Second Order Optimality in Markov and Semi-Markov Decision Processes

Karel Sladký[1]

Abstract. Semi-Markov decision processes can be considered as an extension of discrete- and continuous-time Markov reward models. Unfortunately, traditional optimality criteria as long-run average reward per time 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 average or total reward and its variance. Such approach has been already studied for very special classes of semi-Markov decision processes, in particular, for Markov decision processes in discrete- and continuous-time setting. In this note these approaches are summarized and possible extensions to the wider class of semi-Markov decision processes is discussed. Attention is mostly restricted to uncontrolled models in which the chain is aperiodic and contains a single class of recurrent states. In this case growth rate of total reward and the variance is again asymptotically linear in time and is independent of the starting state.

Keywords: Semi-Markov processes with rewards, discrete- and continuous-time Markov reward chains, average reward and variance over time, risk-sensitive optimality, policy iterations.

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

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 robustness of 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 stability and variability-risk features of the problem. Hence robustness and risk control are also important issues in practical applications. As well known one of the common and popular risk measure is the variance which is often used to characterize the stability of the system. Perhaps the best known approaches stem from the classical work of Markowitz (cf. [6]) on mean variance selection rules, i.e. we optimize the weighted sum of average or total reward and its variance. Higher moments and variance of cumulative rewards in Markov reward chains have been primarily studied for discrete time models. Research in this direction has been initiated in Benito [1], Jaquette [4], Mandl [5] and Sobel [15].

In the paper Van Dijk and Sladký [14] results for the discrete-time case are extended to continuous-time Markov reward chains. As the essential step is an expression for the variance of the undiscounted cumulative reward and its asymptotic behavior. In this note, for the sake of simplicity, the presentation is restricted to the *un*controlled case, the implication for the controlled case is only briefly be referred to. For additional results on the limiting average variance for continuous-time models let us mention the paper by Prieto-Rumeau and Hernández-Lerma [7] along with the monograph [2]. Similar results are also reported in Guo et al [3]. Since no transition rewards are considered, the obtained formula for the limiting variance is a special case of more complicated results reported in [14].

The article is structured as follows. Section 2 contains notations and summary of basic facts on discrete- and continuous-time Markov reward chains and their extensions to semi-Markov reward processes. The heart of the paper are sections 3 and 4. Second order optimality for Markov reward chains is discussed in Section 3, i.e. formulas for total expected reward and for the corresponding variances of total reward are derived for discrete- and continuous-time models. The analysis is limited to finite horizon case, however, the obtained results can be also used for long run discounted (or transient) models. Extensions of presented results to semi-Markov reward processes is contained in Section 4. Conclusions are made in Section 5.

[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

2 Notations and Preliminaries

Semi-Markov processes present an extension of Markov processes considered in discrete- and continuous-time setting. Considering Markov models with rewards we can summarize the following facts.

In the discrete-time case, we consider Markov decision chain $X^{\mathrm{d}} = \{X_n, n = 0, 1, \ldots\}$ 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.

In the continuous-time setting, the development of the considered Markov decision process $X^{\mathrm{c}} = \{X(t), t \geq 0\}$ (with finite state space $\mathcal{I}$) over time is governed by the transition rates $q(j|i, a)$, for $i, j \in \mathcal{I}$, depending on the selected action $a \in \mathcal{A}_i$. For $j \neq i$ $q(j|i, a)$ is the transition rate from state i into state j, $q(i|i, a) = \sum_{j\in\mathcal{I}, j\neq i} q(j|i, a)$ is the transition rate out of state i. Recall that on entering state i the process stays in state i for a random time that is exponentially distributed with parameter $q(i, a) = -q(i|i, a)$ and the next jump to state j occurs with probability $p_{ij}(a) = q(j|i, a)/q(i, a)$. As concerns reward rates, $r(i)$ denotes the rate earned in state $i \in \mathcal{I}$, and $r(i, j)$ is the transition rate accrued to a transition from state i to state j.

A (Markovian) policy controlling the decision process is given by either a sequence of decision at every time point (discrete-time case) or as a piecewise constant right continuous function of time (continuous-time case). In particular, for discrete-time models 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^{d} is in state i.

