למידה במערכות דינמיות אביב תש"פ

תרגיל בית 2

מגישים:

יאיר נחום 034462796

דר ערבה 205874951

1.

a) All probabilities are greater or equal to 0. So, there can't be P_{ij}in a row I that is negative. From the total probability formula on all possible next states we get:

$$p(x_t = i) = \sum_j p(x_t = i, x_{t+1} = j)$$
$$= \sum_j p(x_t = i)p(x_{t+1} = j|x_t = i)$$
$$= p(x_t = i)\sum_j p(x_{t+1} = j|x_t = i)$$
$$\Rightarrow \sum_j p(x_{t+1} = j|x_t = i) \ = \ 1$$

In other words, this is because, given that we are in state *i*, the next state must be one of the possible states. Thus, when we sum over all the possible values of *j*, we should get one. That is, the rows of any state transition matrix must sum to one.

b) From a we get that the sum of each row is 1. Therefore, if we multiply the P matrix with a vector of $\mathbb{1} \ = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$ we get that $P\mathbb{1} \ = \ \mathbb{1}$. Thus, $\mathbb{1}$ is an eigenvector of eigenvalue 1. So, since the probabilities sum up on each row to 1, there is always an eigenvalue 1 that has an eigenvector $\mathbb{1}$.

c) Let's assume the opposite. Meaning, we have an eigenvalue $|\lambda| \ > \ 1$ which has an eigenvector $v \ \in R^d$.
We denote the maximum component of v vector as v_{imax} $s.t$: $imax \ = \ arg\max_{1\le i\le d} v_i$
and the minimum component of v vector as v_{imin} $s.t$: $imin \ = \ arg\min_{1\le i\le d} v_i$
In case $\lambda \ > \ 1$ and $v_{imax} \ \geq \ 0$:
On one hand, for $\lambda v_{imax} \ > \ v_{imax}$ and on the other hand $\lambda v_{imax} = \sum_j P_{imax,j} v_j \leq \sum_j P_{imax,j} v_{imax} = v_{imax} \sum_j P_{imax,j} = v_{imax}$
So, we got a contradiction.
In case $\lambda \ > \ 1$ and $v_{imax} \ < \ 0$:
the $v_{imin} \ < \ 0$ and on one hand, for $\lambda v_{imin} \ < \ v_{imin}$ but on the other hand $\lambda v_{imin} = \sum_j P_{imin,j} v_j \geq \sum_j P_{imin,j} v_{imin} = v_{imin} \sum_j P_{imin,j} = v_{imin}$
So, again we got a contradiction. Thus, it can't be true that $\lambda \ > \ 1$.
We can show the same thing when assuming $\lambda \ < \ -1$ symetrically when the $v_{imin} \ \leq \ 0 \ (\ if \ v_{imin} \ > \ 0 \ we \ look \ at \ v_{imax})$.
That's proves the claim that $|\lambda| \leq \ 1$.

2.

a) We've defined the reward on finite horizon (with T=3 in our case) as follows:

$$\sum_{t=0}^{T-1} r_t\,(s_t, a_t) + r_T(s_T)$$

In our case the reward is random as well, thus we define r_t as follows (depends on given current state and policy):

$$r_t(s_t, a_t) = r_t(s_t) \;= \mathbb{E}^{\pi}(R(s)|s = s_t)\; and\; r_t(s_T) = \mathbb{E}^{\pi}(R(s)|s = s_T)$$

The calculated rewards as defined (with expectation depending on the current state) are:

For π_1:

$$\mathbb{E}^{\pi_1}(R(s)|s = s0) \sim \text{Bernouli}(0.2) = 0.2$$
$$\mathbb{E}^{\pi_1}(R(s)|s = s1) \sim \text{Normal}(1,1) = 1$$
$$\mathbb{E}^{\pi_1}(R(s)|s = s2) \sim \text{Bin}(5,0.1) = 5 * 0.1 \;=\; 0.5$$

For π_2:

$$\mathbb{E}^{\pi_2}(R(s)|s = s0) \sim \text{Bernouli}(0.7) = 0.7$$
$$\mathbb{E}^{\pi_2}(R(s)|s = s1) \sim \text{Normal}(0,1) = 0$$
$$\mathbb{E}^{\pi_2}(R(s)|s = s2) \sim Bin\left(7, \frac{1}{14}\right) \;=\; 0.5$$

We have 8 possible paths (as we have 2^8 permutation on states transitions) according to the policy given (a2, a1, a2, a1).

We also note that the reward at each time step depends only on the current state (As we showed above).

