COMP0080PastExams

1 2022 Exam

1.1 Independence in Graphical Models

(a) **Question** Suppose that X_1 and X_2 are two random variables, where each can take one of three states. Here is a probability table P describing their joint distribution, $P_{ij} = p(X_1 = i, X_2 = j)$:

$$\begin{pmatrix} 0.06 & 0.37 & 0.12 \\ 0.01 & 0.07 & 0.02 \\ 0.03 & 0.26 & 0.06 \end{pmatrix}$$

Are X_1 and X_2 independent? Support your answer and if they are not independent, what is the minimal number of entries in the table that you need to change to make them independent?

Answer:

A method of solving this would be to look at the columns: every column should be a multiple of another column. We see that the first and third column are multiples of each other: their proportion is $1:2$ (just like their marginals are, which are 0.1 and 0.2. We notice that for the first and second column, their marginals are 0.1 and 0.7. But the entries in the columns don't have the same ratio. So in order for them to have the same proportion, we change the entries from $0.37 \to 0.42$ and $0.26 \to 21$. Hence, a minimum of 2 entries would have to be changed to make X_1 and X_2 independent.

(b) **Question** Consider three random variables, A, B, C. Give an example of a distribution in which $A \perp\!\!\!\perp B, B \perp\!\!\!\perp C, C \perp\!\!\!\perp A$ but A and B are not conditionally independent when given C.

Answer:

Let $A = \{\text{flipping coin A}\}$ and $B = \{\text{flipping coin B}\}$ and $C = \begin{cases} 1 & \text{both heads or both tails} \\ 0 & \text{otherwise} \end{cases}$. We have that $P(A = h) = P(A = t) = P(B = h) = P(B = t) = \frac{1}{2}$.

$$\begin{aligned} P(C = 1|A = h) = P(C = 1|B = H) &= \sum_B P(C = 1|A = h, B)P(A = h)P(B) \\ &= P(C = 1|A = h, B = h)\frac{1}{2}\frac{1}{2} + P(C = 1|A = h, B = t)\frac{1}{2}\frac{1}{2} \\ &= (1)\frac{1}{4} + (0)\frac{1}{4} = \frac{1}{4} \end{aligned}$$

We also have that

$$\begin{aligned} P(C = 1) &= \sum_{A,B} P(C = 1|A, B)P(A)P(B) \\ = P(C = 1|A = h, B = h)P(A = h)P(B = h) + P(C = 1|A = h, B = t)P(A = h)P(B = t) + \\ +P(C = 1|A = t, B = h)P(A = t)P(B = h) + P(C = 1|A = t, B = t)P(A = t)P(B = t) \\ &= (1)\frac{1}{4} + (0)\frac{1}{4} + (0)\frac{1}{4} + (1)\frac{1}{4} \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \end{aligned}$$

So we have that $P(C=1)P(A=h) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} = P(C=1|A=h)$

(c) **Question** For the distribution from the previous question, draw the graphical model with which this distribution is compatible and which captures as many independence statements from the previous question as possible

Answer:

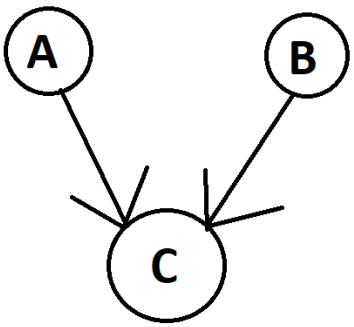

Figure 1

(d) **Question** Consider the following Markov Network on the variables X_i:

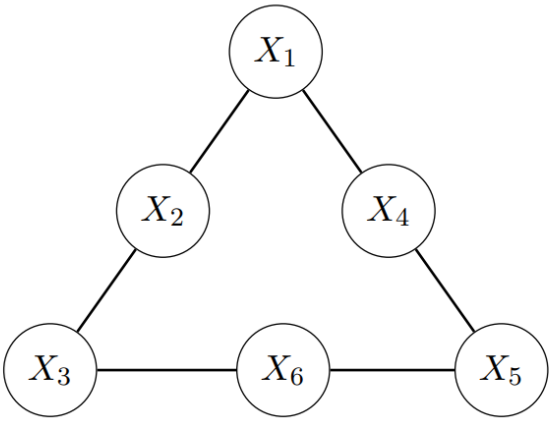

Figure 2

Answer:

We have that

$P(X_1, X_2, X_3, X_4, X_5, X_6) = \frac{1}{Z}\phi(X_1, X_2)\phi(X_1, X_4)\phi(X_2, X_3)\phi(X_3, X_6)\phi(X_5, X_6)$

