Multiobjective Markov decision process with average reward criterion*

S. DURINOVIC
Faculty of Economic Sciences, University of Zagreb, Yugoslavia

H.M. LEE, M.N. KATEHAKIS
Department of Applied Mathematics, SUNY at Stony Brook, New York, NY 11790, U.S.A.

J.A. FILAR
Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD 21218, U.S.A.

Received 1 October 1985
Communicated by P. Borne

We consider a multiobjective, average reward, Markovian decision process and its relationship with an associated multiobjective linear program. In particular, we characterize the complete sets of efficient policies, efficient deterministic policies, and efficient points in the objective space. The latter characterization is achieved with the help of finite sets of basic efficient deterministic policies which can be constructed by known algorithms. In addition, we discuss the relationship between our definition of efficiency which depends on the initial state distribution, and another definition which is independent of the initial state. It turns out that these two definitions are, in a sense, equivalent in the unichain case, but not in the general multichain case. Our definition leads to a useful interpretation of the associated linear programs.

1. Introduction

The purpose of this paper is to clarify the relationship between efficient policies in a multiobjective average reward Markovian decision process (MDP, for short) and the efficient points of a related multiobjective linear program.

For single objective MDPs the relationship between optimal policies and solutions of related linear programs are well known (e.g. see Derman [2], Denardo and Fox [1], and Hordijk and Kallenberg [9]). In the past decade there has developed substantial interest in multiobjective MDPs; for a recent review of many of the relevant researches we refer the reader to Heyman and Sobel [8, ch. 6]. The discounted reward MDP has been studied by Henig [7]; also, Furukawa [6], Shin [13] and White and Kim [14] developed successive approximations and policy improvement type algorithms for constructing efficient policies. The connection with multiobjective linear programming is discussed in Viswanathan et. al. [15] and Durinovic [3].

However, the average reward case is technically more cumbersome, and (to the best of our

*Substantial portions of the results reported here were derived independently by Durinovic and Katehakis (see[3]), and by Filar and Lee (see [4] and [5]). The research of the latter authors was supported in part by NSF Grant #ECS8204677. S. Durinovic and M.N. Katehakis gratefully acknowledge E. Beltrami for bringing this topic to their attention

North-Holland
Large Scale Systems 10 (1986) 215–226

knowledge) has not been treated fully in the literature. In this paper we derive the following main results: (i) the characterization of the whole set of efficient policies and of its relationship with the set of efficient points of a related multiobjective linear program (Theorem 3.5); (ii) the characterization of the whole set of efficient points in the objective space of the multiobjective MDP with the help of a finite set of basic efficient policies (Theorem 3.9); and (iii) the characterization of all efficient deterministic policies (Theorem 3.7).

The above results are obtained by appropriately combining the MDP techniques of Hordijk and Kallenberg [9, 10] and Derman [2] with the techniques of multiobjective linear programming such as those of Iserman [11] and Yu and Zeleny [16]. While in the unichain case our results can be derived in a simple fashion, the possibility of multiple ergodic chains complicates the analysis, and introduces a number of pitfalls, some of which are demonstrated by the counterexamples given in Appendix A.

2. Definitions and preliminaries

A discrete *Markovian decision process* Γ is observed at discrete time points $t=1,2,\ldots$. The state space is denoted by $E=\{1,2,\ldots,N\}$. With each state $i\in E$, we associate a finite action set $A(i)$. At any time point t the system is in one of the states and an action has to be chosen by the decision-maker. If the system is in state i and action $a\in A(i)$ is chosen, then an immediate reward $r(i,a)$ is earned and the process moves to a state $j\in E$ with transition probability $p(j\,|\,i,a)$, where $p(j\,|\,i,a)\geqslant 0$ and $\sum_{j=1}^{N} p(j\,|\,i,a)=1$.

A *decision rule* D^t at time t is a function which assigns a probability to the event that action a is taken at time t. In general, D^t may depend on all realized states up to and including time t and on all realized actions up to time t. A *policy* π is a sequence of decision rules $\pi=(D^1,D^2,\ldots,D^t,\ldots)$. For a *Markov policy* we require that the decision rule at time t depends only on the state at time t, $t=1,2,\ldots$. A policy π is called *stationary* if all decision rules are identical, that is, $D^t\equiv D$ for all t. In this case, D_{ia} denotes the probability of choosing action a in state i. A *deterministic policy* is a stationary policy with nonrandomized decision rules. Let C, $C(M)$, $C(S)$ and $C(D)$ be the sets of all policies, the Markov policies, the stationary policies, and the deterministic policies, respectively.

Let X_t be the state at time t and Y_t be the action at time t and $p_\pi(X_t=j,\,Y_t=a\,|\,X_1=i)$ be the conditional probability that at time t the state is j and the action taken is a, given that the initial state is i and the decision-maker uses a policy π. For any policy π and initial state i, we define the *average expected reward* over the infinite horizon by

$$\phi(i,\pi)=\liminf_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{j,a}p_\pi(X_t=j,\,Y_t=a\,|\,X_1=i)r(j,a)\,.$$

A policy π^* is *average optimal* in Γ if $\phi(i,\pi^*)=\max_C\phi(i,\pi)$ for all $i\in E$. It is well known that there exists an average optimal deterministic policy (e.g. see [2, p. 25]).

