Homework Set 1 - Solution

Problem 1: For a vector $\mathbf{x} \in \mathbb{R}^n$, recall the definitions of $\|\mathbf{x}\|_\infty$, $\|\mathbf{x}\|_1$, and $\|\mathbf{x}\|_2$.

1. Prove that $\|\mathbf{x}\|_\infty$ satisfies the properties of a norm over a vector space. In particular, show that:

   - $\|\mathbf{x}\|_\infty \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, and $\|\mathbf{x}\|_\infty = 0$ if an only if $\mathbf{x} = \mathbf{0}$.
     Answer:

     Proof. Proof for $||\mathbf{x}||_\infty \geq 0$

     $$\begin{aligned} & \forall i \ |x_i| \geq 0 \\ \implies & \max_i |x_i| \geq 0 \\ \iff & ||\mathbf{x}||_\infty \geq 0 \end{aligned}$$ □

     Proof. Proof for $||\mathbf{x}||_\infty = 0 \iff \mathbf{x} = 0$

     $$\begin{aligned} & ||\mathbf{x}||_\infty = 0 \\ \iff & \max_i |x_i| = 0 \\ \iff & x_i = 0 \ \forall i && \text{as } \forall i \ |x_i| \leq 0 \text{ and } |x_i| \geq 0 \\ \iff & \mathbf{x} = 0 && \square \end{aligned}$$

   - $\|\alpha\mathbf{x}\|_\infty = |\alpha| \, \|\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.
     Answer:

     Proof.

     $$\begin{aligned} ||\alpha x||_\infty &= \max_i |\alpha x_i| \\ &= \max_i |\alpha||x_i| \\ &= |\alpha| \max_i |x_i| && \text{as } |\alpha| \text{ is independent of } i \\ &= |\alpha| ||\mathbf{x}||_\infty && \square \end{aligned}$$

   - $\|\mathbf{x}_1 + \mathbf{x}_2\|_\infty \leq \|\mathbf{x}_1\|_\infty + \|\mathbf{x}_2\|_\infty$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$.
     Answer:

     Proof.

     $$\begin{aligned} ||\mathbf{x}_1 + \mathbf{x}_2||_\infty &= \max_i (|\mathbf{x}_{1_i} + \mathbf{x}_{2_i}|) \\ &\leq \max_i (|\mathbf{x}_{1_i}| + |\mathbf{x}_{2_i}|) && \text{by triangle inequality} \\ &\leq \max_i (|\mathbf{x}_{1_i}|) + \max_i (|\mathbf{x}_{1_i}|) && \text{as } |\mathbf{x}_{1_j}| + |\mathbf{x}_{2_j}| \leq \max_i (|\mathbf{x}_{1_i}|) + \max_i (|\mathbf{x}_{2_i}|) \ \forall j \\ &= ||\mathbf{x}_1||_\infty + ||\mathbf{x}_2||_\infty && \square \end{aligned}$$

2. Prove that

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 .$$

Answer:

Proof. Proof for $||\mathbf{x}||_\infty \leq ||\mathbf{x}||_2$.

Let $j = \arg\max_i |\mathbf{x}_i|$

$$\begin{aligned} ||\mathbf{x}||_\infty^2 &= |x_j|^2 \\ &\leq |x_j|^2 + \sum_{i \neq j} |x_i|^2 \\ &= ||\mathbf{x}||_2^2 \\ \implies ||\mathbf{x}||_\infty &\leq ||\mathbf{x}||_2 \qquad \text{as } f : x \mapsto x^2 \text{ is monotonically increasing for } x \geq 0 \end{aligned}$$ □

Proof. Proof for $||\mathbf{x}||_2 \leq ||\mathbf{x}||_1$.

$$\begin{aligned} ||\mathbf{x}||_1^2 &= (\sum_i |x_i|)^2 \\ &= \sum_i |x_i|^2 + \sum_i \sum_{j \neq i} |x_i||x_j| \\ &\geq \sum_i |x_i|^2 \\ &= ||\mathbf{x}||_2^2 \\ \implies ||\mathbf{x}||_2 &\leq ||\mathbf{x}||_1 \end{aligned}$$ □

3. Prove the following special case of the Holder's inequality for two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$:

$$| < \mathbf{x}_1, \mathbf{x}_2 > | \leq \|\mathbf{x}_1\|_\infty \|\mathbf{x}_2\|_1 .$$

Answer:

Proof.

