School of Computing
National University of Singapore
CS5340: Uncertainty Modeling in AI
Semester 1, AY 2018/19

Exercise 2

Question 1

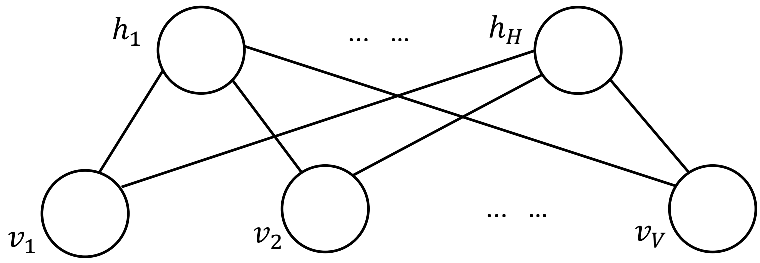


Fig. 1.1

The restricted Boltzmann machine is a Markov Random Field (MRF) defined on a bipartite graph as shown in Fig. 3.1. It consists of a layer of visible variables $\boldsymbol{v} = [v_1, \dots, v_V]^T$ and hidden variables $\boldsymbol{h} = [h_1, \dots, h_H]^T$, where all variables are binary taking states $\{0,1\}$. The joint distribution of the MRF is given by:

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b})} \exp(\boldsymbol{v}^T \boldsymbol{W} \boldsymbol{h} + \boldsymbol{a}^T \boldsymbol{v} + \boldsymbol{b}^T \boldsymbol{h}),$$

where $\theta = \{\boldsymbol{W}_{V\times H}, \boldsymbol{a}_{V\times 1}, \boldsymbol{b}_{H\times 1}\}$ are the parameters of the potential functions, and $Z(.)$ is the partition function.

a) Given that:

$$p(h_i = 1 \mid \boldsymbol{v}) = \sigma\left(b_i + \textstyle\sum_j W_{ji} v_j\right),$$

where $\sigma(x) = \frac{e^x}{1+e^x}$ is the sigmoid activation function. Show that the distribution of hidden units conditioned on the visible units factorizes as:

$$p(\boldsymbol{h} \mid \boldsymbol{v}) = \textstyle\prod_i p(h_i \mid \boldsymbol{v}).$$

Show all your workings clearly.

Answer:

Using product rule, we have:

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{p(\boldsymbol{h},\boldsymbol{v})}{\sum_{\boldsymbol{h}} p(\boldsymbol{h},\boldsymbol{v})} \\
&= \frac{\frac{1}{Z} exp\{\boldsymbol{v}^T\boldsymbol{W}\boldsymbol{h} + \boldsymbol{a}^T\boldsymbol{v} + \boldsymbol{b}^T\boldsymbol{h}\}}{\frac{1}{Z}\sum_{\boldsymbol{h}} exp\{\boldsymbol{v}^T\boldsymbol{W}\boldsymbol{h} + \boldsymbol{a}^T\boldsymbol{v} + \boldsymbol{b}^T\boldsymbol{h}\}} \\
&= \frac{exp\{(\boldsymbol{v}^T\boldsymbol{W} + \boldsymbol{h}^T)\boldsymbol{h}\} exp\{\boldsymbol{a}^T\boldsymbol{v}\}}{\sum_{(\boldsymbol{h}} exp\{\boldsymbol{v}^T\boldsymbol{W} + \boldsymbol{h}^T)\boldsymbol{h}\} exp\{\boldsymbol{a}^T\boldsymbol{v}\}} \\
&= \frac{exp\{(\boldsymbol{v}^T\boldsymbol{W} + \boldsymbol{h}^T)\boldsymbol{h}\}}{\sum_{(\boldsymbol{h}} exp\{\boldsymbol{v}^T\boldsymbol{W} + \boldsymbol{h}^T)\boldsymbol{h}\}}
\end{aligned}
$$

Let $\boldsymbol{m}^T = \boldsymbol{v}^T\boldsymbol{W} + \boldsymbol{b}^T$ and since $\boldsymbol{h} = [h_1, h_2, ...h_H]^T$, we have:

