The approach in Markov decision processes revisited

Eitan Altman, Flos Spieksma

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 1569

Programme 1
Architectures parallèles, Bases de données,
Réseaux et Systèmes distribués

THE LINEAR PROGRAM APPROACH IN MARKOV DECISION PROCESSES REVISITED

Eitan ALTMAN
Flos SPIEKSMA

Décembre 1991

RR_1569

Programmation linéaire appliquée aux problèmes de décision markoviens

Eitan Altman
INRIA
Centre Sophia Antipolis
06565 Valbonne Cedex, France

Flos Spieksma
Institute of Applied Mathematics
University of Leiden
Leiden, The Netherlands

26 novembre 1991

Résumé

Il est reconnu que la programmation linéaire est un outil utile pour la résolution de processus de décision markoviens (MDP). Cette démarche est issue de la programmation dynamique, qui elle aussi permet de résoudre des MDP, mais en l'absence de contraintes. Le but de cet article est de comprendre la signification physique des variables de décision intervenant dans les programmes linéaires.

Mots Clés: problèmes de décision Markoviens, programmation linéaire.

The Linear Program approach in Markov Decision Processes revisited

Eitan Altman

INRIA
Centre Sophia Antipolis
06565 Valbonne Cedex, France

Flos Spieksma

Institute of Mathematics & Computer Science
University of Leiden
P.O. Box 9512, 2300RA Leiden, The Netherlands

ABSTRACT

Linear Programming is known to be an important and useful tool for solving Markov Decision Processes (MDP). Its derivation relies on the Dynamic Programming approach, which also serves to solve MDP. However, for constrained Markov Decision Processes the only available methods are based on Linear Programs. The aim of this paper is to investigate some aspects of such Linear Programs. We first present a stochastic interpretation of the decision variables that appear in the Linear Programs available in the Literature. We then show for the multi-constrained Markov Decision Process that the Linear Program suggested in [7] can be obtained from an equivalent unconstrained Lagrange formulation of the control problem. This shows the connection between the Linear Program approach and the Lagrange approach, that was previously used only for the case of a single constraint [3,11,12].

Keywords: Multi-chain Markov Decision Processes, average cost criterion, state-action frequencies, deviation measure, linear programming, lagrange formulation.

INTRODUCTION

Markov Decision Processes (MDP) have been used extensively in the past to model and solve problems in data communications, computer networks, production etc... We consider in this note MDPs where the optimization criteria are of expected time average type, with finite state and action spaces. A basic method for solving such problems has been Dynamic Programming. Derman [5] obtained a Linear Program formulation from the dynamic program and investigated its dual program DLP. Using DLP and another Linear Program he was able to obtain an optimal control policy. Later on, Hordijk and Kallenberg obtained an optimal policy directly from DLP [6]. One advantage of using this new approach was that it could be extended to Constrained Markov Decision Processes, as Hordijk and Kallenberg did in [7], unlike the Dynamic Programming approach.

The first issue of this paper is to present a stochastic meaning to the two different kinds of decision variables that appear in the Linear Programs of [7].

The first kind is related to the expected frequencies of pairs of states and actions (see e.g. [5]). These frequencies are also known as "occupation measure" (see e.g. [4]) or expected empirical probabilities of state-action pairs. The precise definition of these frequencies is given in the next Section. Derman [5] considered the set of all frequencies that are achieved by different policies, and showed that it is equal to the set of achievable decision variables of the first kind. The properties of these frequencies have been studied in [5], [6] and [7].

However, the stochastic meaning of the second type of decision variables was unknown till now, to the best of our knowledge, except for stationary policies, where the decision variables were related to the deviation matrix by Kallenberg [8].

The importance in understanding the stochastic meaning of the decision variables is that it may enable us to obtain a Linear Program for solving MDP's directly, without requiring to go through the dual problem obtained from the Dynamic Program. This may be crucial in case duality problems may arise, such as the duality gap in infinite Linear Programs. Such a direct approach to obtain the Linear Problem is illustrated in [1]. There Altman and Shwartz obtained a Linear Program for solving constrained MDPs with a countable state space. However, they make strong ergodic assumptions (such as the unichain assumption) under which only decision variables of the first kind are needed. In the case of an infinite state space, it seems that understanding the meaning of all decision variables may enable to obtain Linear Programs with a general multichain structure. This is especially important in the constrained case where no alternative methods are known for solving the problem.

The second issue in this paper is to present an alternative derivation of the Linear Program used for solving the constrained Markov Decision problem [7]. Our derivation is based on a Lagrange formulation which generalizes the one used in the case of a single constraint [3] (later extended in [11],[12] to the countable state space). Our new derivation seems to be more natural and straightforward than the previous method [7], where the paradigm was to introduce the LP and then to prove that the optimal policy and optimal value of the control problem are appropriately related to the optimal solution and value of the LP.

After reviewing in Section 2 the Linear Program for solving MDPs, we introduce in Section 3 the biased deviation measure for arbitrary policies and show that it corresponds to the second kind of decision variables in the following sense. The sets of pairs of frequencies and biased deviation measures obtained by any arbitrary policy form a feasible solution for the Linear Program. In Section 4 we then present some properties of the deviation measures. In Section 5 we present the alternative derivation of the LP used for the constrained case.

1. MODEL AND ASSUMPTIONS.

Consider the basic discrete time process $\{X_t\}_{t=1}^{\infty}$, defined on the finite *state space* $\mathbf{X} = 0, 1, ..., N$; an *action* a belongs to the finite *action space* $\mathbf{A}$, and A_t is the action taken at time t. Without loss of generality, we assume that in any state x all actions in $\mathbf{A}$ are available. $H_t := (X_1, A_1, ..., X_t, A_t)$ is the *history* of the process up to time t. If the state at time t is x and action a is applied, then the next state will be y with probability

$$P_{xay} := P(X_{t+1} = y \mid X_t = x; A_t = a) = P(X_{t+1} = y \mid H_{t-1} = h, X_t = x; A_t = a) \qquad (1.1)$$

A policy u in the *policy space* U is described as $u = \{u_1, u_2, ...\}$, where u_t is applied at time epoch t, and $u_{t+1}(\cdot \mid H_t, X_{t+1})$ is a conditional probability measure over $\mathbf{A}$. Given an initial distribution β on $\mathbf{X}$, each policy u induces a probability measure denoted by P_β^u 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_β^u.

