CS 188 Fall 2013 Introduction to Artificial Intelligence Midterm 2

- You have approximately 2 hours and 50 minutes.
- The exam is closed book, closed notes except your one-page crib sheet.
- Please use non-programmable calculators only.
- Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a *brief* explanation. All short answer sections can be successfully answered in a few sentences AT MOST.

First name	
Last name	
SID	
edX username	

First and last name of student to your left	
First and last name of student to your right	

For staff use only:

Q1.	Probability: Chain Rule Madness	/8
Q2.	Bayes' Nets: Representation	/8
Q3.	Bayes' Nets: Independence	/9
Q4.	Machine Learning: Maximum Likelihood	/5
Q5.	Independence in Hidden Markov Models	/12
Q6.	Perceptron	/7
Q7.	Variable Elimination	/9
Q8.	Reduced Bayes' Nets	/6
Q9.	Finding WALL-E	/7
Q10.	Decision Networks	/14
Q11.	Argg! Sampling for the Legendary Treasure	/15
	Total	/100

THIS PAGE IS INTENTIONALLY LEFT BLANK

Q1. [8 pts] Probability: Chain Rule Madness

Note that the set of choices is the same for all of the following questions. You may find it beneficial to read all the questions first, and to make a table on scratch paper.

(a) [2 pts] Fill in the circles of **all** expressions that are equivalent to $\mathbf{P(A, B \mid C)}$, **given no independence assumptions**:

○ $\frac{P(C|A)\ P(A|B)\ P(B)}{P(C)}$

○ $\frac{P(A|C)\ P(C|B)\ P(B)}{P(B,C)}$

○ $\frac{P(C|A,B)\ P(B|A)\ P(A)}{P(B|C)\ P(C)}$

○ $\frac{P(B,C|A)\ P(A)}{P(B,C)}$

○ $\frac{P(A|C)\ P(B,C)}{P(C)}$

● $P(A \mid B, C)\ P(B \mid C)$

○ None of the above.

(b) [2 pts] Fill in the circles of **all** expressions that are equivalent to $\mathbf{P(A, B \mid C)}$, **given that** $\mathbf{A \perp\!\!\!\perp B \mid C}$:

○ $\frac{P(C|A)\ P(A|B)\ P(B)}{P(C)}$

○ $\frac{P(A|C)\ P(C|B)\ P(B)}{P(B,C)}$

○ $\frac{P(C|A,B)\ P(B|A)\ P(A)}{P(B|C)\ P(C)}$

○ $\frac{P(B,C|A)\ P(A)}{P(B,C)}$

● $\frac{P(A|C)\ P(B,C)}{P(C)}$

● $P(A \mid B, C)\ P(B \mid C)$

○ None of the above.

(c) [2 pts] Fill in the circles of **all** expressions that are equivalent to $\mathbf{P(A \mid B, C)}$, **given no independence assumptions**:

○ $\frac{P(C|A)\ P(A|B)\ P(B)}{P(C)}$

○ $\frac{P(A|C)\ P(C|B)\ P(B)}{P(B,C)}$

● $\frac{P(C|A,B)\ P(B|A)\ P(A)}{P(B|C)\ P(C)}$

● $\frac{P(B,C|A)\ P(A)}{P(B,C)}$

○ $\frac{P(A|C)\ P(B,C)}{P(C)}$

○ $P(A \mid B, C)\ P(B \mid C)$

○ None of the above.

(d) [2 pts] Fill in the circles of **all** expressions that are equivalent to $\mathbf{P(A \mid B, C)}$, **given that** $\mathbf{A \perp\!\!\!\perp B \mid C}$:

○ $\frac{P(C|A)\ P(A|B)\ P(B)}{P(C)}$

● $\frac{P(A|C)\ P(C|B)\ P(B)}{P(B,C)}$

● $\frac{P(C|A,B)\ P(B|A)\ P(A)}{P(B|C)\ P(C)}$

● $\frac{P(B,C|A)\ P(A)}{P(B,C)}$

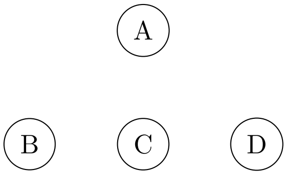
○ $\frac{P(A|C)\ P(B,C)}{P(C)}$

○ $P(A \mid B, C)\ P(B \mid C)$

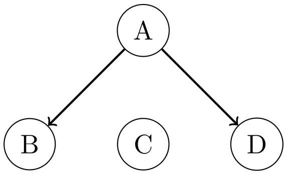
○ None of the above.

To approach this question, try to get from each expression given as an option to the expression given as the query (e.g. $\mathbf{P(A,\ B \mid C)}$) via application of the rules of probability. Think about what additional rules become true when a conditional independence is asserted.

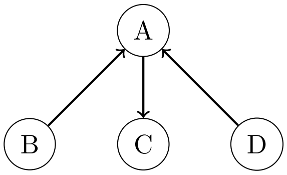
Q2. [8 pts] Bayes' Nets: Representation



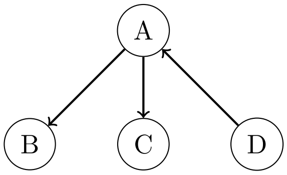
$\mathbf{G_1}$



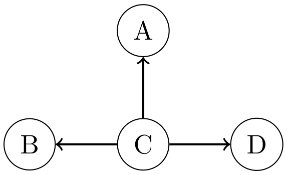
$\mathbf{G_2}$



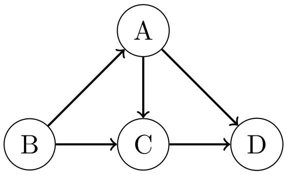
$\mathbf{G_3}$



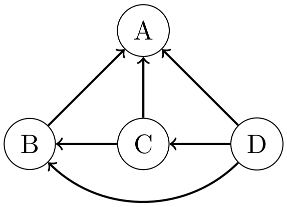
$\mathbf{G_4}$



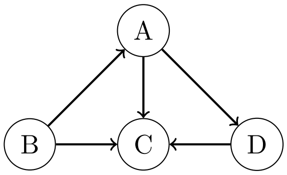
$\mathbf{G_5}$



$\mathbf{G_6}$

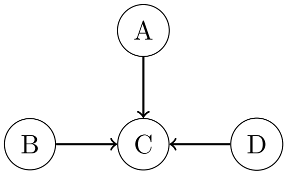


$\mathbf{G_7}$



$\mathbf{G_8}$

(a) Consider the Bayes' Net $\mathbf{B_1}$ below, and fill in **all the circles** (or select *None of the above*) corresponding to the Bayes' Nets $\mathbf{G_1}$ through $\mathbf{G_8}$ that...



$\mathbf{B_1}$

(i) [2 pts] ...are able to represent **at least one distribution** that $\mathbf{B_1}$ is able to represent.

● $\mathbf{G_1}$ ● $\mathbf{G_2}$ ● $\mathbf{G_3}$ ● $\mathbf{G_4}$

● $\mathbf{G_5}$ ● $\mathbf{G_6}$ ● $\mathbf{G_7}$ ● $\mathbf{G_8}$

○ None of the above.