$$\begin{aligned} | < \mathbf{x}_1, \mathbf{x}_2 > | &= \left| \sum_i x_{1_i} x_{2_i} \right| \\ &\leq \sum_i |x_{1_i} x_{2_i}| \qquad \text{by triangle inequality} \\ &= \sum_i |x_{1_i}||x_{2_i}| \\ &\leq \sum_i \max_j |x_{1_j}||x_{2_i}| \\ &= \max_j |x_{1_j}| \sum_i |x_{2_i}| \\ &= \|\mathbf{x}_1\|_\infty \|\mathbf{x}_2\|_1 \end{aligned}$$ □

Problem 2: Consider two discrete random variables X and Y with probability distributions $\mathbb{P}[X = x]$ defined for all $x \in \mathcal{X}$, $\mathbb{P}[Y = y]$ defined for all $y \in \mathcal{Y}$, and $\mathbb{P}[X = x, Y = y]$ defined for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Prove the law of total expectation (tower rule) for X and Y, i.e., show that the following holds:

$$\mathbb{E}_Y \left[\mathbb{E}_X \left[X|Y\right]\right] = \mathbb{E}_X \left[X\right],$$

where the subscripts denote the randomness over which the expectations are computed.

Answer:

Proof.

$$\begin{aligned}
\mathbb{E}_Y \left[\mathbb{E}_X \left[X|Y\right]\right] &= \sum_{y\in\mathcal{Y}} \left(\sum_{x\in\mathcal{X}} x\mathbb{P}[X = x|Y = y] \right) \mathbb{P}[Y = y] \\
&= \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} x\mathbb{P}[X = x, Y = y] \\
&= \sum_{x\in\mathcal{X}} x \left(\sum_{y\in\mathcal{Y}} \mathbb{P}[X = x, Y = y] \right) \\
&= \sum_{x\in\mathcal{X}} x\mathbb{P}[X = x] \\
&= \mathbb{E}_X \left[X\right]
\end{aligned}$$

□

Note: We use the following in the proof steps above:

- Expectation of a discrete random variable: $\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x\mathbb{P}[X = x]$
- Relationship between conditional probability and joint probability for discrete random variables:

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x|Y = y) \cdot \mathbb{P}(Y = y) = \mathbb{P}(Y = y|X = x) \cdot \mathbb{P}(X = x)$$

Problem 3: Consider a discrete-time, time-homogeneous Markov chain $(X_t)_{t=0}^{T}$ that satisfies the Markov property, i.e.,

$$\mathbb{P}[X_{t+1}|X_0, X_1, \ldots, X_t] = \mathbb{P}[X_{t+1}|X_t]\,, \qquad \forall t \in \mathbb{N}_0.$$

1. Prove the following holds:

$$\mathbb{P}[X_{t+1}, X_{t+2}, \ldots, X_{t+m}|X_0, X_1, \ldots, X_t] = \mathbb{P}[X_{t+1}, X_{t+2}, \ldots, X_{t+m}|X_t]\,, \quad \forall t \in \mathbb{N}_0\,,\ \forall m \in \mathbb{N}.$$

Answer:

Proof.

$$\begin{aligned}
\mathbb{P}(X_{t+1}, \ldots, X_{t+m}|X_0, \ldots, X_t) &= \mathbb{P}(X_{t+1}|X_0, X_1, \ldots, X_t) \cdot \mathbb{P}(X_{t+2}|X_0, X_1, \ldots, X_t, X_{t+1}) \ldots \\
&\quad \ldots \mathbb{P}(X_{t+m}|X_0, X_1, \ldots, X_t, X_{t+1}, X_{t+2}, \ldots, X_{t+m-1}) \quad (\#1) \\
&= \mathbb{P}(X_{t+1}|X_t) \cdot \mathbb{P}(X_{t+2}|X_{t+1}) \ldots \mathbb{P}(X_{t+m}|X_{t+m-1}) \quad (\#2) \\
&= \mathbb{P}(X_{t+1}|X_t) \cdot \mathbb{P}(X_{t+2}|X_{t+1}, X_t) \ldots \mathbb{P}(X_{t+m}|X_{t+m-1}, X_t) \quad (\#3) \\
&= \mathbb{P}(X_{t+1}, \ldots, X_{t+m}|X_t) \quad (\#4) \\
&\forall t \in \mathbb{N}_0 \quad \forall m \in \mathbb{N}.
\end{aligned}$$

