J. Appl. Prob. **19**, 794–802 (1982)
Printed in Israel
0021–9002/82/040794–694–09 $01.15

THE VARIANCE OF DISCOUNTED MARKOV DECISION PROCESSES

MATTHEW J. SOBEL,* *Georgia Institute of Technology*

Abstract

Formulae are presented for the variance and higher moments of the present value of single-stage rewards in a finite Markov decision process. Similar formulae are exhibited for a semi-Markov decision process. There is a short discussion of the obstacles to using the variance formula in algorithms to maximize the mean minus a multiple of the standard deviation.

MARKOV DECISION PROCESS; VARIANCE; DISCOUNTED RETURN; POLICY IMPROVEMENT

1. Introduction and notation

The usual optimization criterion for a discounted Markov decision process (MDP) is to maximize the expected value of the sum of discounted rewards. In several kinds of applications (cf. Mendelssohn (1980)) practitioners are concerned with the variance of the sum as well as with its expected value. A formula for the variance is presented in Section 2. An analogous formula for the semi-Markov decision process is presented in Section 3.

Many authors have written about alternative approaches to making decisions under uncertainty. Besides the MDP literature and its direct forebears, major research efforts include the approaches to sequential decisions examined in the literatures on stochastic programs with recourse and chance-constrained programs. See Stancu-Minasian and Wets (1976) for an exhaustive bibliographic guide to this research. In recent years, several authors have written about choice over time from the point of view of utility theory; see Ferejohn and Page (1978) and Kreps and Porteus (1978) and its references. In the MDP literature, Derman (1970), Kushner (1971), and Mine and Osaki (1970) explain how probabilistic constraints may be incorporated into linear programs for MDP. White (1974) discusses the use of Lagrange multipliers for the inclusion of probabilistic constraints or variances in the optimization of MDP. Mandl (1971) and Jaquette

Received 7 April 1981; revision received 12 November 1981.

* Postal address: College of Management, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

(1973) propose minimization of variance in order to resolve ties amongst policies which maximize mean values.

The following Markov decision process is a standard model. See Denardo (1971) for its genealogy.

Let $\mathcal{S}$ be the state space, A_s be the set of actions available in state s, and $\mathcal{C} = \{(s, a) : a \in A_s,\ s \in \mathcal{S}\}$ which is assumed to contain only finitely many elements. Let s_n and a_n indicate the state and action in the nth period. The reward in the nth period is specified by $r(s_n, a_n, s_{n+1})$. Let

$$p_{sj}^a = P\{s_{n+1} = j \mid s_n = s,\ a_n = a\}$$

and let $0 < \beta < 1$ be the single-period discount factor. Since $\mathcal{C}$ is a finite set, there is no loss of generality in the assumption, made here, that $0 \leqq r(s, a, j)$ for all $(s, a) \in \mathcal{C}$ and $j \in \mathcal{S}$.

The present value of the single-period rewards is

$$B = \sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n, s_{n+1}).$$

Suppose that the stationary policy δ is used to choose actions, i.e. $a_n = \delta(s_n)$ for all n. Let B_s denote the random variable B if $s_1 = s$ and $a_n = \delta(s_n)$ for all n.

Let

$$F_s(x) = P\{B_s \leqq x\}$$

$$v_s^{(m)} = \int_0^{\infty} x^m dF_s(x), \qquad v_s = v_s^{(1)},$$

and

$$V_s = v_s^{(2)} - v_s^2$$

be the distribution function, mth moment, and variance of B_s.

2. Formulae

Kemeny and Snell (1960) have a formula for the second moment of first-passage times in Markov chains. Platzman (1978) presents formulae for the second moment of the total reward accumulated in transient states of a Markov chain with rewards. Formula (4) with (2), below, is very similar to those of Kemeny and Snell and Platzman and could be derived in the same way. Instead, the proof uses the lemma below which has independent interest. The lemma is closely related to Theorem 1 in Mandl (1971).

Let $r_{sj} = r[s, \delta(s), j]$ and $p_{sj} = p_{sj}^{\delta(s)}$.

