Mathematical Methods in Operations Research 2005
Special issue in honor of *Arie Hordijk*

Remarks on old results

Lodewijk Kallenberg

Mathematical Institute, University of Leiden, P.O. Box 9512, 2300 RA Leiden, The Netherlands

Received: January 2005

Abstract Linear programming can be used for the solution of Markov decision problems (MDPs), both for the discounted and for the average reward criterion. For the average reward criterion, there are essential differences in the solution of irreducible, unichained and multichained MDPs. The basic results for irreducible MDPs were proven already in 1960. In 1968, Denardo and Fox obtained essential results for the multichain case and simplified algorithms for the unichain case. Finally, in 1979, Hordijk and this author have shown a theorem for the computation of optimal policies in multichained MDPs. In the present paper several examples are given to illustrate the differences in handling irreducible, unichained and multichained MDPs.

1 Introduction

When Hordijk was appointed at Leiden University in 1976, I became his first PhD student. Looking for a PhD project Hordijk suggested linear programming (LP) methods for the solution of MDPs. LP for MDPs was introduced by D'Epenoux ([3]) for the discounted case. De Ghellinck ([1]) as well as Manne ([6]) obtained LP formulations for the average reward criterion in the irreducible case. The first analysis of LP for the multichain case was given by Denardo and Fox ([2]). Our interest was raised by Derman's remark ([4] p. 84): "No satisfactory treatment of the dual program for the multiple class case has been published".

We started to work on this subject. We succeeded in proving a theorem, which shows how to obtain an optimal policy from the dual program ([5]).

The main difference in the linear programming approach between on one the side discounted and average irreducible MDPs and on the other side unichained and multichained MDPs is the fact that the one-to-one cor-

respondence between stationary policies and feasible solutions of the linear program does not hold in the last cases.

In this paper we give some examples for the following properties in the multichain case:

1. An extreme optimal solution of the dual program has in some state more than one positive variable.
2. An extreme feasible solution of the dual program is mapped on a non-deterministic stationary policy.
3. Two different solutions are mapped on the same deterministic and stationary policy.
4. An non-optimal solution of the dual program is mapped on an optimal deterministic and stationary policy.
5. The results of the unichain case cannot be generalized to the general single chain case.

In Section 2 the notation of the MDP model and some results of [5] are given. Section 3 presents the examples.

2 Notation and properties

Let S be the finite *state space* and $A(i)$ the finite *action set* in state $i \in S$. If in state i action $a \in A(i)$ is chosen, the a *reward* $r_i(a)$ is earned and $p_{ij}(a)$ is the *transition probability* that the next state is state j.

Given starting state i and policy R, the average expected reward is denoted by $\phi_i(R)$. The *value-vector* ϕ is defined by $\phi_i = sup_R\, \phi_i(R),\ i \in S$. Policy R^* is an *average optimal policy* if $\phi_i(R^*) = \phi_i,\ i \in S$.

An MDP is called *irreducible* if all states belong to a single ergodic class under all policies; *unichained* if any policy produces a single ergodic class plus a (perhaps empty) policy-dependent set of transient states; *multichained* if there may be several ergodic classes and some transient states, and these classes may vary from policy to policy.

The primal and dual linear programs for multichained MPDS are (β is an arbitrary vector with $\beta_j > 0,\ j \in S$):

$$min \left\{ \sum_j \beta_j v_j \;\middle|\; \begin{array}{r} \sum_j \{\delta_{ij} - p_{ij}(a)\} v_j \geq 0 \qquad \forall (i,a) \\ v_i + \sum_j \{\delta_{ij} - p_{ij}(a)\} u_j \geq r_i(a)\ \forall (i,a) \end{array} \right\} \tag{1}$$

and

$$max \left\{ \sum_{(i,a)} r_i(a) x_i(a) \;\middle|\; \begin{array}{r} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) = 0 \quad \forall j \\ \sum_a x_j(a) + \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} y_i(a) = \beta_j\ \forall j \\ x_i(a), y_i(a) \geq 0\ \forall (i,a) \end{array} \right\} \tag{2}$$

In [2] it was shown that if (v, u) is an optimal solution of the primal problem (1), than $v = \phi$, the value vector. In [5] the following result was proven.

