Infinite-Horizon Average Reward Markov Decision Processes

Dan Zhang
Leeds School of Business
University of Colorado at Boulder

Outline

- The average reward
- Classification of MDPs
- Optimality equations
- Value iteration in unichain models
- Policy iteration in unichain models
- Linear Programming in unichain models

Average Reward Criterion

- Let $\pi = (d_1, d_2, \dots) \in \Pi^{HR}$
- 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 $\{R_t \equiv r_t(X_t, Y_t) : t = 1, 2, \dots\}$.
- The average reward (or gain) from policy $\pi \in \Pi^{HR}$ starting in state s is given by

$$g^{\pi}(s) \equiv \lim_{N\to\infty} \frac{1}{N} \mathrm{E}_s^{\pi} \left[\sum_{t=1}^{N} r(X_t, Y_t) \right].$$

The limit above may not exist, in which case we define

$$g_{-}^{\pi}(s) \equiv \liminf_{N\to\infty} \frac{1}{N} \mathrm{E}_s^{\pi} \left[\sum_{t=1}^{N} r(X_t, Y_t) \right],$$

$$g_{+}^{\pi}(s) \equiv \limsup_{N\to\infty} \frac{1}{N} \mathrm{E}_s^{\pi} \left[\sum_{t=1}^{N} r(X_t, Y_t) \right].$$

Optimality Criteria

- When $g^{\pi}(s)$ exists for all $s \in S$ and $\pi \in \Pi^{HR}$, a policy π^* is average optimal if

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

- The value (or optimal gain) is defined by

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

- Let π^* be an average optimal policy, then $g^{\pi^*}(s) = g^*(s)$ for all $s \in S$.

Theorem

Suppose $\pi \in \Pi^{HR}$. For each $s \in S$, there exists a $\pi' \in \Pi^{MR}$ (which possibly varies with s) for which

$$g_+^{\pi'} = g_+^{\pi},$$
$$g_-^{\pi'} = g_-^{\pi},$$
$$g^{\pi'} = g^{\pi} \text{ whenever } g_+^{\pi'} = g_-^{\pi'}, g_+^{\pi} = g_-^{\pi}.$$

Assumptions

- Stationary rewards and transition probabilities: $r(s,a)$ and $p(j|s,a)$ do not vary with time
- Bounded rewards: $|r(s,a)| \leq M < \infty$
- Finite state spaces
- Unichain: the transition matrix corresponding to every deterministic stationary policy is unichain (i.e., it consists of a single recurrent class plus a possibly empty set of transient states).

The Average Reward Optimality Equation – Unichain Models

- For unichain models, it can be shown that all stationary policies have constant gain g.
- Optimality equations:

$$0 = \max_{a \in A_s} \left[r(s,a) - g + \sum_{j \in S} p(j|s,a) h(j) - h(s) \right].$$

In matrix notation:

$$0 = \max_{d \in D} \{ r_d - ge + (P_d - I)h \} \equiv B(g,h).$$

The Average Reward Optimality Equation – Unichain Models

Theorem

Suppose S is countable.

(i) If there exists a scalar g and an $h \in V$ which satisfy $B(g,h) \leq 0$, then $ge \geq g_+^*$;

(ii) If there exists a scalar g and an $h \in V$ which satisfy $B(g,h) \geq 0$, then $ge \leq \sup_{d \in D^{MD}} g_-^{d^\infty} \leq g_-^*$;

(iii) If there exists a scalar g and an $h \in V$ which satisfy $B(g,h) = 0$, then $ge = g^* = g_+^* = g_-^*$.

Existence of Solutions to the Optimality Equation – Unichain Models

Theorem

Suppose S and A_s are finite, $|r(s,a)| \leq M < \infty$ for all s, a, and the model is unichain.

(i) There exists a $g \in R^1$ and $h \in V$ for which

$$0 = \max_{d \in D}\{r_d - ge + (P_d - I)h\};$$

(ii) If (g', h') is any other solution of the average reward optimality equation, then $g = g'$.

Existence of Optimal Policies – Unichain Models

- A decision d_h is h-improving if $d_h \in \text{argmax}_{d \in D}\{r_d + P_d h\}$.

