Review of Chapter2 $\sim$ Chapter6

Markov Decision Processes:Discrete Stochastic Dynamic Programming

Peng Lingwei

August 13, 2019

.

Chapter2:Model Formulation

$$Model : \{\mathcal{T}, \mathcal{S}, \mathcal{A}, \underbrace{\mathcal{R}(s,a), \mathcal{P}(s'|s,a)}_{Markovian}\}$$

1. $\mathcal{T} = (1, 2, 3, \ldots)$: finite horizon, infinite horizon;
2. $\tau = (s_1, a_1, r(s_1, a_1), s_2, a_2, r(s_2, a_2), \ldots)$;
3. $h_t = (s_1, a_1, s_2, a_2, \ldots) \in H_t = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A} \ldots$
4. $\mathcal{S}, \mathcal{A}$:
   - finite sets;
   - countable infinite sets;
   - compact sets of finite dimensional Euclidean space;
   - non-empty Borel subsets of complete, seperable metric spaces.
5. $\int \mathcal{R}(s, a, s')\mathcal{P}(s'|s, a)ds' = \mathcal{R}(s, a)$.

Chapter2:Model Formulation

$$Decider : \{D^k, \Pi^k, k = \{HR, HD, MR, MD\}\}$$

1. $d \in D^k : d^{HR}(h_t) = P(\mathcal{A}_{s_t}), d^{HD}(h_t) = a_t, d^{MR}(s_t) = P(\mathcal{A}_{s_t}), d^{MD}(s_t) = a_t;$
2. $\pi = (d_1, d_2, \ldots) \in \Pi^K = D^K \times D^K \times \ldots;$
3. Stateionary policy: $\pi_d = (d, d, \ldots)$.

$$\text{Optimality criterion} : \{V^\pi(s), Q^\pi(s,a), V^*, Q^*\}$$

1. $V^\pi = \left\{V_N^\pi, V_1^\pi, V_\lambda^\pi, V_{avg}^\pi\right\}$

$$Model + Decider = \text{Markov Decision Process}$$

$$Model + Decider + \text{Optimality criterion} = \text{Markov Decision Problem}$$

Chapter4:Finite-Horizon Markov Decision Process

1. We only assume that $\mathcal{S}$ is discrete.
2. $T = (1, 2, \ldots, N-1, N)$
3. $\tau = (s_1, a_1, r(s_1, a_1), \ldots, s_{N-1}, a_{N-1}, r(s_{N-1}, a_{N-1}), s_N, r(s_N))$
4. Optimal criterion:
   - Value function: $V_N^\pi(s) = \mathbb{E}_{\tau\sim\mathcal{P}}\left\{\sum_{t=1}^{N-1} r(s_t, a_t) + r_N(s_N)\right\}$;
   - Optimal value: $V_N^* = \sup_{\pi\in\Pi^{HR}} V_N^\pi$;
   - Optimal policy π^*: $\forall\pi\in\Pi^{HR}, V_N^{\pi^*} \succeq V_N^\pi$;
   - $\epsilon -$ *Optimal* policy π_ϵ^*: $\forall\pi\in\Pi^{HR}, V_N^{\pi_\epsilon^*} + \epsilon \succeq V_N^\pi$.
5. Initial state distribution $\mathcal{P}_1 = P\{s_1 = s\}$,
   $V_N^{\pi,\mathcal{P}_1} = \sum_{s=S} v_N^\pi(s) P\{s_1 = s\}$.

Policy Evaluation

Define:
$u_t^\pi(h_t) = \mathbb{E}_{h_t}^\pi \left\{ \sum_{n=t}^{N-1} r_n(X_n, Y_n) + r_N(X_N) \right\}, u_N^\pi(h_N) = r(s_N).$
For $\hat{u}_N^\pi(h_N) = u_N^\pi(h, N) = r(s_N)$, we can use backward induction to calculate any policy $\pi = (d_1, d_2, \dots, d_{N-1})$'s value:

1. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{HR}$):
$$\hat{u}_t^\pi(h_t) = \sum_{a \in A_{s_t}} q_{d_t(h_t)}(a) \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s'|s_t, a) \hat{u}_{t+1}^\pi(h_t, a, s'). \right\}$$
2. Finite horizon-policy evaluation algorithm ($\pi \in \Pi^{MD}$):
$$\hat{u}_t^\pi(s_t) = r_t(s_t, d_t(s_t)) + \sum_{s' \in S} p_t(s'|s_t, d_t(s_t)) \hat{u}_{t+1}^\pi(s').$$
3. $\pi \in \Pi^{HR} : \sum_{i=1}^N (|\mathcal{S}||\mathcal{A}|)^i$;
4. $\pi \in \Pi^{MD} : (N-1)|\mathcal{S}|^2|\mathcal{A}|$

Optimal Equation

Definition

(Optimal equation or bellman equaiton).

$$\hat{u}_t(h_t) = \sup_{a \in A_{s_t}} \left\{ r_t(s_t, a) + \sum_{s' \in S} p_t(s' | s_t, a) \hat{u}_{t+1}(h_t, a, s') \right\},$$

$$s.t.\ \hat{u}_N(h_N) = r_N(s_N).$$

Definition

$$u_t^*(h_t) = \sup_{\pi \in \Pi^{HR}} u_t^\pi(h_t)$$

Theorem

$$\forall h_t \in H_t, \hat{u}_t(h_t) = u_t^*(h_t)$$

Chapter4:Infinite-Horizon Models:Foundations

1. From optimal equation, we can get
   $v_N^* = \sup_{\pi \in \Pi^{HR}} v_N^\pi = \sup_{\pi \in \Pi^{HD}} v_N^\pi$.
2. $v_N^* = \sup_{\pi \in \Pi^{HR}} v_N^\pi = \sup_{\pi \in \Pi^{MD}} v_N^\pi$
3. The condition of attainable optimal equation:
   - $\mathcal{S}$ is discrete;
   - $\mathcal{A}$:
     - A_s is countable, or;
     - A_s is compact; $\mathcal{P}(s'|s,a)$ is lower semi-continuous in a and $r(s,a)$ is upper semicontinuous in a and $|r(s,a)| \leq M$.

Chapter5: Infinite-horizon Models

Optimality criterion:

1. Expected total reward: $v_1^{\pi}(s) = \mathbb{E}_S^{\pi}\left\{\sum_{t=1}^{\infty} r(X_t, Y_t)\right\}$;
2. Expected total discounted reward:
   $v_{\lambda}^{\pi}(s) = \mathbb{E}_S^{\pi}\left\{\sum_{t=1}^{\infty} \lambda^{t-1} r(X_t, Y_t)\right\}$;
3. Average reward:
   $V_{avg}^{\pi}(s) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_S^{\pi}\left\{\sum_{t=1}^{n} r(X_t, Y_t)\right\} = \lim_{n\to\infty} \frac{1}{n} v_{n+1}^{\pi}(s)$

Markov policies:

$$v_{\lambda}^{\pi}(s_1) = \sum_{t=1}^{\infty} \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}_s'} \lambda^{t-1} r(s', a) P^{\pi}\left\{S_t = s', A_t = a | S_1 = s_1\right\}$$

Target:

$$P^{\pi}\left\{S_t = s', A_t = a | S_1 = s_1\right\} = P^{\pi'}\left\{S_t = s', A_t = a | S_1 = s_1\right\}$$

Method:

$$\pi' = \left\{d_1', d_2', \ldots\right\}, q_{d_t'(s')}(a) = P^{\pi}\left\{A_t = a | S_t = s', S_1 = s_1\right\}$$

Proof.

(which means $v^* = \sup_{\pi \in \Pi^{HR}} v^\pi = \sup_{\pi \in \Pi^{MR}} v^\pi$).

$$
\begin{aligned}
& P^\pi \left\{ S_t = s' | S_1 = s_1 \right\} \\
=& \sum_{s \in S} \sum_{a \in A_s} P^\pi \left\{ S_{t-1} = s, A_{t-1} = a | S_1 = s_1 \right\} p(s'|s,a) \\
=& \sum_{s \in S} \sum_{a \in A_s} P^{\pi'} \left\{ S_{t-1} = s, A_{t-1} = a | S_1 = s_1 \right\} p(s'|s,a) \\
=& P^{\pi'} \left\{ S_t = s' | S_1 = s_1 \right\}
\end{aligned}
$$

$$
\begin{aligned}
& P^\pi \left\{ S_t = s', A_t = a | S_1 = s_1 \right\} \\
=& P^\pi \left\{ A_t = a | S_t = s', S_1 = s_1 \right\} P^\pi \left\{ S_t = s' | S_1 = s_1 \right\} \\
=& P^{\pi'} \left\{ A_t = a | S_t = s', S_1 = s_1 \right\} P^{\pi'} \left\{ S_t = s' | S_1 = s_1 \right\} \\
=& P^{\pi'} \left\{ S_t = s', A_t = a | S_1 = s_1 \right\}
\end{aligned}
$$

