Continuous Time Markov Decision Processes with Expected Discounted Total Rewards*

Qiying Hu[1], Jianyong Liu[2], and Wuyi Yue[3]

[1] College of International Business & Management,
Shanghai University, Shanghai 201800, China.
qyhu@mail.shu.edu.cn
[2] Institute of Applied Mathematics, Academia Sinica, Beijing 100080, China.
[3] Dept. of Information Science and Systems Engineering
Konan University, Kobe 658-8501, JAPAN
yue@konan-u.ac.jp

Abstract. This paper discusses continuous time Markov decision processes with criterion of expected discounted total rewards, where the state space is countable, the reward rate function is extended real-valued and the discount rate is a real number. Under necessary conditions that the model is well defined, the state space is partitioned into three subsets, on which the optimal value function is positive infinity, negative infinity, or finite, respectively. Correspondingly, the model is reduced into three submodels, by generalizing policies and eliminating some worst actions. Then for the submodel with finite optimal value, the validity of the optimality equation is shown and some its properties are obtained.

1 Introduction

Markov decision processes (MDP) have been studied well since its beginning in 1960s. While continuous time MDP (CTMDP) [1], as one of its three basic models, was also studied well though has some delay with respect to the other two basic models, discrete time MDP (DTMDP) [2], and semi-Markov decision processes (SMDP) [3]. A new area is the hybrid system which combines event-driven dynamics and time-driven dynamics, e.g., see [4]. The criterions include discounted criterion, average criterion, expected total rewards and mixed criterion, etc. The standard results in MDP with discounted criterion include the following three aspects. 1) The model is well defined. 2) The optimality equation holds. 3) A stationary policy achieving the supermum of the optimality equation will be optimal. In order to obtain these standard results, some conditions should be required. The general and the most usual method to study a MDP model is first to present a set of conditions for the model, and then, based on the conditions, show the standard results 1), 2) and 3) successively.

There are various conditions presented in literature, especially for DTMDP and SMDP. As for CTMDP with discounted criterion, [5] studied it with unbounded transition rates by using the general method. In [6], the author studied the CTMDP also with unbounded transition rates but by using a transformation method, which can transform the CTMDP into a DTMDP under the discounted criterion. Under this transformation, the corresponding optimality equations and discounted objective functions for the stationary

* This research was supported by the National Natural Science Foundation of China, and by Institute of Applied Mathematics, Academia Sinica and by GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No.13650440), Japan.

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2658, pp. 64–73, 2003.

policies in the CTMDP model and the DTMDP model are equivalent. So the results for CTMDP can be obtained directly from that for DTMDP. In [7], the author studied CTMDP with bounded transition rates also by a transformation, but under which only the discounted objectives for stationary policies in the CTMDP model and the DTMDP model are equivalent. On the other hand, in [8] the author presented a set of conditions for unbounded reward rate. Recently, in [9] the authors discussed a denumerable-state CTMDP with unbounded transition and reward rates. But the method they used is a combination of that of [6] and [8]. In [10], the authors discussed the same model and same conditions but on average criterion.

But there are few studies about expected total rewards criterion. Though the methods presented in [6] and [7] may be used to study it with nonpositive or nonnegative rewards, it is restricted to the stationary policies and can not deal with the general reward rate function or the negative discount rate.

The various conditions presented in literature are only sufficient for MDP. On the contrary, we try to study the necessary conditions, i.e., we want to see what results can be obtained under the condition that the MDP model is well defined. This condition is only the standard result 1), and is obviously the precondition for studying MDP. It is interesting to see if the standard results 2) and 3) can be implied by it. In [11], we studied it for DTMDP with expected discounted total rewards.

This paper is a subsequent one to [11] for CTMDP, where the state space is countable, the reward rate function is extended real-valued and the discount rate may be any real number. The criterion is the discounted expected total rewards with no limits on the discount factor. So it includes the traditional discounted criterion and the expected total rewards criterion. We first generalize the general Markov policies into piecewise semi-Markov policies. Then under the condition that the model is well defined, we show that after eliminating some worst actions, the state space S can be partitioned into three subsets, on which the optimal value function equals $+\infty$, $-\infty$, or is finite, respectively. According to it, the original MDP model can be decomposed into three corresponding sub-models. In the one with finite optimal value, the reward rate function is finite and bounded above at each state, and the validity of the optimality equation is discussed.

The remainder of the paper is organized as follows. Section 2 gives the formulation of the model and presents two conditions, under which Section 3 decomposes the state space and the MDP model. Section 4 discusses some properties of the CTMDP model. In Section 5, the validity of the optimality equation with finite optimal value is shown and several its properties are discussed. While Section 6 is a concluding section.

