Assignment 2

Electrical and Electronics Engineering Department, Bilkent University

Instructor: Muhammed O. Sayin

Posted on Oct. 9 **Due Date: Oct. 27**

Disclaimer: *These assignments shall not be distributed outside this class.*

Content: Markov Decision Processes and Dynamic Programming
Recommended Reading: Sutton and Barto: Chapters 3 and 4

Problem 1. (40pt) The Fibonacci numbers, commonly denoted by $F(n)$, form a sequence, called the **Fibonacci sequence**, such that each number is the sum of the two preceding ones, starting from 0 and 1. That is

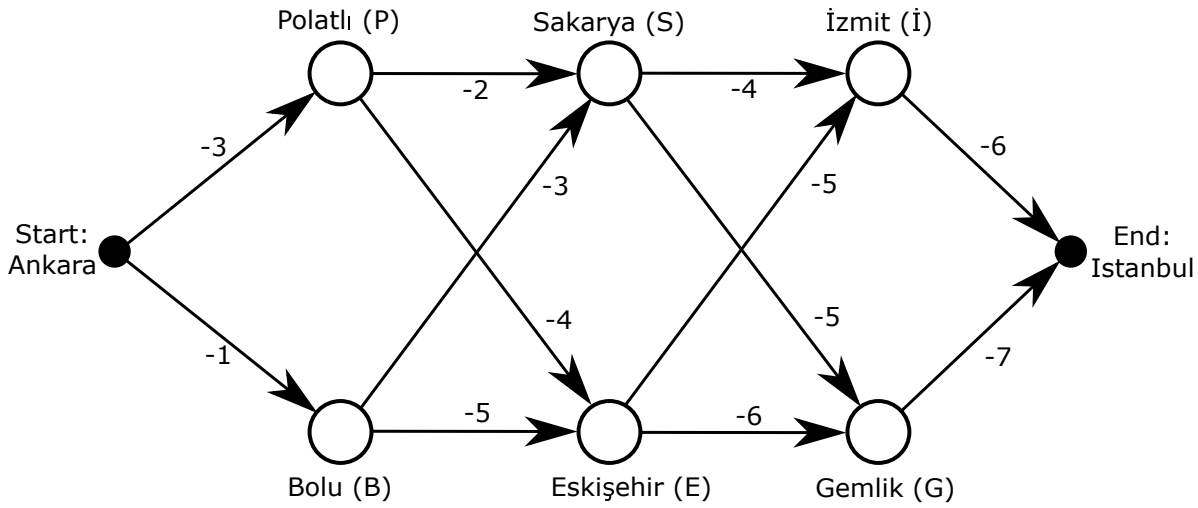
$$F(0) = 0, \quad F(1) = 1$$
$$F(n) = F(n-1) + F(n-2), \quad \forall n > 1.$$

Write the pseudocode of an algorithm that takes n as an input and outputs the nth Fibonacci number in the following three ways:

- The naïve solution with recursion
- The bottom-up solution, where you compute smaller values first and then build larger values from them
- The top-down solution with memoization, where you first break down the problem into subproblems and then calculate and store values

Implement these algorithms in whatever computational software or programming language you are comfortable with. **Compute** the 45th Fibonacci number via these algorithms. **Report and compare** the computation times.

Problem 2. (40pt) Imagine you are planning a trip from Ankara to Istanbul with your friends. You are evaluating which route to take. The following is an illustration of the problem:



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

- **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., "Sakarya", there is a 0.5 probability to go to "İzmit" and 0.5 probability to go to "Gemlik").
- **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?
- **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 can we use these values to express the optimal policy?

Problem 3. (20pt) Consider an infinite horizon Markov decision process characterized by the tuple $\langle S, A, R, P, \gamma \rangle$. We would like to solve this problem by using finite horizon approximation. Derive the horizon length K guaranteeing $\epsilon > 0$ approximation error. You will find a lower bound on K in terms of γ, ϵ, and the reward function $R(\cdot)$.

For example, given K, let $\tilde{\pi} = \{\tilde{\pi}_k\}_{k=0}^{\infty}$ be a policy defined such that $\{\tilde{\pi}_k\}_{k=0}^{K-1}$ is the best strategy for the objective

$$\mathrm{E}\left[\sum_{k=0}^{K-1} \gamma^k R(s_k, a_k) \mid s_0 = 0\right]$$

and $\{\tilde{\pi}_k\}_{k=K}^{\infty}$ is arbitrary. Then, K derived must lead to

$$\max_{\pi} \mathrm{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k) \mid s_0 = 0\right] - \epsilon \leq \mathrm{E}_{\tilde{\pi}}\left[\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k) \mid s_0 = 0\right],$$

where $\mathrm{E}_{\pi}[\cdot]$ (and $\mathrm{E}_{\tilde{\pi}}[\cdot]$) denotes the expectation taken with respect to the randomness on (s_k, a_k) induced by the policy π (and the policy $\tilde{\pi}$) and the underlying transition kernel.