(#1) is obtained by resolving the joint-probability distribution term into individual terms, conditioned on the respective dependent random variables. (#2) is obtained by applying the Markov property. (#3) is obtained by introducing X_t into the conditional probabilities, which is equivalent to the expression in (#2). Finally, (#3) is obtained by sequentially combining the conditional probability terms into joint probabilities, from $t+m$ to $t+1$. □

2. Prove the following holds:

$$\mathbb{P}[X_{t+k}, X_{t+k+1}, \ldots, X_{t+m}|X_0, X_1, \ldots, X_t] = \mathbb{P}[X_{t+k}, X_{t+k+1}, \ldots, X_{t+m}|X_t]\,, \quad \forall t \in \mathbb{N}_0\,,\ \forall k, m \in \mathbb{N}\,,\ m \geq k.$$

Answer:

Proof. We first establish the following claim: $\mathbb{P}[X_{t+k}|X_0, X_1, \ldots, X_t] = \mathbb{P}[X_{t+k}|X_t], k \in \mathbb{N}$.

The above statement can be proved by induction. For $k = 1$, the statement is just the definition of Markov chain (as stated in the problem), and thus holds true. Now we assume for $k = d$, we have the following

$$\mathbb{P}(X_{t+d}|X_0, \ldots, X_t) = \mathbb{P}(X_{t+d}|X_t).$$

We consider the case when $k = d + 1$. We have

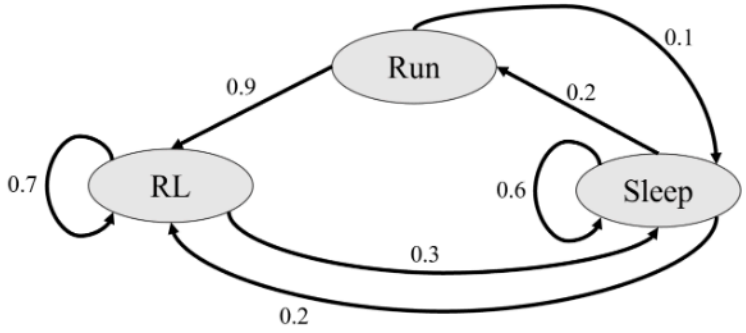
$$\begin{aligned}
\mathbb{P}(X_{t+d+1}|X_0, \ldots, X_t) &= \sum_j \mathbb{P}(X_{t+d+1}, X_{t+d} = j|X_0, \ldots, X_t) \\
&= \sum_j \mathbb{P}(X_{t+d+1}|X_{t+d} = j, X_0, \ldots, X_t)\mathbb{P}(X_{t+d} = j|X_0, \ldots, X_t) \\
&= \sum_j \mathbb{P}(X_{t+d+1}|X_{t+d} = j)\mathbb{P}(X_{t+d} = j|X_t) \quad \text{(Markov property \& ind'n hypothesis)} \\
&= \sum_j \mathbb{P}(X_{t+d+1}|X_{t+d} = j, X_t)\mathbb{P}(X_{t+d} = j|X_t) \\
&= \sum_j \mathbb{P}(X_{t+d+1}, X_{t+d} = j|X_t) \quad \text{(joint prob. in terms of conditional prob.)} \\
&= \mathbb{P}(X_{t+d+1}|X_t).
\end{aligned}$$

Therefore, by induction, we prove that the statement holds for all $t \in \mathbb{N}_0$ and for all $k \in \mathbb{N}$.
We will now use this to establish the proof as follows.

$$\begin{aligned}
\mathbb{P}(X_{t+k}, \ldots, X_{t+m}|X_0, \ldots, X_t) &= \mathbb{P}(X_{t+k}|X_0, X_1, \ldots, X_t) \cdot \mathbb{P}(X_{t+k+1}|X_0, X_1, \ldots, X_t, X_{t+k}) \ldots \\
&\quad \ldots \mathbb{P}(X_{t+m}|X_0, X_1, \ldots, X_t, X_{t+k}, X_{t+k+1}, \ldots, X_{t+m-1}) \\
&= \mathbb{P}(X_{t+k}|X_t) \cdot \mathbb{P}(X_{t+k+1}|X_{t+k}) \ldots \mathbb{P}(X_{t+m}|X_{t+m-1}) \\
&= \mathbb{P}(X_{t+k}|X_t) \cdot \mathbb{P}(X_{t+k+1}|X_{t+k}, X_t) \ldots \mathbb{P}(X_{t+m}|X_{t+m-1}, X_t) \\
&= \mathbb{P}(X_{t+k}, \ldots, X_{t+m}|X_t) \\
&\forall t \in \mathbb{N}_0 \quad : \forall k, m \in \mathbb{N} \quad m \geq k
\end{aligned}$$

Note that the arguments made above hold true since we have $m \geq k$. □

Problem 4: Consider the discrete-time, time-homogeneous Markov chain $(X_t)_{t=0}^{T}$ depicted in the figure below. It represents the activities that a student will do in every hour of the day and the transitions from one activity to another one. Assume that the initial state will be *sleep*.



