HSVI Can Solve Zero-Sum Partially Observable Stochastic Games

Aurélien Delage, Olivier Buffet, Jilles Dibangoye, Abdallah Saffidine

HSVI CAN SOLVE ZERO-SUM PARTIALLY OBSERVABLE STOCHASTIC GAMES

A PREPRINT

Aurélien Delage
CITI
INSA Lyon
Villeurbanne
aurelien.delage@insa-lyon.fr

Olivier Buffet
INRIA - CNRS
Université de Lorraine
Villers-lès-Nancy
olivier.buffet@inria.fr

Jilles S. Dibangoye
CITI
INSA Lyon
Villeurbanne
jilles-steeve.dibangoye@inria.fr

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

October 27, 2022

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.

1 Introduction

Solving imperfect information sequential games is a challenging field with many applications from playing Poker [Kuhn, 1950] to security games [Basilico et al., 2016]. 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) [Oliehoek and Vlassis, 2006][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) [Koller et al., 1996, von Stengel, 1996, Bošanský et al., 2014], 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) [Zinkevich et al., 2007, Brown and Sandholm, 2018], or a first-order method [Hoda et al., 2010, Kroer et al., 2020], 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 [Moravčík et al., 2017, Brown and Sandholm, 2018, Brown et al., 2020]. A third approach, proposed by Wiggers [2015], 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.

[1] Note: POSGs are equivalent to the large class of "well-behaved" EFGs as defined by Kovařík et al. [2019].

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 [Åström, 1965, Smith, 2007], Dec-POMDP [Szer et al., 2005, Dibangoye et al., 2016], or subclasses of zs-POSGs with simplifying observability assumptions [Ghosh et al., 2004, Chatterjee and Doyen, 2014, Basu and Stettner, 2015, Horák et al., 2017, Cole and Kocherlakota, 2001, Horák and Bošanský, 2019]). 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. [2016b,a] 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.

Section 2 presents some necessary background, including the concept of *occupancy state* [Dibangoye et al., 2016, Wiggers et al., 2016a] (*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' [Smith and Simmons, 2005] 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. Section 5 discusses similarities and differences of our work with CFR-based continual resolving methods before concluding.

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:

θ^i_τ $= (a^i_0, z^i_1, \dots, a^i_{\tau-1}, z^i_\tau)$ 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 such 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.

$\beta^i_{\tau:\tau'} = (\beta^i_\tau, \dots, \beta^i_{\tau'})$ 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*.

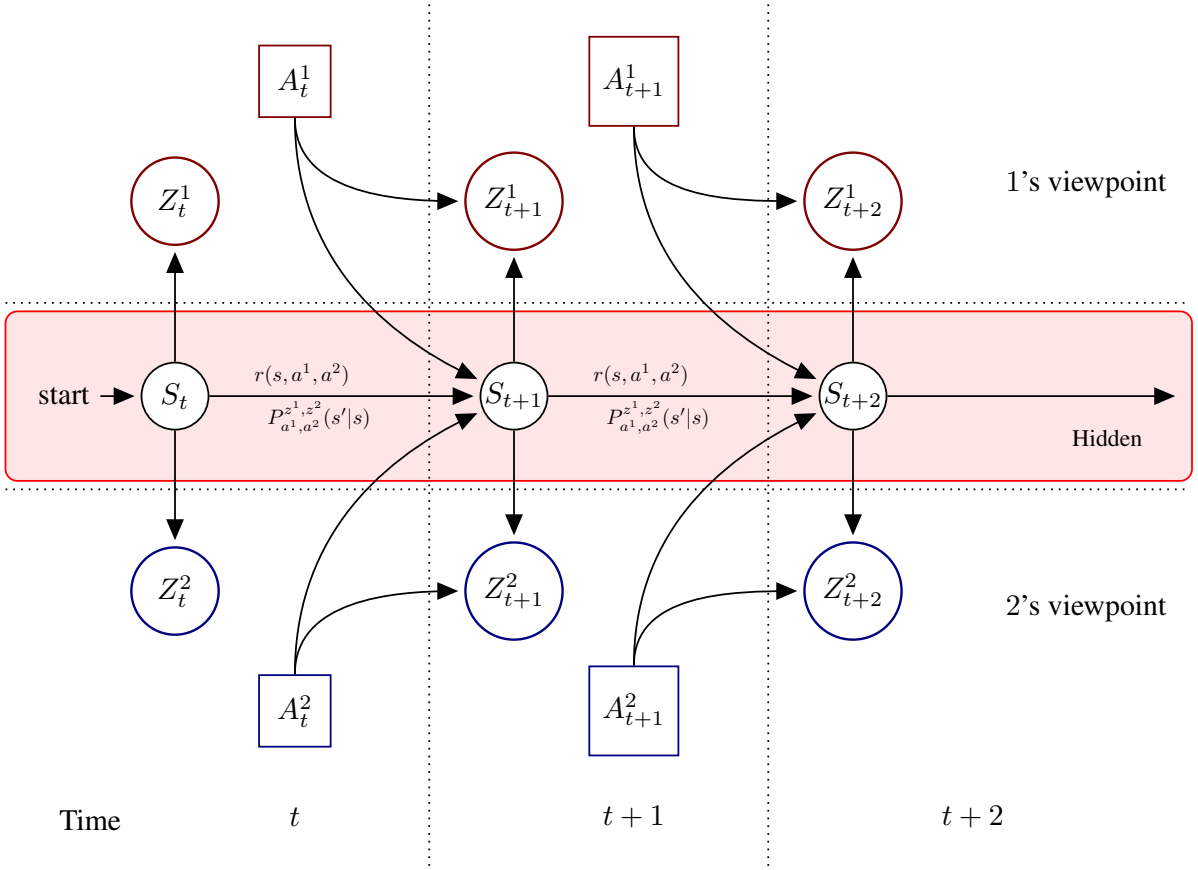

Figure 1: Dynamic influence diagram representing the evolution of a zs-POSG

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

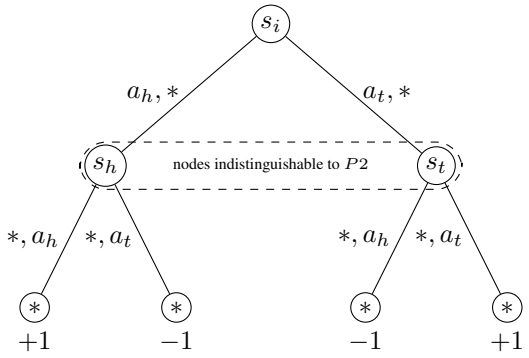

Figure 2: Simplified tree representation of the sequentialized Matching Pennies game. Irrelevant actions, noted $*$, allow merging edges with the same action for (i) player 2 at $t = 0$, and (ii) player 1 at $t = 1$. Notes: (a) Due to irrelevant actions, this game can be seen as an Extensive Form Game,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

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 $\neg i$.

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 penny's 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;
- $P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) = T(s, \boldsymbol{a}, s') \cdot \mathcal{O}(\boldsymbol{a}, s', \boldsymbol{z})$, using the next two definitions;
- T is deterministic and such that ($\cdot$ is used to denote "for all")
  - $T(\cdot, \cdot, a_h) = s_h$,
  - $T(\cdot, \cdot, a_t) = s_t$;
- $\mathcal{O}$ 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 $t = 1$). Then, we face a "subgame" at $t = 1$ 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 **?**Schmid [2021] (also called *safety issues* Burch et al. [2014]), 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.

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. [2016a], 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. [Dibangoye et al., 2016, Thm. 1] – Proof in 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* [Wiggers et al., 2016a, 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:

- $\mathcal{O}^\sigma (= \cup_{t=0}^{H-1} \mathcal{O}^\sigma_t)$ 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.

[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.

Bellman Optimality Equation

Despite the OS at $\tau > 0$ not being accessible to any player, let us define a *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* $\sigma^{m,1}_\tau$ and a *conditional term* $\sigma^{c,1}_\tau$, where

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

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

Now, if 1 faces AOH θ^1_τ, knows 2's future strategy $\beta^2_{\tau:H-1}$, and has access to $\sigma^{c,1}_\tau(\theta^2_\tau | \theta^1_\tau)$ for any θ^2_τ, then she faces a POMDP whose optimal value we denote $\nu^2_{[\sigma^{c,1}_\tau, \beta^2_{\tau:H-1}]}(\theta^1_\tau)$. This leads to defining the best-response value vector $\nu^2_{[\sigma^{c,1}_\tau, \beta^2_{\tau:H-1}]}$, which contains one component per AOH θ^1_τ, and writing the value of 1's best response against $\beta^2_{\tau:H-1}$ under σ_τ as $\sigma^{m,1}_\tau \cdot \nu^2_{[\sigma^{c,1}_\tau, \beta^2_{\tau:H-1}]}$. 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 ([Wiggers et al., 2016a, Thm. 2])**.** *For any $\tau \in \{0 \,..\, H-1\}$, V^*_τ is (i) concave w.r.t. $\sigma^{m,1}_\tau$ for a fixed $\sigma^{c,1}_\tau$, and (ii) convex w.r.t. $\sigma^{m,2}_\tau$ for a fixed $\sigma^{c,2}_\tau$. More precisely,*

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

$\nu^2_{[\sigma^{c,1}_\tau, \beta^2_{\tau:H-1}]}(\theta^1_\tau) \stackrel{\text{def}}{=} \max_{\beta^1_{\tau:H-1}} \mathbb{E}_{\theta^2_\tau \sim \sigma^{c,1}_\tau(\theta^1_\tau)} \left\{\sum_{t=\tau}^{H-1} \gamma^{t-\tau} r(S_t, A^1_t, A^2_t) \mid \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}\right\}$.

Proof. (Sketch) We start from von Neumann's Minimax theorem [von Neumann, 1928] giving the following equation:

$$\begin{aligned} V^*_\tau(\sigma_\tau) &= \min_{\beta^2_{\tau:H-1}} \max_{\beta^1_{\tau:H-1}} \left[V_\tau(\sigma_\tau, \beta^1_{\tau:H-1}, \beta^1_{\tau:H-1})\right] \\ &= \min_{\beta^2_{\tau:H-1}} \max_{\beta^1_{\tau:H-1}} \left[\mathbb{E}\left\{\sum_{t=\tau}^{H-1} \gamma^{t-\tau} r(S_t, A^1_t, A^2_t) \mid \theta^1_\tau, \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}, \sigma^{c,1}_\tau\right\}\right], \end{aligned}$$

then, observing that 1's best response to $\beta^2_{\tau:}$ can be computed for each AOH θ^1_τ independently, we can swap the $\max$ operator and part of the expectation one ($\mathbb{E}$) as follows:[4]

$$= \min_{\beta^2_{\tau:H-1}} \mathbb{E}_{\theta^1_\tau \sim \sigma^{m,1}_\tau} \underbrace{\left\{\max_{\beta^1_{\tau:H-1}} \mathbb{E}_{\theta^2_\tau \sim \sigma^{c,1}_\tau(\theta^1)} \left[\sum_{t=\tau}^{H-1} \gamma^{t-\tau} r(S_t, A^1_t, A^2_t) \mid \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}\right]\right\}}_{\text{best-response of 1 to } \beta^2_{\tau:} \text{ under } \theta^1_\tau}$$

and, recognizing the components of vector $\nu^2_{[\sigma^{c,1}_\tau, \beta^2_{\tau:H-1}]}$ and writing the expectation over AOHs $\theta^1_{\tau:}$ as a scalar product:

$$= \min_{\beta^2_{\tau:H-1}} \left[\sigma^{m,1}_\tau \cdot \nu^2_{[\sigma^{c,1}_\tau, \beta^2_{\tau:H-1}]}\right].$$

□

[3] We will come back to the validity of this point in Section 3.1.

An important observation that ensues from this theorem is that V^*_τ is concave in $\sigma^{m,1}_\tau$ and convex in $\sigma^{m,2}_\tau$. In practice, however, such continuity properties alone only allow upper-bounding V^*_τ for finitely many conditional terms $\sigma^{c,i}_\tau$, 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.

Note: To help the reader, Appendix A provides two synthetic tables: Table 3 (p. 23) to sum up various theoretical properties that are stated in this paper (assuming a finite temporal horizon), and Table 4 (p. 24) to sum up the notations used in this paper, including some that are used only in the appendix.

Also, 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. [Wiggers et al., 2016a, 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 [von Neumann, 1928] applies when in σ_τ, thus justifying the definition of the optimal (Nash equilibrium) value of a subgame.

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

$$V^*_\tau(\sigma_\tau) \overset{\text{def}}{=} \max_{\beta^1_{\tau:H-1}} \min_{\beta^2_{\tau:H-1}} V_\tau(\sigma_\tau, \beta^1_{\tau:H-1}, \beta^2_{\tau:H-1}). \tag{4}$$

$V(\sigma_\tau,\cdot,\cdot)$ not being bilinear in the space of behavioral strategies (Appendix 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 [Kuhn, 1950] to the subgames.

Then, defining the optimal action-value function:

$$Q^*_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau) \overset{\text{def}}{=} r(\sigma_\tau, \boldsymbol{\beta}_\tau) + \gamma V^*_{\tau+1}(T(\sigma_\tau, \boldsymbol{\beta}_\tau)), \tag{5}$$

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 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.

[4]Note that this property is well known in Bayesian games, where AOHs correspond to *types*, *cf.* [Harsanyi, 1968, Th. 1, p. 321].

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.* App. D.2), for instance:

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

where $\overline{\nu}^2_\tau$ component-wise upper-bounds $\nu^2_{[\tilde{\sigma}^{c,1}_\tau, \beta^2_{\tau,:}]}$ 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 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 $\max\text{-}\min$ 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:

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

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 $w = \langle \tilde{\sigma}_\tau, \beta^2_\tau, \langle \overline{\nu}^2_{\tau+1}, \beta^2_{\tau+1:} \rangle \rangle$ stored in sets $\overline{\mathcal{I}}_\tau$, and where $\overline{\nu}^2_{\tau+1}$ upper-bounds $\nu^2_{[\tilde{\sigma}^{c,1}_{\tau+1}, \beta^2_{\tau+1:}]}$.

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

$$\begin{aligned} \overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \overset{\text{def}}{=} \min_{\langle \tilde{\sigma}_\tau, \beta^2_\tau, \langle \overline{\nu}^2_{\tau+1}, \beta^2_{\tau+1:} \rangle \rangle \in \overline{\mathcal{I}}_\tau} & \left[r(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) + \gamma T^1_m(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) \cdot \overline{\nu}^2_{\tau+1} \right. \\ & \left. + \lambda_{\tau+1} \| T(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) - T^1_m(\sigma_\tau, \beta^1_\tau, \beta^2_\tau) T^1_c(\tilde{\sigma}^{c,1}_\tau, \beta^2_\tau) \|_1 \right] \end{aligned} \tag{6}$$

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. (6) allow writing $\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) = \min_{w \in \overline{\mathcal{I}}_\tau} {\beta^1_\tau}^\top \cdot M^{\sigma_\tau}_{(\cdot,w)}$, 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 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 App. D.3.1)**.** *For any given σ_τ and any set $\overline{\mathcal{I}}_\tau$ of tuples $w = \langle \tilde{\sigma}_\tau, \beta^2_\tau, \langle \overline{\nu}^2_{\tau+1}, \beta^2_{\tau+1:}\rangle\rangle$, $\max_{\beta^1_\tau} \overline{W}_\tau(\sigma_\tau, \beta^1_\tau)$ is equivalent to the LP and dual LP:*

$$
\begin{array}{lllllll}
\mathrm{LP}\overline{W}_\tau(\sigma_\tau) : & \max\limits_{\beta^1_\tau, v} v & \textit{s.t.} & \textit{(i)} & \forall w \in \overline{\mathcal{I}}_\tau, & v \le {\beta^1_\tau}^\top \cdot M^{\sigma_\tau}_{(\cdot,w)} & \textit{and} \\
 & & & \textit{(ii)} & \forall \theta^1_\tau \in \Theta^1_\tau, & \sum\limits_{a^1} \beta^1_\tau(a^1|\theta^1_\tau) = 1, & \\
\mathrm{DLP}\overline{W}_\tau(\sigma_\tau) : & \min\limits_{\psi^2_\tau, v} v & \textit{s.t.} & \textit{(i)} & \forall (\theta^1_\tau, a^1), & v \ge M^{\sigma_\tau}_{((\theta^1_\tau,a^1),\cdot)} \cdot \psi^2_\tau & \textit{and} \\
 & & & \textit{(ii)} & & \sum\limits_{w \in \overline{\mathcal{I}}_\tau} \psi^2_\tau(w) = 1. &
\end{array}
\tag{7}
$$

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, as illustrated by Fig. 3, which can be turned into a behavioral strategy $\beta^2_{\tau:H-1}$ (more details in 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.

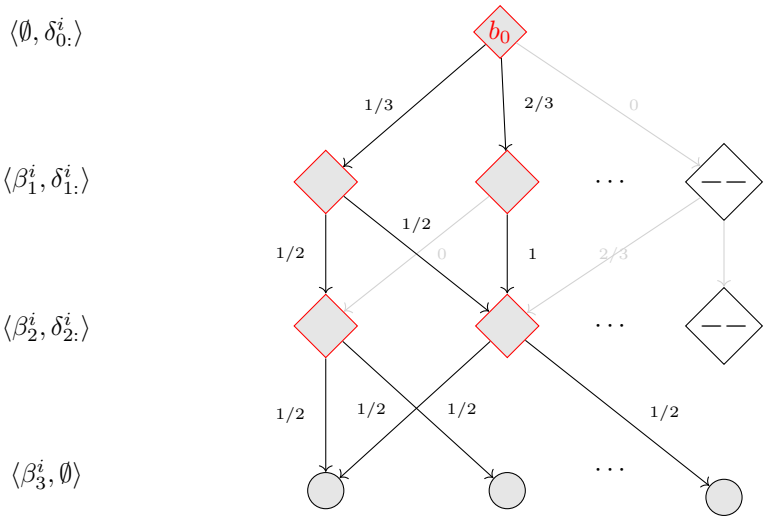

Figure 3: Representation of the strategy recursively induced by some ψ^1_0. At each time step τ, one must (i) sample a next tuple/node w^1_τ from current distribution ψ^1_τ, (ii) apply DR $\beta^1_\tau[w^1_\tau]$, and (iii) make $\psi^1_{\tau+1}[w^1_\tau]$ the new current distribution (unless reaching a leaf).

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 $\overline{\nu}^2_\tau$.

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

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

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 $\mathrm{DLP}\overline{W}_\tau(\sigma_\tau)$ provides both*
  - *a tree strategy ψ^2_τ (as the main solution of the DLP), and*
  - *a vector $\overline{\nu}^2_\tau = \frac{1}{\sigma^{m,1}_\tau} M^{\sigma_\tau}_{((\theta^1_\tau,a^1),\cdot)} \cdot \psi^2_\tau$ (as a by-product).*

Then,

1. *$\overline{\mathcal{I}}_{\tau-1} \leftarrow \overline{\mathcal{I}}_{\tau-1} \cup \{\langle \sigma^{c,1}_{\tau-1}, \beta^2_{\tau-1}, \langle \overline{\nu}^2_\tau, \psi^2_\tau\rangle\rangle\}$ is a valid update operator in the sense that it preserves $\overline{W}_\tau$'s upper-bounding property, and*
2. *similarly, $\overline{\mathcal{J}}_\tau \leftarrow \overline{\mathcal{J}}_\tau \cup \{\langle \sigma^{c,1}_\tau, \langle \overline{\nu}^2_\tau, \psi^2_\tau\rangle\rangle\}$ 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 $w_{\tau-1,init} = \langle \sigma^{c,1}_{\tau-1}, \beta^2_{\tau-1}, \langle \overline{\nu}^2_\tau, \psi^2_\tau \rangle \rangle$, 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
- $\overline{\nu}^2_\tau(\theta^1_\tau) = r_{max} \cdot (H - \tau)$ 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 $\tau \geq 0$, we similarly create tuples $\langle \sigma^{c,1}_\tau, \langle \overline{\nu}^2_\tau, \psi^2_{\tau:} \rangle \rangle$ 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 withing ϵ from each other, in particular at $\tau = 0$.

