ON THE EXISTENCE OF A LAURENT EXPANSION OF THE VALUE FUNCTION IN CONSTRAINED CONTROL OF MARKOV CHAINS

E. ALTMAN
INRIA, Centre Sophia-Antipolis
2004 Route des Lucioles, B.P.93
06902 Sophia-Antipolis Cedex
France

A. HORDIJK and L. Kallenberg
Dept. of Mathematics and Computer Science
Leiden University
P.O.Box 9512, 2300RA Leiden
The Netherlands

April 1995

Abstract

The existence of a Laurent expansion of the value function in controlled Markov chains with finitely many states and actions is well known. We show in this note that this property also holds for the optimal value of controlled Markov chains with additional constraints (CMDP). More precisely, we show that (i) the interval [0,1] can be divided into a finite number of sub-intervals $< 0 = \alpha_0, \alpha_1 >, < \alpha_1, \alpha_2 >, ..., < \alpha_s, \alpha_{s+1} = 1 >$ such that in any interval i, either the CMDP is not feasible, or the value function is a rational function of the discount factor α. The intervals may be closed or open in either sides. (ii) For any $\alpha \in (0, 1]$ there exists some $0 \leq \alpha^- < \alpha$; the value function has a Laurent expansion in the whole interval (α^-, α), or CMDP is not feasible in this interval. Similarly, for every $\alpha \in [0, 1)$ there exists some $\alpha^+ > \alpha$ such that either the value function has a Laurent expansion in the whole interval (α, α^+), or CMDP is not feasible in this interval.

Key words: Constrained Markov decision processes, Laurent expansion.

1 Introduction

In many applications of Markov decision processes (MDPs), the discount factor may be known only within some range. A natural question is then how sensitive is the value to changes in the discount factor α. In particular, an interesting question is the behavior of the optimal values and policies in the neighborhood of $\alpha = 1$, and the convergence of the values of the discounted problems to the expected average cost MDP, as the discount factor α approaches 1. We shall use below the scaled version of the discounted cost (i.e. the total expected discounted cost scaled by 1 - α). Let us agree, for simplicity of the presentation, that by a cost with discount factor of 1 we mean the expected average cost.

For unconstrained MDPs the following is known: The interval [0,1] of discount factors can be divided into a finite number of closed (non-empty) sub-intervals $[0 = \alpha_0, \alpha_1], [\alpha_1, \alpha_2], ..., [\alpha_s, \alpha_{s+1} = 1]$ such that there exists a deterministic stationary policy g_i, $0 \leq i \leq s$, where g_i is α-optimal for all $\alpha \in [\alpha_i, \alpha_{i+1}]$ (see e.g. Blackwell [7], Smallhood [17]). This implies in particular that the value function has a Laurent expansion to the right and to the left of any α in $[0, 1]$.

When there is more than one objective and more than one controller, then the asymptotic behavior of the value function may exhibit a more complex behavior. For Markov games with finitely many states and actions, the values cannot be expressed in general as a Laurent series in the discount factor. In particular, for zero-sum Markov games (two controllers with opposite objectives), Bewley and Kohlberg showed that the value function has in fact a fractional Puiseux expansion in the discount factor [6].

The purpose of this note is to show that the value function of constrained Markov decision processes (CMDP) has a Laurent expansion in the discount factor. More precisely, we show that
(i) the interval [0,1] can be divided into a finite number of sub-intervals $< 0 = \alpha_0, \alpha_1 >, < \alpha_1, \alpha_2 >, ..., < \alpha_s, \alpha_{s+1} = 1 >$ such that in any interval i, either the CMDP is not feasible, or the value function is a rational function of the discount factor α. The intervals may be closed or open in either sides. As a corollary, we have:
(ii) For any $\alpha \in (0, 1]$ there exists some $0 \leq \alpha^- < \alpha$; the value function has a Laurent expansion in the whole interval (α^-, α), or CMDP is not feasible in this interval. Similarly, for every $\alpha \in [0, 1)$ there exists some $\alpha^+ > \alpha$ such that either the value function has a Laurent expansion in the whole interval (α, α^+), or CMDP is not feasible in this interval.