2 Model and Conditions

The CTMDP model discussed here is

$$\{S, A(i), q_{ij}(a), r(i,a), U_\alpha\} \tag{1}$$

where the state space S and the action set $A(i)$, available at state i, are countable; $\{q_{ij}(a)|i,\ j{\in}S, a{\in}A(i)\}$ is the state transition rate family satisfying $q_{ij}(a) \geq 0$ for $i \neq j$ and $\sum_j q_{ij}(a) = 0$ for $(i,a){\in}\Gamma := \{(i,a)|i{\in}S, a{\in}A(i)\}$, and it is assumed that $\lambda(i) := \sup\{-q_{ii}(a)|a{\in}A(i)\} < \infty$ for $i{\in}S$; the reward rate function $r(i,a)$ is extended real-valued; U_α is the objective function for the criterion of expected discounted total rewards with discount factor $\alpha{\in}(-\infty, +\infty)$, and will be defined below.

We suppose that the measure about the time variable t is the Lebesgue measure.

We define the following policies as in literature, a Markov policy $\pi = (\pi_t, t \geq 0) \in \Pi_m$, a stochastic stationary policy $\pi_0 \in \Pi_s$, a stationary policy $f \in F = \times_i A(i)$. For a policy $\pi = (\pi_t)$ and $s \geq 0$, we define a policy $\pi^s = (\pi_t^*) \in \Pi_m$ by $\pi_t^* = \pi_{s+t}$ for $t \geq 0$. For any policy $\pi = (\pi_t) \in \Pi_m$ and $t \geq 0$, we define a matrix $Q(\pi, t) = (q_{ij}(\pi, t))$ with $q_{ij}(\pi, t) = \sum_{a \in A(i)} q_{ij}(a)\pi_t(a|i)$ and a vector $r(\pi, t) = (r_i(\pi, t))$ with $r_i(\pi, t) = \sum_{a \in A(i)} r(i, a)\pi_t(a|i)$. Thus, $q_{ij}(\pi, t)$ and $r_i(\pi, t)$ are respectively the state transition rate family and the reward rate function under policy π. If $\pi = \pi_0 \in \Pi_s$, then both $Q(\pi_0, t)$ and $r(\pi_0, t)$ are independent of t, and will be denoted respectively by $Q(\pi_0) = (q_{ij}(\pi_0))$ and $r(\pi_0) = (r_i(\pi_0))$.

CONDITION A: For any policy $\pi \in \Pi_m$, the $Q(\pi, t)$-process $\{P(\pi, s, t), 0 \leq s \leq t < \infty\}$ exists uniquely and is the minimal one; moreover, for any $0 \leq s \leq t \leq u < \infty$,

$$\frac{\partial}{\partial t} P(\pi, s, t) = P(\pi, s, t)Q(\pi, t),\ P(\pi, s, u) = P(\pi, s, t)P(\pi, t, u),$$

$$\sum_j P_{ij}(\pi, s, t) = 1,\ P_{ij}(\pi, s, s) = \delta_{ij},\ i, j \in S.$$

One can find the constructing algorithm for the minimal Q-process in [12] (II. 17) for stationary case and in [1] for nonstationary case. Condition A is true when $q_{ij}(a)$ is bounded, or under the assumptions presented in [5] when $q_{ij}(a)$ is unbounded.

Now, we generalize the concept of policies. Let $Y(t)$ be the state of the process at time t. Given any integer N, real numbers $\{t_i, i = 1, 2, \ldots, N\}$ with $0 = t_0 < t_1 < \ldots < t_N < t_{N+1} = \infty$, and Markov policies $\{\pi^{n,i}, n = 0, 1, 2, \ldots, N, i \in S\} \subset \Pi_m$, we define a policy $\pi = (\pi^{n,i}; n = 0, 1, 2, \ldots, N, i \in S)$ as follows: for $n = 0, 1, 2, \ldots, N$, if $Y(t_n) = i$, then $\pi^{n,i}$ is used in time interval $[t_n, t_{n+1})$, i.e., the action is chosen according to $\pi^{n,i}_{t-t_n}(\cdot|j)$ at time $t \in [t_n, t_{n+1}]$ if $Y(t) = j \in S$. Such a policy, denoted by $\pi = (\pi^{n,i})$ for short, is called a (finite) *piecewise semi-Markov policy*, the set of which is denoted by $\Pi_m(s)$. If all $\pi^{n,i} = f^{n,i} \in F$, then $\pi = (f^{n,i})$ is called a piecewise semi-stationary policy, the set of which is denoted by $\Pi_s^d(s)$.

For such a policy π, if $Y(t_n) = i$, then the system in $[t_n, t_{n+1}]$ is a Markov process with transition probability matrix $P(\pi^{n,i}, s, t)$. So, the system under it is a special case of piecewise Markov process (see [13]). In details, for each s and t with $0 \leq s \leq t$ and $i, j \in S$, suppose that $s \in [t_m, t_{m+1}]$ and $t \in [t_n, t_{n+1}]$ for some $m \leq n$, then the state transition probability that the system will be in state j at time t provided that the system is in state i at time s and in state k at time t_m is

$$\begin{aligned} P_{ij}^k(\pi, s, t) &:= P_\pi\{Y(t) = j | Y(s) = i, Y(t_m) = k\} \\ &= \sum_{j_1} P_{ij_1}(\pi^{m,k}, s - t_m, t_{m+1} - t_m) \\ &\quad \cdot \sum_{j_{n-m}} P_{j_{n-m-1}j_{n-m}}(\pi^{n-1,j_{n-m-1}}, 0, t_n - t_{n-1}) \\ &\quad \cdot P_{j_{n-m}j}(\pi^{n,j_{n-m}}, t_n, t). \end{aligned} \tag{2}$$

For $i, j \in S$, let $P_{ij}(\pi, t) = P_{ij}^i(\pi, 0, t)$.

