at τ is a probability simplex, for any o and o' in this space, $\|o - o'\|_1 \leq 2$. As a consequence, such a value function being linear in o_τ (*cf.* Lemma 4), it is also $\lambda_{H-\tau}$-LC, *i.e.*,

$$|V_{\boldsymbol{\beta}_{\tau:H-1}}(o) - V_{\boldsymbol{\beta}_{\tau:H-1}}(o')| \leq \lambda_{H-\tau}\|o-o'\|_1 \quad (\forall o, o'),$$
$$\text{with } \lambda_{H-\tau} = \frac{V^{\max}_{H-\tau} - V^{\min}_{H-\tau}}{2}.$$

Considering now optimal solutions, this means that, at depth τ and for any $(o,o') \in O_\tau$:

$$\begin{aligned} V^*_\tau(o) - V^*_\tau(o') &= \max_{\beta^1_{\tau:H-1}} \min_{\beta^2_{\tau:H-1}} V_\tau(o, \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}) - \max_{\beta'^1_{\tau:H-1}} \min_{\beta'^2_{\tau:H-1}} V_\tau(o', \beta'^1_{\tau:H-1}, \beta'^2_{\tau:H-1}) \\ &\leq \max_{\beta^1_{\tau:H-1}} \min_{\beta^2_{\tau:H-1}} \left[V_\tau(o', \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}) + \lambda_{H-\tau}\|o-o'\|_1 \right] - \max_{\beta'^1_{\tau:H-1}} \min_{\beta'^2_{\tau:H-1}} V_\tau(o', \beta'^1_{\tau:H-1}, \beta'^2_{\tau:H-1}) \\ &= \lambda_{H-\tau}\|o-o'\|_1. \end{aligned}$$

Symmetrically, $V^*_\tau(o) - V^*_\tau(o') \geq -\lambda_{H-\tau}\|o-o'\|_1$, hence the expected result:

$$|V^*_\tau(o) - V^*_\tau(o')| \leq \lambda_{H-\tau}\|o-o'\|_1.$$ □

A.6 HSVI for POSGs

The following results help (i) tune zs-OMG-HSVI's radius parameter ρ, ensuring that trajectories will always stop, and (ii) then demonstrate the finite time convergence of this algorithm.

Lemma 5. (originally stated on page 7) *Assuming a single depth-independent Lipschitz constant λ, and noting that*

$$thr(\tau) = \gamma^{-\tau}\epsilon - 2\rho\lambda\frac{\gamma^{-\tau}-1}{1-\gamma}, \tag{3}$$

one can ensure positivity of the threshold at any $\tau \geq 1$ by enforcing $\rho < \frac{1-\gamma}{2\lambda}\epsilon$.

Proof. First, we have (for $\tau \geq 1$):

$$\begin{aligned} thr(\tau) &= \gamma^{-\tau}\epsilon - \sum_{i=1}^{\tau} 2\rho\lambda\gamma^{-i} \\ &= \gamma^{-\tau}\epsilon - 2\rho\lambda\gamma^{-1}\frac{\gamma^{-\tau}-1}{\gamma^{-1}-1} \\ &= \gamma^{-\tau}\epsilon - 2\rho\lambda\frac{\gamma^{-\tau}-1}{1-\gamma}. \end{aligned}$$

Then, let us derive the following equivalent inequalities:

$$\begin{aligned} 0 &< thr(\tau) \\ 2\rho\lambda\frac{\gamma^{-\tau}-1}{1-\gamma} &< \gamma^{-\tau}\epsilon \\ \rho &< \frac{1}{2\lambda}\frac{1-\gamma}{\gamma^{-\tau}-1}\gamma^{-\tau}\epsilon \\ \rho &< \frac{1}{2\lambda}\frac{1-\gamma}{1-\gamma^{\tau}}\epsilon. \end{aligned}$$

To ensure positivity of the threshold for any $\tau \geq 1$, one thus just needs to set ρ as a positive value smaller than $\frac{1-\gamma}{2\lambda}\epsilon$. □

Lemma 6. (originally stated on page 8) *Assuming a depth-independent Lipschitz constant λ, and with $W \stackrel{\text{def}}{=} \|U^{(0)} - L^{(0)}\|_\infty$, the length of trajectories is upper-bounded by*

$$T_{\max} \stackrel{\text{def}}{=} \left\lceil \log_\gamma \frac{\epsilon - \frac{2\rho\lambda}{1-\gamma}}{W - \frac{2\rho\lambda}{1-\gamma}} \right\rceil.$$

