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

Konstantin Avrachenkov, Laura Cottatellucci, Lorenzo Maggi

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

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

Konstantin Avrachenkov — Laura Cottatellucci — Lorenzo Maggi

N° 7355

July 2010

Thème COM

ISSN 0249-6399 ISRN INRIA/RR--7355--FR+ENG

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

Konstantin Avrachenkov* , Laura Cottatellucci† , Lorenzo Maggi‡

Thème COM — Systèmes communicants
Projet Maestro

Rapport de recherche n° 7355 — July 2010 — 22 pages

Abstract: In stochastic games with perfect information, in each state at most one player has more than one action available. We propose two algorithms which find the uniform optimal strategies for zero-sum two-player stochastic games with perfect information. Such strategies are optimal for the long term average criterion as well. We prove the convergence for one algorithm, which presents a higher complexity than the other one, for which we provide numerical analysis.

Key-words: Stochastic games, Perfect information, Uniform optimal strategies, Computation

* INRIA Sophia Antipolis-Méditerranée, France, k.avrachenkov@sophia.inria.fr
† Eurecom, Mobile Communnications, France, laura.cottatellucci@eurecom.fr
‡ Eurecom, Mobile Communnications, France, lorenzo.maggi@eurecom.fr

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 — Télécopie : +33 4 92 38 77 65

Des algorithmes pour trouver les stratégies uniformément optimales dans les jeux stochastiques à somme nulle avec deux joueurs et avec information parfaite

Résumé : Dans les jeux stochastiques à information parfaite, dans chaque etat, au plus, un joueur a plus d'une action disponibles. Nous proposons deux algorithmes qui trouvent les stratégies uniformément optimales pour les jeux stochastiques à somme nulle avec deux joueurs et information parfaite. Ces stratégies sont aussi optimales pour le critère de la moyenne à long terme. Nous prouvons la convergence pour un algorithme, qui a une plus grande complexité que l'autre, pour lequel nous offrons une analyse numérique.

Mots-clés : Jeux stochastiques, Information parfaite, Stratégies uniformément optimales, Calcul

1 Introduction

Stochastic games are multi-stage interactions among several participants in an environment whose conditions change stochastically, influenced by the decisions of the players. Such games were introduced by Shapley (1953), who proved the existence of the discounted value and of the stationary discounted optimal strategies in two-player zero-sum games with finite state and action spaces. The problem of long term average reward games was addressed first by Gillette (1957). Bewley and Kohlberg (1976) proved that the field of real Puiseux series is an appropriate class to study the asymptotic behavior of discounted stochastic game when the discount factor tends to one. Mertens and Neyman (1981) showed the existence of the long term average value of stochastic games. Then, Parthasarathy and Raghavan (1981) first introduced the notion of order field property. This property implies that the solution of a game lies in the same ordered field of the game data. Solan and Vieille (2009) presented an algorithm to find the ε-optimal uniform discounted strategies in two-player zero-sum stochastic games, where $\varepsilon > 0$.

Perfect information games were addressed by several researchers (e.g. see Thuijsman and Raghavan, 1997, Altman and Feinberg, 2000), since they are the most elementary form of stochastic games: the reward and the transition probabilities in each state are controlled at most by one player. Recently, Raghavan and Syed (2002) provided an algorithm which finds the optimal strategies for two-player zero-sum perfect information games under the discounted criterion for a fixed discount factor.

Markov Decision Processes (MDPs) can be seen as stochastic games in which only one player can possess more than one action in each state. It is well known (see e.g. Filar and Vrieze, 1996) that the optimal strategy in an MDP can be computed with the help of a linear programming formulation. Hordijk, Dekker and Kallenberg (1985) proposed to find the Blackwell optimal strategies (uniform optimal discount strategies) for MDPs by using the simplex method in the ordered field of rational functions with real coefficients. Altman, Avrachenkov and Filar (1999) analysed singularly perturbed MDP using the simplex method in the ordered field of rational functions. More generally, Eaves and Rothblum (1994) studied how to solve a vast class of linear problems, including linear programming, in any ordered field.

In this paper we propose two algorithms to determine the uniform optimal discount strategies in two-player zero-sum games with perfect information. Such strategies are optimal in the long run average criterion as well. The proposed approaches generalize the works by Hordijk, Dekker, Kallenberg (1985) and Raghavan, Syed (2003) to the game model in the field $F(\mathbb{R})$ of the non-archimedean ordered field of rational functions with coefficients in $\mathbb{R}$.

Let Γ be a two-player zero-sum stochastic game with perfect information and $\Gamma_i(\mathbf{h}), i = 1,2$ be the MDP that player i faces when the other player fixes his own strategy $\mathbf{h}$. Our first algorithm can be summed up in the following 3 steps:

1. Choose a stationary pure strategy $\mathbf{g}$ for player 2.
2. Find the uniform optimal strategy $\mathbf{f}$ for player 1 in the MDP $\Gamma_1(\mathbf{g})$.

3. Find the *first* state controlled by player 2 in which a change of strategy $\mathbf{g}'$ is a benefit for player 2 for all the discount factors close enough to 1. If it does not exists, then $(\mathbf{f},\mathbf{g})$ are uniform optimal, otherwise set $\mathbf{g}:=\mathbf{g}'$ and go to step 2.

It is evident that player 1 is left totally free to optimize the MDP that he faces at each iteration of the algorithm in the most efficient way.
Our second algorithm is a best response approach, in which the two players alternatively find their own uniform optimal strategies:

1. Choose a stationary pure strategy $\mathbf{g}$ for player 2.
2. Find the uniform optimal strategy $\mathbf{f}$ for player 1 in the MDP $\Gamma_1(\mathbf{g})$.
3. If $\mathbf{g}$ is uniform optimal for player 2 in the MDP $\Gamma_2(\mathbf{f})$, then $(\mathbf{f},\mathbf{g})$ are uniform optimal. Otherwise, find the uniform optimal strategy $\mathbf{g}'$ in $\Gamma_2(\mathbf{f})$, set $\mathbf{g}:=\mathbf{g}'$ and go to step 2.

The convergence in a finite time of the first algorithm is proven, while for the second we provide numerical analysis. We also show that the second algorithm has a lower complexity.

This paper is organized as follows. In section 2 we introduce formally the properties of stochastic games, section 3 is dedicated to the description of the field of rational functions with real coefficients, while in section 4 we recall the linear programming procedures in the field $F(\mathbb{R})$ in order to find a Blackwell optimal policy for MDPs. We present some new useful results on perfect information games in section 5 and section 6 is dedicated to the description and to the validation of our first algorithm. In section 7 we provide a numerical example. In section 8 we introduce an algorithm whose convergence is only conjectured; we report some considerations and numerical results about the complexity of our algorithms in section 8.1.

