# COMPSCI 687 Homework 1 - Fall 2024

Due **September 26, 2024**, 11:55pm Eastern Time

## 1 Instructions

This homework assignment consists of a written portion and a programming portion. While you may discuss problems with your peers (e.g., to discuss high-level approaches), you must answer the questions on your own. In your submission, do explicitly list all students with whom you discussed this assignment. Submissions must be typed (handwritten and scanned submissions will not be accepted). You must use LaTeX. The assignment should be submitted on Gradescope as a PDF with marked answers via the Gradescope interface. The source code should be submitted via the Gradescope programming assignment as a .zip file. Include with your source code instructions for how to run your code. You **must** use Python 3 for your homework code. You may not use any reinforcement learning or machine learning specific libraries in your code, e.g., TensorFlow, PyTorch, or scikit-learn. You *may* use libraries like numpy and matplotlib, though. The automated system will not accept assignments after 11:55pm on September 26. The tex file for this homework can be found here.

## 2 Hints and Probability Review

- **Write Probabilities of Events:** In some of the probability hints below that are not specific to RL, we use expressions like $\Pr(a|b)$, where $a$ and $b$ are events. Remember that in the RL notation used for this class, the values of $\Pr(s_0)$, $\Pr(a_0)$, $\Pr(A_0)$, or $\Pr(A_0|S_0)$ are all undefined, since those are simply states, actions, or random variables (not events). Instead, we ***must*** write about the probabilities of events. For example: $\Pr(A_0 = a_0)$ or $\Pr(A_0 = a_0|S_0 = s_0)$.
- **Bayes' Theorem:** $\Pr(a|b) = \frac{\Pr(b|a)\Pr(a)}{\Pr(b)}$. This is useful for dealing with conditional probabilities $\Pr(a|b)$ if the event $a$ occurs *before* event $b$. For example, it is often difficult to work with an expression like $\Pr(S_0 = s_0|A_0 = a_0)$, because the agent *first* observes the current state, $S_0$, and only afterwards selects an action, $A_0$; in this case, it is much easier to deal with the 3 terms in $\frac{\Pr(A_0=a_0|S_0=s_0)\Pr(S_0=s_0)}{\Pr(A_0=a_0)}$.
- **The law of total probability:** For an event $a$, and a set of events $\mathcal{B}$,

$$\Pr(a) = \sum_{b\in\mathcal{B}} \Pr(b)\Pr(a|b).$$

  See the example below for several useful applications of this property.
- **"Extra" given terms:** Remember that when applying laws of probability, any "extra" given terms stay in the result. For example, applying the law of total probability:

$$\Pr(a|c,d) = \sum_{b\in\mathcal{B}} \Pr(b|c,d)\Pr(a|c,d,b).$$

- **Conditional Probabilities - Useful property #1:** If you need to move terms from the "right-hand side" of a conditional probability to the "left-hand side", you can use the following identity:
  $\Pr(a|b,c) = \frac{\Pr(a,b|c)}{\Pr(b|c)}$
- **Conditional Probabilities - Useful property #2:** If you need to move terms from the "left-hand side" of a conditional probability to the "right-hand side", you can use the following identity:
  $\Pr(a,b|c) = \Pr(a|b,c)P(b|c)$.
- **Expected Values:** The expected value of a random variable $X$ with possible outcomes in $\mathcal{X}$ is

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x\Pr(X = x).$$

- **Conditional Expected Values:** The expected value of a random variable $X$ with possible outcomes in $\mathcal{X}$, conditioned on an event $A = a$, is

$$\mathbb{E}[X \mid A = a] = \sum_{x\in\mathcal{X}} x\Pr(X = x \mid A = a).$$

- **Example problem:** The probability that the state at time $t = 1$ is $s \in \mathcal{S}$.

$$\Pr(S_1 = s) = \sum_{s_0 \in \mathcal{S}} \Pr(S_0 = s_0) \Pr(S_1 = s | S_0 = s_0) \tag{1}$$

$$= \sum_{s_0 \in \mathcal{S}} d_0(s_0) \Pr(S_1 = s | S_0 = s_0) \tag{2}$$

$$= \sum_{s_0 \in \mathcal{S}} d_0(s_0) \sum_{a_0 \in \mathcal{A}} \Pr(A_0 = a_0 | S_0 = s_0) \times \Pr(S_1 = s | S_0 = s_0, A_0 = a_0) \tag{3}$$

$$= \sum_{s_0 \in \mathcal{S}} d_0(s_0) \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0)\, p(s_0, a_0, s). \tag{4}$$

---

# Part One: Written (55 Points Total)

1. ***(Your grade will be a zero on this assignment if this question is not answered correctly)*** *Read the class syllabus carefully, including the academic honesty policy. To affirm that you have read the syllabus, type your name as the answer to this problem.* SHREYA BIRTHARE

2. (**18 Points [Total]**) Given an MDP $M = (\mathcal{S}, \mathcal{A}, p, R, d_0, \gamma)$ and a fixed policy, $\pi$, the probability that the action at time $t = 0$ is $a \in \mathcal{A}$ is:

$$\Pr(A_0 = a) = \sum_{s \in \mathcal{S}} d_0(s) \pi(s, a). \tag{5}$$

Write similar expressions (using only $\mathcal{S}, \mathcal{A}, p, R, d_0, \gamma$, and $\pi$) for the following problems.

**Important:**

- *Assume, below, that the reward function will be in the form $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. That is, the reward at time $t$ depends only on the state at time $t$ and action at time $t$.*
- *All solutions below need to be derived from "first principles": you should repeatedly apply definitions and properties of probability distributions such as the ones discussed in Section 2, as well as the Markov Property (when appropriate), and then replace the relevant quantities with their corresponding definitions in RL (e.g., you can substitute $\Pr(A_0 = a | S_0 = s)$ with $\pi(s, a)$).*
- *Remember that the Markov Property allows you to ignore history information, prior to time $t$, if you know $S_t$ (that is, if the probability term is conditioned on $S_t$). It does not allow you to ignore variables associated with time $t$ or any future times ($t + 1$, $t + 2$, etc). For instance:*

$$\Pr(S_1 = s_1 | A_1 = a_1, S_0 = s_0) \neq \Pr(S_1 = s_1 | S_0 = s_0)$$

*and*

$$\Pr(S_2 = s_2 | S_4 = s_4, S_1 = s_1) \neq \Pr(S_2 = s_2 | S_1 = s_1).$$

- *When writing the final answers to the problems below (2a–2d), please reorganize your terms and summations in "temporal" order. For instance, instead of presenting your final answer as*

$$\sum_{s_1} p(s_1, a_1, s_2) \pi(s_1, a_1) \sum_{a_0} p(s_0, a_0, s_1) \pi(s_0, a_0)$$

