CSCI 4962 Homework 6
Siddha Kilaru

Task 1

Generally, reinforcement learning problems have three main factors, the agent, policy, and environment. The environment is essentially the specific game the agent is playing, such as tic-tac-toe or a robot navigating a course. At any specific time, the agent is in a specific state in the environment, and the agent refers to its policy to decide what to do. In other words, the policy is a function that maps states to actions. One way to mathematically formalize and model this problem is to use a Markov decision process (MDP). An MDP can be fully characterized by just five things, a set of states (S), a set of actions (A), a distribution containing the probabilities that taking $a \in A$ in $s \in S$ will lead to some $s' \in S$, a reward function for transitioning from $s \in S$ to $s' \in S$, and lastly a discount factor γ that decays the rewards over time.

Let us consider a reinforcement learning problem where a rover (robot) with a limited battery life must collect as many eggs as possible. The eggs are randomly distributed around the course. If the rover's battery is running low, the rover can return to the home position and recharge. The game is over if the rover runs out of battery in the middle of the course. An MDP can model this game. First, the set of states is a battery level greater than 20% and a battery level less than or equal to 20%. The set of actions is either search for eggs or hold the current position. The reward function is proportional to the number of eggs collected. A high-level overview of the transition model is as follows: if the battery is at a high state, the rover can either search or hold position, and if the battery is at a low state, the rover can either search or hold the position. From this, one can immediately see that if the rover is in a low battery state, it is optimal to hold the position and recharge. Conversely, if the rover is in a high battery state, it is optimal to search. A more complex optimal decision that the rover could make is when it starts to see its battery run low; it starts to search in the direction of the home position, where it can recharge. The rover would learn this over many iterations with the help of the reward function and the discount factor.

Task 2

Reinforcement Learning is highly applicable and related to healthcare problems. The process that doctors in health care undergo is very similar to the kind of decision-making process an agent makes in a reinforcement learning problem. For instance, a doctor would diagnose and suggest specific treatments to a patient, and if the patient's

health improves, the doctor continues to offer the current treatment. If the patient's health worsens, perhaps there was a misdiagnosis, or other treatments are better. This sequential decision process can be modeled by a reinforcement learning problem. More specifically, an MDP can model this kind of process.

Consider this open-source project https://github.com/microsoft/med-deadend.This project is a reinforcement learning model that identifies certain treatments doctors should avoid prescribing because they will lead to medical dead ends. A medical dead end is a state where all further actions/or in-action would lead to a patient's death. The transition model is fairly simple. Essentially, the state space consists of all the possible medical conditions a person has. The action space consists of all the possible prescriptions the doctor can make to a patient. The basic idea for the most optimal transition model is as follows: if a patient is at state s and action a causes the state to transition to a medical dead end with probability p, then the policy picks a at s with probability $1-p$. More formally, this is described by the optimal value functions $Q_D^*(s,a)$ and $Q_R^*(s,a)$. The project above aims to learn these functions and then can be used to identify medical dead ends in practice.

The model was tested on sepsis, and the researchers obtained exciting results. First, the results indicated "that more than 12 percent of treatments given to non-surviving patients could be detrimental 24 hours before death."[1] Also, they identified that "2.7 percent of non-surviving patients entered medical dead-end trajectories with a sharply increasing rate up to 48 hours before death, and close to 10 percent when we slightly relaxed our thresholds for predicting medical dead-ends."[1] These results can significantly impact the real world, potentially saving thousands of lives.

[1] Mehdi Fatemi, Taylor W. Killian, Jayakumar Subramanian, Marzyeh Ghassemi: "Medical Dead-ends and Learning to Identify High-risk States and Treatments", 2021.