Some Remarks on Optimality Conditions for Markovian Decision Process

By

Hisashi MINE* and Yoshio TABATA*

(Received March 31, 1969)

This paper is concerned with a discrete time parameter Markovian decision problem. The expected total returns for infinite horizon are considered as the power series of discount factor. Some optimality criterions, for example, β-optimal, 1-optimal and so on, are discussed from the view point of the theory of infinite series. And a new optimality criterion is introduced. This criterion is valuable to construct an intuitive optimal policy theoretically.

1. Introduction

In recent years much effort has been devoted to the infinite horizon problem of discrete time parameter Markovian sequential decision process with finitely many states and actions. Blackwell [1] discusses the case with a discount factor β, $0 \leqq \beta < 1$, and defines an optimal policy, termed β-optimal, that maximizes the total expected discounted return over an infinite horizon. He shows the existence of a stationary β-optimal policy. Moreover for the case of $\beta = 1$, he defines a 1-optimal policy as the limiting one of β-optimal policy in some sense and shows the existence of a stationary 1-optimal policy.

Veinott [2] establishes the algorithm to find a 1-optimal policy and proposes another optimality criterion, called Veinott's optimal, maximizing the Cesaro sum of the vectors of expected returns received in a finite horizon tending the horizon to infinity. Denardo and Miller [3] verify Veinott's conjecture that there exist optimal stationary policies.

In this paper, some optimality criterions such as β-optimal, 1-optimal, mean-optimal and Veinott's optimal are discussed by applying the theory of infinite series and integral, and the significance of each of the criterions is shown. Moreover, some relations and a new optimality criterion are introduced.

The following mathematical model is considered in this paper. Consider

* Department of Applied Mathematics and Physics

a system that is observed in sequence at the discrete time points labeled 1, 2, $\cdots$, N, $\cdots$. At each point, the system is observed to be in one of S states labeled 1, 2, $\cdots$, S. If the system is in state i, a decision k is selected from a finite set F_i of possible decisions and an immediate expected return $r(i, k)$ is received. $q(j|i, k)$ is the conditional probability that the system is in state j at time $n+1$ $(n=1, 2, \cdots)$, given that the system was in state i and that the decision k was selected at time n. And the data $r(i, k)$ and $q(j|i, k)$ are known to the decision-maker and assumed to depend only on i, j and k.

Let $F=\times_{i=1}^{S}F_i$. A policy π is a sequence $(f_1, f_2, \cdots)$ of the element of F and using a policy π means that the selected decision is $f_n(i)$ given that the system was in state i at time n. Let $f^\infty=(f, f, \cdots)$. f^∞ is called a stationary policy. For any $f\in F$, the $S\times 1$ column vector $r(f)$ whose ith element is $r(i, f(i))$ and the $S\times S$ Markov matrix $Q(f)$ whose (i, j) element is $q(j|i, f(i))$ are introduced. Thus, for the policy $\pi=(f_1, f_2, \cdots)$,

$$Q_n(\pi) = Q(f_1)\,Q(f_2)\cdots Q(f_n)\,.$$

For any two column vectors u, v, they are denoted as $u\geqq v$, if every element of u is at least as large as the corresponding element v. The convergence of the infinite series whose each single term is vector is defined as the convergence of all of its elements.

2. β-optimal policy

Using the policy π, the vector of total expected discounted returns starting from each state is represented by

$$V_\beta(\pi) = \sum_{n=0}^{\infty}\beta^n Q_n(\pi)\,r(f_{n+1}) = \sum_{n=0}^{\infty}A_n(\pi)\beta^n\,, \qquad (1)$$

where $0\leqq\beta<1$ is a discount factor and $Q_0(\pi)$ is the $S\times S$ identity matrix I, and the finite vector $A_n(\pi)$ being independent of β is described by

$$A_n(\pi) = Q_n(\pi)\,r(f_{n+1}) = Q(f_1)\,Q(f_2)\cdots Q(f_n)\,r(f_{n+1})\,. \qquad (2)$$

The equation (1) is considered as a power series of β. The following Lemma 1 holds for the convergence of this power series.

Lemma 1. The total expected discounted return with infinite horizon converges uniformly to $V_\beta(\pi)$ in the interval $0\leqq\beta<1$, when the immediate expected returns $r(f_n)$ are finite for all n.

Proof. As the matrices $Q_n(\pi)$ $(n=0, 1, 2, \cdots)$ are Markov matrices and $r(f_n)$ are finite for all n, there exists a finite number M such as

$$A_n(\pi) = Q_n(\pi) r(f_{n+1}) \leqq M(<\infty) .$$

Then for any β $(0 \leqq \beta < 1)$,

$$\sum A_n(\pi) \beta^n \leqq \sum M \beta^n = \frac{M}{1-\beta} < \infty .$$

Therefore the power series $\sum_{n=0}^{\infty} A_n(\pi)\beta^n$ converges uniformly to $V_\beta(\pi)$.

