NECESSARY CONDITIONS FOR CONTINUOUS
TIME MARKOV DECISION PROCESSES WITH
EXPECTED DISCOUNTED TOTAL REWARDS

Qiying Hu\textsuperscript{1}, Jianyong Liu\textsuperscript{2}, Wuyi Yue\textsuperscript{3} \textsuperscript{6}

\textsuperscript{1}College of International Business and Management
Shanghai University
Shanghai 201800, P.R. CHINA
e-mail: qyhu@mail.shu.edu.cn

\textsuperscript{2}Institute of Applied Mathematics
Academia Sinica
Beijing 100080, P.R. CHINA

\textsuperscript{3}Department of Information Science and Systems Engineering
Faculty of Science and Engineering
Konan University
8-9-1 Okamoto, Higashinada-ku, Kobe 658-8501, JAPAN
e-mail: yue@konan-u.ac.jp

Abstract: This paper discusses a set of necessary conditions for 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 any 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.

AMS Subject Classification: 26A33
Key Words: continuous time Markov decision processes, expected total rewards, model decomposition, necessary conditions

\textsuperscript{6}Correspondence author
1. Introduction

Markov decision processes (MDP) have been studied well since its beginning in 1960s. It has three basic models: discrete time MDP (DTMDP) [1], continuous time MDP (CTMDP) [2] and semi-Markov decision processes (SMDP) [3]. Based on them, several generalized models are presented, such as partially observable MDP, adaptive MDP, multi-objective MDP, etc. While the criteria include expected discounted total rewards, average rewards and mixed criterion, etc. One can see a survey paper [4]. A new area is the hybrid system, which combines event-driven dynamics and time-driven dynamics, e.g., see [5].

The standard results in MDP include the following three aspects: 1) The model is well defined, i.e., the process under consideration is well defined and will happen only a finite number of state transitions in every finite time with probability one, moreover, the objective function is well defined, and often is finite. 2) The optimal value function satisfies the optimality equation (then we say that the optimality equation holds). 3) A stationary policy achieving the ($\varepsilon$-)supremum of the optimality equation will be ($\varepsilon'$-)optimal, where $\varepsilon'$ is a function of $\varepsilon$ and tends to zero when $\varepsilon$ tends to zero. 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, present a set of conditions for the model, and then, based on the conditions, show the standard results 1), 2) and 3) successively.

For the expected discounted total rewards, two simpler cases are the discount factor being in $(0, 1)$ with uniform bounded rewards and the discount factor being nonnegative with nonnegative or nonpositive rewards, but they are too strong. In fact, various conditions are presented in MDP literature to suit the various practical problems, especially for discrete time MDP (DTMDP) and semi-Markov decision processes (SMDP). For example, in [3], the author presented a set of conditions on the unbounded reward functions for SMDP, while [6] and [7] presented that for DTMDP, all of them for the discounted criterion and shown the standard results 1), 2) and 3) in a row, based on the conditions presented.

Continuous time MDP (CTMDP) has been also studied well for discounted criterion with bounded rewards. In [8], the author studied it with unbounded transition rates by using the general method. In [2], the author studied the CTMDP with discounted criterion also with unbounded transition rates but by using a transformation method, which can transform the CTMDP with discounted criterion into a DTMDP with discounted criterion. Under this transformation, the corresponding optimality equations and discounted objective
functions for the stationary 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 [9], the author studied CTMDP with bounded transition rates also by a transformation, but under which only the discounted objective functions for the stationary policies in the CTMDP model and the DTMDP model are equivalent. On the other hand, in [10] the author presented a set of conditions for discounted criterion CTMDP with unbounded reward rate. Her conditions are, in fact, the generalization of that of Lippman’s. Recently, in [11] the authors discussed a denumerable-state CTMDP with the discounted criterion by presenting a set of conditions for the unbounded transition and reward rates, which is weaker than that in literature illustrated by examples. But, the method they used is a combination of that of [2] and [10]. In [12], the authors discussed the same model and same conditions but on average criterion, and the standard results are obtained.

There are few studies about expected total rewards criterion. Though the methods presented in [2] and [9] 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 nonstationary CTMDP models discussed in literature, e.g., in [13], are often corresponding to the stationary one with discounted criterion and bounded rewards.

Certainly, the various conditions presented in literature, e.g., [3]-[7], are only the sufficient conditions 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 [14], we studied it for DTMDP with expected discounted total rewards. By eliminating some worst actions, we show the validity of the optimality equation if its right hand side is well defined, which is ensured in a subset of the state space with finite optimal value.

This paper is a subsequent one to [14] 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, which excludes the confusion of $\infty - \infty$ in the optimality equation. Moreover, each sub-model can be discussed separately. 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. Moreover, some properties of the optimality equation are obtained. The negative discount rate is allowed in this paper, because the method used is also suitable for it. This corresponds to the case with the discount factor being larger than one in discrete time Markov decision processes discussed, e.g., in [15].

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 model of continuous time Markov decision processes discussed in this paper is

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

where the state space $S$ and the action set $A(i)$, available at state $i$, are countable; $\{q_{ij}(a) \mid i, j \in S, a \in A(i)\}$ is the state transition rate family satisfying $q_{ij}(a) \geq 0$ for $i \neq j$ and $\sum_j q_{ij}(a) = 0$ for $(i, a) \in \Gamma := \{(i, a) \mid i \in S, a \in A(i)\}$, and it is assumed that for each $i \in S$, $\lambda(i) := \sup\{-q_{ii}(a) \mid a \in A(i)\} < \infty$; the reward rate function $r(i, a)$ is extended real-valued; $U_\alpha$ is the objective function for the criterion of expected discounted total rewards with discount factor $\alpha \in (-\infty, +\infty)$, and will be defined below. In general, $\alpha \in [0, \infty)$ is required in literature, but is not necessary in this paper.

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

A Markov policy $\pi = (\pi_t, t \geq 0) \in \Pi_m$ means that if the system is in a state $i$ at time $t \geq 0$, then the action chosen is according to a probability distribution $\pi_t(\cdot \mid i)$ in $A(i)$. Here it is assumed that $\pi_t(\cdot \mid i)$ is Lebesgue measurable for each $i$ and $a \in A(i)$. A stochastic stationary policy $\pi_0 \in \Pi_s$ is a Markov policy $\pi = (\pi_t)$ satisfying $\pi_t = \pi_0$ for all $t \geq 0$. A stationary policy $f \in \Pi^d_s$ is a stochastic stationary policy $\pi_0$ such that $\pi_0(f(i) \mid i) = 1$ for some
NECESSARY CONDITIONS FOR CONTINUOUS...

\( f(i) \in A(i), i \in S \). The set of decision functions is defined by \( F = \times_i A(i) \). The details can be found, in [13]. It is obvious that \( \Pi^d \) is equivalent to \( F \). For a policy \( \pi = (\pi_t) \) and \( s \geq 0 \), we define a policy \( \pi^* = (\pi^*_t) \in \Pi_m \) by \( \pi^*_t = \pi_{s+t} \) for \( t \geq 0 \). \( \pi^* \) is, in fact, the policy \( \pi \) but be delayed a time \( s \).

