Policy iteration algorithm for zero-sum stochastic games with mean payoff

Jean Cochet-Terrasson, Stéphane Gaubert

POLICY ITERATION ALGORITHM FOR ZERO-SUM STOCHASTIC GAMES WITH MEAN PAYOFF

JEAN COCHET-TERRASSON AND STÉPHANE GAUBERT

ABSTRACT. We give a policy iteration algorithm to solve zero-sum stochastic games with finite state and action spaces and perfect information, when the value is defined in terms of the mean payoff per turn. This algorithm does not require any irreducibility assumption on the Markov chains determined by the strategies of the players. It is based on a discrete nonlinear analogue of the notion of reduction of a super-harmonic function.

ALGORITHME D'ITÉRATION SUR LES POLITIQUES POUR LES JEUX STOCHASTIQUES À SOMME NULLE AVEC GAIN MOYEN

RÉSUMÉ. Nous donnons un algorithme d'itération sur les politiques pour résoudre les jeux stochastiques à somme nulle, avec espaces d'état et d'action finis, en information parfaite, lorsque la valeur du jeu est définie en termes de gain moyen par tour. Cet algorithme ne demande pas que les chaînes de Markov déterminées par les stratégies des deux joueurs soient irréductibles. Il repose sur un analogue discret non-linéaire de la notion de réduite d'une fonction surharmonique.

VERSION ABRÉGÉE EN FRANÇAIS

Une application f définie sur $\mathbb{R}^n$ est dite *polyédrale* si l'on peut recouvrir $\mathbb{R}^n$ par un nombre fini de polyèdres de sorte que la restriction de f à chacun des polyèdres soit affine. Une application $f : \mathbb{R}^n \to \mathbb{R}^n$ est dite *contractante* pour une norme $\|\cdot\|$ si $\|f(x)-f(y)\| \leq \|x-y\|$. Kohlberg [12] a démontré que si $f : \mathbb{R}^n \to \mathbb{R}^n$ est polyédrale et contractante pour une norme quelconque, on peut trouver des vecteurs v et η dans $\mathbb{R}^n$ tels que $f(v+t\eta) = v + (t+1)\eta$, pour tout réel t assez grand. Nous appellerons *demi-droite* une application de la forme $t \mapsto v + t\eta$, et nous dirons qu'elle est *invariante* (relativement à f) si elle

Date: June 16, 2006.

vérifie la propriété précédente. L'intérêt d'une demi-droite invariante est qu'elle détermine le taux de croissance des orbites de f, $\chi(f) := \lim_{k\to\infty} f^k(x)/k$, où l'on désigne par f^k la k-ième itérée de f, et où x désigne un vecteur quelconque de $\mathbb{R}^n$. On vérifie en effet que $\chi(f) = \eta$.

Dans cette note, nous donnons un algorithme pour calculer une demi-droite invariante lorsque l'application polyédrale f est croissante au sens large (pour l'ordre partiel usuel de $\mathbb{R}^n$) ainsi qu'*additivement homogène*, ce qui signifie qu'elle commute avec l'addition d'un vecteur constant. Voir [8] pour plus de détails sur cette classe d'applications. Ce travail est motivé par l'étude des jeux stochastiques à deux joueurs et à somme nulle, en information parfaite, avec gain moyen. Lorsque l'espace d'états est $\{1, \dots, n\}$, et lorsque les espaces d'actions sont finis, l'opérateur de programmation dynamique du jeu vérifie les hypothèses précédentes. La coordonnée $\chi_i(f)$ fournit la valeur du jeu lorsque l'état initial est i.

Notre algorithme étend l'itération sur les politiques de Howard [10], qui s'applique au cas à un seul joueur. Hoffman et Karp [9] ont donné une première généralisation au cas à deux joueurs, cependant, leur méthode exige que les matrices de Markov induites par un couple quelconque de stratégies stationnaires des deux joueurs soient irréductibles. Sinon, on rencontre des itérations dégénérées, dans lesquelles la nouvelle stratégie qui est choisie n'améliore pas nécessairement la valeur de la précédente. L'algorithme peut alors cycler.

