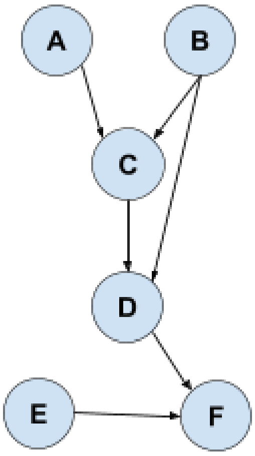
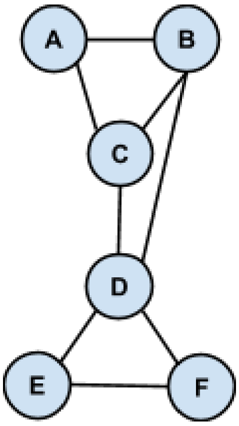
Probabilistic Graphical Models **Final Exam**	**Instructor:** **B. Nasihatkon**	دانشگاه صنعتی خواجه نصیرالدین طوسی K. N. TOOSI UNIVERSITY OF TECHNOLOGY
Name:	**ID:**	**Khordad 1396 - June 2017**

1. Variable Elimination (20 points)



Assume that we are to apply variable elimination to the following Bayesian network.

A) Draw the corresponding Markov network (4 points)



B) If you are to eliminate one variable from the network above, which variable(s) is (are) the best to start with? Which one is the worst? Why? (4 points)

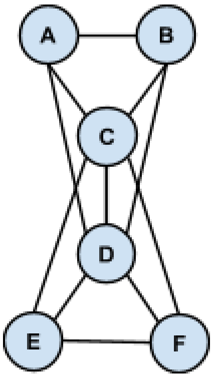
Considering the Markov Network, by eliminating variables A,E and F we sum over a factor of size 3 (a factor of 3 variables), and create no extra links between the nodes. So, these are the best. Considering the Bayesian Network, F is easiest to eliminate since $\sum_F P(F \mid D, E) = 1$. (both answers are acceptable.)

The worst variable to start with is D, as it is connected to four other variables, and creates a link between all pairs among these variables after elimination. Before elimination, it creates a factor of size 5.

C) To compute P(D | F), propose an optimal elimination ordering in terms of algorithm efficiency. Why is this the best (or a best) order? (7 points)

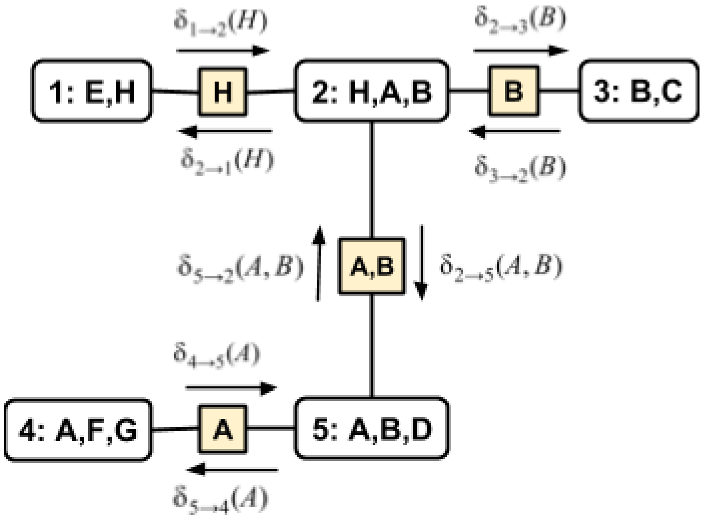
P(D | F) = P(D,F) / P(F). Thus, we need to eliminate A,B,C,E to compute P(D,F) and then D to compute P(F). Therefore, D must always be the last variable to eliminate. An optimal ordering can be A,B,E,C,D. There are other orderings like E,A,B, C,D. By following these ordering schemes no extra links are created, and the induced graph will be equal to the original Markov network. More importantly, the size of intermediate factors created during elimination is minimal.

D) Assume that we want to eliminate B, A, D, F and C in order. Draw the corresponding ***induced graph***. (5 points)



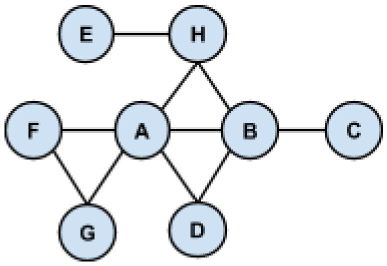
2. Junction Tree (22 points)

Consider the following junction tree (clique tree)



A) Write down the corresponding message beside each arrow. (2 points)

B) Draw a Markov network (MRF) corresponding to the above clique tree (4 points).



