1 Markov Decision Process, MDP

Def [MDP]: $M = (S, A, P, r, \gamma, \mu)$

① state space: S, finite / countably infinite

② action space: A, finite / infinite

③ transition function: $P: S \times A \to \Delta(S)$ distributions over S

④ reward function: $r: S \times A \to [0,1]$ or generally $\Delta([0,1])$

⑤ discount factor: $\gamma \in [0,1)$

⑥ initial state distribution $\mu \in \Delta(S)$ to generate s_0.

Note: (i) for convenience, we assume that:

☆1: A is finite;

☆2: r is deterministic;

☆3: μ is one-point distribution at s_0.

Def [trajectory]: a trajectory τ_t at time t is the interaction record at time t,

i.e. $\tau_t = (s_0, a_0, r_0, \cdots, s_{t-1}, a_{t-1}, r_{t-1}, s_t)$

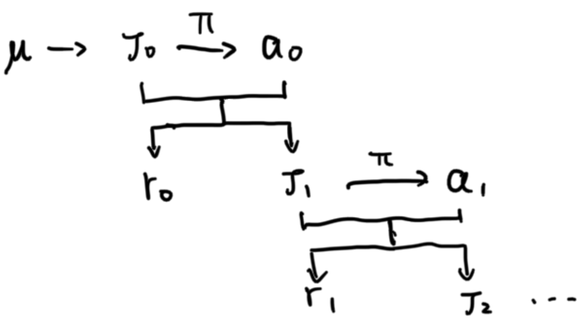
① $H := \{ \tau_t : t \geq 0 \}$

Def [policy]: a policy is such a mapping: $\pi: H \to \Delta(A)$

① a stationary policy: $\pi: S \to \Delta(A)$, only depends on s_t.

② a deterministic stationary policy: $\pi: S \to A$.

Note: MDP with policy π runs in such a flow:



Def [value function] $V_M^\pi : S \to \mathbb{R}$, π : policy , M: MDP.

$$V_M^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s \right]$$

Note: ① since $\gamma \in [0,1)$ and $r \in [0,1]$, $V_M^\pi \in [0, \frac{1}{1-\gamma}]$

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

$$Q_M^\pi(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a \right]$$

Note: ① Q_M^π is also bounded by $\frac{1}{1-\gamma}$.

② we ignore M when it is clear from context.

The goal of a MDP problem is to find a optimal policy π^* s.t.

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

Assertion : there exists a optimal deterministic and stationary policy π^*.

1.2. Bellman Consistency Equations for stationary policies.

Lemma 1.4 Suppose π : stationary policy , then we have

Bellman consistency equations $\Rightarrow$

$$V^\pi(s) = Q^\pi(s, \pi(s)) .$$

$$Q^\pi(s,a) = r(s,a) + \gamma E_{s' \sim P(\cdot|s,a)} [V^\pi(s')] .$$

pf: ① $Q^\pi(s, \pi(s)) = E_{a \sim \pi(\cdot|s)} [Q^\pi(s,a)]$

$$= E_{a \sim \pi(\cdot|s)} \left[E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a \right] \right]$$

rule of conditional expectation $= E\left[E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a \right] \,\middle|\, a \sim \pi(s) \right]$

$$= E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s, \pi \right]$$

$$= V^\pi(s)$$

② $Q^\pi(s,a) = E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a \right]$

$$= E\left[r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, \pi, s_0 = s, a_0 = a \right]$$

$$= r(s, a) + \gamma E\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) \,\Big|\, \pi, s_0 = s, a_0 = a \right]$$

$$= r(s, a) + \gamma E\left[E\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) \,\Big|\, s_1 = s' \right] \Big|\, s' \sim P(s,a) \right]$$

$$= r(s, a) + \gamma E_{s' \sim P(s,a)} [V^{\pi}(s')]$$

It is easy to see that if a is substituded by $\pi(s)$,

we have $Q^{\pi}(s, \pi(s)) = r(s, \pi(s)) + \gamma E_{\substack{a \sim \pi(s), \\ s' \sim P(s,a)}} [V^{\pi}(s')]$ $\square$

Notation: $P_{(s,a), s'} := P(s' | s, a)$

Def [transition matrix on (s, a) with a stational policy π]

$$P^{\pi}_{(s,a),(s',a')} := P(s'|s,a)\,\pi(a'|s')$$

Note: ① for deterministic policies:

$$P^{\pi}_{(s,a),(s',a')} := \begin{cases} P(s'|s,a), & \text{if } a' = \pi(s') \\ 0, & \text{else.} \end{cases}$$

