

By definition:

$$Q^\pi(s, a) = E_{\gamma^\pi} \left[\sum_{t \leq T} \gamma^t r(s_t, a_t) \mid S_0 = s, A_0 = a \right]$$

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

$$Q^\pi(s, a) = E_{\gamma^\pi} \left[\sum_{0 \leq t \leq T} \gamma^t r(s_t, a_t) \mid S_0 = s, A_0 = a \right]$$

$$= E_{\gamma^\pi} \left[r(s, a) + \gamma \sum_{0 \leq t \leq T} \gamma^t r(s_{t+1}, a_{t+1}) \mid S_0 = s, A_0 = a \right]$$

$$= E_{\gamma^\pi} \left[r(s, a) + \gamma E \left[\sum_{0 \leq t \leq T} \gamma^t r(s_{t+1}, a_{t+1}) \mid S_{t+1}, A_{t+1}, S_0 = s, A_0 = a \right] \mid S_0 = s, A_0 = a \right]$$

As the processes $(S_t, A_t)_{t \geq 0}$ is a Markovian, thus the following holds:

$$Q^\pi(s, a) = E_{\gamma^\pi} \left[r(s, a) + \gamma Q^\pi(s_{t+1}, A_{t+1}) \mid S_0 = s, A_0 = a \right]$$

Show that $Q^*(s, a) = E_{(s', a') \sim \pi^*(\cdot | s, a)} [r(s, a) + \gamma \max_a Q^*(s', a')]$

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

$$= \max_{\pi} E_{\gamma^\pi} \left[r(s, a) + \gamma Q^\pi(s_{t+1}, A_{t+1}) \mid S_0 = s, A_0 = a \right]$$

By definition of π^* the optimal policy:

$$Q^*(s, a) = E_{\gamma^{\pi^*}} \left[r(s, a) + \gamma \max_{\pi} Q^\pi(s_{t+1}, A_{t+1}) \mid S_0 = s, A_0 = a \right]$$

$$= E_{\gamma^{\pi^*}} \left[r(s, a) + \gamma \max_{a'} Q^*(s_{t+1}, a') \mid S_0 = s, A_0 = a \right]$$