Theorem 1 *Let (x, y) be an extreme optimal solution of the dual program (2). Then, any stationary deterministic policy f such that*

$\begin{cases} x_i(f(i)) > 0 \text{ if } \sum_a x_i(a) > 0 \\ y_i(f(i)) > 0 \text{ if } \sum_a x_i(a) = 0 \end{cases}$ *is well-defined and is an average optimal policy.*

The correspondence between feasible solutions (x, y) of (2) and randomized stationary policies π is given by the following mappings. For a feasible solution (x, y) the corresponding policy $\pi^{x,y}$ is defined by

$$\pi_i^{x,y}(a) = \begin{cases} \frac{x_i(a)}{\sum_a x_i(a)} \text{ if } \sum_a x_i(a) > 0 \\ \frac{y_i(a)}{\sum_a y_i(a)} \text{ if } \sum_a x_i(a) = 0 \end{cases} \tag{3}$$

Conversely, for a stationary policy π, we define a feasible solution (x^π, y^π) of the dual program by

$$\begin{cases} x_i^\pi(a) = \left\{ \sum_j \beta_j \{P^*(\pi)\}_{ji} \right\} \cdot \pi_i(a) \\ y_i^\pi(a) = \left\{ \sum_j \beta_j \{D(\pi)\}_{ji} + \sum_j \gamma_j \{P^*(\pi)\}_{ji} \right\} \cdot \pi_i(a), \end{cases} \tag{4}$$

where $P^*(\pi)$ and $D(\pi)$ are the stationary and the deviation matrix of the transition matrix $P(\pi)$; $\gamma_j = 0$ on the transient states and constant on each recurrent class under $P(\pi)$ (for the precise definition of γ see [5]).

Example 1

It is well-known that in the irreducible case each extreme optimal solution has exactly one positive x-variable. It is also well known that in other cases some states can have zero positive x-variables. This example gives an MDP with an extreme optimal solution which has two positive x-variables for some state. Hence, the two corresponding deterministic and stationary policies are both optimal. Furthermore, this extreme feasible solution is mapped on a non-deterministic policy.

The dual program (2) of the model of Figure 1 is ($\beta_1 = \beta_2 = \frac{1}{4}, \beta_3 = \frac{1}{4}$):
maximize $x_1(1) + 2x_2(1) + 4x_3(1) + 3x_3(2)$
subject to

$$\begin{aligned}
x_1(1) - x_3(1) &= 0 \\
x_2(1) - x_3(2) &= 0 \\
x_1(1) - x_2(1) + x_3(1) + x_3(2) &= 0 \\
x_1(1) + y_1(1) - y_3(1) &= \tfrac{1}{4} \\
x_2(1) + y_2(1) - y_3(2) &= \tfrac{1}{4} \\
x_3(1) + x_3(2) - y_1(1) - y_2(1) + y_3(1) + y_3(2) &= \tfrac{1}{2} \\
x_1(1), x_2(1), x_3(1), x_3(2), y_1(1), y_2(1), y_3(1), y_3(2) &\geq 0
\end{aligned}$$

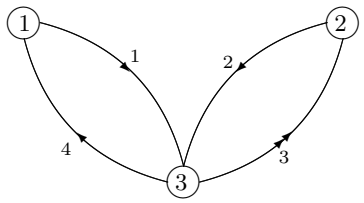

Figure 1

The solution (x, y) given by
$x_1(1) = x_2(1) = x_3(1) = x_3(2) = \frac{1}{4}$,
$y_1(1) = y_2(1) = y_3(1) = y_3(2) = 0$,
is an extreme optimal solution. State 3 has two positive x-variables.

