AN ALGORITHM TO OBTAIN AN OPTIMAL STRATEGY FOR THE MARKOV DECISION PROCESSES, WITH PROBABILITY DISTRIBUTION FOR THE PLANNING HORIZON.

Goulionis E. John
Department of statistics and insurance science
University of Pireas, 80 Karaoli a Dimitriou Street,
18534 Piraeus, Greece

ABSTRACT
In this paper we formulate Markov Decision Processes with Random Horizon. We show the optimality equation for this problem, however there may not exist optimal stationary strategies. For the MDP (Markov–Decision–Process), with probability distribution for the planning horizon with infinite support, we show Turnpike Planning Horizon Theorem. We develop an algorithm obtaining an optimal first stage decision. We give some numerical examples.

MSC: 90C40, 90B50, 90C39, 90C15

KEY WORDS: MDPs, optimization, probabilities and decision making, operation research

RESUMEN
En este trabajo formulamos un Proceso de Decisión Markoviano con Horizonte Aleatorio. Desarrollamos la ecuación de optimalidad para este problema, sin embargo puede no existir estrategias optimales estacionarias. Para el MDP (Proceso de Decisión Markoviano), con distribución de probabilidad para horizonte de planeamiento con soporte infinito, demostramos el Teorema de Horizonte de Planeamiento de Turnpike. Desarrollamos un algoritmo para obtener una decisión de primera etapa optimal. Damos algunos ejemplos numéricos.

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.

If the planning horizon changes may cause a remarkable change of optimal strategy, and the total reward may differ much. Hence, it is necessary to make a decision considering the probability of the time at which the project will end. We formulate these problems using MDPs in which probability distributions for the planning horizon are given in advance, that is, MDP with Random Horizon. In this paper we will consider non – homogeneous MDPs.

It is known one typical example with a geometrically distributed planning horizon, which is equivalent to an ordinary discounted MDP (Ross [9]). We can notice that the discount rate represents an evaluation of uncertainty expected to be happened in the future. Numerous researches have been made for the type of MDP which has variable discount rates (White [12], Puterman [8], Sondik [12].

In the MDP with random horizon there may not exist an optimal stationary strategy. 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. When the support of the probability distribution for the planning horizon is infinite, it is difficult to solve the problem. So we adopt a rolling horizon strategy to obtain an optimal strategy, that is, first we obtain the Turnpike Planning horizon for MDP and solve the problem under its horizon. Shapiro [11] 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]. They treats deterministic decision problems. In addition Hopp, Bean and Smith [7] 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, the model is described in detail and some assumptions are provided. We derive the optimality equation for the MDP with random horizon. In this section the optimal problem is formulated as MDP.

In Section 3 we describe the structure and 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 problem based on the nature derived in section 3. In section 5 some conclusions are provided.

2. MODEL DESCRIPTION AND ASSUMPTIONS.

Let (Ω,F,P) denote the underlying probability space. Let $T=\{0,1,2,\ldots\}$ be the set of nonnegative integers. We consider a discrete – time non – homogeneous Markov Decision Model with

(i) Countable state space S,

(ii) measurable action space A endowed with σ-field, A containing all one-point subsets of A.

(iii) sets of action $A(s)$ available at $s\in S$, where $A(s)$ is an element of A,

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

(v) sets of reward functions $\{r_t(i,a)\}$ at stage t, $t\in T$, 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, $t\in T$, 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,

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

Let a function $a_t: S\to A$, $t\in T$ be a decision function with $a_t(s_t)\in A(s_t)$. The sequence $\delta=(a_t, t\in T)$ is called a strategy. Let Δ denote the set of all strategies. We also use the notation ${}_n\delta=(\delta_0,\delta_1,\ldots,\delta_{n-1})$ to represent first n decisions in δ. In this model, we also set,

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

Also 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 $t\in T$, $p_t(j\mid s',a')=0$, $p_t(s'\mid s',a')=1$,

(v)′ for all $t\in T$, $r_t(s',a')=o$

Let $H_t = S \times \left(A' \times S' \right)^t$ be the space of histories up to the stage $t \in \overline{T} \cup \{\infty\}$, where $A' = A \cup \{a'\}$. If a strategy $\delta \in \Delta$ and initial state s are specified, transition probabilities are determined completely. Accordingly a probability measure P_s^δ is induced.