*rewrite it as follows:*

$$\sum_{a_0} \pi(s_0, a_0) \sum_{s_1} p(s_0, a_0, a_1) \pi(s_1, a_1) p(s_1, a_1, s_2).$$

**Question 2 - Problems:**

- **(Question 2a. 3 Points)** What is the probability that the action at time $t = 2$ is $\alpha$ given that the state at time $t = 0$ is $s_0$?

$$\begin{aligned}
\Pr(A_{t=2} = \alpha | S_{t=0} = s_0) &= \sum_{a_0 \in \mathcal{A}} \Pr(A_0 = a_0 | S_0 = s_0) \Pr(A_2 = \alpha | S_0 = s_0, A_0 = a_0) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \Pr(A_2 = \alpha | S_0 = s_0, A_0 = a_0) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}} \Pr(S_1 = s_1 | S_0 = s_0, A_0 = a_0) \Pr(A_2 = \alpha | S_0 = s_0, A_0 = a_0, S_1 = s1) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}} p(s_0, a_0, s_1) \Pr(A_2 = \alpha | S_0 = s_0, A_0 = a_0, S_1 = s1) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}} \Pr(A_1 = a_1 | S_1 = s_1) \Pr(A_2 = \alpha | S_1 = s1, A_1 = a_1) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}} \pi(s_1, a_1) \Pr(A_2 = \alpha | S_1 = s1, A_1 = a_1) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}} \pi(s_1, a_1) \\
&\quad \sum_{s_2 \in \mathcal{S}} \Pr(S_2 = s_2 | S_1 = s_1, A_1 = a_1) \Pr(A_2 = \alpha | S_1 = s1, A_1 = a_1, S_2 = s_2) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}} \pi(s_1, a_1) \sum_{s_2 \in \mathcal{S}} p(s_1, a_1, s_2) \Pr(A_2 = \alpha | S_2 = s_2) \\
&= \sum_{a_0 \in \mathcal{A}} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}} \pi(s_1, a_1) \sum_{s_2 \in \mathcal{S}} p(s_1, a_1, s_2) \pi(s_2, \alpha)
\end{aligned}$$

- **(Question 2b. 3 Points)** What is the probability that the state at time $t = 50$ is $\beta$ given that the state at time $t = 48$ is $s_{48}$?

$$\begin{aligned}
\Pr(S_{t=50} = \beta | S_{t=48} = s_48) &= \sum_{a_{48} \in \mathcal{A}} \Pr(A_{48} = a_{48} | S_{48} = s_{48}) \Pr(S_{50} = \beta | S_{48} = s_{48}, A_{48} = a_{48}) \\
&= \sum_{a_{48} \in \mathcal{A}} \pi(s_{48}, a_{48}) \Pr(S_{50} = \beta | S_{48} = s_{48}, A_{48} = a_{48}) \\
&= \sum_{a_{48} \in \mathcal{A}} \pi(s_{48}, a_{48}) \sum_{S_{49} \in \mathcal{S}} \Pr(S_{49} = s_49 | S_{48} = s_{48}, A_{48} = a_{48}) \\
&\quad \Pr(S_{50} = \beta | S_{48} = s_{48}, A_{48} = a_{48}, S_{49} = s_49) \\
&= \sum_{a_{48} \in \mathcal{A}} \pi(s_{48}, a_{48}) \sum_{S_{49} \in \mathcal{S}} p(s_{48}, a_{48}, s_{49}) \Pr(S_{50} = \beta | S_{49} = s_{49}) \\
&= \sum_{a_{48} \in \mathcal{A}} \pi(s_{48}, a_{48}) \sum_{S_{49} \in \mathcal{S}} p(s_{48}, a_{48}, s_{49}) \sum_{A_{49} \in \mathcal{A}} \Pr(A_{49} = a_49 | S_{49} = s_{49}) \\
&\quad \Pr(S_{50} = \beta | S_{49} = s_{49}, A_{49} = a_{49}) \\
&= \sum_{a_{48} \in \mathcal{A}} \pi(s_{48}, a_{48}) \sum_{S_{49} \in \mathcal{S}} p(s_{48}, a_{48}, s_{49}) \sum_{A_{49} \in \mathcal{A}} \pi(s_{49}, a_{49}) p(s_{49}, a_{49}, \beta)
\end{aligned}$$

- **(Question 2c. 4 Points)** What is the expected reward at time $t = 14$ given that the action at time $t = 14$ is $a_{14}$ and the state at time $t = 13$ is $s_{13}$?

$$\mathbb{E}[R_{t=14} | A_{t=14} = a_{14}, S_{t=13} = s_{13}]$$

$$
\begin{aligned}
&= \sum_{s_{14}\in\mathcal{S}} \Pr(S_{14}=s_{14}|A_{14}=a_{14},S_{13}=s_{13}) \times \mathbb{E}[R_{14}|A_{14}=a_{14},S_{13}=s_{13},S_{14}=s_{14}] \\
&= \sum_{s_{14}\in\mathcal{S}} \mathbb{E}[R_{14}|A_{14}=a_{14},S_{13}=s_{13},S_{14}=s_{14}] \times \Pr(S_{14}=s_{14}|A_{14}=a_{14},S_{13}=s_{13}) \\
&= \sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14}) \frac{\Pr(A_{14}=a_{14}|S_{13}=s_{13},S_{14}=s_{14})\Pr(S_{14}=s_{14}|S_{13}=s_{13})}{\Pr(A_{14}=a_{14}|S_{13}=s_{13})} \\
&= \sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14})\pi(s_{14},a_{14}) \frac{\Pr(S_{14}=s_{14}|S_{13}=s_{13})}{\Pr(A_{14}=a_{14}|S_{13}=s_{13})} \\
&= \sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14})\pi(s_{14},a_{14}) \frac{\sum_{a_{13}\in\mathcal{A}}\Pr(A_{13}=a_{13}|S_{13}=s_{13})\Pr(S_{14}=s_{14}|S_{13}=s_{13},A_{13}=a_{13})}{\sum_{a_{13}\in\mathcal{A}}\Pr(A_{13}=a_{13}|S_{13}=s_{13})\Pr(A_{14}=a_{14}|S_{13}=s_{13},A_{13}=a_{13})} \\
&= \sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14})\pi(s_{14},a_{14}) \frac{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})p(s_{13},a_{13},s_{14})}{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\Pr(A_{14}=a_{14}|S_{13}=s_{13},A_{13}=a_{13})} \\
&= \frac{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14})\pi(s_{14},a_{14})p(s_{13},a_{13},s_{14})}{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\Pr(A_{14}=a_{14}|S_{13}=s_{13},A_{13}=a_{13})} \\
&= \frac{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14})\pi(s_{14},a_{14})p(s_{13},a_{13},s_{14})}{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\sum_{s_{14}\in\mathcal{S}}\Pr(S_{14}=s_{14}|S_{13}=s_{13},A_{13}=a_{13})\Pr(A_{14}=a_{14}|S_{13}=s_{13},A_{13}=a_{13},S_{14}=s_{14})} \\
&= \frac{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14})\pi(s_{14},a_{14})p(s_{13},a_{13},s_{14})}{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\sum_{s_{14}\in\mathcal{S}} p(s_13,a_13,s_14)\Pr(A_{14}=a_{14}|S_{14}=s_{14})} \\
&= \frac{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\sum_{s_{14}\in\mathcal{S}} R(s_{14},a_{14})\pi(s_{14},a_{14})p(s_{13},a_{13},s_{14})}{\sum_{a_{13}\in\mathcal{A}}\pi(s_{13},a_{13})\sum_{s_{14}\in\mathcal{S}} p(s_{13},a_{13},s_{14})\pi(s_{14},a_{14})}
\end{aligned}
$$

