Transient and Average Markov Reward Chains with Applications to Finance

Karel Sladký [1]

Abstract. The article is devoted to Markov reward chains, in particular, attention is primarily focused on the reward variance arising by summation of generated rewards. Explicit formulae for calculating the variances for transient and average models are reported along with sketches of algorithmic procedures for finding policies guaranteeing minimal variance in the class of policies with a given transient or average reward. Application of the obtained results to financial models is indicated.

Keywords: dynamic programming, transient and average Markov reward chains, reward-variance optimality, optimality in financial models.

JEL classification: C44, C61, C63
AMS classification: 90C40, 60J10, 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. To this end 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 (cf. [2]) on mean variance selection rules, i.e. we optimize the weighted sum of average or total reward and its variance. In the present paper we restrict attention on transient and average models with finite state space.

2 Notation and Preliminaries

In this note, we consider at discrete time points Markov decision process $X = \{X_n, n = 0, 1, \ldots\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$, and compact set $\mathcal{A}_i = [0, K_i]$ 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 chosen, 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, $\pi = (f^0, f^1, \ldots)$, 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 $f^n_i \in \mathcal{A}_i$ is the decision (or action) taken at the nth transition if the chain X is in state i. Let $\pi^m = (f^m, f^{m+1}, \ldots)$, hence $\pi = (f^0, f^1, \ldots, f^{m-1}, \pi^m)$, in particular $\pi = (f^0, \pi^1)$. The symbol E^{π}_i denotes the expectation if $X_0 = i$ and policy $\pi = (f^n)$ is followed, in particular, $\mathsf{E}^{\pi}_i(X_m = j) = \sum_{i_j \in \mathcal{I}} p_{i,i_1}(f^0_i) \ldots p_{i_{m-1},j}(f^{m-1}_{m-1})$; $\mathrm{P}(X_m = j)$ is the probability that X is in state j at time m.

Policy π which selects at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary, hence following policy $\pi \sim (f)$ X is a homogeneous Markov chain with transition probability matrix $P(f)$ whose ij-th element is $p_{ij}(f_i)$. Then $r^{(1)}_i(f_i) := \sum_{j \in \mathcal{I}} p_{ij}(f_i) r_{ij}$ is the expected one-stage reward obtained in state i. Similarly, $r^{(1)}(f)$ is an N-column vector of one-stage rewards whose i-the elements equals $r^{(1)}_i(f_i)$. The symbol I denotes an identity matrix and e is reserved for a unit column vector.

Considering the standard probability matrix $P(f)$ the spectral radius of $P(f)$ is equal to one. Recall that (the Cesaro limit of $P(f)$) $P^*(f) := \lim_{n\to\infty} \frac{1}{n} \sum_{k=}^{n-1} P^k(f)$ (with elements $p^*_{ij}(f)$) exists, and if

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

$P(f)$ is aperiodic then even $P^*(f) = \lim\limits_{k\to\infty} P^k(f)$ and the convergence is geometrical. Then $g^{(1)}(f) = P^*(f)\, r^{(1)}(f)$ is the (column) vector of average rewards, its $i$the entry $g_i^{(1)}(f)$ denotes the average reward if the process starts in state i. Moreover, if $P(f)$ is unichain, i.e. $P(f)$ contains a single class of recurrent states, 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. [3, 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.

3 Reward Variance for Finite and Infinite Time Horizon

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 $\xi_n(\pi)$

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

It is well known from the literature (cf. e.g. [1],[3],[7],[8]) that for the time horizon tending to infinity policies maximizing or minimizing the values $v_i^{(1)}(\pi,n)$ for transient models, as well as policies maximizing or minimizing the value $g^{(1)}(\pi) = \lim_{n\to\infty} n^{-1} v_i^{(1)}(\pi,n)$ can be found in the class of stationary policies, i.e. there exist $f^*, \hat{f}, \bar{f}^*, \bar{f} \in \mathcal{F}$ such that for all $i \in \mathcal{I}$ and any policy $\pi = (f^n)$

$$v_i^{(1)}(f^*) := \lim_{n\to\infty} v_i^{(1)}(f^*,n) \geq \limsup_{n\to\infty} v_i^{(1)}(\pi,n), \quad v_i^{(1)}(\hat{f}) := \lim_{n\to\infty} v_i^{(1)}(\hat{f},n) \leq \liminf_{n\to\infty} v_i^{(1)}(\pi,n), \tag{2}$$

