Karel Sladký
Second Order optimality in Markov decision chains

Kybernetika, Vol. 53 (2017), No. 6, 1086–1099

Persistent URL: http://dml.cz/dmlcz/147086

KYBERNETIKA — VOLUME *53* (2017), NUMBER 6, PAGES 1086–1099

SECOND ORDER OPTIMALITY IN MARKOV DECISION CHAINS

Karel Sladký

The article is devoted to Markov reward chains in discrete-time setting with finite state spaces. Unfortunately, the usual optimization criteria examined in the literature on Markov decision chains, such as a total discounted, total reward up to reaching some specific state (called the first passage models) or mean (average) reward optimality, may be quite insufficient to characterize the problem from the point of a decision maker. To this end it seems that it may be preferable if not necessary to select more sophisticated criteria that also reflect variability-risk features of the problem. Perhaps the best known approaches stem from the classical work of Markowitz on mean variance selection rules, i. e. we optimize the weighted sum of average or total reward and its variance. The article presents explicit formulae for calculating the variances for transient and discounted models (where the value of the discount factor depends on the current state and action taken) for finite and infinite time horizon. The same result is presented for the long run average nondiscounted models where finding stationary policies minimizing the average variance in the class of policies with a given long run average reward is discussed.

Keywords: Markov decision chains, second order optimality, optimality conditions for transient, discounted and average models, policy iterations, value iterations

Classification: 90C40, 93E20

1. INTRODUCTION

The usual optimization criteria examined in the literature on stochastic dynamic programming, such as a total discounted or mean (average) reward structures, may be quite insufficient to characterize the problem from the point of a decision maker. Perhaps the best known approaches stem from the classical work of Markowitz (cf. [5]) on mean variance selection rules, i. e. we optimize the weighted sum of average or total reward and its variance.

To this end it may be preferable if not necessary to select more sophisticated criteria that also reflect variability-risk features of the problem. Most notably, the variance of the cumulative rewards can be indicative and seems of interest. For a detailed discussion of such approaches see the review paper by White [15].

DOI: 10.14736/kyb-2017-6-1086

To the best of our knowledge higher moments and variance of cumulative rewards in Markov reward chains have been systematically studied mostly for discrete time models. Research in this direction has been initiated in Mandl [4], Jaquette [3], and Sobel [12].

Particularly, in these references for controlled discrete-time Markov reward chains, the variance (or second moment) of total expected discounted or average rewards has been considered to select the 'best' policy within the class of discounted (or average) optimal policies to find a lower variance (or lower second moment) of the cumulative reward. Alternatively, also criteria reflecting the variability or risk features for policies not restricted to the class of optimal policies can be analyzed using this approach.

In the present paper we focus attention on discrete-time Markov decision chains with finite state space. We present explicit formulas for the expected total reward and variance for finite horizon models along with their asymptotic behavior for transient and discounted models. As concerns undiscounted models we present explicit formulas for the mean (average) total reward and the corresponding variance. It is indicated how policy and value iteration method can be employed for finding (not necessary optimal) policies minimizing the variance.

The paper is structured as follows. Section 2 presents notations and preliminaries, recursive formulas for finding first and second moment along the corresponding variance are discussed in section 3. Asymptotic properties of total reward of transient (first passage models) and discounted models are studied in section 4. Explicit formulas enable to calculate the difference in the variances for discounted and transient models with the same total rewards. Our approach primarily based on transient models can be easily extended to discounted models where the discount factor depends on the current state and decision taken. Asymptotic behaviour of undiscounted models is studied in section 5. Attention is focused on mean reward and variances in unichain and multichain models. Policy and value iteration method for finding optimal decision are discussed in section 6. Conclusions are made in section 7.

2. NOTATIONS AND PRELIMINARIES

We consider Markov decision chain $X = \{X_n, n = 0, 1, \ldots\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$, and finite set $\mathcal{A}_i = \{1, 2, \ldots, K_i\}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Supposing that in state $i \in \mathcal{I}$ action $a \in \mathcal{A}_i$ is selected, then state j is reached in the next transition with a given probability $p_{ij}(a)$ and one-stage transition reward r_{ij} will be accrued to such transition.

A (Markovian) policy controlling the decision process is given by a sequence of decisions at every time point. In particular, 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 $f^n \in \mathcal{F} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ 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 chain X is in state i. Policy which takes at all times the same decision rule, i. e. $\pi \sim (f)$, is called stationary; $P(f)$ is transition probability matrix with elements $p_{ij}(f_i)$. Obviously, $r_i^{(1)}(f_i) = \sum_{j \in \mathcal{I}_i} p_{ij}(f_i) r_{ij}$ is the expected one-stage reward obtained in state $i \in \mathcal{I}$ and $r^{(1)}(f)$ denotes the corresponding N-dimensional column vector of one-stage rewards. Then $[P(f)]^n \cdot r^{(1)}(f)$ is the (column) vector of

rewards accrued after n transitions, its ith entry denotes expectation of the reward if the process X starts in state i.

Recall that (the Cesaro limit of $P(f)$) $P^*(f) := \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k(f)$ (with elements $p^*_{ij}(f)$) always exists. Moreover, if $P(f)$ is aperiodic then even $P^*(f) = \lim_{k\to\infty} P^k(f)$ and the convergence is geometrical. Then $g^{(1)}(f) = P^*(f)\, r^{(1)}(f)$ is the (column) vector of average rewards, its ith entry $g^{(1)}_i(f)$ denotes the average reward if the process starts in state i. In particular, if $P(f)$ is *unichain* (i. e. $P(f)$ contains a single class of recurrent states) the rows of $P^*(f)$, denoted $p^*(f)$, are identical. Then $p^*_{ij}(f) = p^*_j(f)$, i. e. limiting distribution is independent of the starting state and $g^{(1)}(f)$ is a constant vector with elements $\bar{g}^{(1)}(f)$. It is well-known (cf. e. g. [6, 7]) that also $Z(f)$ (fundamental matrix of $P(f)$), and $H(f)$ (the deviation matrix) exist, where $Z(f) := [I - P(f) + P^*(f)]^{-1}, H(f) := Z(f)\,(I - P^*(f))$.

