Average Reward Optimality in Semi-Markov Decision Processes with Costly Interventions

Karel Sladký [1]

Abstract. In this note we consider semi-Markov reward decision processes evolving on finite state spaces. We focus attention on average reward models, i.e. we establish explicit formulas for the growth rate of the total expected reward. In contrast to the standard models we assume that the decision maker can also change the running process by some (costly) intervention. Recall that the result for optimality criteria for the classical Markov decision chains in discrete- and continuous-time setting turn out to be a very specific case of the considered model. The aim is to formulate optimality conditions for semi-Markov models with intervention and present algorithmic procedures for finding optimal solutions.

Keywords: controlled semi-Markov reward processes, long run optimality, intervention of the decision maker.

JEL classification: C44, C61
AMS classification: 90C40, 60J10, 93E20

1 Formulation and Notation

Consider a controlled semi-Markov reward process $Y = \{Y(t), t \geq 0\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$ along with the embedded Markov chain $X\{X_n, n = 0, 1, \ldots\}$. The development of the process $Y(t)$ over time governed the decision maker is the following: At time $t = 0$ if $Y(0) = i$ the decision maker selects decision from a finite $\mathcal{A}_i = 1, 2, \ldots, S$ or from an infinite (compact) set $\mathcal{A}_i \equiv [0, K_i] \subset \mathbb{R}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Then state j is reached in the next transition with a given probability $p_{ij}(a)$ after random time $\eta_{ij}(a)$. Let $F_{ij}(a, \tau)$ be a non-lattice distribution function (i.e. the discrete probability distribution concentrated on a set of points of the form $a+nh$ where $h > 0$ and $n = 0, +1, -1, +2, -2, ...$) representing the probability $\mathrm{P}(\eta_{ij} \leq \tau)$. We assume that for $\ell = 1, 2, \quad 0 < d_{ij}^{(\ell)} = \int_0^\infty \tau^\ell \, \mathrm{d}F_{ij}(a, \tau) < \infty$ hence also $0 < d_i^{(\ell)} = \sum_{j=1}^N p_{ij}(a) d_{ij}^{(\ell)}(a) < \infty$. Finally, one-stage transition reward $r_{ij} > 0$ will be accrued to transition from state i to state j, and reward rate $r_i(a)$ per unit of time spent in state i is earned. We assume that each $p_{ij}(a)$ and $r_i(a)$ is a continuous function of $a \in \mathcal{A}_i$.

Moreover, since the decision maker has a complete knowledge on the development of the process over time we assume that the decision maker has an option for additional improvement of the system dynamics by paying certain amount, say $c_i(s)$, to guarantee that the system will jump from state i to state s.

If no intervention is applied the development of the considered Markov process over time is the following.
A (Markovian) policy controlling the semi-Markov process Y, say $\pi = (f^0, f^1, \ldots)$, is identified by a sequence of decision vectors $\{f^n, n = 0, 1, \ldots\}$ where $f^n \in \mathcal{F} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ for every $n = 0, 1, 2, \ldots$, and $f_i^n \in \mathcal{A}_i$ is the decision (or action) taken at the nth transition if the embedded Markov chain X is in state i. Let π^k be a sequence of decision vectors starting at the k-th transition, hence $\pi = (f^0, f^1, \ldots f^{k-1}, \pi^k)$. Policy which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary; $P(f)$ is a transition probability matrix with elements $p_{ij}(f_i)$. Stationary policy $\tilde{\pi}$ is randomized if there exist decision vectors $f^{(1)}, f^{(2)}, \ldots, f^{(m)} \in \mathcal{F}$ and on following policy $\tilde{\pi}$ we select in state i action $f_i^{(j)}$ with a given probability $\kappa_i^{(j)}$ (of course, $\kappa_i^{(j)} \geq 0$ with $\sum_{j=1}^N \kappa_i^{(j)} = 1$ for all $i \in \mathcal{I}$). For details see e.g. [1, 5, 6].

For the (random) reward earned up to time t, say $\xi(t)$ we have $\xi(t) := \left[\int_0^t r_{Y(s)} \mathrm{d}s \; + \sum_{k=0}^{N(t)} r_{Y(\tau_k^-), Y(\tau_k^+)}\right]$, with $Y(s)$, denoting the state of the system at time s, $Y(\tau_k^-)$ and $Y(\tau_k^+)$ the state just prior and after the kth jump, $N(t)$ the number of jumps up to time t, and $v_i(\pi, t) := \mathsf{E}_i^\pi \xi(t)$ denote the expected total reward of the semi-Markov process $Y(t)$ up to time t given its initial state at time $t = 0$ if policy $\pi = f^n$ is followed. Hence $g_i(\pi) = \lim_{t\to\infty} \frac{1}{t} v_i(\pi, t)$ seems to be the natural definition of average reward generated by the considered semi-Markov process.