- **(Question 2d. 8 Points)** Consider the MDP shown in Figure 1, inspired in the 687-GridWorld. The agent's goal is to reach the state with a coin ($s_5$) while avoiding the states where it dies (states $s_2$ and $s_3$). The agent always starts in $s_0$ and follows the intuitive policy of moving up twice and then moving left. That is, $\pi(s_0, \text{AU}) = 1.0$, $\pi(s_1, \text{AU}) = 1.0$, and $\pi(s_4, \text{AL}) = 1.0$. However, notice that—similarly to the 687-GridWorld—these actions may fail and lead the agent to a different state than expected. States $s_2$, $s_3$, and $s_5$ are terminal states. If colliding with an obstacle, the agent remains in its current state. The arrows in each state indicate the transition dynamics *under the agent's policy*; for instance, $p(s_4, \text{AL}, s_5) = 0.8$ and $p(s_4, \text{AL}, s_3) = 0.2$. Assume the agent is allowed to take at most 3 actions. Even though it follows a reasonable policy, it may accidentally fail and die, ending up in $s2$ or $s3$ rather than in the coin state. What is the probability that the agent will accidentally die within at most 3 steps? Show every step in your derivation.

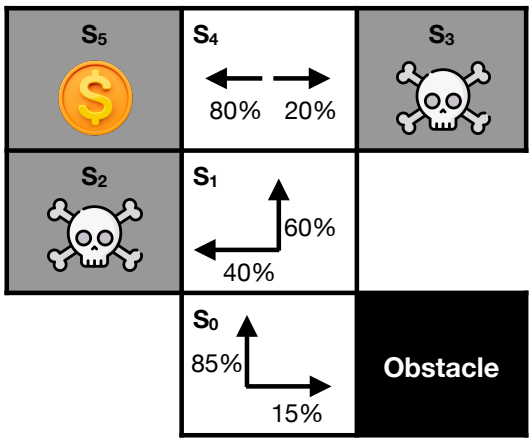


Figure 1: An MDP inspired in the 687-Gridworld. The arrows indicate the transition dynamics under the agent's policy (described in the question).

$$\begin{aligned}\Pr(\text{Die in at most 3 steps}) =& \Pr(\text{Die at } S_2 \text{ in at most 3 steps}) + \Pr(\text{Die at } S_3 \text{ in at most 3 steps}) \\ =& \sum_{t=0}^{3} \Pr(S_t = S_2) + \sum_{t=0}^{3} \Pr(S_t = S_3) \\ =& \Pr(S_{t=0} = S_2) + \Pr(S_{t=1} = S_2) + \Pr(S_{t=2} = S_2) + \Pr(S_{t=3} = S_2) + \\ & \Pr(S_{t=0} = S_3) + \Pr(S_{t=1} = S_3) + \Pr(S_{t=2} = S_3) + \Pr(S_{t=3} = S_3)\end{aligned}$$

Following will be zero:
$\Pr(S_{t=0} = S_2), \Pr(S_{t=1} = S_2), \Pr(S_{t=0} = S_3), \Pr(S_{t=1} = S_3), \Pr(S_{t=2} = S_3) = 0$
Given that it can move up and left only with the following policies: $\pi(s_0, \text{AU}) = 1.0$, $\pi(s_1, \text{AU}) = 1.0$, and $\pi(s_4, \text{AL}) = 1.0$ there is no way we can reach the above states at those timestamps as the agent always starts from state $S_0$. For e.g. there is no way to be at $S_2$ at time t=0.

$$Therefore \Pr(\text{Die in at most 3 steps}) = \Pr(S_{t=2} = S_2) + \Pr(S_{t=3} = S_2) + \Pr(S_{t=3} = S_3) \tag{6}$$

Now calculating $\Pr(S_{t=2} = S_2)$:

$$\begin{aligned}\Pr(S_{t=2} = S_2) =& \sum_{s_0 \in \mathcal{S}} \Pr(S_{t=0} = s_0) \sum_{s_1 \in \mathcal{S}} \Pr(S_{t=1} = s_1 | S_{t=0} = s_0) \Pr(S_{t=2} = s_2 | S_{t=1} = s_1) \\ =& 1 * \sum_{s_1 \in \mathcal{S}} \Pr(S_{t=1} = s_1 | S_{t=0} = s_0) \Pr(S_{t=2} = s_2 | S_{t=1} = s_1)\end{aligned}$$

On expanding the above we find that only the following are valid terms:

$$\begin{aligned}\Pr(S_{t=2} = S_2) =& \Pr(S_{t=1} = s_0 | S_{t=0} = s_0) \Pr(S_{t=2} = s_2 | S_{t=1} = s_0) + \\ & \Pr(S_{t=1} = s_1 | S_{t=0} = s_0) \Pr(S_{t=2} = s_2 | S_{t=1} = s_1) \\ =& (0.15 * 0) + (0.85 * 0.4) \\ =& 0.34\end{aligned}$$