The reward expectation as defined above can be written explicitly as follows:

$$r_0{}^{\pi_2}(s_0 = 0) \;+ \sum_{s' \in S} P(s'|s_0 = 0, a_0 = 2) r_1{}^{\pi_1}(s')$$
$$+ \sum_{s'' \in S} \sum_{s' \in S} P(s''|s', a_1 = 1) P(s'|s_0 = 0, a_0 = 2)\, r_2{}^{\pi_2}(s'')$$
$$+ \sum_{s''' \in S} \sum_{s'' \in S} \sum_{s' \in S} P(s'''|s'', a_2 = 2) P(s''|s', a_1 = 1) P(s'|s_0 = 0, a_0 = 2)\, r_3{}^{\pi_1}(s''')$$

When the final component is the expectation over r_T.

$0.7 + (0.125 \cdot 1 \;+\; 0.875 \cdot 0.5) + (0.125 \cdot (2/3) \cdot 0.7 \;+\; 0.125 \cdot (1/3) \cdot 0.5 \;+\; 0.875 \cdot 0.75 \cdot 0.7 \;+\; 0.875 \cdot 0.25 \cdot 0) + (0.125 \cdot (2/3) \cdot 0.125 \cdot 1 \;+\; 0.125 \cdot (2/3) \cdot 0.875 \cdot 0.5 \;+\; 0.125 \cdot (1/3) \cdot 0.25 \cdot 1 \;+\; 0.125 \cdot (1/3) \cdot 0.75 \cdot 0.2 \;+\; 0.875 \cdot 0.75 \cdot 0.125 \cdot 1 \;+\; 0.875 \cdot 0.75 \cdot 0.875 \cdot 0.5 \;+\; 0.875 \cdot 0.25 \cdot (2/3) \cdot 0.7 \;+\; 0.875 \cdot 0.25 \cdot (1/3) \cdot 0.5) = \; 0.7 \;+ 0.5625 \;+\; 0.53854 \;+\; 0.57122 = 1.80104 \;+\; 0.57122 =$ **2.37226**

b) Both the policy and the MDP are stationary therefore we can induce the homogenous Markov chain by the total probability definition over the possible actions between states.

For example, the p_{01} can be induced as follows:

$P(s' = 1|s_0 = 0) = \sum_{\alpha \in A} P(s'_1 = 1, a_0 = a|s_0 = 0) =$
$\sum_{\alpha \in A} P(s'_1 = 1|s_0 = 0, a_0 = a)\pi(a_0 = a|s_0 = 0)$

$= 0.5 \cdot 0.125 + 0.5 \cdot 0.5 = 0.3125$
$p_{02} = 1 - p_{01} = 1 - 0.3125 = 0.6875$
$p_{10} = 0.5 \cdot (2/3) + 0.5 \cdot 0.5 = 0.58333$
$p_{12} = 1 - p_{10} = 0.41666$
$p_{20} = 0.75$
$p_{21} = 0.25$

$$P = \begin{pmatrix} 0 & 0.3125 & 0.6875 \\ 0.58333 & 0 & 0.41666 \\ 0.75 & 0.25 & 0 \end{pmatrix}$$

The expected reward now can be calculated as follows:

$$\big(P(\pi(s_0) = 1|s_0) \cdot r_0(s_0, a_0 = 1) + P(\pi(s_0) = 2|s_0) \cdot r_0(s_0, a_0 = 2)\big)$$

$$+ \sum_{s' \in S, \alpha \in A} P(s'|s_0) r_1(s', a_1 = a)\, P(\pi(s_1) = a|s_1 = s')$$
$$+ \sum_{s'' \in S, s' \in S, \alpha \in A} P(s''|s')P(s'|s_0)\, r_2(s'', a_2 = a) P(\pi(s_2) = a|s_2 = s'')$$
$$+ \sum_{s''' \in S, s'' \in S, s' \in S, \alpha \in A} P(s'''|s'')P(s''|s')P(s'|s_0)\, r_3(s'', a_3 = a) P(\pi(s_3) = a|s_3 = s''')$$

Since the policy is equally distributed for all states-

$$P(\pi(s_i) = 1|s_i = s) = P(\pi(s_i) = 2|s_i = s) = 0.5 \quad \forall s \in S$$

We get the following expression-

$$\frac{1}{2}\big(r_0(s_0, a_0 = 1) + r_0(s_0, a_0 = 2)\big) + \sum_{s' \in S, \alpha \in A} P(s'|s_0) \frac{1}{2} r_1(s', a_1 = a)$$
$$+ \sum_{s'' \in S, s' \in S, \alpha \in A} P(s''|s')P(s'|s_0) \frac{1}{2} r_2(s'', a_2 = a)$$
$$+ \sum_{s''' \in S, s'' \in S, s' \in S, \alpha \in A} P(s'''|s'')P(s''|s')P(s'|s_0) \frac{1}{2} r_3(s'', a_3 = a)$$

