KYBERNETIKA — VOLUME *46* (2010), NUMBER 3, PAGES 558–570

IDENTIFICATION OF OPTIMAL POLICIES IN MARKOV DECISION PROCESSES

Karel Sladký

In this note we focus attention on identifying optimal policies and on elimination suboptimal policies minimizing optimality criteria in discrete-time Markov decision processes with finite state space and compact action set. We present unified approach to value iteration algorithms that enables to generate lower and upper bounds on optimal values, as well as on the current policy. Using the modified value iterations it is possible to eliminate suboptimal actions and to identify an optimal policy or nearly optimal policies in a finite number of steps without knowing precise values of the performance function.

Keywords: finite state Markov decision processes, discounted and average costs, elimination of suboptimal policies

Classification: 90C40, 60J10, 93E20

1. INTRODUCTION

Finding optimal policies for Markov decision chains may be a computationally difficult task for large-scale models. In this case it may be preferable if not necessary to employ only simple methods that reduce the computational burden compensated by finding either only nearly optimal policies or by identifying optimal policies without knowing exact values of the performance criterion. The origins of such approaches go back to early papers on Markov decision processes. See e. g. Grinold [3], Hastings [4], Hastings and Mello [5, 6], MacQueen [7, 8], Odoni [9], Puterman and Shin [10, 11], Sladký [13] and White [14]; these results are summarized in Chapter 6.7 of Puterman's monography [12].

Recently in a series of papers for models with compact state space it has been possible to establish that, under suitable conditions (in particular, uniqueness of the optimal policy), the value iteration procedure produces a sequence of policies that converges to the optimal policy uniformly over compact sets (see e. g. [1, 2]).

In the present paper, connections between value iterations for discounted and undiscounted models are employed and we are able to generate lower and upper bounds on optimal values. This also helps to eliminate nonoptimal actions and to identify optimal policy without knowing optimized values precisely.

2. NOTATION

In this note we consider a Markov decision chain $X = \{X_n,\ n = 0, 1, \ldots\}$ with finite state space $\mathcal{I}$, and a compact set of possible decisions (actions) $\mathcal{A}(i)$ in state $i \in \mathcal{I}$. We assume that $\mathcal{A}(i)$ is a union of a finite number of closed (bounded) intervals, possibly singletons, from $\mathbb{R}$. Supposing that in state $i \in \mathcal{I}$ action $a \in \mathcal{A}(i) \subset \mathbb{R}$ is selected, then state j is reached in the next transition with given probability $p_{ij}(a)$ and one-stage (nonnegative) cost $c_i(a) \geq 0$ will be accrued to such transition. Both $p_{ij}(a)$ and $c_i(a)$ are assumed to be continuous functions of $a \in \mathcal{A}(i)$.

A (Markovian) policy controlling the chain, $\pi = (f^0, f^1, \ldots)$, is identified by a sequence of decision vectors $\{f^n, n = 0, 1, \ldots\}$ where $f^n \in \mathcal{A}$ for every $n = 0, 1, 2, \ldots$, and the ith element of f^n, denoted $f_i^n \in \mathcal{A}(i)$, is the decision (or action) taken if $X_n = i$. Policy which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary. Stationary policy $\tilde{\pi}$ is randomized if there exist decision vectors $f^{(1)}, f^{(2)}, \ldots, f^{(m)} \in \mathcal{A}$ and on following policy $\tilde{\pi}$ we select in state i action $f_i^{(j)}$ with a given probability $\alpha_i^{(j)}$ (of course, $\alpha_i^{(j)} \geq 0$ with $\sum_{j=1}^N \alpha_i^{(j)} = 1$ for all $i \in \mathcal{I}$). Given the initial state $i \in \mathcal{I}$ any policy defines the unique probability distribution of the state-action process $(X_n, f_{X_n}^n)$.

Let $\xi_{X_0}^n(\pi) = \sum_{k=0}^{n-1} c_{X_k}(f_{X_k}^k)$ (resp. $\xi_{X_0}^{\beta,n}(\pi) = \sum_{k=0}^{n-1} \beta^k\, c_{X_k}(f_{X_k}^k)$) be the the (random) total cost (resp. total β-discounted cost) received in the n next transitions of the considered Markov chain X if policy $\pi = (f^n)$ is followed and the chain starts in state X_0. Then for the total expected cost and for the total expected discounted cost respectively, received in the n next transitions if the chain starts in state $i \in \mathcal{I}$ and policy $\pi = (f^n)$ is followed we have (E_i^π is the expectation if the process starts in state i and policy π is followed)

$$v_i^n(\pi) \ := \ \mathsf{E}\,[\xi_{X_0}^n(\pi)|X_0 = i] = \mathsf{E}_i^\pi \sum_{k=0}^{n-1} c_{X_k}(f_{X_k}^k), \tag{2.1}$$

$$v_i^{\beta,n}(\pi) \ := \ \mathsf{E}\,[\xi_{X_0}^{\beta,n}(\pi)|X_0 = i] = \mathsf{E}_i^\pi \sum_{k=0}^{n-1} \beta^k\, c_{X_k}(f_{X_k}^k) \quad \text{respectively.} \tag{2.2}$$

Since the state space is finite it will be convenient to introduce matrix notations. In particular, $P(f)$ is transition probability matrix with elements $p_{ij}(f_i)$. Then $P^m(\pi) = \prod_{n=0}^{m-1} P(f^n)$ (obviously, $P^{n+1}(\pi) = P^n(\pi)P(f^n))$), for convenience we set $P^0(\pi) = I$, the identity matrix. If $\pi \sim (f)$ (i.e. if π is stationary) then $P^m(\pi) = [P(f)]^m$, and recall that the limit matrix $P^*(f) = \lim_{m\to\infty} m^{-1} \sum_{n=0}^{m-1} [P(f)]^n$ exists and $P^*(f)\, e = e$ (e is reserved for a (column) unit vector). Markov chain is called unichain if it contains a single class of recurrent states, and possibly also transient states; hence if $P(f)$ is unichain the rows of $P^*(f)$, denoted $p^*(f)$, are identical. Similarly, $c(f^n)$ denotes the (column) vector whose ith element equals $c_i(f_i^n)$; $v^n(\pi)$ denotes the (column) vector of expected costs whose ith element equals $v_i^n(\pi)$.

