Appl Math Optim 43:187–201 (2001)
DOI: 10.1007/s00245-001-0007-9

Applied Mathematics & Optimization

Markov Decision Processes with Distribution Function Criterion of First-Passage Time*

Liu Jianyong[1] and Siming Huang[2]

[1]Institute of Applied Mathematics, Academia Sinica,
Beijing 100080, People's Republic of China

[2]Institute of Policy and Management, Academia Sinica,
Beijing 100080, People's Republic of China

Abstract. In this paper we discuss MDP with distribution function criterion of first-passage time. Some properties of several kinds of optimal policies are given. Existence results and algorithms for these optimal policies are given in this paper.

Key Words. Markov decision processes, Distribution function of first-passage time, Optimal policy, Reliability.

AMS Classification. 90C40.

1. Introduction

In the field of Markov decision processes (MDP), the normal first-passage model is discussed in some papers, such as [1]–[6]. The aim of the normal model is to maximize the expected reward of a first passage from an arbitrary state to state 0. Applications of the normal first-passage model include stochastic search for a hidden object, shares deal, secretary problem and so on (see [7]).

In this paper we discuss MDP with distribution function criterion of first-passage time. In this paper we also study the first-passage problem, but our aim (see Definition 1.1 below) is different from the aim of the above normal first-passage model. Our model can be applied in the field of reliability (for further discussion, see Remark 1.1).

* The research of the first author was supported by the National Natural Science Foundation of China. The second author's research was supported by the National Natural Science Foundation of China, Grant No. 19731010.

Lin Yuanlie [8] discussed an optimal model for the first-passage time distribution function with a continuous-time parameter. The set of all stationary policies is only considered in [8]. Algorithms to find some optimal policies are given in [8] when the state space and action space are finite.

The criterion for our model is similar to the criterion in [8]. Because our model deals with the case of discrete time, many results in our paper and the methods used by us are different from those in [8].

Other related work includes [9] and [10] (for details, see Remark 1.2).

In this paper the definition and interpretation of our model are given in Section 1.

Some results of two kinds of optimal policies are given in Section 2. These results include sufficient conditions for these optimal policies (Theorems 2.3 and 2.5), necessary conditions for them (Theorems 2.4 and 2.4′), sufficient and necessary conditions for the existence of these optimal policies (Corollaries 2.1 and 2.1′) and algorithms for them (Remark 2.1).

Results in Section 3 include the existence of an n-optimal policy and a sequence of (n, ε)-optimal policies (Corollary 3.1, Theorem 3.5 and Corollary 3.3), and an algorithm to find a sequence of (n, ε)-optimal policies (Remark 3.3).

Our model is $\{S, (A(i), i \in S), q, D\}$, where the state space $S = \{0, 1, 2, \ldots\}$ is countable. Let $S_0 = \{1, 2, 3, \ldots\} \subset S$. We use $A(i)$ to denote the set of possible actions when the system is in state $i \in S$. All $A(i)$ $(i \in S)$ are countable. The letter q denotes the family of stationary one-step transition laws: when the system is in state i and we take an action $a \in A(i)$, the system moves to a new state j selected according to the conditional probability $q(j \mid i, a)$.

The set of general policies $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ (see [1] for the definition of π) is denoted by Π. A mapping $f\colon S \to \bigcup_{i \in S} A(i)$ satisfying $f(i) \in A(i)$ for all $i \in S$ is called a deterministic decision rule. Let F denote the set of all deterministic decision rules f. Let $f_i \in F$, $i = 0, 1, 2, \ldots$, $\pi = (f_0, f_1, f_2, \ldots)$ is called a Markov policy. Let Π_{m}^d denote the set of all Markov policies. Let $f \in F$, $f^\infty = (f, f, \ldots)$ is called a stationary policy. Let Π_s^d denote the set of all stationary policies. Obviously, $\Pi_s^d \subset \Pi_{\mathrm{m}}^d \subset \Pi$.

At any stage t (≥ 0), X_t and Δ_t denote the state of the system and action taken in that state, respectively.

Definition 1.1. We define

$$D(i, n, \pi) = P_\pi\{\tau \geq n \mid X_0 = i\}, \qquad i \in S_0, \quad \pi \in \Pi, \quad n = 1, 2, \ldots,$$

where τ denotes the smallest integer t (≥ 0) such that $X_t = 0$ when $X_0 = i$, $P_\pi\{\tau \geq n \mid X_0 = i\}$ denotes the probability of "$\tau \geq n$" using the policy π starting from i. Obviously, $D(i, 1, \pi) = 1$, $i \in S_0$, $\pi \in \Pi$.

We define

$$D^*(i, n) = \sup_{\pi \in \Pi} D(i, n, \pi), \qquad i \in S_0, \quad n = 1, 2, \ldots.$$

Obviously, $D^*(i, 1) = 1$, $i \in S_0$.

Definition 1.2. Let $n \geq 1$, $\pi \in \Pi$. π is called an optimal policy (up to n) if

$$D(i, k, \pi) = D^*(i, k), \qquad \text{for all} \quad i \in S_0, \quad k = 1, 2, \ldots, n.$$

π is called an optimal policy if

$$D(i, k, \pi) = D^*(i, k), \qquad \text{for all} \quad i \in S_0, \quad k = 1, 2, \ldots.$$

Remark 1.1. From the viewpoint of reliability we explain the above model as follows. Let the state 0 denote an inefficient state of a system. Then τ denotes the working life (operating life) of the system and $P_\pi\{\tau \geq n \mid X_0 = i\}$ denotes the reliability function of the system using policy π starting from state i. Roughly our aim is to find an optimal policy which maximizes the reliability function of the system. Hence, the background of our model is an optimization problem in the field of reliability.