2 The model

Consider an MDP with a finite state space $\mathbf{X} = \{0, 1, ..., N\}$ and a finite *action space* $\mathbf{A}$. Without loss of generality, we assume that in any state x all actions in $\mathbf{A}$ are available. The probability to go from state x to state y given that action a is used, is given by the transition probability P_{xay}. A policy u in the *policy space* $\mathbf{U}$ is described as $u = \{u_1, u_2, ...\}$, where u_t, applied at time epoch t, is a probability measure over $\mathbf{A}$ conditioned on the whole history of actions and state prior to t, as well as the state at time t. Given an initial distribution β on $\mathbf{X}$, each policy u induces a probability measure denoted by P^u_β on the space of sample paths of states and actions (which serves as the canonical sample space Ω). The corresponding expectation operator is denoted by E^u_β. On this probability space are defined the state and action processes, X_t, A_t, $t = 1, 2, ...$

Any given distribution β for the initial state (at time 1) and a policy u define a probability measure P^u_β on which the stochastic processes X_t and A_t of the states and actions are defined.

Let $c : \mathbf{X} \times \mathbf{A} \to \mathbb{R}$, and $d : \mathbf{X} \times \mathbf{A} \to \mathbb{R}^K$ be immediate cost functions, $d = (d^1, d^2, ..., d^k)$.

Fix some discount factor $\alpha \in [0, 1)$, and defined the normalized discounted costs corresponding to an initial distribution β and a policy u by

$$
\begin{aligned}
C_\alpha(\beta, u) &= (1-\alpha)\sum_{t=1}^{\infty} E_\beta^u \alpha^{t-1} c(X_t, A_t) \\
D_\alpha^k(\beta, u) &= (1-\alpha)\sum_{t=1}^{\infty} E_\beta^u \alpha^{t-1} d^k(X_t, A_t), \qquad k = 1, ..., K.
\end{aligned}
$$

Define the average costs associated with a policy u and with an initial distribution β on $\mathbf{X}$:

$$
\begin{aligned}
C_1(\beta, u) &= \overline{\lim_{t\to\infty}} \frac{1}{t} E_\beta^u \left[\sum_{s=1}^{t} c(X_s, A_s)\right] \\
D_1^k(\beta, u) &= \overline{\lim_{t\to\infty}} \frac{1}{t} E_\beta^u \left[\sum_{s=1}^{t} d^k(X_s, A_s)\right], \qquad k = 1, ..., K.
\end{aligned}
$$

Given a vector $V \in \mathbb{R}^{\mathrm{K}}$, we consider the subset $\Pi_V \subset \mathbf{U}$ of policies satisfying the constraints

$$D(\beta, u) \le V. \tag{1}$$

A policy $u \in \Pi_V$ is called feasible. We introduce the following Constrained Problem CMDP: find a policy u^* that achieves

$$C(\beta) := \inf_{u \in \Pi_V} C(\beta, u). \tag{2}$$

In (1) and (2), the costs stand for either the discounted or the average cost. CMDP is said to be feasible if Π_V is nonempty.

There are several ways to solve (2). For the discounted cost, the solution was given by Kallenberg in [12] using a LP approach. For the expected average cost, there are several possible LP approaches: the one by Hordijk and Kallenberg [11, 12], the one by Feinberg [10], and a related one by Altman and Shwartz [3] and Altman and Tidball [18]; for a definition slightly different than in (2), an efficient LP method for computing ϵ-optimal solutions was obtained by Ross and Varadarajan, see [13, 14]. Lagrangian techniques have also been used to solve constrained MDPs with a single constraint, see Beutler, Ross and Sennott [4, 5, 15, 16].

Denote the simplex $S(L) := \{\gamma \in R^L : \gamma_i \ge 0, i = 1, ..., L, \ \sum_{i=1}^{L} \gamma_i = 1\}$.

The following LP [10, 18] can be used to compute the optimal values (and policies) of the CMDPs. (The alternative LP approach introduced by Kallenberg [12] and Hordijk and Kallenberg [11] is much simpler computationally. However, the LP below has the advantage over the other LP methods that it has the same form for both discounted and expected average costs). Let L stand for the cardinality of the set of stationary deterministic policies. Introduce the following $\mathbf{LP}_\alpha$, $\alpha \in [0, 1]$: Find $\gamma^* \in S(L)$ that achieves:

$$\mathcal{C}_\alpha := \min_{\gamma \in S(L)} \sum_{i=1}^{L} \gamma_i C_\alpha(\beta, u^i), \qquad s.t. \tag{3}$$

$$\sum_{i=1}^{L} \gamma_i D_\alpha^k(\beta, u^i) \le V^k, \ k = 1, ..., K \tag{4}$$

We say that the LP is feasible if the subset of $S(L)$ satisfying the constraints (4) is nonempty.

