1 Markov Decision Process

Concepts:
state(S)-policy(pi)-action(A)-model(p)-state(S')
reward(R') from transition
return(G) from episode
value(V+Q)

$$p\left(s', r \mid s, a\right) = \Pr\left\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right\}$$

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

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

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

2 Bellman Equations

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

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

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 (model p)

Policy Evaluation: (matrix solution vs. iteration solution)

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

Policy Improvement:

$$\pi'(s) = \arg\max_a q_\pi(s, a)$$

Value Iteration:

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

4 Monte Carlo