CS 344 (Spring 2017): Class Test 3

Instructor: Shivaram Kalyanakrishnan

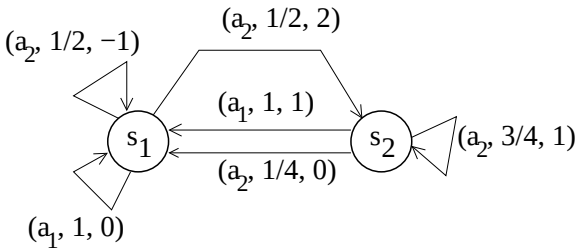
8.30 a.m. – 9.30 a.m., March 21, 2017, 101/103 New CSE Building

Total marks: 15

Note. Provide brief justifications and/or calculations along with each answer to illustrate how you arrived at the answer.

Question 1. Consider an MDP $M = (S, A, R, T, \gamma)$, with a set of states $S = \{s_1, s_2\}$; a set of actions $A = \{a_1, a_2\}$; a transition function T and a reward function R as specified in the table below; and a discount factor $\gamma = \frac{2}{3}$. A state transition diagram corresponding to M is shown alongside the table. In the diagram, each transition is annotated with (action, transition probability, reward). Transitions with zero probabilities are not shown.

Transition probabilities	Rewards
$T(s_1, a_1, s_1) = 1$	$R(s_1, a_1, s_1) = 0$
$T(s_1, a_1, s_2) = 0$	$R(s_1, a_1, s_2) = 0$
$T(s_1, a_2, s_1) = 1/2$	$R(s_1, a_2, s_1) = -1$
$T(s_1, a_2, s_2) = 1/2$	$R(s_1, a_2, s_2) = 2$
$T(s_2, a_1, s_1) = 1$	$R(s_2, a_1, s_1) = 1$
$T(s_2, a_1, s_2) = 0$	$R(s_2, a_1, s_2) = 0$
$T(s_2, a_2, s_1) = 1/4$	$R(s_2, a_2, s_1) = 0$
$T(s_2, a_2, s_2) = 3/4$	$R(s_2, a_2, s_2) = 1$



For $i, j \in \{1, 2\}$, let π^{ij} denote the deterministic policy that takes action i from state s_1 and action j from state s_2.

1a. Calculate $V^{\pi^{21}}(s_1)$ and $V^{\pi^{21}}(s_2)$. [2 marks]

1b. It is a well-known result (a derivative of the Policy Improvement Theorem) that a policy π is optimal if and only if $\forall s \in S, \forall a \in A : Q^{\pi}(s, a) \leq V^{\pi}(s)$. Use this result to ascertain if π^{21} is an optimal policy. [3 marks]

Question 2. An agent interacts with a 3-state, 3-action MDP, in which the discount factor $\gamma = 1/2$. The agent initialises its estimate of the action value function as in the table below.

Q	a_1	a_2	a_3
s_1	5	6	9
s_2	0	-1	2
s_3	6	4	-3

The agent then encounters a trajectory $s_2, a_2, 4, s_2, a_3, 0, s_3, a_2, \ldots$ (4 and 0 are rewards). If the agent was implementing Q-learning with a constant learning rate of $\alpha = 0.1$, what would the action value table be after making the first two learning updates? [2 marks]

Question 3. This question pertains to the use of heuristic functions (or simply *heuristics*) in search.

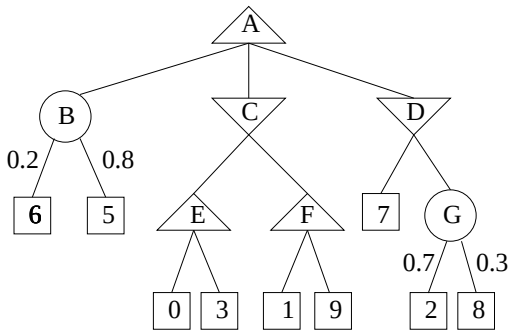
3a. What are the properties of a **consistent** heuristic? [1 mark]

3b. If h_1 is a consistent heuristic and h_2 is a consistent heuristic, are h_3, h_4, and h_5 (defined below) *necessarily* consistent?

- $h_3 = h_1 + h_2$.
- $h_4 = \frac{h_1 + h_2}{3}$.
- $h_5 = \max(h_1, h_2)$.

Support your answer in each case with a short proof. [3 marks]

Question 4. Consider the game tree below, which shows max nodes (triangles pointed upwards), min nodes (triangles pointed downwards), chance nodes (circles), and leaves (rectangles). Probabilities associated with chance events are shown on the corresponding links. The value of each leaf is shown inside it. Since the leaves have distinct values, we also use each leaf's value as its name.



4a. Assuming max and min play optimally, what is the (expectiminimax) value of each internal node in the tree? [2 marks]

4b. What is the sequence of nodes expanded by DFS with Alpha-Beta pruning (also called Alpha-Beta search)? Assume the leftmost unexpanded child is given preference in a tie. [1 mark]

4c. If it is known *a priori* that the values of leaves lie in $[0, 9]$, can the strategy described in 4b be improved such that an optimal action for the root can be determined in even fewer steps? Explain. [1 mark]

Solutions

1a. The set of Bellman's Equations for π^{21} are as follows.

$$V^{\pi^{21}}(s_1) = \frac{1}{2}\left(-1 + \gamma V^{\pi^{21}}(s_1)\right) + \frac{1}{2}\left(2 + \gamma V^{\pi^{21}}(s_2)\right); V^{\pi^{21}}(s_2) = 1 + \gamma V^{\pi^{21}}(s_1).$$

