Probability in the Engineering and Informational Sciences, **14**, 2000, 533–548. Printed in the U.S.A.

VECTOR-VALUED MARKOV DECISION PROCESSES WITH AVERAGE REWARD CRITERION: THE MULTICHAIN CASE

KAZUYOSHI WAKUTA
Nagaoka Technical College
Nagaoka, Niigata 940-8532, Japan
E-mail: wakuta@nagaoka-ct.ac.jp

We study the multichain case of a vector-valued Markov decision process with average reward criterion. We characterize optimal deterministic stationary policies via systems of linear inequalities and discuss a policy iteration algorithm for finding all optimal deterministic stationary policies.

1. INTRODUCTION

We study average reward, vector-valued Markov decision processes (average reward VMDPs, for short). The discounted reward VMDPs have been investigated by many authors (e.g., Viswanathan, Aggarwal, and Nair [17], Furukawa [4,5], White and Kim [21], White [22], Henig [7], Novák [12,13], Wakuta [18,19]). On the other hand, average reward VMDPs do not seem to be treated fully in the literature, except for several authors. Thomas [15] dealt with the completely ergodic case and presented a policy iteration algorithm for finding all (Pareto) optimal deterministic stationary policies by modifying the Furukawa's algorithm [5]. Iki and Furukawa [10] considered the multichain case under different optimality from that of Thomas [15]. Durinovic, Lee, Katehakis, and Filar [3] formulated average reward VMDPs as multiobjective linear programs and characterized optimal deterministic stationary policies in the multichain case. Novák [12,13] solved the completely ergodic case as well as the discounted reward case via multiobjective linear programming. Ghosh [6] developed an algorithm for finding an optimal policy with respect to some utility

© 2000 Cambridge University Press 0269-9648/00 $12.50

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

function in the completely ergodic case. Recently, Wakuta and Togawa [20] have characterized optimal deterministic stationary policies via systems of linear inequalities in the completely ergodic case and have suggested a policy iteration algorithm. The purpose of this paper is to study the multichain case through a similar approach and to discuss a policy iteration algorithm for finding all optimal deterministic stationary policies.

Thomas [15] and Iki and Furukawa [10] restricted themselves to deterministic stationary policies. Thus, they had only to compare the values of policies for dominance if there exist undetermined policies after the policy iteration. However, here we treat all (randomized, history-remembering) policies. Then, policies that are optimal among all deterministic stationary policies but not optimal among all policies should be excluded from the solutions. Hence, their methods do not apply in this case. In the present paper, a policy is said to be optimal if it is nondominated for all initial states. This definition differs from β-efficiency in Durinovic et al. [3]. A policy is β-efficient if it is nondominated for a given initial state distribution β. If a policy is β-efficient for some $\beta > 0$, then it is also optimal. However, the converse does not hold in general. In their Appendix A, Durinovic et al. [3] discussed the relation between β-efficiency and u-efficiency. The u-efficiency was introduced in Hordijk and Kallenberg [9] and is a weaker condition than the optimality in the present paper.

2. DEFINITIONS AND PRELIMINARIES

Let R^m denote the m-dimensional Euclidean space. We shall adopt the standard convention for inequalities: Given $a = (a_1, \ldots, a_m)$, $b = (b_1, \ldots, b_m) \in R^m$, define $a \geqq b$ if and only if $a_k \geqq b_k$, $k = 1, \ldots, m$; $a \geq b$ if and only if $a \geqq b$ and $a \neq b$; $a > b$ if and only if $a_k > b_k$, $k = 1, \ldots, m$. Similarly, $\leqq$, $\leq$, and, $<$ are defined. For a nonempty subset $U \subset R^m$, a point $x \in U$ is said to be an efficient point of U when $x \leqq y$ for some $y \in U$ implies $x = y$. We denote by $e(U)$ the set of all efficient points of U. We apply the relations on R^m to the set $B^m(S)$ of all m-valued functions on S: Given $u, v \in B^m(S)$, define $u \geqq v$ if and only if $u(i) \geqq v(i)$ for all $i \in S$; $u \geq v$ if and only if $u \geqq v$ and $u \neq v$; $u > v$ if and only if $u(i) > v(i)$ for all $i \in S$. A constant function in $B^m(S)$ is said to be a constant vector.

The (average reward) VMDP is defined by the following: $S = \{1, 2, \ldots, N\}$, the state space; A is a finite set, the action space; $A(i)$ is the permissible set of actions in state $i \in S$; $p(j|i,a)$, $i, j \in S$, $a \in A(i)$, is a stochastic kernel on S given $\mathrm{Gr}\, A$, the transition law, where $\mathrm{Gr}\, A = \{(i,a) | i \in S, a \in A(i)\}$; $r(i,a) = (r^1(i,a), \ldots, r^m(i,a)) \in R^m$ is the vector-valued reward function on $\mathrm{Gr}\, A$.

Let $H_1 = S$ and $H_{t+1} = \mathrm{Gr}\, A \times H_t$, $t \geqq 1$. Then, H_t is the set of all possible histories of the system when the tth action must be chosen. A policy π is a sequence $\{\pi_1, \pi_2, \ldots\}$, where π_t is a stochastic kernel on A given H_t. It is assumed that $\pi_t(A(i_t)|h_t) = 1$ for all histories $h_t = (i_1, a_1, i_2, \ldots, i_t) \in H_t$. We denote by Π the set of all policies. A deterministic stationary policy is a sequence $\{f, f, \ldots\}$, $f \in F$, where F is the set of all mappings f from S to A such that $f(i) \in A(i)$, $i \in S$. We denote such

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

a policy by f^∞ and the set of all deterministic stationary policies by Π_D. For $f^\infty \in \Pi_D$, we define the stochastic matrix $P(f) = (p_{ij}(f))$ as $P_{ij}(f) = p(j|i, f(i))$, $i, j \in S$. We denote by $P^*(f) = (p^*_{ij}(f))$ the stationary matrix of $P(f)$ and by $D(f)$ the deviation matrix (see, e.g., Kallenberg [11]). We also define the $N \times m$ reward matrix $R(f) = (r^k_f(i))$ as $r^k_f(i) = r^k(i, f(i))$, $i = 1, \ldots, N$, $k = 1, \ldots, m$, and the $N \times m$ bias-value matrix u_f as $u_f = D(f)R(f)$.

