Journal of the Operations Research
Society of Japan
Vol. 39, No. 4, December 1996

MARKOV DECISION PROCESSES WITH RANDOM HORIZON

Tetsuo Iida Masao Mori
Tokyo Institute of Technology

(Received April 27, 1995; Revised September 11, 1995)

Abstract In this paper we formulate Markov Decision Processes with Random Horizon (MDPRH). We show the optimality equation for the MDPRH, however there may not exist optimal stationary strategies, or ϵ-optimal stationary strategies for the processes. When the MDPRH has the probability distribution for the planning horizon with infinite support, we show Turnpike Planning Horizon Theorem. Then we evaluate rolling strategies and develop an algorithm obtaining an optimal first stage decision. Finally, some numerical experiments on a simple inventory model are done to understand the phenomena.

1. Introduction

A multiperiod optimization problem is often modeled as an infinite horizon problem when its horizon is long sufficiently. We do not necessarily know the horizon of the problem in advance since we can not predict the future precisely. For example, we imagine vaguely that a drastic change of a project may occur some day, and we only believe when its change will occur under a certain probability distribution. Thus it is not appropriate that we simply model the problem as an infinite or a fixed finite horizon case.

A change of planning horizon may cause a remarkable change of optimal strategy, and the total reward may differ much. Therefore it is necessary to make a decision considering the probability of the time at which the project will end.

We will deal with these problems and formulate them using Markov Decision Processes in which probability distributions for the planning horizon are given in advance, that is, MDP with Random Horizon(MDPRH). In this paper we will consider them in the framework of non-homogeneous MDPs.

One typical example of this MDPRH is the MDP with a geometrically distributed planning horizon, which is known to be equivalent to an ordinary Discounted MDP(Ross[6]).

As for the Discounted MDP, numerous researches have been made(Puterman[5]) until now. There are many interpretations for this discount rate, however we can also add an interpretation that discount rate represents an evaluation of uncertainty expected to be happened in the future. Also there is the research for the type of MDP which has variable discount rates (White[9]).

In the MDPRH there may not exist an optimal stationary(nonrandomized) strategy, and perhaps, there may not exist even an $\epsilon-$optimal randomized stationary strategy(see Appendix). The purpose of this paper is to analyse an optimal strategy for the MDPRH and propose the algorithm to obtain it numerically by Turnpike Planning Horizon approach. When the support of the probability distribution for the planning horizon is finite, we can easily get an optimal strategy by solving the corresponding optimality equation for the MDPRH. On the other hand, when the support of the probability distribution for the

planning horizon is infinite, it is much more difficult to solve the problem, because in this case we can not solve the optimality equation by backward induction. So we adopt a rolling horizon strategy to obtain an optimal strategy, that is, first we obtain the Turnpike Planning Horizon for MDPRH and solve the problem under its horizon. Shapiro[8]shows the existence of the Turnpike Planning Horizon for the homogeneous Discounted MDP.

This paper is related to the researches of Bean and Smith[2], Sethi and Bhaskaran[7], and Bes and Sethi[3]. They analyse their problems respectively by means of Forecast and Decision Horizon approach similar to Turnpike Planning Horizon one. Bean and Smith[2] treats deterministic decision problems and Sethi and Bhaskaran[7] and Bes and Sethi[3] consider the discounted MDPs. In addition Alden and Smith[1] discusses about the Rolling Horizon Procedure which leads non-optimal decisions sequentially by solving a fixed finite horizon problem iteratively but they also obtain its error bound. Hopp, Bean and Smith[4] considers the condition for the existence of an optimal strategy for the non-homogeneous non-Discounted MDP under a weak ergodicity assumption.

In section 2 we first formulate the MDPRH precisely and derive the optimality equation for the MDPRH. In section 3 we describe about nature of an optimal strategy in the case that the support of the probability distribution for the planning horizon is infinite. In section 4 we give an algorithm for solving the MDPRH based on the nature derived in section 3. In Section 5 some concluding remarks are stated.