3.2.4 Extracting a NES

Vectors $\overline{\nu}^2_0$ 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 $\overline{V}_0(\sigma_0) - \underline{V}_0(\sigma_0) \leq \epsilon$, then $\arg\max_{\underline{w} \in \underline{\mathcal{J}}_0} \underline{\nu}^2_0$ and $\arg\min_{\overline{w} \in \overline{\mathcal{J}}_0} \overline{\nu}^2_0$ 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 ν 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

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

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

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

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. Appendix 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 (7). At $\tau = H-1$, line 14 selects DRs by solving an exact game, and line 20 returns a distribution reduced to the single element added in line 15.

A key difference with Smith and Simmons's HSVI algorithm [Smith and Simmons, 2005] 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 Horák et al. [2017] 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** $\arg\max_{w_0 \in \underline{\mathcal{J}}_0} \underline{\nu}_0^1$

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 $\langle \overline{\nu}_\tau^2, \psi_\tau^2 \rangle \leftarrow \text{DLP}\overline{W}_\tau(\sigma_\tau,)$
21 $\overline{\mathcal{I}}_{\tau-1} \leftarrow \overline{\mathcal{I}}_{\tau-1} \cup \{\langle \sigma_{\tau-1}^{c,1}, \underline{\beta}_{\tau-1}^2, \langle \overline{\nu}_\tau^2, \psi_\tau^2 \rangle \rangle\}$
22 $\overline{\mathcal{J}}_\tau \leftarrow \overline{\mathcal{J}}_\tau \cup \{\langle \sigma_\tau^{c,1}, \langle \overline{\nu}_\tau^2, \psi_\tau^2 \rangle \rangle\}$

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.* Prop. E.1, App. E.1.1), 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 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 App. E.2.2)**.** *When $\gamma < 1$, the length of trajectories is upper bounded by $T_{\max} \overset{\text{def}}{=} \left\lceil \log_\gamma \frac{\epsilon - \frac{2\rho\lambda^\infty}{1-\gamma}}{W - \frac{2\rho\lambda^\infty}{1-\gamma}} \right\rceil$, where λ^∞ is a depth-independent Lipschitz constant and $W \overset{\text{def}}{=} \|\overline{V}^{(0)} - \underline{V}^{(0)}\|_\infty$ 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 [2015]. 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 $\gamma = 1$. Table 1 gives the cardinal of the state, action and observation sets for each of these problems.

Table 1: Number of states/actions/observations for each benchmark problem

	$\mathcal{S}$	$\mathcal{A}^1$	$\mathcal{A}^2$	$\mathcal{O}^1$	$\mathcal{O}^2$
Competitive Tiger	2	4	4	3	3
Adversarial Tiger	2	3	2	2	2
Recycling Robot	4	3	3	2	2
Mabc	4	2	2	2	2
Matching Pennies	3	2	2	1	1

Algorithms

For conciseness, Algorithm 1 is here denoted HSVI, and compared against *Random* search and *Informed* search Wiggers [2015] (both using Wiggers's implementation (unlicensed and unreleased)), SFLP [Koller et al., 1996], and CFR+ Tammelin [2014] (both using open_spiel [Lanctot et al., 2019] (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 [Dibangoye et al., 2016]. 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 $V_0^*(\sigma_0)$:

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

Table 2: 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]	**10.6981981981982 %**	[OOT]	[OOT]
	6^*	[OOT]	[OOT]	**45.51426426426426 %**	[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

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* $\langle \beta^1_{0:}, \beta^2_{0:} \rangle$:

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

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.*,

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

4.2.1 Comparison with the state of the art

Table 2 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).

We now study the dynamic behavior of the algorithms at hand by providing and analyzing the bounds and exploitability graphs for the same benchmarks.

4.2.2 Bounding Graphs

Left-side graphs in Figures 4 to 9 show how the computed upper- and lower-bounding values $\overline{V}_0(\sigma_0)$ and $\underline{V}_0(\sigma_0)$ (respectively the *dotted* dark and light green curves) evolve as a function of computation time (always given in seconds). The *solid* dark and light green curves show the security levels $\nu^1_{[\sigma_0,\psi^2_0]}$ and $\nu^2_{[\sigma_0,\psi^1_0]}$ of the current returned strategies ψ^2_0 and ψ^1_0. Note that, when best-response computations to obtain security levels are expensive (*e.g.*, for the competitive tiger problem, with $H = 4$), they are performed either periodically (*e.g.* every 10 iterations) or only once, at the end. In the captions, we indicate the (arbitrary) frequency of the POMDP evaluations. For example, $(1, 1, once)$ means that, for the first two horizons, the POMDP evaluations were done after each iteration, and, for the last one, only once (at the end).

Overall, we observe consistent curves with (i) security levels in-between bounds and around the NEV, and (ii) bounds converging monotonically. Note that HSVI stops when the gap between bounds is small enough, while the gap between SLs (used by Informed, Random and CFR, and whose computation can be time-consuming) can be much smaller. As a matter of fact, one can notice that strategies ψ^i_0 returned at each iteration by HSVI are often better (in terms of security level) than their pessimistic lower- or upper-bounding guarantees $\underline{\nu}^1_{[\sigma_0,\psi^1_0]}$ and $\overline{\nu}^2_{[\sigma_0,\psi^2_0]}$.

4.2.3 Exploitability Graphs

Right-side graphs in Figures 4 to 9 show the exploitability of the returned strategy profile as a function of computation time for HSVI, Random, Informed, and CFR+ for the different benchmarks considered. A limit precision of 10^{-7} (chosen empirically, according to the LP solver's precision) was applied to HSVI's exploitability.

As can be observed, Random and Informed tend to produce reasonable strategies quickly, but struggle to improve them so as to converge towards an ϵ-NES with $\epsilon \simeq 0$. In contrast, our algorithm keeps improving as computation time increases. The exploitation graphs support the observed behavior in Table 2 that HSVI converges in reasonable time compared to Wiggers's algorithms. However, the graphs also show that CFR+ essentially outperforms HSVI when the problems are difficult enough (*i.e.*, when the temporal horizon grows) but the traversal of the whole tree still remains tractable (thus allowing CFR+ to perform iterations). An interesting observation is that, on small enough problems, HSVI achieves very low exploitabilities earlier than CFR+.

Finally, HSVI's exploitability graph shares strong similarities with those of Bošanský et al.'s double-oracle algorithms [Bošanský et al., 2014, Fig. 8 and 11]. This can be understood as HSVI iteratively building two sets of strategies, one per player, until they are sufficient to support NES profiles, so that the average exploitability is almost zero. But note that Bošanský et al. construct LPs using pure strategies (deterministic best responses), while HSVI's strategies are stochastic.

Having empirically studied the behavior of HSVI compared to other basic offline solvers, we now provide insight about the connections between HSVI and continual (thus online) resolving methods.

5 Comparison with Continual Resolving

Continual Resolving techniques share some similarities with our approach, but also important differences. The purpose of this section is to clarify these points. It starts with a quick description of CFR, on which Continual Resolving is built.

Counterfactual Regret Minimization (CFR) Zinkevich et al. [2007] belongs to the self-play family of algorithms, which gave rise to several CFR-based approaches [Tammelin, 2014, Burch et al., 2019, Lanctot et al., 2009, Brown et al., 2017]). It iteratively traverses the whole game tree and applies, in each private history, a regret-matching update rule based on a specific type of regret called counterfactual regrets. Iteratively updating an initially uniform strategy asymptotically converges towards a NES. However, the tree traversal becomes intractable when the tree size is large.

Built on top of the CFR framework, approaches based on *limited-lookahead continual resolving* (LLCR) (inspired by Burch et al.'s decomposition Burch et al. [2014]) such as DeepStack Moravčík et al. [2017], Libratus Brown and Sandholm [2018], ReBeL Brown et al. [2020] and Player of Games Schmid et al. [2021], perform well by exploiting a temporal decomposition in subgames, which are specified through knowledge about both players' past strategies. Our approach thus shares similarities with these works.

Yet, as we will see in the next sections, a closer look at LLCR [Schmid, 2021] demonstrates how fundamentally different they are, starting with the fact that LLCR is an *online* search algorithm, *i.e.*, is meant to make good decisions

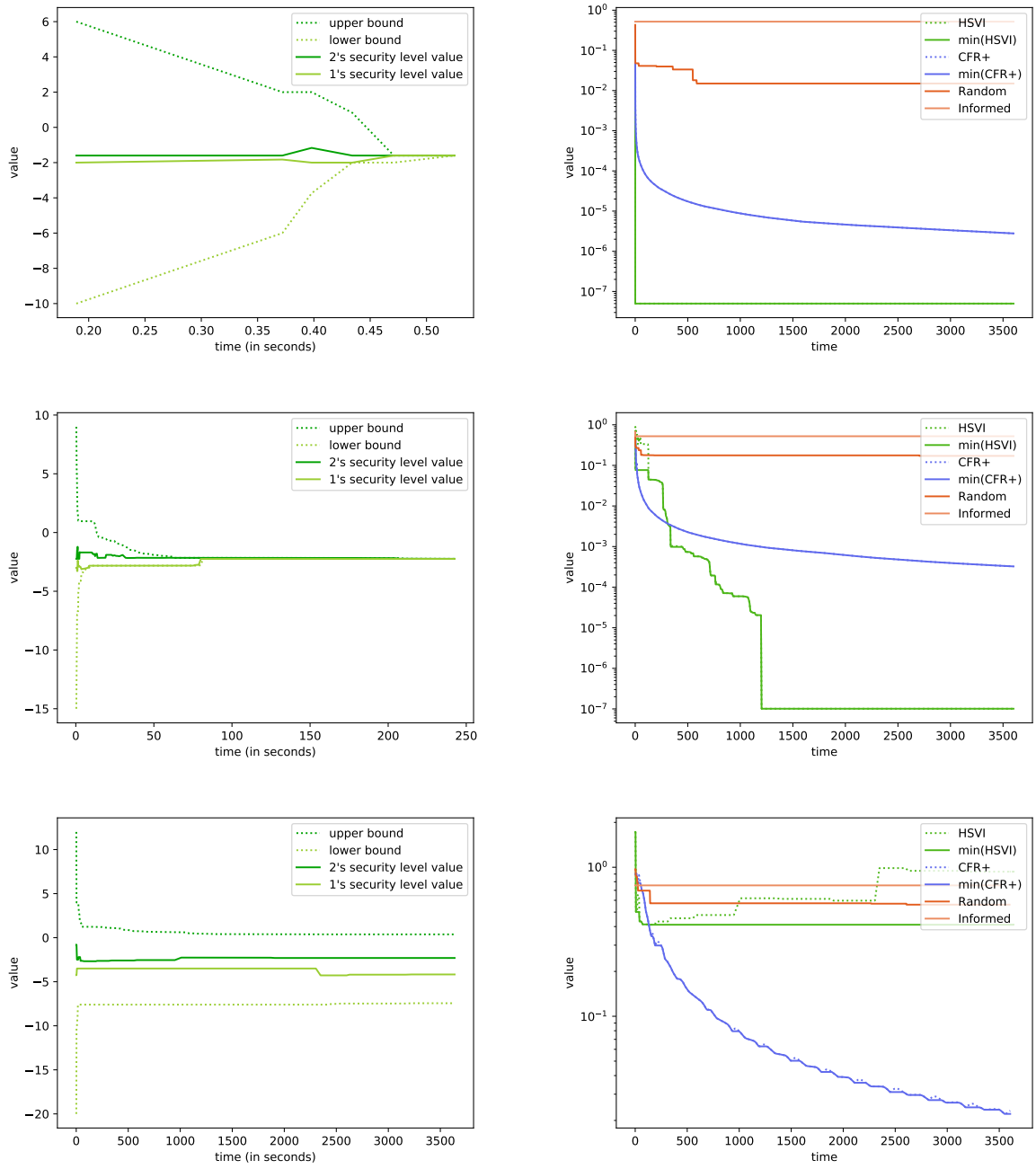

Figure 4: **Adversarial Tiger ($H = 2, 3, 4$) (1,1,10): (left)** Evolution of (in dotted lines) the upper- and lower-bound values, and (in solid lines) the security levels of the returned strategies for HSVI as a function of time (ms). **(right)** Exploitability ($= \frac{\text{SL-gap}}{2}$) as a function of time (s) for Random, Informed, CFR+, and HSVI.

at each time step, based on the current knowledge about the state of the game, while HSVI, as SFLP or CFR (and its variants), is an *offline* algorithm returning a complete solution strategy.

5.1 Continual Resolving

(Continual) Resolving techniques have been the first ingredient to adapt CFR to online settings. They address the problem of solving the complete subgame (down to its end) starting in the current situation at τ, while maintaining the global consistency (aka safety) of the whole strategy, *i.e.*, not making choices that could encourage the opponent to deviate *in the past*, before τ. This is achieved by introducing constraints, called *gadgets*, in a preliminary stage of the subgame that represent possible deviations and their values, but increase the size of the game tree so that

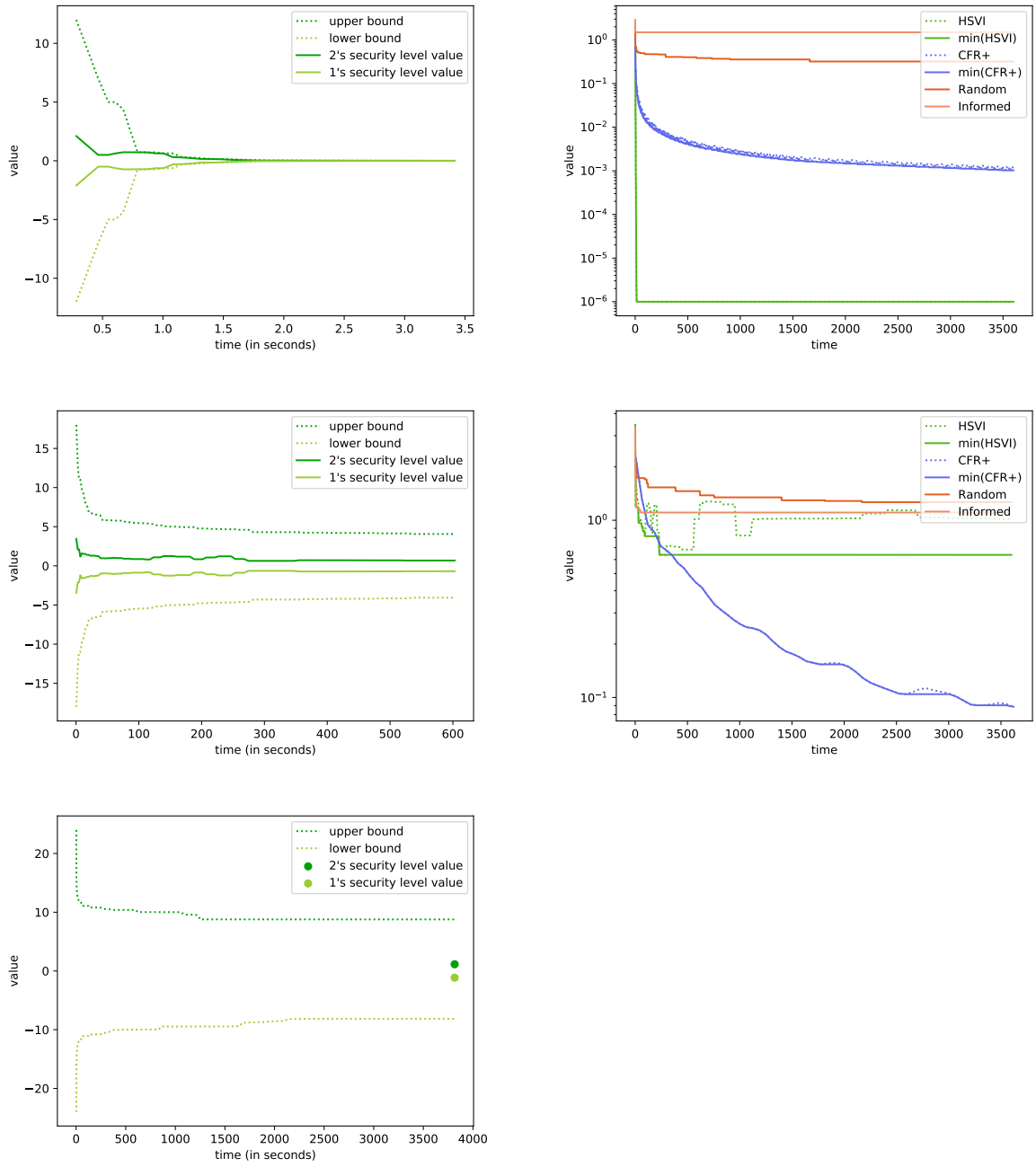

Figure 5: **Competitive Tiger ($H = 2, 3$) (1,1): (left)** Evolution of (in dotted lines) the upper- and lower-bound values, and (in solid lines) the security levels of the returned strategies for HSVI as a function of time (ms). **(right)** Exploitability ($= \frac{\text{SL-gap}}{2}$) as a function of time (s) for Random, Informed, CFR+, and HSVI.

it is practically intractable Moravčík et al. [2017]. For its part, HSVI solves similar subgames, but ensuring only local consistency, *i.e.*, only considering the subgame. Global consistency comes from the way HSVI combines "lower-level" solutions in its backtracking process, without adding any gadget modifying the game.

For both Resolving and HSVI, solving a subgame requires *sufficient statistics* that represent a prefix strategy profile from $t = 0$ to τ. In Resolving's online setting, this may seem surprising, since the opponent's actual strategy is not public. Yet, Resolving does not actually require knowing or guessing the opponent's actual strategy. In Resolving, any opponent Nash equilibrium strategy is appropriate, since the purpose is to verify that the opponent has no incentive to deviate from the Nash equilibrium. A requirement for Resolving is for the sufficient statistics to represent *complete* strategies (given the current public information), so that decisions are anticipated for player i even in AOHs (infostates) not reachable given player $\neg i$'s strategy. This leads to using *ranges* Kovařík et al. [2019], a vector that

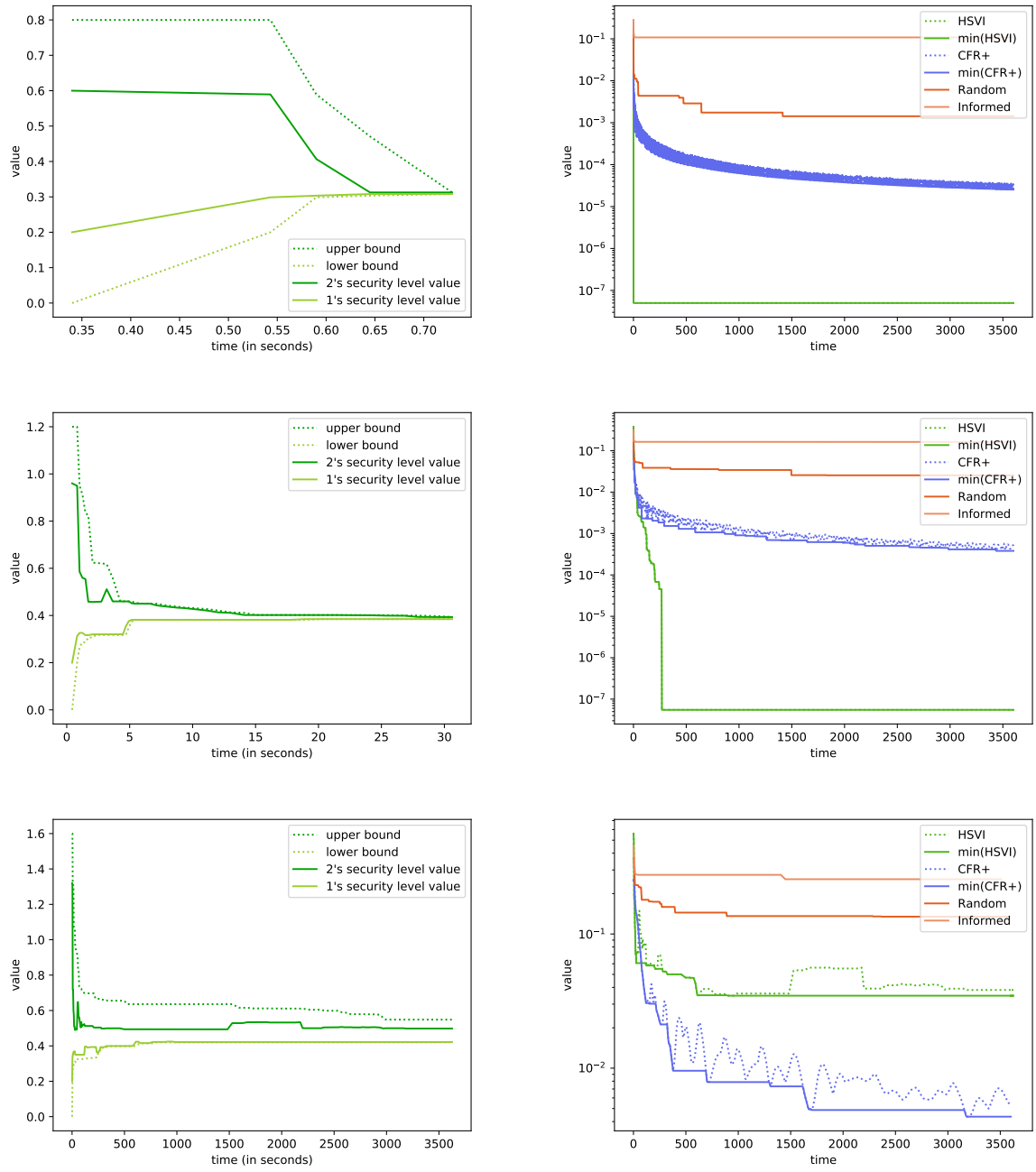

Figure 6: **Mabc** ($H = 2, 3, 4$) **(1,1,10): (left)** Evolution of (in dotted lines) the upper- and lower-bound values, and (in solid lines) the security levels of the returned strategies for HSVI as a function of time (ms). **(right)** Exploitability ($= \frac{\text{SL-gap}}{2}$) as a function of time (s) for Random, Informed, CFR+, and HSVI.

gives, for each player, her contribution to the probability of any history she could face at time step τ. In contrast, HSVI's occupancy state is not necessarily related to a Nash equilibrium strategy in any manner, and leads to ignoring unreachable AOHs, which helps to reduce the size of the decision-making (sub)problem.

