数理解析研究所講究録
978巻 1997年 128-135

A minimum average-variance in Markov decision processes

和歌山大教育　門田 良信
(Yoshinobu KADOTA)*

Abstract†

This paper is concerned with the average-variance of Markov decision processes with countable states and finite actions. Sufficient conditions will be given to assure that there is a stationary deterministic policy which minimizes the average-variance in a class of the mean-optimal policies. The class of the policies is detetermined by the quantity of the actions which do not satisfy the mean-optimal equation.

1. Introduction.

As well as mean (viz. expected average) rewards, variances and their related criteria of Markov decision processes (MDP's) have been studied by many authors. Many of their works are overviewed by White(1988) and seen also in Filar, Kallenberg and Lee(1989).

But to our knowledge, there are a few papers which are concerned with the minimization of the average-variance in the class of mean-optimal policies. The average-variance is the applied form to MDP's of the variance given by Kemeny and Snell(1976). In finite state MDP's, Mandl(1971) investigates the asymptotic behavior of the average-variance in details. He contributes to construct MDP's whose mean rewards are the variances for the given MDP's. Kurano(1987) shows in general state MDP's that there is a stationary deterministic policy which minimizes the average-variance in the restricted class of policies whose actions satisfy the mean-optimal equation.

In this paper, the MDP's have countable states and finite actions. We show the existence of a stationary deterministic policy which minimizes the average-variance in the larger class of mean-optimal policies. The class is determined in relation to the mean-optimal equation. As a corollary, if the state space is finite, the class is given by the set of all mean-optimal policies.

In section 2, the necessary notations and the problem to be examined is stated. Also, under an ergodic condition, the properties of mean-optimal policies are reviewed. In

*Faculty of Education, Wakayama University, Wakayama 640, Japan

†*Keywords and phrases*: average-variance, Markov decision process, mean-optimal, mean-optimal equation.

AMS(MOS) subject classification: 90C40, 60J10.

section 3, applying the equality given by Kurano(1987), we obtain a lemma which describes Mandl(1971)'s idea more generally. Owing to the lemma, the main theorem is proved.

2. Definitions and notation.

Our MDP's are specified by (S, A, p, r), where $S = \{0, 1, 2, \cdots\}$ is the set of states, the subset $A(i)$ of A is the set of actions available at each state $i \in S$, $p = (p(a)_{ij})$ is the matrix of transition probabilities satisfying that $\sum_{j \in S} p(a)_{ij} = 1$ for any $i \in S$ and $a \in A(i)$, and $r(i, a)$ is an immediate reward function defined on $\{(i, a);\ i \in S,\ a \in A(i)\}$. We assume that S is countable, each $A(i)$ is finite and that r is uniformly bounded, i.e., $|r(i, a)| \leq M$ for any $i \in S$ and $a \in A(i)$.

The sample space is the product space $\Omega = (S \times A)^{\infty}$ such that the projection (X_n, Δ_n) to the n-th factor $S \times A$ describes the state and the action at time n of the process respectively for $n = 0, 1, 2, \cdots$.

A policy $\pi = (\pi_0, \pi_1, \cdots)$ is a sequence of conditional probability π_n such that $\pi_n(A(i_n) | i_0, a_0, i_1, \cdots, i_n) = 1$ for any history $(i_0, a_0, \cdots i_n) \in (S \times A)^n \times S$. A policy π is called Markov if $\pi_n(a | i_0, a_0, \cdots, i_n) = \pi_n(a | i_n)$ for any $a \in A(i)$, n and $(i_0, a_0, \cdots i_n)$. A Markov policy π is called deterministic if there is a function f_n on S with $f_n(i) \in A(i)$ for any n such that $\pi_n(\{f_n(i)\} | i_n = i) = 1$ for any $i \in S$. A deterministic policy is called stationary if $f_n = f$ for all $n = 0, 1, \cdots$. The stationary policy is denoted by $\pi = f$. Let Π and Π_S be the sets of all policies and stationary deterministic policies, respectively.

Let $H_n = (X_0, \Delta_0, \cdots, \Delta_{n-1}, X_n)$ for $n = 0, 1, 2, \cdots$. We assume for any $\pi \in \Pi$ with $n = 0, 1, \cdots$, $i, j \in S$ and $a \in A(i)$,

$$P^{\pi}(X_{n+1} = j | H_{n-1}, \Delta_{n-1}, X_n = i, \Delta_n = a) = p(a)_{ij}.$$

An initial state $i \in S$ and a policy $\pi = (\pi_0, \pi_1, \cdots)$ determine the probability measure P_i^{π} on Ω by the usual way. The expectation of a random variable Y with respect to P_i^{π} is denoted by $E_i^{\pi}(Y)$.

