ECE433/COS435 Introduction to RL
Solution of Assignment 1: MDP
Spring 2025

Fill me in

Your name here.

Due February 10, 2025

Question 1. Markov Chain

We have a two-state Markov chain with two states, s_1 and s_2. The probability of transitioning from s_1 to s_2 is $1-p$, and vice versa. We can summarize the transition probabilities in the table shown below.

$$P = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix},$$

the value of $P_{i,j}$ indicates the probability of transiting from state i to state j, for any $i, j \in [1, 2]$.

Question 1.a

Use the principle of induction[1] to show that

$$P^{(n)} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}(2p-1)^n & \frac{1}{2} - \frac{1}{2}(2p-1)^n \\ \frac{1}{2} - \frac{1}{2}(2p-1)^n & \frac{1}{2} + \frac{1}{2}(2p-1)^n \end{bmatrix}.$$

Solution

Base Case ($n = 1$):

$$\begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}(2p-1)^1 & \frac{1}{2} - \frac{1}{2}(2p-1)^1 \\ \frac{1}{2} - \frac{1}{2}(2p-1)^1 & \frac{1}{2} + \frac{1}{2}(2p-1)^1 \end{bmatrix}.$$

[1] `https://en.wikipedia.org/wiki/Mathematical_induction`

Our induction hypothesis is

$$P^{(n)} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}(2p-1)^n & \frac{1}{2} - \frac{1}{2}(2p-1)^n \\ \frac{1}{2} - \frac{1}{2}(2p-1)^n & \frac{1}{2} + \frac{1}{2}(2p-1)^n \end{bmatrix}.$$

And we want to show that the above form holds for $P^{(n+1)}$. Observe that

$$P^{(n+1)} = PP^{(n)} = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix} \begin{bmatrix} \frac{1}{2} + \frac{1}{2}(2p-1)^n & \frac{1}{2} - \frac{1}{2}(2p-1)^n \\ \frac{1}{2} - \frac{1}{2}(2p-1)^n & \frac{1}{2} + \frac{1}{2}(2p-1)^n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} + (p-\frac{1}{2})(2p-1)^n & \frac{1}{2} + (\frac{1}{2}-p)(2p-1)^n \\ \frac{1}{2} + (\frac{1}{2}-p)(2p-1)^n & \frac{1}{2} + (p-\frac{1}{2})(2p-1)^n \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} + \frac{1}{2}(2p-1)^{n+1} & \frac{1}{2} - \frac{1}{2}(2p-1)^{n+1} \\ \frac{1}{2} - \frac{1}{2}(2p-1)^{n+1} & \frac{1}{2} + \frac{1}{2}(2p-1)^{n+1} \end{bmatrix}$$

Question 1.b

Expectation of State Occupancy

- Compute the expected number of times the process is in s_1 after n transitions (including the starting state $t = 0$), starting from s_1.
- Compute the expected number of times the process is in s_2 after n transitions, starting from s_1.
- For $p \neq 0$, discuss how the expectations change as n approaches infinity.

Solution

We have the Markov Chain $(X_t)_{t\in\mathbb{N}}$ with the transition probability matrix P. Let $N_{(j)}(n)$ be the random variable denoting the number of times the process is in state (j) after n steps:

$$\begin{aligned} \mathbb{E}[N_{s_1}(n) \mid X_0 = s_1] = \mathbb{E}\left[\sum_{i=0}^{n} \mathbf{1}_{\{X_i = s_1\}} \mid X_0 = s_1\right] &= \sum_{i=0}^{n} \mathbb{E}\left[\mathbf{1}_{\{X_i=s_1\}} \mid X_0 = s_1\right] \\ &= \sum_{i=0}^{n} \mathbb{P}\left(X_i = s_1 \mid X_0 = s_1\right) \\ &= \sum_{i=0}^{n} P_{11}^{(i)} \\ &= \sum_{i=0}^{n} \frac{1}{2} + \frac{1}{2}(2p-1)^i \\ &= \frac{n+1}{2} + \frac{1}{2} \cdot \frac{1-(2p-1)^{n+1}}{2-2p} \end{aligned}$$

The first two equalities holds by linearity and properties of indicators random variables. The last step contains a geometric sum. Similarly,

$$\begin{aligned}
\mathbb{E}[N_{s_2}(n) \mid X_0 = s_1] = \mathbb{E}\left[\sum_{i=0}^{n} \mathbf{1}_{\{X_i = s_2\}} \mid X_0 = s_1\right] &= \sum_{i=0}^{n} \mathbb{E}\left[\mathbf{1}_{\{X_i = s_2\}} \mid X_0 = s_1\right] \\
&= \sum_{i=0}^{n} \mathbb{P}\left(X_i = s_2 \mid X_0 = s_1\right) \\
&= \sum_{i=0}^{n} P_{12}^{(i)} \\
&= \sum_{i=0}^{n} \frac{1}{2} - \frac{1}{2}(2p-1)^i \\
&= \frac{n+1}{2} - \frac{1}{2} \cdot \frac{1-(2p-1)^{n+1}}{2-2p}
\end{aligned}$$

As $n \to \infty$, the above expectations converge to $\frac{n+1}{2} \pm \frac{1}{2(2-2p)}$ respectively (if $p \neq \frac{1}{2}$).

