Grade 11/12 Math Circles

April 5, 2023

Reinforcement Learning 3 - Solutions

Problem Set 3 Solutions

1. A Lunar Roving Vehicle (LRV) is a battery operated four wheeled rover used on Moon to assist astronauts in their study of the lunar surface. The LRV is typically powered by solar energy. To tap solar energy into its panels, the LRV has to get on top of a hill. But the rover has a tendency to roll downhill and then it needs energy to ride uphill. Specifically, the LRV could be in three situations, namely, at the bottom of the hill, top of the hill or rolling down the hill. In each situation, the LRV could either drive or not drive. If the LRV is at the top of the hill and it is driving, at the next time instant, it could remain at the top of the hill with 0.6 probability or start rolling down the hill with 0.4 probability. If at the top of the hill and the LRV doesn't drive, these probabilities are 0.8 and 0.2 respectively. If the LRV is rolling down and drives, it will roll down with probability 0.3, end at hill top with probability 0.5 or end at the bottom with 0.2 probability in the next time period. If rolling down and doesn't drive, with probability 0.8 it will end at bottom and with 0.2 it will roll down in the next time period. Finally, if the LRV is at bottom hill and drives, it will roll down with probability 0.4 and end up hill with probability 0.6 in the next time step. If at the bottom of the hill and doesn't drive, the LRV will be at bottom of the hill with probability 1. Further, driving always consumes 1 unit of energy. Also, the rover absorbs one, two and three units of energy while at the bottom, rolling down the hill and at top of the hill respectively. For example, if the LRV is at the top of a hill and driving the total reward for the LRV is, 3-1 = 2.

   (a) Provide a graphical representation of the MDP.

   (b) Suggest a deterministic and stochastic policy for the MDP (It need not be optimal).

   (c) For each policy suggested, write down the induced transition matrix and a corresponding Markov chain trajectory.

Solution:

(a) See Fig. 1.

(b) There can be several answers to this part.

(c) The solution to this part will vary according to your solution to part b.

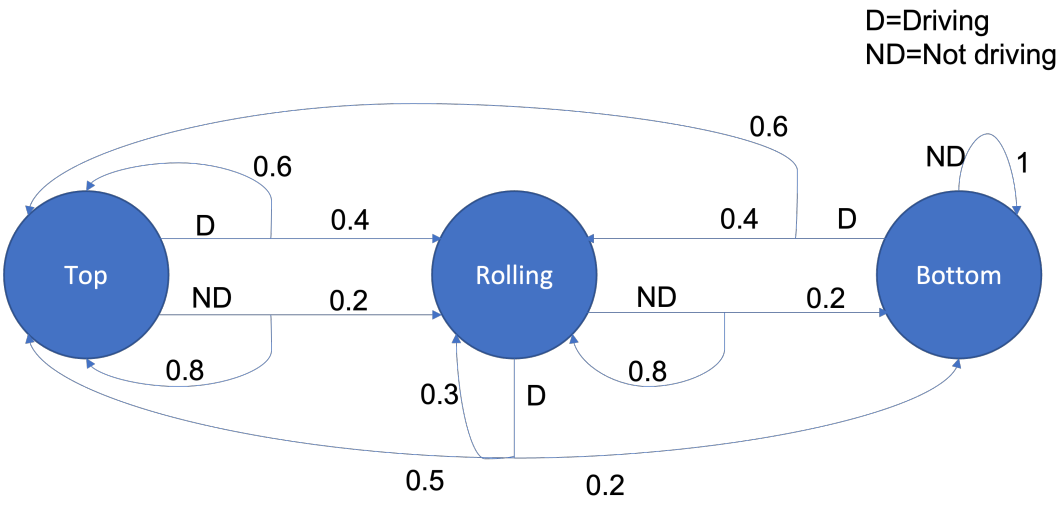

Figure 1: Graphical representation of a MDP

2. Consider the grid depicted in Fig.2.

   If $R_t = -1$ on all transitions and you want to reach the goal state in as minimum plays as possible, suggest an optimal policy to achieve your objective. How many optimal policies are there?

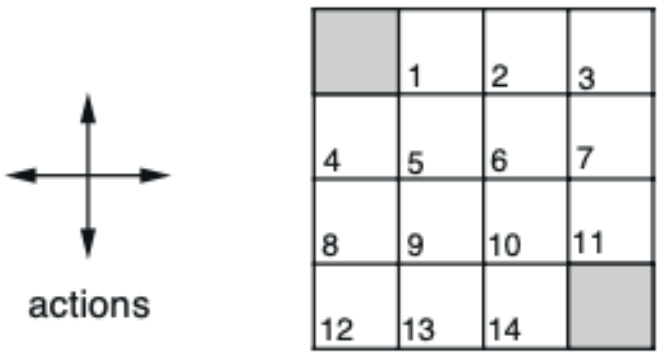

Figure 2: Navigation grid

Solution: There are infinitely many optimal policies. One example of an optimal policy is given in Fig.3.

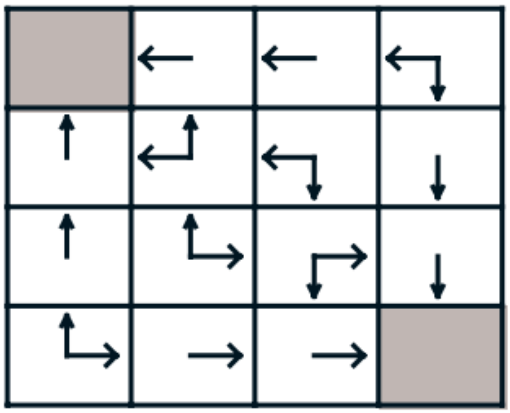

Figure 3: Optimal policy for navigation grid