Journal of Theoretical Probability, Vol. 16, No. 4, October 2003 (© 2003)

Continuity of the Value of Competitive Markov Decision Processes

Eilon Solan[1]

Received November 7, 2001; revised June 19, 2003

We provide a bound for the variation of the function that assigns to every competitive Markov decision process and every discount factor its discounted value. This bound implies that the undiscounted value of a competitive Markov decision process is continuous in the relative interior of the space of transition rules.

KEY WORDS: Competitive Markov decision process; stochastic games; value; sensitivity analysis.

1. INTRODUCTION

A Markov Decision Process (MDP) is given by (i) a finite set of states S and an initial state $s_1 \in S$, (ii) a finite set of actions A, (iii) a cost function $c: S \times A \to \mathbf{R}$, and (iv) a transition rule $p: S \times A \to \Delta(S)$, where $\Delta(S)$ is the space of probability distributions over S.

At every stage $n \in \mathbf{N}$, where $\mathbf{N}$ is the set of positive integers, the process is in some state $s_n \in S$. The decision maker chooses an action $a_n \in A$, and a new state $s_{n+1} \in S$ is chosen according to $p(\cdot \mid s_n, a_n)$. It is assumed that the decision maker remembers the sequence of states the process visited and his past actions.

Denote by $H = \bigcup_{n \in \mathbf{N}} (S \times A)^{n-1} \times S$ the set of all *finite histories*, where by convention, $B^0 = \varnothing$ for every finite set B and we identify $\varnothing \times S$ with S. A *plan* of the decision maker is a function σ which assigns to every finite

[1] Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208-2001, and School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: eilons@post.tau.ac.il, e-solan@kellogg.northwestern.edu

0894-9840/03/1000-0831/0 © 2003 Plenum Publishing Corporation

history $h \in H$ a probability distribution over A. Every plan σ, together with the initial state s_1 and the transition rule p, induces a probability distribution $\mathbf{P}_{s_1, p, \sigma}$ over the space of infinite histories $(S \times A)^{\mathbf{N}}$. The corresponding expectation operator is $\mathbf{E}_{s_1, p, \sigma}$.

For every discount factor $\lambda \in (0, 1]$, the discounted cost of a plan σ (at the initial state s_1, given the transition rule p and the cost function c) is

$$\gamma_\lambda(s_1, p, c; \sigma) = \mathbf{E}_{s_1, p, \sigma}\left[\lambda \sum_{n \in \mathbf{N}} (1-\lambda)^{n-1} c(s_n, a_n)\right],$$

whereas the undiscounted cost is

$$\gamma_0(s_1, p, c; \sigma) = \mathbf{E}_{s_1, p, \sigma}\left[\limsup_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} c(s_n, a_n)\right].$$

The *λ-discounted value* (at the initial state s_1, given the transition rule p and the cost function c) is

$$V_\lambda(s_1, p, c) = \inf_\sigma \gamma_\lambda(s_1, p, c; \sigma), \tag{1}$$

and the *undiscounted value* is

$$V_0(s_1, p, c) = \inf_\sigma \gamma_0(s_1, p, c; \sigma). \tag{2}$$

Every plan σ that achieves the infimum in (1) (resp. (2)) is *λ-discounted optimal* (resp. *optimal*).

A plan σ is *deterministic* if for every finite history $h \in H$, $\sigma(h)$ gives unit weight to some action in A. It is *stationary* if $\sigma(h)$ depends only on the last state of h, for every $h \in H$. A stationary plan can be identified with a vector $x = (x_s^a)_{s \in S}^{a \in A} \in (\Delta(A))^S$, with the understanding that x_s^a is the probability by which the action a is chosen whenever the process visits the state s. It is well known (see, e.g., Ref. 2, Theorems 2.1 and 4.3) that there are optimal and λ-discounted optimal plans which are deterministic and stationary. Moreover, $V_0(s_1, p, c) = \lim_{\lambda \to 0} V_\lambda(s_1, p, c)$ for every initial state s_1, every transition rule p and every cost function c (see, e.g., Ref. 2, Corollary 4.1).

The assumption that the set of available actions is independent of the state simplifies the notations, and is without loss of generality when one is interested in the value function. Indeed, if this is not the case, one can add an absorbing state in which all actions yield an extremely high cost, and, in every state which has "too few" actions, add actions that yield extremely

high cost, and lead deterministically to the new absorbing state. This construction does not affect the discounted or the undiscounted value of the MDP.

Let $\mathscr{P} = (\Delta(S))^{S \times A}$ denote the space of all transition rules. Define an equivalence relation over $\mathscr{P}$ as follows. For every $p, p' \in \mathscr{P}$, $p \sim p'$ if and only if for every $(t, s, a) \in S^2 \times A$, $p(t \mid s, a) > 0 \Leftrightarrow p'(t \mid s, a) > 0$. Let $\mathscr{D}$ be the partition of $\mathscr{P}$ into equivalence classes. $\mathscr{D}$ contains $(2^{|S|} - 1)^{|S| \times |A|}$ sets.[2]

For every stationary plan x and every initial state s_1, the function $(\lambda, p, c) \mapsto \gamma_\lambda(s_1, p, c; x)$ is continuous over $(0, 1] \times \mathscr{P} \times \mathbf{R}^{S \times A}$. Indeed, for every state $s \in S$ and every $n \in \mathbf{N}$, the function $p \mapsto \mathbf{P}_{s_1, p, x}(s_n = s)$ is continuous over $\mathscr{P}$, and moreover

$$\gamma_\lambda(s_1, p, c; x) = \sum_{n=1}^{\infty} \left(\lambda(1-\lambda)^{n-1} \sum_{(s, a) \in S \times A} \mathbf{P}_{s_1, p, x}(s_n = s) \times x_s^a \times c(s, a) \right)$$
$$\leqslant \sum_{n=1}^{\infty} \lambda(1-\lambda)^{n-1} \|c\|_\infty = \|c\|_\infty.$$

Since S and A are finite, $\|c\|_\infty = \max_{(s, a) \in S \times A} |c(s, a)| < +\infty$, so that the Weierstrass M-test yields uniform convergence implying that the function $(\lambda, p, c) \mapsto \gamma_\lambda(s_1, p, c; x)$ is continuous.