A *Markov policy* $u \in U(M)$ is characterized by the dependence of $u_{t+1}(\cdot \mid H_t, X_{t+1})$ on X_{t+1} only; i.e. $u_{t+1}(\cdot \mid H_t, X_{t+1}) = u_{t+1}(\cdot \mid X_{t+1})$. A *stationary policy* $g \in U(S)$ is characterized by a single conditional probability measure $p^g_{\cdot|x}$ over $\mathbf{A}$, so that $p^g_{\mathbf{A}|x} = 1$; under g, X_t becomes a Markov chain with stationary transition probabilities, given by $P^g_{xy} = \sum_{a \in \mathbf{A}} p^g_{a|x} P_{xay}$. The class of

stationary deterministic policies $U(SD)$ is a subclass of $U(S)$, and every $g \in U(SD)$ is characterized by a mapping g:$\mathbf{X} \rightarrow \mathbf{A}$, so that $p^g_{\cdot|x} = \delta_{g(x)}(\cdot)$ is concentrated at the point $g(x)$ for each x.

Let $c : \mathbf{X} \times \mathbf{A} \rightarrow \mathbb{R}$, be a (real valued) cost function and define the average costs associated with a policy u and with an initial distribution β on $\mathbf{X}$:

$$C_\beta(u) = \overline{\lim_{t\to\infty}} \frac{1}{t} E^u_\beta \left[\sum_{s=1}^{t} c(X_s, A_s)\right] \tag{1.2}$$

Denote by OP the problem of finding a policy u that minimizes $C_\beta(u)$ for a given initial distribution β. Let C_β be the optimal value of OP. A policy that achieves $C_\beta(u) = C_\beta$ is said to be optimal for OP. Denote $U(\beta)$ the set of all such policies.

Derman [5] has shown that the cost in (1.2) can be represented in terms of the following *state action frequencies*: let

$\bar{f}^T(\beta, u)$ is the matrix whose elements are given by

$\bar{f}^T(y, a; \beta, u)$:= the frequency of state y and action a till time T when using policy u and initial probability distribution on the state space is $\beta(.)$, i.e.

$$\bar{f}^T(y, a; \beta, u) = \frac{1}{T} \sum_{s=1}^{T} P^u_\beta(X_s = y, A_s = a)$$

Let $\bar{F}(\beta, u)$ be the set of accumulation points of $\bar{f}^T(\beta, u)$. A generic element of $\bar{F}(\beta, u)$ will be denoted by $\bar{f}(\beta, u)$.

Given a class of policies U', define $L_\beta(U') := \cup_{u\in U'} \bar{F}(\beta, u)$. The set of frequencies obtained by all policies is $L_\beta := \cup_{u\in U} \bar{F}(\beta, u)$.

The following notation is used below: $1\{A\}$ is the indicator function of the set A and $\delta_a(x)$ is the Kronecker delta function. We denote by $\overline{B}$ the closure of a set B, and $|B|$ is the cardinality of the set (we shall use this notation for finite sets only, in which case $|B|$ is the number of elements in B).

2. LINEAR PROGRAM FOR SOLVING MDPs

One method of solving OP is based on the solution of a LP which we present below. The importance of this method is lies in the fact that it also enables to deal with optimization problems with additional constraints, where other methods (based on dynamic programming) fail.

Given $\beta \in S(\mathbf{X} \times \mathbf{A})$, define Π_β to be the set of $\{(z,\zeta)\}$, $z,\zeta \in \mathbb{R}^{|\mathbf{X}\times\mathbf{A}|}$, that satisfy

$$\sum_{y\in\mathbf{X}}\sum_{a\in\mathbf{A}}(\delta_y(v) - P_{yav})z(y,a) = 0, \qquad v \in \mathbf{X} \tag{2.1.a}$$

$$\sum_{a\in\mathbf{A}} z(v,a) + \sum_{y\in\mathbf{X}}\sum_{a\in\mathbf{A}}(\delta_y(v) - P_{yav})\zeta(y,a) = \beta(v), \qquad v \in \mathbf{X} \tag{2.1.b}$$

$$z \geq 0 \tag{2.1.c}$$

$$\zeta \geq 0 \tag{2.1.d}$$

Remark: Every $z(\cdot,\cdot) \in \Pi_\beta$ satisfies $\sum_{y,a} z(y,a) = 1$. This can be seen by summing equation (2.1.b) over all $v \in \mathbf{X}$.

Consider the following Linear Program:

LP: Find $z,\zeta \in \mathbb{R}^{|\mathbf{X}\times\mathbf{A}|}$, that minimize $c \cdot z$ subject to $(z,\zeta) \in \Pi_\beta$.

LP is related to OP in the following way (see [6,8]). Given any $(z,\zeta) \in \Pi_\beta$ define the stationary policy $g(z,\zeta)$ by

$$p_{a|y}^{g(z,\zeta)} = \begin{cases} \dfrac{z(y,a)}{\sum_a z(y,a)}, & if \ \sum_a z(y,a) > 0 \\ \dfrac{\zeta(y,a)}{\sum_a \zeta(y,a)}, & if \ \sum_a z(y,a) = 0 \ and \ \sum_a \zeta(y,a) > 0 \\ arbitrary, & otherwise. \end{cases}$$

Lemma 1:

(i) The optimal value of OP and of LP are equal.

(ii) Suppose that (z^*,ζ^*) is any extreme optimal solution of LP. then $g(z^*,\zeta^*)$ is optimal for OP.

The Lemma motivates our interest in studying the properties of LP, as these may relate to properties of Markov Decision Processes. The Lemma also has extensions to constrained Markov

Decision Problems (see [7]). The aim of this paper is to study more about the properties of the decision variables of LP and their relation to quantities that characterize policies.

Denote $\Pi_\beta^z := \{z : (z, \zeta) \in \Pi_\beta\}$.

