# 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 $P_{ij}(a)$, apparently 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 $\Pi$.

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 $\Pi^d$.

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 $\Pi_{sm}$. Similarily, we can define Markov policy $\Pi_m$.

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 $\Pi_s^d$, 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.

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

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

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

where $\alpha$ is called the discount factor. This is called *infinite time discounted performance evaluation.*

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

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 $\sum_{j\in\Phi} a_{ij}(a) = 0$.

In this study, we assume that the decision point is always the time when state transits, then we can have a similar definition as DTMDP.

$$\eta_\alpha^v(i) = E_v\{\int_0^\infty e^{-\alpha t} f(Y_t, v(Y_t))dt|Y_0 = i\}, i \in \Phi \tag{5}$$

We call it *infinite time discounted performance evaluation.*

$$\eta^v(i) = \lim_{T\to\infty} \frac{1}{T} E_v\{\int 0^T f(Y_t, v(Y_t))dt|Y_0 = i\}, i \in \Phi \tag{6}$$

This is called *infinite time average performance evaluation.*

If the Markov process Y is ergodic, then $\eta^v(i)$ is irrelevant to initial state, so we have

$$\eta^v = \sum_{i\in\Phi} p^v(i) f(i, v(i)) = p^v f^v \tag{7}$$

where $p^v(i), i \in \Phi$ is the steady distribution of Markov process $Y$ with policy $v$. We call $\eta^v$ the performance measure under policy $v$.

One objective of MDP is to find $v^* \in \Pi$ so that all performance is optimal.

# 2 DTMDP

## 2.1 Performance potential

For an ergodic Markov Chain $X = \{X_n; n \geq 0\}$, its state space $\Phi = \{1, 2, ..., K\}$, finite action space $A$, policy space determined steady policy set $\Pi_s^d$, with policy $v \in \Pi_s^d$, its transition matrix $P^v = [P_{ij}(v(i))]$, steady distribution $\pi^v = (\pi^v(1), ..., \pi^v(K))$ satisfies

$$\pi^v P^v = \pi^v, \pi^v e = 1 \tag{8}$$

where $e = (1, 1, ..., 1)^\tau$.

We only consider the infinite time average performance evaluation.

$$\eta^v = \sum_{n=0}^{K} \pi^v f(i, v(i)) = \pi^v f^v \tag{9}$$

where $f^v = (f(1, v(1)), ..., f(K, v(K))^\tau$.

Denote

$$g^v(i) = E\{\sum_{n=0}^{\infty}[f(X_n, v(X_n)) - \eta^v]|X_0 = i\}, i = \Phi \tag{10}$$

*Definition* $g^v = (g^v(1), ..., g^v(K))^\tau$ is called performance potential vector, or potential of Markov Chain $X$ w.r.t performance function $f^v$ at state $i$.
*Lemma* potential $g$ satisfies Poisson equation

$$(I - P^v)g^v = f^v - \eta^v e \tag{11}$$

all potentials can be represented as

$$g^v = (I - P^v + e\pi)^{-1} f^v + ce \tag{12}$$

where $c$ can be any constant. *proof*

$$g(i) = \lim_{N\to\infty} E\{\sum_{n=0}^{N}[f(X_n, v(X_n)) - \eta^v]|X_0 = i\} \tag{13}$$

$$= \lim_{N\to\infty} \sum_{n=0}^{N} E[f(X_n, v(X_n)) - \eta^v|X_0 = i] \tag{14}$$

$$= f(i, v(i)) - \eta^v + \lim_{N\to\infty} \sum_{n=1}^{N} E[f(X_n, v(X_n)) - \eta^v|X_0 = i] \tag{15}$$

$$= f(i, v(i)) - \eta^v + \sum_{j\in\Phi} P_{ij}^v \lim_{N\to\infty} \sum_{n=1}^{N} E[f(X_n, v(X_n)) - \eta^v|X_1 = j] \tag{16}$$

$$= f(i, v(i)) - \eta^v + \sum_{j\in\Phi} P_{ij}^v g^v(j) \tag{17}$$

that is,

$$g^v = f^v - \eta^v e + P^v g^v \tag{18}$$
$$(I - P^v)g^v = f^v - \eta^v e \tag{19}$$

Appearently $rank(I - P^v) = K - 1$, so that solution space of equation $(I - P^v)g^v$ is 1-dimensional and $g^v = e$ is one solution. Then we need a particular solution of the equation that satisfies $\pi g = \eta$. That is

$$g^v = (I - P^v + e\pi)^{-1} f^v \tag{20}$$

□

Theoretically performance potential can be solved from Poisson equation, but when state space is too large, it is not very easy to be solved. So we often use sample orbit to estimate it. Denote

$$g_L(i) = E[\sum_{l=0}^{L-1} f(X_l)|X_0 = i] - L\eta \tag{21}$$

$g(i) = \lim_{L\to\infty} g_L(i)$, because of the property of potential, we can ignore constant term $L\eta$.

