HSVI pour zs-POSG usant de propriétés de convexité, concavité, et Lipschitz-continuité

Aurélien Delage, Olivier Buffet, Jilles Dibangoye

HSVI pour zs-POSG usant de propriétés de convexité, concavité, et Lipschitz-continuité

Aurélien Delage[1] Olivier Buffet[2] Jilles Dibangoye[1]

[1] Univ Lyon, INSA Lyon, INRIA, CITI, F-69621 Villeurbanne, France
[2] Université de Lorraine, INRIA, CNRS, LORIA, F-54000 Nancy, France

prénom.nom@inria.fr

Résumé

La résolution d'un jeu stochastique partiellement observable à 2 joueurs et somme nulle (zs-POSG) repose typiquement sur sa transformation en un jeu en forme extensive, perdant ainsi de l'information structurelle contenue dans la représentation originale. Nous évitons une telle perte en transformant le problème plutôt en un jeu de Markov sur les états d'occupation, ce qui permet de tirer parti d'approches de l'état de l'art pour les processus de décision markoviens partiellement observables, les POMDP décentralisés, et un certain nombre de sous-classes de zs-POSG. Pour appliquer le principe d'optimalité de Bellman dans ce cadre, nous (i) exhibons des propriétés de continuité originales de la fonction de valeur optimale ; (ii) dérivons des approximateurs encadrants à base de points ; et (iii) proposons des opérateurs de mise à jour efficaces reposant sur la programmation linéaire. Une variante de l'algorithm HSVI de SMITH *et* SIMMONS *[23] est construite sur la base de ces idées et nous prouvons sa convergence vers une solution ϵ-optimale avant de présenter des résultats expérimentaux.*

Mots Clef

POSG, jeu stochastique partiellement observable, principe d'optimalité de Bellman, Heuristic Search Value Iteration, jeu à information imparfaite.

Abstract

Solving a 2-player zero-sum partially observable stochastic game (zs-POSG) typically relies on turning it into an extensive-form game, thus losing structural information contained in the original representation. We prevent such a loss by turning the problem instead into an occupancy Markov game, which allows building on state-of-the-art approaches for partially observable Markov decision processes, decentralized POMDPs, and a number of subclasses of zs-POSGs. To apply Bellman's optimality principle in this setting, we (i) exhibit novel continuity properties of the optimal value function ; (ii) derive point-based bounding approximators ; and (iii) propose efficient backup operators based on linear programming. A variant of SMITH *et* SIMMONS*'s (2005) HSVI algorithm is built on these ideas and we prove its finite-time convergence to an ϵ-optimal solution before presenting experimental results.*

Keywords

POSG, partially observable stochastic game, Bellman's optimality principle, Heuristic Search Value Iteration, imperfect information game.

1 Introduction

Résoudre des jeux séquentiels à information imparfaite est un sujet difficile avec de nombreuses applications, du jeu du Poker [17] aux jeux de sécurité [2]. Nous nous concentrons sur les jeux stochastiques partiellement observables à 2 joueurs et somme nulle ((2p) zs-POSG), une sous-classe importante venant avec des représentations de problèmes qui sont compactes et permettent d'exploiter leur structure (par exemple, pour dériver des relaxations). Les principales techniques de résolution des zs-POSG requièrent de les transformer en jeux à forme extensive et à somme nulle (zs-EFG) [20]. Ensuite, une première approche repose sur la résolution d'un programme linéaire en *forme séquentielle* [15, 24, 4], ce qui engendre un algorithme exact. Une seconde approche est d'employer un solveur de jeu itératif, c'est-à-dire soit une méthode basée sur le regret contrefactuel [29, 5], soit une méthode du premier ordre [16], les deux venant avec des propriétés de convergence asymptotiques. En revanche, en évitant la transformation POSG→EFG, notre approche préserve et exploite l'information structurelle contenue dans la représentation POSG. Dans cet article, nous exploitons aussi les propriétés de concavité et de convexité de V^* (la fonction de valeur optimale des zs-POSG) introduites par WIGGERS, OLIEHOEK et ROIJERS [28] et dérivons pour la première fois une approche reposant sur le principe d'optimalité de Bellman. Elle est prometteuse, étant similaire à des approches de l'état de l'art dans des classes de problèmes connexes (POMDP [1, 22], Dec-POMDP [25, 9]) ou des sous-classes

des zs-POSG [10, 7, 3, 13, 8, 12]. Comme elles le font, nous : (i) transformons le problème initial en un problème complètement observable pour lequel le principe d'optimalité de Bellman s'applique directement (section 3) ; puis (ii) exhibons de nouvelles propriétés de continuité de V^*, lesquelles permettent de proposer des approximateurs encadrant à base de points et des opérateurs de mise-à-jour et de sélection efficaces reposant sur la programmation linéaire (sect. 4) ; et (iii) montrons comment adapter le schéma algorithmique HSVI de SMITH et SIMMONS [23] pour résoudre le problème ϵ-optimalement en un nombre fini d'itérations (sect. 5). Enfin, la sect. 6 présente des résultats expérimentaux préliminaires pour valider l'algorithme proposé.

2 Travaux connexes

Les POSG à horizon infini sont indécidables [18], ce qui justifie de chercher des solutions quasi-optimales, par exemple à travers des horizons finis (lesquels garantissent d'être à ϵ de l'optimum sous horizon infini), comme nous allons le faire. Peu de travaux se sont intéressés à la résolution de POSG. Une exception est le travail de HANSEN, BERNSTEIN et ZILBERSTEIN [11] sur les POSG à horizon fini, dans lequel des arbres-politiques non-dominés pour chaque joueur sont construits incrémentalement par programmation dynamique (DP), ce qui permet ensuite de dériver un solveur pour POSG à gain commun, c'est-à-dire des processus de décision markoviens partiellement observables (Dec-POMDP). À l'inverse, nous employons le principe d'optimalité de Bellman pour estimer la fonction de valeur optimale.

Dec-POMDP Le principe d'optimalité de Bellman apparaît comme le composant principal d'un solveur de Dec-POMDP quand SZER, CHARPILLET et ZILBERSTEIN [25] adoptent un point de vue centré-planificateur dans lequel le planificateur fournit aux joueurs leurs politiques privées couvrant tous les historiques d'action-observation qu'ils pourraient rencontrer. L'état d'information du planificateur à t contient ainsi l'état de croyance initial et la politique jointe jusqu'à t. Cela conduit à un MDP sur les états d'information avec une dynamique déterministe résolu avec une recherche A* appelée MAA* (*multi-agent A**).
Ensuite, une autre étape importante est quand DIBANGOYE et coll. [9] montrent que (i) l'*état d'occupation*, une statistique employée pour calculer les récompenses espérées dans MAA*, est en fait une statistique suffisant pour la planification, et (ii) la fonction de valeur optimale est linéaire par morceaux et convexe (PWLC) dans l'espace des états d'occupation, ce qui permet d'adapter des solveurs POMDP à base de points en utilisant des approximateurs de V^*.

Sous-classes de POSG Des travaux récents ont abordé des cas particuliers des jeux stochastiques partiellement observables atténués (POSG), à 2 joueurs et somme-0 sauf indication contraire, en exploitant la structure du problème pour le transformer en un problème équivalent dans lequel le principe d'optimalité de Bellman s'applique. GHOSH, MCDONALD et SINHA [10] ont considéré des POSG avec actions publiques et observations partagées, lesquels peuvent être convertis en des jeux stochastiques définis sur l'espace commun des états de croyance, de même que pour les POMDP convertis en MDP sur les états de croyance. CHATTERJEE et DOYEN, BASU et STETTNER, HORÁK, BOŠANSKÝ et PĚCHOUČEK [7, 3, 13] ont considéré des One-Sided POSG, c'est-à-dire des scénarios dans lesquels (le joueur) 2 (sans perte de généralité) n'observe que partiellement l'état du système, et 1 a accès à l'état du système ainsi qu'à l'action et à l'observation de 2. COLE et KOCHERLAKOTA [8] ont considéré des POSG (à n joueurs) avec des états privés indépendants, une observabilité partiellement partagée, et la fonction d'utilité du joueur i dépendant de son état privé et de son observation partagée. HORÁK et BOŠANSKÝ [12] ont considéré des zs-POSG avec des états privés indépendants et des observations publiques, c'est-à-dire des scénarios dans lesquels (i) chaque joueur i a un état privé s_i qu'il observe complètement, et (ii) les deux joueurs reçoivent la même observation publique de l'état privé de chaque joueur. Toute croyance d'un joueur sur l'état privé de l'autre joueur est ainsi une connaissance commune.
Concernant les travaux de HORÁK et BOŠANSKÝ [12] [13], dans les deux cas des propriété de convexité et de concacivté de la fonction de valeur optimale sont obtenus, lesquels permettent de dériver des approximateurs majorant et minorant. Ces approximateurs sont ensuite employés dans des algorithmes reposant sur HSVI. Toutefois, passer des MDP et POMDP (comme dans le travail de SMITH [22]) à ces cadres induit un arbre des futurs possibles dont le facteur de branchement est infini, ce qui requiert des modifications au schéma algorithmique, et donc à l'analyse théorique de la convergence en temps fini. Comme on le verra, le présent travail adopte des modifications similaires.
WIGGERS, OLIEHOEK et ROIJERS [27] [28] démontrent que, en utilisant des représentations appropriées, la fonction de valeur associée à un zs-POSG est convexe pour le joueur (maximisant) 1 et concave pour le joueur (minimisant) 2. Sur cette base, WIGGERS [26] propose des solveurs heuristiques dépourvus de garanties de convergence, une limitation que nous surmontons dans le présent travail.

Jeux à information imparfaite Comme déjà évoqué en introduction, les POSG à horizon fini (et somme générale) peuvent être écrits comme des jeux en forme extensive avec information imparfaite et mémoire parfaite (EFG, souvent désignés comme des *jeux à information imparfaite*) [20], ce qui rend les techniques de résolution pour EFG pertinentes même pour les POSG à horizon infini. Une première approche pour résoudre les EFG est de les convertir en un jeu en forme normale avant de chercher un équilibre de Nash, en ignorant donc l'aspect temporel du problème [21] et en induisant une explosion combinatoire. KOLLER et MEGIDDO [14] proposent une approche par programma-

tion linéaire améliorée pour zs-EFG qui exploite l'aspect temporel à travers le choix de variables de décision (voir aussi [24, 15]). En utilisant une méthode de double-oracle, BOŠANSKÝ et coll. [4] accélèrent ce processus de résolution d'autant plus que les stratégies optimales ont un petit support.

Plus récemment, des méthodes itératives ont été proposées qui fournissent des garanties de convergence asymptotiques et permettent d'aborder des jeux à information imparfaite de grande taille. (i) La minimisation du regret contrefactuel (*Counterfactual Regret minimization* (CFR)) [29] a une vitesse de convergence en $O(\frac{1}{\epsilon^2})$ (avec des versions modernes gagnant désormais contre les meilleurs joueurs humains au poker *heads-up no limit hold'em* [5]), alors que (ii) les méthodes du premier ordre (*first-order methods* (FOMs)) [16] ont une vitesse de convergence en $O(\frac{1}{\epsilon})$. Pour notre part, nous visons un algorithme à erreur bornée ayant des garanties de convergence en temps fini.

