HW2 in 046203 Planning and Reinforcement Learning

Submitter 1: David Valensi 342439643
Submitter 2: Yaniv Galron 206765646

Question 1 (Markov Chain):

1. P is the transition matrix for some Markov Chain (MC), so $p_{i,j}$ is the transition probability from state I to j. Thus, $p_{i,j} \geq 0$ for all i,j.
   Each row in P sums up to 1 (i.e. P is row stochastic):
   The first equality holds since we consider a time-homogeneous MC as defined in class:

$$\forall i: \sum_{j=1}^{n} P_{i,j} = \sum_{j=1}^{n} P(X_1 = j|X_0 = i) = \frac{1}{P(X_0 = i)} \sum_{j=1}^{n} P(X_1 = j, X_0 = i) = \frac{P(X_0 = i)}{P(X_0 = i)} = 1$$

2. We show that $P\vec{1} = \vec{1}$.
   P is row stochastic, then for each row $i, \sum_{j=1}^{n} P_{i,j} = 1 \Longrightarrow P\vec{1} = \lambda\vec{1}, with\ \lambda = 1$. It means that $\lambda = 1$ is one of the eigenvalues of P and the corresponding eigenvector is $\vec{1} = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$. Since the right eigenvalues and left eigenvalues are the same for square matrices, 1 is also a left eigenvalue of P.
   $\Longrightarrow \exists x : x^T P = x^T$.
3. Let λ be a P eigenvalue with x the corresponding eigenvector. We show that $|\lambda| \leq 1$.
   Suppose by contradiction that $|\lambda| > 1$. Let's denote $x_i the\ largest\ element\ in\ x$.
   Since any αx holds the equation, then we assume $x_i > 0$.

$$We\ have\ Px = \lambda x \Rightarrow (Px)_i = \sum_{j=1}^{n} P_{i,j}\, x_j = \lambda x_i > x_i\ (|\lambda| > 1)$$

   But in the second hand, P rows sum to 1 and each element in λx is a convex combination of x. Thus, no entry in λx can be larger than x_i. Contradiction.
   $\Longrightarrow |\lambda| \leq 1$.

Question 3 (The Secretary Problem)

1. $g_t(s=0)$ is the probability that the t^th candidate has the highest score while $s=0$, meaning that the current candidate is not the best. So obviously, the t^th candidate has not the highest score: $g_t(s=0)=0$.
   For s=1, we have seen so far t-1 candidates, and we are interested in the probability that the t^th candidate has highest score. Of course if the t^th candidate has highest score, it has particularly the highest among first t candidates and that's the information we are given (as interviewers)

$$g_t(s=1) = P(t^{th} candidate\ has\ highest\ score \mid t^{th} candidate\ is\ the\ best\ among\ first\ t)$$

   From uniform sampling we have:

$$g_t(s=1) = \frac{P(t^{th} candidate\ has\ highest\ score\ )}{P(\ t^{th} candidate\ is\ the\ best\ among\ first\ t)} = \frac{\frac{1}{N}}{\frac{1}{t}} = \frac{t}{N}$$

2. Now we are interested in $P_t(1|s)$. It is the probability that the t+1^th candidate is the best one given that we already interviewed t candidates. Each candidate is uniformly sampled, thus $P_t(1|s) = \frac{1}{t+1}$.
   And $P_t(0|s) = 1 - P_t(1|s) = \frac{t}{t+1}$.
3. To compute $V_t^*(s)$, we need to consider two options at time t and state s: hire t^th candidate or continue interviewing.
   The first option to pick the t^th candidate after interviewing t candidates and we are in state s, that it $g_t(s)$.
   The second option is to continue searching and act according to $V_{t+1}^*(s_{next})$**.** In the next step, we may be in two different states. So, if we continue searching V holds the following.