For any real number \( c \), we define \( c^\pm := \max\{0, \pm c\} \). For any sequence \( \{c_n, n \geq 0\} \), if \( \sum n c_n^+ \) or \( \sum n c_n^- \) is finite, then we define the series \( \sum n c_n := \sum n c_n^+ - \sum n c_n^- \); otherwise, we say that the series \( \sum n c_n \) is undefined.

For any policy \( \pi = (\pi_t) \in \Pi_m \) and \( t \geq 0 \), we define a matrix \( Q(\pi, t) = (q_{ij}(\pi, t)) \) and a vector \( r(\pi, t) = (r_i(\pi, t)) \) by:

\[
q_{ij}(\pi, t) = \sum_{a \in A(i)} q_{ij}(a) \pi_t(a \mid i), \quad r_i(\pi, t) = \sum_{a \in A(i)} r(a) \pi_t(a \mid i).
\]

Thus, \( q_{ij}(\pi, t) \) and \( r_i(\pi, t) \) are respectively the state transition rate family and the reward rate function under policy \( \pi \). It is apparent that \( \lambda(i) < \infty \) is necessary to ensure the finiteness of \( q_{ij}(\pi, t) \). While for \( r_i(\pi, t) \), we assume that it is well defined first. Lemma 1 below will discuss some for it in detail. 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 \( r(\pi_0) = (r_i(\pi_0)) \). The following condition is about the process under each policy and is assumed throughout the paper.

**Condition A.** For any Markov 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 \( 0 \leq s \leq t \leq u < \infty \),

\[
\frac{\partial}{\partial t} P(\pi, s, t) = P(\pi, s, t)Q(\pi, t), \quad P(\pi, s, u) = P(\pi, s, t)P(\pi, t, u),
\]

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

One can find the constructing algorithm for the minimal \( Q \)-process in [16] (II. 17) for stationary case and in [17] for nonstationary case. Condition A is true when \( q_{ij}(a) \) is bounded, or under the assumptions presented in [8] when \( q_{ij}(a) \) is unbounded. This paper except Section 5 will also deal with the unbounded case, though the boundedness of \( q \) will make our discussions more easier.

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, …, N, i ∈ S} ⊂ Πm, 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 \( \pi^{t_{n+1} - t_n}(\cdot | j) \) 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 \( \Pi_d(s) \). Especially, \( \pi \in \Pi_m \) is a piecewise policy with \( N = 0 \) and \( \pi^{0,i} \) is independent of \( i \).

For such a policy \( \pi \), if \( Y(t_n) = i \), then the system in \([t_n, t_{n+1}]\) is a Markov process with transition probability matrix \( P^{n,i}(s, t) \). So, the system under a piecewise semi-Markov policy is a special case of piecewise Markov process (see [18]). 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

\[
P_{ij}^k(\pi, s, t) := P_{s\{Y(t) = j \mid Y(s) = i, Y(t_m) = k\}} = \sum_{j_1} P_{ij_1}(\pi^{m,k}, s - t_m, t_{m+1} - t_m) \times \sum_{j_2} P_{j_1j_2}(\pi^{m+1,j_1}, 0, t_{m+2} - t_{m+1}) \cdots \times \sum_{j_n} P_{j_{n-1}j_n}(\pi^{n-1,j_{n-1}}, 0, t_n - t_{n-1}) \times P_{j_{n-1}j_n}(\pi^{n,j_{n-1}}, t_n, t).
\]

For \( i, j \in S \), let \( P_{ij}(\pi, t) = P_{ij}^i(\pi, 0, t) \) be the state transition probability to reach the state \( j \) at time \( t \) from the state \( i \) at the initial time 0 under a policy \( \pi \).

Under a piecewise semi-Markov policy, the process is divided into several sub-processes, which make it possible to apply the Bellman optimality principle and then to obtain the optimality equation here. This is important, e.g., in the proof of Theorem 2, 3 and 5 below. But it will be proved that such policy does not improve the optimality under certain conditions. So, the introduced piecewise semi-Markov policies is only for the proof of our results.

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

\[
U_\alpha(\pi) = \int_0^\infty \exp(-\alpha t)P(\pi, t)r(\pi, t)dt ,
\]

(3)
where the integral is the Lebesgue integral. It is the expected discounted total rewards on the whole time axis under \( \pi \). 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 \( \pi \). Similarly, for a piecewise semi-Markov policy \( \pi = (\pi^{n,i}) \in \Pi_m(s) \) with \( \{t_n, n = 1, 2, \cdots, N\} \) and \( t \geq 0 \), we define inductively

\[
U^{n,k}_\alpha(\pi, t, i) = \int_t^{t_{n+1}} \exp(-\alpha(s - t)) \sum_j P_{ij}(\pi^{n,k}, t, s)r_j(\pi^{n,k}, s)ds
\]

\[
+ \exp(-\alpha(t_{n+1} - t)) \times \sum_j P_{ij}(\pi^{n,k}, t, t_{n+1})U^{n+1,j}_\alpha(\pi, t_{n+1}, j),
\]

\[ t \in [t_n, t_{n+1}), \ n = 0, 1, \cdots, N - 1, \ k, i \in S, \]

\[
U^{N,k}_\alpha(\pi, t, i) = U_\alpha(\pi^{N,k}, t - t_N, i), \ t \geq t_N, k, i \in S. \tag{5}
\]

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

Having defined the objective function, we now present the second condition.

**Condition B.** \( U_\alpha(\pi) \) is well defined (may be infinite) for each policy \( \pi \in \Pi_m(s) \).

The meaning of Condition B has three aspects:1) \( \sum_j P_{ij}(\pi, t)r_j(\pi, t) \), and furthermore, the integral in equation (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 equation (5) is well-defined.

The above 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. For example, when \( \alpha > 0 \) and \( r(i, a) \) has a upper bound \( M \), then for any \( \pi \in \Pi_m \),

\[
\int_0^\infty \exp(-\alpha t)[P(\pi, t)r(\pi, t)]^+_i dt \\
\leq \int_0^\infty \exp(-\alpha t) \sum_j P_{ij}(\pi, t)r^+_j(\pi, t)dt \leq \alpha^{-1} M,
\]
which together with the definition of the Lebesgue integral imply that \( U_\alpha(\pi) \) is well defined and bounded above for \( \pi \in \Pi_m \), and thus \( U^{n,k}_\alpha(\pi, t, i) \) in equation (5) is also well defined and bounded above. Condition B is assumed to be true throughout the paper.

Condition A and Condition B say respectively that for each policy \( \pi \), the process and the objective function are well-defined. We say that the CTMDP model equation (1) is well defined if both Condition A and Condition B are true. Surely, it is impossible to discuss the CTMDP equation (1) if one of these two conditions does not hold.