Some notation remarks: the ordering relation between vectors of the same length $\mathbf{a} \geq (\leq) \mathbf{b}$ means that for every component i, $\mathbf{a}(i) \geq (\leq) \mathbf{b}(i)$. The discount factor and the interest rate are barred ($\overline{\beta},\overline{\rho}$) if they are a fixed value; the symbols β,ρ represent the related variables.

2 The model

In a two-player stochastic game Γ we have a set of states $S = \{s_1, s_2, \ldots, s_N\}$, and for each state s the set of actions available to the i-th player is called $A^{(i)}(s) = \{a_1^{(i)}(s), \ldots, a_{m_i(s)}^{(i)}\}$, $i = 1,2$. Each triple (s,a_1,a_2) with $a_1 \in A^{(1)}$, $a_2 \in A^{(2)}$ is assigned an immediate reward $r(s,a_1,a_2)$ for player 1, $-r(s,a_1,a_2)$ for 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 the i-th player determines the probability $u(a|s)$ that in state s player i chooses the actions $a \in [a_1^{(i)}, \ldots, a_{m_i(s)}^{(i)}]$.
We assume that both the number of states and the overall number of available actions are finite.

It is evident that a couple of strategies $\mathbf{f} \in \mathbf{F}_S,\ \mathbf{g} \in \mathbf{G}_S$ for player 1 and 2, respectively, sets up a Markov chain in which the transition probability equals

$$p(s'|s,\mathbf{f},\mathbf{g}) = \sum_{p=1}^{m_1(s)} \sum_{q=1}^{m_2(s)} p(s'|s,a_p^{(1)},a_q^{(2)})\,\mathbf{f}(a_p^{(1)}|s)\,\mathbf{g}(a_q^{(2)}|s)$$

$\forall s,s' \in S$, while the average immediate reward $r(s,\mathbf{f},\mathbf{g})$ equals

$$r(s,\mathbf{f},\mathbf{g}) = \sum_{p=1}^{m_1(s)} \sum_{q=1}^{m_2(s)} r(s,a_p^{(1)},a_q^{(2)})\, f(a_p^{(1)}|s)\, g(a_q^{(2)}|s)$$

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 a column vector of length N such that its i-th component 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 average reward vector associated to the couple of strategies $(\mathbf{f},\mathbf{g})$ respectively.

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

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

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

$$\Phi_{\overline{\beta}}(\mathbf{f}^*_{\overline{\beta}},\mathbf{g}) \geq \Phi_{\overline{\beta}}(\Gamma) \qquad \forall \mathbf{g} \in \mathbf{G}$$

while $\mathbf{g}^*_{\overline{\beta}}$ is optimal for player 2 iff

$$\Phi_{\overline{\beta}}(\mathbf{f},\mathbf{g}^*_{\overline{\beta}}) \leq \Phi_{\overline{\beta}}(\Gamma) \qquad \forall \mathbf{f} \in \mathbf{F}.$$

Let $\Phi(\mathbf{f},\mathbf{g})$ be the long term average value of the game Γ associated to the couple 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 $\Phi(\Gamma)$ be the value vector for the long term 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 (Filar and Vrieze, 1996):

Theorem 2.1. *Under the hypothesis of discounted pay-off, 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 3. *A stationary strategy* $\mathbf{h}$ *is said to be uniformly discount optimal for a player if* $\mathbf{h}$ *is optimal 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 perfect information stochastic games.

Definition 4. *Under the hypothesis of perfect information, in each state at most one player has more than one action available.*

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

3 The ordered field of rational functions with real coefficients

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

Definition 5. *The dominating coefficient of a polynomial* $f = a_0 + a_1 x + \cdots + a_n x^n$ *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 $\mathbb{R}$:

$$f(x) = \frac{c_0 + c_1 x + \cdots + c_n x^n}{d_0 + d_1 x + \cdots + d_m x^m} \qquad f \in F(\mathbb{R})$$

where the operations of sum and product are defined in the usual way (see Hordijk, Dekker and Kallenberg, 1985). Two rational functions $h/g,\ p/q$ are identical (and we say $h/g =_l p/q$) if and only if $h(x)q(x) = p(x)g(x)\ \forall x \in \mathbb{R}$.

The following lemma (Hordijk et al., 1985) introduces the ordering in the field $F(\mathbb{R})$:

Lemma 3.1. *A complete ordering in* $F(\mathbb{R})$ *is obtained by the rule*

$$\frac{p}{q} >_l 0 \iff \mathscr{D}(p)\mathscr{D}(q) > 0 \qquad p, q \in P(\mathbb{R})$$

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

The ordering law defined above is useful when one wants to compare the behavior of rational functions whose indipendent variable is positive and approaches to 0 (see Hordijk et al., 1985).

Lemma 3.2. *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 Application to stochastic games

From the next theorems the reader will start perceiving the importance of dealing with the field $F(\mathbb{R})$ in stochastic games.

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

Proof. For any couple of stationary strategies $(\mathbf{f},\mathbf{g})$, we can write

$$\sum_{s'=1}^{N} [(1+\rho)\delta_{s,s'} - p(s'|s,\mathbf{f},\mathbf{g})]\Phi_\rho(\mathbf{f},\mathbf{g},s') = (1+\rho)r(s,\mathbf{f},\mathbf{g}) \quad s \in [1;N] \tag{2}$$

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

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 Filar and Vrieze, 1996). From Theorems 2.1 and 3.3 it is straightforward to obtain the following important Lemma.

Lemma 3.4. *Let* Γ *be a zero-sum stochastic game which possesses uniform discount optimal strategies for both players. Then, there exist* $\overline{\rho}^* > 0$ *and* $\Phi^*_\rho(\Gamma) \in F(\mathbb{R})$ *such that* $\Phi^*_{\overline{\rho}}(\Gamma)$ *is the discounted optimal value for all the interest rates* $\overline{\rho} \in (0;\overline{\rho}^*]$.

