Second Order Optimality in Transient and Discounted Markov Decision Chains

Karel Sladký[1]

Abstract. The article is devoted to second order optimality in Markov decision processes. Attention is primarily focused on the reward variance for discounted models and undiscounted transient models (i.e. where the spectral radius of the transition probability matrix is less then unity). Considering the second order optimality criteria means that in the class of policies maximizing (or minimizing) total expected discounted reward (or undiscounted reward for the transient model) we choose the policy minimizing the total variance. Explicit formulae for calculating the variances for transient and discounted models are reported along with sketches of algorithmic procedures for finding second order optimal policies.

Keywords: dynamic programming, transient Markov reward chains, discounted Markov reward chains, reward-variance optimality.

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 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 transient and discounted 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 $f_i^n \in \mathcal{A}_i$ 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, $\mathsf{E}_i^\pi(X_m = j) = \sum_{i_j \in \mathcal{I}} p_{i,i_1}(f_i^0) \ldots p_{i_{m-1},j}(f_{m-1}^{m-1})$; $\mathrm{P}(X_m = j)$ 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 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 $r_i^{(1)}(f_i) := \sum_{j \in \mathcal{I}} p_{ij}(f_i) r_{ij}$ is the expected one-stage reward obtained in

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

state i. Similarly, $r^{(1)}(f)$ is an N-column vector of one-stage rewards whose i-the elements equals $r_i^{(1)}(f_i)$. Observe that $\mathsf{E}_i^\pi(X_m = j) = [P^m(f)]_{ij}$ ($[A]_{ij}$ denotes the ij-th element of the matrix A, $A \geq B$, resp. $A > B$ iff for each i,j $[A]_{ij} \geq [B]_{ij}$ resp. $[A]_{ij} \geq [B]_{ij}$ and $[A]_{ij} > [B]_{ij}$ for some i,j). 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 $P(f)$) $P^*(f) := \lim\limits_{n\to\infty} \frac{1}{n}\sum_{k=}^{n-1} P^k(f)$ (with elements $p^*_{ij}(f)$) exists, and if $P(f)$ is aperiodic then even $P^*(f) = \lim\limits_{k\to\infty} P^k(f)$ and the convergence is geometrical. The (row) vector $g^{(1)}(f) = P^*(f)\, r^{(1)}(f)$ is the 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]) 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.

3 Reward Variance for Finite and Infinite Time Horizon

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

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

resp.

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

It is well known from the literature (cf. e.g. [2],[3],[7]) that for the time horizon n tending to infinity policies maximizing or minimizing the values $v_i^{(1)}(\pi,n)$ and $v_i^{\alpha(1)}(\pi,n)$ can be found in the class of stationary policies, i.e. there exist $\hat{f}, \hat{f}^\alpha, \bar{f}, \bar{f}^\alpha \in \mathcal{F}$ such that for all $i \in \mathcal{I}$ and any policy $\pi = (f^n)$

$$v_i^{(1)}(\hat{f}) := \lim_{n\to\infty} v_i^{(1)}(\hat{f},n) \geq \limsup_{n\to\infty} v_i^{(1)}(\pi,n), \quad \lim_{n\to\infty} v_i^{(1)}(\bar{f},n) \leq \liminf_{n\to\infty} v_i^{(1)}(\pi,n), \tag{3}$$

$$v_i^{\alpha(1)}(\hat{f}^\alpha) := \lim_{n\to\infty} v_i^{\alpha(1)}(\hat{f}^\alpha,n) \geq \limsup_{n\to\infty} v_i^{\alpha(1)}(\pi,n), \quad \lim_{n\to\infty} v_i^{\alpha(1)}(\bar{f}^\alpha,n) \leq \liminf_{n\to\infty} v_i^{\alpha(1)}(\pi,n). \tag{4}$$

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 undiscounted 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

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

$$\mathsf{E}_i^\pi[\xi_n]^2 \;=\; \mathsf{E}_i^\pi[\xi_m]^2 + \mathsf{E}_i^\pi\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^\pi[\xi_{n-m}]^2\} + 2\cdot \mathsf{E}_i^\pi[\xi_m]\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^\pi[\xi_{n-m}]. \tag{6}$$

