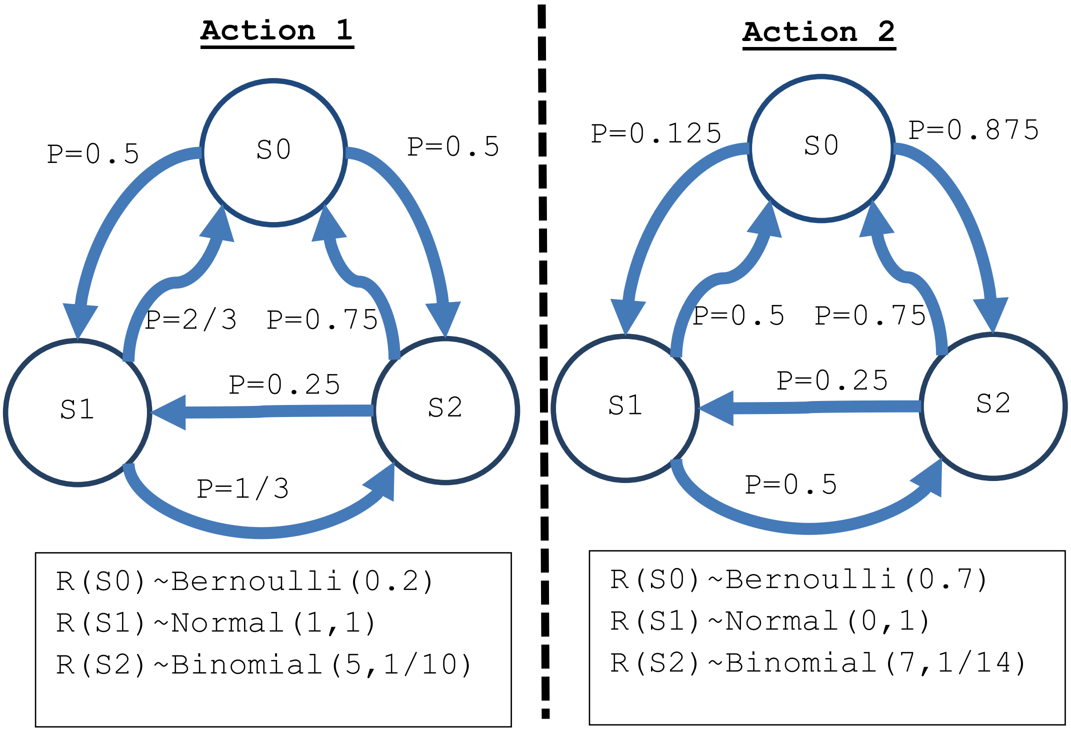
Planning and Learning in Dynamical Systems (046203)
Homework 2

Question 1 (Markov chains)

Let P be a transition matrix for some Markov chain. Show that:
1. All its rows are positive and sum to 1.
2. P has an eigenvalue of 1. What is the corresponding eigenvector?
3. All eigenvalues λ_i of P satisfy $|\lambda_i| \leq 1$.

Question 2

You are inside a shady casino with your not so bright friend Jack. You sit at the first table you see and the dealer offers you the following game: he presents you with a Markov Decision Process where you start at s_0 and can take one of two actions in each state. The transition and rewards for each action are given as follows:



R(S0)~Bernoulli(0.2)
R(S1)~Normal(1,1)
R(S2)~Binomial(5,1/10)

R(S0)~Bernoulli(0.7)
R(S1)~Normal(0,1)
R(S2)~Binomial(7,1/14)

a. You allow Jack to play a few rounds. Since 21 is his favorite number, Jack starts with the action 2, followed by the action 1 then again action 2 and so on. What is Jack's expected reward after 3 rounds (i.e., 3 actions)?

b. Jack changes his strategy and starts a new game (at s_0) choosing the action to be either 1 or 2 with equal probability. What will be Jack's expected reward after 3 rounds now? What is the induced stationary chain over the states?

c. Write and solve the optimal Bellman equation for 3 rounds. What is the optimal policy?

d. Assuming each round there is a β probability of getting thrown out of the casino, write down the infinite horizon cumulative reward. Conclude the connection between the discount factor and the death rate of a process.

e. Write the Bellman equations for the infinite horizon discounted case in this problem.

Question 3 (The Secretary Problem)

In this section you will model and solve the secretary problem by modeling it as an MDP. In the secretary problem, you are faced with N candidates to fill a secretarial position. In each round, a candidate is sampled uniformly from the poll of left candidates and being interviewed. Upon completion of an interview you decide whether to offer the job to the current candidate according to the score of the candidate. If you do not offer the job to the current candidate, the individual seeks employment elsewhere and is extracted out of the poll of candidates. Your goal is to hire the secretary most fit to the position, i.e., with the highest score (assume there is a single best candidate).

Define the state space as $S = \{0,1\}$ where $s = 1$ means that the current candidate has the highest score and $s = 0$ means it does not have the highest score. Define $g_t(s)$ as the probability that the current candidate has the highest score after observing $t - 1$ candidates (remember there are N candidates overall).