It is known [6] that the set of policies U can be related to the set Π_β^z through the state-action frequencies, i.e. $\Pi_\beta^z = L$. This means that for any policy u, $\bar{f}(\beta, u) \in \Pi_\beta^z$. On the other hand, for each $z \in \Pi_\beta^z$ there exists a policy u such that $\bar{f}(\beta, u) = z$.

However, little is known about the "physical" meaning of the variable ζ. For $u \in U(S)$ Kallenberg [8 p. 109] gives an explicit expression $\zeta = \zeta(\beta, u)$ such that $(\bar{f}(\beta, u), \zeta(\beta, u)) \in \Pi_\beta$. The reason that the meaning of the relation between the decision variables of LP and the control problem was not clear (for all policies) before is related to the fact that LP was not obtained directly from the control problem, but rather as the dual problem of another LP, related to dynamical programming. In the following sections we present an interpretation of the variables ζ that is related to the control problem.

3. THE DEVIATION MEASURE

With relation to some $\bar{f}(\beta, u) \in \bar{F}(\beta, u)$ we define the biased total occupation matrix $\bar{r}^T(\beta, u)$ whose elements are given by

$$\bar{r}^T(y, a; \beta, u) := \sum_{s=1}^{T} \left[P_\beta^u\{X_s = y, A_s = a\} - \bar{f}(y, a; \beta, u) \right]$$

Define the average biased occupation matrix $\bar{r}^T(\beta, u)$ whose elements are given by

$$\bar{r}^T(y, a; \beta, u) := \frac{1}{T} \sum_{t=1}^{T} \bar{r}^t(y, a; \beta, u).$$

Let t_n be a subsequence of t along which $\bar{f}^t(\beta, u) \to \bar{f}(\beta, u)$. Pick a further subsequence s_n of t_n along which some limit $\tilde{r}(\beta, u) := \lim_{n \to \infty} \bar{r}^{s_n}(\beta, u)$ exists. We call $\tilde{r}(\beta, u)$ the *deviation measure.*

The following Lemma relates the quantities $\bar{f}(\beta, u)$ and $\tilde{r}(\beta, u)$ to the decision variables in LP, or more precisely to the variables that determine the set Π_β. By appropriately adding a bias factor, we then obtain in Theorem 3 a stronger characterization.

Lemma 2:

Under any policy $u \in U$ and initial distribution β, if the elements of $\tilde{r}(\beta, u)$ are finite then the tuple $(\bar{f}(\beta, u), \tilde{r}(\beta, u))$ satisfies equations (2.1.a), (2.1.b) and (2.1.c).

Proof: We shall use the following:

$$P\{X_t = y\} = E[P\{X_t = y|H_{t-1}\}] = \sum_{a,z} P\{X_{t-1} = z, A_{t-1} = a\}P_{zay} \tag{3.1}$$

By averaging we obtain:

$$\frac{1}{t}\sum_{s=2}^{t} P_\beta^u\{X_s = y\} = \frac{1}{t}\sum_{s=2}^{t}\sum_{a,z} P_\beta^u\{X_{s-1} = z, A_{s-1} = a\}P_{zay} \tag{3.2}$$

Interchanging the order of summation in the right side of (3.2), we get

$$\sum_{y,a} \bar{f}(y,a;\beta,u)\,[\delta_y(z) - P_{zay}] = 0 \tag{3.3}$$

for any accumulation point $\bar{f}(\beta,u)$, which proves that (2.1.a) is satisfied. Moreover, we clearly have

$$\bar{f}(y,a;\beta,u) \geq 0 \tag{3.4}$$

This proves that (2.1.c) is satisfied.

Using (3.1) and (3.3) we get

$$\sum_{s=2}^{t}\sum_{a}\left[P_\beta^u\{X_s = y, A_s = a\} - \bar{f}(y,a;\beta,u)\right] =$$

$$\sum_{s=2}^{t}\sum_{z,a}\left[\left(P_\beta^u\{X_{s-1} = z, A_{s-1} = a\} - \bar{f}(y,a;\beta,u)\right)P_{zay}\right] \tag{3.5}$$

We then obtain by summation:

$$\sum_{a}\bar{r}^t(y,a;\beta,u) + \sum_{a}\bar{f}(y,a;\beta,u) - \beta(y) = \sum_{z,a}\bar{r}^{t-1}(z,a;\beta,u)P_{zay} \tag{3.6}$$

Taking the time average of (3.6) we obtain:

$$\sum_{a}\frac{1}{s-1}\sum_{t=2}^{s}\bar{r}^t(y,a;\beta,u) + \sum_{a}\bar{f}(y,a;\beta,u) - \beta(y)$$

$$= \sum_{z,a} P_{zay}\frac{1}{s-1}\sum_{t=2}^{s}\bar{r}^{t-1}(z,a;\beta,u) \tag{3.7}$$

Equation (3.7) can then be rewritten as:

$$\sum_a \left(\frac{s}{s-1}\bar{r}^s(y,a;\beta,u)\right) - \frac{1}{s-1}\bar{r}^1(y,a;\beta,u) + \sum_a \bar{f}(y,a;\beta,u) - \beta(y)$$

$$= \sum_{z,a} \tilde{r}^{s-1}(z,a;\beta,u)P_{zay} \tag{3.8}$$

Note that the term $\frac{1}{s-1}\bar{r}^1(y,a;\beta,u)$ converges to zero as $s \to \infty$. With the sequence s_n defined above Lemma 2, note that we have

$$\lim_{n\to\infty} \tilde{r}^{s_n-1}(\beta,u) = \tilde{r}(\beta,u) \tag{3.9}$$

which follows from the fact that

$$\frac{t}{t-1}\bar{r}^t(y,a;\beta,u) - \tilde{r}^{t-1}(y,a;\beta,u) = \frac{1}{t-1}\bar{r}^t(y,a;\beta,u) =$$

$$\frac{t}{t-1}\left[\bar{f}^t(y,a;\beta,u) - \bar{f}(y,a;\beta,u)\right] \to 0 \tag{3.10}$$

as $t \to \infty$, along the subsequence s_n. We thus obtain from (3.8)

$$\sum_a \bar{f}(y,a;\beta,u) + \sum_{z,a} \tilde{r}(z,a;\beta,u)\left[\delta_y(z) - P_{zay}\right] = \beta(y) \tag{3.11}$$

and hence also (2.1.b) is satisfied. Note that (3.11) is obtained for any subsequence s_n of t_n. ■

Remark: If the limit $\bar{r}(\beta,u) := \lim_{t\to\infty} r^t(\beta,u)$ exists then (3.11) can be obtained directly from (3.6) with $\bar{r}(\beta,u)$ replacing $\tilde{r}(\beta,u)$.

