STOCHASTIC GAMES AND STATIONARY STRATEGIES

O.J. VRIEZE
Maastricht University
Maastricht, The Netherlands

1. Introduction

In this chapter we treat stationary strategies and the limiting average criterion. Generally, for the average reward criterion, players need history-dependent strategies in playing nearly optimal or in equilibrium. A behavioral strategy may condition its choice of mixed action, at any given stage, on the entire history; and therefore its implementation is often a huge task. However, a stationary strategy conditions its choice of mixed action, at any given state, only on the present state. Since only as many decision rules as states need be remembered, a stationary strategy is preferable. In addition, ε-optimal stationary strategies are also ε-optimal in the model of the game where the players observe at every stage only the current state. These observation make it useful to know in which situations ε-optimal stationary strategies exist. This viewpoint is the main topic of this chapter.

Two-person zero-sum stochastic games with the limiting average criterion were introduced by Gillette [6]. He considered games with perfect information and irreducible stochastic games. For both classes both players possess average optimal stationary strategies. Blackwell and Ferguson [3] introduced the big match which showed that for limiting average stochastic games the value does not need to exist within the class of stationary strategies, and this result shows that history-dependent strategies are indispensable. Hoffman and Karp [7] considered irreducible stochastic games and gave an algorithm that yields ε-optimal stationary strategies. Starting with the paper of Parthasarathy and Raghavan [9], during the eighties several papers on special classes of stochastic games appeared. These classes were defined by conditions on the reward and/or transition structure. These classes were defined on the one hand to represent practical situations and on the other hand to derive classes of games for which the solution is relatively easy, i.e., in terms of stationary strategies. In this spirit we just

mention the papers of Parthasarathy et al. [10] and Vrieze et al. [14]. Finally, we mention the paper of Filar et al. [4], which examines from the computational viewpoint the possibilities of stationary strategies.

The chapter is built up around a limit theorem that connects limits of discounted rewards to average rewards when the discount factor tends to 0. This limit theorem turns out to be quite powerful, since many theorems concerning stationary strategies can now easily be proved. Among them are irreducible stochastic games and existence questions concerning (ε-)easy states. We end the chapter with some considerations regarding the relation between Puisseux series and existence of optimal stationary strategies for the limiting average criterion and the total reward criterion.

2. Stationary Strategies

In this section we consider stationary strategies. It will turn out that for stationary strategies the rewards for the different criteria can be written in a closed form and therefore stationary strategies are relatively easy to analyze.

Let us recall that a stationary strategy, say for player 1, was defined by a t-tuple $\alpha = (\alpha(1), \alpha(2), \ldots, \alpha(t))$ where $\alpha(z) = (\alpha(z,1), \alpha(z,2), \ldots, \alpha(z,m_z))$, for all $z \in S$, is a probability vector in the $I\!R^{m_z}$. The obvious implementation of such a strategy is that whenever the state of the system is in state z, player 1 will choose action a with probability $\alpha(z,a), a = 1, \ldots, m_z$. Here m_z denotes the number of actions of player 1 in state z.

For player 2 a stationary strategy is denoted by $\beta = (\beta(1), \beta(2), \ldots, \beta(t))$ where $\beta(z) =(\beta(z,1),\beta(z,2),\ldots,\beta(z,n_z))$ for all $z \in S$ and where $\beta(z)$ is a probability vector in the $I\!R^{n_z}$. The implementation of such a β is analogous to the implementation of α.

Now we investigate what will happen when the players implement the combination (α, β). Associated to (α, β) are a matrix $P(\alpha, \beta)$ and a vector $r(\alpha, \beta)$, which can be defined as follows:

$$\begin{aligned} P(\alpha,\beta) &= (p(z'|z,\alpha,\beta))_{z=1,z'=1}^{t,t}, \\ \text{where } p(z'|z,\alpha,\beta) &= \sum_{a=1}^{m_z}\sum_{b=1}^{n_z} p(z'|z,a,b)\alpha(z,a)\beta(z,b) \end{aligned}$$

$$\begin{aligned} r(\alpha,\beta) &= (r(1,\alpha,\beta), r(2,\alpha,\beta), \ldots, r(t,\alpha,\beta)) \\ \text{where } r(z,\alpha,\beta) &= \sum_{a=1}^{m_z}\sum_{b=1}^{n_z} r(z,a,b)\alpha(z,a)\beta(z,b). \end{aligned}$$

First observe that $p(z'|z,\alpha,\beta)$ and $r(z,\alpha,\beta)$ only depend on $\alpha(z)$ and $\beta(z)$. Next, by definition it follows that $p(z'|z,\alpha,\beta)$ is the expected probability that the system moves in one step to state z' whenever the system is in state z and the players play according to α and β. Likewise, $r(z,\alpha,\beta)$ denotes the expected payoff in state z.