For a policy π, the long-run mean (or expected average) reward per unit time starting from $i \in S$ is defined by

$$x(i, \pi) = \liminf_{n \to \infty} \frac{1}{n+1} E_i^{\pi}\Big[\sum_{k=0}^{n} r(X_k, \Delta_k)\Big]. \tag{1}$$

The average-variance for the policy is defined, following to Kurano(1987), by

$$\psi(i, \pi) = \limsup_{n \to \infty} \frac{1}{n+1} V_i^{\pi}\Big[\sum_{k=0}^{n} r(X_k, \Delta_k)\Big],$$

where $V_i^\pi[Y] = E_i^\pi\{Y - E_i^\pi(Y)\}^2$ for a random variable Y. The average-variance is the application to MDP's of the variance given in Kemeney and Snell(1976).

Let $x^*(i) = \sup\{x(i,\pi);\ \pi \in \Pi\}$. If a policy π^* satisfies $x(i,\pi^*) = x^*(i)$ for any $i \in S$, we say π^* is mean-optimal. We denote by $\Pi(M)$ the set of mean-optimal policies.

For any $f \in \Pi_S$, let

$$m(f)_{ij} = \sum_{n=1}^{\infty} nP_i^f(\,X_k \neq j \ \text{ for } \ 1 \leq k < n,\ X_n = j),$$

which is called the mean recurrence time to go from $i \in S$ to $j \in S$. We shall set up a condition.

Condition I. There are a state $0 \in S$ and a constant $b > 0$ such that $m(f)_{i0} \leq b$ for all $i \in S$ and $f \in \Pi_S$.

Remark 1. According to Ross(1970), if S is finite and each stationary Markov chain $\{p(f)_{ij};\ i,j \in S\}$ is irreducible, Condition I holds. Federgruen, Hordijk and Tijms(1978) shows that Condition I holds if $\lim_{n\to\infty} p(f)_{ij}^n$ exists independently of $i \in S$ and the limit is approached with exponential speed. Dekker and Hordijk(1992) and Mann(1985) generalize Condition I to the multi-chain case.

Ross(1970) shows under Condition I that (i) $x(i,\ f)$ is independent of $i \in S$, (ii) there exists a bounded vector $v = (v(i))$ satisfying

$$x^* + v(i) = \max_{a\in A(i)}\{r(i,a) + \sum_{j\in S} p(a)_{ij}v(j)\} \quad \text{for any } \ i \in S, \tag{2}$$

where $x^* = \sup\{x(i,\ f);\ f \in \Pi_S\}$, and (iii) If $f \in \Pi_S$ maximizes the term in the brackets of (2) for any $i \in S$, then $f \in \Pi(M)$. Letting

$$v(i,\ f) = \lim_{\beta\nearrow 1} \sum_{n=0}^{\infty} \beta^n\{E_i^f r(X_n,\ \Delta_n) - x(i,\ f)\} \quad \text{for } f \in \Pi_S,$$

Mann(1985) shows that there is $f^* \in \Pi_S$ which satisfies the above (iii) with $v(i,\ f^*) = v(i)$ for $i \in S$.

For the analysis of the mean-optimality, let

$$\varphi(i,\ a) \ = \ r(i,\ a) + \sum_{j\in S} p(a)_{ij}v(j) - x^* - v(i) \quad \text{and} \tag{3}$$

$$K(i) \ = \ \{a \in A(i);\ \varphi(i,\ a) = 0\} \quad \text{for any } \ i \in S. \tag{4}$$

Let $K = \times_{i\in S} K(i)$. Let denote by $\Pi(K)$ the set of policies $\pi = (\pi_0,\ \pi_1, \cdots)$ such that $\pi_n(K(i_n)|\, i_0,\ a_0,\ i_1, \cdots, i_n) = 1$ for any history $(i_0,\ a_0, \cdots i_n)$.