Then by (2.1), (2.2) if policy $\pi = (f^n)$ is followed we immediately have for the vector of total costs $v^n(\pi)$, resp. total discounted costs $v^{\beta,n}(\pi)$, in the n next transitions

$$v^n(\pi) = \sum_{k=0}^{n-1}\prod_{j=0}^{k-1} P(f^j)c(f^k), \qquad \text{resp.} \quad v^{\beta,n}(\pi) = \sum_{k=0}^{n-1}\prod_{j=0}^{k-1} \beta^k P(f^j)c(f^k). \tag{2.3}$$

For $n \to \infty$ elements of $v^n(\pi)$ (resp. $v^{\beta,n}(\pi)$) can be typically infinite (resp. bounded by $M/(1-\beta)$ where $M = \max_i \max_k c_i(k)$). In case that π is stationary, i. e. $\pi \sim (f)$, sometimes we replace in (2.3) $v_i^n(\pi), v_i^n(\pi,\beta)$ by $v_i^n(f), v_i^n(f,\beta)$ respectively. Following stationary policy $\pi \sim (f)$ for n going to infinity there exist vectors of average costs per transition, denoted $g(f)$ (with elements $g_i(f)$ bounded by M) and vector of total discounted costs, denoted $v^\beta(f)$ with elements $v_i^\beta(f)$ being the discounted cost if the process starts in state i, where

$$g(f) \ := \ \lim_{n\to\infty} \frac{1}{n} v^n(f) = P^*(f)c(f) \tag{2.4}$$

$$v^\beta(f) \ := \ \sum_{k=0}^{\infty} [\beta\, P(f)]^k c(f) = [I - \beta P(f)]^{-1}\, c(f) = c(f) + \beta\, v^\beta(f). \tag{2.5}$$

Let for arbitrary policy $\pi = (f^n)$ $\hat{v}^\beta := \inf_\pi v^\beta(\pi)$, $\hat{g} := \inf_\pi \liminf_{n\to\infty} \frac{1}{n} v^n(\pi)$ where $\hat{v}_i^\beta$, resp. $\hat{g}_i$ (the ith element of $\hat{v}^\beta$, resp. of $\hat{g}$) is the minimal β-discounted cost, resp. minimal average cost, if the process starts in state $i \in \mathcal{I}$.

In this note we make the following general assumption.

Assumption GA. There exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{A}$, i. e. for every $f \in \mathcal{A}$ the transition probability matrix $P(f)$ is unichain (i. e. $P(f)$ has no two disjoint closed sets).

Under Assumption GA for every stationary policy $\pi \sim (f)$ the vector $g(f)$ is a constant vector with elements $\bar{g}(f)$ equal to $p^*(f)\, c(f)$.

3. PRELIMINARIES

The following facts are mostly known to workers in stochastic dynamic programming (see e. g. Puterman [12]). In particular, Fact 3.1 summarizes well-known properties of optimal discounted and average policies that can be found in the class of stationary policies. This can be verified by policy iteration. On the other hand deep results on the asymptotic behavior of minimal total costs are summarized in Fact 3.2; in Fact 3.3 we show how to employ results of Fact 3.2 for effective successive approximations of minimal average cost. Finally, in Fact 3.4 we construct simple upper and lower bounds on optimal average and discounted costs.

Fact 3.1. (i) There exists decision vector $\hat{f}^\beta \in \mathcal{A}$ along with (column) vector $\hat{v}^\beta = v(\hat{f}^\beta)$, being the unique solution of

$$v^\beta(f) = \min_{f\in\mathcal{A}} \left[c(f) + \beta P(f)\, v^\beta(f)\right]. \tag{3.1}$$

In particular, for elements of $\hat{v}^{\beta}$, denoted $\hat{v}_i^{\beta}$, we can write

$$\hat{v}_i(\beta) = \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j\in\mathcal{I}} p_{ij}(a)\hat{v}_j(\beta)\Big] = c_i(\hat{f}_i^{\beta}) + \beta \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}_i^{\beta})\hat{v}_j(\beta). \qquad (3.2)$$

(ii) If Assumption GA holds there exists decision vector $\hat{f} \in \mathcal{A}$ along with (column) vectors $\hat{w} = w(\hat{f})$ and $\hat{g} = g(\hat{f})$ (constant vector with elements $\bar{g}(f) = p^*(\hat{f})c(\hat{f})$) being the solution of

$$w(f) + g(f) = \min_{f\in\mathcal{A}} \big[c(f) + P(f)\,w(f)\big] \qquad (3.3)$$

where $w(\hat{f})$ is unique up to additive constant, and unique under the additional normalizing condition

$$P^*(f)\,w(f) = 0. \qquad (3.4)$$

In particular, for elements of $\hat{g} = g(\hat{f})$, and $\hat{w} = w(\hat{f})$, denoted $\bar{g}$ and $\hat{w}_i$, we can write

$$\hat{w}_i + \bar{g} = \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \sum_{j\in\mathcal{I}} p_{ij}(a)\hat{w}_j\Big] = c_i(\hat{f}_i) + \sum_{j\in\mathcal{I}} p_{ij}(\hat{f}_i)\hat{w}_j. \qquad (3.5)$$

