Operations Research Letters 44 (2016) 587–591

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

The value functions of Markov decision processes

Ehud Lehrer [a,b,*], Eilon Solan [a], Omri N. Solan [a]

[a] School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997800, Israel
[b] INSEAD Bd. de Constance, 77305 Fontainebleau Cedex, France

ARTICLE INFO

Article history:
Received 18 January 2016
Received in revised form
27 June 2016
Accepted 27 June 2016
Available online 6 July 2016

Keywords:
Markov decision problems
Value function
Characterization

ABSTRACT

It is known that the value function of a Markov decision process, as a function of the discount factor λ, is the maximum of finitely many rational functions in λ. Moreover, each root of the denominators of the rational functions either lies outside the unit ball in the complex plane, or is a unit root with multiplicity 1. We prove the converse of this result, namely, every function that is the maximum of finitely many rational functions in λ, satisfying the property that each root of the denominators of the rational functions either lies outside the unit ball in the complex plane, or is a unit root with multiplicity 1, is the value function of some Markov decision process. We thereby provide a characterization of the set of value functions of Markov decision processes.

1. Introduction

Markov decision processes (MDP for short) are a standard tool for studying dynamic optimization problems. The discounted value of such a problem is the maximal total discounted amount that the decision maker can guarantee to himself. By Blackwell [1], the function $\lambda \mapsto v_\lambda(s)$ that assigns the discounted value at the initial state s to each discount factor λ is the maximum of finitely many rational functions (with real coefficients). Standard arguments show that the roots of the polynomial in the denominator of these rational functions lie outside the unit ball in the complex plane, or on the boundary of the unit ball, in which case they have multiplicity 1. Using the theory of eigenvalues of stochastic matrices one can show that the roots on the boundary of the unit ball must be unit roots.

In this note we prove the converse result: every function $\lambda \mapsto v_\lambda$ that is the maximum of finitely many rational functions such that each root of the polynomials in the denominators either lies outside the unit ball in the complex plane, or is a unit root with multiplicity 1 is the value of some Markov decision process.

2. The model and the main theorem

Definition 1. A *Markov decision process* is a tuple (S, μ, A, r, q) where

- S is a nonempty finite set of states.
- $\mu \in \Delta(S)$ is the distribution according to which the initial state is chosen, where $\Delta(X)$ is the set of probability distributions over X, for every nonempty finite set X.
- $A = (A(s))_{s \in S}$ is the family of nonempty and finite sets of actions available at each state $s \in S$. Denote $SA := \{(s, a) : s \in S, a \in A(s)\}$.
- $r : SA \to \mathbb{R}$ is a payoff function.
- $q : SA \to \Delta(S)$ is a transition function.

The process starts at an initial state $s_1 \in S$, chosen according to μ. It then evolves in discrete time: at every stage $n \in \mathbb{N}$ the process is in a state $s_n \in S$, the decision maker chooses an action $a_n \in A(s_n)$, and a new state s_{n+1} is chosen according to $q(\cdot \mid s_n, a_n)$.

A *finite history* is a sequence $h_n = (s_1, a_1, s_2, a_2, \ldots, s_n) \in H := \cup_{k=0}^{\infty} (SA)^k \times S$. A *pure strategy* is a function $\sigma : H \to \cup_{s \in S} A(s)$ such that $\sigma(h_n) \in A(s_n)$ for every finite history $h_n = (s_1, a_1, \ldots, s_n)$, and a *behavior strategy* is a function $\sigma : H \to \cup_{s \in S} \Delta(A(s))$ such that $\sigma(h_n) \in \Delta(A(s_n))$ for every such finite history. In other words, a behavior strategy σ assigns to every finite history a distribution over the set of available actions, which we call a *mixed action*. The set of behavior strategies is denoted $\mathcal{B}$. A strategy is *stationary* if for every finite history $h_n = (s_1, a_1, \ldots, s_n)$, the mixed action $\sigma(h_n)$ is a function of s_n and is independent of $(s_1, a_1, \ldots, a_{n-1})$.

Every behavior strategy together with a prior distribution μ over the state space induce a probability distribution $\mathbf{P}_{\mu,\sigma}$ over the space of infinite histories $(SA)^\infty$ (which is endowed with the product σ-algebra). Expectation w.r.t. this probability distribution is denoted $\mathbf{E}_{\mu,\sigma}$.

