RL-hw2

Omer Cohen 308428127
Dvir Marsh 313592685

May 2022

1 Question 1

1.1

By definition: $p_{i,j} = \mathbb{P}(X_t = j | X_{t-1} = i)$

$$\Rightarrow \sum_j p_{i,j} = \sum_j \mathbb{P}(X_t = j | X_{t-1} = i) = 1$$

1.2

Lets look at $u = \begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$

Then: $v = P \cdot u = \begin{pmatrix} \sum\limits_j p_{1,j} \\ \sum\limits_j p_{2,j} \\ \cdot \\ \cdot \\ \cdot \\ \sum\limits_j p_{n,j} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix} = 1 \cdot u$

Hence, u is an eigenvector with eigenvalue $\lambda_u = 1$.

1.3

Lets suppose in the negative that we have an eigenvector of P w with eigenvalue $|\lambda_w| > 1$, let w be normalized so that $\max\limits_i |w_i| = 1$, let u be the vector from last question, so the vector $w + u$ has only positive or zero values, and so the vector $\frac{u+w}{\alpha}$ such that $\alpha = \sum\limits_i (u + w)_i$ is a valid probability vector.

Because $|\lambda_w| > 1$, there exist an even m where $|\lambda_w|^m = \lambda_w^m > \alpha$. Lets look at: $v^m = P^m \cdot \frac{u+w}{\alpha} = \frac{1}{\alpha}(u + \lambda_w^m w)$.
Now, we will look at the index i where $|w_i| = 1$ and assume $w_i = 1$: $v_i^m = \frac{1}{\alpha} + \frac{\lambda_w^m}{\alpha} > \frac{1}{\alpha} + 1 > 1$, so we got that we have state with probability grater then 1 and it is not feasible.
For $w_i = -1$ we could have found m such $v_i^m < 0$ which is also not feasible.

2 Question 2

2.1

Lets define the transition matrix given action $i \in \{1, 2\}$ as P_i. Lets define p_t as the probability distribution over states at time t. Last of all, lets define $R_i, i \in \{1, 2\}$ as expected reward for each state given action i.
$P_2 = \begin{pmatrix} 0 & 0.125 & 0.875 \\ 0.5 & 0 & 0.5 \\ 0.75 & 0.25 & 0 \end{pmatrix}$, $P_1 = \begin{pmatrix} 0 & 0.5 & 0.5 \\ \frac{2}{3} & 0 & \frac{1}{3} \\ 0.75 & 0.25 & 0 \end{pmatrix}$.
$p_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$.
$R_2 = \begin{pmatrix} 0.7 \\ 0 \\ 0.5 \end{pmatrix}$, $R_1 = \begin{pmatrix} 0.2 \\ 1 \\ 0.5 \end{pmatrix}$.
So, the total reward would be: $r = R_2^T \cdot p_0 + R_1^T \cdot P_2 \cdot p_0 + R_2^T \cdot P_1 \cdot P_2 \cdot p_0 = 2.075$

2.2

Now the induced stationary chain is $P = \frac{1}{2}(P_1 + P_2)$, $R = \frac{1}{2}(R_1 + R_2)$.
So, the expected reward would be: $r = R^T \cdot p_0 + R^T \cdot P \cdot p_0 + R^T \cdot P \cdot P \cdot p_0 = 1.66$.

2.3

The optimal bellman equation is:
$V_t(s) = \max_a \left[r(a,s) + \sum_{s'} p(s'|s,a) \cdot V_{t+1}(s')\right]$
$V_2(s_0) = \max\{0.2, 0.7\} = 0.7, a_2$
$V_2(s_1) = \max\{1, 0\} = 1, a_1$
$V_2(s_2) = \max\{0.5, 0.5\} = 0.5, a_1, a2$
$V_1(s_0) = \max\{0.2 + V_2(s1) \cdot 0.5 + V_2(s2) \cdot 0.5, 0.7 + V_2(s1) \cdot 0.125 + V_2(s_2) \cdot -.875\} = 1.265, a_2$
$V_1(s_1) = \max\{1 + V_2(s0) \cdot 0.667 + V_2(s2) \cdot 0.333, 0 + V_2(s0) \cdot 0.5 + V_2(s_2) \cdot 0.5\} = 1.6333, a_1$
$V_1(s_2) = \max\{0.5 + V_2(s0) \cdot 0.75 + V_2(s1) \cdot 0.25, 0.5 + V_2(s0) \cdot 0.75 + V_2(s_1) \cdot 0.25\} = 2.2, (a_1, a_2)$
$V_0(s_0) = \max\{0.2 + V_1(s1) \cdot 0.5 + V_1(s2) \cdot 0.5, 0.7 + V_1(s1) \cdot 0.125 + V_1(s_2) \cdot 0.875\} = 2.829, a_2$