Consequently, for any such policy u, $(\bar{f}(\beta,u), \bar{r}(\beta,u))$ satisfies (2.1.a), (2.1.b) and (2.1.c).

Note that the $\tilde{r}$ need not be positive and hence equation (2.1.d) is not satisfied. In fact, for any $u \in U$ and initial distribution β, $\sum_{y,a} \tilde{r}(y,a;\beta,u) = 0$. This follows from the fact that

$$\sum_{y,a} \tilde{r}(y,a;\beta,u) = \sum_{a,y} \lim_{n\to\infty} \frac{1}{s_n} \sum_{s=1}^{s_n} \sum_{t=1}^{s} \left[P_\beta^u(X_t = y, A_t = a) - \sum_a \bar{f}(y,a;\beta,u)\right] =$$

$$\lim_{n\to\infty} \sum_{a,y} \frac{1}{s_n} \sum_{s=1}^{s_n} \sum_{t=1}^{s} \left[P_\beta^u(X_t = y, A_t = a) - \sum_a \bar{f}(y,a;\beta,u)\right] =$$

$$= \lim_{n\to\infty} \frac{1}{s_n} \sum_{s=1}^{s_n} \frac{1}{s_n} \sum_{t=1}^{s} [1-1] = 0$$

Hence in order that all the $\bar{r}$ be non-negative, so that we have in fact $(\bar{f}(\beta,u), \bar{r}(\beta,u)) \in \Pi_\beta$, another bias factor should be added (see Theorem 3 below).

Let $T \subset \mathbf{X} \times \mathbf{A}$ be the set of pairs $\{y,a\}$ satisfying $\bar{f}(y,a;\beta,u) = 0$. Note that for any $(y,a) \in T$, $\bar{r}(y,a;\beta,u) \geq 0$. Let γ be any real number that satisfies

$$\gamma \geq \min_{k\in\mathbf{R}} \left\{ \bar{r}(y,a;\beta,u) + k\bar{f}(y,a;\beta,u) \geq 0 \text{ for all } y,a \right\}$$

Define:

$$Y(y,a;\beta,u) := \begin{cases} \bar{r}(y,a;\beta,u), & (y,a) \in T \\ \bar{r}(y,a;\beta,u) + \gamma \bar{f}(y,a;\beta,u), & (y,a) \notin T \end{cases}$$

Clearly $Y(y,a;\beta,u)$ are nonnegative for all state-action pairs. Hence $Y(\beta,u)$ satisfies equation (2.1.d). We thus obtain:

Theorem 3: *Under any policy $u \in U$ and initial distribution β, if the elements of $\bar{r}(\beta,u)$ are finite then $(\bar{f}(\beta,u), Y(\beta,u)) \in \Pi_\beta$.*

Proof: The proof that equations (2.1.a) and (2.1.c) are satisfied remains unchanged. Equation (2.1.c) holds by (2.1.a) and (3.11). ■

An open problem is whether for any solution (z,ζ) of LP we can construct a policy u, such that $(z,\zeta) = (\bar{f}(\beta,u), Y(\beta,u))$. A step towards the answer is presented in Theorem 4, where we give a characterization of the polytope Π_β. To do so, we need some notation first.

For sake of convenience we will write $z(y)$, $\zeta(y)$ and $\bar{f}(y;\beta,u)$ for $\sum_a z(y,a)$, $\sum_a \zeta(y,a)$ and $\sum_a \bar{f}(y,a;\beta,u)$. From the context it will always be clear, whether these quantities have to be interpreted as vectors or as matrices.

For any given policy $g \in U(S)$ we denote by $\nu(g)$ the number of closed classes in the Markov chain generated by policy g. We write $B(g)$ for a subset of X that contains precisely one state from each closed class. Clearly $|B(g)| = \nu(g)$. $T(g)$ denotes the set of transient states and $\mathcal{R}_l(g)$, $l = 1,\ldots,\nu(g)$ the closed classes.

$$_M P^g_{\delta_x}\{X_t = y\} := \begin{cases} P^g_{\delta_x}\{X_t = y, X_2, \ldots, X_{t-1} \notin M\}, & t \geq 2 \\ \delta_x(y), & t = 1 \end{cases}$$

stands for the taboo transition probabilities with taboo set M. The probabilities $F_M(\delta_x; g)$ of reaching set M, when the chain initially starts in state x, are equal to

$$F_M(\delta_x; g) = P^g_{\delta_x}\{X_2 \in M\} + \sum_{t\geq 3} P^g_{\delta_x}\{X_t \in M, X_2, \ldots, X_{t-1} \notin M\}$$

$$= \sum_{m\in M}\sum_{t\geq 1}\sum_{y\in \mathbf{X}} {}_M P^g_{\delta_x}\{X_t = y\} P^g_{ym}.$$

Finally we have to introduce the deviation matrix, the entries of which are given by

$$D_{\delta_x}(y, g) = \lim_{\alpha\uparrow 1} \sum_{t=1}^{\infty} \alpha^{t-1}\big(P^g_{\delta_x}\{X_t = y\} - \bar{f}(y; \delta_x, g)\big).$$

This can be rewritten as (cf. Spieksma [13] p. 82)

$$D_{\delta_x}(y, g) = \sum_{v\in \mathbf{X}} \big(\delta_x(v) - \bar{f}(v; \delta_x, g)\big) \sum_{t=1}^{\infty}\sum_{w\in \mathbf{X}} {}_{B(g)} P^g_{\delta_v}\{X_t = w\}\big(\delta_w(y) - \bar{f}(y; \delta_w, g)\big). \tag{3.12}$$