In particular, from (3), (5) and (6) we conclude for $m = 1$

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

$$v_i^{\alpha(2)}(f,n+1) \;=\; r_i^{(2)}(f_i) + 2\cdot\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\; r_{ij}\cdot v_j^{(1)}(f,n) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\, v_j^{(2)}(f,n) \tag{8}$$

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

Since the variance $\sigma_i(f,n) = v_i^{(2)}(f,n) - [v_i^{(1)}(f,n)]^2$ from (7),(8) we get

$$\begin{aligned}
\sigma_i(f,n+1) &= r_i^{(2)}(f_i) + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j(f,n) + 2\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\ r_{ij}\cdot v_j^{(1)}(f,n) \\
&\quad - [v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)[v_j^{(1)}(f,n)]^2 \qquad (9) \\
&= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} +\ v_j^{(1)}(f,n)]^2 - [v_i^{(1)}(f,n+1)]^2 + \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j(f,n). \qquad (10)
\end{aligned}$$

Using matrix notations equations (7),(8) can be written as:

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

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

where $R = [r_{ij}]_{i,j}$ is an $N\times N$-matrix, and
$r^{(2)}(f) = [\ r_i^{(2)}(f_i)]$, $\ v^{(2)}(f,n) = [v_i^{(2)}(f,n)]$, $v^{(1)}(f,n) = [(v_i^{(1)}(f,n)]$ 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 Transient Models

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 $\tilde{P}^*(f) = \lim_{n\to\infty}[\tilde{P}(f)]^n = 0$ and for the fundamental and deviation matrices we get $\tilde{Z}(f) = \tilde{H}(f) = [I - \tilde{P}(f)]^{-1}$.

Then on iterating (11) we easily conclude that there exists $v^{(1)}(f) := \lim\limits_{n\to\infty} v^{(1)}(f,n)$ such that

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

Similarly, from (12) (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 $v^{(2)}(f) = \lim\limits_{n\to\infty} v^{(2)}(f,n)$ such that

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

hence

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

On letting $n\to\infty$ from (9), (10) we get for $\sigma_i(f) := \lim\limits_{n\to\infty}\sigma_i(f,n)$

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

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

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

In particular, if $\tilde{P}(f) := \alpha P(f)$ then (19) reads

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

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

3.3 Discounted Case

From (4) similarly to subsection 3.1 for $n > m$ we can conclude that

$$\begin{aligned} \mathsf{E}_i^\pi[\xi_n^\alpha] &= \mathsf{E}_i^\pi[\xi_m^\alpha] + \alpha^m \mathsf{E}_i^\pi\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^\pi[\xi_{n-m}^\alpha]\}. && (21)\\ \mathsf{E}_i^\pi[\xi_n^\alpha]^2 &= \mathsf{E}_i^\pi[\xi_m^\alpha]^2 + \alpha^{2m} \mathsf{E}_i^\pi\{\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^\pi[\xi_{n-m}^\alpha]^2\} \\ &\quad +2\cdot\alpha^m\cdot \mathsf{E}_i^\pi[\xi_m^\alpha]\sum_{j\in\mathcal{I}} \mathrm{P}(X_m = j)\cdot \mathsf{E}_j^\pi[\xi_{n-m}^\alpha]. && (22)\end{aligned}$$

In particular, from (2), (21) and (22) we conclude for $m = 1$

$$\begin{aligned} v_i^{\alpha(1)}(f,n+1) &= r_i^{(1)}(f_i) + \alpha \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot v_j^{\alpha(1)}(f,n) && (23)\\ v_i^{\alpha(2)}(f,n+1) &= r_i^{(2)}(f_i) + 2\cdot\alpha\cdot \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\ r_{ij}\cdot v_j^{\alpha(1)}(f,n) + \alpha^2\cdot\sum_{j\in\mathcal{I}} p_{ij}(f_i)\, v_j^{\alpha(2)}(f,n) && (24)\end{aligned}$$