$$g_L(i) \approx E[\sum_{l=0}^{L-1} f(X_l)|X_0 = i] \tag{22}$$

denote

$$g_{L,N}(i) = \frac{\sum_{n=0}^{N-L+1} I_i(X_n)[\sum_{l=0}^{L-1} f(X_l)|X_0 = i]}{\sum_{n=0}^{N-L+1} I_i(X_n)} \tag{23}$$

where $I_i(x)$ is characteristic function of state $i$, so we have

$$g_L(i) = \lim_{N\to\infty} g_{L,N}(i), (w.p.1) \tag{24}$$

### 2.2 Performance optimization

**Performance difference** Denote $\eta^u, \pi^u, P^u, g^u$ and $\eta^v, \pi^v, P^v, g^v$ are values under policy $u$ and $v$ respectively, from Poisson equation, we have

$$(I - P^u)g^u = f^u - \eta^u e \tag{25}$$
$$(\pi^v - \pi^v P^u)g^u = \pi^v f^u - \pi^v \eta^u e \tag{26}$$
$$\eta^u = \pi^v f^u - (\pi^v - \pi^v P^u)g^u \tag{27}$$
$$\eta^u = \pi^v f^u - (\pi^v P^v - \pi^v P^u)g^u \tag{28}$$
$$\eta^u = \pi^v f^u - \pi^v (P^v - P^u)g^u \tag{29}$$
$$\tag{30}$$

with $\eta^v = \pi^v f^v$, we have performance difference formula

$$\eta^u - \eta^v = \pi^v f^u - \pi^v (P^v - P^u) g^u - \pi^v f^v \tag{31}$$
$$= \pi^v [(P^u g^u + f^u) - (P^v g^u + f^v)] \tag{32}$$

for $x = \{x_1, ..., x_K\}$ and $y = \{y_1, ..., y_K\}$, $x \prec y$ means for any $i$, we have $x_i \leq y_i$ and there exists at least one $j, 1 \leq j \leq K$ satisfies $x_j < y_j$.

*Lemma*

(1) If

$$P^u g^u + f^u \prec P^v g^u + f^v \tag{33}$$

then $\eta^u < \eta^v$.

(2) $v^*$ is the optimal policy of infinite time average performance if and only if for any $v \in \Pi_s^d$,

$$P^{v^*} g^{v^*} + f^{v^*} \leq P^v g^{v^*} + f^v \tag{34}$$

we call it optimal inequality.

*proof for (2)*

If (34) holds for $\forall v \in \Pi_s^d$, then

$$P^{v^*} g^{v^*} + f^{v^*} \leq P^v g^{v^*} + f^v \tag{35}$$
$$\pi^v \{P^{v^*} g^{v^*} + f^{v^*} - (P^v g^{v^*} + f^v)\} \leq 0 \tag{36}$$
$$\eta^{v^*} - \eta^v \leq 0 \tag{37}$$
$$\tag{38}$$

so that $v^*$ is an optimal policy.

On the other hand, if $v^*$ is a optimal policy, then we can get (34) easily as well. □

*Theorem* $v^*$ is the optimal policy for infinite time average performance if and only if $v^*$ satisfies

$$0 = \min_{v \in \Pi_s^d} \{f^v + (P^v - I) g^{v^*} - \eta^{v^*} e\} \tag{39}$$

*proof* Hint: Combine (34) and (11).

*Algorithm* In policy iteration, we select

$$v_{k+1}(i) = \arg\min_{v(i) \in A(i)} \sum_{j=1}^{K} P_{ij}(v(i)) g^{v_k}(i) + f(i, v(i)) \tag{40}$$

**Performance derivative** If both transition matrix and performance function is related to parameter $\theta \in I$, then we can consider the *parameterizing policy.* Assume $P(\theta), f(\theta)$ are differentiable functions. From Poisson equation (11),

$$-\frac{\partial P}{\partial \theta} g + (I - P) \frac{\partial g}{\partial \theta} = \frac{\partial f}{\partial \theta} - \frac{\partial \eta}{\partial \theta} e \tag{41}$$

notice that $\pi P = \pi, \pi e = 1$, we have

$$-\pi\frac{\partial P}{\partial \theta}g + \pi(I-P)\frac{\partial g}{\partial \theta} = \pi\frac{\partial f}{\partial \theta} - \pi\frac{\partial \eta}{\partial \theta}e \tag{42}$$

$$-\pi\frac{\partial P}{\partial \theta}g = \pi\frac{\partial f}{\partial \theta} - \frac{\partial \eta}{\partial \theta} \tag{43}$$

$$\frac{\partial \eta}{\partial \theta} = \pi\frac{\partial f}{\partial \theta} + \pi\frac{\partial P}{\partial \theta}g \tag{44}$$

$$\tag{45}$$

For $\theta = (\theta_1, ..., \theta_m)$, we have

$$\nabla\eta = \Big(\frac{\partial \eta}{\partial \theta_1}, \frac{\partial \eta}{\partial \theta_2}, ..., \frac{\partial \eta}{\partial \theta_m}\Big) \tag{46}$$

Any optimal policy $\theta^*$ should satisfies

$$\theta^* = \arg_{\theta}\{\nabla\eta = 0\} \tag{47}$$

# 3 CTMDP