Now consider the radius of convergence R of the power series (1). From Cauchy-Hadamard's Theorem, if

$$\limsup_{n\to\infty} \sqrt[n]{|A_n(\pi)|} = l , \tag{3}$$

then the radius of convergence R is given by

(i) $R=+\infty$ if $l=0$

(ii) $R=0$ if $l=\infty$

(iii) $R=\frac{1}{l}$ if $l \neq 0$ and ∞.

Since the power series $\sum_{n=0}^{\infty} A_n(\pi)\beta^n$ converges in the interval $0 \leqq \beta < 1$ by Lemma 1, $R \neq 0$ follows from the definition of the radius of convergence, that is, the case $l=\infty$ of (ii) never appears. On the other hand, discount factor β is not necessarily introduced if the series $\sum_{n=0}^{\infty} A_n(\pi)$ converges to a finite value, that is the power series $\sum_{n=0}^{\infty} A_n(\pi)\beta^n$ converges for $\beta=1$. In this case the optimal policy is defined from the total expected return with no discounting. Therefore this case is not considered in this paper. Therefore, it is assumed that the series $\sum_{n=0}^{\infty} A_n(\pi)$ diverges.

Moreover, this paper is not concerned with the cases where the immediate expected returns $r(f_n)$ are identically zero or all the immediate expected returns after N decision point are zero, for example, a Markov chain has an absorbing state and the immediate return is zero in this state. Under the conditions mentioned above, from $0<|A_n(\pi)|<\infty$ for all n, the following equation holds

$$\limsup_{n\to\infty} \sqrt[n]{|A_n(\pi)|} = 1 = l .$$

Therefore the radius of convergence R of power series (1) is equal to one under the above assumptions.

Consider two stationary policies $\pi=f^\infty$ and $\pi'=g^\infty$. Assume that two total expected discounted returns corresponding to these policies are equal to each other. Then by equality $V_\beta(\pi)=V_\beta(\pi')$, the following equation is satisfied;

$$\sum A_n(\pi)\beta^n = \sum A_n(\pi')\beta^n \tag{4}$$

For $n=0, 1, 2, \cdots$, $A_n(\pi)=A_n(\pi')$, that is to say,

$$Q^n(f)r(f) = Q^n(g)r(g) .$$

Letting $n=0$ and 1, the equations

$$r(f) = r(g) \quad \text{and} \quad Q(f) = Q(g)$$

hold. Therefore, the two stationary policies which give the same total expected discounted returns are regarded as identical policies.

Now let introduce an optimal policy for the total expected discounted returns. The following definition of β-optimal policy by Blackwell [1] is useful for the definition of the optimal policy with a discount factor.

Definition 1.···Blackwell. If the equation $\sum(A_n(\pi^*)-A_n(\pi))\beta^n \geqq 0$ is satisfied for any policy π, then the policy π^* is called β-optimal policy.

It is well-known that the following important Theorem 1 holds for β-optimal policy.

Theorem 1.···Blackwell. There exists a stationary β-optimal policy.

The β-optimal policy is found by Howard's Policy Iteration Method. It is noted that the β-optimal policy is the one which maximizes the sum of the power series (1) for some fixed β.

3. 1-optimal policy

In the case of $\beta=1$, the total expected return diverges in general. That is to say, the power series (1) does not always converge for $\beta=1$. Then it is impossible to find an optimal policy by means of comparison of any two total expected returns with no discoudting from the definition 1. It is possible, however, to find it by considering the limiting case of β-optimal policy as β tends to 1. Then Blackwell [1] has proceeded with his conception in accordance with the following definition of l-optimal policy.

Definition 2. Let π^* be a β-optimal policy. If the equation

$$\lim_{\beta \to 1-0} [V_\beta(\pi) - V_\beta(\pi^*)] = 0 \tag{5}$$

is satisfied for some policy π, that is,

$$\lim_{\beta \to 1-0} \sum_{n=0}^{\infty} [A_n(\pi) - A_n(\pi^*)]\beta^n = 0 , \tag{6}$$

then the policy π is defined to be a 1-optimal policy.

