School of Computing
National University of Singapore
CS5340: Uncertainty Modeling in AI
Semester 1, AY 2020/21

Exercise 2

Question 1

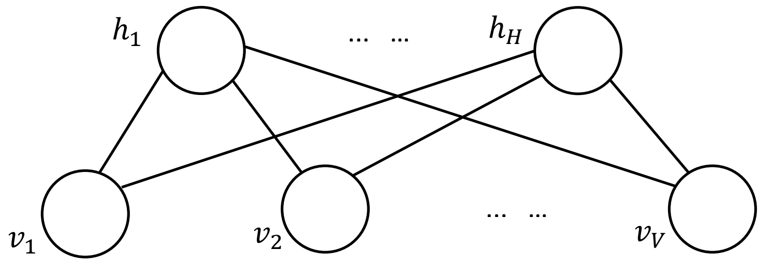


Fig. 1.1

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 $\boldsymbol{v} = [v_1, \ldots, v_V]^T$ and hidden variables $\boldsymbol{h} = [h_1, \ldots, h_H]^T$, 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:

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

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.

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. $\theta = \{\boldsymbol{W}_{2\times1}, \boldsymbol{a}_{2\times1}, b\}$.

Question 2

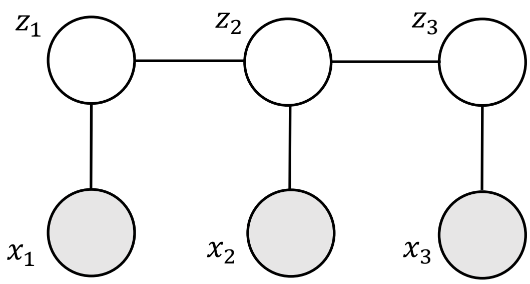


Fig. 2.1

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.

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 $\sum_k z_{nk} = 1$.

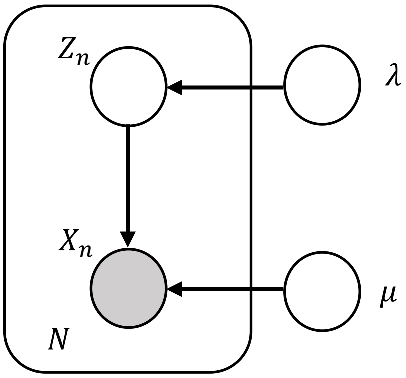


Fig. 3.1

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:

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

(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.

Question 4

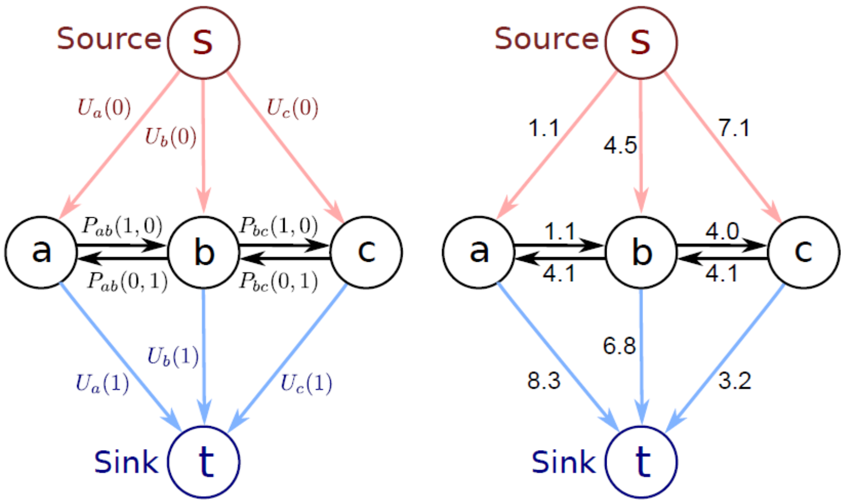


Fig 4.1
(Image source: "Computer Vision: Models, Learning and Inference", Simon Prince)

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

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

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 in order to apply Gibbs sampling to the posterior distribution $p(\mu, \tau \mid x)$.