$$g(\bar{f}^*) := \lim_{n\to\infty} \frac{1}{n} v_i^{(1)}(\bar{f}^*,n) \geq \limsup_{n\to\infty} \frac{1}{n} v_i^{(1)}(\pi,n), \; g(\bar{f}) := \lim_{n\to\infty} \frac{1}{n} v_i^{(1)}(\bar{f},n) \leq \liminf_{n\to\infty} \frac{1}{n} v_i^{(1)}(\pi,n). \tag{3}$$

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

$$\mathsf{E}_i^{\,\pi}[\xi_n] \;=\; \mathsf{E}_i^{\,\pi}[\xi_m] + \mathsf{E}_i^{\,\pi}\big\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^{\,\pi}[\xi_{n-m}]\big\}. \tag{4}$$

$$\mathsf{E}_i^{\,\pi}[\xi_n]^2 \;=\; \mathsf{E}_i^{\,\pi}[\xi_m]^2 + \mathsf{E}_i^{\,\pi}\big\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^{\,\pi}[\xi_{n-m}]^2\big\} + 2\cdot \mathsf{E}_i^{\,\pi}[\xi_m] \sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^{\,\pi}[\xi_{n-m}]. \tag{5}$$

In particular, from (2), (4) and (5) we conclude for $m = 1$

$$v_i^{(1)}(f,n+1) \;=\; r_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot v_j^{(1)}(f,n) \tag{6}$$

$$v_i^{(2)}(f,n+1) \;=\; r_i^{(2)}(f_i) + 2\cdot\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\; r_{ij}\cdot v_j^{(1)}(f,n) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, v_j^{(2)}(f,n) \tag{7}$$

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

Since the variance $\sigma_i(f,n) = v_i^{(2)}(f,n) - [v_i^{(1)}(f,n)]^2$ from (6),(7) we get

$$\begin{aligned}\sigma_i(f,n+1) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j(f,n) + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\ r_{ij}\cdot v_j^{(1)}(f,n) \\ &\quad -[v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[v_j^{(1)}(f,n)]^2 \qquad (8)\\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + v_j^{(1)}(f,n)]^2 - [v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j(f,n). \qquad (9)\end{aligned}$$

Using matrix notations (cf. [5]) equations (6),(7),(8) can be written as:

$$\begin{aligned}v^{(1)}(f,n+1) &= r^{(1)}(f) + P(f)\cdot v^{(1)}(f,n) \qquad (10)\\ v^{(2)}(f,n+1) &= r^{(2)}(f) + 2\cdot P(f)\circ R\cdot v^{(1)}(f,n) + P(f)\ \cdot v^{(2)}(f,n) \qquad (11)\\ \sigma(f,n+1) &= r^{(2)}(f) + P(f)\sigma(f,n) + 2\cdot P(f)\circ R\cdot v^{(1)}(f,n) \\ &\quad -[v^{(1)}(f,n+1)]^2 + P(f)\ \cdot [v^{(1)}(f,n)]^2 \qquad (12)\end{aligned}$$

where $R = [r_{ij}]_{i,j}$ is an $N\times N$-matrix, and $r^{(2)}(f) = [\ r_i^{(2)}(f_i)]$, $v^{(2)}(f,n) = [v_i^{(2)}(f,n)]$, $v^{(1)}(f,n) = [(v_i^{(1)}(f,n)]$, $\sigma(f,n) = [\sigma_i(f,n)]$ are column vectors.
The symbol $\circ$ is used for Hadamard (entrywise) product of matrices. Observe that
$r^{(1)}(f) = (P(f)\circ R)\cdot e$, $\quad r^{(2)}(f) = [P(f)\circ(R\circ R)]\cdot e$.

3.2 Infinite-Time Horizon: Transient Case

In this subsection we focus attention on transient models, i.e. we assume that the transition probability matrix $\tilde{P}(f)$ with elements $p_{ij}(f_i)$ is substochastic and $\rho(f)$, the spectral radius of $\tilde{P}(f)$, is less than unity.