With this notation, we have the following equations:

(H)
$$Q^{\pi} = r + \gamma P V^{\pi} ;$$
$$Q^{\pi} = r + \gamma P^{\pi} Q^{\pi} .$$

Pf: for a stational policy π.

according to Lemma 1.4, we have

$$Q^{\pi}(s, a) = r(s, a) + \gamma E_{s' \sim P(s,a)} [V^{\pi}(s')]$$
$$= r(s,a) + \gamma \sum_{s'} P_{(s,a),s'} V^{\pi}(s') \quad (1)$$

we can rewrite it as

$$Q^{\pi} = r + \gamma P V^{\pi}.$$

since Lemma 1.4, $V^\pi(s') = Q^\pi(s', \pi(s'))$

$$= \sum_{a'} \pi(a'|s') Q^\pi(s', a') \quad (2)$$

Introduce (2) into (1):

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} \sum_{a'} P_{(s,a),s'} \pi(a'|s') Q^\pi(s',a')$$
$$= r(s,a) + \gamma \sum_{(s',a')} P^\pi_{(s,a),(s',a')} Q^\pi(s',a').$$

We can rewrite it as

$$Q^\pi = r + \gamma P^\pi Q^\pi. \qquad \square$$

Corollary 1.5: Suppose π is stationary, we have

$$Q^\pi = (I - \gamma P^\pi)^{-1} r.$$

Pf: only need to show $I - \gamma P^\pi$ is invertible.

for any non-zero vector $x \in \mathbb{R}^{|S|\cdot|A|}$,

$$\|(I - \gamma P^\pi)x\|_\infty = \|x - \gamma P^\pi x\|_\infty$$
$$\geq \|x\|_\infty - \|\gamma P^\pi x\|_\infty$$
$$\geq \|x\|_\infty - \gamma\|x\|_\infty > 0 \quad (\text{since } \gamma < 1 \ \& \ \|x\|_\infty > 0)$$

which implies $I - \gamma P^\pi$ is full rank.

An intuitive way to see this is:

according to Gershgorin circle th.

since P^π is a stochastic matrix, the spectral radius of P^π should be 1, the the spectral radius of γP^π is strictly smaller than 1, which implies $\det(I - \gamma P^\pi) \neq 0$. $\square$

Lemma 1.6.

$$[(1-\gamma)(I - \gamma P^\pi)^{-1}]_{(s,a),(s',a')} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s', a_t = a' \mid s_0 = s, a_0 = a)$$

Pf: $(I-\gamma P^\pi)[I+\gamma P^\pi+\gamma^2(P^\pi)^2+\cdots+\gamma^t(P^\pi)^t]$

$= I-\gamma^{t+1}(P^\pi)^{t+1}$

since $\max (P^\pi)^{t+1} = \max P^\pi (P^\pi)^t \le \max (P^\pi)^t \le \cdots \le \max P^\pi$

and $\gamma \in [0,1)$,

when $t \to \infty$, we have:

① $\sum_{t=0}^{\infty} \gamma^t (P^\pi)^t$ converges

② $(I-\gamma P^\pi)(\sum_{t=0}^{\infty}\gamma^t(P^\pi)^t) = I$

② implies $(I-\gamma P^\pi)^{-1} = \sum_{t=0}^{\infty}\gamma^t (P^\pi)^t$.

On the other hand.

$$
\begin{aligned}
(P^\pi)^t_{(s,a),(s',a')} &= \sum_{(s_1,a_1)} P^\pi_{(s,a),(s_1,a_1)} (P^\pi)^{t-1}_{(s_1,a_1),(s',a')} \\
&= \cdots \\
&= \sum_{(s_{t-1},a_{t-1})} \cdots \sum_{(s_1,a_1)} P^\pi_{(s,a),(s_1,a_1)} \cdots P^\pi_{(s_{t-1},a_{t-1}),(s',a')} \\
&= \mathbb{P}^\pi(s_t=s', a_t=a' \mid s_0=s, a_0=a)
\end{aligned}
$$

then we have

$$[(I-\gamma P^\pi)^{-1}]_{(s,a),(s',a')} = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{P}^\pi(s_t=s', a_t=a' \mid s_0=s, a_0=a) \quad \square$$

1.3 Bellman Optimally Equations.

Now we proof the assertion at the beginning.

Theorem 1.7 Let Π be the set of all non-stationary and randomized policies. Define:

$$V^*(s) := \sup_{\pi\in\Pi} V^\pi(s)$$

$\Rightarrow$ V^*, Q^* are bounded by $\frac{1}{1-\gamma}$.

