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/

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