Remark 1.2. The optimization problem $\inf_\pi P_\pi\{Z_\infty^\pi \leq c \mid X_0 = i\}$ is studied in [9], where Z_∞^π denotes the total discounted reward (infinite horizon) using policy π starting from state i and c is a constant.

The optimization problem $\sup_\pi P_\pi\{R \geq c \mid X_0 = i\}$ is studied in [10], where

$$R = \varliminf_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} r(X_n, \Delta_n)$$

denotes the average reward using policy π starting from state i and c is a constant.

It is easy to see, the above criteria are different from the criteria (Definition 1.1) of our model.

Let $X_0 = i_0$, $\Delta_0 = a_0 \in A(i_0)$, $X_1 = i_1$, $\Delta_1 = a_1 \in A(i_1), \ldots, X_n = i_n$. $h_n = (i_0, a_0, i_1, a_1, \ldots, i_n)$ is called the history up to stage n. H_n ($n \geq 0$) denotes the set of all h_n.

Let $\pi = (\pi_0, \pi_1, \pi_2, \ldots) \in \Pi$, $h_n = (i_0, a_0, i_1, a_1, \ldots, i_n) \in H_n$ ($n \geq 1$). The policy $\pi' = (\pi_0', \pi_1', \ldots) \in \Pi$ is defined as follows: for $\forall t \geq 0$, $\forall h_t \in H_t$, we define

$$\pi_t'(a \mid h_t) = \pi_{n+t}(a \mid i_0, a_0, i_1, a_1, \ldots, a_{n-1}, h_t), \qquad a \in A(i),$$

where the last component of h_t is i.

Write $\pi' = \pi(i_0, a_0, \ldots, i_{n-1}, a_{n-1})$.

2. Some Results of Optimal Policies

Theorem 2.1. *Let $n \geq 2$, $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$, $i \in S_0$, then*

$$D(i, n, \pi) = \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D(j, n-1, \pi(i, a)). \tag{2.1}$$

Proof.

$$\begin{aligned} D(i,n,\pi) &= P_\pi\{\tau \geq n \mid X_0 = i\} \\ &= \sum_{a\in A(i)} P_\pi\{\tau \geq n, \Delta_0 = a \mid X_0 = i\} \\ &= \sum_{a\in A(i)} \pi_0(a \mid i) P_\pi\{\tau \geq n \mid X_0 = i, \Delta_0 = a\} \\ &= \sum_{a\in A(i)} \pi_0(a \mid i) \sum_{j\in S_0} P_\pi\{\tau \geq n, X_1 = j \mid X_0 = i, \Delta_0 = a\} \\ &= \sum_{a\in A(i)} \pi_0(a \mid i) \sum_{j\in S_0} q(j \mid i,a) P_\pi\{\tau \geq n \mid X_0 = i, \Delta_0 = a, X_1 = j\}. \end{aligned}$$

By the definition of $\pi(i,a)$ (see Section 1), it is easy to see that

$$\begin{aligned} &P_\pi\{\tau \geq n \mid X_0 = i, \Delta_0 = a, X_1 = j\} \\ &\quad = P_{\pi(i,a)}\{\tau \geq n-1 \mid X_0 = j\}, \qquad n \geq 2, \quad i, j \in S_0, \quad a \in A(i). \end{aligned}$$

So (2.1) is true. □

Theorem 2.2. *Let $n \geq 2$, $i \in S_0$, then*

$$D^*(i,n) = \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a) D^*(j, n-1).$$

Proof. $\forall \pi \in \Pi$, $i \in S_0$. By Theorem 2.1 we have

$$\begin{aligned} D(i,n,\pi) &= \sum_{a\in A(i)} \pi_0(a \mid i) \sum_{j\in S_0} q(j \mid i,a) D(j, n-1, \pi(i,a)) \\ &\leq \sum_{a\in A(i)} \pi_0(a \mid i) \sum_{j\in S_0} q(j \mid i,a) D^*(j, n-1) \\ &\leq \sum_{a\in A(i)} \pi_0(a \mid i) \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a) D^*(j, n-1) \\ &= \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a) D^*(j, n-1). \end{aligned}$$

So,

$$D^*(i,n) \leq \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a) D^*(j, n-1), \qquad i \in S_0. \tag{2.2}$$

$\forall \varepsilon > 0$, $\forall i \in S_0$. It is evident that there exists ${}^i\pi \in \Pi$ such that $D(i, n-1, {}^i\pi) \geq D^*(i, n-1) - \varepsilon$. The policy $\pi \in \Pi$ is defined as follows: $\pi = {}^i\pi$, when $X_0 = i \in S_0$; π is an arbitrary policy in Π, when $X_0 = 0$. Then

$$D(i, n-1, \pi) \geq D^*(i, n-1) - \varepsilon, \qquad \text{for all} \quad i \in S_0.$$

It is evident that, $\forall i \in S_0$, $\exists a_i \in A(i)$ such that

$$\sum_{j\in S_0} q(j \mid i, a_i) D^*(j, n-1) \geq \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a) D^*(j, n-1) - \varepsilon.$$

We define $f(i) = a_i, i \in S_0$; $f(0) \in A(0)$, then $f \in F$. Let $\hat{\pi} = (f, \pi)$. By Theorem 2.1 we have

$$\begin{aligned} D(i, n, \hat{\pi}) &= \sum_{j \in S_0} q(j \mid i,\ f(i)) D(j, n-1, \pi) \\ &\geq \sum_{j \in S_0} q(j \mid i,\ f(i))(D^*(j, n-1) - \varepsilon) \\ &\geq \sum_{j \in S_0} q(j \mid i,\ f(i)) D^*(j, n-1) - \varepsilon \\ &\geq \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1) - 2\varepsilon, \qquad i \in S_0. \end{aligned}$$

So, $D^*(i, n) \geq \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1) - 2\varepsilon, \qquad i \in S_0.$

Let $\varepsilon \to 0$, we have

$$D^*(i, n) \geq \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1), \qquad i \in S_0. \tag{2.3}$$

From (2.2) and (2.3) we know that Theorem 2.2 is true. □

Theorem 2.3. *Let $f \in F$ satisfy $q(j \mid i,\ f(i)) = \sup_{a \in A(i)} q(j \mid i, a)$ for all $i, j \in S_0$, then f^∞ is an optimal policy.*

Proof. (Apply the induction.) Obviously, $D(i, 1,\ f^\infty) = D^*(i, 1), i \in S_0$.