[*] Corresponding author at: School of Mathematical Sciences, Tel Aviv University, Tel Aviv 6997800, Israel.
E-mail addresses: lehrer@post.tau.ac.il (E. Lehrer), eilons@post.tau.ac.il (E. Solan), omrisola@post.tau.ac.il (O. N. Solan).

http://dx.doi.org/10.1016/j.orl.2016.06.007
0167-6377/© 2016 Elsevier B.V. All rights reserved.

For every discount factor $\lambda \in [0, 1)$, the λ-*discounted payoff* is

$$\gamma_\lambda(\mu, \sigma) := \mathbf{E}_{\mu,\sigma}\left[\sum_{n=1}^{\infty} \lambda^{n-1} r(s_n, a_n)\right].$$

When μ is a probability measure that is concentrated on a single state s we denote the λ-discounted payoff also by $\gamma(s, \sigma)$. The λ-*discounted value* of the Markov decision process, with the prior μ over the initial state is

$$v_\lambda(\mu) := \sup_{\sigma \in \mathcal{B}} \gamma_\lambda(\mu, \sigma). \tag{1}$$

A behavior strategy is λ-*discounted optimal* if it attains the maximum in (1).

Denote by $\mathcal{V}$ the set of all functions $\lambda \mapsto v_\lambda(\mu)$ that are the value function of some Markov decision process starting with some prior $\mu \in \Delta(S)$. The goal of the present note is to characterize the set $\mathcal{V}$.

A Markov decision process is *degenerate* if $|A(s)| = 1$ for every $s \in S$, that is, the decision maker makes no choices along the process. When M is a degenerate Markov decision process we omit the reference to the action in the functions r and q. A degenerate Markov decision process is thus a quadruple (S, μ, r, q), where S is the state space, μ is a probability distribution over S, $r : S \to \mathbb{R}$, and $q(\cdot \mid s)$ is a probability distribution over S for every state $s \in S$.

Denote by $\mathcal{V}_D$ the set of all functions that are payoff functions of some degenerate Markov decision process and by $Max\mathcal{V}_D$ the set of functions that are the maximum of a finite number of functions in $\mathcal{V}_D$. By Blackwell [1] we have $\mathcal{V} = Max\mathcal{V}_D$.

Recall that a complex number $\omega \in \mathbb{C}$ is a *unit root* if there exists $n \in \mathbb{N}$ such that $\omega^n = 1$.

Notation 1. (i) *Denote by $\mathcal{F}$ the set of all rational functions P/Q such that each root of Q is either (a) outside the unit ball, or (b) a unit root with multiplicity 1.*

(ii) *Denote by $Max\mathcal{F}$ the set of functions that are the maximum of a finite number of functions in $\mathcal{F}$.*

The next proposition states that any function in $\mathcal{V}$ is the maximum of a finite number of functions in $\mathcal{F}$.

Proposition 1. *$\mathcal{V}_D \subseteq \mathcal{F}$, and consequently $\mathcal{V} \subseteq Max\mathcal{F}$.*

Proof. Fix a degenerate MDP. For every prior μ, and every discount factor $\lambda \in [0, 1)$, the vector $(\gamma_\lambda(s_1))_{s_1 \in S}$ is the unique solution of a system of $|S|$ linear equations in λ:

$$\gamma_\lambda(s) = r(s) + \lambda \sum_{s' \in S} q(s' \mid s)\gamma_\lambda(s'), \quad \forall s \in S.$$

It follows that

$$\gamma_\lambda = (I - \lambda\mathcal{Q})^{-1} \cdot r,$$

where $\mathcal{Q} = (q(s' \mid s))_{s,s' \in S}$. By Cramer's rule, the function $\lambda \mapsto (I - \lambda\mathcal{Q})^{-1}$ is a rational function whose denominator is $\det(I - \lambda\mathcal{Q})$. In particular, the roots of the denominator are the inverse of the eigenvalues of $\mathcal{Q}$. Since the denominator is independent of s, it is also the denominator of $\gamma_\lambda(\mu) = \sum_{s \in S} \mu(s)\gamma_\lambda(s)$.

