Algorithms for uniform optimal strategies in two-player zero-sum stochastic games with perfect information

Konstantin Avrachenkov[a], Laura Cottatellucci[b], Lorenzo Maggi[b]

[a] *INRIA, BP95, 06902 Sophia Antipolis Cedex, France, e-mail: k.avrachenkov@sophia.inria.fr*
[b] *Eurecom, Mobile Communications, BP193, F-06560 Sophia Antipolis Cedex, France, e-mail: {laura.cottatellucci, lorenzo.maggi}@eurecom.fr*

Abstract

We deal with zero-sum two-player stochastic games with perfect information. We propose two algorithms to find the uniform optimal strategies and one method to compute the optimality range of discount factors. We prove the convergence in finite time for one algorithm. The uniform optimal strategies are also optimal for the long run average criterion and, in transient games, for the undiscounted criterion as well.

Keywords: Stochastic games, multi-agent Markov decision processes, perfect information, uniform optimal strategies, optimality range

1. Introduction

Stochastic games, also called multi-agent Markov Decision Processes, are multi-stage interactions among several participants in an environment whose conditions change stochastically, influenced by the decisions of the players. A detailed survey on this topic can be found in the book [1] by Filar and Vrieze. In this paper we deal with zero-sum stochastic games with two players and with perfect information. Under the perfect information assumption, the reward and the transition probabilities in each state are controlled at most by one player. Our results are grounded on the following references. Filar proved in [2] an ordered field property for the value of switching control stochastic games; the games with perfect information are a specific case of them. Raghavan and Syed provided in [3] a policy improvement algorithm to determine the optimal strategies for two-player zero-sum perfect information games under the discounted criterion, for a fixed discount factor. Inspired by the work of Jeroslow [4], Hordijk, Dekker, and Kallenberg proposed in [5] to find the optimal discount strategies for Markov Decision Processes (MDP's) for all discount factors close enough to 1 by utilizing the simplex method in the ordered field of rational functions with real coefficients. Filar, Altman, and Avrachenkov presented in [6] some algorithms for the computation of uniform optimal strategies in the context of perturbed MDP's; in [7], the same authors proposed an efficient asymptotic simplex method based on Laurent series expansion. Our contribution is organized as follows. We first introduce our stochastic game model in Section 2. In Section 3 we prove that, for all discounted factors close enough to 1, the discounted value belongs to the field of rational functions with real coefficients. Moreover, we summarize the main results of [5]. In Section 4 we present some useful results on uniform optimality in perfect information games. Then, we propose two algorithms which compute a pair of uniform discount optimal strategies $(\mathbf{f}^*, \mathbf{g}^*)$, which are optimal in the long run average criterion as well. The convergence in a finite time of the first algorithm, based on policy improvement, is proven in Section 5. A simple method to find the range of discount factors in which $(\mathbf{f}^*, \mathbf{g}^*)$ are discount optimal is shown in Section 6. We present our second algorithm, which is a best response algorithm, in Section 7. In Section 8 we show by simulation that the second algorithm has a lower complexity than the first one, in terms of number of pivot operations. In Section 9 we finally prove that, for transient stochastic games, $(\mathbf{f}^*, \mathbf{g}^*)$ are optimal under the undiscounted criterion as well.

Some notation remarks: the ordering relation between vectors of the same length $\mathbf{a} \geq (\leq)\mathbf{b}$ means that for every component $\mathbf{a}(i)$ and $\mathbf{b}(i)$, $\mathbf{a}(i) \geq (\leq)\mathbf{b}(i)$. The indicator function is referred to as $\mathbb{1}$. The symbol δ stands for Kronecker delta. The discount factor and the interest rate are barred, i.e. $(\overline{\beta}, \overline{\rho})$, if they represent a fixed real value; the symbols (β, ρ) represent the related real variables.

2. The model

In a two-player stochastic game Γ we have a set of states $S = \{s_1, s_2, \ldots, s_N\}$. For each state s, the set of actions available to Player i is called $A^{(i)}(s) = \{a_1^{(i)}(s), \ldots, a_{m_i(s)}^{(i)}(s)\}$, $i = 1, 2$. In zero-sum games, for each triple (s, a_1, a_2) with $a_1 \in A^{(1)}(s)$, $a_2 \in A^{(2)}(s)$ we assign an immediate reward $\mathbf{r}(s, a_1, a_2)$ to Player 1, $-\mathbf{r}(s, a_1, a_2)$ to Player 2 and a transition probability distribution $p(.|s, a_1, a_2)$ on S.