Let x_t denote the state of process at state t, M denote the random planning horizon with distribution φ_t, A_t denote the action taken at state t, then the expected total reward for the n-horizon problem is given by

$$R_t\left(X_t, A_t\right) = \begin{cases} r_t\left(X_t, A_t\right) & when \quad X_t \in S, t < M \\ c_t\left(X_t, A_t\right) & when \quad X_t \in S, t = M \\ 0 & when \quad X_t \in S' \end{cases}. \tag{2.1}$$

Considering the n-horizon problem, for a fixed n, when the process starts with an initial state s under a strategy δ, the expected total reward for this problem is given by

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

where E_s^δ is the corresponding expectation operator. We notice that the expected reward $V\left(s, \delta, n\right)$ depends only on the first n decisions δ_n in each δ.

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

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

A strategy $\delta^*\left(n\right) \in \Delta$ is called an optimal strategy for the n-horizon problem if for each $s \in S$, $V\left(s, \delta^*\left(n\right), n\right) = \sup_{\delta \in \Delta} V\left(s, \delta, n\right)$.

For the random M-horizon problem, we set

$$V\left(s, \delta\right) = E_s^\delta \sum_{t=0}^{n} R_t\left(X_t, A_t\right). \tag{2.4}$$

It should be also noted that an optimal strategy for n-horizon problem depends only on the first n decisions in each δ.

Similarly a strategy $\delta^* \in \Delta$ is called an optimal strategy for the random M-horizon problem if for each $s \in S$, $V\left(s, \delta^*\right) = \sup_{\delta \in \Delta} V\left(s, \delta\right)$.

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

$$V\left(s, \delta_\varepsilon^*\left(n\right), n\right) \geq V\left(s, \delta^*\left(n\right), n\right) - \varepsilon. \tag{2.5}$$

Now we consider the optimality equation for the MDP with random horizon. Let b_t be a probability which the project is still continuing at stage $\left(t+1\right)$ under condition that it has continued until stage t, that is,

$$b_t = \frac{1-\sum_{\kappa=1}^{t}\phi_\kappa}{1-\sum_{\kappa=1}^{t-1}\phi_\kappa}. \tag{2.6}$$

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

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

Now, let $v_t^*(i)$ denote the 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,

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

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

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

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

3. OPTIMAL STRATEGIES WHEN THE SUPPORT OF THE PROBABILITY DISTRIBUTION FOR THE PLANNING HORIZON IS INFINITE.

In this section we discuss the MDPs with random horizon which have the infinite support of the probability distribution for the planning horizon. We discuss the 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 [7], Bes and Sethi [3] are based on this idea, too.

We now define a metric topology on the set of all strategies Δ. The metric ρ below is the same one which Bean and Smith [2] uses

$$\rho(\delta,\delta') = \sum_{n=1}^{\infty} 2^{-n}\cdot\sigma_n(\delta,\delta'),$$

where $\sigma_n(\delta,\delta') = \begin{cases} 1 & a_n'(x) \neq a_n(x) \quad \exists\, x\in S \\ 0 & a_n'(x) = a_n(x) \quad \forall\, x\in S \end{cases}$

The ρ metric has the property that any two strategies that agree in the first M policies, for any M, are considered closer than any two strategies that do not.

Now we define

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

Proposition 3.1. (Δ,ρ) is a metric space. If $\rho(\delta,\delta')<\varepsilon<1$ for all $n\le -\log_2\varepsilon$ $\alpha_n' = \alpha_n$.

Proof. See Bean and Smith [2] □

Assumption 3.1. We assume that the strategy Space, Δ, is compact in metric space generated by ρ. This assumption precludes the possibility of a sequence of feasible strategies converging to an infeasible strategy. For further discussion of a related problem, see Bean and Smith [2].

Let $\Delta_t = \{a_t\}$ for all $t \in T$. We define the discrete topology $p_t\left(a_t, a_t^{'}\right) = 2^{-t} \cdot \sigma_t\left(a_t, a_t^{'}\right)$ on them. Then the theorem below holds.