Transition probability matrix $\tilde{P}(f)$ is called *transient* if the spectral radius of $\tilde{P}(f)$ is less than unity, i. e. it at least some row sums of $\tilde{P}(f)$ are less than one. Then $\lim_{n\to\infty}[\tilde{P}(f)]^n = 0$, $\tilde{P}^*(f) = 0$, $g^{(1)}(f) = \tilde{P}^*(f)\, r^{(1)}(f) = 0$ and $\tilde{Z}(f) = \tilde{H}(f) = [I - \tilde{P}(f)]^{-1}$. Observe that if $P(f)$ is stochastic and $\alpha \in (0,1)$ then $\tilde{P}(f) := \alpha P(f)$ is transient, however, if $\tilde{P}(f)$ is transient it may happen that some row sums may be even greater than unity. Moreover, for the so-called first passage problem, i. e. if we consider total reward up to the first reaching of a specific state (resp. the set of specific states), the resulting transition matrix is transient if the specific state (resp. the set of specific states) can be reached from any other state.

Let $\xi_n(\pi) = \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$ be the stream of rewards received in the n next transitions of the considered Markov chain X if policy $\pi = (f^n)$ is followed. Supposing that $X_0 = i$, on taking expectation we get for the first and second moments of $\xi_n(\pi)$

$v^{(1)}_i(\pi, n) := \mathrm{E}^{\,\pi}_i(\xi_n(\pi)) = \mathrm{E}^{\,\pi}_i \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$,

$v^{(2)}_i(\pi, n) := \mathrm{E}^{\,\pi}_i(\xi_n(\pi))^2 = \mathrm{E}^{\,\pi}_i (\sum_{k=0}^{n-1} r_{X_k,X_{k+1}})^2$.

It is well known from the literature (cf. e. g. [4, 6, 7, 14]) that for the time horizon tending to infinity policies maximizing or minimizing the values $v^{(1)}_i(\pi, n)$ for transient models, resp. policies maximizing or minimizing for discounted models the values $v^{\alpha(1)}_i(\pi, n) = \mathrm{E}^{\,\pi}_i \sum_{k=0}^{n-1} \alpha^k r_{X_k,X_{k+1}}$, can be found in the class of stationary policies, i. e. there exist $f^*, \hat{f} \in \mathcal{F}$ such that for all $i \in \mathcal{I}$ and any policy $\pi = (f^n)$

$$v^{(1)}_i(f^*) := \lim_{n\to\infty} v^{(1)}_i(f^*, n) \geq \limsup_{n\to\infty} v^{(1)}_i(\pi, n),$$

$$v^{(1)}_i(\hat{f}) := \lim_{n\to\infty} v^{(1)}_i(\hat{f}, n) \leq \liminf_{n\to\infty} v^{(1)}_i(\pi, n).$$

3. FINITE TIME HORIZON

If policy $\pi \sim (f)$ is stationary, the process X is time homogeneous and for $m < n$ we write for the generated random reward $\xi_n = \xi_m + \xi_{n-m}$ (here we delete the symbol π and tacitly assume that $\mathrm{P}(X_m = j)$ and ξ_{n-m} starts in state j). Similarly we conclude that

$[\xi_n]^2 = [\xi_m]^2 + [\xi_{n-m}]^2 + 2 \cdot \xi_m \cdot \xi_{n-m}$. Then for $n > m$ we can conclude that

$$\begin{aligned} \mathrm{E}_i^{\pi}[\xi_n] &= \mathrm{E}_i^{\pi}[\xi_m] + \mathrm{E}_i^{\pi}\big\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j) \cdot \mathrm{E}_j^{\pi}[\xi_{n-m}]\big\} \quad (1) \\ \mathrm{E}_i^{\pi}[\xi_n]^2 &= \mathrm{E}_i^{\pi}[\xi_m]^2 + \mathrm{E}_i^{\pi}\big\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j) \cdot \mathrm{E}_j^{\pi}[\xi_{n-m}]^2\big\} \\ &\quad +2 \cdot \mathrm{E}_i^{\pi}[\xi_m] \sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j) \cdot \mathrm{E}_j^{\pi}[\xi_{n-m}]. \quad (2) \end{aligned}$$

In particular, from (1), (2) we conclude for $m = 1$

$$\begin{aligned} v_i^{(1)}(f, n+1) &= r_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot v_j^{(1)}(f, n) \quad (3) \\ v_i^{(2)}(f, n+1) &= r_i^{(2)}(f_i) + 2 \cdot \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \ r_{ij} \cdot v_j^{(1)}(f, n) \\ &\quad + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, v_j^{(2)}(f, n) \quad (4) \end{aligned}$$

where $r_i^{(1)}(f_i) := \sum_{j\in\mathcal{I}} p_{ij}(f_i)\ r_{ij}$, $\quad r_i^{(2)}(f_i) := \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij}]^2$.

Remark 3.1. If the transition reward $r_{ij} = r_i$, i. e. $r_i^{(1)} = r_i$, $\ r_i^{(2)}(f_i) = [r_i]^2$ (cf. [16]) for all $i, j \in \mathcal{I}$ then the first two terms on the RHS of (4) in virtue of (3) can be replaced by $r_i[r_i + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i) v_j^{(1)}(f, n)] = r_i[2v_i^{(1)}(f, n+1) - r_i]$; hence (4) takes on the form

$$v_i^{(2)}(f, n+1) = r_i^{(1)}(f_i) \cdot [2v_i^{(1)}(f, n+1) - \ r_i^{(1)}(f_i)] + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, v_j^{(2)}(f, n). \quad (5)$$