A stationary strategy $\mathbf{u} \in \mathbf{U}_S$ for Player i determines the probability $\mathbf{u}(a|s)$ that in state s Player i chooses the action $a \in A^{(i)}(s)$. We assume that both the number of states and the overall number of available actions are finite. Let $p(s'|s, \mathbf{f}, \mathbf{g})$ and $\mathbf{r}(s, \mathbf{f}, \mathbf{g})$ be the expectation with respect to the stationary strategies $(\mathbf{f}, \mathbf{g})$ of $p(s'|s)$ and of $\mathbf{r}(s)$, respectively.

Let $\overline{\beta} \in [0;1)$ be the discount factor and $\overline{\rho}$ be the interest rate such that $\overline{\beta}(1+\overline{\rho}) = 1$. Note that when $\overline{\beta} \uparrow 1$, then $\overline{\rho} \downarrow 0$. We define $\Phi_{\overline{\beta}}(\mathbf{f},\mathbf{g})$ as the N-by-1 vector whose i-th component $\Phi_{\overline{\beta}}(s_i,\mathbf{f},\mathbf{g})$ equals the expected $\overline{\beta}$-discounted reward when the initial state of the stochastic game is s_i:

$$\Phi_{\overline{\beta}}(\mathbf{f},\mathbf{g}) = \sum_{t=0}^{\infty} \overline{\beta}^t \mathbf{P}^t(\mathbf{f},\mathbf{g})\mathbf{r}(\mathbf{f},\mathbf{g}),$$

where $\mathbf{P}(\mathbf{f},\mathbf{g})$ and $\mathbf{r}(\mathbf{f},\mathbf{g})$ are the N-by-N transition probability matrix and the N-by-1 state-wise expected reward vector associated to the pair of strategies $(\mathbf{f},\mathbf{g})$, respectively.

Definition 1. *The $\overline{\beta}$-discounted value of the game Γ is such that*

$$\Phi_{\overline{\beta}}(\Gamma) = \sup_{\mathbf{f}} \inf_{\mathbf{g}} \Phi_{\overline{\beta}}(\mathbf{f},\mathbf{g}) = \inf_{\mathbf{g}} \sup_{\mathbf{f}} \Phi_{\overline{\beta}}(\mathbf{f},\mathbf{g}). \tag{1}$$

An optimal strategy $\mathbf{f}^*_{\overline{\beta}}$ ($\mathbf{g}^*_{\overline{\beta}}$) for Player 1 (2) assures to him a reward which is at least (at most) $\Phi_{\overline{\beta}}(\Gamma)$.

Let $\Phi(\mathbf{f},\mathbf{g})$ be the long run average reward of the game Γ associated to the pair of strategies $(\mathbf{f},\mathbf{g})$:

$$\Phi(\mathbf{f},\mathbf{g}) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbf{P}^t(\mathbf{f},\mathbf{g})\mathbf{r}(\mathbf{f},\mathbf{g})$$

and let $\Phi(\Gamma)$ be the value vector for the long run average criterion of the game Γ, defined in an analogous way to expression (1). The existence of optimal strategies in discounted stochastic games is guaranteed by the following Theorem.

Theorem 2.1 ([1]). *Under the hypothesis of discounted reward criterion, stochastic games possess a value, the optimal strategies $(\mathbf{f}^*_{\overline{\beta}},\mathbf{g}^*_{\overline{\beta}})$ exist among stationary strategies and, moreover, $\Phi_{\overline{\beta}}(\Gamma) = \Phi_{\overline{\beta}}(\mathbf{f}^*_{\overline{\beta}},\mathbf{g}^*_{\overline{\beta}})$.*

Definition 2. *A stationary strategy $\mathbf{h}$ is said to be uniform discount optimal (or equivalently uniform optimal) for Player $i=1,2$ if $\mathbf{h}$ is optimal for Player i for every $\overline{\beta}$ close enough to 1 (or, equivalently, for all $\overline{\rho}$ close enough to 0).*

In the present paper we deal with stochastic games with perfect information.

Definition 3. *Under the hypothesis of perfect information, in each state at most one player has more than one action available.*

Let $S_1 = \{s_1,\dots,s_{t_1}\}$ be the set of states controlled by Player 1 and $S_2 = \{s_{t_1+1},\dots, s_{t_1+t_2}\}$ be the set controlled by Player 2, with $t_1+t_2 \le N$.

3. The ordered field of rational functions with real coefficients

Let $P(\mathbb{R})$ be the ring of the polynomials with real coefficients.

