Continuous Time Markov Decision Processes with Discounted Moment Criterion

Qiying Hu

Management Department, Xidian University, Xian, People’s Republic of China

Submitted by Augustine O. Esogbue

Received December 27, 1990

Time Markov decision processes with countable states and actions continuous are discussed with the criterion of discounted moment optimality. They are transformed into a sequence of discrete time Markov decision processes with the criterion of discounted expected total rewards.

1. INTRODUCTION

In Markov decision processes (MDP), the discounted criterion is the usual criterion. Under this criterion, a policy with the largest discounted expected total rewards will be optimal. But it considers only the expectation value of the discounted total rewards and not the risk of randomness. Thus, as its improvement, the discounted moment criterion is presented. This criterion means that for two policies, if their mean rewards are equal, then the policy with smaller variance will be better; if they have the same mean rewards and variances, then the policy with higher 3rd moment will be better, et al. The first author to discuss the discounted moment criterion is Jaquette [1], who dealt with the discrete time Markov decision processes (DTMDP) with finite states and actions and proved that there exist a policy \( f \in F \) and a constant \( \beta_0 \in (0,1) \) such that for every discount factor \( \beta \in (\beta_0,1) \), \( f \) is moment optimal within the stationary policies. Sobel [2] also considered the discounted moment criterion for DTMDP and semi-Markov decision processes with finite states and actions, but only presented some formulae of the variance and higher moments of the discounted total rewards for each stationary policies. Hu [3] discussed the DTMDP model with countable states and actions and proved that the moment optimal problem can be transformed into a sequence of DTMDP with the discounted criterion.
In this paper, we study the discounted moment criterion for continuous time Markov decision processes (CTMDP) with countable states and actions. By using the methods presented in Hu [3, 4] and the properties of optimal policies, we transform the moment optimal problem into a sequence of discrete time Markov decision processes with the discounted criterion.

The remainder of the paper is organized as follows. In Section 2, the model is formulated and some preliminary results are proved. In Section 3, the moment optimal problem is transformed into a sequence of DTMDP with discounted criterion.

2. MODEL AND PRELIMINARIES

The continuous time Markov decision processes model discussed here is \( \{S, (A(i), i \in S), q, r, \alpha, (\mu_{\alpha})\} \), where the state space \( S \) and action sets \( A(i) \) available at state \( i \) are all countable. \( (q_{ij}(a)) \) is the state transition rate family; that is, if the system is in state \( i \) at time \( t \) and action \( a \in A(i) \) is used in time interval \( [t, t + \Delta t] \) for \( \Delta t \) small enough, then the probability that the system will transfer to state \( j \) at time \( t + \Delta t \) is \( \delta_{ij} + q_{ij}(a) \Delta t + o(\Delta t) \). We assume that \( q_{ij}(a) \) satisfies \(-\infty < q_{ij}(a) \leq 0, q_{ij}(a) \geq 0 \) for \( j \neq i; \sum_j q_{ij}(a) = 0 \) for \( i \in S \) and \( a \in A(i) \); the reward rate function \( r(i, a) \) is uniformly bounded, i.e., there exists a positive constant \( M \) such that \( |r(i, a)| \leq M \) for all \( i \in S \) and \( a \in A(i) \). \( \alpha > 0 \) is the discounting factor; \( \mu_{\alpha} \) is the \( n \)th moment of the discounted total rewards which will be defined in (1) below for \( n \geq 1 \).

We consider only the stochastic stationary policies, i.e., the policies having form \( \pi_{\alpha}^{-} \). Using \( \pi_{\alpha}^{-} \) means that when the system is in state \( i \) the action taken is according to the probability distribution \( \pi_{\alpha}^{-}(\cdot | i) \) on \( A(i) \). For simplicity, denote \( \pi_{\alpha}^{-} \) by \( \pi_{\alpha} \). Let \( \Pi_{\alpha} \) be the policies set. If for each \( i \in S, \pi_{\alpha}(\cdot | i) \) is degenerative at \( f(i) \) for some \( f(i) \in A(i) \), then we call \( \pi_{\alpha} \) a stationary policy and write it as \( f \). Such a policy in fact corresponds to a decision function, the set of which is \( F := \times_i A(i) \).