Consider the fully independent joint $P(A, B, C, D) = P(A)P(B)P(C)P(D)$ with uniform distribution (that is, every table entry has probability $\frac{1}{16}$. This can be represented by any Bayes' net. Pick any conditional independence assumption and verify that it is satisfied with by this distribution.

(ii) [2 pts] ...are able to represent **all distributions** that $\mathbf{B_1}$ is able to represent.

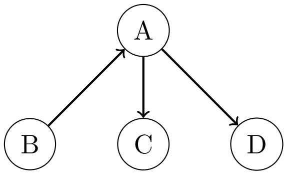
○ $\mathbf{G_1}$ ○ $\mathbf{G_2}$ ○ $\mathbf{G_3}$ ○ $\mathbf{G_4}$

○ $\mathbf{G_5}$ ○ $\mathbf{G_6}$ ● $\mathbf{G_7}$ ● $\mathbf{G_8}$

○ None of the above.

To represent all of the distributions of B_1, a Bayes' net must not make further independence assumptions. A family of distributions that makes only a subset of the independences assumptions of B_1 can represent all of the same distributions as B_1.

(b) Consider the Bayes' Net $\mathbf{B_2}$ below, and fill in **all the circles** (or select *None of the above*) corresponding to the Bayes' Nets $\mathbf{G_1}$ through $\mathbf{G_8}$ that...



$\mathbf{B_2}$

(i) [2 pts] ...are able to represent **at least one distribution** that $\mathbf{B_2}$ is able to represent.

● $\mathbf{G_1}$ ● $\mathbf{G_2}$ ● $\mathbf{G_3}$ ● $\mathbf{G_4}$

● $\mathbf{G_5}$ ● $\mathbf{G_6}$ ● $\mathbf{G_7}$ ● $\mathbf{G_8}$

○ None of the above.

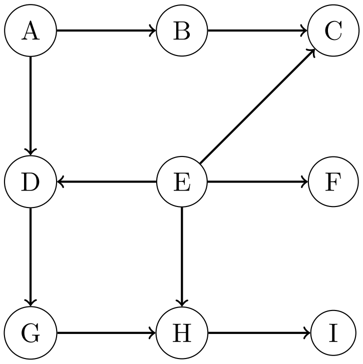
(ii) [2 pts] ...are able to represent **all distributions** that $\mathbf{B_2}$ is able to represent.

○ $\mathbf{G_1}$ ○ $\mathbf{G_2}$ ○ $\mathbf{G_3}$ ● $\mathbf{G_4}$

○ $\mathbf{G_5}$ ● $\mathbf{G_6}$ ● $\mathbf{G_7}$ ● $\mathbf{G_8}$

○ None of the above.

Q3. [9 pts] Bayes' Nets: Independence



Given the above Bayes' Net, select all true statements below. ($\emptyset$ means that no variables are observed.)

● $\mathbf{A} \perp\!\!\!\perp \mathbf{F} \mid \emptyset$ is guaranteed to be true

○ $\mathbf{A} \perp\!\!\!\perp \mathbf{D} \mid \emptyset$ is guaranteed to be true

○ $\mathbf{A} \perp\!\!\!\perp \mathbf{I} \mid \mathbf{E}$ is guaranteed to be true

○ $\mathbf{B} \perp\!\!\!\perp \mathbf{H} \mid \mathbf{G}$ is guaranteed to be true

● $\mathbf{B} \perp\!\!\!\perp \mathbf{E} \mid \mathbf{F}$ is guaranteed to be true

○ $\mathbf{C} \perp\!\!\!\perp \mathbf{G} \mid \mathbf{A}, \mathbf{I}$ is guaranteed to be true

○ $\mathbf{D} \perp\!\!\!\perp \mathbf{H} \mid \mathbf{G}$ is guaranteed to be true

○ $\mathbf{A} \perp\!\!\!\perp \mathbf{F} \mid \emptyset$ is guaranteed to be false

○ $\mathbf{A} \perp\!\!\!\perp \mathbf{D} \mid \emptyset$ is guaranteed to be false

○ $\mathbf{A} \perp\!\!\!\perp \mathbf{I} \mid \mathbf{E}$ is guaranteed to be false

○ $\mathbf{B} \perp\!\!\!\perp \mathbf{H} \mid \mathbf{G}$ is guaranteed to be false

○ $\mathbf{B} \perp\!\!\!\perp \mathbf{E} \mid \mathbf{F}$ is guaranteed to be false

○ $\mathbf{C} \perp\!\!\!\perp \mathbf{G} \mid \mathbf{A}, \mathbf{I}$ is guaranteed to be false

○ $\mathbf{D} \perp\!\!\!\perp \mathbf{H} \mid \mathbf{G}$ is guaranteed to be false

The structure of a Bayes Net cannot make guarantees about *absence* of independence between variables. Therefore, nothing in the right column should ever be marked.

As for the left column, the answers are given by simple, rigourous application of the d-Separation rules.

Q4. [5 pts] Machine Learning: Maximum Likelihood

Identical twins are rare, but just how *un*likely are they? With the help of the sociology department, you have a representative sample of twins to help you answer the question. The twins data gives the following observations (a twin refers to one pair of two people):

- m_i = number of identical male twins and f_i = number of identical female twins
- m_f = number of fraternal male twins and f_f = number of fraternal female twins
- b = number of fraternal opposite gender twins

To model this data, we choose these distributions and parameters:

- Twins are identical with probability θ.
- Given identical twins, the twins are male with probability p.
- Given fraternal twins, the probability of male twins is q^2, probability of female twins is $(1-q)^2$ and probability of oppsite gender twins is $2q(1-q)$.

(a) [2 pts] Write expressions for the likelihood and the log-likelihood of the data as functions of the parameters θ, p, and q for the observations $m_i,\ f_i,\ m_f,\ f_f,\ b$.

The probability of identical male twins is θp, probability of identical female twins is $\theta(1-p)$, probability of fraternal male twins is $(1-\theta)q^2$, probability of fraternal female twins is $(1-\theta)(1-q)^2$ and probability of fraternal opposite gender twins is $(1-\theta)\cdot 2q(1-q)$. Therefore, the likelihood is

$$\begin{aligned} L(\theta, p, q) &= (\theta p)^{m_i} \cdot (\theta(1-p))^{f_i} \cdot ((1-\theta)q^2)^{m_f} \cdot ((1-\theta)(1-q)^2)^{f_f} \cdot ((1-\theta)\cdot 2q(1-q))^b \\ &= \theta^{(m_i+f_i)}(1-\theta)^{(m_f+f_f+b)} \cdot p^{m_i}(1-p)^{f_i} \cdot q^{2m_f}(1-q)^{2f_f}(2q(1-q))^b \end{aligned}$$

And the log likehood is

$$l(\theta, p, q) = (m_i+f_i)\cdot\log\theta + (m_f+f_f+b)\cdot\log(1-\theta) + m_i\cdot\log p + f_i\cdot\log(1-p) + 2m_f\cdot\log q + 2f_f\cdot\log(1-q) + b\cdot\log(2q(1-q))$$

Likelihood $L(\theta, p, q) =$ $\theta^{(m_i+f_i)}(1-\theta)^{m_f+f_f+b} \cdot p^{m_i}(1-p)^{f_i} \cdot q^{2m_f}(1-q)^{2f_f}(2q(1-q))^b$

Log likelihood $l(\theta, p, q) =$ $(m_i+f_i)\log\theta + (m_f+f_f+b)\log(1-\theta) + m_i\log p + f_i\log(1-p) + 2m_f\log(q) + 2f_f\log(1-q) + b\log(2q(1-q))$

(b) [3 pts] What are the maximum likelihood estimates for θ, p and q? (Some work on scratch paper might be helpful.)

