Linear Dependence of Stationary Distributions in Ergodic Markov Decision Processes

Ronald Ortner

Department Mathematik und Informationstechnologie, Montanuniversität Leoben

Abstract

In ergodic MDPs we consider stationary distributions of policies that coincide in all but n states, in which one of two possible actions is chosen. We give conditions and formulas for linear dependence of the stationary distributions of $n+2$ such policies, and show some results about combinations and mixtures of policies.

Key words: Markov decision process; Markov chain; stationary distribution
1991 MSC: Primary: 90C40, 60J10; Secondary: 60J20

1 Introduction

Definition 1 *A* Markov decision process *(MDP) $\mathcal{M}$ on a (finite) set of* states S *with a (finite) set of* actions A *available in each state* $\in S$ *consists of*

(i) an initial distribution μ_0 that specifies the probability of starting in some state in S,
(ii) the transition probabilities $p_a(s,s')$ that specify the probability of reaching state s' when choosing action a in state s, and
(iii) the payoff distributions with mean $r_a(s)$ that specify the random reward for choosing action a in state s.

A (deterministic) policy *on $\mathcal{M}$ is a mapping $\pi : S \to A$.*

Note that each policy π induces a Markov chain on $\mathcal{M}$. We are interested in MDPs, where in each of the induced Markov chains any state is reachable from any other state.

Email address: `ronald.ortner@unileoben.ac.at` (Ronald Ortner).
[1] Ronald Ortner, Department Mathematik und Informationstechnologie, Montanuniversität Leoben, Franz-Josef-Straße 18, 8700-Leoben, Austria

Definition 2 *An MDP $\mathcal{M}$ is called* ergodic, *if for each policy π the Markov chain induced by π is ergodic, i.e. if the matrix $P = (p_{\pi(i)}(i,j))_{i,j\in S}$ is irreducible.*

It is a well-known fact (cf. e.g. [1], p.130ff) that for an ergodic Markov chain with transition matrix P there exists a unique invariant and strictly positive distribution μ, such that independent of the initial distribution μ_0 one has $\mu_n = \mu_0 \bar{P}_n \to \mu$, where $\bar{P}_n = \frac{1}{n}\sum_{j=1}^{n} P^j$. Thus, given a policy π on an ergodic MDP that induces a Markov chain with invariant distribution μ, the *average reward* of that policy can be defined as

$$V(\pi) := \sum_{s\in S} \mu(s) r_{\pi(s)}(s).$$

A policy π° is called *optimal*, if for all policies π: $V(\pi) \leq V(\pi^\circ)$. It can be shown ([2], p.360ff) that the optimal value $V(\pi^\circ)$ cannot be increased by allowing time-dependent policies, as there is always a *deterministic* policy that gains optimal average reward.

2 Main Theorem and Proof

Given n policies $\pi_1, \pi_2, \ldots, \pi_n$ we say that another policy π is a *combination* of $\pi_1, \pi_2, \ldots, \pi_n$, if for each state s one has $\pi(s) = \pi_i(s)$ for some i.

Theorem 3 *Let $\mathcal{M}$ be an ergodic MDP and π_1, π_2,…,π_{n+1} pairwise distinct policies on $\mathcal{M}$ that coincide on all but n states s_1, s_2, …, s_n. In these states each policy applies one of two possible actions, i.e. we assume that for each i and each j either $\pi_i(s_j) = 0$ or $\pi_i(s_j) = 1$. Let π be a combination of the policies π_1, π_2,…,π_{n+1}. We may assume without loss of generality that $\pi(s_j) = 1$ for all j by swapping the names of the actions correspondingly. Let μ_i be the stationary distribution of policy π_i $(i = 1, \ldots, n+1)$, and let S_n be the set of permutations of the elements $\{1, \ldots, n\}$. Then setting*

$$\Gamma_k := \{\gamma \in S_{n+1} \,|\, \gamma(k) = n+1 \textit{ and } \pi_j(s_{\gamma(j)}) = 0 \textit{ for all } j \neq k\}$$

and for all $s \in S$

$$\mu(s) := \frac{\sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma_k} sgn(\gamma)\, \mu_k(s) \prod_{\substack{j=1\\ j\neq k}}^{n+1} \mu_j(s_{\gamma(j)})}{\sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma_k} sgn(\gamma)\, \prod_{\substack{j=1\\ j\neq k}}^{n+1} \mu_j(s_{\gamma(j)})},$$

μ is the stationary distribution of π, provided that $\mu \neq \mathbf{0}$.