2. Definition

Let $(\Omega, \mathcal{F}, \mathcal{P})$ denote the underlying probability space. Let $Z = \{0, 1, 2, \cdots\}$be the set of nonnegative integers. Consider a discrete-time non-homogeneous Markov Decision Model with

(i) countable state space S,

(ii) measurable action space A endowed with σ–*field* $\mathcal{A}$ containing all one-point subsets of A,

(iii) sets of action $A(s)$ available at $s \in S$, where $A(s)$ is a element of $\mathcal{A}$,

(iv) transition probabilities $\{p_t(j|i,a)\}$ at stage t , $t \in Z$, where for each $i, j \in S$, $p_t(j|i,a)$ is nonnegative and measurable in a, and for each $i \in S$, $a \in A(s)$, $\sum_{j \in S} p_t(j|i,a) = 1$, $t \in Z$,

(v) sets of reward functions $\{r_t(i,a)\}$ at stage t, $t \in Z$, where the function $r_t(i,a)$ is measurable in a,

(vi) sets of salvage cost functions $\{c_t(i,a)\}$ when the project end at stage t, $t \in Z$, where the function $c_t(i,a)$ is measurable in a. The salvage cost is the incurred cost to stop the project and may depend on the state and action at that stage.

Assumption 2.1 *For each stage, reward functions and salvage cost functions are assumed to be bounded, that is,*

$$|r_t(s,a)| \leq R < \infty \ , \quad |c_t(s,a)| \leq C < \infty$$

Let a function $u_t : S \to A$, $t \in Z$ be a decision function with $u_t(s_t) \in A(s_t)$. The sequence $u = (u_t, t \in Z)$ is called a strategy. Let Π denote the set of all strategies. We also use the notation ${}_n u = (u_0, u_1, \ldots, u_{n-1})$ to represent first n decisions in u.

In this model, we also set,

(vii) a probability distribution f_t with which the project end at stage t, $t \in Z$.

We consider an absorbing state s' representing the end state of project and let $S' = S \cup \{s'\}$. Then we add next three to the above (iii), (iv) and (v),

(iii)' $A(s') = \{a'\}$,
(iv)' *for any* $j \in S$, *all* $t \in Z$, $p_t(j|s', a') = 0$, $p_t(s'|s', a') = 1$,
(v)' *for all* $t \in Z$, $r_t(s', a') = 0$.

Let $H_t = S \times (A' \times S')^t$ be the space of histories up to the stage $t \in \bar{Z} \cup \{\infty\}$, where $A' = A \cup \{a'\}$. Clearly once a strategy $u \in \Pi$ and initial state s are specified, transition probabilities are determined completely. Accordingly a probability measure P_s^u is induced. We denote the corresponding expectation operator by E_s^u.

Let X_t denote the state of process at stage t, A_t denote the action taken at stage t, and N denote the random planning horizon with distribution f_t. Then the reward $R_t(X_t, A_t)$ which we get at stage t is defined as follows,

$$R_t(X_t, A_t) = \begin{cases} r_t(X_t, A_t) & \text{when } X_t \in S \text{ and } t < N \\ c_t(X_t, A_t) & \text{when } X_t \in S \text{ and } t = N \\ 0 & \text{when } X_t \in S' \end{cases} . \tag{2.1}$$

For a fixed n we consider the n-horizon problem. When the process starts with an initial state s under a strategy u, the expected total reward for the $n-$horizon problem is given by

$$V(s, u, n) = E_s^u \sum_{t=0}^{n} R_t(X_t, A_t). \tag{2.2}$$

Note that the expected reward $V(s, u, n)$ depends only on the first n decisions ${}_n u$ in each u.

Now we can describe the $n-$horizon and $N-$horizon optimal decision problem. The $n-$horizon problem is defined as

$$\sup_{u \in \Pi} \left\{ V(s, u, n) = E_s^u \sum_{t=0}^{n} R_t(X_t, A_t) \right\}, \quad \text{for each } s. \tag{2.3}$$

A strategy $u^*(n) \in \Pi$ is called an optimal strategy for the $n-$horizon problem if for each $s \in S$, $V(s, u^*(n), n) = \sup_{u \in \Pi} V(s, u, n)$. It should be also noted that an optimal strategy for $n-$horizon problem depends only on the first n decisions in each u.

On the other hand, for the random $N-$horizon problem, we set

$$V(s, u) = E_s^u \sum_{t=0}^{N} R_t(X_t, A_t). \tag{2.4}$$

Similarly a strategy $u^* \in \Pi$ is called an optimal strategy for the random $N-$horizon problem if for each $s \in S$, $V(s, u^*) = \sup_{u \in \Pi} V(s, u)$.

