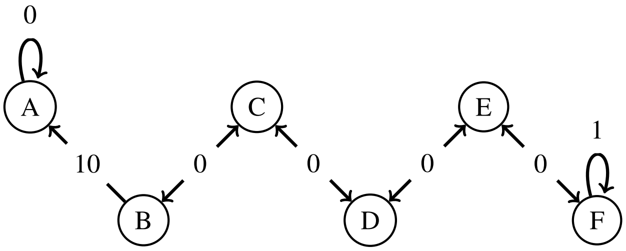
CS 188 Summer 2020 Introduction to Artificial Intelligence

Exam Prep 2

1 MDP



Consider the MDP above, with states represented as nodes and transitions as edges between nodes. The rewards for the transitions are indicated by the numbers on the edges. For example, going from state B to state A gives a reward of 10, but going from state A to itself gives a reward of 0. Some transitions are not allowed, such as from state A to state B. Transitions are deterministic (if there is an edge between two states, the agent can choose to go from one to the other and will reach the other state with probability 1).

label=() For this part only, suppose that the max horizon length is 15. Write down the optimal action at each step if the discount factor is $\gamma = 1$.

lbbel=() Now suppose that the horizon is infinite. For each state, does the optimal action depend on γ? If so, for each state, write an equation that would let you determine the value for γ at which the optimal action changes.

Q2. MDPs and RL: Mini-Grids

The following problems take place in various scenarios of the gridworld MDP (as in Project 3). In all cases, A is the start state and double-rectangle states are exit states. From an exit state, the only action available is *Exit*, which results in the listed reward and ends the game (by moving into a terminal state X, not shown).

From non-exit states, the agent can choose either *Left* or *Right* actions, which move the agent in the corresponding direction. There are no living rewards; the only non-zero rewards come from exiting the grid.

Throughout this problem, assume that value iteration begins with initial values $V_0(s) = 0$ for all states s.

First, consider the following mini-grid. For now, the discount is $\gamma = 1$ and legal movement actions will always succeed (and so the state transition function is deterministic).

+1	A			+10

(a) What is the optimal value $V^*(A)$?

(b) When running value iteration, remember that we start with $V_0(s) = 0$ for all s. What is the first iteration k for which $V_k(A)$ will be non-zero?

(c) What will $V_k(A)$ be when it is first non-zero?

(d) After how many iterations k will we have $V_k(A) = V^*(A)$? If they will never become equal, write *never*.

Now the situation is as before, but the discount γ is less than 1.

(e) If $\gamma = 0.5$, what is the optimal value $V^*(A)$?

(f) For what range of values γ of the discount will it be optimal to go *Right* from A? Remember that $0 \leq \gamma \leq 1$. Write *all* or *none* if all or no legal values of γ have this property.

Finally, consider the following mini-grid (rewards shown on left, state names shown on right).

+4	A	+16

L	A	R

In this scenario, the discount is $\gamma = 1$. The failure probability is actually $f = 0$, but, now we do not actually know the details of the MDP, so we use reinforcement learning to compute various values. We observe the following transition sequence (recall that state X is the end-of-game absorbing state):

s	a	s'	r
A	*Right*	R	0
R	*Exit*	X	16
A	*Left*	L	0
L	*Exit*	X	4
A	*Right*	R	0
R	*Exit*	X	16
A	*Left*	L	0
L	*Exit*	X	4

(g) After this sequence of transitions, if we use a learning rate of $\alpha = 0.5$, what would temporal difference learning learn for the value of A? Remember that $V(s)$ is intialized with 0 for all s.

(h) If these transitions repeated many times and learning rates were appropriately small for convergence, what would temporal difference learning converge to for the value of A?

(i) After this sequence of transitions, if we use a learning rate of $\alpha = 0.5$, what would Q-learning learn for the Q-value of $(A, \textit{Right})$? Remember that $Q(s, a)$ is initialized with 0 for all (s, a).

(j) If these transitions repeated many times and learning rates were appropriately small for convergence, what would Q-learning converge to for the Q-value of $(A, \textit{Right})$?

Q3. Wandering Poet

In country B there are N cities. They are all connected by roads in a circular fashion. City 1 is connected with city N and city 2. For $2 \leq i \leq N-1$, city i is conected with cities $i-1$ and $i+1$.

A wandering poet is travelling around the country and staging shows in its different cities.

He can choose to move from a city to a neighboring one by moving East or moving West, or stay in his current location and recite poems to the masses, providing him with a reward of r_i. If he chooses to travel from city i, there is a probability $1 - p_i$ that the roads are closed because of B's dragon infestation problem and he has to stay in his current location. The reward he is to reap is 0 during any successful travel day, and $r_i/2$ when he fails to travel, because he loses only half of the day.

(a) Let $r_i = 1$ and $p_i = 0.5$ for all i and let $\gamma = 0.5$. For $1 \leq i \leq N$ answer the following questions *with real numbers*:

Hint: Recall that $\sum_{j=0}^{\infty} u^j = \frac{1}{1-u}$ for $u \in (0,1)$.

(i) What is the value $V^{stay}(i)$ under the policy that the wandering poet always chooses to stay?

(ii) What is the value $V^{west}(i)$ of the policy where the wandering poet always chooses west?

(b) Let N be even, let $p_i = 1$ for all i, and, for all i, let the reward for cities be given as

$$r_i = \begin{cases} a & i \text{ is even} \\ b & i \text{ is odd,} \end{cases}$$

where a and b are constants and $a > b > 0$.

(i) Suppose we start at an even-numbered city. What is the range of values of the discount factor γ such that the optimal policy is to stay at the current city forever? Your answer may depend on a and b.

(ii) Suppose we start at an odd-numbered city. What is the range of values of the discount factor γ such that the optimal policy is to stay at the current city forever? Your answer may depend on a and b.

(iii) Suppose we start at an odd-numbered city and γ does not lie in the range you computed. Describe the optimal policy.

(c) Let N be even, $r_i \geq 0$, and the optimal value of being in city 1 be positive, i.e., $V^*(1) > 0$. Define $V_k(i)$ to be the value of city i after the kth time-step. Letting $V_0(i) = 0$ for all i, what is the largest k for which $V_k(1)$ could still be 0? Be wary of off-by-one errors.

(d) Let $N = 3$, and $[r_1, r_2, r_3] = [0, 2, 3]$ and $p_1 = p_2 = p_3 = 0.5$, and $\gamma = 0.5$. Compute:

(i) $V^*(3)$

(ii) $V^*(1)$

(iii) $Q^*(1, stay)$