J. Appl. Prob. **43,** 1044–1052 (2006)
Printed in Israel

ON THE TOTAL REWARD VARIANCE FOR CONTINUOUS-TIME MARKOV REWARD CHAINS

NICO M. VAN DIJK,* *University of Amsterdam*
KAREL SLADKÝ,** *Institute of Information Theory and Automation, Prague*

Abstract

As an extension of the discrete-time case, this note investigates the variance of the total cumulative reward for continuous-time Markov reward chains with finite state spaces. The results correspond to discrete-time results. In particular, the variance growth rate is shown to be asymptotically linear in time. Expressions are provided to compute this growth rate.

Keywords: Continuous-time Markov reward chain; variance of cumulative reward; asymptotic behaviour; uniformization

2000 Mathematics Subject Classification: Primary 90C47
Secondary 60J27

1. Introduction

1.1. Motivation

The usual optimization criteria examined in the literature for the optimization of Markov reward processes, such as a total discounted or average reward structure, can be quite insufficient to fully capture the various aspects considered by a decision maker. It may be preferable, if not necessary, to select or to include more sophisticated criteria that also reflect variability risk features of the problem. Most notably, the variance of the cumulative reward can be indicative and seems of interest. For a detailed discussion of such approaches, see the review paper by White [20].

To the best of the authors' knowledge (and with the exception of [7], which will be discussed below), higher moments and the variance of the cumulative reward in Markov reward chains have been systematically studied only for discrete-time models. Research in this direction was initiated by Mandl [12], Jaquette [5], [6], [8], Benito [1], and Sobel [17]. More recent extensions of these results can be found in [11], [9], and [16]. In particular, in these references the variance (or second moment) of the total expected discounted or average rewards of controlled, discrete-time Markov reward chains was considered, to determine the 'best' policy within the class of discounted (or average) optimal policies and find a smaller variance (or lower second moment) of the cumulative reward.

Alternatively, criteria reflecting the variability or risk features of policies not restricted to the class of optimal policies have been investigated in the literature on Markovian decision models. More precisely, Sobel [18] maximized the ratio of the mean to the standard deviation using the methods of nonlinear and parametric linear programming. Similarly, Kawai [10] considered

Received 3 February 2005; revision received 28 August 2006.

* Postal address: Department of Economic Sciences and Econometrics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. Email address: nivd@fee.uva.nl

** Postal address: Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, PO Box 18, Pod Vodárenskou věží 4, 182 08 Prague 8, Czech Republic. Email address: sladky@utia.cas.cz

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press

the problem of minimizing the variance subject to a lower bound on the average reward. Filar *et al.* [3] proposed a mathematical programming approach for mean–variance Markov decision chains. Huang and Kallenberg [4] unified and extended the formulations and existence results obtained in [3], [10], and [18]. Finally, in [16] it was shown that optimal policies with respect to the standard mean–variance optimality criteria can be found in vertices of a special convex polyhedron, and a policy iteration method was suggested to find these vertices. Here it is important to note that, in these papers for finding the optimum policy with respect to various mean–variance optimality criteria, the 'variance' is considered only with respect to one-stage reward variances and not as the variance of the cumulative reward. To date, however, no results for the continuous-time case seem to have been reported.

The only exception to this seems to be in [7]. That paper, which dealt with the discounted reward case, provided a characterization of moment optimal policies. More precisely, for the discounted reward case it showed that a moment optimal policy can always be found within the class of piecewise-constant policies. However, no explicit expressions for these moments or for the total cumulative reward or its asymptotic behaviour were provided in [7].

1.2. Objective and results

In this note, therefore, we aim to investigate whether the results established for the discrete-time case can be extended to continuous-time Markov reward chains. As the essential step is an expression for the variance of the cumulative reward and its asymptotic behaviour, in this note the presentation will be restricted to the *uncontrolled* case. The implication for the controlled case will be briefly discussed (see Remark 3.3). The formulae obtained are similar to those for the discrete-time case. In addition to reward rates, we also consider transition rewards. In particular, we show that the variance of the total reward has a growth rate that is asymptotically linear in time. Relations are provided to compute this growth rate.