3 État de l'art

Par souci de clarté, les concepts et résultats de la littérature EFG employés dans ce travail sont retranscrits dans le cadre POSG.

Un jeu stochastique partiellement observable (à 2 joueurs et) à somme nulle (zs-POSG) est défini par un tuple $\langle \mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, \mathcal{Z}^1, \mathcal{Z}^2, P, r, H, \gamma, b_0 \rangle$, où

- $\mathcal{S}$ est un ensemble fini d'états ;
- $\mathcal{A}^i$ est l'ensemble fini des actions de i ;
- $\mathcal{Z}^i$ est l'ensemble fini des observations de i ;
- $P^{z^1,z^2}_{a^1,a^2}(s'|s)$ est la probabilité de transiter vers l'état s' et de recevoir les observations z^1 et z^2 quand les actions a^1 et a^2 sont effectuées dans l'état s ;
- $r(s, a^1, a^2)$ est une fonction de récompense (scalaire) ;
- $H \in \mathbb{N}$ est un horizon temporel (fini) ;
- $\gamma \in [0, 1]$ est un facteur d'atténuation ; et
- b_0 est l'état de croyance initial.

Le joueur 1 voudrait maximiser l'espérance du retour, définie comme la somme des récompenses futures atténuées, alors que 2 voudrait la minimiser, ce que nous formalisons plus loin. Du fait du cadre symétrique, certaines définitions ou résultats seront donnés du point de vue d'un joueur lorsque seuls des changements évidents sont requis pour l'autre.

Des littératures Dec-POMDP, POSG et EFG, nous employons les concepts et définitions suivants, où $i \in \{1, 2\}$:

$-i$ est l'adversaire de i ($-1 = 2$ & $-2 = 1$).

$\theta^i_\tau = (a^i_1, z^i_1, \ldots, a^i_\tau, z^i_\tau)$ $(\in \Theta^i = \cup_{t=0}^{H-1} \Theta^i_t)$ est un *historique d'action-observation* (AOH) de longueur τ pour i.

$\boldsymbol{\theta}_\tau = (\theta^1_\tau, \theta^2_\tau)$ $(\in \boldsymbol{\Theta} = \cup_{t=0}^{H-1} \boldsymbol{\Theta}_t)$ est une AOH *jointe* à τ.

[σ_τ] Un *état d'occupation* (OS) σ_τ $(\in \mathcal{O}^\sigma = \cup_{t=0}^{H-1} \mathcal{O}^\sigma_t)$ à τ est une distribution de probabilité sur les AOH joints $\boldsymbol{\theta}_\tau$.

[β^i_τ] Une *règle de décision (comportementale)* (DR) au temps τ pour i est une application β^i_τ des AOH privés dans Θ^i_τ vers les *distributions* sur les actions privées. Nous notons $\beta^i_\tau(\theta^i_\tau, a^i)$ la probabilité de choisir a^i en présence de θ^i_τ.

$\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$ $(\in \mathcal{B} = \cup_{t=0}^{H-1} \mathcal{B}_t)$ est un *profil de règles de décision*.

$\beta^i_{\tau:\tau'} = (\beta^i_\tau, \ldots, \beta^i_{\tau'})$ est une *stratégie comportementale* pour i du pas de temps τ à τ' (inclus).

$\boldsymbol{\beta}_{\tau:\tau'} = \langle \beta^1_{\tau:\tau'}, \beta^2_{\tau:\tau'} \rangle$ est un *profil de stratégies comportementales*.

[$V_0(\sigma_0, \boldsymbol{\beta}_{0:H-1})$] La *valeur* du profile de stratégies comportementales $\boldsymbol{\beta}_{0:H-1}$ dans σ_0 (à partir du pas de temps 0) est

$$V_0(\sigma_0, \boldsymbol{\beta}_{0:H-1}) = E[\sum_{t=0}^{H-1} \gamma^t R_t \mid \sigma_0, \boldsymbol{\beta}_{0:H-1}], \quad (1)$$

où R_t est la variable aléatoire associée à la récompense instantanée à t.

Notes : (i) Par souci de concision, l'indice «$\tau : H-1$» est souvent écrit «$\tau :$». (ii) On pourra écrire $\beta^{1\mathsf{T}}_\tau \cdot f(\cdot)$ toute fonction $f(\beta^1_\tau)$ linéaire en β^1_τ (vu comme un vecteur de $\mathbb{R}^{\Theta^1_\tau \times \mathcal{A}^1}$).

L'objectif premier est ici de trouver une stratégie à l'équilibre de Nash (NES), c'est-à-dire un profil de stratégies comportementales $\boldsymbol{\beta}^*_{0:} = \langle \beta^{1*}_{0:}, \beta^{2*}_{0:} \rangle$ tel qu'aucun joueur n'a intérêt à en dévier, ce qui peut s'écrire :

$$\forall \beta^1, V_0(\sigma_0, \beta^{1*}_{0:}, \beta^{2*}_{0:}) \geq V_0(\sigma_0, \beta^1_{0:}, \beta^{2*}_{0:}), \quad (2)$$
$$\forall \beta^2, V_0(\sigma_0, \beta^{1*}_{0:}, \beta^{2*}_{0:}) \leq V_0(\sigma_0, \beta^{1*}_{0:}, \beta^2_{0:}). \quad (3)$$

Dans un tel jeu à somme-0 et 2 joueurs, tous les NES ont la même valeur (NEV) $V^*_0(\sigma_0) \stackrel{\text{def}}{=} V_0(\sigma_0, \beta^{1*}_{0:}, \beta^{2*}_{0:})$.

Le principe d'optimalité de Bellman ne peut pas s'appliquer directement dans un jeu où les joueurs ne partagent pas leurs historiques individuels, et n'ont ainsi pas la même information sur la situation courante (sauf à $\tau = 0$). Pour remédier à ce problème, nous suivons la même idée que pour les Dec-POMDP ou certaines sous-classes de zs-POSG telles que les One-Sided POSG [13] en considérant un jeu différent dont les nouveaux joueurs (disons $\tilde{1}$ et $\tilde{2}$) n'accèdent pas aux historiques individuels de 1 et 2 en cours de partie, mais doivent (publiquement) fournir à 1 et 2 les règles de décision qu'ils devront exécuter. C'est ce que nous faisons dans la section suivante, en démontrant ultérieurement que l'on peut au final dériver des stratégies solution pour le problème original qui sont robustes aux déviations.

3.1 Résoudre des POSG comme des Occupancy MG

Nous décrivons ici la reformulation d'un zs-POSG en un jeu de Markov à somme nulle différent, complètement observable.

Notons d'abord qu'un profil de stratégies partiel $\boldsymbol{\beta}_{0:\tau-1}$ est associé de manière déterministe à un état d'occupation $\sigma_\tau = \sigma_{\boldsymbol{\beta}_{0:\tau-1}}$, lequel exhibe les propriétés suivantes.

Lemma 1. *$\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$, pris avec $\boldsymbol{\beta}_\tau$, est une statistique suffisante pour calculer (i) le prochain état d'occupation, $\sigma_{\boldsymbol{\beta}_{0:\tau}}$, et (ii) l'espérance de récompense à τ :* $\mathbb{E}\left[R_\tau \mid \boldsymbol{\beta}_{0:\tau-1} \oplus \boldsymbol{\beta}_\tau\right]$.

Cette propriété de Markov sur les états d'occupation et la capacité à estimer la récompense espérée permettent d'introduire un jeu équivalent (employé implicitement par WIGGERS, OLIEHOEK et ROIJERS [27]), appelé *jeu de Markov à somme nulle sur les états d'occupation* (*zs Occupancy Markov Game* (zs-OMG)),[1] défini formellement par le tuple $\langle \mathcal{O}^\sigma, \mathcal{B}, T, r, H, \gamma, b_0 \rangle$, où :

- $\mathcal{O}^\sigma$ est l'ensemble des états d'occupation induits par le zs-POSG ;
- $\mathcal{B}$ est l'ensemble des profils de règles de décision du zs-POSG ;
- T est une fonction de transition déterministe qui associe à chaque couple $(\sigma_\tau, \boldsymbol{\beta}_\tau)$ le (seul) état d'occupation suivant possible $\sigma_{\tau+1}$; formellement, $\forall \theta^1_\tau, a^1, z^1, \theta^2_\tau, a^2, z^2$,

$$T(\sigma_\tau, \boldsymbol{\beta}_\tau)((\theta^1_\tau, a^1, z^1), (\theta^2_\tau, a^2, z^2)) \tag{4}$$

$$\stackrel{\text{def}}{=} Pr((\theta^1_\tau, a^1, z^1), (\theta^2_\tau, a^2, z^2) | \sigma_\tau, \boldsymbol{\beta}_\tau) \tag{5}$$

$$= \beta^1_\tau(\theta^1_\tau, a^1)\beta^2_\tau(\theta^2_\tau, a^2)\sigma_\tau(\boldsymbol{\theta}_\tau) \sum_{s,s'} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) b(s|\boldsymbol{\theta}_\tau), \tag{6}$$

  où $b(s|\boldsymbol{\theta}_\tau)$ est un état de croyance obtenu par filtrage HMM ;
- r est une fonction de récompense induite par le zs-POSG comme récompense espérée pour l'état d'occupation courant et le profil de règle de décision courant :

$$r(\sigma_\tau, \boldsymbol{\beta}_\tau) \stackrel{\text{def}}{=} E[r(S, A^1, A^2) | \sigma_\tau, \boldsymbol{\beta}_\tau] \tag{7}$$

$$= \sum_{s, \boldsymbol{\theta}_\tau, \boldsymbol{a}} \sigma_\tau(\boldsymbol{\theta}_\tau) b(s|\boldsymbol{\theta}_\tau) \beta^1_\tau(a^1|\theta^1) \beta^2_\tau(a^2|\theta^2) r(s, \boldsymbol{a}); \tag{8}$$

  nous employons la même notation r pour les zs-POSG dans la mesure où le contexte indiquera duquel il est question ;
- H, γ, et b_0 sont les mêmes que dans le zs-POSG.

1. Nous utilisons (i) «jeu de Markov» au lieu de «jeu stochastique» parce que la dynamique n'est pas stochastique, et (ii) «partially observable stochastic game» pour correspondre à la littérature.

Ce n'est pas un jeu de Markov à somme nulle fini standard non plus puisque (i) il est non stationnaire, avec des espaces d'états et d'actions différents (et continus) à chaque pas de temps ; (ii) à chaque pas de temps, il y a un nombre infini d'actions, et un mélange de telles actions pures est équivalent à une action pure pré-existante ; et (iii) la dynamique est déterministe (même pour des actions «mixtes»). Mais un bénéfice important de travailler avec un zs-OMG est que $\tilde{1}$ et $\tilde{2}$ connaissent tous deux l'état d'occupation courant, σ_τ, et partagent ainsi la même information sur le jeu. Cela va permettre de trouver un ϵ-équilibre de Nash solution de ce jeu en explorant l'arbre des profils de stratégies comportementales partielles. Cet arbre a un facteur de branchement infini du fait des espaces d'action (les règles de décision) et d'état (les états d'occupation) continus (contrairement à l'arbre exploré lors de la résolution d'un Dec-POMDP à l'aide d'un occupancy MDP).

