Markov Decision Processes: Discrete Stochastic Dynamic Programming

Peng Lingwei

July 24, 2019

Contents

4 Chapter4: Finite-Horizon Markov Decision Processes

4.1 OPTIMALITY CRITERIA

4.1.1 Some Preliminaries

About MDP:

1. $\pi = (d_1, d_2, \ldots, d_{N-1}, \ldots) \in \Pi^{HR}$;
2. $h_N = (s_1, a_1, s_2, \ldots, s_N)$
3. Rewards sequence: $\{r_1(s_1, a_1), r_2(s_2, a_2), \ldots, r_{N-1}(s_{N-1}, a_{N-1}), r_N(s_N)\}$
   - $\pi \in \Pi^{HD}$, $\{r_1(X_1, d_1(h_1)), \ldots, r_{N-1}(X_{N-1}, d_{N-1}(h_{N-1})), r_N(X_N)\}$
   - $\pi \in \Pi^{MD}$, $\{r_1(X_1, d_1(X_1)), \ldots, r_{N-1}(X_{N-1}, d_{N-1}(X_{N-1})), r_N(X_N)\}$
4. $R = (R_1, R_2, \ldots, R_N)$, where $R_t = r_t(X_t, Y_t)$, and $|R_t| \leq M < \infty$.
5. $\mathbb{P}^\pi_R(\rho_1, \rho_2, \ldots, \rho_N) = \mathbb{P}^\pi\left[\{(s_1, a_1, \ldots, s_N) : (r(s_1, a_1), \ldots, r_N(s_N)) = (\rho_1, \ldots, \rho_N)\}\right]$

Definition:

1. The random vairable U is stochastically greater than V:
$$\forall t \in \mathbb{R}, \quad P(V > t) \leq P(U > t).$$
2. Probability distribution P_2 is stochastically greater than P_1 if:
$$\forall t \in \mathbb{R}, \quad \int_t^\infty p_1(t)dt \leq \int_t^\infty p_2(t)dt.$$
3. The random vector $\vec{U} = (U_1, \ldots, U_n)$ is stochastically greater than the random vector $\vec{V} = (V_1, \ldots, V_n)$:
$$\forall f \in \{f : \mathbb{R}^n \to \mathbb{R} | \vec{v} \preceq \vec{u} \Rightarrow f(\vec{v}) \leq f(\vec{u})\}, \quad \mathbb{E}[f(\vec{V})] \leq \mathbb{E}[f(\vec{U})]$$

4.1.2 The Expected Total Reward Criterion

The expected total reward criterion:

1. $\pi \in \Pi^{HR}$: $v^\pi_N(s) = \mathbb{E}^\pi_S \left\{ \sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right\}$.
2. $\pi \in \Pi^{HD}$: $v^\pi_N(s) = \mathbb{E}^\pi_S \left\{ \sum_{t=1}^{N-1} r_t(X_t, d_t(h_t)) + r_N(X_N) \right\}$.
3. Discounted reward:$\pi \in \Pi^{HR}$,
   $v^\pi_{N,\lambda}(s) = \mathbb{E}^\pi_S \left\{ \sum_{t=1}^{N-1} \lambda^{t-1} r_t(X_t, d_t(h_t)) + \lambda^{N-1} r_N(X_N) \right\}$.
   Taking the discount factor into account does not effect any theoretical results or algorithms in the finite-horizon case but might effect the decision maker's preference for policies.

4.1.3 Optimal Policies

Definition:

1. Optimal policy $\pi^* : \forall \pi \in \Pi^{HR}, v_N^{\pi^*} \succeq v_N^{\pi}$.
2. $\epsilon-$optimal policy, $\pi^* : \forall \pi \in \Pi^{HR}, v_N^{\pi_\epsilon^*} + \epsilon \succeq v_N^{\pi}$.
3. Optimal value: $v_N^* = \sup_{\pi \in \Pi^{HR}} v_N^{\pi}$.
4. We can get $v_N^{\pi^*} = v_N^*$ and $v_N^{\pi_\epsilon^*} + \epsilon > v_N^*$.
5. Considering initial state distribution P_1: $v_N^{\pi, P_1} = \sum_{s \in S} v_N^{\pi}(s) P_1\{X_1 = s\}$.

