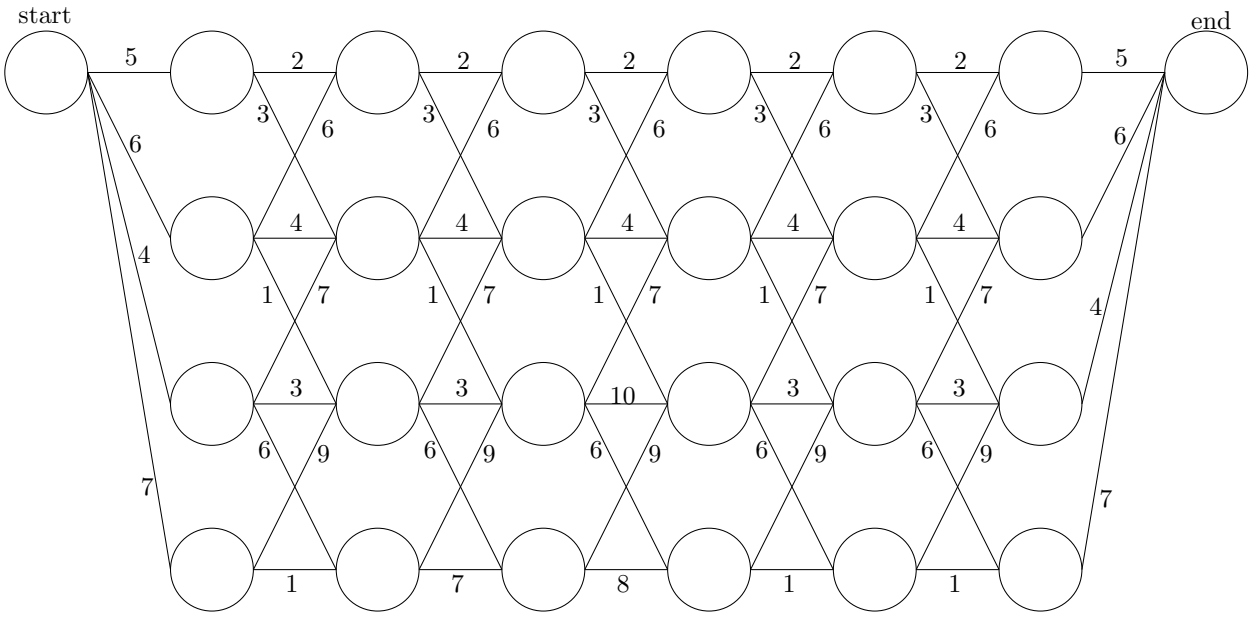
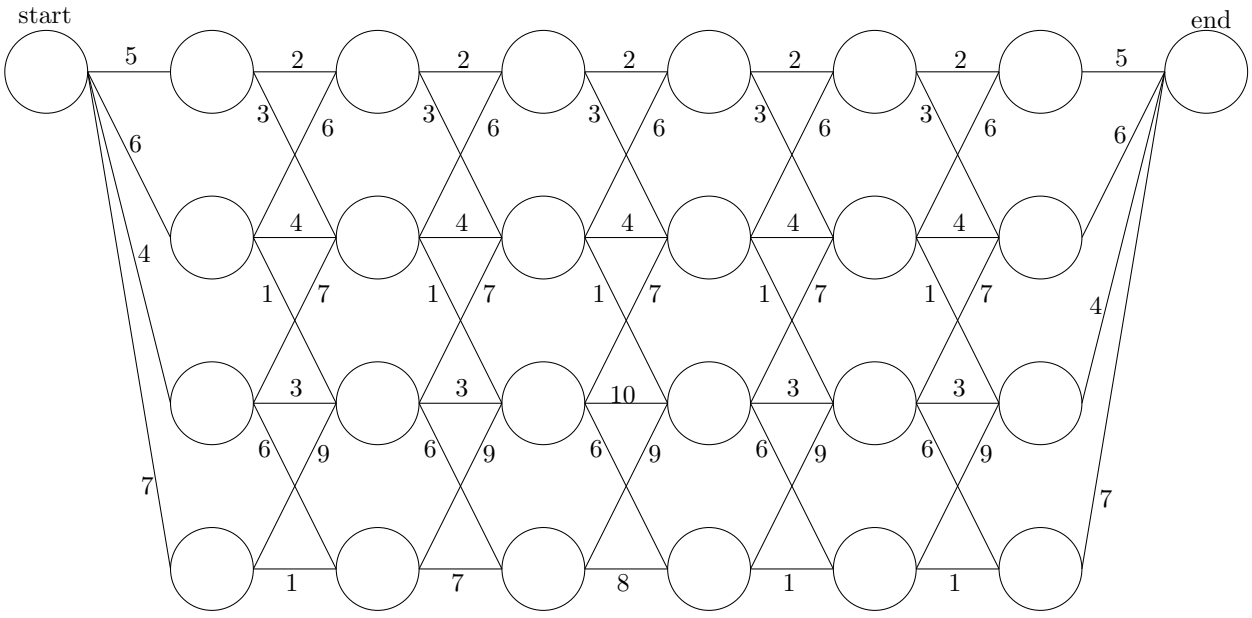
EE 526X Deep Machine Learning: Theory and Practice — Homework 5
Assigned: 11/24. Due: 12/08