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

There exists a stationary and deterministic policy π s.t. for all $s \in S$ and $a \in A$,

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

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

Pf: let $(S_0, A_0, R_0, S_1, A_1, R_1, \cdots, S_{t-1}, A_{t-1}, R_{t-1}, S_t)$ denote a random $\mathcal{T}_t$. where S_i, A_j, R_k are r.v..

① First we show that

$$\sup_{\pi \in \Pi} E\left[\sum_{t=1}^{\infty} \gamma^t r(S_t, a_t) \,\middle|\, \pi, (S_0, A_0, R_0, S_1) = (s, a, r, s')\right] = \gamma V^*(s'). \quad (\text{✪})$$

For $\forall \pi \in \Pi$, define an "offset" policy $\pi_{(s,a,r)}$:

$$\pi_{(s,a,r)}(A_t = a \mid S_0 = s_0, A_0 = a_0, R_0 = r_0, \cdots, S_t = s_t)$$

$$:= \pi(A_{t+1} = a \mid S_0 = s, A_0 = a, R_0 = r, S_1 = s_0, A_1 = a, R_1 = r_0, \cdots, S_{t+1} = s_t)$$

By the Markov property:

$$\text{LHS of } (\text{✪}) = \sup_{\pi \in \Pi} \gamma E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi_{(s,a,r)}, S_0 = s'\right]$$

$$= \sup_{\pi \in \Pi} \gamma V^{\pi_{(s,a,r)}}(s')$$

we have for all (s,a,r), $\{\pi_{(s,a,r)} \mid \pi \in \Pi\} = \Pi$.

$\forall \pi \in \Pi$, let $\pi'_{(s,a,r)} = \pi$

and $\pi'(s)$ is one-point distribution.

then $\pi' \in \Pi \Rightarrow \{\pi_{(s,a,r)} \mid \pi \in \Pi\} \supset \Pi$

We have $\text{LHS of } (\text{✪}) = \gamma \cdot \sup_{\pi \in \Pi} V^{\pi}(s')$

$$= \gamma \cdot V^*(s')$$

$$= \text{RHS of } (\text{✪})$$

② we now show the deterministic and stationary policy $\tilde{\pi}$ is optimal:

$$\tilde{\pi}(s) = \arg\sup_{a \in A} E\left[r(s,a) + \gamma V^*(S_1) \,\middle|\, (S_0, A_0) = (s,a) \right]$$

For this, we have

$$V^*(s_0) = \sup_{\pi \in \Pi} E\left[r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(S_t, a_t) \,\middle|\, \pi, S_0 = s_0 \right]$$

$$= \sup_{\pi \in \Pi} E\left[r(s_0, a_0) + E\left[\sum_{t=1}^{\infty} \gamma^t r(S_t, a_t) \,\middle|\, \pi, \begin{matrix} (S_0, A_0, R_0, S_1) \\ = (s_0, a_0, r_0, s_1) \end{matrix} \right] \right]$$

$$\leq \sup_{\pi \in \Pi} E\left[r(s_0, a_0) + \sup_{\pi' \in \Pi} \left[\sum_{t=1}^{\infty} \gamma^t r(S_t, a_t) \,\middle|\, \pi', \begin{matrix} (S_0, A_0, R_0, S_1) \\ = (s_0, a_0, r_0, s_1) \end{matrix} \right] \right]$$

$$= \sup_{\pi \in \Pi} E\left[r(s_0, a_0) + \gamma V^*(s_1) \right]$$

$$= \sup_{a_0 \in A} E\left[r(s_0, a_0) + \gamma V^*(s_1) \right]$$

$$= E\left[r(s_0, a_0) + \gamma V^*(s_1) \,\middle|\, \tilde{\pi} \right]$$

By applying the argument recursively:

$$V^*(s_0) \leq E\left[r(s_0, a_0) + \gamma V^*(s_1) \,\middle|\, \tilde{\pi} \right]$$

$$\leq E\left[r(s_0, a_0) + \gamma E\left[r(s_1, a_1) + \gamma V^*(s_2) \,\middle|\, \tilde{\pi} \right] \,\middle|\, \tilde{\pi} \right]$$

$$= E\left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 V^*(s_2) \,\middle|\, \tilde{\pi} \right]$$

$$\leq \cdots$$

$$\leq V^{\tilde{\pi}}(s_0)$$

since $V^{\tilde{\pi}}(s_0) \leq V^*(s_0)$ by definition.

we have $V^{\tilde{\pi}} = V^*$ □

Notation ①: $\pi_Q := \arg\max_{a \in A} Q(s,a)$

②: $V_Q(s) := \max_{a \in A} Q(s,a)$

Theorem 1.8 [Bellman optimality equations , BOE]

A vector $Q \in \mathbb{R}^{|S||A|}$ satisfies the BOE if

$$Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [\max_{a' \in A} Q(s',a')] .$$

Then for any $Q \in \mathbb{R}^{|S||A|}$, $Q = Q^*$ $\Leftrightarrow$ Q satisfies BOE.