Let $H' = A \times S \times A \times \cdots$. Then, any policy π defines a probability measure $p_\pi\{\cdot | X_1 = i_1\}$ on H' for any initial state $i_1 \in S$ (see, e.g., Hinderer [8]). For any $\pi \in \Pi$ and any positive integer T, we denote by $x^T[\pi](i_1)$ the expected state-action frequencies in the first T periods when the system starts from the initial state $i_1 \in S$; that is,

$$x^T_{ja}[\pi](i_1) = \frac{1}{T} \sum_{t=1}^{T} P_\pi\{X_t = j, Y_t = a | X_1 = i_1\}, \qquad (j, a) \in \mathrm{Gr}\, A, \tag{2.1}$$

where X_t is the tth state and Y_t is the tth action chosen at the tth stage. We also denote by $X[\pi](i_1)$ the set of all limit points $x[\pi](i_1)$ of the vectors $\{x^T[\pi](i_1), T = 1, 2, \ldots\}$. For any $i_1 \in S$, $\Pi_1(i_1)$ stands for the set of policies such that $X[\pi](i_1)$ is a singleton. Because $\lim_{T\to\infty} x^T_{jf(j)}[f^\infty](i_1) = p^*_{i_1 j}(f)$ for any $f^\infty \in \Pi_D$, $\Pi_D \subset \Pi_1(i_1)$ for all $i_1 \in S$.

When r is scalar-valued, we usually define the expected average reward as

$$\phi_\pi(i_1) = \liminf_{T\to\infty} \sum_{j,a} x^T_{ja}[\pi](i_1) r(j, a), \qquad i_1 \in S, \tag{2.2}$$

and a policy π^* is said to be optimal if $\phi_{\pi^*}(i_1) = \sup_{\pi\in\Pi} \phi_\pi(i_1)$ for all $i_1 \in S$.

For a vector-valued reward r, we analogously define the expected average reward as

$$\phi_\pi(i_1) = (\phi^1_\pi(i_1), \ldots, \phi^m_\pi(i_1)), \tag{2.3}$$

where $\phi^k_\pi(i_1) = \liminf_{T\to\infty} \sum_{j,a} x^T_{ja}[\pi](i_1) r^k(j, a)$, $i_1 \in S$. Obviously, for a policy $\pi \in \bigcap_{i_1\in S} \Pi_1(i_1)$,

$$\phi_\pi(i_1) = \sum_{j,a} x_{ja}[\pi](i_1) r(j, a), \qquad i_1 \in S. \tag{2.4}$$

For each $i_1 \in S$, we set

$$V(i_1) = \bigcup_{\pi\in\Pi} \{\phi_\pi(i_1)\}, \qquad V_1(i_1) = \bigcup_{\pi\in\Pi_1(i_1)} \{\phi_\pi(i_1)\},$$

$$V_D(i_1) = \bigcup_{f^\infty\in\Pi_D} \{\phi_{f^\infty}(i_1)\}. \tag{2.5}$$

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

DEFINITION 2.1: *A policy π^* is said to be optimal for* VMDP *if $\phi_\pi(i_1) \in e(V(i_1))$ for all $i_1 \in S$.*

We also set

$$W(i_1) = \bigcup_{\pi \in \Pi} \left\{ \sum_{j,a} x_{ja}[\pi](i_1) r(j,a) | x[\pi](i_1) \in X[\pi](i_1) \right\}, \qquad i_1 \in S. \tag{2.6}$$

PROPOSITION 2.1: *$W(i_1) = V_1(i_1) = \text{co}\, V_D(i_1)$, $i_1 \in S$, where* co *stands for the convex hull of a set.*

PROOF: Let $L(i_1) = \bigcup_{\pi \in \Pi} X[\pi](i_1)$, $L_1(i_1) = \bigcup_{\pi \in \Pi_1(i_1)} X[\pi](i_1)$, and $L_D(i_1) = \bigcup_{f^\infty \in \Pi_D} X[f^\infty](i_1)$, $i_1 \in S$. Then, $L(i_1) = L_1(i_1) = \text{co}\, L_D(i_1)$, $i_1 \in S$ (see Kallenberg [11, Thm. 4.7.1]). Hence, $W(i_1) = V_1(i_1) = \text{co}\, V_D(i_1)$, $i_1 \in S$. ■

PROPOSITION 2.2: *$e(V(i_1)) = e(V_1(i_1))$, $i_1 \in S$.*

PROOF: The set of all limit points of $\{\sum_{j,a} x_{ja}^T[\pi](i_1) r^k(j,a), T = 1,2,\ldots\}$ is identical to $\{\sum_{j,a} x_{ja}[\pi](i_1) r^k(j,a) | x[\pi](i_1) \in X[\pi](i_1)\}$, and $\phi_\pi^k(i_1)$ is the minimum point of this set. From this fact and Proposition 2.1, $V_1(i_1) \subset V(i_1) \subset V_1(i_1) + R_-^m$, $i_1 \in S$, where R_-^m is the nonpositive orthant of R^m. Thus, the result follows. ■

By virtue of the above propositions, we can restrict ourselves to the set $\Pi_1(i_1)$ of policies for each $i_1 \in S$.

3. OPTIMAL DETERMINISTIC STATIONARY POLICIES

For a $c \in B^m(S)$, we consider a nonstationary dynamic program (NDP(c), for short) defined by the same data as in VMDP except for the reward function $r^c(i_1, i_t, a_t) = \langle c(i_1), r(i_t, a_t)\rangle$, where $\langle x, y\rangle$ denotes the inner product of $x, y \in R^m$. Then, the expected average reward for NDP(c) is defined as

$$J_\pi(i_1) = \sum_{j,a} x_{ja}[\pi](i_1) r^c(i_1, j, a), \qquad \pi \in \Pi_1(i_1),\ i_1 \in S. \tag{3.1}$$

Note that $J_\pi(i_1) = \langle c(i_1), \phi_\pi(i_1)\rangle$. If we choose a constant vector c, then we have the ordinary Markov decision process (MDP(c), for short) with the reward function $r^c(i,a) = \langle c, r(i,a)\rangle$. A policy π^* is called the optimal policy for NDP(c) if $J_{\pi^*}(i_1) \geqq J_\pi(i_1)$ for all $\pi \in \Pi_1(i_1)$ and all $i_1 \in S$.

Because $V_1(i_1)$ is a polytope by Proposition 2.1, the following scalarization holds (cf. Yu [23, Thm. 8.5(ii)]).

PROPOSITION 3.1: *A policy $\pi \in \bigcap_{i_1 \in S} \Pi_1(i_1)$ is optimal for* VMDP *if and only if π is optimal for* NDP(c) *with a $c \in B^m(S)$, $c > 0$.*

PROPOSITION 3.2: *There exists an optimal deterministic stationary policy in* VMDP.

