Finite-Horizon Markov Decision Processes

Dan Zhang
Leeds School of Business
University of Colorado at Boulder

Outline

- Expected total reward criterion
- Optimality equations and the principle of optimality
- Optimality of deterministic Markov policies
- Backward induction
- Applications

Expected Total Reward Criterion

- Let π be a randomized history-dependent policy; i.e., $\pi \in \Pi^{HR}$.
  - $\pi = (d_1, \ldots, d_{N-1})$ where $d_t : H_t \to \mathcal{P}(A)$.
- Starting at a state s, using policy π leads to a sequence of state-action pairs $\{X_t, Y_t\}$. The sequence of rewards is given by $\{R_t \equiv r_t(X_t, Y_t) : t = 1, \ldots, N-1\}$ with terminal reward $R_N \equiv r_N(X_N)$.
- The expected total rewards from policy π starting in state s is given by

$$v_N^\pi(s) \equiv \mathrm{E}_s^\pi \left[\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(x_N) \right].$$

Optimal Policy

- A policy π^* is an optimal policy if

$$v_N^{\pi^*}(s) \geq v_N^{\pi}(s), \quad \forall s \in S, \pi \in \Pi^{HR}.$$

- The value of a Markov decision problem is defined by

$$v_N^*(s) \equiv \sup_{\pi \in \Pi^{HR}} v_N^{\pi}(s), \quad \forall s \in S.$$

- We have $v_N^{\pi^*}(s) = v_N^*(s)$ for all $s \in S$.

Finite-Horizon Policy Evaluation

- Let $\pi \in \Pi^{HR}$ be a randomized history-dependent policy.
- Let $u_t^\pi : H_t \to R$ be the total expected reward obtained by using policy π at decision epochs $t, t+1, \ldots, N-1$.
- Given $h_t \in H_t$ for $t < N$, let

$$u_t^\pi(h_t) = \mathrm{E}_{h_t}^\pi \left[\sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right].$$

- Furthermore, let $u_N^\pi(h_N) = r_N(s)$ for $h_N = (h_{N-1}, a_{N-1}, s)$.
- For given initial state s, we have $u_1^\pi(s) = v_N^\pi(s)$.

The Finite-Horizon Policy Evaluation Algorithm

Assume $\pi \in \Pi^{HD}$.

1. Set $t = N$ and $u_N^{\pi}(h_N) = r_N(s_N)$ for all $h_N = (h_{N-1}, a_{N-1}, s_N) \in H_N$.
2. If $t = 1$, stop; otherwise go to step 3.
3. Substitute $t-1$ for t and compute $u_t^{\pi}(h_t)$ for each $h_t = (h_{t-1}, a_{t-1}, s_t) \in H_t$ by

$$u_t^{\pi}(h_t) = r_t(s_t, d_t(h_t)) + \sum_{j \in S} p_t(j|s_t, d_t(h_t)) u_{t+1}^{\pi}(h_t, d_t(h_t), j).$$

4. Return to 2.

The Principle of Optimality

- Let $u_t^*(h_t) = \sup_{\pi \in \Pi^{HR}} u_t^\pi(h_t)$.
- Consider the following optimality equations:

$$u_t(h_t) = \sup_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}(h_t, a, j) \right],$$

$$\forall t = 1, \dots, N-1, h_t = (h_{t-1}, a_{t-1}, s_t) \in H_t,$$

$$u_N(h_N) = r_N(s_N), \qquad \forall h_N = (h_{N-1}, a_{N-1}, s_N) \in H_N.$$

Theorem

Suppose u_t is a solution to the optimality equations for all t. Then

(a) $u_t(h_t) = u_t^*(h_t)$ for all $h_t \in H_t$, $t = 1, \ldots, N$;

(b) $u_1(s_1) = v_N^*(s_1)$ for all $s_1 \in S$.

Optimality of Deterministic Markov Policies

Theorem

Let u_t^* be a solution to the optimality equations for all t. Then

(a) For each $t = 1, \dots, N$, $u_t^*(h_t)$ depends on h_t only through s_t;

(b) If there exists an $a' \in A_{s_t}$ such that

$$r_t(s_t, a') + \sum_{j \in S} p_t(j|s_t, a') u_{t+1}^*(h_t, a', j)$$
$$= \sup_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(h_t, a, j) \right]$$

for each $s_t \in S$ and $t = 1, \dots, N-1$, there exists an optimal policy which is deterministic and Markovian.

Backward Induction

1. Set $t = N$ and $u_N^*(s_N) = r_N(s_N)$ for all $s_N \in S$.
2. Substitute $t-1$ for t and compute $u_t^*(s_t)$ for each $s_t \in S$ by

$$u_t^*(s_t) = \max_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(j) \right].$$

   Set

$$A_{s_t,t}^* = \arg\max_{a \in A_{s_t}} \left[r_t(s_t, a) + \sum_{j \in S} p_t(j|s_t, a) u_{t+1}^*(j) \right].$$

3. If $t = 1$, stop; otherwise go to step 2.

e-Rite-Way: An MDP Formulation

- Decision epochs: $T = \{1, 2, 3, 4, 5\}$
- States: $S = \{1, 2\}$.
- Actions: $A_s = \{0, 1, 2\}$.
  - 0: Do nothing
  - 1: Gift and minor price promotion
  - 2: Gift and Major price promotion
- Expected rewards: $r_t(s, a)$ (see handout).
- Terminal rewards: $r_N(s) = 0$.
- Transition probabilities:

$$p_t(i|s, a) = \alpha p_{si}^a, \quad \forall i = 1, 2.$$