Let define a function $\tilde{r}(i, a)$ by

$$\tilde{r}(i, a) = \sum_{j \in S} p(a)_{ij}\{v(j) - \sum_{h \in S} p(a)_{ih} v(h)\}^2 \quad \text{for any } i \in S \text{ and } a \in A(i).$$

Notice that $\tilde{r}$ is uniformly bounded. Using $\tilde{r}$ instead of r, $\tilde{x}(i, \pi)$ is defined by *limsup* similarly as (1). For MDP's specified by $(S, K, p, \tilde{r})$, $\tilde{x}_*$ is defined by *inf* similarly as x^*. The bounded vector $\tilde{v} = (\tilde{v}(i))$ exists. Corresponding to (3), let

$$\tilde{\varphi}(i, a) = \tilde{r}(i, a) + \sum_{j \in S} p(a)_{ij}\tilde{v}(j) - \tilde{x}_* - \tilde{v}(i) \quad \text{for any } i \in S \text{ and } a \in A(i).$$

Then, $\tilde{\varphi}(i, a) \geq 0$ for any $i \in S$ and $a \in K(i)$. Corresponding to (4), let $\widetilde{K}(i) = \{a \in K(i);\ \tilde{\varphi}(i, a) = 0\}$. Then, $\widetilde{K}(i) \neq \emptyset$ for all $i \in S$. Let a policy $\tilde{f} \in \Pi(K)_S$ be $\tilde{f}(i) \in \widetilde{K}(i)$ for any $i \in S$.

Kurano(1987) and Mandl(1971) show that $\tilde{f} \in \Pi(K)$ has the minimum average-variance within the class of policies satisfying $\sum_{k=0}^{\infty} P_i^{\pi}(\Delta_k \notin K(X_k)) < \infty$.

3. Class of policies to the minimum average-variance.

This section gives sufficient conditions to determine the class of policies where $\tilde{f} \in \Pi(K)_S$ in the previous section has the minimum average-variance. To the end, two lemmas are prepared.

Next Lemma 3.1 estimates the quantity of the actions which do not satisfy (2) for the mean-optimal policies.

Lemma 3.1. *Suppose that Condition I holds. If $\pi \in \Pi(M)$, it follows that*

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} P_i^{\pi}(X_k = j,\ \Delta_k \notin K(j)) = 0 \quad \text{for any} \quad i, j \in S \tag{5}$$

Proof. We have from (3) that

$$\sum_{k=0}^{n} E_i^{\pi} \varphi(X_k, \Delta_k) = \sum_{k=0}^{n} E_i^{\pi} r(X_k, \Delta_k) - (n+1)x^* - v(i) + E_i^{\pi} v(X_{n+1}). \tag{6}$$

Since $\varphi(i, a) \leq 0$ for any $i \in S$ and $a \in A(i)$, a policy π is mean-optimal if and only if

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} E_i^{\pi}(\varphi(X_k, \Delta_k)) = 0 \quad \text{for} \quad i \in S. \tag{7}$$

Let $\varepsilon_j = \max\{\varphi(j, a);\ a \notin K(j)\}$. Then, $\varepsilon_j < 0$ for any $j \in S$ since $A(j)$ is finite. Take the terms $\varphi(j, a) = 0$ away from (7), we get

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \sum_{j \in S} P_i^{\pi}(X_k = j,\ \Delta_k \notin K(j))\varepsilon_j = 0. \tag{8}$$

(8) implies (5) immediately, completing the proof. □

For the sake of brevity, we shall omit the notation Δ_k freely in all the proofs of the subsequent propositions if no confusion occurs. In paticular, the conditional expectation $E_i^\pi(v(X_{k+1})|X_k, \Delta_k)$ will be denoted by $E_i^\pi(v(X_{k+1})|X_k)$.

In the proof of Lemma 3.2, we shall use the followig equality which is given by Lemma 3.2 in Kurano(1987).

$$E_i^\pi\Big\{\sum_{k=1}^{n}(v(X_k) - E_i^\pi(v(X_k)|X_{k-1}))\Big\}^2 = \sum_{k=0}^{n-1} E_i^\pi \tilde{r}(X_k, \Delta_k). \tag{9}$$

Lemma 3.2. *Suppose Condition I holds. Then, it holds for any policy* $\pi \in \Pi$ *and* $n = 0, 1, 2, \cdots$ *that*

$$\begin{aligned} &\Big|\Big\{V_i^\pi(\sum_{k=0}^{n} r(X_k, \Delta_k)) - V_i^\pi(\sum_{k=0}^{n}\varphi(X_k, \Delta_k))\Big\} - \sum_{k=0}^{n} E_i^\pi \tilde{r}(X_k, \Delta_k)\Big| \\ &\le 2\Big[\Big\{\sum_{k=0}^{n-1} E_i^\pi \tilde{r}(X_k, \Delta_k)\Big\}^{\frac{1}{2}} + M_1\Big]\Big\{V_i^\pi(\sum_{k=0}^{n}\varphi(X_k, \Delta_k))\Big\}^{\frac{1}{2}} + o(n), \end{aligned} \tag{10}$$

where M_1 *is a constant and* $o(n)$ *is a number such that* $\lim_{n\to\infty} o(n)/n = 0$.

