POLICY ITERATION FOR PERFECT INFORMATION STOCHASTIC MEAN PAYOFF GAMES WITH BOUNDED FIRST RETURN TIMES IS STRONGLY POLYNOMIAL

MARIANNE AKIAN AND STÉPHANE GAUBERT

Abstract. Recent results of Ye and Hansen, Miltersen and Zwick show that policy iteration for one or two player (perfect information) zero-sum stochastic games, restricted to instances with a fixed discount rate, is strongly polynomial. We show that policy iteration for mean-payoff zero-sum stochastic games is also strongly polynomial when restricted to instances with bounded first mean return time to a given state. The proof is based on methods of nonlinear Perron-Frobenius theory, allowing us to reduce the mean-payoff problem to a discounted problem with state dependent discount rate. Our analysis also shows that policy iteration remains strongly polynomial for discounted problems in which the discount rate can be state dependent (and even negative) at certain states, provided that the spectral radii of the nonnegative matrices associated to all strategies are bounded from above by a fixed constant strictly less than 1.

1. Introduction

Motivation and earlier works. Policy iteration algorithm is a classical algorithm to solve discounted Markov decision problems (one player games) with finite state and actions spaces. A policy is a map from the set of states to the set of actions, representing a Markovian decision rule. The algorithm constructs a sequence of policies such that the associated sequence of values is strictly decreasing (assuming that the player minimizes her cost function). Hence, its number of iteration is bounded by the number of policies. The method carries over to discounted zero-sum games with perfect information, still with finite state and action spaces. It now makes external iterations in the space of policies of the first player, and at each step, solves an auxiliary Markov decision problem, making then internal iterations in the space of policies of the second player. Again, the first player never selects twice the same policy, which entails that the algorithm does terminate in a time which is bounded by the product of the numbers of policies of both players. This yields an exponential bound on the execution time, as the number of policies of one player can be exponential in the number of states. However, this general exponential bound does not capture the experimental efficiency of the algorithm on most applications.

Some recent results shed light on the behavior of policy iteration as a function of some particular parameters, such as the discount factor. Friedmann constructed in [10] an infinite family of 2-player discounted deterministic games with a discount factor tending to 1, showing that the number of policy iterations can indeed be exponential. Fearnley [8] and Andersson [3] extended his result to 1-player stochastic

Key words and phrases. Stochastic games, policy iteration, mean return time, Doeblin state, cone spectral radius.

games. However, Ye showed in [19] that policy iteration solves 1-player discounted games with a *fixed* discount factor $\lambda < 1$ in *strongly polynomial* time (λ is not part of the input). Then, Hansen, Miltersen and Zwick extended this result in [11] to zero-sum 2-player discounted games with perfect information, and improved Ye's bound. They showed that the number of external iterations of the policy iteration algorithm for 2-player games with a fixed discount factor $\lambda < 1$ is bounded by:

$$(m+1)(1+\frac{\log(n^2/(1-\lambda))}{-\log(\lambda)}) = \mathcal{O}(\frac{m}{1-\lambda}\log\frac{n}{1-\lambda}), \tag{1}$$

where n is the number of states, and m is the *total number of actions of both players*, that is the number of triples (i,a,b) where i is a state, a is an action of first player, and b is an action of second player.

Contribution. We show that policy iteration still has a strongly polynomial behavior for a class of mean payoff games, as well as for a more general class of discounted games.

As a preliminary step, we show that we can improve the bound (1), in the original situation considered in [11]. We replace this bound by the following one (Theorem 5):

$$s_{\max} := (m_1 - n)(1 + \lfloor\frac{\log(1-\lambda)}{\log(\lambda)}\rfloor) = \mathcal{O}(\frac{m_1 - n}{1-\lambda}\log\frac{1}{1-\lambda}), \tag{2}$$

with m_1 the *total number of actions of the first player*, that is the number of couples state-action (i,a). The above new bound is obtained by adapting the technique of Ye and Hansen, Miltersen and Zwick to nonlinear maps which allows us in particular to replace m by m_1. Note that the bound (2) is linear in the size of the input, for a fixed λ.

Then, we consider games with state dependent discount factors, possibly greater than 1 locally. We establish a strongly polynomial bound for the number of iterations (Corollary 10) which differs from (1) and (2) in that the discount factor λ is now replaced by the maximum of the spectral radii of all the transition matrices associated to pairs of policies of both players. We introduce a natural scaling transformation, which has the property of leaving invariant the combinatorial trace of the policy iteration algorithm. This scaling is obtained using techniques of nonlinear Perron-Frobenius theory [16, 2]. An advantage of the present bound is that it is invariant by scaling. For instance, with a state dependent discount factor < 1, it leads to a tighter bound than the one which may be derived from (1) or (2).

Finally, we derive (Corollary 15) a strongly polynomial bound for the subclass of mean payoff games such that there is a distinguished state to which the mean return time is bounded by a constant $K = 1/(1-\lambda)$, for every choice of policies. This condition implies that each transition matrix associated to a pair of policies of both players has a unique recurrence class, and that there is a state which is common to each of these classes.

The paper is organized as follows. We present background materials on zero-sum two-player stochastic games in Section 2, and on policy iterations in Section 3. We state our results in Section 4. Proofs or sketches of proofs, as well as some results of Perron-Frobenius theory on which they are based, can be found in the following sections.

2. Two player zero-sum stochastic games with discrete time and mean payoff

2.1. **The game processes.**

Two player zero-sum stochastic games were introduced by Shapley in the early fifties, see [18]. We recall in this section basic definitions in the case of finite state space and discrete time (for more details see [18, 9]). When there is only one player (the set of actions of one of the two players is reduced to a singleton), such a game is more commonly called a Markov Decision Process (MDP), we refer to [13, 6, 17] for this topic.

We consider the finite state space $[n] := \{1, \dots, n\}$. A stochastic process $(\xi_k)_{k\geq 0}$ on $[n]$ gives the state of the game at each point time k, called *stage*. At each of these stages, two players, called "MIN" and "MAX" (the minimizer and the maximizer) have the possibility to influence the course of the game.

The *stochastic game* $\Gamma(i_0)$ starting from $i_0 \in [n]$ is played in stages as follows. The initial state ξ_0 is equal to i_0 and known by the players. Player MIN plays first, and chooses an action α_0 in a set of possible actions A_{ξ_0}. Then the second player, MAX, chooses an action β_0 in a set of possible actions B_{ξ_0}. The actions of both players and the current state determine the payment $r_{\xi_0}^{\alpha_0\beta_0}$ made by MIN to MAX and the probability distribution $j \mapsto P_{\xi_0 j}^{\alpha_0\beta_0}$ of the new state ξ_1. Then the game continues from state ξ_1, and so on.

At a stage k, each player chooses an action knowing the *history* defined by $\zeta_k = (\xi_0, \alpha_0, \beta_0, \cdots, \xi_{k-1}, \alpha_{k-1}, \beta_{k-1}, \xi_k)$ for MIN and (ζ_k, α_k) for MAX. We call a *strategy* for a player, a rule which tells him the action to choose in any situation. Assume $A_i \subset A$ and $B_i \subset B$ for some sets A and B. We shall consider only *pure Markovian strategies* for MIN (resp. MAX). The latter are sequences $\bar\sigma := (\sigma_0, \sigma_1, \cdots)$ (resp. $\bar\delta := (\delta_0, \delta_1, \cdots))$ where σ_k is a map $[n] \to A$ such that $\sigma_k(i) \in A_i$ for all $i \in [n]$ (resp. δ_k is a map $[n] \times A \to B$ such that $\delta_k(i, a) \in B_i \, \forall i \in [n]$, $a \in A_i$). They are said to be *stationary* if they are independent of k. Then σ_k is also denoted by σ and δ_k by δ. Also $\bar\sigma$ and $\bar\delta$ are identified with σ and δ respectively. A pure Markovian stationary strategy is also called a *feedback policy* or simply a *policy*.