Since the variance $\sigma_i^{(2)}(f, n) = v_i^{(2)}(f, n) - [v_i^{(1)}(f, n)]^2$ from (3),(4) we get

$$\begin{aligned} \sigma_i^{(2)}(f, n+1) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \sigma_j^{(2)}(f, n) + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \ r_{ij} \cdot v_j^{(1)}(f, n) \\ &\quad -[v_i^{(1)}(f, n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[v_j^{(1)}(f, n)]^2 \quad (6) \\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + v_j^{(1)}(f, n)]^2 - [v_i^{(1)}(f, n+1)]^2 \\ &\quad + \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \sigma_j^{(2)}(f, n). \quad (7) \end{aligned}$$

Using matrix notations (cf. [11]) equations (3),(4),(6) can be written as:

$$\begin{aligned} v^{(1)}(f, n+1) &= r^{(1)}(f) + P(f) \cdot v^{(1)}(f, n) \quad (8) \\ v^{(2)}(f, n+1) &= r^{(2)}(f) + 2 \cdot P(f) \circ R \cdot v^{(1)}(f, n) + P(f) \ \cdot v^{(2)}(f, n) \quad (9) \\ \sigma^{(2)}(f, n+1) &= r^{(2)}(f) + P(f) \cdot \sigma^{(2)}(f, n) + 2 \cdot P(f) \circ R \cdot v^{(1)}(f, n) \\ &\quad -[v^{(1)}(f, n+1)]^2 + P(f) \ \cdot [v^{(1)}(f, n)]^2 \quad (10) \end{aligned}$$

where

$R = [r_{ij}]$ is an $N \times N$-matrix, and $r^{(2)}(f) = [\, r_i^{(2)}(f_i)], v^{(2)}(f,n) = [v_i^{(2)}(f,n)]$, $v^{(1)}(f,n) = [(v_i^{(1)}(f,n)]$, $\sigma^{(2)}(f,n) = [\sigma_i^{(2)}(f,n)]$ are column vectors.

The symbol $\circ$ is used for Hadamard (entrywise) product of matrices. Observe that $r^{(1)}(f) = (P(f) \circ R) \cdot e, \quad r^{(2)}(f) = [P(f) \circ (R \circ R)] \cdot e.$

4. INFINITE TIME HORIZON: TOTAL REWARD

4.1. Transient (first passage) models

In this section we focus attention on transient models, i. e. we assume that the transition probability matrix $\tilde{P}(f)$ with elements $\tilde{p}_{ij}(f_i)$ is substochastic and $\rho(f)$, the spectral radius of $\tilde{P}(f)$, is less than unity. Observe that if $\sum_{j\in\mathcal{I}} \tilde{p}_{ij}(f_i) = \alpha_i < 1$ then on reaching state i the process X stops with probability $1-\alpha_i$. Moreover, if $\tilde{P}(f) = \alpha P(f)$ the process X stops in every state with probability $1-\alpha$. Then $\tilde{P}^*(f) = \lim_{n\to\infty}[\tilde{P}(f)]^n = 0$ and for the fundamental and deviation matrices we get $\tilde{Z}(f) = \tilde{H}(f) = [I - \tilde{P}(f)]^{-1}$. Moreover, on iterating (8) we easily conclude that there exists $v^{(1)}(f) := \lim\limits_{n\to\infty} v^{(1)}(f,n)$ such that

$$v^{(1)}(f) = r^{(1)}(f) + \tilde{P}(f) \cdot v^{(1)}(f) \Longleftrightarrow v^{(1)}(f) = [I - \tilde{P}(f)]^{-1} r^{(1)}(f). \tag{11}$$

Similarly, from (4),(9) (since the term $2 \cdot P(f) \circ R \cdot v^{(1)}(f,n)$ is bounded and $\lim\limits_{n\to\infty} v^{(1)}(f,n) = v^{(1)}(f)$) on letting $n \to \infty$ we can also verify existence $v^{(2)}(f) = \lim\limits_{n\to\infty} v^{(2)}(f,n)$ such that

$$v^{(2)}(f) = \; r^{(2)}(f) + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f) + \tilde{P}(f)\, v^{(2)}(f) \tag{12}$$

hence

$$v^{(2)}(f) = [I - \tilde{P}(f)]^{-1} \left\{ \, r^{(2)}(f) + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f) \right\}. \tag{13}$$

Employing (12),(13) arrive at the formula for total variance $\sigma^{(2)}(f)$.

Theorem 4.1.

$$\sigma^{(2)}(f) \;\; = \;\; [I - \tilde{P}(f)]^{-1} \cdot \{ \, r^{(2)}(f) + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f) \} - [v^{(1)}(f)]^2. \tag{14}$$

P r o o f . On letting $n \to \infty$ from (6), (7) we get for $\sigma_i^{(2)}(f) := \lim\limits_{n\to\infty} \sigma_i^{(2)}(f,n)$

$$\begin{aligned} \sigma_i^{(2)}(f) \;\; &= \;\; r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} \tilde{p}_{ij}(f_i) \cdot \sigma_j^{(2)}(f) + 2 \sum_{j\in\mathcal{I}} \tilde{p}_{ij}(f_i) \cdot \; r_{ij} \cdot v_j^{(1)}(f) \\ &\quad - [v_i^{(1)}(f)]^2 + \sum_{j\in\mathcal{I}} \tilde{p}_{ij}(f_i) [v_j^{(1)}(f)]^2 \qquad (15) \\ &= \;\; \sum_{j\in\mathcal{I}} \tilde{p}_{ij}(f_i) [r_{ij} + v_j^{(1)}(f)]^2 - [v_i^{(1)}(f)]^2 + \sum_{j\in\mathcal{I}} \tilde{p}_{ij}(f_i) \cdot \sigma_j^{(2)}(f). \qquad (16) \end{aligned}$$

Hence in matrix notation (15) reads

$$\sigma^{(2)}(f) = \; r^{(2)}(f) + \tilde{P}(f) \cdot \sigma^{(2)}(f) + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f) - [v^{(1)}(f)]^2 + \tilde{P}(f) \cdot [v^{(1)}(f)]^2. \tag{17}$$

(14) follows from (17) after little algebra. □

If the discount factor depends on the current state and action taken, let

$$A(f) = \text{diag}\,[\alpha_1(f_1), \cdots, \alpha_N(f_N)]$$

be a diagonal matrix whose ith element is the value of the discount factor if in state i action f_i is selected. Then $\tilde{P}(f) := A(f)P(f)$ and (14) reads

$$\sigma^{(2)}(f) \;=\; [I - A(f)P(f)]^{-1} \cdot \{\, r^{(2)}(f) + 2 \cdot A(f)P(f) \circ R \cdot v^{(1)}(f)\} - [v^{(1)}(f)]^2 \quad (18)$$

where $r^{(2)}(f) = [P(f) \circ (R \circ R)] \cdot e$.