Let $\beta=[\beta_1,\beta_2,\ldots,\beta_N]$ be a given initial distribution, that is, β_i is the probability that $X_1=i$, where $\beta_i\geqslant 0$ for all $i\in E$ and $\sum_i\beta_i=1$. For any policy π, define $\phi(\beta,\pi)$ by

$$\phi(\beta,\pi)=\sum_{i=1}^{N}\beta_i\phi(i,\pi)\,. \tag{2.1}$$

We shall say that π^0 is *β-optimal* in Γ if $\phi(\beta,\pi^0)\;=\max_C\phi(\beta,\pi)$. Clearly, an average optimal policy is β-optimal, but not conversely.

For any policy $\pi\in C$ and any positive integer T, we denote the *average expected state-action frequencies* in the first T periods by $x^T(\pi)$ according to the definition

$$x^T_{ja}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \beta_i p_\pi(X_t = j, Y_t = a \mid X_1 = i) , \tag{2.2}$$

for all $a \in A(j)$, $j \in E$. Let $X(\pi)$ denote the set of vector-limit-points of the sequence $\{x^T(\pi), T = 1, 2, \ldots\}$. Note that $X(\pi)$ is nonempty. Now, define

$$C_0 = \{\pi \in C \mid X(\pi) \text{ is a singleton}\} .$$

It is well known (e.g. see Kallenberg [12, p.135]) that $C(S) \subset C_0$. If we denote by $x(\pi)$ the unique element of $X(\pi)$ for any $\pi \in C_0$, then from the definition of $\phi(i, \pi)$ and (2.1) we have

$$\phi(\beta, \pi) = \sum_{j,a} x_{ja}(\pi) r(j, a) . \tag{2.3}$$

Let $L(C) = \{x(\pi) \in X(\pi) \mid \pi \in C\}$ and let the sets $L(C(M))$, $L(C_0)$, $L(C(S))$, and $L(C(D))$ be defined analogously.

The linear program usually associated with the average reward MDP is the following (see [1], [2], [9])

$$\max \sum_{i,a} r(i, a) x_{ia}$$

subject to

$$\sum_{i,a} (\delta_{ij} - p(j \mid i, a)) x_{ia} = 0, \quad j \in E ,$$

$$\sum_{a} x_{ia} + \sum_{i,a} (\delta_{ij} - p(j \mid i, a)) y_{ia} = \beta_j , \quad j \in E ,$$

$$x_{ia}, y_{ia} \geqslant 0; \quad i \in E, a \in A(i) . \tag{P}$$

Let F denote the set of all feasible solutions of the above linear program, and for $(x, y) \in F$ define[1] $S_x = \{i \in E \mid \sum_a x_{ia} > 0\}$ and $S_y = \{i \in E \mid \sum_a x_{ia} = 0 \text{ and } \sum_a y_{ia} > 0\}$. In addition, let $X = \{x \mid$ there exists y such that $(x, y) \in F\}$, and $\mathrm{Ext}(F)$, $\mathrm{Ext}(X)$ will denote the sets of extreme points of F and X, respectively. We shall use the following well-known results.

Theorem 2.1 (Derman [2], Hordijk and Kallenberg [9]).

$$\overline{L(C(D))} = \overline{L(C(S))} = L(C(M)) = L(C_0) = L(C) = X ,$$

where $\overline{A}$ denotes the closed convex hull of a set A.

Corollary 2.2. *For every $\pi \in C$ there exists an equivalent or better $\hat{\pi} \in C_0$ in the sense that $\phi(\beta, \pi) \leqslant \phi(\beta, \hat{\pi})$.*

With a given $(x, y) \in F$ we can associate a policy $\pi \in C(S)$ defined by

$$D_{ia}(x, y) = \begin{cases} x_{ia}/x_i, & \text{if } i \in S_x, a \in A(i) , \\ y_{ia}/y_i, & \text{if } i \in S_y, a \in A(i) , \\ \text{arbitrarily}, & \text{if } i \not\in S_x \cup S_y , \end{cases} \tag{2.4}$$

[1] Here $x(y)$ is treated as a vector whose components are $x_{ia}(y_{ia})$, arranged in the natural fashion. We shall not always differentiate between row and column vectors, this identification should be made in a manner consistent in any given equation.

where $x_i = \sum_a x_{ia}$ any $y_i = \sum_a y_{ia}$.

Furthermore, for $(x, y) \in F$ if $i \in S_x$ let $A_x(i) = \{a \in A(i) \mid x_{ia} > 0\}$, and if $i \in S_y$ let $A_y(i) = \{a \in A(i) \mid y_{ia} > 0\}$. Now define a family $C_{xy}(D)$ of deterministic policies $\pi(x, y)$ which chooses an arbitrary element $a \in A_x(i)$ $(A_y(i))$ if $i \in S_x(S_y)$ and an arbitrary $a \in A(i)$ if $i \not\in S_x \cup S_y$.

