Homework 01 - Group 10

Task 01

The game of chess as an Markov Decision Process.

A Markov Decision Process formually consists of two major parts and their interaction. The first being an Agent while the other one is the Environment the Agent responds to/ acts upon. In the case of Chess the Agent would be a player while the Chessboard and the current game represents the Environment. Hence, one time step in the environment would be one move done by the agent, the resulting state and the responding state (being from a real player or another agent).

The state in question is defined by the current disposition of the game. Thus, every piece an its position would be accounted for in the state definition. This can be done multiple ways be it with a 8x8 matrix with each field being represented in the state or with a 2x16 state representation where each piece is denoted with its current position. While the former might be more readable the second might be more computationally efficient.

Since the end goal of chess is to checkmate the opponent's king states depicting this should result in a positive reward (e.g. 1). If the opposite is true (the agent's king is checkmate) a penalty should be given (a negative reward like -1). Since winning the game as soon as possible is preferred a slight action penalty should be given (a small negative reward in each state that does not lead in ending the game e.g. 0.1).

The actions performed by the agent is moving one of the pieces per round. These actions are of course bound to the rules of the environment/game. Thus the pieces can only move in the way they are designed to be.

Thus, the policy for an action in a current state that would result in a checkmate state would be high.

Task 02

The Lunar Landar as a MDP

The same principles from before hold here aswell.
The Agent would be the Lander (or the entitiy steering it).
The Environment is represented in the planet where the agent is landing on.

The first part of the reward is the negative distance to the landing pat thus the objective can be more thought of as minimizing the penalty (then the regular reward maximization). The second is the actual landing of the Lander. If it lands successfully it receives a positive (100 points) reward while a negative one (-100) when it crashes. In addition to this, the agent receives a positive reward of 10 for each leg ground contact aswell as a slight penalty of 0.3 for firing the engine. Solving i.e. landing successfully in the landing pad grants a reward of 200.

The state is defined by the position of the landing pad and the position of the lander itself.

The action space is 4. Namely, the agent can either fire the main engine, do nothing or fire the right/left engine.

Thus, the policy would be the probability of either one of those for action given the current position of the lander and the landing pad.

Task 03

Environment dynamics

a)
The reward function defines the probability of a state with a certain reward given from a previous state after performing an action. For example the reward of performing the action "right engine" in state where the landing pad is right to you is based on the reward the resulting state is associated with (which is higher since you should be now closer to the goal landing point).

The transition function models the probability of encountering a state after performing an action in a previous state.

In general, these functions should model the properties and physics of the environment.

b)
The policy iteration makes use of these environment dynamics. Thus, the environment dynamics have to be modeled beforehand. Oftentimes for this only the goal/end state has to be known. Thus if these and the possible actions are given the problem can be solved with RL. Nonetheless, in some cases the state transition probabilities might be unknown, meaning it is hard to model the exact physics and properties of the environment. This can be a problem in real world applications like autonomous driving, as we might not be able to model all contingencies.
Hence, in some cases model-free learning might be more beneficial since the environment physics don’t have to be known.