A strategy $\bar\sigma = (\sigma_k)_{k\geq 0}$ (resp. $\bar\delta = (\delta_k)_{k\geq 0}$) together with an initial state determines stochastic processes $(\alpha_k)_{k\geq 0}$ for the actions of MIN, $(\beta_k)_{k\geq 0}$ for the actions of MAX and $(\xi_k)_{k\geq 0}$ for the states of the game. For instance, for each pair of feedback policies (σ, δ) of the two players, the state process $(\xi_k)_{k\geq 0}$ is a Markov chain on $[n]$ with transition probability

$$P(\xi_{k+1} = j \,|\, \xi_k = i) \;=\; P_{ij}^{\sigma(i)\delta(i,\sigma(i))} \qquad \text{for } i, j \in [n] \;\; ,$$

and $\alpha_k = \sigma(\xi_k)$ and $\beta_k = \delta(\xi_k, \alpha_k)$.

2.2. **Non-uniformly discounted and mean payoff games.**

The payoff of the game $\Gamma(i)$ starting from i is the expected sum of the rewards at all steps of the game that MAX wants to maximize and MIN to minimize. In this paper we shall consider games with an infinite horizon and a discount factor γ, which is not uniform in that it depends both on the state and actions, $\gamma : [n] \times A \times B \to [0, \infty)$. We allow $\gamma(i, a, b)$ to take values larger that 1 for some (i, a, b). The reward at time k is defined to be the payment made by MIN to MAX multiplied by all discount factors from time 0 to time $k - 1$. Thus, when the strategies $\bar\sigma$ for MAX and $\bar\delta$ for MIN are fixed, the infinite horizon discounted payoff of the game $\Gamma(i, \bar\sigma, \bar\delta)$ starting from i is given by

$$J^\gamma(i,\bar\sigma,\bar\delta) \;=\; \mathbb{E}_i^{\bar\sigma\bar\delta}\left[\sum_{k=0}^{\infty}\Big(\prod_{\ell=0}^{k-1}\gamma(\xi_\ell,\alpha_\ell,\beta_\ell)\Big)r_{\xi_k}^{\alpha_k\beta_k}\right],$$

where $\mathbb{E}_i^{\bar\sigma,\bar\delta}$ denotes the expectation for the probability law determined by the choice of strategies. When $\gamma\leq 1$, meaning that $\gamma(i,a,b)\leq 1$ holds for all $i\in[n]$, $a\in A$, $b\in B$, the above discounted game can be seen equivalently as a game which has, at each stage, a stopping probability equal to $1-\gamma(i,a,b)$.

In all the paper, we shall assume, that

(A1) the action spaces A_i and B_i, $i\in[n]$, are finite sets.

We shall write $\gamma\ll 1$ when the discount factor is such that $\gamma(i,a,b)<1$ holds for all $i\in[n]$, $a\in A$, $b\in B$. This is the case if and only if there exists a scalar λ such that:

(A2) $\gamma(i,a,b)\leq\lambda$, for all $i\in[n]$, $a\in A$, $b\in B$, with $\lambda\in[0,1)$.

Then, one can transform the above discounted game into a game with an additional state (a "cemetery" state) and a discount factor identically equal to λ (independent of state and actions). We can then apply to this situation earlier results concerning constant discount factors.

We shall also consider mean payoff games, defined as follows. When the strategies $\bar\sigma$ for MIN and $\bar\delta$ for MAX are fixed, the *(undiscounted) payoff in finite horizon* τ of the game $\Gamma(i,\bar\sigma,\bar\delta)$ starting from i is

$$J^\tau(i,\bar\sigma,\bar\delta) \;=\; \mathbb{E}_i^{\bar\sigma\bar\delta}\left[\sum_{k=0}^{\tau-1} r_{\xi_k}^{\alpha_k\beta_k}\right],$$

and its *mean payoff* is

$$J(i,\bar\sigma,\bar\delta) \;=\; \limsup_{\tau\to\infty}\frac{1}{\tau}\,J^\tau(i,\bar\sigma,\bar\delta).$$

The discounted infinite horizon game with a discount factor $\gamma\ll 1$, the finite horizon game and the mean payoff game, are all known to have a *value*, denoted respectively by v_i^γ, v_i^τ and ρ_i,

$$v_i^\gamma \;:=\; \inf_{\bar\sigma}\sup_{\bar\delta} J^\gamma(i,\bar\sigma,\bar\delta)=\sup_{\bar\delta}\inf_{\bar\sigma} J^\gamma(i,\bar\sigma,\bar\delta), \tag{3}$$

$$v_i^\tau \;:=\; \inf_{\bar\sigma}\sup_{\bar\delta}\, J^\tau(i,\bar\sigma,\bar\delta)=\sup_{\bar\delta}\inf_{\bar\sigma}\, J^\tau(i,\bar\sigma,\bar\delta), \tag{4}$$

$$\rho_i \;:=\; \inf_{\bar\sigma}\sup_{\bar\delta}\, J(i,\bar\sigma,\bar\delta),=\sup_{\bar\delta}\inf_{\bar\sigma}\, J(i,\bar\sigma,\bar\delta), \tag{5}$$

for all initial states $i\in[n]$, where the infimum is taken among all strategies $\bar\sigma$ for MIN and the supremum is taken over all strategies $\bar\delta$ for MAX (we refer the reader to [18] for finite horizon or discounted infinite horizon games with constant discount factor, and to [14] for mean payoff games).

Optimal strategies for both players (together with the value of the game Γ for every initial state) can be obtained by the dynamic programming approach [18], which we next recall.

2.3. **Dynamic programming equations.** When considering finite horizon or mean-payoff games, we assume that the discount factor $\gamma(i,a,b)$ at every state and node is identically equal to 1, written $\gamma \equiv 1$. To handle in the same setting the discounted and the mean payoff cases, it will be convenient to consider the following unnormalized nonnegative cooefficients, rather than the transition probabilities:

$$M_{ij}^{ab} = \gamma(i,a,b) P_{ij}^{ab} \quad \forall i,j \in [n], a \in A_i,\ b \in B_i \ \ .$$

We wil also use the following notation, for all $i \in [n],\ a \in A_i,\ b \in B_i$ and $v \in \mathbb{R}^n$:

$$F(v;i,a,b) \ = \ \sum_{j\in[n]} M_{ij}^{ab} v_j \ + \ r_i^{ab}; \tag{6a}$$

$$F(v;i,a) \ = \ \max_{b\in B_i} F(v;i,a,b); \tag{6b}$$

$$F(v;i) \ = \ \min_{a\in A_i} F(v;i,a). \tag{6c}$$

The *dynamic programming* or *Shapley operator* associated to all above games is the self-map f of $\mathbb{R}^n$ given by:

$$[f(v)]_i := F(v;i), \qquad \forall i \in [n],\ v \in \mathbb{R}^n. \tag{7}$$

The value $v^\tau = (v_i^\tau)_{i\in[n]}$ of the finite horizon game satisfies the *dynamic programming equation* [18] associated to the operator f:

$$v^{\tau+1} \ = \ f(v^\tau), \qquad \tau = 0,1,\dots$$

with initial condition $v^0 \equiv 0$ ($v_i^0 = 0,\ i \in [n]$).

Similarly, the value $v^\gamma = (v_i^\gamma)_{i\in[n]}$ of the discounted infinite horizon game, with a discount factor $\gamma \ll 1$, is the unique solution $v \in \mathbb{R}^n$ of the (stationary) dynamic programming equation [18]:

$$v \ = \ f(v). \tag{8}$$

Also, optimal strategies are obtained for both players by taking pure Markovian stationary strategies σ for MIN and δ for MAX such that, for all $i \in [n]$, and $a \in A_i$, $\sigma(i)$ attains the minimum in the expression of $F(v;i)$ in (6c), and $\delta(i,a)$ attains the maximum in the expression of $F(v;i,a)$ in (6b).

