ECE433/COS435 Introduction to RL
Assignment 1: MDP
Spring 2024

Fill me in

Your name here.

Due February 11, 2024

Collaborators

Fill me in

Please fill in the names and NetIDs of your collaborators in this section.

Instructions

You should work alone on this assignment. Writeups should be typesetted in Latex and submitted as PDFs. You can work with whatever tool you like for the code, but **please submit the asked-for snippet and answer in the solutions box as part of your writeup. We will only be grading your write-up.** Make sure still also to attach your notebook/code with your submission.

Question 1. Markov Chain

Let the transition probability matrix of a two-state Markov chain be specified by We have a Markov chain with two states, s_1 and s_2. The probability of transitioning from s_1 to s_2 is 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 ($n = 1$):

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

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

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

Question 1.b

Expectation of State Occupancy

- Compute the expected number of times the process is in s_1 after n transitions, starting from s_1.
- Compute the expected number of times the process is in s_2 after n transitions, starting from s_1.
- Discuss how the expectations change as n approaches infinity and the implications for the state occupancy for the above two cases.

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

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

(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 third equality holds by the Chapman-Kolomogrov theorem of Markov Chains (or can also be computed by Bayes Formula). The remainder follows from the previous part and finally 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 $\frac{n}{2} \pm \frac{1}{2(2-2p)}$ respectively (if $p \neq \frac{1}{2}$), indicating that there is some form of stationary distribution in the limit.

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.
- How does this probability change as k increases?

Solution

This is just a geometric random variable, so

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

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.
- Explore how this expectation varies with different values of p.

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,

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

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

By construction, p is the probability of staying at a state. If $p = \frac{1}{2}$, the conditional expectation goes to $\frac{n+1}{2}$ as $n \to \infty$. Otherwise, if $p > \frac{1}{2}$, the expected number of visits is below half, and vice versa.

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. Secretary problem

Suppose you are hiring one secretary and going to interview the candidates one by one sequentially. After an interview is over, you have to decide whether to hire the current candidate. If you hire the current candidate, the whole process stops. Otherwise, the interview continues, but the candidate will not return and cannot be hired. In other words, you can only hire the current candidate but cannot hire the past candidate. In total, there are n candidates, and you have a strict preference among them. It means you can tell who is better than whom, and no two candidates are equal. When meeting a new candidate, you compare with past candidates, but you do not know the ranking of current candidates in all people. For example, if you have interviewed c_1, c_2, c_3, then you can rank these three, say $c_1 > c_3 > c_2$, but you do not know the ranking of c_1 among all n candidates. Candidates come in a uniform random order. The objective is to find a policy (when to stop) in order to maximize the probability of hiring the best candidate. The problem is known as the optimal stopping problem. It can be formulated by MDP.

Specifically, denote t as the time after we interview the t-th candidate, $t = 1, \cdots, n$. We introduce a state variable $s_t \in \{-1, 0, 1\}$. $s_t = -1$ means the position is filled. $s_t = 0$ means the position is not filled, and the current t-th candidate is not the best candidate so far. $s_t = 1$ means the position is not filled, and the current t-th candidate is the best candidate so far. Initially, we set $s_t = 1$ because after interviewing the first candidate, he/she must be the best among past candidates. Suppose we do not hire anyone and keep interviewing.

Question 2.a

Compute the following probabilities with variable t: $P_t(s, 1) = \mathbb{P}(s_{t+1} = 1|s_t = s)$ and $P_t(s, 0) = \mathbb{P}(s_{t+1} = 0|s_t = s)$.

Solution

$$P_t(s,1) = \mathbb{P}(s_{t+1} = 1|s_t = s) = \frac{1}{t+1}, P_t(s,0) = \mathbb{P}(s_{t+1} = 0|s_t = s) = \frac{t}{t+1}.$$

Notice the transition matrix depends on time t but is independent of starting state s.

Question 2.b

At time t, our action is either to hire ($a_t = 1$) or to continue interviewing ($a_t = 0$). Show the value of $P(s_{t+1} = s'|s_t = s, a_t = a)$ for $\forall s, s' \in \{-1, 0, 1\}, a_t \in \{0, 1\}$. (Hint: you may keep $P_t(s, s')$ in the expression)

Solution

for $s, s' \in \{0, 1\}$,

$$\mathbb{P}(s_{t+1} = -1|s_t = s, a_t = 1) = 1,\ \mathbb{P}(s_{t+1} = s'|s_t = s, a_t = 0) = P_t(s, s').$$

For completeness, we define for $a \in \{0, 1\}, \mathbb{P}(s_{t+1} = -1|s_t = -1, a_t = a) = 1$. Therefore, we will find that $\{s_t^a\}_{1\leqslant t\leqslant n}$ can be formalized by an MDP.

Question 3. 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 3.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 3.b

Define the policy.

Solution

See **3.a**.

Question 3.c

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.