[1]Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic, sladky@utia.cas.cz

On the other hand we can very easily calculated reward generated by the embedded Markov chain. Let ξ_n (resp. τ_n) be the cumulative reward obtained (resp. total time spent) in the n first transitions of the considered embedded Markov chain X. Since the process starts in state X_0, $\xi_n = \sum_{k=0}^{n-1}[r_{X_k} \cdot \eta_{X_k,X_{k+1}} + r_{X_k,X_{k+1}}]$ (resp. $\tau_n = \sum_{k=0}^{n-1} \eta_{X_k,X_{k+1}}$). Similarly let $\xi_{(m,n)}$ (resp. $\tau_{(m,n)}$) be reserved for the cumulative (random) reward, obtained (resp. total time spent) from the mth up to the nth transition. Obviously (we tacitly assume that $\xi_{(1,n)}$ starts in state X_1), $\xi_n = r_{X_0} \cdot \eta_{X_0,X_1} + r_{X_0,X_1} + \xi_{(1,n)}, \quad \tau_n = \eta_{X_0,X_1} + \tau_{(1,n)}$.

If the process starts in state i and policy π is followed let $x_i^\pi(n)$ (resp. $\tau_i^\pi(n)$) be the total expected reward (resp. total time spent) in the n next transitions. The growth rate of $x_i^\pi(n)$ (resp. $\tau_i^\pi(n)$) is linear in time, in particular $x_i^\pi(n) = g^\xi \cdot n + o(n)$ (resp. $\tau_i^\pi(n) = g^\tau \cdot n + o(n)$). Let $G_i(\pi, n) := \frac{x_i^\pi(n)}{\tau_i^\pi(n)}$ is the average reward in the n next transitions, so we can conclude that $G_i(\pi, n) \to G_i(\pi) = g^\xi/g^\tau$.

Comparing the average reward generated by $g_i(\pi)$ and $G_i(\pi)$ we can see that $g_i(\pi)$ represents what is usually meant by the expected average reward, that is the expected reward generated by time t. However, $G_i(\pi)$ also represents at least of some sense the average expectation. While $g_i(\pi)$ is clearly more appealing criterion, it turns that it is easier to work with $G_i(\pi)$. Fortunately, it turns out that under certain regenerative condition both criteria are equal. Roughly speaking, a sufficient condition is that for any stationary policy the resulting semi-Markov process is a regenerative process. For more details see e.g.[5] or [6].

The aim of this note is to formulate optimality conditions for semi-Markov models with additional intervention of the decision maker and present algorithmic procedures for finding optimal solutions. To this end using standard policy iteration procedures we find stationary policy yielding maximum average reward of the considered controlled semi-Markov process.

In what follows we show that under some specific conditions (e.g. if the considered Markov chain contains a single class of recurrent states) for $n \to \infty$ the asymptotic value of average reward is the same for the two definitions mentioned above.

2 Analysis of Average Reward Optimality in Semi-Markov Processes

To begin with we focus attention on average reward optimality in semi-Markov processes and present characterization of control policies by discrepancy functions. In contrast to the standard models on control of semi-Markov processes we assume that the decision maker has complete information on random times spent in each state (only on mean time not only on the mean time spent in each state) along with complete information on the progress of the controlled. On analyzing the current state of the process can decide if some (costly) intervention changing the current state of the process is suitable.

We begin our analysis with so called unichain models, i.e. when the underlying Markov chain contains a single class of recurrent states and hence the resulting average reward per unit time is independent of the starting state. Our analysis can be easily extended to a more general multichain model where average reward per time depends on the starting state and the state space can be partitioned on classes with the same average reward.

2.1 Unichain models

To begin with we make

Assumption 1. There exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{F}$.

Obviously, if Assumption 1 holds, then the resulting transition probability matrix $P(f)$ is *unichain* for every $f \in \mathcal{F}$ (i.e. $P(f)$ has no two disjoint closed sets).