PROOF: Consider an MDP(c) with a constant vector $c > 0$. Because S and A are finite, there exists an optimal deterministic stationary policy for MDP(c) (see, e.g., Derman [2]). By Proposition 3.1, it is optimal for VMDP. ■

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

We define

$$S(f^\infty, i_1) = \{j \in S | P_{f^\infty}\{X_t = j | X_1 = i_1\} > 0 \text{ for some } t \geqq 1\}, \qquad i_1 \in S, \quad \textbf{(3.2)}$$

and

$$S(i_1) = \bigcup_{f^\infty \in \Pi_D} S(f^\infty, i_1), \qquad i_1 \in S. \quad \textbf{(3.3)}$$

Then, an optimal deterministic stationary policy is characterized as follows.

THEOREM 3.1: *If a deterministic stationary policy f^∞ is optimal for* VMDP, *then there exist a $u \in B^m(S)$ and a $c \in B^m(S)$, $c > 0$, such that for each $i_1 \in S$ and any $(i_t, a_t) \in \operatorname{Gr} A$ with $i_t \in S(f^\infty, i_1)$,*

$$\langle c(i_1), \phi_{f^\infty}(i_t)\rangle \geqq \sum_{j=1}^{N} p(j | i_t, a_t)\, \langle c(i_1), \phi_{f^\infty}(j)\rangle \quad \textbf{(3.4)}$$

and

$$\langle c(i_1), \phi_{f^\infty}(i_t)\rangle + \langle c(i_1), u(i_t)\rangle \geqq \langle c(i_1), r(i_t, a_t)\rangle + \sum_{j=1}^{N} p(j | i_t, a_t)\, \langle c(i_1), u(j)\rangle. \quad \textbf{(3.5)}$$

PROOF: By Proposition 3.1, f^∞ is optimal for NDP(c) with a $c \in B^m(S)$, $c > 0$. Given an initial state $i_1 \in S$, let $\bar{c} = c(i_1)$ and consider an MDP($\bar{c}$) with the reward function $\bar{r}(i, a) = \langle \bar{c}, r(i, a)\rangle$, $(i, a) \in \operatorname{Gr} A$. Then, f^∞ is i_1-optimal for MDP($\bar{c}$); that is, $\bar{J}_{f^\infty}(i_1) \geqq \bar{J}_\pi(i_1)$ for all $\pi \in \Pi_1(i_1)$. Besides, by Theorem 1 in Schweitzer and Gavish [14], f^∞ is i-optimal for MDP($\bar{c}$) for all $i \in S(f^\infty, i_1)$. Although we restrict ourselves to the set $\Pi_1(i_1)$ of policies, their theorem applies to MDP($\bar{c}$) because there exists an optimal policy in Π_D for MDP($\bar{c}$). For a Blackwell optimal deterministic stationary policy g^∞ for MDP($\bar{c}$), let $\bar{J}_{g^\infty} = P^*(g)\bar{r}(g)$ and $\bar{u}_{g^\infty} = D(g)\bar{r}(g)$. Also, let $J^*(i) = \bar{J}_{g^\infty}(i)$ and $w^*(i) = \bar{u}_{g^\infty}(i) - M\bar{J}_{g^\infty}(i)$, $i \in S$, where M is a number defined as in the proof of Theorem 4.2.2 in Kallenberg [11]. Then,

$$J^*(i) \geqq \sum_{j=1}^{N} p(j | i, a) J^*(j) \quad \textbf{(3.6)}$$

and

$$J^*(i) + w^*(i) \geqq \bar{r}(i, a) + \sum_{j=1}^{N} p(j | i, a) w^*(j), \quad (i, a) \in \operatorname{Gr} A. \quad \textbf{(3.7)}$$

Define $u(i) = u_{g^\infty}(i) - M\phi_{g^\infty}(i)$, $i \in S$, where $u_{g^\infty}(i)$ is the ith row of $U_{g^\infty} = D(g)R(g)$. Note that $\bar{J}_{g^\infty}(i) = \langle \bar{c}, \phi_{g^\infty}(i)\rangle$ and $\bar{u}_{g^\infty}(i) = \langle \bar{c}, u_{g^\infty}(i)\rangle$. Then, $w^*(i) = \langle \bar{c}, u(i)\rangle$, $i \in S$.

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

Now, we have

$$J^*(i) = \bar{J}_{g^\infty}(i) \begin{cases} = \bar{J}_{f^\infty}(i) = \langle \bar{c}, \phi_{f^\infty}(i) \rangle & \text{for all } i \in S(f^\infty, i_1) \\ \geqq \bar{J}_{f^\infty}(i) = \langle \bar{c}, \phi_{f^\infty}(i) \rangle & \text{otherwise.} \end{cases} \tag{3.8}$$

Substituting (3.8) and $w^*(i) = \langle \bar{c}, u(i) \rangle$, $i \in S$, into (3.6) and (3.7), we can obtain the desired result. ■

THEOREM 3.2: *A deterministic stationary policy f^∞ is optimal for* VMDP *if there exist a $u \in B^m(S)$ and a $c \in B^m(S)$, $c > 0$, such that* (3.4) *and* (3.5) *hold for each $i_1 \in S$ and any $(i_t, a_t) \in \mathrm{Gr}\, A$ with $i_t \in S(i_1)$.*

PROOF: Fix an arbitrary initial state $i_1 \in S$ and let g^∞ be an arbitrary deterministic stationary policy. Then, using (3.4), we have

$$\begin{aligned} & E_{g^\infty}[\langle c(X_1), \phi_{f^\infty}(X_t) \rangle | X_1 = i_1] \\ & \quad = E_{g^\infty}[E_{g^\infty}[\langle c(X_1), \phi_{f^\infty}(X_t) \rangle | X_1, X_{t-1}, g(X_{t-1})] | X_1 = i_1] \\ & \quad = E_{g^\infty}\left[\sum_{j=1}^{N} p(j | X_{t-1}, g(X_{t-1})) \langle c(X_1), \phi_{f^\infty}(j) \rangle | X_1 = i_1 \right] \\ & \quad \leqq E_{g^\infty}[\langle c(X_1), \phi_{f^\infty}(X_{t-1}) \rangle | X_1 = i_1]. \end{aligned} \tag{3.9}$$