In this paper, let consider the meaning of Definition 2 from the view point

of the power series (1). Introduce the vector $B_n(\pi, \pi^*)$ such as

$$A_n(\pi)-A_n(\pi^*) = B_n(\pi, \pi^*) \tag{7}$$

for two policies π and π^*. Since the radii of convergence for power series $\sum A_n(\pi)\beta^n$ and $\sum A_n(\pi^*)\beta^n$ are 1, the power series $\sum B_n(\pi, \pi^*)\beta^n$ converges uniformly in the interval $0 \leqq \beta < 1$. And if there exists the finite limit of series $\sum B_n(\pi, \pi^*)\beta^n$ as $\beta \to 1-0$, that is

$$\lim_{\beta \to 1-0} \sum_{n=0}^{\infty} B_n(\pi, \pi^*)\beta^n = C(\pi, \pi^*) . \tag{8}$$

then $C(\pi, \pi^*)$ is called the Abelian Sum. In particular, if π^* is β-optimal policy and the limit $C(\pi, \pi^*)$ is equal to zero, then the left-hand side of the equation (8) is reduced to the 1-optimal policy of Definition 2. Consequently Definition 2 is turned out as follows;

Definition 2′. Let π^* be a β-optimal policy, and for some policy π express $B_n(\pi, \pi^*)$ as follows;

$$B_n(\pi, \pi^*) = A_n(\pi)-A_n(\pi^*) .$$

If the Abelian Sum of series $\sum B_n(\pi, \pi^*)$ is equal to zero, then the policy π is called a l-optimal policy.

In the meantime, let consider the condition under which there exists the Abelian Sum $C(\pi, \pi^*)$ of $\sum B_n(\pi, \pi^*)$ in Definition 2′. For the necessary condition the following Theorem holds.

Theorem 2. Let π and π^* be two policies, and $B_n(\pi, \pi^*)$ denote

$$B_n(\pi, \pi^*) = A_n(\pi)-A_n(\pi^*) .$$

If the series $\sum B_n(\pi, \pi^*)$ converges to $C(\pi, \pi^*)$, then

$$\lim_{\beta \to 1-0} \sum B_n(\pi, \pi^*)\beta^n = C(\pi, \pi^*) . \tag{9}$$

First of all, consider the following Lemma 1 in order to prove Theorem 2. This Lemma 2 is concerned with the concept of α-optimal [1].

Lemma 2. If the series $\sum B_n(\pi, \pi^*)$ converges, then the power series $\sum B_n(\pi, \pi^*)\beta^n$ is uniformly convergent in an interval, $0 \leqq \beta \leqq \alpha$ (for some $\alpha \leqq 1$).

Proof. Let

$$S_{m,k} = B_{m+1}+B_{m+2}+ \cdots +B_k$$

and

$$\sum_{k=m+1}^{n} B_k(\pi, \pi^*)\beta^k = \sum_{k=m+1}^{n-1} S_{m,k}(\beta^k-\beta^{k+1})+S_{m,n}\beta^n .$$

Since the series $\sum B_n(\pi, \pi^*)$ converges in the given interval $0 \leqq \beta \leqq \alpha$, there exists some N which satisfies the inequality

$$|S_{m,k}| < \varepsilon$$

for any $\varepsilon > 0$ and every $k > m > N$. Thus for every m, n such $n > m > N$, the following inequality is satisfied.

$$|\sum_{k=m+1}^{n} B_k(\pi, \pi^*)\beta^k| < 2\varepsilon .$$

Then the power series $\sum B_n(\pi, \pi^*)\beta^n$ uniformly converges in the interval $0 \leqq \beta \leqq \alpha$.

In the second case let us prove Theorem 2.

Proof of Theorem 2. Let $g_n(\beta) = \beta^{n-1}$. By the assumption, the series $\sum B_n(\pi, \pi^*)$ converges. Since the sequence of function $\{g_n(\beta)\}$ is monotone and decreasing in $0 \leqq \beta \leqq 1$, $|g_n(\beta)| = |\beta^{n-1}| \leqq 1$, $\sum_{n=0}^{\infty} B_n(\pi, \pi^*)g_n(\beta) = \sum_{n=0}^{\infty} B_n(\pi, \pi^*)\beta^n$ converges uniformly in $0 \leqq \beta \leqq 1$ by Lemma 2. And the equation

$$\lim_{\beta \to 1-0} B_n(\pi, \pi^*)\beta^n = B_n(\pi, \pi^*)$$

gives the following;

$$\lim_{\beta \to 1-0} \sum B_n(\pi, \pi^*)\beta^n = \sum B_n(\pi, \pi^*) = C(\pi, \pi^*)$$

The proof is complete.