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

By the notation: the optimal policy $\pi^* = \pi_{Q^*}$

$$\mathcal{T}_M : \mathbb{R}^{|S||A|} \to \mathbb{R}^{|S||A|} , \quad \mathcal{T} Q := r + \gamma P V_Q$$

Bellman optimality operator

we can write BOE: $Q = \mathcal{T} Q$. so the theorem states that

$Q = Q^* \Leftrightarrow Q$ is a fixed point of $\mathcal{T}$.

Pf: ① First we want to show $V^*(s) = \max_a Q^*(s,a)$ (A)

Let π^* be an optimal stationary and deterministic policy

$V^*(s) = \sup_{\pi \in \Pi} V^\pi(s)$, consider such a policy π': take action a

then follow π^*. $\pi' \in \Pi$

$$\Rightarrow V^*(s) \geq V^{\pi'}(s) = Q^{\pi^*}(s,a) \overset{\text{Th1.7}}{=\!=} Q^*(s,a) \quad (\forall a \in A)$$

$$\Rightarrow V^*(s) \geq \max_a Q^*(s,a)$$

On the other hand, by Th 1.7 & Lem 1.4 (since π^* is stationary)

$$\begin{aligned} V^*(s) = V^{\pi^*}(s) &= Q^{\pi^*}(s, \pi^*(s)) \\ &\leq \max_a Q^{\pi^*}(s,a) \\ &= \max_a Q^*(s,a) \end{aligned}$$

we prove (A) holds.

② sufficiency : Q^* satisfies $Q^* = \mathcal{T} Q^*$

for all actions $a \in A$:

$$Q^*(s,a) = Q^{\pi^*}(s,a)$$
$$= r(s,a) + \gamma E_{s'\sim P(s,a)}[V^{\pi^*}(s')]$$
$$= r(s,a) + \gamma E_{s'\sim P(s,a)}[V^*(s')]$$
$$= r(s,a) + \gamma E_{s'\sim P(s,a)}[\max_{a'} Q^*(s',a')]$$
$$= \mathcal{T} Q^* .$$

③ necessity: assume $Q = \mathcal{T}Q$, we show that $Q = Q^*$

Let $\pi = \pi_Q$. $Q = \mathcal{T}Q = r + \gamma P V_Q$

since $\pi = \pi_Q = \arg\max_a Q(s,a)$ is stationary and deterministic

$$Q(s,a) = r(s,a) + \gamma \sum_{(s',a')} P_{(s,a),s'} \pi(a'|s') Q(s',a')$$

$$\Rightarrow Q = r + \gamma P^\pi Q$$

$$\Rightarrow Q = (I - \gamma P^\pi)^{-1} r \xlongequal{\text{Cor 1.5}} Q^\pi$$

i.e. Q is action value of π_Q.

$\forall$ stationary and deterministic π', we have

$$[(P^\pi - P^{\pi'})Q^\pi]_{s,a} = E_{s'\sim P(\cdot|s,a)}[Q^\pi(s',\pi(s)) - Q^\pi(s',\pi'(s))] \geq 0$$

$$\Rightarrow Q - Q^{\pi'} = Q^\pi - Q^{\pi'}$$
$$= Q^\pi - (I - \gamma P^{\pi'})^{-1} r$$
$$= (I - \gamma P^{\pi'})^{-1}[(I - \gamma P^{\pi'}) - (I - \gamma P^\pi)] Q^\pi$$
$$= \gamma (I - \gamma P^{\pi'})^{-1}(P^\pi - P^{\pi'}) Q^\pi$$
$$\geq 0$$

By Lem 1.6 $\gamma(I - \gamma P^{\pi'})^{-1} \geq 0$.

Specifically, $Q^\pi = Q \geq Q^{\pi^*} = Q^* \geq Q^\pi$

$$\Rightarrow Q = Q^\pi = Q^* \quad \square$$