by taking the argmax:
$\pi^*_0(s0) = a2$

$$\pi^*_{1,2}(s) = \begin{cases} a_2, & s = s_0 \\ a_1, & s = s1 \\ \textit{dont care}, & s = s_2 \end{cases}$$

2.4

Considering the probability of getting thrown out of the casino is β, cumulative reward is now:

$$V = \mathbb{E}\left[\sum_{t=1}^{\infty} r(s_t, a_t)(1-\beta)^t\right]$$

Defining $\gamma = 1 = \beta$ we have the discount return infinite horizon:

$$V = \mathbb{E}\left[\sum_{t=1}^{\infty} r(s_t, a_t)\gamma^t\right]$$

2.5

the Bellman equation in that case is:

$$V(s) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \cdot V(s')$$

the Bellman optimality equation in this case is:

$$V(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) \cdot V(s')\right]$$

3 Question 3

3.1

$g_t(s=0) = \mathbb{P}(\textit{candidate t has the highest score}|\textit{there is a candidate in } [1, t-1] \textit{ which has higher score then candidate t}) = 0$
Lets define event A to be that candidate t has the; highest score. and B to be the event where candidate t has the highest score compared to all [1,t] candidates.
$g_t(s=1) = \mathbb{P}(\textit{candidate t has the highest score}|\textit{candidate t has the highest score compared to all } [1,t] \textit{ candidates}) = \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{1/N}{1/t} = \frac{t}{N}$.

3.2

$\mathbb{P}_t(1|s)$ is the probability that the t+1 candidate is the best so far given that candidate t was the best in $[1, t]$. Since we sample the candidates independently, those two variables are independent, and so:

$$\mathbb{P}_t(1|s) = \mathbb{P}_t(1) = \frac{1}{t+1}$$

.
And also: $\mathbb{P}_t(0|s) = 1 - \mathbb{P}_t(1|s) = \frac{t}{t+1}$

3.3

Lets denote p_t as the maximum probability of getting the best candidate at a later time step, i.e.: $p_t = \max\{\mathbb{P}(\textit{getting the best candidate at a later time step})\} = V^*_{t+1}(s_{t+1} = 1) \cdot \mathbb{P}_{t+1}(s_{t+1} = 1|s_t) + V^*_{t+1}(s_{t+1} = 0) \cdot \mathbb{P}_{t+1}(s_{t+1} = 0|s_t)$
Hence, we get: $V^*_t(s) = max\{g_t(s), p_t\}$
And we get: $V^*_t(0) = max\{g_t(0), V^*_{t+1}(s_{t+1} = 1) \cdot \mathbb{P}_{t+1}(s_{t+1} = 1|0) + V^*_{t+1}(s_{t+1} = 0) \cdot \mathbb{P}_{t+1}(s_{t+1} = 0|0)\}$

$V_t^*(1) = max\{g_t(1), V_{t+1}^*(s_{t+1} = 1) \cdot \mathbb{P}_{t+1}(s_{t+1} = 1|1) + V_{t+1}^*(s_{t+1} = 0) \cdot \mathbb{P}_{t+1}(s_{t+1} = 0|1)\}$
It is obvious that $V_N^*(1) = 1$ and

$$V_N^*(0) = 0$$

.