Dans un zs-OMG, à cause de l'observabilité partielle, si $\tilde{2}$ dévie de sa stratégie solution obtenue, alors $\tilde{1}$ pourrait s'adapter en-ligne en replanifiant depuis l'état d'occupation (observé) résultant (à supposer qu'il dispose d'un temps de calcul suffisant). Dans le zs-POSG correspondant, l'état d'occupation n'est pas observé, ce qui interdit une telle replanification. Toutefois, comme nous le verrons, on peut quand même dériver une stratégie solution ϵ-optimale pour 1 comme pour 2 (lesquels nous ne distinguerons plus de $\tilde{1}$ et $\tilde{2}$ désormais).

Nous allons étudier les sous-jeux d'un zs-OMG, c'est-à-dire les situations où un état d'occupation σ_τ a été atteint au temps τ, et le solveur cherche des stratégies optimales ($\beta^1_{\tau:}$ et $\beta^2_{\tau:}$) à fournir aux joueurs. σ_τ indique à quels historiques chaque joueur pourrait faire face avec probabilité non-nulle, et sont donc pertinents pour la planification. Nous pouvons alors étendre la définition de la fonction de valeur (optimale) à tout pas de temps τ comme suit :

$$V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:}) \stackrel{\text{def}}{=} E[\sum_{t=\tau}^{\infty} \gamma^{t-\tau} R(S_t, A_t) | \sigma_\tau, \boldsymbol{\beta}_{\tau:}], \tag{9}$$

$$V^*_\tau(\sigma_\tau) \stackrel{\text{def}}{=} \max_{\beta^1_{\tau:}} \min_{\beta^2_{\tau:}} V_\tau(\sigma_\tau, \beta^1_{\tau:}, \beta^2_{\tau:}). \tag{10}$$

Notons qu'une telle extension est non triviale quand on utilise des stratégies mixtes puisque celles-ci sont usuellement définies à partir du pas de temps 0.

Nous regardons maintenant des propriétés structurelles de V^* qui seront utilisées plus tard pour définir des approximateurs de fonction.

3.2 Propriétés de concavité et de convexité de V^*

WIGGERS, OLIEHOEK et ROIJERS [27] (dont nous étendons ici les résultats de $\gamma = 1$ à $\gamma \leq 1$) définissent les distributions marginale et conditionnelle du point de vue de 1 comme $\sigma^{m,1}_\tau(\theta^1_\tau) \stackrel{\text{def}}{=} \sum_{\theta^2_\tau} \sigma(\theta^1_\tau, \theta^2_\tau)$ et $\sigma^{c,1}_\tau(\theta^2_\tau | \theta^1_\tau) \stackrel{\text{def}}{=} \frac{\sigma(\theta^1, \theta^2)}{\sigma^1_{\tau,m}(\theta^1)}$, de sorte que $\sigma_\tau = \sigma^{m,1}_\tau \sigma^{c,1}_\tau$. De plus, $T^1_m(\cdot)$ et

$T_c^1(\cdot)$ dénoteront les états d'occupation marginal et conditionnel de 1 induits par $T(\cdot)$.
Pour un $\beta^2_{\tau:}$ fixé, 1 est confronté à un POMDP, et la valeur (POMDP) optimale dans chaque historique d'action-observation θ^1_τ est donnée par $\nu_{[\sigma^{c,1}_\tau, \beta^2_{\tau:}]}(\theta^1_\tau) \stackrel{\text{def}}{=} \max_{\beta^1_{\tau:}} \mathbb{E}\left\{\sum_{t=\tau}^{H-1} \gamma^{t-\tau} R(S_t, A^1_t, A^2_t) \mid \theta^1_\tau, \beta^1_{\tau:}, \beta^2_{\tau:}, \sigma^{c,1}_\tau\right\}$.
WIGGERS, OLIEHOEK et ROIJERS démontrent ensuite les propriétés de concavité et convexité suivantes de V^*.

Theorem 1 (Concavité et convexité de V^*_τ)**.** *Pour tout $\tau \in \{0\,..\,H-1\}$, V^*_τ est (i) concave en $\sigma^{m,1}_\tau$ pour $\sigma^{c,1}_\tau$ fixé, et (ii) convexe en $\sigma^{m,2}_\tau$ pour $\sigma^{c,2}_\tau$ fixé. Plus précisément,*

$$V^*_\tau(\sigma_\tau) = \min_{\beta^2_{\tau:}} \left[\sigma^{m,1}_\tau \cdot \nu_{[\sigma^{c,1}_\tau, \beta^2_{\tau:}]}\right], \tag{11}$$

$$= \max_{\beta^1_{\tau:}} \left[\sigma^{m,2}_\tau \cdot \nu_{[\sigma^{c,2}_\tau, \beta^1_{\tau:}]}\right]. \tag{12}$$

Toutefois, cette propriété seule ne permet d'approximer la valeur optimale que pour des états d'occupation conditionnels $\sigma^{c,i}_\tau$ fixes (donc en nombre fini), en utilisant pour chacun un ensemble fini de vecteurs $\nu_{[\sigma^{c,i}_\tau, \beta^{-i}_{\tau:}]}$, mais pas pour l'espace d'occupation complet.

3.3 Introduction des jeux locaux

Pour tous τ et σ_τ, définissons (i) $\boldsymbol{\beta}^*_{\tau:}(\sigma_\tau)$ un profil à l'équilibre de Nash pour le *sous*-jeu à σ_τ et (ii) le jeu *local* à σ_τ

$$Q^*_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau) = r(\sigma_\tau, \boldsymbol{\beta}_\tau) + \gamma V^*_{\tau+1}(T(\sigma_\tau, \boldsymbol{\beta}_\tau)), \tag{13}$$

(qui suppose une fonction $V^*_{\tau+1}(\cdot)$ connue). T étant linéaire en β^1_τ et β^2_τ, le théorème 1 conduit au résultat suivant.

Lemma 2. *$Q^*_\tau(\sigma_\tau, \boldsymbol{\beta}_\tau)$ est concave en β^1_τ et convexe en β^2_τ.*

Ce lemme 2 permet d'appliquer le théorème du minimax de VON NEUMANN [19] dans ces jeux locaux. En conséquence, et alors que $Q^*_\tau(\sigma_\tau, \cdot, \cdot)$ peut ne pas être bilinéaire, (i) tous les profils de stratégies à l'équilibre de Nash ont même valeur et sont interchangeables, et (ii) une stratégie à l'équilibre de Nash peut être trouvée par optimisation bi-niveau. On peut donc travailler avec des jeux locaux au lieu des sous-jeux, construisant une solution du zs-OMG par concaténation de solutions de jeux locaux successifs.

4 Propriétés et approximation de V^*

Cette section introduit des approximateurs majorant et minorant de la fonction de valeur exploitant des propriétés de concavité et convexité (CC) et de Lipschitz-continuité (LC).

4.1 Lipschitz-continuité de V^*

L'établissement de la Lipschitz-continuité de V^* part de propriétés de T.

Lemma 3. *À profondeur τ, $T(\sigma_\tau, \boldsymbol{\beta}_\tau)$ est linéaire en β^1_τ, β^2_τ, et σ_τ, où $\boldsymbol{\beta}_\tau = \langle \beta^1_\tau, \beta^2_\tau \rangle$. Elle est plus précisément 1-Lipschitz-continue (1-LC) en σ_τ (en norme 1), c'est-à-dire que, pour tous σ_τ, σ'_τ :*

$$\|T(\sigma'_\tau, \boldsymbol{\beta}_\tau) - T(\sigma_\tau, \boldsymbol{\beta}_\tau)\|_1 \leq 1 \cdot \|\sigma'_\tau - \sigma_\tau\|_1. \tag{14}$$

Aussi, la récompense espérée en tout τ est linéaire en σ_τ (voir anx. A.1), tout comme l'est aussi la valeur espérée d'un profil de stratégies partielles partant de τ.

Lemma 4. *À profondeur τ, $V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:})$ est linéaire en σ_τ.*

Cela conduit à la Lipschitz-continuité de V^*_τ en σ_τ.

Theorem 2. *Soit $h_\tau \stackrel{\text{def}}{=} \frac{1-\gamma^{H-\tau}}{1-\gamma}$ si $\gamma < 1$, sinon $h_\tau \stackrel{\text{def}}{=} H - \tau$ (si $\gamma = 1$). Alors $V^*_\tau(\sigma_\tau)$ est λ_τ-Lipschitz-continue en σ_τ à toute profondeur $\tau \in \{0\,..\,H-1\}$, où $\lambda_\tau = \frac{1}{2} h_\tau (r_{\max} - r_{\min})$.*

4.2 Approximateurs

Un algorithme de résolution peut reposer sur la seule Lipschitz-continuité, comme présenté par BUFFET et coll. [6]. Celui-ci souffre toutefois non seulement d'approximateurs faibles, mais aussi d'employer une optimisation bi-niveau globale dans l'espace (à grande dimension) des règles de décision.
L'exploitation des propriétés de concavité et de convexité de V^*_τ permet de dériver les approximateurs majorant et minorant suivants :

$$\overline{V}_\tau(\sigma_\tau) = \min_{\substack{\langle \tilde{\sigma}^{c,1}_\tau, \nu^2_\tau \rangle \\ \in \overline{bag}V}} \left[\sigma^{m,1}_\tau \cdot \nu^2_\tau + \lambda \|\sigma_\tau - \sigma^{m,1}_\tau \tilde{\sigma}^{c,1}_\tau\|_1\right], \tag{15}$$

$$\underline{V}_\tau(\sigma_\tau) = \max_{\substack{\langle \tilde{\sigma}^{c,2}_\tau, \nu^1_\tau \rangle \\ \in \underline{bag}V}} \left[\sigma^{m,2}_\tau \cdot \nu^1_\tau - \lambda \|\sigma_\tau - \sigma^{m,2}_\tau \tilde{\sigma}^{c,2}_\tau\|_1\right]. \tag{16}$$

Ne disposant pas d'une façon efficace de résoudre, à l'aide de ces approximateurs, les jeux locaux majorant et minorant à σ_τ induits par l'éq. (13), nous introduisons les deux fonctions de valeurs intermédiaires suivantes (entre Q^* et V^*) dont les approximations vont faciliter la résolution des jeux locaux :

$$W^{1,*}_\tau(\sigma_\tau, \beta^1_\tau) \stackrel{\text{def}}{=} \min_{\beta^2_\tau} Q^*_\tau(\sigma_\tau, \beta^1_\tau, \beta^2_\tau), \tag{17}$$

$$W^{2,*}_\tau(\sigma_\tau, \beta^2_\tau) \stackrel{\text{def}}{=} \max_{\beta^1_\tau} Q^*_\tau(\sigma_\tau, \beta^1_\tau, \beta^2_\tau). \tag{18}$$