Iterating (3.9) $t-1$ times, we have

$$E_{g^\infty}[\langle c(X_1), \phi_{f^\infty}(X_t) \rangle | X_1 = i_1] \leqq \langle c(i_1), \phi_{f^\infty}(i_1) \rangle. \tag{3.10}$$

Now, we have

$$E_{g^\infty}\left[\sum_{t=2}^{T+1} \{ \langle c(X_1), u(X_t) \rangle - E_{g^\infty}[\langle c(X_1), u(X_t) \rangle | X_1, X_{t-1}, g(X_{t-1})] \} | X_1 = i_1 \right] = 0. \tag{3.11}$$

By (3.5), we have

$$\begin{aligned} & E_{g^\infty}[\langle c(X_1), u(X_t) \rangle | X_1, X_{t-1}, g(X_{t-1})] \\ & \quad = \sum_{j=1}^{N} p(j | X_{t-1}, g(X_{t-1})) \langle c(X_1), u(j) \rangle \\ & \quad = \langle c(X_1), r(X_{t-1}, g(X_{t-1})) \rangle + \sum_{j=1}^{N} p(j | X_{t-1}, g(X_{t-1})) \langle c(X_1), u(j) \rangle \\ & \qquad - \langle c(X_1), r(X_{t-1}, g(X_{t-1})) \rangle \\ & \quad \leqq \langle c(X_1), \phi_{f^\infty}(X_{t-1}) \rangle + \langle c(X_1), u(X_{t-1}) \rangle - \langle c(X_1), r(X_{t-1}, g(X_{t-1})) \rangle. \end{aligned} \tag{3.12}$$

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

Substituting (3.12) into (3.11), we have

$$E_{g^\infty}\left[\sum_{t=2}^{T+1}\{\langle c(X_1),u(X_t)\rangle-\langle c(X_1),u(X_{t-1})\rangle+\langle c(X_1),r(X_{t-1},g(X_{t-1}))\rangle\}|X_1=i_1\right]$$

$$\leqq E_{g^\infty}\left[\sum_{t=2}^{T+1}\langle c(X_1),\phi_{f^\infty}(X_{t-1})\rangle|X_1=i_1\right]. \tag{3.13}$$

By (3.10), we have

$$E_{g^\infty}\left[\langle c(X_1),u(X_{T+1})\rangle-\langle c(X_1),u(X_1)\rangle+\sum_{t=2}^{T+1}\langle c(X_1),r(X_{t-1},g(X_{t-1}))\rangle|X_1=i_1\right]$$

$$\leqq T\langle c(i_1),\phi_{f^\infty}(i_1)\rangle. \tag{3.14}$$

Dividing both sides of (3.14) by T and letting $T\to\infty$, we have

$$\langle c(i_1),\phi_{f^\infty}(i_1)\rangle \geqq \langle c(i_1),\phi_{g^\infty}(i_1)\rangle; \tag{3.15}$$

that is, $J_{f^\infty}(i_1)\geqq J_{g^\infty}(i_1)$. Note that there exists an i_1-optimal policy in Π_D for NDP(c) because an optimal policy exists in Π_D for MDP$(\bar{c})$ with $\bar{c}=c(i_1)$. As $i_1\in S$ is arbitrary, f^∞ is optimal for NDP(c). By Proposition 3.1, f^∞ is optimal for VMDP. ■

COROLLARY 3.1: *A deterministic stationary policy f^∞ is optimal for* VMDP *if there exist a $u\in B^m(S)$ and a $c\in B^m(S)$, $c>0$, such that* (3.4) *and* (3.5) *hold for all $i_1\in S$ and all $(i_t,a_t)\in \mathrm{Gr}\,A$.*

In Section 4, we shall give a counterexample showing the existence of a gap between the necessary and sufficient conditions. However, if we suppose the complete ergodicity, then this gap can be removed.

Assumption 3.1 (The completely ergodic case): For any $f^\infty\in\Pi_D$, all states belong to a single ergodic set in the Markov chain induced by $P(f)$.

The following is the same as Theorem 3.3 in Wakuta and Togawa [20].

THEOREM 3.3: *Suppose Assumption* 3.1 *is satisfied. Then, a deterministic stationary policy f^∞ is optimal for* VMDP *if and only if there exists a constant vector $c>0$ such that for any $(i,a)\in \mathrm{Gr}\,A$,*

$$\langle c,\phi_{f^\infty}\rangle+\langle c,u_{f^\infty}(i)\rangle \geqq \langle c,r(i,a)\rangle+\sum_{j=1}^{N}p(j|i,a)\langle c,u_{f^\infty}(i)\rangle, \tag{3.16}$$

where ϕ_{f^∞} is a constant vector.

PROOF: The proof can be found in Proposition 3.4 of Wakuta and Togawa [20]. ■

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

Theorems 3.1 and 3.2 give simple conditions for optimality, but as stated earlier, there is a gap between them. To find all optimal deterministic stationary policies, we need further discussion. Assume that some deterministic stationary policies have already been known to be not optimal and denote by B the set of such policies. Let

$$\Pi'_D = \Pi_D \cap B^c \tag{3.17}$$

and

$$V'_D(i_1) = \bigcup_{f^\infty \in \Pi'_D} \{\phi_{f^\infty}(i_1)\}, \qquad i_1 \in S. \tag{3.18}$$

We define

$$\Pi^*_D = \{ f^\infty \in \Pi_D | \, f^\infty \text{ is optimal for MDP}(c) \text{ with some constant vector } c > 0\} \tag{3.19}$$

and

$$V^*_D(i_1) = \bigcup_{f^\infty \in \Pi^*_D} \{\phi_{f^\infty}(i_1)\}, \qquad i_1 \in S. \tag{3.20}$$

LEMMA 3.1:

$$e(\text{co } V^*_D(i_1)) = e(\text{co } V'_D(i_1)) = e(V_1(i_1)), \qquad i_1 \in S. \tag{3.21}$$

PROOF: Choose an arbitrary initial state $i_1 \in S$. Note that there exists an $f^\infty \in \Pi^*_D$ whose value corresponds to an efficient extreme point of $V_1(i_1)$. Hence, $e(V_1(i_1)) \subset \text{co } V^*_D(i_1)$. Accordingly, $e(V_1(i_1)) = e(\text{co } V^*_D(i_1))$. On the other hand, $\Pi^*_D \subset \Pi'_D$ by Proposition 3.1. Then, $\text{co } V^*_D(i_1) \subset \text{co } V'_D(i_1) \subset V_1(i_1)$. Thus, the result follows directly. ■

THEOREM 3.4: *A deterministic stationary policy f^∞ is optimal for* VMDP *if and only if $f^\infty \in \Pi'_D$ and there exists a $c \in B^m(S)$, $c > 0$, such that for each $i_1 \in S$*

$$\langle c(i_1), \phi_{f^\infty}(i_1)\rangle \geqq \langle c(i_1), \phi_{g^\infty}(i_1)\rangle, \qquad g^\infty \in \Pi'_D, \tag{3.22}$$

where Π'_D is the set of policies defined by (3.17).