Theorem 4: $(z, \zeta) \in \Pi_\beta \iff$ *there are* g_1, $g_2 \in U(S)$ *and an initial distribution* $\tilde{\beta}$, *such that* z, ζ *as vectors on* $\mathbf{X}$ *satisfy (3.13) and* $z(y, a) = z(y)P^{g_1}_{a|y}$, $\zeta(y, a) = \zeta(y)P^{g_2}_{a|y}$, $y \in X$.

$$\begin{cases} z(y) = \bar{f}(y; \tilde{\beta}, g_1), \ \forall\, y & (3.13a) \\ \zeta(y) = \sum_x \Big[\beta(x)D(y; \delta_x, g_2) + \big(\bar{f}(x; \beta, g_2) - z(x)\big)\sum_{t\geq 1} {}_{B(g_2)} P^{g_2}_{\delta_x}\{X_t = y\} + \gamma(x)\bar{f}(y; \delta_x, g_2)\Big] & (3.13b) \\ \sum_x \big(\beta(x) - z(x)\big) F_{\mathcal{R}_l(g_2)}(\delta_x, g_2) = 0, \ l = 1, \ldots, \nu(g_2) & (3.13c) \\ \sum_x \big(\beta(x) - z(x)\big) \sum_{t\geq 1} {}_{B(g_2)} P^{g_2}_{\delta_x}\{X_t = y\} \geq 0, \ \ \forall\, y \in \mathcal{T}(g_2), & (3.13d) \end{cases}$$

where γ *is a vector on* X *with*

$$\gamma(y) = \begin{cases} 0, y \in \mathcal{T}(g_2) \\ c_l, y \in \mathcal{R}_l(g_2), \end{cases}$$

with c_l *a constant satisfying*

$$c_l \geq \max_{y\in \mathcal{R}_l(g_2)} \frac{-\sum_x \Big[\beta(x)D(y; \delta_x, g_2) + \big(\bar{f}(x; \beta, g_2) - z(x)\big)\sum_{t\geq 1} {}_{B(g_2)} P^{g_2}_{\delta_x}\{X_t = y\}\Big]}{\sum_{x\in \mathcal{R}_l(g_2)} \bar{f}(y; \delta_x, g_2)}, l = 1, \ldots, \nu(g_2).$$

Notice, that the vector γ is constant on each positive recurrent class in the Markov chain induced by g_2. Equations (3.13c) and (3.13d) are necessary to ensure that we only get positive solutions ζ to system (3.13).

If $\tilde{\beta} = \beta$ and $g_1 = g_2$, then $\zeta(y) = \sum_x \Big[\beta(x)D(y;\delta_x,g_2) + \gamma(x)\bar{f}(y;\delta_x,g_2)\Big]$. This is in fact the formula used by Kallenberg to show, that we can construct a feasible solution (x,ζ) of Π_β for any stationary policy.

Proof of Theorem 4.

First choose $(z,\zeta) \in \Pi_\beta$. By the construction in [8] Π_β is not empty. Define two stationary policies g_1, g_2 in the following way:

$$p^{g_1}_{a|y} := \begin{cases} \dfrac{z(y,a)}{z(y)}, \text{ if } z(y) > 0 \\ \dfrac{\zeta(y,a)}{\zeta(y)}, \text{ if } z(y) = 0, \zeta(y) > 0 \\ \text{arbitrarily ,otherwise,} \end{cases} \qquad p^{g_2}_{a|y} := \begin{cases} \dfrac{\zeta(y,a)}{\zeta(y)}, \text{ if } \zeta(y) > 0 \\ \dfrac{z(y,a)}{z(y)}, \text{ if } \zeta(y) = 0, z(y) > 0 \\ \text{arbitrarily ,otherwise.} \end{cases}$$

So, g_1 and g_2 only differ in states y with $z(y), \zeta(y) > 0$ and $z(y,a)/z(y) \neq \zeta(y,a)/\zeta(y)$, or possibly in states y with $z(y) = \zeta(y) = 0$.

By the construction of g_1 and g_2 the vectors z and ζ satisfy the following linear equations.

$$z(y) - \sum_x z(x)P^{g_1}_{xy} = 0, \; y \in \mathbf{X} \tag{3.14a}$$

$$z(y) + \zeta(y) - \sum_x \zeta(x)P^{g_2}_{xy} = \beta(y), \; y \in \mathbf{X}. \tag{3.14b}$$

Using (3.14b) it is easily checked that $\sum_{y,a} z(y,a) = 1$. Consequently z is an invariant probability measure for the Markov chain generated by policy g_1. So, $z = \bar{f}(\tilde{\beta}, g_1)$ for some initial distribution $\tilde{\beta}$.

Notice, that for any vector η on $\mathbf{X}$, z, $\tilde{\zeta}(\bullet) := \zeta(\bullet) + \sum_x \eta(x)\bar{f}(\bullet;\delta_x,g_2))$ are a solution of (3.14) as well. Let $B(g_2)$ be a set of reference states in the Markov chain generated by g_2. We write $b_l := \mathcal{R}_l(g_2) \cap B(g_2)$, $l = 1,\ldots,\nu(g_2)$. As the vector η we choose $\eta(y) := 0$ for $y \in \mathcal{T}(g_2)$, and $\eta(y) := -\zeta(b_l)/\sum_{x\in\mathcal{R}_l(g_2)} \bar{f}(b_l;\delta_x,g_2)$ for $y \in \mathcal{R}_l(g_2)$. Thus, we have chosen η to be constant on each positive recurrent class and 0 otherwise, in such a way that $\tilde{\zeta}(b_l) = 0$. Note, that $\eta(x)\bar{f}(y;\delta_x,g_2) = 0$ for $y \in \mathcal{R}_l(g_2)$, $x \notin \mathcal{R}_l(g_2)$. Plugging $\tilde{\zeta}$ into equation (3.14b) we get

$$\tilde{\zeta}(y) = \beta(y) - z(y) + \sum_{x\notin B(g_2)} \tilde{\zeta}(x)P^{g_2}_{xy}.$$