Now consider the for a finite time horizon the “opposite time” orientation. In particular, let $V_i(n,\beta)$ be the minimum β discounted expected cost and $V_i(n)$ be the minimum expected cost respectively accrued in the n remaining transition if the Markov chain X is in state i. As well known $V_i(n,\beta)$ (the value iteration function) must fulfill the following dynamic programming recursion

$$V_i(n+1,\beta) = \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j\in\mathcal{I}} p_{ij}(a)V_j(n,\beta)\Big], \qquad (3.6)$$

where the initial condition $V_\ell(0,\beta)$ is usually set equal to 0 for all $\ell \in \mathcal{I}$.

Considering the vector $V(n,\beta) = [V_i(n,\beta)]$ of minimum expected discounted costs then (3.6) can be written as

$$V(n+1,\beta) = \min_{f\in\mathcal{A}} \big[c(f) + \beta P(f)V(n,\beta)\big] = c(\hat{f}(n)) + \beta P(\hat{f}(n))V(n,\beta) \qquad (3.7)$$

where

$\hat{f}(n) \in \mathcal{A}$ denotes the decision vector taken if n transitions are to be left.

Similarly, for undiscounted models $V_i(n)$ (the value iteration function) must fulfill the following dynamic programming recursion

$$V_i(n+1) = \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \sum_{j\in\mathcal{I}} p_{ij}(a)V_j(n)\Big], \qquad (3.8)$$

where the initial condition $V_\ell(0)$ is usually set equal to 0 for all $\ell \in \mathcal{I}$.

(3.8) can be written in matrix form as

$$V(n+1) = \min_{f\in\mathcal{A}} \big[c(f) + P(f)V(n)\big] = c(\hat{f}(n)) + P(\hat{f}(n))V(n) \tag{3.9}$$

where

$V(n) = [V_i(n)]$ is a column vector

$\hat{f}(n) \in \mathcal{A}$ denotes the decision vector taken if n transitions are to be left, and

$V(0)$ is arbitrary (usually we set $V(0) = 0$).

Asymptotic properties of (3.8), (3.9) were studied in many papers. For what follows we shall need the following deep result on asymptotic properties of $V(n)$.

Fact 3.2. (i) For β-discounted models it holds for any $i \in \mathcal{I}$ as $n \to \infty$

$$V_i(n,\beta) \to \hat{v}_i^\beta \qquad \text{independently of initial conditions } \ V_\ell(0,\beta) \text{ with } \ell \in \mathcal{I}. \tag{3.10}$$

(ii) If Assumption GA holds, i. e. if $P(f)$ is unichain, and moreover also aperiodic then as $n \to \infty$ it holds for every $f \in \mathcal{A}$, $i \in \mathcal{I}$

$$V_i(n) \to n\bar{g} + \hat{w}_i \qquad \text{where } \ \hat{w}_i \text{ depends on } V_\ell(0) \text{ with } \ell \in \mathcal{I}. \tag{3.11}$$

From (3.11) we can easily conclude that as $n \to \infty$ for any $i \in \mathcal{I}$

$$B_i(n) \ := \ V_i(n+1) - V_i(n) \ \to \ \bar{g} \tag{3.12}$$

$$V_i(n) - V_j(n) \ \to \ \hat{w}_i - \hat{w}_j \ =: \ \hat{w}_i^{(j)} \quad \text{independently of all } \ V_\ell(0)'\text{s}, \tag{3.13}$$

$$\text{observe that } \ \hat{w}_i^{(i)} = 0.$$

In spite of convergence of the differences of expected total cost to minimal average cost (cf. (3.12)) and to weighting coefficient being a solution to (3.3) (cf. (3.13)), finding optimal values using the dynamic programming recursions (3.8), (3.9) may be awkward since $V_i(n)$ may be large (in fact, $V_i(n)$ grows to infinity as $n \to \infty$) and no bounds on minimal average cost are obtained.

The following results will show how to overcome the above difficulties (cf. [4, 9, 13, 14]). To this end, let for arbitrary $\ell \in \mathcal{I}$

$$w(n) := V(n) - V_\ell(n)\, e, \quad \bar{g}(n+1) := [V_\ell(n+1) - V_\ell(n)], \quad g(n+1) := \bar{g}(n+1)\, e$$

$$\text{hence} \qquad B_i(n) = w_i(n+1) + \bar{g}(n+1) - w_i(n), \quad w_\ell(n) \equiv 0.$$

Fact 3.3. Using the successive approximations given by

$$w(n+1) + g(n+1) = \min_{f\in\mathcal{A}} [c(f) + P(f)\, w(n)], \quad n = 0, 1, \ldots \tag{3.14}$$

with $w_\ell(0) = 0$ and $w_i(0)$ arbitrary for $i \neq \ell$ we can conclude that as $n \to \infty$

$$w(n) \ \to \ \hat{w} \quad \text{where } \hat{w} \text{ is a solution to (3.3) } \ \text{with } \hat{w}_\ell = 0 \tag{3.15}$$

$$g(n) \ \to \ \hat{g} \tag{3.16}$$

$$\bar{g}'(n) := \max_{i\in\mathcal{I}} w_i(n) + \bar{g}(n), \quad \text{resp. } \ \bar{g}'(n) := \min_{i\in\mathcal{I}} w_i(n) + \bar{g}(n)$$

is an upper, resp. lower, bound on elements both of $\hat{g}$, $g(\hat{f}(n))$, and as $n \to \infty$

$$g''(n)\, e \to \hat{g}, \quad g'(n)\, e \to \hat{g} \quad \text{monotonously,} \tag{3.17}$$

i. e. $\{g''(n), n = 0, 1, \ldots\}$ is nonincreasing, and $\{g'(n), n = 0, 1, \ldots\}$ is nondecreasing.