Proof. Let $(\mathbf{f}^*,\mathbf{g}^*)$ be a couple of uniformly discount optimal strategies for players 1 and 2 respectively. Then, by definition, there exists $\overline{\rho}^* > 0$ such that $(\mathbf{f}^*,\mathbf{g}^*)$ are discounted optimal for all the interest rates $\overline{\rho} \in (0;\overline{\rho}^*]$. From Theorem 3.3 we know that $\Phi_\rho(\mathbf{f}^*,\mathbf{g}^*) \in F(\mathbb{R})$ and, from Theorem 2.1, the optimum uniform discounted value $\Phi_{\overline{\rho}}(\Gamma) = \Phi_{\overline{\rho}}(\mathbf{f}^*,\mathbf{g}^*)\ \forall \overline{\rho} \in (0;\overline{\rho}^*]$. So, $\Phi^*_\rho(\Gamma) \in F(\mathbb{R})$ represents the discounted value of Γ for all the interest rates sufficiently close to 0. □

Lemma 3.5. *Let* Γ *be a zero-sum stochastic game which possesses uniform discount optimal strategies* $\mathbf{f}^*,\mathbf{g}^*$ *for players 1 and 2 respectively. Then,*

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

where

$$\Phi^*_\rho(\Gamma) =_l \max_{\mathbf{f}}{}_l \min_{\mathbf{g}}{}_l \Phi_\rho(\mathbf{f},\mathbf{g}) =_l \min_{\mathbf{g}}{}_l \max_{\mathbf{f}}{}_l \Phi_\rho(\mathbf{f},\mathbf{g}). \tag{4}$$

Proof. From Theorem 2.1 and by the definition of uniform discount optimal strategy, we assert that

$$\exists \overline{\rho}^* > 0 :\ \forall \overline{\rho} \in (0;\overline{\rho}^*] \ \Rightarrow\ \Phi_{\overline{\rho}}(\mathbf{f},\mathbf{g}^*) \leq \Phi_{\overline{\rho}}(\mathbf{f}^*,\mathbf{g}^*) \leq \Phi_{\overline{\rho}}(\mathbf{f}^*,\mathbf{g}) \quad \forall \mathbf{f},\mathbf{g}$$

which coincides with (3) for Lemma 3.2. The equation (4) is a direct consequence of (3). □

Definition 6. $\Phi^*_\rho(\Gamma)$*, defined as in* (4)*, is the uniform discount value of the stochastic game* Γ.

4 Computation of Blackwell optimum policy in MDPs

In this section we will discuss about some concepts of linear programming, which can be easily found on any book on linear optimization (e.g. see Luenberger and Ye 2008).

Let Ψ be a Markov Decision Process, which can be seen as a two-player stochastic game in which one of the two players either fixes his own strategy or has only one available action in each state. We call $\Phi_\rho(\mathbf{f})$ the value of the discounted MDP associated to the strategy $\mathbf{f}$ with interest rate variable ρ.

It is known (Puterman, 1994) that the interval of interest rate $(0;\infty)$ can be broken into a finite number n of subintervals, say $(0\equiv\alpha_0;\alpha_1],(\alpha_1;\alpha_2],\ldots,(\alpha_{n-1};\infty)$ in such a way that for each one there exists an optimal pure strategy.

A Blackwell optimal policy is an optimal strategy associated to the first sub-interval.

Definition 7. *We say that the strategy $\mathbf{f}^*$ is Blackwell optimal iff there exists $\bar\rho^*>0$ such that $\mathbf{f}^*$ is optimal in the $(1/\bar\rho-1)$-discounted MDP for all the interest rates $\bar\rho\in(0;\bar\rho^*]$.*

Since for Theorem 3.3 $\Phi_\rho(\mathbf{f})\in F(\mathbb{R})$ for any $\mathbf{f}\in\mathbf{F}_S$, we can say

$$\Phi_\rho(\mathbf{f}^*)\geq_l\Phi_\rho(\mathbf{f})\qquad\forall\mathbf{f}\in\mathbf{F}$$

where $\mathbf{F}$ is the set of all possible strategies.
Hordijk, Dekker and Kallenberg (1985) provided a useful algorithm to compute the Blackwell optimum policy in MDPs. It consists in solving the following parametric linear programming problem:

$$\begin{cases}\max\limits_{\mathbf{x}}{}_l\ \sum_{s=1}^N\sum_{a=1}^{m(s)}x_{sa}(\rho)r(s,a)\\ \sum_{s=1}^N\sum_{a=1}^{m(s)}[(1+\rho)\delta_{s,s'}-p(s'|s,a)]\,x_{s,a}(\rho)=_l 1,\quad s'\in S\\ x_{s,a}(\rho)\geq_l 0,\quad s\in S,\ a\in A(s)\end{cases}\tag{5}$$

in the ordered field of rational functions with real coefficients $F(\mathbb{R})$. This means that

i) ρ is the variable of polynoms;

ii) all the elements of the related simplex tableau belong to $F(\mathbb{R})$;

iii) all the algebraic and ordering operations required by the simplex method are carried out in the field $F(\mathbb{R})$.

The practical technique to solve the linear optimization problem (5) proposed by Hordijk et al. (1985) is the so-called *two-phases method*.
In the *first phase* the artificial variables $z_1,\ldots,z_N$ are introduced as basic variables and the tableau of the following linear programming problem

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

is built. Then, N successive pivot operations on all the artificial variables are carried out so that the feasibility of the solution is preserved. We call *entering variables* the basic variables of the tableau at the end of the first phase. In the *second phase* the columns of the tableau associated to the artificial variables $z_1, \ldots, z_N$ (which are now all non-basic) are removed and the simplex method is performed in the ordered field $F(\mathbb{R})$ on the obtained tableau.

We note that another approach for the solution of the parametric linear program (5) is given by simplex method in the field of Laurent series (see Filar, Altman and Avrachenkov, 2002).

The optimal Blackwell stationary pure strategy $\mathbf{f}^*$ is computed as:

$$\mathbf{f}^*(a|s) = \frac{x^*_{s,a}(\rho)}{\sum_{a=1}^{m(s)} x^*_{s,a}(\rho)} \quad \forall s \in S,\ a \in A(s) \tag{7}$$

where $\{x^*_{s,a}(\rho)\ \forall s,a\}$ is the solution of the optimization problem. The simplex method guarantees that the optimum strategy $\mathbf{f}^*$ is well-defined and pure (see Filar and Vrieze 1996).

5 Uniform optimality in perfect information games

As we said before, in a perfect information game in each state at most one player has more than one action available. A stationary strategy for the player $i = 1,2$ is a function $\mathbf{f}_i : S \to \bigcup_{k=1}^{N} A_i(s_k)$ with $f_i(.|s_t) \in A_i(s_t)$.

Theorem 5.1. *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.*

The Theorem 5.1 (see Filar and Vrieze, 1996) guarantees the existence of the optimal strategies for both players in the average criterion for games with perfect information. Moreover, it suggests that in order to find the optimal strategies for the average criterion one has to find the optimal strategies in the discounted criterion for a discount factor sufficiently close to 1.

Definition 8. *We call two pure stationary strategies adjacent if and only if they differ only in one state.*

Then the following property holds, which proof is analogous to the one in the field of real numbers.

Lemma 5.2. *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 property above allows us to give the following definition.

Definition 9. *Let* $(\mathbf{f}, \mathbf{g})$ *be a pair of pure stationary strategy respectively for player 1 and 2. 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 results.

