CS 344 (Spring 2018): Mid-semester Examination

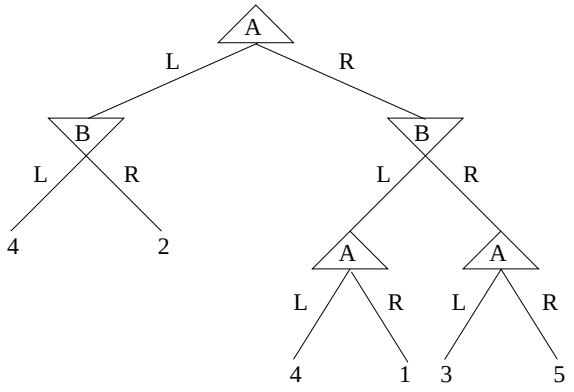
Instructor: Shivaram Kalyanakrishnan

11.00 a.m. – 1.00 p.m., March 1, 2018, 101/103, New CSE Building

Total marks: 20

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

Question 1. The following question pertains to a turn-taking zero sum game played by a "max" agent A and a "min" agent B. The game proceeds according to the tree shown below. The actions available at each node are "left" (L) and "right" (R). Each leaf shows A's utility (which is the negative of B's utility).



Suppose that unfortunately, A has lost the ability to sense its state, and nor does it have any memory (to remember previous actions). In other words, A is constrained to apply the same rule for action-selection from every node in the tree. Your task is to show that A can do strictly better by adopting a *randomised* strategy. Specifically, let A play L with probability p and R with probability $1 - p$ from each of its nodes. Let $minReward(p)$ denote the least expected reward that A will get by playing this way. Note that B is not handicapped like A: it can follow different strategies at each of its nodes. In fact, A obtains $minReward(p)$ when it plays L with probability p and B plays adversarially with respect to A's strategy.

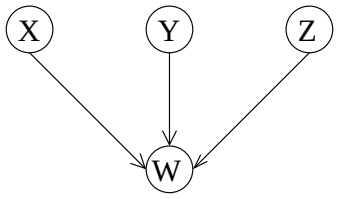
Answer the following questions. In your solutions, name your nodes by the path leading to them from the root. For example, the root node is $\emptyset$, and the right-most leaf is R-R-R.

1a. What are $minReward(0)$ and $minReward(1)$? [1 mark]

1b. What are $p^\star = \text{argmax}_{p \in [0,1]} minReward(p)$ and $minReward(p^\star)$? What are B's actions to restrict A when it plays L with probability $p^\star$? [4 marks]

Question 2. Write down pseudocode for an uninformed search algorithm that is *complete*, as well as *memory-efficient*. Completeness implies that the algorithm is guaranteed to find a path to a goal state should a finite-length path indeed exist. The state space can itself be infinite, though. By memory-efficiency we mean that the algorithm only needs access to a data structure whose size is *linear* in the (unknown) shortest distance to the goal. [3 marks]

Question 3. This question is about the Bayes Net shown below. Answer the two parts of the question from first principles: that is, using only the basic semantics of Bayes Nets and the rules of probability. In particular, do not directly use the rules for D-separation.

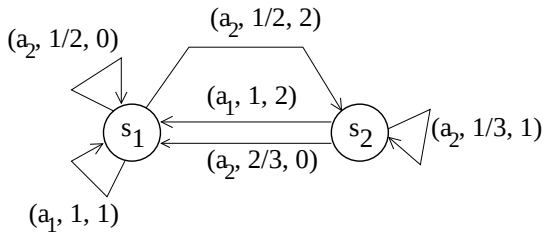


3a. Is it guaranteed that $X \perp Y$? Prove that your answer is correct. [2 marks]

3b. Is it guaranteed that $(X \perp Y)|Z$? Prove that your answer is correct. [2 marks]

Question 4. Consider an MDP $M = (S, A, T, R, \gamma)$, 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 and figure below. The discount factor $\gamma = \frac{1}{2}$. In the figure, each transition is annotated with (action, transition probability, reward). Transitions with zero probabilities are not shown.

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



Let π^{12} be the deterministic policy that takes action a_1 from state s_1 and action a_2 from state s_2; let π^{22} be the deterministic policy that takes action a_2 from state s_1 and action a_2 from state s_2.