For clarification of Theorem 3, we proceed with an example.

Example 4 Let $\mathcal{M}$ be an ergodic MDP with state space $S = \{s_1, \ldots, s_N\}$ and π_{000}, π_{010}, π_{101}, π_{110} policies on $\mathcal{M}$ whose actions differ only in three states s_1, s_2 and s_3. The subindices of a policy correspond to the word $\pi(s_1)\pi(s_2)\pi(s_3)$, so that e.g. $\pi_{010}(s_1) = \pi_{010}(s_3) = 0$ and $\pi_{010}(s_2) = 1$. Now let μ_{000}, μ_{010}, μ_{101}, and μ_{110} be the stationary distributions of the respective policies. As for $n = 3$ the possibility of $\mu = \mathbf{0}$ can be excluded (see sufficient condition (i) of Remark 6 below), by Theorem 3 we may calculate the distributions of all other policies that play in states s_1, s_2, s_3 action 0 or 1, and coincide with the above mentioned policies in all other states. In order to obtain e.g. the stationary distribution μ_{111} of policy π_{111} in an arbitrary state s, first we have to determine the sets Γ_{000}, Γ_{010}, Γ_{101}, and Γ_{110}. This can be done by interpreting the subindices of our policies as rows of a matrix. In order to obtain Γ_k one cancels row k and looks for all possibilities in the remaining matrix to choose three 0s that neither share a row nor a column:

$$
\begin{array}{ccc} \cancel{0\;0\;0} \\ \underline{0}\;1\;0 \\ 1\;\underline{0}\;1 \\ 1\;1\;\underline{0} \end{array}
\qquad\qquad
\begin{array}{ccc} \underline{0}\;0\;0 \\ \cancel{0\;1\;0} \\ 1\;\underline{0}\;1 \\ 1\;1\;\underline{0} \end{array}
\qquad\qquad
\begin{array}{ccc} 0\;\underline{0}\;0 \\ \underline{0}\;1\;0 \\ \cancel{1\;0\;1} \\ 1\;1\;\underline{0} \end{array}
\qquad\qquad
\begin{array}{ccc} \underline{0}\;0\;0 \\ 0\;1\;\underline{0} \\ 1\;\underline{0}\;1 \\ \cancel{1\;1\;0} \end{array}
\qquad\qquad
\begin{array}{ccc} 0\;0\;\underline{0} \\ \underline{0}\;1\;0 \\ 1\;\underline{0}\;1 \\ \cancel{1\;1\;0} \end{array}
$$

Each of the matrices now corresponds to a permutation in Γ_k, where k corresponds to the cancelled row. Thus Γ_{000}, Γ_{010} and Γ_{101} contain only a single permutation, while Γ_{110} contains two. The respective permutation can be read off each matrix as follows: note for each row one after another the position of the chosen 0, and choose $n+1$ for the cancelled row. Thus the permutation for the third matrix is $(2, 1, 4, 3)$. Now for each of the matrices one has a term that consists of four factors (one for each row). The factor for a row j is $\mu_j(s')$, where $s' = s$ if row j was cancelled (i.e. $j = k$), or equals the state that corresponds to the column of row j in which the 0 was chosen. Thus for the third matrix above one gets $\mu_{000}(s_2)\mu_{010}(s_1)\mu_{101}(s)\mu_{110}(s_3)$. Finally, one has to consider the sign for each of the terms which is the sign of the corresponding permutation. Putting all together, normalizing the output vector and abbreviating $a_i := \mu_{000}(s_i)$, $b_i := \mu_{010}(s_i)$, $c_i := \mu_{101}(s_i)$, and $d_i := \mu_{110}(s_i)$, one obtains for all states s_i $(i = 1, \ldots, N)$

$$\mu_{111}(s_i) = \frac{a_i b_1 c_2 d_3 - a_1 b_i c_2 d_3 - a_2 b_1 c_i d_3 + a_1 b_3 c_2 d_i - a_3 b_1 c_2 d_i}{b_1 c_2 d_3 - a_1 c_2 d_3 - a_2 b_1 d_3 + a_1 b_3 c_2 - a_3 b_1 c_2}.$$

Theorem 3 can be obtained from the following more general result where the stationary distribution of a randomized policy is considered.