Given \( \pi_{\alpha} \), we define a vector \( r(\pi_{\alpha}) = (r(\pi_{\alpha}),) \) and a matrix \( Q(\pi_{\alpha}) = (q(\pi_{\alpha}),) \), respectively, by

\[
\begin{align*}
    r(\pi_{\alpha})_i &= \sum_{a \in A(i)} r(i, a) \pi_{\alpha}(a | i), \quad i \in S, \\
    q(\pi_{\alpha})_{ij} &= \sum_{a \in A(i)} q_{ij}(a) \pi_{\alpha}(a | i), \quad i, j \in S.
\end{align*}
\]

Here, \( r(\pi_{\alpha})_i \) and \( q(\pi_{\alpha})_{ij} \) are, respectively, the reward rate at state \( i \) and the state transition family of the system controlled by policy \( \pi_{\alpha} \).
In order to ensure that the CTMDP model is well defined, the following three assumptions are assumed to be true throughout this paper.

**Assumption 1.** \( \lambda(i) = \sup(-q_i(a); i \in S, a \in A(i)) < \infty \), for all \( i \in S \).

**Assumption 2.** For every \( \pi_o \in \Pi_S \), the \( Q(\pi_o) \)-process \( \{Q(\pi_o), t \geq 0\} \) exists uniquely and satisfies (where \( I \) is the unit matrix)

\[
\frac{d}{dt}P(\pi_o, t) = P(\pi_o, t)Q(\pi_o) = Q(\pi_o)P(\pi_o, t), \quad t \geq 0, \\
P(\pi_o, 0) = I.
\]

**Assumption 3.** For every bounded vector \( u \in \mathbb{R}^s \) and \( \pi_o \in \Pi_S \), we have

\[
(P(\pi_o, t)Q(\pi_o))u = P(\pi_o, t)(Q(\pi_o)u), \quad t \geq 0.
\]

Certainly, if \( q \) is uniformly bounded, all the above three assumptions will be satisfied. Song [5] presented some assumptions about non-uniformly bounded \( q \), which imply Assumptions 2 and 3 above.

In order to define the objective function, let \( X(t) \) and \( \Delta(t) \) be the system’s state and action taken at time \( t \), respectively. Define the discounted total rewards \( W := \int_0^\infty \exp(-\alpha t)\pi(X(t), \Delta(t))dt \). Thus, the objective function \( \{\mu_n\} \) can be defined as follows:

\[
\mu_n(\pi_o, i) = E_{\pi_o}[W^n \mid X(0) = i], \quad \pi_o \in \Pi_S, i \in S, n \geq 0. \tag{1}
\]

When the system is controlled by a policy \( \pi_o \), \( X(t) \) is piecewise constant under \( \pi_o \). Thus \( W \) is the sum of an infinite series, which implies that \( W \) is well defined and \( |W| \leq M/\alpha \). Moreover, \( \mu_n(\pi_o, i) \) is also well defined and \( |\mu_n(\pi_o, i)| \leq (M/\alpha)^n \). Denote

\[
\mu^{(n)}(\pi_o) = (\mu_1(\pi_o), -\mu_2(\pi_o), \ldots, (-1)^{n+1}\mu_n(\pi_o)), \\
\mu^{(\infty)}(\pi_o) = (\mu_1(\pi_o), -\mu_2(\pi_o), \ldots, (-1)^{n+1}\mu_n(\pi_o), \ldots).
\]

Now we give the definition of moment optimality. A policy \( \pi_o^* \in \Pi_S \) is said to be (\( \infty \)) moment optimal if \( \mu^{(\infty)}(\pi_o^*) \mu^{(\infty)}(\pi_o) \) for all \( \pi_o \in \Pi_S \), i.e., \( \pi_o^* \) lexicographically maximizes the sequence of the moments of return vectors with alternating signs. In addition, a policy \( \pi_o^* \) is said to be (\( n \)) moment optimal if \( \mu^{(n)}(\pi_o^*) \mu^{(n)}(\pi_o) \) for all \( \pi_o \in \Pi_S \).