Denote the expected payoff at stage n by $x_n := \mathbf{E}_\mu[r(s_n)]$, so that $\gamma_\lambda(\mu) = \sum_{n=1}^{\infty} x_n \lambda^{n-1}$. Since $|x_n| \le \|r\|_\infty := \max_{(s,a) \in SA} |r(s, a)|$ for every $n \in \mathbb{N}$, it follows that the denominator $\det(I - \lambda\mathcal{Q})$ does not have roots in the interior of the unit ball and that all its roots that lie on the boundary of the unit ball have multiplicity 1. These two observations hold since by the triangle inequality we have

$$|\gamma_\lambda(\mu)| = \left|\sum_{n=1}^{\infty} x_n \lambda^{n-1}\right| \le \|r\|_\infty \sum_{n=1}^{\infty} |\lambda|^{n-1} = \frac{\|r\|_\infty}{1 - |\lambda|}. \tag{2}$$

If λ_0 is a root of $\det(I - \lambda\mathcal{Q})$ that lie in the interior of the unit ball, then for the payoff function $r \equiv 1$ we would have that $\gamma_{\lambda_0}(\mu) = \infty$, which violates (2). Similarly, if λ_0 is a root of $\det(I - \lambda\mathcal{Q})$ with multiplicity at least 2 that lies on the boundary of the unit ball, then for the payoff function $r \equiv 1$ Eq. (2) is violated.

Moreover, by, Dmitriev and Dynkin [2] the roots that lie on the boundary of the unit ball must be unit roots. ■

The main result of this note is that the converse holds as well.

Theorem 1. *$\mathcal{V}_D \supseteq \mathcal{F}$, and consequently $\mathcal{V} = Max\mathcal{F}$.*

To avoid cumbersome notations we write $f(\lambda)$ for the function $\lambda \mapsto f(\lambda)$. In particular, $\lambda f(\lambda)$ will denote the function $\lambda \mapsto \lambda f(\lambda)$.

3. Characterizing the set $\mathcal{V}_D$

The following lemma lists several properties of the functions implementable by degenerate Markov decision processes.

Lemma 1. *For every $f \in \mathcal{V}_D$ we have*

(a) *$af(\lambda) \in \mathcal{V}_D$ for every $a \in \mathbb{R}$.*
(b) *$f(-\lambda) \in \mathcal{V}_D$.*
(c) *$\lambda f(\lambda) \in \mathcal{V}_D$.*
(d) *$f(c\lambda) \in \mathcal{V}_D$ for every $c \in [0, 1]$.*
(e) *$f(\lambda) + g(\lambda) \in \mathcal{V}_D$ for every $g \in \mathcal{V}_D$.*
(f) *$f(\lambda^n) \in \mathcal{V}_D$ for every $n \in \mathbb{N}$.*

Proof. Let $M_f = (S_f, \mu_f, r_f, q_f)$ be a degenerate Markov decision process whose value function is f.

To prove Part (a), we multiply all payoffs in M_f by a. Formally, define a degenerate Markov decision process $M' = (S_f, \mu_f, r', q_f)$ that differs from M only in its payoff function: $r'(s) := ar_f(s)$ for every $s \in S_f$. The reader can verify that the value function of M' is $af(\lambda)$.

To prove Part (b), multiply the payoff in even stages by -1. Formally, let $\widehat{S}$ be a copy of S_f; for every state $s \in S_f$ we denote by $\widehat{s}$ its copy in $\widehat{S}$. Define a degenerate Markov decision process $M' = (S_f \cup \widehat{S}, \mu_f, r', q')$ with initial distribution μ_f (whose support is S_f) that visits states in $\widehat{S}$ in even stages and states in S_f in odd stages as follows:

$$r'(s) := r_f(s), \qquad r'(\widehat{s}) := -r_f(s), \quad \forall s \in S_f,$$
$$q'(\widehat{s'} \mid s) = q'(s' \mid \widehat{s}) := q_f(s' \mid s), \quad \forall s, s' \in S_f,$$
$$q'(s' \mid s) = q'(\widehat{s'} \mid \widehat{s}) := 0, \quad \forall s, s' \in S_f.$$

The reader can verify that the value function of M' is $f(-\lambda)$.

To prove part (c), add a state with payoff 0 from which the transition probability to a state in S_f coincides with μ. Formally, define a degenerate Markov decision process $M' = (S_f \cup \{s^*\}, \mu', r', q')$ in which μ' assigns probability 1 to s^*. r' coincides with r_f on S_f, while $r'(s^*) := 0$. Finally, q' coincides with q_f on S_f, while at the state s^*, $q'(\cdot \mid s^*) := \mu$. The value function of M' is $\lambda f(\lambda)$.

A state $s \in S$ is *absorbing* if $q(s \mid s, a) = 1$ for every action $a \in A(s)$. To prove part (d), consider the transition function that at every stage, moves to an absorbing state with payoff 0 with probability $1 - c$, and with probability c continues as in M. Formally, define a degenerate Markov decision process $M' = (S_f \cup \{s^*\}, \mu, r', q')$ in which μ coincides with μ_f, r' and q' coincide with r_f and q_f on S_f, $r'(s^*) := 0$, and $q'(s^* \mid s^*) := 1$ (that is, s^* is an absorbing state), and

$$q'(s^* \mid s) := 1 - c, \qquad q'(s' \mid s) := cq_f(s' \mid s), \quad \forall s, s' \in S_f.$$

The value function of M' at the initial state $s_{1,f}$ is $f(c\lambda)$.