Theorem 5 *Under the assumptions of Theorem 3, the stationary distribution μ of the policy π that plays in state s_i $(i = 1, \ldots, n)$ action 0 with probability*

$\lambda_i \in [0,1]$ *and action 1 with probability* $(1-\lambda_i)$ *is given by*

$$\mu(s) = \frac{\sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} sgn(\gamma)\, \mu_k(s) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j)}{\sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} sgn(\gamma)\, \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j)},$$

provided that $\mu \neq \boldsymbol{0}$*, where* $\Gamma'_k := \{\gamma \in S_{n+1} \,|\, \gamma(k) = n+1\}$ *and*

$$f(i,j) := \begin{cases} \lambda_i\, \mu_j(s_i), & \text{if } \pi_j(i) = 1, \\ (\lambda_i - 1)\, \mu_j(s_i), & \text{if } \pi_j(i) = 0. \end{cases}$$

Theorem 3 follows from Theorem 5 by simply setting $\lambda_i = 0$ for $i = 1,\dots,n$.

Proof of Theorem 5 Let $S = \{1,2,\dots,N\}$ and assume that $s_i = i$ for $i = 1,2,\dots,n$. We denote the probabilities associated with action 0 with $p_{ij} := p_0(i,j)$ and those of action 1 with $q_{ij} := p_1(i,j)$. Furthermore, the probabilities in the states $i = n+1,\dots,N$, where the policies $\pi_1,\dots,\pi_{n+1}$ coincide, are written as $p_{ij} := p_{\pi_k(i)}(i,j)$ as well. Now setting

$$\nu_s := \sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \mathrm{sgn}(\gamma)\, \mu_k(s) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j)$$

and $\nu := (\nu_s)_{s\in S}$ we are going to show that $\nu P_\pi = \nu$, where P_π is the probability matrix of the randomized policy π. Since the stationary distribution is unique, normalization of the vector ν proves the theorem. Now

$$\begin{aligned}(\nu P_\pi)_s &= \sum_{i=1}^{n} \nu_i \Big(\lambda_i p_{is} + (1-\lambda_i) q_{is}\Big) + \sum_{i=n+1}^{N} \nu_i\, p_{is} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \mathrm{sgn}(\gamma)\, \mu_k(i) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j) \Big(\lambda_i p_{is} + (1-\lambda_i) q_{is}\Big) \\ &\quad + \sum_{i=n+1}^{N} \sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \mathrm{sgn}(\gamma)\, \mu_k(i) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j)\, p_{is}.\end{aligned}$$

Since

$$\sum_{i=n+1}^{N} \mu_k(i)\, p_{is} \;=\; \mu_k(s) - \sum_{i:\pi_k(i)=0} \mu_k(i)\, p_{is} - \sum_{i:\pi_k(i)=1} \mu_k(i)\, q_{is},$$

this gives

$$
\begin{aligned}
(\nu P_\pi)_s =& \sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \operatorname{sgn}(\gamma) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j) \Big(\sum_{i=1}^{n} \mu_k(i) \left(\lambda_i p_{is} + (1-\lambda_i) q_{is}\right) \\
&\qquad + \mu_k(s) - \sum_{i:\pi_k(i)=0} \mu_k(i)\, p_{is} - \sum_{i:\pi_k(i)=1} \mu_k(i)\, q_{is} \Big) \\
=& \nu_s + \sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \operatorname{sgn}(\gamma) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j) \Big(\sum_{i:\pi_k(i)=0} \mu_k(i)\, (\lambda_i - 1)(p_{is} - q_{is}) \\
&\qquad + \sum_{i:\pi_k(i)=1} \mu_k(i)\, \lambda_i (p_{is} - q_{is}) \Big) \\
=& \nu_s + \sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \operatorname{sgn}(\gamma) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j) \sum_{i=1}^{n} (p_{is} - q_{is}) f(i,k) \\
=& \nu_s + \sum_{i=1}^{n} (p_{is} - q_{is}) \sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \operatorname{sgn}(\gamma)\, f(i,k) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j)
\end{aligned}
$$

