Markov Decision Processes I

Daisuke Oyama

Topics in Economic Theory

December 10, 2014

Formulation (Finite States/Actions)

- State space: $S = \{0, \ldots, n-1\}$
- Action space: $A = \{0, \ldots, m-1\}$
- Reward function: r

  $r(s, a)$ is the reward for action $a \in A$ when the state is $s \in S$.
- Transition probability function: q

  $q(s'|s, a)$ is the probability that the state in the next period is $s' \in S$ when the current state is $s \in S$ and the action chosen is $a \in A$.
- Feasibility constraints:
  $\Gamma(s) = \{a \in A \mid a \text{ is feasible when the state is } s\}$.

  Embedded in r: $r(s, a) = -\infty$ if $a \notin \Gamma(s)$.
- Discount factor: $\beta \in [0, 1)$.

- A *policy function* (or simply, *policy*) is a function $\sigma\colon S \to A$.

  (We only consider *feasible* policies, σ such that $r(s, \sigma(s)) > -\infty$.)

- A *plan* is a sequence $(\sigma_0, \sigma_1, \ldots)$ of policies.

- For a given plan $\pi = (\sigma_0, \sigma_1, \ldots)$ and an initial state s,

  - the reward at period 0 is $r(s, \sigma_0(s))$;
  - the expected reward at period 1 is $\sum_{s'} q(s'|s, \sigma_0(s)) r(s', \sigma_1(s'))$;
  - the expected reward at period 2 is $\sum_{s',s''} q(s'|s, \sigma_0(s)) q(s''|s', \sigma_1(s')) r(s'', \sigma_2(s''))$, or $(Q_\pi^2 r_{\sigma_2})[s]$, where
    - $Q_\pi^2 \in \mathbb{R}^{n \times n}$: $Q_\pi^2 = Q_{\sigma_0} Q_{\sigma_1}$;
    - $Q_\sigma \in \mathbb{R}^{n \times n}$: $Q_\sigma[s, s'] = q(s'|s, \sigma(s))$;
    - $r_\sigma \in \mathbb{R}^n$: $r_\sigma[s] = r(s, \sigma(s))$.

- Objective:

  Find a plan $\pi = (\sigma_0, \sigma_1, \ldots)$ that maximizes

  $$v_\pi(s) = \sum_{t=0}^{\infty} \beta^t (Q_\pi^t r_{\sigma_t})[s]$$

  for each $s \in S$, where $Q_\pi^0 = I$.

- The *optimal value function* (or simply, *value function*) is the function $v^* : S \to \mathbb{R}$ that satisfies

  $$v^*(s) = \max_\pi v_\pi(s)$$

  for all $s \in S$.

- π^* is an *optimal plan* if $v_{\pi^*} = v^*$.

- σ^* is an *optimal policy* if $(\sigma_0, \sigma_1, \ldots)$ with $\sigma_t = \sigma^*$ for all t is an optimal plan, i.e., $v_{\sigma^*} = v^*$, where $v_{\sigma^*} = v_{(\sigma^*, \sigma^*, \ldots)}$.

Operators

- Bellman operator:

  $T\colon \mathbb{R}^S \to \mathbb{R}^S$ defined by

$$(Tw)(s) = \max_{a \in A} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) w(s').$$

- For a policy σ,

  $T_\sigma\colon \mathbb{R}^S \to \mathbb{R}^S$ defined by

$$(T_\sigma w)(s) = r(s,\sigma(s)) + \beta \sum_{s' \in S} q(s'|s,\sigma(s)) w(s').$$

- By definition, $T_\sigma w \leq Tw$.
- $v_\sigma = T_\sigma v_\sigma$.

Monotonicity

- T and T_σ are monotone,

  i.e., if $v \leq w$, then

  $$Tv \leq Tw,$$
  $$T_\sigma v \leq T_\sigma w.$$

Contraction

- T and T_σ are uniformly contracting with modulus β.
- v_σ is the unique fixed point of T_σ.
- T has a unique fixed point.

Theorem

1. The unique fixed point of T is the value function v^*.
2. σ^* is an optimal policy if and only if

$$\sigma^*(s) \in \arg\max_{a \in A} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v^*(s')$$

   for all $s \in S$, or equivalently, $T_{\sigma^*} v^* = T v^*$.
3. (Another expression:

   σ^* is an optimal policy if and only if

$$\sigma^*(s) \in \arg\max_{a \in A} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v_{\sigma^*}(s')$$

   for all $s \in S$, or equivalently, $T_{\sigma^*} v_{\sigma^*} = T v_{\sigma^*}$.)
4. An optimal policy exists.