Conversely, with every $\pi \in C(s)$ we can associate a point $(x(\pi), y(\pi)) \in F$ defined by

$$x_{ia}(\pi) = [\beta^T P^*(\pi)]_i D_{ia}, \quad i \in E, a \in A(i), \tag{2.5}$$

$$y_{ia}(\pi) = [\beta^T D(\pi) + \gamma^T p^*(\pi)]_i D_{ia}, \quad i \in E, a \in A(i), \tag{2.6}$$

where $P^*(\pi)$ and $D(\pi)$ are the stationary and the deviation matrices, respectively, of the Markov chain induced by π, and γ is an appropriately chosen vector. For detailed definition of this transformation and the discussion of its relationship with the transformation (2.4), we refer the reader to Hordijk and Kallenberg [9].

Theorem 2.3.

(i) *If a policy $\pi_0 \in C_0$ is β-optimal and $x(\pi_0)$ is the corresponding vector limit point of state-action frequencies, then $(x(\pi_0), y)$ is optimal for* (P), *for every y such that $(x(\pi_0), y \in F$.*

(ii) *If $(x, y) \in F$ is an extreme optimal for* (P) *and $\pi \in C(S)$ is constructed via* (2.4), *or if $\pi = \pi(x, y) \in C_{xy}(D)$, then π is a β-optimal policy. In the former case the conclusion holds even if (x, y) is not extreme.*

Remark 2.4. In the case when $\beta_i > 0$, for all $i \in E$, the above theorem is proved in [9]. In the case when $\beta_i = 0$, for some $i \in E$, a detailed proof can be found in Durinovic [3, pp.28–35], and an outline is given in Appendix B. Note also the x and y above are now treated as $\sum_{i=1}^{N} |A(i)|$-dimensional vectors with components x_{ia} and y_{ia}, respectively ($|A(i)|$ is the cardinality of $A(i)$).

Now suppose that instead of one system of rewards we have m systems, that is, $r^k = \{r^k(i, a) \mid (i, a) \in E \times A(i)\}$, where $k \in m = \{1, 2, \ldots, m\}$ and $r^k(i, a)$ denotes the reward in the kth system corresponding to the state-action pair (i, a). Now, for each $k \in m$ we can define $\phi^k(\beta, \pi)$ via (2.1) by simply replacing $r(j, a)$ with $r^k(j, a)$. For each $k \in M$ we can find an average optimal policy π_k^*; however, in general $\pi_k^* \neq \pi_l^*$ if $k \neq l$ and there does not exist one policy that is optimal for all the m systems. We are thus led to consider efficient or nondominated policies in such a multiobjective MDP. In particular, we shall say that a policy $\pi^* \in U$ (where U is some subset of C) is β-*efficient* with respect to U if there does not exist $\pi \in U$ such that

$$\phi^k(\beta, \pi) \geqslant \phi^k(\beta, \pi^*) \quad \text{for all } k \in M,$$

with strict inequality holding for some $k_0 \in M$. We shall denote the set of all such β-efficient policies by $E(U)$.

With the above multiobjective MDP we shall associate the following multiobjective linear program:

$$\max_{(x,y) \in F} (r^1 x, r^2 x, \ldots, r^m x), \tag{MP}$$

where $r^k x = \sum_{i,a} r^k(i, a) x_{ia}$ for every $k \in M$. The point $(x^e, y^e) \in F$ will be called *efficient* for (MP) if there does not exist $(x, y) \in F$ such that $r^k x \geqslant r^k x^e$ for all $k \in M$, with strict inequality holding for some $k \in M$. The set of all efficient points for (MP) will be denoted by $E(F)$; also let $E(X) = \{x \in x \mid (x, y) \in E(F)$ for some $y\}$ (note that X is a bounded polyhedron).

Theorem 2.5. (Isermann [11]). *A point $(x^0, y^0) \in E(F)$ if and only if there exist positive weights $\lambda_k (\sum_k \lambda_k = 1)$ such that (x^0, y^0) is an optimal solution of the linear program*

$$\max \sum_k \lambda_k r^k x \tag{P$_\lambda$}$$

subject to $(x, y) \in F$.

Note that for every $\pi \in C_0$ we know from Theorem 2.1, and (2.3), that $x(\pi) = x$ for some $x \in X$ and hence

$$\phi^k(\beta, \pi) = \sum_{i,a} r^k(i, a) x_{ia}(\pi) = r^k x \,, \tag{2.7}$$

for every $k \in M$. Thus, the values of the objectives in (MP) correspond exactly to the payoffs $\phi^k(\beta, \pi)$ in the multiobjective MDP.

Theorem 2.6. (Yu and Zeleny [16]). *Let* $\mathrm{Ext}(X) \cap E(X) = N_{ex}$, then $\overline{N}_{ex} \supset E(X)$. *Thus, all efficient points are generated by the extreme points of $E(X)$, since they belong to its closed convex hull.*

Remark 2.7. Hordijk and Kallenberg [10, p. 282] used an alternative definition of an efficient policy, namely, they called π^* efficient if there does not exist a policy $\pi \in C$ such that

$$\phi^k(i, \pi) \geq \phi^k(i, \pi^*) \quad \text{for all } k \in M \text{ and } i \in E \,, \tag{2.8}$$

with strict inequality holding for some $k_0 \in M$ and $i_0 \in E$. We shall call such a policy π^* *u-efficient*, where u stands for uniformly. In Appendix A we show that u-efficiency and β-efficiency are different. However, (2.7) suggests that for the purpose of analyzing the multiobjective MDP with the help of the multiobjective linear program (MP), β-efficiency is perhaps the more natural solution concept.