and from (23),(24), for the variance $\sigma_i^\alpha(f,n) := v_i^{\alpha(2)}(f,n) - [v_i^{\alpha(1)}(f,n)]^2$ we get

$$\begin{aligned} \sigma_i^\alpha(f,n+1) &= r_i^{(2)}(f_i) + \alpha^2 \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^\alpha(f,n) + 2\cdot\alpha\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\ r_{ij}\cdot v_j^{\alpha(1)}(f,n) \\ &\quad -[v_i^{\alpha(1)}(f,n+1)]^2 + \alpha^2\sum_{j\in\mathcal{I}} p_{ij}(f_i)[v_j^{\alpha(1)}(f,n)]^2 && (25)\\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + \alpha\cdot v_j^{\alpha(1)}(f,n)]^2 - [v_i^{\alpha(1)}(f,n+1)]^2 + \alpha^2 \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^\alpha(f,n). && (26)\end{aligned}$$

Using matrix notations equations (23), (24) can be written as:

$$\begin{aligned} v^{\alpha(1)}(f,n+1) &= r^{(1)}(f) + \alpha\cdot P(f)\cdot v^{\alpha(1)}(f,n) && (27)\\ v^{\alpha(2)}(f,n+1) &= r^{(2)}(f) + 2\cdot\alpha\cdot P(f)\circ R\cdot v^{\alpha(1)}(f,n) + \alpha^2\cdot P(f)\, v^{\alpha(2)}(f,n) && (28)\end{aligned}$$

recall that $R = [r_{ij}]_{i,j}$ is an $N \times N$-matrix, and $\circ$ is used for Hadamard (entrywise) product of matrices.

On iterating (27) we easily conclude that there exists $v^{\alpha(1)}(f) := \lim_{n\to\infty} v^{\alpha(1)}(f,n)$ such that

$$v^{\alpha(1)}(f) = r^{(1)}(f) + \alpha P(f)\cdot v^{\alpha(1)}(f) \Longleftrightarrow v^{\alpha(1)}(f) = [I - \alpha P(f)]^{-1} r^{(1)}(f). \qquad (29)$$

Similarly to the transient case on letting $n \to \infty$ for discounted models also exists $v^{\alpha(2)}(f) = \lim_{n\to\infty} v^{\alpha(2)}(f,n)$ and by (28)

$$v^{\alpha(2)}(f) = r^{(2)}(f) + 2\cdot\alpha\cdot P(f)\circ R\cdot v^{\alpha(1)}(f) + \alpha^2\cdot P(f)\, v^{\alpha(2)}(f), \qquad (30)$$

so

$$v^{\alpha(2)}(f) = [I - \alpha^2\cdot P(f)]^{-1}\ \Big\{\ r^{(2)}(f) + 2\cdot\alpha\cdot P(f)\circ R\cdot v^{\alpha(1)}(f)\Big\}. \qquad (31)$$

On letting $n \to \infty$ from (25), (26) we get for $\sigma_i^\alpha(f) := \lim_{n\to\infty} \sigma_i^\alpha(f,n)$

$$\begin{aligned} \sigma_i^\alpha(f) &= r_i^{(2)}(f_i) + \alpha^2 \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^\alpha(f) + 2\cdot\alpha\sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\ r_{ij}\cdot v_j^{\alpha(1)}(f) \\ &\quad -[v_i^{\alpha(1)}(f)]^2 + \alpha^2\sum_{j\in\mathcal{I}} p_{ij}(f_i)[v_j^{\alpha(1)}(f)]^2 && (32)\\ &= \sum_{j\in\mathcal{I}} p_{ij}(f_i)[r_{ij} + \alpha\cdot v_j^{\alpha(1)}(f)]^2 - [v_i^{\alpha(1)}(f)]^2 + \alpha^2 \sum_{j\in\mathcal{I}} p_{ij}(f_i)\cdot\sigma_j^\alpha(f). && (33)\end{aligned}$$