Now, we define the objective function, for a Markov policy $\pi \in \Pi_m$, by

$$U_\alpha(\pi) = \int_0^\infty \exp(-\alpha t) P(\pi, t) r(\pi, t) dt \tag{3}$$

where the integral is the Lebesgue integral. It is the expected discounted total rewards on the whole time axis under π. Let $U_\alpha(\pi,t) := U_\alpha(\pi^t)$, for $t \geq 0$. Obviously,

$$U_\alpha(\pi,t) = \int_t^\infty \exp(-\alpha(s-t))P(\pi,t,s)r(\pi,s)ds \tag{4}$$

is the expected discounted, to time t, total rewards on the time axis $[t,\infty)$ under π. And for $\pi = (\pi^{n,i}) \in \Pi_m(s)$ with $\{t_n, n = 1,2,\cdots,N\}$ and $t \geq 0$, we define inductively

$$\begin{aligned} U_\alpha^{n,k}(\pi,t,i) = &\int_t^{t_{n+1}} \exp(-\alpha(s-t)) \sum_j P_{ij}(\pi^{n,k},t,s)r_j(\pi^{n,k},s)ds \\ &+ \exp(-\alpha(t_{n+1}-t)) \sum_j P_{ij}(\pi^{n,k},t,t_{n+1})U_\alpha^{n+1,j}(\pi,t_{n+1},j), \\ &\qquad t \in [t_n,t_{n+1}),\ n = 0,1,\ldots,N-1,\ k,i \in S, \\ U_\alpha^{N,k}(\pi,t,i) = &\ U_\alpha(\pi^{N,k},t-t_N,i),\ t \geq t_N,\ k,i \in S. \end{aligned} \tag{5}$$

Let $U_\alpha^{n,k}(\pi,t_n,i) = 0$ for $t = t_n$ and $k \neq i$, and $U_\alpha(\pi,i) = U_\alpha^{0,i}(\pi,0,i)$. Let $U_\alpha(\pi)$ be the vector with its i-th component $U_\alpha(\pi,i)$.

CONDITION B: $U_\alpha(\pi)$ is well defined (may be infinity) for each $\pi \in \Pi_m(s)$.

This condition means that :1) $\sum_j P_{ij}(\pi,t)r_j(\pi,t)$, and furthermore, the integral in Eq. (3), are well-defined for each $\pi \in \Pi_m$; 2) $\sum_j P_{ij}(\pi,t,s)U_\alpha(\pi',s,j)$ is well-defined for every policies $\pi \in \Pi_m$ and $\pi' \in \Pi_m(s)$; 3) the sum in Eq. (5) is well-defined. The condition is necessary to discuss CTMDP. It is well known that it is true whenever $\alpha > 0$ and $r(i,a)$ is bounded above or below; or $\alpha \geq 0$ and $r(i,a)$ is nonnegative or nonpositive. Condition B is assumed to be true throughout the paper.

Conditions A and B hold mean that the CTMDP model (Eq. (1)) is well defined.

Because a policy $\pi \in \Pi_m$ is also a piecewise semi-Markov policy with arbitrary N and $t_1,t_2,\ldots,t_N$, it follows from Eq. (5) that for $\pi \in \Pi_m, t \geq 0$,

$$U_\alpha(\pi) = \int_0^t \exp(-\alpha s)P(\pi,s)r(\pi,s)ds + \exp(-\alpha t)P(\pi,t)U_\alpha(\pi,t), \tag{6}$$

which means that $P(\pi,t)$ can be put out of the integral $\int_t^\infty$, that is,

$$\begin{aligned} &\int_t^\infty \exp(-\alpha(s-t))P(\pi,s)r(\pi,s)ds \\ &= P(\pi,t)\int_t^\infty \exp(-\alpha(s-t))P(\pi,t,s)r(\pi,s)ds \\ &= P(\pi,t)U_\alpha(\pi,t). \end{aligned}$$

Eq. (6) is still true for policies $\pi \in \Pi_m(s)$ by defining $r(\pi,s)$ adequately.

Let the optimal value function be $U_\alpha^*(i) = \sup\{U_\alpha(\pi,i) | \pi \in \Pi_m(s)\}$ for $i \in S$. For $\varepsilon \geq 0, \pi^* \in \Pi_m(s)$, if $U_\alpha(\pi^*,i) \geq U_\alpha^*(i) - \varepsilon$ (if $U_\alpha^*(i) < +\infty$) or $\geq 1/\varepsilon$ (if $U_\alpha^*(i) = +\infty$), then π^* is called ε-optimal. Here, $1/0 = +\infty$ is assumed. 0-optimal is simply called optimal.