To prove Part (e), we show that $(1/2)f + (1/2)g$ is in $\mathcal{V}_D$ and we use part (a) with $a = 2$. The function $(1/2)f + (1/2)g$ is the value

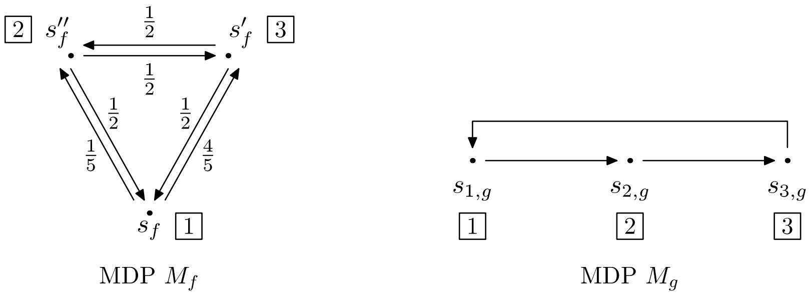

Fig. 1. An example of two MDP's.

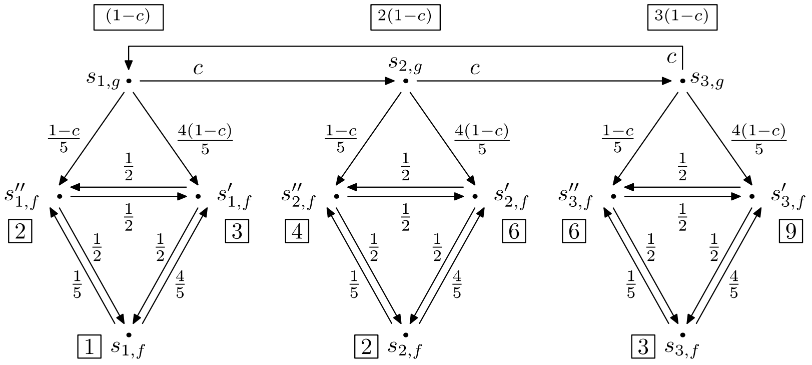

Fig. 2. The degenerate MDP M.

function of the degenerate Markov decision process in which the prior chooses with probability 1/2 one of two degenerate Markov decision processes that implement f and g. Formally, let $M_g = (S_g, \mu_g, r_g, q_g)$ be a degenerate Markov decision process whose value function is g. Let $M = (S_f \cup S_g, \mu', r', q')$ be the degenerate Markov decision process whose state space consists of disjoint copies of S_f and S_g, the functions r' (resp. q') coincide with r_f and r_g (resp. q_f and q_g) of S_f (resp. S_g), and the initial distribution is $\mu' = (1/2)\mu_f + (1/2)\mu_g$. The value function of M is $(1/2)f + (1/2)g$.

To prove Part (f), we space out the Markov decision process in a way that stage k of the Markov decision process that implements f becomes stage $1 + (k-1)n$, and the payoff in all other stages is 0. Formally, let $M' = (S_f \times \{1, 2, \ldots, n\}, \mu', r', q')$ be a degenerate Markov decision process where $\mu' = \mu_f$ and

$$q'((s, k+1) \mid (s, k)) := 1 \quad k \in \{1, 2, \ldots, n-1\},\ s \in S,$$
$$q'(\cdot \mid (s, n)) := q_f(s) \quad s \in S,$$
$$r'((s, 1)) := r_f(s) \quad s \in S,$$
$$r'((s, k)) := 0 \quad k \in \{2, 3, \ldots, n\},\ s \in S.$$

The value function of M' with the prior μ' is $f(\lambda^n)$. ■

In the following lemma and later in the paper, whenever we refer to a polynomial we mean a polynomial with real coefficients.

Lemma 2. (a) *Every polynomial P is in $\mathcal{V}_D$ and if $f \in \mathcal{V}_D$, then $P \cdot f$ is also in $\mathcal{V}_D$.*
(b) *Let P and Q be two polynomials. If $1/Q \in \mathcal{V}_D$ then $P/Q \in \mathcal{V}_D$. In particular, if Q' divides Q and $1/Q \in \mathcal{V}_D$ then $1/Q' \in \mathcal{V}_D$.*
(c) *If Q is a polynomial whose roots are all unit roots of multiplicity 1, then $\frac{1}{Q} \in \mathcal{V}_D$.*

Proof. Part (a) follows from Lemma 1(a,c,e) and the observation that any constant function a is in $\mathcal{V}_D$, which holds since the constant function a is the value function of the degenerate Markov decision process that starts with a state whose payoff is a and continues to an absorbing state whose payoff is 0.