We denote by $P(f) = [p_{ij}(f_i)]$ the $N \times N$ transition matrix of the chain X^{d}. Obviously, the row sums along with the spectral radius of $P(f)$ are equal to one. 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 $\lim_{n\to\infty}[\tilde{P}(f)]^n = 0$. 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.

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 $p_{ij}(f_i)$. Recall that the limiting matrix $P^*(f) = \lim\limits_{m\to\infty} m^{-1} \sum_{n=0}^{m-1} P^n(f)$ exists; in case that the chain is aperiodic even $P^*(f) = \lim\limits_{n\to\infty} (P(f))^n$. In particular, if $P(f)$ is *unichain* (i.e. $P(f)$ contains a single class of recurrent states) the rows of $P^*(f)$, denoted $p^*(f_i)$, are identical. Obviously, $r_i(f_i) = \sum_{j=1}^{N} p_{ij}(f_i) r_{ij}$ is the expected one-stage reward obtained in state $i \in \mathcal{I}$ and $r(f)$ denotes the corresponding N-dimensional column vector of one-stage rewards. Then $v(f) := [P(f)]^n \cdot r(f)$ is the (column) vector of rewards accrued after n transitions, its ith entry $v_i(f)$ denotes expectation of the reward if the process X^{d} starts in state i.

Similarly, for the continuous-time case policy controlling the chain, $\pi = f(t)$, is a piecewise constant, right continuous vector function where $f(t) \in \mathcal{F} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$, and $f_i(t) \in \mathcal{A}_i$ is the decision (or action) taken at time t if the process $X(t)$ is in state i. Since π is piecewise constant, for each π we can identify the time points $0 < t_1 < t_2 \ldots < t_i < \ldots$ at which the policy switches; we denote by $f^i \in \mathcal{F}$ the decision rule taken in the time interval $(t_{i-1}, t_i]$. Policy which takes at all times the same decision rule, i.e. $\pi \sim (f)$, is called stationary.

Let for $f \in \mathcal{F}$ $Q(f) = [q_{ij}(f_i)]$ be an $N \times N$ matrix whose ijth element $q_{ij}(f_i) = q(j|i, f_i)$ for $i \neq j$ and for the iith element we set $q_{ii}(f_i) = -q(i|i, f_i)$ (recall that the row sums of a transition rate matrix $Q(f)$ are equal to null). The sojourn time of the considered process X^{c} in state $i \in \mathcal{I}$ is exponentially distributed with parameter $q(i|i, f_i)$. Hence the expected value of the reward obtained in state $i \in \mathcal{I}$ equals $r_i(f_i) = [q(i|i, f_i)]^{-1} r(i) + \sum_{j\in\mathcal{I}, j\neq i} q(j|i, f_i)\, r(i, j)$ and $r(f) = [r_i(f)]$ is the (column) vector of reward rates at time t. Recall that 0 is an eigenvalue of $Q(f)$, and the real part of any eigenvalue of $Q(f)$ is non-positive. Similarly to discrete-time model $\tilde{Q}(f)$ is transient if the real part of any eigenvalue of $Q(f)$ is negative.

The above two models can be unified and generalized by introducing semi-Markov reward processes. To this end, we shall define semi-Markov reward processes as follows.

Consider a controlled semi-Markov reward process $Y = \{Y(t), t \geq 0\}$ with finite state space $\mathcal{I} = \{1, 2, \ldots, N\}$ along with the embedded Markov chain $X^{\mathrm{d}} = \{X_n, n = 0, 1, \ldots\}$. We assume that X^{d} is unichain for any stationary policy. The development of the process $Y(t)$ over time is the following:

At time $t = 0$ if $Y(0) = i$ the decision maker selects decision from a finite set $\mathcal{A}_i = \{1, 2, \ldots, K_i\}$ of possible decisions (actions) in state $i \in \mathcal{I}$. Then state j is reached in the next transition with a given probability $p_{ij}(a)$ after random time $\eta_i(a)$. Let $F_i(a, \tau)$ be a non-lattice distribution function representing the conditional probability $\mathrm{P}(\eta_i \leq \tau)$. We assume that for $\ell = 1, 2$ and any $i, j = 1, \ldots, N$, $0 < d_i^{(\ell)}(a) = \int_0^\infty \tau^\ell \, \mathrm{d}F_i(a, \tau) < \infty$. Finally, one-stage transition reward $r(i, j) > 0$ will be accrued to such transition and reward rate $r(i)$ per unit of time incurred in state i is earned.

