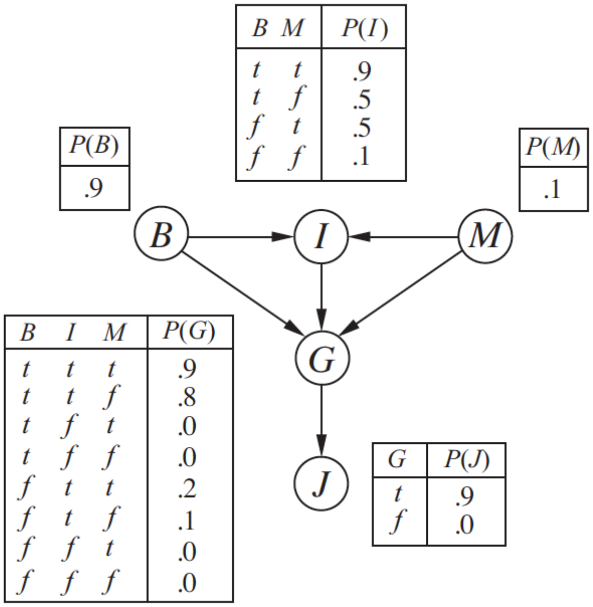
AE8803TSI **Homework #3** **Fall 2019**

Due: November 19 , 2019

1. Consider the Bayesian network shown in the figure below, with Boolean variables: B = BrokeElectionLaw, I = Indicted, M = PoliticallyMotivatedProsecutor, G = FoundGuilty, J = Jailed.



$P(B)$
.9

B	M	$P(I)$
t	t	.9
t	f	.5
f	t	.5
f	f	.1

$P(M)$
.1

B	I	M	$P(G)$
t	t	t	.9
t	t	f	.8
t	f	t	.0
t	f	f	.0
f	t	t	.2
f	t	f	.1
f	f	t	.0
f	f	f	.0

G	$P(J)$
t	.9
f	.0

(a) Which of the following are asserted by the network structure?

i) $P(B,I,M) = P(B)P(I)P(M)$
ii) $P(J|G) = P(J|G,I)$
iii) $P(M|G,B,I) = P(M|G,B,I,J)$

(b) Calculate the value of P(B = true, I = true, M = false, G = true, J = true).

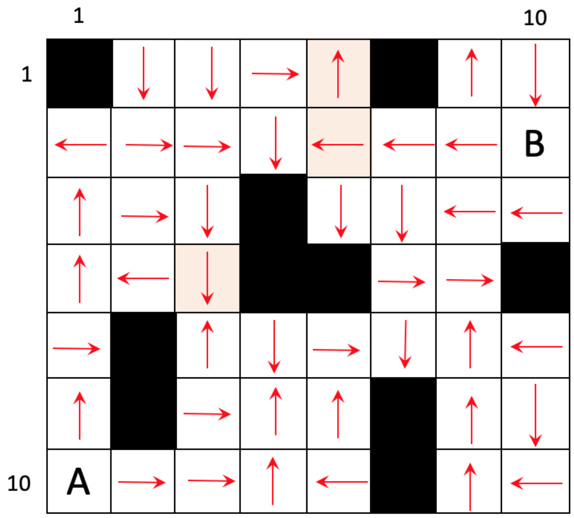
(c) Calculate the probability that someone goes to jail given that they broke the law, have been indicted, and face a politically motivated prosecutor.

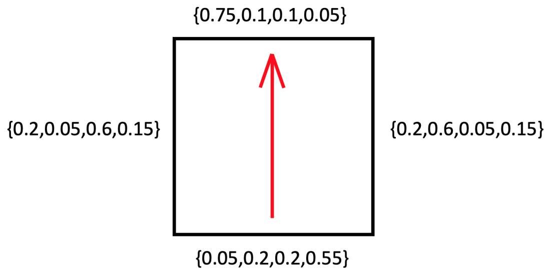
(d) Suppose we want to add the variable P = PresidentialPardon to the network; draw the new network and briefly explain the links you added.