Nous résolvons ici cette difficulté en incorporant dans chaque itération dégénérée un analogue non-linéaire du calcul de la réduite d'une fonction surharmonique, ce qui revient à résoudre un problème auxiliaire d'arrêt optimal (à un joueur), assurant ainsi la terminaison de l'algorithme. Ceci étend au cas stochastique les algorithmes développés par Gunawardena et les auteurs [7, 4] dans le cas des jeux déterministes. La preuve exploite des résultats d'Akian et du second auteur [1], sur la structure de l'ensemble des points fixes d'une application convexe, croissante, additivement homogène. L'existence d'un algorithme polynômial pour calculer $\chi(f)$ est un problème ouvert [5, 2, 11]. Signalons qu'une version préliminaire du présent travail est apparue dans [3].

Présentons maintenant l'analogue non-linéaire de la notion de fonction surharmonique réduite utilisé dans l'algorithme. Soit $g : \mathbb{R}^n \to \mathbb{R}^n$ une application convexe, croissante, additivement homogène, telle que $\chi(g) = 0$. Un vecteur $u \in \mathbb{R}^n$ est dit *harmonique* (relativement à g) si $g(u) = u$. Il est dit *surharmonique* si $g(u) \leq u$. Rappelons quelques définitions et résultats de [1]. Supposons qu'il existe au moins un vecteur harmonique, u. Nous définissons le sous-différentiel de g au point u, $\partial g(u) := \{M \in \mathbb{R}^{n\times n} \mid g(x) - g(u) \geq M(x-u),\ \forall x \in \mathbb{R}^n\}$, et notons que cet ensemble est formé de matrices stochastiques. Un nœud est dit *critique* s'il appartient à une classe de récurrence d'une des matrices $M \in \partial g(u)$. L'ensemble des nœuds critiques est indépendant du choix du vecteur harmonique u. En fait, lorsque g provient d'un problème de contrôle stochastique avec critère ergodique, un nœud est critique si et seulement si il est récurrent pour une stratégie stationnaire optimale. Pour tout sous-ensemble $I \subset \{1,\dots,n\}$, et pour tout vecteur $x \in \mathbb{R}^n$, nous notons $r_I x \in \mathbb{R}^N$ la restriction de x, définie par $(r_I x)_i := x_i$, Nous posons $x_I := r_I x$, Soit $J := \{1,\dots,n\} \setminus I$. Nous définissons l'application $g_I := r_I \circ g$, et désignons par $\imath_I$ l'application identifiant canoniquement $\mathbb{R}^I \times \mathbb{R}^J$ à $\mathbb{R}^n$, laquelle envoie (y,z) vers le vecteur x tel que $x_i = y_i$ pour tout $i \in I$ et $x_i = z_i$ pour tout $i \in J$.

Théorème 1. *Soit $g : \mathbb{R}^n \to \mathbb{R}^n$ une application convexe, croissante, additivement homogène, admettant au moins un vecteur harmonique. Soit C l'ensemble des nœuds critiques de g, $N := \{1,\dots,n\} \setminus C$, et soit u un vecteur surharmonique. L'une quelconque des conditions suivantes définit de manière unique le même vecteur v : (i) v est harmonique et coïncide avec u sur C ; (ii) v est le plus petit vecteur surharmonique majorant u sur C ; (iii) v coïncide avec u sur C, et sa restriction à N est l'unique point fixe de l'application $y \mapsto g_N(\imath_N(y, u_C))$.*

Nous notons $Q_g u$ l'unique vecteur harmonique v défini dans le Théorème 1. Lorsque $g(x) = Mx$ est linéaire, et que la matrice M est stochastique, $Q_g u$ coïncide avec la *réduite* du vecteur surharmonique u relativement à l'ensemble C. Le calcul de $Q_g u$ est équivalent à la résolution d'un problème d'arrêt optimal, qui peut être menée par itération sur les politiques [6].