Let \( \Pi^a = \Pi_S \), \( F^a = F \) and for \( 1 \leq n \leq \infty \)

\[\Pi^a = \{\pi_o \mid \pi_o \in \Pi_S, \pi_o \text{ is (} n \text{) moment optimal}\},\]
\[
F^n = \{ f \mid f \in F, f \text{ is } (n) \text{ moment optimal} \}.
\]

\(\Pi^n\) and \(F^n\) represent the sets of \((n)\) moment optimal policies and \((n)\) moment optimal stationary policies, respectively.

If we set \(R_n(\pi_o) = (R_n(\pi_o)_i)\) with \(R_n(\pi_o)_i = nr(\pi_o)_i, \mu_{n-1}(\pi_o, i)\) for \(i \in S\), then we have the following theorem.

**Theorem 1.** For \(n \geq 1, \pi_o \in \Pi_n, \mu_n(\pi_o)\) is the unique bounded solution of the following equation and \(\mu_n(\pi_o) = [n \alpha]^{-1}Q(\pi_o)R_n(\pi_o)\):

\[
n \alpha u = R_n(\pi_o) + Q(\pi_o)u. \tag{2}
\]

**Proof.** For every \(\pi_o \in \Pi_n, \{X(t), t \geq 0\}\) is a continuous time homogeneous Markov process with rewards under \(\pi_o\). It can also be considered as a semi-Markov process with rewards. Then, using the symbols presented in Sobel [2], its kernel

\[
Q_{ij}(t) = \left[ -q(\pi_o)_{ij}/q(\pi_o)_{ii} \right] \cdot \left[ 1 - \exp\left(q(\pi_o)_{ii}t\right) \right] (1 - \delta_{ij})
\]

and the rewards \(R_{ij}(t) = r(\pi_o)_{ij}t\). Thus from the results in [2] we can get

\[
\mu_n(\pi_o, i) = \sum_{j \neq i} q(\pi_o)_{ij}/(n \alpha - q)(\pi_o)_{ij} \mu_n(\pi_o, j)
\]

\[
+ \sum_{k=0}^{n-1} \frac{n!}{k!} r(\pi_o)_{ij}^{n-k} \left[ 1 - k! \prod_{l=k}^{n-1} (1 - q(\pi_o)_{ii}) \right]^{-1}
\]

\[
\times \sum_{j \neq i} q(\pi_o)_{ij} \mu_k(\pi_o, j), \quad \text{for } n \geq 0, i \in S.
\]

Multiplying the above formula by \(n \alpha - q(\pi_o)_{ij}\), we obtain that

\[
n \alpha \mu_n(\pi_o, i) = \sum_{j \neq i} q(\pi_o)_{ij} \mu_n(\pi_o, j)
\]

\[
+ \sum_{k=0}^{n-1} \frac{n!}{k!} r(\pi_o)_{ij}^{n-k} \left[ 1 - k! \prod_{l=k}^{n-1} (1 - q(\pi_o)_{ii}) \right]^{-1}
\]

\[
\times \sum_{j \neq i} q(\pi_o)_{ij} \mu_k(\pi_o, j), \quad \text{for } n \geq 0, i \in S. \tag{3}
\]

Using the induction method we will prove that \(\mu_n(\pi_o)\) is a solution of (2).

For \(n = 1\), it is obvious. Assume that for some \(n \geq 1, \mu_k(\pi_o)\) is a solution of (2) for all \(k \leq n\), i.e.,

\[
k \alpha u = R_k(\pi_o) + Q(\pi_o)u, \quad \text{for } k \leq n,
\]
or
\[
\sum_{j \neq i} q(\pi_o)_{ij} \mu_k(\pi_o, i)
= [k \alpha - q(\pi_o)_{ii}] \mu_k(\pi_o, i) - kr(\pi_o)_{i} \mu_{k-1}(\pi_o, i), \quad \text{for } i \in S, k \leq n.
\]