Definition 4. *The dominating coefficient of a polynomial $p(x) = a_0 + a_1x + \cdots + a_nx^n$ is the coefficient a_k, where $k = \min\{i : a_i \neq 0\}$ and we denote it with $\mathscr{D}(f)$.*

Let $F(\mathbb{R})$ be the non-Archimedean ordered field of fractions of polynomials with coefficients in $\mathbb{R}$:

$$f(x) = \frac{c_0 + c_1x + \cdots + c_nx^n}{d_0 + d_1x + \cdots + d_mx^m} \qquad f \in F(\mathbb{R}),$$

where the operations of sum and product are defined in the usual way (see [5]). Two rational functions h/g, p/q are identical (and we say $h/g =_l p/q$) if and only if $h(x)q(x) = p(x)g(x)$, $\forall x \in \mathbb{R}$.

Lemma 3.1 ([5]). *A complete ordering in $F(\mathbb{R})$ is obtained by the rule: $p/q >_l 0$ if and only if $\mathscr{D}(p)\mathscr{D}(q) > 0$, where $p,q \in P(\mathbb{R})$.*

In the same way, we also define the operations of maximum ($\max_l$) and minimum ($\min_l$) in $F(\mathbb{R})$.

Lemma 3.2 ([5]). *The rational function p/q is positive $(p/q >_l 0)$ if and only if there exists $x_0 > 0$ such that $p(x)/q(x) > 0$ for every $x \in (0;x_0]$.*

3.1. Computation of Blackwell optimum policy in MDP's

Let us consider a Markov Decision Process (MDP), which can be seen as a two-player stochastic game in which one of the two players fixes its own strategy. Let $A(s)$ be the finite action space available in state $s \in S$. Let $m(s) = |A(s)|$.

Definition 5. *The strategy $\mathbf{f}^*$ is Blackwell optimal if and only if there exists $\overline{\rho}^* > 0$ such that $\mathbf{f}^*$ is optimal in the $(\overline{\rho}+1)^{-1}$-discounted MDP for all the interest rates $\overline{\rho} \in (0;\overline{\rho}^*]$.*

In [5] the authors provide an algorithm to compute the Blackwell optimal policy in MDP's. It consists in solving the following parametric linear programming model:

$$\begin{cases} \max_{\mathbf{x}}{}_l \sum_{s=1}^{N} \sum_{a=1}^{m(s)} x_{s,a}(\rho)\mathbf{r}(s,a) \\ \sum_{s=1}^{N} \sum_{a=1}^{m(s)} [(1+\rho)\delta_{s,s'} - p(s'|s,a)]\, x_{s,a}(\rho) =_l 1,\ s' \in S \\ x_{s,a}(\rho) \ge_l 0, \quad s \in S,\ a \in A(s) \end{cases} \tag{2}$$

in the ordered field of rational functions with real coefficients $F(\mathbb{R})$. The Blackwell optimal strategy is computed as $\mathbf{v}^*(a|s) := \mathbb{1}\left(x^*_{s,a}(\rho) >_l 0\right)$ for all $s \in S,\ a \in A(s)$, where $\{x^*_{s,a}(\rho),\ \forall s,a\}$ is the solution of (2).

3.2. Application to stochastic games

In this section we will introduce the ordered field $F(\mathbb{R})$ in stochastic games, not necessarily with perfect information.

Theorem 3.3. *Let $\mathbf{f},\mathbf{g}$ be two stationary strategies for Players 1 and 2, respectively. Let $\Phi_\rho(\mathbf{f},\mathbf{g}) : \mathbb{R} \to \mathbb{R}^N$ be the discounted reward associated to $(\mathbf{f},\mathbf{g})$, expressed as a function of ρ. Then, $\Phi_\rho(\mathbf{f},\mathbf{g}) \in F(\mathbb{R})$.*

Proof. For any pair of stationary strategies $(\mathbf{f},\mathbf{g})$, we can write $\forall s \in [1;N]$:

$$\sum_{s'=1}^{N} [(1+\rho)\delta_{s,s'} - p(s'|s,\mathbf{f},\mathbf{g})]\Phi_\rho(s',\mathbf{f},\mathbf{g}) = (1+\rho)\mathbf{r}(s,\mathbf{f},\mathbf{g}),$$

where ρ is a variable. By solving the above system of equations in the unknown Φ_ρ by Cramer rule, it is evident that $\Phi_\rho(\mathbf{f},\mathbf{g}) \in F(\mathbb{R})$. □

From Theorems 2.1 and 3.3 we obtain the following Lemma, which ensures that, for discounted factors close enough to 1, the discounted value exists and belongs to the field of rational functions with real coefficients.