Let ϵ be an arbitrary nonnegative constant. Then a strategy $u_\epsilon^*(n)$ is called $\epsilon-$optimal strategy for the $n-$horizon problem if for each $s \in S$,

$$V(s, u_\epsilon^*(n), n) \geq V(s, u^*(n), n) - \epsilon. \tag{2.5}$$

Now we consider the optimality equation for the MDPRH. Let α_t be a probability which the project is still continuing at stage $(t+1)$ under condition that it has continued until stage t, that is,

$$\alpha_t = \frac{1 - \sum_{k=1}^{t} f_k}{1 - \sum_{k=1}^{t-1} f_k}. \tag{2.6}$$

When the process is in state i and action a is used at stage t, the expected reward we get is

$$b_t(i,a) = \alpha_t r_t(i,a) + (1-\alpha_t)c_t(i,a). \tag{2.7}$$

Let $v_t^*(i)$ denote a maximal value which we can get after the stage t when the process is in state i at stage t. Therefore we can get the optimality equation as follows,

$$v_t^*(i) = \max_{a\in A}\left\{ b_t(i,a) + \alpha_t \sum_{j\in S} p_t(j|i,a)v_{t+1}^*(j)\right\}. \tag{2.8}$$

When the support of the probability distribution for the plannig horizon is finite, we can easily obtain the solution of the problem, analogously as in an ordinary finite horizon problem, by applying the backward induction method to the optimality equation(2.8) with setting

$$v_n^*(i) = \max_{a\in A(i)} c_n(i,a)\ , \quad \text{for all } i\in S, \tag{2.9}$$

where n is a maximal value of the support of $\{f_t, t\in Z\}$.

3. Optimal Strategies when the support of the probability distribution for the planning horizon is infinite

In this section we discuss the MDPRHs which have the infinite support of the probability distribution for the planning horizon. When the support is finite, we can obtain an optimal strategy by applying the backward induction method to the optimality equation. However, we can not use such a way in the case that the support is infinite. Therefore we discuss this problem based on the idea that if the optimal strategies for the finite horizon problem approach a particular strategy for the infinite support problem, we will consider that strategy as the optimal one. Works of Hopp, Bean and Smith[4], Bes and Sethi[3] are based on this idea, too.

In order to discuss above, we now define a metric topology on the set of all strategies Π. The metric ρ below is the same one which Bean and Smith[2] uses.

$$\rho(u,u') = \sum_{n=1}^{\infty} 2^{-n}\phi_n(u,u'),$$

where

$$\phi_n(u,u') = \begin{cases} 1 & u_n'(x) \neq u_n(x) \quad \exists x\in S \\ 0 & u_n'(x) = u_n(x) \quad \forall x\in S \end{cases}$$

Now we introduce the optimality criterion of periodic forecast horizon (PFH) optimality which is defined in Hopp, Bean and Smith[4].

Definition 3.1 *A strategy $u^\natural \in \Pi$ is periodic forecast horizon (PFH) optimal if for some subsequence of the integers, $\{N_m\}_{m=1}^{\infty}$, $u^*(N_m) \to u^\natural$ in the ρ metric as $m\to\infty$.*

Proposition 3.1 *(Π,ρ) is a metric space. Also, if $\rho(u,u') < \epsilon < 1$, for all $n \leq -\log_2\epsilon$ $u_n' = u_n$.*

proof. See Bean and Smith[2]. □

Assumption 3.1 *(Π,ρ) is a compact metric space.*

Let $\Pi_t = \{u_t\}$ for all $t \in Z$. We define the discrete topology $\rho_t(u_t, u_t') = 2^{-t}\phi_t(u_t, u_t')$ on them. Then the theorem below holds.

Theorem 3.2 *Π is compact if and only if $\forall t \in Z$, Π_t are finite sets. If Π is compact, then each cylinder subset of Π is compact.*

proof. See [3] and its references. □

When S and A are finite sets, a compactness of Π is ensured. Let $u^*(n) \in \Pi$ be an optimal strategy for n-horizon problem and $\Pi^\sharp$ be a set of cluster points of all the sequences $\{u^*(n)\}$, that is, a set of PFH-optimal strategies, and let ${}_t\Pi^\sharp = \{{}_t u | u \in \Pi^\sharp\}$ $t \in Z$. Note that since $V(s, u, n)$ is continuous in u and Π is compact, $\Pi^\sharp$ is a nonempty set.

From the definition of $V(s, u)$, we have the following proposition.

Proposition 3.3 *When the expectation of the planning horizon is finite, the total expected reward is finite.*

proof. Since the expectation of the planning horizon, $E[N]$, is finite,

$$\sum_{t=1}^{\infty} t f_t < +\infty. \tag{3.1}$$

Then for $\forall u \in \Pi$,

$$\begin{aligned} V(s,u,N) &= E_s^u\left[\sum_{t=0}^{N} R_t(X_t, A_t)\right] \\ &= \sum_{t=1}^{\infty} E_s^u\left[r_t(X_t, A_t) | N > t\right] P\left[N > t\right] + \sum_{t=1}^{\infty} E_s^u\left[c_t(X_t, A_t) | N = t\right] P\left[N = t\right] \\ &< \max\{R, C\} \sum_{t=1}^{\infty}\left(1 - \sum_{k=1}^{t-1} f_k\right) \\ &= \max\{R, C\} \sum_{t=1}^{\infty} t f_t \end{aligned}$$

Thus from (3.1), $V(s, u, N) < +\infty$ □

Remark 3.4 *When for all i , a , $r_t(i, a) > 0$, the converse of Prop.3.3 holds.*

Hereafter we assume the following.