After opening the sum by a -

$$\frac{1}{2}\big(r_0(s_0, a_0 = 1) + r_0(s_0, a_0 = 2)\big) +$$
$$\sum_{s' \in S} P(s'|s_0)\frac{1}{2}(r_1(s', a_1 = 1) + r_1(s', a_1 = 2))$$
$$+ \sum_{s'' \in S, s' \in S} P(s''|s')P(s'|s_0)\frac{1}{2}(r_2(s'', a_2 = 1) + r_2(s'', a_2 = 2))$$
$$+ \sum_{s''' \in S, s'' \in S, s' \in S} P(s'''|s'')P(s''|s')P(s'|s_0)\frac{1}{2}(r_3(s'', a_3 = 1) + r_3(s''', a_3 = 2))$$

We calculate the P matrix powers of 2 and 3 in order to simplify the probabilities between states:

$P^2 =$

$$\begin{matrix} 0.6979 & 0.1719 & 0.1302 \\ 0.3125 & 0.2865 & 0.4010 \\ 0.1458 & 0.2344 & 0.6198 \end{matrix}$$

$P^3 =$

$$\begin{matrix} 0.1979 & 0.2507 & 0.5514 \\ 0.4679 & 0.1979 & 0.3342 \\ 0.6016 & 0.2005 & 0.1979 \end{matrix}$$

Therefore, the calculation is as follows:

$(0.2 + 0.7) \cdot 0.5 + (0.3125 \cdot 1 \cdot 0.5 + 0.3125 \cdot 0 \cdot 0.5 + 0.6875 \cdot 0.5 \cdot 0.5 + 0.6875 \cdot 0.5 \cdot 0.5\,) + (0.6979 \cdot (0.2 + 0.7) \cdot 0.5 + 0.1719 \cdot (1 + 0) \cdot 0.5 + 0.1302 \cdot (0.5 + 0.5) \cdot 0.5) + (0.1979 \cdot (0.2 + 0.7) \cdot 0.5 + 0.2507 \cdot (1 + 0) \cdot 0.5 + 0.5514 \cdot (0.5 + 0.5) \cdot 0.5) = 0.45 + 0.5 + 0.465105 + 0.490105 = \mathbf{1.905}$

Or in vector representation –

$$\vec{v}_{s0}(P^0\vec{r} + P^1\vec{r} + P^2\vec{r} + P^3\vec{r})$$

When we define $\vec{r}$ as-

$$\vec{r} = \vec{r}_t = \begin{pmatrix} \frac{1}{2}(r_t(s0, a_t = 1) + r_t(s0, a_t = 2)) \\ \frac{1}{2}(r_t(s1, a_t = 1) + r_t(s1, a_t = 2)) \\ \frac{1}{2}(r_t(s2, a_t = 1) + r_t(s2, a_t = 2)) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 0.2 + 0.7 \\ 1 + 0 \\ 0.5 + 0.5 \end{pmatrix} = \begin{pmatrix} 0.45 \\ 0.5 \\ 0.5 \end{pmatrix}$$

And the initial distribution-

$$\vec{v}_{s0} = \begin{pmatrix} P(s_0 = s0) \\ P(s_0 = s1) \\ P(s_0 = s2) \end{pmatrix}^T = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}^T$$

After calculating via MATLAB (Appendix B) we get that the expectancy of the cumulative reward is – $\mathbf{1.904}$.

c) The Bellman equation for the finite horizon(T=3) problem-

$$V_3(s) = r_3(s) = \max_{a \in A}\{r_3(s, a_3 = a)\}\ \forall s \in S$$
$$\rightarrow V_3 = \begin{pmatrix} 0.7 \\ 1 \\ 0.5 \end{pmatrix} \pi_3^* = \begin{pmatrix} a_2 \\ a_1 \\ a_2 \end{pmatrix}$$

$$for\ k = T - 1, \dots 0$$
$$\boldsymbol{V_k(s) = \max_{a \in A}\{r_k(s, a) + \sum_{s' \in S} P(s'|s, a) * V_{k+1}(s')\}}$$
$$\boldsymbol{\pi_k^*(s) = \operatorname{argmax}_{a \in A}\{r_k(s, a) + \sum_{s' \in S} P(s'|s, a) * V_{k+1}(s')\}}$$

$$r_k(s, a) = \mathbb{E}[R(s)|s, a]$$

Using MATLAB code in Appendix B we can calculate the optimal value function and the matching optimal policy that gives it:

$V^* =$

2.6066	2.0198	1.2625	0.7000
2.9649	2.2667	1.6333	1.0000
2.5815	1.8552	1.2750	0.5000

$\pi^* =$

2	2	2	2
1	1	1	1
1	1	1	2

Each column is a different time index (0,1,2,3) and each row index is per state (0,1,2). We can see the policy is about the same except for the end time we defined a policy on to get the maximum reward on the last time index.
For initial state $s_0 = s0$ –

$$\pi_{t=0\dots3}^* = \begin{cases} a2 & t = 0 \\ a1 & t = 1 \\ a2\ if\ s_2 = s0,\ \ a1\ \ else & t = 2 \\ a1\ if\ s_3 = s1,\ \ a2\ else & t = 3 \end{cases}$$

d) The probability of being thrown out of the casino in each round is – β (equal chance in each round). Therefore, the chance of staying in the casino after k rounds is $(1-\beta)^k$.

The infinite horizon cumulative reward is-

$$\boldsymbol{J^{\pi}(s) = \mathbb{E}^{\pi}[\sum_{t=0}^{\infty} P(Jack\ still\ in\ casino\ at\ round\ t) * \underbrace{r_t(s_t, a_t)}_{reward\ at\ round\ t} |s_0 = s)]}$$

$$\boldsymbol{= \mathbb{E}^{\pi}[\sum_{t=0}^{\infty} \underbrace{(1-\beta)^t}_{discount\ factor:\ \gamma^t} * r_t(s_t, a_t)|s_0 = s)]}$$

The connection between the death rate and the discount factor is - $\boldsymbol{(1-\beta) = \gamma}$

e) The Bellman equations for the infinite horizon problem-

$$\boldsymbol{V(s) = \max_{a \in \{a_1, a_2\}} \{r(s,a) + \underbrace{(1-\beta)}_{\gamma} \sum_{s' \in \{s0, s1, s2\}} P(s'|s,a) * V(s')\}}$$

$$\boldsymbol{\pi_k^*(s) = \underset{a \in \{a_1, a_2\}}{\mathrm{argmax}} \{r(s,a) + \underbrace{(1-\beta)}_{\gamma} \sum_{s' \in \{s0, s1, s2\}} P(s'|s,a) * V(s')\}}$$

As proved in lecture, solution of this non-linear equation is optimal value function V^* (theorem 1).

3.

a) Let $g_t(s)$ be the probability that after observing $t-1$ candidates the tth candidate has the highest score (of all N candidates).
Let c_t be the score of the tth candidate.
By observing $t-1$ candidates we know either if-

a. One of them has a higher score than c_t and therefore c_t is not the highest score seen so far $\leftrightarrow s = 0$.

Or -

b. c_t is higher than all the scores seen so far one $\leftrightarrow s = 1$.

Therefore -

$$\boldsymbol{g_t(s=0)} = P(c_t \text{ is highest score } | c_t \text{ is not highest in } t)$$
$$= P(c_t > c_i \ \forall i \in [1,N] | c_i > c_t \ \exists i \in [1, t-1]) = \mathbf{0}$$

$$\boldsymbol{g_t(s=1)} = P(c_t \text{ is highest score } | c_t \text{ is highest in } t)$$

$$= P(c_t > c_i \ \forall i \in [1,N] \ | c_i < c_t \ \forall i \in [1, t-1])$$
$$= P(\text{ highest score in } t \ \cap c_t \text{ is highest in } t \ | \ c_t \text{ is highest in } t \ )$$
$$= P(\text{highest score in } t \ )$$
$$= \frac{\text{Number of candidates subsets of size } t \text{ including highest scoring candidate}}{\text{Number of candidates subsets of size } t}$$
$$= \frac{\binom{N-1}{t-1}}{\binom{N}{t}} = \frac{\frac{N-1!}{N-1-t+1!\,t-1!}}{\frac{N!}{N-t!\,t!}} = \boldsymbol{\frac{t}{N}}$$

Another way to think of it, is as follows:

$$\boldsymbol{g_t(s=1)} = P(c_t \text{ is highest score } | c_t \text{ is highest in } t)$$
$$= P(c_t \text{ is highest score} \cap c_t \text{ is highest in } t)/P(c_t \text{ is highest in } t)$$
$$= P(c_t \text{ is highest score})/P(c_t \text{ is highest in } t)$$
$$= \frac{1/N}{1/t} = t/N$$

b) Let $P_t(1|s)$ be the transition probability of the following candidate be the highest scoring candidate seen so far (with knowing s of the current state).
In the same way $P_t(0|s)$ will be the transition probability of the following candidate to **not** be the highest scoring candidate seen so far (with knowing s of the current state).
Clearly $P_t(j|s) = P_t(j)$ since the transition probability in independent of the current state. For the candidate selected at time $t+1$ (the next state) to be $s = 1$ it must be larger than all t scores prior to it. The probability $P_t(j)$ is the same whether c_t was the largest(current state $s = 1$) or c_i ($i \in [1, t-1]$) was the largest of the first t states(current state - $s = 0$).