In the following, consider the conditions that the series $\sum B^n(\pi, \pi^*)$ converges. One of the necessary conditions is that the following equation holds.

$$\lim_{n \to \infty} B_n(\pi, \pi^*) = 0 .$$

Especially, for two stationary policies $\pi = f^\infty$ and $\pi^* = g^\infty$, the equation described above is reduced to the following.

$$\lim_{n \to \infty} (Q^n(f)r(f) - Q^n(g)r(g)) = 0 .$$

And if the Markov chains $Q(f)$ and $Q(g)$ are completely ergodic, then the necessary condition is represented as

$$Q^*(f)r(f) = Q^*(g)r(g) .$$

where matrices $Q^*(f)$ and $Q^*(g)$ are given by

$$\lim_{n \to \infty} Q^n(f) = Q^*(f)$$
$$\lim_{n \to \infty} Q^n(g) = Q^*(g) .$$

In practical situations, this necessary condition is very useful because of the simplicity when the consideration of the relations between β-optimal and 1-optimal policies is restricted in a class of stationary policies.

On the other hand, the following Lemma 3 is well known as the necessary and sufficient condition for the convergence of the series $\sum B_n(\pi, \pi^*)$.

Lemma 3. The necessary and sufficient condition for the convergence of the series $\sum B_n(\pi, \pi^*)$ is that there exists a positive number N which satisfies the inequality

$$|B_{m+1}+B_{m+2}+\cdots+B_n|<\varepsilon$$

for any positive number ε and all n, m such as $n>m>N$.

Many authors established the relations between β-optimal and 1-optimal policies in a class of the stationary plicies. But in this paper, Theorem 2 and Lemma 3 represented above shows the relations between β-optimal and 1-optimal policies in a class of all policies.

Now let consider the contrary case.

Assume that there exists a limit

$$\lim_{\beta\to 1-0}\sum_{n=0}^{\infty} B_n(\pi, \pi^*)\beta^n = C(\pi, \pi^*)\ .$$

Even if a limit $C(\pi, \pi^*)$ exists, the series $\sum B_n(\pi, \pi^*)$ does not always converge to a finite value. The following Theorem 3 which modifies slightly Tauber's Theorem is useful to the conditions of the convergence for the series $\sum B_n(\pi, \pi^*)$.

Theorem 3. Let π^* be a β-optimal policy and π be a 1-optimal policy. If

$$B_n(\pi, \pi^*) = 0(1/n)\ , \tag{10}$$

then the series $\sum B_n(\pi, \pi^*)$ converges to zero.

Proof. Define $S_n(\pi, \pi^*)$ and $f(\beta)$ by the following equation.

$$S_n(\pi, \pi^*) = B_0+B_1+B_2+\cdots+B_n\ ,$$

$$f(\beta) = \sum_{k=0}^{\infty} B_k(\pi, \pi^*)\beta^k\ .$$

Then $$|S_n(\pi, \pi^*)-f(\beta)| \leqq |\sum_{k=1}^{n} B_k(1-\beta^k)| + |\sum_{k=n+1}^{\infty} B_k\beta^k|\ .$$

Note that, for $0\leqq\beta<1$,

$$1-\beta^k = (1-\beta)(1+\beta+\beta^2+\cdots+\beta^{k-1}) \leqq (1-\beta)^k$$

Then, it follows that

$$|S_n(\pi, \pi^*)-f(\beta)| \leqq (1-\beta)\sum_{k=1}^{n} k|B_k| + \sum_{k=n+1}^{\infty}|B_k|\beta^k\ .$$

From the assumption, $n|B_n| \to 0$. Then there exists $m>0$ such that for any $\varepsilon>0$ and all $n>m$,

$$n|B_n| < \varepsilon, \quad \text{and} \quad (1/n)\sum_{k=1}^{n} k|B_k| < \varepsilon .$$

Thus, it holds that, for $n>m$,

$$\sum_{k>m} |B_k|\beta^k = \sum_{k>m} k|B_k|\frac{1}{k}\beta_k \leqq \sum \frac{\varepsilon}{n}\beta^k < \frac{\varepsilon \cdot 1}{n(1-\beta)}$$

Therefore, the following inequality is implied for

$$\left|S_n(\pi, \pi^*) - f\left(1-\frac{1}{n}\right)\right| < \frac{1}{n}\sum_{k=1}^{n} k|B_k| + \varepsilon < \varepsilon + \varepsilon = 2\varepsilon ,$$

where $\beta = 1-1/n$.

Then $f(1-1/n)\to 0$ implies $S_n \to 0$. Note that, from the definition 2 of 1-optimal, $f(1-1/n)\to 0$ as $n\to\infty$. Then the proof is complete.

4. Mean-optimal policy

Let $V_n(\pi)$ be the vector of total expected returns with no discounting (that is $\beta=1$) for epochs 1 through n using policy π. Then

$$V_n(\pi) = \sum_{k=0}^{n-1} Q_k(\pi) r(f_{k+1}) = \sum_{k=0}^{n-1} A_k(\pi) , \tag{12}$$

where $Q_0(\pi) = I$.