Hence in matrix notation

$$\sigma^\alpha(f) = r^{(2)}(f) + \alpha^2\cdot P(f)\cdot\sigma^\alpha(f) + 2\cdot\alpha\cdot P(f)\circ R\cdot v^{\alpha(1)}(f) - [v^{\alpha(1)}(f)]^2 + \alpha^2\cdot P(f)\cdot[v^{\alpha(1)}(f)]^2. \qquad (34)$$

After some algebra (34) can be also written as

$$\sigma^{\alpha}(f) \quad = \quad [I - \alpha^2 \cdot P(f)]^{-1} \cdot \{\, r^{(2)}(f) + 2 \cdot \alpha \cdot P(f) \circ R \cdot v^{\alpha(1)}(f)\} - [v^{\alpha(1)}(f)]^2. \qquad (35)$$

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

(35) is similar to the formula for the variance of discounted rewards obtained by Sobel [6] by different methods (see also [5]).

3.4 Comparison of Transient and Discounted Models

Let us consider transient model where the transient transition probability matrix $\tilde{P}(f) := \alpha P(f)$, called the α-transient model. Then by (20) the variance

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

On the other hand for the α-discounted model we get by (35)

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

Since $[I - \alpha P(f)]^{-1} > [I - \alpha^2 P(f)]^{-1}$ from (15),(31) and from (19),(20) we immediately conclude that

$$v^{\alpha(2)}(f) - \tilde{v}^{\alpha(2)}(f) \;=\; \{[I - \alpha P(f)]^{-1} - [I - \alpha^2 P(f)]^{-1}\} \cdot \{\, r^{(2)}(f) + 2\alpha P(f) \circ R \cdot \tilde{v}^{\alpha(1)}(f)\} > 0, \quad (36)$$

$$\tilde{\sigma}^{\alpha}(f) - \sigma^{\alpha}(f) \;=\; \{[I - \alpha P(f)]^{-1} - [I - \alpha^2 P(f)]^{-1}\} \cdot \{\, r^{(2)}(f) + 2\alpha P(f) \circ R \cdot \tilde{v}^{\alpha 1)}(f) > 0. \quad (37)$$

Observe that

$$[I - \alpha P(f)]^{-1} - [I - \alpha^2 P(f)]^{-1} = \sum_{k=0}^{\infty} \{(\alpha P(f))^k - (\alpha^2 P(f))^k\} = \sum_{k=0}^{\infty} \{(1 - \alpha^k)\alpha^k (P(f))^k\} > 0.$$

In words:

Total reward of α-transient Markov reward chain is greater the that of α-discounted Markov reward chain (this fact was also mentioned in [1] and in [4] using approached different to ours).

4 Finding Optimal Policies

For finding second order optimal policies, at first it is necessary to construct the set of optimal transient or optimal α-discounted policies. Recalling (3),(4) for the both optimality criteria mentioned above optimal policies can be found in the class of stationary policies, i.e. there exists $\hat{f}, \hat{f}^{\alpha} \in \mathcal{F}$ such that

$$v^{(1)}(\hat{f}) \ge v^{(1)}(\pi) \quad \text{resp.} \quad v^{\alpha(1)}(\hat{f}) \ge v^{\alpha(1)}(\pi) \quad \text{for every policy} \quad \pi = (f^n). \qquad (38)$$

Let $\mathcal{F}^{(0)} \subset \mathcal{F}$ be the set of all transient optimal stationary policies, $\mathcal{F}^{\alpha(0)} \subset \mathcal{F}$ be the set of all α-discounted optimal stationary policies.

Obviously, following policy $\hat{f} \in \mathcal{F}^{(0)}$, resp. $\hat{f} \in \mathcal{F}^{\alpha(0)}$, the corresponding action $\hat{f}_i \in \mathcal{A}_i^{(0)} \subset \mathcal{A}_i$, resp. $\hat{f}_i^{\alpha} \in \mathcal{A}_i^{\alpha(0)} \subset \mathcal{A}_i$. For finding the sets $\mathcal{F}^{(0)}$, $\mathcal{F}^{\alpha(0)}$ of optimal policies well-known policy iteration algorithms, value iteration algorithms or modified value iteration algorithms can be used.