Nous définissons maintenant un opérateur analogue à Q_g agissant sur les demi-droites. Nous supposons pour cela que $g : \mathbb{R}^n \to \mathbb{R}^n$ est polyédrale, convexe, croissante, et additivement homogène, et que sa i-ième coordonnée est donnée par l'expression (1) ci-dessous, dans laquelle B_i est un ensemble fini, $r_i^b \in \mathbb{R}$, and P_i^b est un vecteur (ligne) stochastique. Soit $\eta := \chi(g)$. Définissons l'application $\bar{g}(x) := \lim_{t\to\infty} g(x+t\eta) - (t+1)\eta$. Les coordonnées de $\bar{g}$ sont données par $\bar{g}_i(x) = \max_{b\in\bar{B}_i} -\eta_i + r_i^b + P_i^b x$, où $\bar{B}_i$ est l'ensemble des b atteignant le maximum dans l'expression $\max_{b\in B_i} P_i^b\eta$. L'ensemble des *nœuds critiques* de g, $C(g)$, est par définition l'ensemble des nœuds critiques de $\bar{g}$. Nous dirons qu'une demi-droite $w(t)$ est *surinvariante* si $g \circ w(t) \leq w(t+1)$, pour t assez grand.

Corollaire 1. *Soit $g : \mathbb{R}^n \to \mathbb{R}^n$ une application polyédrale, convexe, croissante, et additivement homogène. Soit $w(t) = v + t\eta$ une demi-droite surinvariante de g, telle que $\eta = \chi(g)$. Il existe une unique demi-droite invariante de g qui coïncide avec w sur l'ensemble des nœuds critiques de g. Elle est donnée par $t \mapsto Q_{\bar{g}}v + t\eta$.*

Nous posons $Q_g w(t) := Q_{\bar{g}}v + t\eta$. Dans la suite, nous supposons que chaque coordonnée f_i de f est donnée par (2), où A_i est un ensemble fini, et où chaque f_i^a est une application polyédrale croissante, additivement homogène, et convexe, de $\mathbb{R}^n$ dans $\mathbb{R}$. Une *stratégie* (stationnaire, en boucle fermée) est une application σ de $\{1,\dots,n\}$ dans $\cup_{1\leq i\leq n} A_i$ telle que $\sigma(i) \in A_i$. Pour toute stratégie σ, nous désignons par $f^{(\sigma)}$ l'application de $\mathbb{R}^n$ dans lui même dont la i-ième coordonnée est donnée par $f_i^{(\sigma)} = f_i^{\sigma(i)}$.

Algorithme 1 (Itération sur les politiques pour les jeux stochastiques)**.** Donnée *: Une application f dont les coordonnées sont de la forme* (2). Résultat *: Une demi-droite invariante de f.*

(1) Initialisation. *Choisir une stratégie arbitraire σ_1, et calculer une demi-droite invariante de $f^{(\sigma_1)}$, $w^{(1)}(t) = v^{(1)} + t\eta^{(1)}$. Poser $k = 1$.*

(2) *Si $f \circ w^k(t) = w^k(t+1)$ a lieu pour t assez grand, l'algorithme s'arrête.*

(3) *Sinon,* améliorer la stratégie *en sélectionnant une nouvelle strategie σ_{k+1} telle que $f \circ w^k(t) = f^{(\sigma_{k+1})} \circ w^k(t)$ ait lieu pour t assez grand. Le choix de σ_{k+1}*

doit être effectué de manière conservatrice, ce qui signifie que $\sigma_{k+1}(i) = \sigma_k(i)$ *si* $f_i \circ w^k(t) = f_i^{(\sigma_k)} \circ w^k(t)$ *pour* t *assez grand.*

(4) *Calculer une demi-droite invariante arbitraire* $w(t) = v + t\eta$ *de* $f^{(\sigma_{k+1})}$. *Si* $\eta \neq \eta^{(\sigma_k)}$, *on pose* $w^{k+1} = w$. *Si* $\eta = \eta^{(\sigma_k)}$, *l'itération est qualifiée de* dégénérée, *et on pose* $w^{k+1} = Q_{f^{(\sigma_{k+1})}} w^k$. *On définit* v^{k+1} *et* $\eta^{(\sigma_{k+1})}$ *par* $w^{k+1}(t) = v^{k+1} + t\eta^{(\sigma_{k+1})}$.