The dynamic programming operator f is always *order-preserving*, i.e., $v \leq w \implies f(v) \leq f(w)$ where $\leq$ denotes the partial ordering of $\mathbb{R}^n$ ($v \leq w$ if $v_i \leq w_i$ for all $i \in [n]$). When $\gamma \leq 1$, f is also *additively subhomogeneous*, meaning that it satisfies $f(\lambda + v) \leq \lambda + f(v)$ for all $\lambda \in \mathbb{R}$ nonnegative ($\lambda \geq 0$) and $v \in \mathbb{R}^n$, where $\lambda + v := (\lambda + v_i)_{i\in[n]}$. This implies that f is nonexpansive in the sup-norm. When in addition Assumption (A2) holds, the map f is contracting in the sup-norm with contraction factor λ, that is:

$$\|f(v) - f(w)\| \leq \lambda \|v - w\| \ \ ,$$

where $\|\cdot\|$ denotes the sup-norm of $\mathbb{R}^n$ ($\|v\| = \max\{|v_i| \mid i \in [n]\}$). Then, one can solve the fixed point equation (8) of f by using the fixed point iterations, also called *value iterations* in the optimal control or game context: $v^{k+1} = f(v^k)$. They will converge geometrically towards the solution v with factor λ: $\lim_{k\to\infty} \|v^k - v\|^{1/k} \leq \lambda$. However the complexity of this algorithm is known to be only pseudo polynomial. Indeed the number of necessary iterations will depend on the norm of the solution, which depends itself on the modulus of the parameters.

When now $\gamma \equiv 1$, f is *additively homogeneous*, meaning that it commutes with the addition of a constant vector, i.e., that $f(\lambda + v) = \lambda + f(v)$ for all $\lambda \in \mathbb{R}$ and

$v \in \mathbb{R}^n$. Then, the mean payoff of the game can be studied through the following additive eigenproblem

$$\eta + v = f(v) \ . \tag{9}$$

Here, the vector $v \in \mathbb{R}^n$ is called an *additive eigenvector* of f associated to the *additive eigenvalue* $\eta \in \mathbb{R}$. If such an additive eigenpair exists, then, the value of the mean payoff game represented by f is equal to η for all initial states. Optimal strategies are obtained in the same way as for the discounted infinite horizon problem.

3. Policy iteration algorithm: presentation and preliminary properties

3.1. **Assumptions and notations.** Assume first that f is given by (7), with F as in (6), M_{ij}^{ab} nonnegative scalars, and A_i and B_i finite sets (Assumption (A1)). Then, the sets of feedback policies $A_{\mathrm{M}} := \{\sigma : [n] \to A \mid \sigma(i) \in A_i \, \forall i \in [n]\}$ and $B_{\mathrm{M}} := \{\delta : [n] \times A \to B \mid \delta(i,a) \in B_i \, \forall i \in [n],\, a \in A_i\}$ are finite. Now for every pair of policies $\sigma \in A_{\mathrm{M}}$ and $\delta \in B_{\mathrm{M}}$ of the first and second players, we define the following matrices and vectors:

$$M^{(\sigma\delta)} = (M_{ij}^{\sigma_i\delta_i})_{ij=1,\dots,n}, \quad \text{and } r^{(\sigma\delta)} = (r_i^{\sigma_i\delta_i})_{i=1,\dots,n} \ ,$$

and respectively the affine and nonlinear maps $f^{(\sigma\delta)}$ and $f^{(\sigma)}$ from $\mathbb{R}^n$ to itself which coordinates are given, for all $v \in \mathbb{R}^n$, by:

$$f_i^{(\sigma\delta)}(v) = F(v; i, \sigma_i, \delta_i), \tag{10a}$$

$$f_i^{(\sigma)}(v) = F(v; i, \sigma_i). \tag{10b}$$

Then, we can write, for all $v \in \mathbb{R}^n$,

$$f^{(\sigma\delta)}(v) = M^{(\sigma\delta)} v + r^{(\sigma\delta)}, \tag{11a}$$

$$f^{(\sigma)}(v) = \max_{\delta \in B_{\mathrm{M}}} f^{(\sigma\delta)}(v), \tag{11b}$$

$$f(v) = \min_{\sigma \in A_{\mathrm{M}}} f^{(\sigma)}(v) \ , \tag{11c}$$

where in these expressions, the maximum and the minimum mean the supremum and infimum with respect to the partial order of $\mathbb{R}^n$. Note that it is attained for an element of B_{M} and A_{M} respectively. Indeed, from (10), the i-th entry of $f^{(\sigma\delta)}$ and $f^{(\sigma)}$ depends only on the policy at state i. We shall say that a set of vectors $V \subset \mathbb{R}^n$ is *rectangular* if $V = \pi_1(V) \times \cdots \times \pi_n(V)$, where $\pi_i : \mathbb{R}^n \to \mathbb{R}$ denotes the projection on the ith coordinates. It follows that the set of vectors $\{f^{(\sigma)}(v) \mid \sigma \in A_{\mathrm{M}}\}$ is rectangular, and that for each σ, the set $\{f^{(\sigma\delta)}(v) \mid \delta \in B_{\mathrm{M}}\}$ is also rectangular.

The maps $f^{(\sigma\delta)}$ and $f^{(\sigma)}$ satisfy the same properties as the ones stated in Section 2.3 for f. They are all order preserving. When $\gamma \leq 1$ (resp. $\gamma \equiv 1$), they are additively subhomogeneous (resp. homogeneous), hence nonexpansive in the sup-norm. When Assumption (A2) holds, these maps are contracting in the sup-norm with contraction factor λ.

3.2. **Policy iteration algorithm for discounted games.** Here we are interested in solving Equation (8) by using the policy iteration algorithm for discounted games, introduced by Howard [13] for 1-player games, and by Hoffman and Karp [12], and Denardo [7] for 2-player games. It will be convenient to consider the following general algorithm.

Algorithm 1 (General policy iteration algorithm)**.**
Input: A set A_{M}, and maps f and $f^{(\sigma)}$, from $\mathbb{R}^n$ to itself, for $\sigma \in A_{\mathrm{M}}$, satisfying (11c), for all $v \in \mathbb{R}^n$.
Output: A fixed point v of f and a policy $\sigma \in A_{\mathrm{M}}$ such that $f(v) = f^{(\sigma)}(v)$.

(1) *Initialization*: Set $s = 0$. Select an arbitrary strategy $\sigma^0 \in A_{\mathrm{M}}$.
(2) Compute the fixed point v^s of $f^{(\sigma^s)}$.
(3) Improve the policy: choose an optimal policy for v^s, that is $\sigma^{s+1} \in A_{\mathrm{M}}$ such that $f(v^s) = f^{(\sigma^{s+1})}(v^s)$, with $\sigma^{s+1} = \sigma^s$ as soon as this is possible.
(4) If $\sigma^{s+1} = \sigma^s$, then the algorithm stops and returns v^s and σ^s. Otherwise, increment s by one and go to Step 2.

When A_{M} is as in Section 3.1, and $f^{(\sigma)}$ satisfies also (10b), Step 3 is equivalent to

$$\sigma_i^{s+1} \in \operatorname*{argmin}_{a \in A_i} F(v^s; i, a), \quad i \in [n],$$

and we can also choose σ^{s+1} in a conservative way, that is such that, for all $i \in [n]$, $\sigma_i^{s+1} = \sigma_i^s$ as soon as this is possible. Algorithm 1 can also be applied to a map f satisfying (11c) with the min operation replaced by the max operation.

When f is as in Section 3.1, the policy iteration algorithm for 2-player games consists in two levels of nested instances of the previous algorithm.

Algorithm 2 (Policy iteration algorithm for 2-player games)**.**
Input: A map f given as in Section 3.1.
Output: The value v of the game associated to f and an optimal policy $\sigma \in A_{\mathrm{M}}$.

- Apply Algorithm 1 (that is construct the sequences σ^s of policies and v^s of values, $s \geq 0$).
- The solution v^s in Step 2 is the value of the game with fixed policy σ^s. It is computed as follows:
  - Apply Algorithm 1 to the set B_{M} instead of A_{M}, the map $f^{(\sigma_s)}$ instead of f and the maps $f^{(\sigma^s \delta)}$ with $\delta \in B_{\mathrm{M}}$ instead of the maps $f^{(\sigma)}$ with $\sigma \in A_{\mathrm{M}}$. This constructs sequences of policies $\delta^{s,l}$ and values $v^{s,l}$, with $l \geq 0$.
  - When the latter algorithm stops, put $v^s = v^{s,l}$. Then $\delta^{s,l}$ is an optimal policy of the second player of the game with fixed policy σ^s for the first player.
- When the algorithm stops, return v^s, σ^s and $\delta^{s,l}$ with s equal to the final index of the external iteration of Algorithm 1, and l the final index of the internal iteration of Algorithm 1.

Note that in the nested application of Algorithm 1, Step 2 consists in solving a linear system, which can be done either by a direct linear solver, or approximately, by an iterative method. In the present paper, we require an exact solution.

