The Variance of Discounted Rewards in Markov Decision Processes: Laurent Expansion and Sensitive Optimality

Karel Sladký[1]

Abstract. In this paper we consider discounted Markov decision processes with finite state space and compact actions spaces. We present formulas for the variance of total expected discounted rewards along with its partial Laurent expansion. This enables to compare the obtained results with similar results for undiscounted models.

Keywords: discrete-time Markov decision chains, variance of total discounted rewards, Laurent expansion, mean-variance optimality

JEL classification: C44
AMS classification: 90C15

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 on mean variance selection rules, i.e. we optimize the weighted sum of the expected total (or average) reward and its variance.

In the present paper we restrict attention on unichain models with finite state space. At first we rederive recursive formulas for total undiscounted and discounted reward variance. The heart of this article is the partial Laurent expansion of the variance of discounted rewards and analysis of this behaviour for the discount factor tending to unity.

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 $f_i^n \in \mathcal{A}_i$ 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_i^π denotes the expectation if $X_0 = i$ and policy $\pi = (f^n)$ is followed, in particular, $\mathrm{E}_i^\pi(X_m = j) = \sum_{i_j \in \mathcal{I}} p_{i,i_1}(f_i^0) \ldots p_{i_{m-1},j}(f_{m-1}^{m-1})$; $\mathrm{P}(X_m = j)$ 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 X is a homogeneous Markov chain with transition probability matrix $P(f)$ whose ij-th element equals $p_{ij}(f_i)$; $\mathrm{E}_i^\pi(X_m = j) = [P^m(f)]_{ij}$ (symbol $[A]_{ij}$ denotes the ij-th element of the matrix A) and $r_i(f_i) := \sum_{j \in \mathcal{I}} p_{ij}(f_i) r_{ij}$ is the expected reward obtained in state i. Similarly, $r(f)$ is an N-column vector of

[1]Institute of Information Theory and Automation of the AS CR, Department of Econometrics, Praha 8, Pod vodárenskou věží 4, e-mail: sladky@utia.cas.cz

one-stage rewards whose i-the elements equals $r_i(f_i)$. The symbol I denotes an identity matrix and e is reserved for a unit column vector.

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 $P(f)$ is aperiodic then even $P^*(f) = \lim_{k\to\infty} P^k(f)$ and the convergence is geometrical. 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.

It is well-known (cf. e.g. [1], [6]) that also $Z(f)$ (fundamental matrix of $P(f)$), and $H(f)$ (the deviation matrix) exist, where

$$Z(f) := (I - P(f) + P^*(f))^{-1}, \quad H(f) := Z(f)\,(I - P^*(f)) \quad \text{satisfy}$$

$$P^*(f)Z(f) = Z(f)P^*(f) = P^*(f), \quad H(f) = Z(f) - P^*(f)$$

$$P^*(f)H(f) = H(f)P^*(f) = 0, \quad (I - P(f))\,Z(f) = Z(f)\,(I - P(f)) = I - P^*(f).$$

As it is known from the literature (see e.g. [3],[12]), when λ is not an eigenvalue of $P(f)$, there exists $\mathrm{R}(\lambda, P(f)) = (\lambda I - P(f))^{-1}$ (called the resolvent of $P(f)$) and for λ sufficiently close to 1 $\mathrm{R}(\lambda, P(f))$ has the following Laurent series expansion:

$$\mathrm{R}(\lambda, P(f)) = \frac{1}{\lambda - 1}P^*(f) + H(f) + \sum_{k=1}^{\infty}(-1)^k(\lambda - 1)^k H^{k+1}(f). \tag{1}$$

Observe that the infinite series on the RHS of (1) converges if $|\lambda - 1| < 1/\rho(H(f))$ where $\rho(H(f))$ denotes the spectral radius of $H(f)$.