Lemma 5.3. *Let Γ be a perfect information stochastic game. A couple 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.*

Proof. The *only if* implication is obvious. If the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are such that no uniform adjacent improvements are possible for both players, then no improvements are possible also for the first stage of the game too, that is

$$\mathbf{f}^*(s) = \underset{a \in A_1(s)}{\operatorname{argmax}}_l \left\{ r(s,a) + (1+\rho)^{-1} \sum_{s'=1}^{N} p(s'|s,a) \Phi_\rho(s', \mathbf{f}^*, \mathbf{g}^*) \right\} \quad s \in S_1$$

$$\mathbf{g}^*(s) = \underset{a \in A_2(s)}{\operatorname{argmin}}_l \left\{ r(s,a) + (1+\rho)^{-1} \sum_{s'=1}^{N} p(s'|s,a) \Phi_\rho(s', \mathbf{f}^*, \mathbf{g}^*) \right\} \quad s \in S_2$$

It is known (see Filar and Vrieze, 1996) that if the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ satisfy such equations then they are uniform discount optimal. □

In perfect information games, the following result (see Raghavan and Syed, 2002) holds

Lemma 5.4. *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 5.5. *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. The *if* condition is less obvious. 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 5.4, we can say that $\forall \overline{\rho} \in (0; \rho^*]$ the strategies $\mathbf{f}^*, \mathbf{g}^*$ are optimal in the discounted game Γ, which means that they are discount optimal. □

Let s_t be a state controlled by player i ($i = 1, 2$) 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. Analogously to the result of Raghavan and Syed (2002), we propose the following Lemma.

Lemma 5.6. *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 $\Phi^*_\rho(\Gamma^t_{X \cup Y}) \in F(\mathbb{R})$, which is the uniform value of the game $\Gamma^t_{X \cup Y}$, equals*

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

Proof. Let us suppose that the state s_t is controlled by player 2. We indicate with $\mathbf{G}_X^t$ the set of pure stationary strategies in which the choice in state s_t is restricted to the set X. We note that the restriction in state s_t does not affect player 1. Thus, $\mathbf{F}_X^t = \mathbf{F}$.
If it is possible to find optimal strategies for player 2 both in $\mathbf{G}_X^t$ and in $\mathbf{G}_Y^t$, then $\Phi_\rho^*(\Gamma_X^t) =_l \Phi_\rho^*(\Gamma_Y^t) =_l \Phi_\rho^*(\Gamma_{X\cup Y}^t)$ for Lemma 5.5.
Otherwise, the uniform discount pure strategy of game $\Gamma_{X\cup Y}^t$ for player 2 belongs either to $\mathbf{G}_X^t$ or to $\mathbf{G}_Y^t$. For example, let us suppose that the optimal discount strategy in the stochastic game $\Gamma_{X\cup Y}^t$ for player 2 is found in Y. Then we have

$$\begin{aligned}
\Phi_\rho^*(\Gamma_Y^t) &=_l \Phi_\rho^*(\Gamma_{X\cup Y}^t) \\
&=_l \min_{\mathbf{g}\in\mathbf{G}}{}_l \max_{\mathbf{f}\in\mathbf{F}}{}_l \Phi_\rho(\mathbf{f},\mathbf{g}) \\
&\leq_l \min_{\mathbf{g}\in\mathbf{G}_\mathbf{X}^\mathbf{t}}{}_l \max_{\mathbf{f}\in\mathbf{F}}{}_l \Phi_\rho(\mathbf{f},\mathbf{g}) \\
&=_l \Phi_\rho^*(\Gamma_X^t)
\end{aligned}$$

The proof for the situation in which $s_t \in S_1$ is analogous. □

6 Algorithm description

Our task is to find an algorithm which allows to find the uniform discount optimal strategies for both players in a perfect information stochastic game Γ, which coincide with the optimal strategies for the long term average criterion for Theorem 5.1. Following the lines of the algorithm of Raghavan and Syed (2002) for optimal discount strategy, we propose an algorithm suitable to the ordered field $F(\mathbb{R})$.

Let Γ be a zero-sum two-player stochastic game with perfect information.

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

Remark 1. *Unlike Raghavan and Syed's solution, the algorithm* **??** *does not require the strategy search for player 1 to be lexicographic. Player 1, in fact, faces in step 2 a classic Blackwell optimization.*

Remark 2. *Obviously, the roles of player 1 and 2 can be swapped in the algorithm* **??**. *For simplicity, throughout the paper the player 1 will be assigned to step 2.*

Remark 3. *In step 3, 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 reduced cost $c_{s_{t_1+k},a} \leq_l 0$.*

Now, we prove the appropriateness of the algorithm **??**. The proof is analogous to the one by Raghavan and Syed (2002).

Step 1 Choose 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:

$$\begin{cases} \max\limits_{\mathbf{x}}{}_l \sum_{s=1}^{N} \sum_{a=1}^{m_1(s)} x_{s,a}(\rho) r(s,a,\mathbf{g}) \\ \sum_{s=1}^{N} \sum_{a=1}^{m_1(s)} [(1+\rho)\delta_{s,s'} - p(s'|s,a,\mathbf{g})] x_{s,a}(\rho) =_l 1, \quad s' \in S \\ x_{s,a}(\rho) \geq_l 0, \quad s \in S,\ a \in A_1(s) \end{cases} \tag{8}$$

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

$$f(a|s) = \frac{x^*_{s,a}(\rho)}{\sum_{a=1}^{m_1(s)} x^*_{s,a}(\rho)} \qquad \forall s \in S,\ a \in A_1(s) \tag{9}$$

where $\{x^*_{s,a}(\rho),\ \forall s,a\}$ is the solution of (8).

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:

$$\begin{cases} \max\limits_{\mathbf{x}}{}_l -\sum_{s=1}^{N} \sum_{a=1}^{m_2(s)} x_{s,a}(\rho) r(s,\mathbf{f},a) \\ \sum_{s=1}^{N} \sum_{a=1}^{m_2(s)} [(1+\rho)\delta_{s,s'} - p(s'|s,\mathbf{f},a)] x_{s,a}(\rho) =_l 1, \quad s' \in S \\ x_{s,a}(\rho) \geq_l 0, \quad s \in S,\ a \in A_2(s) \end{cases} \tag{10}$$

where the entering variables are $\{x_{s,a} :\ g(a|s) = 1,\ \forall s\}$.
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 term average optimal in the stochastic game Γ respectively for player 1 and player 2.

□

Theorem 6.1. *The algorithm stops in a finite time and the couple of strategies* $(\mathbf{f}^*, \mathbf{g}^*)$ *are uniform discount optimal in the stochastic game* Γ.

Proof. We assume that the overall number of actions

$$\mu = \sum_{k=1}^{t_1} m_1(s_k) + \sum_{k=1}^{t_2} m_2(s_{k+t_1})$$

is finite.
Without loss of generality, let us reorder the states so that in the first t_1 states player 1 has more than one action and the second t_2 states are controlled by player 2. Of course, $t_1+t_2 \leq N$.