1.3. Formulation

Consider a continuous-time Markov reward process with finite state space $\mathcal{S} = \{1, 2, \dots, N\}$ and a transition and reward structure characterized by

q_{ij}, the transition rate for a transition from i to j $(i, j \in \mathcal{S}, j \neq i)$,
with $q_{ii} = -\sum_{j\in\mathcal{S},\, j\neq i} q_{ij}$,

r_{ij}, the instantaneous transition reward for a transition from i to j, and

r_i, the reward rate in state i.

Let the vectors $\boldsymbol{R}(t)$, $\boldsymbol{S}(t)$, and $\boldsymbol{V}(t)$ respectively denote the first moment, the second moment, and the variance of the total reward up to time t, given its initial state at time $t = 0$. More precisely,

$$R_i(t) = \mathrm{E}[\xi(t) \mid X(0) = i],$$
$$S_i(t) = \mathrm{E}[\xi^2(t) \mid X(0) = i],$$
$$V_i(t) = \sigma^2[\xi(t) \mid X(0) = i],$$

where

$$\xi(t) = \int_0^t r_{X(s)}\,\mathrm{d}s + \sum_{k=0}^{N(t)} r_{X(\tau_k^-),X(\tau_k^+)},$$

with $X(s)$ denoting the state of the system at time s, $X(\tau_k^-)$ and $X(\tau_k^+)$ the states just before and after the kth jump, respectively, and $N(t)$ the number of jumps up to time t.

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press

In the literature on dynamic programming (e.g. [14] and [15]) the vector $\boldsymbol{R}(\cdot)$ is well known as the value function, which, because of the additive reward structure, can also be written as

$$\boldsymbol{R}(t) = \int_0^t \boldsymbol{P}(s)\tilde{\boldsymbol{r}}\,\mathrm{d}s \qquad \text{or} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{R}(t) = \tilde{\boldsymbol{r}} + \boldsymbol{Q}\boldsymbol{R}(t) \quad \text{with} \quad \boldsymbol{R}(0) = \boldsymbol{0}, \tag{1.1}$$

where

$\boldsymbol{P}(t) = [p_{ij}(t)]$, the transition probability matrix over time t,

$\boldsymbol{Q} = [q_{ij}]$, the infinitesimal generator, with $\sum_{j\in\mathcal{S}} q_{ij} = 0$, and

$\tilde{\boldsymbol{r}} = [\tilde{r}_i]$, the column N-vector of expected reward rates, with $\tilde{r}_i = r_i + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}r_{ij}$.

Remark 1.1. (*Transition rewards.*) 1. Note that the transition rewards r_{ij} are less natural as instantaneous rewards in a discrete-time setting than in a continuous-time setting. Furthermore, as shown above, they can be included in an expected reward rate $\tilde{\boldsymbol{r}}$ for determining $\boldsymbol{R}(\cdot)$.

2. In contrast, the actual state transition of the state process itself, and its corresponding reward consequences, will be of influence on the second moment and the variance of the total reward. In analysing the variance, therefore, the instantaneous transition rewards cannot be included in the reward rate and are to be kept separate. This will also become apparent in the expressions that will be derived below.

Remark 1.2. (*Exponential convergence.*) By $\varepsilon(t)$ we denote a function of t such that $\varepsilon(t) \to 0$ exponentially quickly as $t \to \infty$, i.e. for some α and β, $|\varepsilon(t)| \leq \alpha \mathrm{e}^{-\beta t}$. By $\boldsymbol{\varepsilon}(t)$ we denote a vector function such that, for all i, $\varepsilon_i(t) \to 0$ exponentially quickly as $t \to \infty$. Furthermore, for any γ we write $\boldsymbol{\gamma}$ for the vector with $\gamma_i = \gamma$ for all i. By $\boldsymbol{I}$ we denote the identity matrix, and by $\boldsymbol{\pi}$ the row vector of steady state probabilities determined by $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$.