(5) *Augmenter* k *d'une unité, et retourner à l'étape 2.*

Théorème 2. *La même stratégie n'est jamais sélectionnée deux fois et donc l'algorithme s'arrête.*

Ce résultat est démontré à l'aide du Théorème 1. On montre notamment que $\chi(f^{(\sigma_{k+1})}) \leq \chi(f^{(\sigma_k)})$, et que si $\chi(f^{(\sigma_{k+1})}) = \chi(f^{(\sigma_k)})$, on a $w^{k+1} \leq w^k$, $w_i^{k+1} = w_i^k$ pour tout $i \in C(f^{(\sigma_{k+1})})$, et $C(f^{(\sigma_{k+1})}) \subset C(f^{(\sigma_k)})$.

---

1. INTRODUCTION

A map f defined on $\mathbb{R}^n$ is *polyhedral* if there is a covering of $\mathbb{R}^n$ by finitely many polyhedra such that the restriction of f to any of these polyhedra is affine. A self-map f of $\mathbb{R}^n$ is *nonexpansive* in a norm $\|\cdot\|$ if $\|f(x) - f(y)\| \leq \|x - y\|$. Kohlberg [12] showed that if f is a polyhedral self-map of $\mathbb{R}^n$ that is nonexpansive in some norm, then, there exist two vectors v and η in $\mathbb{R}^n$ such that $f(v + t\eta) = v + (t+1)\eta$, for all $t \in \mathbb{R}$ large enough. A map of the form $t \mapsto v + t\eta$ is called a *half-line*. It is *invariant* if it satisfies the latter property. The interest of an invariant half-line is that its linear part determines the growth rate of the orbits of f, $\chi(f) := \lim_{k\to\infty} f^k(x)/k$. Here, f^k denotes the k-th iterate of f, and x is an arbitrary vector of $\mathbb{R}^n$. If f has an invariant half-line $t \mapsto v + t\eta$, then, $\chi(f)$ does exist and is equal to η.

In this note, we give an algorithm to compute an invariant half-line, when the polyhedral map f satisfies the following conditions. We require f to be *order-preserving*, meaning

that $x \leq y \implies f(x) \leq f(y)$, where $\leq$ denotes the standard ordering of $\mathbb{R}^n$. We also require f to be *additively homogeneous*, meaning that it commutes with the addition of a constant vector. These two conditions imply that f is nonexpansive in the sup-norm. (See [8] for more background on this class of nonlinear maps.)

This work is motivated by the study of zero-sum two players stochastic games with perfect information and mean payoff. When the state space is $\{1, \dots, n\}$, and when the action spaces are finite, the dynamic programming operator of the game satisfies the previous assumptions. The coordinate $\chi_i(f)$ gives the value of an infinite game with *mean payoff*, in which the initial state is i and the payoff of the infinite trajectory induced by a pair of strategies of the two players is defined as the Cesaro limit of the expectations of the payoffs of the successive transitions.

Our algorithm extends Howards' policy iteration algorithm [10], which applies to the one player case. Hoffman and Karp [9] gave a partial extension to the two players case. However, their method requires every Markov chain associated to a pair of stationary feedback strategies of the two players to be irreducible. If this assumption does not hold, degenerate iterations may occur, in which the new strategy which is selected may not have an improved value. Then, the algorithm may cycle.