3 Eliminating the Worst Actions

First, we introduce some concepts. State j can be reached from state i (and write $i \to j$) if there are a policy $\pi \in \Pi_m(s)$ and $t \geq 0$ such that $P_{ij}(\pi,t) > 0$. It is easy to see that

$i \to j$ iff there are $\pi \in \Pi_m$ and $t \geq 0$ such that $P_{ij}(\pi, t) > 0$, or equivalently there are $n \geq 0$, states $j_1, j_2, \ldots, j_n \in S$ and $f \in F$ such that $q_{ij_1}(f) q_{j_1 j_2}(f) \ldots q_{j_n j}(f) > 0$. It is apparent that if $i \to j$ and $j \to k$, then $i \to k$. For a subset $S_0 \subset S$ and a state i, if there is a state $j \in S_0$ such that $i \to j$, then we say that S_0 can be reached from state i, which is denoted by $i \to S_0$. Let $S_0^* = \{i | i \to S_0\}$ be a set of states that can reach S_0. Because $i \to i$, so $S_0 \subset S_0^*$. A subset S_0 of S is called a closed (state) set if $q_{ij}(a) = 0$ for all $i \in S_0, a \in A(i)$ and $j \bar{\in} S_0$, or equivalently, $(S - S_0)^* = S - S_0$. Similarly as above, S_0 is closed iff $P_{ij}(\pi, t) = 0$ for all $i \in S_0, \pi \in \Pi_m(s), j \bar{\in} S_0$ and $t \geq 0$.

For any closed subset S_0, if the system's initial state $i \in S_0$, then the system will remain in S_0 irrespective of the policies used. Thus, the restriction of CTMDP to S_0,

$$S_0\text{-CTMDP} := \{S_0, (A(i), i \in S_0), p_{ij}(a), r(i, a), U_\alpha\}$$

is also a CTMDP, which is called an induced sub-CTMDP by S_0. Its policies are restriction of the original policies to S_0. It is clear that Condition A and B are also true for S_0-CTMDP. Let its objective function be $U_\alpha^{S_0}(\pi)$.

THEOREM 1: For any closed subset $S_0 \subset S, \pi \in \Pi_m(s)$ and $i \in S_0$, $U_\alpha(\pi, i) = U_\alpha^{S_0}(\pi, i)$.

The theorem says that the induced sub-CTMDP by a closed set S_0 is equivalent to the original CTMDP in subset S_0. So, if both S_0 and $S - S_0$ are closed, then CTMDP can be partitioned into two smaller parts:S_0-CTMDP and $(S - S_0)$-CTMDP. On the other hand, if S_0 is closed while $U_\alpha^*(i)$ for $i \in S - S_0$ is known, or a (ε-)optimal policy can be obtained in $S - S_0$, then one need to discuss only S_0-CTMDP. Thus the state space is partitioned and reduced.

On the other hand, some actions may be eliminated with no influence on the essential of the model.

DEFINITION 1: Suppose that $A_1(i) \subset A(i)$ for $i \in S$. We denote by CTMDP$'$ the CTMDP with $A(i)$ being replaced by $A_1(i)$ (a symbol " $'$ " is added). If for any policy π of the (original) CTMDP there is a policy π' of the CTMDP$'$ such that $U_\alpha(\pi, i) \leq U'_\alpha(\pi', i)$ for all i, then the CTMDP is equivalent to the CTMDP$'$, and we say that $A(i)$ can be reduced as $A_1(i)$ for $i \in S$, or $a \in A(i) - A_1(i)$ can be eliminated for $i \in S$.

Now, we denote by $U(i) = \sup\{r(i, a) | a \in A(i)\}$ and $L(i) = \inf\{r(i, a) | a \in A(i)\}$ respectively the supremum and infimum of the reward rate function $r(i, a)$ over the action set $A(i)$ for $i \in S$. Let $S_U = \{i | U(i) = +\infty\}, S_{=\infty} = \{i |$there is $\pi \in \Pi_m(s)$ such that $U_\alpha(\pi, i) = +\infty\}, S_\infty = \{i | U_\alpha^*(i) = +\infty\} - S_{=\infty}, S_{-\infty} = \{i | U_\alpha^*(i) = -\infty\}, S_0 = S - S_{=\infty} - S_\infty - S_{-\infty} = \{i | -\infty < U_\alpha^*(i) < \infty\}$. These state subsets have obvious meanings.

LEMMA 1:

1) For $i \in S_U$, there is a policy $\pi_0 \in \Pi_s$ such that $r_i(\pi_0) = +\infty$. So $U_\alpha(\pi_0, i) = +\infty$ and $S_U \subset S_{=\infty}$.

2) For $i \in S$ with $L(i) = -\infty$, there is a policy $\pi_0 \in \Pi_s$ such that $r_i(\pi_0) = -\infty$ and then $U_\alpha(\pi_0, i) = -\infty$.

3) For $i \in S, L(i) = -\infty$ and $U(i) = +\infty$ can not be true simultaneously.

THEOREM 2:

1) $S_{=\infty}^* = S_{=\infty}$ and so $S' := S - S_{=\infty}$ is closed.