4.2. Discounted models

From (3),(4) we conclude that

$$v_i^{\alpha(1)}(f, n+1) \;=\; r_i^{(1)}(f_i) + \alpha_i(f_i) \sum_{j \in \mathcal{I}} p_{ij}(f_i) \cdot v_j^{\alpha(1)}(f, n) \quad (19)$$

$$\begin{aligned} v_i^{\alpha(2)}(f, n+1) \;=\; & r_i^{(2)}(f_i) + 2 \cdot \alpha_i(f_i) \cdot \sum_{j \in \mathcal{I}} p_{ij}(f_i) \cdot\; r_{ij} \cdot v_j^{\alpha(1)}(f, n) \\ & + [\alpha_i(f_i)]^2 \cdot \sum_{j \in \mathcal{I}} p_{ij}(f_i)\, v_j^{\alpha(2)}(f, n) \end{aligned} \quad (20)$$

and from (19),(20), for the variance $\sigma_i^{\alpha(2)}(f, n) := v_i^{\alpha(2)}(f, n) - [v_i^{\alpha(1)}(f, n)]^2$ we get

$$\begin{aligned} \sigma_i^{\alpha(2)}(f, n+1) = \; & r_i^{(2)}(f_i) + [\alpha_i(f_i)]^2 \sum_{j \in \mathcal{I}} p_{ij}(f_i) \cdot \sigma_j^{\alpha(2)}(f, n) - [v_i^{\alpha(1)}(f, n+1)]^2 \\ & + 2\alpha_i(f_i) \sum_{j \in \mathcal{I}} p_{ij}(f_i)\; r_{ij} \cdot v_j^{\alpha(1)}(f, n) + [\alpha_i(f_i)]^2 \sum_{j \in \mathcal{I}} p_{ij}(f_i) [v_j^{\alpha(1)}(f, n)]^2 \quad (21) \\ = \; & \sum_{j \in \mathcal{I}} p_{ij}(f_i) [r_{ij} + \alpha_i(f_i) \cdot v_j^{\alpha(1)}(f, n)]^2 - [v_i^{\alpha(1)}(f, n+1)]^2 \\ & + [\alpha_i(f_i)]^2 \sum_{j \in \mathcal{I}} p_{ij}(f_i) \cdot \sigma_j^{\alpha(2)}(f, n). \quad (22) \end{aligned}$$

Using matrix notations equations (21), (22) can be written as:

$$v^{\alpha(1)}(f, n+1) = \; r^{(1)}(f) + A(f)P(f) \cdot v^{\alpha(1)}(f, n) \quad (23)$$

$$v^{\alpha(2)}(f, n+1) = \; r^{(2)}(f) + 2A(f)P(f) \circ R \cdot v^{\alpha(1)}(f, n) + [A(f)]^2 P(f) \cdot v^{\alpha(2)}(f, n) \quad (24)$$

recall that $R = [r_{ij}]$ is an $N \times N$-matrix, and $\circ$ is used for Hadamard (entrywise) product of matrices (observe that $[A(f)]^2 = A(f) \cdot A(f) = A(f) \circ A(f)$ since $A(f)$ is diagonal).

On iterating (23) we conclude that $v^{\alpha(1)}(f) := \lim_{n \to \infty} v^{\alpha(1)}(f, n)$ exists and

$$v^{\alpha(1)}(f) = r^{(1)}(f) + A(f)P(f) \cdot v^{\alpha(1)}(f) \iff v^{\alpha(1)}(f) = [I - A(f)P(f)]^{-1} r^{(1)}(f). \quad (25)$$

Similarly to the transient case on letting $n \to \infty$ for discounted models also $v^{\alpha(2)}(f) = \lim_{n\to\infty} v^{\alpha(2)}(f,n)$ exists and by (24)

$$v^{\alpha(2)}(f) = r^{(2)}(f) + 2 \cdot A(f) \cdot P(f) \circ R \cdot v^{\alpha(1)}(f) + [A(f)]^2 \cdot P(f)\, v^{\alpha(2)}(f), \tag{26}$$

so

$$v^{\alpha(2)}(f) = \{[I - [A(f)]^2 \cdot P(f)\}^{-1} [\, r^{(2)}(f) + 2 \cdot A(f) \cdot P(f) \circ R \cdot v^{\alpha(1)}(f)]. \tag{27}$$

Now we are in a position to present explicit formula for the limiting discounted variance.

Theorem 4.2.

$$\begin{aligned} \sigma^{\alpha(2)}(f) &= \{I - [A(f)]^2 \cdot P(f)\}^{-1} \cdot \{\, r^{(2)}(f) + 2 \cdot A(f) \cdot P(f) \circ R \cdot v^{\alpha(1)}(f)\} \\ &\quad - [v^{\alpha(1)}(f)]^2 \end{aligned} \tag{28}$$

where $r^{(2)}(f) = [P(f) \circ (R \circ R)] \cdot e$.