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 equation (5) that for \( \pi \in \Pi_m \) and \( t \geq 0 \),

\[
U_\alpha(\pi) = \int_0^t \exp(-\alpha s)P(\pi, s)r(\pi, s)ds + \exp(-\alpha t) \times P(\pi, t)U_\alpha(\pi, t),
\]

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

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

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

Let the optimal value function be \( U^*_\alpha(i) = \sup\{U_\alpha(\pi, i) \mid \pi \in \Pi_m(s)\} \) 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 \( \pi^* \) is called \( \varepsilon \)-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 \( i \rightarrow j \)) 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 \( i \rightarrow j \) 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 \( q_{ij_1}(f)q_{ij_2}(f)\ldots q_{ijn}(f) > 0 \). It is apparent that if \( i \rightarrow j \) and \( j \rightarrow k \), 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 $S_0^* = \{i \mid i \to S_0\}$ 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 $j \notin S_0$, or equivalently, $(S - S_0)^* = S - S_0$. Similarly as above, $S_0$ is closed iff $P_{ij}(\pi, t) = 0$ for all $i \in S_0, \pi \in \Pi_m(s), j \notin S_0$ 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 Condition B are also true for $S_0^{-}\text{CTMDP}$. Let its objective function be $U_\alpha^{S_0}(\pi)$.

We have the following obvious theorem.

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

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: $S_0^{-}\text{CTMDP}$ and $(S - S_0)^{-}\text{CTMDP}$. On the other hand, if $S_0$ is closed while $U_\alpha^*(i)$ for $i \in S - S_0$ is known, or a $(\varepsilon)$-optimal policy can be obtained in $S - S_0$, then one need to discuss only $S_0^{-}\text{CTMDP}$. 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 $\pi$ of the (original) CTMDP there is a policy $\pi'$ 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 $i \in S$.

Surely, any policy $\pi'$ for CTMDP' is also a policy for the original CTMDP and $U_\alpha'(\pi, i) = U_\alpha(\pi, i)$ for all $i$. So, when $A(i)$ can be reduced as $A_1(i)$, the optimal value function of CTMDP equals obviously to that of CTMDP'. Thus, we say that they are equivalent.

Now, for $i \in S$, we denote by $U(i) = \sup\{r(i, a) \mid a \in A(i)\}$ and $L(i) = \inf\{r(i, a) \mid a \in A(i)\}$ respectively the supremum and infimum of the reward rate
function $r(i, a)$ over the action set $A(i)$. Let $S_U = \{i \mid U(i) = +\infty\}$, $S_{-\infty} = \{i \mid r(i, a) = +\infty\}$, there is $\pi \in \Pi_m(s)$ such that $U_\alpha(\pi, i) = +\infty$, $S_\infty = \{i \mid U_\alpha^*(i) = +\infty\} - S_{-\infty}$, $S_{-\infty} = \{i \mid U_\alpha^*(i) = -\infty\}$, $S_0 = S - S_{-\infty} - S_\infty - S_{-\infty} = \{i \mid -\infty < U_\alpha^*(i) < \infty\}$. 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 $U(i) = +\infty$ cannot be true simultaneously.

**Proof.** 1) For $i \in S_U$, if there is $a \in A(i)$ such that $r(i, a) = +\infty$, then we define $\pi_0(a \mid i) = 1$. Otherwise, there is a subset $\{a_n, n \geq 1\} \subset A(i)$, which are different with each other, such that $r(i, a_n) \geq n$. Fixing a constant $\delta \in (0, 1)$, let $c = \sum_{n=1}^{\infty} n^{-(1+\delta)} < \infty$, and define $\pi_0(a_n \mid i) = (cn^{1+\delta})^{-1}$ for $n \geq 1$. For $i \in S_U, \pi_0(\cdot \mid i)$ can be defined arbitrarily. Then, it is easy to prove that for any $i \in S_U$, $r_i(\pi_0) = +\infty$, and, so $U_\alpha(\pi_0, i) = +\infty$ by $-q_{ii}(\pi_0) < +\infty$.

2) can be proved similarly as 1).

3) If $L(i) = -\infty$ and $U(i) = +\infty$ for some $i \in S$, let $\pi_0^{(1)}$ and $\pi_0^{(2)}$ be respectively the policies in 1) and 2). Then for a policy $\pi_0$ defined by $\pi_0(\cdot \mid i) := 0.5\pi_0^{(1)}(\cdot \mid i) + 0.5\pi_0^{(2)}(\cdot \mid i), r_i(\pi_0)$ is undefined, and so $U_\alpha(\pi_0, i)$ is also undefined, which contradicts Condition B. □

**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

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

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

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

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

**Proof.** 1) For any state $i \in S_{-\infty}^*$, there are a state $j_0 \in S_{-\infty}$, a policy $\pi^* \in \Pi_m$ and $t^* \geq 0$ such that $P_{ij_0}(\pi^*, t^*) > 0$ by the definition. Taking any policy $\pi' \in \Pi_m(s)$ with $U_\alpha(\pi', j_0) = +\infty$, we define a policy $\pi$ by using $\pi^*$ in
[0, t^*) and \( \pi' \) in \([t^*, \infty)\). Then it is easy to see that \( U_\alpha(\pi, i) = +\infty \) and \( i \in S_{-\infty} \).

So, \( S_{-\infty} = S_{\infty} \) and \( S' := S - S_{-\infty} \) is closed.

2) For any policy \( \pi \in \Pi_m \), state \( i \in S' - S_{-\infty} \) and action \( a \in A(i) \), it can be assumed by equation (5) that there is \( t^* > 0 \) with \( \pi_t(a \mid i) > 0 \) for \( t \leq t^* \).

If \( r(i, a) = -\infty \), then \( r_i(\pi, t) = -\infty \) for \( t \leq t^* \), which implies that \( U_\alpha(\pi, i) = -\infty \) from equation (5).

If there is a state \( j_0 \in S_{-\infty} \) with \( q_{ij_0}(a) > 0 \), then \( P_{ij_0}(\pi, t^*) > 0 \) by the construction of the minimal \( Q \)-process, which together with \( U_\alpha^{j_0}(j_0) = -\infty \) and equation (5) also imply \( U_\alpha(\pi, i) = -\infty \).

So, \( A(i) \) can be reduced as \( A_1(i) \) for \( i \in S' - S_{-\infty} \). It is apparent that \( S^*_{-\infty} = S_{-\infty} \) after this reduction.