Then, by (3), the induction assumption, one can get that for \(n + 1\),
\[
(n + 1) \alpha \mu_{n+1}(\pi_o, i) - \sum_j q(\pi_o)_{ij} \mu_{n+1}(\pi_o, j)
= \sum_{k=0}^{n} \frac{(n + 1)!}{k!} r(\pi_o)_{i}^{n+1-k} \prod_{l=k}^{n} (l \alpha - q(\pi_o)_{ii})^{-1} \sum_{j \neq i} q(\pi_o)_{ij} \mu_k(\pi_o, j)
\]
\[
= \sum_{k=0}^{n} \frac{(n + 1)!}{k!} r(\pi_o)_{i}^{n+1-k} \prod_{l=k}^{n} (l \alpha - q(\pi_o)_{ii})^{-1} \left[ (k \alpha - q(\pi_o)_{ii}) \mu_k(\pi_o, i) - kr(\pi_o)_{i} \mu_{k-1}(\pi_o, i) \right]
\]
\[
= \sum_{k=0}^{n} \frac{(n + 1)!}{k!} r(\pi_o)_{i}^{n+1-k} \prod_{l=k+1}^{n} (l \alpha - q(\pi_o)_{ii})^{-1} \mu_k(\pi_o, i)
- \sum_{k=0}^{n-1} \frac{(n + 1)!}{k!} r(\pi_o)_{i}^{n+1-k} \prod_{l=k+1}^{n} (l \alpha - q(\pi_o)_{ii})^{-1} \mu_k(\pi_o, i)
= (n + 1)r(\pi_o)_{i} \mu_n(\pi_o, i) = R_n(\pi_o)_{i}.
\]

By Lemma 2 in [4], the bounded solution of (2) is unique and
\[
\mu_n(\pi_o) = [n \alpha I - Q(\pi_o)]^{-1} R_n(\pi_o), \quad \text{for } n \geq 1.
\]

This completes the proof.

Remark. The relevant results in [2] are about a finite MDP, but it is easy to generalize the results into the countable MDP.
From (2) and (3), we have

\[
R_n(\pi_o)_i = \sum_{k=0}^{n-1} \frac{n!}{k!} r(\pi_o)_i^{n-k} \left[ \prod_{l=k}^{n-1} (l \alpha - q(\pi_o)_i) \right]^{-1} \times \sum_{j \neq i} q(\pi_o)_{ij} \mu_k(\pi_o, j).
\]  

(4)

The meaning of \( R_n(\pi_o) \) is as follows. Consider a semi-Markov process with rewards, where the semi-Markov process is exactly the system controlled by \( \pi_o \), but its reward rate is \( R_p(\pi_o) \), at state \( i \). Then the expected discounted total rewards, with discounting factor \( n \alpha \), received by this system is \( R_p(\pi_o) \).

Denote \( N_1 = \sup\{n \mid n \geq 1, \Pi^n \neq \emptyset\} \), \( \mu^*_n = \sup\{(-1)^{n+1} \mu_n(\pi_o) \mid \pi_o \in \Pi^{n+1}\} \) for \( n < N_1 + 2 \). Thus, by the definition of \((n)\) moment optimality, we can immediately obtain the following lemma.

**Lemma 1.** For \( n < N_1 + 1 \), \( \pi_o \in \Pi^n \) iff \( \mu_k(\pi_o) = (-1)^{k+1} \mu^*_n \) for \( k = 1, 2, \ldots, n \).

3. TRANSFORMATION

Now, we will define a sequence of discounted DTMDP models

\[
\{S, (A_n(i), i \in S), p, r_n, \beta_n, V_n\}, \quad 1 \leq n < N_1 + 1
\]

(denoted by DTMDP\((n)\)), such that finding the \((n)\) moment optimal policies is equivalent to finding the optimal policies for DTMDP\((n)\). Thus the defined model DTMDP\((n)\) should satisfy that its policy set is \( \Pi^{n-1} \) and its discounted objective is \((-1)^{n+1} \mu_n(\pi_o)\), which together with Theorem 1 motivates us to define the discounted DTMDP\((n)\) as follows.

The state space \( S \) is the same as that in the CTMDP model; \( A_n(i) \), the action sets available at state \( i \), will be defined by the introduction method below; the state transition probabilities \( p_{ij}(a) = q_{ij}(a)/\lambda(i) + \delta_{ij} \); the reward function

\[
r_n(i, a) = -nr(i, a) \mu_{n-1}(i)/(\lambda(i) + n \alpha); \quad \beta_n(i) = \lambda(i)/(\lambda(i) + n \alpha)
\]

is the state-dependent discounting factor, which means that if the state is \( i \) at time period \( m \), then the unit reward obtained at the time period \( m + 1 \) will be only worth \( \beta_n(i) \) at \( m \).