Lemma.

$$F_s(x) = \sum_{j \in \mathcal{S}} p_{sj} F_j\left[\frac{x - r_{sj}}{\beta}\right]. \tag{1}$$

Proof.

$$F_s(x) = P\left\{\sum_{n=1}^{\infty} \beta^{n-1} r_{s_n s_{n+1}} \leqq x \mid s_1 = s\right\}$$

$$= \sum_{j \in \mathcal{S}} p_{sj} P\left\{r_{sj} + \beta \sum_{n=1}^{\infty} \beta^{n-1} r_{s_{n+1} s_{n+2}} \leqq x \mid s_1 = s, s_2 = j\right\}$$

$$= \sum_{j \in \mathcal{S}} p_{sj} P\left\{\sum_{n=1}^{\infty} \beta^{n-1} r_{s_{n+1} s_{n+2}} \leqq (x - r_{sj})/\beta \mid s_1 = s, s_2 = j\right\}$$

$$= \sum_{j \in \mathcal{S}} p_{sj} F_j[(x - r_{sj})/\beta].$$

Let $\boldsymbol{\theta}$ denote the vector whose sth component is

$$\theta_s = \sum_{j \in \mathcal{S}} p_{sj} (r_{sj} + \beta v_j)^2 - v_s^2. \tag{2}$$

Let P denote the matrix whose (s, j)th component is p_{sj} and let $\boldsymbol{r}$ denote the vector whose sth component is

$$r_s = \sum_{j \in \mathcal{S}} p_{sj} r_{sj}.$$

Let $\boldsymbol{v}$ and $\boldsymbol{V}$ denote the vectors whose sth components are v_s and V_s, respectively.

Theorem 1.

$$\boldsymbol{v} = \boldsymbol{r} + \beta P \boldsymbol{v} = (I - \beta P)^{-1} \boldsymbol{r}, \tag{3}$$

$$\boldsymbol{V} = \boldsymbol{\theta} + \beta^2 P \boldsymbol{V} = (I - \beta^2 P)^{-1} \boldsymbol{\theta}, \tag{4}$$

and

$$v_s^{(m)} = \sum_{i=0}^{m-1} \binom{m}{i} \beta^i \sum_{j \in \mathcal{S}} p_{sj} r_{sj}^{m-i} v_j^{(i)} + \beta^m \sum_{j \in \mathcal{S}} p_{sj} v_j^{(m)}. \tag{5}$$

Comments. (i) Formula (3) is well known and is included as a consistency check on the method of proof. (ii) The system (4) involves inversion of two matrices, namely of $I - \beta P$ to compute $\boldsymbol{v}$ and then of $I - \beta^2 P$. If for all $(s, a) \in \mathscr{C}$ $r(s, a, j)$ does not depend on j, let $\rho_s \equiv r[s, \delta(s), \cdot]$. Then (2) and (5) become

$$\theta_s = \beta^2 \sum_{j \in \mathcal{S}} p_{sj} v_j^2 - (v_s - \rho_s)^2 = \beta^2 \left[\sum_{j \in \mathcal{S}} p_{sj} v_j^2 - \left(\sum_{j \in \mathcal{S}} p_{sj} v_j\right)^2\right]. \tag{6}$$

and

$$v_s^{(m)} = \sum_{i=0}^{m-1} \binom{m}{i} \beta^i r_s^{m-i} \sum_{j \in \mathcal{S}} p_{sj} v_j^{(i)} + \beta^m \sum_{j \in \mathcal{S}} p_{sj} v_j^{(m)}.$$

(iii) The conditional variance formula

$$\mathrm{Var}(X) = \mathrm{Var}[E(X \mid Y)] + E[\mathrm{Var}(X \mid Y)]$$

can be applied with $X = B_s$ and $Y = s_2$. Then $\theta_s = \mathrm{Var}[E(B_s \mid s_2)]$ and

$$\beta^2 \sum_{j \in \mathcal{S}} p_{sj} V_j = E[\mathrm{Var}(B_s \mid s_2)].$$

Dr Daniel P. Heyman observes that a proof of (4) can be constructed with the conditional variance formula.

