MDP:Preliminaries

Peng Lingwei

July 16, 2019

Contents

1 Basic model

1. State set: $S = \{s_1, s_2, \ldots, s_n\}$
2. Action set: $A = \{a_1, a_2, \ldots, a_m\}$
3. State transition matrix:
$$P = \begin{pmatrix} \vec{s_{11}} & \vec{s_{12}} & \cdots & \vec{s_{1m}} \\ \vec{s_{21}} & \vec{s_{22}} & \cdots & \vec{s_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{s_{n1}} & \vec{s_{n2}} & \cdots & \vec{s_{nm}} \end{pmatrix}$$
4. Reward matrix:
$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{pmatrix}$$
5. Decision matrix:
$$d_t = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{pmatrix}$$
6. Policy process: $\pi = \{d_0, d_1, \ldots, d_t, \ldots\}$
7. Matrix product:
$$\left\langle \begin{pmatrix} \vec{a_1} \\ \vec{a_2} \\ \vdots \\ \vec{a_n} \end{pmatrix}, \begin{pmatrix} \vec{b_1} \\ \vec{b_2} \\ \vdots \\ \vec{b_n} \end{pmatrix} \right\rangle = \begin{pmatrix} \langle \vec{a_1}, \vec{b_1} \rangle \\ \langle \vec{a_2}, \vec{b_2} \rangle \\ \vdots \\ \langle \vec{a_n}, \vec{b_n} \rangle \end{pmatrix}$$

8. Process' state transition:

$$\begin{aligned} P^{\pi} =& \{P_0, P_1, \ldots, P_t, \ldots\} \\ =& \{\langle P, d_0\rangle, \langle P, d_1\rangle, \ldots, \langle P, d_t\rangle, \ldots\} \end{aligned}$$

9. Process' reward:

$$\begin{aligned} R^{\pi} =& \{R_0, R_1, \ldots, R_t, \ldots\} \\ =& \{\langle R, d_0\rangle, \langle R, d_1\rangle, \ldots, \langle R, d_t\rangle, \ldots\} \end{aligned}$$

10. Process's value function:

$$\begin{aligned} V^{\pi} =& \mathbb{E}\left\{\sum_{t=0}^{\infty} \alpha^t r(s_t, a_t, s_{t+1})\right\} \\ =& \sum_{t=0}^{\infty} \alpha^t \prod_{k=0}^{t-1} P_k R_t \end{aligned}$$

11. Optimal process' value: $V^* = \sup_\pi V$, corresponding policy π^* is called $\alpha - \mathit{optimal}$ policy.

2 Basic theorem

Theorem 2.1.

$$V^* = \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^*)$$

$$V^{\pi^* = \{d_0^*, d_1^*, \ldots\}} = \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^{\pi^* = \{d_0^*, d_1^*, \ldots\}})$$

Proof. $\forall \pi = \{d_0, d_1, \ldots\}$, we have

$$\begin{aligned} V^{\pi = \{d_0, d_1, \ldots\}} =& R_0 + \alpha P_0 V^{\pi' = \{d_1, d_2, \ldots\}} \\ \preceq& R_0 + \alpha P_0 V^* \\ \preceq& \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^*) \\ V^* \preceq& \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^*) \end{aligned}$$

Let $a_0 = \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^*)$, and $d = \mathbf{1}\{a = a_0\}$, so we construct that: $\pi = \{d, d_1, \ldots\}$, and $\pi' = \{d_1, d_2, \ldots\}$, then:

$$\forall \epsilon, \exists \pi' = \{d_1, d_2, \ldots\}, \quad V^{\pi' = \{d_1, d_2, \ldots\}} \succeq V^* - \epsilon$$

$$\begin{aligned} V^* \succeq V^{\pi} =& R_0^* + \alpha P_0^* V^{\pi'} \\ \succeq& R_0^* + \alpha P_0^* V^* - \alpha \epsilon \end{aligned}$$

because ϵ is arbitrary,so

$$V^* \succeq \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^*)$$

□

Theorem 2.2. *$\exists \pi = \{d, d, \ldots, \}$ is $\alpha -$ optimal policy.*

Proof. Let $d = \mathbf{1}\{a = \max_{a'}(R(\cdot, a') + \alpha P(\cdot, a')V^*)\}$ (This construction maybe problematic.) From preceeding theorem, we can get:

$$
\begin{aligned}
V^{\pi^*=\{d_0^*, d_1^*, \ldots\}} &= \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^{\pi^* = \{d_1^*, d_2^*, \ldots\}}) \\
&= \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^{\pi^* = \{d_0^*, d_1^*, \ldots\}}) \\
&= \langle R, d \rangle + \alpha \langle P, d \rangle V^{\pi^* = \{d_0^*, d_1^*, \ldots\}} \\
&= \sum_{t=0}^{n} \alpha^t \langle P, d \rangle^t \langle R, d \rangle + \alpha^n \langle P, d \rangle^n V^{\pi^* = \{d_0^*, d_1^*, \ldots\}} \\
n \to \infty, \quad V^* &= V^\pi
\end{aligned}
$$