Since there is a λ-discounted optimal deterministic stationary plan, and since the number of deterministic stationary plans is finite, it follows that the function $(\lambda, p, c) \mapsto V_\lambda(s_1, p, c)$, as the minimum of finitely many continuous functions, is continuous over $(0, 1] \times \mathscr{P} \times \mathbf{R}^{S \times A}$.

Now we return to the undiscounted case. For every cost function c, every stationary plan x, and every initial state s_1, the function $p \mapsto \gamma_0(s_1, p, c; x)$ is continuous over every $P \in \mathscr{D}$. Indeed, x induces a Markov chain over S with transition rule q that is defined by $q(t \mid s) = \sum_{a \in A} x_s^a p(t \mid s, a)$. The ergodic structure of the induced Markov chain is constant over every $P \in \mathscr{D}$, and therefore, by Schweitzer,[11] the stationary distribution $\mu_{s_1, p}$ determined by the initial state s_1 is a continuous function of p over every $P \in \mathscr{D}$. Finally,

$$\gamma_0(s_1, p, c; x) = \sum_{(s, a) \in S \times A} \mu_{s_1, p}[s] \times x_s^a \times c(s, a).$$

Since there is an optimal deterministic stationary plan, it follows that the function $p \mapsto V_0(s_1, p, c)$ is continuous over every $P \in \mathscr{D}$.

[2] Indeed, for every $(s, a) \in S \times A$ and every $p \in \mathscr{P}$ denote $\operatorname{supp}(p(\cdot \mid s, a)) = \{t \in S \mid p(t \mid s, a) > 0\}$. Then $p \sim p'$ if and only if $\operatorname{supp}(p(\cdot \mid s, a)) = \operatorname{supp}(p'(\cdot \mid s, a))$ for every $(s, a) \in S \times A$. Since $\operatorname{supp}(p(\cdot \mid s, a))$ is a non-empty subset of S for every $(s, a) \in S \times A$, and since there are $2^{|S|} - 1$ such subsets, the number of equivalence classes is $(2^{|S|} - 1)^{|S| \times |A|}$.

As the following example shows, the function $p \mapsto V_0(s_1, p, c)$ is *not* continuous over all of $\mathscr{P}$. Take $S = \{s, s'\}$, $A = \{a\}$, $c(s, a) = 1$ and $c(s', a) = 0$. Define for every $k \in \mathbf{N}$ a transition rule p_k by $p_k(s' \mid s, a) = 1/k$ and $p_k(s' \mid s', a) = 1$. The sequence $(p_k)_{k \in \mathbf{N}}$ converges to the transition rule p that is defined by $p(s \mid s, a) = p(s' \mid s', a) = 1$. However, $V_0(s, p_k, c) = 0$ for every $k \in \mathbf{N}$, while $V_0(s, p, c) = 1$.

The goal of the present article is to extend this line of investigation to *competitive Markov decision processes*, or *stochastic games*.

Definition 1. A *competitive Markov decision process* (or a *stochastic game*) is a tuple $\Gamma = (S, s_1, A, B, p, c)$, where (i) S is a finite set of states, and $s_1 \in S$ is the initial state, (ii) A and B are two finite sets of actions for the two decision makers DM1 and DM2 respectively, (iii) $p: S \times A \times B \to \Delta(S)$ is a transition rule, and (iv) $c: S \times A \times B \to \mathbf{R}$ is a cost function.

The process proceeds as follows. At every stage $n \in \mathbf{N}$, knowing the past history $(s_1, a_1, b_1, s_2, a_2, b_2, \ldots, s_n)$, the two decision makers choose, independently and simultaneously, actions $a_n \in A$ and $b_n \in B$ respectively. DM1 pays DM2 the amount $c(s_n, a_n, b_n)$, and a new state s_{n+1} is chosen according to the probability distribution $p(\cdot \mid s_n, a_n, b_n)$.

Competitive MDPs are useful to model situations when two (or more) strategic decision makers control the evolution of a system. They have been applied in various contexts, ranging from arms races[14] to optimal inspection models[5] and resource extraction models.[1, 8]

A *plan* of DM1 is a function σ from the set of all finite histories $H = \bigcup_{n \in \mathbf{N}} (S \times A \times B)^{n-1} \times S$ to $\Delta(A)$. Plans of DM2 are functions $\tau: H \to \Delta(B)$. A plan σ (resp. τ) is *stationary* if $\sigma(h)$ (resp. $\tau(h)$) depends only on the last state of h, for every $h \in H$. Stationary plans of the two decision makers are denoted by $x \in (\Delta(A))^S$ and $y \in (\Delta(B))^S$ respectively.

Every pair of plans (σ, τ), together with an initial state $s_1 \in S$ and a transition rule p, induces a probability measure $\mathbf{P}_{s_1, p, \sigma, \tau}$ over the space $(S \times A \times B)^{\mathbf{N}}$ of infinite histories. The corresponding expectation operator is $\mathbf{E}_{s_1, p, \sigma, \tau}$.