Similarly calculating $\Pr(S_{t=3} = S_2)$:

$$\begin{aligned}\Pr(S_{t=3} = S_2) =& \sum_{s_0 \in \mathcal{S}} \Pr(S_{t=0} = s_0) \sum_{s_1 \in \mathcal{S}} \Pr(S_{t=1} = s_1 | S_{t=0} = s_0) \sum_{s_2 \in \mathcal{S}} \Pr(S_{t=2} = s_2 | S_{t=1} = s_1) \times \\ & \Pr(S_{t=3} = s_2 | S_{t=2} = s_2) \\ =& (0.85 * 0.4 * 0) + (0.85 * 0.6 * 0) + (0.15 * 0.85 * 0.4) + (0.15 * 0.15 * 0) \\ & \text{(Similar to how we expanded the valid terms above)} \\ =& 0.15 * 0.85 * 0.4 \\ =& 0.051\end{aligned}$$

Finally calculating $\Pr(S_{t=3} = S_3)$:

$$\begin{aligned}\Pr(S_{t=3} = S_2) =& 0.85 * 0.6 * 0.2 \\ =& 0.102\end{aligned}$$

Substituting these values in equation 6:

$$\begin{aligned}\Pr(\text{Die in at most 3 steps}) =& 0.34 + 0.051 + 0.102 \\ =& 0.493\end{aligned}$$

3. (**14 Points [Total]**) In class we discussed how reward functions can be used to specify what is the "goal" (or objective) of the agent. We presented three ways in which the reward function can be specified: some are extremely general, but not necessarily easy to define in practice; and some are less general but can be defined more intuitively in real-world problems:

   - The most general formulation of the reward function is given by $d_R$, which specifies an arbitrary distribution over rewards given that the agent is in some state $s$, executes an action $a$, and transitions to some state $s'$.

- Alternatively, in some problems the reward function can be defined in a way that it (deterministically) returns a scalar number based on $s$, $a$, and $s'$. That is, $R$ can be defined as a function of the form $R(s, a, s')$.
- Finally, an even simpler formulation of reward functions can be constructed that depends only on the state of the agent ($s$) and the action that it executed ($a$). That is, $R$ can be defined as a function of the form $R(s, a)$.

**(Question 3a. 10 Points)** In some problems, the reward depends only on where the agent was and the state that it reached, but not the particular action it took to reach that state. For example, consider a runner in the Olympics. As long as they were at the starting line and transitioned to winning the race, the particular action they took between these moments is irrelevant. Show, from "first principles", *step by step*, how to derive an equation for $R(s, s')$ in terms of $d_R$. Recall that, by definition, $R(s, s') = \mathbb{E}[R_t|S_t = s, S_{t+1} = s']$.

$$
\begin{aligned}
R(s,s') =& \mathbb{E}[R_t|S_t = s, S_{t+1} = s'] \\
=& \sum_{r\in\mathcal{R}} r * \Pr(R_t = r|S_t = s, S_{t+1} = s') \\
=& \sum_{a_t\in\mathcal{A}} \Pr(A_t = a_t|S_t = s, S_{t+1} = s') \sum_{r\in\mathcal{R}} r * \Pr(R_t = r|S_t = s, S_{t+1} = s', A_t = a_t) \\
=& \sum_{a_t\in\mathcal{A}} \Pr(A_t = a_t|S_t = s, S_{t+1} = s') \sum_{r\in\mathcal{R}} r * d_R \\
=& \sum_{a_t\in\mathcal{A}} \frac{\Pr(S_{t+1} = s'|S_t = s, A_t = a_t)\Pr(A_t = a_t|S_t = t)}{\Pr(S_{t+1} = s'|S_t = s)} \sum_{r\in\mathcal{R}} r * d_R \\
=& \sum_{a_t\in\mathcal{A}} \frac{p(s, a_t, s')\pi(s, a_t)}{\Pr(S_{t+1} = s'|S_t = s)} \sum_{r\in\mathcal{R}} r * d_R \\
=& \sum_{a_t\in\mathcal{A}} \frac{p(s, a_t, s')\pi(s, a_t)}{\sum_{a_t\in\mathcal{A}} \Pr(A_t = a_t|S_t = s)\Pr(S_{t+1} = s'|S_t = s, A_t = a_t)} \sum_{r\in\mathcal{R}} r * d_R \\
=& \sum_{a_t\in\mathcal{A}} \frac{p(s, a_t, s')\pi(s, a_t)}{\sum_{a_t\in\mathcal{A}} \pi(s, a_t)p(s, a_t, s')} \sum_{r\in\mathcal{R}} r * d_R \qquad (7)
\end{aligned}
$$

**(Question 3b. 4 Points)** Still regarding the example above, notice that *(i)* the reward does not depend on the particular action that was taken by the agent; and *(ii)* the reward is *deterministic*. For example, if the agent was at the starting line and then transitioned to winning the race, it will deterministically win a medal. To model this, suppose $d_R$ is deterministic and independent of the action; that is, $d_R(c; s, a_i, s') := \Pr(R_t = c|S_t = s, A_t = a_i, S_{t+1} = s') = 1$ for all actions $a_i \in \mathcal{A}$, and where $c$ is a constant. How does this assumption change/simplify the expression for $R(s, s')$ you derived in the previous question? Show your derivation step-by-step.

- **Hint**: *In this question, "Useful property #1" and "Useful property #2", shown in Section 2, may be useful.*

Given that $d_R$ is deterministic and independent of the action; that is, $d_R(c; s, a_i, s') := \Pr(R_t = c|S_t = s, A_t = a_i, S_{t+1} = s') = 1$ It means that only for reward r=c will $d_R$ be 1. For all other values for r, the reward will be 0.

This means $\sum_{r\in\mathcal{R}} r * d_R = c$

So in equation 7, we can cancel the terms $p$ and $\pi$ from the numerator and denominator as the reward is independent of the action taken and that it is deterministic. No matter whichever actions we take, reward will be received only when the race is won.
So,

$$
\begin{aligned}
R(s,s') &= \sum_{a_t\in\mathcal{A}} \frac{p(s, a_t, s')\pi(s, a_t)}{\sum_{a_t\in\mathcal{A}} \pi(s, a_t)p(s, a_t, s')} \sum_{r\in\mathcal{R}} r * d_R \\
&= \sum_{r\in\mathcal{R}} r * d_R \\
&= c
\end{aligned}
$$

4. (**5 Points**) Suppose you wish to identify an optimal *deterministic* policy for a variant of the 687-GridWorld MDP that consists of 2 states and 5 actions. One approach to finding the optimal policy is through a brute-force search. This involves evaluating the performance, $J(\pi)$, for *all* possible deterministic policies $\pi$ and selecting the one with the highest expected return. Assume only one optimal policy exists and that it takes your computer 300 seconds to evaluate $J(\pi)$ for any policy $\pi$. Suppose your algorithm saves all possible deterministic policies in a list, shuffles it, and then performs brute force on the elements of this list. Importantly, because you are performing such a policy optimization process on a cluster, your brute force approach can run for only 40 minutes. When this time is reached, your algorithm returns the best policy out of the ones it managed to evaluate. What is the probability that this approach would return and identify the optimal policy?