Induction hypothesis: $D(i, n,\ f^\infty) = D^*(i, n), i \in S_0$.

We have by Theorem 2.1 and the induction hypothesis,

$$\begin{aligned} D(i, n+1,\ f^\infty) &= \sum_{j \in S_0} q(j \mid i,\ f(i)) D(j, n,\ f^\infty) \\ &= \sum_{j \in S_0} q(j \mid i,\ f(i)) D^*(j, n) \\ &\geq \sum_{j \in S_0} q(j \mid i, a) D^*(j, n), \quad i \in S_0,\ a \in A(i). \end{aligned}$$

So, we have by Theorem 2.2,

$$D(i, n+1,\ f^\infty) \geq \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, n) = D^*(i, n+1), \qquad i \in S_0.$$

That is $D(i, n+1,\ f^\infty) = D^*(i, n+1), i \in S_0$. □

Example 2.1. Let $S = \{0, 1, 2\}$, $S_0 = \{1, 2\}$; $A(0) = A(1) = \{1\}$, $A(2) = \{1, 2\}$; $q(0 \mid 0, 1) = 1$, $q(0 \mid 1, 1) = 0.25$, $q(1 \mid 1, 1) = 0.5$, $q(2 \mid 1, 1) = 0.25$, $q(0 \mid 2, 1) = 0.35$, $q(1 \mid 2, 1) = 0.15$, $q(2 \mid 2, 1) = 0.5$, $q(0 \mid 2, 2) = 0.15$, $q(1 \mid 2, 2) = 0.3$, $q(2 \mid 2, 2) = 0.55$. It is easy to see that $F = \{f, g\}$, where $f(2) = 1$, $g(2) = 2$. Obviously, $q(j \mid i, g(i)) = \sup_{a \in A(i)} q(j \mid i, a)$, $i, j \in S_0$. So, g^∞ is an optimal policy by Theorem 2.3.

Lemma 2.1. *Let $n \geq 2$, $i \in S_0$. Let $\pi = (\pi_0, \pi_1, \pi_2, \ldots) \in \Pi$ satisfy $D(i, n, \pi) = D^*(i, n)$, then $D^*(i, n) = \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1)$.*

Proof. By Theorems 2.1 and 2.2, we have

$$\begin{aligned} D^*(i, n) = D(i, n, \pi) &= \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D(j, n-1, \pi(i, a)) \\ &\leq \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1) \\ &\leq \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1) \\ &= D^*(i, n). \end{aligned}$$

So, $D^*(i, n) = \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1)$. □

We define

$$A_n^*(i) = \left\{ a \in A(i) \mid \sum_{j \in S_0} q(j \mid i, a) D^*(j, n-1) = D^*(i, n) \right\}, \quad i \in S_0, \quad n \geq 2.$$

If all $A(i)$ ($i \in S_0$) are finite, then by Theorem 2.2 we know that $A_n^*(i) \neq \emptyset$, $i \in S_0$, $n \geq 2$.

Theorem 2.4. *Let $n \geq 2$. $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$ is an optimal policy (up to n). Then*

$$\{a \in A(i) \mid \pi_0(a \mid i) > 0\} \subset \bigcap_{k=2}^{n} A_k^*(i), \qquad i \in S_0.$$

Proof. For $2 \leq k \leq n$, by Theorem 2.1, Lemma 2.1 and Theorem 2.2 we have

$$\begin{aligned} D^*(i, k) = D(i, k, \pi) &= \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D(j, k-1, \pi(i, a)) \\ &= \sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1) \\ &\leq \sum_{a \in A(i)} \pi_0(a \mid i) \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1) \\ &= D^*(i, k), \qquad i \in S_0. \end{aligned}$$

So,

$$\begin{aligned} &\sum_{a \in A(i)} \pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1) \\ &\quad = \sum_{a \in A(i)} \pi_0(a \mid i) \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1), \qquad i \in S_0. \end{aligned}$$

On the other hand,

$$\begin{aligned} &\pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1) \\ &\quad \leq \pi_0(a \mid i) \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1), \qquad i \in S_0, \quad a \in A(i). \end{aligned}$$

So,

$$\pi_0(a \mid i) \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1)$$
$$= \pi_0(a \mid i) \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1), \qquad i \in S_0, \quad a \in A(i).$$

If $i \in S_0$, $a \in A(i)$ and $\pi_0(a \mid i) > 0$, then

$$\sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1) = \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j, k-1) = D^*(i, k),$$

that is, $a \in A_k^*(i)$. So $\{a \in A(i) \mid \pi_0(a \mid i) > 0\} \subset A_k^*(i),\ i \in S_0$. Hence Theorem 2.4 is true. □

Similarly, we have Theorem 2.4′.

Theorem 2.4′. *Let $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$ be an optimal policy, then*

$$\{a \in A(i) \mid \pi_0(a \mid i) > 0\} \subset \bigcap_{k=2}^{\infty} A_k^*(i), \qquad i \in S_0.$$

Theorem 2.5. *Let $n \geq 2$ and $f \in F$ satisfy $f(i) \in \bigcap_{k=2}^{n} A_k^*(i)$ for all $i \in S_0$, then f^∞ is an optimal policy* (*up to n*).

Proof. (Apply the induction). Obviously, $D(i, 1, f^\infty) = D^*(i, 1)$, $i \in S_0$.