A (Markovian) policy controlling the semi-Markov process Y, $\pi = (f^0, f^1, \ldots)$, is identified by a sequence of decision vectors $\{f^n, n = 0, 1, \ldots\}$ where $f^n \in \mathcal{F} \equiv \mathcal{A}_1 \times \ldots \times \mathcal{A}_N$ for every $n = 0, 1, 2, \ldots$, and $f_i^n \in \mathcal{A}_i$ is the decision (or action) taken at the nth transition if the embedded Markov chain X^{d} is in state i. Let π^k be a sequence of decision vectors starting at the k-th transition, hence $\pi = (f^0, f^1, \ldots f^{k-1}, \pi^k)$.

Let ξ_n be the cumulative random reward obtained in the n first transitions of the considered embedded Markov chain X^{d} and $\xi(t)$ denotes the (random) reward earned up to time t, i.e.

$$\xi_n = \sum_{k=0}^{n-1}[r(X_k) \cdot \eta_{X_k, X_{k+1}} + r(X_k, X_{k+1})], \qquad \xi(t) := \left[\int_0^t r(Y(s))\mathrm{d}s \, + \sum_{k=0}^{N(t)} r(Y(\tau_k^-), Y(\tau_k^+))\right] \tag{1}$$

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

Obviously, discrete-time Markov reward chain is a very special case of semi-Markov reward process where all holding times are non-random and equal to one, and one-stage rewards depend only on the labels of the consecutive two states. Moreover, continuous-time Markov reward chains can be considered as a very specific case of semi-Markov reward processes where holding times are exponentially distributed. In particular, considering continuous-time Markov reward process with transition rates $q(j|i,a)$, reward rates $r(i)$ in state i and rewards per transition $r(i,j)$, the process can be treated as a semi-Markov process with transition probabilities $p_{ij}(a) = q(j|i,a)/q(i,a)$ and exponentially distributed holding times with parameter $q(i,a)$.

It is well-known that the long-run average reward of the considered semi-Markov process Y can be calculated using the embedded Markov chain X^{d}. In particular, if stationary policy $\pi \sim (f)$ is followed, on recalling that X^{d} is unichain the limiting matrix $P^*(f)$ has identical rows, i.e. $p_j^*(f)$ is the jth entry of each row of $P^*(f)$. Moreover, for the long run models also the fraction of time spent by the semi-Markov process Y in state i can be easily calculated (see e.g. [8, 9]). Then the average reward per unit of time, say $\bar{g}(f)$, generated by the semi-Markov process Y is independent of the starting state and can be calculated as

$$\bar{g}(f) = \sum_{j \in \mathcal{I}} \bar{p}_j^*(f) \cdot r_j(f), \text{ where } \bar{p}_i^*(f) = \frac{p_i^*(f) \cdot d_i(f)}{\sum_{j \in \mathcal{I}} p_{j\ell}^*(f) \cdot d_j(f)}, \; r_j(f) = d_j(f) \cdot r(j) + \sum_{\ell \in \mathcal{I}} p_{j\ell}(f_i) \cdot r(j,\ell). \tag{2}$$

Similarly it is possible to extend the presented discrete-time Markov reward chain model to a more general model of semi-Markov reward processes. To this end, let $\eta_i(a)$ be the random time spent in state i with expectation $d_i(a)$ if action a is chosen. Obviously, it suffices to add in each state $i \in \mathcal{I}$ to the one-stage rewards r_{ij} the following term $r(i) \cdot \eta_i(a)$ representing additional reward earned during random stay of the process X in state i. Hence the expected value and the second moment of the total reward earned in state i are $r(i) \cdot d_i^{(1)}(a) + \sum_{i \in \mathcal{I}} p_{ij}(a) \cdot r(i,j)$ and $\sum_{i \in \mathcal{I}} p_{ij}(a) \cdot \mathsf{E}\,[r(i) \cdot \eta_i(a) + r_{ij}]^2$ respectively.