Ensuite, $V^*_{\tau+1}$ étant Lipschitz-continue en $\sigma_{\tau+1}$ (théorème 2), et en exploitant les propriétés de linéarité et d'indépendance de $T^1_m(\sigma_\tau, \boldsymbol{\beta}_\tau)$ et $T^1_c(\sigma_\tau, \boldsymbol{\beta}_\tau)$ (lemmes 7+8, anx. B.2), nous pouvons dériver un approximateur *majorant* $\overline{W}^1_\tau$ de $W^{1,*}_\tau$ (et réciproquement un approximateur *minorant* $\underline{W}^2_\tau$ de $W^{2,*}_\tau$) en utilisant un nombre fini de

tuples $\langle \tilde{\sigma}_\tau, \beta^2_\tau, \nu^2_{\tau+1} \rangle$ stockés dans un ensemble $\overline{bagW}_\tau$ (*cf.* anx. B.2) :

$$\overline{W}^1_\tau(\sigma_\tau, \beta^1_\tau) = \min_{\substack{\langle \tilde{\sigma}_\tau, \beta^2_\tau, \nu^2_{\tau+1} \rangle \\ \in \overline{bagW}_\tau}} \Big(\beta^{1\intercal}_\tau \cdot \big[r(\sigma_\tau, \cdot, \beta^2_\tau) \tag{19}$$

$$+ \gamma T^1_m(\tilde{\sigma}_\tau, \cdot, \beta^2_\tau) \cdot \nu^2_{\tau+1} \big] + \gamma \lambda_\tau \cdot \|\sigma_\tau - \tilde{\sigma}_\tau\|_1 \Big) . \tag{20}$$

Comme détaillé dans **??**, étant donnée une distribution δ^2_τ sur des tuples $w = \langle \tilde{\sigma}_\tau, \beta^2_\tau, \nu^2_{\tau+1} \rangle$ de $\overline{W}_\tau$, nous pouvons maintenant majorer la valeur du «profil» $\langle \beta^1_\tau, \delta^2_\tau \rangle$ dans σ_τ comme

$$\sum_{w \in \overline{W}_\tau} \beta^{1\intercal}_\tau \cdot \Big[r(\sigma_\tau, \cdot, \beta^2_\tau[w]) \tag{21}$$

$$+ \gamma T^1_m(\tilde{\sigma}_\tau[w], \cdot, \beta^2_\tau[w]) \cdot \nu^2_{\tau+1}[w] \tag{22}$$

$$+ \gamma \lambda_\tau \cdot \sigma^{m,1}_\tau(\Theta^1_\tau \times \mathcal{A}^1) \cdot \|\sigma_\tau - \tilde{\sigma}_\tau[w]\|_1 \Big] \delta^2_\tau(w) \tag{23}$$

(où $x[w]$ dénote le champ x du tuple w, et $\sigma^{m,1}_\tau(\Theta^1_\tau \times \mathcal{A}^1)$ est un vecteur indexé par des couples (θ^1_τ, a^1), et dont le composant $\sigma^{m,1}_\tau(\theta^1_\tau, a^1)$ a comme valeur $\sigma^{m,1}_\tau(\theta^1_\tau)$)

$$= \beta^{1\intercal}_\tau \cdot M^{\sigma_\tau} \cdot \delta^2_\tau, \tag{24}$$

avec M^{σ_τ} une matrice de taille $|\Theta^1_\tau \times \mathcal{A}^1| \times |\overline{bagW}^1_\tau|$, où

$$M^{\sigma_\tau}_{(\langle \theta^1_\tau, a^1 \rangle, w)} = \sum_{s, \theta^2_\tau, a^2} \sigma_\tau(\boldsymbol{\theta}_\tau) b(s|\boldsymbol{\theta}_\tau) \beta^2_\tau[w](a^2|\theta^2) r(s, \boldsymbol{a}) \tag{25}$$

$$+ \sum_{z^1} \Big[\sum_{\theta^2_\tau, a^2} \beta^2_\tau[w](a^2|\theta^2_\tau) \sum_{s, s', z^2} P^{\boldsymbol{z}}_{\boldsymbol{a}}(s'|s) b(s|\boldsymbol{\theta}_\tau) \tag{26}$$

$$\cdot \tilde{\sigma}_\tau[w](\boldsymbol{\theta}_\tau) \Big] \cdot \nu^2_{\tau+1}[w](\theta^1_\tau, a^1, z^1) \tag{27}$$

$$+ \gamma \lambda_{(\tau+1)} \cdot \sigma^{m,1}_\tau(\theta^1_\tau) \cdot \|\sigma_\tau - \tilde{\sigma}_\tau[w]\|_1. \tag{28}$$

Ainsi, la résolution de $\max_{\beta^1_\tau} \overline{W}^1_\tau(\sigma_\tau, \beta^1_\tau)$ peut être réécrite comme la résolution d'un jeu de matrice à somme nulle dont les stratégies pures sont : pour 1, le choix de $|\Theta^1_\tau|$ actions et, pour 2, le choix d'1 élément de $\overline{bagW}^1_\tau$. Le LP correspondant est :

$$\max_{\beta^1_\tau, v} v \text{ s.t. } \quad \forall w \in \overline{bagW}_\tau, \quad v \leq \beta^{1\intercal}_\tau \cdot M^{\sigma_\tau}_{(\cdot, w)} \tag{29}$$

$$\forall \theta^1_\tau \in \Theta^1_\tau, \ \textstyle\sum_{a^1} \beta^1_\tau(a^1|\theta^1_\tau) = 1, \tag{30}$$

et le LP dual :

$$\min_{\delta^2_\tau, v} v \text{ s.t. } \quad \forall (\theta^1_\tau, a^1), \quad v \leq M^{\sigma_\tau}_{((\theta^1_\tau, a^1), \cdot)} \cdot \delta^2_\tau \tag{31}$$

$$\textstyle\sum_{w \in \overline{bagW}_\tau} \delta^2_\tau(w) = 1. \tag{32}$$

Nous regardons maintenant les opérateurs employés pour manipuler ces approximateurs.

4.3 Opérateurs associés

Pour $\tau \geq 1$, $\overline{V}_\tau$ et $\overline{W}^1_{\tau-1}$ reposent tous deux essentiellement sur la même information et sont fortement reliés, de sorte que nous les discuterons conjointement. $\overline{bagV}_\tau$ contient des tuples $\langle \sigma^{c,1}_\tau, \langle \nu^2_\tau, \delta^2_\tau \rangle \rangle$, et $\overline{bagW}^1_{\tau-1}$ (pour $\tau \geq 1$) les tuples associés $\langle \sigma_{\tau-1}, \beta^2_{\tau-1}, \langle \nu^2_\tau, \delta^2_\tau \rangle \rangle$. Ici, δ^2_τ, en tant que distribution sur les tuples de $\overline{bagW}^1_\tau$, induit une stratégie définie récursivement pour 2 comme (i) une mixture de règles de décisions comportementales à τ :

$$\beta^2_\tau[\delta^2_\tau] \stackrel{\text{def}}{=} \sum_{\tilde{\beta}^2_\tau \in \overline{bagW}^1_\tau} \delta^2_\tau(\tilde{\beta}^2_\tau) \cdot \tilde{\beta}^2_\tau, \tag{33}$$

et (ii) une mixture d'autres stratégies mélangées pour les pas de temps $\tau + 1$ et suivants :

$$\delta^2_{\tau+1}[\delta^2_\tau] \stackrel{\text{def}}{=} \sum_{\tilde{\delta}^2_{\tau+1} \in \overline{bagW}^1_\tau} \delta^2_\tau(\tilde{\delta}^2_{\tau+1}) \cdot \tilde{\delta}^2_{\tau+1}, \tag{34}$$

jusqu'à atteindre l'horizon. Note : Par commodité, nous pourrons remplacer un tuple complet par les quelques éléments d'intérêt.

Initialisations On peut chercher un majorant de V^*, c'est-à-dire une borne optimiste (une heuristique admissible) pour le joueur (maximisant) 1, en relâchant le problème auquel 1 fait face. Dans ce but, nous résolvons ici le POMDP induit quand est affecté à 2 un $\beta^{2,\oslash}_{0:}$ uniformément aléatoire, la meilleure réponse résultante étant notée $\beta^{1,\otimes}_{0:}$. À profondeur τ, $\beta^{2,\oslash}_{0:}$ et $\beta^{1,\otimes}_{0:}$ induisent (i) un état d'occupation OS σ_τ et (ii) un vecteur $\overline{\nu}_\tau$, où $\overline{\nu}_\tau(\theta^1_\tau)$ est la valeur de $\beta^{1,\otimes}_{0:}$ en θ^1_τ (contre $\beta^{2,\oslash}_{0:}$ et sous $\sigma^{c,1}_\tau$).
Étant données ces stratégies, chaque $\overline{bagV}_\tau$ (respectivement $\overline{bagW}^1_{\tau-1}$) est initialisé avec $\langle \sigma^{c,1}_\tau, \langle \overline{\nu}_\tau, \delta^{2,\oslash}_\tau \rangle \rangle$ (resp. $\langle \sigma_{\tau-1}, \beta^{2,\oslash}_{\tau-1}, \langle \overline{\nu}_\tau, \delta^{2,\oslash}_\tau \rangle \rangle$), où $\delta^{2,\oslash}_\tau$ est une distribution dégénérée sur le seul élément dans $\overline{bagW}^1_\tau$.
Des initialisations plus avancées pourraient être proposées en améliorant les stratégies de 2, en utilisant une relaxation One-Sided zsPOSG à la place, ou en combinant plusieurs initialisations.

Mise-à-jour de $\overline{V}_\tau$ et $\overline{W}^1_{\tau-1}$ Pour mettre à jour $\overline{bagV}_\tau$ et (si $\tau \geq 1$) $\overline{bagW}^1_{\tau-1}$ simultanément, nous considérons un tuple $\langle \sigma_\tau, \sigma^{c,1}_{\tau-1}, \beta^2_{\tau-1} \rangle$ (partiellement indéfini si $\tau = 0$), et obtenons un couple approprié $\langle \nu^2_\tau, \overline{\delta}^2_\tau \rangle$ en résolvant le LP dual (31), noté $\arg\min_{\delta^2_\tau} \overline{W}^1_\tau(\sigma_\tau, \delta^2_\tau)$, donc reposant sur $\overline{bagV}_{\tau+1}$.
Pour un $\overline{\delta}^2_\tau$ donné, $\nu^2_\tau(\theta^1_\tau, \overline{\delta}^2_\tau)$ est la valeur de la meilleure action a^1 de 1 en supposant que (i) 2 suit la stratégie $\overline{\delta}^2_\tau$ et (ii) le retour espéré à partir de $\tau + 1$ est donné par $\overline{V}_{\tau+1}$ ($= 0$ si $\tau + 1 = H$) :

$$\nu^2_\tau(\theta^1_\tau, \overline{\delta}^2_\tau) = \frac{1}{\sigma^1_{\tau,m}(\theta^1_\tau)} \max_{a^1 \in \mathcal{A}^1} M^{\sigma_\tau}_{((\theta^1_\tau, a^1), .)} \cdot \overline{\delta}^2_\tau. \tag{35}$$

On a alors besoin d'ajouter $\langle \sigma_\tau^{c,1}, \langle \nu_\tau^2(\underline{\delta}_\tau^2), \underline{\delta}_\tau^2 \rangle \rangle$ to $\overline{bagV}_\tau$, et (si $\tau \geq 1$) $\langle \sigma_{\tau-1}, \beta_{\tau-1}^2 \langle \nu_\tau^2(\underline{\delta}_\tau^2), \underline{\delta}_\tau^2 \rangle, \rangle$ à $\overline{bagW}^1{}_{\tau-1}$.
Ce processus est résumé en ligne 14 sqq. de l'algorithme 1.