PROOF: Assume that $f^\infty \in \Pi'_D$ and that there exists a $c \in B^m(S)$, $c > 0$, satisfying (3.22). Then, for any $\nu \in \text{co } V'_D(i_1)$

$$\langle c(i_1), \phi_{f^\infty}(i_1)\rangle \geqq \langle c(i_1), v\rangle, \qquad i_1 \in S. \tag{3.23}$$

Hence,

$$\phi_{f^\infty}(i_1) \in e(\text{co } V'_D(i_1)), \qquad i_1 \in S. \tag{3.24}$$

By Lemma 3.1, f^∞ is optimal for VMDP. The converse is obvious from Proposition 3.1. ■

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

4. CHARACTERIZATION OF OPTIMAL POLICIES VIA SYSTEMS OF LINEAR INEQUALITIES

Given an initial state $i_1 \in S$, let $x_k = c^k(i_1)$ and $y_i = \sum_{k=1}^m u^k(i)x_k$, where c and u are defined in Theorem 3.1. Then, replacing (i_t, a_t) with (i, a), the conditions in Theorem 3.1 can be written in the form

$$(S_{i_1}):\begin{cases} \sum_{k=1}^m \phi_{f^\infty}^k(i)x_k \geqq \sum_{k=1}^m \sum_{j=1}^N p(j|i,a)\phi_{f^\infty}^k(j)x_k \\ \sum_{k=1}^m \phi_{f^\infty}^k(i)x_k + y_i \geqq \sum_{k=1}^m r^k(i,a)x_k + \sum_{j=1}^N p(j|i,a)y_j, \quad (i,a) \in \operatorname{Gr} A,\ i \in S(f^\infty, i_1) \\ x_k > 0, \qquad k = 1,\ldots,m. \end{cases} \tag{4.1}$$

Conversely, assume that (S_{i_1}) has a solution (x, y). Then, decomposing y_i as $y_i = \sum_{k=1}^m y_i^k$ and letting $u^k(i) = y_i^k/x_k$, we have $y_i = \sum_{k=1}^m u^k(i)x_k$. Hence, the conditions in Theorem 3.1 hold. Similarly, replacing $S(f^\infty, i_1)$ in (4.1) with $S(i_1)$ and with S, respectively, the conditions in Theorem 3.2 and Corollary 3.1 can also be written in the systems of linear inequalities, which we denote by (T_{i_1}) and (S_0). Thus, Theorems 3.1 and 3.2 and Corollary 3.1 turn to the following.

THEOREM 4.1: *If a deterministic stationary policy f^∞ is optimal for* VMDP, *then each system $(S_1),\ldots,(S_N)$ of linear inequalities has a solution (x, y).*

THEOREM 4.2: *A deterministic stationary policy f^∞ is optimal for* VMDP *if each system $(T_1),\ldots,(T_N)$ of linear inequalities has a solution (x, y).*

COROLLARY 4.1: *A deterministic stationary policy f^∞ is optimal for* VMDP *if the system (S_0) of linear inequalities has a solution (x, y).*

Remark 4.1: In the completely ergodic case, the conditions can also be represented via the system of linear inequalities (see Thm. 3.4 in Wakuta and Togawa [20]).

Given an initial state $i_1 \in S$, let $x_k = c^k(i_1)$, $k = 1, 2, \ldots, m$, where $c \in B^m(S)$, $c > 0$. Then, the conditions in Theorem 3.4 can be written in the form

$$(U_{i_1}):\begin{cases} \sum_{k=1}^m \phi_{f^\infty}^k(i_1)x_k \geqq \sum_{k=1}^m \phi_{g^\infty}^k(i_1)x_k, \quad g^\infty \in \Pi'_D \\ x_k > 0, \qquad k = 1,\ldots,m. \end{cases} \tag{4.2}$$

Then, Theorem 3.4 turns to

THEOREM 4.3: *A deterministic stationary policy f^∞ is optimal for* VMDP *if and only if $f^\infty \in \Pi'_D$ and each system $(U_1),\ldots,(U_N)$ of linear inequalities has a solution x.*

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

Now, we shall show a counterexample exhibiting a policy that satisfies the necessary condition but not the sufficient condition.

Counterexample: Consider the following VMDP:

$$S = \{1,2,3\}, \quad A(1) = A(2) = \{1,2\}, \quad A(3) = \{1\};$$
$$p(1|1,1) = p(2|1,2) = 1, \quad p(2|2,1) = p(3|2,2) = 1, \quad p(3|3,1) = 1;$$
$$r(1,1) = (4,4), \quad r(1,2) = (0,0), \quad r(2,1) = (6.1),$$
$$r(2,2) = (0,0), \quad r(3,1) = (5,3).$$

There are four deterministic stationary policies $\alpha^\infty, \beta^\infty, \gamma^\infty$, and δ^∞ defined as follows:

$$\alpha^\infty: \alpha(1) = 1, \quad \alpha(2) = 1; \qquad \beta^\infty: \beta(1) = 1, \quad \beta(2) = 2;$$
$$\gamma^\infty: \gamma(1) = 2, \quad \gamma(2) = 1; \qquad \delta^\infty: \delta(1) = 2, \quad \delta(2) = 2.$$

The values of these policies can be calculated directly from the above data as

$$\phi_{\alpha^\infty}(1) = (4,4), \quad \phi_{\alpha^\infty}(2) = (6,1), \quad \phi_{\alpha^\infty}(3) = (5,3);$$
$$\phi_{\beta^\infty}(1) = (4,4), \quad \phi_{\beta^\infty}(2) = (5,3), \quad \phi_{\beta^\infty}(3) = (5,3);$$
$$\phi_{\gamma^\infty}(1) = (6,1), \quad \phi_{\gamma^\infty}(2) = (6,1), \quad \phi_{\gamma^\infty}(3) = (5,3);$$
$$\phi_{\delta^\infty}(1) = (5,3), \quad \phi_{\delta^\infty}(2) = (5,3), \quad \phi_{\delta^\infty}(3) = (5,3).$$

Obviously, all deterministic stationary policies are optimal. Let us examine α^∞. From the transition law, $S(\alpha^\infty,1) = \{1\}$, $S(\alpha^\infty,2) = \{2\}$, $S(\alpha^\infty,3) = \{3\}$, $S(1) = \{1,2,3\}$, $S(2) = \{2,3\}$, and $S(3) = \{3\}$. Substituting the above data into (4.1), we can easily see that each system (S_1), (S_2), and (S_3) has a solution. However, (S_1) is inconsistent with (S_2). Hence, $(T_1) = (S_1) \cup (S_2) \cup (S_3)$ has no solution. Therefore, α^∞ satisfies the necessary condition but not the sufficient condition.