Furthermore, if $P(\bar{f})$ is another $N \times N$ transition probability matrix then for any λ sufficiently close to 1, the so-called second Laurent expansion holds

$$\mathrm{R}(\lambda, P(\bar{f})) - \mathrm{R}(\lambda, P(f)) = \mathrm{R}(\lambda, P(\bar{f}))\,(P(\bar{f}) - P(f))\,\mathrm{R}(\lambda, P(f)). \tag{2}$$

3 Reward Variance Under Stationary Policies

Let $\xi^\alpha_n(\pi) = \sum_{k=0}^{n-1}\alpha^k r_{X_k,X_{k+1}}$ with $\alpha \in (0,1)$, resp. $\xi_n(\pi) = \sum_{k=0}^{n-1} r_{X_k,X_{k+1}}$, be the stream of α-discounted, resp. undiscounted, 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 $\xi^\alpha_n(\pi)$,

$$v_i^{\alpha(1)}(\pi, n) := \mathrm{E}^\pi_i(\xi^\alpha_n(\pi)) = \mathrm{E}^\pi_i\sum_{k=0}^{n-1}\alpha^k r_{X_k,X_{k+1}}, \quad v_i^{\alpha(2)}(\pi, n) := \mathrm{E}^\pi_i(\xi^\alpha_n(\pi))^2 = \mathrm{E}^\pi_i(\sum_{k=0}^{n-1}\alpha^k r_{X_k,X_{k+1}})^2. \tag{3}$$

If policy $\pi \sim (f)$ is stationary, the process X is time homogeneous and for $m < n$ we write for undiscounted, resp. α-discounted, random reward $\xi_n = \xi_m + \xi_{n-m}$, resp. $\xi^\alpha_n = \xi^\alpha_m + \alpha^m\xi^\alpha_{n-m}$ (here we delete the symbol π and tacitly assume that $\mathrm{P}(X_m = j)$ and ξ_{n-m} starts in state j). Hence $[\xi^\alpha_n]^2 = [\xi^\alpha_m]^2 + \alpha^{2m}\cdot[\xi^\alpha_{n-m}]^2 + 2\cdot\alpha^m\cdot\xi^\alpha_m\cdot\xi^\alpha_{n-m}$. Then for $n > m$ we can conclude that

$$\mathrm{E}^\pi_i[\xi^\alpha_n] = \mathrm{E}^\pi_i[\xi^\alpha_m] + \alpha^m\mathrm{E}^\pi_i\big\{\sum_{j\in\mathcal{I}}\mathrm{P}(X_m = j)\cdot\mathrm{E}^\pi_j[\xi^\alpha_{n-m}]\big\}. \tag{4}$$

$$\mathrm{E}^\pi_i[\xi^\alpha_n]^2 = \mathrm{E}^\pi_i[\xi^\alpha_m]^2 + \alpha^{2m}\mathrm{E}^\pi_i\big\{\sum_{j\in\mathcal{I}}\mathrm{P}(X_m = j)\cdot\mathrm{E}^\pi_j[\xi^\alpha_{n-m}]^2\big\}$$
$$+2\cdot\alpha^m\cdot\mathrm{E}^\pi_i[\xi^\alpha_m]\sum_{j\in\mathcal{I}}\mathrm{P}(X_m = j)\cdot\mathrm{E}^\pi_j[\xi^\alpha_{n-m}]. \tag{5}$$

Using the more appealing notation introduced in (3), from (4) and (5) we conclude for $m = 1$

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

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

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

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

Using matrix notations equations (6), (7) can be written as:

$$v^{\alpha(1)}(f,n+1) = r^{(1)}(f) + \alpha P(f)\cdot v^{\alpha(1)}(f,n) \qquad (10)$$

$$v^{\alpha(2)}(f,n+1) = r^{(2)}(f) + 2\cdot\alpha\cdot P(f)\circ R\cdot v^{\alpha(1)}(f,n) + \alpha^2\cdot P(f)\, v^{\alpha(2)}(f,n) \qquad (11)$$