P r o o f . On letting $n \to \infty$ from (21), (22) we get for $\sigma_i^{\alpha(2)}(f) := \lim_{n\to\infty} \sigma_i^{\alpha(2)}(f,n)$

$$\begin{aligned} \sigma_i^{\alpha(2)}(f) &= r_i^{(2)}(f_i) + [\alpha(f_i)]^2 \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \sigma_j^{\alpha(2)}(f) \\ &\quad + 2 \cdot \alpha(f_i) \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot\ r_{ij} \cdot v_j^{\alpha(1)}(f) - [v_i^{\alpha(1)}(f)]^2 \\ &\quad + [\alpha_i(f_i)]^2 \sum_{j\in\mathcal{I}} p_{ij}(f_i) [v_j^{\alpha(1)}(f)]^2 \qquad (29) \\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i) [r_{ij} + \alpha_i(f_i) \cdot v_j^{\alpha(1)}(f)]^2 - [v_i^{\alpha(1)}(f)]^2 \\ &\quad + [\alpha_i(f_i)]^2 \sum_{j\in\mathcal{I}} p_{ij}(f_i) \cdot \sigma_j^{\alpha(2)}(f). \qquad (30) \end{aligned}$$

Hence in matrix notation

$$\begin{aligned} \sigma^{\alpha(2)}(f) &= r^{(2)}(f) + [A(f)]^2 \cdot P(f) \cdot \sigma^{\alpha(2)}(f) + 2 \cdot A(f) \cdot P(f) \circ R \cdot v^{\alpha(1)}(f) \\ &\quad - [v^{\alpha(1)}(f)]^2 + [A(f)]^2 \cdot P(f) \cdot [v^{\alpha(1)}(f)]^2 \end{aligned} \tag{31}$$

(28) follows immediately from (31) after some algebra. □

Remark 4.3. In particular, if the discount factor is independent of the current state and action taken, (28) takes the form

$$\sigma^{\alpha(2)}(f) = [I - \alpha^2 \cdot P(f)]^{-1} \cdot \{\, r^{(2)}(f) + 2 \cdot \alpha \cdot P(f) \circ R \cdot v^{\alpha(1)}(f)\} - [v^{\alpha(1)}(f)]^2. \tag{32}$$

(32) is similar to the formula for the variance of discounted rewards obtained by Sobel [12] by different methods (see also [11]). The formula also follows directly by subtracting $[v^{\alpha(1)}(f)]^2$ from

$$v^{\alpha(2)}(f) = [I - \alpha^2 \cdot P(f)]^{-1} \left\{ r^{(2)}(f) + 2\alpha \cdot P(f) \circ R \cdot v^{\alpha(1)}(f) \right\} \tag{33}$$

(i. e. special case of (27) iff $\alpha(f_i) = \alpha$ for all $i \in \mathcal{I}$.)

Comparing (14) and (32) both formulas differ only in the first term on the right hand-side. Since the expectation of α-discounted model is the same as for the transient model with probability α of stopping, the corresponding variance are different (see [1, 8, 11]).

Remark 4.4. If the transition reward $r_{ij} = r_i$, i. e. $r_i^{(1)} = r_i$, $r_i^{(2)}(f_i) = [r_i]^2$ for all $i, j \in \mathcal{I}$ (see Remark 3.1 and Eq. (5)) then (12) takes on the form

$$v^{(2)}(f) = r^{(1)} \circ [2v^{(1)}(f) - r^{(1)}] + \tilde{P}(f)\, v^{(2)}(f) \tag{34}$$

and similarly (14) can be written as

$$\sigma^{(2)}(f) \quad = \quad [I - \tilde{P}(f)]^{-1} \cdot \{r^{(1)} \circ [2v^{(1)}(f) - r^{(1)}]\} - [v^{(1)}(f)]^2. \tag{35}$$

(35) is similar to the formula for the variance of transient model reported in [16].

5. INFINITE-TIME HORIZON: AVERAGE CASE

5.1. Unichain model

We make the following

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 for every $f \in \mathcal{F}$ the transition probability matrix $P(f)$ is *unichain* (i. e. $P(f)$ have no two disjoint closed sets). In particular, transition probability matrix $P(f)$ and the state space $\mathcal{I}$ can be decomposed as

$$P(f) = \begin{bmatrix} P_{\mathrm{TT}}(f) & P_{\mathrm{TR}}(f) \\ 0 & P_{\mathrm{RR}}(f) \end{bmatrix}, \qquad \mathcal{I} = \mathcal{I}_{\mathrm{T}}(f) \cup \mathcal{I}_{\mathrm{R}}(f),$$

where $\mathcal{I}_{\mathrm{T}}(f)$ resp. $\mathcal{I}_{\mathrm{R}}(f)$, contains all transient (resp. recurrent) states of matrix $P(f)$.

As well known from the literature (see e. g. [6]), if Assumption 1 holds, then the growth rate of $v^{(1)}(f, n)$ is linear and independent of the starting state. In particular, there exists constant vector $g^{(1)}(f) = P^*(f)r^{(1)}(f)$ (with elements $\bar{g}^{(1)}(f)$) along with vector $w^{(1)}(f)$ (unique up to additive constant) such that

$$w^{(1)}(f) + g^{(1)}(f) = r^{(1)}(f) + P(f)w^{(1)}(f). \tag{36}$$

In particular, it is possible to select $w^{(1)}(f)$ such that $P^*(f)w^{(1)}(f) = 0$. Then $w^{(1)}(f) = H(f)r^{(1)}(f) = Z(f)r^{(1)}(f) - P^*(f)r^{(1)}(f)$. On iterating (36) we can conclude that

$$v^{(1)}(f, n) = g^{(1)}(f) \cdot n + w^{(1)}(f) - [P(f)]^n w^{(1)}(f). \tag{37}$$

To simplify the limiting behavior we make also

Assumption 2. The matrix $P(f)$ is aperiodic, i. e. $\lim_{n\to\infty}[P(f)]^n = P^*(f)$.

If Assumption 2 holds and $P^*(f)w^{(1)}(f) = 0$ then for n tending to infinity $v^{(1)}(f,n) - ng^{(1)}(f) - w^{(1)}(f)$ tends to the null vector and the convergence is geometrical. In particular, by (37) we can conclude that for $\varepsilon(n) = P(f)^n w^{(1)}(f)$

$$v^{(1)}(f,n) = g^{(1)}(f)\cdot n + w^{(1)}(f) + \varepsilon(n). \tag{38}$$

In what follows the symbol $\varepsilon(n)$ is reserved for any column vector of appropriate dimension with elements elements $\bar{\varepsilon}(n)$ that converge geometrically to the null vector.