1. What is the probability that the following trajectory (sequence of states) is realized after 4 transitions?

$$(sleep, run, RL, sleep, sleep)$$

Answer:

The probability of this path is given by: $1.0 \times 0.2 \times 0.9 \times 0.3 \times 0.6 = 0.0324$.

2. What is the probability that the following trajectory is realized if the process runs till infinity?

$$(sleep, run, RL, run, RL, run, RL, \ldots)$$

where $X_{2k-1} = run$ and $X_{2k} = RL$ for all $k \in \mathbb{N} = \{1, 2, 3, \ldots\}$.

Answer:

The probability of this path is given by: $1.0 \times 0.2 \times 0.9 \times 0 \times \ldots = 0$.

Note that the probability of state transition from RL to run is 0 (i.e., this transition is not possible), and hence the probability of this trajectory being realized is 0.

3. Define the state space and the transition function (matrix) of this Markov chain.

Answer:

The state space is defined as $\mathcal{S} := \{sleep, run, RL\}$.

The transition function (matrix) can be defined as follows: (state order $\rightarrow sleep, run, RL$)

$$P = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0 & 0.9 \\ 0.3 & 0 & 0.7 \end{bmatrix}$$

4. What will be the state distribution after two transitions, i.e., μ_2?

Answer:

We know that $\mu_0 = [1 \quad 0 \quad 0]$. We can obtain μ_2 as follows:

$$\mu_2 = \mu_0 P^2 = [0.44 \quad 0.12 \quad 0.44]$$

5. Compute the stationary (steady-state) distribution $\bar{\mu}$ of this Markov chain, by hand, calculator, or code; show the steps you use in the computation. Notice that since this is an ergodic Markov chain, it has a unique stationary distribution.

   Answer:

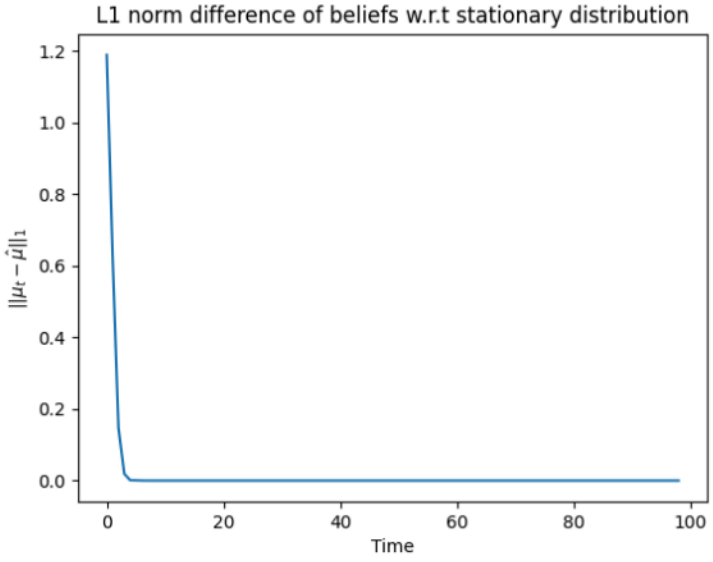
   The stationary distribution $\bar{\mu} = [\bar{\mu}_1 \quad \bar{\mu}_2 \quad \bar{\mu}_3]$ is the left eigenvector corresponding to the eigenvalue 1 of the transition matrix P. It can be computed analytically as follows:

$$[\bar{\mu}_1 \quad \bar{\mu}_2 \quad \bar{\mu}_3] \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.1 & 0 & 0.9 \\ 0.3 & 0 & 0.7 \end{bmatrix} = [\bar{\mu}_1 \quad \bar{\mu}_2 \quad \bar{\mu}_3];$$
$$\bar{\mu}_1 + \bar{\mu}_2 + \bar{\mu}_3 = 1.$$

   By solving the above system of equations, we obtain $\bar{\mu} = [0.4054 \quad 0.0810 \quad 0.5136]$.