3) First, it should be noted that equation (8) is also true for \( \pi \in \Pi_m(s) \), thus it is apparent that \( A_1(i) \) can be reduced as \( A_2(i) \). After this reduction, for any state \( i \in S^*_s \), if there are a state \( j_0 \in S_{\infty} \) and an action \( a \in A_2(i) \) such that \( q_{ij_0}(a) > 0 \), then there is a policy \( \pi \) with \( U_\alpha(\pi, i) > -\infty \) from the definition of \( A_2(i) \), and there is \( t^* > 0 \) such that \( P_{ij_0}(\pi, t^*) > 0 \) by the construction of the minimal \( Q \)-process. Thus, we can get by equation (5) that

\[
c := \int_0^{t^*} \exp(-\alpha s) \sum_j P_{ij}(\pi, s)r_j(\pi, s)ds + \exp(-\alpha t^*) \sum_{j \neq j_0} P_{ij}(\pi, t^*)U_\alpha(\pi, t^*, j) > -\infty.
\]

Now, for any constant \( M > 0 \), taking any policy \( \pi^M \in \Pi_m(s) \) with \( U_\alpha(\pi^M, j_0) > M \), we define a policy \( \pi^* = (\pi^0_j, \pi^1_j; j \in S) \) by \( \pi^0_j = \pi, \pi^1_j = \pi^* \) for \( j \neq j_0, \pi^{1j_0} = \pi^M \), and \( t_1 = t^* \). Then, by equation (5) we have

\[
U_\alpha(\pi^*M, i) \geq c + \exp(-\alpha t^*)P_{ij_0}(\pi, t^*)M.
\]

Letting \( M \to \infty \), we get that \( U_\alpha^*(i) = +\infty \), i.e., \( i \in S_{\infty} \). So, \( S^*_{\infty} = S_{\infty} \). \( \square \)

By Theorem 1 and Theorem 2, the state space \( S \) 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 \( \pi \), 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:

\[
S_0-\text{CTMDP} = \{S_0, A_2(i), q_{ij}(a), r(i, a), U_\alpha\}.
\]
Because $i \in S_{-\infty}$ when $A_2(i) = \emptyset$, equation (9) is a CTMDP. Furthermore, we have
\[ -\infty < U^*_\alpha(i) < +\infty, \quad -\infty < r(i, a) \leq U(i) < +\infty, \quad \forall i, a. \] (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, unless a special statement is given.

4. Some Properties

This section discusses some properties of $S_0$-CTMDP equation (9) and simplifies the expression of $A_2(i)$. First, the following lemma is from [16] (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
\[ p_{ij}(t) = \exp(-q_it) \int_0^t \exp(q_is)q_i g_{ij}(s)ds + \exp(-q_it)\delta_{ij}, \]
\[ i, j \in S, \quad t \geq 0 \]
where $\delta_{ij}$ denotes the Kronecker delta function, and for $s > 0, t \geq 0$,
\[ \lim_{s \to 0^+} g_{ij}(s) = (1 - \delta_{ij})q_{ij}/q_i, \quad \sum_j g_{ij}(s) = 1, \quad g_{ij}(s+t) = \sum_k g_{ik}(s)p_{kj}(t). \]

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
\[ h_i(t) := q_i \exp(q_it) \sum_{j \in Z} g_{ij}(t) u_j \]
is finite and continuous in $[0, t^*)$, and $\sum_{j \in Z} q_{ij} u_j < \infty$; otherwise, $h_i(t) = +\infty$ for all $t > 0$. 
Proof. It follows from Lemma 2 that
\[\sum_{j \in Z} p_{ij}(t)s u_j = \exp(-q_i t^*) \int_0^{t^*} h_i(s) ds + \exp(-q_i t^*) u_i \chi_Z(i)\]
where \(\chi_Z\) is the indicator function of the set \(Z\). Again, it follows from Lemma 2 that \(g_{ij}(s + t) \geq g_{ij}(s)p_{ij}(t)\) for \(s > 0, t \geq 0\) and \(j \in S\), which is also true for \(s = 0\) by the continuity of \(g_{ij}(t)\). For the boundedness of \(g_i\) and [16] (equation (6) in pp. 130), one knows that for any \(\varepsilon \in (0, 1)\), there is a constant \(\delta' > 0\) such that \(p_{ij}(t) > 1 - \varepsilon\) for each \(j \in S\) and \(t < \delta'\). So,
\[
g_{ij}(s + t) \geq (1 - \varepsilon) g_{ij}(s), \quad s \geq 0, t < \delta', j \in S, \tag{11}\]
and therefore
\[
h_i(s + t) \geq (1 - \varepsilon) h_i(s), \quad s \geq 0, t < \delta'. \tag{12}\]

1) Suppose that \(\sum_{j \in Z} p_{ij}(t^*) u_j\) is finite, then \(\int_0^{t^*} h_i(s) ds\) is finite, and so \(h_i(t)\) is finite for a.e. \(t \in [0, t^*]\), which implies that there is a constant \(\delta < \delta'\) such that \(h_i(\delta)\) is finite. By equation (11)
\[
g_{ij}(t) \leq \frac{1}{1 - \varepsilon} g_{ij}(t + (\delta - t)) = \frac{1}{1 - \varepsilon} g_{ij}(\delta), \quad j \in S, t < \delta.
\]
So, \(\sum_{j \in Z} g_{ij}(t) u_j\), as a series, is uniformly convergent in \([0, \delta]\) by the finiteness of \(h_i(\delta)\). It can be proved similarly that this series is uniformly convergent in any subinterval of \([0, t^*]\) with its length being less than or equal to \(\delta'\), which results in that the series is uniformly convergent in \([0, t^*]\). But \(g_{ij}(t)\) is continuous. So, \(h_i(t)\) is also continuous in \([0, t^*]\). Now, \(\sum_{j \in Z} q_i u_j < \infty\) follows \(h_i(0) < \infty\).

2) If \(\int_0^t h_i(t) dt = \infty\) for each \(t > 0\), then there is a decreasing sequence \(t_n \to 0\) such that \(h_i(t_n) \to +\infty\) as \(n \to +\infty\). Fixing any \(0 < t < \delta'\), one has that \(t_n < t\) for sufficiently large \(n\), and thus by equation (12),
\[
h_i(t) = h_i(t_n + (t - t_n)) \geq (1 - \varepsilon) h_i(t_n), \quad 0 < t_n < t.
\]
Letting \(n \to \infty\) implies that \(h_i(t) = \infty\), which is still true for all \(t > 0\) by equation (12). \(\square\)

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
\[
\frac{d}{dt} \{\sum_j p_{ij}(t) u_j\} = \sum_j \frac{d}{dt} p_{ij}(t) u_j = \sum_j \{-q_i p_{ij}(t) + q_i g_{ij}(t)\} u_j.
\]
Proof. Applying Lemma 2 respectively to the positive and negative parts of $\sum_j p_{ij}(t)u_j$ result in that

$$\sum_j p_{ij}(t)u_j = \exp(-q_i t)\left\{ \int_0^t \exp(q_i s)q_i \sum_j g_{ij}(s)u_j ds + u_i \right\}, \quad t \in [0, t^*].$$

But, it follows from Lemma 3 that the integrand above is continuous. So, $\sum_j p_{ij}(t)u_j$ is differentiable and its derivative

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

is continuous in $[0, t^*]$. On the other hand, by Lemma 2 again, one gets that

$$\sum_j \frac{d}{dt}p_{ij}(t)u_j = \sum_j \{ -q_i p_{ij}(t) + q_i g_{ij}(t) \} u_j, \quad t \in [0, t^*]. \quad \square$$

Having the above several lemmas for preparation, we can now prove the following theorem, where $S^+ := \{ i \in S_0 \mid U^*_\alpha(i) \geq 0 \}$ and $S^- := \{ i \in S_0 \mid U^*_\alpha(i) < 0 \}$ are the state subsets with nonnegative and negative optimal values respectively.