2) For $i \in S' - S_{-\infty}$, $A(i)$ can be reduced as

$$A_1(i) = \{a \in A(i) | r(i,a) > -\infty \quad \text{and} \sum_{j \in S_{-\infty}} q_{ij}(a) = 0\}. \tag{7}$$

After the reduction, $S^*_{-\infty} = S_{-\infty}$ and so $S'' := S' - S_{-\infty}$ becomes closed.

3) For $i \in S''$, $A_1(i)$ can further be reduced as

$$A_2(i) = \{a \in A_1(i) | \exists \pi \in \Pi_m \text{ with } U_\alpha(\pi, i) > -\infty \text{ and the Lebesgue measure of } \{s \in [0,t] | \pi_s(a|i) > 0\} \text{ is positive for each } t > 0\}. \tag{8}$$

After this reduction, $S^*_\infty = S_\infty$, and so $S_0 := S'' - S_\infty$ is closed.

By Theorem 1 and 2, S can be partitioned into four subsets: $S_{-\infty}, S_{=\infty}, S_\infty, S_0$. In $S_{-\infty}$, each policy is optimal; in $S_{=\infty}$, there is an optimal policy (in fact, there is a stochastic stationary optimal policy in S_U); in $S_\infty, U_\alpha(\pi, i) < \infty$ for each π, while $U^*_\alpha(i) = \infty$, and thus there is no optimal policy, and in $S_0, U^*_\alpha(i)$ is finite, and S_0 is closed after eliminating some worst actions. So, one can consider only the following CTMDP:

$$S_0\text{-CTMDP} = \{S_0, A_2(i), q_{ij}(a), r(i,a), U_\alpha\}. \tag{9}$$

Because $i \in S_{-\infty}$ when $A_2(i) = \emptyset$, Eq. (9) is a CTMDP; furthermore, we have

$$-\infty < U^*_\alpha(i) < +\infty, -\infty < r(i,a) \leq U(i) < +\infty, \qquad \forall i, a. \tag{10}$$

It is easy to see that all the above results restricted to $\Pi_s(s)$ are also true.

In the remaining of this paper, we discuss mainly the S_0-CTMDP, and so will write S_0 and $A_2(i)$ by S and $A(i)$ respectively for convenience.

4 Some Properties

This section discusses some properties of S_0-CTMDP (Eq. (9)) and simplifies the expression of $A_2(i)$. First, the following lemma is from [12] (II. 15-17).

LEMMA 2: Suppose that $P(t) = (p_{ij}(t))$ is a homogeneous state transition probability matrix on a countable state space S with a finite transition rate family $Q = (q_{ij})$. Let $q_i = -q_{ii}$. Then there are nonnegative continuous functions $g_{ij}(t)$ for $i, j \in S$, on $[0, \infty)$, such that

$$p_{ij}(t) = \exp(-q_i t) \int_0^t \exp(q_i s) q_i g_{ij}(s) ds + \exp(-q_i t) \delta_{ij}, \qquad i, j \in S, t \geq 0$$

where δ_{ij} denotes the Kronecker delta function, and for $s > 0, t \geq 0$,

$$\lim_{s \to 0^+} g_{ij}(s) = (1 - \delta_{ij}) q_{ij}/q_i, \ \sum_j g_{ij}(s) = 1, \ g_{ij}(s+t) = \sum_k g_{ik}(s) p_{kj}(t).$$

Based on the above lemma, one can proved the following two lemmas.

LEMMA 3: Suppose that $P(t) = (p_{ij}(t)), Q$ and $g_{ij}(t)$ are as in Lemma 2, $\sup_i q_i < \infty, u$ is a finite nonnegative function in $S, Z \subset S, i \in S$. If $\sum_{j\in Z} p_{ij}(t^*)u_j$ is finite for some $t^* > 0$, then $h_i(t) := q_i \exp(q_i t) \sum_{j\in Z} g_{ij}(t)u_j$ is finite and continuous in $[0, t^*)$, and $\sum_{j\in Z} q_{ij}u_j < \infty$; otherwise, $h_i(t) = +\infty$ for all $t > 0$.

LEMMA 4: Using the symbols in Lemma 2, Suppose that $\sup_i q_i < \infty, u$ is a finite function in $S, t^* > 0$ and $i \in S$. If $\sum_j p_{ij}(t)u_j$ is finite in $[0, t^*]$, then its derivative is well-defined and continuous in $[0, t^*)$, and

$$\frac{d}{dt}\Big\{\sum_j p_{ij}(t)u_j\Big\} = \sum_j \frac{d}{dt}p_{ij}(t)u_j = \sum_j \{-q_i p_{ij}(t) + q_i g_{ij}(t)\}u_j.$$

Having the above several lemmas for pre preparation, now we can prove the following theorem, where $S^+ := \{i \in S_0 | U^*_\alpha(i) \geq 0\}$ and $S^- := \{i \in S_0 | U^*_\alpha(i) < 0\}$.

THEOREM 3: 1) $P(\pi, t)U^*_\alpha < \infty$ is well defined for each $\pi \in \Pi_m(s)$ and $t > 0$.