In general, the equation (12) does not always converge as $n\to\infty$. Consequently it is difficult to compare the two policies from the viewpoint of the total expected return. In this case, the most standard optimal criterion is the average rate of gain $u_n = n^{-1}\cdot V_n(\pi)$ for the first epochs using policy π.

Now let

$$B_n(\pi) = A_n(\pi) - A_{n-1}(\pi) .$$

Then

$$A_n(\pi) = B_0(\pi) + B_1(\pi) + \cdots + B_n(\pi) ,$$

where $A_0(\pi) = B_0(\pi)$.

Thus, the average rate of gain u_n is given by

$$u_n(\pi) = n^{-1}V_n(\pi) = \frac{1}{n}\sum_{k=0}^{n-1} A_k(\pi) .$$

In case of infinite horizon, the limiting value of the average rate of ggin

$$\lim_{n\to\infty} u_n(\pi) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} A_k(\pi) = u_\infty(\pi)$$

expresses a Cesaro sum of the 1st degree, if u_∞ exists. And the existence of the average rate of gain u_∞ shows that the series $\sum B_n(\pi)$ is Cesaro summable. When $A_n(\pi) \geqq 0$, the following Abelian and Tauberian Theorem is useful to the summable condition.

Theorem 4. The average rate of gain $u_n(\pi)$ exists if and only if

$$\lim_{\beta \to 1-0} (1-\beta) V_\beta(\pi) = u_\infty(\pi) , \tag{13}$$

where $A_n(\pi) \geqq 0$.

It is noted that, for a stationary policy $\pi = f$,

$$u_\infty(f) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} A_k(f) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} Q^k(f) r(f)$$

$$= Q^*(f) r(f) ,$$

using

$$A_n(\pi) = Q^n(f) r(f) ,$$

and

$$\lim_{n\to\infty} \frac{I + Q(f) + \cdots + Q^n(f)}{n} = Q^*(f) . \tag{14}$$

Then the above condition (13) is reduced to

$$\lim_{\beta \to 1-0} (1-\beta) V_\beta(f) = Q^*(f) r(f) . \tag{15}$$

This fact exactly shows the existence of an average rate of gain.

Theorem 4 supposes $A_n(\pi) \geqq 0$, but generally it is not true. In general, the necessary and sufficient condition that the series $\sum B_n(\pi)$ is Cesaro summable is obtained by applying the following modified Hardy's Theorem.

Theorem 5. The average rate of gain with no discounting exists if and only if the series $\sum c_n(\pi)$ is Ceasro summable of 0th degree, where the sequence $\{c_n(\pi)\}$ satisfies the following recursive relations.

$$B_n(\pi) = (n+1)(C_n(\pi) - C_{n+1}(\pi)) \qquad (n = 0, 1, \cdots) \tag{16}$$

In other words, the condition is that the limit

$$\lim_{n\to\infty} D_n(\pi) \tag{17}$$

exists, where $D_n(\pi) = \sum_{k=0}^{n} C_k(\pi)$.

If the consideration is restricted in a class of stationary policies, there exists an Abelian Sum of $\sum B_n(f)$ because the series $\sum B_n(f)$ is Cesaro summable. Then

$$\lim_{\beta \to 1-0} \sum B_n(f) \beta^n = u_n(f) ,$$

that is to say, the average rate of gain is given by

$$u_n(f) = \lim_{\beta \to 1-0} \sum (A_n(f) - A_{n-1}(f))\beta^n$$
$$= \lim_{\beta \to 1-0} \sum Q^{n-1}(f)(Q(f)-I)r(f)\beta^n . \tag{18}$$

Conversely, if the series $\sum B_n(\pi)$ has an Abelian Sum, the series $\sum B_n(\pi)$ is Cesaro summable of the 1st degree when

$$A_n(\pi) = B_0(\pi) + B_1(\pi) + \cdots + B_n(\pi) = Q_n(\pi) r(f_{n+1}) > M \quad (n=0, 1, 2, \cdots) ,$$

where M is an arbitrary positive constant.

That is, the average rate of gain $u_n(\pi) = \lim_{n \to \infty} n^{-1} \sum_{k=0}^{n-1} A_k(\pi)$ esists. Futhermore, by Chapman's Theorem [6], if the average rate of gain exists, then $B_n(f) = o(n)$ or

$$\lim_{n \to \infty} \frac{Q^{n-1}(f)(Q(f)-I)\, r(f)}{n} = 0 . \tag{19}$$

5. Veinott's criterion

One of the standard optimal criterions for the problem with no discounting is the one which selects the policy π maximizing $n^{-1}V_n(\pi)$ as $n \to \infty$. As Denardo and Miller [3] pointed out, this criterion is useful to the stationary policy. It is noted, however, that the criterion is rather unselective when the average depends only on the tail of the return stream and not on the return in the first millennium. In this case, it is recomended to use a policy π' such that $\liminf [V_n(\pi') - V_n(\pi)] \geqq 0$ for any policy π instead of maximizing $n^{-1}V_n(\pi)$, and π' is called an optimal policy. Unfortunately, it is known that there are some examples where two policies π and π' may have $V_n(\pi') - V_n(\pi)$ oscillating around zero as $n \to \infty$. Veinott [2] gives the criterion using Cesaro summation to damp down such oscillations. That is, π^* is called Veinott's optimal policy if