Now we focus attention on asymptotic behaviour of recursive formula (6) if the underlying Markov process is unichained and the time horizon tends to infinity. To this end, we need some facts on the asymptotic properties of total expected reward.

In particular, we can conclude that by (3),(36),(37)

$$\begin{aligned} v_i^{(1)}(f,n+1) + v_j^{(1)}(f,n) &= r_i^{(1)}(f) + \sum_{k\in\mathcal{I}} p_{ik}(f)\cdot v_k^{(1)}(f,n) + v_j^{(1)}(f,n) \\ &= r_i^{(1)}(f) + 2n\bar{g}^{(1)}(f) + \sum_{k\in\mathcal{I}} p_{ik}(f) w_k^{(1)}(f) + w_j^{(1)}(f) + \bar{\varepsilon}(n) \\ &= (2n+1)\bar{g}^{(1)}(f) + w_i^{(1)}(f) + w_j^{(1)}(f) + \bar{\varepsilon}(n) \end{aligned} \tag{39}$$

$$\begin{aligned} v_i^{(1)}(f,n+1) - v_j^{(1)}(f,n) &= r_i^{(1)}(f) + \sum_{k\in\mathcal{I}} p_{ik}(f)\cdot v_k^{(1)}(f,n) - v_j^{(1)}(f,n) \\ &= r_i^{(1)}(f) + \sum_{k\in\mathcal{I}} p_{ik}(f) w_k^{(1)}(f) - w_j^{(1)}(f) + \bar{\varepsilon}(n) \\ &= \bar{g}^{(1)}(f) + w_i^{(1)}(f) - w_j^{(1)}(f) + \bar{\varepsilon}(n). \end{aligned} \tag{40}$$

From (38),(39),(40) we get

$$\begin{aligned} \sum_{j\in\mathcal{I}} p_{ij}(f) & [v_i^{(1)}(f,n+1) + v_j^{(1)}(f,n)][v_i^{(1)}(f,n+1) - v_j^{(1)}(f,n)] \\ &= \sum_{j\in\mathcal{I}} p_{ij}(f)[2n\bar{g}^{(1)}(f) + \bar{g}^{(1)}(f) + w_i^{(1)}(f) + w_j^{(1)}(f)] \\ &\quad \times[\bar{g}^{(1)}(f) + w_i^{(1)}(f) - w_j^{(1)}(f)] + \bar{\varepsilon}(n) \\ &= 2n\bar{g}^{(1)}(f)\sum_{j\in\mathcal{I}} p_{ij}(f)[\bar{g}^{(1)}(f) + w_i^{(1)}(f) - w_j^{(1)}(f)] \\ &\quad + \sum_{j\in\mathcal{I}} p_{ij}(f)\left\{[\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 - [w_j^{(1)}(f)]^2\right\} + \bar{\varepsilon}(n) \\ &= 2n\bar{g}^{(1)}(f)\cdot r_i^{(1)}(f) \\ &\quad + \sum_{j\in\mathcal{I}} p_{ij}(f)\left\{[\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 - [w_j^{(1)}(f)]^2\right\} + \bar{\varepsilon}(n). \end{aligned} \tag{41}$$

Similarly by (38) for the third term on the RHS of (6) (and also for the third term on the RHS of (10)), we have

$$\sum_{j\in\mathcal{I}} p_{ij}(f)\cdot r_{ij}\cdot v_j^{(1)}(f,n) = \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot r_{ij}\cdot [n\cdot \bar{g}^{(1)}(f) + w_j^{(1)}(f) + \bar{\varepsilon}(n)]$$

$$= \quad n\cdot \bar{g}^{(1)}(f)\cdot r_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot r_{ij}\cdot w_j^{(1)}(f) + \bar{\varepsilon}(n). \tag{42}$$

Substitution from (41), (42) into (6) yields after some algebra

$$\begin{aligned}\sigma_i^{(2)}(f,n+1) &= \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot \sigma_j^{(2)}(f,n) + r_i^{(2)}(f_i) + 2\cdot \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot r_{ij}\cdot w_j^{(1)}(f)\\ &\quad + \sum_{j\in\mathcal{I}} p_{ij}(f)[w_j^{(1)}(f)]^2 - [\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 + \bar{\varepsilon}(n)\\ &= \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot \{\sigma_j^{(2)}(f,n) + [r_{ij} + w_j^{(1)}(f)]^2\}\\ &\quad - [\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 + \bar{\varepsilon}(n). \end{aligned} \tag{43}$$

Hence, in matrix form we have:

$$\sigma^{(2)}(f,n+1) = P(f)\,\sigma^{(2)}(f) + s(f) + \varepsilon(n), \tag{44}$$

where elements $s_i(f)$ of the (column) vector $s(f)$ are equal to

$$s_i(f) = \sum_{j\in\mathcal{I}} p_{ij}(f)[r_{ij} + w_j^{(1)}(f)]^2 - [\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 \tag{45}$$

$$= \sum_{j\in\mathcal{I}} p_{ij}(f)[r_{ij} + w_j^{(1)}(f) - \bar{g}^{(1)}(f)]^2 - [w_i^{(1)}(f)]^2. \tag{46}$$

Observe that (46) follows immediately from (45) since by (36)
$-2\sum_{j\in\mathcal{I}} p_{ij}(f)[r_{ij} + w_j^{(1)}(f)]\bar{g}^{(1)}(f) - [\bar{g}^{(1)}(f)]^2 \;=\; -2w_i^{(1)}(f)\bar{g}^{(1)}(f) - [\bar{g}^{(1)}(f)]^2.$
Employing (37) and the analogy between (7) and (44) we arrive at

Theorem 5.1.

$$G(f) := \lim_{n\to\infty} \frac{1}{n}\sigma^{(2)}(f) = P^*(f)s(f) \tag{47}$$

is the average variance corresponding to policy $\pi \sim (f)$.