2. Total reward variance for finite horizon

Let E_i denote the conditional expectation given that $X(0) = i$, and note that $\xi(t+\Delta) = \xi(\Delta) + \xi^{(\Delta,t+\Delta)}$, where $\xi^{(\Delta,t+\Delta)}$ denotes the total (random) reward obtained in the time interval $[\Delta, t+\Delta)$. Hence,

$$\mathrm{E}_i[\xi(t+\Delta)] = \mathrm{E}_i[\xi(\Delta)] + \mathrm{E}_i[\xi^{(\Delta,t+\Delta)}],$$
$$\mathrm{E}_i[\xi(t+\Delta)]^2 = \mathrm{E}_i[\xi(\Delta)]^2 + \mathrm{E}_i[\xi^{(\Delta,t+\Delta)}]^2 + 2\,\mathrm{E}_i[\xi^{(\Delta,t+\Delta)}\xi(\Delta)].$$

Then, since $\boldsymbol{P}(\Delta) = \boldsymbol{I} + \Delta\boldsymbol{Q} + \boldsymbol{o}(\Delta^2)$; since the probability of more than one transition occurring in time Δ is of order Δ^2; since in the case of a transition in time Δ, say from i to j (for which the probability is of order Δ), the reward incurred during that interval is of the form $r_{ij} + o(\Delta)$; and since the continuous-time Markov reward process considered is time homogeneous, we obtain

$$\begin{aligned} R_i(t+\Delta) &= \Delta r_i + (1+\Delta q_{ii})R_i(t) \\ &\qquad + \Delta \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij} + R_j(t)\} + o(\Delta^2), \\ S_i(t+\Delta) &= (1+\Delta q_{ii})\{2\Delta r_i R_i(t) + S_i(t)\} \\ &\qquad + \Delta \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij}^2 + 2r_{ij}R_j(t) + S_j(t)\} + o(\Delta^2). \end{aligned}$$

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}R_i(t) = r_i + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}r_{ij} + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{R_j(t) - R_i(t)\}, \tag{2.1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S_i(t) = 2r_iR_i(t) + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij}^2 + 2r_{ij}R_j(t)\} + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{S_j(t) - S_i(t)\}. \tag{2.2}$$

From $V_i(t) = S_i(t) - R_i(t)^2$ we thus obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V_i(t) &= \frac{\mathrm{d}}{\mathrm{d}t}S_i(t) - 2R_i(t)\frac{\mathrm{d}}{\mathrm{d}t}R_i(t)\\ &= 2r_iR_i(t) + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij}^2 + 2r_{ij}R_j(t)\} + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{S_j(t) - S_i(t)\}\\ &\quad - 2R_i(t)\biggl(r_i + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}r_{ij} + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{R_j(t) - R_i(t)\}\biggr).\end{aligned} \tag{2.3}$$

By making the substitutions $S_j(t) = V_j(t) + R_j(t)^2$ and $-\sum_{j\in\mathcal{S},\, j\neq i} q_{ij} = q_{ii}$ in (2.3), it can be rewritten as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}V_i(t) &= 2r_iR_i(t) + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{[r_{ij} + R_j(t)]^2 - R_i(t)^2\} + \sum_{j\in\mathcal{S}} q_{ij}V_j(t)\\ &\quad - 2R_i(t)\biggl(r_i + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}r_{ij} + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{R_j(t) - R_i(t)\}\biggr)\\ &= \sum_{j\in\mathcal{S}} q_{ij}V_j(t) - 2R_i(t)\biggl(\sum_{j\in\mathcal{S},\, j\neq i} q_{ij}r_{ij} + \sum_{j\in\mathcal{S}} q_{ij}R_j(t)\biggr)\\ &\quad + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij} + R_j(t) + R_i(t)\}\{r_{ij} + R_j(t) - R_i(t)\}.\end{aligned} \tag{2.4}$$