Now it is easy to see that $\sum_{k=1}^{n+1} \sum_{\gamma\in\Gamma'_k} \operatorname{sgn}(\gamma)\, f(i,k) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j) = 0$: fix k and some permutation $\gamma \in \Gamma'_k$, and let $l := \gamma^{-1}(i)$. Then there is exactly one permutation $\gamma' \in \Gamma'_l$, such that $\gamma'(j) = \gamma(j)$ for $j \neq k, l$ and $\gamma'(k) = i$. The pairs (k,γ) and (l,γ') correspond to the same summands

$$
f(i,k) \prod_{\substack{j=1\\ j\neq k}}^{n+1} f(\gamma(j),j) \;=\; f(i,l) \prod_{\substack{j=1\\ j\neq l}}^{n+1} f(\gamma'(j),j)
$$

– yet, since $\operatorname{sgn}(\gamma) = -\operatorname{sgn}(\gamma')$, they have different sign and cancel out each other, which finishes the proof. □

Remark 6 *The condition $\mu \neq \boldsymbol{0}$ in the theorems is (unfortunately) necessary, as there are some degenerate cases where the formula gives $\boldsymbol{0}$. Choose e.g. an MDP and policies that violate condition (iii) below, such that the stationary distribution of each policy is the uniform distribution over the states. Then in the setting of Theorem 3, one obtains $\mu = \boldsymbol{0}$. It is an open question whether these "singularities" can be characterized or whether there is a more general formula that avoids them. At the moment, only the following sufficient conditions guaranteeing $\mu \neq \boldsymbol{0}$ are available:*

(i) $n < 4$,
(ii) the distributions μ_i are linearly independent,
(iii) the $n \times (n+1)$ matrix $\left(\pi_j(s_i)\right)_{ij}$ has rank n.

While (ii) is trivial, (i) and (iii) can be verified by noting that $\mu(s)$ can be *written as the following determinant:*

$$\mu(s) = \begin{vmatrix} f(1,1) & f(1,2) & \cdots & f(1,n+1) \\ f(2,1) & f(2,2) & \cdots & f(2,n+1) \\ \vdots & \vdots & \ddots & \vdots \\ f(n,1) & f(n,2) & \cdots & f(n,n+1) \\ \mu_1(s) & \mu_2(s) & \cdots & \mu_{n+1}(s) \end{vmatrix}$$

Sometimes (iii) can be guaranteed by simply swapping names of the actions, but there are cases where this doesn't work.

3 Applications

3.1 Not All Combined Policies are Worse

A consequence of Theorem 3 is that given two policies π_1, π_2, not all combined policies can be worse than π_1 and π_2. For the case where π_1 and π_2 differ only in two states this is made more precise in the following proposition.

Proposition 7 *Let $\mathcal{M}$ be an ergodic MDP. Let $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$ be four policies on $\mathcal{M}$ that coincide in all but two states s_1, s_2, in which either action 0 or 1 (according to the subindices) is chosen. We denote the average rewards of the policies by $V_{00}, V_{01}, V_{10}, V_{11}$. Let $\alpha, \beta \in \{0,1\}$ and set $\neg z := 1-z$. Then it cannot be the case that both, $V_{\alpha\beta} > V_{\neg\alpha,\beta}, V_{\alpha,\neg\beta}$ and $V_{\neg\alpha,\neg\beta} \geq V_{\neg\alpha,\beta}, V_{\alpha,\neg\beta}$. Analogously, it cannot hold that both, $V_{\alpha\beta} < V_{\neg\alpha,\beta}, V_{\alpha,\neg\beta}$ and $V_{\neg\alpha,\neg\beta} \leq V_{\neg\alpha,\beta}, V_{\alpha,\neg\beta}$.*

Proof Let the invariant distributions of the policies $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$ be $\mu_{00} = (a_i)_{i\in S}$, $\mu_{01} = (b_i)_{i\in S}$, $\mu_{10} = (c_i)_{i\in S}$, $\mu_{11} = (d_i)_{i\in S}$. Assume without loss of generality that $\alpha = \beta = 0$ and $V_{00} \leq V_{11}$. We show that if $V_{00} > V_{01}, V_{10}$, then $V_{00} > V_{11}$, contradicting our assumption.