Lemma 3.4. *Let Γ be a zero-sum stochastic game possessing uniform discount optimal strategies $\mathbf{f}^*$ and $\mathbf{g}^*$ for Players 1 and 2, respectively. Then, there exists $\Phi^*_\rho(\Gamma) \in F(\mathbb{R})$ such that:*

$$\Phi_\rho(\mathbf{f},\mathbf{g}^*) \leq_l \Phi_\rho(\mathbf{f}^*,\mathbf{g}^*) =_l \Phi^*_\rho(\Gamma) \leq_l \Phi_\rho(\mathbf{f}^*,\mathbf{g}), \quad \forall \mathbf{f},\mathbf{g}. \quad (3)$$

Proof. By hypothesis, there exists $\overline{\rho}^* > 0$ such that $(\mathbf{f}^*,\mathbf{g}^*)$ are discounted optimal for all the interest rates $\overline{\rho} \in (0;\overline{\rho}^*]$. For Theorem 3.3, $\Phi_\rho(\mathbf{f}^*,\mathbf{g}^*) \in F(\mathbb{R})$ and, from Theorem 2.1, the uniform optimum value $\Phi_{\overline{\rho}}(\Gamma) = \Phi_{\overline{\rho}}(\mathbf{f}^*,\mathbf{g}^*)$ $\forall \overline{\rho} \in (0;\overline{\rho}^*]$. Hence, the saddle point relation in (3) holds. □

Definition 6. *$\Phi^*_\rho(\Gamma)$, defined as in (3), is the uniform discount value of the stochastic game Γ.*

4. Uniform optimality in perfect information games

In a perfect information game, a pure stationary strategy for Player i is a probability distribution $\mathbf{f}(.|s)$ on the action space $A_i(s)$ such that there exists $a \in A_i(s)$ for which $\mathbf{f}(a|s) = 1$, for every $s \in S$.

Theorem 4.1 ([1]). *For a stochastic game with perfect information, both players possess uniform discount optimal pure stationary strategies, which are optimal for the average criterion as well.*

Definition 7. *We call two pure stationary strategies adjacent if and only if they differ only in one state.*

The following property holds, whose proof is analogous to the one in the field of real numbers in [3].

Lemma 4.2. *Let $\mathbf{g}$ be a strategy for Player 2 and $\mathbf{f},\mathbf{f}_1$ be two adjacent strategies for Player 1. Then, either $\Phi_\rho(\mathbf{f}_1,\mathbf{g}) \geq_l \Phi_\rho(\mathbf{f},\mathbf{g})$ or $\Phi_\rho(\mathbf{f}_1,\mathbf{g}) \leq_l \Phi_\rho(\mathbf{f},\mathbf{g})$, which means that the two vectors are partially ordered.*

The Lemma 4.2 allows us to give the following definition.

Definition 8. *Let $(\mathbf{f},\mathbf{g})$ be a pair of pure stationary strategies for Player 1 and 2, respectively. We call $\mathbf{f}_1$ $(\mathbf{g}_1)$ a uniform adjacent improvement for Player 1 (2) in state s_t if and only if $\mathbf{f}_1$ $(\mathbf{g}_1)$ is a pure stationary strategy which differs from $\mathbf{f}$ $(\mathbf{g})$ only in state s_t and $\Phi_\rho(\mathbf{f}_1,\mathbf{g}) \geq_l \Phi_\rho(\mathbf{f},\mathbf{g})$ $(\Phi_\rho(\mathbf{f},\mathbf{g}_1) \leq_l \Phi_\rho(\mathbf{f},\mathbf{g}))$, where the strict inequality holds in at least one component.*

As in the case in which the discount interest rate is fixed, we achieve the following result. Its proof directly stems from the Bellman optimality equations in the ordered field $F(\mathbb{R})$.

Lemma 4.3. *Let Γ be a stochastic game with perfect information. A pair of pure stationary strategies $(\mathbf{f}^*,\mathbf{g}^*)$ is uniform discount optimal if and only if no uniform adjacent improvement is possible for both players.*

In perfect information games, the following result holds.

Lemma 4.4 ([3]). *In a zero-sum, perfect information, two-player discounted stochastic game Γ with interest rate $\overline{\rho} > 0$, a pair of pure stationary strategies $(\mathbf{f}^*,\mathbf{g}^*)$ is optimal if and only if $\Phi_{\overline{\rho}}(\mathbf{f}^*,\mathbf{g}^*) = \Phi_{\overline{\rho}}(\Gamma)$, the value of the discounted stochastic game Γ.*

From the above result we can easily derive the analogous property in the ordered field $F(\mathbb{R})$.

