CSE 6369 - *Reinforcement Learning*

Homework 1- Spring 2020

Due Date: Feb. 11 2020

N-Armed Bandits

1. Consider the following (unrealistic) elevator scheduling problem. In a building there is a single elevator that moves between the 6 floors of the building. The time it takes the elevator to travel one floor is 7s and the time to get a person into and off the elevator is also 7s. The elevator is very lightly used and in particular, it never happens that two persons call the elevator any closer than 100s apart (i.e. the elevator has sufficient time to move from any floor to the person, load the person, bring them to the exit, let them out, and move to an arbitrary floor before the next person presses the button).

   At a particular time of the day, people call the elevator from "call-floors", S_c, with a particular probability distributions, $P_t(S_c)$, and want to go to "exit floors", S_e, with a particular probability distribution, $P_t(S_e|S_c)$.

   a) Assuming that the only important factor is the time a person has to wait until they arrive at their destination (i.e. that the energy the elevator uses is not relevant and humans do not distinguish between waiting inside and outside of the elevator) and that the "penalty" of waiting increases linearly with the wait time, design a utility function for the problem as a function of the starting floor of the elevator and the probability distribution that persons will call the elevator from a particular floor, $P(S_c)$ and want to exit at any other floor, $P(S_c|S_e)$.

   b) Formulate an n-armed bandit problem to determine the ideal (fixed) starting floor for the elevator. Indicate all the parts of your formulation of the n-armed bandit problem.

   c) Solve your n-armed bandit problem from part b) for the following cases: i) in the early morning all persons come in on the first floor and exit on each of the other floors (2-6) with uniform probability; ii) at noon people call the elevator from floors 2-6 with a uniform distribution and all exit on the first floor, and iii) in the afternoon, 50% of the time the elevator is called from the second floor where people want to go to the other floors with a uniform distribution while the other 50% of the time the elevator is called from floors 2-6 with a uniform distribution with persons always exiting on the 1st floor.

   d) Implement an incremental averaging solution to your n-armed bandit for the (fixed starting floor) elevator scheduling problem for arbitrary "call-floor" and "exit floor" probability distributions and use it to determine the solution for the following situations: persons call the elevator from each of the floors with a uniform probability and are going to travel to other floors with a probability distribution where when going down the probability to travel to each of the other floors is proportional to the number of floors that they cross (i.e. that, e.g., from the fourth floor it is twice as likely to want to go to the second foor - moving 2 floors down - than to want to go to floor 3 - moving down by 1 floor), and when going up the probability is equal to the probability of moving 5 floors down for all the floors above. This represents the situation where people sometimes take the stairs for short trips down but not up.

      Make sure to submit the code as well as a graph showing the estimated utilities for the actions in your n-armed bandit at each iteration (i.e. after each person used the elevator) for the first 500 iterations. You have to implement your n-armed bandit from scratch (you can use math libraries if you want) and can not use any pre-built n-armed bandit libraries and functions.

e) Use your learning implementation from part d) to solve the same problem for 3 more distributions of your own choice and discuss the behavior of the n-armed bandit learner.

2. In class we saw that the utility of money does not behave linearly in the amount of money. Consider that this is also the case for the utility of waiting time and assume that the perceived ”penalty” of waiting instead behaves quadratic in terms of the wait time (i.e. additional wait time becomes increasingly bad over time).

   a) Assuming the same elevator scheduling problem as in Question 1 and the quadratic ”penalty” for waiting, design a new utility function for the problem as a function of the starting floor of the elevator and the probability distribution that persons will call the elevator from a particular floor, $P(S_c)$ and want to exit at any other floor, $P(S_c|S_e)$.

   b) Modify your n-armed bandit code to use this utility function and re-run the experiments from parts 1 d) and 1 e). Show the results and discuss the differences in behavior of the elevator.

      Again, make sure to submit the code as well as a graph showing the estimated utilities for the actions in your n-armed bandit at each iteration (i.e. after each person used the elevator) for the first 500 iterations.