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{exp\{[m_1, m_2...m_H][h_1, h_2...h_H]^T\}}{\sum_{\boldsymbol{h}} exp\{[m_1, m_2...m_H][h_1, h_2...h_H]^T\}} \\
&= \frac{exp\{m_1h_1, m_2h_2...m_Hh_H\}}{\sum_{\boldsymbol{h}} exp\{m_1h_1, m_2h_2...m_Hh_H\}} \\
&= \frac{exp(m_1h_1)exp(m_2h_2)...exp(m_Hh_H)}{\sum_{h_1}\sum_{h_2}...\sum_{h_H} exp(m_1h_1)exp(m_2h_2)...exp(m_Hh_H)} \\
&= \frac{exp(m_1h_1)}{\sum_{h_1} exp(m_1h_1)} \frac{exp(m_2h_2)}{\sum_{h_2} exp(m_2h_2)} ... \frac{exp(m_Hh_H)}{\sum_{h_H} exp(m_Hh_H)} \\
&= \Pi_i \frac{exp(m_ih_i)}{\sum_{h_i} exp(m_ih_i)} \\
&= \Pi_i \frac{exp(m_ih_i)}{exp(m_ih_i = 0) + exp(m_ih_i = 1)} \\
&= \Pi_i \frac{exp(m_ih_i)}{1 + exp(m_ih_i = 1)} \\
&= \Pi_i p(h_i|\boldsymbol{v})
\end{aligned}
$$

b) Assuming that the restricted Boltzmann machine consists of only 2 visible and 1 hidden variables, and the joint distribution of the MRF is given by:

h	v_1	v_2	$\exp(\boldsymbol{v}^T \boldsymbol{W} h + \boldsymbol{a}^T \boldsymbol{v} + bh)$
0	0	0	1.00
0	0	1	2.13
0	1	0	4.65
0	1	1	9.90
1	0	0	3.65
1	0	1	8.66
1	1	0	4.22
1	1	1	10.01

Find the unknown parameters, i.e. $\theta = \{\boldsymbol{W}_{2\times 1}, \boldsymbol{a}_{2\times 1}, b\}$.

Answer:

$$exp\{[v_1\ v_2]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} h + [a_1\ a_2]\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + bh\}$$

Case 1: $h = 0, v_1 = 0, v_2 = 1$

$$exp\{[0\ 1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} 0 + [a_1\ a_2]\begin{bmatrix} 1 \\ 0 \end{bmatrix} + b\ 0\} = 2.13$$

$$\Rightarrow exp(a_2) = 2.13 \Rightarrow a_2 = 0.756$$

Case 2: $h = 0, v_1 = 1, v_2 = 0$

$$exp\{[0\ 1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} 0 + [a_1\ a_2]\begin{bmatrix} 0 \\ 1 \end{bmatrix} + b\ 0\} = 4.65$$

$$\Rightarrow exp(a_1) = 4.65 \Rightarrow a_1 = 1.537$$

Case 3: $h = 1, v_1 = 0, v_2 = 0$

$$exp\{[0\ 0]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} 1 + [a_1\ a_2]\begin{bmatrix} 0 \\ 0 \end{bmatrix} + b\ 1\} = 3.65$$

$$\Rightarrow exp(b) = 3.65 \Rightarrow b = 1.2947$$

Case 4: $h = 1, v_1 = 0, v_2 = 1$

$$exp\{[0\ 1]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} 1 + [a_1\ a_2]\begin{bmatrix} 0 \\ 1 \end{bmatrix} + b\ 1\} = 8.66$$

$$\Rightarrow exp(w_2 + a_2 + b) = 8.66 \Rightarrow exp(w_2 + 0.756 + 1.2947) = 8.66$$

$$\Rightarrow w_2 + 2.0507 = 2.1587 \Rightarrow w_2 = 0.1080$$

Case 5: $h = 1, v_1 = 1, v_2 = 0$

$exp\{[1\ 0]\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} 1 + [a_1\ a_2]\begin{bmatrix} 1 \\ 0 \end{bmatrix} + b\ 1\} = 4.22$