Proof. Let $\pi \in \Pi$. For a random variable Y and a constant μ, it follows that $E(Y - E(Y))^2 = E(Y - \mu)^2 - (E(Y) - \mu)^2$. Substitute $Y = \sum_{k=0}^{n} r(X_k)$ and $\mu = \sum_{k=0}^{n}(x^* + E_i^\pi\varphi(X_k))$ to the equality. Using (6), we have

$$\begin{aligned} V_i^\pi\{\sum_{k=0}^{n} r(X_k)\} &= E_i^\pi\Big\{\sum_{k=0}^{n}(r(X_k) - x^* - E_i^\pi\varphi(X_k))\Big\}^2 \\ &\quad - \{v(i) - E_i^\pi v(X_{n+1})\}^2. \end{aligned} \tag{11}$$

On the other-hand, we have from the definition of $\tilde{r}$ and (9) that

$$\sum_{k=0}^{n} E_i^\pi(\tilde{r}(X_k)) = E_i^\pi\Big\{\sum_{k=0}^{n}(v(X_k) - E_i^\pi(v(X_{k+1})|X_k)) - (v(X_0) - v(X_{n+1}))\Big\}^2. \tag{12}$$

Expand (12) with the form $a^2 - 2ab + b^2$. Substitute (3) to a^2. Rearrange the terms in $2ab$ following:

(13)

$$E_i^\pi\Big\{v(X_{n+1})\sum_{k=1}^{n}(v(X_k) - E_i^\pi(v(X_k)|X_{k-1})) + v(X_{n+1})(v(X_0) - E_i^\pi(v(X_{n+1})|X_n))\Big\}.$$

After sharing E_i^π, apply the Schwarz inequality to (13) and use (9). Since v is bounded, (13) is represented by $o(n)$. Then, (12) turns out to be

$$\sum_{k=0}^{n} E_i^\pi \tilde{r}(X_k) = E_i^\pi\{\sum_{k=0}^{n}(r(X_k) - x^* - \varphi(X_k))\}^2 + o(n). \tag{14}$$

Comparing (14) with (11), we have

$$\begin{aligned}(15)\qquad &\sum_{k=0}^{n} E_i^{\pi}\tilde{r}(X_k) = V_i^{\pi}(\sum_{k=0}^{n} r(X_k)) + V_i^{\pi}(\sum_{k=0}^{n} \varphi(X_k)) \\ &-2E_i^{\pi}\Big\{\sum_{k=0}^{n}(r(X_k) - x^* - E_i^{\pi}\varphi(X_k))\sum_{\ell=0}^{n}(\varphi(X_\ell) - E_i^{\pi}\varphi(X_\ell))\Big\} + o(n).\end{aligned}$$

Substitute (3) to $r(X_k) - x^*$ in (15). In the same way as (13), apply the Schwarz inequalities to both terms and use (9). The constant M_1 should be taken to satisfy

$$M_1 \geq \{E_i^{\pi}(v(X_0) - E_i^{\pi}(v(X_{n+1})|X_n))^2\}^{\frac{1}{2}} \quad \text{for all } n.$$

Since v is bounded, such M_1 exists. This completes the proof. □

Theorem 3.3. *Suppose Condition I holds. Let $\pi \in \Pi$ satisfy*

$$(16)\qquad \lim_{n\to\infty} \frac{1}{n+1} V_i^{\pi}(\sum_{k=0}^{n}(\varphi(X_k, \Delta_k)) = 0 \quad \textit{and}$$

$$(17)\qquad \limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} E_i^{\pi}(\tilde{\varphi}(X_k, \Delta_k)) \geq 0 \quad \textit{for any } i \in S.$$

Then, it holds that

$$(18)\qquad \psi(i, \tilde{f}) = \tilde{x}(i, \tilde{f}) \leq \tilde{x}(i, \pi) = \psi(i, \pi) \quad \textit{for any } i \in S.$$

Proof. Dividing (10) by $n+1$, let n tend to infinity. We have from (16) that $\tilde{x}(i, \pi) = \psi(i, \pi)$ for any $i \in S$. In particular, $\tilde{x}_* = \tilde{x}(i, \tilde{f}) = \psi(i, \tilde{f})$, since $\varphi(j, \tilde{f}(j)) = 0$ for any $j \in S$.