Iterating this we obtain

$$\tilde{\zeta}(y) = \beta(y) - z(y) + \sum_{x \notin B(g_2)} \Big[\beta(x) - z(x) + \sum_{v \notin B(g_2)} \tilde{\zeta}(v) P^{g_2}_{vx}\Big] P^{g_2}_{xy}$$

$$= \beta(y) - z(y) + \sum_{x \notin B(g_2)} (\beta(x) - z(x)) P^{g_2}_{xy}$$

$$+ \sum_{v \notin B(g_2)} \tilde{\zeta}(v) P^{g_2}_{\delta_v}\{X_3 = y, X_2 \notin B(g_2)\}$$

$$= \beta(y) - z(y) + \sum_{x \notin B(g_2)} (\beta(x) - z(x)) \sum_{t=2}^{T} P^{g_2}_{\delta_x}\{X_t = y, X_2, \ldots, X_{t-1} \notin B(g_2)\}$$

$$+ \sum_{v \notin B(g_2)} \tilde{\zeta}(v) P^{g_2}_{\delta_v}\{X_{T+1} = y, X_2, \ldots, X_T \notin B(g_2)\}.$$

Set $B(g_2)$ is reached with probability 1 from any initial state, when policy g_2 is used. So, taking the limit for $T \to \infty$ in the last equation yields

$$\tilde{\zeta}(y) = \beta(y) - z(y) + \sum_{x \notin B(g_2)} (\beta(x) - z(x)) \sum_{t=2}^{\infty} P^{g_2}_{\delta_x}\{X_t = y, X_2, \ldots, X_{t-1} \notin B(g_2)\}. \tag{3.15}$$

For $y \notin B(g_2)$ the probabilities in the last expression equal the taboo transition probabilities for the taboo set $B(g_2)$. For $y \in B(g_2)$ they equal the first hitting probabilities of the states contained in set $B(g_2)$. As a consequence we obtain the following expressions for $\tilde{\zeta}$.

$$\tilde{\zeta}(y) = \begin{cases} \displaystyle\sum_{x \notin B(g_2)} (\beta(x) - z(x)) \sum_{t=1}^{\infty} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = y\},\ y \notin B(g_2) & (3.16a) \\ \displaystyle\sum_{x} (\beta(x) - z(x)) F_y(\delta_x, g_2),\ y \in B(g_2). & (3.16b) \end{cases}$$

As $F_{b_l}(\delta_x, g_2) = F_{\mathcal{R}_l(g_2)}(\delta_x, g_2)$ for $x \in \mathrm{X}$, $F_{b_l}(\delta_x, g_2) = 1$ for $x \in \mathcal{R}_l(g_2)$ and $\tilde{\zeta}(b_l) = 0$ (by definition), (3.13c) follows directly from (3.16b).

Obviously (3.16a) equals 0 for $y \in B(g_2)$. Consequently (3.16a) holds for *all* states. Thus we

get for $y \in \mathcal{R}_l(g_2)$:

$$\begin{aligned}
\tilde{\zeta}(y) &= \sum_{x\in\mathbf{X}} \big(\beta(x) - z(x)\big) \sum_{t=1}^{\infty} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = y\} - \big(\beta(b_l) - z(b_l)\big) \sum_{t=1}^{\infty} {}_{B(g_2)}P^{g_2}_{\delta_{b_l}}\{X_t = y\} \\
&= \sum_{x\in\mathbf{X}} \big(\beta(x) - z(x)\big) \sum_{t=1}^{\infty} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = y\} - \big(\beta(b_l) - z(b_l)\big) \frac{\bar{f}(y;\delta_{b_l},g_2)}{\bar{f}(b_l;\delta_{b_l},g_2)} \\
&= \sum_{v\in\mathbf{X}} \beta(v) \sum_{x} \big(\delta_x(v) - \bar{f}(x;\delta_v,g_2)\big) \sum_{t} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = y\} \\
&\quad + \sum_{x\in\mathbf{X}} \big(\bar{f}(x;\beta,g_2) - z(x)\big) \sum_{t} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = y\} - \frac{\beta(b_l) - z(b_l)}{\bar{f}(b_l;\delta_{b_l},g_2)} \bar{f}(y;\delta_{b_l},g_2) \\
&= \sum_{x\in\mathbf{X}} \beta(x) D(y;\delta_x,g_2) + \sum_{x\in\mathbf{X}} \big(\bar{f}(x;\beta,g_2) - z(x)\big) \sum_{t} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = y\} + \\
&\quad \Big\{ \sum_{v\in\mathbf{X}} \beta(v) \sum_{x\in\mathbf{X}} \big(\delta_x(v) - \bar{f}(x;\delta_v,g_2)\big) \sum_{t} \sum_{w\in\mathbf{X}} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = w\} F_{\mathcal{R}_l(g_2)}(\delta_w;g_2) \\
&\qquad - \frac{\beta(b_l) - z(b_l)}{\bar{f}(b_l;\delta_{b_l},g_2)} \Big\} \bar{f}(y;\delta_{b_l},g_2).
\end{aligned} \tag{3.17}$$

For the second equality we use the renewal theorem, if $y \neq b_l$, and if $y = b_l$ we use ${}_{B(g_2)}P^{g_2}_{\delta_{b_l}}\{X_2 = y\} = 1 = \bar{f}(y;\delta_{b_l},g_2)/\bar{f}(b_l,\delta_{b_l},g_2)$. For the 4th equality we use formula (3.12) for the deviation matrix. Obviously, the same expressions hold for $y \in \mathcal{T}(g_2)$, since transient states cannot be reached from any positive recurrent state, so that only expressions equalling 0 are added in the first equality. The derivation of the other equalities is similar to the foregoing ones.