In the following, we write for the rewards of the policy π_{00} simply r_i instead of $r_{\pi_{00}(i)}(i)$. For the deviating rewards in state s_1 under policies π_{10}, π_{11} and in state s_2 under π_{01}, π_{11} we write r'_1 and r'_2, respectively. Then we have

$$V_{00} = \sum_{i\in S} a_i r_i, \qquad V_{01} = b_2 r'_2 + \sum_{i\in S\setminus\{2\}} b_i r_i, \qquad V_{10} = c_1 r'_1 + \sum_{i\in S\setminus\{1\}} c_i r_i,$$
$$V_{11} = d_1 r'_1 + d_2 r'_2 + \sum_{i\in S\setminus\{1,2\}} d_i r_i.$$

If we now assume that $V_{00} > V_{01}, V_{10}$, the first three equations yield

$$b_2 r_2' \,<\, a_2 r_2 + \sum_{i \in S \setminus \{2\}} (a_i - b_i) r_i \quad \text{and} \quad c_1 r_1' \,<\, a_1 r_1 + \sum_{i \in S \setminus \{1\}} (a_i - c_i) r_i, \tag{1}$$

while applying Theorem 3 for $n = 2$ together with sufficient condition (i) of Remark 6 to the fourth equation gives

$$V_{11} = \frac{1}{D}\Big(a_2 b_1 c_1 r_1' + a_1 b_2 c_2 r_2' + \sum_{i \in S \setminus \{1,2\}} (a_2 b_1 c_i - a_i b_1 c_2 + a_1 c_2 b_i) r_i \Big),$$

where $D = a_2 b_1 - b_1 c_2 + a_1 c_2$. Substituting according to (1) then yields

$$\begin{aligned}
V_{11} < {} & \frac{a_2 b_1}{D}\Big(a_1 r_1 + \sum_{i \in S \setminus \{1\}} (a_i - c_i) r_i \Big) + \frac{a_1 c_2}{D}\Big(a_2 r_2 + \sum_{i \in S \setminus \{2\}} (a_i - b_i) r_i \Big) \\
& + \frac{a_2 b_1}{D} \sum_{i \in S \setminus \{1,2\}} c_i r_i - \frac{b_1 c_2}{D} \sum_{i \in S \setminus \{1,2\}} a_i r_i + \frac{a_1 c_2}{D} \sum_{i \in S \setminus \{1,2\}} b_i r_i \\
= {} & \frac{1}{D}\Big(a_1 a_2 b_1 r_1 + a_2 b_1 (a_2 - c_2) r_2 + a_1 a_2 c_2 r_2 + a_1 c_2 (a_1 - b_1) r_1 \\
& + (a_2 b_1 + a_1 c_2 - b_1 c_2) \sum_{i \in S \setminus \{1,2\}} a_i r_i \Big) \\
= {} & \frac{a_2 b_1 - b_1 c_2 + a_1 c_2}{D}\Big(a_1 r_1 + a_2 r_2 + \sum_{i \in S \setminus \{1,2\}} a_i r_i \Big) \;=\; V_{00}. \qquad \square
\end{aligned}$$

Corollary 8 *Let $V_{00}, V_{01}, V_{10}, V_{11}$ and α, β be as in Proposition 7. Then the following implications hold:*

(i) $V_{\alpha\beta} < V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta} \implies V_{\neg\alpha,\neg\beta} > \min(V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta})$.

(ii) $V_{\alpha\beta} > V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta} \implies V_{\neg\alpha,\neg\beta} < \max(V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta})$.