3. Main results

Our aim is to characterize $E(C)$ and, equally importantly, the set $E_0(C)$ which is its image in the objective space. More precisely, $E_0(C) = \{\phi = (\phi^1, \ldots, \phi^m) \mid \text{there exists } \pi \in E(C) \text{ such that } \phi^k(\beta, \pi) = \phi^k \text{ for all } k \in M\}$. It will be convenient to introduce the following notation and conventions. We shall say that a p-dimensional vector $v \geq (>) 0$ if every component $v_i \geq (>) 0$, $i = 1, 2, \ldots, p$. Furthermore, we shall consider an $m \times \sum_{i=1}^N |A(i)|$-dimensional matrix R whose rows are $r^1, r^2, \ldots, r^m$, the rewards of the m single-objective MDPs (treated as $\sum_{i=1}^N |A(i)|$-dimensional vectors). Now, with every vector $\lambda = (\lambda_1 \ldots \lambda_m) \geq 0$, $(\sum_k \lambda_k = 1)$ we can associate a weighted MDP, denoted by $\Gamma(\lambda)$, which is identical to the original MDP in the transition law, but whose reward vector is defined by

$$r^\lambda = \lambda R = \sum_k \lambda_k r^k \,. \tag{3.1}$$

The payoff function $\phi^\lambda(\beta, \pi)$ of $\Gamma(\lambda)$ is defined as in (2.1).

Lemma 3.1. *Let $\lambda = (\lambda_1 \ldots \lambda_m) > 0$ and $\sum_{k=1}^N \lambda_k = 1$, and π^0 be a β-optimal policy of $\Gamma(\lambda)$, then $\pi^0 \in E(C)$. (Established independently in* [10]).

Proof. Suppose not; then there exists a $\hat{\pi} \in C$ such that $\phi^k(\beta, \hat{\pi}) \geq \phi^k(\beta, \pi^0)$ for all $k \in M$, with strict inequality holding for some $k_0 \in M$. It is easy to see (with the help of Theorem 2.1) that this leads to a contradiction of β-optimality in $\Gamma(\lambda)$. □

We can now prove the following result.

Theorem 3.2.
(i) If $(x^0, y^0) \in E(F)$ and $x^0 = x(\pi^0)$ for some $\pi^0 \in C_0$, then $\pi^0 \in E(C)$.
(ii) If $\pi^0 \in E(C) \cap C_0$, then $(x(\pi^0), y) \in E(F)$ for all y such that $(x(\pi^0), y) \in F$.

Proof.
(i) By Theorem 2.5 there exists $\lambda = (\lambda_1 \ldots \lambda_m) > 0$, with $\sum_\lambda \lambda_k = 1$ such that

$$\lambda R x^0 \geqslant \lambda R x \quad \text{for all } x \in X.$$

Hence by Theorem 2.1, and (2.7),

$$\phi^\lambda(\beta, \pi^0) \geqslant \phi^\lambda(\beta, \pi), \quad \text{for all } \pi \in C_0.$$

Now, π^0 is β-optimal in $\Gamma(\lambda)$ because otherwise there would exist some $\pi \in C$ such that $\phi^\lambda(\beta, \pi) > \phi^\lambda(\beta, \pi^0)$ which, in view of the above and Corollary 2.2, is impossible. Hence, $\pi^0 \in E(C)$ by Lemma 3.1.

(ii) The existence of a vector y such that $(x(\pi^0), y) \in F$ follows from Theorem 2.1. Suppose $(x(\pi^0), y) \not\in E(F)$, then there exists $(\hat{x}, \hat{y}) \in F$ such that

$$r^k \hat{x} \geqslant r^k x(\pi^0) \quad \text{for all } k \in M,$$

with strict inequality holding for some $k_0 \in M$. Again, by Theorem 2.1 there exists a policy $\hat{\pi} \in C_0$ such that $\hat{x} = x(\hat{\pi})$; hence, the above and (2.7) imply that $\phi^k(\beta, \hat{\pi}) \geqslant \phi^k(\beta, \pi^0)$ for all $k \in M$, with strict inequality holding for some k_0. Hence, $\pi^0 \not\in E(C)$ which yields the desired contradiction. □

One consequence of the preceding theorem is that the transformation (2.4) and its inverse, (2.5)–(2.6), preserve efficiency, which can be seen from the following result.

Corollary 3.3.
(i) *If $\pi^0 \in E(C) \cap C(S)$ and $(x^0, y^0) \in F$ is constructed from π^0 as in* (2.5)–(2.6), *then $(x^0, y^0) \in E(F)$.*
(ii) *Conversely, if $(x, y) \in E(F)$ and $\pi(x, y)$, as in* (2.4), *then $\pi(x, y) \in E(C)$.*

Proof. Part (i) follows from part (ii) of Theorem 3.2. Part (ii) can be easily proven with the help of Theorems 2.3, 2.5, and Lemma 3.1. □

Lemma 3.4. *For every $\pi \in E(C)$ there exists an equivalent $\hat{\pi} \in C_0 \cap C(M)$ in the sense that $\phi^k(\beta, \pi) = \phi^k(\beta, \hat{\pi})$ for all $k \in M$.*

Proof. Note that for any $k \in M$ and $x(\pi) \in X(\pi)$

$$\phi^k(\beta, \pi) \leqslant r^k x(\pi) = r^k x(\overline{\pi}) = \Phi^k(\beta, \overline{\pi}) \quad \text{for some } \overline{\pi} \in C_0,$$

where the first equality above follows from Theorem 2.1, while the second equality follows from (2.7). Now, since $\pi \in E(C)$ a strict inequality in the above is impossible. For every $\overline{\pi} \in C_0$ there exists a $\hat{\pi} \in C(M)$ such that $x(\overline{\pi}) = x(\hat{\pi})$ (see [2], [10]) which completes the proof. □

The next result characterizes the set $E(C)$ of all efficient policies, and is analogous to Theorem 2.5.