$\Rightarrow exp(w_1 + a_1 + b) = 4.22 \Rightarrow exp(w_1 + 1.537 + 1.2947) = 4.22$

$\Rightarrow w_1 + 2.8317 = 1.4398 \Rightarrow w_1 = -1.3919$

Verifications:

Case 1: $h = 0, v_1 = 0, v_2 = 0 \Rightarrow exp(0) = 1.00$

Case 2: $h = 0, v_1 = 1, v_2 = 1 \Rightarrow exp(a_1 + a_2) = exp(1.537 + 0.756) = 9.90$

Case 3: $h = 1, v_1 = 1, v_2 = 1$

$\Rightarrow exp(v_1 W_1 h + v_2 W_2 h + a_1 v_1 + a_2 v_2 + bh) = exp(w_1 + w_2 + a_1 + a_2 + b)$

$= exp(-1.3919 + 0.1080 + 1.537 + 0.756 + 1.2947) = 10.0122$

Question 2

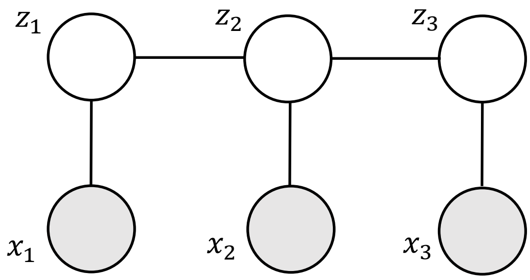


Fig. 2.1

Fig. 4.1 shows a Markov Random Field (MRF) representation of a Hidden Markov Model (HMM) over three time steps. The hidden variables z_1, z_2, z_3 are discrete random variables that take three possible states $z_n \in \{F, H, M\}$, and x_1, x_2, x_3 are the observed variables that take on real values $x_n \in \mathbb{R}$. The joint distribution is given by:

$$p(z_1, z_2, z_3, x_1, x_2, x_3) = \frac{1}{Z} \prod_{n=2}^{3} \psi_t(z_n, z_{n-1}) \prod_{n=1}^{3} \psi_e(x_n, z_n),$$

where Z is the partition function, and the transition potential $\psi_t(z_n, z_{n-1})$ and the emission potentials $\psi_e(x_n, z_n)$ are given by:

$\boldsymbol{\psi_t(z_n, z_{n-1})}$	$z_n = F$	$z_n = H$	$z_n = M$
$z_{n-1} = F$	2.0	3.0	5.0
$z_{n-1} = H$	1.0	6.0	3.0
$z_{n-1} = M$	4.5	2.0	2.5

z_1	$\psi_e(x_1, z_1)$
F	1.0
H	8.0
M	1.0

z_2	$\psi_e(x_2, z_2)$
F	7.0
H	1.0
M	2.0

z_3	$\psi_e(x_3, z_3)$
F	2.0
H	3.0
M	5.0

Decode the message that corresponds to the states of the hidden variables that give the maximal probability. Show all your workings clearly.

Answer:

The solution can be evaluated as:

$$\max_{z_1,z_2,z_3} \psi(z_3,x_3)\psi(z_2,z_3)\psi(z_2,x_2)\,\psi(z_1,z_2)\psi(z_1,x_1) =$$