Markov decision problem = Markov decision process + Optimality criteria

4.2 FINITE-HORIZON POLICY EVALUATION

1. $\pi = (d_1, d_2, \ldots, d_{N-1}) \in \Pi^{HR}$
2. Define: $u_t^{\pi}(h_t) = \mathbb{E}_{h_t}^{\pi} \left\{ \sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\}$, $(u_t^{\pi} : H_t \to \mathbb{R})$.
   And we define $U_N^{\pi}(h_N) = r_N(s_N)$.
3. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{HD}$):

$$\begin{aligned} \hat{u}_t^{\pi}(h_t) =& r_t(s_t, d_t(h_t)) + \sum_{s' \in S} p_t(s'|s_t, d_t(h_t)) \hat{u}_{t+1}^{\pi}(h_t, d_t(h_t), s'). \quad ((h_t, d_t(h_t), s') \in H_{t+1}) \\ =& r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^{\pi} \left\{ \hat{u}_{t+1}^{\pi}(h_t, d_t(h_t), X_{t+1}) \right\} \end{aligned}$$

   Proof. Part proof with backward induction hypothesis ($u_{h_{t+1}}^{\pi} = \hat{u}_{h_{t+1}}^{\pi}$):

$$\begin{aligned} \hat{u}_t^{\pi}(h_t) =& r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^{\pi} \left\{ u_{t+1}^{\pi}(h_t, d_t(h_t), X_{t+1}) \right\} \\ =& r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^{\pi} \left\{ \mathbb{E}_{h_{t+1}}^{\pi} \left\{ \sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\} \right\} \\ =& r_t(s_t, d_t(h_t)) + \mathbb{E}_{h_t}^{\pi} \left\{ \sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\} \\ =& \mathbb{E}_{h_t}^{\pi} \left\{ r_t(s_t, d_t(h_t)) + \sum_{n=t+1}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\} = u_t^{\pi}(h_t) \end{aligned}$$

   □

4. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{HR}$):

$$\hat{u}_t^{\pi}(h_t) = \sum_{a \in A_{s_t}} q_{d_t(h_t)}(a) \left\{ r_t(s_t, d_t(h_t)) + \sum_{s' \in S} p_t(s'|s_t, d_t(h_t)) \hat{u}_{t+1}^{\pi}(h_t, d_t(h_t), s'). \right\}$$

5. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{MR}$):

$$\hat{u}_t^{\pi}(s_t) = r_t(s_t, d_t(s_t)) + \sum_{s' \in S} p_t(s'|s_t, d_t(s_t)) \hat{u}_{t+1}^{\pi}(s').$$

6. The computation complexity. There are K states and L actions, then:
   - If $\pi \in \Pi^{HD}$, then requiring $K^2 \sum_{i=0}^{N-1} (KL)^i$ multiplications.
   - If $\pi \in \Pi^{MD}$, then requiring $(N-1)K^2$ multiplications.

4.3 OPTIMALITY EQUATIONS AND THE PRINCIPLE OF OPTIMALITY

Optimality equations (Bellman equations or functional equations).
We start study this equation:

$$u_t^*(h_t) = \sup_{\pi \in \Pi^{HR}} u_t^\pi(h_t)$$

When minimizing costs instead of maximizing rewards, we sometimes refer to u_t^* as a **cost-to-go** funciton.

Definition 1. ***(Optimality equations).***

$$\hat{u}_t(h_t) = \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a)\hat{u}_{t+1}(h_t, a, s') \right\}, \quad s.t.\ \hat{u}_N(h_N) = r_N(s_N). \tag{1}$$

If A_{s_t} is finite, it can be replaced by max*. Then, $\forall h_t, \hat{u}_t(h_t) = u_t^*(h_t)$.*

Proof. The proof is in two parts.
Let arbitrary $\pi' = (d_1', d_2', \ldots, d_{N-1}') \in \Pi^{HR}$.
Step1:
First, we have $u_N^{\pi'}(h_N) = \hat{u}_N(h_N) = u_N^*(h_N)$.
Then, because we take the operation sup, we reasonably have $\hat{u}_{N-1}(h_{N-1}) \geq u_{N-1}^*(h_{N-1})$.
Assuming that $\forall h_t \in H_t$, and $t = n+1, \ldots, N$, we have $\hat{u}_t(h_t) \geq u_t^*(h_t)$.

$$\begin{aligned}
\hat{u}_n(h_n) &= \sup_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a)\hat{u}_{n+1}(h_n, a, s') \right\} \\
&\geq \sup_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a)u_{n+1}^*(s_n, a, s') \right\} \\
&\geq \sum_{a \in A_{s_n}} q_{d_n'(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a)u_{n+1}^{\pi'}(s_n, a, s') \right\} \\
&\geq u_n^{\pi'}(h_n)
\end{aligned}$$