For details see [9, 10, 13].

5.2. Multichain model

First observe that under suitable labelling the states (or suitable changes rows and corresponding columns) state space $\mathcal{I}$ can be partitioned as

$$\mathcal{I} = \mathcal{I}_0(f) \cup \mathcal{I}_1(f) \cup \mathcal{I}_2(f) \cup \ldots \mathcal{I}_s(f)$$

where $\mathcal{I}_0(f)$ contains all transient states having access at least to two recurrent classes and for $\ell = 1, \dots, s$ each class $\mathcal{I}_\ell(f)$ contains single recurrent class along with transient states having access only to this recurrent class.

Similarly, the transition probability matrix $P(f)$ can be decomposed in the following block triangular form

$$P(f) = \begin{bmatrix} P_{00}(f) & P_{01}(f) & P_{02}(f) & \dots & P_{0s}(f) \\ 0 & P_{11}(f) & 0 & \dots & 0 \\ 0 & 0 & P_{22}(f) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & P_{ss}(f) \end{bmatrix}. \tag{48}$$

Observe that for $\ell = 1, \dots, s$ each class $P_{\ell\ell}(f)$ can be treated as unichained Markov chain. Hence, if the process starts in state $i \in \mathcal{I}_\ell$ then the average variance $G_\ell(f) = P^*_{\ell\ell}(f) \cdot s^\ell(f)$; elements of the row vector $s^\ell(f)$, say $s_i(f)$ are selected for $i \in \mathcal{I}_\ell(f)$.

6. FINDING SECOND ORDER OPTIMAL POLICIES

For finding second order optimal policies, at first it is necessary to construct the set of all transient optimal discounted optimal policies, resp. the set of all average optimal policies (cf. e. g. [4, 6, 7]). Since optimal policies can be found in the class of stationary policies, i. e. there exist $f^*, \bar{f}^* \in \mathcal{F}$ such that

$$v^{(1)}(f^*) \geq v^{(1)}(\pi) \quad \text{resp.} \quad v^{\alpha(1)}(\bar{f}^*) \geq v^{\alpha(1)}(\pi) \quad \text{for every policy} \quad \pi = (f^n) \tag{49}$$

where

$$v^{(1)}(f^*) = r^{(1)}(f^*) + \tilde{P}(f^*) \cdot v^{(1)}(f^*) \geq r^{(1)}(f) + \tilde{P}(f) \cdot v^{(1)}(f) \tag{50}$$

$$v^{(1)}(f^*) = r^{(1)}(f^*) + A(f^*)P(f^*) \cdot v^{(1)}(f^*) \geq r^{(1)}(f) + A(f)P(f) \cdot v^{(1)}(f^*). \tag{51}$$

Let $\mathcal{F}^* \subset \mathcal{F}$ be the set of all transient optimal or discounted optimal stationary policies. Stationary optimal policies minimizing total or discounted variance can be constructed by standard policy or value iteration procedures in the class of policies from $\mathcal{F}^*$.

Similarly, considering average reward of undiscounted models, there exists $\hat{f} \in \mathcal{F}$ such that for all $f \in \mathcal{F}$

$$w^{(1)}(\hat{f}) + g^{(1)}(\hat{f}) = r^{(1)}(\hat{f}) + P(\hat{f})w^{(1)}(\hat{f}) \geq r^{(1)}(f) + P(f)w^{(1)}(\hat{f}). \tag{52}$$

Let $\hat{\mathcal{F}} \subset \mathcal{F}$ be the set of all average optimal stationary policies. Stationary optimal policies minimizing average variance can be constructed on applying standard policy or value iteration procedures in the class of policies from $\hat{\mathcal{F}}$. Of course, it is necessary to consider one-stage rewards $s(f)$, see (44), (45), instead of $r^{(1)}(f)$.

Up to now we looked for policies with minimal variance in the class of policies with maximal total reward, resp. maximal average reward. However, considering stationary

policy (even randomized) not maximizing total reward, resp. average reward, say $f^{\mathrm{d}} \in \mathcal{F}^{\mathrm{d}}$, policy and value iteration can be used for finding policies guaranteeing total reward $v^{(1)}(f^{\mathrm{d}})$, resp. average reward $g^{(1)}(f^{\mathrm{d}})$ with minimal possible variances. To this end, we can use standard algorithmic procedures where $v^{(1)}(f^*)$ is replaced by $v^{(1)}(f^{\mathrm{d}})$, resp. $g^{(1)}(\hat{f})$ is replaced by $g^{(1)}(f^{\mathrm{d}})$ and $\mathcal{F}^*$ by $\mathcal{F}^{\mathrm{d}}$, resp. $\hat{\mathcal{F}}$ by $\mathcal{F}^{\mathrm{d}}$.

For the sake of completeness we present in detail policy and value iteration algorithms for finding optimal policies with minimal variances in transient models.

Algorithm 6.1. (Policy iterations for finding optimal policy with minimal variances.) Construct a sequence of decisions $f^{*(k)} \in \mathcal{F}^*$, along with a sequence of transient matrices $\tilde{P}(f^{*(k)})$, $k = 0, 1, \ldots$ such that

Step 0. Select matrix $\tilde{P}(f^{*(0)})$ with $f^{*(0)} \in \mathcal{F}^*$.