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

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

$$v^{\alpha(1)}(f) = r^{(1)}(f) + \alpha P(f)\cdot v^{\alpha(1)}(f) \Longleftrightarrow v^{\alpha(1)}(f) = [I-\alpha P(f)]^{-1} r^{(1)}(f). \qquad (12)$$

Finally, for discounted models on letting $n\to\infty$ there also exists $v^{\alpha(2)}(f) = \lim_{n\to\infty} v^{\alpha(2)}(f,n)$ and by (11)

$$v^{\alpha(2)}(f) = r^{(2)}(f) + 2\cdot\alpha\cdot P(f)\circ R\cdot v^{\alpha(1)}(f) + \alpha^2\cdot P(f)\, v^{\alpha(2)}(f), \qquad (13)$$

hence

$$v^{\alpha(2)}(f) = [I-\alpha^2\cdot P(f)]^{-1}\left\{\ r^{(2)}(f) + 2\cdot\alpha\cdot P(f)\circ R\cdot v^{\alpha(1)}(f)\right\}. \qquad (14)$$

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

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

Hence in matrix notation

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

After some algebra (17) can be also written as

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

(18) is similar to the formula for the variance of discounted rewards obtained by Sobel [11] by different methods.

4 Laurent Expansions of Discounted Variance

To begin with, first observe that for $\alpha := \lambda^{-1}$ we have $(\lambda I - P(f))^{-1} = \alpha(I-\alpha P(f))^{-1}$ and (1) takes on the form

$$\alpha(I-\alpha P(f))^{-1} = \frac{\alpha}{1-\alpha}P^*(f) + H(f) + \sum_{k=1}^{\infty}(-1)^k\left(\frac{1-\alpha}{\alpha}\right)^k H^{k+1}(f) \qquad (19)$$

Introducing $\rho := \frac{1-\alpha}{\alpha} \Leftrightarrow \alpha = \frac{1}{1+\rho}$ from (19) we get

$$\alpha(I - \alpha P(f))^{-1} = \rho^{-1}P^*(f) + H(f) + \sum_{k=1}^{\infty}(-1)^k \rho^k H^{k+1}(f). \tag{20}$$

For what follows we shall also need Laurent expansion of $(I - \alpha^2 P(f))^{-1}$. To this end, let $\beta := \alpha^2$ and $\bar{\rho} := \frac{1-\beta}{\beta} \Leftrightarrow \beta = \frac{1}{1+\bar{\rho}}$. Then, in analogy of (19) Laurent expansion of $(I - \beta P(f))^{-1}$ takes on the form

$$\beta(I - \beta P(f))^{-1} = \frac{\beta}{1-\beta}P^*(f) + H(f) + \sum_{k=1}^{\infty}(-1)^k \left(\frac{1-\beta}{\beta}\right)^k H^{k+1}(f) \tag{21}$$

$$(I - \beta P(f))^{-1} = \frac{(\bar{\rho}+1)}{\bar{\rho}}P^*(f) + (\bar{\rho}+1)H(f) + (\bar{\rho}+1)\sum_{k=1}^{\infty}(-1)^k \bar{\rho}^k H^{k+1}(f) \tag{22}$$

Assumption A. There exists state $i_0 \in \mathcal{I}$ that is accessible from any state $i \in \mathcal{I}$ for every $f \in \mathcal{F}$, i.e. for every $f \in \mathcal{F}$ the transition probability matrix $P(f)$ is *unichain*.

Lemma 3.1. If Assumption A holds then

$$\alpha\, v^{\alpha(1)}(f) = \rho^{-1}\, \bar{g}^{(1)}(f) \cdot e + w^{(1)}(f) + \sum_{k=1}^{\infty}(-\rho)^k\, w^{(k,1)}(f) \tag{23}$$

where $\bar{g}^{(1)}(f) = p^*(f) \cdot r^{(1)}(f)$, $w^{(1)}(f) = H(f) \cdot r^{(1)}(f)$, and $w^{(k,1)}(f) = H^{k+1}(f) \cdot r^{(1)}(f)$, for $k = 1, 2, \ldots$. In particular, for the ith element of $v^{\alpha(1)}(f)$ it holds

