1 Markov Decision Process

state(S)-policy(pi)-action(A)-model(p)-state(S')

reward(R') from transition

return(G) from episode

value(V+Q)

$$p\left(s^{\prime}, r \mid s, a\right)=\operatorname{Pr}\left\{S_t=s^{\prime}, R_t=r \mid S_{t-1}=s, A_{t-1}=a\right\}$$

$$\sum_{s^{\prime} \in S} \sum_{r \in R} p\left(s^{\prime}, r \mid s, a\right)=1$$

$$p\left(s^{\prime} \mid s, a\right)=\sum_{r \in R} p\left(s^{\prime}, r \mid s, a\right)$$

$$r(s, a)=\sum_{s^{\prime} \in S} \sum_{r \in R}\left(p\left(s^{\prime}, r \mid s, a\right) * r\right)$$

2 Bellman Equations

描述状态之间的静态关系

$$
\begin{aligned}
v_\pi(s) &= E_\pi\left[G_t \mid S_t = s\right] \\
&= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right] \\
&= \sum_{a \in A} \pi(a \mid s) \sum_{s', r} p\left(s', r \mid s, a\right) * \left[r + \gamma E_\pi\left[G_{t+1} \mid S_{t+1} = s'\right]\right] \\
&= \sum_{a \in A} \pi(a \mid s) \sum_{s', r} p\left(s', r \mid s, a\right) * \left[r + \gamma v_\pi\left(s'\right)\right] \\
&= \sum_{a \in A} \left(\pi(a \mid s) * q_\pi(s, a)\right)
\end{aligned}
$$

$$
\begin{aligned}
q_\pi(s, a) &= E_\pi\left[G_t \mid S_t = s, A_t = a\right] \\
&= E_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] \\
&= \sum_{s', r} p\left(s', r \mid s, a\right) * \left[r + \gamma E_\pi\left[G_{t+1} \mid S_{t+1} = s'\right]\right] \\
&= \sum_{s', r} p\left(s', r \mid s, a\right) * \left[r + \gamma v_\pi\left(s'\right)\right] \\
&= \sum_{s', r} p\left(s', r \mid s, a\right) * \left[r + \gamma \sum_{a' \in A} \left(\pi\left(a' \mid s'\right) * q_\pi\left(s', a'\right)\right)\right]
\end{aligned}
$$

policy-comparison:

$$\pi' \geq \pi \quad \longleftrightarrow \quad v_{\pi'}(s) \geq v_\pi(s) \quad \forall s \in S$$

policy-improvement:

$$E_{\pi'}\left[q_\pi\left(s, \pi'(s)\right)\right] \geq v_\pi(s) = E_\pi\left[q_\pi\left(s, \pi(s)\right)\right] \quad \forall s \in S$$

optimal-policy:

$$v_*(s) = \max_\pi v_\pi(s) = \max_{a \in A} q_{\pi *}(s, a) \quad \forall s \in S$$

3 Dynamic Programming

(1) 基于模型 p 的策略 pi 迭代（策略估计、策略改进）：

Policy Evaluation: (matrix solution vs. iteration solution)

$$
\begin{aligned}
v_{k+1}(s) &= E_\pi\left[R_{t+1}+\gamma v_k\left(S_{t+1}\right) \mid S_t=s\right] \\
&= \sum_a \pi(a \mid s) \sum_{s', r} p\left(s', r \mid s, a\right)\left[r+\gamma v_k\left(s'\right)\right]
\end{aligned}
$$

Policy Improvement: (greedy)

$$
\pi'(s) = \underset{a}{\arg\max}\ q_\pi(s, a)
$$

(2) 基于模型 p 的价值 v 迭代:

Value Evaluation + Policy Improvement:

$$
v_{k+1}(s) = \max_a \mathbb{E}\left[R_{t+1}+\gamma v_k\left(S_{t+1}\right) \mid S_t=s, A_t=a\right]
$$

4 Monte Carlo