Theorem 3.5. *A policy $\pi^0 \in E(C)$ if and only if there exists a positive vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ with $\sum_k \lambda_k = 1$, such that π^0 is β-optimal for the single objective MDP with rewards r^λ.*

Proof. Sufficiency follows simply from Lemma 3.1 (and was established independently in [10]). To establish the necessity take any $\pi^0 \in E(C)$, then by Lemma 3.4 there exists an equivalent policy $\pi^* \in C_0$ which must also be β-efficient. Hence, by Theorem 3.2 $(x(\pi^*), y) \in E(F)$ whenever $(x(\pi^*), y) \in F$, and now Theorem 2.5 implies that $(x(\pi^*), y)$ is optimal for (P_λ) for some $\lambda > 0$ $(\sum_k \lambda_k = 1)$. Hence, by an argument such as that used in the proof of Theorem 3.2(i) π^* is β-optimal in $\Gamma(\lambda)$. Thus, π^0 is also β-optimal in $\Gamma(\lambda)$, as required. □

Corollary 3.6. *Suppose that $\pi^0 \in E(C)$ and that the initial distribution vector $\beta^0 > 0$, then π^0 remains β-efficient for any $\beta \geqslant 0$ with $\sum_i \beta_i = 1$.*

Proof. Since $\beta_i^0 > 0$ for every $i \in E$, we have that π^0 is uniformly optimal (i.e. independently of the initial state) in some weighted single objective MDP with all weights strictly positive; hence, π^0 is β-optimal in this MDP and the result follows from Theorem 3.5. □

The next two theorems show it is sufficient to consider a finite set of efficient policies. In particular let $\mathrm{Ext}(F) \cap E(F) = \{(x_1, y_1), \ldots, (x_p, y_p)\}$ and $\mathrm{D} = \bigcup_{s=1}^{p} C_{x_s y_s}(D)$ (see Section 2). Note that $D \subset C(D)$. The next theorem shows that D constitutes the set of *all* deterministic efficient policies.

Theorem 3.7. *$D = E(C) \cap C(D)$.*

Proof.

(i) By construction, if $\hat{\pi} \in D$, then $\hat{\pi} \in c_{x_s y_s}(D) \subset C(D)$ for some s. Also by Theorem 2.5 (x_s, y_s) is optimal for (P_λ) for some $\lambda > 0$. Now by Theorem 2.3(ii) $\hat{\pi}$ is β-optimal in $\Gamma(\lambda)$; hence, it follows from Theorem 3.5 that $\hat{\pi} \in E(C)$. Thus, $D \subset E(C) \cap C(D)$.

(ii) Suppose that $\hat{\pi} \in E(C) \cap C(D)$, and let $(\hat{x}, \hat{y}) \in F$ be constructed via (2.5)–(2.6). It can be easily checked that the proof of Theorem 4.3.4 of Kallenberg [12] remains valid even when $\beta \geqslant 0$. Thus, $(\hat{x}, \hat{y}) \in \mathrm{Ext}(F)$. Next we shall prove that $(\hat{x}, \hat{y}) \in E(F)$. If the latter were not so, then there would be some $(\overline{x}, \overline{y}) \in F$ such that $r^k\overline{x} \geqslant r^k\hat{x}$ for all $k \in M$, with strict inequality holding for some $k_0 \in M$. Thus, for any vector $\lambda > 0 (\sum_k \lambda_k = 1)$ we have

$$r^\lambda \overline{x} > r^\lambda \hat{x} .$$

But it follows from (2.5) that $\hat{x} = x(\hat{\pi})$, and there exists some $\overline{\pi} \in C_0$ such that $x(\pi) = \overline{x}$. Substituting these in the above yields $\phi^\lambda(\beta, \overline{\pi}) > \phi^\lambda(\beta, \hat{\pi})$, thus contradicting the hypothesis $\hat{\pi} \in E(C)$. Hence, $(\hat{x}, \hat{y}) \in \mathrm{Ext}\ (F) \cap E(F)$, and by definition of D, $C_{\hat{x}\hat{y}}(D) \subset D$. However, since $\hat{\pi} \in C(D)$ it follows from (2.5)–(2.6) that $\hat{\pi} \in C_{\hat{x}\hat{y}}(D)$, implying that $\hat{\pi} \in D$. This completes the proof. □

The final results derived below show that there are finite subsets of deterministic policies which, in a special sense, generate the whole efficient set in the objective space.