5.2 Limited Lookahead

Continual Resolving alone solves complete subgames, thus larger problems at early stages of the game than at the end, which is not appropriate in an online setting. To address this issue through limiting the lookahead of subgames, one needs to estimate the value of the leaves of any truncated subgame. This is achieved through learning offline, for each player i, deep networks that, given the current public belief state, map each AOH to its value under some Nash

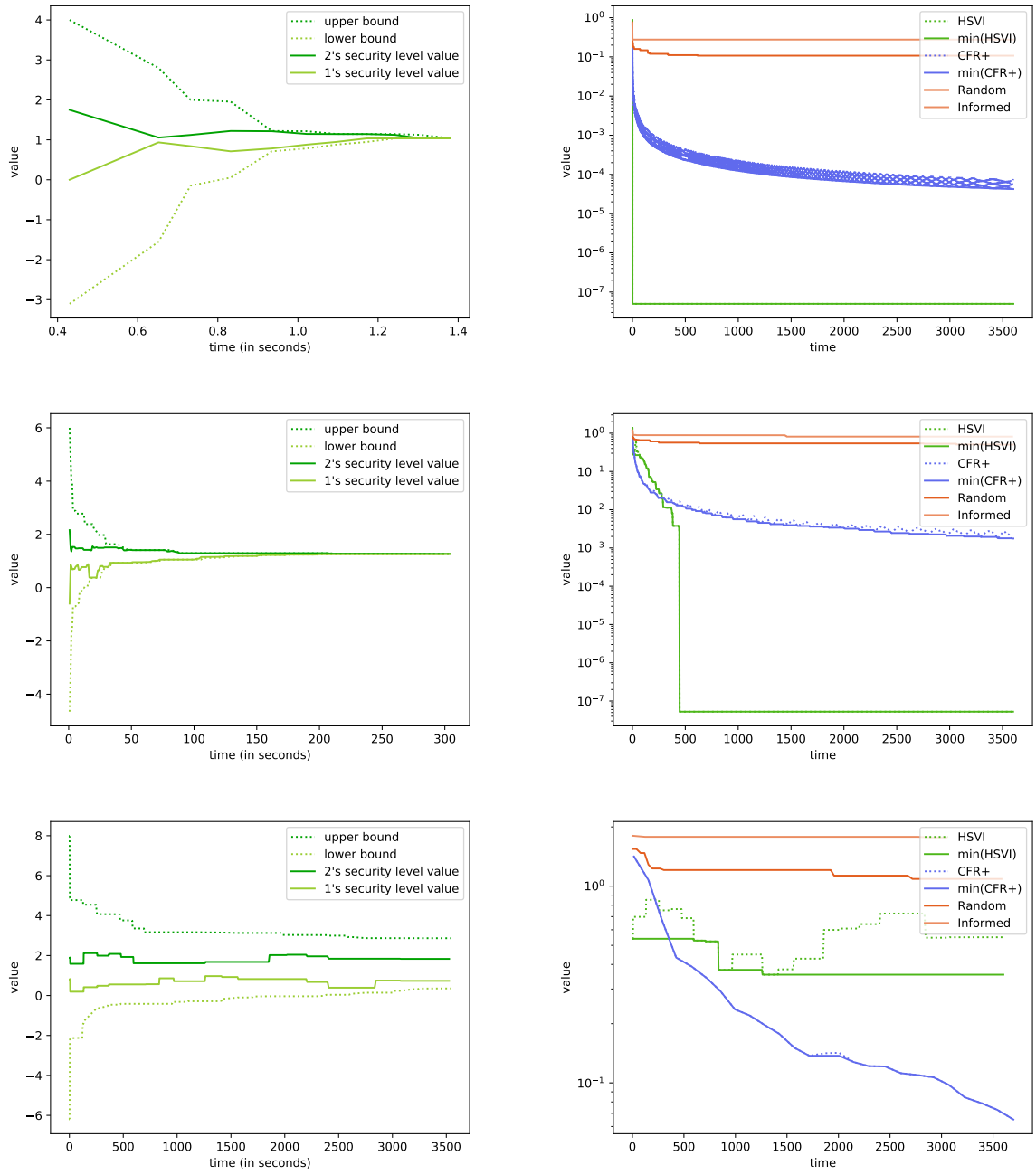

Figure 7: **Recycling Robot** ($H = 2, 3, 4$) **(1,1,10): (left)** Evolution of (in dotted lines) the upper- and lower-bound values, and (in solid lines) the security levels of the returned strategies for HSVI as a function of time (ms). **(right)** Exploitability ($= \frac{\text{SL-gap}}{2}$) as a function of time (s) for Random, Informed, CFR+, and HSVI.

equilibrium strategy profile. Note that the target function is not unique [Kovařík et al., 2019, Proposition A.1], since each NES profile maps to different value vectors. Still, according to Kovařík et al. [2019], this does not seem to cause problems in practice.

In contrast, the individual value functions HSVI considers (the "ν" functions) are uniquely defined since they correspond to the best responses to given (not necessarily Nash equilibrium) strategies of the opponent.

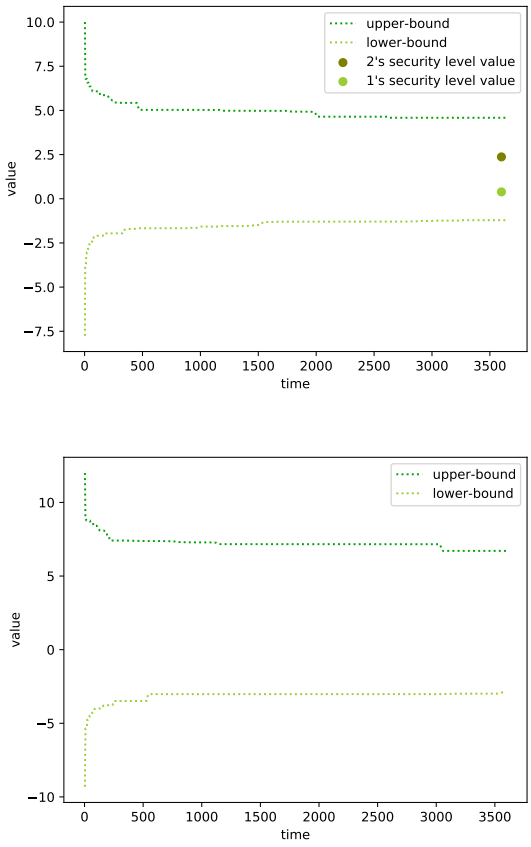

Figure 8: **Recycling Robot ($H = 5, 6$) (once,none): (left)** Evolution of (in dotted lines) the upper- and lower-bound values, and (in solid lines) the security levels of the returned strategies for HSVI as a function of time (ms).

5.3 Limited Lookahead Continual Resolving as a General Scheme?

The previous subsections highlight to what extent HSVI and LLCR are fundamentally different, in particular because they are not on the same algorithmic level. LLCR should be seen as a general scheme in which the subgame solver used, namely CFR, could be replaced by other "basic" offline algorithms such as HSVI or SFLP. But we leave further investigation on this topic for future work.

6 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.

Despite some similarities, our (offline) approach is fundamentally different from (online) continual resolving approaches. The latter could even possibly be adapted to use other offline methods than CFR-based ones, including HSVI.

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

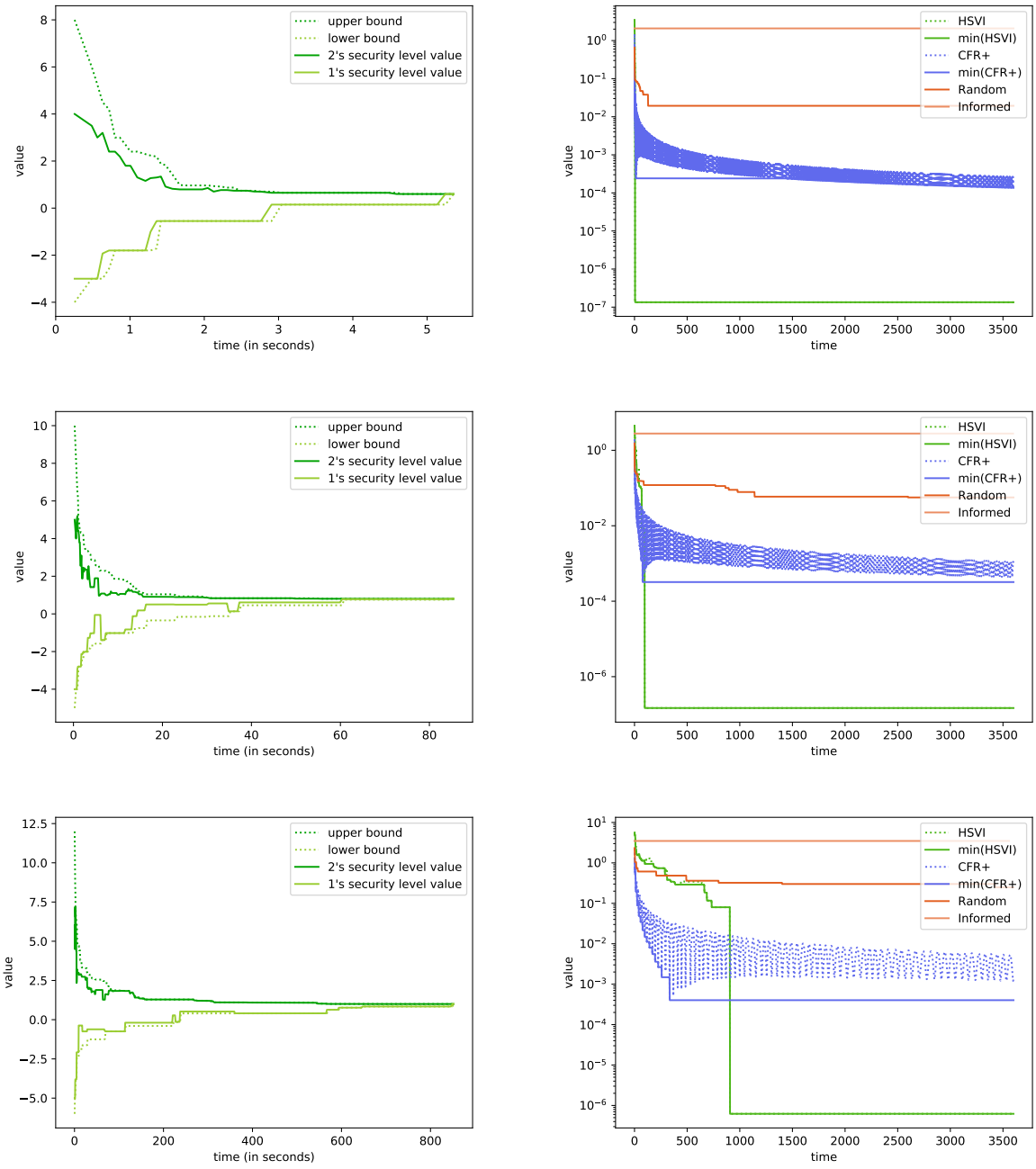

Figure 9: **Matching Pennies** ($H = 4, 5, 6$) **(1,1,1): (left)** Evolution of (in dotted lines) the upper- and lower-bound values, and (in solid lines) the security levels of the returned strategies for HSVI as a function of time (ms). **(right)** Exploitability ($= \frac{\text{SL-gap}}{2}$) as a function of time (s) for Random, Informed, CFR+, and HSVI.

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 [Bono et al., 2018]. 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.

References

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.

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.

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.

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.

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

Noam Brown, Christian Kroer, and Tuomas Sandholm. Dynamic thresholding and pruning for regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

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.

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

Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting cfr+ and alternating updates. *Journal of Artificial Intelligence Research*, 64:429–443, 2019.

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

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

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.

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.

John C. Harsanyi. Games with incomplete information played by "Bayesian" players, I-III. part II. Bayesian equilibrium points. *Management Science*, 14(5):320–334, January 1968. URL `http://www.jstor.org/stable/2628673`.

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`.

Karel Horák. *Scalable Algorithms for Solving Stochastic Games with Limited Partial Observability*. PhD thesis, Czech Technical University in Prague, Faculty of Electrical Engineering, 2019.

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.

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.

Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Fast algorithms for finding randomized strategies in game trees. In *Proceedings of the 26th ACM Symposium on the Theory of Computing (STOC'94)*, pages 750–759, 1994.

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.

Vojtěch Kovařík, Dominik Seitz, Viliam Lisỳ, Jan Rudolf, Shuo Sun, and Karel Ha. Value functions for depth-limited solving in imperfect-information games. *arXiv preprint arXiv:1906.06412*, 2019.

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.

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.

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.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. *Advances in neural information processing systems*, 22, 2009.

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. URL `http://arxiv.org/abs/1908.09453`.

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.

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.

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

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

Martin Schmid, Matej Moravcik, Neil Burch, Rudolf Kadlec, Joshua Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, Zach Holland, Elnaz Davoodi, Alden Christianson, and Michael Bowling. Player of games. *CoRR*, abs/2112.03178, 2021. URL `https://arxiv.org/abs/2112.03178`.

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

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.

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.

Oskari Tammelin. Solving large imperfect information games using cfr+. *arXiv preprint arXiv:1407.5042*, 2014.

John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100, 1928.

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

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

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, 2016a.

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, 2016b. doi: 10.3233/978-1-61499-672-9-1628.

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.

A Synthetic Tables

For convenience, we provide two synthetic tables: Table 3 to sum up various theoretical properties that are stated in this paper (assuming a finite temporal horizon), and Table 4 to sum up the notations used in this paper, adding some notations that appear only in the appendix.

More precisely, Table 3 indicates, for various functions f and variables x, properties that f is known to exhibit with respect to x. We denote by

-	a function with no known (or used) property (see also comment below);
N/A	a non-applicable case;
Lin	a linear function;
LC	a Lipschitz-continuous function;
Cv	(resp. Cc) a convex (resp. concave) function;
$PWLCv$	(resp. $PWLCc$) a piecewise linear and convex (resp. concave) function;
$\perp\!\!\!\perp$	the function being independent of the variable;
$\neg P$	the negation of some property P (*i.e.*, P is known not to hold).

Note also that, as $\sigma_\tau = \sigma_\tau^{c,1}\sigma_\tau^{m,1}$, the linearity or Lipschitz-continuity properties of any function w.r.t. σ_τ extends to both $\sigma_\tau^{c,1}$ and $\sigma_\tau^{m,1}$. Reciprocally, related negative results extend from $\sigma_\tau^{c,1}$ or $\sigma_\tau^{m,1}$ to σ_τ. In these three columns, we just indicate results that cannot be derived from one of the two other columns.

Table 3: Known properties of various functions appearing in this work

	σ_τ	$\sigma_\tau^{m,1}$	$\sigma_\tau^{c,1}$	β_τ^i	$\beta_\tau^{\neg i}$
$T(\sigma_\tau, \boldsymbol{\beta}_\tau)$	Lin (prop. 2.3, p. 5)	-	-	Lin (prop. 2.3, p. 5)	Lin (prop. 2.3, p. 5)
$T_m^i(\sigma_\tau, \boldsymbol{\beta}_\tau)$	Lin (lem. 3, p. 28)	-	-	Lin (lem. 3, p. 28)	Lin (lem. 3, p. 28)
$T_c^i(\sigma_\tau, \boldsymbol{\beta}_\tau)$	-	$\perp\!\!\!\perp$ (lem. 4, p. 29)	-	$\perp\!\!\!\perp$ (lem. 4, p. 29)	-
$V_\tau^*(\sigma_\tau)$	LC (app. D.1.3, p. 30)	$PWLCv$ (thm. 2.5, p. 6)	-	N/A	N/A
$W_\tau^{i,*}(\sigma_\tau, \beta_\tau^i)$	LC (from $Q_{\tau+1}^*$ LC)	-	-	$\neg Lin$ (from Q^* $\neg Lin$)	N/A
$\nu^2_{[\sigma_\tau^{c,1}, \beta^2_{\tau:}]}$	N/A	N/A	LC (lem. 7, p. 31)	N/A	-

B Background

B.1 Occupancy States

The following result shows that the occupancy state is (i) Markovian, *i.e.*, its value at τ only depends on its previous value $\sigma_{\tau-1}$, the system dynamics $P_{a^1,a^2}^{z^1,z^2}(s'|s)$, and the last behavioral decision rules $\beta_{\tau-1}^1$ and $\beta_{\tau-1}^2$, and (ii) sufficient to estimate the expected reward. Note that it holds for general-sum POSGs with any number of agents, and as many reward functions; similar results have already been established, *e.g.*, for Dec-POMDPs (*cf.* [Dibangoye et al., 2016, Theorem 1]).