For every pair of plans (σ, τ), and every discount factor $\lambda \in (0, 1]$, we define the *λ-discounted cost* (at the initial state s_1, given the transition rule p and the cost function c) by

$$\gamma_\lambda(s_1, p, c; \sigma, \tau) = \mathbf{E}_{s_1, p, \sigma, \tau}\left[\lambda \sum_{n \in \mathbf{N}} (1-\lambda)^{n-1} c(s_n, a_n, b_n)\right], \tag{3}$$

and the *undiscounted cost* by

$$\gamma_0(s_1, p, c; \sigma, \tau) = \mathbf{E}_{s_1, p, \sigma, \tau}\left[\limsup_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} c(s_n, a_n, b_n)\right]. \tag{4}$$

If for a given initial state s_1, transition rule p and cost function c the equality

$$\inf_{\sigma} \sup_{\tau} \gamma_\lambda(s_1, p, c; \sigma, \tau) = \sup_{\tau} \inf_{\sigma} \gamma_\lambda(s_1, p, c; \sigma, \tau) \tag{5}$$

holds, then the common value is the *λ-discounted value* and it is denoted by $v_\lambda(s_1, p, c)$. Similarly, if for a given initial state s_1, transition rule p and cost function c the equality

$$\inf_{\sigma} \sup_{\tau} \gamma_0(s_1, p, c; \sigma, \tau) = \sup_{\tau} \inf_{\sigma} \gamma_0(s_1, p, c; \sigma, \tau) \tag{6}$$

holds, then the common value is the *undiscounted value* and it is denoted by $v_0(s_1, p, c)$.

The assumption that the action sets are independent of the state is without loss of generality when one studies the value of competitive MDPs as well; one first adds an absorbing state that yields extremely high cost, and adds actions to DM1 that lead deterministically to this absorbing state, and then one adds another absorbing state that yields extremely low cost, and adds actions to DM2 that lead deterministically to the second absorbing state.

Shapley[12] proved that for every discount factor $\lambda \in (0, 1]$, the λ-discounted value exists, and that moreover both decision makers have optimal stationary plans; that is, a stationary plan σ (resp. τ) that achieves the infimum in the left-hand side (resp. the supremum in the right-hand side) in (5) for every initial state s_1.

The first step towards studying continuity properties of the discounted value was done by Bewley and Kohlberg.[3] Their analysis implies that for every initial state s_1 the function $(\lambda, p, c) \mapsto v_\lambda(s_1, p, c)$ is semi-algebraic over $(0, 1] \times \mathscr{P}^* \times \mathbf{R}^{S \times A \times B}$, where $\mathscr{P}^* = (\Delta(S))^{S \times A \times B}$ is the space of transition rules. In particular, this function is piecewise continuous.

Filar and Vrieze (Ref. 6, Eq. (4.19)) proved that for every initial state s_1, every pair of transition rules (p, p'), every pair of cost functions (c, c'), and every pair of discount factors $\lambda, \lambda' \in (0, 1]$,[3]

$$\begin{aligned} &|v_\lambda(s_1, p, c) - v_{\lambda'}(s_1, p', c')| \\ &\quad \leqslant \|c - c'\|_\infty + \frac{1-\lambda}{\lambda} \|p - p'\|_1 \times \|c'\|_\infty + 2 \frac{|\lambda - \lambda'|}{\lambda} \|c'\|_\infty. \end{aligned} \tag{7}$$

In particular, for every initial state s_1, the function $(\lambda, p, c) \mapsto v_\lambda(s_1, p, c)$ is continuous over $(0, 1] \times \mathscr{P}^* \times \mathbf{R}^{S \times A \times B}$.

[3] Filar and Vrieze use the quantity $\beta = 1 - \lambda$ rather than λ. Also, the metric they use to measure the distance between transition rules is slightly different than the L_1-norm.

Milman[10] further studied continuity properties of the value function. He proved that $|\frac{\partial v_\lambda(s_1, p, c)}{\partial \lambda}| \leqslant \frac{\|c\|_\infty f(p)}{\lambda^{1-1/M}}$, where M is some positive integer that depends on $|S|$, $|A|$, and $|B|$, and f is some positive (but not necessarily bounded) function.

The carryover of such continuity results to the undiscounted case would seem promising, since Mertens and Neyman[9] proved that for every initial state s_1, every transition rule p, and every cost function c, the undiscounted value exists, and that moreover

$$v_0(s_1, p, c) = \lim_{\lambda \to 0} v_\lambda(s_1, p, c). \tag{8}$$

But since in (7) the discount factor appears in the denominator, (7) and (8) together do not imply that for a given initial state s_1 and a given cost function c the function $p \mapsto v_0(s_1, p, c)$ is continuous over every $P \in \mathscr{D}^*$, where $\mathscr{D}^*$ is the partition of the space of transition rules which is analogous to $\mathscr{D}$.

One application of (7) is in estimating the discounted value of a competitive MDP, for which the data are not precisely known, due to, e.g., rounding errors, since the model under consideration is a simplification of a more complex model, or since the data are estimated by various statistical methods. Equation (7) relates, in such a case, the precision of the data to the precision of the value. Unfortunately, for any fixed level of desired precision in the discounted value, the required precision in λ and p according to (7) increases as the discount factor goes to 0. Furthermore, Eq. (7) cannot be used to estimate the undiscounted value when the data are not precisely known.