The usual assumption for the validity of the above algorithms is Assumption (A2). Under this assumption, one can show (see for instance [4] for one-player games and [7] for 2-player games) that the sequence of values $(v^s)_{s\geq 0}$ (resp. $(v^{s,l})_{l\geq 0}$ for some fixed $s\geq 0$) of Algorithm 2 is nonincreasing (resp. nondecreasing) and converges towards the unique fixed point v of f (resp. v^s of $f^{(\sigma^s)}$), and deduce in particular that Algorithm 2 (resp. each nested application of Algorithm 1 in Algorithm 2) never visits twice the same policy of the first (resp. second) player (except before stopping). Then, since the action spaces are finite, the algorithm (resp. nested policy iterations) stops after a finite time.

These properties can indeed be obtained from the following result concerning Algorithm 1.

Proposition 1. *Let A_{M}, f and $f^{(\sigma)}$ be as in Algorithm 1. Assume that A_{M} is finite, that the maps $f^{(\sigma)}$ are order preserving and contracting in the sup-norm with the same contraction factor λ. We have:*

1. *f is also order preserving and contracting in the sup-norm with contraction factor λ;*
2. *the iterations of Algorithm 1 are well defined;*
3. *the sequence v^s is nonincreasing and converges towards the unique fixed point v of f;*
4. *more precisely: $v\leq v^{s+1}\leq f(v^s)\leq v^s$;*
5. *the sequence σ^s never visits twice the same policy (except when the stopping condition is satisfied);*
6. *hence Algorithm 1 stops after a finite time.*

From Point 4, we see that the sequence $(v^s)_{s\geq 0}$ of policy iteration algorithm 2 converges faster towards v than the value iteration algorithm starting from v^0. One can also deduce the following contraction property (see for instance [11]).

Corollary 2. *Under the assumptions of Proposition 1, the sequence v^s satisfies the following contraction property in the sup-norm: $\|v^{s+1}-v\|\leq\lambda\|v^s-v\|$.*

3.3. **Policy iteration algorithm for mean-payoff games.** Now, we assume that $\gamma\equiv 1$, and are interested in solving the optimality equation of the mean payoff problem, Equation (9), by policy iteration. The first algorithm doing so was introduced by Hoffman and Karp [12], assuming all the matrices $M^{(\sigma\delta)}$ to be irreducible. This algorithm is very similar to the algorithm for discounted games, so we only present here the differences.

Algorithm 3 (General policy iteration algorithm for the mean payoff additive eigenproblem)**.**

Same as Algorithm 1, except that

- f and $f^{(\sigma)}$ are assumed to be additively homogeneous;
- the initialization includes the selection of an arbitrary state $c\in[n]$;
- instead of a fixed point v of f, the algorithm is returning an additive eigenvector v and associated eigenvalue η of f ($\eta+v=f(v)$) such that $v_c=0$;
- the computation of a fixed point v^s of $f^{(\sigma^s)}$ is replaced by the computation of an additive eigenvector v^s and associated eigenvalue η^s of $f^{(\sigma^s)}$ ($\eta^s+v^s=f^{(\sigma^s)}(v^s)$), such that $v^s_c=0$.

Note that since the maps f and $f^{(\sigma)}$ are additively homogeneous, changing c into another state $\tilde{c}$ does not change the admissible sequences of policies, and only modifies the additive eigenvectors by additive constants. Indeed, for any given input, σ^s, η^s and v^s are respectively admissible sequences of policies, additive eigenvalues, and additive eigenvectors with c, if and only if σ^s, η^s and $\tilde{v}^s = v^s - v^s_{\tilde{c}}$ are respectively admissible sequences of policies, additive eigenvalues, and additive eigenvectors with $\tilde{c}$.

Algorithm 4 (Hoffman and Karp policy iteration algorithm for 2-player mean–payoff games)**.**

Same as Algorithm 2, except that

- we assume that $\gamma \equiv 1$;
- instead of the value v of the game associated to f, the algorithm is returning the value η and a bias v such that $v_c = 0$;
- Algorithm 1 is replaced by Algorithm 3;
- the algorithm constructs the sequences of values η^s and bias v^s of the game with a fixed policy σ^s, with $v^s_c = 0$, and for each $s \geq 0$, it constructs the sequences of values $\eta^{s,l}$ and bias $v^{s,l}$, $l \geq 0$, of the game with fixed policies σ^s and $\delta^{s,l}$ of the first and second player, with $v^{s,l}_c = 0$.

Some or all of the following properties can be found in [4] for one-player games and [12] for 2-player games.

Proposition 3. *Assume all the matrices $M^{(\sigma\delta)}$, $\sigma \in A_{\mathrm{M}}$, $\delta \in B_{\mathrm{M}}$, are irreducible. Then, the sequence of values $(\eta^s)_{s\geq 0}$ (resp. $(\eta^{s,l})_{l\geq 0}$ for some fixed $s \geq 0$) of Algorithm 4 is nonincreasing (resp. nondecreasing) and converges towards the unique eigenvalue η of f (resp. η^s of $f^{(\sigma^s)}$). Also the sequence of bias $(v^s)_{s\geq 0}$ (resp. $(v^{s,l})_{l\geq 0}$ for some fixed $s \geq 0$) of Algorithm 4 converges towards the unique bias v of f such that $v_c = 0$ (resp. v^s of $f^{(\sigma^s)}$ such that $v^s_c = 0$).*

Corollary 4. *Assume all the matrices $M^{(\sigma\delta)}$, $\sigma \in A_{\mathrm{M}}$, $\delta \in B_{\mathrm{M}}$, are irreducible. Then, Algorithm 4 (resp. each nested application of Algorithm 3 in Algorithm 4) never visits twice the same policy of the first (resp. second) player (except when the stopping condition is verified). Hence, the policy iterations (resp. nested policy iterations) stop after a finite time.*

Note that the above algorithms cannot be applied to multichain games (such that some matrices $M^{(\sigma\delta)}$ have at least two final classes), since then, the value of the game is not any more given by a constant η independent of the initial state. See [5, 1] for a discussion of the multichain case.

4. Bounds on the number of policy iterations

In the sequel, we shall state as far as possible our results in the framework of the general policy iteration algorithms 1 and 3, the application of these results to the zero-sum two-player game policy iteration algorithms 2 and 4 being immediate.

4.1. Revisiting the bound of Ye and Hansen, Miltersen and Zwick with non linear maps.

The following improvement of [11] is obtained by the same arguments as in [11], except that we use the nonlinear maps $f^{(\sigma)}$ directly instead of the affine maps $f^{(\sigma\delta)}$, and that we use only sup-norms, whereas ℓ^1 norms were used in some places in [11].

Theorem 5. *Let A_{M}, f, $f^{(\sigma)}$ and λ be as in Proposition 1. Assume also that A_{M} is as in Section 3.1, and that $f^{(\sigma)}$ satisfies* (10b).

Then the policy iteration algorithm 1 stops after at most $s_{\max}$ iterations, where $s_{\max} := (m_1 - n)(1 + \lfloor \frac{\log(1-\lambda)}{\log(\lambda)} \rfloor)$ and m_1 is the cardinality of $\mathsf{SA} := \{(i, a) \mid i \in [n],\ a \in A_i\}$.

4.2. **Discounted games with state dependent discount factors.** We denote by $r(M)$ the spectral radius of a $n \times n$ matrix M, that is the maximum of the moduli of its eigenvalues. When $\varphi \in \mathbb{R}^n$ has strictly positive coordinates and $v \in \mathbb{R}^n$, we set $\varphi^{-1} := (\varphi_i^{-1})_{i \in [n]} \in \mathbb{R}^n$ and $\varphi v := (\varphi_i v_i)_{i \in [n]} \in \mathbb{R}^n$ (these are the usual notations, if we identify $\mathbb{R}^n$ to the set of functions from $[n]$ to $\mathbb{R}$). For all self-maps f of $\mathbb{R}^n$, we denote by $\mathcal{S}_\varphi(f)$ its scaling by φ, which is the map $\mathcal{S}_\varphi(f) : v \mapsto \varphi^{-1} f(\varphi v)$. It is easy to see that if f is order preserving so is $\mathcal{S}_\varphi(f)$.a The following result shows that these scalings leave invariant the sequences of policies generated by the policy iteration algorithm. A sequence of policies and fixed points will be said to be *admissible* for a given input if there is a valid run of the algorithm on this input producing this sequence.