Proposition 2.3. (originally stated on page 5) *$\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$, 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.*

Proof. Let us first derive a recursive way of computing $\sigma_{\boldsymbol{\beta}_{0:\tau}}(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau, \boldsymbol{z}_{\tau+1})$:

$$
\begin{aligned}
&\sigma_{\boldsymbol{\beta}_{0:\tau}}(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau, \boldsymbol{z}_{\tau+1}) \overset{\text{def}}{=} Pr(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau, \boldsymbol{z}_{\tau+1} \mid \boldsymbol{\beta}_{0:\tau}) \\
&= \sum_{s_\tau, s_{\tau+1}} Pr(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau, \boldsymbol{z}_{\tau+1}, s_\tau, s_{\tau+1} \mid \boldsymbol{\beta}_{0:\tau}) \\
&= \sum_{s_\tau, s_{\tau+1}} Pr(\boldsymbol{z}_{\tau+1}, s_{\tau+1} \mid \boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau, s_\tau, \boldsymbol{\beta}_{0:\tau}) Pr(\boldsymbol{a}_\tau \mid \boldsymbol{\theta}_\tau, s_\tau, \boldsymbol{\beta}_{0:\tau}) \\
&\qquad Pr(s_\tau \mid \boldsymbol{\theta}_\tau, \boldsymbol{\beta}_{0:\tau}) Pr(\boldsymbol{\theta}_\tau \mid \boldsymbol{\beta}_{0:\tau})
\end{aligned}
$$

Table 4: Various notations used in this work

$\neg i \stackrel{\text{def}}{=}$	i's opponent. Thus: $\neg 1 = 2$, and $\neg 2 = 1$.
	Histories and occupancy states
$\theta^i_\tau \stackrel{\text{def}}{=}$	$(a^i_0, z^i_1, \dots, a^i_{\tau-1}, z^i_\tau)$ ($\in \Theta^i = \cup_{t=0}^{H-1}\Theta^i_t$) is a length-$\tau$ *action-observation history* (AOH) for i.
$\boldsymbol{\theta}_\tau \stackrel{\text{def}}{=}$	$(\theta^1_\tau, \theta^2_\tau)$ ($\in \boldsymbol{\Theta} = \cup_{t=0}^{H-1}\boldsymbol{\Theta}_t$) is a *joint* AOH at τ.
$\sigma_\tau(\boldsymbol{\theta}_\tau) \stackrel{\text{def}}{=}$	*Occupancy state* (OS) σ_τ ($\in \mathcal{O}^\sigma = \cup_{t=0}^{H-1}\mathcal{O}^\sigma_t$, where $\mathcal{O}^\sigma_\tau \stackrel{\text{def}}{=} \Delta(\boldsymbol{\Theta}_\tau)$), *i.e.*, probability distribution over joint AOHs $\boldsymbol{\theta}_\tau$ (typically for some applied $\boldsymbol{\beta}_{0:\tau-1}$).
$\sigma^{m,i}_\tau(\theta^i_\tau) \stackrel{\text{def}}{=}$	*Marginal term* of σ_τ from player i's point of view ($\sigma^{m,i}_\tau \in \Delta(\Theta^i_\tau)$).
$\sigma^{c,i}_\tau(\theta^{\neg i}_\tau \mid \theta^i_\tau) \stackrel{\text{def}}{=}$	*Conditional term* of σ_τ from i's point of view ($\sigma^{c,i}_\tau : \Theta^i_\tau \mapsto \Delta(\Theta^{\neg i}_\tau)$).
$b(s \mid \boldsymbol{\theta}_\tau) \stackrel{\text{def}}{=}$	*Belief state*, *i.e.*, probability distribution over states given a joint AOH ($b(s \mid \boldsymbol{\theta}_\tau) : \mathcal{S} \times \boldsymbol{\Theta}_\tau \mapsto \mathbb{R}$). Can be computed by an HMM filtering process.
$o_\tau \stackrel{\text{def}}{=}$	*Full occupancy state* o_τ ($\in \Delta(\mathcal{S} \times \boldsymbol{\Theta}_\tau)$), *i.e.*, $Pr(s, \boldsymbol{\theta}_\tau)$ for the current $\boldsymbol{\beta}_{0:\tau-1}$, and thus verifies $\sigma_\tau(\boldsymbol{\theta}_\tau) = \sum_{s \in \mathcal{S}} o_\tau(s, \boldsymbol{\theta}_\tau)$. Is used in the implementation to simplify computations (*e.g.*, of r_t and $\sigma_{\tau+1}$ through b).
	Decision rules and strategies
$\pi^i_{0:\tau} \stackrel{\text{def}}{=}$	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 in $\mathcal{A}^i$. By default, $\pi^i \stackrel{\text{def}}{=} \pi^i_{0:H-1}$.
$\mu^i_{0:\tau} \stackrel{\text{def}}{=}$	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 it until $t = \tau$.
$\mu^i_{0:\tau'\mid\sigma_\tau\rangle} \stackrel{\text{def}}{=}$	$(\tau \leq \tau')$ is a mixed strategy *compatible* with some OS σ_τ, *i.e.*, that could induce this OS at τ (assuming an appropriate complementary $\mu^{\neg i}_{0:\tau'\mid\sigma_\tau\rangle}$).
$\beta^i_\tau \stackrel{\text{def}}{=}$	A *(behavioral) decision rule* (DR) at time τ for i is a mapping β^i_τ from private AOHs in Θ^i_τ to *distributions* over private actions. We note $\beta^i_\tau(\theta^i_\tau, a^i)$ the probability to pick a^i when facing θ^i_τ.
$\beta^i_{\tau:\tau'} \stackrel{\text{def}}{=}$	$(\beta^i_\tau, \dots, \beta^i_{\tau'})$ is a *behavioral strategy* for i from time step τ to τ' (included).
$rw^i(\theta^i_\tau, a^i_\tau) \stackrel{\text{def}}{=}$	$\prod_{t=0}^{\tau} \beta^i_{0:}(a^i_t \mid a^i_0, z^i_1, a^i_1, \dots, z^i_t)$ is the *realization weight* (RW) of sequence $a^i_0, z^i_1, a^i_1, \dots, a^i_\tau (= \theta^i_\tau, a^1_\tau)$ under strategy $\beta^i_{0:}$.
$rw^i(\phi^i_{\tau:\tau'} \mid \theta^i_\tau) \stackrel{\text{def}}{=}$	$\prod_{t=\tau}^{\tau'} \beta^i_{0:}(a^i_t \mid \theta^i_\tau, a^i_\tau, \dots, z^i_t)$ is the RW of a *suffix sequence* $\phi^i_{\tau:\tau'} = a^i_\tau, \dots, a^i_{\tau'}$ "conditioned" on a *prefix sequence*/AOH θ^i_τ.
$\boldsymbol{\pi}_{0:\tau} \stackrel{\text{def}}{=}$	is a *pure strategy profile*.
$\boldsymbol{\mu}_{0:\tau} \stackrel{\text{def}}{=}$	is a *mixed strategy profile*.
$\boldsymbol{\mu}_{0:\tau'\mid\sigma_\tau\rangle} \stackrel{\text{def}}{=}$	$(\tau \leq \tau')$ is a mixed strategy profile *compatible* with some OS σ_τ, *i.e.*, that could induce this OS at τ.
$\boldsymbol{\beta}_\tau \stackrel{\text{def}}{=}$	$\langle \beta^1_\tau, \beta^2_\tau \rangle$ ($\in \mathcal{B} = \cup_{t=0}^{H-1}\mathcal{B}_t$) is a *decision rule profile*.
$\boldsymbol{\beta}_{\tau:\tau'} \stackrel{\text{def}}{=}$	$\langle \beta^1_{\tau:\tau'}, \beta^2_{\tau:\tau'} \rangle$ is a *behavioral strategy profile*.

$$= \sum_{s_\tau, s_{\tau+1}} \underbrace{Pr(\boldsymbol{z}_{\tau+1}, s_{\tau+1} \mid \boldsymbol{a}_\tau, s_\tau)}_{=P^{\boldsymbol{z}_{\tau+1}}_{\boldsymbol{a}_\tau}(s_{\tau+1}\mid s_\tau)} \underbrace{Pr(\boldsymbol{a}_\tau \mid \boldsymbol{\theta}_\tau, \boldsymbol{\beta}_\tau)}_{=\boldsymbol{\beta}(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau)} \underbrace{Pr(s_\tau \mid \boldsymbol{\theta}_\tau, \boldsymbol{\beta}_{0:\tau})}_{=b(s_\tau\mid\boldsymbol{\theta}_\tau)} \underbrace{Pr(\boldsymbol{\theta}_\tau \mid \boldsymbol{\beta}_{0:\tau-1})}_{=\sigma_{\boldsymbol{\beta}_{0:\tau-1}}(\boldsymbol{\theta}_\tau)},$$

(where $b(s \mid \boldsymbol{\theta}_\tau)$ is the belief over states obtained by a usual HMM filtering process)

$$= \sum_{s_\tau, s_{\tau+1}} P^{\boldsymbol{z}_{\tau+1}}_{\boldsymbol{a}_\tau}(s_{\tau+1}\mid s_\tau)\boldsymbol{\beta}(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau) b(s_\tau \mid \boldsymbol{\theta}_\tau)\sigma_{\boldsymbol{\beta}_{0:\tau-1}}(\boldsymbol{\theta}_\tau).$$

$\sigma_{\boldsymbol{\beta}_{0:\tau}}$ can thus be computed from $\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$ and $\boldsymbol{\beta}_\tau$ without explicitly using $\boldsymbol{\beta}_{0:\tau-1}$ or earlier occupancy states.

Then, let us compute the expected reward at τ given $\boldsymbol{\beta}_{0:\tau}$:

$$\begin{aligned} & E[r(S_\tau, A^1_\tau, A^2_\tau) \mid \boldsymbol{\beta}_{0:\tau}] \\ &= \sum_{s_\tau, \boldsymbol{a}_\tau} r(s_\tau, \boldsymbol{a}_\tau) Pr(s_\tau, \boldsymbol{a}_\tau \mid \boldsymbol{\beta}_{0:\tau}) \\ &= \sum_{s_\tau, \boldsymbol{a}_\tau} \sum_{\boldsymbol{\theta}_\tau} r(s_\tau, \boldsymbol{a}_\tau) Pr(s_\tau, \boldsymbol{a}_\tau, \boldsymbol{\theta}_\tau \mid \boldsymbol{\beta}_{0:\tau}) \end{aligned}$$

Rewards and value functions

$r_{\max} \stackrel{\text{def}}{=} \max_{s,\boldsymbol{a}} r(s,\boldsymbol{a})$ Maximum possible reward.

$r_{\min} \stackrel{\text{def}}{=} \min_{s,\boldsymbol{a}} r(s,\boldsymbol{a})$ Minimum possible reward.

$V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:}) \stackrel{\text{def}}{=} E[\sum_{t=\tau}^{H-1} \gamma^t R_t \mid \sigma_\tau, \boldsymbol{\beta}_{\tau:}]$, *Value* of $\boldsymbol{\beta}_{\tau:H-1}$ in OS σ_τ.
where R_t is the random var. for the reward at t.

$V^*_\tau(\sigma_\tau) \stackrel{\text{def}}{=} \max_{\beta^1_{\tau:}} \min_{\beta^2_{\tau:}} V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:})$ *Optimal value function*

$Q^*_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau) \stackrel{\text{def}}{=} r(\sigma_\tau, \boldsymbol{\beta}_\tau) + \gamma V^*_{\tau+1}(T(\sigma_\tau, \boldsymbol{\beta}_\tau))$ *Opt. (joint) action-value fct.*

$W^{i,*}_\tau(\sigma_\tau, \beta^i_\tau) \stackrel{\text{def}}{=} opt_{\beta^{-i}_\tau} Q^*_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau)$, *Opt. (individual) action-value fct.*
where $opt = \max$ if $i = 1$, $\min$ otherwise.

$\nu^2_{[\sigma^{c,1}_\tau, \beta^2_{\tau:}]} \stackrel{\text{def}}{=}$ Vector of values (one component per AOH θ^1_τ) for 1's best response to $\beta^2_{\tau:}$ assuming $\sigma^{c,1}_\tau$. This solution of a POMDP allows computing V^*_τ (see Theorem 2.5).

Approximations

$\overline{V}_\tau(\sigma_\tau) \stackrel{\text{def}}{=}$ Upper bound approximation of $V^*_\tau(\sigma_\tau)$; relies on data set $\overline{\mathcal{I}}_{\tau-1}$.

$\underline{V}_\tau(\sigma_\tau) \stackrel{\text{def}}{=}$ Lower bound approximation of $V^*_\tau(\sigma_\tau)$; relies on data set $\underline{\mathcal{I}}_{\tau-1}$.

$\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \stackrel{\text{def}}{=}$ Upper bound approximation of $W^{*,1}_\tau(\sigma_\tau, \beta^1_\tau)$; relies on data set $\overline{\mathcal{I}}_\tau$.

$\underline{W}_\tau(\sigma_\tau, \beta^2_\tau) \stackrel{\text{def}}{=}$ Lower bound approximation of $W^{*,2}_\tau(\sigma_\tau, \beta^2_\tau)$; relies on data set $\underline{\mathcal{I}}_\tau$.

$\overline{\nu}^2_\tau \stackrel{\text{def}}{=}$ Vector (with one component per AOH θ^1_τ) used in $\overline{V}_\tau$ and $\overline{W}_{\tau-1}$ (if $\tau \geq 1$).

Miscellaneous

$w_\tau \stackrel{\text{def}}{=}$ Denotes a triplet $\langle \sigma^{c,1}_{\tau-1}, \beta^1_{\tau-1}, \overline{\nu}^2_\tau, \rangle \in \overline{\mathcal{I}}_\tau$ (or a triplet in $\underline{\mathcal{I}}_\tau$).

$\psi^2_\tau \stackrel{\text{def}}{=}$ Distribution over triplets $w_{\tau+1} \in \overline{\mathcal{I}}_{\tau+1}$ (inducing a recursively defined strategy from τ to $H-1$). Often denotes the strategy it induces.

$x^\top \stackrel{\text{def}}{=}$ The transpose of a (usually column) vector x of $\mathbb{R}^n$.

$c[y] \stackrel{\text{def}}{=}$ Denotes field c of object/tuple y.

$Supp(d) \stackrel{\text{def}}{=}$ Support of distribution d, *i.e.*, set of its non-zero probability elements.

$$
\begin{aligned}
&= \sum_{s_\tau, \boldsymbol{a}_\tau} \sum_{\boldsymbol{\theta}_\tau} r(s_\tau, \boldsymbol{a}_\tau) Pr(s_\tau, \boldsymbol{a}_\tau \mid \boldsymbol{\theta}_\tau, \boldsymbol{\beta}_{0:\tau}) Pr(\boldsymbol{\theta}_\tau \mid \boldsymbol{\beta}_{0:\tau}) \\
&= \sum_{s_\tau, \boldsymbol{a}_\tau} \sum_{\boldsymbol{\theta}_\tau} r(s_\tau, \boldsymbol{a}_\tau) \underbrace{Pr(\boldsymbol{a}_\tau \mid \boldsymbol{\theta}_\tau, \boldsymbol{\beta}_{0:\tau})}_{\boldsymbol{\beta}_\tau(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau)} \underbrace{Pr(s_\tau \mid \boldsymbol{\theta}_\tau, \boldsymbol{\beta}_{0:\tau})}_{b(s_\tau \mid \boldsymbol{\theta}_\tau)} \underbrace{Pr(\boldsymbol{\theta}_\tau \mid \boldsymbol{\beta}_{0:\tau})}_{\sigma_{\boldsymbol{\beta}_{0:\tau-1}}(\boldsymbol{\theta}_\tau)} \\
&= \sum_{s_\tau, \boldsymbol{a}_\tau} \sum_{\boldsymbol{\theta}_\tau} r(s_\tau, \boldsymbol{a}_\tau) \boldsymbol{\beta}_\tau(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau) b(s_\tau \mid \boldsymbol{\theta}_\tau) \sigma_{\boldsymbol{\beta}_{0:\tau-1}}(\boldsymbol{\theta}_\tau).
\end{aligned}
$$

The expected reward at τ can thus be computed from $\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$ and $\boldsymbol{\beta}_\tau$ without explicitly using $\boldsymbol{\beta}_{0:\tau-1}$ or earlier occupancy states. □

C Occupancy Markov Games: Definition and Preliminary Properties

C.1 Properties of V^*

Before proving the postulates implicitly used by [Wiggers et al., 2016a], we need to show that we can reason with mixed strategies in subgames as is usually done on full games.

C.1.1 V^* is not linear in β (behavioral strategies)

Let us consider the following (finite-horizon, deterministic) Non-Observable MDP:

$\mathcal{S} \stackrel{\text{def}}{=} \{-2, -1, 0, +1, +2\}, \quad b_0(0) = 1,$ (always start in $s = 0$)

$\mathcal{A} \stackrel{\text{def}}{=} \{-1, +1\},$ (moves = add or subtract 1)

$T(s,a) \stackrel{\text{def}}{=} \min\{+2, \max\{-2, s+a\}\},$ (dép. de $+1$ ou -1 dans $\mathcal{S}$)

$\mathcal{Z} \stackrel{\text{def}}{=} \{none\}, \quad O(none) \stackrel{\text{def}}{=} 1,$ (no observation)

$$r(s) \stackrel{\text{def}}{=} \begin{cases} +1 & \text{si } s \in \{-2, +2\} \\ 0 & \text{sinon,} \end{cases} \qquad (|s| = 2 : \text{victoire !})$$
$$\gamma \stackrel{\text{def}}{=} 1, \quad H \stackrel{\text{def}}{=} 2.$$

Let us then consider two particular behavioral strategies:

$$\forall \theta,\ \beta^+(A = +1|\theta) = 1 \qquad \text{(always } +1\text{), and}$$
$$\forall \theta,\ \beta^-(A = -1|\theta) = 1 \qquad \text{(always } -1\text{).}$$

These two strategies are optimal, with an expected return of $+1$, because, at $t = H = 2$, β^+ reaches $+2$ w.p. 1, and β^- reaches -2 w.p. 1:

$$V(\beta^+) = V(\beta^-) = +1.$$

Let us now consider their linear combination $\beta^\pm \stackrel{\text{def}}{=} \frac{1}{2}\beta^+ + \frac{1}{2}\beta^-$:

$$\forall \theta,\ \beta^\pm(A = -1|\theta) = 0.5,$$
$$\forall \theta,\ \beta^\pm(A = +1|\theta) = 0.5.$$

Here, the probability to reach $s = -2$ or $s = +2$ at the last time step is much lower, and gives the value of that strategy:

$$\begin{aligned} V(\beta^\pm) &= Pr(s_2 = +2|\beta^\pm) + Pr(s_2 = +2|\beta^\pm) \\ &= Pr(a_0 = +1|\beta^\pm) \cdot Pr(a_1 = +1|\beta^\pm) \\ &+ Pr(a_0 = -1|\beta^\pm) \cdot Pr(a_1 = -1|\beta^\pm) \\ &= \underbrace{0.5 \cdot 0.5}_{0.25} + \underbrace{0.5 \cdot 0.5}_{0.25} = 0.5. \end{aligned}$$

C.1.2 Back to Mixed Strategies

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

First, for a given σ_τ and $\tau \leq \tau'$, let $\boldsymbol{\mu}_{0:\tau'-1|\sigma_\tau\rangle}$ denote a mixed strategy profile that is defined over $0 : \tau' - 1$, and induces (/is *compatible* with) σ_τ at time τ. Then, to complete a given mixed *prefix* strategy $\boldsymbol{\mu}_{0:\tau'-1|\sigma_\tau\rangle}$ (here $\tau = \tau'$), the solver 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 $\mu^i_{0:\tau-1|\sigma_\tau\rangle}$ (belonging to a set denoted $\Pi^i_{0:\tau-1|\sigma_\tau\rangle}$) 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 $M^i_{0:H-1|\sigma_\tau\rangle}$ be the set of mixed *full* strategies $\mu^i_{0:H-1|\sigma_\tau\rangle}$ obtained by considering the distributions over $\bigcup_{\pi^i_{0:\tau-1} \in \Pi^i_{0:\tau-1|\sigma_\tau\rangle}} \Pi^i_{0:H-1}(\pi^i_{0:\tau-1})$ that verify, $\forall \pi^i_{0:\tau-1}$,

$$\sum_{\substack{\pi^i_{0:H-1} \in \\ \Pi^i_{0:H-1}(\pi^i_{0:\tau-1})}} \mu^i_{0:H-1|\sigma_\tau\rangle}(\pi^i_{0:H-1}) = \mu^i_{0:\tau-1|\sigma_\tau\rangle}(\pi^i_{0:\tau-1}). \tag{8}$$

This is the set of mixed strategies compatible with σ_τ.

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

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

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

$$+ (1-\alpha)\left[\sum_{\pi^i_{0:H-1}\in\Pi^i_{0:H-1}(\pi^i_{0:\tau-1})} \nu^i_{0:H-1|\sigma_\tau\rangle}(\pi^i_{0:H-1})\right]$$

(because both mixed strategies are compatible with σ_τ (eq. 8, p. 26):)

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

Eq. 8 thus also applies to $\alpha\cdot\mu^i_{0:H-1|\sigma_\tau\rangle} + (1-\alpha)\cdot\nu^i_{0:H-1|\sigma_\tau\rangle}$, proving that it belongs to $M^i_{0:H-1|\sigma_\tau\rangle}$ 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 $\theta^i_{0:\tau-1}$. □

While only future rewards are relevant when making a decision at τ, reasoning with mixed strategies defined from $t=0$ will be convenient because $V_\tau(\sigma_\tau,\cdot,\cdot)$ is linear in $\mu^i_{0:H-1|\sigma_\tau\rangle}$, 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 $M^i_{0:H-1|\sigma_\tau\rangle}$, set which we now note $M^i_{|\sigma_\tau\rangle}$. Also, since we shall work with local game $Q^*_\tau(\sigma_\tau,\boldsymbol{\beta}_\tau)$, let us define: $M^i_{|\sigma_\tau,\beta^j_\tau\rangle}$ the set of i's mixed strategies compatible with occupancy states reachable given σ_τ and β^j_τ (with either $j=i$ or $j=-i$). Then, $M^i_{|T(\sigma_\tau,\boldsymbol{\beta}_\tau)\rangle}\subseteq M^i_{|\sigma_\tau,\beta^j_\tau\rangle}\subseteq M^i_{|\sigma_\tau\rangle}$ (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 σ_τ:

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

C.1.3 Von Neumann's Minimax Theorem for Subgames and a Bellman Optimality Equation

Using the previous results, one can show that von Neumann's minimax theorem applies in any subgame, allowing to swap operators $\max$ and $\min$.

