
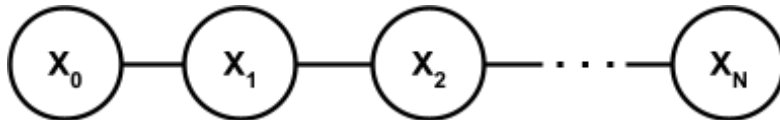


<b>Probabilistic Graphical Models Final Exam - Spring 1398 (2019)</b>	<b>Instructor: B. Nasihatkon</b>	 دانشگاه صنعتی خواجه نصیرالدین طوسی K. N. TOOSI UNIVERSITY OF TECHNOLOGY
<b>Name:</b>	<b>ID:</b>	<b>Khordad 1398 - June 2019</b>

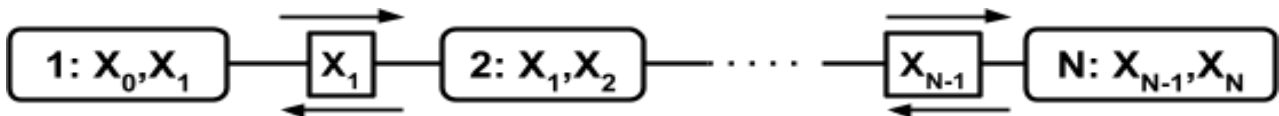
## Question 0- Junction-tree / MAP Inference / Max-Sum Message Passing (46 points)



The above is an MRF chain over binary variables  $X_t \in \{0, 1\}$  with

$$P(X_0, X_1, \dots, X_N) = \frac{1}{Z} \exp\left(\sum_{t=1}^N \theta_t(X_{t-1}, X_t)\right), \text{ where } \theta_t(X_{t-1}, X_t) = t X_{t-1} X_t,$$

that is  $\theta_1(X_0, X_1) = X_0 X_1$ ,  $\theta_2(X_1, X_2) = 2 X_1 X_2$ , and so on (hence the chain is non-stationary). We intend to perform MAP inference using **max-sum** message passing in the following junction-tree (clique-tree).



**\* In all cases, simplify your answers as much as you can.**

A) Compute the message  $\delta_{1 \rightarrow 2}(X_1)$ . Show that  $\delta_{1 \rightarrow 2}(1) > \delta_{1 \rightarrow 2}(0)$ . (3 points)

$$\delta_{1 \rightarrow 2}(X_1) = \max_{X_0} \theta_1(X_0, X_1) = \max_{X_0} X_0 X_1 = X_1$$

$$1 = \delta_{1 \rightarrow 2}(1) > \delta_{1 \rightarrow 2}(0) = 0$$

B) Assume that  $\delta_{t-1 \rightarrow t}(1) > \delta_{t-1 \rightarrow t}(0)$ . Obtain a recursive formula for  $\delta_{t \rightarrow t+1}(X_t)$  in terms of  $\delta_{t-1 \rightarrow t}(\cdot)$ . Your final formula must not include the max operator. (6 points)

$$\begin{aligned} \delta_{t \rightarrow t+1}(X_t) &= \max_{X_{t-1}} \delta_{t-1 \rightarrow t}(X_{t-1}) + \theta_t(X_{t-1}, X_t) \\ &= \max_{X_{t-1}} \delta_{t-1 \rightarrow t}(X_{t-1}) + t X_{t-1} X_t \\ &= \delta_{t-1 \rightarrow t}(1) + t X_t \end{aligned}$$

C) Using parts A and B, prove that  $\delta_{t \rightarrow t+1}(1) > \delta_{t \rightarrow t+1}(0)$  for all  $t$  (by induction). (4 points)

For  $t=1$  we have  $\delta_{1 \rightarrow 2}(1) > \delta_{1 \rightarrow 2}(0)$  from part A.

For  $t>1$  assume  $\delta_{t-1 \rightarrow t}(1) > \delta_{t-1 \rightarrow t}(0)$ , then from part B we have

$$\begin{aligned} \delta_{t \rightarrow t+1}(X_t) &= \delta_{t-1 \rightarrow t}(1) + t X_t \\ \delta_{t-1 \rightarrow t}(1) + t &= \delta_{t \rightarrow t+1}(1) > \delta_{t \rightarrow t+1}(0) = \delta_{t-1 \rightarrow t}(1) \text{ (claim proved by} \\ &\text{induction).} \end{aligned}$$

D) Obtain an explicit formula for  $\delta_{t \rightarrow t+1}(X_t)$ . (7 points)

$$\begin{aligned} \delta_{t \rightarrow t+1}(X_t) &= \delta_{t-1 \rightarrow t}(1) + t X_t \Rightarrow \delta_{t \rightarrow t+1}(1) = \delta_{t-1 \rightarrow t}(1) + t \\ \Rightarrow \text{(by induction)} \\ \delta_{t \rightarrow t+1}(1) \\ &= \delta_{t-2 \rightarrow t-1}(1) + (t-1) + t = \delta_{1 \rightarrow 2}(1) + 2 + 3 + \dots + (t-1) + t \\ &= 1 + 2 + 3 + \dots + (t-1) + t = t(t+1)/2 \\ \delta_{t \rightarrow t+1}(X_t) &= \delta_{t-1 \rightarrow t}(1) + t X_t = (t-1)t/2 + t X_t. \end{aligned}$$