3 Second Order Optimality in Markov Reward Chains

Considering discrete-time models, 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 $\xi_n(\pi)$

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

If policy $\pi \sim (f)$ is stationary, the process X^{d} 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

$$\mathsf{E}_i^\pi[\xi_n] \;=\; \mathsf{E}_i^\pi[\xi_m] + \mathsf{E}_i^\pi\Big\{\sum_{j \in \mathcal{I}} \mathsf{P}(X_m = j) \cdot \mathsf{E}_j^\pi[\xi_{n-m}]\Big\}. \tag{3}$$

$$\mathsf{E}_i^\pi[\xi_n]^2 \;=\; \mathsf{E}_i^\pi[\xi_m]^2 + \mathsf{E}_i^\pi\Big\{\sum_{j \in \mathcal{I}} \mathsf{P}(X_m = j) \cdot \mathsf{E}_j^\pi[\xi_{n-m}]^2\Big\} + 2 \cdot \mathsf{E}_i^\pi[\xi_m] \sum_{j \in \mathcal{I}} \mathsf{P}(X_m = j) \cdot \mathsf{E}_j^\pi[\xi_{n-m}]. \tag{4}$$

In particular, from (3), (4) we conclude for $m = 1$

$$v_i^{(1)}(f, n+1) \;=\; r_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot v_j^{(1)}(f,n) \tag{5}$$

$$v_i^{(2)}(f, n+1) \;=\; r_i^{(2)}(f_i) + 2\cdot\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot r_{ij}\cdot v_j^{(1)}(f,n) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, v_j^{(2)}(f,n) \tag{6}$$

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

Since the variance $\sigma_i^{(2)}(f,n) = v_i^{(2)}(f,n) - [v_i^{(1)}(f,n)]^2$ from (5),(6) we get

$$\begin{aligned}\sigma_i^{(2)}(f,n+1) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^{(2)}(f,n) + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot r_{ij}\cdot v_j^{(1)}(f,n)\\ &\quad -[v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot[v_j^{(1)}(f,n)]^2 \qquad (7)\\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot[r_{ij}+v_j^{(1)}(f,n)]^2 - [v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^{(2)}(f,n). \qquad (8)\end{aligned}$$

Using matrix notations (cf. [12, 13]) equations (5),(6),(20) can be written as:

$$v^{(1)}(f,n+1) \;=\; r^{(1)}(f) + P(f)\cdot v^{(1)}(f,n) \tag{9}$$

$$v^{(2)}(f,n+1) \;=\; r^{(2)}(f) + 2\cdot P(f)\circ R\cdot v^{(1)}(f,n) + P(f)\;\cdot v^{(2)}(f,n) \tag{10}$$

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

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 $r^{(1)}(f) = (P(f)\circ R)\cdot e,\;\; r^{(2)}(f) = [P(f)\circ(R\circ R)]\cdot e$ (e is reserved for unit column vector).

Similarly, considering Markov reward chains in continuous time the expected reward $v_i(t,\pi)$ can be considered as the first moment of the random variable $\xi(t)$ if the starting state $X(0) = i$ policy $\pi = f(t)$ is followed. Similarly, the corresponding second moment and variance are given by $v_i^{(2)}(t,\pi) := \mathsf{E}_i^{\pi}[\xi(t)]^2,\;\; \sigma_i^{(2)}(t,\pi) := v_i^{(2)}(t,\pi) - [v_i(t,\pi)]^2$.

Considering stationary policy $\pi\sim(f)$, let $\xi(t+\Delta) = \xi(\Delta) + \xi^{(\Delta,t+\Delta)}$ where $\xi^{(\Delta,t+\Delta)}$ is reserved for the total (random) reward obtained in the time interval $[\Delta, t+\Delta)$. Then $\xi(\Delta)+\xi^{(\Delta,t+\Delta)}$ and $[\xi(t+\Delta)]^2 = [\xi(\Delta)]^2 + [\xi^{(\Delta,t+\Delta)}]^2 + 2[\xi(\Delta)][\xi^{(\Delta,t+\Delta)}]$ and hence

$$\mathsf{E}_i^{\pi}[\xi(t+\Delta)] = \mathsf{E}_i^{\pi}[\xi(\Delta)] + \mathsf{E}_i^{\pi}[\xi^{(\Delta,t+\Delta)}] \tag{12}$$

$$\mathsf{E}_i^{\pi}[\xi(t+\Delta)]^2 = \mathsf{E}_i^{\pi}[\xi(\Delta)]^2 + \mathsf{E}_i^{\pi}[\xi^{(\Delta,t+\Delta)}]^2 + 2\cdot\mathsf{E}_i^{\pi}[\xi(\Delta)][\xi^{(\Delta,t+\Delta)}] \tag{13}$$