Assumption 3.2 *The expectation of the planning horizon is finite.*

Now we discuss the existence of optimal strategy for the MDPRH with infinite support. Before the discussion we show the following lemma.

Lemma 3.5 *$V(s, u)$ is continuous in $u \in \Pi$.*

proof. For any $\epsilon > 0$, there exists M, such that $\max\{R, C\} \sum_{t=M+1}^{\infty} \prod_{k=1}^{t-1} \alpha_k < \frac{\epsilon}{2}$. Therefore we get a δ such that $M \leq -\log_2 \delta$. Then for any $u' \in \Pi$ such that $\rho(u, u') < \delta$,

$$\begin{aligned} &\left|V(s,u) - V(s,u')\right| \\ &= \left|E_s^u\left[\sum_{t=1}^{N} R_t(X_t, A_t)\right] - E_s^{u'}\left[\sum_{t=1}^{N} R_t(X_t, A_t)\right]\right| \\ &= \left|\sum_{t=M+1}^{\infty} E_s^u\left[r_t(X_t, A_t) | N > t\right] P\left[N > t\right] + \sum_{t=M+1}^{\infty} E_s^u\left[c_t(X_t, A_t) | N = t\right] P\left[N = t\right]\right. \end{aligned}$$

$$\left| - \sum_{t=M+1}^{\infty} E_s^{u'} [r_t(X_t, A_t)|N > t] P[N > t] - \sum_{t=M+1}^{\infty} E_s^{u'} [c_t(X_t, A_t)|N = t] P[N = t] \right|$$

$$\leq 2\max\{R, C\} \sum_{t=M+1}^{\infty} \prod_{k=1}^{t-1} \alpha_k < \epsilon.$$

□

Theorem 3.6 (existence)
Under assumptions 2.1,3.1 and 3.2, there exists a PFH-optimal strategy for the MDPRH.

proof. Since Π is compact, $V(s,u)$ is uniformly continuous on Π. Thus there exists an strategy $u^* \in \Pi$ such that $V(s,u^*) = \max_{v\in\Pi} V(s,v)$ Therefore there exists a PFH-optimal strategy for the MDPRH $u^* \in \Pi$. □

Next we show the two lemmas, which lead to Turnpike Planning Horizon Theorem.

Lemma 3.7 $\lim_{n\to\infty} \max_{u\in\Pi} V(s,u,n) = \max_{u\in\Pi} V(s,u)$

proof. Set $K_n = \max\{R, C\} \sum_{t=n+1}^{\infty} \prod_{k=1}^{t-1} \alpha_k$, and we have

$$\max_{u\in\Pi} V(s,u) - K_n \leq \max_{u\in\Pi} V(s,u,n) \leq \max_{u\in\Pi} V(s,u) + K_n$$

Thus since $K_n \to 0$ *as* $n \to \infty$,

$$\lim_{n\to\infty} \max_{u\in\Pi} V(s,u,n) = \max_{u\in\Pi} V(s,u)$$

□

Let Π^* denotes a set of all optimal strategies $u^* \in \Pi$ for the MDPRH.

Lemma 3.8 $\Pi^\sharp \subset \Pi^*$

proof. Let $u^* \in \Pi^\sharp$. From the definition there exists a sequence of strategies $\{u^*(m(j))\}_{j\in Z}$ such that $\lim_{j\to\infty} u^*(m(j)) = u^*$. Thus $\lim_{j\to\infty} V(s, u^*(m(j)), m(j)) = V(s,u^*)$. since from lemma 3.7 $V(s,u^*) = \max_{u\in\Pi} V(s,u)$, $u^* \in \Pi^*$.

□

There may not necessarily exists a stationary deterministic strategy or stationary randomized strategy for the MDPRH (see Appendix). An optimal strategy we wish to know may be non-stationary, so it is difficult to get it directly. Therefore we begin with showing that a theorem similar to Turnpike Planning Horizon Theorem which Shapiro[8] shows for the homogeneous Discounted MDP holds for this MDPRH. We introduce the following two notations,

$F = \{_1u : u \in \Pi^*\}$: a set of optimal decisions at the first stage for the random $N-$horizon problem, and

$F(n) = \{_1u : u = u^*(n)\}$: a set of optimal decisions at the first stage for the n-horizon problem.