Lemma 3.8. *For every $x_s \in \mathrm{Ext}\ (X) \cap E(X)$ there exists a policy $\pi_s \in D$ such that $x_s = x(\pi_s)$.*

Proof. By Kallenberg [12, p.138] there exists $\pi_s \in C(D)$ such that $x_s = x(\pi_s)$. Since $x_s \in E(X)$ by Theorem 3.2(i) we have $\pi_s \in E(C)$, and hence by Theorem 3.7, $\pi_s \in D$. □

Now we would like to construct a deterministic policy π from a given point $x \in \mathrm{Ext}(X) \cap E(X)$. Since we know that $x = x(\pi)$ for some $\pi \in D$ and

$$x_{ja} = x_{ja}(\pi) = [\beta^T P^*(\pi)]_j \pi_{ja} , \tag{3.2}$$

Where $P^*(\pi)$ is the stationary matrix induced by π, it follows from (3.2) that if $j \in S_x$, then $x_{ja} = 0$ for all $a \in A(j)$ except one a_j. Hence, the policy $\pi \in D$ has to choose the action a_j in the state j for every $j \in S_x$. For $j \not\in S_x$ we have that $x_{ja} = 0$ for all $a \in A(j)$. In order to find the policy's rule at state $j \not\in S_x$, we have to examine the set $\{\pi \in C(D) \mid \pi_{ja_j} = 1 \text{ if } j \in S_x\}$, if for some π in this set $[\beta^T P^*(\pi)]_j \pi_{ja} = 0$ for all $a \in A(j)$ and $j \not\in S_x$ then π is the policy required, that is, $x = x(\pi)$.

Without loss of generality we let (see Theorem 2.6) $N_{ex} = \{x_1, \ldots, x_n\}$, with $n \leqslant p$, and for each $s = 1, \ldots, n$ we let π_s be constructed from x_s via the above. Now we shall refer to a set $D_b = \{\pi_1, \ldots, \pi_n\} \subset D$ as a *basic efficient set* of policies (recall that $D = C(D) \cap E(C)$).

Theorem 3.9 *Let $D_b = \{\pi_1, \pi_2, \ldots, \pi_n\}$ be a basic efficient set, as defined above, and choose an arbitrary $\phi = (\phi^1, \ldots, \phi^m) \in E_0(C)$, then there exist non-negative numbers a_s, $s = 1, \ldots, n$, such that $\sum_{s=1}^n a_s = 1$, and*

$$\phi^k = \sum_{s=1}^{n} a_s \phi^k(\beta, \pi_s) \quad \text{for all } k \in m.$$

Proof. By definition of $E_0(C)$ and Lemma 3.4 there exist $\pi \in E(C)$ and $\hat{\pi} \in C_0$ such that for every $k \in M$

$$\phi^k = \phi^k(\beta, \pi) = \phi^k(\beta, \hat{\pi}) = r^k x(\hat{\pi}) . \tag{3.3}$$

Thus, $\hat{\pi} \in E(C)$. Since $x(\hat{\pi}) \in X$, there exists y such that $(x(\hat{\pi}), y) \in F$, and by Theorem 3.2(ii) it is also in $E(F)$. Now, by Theorem 2.6 there exist non-negative numbers a_s, $s = 1, \ldots, n$, such that $\sum_{s=1}^n a_s = 1$ and for $x_s \in N_{\text{ex}}$ the relation

$$x(\hat{\pi}) = \sum_{s=1}^{n} a_s x_s \tag{3.4}$$

is satisfied. Hence from (3.3) and (3.4) we have

$$\begin{aligned} \phi^k &= \sum_{s=1}^{n} a_s (r^k x_s) = \sum_{s=1}^{n} a_s (r^k x(\pi_s)) \\ &= \sum_{s=1}^{n} a_s \phi^k(\beta, \pi_s) \quad \text{for all } k \in M . \end{aligned} \tag{3.5}$$

The fact that $x_s = x(\pi_s)$ follows from the definition of D_b, and the last equality is simply (2.7). This completes the proof. □

Remark 3.10. The above results indicate that the difficulty of characterizing either $E(C) \cap C(D)$ or $E_0(C)$ is comparable to the difficulty of generating all extreme efficient points of the multiobjective linear program (MP) via one of the standard algorithms for this purpose (e.g. see Yu and Zeleny [16]). However, in the two-objective case a more efficient algorithm for characterizing $E_0(C)$ has recently been proposed by Filar and Lee [5].

Appendix A

In Section 2 we introduced two concepts of efficiency in multiobjective Markovian decision processes:

the u-efficient policy and the β-efficient policy. We shall now comment on the relationships between these two definitions.

Remark A.1. Let $\beta=(\beta_2,\ldots,\beta_N)$ with $\beta_i>0$ for all $i=1,2,\ldots,N$, then it can be easily seen that a β-efficient policy is also a u-efficient policy.

Remark A.2. If $\beta_i=0$ for some i, then Remark A.1 no longer holds, as can be seen from the following example. Suppose $\beta=(0,1)$ and we have the two MDP processes:

$$k=1 \qquad \begin{array}{cc} i=1 & i=2 \\ \begin{pmatrix}3\\3\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}1,0\\1,0\end{pmatrix} & \begin{pmatrix}2\\2\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}0,1\\0,1\end{pmatrix} \end{array}$$

$$k=2 \qquad \begin{array}{cc} i=1 & i=2 \\ \begin{pmatrix}1\\3\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}1,0\\1,0\end{pmatrix} & \begin{pmatrix}1\\1\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}0,1\\0,1\end{pmatrix} \end{array}$$

The notation $\to(\cdot,\cdot)$ gives the probability transitions when a given action is chosen (e.g. if action 2 is chosen in state 1 in the process $k=2$, the reward of 3 is earned and the system remains in state 1). Let $\pi^*\in C(D)$ always choose action 1 in states 1 and 2, π^* is a β-efficient policy but not a u-efficient policy, since policy π^0 which always chooses action 2 in states 1 and 2 dominates policy π^*.