Theorem

Suppose there exists a scalar g^* and an $h^* \in V$ for which $B(g^*, h^*) = 0$. Then if d^* is h^*-improving, $(d^*)^\infty$ is average optimal.

Existence of Optimal Policies – Unichain Models

Theorem

Suppose S and A_s are finite, $r(s,a)$ is bounded, and the model is unichain. Then

(i) there exists a stationary average optimal policy;

(ii) there exists a scalar g^* and an $h^* \in V$ for which $B(g^*, h^*) = 0$;

(iii) any stationary policy derived from an h^*-improving decision rule is average optimal;

(iv) $g^* e = g^*_+ = g^*_-$.

Value Iteration

1. Select $v^0 \in V$, specify $\epsilon > 0$, and set $n = 0$.
2. For each $s \in S$, compute $v^{n+1}(s)$ by

$$v^{n+1}(s) = \max_{a \in A_s} \left[r(s,a) + \sum_{j \in S} p(j|s,a) v^n(j) \right].$$

3. If $sp(v^{n+1} - v^n) < \epsilon$, go to step 4. Otherwise, increment n by 1 and return to step 2.
4. For each $s \in S$, choose

$$d_\epsilon(s) \in \operatorname*{argmax}_{a \in A_s} \left[r(s,a) + \sum_{j \in S} p(j|s,a) v^{n+1}(j) \right]$$

and stop.

Relative Value Iteration

1. Select $u^0 \in V$, choose $s^* \in S$, specify $\epsilon > 0$, set $w^0 = u^0 - u^0(s^*)e$, and set $n = 0$.
2. For each $s \in S$, compute $u^{n+1}(s)$ by

$$u^{n+1}(s) = \max_{a \in A_s} \left[r(s,a) + \sum_{j \in S} p(j|s,a) w^n(j) \right].$$

   Let $w^{n+1} = u^{n+1} - u^{n+1}(s^*)e$.
3. If $sp(u^{n+1} - u^n) < \epsilon$, go to step 4. Otherwise, increment n by 1 and return to step 2.
4. For each $s \in S$, choose

$$d_\epsilon(s) \in \operatorname*{argmax}_{a \in A_s} \left[r(s,a) + \sum_{j \in S} p(j|s,a) u^n(j) \right]$$

   and stop.

Policy Iteration

1. Set $n = 0$ and select an arbitrary decision rule $d_0 \in D$.
2. (Policy evaluation) Obtain a scalar g_n and an $h_n \in V$ by solving
$$0 = r_{d_n} - ge + (P_{d_n} - I)h.$$
3. (Policy improvement) Choose d_{n+1} satisfy
$$d_{n+1} \in \operatorname*{argmax}_{d \in D}[r_d + P_d h_n].$$
Setting $d_{n+1} = d_n$ if possible.
4. If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise increment n by 1 and return to step 2.

Policy Iteration

1. Set $n = 0$ and select an arbitrary decision rule $d_0 \in D$.
2. (Policy evaluation) Obtain a scalar g_n and an $h_n \in V$ by solving
$$0 = r_{d_n} - ge + (P_{d_n} - I)h.$$
3. (Policy improvement) Choose d_{n+1} satisfy
$$d_{n+1} \in \operatorname*{argmax}_{d \in D}[r_d + P_d h_n].$$
Setting $d_{n+1} = d_n$ if possible.
4. If $d_{n+1} = d_n$, stop and set $d^* = d_n$. Otherwise increment n by 1 and return to step 2.

Practical consideration: set $h_n(s_0) = 0$ for some fixed $s_0 \in S$.

Linear Programming

- Primal linear program is given by

$$\min_{g,h} g$$

$$g + h(s) - \sum_{j \in S} p(j|s,a)h(j) \geq r(s,a), \qquad \forall s \in S, a \in A_s.$$

- Dual linear program is given by

$$\max_{x} \sum_{s \in S} \sum_{a \in A_s} r(s,a)x(s,a)$$

$$\sum_{a \in A_j} x(j,a) - \sum_{s \in S} \sum_{a \in A_s} \lambda p(j|s,a)x(s,a) = 0, \quad \forall j \in S,$$

$$\sum_{s \in S} \sum_{a \in A_s} x(s,a) = 1,$$

$$x(s,a) \geq 0, \qquad \forall s \in S, a \in A_s.$$