Denote the decision function set \( F_n : \times i A_n(i), \Pi(n) \) the set of stochastic stationary policies of DTMDP\((n)\), which we only consider here. From
the definition of \( A_n(i) \) below, \( A_n(i) \subset A(i) \), and the forms of stochastic stationary policies for DTMDP and CTMDP are the same, so we have \( \Pi_n(n) \subset \Pi_i \) for all \( n \geq 1 \). We define the discounted objective function of DTMDP by

\[
V_n(\pi_o, i) = E_{\pi_o} \left[ r_n(X_o, \Delta_o) + \sum_{k=1}^{\infty} \beta_n(X_o) \beta_n(X_1) \right. \\
\ldots \beta_n(X_{k-1}) r_n(X_k, \Delta_k) | X_o = i \] 

for \( \pi_o \in \Pi_i(n), i \in S \),

where \( X_k, \Delta_k \) denote the state and action taken at time period \( k \), respectively. Define

\[
V^*_n(i) = \sup \{ V_n(\pi_o, i) \mid \pi_o \in \Pi_i(n) \}, \quad i \in S.
\]

Now we define \( A_n(i) \) by induction. For \( n = 1 \), define \( A_1(i) = A(i) \) for \( i \in S \). Assume that \( A_n(i) \subset A(i) \) (\( i \in S \)) are well defined for some \( n < N_1 + 1 \); then the DTMDP is well defined. Thus, define

\[
A^*_n(i) = \{ a \mid a \in A_n(i) \}
\]

and \( r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a) V^*_n(j) = V^*_n(i) \), \( i \in S \).

It will be proved (see Lemma 3 below) that any optimal policy of DTMDP must choose the actions from \( A^*_n(i) \) and vice versa. So, we should call \( A^*_n(i) \) the optimal action set available at state \( i \) for DTMDP. If \( A^*_n(i) \neq \phi \) for every \( i \in S \), then there exist optimal policies for DTMDP and we need to define DTMDP \( n + 1 \) to compare them. Define \( A^*_{n+1}(i) = A^*_n(i) \) for \( i \in S \); DTMDP \( n + 1 \) is thus well defined. If there is \( i \in S \) such that \( A^*_n(i) = \phi \), then let \( N = n \) and stop defining DTMDP \( n + 1 \).

If \( N_1 < \infty \) and \( A^*_{N_1 + 2}(i) = \phi \) for all \( i \), then let \( N = N_1 + 1 \) and do not define DTMDP \( n \geq N + 1 \). If \( N_1 > \infty \) and \( A^*_n(i) \neq \phi \) for all \( i \in S, n \geq 1 \), then let \( N = \infty \).

By induction, we have defined a sequence of models (DTMDP \( n: n < N + 1 \)). But here sup, \( \beta, i \) may be equal to one, so the existence and boundedness of \( V_n(\pi_o) \) should be further discussed. First we prove the following lemma.

**Lemma 2.** Suppose \( n < N + 1 \); then:

(i) For \( \pi_o \in \Pi_i(n), V_n(\pi_o) \) exists and is the unique bounded solution of the equation

\[
u_n(i) = r_n(\pi_o)_i + \beta_n(i) \sum_j P(\pi_o)_i j \nu_n(j), \quad i \in S, \quad (5)
\]

where \( r_n(\pi_o) \) and \( P(\pi_o) \) are defined exactly as \( r(\pi_o) \) and \( Q(\pi_o) \), respectively.
(ii) \( V_n^* \) is the unique bounded solution of the optimality equation of DTMDP \( (n) \)

\[
V_n(i) = \sup_{a \in A_n(i)} \left\{ r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a) V_n(j) \right\}, \quad i \in S. \quad (6)
\]

(iii) \( f \) attains the supremum of the right-hand side of (6) iff \( f \) is optimal for DTMDP \( (n) \).