At first we focus attention on the embedded Markov chains and slightly extend some results reported in [8]. To this end, on introducing for arbitrary $g, w_j \in \mathbb{R}$ $(i, j \in \mathcal{I})$ and decision $f \in \mathcal{F}$, the discrepancy functions

$$\varphi_{i,j}^{\mathrm{c}}(w^{\mathrm{c}}, g^{\mathrm{c}}, f) := d_i(f_i) \cdot r(i) + r_{ij} - w_i^{\mathrm{c}} + w_j^{\mathrm{c}} - g^{\mathrm{c}}, \tag{1}$$

$$\varphi_{i,j}^{\mathrm{t}}(w^{\mathrm{t}}, g^{\mathrm{t}}, f) := d_i(f_i) - w_i^{\mathrm{t}} + w_j^{\mathrm{t}} - g^{\mathrm{t}} \tag{2}$$

for the random reward obtained, resp. time elapsed, up to the nth transition we have

$$\xi_n = ng^{\mathrm{c}} + w^{\mathrm{c}}_{X_0} - w^{\mathrm{c}}_{X_n} + \sum_{k=0}^{n-1} \varphi^{\mathrm{c}}_{X_k,X_{k+1}}(w^{\mathrm{c}}, g^{\mathrm{c}}, f), \tag{3}$$

$$\tau_n = ng^{\mathrm{t}} + w^{\mathrm{t}}_{X_0} - w^{\mathrm{t}}_{X_n} + \sum_{k=0}^{n-1} \varphi^{\mathrm{t}}_{X_k,X_{k+1}}(w^{\mathrm{t}}, g^{\mathrm{t}}, f). \tag{4}$$

Hence by (3), (5) for the expectation of ξ_n, $\mathsf{E}^{\,\pi}_i \xi_n =: v^{\pi}_i(n)$, resp. of τ_n, with $\mathsf{E}^{\,\pi}_i \tau_n =: t^{\pi}_i(n)$, we get

$$v^{\pi}_i(n) = ng^{\mathrm{c}} + w^{\mathrm{c}}_i + \mathsf{E}^{\,\pi}_i\{\sum_{k=0}^{n-1} \varphi^{\mathrm{c}}_{X_k,X_{k+1}}(w^{\mathrm{c}}, g^{\mathrm{c}}, f) - w^{\mathrm{c}}_{X_n}\}, \tag{5}$$

$$t^{\pi}_i(n) = ng^{\mathrm{t}} + w^{\mathrm{t}}_i + \mathsf{E}^{\,\pi}_i\{\sum_{k=0}^{n-1} \varphi^{\mathrm{t}}_{X_k,X_{k+1}}(w^{\mathrm{t}}, g^{\mathrm{t}}, f) - w^{\mathrm{t}}_{X_n}\}. \tag{6}$$

Now we show how to express average reward generated by the semi-Markov process $Y(t), t \geq 0$ in terms of the embedded Markov chain X_n. Considering policy $\pi \sim (f)$, let

$$\varphi^{\mathrm{c}}_i(w^{\mathrm{c}}, g^{\mathrm{c}}, f) \;:=\; \sum_{j\in\mathcal{I}} p_{ij}(f_i)\varphi^{\mathrm{c}}_{i,j}(w^{\mathrm{c}}, g^{\mathrm{c}}, f) = \sum_{j\in\mathcal{I}} p_{ij}(f_i)[d_i(f_i)\cdot r(i) + r_{ij} - w^{\mathrm{c}}_i + w^{\mathrm{c}}_j - g^{\mathrm{c}}], \tag{7}$$

$$\varphi^{\mathrm{t}}_i(w^{\mathrm{t}}, g^{\mathrm{t}}, f) \;:=\; \sum_{j\in\mathcal{I}} p_{ij}(f_i)\varphi^{\mathrm{t}}_{i,j}(w^{\mathrm{t}}, g^{\mathrm{t}}, f) = \sum_{j\in\mathcal{I}} p_{ij}(f_i)[d_i(f_i) - w^{\mathrm{t}}_i + w^{\mathrm{t}}_j - g^{\mathrm{t}}] \tag{8}$$

It is well-known from the dynamic programming literature (cf. e.g. [1, 3, 5, 6]) that for every $f \in \mathcal{F}$ and arbitrary transition costs $s_{ij}(f) = d_i(f_i)r(i) + r_{ij}, i, j, \in \mathcal{I}$, there exist numbers $g(f)$ and $w_i(f), i \in \mathcal{I}$ (unique up to additive constant) such that

$$w_i(f) + g(f) = d_i(f_i)r(i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + w_j(f)], \;\; (i \in \mathcal{I}), \;\; \text{i.e.} \tag{9}$$