Define a function $d: \mathscr{P}^* \times \mathscr{P}^* \to [0, +\infty]$ as follows. For every $p, p' \in \mathscr{P}^*$

$$d(p, p') = \max \left\{ \frac{p(t \mid s, a, b)}{p'(t \mid s, a, b)}, \frac{p'(t \mid s, a, b)}{p(t \mid s, a, b)} \mid s, t \in S, a \in A, b \in B \right\} - 1, \tag{9}$$

where by convention $x/0 = +\infty$ for $x > 0$, and $0/0 = 1$. A few simple properties of the non-negative valued function $d(\cdot, \cdot)$ are:

(A.i) $d(p, p') = 0$ if and only if $p = p'$,

(A.ii) $d(p, p') = d(p', p)$,

(A.iii) $d(p, p') < +\infty$ if and only if p and p' lie in the same element of $\mathscr{D}^*$, and

(A.iv) $d(p_n, p) \to 0$, whenever $p, (p_n)_{n \in \mathbf{N}}$ lie in the same element of $\mathscr{D}^*$ and $p_n \to p$ in the Euclidean norm.

As the following example shows, $d(\cdot, \cdot)$ is *not* a metric.

Example 1. Let $S=\{s, s'\}$, and take $|A|=|B|=1$, so that the competitive MDP is reduced to a Markov chain. Fix $\epsilon \in (0, 1/7)$. For $k=1, 2, 5$, define a transition rule $p_k: S \to \Delta(S)$ by: $p_k(s \mid t)=1-p_k(s' \mid t)=k\epsilon$ for each $t \in S$. Since $\epsilon \in (0, 1/7)$ one has $d(p_1, p_2)=1$, $d(p_2, p_5)=3/2$, and $d(p_1, p_5)=4$, so that $d(p_1, p_2)+d(p_2, p_5)<d(p_1, p_5)$, and $d(\cdot, \cdot)$ is not a metric.

The main result to be presented below is that for every initial state s_1, every pair of transition rules (p, p'), every pair of cost functions (c, c'), and every $\lambda \in [0, 1]$ (including the case $\lambda=0$)

$$-4\,|S|\,d(p, p')\,\|c\|_\infty - \|c-c'\|_\infty \leqslant v_\lambda(s_1, p', c') - v_\lambda(s_1, p, c) \leqslant \frac{4\,|S|\,d(p, p')}{(1-2\,|S|\,d(p, p'))^+}\,\|c\|_\infty + \|c-c'\|_\infty, \quad (10)$$

where, for every $x \in \mathbf{R}$, $x^+=\max\{x, 0\}$, $+\infty+x=+\infty$, and $+\infty \times 0=1$. The proof of (10) uses a graph-theoretic approach, initiated by Friedlin and Wentzell[7] for MDPs with rare transitions, rather than the matrix analysis approach, which is the standard approach to studying Markov chains.

The estimate given in (10) is uniform over $\lambda \in [0, 1]$. We show below that this, together with (7) and (8), implies that the function $(\lambda, p, c) \mapsto v_\lambda(s_1, p, c)$ is continuous over $[0, 1] \times P \times \mathbf{R}^{S \times A \times B}$, for every $P \in \mathscr{D}^*$. As Milman[10] remarks, our analysis can be used to improve his results. Finally, observe that (10) is neither stronger nor weaker than (7).[4]

2. ON MARKOV CHAINS

In the present section we recall a result due to Freidlin and Wentzell,[7] and we apply it to competitive MDPs.

2.1. A Result of Freidlin and Wentzell

Let (S, q) be a Markov chain; that is, S is a finite set of states, and $q: S \to \Delta(S)$ is a transition function. Let $s_1 \in S$ be the initial state.

[4] Indeed, when c, c', p and p' are fixed, the bound in (10) is independent of λ, whereas, if $p \neq p'$, the bound in (7) goes to $+\infty$ as λ goes to 0. Hence (10) is not weaker than (7). On the other hand, one may find two transition rules p and p' such that $d(p, p') \geqslant 1$ (see the transition rules in Example 1). In such a case, when $c \neq 0$, the upper bound in (10) is infinite, while the bound in (7) is finite, so that (10) is not stronger than (7).

Let C be a proper subset of S. Denote by $e_C = \min\{n \in \mathbf{N} \mid s_n \notin C\}$ the first hitting time of the complement of C.[5] Recall that $S^C = \{f: C \to S\}$ is the set of all functions from C to S. Every $f \in S^C$ naturally defines a directed graph over S: the graph contains the edge $(s \to s')$ if and only if $s' = f(s)$. We set $\alpha_f = 1$ if this directed graph has *no* directed cycles, and $\alpha_f = 0$ if it has at least one directed cycle. Since f is a function, there is exactly one directed path that leaves each $s \in C$. For every $s \in C$ and every $s' \in S$ we set $\beta_f(s \to s') = 1$ if the directed path that leaves s in the directed graph induced by f reaches s', and $\beta_f(s \to s') = 0$ otherwise.

The following lemma is a special case of a result due to Freidlin and Wentzell[7] (see also Ref. 4).

Lemma 2 (Ref. 10, Lemma 6.3.3). Let (S, q) be a Markov chain, and let C be a proper subset of S such that $\mathbf{P}_{s,q}(e_C < +\infty) > 0$ for every $s \in C$. Then for every initial state $s_1 \in C$, and every $r \notin C$,

$$\mathbf{P}_{s_1,q}(s_{e_C} = r) = \frac{\sum_{f \in S^C} (\beta_f(s_1 \to r) \prod_{s \in C} q(f(s) \mid s))}{\sum_{f \in S^C} \alpha_f \prod_{s \in C} q(f(s) \mid s)}. \tag{11}$$

We now show that the hypothesis that $\mathbf{P}_{s,q}(e_C < +\infty) > 0$ for every $s \in C$ ensures that the denominator in (11) is positive. Since all terms in the summation of the denominator in (11) are non-negative, it suffices to exhibit an $f_* \in S^C$ with $\alpha_{f_*} = 1$ that satisfies $\prod_{s \in C} q(f_*(s) \mid s) > 0$.