Lemma 1 *Let $P^n(\alpha,\beta) := P(\alpha,\beta)(P^{n-1}(\alpha,\beta))$ with $n = 1,2,\ldots$, and where $P^0(\alpha,\beta) := I$. Then $p^n(z'|z,\alpha,\beta)$, being the (z,z')-th element of $P^n(\alpha,\beta)$, equals the probability that the system is in state z' after n steps, when the starting state was state z.*

Proof. By induction. For $n = 1$ the lemma is true by definition. Suppose the lemma is true for $n-1$. Observe that by definition

$$p^n(z'|z,\alpha,\beta) = \sum_{\widetilde{z}=1}^{t} p(\widetilde{z}|z,\alpha,\beta)p^{n-1}(z'|\widetilde{z},\alpha,\beta).$$

Given the induction hypothesis this expression denotes the sum of the probabilities of all the trajectories from z to z', where in the first step the system moves from state z to $\widetilde{z}$ and in the next $n-1$ steps from $\widetilde{z}$ to z'. Hence the lemma is true for n, which completes the proof. ∎

Now that we know the probabilities on the future states, the future expected payoffs can be expressed straightforwardly. First observe that in the z-th row of $P^n(\alpha,\beta)$ we find the n-step probabilities for the game that starts in state z. Then, the z-th component of the vector

$$P^n(\alpha,\beta)r(\alpha,\beta)$$

denotes the expected payoff at stage n when the starting state was z and (α,β) were implemented.

For the different criteria the expressions for the overall expected payoff follow easily:

Discounted rewards:

$$\begin{aligned}\gamma_\lambda(\alpha,\beta) &= \lambda\sum_{n=0}^{\infty}(1-\lambda)^n P^n(\alpha,\beta)r(\alpha,\beta)\\ &= \lambda(I-(1-\lambda)P(\alpha,\beta))^{-1}r(\alpha,\beta).\end{aligned}$$

Average rewards:

$$\begin{aligned}\gamma(\alpha,\beta)) &= \lim_{N\to\infty}\frac{1}{N+1}\sum_{n=0}^{N}P^n(\alpha,\beta)r(\alpha,\beta)\\ &= Q(\alpha,\beta)r(\alpha,\beta).\end{aligned}$$

Total rewards in case of finiteness:

$$\begin{aligned}\gamma_T(\alpha,\beta) &= \lim_{N\to\infty}\frac{1}{N+1}\sum_{n=0}^{N}\sum_{k=0}^{n}P^k(\alpha,\beta)r(\alpha,\beta)\\ &= (I-P(\alpha,\beta)+Q(\alpha,\beta))^{-1}r(\alpha,\beta).\end{aligned}$$

Weighted rewards, $\delta\in[0,1]$:

$$\gamma_\delta(\alpha,\beta)=\delta\gamma_\lambda(\alpha,\beta)+(1-\delta)\gamma(\alpha,\beta).$$

Some comments might be in order.

(i) That $\sum\limits_{n=o}^{\infty}(1-\lambda)^nP^n(\alpha,\beta)=(I-(1-\lambda)P(\alpha,\beta))^{-1}$ follows from the fact that

$$(I-(1-\lambda)P(\alpha,\beta))(\sum_{n=0}^{N}(1-\lambda)^nP^n(\alpha,\beta)) \;=$$

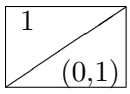
$$(\sum_{n=0}^{N}(1-\lambda)^nP^n(\alpha,\beta))(I-(1-\lambda)P(\alpha,\beta)) \;=\; I-(1-\lambda)^{N+1}P^{N+1}(\alpha,\beta)$$

while $\lim_{N\to\infty}(1-\lambda)^{N+1}P^{N+1}(\alpha,\beta)=0$.

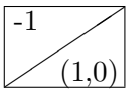
(ii) By definition $Q(\alpha,\beta):=\lim\limits_{N\to\infty}\frac{1}{N+1}\sum\limits_{n=0}^{N}P^n(\alpha,\beta)$, which is called the Cesaro limit.

For later use, it easily follows that $P(\alpha,\beta)Q(\alpha,\beta)=Q(\alpha,\beta)P(\alpha,\beta)=Q(\alpha,\beta)$.