Theorem 3.1. (originally stated on page 7) *The subgame defined in Eq.* (3) *admits a unique NEV*

$$V^*_\tau(\sigma_\tau) \overset{\text{def}}{=} \max_{\beta^1_{\tau:H-1}}\min_{\beta^2_{\tau:H-1}} V_\tau(\sigma_\tau,\beta^1_{\tau:H-1},\beta^2_{\tau:H-1}). \tag{4}$$

Proof. For any occupancy state σ_τ,

$$\begin{aligned}\max_{\beta^1_{\tau:}}\min_{\beta^2_{\tau:}} V(\sigma_\tau,\boldsymbol{\beta}_{\tau:}) &= \max_{\mu^1_{\tau:}}\min_{\mu^2_{\tau:}} V(\sigma_\tau,\boldsymbol{\mu}_{\tau:}) && \text{(Kuhn's theorem (generalized))} && (9)\\ &= \min_{\mu^2_{\tau:}}\max_{\mu^1_{\tau:}} V(\sigma_\tau,\boldsymbol{\mu}_{\tau:}) && \text{(von Neumann's theorem)} && (10)\\ &= \min_{\beta^2_{\tau:}}\max_{\beta^1_{\tau:}} V(\sigma_\tau,\boldsymbol{\beta}_{\tau:}) && \text{(again Kuhn's theorem (generalized))} && (11)\end{aligned}$$

□

One can also show that a Bellman optimality equation allows relating optimal values in subgames at τ and $\tau+1$, leading to a recursive expression of V^*_0.

Theorem 3.2. (originally stated on page 7) $V^*_\tau(\sigma_\tau)$ *satisfies the following functional equation:*

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

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

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

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

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

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

(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:)

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

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

$$= \max_{\beta^1_\tau} \min_{\mu^2 \in M^2_{|\sigma_\tau, \beta^1_\tau\rangle}} \max_{\mu^1 \in M^1_{|\sigma_\tau, \beta^1_\tau, \beta^2_\tau(\mu^2)\rangle}}$$
$$\left[r(\sigma_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)) + \gamma V_{\tau+1}(T(\sigma_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)), \mu^1, \mu^2)\right] \tag{12}$$

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

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

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

$$= \max_{\beta^1_\tau} \max_{\mu^1 \in M^1_{|\sigma_\tau, \beta^1_\tau\rangle}} \min_{\mu^2 \in M^2_{|\sigma_\tau\rangle}} \left[r(\sigma_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)) + \gamma V_{\tau+1}(T(\sigma_\tau, \beta^1_\tau, \beta^2_\tau(\mu_2)), \mu^1, \mu^2)\right] \tag{13}$$

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

$$= \max_{\mu^1 \in M^1_{|\sigma_\tau\rangle}} \min_{\mu^2_\tau \in M^2_{\tau|\sigma_\tau}} \left[r(\sigma_\tau, \beta^1_\tau(\mu_1), \beta^2_\tau(\mu_2)) + \gamma V_{\tau+1}(T(\sigma_\tau, \beta^1_\tau(\mu^1), \beta^2_\tau(\mu_2)), \mu^1, \mu^2)\right] \tag{14}$$

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

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

More precisely, line 12 (and, similarly, line 14) is obtained by observing that

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

Also, line 13 results from the observation that, while $M^1_{|\sigma_\tau, \beta^1_\tau\rangle}$ and $M^2_{|\sigma_\tau\rangle}$ allow to actually make decisions over different time intervals, we are here minimizing over μ^2 while maximizing over μ^1 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. □

D Solving zs-OMGs

D.1 Preliminary Properties

D.1.1 Properties of T^1_m and T^1_c

The first two lemmas below present properties of T^1_m and T^1_c that will be useful afterwards.

Lemma 3. *$T^1_m(\sigma_\tau, \boldsymbol{\beta}_\tau)$ is linear in σ_τ, β^1_τ, and β^2_τ.*

Proof.

$$T^1_m(\sigma_\tau, \boldsymbol{\beta}_\tau)(\theta^1_\tau, a^1, z^1) \tag{15}$$

$$
\begin{aligned}
&= \sum_{\theta^2_\tau, a^2, z^2} T(\sigma_\tau, \boldsymbol{\beta}_\tau)((\theta^1_\tau, a^1, z^1), (\theta^2_\tau, a^2, z^2)) \qquad \text{(from Eq. (1))} \\
&= \sum_{s', \theta^2_\tau, a^2, z^2} \beta^1_\tau(\theta^1_\tau, a^1)\beta^2_\tau(\theta^2_\tau, a^2) \sum_s P^{z^1,z^2}_{a^1,a^2}(s'|s) b(s|\theta^1_\tau, \theta^2_\tau)\sigma_\tau(\theta^1_\tau, \theta^2_\tau) \\
&= \beta^1_\tau(\theta^1_\tau, a^1) \sum_{\theta^2_\tau, a^2} \beta^2_\tau(\theta^2_\tau, a^2) \sum_{s, s', z^2} P^{z^1,z^2}_{a^1,a^2}(s'|s) b(s|\theta^1_\tau, \theta^2_\tau)\sigma_\tau(\theta^1_\tau, \theta^2_\tau). \qquad (16)
\end{aligned}
$$

□

Lemma 4. *$T^1_c(\sigma_\tau, \boldsymbol{\beta}_\tau)$ is independent of β^1_τ and $\sigma^{m,1}_\tau$.*

Proof. See Wiggers [2015], Lemma 4.2.3. □

D.1.2 Linearity and Lipschitz-continuity of $T(\sigma_\tau, \beta^1_\tau, \beta^2_\tau)$

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

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

Proof. Let σ be an occupancy state at time τ and $\boldsymbol{\beta}_\tau$ be a decision rule. Then, as seen in the proof of Proposition 2.3, the next occupancy state $\sigma' = T(\sigma, \boldsymbol{\beta}_\tau)$ satisfies, for any s' and $(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z})$:

$$
\begin{aligned}
\sigma'(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z}) &\overset{\text{def}}{=} Pr(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{z}|\sigma, \beta^1_\tau, \beta^2_\tau) \\
&= \beta^1_\tau(\theta^1, a^1)\beta^2_\tau(\theta^2, a^2) \left[\sum_{s', s \in \mathcal{S}} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) b(s|\boldsymbol{\theta}) \right] \sigma(\boldsymbol{\theta}).
\end{aligned}
$$

$b(s|\boldsymbol{\theta})$ depending only on the model (transition function and initial belief), the next occupancy state σ' thus evolves linearly w.r.t. (i) *private* decision rules β^1_τ and β^2_τ, and (ii) the occupancy state σ.

The 1-Lipschitz-continuity holds because each component of vector σ_τ is distributed over multiple components of σ'. 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 σ into $\sigma' \overset{\text{def}}{=} T(\sigma_\tau, \boldsymbol{\beta}_\tau)$). Then,

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

□

D.1.3 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-bound approximations.

Lemma 6. *At depth τ, $V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:})$ is linear w.r.t. σ_τ and $\boldsymbol{\beta}_{\tau:}$.*

Note: This result in fact applies to any reward function of a general-sum POSG with any number of agents (here N), *e.g.*, to a Dec-POMDP. The following proof handles the general case (with $\boldsymbol{\beta}_\tau \overset{\text{def}}{=} \langle \beta^1_\tau, \ldots, \beta^N_\tau \rangle$, and $\boldsymbol{\beta}_\tau(\boldsymbol{a}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \beta^i_\tau(a^i, \theta^1)$).

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

$$\begin{aligned}
V_{H-1}(\sigma_{H-1}, \boldsymbol{\beta}_{H-1:}) &= r(\sigma_{H-1}, \boldsymbol{\beta}_{H-1}) \\
&= \sum_{s,\boldsymbol{a}} \left(\sum_{\boldsymbol{\theta}} Pr(s, \boldsymbol{a}|\boldsymbol{\theta}) \sigma_{H-1}(\boldsymbol{\theta}) \right) r(s, \boldsymbol{a}) \\
&= \sum_{s,\boldsymbol{a}} \left(\sum_{\boldsymbol{\theta}} b(s|\boldsymbol{\theta}) \boldsymbol{\beta}_\tau(\boldsymbol{a}|\boldsymbol{\theta}) \sigma_{H-1}(\boldsymbol{\theta}) \right) r(s, \boldsymbol{a}) \\
&= \sum_{s,\boldsymbol{\theta}} b(s|\boldsymbol{\theta}) \sigma_{H-1}(\boldsymbol{\theta}) \left(\sum_{\boldsymbol{a}} \boldsymbol{\beta}_\tau(\boldsymbol{a}|\boldsymbol{\theta}) r(s, \boldsymbol{a}) \right).
\end{aligned}$$

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

$$\begin{aligned}
V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:}) &= \sum_{s,\boldsymbol{a}} \Big(\sum_{\boldsymbol{\theta}} b(s|\boldsymbol{\theta}) \boldsymbol{\beta}_\tau(\boldsymbol{a}|\boldsymbol{\theta}) \sigma_\tau(\boldsymbol{\theta}) \Big) r(s, \boldsymbol{a}) + \gamma V_{\tau+1} \left(T(\sigma_\tau, \boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:} \right) \\
&= \sum_{s,\boldsymbol{\theta}} b(s|\boldsymbol{\theta}) \sigma_\tau(\boldsymbol{\theta}) \Big(\sum_{\boldsymbol{a}} \boldsymbol{\beta}_\tau(\boldsymbol{a}|\boldsymbol{\theta}) r(s, \boldsymbol{a}) \Big) + \gamma V_{\tau+1} \left(T(\sigma_\tau, \boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:} \right).
\end{aligned}$$

As

- $T(\sigma_\tau, \boldsymbol{\beta}_\tau)$ is linear in σ_τ (Lemma 5) and
- $V_{\tau+1}(\sigma_{\tau+1}, \boldsymbol{\beta}_{\tau+1:})$ is linear in $\sigma_{\tau+1}$ (induction hypothesis),

their composition, $V_{\tau+1}(T(\sigma_\tau, \boldsymbol{\beta}_\tau), \boldsymbol{\beta}_{\tau+1:})$, is also linear in σ_τ, and so is $V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:})$. Similarly, $V_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau)$ is linear in $\boldsymbol{\beta}_\tau$ for any σ_τ. □

Theorem 3.3. (originally stated on page 7) *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* $\lambda_\tau = \frac{1}{2} h_\tau \left(r_{\max} - r_{\min} \right)$.

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

$$\begin{aligned}
V_\tau^{\max} &\overset{\text{def}}{=} h_\tau r_{\max}, \quad \text{where } r_{\max} \overset{\text{def}}{=} \max_{s,\boldsymbol{a}} r(s, \boldsymbol{a}), \text{ and} \\
V_\tau^{\min} &\overset{\text{def}}{=} h_\tau r_{\min}, \quad \text{where } r_{\min} \overset{\text{def}}{=} \min_{s,\boldsymbol{a}} r(s, \boldsymbol{a}).
\end{aligned}$$

Thus, $V_{\boldsymbol{\beta}_{\tau:}}$ being a linear function defined over a probability simplex ($\mathcal{O}_\tau^\sigma$) (*cf.* Appendix D.1.3) and bounded by $[V_\tau^{\min}, V_\tau^{\max}]$, we can apply Horák's PhD thesis' Lemma 3.5 (p. 33) 2019 to establish that it is also λ_τ-LC, *i.e.*,

$$\begin{gathered}
|V_{\boldsymbol{\beta}_{\tau:}}(\sigma) - V_{\boldsymbol{\beta}_{\tau:}}(\sigma')| \leq \lambda_\tau \|\sigma - \sigma'\|_1 \quad (\forall \sigma, \sigma'), \\
\text{with } \lambda_\tau = \frac{V_\tau^{\max} - V_\tau^{\min}}{2}.
\end{gathered}$$

Considering now optimal solutions, this means that, at depth τ and for any $(\sigma, \sigma') \in \mathcal{O}_\tau^\sigma$:

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

Symmetrically, $V_\tau^*(\sigma) - V_\tau^*(\sigma') \geq -\lambda_\tau \|\sigma - \sigma'\|_1$, hence the expected result:

$$|V_\tau^*(\sigma) - V_\tau^*(\sigma')| \leq \lambda_\tau \|\sigma - \sigma'\|_1.$$

□

As it will be used later, let us also present the following lemma.

Lemma 7. *Let us consider $\tau \in \{0\,..\,H-1\}$, θ^1_τ, and ψ^2_τ. Then $\nu^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]}(\theta^1_\tau)$ is λ_τ-LC in $\sigma^{c,1}_\tau(\cdot|\theta^1_\tau)$.*

Equivalently, we will also write that $\nu^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]}$ is λ_τ-LC in $\sigma^{c,1}_\tau$ in vector-wise *1-norm,* i.e.*:*

$$\overrightarrow{|\nu^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]} - \nu^2_{[\tilde{\sigma}^{c,1}_\tau, \psi^2_\tau]}|_1} \overset{\rightarrow}{\leq} \lambda_\tau \overrightarrow{\|\sigma^{c,1}_\tau - \tilde{\sigma}^{c,1}_\tau\|_1},$$

where (i) the absolute value of a vector is obtained by taking the absolute value of each component; and (ii) the vector-wise 1-norm of a matrix is a vector made of the 1-norm of each of its component vectors.

Proof. For any θ^1_τ, $\sigma^{c,1}_\tau$ and ψ^2_τ induce a POMDP for Player 1 from τ on, where (i) the state at any $t \in \{\tau\,..\,H-1\}$ corresponds to a pair $\langle s, \theta^2_t\rangle$, and (ii) the initial belief is derived from $\sigma^{c,1}_\tau(\cdot|\theta^1_t)$. The belief state at t thus gives:

$$b_{\theta^1_t}(s, \theta^2_t) \overset{\text{def}}{=} Pr(s, \theta^2_t|\theta^1_t) = \underbrace{Pr(s|\theta^2_t, \theta^1_t)}_{b^{\text{HMM}}_{\theta^2_t, \theta^1_t}(s)} \cdot \underbrace{Pr(\theta^2_t|\theta^1_t)}_{\sigma^{c,1}_t(\theta^2_t|\theta^1_t)}.$$

So,

- the value function of any behavioral strategy $\beta^1_{\tau:}$ is linear at t in $b_{\theta^1_t}$, thus (in particular) in $\sigma^{c,1}_t(\cdot|\theta^1_t)$; and
- the optimal value function is LC at t also in $b_{\theta^1_t}$ (with the same depth-dependent upper-bounding Lipschitz constant λ_t as in the proof of Theorem 3.3),[5] thus (in particular) in $\sigma^{c,1}_t(\cdot|\theta^1_t)$.

Using $t = \tau$, the optimal value function is $\nu^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]}(\theta^1_\tau)$, which is thus λ_τ-LC in $\sigma^{c,1}_\tau(\cdot|\theta^1_\tau)$. □

D.2 Bounding Approximations of V^*, $W^{1,*}$ and $W^{2,*}$

D.2.1 $\overline{V}_\tau$ and $\underline{V}_\tau$

To find a form that could be appropriate for an upper bound approximation of V^*_τ, let us consider an OS σ_τ and a single tuple $\langle \tilde{\sigma}_\tau, \nu^2_{[\tilde{\sigma}^{c,1}_\tau, \beta^2_{\tau:}]}\rangle$, and define $\zeta_\tau \overset{\text{def}}{=} \sigma^{m,1}_\tau \tilde{\sigma}^{c,1}_\tau$. Then,

$$\begin{aligned}
V^*(\sigma_\tau) &\leq V^*(\zeta_\tau) + \lambda_\tau \|\sigma_\tau - \zeta_\tau\|_1 && \text{(LC, } \textit{cf.} \text{ Theorem 3.3)}\\
&= V^*(\sigma^{m,1}_\tau \tilde{\sigma}^{c,1}_\tau) + \lambda_\tau \|\sigma_\tau - \zeta_\tau\|_1 \\
&\leq \sigma^{m,1}_\tau \cdot \nu^2_{[\tilde{\sigma}^{c,1}_\tau, \beta^2_{\tau:}]} + \lambda_\tau \|\sigma_\tau - \sigma^{m,1}_\tau \tilde{\sigma}^{c,1}_\tau\|_1. && \text{(Cvx, } \textit{cf.} \text{ Theorem 2.5)}
\end{aligned}$$

Notes:

- $\tilde{\sigma}^{m,1}_\tau$ does not appear in the resulting upper bound, thus will not need to be specified.
- For $\tau = H-1$, $\nu^2_{[\tilde{\sigma}^{c,1}_\tau, \beta^2_{\tau:}]}$ is a simple function of r, $\tilde{\sigma}^{c,1}_\tau$, $\beta^2_{\tau:}$, and the dynamics of the system, as described in Eq. (9) of Wiggers et al. [2016a].

From this, we can deduce the following appropriate forms of upper and (symmetrically) lower bound function approximations for V^*_τ:

$$\begin{aligned}
\overline{V}_\tau(\sigma_\tau) &= \min_{\langle \tilde{\sigma}^{c,1}_\tau, \langle \overline{\nu}^2_\tau, \psi^2_{\tau:}\rangle\rangle \in \overline{\mathcal{J}}_\tau} \left[\sigma^{m,1}_\tau \cdot \overline{\nu}^2_\tau + \lambda_\tau \|\sigma_\tau - \sigma^{m,1}_\tau \tilde{\sigma}^{c,1}_\tau\|_1\right], \text{ and}\\
\underline{V}_\tau(\sigma_\tau) &= \max_{\langle \tilde{\sigma}^{c,2}_\tau, \langle \underline{\nu}^1_\tau, \psi^1_{\tau:}\rangle\rangle \in \underline{\mathcal{J}}_\tau} \left[\sigma^{m,2}_\tau \cdot \underline{\nu}^1_\tau - \lambda_\tau \|\sigma_\tau - \sigma^{m,2}_\tau \tilde{\sigma}^{c,2}_\tau\|_1\right],
\end{aligned}$$

which are respectively concave in $\sigma^{m,1}_\tau$ and convex in $\sigma^{m,2}_\tau$, and which both exploit the Lipschitz continuity.

D.2.2 $\overline{W}_\tau$ and $\underline{W}_\tau$

Note: We discuss all depths from 0 to $H-1$, even though we do not need these approximations at $\tau = H-1$.

Let us first see how concavity-convexity properties affect $W^{*,1}_\tau$.

Lemma 8. *Considering that vectors $\nu^2_{[\sigma^{c,1}_H, \beta^2_{H:}]}$ are null vectors, we have, for all $\tau \in \{0\,..\,H-1\}$:*

$$\begin{aligned}
W^{1,*}_\tau(\sigma_\tau, \beta^1_\tau) = \min_{\beta^2_\tau, \langle \beta^2_{\tau+1:}, \nu^2_{[T^1_c(\sigma_\tau, \beta^2_\tau), \beta^2_{\tau+1:}]}\rangle} \beta^1_\tau \cdot \Big[& r(\sigma_\tau, \cdot, \beta^2_\tau)\\
& + \gamma T^1_m(\sigma_\tau, \cdot, \beta^2_\tau) \cdot \nu^2_{[T^1_c(\sigma_\tau, \beta^2_\tau), \beta^2_{\tau+1:}]}\Big].
\end{aligned}$$

[5] The proof process is similar. The only difference lies in the space at hand, but without any impact on the resulting formulas.