$$\alpha v_i^{\alpha(1)}(f) = \rho^{-1}\bar{g}^{(1)}(f) + w_i^{(1)}(f) + \sum_{k=1}^{\infty}(-\rho)^k\, w_i^{(k,1)}(f), \quad \text{hence}$$

$$v_i^{\alpha(1)}(f) = (1+\rho)\left[\rho^{-1}\bar{g}^{(1)}(f) + w_i^{(1)}(f) - \rho w_i^{(1,1)}(f) + \rho^2 w_i^{(2,1)}(f) + o(\rho^2)\right] \tag{24}$$

where $\lim\limits_{\rho \downarrow 0} o(\rho^2) = 0$.

In what follows we construct partial Laurent expansion of discounted variance $\sigma_i^{\alpha}(f)$. To this end from (16),(17) we conclude that

$$\sigma_i^{\alpha}(f) - \alpha^2 \sum_{j \in \mathcal{I}} p_{ij}(f_i) \cdot \sigma_j^{\alpha}(f) = \sum_{j \in \mathcal{I}} p_{ij}(f_i)[r_{ij} + \alpha \cdot v_j^{\alpha(1)}(f)]^2 - [v_i^{\alpha(1)}(f)]^2 \tag{25}$$

To simplify the RHS of (25) the following facts will be extremely useful. Observe that (26), (27) follow from (24), and (28) follows from (24), (26) and (27) after some algebraic manipulations.

Lemma 3.2. If Assumption A holds then

$$\begin{aligned}\alpha^2[v_i^{\alpha(1)}(f)]^2 = {} & [\rho^{-1} \cdot \bar{g}^{(1)}(f) + w_i^{(1)}(f)]^2 - 2 \cdot \bar{g}^{(1)}(f) \cdot w_i^{(1,1)}(f) \\ & + 2\rho \cdot [\bar{g}^{(1)}(f) \cdot w_i^{(2,1)}(f) + w_i^{(1)}(f) \cdot w_i^{(1,1)}(f)] + o(\rho^2)\end{aligned} \tag{26}$$

$$\begin{aligned}(v_i^{\alpha(1)}(f))^2 = {} & \rho^{-2}[\bar{g}^{(1)}(f)]^2 + 2\rho^{-1}\bar{g}^{(1)}(f)[\bar{g}^{(1)}(f) + w_i^{(1)}(f)] + [\bar{g}^{(1)}(f)]^2 + [w_i^{(1)}(f)]^2 + \\ & \bar{g}^{(1)}(f)[4w_i^{(1)}(f) - 2w_i^{(1,1)}(f)] + 2\rho \cdot [\bar{g}^{(1)}(f) \cdot w_i^{(2,1)}(f) - w_i^{(1)}(f) \cdot w_i^{(1,1)}(f) \\ & + [w_i^{(1)}(f)]^2 - 2 \cdot \bar{g}^{(1)}(f) \cdot w_i^{(1,1)}(f) + \bar{g}^{(1)}(f) \cdot w_i^{(1)}(f)] + o(\rho^2)\end{aligned} \tag{27}$$

$$\begin{aligned}[r_{ij} + \alpha v_j^{\alpha(1)}(f)]^2 = {} & [r_{ij}]^2 + 2 \cdot \alpha \cdot v_j^{\alpha(1)}(f) \cdot r_{ij} + \alpha^2[v_j^{\alpha(1)}(f)]^2 \\ = {} & [\rho^{-1}\bar{g}^{(1)}(f) + w_j^{(1)}(f)]^2 - 2 \cdot \bar{g}^{(1)}(f) \cdot w_j^{(1,1)}(f) + [r_{ij}]^2 + 2\rho \cdot [\bar{g}^{(1)}(f) \cdot w_j^{(2,1)}(f) \\ & + w_j^{(1)}(f) \cdot w_j^{(1,1)}(f)] + 2 \cdot r_{ij} \cdot [\rho^{-1}\bar{g}^{(1)}(f) + w_j^{(1)}(f) + \rho w_j^{(1,1)}(f)] + o(\rho^2)\end{aligned} \tag{28}$$