(iii) For the total rewards the expression $\frac{1}{N+1}\sum\limits_{n=0}^{N}\sum\limits_{k=0}^{n}P^k(\alpha,\beta)r(\alpha,\beta)$ denotes the average of the first N partial sums of expected payoffs, while $\gamma_T(\alpha,\beta)$ is the limit of these numbers. That we use this definition for the total rewards and not simply $\lim\limits_{N\to\infty}\sum\limits_{n=0}^{N}P^n\,(\alpha,\beta)r(\alpha,\beta)$ is motivated, among other things, by the following zero-sum example:



state 1



state 2

Obviously, $\sum_{n=0}^{N}P^n(\alpha,\beta)r(\alpha,\beta)$ equals $(1,-1)$ for N even and (0,0) for N odd. Hence the limit does not exist, while $\gamma_T(\alpha,\beta)$ as defined above equals $(\frac{1}{2},-\frac{1}{2})$, which denotes the average possession of player 1.

Whenever the average reward is unequal to 0 the total reward will be $+\infty$ or $-\infty$ depending on the sign of the average reward. When the average reward for a pair of stationary strategies equals 0 we have a nice expression for $\gamma_T(\alpha,\beta)$ as given above, and which is proved in the next lemma.

Lemma 2 *Whenever* $Q(\alpha,\beta)r(\alpha,\beta)=0$ *we have that*

$$\gamma_T(\alpha,\beta)=(I-P(\alpha,\beta)+Q(\alpha,\beta))^{-1}r(\alpha,\beta) \quad .$$

Proof. In the proof we suppress the dependency of the variables on α and β. Let $\widetilde{\gamma}_T=(I-P+Q)^{-1}r$ or $r=(I-P+Q)\widetilde{\gamma}_T$. Multiplying this equation by Q gives $Q\widetilde{\gamma}_T=Qr$ (since $PQ=Q$ and $QQ=Q$), hence $Q\widetilde{\gamma}_T=0$. Then $(I-P)\widetilde{\gamma}_T=r$, or $\widetilde{\gamma}_T=r+P\widetilde{\gamma}_T$. Repeatedly iterating this last equation shows that

$$\widetilde{\gamma}_T=\sum_{k=0}^{n}P^k r+P^{n+1}\widetilde{\gamma}_T \qquad n=0,1,2,\ldots$$

and after averaging:

$$\widetilde{\gamma}_T=\frac{1}{N+1}\sum_{n=0}^{N}\sum_{k=0}^{n}P^k r+\frac{1}{N+1}\sum_{n=0}^{N}P^{n+1}\widetilde{\gamma}_T.$$

Taking limits we get

$$\widetilde{\gamma}_T=\gamma_T+Q\widetilde{\gamma}_T=\gamma_T.$$

■

(iv) The weighted reward is a convex combination of the discounted reward and the average reward. It can be interpreted as a balance between long-run incentives (the average reward) and short-sighted income (the discounted reward).

2.1. A LIMIT THEOREM

If we rearrange the states then the stochastic matrix $P(\alpha,\beta)$ can be written as

$$P(\alpha,\beta)=\begin{pmatrix} P_1(\alpha,\beta) & 0 & \cdots & \cdot & 0 \\ 0 & P_2(\alpha,\beta) & \cdots & \cdot & \cdot \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdot & \cdot & P_L(\alpha,\beta) & 0 \\ P_{L+11}(\alpha,\beta) & P_{L+12}(\alpha,\beta) & \cdots & P_{L+1L}(\alpha,\beta) & P_{L+1}(\alpha,\beta) \end{pmatrix}$$

For each $l \in \{1, \ldots, L\}$, $P_l(\alpha, \beta)$ is a square matrix and the states corresponding to this matrix form an ergodic class. We suppose that the Markov chain has L ergodic classes.

The remaining states, corresponding to the bottom block, are the transient states.

The Cesaro limit $Q(\alpha, \beta)$ has a similar shape:

$$Q(\alpha,\beta) = \begin{pmatrix} Q_1(\alpha,\beta) & 0 & \cdots & \cdot & 0 \\ 0 & Q_2(\alpha,\beta) & \cdots & \cdot & \cdot \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdot & \cdots & Q_L(\alpha,\beta) & 0 \\ Q_{L+11}(\alpha,\beta) & Q_{L+12}(\alpha,\beta) & \cdots & Q_{L+1L}(\alpha,\beta) & 0 \end{pmatrix}$$

In any textbook on Markov chains the following lemma can be found.

Lemma 3

(i) $Q_l(\alpha,\beta) = \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} P_l^n(\alpha,\beta)$.

(ii) $Q_l(\alpha,\beta)$ *has identical rows, say* $q_l(\alpha,\beta)$, *and* $q_l(\alpha,\beta)$ *is the unique probability vector that satisfies the equation* $q^T P_l(\alpha,\beta) = q^T$.