Then on iterating (10) we easily conclude that there exists $v^{(1)}(f) := \lim_{n\to\infty} v^{(1)}(f,n)$ such that

$$v^{(1)}(f) = r^{(1)}(f) + \tilde{P}(f)\cdot v^{(1)}(f) \Longleftrightarrow v^{(1)}(f) = [I-\tilde{P}(f)]^{-1}r^{(1)}(f). \qquad (13)$$

Similarly, from (11) (since the term $2\cdot P(f)\circ R\cdot v^{(1)}(f,n)$ must be bounded) on letting $n\to\infty$ we can also verify existence $v^{(2)}(f) = \lim_{n\to\infty} v^{(2)}(f,n)$ such that

$$v^{(2)}(f) = \ r^{(2)}(f) + 2\cdot\tilde{P}(f)\circ R\cdot v^{(1)}(f) + \tilde{P}(f)\, v^{(2)}(f) \qquad (14)$$

hence

$$v^{(2)}(f) = [I-\tilde{P}(f)]^{-1}\ \Big\{\ r^{(2)}(f) + 2\cdot\tilde{P}(f)\circ R\cdot v^{(1)}(f)\Big\}. \qquad (15)$$

On letting $n\to\infty$ from (8), (9) we get for $\sigma_i(f) := \lim_{n\to\infty}\sigma_i(f,n)$

$$\begin{aligned}\sigma_i(f) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j(f) + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\ r_{ij}\cdot v_j^{(1)}(f) \\ &\quad -[v_i^{(1)}(f)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[v_j^{(1)}(f)]^2 \qquad (16)\\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + v_j^{(1)}(f)]^2 - [v_i^{(1)}(f)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j(f). \qquad (17)\end{aligned}$$

Hence in matrix notation

$$\sigma(f) = \ r^{(2)}(f) + \tilde{P}(f)\cdot\sigma(f) + 2\cdot\tilde{P}(f)\circ R\cdot v^{(1)}(f) - [v^{(1)}(f)]^2 + \tilde{P}(f)\cdot[v^{(1)}(f)]^2. \qquad (18)$$

After some algebra (18) can be also written as

$$\sigma(f) \ = \ [I-\tilde{P}(f)]^{-1}\cdot\{\ r^{(2)}(f) + 2\cdot\tilde{P}(f)\circ R\cdot v^{(1)}(f) - [v^{(1)}(f)]^2\}. \qquad (19)$$

In particular, for the discounted case, i.e. if for some discount factor $\alpha\in(0,1)$ the transient matrix $\tilde{P}(f) := \alpha P(f)$ then (19) reads

$$\sigma(f) \ = \ [I-\alpha P(f)]^{-1}\cdot\{\ r^{(2)}(f) + 2\cdot\alpha P(f)\circ R\cdot v^{(1)}(f) - [v^{(1)}(f)]^2\}. \qquad (20)$$

(20) is similar to the formula for the variance of discounted rewards obtained by Sobel [6] using different methods.

3.3 Infinite-Time Horizon: Average Case

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).

As well known from the literature (see e.g. [3]), 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 $g^{(1)}(f) = P^*(f)r(f)$ (with elements $\bar{g}^{(1)}(f)$) along with vector $w^{(1)}(f)$ (unique up to an additive constant) such that

$$w^{(1)}(f) + g^{(1)}(f) = r(f) + P(f)w^{(1)}(f). \tag{21}$$

In particular, it is possible to select $w^{(1)}(f)$ such that $P^*(f)w^{(1)}(f) = 0$. Then $w^{(1)}(f) = H(f)r(f) = Z(f)r(f) - P^*(f)r(f)$. On iterating (21) we can conclude that

$$v^{(1)}(f,n) = g^{(1)}(f) \cdot n + w^{(1)}(f) + [P(f)]^n w^{(1)}(f) \tag{22}$$

To simplify the limiting behavior we make also

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

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 geometric. In particular, by (22) we can conclude that for $\varepsilon(n) = P(f)^n w^{(1)}(f)$

$$v^{(1)}(f,n) = g^{(1)}(f) \cdot n + w^{(1)}(f) + \varepsilon(n) \tag{23}$$

where $\varepsilon(n)$ tends to the null vector and the convergence is geometrical. In what follows the symbol $\varepsilon(n)$ is reserved for any column vector of appropriate dimension whose elements converge geometrically to the null vector.