Remark A.3. A u-efficient policy is not always a β-efficient policy, as can be seen from the following example:

$$k=1 \qquad \begin{array}{cc} i=1 & i=2 \\ \begin{pmatrix}5\\4\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}1,0\\1,0\end{pmatrix} & \begin{pmatrix}5\\7\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}0,1\\0,1\end{pmatrix} \end{array}$$

$$k=2 \qquad \begin{array}{cc} i=1 & i=2 \\ \begin{pmatrix}2\\7\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}1,0\\1,0\end{pmatrix} & \begin{pmatrix}7\\5\end{pmatrix} \begin{matrix}\to\\ \to\end{matrix} \begin{pmatrix}0,1\\0,1\end{pmatrix} \end{array}$$

Let $\pi^*\in C(D)$ always choose action 1 in states 1 and 2, then π^* is a u-efficient policy but it is not always β-efficient. This is because policy π^0, which always chooses action 2 in states 1 and 2, dominates policy π^* whenever $\beta=(\beta_1,\beta_2)$ and $\beta_1\in(2/7,2/3)$.

However, in the important unichain case, i.e. when for any $\pi\in C(D)$ the Markov chain induced by π has exactly one ergodic set plus a (perhaps empty) set of transient states (it is straightforward to show that this property also holds for any stationary policy), there is a close relationship between the concepts of u-efficiency and β-efficiency.

Theorem A.4. *In the unichain case, $\pi\in C(S)$ is u-efficient if and only if π is β-efficient for any initial distribution β.*

Proof. Since the Markov chain induced by $\pi\in C(S)$ has one ergodic set, the vector $\phi^k(\pi)=(\phi^k(1,\pi),\ldots,\phi^k(N,\pi))$ has identical components.

(i) Suppose that π is u-efficient but not β-efficient for some β, then there exists $\pi^0\in C$ such that

$$\beta\phi^k(\pi^0)\geqslant\beta\phi^k(\pi)\quad\text{for all } k,$$

with strict inequality holding for some k. That is, for some $x(\pi^0) \in X(\pi^0) \subset X$, and an arbitrary fixed i,

$$r^k x(\pi^0) \geqslant \phi^k(i, \pi) \quad \text{for all } k,$$

with strict inequality holding for some k. Since $X = L(S)$ (see [12, pp. 155–156]) there exists a policy $\pi^* \in C(S)$ such that $x(\pi^*) = x(\pi^0)$, and hence for an arbitrary fixed i

$$r^k x(\pi^*) = \phi^k(i, \pi^*) \geqslant \phi^k(i, \pi) \quad \text{for all } k,$$

with strict inequality holding for some k. This contradicts the u-efficiency of π.

(ii) Suppose that π is β-efficient but not u-efficient, then there exists some policy $\pi^0 \in C$ such that, for all i and k,

$$\phi^k(i, \pi^0) \geqslant \phi^k(i, \pi),$$

with strict inequality holding for some $i = e$ and $k = k_e$. However, we know that $\phi^k(e, \pi^0) \leqslant r^k x(\pi^0)$ for any $x(\pi^0) \in X(\pi^0)$ (since $X = L(S) \supset X(\pi^0)$ and in computing $x(\pi^0)$ we assume the initial state is e). Thus, there exists a policy $\pi^* \in C(S)$ such that $x(\pi^*) = x(\pi^0)$ which satisfies

$$\phi^k(e, \pi^*) \geqslant \phi^k(e, \pi) \quad \text{for all } k,$$

with strict inequality holding for $k = k_e$. Since $\pi^*, \pi \in C(S)$ we obtain:

$$\phi^k(e, \pi^*) = \beta\phi^k(\pi^*) \geqslant \beta\phi^k(\pi) \quad \text{for all } k,$$

with strict inequality holding for $k = k_e$ for every $\beta \geqslant 0$. This contradicts the β-efficiency of π.

Appendix B

We now give an outline of the proof of Theorem 2.3. As mentioned earlier, the case where $\beta > 0$ was proved in Hordijk and Kallenberg [9]. In the case where $\beta \geqslant 0$, the proof of part (i) can again be established by an argument such as that used to prove Theorem 8(a) in [9]; however, the proof of part (ii) requires the resolution of a number of technical difficulties.