Proposition 6 (Scaling Invariance). *Let A_{M}, f and $f^{(\sigma)}$ be as in Algorithm 1, and let $\varphi \in \mathbb{R}^n$ have strictly positive coordinates. Denote $\tilde{f} := \mathcal{S}_\varphi(f)$ and $\tilde{f}^{(\sigma)} = \mathcal{S}_\varphi(f^{(\sigma)})$. Then, A_{M}, $\tilde{f}$ and $\tilde{f}^{(\sigma)}$ constitute a valid input of Algorithm 1. Moreover, σ^s and $\tilde{v}^s = \varphi^{-1} v^s$ constitute an admissible sequence of policies and fixed points for this input, if and only if σ^s and v^s constitute an admissible sequence of policies and fixed points for the original input A_{M}, f and $f^{(\sigma)}$.*

For a set $\mathcal{M}$ of $n \times n$ matrices $\mathcal{M} \subset \mathbb{R}^{n \times n}$, we shall define its *rectangular hull*, denoted $\mathrm{rec}(\mathcal{M})$, as the set of matrices N such that, for all $i \in [n]$, the row i of N coincides with the row i of some element M of $\mathcal{M}$. When g is a polyhedral self-map of $\mathbb{R}^n$, so that $\mathbb{R}^n$ can be covered by finitely many polyhedra on which g is affine, we shall denote by $\mathrm{imD}(g)$ the finite set of matrices representing the differential of g in each of these polyhedra.

The proof of the following result is based on nonlinear Perron-Frobenius theory and in particular on some results in [16, 2].

Theorem 7. *Let A_{M}, f and $f^{(\sigma)}$ be as in Algorithm 1. Assume that A_{M} is as in Section 3.1, that $f^{(\sigma)}$ satisfies* (10b), *and that the maps $f^{(\sigma)}$ are order preserving and polyhedral. Let $\mathcal{M}(\sigma) = \mathrm{rec}(\mathrm{imD}(f^{(\sigma)}))$ and $\mathcal{M} = \cup_{\sigma \in A_{\mathrm{M}}} \mathcal{M}(\sigma)$. Assume that the spectral radii of all the matrices M in $\mathcal{M}$ are strictly less than 1, and denote by ω the maximum of these spectral radii. Then for all λ such that $\omega < \lambda < 1$, there exists $\varphi \in \mathbb{R}^n$ with strictly positive coordinates such that the scaled maps $\tilde{f} := \mathcal{S}_\varphi(f)$ and $\tilde{f}^{(\sigma)} = \mathcal{S}_\varphi(f^{(\sigma)})$ are contracting in the sup-norm with contraction factor λ.*

Using Theorem 7 and Proposition 6, we obtain:

Corollary 8. *Under the assumptions of Theorem 7, the conclusion of Proposition 1 holds.*

Applying Theorem 7, Proposition 6, and Theorem 5 to all λ such that $\omega < \lambda < 1$, we obtain:

Corollary 9. *Under the assumptions of Theorem 7, the conclusion of Theorem 5 holds with $\lambda = \omega$.*

Corollary 10. *Let A_{M}, B_{M} and f be given as in Section 3.1. Assume that the spectral radii of all the matrices $M^{(\sigma\delta)}$, $\sigma \in A_{\mathrm{M}}$, $\delta \in B_{\mathrm{M}}$, are strictly less than*

1, so that $\bar{\omega} := \max_{\sigma\in A_{\mathrm{M}},\delta\in B_{\mathrm{M}}} r(M^{(\sigma\delta)}) < 1$. Then the conclusion of Theorem 5 holds for the policy iteration algorithm for 2-player games, Algorithm 2 (instead of Algorithm 1), with $\lambda = \bar{\omega}$.

Corollary 11. *Let $\lambda \in [0, 1)$ be fixed. Then, the policy iteration algorithm solves in strongly polynomial time the instances of zero-sum 2-player "discounted" stochastic games with perfect information and state dependent discount factors (possibly locally greater than 1) that are such that the spectral radii of the transition matrices associated to every pair of policies of the two players is bounded by λ.*

4.3. **Mean-payoff games with a renewal state.** For a Markov matrix M and states i, j, we shall denote:

$$\mathcal{T}_{ij}(M) = \mathbb{E}[\inf\{k \geq 1 \mid X_k = j\} \mid X_0 = i] \enspace ,$$

the expected first mean return time to state j of a Markov chain X_k with transition matrix M and initial state i. It is easy to see that $\mathcal{T}_{ic}(M) < +\infty$ for all $i \in X$ if and only if M has a unique final (recurrent) class and that c belongs to this class. The state c is called a *renewal state*.

The following transformation will allow us to replace a self-map f of $\mathbb{R}^n$ by a sup-norm contraction. This will play a similar role to the scaling transformation used in the discounted case.

Let $\varphi \in \mathbb{R}^n$ have positive coordinates and $c \in [n]$. Then, the map L_φ which to a couple (η, v), with $\eta \in \mathbb{R}$ and $v \in \mathbb{R}^n$ such that $v_c = 0$, associates the vector $w = \eta + \varphi^{-1}v \in \mathbb{R}^n$ is an affine isomorphism, with inverse given by: $\eta = w_c$ and $v = \varphi(w - w_c)$. For all self-maps f of $\mathbb{R}^n$, we shall denote by $\mathcal{L}_\varphi(f)$ the self-map of $\mathbb{R}^n$, such that for all $w, v \in \mathbb{R}^n$ and $\eta \in \mathbb{R}$ with $v_c = 0$ and $w = \eta + \varphi^{-1}v$, we have $\mathcal{L}_\varphi(f)(w) = \varphi^{-1}(\eta(\varphi - 1) + f(v))$.

Proposition 12. *Let A_{M}, f, and $f^{(\sigma)}$ be as in Algorithm 3, and let $\varphi \in \mathbb{R}^n$ have strictly positive coordinates and $c \in [n]$. Denote $\tilde{f} := \mathcal{L}_\varphi(f)$ and $\tilde{f}^{(\sigma)} = \mathcal{L}_\varphi(f^{(\sigma)})$. Then, A_{M}, $\tilde{f}$ and $\tilde{f}^{(\sigma)}$ are a valid input of Algorithm 1. Moreover σ^s and $\tilde{v}^s = \eta^s + \varphi^{-1}v^s$ constitute an admissible sequence of policies and fixed points for Algorithm 1 on this input, if and only if σ^s, η^s and v^s constitute an admissible sequence of policies, additive eigenvalues and additive eigenvectors for Algorithm 3 on the original input A_{M}, f, $f^{(\sigma)}$, when c is chosen.*

Theorem 13. *Let A_{M}, f, and $f^{(\sigma)}$ be as in Algorithm 3. Assume that A_{M} is as in Section 3.1, that $f^{(\sigma)}$ satisfies* (10b), *and that the maps $f^{(\sigma)}$ are order preserving and polyhedral. Let $\mathcal{M}(\sigma) = \mathrm{rec}(\mathrm{imD}(f^{(\sigma)}))$ and $\mathcal{M} = \cup_{\sigma\in A_{\mathrm{M}}}\mathcal{M}(\sigma)$. Then, all matrices M in $\mathcal{M}$ are Markov matrices. Assume that they all have a unique final class, and there there is a state $c \in [n]$ which is common to each of these classes, so that*

$$\mathcal{T}_{ic} := \max_{M\in\mathcal{M}} \mathcal{T}_{ic}(M) < +\infty \quad \forall i \in [n] \enspace .$$

Let $\varphi \in \mathbb{R}^n$ be the vector with coordinates $\varphi_i = \mathcal{T}_{ic} \geq 1$, and $K = \max_{i\in[n]} \mathcal{T}_{ic}$. Then, the transformed maps $\tilde{f} := \mathcal{L}_\varphi(f)$ and $\tilde{f}^{(\sigma)} = \mathcal{L}_\varphi(f^{(\sigma)})$ are order-preserving and contracting in the sup-norm with contraction factor $\lambda = (K - 1)/K$.