Which means that, $\forall \pi \in \Pi^{HR}, \hat{u}_n(h_n) \geq u_n^\pi(h_n)$.
Step2:
$\forall \epsilon$, we can construct $\pi' \in \Pi^{HR}$ for which: $u_n^{\pi'}(h_n) + (N-n)\epsilon \geq \hat{u}_n(h_n)$.
To do this, construct a policy $\pi' = (d_1', d_2', \ldots, d_{N-1}') \in \Pi^{HR}$ by choosing $d_n(h_n)$ to satisfy

$$\sum_{a \in A_{S_t}} q_{d_n'(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a)\hat{u}_{n+1}(s_n, a, s') \right\} + \epsilon \geq \hat{u}_n(h_n).$$

First, we have $u_N^{\pi'}(h_N) = u_N(h_N)$.
Then, we assume that $u_t^{\pi'}(h_t) + (N-t)\epsilon \geq u_t(h_t)$ for $t = n+1, \dots, N$.

$$\begin{aligned}
u_n^{\pi'}(h_n) =& \sum_a q_{d'_n(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) u_{n+1}^{\pi'}(s_n, a, s') \right\} \\
\geq& \sum_a q_{d'_n(h_n)}(a) \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) \hat{u}_{n+1}(s_n, a, s') \right\} - (N-n-1)\epsilon \\
\geq& \hat{u}_n(h_n) - (N-n)\epsilon
\end{aligned}$$

Step3: $u_n^*(h_n) + (N-n)\epsilon \geq u_n^{\pi'}(h_n) + (N-n)\epsilon \geq u_n(h_n) \geq u_n^*(h_n)$. □

Theorem 1. *Suppose $u_t^*, t = 1, \dots, N$ are solutions of the optimality equation (max version). Then we can construct a corresponding policy $\pi^* = (d_1^*, d_2^*, \dots, d_{N-1}^*) \in \Pi^{HD}$ satisfies*

$$d_t^*(h_t) \in \arg\max_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) u_{t+1}^*(h_t, a, s') \right\}$$

for $t = 1, \dots, N-1$. Then

1. $u_t^{\pi^*}(h_t) = u_t^*(h_t), \quad h_t \in H_t$.
2. $v_N^{\pi^*}(s) = v_N^*(s), \quad s \in S$.

Proof. Clearly, $u_N^{\pi^*}(h_n) = u_N^*(h_n), h_n \in H_n$.
We assume that $u_{n+1}^{\pi^*}(h_{n+1}) = u_{n+1}^*(h_{n+1})$,

$$\begin{aligned}
u_n^*(h_n) =& \max_{a \in A_{s_n}} \left\{ r_n(s_n, a) + \sum_{s' \in S} p_n(s'|s_n, a) u_{n+1}^*(h_n, a, s') \right\} \\
=& r_n(s_n, d_n^*(h_n)) + \sum_{s' \in S} p_n(s'|s_n, d_n^*(h_n)) u_{n+1}^*(h_n, d_n^*(h_n), s') \\
=& u_n^{\pi^*}(h_n)
\end{aligned}$$

□

Theorem 2. *Let $\epsilon > 0$ be arbitrary and suppose $u_t^*, t = 1, \dots, N$ are solutions of the optimality equation (sup version). Then we can construct a corresponding policy $\pi^\epsilon = (d_1^\epsilon, d_2^\epsilon, \dots, d_{N-1}^\epsilon) \in \Pi^{HD}$ satisfies*

$$\begin{aligned}
&\left\{ r_t(s_t, d_t^\epsilon) + \sum_{s' \in S} p_t(s'|s_t, d_t^\epsilon) u_{t+1}^*(h_t, d_t^\epsilon, s') \right\} + \frac{\epsilon}{N-1} \\
&\geq \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) u_{t+1}^*(h_t, a, s') \right\}
\end{aligned}$$

for $t = 1, \dots, N-1$. Then

1. $u_t^{\pi^\epsilon}(h_t) + (N-t)\frac{\epsilon}{N-1} \geq u_t^*(h_t), \quad h_t \in H_t$.
2. $v_N^{\pi^\epsilon}(s) + \epsilon = v_N^*(s), \quad s \in S$.

The proof is analogous.