$$\boldsymbol{P_t(1|s)} = P_t(1) = P(c_{t+1}\ is\ highest\ score\ seen\ till\ now\ )$$
$$= P(c_i < c_{t+1}\ \forall i \in [1,t]) \underset{(*)}{=} \boldsymbol{\frac{1}{t+1}}$$

$$\boldsymbol{P_t(0|s)} = P_t(0) = P(c_{t+1}\ is\ not\ highest\ score\ seen\ till\ now\ )$$
$$= 1 - P(c_{t+1}\ is\ highest\ score\ seen\ till\ now\ ) = \boldsymbol{\frac{t}{t+1}}$$

$(*)$ – One of the $t+1$ candidates has the highest score. Since they are randomly(uniformly) selected, each has same probability of being the largest in the group.
One can think of it as selecting the last place in the t+1 places for the largest number in these t+1 places.

c) 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.

$$\mathrm{V}_t^*\ (1)\ =\ \max\{\ \underbrace{\mathrm{g}_t(1)}_{choosing\ c_t\ as\ best\ candidate}\ ,\ \underbrace{\mathrm{P_t}\ (1|1)\mathrm{V}_{t+1}^*(1)\ +\ \mathrm{P}_t\ (0|\ 1)V^*{}_{t+1}\ (0)}_{choosing\ to\ discard\ c_t}\}$$

$$\mathrm{V}_t^*\ (0)\ =\ \max\{\ \underbrace{\mathrm{g}_t(0)}_{choosing\ c_t\ as\ best\ candidate}\ ,\ \underbrace{\mathrm{P_t}\ (1|0)\mathrm{V}_{t+1}^*(1)\ +\ \mathrm{P}_t\ (0|\ 0)V^*{}_{t+1}\ (0)}_{choosing\ to\ discard\ c_t}\}$$

$V_t^*(s)$ is actually the maximum probability to select the best candidate given we are in state s. we can think of it as computing this value (probability function) in DP and deciding when to stop (optimal stop time) according to it. If we're in s = 1 for example, we select between the immediate reward of $\mathrm{g}_t(1)$ compared to selecting the possible future reward. The future reward is the expected value of $V_{t+1}^*(s)$. Meaning, $\sum_{s' \in S} p(s_{t+1} = s' | s_t = s)\ \mathrm{V}_{t+1}^*\ (s')$.
The action space has only one 2 actions, stop or continue and we max over it.

If we got to the last candidate (and haven't chosen one of the $N-1$ candidates beforehand)-

$\mathbf{g_N(1) = 1}$ since all candidates have been reviewed -
$P(c_N\ is\ highest\ score\ |c_N\ is\ highest\ score\ seen\ till\ now) = 1$

Therefore, the maximal probability is $\mathbf{V_N^*\ (1) = 1.}$

$\mathbf{g_N(0) = 0}$ always, and since we have reached the final candidate, choosing the highest scoring candidate is impossible - $\mathbf{V_N^*\ (0) = 0.}$