Corollary 14. *Under the assumptions of Theorem 13, Assertions 2,3,5, and 6 of Proposition 1 hold for Algorithm 3 instead of Algorithm 1, with v^s replaced by $\eta^s + \varphi^{-1}v^s$, and v replaced by $\eta + \varphi^{-1}v$.*

Applying Theorem 13, Proposition 12, and Theorem 5, we obtain:

Corollary 15. *Under the assumptions of Theorem 13, the policy iteration algorithm 3 stops after at most $s_{\max}$ iterations, where $s_{\max} := (m_1-n)(1+\lfloor\frac{\log(K)}{\log(K/(K-1))}\rfloor) = \mathcal{O}((m_1-n)K\log K)$, $K=\max_{i\in[n]}\mathcal{T}_{ic}$, and m_1 is the cardinality of SA.*

Corollary 16. *Let $A_{\rm M}$ and f be given as in Section 3.1, with $\gamma\equiv 1$. Assume that every matrix $M^{(\sigma\delta)}$, $\sigma\in A_{\rm M}$, $\delta\in B_{\rm M}$, has a unique final class, and there is a state $c\in[n]$ which is common to each of these classes, so that*

$$\bar{\mathcal{T}}_{ic} := \max_{\sigma\in A_{\rm M},\delta\in B_{\rm M}} \mathcal{T}_{ic}(M^{(\sigma\delta)}) < +\infty \quad \forall i\in[n] \enspace .$$

Then, the conclusion of Corollary 15 holds for the Hoffman and Karp policy iteration algorithm for 2-player mean-payoff games, Algorithm 4 (instead of Algorithm 3) with $K=\max_{i\in[n]}\bar{\mathcal{T}}_{ic}$.

Corollary 17. *Let $K\in[1,+\infty)$ be fixed. Then, the Hoffman and Karp policy iteration algorithm solves in strongly polynomial time the instances of zero-sum 2-player stochastic mean-payoff games with perfect information having a distinguished state to which the mean return time is bounded by K for all choices of policies of both players.*

5. Proof of the preliminary results of Section 3

The following proof is similar to the proofs of the same properties for Algorithm 2 that can be found for instance in [4].

Proof of Proposition 1. Let $A_{\rm M}$, f and $f^{(\sigma)}$ be as in the proposition. From (11c), f is order preserving and contracting in the sup-norm with contraction factor λ.

Hence the maps $f^{(\sigma)}$ and f have a unique fixed point, which implies that Step 2 of Algorithm 1 is well defined.

Let $(v^s)_{s\geq 1}$ be the sequence of Algorithm 1. We have $v^s = f^{(\sigma^s)}(v^s) \geq f(v^s) = f^{(\sigma^{s+1})}(v^s)$. In particular, $v^s \geq f^{(\sigma^{s+1})}(v^s)$, which implies that the sequence $(f^{(\sigma^{s+1})})^k(v^s)$ is nonincreasing. By the fixed point theorem for the contracting map $f^{(\sigma^{s+1})}$, the former sequence converges towards the unique fixed point, which by definition is v^{s+1}. This implies in particular that $v^s\geq f^{(\sigma^{s+1})}(v^s)\geq v^{s+1}$, so that the sequence $(v^s)_{s\geq 1}$ is nonincreasing.

Moreover, from the above equations, we deduce that $v^s\geq f(v^s)\geq v^{s+1}$. In particular, the sequence $f^k(v^s)$ is nonincreasing. Again, by the fixed point theorem for the contracting map f, the former sequence converges towards the unique fixed point v of f, hence $v^s\geq v$ for all s. Since the sequence $(v^s)_{s\geq 1}$ is nonincreasing and lower bounded by v, it converges towards some vector $w\geq v$. Then, from the above equations, we also get that $v^s\geq f(v^s)\geq v^{s+1}\geq v$, for all s, passing to the limit and using the continuity of f, we deduce that $w=f(w)$, and since f has a unique fixed point, we deduce that $w=v$.

Assume by contradiction that the sequence σ^s visits twice the same policy. This means that $\sigma^{s'}=\sigma^s$ for some $s'>s\geq 0$. Since the map $f^{(\sigma^s)}=f^{(\sigma^{s'})}$ has a unique fixed point, we get that $v^s=v^{s'}$. Since we already proved that the sequence $(v^s)_{s\geq 0}$ is nonincreasing, we obtain $v^s\geq v^{s+1}\geq v^{s'}$. This implies that $v^s=v^{s+1}$, hence $v^s=f(v^s)$, so that, by definition, the algorithm necessarily stops at iteration s if it did not stopped before, hence the iteration s' does not occur, and $\sigma^{s'}$ is computed only if $s' = s+1$ and so the algorithm cannot visits twice the same policy, except when the stopping condition is verified.

This implies that Algorithm 1 stops after at most a number of iterations equal to the cardinality of the set A_{M}. □

Proof of Corollary 2. From $v^s \geq f(v^s) \geq v^{s+1} \geq v$, we get that $\|v^{s+1} - v\| \leq \|f(v^s) - v\|$ and since f is contracting with factor λ, we deduce that $\|v^{s+1} - v\| \leq \lambda \|v^s - v\|$. □

6. Proof of Theorem 5

Let v denote the unique fixed point of f, and for all $\sigma \in A_{\mathrm{M}}$, denote by $R^{(\sigma)} = f^{(\sigma)}(v) - v$ the residual induced by v on the fixed point equation of $f^{(\sigma)}$. By (11c), we have $\min_{\sigma \in A_{\mathrm{M}}} R^{(\sigma)} = 0$, so $R^{(\sigma)} \geq 0$ for all $\sigma \in A_{\mathrm{M}}$. Moreover, σ is an optimal policy of the game if and only if $R^{(\sigma)} = 0$, or equivalently $\|R^{(\sigma)}\| = 0$. Finally, by (10b), $R_i^{(\sigma)} = R_i^{\sigma_i}$ for all $i \in [n]$, where $R_i^a = F(v; i, a) - v_i$ for all $i \in [n]$ and $a \in A_i$ plays the role of a new reward such that the value of the dynamic programming equation is identically equal to zero.

Let v^s and σ^s be the sequences of values and policies constructed in Algorithm 1. Since $v^s \geq v$, $v^s = f^{(\sigma^s)}(v^s)$, and $f^{(\sigma^s)}$ is order preserving, we get that $v^s \geq f^{(\sigma^s)}(v) \geq f(v) = v$, hence $0 \leq R^{(\sigma^s)} \leq v^s - v$ and taking the supremum over all coordinates (or states), we get that $\|R^{(\sigma^s)}\| \leq \|v^s - v\|$. Now, since v^s is the unique fixed point of the λ-contracting map $f^{(\sigma^s)}$, we get that $\|v^s - f^{(\sigma^s)}(v)\| \leq \lambda \|v^s - v\|$. Then, $\|v^s - v\| \leq \|v^s - f^{(\sigma^s)}(v)\| + \|R^{(\sigma^s)}\| \leq \lambda \|v^s - v\| + \|R^{(\sigma^s)}\|$. From all the above inequalities, we get that

$$\|R^{(\sigma^s)}\| \leq \|v^s - v\| \leq \frac{1}{1-\lambda} \|R^{(\sigma^s)}\| \ .$$

Combining these inequalities with the contraction of policy iterations shown in Corollary 2 ($\|v^{s+1} - v\| \leq \lambda \|v^s - v\|$), we obtain that for all $t \geq s + p$,

$$\|R^{(\sigma^t)}\| \leq \mu \|R^{(\sigma^s)}\|, \quad \text{with} \quad \mu = \frac{1}{1-\lambda} \lambda^p \ .$$

Moreover when $p = 1 + \lfloor \log(1-\lambda)/\log(\lambda) \rfloor$ (which is the least integer such that $p > \log(1-\lambda)/\log(\lambda)$), we have $\mu < 1$.

