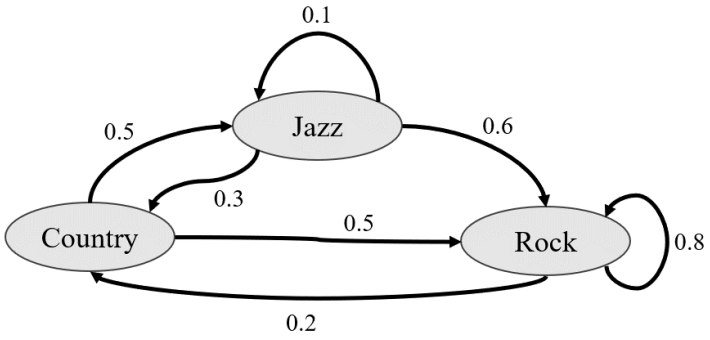
Homework Set 2

Problem 1: Consider a jukebox that plays songs from three genres of music: "Jazz," "Rock," and "Country." Once started, it plays a Jazz song and then, switches between the genres according to the time-homogeneous Markov chain $(X_t)_{t=0}^{\infty}$, depicted below.



1. Specify the state space $\mathcal{S}$, the initial distribution μ_0, and the transition function (matrix) P of this Markov chain.

2. What is the probability that the following sequence of genres is played once the jukebox starts?

   (a) (*Jazz*, *Country*, *Country*, *Rock*)

   (b) (*Jazz*, *Rock*, *Rock*, *Country*, *Jazz*)

   (c) (*Jazz*, *Rock*, *Country*, *Rock*, *Country*, ..., *Rock*, *Country*)
   where $X_0 = \textit{Jazz}$, $X_{2k-1} = \textit{Rock}$, and $X_{2k} = \textit{Country}$ for all $k \in \{1, 2, 3, \dots\ m\}$.

   (d) (*Jazz*, *Rock*, *Country*, *Rock*, *Country*, ...)
   where $X_0 = \textit{Jazz}$, $X_{2k-1} = \textit{Rock}$, and $X_{2k} = \textit{Country}$ for all $k \in \mathbb{N} = \{1, 2, 3, \dots\}$, i.e., as $m \to \infty$ in the previous part.

3. With what probabilities will the third song (at $t = 2$) belong to each of these three genres?

4. This Markov chain is ergodic and hence, has a unique stationary (steady-state) distribution $\bar{\mu}$. Compute $\bar{\mu}$ by hand, calculator, or code; show the steps you use in the computation.

5. 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.

6. If we start the jukebox and let it play for a very long time, which genre do you expect to be played most often? Explain your reasoning.

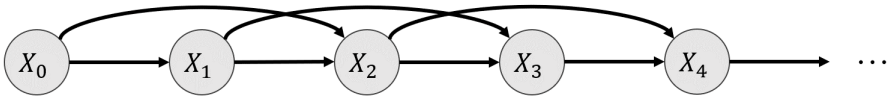
[**Note:** For this problem, if you have any code in a programming language of your choice, please attach it to the end of your submission.]

Problem 2: A Markov chain of order k defines a stochastic process $(X_t)_{t=0}^{\infty}$ such that

$$\mathbb{P}\left(X_t|X_0, X_1, X_2, \ldots, X_{t-1}\right) = \mathbb{P}\left(X_t|X_{t-k}, X_{t-k+1}, \ldots, X_{t-1}\right) \quad \text{for all } t \in \mathbb{N}_0, t \geq k,$$

i.e., the probability of moving to the next state depends on the past k states. Higher-order Markov chains are more expressive than simple Markov chains; however, they can be transformed into an equivalent simple Markov chain (Markov chain of order 1) at the expense of increasing the size of the state space.

1. Consider a Markov model of order 2 as shown in the graphical model below. Let the state



   space be $\{1, 2\}$ and the transition probabilities be

$$\mathbb{P}\left(X_t = 1|X_{t-2} = 1, X_{t-1} = 1\right) = 0.8\,, \quad \mathbb{P}\left(X_t = 1|X_{t-2} = 1, X_{t-1} = 2\right) = 0.1,$$
$$\mathbb{P}\left(X_t = 1|X_{t-2} = 2, X_{t-1} = 1\right) = 0.3\,, \quad \mathbb{P}\left(X_t = 1|X_{t-2} = 2, X_{t-1} = 2\right) = 0.7,$$

   if $t \geq 2$, and

$$\mathbb{P}\left(X_1 = 1|X_0 = 1\right) = 0.2\,, \quad \mathbb{P}\left(X_1 = 1|X_0 = 2\right) = 0.4,$$

   if $t = 1$. Let the initial state distribution be $\mathbb{P}\left(X_0 = 1\right) = 0.5$. Create an equivalent simple Markov chain over the larger state space $\{1, 2\}^2 = \{(1, 1), (1, 2), (2, 1), (2, 2)\}$. Write all transition probabilities between the states and the initial state distribution, or alternatively, show them on a graphical representation.

2. Based on the example in the previous part, how can we transform a Markov chain of order k into an equivalent simple Markov chain in general? Describe the process.

Problem 3: 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

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 B, 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 the state space, initial state distribution, action space, transition function, and reward function (let the reward value be bounded between $[-1, +1]$).

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).

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.

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 desired genre before that. Show a graphical representation of the temporal evolution of the model in this case.

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

Problem 4: Consider two MDPs $\mathcal{M}_1 = (\mathcal{S}, \mu_0, \mathcal{A}, P, R, \gamma)$ and $\mathcal{M}_2 = (\mathcal{S}, \mu_0, \mathcal{A}, P, R', \gamma)$ in the infinite-horizon discounted setting. $\mathcal{M}_2$ is identical to $\mathcal{M}_1$, except that its reward function R' is the result of a positive affine transformation over R, that is $R'(s,a) = \alpha R(s,a) + \beta$, where $\alpha > 0$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

1. For a stationary policy π, let $V_1^\pi(s)$ and $V_2^\pi(s)$ denote its value function over $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Assuming $V_1^\pi(s)$ is provided for all $s \in \mathcal{S}$, how can we write $V_2^\pi(s)$ based on $V_1^\pi(s)$ for all $s \in \mathcal{S}$? Show all steps and simplify your final answer.

2. Given the result in Part 1, argue about the relationship between the set of optimal stationary policies for $\mathcal{M}_1$ and the set of optimal stationary policies for $\mathcal{M}_2$. Explain your reasoning.

Problem 5: 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.
   [Hint: Recall that the transition matrix of a Markov chain is a row stochastic matrix whose maximum eigenvalue is 1*.]*

2. Derive a similar equation for policy evaluation for a stochastic, stationary policy. Specify clearly how each variable (P^π and R^π) in this equation can be determined.
   [Hint: Start with the Bellman consistency equation and follow similar steps as we did in class.]

3. Derive a similar equation for policy evaluation for a stochastic, stationary policy when the reward function is stochastic, i.e., $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$. Specify clearly how each variable (P^π and R^π) in this equation can be determined.
   [Hint: Start with the definition of the value function in this setting to arrive at a slightly modified Bellman consistency equation and then follow similar steps as we did in class.]