Theorem 3.2. Δ is compact if and only if $\forall t \in T$, Δ_t are finite sets. If Δ is compact, then each cylinder subset of Δ is compact.

Proof. See [3] □

Theorem 3.3.A periodic forecast horizon optimal strategy exists for the nonhomogeneous Markov decision process.

Proof.

Compactness of Δ implies that the sequence $\left\{\delta^*\left(M_m\right)\right\}_{m=1}^{\infty}$ has a convergent subsequence. The limit of such a sequence is PFH optimal by definition (3.1). When S and A are finite sets, a compactness of Δ is ensured.

From the definition of $V(s,a)$, 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, $E(M)$, is finite,

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

Then, $\forall \delta \in \Delta$,

$$V(s,\delta,M) = E_s^{\delta}\left[\sum_{t=0}^{N} R_t\left(X_t, A_t\right)\right]$$

$$= \sum_{t=1}^{\infty} E_s^{\delta}\left[r_t\left(X_t, A_t\right) \middle| M > t\right] \cdot P[M > t] + \sum_{t=1}^{\infty} E_s^{\delta}\left[c_t\left(X_t, A_t\right) \middle| M = t\right] \cdot P[M = t]$$

$$< \max\{R, C\} \cdot \sum_{t=1}^{\infty}\left(1 - \sum_{\kappa=1}^{t-1} \varphi_{\kappa}\right) = \max\{R, C\} \cdot \sum_{t=1}^{\infty}\left(t \cdot \varphi_{\kappa}\right).$$

Thus from (3.1), $V(s,\delta,M) < +\infty$. □

Assumption 3.4. The expectation of the planning horizon is finite.

Now we can discuss the existence of optimal strategy for the MDP with random horizon and infinite support.

Lemma 3.5. $V(s,\delta)$ is continuous in $\delta \in \Delta$.

Proof.

For any $\varepsilon > 0$, there exists Λ, such that $\max\{R,C\} \cdot \sum_{t=\Lambda+1}^{\infty} \prod_{\kappa=1}^{t-1} b_\kappa < \frac{\varepsilon}{2}$.

Therefore we get a v such that $\Lambda \le -\log_2 v$. Then for any $\delta' \in \Delta$ such that $\rho(\delta, \delta') < v$,

$$\left|V(s,\delta) - V(s,\delta')\right| = \left| E_s^{\delta}\left[\sum_{t=1}^{M} R_t(X_t, A_t)\right] - E_s^{\delta'}\left[\sum_{t=1}^{M} R_t(X_t, A_t)\right]\right|$$

$$= \left| \sum_{t=\mathrm{M}+1}^{\infty} E_s^{\delta}\left[r_t(X_t, A_t)\middle| M > t\right] \cdot P[M > t] + \sum_{t=\mathrm{M}+1}^{\infty} E_s^{\delta}\left[c_t(X_t, A_t)\middle| M = t\right] \cdot P[M = t] \right.$$

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

$$\le 2\max\{R,C\} \sum_{t=M+1}^{\infty} \prod_{\kappa=1}^{t-1} b_\kappa < \varepsilon.$$

Let now $\delta^*(n) \in \Delta$ be an optimal strategy for n-horizon problem and $\hat{\Delta}$ be a set of cluster points of all the sequences $\{\delta^*(n)\}$, that is, a set of PFH – optimal strategies and let $\hat{\Delta}_t = \{\alpha_t \mid \delta \in \hat{\Delta}\}$, $t \in T$. Note that since $V(s,\delta,n)$ is continuous in n and Δ is compact, $\hat{\Delta}$ is a nonempty set.

Theorem 3.6 (existence). Under assumptions 2.1, 3.1 and 3.4 there exists a PFH – optimal strategy for the MDP with random horizon.

Proof.

Since Δ is compact, $V(s,\delta)$ is uniformly continuous on Δ. Thus there exists a strategy $\delta^* \in \Delta$ such that $V(s,\delta^*) = \max_{\delta\in\Delta} V(s,\delta)$. Therefore there exists a PFH – optimal strategy for the MDP with random horizon $\delta^* \in \Delta$.