Then the following theorem holds.

Theorem 3.9 (Turnpike Planning Horizon Theorem) *There exists some L such that for any $n \geq L$, $F(n) \subset F$.*

proof. Assume as the contrary that there does not exist such a number L. Then there exists an integer M_1 such that the first decision of some optimal strategy for the M_1 horizon problem is not contained in F, and there exists an integer $M_2(> M_1)$ similarly. So we obtain a sequence of strategies $\{u^*(M_i)\}$ such that $_1u^*(M_i) \notin F$ for all i. Since Π is compact, there

exists a subsequence $\{u^*(m(M_i))\}$ such that its limit is $u^{**} \in \Pi$. Thus for sufficient large $m(M_i)^*$, $\rho(u^{**}, u^*(m(M_i))) < \epsilon$, so ${}_1u^{**} = {}_1u^*(m(M_i))$. Therefore ${}_1u^{**} \notin F$. On the other hand, from the definition $u^{**} \in \tilde{\Pi}^*$, and from the lemma 3.8 $u^{**} \in \Pi^*$. Thus ${}_1u^{**} \in F$, which is a contradiction. □

From the above theorem we can make a first optimal decision by solving the sufficient large n-horizon problem. It should be noted that there exits an optimal rolling strategy.

4. Algorithm for finding an optimal first decision

Although the Turnpike Planning Horizon Theorem in the above section states the existence of the turnpike horizon, the theorem shows no way for finding it. Hence in this section we investigate an algorithm for finding an optimal first decision or ϵ-optimal first decision. If we can find an optimal first decision, next we pay attention to the second stage, that is, we consider the second stage as the first stage, and then apply the same algorithm to it. By means of continuing this procedure at third, fourth, ... stage, we can find a sequence of optimal decisions one by one, that is, an optimal rolling strategy. Above procedures corresponds to identifing the PFH-optimal strategy gradually, that is, making the neighborhood of PFH-optimal strategy small.

Let $\hat{\Pi}^n$ denotes a set of strategies such that its first decision is not included in $F(n)$, that is, $\hat{\Pi}^n = \{u \in \Pi : {}_1u \notin F(n)\}$.

Then the following theorem holds.

Theorem 4.1 *For any $u \in \hat{\Pi}^n$, if u satisfies the following condition $(*)$,*

$$(4.1) \qquad (*) \quad \max_{u \in \Pi} V(s,u,n) - V(s,u,n) > 2\max\{R,C\} \sum_{t=n+1}^{\infty} \prod_{k=1}^{t} \alpha_k,$$

u is not optimal for the problem with infinite support.

proof. Let $u' \in \Pi$ be a strategy satisfying a condition $(*)$. Set $K_n = \max\{R,C\} \sum_{t=n+1}^{\infty} \prod_{k=1}^{t} \alpha_k$, then from the condition $(*)$,

$$(4.2) \qquad \max_{u \in \Pi} V(s,u,n) - V(s,u',n) > 2K_n,$$

and

$$(4.3) \qquad \max_{u \in \Pi} V(s,u,n) - K_n \leq \max_{u \in \Pi} V(s,u).$$

Therefore from (4.2) and (4.3)

$$\max_{u \in \Pi} V(s,u) - V(s,u',n) > K_n.$$

Thus $u' \in \Pi$ is not optimal. □

Remark 4.2 If $F(n)$ is singleton and a condition of theorem 4.1 holds for any $u \in \hat{\Pi}^n$, ${}_1u \in F(n)$ is an optimal first decision.

From the above theorem we can find a first decision which is not optimal and then remove it. In consequence we propose an algorithm which decreases the number of decisions possible to be optimal by iterating the above check. The following algorithm finds either an optimal first decision or an ϵ-optimal decision.

Algorithm 4.3

step 1. Set t = 1.

step 2. Let $\delta_t = \max\{R, C\} \sum_{n=t+1}^{\infty} \prod_{k=1}^{n} \alpha_k$.

step 3. $\forall a \in A$, compute $\xi_t^a = \max_{u \in \Pi} V(s,u,t) - V(s,u_a,t)$, where ${}_1u_a = a$. If $\xi_t^a > 2\delta_t$ and F^t is singleton, Stop. Its decision is an optimal first one.

step 4. If $\delta_t \leq \epsilon$, Stop. Its decision is an ϵ-optimal first one.

step 5. $t = t + 1$, and goto step 2.