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\,\varphi^{\mathrm{c}}_{i,j}(w, g, f) = 0 \quad \text{where} \;\; \varphi^{\mathrm{c}}_{i,j}(w, g, f) := d_i(f_i)r(i) + r_{ij} - w_i(f) + w_j(f) - g(f).$$

In particular, for suitable selected $w^{\mathrm{c}}_j(f)$, resp. $w^{\mathrm{t}}_j(f)$, we have

$$v^{\pi}_i(n) = ng^{\mathrm{c}}(f) + w^{\mathrm{c}}_i(f) - \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot w^{\mathrm{c}}_j(f), \quad \text{where} \tag{10}$$

$$w^{\mathrm{c}}_i(f) + g^{\mathrm{c}}(f) = d_i(f_i)\cdot r(i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + w^{\mathrm{c}}_j(f)],$$

Similarly, for suitable selected $w^{\mathrm{t}}_j(f)$, we have

$$t^{\pi}_i(n) = ng^{\mathrm{t}} + w^{\mathrm{t}}_i - \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot w^{\mathrm{t}}_j(f), \quad \text{where} \tag{11}$$

$$w^{\mathrm{t}}_i(f)\cdot\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)} + g^{\mathrm{c}}(f) = d_i(f_i)\cdot\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)} + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot w^{\mathrm{t}}_j(f)\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)} \tag{12}$$

and by subtracting (12) from (11) we get

$$w_i(f) = \bar{r}_i(f) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)w_j(f) - d_i(f_i)g(f) \tag{13}$$

with

$$w_i(f) := w^{\mathrm{c}}_i(f) - w^{\mathrm{t}}_i(f)\cdot\frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)}, \quad g(f) := \frac{g^{\mathrm{c}}(f)}{g^{\mathrm{t}}(f)}, \quad \bar{r}_i(f) = d_i(f_i)\cdot r(i) + \sum_{j\in\mathcal{I}, j\neq i} p_{ij}(f_i)r_{ij}.$$

On introducing matrix notations

$$P(f) = [p_{ij}(f_i)], \quad D(f) = \text{diag}\,[d_i(f_i)], \ \text{(square matrices)}$$

$$\bar{r}(f) = [\bar{r}_i(f)], \quad w(f) = [w_i(f)], \quad \bar{g}(f) = [g(f)] \ \text{(column vectors)}$$

equation (11) can be written as

$$w(f) = \bar{r}(f) + P(f)w(f) - D(f)\bar{g}(f) \Rightarrow \bar{g}(f) = D^{-1}(f)\bar{r}(f) + [D^{-1}(f)P(f) - I] \cdot w(f)]. \tag{14}$$

Let

$$\tilde{r}(f) := D^{-1}(f)\bar{r}(f), \qquad \tilde{w}(f) := D^{-1}(f)w(f), \qquad \tilde{P}(f) := D^{-1}(f) \cdot P(f) \cdot D(f)$$

Then

$$\bar{g}(f) = \bar{r}(f) + [\bar{P}(f) - I] \cdot \bar{w}(f)]. \tag{15}$$

and for elements of $\tilde{r}(f),\ \tilde{w}(f),\ \tilde{P}(f)$ we have

$$\tilde{r}_i(f) = \bar{r}(i) + [d_i(f_i)]^{-1}r_{ij}, \quad \tilde{p}_{ij}(f_i) := p_{ij}(f_i)\frac{[d_j(f_j)]}{[d_i(f_i)]}, \quad \tilde{w}_i(f) := [d_i(f_i)]^{-1}w_i(f)$$

(observe that $\bar{g}(f)$ is a constant vector with elements $g(f) = \frac{g^c(f)}{g^t(f)}$).

In particular, let us consider continuous-time Markov decision chain with transition intensities $\mu_{ij}(f_i)$, where $\sum_{j\in\mathcal{I},j\neq i}\mu_{ij}(f_i) = -\mu_{ii}(f_i)$ and $\mu_i(f_i) = -\mu_{ii}(f_i)$ is the is the intensity of jumps from state i. Obviously, this is a very special case of semi-Markov processes with transition probabilities $p_{ij}(f) = \frac{\mu_{ij}(f_i)}{\mu_i(f_i)}$, and expected holding time $d_i(f_i) = \frac{1}{\mu_i(f_i)}$ in state i. Then on replacing in (14) transition probabilities and expected holding times by transition intensities for the average reward per unit of time of the considered continuous-time Markov process we conclude that

$$g(f) = r(i) + \sum_{j\neq i}\mu_{ij}(f_i)r_{ij} + \sum_j \mu_{ij}(f_i)w_j(f) \tag{16}$$

and policy $f \in \mathcal{F}$ appears in equation for average reward of a continuous time Markov reward chain (cf. e.g. [3]).