1. Show that (hint: remember the candidates are uniformly sampled)

$$g_t(s = 0) = 0,$$

$$g_t(s = 1) = P(Best\ object\ is\ in\ first\ t\ steps) = \frac{t}{N}.$$

2. Show that (hint: remember the candidates are uniformly sampled)

$$P_t(1|\,s) = \frac{1}{t+1},$$

$$P_t(0|\,s) = \frac{t}{t+1},$$

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

3. Let $V_t^*(s)$ denote the maximal probability of choosing the best candidate from state s at time t assuming no candidate had been chosen so far. At time step t we are left with two options: either we pick the current candidate, or we discard the current one and continue to the next candidate. Using this observation and following similar reasoning as in Backward-Induction show that:
$$V_t^*(1) = \max\{g_t(1), P_t(1|\,1)V_{t+1}^*(1) + P_t(0|\,1)V_{t+1}^*(0)\}$$
$$V_t^*(0) = \max\{g_t(0), P_t(1|\,0)V_{t+1}^*(1) + P_t(0|\,0)V_{t+1}^*(0)\}\,.$$
What are $V_{t=N}^*(1)$ and $V_{t=N}^*(0)$?
4. Show that these equations can be written as
$$V_t^*(1) = \max\{\frac{t}{N}, V_t^*(0)\}$$
$$V_t^*(0) = \frac{1}{t+1}V_{t+1}^*(1) + \frac{t}{t+1}V_{t+1}^*(0)\,.$$
Solve the induction numerically for $N = 10$ and plot $V_t^*(1), V_t^*(0)$ versus t.
5. Explain why section (4) implies that the optimal strategy is observing τ candidates and then select the first candidate who is better than all the previous ones.

<u>Question 4 (From the lecture)</u>

Prove the following equality

$$V^{\pi}(s) \,@ E^{\pi}\Big(\sum_{t=0}^{\infty}\gamma^{t} r(s_t, a_t) \,|\, s_0 = s\Big)$$
$$= E^{\pi}\Big(\sum_{t=1}^{\infty}\gamma^{t-1} r(s_t, a_t) \,|\, s_1 = s\Big)$$

Question 5 (MDPs with Non-Linear Objectives)

Consider an MDP $\mathcal{M}$ with finite state space $\mathcal{S}$ and finite actions space $\mathcal{A}$, transitions $P(s'|s,a)$, discount factor $\gamma \in (0,1)$, a **fixed** initial state s_{init}, and **two** reward functions: $r_1(s)$ and $r_2(s)$. In this question we will consider objectives that are a function of both r_1 and r_2.

For a Markov policy π, we denote the discounted returns J_1^π and J_2^π as:

$$J_1^\pi = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_1(s_t) \middle| s_0 = s_{init} \right],$$
$$J_2^\pi = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_2(s_t) \middle| s_0 = s_{init} \right].$$

Note that the initial state is fixed, and that J_1^π, J_2^π denote scalar returns and not value functions.

Let $f(x,y)$ be some function of two variables. For some policy π we denote $J^\pi = f(J_1^\pi, J_2^\pi)$. We wish to find a policy π^* that maximizes J^π:

$$J^* = \max_\pi J^\pi, \quad \pi^* \in \operatorname*{argmax}_\pi J^\pi.$$

a. For $f(x,y) = \alpha x + \beta y$, propose a standard MDP with a single reward $\hat{r}$ such that it's optimal policy is π^*. Explain.

For the rest of this question, we consider the function $f(x,y) = \frac{x}{y}$. Furthermore, we assume the following bounds on the rewards: $0 < r_{min} \le r_1(s) < r_2(s) \le r_{max}$, for all $s \in \mathcal{S}$.

b. Can the standard MDP solution approaches (value iteration, policy iteration) be used to find π^* in this case? Explain (no need to prove formally).

For some $\rho \in [0,1]$, consider a standard MDP $\mathcal{M}_\rho$ with the same $\mathcal{S}, \mathcal{A}, P, \gamma$ as $\mathcal{M}$ and reward $\hat{r}(s) = r_1(s) - \rho r_2(s)$. Denote the discounted reward for a policy π in $\mathcal{M}_\rho$ as $J_\rho^\pi = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t \hat{r}(s_t) | s_0 = s_{init} \right]$.

c. Assume that for some policy π, we have that $J_\rho^\pi = 0$. What is J^π?

d. Let π_ρ^* be an optimal policy in $\mathcal{M}_\rho$, that also satisfies $J_\rho^{\pi_\rho^*} = 0$. Show that π_ρ^* is optimal also in $\mathcal{M}$.

Hint: assume that for some π', $J^{\pi'} > \rho$, and show a contradiction.

e. Show that for $\rho = 0$, $J_\rho^\pi > 0$ for any π.

f. Show that for $\rho = 1$, $J_\rho^\pi < 0$ for any π.

g. Let J_ρ^* denote the optimal value in $\mathcal{M}_\rho$. Show that J_ρ^* is monotonically decreasing in ρ.

Hint: start by showing monotonicity for a fixed policy.

h. Based on (d-g), propose an approach for finding the optimal policy π^*. Technically, you can assume that J_ρ^* is continuous in ρ, and you can invoke any standard MDP solver in your solution.