$$V_t^*(1) = \ P_t(1|1)V_{t+1}^*(1) + P_t(0|1)V_{t+1}^*(0)$$
$$V_t^*(0) = \ P_t(1|0)V_{t+1}^*(1) + P_t(0|0)V_{t+1}^*(0)$$

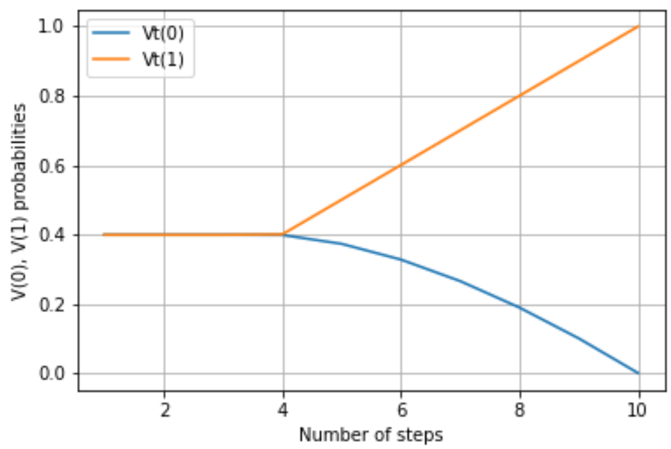
   Of course, we'll act greedily at each time step and take the maximum between the two options:

$$V_t^*(1) = \max\{g_t(1), P_t(1|1)V_{t+1}^*(1) + P_t(0|1)V_{t+1}^*(0)\}$$
$$V_t^*(0) = \max\{g_t(0), P_t(1|0)V_{t+1}^*(1) + P_t(0|0)V_{t+1}^*(0)\}$$

   $V_{t=N}^*(1) = 1$ since at time N if we have not chosen any candidate and the last one is the best, then the probability to choose the best is 1.
   $V_{t=N}^*(0) = 0$ from similar considerations, at last time step, if the last candidate is not the best, then the probability to choose the best is 0.
4. Since $g_t(s=0)=0$:
   $V_t^*(0) = \max\{g_t(0), P_t(1|0)V_{t+1}^*(1) + P_t(0|0)V_{t+1}^*(0)\}$
   $V_t^*(0) = P_t(1|0)V_{t+1}^*(1) + P_t(0|0)V_{t+1}^*(0)$
   $V_t^*(0) = \dfrac{1}{t+1}V_{t+1}^*(1) + \dfrac{t}{t+1}V_{t+1}^*(0)$

$$V_t^*(1) = \max\{g_t(1), P_t(1|1)V_{t+1}^*(1) + P_t(0|1)V_{t+1}^*(0)\} = \max\left\{\frac{t}{N}, V_t^*(0)\right\}$$

The code of the plot appears at the end. The plot if V values for N=10 is:



5. We observe that the values of Vt(1) and Vt(0) are the same until $\tau = t = 4$. It means that there is no value to choose a candidate before interviewing at least 4 candidates. Then $V_t^*(1) = \frac{t}{N}$ meaning that the best option was to choose the t^th candidate if he's better than the previous ones.

```
import numpy as np
import matplotlib.pyplot as plt

v1 = [1]
v0 = [0]

N=10

for t in range(N-1, 0, -1):
    v0.append(1/(t+1)* v1[-1] + t/(t+1)* v0[-1])
    v1.append(max(t/N, v0[-1]))

v0.reverse()
v1.reverse()
x_axis = range(1, N+1)
fig, ax = plt.subplots()
ax.plot(x_axis, v0, label='Vt(0)')
ax.plot(x_axis, v1, label='Vt(1)')
ax.set(xlabel='Number of steps', ylabel='V(0), V(1) probabilities')
ax.grid()
ax.legend()
fig.savefig("v_over_time.png")

plt.show()
```

Question 5 (MDP with Non-Linear Objectives)

a) Let's suggest the following reward function. The rest of the MDP parameters are the same as the original MDP.