Part (b) follows from Part (a).

We turn to prove Part (c). The degenerate Markov decision process with a single state in which the payoff is 1 yields payoff $1/(1-\lambda)$, and therefore $1/(1-\lambda) \in \mathcal{V}_D$. By Lemma 1(f) it follows that $1/(1-\lambda^n) \in \mathcal{V}_D$, for every $n \in \mathbb{N}$. Let n be large enough such that Q divides $1 - \lambda^n$. The result now follows by Part (b) of this lemma. ■

To complete the proof of Theorem 1 we characterize the polynomials Q that satisfy $1/Q \in \mathcal{V}_D$. To this end we need the following property of $\mathcal{V}_D$.

Lemma 3. *If $f, g \in \mathcal{V}_D$ then $f(\lambda)g(\lambda c) \in \mathcal{V}_D$ for every $c \in (0, 1)$.*

The lemma holds for $c = 0$ and does not hold for $c = 1$. Indeed, for $c = 0$ we have by Lemma 1(a) $f(\lambda)g(\lambda c) = g(0)f(\lambda) \in \mathcal{V}_D$; for $c = 1$ we already saw that $1/(1-\lambda) \in \mathcal{V}_D$ while Theorem 1 implies that $1/(1-\lambda)^2 \notin \mathcal{V}_D$.

Proof. The proof of Lemma 3 is the most intricate part of the proof of Theorem 1. We start with an example, that will help us illustrate the formal definition of the degenerate MDP that implements $f(\lambda)g(\lambda c)$.

Let M_f and M_g be the degenerate Markov decision processes that are depicted in Fig. 1 with the initial distributions $\mu_f(s_f) = 1$ and $\mu_g(s_{1,g}) = 1$, in which the payoff at each state appears in a square next to the state. Denote by f and g the value functions of M_f and M_g, respectively.

Consider the degenerate Markov decision process M depicted in Fig. 2, where $c \in (0, 1)$ and the initial state is $s_{1,g}$. The MDP M is composed of one copy of M_g, and for every state in S_g it contains one copy of M_f. It starts at $s_{1,g}$, the initial state of M_g. Then, at every stage, with probability c it continues as in M_g, and with probability $(1-c)$ it moves to a copy of M_f. In case a transition to a copy of M_f occurs, the new state is chosen according to the transitions $q_f(\cdot \mid s_f)$. This induces a distribution similar to that of the second stage of M_f.

The payoff in each of the copies of M_f is the product of the payoff in M_f times the payoff of the state in M_g that has been assigned to that copy. The payoff in each state $s_g \in S_g$ is $(1-c)r_g(s_g)$ times the expected payoff in the first stage of M_f.

Thus, each state in $s_g \in S_g$ serves three purposes (see Fig. 2). First, it is a regular state in the copy of M_g. Second, once a transition from M_g to a copy of M_f occurs (at each stage it occurs with probability $(1-c)$), it serves as the first stage in M_f. Finally, once a transition from s_g to a copy of M_f occurs, the payoffs in the copy are set to the product of the original payoffs in M_f times $r_g(s_g)$.

We now turn to the formal construction of M. Let $f, g \in \mathcal{V}_D$ and let $M_f = (S_f, \mu_f, r_f, q_f)$ (resp. $M_g = (S_g, \mu_g, r_g, q_g)$) be the degenerate Markov decision process that implements f (resp. g). Define the following degenerate Markov decision process $M = (S, \mu, r, q)$:

- The set of states is $S = (S_g \times S_f) \cup S_g$. In words, the set of states contains a copy of S_g, and for each state $s_g \in S_g$ it contains a copy of S_f.
- The initial distribution is μ_g:

$$\mu(s_g) = \mu_g(s_g), \quad \forall s_g \in S_g,$$
$$\mu(s_g, s_f) = 0, \quad \forall (s_g, s_f) \in S_g \times S_f.$$