Let $s \in C$ be arbitrary. Since $\mathbf{P}_{s,q}(e_C < +\infty) > 0$, there is $K > 1$ and a sequence $s = s_1, s_2, \ldots, s_K$ such that $s_2, \ldots, s_{K-1} \in C$, $s_K \notin C$, and $q(s_{k+1} \mid s_k) > 0$ for $k = 1, \ldots, K-1$. Denote by K_s the length of a shortest such sequence. Then either (i) $K_s = 2$, so that there exists $r \notin C$ with $q(r \mid s) > 0$, or (ii) $K_s > 2$, which implies that there exists $r \in C$ such that $q(r \mid s) > 0$ and $K_s = K_r + 1$. We set $f_*(s) = r$, where if $K_s = 2$ then r is any state in $S \setminus C$ that satisfies $q(r \mid s) > 0$, and if $K_s > 2$ then r is any state in C that satisfies $q(r \mid s) > 0$ and $K_s = K_r + 1$.

Since s is arbitrary, f_* is a function from C to S, and by construction $\prod_{s \in C} q(f_*(s) \mid s) > 0$. Since for every $s \in C$, if $f_*(s) \in C$ then $K_{f_*(s)} + 1 = K_s$, it follows that the directed graph induced by f_* has no cycles, that is, $\alpha_{f_*} = 1$.

[5] By convention, the minimum of an empty set is $+\infty$.

2.2. The Mean Discounted Time

For every discount factor $\lambda \in (0, 1]$ and every state $s \in S$, the *mean λ-discounted time* the process spends in state s, if the initial state is s_1 and the transition function is q, is given by:

$$t_\lambda(s_1, q; s) = \mathbf{E}_{s_1, q}\left[\lambda \sum_{n \in \mathbf{N}} (1-\lambda)^{n-1} \mathbf{1}_{s_n = s}\right],$$

where $\mathbf{1}_{s_n = s}$ is the indicator function.

Our basic observation is the following.

Proposition 3. For every initial state s_1, every state $s \in S$, and every discount factor $\lambda \in (0, 1]$, there exist two polynomials $h_1(q)$ and $h_2(q)$ in the $|S|^2$ variables $(q(t \mid r))_{r, t \in S}$ that satisfy (i) both polynomials have degree at most $|S|$ and non-negative coefficients, and (ii) $t_\lambda(s_1, q; s) = h_1(q)/h_2(q)$ for every transition function q over S.

Proof. Fix $\lambda \in (0, 1]$, but do not yet fix s_1 or s. Define an auxiliary Markov chain $(\hat{S}, \hat{q})$ as follows.

1. The state space is $\hat{S} = S' \cup S''$, where S' and S'' are two disjoint copies of S. For every $s \in S$, we denote by s' and s'' the corresponding states in S' and S'' respectively.
2. Every state $s'' \in S''$ is absorbing: $\hat{q}(s'' \mid s'') = 1$ for every $s'' \in S''$.
3. The transition function from every state $s' \in S'$ is as follows.

$$\begin{aligned} &\hat{q}(s'' \mid s') = \lambda, \\ &\hat{q}(t' \mid s') = (1-\lambda)\, q(t \mid s) \qquad \forall t \in S, \qquad \text{and} \\ &\hat{q}(t'' \mid s') = 0 \qquad \forall t \in S \setminus \{s\}. \end{aligned}$$

We first claim that

$$t_\lambda(s_1, q; s) = \mathbf{P}_{s'_1, \hat{q}}(s_{e_{S''}} = s'') \qquad \forall s, s_1 \in S. \tag{12}$$

Indeed, one can verify that for every $s \in S$, both $(t_\lambda(s_1, q; s))_{s_1 \in S}$ and $(\mathbf{P}_{s'_1, \hat{q}}(s_{e_{S''}} = s''))_{s_1 \in S}$ are solutions of the system of linear equations

$$x(s_1) = \lambda \mathbf{1}_{s_1 = s} + (1-\lambda) \sum_{r \in S} q(r \mid s_1)\, x(r) \qquad \forall s_1 \in S.$$

Moreover, this system of linear equations has a unique solution. Indeed, let $x = (x(r))_{r \in S}$ and $y = (y(r))_{r \in S}$ be two solutions of this system. Choose

$s_1 \in S$ such that the quantity $|x(s_1)-y(s_1)|$ is maximal, and set $\delta = |x(s_1)-y(s_1)|$. Since both x and y are solutions, and by the triangle inequality,

$$0 \leqslant \delta = |x(s_1)-y(s_1)| \leqslant (1-\lambda) \sum_{r \in S} q(r \mid s_1)\, |x(r)-y(r)|$$
$$\leqslant (1-\lambda) \sum_{r \in S} q(r \mid s_1)\, \delta = \delta(1-\lambda). \tag{13}$$

Since $\lambda > 0$, (13) implies that $\delta = 0$, so that the two solutions coincide. In particular (12) holds.