Chapter6:Discounted Markov Decision Problems

Key assumption:

1. Stationary rewards and transition probabilities;
2. $|r(s,a)| \leq M < \infty$;
3. $0 \leq \lambda < 1$;
4. Discrete state space $\mathcal{S}$.

Policy evaluation (Stationary policy):

$$v_\lambda^{\pi_d} = \mathbb{E}^{\pi_d}\left\{\sum_{t=1}^{\infty} \lambda^{t-1} r(s_t, a_t)\right\} = r_d + \lambda P_d v_\lambda^{\pi_d} = (I - \lambda P_d)^{-1} r_d$$

Optimal equation (Bellman equation):

$$v(s) = \sup_{a \in A_s}\left\{r(s,a) + \sum_{s' \in S} \lambda p(s'|s,a) v(s')\right\}$$

Optimal equations

Theorem

$\forall v \in V, 0 \leq \lambda < 1$,

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

Proof.

First, $D^{MD} \subset D^{MR}$, *so*

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

Second, $\forall d^{MR} \in D^{MR}$,

$$\sum_{a \in A_s} q_{d^{MR}}(a) \left[r(s,a) + \sum_{s' \in S} \lambda p(s'|s,a) v(s') \right]$$

$$\leq \sup_{a \in A_s} \left\{ r(s,a) + \sum_{s' \in S} \lambda p(s'|s,a) v(s') \right\}$$

$\square$

Optimal equations

Definition

(Bellman operator)

$$\mathcal{L}v = \sup_{d \in D^{MD}} \{r_d + \lambda P_d v\}, \quad Lv = \max_{d \in D^{MD}} \{r_d + \lambda P_d v\}$$

Theorem

1. $v \succeq \mathcal{L}v \Rightarrow v \succeq v_\lambda^*$;
2. $v \preceq \mathcal{L}v \Rightarrow v \preceq v_\lambda^*$;
3. $v = \mathcal{L}v \Rightarrow v$ *is unique and* $v = v_\lambda^*$.

Solving bellman equation:

$$v^{n+1} = \mathcal{L}v^n, \quad \lim_{n \to \infty} v^n = v_\lambda^*$$

(The proof need: banach fixed-point theorem, contraction mapping.)

Optimal equations

Theorem

Assume $\mathcal{S}$ is discrete, and either

1. *A_s is finite for each $s \in S$, or;*
2. *A_s is compact, $r(s, a)$ is continuous in a for each $s \in S$, and for each $s' \in S$ and $s \in S$, $p(s'|s, a)$ is continuous in a, or;*
3. *A_s is compact, $r(s, a)$ is upper semicontinuous in a for each $s \in S$, and for each $s' \in S$ and $s \in S$, $p(s'|s, a)$ is lower semicontinuous in a.*

Then there exists an optimal deterministic stationary policy.

Value Iteration

Algorithm 1 Value Iteration Algorithm

Require: $\epsilon > 0$

Ensure: $v^0 \in V$

for $n = 1, 2, \ldots$ **do**

$\forall s \in S, v^{n+1}(s) = \max_{a \in A_s} \left\{ r(s,a) + \sum_{s' \in S} \lambda p(s'|s,a) v^n(s') \right\}$

if $\|v^{n+1} - v^n\| < \epsilon(1-\lambda)/(2\lambda)$ **then**

break.

end if.

end for.