Remark 4.4 From the theorem 3.6 the above algorithm stops in a finite number of steps.

Remark 4.5 If Π^* is singleton, the above algorithm can find an optimal first decision in a finite number of steps. It is discussed by Bes and Sethi[3] that Π^* is not rarely singleton.

As a numerical example, we consider an following inventory problem. An item has a lifetime distribution an account of its lifecycle or appearance of a new item. We consider that this distribution corresponds to the random horizon previously stated. We denote its distribution by $\{f_t\}$. When the project end, all remaining items may be sent back at a salvage cost per unit.

Here we assume that one-period demand, η_t, follows i.i.d.Poisson distribution. Let a_t denotes the amount of order. So the amount of stock satisfies a following relation,

$$s_t = s_{t-1} + a_t - \eta_t, \tag{4.4}$$

where the initial stock, s_0, is given. We assume that $\underline{S} \leq s_t \leq \bar{S}$, that is, an upper bound and a lower bound of the stock is given. The cost we consider are following,

$k_t(a_t)$: the order cost in the period t when a_t items are ordered,
$c_t(s_t)$: the holding cost in the period t when $s_t \geq 0$,
the backlogging cost in the period t when $s_t < 0$,
$r_t(x_t)$: the income in the period t when x_t items are sold,
where $x_t = \max\{\min\{\eta_t, s_{t-1}\}, 0\}$.

Accordingly the problem is to maximize the total expected reward :

$$\text{Maximize } E_{s_0}^a \left[\sum_{t=1}^{\infty} \{r_t(X_t) - k_t(A_t) - c_t(S_t)\} \right]. \tag{4.5}$$

Now we assume that the data are as follows,
$s_0 = 5$, $\underline{S} = -5$, $\bar{S} = 20$. The expected value of the demands in one-period is 7.

$$r_t(x) = \begin{cases} 10x & (x \geq 0) \\ 0 & (x < 0) \end{cases} \quad k_t(a) = \begin{cases} 8 + 5a & (a \geq 0) \\ 0 & (a < 0) \end{cases} \quad c_t(s) = \begin{cases} 2s & (s \geq 0) \\ 4s & (s < 0) \end{cases} \tag{4.6}$$

Then let the salvage cost per unit be 7.

We examine how the probability distribution for the planning horizon cause the change of the first optimal decisions. We use the following composite distribution of Poisson distributions,

$$P[N = t] = 0.5P_{\lambda_1}[N = t] + 0.5P_{\lambda_2}[N = t] \tag{4.7}$$

which enables us to arrange various combinations of values of the mean and coefficient of variation of the distribution by changing λ_1 and λ_2. We calculate the optimal decisions for amount of orders at the first stage and the Turnpike planning horizons for the cases in which means are $2, 3, 5, 10, 15, 20, 30, 50$, and coefficients of variation are $0.5, 0.6, 0.7, 0.8, 0.9, 1.0$. The results of caluculations are shown in Table 1.

Table 1. *Optimal First Decisions and Turnpike Planning Horizons*

CV / *Mean*	0.5	0.6	0.7	0.8	0.9	1.0
2	–	–	–	6	6	5
	–	–	–	10	9	10
	–	–	–	(1.25, 2.75)	(0.886, 3.11)	(0.589, 3.41)
3	–	8	8	7	6	5
	–	11	15	12	13	14
	–	(2.51, 3.49)	(1.81, 4.19)	(1.34, 4.66)	(0.929, 5.07)	(0.551, 5.50)
5	13	9	8	7	6	5
	16	20	19	21	22	20
	(3.88, 6.12)	(3.00, 7.00)	(2.31, 7.69)	(1.68, 8.32)	(1.09, 8.91)	(0.528, 9.47)
10	15	15	13	9	7	5
	29	32	33	34	34	34
	(6.13, 13.9)	(4.90, 15.1)	(3.76, 16.2)	(2.65, 17.3)	(1.57, 18.4)	(0.513, 19.5)
15	15	15	15	13	8	5
	39	42	45	47	47	47
	(8.58, 21.4)	(6.88, 23.1)	(5.24, 24.8)	(3.64, 26.4)	(2.07, 27.9)	(0.509, 29.5)
20	15	15	15	14	9	5
	49	52	56	62	62	60
	(11.1, 28.9)	(8.86, 31.1)	(6.73, 33.3)	(4.64, 35.4)	(2.56, 37.4)	(0.506, 39.5)
30	15	15	15	15	13	5
	69	73	77	81	87	85
	(16.0, 44.0)	(12.9, 47.1)	(9.73, 50.3)	(6.63, 53.4)	(3.56, 56.4)	(0.504, 59.5)
50	15	15	15	15	15	5
	107	113	119	125	132	132
	(26.0, 74.0)	(20.8, 79.2)	(15.7, 84.3)	(10.6, 89.4)	(5.56, 94.4)	(0.503, 99.5)