6. Consider the consecutive update of the state distribution μ_t from $t = 0$ to $t = 100$. Plot the sequence of differences between the state distribution μ_t and the stationary distribution $\bar{\mu}$, measured using L1 norm, with respect to the time steps.

   Answer:



[**Note:** For this problem, if you have any code, please attach it to the end of your submission.]

Problem 5: Consider a movie recommendation system, interacting with a human user, that aims to suggest movies that the user would like to watch. In each step of the interaction, the human communicates its current desired movie genre to the system, which may be *Action*, *Horror*, or *Comedy*. Upon receiving this communication, the system selects a movie among *Movie A*, *Movie B*, *Movie C*, and *Movie D*. The genre of these movies is listed below, where 1 shows that the movie is categorized under that genre while 0 shows that it is not categorized under that genre.

	Action	Horror	Comedy
Movie A	1	0	1
Movie B	1	1	0
Movie C	0	1	1
Movie D	0	1	0

Table 1: Movie's genres

Upon receiving the recommendation, the user watches the movie and provides a *like* if the movie belongs to the desired genre and provides a *dislike* otherwise. If the movie belongs to the desired genre, in the next interaction, the user's desired movie genre will be selected uniformly at random from the other two genres. For instance, if the user wanted Horror in this interaction and Movie A, C, or D was suggested, the user will want Action or Comedy in the next interaction with equal probability. However, if the movie does not belong to the desired genre, in the next interaction, the user's desired movie genre will remain the same. At the beginning of the interaction, the user may select any of the genres uniformly at random.

1. Define an MDP that models this sequential decision-making scenario; in particular, specify all elements of the MDP. Consider an infinite-horizon setting with a discount factor of $\gamma = 0.98$. Let the reward value be bounded between $[-1, +1]$.

   Answer:

   We define every element of the MDP, the state space $\mathcal{S}$, the action space $\mathcal{A}$, the initial probability μ_0, the reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the transition probability $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$.

   - The state space $\mathcal{S}$ represents the three possible genres the user can ask for which we notate as $\mathcal{S} = \{s_A, s_H, s_C\}$ representing action, horror, and comedy in that order.
   - The action space $\mathcal{A}$ represents the four possible movies the recommendation agent can provide, which we notate as $\mathcal{A} = \{a_A, a_B, a_C, a_D\}$ representing movies A, B, C, and D in that order.
   - The initial state probability is uniform, hence $\mu_0(s) = \frac{1}{3}\ \forall\ s \in \mathcal{S}$.
   - The reward function $R(s, a) = 1$ for state and action pairs with 1 in the corresponding cell in table 1 and -1 otherwise. The agent gets a $+1$ reward when recommending a movie of the correct genre, and a -1 when recommending a movie of the wrong genre.
   - The transition probability

$$\mathcal{P}(s_{t+1} = s' | s_t = s, a_t = a) = \begin{cases} 1, & \text{if } s' = s \text{ and } a \text{ is not of the same genre as } s \\ 0, & \text{if } s' = s \text{ and } a \text{ is of the same genre as } s \\ 0, & \text{if } s' \neq s \text{ and } a \text{ is not of the same genre as } s \\ \frac{1}{2}, & \text{if } s' \neq s \text{ and } a \text{ is of the same genre as } s \end{cases}$$

2. Consider the following recommendation strategy by the system:

   - If the user asks for Action movies $\longrightarrow$ The system will recommend Movies A or B, each with probability of 0.4, or Movie D with probability of 0.2.
   - If the user asks for Horror movies $\longrightarrow$ The system will recommend any of the four movies with probability 0.25.
   - If the user asks for Comedy movies $\longrightarrow$ The system will recommend Movie B with probability of 0.1, or Movie C with probability of 0.9.

   Determine whether this policy is stationary or not. Also, determine whether this policy is deterministic or randomized (stochastic).

   Answer:

   The given policy is a stationary policy since it only depends on the current state. The agent will take the same decision whenever certain state is visited. However, a stationary policy may be deterministic or probabilistic (non-deterministic/stochastic/randomized), i.e., given a state, the agent may deterministically choose an action or sample an action from a probability distribution. In our case, the agent chooses actions (movies) probabilistically. Thus, the given policy is stationary and stochastic.