(iii) $I_{L+1} - P_{L+1}(\alpha,\beta)$ *is nonsingular.*

(iv) $Q_{L+1l}(\alpha,\beta) = (I_{L+1} - P_{L+1}(\alpha,\beta))^{-1} P_{L+1l}(\alpha,\beta) Q_l(\alpha,\beta)$, $l = 1, \ldots, L$.

(v) The solution set of the equation $q^T P(\alpha,\beta) = q^T$, *where* q *is required to be a probability vector, consists of the set*

$$\left\{ \sum_{l=1}^{L} \xi_l \tilde{q}_l(\alpha,\beta); \sum_{l=1}^{L} \xi_l = 1, \xi_l \geq 0 \right\}$$

where $\tilde{q}_l(\alpha,\beta)$ *is the extension of* $q_l(\alpha,\beta)$ *with zeros at the appropriate places in order to get a vector of length* t, *which is the number of states. Such a solution* q *is called an invariant distribution, and apparently the set of invariant distributions equals the convex hull of the vectors* $\tilde{q}_l(\alpha,\beta)$.

Now we are going to prove our limit theorem.

Theorem 1 *Let* $\lambda_n, n = 1, 2, \ldots$ *be a sequence of discount factors with* $\lim_{n\to\infty} \lambda_n = 0$. *Let* $(\alpha_{\lambda_n}, \beta_{\lambda_n}), n = 1, 2, \ldots$ *be a sequence of stationary strategies, such that, for each* z, *either* $\alpha_{\lambda_n}(z,a) = 0$ *for all* n *or* $\alpha_{\lambda_n}(z,a) > 0$ *for all* n *and likewise for* $\beta_{\lambda_n}(z,b)$, *and such that* $\lim_{n\to\infty} (\alpha_{\lambda_n}, \beta_{\lambda_n}) = (\alpha,\beta)$ *exists. Further assume that* $\gamma := \lim_{n\to\infty} \gamma_{\lambda_n}(\alpha_{\lambda_n}, \beta_{\lambda_n})$ *exists. Then we can write the* z*-th component of this limit vector as* $\gamma(z) = \sum_{l=1}^{L} \xi_{zl} \widetilde{\gamma}_l(\alpha,\beta)$ *with* $\xi_{zl} \geq 0$ *and* $\sum_{l=1}^{L} \xi_{zl} = 1$. *Here* $\widetilde{\gamma}_l(\alpha,\beta)$ *is the average reward (a number) for ergodic class* l.

Proof. Let $T(\lambda_n) = \lambda_n(I-(1-\lambda_n)P(\alpha_{\lambda_n},\beta_{\lambda_n}))^{-1}$. So $\gamma_{\lambda_n}(\alpha_{\lambda_n},\beta_{\lambda_n}) = T(\lambda_n)r(\alpha_{\lambda_n}\beta_{\lambda_n})$ and $\gamma = \lim_{n\to\infty} T(\lambda_n)r(\alpha_{\lambda_n},\beta_{\lambda_n})$. We can write $T(\lambda_n)(I-(1-\lambda_n)P(\alpha_{\lambda_n},\beta_{\lambda_n})) = \lambda_n I$. Since $\lim_{n\to\infty}\gamma_{\lambda_n}(\alpha_{\lambda_n},\beta_{\lambda_n})$ exists, it follows that $T = \lim_{n\to\infty} T(\lambda_n)$ exists.

Then we see that $T(I-P(\alpha,\beta)) = 0$. Hence each row of T is an invariant distribution with respect to $P(\alpha,\beta)$, as in (v) of the previous lemma. But then for suitable ξ_{zl}:

$$\gamma(z) = (Tr(\alpha,\beta))(z) = \sum_{l=1}^{L} \xi_{zl}\widetilde{q}_l(\alpha,\beta)r(\alpha,\beta)$$
$$= \sum_{l=1}^{L} \xi_{zl}\widetilde{\gamma}_l(\alpha,\beta).$$

∎

As a consequence of this limit theorem we can formulate the following theorem, which can first be found in Schweitzer [11].

