ECE/CS 559 - Fall 2021 - HW#8
Due: 11/25/2021, 11:00pm Chicago time.

Q1. (100 pts) Consider an $n \times n$ environment with locations (i, j), $i = 1, \ldots, n$, $j = 1, \ldots, n$. A miner named Sue begins at location $(1, 1)$ and can travel up, left, down, or right, provided that she remains within the environment at all times. There is a gold mine at location $(n, 1)$ and miner's home at location $(1, n)$. Sue begins with 0 coins in her backpack. Each time she visits the mine, she can get one gold. The maximum amount of gold that Sue can carry is given by some constant integer $G > 0$. Sue and the gold mine cannot be on the same location $(n, 1)$; instead Sue "visits" the gold mine by first traveling to a location that is adjacent to the gold mine (either $(n - 1, 1)$ or $(n, 2)$) and then taking an action that would place herself on the mine. This results in Sue acquiring one more gold provided that she is not carrying G golds already, and meanwhile, Sue's location remains the same. Sue can visit her home whenever she wants, similarly by first traveling to a location that is adjacent to her home, and taking an action that would place herself on top of her home. This results in Sue unloading all the gold she carries to her home. She receives the amount of gold she carries as the reward. Similarly, her location remains the same in this process.

An example scenario for $n = 3$ and $G = 2$ is as follows. Given the initial location $(1, 1)$, consider the sequence of actions RDRRRUULDDRUUL$\cdots$ The first action R will move Sue from (1,1) to (2,1). She receives no reward. Next action D attempts to place Sue outside the environment. Her location will remain the same and she will again receive no rewards. Next action, R, will acquire one gold from the mine, Sue will again stay at (2,1), no reward is received. Next action, R, will acquire another gold from the mine, Sue will again stay at (2,1), no reward is received. Next action, R, will not acquire another gold from the mine because Sue already has the maximum amount of golds she can carry, she again stays at (2,1), and no reward is received. Next two actions UU will move Sue to location (2,3). Subsequent action L will unload all golds to home. Sue's location will again be (2,3), but she will receive a reward of 2. Note that events where gold is unloaded to home are the only events for which the environment gives rewards. The next sequence of actions DDRUUL will go back to the square adjacent to the mine, pick up one gold, and drop it to home, receiving a reward of 1 only.

The cumulative reward is calculated in the standard manner. Let γ be the discount factor. Then, we define the cumulative reward as $\sum_{t=0}^{\infty} \gamma^t R_t$, where R_t is the reward at time t. The cumulative reward for the example above is then $2\gamma^7 + \gamma^{13} + \cdots$.

In the following, let $n = 5$ and $G = 3$. Find the optimal policies using (standard) Q learning. Initialize your Q table entries to be independent and identically distributed Gaussian random variables with zero mean and variance 1.

(a) How many states and actions does the problem have? Justify your answer. A rough answer (ignoring a few inadmissable states) is sufficient.

(b) Let $\gamma = 0.9$. Consider a pure-exploitation strategy that always chooses the best action during training.

i. Describe the policy learned by your algorithm, i.e. what does the miner do under this policy? Also, write the first 40 actions with the policy.

ii. What is the cumulative reward of the policy? Write down the value you obtained through your simulations.

iii. Do you think this is the optimal policy that maximizes the cumulative reward for the given γ? Justify your answer.

(c) Let $\gamma = 0.9$. Consider instead an exploration/exploitation strategy that chooses random actions with some probability, instead of the best actions.

i. Describe the policy learned by your algorithm, i.e. what does the miner do under this policy? Also, write the first 40 actions with the policy.

ii. What is the cumulative reward of the policy? Write down the value you obtained through your simulations.

iii. Do you think this is the optimal policy that maximizes the cumulative reward for the given γ? Justify your answer.

(d) Let $\gamma = 0.6$. Consider exploration/exploitation training.

i. Describe the policy learned by your algorithm, i.e. what does the miner do under this policy? Also, write the first 40 actions with the policy. Attach the code that generates the policy to the end of this exam file. Also, upload the code to the Box as 08-IDNumber-LastName.py. Upload one and only one file only. Without a functioning code for this part, answers to above questions will also not be given credit.

ii. What is the cumulative reward of the policy? Write down the value you obtained through your simulations.

iii. Compare and contrast the policy to the policy obtained in (c). If they are the same, explain why they should be the same. If they are different, explain why they should be different.