1.2 Finite- Horizon MDP

Def [finite-horizon, time-dependent MDP]

$$M = (S, A, \{P\}_h, \{r\}_h, H, \mu)$$

① state space S : finite or infinite

② action space A : discrete or finite

③ time-depend transition function $P_h : S \times A \to \Delta(S)$

h : time step h.

④ time-depend reward function $r_h : S \times A \to [0,1]$

⑤ integer H : horizon length.

⑥ initial state distribution $\mu \in \Delta(S)$

Def [value function] $V_h^\pi : S \to \mathbb{R}$

$$V_h^\pi(s) = E\left[\sum_{t=h}^{H-1} r_h(S_t, a_t) \,\middle|\, \pi, S_h = s \right]$$

Note: ① the expectation is w.r.t. the randomness of trajectory.

② $V^\pi(s) := V_0^\pi(s)$

Def [state-action value function] $Q_h^\pi : S \times A \to \mathbb{R}$

$$Q_h^\pi(s,a) = E\left[\sum_{t=h}^{H-1} r_h(S_t, a_t) \,\middle|\, \pi, S_h = s, a_h = a \right]$$

The optimization problem is to find such π^* s.t.

$$V^{\pi^*}(s) = \max_\pi V^\pi(s)$$

Theorem 1.9 [Bellman optimality equations]

Define $Q_h^*(s,a) = \sup_{\pi \in \Pi} Q_h^\pi(s,a)$,

suppose $Q_H(s,a) = 0$,

we have $Q_h = Q_h^*$ for all $h \in [H]$

$\Leftrightarrow$ for all $h \in [H]$,

$$Q_h(s,a) = r_h(s,a) + E_{s' \sim P_h(s,a)}\left[\max_{a' \in A} Q_{h+1}(s',a')\right].$$

Furthermore, $\pi(s,h) = \arg\max_{a \in A} Q_h^*(s,a)$ is an optimal policy.

Pf: ($\Rightarrow$) Let π^* be a stationary and deterministic policy

s.t. $\pi^*(s_h) = \arg\max_{a \in A} Q_h^*(s_h, a)$

then we have

$$\begin{aligned}
Q_h^*(s,a) &= \sup_{\pi \in \Pi} Q_h^\pi(s,a) \\
&= r(s,a) + \sup_{\pi \in \Pi} E_{s' \sim p(s,a)}\left[\sum_{t=h+1}^{H-1} r(s_t, a_t) \,\middle|\, \pi, s_{t+1} = s'\right] \\
&= r(s,a) + \sup_{\pi \in \Pi} E_{s' \sim p(s,a)}\left[V_{h+1}^\pi(s')\right] \\
&= r(s,a) + \sup_{\pi \in \Pi} E_{s' \sim p(s,a)}\left[Q_{h+1}^\pi(s', \pi(s'))\right] \\
&= r(s,a) + E_{s' \sim p(s,a)}\left[Q_{h+1}^{\pi^*}(s', \pi^*(s'))\right] \\
&= r(s,a) + E_{s' \sim p(s,a)}\left[\max_{a' \in A} Q_{h+1}^*(s', a')\right]
\end{aligned}$$

($\Leftarrow$) Let π be a stationary and deterministic policy

s.t. $\pi(s_h) = \arg\max_{a \in A} Q_h(s,a)$

$a \in A$

then we have

$$Q_h(s,a) = r_h(s,a) + E_{s' \sim P(s,a)}\left[\max_{a' \in A} Q_{h+1}(s',a')\right]$$

$$= r_h(s,a) + E_{s' \sim P(s,a)}\left[Q_{h+1}(s', \pi(s'))\right]$$

for all $t \geq h$. which implies $Q_{t \geq h}$ can be calculated following the policy π.

Thus we get note that π is deterministic.

$$Q_h(s,a) = r_h(s,a) + E_{s' \sim P(s,a)}\left[Q^{\pi}_{h+1}(s', \pi(s'))\right]$$

$$= Q^{\pi}_h(s,a)$$

for any other deterministic and stationary policy π':

$$\left[(P^{\pi}_h - P^{\pi'}_h) Q^{\pi}_h\right]_{s,a} = E_{s' \sim P(s,a)}\left[Q^{\pi}_{h+1}(s', \pi(s')) - Q^{\pi}_{h+1}(s', \pi'(s'))\right]$$

$$= E_{s' \sim P(s,a)}\left[Q_{h+1}(s', \pi(s')) - Q_{h+1}(s', \pi'(s'))\right]$$

$$\geq 0$$

$$\Rightarrow Q_h - Q^{\pi'}_h = Q^{\pi}_h - Q^{\pi'}_h$$

$$= Q^{\pi}_h - (I - P^{\pi'}_h)^{-1} r_h$$

$$= (I - P^{\pi'}_h)^{-1}\left[(I - P^{\pi'}_h) - (I - P^{\pi}_h)\right] Q^{\pi}_h$$

$$= (I - P^{\pi'}_h)^{-1} (P^{\pi}_h - P^{\pi'}_h) Q^{\pi}_h$$

By lemma 1.6, $(I - P^{\pi'}_h)^{-1} \geq 0$, then we have

$$Q_h - Q^{\pi'}_h = Q^{\pi}_h - Q^{\pi'}_h \geq 0$$

Analogously we can prove $Q_h^* = Q_h^{\pi^*}$

we have $Q_h^* \geq Q_h^{\pi} \geq Q_h^{\pi^*} = Q_h^*$

$\Rightarrow Q_h = Q_h^{\pi} = Q_h^{\pi^*} = Q_h^*$ $\square$

Discussion:

① Time-dependent MDPs are more convenient for analysis

② Time-dependent MDPs cost $O(H)$ more memory which makes it less utilized.

③ In practice, temporal info would alway be incorporated into states.