Employing the above facts we can conclude that by (6),(21),(22)

$$\begin{aligned}
v_i^{(1)}(f,n+1) + v_j^{(1)}(f,n) &= r_i(f) + \sum_{k\in\mathcal{I}} p_{ik}(f) \cdot v_k^{(1)}(f,n) + v_j^{(1)}(f,n) \\
&= r_i(f) + 2n\bar{g}^{(1)}(f) + \sum_{k\in\mathcal{I}} p_{ik}(f) w_k^{(1)}(f) + w_j^{(1)}(f) + \varepsilon(n) \\
&= (2n+1)\bar{g}^{(1)}(f) + w_i^{(1)}(f) + w_j^{(1)}(f) + \varepsilon(n) \qquad (24) \\
v_i^{(1)}(f,n+1) - v_j^{(1)}(f,n) &= r_i(f) + \sum_{k\in\mathcal{I}} p_{ik}(f) \cdot v_k^{(1)}(f,n) - v_j^{(1)}(f,n) \\
&= r_i(f) + \sum_{k\in\mathcal{I}} p_{ik}(f) w_k^{(1)}(f) - w_j^{(1)}(f) + \varepsilon(n) \\
&= \bar{g}^{(1)}(f) + w_i^{(1)}(f) - w_j^{(1)}(f) + \varepsilon(n) \qquad (25)
\end{aligned}$$

From (23),(24),(25) we get

$$\begin{aligned}
\sum_{j\in\mathcal{I}} p_{ij}(f) \quad & [v_i^{(1)}(f,n+1) + v_j^{(1)}(f,n)][v_i^{(1)}(f,n+1) - v_j^{(1)}(f,n)] \\
&= \sum_{j\in\mathcal{I}} p_{ij}(f)[2n\bar{g}^{(1)}(f) + \bar{g}^{(1)}(f) + w_i^{(1)}(f) + w_j^{(1)}(f)][\bar{g}^{(1)}(f) + w_i^{(1)}(f) - w_j^{(1)}(f)] + \varepsilon(n) \\
&= 2n\bar{g}^{(1)}(f) \sum_{j\in\mathcal{I}} p_{ij}(f)[\bar{g}^{(1)}(f) + w_i^{(1)}(f) - w_j^{(1)}(f)] \\
&+ \sum_{j\in\mathcal{I}} p_{ij}(f) \left\{ [\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 - [w_j^{(1)}(f)]^2 \right\} + \varepsilon(n) \\
&= 2n\bar{g}^{(1)}(f) r_i(f) + \sum_{j\in\mathcal{I}} p_{ij}(f) \left\{ [\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 - [w_j^{(1)}(f)]^2 \right\} + \varepsilon(n) \qquad (26)
\end{aligned}$$

Similarly by (23) for the third term on the RHS of (8) (and also for the third term on the RHS of (12)), we have

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

$$= \quad n\cdot \bar{g}^{(1)}(f)\cdot r_i(f) + \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot r_{ij}\cdot w_j^{(1)}(f) + \varepsilon(n) \tag{27}$$

Substitution from (26), (27) into (8) yields after some algebra

$$\begin{aligned}\sigma_i(f,n+1) &= \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot\sigma_j(f,n) + r_i^{(2)}(f) + 2\cdot\sum_{j\in\mathcal{I}} p_{ij}(f)\cdot r_{ij}\cdot w_j^{(1)}(f)\\ &\quad + \sum_{j\in\mathcal{I}} p_{ij}(f)[w_j^{(1)}(f]^2 - [\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 + \varepsilon(n)\\ &= \sum_{j\in\mathcal{I}} p_{ij}(f)\cdot\{\sigma_j(f,n) + [r_{ij} + w_j^{(1)}(f)]^2\} - [\bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 + \varepsilon(n) \qquad (28)\end{aligned}$$

Hence, in matrix form we have:

$$\sigma(f,n+1) = \sigma(f) + s(f) + \varepsilon(n) \tag{29}$$

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

$$s_i(f) = \sum_{j\in\mathcal{I}} p_{ij}(f)[r_{ij} + w_j^{(1)}(f)]^2 - [g^{(1)}(f) + w_i^{(1)}(f)]^2 \tag{30}$$

$$= \sum_{j\in\mathcal{I}} p_{ij}(f)[r_{ij} + w_j^{(1)}(f) - g^{(1)}(f)]^2 - [w_i^{(1)}(f)]^2 \tag{31}$$