Élagage Parce qu'elles ont des formes différentes, $\overline{V}_\tau$ et $\overline{bagW}^1{}_{\tau-1}$ doivent être élaguées indépendamment. Aussi, l'élagage ne devrait pas casser les stratégies définies récursivement δ_τ^2. Les tuples élagués devraient donc rester stockés en mémoire (si utilisés).
$\overline{V}_\tau$ repose sur une représentation «min-surfaces» (plutôt que «min-planes»), où chaque surface est linéaire en $\sigma_\tau^{m,1}$ et exploite la Lipschitz-continuité. Cela permet d'exploiter (en les inversant) les méthodes d'élagage max-planes pour POMDP comme expliqué dans le résultat suivant.

Theorem 3. *Soit P un opérateur d'élagage* min*-planes (inverse d'un élagage* max*-planes pour POMDP), et $\langle \nu_\tau, \sigma_\tau^{c,1} \rangle \in \overline{bagV}_\tau$. Si P identifie correctement ν_τ comme non-dominé (ou respectivement dominé) sous $\sigma_\tau^{c,1}$ fixe, alors $\langle \nu_\tau, \sigma_\tau^{c,1} \rangle$ est non-dominé (ou resp. dominé) dans $\mathcal{O}_\tau^\sigma$.*

Le fait qu'un test induise des faux positifs (élaguant des éléments non-dominés) ou des faux négatifs (n'élaguant pas des éléments dominés) se propage donc du cadre min-planes au cadre min-surfaces.

Pour sa part, $\overline{W}_\tau^1$ impliquant un terme de récompense qui est bilinéaire (linéaire à la fois en σ_τ et β_τ^1), dériver des techniques d'élagage n'est pas si direct. Cela peut requérir d'utiliser de l'optimisation quadratique, **ou** de majorer le terme de récompense bilinéaire avec un linéaire pour retomber sur le cadre «min-surface» précédent. Alors qu'on s'attend à ce que la résolution de jeux locaux bénéficie de manière significative de l'élagage de $\overline{W}_\tau^1$, nous laissons ce sujet pour des travaux futurs.

5 HSVI pour zs-POSG

Dans cette section, nous cherchons des solutions ϵ-optimales.

5.1 Algorithme

HSVI pour zs-OMG est décrit dans l'algorithme 1. Comme le HSVI de base, il repose sur (i) la génération de trajectoires en agissant avec optimisme (lignes 9+10), c'est-à-dire que le joueur 1 (resp. 2) agit de manière «gourmande» par rapport à $\overline{W}_\tau$ (resp. $\underline{W}_\tau$), et (ii) la mise-à-jour locale des approximateurs majorant et minorant (lignes 12+13). Ici, les calculs des mises-à-jour des valeurs et des stratégies repose sur la résolution de jeux en forme normale décrits par le LP (29). Notons que l'implémentation maintient des états d'occupation *complets* $o_\tau \in \Delta(\mathcal{S} \times \mathbf{\Theta}_\tau)$, ce qui permet de facilement retrouver à la fois les états d'occupation «simples» $\sigma_\tau \in \mathcal{O}_\tau^\sigma = \Delta(\mathbf{\Theta}_\tau)$ et les «croyances» $b(s|\boldsymbol{\theta}_\tau)$. Une différence clef avec l'algorithme de SMITH et SIMMONS réside dans le critère d'arrêt des trajectoires. Dans l'algorithme HSVI de base (pour POMDP), le facteur de branchement fini permet de regarder la convergence de $\overline{V}$ et $\underline{V}$ en chaque point atteignable sous une stratégie optimale. Pour garantir l'ϵ-convergence en σ_0, les trajectoires doivent juste être interrompues quand la largeur courante en σ_τ ($\stackrel{\text{def}}{=} \overline{V}_\tau(\sigma_\tau) - \underline{V}_\tau(\sigma_\tau)$) est plus petite qu'un seuil $\gamma^{-\tau}\epsilon$. (Cela arrive même si $\gamma = 1$ parce que la largeur de l'approximateur est nulle au-delà de H.) Ici, ayant à faire à un facteur de branchement infini, on peut converger vers une solution optimale tout en visitant toujours de nouveaux points de l'espace des états d'occupation. Pour contrer cela, nous bornons la largeur à l'intérieur de boules autour des points visités en exploitant la Lipschitz-continuité de V^*. Ceci est accompli en ajoutant un terme $-\sum_{i=1}^{\tau} 2\rho\lambda_{\tau-i}\gamma^{-i}$ [13] (même si $\gamma = 1$) pour garantir que la largeur est inférieure à $\gamma^\tau\epsilon$ à l'intérieur d'une boule de rayon ρ autour du point courant (ici σ_τ). D'où le seuil

$$thr(\tau) \stackrel{\text{def}}{=} \gamma^{-\tau}\epsilon - \sum_{i=1}^{\tau} 2\rho\lambda_{\tau-i}\gamma^{-i}. \tag{36}$$

Algorithme 1 : zs-OMG-HSVI($b_0, [\epsilon, \rho]$)
[here returning solution strategy $\underline{\delta}_0^1$ for Player 1]

1 **Fct** `zs-OMG-HSVI` $(b_0 \simeq \sigma_0)$
2 Initialize $\overline{V}_{\dots}$ & $\underline{V}_{\dots}$
3 **while** $\left[\overline{V}_0(\sigma_0) - \underline{V}_0(\sigma_0) > thr(0)\right]$ **do**
4 `Explore` $(\sigma_0, 0, -, -)$
5 $\underline{\delta}_0^1 \leftarrow \underset{\langle \tilde{\sigma}_0 = \sigma_0, \langle \nu_0^1, \underline{\delta}_0^1 \rangle, - \rangle \in \underline{bagV}_0}{\arg\max} \left(\sigma_0^{m,2} \cdot \nu_0^1 + \lambda_\tau \cdot 0 \right)$
6 **return** $\underline{\delta}_0^1$
7 **Fct** `Explore` $(\sigma_\tau, \tau, \sigma_{\tau-1}, \boldsymbol{\beta}_{\tau-1})$
8 **if** $\left[\overline{V}_\tau(\sigma) - \underline{V}_\tau(\sigma) > thr(\tau)\right]$ **then**
9 $\overline{\beta}_\tau^1 \leftarrow \arg\max_{\beta_\tau^1} \overline{W}_\tau^1(\sigma, \beta_\tau^1)$
10 $\underline{\beta}_\tau^2 \leftarrow \arg\min_{\beta_\tau^2} \underline{W}_\tau^2(\sigma, \beta_\tau^2)$
11 `Explore` $(T(\sigma_\tau, \overline{\beta}_\tau^1, \underline{\beta}_\tau^2), \tau{+}1, \sigma_\tau, \langle \overline{\beta}_\tau^1, \underline{\beta}_\tau^2 \rangle)$
12 $\overline{\textbf{Update}}(\overline{V}_\tau, \overline{W}^1{}_{\tau-1}, \langle \sigma_\tau, \sigma_{\tau-1}^{c,1}, \underline{\beta}_{\tau-1}^2 \rangle)$
13 $\underline{\textbf{Update}}(\underline{V}_\tau, \underline{W}^2{}_{\tau-1}, \langle \sigma_\tau, \sigma_{\tau-1}^{c,2}, \overline{\beta}_{\tau-1}^1 \rangle)$
14 **Fct** $\overline{\textbf{Update}}$ $(\overline{V}_\tau, \overline{W}^1{}_{\tau-1}, \langle \sigma_\tau, \sigma_{\tau-1}^{c,1}, \beta_{\tau-1}^2 \rangle)$
15 $\overline{\delta}_\tau^2 \leftarrow \arg\min_{\delta_\tau^2} \overline{W}_\tau^1(\sigma_\tau, \delta_\tau^2)$
16 $\nu_\tau^2 \leftarrow \nu^2(\overline{\delta}_\tau^2)$
17 $\overline{bagV}_\tau \leftarrow \overline{bagV}_\tau \cup \{\langle \sigma_\tau^{c,1}, \langle \nu_\tau^2, \overline{\delta}_\tau^2 \rangle \rangle\}$
18 $\overline{bagW}^1{}_{\tau-1} \leftarrow \overline{bagW}^1{}_{\tau-1} \cup \{\langle \sigma_{\tau-1}, \beta_{\tau-1}^2, \langle \nu_\tau^2, \overline{\delta}_\tau^2 \rangle \rangle\}$

Régler ρ Comme on peut l'observer, cette fonction seuil devrait toujours retourner des valeurs positives, ce qui requiert un ρ suffisamment petit. Pour un problème donné, la valeur maximum possible de ρ dépend des constantes de Lipschitz à chaque pas de temps, lesquelles dépendent

elles-mêmes des majorant et minorant initiaux de la fonction de valeur optimale.

Lemma 5. *En bornant λ_τ par $\lambda^\infty = \frac{1}{2}\frac{1}{1-\gamma}[r_{\max} - r_{\min}]$ quand $\gamma < 1$, et en notant que*

$$thr(\tau) = \begin{cases} \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty \frac{\gamma^{-\tau}-1}{1-\gamma} \text{ if } \gamma < 1, \text{ else} \\ \epsilon - \rho(r_{\max} - r_{\min})(2H + 1 - \tau)\tau, \end{cases} \quad (37)$$

on peut garantir la positivité du seuil en tout $\tau \in 1..H-1$ en contraignant

$$0 < \rho < \begin{cases} \frac{1-\gamma}{2\lambda^\infty}\epsilon & \text{if } \gamma < 1, \\ \frac{\epsilon}{(r_{\max}-r_{\min})(H+1)H} & \text{if } \gamma = 1. \end{cases} \quad (38)$$

Mais quel est l'effet de régler ρ à de petites ou grandes valeurs ?

— Plus ρ est petit, plus $thr(\tau)$ est grand, plus courtes sont les trajectoires, mais plus petites sont les boules et plus grand devient la densité de points nécessaire autour de la trajectoire optimale, ce qui implique un plus grand nombre de trajectoires requises pour mener à la convergence.

— Plus ρ est grand, plus $thr(\tau)$ est petit, plus longues sont les trajectoires, mais plus grandes sont les boules et moins importante devient la densité de points nécessaire autour de la trajectoire optimale, ce qui implique un plus petit nombre de trajectoires pour mener à la convergence.