Example 2

The next example (see Figure 2) shows that the mapping (3) is not one-to-one. Since the rewards are not important for this property, we drop these numbers in the picture. The constraints of the dual program are ($\beta_i = \frac{1}{4},\ 1 \le i \le 4$):

$$
\begin{aligned}
x_1(1) - x_3(2) &= 0\\
-x_1(1) + x_2(1) + x_2(2) &= 0\\
-x_2(1) + x_3(2) &= 0\\
-x_2(2) &= 0\\
x_1(1) + y_1(1) - y_3(2) &= \tfrac{1}{4}\\
x_2(1) + x_2(2) - y_1(1) + y_2(1) + y_2(2) &= \tfrac{1}{4}\\
x_3(1) + x_3(2) - y_2(1) + y_3(2) &= \tfrac{1}{4}\\
x_4(1) - y_2(2) &= \tfrac{1}{4}\\
x_1(1), x_2(1), x_2(2), x_3(1), x_3(2), x_4(1), y_1(1), y_2(1), y_2(2), y_3(2 &\ge 0
\end{aligned}
$$

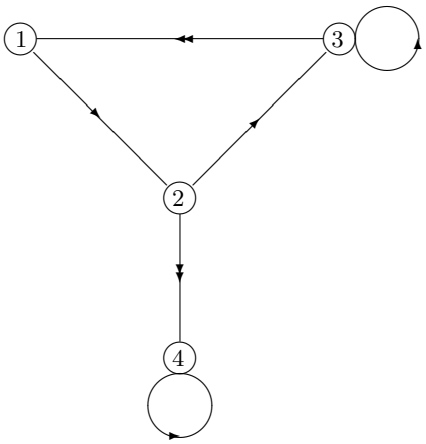

Figure 2

The solutions (x^1, y^1) given by
$x_1^1(1) = \frac{1}{4},\ x_2^1(1) = \frac{1}{4},\ x_2^1(1) = 0$,
$x_3^1(1) = 0,\ x_3^1(2) = \frac{1}{4},\ x_4^1(1) = \frac{1}{4}$,
$y_1^1(1) = y_2^1(1) = y_2^1(2) = y_3^1(2) = 0$,
and (x^2, y^2), where $x_1^2(1) = \frac{1}{6}$,
$x_2^2(1) = \frac{1}{6},\ x_2^2(1) = x_3^2(1) = 0$,
$x_3^2(2) = \frac{1}{6},\ x_4^2(1) = \frac{1}{2},\ y_1^2(1) = \frac{1}{6}$,
$y_2^2(1) = 0,\ y_2^2(2) = \frac{1}{4},\ y_3^2(2) = \frac{1}{12}$,
are two feasible solutions which are mapped on the same deterministic and stationary policy f, where
$f(1) = f(2) = 1, f(3) = 2, f(4) = 1$.

Example 3

In this example we present a feasible non-optimal solution which is mapped on an optimal policy. The dual program for the model of Figure 3 is ($\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$):

maximize $x_1(1)$

subject to

$$\begin{array}{rl}
x_1(1) + x_1(2) - x_2(1) & = 0 \\
x_2(1) - x_1(1) & = 0 \\
-x_1(2) & = 0 \\
x_1(1) + x_1(2) + y_1(1) + y_1(2) - y_2(1) & = \frac{1}{3} \\
x_2(1) + x_2(2) - y_1(1) + y_2(1) & = \frac{1}{3} \\
x_3(1) - y_1(2) & = \frac{1}{3} \\
x_1(1), x_1(2), x_2(1), x_2(2), x_3(1), y_1(1), y_1(2), y_2(1) & \geq 0
\end{array}$$

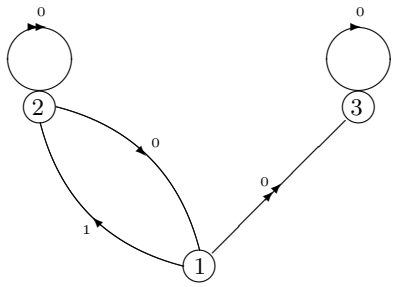

Figure 3

The solution (x, y) given by $x_1(1) = \frac{1}{6}$, $x_1(2) = 0,\ x_2(1) = \frac{1}{6},\ x_2(2) = 0,$ $x_3(1) = \frac{2}{3},\ y_1(1) = 0,\ y_1(2) = \frac{1}{3},$ $y_2(1) = \frac{1}{6}$ is a feasible solution, but not an optimal solution ($x_1(1) = x_2(1) = x_3(1) = \frac{1}{3}$ and all other variables 0 is the optimal solution). However, the corresponding policy f has actions $f(1) = 1,\ f(2) = 2,\ f(3) = 1$ is an optimal policy.