3. Infinite horizon

Assume that the Markov chain has a single class of recurrent states. The average reward, g, is then well defined, independently of the initial state i at time 0, by

$$g = \lim_{t\to\infty} \frac{1}{t}R_i(t).$$

In addition, by dynamic programming (see, e.g. [14] and [15]) it is well known that there exists a vector $\boldsymbol{w}$ such that

$$\boldsymbol{R}(t) = \boldsymbol{g}t + \boldsymbol{w} + \boldsymbol{\varepsilon}(t); \tag{3.1}$$

hence, $\boldsymbol{R}(t)$ has a growth rate linear in t up to a vector $\boldsymbol{w}$ and a term converging exponentially quickly to 0 as $t \to \infty$. The vector $\boldsymbol{w}$ is the relative gain (or bias) vector, determined by

$$g = r_i + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij} + w_j - w_i\}. \tag{3.2}$$

Note that the w_j are uniquely determined by (3.2) up to an additive constant. Under the additional condition $\boldsymbol{\pi}\boldsymbol{w} = 0$, the w_j are the unique solution to (3.2).

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press

Now, by

$$\begin{aligned} R_j(t) + R_i(t) &= 2gt + w_j + w_i + \varepsilon(t), \\ R_j(t) - R_i(t) &= w_j - w_i + \varepsilon(t), \end{aligned} \tag{3.3}$$

for the last term in (2.4) we can write

$$\begin{aligned} &\sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij} + R_j(t) + R_i(t)\}\{r_{ij} + R_j(t) - R_i(t)\} \\ &\quad = \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{[r_{ij} + w_j]^2 - w_i^2\} + 2gt \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij} + w_j - w_i\} + \varepsilon(t) \\ &\quad = \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{[r_{ij} + w_j]^2 - w_i^2\} + 2g^2 t - 2gtr_i + \varepsilon(t). \end{aligned}$$

Furthermore, by again using (3.3) and (3.2), for the second term of (2.4) we obtain

$$\begin{aligned} &2R_i(t)\bigg(\sum_{j\in\mathcal{S},\, j\neq i} q_{ij}r_{ij} + \sum_{j\in\mathcal{S}} q_{ij}R_j(t)\bigg) \\ &\quad = 2R_i(t)\bigg(\sum_{j\in\mathcal{S},\, j\neq i} q_{ij}r_{ij} + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{R_j(t) - R_i(t)\}\bigg) \\ &\quad = 2(gt + w_i + \varepsilon(t))\bigg(\sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij} + w_j - w_i\} + \varepsilon(t)\bigg) \\ &\quad = 2g^2 t - 2gr_i t + 2w_i(g - r_i) + \varepsilon(t). \end{aligned}$$

Substitution into (2.4) yields

$$\frac{\mathrm{d}}{\mathrm{d}t}V_i(t) = \sum_{j\in\mathcal{S}} q_{ij}V_j(t) + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{[r_{ij} + w_j]^2 - w_i^2\} + 2(r_i - g)w_i + \varepsilon(t).$$

Hence, in matrix form, and with the vector $\boldsymbol{s}$ defined by

$$s_i = \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{[r_{ij} + w_j]^2 - w_i^2\} + 2(r_i - g)w_i,$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{V}(t) = \boldsymbol{s} + \boldsymbol{Q}\boldsymbol{V}(t) + \boldsymbol{\varepsilon}^{(1)}(t), \tag{3.4}$$

where all elements of the column N-vector $\boldsymbol{\varepsilon}^{(1)}(t)$ converge to 0 exponentially quickly, implying that

$$\|\boldsymbol{\varepsilon}^{(1)}(t)\| \leq c\mathrm{e}^{-\delta t} \tag{3.5}$$

for some numbers $c > 0$ and $\delta > 0$, where $\|\cdot\|$ is the standard ∞ norm.