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \{V_k(\pi^*) - V_k(\pi)\} \geqq 0 , \tag{20}$$

for any policy π, where

$$V_N(\pi) = \sum_{k=0}^{N-1} A_k(\pi) = \sum_{k=0}^{N-1} Q_k(\pi) r(f_{k+1}) . \tag{21}$$

In this paper, let discuss the meaning of Veinott's criterion, using not the concept of Cesaro sum in the divergent series, but the integral theory. It is noted that the expected return $V_\pi(n) = V_n(\pi)$, for epochs 1 through n using policy π, is a step function of n as shown in Figure 1.

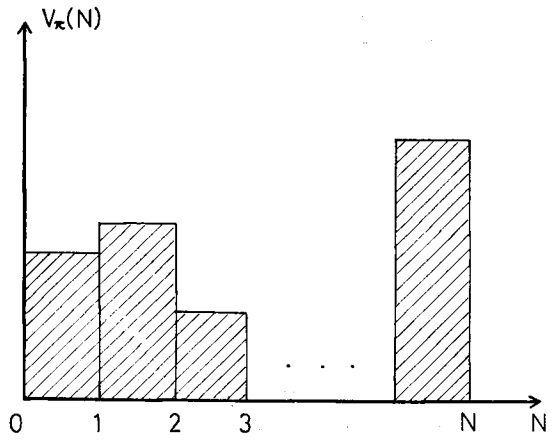

Fig. 1.

Let $I_0, I_1, I_2, \cdots$ be the time intervals between the two successive decision points respectively. Then $\{I_i\}$ is the disjoint measurable set and Lebesgue measure of each interval is one.

Let $E=\sum I_i$, then, from σ-additiveity,

$$\mu E = \mu I_0+\mu I_1+\mu I_2+ \cdots ,$$

where μ denotes Lebesgue measure.

A step function $V_\pi(n)$ is measurable on the set E and $V_\pi(n)$ is decomposed as

$$V_\pi(N) = V_\pi^+(n)-V_\pi^-(n) ,$$

where $$V_\pi^+(n) = \max [V_\pi(n), 0] \geqq 0 ,$$

and $$V_\pi^-(n) = \max [-V_\pi(n), 0] \geqq 0 .$$

The integral of $V_\pi(n)$ on the set E is described as follows;

$$\int_E V_\pi(n)\, d\mu = \sum_{I_i \in E} V_i(\pi)\, \mu(I_i) = \int_E V_\pi^+(n)\, d\mu - \int_E V_\pi^-(n)\, d\mu$$
$$= \sum_{I_i \in E} [V_i^+(\pi)-V_i^-(\pi)] = S(\pi) .$$

In this case, $S(\pi)$ expresses the area of the shaded portion in Figure 1.

Consider the above integral for two policies π and π^* as follows;

$$S(\pi^*)-S(\pi) = \sum_{i=1}^{N} [(V_i^+(\pi^*)-V_i^-(\pi^*))-(V_i^+(\pi)-V_i^-(\pi))]$$
$$= \sum_{i=1}^{N} [V_i(\pi^*)-V_i(\pi)] = \int_E (V_{\pi^*}(n)-V_\pi(n))\, d\mu .$$

Then Veinott's criterion is given by

$$\liminf_{N\to\infty} \frac{1}{N} \int_E (V_{\pi^*}(n) - V_{\pi}(n))\, d\mu \geqq 0 .$$

That is to say, the Veinott's criterion is considered as the average of the difference of the area $S(\pi) - S(\pi^*)$ intuitively.

6. New criterion

In this section, a new optimality criterion applying the concept of a convergence-speed appeared in the theory of infinite series is introduced. For two divergent series, for example $\sum 2^{n-1}$ and $\sum 1$, let

$$S_n = 1+2+ \cdots +2^{n-1} = 2^n - 1 ,$$
$$S'_n = 1+1+ \cdots +1 = n .$$

Since

$$\frac{S_n}{S'_n} \to \infty \quad \text{or} \quad \frac{S'_n}{S_n} \to 0 ,$$

S_n is far larger than S'_n. Thus it is said that the series $\sum 2^{n-1}$ diverges faster than the series $\sum 1$.