□

Theorem 2.3. *Let $T_{\pi=\{d,d,\ldots\}}V = \langle P, d \rangle + \alpha \langle P, d \rangle V$, d and V are arbitrary, then*

$$n \to \infty, \quad T_\pi^n V = V^\pi$$

Theorem 2.4. *If $T_{\pi_2} V^{\pi_1} = \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^{\pi_1})$, then $V^{\pi_2} \succeq V^{\pi_1}$.*

Proof.

$$
\begin{aligned}
T_{\pi_2} V^{\pi_1} = \max_a (R(\cdot, a) + \alpha P(\cdot, a) V^{\pi_1}) \succeq T_{\pi_1} V^{\pi_1} = V^{\pi_1} \\
n \to \infty, \quad V^{\pi_2} = T_{\pi_2}^n V^{\pi_1} \succeq V^{\pi_1}
\end{aligned}
$$

□

Theorem 2.5. *If $U \succeq \max_a (R(\cdot, a) + \alpha P(\cdot, a) U)$, then $U \succeq V^*$.*

Proof. $U \succeq V_n^* + \alpha^n \mathbb{E}^\pi [U(s_t)|s_0] \to V^*, as\ n \to \infty$ (Using next section's proof, value Improvement). □

Theorem 2.6. *The equation of V has unique solution.*

$$V = \max_a [R(\cdot, a) + \alpha P(\cdot, a) V]$$

Proof. Assuming that the equation has two solution U, V, then

$$
\begin{aligned}
U - V &= \max_a [R(\cdot, a) + \alpha P(\cdot, a) U] - \max_a [R(\cdot, a) + \alpha P(\cdot, a) V] \\
&= [R(\cdot, a_U) + \alpha P(\cdot, a_U) V] - \max_a [R(\cdot, a) + \alpha P(\cdot, a) V] \\
&\preceq \alpha P(\cdot, a_U)[U - V] \\
&\preceq \alpha \sup |U - V| \cdot \vec{e}, \quad (\vec{e} = [1, 1, 1, \ldots, 1]^T)
\end{aligned}
$$

Similarly,

$$V - U \preceq \alpha \sup |U - V| \cdot \vec{e}$$

So, $\sup |U - V| = 0$, $U = V$ □

Theorem 2.7. *The equation of V about $\pi = \{d, d, \ldots\}$ has unique solution, and the solution is V^π.*

$$V = \langle R, d \rangle + \alpha \langle P, d \rangle V$$

Theorem 2.8.

3 Value Improvement Method

Definition 3.1. ***Policy Improvement Method***

1. *Step 1: Arbitrary state value:* V_0*;*
2. *Step 2:* $V_n = \max_a[R(\cdot, a) + \alpha P(\cdot, a)V_{n-1}]$.

Proof. Here.
If $V_0 = \vec{0}$, then

$$\begin{aligned} V_n = V_n^* &= \max_{\pi} \mathbb{E}_n^{\pi=\{d_0,d_1,\ldots,d_{n-1}\}}\{\sum_{t=0}^{n-1} \alpha^t r(s_t, a)|s_0\} \\ &= \max_{\pi} \sum_{t=1}^{n} \alpha^t \prod_{k=0}^{t-1} \langle P, d_k\rangle \langle R, d_t\rangle \end{aligned}$$

It can be proofed by induction, but here is intuitive description:

$$V_3 = \langle R, d_0\rangle + \alpha\langle P, d_0\rangle(\langle R, d_1\rangle + \alpha\langle P, d_1\rangle\langle R, d_2\rangle)$$

The process can be saw as that: V_1 gets optimal decision d_2^*, then V_2 gets optimal decision d_1^*, and V_3 gets optimal decision d_0^*, therefore, $V_3 = V_3^*$.

If $|r(s_t, a_t)| \leq B$, then

$$|V^* - V_n| \preceq \left|\mathbb{E}^{\pi}[\sum_{t=n+1}^{\infty} \alpha^t r(s_t, a_t)|s_0]\right| \preceq \alpha^{n+1}B/(1-\alpha)$$

If $V_0 \neq \vec{0}$, then we let V_n^0 denote V_n when $V_0 = 0$, then

$$V_n = V_n^0 + \alpha^n \prod_{k=0}^{n-1} \langle P, d_k\rangle V_0$$

$$|V_n - V_n^0| = \left|\alpha^n \prod_{k=0}^{n-1} \langle P, d_k\rangle V_0\right| \preceq \alpha^n \sup |V_0|\vec{e}$$

Then

$$n \to \infty, \quad V_n^0 \to V^*, V_n \to V_n^0$$

$\square$