Proof. Considering that vectors $\nu^2_{[\sigma_H^{c,1},\beta^2_{H:}]}$ are null vectors, we have, for all $\tau \in \{0\,..\,H-1\}$:

$$W_\tau^{1,*}(\sigma_\tau,\beta_\tau^1) = \min_{\beta_\tau^2} Q_\tau^*(\sigma_\tau,\beta_\tau^1,\beta_\tau^2) = \min_{\beta_\tau^2}\left[r(\sigma_\tau,\boldsymbol{\beta}_\tau) + \gamma V_{\tau+1}^*(T(\sigma_\tau,\boldsymbol{\beta}_\tau))\right]$$

(Line below exploits Theorem 2.5 (p. 6) and T_c^1's independence from β_τ^1 (Lemma 4).)

$$= \min_{\beta_\tau^2}\left[r(\sigma_\tau,\boldsymbol{\beta}_\tau) + \gamma \min_{\langle\beta^2_{\tau+1:},\nu^2_{[T_c^1(\sigma_\tau,\beta_\tau^2),\beta^2_{\tau+1:}]}\rangle}\left[T_m^1(\sigma_\tau,\boldsymbol{\beta}_\tau)\cdot\nu^2_{[T_c^1(\sigma_\tau,\beta_\tau^2),\beta^2_{\tau+1:}]}\right]\right]$$
$$= \min_{\beta_\tau^2,\langle\beta^2_{\tau+1:},\nu^2_{[T_c^1(\sigma_\tau,\beta_\tau^2),\beta^2_{\tau+1:}]}\rangle}\left[r(\sigma_\tau,\boldsymbol{\beta}_\tau) + \gamma T_m^1(\sigma_\tau,\boldsymbol{\beta}_\tau)\cdot\nu^2_{[T_c^1(\sigma_\tau,\beta_\tau^2),\beta^2_{\tau+1:}]}\right]$$

(Line below exploits r and T_m^1's linearity in β_τ^1 (Lemma 3).)

$$= \min_{\beta_\tau^2,\langle\beta^2_{\tau+1:},\nu^2_{[T_c^1(\sigma_\tau,\beta_\tau^2),\beta^2_{\tau+1:}]}\rangle} {\beta_\tau^1}^\top\cdot\Big[r(\sigma_\tau,\cdot,\beta_\tau^2) + \gamma T_m^1(\sigma_\tau,\cdot,\beta_\tau^2)\cdot\nu^2_{[T_c^1(\sigma_\tau,\beta_\tau^2),\beta^2_{\tau+1:}]}\Big].$$

□

Note that, since $V_H^* = 0$, $\tau = H-1$ is a particular case which can be simply re-written:

$$W_\tau^{1,*}(\sigma_\tau,\beta_\tau^1) = \min_{\beta_\tau^2} {\beta_\tau^1}^\top\cdot r(\sigma_\tau,\cdot,\beta_\tau^2).$$

To find a form that could be appropriate for an upper bound approximation of $W_\tau^{*,1}$, let us now consider an OS σ_τ and a single tuple $\langle\tilde\sigma_\tau,\tilde\beta_\tau^2,\nu^2_{[T_c^1(\tilde\sigma_\tau,\tilde\beta_\tau^2),\tilde\beta^2_{\tau+1:}]}\rangle$. Then,

$$W_\tau^{1,*}(\sigma_\tau,\beta_\tau^1) = \min_{\beta_\tau^2}\left[r(\sigma_\tau,\boldsymbol{\beta}_\tau) + \gamma V_{\tau+1}^*(T(\sigma_\tau,\boldsymbol{\beta}_\tau))\right]$$
$$\leq r(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2) + \gamma V_{\tau+1}^{BR,1}(T(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2)|\tilde\beta^2_{\tau+1:})$$

(Use $\tilde\beta_\tau^2$ & $\tilde\beta^2_{\tau+1:}$ instead of mins)

(where $V_{\tau+1}^{BR,1}(T(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2)|\tilde\beta^2_{\tau+1:})$ is the value of 1's best response to $\tilde\beta^2_{\tau+1:}$ if in $T(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2)$)

$$= r(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2) + \gamma T_m^1(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2)\cdot\underbrace{\nu^2_{[T_c^1(\sigma_\tau^{c,1},\tilde\beta_\tau^2),\tilde\beta^2_{\tau+1:}]}}$$

(Lem. 3 of Wiggers et al. [2016a])

$$\leq r(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2) + \gamma T_m^1(\sigma_\tau,\beta_\tau^1,\tilde\beta_\tau^2)\cdot\Big(\nu^2_{[T_c^1(\tilde\sigma_\tau^{c,1},\tilde\beta_\tau^2),\tilde\beta^2_{\tau+1:}]} + \lambda_{\tau+1}\cdot\overrightarrow{\|T_c^1(\sigma_\tau^{c,1},\tilde\beta_\tau^2) - T_c^1(\tilde\sigma_\tau^{c,1},\tilde\beta_\tau^2)\|_1}\Big) \tag{17}$$

(Lem. 7: $\lambda_{\tau+1}$-LC of $\nu^2_{[T_c^1(\sigma_\tau^{c,1},\tilde\beta_\tau^2),\tilde\beta^2_{\tau+1:}]}$)

$$= {\beta_\tau^1}^\top\cdot\Big[r(\sigma_\tau,\cdot,\tilde\beta_\tau^2) + \gamma T_m^1(\sigma_\tau,\cdot,\tilde\beta_\tau^2)\cdot\Big(\nu^2_{[T_c^1(\tilde\sigma_\tau^{c,1},\tilde\beta_\tau^2),\tilde\beta^2_{\tau+1:}]} + \lambda_{\tau+1}\cdot\overrightarrow{\|T_c^1(\sigma_\tau^{c,1},\tilde\beta_\tau^2) - T_c^1(\tilde\sigma_\tau^{c,1},\tilde\beta_\tau^2)\|_1}\Big)\Big] \tag{18}$$

(Linearity in β_τ^1)

$$= {\beta_\tau^1}^\top\cdot\Big[r(\sigma_\tau,\cdot,\tilde\beta_\tau^2) + \gamma T_m^1(\sigma_\tau,\cdot,\tilde\beta_\tau^2)\cdot\nu^2_{[T_c^1(\tilde\sigma_\tau^{c,1},\tilde\beta_\tau^2),\tilde\beta^2_{\tau+1:}]} + \gamma\lambda_{\tau+1}\cdot\|T(\sigma_\tau,\cdot,\tilde\beta_\tau^2) - T_m^1(\sigma_\tau,\cdot,\tilde\beta_\tau^2)T_c^1(\tilde\sigma_\tau^{c,1},\tilde\beta_\tau^2)\|_1\Big]$$

(Alternative writing)

From this, we can deduce the following appropriate forms of (i) upper bounding approximation for $W_\tau^{1,*}$ and (ii) (symmetrically) of lower bound approximation for $W_\tau^{2,*}$:

$$\overline{W}_\tau(\sigma_\tau, \beta_\tau^1) = \min_{\langle \tilde{\sigma}_\tau^{c,1}, \beta_\tau^2, \overline{\nu}_{\tau+1}^2 \rangle \in \overline{\mathcal{I}}_\tau} {\beta_\tau^1}^\top \cdot \Big[r(\sigma_\tau, \cdot, \beta_\tau^2) + \gamma T_m^1(\sigma_\tau, \cdot, \beta_\tau^2) \cdot \overline{\nu}_{\tau+1}^2 + \gamma \lambda_{\tau+1} \cdot \|T(\sigma_\tau, \cdot, \beta_\tau^2) - T_m^1(\sigma_\tau, \cdot, \beta_\tau^2) T_c^1(\tilde{\sigma}_\tau^{c,1}, \beta_\tau^2)\|_1 \Big], \text{ and}$$

$$\underline{W}_\tau(\sigma_\tau, \beta_\tau^2) = \max_{\langle \tilde{\sigma}_\tau^{c,2}, \beta_\tau^1, \underline{\nu}_{\tau+1}^1 \rangle \in \underline{\mathcal{I}}_\tau} {\beta_\tau^2}^\top \cdot \Big[r(\sigma_\tau, \beta_\tau^1, \cdot) + \gamma T_m^2(\sigma_\tau, \beta_\tau^1, \cdot) \cdot \underline{\nu}_{\tau+1}^1 - \gamma \lambda_{\tau+1} \cdot \|T(\sigma_\tau, \beta_\tau^1, \cdot) - T_m^2(\sigma_\tau, \beta_\tau^1, \cdot) T_c^2(\tilde{\sigma}_\tau^{c,2}, \beta_\tau^1)\|_1 \Big],$$

where $\overline{\nu}_{\tau+1}^2$ and $\underline{\nu}_{\tau+1}^1$ respectively upper and lower bound the actual vectors associated to the players' future strategies (resp. of 2 and 1).

Again, $\tau = H - 1$ is a particular case where only the reward term is preserved.

This constitutes the proof to the following proposition

Proposition 3.4. (originally stated on page 8) *Let $\overline{\mathcal{I}}_\tau$ be a set of tuples $w = \langle \tilde{\sigma}_\tau, \beta_\tau^2, \langle \overline{\nu}_{\tau+1}^2, \beta_{\tau+1:}^2 \rangle \rangle$. Then,*

$$\overline{W}_\tau(\sigma_\tau, \beta_\tau^1) \stackrel{\text{def}}{=} \min_{\langle \tilde{\sigma}_\tau, \beta_\tau^2, \langle \overline{\nu}_{\tau+1}^2, \beta_{\tau+1:}^2 \rangle \rangle \in \overline{\mathcal{I}}_\tau} \big[r(\sigma_\tau, \beta_\tau^1, \beta_\tau^2) + \gamma T_m^1(\sigma_\tau, \beta_\tau^1, \beta_\tau^2) \cdot \overline{\nu}_{\tau+1}^2 + \lambda_{\tau+1} \|T(\sigma_\tau, \beta_\tau^1, \beta_\tau^2) - T_m^1(\sigma_\tau, \beta_\tau^1, \beta_\tau^2) T_c^1(\tilde{\sigma}_\tau^{c,1}, \beta_\tau^2)\|_1 \big] \tag{6}$$

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

D.3 Related Operators

D.3.1 Selection Operator: Solving for β_τ^1 as an LP

Proposition D.1. *Using now a distribution ψ_τ^2 over tuples $w = \langle \tilde{\sigma}_\tau^{c,1}, \beta_\tau^2, \overline{\nu}_{\tau+1}^2 \rangle \in \overline{\mathcal{I}}_\tau^1$, the corresponding upper-bounding value for "profile" $\langle \beta_\tau^1, \psi_\tau^2 \rangle$ when in σ_τ can be written as an expectancy:*

$${\beta_\tau^1}^\top \cdot M^{\sigma_\tau} \cdot \psi_\tau^2,$$

where M^{σ_τ} is an $|\Theta_\tau^1 \times \mathcal{A}^1| \times |\overline{\mathcal{I}}_\tau^1|$ matrix.

Proof. From the right-hand side term in (Equation (18)), the upper-bounding value associated to σ_τ, β_τ^1 and a tuple $\langle \tilde{\sigma}_\tau^{c,1}, \beta_\tau^2, \overline{\nu}_{\tau+1}^2 \rangle \in \overline{\mathcal{I}}_\tau^1$ can be written:

$${\beta_\tau^1}^\top \cdot \Big[r(\sigma_\tau, \cdot, \beta_\tau^2) + \gamma T_m^1(\sigma_\tau, \cdot, \beta_\tau^2) \cdot \Big(\overline{\nu}_{\tau+1}^2 + \lambda_{\tau+1} \cdot \overrightarrow{\|T_c^1(\sigma_\tau^{c,1}, \tilde{\beta}_\tau^2) - T_c^1(\tilde{\sigma}_\tau^{c,1}, \tilde{\beta}_\tau^2)\|_1} \Big) \Big].$$

Using now a distribution ψ_τ^2 over tuples $w = \langle \tilde{\sigma}_\tau^{c,1}, \beta_\tau^2, \overline{\nu}_{\tau+1}^2 \rangle \in \overline{\mathcal{I}}_\tau^1$, the corresponding upper-bounding value for "profile" $\langle \beta_\tau^1, \psi_\tau^2 \rangle$ when in σ_τ can be written as an expectancy:

$$\sum_{w \in \overline{W}_\tau} {\beta_\tau^1}^\top \cdot \Big[r(\sigma_\tau, \cdot, \beta_\tau^2[w]) + \gamma T_m^1(\sigma_\tau, \cdot, \beta_\tau^2[w]) \cdot \Big(\overline{\nu}_{\tau+1}^2[w] + \lambda_{\tau+1} \cdot \overrightarrow{\|T_c^1(\sigma_\tau^{c,1}, \beta_\tau^2[w]) - T_c^1(\tilde{\sigma}_\tau^{c,1}[w], \beta_\tau^2[w])\|_1} \Big) \Big] \cdot \psi_\tau^2(w)$$

(where $x[w]$ denotes the field x of tuple w)

$$= {\beta_\tau^1}^\top \cdot M^{\sigma_\tau} \cdot \psi_\tau^2,$$

where M^{σ_τ} is an $|\Theta_\tau^1 \times \mathcal{A}^1| \times |\overline{\mathcal{I}}_\tau^1|$ matrix. □

For implementation purposes, using Eqs. (2) and (16) (to develop respectively $r(\cdot, \cdot, \cdot)$ and $T_m^1(\cdot, \cdot, \cdot)$), we can derive the expression of a component, *i.e.*, the upper-bounding value if a^1 is applied in θ_τ^1 while w is chosen:

$$M^{\sigma_\tau}_{(\langle \theta_\tau^1, a^1 \rangle, w)} \stackrel{\text{def}}{=} r(\sigma_\tau, \cdot, \beta_\tau^2[w]) + \gamma T_m^1(\sigma_\tau, \cdot, \beta_\tau^2[w]) \cdot \Big(\overline{\nu}_{\tau+1}^2[w] + \lambda_{\tau+1} \cdot \overrightarrow{\|T_c^1(\sigma_\tau^{c,1}, \beta_\tau^2[w]) - T_c^1(\tilde{\sigma}_\tau^{c,1}[w], \beta_\tau^2[w])\|_1} \Big)$$

$$
\begin{aligned}
&= \sum_{s,\theta^2_\tau,a^2} \sigma_\tau(\boldsymbol{\theta}_\tau) b(s|\boldsymbol{\theta}_\tau) \beta^2_\tau[w](a^2|\theta^2) r(s,\boldsymbol{a}) \\
&\quad + \gamma \sum_{z^1} \left[\sum_{\theta^2_\tau,a^2} \beta^2_\tau[w](a^2|\theta^2_\tau) \sum_{s,s',z^2} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) b(s|\boldsymbol{\theta}_\tau) \sigma_\tau(\boldsymbol{\theta}_\tau) \right] \cdot \\
&\quad \Big(\overline{\nu}^2_{\tau+1}[w](\theta^1_\tau, a^1, z^1) \\
&\quad + \lambda_{\tau+1} \cdot \overrightarrow{\| T^1_c(\sigma^{c,1}_\tau, \beta^2_\tau[w]) - T^1_c(\tilde{\sigma}^{c,1}_\tau[w], \beta^2_\tau[w]) \|_1}(\theta^1_\tau, a^1, z^1) \Big) \\
&= \sum_{\theta^2_\tau} \sigma_\tau(\boldsymbol{\theta}_\tau) \sum_{a^2} \beta^2_\tau[w](a^2|\theta^2_\tau) \\
&\quad \cdot \Bigg(\sum_s b(s|\boldsymbol{\theta}_\tau) r(s,\boldsymbol{a}) \\
&\quad + \gamma \sum_{z^1} \left[\sum_{s,s',z^2} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) b(s|\boldsymbol{\theta}_\tau) \right] \cdot \Big(\overline{\nu}^2_{\tau+1}[w](\theta^1_\tau, a^1, z^1) \\
&\quad + \lambda_{\tau+1} \cdot \overrightarrow{\| T^1_c(\sigma^{c,1}_\tau, \beta^2_\tau[w]) - T^1_c(\tilde{\sigma}^{c,1}_\tau[w], \beta^2_\tau[w]) \|_1}(\theta^1_\tau, a^1, z^1) \Big) \Bigg).
\end{aligned}
$$

Then, solving $\max_{\beta^1_\tau} \overline{W}_\tau(\sigma_\tau, \beta^1_\tau)$ can be rewritten as solving a zero-sum game where pure strategies are:

- for Player 1, the choice of not 1, but $|\Theta^1_\tau|$ actions (among $|\mathcal{A}^1|$) and,
- for Player 2, the choice of 1 element of $\overline{\mathcal{I}}^1_\tau$.

One can view it as a Bayesian game with one type per history θ^1_τ for 1, and a single type for 2.

With our upper bound, $\max_{\beta^1_\tau} \overline{W}_\tau(\sigma_\tau, \beta^1_\tau)$ can thus be solved as the LP:

$$
\begin{aligned}
\max_{\beta^1_\tau, v} v \quad \text{s.t.} \quad & \text{(i)} \ \forall w \in \overline{\mathcal{I}}^1_\tau, \quad v \leq {\beta^1_\tau}^\top \cdot M^{\sigma_\tau}_{(\cdot,w)} \\
& \text{(ii)} \ \forall \theta^1_\tau \in \Theta^1_\tau, \quad \sum_{a^1} \beta^1_\tau(a^1|\theta^1_\tau) = 1,
\end{aligned}
$$

whose dual LP is given by

$$
\begin{aligned}
\min_{\psi^2_\tau, v} v \quad \text{s.t.} \quad & \text{(i)} \ \forall (\theta^1_\tau, a^1) \in \Theta^1_\tau \times \mathcal{A}^1, \quad v \geq M^{\sigma_\tau}_{((\theta^1_\tau,a^1),\cdot)} \cdot \psi^2_\tau \\
& \text{(ii)} \qquad\qquad\qquad\qquad\qquad \sum_{w \in \overline{\mathcal{I}}^1_\tau} \psi^2_\tau(w) = 1.
\end{aligned}
$$

As can be noted, M^{σ_τ}'s columns corresponding to 0-probability histories θ^1_τ in $\sigma^{m,1}_\tau$ are empty (full of zeros), so that the corresponding decision rules (for these histories) are not relevant and can be set arbitrarily. The actual implementation thus ignores these histories, whose corresponding decision rules also do not need to be stored.

Remark D.2 (Interpretation of M^{σ_τ}). The content of this matrix can be interpreted by noting that, a given w containing a behavioral strategy $\beta^2_{\tau:H-1}$ and an OS $\tilde{\sigma}_\tau$, a pair $\langle w, \theta^1_\tau \rangle$ induces a POMDP for player 1 whose state space is made of pairs $\langle s, \theta^2_t \rangle$, and whose initial belief b_τ depends on $\tilde{\sigma}_\tau$ and θ^1_τ. Solving this POMDP amounts to finding a best response of player 1 to $\beta^2_{\tau:H-1}$. In this setting, an element $M^{\sigma_\tau}_{((\theta^1_\tau,a^1),w)}$ is an upper-bound of the optimal (POMDP) Q-value when player 1 performs a^1 while facing b_τ ($Q^*_{\text{POMDP}}(b_\tau, a^1)$).

D.3.2 Upper Bounding $\nu^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]}$

Adding a new complete tuple to $\overline{\mathcal{I}}^1_\tau$ requires a new vector $\overline{\nu}^2_\tau$ that upper bounds the vector $\nu^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]}$ associated to the strategy induced by ψ^2_τ. We can obtain one in a recursive manner (not solving the induced POMDP).