4a. Consider an agent A that starts in state s_1. The agent has decided to take action a_2 the very first time it acts (at $t = 0$), and thereafter (for $t \geq 1$), to follow policy π^{12}. What is the expected long-term discounted reward that the agent will accrue? In other words, if $s^0, a^0, r^0, s^1, a^1, r^1, s^2, a^2, r^2, \ldots$ is A's trajectory over time, what is

$$\mathbb{E}[r^0 + \gamma r^1 + \gamma^2 r^2 + \ldots | s^0 = s_1; a^0 = a_2; \text{ for } t \geq 1, a^t = \pi^{12}(s^t)]?$$ [3 marks]

4b. Consider a second agent B that is known to follow π^{22}. Unfortunately, it is not reliably known to the observer in which state (s_1 or s_2) the agent currently resides. The only clues available to the observer are messages that B broadcasts at each step, before it takes an action. In these messages, B announces its state (as "I am in s_1" or "I am in s_2"). However, these messages are not guaranteed to be correct. Precisely, they will be correct with probability $p \in [0.5, 1]$ and incorrect with probability $1 - p$. For example, one possible sequence of events is as follows.

- $t = 0$: B starts in state s_2, broadcasts "I am in state s_2.", and takes action a_2.
- $t = 1$: B is in state s_1, broadcasts "I am in state s_2.", and takes action a_2.

The first message is correct, while the second one is incorrect.

Draw a Dynamic Bayes Net with nodes $State^t$ and $Message^t$ for $t \geq 0$, where these random variables correspond to B's true state at t and message broadcast at t, respectively. You should show both the topology and the conditional probability tables (which may depend on p). [1 mark]

4c. Continuing from 4b, suppose the observer believes that to begin, Agent B is equally likely to be in state s_1 or state s_2. Also suppose that the observer receives messages exactly as in the trace given in 4b: that is, two successive broadcasts of "I am in state s_2". If the observer performs Bayesian inference, what is its belief about B's state at $t = 1$, after it has heard the first two messages? In other words, what is $\mathbb{P}\{State^1 | Message^0 = \text{"I am in state } s_2\text{"}, Message^1 = \text{"I am in state } s_2\text{"}\}$? [4 marks]

Solutions

1a. If A always plays L, B will play R from node L to restrict A's reward to 2. If A always plays R, B will play L from node R to restrict A's reward to 1. Thus: $minReward(0) = 1$ and $minReward(1) = 2$. In the next part, we provide the answer for general $p \in [0, 1]$.

1b. From node L, B will play R to keep A's reward down to 2. Assume that A plays L with probability p. Then A's expected reward from R-L is $4p + 1 - p = 3p + 1$, and A's expected reward from R-R is $3p + 5(1 - p) = 5 - 2p$. For $p \in [0, 0.8]$, $3p + 1 \leq 5 - 2p$, and for $p \in [0.8, 1]$, $3p + 1 \geq 5 - 2p$. Hence, if $p < 0.8$, B will play L from node R; for $p > 0.8$, B will play R from node R. If $p = 0.8$, either choice yields the same expected reward for A, of 3.4.

A's expected reward from playing L with probability p, assuming B plays adversarially for the corresponding value of p is computed for two cases. When $p \leq 0.8$, A's expected reward is

$$p(2) + (1 - p)(3p + 1) = 1 + 4p - 3p^2,$$

and when $p \geq 0.8$, A's expected reward is

$$p(2) + (1 - p)(5 - 2p) = 5 - 5p + 2p^2.$$

We verify that we obtain the answers for 1a by setting $p = 0$ and $p = 1$.

The maximum A can assure itself with $p \leq 0.8$ is achieved at $p = \frac{2}{3}$, and the maximum it can achieve with $p \geq 0.8$ is by taking $p = 0.8$. The first choice yields a reward of $1 + 4(\frac{2}{3}) - 3(\frac{2}{3})^2 = \frac{7}{3}$, while the second choice yields $5 - 5(0.8) + 2(0.8)^2 = 2.28$.

Thus, $p^\star = \frac{2}{3}$ and $minReward(p^\star) = \frac{7}{3}$. When A plays L with probability $p^\star$, B's adversarial strategy is to play R from node L, and L from node R.