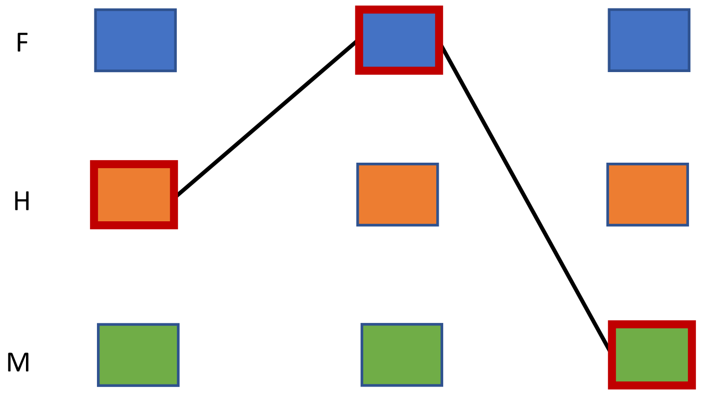
$$\max_{z_3} \psi(z_3,x_3) \max_{z_2} \psi(z_2,z_3)\psi(z_2,x_2) \max_{z_1} \psi(z_1,z_2)\psi(z_1,x_1)$$

Z_2	$\max_{z_1} \psi(z_1,z_2)\psi(z_1,x_1) = z_2^{max}(z_1)$	$\delta^{max}(z_1)$
F	$\max(2.0\times1.0, 1.0\times8.0, 4.5\times1.0) = \max(2.0,8.0,4.5) = 8.0$	H
H	$\max(3.0\times1.0, 6.0\times8.0, 2.0\times1.0) = \max(3.0,48.0,2.0) = 48.0$	H
M	$\max(5.0\times1.0, 3.0\times8.0, 2.5\times1.0) = \max(5.0,24.0,2.5) = 24.0$	H

Z_3	$\max_{z_2} \psi(z_2,z_3)\psi(z_2,x_2)\, z_2^{max}(z_1) = z_3^{max}(z_2)$	$\delta^{max}(z_2)$
F	$\max(2.0\times7.0\times8.0, 1.0\times1.0\times48.0, 4.5\times2.0\times24.0)$ $= \max(112.0,48.0,216.0) = 216.0$	M
H	$\max(3.0\times7.0\times8.0, 6.0\times1.0\times48.0, 2.0\times2.0\times24.0)$ $= \max(168.0,288.0,96.0) = 288.0$	H
M	$\max(5.0\times7.0\times8.0, 3.0\times1.0\times48.0, 2.5\times2.0\times24.0)$ $= \max(280.0,144.0,120.0) = 280.0$	F

$\max_{z_3} \psi(z_3,x_3)\, z_3^{max}(z_2)$	$\delta^{max}(z_3)$
$\max(216\times2.0, 288.0\times3.0, 280\times5.0)$ $= \max(432.0,864.0,1400.0) = 1400.0$	M

Backtracking:



The code is: HFM

Question 3

Fig. 3.1 shows a Bayesian network of the mixture of Bernoulli Distribution. X_n is a binary random variable, i.e. $x_n \in \{0,1\}$. N is the total number of observations. Z_n is the 1-of-k indicator random variable, $z_{nk} = 1 \Rightarrow z_{n,j\neq k} = 0$ indicates the assignment of the random variable x to the k^{th} Bernoulli density. $z_{nk} \in \{0,1\}$ and $\sum_k z_{nk} = 1$.

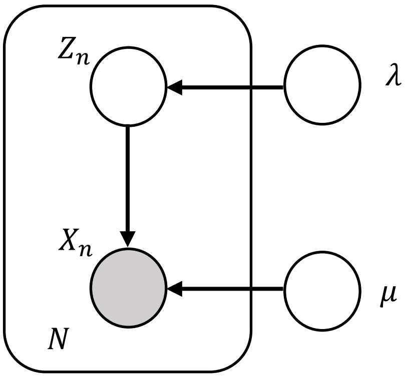


Fig. 3.1

Given the expressions for the Bernoulli distribution:

$$p(\mathrm{x} \mid \mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{(1-x_n)} ,$$

and marginal distribution of Z_n, which is a categorical distribution specified in terms of the mixing coefficients λ_k:

$$p(\mathrm{z_n}) = \prod_{k=1}^{K} \lambda_k^{z_{nk}} = \mathrm{cat}_{\mathrm{z_n}}[\lambda] \text{ , where } 0 \leq \lambda_k \leq 1 \text{ and } \sum_k \lambda_k = 1.$$