Recall that $\zeta(y) = \tilde{\zeta}(y) - \sum_{x\in\mathbf{X}} \eta(x)\bar{f}(y;\delta_x,g_2)$. So, we choose the vector γ equal to

$$\gamma(y) := \begin{cases} -\eta(y) + \dfrac{1}{|\mathcal{R}_l(g_2)|} \Big\{ \displaystyle\sum_{v\in\mathbf{X}} \beta(v) \sum_{x\in\mathbf{X}} \big(\delta_x(v) - \bar{f}(x;\delta_v,g_2)\big) \times \\ \qquad \times \displaystyle\sum_{t} \sum_{w\in\mathbf{X}} {}_{B(g_2)}P^{g_2}_{\delta_x}\{X_t = w\} F_{\mathcal{R}_l(g_2)}(\delta_w;g_2) - \dfrac{\beta(b_l) - z(b_l)}{\bar{f}(b_l;\delta_{b_l},g_2)} \Big\}, & y \in \mathcal{R}_l(g_2) \\ -\eta(y) = 0, & y \in \mathcal{T}(g_2). \end{cases}$$

Then by virtue of (3.17) and our choice of η, z satisfies (3.13b) for this choice of γ. Moreover, $\gamma(y)$ is constant on each positive recurrent class. As $\zeta(y) \geq 0$ by assumption, we have obtained that γ satisfies the conditions of the Theorem. For transient states y, $\tilde{\zeta}(y) = \zeta(y) \geq 0$. Using expression (3.16a) for $\tilde{\zeta}(y)$ gives (3.13d).

The proof of the reverse implication is straightforward. Here we have to use, that $\sum_v D(v;\delta_x,g_2)\cdot(\delta_y(v)-P^{g_2}_{vy})=\delta_y(x)-\bar{f}(y;\delta_x,g_2)$ (cf. Kallenberg [8, p. 29], Spieksma [13, p. 68]). ■

4. PROPERTIES OF THE DEVIATION MEASURE

4.1. Bounds

Unlike the state-action frequencies, which sum to one under any arbitrary policy, the deviation measures are not bounded in the class of policies. To see that, assume that there exists a state v that is recurrent under all the stationary deterministic policies. Consider the following set of policies $\{u(t)\}_{t=1}^{\infty}$: let $u(t)$ be a policy that follows one stationary policy till time t and then switches to another one with a different stationary probability distribution. It follows that $|\bar{r}^t(y,a;\beta,u(t))|$ grows like $O(t)$ and is thus unbounded.

4.2. Uniqueness of the deviation measure

Lemma 4: *Assume that under every deterministic stationary policy there exists one recurrent class (plus possibly a transient set). Let u and v be two policies that achieve along some subsequence of t:*

$$\bar{f}(y,a;\beta,v)=\bar{f}(y,a;\beta,u)$$

for all y,a. Assume moreover that along that subsequence we have

$$\frac{\bar{r}(y,a;\beta,v)}{\sum_a \bar{r}(y,a;\beta,v)}=\frac{\bar{r}(y,a;\beta,u)}{\sum_a \bar{r}(y,a;\beta,u)}=:\gamma_y^a$$

Then $\bar{r}(y,a;\beta,u)=\bar{r}(y,a;\beta,v)$.

Proof: Let P^* be the matrix whose components are given by: $P^*_{xy}=\sum_a \gamma_y^a P_{xay}$. In the previous Section we showed that

$$0=\sum_a \bar{f}(y,a;\beta,u)+\sum_{z,a}\bar{r}(z,a;\beta,u)\left[\delta_y(z)-P_{zay}\right]-\beta(y)$$

which can be restated in the following form:

$$0=\bar{f}(y;\beta,u)+\sum_z \bar{r}(z;\beta,u)\left[\delta_y(z)-P^*_{zy}\right]-\beta(y)$$

(where $\bar{r}(z;\beta,u):=\sum_a \bar{r}(z,a;\beta,u)$). Since the same holds for v too, we obtain by subtraction:

$$\sum_z[\bar{r}(z;\beta,u)-\bar{r}(z;\beta,v)]\left[\delta_y(z)-P^*_{zy}\right]=0$$

Hence $[\tilde{r}(\beta,u)-\tilde{r}(\beta,v)]$ is an eigenvector of the matrix $\{P^*_{zy}\}$ which corresponds to the eigenvalue 1. It follows from the Perron - Frobenius Theorem (see e.g. [9 p. 3]) that there exists some constant c such that $\tilde{r}(z;\beta,u)-\tilde{r}(z;\beta,v)=c\pi^*(z)$ where π^* is the vector of stationary distribution that corresponds to P^*. But since $\tilde{r}(z;\beta,u)$ sum to 0, it follows that $c=0$ from which the Lemma follows. ■

5. LINEAR PROGRAM FOR CONSTRAINED MDPs

In this Section we study the Linear Program method applied for the constrained Markov Decision Process [7]. An alternative method based on a Lagrange approach for solving such problems in the case of a single constraint was introduced in [3] (later extended to countable state space in [11],[12]).

The aim of this Section is to generalize the formulation of the Lagrange method to several constraints and then to use it for obtaining a new derivation of the Linear Program method of Hordijk and Kallenberg [7]. Our new derivation seems to be more natural and straightforward than the previous method [7], where the paradigm was to introduce the LP and then to prove that the optimal policy and optimal value of the control problem are appropriately related to the optimal solution and value of the LP.

To define the constrained problem, we define a set of K (real valued) cost functions d^k : $\mathbf{X}\times\mathbf{A}\to\mathcal{R}$, $k=1,...,K$, and define the average costs associated with a policy u and with an initial distribution β on $\mathbf{X}$:

$$D^k_\beta(u)=\overline{\lim_{t\to\infty}}\frac{1}{t}E^u_\beta\left[\sum_{s=1}^{t}d^k(X_s,A_s)\right]$$

For a given vector of real numbers V_k, $k=1,...,K$ we define the constrained problem **COP**: find a policy $u\in U$ that minimizes $C_\beta(u)$ subject to $D^k_\beta(u)\le V_k$, $k=1,...,K$, for a given initial distribution β.

A policy that satisfies $D^k_\beta(u)\le V_k$, $k=1,...,K$ is said to be feasible. Let C_β be the optimal value of COP. A feasible policy u that achieves $C_\beta(u)=C_\beta$, is said to be optimal for COP.