Example 4

Sometimes, the notion of a *general unichained* MDP is used. In the general unichain case, there exists an *optimal* policy with a single ergodic class plus a (perhaps empty) policy-dependent set of transient states. In this last example, we show that the general unichained model needs another approach than the unichain case; even an additional search procedure is not sufficient.

In the unichain case, the value vector is a constant vector and the linear programs simplify to

$$min \left\{ v \,\middle|\, v + \textstyle\sum_j \{\delta_{ij} - p_{ij}(a)\} u_j \geq r_i(a)\ \forall (i,a) \right\} \tag{5}$$

and

$$max \left\{ \sum_{(i,a)} r_i(a) x_i(a) \,\middle|\, \begin{array}{ll} \sum_{(i,a)} \{\delta_{ij} - p_{ij}(a)\} x_i(a) & = 0\ \forall j \\ \sum_{(i,a)} x_i(a) & = 1 \\ x_i(a) \geq 0\ \forall (i,a) & \end{array} \right\} \tag{6}$$

Furthermore, if x is an extreme optimal solution of the dual program (6), then any stationary deterministic policy f such that

$\begin{cases} x_i(f(i)) > 0 \text{ if } \sum_a x_i(a) > 0 \\ \text{arbitrarily} \quad \text{if } \sum_a x_i(a) = 0, \end{cases}$ is well-defined and is an average optimal policy.

The model of Figure 4 is general unichained since the policy which chooses in states 2 and 3 action 2 (state 1 has only one action) is an optimal policy and has a single chain structure. The dual program (6) of this model is:

maximize $x_2(2) + x_3(1)$

subject to

$$\begin{aligned} x_1(1) - x_2(1) &= 0 \\ -x_1(1) + x_2(1) - x_3(2) &= 0 \\ -x_3(2) &= 0 \\ x_1(1) + x_2(1) + x_2(2) + x_3(1) + x_3(2) &= 1 \\ x_1(1), x_2(1), x_2(2), x_3(1), x_3(1) &\geq 0 \end{aligned}$$

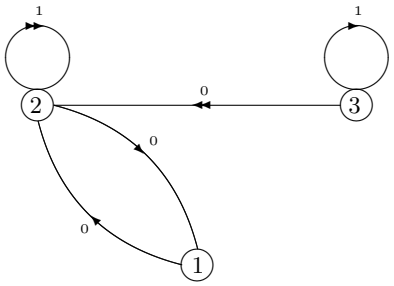

Figure 4

x given by $x_1(1) = x_2(1) = x_2(2) = x_3(2) = 0, x_3(1) = 1$ is an extreme optimal solution. In state 3, the policy corresponding to x chooses action 1. The choice in state 2 for an optimal policy is not arbitrary. Since the set of the states 1 and 2 is closed under any policy, it is impossible to search for actions in these states with transitions to the absorbing state 3.

Acknowledgement
I am very grateful to Arie Hordijk for the fruitful discussions and joint research over many years.

References

1. De Ghellinck, GT, Les problèmes de décisions sequentielles, Cahiers du Centre d'Etudes de Recherche Opérationelle, **2** (1960) 161–179.
2. Denardo, EV and Fox, BL, Multichain Markov renewal programs, SIAM Journal on Applied Mathematics **16** (1968) 468–487.
3. D'Epenoux, F, Sur un problème de production et de stockage dans l'aléatoire, Revue Française de Recherche Opérationelle **14,** (1960) 3–16.
4. Derman, C, *Finite state Markovian decision processes* (Academic Press, New York, 1970).
5. Hordijk, A and Kallenberg, LCM, Linear programming and Markov decision chains, Management Science, **25** (1979) 352–362.
6. Manne, AS, Linear programming and sequential decisions, Management Science, **6** (1960) 259–267.