Proof. First, we define a sequence of CTMDP models

CTMDP \( (n): (S, (A_n(i), i \in S), q, r_n h, V_n \alpha) \) \( (n < N + 1) \). Here, the state space \( S \), the action sets \( A_n(i) \), and the state transition rate family \( q \) are as before; the reward rate function \( r_n(i, a) = -nr(i, a) \mu_n^{-1}(i) \); \( V_n \alpha(\pi_o) \) are the expected discounted total rewards of \( \pi_o \in \Pi_s(n) \) with the discount factor \( \alpha \), i.e.,

\[
V_n \alpha(\pi_o) = \int_0^\infty \exp(-\alpha t) P(\pi_o, t) \tilde{r}_n(\pi_o) dt, \quad \pi_o \in \Pi_s(n).
\]

From Assumption 2, the \( Q(\pi_o) \)-process \( \{P(\pi_o, t), t \geq 0\} \) exists uniquely for every \( \pi_o \in \Pi_s(n) \). Because \( r_n(i, a) \) is uniformly bounded in \( i, a \), \( V_n \alpha(\pi_o, i) \) is also uniformly bounded in \( \pi_o, i \). By Theorem 3 in Hu [4], the CTMDP \( (n) \) is equivalent to the DTMDP \( (n) \) for \( n < N + 1 \) in the following meaning:

(a) For \( \pi_o \in \Pi_s(n), V_n \alpha(\pi_o) = V_n^*(\pi_o) \) and is the unique bounded solution of (5).

(b) Eq. (6) is equivalent to the optimality equation of CTMDP \( (n) \)

\[
\alpha V_n^*(i) = \sup_{a \in A_n(i)} \left\{ r_n(i, a) + \sum_j q_{ij}(a) V_n^*(j) : a \in A_n(i) \right\}, \quad i \in S. \quad (7)
\]

and \( V_n^* \) is the unique bounded solution of (6) or (7).

(c) \( f \) attains the supremum in (6) or (7) iff \( f \) is optimal for DTMDP \( (n) \) and CTMDP \( (n) \). This completes the proof.

When \( \lambda(i) \) is uniformly bounded, sup, \( \beta(i) < 1 \), and Lemma 2 is the standard result. But when \( \lambda(i) \) is non-uniformly bounded, for example, \( \lim_{i \to \infty} \lambda(i) = \infty \), \( \lim_{i \to \infty} \beta(i) = 1 \). Fortunately, in this case, \( \lim_{i \to \infty} r_n(i, a) = 0 \), so \( V_n(\pi_o) \) may still exist. In Lemma 2, we proved that for the particular structure of DTMDP \( (n) \), \( V_n(\pi_o) \) exists and is uniformly bounded. The following lemma about DTMDP \( (n) \) is also a generalization of the results in [6] with constant discount factor \( \beta < 1 \).

Lemma 3. For \( n < N, \pi_o \in \Pi_s(n) \) is optimal for DTMDP \( (n) \) iff \( \pi_o(\Lambda_n(i) \mid i) = 1 \) for all \( i \in S \).
Proof. Suppose $\pi_0 \in \Pi_1(n)$ is optimal for DTM DP($n$), i.e., $V_n(\pi_0) = V_n^*$. From (i) and (ii) of Lemma 2 we get

$$V_n^*(i) = V_n(\pi_0, i) = \sum_{a \in A_n(i)} \pi_0(a | i) \times \left[ r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a)V_n^*(j) \right]$$

$$= \sum_{a \in A_n(i)} \pi_0(a | i) \left[ r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a)V_n^*(j) \right].$$

Therefore by the definition of $A^*_n(i)$

$$0 = \sum_{a \in A_n(i)} \pi_0(a | i) \left[ V_n^*(i) - \left( r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a)V_n^*(j) \right) \right]$$

$$= \sum_{a \in A_n(i) \setminus A^*_n(i)} \pi_0(a | i) \left[ V_n^*(i) - \left( r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a)V_n^*(j) \right) \right].$$

Because

$$V_n^*(i) - \left[ r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a)V_n^*(j) \right]$$

is nonnegative, $\pi_0(a | i) = 0$ for all $a \in A_n(i) \setminus A^*_n(i)$ and $i \in S$, i.e., $\pi_0(A^*_n(i) | i) = 1$ for all $i \in S$.