Theorem 2 *Let $\lambda_n \to 0, (\alpha_{\lambda_n},\beta_{\lambda_n}) \to (\alpha,\beta)$ and $\gamma_{\lambda_n}(\alpha_{\lambda_n},\beta_{\lambda_n}) \to \gamma$ as $n\to\infty$. Suppose that the ergodic classes corresponding to $P(\alpha,\beta)$ are the same as those of $P(\alpha_{\lambda_n},\beta_{\lambda_n})$ for all n. Then $\lim_{n\to\infty}\gamma_{\lambda_n}(\alpha_{\lambda_n},\beta_{\lambda_n}) = \gamma(\alpha,\beta)$.*

Proof. The proof needs some work and will not be given in detail here. By realizing that

$$T(\beta) = \lambda\sum\nolimits_{n=0}^{\infty}(1-\lambda)^n P^n(\alpha_{\lambda_n},\beta_{\lambda_n})$$

and that

$$(P^n(\alpha_{\lambda_n},\beta_{\lambda_n}))(z,z') = 0$$

whenever $z \in$ ergodic class l and $z' \notin$ ergodic class l, it follows that $\xi_{z\widetilde{l}} = 0$ for $z \in$ ergodic class l and $\widetilde{l} \neq l$. Hence $\xi_{zl} = 1$ and the theorem is correct for the recurrent states. That the theorem is correct for the transient states as well follows from continuity arguments. ∎

3. Unichain Stochastic Games

Unichain stochastic games are defined as games where for each pair of stationary strategies there is exactly one ergodic class. Notice that transient states are allowed. A special case of the class of unichain stochastic games is the irreducible games where the one ergodic class consists of the whole state space.

For unichain stochastic games the main existence theorems can be derived straightforwardly by considering the limit process of discounted games when λ_n goes to 0.

The next lemma is an easy consequence of the limit theorem of the previous section.

Lemma 4 *Consider a unichain stochastic game. Let*

$$\lambda_n \to 0 \text{ and } (\alpha_{\lambda_n}, \beta_{\lambda_n}) \to (\alpha, \beta) \text{ when } n \to \infty.$$

Then $\gamma_{\lambda_n}(\alpha_{\lambda_n},\beta_{\lambda_n}) \to \gamma(\alpha, \beta)$.

The proof of the next theorem is now immediate.

Theorem 3 *Consider a unichain stochastic game. Let*

$$\lambda_n \to 0 \text{ and } (\alpha_{\lambda_n}, \beta_{\lambda_n}) \to (\alpha, \beta) \text{ when } n \to \infty,$$

where α_{λ_n} *and* β_{λ_n} *are optimal for the players for the* λ_n*-discounted reward stochastic game. Then* α *and* β *are optimal for the average reward stochastic game.*

Proof. Let α_{λ_n} and β_{λ_n} be λ_n-discounted optimal. Then $\gamma_{\lambda_n}(\widetilde{\alpha}, \beta_{\lambda_n}) \leq \gamma_{\lambda_n} = \gamma_{\lambda_n}(\alpha_{\lambda_n},\beta_{\lambda_n}) \leq v(\alpha_{\lambda_n}, \widetilde{\beta})$ for all $\widetilde{\alpha}, \widetilde{\beta}$. Taking limits yields $\gamma(\widetilde{\alpha},\beta) \leq v = \gamma(\alpha, \beta) \leq \gamma(\alpha, \widetilde{\beta})$ for all $\widetilde{\alpha}, \widetilde{\beta}$. Which shows that the value of the average reward game equals $\gamma(\alpha, \beta)$ and that α and β are optimal. In fact, in this last conclusion we used the fact that a best response of a player to a stationary strategy of the other player can be found among his stationary strategies (Hordijk et al. [8]). ∎

The next non-zero-sum version of the previous theorem can be proved along the same lines.

Theorem 4 *Consider a unichain stochastic game. Let*

$$\lambda_n \to 0 \text{ and } (\alpha_{\lambda_n}, \beta_{\lambda_n}) \to (\alpha, \beta) \text{ when } n \to \infty,$$

where $(\alpha_{\lambda_n},\beta_{\lambda_n})$ *is an equilibrium point with respect to the* λ_n*-discounted reward stochastic game. Then* (α, β) *is an equilibrium point with respect to the average reward criterion.*

Obviously, since $Q(\alpha, \beta)$ has identical rows, it follows that $\gamma(\alpha, \beta)$ has identical components. So any starting state gives the same future prospects. This insight can be used in the analysis of the weighted reward stochastic game. We can state the following theorem.

Theorem 5 *Consider a weighted reward unichain stochastic game. Let* γ_λ *be the discounted reward value and* γ *the average reward value. Then the value for the weighted reward game equals* $\delta\gamma_\lambda + (1 - \delta)\gamma$.

Proof. Let α_λ be discounted optimal and α be average optimal. Consider the strategy $\sigma(N)$ which is defined as playing α_λ at the first N stages and playing α thereafter. It can easily be checked that, for any $\varepsilon > 0$, $\sigma(N)$ guarantees $\delta\gamma_\lambda + (1-\delta)\gamma$ up to ε when N is large enough. ∎

In general, optimal strategies will not exist for weighted reward games. The previous theorem can in an obvious way be extended to non-zero-sum weighted reward unichain stochastic games, by combining a λ-discounted equilibrium pair $(\alpha_\lambda,\beta_\lambda)$ with an average equilibrium pair (α, β).