Proof (1/2)

1.
   - Let w^* be the unique fixed point of T: $Tw^* = w^*$.
   - Let σ^* be such that $T_{\sigma^*} w^* = Tw^*$ (such a policy exists), for which we have $T_{\sigma^*} w^* = w^*$.
   - Since v_{σ^*} is the unique fixed point of T_{σ^*}, we must have $w^* = v_{\sigma^*}$.
   - Take any plan $\pi = (\sigma_0, \sigma_1, \ldots)$.
   - Since for any policy σ, we have $w^* = Tw^* \geq T_\sigma w^*$, and T_σ is monotone, we have $w^* \geq T_{\sigma_0} w^* \geq T_{\sigma_0} T_{\sigma_1} w^* \geq T_{\sigma_0} T_{\sigma_1} T_{\sigma_2} w^* \geq \cdots \searrow v_\pi$.
   - This means that $w^* = \max_\pi v_\pi$.

Proof (2/2)

2. $v^* = v_\sigma$
   $\iff T_\sigma v^* = v^*$ ($\because v_\sigma$ is the unique fixed point of T_σ)
   $\iff T_\sigma v^* = T v^*$ ($\because v^*$ is a fixed point of T)

3. $v^* = v_\sigma$
   $\iff T v_\sigma = v_\sigma$ ($\because v^*$ is the unique fixed point of T)
   $\iff T v_\sigma = T_\sigma v_\sigma$ ($\because v_\sigma$ is a fixed point of T_σ)

Value Iteration

- Take any v_0.
- Let $v_{i+1} = Tv_i$.
- Then $v_i \to v^*$ as $i \to \infty$.

Policy Iteration

- Take any σ_0.
- [Policy evaluation] Compute v_{σ_i}, which satisfies $v_{\sigma_i} = T_{\sigma_i} v_{\sigma_i}$.
- [Policy improvement] Compute σ_{i+1} such that

$$\sigma_{i+1}(s) \in \arg\max_{a \in A} r(s,a) + \beta \sum_{s' \in S} q(s'|s,a) v_{\sigma_i}(s')$$

for all $s \in S$, or

$$T_{\sigma_{i+1}} v_{\sigma_i} = T v_{\sigma_i}.$$

Proposition 1

1. $v_{\sigma_i} = T_{\sigma_i} v_{\sigma_i} \leq T v_{\sigma_i} = T_{\sigma_{i+1}} v_{\sigma_i} \leq T_{\sigma_{i+1}}^2 v_{\sigma_i} \leq \cdots \nearrow v_{\sigma_{i+1}}$.
2. $T^i v_{\sigma_0} \leq v_{\sigma_i}$ $(\leq v^*)$.

   (∵ By induction: $T^{i+1} v_{\sigma_0} \leq T v_{\sigma_i} \leq v_{\sigma_{i+1}}$.*)*
3. *Hence,* $v_{\sigma_i} \to v^*$ *as* $i \to \infty$.
4. *If* $v_{\sigma_i} = v_{\sigma_{i+1}}$*, then* $v_{\sigma_i} = T v_{\sigma_i}$ *and hence* σ_i *is optimal;*

   if $v_{\sigma_i} \neq v_{\sigma_{i+1}}$*, then* $v_{\sigma_i} \neq T v_{\sigma_i}$ *and hence* σ_i *is not optimal.*

Policy Evaluation (1): Solving a Linear Equation

- Given σ, solve the system of linear equations

$$v(s) = r(s, \sigma(s)) + \beta \sum_{s' \in S} q(s'|s, \sigma(s))v(s'),$$

or

$$(I - \beta Q_\sigma)v = r_\sigma,$$

where

  - $Q_\sigma \in \mathbb{R}^{n \times n}$: $Q_\sigma[s, s'] = q(s'|s, \sigma(s))$;
  - $r_\sigma \in \mathbb{R}^n$: $r_\sigma[s] = r(s, \sigma(s))$.

Policy Evaluation (2): Approximation by Iteration

- For any w, $T_\sigma^i w \to v_\sigma$.

  Thus, for large k, $T_\sigma^k w = \sum_{t=0}^{k-1} \beta^t Q_\sigma^t r_\sigma + \beta^k Q_\sigma^T w \approx v_\sigma$.

- In particular,
  given σ_{i+1}, where $T_{\sigma_{i+1}} v_{\sigma_i} = T v_{\sigma_i}$, for $w = v_{\sigma_i}$,
  $T_{\sigma_{i+1}} v_{\sigma_i} \leq T_{\sigma_{i+1}}^2 v_{\sigma_i} \leq \cdots \nearrow v_{\sigma_{i+1}}$.

  Thus, for large k,

$$T_{\sigma_{i+1}}^k v_{\sigma_i} = \sum_{t=0}^{k-1} \beta^t Q_{\sigma_{i+1}}^t r_{\sigma_{i+1}} + \beta^k Q_{\sigma_{i+1}}^k w \approx v_{\sigma_{i+1}}.$$

- "Modified" policy iteration.
- If $k = 1$, then $T_{\sigma_{i+1}} v_{\sigma_i} = T v_{\sigma_i}$ $\cdots$ value iteration.

  If $k \to \infty$, then $T_{\sigma_{i+1}}^k v_{\sigma_i} \to v_{\sigma_{i+1}}$
  $\cdots$ "exact" policy iteration.