$$\hat{r}(s) = \alpha r_1(s) + \beta r_2(s)$$

From Expectation linearity:

$$J^\pi = E^\pi\left[\sum_{t=0}^{\infty} \gamma^t\big(\alpha r_1(s_t) + \beta r_2(s_t)\big) |\, s_0 = s_{init}\right] = \alpha J_1^\pi + \beta J_2^\pi$$

$$J^* = \max f(J_1^\pi, J_2^\pi) = \max \alpha J_1^\pi + \beta J_2^\pi$$

We got the same discounted reward $\alpha J_1^\pi + \beta J_2^\pi$ with the suggested reward function $\hat{r}(s)$.

Thus, the optimal policy for the suggested MDP is as expected:

$$\pi^* \in argmax\, J^\pi$$

b) With $J^\pi = f(J_1^\pi, J_2^\pi) = \frac{J_1^\pi}{J_2^\pi}$, the standard approaches like VI or PI cannot be applied.

We know $r_1 < r_2$, then $J_1^\pi \leq J_2^\pi$. Thus, we also have that $\frac{J_1^\pi}{J_2^\pi} \leq 1$.

In fact, maximizing $\frac{J_1^\pi}{J_2^\pi}$, means that we want each of the discounted rewards J_1^π ,J_2^π to be as close as possible to the other. The approaches maximize the discounted reward with respect to a specific reward function and not with respect to the whole discounted reward. When we optimize the factor $\frac{J_1^\pi}{J_2^\pi}$, the game is different: if we increase J_1^π, we will also increase J_2^π according to the observation above. But the greater J_2^π , the lower $\frac{J_1^\pi}{J_2^\pi}$.

c) $J_\rho^\pi = J_1^\pi - \rho J_2^\pi = 0 \rightarrow J_1^\pi = \rho J_2^\pi \rightarrow J^\pi = \frac{J_1^\pi}{J_2^\pi} = \rho$

d) We want to prove $\pi_\rho^* \in \underset{\pi}{argmax} \frac{J_1^\pi}{J_2^\pi}$

Let's assume in contrary that $\exists \pi' s.t. J^{\pi'} = \frac{J_1^{\pi'}}{J_2^{\pi'}} > \rho$:

$\frac{J_1^{\pi'}}{J_2^{\pi'}} > \rho \rightarrow J_1^{\pi'} > \rho J_2^{\pi'} \rightarrow J_\rho^{\pi'} = J_1^{\pi'} - \rho J_2^{\pi'} > 0$. In contradiction with optimality of π_ρ^*.

Thus, $.J^{\pi'} = \frac{J_1^{\pi'}}{J_2^{\pi'}} \leq \rho$. Under the same policy π_ρ^*, $J^{\pi_\rho^*} = \rho$ as we proved.

Thus $\pi_\rho^* \in \underset{\pi}{argmax} \frac{J_1^\pi}{J_2^\pi}$.

e) We are given that $0 < r_{min} < r_1(s)$

$$\Rightarrow \forall \pi, J_{\rho=0}^\pi = E^\pi\left[\sum_{t=0}^{\infty} \gamma^t\big(r_1(s_t)\big) |\, s_0 = s_{init}\right] \geq E^\pi\left[\sum_{t=0}^{\infty} \gamma^t r_{min} |\, s_0 = s_{init}\right] > 0$$

f) We are given that $\forall s,\ 0 < r_{min} < r_1(s) < r_2(s) \Rightarrow r_1(s) - r_2(s) < 0$.