2.2 Multichain models

Considering transition probability matrix $P(f)$ for fixed $f \in F$, as well-known it is possible decompose the considered Markov chain such that

$$P(f) = \begin{bmatrix} P_{00}(f) & P_{01}(f) & P_{02}(f) & \dots & P_{0r}(f) \\ 0 & P_{11}(f) & 0 & \dots & 0 \\ 0 & 0 & P_{22}(f) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & \dots & P_{rr}(f) \end{bmatrix} \tag{17}$$

where existence of at least one recurrent state accessible from any state of $P_{ii}(f)$ (cf. Assumption 1) is fulfilled for diagonal submatrices $P_{ii}(f)$, $i = 1, \dots, r$, and $P_{00}(f)$ contains transient states that have access at least to two diagonal submatrices $P_{ii}(f)$, $i = 1, \dots, r = r(f)$.Observe that each diagonal block of $P(f)$ contains a single class of recurrent states possibly along with transient states accessible only to recurrent states of the same diagonal block.

In what follows we show that the decomposition depicted in (17) can be extended to the set of all matrices $P(f)$ with $f \in \mathcal{F}$. To this end, on recalling the notions of accessibility and communication for states of Markov chains, on considering stationary policies, say $f^{(1)}, f^{(2)}$, then for any $\kappa_i \in (0, 1)$ there exists $f \in \mathcal{F}$ such that for any $i, j \in \mathcal{I}$ it holds $p_{ij}(f) = \kappa_i p_{ij}(f^{(1)}) + (1 - \kappa_i)p_{ij}(f^{(2)})$. Hence any recurrent class of $P(f^{(1)})$ or $P(f^{(2)})$ must be contained in some recurrent class of $P(f)$.

Repeating this analysis for all admissible stationary policies we can conclude existence of (possibly randomized)

stationary policy, say $\tilde{f}$, such that

$$P(\tilde{f}) = \begin{bmatrix} P_{00}(\tilde{f}) & P_{01}(\tilde{f}) & P_{02}(\tilde{f}) & \dots & P_{0r}(\tilde{f}) \\ 0 & P_{11}(\tilde{f}) & 0 & \dots & 0 \\ 0 & 0 & P_{22}(\tilde{f}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & \dots & P_{rr}(\tilde{f}) \end{bmatrix} \tag{18}$$

with $r = \min r(\tilde{f})$ for any $f \in \mathcal{F}$.

Recall that Assumption 1 of Section 2.1 is fulfilled for each diagonal submatrix $P_{ii}(\tilde{f})$, $i = 1, \dots, r$ and $P_{00}(\tilde{f})$ contains transient states that have access at least to two diagonal submatrices $P_{ii}(\tilde{f})$, $i = 1, \dots, r = r(\tilde{f})$. Moreover observe that for $i = 1, \dots, r$ submatrix $P_{ii}(\tilde{f})$, contains recurrent class that is final, i.e. elements of this class have no access to states not belonging to this class, in contrast to elements of $P_{00}(\tilde{f})$. Employing the decomposition according to (18) we can employ policy iteration to find stationary policy, say $\hat{\pi} \sim \hat{f}$, yielding maximal average reward, say g_i^*, for elements of every diagonal submatrix $P_{ii}(\tilde{f})$. For details see [3].

Using this approach we finally arrive to the following decomposition

$$P(\hat{f}) = \begin{bmatrix} P_{00}(\hat{f}) & P_{01}(\hat{f}) & P_{02}(\hat{f}) & \dots & P_{0r}(\hat{f}) \\ 0 & P_{11}(\hat{f}) & 0 & \dots & 0 \\ 0 & 0 & P_{22}(\hat{f}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & \dots & P_{rr}(\hat{f}) \end{bmatrix} \tag{19}$$

with $r = \min r(\hat{f})$ for any $f \in \mathcal{F}$.