Then, by using that $P(\Delta,f) = I + \Delta Q(f) + o(\Delta^2)$ and that the probability for more than one transition in time Δ is of order Δ^2, for Δ tending to zero we obtain

$$\begin{aligned}\frac{\mathrm{d}v_i(t,f)}{\mathrm{d}t} &= r(i) + \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\cdot r(i,j) + \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\cdot[v_j(t,f)-v_i(t,f)]\\ &= r_i(f_i) + \sum_{j\in\mathcal{I}} q_{ij}(f_i)\cdot v_j(t,f) \qquad (14)\end{aligned}$$

$$\frac{\mathrm{d}v_i^{(2)}(t,f)}{\mathrm{d}t} = 2\cdot r(i)\cdot v_i(t,f) + \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\left\{[r(i,j)]^2 + 2\cdot r(i,j)\cdot v_j(t,f)\right\} + \sum_{j\in\mathcal{I}} q_{ij}(f_i)\cdot v_j^{(2)}(t,f) \tag{15}$$

By $\sigma_i^{(2)}(t,f) = v_i^{(2)}(t,f) - [v_i(t,f)]^2$ we thus obtain:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\sigma_i^{(2)}(t,f) &= \frac{\mathrm{d}}{\mathrm{d}t}v_i^{(2)}(t,f) - 2\cdot v_i(t,f)\frac{\mathrm{d}}{\mathrm{d}t}v_i(t,f)\\ &= 2\cdot r(i)\cdot v_i(t,f) + \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\left\{[r(i,j)]^2 + 2\cdot r(i,j)\cdot v_j(t,f)\right\} + \sum_{j\in\mathcal{I}} q_{ij}(f_i)\cdot v_j^{(2)}(t,f)\\ &\quad -2\cdot v_i(t,f)\cdot r(i)(f_i) + \sum_{j\in\mathcal{I}} q_{ij}(f_i)\cdot v_j(t,f) \end{aligned}\tag{16}$$

Using matrix notations equations (14),(15) can be written as:

$$\frac{\mathrm{d}}{\mathrm{d}t}v(t,f) = r(f) + Q(f)\cdot v(t,f), \qquad \frac{\mathrm{d}}{\mathrm{d}t}v^{(2)}(t,f) = r^{(2)}(t,f) + Q(f)\cdot v^{(2)}(t,f). \tag{17}$$

where $r(f) = [r_i(f)]$, $r^{(2)}(t,f) = [r_i^{(2)}(t,f)]$, $v(t,f) = [v_i(t,f)]$, $v^{(2)}(t,f) = [v_i^{(2)}(t,f)]$, are column vectors with elements $r_i(f_i) = r(i) + \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)r(i,j)$, $r_i^{(2)}(t,f) = 2r(i)v_i(t,f) + \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)\{[r(i,j)]^2 + 2r(i,j)v_j(t,f)\}$.

Similarly after some algebra (16) can be also written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma^{(2)}(t,f) = r^{(2\sigma)}(t,f) + Q(f)\sigma^{(2)}(t,f) \qquad \text{where}$$

$\sigma^{(2)}(t,f) = [\sigma_i^{(2)}(t,f)]$, $r^{(2\sigma)}(t,f) = [r_i^{(2\sigma)}(t,f)]$, $r_i^{(2\sigma)}(t,f) = \sum_{j\in\mathcal{I},j\neq i} q_{ij}(f_i)[r(i,j) + v_j(t,f) - v_i(t,f)]^2$