(a) Show that the mixture of Bernoulli distribution is given by:

$$p(\mathrm{x} \mid \mu, \lambda) = \prod_{n=1}^{N} \sum_{k=1}^{K} \lambda_k \mu_k^{x_n}(1-\mu_k)^{(1-x_n)}.$$

(b) Derive the responsibility $\gamma(z_{nk}) = p(z_{nk} = 1 \mid \mathrm{x})$, and show that the updates for the unknown parameters μ and λ in the maximization step of the EM algorithm are given by:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n,$$

$$\lambda_k = \frac{N_k}{N} \text{ , where } N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

Show all your workings clearly.

Answer:

Refer to Section 9.3.3 in "Pattern Recognition and Machine Learning", Christopher Bishop.

Question 4

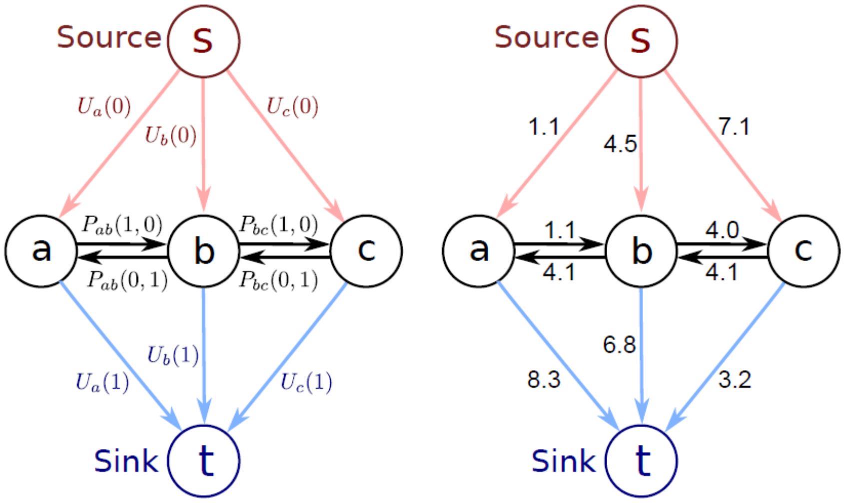


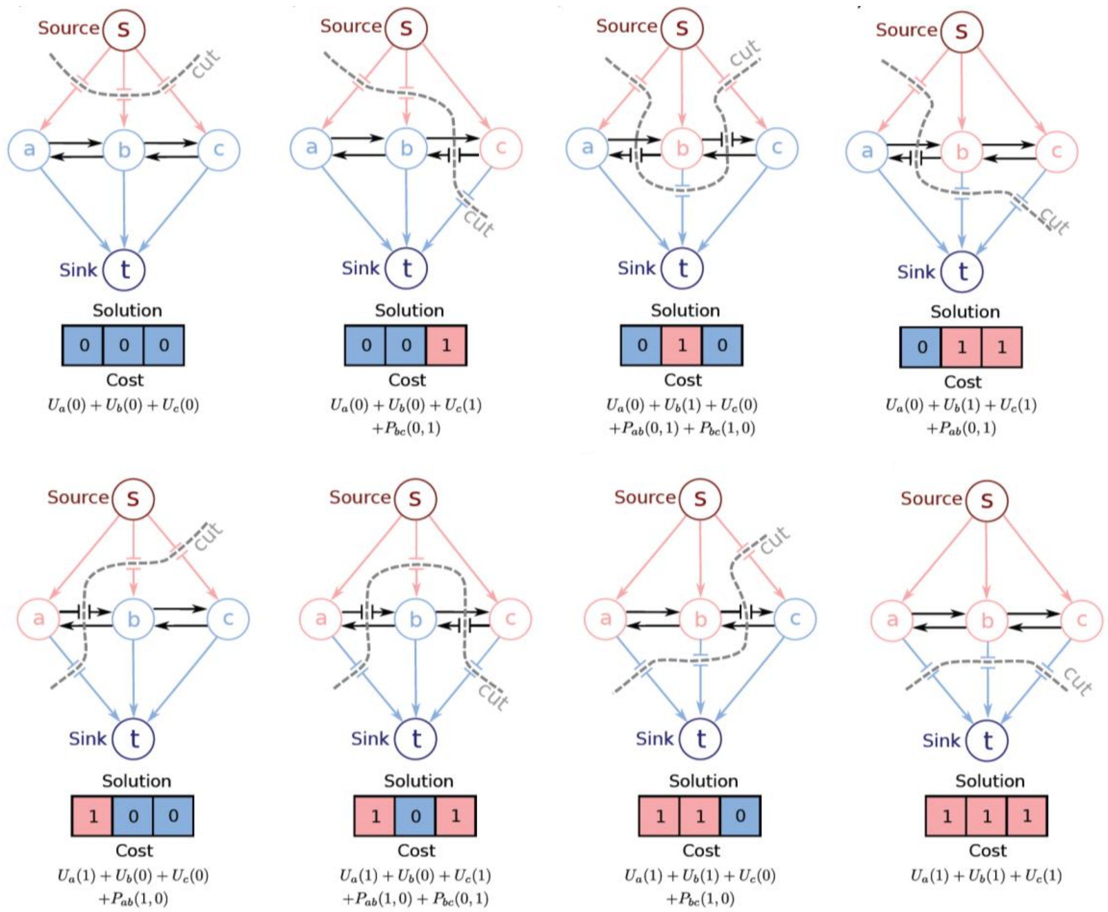
Fig 4.1

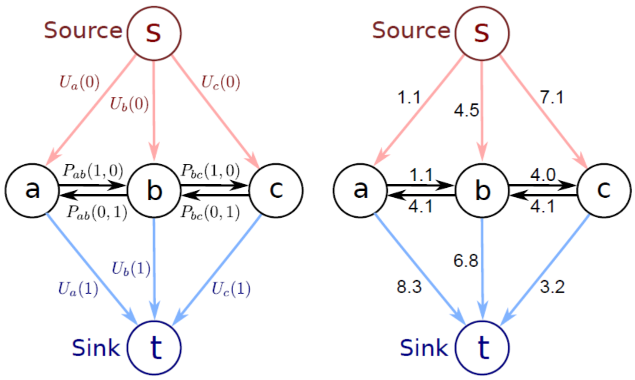
(Image source: “Computer Vision: Models, Learning and Inference”, Simon Prince)

Compute the **MAP solution** to the three-pixel graph cut problem in Fig. 4.1 by

(i) computing the cost of all eight possible solutions explicitly and finding the one with the minimum cost, and

Answer:



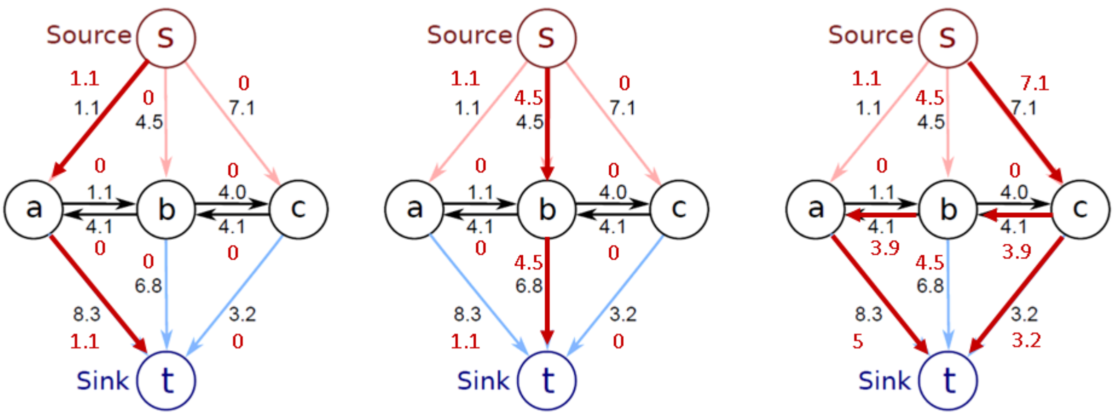


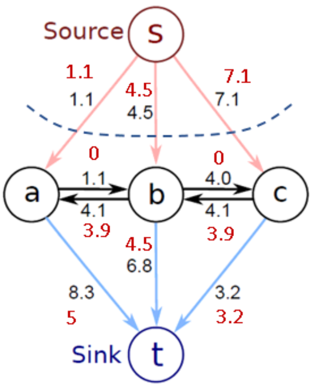
$\boldsymbol{U_a(0) + U_b(0) + U_c(0) = 1.1 + 4.5 + 7.1 = 12.7}$
$U_a(0) + U_b(0) + U_c(1) + P_{bc}(0{,}1) = 1.1 + 4.5 + 3.2 + 4.1 = 12.9$
$U_a(0) + U_b(1) + U_c(0) + P_{ab}(0{,}1) + P_{bc}(1{,}0) = 1.1 + 6.8 + 7.1 + 4.1 + 4.0 = 23.1$
$U_a(0) + U_b(1) + U_c(1) + P_{ab}(0{,}1) = 1.1 + 6.8 + 3.2 + 4.1 = 15.2$

$U_a(1) + U_b(0) + U_c(0) + P_{ab}(1{,}0) = 8.3 + 4.5 + 7.1 + 1.1 = 21$
$U_a(1) + U_b(0) + U_c(1) + P_{ab}(1{,}0) + P_{bc}(0{,}1) = 8.3 + 4.5 + 3.2 + 1.1 + 4.1 = 21.2$
$U_a(1) + U_b(1) + U_c(0) + P_{bc}(1{,}0) = 8.3 + 6.8 + 7.1 + 4.0 = 26.2$
$U_a(1) + U_b(1) + U_c(1) = 8.3 + 6.8 + 3.2 = 18.3$

(ii) running the augmenting paths algorithm on this graph by hand and interpreting the minimum cut.

Answer:





Question 5

Consider the simple 3-node graph shown in Fig. 5.1 in which the observed node X is given by a Gaussian distribution $\mathcal{N}(x|\mu,\tau^{-1})$ with mean μ and precision τ. Suppose that the marginal distributions over the mean and precision are given by $\mathcal{N}(\mu|\mu_0,s_0)$ and $\mathrm{Gam}(\tau|a,b)$, where $\mathrm{Gam}(.|.,.)$ denotes a gamma distribution. Write down expressions for the conditions distributions for the conditional distributions $p(\mu|x,\tau)$ and $p(\tau|x,\mu)$ that would be required to apply Gibbs sampling to the posterior distribution $p(\mu,\tau \mid x)$.

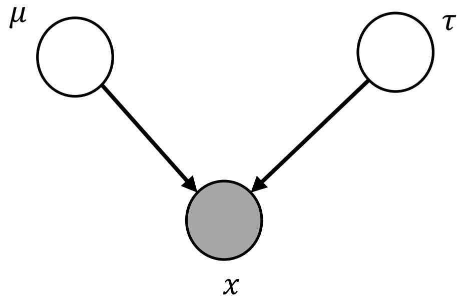


Fig. 5.1

Answer:

$$p(\mu|x,\tau) = \frac{p(\mu,x,\tau)}{\int p(\mu,x,\tau)d\mu} = \frac{p(\mu)p(\tau)p(x|\mu,\tau)}{\int p(\mu)p(\tau)p(x|\mu,\tau)d\mu} = \frac{p(\mu)p(x|\mu,\tau)}{\int p(\mu)p(x|\mu,\tau)d\mu}$$

$$p(x \mid \mu,\tau) = C_x \exp\{-0.5\tau(x-\mu)^2\}$$
$$p(\mu \mid \mu_0,s_0) = C_\mu \exp\{-0.5 s_0(\mu_0-\mu)^2\}$$
$$p(\mu)p(x|\mu,\tau) = C_x C_\mu \exp\{-0.5\,[\mu^2(\tau+\tau_0) - 2\mu(\tau x - \tau_0\mu_0) + (\tau x^2 + \tau_0\mu_0^2)]\}$$