We solve this difficulty by computing a non-linear analogue of a reduced super-harmonic function, at each degenerate iteration, which requires solving an auxiliary (one player) optimal stopping problem. This is intimately related with Perron's method in the study of the Dirichlet problem. The present algorithm extends the ones which have been developed by the authors and Gunawardena [7, 4] in the case of deterministic games. Its proof exploits earlier results of Akian and the second author [1], on the structure of the fixed point set of a convex order-preserving additively homogeneous map. The existence of a polynomial time algorithm to compute $\chi(f)$ is an open question [5], even in the deterministic case [2, 11]. Finally, we note that a preliminary account of the present work has appeared in [3].

2. REDUCED SUPER-HARMONIC VECTORS AND POLICY ITERATION ALGORITHM

We first present the non-linear analogue of a result of classical potential theory, on which the algorithm relies. Assume that g is a self-map of $\mathbb{R}^n$ that is convex, order preserving, additively homogeneous, with $\chi(g) = 0$. We say that a vector $u \in \mathbb{R}^n$ is *harmonic* (with respect to g) if $g(u) = u$, and that it is *super-harmonic* if $g(u) \leq u$. Let us recall some definitions and results of [1]. Assume that there exists at least one harmonic vector, u. The *subdifferential* of g at point u is defined by $\partial g(u) := \{M \in \mathbb{R}^{n\times n} \mid g(x) - g(u) \geq M(x-u),\ \forall x \in \mathbb{R}^n\}$. This set consists of stochastic matrices. We say that a node is *critical* if it belongs to a recurrence class of some matrix $M \in \partial g(u)$. The set of critical nodes is independent of the choice of the harmonic vector u. Indeed, when g arises from a stochastic control problem with ergodic reward, a node is critical iff it is recurrent for some stationary optimal strategy. If I is any subset of $\{1,\dots,n\}$, we denote by r_I the restriction from $\mathbb{R}^n$ to $\mathbb{R}^I$, such that $(r_I x)_i := x_i$, for all $i \in I$. For all $u \in \mathbb{R}^n$, we define $u_I := r_I u$. Let $J := \{1,\dots,n\}\setminus I$. We define the map $g_I := r_I \circ g$, and we denote by $\imath_I$ the canonical map identifying $\mathbb{R}^I \times \mathbb{R}^J$ to $\mathbb{R}^n$, which sends (y,z) to the vector u such that $u_i = y_i$ for all $i \in I$ and $u_i = z_i$ for all $i \in J$.

Theorem 1. *Let g denote a convex, order preserving, and additively homogeneous self-map of $\mathbb{R}^n$. Assume that g admits at least one harmonic vector. Let C denote the set of critical nodes of g, let N denote the complement of C in $\{1,\dots,n\}$, and let u denote a super-harmonic vector. Then, any of the following conditions defines uniquely the same vector v: (i) v is harmonic and coincides with u on C; (ii) v is the smallest super-harmonic vector that dominates u on C; (iii) v coincides with u on C and its restriction to N is the unique fixed point of the map $y \mapsto g_N(\imath_N(y, u_C))$.*

We denote by $Q_g u$ the unique harmonic vector v defined in Theorem 1. When $g(x) = Mx$ is a linear operator, and M is a stochastic matrix, $Q_g u$ coincides with the *reduced* super-harmonic vector of u with respect to the set C. When g is a max-plus linear operator,

the operator Q_g coincides with the *spectral projector* which has been defined in the max-plus literature, see [4]. For this reason, we call Q_g the (nonlinear) *spectral projector* of g.

Let us now explain how to compute $Q_g u$ when g is polyhedral. Assume that every coordinate g_i is given as follows:

$$g_i(x) = \max_{b \in B_i} r_i^b + P_i^b x \ , \tag{1}$$

where B_i is a finite set, $r_i^b \in \mathbb{R}$, and P_i^b is a stochastic (row) vector. Note first that an invariant half-line of g can be computed by applying the multichain policy iteration algorithm of Denardo and Fox [6]. This provides $\chi(g)$, together with an harmonic vector of g when $\chi(g) = 0$. Once an harmonic vector of g is known, the set of critical nodes can be computed by the algorithm of [1, § 6.3]. Hence, to compute $v := P_g u$, it suffices to apply the standard fixed point iteration to the latter map. Alternatively (experiments indicate this is faster), one may note that the vector y solution of $y = g_N(\imath_N(y, u_C))$ is the value of an optimal stopping problem, in which the process dies when reaching the set C. This vector can be computed by the original policy iteration algorithm of Howard [10].