3.1. Growth rate

Now, in order to investigate the behaviour of $\boldsymbol{V}(t)$ for large t, let

$$\boldsymbol{X}(t) = \boldsymbol{V}(t) - \boldsymbol{W}(t), \tag{3.6}$$

where $\boldsymbol{W}(t)$ is defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{W}(t) = \boldsymbol{s} + \boldsymbol{Q}\boldsymbol{W}(t) \tag{3.7}$$

and $\boldsymbol{W}(0) = \mathbf{0}$. In analogy with (3.4) and in correspondence to (1.1), (3.1), and (3.2), $\boldsymbol{W}(t)$ can thus be regarded as the expected cumulative reward up to time t with reward rate vector $\boldsymbol{s}$, and can be written as

$$\boldsymbol{W}(t) = \int_0^t \boldsymbol{P}(u)\boldsymbol{s}\,\mathrm{d}u = \boldsymbol{\gamma}t + \boldsymbol{h} + \boldsymbol{\varepsilon}^{(2)}(t), \tag{3.8}$$

where the function $\boldsymbol{\varepsilon}^{(2)}(t)$ is again a vector that converges to $\mathbf{0}$ exponentially quickly as $t \to \infty$ and the growth rate γ and the vector $\boldsymbol{h}$ are determined by

$$\gamma = s_i + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}(h_j - h_i), \qquad \sum_{i\in\mathcal{S}} \pi_i h_i = 0, \tag{3.9}$$

with

$$\gamma = \lim_{t\to\infty} \frac{1}{t} W_i(t) \tag{3.10}$$

for any $i \in \mathcal{S}$.

Lemma 3.1. *With $\boldsymbol{\varepsilon}^{(1)}(t)$ as in (3.4),*

$$\boldsymbol{X}(t) = \int_0^t \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u \tag{3.11}$$

and, for some positive constants c and δ,

$$\|\boldsymbol{X}(t)\| \leq \frac{1}{\delta}c[1 - \mathrm{e}^{-\delta t}]. \tag{3.12}$$

Proof. By (3.4), (3.6), and (3.7),

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{X}(t) = \boldsymbol{\varepsilon}^{(1)}(t) + \boldsymbol{Q}\boldsymbol{X}(t), \tag{3.13}$$

and by the uniqueness of its solution (see, e.g. [2, p. 23]) it suffices to show that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^t \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u = \boldsymbol{\varepsilon}^{(1)}(t) + \boldsymbol{Q}\int_0^t \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u.$$

To this end, we can write

$$\begin{aligned}
&\int_0^{t+\Delta} \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t+\Delta-u)\,\mathrm{d}u - \int_0^t \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u \\
&\quad= \int_0^{\Delta} \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t+\Delta-u)\,\mathrm{d}u + \int_{\Delta}^{t+\Delta} \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t+\Delta-u)\,\mathrm{d}u \\
&\qquad - \int_0^t \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u \\
&\quad= \int_0^{\Delta} \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)[1+o(1)]\,\mathrm{d}u + \int_0^t \boldsymbol{P}(u+\Delta)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u \\
&\qquad - \int_0^t \boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u \\
&\quad= \boldsymbol{\varepsilon}^{(1)}(t)[1+o(1)]\Delta + \int_0^t [\boldsymbol{P}(\Delta) - \boldsymbol{I}]\boldsymbol{P}(u)\boldsymbol{\varepsilon}^{(1)}(t-u)\,\mathrm{d}u.
\end{aligned}$$

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press

By dividing the left- and right-hand sides of this equation by Δ, letting $\Delta \to 0$, and using $[\boldsymbol{P}(\Delta)-\boldsymbol{I}]/\Delta \to \boldsymbol{Q}$ (in the strong sense (see [2, p. 23])), we can show that (3.11) satisfies (3.13). Inequality (3.12) follows directly from the exponential convergence (3.5), since

$$\|\boldsymbol{X}(t)\| \leq \int_0^t c\mathrm{e}^{-\delta(t-u)}\,\mathrm{d}u = \frac{1}{\delta}c[1-\mathrm{e}^{-\delta t}].$$

Theorem 3.1. *For a (constant) growth rate vector $\boldsymbol{\gamma}$ and vector $\boldsymbol{h}$ as determined by (3.9) and (3.10), some constant C, and some exponentially quickly converging vector function $\boldsymbol{\varepsilon}(t)$, we have*

$$\boldsymbol{V}(t) = \boldsymbol{\gamma}t + \boldsymbol{h} + \boldsymbol{c}(t) + \boldsymbol{\varepsilon}(t) \quad \textit{with } \|\boldsymbol{c}(t)\| \leq C \textit{ for all } t\,. \tag{3.14}$$

Proof. Equation (3.14) follows directly by combining (3.4), (3.8), and (3.12) and combining the exponentially converging terms $\boldsymbol{\varepsilon}^{(1)}(t)$ and $\boldsymbol{\varepsilon}^{(2)}(t)$ for $\boldsymbol{X}(t)$ and $\boldsymbol{W}(t)$.