3.4

Since $g_t(0) = 0$, we get that: $V_t^*(0) = V_{t+1}^*(s_{t+1} = 1) \cdot \mathbb{P}_{t+1}(s_{t+1} = 1|0) + V_{t+1}^*(s_{t+1} = 0) \cdot \mathbb{P}_{t+1}(s_{t+1} = 0|0)$
We know that $\mathbb{P}_{t+1}(s_{t+1}|s_t) = \mathbb{P}_{t+1}(s_{t+1})$ and hence we get:
$V_t^*(1) = max\{\frac{t}{N}, V_{t+1}^*(s_{t+1} = 1) \cdot \mathbb{P}_{t+1}(s_{t+1} = 1) + V_{t+1}^*(s_{t+1} = 0) \cdot \mathbb{P}_{t+1}(s_{t+1} = 0)\} = \max\{\frac{t}{N}, V_t^*(0)\}$
and: $V_t^*(0) = \frac{1}{t+1} \cdot V_{t+1}^*(s_{t+1} = 1) + \frac{t}{t+1} \cdot V_{t+1}^*(s_{t+1} = 0)$

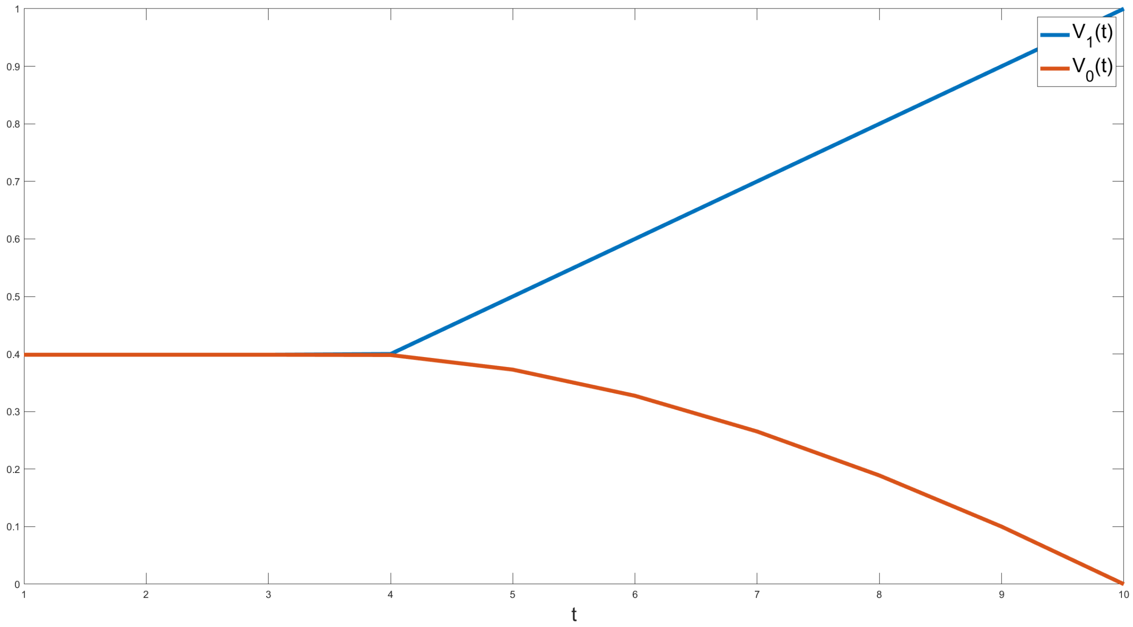

Figure 1: $V_1(t), V_0(t)$

3.5

4 Question 4

Lets define $G_t = \sum\limits_{k=t}^{\infty} \gamma^t \cdot r(s_t, a_t)$.
Since the policy is stationary $(r(s_t = s, a_t = a) = r(s_{t+\tau} = s, a_{t+\tau} = a))$, and the policy is stationary $(\pi(a_t = a|s_t = s) = \pi(a_{t+\tau} = a|s_{t+\tau} = s)$ the following holds: $E^\pi\left[G_t|s_t = s\right] = E^\pi\left[G_{t'}|s_{t'} = s\right]$.
Hence, for $t = 1, t' = 0$ we get:
$E^\pi\left[\sum\limits_{t=1}^{\infty} \gamma^t \cdot r(s_t, a_t)|s_1 = s\right] = E^\pi\left[\sum\limits_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t)|s_0 = s\right] = V^\pi(s)$