Solving, we get:

$$V^{\pi^{21}}(s_1) = \frac{15}{8}; V^{\pi^{21}}(s_2) = \frac{9}{4}.$$

1b. $Q^{\pi^{21}}(\cdot,\cdot)$ is calculated as follows.

$$Q^{\pi^{21}}(s_1, a_1) = 0 + \gamma V^{\pi^{21}}(s_1) = \frac{5}{4}.$$

$$Q^{\pi^{21}}(s_1, a_2) = V^{\pi^{21}}(s_1) = \frac{15}{8}.$$

$$Q^{\pi^{21}}(s_2, a_1) = V^{\pi^{21}}(s_2) = \frac{9}{4}.$$

$$Q^{\pi^{21}}(s_2, a_2) = \frac{1}{4}\left(0 + \gamma V^{\pi^{21}}(s_1)\right) + \frac{3}{4}\left(1 + \gamma V^{\pi^{21}}(s_2\right) = \frac{35}{16}.$$

Applying the result provided, we conclude that π^{21} is indeed an optimal policy.

2. Let the table given correspond to $Q_0(\cdot,\cdot)$. We get

$$Q_1(s_2, a_2) = Q_0(s_2, a_2)(1-\alpha) + \alpha(4 + \gamma Q_0(s_2, a_3)) = -1 \times 0.9 + 0.1 \times (4 + 0.5 \times 2) = 0.4.$$

If $s \neq s_2$ or $a \neq a_2$, $Q_1(s,a) = Q_0(s,a)$. From the second update, we get

$$Q_2(s_2, a_3) = Q_1(s_2, a_3)(1-\alpha) + \alpha(0 + \gamma Q_1(s_3, a_1)) = 2 \times 0.9 + 0.1 \times (0 + 0.5 \times 6) = 2.1,$$

and again, if $s \neq s_2$ or $a \neq a_3$, $Q_2(s,a) = Q_1(s,a)$. Thus, $Q_2(\cdot,\cdot)$ is as follows.

Q	a_1	a_2	a_3
s_1	5	6	9
s_2	0	-0.4	2.1
s_3	6	4	-3

3a. Let n and n' be *arbitrary* nodes in a search tree, wherein n' is reached by taking action a from n. Let g be an arbitrary node that corresponds to a goal state. We make the standard assumption that costs are non-negative.

A heuristic h is consistent if (1) $h(n) \leq cost(n, a, n') + h(n')$, and (2) $h(g) = 0$. Together these conditions imply (3) $h(n) \leq$ cost-to-goal(n), and in fact (1) and (3) imply (2). Thus, to show consistency, either (1) and (2), or (1) and (3) must be shown to be satisfied. If any of (1), (2), and (3) is violated by a heuristic, then the heuristic is not consistent.

- $h_3 = h_1 + h_2$ need not be consistent. For example, consider $h_1(n) = h_2(n) =$ cost-to-goal(n). Clearly cost-to-goal(n) is a consistent heuristic. Then $h_3(n) = 2 \times$ cost-to-goal(n), which violates (3) if cost-to-goal(n) is positive.

- $h_4 = \frac{h_1+h_2}{3}$ is necessarily consistent, as argued below.
  $h_4(n) = \frac{h_1(n)+h_2(n)}{3} \leq \frac{2}{3} cost(n, a, n') + \frac{h_1(n')+h_2(n')}{3} \leq cost(n, a, n') + h_4(n')$.
  $h_4(n) = \frac{h_1(n)+h_2(n)}{3} \leq \frac{2}{3}$cost-to-goal$(n) \leq$ cost-to-goal(n).

- $h_5 = \max(h_1, h_2)$ is necessarily consistent.
  $h_5(n) = \max(h_1(n), h_2(n)) \leq \max(cost(n, a, n') + h_1(n'), cost(n, a, n') + h_2(n')) = cost(n, a, n') +$ $\max(h_1(n'), h_2(n')) = cost(n, a, n') + h_4(n')$.
  $h_5(n) = \max(h_1(n), h_2(n)) \leq \max($cost-to-goal$(n),$ cost-to-goal$(n)) =$ cost-to-goal(n).

4a.

$$val(B) = 0.2 \times 6 + 0.8 \times 5 = 5.2.$$

$$val(E) = \max(0, 3) = 3.$$

$$val(F) = \max(1, 9) = 9.$$

$$val(C) = \min(val(E), val(F)) = 3.$$

$$val(G) = 0.7 \times 2 + 0.3 \times 8 = 3.8.$$

$$val(D) = \min(7, val(G) = 3.8.$$

$$val(A) = \max(val(B), val(C), val(D)) = 5.2.$$

4b. Observe that $val(B) \geq val(E)$. Since C is a min node, it can infer based on the evaluations of B and E (which precede it in the DFS order) that F need not be evaluated. Whatever F's evaluation, $val(C) = \min(val(E), val(F))$ is guaranteed not to exceed $val(E) = 3$. Therefore, A will not pick the action leading to C. No other nodes can be Alpha-Beta pruned. Thus, the required expansion order is: $AB65CE03D7G28$.

4c. If the values of leaves are bounded in $[0, 9]$, so must the values in each node, which are obtained using expectation, min, and max operations. Hence, after 2 is evaluated, we can conclude that $val(G)$ is upper-bounded by $0.7 \times 2 + 0.3 \times 9 = 4.1$. If D, which is a min player, plays optimally, it will therefore take the action to reach G. Anticipating that D will do so, A, a max node, will not take the action leading to G, but rather proceed to B, which assures it 5.2 in expectation. Hence, we can eliminate the expansion of 8 from what we performed in 4b; the overall expansion sequence is $AB65CE03D7G2$.