C) Write down the potential functions for the markov graph in part (B). Assume that there are **only binary and ternary** potentials and each cluster corresponds to **exactly one** potential function.For each potential function write down the corresponding cluster number (1,2,3,4 or 5). (3 points)

cluster 1: $\phi_1(E,H)$, cluster 2: $\phi_2(H,A,B)$, cluster 3: $\phi_3(B,C)$,
cluster 4: $\phi_4(A,B,D)$, cluster 5: $\phi_5(A,F,G)$

D) What is the minimum number of message computations needed for the Belief Propagation algorithm to converge to the right solution? Write one such ordering of messages. (6 points).

A minimum of 8 messages. One such ordering is $\delta_{1\to2}(H)$, $\delta_{3\to2}(B)$, $\delta_{2\to5}(A,B)$, $\delta_{5\to4}(A)$, $\delta_{4\to5}(A)$, $\delta_{5\to2}(A,B)$, $\delta_{2\to1}(H)$, $\delta_{2\to3}(B)$.

E) Assume that all variables are binary (∈ $\{0,1\}$), $\phi_1(E,H) = exp(1(E=H))$, where 1(.) is the indicator function, $\phi_2(H,A,B) = exp(3\ A\ B\ H + 2A - BH)$, $\phi_3(B,C) = exp(2\ B\ C + B)$, and we are to perform **max-sum message passing** for **MAP** estimation. Derive $\delta_{1\to2}(H)$, $\delta_{3\to2}(B)$, **and then** $\delta_{2\to5}(A,B)$. Notice that the functions $\delta_{i\to j}$ are **max-sum** messages. You can either write a formula or a tabular representation. (7 points)

$\theta_1(E,H) = log\ \phi_1(E,H) = 1(E=H)$

$\theta_2(H,A,B) = \ log\ \ \phi_2(H,A,B) = 3\ A\ B\ H + 2A - BH$

$\theta_3(B,C)\ = log\ \phi_3(B,C)\ = 2\ B\ C\ +\ B$

$\delta_{1\to2}(H)\ =\ \max_E \theta_1(E,H)\ =\ \max_E 1(E=H)\ =\ 1$

$\delta_{3\to2}(B)\ =\ \max_C \theta_2(H,A,B)\ =\ \max_C 2\ B\ C\ +\ B\ =\ 3B$

$$\delta_{2\to5}(A,B) = \max_H \theta_2(H,A,B) + \delta_{1\to2}(H) + \delta_{3\to2}(B) = \max_H 3ABH + 2A - BH + 1 + 3B$$

$$= \max_H (3AB - B)H + 2A + 3B + 1$$

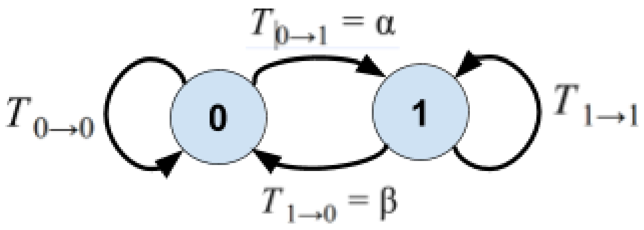
therefore:

$if\ A = 0 \Rightarrow \delta_{2\to5}(A,B) = 3B + 1$

$if\ A = 1 \Rightarrow \delta_{2\to5}(A,B) = 5B + 3$

(You could also use table representations)

3. Random walk / MCMC (22 points)



Consider a markov chain with the following transition model for a 1D distribution $P(X)$ with a binary variable $X \in \{0,1\}$, in which $\alpha = T(0 \to 1)$ and $\beta = T(1 \to 0)$.