Question 1.c

Probability of First Visit

- Compute the probability that the process visits s_2 for the first time on the k-th transition, given it starts in s_1. What happens if $k \to \infty$?

Solution

This is just a geometric random variable, so

$$\mathbb{P} = \mathbb{P}\left(\text{in } s_1 \text{ for } k-1 \text{ steps}\right) \mathbb{P}\left(\text{in } s_2 \text{ on the kth step}\right) = p^{k-1}(1-p)$$

As $k \to \infty$, this probability goes to 0.

Question 1.d

Conditional Expectations

- Given that the chain is in s_2 at the n-th step, compute the conditional expectation of the number of visits to s_1 in the next m steps.

Solution

By Markov property and the fact that our transition matrix is symmetric, this is similar to our solution to (1.b). In other words,

$$\mathbb{E}\left[N_{s_2}(n+m) - N_{s_2}(n)|X_n = s_2\right] = \mathbb{E}\left[N_{s_2}(m)|X_0 = s_2\right] = \sum_{i=0}^{m} P_{21}^{(i)} = \sum_{i=0}^{m} \frac{1}{2} - \frac{1}{2}(2p-1)^i$$

$$= \frac{m+1}{2} - \frac{1}{2} \cdot \frac{1-(2p-1)^{m+1}}{2-2p}.$$

Question 1.e

Expected rewards. When transiting from one state to another, assume we receive a reward of 1 for reaching s_2 and -1 for reaching s_1.

- Compute the expected total reward after n transitions (i.e., the summation of rewards), starting from s_1.

Solution

Using our answer in (1.b), we can assign a positive reward for being in s_2 and a negative reward for being in s_1. So the expected total reward $\mathcal{R}(n)$ after n steps is

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(n)|X_0 = s_1\right] &= 1 - 1 \cdot \mathbb{E}\left[N_{s_1}(n)|X_0 = s_1\right] + 1 \cdot \mathbb{E}\left[N_{s_2}(n)|X_0 = s_1\right] \\
&= 1 - \frac{n+1}{2} - \frac{1}{2} \cdot \frac{1-(2p-1)^{n+1}}{2-2p} + \frac{n+1}{2} - \frac{1}{2} \cdot \frac{1-(2p-1)^{n+1}}{2-2p} \\
&= 1 - \frac{1-(2p-1)^{n+1}}{2-2p}
\end{aligned}$$

Question 2. Grid World Example

In this exercise, you will work with a simple reinforcement learning environment called "Gridworld." Gridworld is a 4x4 grid where an agent moves to reach a goal state. The agent can take four actions at each state (up, down, left, right) and receive a reward for each action. Moving into a wall (the edge of the grid) keeps the agent in its current state.

Grid Layout:

- The grid is a 4x4 matrix.
- Start state (S): Top left cell (0,0).
- Goal state (G): Bottom right cell (3,3).

The agent receives a reward of -1 for each action until it reaches the goal state.

Question 2.a

Formulate the problem as a Markov Decision Process (MDP). Define the states, actions, transition probabilities (assume deterministic transitions), rewards, and policy.

Solution

States:

Positions in the grid, i.e., (i, j) where $i, j \in \{0, 1, 2, 3\}$. There are 16 states in total.

Actions:

At each state, the agent can choose from four actions: up (U), down (D), left (L), and right (R).

Transitions:

From state s and action a, if transiting to s' is permitted (without hitting walls), then

$$P(s'|s, a) = 1.$$

If hitting walls, then

$$P(s|s, a) = 1.$$

For all other s',

$$P(s'|s, a) = 0.$$

Rewards:

Before reaching the goal state (G), we have

$$R(s, a) = -1, \forall s \in S, a \in A.$$

Policy:

A policy $\pi : S \to A$ is the action an agent should take when in a given state. Here we define a uniform policy: It randomly chooses each action

$$\pi(s) = \begin{cases} U & \text{w.p. } 1/4 \\ D & \text{w.p. } 1/4 \\ R & \text{w.p. } 1/4 \\ L & \text{w.p. } 1/4. \end{cases}$$

Question 2.b

How many unique (deterministic) policies are there in total?

Solution

The goal state terminates the game, so no actions can be taken there. We can take 4 actions in every state. Thus, we have 4^{15} possible policies.

Question 3

Solution

Prescribing medication throughout a treatment plan for a patient is inherently sequential and involves significant considerations for the future state of the patient. Each intervention has consequences that affect the patient's health, influencing the appropriateness of future interventions. This is a classic scenario for general RL, where each action (medical intervention) affects the environment (patient's health state) in a way that must be considered for future decisions. The objective is not just to maximize the immediate outcome of an intervention, but to optimize the patient's health over the entire treatment plan. This requires a model that can account for the temporal dependencies between actions and their long-term effects on the state, making general RL the suitable approach.