Question 4:

we want to show that

$$V^{\pi}(s) \triangleq E^{\pi}\left(\sum_{t=0}^{\infty} \gamma^{t} r(s_t, a_t) \middle| s_0 = s\right)$$

$$= E^{\pi}\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \middle| s_1 = s\right)$$

we can prove this lemma by assumption that π is stationary and changing the sum index

$$E^{\pi}\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \middle| s_1 = s\right) = E^{\pi}\left(\sum_{t=0}^{\infty} \gamma^{t} r(s_{t+1}, a_{t+1}) \middle| s_1 = s\right)$$

$$= E^{\pi}\left(\sum_{t=0}^{\infty} \gamma^{t} \left(r(s_{t+1}, \pi(s_{t+1})\right) \middle| s_1 = s\right) = E^{\pi}\left(\sum_{t=0}^{\infty} \gamma^{t} r(s_t, \pi(s_t)) \middle| s_0 = s\right)$$

$$= V^{\pi}(s)$$

Question 2

1. we can see that

$$r(s_0, a_1) = E[\text{bernoulli}(0.2)] = 0.2$$

$$r(s_1, a_1) = E[\text{Normal}(1,1)] = 1$$

$$r(s_2, a_1) = 0.5$$

$$r(s_0, a_2) = 0.1 \qquad r(s_1, a_2) = 0$$

$$r(s_2, a_2) = 0.5$$

Jacks expected reward After 3 rounds is:

$$E^{\pi_{2,1}}\left|\sum_{t=0}^{3} R_t\right| = r(s_0, a_2) + \left(0.125\, r(s_1, a_1) + 0.875\, r(s_2, a_1)\right)$$

$$+ \Big(\left(0.125 \cdot \tfrac{2}{3} + 0.875 \cdot 0.75\right) r(s_0, a_2) + 0.875 \cdot 0.25\, r(s_1, a_2)$$

$$+ 0.125 \cdot \tfrac{1}{3} r(s_2, a_2)\Big) = 1.801$$

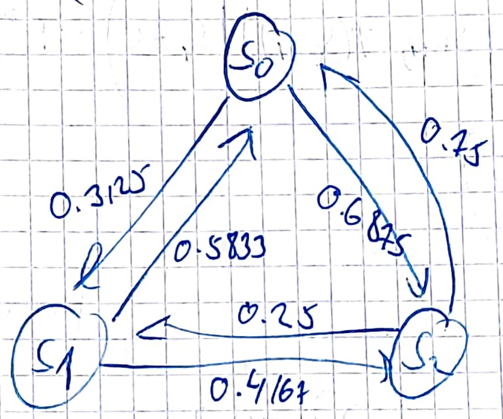
b. now we get

$$P_{i,j} = P(S_{t+1} = S_j \mid S_t = S_i) = P(S_j \mid S_i, a_1)\,P(a_1) + P(S_j \mid S_i, a_2)\,P(a_2)$$

$$= 0.5\left(P(S_j \mid S_i, a_1) + P(S_j \mid S_i, a_2)\right)$$

and with that we get

$$P = \begin{pmatrix} 0 & 0.3125 & 0.6875 \\ 0.5833 & 0 & 0.4167 \\ 0.75 & 0.25 & 0 \end{pmatrix}$$



now the expected reward for each state is

$$r(S_i) = r(S_i, a_1)\,P(a_1) + P(a_2)\,r(S_i, a_2) = 0.5\left(r(S_i, a_1) + r(S_i, a_2)\right)$$

$$r = [0.45, 0.5, 0.5]$$

and the total expected reward is

$$E^{\pi}\left[\sum_{t=0}^{3} R_t\right] = r(S_0) + \left(0.3125\, r(S_1) + 0.6875\, r(S_2)\right)$$

$$+ \left(\left(0.3125 \cdot 0.5833 + 0.6875 \cdot 0.75\right) r(S_0) + 0.6875 \cdot 0.25\, r(S_1) + 0.3125 \cdot 0.4167\, r(S_2)\right) = 1.415$$

c. lets write the bellman equation for 3 rounds:

$$V_3(s) = 0$$

$$V_k(s) = \max_{a \in \{a_1, a_2\}} \left\{ r(s,a) + \sum_{s' \in \{s_0, s_1, s_2\}} P(s'|s,a) \cdot V_{k+1}(s') \right\}$$

we get:

$$V_2(s_0) = \max\{0.2, 0.7\} = 0.7 \ \{a_2\}$$

$$V_2(s_1) = \max\{1, 0\} = 1 \ \{a_1\}$$

$$V_2(s_2) = \max\{0.5, 0.5\} = 0.5 \ \{a_1/a_2\}$$

$$V_1(s_0) = \max\{0.2 + 0.5 \cdot 1 + 0.5 \cdot 0.5,\ 0.7 + 0.125 \cdot 1 + 0.875 \cdot 0.5\}$$

$$= 1.2625 \ \{a_2\}$$

$$V_1(s_1) = \max\left\{1 + \frac{2}{3} \cdot 0.7 + \frac{1}{3} \cdot 0.5,\ 0 + 0.5 \cdot 0.7 + 0.5 \cdot 0.5\right\}$$

$$= 1.633 \ \{a_1\}$$

$$V_1(s_2) = \max\{0.5 + 0.75 \cdot 0.7 + 0.25 \cdot 1,\ 0.5 + 0.75 \cdot 0.7 + 0.25 \cdot 1\}$$

$$= 2.2 \ \{a_1/a_2\}$$

$$V_0(s_0) = \max\{0.2 + 0.5 \cdot 1.633 + 0.5 \cdot 2.2,\ 0.7 + 0.125 \cdot 1.633 + 0.875 \cdot 2.2\}$$

$$= 2.83 \ \{a_2\}$$

and the optimal policy

$$\pi_0^*(s_0) = a_2 \quad \pi_1^*(s) = \pi_2^*(s) = \begin{cases} a_2 & s = s_0 \\ a_1 & s = s_1 \\ a_1 \text{ or } a_2 & s = s_2 \end{cases}$$

d. The probability to stay in the casino After t rounds is $(1-\beta)^t$. hence the infinite horizon cumulative reward:

$$J_\beta^\pi(s) = E^{\pi, s_0}\left(\sum_{t=0}^{\infty} P(\text{Stay After } t \text{ rounds}) \cdot R_t\right)$$

$$= E^{\pi, s_0}\left(\sum_{t=0}^{\infty} (1-\beta)^t R_t\right) = E^{\pi, s_0}\left(\sum_{t=0}^{\infty} (1-\beta)^t r(s_t, a_t)\right)$$

$\gamma \triangleq 1-\beta$. defines the connection between the discount factor and the death rate

e. for the infinite horizon case we have the following bellman equations:

$$V(s) = \max_{a \in \{a_1, a_2\}} \left\{ r(s,a) + (1-\beta) \sum_{s' \in \{s_0, s_1, s_2\}} P(s'|s,a) \cdot V(s') \right\}$$

$$\pi^*(s) = \arg\max_{a \in \{a_1, a_2\}} \left\{ \cdots\cdots \right\}$$