Fact 3.4. Moreover, for arbitrary numbers h_j's ($j \in \mathcal{I}$) and the accompanying vector $h = [h_j]$, let

$$\begin{aligned} B_i[h] &:= \min_{a \in \mathcal{A}(i)} \Big[c_i(a) + \sum_{i \in \mathcal{I}} p_{ij}(a) h_j\Big] - h_i, \\ B_i^{\beta}[h] &:= \min_{a \in \mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{i \in \mathcal{I}} p_{ij}(a) h_j\Big] - h_i. \end{aligned}$$

Then

$$\begin{aligned} \min_{i \in \mathcal{I}} B_i[h] &\le \bar{g} \le \max_{i \in \mathcal{I}} B_i[h], && (3.18) \\ (1-\beta)^{-1} \min_{i \in \mathcal{I}} B_i^{\beta}[h] &\le \hat{v}_i(\beta) + h_i \le (1-\beta)^{-1} \max_{i \in \mathcal{I}} B_i^{\beta}[h]. && (3.19) \end{aligned}$$

To verify (3.18) recall that average cost of Markov chain with one stage cost vectors $c(f)$ and $c(f) - (I - P(f))h$ must be the same if stationary policy $\pi \sim (f)$ is followed and bounded by (3.18). Furthermore, (3.18) also follows by a direct application of Fact 3.3 (it suffices to choose in (3.9) $V(0) = h$). To verify (3.19) observe that for stationary policy $\pi \sim (f)$ by a direct calculation total discounted costs of Markov chains with one-stage cost vectors $c(f)$ and $c(f) - (I - \beta P(f))h$ differ by h and for the latter case total β-discounted costs are bounded in accordance of (3.19).

4. VALUE ITERATION METHODS: A UNIFIED APPROACH

In this section we discuss connections between value iteration methods for discounted and undiscounted models. For the discounted case we rederive so-called modified value iteration method (originally reported by MacQueen in [7]) using a simple transformation of discounted model into the undiscounted unichain case and by applying value iterations to the resulting undiscounted model.

To this end, for a fixed value of the discount factor β and arbitrary policy $\pi = (f^n)$ let (cf. (2.3), (2.5)) $v^{\beta}(\pi)$ be the vector of β-discounted costs over an infinite time horizon (with elements $v_i^{\beta}(\pi)$). Recalling (3.1), (3.2) stationary policy $\hat{\pi} \sim (\hat{f})$ minimizing discounted costs along with the vector of minimal discounted costs $\hat{v}_i^{\beta} := v_i^{\beta}(\hat{\pi})$ can be found as the (unique) solution of the following set of nonlinear equations

$$v_i^{\beta} = \min_{a \in \mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j \in \mathcal{I}} p_{ij}(a) v_j^{\beta}\Big], \qquad i \in \mathcal{I} \tag{4.1}$$

written in matrix notation as

$$v^{\beta} = \min_{f \in \mathcal{A}} [c(f) + \beta P(f) v^{\beta}]. \tag{4.2}$$

Lemma 4.1. If the Markov decision processes identified by one-stage cost vector $c(f)$ and transition probability matrix $P(f)$ starts in state $\ell \in \mathcal{I}$ then the minimal total β-discounted costs $\hat{v}^{\beta}_{\ell}$ can be calculated by means of minimal average cost of an undiscounted unichain Markov decision process identified by one-stage cost vector $c(f)$ and transition probability matrix $P^{\ell}(f) = \beta P(f) + A^{(\ell)}$ where $A^{(\ell)}$ is a square matrix such that the ℓth column is equal to $(1-\beta)$, and elements of the remaining columns equal zero. Moreover, observe that $P^{\ell}(f) = [p^{(\ell)}_{ij}(f_i)]$ is an aperiodic transition probability matrix with absorbing state ℓ and all remaining states transient.

Proof. We introduce the following notations (for $i, j, \ell \in \mathcal{I}$):

$$\begin{aligned} h^{(\ell)}_j(\pi) &:= v^{\beta}_j(\pi) - v^{\beta}_{\ell}(\pi), \qquad \hat{h}^{(\ell)}_j := \hat{v}^{\beta}_j - \hat{v}^{\beta}_{\ell} \\ p^{(\ell)}_{ij}(a) &:= \begin{cases} \beta p_{ij}(a) & \text{for } j \neq \ell \\ \beta p_{ij}(a) + (1-\beta) & \text{for } j = \ell. \end{cases} \end{aligned}$$

(Observe that $p^{(\ell)}_{ij}(a) \geq 0,\ \sum_{j \in \mathcal{I}} p^{(\ell)}_{ij}(a) = 1$ for any $i, \ell \in \mathcal{I}$ and $a \in \mathcal{A}(i)$.)