To get the maximum likelihood estimate, we have to maximize the log likelihood by taking partial derivatives,

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= \frac{m_i+f_i}{\theta} - \frac{m_f+f_f+b}{1-\theta} = 0 & \theta_{\mathrm{ML}} &= \frac{m_i+f_i}{m_i+f_i+m_f+f_f+b} \\ \frac{\partial l}{\partial p} &= \frac{m_i}{p} - \frac{m_i+f_i}{1-p} = 0 & p_{\mathrm{ML}} &= \frac{m_i}{m_i+f_i} \\ \frac{\partial l}{\partial q} &= \frac{2m_f+b}{q} - \frac{2f_f+b}{1-q} = 0 & q_{\mathrm{ML}} &= \frac{2m_f+b}{2m_f+2f_f+2b} \end{aligned}$$

$$\theta_{\mathrm{ML}} = \frac{m_i + f_i}{m_f + f_f + b + m_i + f_i}$$

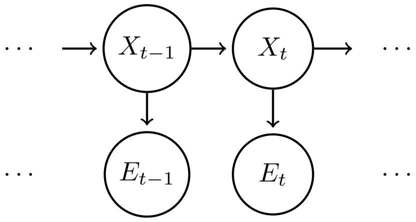
$$p_{\mathrm{ML}} = \frac{m_i}{m_i + f_i}$$

$$q_{\mathrm{ML}} = \frac{2m_f + b}{2m_f + 2f_f + 2b}$$

Q5. [12 pts] Independence in Hidden Markov Models

Below is a full derivation of the forward algorithm updates for Hidden Markov Models. As seen in lecture, we used $e_{1:t}$ to denote all the evidence variables $e_1, e_2, \ldots, e_t$. Similarly, $e_{1:t-1}$ denotes $e_1, e_2, \ldots, e_{t-1}$. For reference, the Bayes net corresponding to the usual Hidden Markov Model is shown on the right side of the derivation below.

$$\begin{aligned}
P(x_t|e_{1:t}) &\propto P(x_t, e_{1:t}) && (1)\\
&= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t}) && (2)\\
&= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t-1}, e_t) && (3)\\
&= \sum_{x_{t-1}} P(e_t \mid x_{t-1}, x_t, e_{1:t-1}) P(x_{t-1}, x_t, e_{1:t-1}) && (4)\\
&= \sum_{x_{t-1}} P(e_t \mid x_t) P(x_{t-1}, x_t, e_{1:t-1}) && (5)\\
&= \sum_{x_{t-1}} P(e_t \mid x_t) P(x_t \mid x_{t-1}, e_{1:t-1}) P(x_{t-1}, e_{1:t-1}) && (6)\\
&= \sum_{x_{t-1}} P(e_t \mid x_t) P(x_t \mid x_{t-1}) P(x_{t-1}, e_{1:t-1}) && (7)\\
&= P(e_t \mid x_t) \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1}, e_{1:t-1}) && (8)
\end{aligned}$$



(a) [2 pts] The following assumption(s) are needed to justify going from step (4) to step (5):
(select all that apply) also correct to select first on left and first on right

● $E_t \perp\!\!\!\perp X_{t-1} \mid X_t$
● $E_t \perp\!\!\!\perp E_k \mid X_t$ for all $1 \leq k \leq t-1$
○ $E_t \perp\!\!\!\perp E_k$ for all $1 \leq k \leq t-1$
○ $E_t \perp\!\!\!\perp E_{t+1} \mid X_t$
○ $E_t \perp\!\!\!\perp E_k \mid X_{t-1}$ for all $1 \leq k \leq t-1$
○ $X_t \perp\!\!\!\perp E_{t+1} \mid X_{t+1}$
○ $X_t \perp\!\!\!\perp E_k \mid X_{t-1}$ for all $1 \leq k \leq t-1$
○ none

(b) [2 pts] The following assumption(s) are needed to justify going from step (5) to step (6):
(select all that apply)

○ $E_t \perp\!\!\!\perp X_{t-1} \mid X_t$
○ $E_t \perp\!\!\!\perp E_k \mid X_t$ for all $1 \leq k \leq t-1$
○ $E_t \perp\!\!\!\perp E_k$ for all $1 \leq k \leq t-1$
○ $E_t \perp\!\!\!\perp E_{t+1} \mid X_t$
○ $E_t \perp\!\!\!\perp E_k \mid X_{t-1}$ for all $1 \leq k \leq t-1$
○ $X_t \perp\!\!\!\perp E_{t+1} \mid X_{t+1}$
○ $X_t \perp\!\!\!\perp E_k \mid X_{t-1}$ for all $1 \leq k \leq t-1$
● none

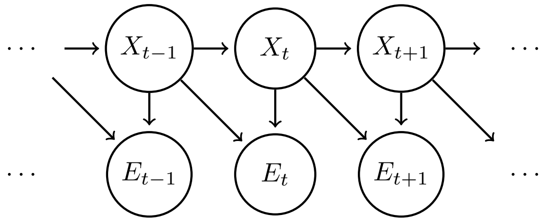
(c) [2 pts] The following assumption(s) are needed to justify going from step (6) to step (7):
(select all that apply)

○ $E_t \perp\!\!\!\perp X_{t-1} \mid X_t$
○ $E_t \perp\!\!\!\perp E_k \mid X_t$ for all $1 \leq k \leq t-1$
○ $E_t \perp\!\!\!\perp E_k$ for all $1 \leq k \leq t-1$
○ $E_t \perp\!\!\!\perp E_{t+1} \mid X_t$
○ $E_t \perp\!\!\!\perp E_k \mid X_{t-1}$ for all $1 \leq k \leq t-1$
○ $X_t \perp\!\!\!\perp E_{t+1} \mid X_{t+1}$
● $X_t \perp\!\!\!\perp E_k \mid X_{t-1}$ for all $1 \leq k \leq t-1$
○ none

Hidden Markov Models can be extended in a number of ways to incorporate additional relations. Since the independence assumptions are different in these extended Hidden Markov Models, the forward algorithm updates will also be different.

