Problem Set 2
INF368A - Reinforcement Learning
Spring 2024

Lars, Runar and Årne

February 2024

1 Markov Models and Bellman Equations

Exercise 1 Markov Processes

Consider the following scenario: a Mars Rover module has been designed to autonomously explore the surface of Mars. Every ten minutes, the module runs its program to transition to a new state. If it is in the normal operative mode of exploration, it will remain in that state with 70% chance and it will keep exploring for the next ten minutes; otherwise it will select uniformly at random among the following three alternatives: taking pictures of the sky, taking samples from the ground, recharging. If it selects taking pictures, it will spend ten minutes taking snapshots of the skies; then, it will revert to exploration with 60% chance or it will take samples from the ground with the remaining probability. If it ends up taking samples from the ground, it will dig for ten minutes; afterward, it will go back to normal operation mode with 35% chance, recharge with 30% chance, dig for more samples with 30% chance, or experience a malfunctioning with 5% chance. If it selects recharging, it will stay put for ten minutes collecting energy through its solar panels; after that, it will revert to normal exploration mode. If it experiences a malfunction, the robot will just shutdown.

1. (*) How can this model comply with the assumption of Markovianity?

The problem is already markovian since the available actions always only depend on the current state. So by simply modelling our states as is the state at time t + 1 only depends on the state at time t.

2. (*) Express the problem as a Markov Process. Write down its definition and draw its graph.

E = explore
P = taking photos

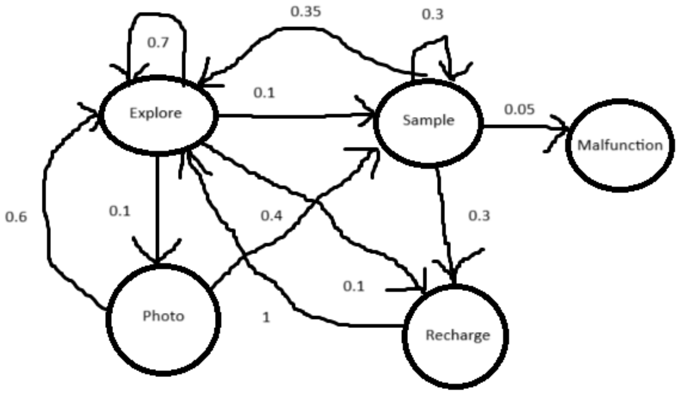

Figure 1: Markov process graph

S = taking samples
R = recharge
M = malfunction

	E	P	S	R	M
E	0.7	0.6	0.35	1	0
P	0.1	0	0	0	0
S	0.1	0.4	0.3	0	0
R	0.1	0	0.3	0	0
M	0	0	0.05	0	0

3. If the robot is in the normal operative mode, are all the states reachable? Is there any state that compromises the reachability of other states?

All states are not reachable. Malfunctioning is not reachable from the operative mode as it is only reachable from the taking samples from the ground state.

4. If the robot is in the normal operative mode, what is the probability of malfunctioning in the next 10 minutes? In the next 20 minutes? In the next 30

minutes?

In the next 10 minutes there is a 0% chance of malfunctioning.
In the next 20 minutes there is a 0.5% chance of malfunctioning.
In the next 30 minutes there is a 1.2% chance of malfunctioning.

5. (*) How would the model change if you were to be told that the probability of experiencing a malfunction in the digging state would depend on the number of times the robot has been extracting samples?

If the probability changes based on the number of times the robot has sampled the current model would no longer be markovian. To remedy this we can encode the number of times the robot has sampled together with the current action in each state S.

Consider the tic-tac-toe1 game, where the players are hard-coded computer programs.

Exercise 2 Markov Reward Processes

Enrich the previous Mars Rover scenario with the following rewards: 0 for normal exploration, +5 for taking pictures of the skies, +20 for collecting samples from the ground, 0 for recharging, -100 for malfunctioning.

1. (*) Redraw the graph for the previous MP as a graph for this new MRP.

See figure 2.

2. If the robot is in the normal operative mode, what is the reward you would expect it to collect in the next 10 minutes? What is the probability of a negative reward within 20 minutes?

To calculate the expected reward for the next 10 minutes starting from normal operative mode, we can calculate all outgoing branches from that state.
E(reward for next 10 minutes | normal operative mode) =

$$0 * 0.7 + 0.1 * 20 + 0.1 * 0 + 0.1 * 5 = 2.5$$

Expected reward in the next 10 minutes staring from state normal operative mode is 2.5

To find out the probability getting a negative reward within 20 minutes starting from normal operative mode, we need to calculate the probability of all paths that makes negative reward in the next 2 steps. This is only normal operative mode to sample to malfunctioning. We get the following:
P(Negative reward within 20 minutes | normal operative mode) =

$$0.1 * 0.05 = 0.005 = 0.5\%$$

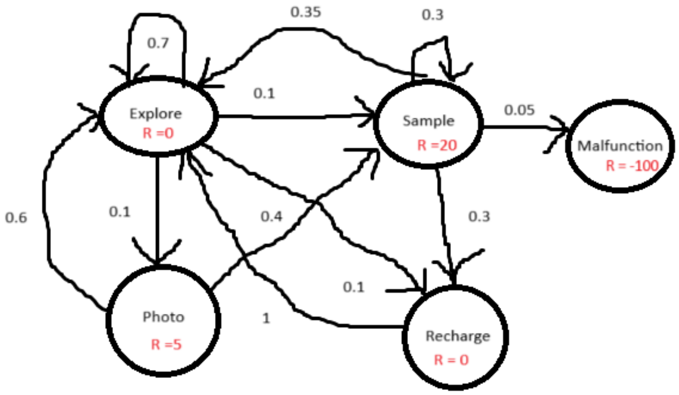

Figure 2: MRP graph

Exercise 3 Markov Decision Processes

Reconsider the previous MRP for the Mars Rover scenario.

