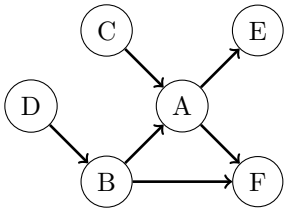
ASEN 5264 Decision Making under Uncertainty
Quiz 1: Probabilistic Models and MDPs

Clearly indicate your final answers and briefly justify numerical answers with text or mathematical expressions.
If you do not understand how to do a problem, skip it and move on so that you have time to attempt all problems.
You may consult any static source, but you may NOT communicate with any person except the instructor or TA,
and you may not use LLMs such as ChatGPT.

Question 1. (30 pts) Consider the following Bayesian network made up of binary random variables: 13:00



a) Is it possible to conclude from the structure only that $E \perp C \mid A$?

b) Is it possible to conclude from the structure only that $F \perp C \mid A$?

c) Suppose $P(E = 1 \mid A = 1) = 0.8$. Find $P(E = 0 \mid A = 1, C = 1)$ if possible with the given information.

d) Suppose $P(E = 1 \mid A = 1) = 0.8$. Find $P(E = 1 \mid A = 0, C = 1)$ if possible with the given information.

e) Suppose $P(F = 1 \mid A = 1) = 0.7$. Find $P(F = 0 \mid A = 1, C = 1)$ if possible with the given information.

a) The only non-cyclic path is

$C \rightarrow A \rightarrow E$

A is in the evidence, so by rule 1, path is d-separated

All paths are d-separated, so we can conclude $E \perp C \mid A$

b) paths:

$C \rightarrow A \leftarrow B \rightarrow F$

$C \rightarrow A \rightarrow F$

$C \rightarrow A \leftarrow B$ is a v-structure, but A is in the evidence

$A \leftarrow B \rightarrow F$ is a fork, but B is not in the evidence.

This path is not d-separated

It is **not possible** to conclude that $F \perp C \mid A$

c) Since we know that $E \perp C \mid A$

$P(E \mid A) = P(E \mid A, C)$

$P(E=0 \mid A=1, C=1) = P(E=0 \mid A=1) = 1 - P(E=1 \mid A=1)$

$= 0.2$

d) Since we do not have $P(E \mid A=0)$ it is **not possible** to compute this.

e) Unlike above, since we do not know $F \perp C \mid A$, and we don't know $P(F \mid A, C)$, it is **not possible** to compute this.

Question 2. (30 pts) Consider the following one player game. During the game, you will flip two fair coins, one after the other. Before the first flip, your friend guesses the outcome, heads or tails. If your friend guesses correctly on the first flip, you pay her \$2. Before the second flip, your friend again guesses heads or tails, but this time you pay her \$1 for each coin that has landed that way. For example if she guesses heads on the second step and both coins were heads, you again pay her \$2.

Formulate this as a Markov decision process from your friend's perspective by writing down the state space, action space, transition probabilities, reward function, and discount factor. Recall that the reward function can be expressed as $R(s, a, s')$.

$S = \{\emptyset, H, T, HH, TT, HT\}$

$A = \{H, T\}$

$\gamma = 1$

$$R(s,a,s') = \begin{cases} 0 & \text{if } s \text{ is terminal} \\ +2 & \text{if } a = s' \\ +1 & \text{if } s' = HT \\ +2 & \text{if } a = H \text{ and } s' = HH \\ +2 & \text{if } a = T \text{ and } s' = TT \\ 0 & \text{o.w.} \end{cases}$$

$$T^H = T^T = \begin{array}{c} \begin{matrix} \emptyset & H & T & HH & TT & HT \end{matrix} \\ \left[\begin{array}{cccccc} 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{matrix} \emptyset \\ H \\ T \\ HH \\ TT \\ HT \end{matrix} \end{array} \quad \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} \text{terminal}$$

Question 3. (30 pts) Consider the following MDP: $S = A = \{1, 2\}$, $R(s, a) = \begin{cases} 2 & \text{if } s = a \\ -1 & \text{otherwise} \end{cases}$, $\gamma = 0.9$. 7:45

If the action is odd, there is a 90% chance of transitioning to the odd state. If the action is even, there is an 80% chance of transitioning to the even state. Suppose that you are performing policy iteration and have computed the value function for the current policy π as $U^\pi = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$.

a) Perform the policy update/improvement step of policy iteration to find a new policy π'.

b) Give a matrix equation for the value of the new policy π'. Fill in numbers in all vectors and matrices except $U^{\pi'}$, but you do not need to solve the equation.

a)

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \left(R(s,a) + \gamma \sum_{s'} T(s'|s,a)\, U^\pi(s') \right)$$

$$\pi'(1) = \operatorname{argmax}\Big(\underset{a=1}{R(1,1)} + \gamma(0.9\, U^\pi(1) + 0.1\, U^\pi(2)),\ \underset{a=2}{R(1,2)} + \gamma(0.8\, U^\pi(2) + 0.2\, U^\pi(1)) \Big)$$

$$2 + 0.9(0.9 \cdot 1 + 0.1 \cdot 5) \qquad -1 + 0.9(0.8 \cdot 5 + 0.2 \cdot 1)$$

$$3.26 \qquad 2.78$$

$\boxed{\pi'(1) = 1}$

$$\pi'(2) = \operatorname{argmax}\Big(\underset{a=1}{R(2,1)} + \gamma(0.9\, U^\pi(1) + 0.1\, U^\pi(2)),\ \underset{a=2}{R(2,2)} + \gamma(0.8\, U^\pi(2) + 0.2\, U^\pi(1)) \Big)$$

$$0.26 \qquad 5.78$$

$\boxed{\pi'(2) = 2}$

b)

$$U^{\pi'} = (I - \gamma T^{\pi'})^{-1} R^{\pi'}$$

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \right)^{-1} \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

5:15

Question 4. (10 pts) Suppose that a 21 state MDP has state space consisting of consecutive integers, $S = \{-10, -9, \ldots, 9, 10\}$, action space $\{-1, 1\}$, reward function $R(s, a) = s^2$, and discount $\gamma = 0.96$. I claim to have an optimal policy π that achieves an expected cumulative discounted reward of 2500 if the initial state is 1, that is $U^\pi(1) = 2500$. Is it possible to prove or disprove my claim with the given information? Explain.

$U^\pi(1) = 2500$ is impossible because the maximum reward that any policy could achieve starting from state 1 is

$$\max_a R(1,a) + \sum_{t=1}^{\infty} \gamma^t \bar{R} \quad \text{where} \quad \bar{R} = \max_{s,a} R(s,a) = 10^2$$

$$= 1 + \gamma \sum_{t=0}^{\infty} \gamma^t \bar{R} = 1 + \gamma \frac{\bar{R}}{1-\gamma} = 2401$$