The problem of whether the system (S_{i_1}) has a solution (x, y) can be formulated by the following linear program:

$$P(S_{i_1}): \text{Max } z$$

subject to

$$\begin{cases} x_1 \geqq z, \ldots, x_m \geqq z \\ \displaystyle\sum_{k=1}^{m} \phi_{f^\infty}^k(i) x_k \geqq \sum_{k=1}^{m} \sum_{j=1}^{N} p(j|i,a) \phi_{f^\infty}^k(j) x_k \\ \displaystyle\sum_{k=1}^{m} \phi_{f^\infty}^k(i) x_k + y_i \geqq \sum_{k=1}^{m} r^k(i,a) x_k + \sum_{j=1}^{N} p(j|i,a) y_j, \\ \qquad\qquad (i,a) \in \operatorname{Gr} A,\ i \in S(f^\infty, i_1) \\ x_k \geqq 0, \qquad k = 1, \ldots, m,\ z \geqq 0. \end{cases} \tag{4.3}$$

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

Linear programs $P(T_{i_1})$, $P(S_0)$, and $P(U_{i_1})$ are defined similarly. Obviously, $P(S_{i_1})$, $P(T_{i_1})$, $P(S_0)$, and $P(U_{i_1})$ are feasible and have nonnegative objectives. Then, we immediately have the following theorem:

THEOREM 4.4:

(i) *The systems* (S_{i_1}), (T_{i_1}), (S_0), *and* (U_{i_1}) *have solutions if and only if* $P(S_{i_1})$, $P(T_{i_1})$, $P(S_0)$, *and* $P(U_{i_1})$ *have positive maximum values, respectively. In these cases, in fact,* $P(S_{i_1})$, $P(T_{i_1})$, $P(S_0)$, *and* $P(U_{i_1})$ *have unbounded objectives.*

(ii) *The systems* (S_{i_1}), (T_{i_1}), (S_0), *and* (U_{i_1}) *have no solution if and only if* $P(S_{i_1})$, $P(T_{i_1})$, $P(S_0)$, *and* $P(U_{i_1})$ *have maximum values of zero, respectively.*

5. POLICY ITERATION ALGORITHM

In this section, we denote by f a deterministic stationary policy f^∞ for simplicity. We set $H_f^g = P(g)\phi_f$ and $I_f^g = R(g) + P(g)u_f$, where $\phi_f = (\phi_f^k(i))$ is the $N \times m$ matrix and U_f is the bias-value matrix defined in Section 2. Note that $H_f^f = \phi_f$ and $I_f^f = \phi_f + u_f$. Now, let us consider the following relations associated with $f, g \in \Pi_D$:

i. $H_f^g \geqq H_f^f$, and for any (i,k) for which $(H_f^g)_{ik} = (H_f^f)_{ik}$, $(I_f^g)_{ik} \geqq (I_f^f)_{ik}$. Moreover, for at least one (i,k), $(H_f^g)_{ik} > (H_f^f)_{ik}$ or $(I_f^g)_{ik} > (I_f^f)_{ik}$. **(5.1)**

ii. $H_f^g \leqq H_f^f$, and for any (i,k) for which $(H_f^g)_{ik} = (H_f^f)_{ik}$, $(I_f^g)_{ik} \leqq (I_f^f)_{ik}$. Moreover, for at least one (i,k), $(H_f^g)_{ik} < (H_f^f)_{ik}$ or $(I_f^g)_{ik} < (I_f^f)_{ik}$. **(5.2)**

iii. $H_f^g = H_f^f$ and $I_f^g = I_f^f$. **(5.3)**

LEMMA 5.1:

(i) *If* (5.1) *holds, then* $\phi_g \geqq \phi_f$.

(ii) *If* (5.2) *holds, then* $\phi_g \leqq \phi_f$.

(iii) *If* (5.3) *holds, then* $\phi_g = \phi_f$.

PROOF: (i) Suppose (5.1) holds. Let k satisfy that for some $i \in S$, $(H_f^g)_{ik} > (H_f^f)_{ik}$ or $(I_f^g)_{ik} > (I_f^f)_{ik}$. Then, by Corollary 1 in Veinott [16], $\phi_g^k(i) \geqq \phi_f^k(i)$ for all $i \in S$. If k satisfies that $(H_f^g)_{ik} = (H_f^f)_{ik}$ and $(I_f^g)_{ik} = (I_f^f)_{ik}$ for all $i \in S$, then $\phi_g^k(i) = \phi_f^k(i)$ for all $i \in S$ by Lemma 2 in Blackwell [1]. Thus, $\phi_g \geqq \phi_f$. The proof of (ii) is similar. The proof of (iii) is straightforward from Lemma 2 in Blackwell [1]. ■

5.1. Policy Iteration Algorithm

By E_n, F_n, and G_n, $n \geqq 0$, we denote respectively the set of deterministic stationary policies that cannot be dominated by any other policy via policy iteration until the nth stage, the set of deterministic stationary policies that have been dominated by some other policy via policy iteration until the nth stage, and the set of deterministic stationary policies that have been determined to be optimal until the nth stage. The policies in E_n could be optimal, but those in F_n are no longer optimal.

