CHENNAI MATHEMATICAL INSTITUTE

Reinforcement learning

Assignment 2.
Date: March 5, 2021. Due date: Mar 18, 2021.

(1) Consider a 6x6 grid with states (i, j), $1 \leq i, j \leq 6$. Suppose states $(4, 3)$ and $(5, 3)$ are removed (you can think of these as holes). At each square you are allowed the following actions attempt left, attempt right, attempt up and attempt down. When you attempt an action, you can succeed with probability 0.8. With probability 0.1 you stay where you are. With probability 0.05 you move in a direction $+90^\circ$ (closkwise) to the direction attempted, and -90° to the direction attempted. If ever an attempted move takes you out of the grid or takes you into a hole, you remain where are. Also assume that if your actions result in visiting $(6, 3)$ you get a reward of -15 and if your action results in visiting $(6, 6)$ you get +15. $(6, 6)$ is a terminal state and once you visit that state the episode ends. Assume discount parameter is 0.9
   a Write a program to compute the value function and Q function for the policy which selects each action with probability 0.25.
   b Write a program which computes the optimal policy by policy iteration. Start with the above random policy and iterate. Run the algorithm for 200 episodes and for 200 trials. For each trial keep track of the value of state $(1, 1)$ in each episode and plot the value of this averaged over the 200 trials, as a function of the episode number. Plot the optimal action from each state on the 6x6 grid, using letters L, R, U, D to indicate the optimal action. If there are multiple optimal actions, write all the letters.
   c Next write a program which computes something close to the optimal policy using cross entropy method (this will be described soon in class) using 6x4=24 parameters. You may need to optimize hyperparameters to get convergence. Describe how you selected hyperparameters. For the best hyperparameters, in each trial you should run about 300 episodes. So each trial will result in a return after the first episode, the second episode and till the 300th episode. You run this for about 500 trials and plot a curve of the average return as a function of the number of episodes.

(2) Repeat part a using Monte Carlo simulations. To simulate a random coin for say action attempt up, do the following: Assume that the interval $[0, 1]$ is divided into $[0, 0.05], [0, 05, 0, 1], [0, 1, 0.2], [0.2, 1]$ corresponding to actually performing the action move right $(+90^\circ)$, move left (-90°), stay, and go up. Now pick a random number in

the interval $[0, 1]$ and perform the action determined by which of the 4 intervals the random number selected lies in. Do both first visit MC and every visit with updates as suggested in class. Run 200 episodes of Monte simulations, and 200 trials. And plot the value of the states $(1, 1)$ as a function of the number of episodes. Does it converge to the value computed in part a above for state $(1, 1)$.