3 Improving Reward Optimality by Decision Maker Interventions

Stationary policy $\hat{\pi} \sim \hat{f}$ constructed in the previous section maximizes long-run average reward of the considered semi-Markov process. Since the decision maker has complete information on the current states in any time instant, and can change the action with respect the current state of the process, the following question can be raised:

Supposing that the decision maker considers that the current state of the process is suitable, is it suitable to change the current state by transfer the process to a more suitable state. If course, such a transfer is costly and the following question can be raised: Is it suitable for a given penalty cost to transfer the process to another state. To this end, we construct an improved stationary policy and compare long-run average reward of the original and improved policies.

Illustrative example.

Consider the controlled process with 6 states and suppose that equation (19) can be decomposed such that

$$P(\hat{f}) = \begin{bmatrix} p_{11}(\hat{f}) & p_{12}(\hat{f}) & p_{13}(\hat{f}) & p_{14}(\hat{f}) & p_{15}(\hat{f}) & p_{16}(\hat{f}) \\ p_{21}(\hat{f}) & p_{22}(\hat{f}) & p_{23}(\hat{f}) & p_{24}(\hat{f}) & p_{25}(\hat{f}) & p_{26}(\hat{f}) \\ 0 & 0 & p_{33}(\hat{f}) & p_{34}(\hat{f}) & 0 & 0 \\ 0 & 0 & p_{43}(\hat{f}) & p_{44}(\hat{f}) & 0 & 0 \\ 0 & 0 & 0 & 0 & p_{55}(\hat{f}) & p_{56}(\hat{f}) \\ 0 & 0 & 0 & 0 & p_{56}(\hat{f}) & p_{66}(\hat{f}) \end{bmatrix} \tag{20}$$

Using the decomposition according to (17), optimal policy contains along with transient states also two (final) classes of recurrent states, as shown in the following display.

In particular,

$$P(\hat{f}) = \begin{bmatrix} P_{00}(\hat{f}) & P_{01}(\hat{f}) & P_{02}(\hat{f}) \\ 0 & P_{11}(\hat{f}) & 0 \\ 0 & 0 & P_{22}(\hat{f}) \end{bmatrix} \tag{21}$$

where

$$P_{00}(\hat{f}) = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}, P_{01}(\hat{f}) = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 0 \end{bmatrix}, P_{02}(\hat{f}) = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 0 \end{bmatrix}, P_{11}(\hat{f}) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{6} & \frac{5}{6} \end{bmatrix}, P_{22}(\hat{f}) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

We assume that the submatrix $P_{11}(\hat{f})$, resp. submatrix $P_{22}(\hat{f})$, generates average reward $G_1(\hat{f})$, resp. $G_2(\hat{f})$. Hence if the process starts in state 1 or 2, the resulting average reward is a suitable linear combination of the values of average rewards $G_1(\hat{f})$ and $G_2(\hat{f})$.

Supposing that the considered process starts in state 1 or 2, if $G_1(\hat{f}) > G_2(\hat{f})$ it is possible by the decision maker's intervention to stop reaching states 5 and 6 by changing submatrices $P_{01}(\hat{f})$, $P_{01}(\hat{f})$ to

$$P_{01}(\hat{f}) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad P_{02}(\hat{f}) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

References

[1] Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control,* Volume 2, Third Edition. Belmont, Mass.: Athena Scientific.

[2] Gantmakher, F. R. (1959). *The Theory of Matrices.* London: Chelsea.

[3] Howard, R. A. (1960). *Dynamic Programming and Markov Processes.* Cambridge, Mass.: MIT Press.

[4] Howard, R. A. & Matheson, J. (1972). Risk-sensitive Markov decision processes, *Manag. Sci.*, 23 , 356–369.

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

[6] Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming.* New York: Academic Press.

[7] Sladký, K. (1973). Necessary and sufficient optimality conditions for average reward of controlled Markov chains, *Kybernetika*, 9, 124–137.

[8] Sladký, K. (2005). On mean reward variance in semi-Markov processes, *Math. Methods Oper. Res.*, 62, 387–39

[9] Sladký, K. (2012). Risk-sensitive and average optimality in Markov decision processes. In J. Ramík & D. Stavárek (Eds.), *Proc. 30th Internat. Conference Mathematical Methods in Economics 2012, Part II* (pp. 799–804). Karviná: Silesian University, School of Business Administrations.

[10] van Dijk, N. M. & Sladký, K. (2006). On the total reward variance for continuous-time Markov reward chains, *J. Appl. Probab.*, 43, 1044–1052.