Step 1. For the matrix $\tilde{P}(f^{*(k)})$ calculate the vector of second moments of total rewards $v^{(2)}(f^{*(k)})$ such that (cf. (12),(13))

$$v^{(2)}(f^{*(k)}) = r^{(2)}(f^{*(k)}) + 2 \cdot \tilde{P}(f^{*(k)}) \circ R \cdot v^{(1)}(f^*) + \tilde{P}(f^{*(k)}) \circ R \cdot v^{(2)}(f^{*(k)}). \quad (53)$$

Step 2. Construct (if possible) matrix $\tilde{P}(f^{*(k+1)})$ with $f^{*(k+1)} \in \mathcal{F}^*$, such that

$$r^{(2)}(f^{*(k+1)}) + \tilde{P}(f^{*(k+1)}) \cdot v^{(1)}(f^*) \leq v^{(2)}(f^{*(k)}). \quad (54)$$

To this end for each action set $\mathcal{A}_i^*$ select $f_i \in \mathcal{A}_i^*$ such that

$$r^{(2)}(f^{*(k+1)}) + 2 \cdot \tilde{P}(f^{*(k+1)}) \circ R \cdot v^{(1)}(f^*) = \min_{f \in \mathcal{F}^*} [r^{(2)}(f) + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f^*)]$$

(observe that the vectorial minimum exists).

Step 3. If $v^{(2)}(f^{*(k+1)}) = v^{(2)}(f^{*(k)})$ then go to Step 4, else go to Step 1.

Step 4. Set $\tilde{P}(\bar{f}^*) := \tilde{P}(f^{*(k+1)})$, find also all decisions $f \in \mathcal{F}^* \subset \mathcal{F}$ such that

$$f \in \mathcal{F}^* \Rightarrow v^{(2)}(\bar{f}^*) = r^{(2)}(f^*) + 2 \cdot \tilde{P}(f^*) \circ R \cdot v^{(1)}(f^*) + \tilde{P}(f^*) \cdot v^{(2)}(\bar{f}^*), \quad (55)$$

then stop.

Policy $\bar{f}^*$ is the policy minimizing total variances in the class of all policies maximizing total expected reward.

Algorithm 6.2. (Value iterations for finding optimal policies with minimal variances.) Construct (recursively) the sequence of (column) vectors $\{v^{(2)}(n), n = 0, 1, \ldots\}$ (where $v^{(2)}(0) := 0$, $f^* \in \mathcal{F}^*$)

$$\begin{aligned} v^{(2)}(n+1) &= \min_{f \in \mathcal{F}^*} [r^{(2)}(f) + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f^*) + \tilde{P}(f) \cdot v^{(2)}(n)], \\ &= [r^{(2)}(f^{(n)}) + 2 \cdot \tilde{P}(f^{(n)}) \circ R \cdot v^{(1)}(f^*) + \tilde{P}(f^{(n)}) \cdot v^{(2)}(n)] \quad (56) \end{aligned}$$

(observe that the vectorial minimum in (56) exists).
If for a given $\varepsilon > 0$ the norm of $\|v^{(2)}(n+1) - v^{(2)}(n)\| < \varepsilon$ for policy $f^{(n)} \in \mathcal{F}$ then stop and set $f^* := f^{(n)}$.

Policy $\bar{f}^*$ is the policy minimizing total variance in the class of all policies maximizing total expected reward (up to a given possible error equal to ε.)

7. CONCLUSIONS

The paper extends formulas for total reward of discounted and first passage models, as well as for long run average reward of undiscounted models, in Markov decision chains under specific forms of one-step rewards and non-constant values of discount factor. In particular, using the recursive formula for total reward for finite time horizon the results of [1, 11, 12, 16] are amplified and extended to transient and discounted models with time-varying discount factor. The results on undiscounted models reported in section 5 are based on [11] where only *un*controlled unichain models were studied. For finding policies with maximum total or average rewards policy and value iterations method can be used. Similarly, as indicated in section 6, also in the class of optimal policies methods of policy and value iterations can be used to find policies minimizing total variance of optimal values.

ACKNOWLEDGEMENT

This work was partially supported by the Czech Science Foundation under Grant 15-10331S.

(Received January 23, 2017)

REFERENCES

[1] E. A. Feinberg and J. Fei: Inequalities for variances of total discounted costs. J. Appl. Probab. *46* (2009), 1209–1212. DOI:10.1239/jap/1261670699

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

[3] S. C. Jaquette: Markov decision processes with a new optimality criterion: Discrete time. Ann. Statist. *1* (1973), 496–505. DOI:10.1214/aos/1176342415

[4] P. Mandl: On the variance in controlled Markov chains. Kybernetika *7* (1971), 1–12.

[5] H. Markowitz: Portfolio Selection – Efficient Diversification of Investments. Wiley, New York 1959.

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

[7] N. Bäuerle and U. Rieder: Markov Decision Processes with Application to Finance. Springer–Verlag, Berlin 2011.

[8] R. Righter: Stochastic comparison of discounted rewards. J. Appl. Probab. *48* (2011), 293–294. DOI:10.1017/S0021900200007786

[9] K. Sladký: On mean reward variance in semi-Markov processes. Math. Meth. Oper. Res. *62* (2005), 387–397. DOI:10.1007/s00186-005-0039-z

[10] K. Sladký: Risk-sensitive and mean variance optimality in Markov decision processes. Acta Oeconomica Pragensia *7* (2013), 146–161.

[11] K. Sladký: Second order optimality in transient and discounted Markov decision chains. In: Proc. 33th Internat. Conf. Math. Methods in Economics MME 2015 (D. Martinčík, ed.), University of West Bohemia, Plzeň 2015, pp. 731–736.

[12] M. Sobel: The variance of discounted Markov decision processes. J. Appl. Probab. *19* (1982), 794–802. DOI:10.2307/3213832

[13] N. M. Van Dijk and K. Sladký: On the total reward variance for continuous-time Markov reward chains. J. Appl. Probab. *43* (2006), 1044–1052. DOI:10.1017/s0021900200002412

[14] A. F. Veinott, Jr: Discrete dynamic programming with sensitive discount optimality criteria. Ann. Math. Statist. *13* (1969), 1635–1660. DOI:10.1214/aoms/1177697379

[15] D. J. White: Mean, variance and probability criteria in finite Markov decision processes: A review. J. Optimizat. Th. Appl. *56* (1988), 1–29. DOI:10.1007/bf00938524

[16] X. Wu and X. Guo: First passage optimality and variance minimisation of Markov decision processes with varying discount factors. J. Appl. Probab. *52* (2015), 441–456. DOI:10.1017/s0021900200012560

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