Proposition D.3. *For each ψ^2_τ obtained as the solution of the aforementioned (dual) LP in σ_τ, and each θ^1_τ, $\nu^2_{[\sigma^{c,1}_\tau, \psi^2_\tau]}(\theta^1_\tau)$ is upper bounded by a value $\overline{\nu}^2_\tau(\theta^1_\tau)$ that depends on vectors $\overline{\nu}^2_{\tau+1}$ in the support of ψ^2_τ. In particular, if $\theta^1_\tau \in Supp(\sigma^{m,1}_\tau)$, we have:*

$$
\overline{\nu}^2_\tau(\theta^1_\tau) \stackrel{\text{def}}{=} \frac{1}{\sigma^1_{\tau,m}(\theta^1_\tau)} \max_{a^1 \in \mathcal{A}^1} M^{\sigma_\tau}_{((\theta^1_\tau,a^1),.)} \cdot \psi^2_\tau.
$$

Proof. For a newly derived ψ_τ^2, as $\nu^2_{[\sigma_\tau^{c,1},\psi_\tau^2]}(\theta_\tau^1)$ is the value of 1's best action ($\in \mathcal{A}^1$) if 1 (i) observes θ_τ^1 while in $\sigma_\tau^{c,1}$ and (ii) 2 plays ψ_τ^2, we have:

$$
\begin{aligned}
&\nu^2_{[\sigma_\tau^{c,1},\psi_\tau^2]}(\theta_\tau^1) \overset{\text{def}}{=} V^\star_{[\sigma_\tau^{c,1},\psi_\tau^2]}(\theta_\tau^1) \qquad \text{(optimal POMDP value function)} \\
&= \max_{\beta^1_{\tau:}} \mathbb{E}\left[\sum_{t=\tau}^{H} \gamma^{t-\tau} R_t \mid \beta^1_{\tau:}, \theta^1_\tau, \sigma^{c,1}_\tau, \psi^2_\tau\right] \\
&= \max_{a^1} \mathbb{E}\left[R_\tau + \gamma \max_{\beta^1_{\tau+1:}} \mathbb{E}\left[\sum_{t=\tau+1}^{H} \gamma^{t-(\tau+1)} R_t \mid \beta^1_{\tau+1:}, \langle \theta^1_\tau, a^1, Z^1\rangle, \sigma^{c,1}_{\tau+1}, \psi^2_{\tau+1}\right]\right. \\
&\qquad \left. \Big| a^1, \theta^1_\tau, \sigma^{c,1}_\tau, \psi^2_\tau\right] \\
&= \max_{a^1} \mathbb{E}\left[R_\tau + \gamma V^\star_{[\sigma^{c,1}_{\tau+1},\psi^2_{\tau+1}]}(\theta^1_\tau, a^1, Z^1) \mid a^1, \theta^1_\tau, \sigma^{c,1}_\tau, \psi^2_\tau\right] \\
&= \max_{a^1} \sum_{w,\theta^2_\tau,a^2,z^1} \underbrace{Pr(w, \theta^2_\tau, z^1, a^2 \mid a^1, \theta^1_\tau, \sigma^{c,1}_\tau, \psi^2_\tau)}_{} \\
&\quad \cdot \Big(r(\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau) + \gamma \nu^2_{[\sigma^{c,1}_{\tau+1},\psi^2_{\tau+1}[w]]}(\theta^1_\tau, a^1, z^1)\Big)
\end{aligned}
$$

(where $\sigma^{c,1}_{\tau+1} = T^1_c(\sigma^{c,1}_\tau, \beta^2_\tau[w])$ (Lemma 4, p. 29))

$$
\begin{aligned}
&= \max_{a_1} \sum_{w,\theta^2_\tau,a^2,z^1} \underbrace{Pr(w|\psi^2_\tau)} \cdot \underbrace{Pr(\theta^2_\tau|\theta^1_\tau, \sigma^{c,1}_\tau)} \cdot \underbrace{Pr(a^2|\beta^2_\tau[w], \theta^2_\tau)} \cdot \underbrace{Pr(z^1|\boldsymbol{\theta}_\tau, \boldsymbol{a}_\tau)} \\
&\quad \cdot \Big(r(\boldsymbol{\theta}_\tau, \boldsymbol{a}) + \gamma \nu^2_{[\sigma^{c,1}_{\tau+1},\psi^2_{\tau+1}[w]]}(\theta^1_\tau, a^1, z^1)\Big) \\
&= \max_{a_1} \sum_w \psi^2_\tau(w) \sum_{\theta^2_\tau} \sigma^{c,1}_\tau(\theta^2_\tau|\theta^1_\tau) \sum_{a^2} \beta^2_\tau[w](a^2|\theta^2_\tau) \\
&\quad \cdot \left(r(\boldsymbol{\theta}_\tau, \boldsymbol{a}) + \gamma \sum_{z^1} Pr(z^1|\boldsymbol{\theta}_\tau, \boldsymbol{a}) \underbrace{\nu^2_{[\sigma^{c,1}_{\tau+1},\psi^2_{\tau+1}[w]]}(\theta^1_\tau, a^1, z^1)}\right)
\end{aligned}
$$

then, as $\nu^2_{[\sigma^{c,1}_{\tau+1},\psi^2_{\tau+1}[w]]}$ is $\lambda_{\tau+1}$-LC in (any) $\sigma^{c,1}_{\tau+1}$ (Lemma 7),

$$
\begin{aligned}
&\leq \max_{a_1} \sum_w \psi^2_\tau(w) \sum_{\theta^2_\tau} \sigma^{c,1}_\tau(\theta^2_\tau|\theta^1_\tau) \sum_{a^2} \beta^2_\tau[w](a^2|\theta^2_\tau) \cdot \left(r(\boldsymbol{\theta}_\tau, \boldsymbol{a}) + \gamma \sum_{z^1} Pr(z^1|\boldsymbol{\theta}_\tau, \boldsymbol{a})\right. \\
&\quad \left. \cdot \left[\overbrace{\underbrace{\nu^2_{[\tilde{\sigma}^{c,1}_{\tau+1}[w],\psi^2_{\tau+1}[w]]}(\theta^1_\tau, a^1, z^1)} + \lambda_{\tau+1} \overrightarrow{\|\sigma^{c,1}_{\tau+1} - \tilde{\sigma}^{c,1}_{\tau+1}[w]\|_1}(\theta^1_\tau, a^1, z^1)}^{}\right]\right) \\
&\leq \max_{a_1} \sum_w \psi^2_\tau(w) \sum_{\theta^2_\tau} \sigma^{c,1}_\tau(\theta^2_\tau|\theta^1_\tau) \sum_{a^2} \beta^2_\tau[w](a^2|\theta^2_\tau) \cdot \left(\underbrace{r(\boldsymbol{\theta}_\tau, \boldsymbol{a})} + \gamma \sum_{z^1} \underbrace{Pr(z^1|\boldsymbol{\theta}_\tau, \boldsymbol{a})}\right. \\
&\quad \left. \cdot \left[\overbrace{\overline{\nu}^2_{\tau+1}[w](\theta^1_\tau, a^1, z^1)} + \lambda_{\tau+1} \overrightarrow{\|\sigma^{c,1}_{\tau+1} - \tilde{\sigma}^{c,1}_{\tau+1}[w]\|_1}(\theta^1_\tau, a^1, z^1)\right]\right) \\
&= \max_{a_1} \sum_w \psi^2_\tau(w) \sum_{\theta^2_\tau} \sigma^{c,1}_\tau(\theta^2_\tau|\theta^1_\tau) \sum_{a^2} \beta^2_\tau[w](a^2|\theta^2_\tau) \cdot \left(\overbrace{\sum_s b(s|\boldsymbol{\theta}_\tau) r(s, \boldsymbol{a})}\right. \\
&\quad + \gamma \sum_{z^1} \left(\overbrace{\sum_s b(s|\boldsymbol{\theta}_\tau) \underbrace{Pr(z^1|s, \boldsymbol{a})}}\right) \cdot \Big[\overline{\nu}^2_{\tau+1}[w](\theta^1_\tau, a^1, z^1) \\
&\quad \left. + \lambda_{\tau+1} \underbrace{\overrightarrow{\|\sigma^{c,1}_{\tau+1} - \tilde{\sigma}^{c,1}_{\tau+1}[w]\|_1}(\theta^1_\tau, a^1, z^1)}\Big]\right) \\
&= \max_{a_1} \sum_w \psi^2_\tau(w) \sum_{\theta^2_\tau} \sigma^{c,1}_\tau(\theta^2_\tau|\theta^1_\tau) \sum_{a^2} \beta^2_\tau[w](a^2|\theta^2_\tau)
\end{aligned}
$$

$$\cdot \left(\sum_s b(s|\boldsymbol{\theta}_\tau) r(s, \boldsymbol{a}) + \gamma \sum_{z^1} \left(\overbrace{\sum_{s,s',z^2} b(s|\boldsymbol{\theta}_\tau) P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s)}^{} \right) \cdot \Big[\overline{\nu}^2_{\tau+1}[w](\theta^1_\tau, a^1, z^1) \right.$$
$$\left. + \lambda_{\tau+1} \overbrace{\|T^1_c(\sigma^{c,1}_\tau, \beta^2_\tau[w]) - T^1_c(\tilde{\sigma}^{c,1}_\tau[w], \beta^2_\tau[w])\|_1}^{\rightarrow}(\theta^1_\tau, a^1, z^1) \Big] \right)$$
$$= \frac{1}{\sigma^1_{\tau,m}(\theta^1_\tau)} \max_{a^1 \in \mathcal{A}^1} M^{\sigma_\tau}_{((\theta^1_\tau, a^1),.)} \cdot \psi^2_\tau.$$

□

D.3.3 Strategy Conversion

Firstly, we give details regarding solutions of Dual LPs (Equation (7)) inducing behavioral strategies. As suggested in Section 3.2.2, one can show by induction that for any timestep τ, each ψ^2_τ is actually equivalent to an element of $\Delta(\mathcal{B}^2_{\tau:})$. The following lemma shows that for any timestep τ, each element of $\Delta(\mathcal{B}^2_{\tau:})$ induces an element of $\mathcal{B}^2_{\tau:}$.

Lemma 9. *Each $\psi^i_\tau \in \Delta(\mathcal{B}^i_{\tau:})$ induces a behavioral strategy. More precisely, we prove that (i) there is a natural injection from the set $\mathcal{B}^i_{\tau:}$ to the set of distributions $\Delta(\mathcal{B}^i_{\tau:})$ and (ii) there is a surjection from the set $\Delta(\mathcal{B}^i_{\tau:})$ to $\mathcal{B}^i_{\tau:}$.*

Proof. By induction on $\tau \in \{0, \dots, H-1\}$, we prove that $\cup_{t=\tau}^H \mathcal{B}^i_t \subset \cup_{t=\tau}^H \Delta(\mathcal{B}^i_t)$. Firstly, for $\tau = H-1$, for all $\beta_{H-1} \in \mathcal{B}_{H-1}$, one can pick the degenerate distribution $\psi_{H-1} = \beta_{H-1}$ which is in $\Delta(\mathcal{B}^i_{H-1})$. Next, assume that $\cup_{t=\tau+1}^{H-1} \mathcal{B}^i_t \subset \cup_{t=\tau+1}^{H-1} \Delta(\mathcal{B}^i_t)$ for some $\tau \in \{0, \dots, H-2\}$, then for all $\beta_{\tau:H-1},\ \beta_{\tau:H-1} = \beta_\tau \oplus \beta_{\tau+1:H-1}$. By the induction hypothesis, there is $\psi_{\tau+1:H-1} \in \Delta(\mathcal{B}_{\tau+1:H-1})$ equal to $\beta_{\tau+1:H-1}$. Thus, we define $\psi_{\tau:H-1} = \beta_\tau \oplus \psi_{\tau+1:H-1} = \beta_\tau \oplus \beta_{\tau+1:H-1} = \beta_{\tau:H-1}$ which is in $\mathcal{B}_{\tau:H-1}$. From this follows a natural injection from the behavioral strategies' set to the set of distributions over behavioral strategies.

The surjection from $\Delta(\mathcal{B}^i_{\tau:})$ to $\mathcal{B}^i_{\tau:}$ is given by the realization weight computation algorithm detailed in Algorithm 2. □

As discussed in Appendix D.3.3, no effort is required to extract a solution strategy for a player from the lower bound (for 1) or the upper bound (for 2), but that strategy is in an unusual recursive form. We will here see (in the finite horizon setting) how to derive a (unique) equivalent behavioral strategy $\beta^i_{0:}$ using realization weights [Koller et al., 1994] in intermediate steps. To that end, we first define these realization weights in the case of a behavioral strategy (rather than for a mixed strategy as done by Koller et al.) and present some useful properties.

About Realization Weights Let us denote $rw^i(a^i_0, z^i_1, a^i_1, \dots, a^i_\tau)$ the *realization weight* (RW) of sequence $a^i_0, z^i_1, a^i_1, \dots, a^i_\tau$ under strategy $\beta^i_{0:}$, defined as

$$rw^i(a^i_0, z^i_1, a^i_1, \dots, a^i_\tau) \overset{\text{def}}{=} \prod_{t=0}^{\tau} \beta^i_{0:}(a^i_t | a^i_0, z^i_1, a^i_1, \dots, z^i_t) \tag{19}$$
$$= rw^i(a^i_0, z^i_1, a^i_1, \dots, a^i_{\tau-1}) \cdot \beta^i_{0:}(a^i_\tau | \underbrace{a^i_0, z^i_1, a^i_1, \dots, z^i_\tau}_{\theta^i_\tau}). \tag{20}$$

This definition already leads to useful results such as:

$$\beta^i_{0:}(a^i_\tau | \theta^i_\tau) = \frac{rw^i(\theta^i_{\tau-1}, a^i_{\tau-1}, z^i_\tau, a^i_\tau)}{rw^i(\theta^i_{\tau-1}, a^i_{\tau-1})}, \tag{21}$$

and

$$\forall z^i_\tau, \quad rw^i(\theta^i_{\tau-1}, a^i_{\tau-1}) = rw^i(\theta^i_{\tau-1}, a^i_{\tau-1}) \cdot \underbrace{\sum_{a^i_\tau} \beta(a^i_\tau | \theta^i_{\tau-1}, a^i_{\tau-1}, z^i_\tau)}_{=1} \tag{22}$$
$$= \sum_{a^i_\tau} rw^i(\theta^i_{\tau-1}, a^i_{\tau-1}) \cdot \beta(a^i_\tau | \theta^i_{\tau-1}, a^i_{\tau-1}, z^i_\tau) \tag{23}$$
$$= \sum_{a^i_\tau} rw^i(\theta^i_{\tau-1}, a^i_{\tau-1}, z^i_\tau, a^i_\tau). \tag{24}$$

We now extend Koller et al.'s definition by introducing *conditional realization weights*, where the realization weight of a *suffix sequence* is "conditioned" on a *prefix sequence*:

$$rw^i(\underbrace{a^i_\tau, \dots, a^i_{\tau'}}_{\text{suffix seq.}} | \underbrace{a^i_0, \dots, z^i_\tau}_{\text{prefix seq.}}) \overset{\text{def}}{=} \prod_{t=\tau}^{\tau'} \beta^i_{0:}(a^i_t | a^i_0, \dots, z^i_\tau, a^i_\tau, \dots, z^i_t) \tag{25}$$

$$= \beta^i_{0:}(a^i_\tau | a^i_0, \dots, z^i_\tau) \cdot rw^i(a^i_{\tau+1}, \dots, a^i_{\tau'} | a^i_0, \dots, z^i_{\tau+1}). \tag{26}$$

As can be noted, this definition only requires the knowledge of a partial strategy $\beta^i_{\tau:}$ rather than a complete strategy $\beta^i_{0:}$.

Mixing Realization Weights Let $\tau' \geq \tau + 1$, and $rw^i[w]$ denote the realization weights of some element w at $\tau + 1$. Then, for some ψ^i_τ, we have

$$rw[\psi^i_\tau](a^i_{\tau+1}, \dots, a^i_{\tau'} | a^i_0, \dots, z^i_{\tau+1}) \tag{27}$$

$$= \sum_w \psi^i_\tau(w) \cdot rw[w](a^i_{\tau+1}, \dots, a^i_{\tau'} | a^i_0, \dots, z^i_{\tau+1}). \tag{28}$$

Algorithm 2: Extracting $\beta^i_{0:}$ from w^i_0

Fct Extract(w^i_0)
  /* Step 1., keeping only $rw(\theta^i_{0:H-1})$ for all $\theta^i_{0:H-1}$ */
  $\left(rw(\theta^i_{0:H-1})\right)_{\theta^i_{0:H-1}} \leftarrow$ **RecGetRWMix**$(0, w^i_0)$
  /* Step 2. */
  for $t = H-2, \dots, 0$ **do**
    forall $\theta^i_{0:t}, a^i_t$ **do**
      $z^i_{t+1} \leftarrow z^i$ s.t. $\beta_t(\cdot | \theta^i_{0:t}, a^i_t, z^i)$ is defined
      $rw(\theta^1_{0:t}, a^i_t) \leftarrow \sum_{a^i_{t+1}} rw(\theta^i_{0:t}, a^i_t, z^i_{t+1}, a^i_{t+1} | -)$
  /* Step 3. */
  for $t = H-1, \dots, 0$ **do**
    forall $\theta^i_{0:t}, a^i_t$ **do**
      $\beta^i_t(a^i_t | \theta^i_{0:t}) \leftarrow \frac{rw^i(\theta^i_{0:t-1}, a^i_{t-1}, z^i_t, a^i_t)}{rw^i(\theta^i_{0:t-1}, a^i_{t-1})}$
  return $\beta^i_{0:}$
Fct RecGetRWMix$(t, w = \langle \beta^i_t, \psi^i_t \rangle)$
  for w' *s.t.* $\psi^i_t(w') > 0$ **do**
    $rwCat[w'] \leftarrow$ **RecGetRWCat**(t, w')
  forall $(a^i_0, \dots, a^i_{H-1})$ **do**
    $rwMix[w](a^i_t, \dots, a^i_{H-1} | a^i_0, \dots, z^i_t)$
    $\leftarrow \sum_{w'} \psi^i_t(w') \cdot rwCat[w'](a^i_{t+1}, \dots, a^i_{H-1} | a^i_0, \dots, z^i_{t+1})$
  return $rwMix[w]$
Fct RecGetRWCat$(t, w = \langle \beta^i_t, \psi^i_t \rangle)$
  if $t = H-1$ **then**
    forall $(a^i_0, \dots, a^i_{H-1})$ **do**
      $rwCat[w](a^i_{H-1} | a^i_0, \dots, z^i_{H-1}) \leftarrow \beta^i_t(a^i_{H-1} | a^i_0, \dots, z^i_{H-1})$
  else
    $rwMix[w] \leftarrow$ **RecGetRWMix**(t, w)
    forall $(a^i_0, \dots, a^i_{H-1})$ **do**
      $rwCat[w](a^i_t, \dots, a^i_{H-1} | a^i_0, \dots, z^i_t) \leftarrow$
       $\beta^i_t(a^i_t | a^i_0, \dots, z^i_t) \cdot rwMix[w](a^i_{t+1}, \dots, a^i_{H-1} | a^i_0, \dots, z^i_{t+1})$
  return $rwCat[w]$

From w^i_0 to $\beta^i_{0:}$ Using the above results, function **Extract** in Algorithm 2 derives a behavioral strategy $\beta^i_{0:}$ equivalent to the recursive strategy induced by some tuple w^i_0 in 3 steps as follows:

1. **From w^i_0 to $rw(\theta^i_{0:H-1}, a^i_{H-1})$ ($\forall(\theta^i_{0:H-1}, a^i_{H-1})$) —** These (classical) realization weights are obtained by recursively going through the directed acyclic graph describing the recursive strategy, computing *full length* (conditional) realization weights $rw(\theta^i_{t:H-1}, a^i_{H-1} | \theta^i_{0:t})$ (for $t = H-1$ down to 0).

   When in a leaf node, at depth $H-1$, the initialization is given by Equation (25) when $\tau = \tau' = H-1$:

$$rw^i(a^i_{H-1} | a^i_0, \dots, z^i_{H-1}) \stackrel{\text{def}}{=} \prod_{t=H-1}^{H-1} \beta^i(a^i_t | a^i_0, \dots, z^i_t)$$

$$= \beta^i(a^i_{H-1}|a^i_0, \ldots, z^i_{H-1}).$$

Then, in the backward phase, we can compute full length realization weights $rw(\theta^i_{t+1:H-1}, a^i_{H-1}|\theta^i_{0:t})$ with increasingly longer suffixes (thus shorter prefixes) using (i) Equation (28) (in function **RecGetRWMix**, line 25) to "mix" several strategies using the distribution ψ^i_t attached to the current w, and (ii) Equation (26), with $\tau' = H-1$, (in function **RecGetRWCat**, line 16) to concatenate the behavioral decision rule β^i_t attached to the current w in front of the strategy induced by the distribution ψ^i_t also attached to w. Note: Memoization can here be used to avoid repeating the same computations.

2. **Retrieving (classical) realization weights** $rw(\theta^i_{0:t}, a^i_t|-)$ **(**$\forall t$**) —** We can now compute realization weights $rw(\theta^i_{0:t}, a^i_t|-)$ for all t's using Equation (24) (line 6).
3. **Retrieving behavioral decision rules** β^i_t **—** Applying Equation (21) (line 9) then provides the expected behavioral decision rules.

In practice, lossless compressions are used to reduce the dimensionality of the occupancy state (*cf.* Section 4.1), which are currently lost in the current implementation of the conversion. Ideally, one would like to preserve compressions whenever possible or at least retrieve them afterwards, and possibly identify further compressions in the solution strategy.