2. For achieving memory-efficiency, the natural choice is to use Depth First Search (DFS). However, DFS is not complete. To ensure completeness, we can perform *iterative deepening*: that is, to perform DFS within an outer loop that increments the depth d. Within each loop, DFS expands paths up to a maximum depth of d. Hence, all 1-length paths get covered, then all 2-length paths, and so on until at some depth $d^\star$, a goal state is encountered and the search terminates. The memory requirement is linear in d. Below is pseudocode for a possible implementation.

```
DFSWithIterativeDeepening(startState)

For depth = 0, 1, 2, ...
    x ← DFS(depth, startState, "").
    If x ≠ ∅
        Return x.
```

```
DFS(depth, state, pathToState)

If isGoal(state)
    Return pathToState.
If depth = 0
    Return ∅.
For a ∈ Actions(state)
    x ← DFS(depth − 1, nextState(state, a), append(pathToState, a)).
    If x ≠ ∅
        Return x.
Return ∅.
```

3a. Yes: $X \perp Y$ since

$$\begin{aligned}
\mathbb{P}\{X, Y\} &= \sum_{z \in Z} \sum_{w \in W} \mathbb{P}\{X, Y, z, w\} \\
&= \sum_{z \in Z} \sum_{w \in W} \mathbb{P}\{X\}\mathbb{P}\{Y\}\mathbb{P}\{z\}\mathbb{P}\{w|X, Y, z\} \\
&= \mathbb{P}\{X\}\mathbb{P}\{Y\} \sum_{z \in Z} \mathbb{P}\{z\} \sum_{w \in W} \mathbb{P}\{w|X, Y, z\} \\
&= \mathbb{P}\{X\}\mathbb{P}\{Y\} \sum_{z \in Z} \mathbb{P}\{z\} \\
&= \mathbb{P}\{X\}\mathbb{P}\{Y\}.
\end{aligned}$$

3b. Yes: $(X \perp Y)|Z$. First, observe that

$$\begin{aligned}
\mathbb{P}\{X, Y|Z\} &= \frac{\mathbb{P}\{X, Y, Z\}}{\mathbb{P}\{Z\}} \\
&= \frac{\sum_{w \in W} \mathbb{P}\{X, Y, Z, w\}}{\mathbb{P}\{Z\}} \\
&= \frac{\sum_{w \in W} \mathbb{P}\{X\}\mathbb{P}\{Y\}\mathbb{P}\{Z\}\mathbb{P}\{w|X, Y, Z\}}{\mathbb{P}\{Z\}} \\
&= \mathbb{P}\{X\}\mathbb{P}\{Y\} \sum_{w \in W} \mathbb{P}\{w|X, Y, Z\} \\
&= \mathbb{P}\{X\}\mathbb{P}\{Y\}.
\end{aligned}$$

Now,

$$\mathbb{P}\{X|Z\} = \frac{\mathbb{P}\{X,Z\}}{\mathbb{P}\{Z\}} = \sum_{y\in Y}\sum_{w\in W}\frac{\mathbb{P}\{X,y,Z,w\}}{\mathbb{P}\{Z\}} = \sum_{y\in Y}\sum_{w\in W}\frac{\mathbb{P}\{X\}\mathbb{P}\{y\}\mathbb{P}\{Z\}\mathbb{P}\{w|X,y,Z\}}{\mathbb{P}\{Z\}} = \mathbb{P}\{X\}$$

and

$$\mathbb{P}\{Y|Z\} = \frac{\mathbb{P}\{Y,Z\}}{\mathbb{P}\{Z\}} = \sum_{x\in X}\sum_{w\in W}\frac{\mathbb{P}\{x,Y,Z,w\}}{\mathbb{P}\{Z\}} = \sum_{x\in X}\sum_{w\in W}\frac{\mathbb{P}\{x\}\mathbb{P}\{Y\}\mathbb{P}\{Z\}\mathbb{P}\{w|x,Y,Z\}}{\mathbb{P}\{Z\}} = \mathbb{P}\{Y\}.$$

Hence, we have shown $\mathbb{P}\{X,Y|Z\} = \mathbb{P}\{X|Z\}\mathbb{P}\{Y|Z\}$, which completes the proof.