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

5.1.1. Phase I

1. Set $E_0 = F_0 = G_0 = \phi$ and choose an $f_1 \in \Pi_D$ arbitrarily.
2. For f_n, $n \geqq 1$, determine $\tilde{\phi} = \phi_{f_n}$ and $\tilde{u} = u_{f_n}$ by solving
$$(I - P(f_n))\tilde{\phi} = O,$$
$$\tilde{\phi} + (I - P(f_n))\tilde{u} = r(f_n),$$
$$\tilde{u} + (I - P(f_n))\tilde{z} = O,$$
where I is the identity matrix and O is the zero matrix (see Kallenberg [11, Thm. 4.4.1]).
3. Determine
$$A_{f_n} = \{g \in (\Pi_D \backslash (E_{n-1} \cup F_{n-1} \cup G_{n-1}))|\ (5.1) \text{ holds for } f = f_n\},$$
$$B_{f_n} = \{g \in (\Pi_D \backslash (E_{n-1} \cup F_{n-1} \cup G_{n-1}))|\ (5.2) \text{ holds for } f = f_n\},$$
$$C_{f_n} = \{g \in (\Pi_D \backslash (E_{n-1} \cup F_{n-1} \cup G_{n-1}))|\ (5.3) \text{ holds for } f = f_n\}.$$
4. If $A_{f_n} \neq \phi$, choose an $f_{n+1} \in A_{f_n}$ and return to Step 2.

   (i) If $\phi_{f_n} \leq \phi_{f_{n+1}}$ and $\phi_{f_{n-k}} = \cdots = \phi_{f_n}$ for some k, let
   $$E_n = E_{n-1}, \quad F_n = F_{n-1} \cup (B_{f_{n-k}} \cup \cdots \cup B_{f_n}) \cup (C_{f_{n-k}} \cup \cdots \cup C_{f_n}),$$
   and $G_n = G_{n-1}$ and go to Step 3.

   (ii) If $\phi_{f_n} = \phi_{f_{n+1}}$, let
   $$E_n = E_{n-1}, \qquad F_n = F_{n-1}, \quad \text{and} \quad G_n = G_{n-1}.$$
   Store in memory that $\phi_{f_n} = \phi_{f_{n+1}}$ as well as B_{f_n} and C_{f_n}, and go to Step 3.

   If $A_{f_n} = \phi$, solve $P(S_0)$ for f_n.

   (iii) If f_n can be determined to be optimal and $\phi_{f_{n-k}} = \cdots = \phi_{f_n}$ for some k, define
   $$M_n = \{g \in (B_{f_{n-k}} \cup \cdots \cup B_{f_n}) | \phi_g = \phi_{f_n}\},$$
   $$N_n = \{g \in (B_{f_{n-k}} \cup \cdots \cup B_{f_n}) | \phi_g \leq \phi_{f_n}\}.$$
   Let
   $$E_n = E_{n-1}, \qquad F_n = F_{n-1} \cup N_n, \quad \text{and}$$
   $$G_n = G_{n-1} \cup M_n \cup (C_{n-k} \cup \cdots \cup C_n).$$
   Choose an $f_{n+1} \in (\Pi_D \backslash (E_n \cup F_n \cup G_n))$ and return to Step 2.

   (iv) If f_n cannot be determined to be optimal (f_n may still be optimal) and $\phi_{f_{n-k}} = \cdots = \phi_{f_n}$ for some k, let
   $$E_n = E_{n-1} \cup (B_{f_{n-k}} \cup \cdots \cup B_{f_n}) \cup (C_{f_{n-k}} \cup \cdots \cup C_{f_n}),$$
   $$F_n = F_{n-1}, \quad \text{and} \quad G_n = G_{n-1}.$$
   Choose an $f_{n+1} \in (\Pi_D \backslash (E_n \cup F_n \cup G_n))$ and return to Step 2.
5. Stop Phase I when $E_n \cup F_n \cup G_n = \Pi_D$.

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

Set $E^* = E_n$, $F^* = F_n$, and $G^* = G_n$, and denote by G the set of all optimal deterministic stationary policies. If $E^* = \phi$, then $G = G^*$ and the procedure ends; otherwise, go to Phase II.

5.1.2. Phase II. For each $f \in E^*$, solve $P(U_1), \ldots, P(U_N)$ with $\Pi'_D = E^* \cup G^*$ one by one. If each $P(U_i)$ is unbounded, then f is optimal; otherwise, f is not optimal. Thus, G can be obtained.

Remark 5.1:

(i) The algorithm does not necessarily call for the values of all deterministic stationary policies; if $(B_{f_{n-k}} \cup \cdots \cup B_{f_n}) \neq \phi$ in Step 4(i) of Phase I, then one need not calculate the values of those policies since then.
(ii) If Assumption 3.1 is satisfied, then Theorem 3.3 provides a necessary and sufficient condition, and equality never holds in Lemma 5.1 (i) and (ii). In this case, the algorithm can be very simplified.

6. A NUMERICAL EXAMPLE

We shall give a numerical example illustrating the policy iteration algorithm presented in Section 5. Consider the following VMDP:

$S = \{1,2\}, A = A(1) = A(2) = \{1,2\}$

$p(1|1,1) = 1, p(2|1,1) = 0$

$p(1|1,2) = \frac{1}{2}, p(2|1,2) = \frac{1}{2}$

$p\,(1|2,1) = 0, p(2|2,1) = 1$

$p(1|2,2) = \frac{1}{2}, p(2|2,2) = \frac{1}{2}$

$r(1,1) = (0,0), r(1,2) = (1,-1)$

$r(2,1) = (2,1), r(2,2) = (1,3).$

There are four deterministic stationary policies α, β, γ, and δ defined as follows: α: $\alpha(1) = 1$, $\alpha(2) = 1$; β: $\beta(1) = 1$, $\beta(2) = 2$; γ: $\gamma(1) = 2$, $\gamma(2) = 1$; δ: $\delta(1) = 2$, $\delta(2) = 2$.

6.1. Phase I

1. Choose $f_1 = \alpha$ and calculate ϕ_α and u_α. Then,

$$\phi_\alpha = \begin{bmatrix} 0 & 0 \\ 2 & 1 \end{bmatrix}, \qquad u_\alpha = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

Calculate H_α^α, I_α^α, H_α^β, I_α^β, H_α^γ, I_α^γ, H_α^δ, and I_α^δ. Then,

$$H_\alpha^\alpha = \begin{bmatrix} 0 & 0 \\ 2 & 1 \end{bmatrix}, \qquad I_\alpha^\alpha = \begin{bmatrix} 0 & 0 \\ 2 & 1 \end{bmatrix},$$

$$H_\alpha^\beta = \begin{bmatrix} 0 & 0 \\ 1 & \frac{1}{2} \end{bmatrix}, \qquad I_\alpha^\beta = \begin{bmatrix} 0 & 0 \\ 1 & 3 \end{bmatrix},$$

$$H_\alpha^\gamma = \begin{bmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{bmatrix}, \qquad I_\alpha^\gamma = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix},$$

$$H_\alpha^\delta = \begin{bmatrix} 1 & \frac{1}{2} \\ 1 & \frac{1}{2} \end{bmatrix}, \qquad I_\alpha^\delta = \begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix}.$$

Hence, $A_\alpha = \{\gamma\}$, $B_\alpha = \{\beta\}$, and $C_\alpha = \{\alpha\}$.