Apply the fact described above to a Markovian decision process with no discounting and introduce the new optimality criterion as follows;

Definition 3. Let $\sum_{n=0}^{\infty} A_n(\pi)$ and $\sum_{n=0}^{\infty} A_n(\pi^*)$ be the total expected returns with no discounting corresponding to policy π and policy π^* respectively.

$$V^N(\pi) = \sum_{n=0}^{N-1} A_n(\pi) ,$$

and

$$V^N(\pi^*) = \sum_{n=0}^{N-1} A_n(\pi^*) .$$

If $\dfrac{V^N(\pi)}{V^N(\pi^*)} \to \infty$, then the policy π is called a better policy than policy π^*.

Definition 4. If $\dfrac{V^N(\pi^*)}{V^N(\pi)} \to \infty$ for any policy π, then the policy π^* is called optimal.

Definition 5. If $\dfrac{V^N(\pi^*)}{V^N(\pi)} \to l \neq 0$ for two policies π and π^*, then these two policies are defined to be of the same degree.

The optimal policy is able to be found by the above definitions, but it is complicated to compute $V^N(\pi)$ the vector of expected total returns for epochs 1 through N using π. Therefore, the following Theorem 6 is useful as the easy method of computation.

Theorem 6. If $\frac{A_n(\pi)}{A_n(\pi^*)} \to +\infty$ for $\sum A_n(\pi)$ and $\sum A_n(\pi^*)$, then the policy π is better than policy π^*.

If $\frac{A_n(\pi)}{A_n(\pi^*)} \to +l \neq 0$, then the policy π and the policy π^* are of same degree.

Proof. Assume that

$$\frac{A_n(\pi)}{A_n(\pi^*)} \to l' \quad (n>m) ,$$

then for all n corresponding to an any positive vector ε there exists a positive integer m such as

$$(l'-\varepsilon)_i [A_n(\pi^*)]_i < [A_n(\pi)]_i < (l'+\varepsilon)_i [A_n(\pi^*)]_i .$$

Therefore, since the series $\sum A_n(\pi^*)$ diverges to $+\infty$,

$$\begin{aligned} \frac{V^n(\pi)}{V^n(\pi^*)} &= \frac{V^m(\pi)+A_{m+1}(\pi)+\cdots+A_n(\pi)}{V^m(\pi^*)+A_{m+1}(\pi^*)+\cdots+A_n(\pi^*)} \\ &< \frac{V^m(\pi)+(l'+\varepsilon)(A_{m+1}(\pi^*)+\cdots+A_n(\pi^*))}{V^m(\pi^*)+A_{m+1}(\pi^*)+\cdots+A_n(\pi^*)} \\ &< l'+2\varepsilon \qquad n>m' \ (>m) \end{aligned}$$

Similarly,

$$\frac{V^n(\pi)}{V^n(\pi^*)} > l'-2\varepsilon \qquad n<m' \ (<m) .$$

Then

$$\frac{V^n(\pi)}{V^n(\pi^*)} \to +l' .$$

Accordingly, if $A_n(\pi)/A_n(\pi^*) \to 0$, that is to say, $l'=0$, then $V^n(\pi)/V^n(\pi^*) \to 0$.

On the other hand, when $l \neq 0$, let $l'=1/l$, then

$$V^n(\pi^*)/V^n(\pi) \to l ,$$

that is to say, if

$$\frac{A_n(\pi)}{A_n(\pi^*)} \to l' ,$$

then

$$\frac{V^n(\pi)}{V^n(\pi^*)} \to l' ,$$

i.e.

$$\frac{V^n(\pi^*)}{V^n(\pi)} \to l .$$

The proof is complete.

Generally speaking, the total expected return with no discounting becomes larger with epochs n. When the total expected return $\sum A_n(\pi)$ corresponding to the policy π diverges to ∞, the relation between the case where a discounting factor β is introduced and the case where $\beta \to 1$, is given by the following Theorem 7.

Theorem 7. Consider two policies π_1 and π_2, and assume that $A_n(\pi_1) > 0$, $\sum A_n(\pi_1) = \infty$ for π_1. If the two policies π_1 and π_2 are of same degree, that is

$$\lim_{n\to\infty} \frac{A_n(\pi_2)}{A_n(\pi_1)} = A < \infty ,$$

then

$$\lim_{\beta\to 1} \frac{V_\beta(\pi_2)}{V_\beta(\pi_1)} = \lim_{\beta\to 1} \frac{\sum A_n(\pi_2)\beta^n}{\sum A_n(\pi_1)\beta^n} = A .$$

Proof. Without loss of generality, suppose that $A=1$. From the assumption that the policy π_1 and the policy π_2 are of same degree,

$$1-\varepsilon < \frac{A_n(\pi_2)}{A_n(\pi_1)} < 1+\varepsilon$$

for $n < N = N(\varepsilon)$.