Observe that by (31) follows immediately from (30) since by (21)
$-2\sum_{j\in\mathcal{I}} p_{ij}(f)(r_{ij} + w_j^{(1)}(f))g^{(1)}(f) - [g^{(1)}(f)]^2 = -2w_i^{(1)}(f)g^{(1)}(f) - [g^{(1)}(f)]^2$.
Employing (22) and the analogy between (9) and (29) we can conclude that

$$G(f) = \lim_{n\to\infty}\frac{1}{n}\sigma(f) = P^*(f)s(f) \tag{32}$$

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

4 Finding Optimal Policies

For finding second order optimal policies, at first it is necessary to construct the set of optimal transient or optimal average policies (cf. e.g,[1, 3, 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

$$v^{(1)}(f^*) \geq v^{(1)}(\pi) \quad \text{resp.} \quad g^{(1)}(\bar{f}^*) \geq g^{(1)}(\pi) \quad \text{for every policy} \quad \pi = (f^n). \tag{33}$$

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

5 Specific Example: Credit Management

The state of the bank is determined by the bank liabilities, i.e. deposits and the capital and is also influenced by the current state of the economy (cf. [7, 9]). It is a task for expert to evaluate each possible state of the bank by some value, say $i\in\mathcal{I}$, we assume that the set $\mathcal{I}$ is finite. A subset of the state space $\mathcal{I}$, say $\mathcal{I}^*$ is called optimal, the decision maker tries to reach this set. To this end at each time point the decision maker receives some money amount depending on the current state of the bank, say c_i, to improve the state of the bank. The decision maker has the following options:

1. advertise the activity of the bank,
2. assign small reward as a courtesy to the non-problematic credit holders,
3. warn and penalized the problematic credit holders.

Based on the experience of the bank, suitable advertising improves the state of the bank by reaching from state i some more suitable state $j \in \mathcal{I}$ with probability $p_{ij}(1)$. Similarly a courtesy reward in the total amount c_i can help to reach a more suitable state $j \in \mathcal{I}$ with probability $p_{ij}(2)$. Finally, warning and penalizing the problematic credit holders changes the state by reaching state j with probability $p_{ij}(3)$.

Using the above mentioned approach the problem of optimal credit-granting policy is formulated as a problem of finding optimal policy of a controlled Markov chain. Observe that the transient model can also grasp models with discount factor depending on the current state. Moreover, if the discount factor is very close to unity we try to optimize the long run average reward.

6 Conclusions

The article is devoted to Markov decision chains, in particular, attention is primarily focused on the optimal policies for transient and average Markov reward chains. In the class of optimal policies procedures for finding policies with minimal total or average variance are discussed. Application of the obtained results for finding an optimal credit-granting policy of a bank is discussed.

Acknowledgements

This research was supported by the Czech Science Foundation under Grant 15-10331S and by CONACyT (Mexico) and AS CR (Czech Republic) under Project 171396.

References

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

[2] Markowitz, H.: *Portfolio Selection - Efficient Diversification of Investiments.* Wiley, New York 1959.

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

[4] Sladký, K.: *On mean reward variance in semi-Markov processes.* Mathematical Methods of Operations Research *62* (2005), No. 3, pp. 387–397.

[5] Sladký, K.: *Second order optimality in transient and discounted Markov Decision Chains.* In: Proceedings of the 33th International Conference Mathematical Methods in Economics 2015, (D. Martinčák et al., eds.). University of West Bohemia, Plzeň, Cheb 2015, pp. 731–736.

[6] Sobel, M.: *The variance of discounted Markov decision processes.* J. Applied Probability **19** (1982), 794–802.

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

[8] Veinott, A.F.Jr.: *Discrete dynamic programming with sensitive discount optimality criteria.* Annals Math. Statistics **40** (1969), 1635–1660.

[9] Waldmann, K.-H.: *On granting credit in a random environment.* Mathem. Methods of Operations Research **47** (2005), 99–115.