Induction hypothesis: for $1 \leq k < n$ we have $D(i, k, f^\infty) = D^*(i, k)$, $i \in S_0$.

Because $f(i) \in A_{k+1}^*(i)$, $i \in S_0$, we have by Theorem 2.1 and the induction hypothesis,

$$\begin{aligned} D(i, k+1, f^\infty) &= \sum_{j \in S_0} q(j \mid i, f(i)) D(j, k, f^\infty) \\ &= \sum_{j \in S_0} q(j \mid i, f(i)) D^*(j, k) \\ &= D^*(i, k+1), \qquad i \in S_0. \end{aligned}$$

So, $D(i, k, f^\infty) = D^*(i, k)$, $i \in S_0$, $k = 1, 2, \ldots, n$. □

Similarly, we have Theorem 2.5′.

Theorem 2.5′. *Let $f \in F$ satisfy $f(i) \in \bigcap_{k=2}^{\infty} A_k^*(i)$ for all $i \in S_0$, then f^∞ is an optimal policy.*

Theorem 2.6. *Let $n \geq 2$. If there exists an optimal policy* (*up to n*), *then there exists $f^\infty \in \Pi_s^d$ which is an optimal policy* (*up to n*).

Proof. This follows immediately from Theorem 2.4 and Theorem 2.5. □

Similarly, we have Theorem 2.6′.

Theorem 2.6′. *If there exists an optimal policy, then there exists* $f^\infty \in \Pi_s^d$ *which is an optimal policy.*

Must f^∞ be an optimal policy (up to n) if $f \in F$ satisfies $D(i,k,f^\infty) = \sup_{g\in F} D(i,k,g^\infty)$ for all $i \in S_0$, $k = 1,2,\ldots,n$? The answer is negative.

Example 2.2. Let S, S_0, $A(i)$ $(i \in S)$ and $q(j \mid 1,1)$ $(j \in S)$ be the same as those in Example 2.1. Let $q(0 \mid 0,1) = 1$, $q(0 \mid 2,1) = 0.15$, $q(1 \mid 2,1) = 0.15$, $q(2 \mid 2,1) = 0.7$, $q(0 \mid 2,2) = 0.1$, $q(1 \mid 2,2) = 0.6$, $q(2 \mid 2,2) = 0.3$. It is easy to see that $F = \{f,g\}$, where $f(2) = 1$, $g(2) = 2$. It is easy to see that by Theorem 2.1, $D(i,k,g^\infty) = \sum_{j\in S_0} q(j \mid i, g(i)) D(j,k-1,g^\infty)$, $i \in S_0$, $k \geq 2$. So, $D(1,2,g^\infty) = 0.75$, $D(2,2,g^\infty) = 0.9$, $D(1,3,g^\infty) = 0.6$, $D(2,3,g^\infty) = 0.72$. Similarly, we have $D(1,2,f^\infty) = 0.75$, $D(2,2,f^\infty) = 0.85$, $D(1,3,f^\infty) = 0.5875$, $D(2,3,f^\infty) = 0.7075$. So, $D(i,k,g^\infty) \geq D(i,k,f^\infty)$, $i \in S_0$, $k = 1,2,3$.

On the other hand, we define $\pi = (f, g^\infty)$. It is easy to see that by Theorem 2.1,

$$D(2,3,\pi) = \sum_{j\in S_0} q(j \mid 2, f(2)) D(j,2,g^\infty) = 0.7425 > 0.72 = D(2,3,g^\infty),$$

hence g^∞ is not an optimal policy (up to 3).

To sum up, we have Corollary 2.1 and Corollary 2.1′.

Corollary 2.1. *Let* $n \geq 2$. *The following three conditions are equivalent*:

(1) *There exists an optimal policy* (*up to* n).
(2) *There exists* $f^\infty \in \Pi_s^d$ *which is an optimal policy* (*up to* n).
(3) $\bigcap_{k=2}^n A_k^*(i) \neq \emptyset$ *for all* $i \in S_0$.

Corollary 2.1′. *The following three conditions are equivalent*:

(1) *There exists an optimal policy.*
(2) *There exists* $f^\infty \in \Pi_s^d$ *which is an optimal policy.*
(3) $\bigcap_{k=2}^\infty A_k^*(i) \neq \emptyset$ *for all* $i \in S_0$.

Remark 2.1. To sum up, when S and all $A(i)$ $(i \in S_0)$ are finite an algorithm can be stated as follows: using Theorem 2.2 we can successively find $D^*(i,k)$ and $A_k^*(i)$, $i \in S_0$, $k = 2,3,\ldots,n$. If $\bigcap_{k=2}^n A_k^*(i) \neq \emptyset$ for all $i \in S_0$, then we can find $f^\infty \in \Pi_s^d$ which is an optimal policy (up to n) by Theorem 2.5. If $i_0 \in S_0$ and k_0 $(2 \leq k_0 \leq n)$ such that $\bigcap_{k=2}^{k_0} A_k^*(i_0) = \emptyset$ exist, then there is no optimal policy (up to n). Using this algorithm we can judge whether an optimal policy (up to $n \geq 2$) exists and can find an optimal stationary policy (up to $n \geq 2$) if there exists an optimal policy (up to $n \geq 2$). Similarly, an algorithm to find an optimal stationary policy can be given.

Remark 2.2. In Example 2.2 it is easy to see that $A_2^*(2) = \{2\}$, $A_3^*(2) = \{1\}$. So $A_2^*(2) \cap A_3^*(2) = \emptyset$. Hence there is no optimal policy (up to 3) or optimal policy for Example 2.2.

In the normal optimal first-passage model (see [2]) we define the one-step reward $r(i, a) = 1$ for all $i \in S_0$ and $a \in A(i)$, then we can obtain the following optimization problem: $\sup_{\pi\in\Pi} E_\pi\{\tau \mid X_0 = i\}$, $i \in S_0$.