2. Consider the 10×10 environment shown in the figure below. The agent is initially at cell A = (10,1) and the goal location is at cell B = (2,10). The action set of the agent is $\mathcal{A} = \{L,R,U,D\}$. Let $\mathcal{O}$ denote the cells with obstacles. The reward for each state is follows

$$r(s) = \begin{cases} +10, & s = s_{\text{goal}}, \\ -1, & s \neq s_{\text{goal}}, \\ -10, & s \in \mathcal{O}. \end{cases}$$

In each cell there is a disturbance (e.g, wind) acting in the direction shown by the corresponding arrows. This disturbance causes the agent to move to an adjacent cell according to the direction of the disturbance and the action chosen by the agent as shown in the figure, where the first entry denotes the probability to move to the next cell in the direction of the arrow, the last entry denotes the probability to move to the next cell that is opposite to the direction of the arrow, and the second and third entries denote the probabilities to move to the cell that is on the right or the left of the arrow direction, respectively, according to the action set $\mathcal{A} = \{U, R, L, D\}$, in this order.





(a) Implement the value iteration for this world using a discount factor of $\gamma = 0.95$.

(b) Repeat the previous problem when the reward for the cells at locations (4,3), (1,5) and (2,5) have the following values a) $r(s) = 0$; b) $r(s) = 100$; c) $r(s) = -3$. Show the policy for each case. Explain intuitively why these values lead to each resulting policy.

(c) Run 50 instances of your computed policy and record the number of times the agent managed to reach the goal state, along with the accumulated reward for each instance. Compare the results and briefly explain if what you observe is what you expected.

3. Repeat the previous problem using policy iteration. Compare the two approaches and comment on the pros and cons for each method.

For your assistance, the pseudo-codes for both VI and PI are given on the next page.

```
function VALUE-ITERATION($mdp, \epsilon$) returns a utility function
  inputs: $mdp$, an MDP with states $S$, actions $A(s)$, transition model $P(s' \mid s, a)$,
              rewards $R(s)$, discount $\gamma$
          $\epsilon$, the maximum error allowed in the utility of any state
  local variables: $U$, $U'$, vectors of utilities for states in $S$, initially zero
                   $\delta$, the maximum change in the utility of any state in an iteration

  repeat
      $U \leftarrow U'; \delta \leftarrow 0$
      for each state $s$ in $S$ do
          $U'[s] \leftarrow R(s) \ + \ \gamma \ \max_{a \in A(s)} \sum_{s'} P(s' \mid s, a) \ U[s']$
          if $|U'[s] \ - \ U[s]| \ > \ \delta$ then $\delta \leftarrow |U'[s] \ - \ U[s]|$
  until $\delta \ < \ \epsilon(1-\gamma)/\gamma$
  return $U$
```

```
function POLICY-ITERATION($mdp$) returns a policy
  inputs: $mdp$, an MDP with states $S$, actions $A(s)$, transition model $P(s' \mid s, a)$
  local variables: $U$, a vector of utilities for states in $S$, initially zero
                   $\pi$, a policy vector indexed by state, initially random

  repeat
      $U \leftarrow$ POLICY-EVALUATION($\pi$, $U, mdp$)
      $unchanged?$ $\leftarrow$ true
      for each state $s$ in $S$ do
          if $\max_{a \in A(s)} \sum_{s'} P(s' \mid s, a) \ U[s'] \ > \ \sum_{s'} P(s' \mid s, \pi[s]) \ U[s']$ then do
              $\pi[s] \leftarrow \operatorname{argmax}_{a \in A(s)} \sum_{s'} P(s' \mid s, a) \ U[s']$
              $unchanged?$ $\leftarrow$ false
  until $unchanged?$
  return $\pi$
```