CS 4122: Reinforcement Learning

Practice Problem Set 4

Release date: 28th March, 2025

NOTE: This problem set is quite important from exam perspective. Not that other problem sets are not important. They are but from the perspective of developing clear concepts. This problem set however is reflective of the kind of problems that will be asked during exams.

Problem 1: Rescue Operation in GridWorld

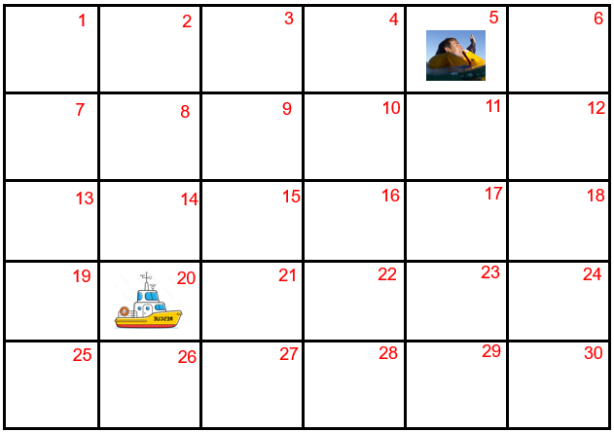

Figure 1: Gridworld showing a rescue boat (grid 20) and a floating survivor wearing life jacket (grid 5).

Consider the gridworld shown in Figure 1. This gridworld depicts a particular portion of a sea. The grid numbers are shown in the upper-right hand side of the grids. *Please Note: The number of grids does not really matter. The figure is just for your visual reference.*

We consider that time is divided into slots. At a given time slot, a rescue boat that is patrolling the sea receives a distress signal from a survivor who is floating in a life jacket. The survivor does not know swimming and hence drifting in the sea. Let $p_s(x^+ \mid x)$ denote the probability of the survivor being in grid number x^+ in the next time slot provided that its grid number is x in the current time slot. We assume that the rescue boat gets continuous update about its location in the sea as well as the location of the survivor.

The rescue boat has to decide whether to go up, down, left, or right in order to minimize the expected search time. One proxy to minimize the expected search time is to minimize the β-discounted cost,

$$E\left[\sum_{t=0}^{T} \beta^t c_t\right]$$

where $c_t = 1$ if the grid of the survivor and the boat is not same and $c_t = 0$ otherwise. In the above equation, T is the first time slot when the grid of the survivor and the boat is the same. Finally, the motion of the rescue boat following an action $a \in \{up, down, left, right\}$ is captured by the

probability distribution $p_b(y^+ \mid y, a)$ which is the probability that the boat will be in grid number y^+ in the next time slot given that its current grid number and current action are y and a respectively. Answer the following questions:

(a) Is this a continuing or an episodic task?
(b) What are the state and state space?
(c) What is the immediate average reward of every state-action pair?
(d) What is the Bellman optimality equation for this problem?
(e) Write the equation for value iteration in order to solve the Bellman optimality equation.

Problem 2: Mars Rover Path Planning

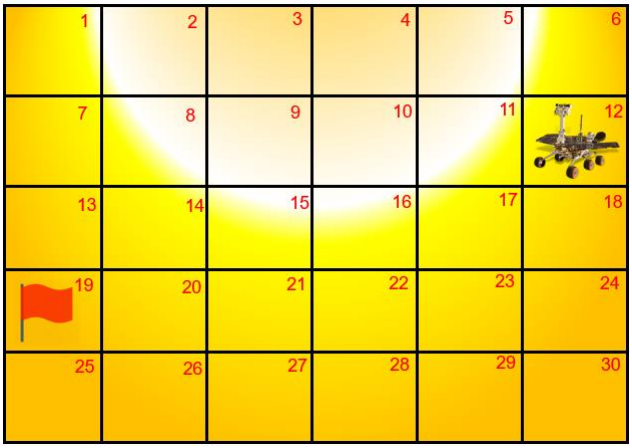

Figure 2: Gridworld showing mars rover (grid 12) and its target grid (grid 19). The shading of the grids shows the level of sunlight the grid receives.

