Infinite-Horizon Discounted Markov Decision Processes

Dan Zhang
Leeds School of Business
University of Colorado at Boulder

Outline

- The expected total discounted reward
- Policy evaluation
- Optimality equations
- Value iteration
- Policy iteration
- Linear Programming

Expected Total 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 $\{R_t \equiv r_t(X_t, Y_t) : t = 1, 2, \dots\}$.
- Let $\lambda \in [0, 1)$ be the discount factor
- The expected total rewards from policy π starting in state s is given by

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

The limit above exists when $r(\cdot)$ is bounded; i.e., $\sup_{s\in S, a\in A_s} |r(s, a)| = M < \infty$.

- Under suitable conditions (such as the boundedness of $r(\cdot)$), we have

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

- Let

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

We have $v^\pi(s) = \lim_{\lambda\uparrow 1} v_\lambda^\pi(s)$ whenever $v^\pi(s)$ exists.

Optimality Criteria

- A policy π is discount optimal for $\lambda \in [0,1)$ if

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

- The value of a discounted MDP is defined by

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

- Let π^* be a discount optimal policy. Then $v_\lambda^{\pi^*}(s) = v_\lambda^*(s)$ for all $s \in S$.

Vector Notation for Markov Decision Processes

- Let V denote the set of bounded real valued functions on S with componentwise partial order and norm $||v|| \equiv \sup_{s \in S} |v(s)|$.
- The corresponding matrix norm is given by

$$||H|| \equiv \sup_{s \in S} \sum_{j \in S} |H(j|s)|,$$

where $H(j|s)$ denotes the (s, j)-th component of H.

- Let $e \in V$ denote the function with all components equal to 1; that is, $e(s) = 1$ for all $s \in S$.

Vector Notation for Markov Decision Processes

- For $d \in D^{MD}$, let

$$r_d(s) \equiv r(s, d(s)) \text{ and } p_d(j|s) \equiv p(j|s, d(s)).$$

  Similarly, for $d \in D^{MR}$, let

$$r_d(s) \equiv \sum_{a \in A_s} q_{d(s)}(a) r(s, a), \quad p_d(j|s) \equiv \sum_{a \in A_s} q_{d(s)}(a) p(j|s, a).$$

- Let r_d denote the $|S|$-vector, with the s-th component $r_d(s)$ and P_d the $|S| \times |S|$ matrix with (s, j)-th entry $p_d(j|s)$. We refer to r_d as the *reward vector* and P_d as the *transition probability matrix*.

Vector Notation for Markov Decision Processes

- $\pi = (d_1, d_2, \dots) \in \Pi^{MR}$. The (s, j) component of the t-step transition probability matrix $P_\pi^t(j|s)$ satisfies

$$P_\pi^t(j|s) = [P_{d_1} \dots P_{d_{t-1}} P_{d_t}](j|s) = P^\pi(X_{t+1} = j | X_1 = s).$$

- For $v \in V$,

$$E_s^\pi[v(X_t)] = \sum_{j \in S} P_\pi^{t-1}(j|s) v(j).$$

- We also have

$$v_\lambda^\pi = \sum_{t=1}^\infty \lambda^{t-1} P_\pi^{t-1} r_{d_t}.$$

Assumptions

- Stationary rewards and transition probabilities: $r(s, a)$ and $p(j|s, a)$ do not vary with time
- Bounded rewards: $|r(s, a)| \leq M < \infty$
- Discounting: $\lambda \in [0, 1)$.
- Discrete state space: S is finite or countable

Policy Evaluation

Theorem

Let $\pi = (d_1, d_2, \ldots) \in \Pi^{HR}$. Then for each $s \in S$, there exists a policy $\pi' = (d'_1, d'_2, \ldots) \in \Pi^{MR}$, satisfying

$$P^{\pi'}(X_t = j, Y_t = a | X_1 = s) = P^{\pi}(X_t = j, Y_t = a | X_1 = s), \quad \forall t.$$

$\Longrightarrow$ Suppose $\pi \in \Pi^{HR}$, then for each $s \in S$, there exists a policy $\pi' \in \Pi^{MR}$ such that $v_\lambda^{\pi'}(s) = v_\lambda^{\pi}(s)$.

$\Longrightarrow$ It suffices to consider Π^{MR}.

$$v_\lambda^*(s) = \sup_{\pi \in \Pi^{HR}} v_\lambda^\pi(s) = \sup_{\pi \in \Pi^{MR}} v_\lambda^\pi(s).$$

Policy Evaluation

- Let $\pi = (d_1, d_2, \dots) \in \Pi^{MR}$. Then

$$v_\lambda^\pi(s) = E_s^\pi\left[\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t)\right].$$

In vector notation, we have

$$\begin{aligned} v_\lambda^\pi &= \sum_{t=1}^{\infty} \lambda^{t-1} P_\pi^{t-1} r_{d_t} \\ &= r_{d_1} + \lambda P_\pi^1 r_{d_2} + \lambda^2 P_\pi^2 r_{d_3} + \dots \\ &= r_{d_1} + \lambda P_{d_1} r_{d_2} + \lambda^2 P_{d_1} P_{d_2} r_{d_3} + \dots \\ &= r_{d_1} + \lambda P_{d_1} (r_{d_2} + \lambda P_{d_2} r_{d_3} + \dots) \\ &= r_{d_1} + \lambda P_{d_1} v_\lambda^{\pi'}, \end{aligned}$$

where $\pi' = (d_2, d_3, \dots)$.