We now claim that $\mathbf{P}_{s', \hat{q}}(e_{S'} < +\infty) > 0$ for every $s' \in S'$. Indeed, since $\hat{q}(s'' \mid s') = \lambda$, and since $s'' \notin S'$,

$$\mathbf{P}_{s', \hat{q}}(e_{S'} < +\infty) \geqslant \mathbf{P}_{s', \hat{q}}(e_{S'} = 2) = \hat{q}(s'' \mid s') = \lambda > 0.$$

By the last claim one can apply Lemma 2 with $C = S'$, which implies the result. Indeed, the terms α_f and $\beta_f(s_1 \to r)$ in (11) are independent of q, and the two products in (11) each contain $|C| = |S|$ terms of the form $\hat{q}(\hat{t} \mid \hat{s})$ for $\hat{s}, \hat{t} \in \hat{S}$. The result follows by the definition of $\hat{q}$. □

Corollary 4. For every initial state $s_1 \in S$, every discount factor $\lambda \in (0, 1]$, and every collection of non-negative scalars $(\theta_s)_{s \in S}$, the function $q \mapsto \sum_{s \in S} \theta_s t_\lambda(s_1, q; s)$ is the ratio of two polynomials in the variables $(q(t \mid r))_{r, t \in S}$ of degree at most $|S|$ with non-negative coefficients.

Proof. By Proposition 3, for every $s \in S$ the function $q \mapsto t_\lambda(s_1, q; s)$ is the ratio of two polynomials in $(q(t \mid r))_{r, t \in S}$. By Lemma 2 and the proof of Proposition 3, all these ratios have the same denominator; it has non-negative coefficients and degree at most $|S|$, as do each of the numerators. Since $(\theta_s)_{s \in S}$ are non-negative scalars the result follows. □

3. COMPETITIVE MARKOV DECISION PROCESSES

Throughout this section we fix the set of states S, the initial state $s_1 \in S$, and the sets of actions A and B.

Equations (3) and (4) and the definitions of the discounted value and undiscounted value readily imply that the function $c \mapsto v_\lambda(s_1, p, c)$ is Lipschitz-1 in the cost function for every $\lambda \in [0, 1]$ (including the case $\lambda = 0$):

Lemma 5. For every initial state s_1, every transition rule p, every pair of cost functions (c, c'), and every $\lambda \in [0, 1]$,

$$|v_\lambda(s_1, p, c) - v_\lambda(s_1, p, c')| \leqslant \|c - c'\|_\infty.$$

Our main result is the following theorem. Note that the estimate provided by the theorem is uniform for every $\lambda \in [0, 1]$ (including $\lambda = 0$).

Theorem 6. Let (S, s_1, A, B, p, c) be a competitive MDP, let p': $S \times A \times B \to \Delta(S)$ be an arbitrary transition rule, and let c': $S \times A \times B \to \mathbf{R}$ be an arbitrary cost function. Then for every $\lambda \in [0, 1]$,

$$-4\,|S|\,d(p, p')\,\|c\|_\infty - \|c - c'\|_\infty \leqslant v_\lambda(s_1, p', c') - v_\lambda(s_1, p, c) \leqslant \frac{4\,|S|\,d(p, p')}{(1 - 2\,|S|\,d(p, p'))^+}\,\|c\|_\infty + \|c - c'\|_\infty. \quad (14)$$

We are going to use the following observation.

Lemma 7. Let $f(x_1, \ldots, x_k)$ be a polynomial in $x_1, \ldots, x_k$ with non-negative coefficients and degree at most n, and let $\epsilon \geqslant 0$. Let $y, y' \in \mathbf{R}^k$ be two non-negative vectors such that $1/(1+\epsilon) \leqslant y_i / y'_i \leqslant 1 + \epsilon$ for every $i = 1, \ldots, k$. Then $(1+\epsilon)^{-n} \leqslant f(y)/f(y') \leqslant (1+\epsilon)^n$.[6]

Proof. Denote $f(x) = \sum_{i=1}^{I} a_i \prod_{j=1}^{n_i} x_{k_{i,j}}$, where $I \in \mathbf{N}$, and for every $i = 1, \ldots, I$, $a_i \geqslant 0$, $0 \leqslant n_i \leqslant n$, and $1 \leqslant k_{i,j} \leqslant k$ for each $j = 1, \ldots, n_i$. By assumption, for every $i = 1, \ldots, I$,

$$\frac{1}{(1+\epsilon)^n} \prod_{j=1}^{n_i} y'_{k_{i,j}} \leqslant \prod_{j=1}^{n_i} y_{k_{i,j}} \leqslant (1+\epsilon)^n \prod_{j=1}^{n_i} y'_{k_{i,j}}. \quad (15)$$

Since $(a_i)_{i=1}^{I}$ are non-negative, multiplying (15) by a_i and summing over $i = 1, \ldots, I$ yields the desired result. □

Proof of Theorem 6. In view of Lemma 5, it is sufficient to prove that

$$-4\,|S|\,d(p, p')\,\|c\|_\infty \leqslant v_\lambda(s_1, p', c) - v_\lambda(s_1, p, c) \leqslant \frac{4\,|S|\,d(p, p')}{(1 - 2\,|S|\,d(p, p'))^+}\,\|c\|_\infty. \quad (16)$$

[6] Recall that $\frac{0}{0} = 1$, so that $y_i = 0$ if and only if $y'_i = 0$. For the same reason, the Lemma trivially holds when f is identically zero.

This inequality trivially holds when $d(p, p') \geqslant 1/2\,|S|$. Indeed, in this case the denominator in the right-hand side vanishes, while the left-hand side is at most $-2\,\|c\|_\infty$, which is a lower bound for $v_\lambda(s_1, p', c) - v_\lambda(s_1, p, c)$.

We therefore assume from now on that $d(p, p') < 1/2\,|S| < 1$. In particular, p and p' lie in the same element of the partition $\mathscr{D}^*$.