On the other hand,

$$\begin{aligned} V_\beta(\pi_2) &= \sum_{n=0}^{N} A_n(\pi_2)\beta^n + \sum_{n=N+1}^{\infty} A_n(\pi_2)\beta^n \\ &\leqq (1+\varepsilon) V_\beta(\pi_1) + \sum_{n=0}^{N} |A_n(\pi_2)|\beta^n , \end{aligned}$$

and

$$V_\beta(\pi_2) \geqq (1-\varepsilon) V_\beta(\pi_1) - \sum_{n=0}^{N} A_n(\pi_1)\beta^n - \sum_{n=0}^{N} |A_n(\pi_2)|\beta^n .$$

Since

$$\lim_{\beta\to 1} V_\beta(\pi_1) \geqq \lim_{\beta\to 1} \sum_{n=0}^{N} A_n(\pi_1)\beta^n \quad (=\text{finite}) ,$$

and

$$V_\beta(\pi_1) \to \infty \qquad (\beta \to 1) ,$$

$$\limsup_{\beta\to 1} \frac{V_\beta(\pi_2)}{V_\beta(\pi_1)} \leqq 1+\varepsilon ,$$

$$\liminf_{\beta\to 1} \frac{V_\beta(\pi_2)}{V_\beta(\pi_1)} \geqq 1-\varepsilon .$$

Therefore

$$\lim \frac{V_\beta(\pi_1)}{V_\beta(\pi_2)} = 1$$

is introduced and the proof is complete.

In the field of Markovian decision problem, nothing is more important than to consider the existence of a stationary optimal policy. Assume that $A_n(\pi)>0$. Let F^∞ be a class of the stationary policies. The set F^∞ is a finite set including the $F_1F_2\cdots F_s$ elements, and $F^\infty \subset F$. Then the following Theorem 8 is introduced.

Theorem 8. There exists at least an optimal stationary policy.

Proof. As the number of stationary policies are finite, two policies f^∞ and g^∞ are of the same degree or policy f^∞ is better policy than g^∞ (or g^∞ is better than f^∞) by Definition 3 for some stationary policy f^∞ and any stationary plicy $g^\infty \in F^\infty$.

If f^∞ is better than g^∞, then policy f^∞ is the unique optimal stationary policy since policy g^∞ is selected arbitrarily from the set F^∞. If f^∞ and g^∞ are of same degree, then f^∞ and g^∞ are optimal stationary policies. If one of the stationary policies g^∞ is better than f^∞, consideration of the different policy g'^∞ from g^∞ such as

$$g'^\infty \in F^\infty - f^\infty$$

and the same discussion assures the existence of the optimal stationary policies since the set F^∞ is finite.

7. Concluding remarks

In the problem of a Markovian decision process, the vector of expected total returns for infinite horizon has been considered as the series of the decision point n and the optimal criterions to decide an optimal policy have been discussed by means of the thoery on the infinite series.

When a discount factor β is considered, the total expected return is regarded as the power series of β. And it has been shown that the criterion of l-optimal policy is introduced by considering the Abelian Sum as $\beta \to 1$. Moreover, the necessary and sufficient condition that the Abelian Sum might exist has been introduced. For the mean-optimal policy, Cesaro sum corresponds to it and the necessary and sufficient condition for the average rate of gain to exist has been given. Commonly, Veinott's criterion is considered as the application of Cesaro summation, but in this paper, it has been discussed in the point of the theory on integral.

At last, in addition to the optimal criterions described above, a new criterion to apply the divergence-speed of the infinite series has been introduced. For a problem in which the total expected returns become infinity with n, this new criterion is valuable to construct an intuitive optimal policy theoretically. And

an algorithm to find the optimal policy is not exactly given, but Theorem 6 and Theorem 7 may give an easy procedure to find the optimal policy.

References

1) D. Blackwell; Discrete Dynamic Programming, Ann. Math, Statist., 33 (1962)
2) A.F. Veinott; On the Finding Optimal Policies in Discrete Dynamic Programming with No Discounting, Ann. Math. Statist., 37 (1966)
3) E.V. Denardo and B.L. Miller; An Optimality Condition for Discrete Dynamic Programming with No Discounting, Ann. Math. Statist., 39 (1968)
4) R.A. Howard; Dynamic Programming and Markov Processes, M.I.T. Press, (1960)
5) S.A. Lippman; On the Set of Optimal Policies in Discrete Dynamic Programming, J. Math. Anal. Appl., 24 (1968)
6) G.H. Hardy; Divergent Series, Oxford, (1949)