   Given 5 actions and 2 states, there can be a total of $5^2 = 25$ policies.
   Each policy takes 300 seconds (5 mins) to run.
   So total policies that can be run in 40 mins = 8
   Therefore the probability that it returns an optimal policy = $8/25 = 32\%$

5. (**6 Points**) To fully specify an MDP, we must define $\mathcal{S}$, $\mathcal{A}$, $p$, $R$, $d_0$, and $\gamma$. We also need to encode the agent's policy, $\pi$, which is $\Pr(A_t = a \mid S_t = s)$. That is, the policy represents the action $a$ the agent should take when in state $s$. However, in many real-world applications, it may not be feasible to store the policy in tables, as discussed in class, especially when the number of states is extremely large or when states and actions may be continuous. Consider a robot tasked with moving to various locations on campus to clean them up. The state might consist of continuous variables, such as its $x$ and $y$ coordinates and the battery level $b$. The robot's actions could combine continuous values (the robot's desired velocity $v$ and the angle $\alpha$ it wishes to turn to) and discrete values (whether or not to turn on its vacuum cleaner, indicated by a binary variable $c$). Suppose someone manually controls the robot using a remote control and logs the value of these variables for one hour. Describe in detail how you would use a machine learning technique to process this data (i.e., samples of the states the robot was in and the actions that it took) to encode/represent a corresponding policy, $\pi$. Specify which ML algorithm you would use, its inputs and outputs, and how it would be trained to produce actions similar to those taken by the human operator.

   For this scenario where we have continuous variables like the coordinates of the robot, like mentioned it is not feasible to store the data in tables.
   We can use neural nets for this scenarios. Given that we already have the data collected for 1 hour, we have information about the coordinates and the routes it took, along with the set of actions it took for the corresponding inputs.

   Input: $f(x, y, b)$ where $x$ : robot's x-coordinate, $y$ : robot's y-coordinate, $b$ : robot's battery percentage.
   These make up the input vector for the model.
   Output: $g(v, \alpha, c)$ where $v$ : robot's velocity, $\alpha$ : robot's angle of turning, $c$ : turn vaccum cleaner on/off.
   Hidden Layers: It allows the neural network to learn complex relationships between the state variables (e.g., coordinates and battery level) and the actions it needs to take (e.g., desired velocity and turning angle). As data flows through the hidden layers, the network can extract increasingly abstract features. The first hidden layer might respond to specific states, like being low on battery, while subsequent layers might combine these features to form more complex decisions, like determining the best route to take based on multiple factors (e.g., distance, battery, and obstacles).

   We split the data collected into training and testing data and train and test our models suitably, preprocess the data and normalize it .
   We set the hyperparameters such as number of hidden layers, activation function (e.g ReLu), bias, weights. We also choose a loss function (M.S.E) and optimiser (e.g. Adam) to evaluate and minimize the losses
   We train our model over multiple epochs, backpropagate to adjust weights and test with the test data to fine tune the hyper-parameters suitably to obtain optimal policies and results.

6. (**7 Points**) We refer to the discounted sum of rewards, $\sum_{t=0}^{\infty} \gamma^t R_t$, as the *return*. Let M be an MDP that has two optimal policies. Can their expected returns differ? If so, give an example of such an MDP, defining it precisely (i.e., specifying its state and action space, transition function, etc.). If not, explain why. Secondly, can the *variance* of their returns differ? If so, give an example. If not, explain why.

   - **Hint**: *In this question (and others), you should be able to design sample MDPs that are very small—MDPs*

*with very few states and actions and a relatively simple reward function. If the MDPs you are creating feel overly complex, you are probably overthinking.*

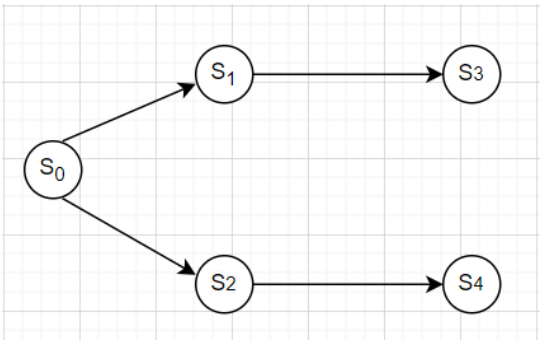


Figure 2: An MDP with $S_0$ as initial state, $S_1$, $S_2$ as intermediate states and $S_3$, $S_4$ are the terminal states

Let the above figure represent out MDP where:
$R(S_0, a_1) = 0.5$, $R(S_0, a_2) = 1$, $R(S_1, a_1) = 0.5$ and $R(S_2, a_1) = 0$
$p(S_0, a_1, S_1) = 1$, $p(S_0, a_2, S_1) = 1$, $p(S_1, a_1, S_3) = 1$ and $p(S_2, a_1, S_4) = 1$
Let there be 2 deterministic policies $\pi_1$ and $\pi_2$ such that:
$\pi_1$:= $\pi(S_0, a_1) = 1$, $\pi(S_1, a_1) = 1$
$\pi_2$:= $\pi(S_0, a_2) = 1$, $\pi(S_2, a_2) = 1$
*(Note that rest all values that are not mentioned are 0.)*
Let $\gamma = 1$.
Now,
$\mathbb{E}[R_{\pi_1}] = (0.5 * \gamma^0) + (0.5 * \gamma^1) = (0.5 * 1^0) + (0.5 * 1^1) = 1$
Similarly,
$\mathbb{E}[R_{\pi_2}] = (1 * \gamma^0) + (0 * \gamma^1) = (1 * 1^0) + (0 * 1^1) = 1$
We can observe that as $\mathbb{E}[R_{\pi_1}] = \mathbb{E}[R_{\pi_2}] = 1$, both $\pi_1$ and $\pi_2$ are optimal policies.