E HSVI for zs-POSGs

This section presents results that 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.

E.1 Algorithm

E.1.1 Setting ρ

Proposition E.1. *Bounding λ_τ by $\lambda^\infty = \frac{1}{2}\frac{1}{1-\gamma}[r_{\max} - r_{\min}]$ when $\gamma < 1$, and noting that*

$$thr(\tau) = \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty \frac{\gamma^{-\tau}-1}{1-\gamma} \quad \text{if } \gamma < 1 \tag{29}$$
$$(= \epsilon - \rho(r_{\max} - r_{\min})(2H+1-\tau)\tau \quad \text{if } \gamma = 1),$$

one can ensure positivity of the threshold at any $\tau \in 1\,..\,H-1$ by enforcing $0 < \rho < \frac{1-\gamma}{2\lambda^\infty}\epsilon$ (or $0 < \rho < \frac{\epsilon}{(r_{\max}-r_{\min})(H+1)H}$ if $\gamma = 1$).

Proof. Let us first consider the case $\gamma < 1$.
We have (for $\tau \in \{1\,..\,H-1\}$):

$$\begin{aligned}
thr(\tau) &= \gamma^{-\tau}\epsilon - \sum_{i=1}^{\tau} 2\rho\lambda^\infty\gamma^{-i} \\
&= \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty \sum_{i=1}^{\tau}\gamma^{-i} \\
&= \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty \left(\gamma^{-1} + \gamma^{-2} + \cdots + \gamma^{-\tau}\right) \\
&= \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty\gamma^{-1}\left(\gamma^{0} + \gamma^{-1} + \cdots + \gamma^{-(\tau-1)}\right) \\
&= \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty\gamma^{-1}\frac{\gamma^{-\tau}-1}{\gamma^{-1}-1} \\
&= \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty\frac{\gamma^{-\tau}-1}{1-\gamma}.
\end{aligned}$$

Then, let us derive the following equivalent inequalities:

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

$$\rho < \frac{1}{2\lambda^{\infty}} \frac{1-\gamma}{1-\gamma^{\tau}} \epsilon.$$

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

Let us now consider the case $\gamma = 1$.
We have (for $\tau \in \{1, \ldots, H-1\}$):

$$\begin{aligned}
thr(\tau) &\stackrel{\text{def}}{=} \epsilon - \sum_{i=1}^{\tau} 2\rho\lambda_{\tau-i} \\
&= \epsilon - \sum_{i=1}^{\tau} 2\rho(H-(\tau-i)) \cdot (r_{\max} - r_{\min}) \\
&= \epsilon - 2\rho(r_{\max} - r_{\min}) \left[\tau(H-\tau) + \sum_{i=1}^{\tau} i\right] \\
&= \epsilon - 2\rho(r_{\max} - r_{\min}) \left[\tau H - \tau^2 + \frac{1}{2}\tau(\tau+1)\right] \\
&= \epsilon - 2\rho(r_{\max} - r_{\min}) \left[(H+\frac{1}{2})\tau - \frac{1}{2}\tau^2\right] \\
&= \epsilon - \rho(r_{\max} - r_{\min}) \left[(2H+1)\tau - \tau^2\right] \\
&= \epsilon - \rho(r_{\max} - r_{\min}) \left[(2H+1-\tau)\tau\right].
\end{aligned}$$

Then, let us derive the following equivalent inequalities:

$$\begin{aligned}
0 &< thr(\tau) \\
\rho(r_{\max} - r_{\min})(2H+1-\tau)\tau &< \epsilon \qquad \text{(holds when } \tau = 0 \text{ and } \tau = H+1\text{)} \\
\rho &< \frac{\epsilon}{(r_{\max} - r_{\min})(2H+1-\tau)\tau} \text{ (when } \tau \in \{0\,..\,H+1\}\text{)}.
\end{aligned}$$

The function $f : \tau \mapsto \frac{\epsilon}{(r_{\max} - r_{\min})(2H+1-\tau)\tau}$ reaches its minimum (for $\tau \in (0, H+1)$) when $\tau = H + \frac{1}{2}$. To ensure positivity of the threshold for any $\tau \in \{1\,..\,H-1\}$, one thus just needs to set ρ as a positive value smaller than $\frac{\epsilon}{(r_{\max} - r_{\min})(H+1)H}$. □

E.2 Finite-Time Convergence

E.2.1 Convergence Proof

Proving the finite-time convergence of zs-oMG-HSVI to an error-bounded solution requires some preliminary lemmas.

Lemma 10. *Let $(\sigma_0, \ldots, \sigma_{\tau+1})$ be a full trajectory generated by zs-oMG-HSVI and $\boldsymbol{\beta}_\tau$ the behavioral* DR *profile that induced the last transition,* i.e.*, $\sigma_{\tau+1} = T(\sigma_\tau, \boldsymbol{\beta}_\tau)$. Then, after updating $\overline{W}_\tau$ and $\underline{W}_\tau$, we have that $\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) - \underline{W}_\tau(\sigma_\tau, \beta^2_\tau) \leq \gamma thr(\tau+1)$.*

Proof. By definition,

$$\begin{aligned}
\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) = \min_{\substack{\langle \tilde{\sigma}^{c,1}_\tau, \tilde{\beta}^2_\tau, \langle \overline{\nu}^2_{\tau+1}, \psi^2_{\tau+1:} \rangle \rangle \\ \in \overline{\mathcal{I}}^1_\tau}} & \beta^1_\tau \cdot \Big(r(\sigma_\tau, \cdot, \tilde{\beta}^2_\tau) + \gamma T^1_m(\sigma_\tau, \cdot, \tilde{\beta}^2_\tau) \\
& \cdot \left[\overline{\nu}^2_{\tau+1} + \lambda_{\tau+1} \overrightarrow{\|T^1_c(\sigma^{c,1}_\tau, \beta^2_\tau) - T^1_c(\tilde{\sigma}^{c,1}_\tau, \beta^2_\tau)\|_1} \right] \Big).
\end{aligned}$$

Therefore, after the update (β^2_τ and β^1_τ being added to their respective bags ($\overline{\mathcal{I}}^1_\tau$ and $\underline{\mathcal{I}}^2_\tau$) along with vectors $\overline{\nu}^2_{\tau+1}$ and $\underline{\nu}^1_{\tau+1}$),

$$\begin{aligned}
\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) &\leq \beta^1_\tau \cdot \left[r(\sigma_\tau, \cdot, \beta^2_\tau) + \gamma T^1_m(\sigma_\tau, \cdot, \beta^2_\tau) \cdot \overline{\nu}^2_{\tau+1} \right], \text{ and} \\
\underline{W}_\tau(\sigma_\tau, \beta^2_\tau) &\geq \beta^2_\tau \cdot \left[r(\sigma_\tau, \beta^1_\tau, \cdot) + \gamma T^2_m(\sigma_\tau, \beta^1_\tau, \cdot) \cdot \underline{\nu}^1_{\tau+1} \right].
\end{aligned}$$

Then,

$$\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) - \underline{W}_\tau(\sigma_\tau, \beta^2_\tau) \leq \left[\cancel{r(\sigma_\tau, \beta^1_\tau, \beta^2_\tau)} + \gamma T^1_m(\sigma_\tau, \boldsymbol{\beta}_\tau) \cdot \overline{\nu}^2_{\tau+1} \right]$$

$$
\begin{aligned}
&\quad - \left[\cancel{r(\sigma_\tau, \beta^1_\tau, \beta^2_\tau)} + \gamma T^2_m(\sigma_\tau, \boldsymbol{\beta}_\tau) \cdot \underline{\nu}^1_{\tau+1}\right] \\
&= \gamma \left[\overline{V}(T(\sigma_\tau, \boldsymbol{\beta}_\tau)) - \underline{V}(T(\sigma_\tau, \boldsymbol{\beta}_\tau))\right] \\
&\leq \gamma thr(\tau+1) \qquad \text{(Holds at the end of any trajectory.)}
\end{aligned}
$$

□

Lemma 11 (Monotonic evolution of $\overline{W}_\tau$ and $\underline{W}_\tau$)**.** *Let $K\overline{W}_\tau$ and $K\underline{W}_\tau$ be the approximations after an update at σ_τ with behavioral* DR *$\langle \overline{\beta}^1_\tau, \underline{\beta}^2_\tau \rangle$ (respectively associated to vectors $\overline{\nu}^2_{\tau+1}$ and $\underline{\nu}^1_{\tau+1}$). Let also $K^{(n+1)}\overline{W}_\tau$ and $K^{(n+1)}\underline{W}_\tau$ be the same approximations after n other updates (in various* OS*s). Then,*

$$
\begin{aligned}
&\max_{\beta^1_\tau} K^{(n+1)}\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \leq \max_{\beta^1_\tau} K\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \leq \overline{W}_\tau(\sigma_\tau, \overline{\beta}^1_\tau) \quad \textit{and} \\
&\min_{\beta^2_\tau} K^{(n+1)}\underline{W}_\tau(\sigma_\tau, \beta^2_\tau) \geq \min_{\beta^2_\tau} K\underline{W}_\tau(\sigma_\tau, \beta^2_\tau) \geq \underline{W}_\tau(\sigma_\tau, \underline{\beta}^2_\tau).
\end{aligned}
$$

Proof. Starting from the definition,

$$
\begin{aligned}
&\max_{\beta^1_\tau} K\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \\
&= \max_{\substack{\beta^1_\tau \\ \overline{\mathcal{I}}^1_\tau \cup \{\langle \sigma^{c,1}_\tau, \underline{\beta}^2_\tau, \langle \overline{\nu}^2_{\tau+1}, \psi^2_{\tau+1:} \rangle \rangle\}}} \min_{\langle \tilde{\sigma}^{c,1}_\tau, \beta^2_\tau, \langle \overline{\nu}^2_{\tau+1}, \psi^2_{\tau+1:} \rangle \rangle \in} \beta^1_\tau \cdot \Bigg[r(\sigma_\tau, \cdot, \beta^2_\tau) \\
&\qquad + \gamma T^1_m(\sigma_\tau, \cdot, \beta^2_\tau) \cdot \left(\overline{\nu}^2_{\tau+1} + \lambda_{\tau+1} \overrightarrow{\|T^1_c(\sigma^{c,1}_\tau, \beta^2_\tau) - T^1_c(\tilde{\sigma}^{c,1}_\tau, \beta^2_\tau)\|_1} \right) \Bigg] \\
&\leq \max_{\beta^1_\tau} \min_{\langle \tilde{\sigma}^{c,1}_\tau, \beta^2_\tau, \langle \overline{\nu}^2_{\tau+1}, \psi^2_{\tau+1:} \rangle \rangle \in \overline{\mathcal{I}}^1_\tau} \beta^1_\tau \cdot \Bigg[r(\sigma_\tau, \cdot, \beta^2_\tau) \\
&\qquad + \gamma T^1_m(\sigma_\tau, \cdot, \beta^2_\tau) \cdot \left(\overline{\nu}^2_{\tau+1} + \lambda_{\tau+1} \overrightarrow{\|T^1_c(\sigma^{c,1}_\tau, \beta^2_\tau) - T^1_c(\tilde{\sigma}^{c,1}_\tau, \beta^2_\tau)\|_1} \right) \Bigg] \\
&= \max_{\beta^1_\tau} \overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \\
&= \overline{W}_\tau(\sigma_\tau, \overline{\beta}^1_\tau).
\end{aligned}
$$

Then, this upper bound approximation can only be refined, so that, for any $n \in \mathbf{N}$,

$$
\begin{aligned}
&\forall \beta^1_\tau, \quad K^{(n+1)}\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \leq K\overline{W}_\tau(\sigma_\tau, \beta^1_\tau), \\
&\text{thus, } \min_{\beta^1_\tau} K^{(n+1)}\overline{W}_\tau(\sigma_\tau, \beta^1_\tau) \leq \min_{\beta^1_\tau} K\overline{W}_\tau(\sigma_\tau, \beta^1_\tau).
\end{aligned}
$$

The expected result thus holds for $\overline{W}_\tau$, and symmetrically for $\underline{W}_\tau$. □

Lemma 12. *After updating, in order, $\overline{W}_\tau$ and $\overline{V}_\tau$, we have*

$$K\overline{V}_\tau(\sigma_\tau) \leq \max_{\beta^1_\tau} K\overline{W}_\tau(\sigma_\tau, \beta^1_\tau).$$

After updating, in order, $\underline{W}_\tau$ and $\underline{V}_\tau$, we have

$$K\underline{V}_\tau(\sigma_\tau) \geq \min_{\beta^2_\tau} K\underline{W}_\tau(\sigma_\tau, \beta^2_\tau).$$

Proof. After updating $\overline{\mathcal{I}}^1_\tau$, the algorithm computes (Algorithm 1, line 20) a new solution $\overline{\psi}^2_\tau$ of the dual LP (at σ^1_τ) and the associated vector $\overline{\nu}^2_\tau$, so that

$$\max_{\beta^1_\tau} K\overline{W}_\tau(\sigma^1_\tau, \beta^1_\tau) = \sigma^{m,1}_\tau \cdot \overline{\nu}^2_\tau.$$

This vector will feed $\overline{bagV}_\tau$ along with σ^1_τ, so that

$$K\overline{V}_\tau(\sigma_\tau) \leq \sigma^{m,1}_\tau \cdot \overline{\nu}^2_\tau.$$

As a consequence,

$$K\overline{V}_\tau(\sigma_\tau) \leq \max_{\beta^1_\tau} K\overline{W}_\tau(\sigma_\tau, \beta^1_\tau).$$

The symmetric property holds for $K\underline{V}_\tau$ and $K\underline{W}_\tau$, which concludes the proof. □

Theorem 3.9. (originally stated on page 11) *zs-oMG-HSVI (Alg. 1) terminates in finite time with an ϵ-approximation of $V_0^*(\sigma_0)$ that statisfies Theorem 3.8.*

Proof. We will prove by induction from $\tau = H$ to 0, that the algorithm stops expanding OSs at depth τ after finitely many iterations (/trajectories).

First, by definition of horizon H, no OS σ_H is ever expanded. The property thus holds at $\tau = H$.

Let us now assume that the property holds at depth $\tau + 1$ after $N_{\tau+1}$ iterations. By contradiction, let us assume that the algorithm generates infinitely many trajectories of length $\tau + 1$. Then, because $\mathcal{O}_\tau^\sigma \times \mathcal{B}_\tau$ is compact, after some time the algorithm will have visited $\langle \sigma_\tau, \boldsymbol{\beta}_\tau \rangle$, then, some iterations later, $\langle \sigma_\tau^{'}, \boldsymbol{\beta}_\tau^{'} \rangle$, such that $\|\sigma_\tau - \sigma_\tau^{'}\|_1 \leq \rho$. Let us also note the corresponding terminal OSs (because trajectories beyond iteration $N_{\tau+1}$ do not go further) $\sigma_{\tau+1} = T(\sigma_\tau, \beta_\tau)$ and $\sigma_{\tau+1}^{'} = T(\sigma_\tau^{'}, \boldsymbol{\beta}_\tau^{'})$.

Now, we show that the second trajectory should not have happened, *i.e.*, $\overline{V}(\sigma_\tau^{'}) - \underline{V}(\sigma_\tau^{'}) \leq thr(\tau)$.

Combining the previous lemmas,

$$\begin{aligned}
\overline{V}(\sigma_\tau^{'}) &\leq \overline{V}(\sigma_\tau) + \lambda_\tau \|\sigma_\tau - \sigma_\tau^{'}\|_1 && \text{(By Lipschitz-Continuity)} \\
&\leq \max_{\tilde{\beta}_\tau^1} \overline{W}_\tau(\sigma_\tau, \tilde{\beta}_\tau^1) + \lambda_\tau \|\sigma_\tau - \sigma_\tau^{'}\|_1 && \text{(Lemma 12)} \\
&\leq \overline{W}_\tau(\sigma_\tau, \beta_\tau^1) + \lambda_\tau \|\sigma_\tau - \sigma_\tau^{'}\|_1 && \text{(Lemma 11)} \\
&= \overline{W}_\tau(\sigma_\tau, \beta_\tau^1) + \lambda_\tau \rho.
\end{aligned}$$

Symmetrically, we also have

$$\underline{V}(\sigma_\tau^{'}) \geq \underline{W}_\tau(\sigma_\tau, \beta_\tau^2) - \lambda_\tau \rho.$$

Hence,

$$\begin{aligned}
\overline{V}(\sigma_\tau^{'}) - \underline{V}(\sigma_\tau^{'}) &\leq \left(\overline{W}_\tau(\sigma_\tau, \beta_\tau^1) + \lambda_\tau \rho\right) - \left(\underline{W}_\tau(\sigma_\tau, \beta_\tau^2) - \lambda_\tau \rho\right) \\
&= \left(\overline{W}_\tau(\sigma_\tau, \beta_\tau^1) - \underline{W}_\tau(\sigma_\tau, \beta_\tau^2)\right) + 2\lambda_\tau \rho \\
&\leq \gamma thr(\tau + 1) + 2\lambda_\tau \rho && \text{(Lemma 10)} \\
&= \gamma \left(\gamma^{-(\tau+1)} \epsilon - \sum_{i=1}^{\tau+1} 2\rho \lambda_{\tau+1-i} \gamma^{-i} \right) + 2\lambda_\tau \rho \\
&= \gamma^{-\tau} \epsilon - \sum_{i=1}^{\tau+1} 2\rho \lambda_{\tau+1-i} \gamma^{-i+1} + 2\lambda_\tau \rho \\
&= \gamma^{-\tau} \epsilon - \sum_{j=0}^{\tau} 2\rho \lambda_{\tau-j} \gamma^{-j} + 2\lambda_\tau \rho \\
&= \gamma^{-\tau} \epsilon - \cancel{2\rho \lambda_{\tau-0} \gamma^{-0}} - \sum_{j=1}^{\tau} 2\rho \lambda_{\tau-j} \gamma^{-j} + \cancel{2\lambda_\tau \rho} = thr(\tau).
\end{aligned}$$

Therefore, $\sigma_\tau^{'}$ should not have been expanded. This shows that the algorithm will generate only a finite number of trajectories of length τ. □

E.2.2 Handling Infinite Horizons

Proposition 3.10. (originally stated on page 12) *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.*

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

$$\begin{aligned}
thr(\tau) &< W \\
\gamma^{-\tau} \epsilon - 2\rho\lambda^\infty \frac{\gamma^{-\tau} - 1}{1 - \gamma} &< W && \text{(from Eq. (29))}
\end{aligned}$$

$$
\begin{aligned}
\gamma^{-\tau}\epsilon - 2\rho\lambda^{\infty}\frac{\gamma^{-\tau}}{1-\gamma} - 2\rho\lambda^{\infty}\frac{-1}{1-\gamma} &< W \\
\gamma^{-\tau}\underbrace{\left(\epsilon - \frac{2\rho\lambda^{\infty}}{1-\gamma}\right)}_{>0 \quad \text{(Prop. E.1)}} &< W - \frac{2\rho\lambda^{\infty}}{1-\gamma} \\
\gamma^{-\tau} &< \frac{W - \frac{2\rho\lambda^{\infty}}{1-\gamma}}{\epsilon - \frac{2\rho\lambda^{\infty}}{1-\gamma}} \\
\exp(-\tau\ln(\gamma)) &< \exp\left(\ln\left(\frac{W - \frac{2\rho\lambda^{\infty}}{1-\gamma}}{\epsilon - \frac{2\rho\lambda^{\infty}}{1-\gamma}}\right)\right) \\
-\tau\ln(\gamma) &< \ln\left(\frac{W - \frac{2\rho\lambda^{\infty}}{1-\gamma}}{\epsilon - \frac{2\rho\lambda^{\infty}}{1-\gamma}}\right) \\
\tau\ln(\gamma) &> \ln\left(\frac{\epsilon - \frac{2\rho\lambda^{\infty}}{1-\gamma}}{W - \frac{2\rho\lambda^{\infty}}{1-\gamma}}\right) \\
\tau &< \log_{\gamma}\left(\frac{\epsilon - \frac{2\rho\lambda^{\infty}}{1-\gamma}}{W - \frac{2\rho\lambda^{\infty}}{1-\gamma}}\right).
\end{aligned}
$$

□

Even if the problem horizon is infinite, trajectories will thus have bounded length. Then, everything beyond this *effective* horizon will rely on the upper- and lower-bound initializations and the corresponding strategies.