2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

Continuous Time Markov Decision Processes with Expected Discounted Total Rewards

Qiying Hu
School of Economics & Management,Xidian University,Xian,P.R.China.
Jianyong Liu
Institute of Applied Mathematics,Academia Sinica,Bejing,P.R.China.

This paper discusses continuous time Markov decision processes with criterion of expected discounted total rewards. Under the necessary condition that the model is well defined, the model is reduced into a small submodel with finite optimal value,by generalizing policies and eliminating some worst actions. Then the validity of the optimality equation is shown and some its properties are discussed about the submodel.

Markov decision processes (MDP) have been studied well since its beginning in 1960s.The standard results include three aspects: 1)the model is well defined,eg., the stochastic process under any policy is well defined, the objective function is well defined,and often is finite; 2) the optimality equation holds and the optimal value function satisfies it, or is a unique solution of it ;3)any stationary policy achieving the (ε-)supremum of the optimality equation will be (ε'-) optimal.

In literature, a MDP model was studied by first presenting some conditions, under which,the standard results 1),2) and 3) are studied successively. So one can say that these various conditions are only sufficient conditions for MDP. For the expected discounted total rewards, two simpler cases are that discount rate belongs to $(0,1)$ with uniform bounded rewards,and the discount rate is nonnegative with nonnegative or non-positive rewards. 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 we can obtain the standard results 2) and 3) by 1). We have studied it for discrete time MDP with expected discounted total rewards.

We consider the following CTMDP model $\{S, A(i), q_{ij}(a), r(i,a), U_\alpha\}$, where S is countable, $r(i,a)$ is extended real-valued and the discount rate α is a real number. The necessary conditions are as follows.

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 $0 \leq s \leq t \leq u < \infty$,

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

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

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

For any subset $S' \subset S$, define S'^* be a set of states that can reach S' and S' is called closed if its state can not reach out of S'. Let S'-CTMDP be the CTMDP restricted in S' with its objective being $U^{S'}(\pi, i)$.

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

Let $S_\infty = \{i | U_\alpha^*(i) = +\infty\}, S_{-\infty} = \{i | U_\alpha^*(i) = -\infty\}, S_0 = \{i | -\infty < U_\alpha^*(i) < \infty\}$.

THEOREM 2. For $i \in S_0, A(i)$ can be reduced as

$$A'(i) = \{a \in A(i) \mid r(i,a) > -\infty \text{ and } \sum_{j \in S_{-\infty}} q_{ij}(a) = 0\},$$
$$\text{and there is } \pi \in \Pi_m \text{ such that the Lebesgue measure of}$$
$$E(\pi, i, a) \cap [0, t] \text{ is positive for each } t > 0, U_\alpha(\pi, i) > -\infty\}$$

where $E(\pi, i, a) = \{t | \pi_t(a|i) > 0\}$.After this reducing,$S_0$ is closed.

So the state space S is divided into three disjoint subset S_∞, S_∞ and S_0 where the optimal value function is finite and the action set is $A'(i)$. So we can consider this sub-CTMDP in the following. Under some mild conditions, we have

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

CONDITION C. For any given $i \in S$ and $a \in A(i)$,there are f and $t > 0$ such that $f(i) = a$ and $U_\alpha(f, t, i) > -\infty$.

THEOREM 3. Under Condition C,U_α^* satisfies the following optimality equation:

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

THEOREM 4. Suppose that $u \in W$ is a solution of the optimality equation and $i \in S$.

1) if there is a policy $\pi \in \Pi_m(s)$ with $U_\alpha(\pi, i) > -\infty$,then $u(i) \leq U_\alpha^*(i)$;

2) if for each $\pi \in \Pi_m(s)$ with $U_\alpha(\pi, i) > -\infty, u$ satisfies

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

then $u(i) \geq U_\alpha^*(i)$.