3. Define the Markov chain that is induced by the policy given in Part 2 over the MDP you introduced in Part 1; in particular, specify all elements of the Markov chain.

   Answer:

   The policy is defined as follows:

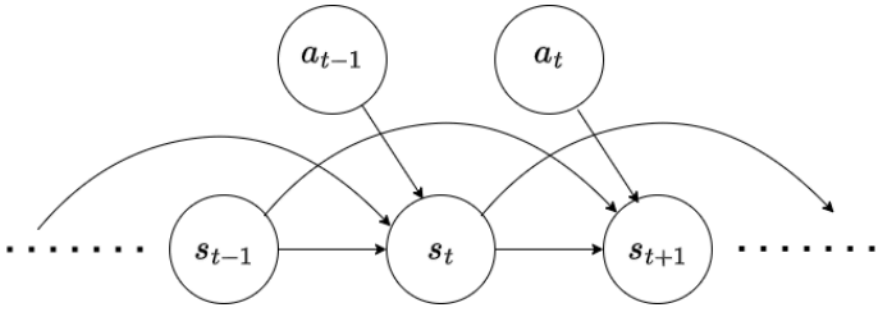
   - $\pi(a_A|s_A) = \pi(a_B|s_A) = 0.4, \quad \pi(a_C|s_A) = 0, \quad \pi(a_D|s_A) = 0.2$
   - $\pi(a_A|s_H) = \pi(a_B|s_H) = \pi(a_C|s_H) = \pi(a_D|s_H) = 0.25$
   - $\pi(a_A|s_C) = 0, \quad \pi(a_B|s_C) = 0.1, \quad \pi(a_C|s_C) = 0.9, \quad \pi(a_D|s_C) = 0$

   The state space $\mathcal{S}$ is the same as that of the MDP, i.e., $\mathcal{S} = \{s_A, s_H, s_C\}$. The matrix $P^\pi \in \mathbb{R}^{3\times3}$ is defined such that $P^\pi(s, s') = \sum_{a\in\mathcal{A}} \pi(a|s)\mathcal{P}(s'|s, a)$. The elements of the matrix are listed below.

   - $P^\pi(s_A, s_A) = (0.4 + 0.4) * 0 + 0.2 * 1 = 0.2$
   - $P^\pi(s_H, s_A) = (0.4 + 0.4) * 0.5 + 0.2 * 0 = 0.4$
   - $P^\pi(s_C, s_A) = (0.4 + 0.4) * 0.5 = 0.4$
   - $P^\pi(s_A, s_H) = 0.25 * 3 * 0.5 = 0.375$
   - $P^\pi(s_H, s_H) = 0.25$
   - $P^\pi(s_C, s_H) = 0.25 * 3 * 0.5 = 0.375$
   - $P^\pi(s_A, s_C) = 0.9 * 0.5 = 0.45$
   - $P^\pi(s_H, s_C) = 0.9 * 0.5 = 0.45$
   - $P^\pi(s_C, s_C) = 0.1 * 1 = 0.1$

4. Now, suppose the user's next desired movie genre was dependent not only on its last desired genre (and the recommended movie) but also the one before that. Show a graphical representation of the temporal evolution of the model in this case.

   Answer:



Problem 6: In the infinite-horizon discounted setting, we derived the equation

$$V^\pi = (I - \gamma P^\pi)^{-1} R^\pi$$

to perform policy evaluation for a deterministic, stationary policy π. Here, $V^\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the value function for all states $s \in \mathcal{S}$ represented as a column vector, $I \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is an identity matrix, $P^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is a probability transition matrix for the induced Markov chain under policy π, and $R^\pi \in \mathbb{R}^{|\mathcal{S}|}$ is the immediate reward for all states $s \in \mathcal{S}$ under policy π.

1. Prove that $I - \gamma P^\pi$ is an invertible matrix.

   Answer:

   Proof. Assume $(I - \gamma P^\pi)$ is singular, then there exists a non-zero vector $\mathbf{x}$ such that $(I - \gamma P^\pi)\mathbf{x} = 0$

   $$\begin{aligned} \mathbf{x} - \gamma P^\pi \mathbf{x} &= 0 \\ \gamma P^\pi \mathbf{x} &= \mathbf{x} \\ P^\pi \mathbf{x} &= \frac{1}{\gamma}\mathbf{x} \end{aligned}$$

   $\implies \frac{1}{\gamma}$ is an eigenvalue of P^π, but the largest eigenvalue of P^π is 1 and $\frac{1}{\gamma} > 1$ which is a contradiction.