Introduce the Lagrange minimax problem $\mathbf{L_{minimax}}$:

$$\min_u\max_{\nu\le 0}\left[C_\beta(u)-\sum_k\nu_k(D^k_\beta(u)-V_k)\right]$$

Let $C_\beta^{minimax}$ be the optimal value of $L_{minimax}$. It is easily seen that $L_{minimax}$ is equivalent to COP (in the sense that the optimal policy and optimal value are the same).

Next we introduce the Linear Program used for calculating the optimal value and policy for COP [7]:

LPC: Find $z, \zeta \in \mathcal{R}^{|\mathbf{X}\times\mathbf{A}|}$, that minimize $c \cdot z$ subject to

$$(z, \zeta) \in \Pi_\beta,$$

$$d^k \cdot z \leq V_k$$

where Π_β is given in Section 2. It is shown in [7] that the optimal value of COP is equal to the optimal value of LPC, and an optimal Markovian policy is constructed from the optimal solution of LPC.

We show below that LPC can be derived from $L_{minimax}$. We first note that we may restrict the minimization in COP to the class of policies U^* under which $\bar{F}(\beta, u)$ is a singleton, see e.g. [7]. Note that this class contains the stationary policies.

Define $\mathbf{L_{maximin}}$:

$$\max_{\nu \leq 0} \min_{u \in U^*} \left[C_\beta(u) - \sum_k \nu_k (D_\beta^k(u) - V_k) \right]$$

Let $C_\beta^{maximin}$ be the optimal value of $L_{minimax}$. It can be shown that the min and max can be interchanged, so that $C_\beta^{maximin} = C_\beta^{minimax}$. We leave the proof to the appendix. In order to obtain LPC we first consider the problem of $\min_{u \in U^*} \left[C_\beta(u) - \sum_k \nu_k (D_\beta^k(u) - V_k) \right]$ for a given ν. It follows from [6] that an optimal stationary deterministic policy for that problem can be obtained and the optimal value is given by **LP'**: find $\lambda_y, g_y,\ y \in \mathbf{X}$, that

$$\max_{\lambda \leq 0} \left[\sum_y \beta(y) g_y + \sum_k \nu_k V_k \right] \tag{5.1}$$

s.t.

$$\lambda_y + g_y \leq c(y, a) - \sum_k \nu_k d^k(y, a) + \sum_v P_{yav} \lambda_v \qquad \forall y, a \tag{5.2}$$

$$g_y \leq \sum_{v \in \mathbf{X}} P_{yav} g_v \qquad \forall y, a \tag{5.3}$$

We may conclude that the optimal value of COP is given by solving the LP: find $\nu_k, k = 1, ..., K$, λ_y, and g_y, $y \in X$, that

$$\max_{\lambda,\nu \le 0} \left[\sum_y \beta(y) g_y + \sum_k \nu_k V_k \right]$$

s.t.

$$\lambda_y + g_y \le c(y,a) - \sum_k \nu_k d^k(y,a) + \sum_v P_{yav} \lambda_v \qquad \forall y, a$$

$$g_y \le \sum_{v \in \mathbf{X}} P_{yav} g_v \qquad \forall y, a$$

The dual of this LP is LPC.

Remark: The results of this Section extend easily to the discounted criteria, even with countable state space. They also extends easily to the countable case under a strong ergodic assumption (the unichain assumption). In both cases, there is no need of equation (5.3), and of (2.1.b) and (2.1.d). Thus we need not consider the decision variables ζ in the Linear Program (2.1). (See e.g. [1],[2],[13] for alternative derivations of such LPs).

APPENDIX

Interchanging the min and max in Section 5.

We show in this appendix that $C_\beta^{minimax} = C_\beta^{maximin}$. Since we restrict to $u \in U^*$ it easily follows that

$$C_\beta(u) - \sum_k \nu_k (D_\beta^k(u) - V_k) = \sum_{y,a} \bar{f}(y,a;\beta,u) \left[c(y,a) - \sum_k \nu_k (c(y,a) - V_k) \right]$$

and hence

$$\min_u \max_{\nu \le 0} \left[C_\beta(u) - \sum_k \nu_k (D_\beta^k(u) - V_k) \right] = \min_{z \in L_\beta(U^*)} \max_{\nu \le 0} \sum_{y,a} \bar{f}(y,a;\beta,u) \left[c(y,a) - \sum_k \nu_k (c(y,a) - V_k) \right]$$

Since $L_\beta = L_\beta(U^*)$ is convex and compact (this is a straightforward generalization of [5] p. 93-94, [7] who prove it for β of the form of δ_x), and since the set $\nu \le 0$ is convex and compact in the

compactified space $[\mathcal{R} \cup \{\infty\}]^K$, it follows from the minimax Theorem (e.g. [10] p. 208) that the min and max can be interchanged from which it finally follows that $C_\beta^{minimax} = C_\beta^{maximin}$.

REFERENCES

[1] E. Altman and A. Shwartz, "Markov decision problems and state-action frequencies", EE. PUB. No. 693, Technion, November 1988, to appear in *SIAM J. Control and Optimization.*

[2] E. Altman, "Denumerable Constrained Markov Decision Problems and Finite Approximations", 1991, submitted to *Math. of O.R.*

[3] F. J. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint", *Math. Anal. Appl.* Vol. 112, pp. 236-252, 1985.

[4] V. S. Borkar, "A convex analytic approach to Markov decision processes", *Probab. Th. Rel. Fields*, Vol. 78, pp. 583-602, 1988.

[5] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, 1970.

[6] A. Hordijk and L. C. M. Kallenberg, "Linear programming and Markov decision chains", *Management Science*, Vol. 25 no. 4 pp. 352-362, April 1979.

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

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

[10] D. G. Luenberger, *Optimization by vector space methods*, John Wiley, 1968.

[9] E. Seneta, *Non-negative Matrices and Markov Chains*, Springer-Verlag, 1981.

[11] L. I. Sennott "Constrained Discounted Markov Decision Chains", Nov. 1990, preprint.

[12] L. I. Sennott "Constrained Average Decision Chains", Dec. 1990, preprint.

[13] F. Spieksma, *Geometrically ergodic Markov chains and the optimal control of queues*, Ph.D. thesis, Leiden 1990.

ISSN 0249-6399