Lemma 3.7. $\lim_{n\to\infty} \max_{\delta\in\Delta} V(s,\delta,n) = \max_{\delta\in\Delta} V(s,\delta)$

Proof.

Let $K_n = \max\{R,C\} \sum_{t=n+1}^{\infty} \prod_{\kappa=1}^{t-1} b_\kappa$, then we have

$$\max_{\delta\in\Delta} V(s,\delta) - K_n \le \max_{\delta\in\Delta} V(s,\delta,n) \le \max_{\delta\in\Delta} V(s,\delta) + K_n$$

Thus since $K_n \to 0$ as $n \to \infty$, $\lim_{n\to\infty} \max_{\delta\in\Delta} V(s,\delta,n) = \max_{\delta\in\Delta} V(s,\delta)$ □

Let Δ^* denotes a set of all optimal strategies $\delta^* \in \Delta^*$ for the MDP with random horizon.

Lemma 3.8. $\hat{\Delta} \subset \Delta^*$

Proof.

Let $\delta^* \in \hat{\Delta}$. From the definition there exists a sequence of strategies $\{\delta^*(m(j))\}_{j\in T}$ such that

$\lim_{j\to\infty} \delta^*(m(j)) = \delta^*$.

Thus $\lim_{j\to\infty} V(s, \delta^*(m(j)), m(j)) = V(s,\delta^*)$.

Since from lemma (3.7) $V\left(s,\delta^*\right)=\max_{\delta\in\Delta}V\left(s,\delta\right)$, $\delta^*\in\Delta^*$. □

There may not necessarily exists a stationary deterministic strategy or stationary randomized strategy for the MDP with random horizon. There may not exist even an ε - optimal randomized tationary strategy. We show an example.

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

$p\left(1\mid s,a\right)=p\left(2\mid s,b\right)=1$, for $s=1,2$.

$r\left(1,a\right)=1$, $r\left(1,b\right)=r\left(2,a\right)=0$, $r\left(2,b\right)=2$.

$c\left(s,x\right)=0$, for $s=1,2$, $x=a,b$.

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

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

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,

δ'**:** keep your state (use action at state 1 and action b at state 2),
δ''**:**move to state 2 and keep it (use only action b).

We shall examine an optimal randomized stationary strategy δ is defined as

$\delta\left(a\mid 1\right)=a$, $\quad\delta\left(a\mid 2\right)=\tau$

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

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

Solving the equations, we have

(a)
$$\begin{pmatrix}v_2\left(1,\delta\right)\\ v_2\left(2,\delta\right)\end{pmatrix}=\frac{1-\beta_2}{\beta_2\left(1-\alpha\left(1-\beta_2\right)+\tau\left(1-\beta_2\right)\right)}\times \begin{pmatrix}a-a\left(1-\tau\right)\left(1-\beta_2\right)+2\left(1-a\right)\left(1-\tau\right)\left(1-\beta_2\right)\\ a\tau\left(1-\beta_2\right)+2\left(1-\tau\right)-2a\left(1-\tau\right)\left(1-\beta_2\right)\end{pmatrix}$$

Similarly,

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

so that

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

Since from (a) we have

$$\frac{\vartheta v_2(1,\delta)}{\vartheta\tau} \le 0, \qquad \frac{\vartheta v_2(2,\delta)}{\vartheta\tau} \le 0, \qquad v_2(1,\delta) - v_2(2,\delta) \le 2,$$

and we obtain

$$\frac{\vartheta v_0(1,\delta)}{\vartheta\tau} \le 0.$$

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

$$v_0(1,\delta') = \frac{2}{5} + \left(\frac{2}{5}\right)^2 + \left(\frac{2}{5}\right)^2\frac{3}{5} + \left(\frac{2}{5}\right)^2\left(\frac{3}{5}\right)^2 + \ldots = \frac{4}{5},$$

$$v_0(1,\delta'') = 2\left(\frac{2}{5}\right)^2 + 2\left(\frac{2}{5}\right)^2\frac{3}{5} + 2\left(\frac{2}{5}\right)^2\left(\frac{2}{5}\right)^2 + \ldots = \frac{4}{5},$$

and the expected reward of randomized stationary strategy π is

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

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

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

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

Define the strategy as follows,

$$\delta_t = \begin{cases} \delta' & if \quad t = 0 \\ \delta'' & if \quad t \ge 1 \end{cases}$$

The expected reward of this strategy is

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

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

We continue now with showing that a theorem similar to Turnpike Planning Horizon Theorem which Shapiro [11]shows for the homogeneous discounted MDP holds for this MDP with random horizon. Because, there may not necessarily exists a stationary deterministic strategy or stationary randomized strategy for MDP with random horizon, an optimal strategy we wish to know may be non – stationary, so it's difficult to get it directly.