- The transition is as follows:
  - In each copy of S_f, the transition is the same as the transition in M_f:
  $q((s_g, s'_f) \mid (s_g, s_f)) := q_f(s'_f \mid s_f), \quad \forall s_g \in S_g, s_f, s'_f \in S_f,$
  $q((s'_g, s'_f) \mid (s_g, s_f)) := 0, \quad \forall s_g \neq s'_g \in S_g, s_f, s'_f \in S_f.$
  - In the copy of S_g, with probability c the transition is as in M_g, and with probability $(1-c)$ it is as in M_f starting with the initial distribution μ_f:
  $q(s'_g \mid s_g) := c q_g(s'_g \mid s_g), \quad \forall s_g, s'_g \in S_g,$
  $$q((s_g, s_f) \mid s_g) := (1-c) \sum_{s'_f \in S_f} \mu(s'_f) q_f(s_f \mid s'_f),$$
  $\forall s_g \in S_g, s_f \in S_f,$
  $q((s'_g, s_f) \mid s_g) := 0 \quad \forall s_g \neq s'_g \in S_g, s_f \in S_f.$
- The payoff function is as follows:

$$r(s_g) := (1-c) r_g(s_g) \sum_{s_f \in S_f} \mu_f(s_f) r_f(s_f), \quad \forall s_g \in S_g,$$
$$r(s_g, s_f) := r_g(s_g) r_f(s_f), \quad \forall s_g \in S_g, s_f \in S_f.$$

We will now calculate the value function of M. Denote by

$$\mathbf{E}_{\mu_f}[r_f] := \sum_{s_f \in S_f} \mu_f(s_f) r_f(s_f)$$

the expected payoff in M_f at the first stage and by R the expected payoff in M_f from the second stage and on. Then $f(\lambda) = \mathbf{E}_{\mu_f}[r_f] + \lambda R$.

At every stage, with probability c the process remains in S_g and with probability $(1-c)$ the process leaves this set. In particular, the probability that at stage n the process is still in S_g is c^{n-1}, in which case (a) the payoff is $(1-c) r_g(s_n) \mathbf{E}_{\mu_f}[r_f]$, and (b) with probability $(1-c)$ the process moves to a copy of M_f, and the total discounted payoff from stage $n+1$ and on is R. It follows that the total discounted payoff is

$$\mathbf{E}\left[\sum_{n=1}^{\infty} c^{n-1} \lambda^{n-1} \big((1-c) r_g(s_n) \mathbf{E}_{\mu_f}(r_f) + (1-c)\lambda r_g(s_n) R\big)\right]$$
$$= \mathbf{E}\left[(1-c) \sum_{n=1}^{\infty} c^{n-1} \lambda^{n-1} r_g(s_n) f(\lambda)\right]$$
$$= (1-c) g(c\lambda) f(\lambda).$$

The result follows by Lemma 1(a). ■

Lemma 4. *Let $\omega \in \mathbb{C}$ be a complex number that lies outside the unit ball.*

(a) *If $\omega \in \mathbb{C} \setminus \mathbb{R}$ then $\frac{1}{(\omega-\lambda)(\overline{\omega}-\lambda)} \in \mathcal{V}_D$, where $\overline{\omega}$ is the conjugate of ω.*

(b) *If $\omega \in \mathbb{R}$ then $1/(\omega - \lambda) \in \mathcal{V}_D$.*

Proof. We start by proving part (a). For every complex number $\omega \in \mathbb{C} \setminus \mathbb{R}$ that lies outside the unit ball there are three natural numbers $k < l < m$ and three nonnegative reals $\alpha_1, \alpha_2, \alpha_3$ that sum up to 1 such that $1 = \alpha_1 \omega^k + \alpha_2 \omega^l + \alpha_3 \omega^m$. Indeed, denote the convex hull of a nonempty set $X \subseteq \mathbb{R}^n$ by conv$X$. Since ω is not a real number, there is $k \in \mathbb{N}$ such that the imaginary part of ω^k is negative. It follows that the origin is in the interior of conv$\{\omega^k, \omega^{-k}, 1\}$, and by multiplying these three terms by ω^{k+1} we obtain that the origin is also in conv$\{\omega, \omega^{k+1}, \omega^{2k+1}\}$. Therefore, the set $E :=$ conv$\{\omega^k : k \in \mathbb{N}\}$ contains an open ball around the origin and it contains ωE. Since $|\omega| > 1$, this implies that E is the whole plane, and in particular contains 1.