(upper) *optimal first decision*
(middle) *Turnpike Planning Horizon*
(lower) (λ_1, λ_2)

From Table 1. we can see two tendencies in this inventory problem, one is that quatity of order at the first stage increases as the mean horizon increases, and the other is that it decreases as the coefficient of variation increases. The numerical result shows the interesting behaviour that when the coefficient of variation is 1.0, the first optimal decisions are always 5. In this numerical example, when the coefficient of variation is 1.0, λ_1 becomes very small for each emans, which suggests the probability that the project will end soon is fairly large. Thus the first decision for amount of order is expected to become small. From theses results the optimal first decisions are considered to depend on the shape of the probability distribution for the planning horizon much.

5. Conclusion

In this paper we formulate Markov Decision Processes with Random Planning Horizon, which are described analogously as Markov Decision Processes with time varying discount rates. For the processes there may not exist optimal stationary strategies, so we evaluate rolling strategies derived by using the result of Turnpike Horizon Theorem. An algorithm obtaining an optimal first stage decision is proposed and some numerical experiments on a simple inventory model with random planning horizon are done to understand the phenomena. As a result of numerical experiments, we found certainly that the optimal first decisions depend on the shape of the probability distribution for the planning horizon.

Appendix

In the section 1 and 3 we state that in the MDPRH there may not exist an optimal stationary(nonrandomized) strategy, and perhaps, there may not exist even an ϵ–optimal randomized stationary strategy. We show an example as follows.

Example. Consider the homogeneous model with $S = \{1, 2\}$ and $A = \{a, b\}$. Let

$$p(1|s,a) = p(2|s,b) = 1, \qquad for \quad s = 1, 2.$$
$$r(1,a) = 1, \quad r(1,b) = r(2,a) = 0, \quad r(2,b) = 2.$$
$$c(s,x) = 0, \qquad for \quad s = 1, 2, \quad x = a, b.$$

We denote the probability distribution for the planning horizon $\{f_t\}$ as follows,

$$f_t = \begin{cases} \beta_1 & (when \quad t = 0) \\ (1-\beta_1)\beta_1 & (when \quad t = 1) \\ (1-\beta_1)^2(1-\beta_2)^{t-2}\beta_2 & (when \quad t \geq 2) \end{cases},$$

that is, the geometric distribution of which parameter changes to β_2 from β_1 at stage 2.

In this model, it is clear that action b is optimal at state 2. Thus there are two candidates for optimal deterministic stationary strategy as follows,

u' : keep your state (use action a at state 1 and action b at state 2),

u'' : move to state 2 and keep it (use only action b).

We shall examine an optimal randomized stationary strategy for this model. A randomized stationary strategy π is defined as

$$\pi(a|1) = \alpha, \qquad \pi(a|2) = \delta.$$

When $t \geq 2$, this model is equivalent to the MDP with discount rate $1 - \beta_2$. Therefore the expected reward $v_2(s, \pi), s = 1, 2$, is the unique soution of the system of the following linear equations,

$$\begin{cases} v_2(1,\pi) = & (1-\beta_2)\{\alpha(1 + v_2(1,\pi)) + (1-\alpha)v_2(2,\pi)\} \\ v_2(2,\pi) = & (1-\beta_2)\{\delta v_2(1,\pi) + (1-\delta)(2 + v_2(2,\pi))\} \end{cases}$$

Solving this equations, we have

$$\text{(a)} \quad \begin{pmatrix} v_2(1,\pi) \\ v_2(2,\pi) \end{pmatrix} = \frac{1-\beta_2}{\beta_2(1-\alpha(1-\beta_2)+\delta(1-\beta_2))} \times \begin{pmatrix} \alpha - \alpha(1-\delta)(1-\beta_2) + 2(1-\alpha)(1-\delta)(1-\beta_2) \\ \alpha\delta(1-\beta_2) + 2(1-\delta) - 2\alpha(1-\delta)(1-\beta_2) \end{pmatrix}.$$

