1. **Recursive Bellman equations**

   Prove the recursive Bellman expectation equations for the value function v_π and the action value function q_π using the state transition function $\mathcal{P}$ and the reward function $\mathcal{R}$. You are allowed to use the equations from Theorem 1 in Section 4.

   (a) $v_\pi(s) = \sum_a \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v_\pi(s') \right)$

   (b) $q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a')$

(a) $v_\pi(s) = E[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$

$$= \sum_a \pi(a|s)(R(s,a) + \gamma v_\pi(S_{t+1}))$$

$$= \sum_a \pi(a|s)(R(s,a) + \gamma \sum_{s'} P(s'|s,a) v_\pi(s'))$$

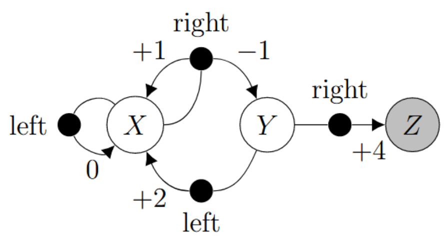
(b) $q_\pi(s,a) = E[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a]$

$$= R(s,a) + \gamma \underbrace{\sum_{a'} \sum_{s'} P(s'|s,a)\, \pi\, q_\pi(s',a')}_{\text{I don't get it.}}$$

2. **Action value functions**

   (a) For any given MDP, policy π, *terminal state* E and action a, what is $q_\pi(E,a)$? All transitions from a terminal state are back to itself with a reward of 0.

   (b) Consider the MDP and policy π_1 from the previous exercise sets. Note that if action *right* is taken in state X, then the transitions to X and Y occur with probabilities 0.75 and 0.25, respectively. The deterministic policy π_1 is defined as $\pi_1(X) = \text{right}$, $\pi_1(Y) = \text{right}$.



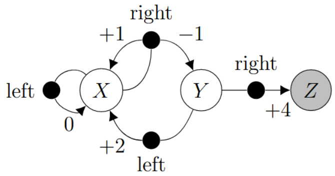
(a) $v_\pi(E) = 0$

$$q_\pi(E,a) = R(E,a) + \gamma \sum_{s'} P(s'|s,a) v_\pi(s')$$

$$= \underbrace{R(E,a)}_{0} + \gamma \underbrace{v_\pi(E)}_{0} = 0$$

(b)

(b) Consider the MDP and policy π_1 from the previous exercise sets. Note that if action *right* is taken in state X, then the transitions to X and Y occur with probabilities 0.75 and 0.25, respectively. The deterministic policy π_1 is defined as $\pi_1(X) = \text{right}$, $\pi_1(Y) = \text{right}$.



Compute the action value of state X and action *left* under policy π_1, i.e. $q_{\pi_1}(X, \text{left})$, using *only* the action value function (don't use the values from the last exercise set). The discount factor is $\gamma = 0.9$.

$$q_{\pi_1}(X, left) = R(X, left) + \gamma \sum_{s'} P(s'|s,a) \sum_{a'} \pi(a'|s') q_{\pi_1}(s',a')$$

$$= \underbrace{R(X, left)}_{0} + \gamma q_{\pi_1}(X, right)$$

It doesn't care where it comes from

$$= \gamma q_{\pi_1}(X, right)$$

$$q_{\pi_1}(X, right) = R(X, right) + \gamma (P(X|X, right) q_{\pi_1}(X, right) + P(Y|X, right) q_{\pi_1}(X, right))$$

$$R(X, right) = (+1) \cdot P(X, +1 | X, right) + (-1) \cdot P(Y, -1 | X, right)$$

$$= 0.75 - 0.25 = 0.5$$

$$q_{\pi_1}(X, right) = 0.5 + \gamma q_{\pi_1}(X, right) \qquad q_{\pi_1}(X, right) = 5$$

then $q_{\pi_1}(X, left) = \gamma q_{\pi_1}(X, right) = 0.9 \cdot 0.5 = 0.45$

(c) In the lecture we defined the policy iteration algorithm to find the optimal policy using value functions. Write down a modified version of policy iteration that finds the optimal policy using action value functions (known as Q-Policy iteration).

pol eval:

$$q_{k+1}(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) \sum_{a'} \pi(a'|s') q_k(s',a')$$

pol improve:

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

$$v_{k+1}(s) = \sum_a \pi(a|s) q_k(s,a)$$

3. **Value iteration**

(a) Perform two steps of value iteration for the MDP from exercise 2 (b), i.e. calculate $v_1(s)$ and $v_2(s)$ for $s \in \{X, Y\}$. Initialize the values with $v_0(X) = 0$ and $v_0(Y) = 0$. You can assume that the value of the terminal state Z is zero in each step.

$k=0$: $\quad v_0(X) = v_0(Y) = v_0(Z) = 0$

$v_1(X) = \max_a (R(X,a)) = R(X, \text{right}) = 0.5$

$v_1(Y) = \max_a (R(Y,a)) = R(Y, \text{right}) = 4$

$k=1$ $\quad v_1(X) = 0.5,\ v_1(Y) = 4,\ v_1(Z) = 0$

$v_2(X) = \max_a (R(X,a) + \gamma P(X|X,a) v_1(X) + \gamma P(Y|X,a) v_1(Y))$

$= \max_a (R(X,a) + 0.45 P(X|X,a) + 0.81 P(Y|X,a))$

$a = \text{left}$: $\quad 0 + 0.45 = 0.45$

$a = \text{right}$: $\quad 0.5 + 0.45 \cdot 0.75 + 0.81 \cdot 0.25 = 1.04$, pick right

$v_2(X) = 1.04$

$v_2(Y) = \max_a (R(Y,a) + \gamma P(X|Y,a) v_1(X) + \underbrace{\gamma P(Z|Y,a) v_1(Z)}_{0}))$

$a = \text{left}$: $\quad 2 + 0.9 \times 0.5 = 2 + 0.45 = 2.45$

$a = \text{right}$: $\quad 4$ $\qquad$, pick right.