(iii) $V_{\alpha\beta} \leq V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta} \implies V_{\neg\alpha,\neg\beta} \geq \min(V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta})$.

(iv) $V_{\alpha\beta} \geq V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta} \implies V_{\neg\alpha,\neg\beta} \leq \max(V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta})$.

(v) $V_{\alpha\beta} = V_{\alpha,\neg\beta} = V_{\neg\alpha,\beta} \implies V_{\neg\alpha,\neg\beta} = V_{\alpha\beta}$.

(vi) $V_{\alpha\beta}, V_{\neg\alpha,\neg\beta} \geq V_{\alpha,\neg\beta}, V_{\neg\alpha,\beta} \implies V_{\alpha\beta} = V_{\alpha,\neg\beta} = V_{\neg\alpha,\beta} = V_{\neg\alpha,\neg\beta}$.

Proof (i)–(iv) are mere reformulations of Proposition 7, while (vi) is an easy consequence. Thus let us consider (v). If $V_{\neg a,\neg b}$ were $< V_{a,\neg b} = V_{\neg a,b}$, then by Proposition 7, $V_{ab} > \min(V_{\neg a,b}, V_{a,\neg b})$, contradicting our assumption. Since a similar contradiction crops up if we assume that $V_{\neg a,\neg b} > V_{a,\neg b} = V_{\neg a,b}$, it follows that $V_{\neg a,\neg b} = V_{a,\neg b} = V_{\neg a,b} = V_{ab}$. $\square$

Actually, Proposition 7 is a special case of the following theorem.

Theorem 9 *Let $\mathcal{M}$ be an ergodic MDP and π_1, π_2 be two policies on $\mathcal{M}$. If there is a combined policy that has lower average reward than π_1 and π_2, then* *there must be another combined policy that has higher average reward than π_1 and π_2.*

3.1.1 Proof of Theorem 9

We show the following implication, from which the theorem follows:

$(*)$ If $V(\pi_1)$, $V(\pi_2) \geq V(\pi)$ for each combination π of π_1, π_2, then $V(\pi_1) = V(\pi_2) = V(\pi)$ for all combinations π of π_1 and π_2 .

We assume without loss of generality that $V(\pi_2) \geq V(\pi_1)$. Furthermore, we ignore all states where π_1, π_2 coincide. For the remaining n states we denote the actions of π_1 by 0 and those of π_2 by 1. Thus any combination of π_1 and π_2 can be expressed as a sequence of n elements $\in \{0, 1\}$, where we assume an arbitrary order on the set of states (take e.g. the one used in the transition matrices). We now define sets of policies or sequences, respectively, as follows: First, let Θ_i be the set of policies with exactly i occurrences of 1. Then set $\Pi_0 := \Theta_0 = \{00 \ldots 0\}$, and for $1 \leq i \leq n$

$$\Pi_i := \{\pi \in \Theta_i \mid d(\pi, \pi^*_{i-1}) = 1\},$$

where d denotes the Hamming distance, and π^*_i is a (fixed) policy in Π_i with $V(\pi^*_i) = \max_{\pi \in \Pi_i} V(\pi)$. Thus, a policy is $\in \Pi_i$ if and only if it can be obtained from π^*_{i-1} by replacing a 0 with a 1.

Lemma 10 *If $n \geq 2$, then $V(\pi^*_{i-1}) \geq V(\pi^*_i)$ for $1 \leq i \leq n$.*

Proof The lemma obviously holds for $i = 1$, since $\pi^*_0 = 00 \ldots 0 = \pi_1$ and by assumption $V(\pi_1) \geq V(\pi)$ for $\pi \in \Pi_1$ (presupposed that $n \geq 2$). Proceeding by induction, let $i > 1$ and assume that $V(\pi^*_{i-2}) \geq V(\pi^*_{i-1})$. By construction of the elements in each Π_j, the policies π^*_{i-2}, π^*_{i-1} and π^*_i differ in at most two states, i.e. the situation is as follows:

$$\begin{aligned} \pi^*_{i-2} &= \ldots 0 \ldots 0 \ldots \\ \pi^*_{i-1} &= \ldots 1 \ldots 0 \ldots \\ \pi^*_{i} &= \ldots 1 \ldots 1 \ldots \\ \pi' &= \ldots 0 \ldots 1 \ldots \end{aligned}$$