Similarly,

$$\begin{cases} v_1(1,\pi) = & (1-\beta_1)\{\alpha(1 + v_2(1,\pi)) + (1-\alpha)v_2(2,\pi)\} \\ v_1(2,\pi) = & (1-\beta_1)\{\delta v_2(1,\pi) + (1-\delta)(2 + v_2(2,\pi))\} \end{cases}.$$

so that

$$\begin{aligned} v_0(1,\pi) &= (1-\beta_1)\{\alpha(1 + v_1(1,\pi)) + (1-\alpha)v_1(2,\pi)\} \\ &= \alpha(1-\beta_1) + \alpha^2(1-\beta_1)^2 + 2(1-\alpha)(1-\delta)(1-\beta_1)^2 \\ &\quad +(1-\beta_1)^2(\alpha^2 + (1-\alpha)\delta)v_2(1,\pi) + (1-\alpha)(1-\beta_1)^2(\alpha + (1-\delta))v_2(1,\pi) \end{aligned}$$

Since from (a) we have

$$\frac{\partial v_2(1,\pi)}{\partial \delta} \leq 0, \qquad \frac{\partial v_2(2,\pi)}{\partial \delta} \leq 0, \qquad v_2(1,\pi) - v_2(2,\pi) \leq 2,$$

and we obtain

$$\frac{\partial v_0(1,\pi)}{\partial \delta} \leq 0.$$

Therefore it is seen formally that action b is optimal at state 2.

Now fix $\beta_1 = \frac{3}{5}, \beta_2 = \frac{2}{5}$ so that the expected rewards of deterministic stationary strategies u', u'' are

$$\begin{aligned} v_0(1,u') &= \frac{2}{5} + (\frac{2}{5})^2 + (\frac{2}{5})^2\frac{3}{5} + (\frac{2}{5})^2(\frac{3}{5})^2 + \cdots = \frac{4}{5}, \\ v_0(1,u'') &= 2(\frac{2}{5})^2 + 2(\frac{2}{5})^2\frac{3}{5} + 2(\frac{2}{5})^2(\frac{2}{5})^2 + \cdots = \frac{4}{5}, \end{aligned}$$

and the expected reward of randomized stationarystrategy π is

$$v_0(1,\pi) = \frac{2(10 - 5\alpha - \alpha^2)}{5(5 - 3\alpha)},$$

which is maximized at $\alpha^* = \frac{5-\sqrt{10}}{3}$. The expected reward associated with this α^* is

$$v_0(1,\pi(\alpha^*)) \approx 0.830019.$$

Given initial state 1, the randomized stationary strategy π^* associated with α^* is the best among all stationary strategies.

Define the strategy u as follows,

$$u_t = \begin{cases} u' & if \quad t = 0, \\ u'' & if \quad t \geq 1. \end{cases}$$

The expected reward of this strategy is

$$v_0(1,u) = \frac{22}{25} = 0.88.$$

From the fact mentioned above, it is seen that for $\epsilon < 0.88 - 0.830019$ there does not exist an ϵ-optimal randomized stationary strategy.

Acknowledgement

The authors would like to thank referees for their very helpful comments and for suggesting another approach using the adaptive Markov decision model.

References

[1] Alden,J.M. and R.Smith : Rolling horizon procedures in nonhomogeneous Markov decision processes. *Operations Research*, 40, No.2, 183-194 (1992).

[2] Bean,J. and R.Smith : Conditions for the existence of planning horizons. *Math. of Operations Research*, 9, No.3, 391-401 (1984).

[3] Bes,C. and S.Sethi : Concepts of forecast and decision horizons : applications to dynamic stochastic optimization problems. *Math. of Operations Research*, 13, No.2, 295-310 (1984).

[4] Hopp,W.J.,J.C.Bean and R.Smith : A new optimality criterion for nonhomogeneous Markov decision processes. *Operations Research*, 35, No.6, 875-883 (1987).

[5] Puterman,M.L. : *Markov Decision Processes : Discrete Stochastic Dynamic Programming.* John Wiley and Sons, New York, 1994.

[6] Ross,S.M. : *Introduction to Stochastic Dynamic Programming.* Academic Press, New York, 1984.

[7] Sethi,S and Bhaskaran,S. : Conditions for the Existence of Decision Horizons for Discounted Problems in a Stochastic Environment. *Operations Research Letters*, 4, No.2, 61-64 (1985).

[8] Shapiro,J.F. : Turnpike planning horizons for a Markovian decision model. *Management Sci.*, 14, No.5, 292-300 (1968).

[9] White,D.J. : Infinite horizon Markov decision processes with unknown or variable discount factors. *Euro. J. Operations Research*, 28, No.1, 96-98 (1987).

Tetsuo IIDA:
Department of Industrial Engineering and Management
Graduate School of Decision Science and Technology
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku
Tokyo 152, Japan
E-mail: tetsuo@me.titech.ac.jp