We can proceed by induction on μ. Trivially $\mu \geq 2N$, because $\mu = 2N$ is equivalent to the situation $t_1 = t_2 = 0$. In this case the algorithm finds the average optimal couple of strategies because it is the only existing.

Now we suppose by induction that the algorithm finds *without cycling* (that is, all pure stationary strategies are visited at most once) the couple of uniform optimal strategies when the number of actions is $\overline{\mu} \geq 2N$. We have to prove that the thesis is valid when the number of actions equals $\overline{\mu}+1$.

If $t_2 = 0$, then again there is nothing to prove, because, as we showed in section 4, the step 1 of our algorithm finds the Blackwell optimal policy $\mathbf{f}^*$ for player 1 in the MDP $\Gamma_1(g)$.

If $t_2 \geq 0$, then we focus on the state $s_{t_1+t_2} = s_\tau$, which is the last examined by our algorithm. The actions available in state s_τ are $A_2(s_\tau) \equiv X \cup a_i$, where $X = \{a_1 \dots a_{i-1}, a_{i+1} \dots a_n\}$ and $n \geq 2$ by hypothesis. By induction hypothesis, we suppose that the algorithm finds the uniform discount optimal strategies for both players in the game Γ_X^τ without cycling. Since no uniform improvements are possible in Γ_X^τ by definition of uniform optimal strategies, then the algorithm looks for an uniform adjacent improvement $\mathbf{g}'$, where $\mathbf{g}'(a_i|s_\tau) = 1$. There are now two possibilities.

If the uniform optimal strategy $\mathbf{g}$ for player 2 found in Γ_X^t is also optimal in Γ, then the algorithm terminates because still no adjacent improvements are possible for player 2 in s_t.

Otherwise, any uniform optimal strategy $\mathbf{g}^*$ for player 2 in Γ includes the action a_τ and the algorithm necessarily finds an adjacent improvement in state s_τ for Theorem 5.3 and it finds by induction hypothesis the uniform discount optimal strategies in the game $\Gamma_{a_n}^t$. So we have

$$\Phi_\rho(\Gamma) =_l \min_l \{\Phi_\rho(\Gamma_X^t), \Phi_\rho(\Gamma_{a_n}^t)\} =_l \Phi_\rho(\Gamma_{a_n}^t)$$

where the second equality holds because otherwise the optimal strategies of Γ_X^t would be uniform optimal in the game Γ for Lemma 5.5. Again thanks to Lemma 5.5, we can assert that the uniform discount optimal strategies $(\mathbf{f}^*, \mathbf{g}^*)$ found in $\Gamma_{a_n}^t$ are optimal also for Γ, because $\Phi_\rho(\mathbf{f}^*, \mathbf{g}^*) = \Phi_\rho^*(\Gamma)$, which is the uniform discount value of the game.
Moreover, the algorithm terminates because for Theorem 5.3 no improvements are available to both players.

We gave a constructive proof of the fact that the algorithm passes through a path of pure strategies, it never cycles and it finds the uniform discount optimal strategies for both players. Since the

overall number of actions is finite, then also the cardinality of pure strategies is finite; hence, the algorithm must terminate in a finite time and the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are uniform discount optimal, and for Theorem 5.1 they are long term average optimal as well. □

6.1 Computing the optimality range factor

The algorithm presented in section 6 suggests a way to determine the range of discount factor in which the long term average optimal strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are also optimal in the discounted game. Before, we report the analogous result to Lemma 5.3 when the discount factor is fixed (see Raghavan and Syed, 2002).

Lemma 6.2. *Let Γ be a perfect information stochastic game and $\overline{\beta} \in [0;1)$. 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 $\frac{df_\rho}{d\rho}\big|_{\rho=u} < 0$, $\forall u \in \zeta(f_\rho)$. Now we are ready to state the following Lemma.

Lemma 6.3. *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 the algorithm* **??** *and*

$$\overline{\rho}^* = \min_c \zeta(c), \quad c \in C.$$

Then, $\overline{\beta}^* = (1+\overline{\rho}^*)^{-1}$ is the smallest value such that the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are $\overline{\beta}$-discount optimal in the game Γ, $\forall \overline{\beta} \in [\overline{\beta}^*; 1)$.

Proof. The existence of such $\overline{\rho}^*$ is guaranteed by Theorem 5.1. For all the value of the interest factor $\overline{\rho} \in (0; \overline{\rho}^*]$, the reduced costs are positive, hence no adjacent improvements are possible for both players. So, for Lemma 6.2 they are discounted optimal. If $\overline{\rho} > \overline{\rho}^*$ and $\overline{\rho}^* < \infty$, 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 $\overline{\beta} = (1+\overline{\rho})^{-1}$. □

6.2 Round-off errors sensitivity

The role of the first non-null coefficients of the polynomials (numerator and denominator) of the tableaux obtained throughout the algorithm unfolding is essential: they determine the positiveness of the elements of the tableaux themselves in the field $F(\mathbb{R})$. This knowledge is fundamental to choose the most suitable pivot elements.
The reader can easily understand that the algorithm is highly sensitive to the round-off errors that affect the null coefficients.

If the data of the problem (rewards and transition probabilities for each strategy) 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.

Table 1: *Immediate rewards and transition probabilities for each player, state and strategy.*

	(s,a)	r	$p(s_1\|s)$	$p(s_2\|s)$	$p(s_3\|s)$	$p(s_4\|s)$	$p(s_5\|s)$
pl. 1	(1,1)	5	0	0	0	0	1
	(1,2)	4	0	0	0.2	0	0.8
	(1,3)	3	0	0	0.6	0	0.4
	(2,1)	6	0	0	0	0	0.1
	(2,2)	1	1	0	0	0	0
	(2,3)	0	0	0	0.1	0	0
pl. 2	(3,1)	4	0	0	0	0.9	0.1
	(3,2)	2	0.1	0	0	0	0
	(3,3)	0	0.3	0	0.2	0.5	0
	(4,1)	2	0	0.1	0.6	0.3	0
	(4,2)	2	0.2	0	0.4	0.4	0
	(4,3)	3	0	0	0	0.9	0.1
	5	0	0	0.1	0.2	0.3	0.4

Instead, if the data are irrational, a simple way to circumvent the round-off errors is to fix a tolerance value ε, and set to 0 all the polynomial coefficients of the tableaux obtained throughout the algorithm whose absolute value is smaller than ε.

7 An example

Here we present a run of our algorithm **??**, where the input data are taken from Raghavan and Syed (2002). There are 5 states, the first two are controlled by player 1 and states 3 and 4 are for player 2; in the final state both players have no action choice. The immediate rewards and the probability transitions for every couple (state,action) for both players are shown in table 1.