A) What are the values of $T(0 \to 0)$ and $T(1 \to 1)$ in terms of α and β. (1 point)

$T(0 \to 0) = 1 - \alpha$ and $T(1 \to 1) = 1 - \beta$

B) Assume that the transition probabilities α and β are given. Derive the corresponding stationary distribution $P^\infty(0) = \pi(0)$ and $P^\infty(1) = \pi(1)$ in terms of α and β. Write down the full derivations. (8 points)

The equations are:

$$P^\infty(0) = T_{0\to0} P^\infty(0) + T_{1\to0} P^\infty(1)$$
$$P^\infty(1) = T_{0\to1} P^\infty(0) + T_{1\to1} P^\infty(1)$$
$$P^\infty(0) + P^\infty(1) = 1$$

Let $p = P^\infty(0)$ and $q = 1 - p = P^\infty(1)$. Then we have

$$p = (1-\alpha)\, p + \beta\, q$$
$$q = \alpha\, p + (1-\beta)\, q$$
$$p + q = 1$$

Solving the above gives $p = P^\infty(0) = \beta/(\alpha+\beta)$ and $q = P^\infty(1) = \alpha/(\alpha+\beta)$

C) Assume $\alpha = 0.3, \beta = 0.8$. What is the corresponding stationary distribution $P^\infty(X)$? (2 points)

$P^\infty(0) = \beta/(\alpha+\beta) = 0.8/(0.3+0.8) = 8/11$ and $P^\infty(1) = \alpha/(\alpha+\beta) = 3/11$

D) Now, we want to solve the inverse problem. Assume that a special stationary distribution $P^{\infty}(X)$ is desired, that is, $P^{\infty}(0) = p$ and $P^{\infty}(1) = 1-p$ for a given p. We want to determine α and β in terms of p. Using the result from part (B), determine the ratio α/β in terms of p. Show that every solution $\alpha, \beta \in [0,1]$ with this ratio is an answer to our problem, and therefore for a given stationary distribution the solution (α, β) is not unique. (Assume that $0 < p < 1$) (8 points)

In part (B) we showed that $p = P^{\infty}(0) = \beta/(\alpha+\beta)$. Since p > 0 then β must be nonzero. Therefore, we have $p = 1/(1+\alpha/\beta)$. Hence, any (α, β) for which $\alpha/\beta = (1-p)/p$ is a solution resulting in $P^{\infty}(0) = p,\ P^{\infty}(1) = 1-p$.

E) Assume that we want to design a markov chain for which $P^{\infty}(0) = p = 0.4$. Obtain two different solutions (α, β) such that for the first one $\beta = 0.1$ and for the second one $\beta = 0.2$ (3 points).

$\alpha/\beta = (1-p)/p = 0.6/0.4 = 3/2$

$\beta = .1 \Rightarrow \alpha = 0.15 \Rightarrow (\alpha, \beta) = (0.15, 0.1)$

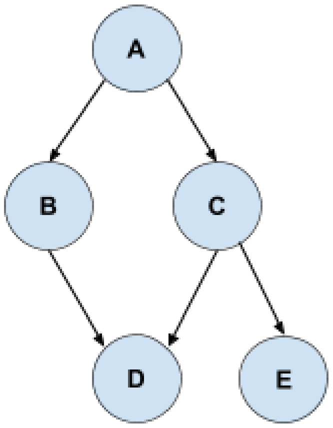
$\beta = .2 \Rightarrow \alpha = 0.3 \Rightarrow (\alpha, \beta) = (0.3, 0.2)$

F) ** Which of the two solutions in part (E) do you think gives a better random walk algorithm in terms of mixing more quickly? Give an intuitive explanation? Can you give an optimal solution (α, β) for part (E)? **(3+3 extra points)**

With $(\alpha, \beta) = (0.3, 0.2)$ we have a higher probability of moving among states (not staying in the same state). Thus, the probability converges more quickly to the stationary distribution.

The best solution is when α and β are as large as possible. But since both of them have to be less than 1, the best solution is $(\alpha, \beta) = (1, 2/3)$.

4. Parameter learning Bayesian Networks (16 points)



Consider the following Bayesian network with all-binary variables ($\in \{0,1\}$). Assume that the training data $X^1, X^2, \ldots, X^N$ is available, where $X^i = (a^i, b^i, c^i, d^i, e^i)$.