2) For $\pi \in \Pi_m(s), t > 0$ and $i \in S$, if $\sum_j P_{ij}(\pi, t)U^*_\alpha(j) = -\infty$, then $U_\alpha(\pi^*, i) = -\infty$ for any piecewise semi-Markov policy $\pi^* = (\pi^{n,j}) \in \Pi_m(s)$ with $\pi^{0,i} = \pi$ and $t_1 = t$, especially, $U_\alpha(\pi, i) = -\infty$.

COROLLARY 1: Suppose that there is $f^* \in F$ such that $Q(f^*)$ is bounded, then $\sum_j q_{ij}(a)U^*_\alpha(j) < \infty$ is well defined for any $i \in S$ and $a \in A(i)$.

COROLLARY 2: Suppose that $f \in F$ with bounded $Q(f), i \in S, t^* > 0, [P(f, t)U^*_\alpha]_i$ is finite in $[0, t^*]$, then $[P(f, t)U^*_\alpha]_i$ is differentiable in $[0, t^*)$, its derivative is continuous and

$$\frac{d}{dt}\Big\{\sum_j P_{ij}(f, t)U^*_\alpha(j)\Big\} = \sum_j \frac{d}{dt}P_{ij}(f, t)U^*_\alpha(j),$$

$$\sum_j [P(f, t)Q(f)]_{ij}U^*_\alpha(j) = \sum_j P_{ij}(f, t)[Q(f)U^*_\alpha(j)]_j, \qquad t \in [0, t^*). \tag{11}$$

We conjecture that the result in Corollary 2 is also true for $\pi \in \Pi_m$, but it needs that Lemma 2 holds for a nonhomogeneous Markov process, which is not known to us.

To conclude equations in this section, we give the following theorem on a simplified expression for $A_2(i)$.

THEOREM 4: If $q_{ij}(a)$ is uniformly bounded, then

$$A_2(i) \subset \{a \in A_1(i) | \sum_j q_{ij}(a)U^*_\alpha(j) > -\infty\}, \qquad i \in S; \tag{12}$$

moreover, if $h_i(t)$ is finite and continuous whenever $\sum_{j\in Z} q_{ij}u_j$ is finite in Lemma 3, then

$$A_2(i) = \{a \in A_1(i) | \sum_j q_{ij}(a)U^*_\alpha(j) > -\infty\}, \qquad i \in S. \tag{13}$$

Remark 1: 1) By the above theorem, if S^- is finite, or $U^*_\alpha(i)$ is bounded below, then Eq. (13) is true when $q_{ij}(a)$ is uniformly bounded; 2) S^- is empty if U^*_α is nonnegative, especially, if the reward function is nonnegative; 3) $U^*_\alpha(i)$ is bounded below if $\alpha > 0$ and the reward function is bounded below.

5 Optimality Equation

This section shall deal with the standard results 2) and 3) (see Section 1), that is, we shall show the optimality equation and the optimality of policies achieving the optimality equation for S_0-CTMDP, under the assumption that $\{q_{ij}(a)\}$ is uniformly bounded, i.e., $\lambda = \sup\{-q_{ii}(a)|i \in S, a \in A(i)\} < \infty$. For $\pi \in \Pi_m(s), t \geq 0$ and a finite function $u = (u(i))$ on S, we define

$$U_\alpha(\pi, t, u) = \int_0^t \exp(-\alpha s) P(\pi, s) r(\pi, s) ds + \exp(-\alpha t) P(\pi, t) u$$

whenever the right hand side is well-defined. Denote $U_\alpha^*(\pi, t) = U_\alpha(\pi, t, U_\alpha^*)$ for short, which is well-defined by Theorem 3. Certainly, $U_\alpha^*(\pi, t)$ is the expected discounted total rewards if π is used in $[0, t]$ and then an optimal policy is used from t.

LEMMA 5: $U_\alpha^* = \sup\{U_\alpha^*(\pi, t)|\pi \in \Pi_m(s)\}$ for $t \geq 0$, and $U_\alpha^*(\pi, t)$ is nonincreasing in t for any $\pi \in \Pi_m(s)$.

Now, we introduce our third condition.

CONDITION C: For each $i \in S$ and $a \in A(i)$, there is f and $t > 0$ such that $f(i) = a$ and $U_\alpha^*(f, t, i) > -\infty$.

Remark 2: Two sufficient conditions for Condition C are as follows: 1) the conditions given in Theorem 4, especially, when S^- is finite or U_α^* is bounded below (see Remark 1); 2) for each $i \in S$, $A(i)$ can be reduced as

$$A'(i) = \{a \in A(i)| \sup_{f \in F: f(i) = a} U_\alpha(f, i) > -\infty\},$$

which means that any action $a \in A(i)$ should be eliminated if any stationary policy f using it will have negative infinite objective value. In fact, if $A(i)$ can be reduced as $A'(i)$, then it follows Lemma 5 that $U_\alpha^* \geq U_\alpha^*(f, t) \geq U_\alpha(f) > -\infty$ for each $f \in F$ and $t > 0$.