Consider the degenerate Markov decision process that is depicted in Fig. 3. That is, the set of states is $S_f := \{s_1, s_2, \ldots, s_m\}$, the payoff function is

$$r(s_m) := 1, \qquad r(s_j) := 0, \quad 1 \leq j < m,$$

and the transition function is

$$q(s_{m-k+1} \mid s_m) := \alpha_1, \qquad q(s_{m-l+1} \mid s_m) := \alpha_2,$$
$$q(s_1 \mid s_m) := \alpha_3,$$
$$q(s_{j+1} \mid s_j) = 1, \quad 1 \leq j < m.$$

The discounted value satisfies

$$v_\lambda(s_j) = \lambda v_\lambda(s_{j+1}), \quad 0 \leq j < m,$$
$$v_\lambda(s_m) = 1 + \lambda\big(\alpha_1 v_\lambda(s_{m-k+1}) + \alpha_2 v_\lambda(s_{m-l+1}) + \alpha_3 v_\lambda(s_1)\big).$$

It follows that at the initial state is s_m,

$$v_\lambda(s_m) = \frac{1}{1-\alpha_1\lambda^k-\alpha_2\lambda^l-\alpha_3\lambda^m}.$$

Hence, this function is in $\mathcal{V}_D$. Since ω is one of the roots of the denominator, by Lemma 2(b) we obtain that $\frac{1}{(\omega-\lambda)(\overline{\omega}-\lambda)} \in \mathcal{V}_D$, as desired.

We turn to prove part (b). Let ω be a real number with $\omega > 1$. As mentioned in the proof of Lemma 2, $1/(1-\lambda) \in \mathcal{V}_D$. By Lemma 1(d), $1/(1-\lambda/\omega) \in \mathcal{V}_D$, and by Lemma 1(a), $1/(\omega-\lambda) \in \mathcal{V}_D$. By Lemma 1(a,b), $1/(-\omega-\lambda) \in \mathcal{V}_D$, which complete the proof of part (b). ■

We are finally ready to prove Theorem 1.

Proof of Theorem 1. Let $Q \neq 0$ be a polynomial with real coefficients whose roots are either outside the unit ball or unit roots with multiplicity 1. To complete the proof of Theorem 1 we prove that $1/Q \in \mathcal{V}_D$. Denote by Ω_1 the set of all roots of Q that are unit roots, by Ω_2 the set of all roots of Q that lie outside the unit ball and have a positive imaginary part, and by Ω_3 the set of all real roots of Q that lie outside the unit ball. If some roots have multiplicity larger than 1, then they appear several times in Ω_2 or Ω_3.

For $i = 1, 3$ denote $Q_i := \prod_{\omega \in \Omega_i}(\omega - \lambda)$ and set $Q_2 := \prod_{\omega \in \Omega_2}(\omega - \lambda)(\overline{\omega} - \lambda)$; when $\Omega_i = \emptyset$ we set $Q_i = 1$. Note that $Q = Q_1 \cdot Q_2 \cdot Q_3$. If $\Omega_1 \neq \emptyset$, then by Lemma 2(c) we have $1/Q_1 \in \mathcal{V}_D$. Otherwise $Q_1 = 1$, in which case, $1/Q_1 \in \mathcal{V}_D$ by Lemma 2(a).

Fix $\omega \in \Omega_2$ and let $c \in \mathbb{R}$ be such that $1 < c < |\omega|$. Since ω/c lies outside the unit ball, Lemma 4(a) implies that $g_\omega(\lambda) := \frac{1}{(\frac{\omega}{c}-\lambda)(\overline{\frac{\omega}{c}}-\lambda)}$ is in $\mathcal{V}_D$. By Lemma 3, $g_\omega(\frac{1}{c} \cdot \lambda) \cdot \frac{1}{Q_1} = \frac{c^2}{(\omega-\lambda)(\overline{\omega}-\lambda)} \cdot \frac{1}{Q_1} \in \mathcal{V}_D$. By Lemma 1(a), $\frac{1}{(\omega-\lambda)(\overline{\omega}-\lambda)} \cdot \frac{1}{Q_1} \in \mathcal{V}_D$. Applying successively this argument for the remaining roots in Ω_2, one obtains that $\frac{1}{Q_2} \cdot \frac{1}{Q_1} \in \mathcal{V}_D$.

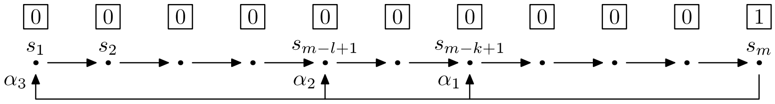

Fig. 3. The degenerate MDP in the proof of Lemma 4.