Proof. For (3), the lemma yields

$$v_s = \int_0^\infty x dF_s(x) = \sum_{j \in \mathcal{S}} p_{sj} \int_0^\infty dF_j[(x - r_{sj})/\beta]$$

$$= \sum_{j \in \mathcal{S}} p_{sj} \int_0^\infty (r_{sj} + \beta u) dF_j(u)$$

$$= \sum_{j \in \mathcal{S}} p_{sj}(r_{sj} + \beta v_j) = r_s + \beta \sum_{j \in \mathcal{S}} p_{sj} v_j.$$

The non-singularity of $I - \beta P$ due to $0 < \beta < 1$ is well known.

For (4), observe that B_s has the same distribution as $r_{ss_2} + \beta B_{s_2}$. Therefore,

$$V_s = E[(r_{ss_2} + \beta B_{s_2})^2] - v_s^2$$

$$= \sum_{j \in \mathcal{S}} p_{sj}[r_{sj}^2 + \beta^2 E(B_j^2) + 2\beta r_{sj} v_j] - v_s^2$$

$$= \sum_{j \in \mathcal{S}} p_{sj}[r_{sj}^2 + \beta^2 (V_j + v_j^2) + 2\beta r_{sj} v_j] - v_s^2$$

$$= \beta^2 \sum_{j \in \mathcal{S}} p_{sj} V_j + \theta_s$$

so $\boldsymbol{V} = \boldsymbol{\theta} + \beta^2 P \boldsymbol{V}$. Again, $I - \beta^2 P$ is non-singular because $0 < \beta < 1$ so $\boldsymbol{V} = (I - \beta^2 P)^{-1} \boldsymbol{\theta}$.

From the lemma and the binomial theorem,

$$v_s^{(m)} = \int_0^\infty x^m dF_s(x) = \sum_{j \in \mathcal{S}} p_{sj} \int_0^\infty x^m dF_j[(x - r_{sj})/\beta]$$

$$= \sum_{j \in \mathcal{S}} p_{sj} \int_0^\infty (r_{sj} + \beta u)^m dF_j(u)$$

$$= \sum_{j \in \mathcal{S}} p_{sj} \sum_{i=0}^{m} \binom{m}{i} r_{sj}^{m-i} \beta^i v_j^{(i)}$$

$$= \sum_{i=0}^{m} \binom{m}{i} \beta^i \sum_{j \in \mathcal{S}} p_{sj} r_{sj}^{m-i} v_j^{(i)}.$$

Finite-horizon versions of the lemma and Theorem 1 can be obtained with minor changes in the proofs. We skip the details and present only the formulae.

Let $L(\cdot)$ be a real-valued salvage value function on $\mathscr{S}$, $(\cdots, \delta_3, \delta_2, \delta_1)$ a sequence of single-stage decision rules (i.e. $\delta_n(s) \in A_s$ for all $s \in \mathscr{S}$ and $n = 1, 2, \cdots$), and

$$B^N = \sum_{n=1}^{N} \beta^{n-1} r[s_n, \delta_n(s_n), s_{n+1}] + \beta^N L(s_{N+1})$$

for N a positive integer. Let $p_{sj}(n)$ and $r_{sj}(n)$ denote $p_{sj}^{\delta_n(s)}$ and $r[s, \delta_n(s), j]$, respectively. Let $F_s^N(\cdot)$, v_{sN}, $v_{sN}^{(2)}$, and V_{sN} denote the distribution function, mean, mean square, and variance of B^N, respectively, when $s_1 = s$. Here are the formulae analogous to (1), (3), and (4):

$$F_s^N(x) = \sum_{j \in \mathscr{S}} p_{sj}(N) F_j^{N-1}\{[x - r_{sj}(N)]/\beta\},$$

$$\boldsymbol{v}_N = \boldsymbol{r}_N + \beta P_N \boldsymbol{v}_{N-1},$$

$$V_{sN} = \beta^2 \sum_{j \in \mathscr{S}} p_{sj}(N) V_{j,N-1} + \sum_{j \in \mathscr{S}} p_{sj}(N)[r_{sj}(N) + \beta v_{j,N-1}]^2 - (v_{sN})^2, \quad s \in \mathscr{S},$$

where $\boldsymbol{v}_N$ and $\boldsymbol{r}_N$ are the vectors with components v_{sN} and $\Sigma_{j \in \mathscr{S}} p_{sj}(N) r_{sj}(N)$, respectively, $v_{s0} = L(s)$, $V_{s0} = 0$, and P_N is the matrix with elements $p_{sj}(N)$.