5.2 Convergence en temps fini

Theorem 4. *zs-OMG-HSVI (algorithme 1) termine en temps fini avec une ϵ-approximation de $V_0^*(\sigma_0)$.*

Démonstration. (ébauche adaptée de HORÁK et BOŠANSKÝ [12]) Supposons que l'algorithme ne termine pas et génère un nombre infini d'essais (trajectoires) exploratoires. Alors, le nombre d'essais de longueur T (pour un certain $0 \leq T \leq H$) doit être infini. Il est impossible de faire tenir un nombre infini de points d'occupation σ_T satisfaisant $\|\sigma_T - \sigma'_T\|_1 > \rho$ à l'intérieur de $\mathcal{O}_T^\sigma$. Il doit donc y avoir deux essais de longueur T, $\{\sigma_{\tau,1}\}_{\tau=0}^T$ et $\{\sigma_{\tau,2}\}_{\tau=0}^T$, tels que $\|\sigma_{T-1,1} - \sigma_{T-1,2}\|_1 \leq \rho$. On peut ainsi montrer (comme fait en annexe C.1) que le second essai ne devrait pas avoir eu lieu. □

Note : La borne sur le nombre d'itérations dépend du nombre de boules de rayon ρ requis pour couvrir le simplexe des états d'occupation à chaque profondeur.
Par ailleurs, le lemme suivant permet de résoudre aussi des problèmes à horizon infini (quand $\gamma < 1$) en bornant la longueur des trajectoires utilisant la largeur bornée de $\hat{V}$ et la croissance exponentielle de $thr(\tau)$.

Lemma 6. *Quand $\gamma < 1$, en utilisant la constante de Lipschitz indépendante de la profondeur λ^∞, et avec la largeur maximale entre initialisations $W \stackrel{\text{def}}{=} \|\overline{V}^{(0)} - \underline{V}^{(0)}\|_\infty$, la longueur des trajectoires est majorée par*

$$T_{\max} \stackrel{\text{def}}{=} \left\lceil \log_\gamma \frac{\epsilon - \frac{2\rho\lambda^\infty}{1-\gamma}}{W - \frac{2\rho\lambda^\infty}{1-\gamma}} \right\rceil. \quad (39)$$

5.3 Exécution

Comme déjà mentionné, toute stratégie δ_τ^2 dans un tuple stocké garantit au plus (c'est-à-dire «au pire» du point de vue de 2) un retour espéré $\sigma_\tau^{m,1} \cdot \overline{\nu}_{\tau:}^2$ si dans l'état d'occupation associé σ_τ, quelle que soit la stratégie de 1. En particulier, δ_0^2 induit une stratégie qui garantit au pire le retour espéré $\sigma_0^{m,1} \cdot \overline{\nu}_0^2$ si en σ_0 (qui correspond toujours à l'état d'occupation initial). Au moment de l'exécution, si 2 emploie une stratégie $\delta_0^{2,\star} \in \arg\max_{\langle\sigma_0,\langle\overline{\nu}_0,\delta_0^2\rangle,-,-\rangle} \sigma_0^{m,1}\cdot\overline{\nu}_0$, alors son retour espéré est au plus $\overline{V}_0(\sigma_0)$ ($\leq V_0^*(\sigma_0) + \epsilon$) (quelle que soit la stratégie de 1).
Ainsi, résoudre le zs-OMG dérivé fournit à chaque joueur une stratégie solution du zs-POSG original, et chaque joueur peut dériver sa stratégie par lui-même puisque les stratégies à l'équilibre de Nash sont interchangeables (pas besoin de la coordination d'un planificateur central comme pour les Dec-POMDP). Par exemple, l'algo. 1 renvoie une stratégie solution seulement pour 1.

6 Expérimentations

Cette section présente des expérimentations préliminaires pour valider l'approche proposée avec une première version de zsOMG-HSVI. Des résultats complémentaires apparaissent aussi en annexe D.

6.1 Configuration

Bancs d'essai Quatre bancs d'essai ont été utilisés. Recycling Robot et Mabc sont des bancs d'essai Dec-POMDP bien connus (*cf.* `http://masplan.org`) et ont été adaptés à notre cadre compétitif en faisant minimiser (plutôt que maximiser) la fonction objectif par le joueur 2. Competitive Tiger et Adversarial Tiger ont été introduits par WIGGERS [26]. Nous ne considérons que des horizons finis et $\gamma = 1$.

Algorithmes Nous comparons nos algorithmes, c'est-à-dire la version reposant seulement sur la Lipschitz continuité (dénotée OMGHSVI$^{\text{LC}}$) et la version reposant aussi sur la concavité et la convexité (dénotée OMGHSVI$_{\text{CC}}^{\text{LC}}$), avec l'algorithme de l'état de l'art Sequence Form [15], et deux approches heuristiques, *Informed* et *Random*, proposées par WIGGERS [26] et s'appuyant sur les propriétés de concavité et de convexité de la fonction de valeur.
OMGHSVI$^{\text{LC}}$ a été exécuté avec une erreur cible ϵ spécifiée dans la table, $\lambda_\tau = H \cdot (r_{\max} - r_{\min})$, et sans élagage. OMGHSVI$_{\text{CC}}^{\text{LC}}$ a été exécuté avec une erreur $\epsilon = 0.0001$, $\lambda_\tau = H \cdot (r_{\max} - r_{\min})$, ρ au milieu de son intervalle de faisabilité, et un élagage de $\overline{V}_\tau$ et $\underline{V}_\tau$, mais pas de $\overline{W}_\tau^1$ ou $\underline{W}_\tau^2$. Pour les Dec-POMDP, un ingrédient clef du passage à l'échelle de FB-HSVI est la compression sans perte des historiques d'action-observation équivalents dans les états

d'occupation, ce qui réduit leur dimensionalité [9]. Dans notre contexte, nous appliquons plus précisément la compression LPE à σ_τ, laissant la compression plus forte pour des travaux futurs.

Nous avons lancé les expérimentations sur une machine Ubuntu avec processeur Intel i7-10810U 1.10 GHz et 16 GB de mémoire vive disponible. Le code source sera rendu disponible en source ouverte.

6.2 Résultats

Une première observation est que OMGHSVI$^{\text{LC}}_{\text{CC}}$ et OMGHSVI$^{\text{LC}}$ maintiennent tous deux des majorant et minorant valides de la valeur optimale en σ_0, et réduisent l'écart progressivement[2] (*cf.* figures en anx. D). Le tableau 1 montre que (i) OMGHSVI$^{\text{LC}}_{\text{CC}}$ est toujours meilleur que OMGHSVI$^{\text{LC}}$ et les heuristiques de WIGGERS [26], et (ii) à moins de manquer de mémoire, Sequence Form surpasse toujours OMGHSVI$^{\text{LC}}_{\text{CC}}$. Par contre, la compression LPE permet à OMGHSVI$^{\text{LC}}_{\text{CC}}$ d'exploiter la structure de certains jeux et ainsi de générer des trajectoires même pour de grands horizons (par exemple dans Recycling Robot pour $H = 6$, *cf.* anx. D). Plus généralement, nous observons que le nombre d'itérations effectuées par OMGHSVI$^{\text{LC}}_{\text{CC}}$ en 24 h est fortement corrélé avec la qualité de la compression LPE (la compression est maximale pour Recycling Robot compresses et minimale pour Adversarial Tiger). Comme attendu, OMGHSVI$^{\text{LC}}$ s'avère très lent, ne terminant même pas la première itération dans la plupart des cas.

7 Discussion

Inspiré par des techniques de résolution de l'état de l'art pour POMDP et Dec-POMDP, nous résolvons ici des zs-POSG en le convertissant en des jeux de Markov à somme nulle sur les états d'occupation, c'est-à-dire des jeux complètement observables qui permettent d'exploiter le principe d'optimalité de Bellman. Nous étendons les propriétés de concavité, convexité et Lipschitz-continuité de V^* et Q^*, et en tirons partie pour proposer des approximateurs à base de points majorant et minorant, ainsi que des opérateurs de mise-à-jour efficace reposant sur la programmation linéaire. Cela permet de dériver une variante de HSVI qui converge en temps fini vers une solution ϵ-optimale, fournissant des stratégies solution (sûres) sous une forme récursive comme sous-produit du processus de résolution. Les expérimentations confirment la faisabilité de cette approche et montrent des résultats améliorés par rapport à des heuristiques apparentées (s'appuyant aussi sur les propriétés de concavité et de convexité).

Cette approche ouvre la voie à une vaste famille de solveurs puisque de nombreuses variantes pourraient être envisagées, par exemple, en employant différents schémas algorithmiques, différents approximateurs, différents opérateurs de mise-à-jour ou différentes techniques d'élagage.

2. Notons que les écarts ne sont pas comparables entre bancs d'essai puisqu'ils ne sont pas normalisés.

TABLE 1 – Expérimentations comparant 4 solveurs sur divers bancs d'essai. Les valeurs reportées sont les temps de calcul, ou les [écarts] $[\overline{V}_0(\sigma_0) - \underline{V}_0(\sigma_0)]$ si la limite de 24 h est atteinte. «(ni)» (*no improvement*) indique aucune amélioration par rapport aux initialisations en 24 h . «xx» indique un dépassement mémoire. «n/a» indique un résultat non disponible.

Competitive Tiger	H=2	H=3	H=4	H=5
Wiggers Random	[0.56]	[2.67]	[5.81]	[6.97]
Wiggers Informed	[2.07]	[2.33]	[3.61]	xx
OMGHSVI$^{\text{LC}}(1.0)$	[2.8]	(ni)	(ni)	(ni)
OMGHSVI$^{\text{LC}}_{\text{CC}}$	2s	[0.02]	[2.55]	[5.82]
Sequence Form	0.14 s	48 s	14 min	2.5 h
Adversarial Tiger	**H=2**	**H=3**	**H=4**	**H=5**
Wiggers Random	[0.04]	[0.38]	[0.92]	[2.07]
Wiggers Informed	[0.59]	[1.32]	[1.79]	[3.34]
OMGHSVI$^{\text{LC}}(0.1)$	22 min	(ni)	(ni)	(ni)
OMGHSVI$^{\text{LC}}_{\text{CC}}$	1 s	10 s	[0.4]	[1.70]
Sequence Form	0.02 s	0.17 s	3 s	107 s
Recycling Robot	**H=3**	**H=4**	**H=5**	**H=6**
OMGHSVI$^{\text{LC}}$	(ni)	(ni)	(ni)	(ni)
OMGHSVI$^{\text{LC}}_{\text{CC}}$	3 min	[0.01]	[0.93]	[2.67]
Sequence Form	1 s	10 s	1.5 h	xx
Mabc	**H=2**	**H=3**	**H=4**	**H=5**
OMGHSVI$^{\text{LC}}(0.05)$	2s	(ni)	(ni)	(ni)
OMGHSVI$^{\text{LC}}_{\text{CC}}$	0.6 s	17 min	[0.11]	n/a
Sequence Form	0.1 s	1 s	3 s	181 s

Nous avons ainsi aussi évalué une variante ne reposant que sur la Lipschitz-continuité de V^*.

Les travaux futurs incluent : chercher de meilleures initialisations et constantes de Lipschitz, par exemple avec des initialisations POMDP plus avancées ou en s'appuyant sur les One-Sided zsPOSG ; proposer une méthode d'élagage pour $\overline{W}^1_\tau$ et $\underline{W}^2_\tau$; employer des méthodes d'oracle ou d'autres heuristiques pour résoudre des jeux locaux plus vite ; exploiter une compression TPE plutôt que LPE ; et brancher sur des observations publiques (voire sur des informations publiques révélées par la structure de l'état d'occupation).

Références

[1] K. ÅSTRÖM. « Optimal control of Markov processes with incomplete state information ». *Journal of Mathematical Analysis and Applications* 10.1 (1965), p. 174-205. ISSN : 0022-247X.