Lemma 2.1 *(Relation between LP and the constrained MDP).*
For any $\alpha \in [0,1]$, $CMDP_\alpha$ is feasible if and only if LP_α is, and the optimal values are the same: $\mathcal{C}_\alpha = C_\alpha(\beta)$.

In order to apply standard LP solution methods, we shall rewrite $\mathbf{LP}_\alpha$ in the following generic (matrix) form (where additional decision variables are added in the standard way, so as to obtain equality constraints):

$$\mathcal{C}_\alpha := \inf c \cdot z$$

$$s.t.\ \ \mathcal{A}(\alpha)z = b(\alpha), \quad z \geq 0.$$

For any fixed stationary policy g, the cost $C_\alpha(\beta, g)$ is a rational function in α in the interval $[0,1]$ and therefore has a Laurent expansion for all $\alpha \in [0,1]$ (see Dekker [8] p. 149). It follows that $\mathcal{A}$ and b above are rational functions in α over $[0,1]$.

Our main result for CMDPs is:

Theorem 2.1 *The interval [0,1] can be divided into a finite number of sub-intervals $< 0 = \alpha_0, \alpha_1 >$, $< \alpha_1, \alpha_2 >, ..., < \alpha_s, \alpha_{s+1} = 1 >$ such that in any interval i, either the CMDP is not feasible, or the value function is a rational function of α.*

The above intervals may be closed or open in either sides. In particular, there may be singular points α_i where the value is not continuous to the right nor to the left. Conditions for continuity at a point are given in [1, 2]. A characterization of discontinuity (singular perturbations) are given in [1].

Proof. In order to prove Theorem 2.1 we shall show the following: Assume that we solve $\mathbf{LP}_\alpha$ using the simplex method. The following holds:

- The interval [0,1] can be divided into a finite number of sub-intervals $< 0 = \alpha_0, \alpha_1 >, < \alpha_1, \alpha_2 >, ..., < \alpha_s, \alpha_{s+1} = 1 >$; for each i, $\mathbf{LP}_\alpha$ is either feasible or unfeasible for all α in the interval. If it is feasible in interval i then there is a basis B_i such that for any α in interval i, an optimal solution of $\mathbf{LP}_\alpha$ is obtained using the same basis B_i, i.e.

$$\mathcal{C}_\alpha = cB_i^{-1}(\alpha)b(\alpha) \qquad \text{for all } \alpha \in < \alpha_i, \alpha_{i+1} > .$$

Assume that the above does not hold. Then either
(i) There exists some basis B and a sequence α_n such that $B^{-1}(\alpha_{2n})b(\alpha_{2n}) = 0$ and such that $\mathbf{LP}_{\boldsymbol{\alpha_{2n+1}}}$ are not feasible. This can be seen to imply, in particular, that $B^{-1}(\alpha_{2n+1})b(\alpha_{2n+1}) \neq 0$.
(ii) there are two basis B_1 and B_2 and there exists an infinite converging sequence α_n such that B_1 is optimal for all $\alpha = \alpha_{2n}$, $n \in \mathbb{N}$, and B_2 is optimal for all $\alpha = \alpha_{2n+1}$, $n \in \mathbb{N}$. Moreover, B_2 is suboptimal for all $\alpha = \alpha_{2n}$, $n \in \mathbb{N}$, and B_1 is suboptimal for all $\alpha = \alpha_{2n+1}$, $n \in \mathbb{N}$. The values of the LP satisfy

$$\mathcal{C}_{\alpha(2n)} = cB_1^{-1}(\alpha_{2n})b(\alpha_{2n}), \qquad \mathcal{C}_{\alpha(2n+1)} = cB_2^{-1}(\alpha_{2n+1})b(\alpha_{2n})$$