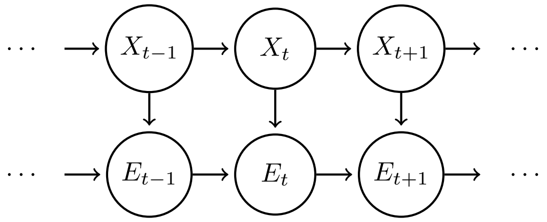
Complete the forward algorithm updates for the extended Hidden Markov Models specified by the following Bayes nets:

(d) [2 pts] $P(x_t|e_{1:t}) \propto \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1}) \cdot \underline{P(e_t \mid x_t, x_{t-1})P(x_t \mid x_{t-1})}$



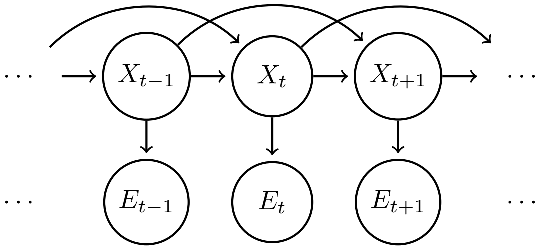
$$\begin{aligned}
P(x_t|e_{1:t}) \propto P(x_t, e_{1:t}) &= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t-1}, e_t) \\
&= \sum_{x_{t-1}} P(e_t \mid x_{t-1}, x_t, e_{1:t-1})P(x_{t-1}, x_t, e_{1:t-1}) \\
&= \sum_{x_{t-1}} P(e_t \mid x_t, x_{t-1})P(x_{t-1}, x_t, e_{1:t-1}) \\
&= \sum_{x_{t-1}} P(e_t \mid x_t, x_{t-1})P(x_t \mid x_{t-1}, e_{1:t-1})P(x_{t-1}, e_{1:t-1}) \\
&= \sum_{x_{t-1}} P(e_t \mid x_t, x_{t-1})P(x_t \mid x_{t-1})P(x_{t-1}, e_{1:t-1})
\end{aligned}$$

(e) [2 pts] $P(x_t|e_{1:t}) \propto \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1}) \cdot \underline{P(e_t \mid x_t, e_{t-1})P(x_t \mid x_{t-1})}$



$$\begin{aligned} P(x_t|e_{1:t}) \propto P(x_t, e_{1:t}) &= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t-1}, e_t) \\ &= \sum_{x_{t-1}} P(e_t \mid x_{t-1}, x_t, e_{1:t-1}) P(x_{t-1}, x_t, e_{1:t-1}) \\ &= \sum_{x_{t-1}} P(e_t \mid x_t, e_{t-1}) P(x_{t-1}, x_t, e_{1:t-1}) \\ &= \sum_{x_{t-1}} P(e_t \mid x_t, e_{t-1}) P(x_t \mid x_{t-1}, e_{1:t-1}) P(x_{t-1}, e_{1:t-1}) \\ &= \sum_{x_{t-1}} P(e_t \mid x_t, e_{t-1}) P(x_t \mid x_{t-1}) P(x_{t-1}, e_{1:t-1}) \end{aligned}$$

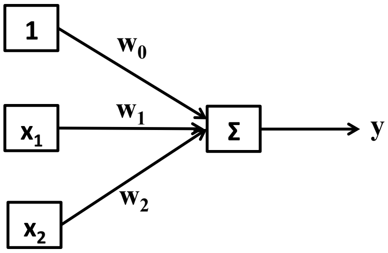
(f) [2 pts] $P(x_t, x_{t+1}|e_{1:t}) \propto \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t-1}) \cdot \underline{P(e_t \mid x_t)P(x_{t+1} \mid x_{t-1}, x_t)}$



$$\begin{aligned} P(x_t, x_{t+1}|e_{1:t}) \propto P(x_t, x_{t+1}, e_{1:t}) &= \sum_{x_{t-1}} P(x_{t-1}, x_t, x_{t+1}, e_{1:t-1}, e_t) \\ &= \sum_{x_{t-1}} P(e_t \mid x_{t-1}, x_t, x_{t+1}, e_{1:t-1}) P(x_{t-1}, x_t, x_{t+1}, e_{1:t-1}) \\ &= \sum_{x_{t-1}} P(e_t \mid x_t) P(x_{t-1}, x_t, x_{t+1}, e_{1:t-1}) \\ &= \sum_{x_{t-1}} P(e_t \mid x_t) P(x_{t+1} \mid x_t, x_{t-1}, e_{1:t-1}) P(x_t, x_{t-1}, e_{1:t-1}) \\ &= \sum_{x_{t-1}} P(e_t \mid x_t) P(x_{t+1} \mid x_t, x_{t-1}) P(x_t, x_{t-1}, e_{1:t-1}) \end{aligned}$$

Q6. [7 pts] Perceptron

(a) Consider the following perceptron, for which the inputs are the always 1 feature and two binary features $x_1 \in \{0,1\}$ and $x_2 \in \{0,1\}$. The output $y \in \{0,1\}$.



$$y = \begin{cases} 1 & \text{if } (w_0 + w_1 \cdot x_1 + w_2 \cdot x_2) > 0 \\ 0 & \text{otherwise} \end{cases}$$

(i) [2 pts] Which one(s) of the following choices for the weight vector $[w_0\ w_1\ w_2]$ can classify y as $y = (x_1$ XOR $x_2)$? XOR is the logical exclusive or operation, which equals to zero when x_1 equals to x_2 and equals to one when x_1 is different from x_2.

○ [1 1 0]
○ [-1.5 1 1]
○ [-2 1 1.5]
○ Any weights that satisfy $(-w_1 - w_2) < w_0 < \min(0, -w_1, -w_2)$.
● No weights can compute the XOR logical relation.

(ii) [2 pts] Which one(s) of the following choices for the weight vector $[w_0\ w_1\ w_2]$ can classify y as $y = (x_1$ AND $x_2)$? Here AND refers to the logical AND operation.

○ [1 1 0]
● [-1.5 1 1]
● [-2 1 1.5]
● Any weights that satisfy $(-w_1 - w_2) < w_0 < \min(0, -w_1, -w_2)$.
○ No weights can compute the logical AND relation.

The truth table for XOR and AND logical operations is:

x_1	x_2	XOR	AND
1	1	0	1
1	0	1	0
0	1	1	0
0	0	0	0

In order to classify y as $y = x_1$ XOR x_2, we need to have $\begin{cases} w_0 + w_1 \cdot 1 + w_2 \cdot 1 \le 0 \\ w_0 + w_1 \cdot 1 + w_2 \cdot 0 > 0 \\ w_0 + w_1 \cdot 0 + w_2 \cdot 1 > 0 \\ w_0 + w_1 \cdot 0 + w_2 \cdot 0 \le 0 \end{cases}$. It is equivalent to $\begin{cases} w_0 \le -w_1 - w_2 \\ w_0 > -w_1 \\ w_0 > -w_2 \\ w_0 \le 0 \end{cases}$, which is impossible. The reason is $w_1 > 0$, $w_2 > 0$ and $-w_1 < w_0 \le -w_1 - w_2$, which makes $w_2 < 0$. Contradiction! So no weights can classify y as x_1 XOR x_2.

Similarly, to classify y as $y = x_1$ AND x_2, we need to have $\begin{cases} w_0 + w_1 \cdot 1 + w_2 \cdot 1 > 0 \\ w_0 + w_1 \cdot 1 + w_2 \cdot 0 \le 0 \\ w_0 + w_1 \cdot 0 + w_2 \cdot 1 \le 0 \\ w_0 + w_1 \cdot 0 + w_2 \cdot 0 \le 0 \end{cases}$. It is equivalent to $\begin{cases} w_0 > -w_1 - w_2 \\ w_0 \le -w_1 \\ w_0 \le -w_2 \\ w_0 \le 0 \end{cases}$, which is $(-w_1 - w_2) < w_0 \le \min(0, -w_1, -w_2)$. So weight [-1.5 1 1] and [-2 1 1.5] and any weights that satisfy $(-w_1 - w_2) < w_0 \le \min(0, -w_1, -w_2)$ can be used to classify $y = x_1$ AND x_2.