$$p(\mu|x,\tau) = \frac{p(\mu)p(x|\mu,\tau)}{\int p(\mu)p(x|\mu,\tau)d\mu}$$

$$= \frac{\exp\{-0.5\,[\mu^2(\tau+\tau_0)-2\mu(\tau x-\tau_0\mu_0)+(\tau x^2+\tau_0\mu_0^2)]\}}{\int \exp\{-0.5\,[\mu^2(\tau+\tau_0)-2\mu(\tau x-\tau_0\mu_0)+(\tau x^2+\tau_0\mu_0^2)]\}d\mu}$$

$$= \frac{\exp\{-0.5\,[\mu^2(\tau+\tau_0)-2\mu(\tau x-\tau_0\mu_0)]\}}{\int \exp\{-0.5\,[\mu^2(\tau+\tau_0)-2\mu(\tau x-\tau_0\mu_0)]\}d\mu}$$

$$= \frac{\exp\{-\alpha\mu^2+\beta\mu\}}{\int \exp\{-\alpha\mu^2+\beta\mu\}d\mu}, \quad \text{where} \quad \alpha = 0.5(\tau+\tau_0) \text{ and } \beta = \tau x - \tau_0\mu_0\ .$$

Since $\int_{-\infty}^{+\infty} \exp\{-\alpha x^2 + \beta x\}\,dx = \sqrt{\frac{\pi}{\alpha}} \exp\left\{\frac{\beta^2}{4\alpha}\right\}$,

$$p(\mu|x,\tau) = \frac{\exp\{-\alpha\mu^2 + \beta\mu\}}{\sqrt{\frac{\pi}{\alpha}}\exp\left\{\frac{\beta^2}{4\alpha}\right\}}$$

$$p(\tau|x,\mu) = \frac{p(\mu,x,\tau)}{\int p(\mu,x,\tau)d\tau} = \frac{p(\mu)p(\tau)p(x|\mu,\tau)}{\int p(\mu)p(\tau)p(x|\mu,\tau)d\tau} = \frac{p(\tau)p(x|\mu,\tau)}{\int p(\tau)p(x|\mu,\tau)d\tau}$$

$p(x \mid \mu,\tau) = C_x \exp\{-0.5\tau(x-\mu)^2\}$
$p(\tau \mid a,b) = C_\tau \tau^{a_0-1}\, exp(-b_0\tau)$
$p(\tau)p(x|\mu,\tau) = C_x C_\tau \tau^{a_0-1} \exp\{\tau[-0.5(x-\mu)^2 - b_0]\}$

$$p(\tau|x,\mu) = \frac{p(\tau)p(x|\mu,\tau)}{\int p(\tau)p(x|\mu,\tau)d\tau} = \frac{\tau^{a_0-1}\exp\{\tau[-0.5(x-\mu)^2 - b_0]\}}{\int \tau^{a_0-1}\exp\{\tau[-0.5(x-\mu)^2 - b_0]\}\,d\tau}$$

$$= \frac{\tau^n \exp\{-\alpha\tau\}}{\int \tau^n \exp\{-\alpha\tau\}d\tau}, \quad \text{where } n = a_0 - 1 \text{ and } \alpha = 0.5(x-\mu)^2 + b_0.$$

Since $\int_0^\infty x^n \exp\{-\alpha x\}\,dx = \begin{cases} \frac{\Gamma(n+1)}{\alpha^{n+1}}, & (n > -1, \alpha > 0) \\ \frac{n!}{a^{n+1}}, & (n = 0,1,2,\dots,\alpha > 0) \end{cases}$,

$$p(\tau|x,\mu) = \frac{\tau^n \exp\{-\alpha\tau\}}{\frac{\Gamma(n+1)}{\alpha^{n+1}}}, \quad \text{since } a_0 > 0.$$

--End--