3. Semi-Markov decision process

This section states analogues of (1), (3), (4), and (5) for a discounted finite semi-Markov decision process. The proofs are omitted because they are merely cumbersome replicas of those in Section 2.

See Denardo (1971) for the details and genealogy of the following model. There is a sequence $(s_1, t_1), (s_2, t_2), \cdots$ of pairs of random variables with s_n the nth observed physical state and t_n the time at which s_n is first observed. As with the MDP, for each n, $s_n \in \mathscr{S}$ and an action a_n is taken while the observed physical state is s_n. The constraint is $a_n \in A_{s_n}$ for each n. Let

$$\mathscr{C} = \{(s, a) : a \in A_s, s \in \mathscr{S}\}$$

which is assumed to be a finite set.

A stationary policy δ satisfies $\delta(s) \in A_s$ for all $s \in \mathscr{S}$, and $a_n = \delta(s_n)$ for all n. For $(s, a) \in \mathscr{C}$, $j \in \mathscr{S}$, and $x \geqq 0$, let

$$Q_{sj}^a(x) = P\{s_{n+1} = j, t_{n+1} \leqq t + x \mid s_n = s, t_n = t, a_n = a\}.$$

It is assumed that a stationary policy δ is in use such that $E(t_n \mid s_1 = s) > 0$ for all n and s, and

$$\lim_{t \to \infty} \sum_{j \in \mathscr{S}} Q_{sj}^{\delta(s)}(t) = 1, \qquad s \in \mathscr{S}.$$

The notation below uses $Q_{sj}(t)$ in place of $Q_{sj}^{\delta(s)}(t)$.

There is an instantaneous discount factor $\gamma > 0$. Let $r_{sj}(t)$ be the total income earned during $[0, t]$ given that $t \leqq t_1$, $s_1 = s$, $s_2 = j$, and $a_1 = \delta(s_1)$. For each s and j, it is assumed that $r_{sj}(\cdot)$ has bounded variation on $[0, \infty)$. Let

$$R_{sj}(t) = \int_0^t e^{-\gamma\tau} dr_{sj}(\tau), \qquad r_s = \sum_{j \in \mathcal{S}} \int_0^\infty R_{sj}(t) dQ_{sj}(t),$$

$$p_{sj} = \int_0^\infty e^{-\gamma t} dQ_{sj}(t), \qquad q_{sj} = \int_0^\infty e^{-2\gamma t} dQ_{sj}(t),$$

$$r_s^{(2)} = \sum_{j \in \mathcal{S}} \int_0^\infty [R_{sj}(t)]^2 dQ_{sj}(t) \quad \text{and} \quad \rho_{sj} = \int_0^\infty R_{sj}(t) e^{-\gamma t} dQ_{sj}(t).$$

Let $B_s(t)$ denote the total income earned during $[0, t]$ if $s_1 = s$ and $a_n = \delta(s_n)$ for all n. The present value of the income and its distribution function are

$$B_s = \int_0^\infty e^{-\gamma t} dB_s(t) \quad \text{and} \quad F_s(x) = P\{B_s \leqq x\}.$$

The mth moment, first moment, and variance of B_s are

$$v_s^{(m)} = E(B_s^m), \quad v_s = v_s^{(1)}, \quad \text{and} \quad V_s = v_s^{(2)} - v_s^2.$$

Let

$$\theta_s = r_s^{(2)} + \sum_{j \in \mathcal{S}} q_{sj} v_j^2 + 2 \sum_{j \in \mathcal{S}} \rho_{sj} v_j - v_s^2.$$

With this notation, the generalizations of (1), (3), (4), and (5) are listed below:

$$F_s(x) = \sum_{j \in \mathcal{S}} \int_0^\infty F_j\{[x - R_{sj}(t)]/e^{-\gamma t}\} dQ_{sj}(t),$$

$$v_s = r_s + \sum_{j \in \mathcal{S}} p_{sj} v_j,$$

$$V_s = \theta_s + \sum_{j \in \mathcal{S}} q_{sj} V_j,$$

and

$$v_s^{(m)} = \sum_{i=0}^{m-1} \binom{m}{i} \sum_{j \in \mathcal{S}} v_j^{(i)} \int_0^\infty e^{-i\gamma t} [R_{sj}(t)]^{m-i} dQ_{sj}(t) + \sum_{j \in \mathcal{S}} v_j^{(m)} \int_0^\infty e^{-m\gamma t} dQ_{sj}(t).$$