E) Taking the steps in parts A-D, the backward messages  $\delta_{t+1 \rightarrow t}(X_t)$  are obtained as

$$\delta_{t+1 \rightarrow t}(X_t) = N(N+1)/2 - (t+1)(t+2)/2 + (t+1)X_t$$

check at the above is correct for the right-most messages  $\delta_{N \rightarrow N-1}(X_{N-1})$  and

$$\delta_{N-1 \rightarrow N-2}(X_{N-2}) \text{ (4 points)}$$

$$\delta_{N \rightarrow N-1}(X_{N-1}) = \max_{X_N} \theta_N(X_{N-1}, X_N) = \max_{X_N} N X_{N-1} X_N = N X_{N-1}$$

F) Obtain the max-sum belief  $\beta_t(X_{t-1}, X_t)$  for cluster  $1 < t < N$ . (4 points)

$$\begin{aligned} \beta_t(X_{t-1}, X_t) &= \theta_t(X_{t-1}, X_t) + \delta_{t+1 \rightarrow t}(X_{t+1}) + \delta_{t-1 \rightarrow t}(X_{t-1}) \\ &= t X_{t-1} X_t + N(N+1)/2 - (t+1)(t+2)/2 + (t+1)X_t + (t-2)(t-1)/2 \\ &= t X_{t-1} X_t + N(N+1)/2 - (t+1)(t+2)/2 + (t+1)X_t + (t-2)(t-1)/2 \\ &= t X_{t-1} X_t + (t+1)X_t + (t-1)X_{t-1} + N(N+1)/2 + 2t + 1 \end{aligned}$$

G) Obtain the max-sum sepset belief  $\beta_t(X_t)$ . (4 points)

$$\begin{aligned}\beta_t(X_t) &= \delta_{t \rightarrow t+1}(X_t) + \delta_{t+1 \rightarrow t}(X_t) \\ &= (t-1)t/2 + tX_t + N(N+1)/2 - (t+1)(t+2)/2 + (t+1)X_t \\ &= (2t+1)X_t - 2t - 1 + N(N+1)/2\end{aligned}$$

H) Find the MAP solution **using part G**. (4 points)

The junction-tree algorithm gives the exact max-marginals. Thus, the optimal  $X_t$  is

$$X_t^* = \arg \max_{X_t} \beta_t(X_t) = \arg \max_{X_t} (2t+1)X_t - 2t - 1 + N(N+1)/2 = 1.$$

Therefore, the MAP solution is when all the variables  $X_t$  are set to 1.

## Question 2 - Gibbs Sampling (14 points)

We use the following Gibbs Sampling algorithm to get samples from the joint distribution of the MRF in Question 1.

```

set t = 0
set  $X_i^0 = 0$  for  $i = 0, 1, \dots, N$ 
while not mixed
  for i = 0 to N
    set  $X_i^{t+1}$  = a sample the distribution  $P_i^t(X_i)$ 
  end
  set t = t+1
end
output  $X_0^t, X_1^t, \dots, X_N^t$ 

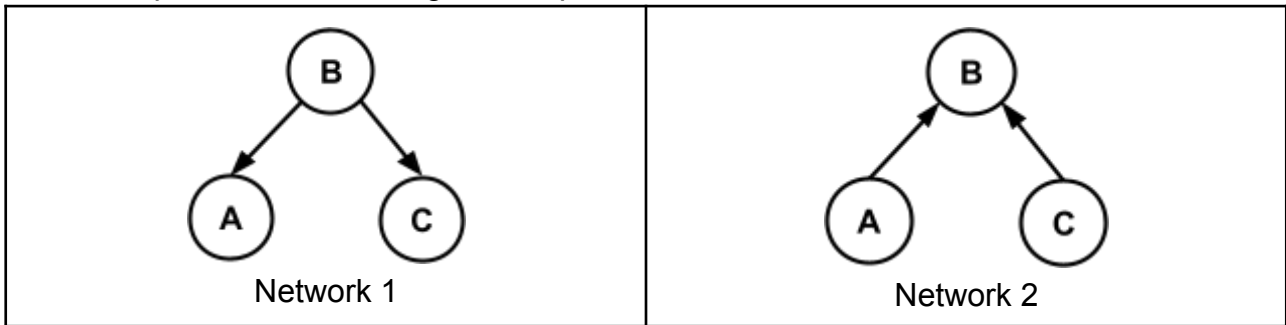
```

Obtain explicit formulae for  $P_1^t(X_1)$  and  $P_N^t(X_N)$  and  $P_i^t(X_i)$  for  $1 < i < N$ ? Write down full derivations and simplify your answers as much as possible. Notice that these quantities can be functions of  $X_0^t, X_1^t, \dots, X_N^t$  and  $X_0^{t+1}, X_1^{t+1}, \dots, X_N^{t+1}$  (14 Points)



# Bayes Nets Parameter/Structure Learning

Consider the following Bayesian networks with binary variables  $A, B, C \in \{0, 1\}$ , where the CPDs are parameterized using table representation.



- In all cases, simplify your answer as much as you can.

A) Parameterize each network using table representation. How many independent parameters each network has? (8 points)

B) Given the training data on the right, write down the log-likelihood function in terms of the network parameters for each network. (9 points)

$a^i$	$b^i$	$c^i$
0	0	0
1	0	1
1	1	0
0	1	1
0	0	1

C) Obtain the optimal parameters for each network in terms of the log-likelihood. (8 points)

D) Compute the likelihood score for each network *by substituting the optimal parameters in part C in the log-likelihood function in part B*. Which model is preferred by the likelihood score? (9 points)

E) Compute the BIC score for each model. Which model is preferred by the BIC score? ( $\log(5) \approx 1.6$ ) (6 points).

$$BIC(data) = \text{likelihoodscore}(data) - 1/2 \log(M) \#(\text{independent parameters})$$