We introduce the following two notations,

$F = \left\{ {}_1\delta : \delta \in \Delta^* \right\}$: a set of optimal decisions at the first state for the random M- horizon problem, and

$F(n) = \left\{ {}_1\delta : \delta = \delta^*(n) \right\}$:a set of optimal decisions at the first state for the n - horizon problem.

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 $\left\{ \delta^*(M_i) \right\}$ such that ${}_1\delta^*(M_i) \notin F$ for all i. Since Δ is compact, there exists a subsequence $\left\{ \delta^*(m(M_i)) \right\}$ such that its limit is $\delta^{**} \in \Delta$ .Thus for sufficient large $m(M_i)^*$, $p(\delta^{**}, \delta^*(m(M_i))) < \varepsilon$, so ${}_1\delta^{**} = {}_1\delta^*(m(M_i))$. Therefore ${}_1\delta^{**} \notin F$. On the other hand, from definition $\delta^{**} \in \tilde{\Delta}^*$, and from the lemma 3.8 $\delta^{**} \in \Delta^*$. Thus ${}_1\delta^{**} \in F$, 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 exists an optimal rolling strategy.

4. ALGORITHM FOR FINDING AN OPTIMAL FIRST DECISION

Although the Turnpike Planning 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 identifying the PFH–optimal strategy gradually, that is, making the neighborhood of PFH-optimal strategy small.

Let $\hat{\Delta}^n$ denotes a set of strategies such that its first decision is not included in $F(n)$, that is $\hat{\Delta}^n = \left\{ \delta \in \Delta : {}_1\delta \notin F(n) \right\}$.

Theorem 4.1. For any $\delta \in \hat{\Delta}^n$, if δ satisfies the following condition (*),

$$(*) \qquad \max_{\delta \in \Delta} V(s, \delta, n) - V(s, \delta', n) > 2 \max\{R, C\} \sum_{t=n+1}^{\infty} \prod_{\kappa=1}^{t} b_\kappa \,, \tag{4.1}$$

δ is not optimal for the problem with infinite support.

Proof.

Let $\delta' \in \Delta$ be a strategy satisfying a condition (*).

Set $K_n = \max\{R, C\} \sum_{t=n+1}^{\infty} \prod_{\kappa=1}^{t} b_\kappa$, then from the condition (*),

$$\max_{\delta \in \Delta} V(s, \delta, n) - V(s, \delta', n) > 2K_n \tag{4.2}$$

and

$$\max_{\delta\in\Delta} V\left(s,\delta,n\right) - K_n \le \max_{\delta\in\Delta} V\left(s,\delta\right). \tag{4.3}$$

Therefore from (4.2) and (4.3)

$$\max_{\delta\in\Delta} V\left(s,\delta\right) - V\left(s,\delta',n\right) > K_n .$$

Thus $\delta' \in \Delta$ is not optimal. □

Remark 4.2. If $F\left(n\right)$ is singleton and a condition of theorem 4.1 holds for any $\delta \in \hat{\Delta}^n$, ${}_1\delta \in F\left(n\right)$ 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 $u_t = \max\left\{R, C\right\} \sum_{n=t+1}^{\infty} \prod_{\kappa=1}^{n} b_\kappa$.

Step 3. $\forall a \in A$, compute $\xi_t^a = \max_{\delta\in\Delta} V\left(s,\delta,t\right) - V\left(s,\delta_a,t\right)$, where, ${}_1\delta_\alpha = \alpha$. If $\xi_t^a > 2\delta_t$ and F^t is singleton, Stop. Its decision is an optimal first one.