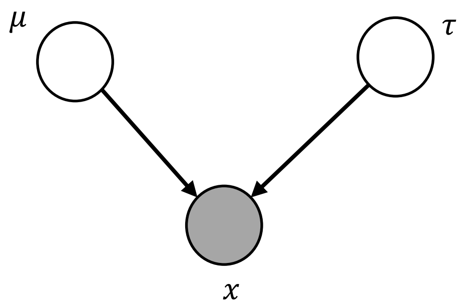


Fig. 5.1

Question 6

Figure 6.1 shows a Markov Random Field (MRF) with two random variables X_1 and X_2, where $x_i \in \{0,1\}$. Furthermore, let $\phi_1(x_1)$ and $\phi_2(x_2)$ denote the unary potentials, and $\psi_{12}(x_1, x_2)$ denotes the pairwise potential. Given the observations over 14 trials as shown in Table 6.1, find the unknown value of $\psi_{12}(x_1 = 0, x_2 = 0)$ in the potential tables shown in Table 6.2. Show all your workings clearly.

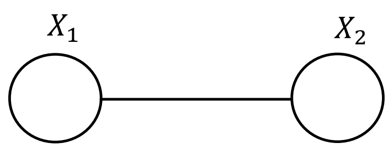


Figure 6.1

Trial Number	Outcomes	
	X_1	X_2
1	0	0
2	1	0
3	1	1
4	1	0
5	0	0
6	0	1
7	1	1
8	0	0
9	1	0
10	1	1
11	0	0
12	0	0
13	1	0
14	1	1

Table 6.1

X_1	$\phi_1(x_1)$
0	2
1	1

X_2	$\phi_2(x_2)$
0	1
1	2

X_1	X_2	$\psi_{12}(x_1, x_2)$
0	0	$\psi_{12}(x_1 = 0, x_2 = 0)$
0	1	1
1	0	2
1	1	2

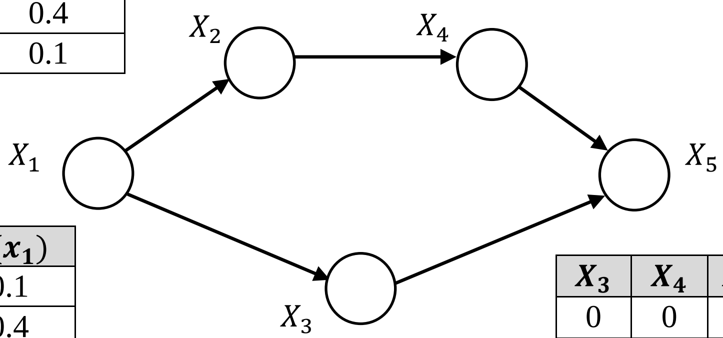
Table 6.2

Question 7

The Bayesian network shown in Figure 7.1 has five random variables X_1, X_2, X_3, X_4, X_5, where $x_i \in \{0,1,2\}$ for $i = 1, 2$ and $x_i \in \{0,1\}$ for $i = 3, 4, 5$.

X_1	X_2	$p(x_2\|x_1)$
0	0	0.1
1	0	0.4
2	0	0.8
0	1	0.2
1	1	0.4
2	1	0.1

X_2	X_4	$p(x_4\|x_2)$
0	0	0.2
1	0	0.5
2	0	0.7



X_1	$p(x_1)$
0	0.1
1	0.4

X_3	X_4	X_5	$p(x_5\|x_3, x_4)$
0	0	0	0.6
0	1	0	0.3
1	0	0	0.1
1	1	0	0.8

X_1	X_3	$p(x_3\|x_1)$
0	0	0.3
1	0	0.4
2	0	0.1

Figure 7.1

(a) Given the following numbers drawn from a uniform distribution $u \sim \mathcal{U}(0,1)$:

$$u = [\ 0.4387 \quad 0.4898 \quad 0.7513 \quad 0.4984 \quad 0.2760\ ],$$

generate one set of samples from the joint distribution $p(x_1, x_2, x_3, x_4, x_5)$ using Gibbs sampling. Use $x_1 = 0$, $x_2 = 0$, $x_3 = 0$, $x_4 = 0$, $x_5 = 0$ as the initialization. Show all your workings clearly.