Lemma 4.5. *In a zero-sum, two-player stochastic game Γ with perfect information, a pair of pure stationary strategies $(\mathbf{f}^*,\mathbf{g}^*)$ are uniform discount optimal if and only if $\Phi_\rho(\mathbf{f}^*,\mathbf{g}^*) =_l \Phi^*_\rho(\Gamma) \in F(\mathbf{R})$, where $\Phi^*_\rho(\Gamma)$ is the uniform discount value of Γ.*

Proof. The *only if* statement coincides with the assertion of Theorem 2.1. Conversely, if a pair of strategies $(\mathbf{f}^*,\mathbf{g}^*)$ has the property $\Phi_\rho(\mathbf{f}^*,\mathbf{g}^*) =_l \Phi^*_\rho(\Gamma)$, then there exists $\rho^* > 0$ such that $\forall \overline{\rho} \in (0;\rho^*]$, $\Phi_{\overline{\rho}}(\mathbf{f}^*,\mathbf{g}^*)$ coincides with the value of the game Γ, $\forall \overline{\rho} \in (0;\rho^*]$. Then, thanks to Lemma 4.4, we can say that the strategies $\mathbf{f}^*,\mathbf{g}^*$ are optimal in the discounted game Γ for any $\overline{\rho} \in (0;\rho^*]$, which means that they are uniform optimal. □

Remark 1. *Generally, the discounted value of a stochastic game for all the interest rates close enough to 0 belongs to the field of real Puiseux series (see [1]).*

Let s_t be a state controlled by Player $i = 1,2$ and $X \subset A_i(s_t)$. Let us call Γ^t_X the stochastic game which is equivalent to Γ except in state s_t, where Player i has only the actions X available. We present the following Lemma, whose proof is analogous to the one in the real field (see [3]).

Lemma 4.6. *Let $i = 1,2$ and $s_t \in S_i$, $X \subset A_i(s_t)$, $Y \subset A_i(s_t)$, $X \cap Y = \emptyset$. Then $\Phi^*_\rho(\Gamma^t_{X\cup Y}) \in F(\mathbb{R})$, the uniform value of the game $\Gamma^t_{X\cup Y}$, equals*

$$\Phi^*_\rho(\Gamma^t_{X\cup Y}) =_l \max{}_l \{\Phi^*_\rho(\Gamma^t_X), \Phi^*_\rho(\Gamma^t_Y)\} \qquad \text{if } i = 1,$$
$$\Phi^*_\rho(\Gamma^t_{X\cup Y}) =_l \min{}_l \{\Phi^*_\rho(\Gamma^t_X), \Phi^*_\rho(\Gamma^t_Y)\} \qquad \text{if } i = 2.$$

5. Policy improvement algorithm

In this section we find a policy improvement algorithm which allows to find the uniform discount optimal strategies for both players in a stochastic game with perfect information. Such strategies coincide with the optimal strategies for the long run average criterion, for Theorem 4.1. Following the lines of Raghavan and Syed's algorithm [3] for a fixed discount factor, we propose an algorithm suitable for the ordered field $F(\mathbb{R})$.

Let Γ be a zero-sum two-player stochastic game with perfect information. Let $\Gamma_i(\mathbf{f})$ be the MDP faced by Player i when the opponent fixes its own strategy $\mathbf{g}$.

Algorithm 5.1.

Step 1 *Select randomly a stationary deterministic pure strategy* $\mathbf{g}$ *for Player 2.*

Step 2 *Find the Blackwell optimal strategy for Player 1 in the MDP* $\Gamma_1(\mathbf{g})$ *by solving within the field* $F(\mathbb{R})$ *the following linear programming model:*

$$\begin{cases} \max\limits_{\mathbf{x}}{}_l \sum_{s=1}^N \sum_{a=1}^{m_1(s)} x_{s,a}(\rho)\mathbf{r}(s,a,\mathbf{g}) \\ \sum_{s=1}^N \sum_{a=1}^{m_1(s)} [(1+\rho)\delta_{s,s'} - p(s'|s,a,\mathbf{g})]\, x_{s,a}(\rho) =_l 1,\ s' \in S \\ x_{s,a}(\rho) \geq_l 0, \quad s \in S,\ a \in A_1(s) \end{cases} \tag{4}$$