So **No**, if there are 2 optimal policies, then the expected returns cannot differ. As optimal policies by definition means that the policy at which we get maximum expected rewards. We cannot have 2 maximum expected rewards, different policies may yield to the same maximum expected rewards however as shown in the above example.
The variance of their returns **can differ** however. To prove this we can use the above example to calculate the variance.
We know that, $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. Using this,
$Var(\pi_1) = (0.5 - 1)^2 + (0.5 - 1)^2 = 0.5$
$Var(\pi_2) = (1 - 1)^2 + (0 - 1)^2 = 1$
We can observe that the variance of the returns for both these policies are different.

7. (**5 Points**) Non-stationary MDPs have the property that either their transition or reward functions can change over time. Consider a self-driving car. Describe one possible representation of its state space and action space. Would this state representation lead to a stationary transition function? Explain why. Now describe a new sensor that could be added to the car and whose readings *could* be included in the state. Explain whether the problem would be stationary or non-stationary depending on whether the agent can access that sensor's readings. Recall that if the transition function is non-stationary, $\Pr(S_{t+1} = s' | S_t = s, A_t = a) \neq \Pr(S_{t+k+1} = s' | S_{t+k} = s, A_{t+k} = a)$, for some $k > 0$. Intuitively, when facing the same state $s$ and executing the same action $a$, the distribution over the agent's next states would differ between time $t$ and time $t + k$. Finally, assuming the agent does have access to the sensor's readings (i.e., they are included in the agent's state), explain how this could impact the stationarity of the initial state distribution, $d_0$; that is, the distribution of states in which the agent might be initialized at the beginning of every new episode.

For a self driving car, let the following be its state space: $S := (X, Y, v, d, b)$ where X,Y are the x and y coordinates of the car, v is the velocity, d is the direction and b is the battery % of the car. And let the following be its action space: $A := (\alpha, \beta, \gamma, \delta)$ where $\alpha$ is the , $\beta$ is the brake, $\gamma$ is steering direction (left/right/U-Turn) and $\delta$ is horn.

This state representation **does not** lead to stationary transition function. For instance if there is some traffic or construction or any new obstacle, then the car will have to take a new direction/route or the car may run out of battery. Or let's say that if the car tyres wear out then the car may skid into a new direction all together thereby transitioning to a different state than expected.
Now let's say that we have a sensor that measures how much the car tyres have weared out. Now if we have access to the readings of the sensor then, the problem may or may not be stationary. Say that if rest of the environment factors such as traffic, etc. do not change then, given the sensor readings it MAY be possible to may the transition function stationary but the rewards and the initial distribution function will still be non stationary as tyres wearing out change the dynamics of driving the car. So generally speaking, even if we have access to sensor readings, given how the environments also changes dynamically, the problem would still remain non stationary as although we may have access to the tyre wear reading and do not need to store previous reading for it but it doesn't guarantee stationary MDP. If we dont have access to the sensor readings then the problem is definitely non stationary, as we need to store the previous values of tyre wearout as well and the environment factors will still be dynamic leading to non stationary initial distribution, transitions and rewards.

# Part Two: Programming (45 Points Total)

Consider the following MDP:

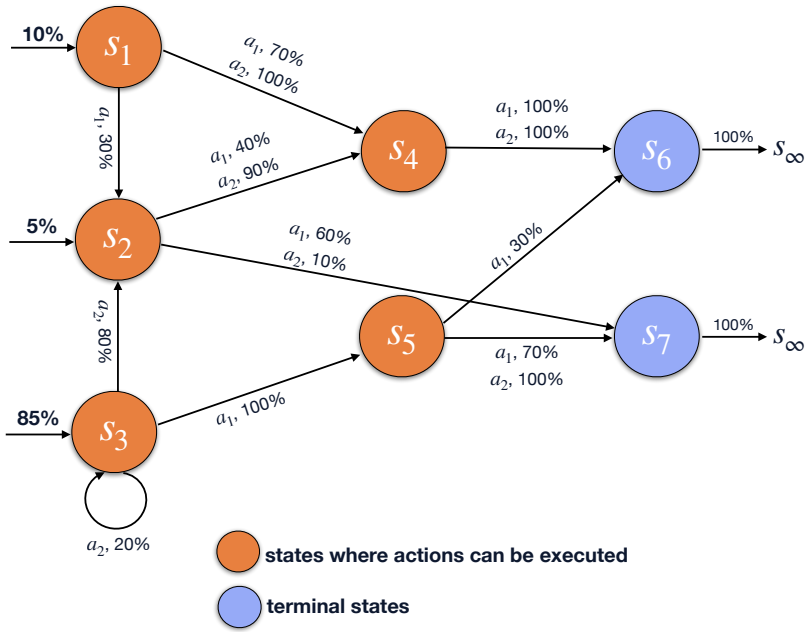


Figure 3: An MDP.

Assume the following initial state distribution, $d_0$, and transition function, $p$, where all transition probabilities not indicated in the table below are 0.

| | | |
|---|---|---|
| $d_0(s_1) = 0.1$ | $d_0(s_2) = 0.05$ | $d_0(s_3) = 0.85$ |

| | |
|---|---|
| $p(s_1, a_1, s_4) = 0.7$ | $p(s_1, a_1, s_2) = 0.3$ |
| $p(s_1, a_2, s_4) = 1.0$ | |
| $p(s_2, a_1, s_4) = 0.4$ | $p(s_2, a_1, s_7) = 0.6$ |
| $p(s_2, a_2, s_4) = 0.9$ | $p(s_2, a_2, s_7) = 0.1$ |
| $p(s_3, a_1, s_5) = 1.0$ | |
| $p(s_3, a_2, s_3) = 0.2$ | $p(s_3, a_2, s_2) = 0.8$ |
| $p(s_4, a_1, s_6) = 1.0$ | |
| $p(s_4, a_2, s_6) = 1.0$ | |
| $p(s_5, a_1, s_6) = 0.3$ | $p(s_5, a_1, s_7) = 0.7$ |
| $p(s_5, a_2, s_7) = 1.0$ | |

Assume the following reward function:

| | |
|---|---|
| $R(s_1, a_1) = 4$ | $R(s_1, a_2) = 3$ |
| $R(s_2, a_1) = 6$ | $R(s_2, a_2) = 1$ |
| $R(s_3, a_1) = -1$ | $R(s_3, a_2) = -10$ |
| $R(s_4, a_1) = 14$ | $R(s_4, a_2) = 8$ |
| $R(s_5, a_1) = -2$ | $R(s_5, a_2) = 2$ |