Moreover,

$$cB_1^{-1}(\alpha_i)b(\alpha_i) - cB_2^{-1}(\alpha_{i+1})b(\alpha_{2n}) < 0,\ i = 2n$$

$$cB_1^{-1}(\alpha_i)b(\alpha_i) - cB_2^{-1}(\alpha_{i+1})b(\alpha_{2n}) > 0,\ i = 2n+1$$

(i) cannot hold, since $B^{-1}b$ is a rational function in α, so if $B^{-1}(\alpha_{2n})b(\alpha_{2n}) = 0$ for infinitely many α then it is identically 0 for all $\alpha \in [0,1]$. (That B^{-1} is rational in α follows from Cramer's rule, and from the fact that $\mathcal{A}$ are rational in α).
(ii) cannot hold from similar arguments since $cB_1^{-1}b$ is rational in α.

The theorem easily follows. ∎

Theorem 2.1 implies:

Corollary 2.1 *For any $\alpha \in (0,1]$ there exists some $0 \le \alpha^- < \alpha$; the value function has a Laurent expansion in the whole interval (α^-, α), or the CMDP is not feasible in this interval. Similarly, for every $\alpha \in [0,1)$ there exists some $\alpha^+ > \alpha$ such that either the value function has a Laurent expansion in the whole interval (α, α^+), or the CMDP is not feasible in this interval.*

References

[1] E. Altman and V. A. Gaitsgory, "Stability and Singular Perturbations in Constrained Markov Decision Problems", *IEEE Trans. Auto. Control*, **38**, No. 6, pp. 971-975, 1993.

[2] E. Altman and A. Shwartz, "Sensitivity of constrained Markov Decision Problems", *Annals of Operations Research*, **32**, pp. 1-22, 1991.

[3] E. Altman and A. Shwartz, "Time-sharing policies for controlled Markov chains", *Operations Research*, **41**, No. 6, pp. 1116-1124, 1993.

[4] F. J. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint", *J. Mathematical Analysis and Applications* **112**, 236-252, 1985.

[5] F. J. Beutler and K. W. Ross, "Time-Average Optimal Constrained Semi-Markov Decision Processes", *Advances of Applied Probability* **18**, No. 2, pp. 341-359, 1986.

[6] T. Bewley and E. Kohlberg, "The asymptotic Theory of Stochastic Games" *Math of Oper. Res.*, **1**, pp. 197-208, 1976.

[7] D. Blackwell, "Discrete dynamic programming", *Ann. Math. Statist.* **33**, pp. 719-726, 1962.

[8] R. Dekker, *Denumerable Markov decision chains: optimal policies for small interest rates*, Thesis, Inst. for Applied Math. and Comp. Science, University of Leiden, 1984.

[9] R. Dekker, A. Hordijk and L. Kallenberg, "Sensitivity-analysis in discounted Markovian decision problems", *OR Spectrum* **7**, pp. 143-151, 1985.

[10] E. A. Feinberg, "Constrained Semi-Markov Decision Processes With Average Rewards", submitted to *ZOR*.

[11] A. Hordijk and L. C. M. Kallenberg, "Constrained undiscounted stochastic dynamic programming", *Mathematics of Operations Research,* **9**, No. 2, May 1984.

[12] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Amsterdam, 1983.

[13] K. Ross and R. Varadarajan, 'Markov Decision Processes with Sample path constraints: the communicating case", *Operations Research*, **37**, No. 5, pp. 780-790, 1989.

[14] K. Ross and R. Varadarajan, "Multichain Markov Decision Processes with a Sample Path Constraint: A Decomposition Approach", *MOR*, Vol. 16 No. 1, pp. 195-207, 1991.

[15] L. I. Sennott, "Constrained discounted Markov decision chains", *Probability in the Engineering and Informational Sciences*, **5**, pp. 463-475, 1991.

[16] L. I. Sennott, "Constrained average cost Markov decision chains", to appear in *Probability in the Engineering and Informational Sciences.*

[17] R. D. Smallhood, "Optimal policy regions for Markov processes with discounting", *Operations Research* **14**, pp. 658-669, 1966.

[18] M. Tidball and A. Altman, "Continuity of optimal values and solutions of convex optimization, and constrained control of Markov chains", submitted to *SIAM*, 1994.