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

Convergence rate of value iteration:

$\|v^{n+1} - v^*_\lambda\| = \|Lv^n - Lv^*_\lambda\| \leq \lambda \|v^n - v^*_\lambda\|$

Policy Iteration

Algorithm 2 Policy Iteration Algorithm

Select an arbitrary rule $d_0 \in D^{MD}$.
for $n = 1, 2, \ldots$ **do**
  Policy evaluation: $v^n = (I - \lambda P_{d_n})^{-1} r_{d_n}$
  Policy improvement: $d_{n+1} \in \arg\max_{d \in D^{MD}} \{r_d + \lambda P_d v^n\}$
  if $d_{n+1} = d_n$ **then**
    break.
  end if.
end for.
return d_{n+1}

Convergence rate of policy iteration: If policy iteration's sequence $\{v^n\}$ satisfies $\|P_{d_{v^n}} - P_{d_{v^*_\lambda}}\| \leq K\|v^n - v^*_\lambda\|$ (for some K),then

$$\|v^{n+1} - v^*_\lambda\| \leq \frac{K\lambda}{1-\lambda}\|v^n - v^*_\lambda\|^2$$

Policy Iteration

1. $v^{n+1} \succeq v^n$

$$r_{d_{n+1}} + \lambda P_{d_{n+1}} v^n \succeq r_{d_n} + \lambda P_{d_n} v^n = v^n$$

$$v^{n+1} = (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} \succeq v^n$$

2. Let $Bv = Lv - v$, then $\forall u, v \in V$ and $d_v \in D_v$.

$$Bu \geq Bv + (\lambda P_{d_v} - I)(u - v) \Rightarrow (\lambda P_{d_v} - I) \in \partial_v(Bv)$$

3. $v^{n+1} = v^n - (\lambda P_{d_{v^n}} - I)^{-1} B v^n$.

$$\begin{aligned} v^{n+1} =& (I - \lambda P_{d_{v^n}})^{-1} r_{d_{v^n}} - v^n + v^n \\ =& v^n - (\lambda P_{d_{v^n}} - I)^{-1} \left[r_{d_{v^n}} + (\lambda P_{d_{v^n}} - I) v^n \right] \\ =& v^n - (\lambda P_{d_{v^n}} - I)^{-1} B v^n \end{aligned}$$

Modified Policy Iteration

In policy iteration, we have

$$v^{n+1} = v^n - (\lambda P_{d_{v^n}} - I)^{-1} B v^n = v^n + \sum_{k=0}^{\infty} \left(\lambda P_{d_{n+1}}^k B v^n \right)$$

In modified policy iteration

$$v^{n+1} = v^n - (\lambda P_{d_{v^n}} - I)^{-1} B v^n = v^n + \sum_{k=0}^{m_n} \left(\lambda P_{d_{n+1}}^k B v^n \right)$$

$$\begin{aligned} v^{n+1} =& v^n + \sum_{k=0}^{m_n} \left(\lambda P_{d_{n+1}} \right)^k \left[r_{d_{n+1}} + \lambda P_{d_{n+1}} v^n - v^n \right] \\ =& r_{d_{n+1}} + \lambda P_{d_{n+1}} r_{d_{n+1}} + \cdots + \left(\lambda P_{d_{n+1}} \right)^{m_n} r_{d_{n+1}} + \left(\lambda P_{d_{n+1}} \right)^{m_n+1} v^n \\ =& (L_{d_{n+1}})^{m_n+1} v^n \end{aligned}$$

Modified Policy Iteration

Algorithm 3 Modified Policy Iteration Algorithm (MPI)

Require: $\epsilon > 0, \{m_0, m_2, \dots\}$.
Ensure: $v^0 \in V_B$.
  for $n = 0, 1, \dots$ **do**
    $d_{n+1} \in \arg\max_{d \in D} \{r_d + \lambda P_d v^n\}$
    $u_n^0 = r_{d_{n+1}} + \lambda P_{d_{n+1}} v^n$
    if $\|u_n^0 - v^n\| < \epsilon(1-\lambda)/(2\lambda)$ **then** break
    end if.
    for $k = 0, 1, \dots, m_n$ **do**
      $u_n^{k+1} = r_{d_{n+1}} + \lambda P_{d_{n+1}} u_n^k = L_{d_{n+1}} u_n^k$
    end for.
    $(v^{n+1} = L_{d_{n+1}}^{m_n+1} v^n)$
  end for.
  return d_{n+1}

Convergence rate of modified policy iteration:

$$\|v^{n+1} - v_\lambda^*\| \leq \left(\frac{\lambda(1-\lambda^{m_n})}{1-\lambda} \|P_{d_n} - P_{d^*}\| + \lambda^{m_n+1} \right) \|v^n - v_\lambda^*\|.$$