Then by (4.1) we have (recall that $h^{(\ell)}_{\ell}(\pi) = \hat{h}^{(\ell)}_{\ell} = 0$ for any $i, j, \ell \in \mathcal{I}$)

$$\hat{v}^{\beta}_i - \hat{v}^{\beta}_{\ell} = \min_{a \in \mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j \in \mathcal{I}} p_{ij}(a)(\hat{v}^{\beta}_j - \hat{v}^{\beta}_{\ell}) \Big] - (1-\beta)\hat{v}^{\beta}_{\ell} \tag{4.3}$$

$$\Updownarrow$$

$$\hat{v}^{\beta}_{\ell} = \frac{1}{1-\beta} \min_{a \in \mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j \in \mathcal{I}} p_{ij}(a)\hat{h}^{(\ell)}_j + (1-\beta)\hat{h}^{(\ell)}_{\ell} - \hat{h}^{(\ell)}_i \Big] \tag{4.4}$$

$$\Updownarrow$$

$$\hat{v}^{\beta}_{\ell} = \frac{1}{1-\beta} \min_{a \in \mathcal{A}(i)} \Big[c_i(a) + \sum_{j \in \mathcal{I}} p^{(\ell)}_{ij}(a)\hat{h}^{(\ell)}_j - \hat{h}^{(\ell)}_i \Big]. \tag{4.5}$$

Moreover, let $h^{(\ell)}(\pi) = [h^{(\ell)}_j(\pi)],\ \hat{h}^{(\ell)} = [\hat{h}^{(\ell)}_j]$. Then by (4.3), (4.4) we get

$$\hat{h}^{(\ell)} + (1-\beta)\hat{v}^{\beta}_{\ell}\, e = \min_{f \in \mathcal{A}} \big[c(f) + \beta P(f)\hat{h}^{(\ell)} \big] + A^{(\ell)}\hat{h}^{(\ell)} \tag{4.6}$$

or

$$\hat{v}^{\beta}_{\ell}\, e = \frac{1}{1-\beta} \Big\{ \min_{f \in \mathcal{A}} \big[c(f) + P^{(\ell)}(f)\hat{h}^{(\ell)} \big] - \hat{h}^{(\ell)} \Big\}. \tag{4.7}$$

Since $\hat{w}_j$'s in (3.5) are unique up to additive constant, on comparing (3.5) and (4.3) – (4.5) (or (3.3) and (4.6) – (4.7)) we can easily see that (4.6) or (4.7) define minimal average cost $\bar{g} = (1-\beta)\hat{v}_{\ell}$ of a controlled chain with unichain transition probability matrix $P^{(\ell)}(f) = \beta P(f) + A^{(\ell)}$ and one-stage cost vector $c(f)$; the "weighting coefficients" $\hat{h}^{(\ell)}_i$ are unique up to additive constant. □

On finding solution of (4.4) or (4.6) by value iteration in virtue of (3.8) or (3.9), we iterate (δ_{ij} is the Kronecker symbol)

$$H_i^{(\ell)}(n+1) = \min_{a\in\mathcal{A}(i)} \Big\{c_i(a) + \sum_{j\in\mathcal{I}}[\beta p_{ij}(a) + (1-\beta)\delta_{\ell,j}]H_j^{(\ell)}(n)\Big\} \qquad (4.8)$$

or in matrix form for $H^{(\ell)}(n) = [H_i^{(\ell)}(n)]$

$$\begin{aligned} H^{(\ell)}(n+1) &= \min_{f\in\mathcal{A}}\Big\{c(f) + \big[\beta P(f) + (1-\beta)A^{(\ell)}\big]H^{(\ell)}(n)\Big\} \\ &= \min_{f\in\mathcal{A}}\Big\{c(f) + P^{(\ell)}(f)H^{(\ell)}(n)\Big\} \qquad (4.9)\end{aligned}$$

with $H^{(\ell)}(0)$ arbitrary, preferably with $H^{(\ell)}(0) = 0$.

By a direct application of Fact 3.2 we get the following

Theorem 4.2.

$$H_i^{(\ell)}(n) \to n\,(1-\beta)\hat{v}_\ell + \hat{h}_i^{(\ell)} \qquad \text{where } \hat{h}_i^{(\ell)} \text{ depends on } H^\ell(0), \qquad (4.10)$$

hence for any $i \in \mathcal{I}$

$$B_i^{(\ell)}(n) := H_i^{(\ell)}(n+1) - H_i^{(\ell)}(n) \ \to\ (1-\beta)\hat{v}_\ell^\beta \qquad (4.11)$$

$$H_i^{(\ell)}(n) - H_j^{(\ell)}(n) \ \to\ \hat{h}_i^{(\ell)} - \hat{h}_j^{(\ell)} \ =\ \hat{v}_i - \hat{v}_j \qquad (4.12)$$

$$\text{independently of all } H^\ell(0), \quad \text{observe that } \hat{h}_\ell^{(\ell)} = 0.$$

In virtue of (4.10) $H_i^{(\ell)}(n)$ grows to infinity as $n \to \infty$. However, elements of

$$h^{(\ell)}(n) := H^{(\ell)}(n) - H_\ell^{(\ell)}(n)\,e \quad \text{(observe that } h_\ell^{(\ell)}(n) \equiv 0),$$

as well as

$$\tilde{v}_\ell(n+1) := H_\ell^{(\ell)}(n+1) - H_\ell^{(\ell)}(n),$$

are bounded. Moreover, since for any real number $k(n)$

$$H_i^{(\ell)}(n+1) - H_i^{(\ell)}(n) = \min_{f\in\mathcal{A}}\Big[c(f) + (P^{(\ell)}(f) - I)(H^{(\ell)}(n) + k(n)\,e)\Big]$$

we conclude that

$$B_i^{(\ell)}(n) = h_i^{(\ell)}(n+1) + \tilde{v}_\ell^\beta(n+1) - h_i^{(\ell)}(n) = H_i^{(\ell)}(n+1) - H_i^{(\ell)}(n). \qquad (4.13)$$