(b) [3 pts] Consider a multiclass perceptron with initial weights $w_A = [1\ 0\ 0]$, $w_B = [0\ 1\ 0]$ and $w_C = [\ 0\ 0\ 1]$. For prediction, if there is a tie, A is chosen over B over C. The following table gives a sequence of three training examples to be incorporated. When incorporating the second training example, start from the weights obtained from having incorporated the first training example. Similarly, when incorporating the third training example, start from the weights obtained from having incorporated the first training example and the second training example. Fill in the resulting weights in each row.

feature vector	label	w_A	w_B	w_C
		[1 0 0]	[0 1 0]	[0 0 1]
[1 -2 3]	A	[2 -2 3]	[0 1 0]	[-1 2 -2]
[1 1 -2]	B	[2 -2 3]	[1 2 -2]	[-2 1 0]
[1 -1 -4]	B	[2 -2 3]	[1 2 -2]	[-2 1 0]

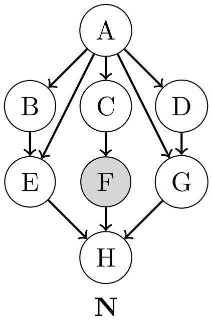
Initial weights: $w_A = [1\ 0\ 0]$, $w_B = [0\ 1\ 0]$ and $w_C = [0\ 0\ 1]$.

After first training example with feature vector [1 -2 3], the algorithm will predict $y = argmax_y(w_y \cdot f) = C$, while the true label is $y^* = A$, so we update $w_A \leftarrow w_A + f =$ [2 -2 3] and $w_C \leftarrow w_C - f =$ [-1 2 -2].

After second training example with feature vector [1 1 -2], the algorithm will predict $y = argmax_y(w_y \cdot f) = C$, while the true label is $y^* = B$, so $w_B \leftarrow w_B + f =$ [1 2 -2] and $w_C \leftarrow w_C - f =$ [-2 1 0].

After third training example with feature vector [1 -1 -4], the algorithm will predict $y = argmax_y(w_y \cdot f) = B$, while the true label is $y^* = B$, so no weight updating is needed.

Q7. [9 pts] Variable Elimination



Variable Elimination. Carry out variable elimination inference for $P(H|+f)$ in the Bayes' net N (shown to the right). Eliminate variables in the order prescribed below. All variables in this question are binary.

Answer format. To write factors made during elimination, include the factor number, arguments, elimination sum, and joined factors like so: $f_1(X,+y) = \sum_z p(z|+y)p(X|z)$.

(a) [1 pt] What factors do we start with after incorporating evidence?

$p(A), p(B|A), p(C|A), P(D|A), P(E|A,B), P(+f|C), P(G|A,D), P(H|E,+f,G)$

(b) [1 pt] Eliminate A to make a new factor f_1.

$f_1(B,C,D,E,G) = \sum_a p(a)p(B|a)p(C|a)P(D|a)P(E|a,B)P(G|a,D)$

(c) [1 pt] Eliminate B to make a new factor f_2.

$f_2(C,D,E,G) = \sum_b f_1(b,C,D,E,G)$

(d) [1 pt] Eliminate C to make a new factor f_3.

$f_3(D,E,+f,G) = \sum_c f_2(c,D,E,G)p(+f|c)$

(e) [1 pt] Eliminate D to make a new factor f_4.

$f_4(+f,E,G) = \sum_d f_3(d,E,G,+f)$

(f) [1 pt] Eliminate E to make a new factor f_5.

$f_5(+f,G,H) = \sum_e f_4(e,G,+f)p(H|e,+f,G)$

(g) [1 pt] Eliminate G to make a new factor f_6.

$f_6(+f,H) = \sum_g f_5(+f,g,H)$

(h) [1 pt] Briefly explain how to compute $p(+h|+f)$ from the set of factors returned by variable elimination.

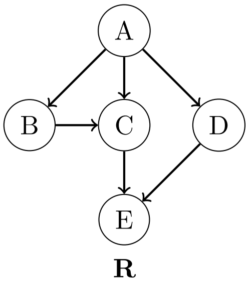
Normalize: $p(+h|+f) = \frac{f6(+f,+h)}{f6(+f,+h)+f6(+f,-h)}$.

(i) [1 pt] Which factor generated during variable elimination has the largest number of unobserved variables?
$f_1(B,C,D,E,G)$

How many table entries are needed to represent it? (Recall all variables are binary.) 2^5

Q8. [6 pts] Reduced Bayes' Nets

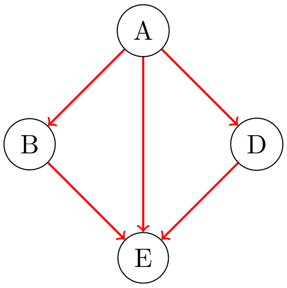
Marginalization and conditioning reduce Bayes' nets.



The Bayes' net R represents the joint distribution $P(A, B, C, D, E)$. Observing or summing out a variable results in a reduced Bayes' net R' without that variable.

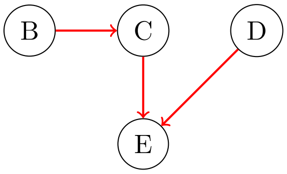
Important: these questions are about separate reductions of the joint distribution. Your answer to one should not affect the others.

(a) [2 pts] Consider the reduced distribution $P(A, B, D, E) = \sum_c P(A, B, c, D, E)$ after C has been summed out. Draw a Bayes' net structure with the minimal number of edges that can represent this distribution. If no edges are needed, write "NONE" next to the diagram.



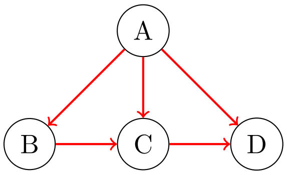
Summing out C connects its neighbors. Multiple solutions exist: $A \to B$ *or* $A \to D$ but *not both* can be flipped. In general, one must take care to not create cycles or independences (conditional or absolute) that are not present in the original Bayes' net R.

(b) [2 pts] Consider the reduced distribution $P(B, C, D, E| + a)$ when A has been observed. Draw a Bayes' net structure with the minimal number of edges that can represent this distribution. If no edges are needed, write "NONE" next to the diagram.



Observing A severs the edges to its children and introduces no new dependence. Multiple solutions exist: $B \to C$ can be flipped.

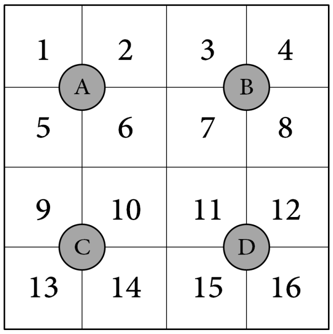
(c) [2 pts] Consider the reduced distribution $P(A, B, C, D | -e)$ when E has been observed.
Draw a Bayes' net structure with the minimal number of edges that can represent this distribution. If no edges are needed, write "NONE" next to the diagram.

