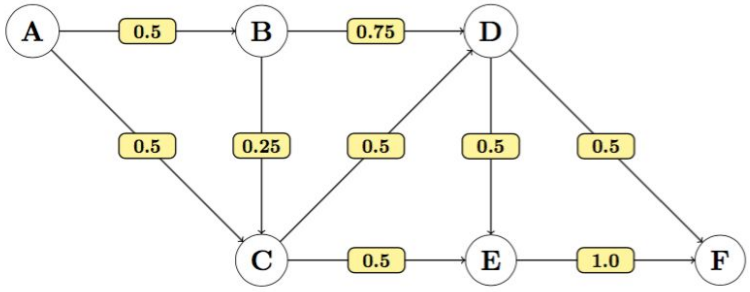
PROBLEM 11

Problem 7 The following graph specifies the states and transition probabilities for a Markov Decision Process (MDP). There are only two actions in this MDP: $a = S$ (stay) or $a = M$ (move). If you elect to stay, you remain in the same state with probability one. If you move, you change states according to the probabilities specified below. F is the only terminal state where you don't move even if you select $a = M$.



The rewards in this MDP are associated with state transitions such that

$R(B \to D) = -1$, $R(D \to F) = +10$, $R(A \to C) = -2$,

and all the remaining rewards are zero. The discount factor is $\gamma = 0.5$.

(7.1) **(4 points)** Suppose we initialize the values as

$V_0(A) = -1$, $V_0(B) = 2$, $V_0(C) = 1$, $V_0(D) = 1$, $V_0(E) = 0$, $V_0(F) = 3$

What would be the resulting action to take in states A and C?

For state A: Move (M)
For state C: Stay (S)

Explanation: If a state has a negative initial value it is advantageous to move to a state with a positive value in order to maximize reward (i.e. can move from A to B or C). Conversely, if a state has a positive initial value such as state C, then it is okay to stay because the payoff is already positive.

(7.2) **(3 points)** Calculate $V_1(C)$ after one value iteration.

$V_1(C) = \max_{a \in A} \sum_{s' \in S} T(s, a, s')[R(s, a, s') + \gamma V_0(s')]$
$a = S \implies$ the equation evaluates to: .5
$a = M \implies$ the equation evaluates to: .25
Therefore, $V_1(C) = 0.5$.

(7.3) **(2 points)** Suppose we perform value iteration until convergence obtaining $V^*(s)$, $s = A, B, C, D, E, F$. What is the resulting $V^*(F)$? **(0)**

(7.4) **(3 points)** Are we guaranteed to get the cumulative discounted reward equal to $V^*(D)$ if we begin in state D and act optimally according to the converged values? (Y/N) **(N)**

Additional explanation:

11) a) Move if in A, stay if in C. If a state has a negative initial value it is advantageous to move to a state with positive value to increase reward, so we would move from A to B or C. If a state has positive initial value, like C, then it will stay because the payoff is already positive.

b) $V_1(C) = max_{a \in A} \sum_{s' \in S} T(s, a, s')[R(s, a, s') + \gamma V_0(s')]$
$a = S$: equation evaluates to 0.5
$a = M$: equation evaluates to 0.25
so $V_1(C) = 0.5$

c) Note that F is a terminal state, so we cannot transition out of it to get any reward. If we perform value iteration until convergence to find $V^*(s)$ for all s, we get that $V^*(F) = 0$.

d) We are not guaranteed to get the cumulative discounted reward equal to $V^*(D)$ if we begin in state D and act optimally according to the converged values because the optimal value function represents an expectation, not a guarantee. In other words, if you could reset the agent to state D and repeatedly run the simulation, then your empirical average cumulative discounted reward would converge to $V^*(D)$. This does not guarantee that any single run will give you that amount of discounted reward.