

Indian Statistical Institute

Reinforcement Learning — M.Tech. CS

End-Semester Examination: 2025–26 Semester I

Date: 20 November 2025

Full Marks: 100 Time: 3 Hours

Instructions: Questions 1-10 carry 7 marks each, and Questions 11-13 carry 10 marks each.

1. Suppose, in a deterministic MDP, two different deterministic optimal policies π_1 and π_2 exist. Let s be a state from which the two policies take different actions, i.e., $\pi_1(s) = a_1$ and $\pi_2(s) = a_2$. Then, prove or disprove: from state s both actions a_1 and a_2 must get the same immediate reward.
2. Why do Monte Carlo methods tend to have higher variance than temporal difference methods? Explain with an example.
3. Give an example of an episodic MDP where a Monte Carlo method updates action values for many state-action pairs after one episode, but TD(0) updates only one. Briefly explain why.
4. In the well-known cliff-walking MDP, ϵ -greedy SARSA avoids the path along the cliff even though it is optimal. Explain why. If ϵ is reduced gradually, will it cause any difference?
5. Consider the statement: A function approximation method may not be able to learn the exact value, or action-value functions, whereas a tabular method can. Show an example to explain this statement.
6. Prove or disprove the statement: Whenever a function approximation method fails to learn the exact value / action-value function, it will also fail to learn the optimal policy.
7. Explain, for which values of λ , the λ -return is equivalent to the one-step TD return and the Monte Carlo (MC) return.
8. Explain how planning (as in Dyna-Q) can accelerate learning compared to a purely model-free algorithm. Mention one example scenario where planning might not help much, or even may hurt.
9. Explain how experience replay improves both stability and sample efficiency (the ability to learn faster without many real samples) in DQN.
10. Why does an actor-critic algorithm need to learn both the policy and a value function, whereas REINFORCE needs to learn only the policy function?

11. As we see in Figure 1, Indiana Jones is standing at position 0 and there is a crown of high archaeological value at position 6. The tiles along positions 1-5 are made of glass. Some tiles (in this picture tiles 1,3 and 5) are fragile. Stepping on a fragile tile or jumping to some position greater than 6 means Dr. Jones falls into the gutter and has to start again from 0. Since this is a reinforcement learning assignment, we make him start over instead of killing him. He can see that the crown is at position 6, but he cannot understand whether a tile is fragile without actually stepping on it.

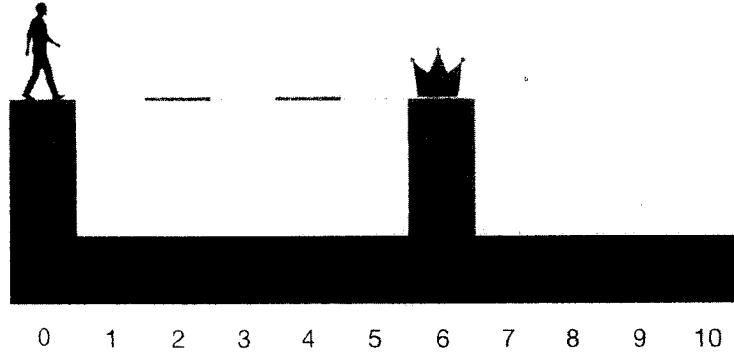


Figure 1: Indiana Jones and the Crown

The rules of the adventure are as follows. Dr. Jones has to jump from one tile to another, but he cannot just choose his length of jump. At each step, he has two choices, namely, to toss a coin or to roll a die. In case of the coin, he has to jump one position forward if head occurs, two positions otherwise. If he rolls the die, it will show a number $i \in \{1, 2, \dots, 6\}$, and he has to jump i positions forward. In this process, if he manages to reach 6 (i.e., does not fall down before 6 and does not go overboard to position 7 or more), he gets the crown (reward 1). All other positions have reward 0. Assume that the coin and the die are unbiased. Answer the following questions based on this scenario.

- Compute the optimal value function $v_*(s)$ and optimal action-value function $q_*(s, a)$ for all non-terminal states s and all actions $a \in \{\text{toss}, \text{roll}\}$. Since you know the dynamics of the environment, you may use that for this question.
- From the optimal action-value function, determine an optimal policy. How many optimal policies are possible for this scenario?
- If Dr. Jones follows the optimal policy for one episode, what is the probability that he will get the crown?

$$[(3 + 3) + (2 + 1) + 1 = 10]$$

12. Assume that in an gridworld, each action generates a constant reward R for reaching all non-terminal states and R_T for reaching the terminal state. The goal of an agent is to learn the shortest path from any state to a terminal state. Which of the following setups will make sure that maximizing the cumulative return $G_t = \sum_{k=t}^{T-1} \gamma^{k-t} R_{k+1}$ aligns with the goal, i.e., learns the shortest path? Explain your answers briefly.

- (a) $R = -1, R_T = 0, \gamma = 1.$
- (b) $R = -1, R_T = -1, \gamma = 0.9.$
- (c) $R = 0, R_T = 1, \gamma = 1.$
- (d) $R = 0, R_T = 1, \gamma = 0.8.$
- (e) $R = 1, R_T = 1, \gamma = 0.9.$

[2 × 5 = 10]

13. Briefly describe how the meta-RL algorithm RL^2 works using a diagram. Mark how the trials, episodes, states, rewards, hidden states and parameters of the network change.

[10]