We assume that $P_\pi\{\tau = \infty \mid X_0 = i\} \equiv P_\pi\{X_n \neq 0, n = 1, 2, 3, \ldots \mid X_0 = i\} = 0$ and $E_\pi\{\tau \mid X_0 = i\} < +\infty$ for all $i \in S_0$ and $\pi \in \Pi$. If $\pi^* \in \Pi$ such that $E_{\pi^*}\{\tau \mid X_0 = i\} = \sup_{\pi\in\Pi} E_\pi\{\tau \mid X_0 = i\}$ for all $i \in S_0$, then π^* is called an expectation life optimal policy.

Theorem 2.7. *Let $P_\pi\{\tau = \infty \mid X_0 = i\} = 0$ and $E_\pi\{\tau \mid X_0 = i\} < +\infty$ for all $i \in S_0$ and $\pi \in \Pi$. If $\pi^* \in \Pi$ is an optimal policy (see Definition* 1.2*), then π^* is also an expectation life optimal policy.*

Proof. For all $\pi \in \Pi$,

$$\begin{aligned} E_{\pi^*}\{\tau \mid X_0 = i\} &= \sum_{n=1}^{\infty} n P_{\pi^*}\{\tau = n \mid X_0 = i\} \\ &= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P_{\pi^*}\{\tau = k \mid X_0 = i\} \\ &= \sum_{n=1}^{\infty} D(i, n, \pi^*) \\ &\geq \sum_{n=1}^{\infty} D(i, n, \pi) = E_\pi\{\tau \mid X_0 = i\}, \qquad i \in S_0. \end{aligned}$$

Hence π^* is an expectation life optimal policy. □

Remark 2.3. From Corollary 2.1 in [2] we know that an expectation life optimal policy in Example 2.2 exists. However there is no optimal policy in Example 2.2 (see Remark 2.2). Hence an expectation life optimal policy need not be an optimal policy in the general case.

3. Existence Results on an n-Optimal Policy and Relevant Results

From Example 2.2 we know that an optimal policy (up to $n \geq 2$) need not exist in the general case. We discuss existence on an n-optimal policy and relevant problems in this section.

Definition 3.1. Let $n \geq 1$ and $\pi \in \Pi$. If $D(i, n, \pi) = D^*(i, n)$ for all $i \in S_0$, then π is called an n-optimal policy.

Let all $A(i)$ $(i \in S_0)$ be finite. It is evident that $\forall n \geq 1$, $\exists f_n \in F$ such that

$$\sum_{j \in S_0} q(j \mid i,\ f_n(i)) D^*(j,n) = \max_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j,n), \qquad i \in S_0. \tag{3.1}$$

Theorem 3.1. *Let $n \geq 2$. Then $\pi^{(n)} = (f_{n-1},\ f_{n-2}, \ldots,\ f_2,\ f_1^\infty) \in \Pi_{\mathrm{m}}^d (\pi^{(2)} = f_1^\infty)$ is an n-optimal policy, where f_k is defined by* (3.1).

Proof. (Apply the induction.) By Theorem 2.1, (3.1) and Theorem 2.2 we have

$$D(i, 2, \pi^{(2)}) = D(i, 2,\ f_1^\infty) = \sum_{j \in S_0} q(j \mid i,\ f_1(i)) = D^*(i, 2), \qquad i \in S_0.$$

Induction hypothesis: $D(i, n, \pi^{(n)}) = D^*(i, n), \quad i \in S_0$.

Obviously, $\pi^{(n+1)} = (f_n,\ f_{n-1},\ f_{n-2}, \ldots,\ f_2,\ f_1^\infty) = (f_n, \pi^{(n)})$. By Theorem 2.1, the induction hypothesis, (3.1) and Theorem 2.2 we have

$$\begin{aligned} D(i, n+1, \pi^{(n+1)}) &= \sum_{j \in S_0} q(j \mid i,\ f_n(i)) D(j, n, \pi^{(n)}) \\ &= \sum_{j \in S_0} q(j \mid i,\ f_n(i)) D^*(j, n) \\ &= D^*(i, n+1), \qquad i \in S_0. \end{aligned}$$

□

Corollary 3.1. *Let $n \geq 2$. If all $A(i) (i \in S_0)$ are finite, then there exists $\pi^{(n)} \in \Pi_{\mathrm{m}}^d$ which is a n-optimal policy, where the definition of $\pi^{(n)}$ can be found in Theorem* 3.1.

Remark 3.1. From Corollary 3.1 we know that there exists $\pi^{(3)} \in \Pi_{\mathrm{m}}^d$ which is a 3-optimal policy in Example 2.2. However, there is no stationary policy which is a 3-optimal policy in Example 2.2. Hence in general case a stationary policy which is an $n(\geq 2)$-optimal policy need not exist.

By the definition of $D^*(i, n)$ it is easy to see that $D^*(i, n) \geq D^*(i, n+1) \geq 0$ for all $i \in S_0$ and $n = 1, 2, \ldots$. So we can define $D^*(i) = \lim_{n \to \infty} D^*(i, n)$, $i \in S_0$.

Theorem 3.2.