Now if for some $\pi_0$, $\pi_0(A^*_n(i) | i) = 1$ for all $i$, then $\pi_0 \in \Pi_1(n)$ and

$$V_n^*(i) = \pi_0(A^*_n(i) | i) V_n^*(i)$$

$$= \sum_{a \in A_n(i)} \pi_0(a | i) \left[ r_n(i, a) + \beta_n(i) \sum_j p_{ij}(a)V_n^*(j) \right]$$

$$= r_n(\pi_0)_i + \beta_n(i) \sum_j p(\pi_0)_j V_n^*(j), \quad i \in S.$$ 

By this and Lemma 2 we have $V_n(\pi_0) = V_n^*$; i.e., $\pi_0$ is optimal for DTM DP($n$). This completes the proof.

Lemma 3 characterizes a property of the optimal policies. One can see [8] for details about the properties of the optimal policies in the usual DTM DP with discounted criterion.

When $\pi_0 = f$, the above lemma is exactly (iii) of Lemma 2.

By Lemma 3, the following corollary is obvious.

**Corollary.** For $n < N$, $\Pi_1(n + 1)$ is the set of optimal policies for DTM DP($n$).

Now we can prove the main theorem of this paper.
Theorem 2. For \( n < N + 1 \), we have

(i) for \( \pi_o \in \Pi_o(n) \), \( V_n(\pi_o) = (-1)^{n+1} \mu_n(\pi_o) \);
(ii) \( V_n^* = \mu_n^* \);
(iii) \( \Pi^o = \Pi_o(n + 1) \), \( F_n = F_{n+1} \) (for \( n < N \)).

Proof. The theorem will be proved by the induction method. For \( n = 1 \):

(i) \( \Pi_o(1) = \Pi_o \), Theorem 1, and Lemma 2, one can get immediately that \( V_1(\pi_o) = \mu_1(\pi_o) \);
(ii) it follows from (i), \( \Pi^o = \Pi_o(1) \), and the definition of \( V_1^*, \mu_1^* \), that \( V_1^* = \mu_1^* \);
(iii) by Lemma 1, Lemma 2, and (i), (ii), for \( n = 1 \), one has
\[
\text{if } f \in F^1 \text{ iff } f \text{ is (1) moment optimal (by definition)}
\]
\[
\text{iff } \mu_1(f) = \mu_1^* \text{ (by Lemma 1)}
\]
\[
\text{iff } V_1^*(f) = V_1^* \text{ (by (i) and (ii) for } n = 1)\]
\[
\text{iff } f \text{ is optimal for DTDMP (1)}
\]
\[
\text{iff } f(i) \in A_1^*(i) \text{ for all } i \text{ (by Lemma 3)}
\]
\[
\text{iff } f \in F_2. \text{ (by the definition of } F_2)\]

So \( F^1 = F_2, \Pi^1 = \Pi_o(2) \) can be proved in a similar way.

Assume that for some \( k < N_o(i) \)–(iii) are true for \( n = 1, 2, \ldots k \).

(i) By induction assumption, \( \Pi^k = \Pi_o(k + 1) \). So for \( \pi_o \in \Pi^k \), it follows from Lemma 1 that \( \mu^k(\pi_o) = (-1)^{k+1} \mu^k_o \). Now assume that \( u = (u_i, i \in S) \) is a bounded vector. Then for \( n \in \Pi_o(k + 1) \), \( u \) is a bounded solution of (2) for \( n = k + 1 \), i.e.,
\[
(k + 1) \alpha u = R_{k+1}(\pi_o) + Q(\pi_o) u,
\]
\[
\text{iff } (k + 1) \alpha u_i = (k + 1) r(\pi_o)(i) (-1)^{k+1} \mu^k_o(i) + \sum_j q(\pi_o)_{ij} \mu_j, \quad i \in S,
\]
\[
\text{iff } (k + 1) \alpha u_i = \sum_{a \in A_i(o)} \pi_o(a \mid i) \left\{ (k + 1) r(i, a) (-1)^{k+1} \mu^k_o(i)
\]
\[
\quad + \sum_j q_{ij}(a) u_j \right\}, \quad i \in S.
\]
\[
\text{iff } (-1)^{k+2} u_i [\lambda(i) + (k+1) \alpha] = \sum_{a \in A_i(o)} \pi_o(a \mid i) \left\{ (-k+1) r(i, a) \mu^k_o(i)
\]
\[
\quad + \lambda(i) \sum_j \left[ q_{ij}(a) / \lambda(i) + \delta_{ij} \right] (-1)^{k+2} u_j \right\}, \quad i \in S,
\]
\[
\text{iff } (-1)^{k+2} u_i = r_{k+1}(\pi_o) i + \beta_{k+1}(i) \sum_j P(\pi_o)_{ij} (-1)^{k+2} u_j, \quad i \in S;
\]
\[
i.e., v = (-1)^{k+2} u \text{ is a bounded solution of (5) for } n = k + 1. \text{ Therefore } V_{k+1}(\pi_o) = (-1)^{k+2} \mu_{k+1}(\pi_o).
(ii) Because (i) is true for \( n = k + 1 \) and \( \Pi^k = \Pi, (k + 1) \), so
\( V_{k+1}^* = \mu_{k+1}^* \).