and compute the pure strategy $\mathbf{f}$ *as*

$$\mathbf{f}(a|s) := \mathbb{1}\left(x^*_{s,a}(\rho) >_l 0\right) \qquad \forall s \in S,\ a \in A_1(s), \tag{5}$$

where $\{x^*_{s,a}(\rho),\ \forall s,a\}$ *is the solution of (4).*

Step 3 *Find the minimum k such that in* $s_{t_1+k} \in \{s_{t_1+1}, \ldots, s_{t_1+t_2}\}$ *there exists an adjacent improvement* $\mathbf{g}'$ *for Player 2, with the help of the simplex tableau associated to the following linear programming model:*

$$\begin{cases} \min\limits_{\mathbf{x}}{}_l \sum_{s=1}^N \sum_{a=1}^{m_2(s)} x_{s,a}(\rho)\mathbf{r}(s,\mathbf{f},a) \\ \sum_{s=1}^N \sum_{a=1}^{m_2(s)} [(1+\rho)\delta_{s,s'} - p(s'|s,\mathbf{f},a)]\, x_{s,a}(\rho) =_l 1,\ s' \in S \\ x_{s,a}(\rho) \geq_l 0, \quad s \in S,\ a \in A_2(s) \end{cases} \tag{6}$$

where the entering variables are $\{x_{s,a} :\ \mathbf{g}(a|s) = 1,\ \forall s\}$. *If no such improvement for Player 2 is possible, then go to step 4, otherwise set* $\mathbf{g} := \mathbf{g}'$ *and go to step 2.*

Step 4 *Set* $(\mathbf{f}^*, \mathbf{g}^*) := (\mathbf{f}, \mathbf{g})$ *and stop. The strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are uniform discount and long run average optimal in the stochastic game* Γ *for Player 1 and Player 2, respectively.*

Note that all the algebraic operations and the order signs are to be intended in the field $F(\mathbb{R})$.

Remark 2. *Unlike the solution in [3], Algorithm 5.1 does not require the strategy search for Player 1 to be lexicographic. In fact, Player 1 faces in step 2 a classic Blackwell optimization.*

Remark 3. *Clearly, the roles of Player 1 and 2 can be swapped in Algorithm 5.1. For simplicity, throughout the paper the Player 1 will be assigned to step 2.*

Remark 4. *In step 3 of Algorithm 5.1, once the state* s_{t_1+k} *is found, the adjacent improvement involves the pivoting of any of the non basic variable* $x_{s_{t_1+k},a}$ *to which corresponds a non positive (in the field* $F(\mathbb{R})$*) reduced cost.*

Let us prove the convergence in finite time of Algorithm 5.1.

Theorem 5.2. *Algorithm 5.1 stops in a finite time and the pair of strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are both uniform discount and long run average optimal in the stochastic game* Γ.

Proof. The proof follows the lines of the analogous one in the real field (see [3]). It proceeds by induction on the overall number of actions and it exploits Lemmas 4.3 and 4.6. The main difference from [3], that does not affect the correctness of the proof, is that Player 1 is not constrained in optimizing lexicographically the MDP $\Gamma_1(\mathbf{g})$. For Theorem 4.1, $(\mathbf{f}^*, \mathbf{g}^*)$ are long run average optimal as well. □

5.1. Round-off errors sensitivity

The first non-zero coefficients of the polynomials (numerator and denominator) of the tableaux obtained throughout the algorithm unfolding determine the positiveness of the elements of the tableaux themselves. Hence, Algorithm 5.1 is highly sensitive to the round-off errors that affect the null coefficients.

If the rewards and the transition probabilities for each pair of strategies are rational, then it is possible to work in the exact arithmetic and such unconveniences are completely avoided. In fact, if all the input data are rational, they will stay rational after the algorithm execution.

6. Computation of the optimality range factor

Let us report the analogous result to Lemma 4.3 when the discount factor is fixed.

Lemma 6.1 **([3]).** *Let* Γ *be a stochastic game with perfect information. Let* $\overline{\beta} \in [0;1)$. *The pure stationary strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are* $\overline{\beta}$*-discount optimal if and only if no uniform adjacent improvements are possible for both players in the* $\overline{\beta}$*-discounted stochastic game* Γ.

Let us define with $\zeta(f_\rho)$, where $f_\rho \in F(\mathbb{R})$, the set of positive roots of f_ρ such that $d f_\rho / d\rho|_{\rho=u} < 0,\ \forall u \in \zeta(f_\rho)$. The following Lemma suggests how to compute the optimality range of discount factors.