3.1. EASY INITIAL STATES

For average reward stochastic games generally optimal or nearly optimal stationary strategies will not exist. This raises the question whether there are starting states such that a player can guarantee the value for these starting states, using stationary strategies. This question can be answered positively, again using the limit theory. A state is called (ε-)easy for a player if the player can guarantee the value for this game (up to ε) using stationary strategies.

Theorem 6 *Let $S_{\max}$ be the subset of states for which γ, the average reward value, is maximal and let $S_{\min}$ be the subset of state for which γ is minimal.*

(i) The states $S_{\max}$ are ε-easy for player 2 and some of the states of $S_{\max}$ are easy for player 1.

(ii) The states $S_{\min}$ are ε-easy for player 1 and some of the states of $S_{\min}$ are easy for player 2.

Proof. We show (i). Let α_{λ_n} be λ_n-discounted optimal and let $\alpha = \lim_{n\to\infty} \alpha_{\lambda_n}$ while $\lim_{n\to\infty} \lambda_n = 0$. Let $\overline{\beta}$ be an average reward best response to α. Then $\gamma_{\lambda_n}(\alpha_{\lambda_n}, \overline{\beta}) \geq \gamma_{\lambda_n}$ and by the limit theorem we derive

$$\sum_{l=1}^{L} \xi_{zl} \widetilde{\gamma}_l(\alpha, \overline{\beta}) \geq \gamma(z).$$

Since $\widetilde{\gamma}_l(\alpha, \overline{\beta}) \leq \max_z \gamma(z)$ this means that there exists an ergodic class l with $\widetilde{\gamma}_l(\alpha, \overline{\beta}) \geq \max_z \gamma(z)$. So α is optimal for this ergodic class. This shows half of the statement.

Let β_λ be λ-discounted optimal. For all α we derive from

$$\lambda(I - (1-\lambda)P(\alpha, \beta_\lambda))^{-1} r(\alpha, \beta_\lambda) = \gamma_\lambda(\alpha, \beta_\lambda)$$

that

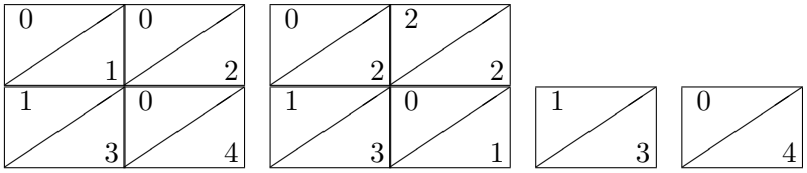
$$\lambda r(\alpha, \beta_\lambda) = (I - (1-\lambda)P(\alpha, \beta_\lambda))\gamma_\lambda(\alpha, \beta_\lambda)$$

which yields after multiplication by $Q(\alpha\beta_\lambda)$ that

$$\gamma(\alpha,\beta_\lambda) = Q(\alpha,\beta_\lambda)r(\alpha,\beta_\lambda) = Q(\alpha,\beta_\lambda)\gamma_\lambda(\alpha,\beta_\lambda) \leq Q(\alpha,\beta_\lambda)\gamma_\lambda.$$

Since $\| \gamma_\lambda - \gamma \| \leq \varepsilon$, for λ close enough to 0 and since the row sums of $Q(\alpha,\beta_\lambda)$ equal 1, we find $\gamma(z,\alpha,\beta_\lambda) \leq \max_z \gamma(z) - \varepsilon$, for λ close enough to 0. Hence β_λ is ε-optimal for the states $z \in S_{\max}$. ∎

The following example, due to Thuijsman [12], shows that not all states in $S_{\max}$ need to be easy.



It can be verified that $\gamma = (1,1,1,0)$, so $S_{\max} = \{1,2,3\}$ and that

$$\min_\beta \gamma_1(\alpha,\beta) = \min_\beta \gamma_2(\alpha,\beta) = 0$$

for all α. Thus only state 3 is (ε-)easy for player 1.

We close this section with the observation that if $S_{\max} = S_{\min} = S$ (so the average value is independent of the initial state), then, for both players, λ-discounted optimal stationary strategies are average ε-optimal as well.

For obvious reasons stochastic games for which the players possess (ε-) optimal stationary strategies are favorable from a practical viewpoint. It would be advantageous to determine the (ε-) easy states of a game. However, we know of no algorithm that specifies the (ε-)easy states. Finally, without proof, we give a set of functional equations; the existence of a solution to it is equivalent to the existence of optimal stationary strategies for both players.