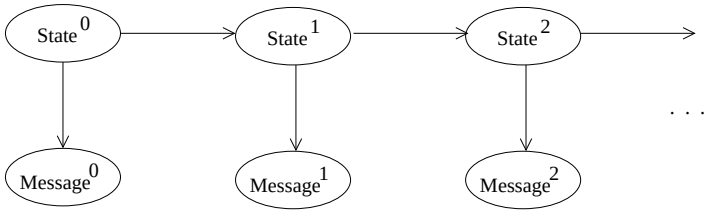
4a. Bellman's Equations for π^{12} are

$$V^{\pi^{12}}(s_1) = 1 + \gamma V^{\pi^{12}}(s_1) \text{ and } V^{\pi^{12}}(s_2) = \frac{2}{3}\{0 + \gamma V^{\pi^{12}}(s_1\} + \frac{1}{3}\{1 + \gamma V^{\pi^{12}}(s_2\}),$$

which yields $V^{\pi^{12}}(s_1) = 2$ and $V^{\pi^{12}}(s_2) = \frac{6}{5}$. Since the agent takes action a_2 from state s_1, and thereafter follows π^{12}, its expected long-term discounted reward is

$$\frac{1}{2}\{0 + \gamma V^{\pi^{12}}(s_1)\} + \frac{1}{2}\{2 + \gamma V^{\pi^{12}}(s_2)\} = \frac{1}{2}(1) + \frac{1}{2}(\frac{13}{5}) = 1.8.$$

4b. The figure below shows the Bayes Net modeling the interaction between agent and observer.



$\mathbb{P}\{State^0 = s_1\}$ and $\mathbb{P}\{State^0 = s_2\}$ constitute the observer's prior belief. These can be arbitrary; in 4c they are assumed to each be $\frac{1}{2}$.

Since the agent follows policy π^{22}, we obtain that for $t \geq 0$, $\mathbb{P}\{State^{t+1}|State^t\}$ is as follows.

$State^t$	$State^{t+1}$	$\mathbb{P}\{State^{t+1}\|State^t\}$
s_1	s_1	$1/2$
s_1	s_2	$1/2$
s_2	s_1	$2/3$
s_2	s_2	$1/3$

From the description of the problem, $\mathbb{P}\{Message^t|State^t\}$ is as follows for $t \geq 0$.

$State^t$	$Message^t$	$\mathbb{P}\{Message^t\|State^t\}$
s_1	s_1	p
s_1	s_2	$1-p$
s_2	s_1	$1-p$
s_2	s_2	p

4c.

$$
\begin{aligned}
&\mathbb{P}\{State^1 | Message^0, Message^1\} \\
&\propto \mathbb{P}\{State^1, Message^0, Message^1\} \\
&= \sum_{s \in State^0} \mathbb{P}\{State^0 = s, State^1, Message^0, Message^1\} \\
&= \sum_{s \in State^0} \mathbb{P}\{State^0 = s\}\mathbb{P}\{State^1 | State^0 = s\}\mathbb{P}\{Message^0 | (State^0 = s)\}\mathbb{P}\{Message^1 | State^1\}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\{State^1 = s_1 | Message^0 = \text{“I am in } s_2\text{”}, Message^1 = \text{“I am in } s_2\text{”}\} \\
&\propto \frac{1}{2} \cdot \frac{1}{2} \cdot (1-p) \cdot (1-p) + \frac{1}{2} \cdot \frac{2}{3} \cdot p \cdot (1-p) \\
&= \frac{3(1-p)^2 + 4p(1-p)}{12}.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\{State^1 = s_2 | Message^0 = \text{“I am in } s_2\text{”}, Message^1 = \text{“I am in } s_2\text{”}\} \\
&\propto \frac{1}{2} \cdot \frac{1}{2} \cdot (1-p) \cdot p + \frac{1}{2} \cdot \frac{1}{3} \cdot p \cdot p \\
&= \frac{3p(1-p) + 2p^2}{12}.
\end{aligned}
$$

Normalising, we get

$$\mathbb{P}\{State^1 = s_1 | Message^0 = \text{“I am in } s_2\text{”}, Message^1 = \text{“I am in } s_2\text{”}\} = \frac{3 - 2p - p^2}{3 + p - 2p^2}$$

and

$$\mathbb{P}\{State^1 = s_2 | Message^0 = \text{“I am in } s_2\text{”}, Message^1 = \text{“I am in } s_2\text{”}\} = \frac{3p - p^2}{3 + p - 2p^2}.$$