For all $\sigma \in A_{\mathrm{M}}$, let us denote by $\mathcal{G}(\sigma)$ the graph of σ: $\mathcal{G}(\sigma) = \{(i, \sigma_i) \mid i \in [n]\}$. Since $R_i^{(\sigma)} = R_i^{\sigma_i}$ for all $i \in [n]$, we get that $\|R^{(\sigma)}\| = \max_{(i,a) \in \mathcal{G}(\sigma)} R_i^a$. Assume σ^s is not optimal, then $\|R^{(\sigma^s)}\| > 0$ and let (i,a) realizes the maximum of R_i^a on $\mathcal{G}(\sigma^s)$. If $t \geq s+p$, with p as before, and $(i,a) \in \mathcal{G}(\sigma^t)$, we get that $R_i^a \leq \|R^{(\sigma^t)}\| \leq \mu \|R^{(\sigma^s)}\| = \mu R_i^a$ with $\mu < 1$ and $R_i^a > 0$, which is impossible. This shows that $(i,a) \notin \mathcal{G}(\sigma^t)$, hence $\mathcal{G}(\sigma^t) \subset \mathsf{SA} \setminus \{(i,a)\}$, for all $t \geq s + p$. Let us construct a sequence SA_s of subsets of SA, equal to the empty set for all $s < p$, and such that for all $s \geq p$, SA_s is the union of SA_{s-1} with the set of couples (i,a) realizing the maximum of R_i^a on $\mathcal{G}(\sigma^{s-p})$. We get that $\mathcal{G}(\sigma^t) \subset \mathsf{SA} \setminus \mathsf{SA}_t$, for all $t \geq 0$ and that for all $s \geq p$, there exist $(i,a) \in \mathsf{SA}_s \setminus \mathsf{SA}_{s-p}$, as long as Algorithm 1 did not stop, so that the cardinality of SA_s increases at least by one after each group of p iterations. Hence, $\mathsf{SA} \setminus \mathsf{SA}_{p(m_1-n)}$ has at most n elements, and since, for all $t \geq p(m_1 - n)$, $\mathcal{G}(\sigma^t) \subset \mathsf{SA} \setminus \mathsf{SA}_{p(m_1-n)}$ and $\mathcal{G}(\sigma^t)$ has exactly n elements, we deduce that, if the algorithm did not stop before iteration number t, there is only one choice for $\mathcal{G}(\sigma^t)$

with $t \geq p(m_1 - n)$, hence $\sigma^t = \sigma^{t+1}$, and the algorithm stops at iteration number t.

7. Spectral radius notions and the results of Section 4.2

Let C be a closed convex cone of $\mathbb{R}^n$, let $\overset{\circ}{C}$ denote its interior, and let h be a nonlinear continuous positively homogeneous map from C to itself ($h(\lambda v) = \lambda h(v)$ for all $\lambda > 0$ and $v \in C$). The following definitions are taken from [15]:

- v is an eigenvector of h in C, and λ is an eigenvalue associated to v, if $h(v) = \lambda v$.
- The *cone eigenvalue spectral radius* of h is the supremum of its eigenvalues in C:
$$\hat{r}_C(h) := \sup\{\lambda \geq 0 \mid \exists v \in C\backslash\{0\} \text{ such that } h(v) = \lambda v\} \enspace .$$
- The *Collatz-Wielandt number* of h is defined as:
$$\mathrm{cw}_C(h) := \inf\{\lambda > 0 \mid \exists v \in \overset{\circ}{C} \text{ such that } h(v) \leq \lambda v\} \enspace .$$
- The *Bonsall's spectral radius* of h is defined as:
$$r_C(h) := \inf_{k\geq 1} \|h^k\|_C^{1/k}, \quad \text{with} \quad \|h\|_C := \sup_{x\in C,\ \|x\|=1} \|h(x)\| \enspace ,$$
for any given norm $\|\cdot\|$ on $\mathbb{R}^n$.

The equality $\hat{r}_{\mathbb{R}_+}(h) = \mathrm{cw}_{\mathbb{R}_+}(h)$ in the following result was established by Nussbaum [16, Theorem 3.1]. The last equality is done in [2] in a more general infinite dimensional context, together with the first one.

Theorem 18 ([16, Theorem 3.1], and [2])**.** *For a continuous, positively homogeneous, order preserving selfmap h of $C = \mathbb{R}^n_+$, all the above spectral radius notions of h coincide:*
$$\hat{r}_{\mathbb{R}_+}(h) = \mathrm{cw}_{\mathbb{R}_+}(h) = r_{\mathbb{R}_+}(h) \enspace .$$
We denote by $r(h)$ this constant.

The following result can be deduced easily from Theorem 18. It is also proved in an infinite dimensional context in [2].

Proposition 19. *Let Π be a set, and h and h_π, $\pi \in \Pi$, be continuous, positively homogeneous, order preserving selfmaps of $\mathbb{R}^n_+$. Assume that for all $v \in \mathbb{R}^n_+$, $h(v) = \max_{\pi\in\Pi} h_\pi(v)$, meaning that $h(v) \geq h_\pi(v)$ for all $\pi \in \Pi$, and that there exists $\pi \in \Pi$ such that $h(v) = h_\pi(v)$. Then*
$$r(h) = \max_{\pi\in\Pi} r(h_\pi) \enspace .$$

Proof of Theorem 7. Since f is order preserving, so is $\mathcal{S}_\varphi(f)$. If f is a polyhedral map such that all the matrices $M \in \mathrm{imD}(f)$ satisfy $M\varphi \leq \lambda\varphi$, then, all the matrices $M' \in \mathrm{imD}(\mathcal{S}_\varphi(f))$ satisfy $M'\mathbb{1} = \varphi^{-1}M\varphi \leq \lambda\mathbb{1}$, where $\mathbb{1}$ is the vector with all coordinates equal to 1. Then, since M' has also nonnegative coordinates, because $\mathcal{S}_\varphi(f)$ is order preserving, we get that M' is contracting in the sup-norm with contraction factor λ. Then, using the polyhedral and continuity properties of f, it is easy to see that $\mathcal{S}_\varphi(f)$ is also contracting in the sup-norm with contraction factor λ.

Let us show the above property for all maps $f^{(\sigma)}$. For this, consider the self-map $\bar{f}$ of $\mathbb{R}^n$ given by:

$$\bar{f}(v) := \sup_{M \in \mathcal{M}} (Mv) \ , \tag{12}$$

where $\mathcal{M}$ is as in the theorem. Since all the matrices involved in the previous formula have nonnegative entries, the corresponding self-maps of $\mathbb{R}^n$ are order-preserving. Since A_{M} is finite and the maps $f^{(\sigma)}$ are polyhedral, the set $\mathcal{M}$ is finite. Since in addition A_{M} is as in Section 3.1, the set $\mathcal{M}$ is the Cartesian product of the sets of its rows, hence the supremum in (12) is a maximum. Then, applying Proposition 19, we get that

$$r(\bar{f}) = \max_{M \in \mathcal{M}} r(M) = \omega.$$

In particular the maximum is attained, hence < 1. Now, from Theorem 18, we get that $\omega = r(\bar{f}) = \mathrm{cw}_{\mathbb{R}_+}(\bar{f})$, hence for all $\lambda > \omega$, there exists $\varphi \in \mathbb{R}^n$ with positive coefficients such that $\bar{f}(\varphi) \leq \lambda\varphi$. This implies that all the matrices $M \in \mathrm{imD}(f^{(\sigma)})$ satisfy $M\varphi \leq \lambda\varphi$, which by the above arguments implies that $\mathcal{S}_\varphi(f^\sigma)$ is contracting in the sup-norm with contraction factor λ. □

Proof of Proposition 6. By definition of $\mathcal{S}_\varphi(f)$, we have $v = f(v)$ if and only if $w = \mathcal{S}_\varphi(f)(w)$ for $w = \varphi^{-1}v$. Moreover, the transformation of maps f, $\mathcal{S}_\varphi(f)(w) = \varphi^{-1}f(\varphi w)$ preserves the order on the maps f. Hence if v^s and σ^s are respectively sequences of fixed points, and policies of Algorithm 1 for f and $f^{(\sigma)}$, then $w^s = \varphi^{-1}v^s$ and σ^s are respectively sequences of fixed points and policies of Algorithm 1 for $\mathcal{S}_\varphi(f)$ and $\mathcal{S}_\varphi(f^{(\sigma)})$. □

8. The results of Section 4.3

Lemma 20. *Let M be a $n \times n$ Markov matrix with a unique final class, and let $c \in [n]$ belong to this final class. Denote by $M_{(c)}$ the matrix obtained from M by putting to zero all entries in the c-th column. Then, the vector $\varphi \in \mathbb{R}^n$ with coordinates $\varphi_i = \mathcal{T}_{ic}(M)$ satisfies $\varphi = 1 + M_{(c)}\varphi$.*