(b) Table 7.1 shows 10 sets of samples drawn from Gibbs sampling. Ignoring the burn-in effect and initialization, find the approximation for the following probabilities using the generated samples:

i. $p(x_2)$
ii. $p(x_3, x_5)$
iii. $p(x_3, x_4 = 1, x_5 = 1)$
iv. $p(x_3|x_2 = 1)$

Sample #	X_1	X_2	X_3	X_4	X_5
0	0	0	0	0	0
1	2	0	1	1	0
2	2	0	0	1	0
3	0	0	0	1	1
4	1	1	1	0	0
5	2	2	1	1	0
6	2	0	1	0	1
7	1	2	0	0	0
8	2	1	0	0	0
9	1	0	1	1	0
10	1	0	1	1	1

Table 7.1

Question 8

a. Figure 8.1 shows a homogeneous hidden Markov Model (HMM) over three time steps. The latent random variables are Y_1, Y_2, Y_3, where $Y_n \in \{0, 1, 2\}$, and the observed random variables are X_1, X_2, X_3, where $X_n \in \mathbb{R}$.

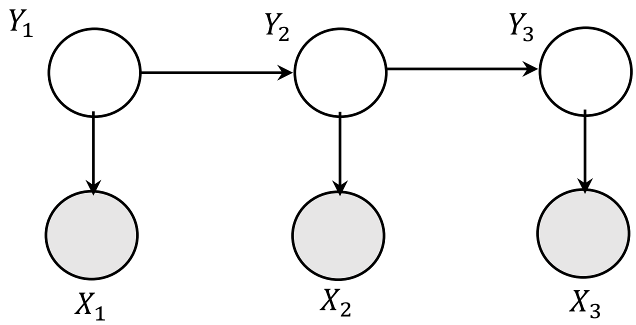


Figure 2.1

The prior probability of the random variable Y_1 is $p(Y_1 \mid \pi) = \prod_k \pi_k^{y_{1k}}$, where $\pi = \{0.2, 0.5, 0.3\}$. Furthermore, the transition probability is given by:

$$p(Y_n \mid Y_{n-1}, A) = \prod_k \prod_j A_{jk}^{y_{n-1,j}y_{nk}} \text{, where } A = \begin{bmatrix} 0.2 & \alpha & \beta \\ 0.1 & 0.6 & 0.3 \\ 0.4 & 0.5 & 0.1 \end{bmatrix}, \text{ and}$$

the emission probabilities of the respective observed random variables X_n are shown in Table 8.1.

	$k = 0$	$k = 1$	$k = 2$
X_1	0.3	0.6	0.4
X_2	0.5	0.4	0.4
X_3	0.3	0.8	0.5

Table 8.1

Given that the minimum probability of the joint distribution $p(Y_1, Y_2, Y_3, X_1, X_2, X_3)$ is 0.000216 and occurs at $Y_1 = 0, Y_2 = 1, Y_3 = 0$, find the unknown values α and β in the transition probability.

b. Figure 8.2 shows an undirected graphic model with six random variables X_1, X_2, X_3, X_4, X_5 and X_6, where $X_i \in \{0,1,2\}$. The potential $\psi(X_i, X_j)$ between any pair of nodes X_i and X_j, where $i < j$ is given in Table 2.2. Given $X_1 = 0, X_3 = 1$ and $X_5 = 2$, find the states of X_2, X_4 and X_6 that maximizes the joint distribution $p(X_1, X_2, X_3, X_4, X_5, X_6)$.

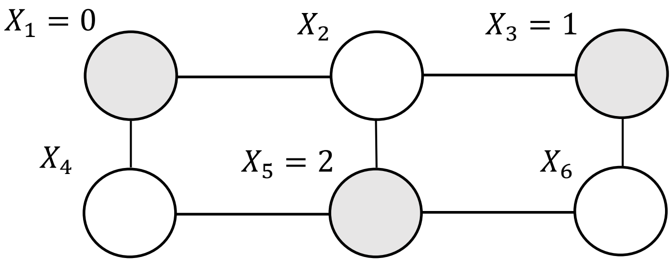


Figure 8.2

X_i	X_j	$\psi(X_i, X_j)$
0	0	1
0	1	5
0	2	7
1	0	2
1	1	4
1	2	8
2	0	3
2	1	6
2	2	9

Table 8.2

Question 9

Figure 9.1 shows a Bayesian network with both binary and continuous state latent random variables, i.e., $Z \in \{0,1\}$ and $T \in \mathbb{R}$. In addition, $X = 0.5$ is the observed random variable. The maximum log-likelihood of T:

$$\underset{T}{\operatorname{argmax}} \log p(T \mid X),$$

can be obtained from the Expectation-Maximization (EM) algorithm. The EM algorithm iterates between the Expectation step that evaluates the expected complete data log-likelihood with respect to $p(Z \mid X, T^{old})$ and the Maximization step that maximizes T over the expected complete data log-likelihood with respect to $p(Z \mid X, T^{old})$. T^{old} is the value of T from the previous iteration of the EM algorithm. $\{\lambda = 0.1, w_{a0} = 0.5, w_{a1} = 0.5, w_{b0} = 0.8, w_{b1} = 0.2, \tau_a = 1.0, \tau_b = 1.2, U = 0.6\}$ are known hyperparameters of the following distributions:

$$p(Z) = \lambda^Z (1-\lambda)^{(1-Z)},$$

$$p(X \mid T, Z) = \mathcal{N}(X \mid w_{a0} + w_{a1}T, \tau_a)^Z \mathcal{N}(X \mid w_{b0} + w_{b1}T, \tau_b)^{(1-Z)},$$

$$\mathcal{N}(X \mid w_0 + w_1 T, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\{-0.5\tau(X - w_0 - w_1 T)^2\},$$

$$p(T) = U.$$

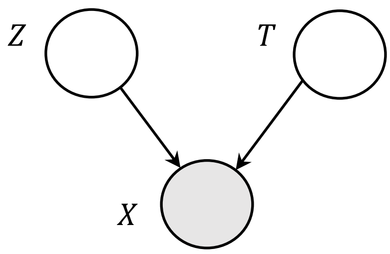


Figure 3.1

a. Derive the expression for the posterior $p(Z \mid X, T^{old})$ from the Bayesian Network.

b. Derive the expression for T that maximizes the expected complete data log-likelihood with respect to $p(Z \mid X, T^{old})$.

c. Given the initial value of $T = 2.0$, find the the value of T in the next EM iteration.

Question 10

a. The objective of image denoising is to recover the clean image (noise-free) from a given noisy image. Figure 10.1 shows a Markov Random Field (MRF) to solve a four-pixel binary image denoising problem. The latent random variable $X_i \in \{-1, +1\}$ represents the pixels of the desired clean image, and the observed random variables $Y_i \in \{-1, +1\}$ represents the pixels of the noisy image. We use the Ising model, i.e., $\psi(X_i, X_j) = \exp(JX_iX_j)$ as the edge potentials, where J is the coupling strength of the smoothness prior between neighboring pixels X_i and X_j. The observation model follows a Gaussian distribution: $p(Y_i \mid X_i) = \mathcal{N}(Y_i \mid X_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-0.5 \frac{(Y_i - X_i)^2}{\sigma^2}\right\}$.

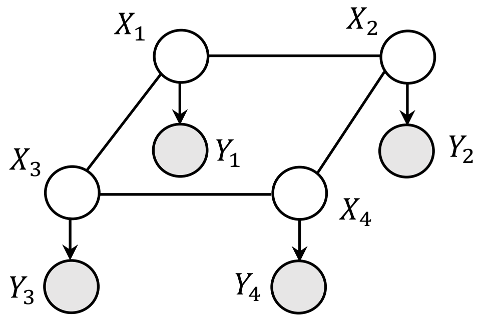


Figure 10.1

Given that we observe $Y_1 = +1, Y_2 = -1, Y_3 = -1, Y_4 = +1$, and the following random numbers are drawn from a uniform distribution $u \sim \mathcal{U}(0,1)$:

$$u = [\ 0.6557 \quad 0.0357 \quad 0.9340 \quad 0.8491\ ],$$

generate one set of samples from the joint distribution $p(X, Y)$ using Gibbs sampling. Use $X_1 = -1, X_2 = -1, X_3 = -1, X_4 = -1$ as the initialization and set $J = 0.01$, $\sigma^2 = 1.0$. Show all workings clearly.

b. Draw the Bayesian Network and write down the factorized joint probability distribution that encodes all the following conditional independences:

1. $X_4 \perp \{X_1, X_2, X_5\} \mid X_3$
2. $X_5 \perp \{X_1, X_3, X_4\} \mid X_2$
3. $X_3 \perp X_5 \mid \{X_1, X_2\}$
4. $X_1 \perp X_2 \mid \emptyset$
5. $X_1 \perp \{X_2, X_5\} \mid \emptyset$

--End--