A) Write down the log-likelihood function in terms of the **logarithm of CPDs**. (6 points)

$$log \prod_{i=1}^{N} P(X^i) = log \prod_{i=1}^{N} P(a^i)\, P(b^i \mid a^i)\, P(c^i \mid a^i)\, P(d^i \mid b^i, c^i)\, P(e^i \mid c^i)$$

$$= \sum_{i=1}^{N} log\, P(a^i) + log\, P(b^i \mid a^i) + log\, P(c^i \mid a^i) + log\, P(d^i \mid b^i, c^i) + log\, P(e^i \mid c^i)$$

B) Assume that all the CPDs in this network are parameterized independently, except P(C | A) and P(E | C) which have shared parameters, i.e. share the same table, that is P(C = x | A = y) = P(E = x| C = y). Consider the following training data. Write the tabular representation of the Maximum Likelihood solution for each of the CPDs. (10 points)

	a^i	b^i	c^i	d^i	e^i
X^1	0	1	0	0	0
X^2	0	0	0	1	1
X^3	1	0	0	1	0
X^4	1	1	0	0	1
X^5	0	1	0	1	0
X^6	0	0	1	0	1

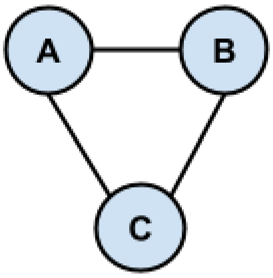
a	$P(A=a)$
0	0.667
1	0.333

a	b	$P(B=b \mid A=a)$
0	0	0.5
0	1	0.5
1	0	0.5
1	1	0.5

b	c	d	$P(D=d \mid B=b, C=c)$
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	0.667
1	0	1	0.333
1	1	0	cannot be computed
1	1	1	cannot be computed

a	e	$P(C=e \mid A=a) = P(E=e \mid C=a)$
0	0	0.667
0	1	0.333
1	0	0.667
1	1	0.333

5. Parameter learning MRFs (20 points)



Consider the following MRF, on **binary** variables **A, B, C** $\in \{0,1\}$ with joint distribution

$$P(A,B,C) = \frac{1}{Z} exp(w_1\ 1(A = B) + w_2\ 1(B = C) + w_3\ 1(C = A)\ ),$$

where **1(X = Y)** is equal to 1 if X = Y and zero otherwise.

A) Derive the partition function Z as a function of w_1, w_2, w_3. (4 points)

$$\begin{aligned} Z &= \sum_{A,B,C} exp(w_1\ 1(A = B) + w_2\ 1(B = C) + w_3\ 1(C = A)\ ) \\ &= exp(w_1\ 1(0 = 0) + w_2\ 1(0 = 0) + w_3\ 1(0 = 0)\ ) \\ &+ exp(w_1\ 1(0 = 0) + w_2\ 1(0 = 1) + w_3\ 1(1 = 0)\ ) \\ &+ exp(w_1\ 1(0 = 1) + w_2\ 1(1 = 0) + w_3\ 1(0 = 0)\ ) \\ &+ exp(w_1\ 1(0 = 1) + w_2\ 1(1 = 1) + w_3\ 1(1 = 0)\ ) \\ &+ exp(w_1\ 1(1 = 0) + w_2\ 1(0 = 0) + w_3\ 1(0 = 1)\ ) \\ &+ exp(w_1\ 1(1 = 0) + w_2\ 1(0 = 1) + w_3\ 1(1 = 1)\ ) \\ &+ exp(w_1\ 1(1 = 1) + w_2\ 1(1 = 0) + w_3\ 1(0 = 1)\ ) \\ &+ exp(w_1\ 1(1 = 1) + w_2\ 1(1 = 1) + w_3\ 1(1 = 1)\ ) \\ &= 2\ exp(w_1 + w_2 + w_3) + 2\ exp(w_1) + 2\ exp(w_2) + 2\ exp(w_3) \end{aligned}$$

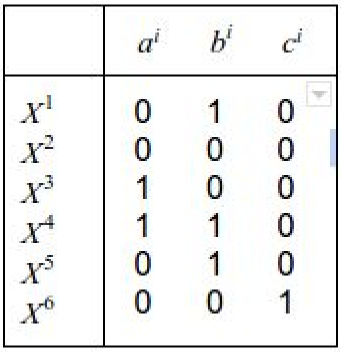
B) Consider the training data $X^1, X^2, \ldots, X^N$, where $X^i = (a^i, b^i, c^i)$. Write down the log-likelihood function in terms of the data (a^i, b^i, c^i) and the weights w_1, w_2, w_3. Simplify your answer as much as possible (4 points)

$$L(w_1, w_2, w_3) = log \prod_{i=1}^{N} P(a^i, b^i, c^i) = \sum_{i=1}^{N} (w_1\ 1(a^i = b^i) + w_2\ 1(b^i = c^i) + w_3\ 1(c^i = a^i)\ ) - N log(Z)$$