Finally, consider the following stochastic policy, $\pi$:

| | |
|---|---|
| $\pi(s_1, a_1) = 0.3$ | $\pi(s_1, a_2) = 0.7$ |
| $\pi(s_2, a_1) = 0.5$ | $\pi(s_2, a_2) = 0.5$ |
| $\pi(s_3, a_1) = 0.6$ | $\pi(s_3, a_2) = 0.4$ |
| $\pi(s_4, a_1) = 0.25$ | $\pi(s_4, a_2) = 0.75$ |
| $\pi(s_5, a_1) = 0.6$ | $\pi(s_5, a_2) = 0.4$ |

---

**(Question 1. 15 Points)** Find an *analytic, closed-form* expression for

$$J(\pi) \quad = \quad \mathbb{E}\left[\sum_{t=0}^{1} \gamma^t R_t\right]$$

as a function of $\gamma$. To do this, you *may* choose do it from "first principles", by repeatedly using the properties of probability distributions and expected values introduced in Section 2. If you do not wish to derive a closed-form expression for $J(\pi)$ this way, you are also allowed to write it directly as a function of $d_0$, $p$, $\pi$, and $R$, similarly to the final equation described in the "**Example problem**" introduced in Section 2. You must present your derivation step-by-step.

- **Hint:** *Start by applying the property of linearity of expectation to the definition of $J(\pi)$ and then derive separate equations for each of the resulting terms.*

Your final answer **must** be in the form of $J(\pi) = c_1 + \gamma c_2$, where each $c_i$ is a real-valued constant. To obtain an equation of this form, start from the analytic, closed-form expression you derived earlier and plug in the particular transition probabilities, rewards, and policy probabilities specified above. You must present your final answer in the form discussed above, showing your work step-by-step.

$$\mathbb{E}\left[\sum_{t=0}^{1} \gamma^t R_t\right] = \mathbb{E}[R_0] + \gamma \mathbb{E}[R_1] \quad \text{(By Linearity of Expectation and taking the constant 'gamma' out)} \tag{8}$$

Evaluating $\mathbb{E}[R_0]$:

$$
\begin{aligned}
\mathbb{E}[R_0] &= \sum_{r_0 \in \mathcal{R}_0} r_0 \Pr(R_0 = r_0) \\
&= \sum_{r_0 \in \mathcal{R}_0} r_0 \sum_{s_0 \in \mathcal{S}_0} \Pr(S_0 = s_0) \Pr(R_0 = r_0 | S_0 = s_0) \\
&= \sum_{r_0 \in \mathcal{R}_0} r_0 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \Pr(R_0 = r_0 | S_0 = s_0) \\
&= \sum_{r_0 \in \mathcal{R}_0} r_0 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \Pr(A_0 = a_0 | S_0 = s_0) \Pr(R_0 = r_0 | S_0 = s_0, A_0 = a_0) \\
&= \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \sum_{r_0 \in \mathcal{R}_0} r_0 \Pr(R_0 = r_0 | S_0 = s_0, A_0 = a_0) \\
&= \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) R(s_0, a_0) \\
&= d_0(s_1) * \pi(s_1, a_1) * R(s_1, a_1) + d_0(s_1) * \pi(s_1, a_2) * R(s_1, a_2) + \\
&\quad d_0(s_2) * \pi(s_2, a_1) * R(s_2, a_1) d_0(s_2) * \pi(s_2, a_2) * R(s_2, a_2) + \\
&\quad d_0(s_3) * \pi(s_3, a_1) * R(s_3, a_1) + d_0(s_3) * \pi(s_3, a_2) * R(s_3, a_2) \\
&\quad \textit{(As we can start only from states } S_0, \ S_1, \ S_2\textit{)} \\
&= (0.1 * 0.3 * 4) + (0.1 * 0.7 * 3) \\
&\quad (0.05 * 0.5 * 6) + (0.05 * 0.5 * 1) \\
&\quad (0.85 * 0.6 * -1) + (0.85 * 0.4 * -10) \\
&= -3.405
\end{aligned}
$$

Evaluating $\mathbb{E}[R_1]$:

$$
\begin{aligned}
\mathbb{E}[R_1] &= \sum_{r_1 \in \mathcal{R}_1} r_1 \Pr(R_1 = r_1) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} \Pr(S_0 = s_0) \Pr(R_1 = r_1 | S_0 = s_0) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \Pr(R_1 = r_1 | S_0 = s_0) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \Pr(A_0 = a_0 | S_0 = s_0) \Pr(R_1 = r_1 | S_0 = s_0, A_0 = a_0) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \Pr(R_1 = r_1 | S_0 = s_0, A_0 = a_0) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}_1} \Pr(S_1 = s_1 | S_0 = s_0, A_0 = a_0) \Pr(R_1 = r_1 | S_0 = s_0, A_0 = a_0, S_1 = s_1) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}_1} p(s_0, a_0, s_1) \Pr(R_1 = r_1 | S_1 = s_1) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}_1} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}_1} \Pr(A_1 = a_1 | S_1 = s_1) \Pr(R_1 = r_1 | S_1 = s_1, A_1 = a_1) \\
&= \sum_{r_1 \in \mathcal{R}_1} r_1 \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}_1} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}_1} \pi(s_1, a_1) \Pr(R_1 = r_1 | S_1 = s_1, A_1 = a_1) \\
&= \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}_1} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}_1} \pi(s_1, a_1) \sum_{r_1 \in \mathcal{R}_1} r_1 \Pr(R_1 = r_1 | S_1 = s_1, A_1 = a_1) \\
&= \sum_{s_0 \in \mathcal{S}_0} d_0(s_0) \sum_{a_0 \in \mathcal{A}_0} \pi(s_0, a_0) \sum_{s_1 \in \mathcal{S}_1} p(s_0, a_0, s_1) \sum_{a_1 \in \mathcal{A}_1} \pi(s_1, a_1) R(s_1, a_1) \\
&= d_0(S_1)\pi(S_1, a_1)p(S_1, a_1, S_4)\pi(S_4, a_1)R(S_4, a_1) + \\
&\quad d_0(S_1)\pi(S_1, a_1)p(S_1, a_1, S_4)\pi(S_4, a_2)R(S_4, a_2) + \\
&\quad d_0(S_1)\pi(S_1, a_1)p(S_1, a_1, S_2)\pi(S_2, a_1)R(S_2, a_1) + \\
&\quad d_0(S_1)\pi(S_1, a_1)p(S_1, a_1, S_2)\pi(S_2, a_2)R(S_2, a_2) + \\
&\quad . \\
&\quad . \\
&\quad . \\
&\quad d_0(S_2)\pi(S_2, a_1)p(S_2, a_1, S_4)\pi(S_4, a_1)R(S_4, a_1) + \\
&\quad d_0(S_2)\pi(S_2, a_1)p(S_2, a_1, S_4)\pi(S_4, a_2)R(S_4, a_2) + \\
&\quad . \\
&\quad . \\
&\quad . \\
&\quad d_0(S_3)\pi(S_3, a_2)p(S_3, a_2, S_2)\pi(S_2, a_2)R(S_2, a_1) \\
&= 1.63995
\end{aligned}
$$