Theorem 7 *The following two assertions are equivalent.*

(i) Both players possess optimal stationary strategies.
(ii) The following set of functional equations in the variables $v, w^1, w^2 \in IR^t$ has a solution:

For all $z \in S$:

$$\begin{aligned} val_{A^1(z)\times A^2(z)}\left[\sum_{z'=1}^{t} p(z'|z,a,b)v(z')\right] &= v(z) \\ val_{E^1(z)\times A^2(z)}\left[r(z,a,b)+\sum_{z'=1}^{t} p(z'|z,a,b)w^1(z')\right] &= v(z)+w^1(z) \\ val_{A^1(z)\times E^2(z)}\left[r(z,a,b)+\sum_{z'=1}^{t} p(z'|z,a,b)w^2(z')\right] &= v(z)+w^2(z). \end{aligned}$$

(Here $val_{C\times D}[\cdots]$ denotes the matrix game value over the pure action sets C and D. Further, $E^k(z)$, $k=1,2$, consists of the extreme points of the polytope of optimal actions for player k for the first equation.)

One should notice that for any solution (v, w^1, w^2) to the above set of functional equations we have $v = \gamma$ while an optimal stationary strategy for player 1 (player 2) can be derived by taking optimal actions in the second (third) equations.

3.2. PUISSEUX SERIES AND OPTIMAL STATIONARY STRATEGIES

In a nice series of papers Bewley and Kohlberg [1], [2] showed how, for any stochastic game, the λ-discounted value can be expressed as a function of λ. They showed that there exists an open interval $(0, \widetilde{\lambda})$ such that $\gamma_\lambda = \sum_{k=0}^{\infty} c_k \lambda^{k/M}$ for suitable $M \in IN$ and $c_k \in IR^t, k = 0, 1, 2, \ldots$ for all $\lambda \in (0, \widetilde{\lambda})$. Such a series is called a Puisseux series.

Since $\lim_{\lambda\to 0} \gamma_\lambda = \gamma$, the average reward value, it follows that $c_0 = \gamma$.

Bewley and Kohlberg showed that $v = \sum_{k=0}^{\infty} c_k \lambda^{k/M}$ is the solution of the set of equations

$$v(z) = val_{A^1(z)xA^2(z)}\left[\lambda r(z,.,.) + (1-\lambda)\sum_{z'=1}^{t} p(z'|z,.,.)v(z')\right], \forall z \in S,$$

for all λ close enough to 0.

Now observe that for Markov decision problems, there is only one player that can influence the outcome of the game, so for Markov decision problems the *val*-operator has to be replaced by either the min- or the max-operator, depending on whether we have a minimizing or a maximizing problem. But then the minimum (or maximum) in the above equation is found for a pure action. Since there are only finitely many different actions, the same action

can be taken for all $\lambda \in (0, \widetilde{\lambda})$. Hence the above set of equations reduces to

$$v(z) = \lambda r(z, i^*) + (1-\lambda) \sum_{z'=1}^{t} p(z'|z, i^*) v(z') \ \ , \ \ \forall z \in S$$

which, by the linearity in λ has a power series as general solution. We deduce that for Markov decision problems the above Puisseux series can be reduced to a power series of the type $\sum_{k=0}^{\infty} c_k \lambda^k$.

This result can be used in proving the next theorem.

Theorem 8 *If for a stochastic game both players possess optimal stationary strategies, then* $c_1 = c_2 = \cdots = c_{M-1} = 0$.

We will not prove this theorem rigorously, but indicate how the above-mentioned result for MDP's can be used. Let α be average reward optimal for player 1 and consider the minimizing MDP that results for player 2 when he gets to know α in advance of the play. For $MDP(\alpha)$ we know by the above result that $\gamma_\lambda(\alpha) = \gamma + O(\lambda)$, hence $\gamma_\lambda \geq \gamma + O(\lambda)$. Likewise, for a stationary strategy β that is average reward optimal for player 2, we find that $\gamma_\lambda \leq \gamma + O(\lambda)$. Hence $\gamma_\lambda = \gamma + O(\lambda)$.

A stationary strategy is called uniform discount optimal if it is discount optimal for all λ close enough to 0. Using the limit theorem it follows that a uniform discount optimal strategy is average optimal as well. The following theorem characterizes uniform discount optimal stationary strategies.

