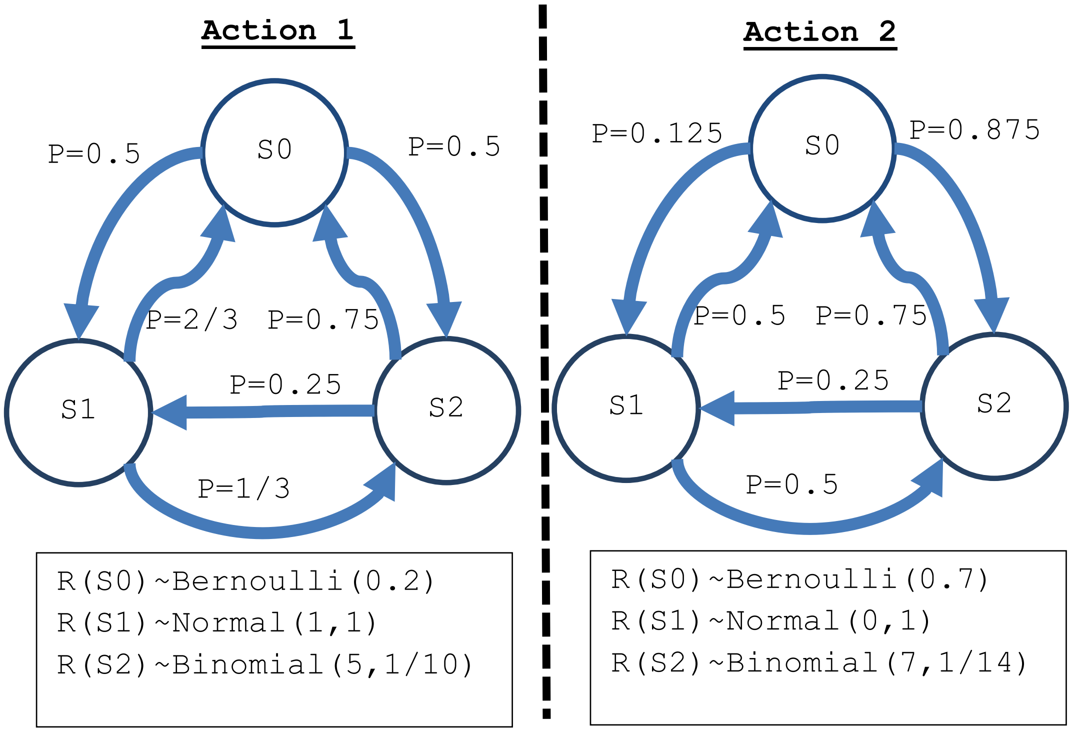
Planning and Learning in Dynamical Systems (046203)
Homework 2, due 20/5/2020

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:



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.

<u>Question 3 (The Secretary Problem)</u>

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.

Question 4 (From the lecture)

Prove the following equality

$$V^{\pi}(s) \triangleq E^{\pi}\Big(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\Big)$$

$$= E^{\pi}\Big(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \mid s_1 = s\Big)$$

Question 5 - Second moment and variance of return

In the lectures we have defined the value function $V^{\pi}(s)$ as the *expected* discounted return when starting from state s and following policy π,

$$V^{\pi}(s) = E^{\pi,s}\Big(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big) .$$

We have seen that $V^{\pi}(s)$ satisfies a set of $|S|$ linear equations (Bellman equation)

$$V^{\pi}(s) = r(s,\pi(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi(s)) V^{\pi}(s'), \quad s \in S .$$

We now define $M^{\pi}(s)$ as the *second moment* of the discounted return when starting from state s and following policy π,

$$M^{\pi}(s) = E^{\pi,s}\left(\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right)^2\right) .$$

a. We will show that $M^{\pi}(s)$ satisfies a 'Bellman like' set of equations. Write an expression for $M^{\pi}(s)$ that has a linear dependence on M^{π} and V^{π}.

Hint: start by following the derivation of the Bellman equation for V^{π}.

b. How many equations are needed to solve in order to calculate $M^{\pi}(s)$ for all $s \in S$?

c. We now define $W^{\pi}(s)$ as the *variance* of the discounted return when starting from state s and following policy π,

$$W^{\pi}(s) = \mathrm{Var}^{\pi,s}\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right) .$$

Explain how $W^{\pi}(s)$ may be calculated.