(iii) It can be proved similarly as for \( n = 1 \).

This completes the proof.

Theorem 2 gives an equality between the set of \((n)\) moment optimal policies in CTMDP with the sets of optimal policies in DTMDP\((n)\).

From Theorem 2 and Lemma 3, we can immediately obtain the following two theorems.

**Theorem 3.** For \( n < N + 1 \), the following four statements are equivalent.

(i) \( \pi_o \in \Pi^n \);

(ii) \( \pi_o \) is optimal for DTMDP\((k), k \leq n \);

(iii) \( \pi_o \) is optimal for DTMDP\((n)\);

(iv) \( \pi_o(A_n^*(i) \mid i) = 1 \) for all \( i \in S \).

By the above theorem, we reduce the problem of finding \((n)\) moment optimal policies in CTMDP to a problem of finding optimal policies in DTMDP\((n)\). About moment optimal policies, we have the following similar result.

**Theorem 4.** The following four statements are equivalent:

(i) \( \pi_o \in \Pi^n \);

(ii) \( \pi_o \in \Pi^n \) for every \( n \geq 1 \);

(iii) \( \pi_o \) is optimal for DTMDP\((n), n \geq 1 \);

(iv) \( \pi_o(A^*(i) \mid i) = 1 \) for all \( i \in S \), where \( A^*(i) = \bigcap_{n=1}^\infty A_n^*(i) \).

When defining DTMDP\((n)\) by induction, we assumed that if \( N_1 < \infty \) and \( A_{N_1}^* + 1(i) \neq \phi \) for all \( i \in S \), then we stop defining DTMDP\((N_1 + 2)\). In fact, we can prove that \( N = N_1 \).

**Theorem 5.** \( N = N_1 \); i.e., the following two statements hold:

(i) when \( N_1 = \infty \), \( A_n^*(i) \neq \phi \) for all \( n, i \);

(ii) when \( N_1 < \infty \), \( A_n^*(i) \neq \phi \) for every \( i \in S \) and \( n \leq N_1 \), and there exists \( i_o \in S \) such that \( A_{N_1}^* + 1(i) = \phi \).

**Proof.** (i) When \( N_1 = \infty \), there exists \( \pi_o \in \Pi^\infty \). By Theorem 4,

\[ \pi_o(A^*(i) \mid i) = 1 \]

for every \( i \in S \); thus \( A^*(i) \neq \phi \).
(ii) When \( N_1 < \infty \), there exists \( \pi_n \in \Pi_{N_1} \). By Theorem 3,

\[
\pi_n(A_n^{+}(i) \mid i) = 1
\]

for every \( i \in S \); thus \( A_n^{+}(i) \neq \emptyset \) for every \( i \in S \) and \( n \leq N_1 \). Now if \( A_{N_1}^{+}(i) = \emptyset \) for every \( i \in S \) then there exists \( f \in \Pi_{N_1}(N_1 + 1) \) such that \( f \) is optimal for \( \text{DTMDP}(N_1 + 1) \), and so \( f \in \Pi_{N_1+1} \) by Theorem 3. This contradicts the definition of \( N_1 \). This completes the proof.

ACKNOWLEDGMENTS

The author thanks the referee for helpful comments and suggestions. This project was supported by the National Natural Science Foundation of China, Grant 19001025.

REFERENCES

3. Q. Hu, Discrete time Markov decision programming model with discounted moment criterion *J. Systems Engrg.* 5 (1990), 64–74. [in Chinese]