**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, $U_\alpha(\pi, i) = -\infty$.

Proof. 1) For $\pi \in \Pi_m, t > 0$ and any $\pi^* = (\pi^{0,i}, \pi^{1,i}; i \in S)$ with $\pi^{0,i} = \pi$ and $t_1 = t$,

$$U_\alpha(\pi^*, i) = \int_0^t \exp(-\alpha s) \sum_j P_{ij}(\pi, s)r_j(\pi, s)ds + \exp(-\alpha t) \sum_j P_{ij}(\pi, t)U_\alpha(\pi^{1,j}, j) \leq U^*_\alpha(i) < \infty$$

is well defined for $i \in S$, which implies that the second term above, $\sum_j P_{ij}(\pi, t)U_\alpha(\pi^{1,j}, j)$, is also well defined and less than infinity. Now, for any $\varepsilon > 0$ and
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 \( a \in A(i) \).

Proof. For any given \( i \in S \) and \( a \in A(i) \), let \( f \) satisfy \( f(i) = a \) and \( f(j) = f^*(j) \) for \( j \neq i \). Then \( Q(f) \) is bounded and \( \sum_{j \in S^+} P_{ij}(f, t)U^*_\alpha(j) < \infty \) by Theorem 3. So, the corollary follows from Lemma 3. \( \square \)

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

\[
\frac{d}{dt} \left[ \sum_{j} P_{ij}(f, t)U^*_\alpha(j) \right] = \sum_{j} \frac{d}{dt} P_{ij}(f, t)U^*_\alpha(j).
\]

\[
\sum_{j} [P(f, t)Q(f)]_{ij}U^*_\alpha(j) = \sum_{j} P_{ij}(f, t)[Q(f)U^*_\alpha], \quad t \in [0, t^*]. \tag{13}
\]

Proof. The results except equation (13) follow Lemma 4 and Corollary 1. Equation (13) follows the arguments below. It can be seen that

\[
\sum_{k \in S} P_{ik}(f, t)\{ \sum_{j \in S^+} q_{kj}(f)U^*_\alpha(j) \}
\]

\[
= \sum_{k \in S^+} P_{ik}(f, t)\{ \sum_{j \in S^+} q_{kj}(f)U^*_\alpha(j) \} + \sum_{k \in S^-} P_{ik}(f, t) \sum_{j \in S^+} q_{kj}(f)U^*_\alpha(j)
\]

\[
= \sum_{j \in S^+} \{ \sum_{k \in S^+} P_{ik}(f, t)q_{kj}(f) + \sum_{k \in S^-} P_{ik}(f, t)q_{kj}(f) \}U^*_\alpha(j)
\]

\[
= \sum_{j \in S^+} \{ \sum_{k} P_{ik}(f, t)q_{kj}(f) \}U^*_\alpha(j) = \sum_{j \in S^+} \frac{d}{dt} P_{ij}(f, t)U^*_\alpha(j)
\]
is finite. Similarly,
\[ \sum_{k \in S} P_{ik}(f, t) \{ \sum_{j \in S^-} q_{kj}(f) U_{ij}^*(j) \} = \sum_{k \in S^-} \{ \sum_{j \in S^-} P_{ik}(f, t) q_{kj}(f) \} U_{ij}^*(j) \]
is also finite. Subtracting the latter one from the former one results that
\[ \sum_{k \in S} P_{ik}(f, t) \{ \sum_{j \in S^-} q_{kj}(f) U_{ij}^*(j) \} = \sum_{j \in S^-} \{ \sum_{k \in S^-} P_{ik}(f, t) q_{kj}(f) \} U_{ij}^*(j) \]
is well-defined and is finite. So, equation (13) is true. □

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 this section, we discuss a simplified expression for \( A_2(i) \) defined in equation (8), and the symbols \( A_1(i), A_2(i), S_0, \) etc., are as in the former sections.

**Theorem 4.** If \( q_{ij}(a) \) is uniformly bounded, then
\[ A_2(i) \subseteq \{ a \in A_1(i) \mid \sum_{j} q_{ij}(a) U_{ij}^*(j) > -\infty \}, \quad i \in S. \] (14)

Moreover, if \( h_i(t) \) is finite and continuous whenever \( \sum_{j \in Z} q_{ij} u_j \) is finite in Lemma 3, then
\[ A_2(i) = \{ a \in A_1(i) \mid \sum_{j} q_{ij}(a) U_{ij}^*(j) > -\infty \}, \quad i \in S. \] (15)

**Proof.** First, we show that \( Q(\pi, t) U_{ij}^* < +\infty \) is well defined. If there is a policy \( \pi \) and a state \( i \) such that \( \sum_{j \in S^+} q_{ij}(\pi, t) U_{ij}^*(j) = +\infty \) in a set with positive measure, e.g., in \([0, t^*]\) for some \( t^* > 0 \), then by the construction of the minimal \( Q \)-process one can get that
\[ \sum_{j \in S^+} P_{ij}(\pi, t) U_{ij}^*(j) \geq +\infty, \quad t > 0, \]
which contradicts 1) of Theorem 3. So, \( Q(\pi, t) U_{ij}^* < +\infty \) is well defined.

Now, for any \( i \in S \) and \( a \in A_1(i) \), if \( \sum_{j} q_{ij}(a) U_{ij}^*(j) = -\infty \), then \( \sum_{j \in S^+} q_{ij}(a) U_{ij}^*(j) < +\infty \) and \( \sum_{j \in S^-} q_{ij}(a) U_{ij}^*(j) = -\infty \). Thus, for any policy \( \pi \) satisfying the condition in equation (8),
\[ \sum_{j} q_{ij}(\pi, t) U_{ij}^*(j) = -\infty, \quad \forall t \in E(\pi, i, a). \]
NECESSARY CONDITIONS FOR CONTINUOUS...

Then $\sum_j P_{ij}(\pi, t)U^*_\alpha(j) = -\infty$, which together with Theorem 3 imply that $U^*_\alpha(\pi, i) = -\infty$. So, equation (14) is true.

Now, suppose that $h_i(t)$ is finite and continuous whenever $\sum_{j \in Z} q_{ij}u_j$ is finite in Lemma 3. For any $i \in S$ and $a \in A_1(i)$ with $\sum_j q_{ij}(a)U^*_\alpha(j) > -\infty$, we say that there is $f^* \in F$ with

$$\sum_j q_{ij}(a)r(j, f^*) > -\infty. \tag{16}$$

Otherwise, $\sum_j q_{ij}(a)r(j, f) = -\infty$ for each $f$, which implies that for any $\pi$ and $t$ we have that $\sum_j q_{ij}(a)r_j(\pi, t) = -\infty$. Again by the construction of the minimal $Q$-process, $\sum_j P_{ij}(\pi, t)r_j(\pi, t) = -\infty$ for $t > 0$. So, $U^*_\alpha(\pi, i) = -\infty$ for each $\pi$ by equation (5), which results in that $A_2(i)$ is empty, a contradiction. Thus, there is $f^*$ satisfying equation (16). Now, let $f$ with $f(i) = a$ and $f(j) = f^*(j)$ for $j \neq i$. Then both $\sum_j q_{ij}(f)r(j, f)$ and $\sum_j q_{ij}(f)U^*_\alpha(j)$ are finite, which imply that both $\sum_j P_{ij}(f, t)U^*_\alpha(j)$ and $\sum_j P_{ij}(f, t)r(j, f)$ are finite, and thus continuous, by Lemma 3. So, $U^*_\alpha(f, i)$ exists and is finite by equation (5), and $a \in A_2(i)$.