We choose the initial strategy $(g(a_2|s_3) = 1, g(a_3|s_4) = 1)$ for player 2. We report the optimum tableau obtained by player 1 at the end of step 2 of the first iteration of our algorithm (tab.4) and the tableau of player 2 after the first improvement at step 3 (tab.5). Analogously, the tableaux 6 and 7 are associated to the second and last iteration of our algorithm. It is known (see Hordijk et al. 1985) that all the elements of simplex tableaux have a common denominator, stored in the top left-hand box. The last column of each tableau contains the numerator of the value of the basic variables, which are listed in the first column. The last row indicates the numerator of the reduced costs.

The optimum long term average strategy for player 1 is $f^*(a_1|s_1) = 1, f^*(a_2|s_2) = 1$, and for player 2 is $g^*(a_2|s_3) = 1, g^*(a_1|s_4) = 1$.

By computing the first positive root of the reduced costs of the two last optimal tableaux we find that the strategies $(\mathbf{f}^*, \mathbf{g}^*)$ are also $\overline{\beta}$-discount optimal for all the discount factor $\beta \in [\overline{\beta}^*; 1)$, where $\overline{\beta}^* \cong 0.74458$.

Note that the optimal strategies differ from the ones of Raghavan and Syed (2002), in which the discount factor is set to 0.999. We suspect that this is due to some clerical errors.

8 A lower complexity algorithm

Let Γ be a zero-sum two-player stochastic game with perfect information. Consider the following algorithm: This is essentially a best reponse algorithm, in which at each step each player alternatively

---

Step 1 Choose 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.

---

looks for his own Blackwell optimal strategy.
Obviously, if the above algorithm stops, $(\mathbf{f}^*, \mathbf{g}^*)$ forms a couple of uniform discount and long term average optimal strategies, since they are both Blackwell optimal in the respective MDPs, $\Gamma_1(\mathbf{g}^*)$ and $\Gamma_2(\mathbf{f}^*)$.
The proof that the algorithm **??** never cycles is still an open problem. It is quite natural to try to prove that $\Phi_\rho(\mathbf{f}_{k+1}, \mathbf{g}_{k+1}) \leq_l \Phi_\rho(\mathbf{f}_k, \mathbf{g}_k)$, but it is not difficult to find a counterexample.
Raghavan and Syed (2002) conjecture as follows:

Conjecture 8.1. *Let Γ be a two-player zero-sum stochastic game with perfect information and $\alpha = (\mathbf{f}, \mathbf{g})$ a couple of pure stationary strategies for the 2 players. For every discount factor $\overline{\beta} \in [0; 1)$, there are no sequences $\alpha_0, \alpha_1, \ldots, \alpha_k$ such that $\Phi_{\overline{\beta}}(\alpha_k) = \Phi_{\overline{\beta}}(\alpha_0)$, where α_i is an adjacent improvement with respect to α_{i-1} in the $\overline{\beta}$-discounted stochastic game Γ for only one player for any $i > 0$.*

If Conjecture 8.1 were valid, then we could conclude that the algorithm **??** terminates in finite time.

8.1 Complexity

In our first algorithm **??**, player 1 faces at each step 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 time.

Player 2 lexicographically expands his search of his optimum strategy, and at the k-th iteration the two players find the solution of a subgame Γ_k which monotonically tends to the entire stochastic game Γ.

Analogously to what Raghavan and Syed (2002) remark, we can assert that the efficiency of our algorithm **??** is mostly due to the fact that most of the actions dominate totally other actions. In other words, it occurs very often that the optimum action $a^* \in A(s),\ s \in S$, found in an iteration k such that $A(s) \subset \Gamma_k$, is optimum also in Γ, and consequently remains the same in all the remaining iterations. This exponentially reduces the policy space in which the algorithm needs to search.

Remark 4. *As discussed in section 6, in the algorithm* **??** *players' roles are interchangeble. 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 greater.*

Differently from Raghavan and Syed (2002), the search for player 1 does not need to be lexicographic, and player 1 is left totally free to optimize the MDP that he faces at each iteration of the algorithm in the most efficient way.

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

M$_1$: Algorithm **??**, in which in step 2 player 1 pivots with respect to the variable with the minimum reduced cost until he finds his own Blackwell optimal strategy.

M$_2$: Algorithm **??**, 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 by Raghavan and Syed (2002), but in the field $F(\mathbb{R})$.

M$_3$: Algorithm **??**.

The results are shown in tables 2 and 3. The simulations were carried out on 10000 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.

Table 2: *Average number of pivotings for the 3 methods.*

	n. pivoting
M_1	40.59
M_2	41.87
M_3	24.93

It is evident that the algorithm M_3 is much faster than the other two, but unfortunately its convergence is not proven yet. However, in our numerical experiment with 10000 randomly generated stochastic games, it never cycles. The difference between M_1 and M_2 is due to the more efficient simplex method used by player 1 in M_1.

Table 3: *$M_i > M_j$ when, fixing the game, the number of pivotings in M_i is strictly smaller than the number of pivoting in M_j.*

$>$ (%)	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 4: *Optimum tableau for player 1 at the first iteration.*

$0.018+0.658\rho+3.07\rho^2+5.13\rho^3+3.7\rho^4+\rho^5$	$x_{1,2}$	$x_{1,3}$	$x_{2,1}$	$x_{2,3}$	
$x_{1,1}$	$0.0198+0.6698\rho+3.06\rho^2+5.11\rho^3+3.7\rho^4+\rho^5$	$0.0234+0.6934\rho+3.04\rho^2+5.07\rho^3+3.7\rho^4+\rho^5$	$0.0288+0.7468\rho+2.418\rho^2+2.7\rho^3+\rho^4$	$0.0297+0.7527\rho+2.413\rho^2+2.69\rho^3+\rho^4$	$0.087+1.707\rho+4.42\rho^2+3.8\rho^3+\rho^4$
$x_{2,2}$	$0.0018+0.022\rho+0.042\rho^2+0.02\rho^3$	$0.0054+0.066\rho+0.126\rho^2+0.06\rho^3$	$0.027+0.756\rho+3.149\rho^2+5.12\rho^3+3.7\rho^4+1\rho^5$	$0.0279+0.767\rho+3.17\rho^2+5.13\rho^3+3.7\rho^4+1\rho^5$	$0.059+\rho+2.75\rho^2+2.8\rho^3+\rho^4$
$x_{3,1}$	$-0.084\rho-0.402\rho^2-0.5\rho^3-0.2\rho^4$	$-0.252\rho-1.206\rho^2-1.5\rho^3-0.6\rho^4$	$0.018+0.196\rho+0.158\rho^2-0.02\rho^3$	$0.018+0.154\rho-0.043\rho^2-0.27\rho^3-0.1\rho^4$	$0.1+1.36\rho+3.07\rho^2+2.9\rho^3+\rho^4$
$x_{4,1}$	$0.054+0.174\rho+0.18\rho^2+0.06\rho^3$	$0.162+0.522\rho+0.54\rho^2+0.18\rho^3$	$0.27+0.51\rho+0.21\rho^2-0.03\rho^3$	$0.297+0.597\rho+0.3\rho^2$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+1\rho^4$
$x_{5,1}$	$0.018+0.238\rho+0.64\rho^2+0.62\rho^3+0.2\rho^4$	$0.054+0.714\rho+1.92\rho^2+1.86\rho^3+0.6\rho^4$	$0.09+1.07\rho+1.77\rho^2+0.69\rho^3-0.1\rho^4$	$0.099+1.189\rho+2.09\rho^2+\rho^3$	$0.41+4.01\rho+6.8\rho^2+4.2\rho^3+\rho^4$
	$0.1908+1.2838\rho+3.891\rho^2+7.028\rho^3+7.53\rho^4+4.3\rho^5+1\rho^6$	$0.5544+3.1754\rho+7.945\rho^2+12.884\rho^3+13.76\rho^4+8.2\rho^5+2\rho^6$	$0.909+3.373\rho+0.229\rho^2-14.525\rho^3-25.79\rho^4-18.5\rho^5-5\rho^6$	$1.1034+7.7329\rho+22.6785\rho^2+34.089\rho^3+26.54\rho^4+9.5\rho^5+1\rho^6$	$4.924+30.709\rho+74.775\rho^2+88.29\rho^3+50.3\rho^4+11\rho^5$