This problem deals with path planning for mars rover. For most practical scenarios, this problem can be visualized as navigating in a gridworld as shown in Figure 2. The mars rover has to go from its current position to a target destination as quickly as possible in order to conduct some experiment. In general, the quickest path is a straight-line path. However, the twist is that the mars rover is powered solely using solar energy and different grids receive different level of sunlight as shown in Figure 2. Hence, the mars rover may have to consider a longer path in order a reach the destination quickly! *Please Note: The number of grids does not really matter. The figure is just for your visual reference.*

We consider that time is divided into slots. The mars rover has to reach a target grid $\bar{x}$ starting from its current grid x_0. The mars rover is equipped with a battery of infinite capacity. Let b_t denote the units of battery power left at time slot t. At any given time slot, the battery power of the rover increases by an amount δ due to solar energy. δ is a random variable with probability distribution $p_s(\delta \mid x)$ where x is the grid location of the rover in that time slot. We assume that the initial battery power and δ are both integers. At any given time slot, the rover can decide the following:

1. To not move. If the rover does not move, it does not consume any battery power.
2. To move. This means that the rover has to decide the direction and the speed. The direction can be up, down, left, or right. The speed can be high or low. If the speed is low, the rover moves by

one grid in the chosen direction and consumes a battery power of θ_l units. If the speed is high, the rover moves by two grids in the chosen direction and consumes a battery power of θ_h units where $\theta_h > \theta_l$. Off course, the rover cannot consume more battery power than it currently stores.

The objective is to minimize the β-discounted cost,

$$E\left[\sum_{t=0}^{T} \beta^t c_t\right]$$

where $c_t = 1$ if the rover's grid at time t is not equal to the target grid $\bar{x}$ and $c_t = 0$ otherwise. Answer the following questions:

(a) Is this a continuing or an episodic task?
(b) What are the state and state space?
(c) What is the immediate average reward of every state-action pair?
(d) What is the Bellman optimality equation for this problem?

Problem 3: Just like problems 1 and 2, consider a 2D-Gridworld of dimension $N \times N$. We consider a time slotted model. A robot operates on this grid. The actions a available to the robot is up, down, left, and right. The dynamics of the robot is captured by the probability distribution $p(x^+ \mid x, a)$ which is the probability that the boat will be in grid x^+ in the next time slot given that its current grid and action are x and a respectively. The objective of a robot is to reach a destination grid x_f **via an intermediate grid x_e** in the shortest possible time. In other words, wherever the robot starts, the robot has to first visit x_e and then x_f and this entire process has to be done in the shortest time.

(a) This is not a discounted reward problem; here you have to use sum of rewards (refer to slide 7 of lectures 17 to 20). Go through the steps of derivation of both the Bellman equation and the Bellman optimality equation and convince yourself that the steps of derivation remains the same; it is just that we have to use discount factor $\beta = 1$ and he concept of "end state" mentioned in lecture 13.
(b) What is the Bellman optimality equation for this problem to minimize the expected travel time to the destination grid. *HINT: Use a state that indicates if the intermediate grid is visited.*

Problem 4: Many a times we have to switch between WiFi and mobile data based on availability. This problem is motivated by this broad idea. Whenever we are using WiFi or mobile data, we are selecting a wireless channel to transmit on. Let's say that WiFi and mobile data are associated with channels 1 and 2 respectively.

Let $i \in \{1,2\}$ denote the channel index. Channel i is either available for not available in a time slot. Let $s_{i,t} \in \{0,1\}$ denote whether channel i is available ($s_{i,t} = 1$) or not available ($s_{i,t} = 0$) in time slot t. If channel i is <u>not available</u> in time slot t, then the probability that it is <u>not available</u> in next time slot $t + 1$ is α_i. If channel i is <u>available</u> in time slot t, then the probability that it is <u>available</u> in next time slot $t + 1$ is θ_i. If a channel is not available, the number of data packets that we can send through it is obviously zero. If channel i is available, we can set $d \in \{0,1,\cdots,D\}$ packets through it where the probability of sending d packets is $p_{i,d}$.

If we switch to channel i, we can't start to use it even if it is available. We have to wait for τ_i time slots for the device to establish connection with the channel. For this τ_i time slots, the number of data packets that we can send is zero. The objective is to maximize the discounted sum of the number of data packets transmitted over an infinite horizon. Answer the following questions:

(a) Formulate this problem as a Markov decision process.
(b) Give a mathematical expression of a greedy policy for this problem. A greedy policy maximizes the immediate average reward.
(c) Qualitatively speaking, why planning is likely to be beneficial for this problem? In other words, why it is important to think ahead for this particular problem?
(d) Write the Bellman equation for this problem for the greedy policy derived in part (b).
(e) Write the Bellman optimality equation for this problem.

Problem 5: Solve the last two years problems related to application of Markov decision process. These problems are available in the Dropbox link in the subfolder titled “previous year exams”. More specifically:

(a) Fall 2023, Mock exam of minor 2, Q2.
(b) Fall 2023, Minor 2, Q2.
(c) Fall 2023, Mock exam of end-sem, Q2.
(d) Fall 2023, End-sem, Q2.
(e) Fall 2024, Mock exam of minor 2, Q3.
(f) Fall 2024, Minor 2, Q3.
(g) Fall 2024, Mock exam of end-sem, Q4.
(h) Fall 2024, End-sem, Q4.