We now define a spectral projector acting on half-lines. We assume that g is a polyhedral, convex, order preserving, and additively homogeneous self-map of $\mathbb{R}^n$. Let $\eta := \chi(g)$, and define $\bar{g}(x) := \lim_{t\to\infty} g(x + t\eta) - (t+1)\eta$. Since g is polyhedral, the limit is attained for all sufficiently large t. Indeed, if the coordinates of g are of the form (1), we have $\bar{g}_i(x) = \max_{b \in \bar{B}_i} -\eta_i + r_i^b + P_i^b x$, where $\bar{B}_i$ is the set of actions b attaining the maximum in $\max_{b \in B_i} P_i^b \eta$. We define the set of *critical nodes* of g, $C(g)$, to be the set of critical nodes of $\bar{g}$. An half-line $w(t) = v + t\eta$ is *super-invariant* if $g \circ w(t) \leq w(t+1)$, for t large enough.

Corollary 1. *Assume that g is a polyhedral, convex, order preserving, and additively homogeneous self-map of $\mathbb{R}^n$. Assume that $w(t) = v + t\eta$ is a super-invariant half-line of g with $\eta = \chi(g)$. Then, there exists a unique invariant half-line of g which coincides with w on the set of critical nodes of g. It is given by $t \mapsto Q_{\bar{g}} v + t\eta$.*

We define $Q_g w$ to be the map $t \mapsto Q_{\bar{g}} v + t\eta$. In order to present the algorithm, we assume that every coordinate of f is given by:

$$f_i(x) = \inf_{a \in A_i} f_i^a(x) \ , \tag{2}$$

where A_i is a finite set, and f_i^a is a polyhedral order preserving, additively homogeneous, and convex map from $\mathbb{R}^n$ to $\mathbb{R}$. We call (stationary, feedback) *strategy* a map σ from $\{1, \dots, n\}$ to $\cup_{1 \leq i \leq n} A_i$ such that $\sigma(i) \in A_i$. For all strategies σ, we denote by $f^{(\sigma)}$ the self-map of $\mathbb{R}^n$ the i-th coordinate of which is given by $f_i^{(\sigma)} = f_i^{\sigma(i)}$.

Algorithm 1 (Policy iteration for stochastic games)**.** Input*: A map f the coordinates of which are of the form* (2)*.* Output*: An invariant half-line of f.*

(1) Initialisation. *Select an arbitrary strategy σ_1. Compute an invariant half-line of $f^{(\sigma_1)}$, $w^{(1)}(t) = v^{(1)} + t\eta^{(1)}$. Set $k = 1$.*
(2) *If $f \circ w^k(t) = w^k(t+1)$ holds for t large enough, the algorithm halts.*
(3) *Otherwise,* improve the strategy*, by selecting a strategy σ_{k+1} such that $f \circ w^k(t) = f^{(\sigma_{k+1})} \circ w^k(t)$, for t large enough. The choice of σ_{k+1} must be conservative, meaning that $\sigma_{k+1}(i) = \sigma_k(i)$ if $f_i \circ w^k(t) = f_i^{(\sigma_k)} \circ w^k(t)$, for t large enough.*
(4) *Compute an arbitrary invariant half-line $w(t) = v + t\eta$ of $f^{(\sigma_{k+1})}$. If $\eta \neq \eta^{(\sigma_k)}$, we set $w^{k+1} = w$. If $\eta = \eta^{(\sigma_k)}$, we say that the iteration is* degenerate*. Set $w^{k+1} = Q_{f^{(\sigma_{k+1})}} w^k$, and define v^{k+1} and $\eta^{(\sigma_{k+1})}$ by $w^{k+1}(t) = v^{k+1} + t\eta^{(\sigma_{k+1})}$.*
(5) *Increment k by one and go to step 2.*