To complete the proof we apply a similar idea to $\omega \in \Omega_3$. Fix $\omega \in \Omega_3$ and let $c \in \mathbb{R}$ be such that $1 < c < |\omega|$. By Lemma 4(b), $\frac{1}{\frac{\omega}{c}-\lambda} \in \mathcal{V}_D$ and again by Lemmas 1 and 3(a), $\frac{1}{(\omega-\lambda)} \cdot \frac{1}{Q_2} \cdot \frac{1}{Q_1} \in \mathcal{V}_D$. By iterating this argument for every $\omega \in \Omega_3$ one obtains that $\frac{1}{Q_3} \cdot \frac{1}{Q_2} \cdot \frac{1}{Q_1} \in \mathcal{V}_D$, as desired. ■

4. Final remarks

4.1. *MDP's with an initial state*

The set $\mathcal{V}$ contains all value functions of MDP's in which the state at the first stage is chosen according to a probability distribution μ. One can wonder whether the set of implementable value functions shrinks if one restricts attention to MDP's in which the first stage is given; that is, μ assigns probability 1 to one state. The answer is negative: the value function of any MDP in which the initial state is chosen according to a probability distribution (prior) can be obtained as the value function of an MDP in which the initial state is deterministic. Indeed, let M be an MDP with a prior. One can construct an MDP M' by adding to M an initial state s' in which the payoff is the expected payoff at the first stage of M and the transitions are the expected transitions after the first stage of M.

4.2. *The size of the MDP that implements a given value function*

A colleague who read our paper came up with an alternative proof that uses algebraic tools. The advantage of our proof is that it is constructive. Unfortunately our method requires prohibitively large MDP's to implement functions in $\mathcal{V}$. Denote by size(f) the size of the smallest MDP needed to implement f as a value function. Our results provide upper bounds on size(f):

- If all roots of Q are unit roots with multiplicity 1, then $\text{size}(1/Q) \leq (n+1)^2$, where n is sufficiently large so that Q divides $1-\lambda^n$.
- If ω is a real number outside the unit ball, then $\text{size}(1/(\lambda-\omega)) \leq 2$, and one of the states in the implementing MDP is an absorbing state with payoff 0.
- If $\omega \in \mathbb{C}\setminus\mathbb{R}$ lies outside the unit ball, then $\text{size}(\frac{1}{(\lambda-\omega)(\lambda-\overline{\omega})}) \leq m$, where m is sufficiently large so that $1 \in \text{conv}\{\omega, \omega^2, \ldots, \omega^m\}$.
- If $f \in \mathcal{V}$ and P is a polynomial, then $\text{size}(P \cdot f) \leq 1 + \text{size}(f) \times (\deg(P)+1)$.
- If $f, g \in \mathcal{V}$ then $\text{size}(f(\lambda)g(c\lambda)) \leq \text{size}(g) \times (\text{size}(f)+1)$.
- If $f_1, \ldots, f_K \in \mathcal{V}_D$ then $\text{size}(\max_{k=1,\ldots,K} f_k) \leq 1 + \sum_{k=1}^{K} \text{size}(f_k)$. Moreover, there is an MDP that implements the function $\max_{k=1,\ldots,K} f_k$ in which the number of actions in all but one state is 1, and the number of actions in the initial state is K.

Suppose that $P/Q \in \mathcal{V}_D$ and define the sets of roots $(\Omega_i)_{i=1}^3$ and the polynomials $(Q_i)_{i=1}^3$ as in the proof of Theorem 1. Denote $\Omega_2 = \{\omega_1, \omega_2, \ldots, \omega_L\}$ and for every l, $1 \leq l \leq L$, denote $m_l := \text{size}(\frac{1}{(\lambda-\omega_l)(\lambda-\overline{\omega}_l)})$. Then

$$\text{size}(P/Q) \leq (\deg(P)+1) \times \big(1 + (((((n+1)^2 + |\Omega_3| \\ + 1)m_1 + 1)m_2 + 1)m_3 + 1) \cdots m_L\big),$$

where n is large enough so that Q_1 divides $1-\lambda^n$. An interesting open problem is the identification of the smallest MDP that can implement any given function in $\mathcal{V}$.

Acknowledgments

We thank Ron Peretz and an anonymous referee for commenting on an earlier version of the paper. Lehrer acknowledges the support of the Israel Science Foundation, Grant #963/15. Solan acknowledges the support of the Israel Science Foundation, Grant #323/13.

References

[1] D. Blackwell, Discrete dynamic programming, Ann. Math. Stat. 33 (1962) 719–726.

[2] N.A. Dmitriev, E. Dynkin, On characteristic roots of stochastic matrices, Izv. Akad. Nauk SSSR Ser. Mat. 10 (1946) 167–184. Translated in Eleven Papers Translated from the Russian (American Mathematical Society Translations), ser. 2, 140, 1988, 57–78.