THEOREM 5: Under Condition C, U_α^* satisfies the following optimality equation:

$$\alpha U_\alpha^*(i) = \sup_{a \in A(i)} \{r(i, a) + \sum_j q_{ij}(a) U_\alpha^*(j)\}, \qquad i \in S. \tag{14}$$

The policy set is generalized here, but it is often our pleasure to restrict an ($\varepsilon \geq 0$) optimal policy to a smaller and simpler policy set. To do this, our first result is the following theorem, which says that the optimality can be restricted to Π_m, the set of Markov policies, iff the optimal value function restricted to Π_m also satisfies the optimality Eq. (14). Let $U_\alpha^m = \sup\{U_\alpha(\pi)|\pi \in \Pi_m\}$. We affirm that U_α^m is finite. In fact, if $U_\alpha^m(i_0) = -\infty$ for some $i_0 \in S$, then it is easy to see from Eq. (5) that $U_\alpha(\pi, i_0) = -\infty$ for each $\pi \in \Pi_m(s)$. Thus $U_\alpha^*(i_0) = -\infty$, which is a contradiction. But $U_\alpha^m \leq U_\alpha^* < \infty$, so, U_α^m is finite.

THEOREM 6: $U_\alpha^* = U_\alpha^m$ iff U_α^m is a solution of the optimality Eq. (14).

In order to obtain some properties for the optimality Eq. (14), we define a set, denoted by W, of finite functions $u = (u(i))$ on S satisfying the following conditions: for each $\pi \in \Pi_m(s)$ and $i \in S$, $\sum_j P_{ij}(\pi, t) u(j) < \infty$ is well-defined for all $t \geq 0$. Moreover, $\sum_j P_{ij}(\pi, t) u(j) > -\infty$ whenever $\sum_j P_{ij}(\pi, t) U_\alpha^*(j) > -\infty$ for each $t \geq 0$ and $i \in S$. W is nonempty for $U_\alpha^* \in W$. It is clear that $U_\alpha(\pi, t, u) < \infty$ is well-defined for each $u \in W$.

LEMMA 6: Suppose that $\varepsilon \geq 0, \beta + \alpha \geq 0, u \in W$, $\pi \in \Pi_m$ and $i \in S$. If π and u satisfy the following two conditions, then $u(i) \leq U_\alpha(\pi, i) + (\beta + \alpha)^{-1}\varepsilon$.

$$\alpha u \leq r(\pi, t) + Q(\pi, t)u + \exp(-\beta t)\varepsilon e, \text{ a.e. } t \geq 0, \tag{15}$$

$$\liminf_{t\to\infty} \exp(-\alpha t) \sum_j P_{ij}(\pi, t)u(j) \leq 0. \tag{16}$$

THEOREM 7: Suppose that $u \in W$ is a solution of the optimality Eq. (14) and $i \in S$.

1) if for some $\beta > -\alpha$ and each $\varepsilon > 0$, there is a policy $\pi \in \Pi_m(s)$ with $U_\alpha(\pi, i) > -\infty$ satisfying Eq. (15) and Eq. (16), then $u(i) \leq U^*_\alpha(i)$;
2) if u satisfies the following (23) for each $\pi \in \Pi_m(s)$ with $U_\alpha(\pi, i) > -\infty$, then $u(i) \geq U^*_\alpha(i)$,

$$\limsup_{t\to\infty} \exp(-\alpha t) \sum_j P_{ij}(\pi, t)u(j) \geq 0. \tag{17}$$

It is clear that there is often a policy $\pi = (f_t) \in \Pi^d_m$ satisfying Eq. (15), but it may be not true that $U_\alpha(\pi, i) > -\infty$. On the other hand, U^*_α often satisfies Eq. (17) for $\pi \in \Pi_m(s)$ with $U_\alpha(\pi, i) > -\infty$. In fact, by Eq. (6) we know that if $U_\alpha(\pi, i) > -\infty$, then $\sum_j P_{ij}(\pi, t)U^*_\alpha(j)$ is also finite for each $t \geq 0$, and

$$\limsup_{t\to\infty} \exp(-\alpha t) \sum_j P_{ij}(\pi, t)U^*_\alpha(j) \geq \limsup_{t\to\infty} \exp(-\alpha t) \sum_j P_{ij}(\pi, t)U_\alpha(\pi, t, j) = 0.$$

The following corollary can be proved easily by Theorem 7 and Lemma 6.

COROLLARY 3: Provided that Eq. (14) holds,

1) for any given $f \in F$, if f attains supremum of Eq. (14), f and U^*_α satisfy Eq. (16) and $U_\alpha(f) > -\infty$, then f is optimal;
2) for some $\pi^* \in \Pi_m(s)$, if $U_\alpha(\pi^*)$ is a solution Eq. (14), then π^* is optimal;
3) if for any $\varepsilon > 0$, there is a Markov policy $\pi \in \Pi^d_m$ with $U_\alpha(\pi) > -\infty, \pi$ and U^*_α satisfy Eq. (15) and Eq. (16) for each $i \in S$, then $U^*_\alpha = \sup\{U_\alpha(\pi) | \pi \in \Pi^d_m\}$;
4) if $\alpha > 0, \varepsilon \geq 0, f \in F$ attains the ε-supremum of Eq. (14), f and U^*_α satisfy Eq. (16), $U_\alpha(f) > -\infty$, then f is $\alpha^{-1}\varepsilon$-optimal; moreover, if such f exists for each $\varepsilon > 0$, then $U^*_\alpha = \sup\{U_\alpha(f) | f \in F\}$;
5) if $U^*_\alpha \leq 0$, then U^*_α is the largest solution of Eq. (14) in W satisfying conditions given in 1) of Theorem 7;
6) U^*_α is the smallest solution of Eq. (14) in W satisfying Eq. (17) for $\pi \in \Pi_m(s)$ and $i \in S$ with $U_\alpha(\pi, i) > -\infty$.