d) From all previous results-

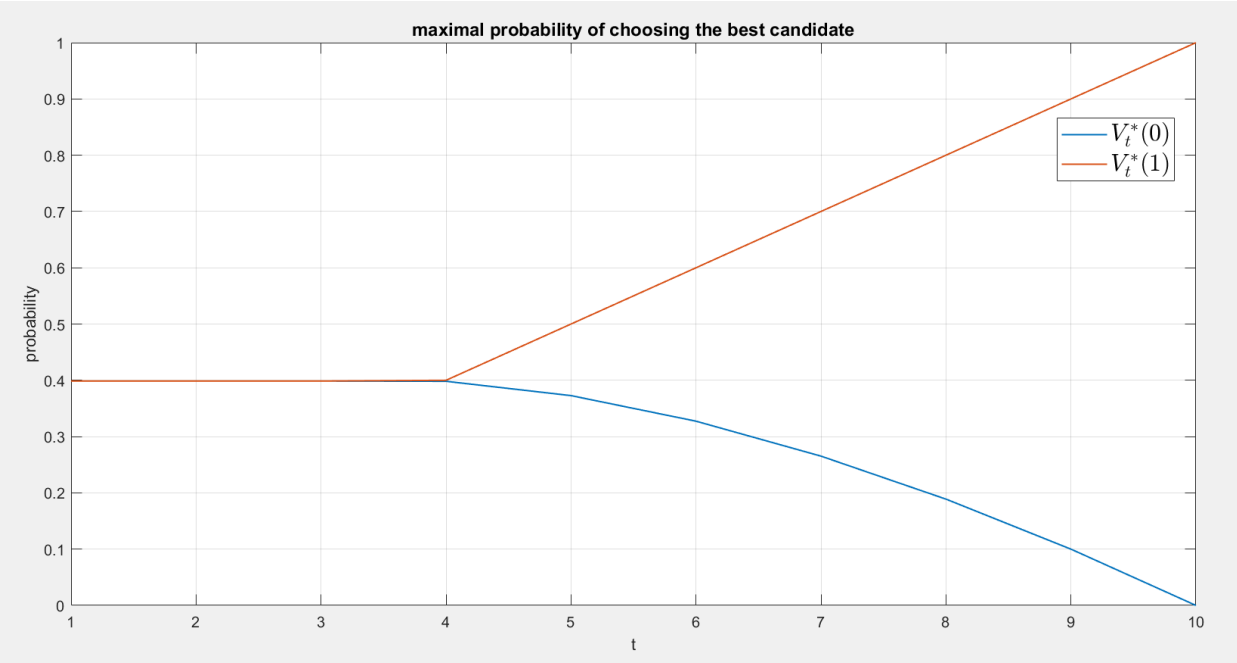
$$\mathrm{g}_t(1) = \frac{t}{N}, \qquad \mathrm{g}_t(0) = 0, \qquad P_t(1|s) = \frac{1}{t+1}, \qquad P_t(0|s) = \frac{t}{t+1}$$

We get-

$$\mathrm{V}_t^*\ (0) = \max\left\{0,\ \frac{1}{t+1}\mathrm{V}_{t+1}^*(1) + \frac{t}{t+1}V^*{}_{t+1}\ (0)\right\} = \boldsymbol{\frac{1}{t+1}\mathrm{V}_{t+1}^*(1) + \frac{t}{t+1}V^*{}_{t+1}\ (0)}$$

$$\mathrm{V}_t^*\ (1) = \max\left\{\frac{t}{N},\ \underbrace{\frac{1}{t+1}\mathrm{V}_{t+1}^*(1) + \frac{t}{t+1}V^*{}_{t+1}\ (0)}_{\mathrm{V}_t^*\ (0)}\right\} = \mathbf{max}\left\{\boldsymbol{\frac{t}{N}}, \mathbf{V}_t^*\ (\mathbf{0})\right\}$$

After solving the induction numerically for N=10 via MATLAB (see code in Appendix A), we got the following result:



e) The optimal strategy for choosing a candidate is the strategy that chooses a candidate at moment t when $\mathrm{V}_t^*\ (\mathrm{s})$ is maximal (highest probability of being the best of all).
As seen from section (4), there in an initial period where $\mathrm{V}_t^*\ (1) = \mathrm{V}_t^*\ (0)$, meaning that the maximal probability of choosing the best candidate is the same regardless to whether or not the current candidate is the best seen so far.
After this first initial period, $\mathrm{V}_t^*\ (1)$ in monotonal increasing while $\mathrm{V}_t^*\ (0)$ is monotonal decreasing. In order to choose the candidate at the highest $\mathrm{V}_t^*(s)$ we want to choose at a point in time where $s = 1$(the current candidate is best seen). This way $\mathrm{V}_t^*\ (\mathrm{s}) = \mathrm{V}_t^*\ (1)$ and we are located on the increasing probability plot.
After waiting an initial period, we are at risk that the best candidate is chosen in the initial discarded group (and then for all the remaining candidates $s = 0 \leftrightarrow$ decreasing $\mathrm{V}_t^*\ (\mathrm{s})$). In this case the strategy will wait for $s = 1$, get to the last candidate and be forced to hire him.
The initial period τ is as seen since there is a tradeoff between the probability that the best candidate will be chosen in the τ period will be low (short τ) and the probability that the best candidate will remain unseen is low and we get a better chance to hire someone good(long τ).
It turns out that this period converges to $\tau \approx N/e$.