(1) *Let $A(i)$ be finite for some $i \in S_0$, then*

$$D^*(i) = \sup_{a \in A(i)} \sum_{j \in S_0} q(j \mid i, a) D^*(j). \tag{3.2}$$

(2) *If S is finite, then* (3.2) *is true for all $i \in S_0$.*

Proof. (1) Let $a \in A(i)$. $\forall \varepsilon > 0$. Because $\sum_{j \in S_0} q(j \mid i, a) \leq 1$, there exists a positive integer N such that $\sum_{j=N+1}^{\infty} q(j \mid i, a) \leq \varepsilon$. It is evident that there exists a positive integer $\tilde{N}$ such that

$$|D^*(j, n) - D^*(j)| \leq \varepsilon, \qquad n \geq \tilde{N}, \quad 1 \leq j \leq N.$$

So, when $n \geq \tilde{N}$,

$$\begin{aligned}
&\left|\sum_{j\in S_0} q(j|i,a)D^*(j,n) - \sum_{j\in S_0} q(j \mid i,a)D^*(j)\right| \\
&\quad \leq \sum_{j\in S_0} q(j \mid i,a)|D^*(j,n) - D^*(j)| \\
&\quad = \sum_{j=1}^{N} q(j|i,a)|D^*(j,n) - D^*(j)| + \sum_{j=N+1}^{\infty} q(j \mid i,a)|D^*(j,n) - D^*(j)| \\
&\quad \leq \varepsilon + 2\varepsilon = 3\varepsilon.
\end{aligned}$$

Hence $\lim_{n\to\infty} \sum_{j\in S_0} q(j \mid i,a)D^*(j,n) = \sum_{j\in S_0} q(j \mid i,a)D^*(j)$.

By Theorem 2.2 we have

$$\begin{aligned}
D^*(i) = \lim_{n\to\infty} D^*(i,n+1) &= \lim_{n\to\infty} \max_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j,n) \\
&= \max_{a\in A(i)} \lim_{n\to\infty} \sum_{j\in S_0} q(j \mid i,a)D^*(j,n) \\
&= \max_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j).
\end{aligned}$$

(2) $\forall \varepsilon > 0$. It is evident that there exists a positive integer N such that

$$|D^*(j,n) - D^*(j)| \leq \varepsilon, \qquad n \geq N, \quad j \in S_0.$$

So, when $n \geq N$,

$$\begin{aligned}
&\left|\sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j,n) - \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j)\right| \\
&\quad \leq \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)|D^*(j,n) - D^*(j)| \\
&\quad \leq \varepsilon, \qquad i \in S_0.
\end{aligned}$$

So

$$\lim_{n\to\infty} \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j,n) = \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j), \qquad i \in S_0.$$

Hence

$$\begin{aligned}
D^*(i) = \lim_{n\to\infty} D^*(i,n+1) &= \lim_{n\to\infty} \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j,n) \\
&= \sup_{a\in A(i)} \sum_{j\in S_0} q(j \mid i,a)D^*(j), \qquad i \in S_0.
\end{aligned}$$

□

We discuss sufficient conditions of $D^*(i) \equiv 0$ below.

Assumption A. There exist a real number $\alpha > 0$ and a positive integer N, such that $P_\pi\{\tau \le N \mid X_0 = i\} \ge \alpha$ for all $i \in S_0$ and $\pi \in \Pi$.

If τ is viewed as the operating life of the system, then Assumption A is satisfied easily in many actual instances. In many actual instances, such as the life of a man or the operating life of a machine and so on, there is a limit to the operating life τ of the system, so Assumption A is satisfied.

Lemma 3.1. *Under Assumption A,* $P_\pi\{\tau > kN \mid X_0 = i\} \le (1-\alpha)^k$, $i \in S_0$, $\pi \in \Pi$, $k = 1, 2, \ldots$.

Proof. (Apply the induction.) Obviously, $P_\pi\{\tau > N \mid X_0 = i\} = 1 - P_\pi\{\tau \le N \mid X_0 = i\} \le 1 - \alpha$, $i \in S_0$, $\pi \in \Pi$. So the proposition (Lemma 3.1) is true for $k = 1$.

Induction hypothesis: $P_\pi\{\tau > nN \mid X_0 = i\} \le (1-\alpha)^n$, $i \in S_0$, $\pi \in \Pi$.

For $i \in S_0$ and $\pi \in \Pi$,

$$
\begin{aligned}
&P_\pi\{\tau > (n+1)N \mid X_0 = i\} \\
&\quad = P_\pi\{X_1 \neq 0, X_2 \neq 0, \ldots, X_{nN} \neq 0, X_{nN+1} \neq 0, \ldots, X_{nN+N} \neq 0 \mid X_0 = i\} \\
&\quad = \sum_{\substack{a_0 \in A(i), a_1 \in A(i_1), \ldots, a_{nN-1} \in A(i_{nN-1}), \\ i_1 \in S_0, i_2 \in S_0, \ldots, i_{nN} \in S_0}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, X_2 = i_2, \ldots, \\
&\qquad\qquad \Delta_{nN-1} = a_{nN-1}, X_{nN} = i_{nN}, X_{nN+1} \neq 0, \ldots, \\
&\qquad\qquad X_{nN+N} \neq 0 \mid X_0 = i\} \\
&\quad = \sum_{\substack{a_0 \in A(i), a_1 \in A(i_1), \ldots, a_{nN-1} \in A(i_{nN-1}), \\ i_1 \in S_0, i_2 \in S_0, \ldots, i_{nN} \in S_0}} P_\pi\{X_{nN+1} \neq 0, \ldots, X_{nN+N} \neq 0 \mid X_0 = i, \\
&\qquad\qquad \Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, \\
&\qquad\qquad X_2 = i_2, \ldots, \Delta_{nN-1} = a_{nN-1}, X_{nN} = i_{nN}\} \\
&\qquad \times P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, X_2 = i_2, \ldots, \Delta_{nN-1} = a_{nN-1}, \\
&\qquad\quad X_{nN} = i_{nN} \mid X_0 = i\}.
\end{aligned}
$$