Lemma 6.2. *Let C be the set of the reduced costs associated to the two optimal tableaux obtained at the step 2 and 3 of the last iteration of Algorithm 5.1. Let* $\overline{\rho}^* = \min_c \zeta(c)$, *where* $c \in C$. *If* $\overline{\rho}^*$ *does not exist, then the uniform optimal strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are optimal for all* $\overline{\beta} \in [0;1)$. *Otherwise,* $\overline{\beta}^* = (1+\overline{\rho}^*)^{-1}$ *is the smallest value such that* $(\mathbf{f}^*, \mathbf{g}^*)$ *are* $\overline{\beta}$*-discount optimal in the game* Γ, *for all* $\overline{\beta} \in [\overline{\beta}^*; 1)$.

Proof. If $\overline{\rho}^*$ does not exist, then the reduced costs are nonnegative for any $\overline{\rho} > 0$. Hence, $(\mathbf{f}^*, \mathbf{g}^*)$ are optimal $\forall \overline{\beta} \in [0;1)$. Otherwise, $\forall \overline{\rho} \in (0; \overline{\rho}^*]$, the reduced costs are positive, hence no adjacent improvements are possible for both players. So, for Lemma 6.1 they are discounted optimal. If $\overline{\rho} > \overline{\rho}^*$ and $\overline{\rho}^* \in \mathbb{R}$, then at least one reduced cost is negative, hence at least an adjacent improvement is possible and $(\mathbf{f}^*, \mathbf{g}^*)$ are not $\overline{\beta}$-discount optimal, where $\overline{\beta} = (1+\overline{\rho})^{-1}$. □

7. Best response algorithm

Let Γ be a zero-sum two-player stochastic game with perfect information. Consider the following best-response algorithm.

Algorithm 7.1.

Step 1 *Select a stationary pure strategy* $\mathbf{g}_0$ *for Player 2. Set* $k := 0$.

Step 2 *Find the Blackwell optimal strategy* $\mathbf{f}_k$ *for Player 1 in the MDP* $\Gamma_1(\mathbf{g}_k)$.

Step 3 *If* $\mathbf{g}_k$ *is Blackwell optimal in* $\Gamma_2(\mathbf{f}_k)$*, then set* $(\mathbf{f}^*, \mathbf{g}^*) := (\mathbf{f}_k, \mathbf{g}_k)$ *and stop. Otherwise, find the Blackwell optimal strategy* $\mathbf{g}_{k+1}$ *for Player 2 in the MDP* $\Gamma_2(\mathbf{f}_k)$*, set* $k := k+1$ *and go to step 2.*

Obviously, if Algorithm 7.1 stops, $(\mathbf{f}^*, \mathbf{g}^*)$ is a pair of uniform discount and long run average optimal strategies, since they are both Blackwell optimal in the respective MDP's, $\Gamma_1(\mathbf{g}^*)$ and $\Gamma_2(\mathbf{f}^*)$.
The proof that Algorithm 7.1 never cycles is still an open problem. We found that $\Phi_\rho(\mathbf{f}_{k+1}, \mathbf{g}_{k+1}) \leq_l \Phi_\rho(\mathbf{f}_k, \mathbf{g}_k)$, is not true in general. However, if the conjecture in [3] were valid, then we could conclude that Algorithm 7.1 terminates in a finite time.

8. Complexity: simulation results

In Algorithm 5.1, Player 1 faces at each iteration an MDP optimization problem in the field of rational functions with real coefficients, which is solvable in polynomial time. Player 2, instead, is involved in a lexicographic search throughout the algorithm unfolding, whose complexity is at worst exponential in the number of states N. Player 2 lexicographically expands its search of its optimal strategy, and at the k-th iteration the two players find the solution of a subgame $\Gamma^{(k)}$ which monotonically tends to the entire stochastic game Γ.
The efficiency of Algorithm 5.1 is mostly due to the fact that most of the actions totally dominate other actions. In other words, it often occurs that an optimal action found in the subgame $\Gamma^{(k)}$, is optimum also in Γ, and consequently remains the same during all the remaining iterations. This exponentially reduces the policy space in which the algorithm needs to search.

Remark 5. *Since in Algorithm 5.1 players' roles are interchangeble and since most of the actions dominate totally other actions, we suggest to assign the step 2 of the algorithm to the player whose total number of available actions is bigger.*

Differently from [3], the search for Player 1 does not need to be lexicographic, and Player 1 is left totally free to optimize the MDP that it faces at each iteration of the algorithm in the most efficient way.