Substituting values of $\mathbb{E}[R_0]$ and $\mathbb{E}[R_1]$ in equation 8,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=0}^{1} \gamma^t R_t\right] &= \mathbb{E}[R_0] + \gamma \mathbb{E}[R_1] \\
&= -3.405 + \gamma * 1.63995 \qquad (9)
\end{aligned}
$$

**(Question 2 - *Programming Question*. 30 Points [Total])**

In this question, you will write a program to estimate $J(\pi)$ by simulating many of the possible outcomes (returns) that might result from running $\pi$ on the previously-defined MDP. Each simulation will produce a particular sequence of

states, actions, and rewards, and, thus, a particular discounted return. Since $J(\pi)$ is defined as the *expected* discounted return, you can construct an estimate of $J(\pi)$, $\hat{J}(\pi)$, by averaging the discounted returns observed across $N$ simulations.

In particular:

- To run one simulation (or episode, or trial), you should follow the "Agent-Environment Interaction" procedure introduced in Lecture #3.
- Start by creating a function called *runEpisode* that takes as input a policy and a value of $\gamma$, and that returns the empirical discounted return resulting from that episode.
- Let $G^i$ be the discounted return of the $i^{th}$ episode. You will estimate $J(\pi)$ by computing $\hat{J}(\pi) := \frac{1}{N}\sum_{i=1}^{N} G^i$.

**(Question 2a. 8 Points)**. Construct $\hat{J}(\pi)$ by running 150,000 simulations/episodes. You should then create a graph where the $x$ axis shows the number of episodes, and the $y$ axis presents the estimate $\hat{J}(\pi)$ constructed based on all episodes executed up to that point. As an example, the point $x = 100$ in this graph should have as its corresponding $y$ coordinate the estimate $\hat{J}(\pi)$ built using the discount returns from the first 100 simulations. Your plot should include the results of this simulation up to 150,000 simulations/episodes. You should use $\gamma = 0.9$.

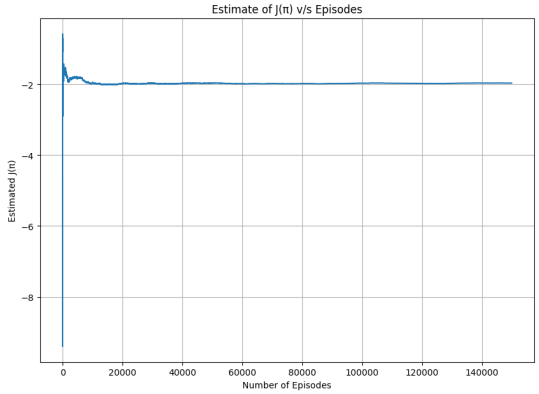


Figure 4: Q(2a.) Expected Discounted Returns v/s Number of Simulations

**(Question 2b. 5 Points)**. Report the average discounted return, as well as its variance, at the end of this process. In other words, report the value of $\hat{J}(\pi)$ after executing 150,000 episodes, as well as the variance of the return of all simulations used to compute this last estimate of the return of the policy.

```
Estimated J(π) after 150000 episodes: -1.96
Variance of the returns: 46.04
```

Figure 5: Q(2b.) Expected Discounted Returns and Variance

**(Question 2c. 5 Points)**. Estimate $\hat{J}(\pi)$ using different discount rates: $\gamma \in \{0.25, 0.5, 0.75, 0.99\}$. Compare these estimates with their true values, computed according to the closed-form solution for $J(\pi)$ found in the first part of this question. These values should approximately match (i.e., $J(\pi) \approx \hat{J}(\pi)$).
I used the formula derived in equation 9 and found the returns using code and printed both the values for different values of $\gamma$. Please see code for more details.

```
Gamma:  0.25
Expected J(π) using formula:  -2.9950124999999996
Computed J(π) using code:  -3.0124566666666666
Gamma:  0.5
Expected J(π) using formula:  -2.585025
Computed J(π) using code:  -2.58751
Gamma:  0.75
Expected J(π) using formula:  -2.1750374999999997
Computed J(π) using code:  -2.19154
Gamma:  0.99
Expected J(π) using formula:  -1.7814495
Computed J(π) using code:  -1.8093945333333332
```

Figure 6: Q(2c.) Expected Discounted Returns with formula and code computation

**(Question 2d. 12 Points)**. Next, you will use your *runEpisode* function to estimate the performance of *different* policies (other than the one we introduced/proposed) in order to search for the policy with the highest performance. You may use any optimization method you want to implement this step (even, e.g., brute force search). To simplify this process, you *should* restrict your search to deterministic policies. You should use $\gamma = 0.75$. Describe how the policy search method you used works. Report the best policy, $\hat{\pi}^*$, identified by the process above, and present its estimated performance, $\hat{J}(\hat{\pi}^*)$, computed using $N = 350,000$ simulations. Here rows indicate the state from $S_1$ to $S_7$

```
Best Policy (binary representation for actions in non-terminal states):
[[0. 1.]
 [0. 1.]
 [1. 0.]
 [1. 0.]
 [0. 1.]
 [0. 0.]
 [0. 0.]]
Maximum Expected Reward: 2.2942642857142856
```

Figure 7: Q(2d.) Optimal policy and expected discounted return

and the columns represent the actions $a_1$ and $a_2$. So,
Estimated performance $\hat{J}(\hat{\pi}^*) = 2.2942642857142856$
Optimal policy:

| | |
|---|---|
| $\pi(s_1, a_1) = 0$ | $\pi(s_1, a_2) = 1$ |
| $\pi(s_2, a_1) = 0$ | $\pi(s_2, a_2) = 1$ |
| $\pi(s_3, a_1) = 1$ | $\pi(s_3, a_2) = 0$ |
| $\pi(s_4, a_1) = 1$ | $\pi(s_4, a_2) = 0$ |
| $\pi(s_5, a_1) = 0$ | $\pi(s_5, a_2) = 1$ |

Table 1: Optimal Policy