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 $\pi = (d_1, d_2, \ldots) \in \Pi^{HR}$

Starting at a state $s$, using policy $\pi$ leads to a sequence of state-action pairs $\{X_t, Y_t\}$. The sequence of rewards is given by $\{R_t \equiv r_t(X_t, Y_t) : t = 1, 2, \ldots\}$.

Let $\lambda \in [0, 1)$ be the discount factor

The expected total rewards from policy $\pi$ starting in state $s$ is given by

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

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

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

Let

\[
\nu^\pi(s) \equiv E_s^\pi \left[ \sum_{t=1}^{\infty} r(X_t, Y_t) \right].
\]

We have \( \nu^\pi(s) = \lim_{\lambda \uparrow 1} \nu_\lambda^\pi(s) \) whenever \( \nu^\pi(s) \) exists.
A policy $\pi$ 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 $\pi^*$ be a discount optimal policy. Then $v_{\lambda}^{\pi^*}(s) = v_{\lambda}^*(s)$ for all $s \in S$. 
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 $s \in S$. 

Dan Zhang, Spring 2012
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*. 
\( \pi = (d_1, d_2, \ldots) \in \Pi^{MR}. \) The \((s, j)\) component of the \(t\)-step transition probability matrix \( P^t_\pi(j|s) \) satisfies

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

- For \( v \in V, \)
  \[
  E^\pi_s[v(X_t)] = \sum_{j \in S} P^{t-1}_\pi(j|s)v(j).
  \]

- We also have
  \[
  v^\pi_\lambda = \sum_{t=1}^{\infty} \lambda^{t-1} P^{t-1}_\pi r_{d_t}.
  \]
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
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

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

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

$\implies$ It suffices to consider $\Pi^{MR}$.

$$v^*_\lambda(s) = \sup_{\pi \in \Pi^{HR}} v^\pi_\lambda(s) = \sup_{\pi \in \Pi^{MR}} v^\pi_\lambda(s).$$
Let $\pi = (d_1, d_2, \ldots) \in \Pi^{MR}$. 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

$$v_{\lambda}^\pi = \sum_{t=1}^{\infty} \lambda^{t-1} P_{\pi}^{t-1} r_d = r_d_1 + \lambda P_{\pi}^1 r_d_2 + \lambda^2 P_{\pi}^2 r_d_3 + \ldots$$

$$= r_d_1 + \lambda P_{d_1} r_d_2 + \lambda^2 P_{d_1} P_{d_2} r_d_3 + \ldots$$

$$= r_d_1 + \lambda P_{d_1} (r_d_2 + \lambda P_{d_2} r_d_3 + \ldots)$$

$$= r_d_1 + \lambda P_{d_1} v_{\lambda}^{\pi'},$$

where $\pi' = (d_2, d_3, \ldots)$.
When $\pi$ is stationary, $\pi = (d, d, \ldots) \equiv d^\infty$ and $\pi' = \pi$.

It follows that $v^d_\lambda$ satisfies

$$v^d_\lambda = r_{d_1} + \lambda P_d v^d_\lambda \equiv L_d v^d_\lambda,$$

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

**Theorem**

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

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

Furthermore, $v^d_\lambda$ may be written as

$$v^d_\lambda = (I - \lambda P_d)^{-1} r_d.$$
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 $\nu \in V$, let

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

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

$$
v = Lv.
$$
Suppose \( \nu \in \mathcal{V} \).

(i) If \( \nu \geq \mathcal{L} \nu \), then \( \nu \geq \nu^*_\lambda \);

(ii) If \( \nu \leq \mathcal{L} \nu \), then \( \nu \leq \nu^*_\lambda \);

(iii) If \( \nu = \mathcal{L} \nu \), then \( \nu \) is the only element of \( \mathcal{V} \) with this property and \( \nu = \nu^*_\lambda \).
Let $U$ be a Banach space (complete normed linear space).
- Special case: space of bounded measurable real-valued functions
- An operator $T : U \rightarrow U$ is a contraction mapping if there exists a $\lambda \in [0, 1)$ such that $||Tv - Tu|| \leq \lambda ||v - u||$ for all $u$ and $v$ in $U$.

**Theorem [Banach Fixed-Point Theorem]**
Suppose $U$ is a Banach space and $T : U \rightarrow U$ 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.

1. There exists a $v^* \in V$ satisfying $Lv^* = v^*$ ($\mathcal{L}v = v^*$). Furthermore, $v^*$ is the only element of $V$ with this property and equals $v^*_\lambda$.
2. For each $d \in D^{MR}$, there exists a unique $v \in V$ satisfying $L_d v = v$. Furthermore, $v = v^d_\lambda$.
A decision rule is $d^*$ is conserving if

$$d^* \in \arg\max_{d \in D} \{ r_d + \lambda P_d v^*_\lambda \}.$$
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

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

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

\[
(l - \lambda P_{d_0})v = r_{d_0}.
\]

3. (Policy improvement) Choose $d_{n+1}$ satisfy

\[
d_{n+1} \in \arg \max_{d \in D} [r_d + \lambda P_d v^n].
\]

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.
Let $\alpha(s)$ be positive scalars such that $\sum_{s \in S} \alpha(s) = 1$.

Primal linear program is given by

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

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

Dual linear program is given by

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

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

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