References

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

Table 5: *Optimum tableau for player 2 at the first iteration.*

$0.288+2.308\rho+6.04\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$x_{3,1}$	$x_{3,3}$	$x_{4,3}$	$x_{4,2}$	
$x_{1,1}$	$-0.0576-0.0416\rho+0.246\rho^2+0.33\rho^3+0.1\rho^4$	$-0.1404-0.5904\rho-0.93\rho^2-0.68\rho^3-0.2\rho^4$	$-0.0036+0.0964\rho+0.26\rho^2+0.16\rho^3$	$-0.0432-0.2472\rho-0.544\rho^2-0.54\rho^3-0.2\rho^4$	$1.11+4.54\rho+6.83\rho^2+4.4\rho^3+1\rho^4$
$x_{2,1}$	$-0.0576-0.208\rho-0.254\rho^2-0.1\rho^3$	$-0.054-0.152\rho-0.15\rho^2-0.05\rho^3$	$-0.0036+0.056\rho+0.216\rho^2+0.26\rho^3+0.1\rho^4$	$0.0144+0.152\rho+0.356\rho^2+0.32\rho^3+0.1\rho^4$	$0.662+2.85\rho+4.68\rho^2+3.5\rho^3+\rho^4$
$x_{3,2}$	$1.088\rho+4.584\rho^2+6.76\rho^3+4.3\rho^4+1\rho^5$	$1.136\rho+4.416\rho^2+6.36\rho^3+4.1\rho^4+1\rho^5$	$0.368\rho+1.404\rho^2+1.6\rho^3+0.6\rho^4$	$0.192\rho+0.608\rho^2+0.6\rho^3+0.2\rho^4$	$1.6+4.92\rho+6.5\rho^2+4.1\rho^3+\rho^4$
$x_{4,1}$	$-0.432-2.442\rho-4.38\rho^2-3.27\rho^3-0.9\rho^4$	$-0.306-1.466\rho-2.46\rho^2-1.8\rho^3-0.5\rho^4$	$0.018+0.658\rho+3.07\rho^2+5.13\rho^3+3.7\rho^4+\rho^5$	$0.216+1.956\rho+5.5\rho^2+6.96\rho^3+4.2\rho^4+\rho^5$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+\rho^4$
$x_{5,1}$	$-0.144-0.214\rho-0.24\rho^2-0.27\rho^3-0.1\rho^4$	$-0.234-0.594\rho-0.56\rho^2-0.2\rho^3$	$-0.054-0.134\rho-0.35\rho^2-0.37\rho^3-0.1\rho^4$	$-0.072-0.292\rho-0.42\rho^2-0.2\rho^3$	$2.33+7.53\rho+8.9\rho^2+4.7\rho^3+1\rho^4$
	$2.3616+14.7176\rho+35.132\rho^2+43.526\rho^3+30.65\rho^4+11.9\rho^5+2\rho^6$	$1.368+5.132\rho+4.652\rho^2-4.782\rho^3-11.87\rho^4-8.2\rho^5-2\rho^6$	$0.8496+5.1836\rho+11.99\rho^2+15.096\rho^3+11.64\rho^4+5.2\rho^5+1\rho^6$	$0.3456+1.7496\rho+3.632\rho^2+4.128\rho^3+2.6\rho^4+0.7\rho^5+2.3008e-006\rho^6$	$-12.232-56.642\rho-108.24\rho^2-105.33\rho^3-51.5\rho^4-10\rho^5$

Table 6: *Optimum tableau for player 1 at the second iteration.*

$0.288+2.308\rho+6.04\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$x_{1,2}$	$x_{1,3}$	$x_{2,1}$	$x_{2,3}$	
$x_{1,1}$	$0.306+2.324\rho+6.018\rho^2+7.3\rho^3+4.3\rho^4+\rho^5$	$0.342+2.356\rho+5.974\rho^2+7.26\rho^3+4.3\rho^4+\rho^5$	$0.4068+2.1148\rho+4.008\rho^2+3.3\rho^3+\rho^4$	$0.4158+2.1228\rho+3.997\rho^2+3.29\rho^3+\rho^4$	$1.11+4.54\rho+6.83\rho^2+4.4\rho^3+\rho^4$
$x_{2,2}$	$0.018+0.058\rho+0.06\rho^2+0.02\rho^3$	$0.054+0.174\rho+0.18\rho^2+0.06\rho^3$	$0.378+2.478\rho+6.11\rho^2+7.31\rho^3+4.3\rho^4+\rho^5$	$0.387+2.507\rho+6.14\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$0.662+2.85\rho+4.68\rho^2+3.5\rho^3+\rho^4$
$x_{3,1}$	$-0.24\rho-0.66\rho^2-0.62\rho^3-0.2\rho^4$	$-0.72\rho-1.98\rho^2-1.86\rho^3-0.6\rho^4$	$0.288+0.436\rho+0.128\rho^2-0.02\rho^3$	$0.288+0.316\rho-0.202\rho^2-0.33\rho^3-0.1\rho^4$	$1.6+4.92\rho+6.5\rho^2+4.1\rho^3+\rho^4$
$x_{4,1}$	$0.054+0.174\rho+0.18\rho^2+0.06\rho^3$	$0.162+0.522\rho+0.54\rho^2+0.18\rho^3$	$0.27+0.51\rho+0.21\rho^2-0.03\rho^3$	$0.297+0.597\rho+0.3\rho^2$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+\rho^4$
$x_{5,1}$	$0.126+0.586\rho+\rho^2+0.74\rho^3+0.2\rho^4$	$0.378+1.758\rho+3\rho^2+2.22\rho^3+0.6\rho^4$	$0.63+2.09\rho+2.19\rho^2+0.63\rho^3-0.1\rho^4$	$0.693+2.383\rho+2.69\rho^2+\rho^3$	$2.33+7.53\rho+8.9\rho^2+4.7\rho^3+\rho^4$
	$0.504+2.818\rho+7.344\rho^2+11.15\rho^3+10.02\rho^4+4.9\rho^5+\rho^6$	$1.224+5.858\rho+13.684\rho^2+20.09\rho^3+18.44\rho^4+9.4\rho^5+2\rho^6$	$1.8+2.896\rho-8.318\rho^2-29.624\rho^3-36.71\rho^4-21.5\rho^5-5\rho^6$	$3.636+18.583\rho+41.268\rho^2+49.431\rho^3+32.21\rho^4+10.1\rho^5+\rho^6$	$12.232+56.642\rho+108.24\rho^2+105.33\rho^3+51.5\rho^4+10\rho^5$