4. Need to prove the following equation:

$$V^{\pi}(s) = \mathbb{E}^{\pi}[(\textstyle\sum_{t=0}^{\infty} \gamma^{t} r(s_t, a_t))|\, s_0 = s] \;=\; \mathbb{E}^{\pi}[(\textstyle\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t))|\, s_1 = s]$$

$$V^{\pi}(s) = \mathbb{E}^{\pi}\left[\left(\sum_{t=0}^{\infty} \gamma^{t} r(s_t, a_t)\right)|\, s_0 = s\right] \;=\; \mathbb{E}^{\pi}\left[\left(\sum_{k=1}^{\infty} \gamma^{k-1} r(s_{k-1}, a_{k-1})\right)|\, s_0 = s\right]$$

Since the policy is a given $r\big(s_i = s, a_i = \pi(s)\big) \;= r(s_i = s)$ and we actually have a regular markov chain w/ stationary dynamics. We can prove with induction that every component in the summary of the rewards is distributed the same relative to its distance from the initial stage (s0). Thus, "moving" the difference in time in this homogeneous MC has the same probability and thus the same expectation.

We need to prove:

$$\mathbb{E}^{\pi}\left[\left(\sum_{k=1}^{\infty} \gamma^{k-1} r(s_{k-1}, a_{k-1})\right)|\, s_0 = s\right] = \mathbb{E}^{\pi}\left[\left(\sum_{k=1}^{\infty} \gamma^{k-1} r(s_k, a_k)\right)|\, s_1 = s\right]$$

We will show that the distribution of $r(s_{k-1})|\, s_0 = s \;\; is\ distributed\ the\ same\ as\ r(s_k)||\, s_1 = s$ and thus the expectation over the summary above is the same (as expectation and sum can be swapped).

Base of the induction:

$r(s_0 = s) = r(s_1 = s)$ as the start state is the same and it's deterministic.

Step of the induction:

we assume that up to N the sum has the same expectation, thus we need to show (ignoring the γ^{k-1} as its constant relative to the expectation) that:

$$\mathbb{E}^{\pi}\big[\big(r(s_N)\big)|\, s_0 = s\big] = \mathbb{E}^{\pi}\big[\big(r(s_{N+1})\big)|\, s_1 = s\big]$$

This is the same as showing that the transition probabilities to reach from the start stage to the current stage are the same.

This is true due to the fact that we talk about stationary MDP and a fixed policy. Which gives a regular homogeneous MC (as we saw in the lectures)

$$P^{\pi}\big[\big(r(s_N)\big)|\, s_0 = s\big] = P_{ss'}^{(N)} = P_{ss'}^{N} = P^{\pi}\big[\big(r(s_{N+1})\big)|\, s_1 = s\big]$$

Therefore, for every component of the infinite sum we get the same distribution and expectation:

$$\mathbb{E}^{\pi}\left[\left(\sum_{k=1}^{\infty} \gamma^{k-1} r(s_{k-1}, a_{k-1})\right)|\, s_0 = s\right] \;=\; \mathbb{E}^{\pi}\left[\left(\sum_{k=1}^{\infty} \gamma^{k-1} r(s_k, a_k)\right)|\, s_1 = s\right]$$

■

5.

a) Let $M^{\pi}(s)$ be the second moment of the discounted return when starting from state s and following policy π.

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

$$\mathbb{E}^{\pi,s}\left[\left(\gamma^0 r(s_0,a_0) + \sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\right)^2\right] =$$

$$\mathbb{E}^{\pi}\left[\left(r(s_0,a_0)\right)^2 + \left(\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\right)^2 + 2r(s_0,a_0)\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t) \mid s_0 = s\,, a_0 = \pi_0(s)\right] =$$

$$\left(r\left(s,\pi_0(s)\right)\right)^2 + \underbrace{\mathbb{E}^{\pi}\left[\left(\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\right)^2 \mid s_0 = s\,, a_0 = \pi_0(s)\right]}_{\mathcal{A}}$$

$$+ 2r(s,a_0)\underbrace{\mathbb{E}^{\pi}\left[\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t) \mid s_0 = s\,, a_0 = \pi_0(s)\right]}_{\mathcal{B}}$$

$$\mathcal{A} \underset{(*)}{=} \mathbb{E}^{\pi}\left[\underbrace{\mathbb{E}^{\pi}\left[\left(\sum_{t=1}^{\infty}\gamma^t r(s_t,a_t)\right)^2 \mid s_1 = s'\right]}_{=\gamma^2 M^{\pi}(s')} \mid s_0 = s\,, a_0 = \pi_0(s)\right]$$

$$= \gamma^2 \mathbb{E}^{\pi}[M^{\pi}(s') \mid s_0 = s\,, a_0 = \pi_0(s)]$$

$$= \gamma^2 \sum_{s'} P(s' \mid s_0 = s\,, a_0 = \pi_0(s)) M^{\pi}(s')$$

$(*) - smoothing\ theorem$

Similarly -

$$\mathcal{B} = \gamma \sum_{s'} P(s' \mid s_0 = s\,, a_0 = \pi_0(s)) V^{\pi}(s')$$