In virtue of (4.9), (4.13) $h_i^{(\ell)}(n)$ can be calculated using the following dynamic programming recursions (this well corresponds to the modified value iteration algorithm reported by MacQueen in [7])

$$\begin{aligned} h_i^{(\ell)}(n+1) + \tilde{v}_\ell(n+1) &= \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \sum_{j\in\mathcal{I}\setminus\{\ell\}} p_{ij}^{(\ell)}(a) h_j^{(\ell)}(n)\Big] \\ &= \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j\in\mathcal{I}\setminus\{\ell\}} p_{ij}(a) h_j^{(\ell)}(n)\Big] \qquad (4.14) \\ \tilde{v}_\ell^\beta(n+1) &= \min_{a\in\mathcal{A}(\ell)} \Big[c_\ell(a) + \sum_{j\in\mathcal{I}\setminus\{\ell\}} p_{\ell j}^{(\ell)}(a) h_j^{(\ell)}(n)\Big] \\ &= \min_{a\in\mathcal{A}(\ell)} \Big[c_\ell(a) + \beta \sum_{j\in\mathcal{I}\setminus\{\ell\}} p_{\ell j}(a) h_j^{(\ell)}(n)\Big], \qquad (4.15) \end{aligned}$$

written in matrix form as

$$\begin{aligned} h^{(\ell)}(n+1) + \tilde{v}_\ell^\beta(n+1)\, e &= \min_{f\in\mathcal{A}} \big[c(f) + P^{(\ell)}(f) h^{(\ell)}(n)\big] \\ &= c(\hat{f}(n)) + P^{(\ell)}(\hat{f}(n)) h^{(\ell)}(n). \qquad (4.16) \end{aligned}$$

In analogy to Fact 3.3 we arrive at

Lemma 4.3. Using the successive approximations given by (4.14) – (4.15), or in a more compact form by (4.16), we can conclude that as $n \to \infty$

$$h^{(\ell)}(n) \;\to\; \hat{h}^{(\ell)} \quad \text{where } \hat{h}^{(\ell)} \text{ is a solution to (4.6) with } \hat{h}_\ell^{(\ell)} = 0 \qquad (4.17)$$

$$\tilde{v}_\ell(n) \;\to\; (1-\beta)\,\hat{v}_\ell. \qquad (4.18)$$

In addition,

$$v_\ell''(n) := \frac{1}{1-\beta} \max_{i\in\mathcal{I}} B_i^{(\ell)}(n), \qquad \text{resp.} \quad v_\ell'(n) := \frac{1}{1-\beta} \min_{i\in\mathcal{I}} B_i^{(\ell)}(n) \qquad (4.19)$$

is an upper, resp. lower, bound on both of $\hat{v}_\ell^\beta$, $v_\ell^\beta(\hat{f}(n))$, and as $n \to \infty$

$$v_\ell''(n) \to \hat{v}_\ell^\beta, \quad v_\ell'(n) \to \hat{v}_\ell^\beta \quad \text{monotonously,} \qquad (4.20)$$

i. e. $\{v_\ell''(n), n = 0, 1, \ldots\}$ is nonincreasing, and $\{v_\ell'(n), n = 0, 1, \ldots\}$ is nondecreasing.

Up to now we have constructed at the nth iteration of the modified dynamic programming recursion only lower and upper bounds $v_\ell'(n)$, $v_\ell''(n)$ on $\hat{v}_\ell^\beta$ by (4.19). Since by (4.14) – (4.16) formulas for different ℓ's differ only in last term on the RHS, on employing $v_\ell'(n)$, $v_\ell''(n)$ along with $h^{(\ell)}(n)$ we can construct the corresponding bounds on each $\hat{v}_i^\beta$.

Lemma 4.4. If we calculate lower and upper bounds on $\hat{v}_\ell^\beta$ using the dynamic programming recursions (4.14) – (4.16) we can construct lower and upper bounds on all $\hat{v}_k^\beta$'s ($k \in \mathcal{I}$) converging monotonously to $\hat{v}_k^\beta$ using the formulas

$$v_k'(n) := v_\ell'(n) + h_k^{(\ell)}(n) \le \hat{v}_k^\beta \le v_k''(n) := v_\ell''(n) + h_k^{(\ell)}(n). \qquad (4.21)$$

Proof. Since for any $k, \ell \in \mathcal{I}$

$$v'_\ell(n) = \frac{1}{1-\beta} \min_{i\in\mathcal{I}} B_i[h^{(\ell)}(n)], \qquad v'_k(n) = \frac{1}{1-\beta} \min_{i\in\mathcal{I}} B_i[h^{(k)}(n)],$$
$$v''_\ell(n) = \frac{1}{1-\beta} \max_{i\in\mathcal{I}} B_i[h^{(\ell)}(n)], \qquad v''_k(n) = \frac{1}{1-\beta} \max_{i\in\mathcal{I}} B_i[h^{(k)}(n)]$$

it suffices to verify that for $h^{(\ell)} := h^{(\ell)}(n)$, $h^{(k)} := h^{(k)}(n)$ (cf. Fact 3.4)

$$\begin{aligned} B_i[h^{(\ell)}] &= \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j\in\mathcal{I}} p_{ij}(a) h_j^{(\ell)}\Big] - h_i^{(\ell)} \\ &= \min_{a\in\mathcal{A}(i)} \Big[c_i(a) + \beta \sum_{j\in\mathcal{I}} p_{ij}(a)(h_j^{(\ell)} - h_k^{(\ell)})\Big] - [h_i^{(\ell)} - h_k^{(\ell)}] - (1-\beta) h_k^{(\ell)} \\ &= B_i[h^{(k)}] - (1-\beta)\, h_k^{(\ell)}. \end{aligned} \tag{4.22}$$