[2] N. BASILICO, G. DE NITTIS et N. GATTI. « A Security Game Combining Patrolling and Alarm–Triggered Responses Under Spatial and Detection Uncertainties ». Dans : *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[3] A. BASU et L. STETTNER. « Finite- and Infinite-Horizon Shapley Games with Nonsymmetric Partial Observation ». *SIAM Journal on Control and Optimization* 53.6 (2015), p. 3584-3619.

[4] B. BOŠANSKÝ, C. KIEKINTVELD, V. LISÝ et M. PĚCHOUČEK. « An Exact Double-Oracle Algorithm for Zero-Sum Extensive-Form Games with Imperfect Information ». *Journal of Artificial Intelligence Research* 51 (2014), p. 829-866. DOI : 10.1613/jair.4477.

[5] N. BROWN et T. SANDHOLM. « Superhuman AI for Heads-Up No-Limit Poker : Libratus Beats Top Professionals ». *Science* 359.6374 (2018), p. 418-424.

[6] O. BUFFET, J. DIBANGOYE, A. DELAGE, A. SAFFIDINE et V. THOMAS. « On Bellman's Optimality Principle for zs-POSGs ». *Computing Research Repository* abs/2006.16395 (2020). arXiv : 2006.16395 [cs.AI].

[7] K. CHATTERJEE et L. DOYEN. « Partial-Observation Stochastic Games : How to Win When Belief Fails ». *ACM Transactions on Computational Logic* 15.2 (2014), p. 16.

[8] H. L. COLE et N. KOCHERLAKOTA. « Dynamic Games with Hidden actions and Hidden States ». *Journal of Economic Theory* 98.1 (2001), p. 114-126.

[9] J. DIBANGOYE, C. AMATO, O. BUFFET et F. CHARPILLET. « Optimally Solving Dec-POMDPs as Continuous-State MDPs ». *Journal of Artificial Intelligence Research* 55 (2016), p. 443-497.

[10] M. K. GHOSH, D. R. MCDONALD et S. SINHA. « Zero-Sum Stochastic Games with Partial Information ». *Journal of Optimization Theory and Applications* 121.1 (avr. 2004), p. 99-118.

[11] E. A. HANSEN, D. BERNSTEIN et S. ZILBERSTEIN. « Dynamic Programming for Partially Observable Stochastic Games ». Dans : *Proceedings of the Nineteenth National Conference on Artificial Intelligence*. San Jose, CA, 2004.

[12] K. HORÁK et B. BOŠANSKÝ. « Solving Partially Observable Stochastic Games with Public Observations ». Dans : *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. 2019, p. 2029-2036.

[13] K. HORÁK, B. BOŠANSKÝ et M. PĚCHOUČEK. « Heuristic Search Value Iteration for One-Sided Partially Observable Stochastic Games ». Dans : *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. 2017, p. 558-564.

[14] D. KOLLER et N. MEGIDDO. « The Complexity of Two-Person Zero-Sum Games in Extensive Form ». *Games and Economic Behavior* 4.4 (1992), p. 528-552. DOI : https://doi.org/10.1016/0899-8256(92)90035-Q.

[15] D. KOLLER, N. MEGIDDO et B. von STENGEL. « Efficient Computation of Equilibria for Extensive Two-Person Games ». *Games and Economic Behavior* 14.51 (1996), p. 220-246.

[16] C. KROER, K. WAUGH, F. KILINÇ-KARZAN et T. SANDHOLM. « Faster Algorithms for Extensive-Form Game Solving via Improved Smoothing Functions ». *Mathematical Programming* 179 (2020), p. 385-417. DOI : 10.1007/s10107-018-1336-7.

[17] H. W. KUHN. « Simplified Two-Person Poker ». Dans : *Contributions to the Theory of Games*. Sous la dir. de H. W. KUHN et A. W. TUCKER. T. 1. Princeton University Press, 1950.

[18] O. MADANI, S. HANKS et A. CONDON. « On the Undecidability of Probabilistic Planning and Infinite-Horizon Partially Observable Markov Decision Problems ». Dans : *Proceedings of the Sixteenth National Conference on Artificial Intelligence*. 1999.

[19] J. von NEUMANN. « Zur Theorie der Gesellschaftsspiele ». *Mathematische Annalen* 100 (1928).

[20] F. OLIEHOEK et N. VLASSIS. *Dec-POMDPs and extensive form games : equivalence of models and algorithms*. Rapp. tech. IAS-UVA-06-02. Intelligent Systems Laboratory Amsterdam, University of Amsterdam, 2006.

[21] Y. SHOHAM et K. LEYTON-BROWN. *Multiagent Systems : Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009. ISBN : 978-0-521-89943-7.

[22] T. SMITH. « Probabilistic Planning for Robotic Exploration ». Thèse de doct. The Robotics Institute, Carnegie Mellon University, 2007.

[23] T. SMITH et R. SIMMONS. « Point-Based POMDP Algorithms : Improved Analysis and Implementation ». Dans : *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. 2005, p. 542-549.

[24] B. von STENGEL. « Efficient Computation of Behavior Strategies ». *Games and Economic Behavior* 14.50 (1996), p. 220-246.

[25] D. SZER, F. CHARPILLET et S. ZILBERSTEIN. « MAA* : A Heuristic Search Algorithm for Solving Decentralized POMDPs ». Dans : *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. 2005, p. 576-583.

[26] A. WIGGERS. « Structure in the Value Function of Two-Player Zero-Sum Games of Incomplete Information ». Mém. de mast. University of Amsterdam, 2015.

[27] A. WIGGERS, F. OLIEHOEK et D. ROIJERS. « Structure in the Value Function of Two-Player Zero-Sum Games of Incomplete Information ». *Computing Research Repository* abs/1606.06888 (2016).

[28] A. WIGGERS, F. OLIEHOEK et D. ROIJERS. « Structure in the Value Function of Two-player Zero-sum Games of Incomplete Information ». Dans : *Proceedings of the Twenty-second European Conference on Artificial Intelligence*. The Hague, The Netherlands, 2016, p. 1628-1629. DOI : 10.3233/978-1-61499-672-9-1628.

[29] M. ZINKEVICH, M. JOHANSON, M. BOWLING et C. PICCIONE. « Regret Minimization in Games with Incomplete Information ». Dans : *Advances in Neural Information Processing Systems 20*. 2007.

Les annexes qui suivent fournissent :

— quelques résultats théoriques utiles (sans démonstrations et sans commentaires), y compris en répétant des résultats présentés dans le corps de l'article, et

— quelques résultats expérimentaux supplémentaires.

A État de l'art

A.1 Résoudre des POSG comme des Occupancy MG

Lemma 1. (énoncé initialement en page 4) *$\sigma_{\boldsymbol{\beta}_{0:\tau-1}}$, pris avec $\boldsymbol{\beta}_\tau$, est une statistique suffisante pour calculer (i) le prochain état d'occupation, $\sigma_{\boldsymbol{\beta}_{0:\tau}}$, et (ii) l'espérance de récompense à τ :* $\mathbb{E}\left[R_\tau \mid \boldsymbol{\beta}_{0:\tau-1} \oplus \boldsymbol{\beta}_\tau\right]$.

A.2 Introduction des jeux locaux

Lemma 7. *$T_m^1(\sigma_\tau, \boldsymbol{\beta}_\tau)$ est linéaire en σ_τ et β_τ^1.*

Lemma 8. *$T_c^1(\sigma_\tau, \boldsymbol{\beta}_\tau)$ est indépendant de β_τ^1.*

Lemma 2. (énoncé initialement en page 5) *$Q_\tau^*(\sigma_\tau, \boldsymbol{\beta}_\tau)$ est concave en β_τ^1 et convexe en β_τ^2.*

B Propriétés et approximation de V^*

B.1 Propriétés de V^*

Linéarité et Lipschitz-continuité de $T(\sigma_\tau, \beta_\tau^1, \beta_\tau^2)$.

Lemma 3. (énoncé initialement en page 5) *À profondeur τ, $T(\sigma_\tau, \boldsymbol{\beta}_\tau)$ est linéaire en β_τ^1, β_τ^2, et σ_τ, où $\boldsymbol{\beta}_\tau = \langle \beta_\tau^1, \beta_\tau^2 \rangle$. Elle est plus précisément 1-Lipschitz-continue (1-LC) en σ_τ (en norme 1), c'est-à-dire que, pour tous σ_τ, σ'_τ :*

$$\|T(\sigma'_\tau, \boldsymbol{\beta}_\tau) - T(\sigma_\tau, \boldsymbol{\beta}_\tau)\|_1 \leq 1 \cdot \|\sigma'_\tau - \sigma_\tau\|_1. \tag{14}$$

Lipschitz-Continuité de V^*.

Lemma 4. (énoncé initialement en page 5) *À profondeur τ, $V_\tau(\sigma_\tau, \boldsymbol{\beta}_{\tau:})$ est linéaire en σ_τ.*

Theorem 2. (énoncé initialement en page 5) *Soit $h_\tau \stackrel{\text{def}}{=} \frac{1-\gamma^{H-\tau}}{1-\gamma}$ si $\gamma < 1$, sinon $h_\tau \stackrel{\text{def}}{=} H - \tau$ (si $\gamma = 1$). Alors $V_\tau^*(\sigma_\tau)$ est λ_τ-Lipschitz-continue en σ_τ à toute profondeur $\tau \in \{0\,..\,H-1\}$, où $\lambda_\tau = \frac{1}{2} h_\tau \, (r_{\max} - r_{\min})$.*

B.2 Approximateurs

$\overline{W}_\tau^1$ et $\underline{W}_\tau^2$.

