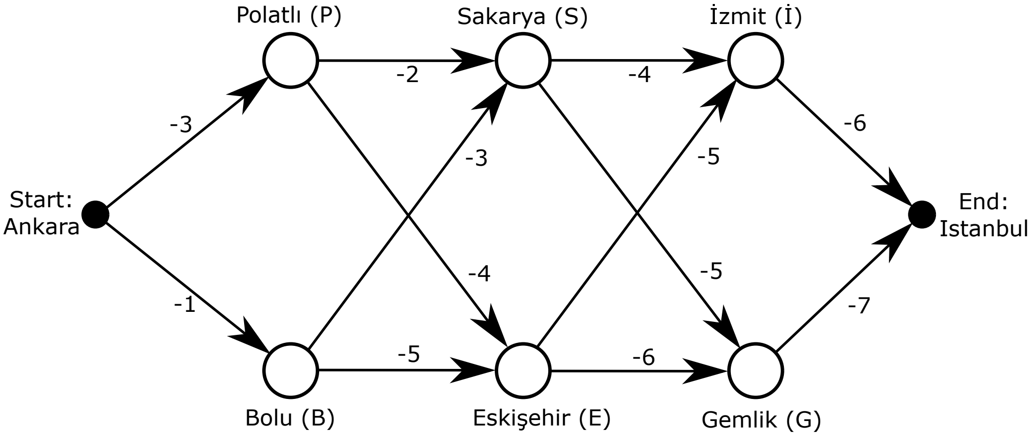
IE-456/556 & EEE-448/548
Reinforcement Learning and Dynamic Programming
Homework Assignment 2 – Due June 27 23:59

1. (60 points) Imagine you are planning a trip to Istanbul with your friends during the upcoming holidays and are evaluating which route to take. The problem is depicted in the following picture:



Consider the state space as {Start:Ankara, Polatlı, Bolu, Sakarya, Eskişehir, İzmit, Gemlik, End:Istanbul} and the rewards are given as negative numbers next to the arrows and represent the satisfaction of traveling between two cities.

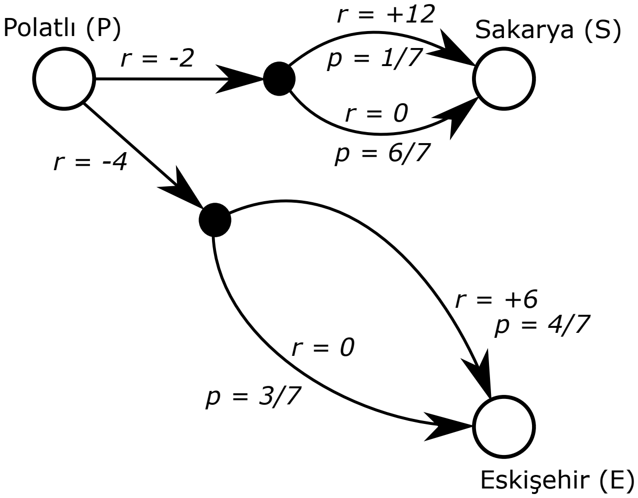
(a) **Policy Evaluation**: With the help of matrix inversion, find the state-value function of each state for $\gamma = 0.9$ for the uniform random policy (50/50 policy: After visiting a city - e.g. "Polatlı", there is a 0.5 probability to go *up* to the "Sakarya" and 0.5 probability to go *down* to "Eskişehir").

(b) **Exhaustive Search**: Enumerate all deterministic policies for this problem (all possible path-permutations) and evaluate them all for $\gamma = 1$. Which one is the best path concerning the satisfaction in the travel?

(c) **Dynamic Programming**: Use value iteration algorithm to find the best policy for $\gamma = 1$. After how many iterations were you able to find the optimal values? How we can use these values to express the optimal policy?

2. (40 points) For $\gamma = 1$, suppose in the previous question that there are some special events that occur in these cities on certain days of the week. However, you don't know which day of the week each event will take place. You only know, for example, that "Sakarya" has one special event per week, but "Eskişehir" has four special events per week. So, the chance to get a positive reward in the "Eskişehir" is higher than in the "Sakarya".

In our problem we have deterministic state transitions because we reach the city we plan to visit for sure. But the reward in our case has a stochastic offset. If it is the day with a special event (randomly, we do not know the weekday), we get an additional positive reward. The probability to get that is dependent on the number of special events per week. For example:



As can be seen, in "Sakarya" we have in 1 of 7 cases (days) an additional special-event-reward and in "Eskişehir" in 4 of 7 cases.

The probability to get the positive reward (special event) is defined by:

$$p_{xr_+} = \begin{bmatrix} \frac{3}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & 0 & 0 \\ \frac{6}{7} & \frac{4}{7} & \frac{4}{7} & \frac{5}{7} & \frac{5}{7} & 0 & 0 \end{bmatrix}$$

$$r_+ = \begin{bmatrix} 12 & 16 & 16 & 16 & 16 & 0 & 0 \\ 6 & 10 & 10 & 8 & 8 & 0 & 0 \end{bmatrix}$$

The probability to get the no extra reward is $p_{xr_-} = 1 - p_{xr_+}$. In that case you take the "no extra reward" r_-, which is zero in all cases. The actions to choose are *up* and *down* (for states İ and G, *up* and *down* mean the same).

So if you are at "Ankara" and you go *up* you get a fixed reward of -3 on the way. With a probability of $p_{x=0,r_+} = \frac{3}{7}$ there is a special event in "Polatlı" and you get an additional reward of $+12$.

Apply **policy iteration** algorithm to find the optimal policy and optimal value functions starting from the policy prescribing *up* in all states. Report the policies and the value functions you find in all iterations.