(4.21) follows then immediately if we take in (4.22) minima and maxima over i. □

Summarizing we have arrived at

Theorem 4.5. Using the successive approximations given by (4.14) – (4.15), or in a more compact form by (4.16), then $v'_i(n)$, $v''_i(n)$ given by (4.19) for $i = \ell$ and by (4.21) for $i \neq \ell$ is is an upper, resp. lower, bound on both of $\hat{v}^\beta_\ell$, $v^\beta_\ell(\hat{f}(n))$, and $v'_i(n) \to \hat{v}^\beta_i$, $v''_i(n) \to \hat{v}^\beta_i$ monotonously as $n \to \infty$. Moreover, for arbitrary numbers h_j's ($j \in \mathcal{I}$) we can generate upper and lower bounds on $\hat{v}^\beta_\ell$ using (3.19).

5. IDENTIFICATION OF OPTIMAL POLICIES

In this section we employ bounds on discounted costs for eliminating suboptimal actions and identification of optimal policies without knowing its precise value. Finally, we indicate how under some additional condition this elimination procedure can also work for non-discounted models. The obtained results slightly extend the original results reported in [8].

To begin with, let us consider the most natural (0-order) bounds obtained by selecting in every state $i \in \mathcal{I}$ actions minimizing one-stage costs. In particular, let

$$c'(i) := c_i(a'_i) = \min\nolimits_{a\in\mathcal{A}(i)} c(i;a), \quad m' := \min\nolimits_{i\in\mathcal{I}} c'(i), \quad M' := \max\nolimits_{i\in\mathcal{I}} c'(i)$$

be the minimum one-stage cost in state $i \in \mathcal{I}$ and the minimum and maximum value of all minimum one-stage costs respectively.

Then following policy that selects action a'_i in state $i \in \mathcal{I}$, total β-discounted cost over an infinite time horizon must be nonsmaller then $v' := (1-\beta)^{-1}\, m'$ and nongreater than $v'' := (1-\beta)^{-1}\, M'$. Moreover, keeping such policy total β-discounted costs generated from the first transition cannot exceed the value $\beta\,(1-\beta)^{-1}\, M'$, in general, total β-discounted costs generated from the nth transition cannot exceed the value $\beta^n\,(1-\beta)^{-1}\, M'$. Hence we have arrived at

Lemma 5.1. β-optimal policy cannot select in state $i \in \mathcal{I}$ any action $a \in \mathcal{A}(i)$ such that
$c_i(a) > (1-\beta)^{-1} M' =: v''$ or $c_i(a) + \beta\,(1-\beta)^{-1} m' > c'(i) + \beta\,(1-\beta)^{-1} M' =: v''_i$.

Moreover, let $\mathcal{A}^{(1)}(i) \subset \mathcal{A}^{(0)}(i) \subset \mathcal{A}(i)$ be such that

$$c_i(a) \;\le\; \frac{1}{1-\beta} M' \quad \text{for any} \quad a \in \mathcal{A}^{(0)}(i) \tag{5.1}$$

$$c_i(a) \;\le\; c'(i) + \frac{\beta}{1-\beta}(M' - m') \quad \text{for any} \quad a \in \mathcal{A}^{(1)}(i). \tag{5.2}$$

Then actions yielding minimal β-discounted expected costs in state $i \in \mathcal{I}$ can be selected only from the set $\mathcal{A}^{(1)}(i)$.

Up to now we haven't employ finer lower and upper bounds on total expected β-discounted costs from the starting state $i \in \mathcal{I}$, denoted v'_i, v''_i. Recall that the modified value iteration method can generate the lower and upper bounds on every $\hat{v}^{\beta}_i$, $i \in \mathcal{I}$ in each iteration step, denoted by $v'_i(n)$ and $v''_i(n)$ respectively.

Supposing that the n steps of the modified value iteration procedure have been performed (cf. (4.14) – (4.17)) in virtue of (4.19), (4.21) we are able to generate lower and upper bounds on each $\hat{v}^{\beta}_i$, denoted $v'_i(n)$ and $v''_i(n)$ respectively. To eliminate suboptimal actions, starting with actions sets $\mathcal{A}^{(0)}(i)$ selecting actions given by (5.1) we define recursively action sets $\mathcal{A}^{(n)}(i)$ by (obviously, $\mathcal{A}^{(1)}(i)$ is given also by (5.4)

$$\mathcal{A}^{(n)}(i) := \Big\{ a \in \mathcal{A}^{(n-1)}(i) : c_i(a) + \beta \sum_{j \in \mathcal{I}} p_{ij}(a) v'_j(n) \le v''_i(n+1) \Big\}. \tag{5.3}$$

To verify elimination of suboptimal actions, if for some action, say $a' \in \mathcal{A}^{(n-1)}(i)$ it happens that $v''_i(n+1) < c_i(a') + \beta \sum_{j \in \mathcal{I}} p_{ij}(a') v'_j(n)$ then it must also hold $\hat{v}^{\beta}_i < c_i(a') + \beta \sum_{j \in \mathcal{I}} p_{ij}(a') \hat{v}^{\beta}_j$ that contradicts (4.1).