**Remark 1.** 1) By the above theorem, if $S^-$ is finite, or $U^*_\alpha(i)$ is bounded below, then equation (15) is true when $q_{ij}(a)$ is uniformly bounded; 2) $S^-$ is empty if $U^*_\alpha$ 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., $\lambda = \sup\{-q_{ii}(a) \mid i \in S, a \in A(i)\} < \infty$. For $\pi \in \Pi_\alpha(s), t \geq 0$ 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 by $U^*_\alpha(\pi, t) = U^*_\alpha(\pi, t, U^*_\alpha)$ for short, which is well-defined by Theorem 3. Certainly, $U^*_\alpha(\pi, t)$ is the expected discounted total rewards if $\pi$ is used in $[0, t]$ and then an optimal policy is used from $t$.

**Lemma 5.** $U^*_\alpha = \sup\{U^*_\alpha(\pi, t) \mid \pi \in \Pi_\alpha(s)\}$ for $t \geq 0$, and $U^*_\alpha(\pi, t)$ is nonincreasing in $t$ for any $\pi \in \Pi_\alpha(s)$. 

Proof. 1) It follows from equation (6) that \( U^*_\alpha \leq \sup \{ U^*_\alpha(\pi, t) \mid \pi \in \Pi_m(s) \} \). On the other hand, for any given \( t > 0 \) and \( \varepsilon > 0 \), we take \( \pi^i \in \Pi_m(s) \) with \( U_\alpha(\pi^i, i) \geq U^*_\alpha(i) - \varepsilon \) for \( i \in S \). Then, we define for each \( \pi \) a piecewise semi-Markov policy \( \pi(\varepsilon) \) by using \( \pi \) in \([0, t)\) and using \( \pi^i \) in \([t, +\infty)\) if \( Y(t) = i \). Thus,

\[
U^*_\alpha(\pi, t) \leq U_\alpha(\pi(\varepsilon)) + \exp(-\alpha t)\varepsilon \leq U^*_\alpha + \exp(-\alpha t)\varepsilon \cdot e,
\]

where \( e \) is the vector with all components being one. This results in that \( \sup_{\pi} U^*_\alpha(\pi, t) \leq U^*_\alpha \) by the arbitrariness of \( \pi \) and \( \varepsilon \). So, the first result is true.

2) It follows from equations (5) and (6) that for each \( \pi \in \Pi_m \) and \( t' < t \),

\[
\int_{t'}^t \exp(-\alpha s)P(\pi, t')P(\pi, t', s)r(\pi, s)ds = P(\pi, t') \int_{t'}^t \exp(-\alpha s)P(\pi, t', s)r(\pi, s)ds,
\]

which together with 1) and equation (6) imply that for \( t' < t \),

\[
U^*_\alpha(\pi, t) = \int_0^{t'} \exp(-\alpha s)P(\pi, s)r(\pi, s)ds + \exp(-\alpha t')P(\pi, t')U^*_\alpha(\pi, t') \leq \int_0^{t'} \exp(-\alpha s)P(\pi, s)r(\pi, s)ds + \exp(-\alpha t')P(\pi, t')U^*_\alpha(\pi, t') = U^*_\alpha(\pi, t').
\]

Obviously, the above holds also for \( \pi \in \Pi_m(s) \) by equation (5). □

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 \( U^*_\alpha \) is bounded below (see Remark 1); 2) for each \( i \in S \), \( A(i) \) can be reduced as

\[
A'(i) = \{ a \in A(i) \mid \sup_{f \in F; f(i) = a} U_\alpha(f, i) > -\infty \},
\]
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 from 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, $U^*_\alpha$ satisfies the following optimality equation:

$$
\alpha U^*_\alpha(i) = \sup_{a \in A(i)} \{r(i, a) + \sum_j q_{ij}(a)U^*_\alpha(j)\}, \quad i \in S.
$$

(17)

**Proof.** 1) For any given $i$ and $a$, it follows from Condition C that there are $f$ and $t^* > 0$ such that $f(i) = a$ and $U^*_\alpha(f, t^*, i) > -\infty$, which together with Lemma 5 that $U^*_\alpha(f, t, i) > -\infty$ for $t \leq t^*$. Thus, by Lemma 5 and Corollary 2 one can get that for $t \in [0, t^*)$,

$$
0 \geq \frac{d}{dt} U^*_\alpha(f, t, i) = \exp(-\alpha t) \sum_j P_{ij}(f, t)\{r(j, f) + \sum_k q_{jk}(f)U^*_\alpha(k) - \alpha U^*_\alpha(j)\}.
$$

Its right hand side is continuous and so it is true for all $t < t^*$. Taking $t = 0$, one gets that

$$
\alpha U^*_\alpha(i) \geq r(i, a) + \sum_j q_{ij}(a)U^*_\alpha(j).
$$

By the arbitrariness of $i$ and $a$, one gets further

$$
\alpha U^*_\alpha(i) \geq \sup_{a \in A(i)} \{r(i, a) + \sum_j q_{ij}(a)U^*_\alpha(j)\}, \quad i \in S.
$$

2) If the above inequality is strict for some $i_0$, then there is $\varepsilon^* > 0$ such that

$$
\alpha U^*_\alpha(i) \geq r(i, a) + \sum_j q_{ij}(a)U^*_\alpha(j) + \varepsilon_i, \quad \forall i, a,
$$

where $\varepsilon_{i_0} = \varepsilon^*$ and $\varepsilon_i = 0$ for $i \neq i_0$. So,

$$
\alpha U^*_\alpha \geq r(\pi, t) + Q(\pi, t)U^*_\alpha + \varepsilon, \quad \pi \in \Pi_m(s), t \geq 0.
$$

(18)