Lemma 3.3. If Assumption A holds then

$$\sum_{j\in\mathcal{I}} p_{ij}(f_i)\left[r_{ij}+\alpha\cdot v_j^{\alpha(1)}(f)\right]^2-[v_i^{\alpha(1)}(f)]^2=\sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij}+w_j]^2-[\bar{g}^{(1)}(f))-w_i]^2+O(\rho)+o(\rho^2) \quad (29)$$

where

$$O(\rho)=2\rho\sum_{j\in\mathcal{I}} p_{ij}(f_i)w_j^{(1,1)}(f)[w_j^{(1)}(f)+r_{ij}]+[w_i^{(1)}(f)+\bar{g}^{(1)}(f)]w_i^{(1,1)}(f)+[w_i^{(1)}(f)-\bar{g}^{(1)}(f)]w_i^{(1)}(f). \quad (30)$$

Proof. In virtue of (27),(28) we can conclude that

$$\begin{aligned}
&\sum_{j\in\mathcal{I}} p_{ij}(f_i)\left[r_{ij}+\alpha\cdot v_j^{\alpha(1)}(f)\right]^2-[v_i^{\alpha(1)}(f)]^2=\sum_{j\in\mathcal{I}} p_{ij}(f_i)\left\{\left[r_{ij}+\alpha\cdot v_j^{\alpha(1)}(f)\right]^2-[v_i^{\alpha(1)}(f)]^2\right\}\\
&=\sum_{j\in\mathcal{I}} p_{ij}(f_i)\left\{[\rho^{-1}\bar{g}^{(1)}(f)+w_j^{(1)}(f)]^2-\ 2\cdot\bar{g}^{(1)}(f)\cdot w_j^{(1,1)}(f)+[r_{ij}]^2\right.\\
&+2\rho\cdot[\bar{g}^{(1)}(f)\cdot w_j^{(2,1)}(f)+w_j^{(1)}(f)\cdot w_j^{(1,1)}(f)]+\ 2\cdot r_{ij}\cdot[\rho^{-1}\bar{g}^{(1)}(f)+w_j^{(1)}(f)+\rho w_j^{(1,1)}(f)]\\
&-\rho^{-2}[\bar{g}^{(1)}(f)]^2-2\rho^{-1}\bar{g}^{(1)}(f)[\bar{g}^{(1)}(f)+w_i^{(1)}(f)]-[\bar{g}^{(1)}(f)]^2-[w_i^{(1)}(f)]^2\\
&-\bar{g}^{(1)}(f)[4w_i^{(1)}(f)-2w_i^{(1,1)}(f)]-2\rho\cdot[\bar{g}^{(1)}(f)\cdot w_i^{(2,1)}(f)-w_i^{(1)}(f)\cdot w_i^{(1,1)}(f)\\
&\left.+[w_i^{(1)}(f)]^2-2\cdot\bar{g}^{(1)}(f)\cdot w_i^{(1,1)}(f)-\bar{g}^{(1)}(f)\cdot w_i^{(1)}(f)]\right\}\ +o(\rho^2) \qquad (31)\\
&\text{Since}\qquad \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij}+w_j^{(1)}(f)-w_i^{(1)}(f)-\bar{g}^{(1)}(f)]=0\\
&\sum_{j\in\mathcal{I}} p_{ij}(f_i)[-w_i^{(1)}(f)+w_j^{(1,1)}(f)-w_i^{(1,1)}(f)]=0,\quad \sum_{j\in\mathcal{I}} p_{ij}(f_i)[w_i^{(1,1)}(f)+w_j^{(2,1)}(f)-w_i^{(2,1)}(f)]=0
\end{aligned}$$

(29) follows from (31) after some algebra.

From (25),(29),(30) we immediately get the following lemma.