   Therefore, $(I - \gamma P^\pi)$ is invertible. □

2. Derive a similar equation for policy evaluation for a stochastic, stationary policy. Specify clearly how each variable in the new equation can be determined.

   Answer:

   Starting from the bellman consistency equation. Let $\pi(a|s)$ be the probability of picking action a in state s according to policy π.

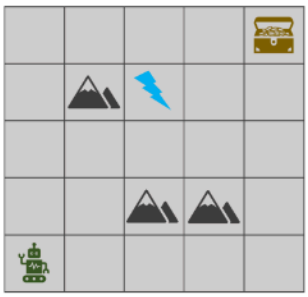
   $$\begin{aligned} V^\pi(s) &= \mathbb{E}_{a\sim\pi(.|s)}\left[R(s,a) + \gamma \mathbb{E}_{s'\sim P(.|s,a)}\left[V^\pi(s')\right]\right] \\ &= \sum_{a\in A} \pi(a|s) R(s,a) + \gamma \sum_{a\in A}\sum_{s'\in S} \pi(a|s) P(s'|s,a) V^\pi(s') \end{aligned}$$

   Let $R^\pi(s) = \sum_{a\in A} \pi(a|s) R(s,a)$ $\quad R^\pi$is an $|S| \times 1$ vector

   Let $P^\pi(s,s') = \sum_{a\in A} \pi(a|s) P(s'|s,a)$ $\quad P^\pi$is an $|S| \times |S|$ matrix

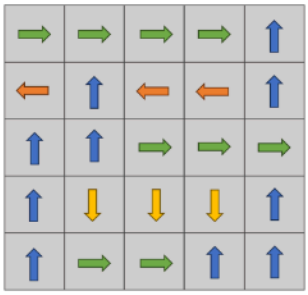
   $$\begin{aligned} \text{Then, } V^\pi &= R^\pi + \gamma P^\pi V^\pi \\ V^\pi &= (I - \gamma P^\pi)^{-1} R^\pi \end{aligned}$$

Problem 7: Consider an MDP over the grid-world illustrated below, where the goal is to take the agent to the treasure chest while avoiding the lightning.



- The state space contains the cells of the grid-world.
- The agent starts at the bottom left corner.
- The action space is $\{\text{up}, \text{down}, \text{left}, \text{right}\}$.
- The agent can only move to its adjacent cells, i.e., the cells that are above, below, to the left, or to the right of its current cell. If the agent is not at the boundary cells, each action will take the agent to the expected cell with probability 0.85 and to one of the remaining three cells, each with probability 0.05. If the agent is at the boundary, it will remain at its current cell with the sum of probabilities that would have taken it outside the grid-world. The cell with a mountain cannot be accessed, i.e., if adjacent to the mountain, the agent will remain at its current cell with the probability that would have taken it to the mountain. All actions in the cell with the lightning bolt and the one with the treasure chest will keep the agent in its current cell.
- The agent receives a reward of 0 in every cell for all actions except two cells. If at the cell with a lightning bolt, it will receive a reward of -1 for all actions, and if at the cell with the treasure chest, it will receive a reward of $+1$ for all actions.
- The discount factor is 0.95.

Evaluate the deterministic, stationary policy shown in the figure below, where each arrow represents the prescribed action in each cell.



1. Apply the analytical solution for policy evaluation. Report the value function at all states.

   Answer: The code for this problem is uploaded as a separate file.

The value function at each state (rounded to the 3rd decimal) is shown in the following matrix, with each value corresponding to the cell position:

$$\begin{pmatrix} 12.02 & 13.141 & 14.021 & 16.942 & 20. \\ 5.623 & 10.064 & -20. & -14.296 & 16.596 \\ 5.144 & 2.851 & 3.795 & 5.485 & 7.088 \\ 4.707 & 2.616 & 2.614 & 3.045 & 6.643 \\ 4.339 & 2.641 & 2.706 & 2.878 & 6.079 \end{pmatrix}$$

2. Implement and run the iterative solution for approximate policy evaluation with a zero initialization for the value function. Pick the number of iterations in a way to ensure 0.01 accuracy in the final computed value function, i.e., $\|V_T - V^\pi\|_\infty \leq 0.01$ without using the results of the previous part. Justify how you picked the number of iterations. Report the value function at all states.

