Markov Decision Processes Exercises

1 Hacked Robot

Assume a robot is acting in a given MDP $M = (S, A, T, R, \gamma, s_0)$, where S is the discrete set of states, A is the discrete set of actions, T is the transition functions which maps current state, action, next state tuples to their probabilities, R is a reward function which maps current state, action, next state tuples to the reward, γ is a discount factor and s_0 the start state.

Question 1.1: For any particular state s and action a, what does the following expression evaluate to?

$$\sum_{s' \in S} T(s, a, s') =$$

The robot unfortunately gets hacked by a malicious virus which affects how the robot acts as follows: at each time step, with probability β the virus does not affect the robot and the robot is free to act according to its policy. With probability $1 - \beta$, however, the virus overrides the robot's decision and instead, makes the robot perform an action at random (unknown to the robot), sampled uniformly from available actions.

Question 1.2: Can you specify a modified MDP $M' = (S', A', T', R', \gamma', s'_0)$ for which the optimal policy maximizes the expected discounted sum of rewards under the specified restrictions on your ability to choose actions? (Hint: not all components of the MDP will need to change)

$S' =$

$A' =$

$T' =$

$R' =$

$\gamma' =$

$s_0 =$

Question 1.3 Let $V_\pi(s)$ be the value function for a deterministic optimal policy π in M and let $V_{\pi'}(s)$ be the value function for a deterministic optimal policy π' in M'.

Are π and π' identical? Why or why not? (Recall that a policy is a mapping from current states to actions)

Question 1.4 Next, consider a class of MDPs in which N grid cells ($N > 3$) are aligned in a corridor (e.g., going from left to right), where the agent starts at the left-most cell, can move either left or right (if a free cell is available in that direction), and received a fixed reward signal (e.g., +10) upon reaching the right-most cell, at which point the episode ends.

Assuming $\beta = 0.5$ and $\gamma = 0.9$, are there states s for which $V_\pi(s) > V_{\pi'}(s)$? Are there states for which $V_\pi(s) = V_{\pi'}(s)$? Explain or construct a simple example (e.g., 3 state corridor MDP) in which you compute the values for the two different MDPs.

Finally, what must we assume about the corridor-style MDPs M and M' in order for the following equality to hold for all states s, $V_\pi(s) = V_{\pi'}(s)$? (Hint: the assumption deals with one or more of the components of an MDP as defined in the beginning of the exercise)