The relation between $\sum_{k=0}^{n} E_i^{\pi}\tilde{r}(X_k)$ and $\sum_{k=0}^{n} E_i^{\pi}\tilde{\varphi}(X_k)$ is given similarly as (6). Then (17) implies

$$\tilde{x}(i, \pi) = \tilde{x}_* + \limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} E_i^{\pi}(\tilde{\varphi}(X_k)) \geq \tilde{x}_*.$$

Thus (18) is obtained, completing the proof. □

Corollary 3.4. *Suppose Condition I holds. If a policy $\pi \in \Pi$ satisfies*

$$(19)\qquad \lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} P_i^{\pi}(\Delta_k \notin K(X_k)) = 0, \quad \textit{for any } i \in S,$$

then $\pi \in \Pi(M)$ and satisfies (16) and (17).

Proof. The equality (19) is written by

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n}\sum_{j\in S}\sum_{a\notin K(j)}P_i^{\pi}(X_k=j,\ \Delta_k=a)=0. \tag{20}$$

Notice that $\varphi(j,a)=0$ for $a\in K(j)$. By the bounded convergence theorem, (20) implies (7), so that $\pi\in\Pi(M)$. Since $\tilde{\varphi}(j,\ a)\geq 0$ for $a\in K(j)$, (20) implies (17) similarly.

Let $|\varphi(i,a)|\leq M$ for some $M>0$ and for any $i\in S$ and $a\in A(i)$. Notice that

$$E_i^{\pi}(\sum_{k=0}^{n}\varphi(X_k))^2=\sum_{k=0}^{n}E_i^{\pi}(\varphi(X_k)^2)+2\sum_{k=1}^{n}\sum_{\ell=0}^{k-1}E_i^{\pi}|\varphi(X_\ell)\varphi(X_k)| \quad \text{and}$$
$$E_i^{\pi}|\varphi(X_\ell)\varphi(X_k)|\leq M^2P_i^{\pi}(\Delta_\ell\notin K(X_\ell) \ \text{ and } \ \Delta_k\notin K(X_k))$$

for $0\leq\ell\leq k$ and $k=0,1,2\cdots$. Then, (19) implies $\lim_{n\to\infty}E_i^{\pi}(\sum_{k=0}^{n}\varphi(X_k))^2/(n+1)=0$, so that (16) holds. The proof is complete. □

Corollary 3.5. *Suppose Condition I holds. If S is finite, (18) holds for any $\pi\in\Pi(M)$.*

Proof. By summing (5) in $j\in S$, we see that $\pi\in\Pi(M)$ satisfies (19). Then, Corollary 3.4 and Theorem 3.3 implies (18). The proof is complete. □

References

[1] R. Dekker and A. Hordijk(1992), Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains, *Mathematics of Operations Reseach*, **17**, 2, 271–289.

[2] A. Federgruen, A. Hordik and H. C. Tijms(1978), Recurrence conditions in denumerable state Markov decision processes, *Dynamic Programming and Its Applications*, edited by M. L. Putterman, 3–22.

[3] J.A. Filar, L.C.M. Kallenberg and H.M. Lee(1989), Variance-penalized Markov decision processes, *Mathematics of Operations Reseach*, **14**, 1, 147–161.

[4] J.G. Kemeny and J.L. Snell(1976), Finite Markov Chains, Springer-Verlag, New York.

[5] M. Kurano(1987), Markov decision processes with a minimum-variance criterion, *Journal of Mathematical Analysis and Applications.*, **123**, 572–583.

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

[7] E. Mann(1985), Optimality equations and sensitive optimality in bounded Markov decision processes. *Optimization*, **16**, 5, 767–781.

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

[9] D.J. White(1988), Mean, variance, and probabilistic criteria in finite Markov decision processes: A review, *Journal of Optimization Theory and Applications*, **56**, 1, 1–29.