$$= \sum_{i=1}^{N} (w_1\ 1(a^i = b^i) + w_2\ 1(b^i = c^i) + w_3\ 1(c^i = a^i)\ ) - N\ log(e^{w_1+w_2+w_3} + e^{w_1} + e^{w_2} + e^{w_3}) - log(2)\ N$$

$$= w_1 \sum_{i=1}^{N} 1(a^i = b^i) + w_2 \sum_{i=1}^{N} 1(b^i = c^i) + w_3 \sum_{i=1}^{N} 1(c^i = a^i) - N\ log(e^{w_1+w_2+w_3} + e^{w_1} + e^{w_2} + e^{w_3}) - log(2)\ N$$

C) Derive the log-likelihood function for the following training data. Simplify your result as much as you can. (3 points)



	a^i	b^i	c^i
X^1	0	1	0
X^2	0	0	0
X^3	1	0	0
X^4	1	1	0
X^5	0	1	0
X^6	0	0	1

$$L(w_1, w_2, w_3) = 3\,w_1 + 2\,w_2 + 3\,w_3 - 6\,log(e^{w_1+w_2+w_3} + e^{w_1} + e^{w_2} + e^{w_3}) - 6\,log\,2$$

D) Which of the following assignments to w_1, w_2, w_3 better describes the data? Why? (3 points)

a) $w_1, w_2, w_3 = (2, 2, 1)$

b) $w_1, w_2, w_3 = (2, 1, 2)$

$L(2, 2, 1) = 13 - 6\,log(e^5 + e^2 + e^2 + e^1) - 6\,log\,2$

$L(2, 1, 2) = 14 - 6\,log(e^5 + e^2 + e^1 + e^2) - 6\,log\,2$

$L(2, 1, 2) > L(2, 2, 1)$, thus the second choice of weight is better in terms of maximum-likelihood.

E) Take derivatives of the log-likelihood function of part (C) with respect to w_1 , w_2 and w_3 , and set them equal to zero. Let $a = e^{w_1}$, $b = e^{w_2}$, and $c = e^{w_3}$. Derive polynomial equations in terms of a, b, c for optimal w_1, w_2, w_3. (4 points)

$$\begin{aligned}
\partial L(w_1, w_2, w_3) / \partial w_1 &= 3 - 6\,(e^{w_1+w_2+w_3} + e^{w_1}) / (e^{w_1+w_2+w_3} + e^{w_1} + e^{w_2} + e^{w_3}) \\
&= 3 - 6\,(abc + a)/(abc + a + b + c) \\
\partial L(w_1, w_2, w_3) / \partial w_2 &= 2 - 6\,(e^{w_1+w_2+w_3} + e^{w_2}) / (e^{w_1+w_2+w_3} + e^{w_1} + e^{w_2} + e^{w_3}) \\
&= 2 - 6\,(abc + b)/(abc + a + b + c) \\
\partial L(w_1, w_2, w_3) / \partial w_3 &= 3 - 6\,(e^{w_1+w_2+w_3} + e^{w_3}) / (e^{w_1+w_2+w_3} + e^{w_1} + e^{w_2} + e^{w_3}) \\
&= 3 - 6\,(abc + c)/(abc + a + b + c)
\end{aligned}$$

$$\begin{aligned}
\partial L(w_1, w_2, w_3) / \partial w_1 &= 0 \Rightarrow abc + a + b + c = 2\,abc + 2a \Rightarrow abc + a - b - c = 0 \quad (1) \\
\partial L(w_1, w_2, w_3) / \partial w_2 &= 0 \Rightarrow abc + a + b + c = 3\,abc + 3b \Rightarrow 2\,abc - a + 2b - c = 0 \quad (2) \\
\partial L(w_1, w_2, w_3) / \partial w_1 &= 0 \Rightarrow abc + a + b + c = 2\,abc + 2c \Rightarrow abc - a - b + c = 0 \quad (3)
\end{aligned}$$

F) Using the result of part (E) prove that for optimal parameters we have $w_1 = w_3$. (2 points)

Subtracting equation (3) from (1) in part (E) gives

$2\,a - 2\,c = 0 \Rightarrow a = c \Rightarrow e^{w_1} = e^{w_2} \Rightarrow w_1 = w_2$