Lemma 3.4. If Assumption A holds then

$$\sigma_i^{\alpha}(f)-\alpha^2\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^{\alpha}(f)=s_i(f_i)+O(\rho)+o(\rho)\quad\text{for}\quad i=1,2,\ldots,N \qquad (32)$$

where

$$s_i(f_i):=\sum_{j\in\mathcal{I}} p_{ij}(f_i)\left\{[r_{ij}+w_j^{(1)}(f)]^2-[\bar{g}^{(1)}(f)+w_i^{(1)}(f)]^2\right\}+O(\rho)+o(\rho)$$

For what follows it will be useful to rewrite (32) in matrix notation. On introducing column vector $s(f)=[s_i(f_i)]_{i=1,\ldots,N}$ (32) can written as

$$\sigma^{\alpha}(f)=(I-\alpha^2P(f))^{-1}s(f)+O(\rho)+o(\rho) \qquad (33)$$

Next lemma adapts Laurent expansion of $(I-\alpha P(f))^{-1}$ to $(I-\alpha^2P(f))^{-1}$. Recall that $\bar{\rho}:=\frac{1-\alpha^2}{\alpha^2}$.

Lemma 3.5. If Assumption A holds then

$$(I-\alpha^2P(f))^{-1}\quad=\quad\frac{(\bar{\rho}+1)}{\bar{\rho}}P^*(f)+(\bar{\rho}+1)H(f)+o(\bar{\rho}) \qquad (34)$$

Observe that $\alpha\to 1\Rightarrow\bar{\rho}\to 0$. In particular, $\lim_{\alpha\to 1}(1-\alpha^2)(I-\alpha^2P(f))^{-1}=P^*(f)$.

Proof. The proof follows immediately from (22).

From (33), (34) we immediately get

Theorem 3.6. If Assumption A holds then

$$\sigma^{\alpha}(f) = \frac{(\bar{\rho}+1)}{\bar{\rho}} P^*(f) + (\bar{\rho}+1)s(f) + O(\rho) + o(\rho^2). \tag{35}$$

Moreover, from the well-known Tauberian theorems it holds for undiscounted variance obtain after n transitions

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \sigma(f,n) = \lim_{\alpha\to 1}(1-\alpha)\sigma^{\alpha}(f) = P^*(f)s(f).$$

5 Conclusions

We have received formulas for the variance of discounted rewards in Markov decision chains along with it partial Laurent expansions. Attention was focused only on unichain models and initial terms of the corresponding Laurent expansion that also enables to find formulas for mean variances of undiscounted rewards.

Acknowledgements

This research was supported by the Czech Science Foundation under Grant 13-14445S and by CONACyT (Mexico) and AS CR (Czech Republic) under Project 171396.

References

[1] Berman, A. and Plemmons, R. J.: *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York 1979.

[2] Feinberg, E.A. and Fei, J.: *Inequalities for variances of total discounted costs.* J. Applied Probability **46** (2009), 1209–1212.

[3] Kato, T.: *Perturbation Theory for Linear Operators.* Springer Verlag, Berlin 1966.

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

[5] Miller, B.L. and Veinott, A.F.Jr.: *Discrete dynamic programming with a small interest rate.* Annals Math. Statistics **39** (1968), 366–370.

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

[7] Righter, R.: *Stochastic comparison of discounted rewards.* J. Applied Probability **48** (2011), 293–294.

[8] Sladký, K.: *On the set of optimal controls for Markov chains with rewards.* Kybernetika **10** (1974), 526–547.

[9] Sladký, K.: *On mean reward variance in semi-Markov processes.* Math. Methods Oper. Res. **62** (2005), 387–397.

[10] Sladký, K.: *Risk sensitive and risk-neutral optimality in Markov decision chains; a unified approach.* In: Proceedings of the International Scientific Conference Quantitative Methods in Economics (Multiple Criteria Decision Making XVI), M. Reiff, ed., University of Economics, Bratislava 2012, pp. 201–205.

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

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