CS 188 Spring 2024 | Introduction to Artificial Intelligence | HW10B Sol.

Solutions for HW 10B

Q1. [18 pts] MDPs and RL

The agent is in a 2×4 gridworld as shown in the figure. We start from square 1 and finish in square 8. When square 8 is reached, we receive a reward of $+10$ at the game end. For anything else, we receive a constant reward of -1 (you can think of this as a time penalty).

1	2	3	4
5	6	7	8

The actions in this MDP include: up, down, left and right. The agent cannot take actions that take them off the board. In the table below, we provide initial non-zero estimates of Q values (Q values for invalid actions are left as blanks):

Table 1

	action=up	action=down	action=left	action=right
state=1		Q(1, down)=4		Q(1, right)=3
state=2		Q(2, down)=6	Q(2, left)=4	Q(2, right)=5
state=3		Q(3, down)=8	Q(3, left)=5	Q(3, right)=7
state=4		Q(4, down)=9	Q(4, left)=6	
state=5	Q(5, up)=5			Q(5, right)=6
state=6	Q(6, up)=4		Q(6, left)=5	Q(6, right)=7
state=7	Q(7, up)=6		Q(7, left)=6	Q(7, right)=8

(a) [2 pts] Suppose that you perform actions in this grid and observe the following episode: 3, right, 4, down, 8 (terminal).

With learning rate $\alpha = 0.2$, discount $\gamma = 0.8$, perform an update of $Q(3, right)$ and $Q(4, down)$. Note that here, we update Q values based on the sampled actions as in TD learning, rather than the greedy actions.

$Q(3, right)$: 7*(1-0.2)+0.2*(-1+0.8*9)
$Q(4, down)$: 9*(1-0.2)+0.2*(10)

Please note that the -1 and 10 comes from the first paragraph of the problem statement (which describes the R(s, a, s')

(b) Your friend Adam guesses that the actions in this MDP are fully deterministic (e.g. taking down from 2 will land you in 6 with probability 1 and everywhere else with probability 0). Since we have full knowledge of T and R, we can thus use the Bellman equation to improve (i.e., further update) the initial Q estimates.

Adam tells you to use the following update rule for Q values, where he assumes that your policy is greedy and thus does $\max_a Q(s, a)$. The update rule he prescribes is as follows:

$$Q_{k+1}(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q_k(s',a')]$$

(i) [1 pt] Perform one update of $Q(3, \text{left})$ using the equation above, where $\gamma = 0.8$. You may break ties in any way.

$-1 + 0.8 \times (1 \times 6)$ because we are in a deterministic grid world with a greedy policy.

(ii) [3 pts] For the Q update rule prescribed above, how is it different from the Q learning update that we saw in lecture, which is $Q_{k+1}(s,a) = (1-\alpha)Q_k(s,a) + \alpha * \text{sample}$?

Adam's Q update rule is the update rule for Q-value iteration (Bellman update). The difference between Q-value iteration and Q learning is that Q learning does not require knowing the transition function T.

(c) After observing the agent for a while, Adam realized that his assumption of T being deterministic is wrong in one specific way: **when the agent tries to legally move down, it occasionally ends up moving**

left instead (except from grid 1 where moving left results in out-of-bound). All other movements are still deterministic.

Suppose we have run the Q updates outlined in the equation above until convergence, to get $Q^*_{wrong}(s,a)$ under the original assumption of the wrong (deterministic) T. Suppose $Q^*_{correct}(s,a)$ denotes the Q values under the new correct T. Note that you don't explicitly know the exact probabilities associated with this new T, but you know that it qualitatively differs in the way described above. As prompted below, list the set of (s,a) pairs where $Q^*_{wrong}(s,a)$ is either an over-estimate or under-estimate of $Q^*_{correct}(s,a)$.

(i) [3 pts] List of (s,a) where $Q^*_{wrong}(s,a)$ is an over-estimate. Explain why.

There are two types of (s,a) pairs that are over-estimated.

First, all the (s,a) pairs that have non-zero probability of landing in $s' = 2$, 3, or 4. This is because $V(2)$, $V(3)$, $V(4)$ will all end up being overestimated. So all (s,a) pairs that use 2,3,4 as s' will be overestimation.

Second, all the (s,a) pairs that start from state $s = 2$, 3, 4 with $a =$ down. This is because they have probability of moving left to farther away from 8.

To sum up, the (s,a) pairs are: $(1, right)$, $(2, right)$, $(2, down)$, $(3, left)$, $(3, right)$, $(3, down)$, $(4, left)$, $(4, down)$, $(6, up)$, $(7, up)$.

Side note: the state values for 1, 5, 6, 7, 8 are not affected because the optimal value of those state can be obtained through a sequence of nodes without having to take a "down" action that is affected by this noisy failure

(ii) [3 pts] List of (s,a) where $Q^*_{wrong}(s,a)$ is an under-estimate (and why):

None

(d) [2 pts] While watching the agent, Adam observes 1000 episodes where the agent was in state 3 and selected action Down.

In 99 episodes, the agent landed in state 2, i.e. $(s = 3, a = \text{Down}, s' = 2)$.

In the other 901 episodes, the agent landed in state 7, i.e. $(s = 3, a = \text{Down}, s' = 7)$.

(i) [1 pt] From Adam's samples, what is the estimated probability that the agent moves left instead of down?

0.099. 99/1000 episodes involved the agent moving left instead of down.

(ii) [1 pt] Perform one update of $Q(3, \text{down})$ using Adam's suggested update, where $\gamma = 0.8$.

$Q(3, \text{down}) = -1 + \gamma(0.099 \cdot Q(s = 2, down) + 0.901 \cdot Q(s = 7, right)) = -1 + 0.8 \cdot (0.099 \cdot 6 + 0.901 \cdot 8) = 5.2416$

(e) Instead of using the "ϵ-greedy" algorithm, we will now do some interesting exploration with softmax. We first introduce a new type of policy: A stochastic policy $\pi(a|s)$ represents the probability of action a being prescribed, conditioned on the current state. In other words, the policy is a now a distribution over possible actions, rather than a function that outputs a deterministic action.

Let's define a new policy as follows:

$$\pi(a|s) = \frac{e^{Q(s,a)}}{\sum_{a'} e^{Q(s,a')}}$$

(i) [2 pts] Suppose we are at square 3 in the grid and we want to use the originally provided Q values from the table. What is the probability that this policy will tell us to go right? What is the probability that this policy will tell us to go left? Note that the sum over actions prescribed above refers to a sum over legal actions. You may leave your answer in terms of e.

$\pi(3, right) = \frac{e^7}{e^8+e^5+e^7} = \frac{e^2}{e^3+e^2+1} \approx 0.259$

$\pi(3, left) = \frac{e^5}{e^8+e^5+e^7} = \frac{1}{e^3+e^2+1} \approx 0.035$

(ii) [2 pts] How is this exploration strategy qualitatively different from "ϵ-greedy"?

This exploration is guided by Q value rather than purely random, so you can explore while still taking some amount of goodness (value) into account.