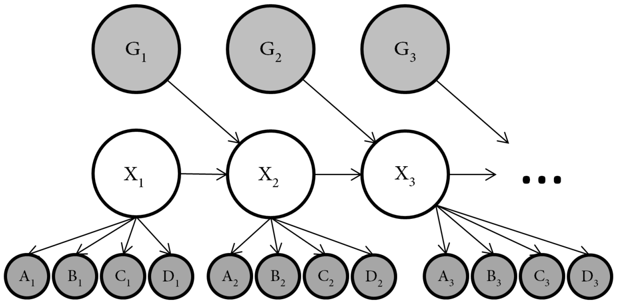


E is the common effect of C and D in R, so observing E requires connecting its parents to represent the activated v-structure. There are two natural solutions, one for each direction of the edge introduced between C and D. Further solutions exist, but care is needed to not introduce cycles or additional independence assumptions not given in R. For instance, the edge $A \to B$ could be reversed.

Q9. [7 pts] Finding WALL-E

Buy N Large, a multi-million dollar corporation, has created a line of garbage collecting robots. They would like to track the location of their friendly garbage collecting robot, WALL-E. WALL-E lives in a 4x4 Manhattan grid city, as shown below. The structure of the HMM is given below, which includes X, the position of WALL-E; G, the readings from a garbage sensor; and (A,B,C,D), readings from the motion sensors.





The garbage sensor G takes on a value in $\{1, 2, ..., 16\}$ corresponding to the square with the most garbage at time t. WALL-E is programmed to move toward the square with the most garbage, but he will only take an optimal action with probability 0.9. In each time step, WALL-E can either stay in the same square, or he can move to an adjacent square. In the case where multiple actions would move you equally close to the desired position, WALL-E has an equal probability of taking any of these actions. In the case that WALL-E fails to take an optimal action, he has an equal probability of taking any of the non-optimal actions. For example, if WALL-E is in square 2, the actions available to him are (EAST, SOUTH, WEST, STOP). If $G_t = 15$, the transition model will look like this:

X_{t+1}	$P(X_{t+1}\|X_t = 2, G_t = 15)$
1	0.05
2	0.05
3	0.45
6	0.45

The motion sensors, (A, B, C, D), take on a value in $\{ON, OFF\}$. At a time t, the sensor adjacent to the square that WALL-E is on always outputs ON. Otherwise, the sensor will output ON or OFF with equal probability. For example, the sensor tables would look like this if $X = 6$:

A	$P(A\|X = 6)$
ON	1
OFF	0

B	$P(B\|X = 6)$
ON	0.5
OFF	0.5

C	$P(C\|X = 6)$
ON	0.5
OFF	0.5

D	$P(D\|X = 6)$
ON	0.5
OFF	0.5

(a) [2 pts] Initially, at $t = 1$, we have particles $[X = 4, X = 2, X = 15]$. We observe that $G_1 = 6$. Use the following random numbers to apply the time update to each of the particles. **Assign square numbers to sample spaces in numerical order.**

[0.7349, 0.5324, 0.1670]

Particle at t=1	Particle after time update
$X = 4$	8
$X = 2$	6
$X = 15$	11

(b) To decouple this question from the previous question, let's say the new particles you have after the time update are $[X = 8, X = 14, X = 11]$. You get the following readings from your sensors $[A = OFF, B = ON, C = ON, D = OFF]$.

(i) [2 pts] What is the weight for each particle?

Particle	Weight
$X = 8$	P(A=OFF|X=8)P(B=ON|X=8)P(C=ON|X=8)P(D=OFF|X=8) = (0.5)(1)(0.5)(0.5) = 0.125
$X = 14$	P(A=OFF|X=14)P(B=ON|X=14)P(C=ON|X=14)P(D=OFF|X=14) = (0.5)(0.5)(1)(0.5) = 0.125
$X = 11$	P(A=OFF|X=11)P(B=ON|X=11)P(C=ON|X=11)P(D=OFF|X=11) = (0.5)(0.5)(0.5)(0) = 0

(ii) [2 pts] It seems, much to your dismay, that sensor C is broken, and will always give a reading of ON. Recalculate the weights with this new knowledge.

Particle	Weight
$X = 8$	P(A=OFF|X=8)P(B=ON|X=8)P(C=ON|X=8)P(D=OFF|X=8) = (0.5)(1)(1)(0.5) = 0.25
$X = 14$	P(A=OFF|X=14)P(B=ON|X=14)P(C=ON|X=14)P(D=OFF|X=14) = (0.5)(0.5)(1)(0.5) = 0.125
$X = 11$	P(A=OFF|X=11)P(B=ON|X=11)P(C=ON|X=11)P(D=OFF|X=11) = (0.5)(0.5)(1)(0) = 0

(iii) [1 pt] To decouple this question from the previous question, let's say that the weights you found for each particle are as follows.

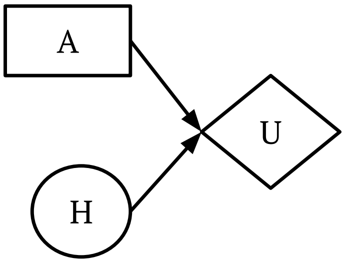
Particle	Weight
$X = 8$	0.225
$X = 14$	0.1
$X = 11$	0.175

If you were to resample 100 new particles, what is the expected number of particles that will be $X = 11$?

Expected number of particles $= \frac{0.175}{0.225+0.1+0.175} \times 100 = 35$

Q10. [14 pts] Decision Networks

After years of battles between the ghosts and Pacman, the ghosts challenge Pacman to a winner-take-all showdown, and the game is a coin flip. Pacman has a decision to make: whether to accept the challenge (*accept*) or decline (*decline*). If the coin comes out heads ($+h$) Pacman wins. If the coin comes out tails ($-h$), the ghosts win. No matter what decision Pacman makes, the outcome of the coin is revealed.



H	$P(H)$
+h	0.5
-h	0.5

H	A	U(H,A)
+h	*accept*	100
-h	*accept*	-100
+h	*decline*	-30
-h	*decline*	50

(a) [5 pts] **Maximum Expected Utility**

Compute the following quantities:

$EU(accept) = P(+h)U(+h, accept) + P(-h)U(-h, accept) = 0.5 * 100 + 0.5 * -100 = 0$

$EU(decline) = P(+h)U(+h, decline) + P(-h)U(-h, decline) = 0.5 * -30 + 0.5 * 50 = 10$

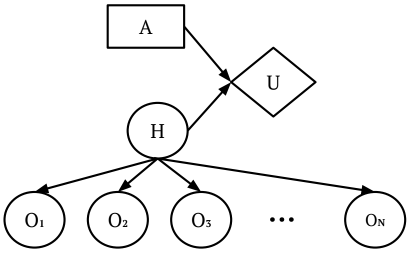
$MEU(\{\}) = max(0, 10) = 10$

Action that achieves $MEU(\{\}) =$ *decline*

(b) VPI relationships When deciding whether to accept the winner-take-all coin flip, Pacman can consult a few fortune tellers that he knows. There are N fortune tellers, and each one provides a prediction O_n for H.