Proof. (detailed version) Since W is the largest possible width, any trajectory stops in the worst case at depth τ such that

$$\begin{aligned} thr(\tau) &< W \\ \gamma^{-\tau}\epsilon - 2\rho\lambda\frac{\gamma^{-\tau}-1}{1-\gamma} &< W \qquad \text{(from Eq. (3))} \\ \gamma^{-\tau}\underbrace{\left(\epsilon - \frac{2\rho\lambda}{1-\gamma}\right)}_{>0 \quad \text{(Lem. 5)}} &< W - \frac{2\rho\lambda}{1-\gamma} \\ \gamma^{-\tau} &< \frac{W - \frac{2\rho\lambda}{1-\gamma}}{\epsilon - \frac{2\rho\lambda}{1-\gamma}} \\ -\tau\ln(\gamma) &< \ln\left(\frac{W - \frac{2\rho\lambda}{1-\gamma}}{\epsilon - \frac{2\rho\lambda}{1-\gamma}}\right) \\ \tau &< \log_\gamma\left(\frac{\epsilon - \frac{2\rho\lambda}{1-\gamma}}{W - \frac{2\rho\lambda}{1-\gamma}}\right). \end{aligned}$$

□

Theorem 4. (originally stated on page 8) *Consider a trial $(o_0, \ldots, o_\tau)$ of length τ and consider that the backward updates of $U_{\tau-1}$ and $L_{\tau-1}$ have* **not yet** *been performed. Then*

1. *$width(\mathcal{K}_{o_{\tau-1}}\hat{V}(o_{\tau-1})) \leq thr(\tau-1) - 2\rho\lambda_\tau$, and*
2. *for every $o'_{\tau-1}$ satisfying $\|o'_{\tau-1} - o_{\tau-1}\|_1 \leq \rho$, it holds: $width(\mathcal{K}_{o_{\tau-1}}\hat{V}(o'_{\tau-1})) \leq thr(\tau-1)$.*

Proof. Observe that from the definition of the sequence $thr(\tau)$ in Equation (2) it follows that

$$\gamma thr(\tau) = thr(\tau-1) - 2\rho\lambda_\tau.$$

Moreover, the trial terminated at depth τ. Therefore, $o' = o_\tau$ (the only occupancy state that can be reached from $o_{\tau-1}$ when following $\beta^{U,1}, \beta^{L,2}$ from line 11 of Alg. 1) must satisfy

$$width(\hat{V}(o_\tau)) \leq thr(\tau).$$

Then:

$$\begin{aligned} width(\mathcal{K}_{o_{\tau-1}}\hat{V}(o_{\tau-1})) &= width(\mathcal{H}\hat{V}(o_{\tau-1})) \\ &= \mathcal{H}U(o_{\tau-1}) - \mathcal{H}L(o_{\tau-1}) && \text{(def. of } width(\cdot)) \\ &= \textsc{nev}(\Gamma^{o_{\tau-1}}(U)) - \textsc{nev}(\Gamma^{o_{\tau-1}}(L)) && \text{(def. of } \mathcal{H}) \end{aligned}$$

$$
\begin{aligned}
&= \max_{\beta^1} \min_{\beta^2} Q^{*,U}(o_{\tau-1}, \beta^1, \beta^2) - \min_{\beta^2} \max_{\beta^1} Q^{*,L}(o_{\tau-1}, \beta^1, \beta^2) \\
&= \min_{\beta^2} Q^{*,U}(o_{\tau-1}, \beta^{U,1}, \beta^2) - \max_{\beta^1} Q^{*,L}(o_{\tau-1}, \beta^1, \beta^{L,2}) \\
&\leq Q^{*,U}(o_{\tau-1}, \beta^{U,1}, \beta^{L,2}) - Q^{*,L}(o_{\tau-1}, \beta^{U,1}, \beta^{L,2}) \\
&= \gamma(U(T(o_{\tau-1}, \beta^{U,1}, \beta^{L,2})) - L(T(o_{\tau-1}, \beta^{U,1}, \beta^{L,2}))) \\
&= \gamma \mathit{width}(\hat{V}(o_\tau)) \\
&\leq \gamma thr(\tau) \\
&= thr(\tau - 1) - 2\rho\lambda_\tau.
\end{aligned}
$$

This proves the first point.
Now, the updated approximators $\mathcal{K}_{o_{\tau-1}} U_{\tau-1}$ and $\mathcal{K}_{o_{\tau-1}} L_{\tau-1}$ are both λ_τ-LC, which immediately gives the second point. □