Theorem 2. *A strategy cannot be selected twice, and so, the algorithm terminates.*

This result is proved using Theorem 1. We show, as intermediate results, that $\chi(f^{(\sigma_{k+1})}) \leq \chi(f^{(\sigma_k)})$, and that, when $\chi(f^{(\sigma_{k+1})}) = \chi(f^{(\sigma_k)})$, we have $w^{k+1} \leq w^k$, $w_i^{k+1} = w_i^k$ for all $i \in C(f^{(\sigma_{k+1})})$, and $C(f^{(\sigma_{k+1})}) \subset C(f^{(\sigma_k)})$.

Let us give the missing details in the implementation of Algorithm 1, when every map f_i^a is given by $f_i^a = \sup_{b \in B_i^a} r_i^{ab} + P_i^{ab} x$, where B_i^a is a finite set, $r_i^{ab} \in \mathbb{R}$, and P_i^{ab} is a stochastic vector. Then, for all strategies σ, f^σ is the dynamic programming operator of a Markov decision process with finite state and action spaces. Recall that an invariant

half-line of $f^{(\sigma)}$ can be computed by applying the policy iteration algorithm for multichain Markov decision processes [6]. The next state is to evaluate the asymptotic behaviour of $f \circ w(t)$, where $w(t)$ is an half-line, as t tends to infinity. Computing the sum of two half-lines, multiplying an half-line by a scalar, and computing the infimum or supremum of two half-lines when t tends to infinity, are linear time operations. It follows that $f \circ w(t)$ can be evaluated (Step 2) in a time which is linear in the size of the coding of the operator f, and a strategy σ_{k+1} satisfying the conditions of Step 3 is obtained as a byproduct of this evaluation.

Example 1. Consider a directed graph, with set of nodes $1, \dots, n$ and set of arcs $E \subset \{1, \dots, n\}^2$, in which each arc (i, j) is equipped with a weight $r_{ij} \in \mathbb{R}$. The map f defined by:

$$f_i(x) = \frac{1}{2}\big(\max_{j:\ (i,j)\in E} r_{ij} + x_j \ + \ \min_{j:\ (i,j)\in E} r_{ij} + x_j \big)$$

arises as the dynamic programming operator of a variant with mean payoff of the "tug of war" game [14] and in a class of auction games [13]. Let us apply the algorithm to the complete graph with 3 nodes, with the weights $r_{11} = 1$, $r_{12} = 2$, $r_{13} = 7$, $r_{21} = 3$, $r_{22} = 3$, $r_{23} = 4$, $r_{31} = 8$, $r_{32} = 5$, $r_{33} = 1$. The action space in every state i can be identified with $\{1, \dots, 3\}$. Let us choose the greedy strategy σ_1, such that $\sigma_1(1) = 1$, $\sigma_1(2) = 1$, $\sigma_1(3) = 3$. The corresponding operator is given by $f^{(\sigma_1)}(x) = \big((1 + x_1 + \max(1 + x_1, 2 + x_2, 7 + x_3))/2, (3 + x_1 + \max(3 + x_1, 3 + x_2, 4 + x_3))/2, (1 + x_3 + \max(8 + x_1, 5 + x_2, 1 + x_3))/2\big)^T$. We compute an invariant half-line of $f^{(\sigma_1)}$ by the algorithm of [6]. We obtain for instance $w^1 = v^1 + t\eta^1$, where $v^1 = (0.5, 0, 1)^T$ and $\eta^1 = (\lambda, \lambda, \lambda)^T$, with $\lambda := 4.25$. Since $f(w^1(t)) < f^{(\sigma_1)}(w^1(t))$, we improve the policy. We get $\sigma_2(1) = 1$, $\sigma_2(2) = 2$, $\sigma_2(3) = 3$. One can check, again by the algorithm of [6], that $\chi(f^{(\sigma_2)}) = (\lambda, \lambda, \lambda)^T$, and so, the iteration is degenerate. Using the algorithm of [1, § 6.3], we get that the set C of critical nodes of $f^{(\sigma_2)}$ is $\{1, 3\}$, and so, $N = \{2\}$. Let $g := f^{(\sigma_2)}$. By Corollary 1, the image of w^1 by the spectral projector Q_g is of the form $z + t\eta^{(1)}$, where $z_1 = v^1_1$, $z_3 = v^1_3$, and z_2 is obtained by solving the equation $4.25 + z_2 = (3 + z_2 + \max(3 + z_1, 3 + z_2, 4 + z_3))/2$. The unique solution is