We got a linear dependency of M^{π} on M^{π} & V^{π}:

$$\boldsymbol{M^{\pi}(s) = \left(r\left(s,\pi_0(s)\right)\right)^2 + \gamma^2 \sum_{s'} P(s_1 = s' \mid s_0 = s\,, a_0 = \pi_0(s)\, M^{\pi}(s')}$$

$$\boldsymbol{+ 2\gamma r(s,a_0) \sum_{s'} P(s_1 = s' \mid s_0 = s\,, a_0 = \pi_0(s)) V^{\pi}(s')}$$

b) In order to calculate $M^{\pi}(s)$ for all states, we solve the equation above and the equation for $V^{\pi}(s)$:

$$(1): \; M^{\pi}(s) = \left(r\left(s,\pi_0(s)\right)\right)^2 + \gamma^2 \sum_{s'} P(s_1 = s' | s_0 = s\,, a_0 = \pi_0(s)\, M^{\pi}(s')$$
$$+ 2\gamma r(s,a_0) \sum_{s'} P(s_1 = s' | s_0 = s\,, a_0 = \pi_0(s)) V^{\pi}(s')$$

$$(2): \quad V^{\pi}(s) = r\left(s,\pi_0(s)\right) + \gamma \sum_{s'} P(s_1 = s' | s_0 = s\,, a_0 = \pi_0(s)) V^{\pi}(s')$$

A total of $\mathbf{2|S|}$ **equations** for $2|S|$ variables-$\{V^{\pi}(s) : s \in S\}$ & $\{M^{\pi}(s) : s \in S\}$.

c) Let $W^{\pi}(s)$ be the variance of the discounted return when starting from state s and following policy π.

$$W^{\pi}(s) = Var^{\pi,s}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$
$$= \underbrace{\mathbb{E}^{\pi,s}\left[\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right)^2\right]}_{=M^{\pi}(s)} - \underbrace{\left(\mathbb{E}^{\pi,s}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]\right)^2}_{=\left(V^{\pi}(s)\right)^2} = M^{\pi}(s) - \left(V^{\pi}(s)\right)^2$$

In order to calculate $W^{\pi}(s)$, we first calculate $M^{\pi}(s)$ and $V^{\pi}(s)$ (as described in equations form previous section). Secondly, we subtract $\left(V^{\pi}(s)\right)^2$ from $M^{\pi}(s)$.

Appendix A

Code for question 3 section 4.

```
clc; %clear all;

N = 10;
t = 1:1:N;
Vt0 = zeros(1,N);
Vt1 = zeros(1,N);

Vt0(N) = 0;
Vt1(N) = 1;

for i = (N-1):-1:1
    Vt0(i) = (1/(i+1))*Vt1(i+1)+(i/(i+1))*Vt0(i+1);
    Vt1(i) = max(i/N,Vt0(i));
end

plot(t,Vt0,t,Vt1,'LineWidth',1); grid on;
title('maximal probability of choosing the best candidate',
'fontsize', 12);
ylabel('probability');
xlabel('t');
legend('$V^*_t(0)$','$V^*_t(1)$', 'Interpreter','latex', 'fontsize',
16);
```

Appendix B

Code for question 2 section b and c.

% for fixed policy section b

```
P = [0 0.3125 0.6875; 0.58333 0 0.41666; 0.75 0.25 0];
r = [0.45 0.5 0.5]';
v0= [1 0 0];
res = (v0+v0*P+v0*(P^2)+v0*(P^3))*r;
```

% for optimal policy section c

```
P_a1 = [    0,      0.5,    0.5;
            2/3,    0,      1/3;
            0.75,   0.25,   0];

P_a2 = [    0,      0.125,  0.875;
            0.5,    0,      0.5;
            0.75,   0.25,   0];

T=3;
V = zeros(3,T+1);
VT = [ 0.7; 1; 0.5];
R_a1 = [ 0.2; 1; 0.5];
R_a2 = [ 0.7; 0; 0.5];
PiT = [ 2; 1; 2];
V(:,T+1) = VT;
Pi(:,T+1) = PiT;
for t=T:-1:1
    a1_val = R_a1 + P_a1 * V(:,t+1);
    a2_val = R_a2 + P_a2 * V(:,t+1);
    A_val(:,1) = a1_val;
    A_val(:,2) = a2_val;
    [V(:,t), Pi(:,t)] = max(A_val,[],2);
end

V
Pi
```