4. Mean-variance tradeoff

It is natural to attempt to use formula (4) in order to optimize MDP with a criterion which involves the variance of B_s. This section describes an obstacle to

such attempts.† For example, the criterion $E(B_s) - \lambda\sqrt{\text{variance of } B_s}$ $(\lambda > 0)$ may not be amenable to optimization with a policy iteration algorithm.

The following notation and terminology occur often in the literature on MDP. Let $\pi = (\delta_1, \delta_2, \cdots)$ be a policy, namely a sequence of single-stage decision rules. For a single-stage decision rule δ, let δ^∞ denote the policy $(\delta, \delta_1, \delta_2, \cdots)$ in which δ delays the use of π for one period. Let $V_s(\pi)$ and $v_s(\pi)$ denote the variance and expected present value, respectively, of B_s induced by π.

Suppose π and π' are two policies such that

$$v_j(\pi) \geqq v_j(\pi') \qquad \text{for all } j \in \mathcal{S}. \tag{7a}$$

It is well known that (7a) implies

$$v_s(\delta, \pi) \geqq v_s(\delta, \pi') \qquad \text{for all } s \in \mathcal{S} \tag{7b}$$

for all single-period decision rules δ. This property, called ***consistent choice***, ***temporal persistence***, ***stationarity***, and ***monotonicity*** by various authors, has been exploited to prove the existence of an optimal stationary policy (Denardo (1967)) and the convergence of a policy improvement algorithm to an optimum (Sobel (1975)). Unfortunately, the variance lacks this property as the following simple example demonstrates.

The analogue of (7a, b) for the variance is

$$V_j(\pi) \geqq V_j(\pi') \text{ for all } j \in \mathcal{S} \Rightarrow V_s(\delta, \pi) \geqq V_s(\delta, \pi') \text{ for all } s \in \mathcal{S} \tag{8}$$

for all single-period decision rules δ. Let $\beta = 0.5$,

$$\mathcal{S} = \{1, 2\}, \quad A_1 = \{1, 2, 3\}, \quad A_2 = \{1\}, \quad p_{11}^1 = p_{22}^1 = p_{12}^2 = 1,$$

$$p_{11}^3 = 1 - \alpha, \quad p_{12}^3 = \alpha, \quad r(1, \cdot, \cdot) \equiv 0, \quad \text{and} \quad r(2, \cdot, \cdot) \equiv 1.$$

Let π and π' denote the policies which always take actions 2 and 1, respectively, in state 1. Let δ be the single-state decision rule which takes action 3 in state 1. Straightforward calculations yield $v_1(\pi) = 1$, $v_2(\pi) = v_2(\pi') = v_2(\delta, \pi) = v_2(\delta, \pi') = 2$, $v_1(\pi') = 0$, $v_1(\delta, \pi) = (1 + \alpha)/2$, $v_1(\delta, \pi') = \alpha$, $V_1(\pi) = V_2(\pi) = V_1(\pi') = V_2(\pi') = V_2(\delta, \pi) = V_2(\delta, \pi') = 0$, $V_1(\delta, \pi) = \alpha(1 - \alpha)/4$, and $V_1(\delta, \pi') = \alpha(1 - \alpha)$. These values satisfy the hypothesis of (8) but violate its conclusion:

$$V_j(\pi) \geqq V_j(\pi') \quad \text{for } j = 1, 2, \quad \text{but} \quad V_1(\delta, \pi) < V_1(\delta, \pi').$$

In spite of the preceding counterexample, one could perform the calculations of the policy improvement algorithm (or other kinds of algorithms) while striving to optimize a criterion such as $E(B_s) - \lambda\sqrt{\text{variance of } B_s}$. However, the

† The author is grateful to Eric V. Denardo for comments on an earlier draft of this section which focused on the criterion $E(B_s) - \lambda\sqrt{\text{variance of } B_s}$ with $\lambda > 0$.

counterexample should make one question both finite termination of the algorithm and, if so, optimality of the terminal policy.