5 Question 5

5.1

The MDP would be defined by the original state and action spaces $\mathcal{S}$, $\mathcal{A}$, the original transition distribution $P(s'|s,a)$ and the reward would be defined as $\hat{r}(s) = \alpha \cdot r_1(s) + \beta \cdot r_2(s)$.
Our cumulative expected reward would be: $\mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty} \gamma^t \hat{r}(s_t)|s_0 = s_{init}\right] = \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty} \gamma^t (\alpha \cdot r_1(s_t) + \beta \cdot r_2(s_t))|s_0 = s_{init}\right] = \alpha\mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty} \gamma^t r_1(s_t)|s_0 = s_{init}\right] + \beta\mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty} \gamma^t r_2(s_t)|s_0 = s_{init}\right] = \alpha J_1^\pi + \beta J_2^\pi = f(J_1^\pi, J_2^\pi) = J^\pi$.
Hence, we get that the optimal policy π^* which maximizes the expected cumulative regret defined by $\hat{r}$, maximizes as well J^π.

5.2

The standard MDP solution approaches is relevent only for the linear objectives. The new objective that was defined for the question is not linear, and so, the standard approaches are not suitable for this kind of problems.

5.3

$J_\rho^\pi = \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t\hat{r}(s_t)|s_0 = s_{init}\right] = \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t(r_1(s_t) - \rho\cdot r_2(s_t))|s_0 = s_{init}\right] = \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t r_1(s_t)|s_0 = s_{init}\right] - \rho\mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t r_2(s_t)|s_0 = s_{init}\right] = J_1^\pi - \rho J_2^\pi$

Hence,for policy π which satisfies $J_\rho^\pi = 0$ we get:

$$J_1^\pi = \rho J_2^\pi$$

$$\Rightarrow J^\pi = \frac{J_1^\pi}{J_2^\pi} = \frac{\rho J_2^\pi}{J_2^\pi} = \rho$$

5.4

Lets assume in contradiction that we have a policy π' that satisfies $J^{\pi^{'}} > \rho$.

$$\Rightarrow J_1^{\pi^{'}} > \rho J_2^{\pi^{'}}$$

$$\Rightarrow J_\rho^{\pi^{'}} = J_1^{\pi^{'}} - \rho J_2^{\pi^{'}} > \rho J_2^{\pi^{'}} - \rho J_2^{\pi^{'}} > 0$$

. In contradiction to the fact that π^* is the optimal policy of J_ρ^π with $J_\rho^{\pi^*} = 0$.
Therefor π^* is the optimal policy for J^π

5.5

For $\rho = 0$ we get $J_0^\pi = J_1^\pi = \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t r_1(s_t)|s_0 = s_{init}\right] \geq \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t r_{min}|s_0 = s_{init}\right] = \frac{r_{min}}{1-\gamma} > 0$

5.6

For $\rho = 1$ we get $J_{\rho=1}^\pi = J_1^\pi - J_2^\pi = \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t(r_1(s_t) - r_2(s_t))|s_0 = s_{init}\right] < \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t(r_2(s_t) - r_2(s_t))|s_0 = s_{init}\right] = 0$

5.7

Since $r_2(s) > 0\ \forall s$ we can deduce that $J_2^\pi = \mathbb{E}^\pi\left[\sum\limits_{t=0}^{\infty}\gamma^t r_2(s_t)|s_0 = s_{init}\right] > 0\ \forall\pi$
For any policy π we get that for any $\rho_1 < \rho_2$:

$$J_{\rho_2}^\pi = J_1^\pi - \rho_2\cdot J_2^\pi < J_1^\pi - \rho_1\cdot J_2^\pi = J_{\rho_1}^\pi$$

Lets denote $\pi^*_{\rho_2}$ as the optimal policy for $J_{\rho_2}^\pi$, we get:

$$J_{\rho_2}^{\pi^*_{\rho_2}} < J_{\rho_1}^{\pi^*_{\rho_2}} \leq J_{\rho_1}^{\pi^*_{\rho_1}}$$

Hence, J_ρ^* is monotonically decreasing with ρ

5.8

Since $J_{\rho=0}^* > 0,\ J_{\rho=1}^* < 0$ and J_ρ^* is monotonic with ρ there exist $\rho \in [0,1]$ that satisfies $J_\rho^* = 0$.
We can search for ρ that satisfies $J_\rho^* = 0$ and as we have proved in prior sections the policy which is optimal for this specific ρ will also be optimal for J^π.