Suppose then that (x^0, y^0) is optimal for the linear program (P) and $\pi^0 \in C_{x^0 y^0}(D)$. Let $P^*(\pi^0) = (p^*_{ij}(\pi^0))^N_{i,j=1}$ be the stationary Markov matrix induced by π^0. Using an argument similar to that in Proposition 1 of [9], and the constraints of (P), we can show that

$$p_{ij}(\pi^0) \equiv p^*_{ij}(\pi^0) \equiv 0 \quad \text{for all } i \in S_{x^0},\ j \notin S_{x^0}, \tag{B.1}$$

and

$$p_{ij}(\pi^0) \equiv p^*_{ij}(\pi^0) \equiv 0 \quad \text{for all } i \in S_{x^0} \cup S_{y^0},\ j \notin S_{x^0} \cup S_{y^0}. \tag{B.2}$$

Next, it can be shown, as in Proposition 3 of [9], that all states of S_{y^0} are transient in the Markov chain induced by π^0. Hence, we have that

$$p^*_{ij}(\pi^0) \equiv 0 \quad \text{for all } j \in S_{y^0}. \tag{B.3}$$

Let π^* be an average optimal policy and let $\phi_i = \phi(i, \pi^*)$ for every $i \in E$. It now follows from the definition of π^0, the complementary slackness conditions applied to (P) and its dual, and the above that

$$\phi_i = \sum_{j \in S_{x^0}} p^*_{ij}(\pi^0)\phi_j, \quad i \in S_{x^0} \cup S_{y^0}, \tag{B.4}$$

and

$$\phi_i + \sum_{j \in S_{x^0}} (\delta_{ij} - p^*_{ij}(\pi^0))u_j = r(i, a_i), \quad i \in S_{x^0}, \tag{B.5}$$

where a_i is the action selected by π^0 in state i. It now follows from (B.1)–(B.5) that for every $i \in S_{x^0} \cup S_{y^0}$:

$$\begin{aligned} \phi(i, \pi^0) &= \sum_{j \in E} r(j, a_j) p^*_{ij}(\pi^0) \\ &= \sum_{j \in S_{x^0}} \Big[\phi_j + u_j - \sum_{k \in S_{x^0}} p^*_{ij}(\pi^0) u_k\Big] p^*_{ij}(\pi^0) \\ &= \phi_i + \sum_{j \in S_{x^0}} p^*_{ij}(\pi^0) u_j - \sum_{k \in S_{x^0}} p^*_{ik}(\pi^0) u_k \\ &= \phi_i . \end{aligned} \tag{B.6}$$

Hence,

$$\begin{aligned} \phi(\beta, \pi^0) &= \sum_{i \in B} \beta_i \phi(i, \pi^0) = \sum_{i \in S_{x^0} \cup S_{y^0}} \beta_i \phi(i, \pi^0) \\ &= \sum_{i \in S_{x^0} \cup S_{y^0}} \beta_i \phi_i = \sum_{i \in E} \beta_i \phi_i , \end{aligned} \tag{B.7}$$

where the second and the last equality above follow from the fact that $B = \{ j \in E \mid \beta_j > 0 \} \subset S_{x^0} \cup S_{y^0}$. But (B.7) shows that $\phi(\beta, \pi^0) \geqslant \phi(\beta, \pi)$ for $\pi \in C$, as required.

If π^0 is constructed from (x^0, y^0) via the transformation (2.4), the β-optimality of π^0 can be proved along the same lines but with some additional technical difficulties. For details of this argument we refer the reader to Durinovic [3, pp. 24–32].

References

[1] E.V. Denardo and B.L. Fox, Multichain Markov renewal programs, SIAM J. Appl. Math. 16 (1968) 468–487.

[2] C. Derman, Finite State Markovian Decision Processes (Academic Press, New York, 1970).

[3] S. Durinovic, On multiple objective Markov decision processes, Ph.D. Thesis, SUNY at Stony Brook (1983).

[4] J.A. Filar and H.M. Lee, Efficient policies in a multiobjective Markovian decision process with average rewards, Technical Report #375, Department of Mathematical Sciences, The Johns Hopkins University (1983).

[5] J.A. Filar and H.M. Lee, Variability in undiscounted Markov decision processes, Technical Report #400, Department of Mathematical Sciences, The Johns Hopkins University (1984).

[6] N. Furukawa, Vector valued Markov decision processes with countable state space, in: R. Hartley, L.C. Thomas and D.J. White (eds.), Recent Developments in Markov Decision Processes (Academic Press, New York, 1980) pp. 205–223.

[7] M.I. Henig, Vector-valued dynamic programming, SIAM J. Control and Optimization (1983) 490–499.

[8] D.P. Heyman and M.J. Sobel, Stochastic Models in Operations Research Vol. II (McGraw-Hill New York, 1984).

[9] A. Hordijk and L.C.M. Kallenberg, Linear programming and Markov decision chains, Management Sci. 25 (1979) 352–362.

[10] A. Hordijk and L.C.M. Kallenberg, Constrained undiscounted dynamic programming, Math. Op. Res. 9 (1984) 276–289.

[11] M. Isermann, Proper efficiency ant the linear vector maximization problem, Op. Res. 22 (1974) 189–191.

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

[13] M.C. Shin, Computational methods for Markov decision problems, Ph.D. Thesis, University of British Columbia, Vancouver B.C., Canada (1980).

[14] C.C. White III and K.W. Kim, Solution procedures for vector criterion Markov decision processes, Large Scale Systems 1 (1980) 129–140.

[15] B. Viswanathan, V.V. Aggarwal and K.P.K. Nair, Multiple criteria Markov decision processes, TIMS Studies in Management Sciences 6 (1977) 263–272.

[16] P. Yu and M. Zeleny, The set of all nondominated solutions in linear cases and multi-criteria simplex method, J. Math. Analysis Appl. 49 (1975) 430–468.