Assume that for $\hat{f}^{(1)}, \hat{f}^{(2)} \in \mathcal{F}^{(0)}$ resp. $\hat{f}^{\alpha(1)}, \hat{f}^{\alpha(2)} \in \mathcal{F}^{\alpha(0)}$

$$\hat{v}^{(1)} := v^{(1)}(\hat{f}^{(1)}) = v^{(1)}(\hat{f}^{(2)}) \;\; \text{resp.} \;\; \hat{v}^{\alpha(1)} := v^{\alpha(1)}(\hat{f}^{\alpha(1)}) = v^{\alpha(1)}(\hat{f}^{\alpha(2)}),$$

however for the corresponding variances $\sigma_i(\hat{f}^{(1)}) \ne \sigma_i(\hat{f}^{(2)}),\;\; \sigma_i(\hat{f}^{\alpha}) \ne \sigma_i(\hat{f}^{\alpha})$.

In what follows we show existence of $f^* \in \mathcal{F}^{(0)}$, $f^{\alpha *} \in \mathcal{F}^{\alpha(0)}$, such that

$$\sigma(f^*) \le \sigma(\pi) \quad \text{resp.} \quad \sigma^{\alpha}(f^*\alpha) \le \sigma^{\alpha}(\pi) \qquad (39)$$

for every policy $\pi = (f^n), f^n \in \mathcal{F}^{(0)}$ resp. $\pi = (f^n), f^n \in \mathcal{F}^{\alpha(0)}$.

To this end from (18), resp. from (34), we have for transient, resp. discounted, case

$$\sigma(f) = \tilde{P}(f) \circ R \circ R \cdot e + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f) - [v^{(1)}(f)]^2 + \tilde{P}(f) \cdot [v^{(1)}(f)]^2 + \tilde{P}(f) \cdot \sigma(f), \tag{40}$$

resp. for every $f \in \mathcal{F}^{\alpha(0)}$

$$\sigma^\alpha(f) = P(f) \circ R \circ R \cdot e + 2 \cdot \alpha \cdot P(f) \circ R \cdot v^{\alpha(1)}(f) - [v^{\alpha(1)}(f)]^2 + \alpha^2 \cdot P(f) \cdot [v^{\alpha(1)}(f)]^2 + \alpha^2 \cdot P(f) \cdot \sigma^\alpha(f). \tag{41}$$

Now let $h(f), h^\alpha(f)$ equal the first three terms on the RHS of (40) and (41), i.e.

$$\begin{aligned} h(f) &:= \tilde{P}(f) \circ R \circ R \cdot e + 2 \cdot \tilde{P}(f) \circ R \cdot v^{(1)}(f) - [v^{(1)}(f)]^2 + \tilde{P}(f) \cdot [v^{(1)}(f)]^2, \\ h^\alpha(f) &:= P(f) \circ R \circ R \cdot e + 2 \cdot \alpha \cdot P(f) \circ R \cdot v^{\alpha(1)}(f) - [v^{\alpha(1)}(f)]^2 + \alpha^2 \cdot P(f) \cdot [v^{\alpha(1)}(f)]^2. \end{aligned}$$

Observe that $h(f), h^\alpha(f)$ well correspond to one-stage rewards in (13) and in (29).

On comparing (40) and (13), resp. (41) and (29), we can conclude that the minimal values $\sigma(f^*)$, resp. $\sigma^\alpha(f^*)$ with $f \in \mathcal{F}^{(0)}$, resp. with $f \in \mathcal{F}^{\alpha(0)}$, exist.

5 Conclusions

We have received formulas for the variance of total rewards for transient and discounted Markov reward chains. This enables in the class of optimal policies select policies minimizing variance of total expected reward.

Acknowledgements

This research was supported by the Czech Science Foundation under Grants 13-14445S, 15-10331S and by CONACyT (Mexico) and AS CR (Czech Republic) under Project 171396.

References

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

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

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

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

[5] Sladký, K.: *The variance of discounted rewards in Markov decision processes: Laurent expansion and sensitive optimality.* In: Proceedings of the 32th International Conference Mathematical Methods in Economics 2014, Palacký University, Olomouc, pp. 908–913.

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

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