$z_2 = -0.5$ (in general, this solution could be found by the basic policy iteration algorithm of [10]). So, $w^{(2)}(t) = (0.5, -0.5, 1)^T + t(\lambda, \lambda, \lambda)^T$ is an invariant half-line of $f^{(\sigma_2)}$. Since $f(w^2(t)) = w^2(t+1)$, the algorithm stops, showing that $\chi(f) = (\lambda, \lambda, \lambda)^T$.

REFERENCES

[1] M. Akian and S. Gaubert. Spectral theorem for convex monotone homogeneous maps, and ergodic control. *Nonlinear Analysis. Theory, Methods & Applications*, 52(2):637–679, 2003.

[2] H. Bjorklund, S. Sandberg, and S. Vorobyov. A combinatorial strongly subexponential strategy improvement algorithm for mean payoff games. Technical Report 05, DIMACS, 2004.

[3] J. Cochet-Terrasson. *Algorithmes d'itération sur les politiques pour les applications monotones contractantes*. Thèse, École des Mines de Paris, 2001.

[4] J. Cochet-Terrasson, S. Gaubert, and J. Gunawardena. A constructive fixed point theorem for min-max functions. *Dynamics and Stability of Systems*, 14(4):407–433, 1999.

[5] A. Condon. The complexity of stochastic games. *Inform. and Comput.*, 96(2):203–224, 1992.

[6] E. V. Denardo and B. L. Fox. Multichain Markov Renewal Programs. *SIAM J.Appl.Math*, 16:468–487, 1968.

[7] S. Gaubert and J. Gunawardena. The duality theorem for min-max functions. *C.R. Acad. Sci.*, 326(1):43–48, 1998.

[8] S. Gaubert and J. Gunawardena. The Perron-Frobenius theorem for homogeneous, monotone functions. *Trans. of AMS*, 356(12):4931–4950, 2004.

[9] A. J. Hoffman and R. M. Karp. On nonterminating stochastic games. *Management sciences*, 12(5):359–370, 1966.

[10] R. Howard. *Dynamic Programming and Markov Processes*. Wiley, 1960.

[11] M. Jurdziński, M. Paterson, and U. Zwick. A deterministic subexponential algorithm for solving parity games. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA 2006)*, January 2006.

[12] E. Kohlberg. Invariant half-lines of nonexpansive piecewise-linear transformations. *Math. Oper. Res.*, 5(3):366–372, 1980.

[13] A. J. Lazarus, D. E. Loeb, J. G. Propp, W. R. Stromquist, and D. H. Ullman. Combinatorial games under auction play. *Games Econom. Behav.*, 27(2):229–264, 1999.

[14] Y. Peres, O. Schramm, S. Sheffield, and D. Wilson. Tug-of-war and the infinity Laplacian. arXiv:math.AP/0605002, 2006.

JEAN COCHET-TERRASSON, CGA, 14 RUE SAINT DOMINIQUE, 75007 PARIS, FRANCE.

STÉPHANE GAUBERT (AUTHOR FOR CORRESPONDENCE). INRIA, DOMAINE DE VOLUCEAU, B.P. 105, 78153 LE CHESNAY CEDEX, FRANCE. TEL: 01 39 63 52 58, FAX: 01 39 63 57 86

E-mail address: Stephane.Gaubert@inria.fr