Let $\pi' = \pi(i, a_0, i_1, a_1, \ldots, i_{nN-1}, a_{nN-1})$ (see Section 1). By the definition of π' it is easy to see that

$$
\begin{aligned}
&P_\pi\{X_{nN+1} \neq 0, \ldots, X_{nN+N} \neq 0 \mid X_0 = i, \Delta_0 = a_0, X_1 = i_1, \\
&\quad \Delta_1 = a_1, \ldots, \Delta_{nN-1} = a_{nN-1}, X_{nN} = i_{nN}\} \\
&\quad = P_{\pi'}\{X_1 \neq 0, \ldots, X_N \neq 0 \mid X_0 = i_{nN}\} \\
&\quad = P_{\pi'}\{\tau > N \mid X_0 = i_{nN}\}, \qquad i_{nN} \in S_0.
\end{aligned}
$$

So, we have by the induction hypothesis,

$$
\begin{aligned}
&P_\pi\{\tau > (n+1)N \mid X_0 = i\} \\
&\quad = \sum_{\substack{a_0 \in A(i), a_1 \in A(i_1), \ldots, a_{nN-1} \in A(i_{nN-1}), \\ i_1 \in S_0, i_2 \in S_0, \ldots, i_{nN} \in S_0}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, X_2 = i_2, \ldots, \\
&\qquad\qquad \Delta_{nN-1} = a_{nN-1}, X_{nN} = i_{nN} \mid X_0 = i\}
\end{aligned}
$$

$$\times P_{\pi'}\{\tau > N \mid X_0 = i_{nN}\}$$
$$\leq (1-\alpha) \sum_{\substack{a_0 \in A(i), a_1 \in A(i_1), \ldots, a_{nN-1} \in A(i_{nN-1}), \\ i_1 \in S_0, i_2 \in S_0, \ldots, i_{nN} \in S_0}} P_\pi\{\Delta_0 = a_0, X_1 = i_1, \Delta_1 = a_1, X_2 = i_2, \ldots,$$
$$\Delta_{nN-1} = a_{nN-1}, X_{nN} = i_{nN} \mid X_0 = i\}$$
$$= (1-\alpha) \sum_{i_1 \in S_0, i_2 \in S_0, \ldots, i_{nN} \in S_0} P_\pi\{X_1 = i_1, X_2 = i_2, \ldots, X_{nN} = i_{nN} \mid X_0 = i\}$$
$$= (1-\alpha) P_\pi\{\tau > nN \mid X_0 = i\}$$
$$\leq (1-\alpha)(1-\alpha)^n = (1-\alpha)^{n+1}. \qquad \square$$

Theorem 3.3. *Under Assumption A we have $D^*(i) = 0$ for all $i \in S_0$.*

Proof. By Lemma 3.1,

$$D(i, kN, \pi) = P_\pi\{\tau \geq kN \mid X_0 = i\}$$
$$\leq P_\pi\{\tau > (k-1)N \mid X_0 = i\}$$
$$\leq (1-\alpha)^{k-1}, \qquad i \in S_0, \quad \pi \in \Pi, \quad k = 2, 3, \ldots.$$

So, $D^*(i, kN) \leq (1-\alpha)^{k-1}$, $i \in S_0$, $k = 2, 3, \ldots$. Obviously, $0 \leq 1-\alpha < 1$. Hence $D^*(i) = \lim_{n\to\infty} D^*(i, kN) = 0$, $i \in S_0$. $\square$

Corollary 3.2. *If there exists $\beta > 0$ such that $q(0 \mid i, a) \geq \beta$ for all $i \in S_0$ and $a \in A(i)$, then the assumption A is true, so $D^*(i) = 0$ for all $i \in S_0$.*

Proof.

$$P_\pi\{\tau \leq 1 \mid X_0 = i\} = P_\pi\{X_1 = 0 \mid X_0 = i\}$$
$$= \sum_{a \in A(i)} \pi_0(a \mid i) q(0 \mid i, a)$$
$$\geq \beta, \qquad i \in S_0, \quad \pi = (\pi_0, \pi_1, \ldots) \in \Pi.$$

That is, Assumption A is true. From Theorem 3.3 we know that $D^*(i) = 0$ for all $i \in S_0$. $\square$

In the case of S and all $A(i)$ $(i \in S)$ being countable Assumption A is proposed. When S and all $A(i)$ $(i \in S)$ are finite we propose the following assumption:

Assumption B. The following are true:

(1) S and all $A(i)$ $(i \in S)$ are finite.
(2) $q(0 \mid 0, a) = 1$ for all $a \in A(0)$.
(3) $P_{f^\infty}\{\exists t > 0$, such that $X_t = 0 \mid X_0 = i\} > 0$ for all $i \in S$ and $f^\infty \in \Pi_s^d$.

Assumption B is from [1, p. 33]. Note that we do not require $q(0 \mid 0, a) = 1$ for all $a \in A(0)$ in Assumption A.

Theorem 3.4. *Under Assumption B we have $D^*(i) = 0$ for all $i \in S_0$.*

Proof. From [1, p. 33] we know that $\lim_{t\to\infty}\sup_{\pi\in\Pi} P_\pi\{X_t \neq 0 \mid X_0 = i\} = 0,\ i \in S$. For $i \in S_0,\ \forall\varepsilon > 0$, there exists a positive integer N such that $\sup_{\pi\in\Pi} P_\pi\{X_t \neq 0 \mid X_0 = i\} < \varepsilon, t \geq N$. So, when $t \geq N$,

$$D^*(i, t+1) \leq \sup_{\pi\in\Pi} P_\pi\{\tau > t \mid X_0 = i\} \leq \sup_{\pi\in\Pi} P_\pi\{X_t \neq 0 \mid X_0 = i\} < \varepsilon.$$

Hence $D^*(i) = \lim_{t\to\infty} D^*(i, t) = 0$. □

Remark 3.2. It is evident that Example 2.2 satisfies Assumptions A and B, but there is no optimal policy (up to 3) in Example 2.2(see Remark 2.2). Hence, an optimal policy (up to $n \geq 2$) need not exist under Assumption A or B.

