Jason Park
ECE 473
HW8 Written

Problem 1

$$V^{*(i+1)}(s) = max_a \Sigma_{s'} T(s,a,s')[R(s,a,s') + \gamma V^{*i}]$$

Sample calculation for $V^{*(1)}(0)$:

Action = -1

$$\Sigma_{s'} T(0,-1,s')[R(0,-1,s') + \gamma V^{*(0)}(s')]$$

$$T(0,-1,-1)\big[R(0,-1,-1) + \gamma V^{*(0)}(-1)\big] + T(0,-1,\ 1)[R(0,-1,1) + \gamma V^{*(0)}(1)]$$

(0.8)[-5 + 0]+(0.7)[-5 + 0]

Action = +1

$$\Sigma_{s'} T(0,+1,s')[R(0,+1,s') + \gamma V^{*(0)}(s')]$$

$$T(0,+1,-1)\big[R(0,+1,-1) + \gamma V^{*(0)}(-1)\big] + T(0,+1,\ 1)[R(0,+1,1) + \gamma V^{*(0)}(1)]$$

(0.2)[-5 + 0] + (0.3)[-5 + 0]

Max(-7.5, -2.5)

$V^{*(1)}(0) = -2.5$

Repeating for i ϵ [0, 1, 2] and s ϵ [-2, -1, 0, 1, 2]

$V^{*(0)}(s)$ = [0, 0, 0, 0, 0]

$V^{*(1)}(s)$ = [0, 12.5, -2.5, 66, 0]

$V^{*(2)}(s)$ = [0, 10.75, 50.7, 64, 0]

Problem 2

a. It is not always the case that $V_1(S_{start}) >= V_2(S_{start})$. For example, if the randomness decided to randomly move to a state with a very high reward, this would mean that $V_2(S_{start})$ would actually be > $V_1(S_{start})$.
b. An algorithm that would allow us to compute V* for each node with only a single pass over all the (s, a, s') triples would be a Topological Sort. We are able to use this algorithm because value iteration carries the properties of being directed and acyclic. This would allow us to compute the values and "direction" for each possible state, starting with the start state.
c. Since we are essentially decreasing the discount by an amount 0 < amount < 1. In order to keep the optimal values equal for all possible states, we must increase (by multiplication) the reward for each state by the same amount we have removed from the discount. To put it in more of an analogy from Piazza, with no discount, it is like we are going through life with a 100% probability that we will continue living the next day. Because of this, we are constantly reaping rewards with every day. However, if we decreased the probability that we will continue living the next day, it means that we are not 100% certain that we will continue living the next day. In order to keep the same overall value we get from each day, we must increase the rewards we receive per day, since value is a function of probability and reward.
(If needed, the next page has a more formal explanation)

Jason Park
ECE 473
HW8 Written

$$T'(s, a, s') = \begin{cases} T(s, a, s')(1-\gamma) & \text{If } s \text{ is o (new state)} \\ T(s, a, s') & \text{Else} \end{cases}$$

$$R'(s, a, s') = \begin{cases} R(s, a, s')(1-\gamma) & \text{If } s \text{ is o (new state)} \\ R(s, a, s') & \text{Else} \end{cases}$$