For each of the questions below, circle **all** of the VPI relations that are guaranteed to be true, or select *None of the above.*

(i) [3 pts] In this situation, the fortune tellers give perfect predictions.
Specifically, $P(O_n = +h \mid H = +h) = 1$, $P(O_n = -h \mid H = -h) = 1$, for all n from 1 to N.



○ $\text{VPI}(O_1,\ O_2) \geq \text{VPI}(O_1) + \text{VPI}(O_2)$

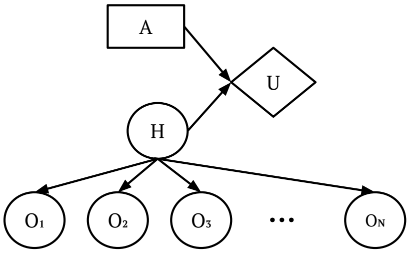
● $\text{VPI}(O_i) = \text{VPI}(O_j)$ where $i \neq j$

○ $\text{VPI}(O_3 \mid O_2,\ O_1) > \text{VPI}(O_2 \mid O_1)$.

○ $\text{VPI}(H) > \text{VPI}(O_1, O_2, \ldots O_N)$

○ None of the above.

(ii) [3 pts] In another situation, the fortune tellers are pretty good, but not perfect.
Specifically, $P(O_n = +h \mid H = +h) = 0.8$, $P(O_n = -h \mid H = -h) = 0.5$, for all n from 1 to N.



○ $\text{VPI}(O_1,\ O_2) \geq \text{VPI}(O_1) + \text{VPI}(O_2)$

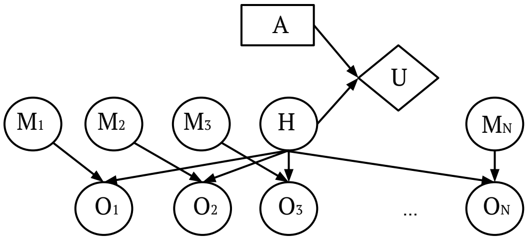
● $\text{VPI}(O_i) = \text{VPI}(O_j)$ where $i \neq j$

○ $\text{VPI}(O_3 \mid O_2,\ O_1) > \text{VPI}(O_2 \mid O_1)$.

● $\text{VPI}(H) > \text{VPI}(O_1, O_2, \ldots O_N)$

○ None of the above.

(iii) [3 pts] In a third situation, each fortune teller's prediction is affected by their mood. If the fortune teller is in a good mood $(+m)$, then that fortune teller's prediction is guaranteed to be correct. If the fortune teller is in a bad mood $(-m)$, then that teller's prediction is guaranteed to be incorrect. Each fortune teller is happy with probability $P(M_n = +m) = 0.8$.



○ $\text{VPI}(M_1) > 0$

● $\forall i\ \ \text{VPI}(M_i|O_i) > 0$

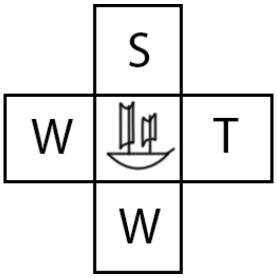
○ $\text{VPI}(M_1,\ M_2,\ \ldots,\ M_N) > \text{VPI}(M_1)$

● $\forall i\ \ \text{VPI}(H) = \text{VPI}(M_i, O_i)$

○ None of the above.

Q11. [15 pts] Argg! Sampling for the Legendary Treasure

Little did you know that Michael and John are actually infamous pirates. One day, they go treasure hunting in the Ocean of Bayes, where rumor says a great treasure lies in wait for explorers who dare navigate in the rough waters. After navigating about the ocean, they are within grasp of the treasure. Their current configuration is represented by the boat in the figure below. They can only make one move, and must choose from the actions: (North, South, East, West). Stopping is not allowed. They will land in either a whirlpool (W), an island with a small treasure (S), or an island with the legendary treasure (T). The utilities of the three types of locations are shown below:

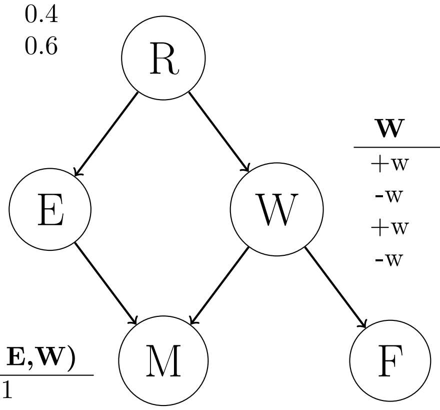


State	**U(State)**
T (Legendary Treasure)	100
S (Small Treasure)	25
W (Whirlpool)	-50

The success of their action depends on the random variable **Movement (M)**, which takes on one of two values: (+m, -m). The Movement random variable has many relationships with other variables: Presence of Enemy Pirates (E), Rain (R), Strong Waves (W), and Presence of Fishermen (F). The Bayes' net graph that represents these relationships is shown below:

R	**P(R)**
+r	0.4
-r	0.6

E	**R**	**P(E | R)**
+e	+r	0.3
-e	+r	0.7
+e	-r	0.6
-e	-r	0.4



W	**R**	**P(W | R)**
+w	+r	0.9
-w	+r	0.1
+w	-r	0.2
-w	-r	0.8

M	**E**	**W**	**P(M | E,W)**
+m	+e	+w	0.1
-m	+e	+w	0.9
+m	+e	-w	0.45
-m	+e	-w	0.55
+m	-e	+w	0.35
-m	-e	+w	0.65
+m	-e	-w	0.9
-m	-e	-w	0.1

F	**W**	**P(F | W)**
+f	+w	0.15
-f	+w	0.85
+f	-w	0.75
-f	-w	0.25

In the following questions we will follow a two-step process:
– (1) Michael and John observed the random variables $R = -r$ and $F = +f$. We then determine the distribution for $P(M| - r, +f)$ via sampling.

– (2) Based on the estimate for $P(M| - r, +f)$, after committing to an action, landing in the intended location of an action successfully occurs with probability $P(M = +m| - r, +f)$. The other three possible landing positions occur with probability $\frac{P(M=-m|-r,+f)}{3}$ each. Use this transition distribution to calculate the optimal action(s) to take and the expected utility of those actions.

(a) **(i)** [1 pt] **Rejection Sampling**: You want to estimate $P(M = +m| - r, +f)$ by rejection sampling. Below is a list of samples that were generated using prior sampling. Cross out those that would be rejected by rejection sampling.

~~$+r$~~	~~$+e$~~	~~$+w$~~	~~$-m$~~	~~$-f$~~	$-r$	$-e$	$+w$	$-m$	$+f$
~~$-r$~~	~~$-e$~~	~~$+w$~~	~~$-m$~~	~~$-f$~~	~~$+r$~~	~~$-e$~~	~~$+w$~~	~~$+m$~~	~~$-f$~~
$-r$	$+e$	$-w$	$-m$	$+f$	$-r$	$-e$	$-w$	$+m$	$+f$
~~$+r$~~	~~$-e$~~	~~$-w$~~	~~$+m$~~	~~$-f$~~	~~$+r$~~	~~$-e$~~	~~$-w$~~	~~$+m$~~	~~$+f$~~
$-r$	$-e$	$-w$	$-m$	$+f$	$-r$	$+e$	$+w$	$-m$	$+f$
$-r$	$+e$	$-w$	$-m$	$+f$	$-r$	$+e$	$-w$	$-m$	$+f$

All samples without the conditioning $-r, +f$ are rejected.