Fixing any $t^* > 0$, for any $\pi \in \Pi_m(s)$, if $U^*_\alpha(\pi, t^*, i_0) > -\infty$, then $U^*_\alpha(\pi, t, i_0)$, and so $\sum_j P_{i_0j}(\pi, t)U^*_\alpha(j)$, are finite in $[0, t^*]$. Due to the uniformly boundedness
of \{q_{ij}(a)\}, we know that \(\sum_j P_{ioj}(\pi, t)q_{jj}(\pi, t) U^*_\alpha(j)\) is finite for \(t \in [0, t^*]\). So,

\[
\sum_j P_{ioj}(\pi, t)\left[\sum_k q_{jk}(\pi, t)U^*_\alpha(k)\right] = \sum_j P_{ioj}(\pi, t)\left[\sum_{k \neq j} q_{jk}(\pi, t)U^*_\alpha(k) + \sum_j P_{ioj}(\pi, t)q_{jj}(\pi, t)U^*_\alpha(j)\right] = \sum_{k \neq j} \sum_j P_{ioj}(\pi, t)q_{jk}(\pi, t)U^*_\alpha(k) + \sum_k P_{iok}(\pi, t)q_{kk}(\pi, t)U^*_\alpha(k)
\]

is well defined. Premultiplying equation (18) by \(\exp(-\alpha t)P(\pi, t)\), one can get that

\[
0 \geq \sum_j \frac{d}{dt} [\exp(-\alpha t)P_{ioj}(\pi, t)U^*_\alpha(j)] + \exp(-\alpha t)\sum_j P_{ioj}(\pi, t)[r_j(\pi, t) + \varepsilon_j], \ t \leq t^*. \tag{19}
\]

Now,

\[
\sum_j \exp(-\alpha t)\frac{d}{dt} P_{ioj}(\pi, t)U^*_\alpha(j) = \sum_j \exp(-\alpha t)\left[\sum_{k \neq j} P_{ioj}(\pi, t)q_{kj}(\pi, t) + P_{ioj}(\pi, t)q_{jj}(\pi, t)\right]U^*_\alpha(j) = \sum_{j \in S^+} \exp(-\alpha t)\sum_{k \neq j} P_{ioj}(\pi, t)q_{kj}(\pi, t)U^*_\alpha(j) + \sum_{j \in S^-} \exp(-\alpha t)\sum_{k \neq j} P_{ioj}(\pi, t)q_{kj}(\pi, t)U^*_\alpha(j)
\]

Because \(\sum_{k \neq j} P_{ioj}(\pi, t)q_{kj}(\pi, t) \geq 0\) and \(q_{jj}(\pi, t) \leq 0\), it follows from the above
formula that
\[ \int_0^{t^*} \sum_j \exp(-\alpha t) \frac{d}{dt} P_{\alpha j}(\pi, t) U_\alpha^*(j) \]
\[ = \sum_j \int_0^{t^*} \exp(-\alpha t) \left[ \sum_{k \neq j} P_{\alpha k}(\pi, t)q_{kj}(\pi, t) + P_{\alpha j}(\pi, t)q_{jj}(\pi, t) \right] U_\alpha^*(j) dt \]
\[ = \sum_j \int_0^{t^*} \exp(-\alpha t) \frac{d}{dt} P_{\alpha j}(\pi, t) U_\alpha^*(j) dt. \]

Similarly, we have
\[ \int_0^{t^*} \sum_j \exp(-\alpha t) P_{\alpha j}(\pi, t) U_\alpha^*(j) dt \]
\[ = \sum_j \int_0^{t^*} \exp(-\alpha t) P_{\alpha j}(\pi, t) U_\alpha^*(j) dt. \]

So,
\[ \int_0^{t^*} \sum_j \frac{d}{dt} \left[ \exp(-\alpha t) P_{\alpha j}(\pi, t) U_\alpha^*(j) \right] dt \]
\[ = \sum_j \int_0^{t^*} \frac{d}{dt} \left[ \exp(-\alpha t) P_{\alpha j}(\pi, t) U_\alpha^*(j) \right] dt \]
\[ = \sum_j \left[ \exp(-\alpha t) P_{\alpha j}(\pi, t) U_\alpha^*(j) \right] |_0^{t^*} \]
\[ = \sum_j \exp(-\alpha t^*) P_{\alpha j}(\pi, t^*) U_\alpha^*(j) - U_\alpha^*(i_0). \]

Integrating equation (19) in \([0, t^*]\) implies that
\[ U_\alpha^*(i_0) \geq U_\alpha^*(\pi, t^*, i_0) + \int_0^{t^*} \exp(-\alpha t) \sum_j P_{\alpha j}(\pi, t) \varepsilon_j dt \]
\[ = U_\alpha^*(\pi, t^*, i_0) + \int_0^{t^*} \exp(-\alpha t) P_{\alpha i_0}(\pi, t) dt e^* \]
\[ \geq U_\alpha^*(\pi, t^*, i_0) + \int_0^{t^*} \exp(-\alpha t) \exp(-\lambda t) dt e^*, \quad (20) \]
where the last inequality is resulted by the construction of the minimal $Q$-process. So, $U_\alpha^*(i_0) > \sup \{U_{\alpha}(\pi, t^*, i) \mid \pi \in \Pi_m(s)\}$, which contradicts Lemma 5. Thus, equation (17) is true. \hfill \Box

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 $\Pi_m$, the set of Markov policies, iff the optimal value function restricted to $\Pi_m$ also satisfies the optimality equation (equation (17)). Let $U_{\alpha}^m = \sup \{U_{\alpha}(\pi) \mid \pi \in \Pi_m\}$. We affirm that $U_{\alpha}^m$ is finite. In fact, if $U_{\alpha}^m(i_0) = -\infty$ for some $i_0 \in S$, then it is easy to see from equation (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, $U_{\alpha}^m$ is finite.

**Theorem 6.** $U_{\alpha}^* = U_{\alpha}^m$ iff $U_{\alpha}^m$ is a solution of the optimality equation (equation (17)).

**Proof.** It suffices to prove the sufficiency. If $U_{\alpha}^m$ is a solution of equation (17), then for each $\pi \in \Pi_m$ and $i \in S$ with $U_{\alpha}(\pi, i) > -\infty$, it follows from equation (6) that $U_{\alpha}(\pi, t, U_{\alpha}^m, i)$ is finite. So, it can be proved as equation (20) that $U_{\alpha}^m(i) \geq U_{\alpha}(\pi, t, U_{\alpha}^m, i)$ for $t \geq 0$. By this, one can conclude from equation (5) that $U_{\alpha}(\pi) \leq U_{\alpha}^m$ for each $\pi \in \Pi_m(s)$. So, $U_{\alpha}^* = U_{\alpha}^m$. \hfill \Box

In order to obtain some properties for the optimality equation (equation (17)), 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, for each $t \geq 0$ and $i \in S$, $\sum_j P_{ij}(\pi, t)u(j) > -\infty$ whenever $\sum_j P_{ij}(\pi, t)u^*(j) > -\infty$. $W$ 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 $\pi$ and $u$ satisfy the following two conditions, then $u(i) \leq U_{\alpha}(\pi, i) + (\beta + \alpha)^{-1}\varepsilon$.

\[
\alpha u \leq r(\pi, t) + Q(\pi, t)u + \exp(-\beta t)\varepsilon e, \text{ a.e. } t \geq 0, \tag{21}
\]

\[
\liminf_{t \to \infty} \exp(-\alpha t) \sum_j P_{ij}(\pi, t)u(j) \leq 0. \tag{22}
\]