Define a policy $\pi' \in \Pi_{i-1}$ as indicated above. Then $V(\pi^*_{i-2}) \geq V(\pi^*_{i-1}) \geq V(\pi')$ by induction assumption and optimality of π^*_{i-1} in Π_{i-1}. Applying (iv) of Corollary 8 yields that $V(\pi^*_i) \leq \max(V(\pi^*_{i-1}), V(\pi')) = V(\pi^*_{i-1})$. □

Since $\pi^*_0 = 00 \ldots 0 = \pi_1$ and $\pi^*_n = 11 \ldots 1 = \pi_2$, it follows from Lemma 10 that $V(\pi_1) \geq V(\pi_2)$. Together with our initial assumption that $V(\pi_2) \geq V(\pi_1)$ this gives $V(\pi_1) = V(\pi_2)$. Eventually, we are ready to prove $(*)$ by induction on the number of states n in which the policies π_1, π_2 differ. For $n = 1$ it is trivial, while for $n = 2$ there are only two combinations of $\pi_1 = \pi_0^*$ and $\pi_2 = \pi_2^*$. One of them is identical to π_1^* and hence has average return $V(\pi_1)$, so that the other one must have average return $V(\pi_1)$ due to Corollary 8 (v).

Thus, let us assume that $n > 2$ and $V(\pi_1), V(\pi_2) \geq V(\pi)$ for each combination π of π_1, π_2. Then we have already shown that the policies π_i^* and hence in particular $\pi_1^* = 00\ldots010\ldots0$ and $\pi_{n-1}^* = 11\ldots101\ldots1$ have average reward $V(\pi_1)$. Since π_1^* and $\pi_n^* = 11\ldots1 = \pi_2$ are policies with average reward $V(\pi_1)$ that share a common digit in some position k, we may conclude by induction assumption that all policies with a 1 in position k yield average reward $V(\pi_1)$. A similar argument applied to the policies $\pi_0^* = 00\ldots0$ and π_{n-1}^* shows that all policies with a 0 in position l (the position of the 0 in π_{n-1}^*) have average reward $V(\pi_1)$ as well. Note that by construction of the sets Π_i, $k \neq l$. Thus, we have shown that all considered policies have average reward $V(\pi_1)$, except those with a 1 in position l and a 0 in position k. However, as all policies of the form

$$\begin{array}{l}
\quad\;\; k \quad\;\; l \\
\ldots 0 \ldots 0 \ldots \\
\ldots 1 \ldots 0 \ldots \\
\ldots 1 \ldots 1 \ldots
\end{array}$$

have average reward $V(\pi_1)$, a final application of Corollary 8 (v) shows that those yield average reward $V(\pi_1)$ as well. □

3.2 Combining and Mixing Optimal Policies

An immediate consequence of Theorem 9 is that combinations of optimal policies in ergodic MDPs are optimal as well.

Theorem 11 *Let $\mathcal{M}$ be an ergodic MDP and π_1°, π_2° optimal policies on $\mathcal{M}$. Then any combination of these policies is optimal as well.*

Proof If any combination of π_1° and π_2° were suboptimal, then by Theorem 9 there would be another combination of π_1° and π_2° with average reward larger than $V(\pi_1^\circ)$, a contradiction to the optimality of π_1°. □

Obviously, if two combined optimal policies are optimal, so are combinations of an arbitrary number of optimal policies. Thus, one immediately obtains that the set of optimal policies is closed under combination.

Corollary 12 *Let $\mathcal{M}$ be an ergodic MDP. A policy π is optimal on $\mathcal{M}$ if and only if for each state s there is an optimal policy π° with $\pi(s) = \pi^\circ(s)$.*

Theorem 11 can be extended to *mixing* optimal policies, that is, our policies are not deterministic anymore, but in each state we choose an action randomly. Building up on Theorem 11, we can show that any mixture of optimal policies is optimal as well.