In words, the theorem states that $\boldsymbol{V}(t)$ has a linear growth rate up to a bounded bias function and an exponential convergence.

3.2. Computation

The growth rate γ can in principle be computed using standard methods as a solution to the set of linear equations (3.9). However, we can also employ successive approximation. This will generate monotone lower and upper bounds converging to γ. To this end, choose a $B < \infty$ such that $\sum_{j\in\mathcal{S},\, j\neq i} q_{ij} < B$ and let the functions $\boldsymbol{W}^{(k)}$, $k = 0, 1, 2, \dots$, be defined recursively by

$$\boldsymbol{W}^{(0)} = \boldsymbol{0}, \qquad \boldsymbol{W}^{(k+1)} = \boldsymbol{s} + \boldsymbol{P}\boldsymbol{W}^{(k)}, \quad k = 0, 1, 2, \dots, \tag{3.15}$$

where the elements of $\boldsymbol{P}$ are defined by

$$p_{ij} = \begin{cases} \dfrac{q_{ij}}{B}, & j \neq i, \\ 1 - \displaystyle\sum_{j\in\mathcal{S},\, j\neq i} \frac{q_{ij}}{B}, & j = i. \end{cases}$$

By the standard step of uniformization (see, e.g. [19, p. 154]) and results for dynamic programming (see, e.g. [13] and [19, p. 207]), the linear growth rate γ of the variance defined by (3.4) can then be approximated as the average reward of the Markov chain with reward rate $\boldsymbol{s}$, by

$$M_n = \max_{i\in\mathcal{S}} |W_i^{(n+1)} - W_i^{(n)}|B, \qquad m_n = \min_{i\in\mathcal{S}} |W_i^{(n+1)} - W_i^{(n)}|B.$$

The values m_n and M_n are then monotonically convergent to γ, and $m_n \leq \gamma \leq M_n$.

Remark 3.1. Since $\gamma = \boldsymbol{\pi s}$ and (recall) $\boldsymbol{\pi Q} = \boldsymbol{0}$, we have $\gamma = \boldsymbol{\pi s}^{(1)} = \boldsymbol{\pi s}^{(2)}$, where the elements of $\boldsymbol{s}^{(1)}$ are defined by

$$s_i^{(1)} = \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{[r_{ij} + w_j]^2 - w_i^2\} + 2r_i w_i$$

and those of $\boldsymbol{s}^{(2)}$ are defined by

$$s_i^{(2)} = \sum_{j\in\mathcal{S},\, j\neq i} q_{ij}\{r_{ij}^2 + 2r_{ij}w_j\} + 2r_i w_i\,.$$

As a consequence, to compute γ we can also replace $\boldsymbol{s}$ by $\boldsymbol{s}^{(1)}$ or $\boldsymbol{s}^{(2)}$ in (3.15).

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press

Remark 3.2. (*Transient and discounted case.*) Assume that, for every $i = 1, \dots, N$, there exist (finite) limits $R_i = \lim_{t\to\infty} R_i(t)$ and $S_i = \lim_{t\to\infty} S_i(t)$ (whence $\boldsymbol{g} = \mathbf{0}$ and, similarly, $\gamma := \lim_{t\to\infty} S_i(t)/t = 0$). Then, from (2.1)–(2.4), we immediately conclude that the limits $V_i = \lim_{t\to\infty} V_i(t)$ also exist and satisfy (see (2.4))

$$0 = \sum_{j\in\mathcal{S}} q_{ij} V_j - 2R_i \Bigg(\sum_{j\in\mathcal{S},\, j\neq i} q_{ij} r_{ij} + \sum_{j\in\mathcal{S}} q_{ij} R_j \Bigg) + \sum_{j\in\mathcal{S},\, j\neq i} q_{ij} \{[r_{ij} + R_j]^2 - R_i^2\}.$$

This situation typically arises when the chain is absorbing in states with $r_i = 0$. A similar expression, with a discounted factor β included, can be derived for the total cumulative discounted reward, $\boldsymbol{R}_\beta$.