COROLLARY 4: For $f \in F$ and $i \in S$ with $U_\alpha(f, i) > -\infty, \sum_j q_{ij}(f)U_\alpha(f, j)$ is finite and

$$\alpha U_\alpha(f, i) = r(i, f) + \sum_j q_{ij}(f)U_\alpha(f, j). \tag{18}$$

To conclude equations in this section, we discuss the CTMDP model (see Eq. (1)) restricted to $\Pi^d_s(s)$, the set of piecewise semi-stationary policies. In this case, Theorem 2 is still true except that " $\leq U(i)$" should be deleted in Eq. (10) and $A_2(i)$, defined by Eq. (8), should be redefined by

$$A_2(i) = \{a \in A_1(i) | \text{there is } f \in F \text{ such that } f(i) = a \text{ and } U_\alpha(f, i) > -\infty\}.$$

Thus, Condition C is trivial. By noting that Corollary 2 and 4 also hold for f. The following theorem can be proved similarly as Theorem 5 and 6.

THEOREM 8: Restricted to $\Pi_s^d(s)$, $U_\alpha^{*d} := \sup\{U_\alpha(\pi)|\pi \in \Pi_s^d(s)\}$ satisfies Eq. (14), moreover, $U_\alpha^s := \sup\{U_\alpha(f)|f \in F\}$ satisfies Eq. (14) iff $U_\alpha^{*d} = U_\alpha^s$.

6 Conclusions

This paper discussed CTMDP with expected discounted total rewards under the necessary conditions that the model is well defined. We partitioned the state space into three subsets, on which the optimal value is negative infinity, positive infinity and finite respectively. Thus the discussion on the CTMDP could be restricted in the sub-state space with finite optimal value (we call it a sub-CTMDP). In fact, the reward rate function of this sub-CTMDP is finite and is bounded above in the action. Finally, we showed, for this sub-CTMDP, its optimality equation and the optimality of policies achieving the optimality equation.

Further research may include if we can deal with the state partition and action elimination directly on the optimality equation such that the optimality equation can be obtained whenever it is well defined. Also, Condition C may be proved.

References

1. Kakumanu, P.V.: Continuous Time Markov Decision Models with Applications to Optimization Problems. Technical Report **63**, Dept. of Oper. Res., Cornell Univ. (1969)
2. Lewis, M.E. and Puterman, M.L.: A Probabilistic Analysis of Bias Optimality in Unichain Markov Decision Processes. IEEE Trans. on Autom. Contr. **46** (2001) 96–100
3. Lippman, S.A.: On Dynamic Programming with Unbounded Rewards, Mgt. Sci. **21** (1975) 1225–1233
4. Cassandras, C.G., Pepyne, D.L. and Wardi, Y.: Optimal control of A Class of Hybrid Systems. IEEE Trans. on AC **46** (2001) 398–415
5. Song, J.: Continuous Time Markov Decision Processes with Nonuniformly Bounded Transition Rate Family, Scientia Sinica Series A, **11** (1988) 1281–1290
6. Hu, Q.: CTMDP and Its Relationship with DTMDP. Chinese Sci. Bull. **35** (1990) 710–714
7. Serfozo, R.F.: An Equivalence Between Continuous and Discrete Time Markov Decision Processes, J. Oper. Res. **27** (1979) 60–70
8. Hou, B.: Continuous-time Markov Decision Processes Programming with Polynomical Reward, Thesis, Institute of Appl. Math. Academic Sinica, Bejing (1986).
9. Guo, X.P. and Zhu, W.P.: Denumerable-state Continuous-time Markov Decision Processes with Unbounded Transition and Reward Rates under the Discounted Criterion. J. Appl. Prob. **39** (2002) 233–250
10. Guo, X.P. and Zhu, W.P.: Denumerable-state Continuous-time Markov Decision Processes with Unbounded Cost and Transition Rates under Average Criterion. ANZIAM J. **43** (2002) 541–557
11. Hu, Q. and Xu, C.: The Finiteness of the Reward Function and the Optimal Value Function in Markov Decision Processes. J. Math. Methods in Ope. Res. **49** (1999) 255–266
12. Chung, K.L.: Markov Chains with Stationary Transition Probabilities. Springer-Verlag (1960)
13. Kuczura, A.: Piecewise Markov Processes. SIAM J. Appl. Math. **24** (1973) 169–181