We first prove that if the cost function is positive, then for every $\lambda \in [0, 1]$,

$$(1-d(p, p'))^{2|S|} \leqslant \frac{v_\lambda(s_1, p', c)}{v_\lambda(s_1, p, c)} \leqslant \frac{1}{(1-d(p, p'))^{2|S|}}. \tag{17}$$

Note that when the cost function is positive, $v_\lambda(s_1, p, c) \geqslant \min\{c(s, a, b) \mid (s, a, b) \in S \times A \times B\} > 0$, so that the denominator in (17) is positive.

Every pair of stationary plans (x, y) naturally defines a Markov chain over S with transition rule q that is defined by

$$q(t \mid s) = \sum_{a \in A} \sum_{b \in B} x_s^a y_s^b p(t \mid s, a, b). \tag{18}$$

In particular, for every $\lambda \in (0, 1]$,

$$\gamma_\lambda(s_1, p, c; x, y) = \sum_{s \in S} \left(t_\lambda(s_1, p, x, y; s) \sum_{a, b} x_s^a y_s^b c(s, a, b) \right), \tag{19}$$

where $t_\lambda(s_1, p, x, y; s)$ is the mean discounted time spent at s in the Markov chain induced by (p, x, y). By (19) and Corollary 4, with $\theta_s = \sum_{a, b} x_s^a y_s^b c(s, a, b)$,

$$\frac{\gamma_\lambda(s_1, p', c; x, y)}{\gamma_\lambda(s_1, p, c; x, y)} = \frac{g_1(p)}{g_1(p')} \times \frac{g_2(p')}{g_2(p)}, \tag{20}$$

where $g_1(p), g_2(p)$ are polynomials in $(p(t \mid s, a, b))_{(t, s, a, b) \in S^2 \times A \times B}$ of degree at most $|S|$ with non-negative coefficients.

Set $\epsilon = d(p, p') < 1/2\,|S| < 1$. By the definition of $d(p, p')$, for every $(t, s, a, b) \in S^2 \times A \times B$ the two quantities $\frac{p(t \mid s, a, b)}{p'(t \mid s, a, b)}$ and $\frac{p'(t \mid s, a, b)}{p(t \mid s, a, b)}$ are between $\frac{1}{1+\epsilon}$ and $1+\epsilon$. It follows by (20) and Lemma 7, with $k = |S|^2 \times |A| \times |B|$, that

$$(1+d(p, p'))^{-2|S|} \leqslant \frac{\gamma_\lambda(s_1, p', c; x, y)}{\gamma_\lambda(s_1, p, c; x, y)} \leqslant (1+d(p, p'))^{2|S|}.$$

Since for every $x \in [0, 1]$ one has $1-x \leqslant 1/(1+x)$ and $1+x \leqslant 1/(1-x)$, one has

$$(1-d(p, p'))^{2|S|} \leqslant \frac{\gamma_\lambda(s_1, p', c; x, y)}{\gamma_\lambda(s_1, p, c; x, y)} \leqslant \frac{1}{(1-d(p, p'))^{2|S|}}. \tag{21}$$

For $\lambda \in (0, 1]$, (17) follows from (21) and the existence of stationary λ-discounted optimal plans. For $\lambda = 0$, (17) follows now from (8).

To finish the proof of the theorem, let $\rho < 0$ satisfy $\rho < c(s, a, b)$ for every $(s, a, b) \in S \times A \times B$. Then $c-\rho$ is a positive cost function. By (17)

$$\begin{aligned} &1-2|S|\, d(p, p') \\ &\quad \leqslant (1-d(p, p'))^{2|S|} \leqslant \frac{v_\lambda(s_1, p', c-\rho)}{v_\lambda(s_1, p, c-\rho)} \\ &\quad \leqslant \frac{1}{(1-d(p, p'))^{2|S|}} \leqslant \frac{1}{1-2|S|\, d(p, p')} = 1+\frac{2|S|\, d(p, p')}{1-2|S|\, d(p, p')}. \end{aligned}$$

In particular,

$$\begin{aligned} -2|S|\, d(p, p')\, v_\lambda(s_1, p, c-\rho) &\leqslant v_\lambda(s_1, p', c-\rho) - v_\lambda(s_1, p, c-\rho) \\ &\leqslant \frac{2|S|\, d(p, p')}{1-2|S|\, d(p, p')}\, v_\lambda(s_1, p, c-\rho). \end{aligned}$$

Equation (16) follows, since $v_\lambda(s_1, p, c-\rho) = -\rho + v_\lambda(s_1, p, c)$, since $v_\lambda(s_1, p', c-\rho) = -\rho + v_\lambda(s_1, p', c)$, since $|v_\lambda(s_1, p, c)| \leqslant \|c\|_\infty$, since ρ can be chosen arbitrarily close to $-\|c\|_\infty$, and since $d(p, p') < 1/2|S|$. □

Corollary 8. For every initial state $s_1 \in S$, the function $(\lambda, p, c) \mapsto v_\lambda(s_1, p, c)$ is continuous over $[0, 1] \times P \times \mathbf{R}^{S \times A \times B}$, for every $P \in \mathscr{D}^*$.

Proof. Let $(\lambda_n)_{n \in \mathbf{N}}$, $(p_n)_{n \in \mathbf{N}}$ and $(c_n)_{n \in \mathbf{N}}$ be converging sequences (in the Euclidean norm) of scalars in $[0, 1]$, transition rules, and cost functions respectively. Denote their limits by λ, p and c respectively. Assume that p and $(p_n)_{n \in \mathbf{N}}$ lie in the same element of $\mathscr{D}^*$. By property (A.iv) of the function $d(\cdot, \cdot)$, $\lim_{n \to \infty} d(p_n, p) = 0$.

