Markov Decision Process

Depu Meng

Oct. 2018

1 Basic Concepts & Examples

Markov Decision Process (MDP) is a Markov Process that decisions are involved. Generally, an MDP can described by a quintuple:

- A Markov Process or an Extended Markov Process to describe the process.
- A state space.
- An action space.
- A state transition function.
- A performance function.

MDP can be divided into Continuous-Time MDP and Discrete-Time MDP according to the time factor of the Markov Process; MDP can also be divided into MDP and Semi-MDP and Partially-Observable MDP.

1.1 Policy and policy space

In this section, we will take policy and policy space in DTMDP as an example.

A DTMDP quintuple can be denoted as $\{X, \Phi, A, P_{ij}(a), f(i,a)\}$, $X = \{X_n; n \geq 0\}$ is a Discrete-Time Markov Process, $\Phi = \{i\}$ and $A = \{a\}$ are state space and action space of this process respectively. For $P_{ij}(a)$, appearently we have $P_{ij}(a) \geq 0$ and $\sum_{j \in \Phi} P_{ij}(a) = 1$. A DTMDP sample orbit can be described as $\{i_0, a_0, i_1, a_1, ...\}$. Denote $h_n = \{i_0, a_0, ..., i_{n-1}, a_{n-1}, i_n\}$ as the history before time n.

A general policy is defined as

$$v = (v_0(a|h_0), v_1(a|h_1), ...) \tag{1}$$

Infact, a general policy is a series of action defined on decision time, which is also a stochastic policy if not specified. The set that contains all policies like (1) is a policy space, denoted as Π.

For a policy, if for each $v_n(a|h_n)$, we select action a w.p.1, then we call it a determined policy, all determined polices is denoted as Π^d.

For a policy, if each $v_n(a|h_n)$ only related to initial state i_0 and state of time n i_n, i.e., for any n, we have $v_n(a|h_n) = v_n(a|i_0, i_n)$, then we call it a semi-Markov policy, denoted as Π_{sm}. Similarily, we can define Markov policy Π_m.

For a Markov policy, if it is also determined policy, and $v_n(a|i_n) = v(a|i)$, then we call it determined steady Markov policy Π_s^d, which is a mapping from state space to action space, i.e., $v : \Phi \to A$.

If not specified, we only need to find the optimal policy in the determined steady policy set.

1.2 Performance evaluation

Performance evaluations of DTMDP DTMDP performance evaluations can be devided into the following three classes.

$$\eta_N^v(i) = \sum_{n=0}^{N} E_v\{f(X_n, v(X_n))|X_0 = i\}, i \in \Phi \quad (2)$$

where E_v means the expectation over policy $v \in \Pi$. This is called *finite time performance evaluation.*

$$\eta_\alpha^v(i) = E_v\{\sum_{n=0}^{\infty} \alpha^n f(X_n, v(X_n))|X_0 = i\}, i \in \Phi \quad (3)$$

where α is called the discount factor. This is called *infinite time discounted performance evaluation.*

$$\eta^v(i) = \lim_{N\to\infty} \frac{1}{N} E_v\{\sum_{n=0}^{N-1} f(X_n, v(X_n))|X_0 = i\}, i \in \Phi \quad (4)$$

This is called *infinite time average performance evaluation.*

Performance evaluations of CTMDP Firstly we show the quintuple representation of a CTMDP. A CTMDP can be represented as $\{Y, \Phi, A, a_{ij}(a), f(i, a)\}$, $Y = \{Y_t; t \geq 0\}$ is a continuous-time Markov process with state space $\Phi = \{i\}$ and action space $A = \{a\}$. The transition speed $a_{ij}(a)$ satisfies for any $i, j \in \Phi, i \neq j, a \in A$, $a_{ij}(a) \geq 0, a_{ii}(a) \leq 0$ and $\sum_{j \in \Phi} a_{ij}(a) = 0$.