Remark 3.3. (*Controlled case.*) In this note we have restricted our attention to uncontrolled, continuous-time Markov reward chains. For controlled models (i.e. with transition rates and transition rewards depending on a decision), the results can be extended immediately for a given stationary policy. Results can then be expected that are related to those on selecting the 'best' optimal stationary policy with smallest (minimal) variance in the discrete-time case (see, e.g. [9], [11], and [12]). However, as this 'optimization' will be notationally more complex and requires a number of technicalities and results from Markov decision theory, the details and results are left for further research. Nevertheless, the essential first step to this end is Theorem 3.1. The situation with nonstationary policies and other optimization criteria also remains a challenging topic for future research.

Acknowledgement

This research was supported by the Grant Agency of the Czech Republic under grants 402/05/0115 and 402/04/1294.

References

[1] Benito, F. (1982). Calculating the variance in Markov-processes with random reward. *Trabajos Estadíst. Investigación Operat.* **33,** 73–85.

[2] Dynkin, E. B. (1965). *Markov Processes*, Vol. I. Springer, Berlin.

[3] Filar, J., Kallenberg, L. C. M. and Lee, H.-M. (1989). Variance penalized Markov decision processes. *Math. Operat. Res.* **14,** 147–161.

[4] Huang, Y. and Kallenberg, L. C. M. (1994). On finding optimal policies for Markov decision chains: a unifying framework for mean-variance-tradeoffs. *Math. Operat. Res.* **19,** 434–448.

[5] Jaquette, S. C. (1972). Markov decision processes with a new optimality criterion: small interest rates. *Ann. Math. Statist.* **43,** 1894–1901.

[6] Jaquette, S. C. (1973). Markov decision processes with a new optimality criterion: discrete time. *Ann. Statist.* **1,** 496–505.

[7] Jaquette, S. C. (1975). Markov decision processes with a new optimality criterion: continuous time. *Ann. Statist.* **3,** 547–553.

[8] Jaquette, S. C. (1976). A utility criterion for Markov decision processes. *Manag. Sci.* **23,** 43–49.

[9] Kadota, Y. (1997). A minimum average-variance in Markov decision processes. *Bull. Inf. Cybernet.* **29,** 83–89.

[10] Kawai, H. (1987). A variance minimization problem for a Markov decision process. *Europ. J. Operat. Res.* **31,** 140–145.

[11] Kurano, M. (1987). Markov decision processes with a minimum-variance criterion. *J. Math. Anal. Appl.* **123,** 572–583.

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

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

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

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press

[15] Ross, S. M. (1970). *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco, CA.

[16] Sladký, K. and Sitař, M. (2004). Optimal solutions for undiscounted variance penalized Markov decision chains. In *Dynamic Stochastic Optimization* (Lecture Notes Econom. Math. Systems **532**), eds K. Marti, Y. Ermoliev and G. Pflug, Springer, Berlin, pp. 43–66.

[17] Sobel, M. J. (1982). The variance of discounted Markov decision processes. *J. Appl. Prob.* **19,** 794–802.

[18] Sobel, M. J. (1985). Maximal mean/standard deviation ratio in an undiscounted MDP. *Operat. Res. Lett.* **4,** 157–159.

[19] Tijms, H. C. (1994). *Stochastic Models. An Algebraic Approach.* John Wiley, Chichester.

[20] White, D. J. (1988). Mean, variance and probability criteria in finite Markov decision processes: A review. *J. Optimization Theory Appl.* **56,** 1–29.

https://doi.org/10.1239/jap/1165505206 Published online by Cambridge University Press