1. (*) Can you model it as a MDP? How would you now model actions, states, transitions and policies?

To model this problem as an MDP we would considere the state transitions as actions the robot can take. We have our set of states S = [E,P,S,R,M] and the actions available at each state are the possible state transitions such as $A_E =$ [E,P,S,R]. Our policies will be the same as the transition probability matrix for the MRP however telling us how likely the robot is to select a given action with no information pertaining to how likely the robot is reach the state. For the most part state transition probabilities will be 1 for all state-action combinations except for when in the sampling state. For simplicity we can model the malfuction as an action the robot can take, which means that all probabilites are still 1. However malfuctioning is likely not an action we want our robot to take but rather an event that is out of our control. In this case we would model the other possible state transitions as actions with all of them having some probability of reaching their desired state and some small probability of reaching the malfunction state instead.

2. Consider the two objectives of maximizing rewards and minimizing the probability of failure. Assuming the same transitions as in the MRP, would the optimal policy for the two objectives be the same? Would the answer depend on the time horizon?

The optimal policy for the two objectives may not always be the same, and it could depend on the time horizon. For shorter periods, the rover might prioritize actions that maximize rewards without considering the risk of malfunctionig. But, for longer periods of time the rover might be more cautious, balancing between maximizing rewards and minimizing the probability of failure to ensure a longer journey.

3. (*) Reconsider the multi-armed bandit problem, for instance, the first problem in Exercise 8 of Problem Assignment 1. Can you formalize it as a MDP? How would it be?

For the MAB in assignment 1 we only have one possible state and three possible actions. We start in a state s, select an action of using medicine A, B or C and then transition back to state s. The probability of each action is learned over time so we can assume a uniform distribution in the beginning. Each action is guaranteed to transition back to s so the state transition probabilities are irrelevant. The reward function is sampling from a normal distribution with variance 1 and mean 1,2 and 3 for A,B and C respectively.

Exercise 4 Bellman equations

1. (*) Can you define Bellman equations for a MP? If yes, draw the corresponding backup diagrams and derive the Bellman equations. If not, explain why not.

No you cannot. The bellman equations consider the rewards of states and actions and for an MP the rewards are not defined.

2. (*) Can you define Bellman expectation equations for a MRP? If yes, draw the corresponding backup diagrams and derive the Bellman equations. If not, explain why not.

Normally it does not inherently make sense to consider the bellman equations in terms of an MRP since we don't consider actions or policies in an MRP however for the sake of the task we can consider an MRP as an MDP with a single action which transitions into one of the possible states. The backup diagram for such a model with three possible state transitions might look something like Figure 3. To find the expected value of the base state we need to consider the reward for the base state as well as potential further reward in later states multiplied by some discount factor γ. So the state value will be given by

$$v(s) = E\left[R(0) + \sum_{i=1}^{T} \gamma^{i} R^{(t+i)} | S^{(0)=s}\right]$$

Or equivalently

$$v(s) = E\left[R(0) + \gamma v(s') | S^{(0)} = s\right]$$

where $v(s')$ is a possible next states value.
Since we only have one action to consider this can be simplified to

$$v(s) = R(0) + \gamma \sum_{s' \in s} P(s'|s) v(s')$$

Action values do not really make sense since MRP's do not consider actions.

3. (*) Can you define Bellman optimality equations for a MRP? If yes, draw the corresponding backup diagrams and derive the Bellman equations. If not, explain why not.

The bellman optimality equations are used to determine the optimal policy for an agent which makes decisions on what actions to take. As such it makes little sense to determine an optimal policy for an MRP model as an MRP has concept of actions and policies. However for the sake of the task we could define it as an optimal trajectory i.e. the trajectory of states which would maximize our gain.

$$v^{*}(s) = R(0) + max_{s' \in S}\left[\gamma v(s')\right]$$

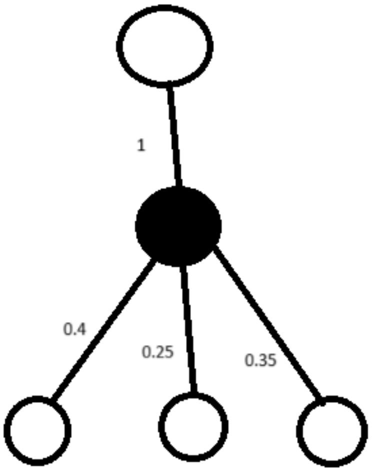

Figure 3: Backup diagram MRP