Let's denote $\epsilon < 0\ s.t. \forall s, r_1(s) - r_2(s) < \epsilon$

$$\Rightarrow \forall\pi, J^{\pi}_{\rho=1} = E^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\big(r_1(s_t) - r_2(s_t)\big) |\, s_0 = s_{init}\right] < E^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t(\epsilon) |\, s_0 = s_{init}\right] < 0$$

g) Let's start for a fixed policy: $\forall\rho' > \rho,\ J^{\pi}_{\rho} > J^{\pi}_{\rho'>\rho}, i.e. J^{\pi}_{\rho} - J^{\pi}_{\rho'>\rho} > 0$

$$J^{\pi}_{\rho} - J^{\pi}_{\rho'>\rho} = E^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\big(r_1(s_t) - \rho r_2(s_t)\big) |\, s_0 = s_{init}\right] - E^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\big(r_1(s_t) - \rho' r_2(s_t)\big) |\, s_0 = s_{init}\right]$$

From expectation linearity:

$$= E^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\big(r_1(s_t) - \rho r_2(s_t) - r_1(s_t) + \rho' r_2(s_t)\big) |\, s_0 = s_{init}\right] =$$

$$E^{\pi}\left[\sum_{t=0}^{\infty}\gamma^t\big((\rho' - \rho) r_2(s_t)\big) |\, s_0 = s_{init}\right] > 0$$

$Since\ \rho' > \rho$

Now we need to prove that for $\forall\rho > \rho',\ J^{*}_{\rho'} > J^{*}_{\rho>\rho'}$

From the fixed policy explanation from above: $J^{*}_{\rho'} = J^{\pi^*}_{\rho'} > J^{\pi^*}_{\rho>\rho'}$

It is obvious that the optimal policy $\pi^{*\prime}\ for\ \rho'\ in\ M_{\rho'}$ gives at least the same reward as the optimal policy $\pi^*\ for\ \rho\ in\ M_{\rho'} : J^{\pi^*}_{\rho'} \le J^{\pi^{*\prime}}_{\rho'}$

Unifying both sides give:

$$J^{\pi^*}_{\rho>\rho'} < J^{\pi^*}_{\rho'} \le J^{\pi^{*\prime}}_{\rho'}$$

As required.

h) We're given that J^{π}_{ρ} is continuous in ρ.

And we showed that for $\rho = 0, J^{\pi}_{\rho} > 0, and\ for\ \rho = 1, J^{\pi}_{\rho} < 0$

From the intermediate value theorem: $\exists\ 0 < \rho < 1 :\ J^{\pi}_{\rho} = 0$

Let's define $0 < \rho < 1$, and let's use one of the standard approaches to solve the M_ρ MDP (Policy Iteration, Value Iteration). Let's denote J^*_ρ the optimal value in M_ρ and π^*_ρ the optimal policy.

If $J^*_\rho = 0$, return π^*_ρ.

Else, let's find $0 < \rho' < 1$ s.t. $J^*_{\rho'} = 0$ using binary search and executing PI or VI algorithms on $M_{\rho'}$. If $J^*_{\rho'} = 0$ return $\pi^*_{\rho'}$.

The binary search will surely find some ρ' s.t. $J^*_{\rho'} = 0$ from the monotonicity and from the intermediate value theorem above.

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:

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

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

$$\left. + 0.125 \cdot \frac{1}{3} r(s_2, a_2)\right) = 1.801$$

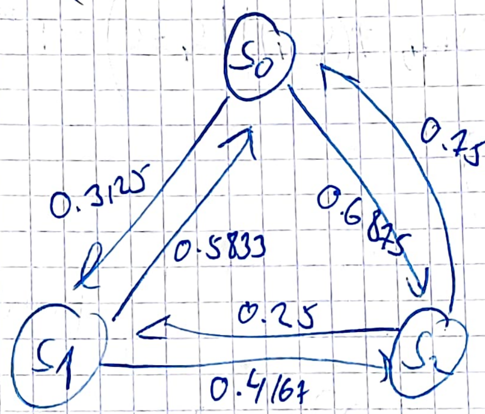
b. now we get

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

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

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

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

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

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:

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

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

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

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

$= 1.2625 \ \{a_2\}$

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

$= 1.633 \ \{a_1\}$

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

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

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

$= 2.83 \ \{a_2\}$

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:

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

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