By the triangle inequality,

$$\begin{aligned} &|v_\lambda(s_1, p, c) - v_{\lambda_n}(s_1, p_n, c_n)| \\ &\quad \leqslant |v_\lambda(s_1, p, c) - v_{\lambda_n}(s_1, p, c)| + |v_{\lambda_n}(s_1, p, c) - v_{\lambda_n}(s_1, p_n, c_n)|. \end{aligned} \tag{22}$$

When $\lambda = 0$ the first term in the right-hand side of (22) goes to zero by (8). When $\lambda > 0$ it goes to zero by (7).[7] The second term in the right-hand side goes to zero by Theorem 6. □

Remark. In the representation of $q \mapsto t_\lambda(s_1, q; s)$ as a ratio of two polynomials in $(q(t \mid r))_{r,t \in S}$, λ contributes to the coefficients of the two polynomials. By adding λ and $\beta = 1 - \lambda$ as (formally independent) variables to these polynomials (so as to obtain polynomials still having non-negative coefficients; see the proof of Proposition 3), one can obtain an estimate to the difference $v_\lambda(s_1, p, c) - v_{\lambda'}(s_1, p', c')$. The estimate is similar to (14), but one should replace all appearances of $d(p, p')$ by $d(p, p') \times d(\lambda, \lambda')$, where $d(\lambda, \lambda') = \max\{\frac{\lambda}{\lambda'}, \frac{\lambda'}{\lambda}, \frac{1-\lambda}{1-\lambda'}, \frac{1-\lambda'}{1-\lambda}\}$.

Remark. As Sylvain Sorin remarked, Theorem 6 applies also to the case where the action sets A and B are Borel spaces, as long as the following conditions hold: (a) the set of states S is finite, (b) for every discount factor the discounted value exists, and both players have discounted stationary ϵ-optimal plans, for every $\epsilon > 0$,[8] (c) the undiscounted value exists and is the limit of the discounted value as the discount factor goes to 0, and (d) the cost function is bounded. The model in this case is similar to the one described in Section 1, except that the actions a_n and b_n of the two decision makers are chosen from Borel spaces A and B respectively. So that the cost of a pair of plans is well defined, a plan of DM1 is a *measurable* function $\sigma: H \to \Delta(A)$, where $\Delta(A)$ is the space of probability distributions over A. Plans of DM2 are defined analogously.

To deal with this more general setup, the proof should be amended as follows. The max in (9) is replaced by sup, and summations in (18) and (19) are replaced by integrals. The results in Section 2 are valid as long as S is finite. The results in Section 3 are valid as long as (a)–(d) hold.

For more details on competitive MDPs with general action sets the reader is referred to Sorin (Ref. 13, Chap. 5), where conditions under which (b) holds are given. Conditions under which (c) holds are given in Mertens and Neyman.[9] In the absence of (c), the bound (14) is still valid for $\lambda \in (0, 1]$.

[7] Alternatively, one can use (3) and the fact that there are stationary discounted optimal plans instead of (7).

[8] A plan σ of player 1 is *discounted ϵ-optimal* if $\sup_\tau \gamma_\lambda(s_1, p, c; \sigma, \tau) \leqslant v_\lambda(s_1, p, c) + \epsilon$. Discounted ϵ-optimal plans of player 2 are defined analogously.

ACKNOWLEDGMENTS

I thank Ehud Lehrer for raising the question and commenting on an earlier version of the paper, Emanuel Milman and Sylvain Sorin for useful discussions, and Ron, who had an inactive yet an important role in the solution. I am indebted to two anonymous referees, who read the paper thoroughly. Their comments were most helpful, and substantially improved the presentation.

REFERENCES

1. Amir, R. (1987). *Sequential Games of Resource Extraction: Existence of Nash Equilibrium*, Cowles Foundation D.P. #825.
2. Arapostathis, A., Borkar, V. S., Fernández-Gaucherand, E., Ghosh, M. K., and Marcus, S. I. (1993). Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM J. Control Optim.* **31**, 282–344.
3. Bewley, T., and Kohlberg, E. (1976). The asymptotic theory of stochastic games. *Math. Oper. Res.* **1**, 197–208.
4. Catoni, O. (1999). *Simulated Annealing Algorithms and Markov Chains with Rare Transitions*, Séminaire de Probabilités, XXXIII, 69–119, Lecture Notes in Mathematics, 1709, Springer, Berlin.
5. Filar, J. A. (1985). Player aggregation in the traveling inspector model, *IEEE Trans. Automatic Control* **AC-30**, 723–729.
6. Filar, J. A., and Vrieze, K. (1997). *Competitive Markov Decision Processes*, Springer-Verlag, New York.
7. Freidlin, M. I., and Wentzell, A. D. (1984). *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin.
8. Levhari, D., and Mirman, L. (1980). The great fish war: An example using a dynamic Cournot–Nash solution. *Bell J. Econ.* **11**, 322–334.
9. Mertens, J. F., and Neyman, A. (1981). Stochastic games. *Int. J. Game theory* **10**, 53–66.
10. Milman, E. (2002). The semi-algebraic theory of stochastic games. *Math. Oper. Res.* **27**, 401–418.
11. Schweizer, P. J. (1968). Perturbation theory and finite Markov chains. *J. Applied Probab.* **5**, 401–413.
12. Shapley, L. S. (1953). Stochastic games. *Proc. Nat. Acad. Sci. U.S.A.* **39**, 1095–1100.
13. Sorin, S. (2002). *A First Course on Zero-Sum Repeated Games*, Mathématiques et Applications, Vol. 37, Springer-Verlag, Berlin.
14. Winston, W. (1978). A stochastic game model of a weapons development competition. *Siam J. Control Optim.* **16**, 411–419.