Theorem 9 *A stationary strategy* $\alpha = (\alpha(1), \alpha(2), \ldots, \alpha(t))$ *is uniform discount optimal for player 1 if and only if, for each* $z \in S$, $\alpha(z)$ *is an optimal action in the matrix game*

$$\left[\lambda r(z, ., .) + (1-\lambda) \sum_{z'=1}^{t} p(z'|z, ., .) (\sum_{k=0}^{\infty} c_k \lambda^{k/M})(z') \right]$$

for all λ *close to 0.*

Proof. The proof follows straightforwardly from discounted stochastic game considerations. ∎

3.3. TOTAL REWARD GAMES AND PUISSEUX SERIES

The total reward criterion makes sense only when the average reward value equals 0 for each initial state. But this condition is not enough. If a player, say player 1, possesses no average optimal strategy, then obviously player 1 cannot guarantee himself anything more than $-\infty$ in the total reward game. Hence we make the additional assumption that both players possess average reward optimal stationary strategies.

The next theorem can be found in Filar and Vrieze [5].

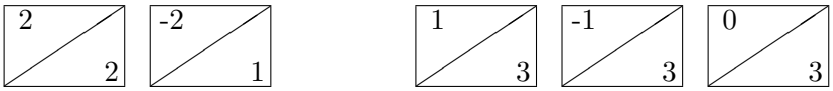
Theorem 10 *If for a stochastic game the average reward value equals 0 and if both players possess average reward optimal stationary strategies, then the total reward value exists and equals* c_{M-1}.

In the previous section we proved that $c_1 = c_2 = \cdots = c_M = 0$ whenever both players possess optimal stationary strategies. A similar statement holds for the total reward games.

Theorem 11 *When for a total reward stochastic game, for which the average reward is* 0 *and both players possess average reward optimal stationary strategies, both players have total reward optimal stationary strategies, then* $c_{M+1} = c_{M+2} = \cdots = c_{2M-1} = 0$.

The proof of this theorem is quite similar to the analogous theorem for the average reward case.

In fact, the total reward criterion should be considered as a refinement in addition to the average reward criterion. This can best be seen from the following two examples:



state 1 state 2 state 1 state 2 state 3

For both games the average reward value equals 0 while the total reward value equals (1,-1), respectively (1,-1,0).

The refinement procedure on top of the average reward criterion we gave here could be repeated in the next levels. For instance, the next extension leads to a criterion which has as value c_{2M}, etc.

References

1. Bewley, T. and Kohlberg, E. (1976) The asymptotic theory of stochastic games, *Mathematics of Operations Research* **1**, 197–208.
2. Bewley, T. and Kohlberg, E. (1978) On stochastic games with stationary optimal strategies, *Mathematics of Operations Research* **3**, 104–123.
3. Blackwell, D. and Ferguson, T. (1968) The big match, *Annals of Mathematical Statistics* **39**, 159–163.
4. Filar, J.A., Schultz, T.A., Thuijsman, F. and Vrieze, O.J. (1991) Nonlinear programming and stationary equilibria of stochastic games, *Mathematical Programming* **5**, 227–237.
5. Filar, J.A. and Vrieze, O.J. (1997) *Competitive Markov Decision Processes*, Springer-Verlag, Berlin.
6. Gillette, D. (1957) Stochastic games with zero stop probabilities, in A.W. Tucker, M. Dresher and P. Wolfe (eds.), *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, Princeton University Press, Princeton, NJ, pp. 179–187.
7. Hoffman, A.J. and Karp, R.M. (1966) On non-terminating stochastic games, *Management Science* **12**, 359–370.

8. Hordijk, A., Vrieze, O.J. and Wanrooij, G.L. (1983) Semi-Markov strategies in stochastic games, *International Journal of Game Theory* **12**, 81–89.
9. Parthasarathy, T. and Raghavan, T.E.S. (1981) An orderfield property for stochastic games when one player controls transition probabilities, *JOTA* **33**, 375–392.
10. Parthasarathy, T., Tijs, S.H. and Vrieze, O.J. (1984) Stochastic games with state independent transitions and separable rewards, in G. Hammer and D. Pallaschke (eds.), *Selected Topics in OR and Mathematical Economics*, Lecture Notes Series 226, Springer-Verlag, Berlin, pp. 262–271.
11. Schweitzer, P. (1968) Perturbation theory and finite Markov chains, *Journal of Applied Probability* **5**, 401–413.
12. Thuijsman, F. (1992) Optimality and equilibria in stochastic games, CWI-Tract 82, CWI, Amsterdam.
13. Vrieze, O.J. (1987) Stochastic games with finite state and actions spaces, CWI-Tract 33, CWI, Amsterdam.
14. Vrieze, O.J., Tijs, S.H., Raghavan, T.E.S. and Filar, J.A. (1983) A finite algorithm for the switching controller stochastic game, *OR Spektrum* **5**, 15–24.