Lemma 21. *Let M be a $n \times n$ Markov matrix with a unique final class, and let $c \in [n]$ belong to this final class. Consider a vector $\varphi \in \mathbb{R}^n$ with positive coordinates such that $\varphi \geq 1 + M_{(c)}\varphi$, and let K be a bound on its coefficients, $K \geq \|\varphi\|$. Construct the $n \times n$ matrix $M_{(c,\varphi)}$ by replacing the c-th column of M with the vector $(1/\varphi_c)(\varphi - 1 - M_{(c)}\varphi)$. Then, $M_{(c,\varphi)}$ has nonnegative entries and satisfies*

$$M_{(c,\varphi)}\varphi = \varphi - 1 \leq \lambda\varphi \ , \tag{13}$$

with $\lambda = (K-1)/K$. Moreover, for all $\eta \in \mathbb{R}$ and $v \in \mathbb{R}^n$ such that $v_c = 0$, we have

$$Mv + \eta(\varphi - 1) = M_{(c,\varphi)}(v + \eta\varphi) \ . \tag{14}$$

Corollary 22. *Under the conditions of Lemma 21, the map $f(v) = Mv$ is such that $\mathcal{L}_\varphi(f)(w) = \varphi^{-1}M_{(c,\varphi)}(\varphi w) = M'w$, for some matrix M' with non negative entries and row sums less or equal to λ. Hence, $\mathcal{L}_\varphi(f)$ is order-preserving and contracting with contraction factor λ.*

Proof. Indeed, $\mathcal{L}_\varphi(f)(w) = \varphi^{-1}M_{(c,\varphi)}(\varphi w)$ follows from (14). Since by (13), $M_{(c,\varphi)}\varphi \leq \varphi$, we deduce that $M'\mathbb{1} \leq \lambda\mathbb{1}$ where $\mathbb{1}$ denotes the the vector with all coordinates equal to 1. □

Proof of Theorem 13. If f is a polyhedral map such that, all matrices $M \in \mathrm{imD}(f)$ satisfy the conditions of Lemma 21 (with the same fixed φ and c), then by Corollary 22, all matrices $M' \in \mathrm{imD}(\mathcal{L}_\varphi(f))$ satisfy the conclusions of Corollary 22. This implies by the continuity of f and $\mathcal{L}_\varphi(f)$, that $\mathcal{L}_\varphi(f)$ is order-preserving and contracting with contraction factor λ.

Let us show the above property for all maps $f^{(\sigma)}$. For this, consider the self-map $\bar{f}$ of $\mathbb{R}^n$ given by:

$$\bar{f}(v) := \max_{M\in\mathcal{M}} (M_{(c)} v) \ . \tag{15}$$

Note that it coincides with map of (12) on the set of vectors v such that $v_i = 0$, but we shall apply it to all vectors. Since all matrices involved in the previous formula have a unique final class and that this class contains c, we get that they have all a spectral radius strictly less than 1. By the same arguments as in the previous section, the map $\bar{f}$ has a spectral radius strictly less than one, so is contracting for the sup-norm after a scaling by some vector ψ (or equivalently is contracting the weighted sup-norm $\|v\|_\psi = \|v\psi^{-1}\|$). In particular the equation $\varphi = 1 + \bar{f}(\varphi)$ has a unique solution φ, and since the set of $\mathcal{M}$ is rectangular, this equation is the dynamic programming equation of an infinite horizon discounted 1-player game problem. The interpretation of φ as the value of this 1-player game problem gives that $\varphi_i = \mathcal{T}_{ic}$ for all $i \in [n]$. Since $\varphi = 1 + \bar{f}(\varphi) \geq 1 + M_{(c)}\varphi$ for all $M \in \mathcal{M}$, and a fortiori for all $M \in \mathrm{imD}(f^{(\sigma)})$ and $\sigma \in A_{\mathrm{M}}$, which implies that M satisfies the conditions of Lemma 21 with φ and c, we get by the above arguments that the maps $\mathcal{L}_\varphi(f^\sigma)$ are contracting in the sup-norm with contraction factor λ. □

Proof of Proposition 12. By definition of $\mathcal{L}_\varphi(f)$, we have $\eta + v = f(v)$ if and only if $w = \mathcal{L}_\varphi(f)(w)$ for $w = \eta + \varphi^{-1}v$. Moreover, the transformation of maps f, $\mathcal{L}_\varphi(f)(w) = \varphi^{-1}(\eta(\varphi - 1) + f(v))$, is preserving the order on the maps f. Hence if η^s, v^s and σ^s are respectively the sequences of eigenvalues, eigenvectors, and policies of Algorithm 3 for f and $f^{(\sigma)}$, then $w^s = \eta^s + \varphi^{-1}v^s$ and σ^s are respectively the sequence of fixed points and policies of Algorithm 1 for $\mathcal{L}_\varphi(f)$ and $\mathcal{L}_\varphi(f^{(\sigma)})$. □

References

[1] M. Akian, J. Cochet-Terrasson, S. Detournay, and S. Gaubert. Policy iteration algorithm for zero-sum multichain stochastic games with mean payoff and perfect information, 2012. `arXiv:1208.0446`.

[2] M. Akian, S. Gaubert, and R. Nussbaum. A Collatz-Wielandt characterization of the spectral radius of order-preserving homogeneous maps on cones. 2011. `arXiv:1112.5968`.

[3] D. Andersson. Extending Friedmann's lower bound to the Hoffman-Karp algorithm. *preprint, June*, 2009.

[4] D. P. Bertsekas. *Dynamic programming.* Prentice Hall Inc., Englewood Cliffs, NJ, 1987. Deterministic and stochastic models.

[5] Jean Cochet-Terrasson and Stéphane Gaubert. A policy iteration algorithm for zero-sum stochastic games with mean payoff. *C. R. Math. Acad. Sci. Paris*, 343(5):377–382, 2006.

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

[7] Eric V. Denardo. Contraction mappings in the theory underlying dynamic programming. *SIAM Review*, 9:165–177, 1967.

[8] John Fearnley. Exponential lower bounds for policy iteration. In *Automata, Languages and Programming*, pages 551–562, 2010.

[9] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes.* Springer-Verlag, New York, 1997.

[10] Oliver Friedmann. An exponential lower bound for the parity game strategy improvement algorithm as we know it. In *LICS*, pages 145–156. IEEE Computer Society, 2009.

[11] T.D. Hansen, P.B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. In *Innovations in Computer Science 2011*, pages 253–263. Tsinghua University Press, 2011.

[12] A. J. Hoffman and R. M. Karp. On nonterminating stochastic games. *Management Science. Journal of the Institute of Management Science. Application and Theory Series*, 12:359–370, 1966.

[13] Ronald A. Howard. *Dynamic programming and Markov processes.* The Technology Press of M.I.T., Cambridge, Mass., 1960.

[14] T. M. Liggett and S. A. Lippman. Stochastic games with perfect information and time average payoff. *SIAM Rev.*, 11:604–607, 1969.

[15] J. Mallet-Paret and Roger Nussbaum. Eigenvalues for a class of homogeneous cone maps arising from max-plus operators. *Discrete and Continuous Dynamical Systems*, 8(3):519–562, July 2002.

[16] R.D. Nussbaum. Convexity and log convexity for the spectral radius. *Linear Algebra Appl.*, 73:59–122, 1986.

[17] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1994.

[18] L. S. Shapley. Stochastic games. In *Stochastic games and applications (Stony Brook, NY, 1999)*, volume 570 of *NATO Sci. Ser. C Math. Phys. Sci.*, pages 1–7. Kluwer Acad. Publ., Dordrecht, 2003. Reprint of Proc. Nat. Acad. Sci. U.S.A. **3**9 (1953), 1095–1100 [0061807].

[19] Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Math. Oper. Res.*, 36(4):593–603, 2011.

Marianne Akian, INRIA Saclay–Île-de-France and CMAP, École Polytechnique. Address: CMAP, École Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France.

E-mail address: `Marianne.Akian@inria.fr`

Stéphane Gaubert, INRIA Saclay–Île-de-France and CMAP, École Polytechnique. Address: CMAP, École Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France.

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