Step 4. If $\delta_t \le \varepsilon$, Stop. Its decision is an ε-optimal first one.

Step 5. $t = t + 1$, and go to Step 2.

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

Remark 4.5. If Δ^* 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 Δ^* 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 $\left\{\varphi_t\right\}$. 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 even. We assume that $\underline{S} \le s_t \le \overline{S}$, that is, an upper bound and a lower bound of the stock is given. The cost we consider are following,

$k_t\left(a_t\right)$: the order cost in the period t when a_t items are ordered,

$c_t\left(s_t\right)$: the holding cost in the period t when $s_t \ge 0$, the backlogging cost in the period t when $s_t < 0$,

$r_t\left(x_t\right)$: the income in the period t when x_t items are sold, where $x_t = \max\left\{\min\left\{\eta_t, s_{t-1}\right\}, 0\right\}$.

Accordingly the problem is to maximize the total expected reward:

$$Maximize\ E^{a}_{s_0}\left[\sum_{t=1}^{\infty}\left\{r_t\left(X_t\right)-\kappa_t\left(\mathrm{A}_t\right)-c_t\left(S_t\right)\right\}\right]. \tag{4.5}$$

Now we assume that the data are as follows,

$s_0=5$, $\underline{S}=-5$, $\overline{S}=20$. The expected value of the demands in one-period is 7.

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

Then let the salvage cost per unit be 7.

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

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

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,

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

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 calculations are shown in Table 1.

From Table 1.we can see two tendencies in this inventory problem, one is that quality 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. CONCLUSIONS

The purpose of this paper is to analyze an optimal strategy for the MDP with random horizon, and purpose the algorithm to obtain it numerically by Turnpike Planning Horizon approach. For the processes there may not exist optimal stationary strategies, so we evaluate rolling strategies, derived by

using the result of Turnpike Horizon Theorem. We develop an algorithm obtaining an optimal first stage decision, and some numerical experiments. As a result of numerical experiments, we take that the optimal first decisions depend on the shape of the probability distribution for the planning horizon.

RECEIVED OCTOBER 2008
REVISED NOVEMBER 2009

REFERENCES

[1] ALDEN ,J.M. and SMITH R.(1992): Rolling horizon procedures in nonhomogeneous Markov decision processes. **Operations Research**, 40, 183-194**.**

[2] BEAN, J.,and SMITH R.(1984): Conditions for the existence of planning horizons. **Math. of Operations Research,** 9, 391-401.

[3] BES C. and SETH, S. (1984): Concepts of forecast and decision horizons : applications to dynamic stochastic optimization problems. **Math., of Operations Research,** 13 , 295-310.

[4] GOULIONIS E.J., (2004): Periodic policies for partially observable Markov decision processes, **Working paper** No. 30, 15-28, University of Piraeus 2004.

[5] GOULIONIS, E.J. (2006): A replacement policy under Markovian deterioration. **Mathematical inspection,** 63, 46-70.

[6] GOULIONIS E.J. (2005): P.O.M.D.Ps with uniformly distributed signal processes. **Spoydai** , 55, 34-55.

[7] HOPP,W.J,.BEAN J.C and SMITH, R (1987): A new optimality criterion for nonhomogeneous Markov decision processes. **Operations Research**, 35,875-883.

[8] PUTERMAN, M.L (1994): **Discrete Stochastic Dynamic Programming**. John Wiley and Sons, New York.

[9] ROSS,S.M. (1984): **Introduction to Stochastic Dynamic Programming**. Academic Press, New York.

[10] SETHI S. and BHASKARAN,S. (1985): Conditions for the Existence of Decision Horizons for Discounted Problems in a Stochastic Environment. **Operations Research Letters**, 4, 61-64.

[11] SHAPIRO,J.F. (1968): Turnpike planning horizons for a Markovian decision model. **Management Sci**., 14, 292-300 .

[12] SONDIK.J.E. (1978): Optimal control of partially observable Markov decision processes over infinite horizon. **Qperations Research** , 26, 282-304.

[13] WHITE,D.J. (1987): Infinite horizon Markov decision processes with unknown variables discount factors. **Euro. J. of Qperations Research**,.28, 96-98.