Hence we have arrived at

Theorem 5.2. At the nth step of the value iteration β-optimal policy can select in state $i \in \mathcal{I}$ only actions from the action set $\mathcal{A}^{(n)}(i)$ such that $\mathcal{A}^{(n+1)}(i) \subset \mathcal{A}^{(n)}(i)$ with $\mathcal{A}^{(0)}(i)$ given by (5.1) and $\mathcal{A}^{(n)}(i)$ for $n = 0, 1, \ldots$ defined recursively by (5.4). Moreover, on conditions that there exists a single optimal policy f^* and the action set is finite, the elimination procedure exclude nonoptimal policies in a finite number of steps.

Lemma 4.1 indicates that the discounted costs can be calculated as average costs after suitable transformation of transition probabilities. Observe that lower and upper bounds on discounted costs are constructed by means of estimates of discounted costs; in contrast to upper and lower bounds on optimal average cost that are calculated using some weighting coefficients. In the rest of this section we indicate how the above elimination of nonoptimal actions used for discounted model can be employed also for average cost case.

This end we make the following condition.

$$\text{There exists state } \ell \in \mathcal{I} \text{ such that } \sum_{j \in \mathcal{I}\setminus(\ell)} p_{ij}(a) \le \beta < 1 \text{ for all } a \in \mathcal{A}(i). \qquad (5.4)$$

Observe that condition (5.4) guarantees fulfilment of Assumption GA and if Assumption GA holds then (5.4) is fulfilled if transition probability matrices are replaced by a suitable product of transition probability matrices $P(f)$ with $f \in \mathcal{A}$.

Under (5.4), let $P(f) = \tilde{P}(f) + A^{(\ell)}$ where $A^{(\ell)}$ is a square matrix such that the ℓth column is equal to $(1-\beta)$ and elements of the remaining columns equal zero. Obviously, for elements of $\tilde{P}(f) = [\tilde{p}_{ij}(f_i)]$ it holds

$$\tilde{p}_{ij}(a) := \begin{cases} p_{ij}(a) \ \text{ for } \ j \ne \ell \\ p_{ij}(a) - (1-\beta) \ \text{ for } \ j = \ell. \end{cases}$$

Hence $\tilde{P}(f)\,e = \beta e$, and the average cost problem with average cost $\bar{g}(f)$ for any $f \in \mathcal{A}$ and minimal average cost $\bar{g}$ can be treated as a discounted cost model where for the β-discounted cost in state ℓ we have $v^{\beta}_{\ell}(f) = (1-\beta)^{-1}\,\bar{g}(f)$ and $\hat{v}^{\beta}_{\ell} = (1-\beta)^{-1}\,\bar{g}$.

Remark 5.3. Observe that the above suboptimal action elimination is not restricted to successive approximation, it can be also employed at each step of policy iteration methods.

ACKNOWLEDGEMENT

This work was supported by the Czech Science Foundation under Grants 402/08/0107 and 402/07/1113.

(Received April 14, 2010)

REFERENCES

[1] D. Cruz-Suárez and R. Montes-de-Oca: Uniform convergence of the value iteration policies for discounted Markov decision processes. Bol. de la Soc. Mat. Mexicana *12* (2006), 133–148.

[2] D. Cruz-Suárez, R. Montes-de-Oca, and F. Salem-Silva: Uniform approximations of discounted Markov decision processes to optimal policies. Proceedings of Prague Stochastics 2006 (M. Hušková and M. Janžura, eds.), Matfyzpress, Prague 2006, pp. 278–287.

[3] J. Grinold: Elimination of suboptimal actions in Markov decision problems. Oper. Res. *21* (1973), 848–851.

[4] N. A. J. Hastings: Bounds on the gain of a Markov decision processes. Oper. Res. *19* (1971), 240–243.

[5] N. A. J. Hastings and J. Mello: Tests for suboptimal actions in discounted Markov programming. Manag. Sci. *19* (1971), 1019–1022.

[6] N. A. J. Hastings and J. Mello: Tests for suboptimal actions in undiscounted Markov decision chains. Manag. Sci. *23* (1976), 87–91.

[7] J. MacQueen: A modified dynamic programming method for Markov decision problems. J. Math. Anal. Appl. *14* (1966), 38–43.

[8] J. MacQueen: A test of suboptimal actions in Markovian decision problems. Oper. Res. *15* (1967), 559–561.

[9] A. R. Odoni: On finding the maximal gain for Markov decision processes. Oper. Res. *17* (1969), 857–860.

[10] M. L. Puterman and M. C. Shin: Modified policy iteration algorithms for discounted Markov decision problems. Manag. Sci. *24* (1978), 1127–1137.

[11] M. L. Puterman and M. C. Shin: Action elimination procedures for modified policy iteration algorithm. Oper. Res. *30* (1982), 301–318.

[12] M. L. Puterman: Markov Decision Processes – Discrete Stochastic Dynamic Programming. Wiley, New York 1994.

[13] K. Sladký: O metodě postupných aproximací pro nalezení optimálního řízení markovského řetězce (On successive approximation method for finding optimal control of a Markov chain). Kybernetika *4* (1969), 2, 167–176.

[14] D. J. White: Dynamic programming, Markov chains and the method of successive approximation. J. Math. Anal. Appl. *6* (1963), 296–306.

Karel Sladký, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: sladky@utia.cas.cz