Let us compare in terms of number of pivotings the following three algorithms:

$\mathbf{M_1}$: Algorithm 5.1, in which in step 2 Player 1 pivots with respect to the variable with the minimum reduced cost until it finds its own Blackwell optimal strategy.

$\mathbf{M_2}$: Algorithm 5.1, in which in step 2 Player 1 pursues a lexicographic search, pivoting iteratively with respect to the *first* non-basic variable with a negative (in the field $F(\mathbb{R})$) reduced cost. This method is analogous to the one shown in [3], but in the field $F(\mathbb{R})$.

$\mathbf{M_3}$: Algorithm 7.1.

The results are shown in Tables 1 and 2. The simulations were carried out on 10^4 randomly generated stochastic games with 4 states, 2 for Player 1 and 2 for Player 2. In each state 5 actions are available for the controlling player.

	n. pivotings
M_1	40.59
M_2	41.87
M_3	24.93

Table 1: Average number of pivotings for the 3 methods.

> (%)	M_1	M_2	M_3
M_1	-	52.85	18.57
M_2	42.18	-	15.26
M_3	80.05	82.75	-

Table 2: $M_i > M_j$ when, fixing the game, the number of pivotings in M_i is strictly smaller than the number of pivoting in M_j.

It is evident that Algorithm M_3 is much faster than the other two. In our numerical experiment, M_3 never cycled. The difference between M_1 and M_2 is due to the more efficient simplex method used by Player 1 in M_1.

9. Transient games

Let $p_t(s'|s)$ be the probability that the process is in state s' at time t given that s is the initial state. Let us give the definition of transient games.

Definition 9. *A stochastic game is transient if and only if* $\sum_{t=0}^{\infty} \sum_{s' \in S} p_t(s'|s, \mathbf{f}, \mathbf{g})$ *is finite for all* $s \in S$ *and for any pair of stationary strategies* $(\mathbf{f}, \mathbf{g})$.

Here we present the result of this section.

Theorem 9.1. *The uniform optimal strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *for a transient stochastic game with perfect information are optimal in the undiscounted criterion, i.e.* $\overline{\beta} = 1$*, as well.*

Proof. The uniform optimal strategies are still optimal when $\overline{\rho} \downarrow 0$, since the reduced costs of the tableaux built at the end of Algorithm 5.1 are non negative when $\overline{\rho} \downarrow 0$. We know from [1] that, for transient stochastic games, the reward associated to each pair of stationary strategies $(\mathbf{f}, \mathbf{g})$ is finite. By invoking Abel's Theorem on power series [8], we claim that the reward associated to any stationary $(\mathbf{f}, \mathbf{g})$ tends to the undiscounted reward when $\overline{\rho} \downarrow 0$. Hence, the saddle-point relation (3) is still valid when $\overline{\rho} = 0$ and $(\mathbf{f}^*, \mathbf{g}^*)$ are optimal in the undiscounted criterion as well. □

Acknowledgements

This reasearch was supported by "Agence Nationale de la Recherche", with reference ANR-09-VERS-001, and by the European research project SAPHYRE, which is partly funded by the European Union under its FP7 ICT Objective 1.1 - The Network of the Future. We would like to thank the anonymous reviewer for very helpful remarks.

References

[1] J. Filar, K. Vrieze, Competitive Markov Decision Processes, Springer, 1996.

[2] J.A. Filar, Ordered Field Property for Stochastic Games When the Player Who Controls Transitions Changes from State to State, Journal of Optimization Theory and Applications. 34, No.4 (1981) 503-513.

[3] T.E.S. Raghavan, Z. Syed, A policy-improvement type algorithm for solving zero-sum two-person stochastic games of perfect information, Mathematical Programming. 95, No.3 (2003) 513-532.

[4] R.G. Jeroslow, Asymptotic Linear Programming, Operations Research. 21 (1973) 1128-1141.

[5] A. Hordijk, R. Dekker, L.C.M. Kallenberg, Sensitivity Analysis in Discounted Markov Decision Processes, OR Spektrum. 7, No.3 (1985) 143-151.

[6] E. Altman, K. Avrachenkov, J.A. Filar, Asymptotic linear programming and policy improvement for singularly perturbed Markov decision processes, ZOR: Mathematical Methods of Operations Research. 49, No.1 (1999) 97-109.

[7] J.A. Filar, E. Altman and K. Avrachenkov, An asymptotic simplex method for singularly perturbed linear programs, Operations Research Letters. 30, No.5 (2002) 295-307.

[8] K. Knopp, Theory and Application of Infinite Series, Dover, 1990.