   Answer: As seen in class, we can upper-bound the maximum value the value function can take using the bounded rewards and the discount factor.

$$\begin{aligned} \sup_{\pi,s} |V^\pi(s)| &= \sup_{\pi,s} \left| \mathbb{E}_{\substack{a_t \sim \pi(.|s_t) \\ s_{t+1} \sim P(.|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s \right] \right| \\ &\leq \sup_{\pi,s} \mathbb{E}_{\substack{a_t \sim \pi(.|s_t) \\ s_{t+1} \sim P(.|s_t,a_t)}} \left[\left| \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right| | s_0 = s \right] && \text{by Jensen's inequality} \\ &\leq \sup_{\pi,s} \mathbb{E}_{\substack{a_t \sim \pi(.|s_t) \\ s_{t+1} \sim P(.|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t \left| R(s_t, a_t) \right| | s_0 = s \right] && \text{by the triangle inequality} \\ &\leq \sup_{\pi,s} \mathbb{E}_{\substack{a_t \sim \pi(.|s_t) \\ s_{t+1} \sim P(.|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t \left| \max_{s,a} R(s,a) \right| | s_0 = s \right] \end{aligned}$$

   Everything in the expectation is now deterministic and does not depend on π or s

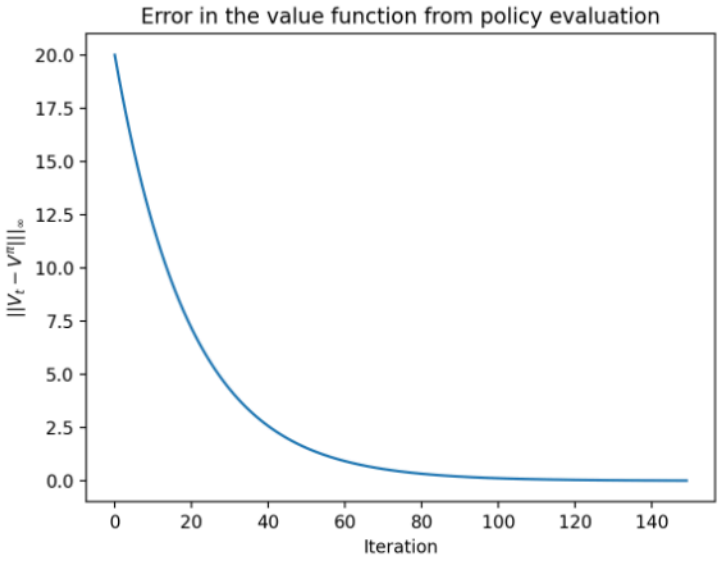
$$\begin{aligned} \sup_{\pi,s} |V^\pi(s)| &\leq \left| \max_{s,a} R(s,a) \right| \sum_{t=0}^{\infty} \gamma^t \\ &= 1 * \frac{1}{1-\gamma} = \frac{1}{1-0.95} = 20 \end{aligned}$$

   For an initial value V_0 of zeros, $||V_0 - V^\pi||_\infty = ||V^\pi||_\infty \leq \sup_{\pi,s} |V^\pi(s)| \leq 20$. Choose T that satisfies $T \geq \frac{\log\left(\frac{||V_0 - V^\pi||_\infty}{0.01}\right)}{\log\left(\frac{1}{\gamma}\right)}$ which is satisfied if $T \geq \frac{\log\left(\frac{20}{0.01}\right)}{\log\left(\frac{1}{\gamma}\right)} = 148.19$. Choose $T = 149$. The value function at each state (rounded to the 3rd decimal) is shown in the following matrix with each value corresponding to the cell position:

$$\begin{pmatrix} 12.013 & 13.133 & 14.013 & 16.933 & 19.99 \\ 5.615 & 10.057 & -19.99 & -14.289 & 16.587 \\ 5.136 & 2.844 & 3.789 & 5.478 & 7.08 \\ 4.7 & 2.608 & 2.607 & 3.037 & 6.635 \\ 4.331 & 2.634 & 2.699 & 2.87 & 6.071 \end{pmatrix}$$

3. Plot the sequence of errors in the value function from the approximate policy evaluation with respect to the iterations. In particular, plot $\|V_t - V^\pi\|_\infty$ against $t \in \mathbb{N}_0$, where V_t is the value function at iteration t and V^π is the value function computed from the analytical solution.

   Answer:



[**Note:** For this problem, attach all your code in a programming language of your choice to the end of your submission.]