5. Numerical example

Let $\beta = 0.5$, $\mathcal{S} = \{1,2,3\}$, $A_1 = A_2 = \{1\}$, $A_3 = \{1,2\}$, $p^1_{11} = p^1_{22} = p^2_{32} = 1$, $p^1_{31} = 0.2$, $p^1_{33} = 0.8$, $r(s,a,\cdot) \equiv r(s,a)$ for all s and a, $r(1,1) = 10$, $r(2,1) = 1.5$, $r(3,1) = 0$, and $r(3,2) = 1.5$. Let δ and γ denote the policies which always take actions 1 and 2, respectively, in state 3.

For γ, the transposes of $\boldsymbol{v}$ and $\boldsymbol{V}$ are $(20,3,3)$ and $(0,0,0)$, respectively. For policy δ, the transposes of $\boldsymbol{v}$ and $\boldsymbol{V}$ are $(20,3,10/3)$ and $(0,0,13.888)$, respectively, as may be verified by direct calculation using the geometric distribution of the number of periods until departure from state 3.

For policy γ, using (6),

$$\theta_1 = (0.5)^2(20)^2 - (20-10)^2 = 0 \quad \text{and}$$

$$\theta_3 = \theta_2 = (0.5)^2(3)^2 - (3-1.5)^2 = 0$$

so each component of $\boldsymbol{V} = (I - \beta^2 P)^{-1}\boldsymbol{\theta}$ is 0 as it should be.

For policy δ, using (6),

$$\theta_1 = (0.5)^2(20)^2 - (20-10)^2 = 0, \qquad \theta_2 = (0.5)^2(3)^2 - (3-1.5)^2 = 0,$$

and

$$\theta_3 = (0.5)^2[(0.2)(20)^2 + (0.8)(10/3)^2] - (10/3-0)^2 = 11.1111.$$

The transpose of the third column of $(I - \beta^2 P)^{-1}$ is $(0,0,5/4)$ so the transpose of $(I - \beta^2 P)^{-1}\boldsymbol{\theta}$ is $(0,0,13.888)$ as it should be.

Acknowledgement

The author is grateful to Professor Eric V. Denardo and Dr Daniel P. Heyman for remarks on an earlier version.

References

DENARDO, E. V. (1967) Contraction mappings in the theory underlying dynamic programming. *SIAM Rev.* **9**, 165–177.

DENARDO, E. V. (1971) Markov renewal programming with small interest rates. *Ann. Math. Statist.* **42**, 477–496.

DERMAN, C. (1970) *Finite State Markovian Decision Processes.* Academic Press, New York.

FEREJOHN, J. AND PAGE, T. (1978) On the foundations of intertemporal choice. *Amer. J. Agricultural Econom.* **60**, 269–275.

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

KEMENY, J. G. AND SNELL, J. L. (1960) *Finite Markov Chains.* Van Nostrand, New York.

KREPS, D. M. AND PORTEUS, E. L. (1978) Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* **46**, 185–200.

KUSHNER, H. (1971) *Introduction to Stochastic Control.* Holt, New York.

MANDL, P. (1971) On the variance of controlled Markov chains. *Kybernetika* **7**, 1–12.

MENDELSSOHN, R. (1980) A systematic approach to determining mean-variance tradeoffs when managing randomly varying populations. *Math. Biosci.* **50**, 75–84.

MINE, H. AND OSAKI, S. (1970) *Markovian Decision Processes.* American Elsevier, New York.

PLATZMAN, L. K. (1978) Mimeographed lecture notes for IOE 315. Dept. of Industrial and Operations Engineering, University of Michigan, Ann Arbor.

SOBEL, M. J. (1975) Ordinal dynamic programming. *Management Sci.* **21**, 967–975.

STANCU-MINASIAN, I. M. AND WETS, M. J. (1976) A research bibliography in stochastic programming. *Operat. Res.* **24**, 1078–1119.

WHITE, D. J. (1974) Dynamic programming and probabilistic constraints. *Operat. Res.* **22**, 654–664.