**Proof.** By the given conditions, it can be proved as equation (20) that

\[
u(i) \leq U_{\alpha}(\pi, t, u, i) + \int_0^t \exp(-\beta + \alpha)s ds \cdot \varepsilon.
\]

By letting $\liminf_{t \to \infty}$ above, the result follows equation (22). \hfill \Box
Theorem 7. Suppose that \( u \in W \) is a solution of the optimality equation (equation (17)) 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 equations (21) and (22), then \( u(i) \leq U_\alpha^*(i) \);

2) if \( u \) satisfies the following equation (23) for each \( \pi \in \Pi_m(s) \) with \( U_\alpha(\pi, i) > -\infty \), then \( u(i) \geq U_\alpha^*(i) \).

\[
\limsup_{t \to \infty} \exp(-\alpha t) \sum_j P_{ij}(\pi, t) u(j) \geq 0.
\] (23)

Proof. 1) \( U_\alpha(\pi, i) > -\infty \) implies that \( U_\alpha^*(\pi, t, i) \), and so \( U_\alpha(\pi, t, u, i) \), are finite for all \( t \geq 0 \) by the definition of \( W \). Thus, it follows from Lemma 6 that

\[ u(i) \leq U_\alpha(\pi, i) + (\beta + \alpha)^{-1} \varepsilon \leq U_\alpha^*(i) + (\beta + \alpha)^{-1} \varepsilon. \]

So, \( u(i) \leq U_\alpha^*(i) \) for the arbitrariness of \( \varepsilon \).

2) Because \( u \) is a solution of equation (17),

\[ \alpha u \geq r(\pi, t) + Q(\pi, t) u, \quad t \geq 0, \pi \in \Pi_m(s). \]

Then, it can be proved as equation (20) that for each \( \pi \in \Pi_m(s) \) with \( U_\alpha(\pi, i) > -\infty \), \( u(i) \geq U_\alpha(\pi, t, u, i) \) for each \( t \geq 0 \). Letting \( \limsup_{t \to \infty} \) implies \( u(i) \geq U_\alpha^*(i) \).

Equation (22) is true if \( u \leq 0 \) or if \( \alpha > 0 \) and \( u \) is bounded above, while equation (23) is true if \( u \geq 0 \) or if \( \alpha > 0 \) and \( u \) is bounded below.

It is clear that there is often a policy \( \pi = (f_t) \in \Pi_m^d \) satisfying equation (21), but it may be not true that \( U_\alpha(\pi, i) > -\infty \). On the other hand, \( U_\alpha^* \) often satisfies equation (23) for \( \pi \in \Pi_m(s) \) with \( U_\alpha(\pi, i) > -\infty \). In fact, by equation (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 equation (17) holds,

1) for any given \( f \in F \), if \( f \) attains supremum of equation (17), \( f \) and \( U_\alpha^* \) satisfy equation (22) and \( U_\alpha(f) > -\infty \), then \( f \) is optimal;
2) for some \( \pi^* \in \Pi_m(s) \), if \( U_\alpha(\pi^*) \) is a solution equation (17), then \( \pi^* \) is optimal;

3) if for any \( \varepsilon > 0 \), there is a policy \( \pi \in \Pi^d_m \) with \( U_\alpha(\pi) > -\infty \), \( \pi \) and \( U_\alpha^* \) satisfy equation (21) and equation (22) for each \( i \in S \), then \( U_\alpha^* = \sup \{ U_\alpha(\pi) \mid \pi \in \Pi^d_m \} \);

4) if \( \alpha > 0, \varepsilon \geq 0, f \in F \) attains the \( \varepsilon \)-supremum of equation (17), \( f \) and \( U_\alpha^* \) satisfy equation (22), \( U_\alpha(f) > -\infty \), then \( f \) is \( \alpha^{-1}\varepsilon \)-optimal; moreover, if such \( f \) exists for each \( \varepsilon > 0 \), then \( U_\alpha^* = \sup \{ U_\alpha(f) \mid f \in F \} \);

5) if \( U_\alpha^* \leq 0 \), then \( U_\alpha^* \) is the largest solution of equation (17) in \( W \) satisfying conditions given in 1) of Theorem 7;

6) \( U_\alpha^* \) is the smallest solution of equation (17) in \( W \) satisfying equation (23) for \( \pi \in \Pi_m(s) \) and \( i \in S \) with \( U_\alpha(\pi, i) > -\infty \).

**Corollary 4.** For each \( f \in F \) and \( i \in S \) with \( U_\alpha(f, i) > -\infty \), we have that \( \sum_j q_{ij}(f)U_\alpha(f, j) \) is finite and

\[
\alpha U_\alpha(f, i) = r(i, f) + \sum_j q_{ij}(f)U_\alpha(f, j). \tag{24}
\]

**Proof.** The equation (6) for \( f \in F \) can be rewritten as

\[
U_\alpha(f, i) = \int_0^t \exp(-\alpha s) \sum_j P_{ij}(f, s)r(j, f)ds + \exp(-\alpha t) \sum_j P_{ij}(f, t)U_\alpha(f, j).
\]

If \( U_\alpha(f, i) > -\infty \), then it follows from Corollary 2 that for \( t \geq 0 \),

\[
0 = \frac{d}{dt} U_\alpha(f, i) = \exp(-\alpha t) \sum_j P_{ij}(f, t)\{r(j, f) + [Q(f)U_\alpha(f)]_j - \alpha U_\alpha(f, j)\}.
\]

Its right hand side is continuous and so it is true for \( t = 0 \), i.e., equation (24) is true. \( \square \)

To conclude this section, we discuss the CTMDP model equation (1) restricted to \( \Pi^d_m(s) \), the set of piecewise semi-stationary policies. In this case, Theorem 2 is still true except that “\( \leq U(i) \)” should be deleted in equation (10) and \( A_2(i) \), defined by equation (8), should be redefined by
A_2(i) = \{a \in A_1(i) \mid \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 Corollary 4 also hold for \( f \). The following theorem can be proved similarly as Theorem 5 and Theorem 6.

**Theorem 8.** If the CTMDP model equation (1) is restricted to \( \Pi^d_s(s) \), then \( U_\alpha^d \) satisfies equation (17). Moreover, \( U_\alpha^s := \sup\{U_\alpha(f) \mid f \in F\} \) satisfies equation (17) iff \( U_\alpha^d = U_\alpha^s \).

### 6. Conclusions

This paper discussed the continuous time Markov decision processes with expected discounted total rewards under the necessary conditions that the model is well defined. By generalizing the concept of policies into piecewise semi-Markov policies and using the elimination of actions and the partition of state space, we first 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 the sub-CTMDP). In fact, the reward rate function of this sub-CTMDP is finite value and is bounded above on the action set for each state. 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.

### Acknowledgement

This research was supported by the National Natural Science Foundation of China (No.69904008), Institute of Applied Mathematics, Academia Sinica, China and GRANT-IN-AID FOR SCIENTIFIC RESEARCH (No. 1365 0440), Japan.

### References