Table 7: *Optimum tableau for player 2 at the second iteration.*

$0.288+2.308\rho+6.04\rho^2+7.32\rho^3+4.3\rho^4+\rho^5$	$x_{3,1}$	$x_{3,3}$	$x_{4,2}$	$x_{4,3}$	
$x_{1,1}$	$-0.0576-0.0416\rho+0.246\rho^2+0.33\rho^3+0.1\rho^4$	$-0.1404-0.5904\rho-0.93\rho^2-0.68\rho^3-0.2\rho^4$	$-0.0432-0.2472\rho-0.544\rho^2-0.54\rho^3-0.2\rho^4$	$-0.0036+0.0964\rho+0.26\rho^2+0.16\rho^3$	$1.11+4.54\rho+6.83\rho^2+4.4\rho^3+\rho^4$
$x_{2,1}$	$-0.0576-0.208\rho-0.254\rho^2-0.1\rho^3$	$-0.054-0.152\rho-0.15\rho^2-0.05\rho^3$	$0.0144+0.152\rho+0.356\rho^2+0.32\rho^3+0.1\rho^4$	$-0.0036+0.056\rho+0.216\rho^2+0.26\rho^3+0.1\rho^4$	$0.662+2.85\rho+4.68\rho^2+3.5\rho^3+\rho^4$
$x_{3,2}$	$1.088\rho+4.584\rho^2+6.76\rho^3+4.3\rho^4+\rho^5$	$1.136\rho+4.416\rho^2+6.36\rho^3+4.1\rho^4+\rho^5$	$0.192\rho+0.608\rho^2+0.6\rho^3+0.2\rho^4$	$0.368\rho+1.404\rho^2+1.6\rho^3+0.6\rho^4$	$1.6+4.92\rho+6.5\rho^2+4.1\rho^3+\rho^4$
$x_{4,1}$	$-0.432-2.442\rho-4.38\rho^2-3.27\rho^3-0.9\rho^4$	$-0.306-1.466\rho-2.46\rho^2-1.8\rho^3-0.5\rho^4$	$0.216+1.956\rho+5.5\rho^2+6.96\rho^3+4.2\rho^4+\rho^5$	$0.018+0.658\rho+3.07\rho^2+5.13\rho^3+3.7\rho^4+\rho^5$	$1.41+4.51\rho+6\rho^2+3.9\rho^3+\rho^4$
$x_{5,1}$	$-0.144-0.214\rho-0.24\rho^2-0.27\rho^3-0.1\rho^4$	$-0.234-0.594\rho-0.56\rho^2-0.2\rho^3$	$-0.072-0.292\rho-0.42\rho^2-0.2\rho^3$	$-0.054-0.134\rho-0.35\rho^2-0.37\rho^3-0.1\rho^4$	$2.33+7.53\rho+8.9\rho^2+4.7\rho^3+\rho^4$
	$2.3616+14.7176\rho+35.132\rho^2+43.526\rho^3+30.65\rho^4+11.9\rho^5+2\rho^6$	$1.368+5.132\rho+4.652\rho^2-4.782\rho^3-11.87\rho^4-8.2\rho^5-2\rho^6$	$0.3456+1.7496\rho+3.632\rho^2+4.128\rho^3+2.6\rho^4+0.7\rho^5$	$0.8496+5.1836\rho+11.99\rho^2+15.096\rho^3+11.64\rho^4+5.2\rho^5+\rho^6$	$-12.232-56.642\rho-108.24\rho^2-105.33\rho^3-51.5\rho^4-10\rho^5$

[2] E. Altman, E. A. Feinberg, A. Shwartz, Weighted discounted stochastic games with perfect information, Annals of the International Society of Dynamic Games, Vol. 5, pp. 303-324 (2000).

[3] T. Bewley, E. Kohlberg, The asymptotic theory of stochastic games, Mathematics of Operations Research, Vol. 1, No. 3, pp. 197-208 (1976).

[4] K. Chatterjee, R. Majumdar, T.A. Henzinger, Stochastic limit-average games are in EXPTIME, International Journal of Game Theory, Vol. 37, No. 2, pp. 219-234 (2008).

[5] B.C. Eaves, U.G. Rothblum, Formulation of linear problems and solution by a universal machine, Mathematical Programming, Vol. 65, No. 1-3, pp. 263-309 (1994).

[6] J. Filar, K. Vrieze, Competitive Markov Decision Processes, Springer (1996).

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

[8] D. Gillette, Stochastic games with zero stop probabilities, Contributions to the theory of games, Princeton University Press, Vol. 3, pp. 179-187 (1957).

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

[10] D.G. Luenberger, Y. Ye, Linear and nonlinear programming (Third ed.), Springer (2008).

[11] J.F. Mertens, A. Neyman, Stochastic games, International Journal of Game Theory, Vol. 10, pp. 53-66 (1981).

[12] T. Parthasarathy, T.E.S. Raghavan, An orderfield property for stochastic games when one player controls transition probabilities, Journal of Optimization Theory and Applications, Vol. 33, No. 3, pp. 375-392 (1981).

[13] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley (1994).

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

[15] L.S. Shapley, Stochastic games, Proceedings of the National Academy of Sciences USA, Vol. 39, pp. 1095-1100 (1953).

[16] E. Solan, N. Vieille, Computing uniformly optimal strategies in two-player stochastic games, Economic Theory, Vol. 42, No. 1, pp. 237-253 (2010).

[17] F. Thuijsman, T.E.S. Raghavan, Perfect information stochastic games and related classes, International Journal of Game Theory, Vol. 26, pp. 403-408 (1997).

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399