(ii) [1 pt] What is the approximation for $P(M = +m| - r, +f)$ using the remaining samples?
$\frac{1}{7}$, the fraction of accepted samples with $+m$ instantiated.

(iii) [1 pt] What are the optimal action(s) for Michael and John based on this estimate of $P(M = +m|-r, +f)$?
South, West. As $p(+m| - r, +f) = \frac{1}{7}$, $p(-m| - r, +f) = \frac{6}{7}$. Michael and John will succeed in the selected action $\frac{1}{7}$ of the time, or take one of the other 3 actions with equal probability of $\frac{2}{7}$. In this case, $p(+m| - r, +f)$ is so low that deciding to head in the direction of the whirlpool actually decreases the chances of landing in it.

(iv) [1 pt] What is the expected utility for the optimal action(s) based on this estimate of $P(M = +m|-r, +f)$?
$\frac{1}{7} * (-50) + \frac{2}{7} * (-50) + \frac{2}{7} * (25) + \frac{2}{7} * (100) = \frac{100}{7}$, the weighted sum of all four outcomes.

(b) **(i)** [2 pts] **Likelihood Weighting:** Suppose instead that you perform likelihood weighting on the following samples to get the estimate for $P(M = +m| - r, +f)$. You receive 4 samples consistent with the evidence.

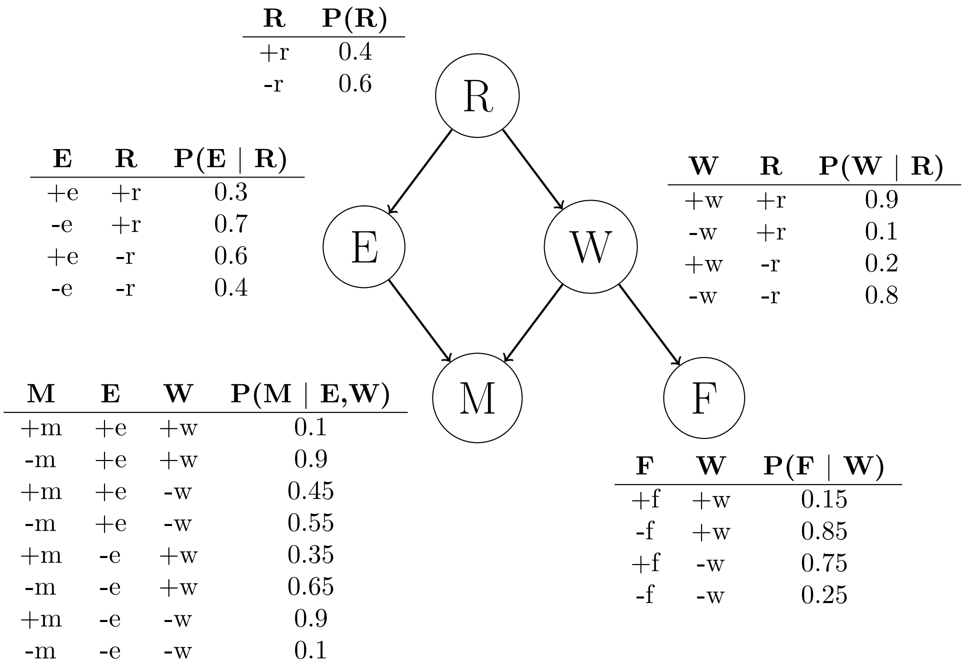
Sample	**Weight**
$-r \quad -e \quad +w \quad +m \quad +f$	$P(-r)P(+f\| + w) = 0.6 * 0.15 = 0.09$
$-r \quad -e \quad -w \quad +m \quad +f$	$P(-r)P(+f\| - w) = 0.6 * 0.75 = 0.45$
$-r \quad -e \quad +w \quad -m \quad +f$	$P(-r)P(+f\| + w) = 0.6 * 0.15 = 0.09$
$-r \quad +e \quad -w \quad -m \quad +f$	$P(-r)P(+f\| - w) = 0.6 * 0.75 = 0.45$

(ii) [1 pt] What is the approximation for $P(M = +m| - r, +f)$ using the samples above?
$\frac{0.09+0.45}{0.09+0.45+0.09+0.45} = \frac{1}{2}$

(iii) [1 pt] What are the optimal action(s) for Michael and John based on this estimate of $P(M = +m|-r, +f)$?
East

(iv) [1 pt] What is the expected utility for the optimal action(s) based on this estimate of $P(M = +m|-r, +f)$?
$\frac{1}{6} * (-50) + \frac{1}{6} * (-50) + \frac{1}{6} * (25) + \frac{1}{2} * (100) = \frac{75}{2}$

Here is a copy of the Bayes' Net, repeated for your convenience.



R	P(R)
+r	0.4
-r	0.6

E	R	P(E | R)
+e	+r	0.3
-e	+r	0.7
+e	-r	0.6
-e	-r	0.4

W	R	P(W | R)
+w	+r	0.9
-w	+r	0.1
+w	-r	0.2
-w	-r	0.8

M	E	W	P(M | E,W)
+m	+e	+w	0.1
-m	+e	+w	0.9
+m	+e	-w	0.45
-m	+e	-w	0.55
+m	-e	+w	0.35
-m	-e	+w	0.65
+m	-e	-w	0.9
-m	-e	-w	0.1

F	W	P(F | W)
+f	+w	0.15
-f	+w	0.85
+f	-w	0.75
-f	-w	0.25

(c) **(i)** [3 pts] **Gibbs Sampling**. Now, we tackle the same problem, this time using Gibbs sampling. We start out with initializing our evidence: $R = -r$, $F = +f$. Furthermore, we start with this random sample:

$-r \quad +e \quad -w \quad +m \quad +f.$

We select variable E to resample. Calculate the numerical value for:
$P(E = +e|R = -r, W = -w, M = +m, F = +f)$.

$P(E = +e|R = -r, W = -w, M = +m, F = +f) = \frac{P(+e|-r)P(+m|+e,-w)}{P(+e|-r)P(+m|+e,-w)+P(-e|-r)P(+m|-e,-w)}$

$= \frac{0.6*0.45}{0.6*0.45+0.4*0.9} = \frac{3}{7}$

We resample for a long time until we end up with the sample:

$-r \quad -e \quad +w \quad +m \quad +f.$

Michael and John are happy for fixing this one sample, but they do not have enough time left to compute another sample before making a move. They will let this one sample approximate the distribution: $P(M = +m| - r, +f)$.

(ii) [1 pt] What is the approximation for $P(M = +m| - r, +f)$, using this one sample?
1

(iii) [1 pt] What are the optimal action(s) for Michael and John based on this estimate of $P(M = +m|-r, +f)$?
East

(iv) [1 pt] What is the expected utility for the optimal action(s) based on this estimate of $P(M = +m|-r, +f)$?
100