Theorem 13 *Let Π^* be a set of deterministic optimal policies on an ergodic MDP $\mathcal{M}$. Then any policy that chooses at each visit in each state s randomly an action a such that there is a policy $\pi \in \Pi^*$ with $a = \pi(s)$, is optimal.*

Proof By Theorem 8.9.3 of [2], the limit points of the *state-action frequencies* (cf. [2], p.399) of a random policy π^R as described in the theorem are contained in the convex hull of the stationary distributions of policies in Π^*. The optimality of π^R follows immediately from the optimality of the policies in Π^*. □

Theorems 11 and 13 can also be obtained by more elementary means than we have applied here. However, in spite of this and the usefulness of the results (see below), as far as we know there is no mention of them in the literature.

3.2.1 Some Remarks

3.2.1.1 No Contexts MDPs are usually presented as a standard example for decision processes with delayed feedback. That is, an optimal policy often has to accept locally small rewards in present states in order to gain large rewards later in future states. One may think that this induces some sort of context in which actions are optimal, e.g. that choosing a locally suboptimal action only “makes sense” in the context of heading to higher reward states. Theorem 13 however shows that this is not the case and optimal actions are rather optimal in any context.

3.2.1.2 An Application of Theorem 13 Consider an algorithm operating on an MDP that every now and then recalculates the optimal policy according to its estimates of the transition probabilities and the rewards, respectively. Sooner or later the estimates are good enough, so that the calculated policy is indeed an optimal one. However, if there is more than one optimal policy, it may happen that the algorithm does not stick to a single optimal policy but starts mixing optimal policies irregularly. Theorem 13 guarantees that the average reward of such a process again is still optimal.

3.2.1.3 Optimality is Necessary Given some policies with equal average reward V, in general, it is not the case that a combination of these policies again has average reward V, as the following example shows. Thus, optimality is a necessary condition in Theorem 11.

Example 14 Let $S = \{s_1, s_2\}$ and $A = \{a_1, a_2\}$. The transition probabilities are given by

$$(p_{a_1}(i,j))_{i,j\in S} = (p_{a_2}(i,j))_{i,j\in S} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

while the rewards are $r_{a_1}(s_1) = r_{a_1}(s_2) = 0$ and $r_{a_2}(s_1) = r_{a_2}(s_2) = 1$. Since the transition probabilities of all policies are identical, policy (a_2, a_2) with an average reward of 1 is obviously optimal. Policy (a_2, a_2) can be obtained as a combination of the policies (a_1, a_2), and (a_2, a_1), which however only yield an average reward of $\frac{1}{2}$.

3.2.1.4 Multichain MDPs Theorem 11 does not hold for MDPs that are multichain as the following simple example demonstrates.

Example 15 Let $S = \{s_1, s_2\}$ and $A = \{a_1, a_2\}$. The transition probabilities are given by

$$(p_{a_1}(i,j))_{i,j\in S} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (p_{a_2}(i,j))_{i,j\in S} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

while the rewards are $r_{a_1}(s_1) = r_{a_1}(s_2) = 1$ and $r_{a_2}(s_1) = r_{a_2}(s_2) = 0$. Then the policies $(a_1, a_1), (a_1, a_2), (a_2, a_1)$ all gain an average reward of 1 and are optimal, while the combined policy (a_2, a_2) yields suboptimal average reward 0.

3.2.1.5 Infinite MDPs Under the strong assumption that there exists a unique invariant and positive distribution for each policy, Theorems 11 and 13 also hold for MDPs with countable set of states/actions. Proofs are identical to the case of finite MDPs (with the only difference that the induction becomes transfinite). However, in general, countable MDPs are much harder to handle as optimal policies need not be stationary anymore (cf. [2], p.413f).

Acknowledgments

The author would like to thank Peter Auer and four anonymous referees for their patience and for valueable comments on prior versions of this paper. This work was supported in part by the the Austrian Science Fund FWF (S9104-N04 SP4) and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] J.G. Kemeny, J.L. Snell, and A.W. Knapp, *Denumerable Markov Chains.* Springer, 1976.

[2] M.L. Puterman, *Markov Decision Processes.* Wiley Interscience, 1994.