Policy Evaluation

- When π is stationary, $\pi = (d, d, \dots) \equiv d^\infty$ and $\pi' = \pi$.
- It follows that $v_\lambda^{d^\infty}$ satisfies

$$v_\lambda^{d^\infty} = r_{d_1} + \lambda P_d v_\lambda^{d^\infty} \equiv L_d v_\lambda^{d^\infty},$$

where $L_d : V \to V$ is a linear transformation.

Theorem

Suppose $\lambda \in [0, 1)$. Then for any stationary policy d^∞ with $d \in D^{MR}$, $v_\lambda^{d^\infty}$ is a solution in V of

$$v = r_d + \lambda P_d v.$$

Furthermore, $v_\lambda^{d^\infty}$ may be written as

$$v_\lambda^{d^\infty} = (I - \lambda P_d)^{-1} r_d.$$

Optimality Equations

- For any fixed n, the finite horizon optimality equation is given by

$$v_n(s) = \sup_{a \in A_s} \left[r(s,a) + \sum_{j \in S} \lambda P(j|s,a) v_{n+1}(j) \right].$$

Taking limits on both sides leads to

$$v(s) = \sup_{a \in A_s} \left[r(s,a) + \sum_{j \in S} \lambda P(j|s,a) v(j) \right].$$

The equations above for all $s \in S$ are *the optimality equations*.

- For $v \in V$, let

$$\mathcal{L}v \equiv \sup_{d \in D^{MD}} [r_d + \lambda P_d v],$$
$$Lv \equiv \max_{d \in D^{MD}} [r_d + \lambda P_d v].$$

Optimality Equations

Proposition

For all $v \in V$ and $\lambda \in [0, 1)$,

$$\sup_{d \in D^{MD}} [r_d + \lambda P_d v] = \sup_{d \in D^{MR}} [r_d + \lambda P_d v].$$

- Replacing D^{MD} with D, the optimality equation can be written as

$$v = \mathcal{L} v.$$

In case supremum can be attained above for all $v \in V$,

$$v = Lv.$$

Solutions of the Optimality Equations

Theorem

Suppose $v \in V$.

(i) If $v \geq \mathcal{L}v$, then $v \geq v_\lambda^*$;

(ii) If $v \leq \mathcal{L}v$, then $v \leq v_\lambda^*$;

(iii) If $v = \mathcal{L}v$, then v is the only element of V with this property and $v = v_\lambda^*$.

Solutions of the Optimality Equations

- Let U be a Banach space (complete normed linear space).
  - Special case: space of bounded measurable real-valued functions
- An operator $T : U \to U$ is a contraction mapping if there exists a $\lambda \in [0, 1)$ such that $||Tv - Tu|| \le \lambda ||v - u||$ for all u and v in U.

Theorem [Banach Fixed-Point Theorem]

Suppose U is a Banach space and $T : U \to U$ is a contraction mapping. Then

(i) There exists a unique v^* in U such that $Tv^* = v^*$;

(ii) For arbitrary $v^0 \in U$, the sequence $\{v^n\}$ defined by $v^{n+1} = Tv^n = T^{n+1}v^0$ converges to v^*.

Solutions of the Optimality Equations

Proposition

Suppose $\lambda \in [0,1)$. Then L and $\mathcal{L}$ are contraction mappings on V.

Theorem

Suppose $\lambda \in [0,1)$, S is finite or countable, and $r(s,a)$ is bounded. The following results hold.

(i) There exits a $v^* \in V$ satisfying $Lv^* = v^*$ ($\mathcal{L}v = v^*$). Furthermore, v^* is the only element of V with this property and equals v_λ^*;

(ii) For each $d \in D^{MR}$, there exists a unique $v \in V$ satisfying $L_d v = v$. Furthermore, $v = v_\lambda^{d^\infty}$.

Existence of Stationary Optimal Policies

- A decision rule is d^* is conserving if

$$d^* \in \underset{d \in D}{\operatorname{argmax}} \{r_d + \lambda P_d v_\lambda^*\}.$$

Theorem

Suppose there exists a conserving decision rule or an optimal policy, then there exists a deterministic stationary policy which is optimal.

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
$$v^{n+1}(s) = \max_{a \in A_s} \left[r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v^n(j) \right].$$
3. If
$$||v^{n+1} - v^n|| < \frac{\epsilon(1-\lambda)}{2\lambda},$$
go to step 4. Otherwise, increment n by 1 and return to step 2.
4. For each $s \in S$, choose
$$d_\epsilon(s) \in \operatorname*{argmax}_{a \in A_s} \left[r(s,a) + \sum_{j \in S} \lambda p(j|s,a) v^{n+1}(j) \right]$$
and stop.

Policy Iteration

1. Set $n = 0$ and select an arbitrary decision rule $d_0 \in D$.
2. (Policy evaluation) Obtain v^n by solving
$$(I - \lambda P_{d_n})v = r_{d_n}.$$
3. (Policy improvement) Choose d_{n+1} satisfy
$$d_{n+1} \in \operatorname*{argmax}_{d \in D} [r_d + \lambda P_d v^n].$$
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.

Linear Programming

- Let $\alpha(s)$ be positive scalars such that $\sum_{s\in S} \alpha(s) = 1$.
- Primal linear program is given by

$$\min_v \sum_{j\in S} \alpha(j)v(j)$$

$$v(s) - \sum_{j\in S} \lambda P(j|s,a)v(j) \geq r(s,a), \qquad \forall s \in S, a \in A_s.$$

- Dual linear program is given by

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

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

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