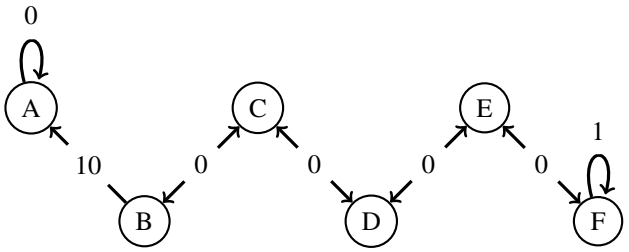
CS 188 Summer 2020 Introduction to Artificial Intelligence

Exam Prep 2 Solutions

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 $\gamma = 1$.

A: Go to A
B: Go to C
C: Go to D
D: Go to E
E: Go to F
F: Go to F

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.

A: Only staying at A is a possible action. For the other states, let n be the number of steps to B, and m be the number of steps to F. Then, the value of going left is $10\gamma^n$ and the value of going right is $\sum_{k=m}^{\infty} \gamma^k = \frac{1}{1-\gamma} - \frac{1-\gamma^m}{1-\gamma}$ because of the geometric series. Now we find the value of γ at which these are equal.

$$\begin{aligned} 10\gamma^n &= \frac{1}{1-\gamma} - \frac{1-\gamma^m}{1-\gamma} = \frac{\gamma^m}{1-\gamma} \\ 10 - 10\gamma &= \gamma^{m-n} \\ \gamma^{m-n} + 10\gamma - 10 &= 0 \end{aligned}$$

The roots of the above polynomial are the points 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).

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

10

Since the discount $\gamma = 1$ and there are no rewards for any action other than exiting, a policy that simply heads to the right exit state and exits will accrue reward 10. This is the optimal policy, since the only alternative reward if 1, and so the optimal value function has value 10.

(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?

2
The first reward is accrued when the agent does the following actions (state transitons) in sequence: Left, Exit. Since two state transitions are necessary before any possible reward, two iterations are necessary for the value function to become non-zero.

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

1
As explained above, the first non-zero value function value will come from exiting out of the left exit cell, which accrues reward 1.

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

4
The value function will equal the optimal value function when it discovers this sequence of state transitions: Right, Right, Right, Exit. This will obviously happen in 4 iterations.

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

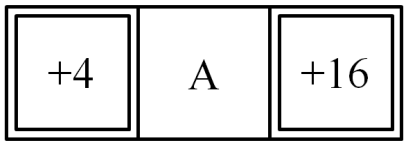
(e) If $\gamma = 0.5$, what is the optimal value $V^*(A)$?

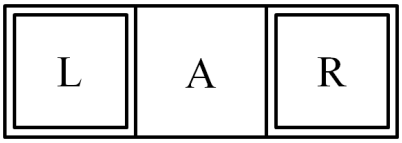
The optimal policy from A is Right, Right, Right, Exit. The rewards accrued by these state transitions are: 0, 0, 0, 10. The discount values are $\gamma^0, \gamma^1, \gamma^2, \gamma^3$, which is 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$. Therefore, $V^*(A) = 0 + 0 + 0 + \frac{10}{8}$.

(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.

The best reward accrued with the policy of going left is $\gamma^1 * 1$. The best reward accrued with the policy of going right is $\gamma^3 * 10$. We therefore have the inequality $10\gamma^3 \geq \gamma$, which simplifies to $\gamma \geq \sqrt{1/10}$. The final answer is $1/\sqrt{10} \leq \gamma \leq 1$

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





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.

3. Remember how temporal difference learning works: upon seeing a s, a, r, s' tuple, we update the value function as $V_{i+1}(s) = (1-\alpha)V_i(s) + \alpha(r + V_i(s'))$. To get the answer, simply write out a table of states, all initially with value 0, and then update it with information in each row of the table above. When all rows have been processed, see what value you ended up with for A.

(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?

10. We are simply updating the value function with the results of following this policy, and that's what we will converge to. For state A, the given tuples show the agent going right as often as it goes left Clearly, if the agent goes left as often as it goes right from A, the value of being in A is only $16/2 + 4/2 = 10$.

(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, Right)? Remember that $Q(s, a)$ is initialized with 0 for all (s, a).

4. The technique is the same as in problem (o), but use the Q-learning update (which includes a max). How do you get the max? Here's an example:

The sample sequence: $(A, \mathit{Right}, R, 0)$.

$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'}(s', a'))$.

$Q(A, right) \leftarrow (1-\alpha)Q(A, right) + \alpha(r + \gamma \max_{a'}(R, a'))$.

But since there is only one exit action from R, then:

$Q(A, right) \leftarrow (1-\alpha)Q(A, right) + \alpha(r + \gamma Q(R, Exit))$.

Note that this MDP is very small – you will finish the game in two moves (assuming you have to move from 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, Right)?

16. Q-learning converges to the optimal Q-value function, if the states are fully explored and the convergence rate is set correctly.

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?

We have that for all i, the Bellman equations for policy evaluation are $V^{stay}(i) = r_i + \gamma V^{stay}(i)$. When $r_i = 1$ and $p_i = 1$ this reduces to $V^{stay}(i) = 1 + 0.5V^{stay}(i)$ which yields $V^{stay}(i) = 2$.

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

$$V^{east}(1) = 0.5(\frac{1}{2} + 0.5V^{east}(1)) + 0.5(\frac{1}{2} + 0.5V^{east}(2)) \tag{1}$$

Since all starting states are equivalent, $V^{east}(1) = V^{east}(2)$. Therefore $V^{east}(1) = V^{east}(2) = \cdots = 1$.

(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.

For all possible values of γ, staying at an even city will be optimal.

(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.

The poet should only move if losing that one extra day fo reward is worth it. So, either he can get the reward of staying for an infinite amount of time at an odd city, which is $b * \frac{1}{1-\gamma}$ or he can move to city a and lose a whole day of as reward, which is $a * \frac{1}{1-\gamma} - a$. He will only stay if the former is greater than the latter, which is only when $\gamma > \frac{b}{a}$

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

The poet should move to an even city and stay there forever.

(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.

Because $V^*(1) > 0$, there must be one $r_i > 0$ for some i. It then follows that $V_1(i) > 0$, $V_2(i-1), V_2(i+1) > 0$ and so on. The worst case is when the diametrically opposite to 1 is the only one having a nonzero r_i. This implies that after $k = M$ steps, $V_{k+1}(1) > 0$ is guaranteed.

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

Notice that $Q^*(1, stay) = \gamma V^*(1)$. Clearly $\pi^*(1) =$ go to 3. $V^*(1) = Q^*(1,$go to $3) = 0.5\gamma V^*(1) + 0.5\gamma V^*(3)$. $V^*(3) Q^*(3, stay) = 3 + \gamma V^*(3)$ Since $\gamma = 0.5$, we have that $V^*(3) = 6$. Therefore $V^*(1) = \frac{4}{3}\frac{1}{4} V^*(3) = 2$. And therefore $Q^*(1, stay) = 1$