Problem 1.

(a) Find the shortest path from `start` to `end` in the following figure.



(b) Find the longest path from `start` to `end` in the following figure.



Problem 2. Consider a MDP with two states S={0, 1}, two actions $A = \{1, 2\}$, and the follow reward function

$$R_s^{(a)} = \begin{cases} 1, & (s,a) = (0,1) \\ 4, & (s,a) = (0,2) \\ 3, & (s,a) = (1,1) \\ 2, & (s,a) = (1,2) \end{cases} \tag{1}$$

and the transition probabilities $P_{ss'}^{(a)}$ as follows:

$$\begin{bmatrix} P_{00}^{(1)} & P_{00}^{(2)} \\ P_{10}^{(1)} & P_{10}^{(2)} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{4} & \frac{2}{3} \end{bmatrix} \tag{2}$$

The other probabilities can be deduced, for example:

$$P_{01}^{(1)} = 1 - P_{00}^{(1)} = 1 - \frac{1}{3} = \frac{2}{3}. \tag{3}$$

The discount factor is

$$\gamma = 3/4. \tag{4}$$

(a) For the policy π that chooses action 1 in state 0, and action 2 in state 1, find the state value function $v_\pi(s)$, by writing out the Bellman's expectation equation, and solve the equation explicitly.

(b) For the same policy π, obtain the state value function using iterative update based on the Bellman's expectation equation. You need to list the first 5 iteration values of $v(s)$.

(c) For the policy π, calculate the $q_\pi(s,a)$ function.

(d) Based on the value function $v_\pi(s)$, obtain an improved policy π' based on

$$\pi'(s) = \arg\max_a q_\pi(s,a). \tag{5}$$

(e) Obtain the optimal value function $v_*(s)$ using value iteration based on the Bellman's optimality equation, with all initial values set to 0.

(f) Obtain the optimal policy.

Problem 3. Consider a MDP with two states S={0, 1}, two actions $A = \{1, 2\}$, and the follow reward function

$$R_s^{(a)} = \begin{cases} 1, & (s,a) = (0,1) \\ 4, & (s,a) = (0,2) \\ 3, & (s,a) = (1,1) \\ 2, & (s,a) = (1,2) \end{cases} \tag{6}$$

and the transition probabilities $P_{ss'}^{(a)}$ as follows:

$$\begin{bmatrix} P_{00}^{(1)} & P_{00}^{(2)} \\ P_{10}^{(1)} & P_{10}^{(2)} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{4} & \frac{2}{3} \end{bmatrix} \tag{7}$$

The other probabilities can be deduced, for example:

$$P_{01}^{(1)} = 1 - P_{00}^{(1)} = 1 - \frac{1}{3} = \frac{2}{3}. \tag{8}$$

The discount factor is

$$\gamma = 3/4. \tag{9}$$

Exercise on model-free prediction:

(a) For the policy π that chooses action 1 in state 0, and action 2 in state 1, starting from state 0, generate one episode E of 10000 triplets of (R_i, S_i, A_i), i=0, 2, ..., 9999, with $R_0 = 0$, $S_0 = 0$.

(b) Based on the episode E, use Monte Carlo policy evaluation to estimate the value function $v_\pi(s)$.

(c) Based on the episode E, use n-step temporal difference policy evaluation to estimate the value function $v_\pi(s)$.

Exercise on model-free control:

(a) Use the SARSA algorithm to estimate the optimal action-value function $q_*(s, a)$, by running the algorithm in Sutton and Barto's book (2nd edition, available online).

(b) Use the Q-learning algorithm to estimate the optimal action-value function $q_*(s, a)$, by running the algorithm in Sutton and Barto's book (2nd edition, available online).

You only need to simulate one episode. In both cases, you will need to decide an appropriate fixed step-size α, and exploration probability ϵ, and number of time steps in the episode.

END OF ASSIGNMENT