4 Mean Reward Variance and Semi-Markov Processes

In this section we extend reported results concerning Markov reward chains to semi-Markov processes. To this end, we focus attention on the embedded Markov chain X^{d} and extend the corresponding formulas presented in Section 3. In particular, we shall assume that if in state $i \in \mathcal{I}$ action a is selected and state j is then reached the one-step (random) reward earned in state i, say $\xi_{i,j}$, is equal to $r(i)\cdot\eta_i(a) + r(i,j)$ and taking the expectation we can conclude that $\mathsf{E}\,\xi_{i,j} = i\sum_{j\in\mathcal{I}} p_{ij}(a)\cdot r(i,j) + r(i)\cdot d_i^{(1)}(a)$.
Unfortunately, $\mathsf{E}\,[\xi_i]^2 = [r(i)]^2\cdot d_i^{(2)}(a) + \sum_{j\in\mathcal{I}} p_{ij}(a)\{2r(i)\cdot d_i^{(1)}(a) + [r(i,j)]^2\}$ and on using this way of reasoning formulas (3)–(6) can be replaced by

$$v_i^{(1)}(f,n+1) = r_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot v_j^{(1)}(f,n) \tag{18}$$

$$v_i^{(2)}(f,n+1) = r_i^{(2)}(f_i) + 2\cdot\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot[r(i,j) + r(i)\cdot d_i(f_i)\cdot v_j^{(1)}(f,n)] + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot v_j^{(2)}(f,n) \tag{19}$$

where $r_i^{(1)}(f_i) := r(i)\cdot d_i^{(1)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot r(i,j)$, $r_i^{(2)}(f_i) := [\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot[r(i,j)]^2$.

Since the variance $\sigma_i^{(2)}(f,n) = v_i^{(2)}(f,n) - [v_i^{(1)}(f,n)]^2$ from (5),(6) we get

$$\begin{aligned}\sigma_i^{(2)}(f,n+1) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^{(2)}(f,n) + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot r(i,j)\cdot v_j^{(1)}(f,n)\\ &\quad -[v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot[v_j^{(1)}(f,n)]^2 \end{aligned}\tag{20}$$

$$= \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot[r_{ij} + v_j^{(1)}(f,n)]^2 - [v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^{(2)}(f,n). \tag{21}$$

The long-run average variance (independent of the starting state) is

$$\bar{g}(f) = \sum_{j\in\mathcal{I}} \bar{p}_j^*(f)\cdot\bar{r}_j(f), \text{ where } \bar{p}_i^*(f) = \frac{p_i^*(f)\cdot d_i(f)}{\sum_{j\in\mathcal{I}} p_{j\ell}^*(f)\cdot d_j(f)},\ \bar{r}_j(f) = d_j(f)\cdot r(j) + \sum_{\ell\in\mathcal{I}} p_{j\ell}(f_i)\cdot r(j,\ell).$$

5 Conclusions

Solving problem on stochastic dynamic programming the decision maker selects by standard policy or value iteration methods the set of all optimal e.g. maximazing average reward. In the next step the decision maker selects in the class of optimal policies policies according to the to second order optimality criterion.

In this note formula for calculating long-run average variance of unichain semi-Markov reward processes is obtained. This also extend results concerning average variance for discrete- a continuous-time Markov reward chains. In particular, solving problems on stochastic dynamic programming at first the decision maker find by standard policy or value iteration methods the set of all optimal policies. In the next step the decision maker selects in the class of optimal policies policies according to the to second order optimality criterion.

Acknowledgement

This research was supported by the Czech Science Foundation under Grant 18-02739S.

References

[1] Benito, F. (1982). Calculating the variance in Markov processes with random reward. *Trabajos de Estadistica y de Investigacion Operativa* 33, 73–85.

[2] Guo, X. and Hernández-Lerma, O. (2009). *Continuous-Time Markov Decision Processes: Theory and Applications.* Berlin: Springer.

[3] Guo, X. and Song, X. (2009). Mean-variance criteria for finite continous-time Markov decision processes. *IEEE Trans. Automat. Control* 54, 2151–2157.

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

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

[6] Markowitz, H. (1952). Portfolio Selection. *Journal of Finance* 7, 77–92.

[7] Prieto-Rumeau, T. and Hernández-Lerma, O. (2009). Variance minimization and the overtaking optimality approach to continuous-time controlled Markov chains. *Math. Method Oper. Res.* 70, 527–540.

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

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

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

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

[12] Sladký, K. (2017). Second order optimality in Markov decision chains. *Kybernetika* 53, 1086–1099.

[13] Sladký, K. (2018). Risk-sensitive and mean variance optimality in continuous-time Markov decision chains. In: 36th Internat. Conf. Mathem. Methods in Economics (L.Váchová, V.Kratochvíl, eds.), Jindřichův Hradec 2018, pp. 497–512.

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

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