2. Choose $f_2 = \gamma \in A_\alpha$ and calculate ϕ_γ and u_γ. Then,

$$\phi_\gamma = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}, \qquad u_\gamma = \begin{bmatrix} -2 & -4 \\ 0 & 0 \end{bmatrix}.$$

Since $\phi_\alpha \leq \phi_\gamma$, $E_1 = \phi$, $F_1 = \{\alpha, \beta\}$, and $G_1 = \phi$. Calculate H_γ^γ, I_γ^γ, H_γ^δ, and I_γ^δ. Then,

$$H_\gamma^\gamma = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}, \qquad I_\gamma^\gamma = \begin{bmatrix} 0 & -3 \\ 2 & 1 \end{bmatrix},$$

$$H_\gamma^\delta = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}, \qquad I_\gamma^\delta = \begin{bmatrix} 0 & -3 \\ 0 & 1 \end{bmatrix}.$$

Hence, $A_\gamma = \phi$, $B_\gamma = \{\delta\}$, and $C_\gamma = \{\gamma\}$.

3. Solve $P(S_0)$ for γ:

$$P(S_0)\text{: Max } z$$

subject to

$$\begin{cases} x_1 \geqq z, x_2 \geqq z \\ -2x_1 - x_2 \leqq 0 \\ -x_1 - 2x_2 - \frac{1}{2}y_1 + \frac{1}{2}y_2 \leqq 0 \\ -2x_1 - y_1 + y_2 \leqq 0 \\ -x_1 + 2x_2 + \frac{1}{2}y_1 - \frac{1}{2}y_2 \leqq 0 \\ x_1 \geqq 0, x_2 \geqq 0, z \geqq 0. \end{cases}$$

By the simplex method, we have Max $z = \infty$. Hence, γ is optimal.

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

4. For $\delta \in B_\gamma$, calculate ϕ_δ and u_δ. Then,

$$\phi_\delta = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad u_\delta = \begin{bmatrix} 0 & -2 \\ 0 & 2 \end{bmatrix}.$$

As $\phi_\delta \leq \phi_\gamma$, $\delta \in N_3$. Hence,

$$E_2 = \phi, \qquad F_2 = \{\alpha, \beta, \delta\}, \quad \text{and} \quad G_2 = \{\gamma\}.$$

5. Stop Phase I because $E_2 \cup F_2 \cup G_2 = \Pi_D$. Set $E^* = E_2$, $F^* = F_2$, and $G^* = G_2$. As $E^* = \phi$, we need not go to Phase II. Thus, we have $G = \{\gamma\}$.

References

1. Blackwell, D. (1962). Discrete dynamic programming. *Annals of Mathematical Statistics* 33: 719–726.
2. Derman, C. (1970). *Finite state Markovian decision processes.* New York: Academic Press.
3. Durinovic, S., Lee, H.M., Katehakis, M.N., & Filar, J.A. (1986). Multiobjective Markov decision process with average reward criterion. *Large Scale Systems* 10: 215–226.
4. Furukawa, N. (1980). Characterization of optimal policies in vector-valued Markovian decision processes. *Mathematics of Operations Research* 5: 271–279.
5. Furukawa, N. (1980). Vector-valued Markovian decision processes with countable state space. In R. Hartley, L.C. Thomas, & D.J. White (eds.), *Recent developments in Markov decision processes.* New York: Academic Press, pp. 205–223.
6. Ghosh, M.K. (1990). Markov decision processes with multiple costs. *Operations Research Letters* 9: 257–260.
7. Henig, M.I. (1983). Vector-valued dynamic programming. *SIAM Journal of Control and Optimization* 21: 490–499.
8. Hinderer, K. (1970). *Foundations of non-stationary dynamic programming with discrete time parameter.* Berlin: Springer-Verlag.
9. Hordijk, A. & Kallenberg, L.C.M. (1984). Constrained undiscounted stochastic dynamic programming. *Mathematics of Operations Research* 9: 276–289.
10. Iki, T. & Furukawa, N. (1984). Vector-valued Markov decision processes with average criterion. *Memoirs of Faculty of Education, Miyazaki University, Series: Natural Sciences.* 54–55: 1–10.
11. Kallenberg, L.C.M. (1983). *Linear programming and finite Markovian control problems*, Mathematical CentreTracts 148. Amsterdam: CWI.
12. Novák, J. (1989). Linear programming in vector criterion Markov and semi-Markov decision processes. *Optimization* 20: 651–670.
13. Novák, J. (1991). VMDP-Program for solving optimality problems in vector criterion Markov and semi-Markov decision processes. *Optimization* 22: 239–248.
14. Schweitzer, P.J. & Gavish, B. (1976). An optimality principle for Markovian decision processes. *Journal of Mathematical Analysis and Applications* 54: 173–184.
15. Thomas, L.C. (1983). Constrained Markov decision processes as multi-objective problems. In S. French, R. Hartley, L.C.M. Thomas, & D.J. White (eds.), *Multi-objective decision making.* New York: Academic Press, pp. 77–94.
16. Veinott, A.F., Jr. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Annals of Mathematical Statistics* 37: 1284–1294.
17. Viswanathan, S., Aggarwal, V.V., & Nair, K.P.K. (1977). Multiple criteria Markov decision processes. *TIMS Studies in the Management Sciences* 6: 263–272.
18. Wakuta, K. (1992). Optimal stationary policies in the vector-valued Markov decision process. *Stochastic Processes and their Applications* 42: 149–156.

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press

19. Wakuta, K. (1995). Vector-valued Markov decision processes and the systems of linear inequalities. *Stochastic Processes and Their Applications* 56: 159–169.
20. Wakuta, K. & Togawa, K. (1998). Solution procedures for multi-objective Markov decision processes. *Optimization* 43: 29–46.
21. White, C.C. & Kim, K.W. (1980). Solution procedures for vector criterion Markov decision processes. *Large Scale Systems* 1: 129–140.
22. White, D.J. (1982). Multi-objective infinite-horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications* 89: 639–647.
23. Yu, P.L. (1985). *Multiple-criteria decision making*. New York: Plenum Press.

https://doi.org/10.1017/S0269964800144092 Published online by Cambridge University Press