Lemma 9. *Considérant que les vecteurs $\nu_{[\sigma_H^{c,1}, \beta_{H:}^2]}$ sont des vecteurs nuls, nous avons, pour tout $\tau \in \{0..H-1\}$:*

$$W_\tau^{1,*}(\sigma_\tau, \beta_\tau^1) = \min_{\beta_\tau^2, \langle \beta_{\tau+1:}^2, \nu_{[T_c^1(\sigma_\tau, \beta_\tau^2), \beta_{\tau+1:}^2]} \rangle} \beta_\tau^1 \cdot \left[r(\sigma_\tau, \cdot, \beta_\tau^2) + \gamma T_m^1(\sigma_\tau, \cdot, \beta_\tau^2) \cdot \nu_{[T_c^1(\sigma_\tau, \beta_\tau^2), \beta_{\tau+1:}^2]} \right]. \tag{40}$$

C HSVI pour zs-POSG

Lemma 5. (énoncé initialement en page 8) *En bornant λ_τ par $\lambda^\infty = \frac{1}{2} \frac{1}{1-\gamma} [r_{\max} - r_{\min}]$ quand $\gamma < 1$, et en notant que*

$$thr(\tau) = \begin{cases} \gamma^{-\tau}\epsilon - 2\rho\lambda^\infty \frac{\gamma^{-\tau}-1}{1-\gamma} \text{ if } \gamma < 1, \text{ else} \\ \epsilon - \rho(r_{\max} - r_{\min})(2H + 1 - \tau)\tau, \end{cases} \tag{37}$$

on peut garantir la positivité du seuil en tout $\tau \in 1\,..\,H-1$ en contraignant

$$0 < \rho < \begin{cases} \frac{1-\gamma}{2\lambda^\infty}\epsilon & \text{if } \gamma < 1, \\ \frac{\epsilon}{(r_{\max} - r_{\min})(H+1)H} & \text{if } \gamma = 1. \end{cases} \tag{38}$$

Lemma 6. (énoncé initialement en page 8) *Quand $\gamma < 1$, en utilisant la constante de Lipschitz indépendante de la profondeur λ^∞, et avec la largeur maximale entre initialisations $W \stackrel{\text{def}}{=} \|\overline{V}^{(0)} - \underline{V}^{(0)}\|_\infty$, la longueur des trajectoires est majorée par*

$$T_{\max} \stackrel{\text{def}}{=} \left\lceil \log_\gamma \frac{\epsilon - \frac{2\rho\lambda^\infty}{1-\gamma}}{W - \frac{2\rho\lambda^\infty}{1-\gamma}} \right\rceil. \tag{39}$$

TABLE 2 – Nombres d'étates/actions/observations pour chaque banc d'essai

	$\mathcal{S}$	$\mathcal{A}^1$	$\mathcal{A}^2$	$\mathcal{O}^1$	$\mathcal{O}^2$
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

C.1 Convergence

Lemma 10. *Soit $(\sigma_0, \ldots, \sigma_{\tau+1})$ une trajectoire complète générée par zs-OMG-HSVI et $\boldsymbol{\beta}_\tau$ la règle de décision comportemental jointe qui a induit la dernière transition, c'est-à-dire que $\sigma_{\tau+1} = T(\sigma_\tau, \boldsymbol{\beta}_\tau)$. Alors, après mise-à-jour de $\overline{W}^1_\tau$ et $\underline{W}^2_\tau$, nous avons que $\overline{W}^1_\tau(\sigma_\tau, \beta^1_\tau) - \underline{W}^2_\tau(\sigma_\tau, \beta^2_\tau) \leq \gamma thr(\tau + 1)$.*

Lemma 11 (Évolution monotone de $\overline{W}^1_\tau$ et $\underline{W}^2_\tau$)**.** *Soient $K\overline{W}^1_\tau$ et $K\underline{W}^2_\tau$ les approximateurs après une mise-à-jour en σ_τ avec les règles de décision comportementales $\langle \overline{\beta}^1_\tau, \underline{\beta}^2_\tau \rangle$ (respectivement associées aux vecteurs $\overline{\nu}^2_{\tau+1}$ et $\underline{\nu}^1_{\tau+1}$). Soient aussi $K^{(n+1)}\overline{W}^1_\tau$ et $K^{(n+1)}\underline{W}^2_\tau$ les mêmes approximateurs après n autres mises-à-jour (dans divers états d'occupation). Alors,*

$$\max_{\beta^1_\tau} K^{(n+1)}\overline{W}^1_\tau(\sigma_\tau, \beta^1_\tau) \leq \max_{\beta^1_\tau} K\overline{W}^1_\tau(\sigma_\tau, \beta^1_\tau) \leq \overline{W}^1_\tau(\sigma_\tau, \overline{\beta}^1_\tau) \quad \textit{et} \tag{41}$$

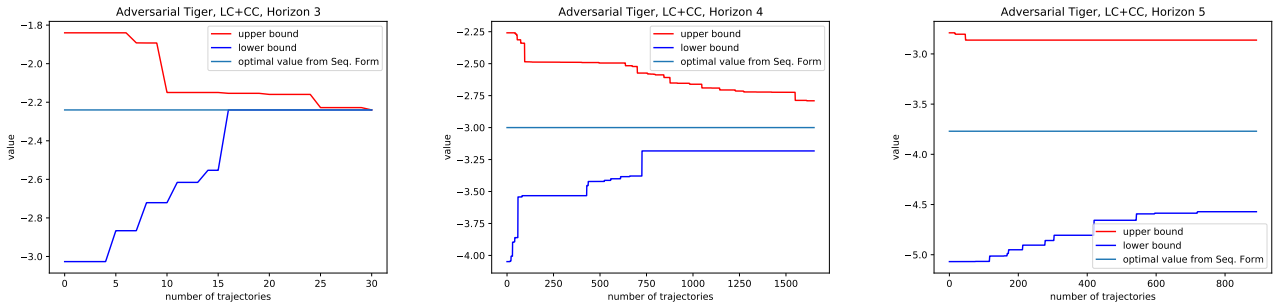
$$\min_{\beta^2_\tau} K^{(n+1)}\underline{W}^2_\tau(\sigma_\tau, \beta^2_\tau) \geq \min_{\beta^2_\tau} K\underline{W}^2_\tau(\sigma_\tau, \beta^2_\tau) \geq \underline{W}^2_\tau(\sigma_\tau, \underline{\beta}^2_\tau). \tag{42}$$

Lemma 12. *Après avoir mis-à-jour, dans l'ordre, $\overline{W}^1_\tau$ et $\overline{V}_\tau$, nous avons $K\overline{V}_\tau(\sigma_\tau) \leq \max_{\beta^1_\tau} K\overline{W}^1_\tau(\sigma_\tau, \beta^1_\tau)$.*
Après avoir mis-à-jour, dans l'ordre, $\underline{W}^2_\tau$ et $\underline{V}_\tau$, nous avons $K\underline{V}_\tau(\sigma_\tau) \geq \min_{\beta^2_\tau} K\underline{W}^2_\tau(\sigma_\tau, \beta^2_\tau)$.

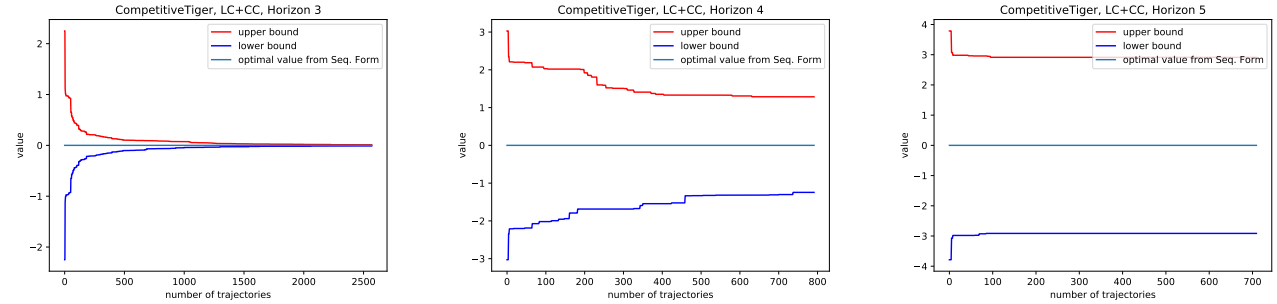
Theorem 4. (énoncé initialement en page 8) *zs-OMG-HSVI (algorithme 1) termine en temps fini avec une ϵ-approximation de $V^*_0(\sigma_0)$.*

D Expérimentations

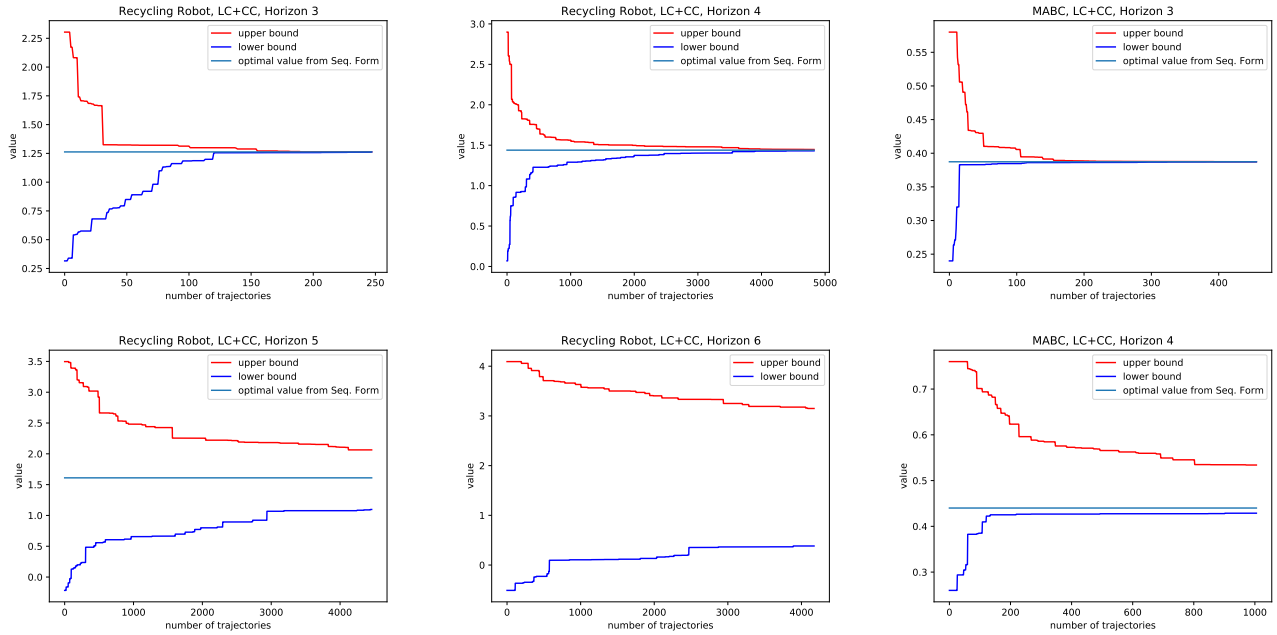
Cette section fournit (i) des informations concernant les bancs d'essai considérés dans le tableau 2 et (ii) des résultats supplémentaires.
Les graphes de la figure 1 montrent comment les valeurs majorantes et minorantes en σ_0 (c'est-à-dire $\overline{V}_0(\sigma_0)$ et $\underline{V}_0(\sigma_0)$) évoluent en fonction du nombre d'itérations, ici en considérant les mêmes bancs d'essai et horizons temporels que dans la section 6 (sauf pour $H = 2$).
Comme attendu, ces bornes convergent de manière monotone vers la valeur optimale (ici fournies par Sequence Form dans tous les cas sauf Recycling Robot pour $H = 6$). Cette convergence serait symétrique dans Competitive Tiger, le seul jeu symétrique, si l'algorithme brise les symétries de manière biaisé quand plusieurs situations équivalentes sont possibles.
Notons aussi qu'on s'attend à ce que les durées des itérations augmentent avec l'horizon temporel, de sorte qu'il est surprenant d'observer, à la fois dans Competitive Tiger et Recycling Robot, des nombres similaires d'itérations en 24 heures pour les horizons au-delà de 4. Ce phénomène est actuellement en cours d'étude.



(a) Adversarial Tiger



(b) Competitive Tiger



(c) Recycling Robot (d) Mabc

FIGURE 1 – Évolution des valeurs majorante et minorante $\overline{V}_0(\sigma_0)$ (en rouge) et $\underline{V}_0(\sigma_0)$ (en bleu) de OMGHSVI$_{\text{CC}}^{\text{LC}}$ pour les différents bancs d'essai en fonction du nombre d'itérations (de trajectoires générées). Valeur optimale trouvée par Sequence Form en vert pour référence (si disponible).