Definition 3.2. Let $n \geq 1$ and $\varepsilon \geq 0$. Let $\pi^{(n,\varepsilon)} \in \Pi$ and $\pi^{(k)} \in \Pi,\ k = 1, 2, 3, \ldots, n$. $\{\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}, \pi^{(n,\varepsilon)}\}$ is called a sequence of (n, ε)-optimal policies if $\pi^{(k)}$ is a k-optimal policy for $k = 1, 2, 3, \ldots, n$ and $D(i, k, \pi^{(n,\varepsilon)}) \geq D^*(i, k) - \varepsilon$ for all $i \in S_0$ and $k = n+1, n+2, \ldots$.

Theorem 3.5. *Let Assumption A be true. Given $\varepsilon > 0$. If all $A(i)$ $(i \in S_0)$ are finite, then there exists $\{\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}, \pi^{(n,\varepsilon)}\}$ which is a sequence of (n, ε)-optimal policies, where $\pi^{(k)} \in \Pi_{\mathrm{m}}^d$, $1 \leq k \leq n$, and $\pi^{(n,\varepsilon)}$ is an arbitrary policy in Π.*

Proof. From the proof of Theorem 3.3 we know that $D^*(i, kN) \leq (1-\alpha)^{k-1}, i \in S_0$, $k = 2, 3, \ldots$. For $\varepsilon > 0$, it is evident that there exists a positive integer $k_0 \geq 2$ such that $D^*(i, k_0 N) \leq (1-\alpha)^{k_0-1} \leq \varepsilon, i \in S_0$.

Let $n = k_0 N$. By (3.1) we can find $f_k \in F,\ 1 \leq k \leq n-1$. We take arbitrarily $\pi^{(1)} \in \Pi_{\mathrm{m}}^d$. Let $\pi^{(k)} = (f_{k-1}, f_{k-2}, \ldots, f_2, f_1^\infty)(\pi^{(2)} = f_1^\infty), 2 \leq k \leq n$. From Th.3.1 we know that $\pi^{(k)} \in \Pi_{\mathrm{m}}^d$ is a k-optimal policy, $1 \leq k \leq n$.

Let $\pi^{(n,\varepsilon)}$ be an arbitrary policy in Π, then

$$D(i, k, \pi^{(n,\varepsilon)}) \leq D^*(i, k) \leq D^*(i, n) \leq \varepsilon, \qquad i \in S_0, \quad k > n.$$

That is, $D(i, k, \pi^{(n,\varepsilon)}) \geq D^*(i, k) - \varepsilon,\ i \in S_0,\ k = n+1, n+2, \ldots$. Hence $\{\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}, \pi^{(n,\varepsilon)}\}$ is a sequence of (n, ε)-optimal policies. □

Theorem 3.6. *Let S and all $A(i)(i \in S_0)$ be finite. Given $\varepsilon > 0$. If $D^*(i) = 0$ for all $i \in S_0$, then there exists $\{\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}, \pi^{(n,\varepsilon)}\}$ which is a sequence of (n, ε)-optimal policies, where $\pi^{(k)} \in \Pi_{\mathrm{m}}^d$, $1 \leq k \leq n$, and $\pi^{(n,\varepsilon)}$ is an arbitrary policy in Π.*

Proof. For $\varepsilon > 0$, it is easy to see that there exists a positive integer $N \geq 2$ such that $D^*(i, N) \leq \varepsilon, i \in S_0$. Let $n = N$. By (3.1) we can find $f_k \in F,\ 1 \leq k \leq n-1$. The remainder of the proof is similar to the proof of Theorem 3.5. □

Corollary 3.3. *Let Assumption B be true. Given $\varepsilon > 0$. Then there exists $\{\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}, \pi^{(n,\varepsilon)}\}$ which is a sequence of (n, ε)-optimal policies, where $\pi^{(k)} \in \Pi_{\mathrm{m}}^d$, $1 \leq k \leq n$ and $\pi^{(n,\varepsilon)}$ is an arbitrary policy in Π.*

Remark 3.3. To sum up, when S and all $A(i)$ ($i \in S_0$) are finite, under Assumption A or B, for an arbitrary $\varepsilon > 0$ we can find a sequence of (n, ε)-optimal policies in finite steps. (These algorithms can be found in the proofs of Theorem 3.5 and 3.6.) In these cases we can say that the optimization problem discussed in this paper has been solved.

References

1. Derman C (1970), Finite State Markovian Decision Processes, Academic Press, New York
2. Liu Jianyong and Liu Ke (1992), Markov decision programming—the first-passage model with denumerable state space, Systems Sci Math Sci 5(4):340–351
3. Liu Jianyong and Liu Ke (1997), Markov decision programming—the moment optimal problem for the first-passage model, J Austral Math Soci Ser B 38:542–562
4. Liu Difen, Liu Jianyong and Liu Ke (1994), Partially observable Markov decision programming: first passage problem (in Chinese), Acta Math Appl Sinica 17(1):44–58
5. Bertsekas DP (1976), Dynamic Programming and Stochastic Control, Academic Press, New York
6. Eaton JH and Zadeh LA (1962), Optimal pursuit strategies in discrete state probabilistic Systems, Trans ASME Ser D J Basic Eng 84:23–29
7. Puterman ML (1994), Markov Decision Processes: Discrete Stochastic Dynamic Programming, Wiley, New York
8. Lin Yuanlie, Tomkins RJ and Chunglie Wang (1994), Optimal odels for the first arrival time distribution function in continuous time—with a special case, Acta Math Appl Sinica 10(2): 194–212
9. White DJ (1993), Minimizing a threshold probability in discounted Markov decision processes, J Math Anal Appl 173:634–646
10. Filar JA, Krass D and Ross KW (1995), Percentile performance criteria for limiting average Markov decision processes, IEEE Trans Automat Control 40(1): 2–10

Accepted 24 *July* 2000. *Online publication* 12 *April* 2001.