And so $P(X_2, X_4, X_6) = \sum_{X_1, X_3, X_5} P(X_1, X_2, X_3, X_4, X_5, X_6) = \phi(X_2, X_4)\phi(X_4, X_6)\phi(X_2, X_6) = \phi(X_2, X_4, X_6)$

So we have the following MN:

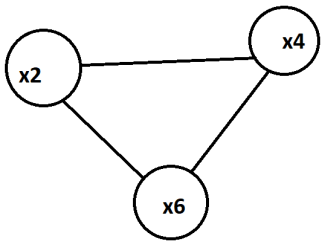

Figure 3

(NOTE: Graphical dependence does **NOT** imply distributional independence ans so there might be independence we can't observe)

We can clearly see that all variables are marginally dependent. Moreover, any variable is conditionally dependent from any other variable given a 3rd variable.

(e) **Question** Consider the following two graphical models:

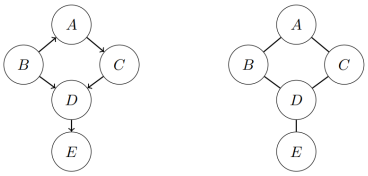

Figure 4

Answer: In order to see which conditional independencies are present in the BN but not in the MN (and vice versa) let's check all of the CI's in both.

First, we will write out some relevant definitions:

Definition of a Collider: Given a path $\mathcal{P}$, a **collider** is a node c on $\mathcal{P}$ with neighbours a and b on $\mathcal{P}$ such that $a \to c \leftarrow b$. **Note** that a collider is path specific!

A Way to determine d-separation (and hence conditional independence): For every variable $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, check every path U between x and y. A path U is said to be **blocked** if there is a node w on U such that **EITHER**

1. w is a collider and neither w nor any of its descendants is in $\mathcal{Z}$, **OR**
2. w is not a collider on U and w is in $\mathcal{Z}$

If all such paths are blocked then $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$. If the variable sets $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$, they are independent conditional on $\mathcal{Z}$.

Now, let's first look at the BN. It's clear that all of the adjacent nodes (basically nodes with a direct link to each other) can't be d-separated, so we look at the other pairs of nodes: $(A, D), (A, E), (B, C), (B, E), (C, E)$.

Let's check for (A, D) the following CI's:

1. $A \perp\!\!\!\perp D | B, C$
2. $A \perp\!\!\!\perp D | B, C, E$

Let's check $A \perp\!\!\!\perp D|B, C$. We have that $A \in \mathcal{X}$, $D \in \mathcal{Y}$ and $\mathcal{Z} = \{B, C\}$. We need to check whether A and B are d-separated by $\mathcal{Z}$.

Now, we have only 2 paths which lead from A to D:

1. $A \leftarrow B \rightarrow D$
2. $A \rightarrow C \rightarrow D$

We need to check whether each path is blocked. For path $A \leftarrow B \rightarrow D$, we check the (1) - we have no colliders in this path. So this already satisfies the definition of a blocked path. If we check (2),then again we see that it's satisfied because each node in this path is a non-colider and B which is in $\mathcal{Z}$ is a non-colider, IN $\mathcal{Z}$. Now we check the path $A \rightarrow C \rightarrow D$. (1) - we have no colliders in this path. —¿ already satisfied "blocked" condition. (2) - we have a non colider $C \in \mathcal{Z}$ so the path is blocked.

Both paths are blocked so this means that A and D are d-separated by B, C, hence $A \perp\!\!\!\perp D|B, C$

Now, is $A \perp\!\!\!\perp D|B, C, E$? We have that E is a collider, but since the paths are the same as above and E is not in them, E has no influence and so $A \perp\!\!\!\perp D|B, C, E$.

We can see that the following are true:

$$(A, D) : (A \perp\!\!\!\perp D \mid B, C),\ (A \perp\!\!\!\perp D \mid B, C, E) \tag{1}$$
$$(A, E) : (A \perp\!\!\!\perp E \mid B, C),\ (A \perp\!\!\!\perp E \mid D, S), \quad \text{with } S \text{ any element of the powerset } \mathcal{P}^{\{B,C\}} \tag{2}$$
$$(B, C) : (B \perp\!\!\!\perp C \mid A), \quad \text{(note the collider (and its child))}, \tag{3}$$
$$(B, E) : (B \perp\!\!\!\perp E \mid D, S), \quad \text{with } S \text{ any element of the powerset } \mathcal{P}^{\{A,C\}} \tag{4}$$
$$(C, E) : (C \perp\!\!\!\perp E \mid D, S), \quad \text{with } S \text{ any element of the powerset } \mathcal{P}^{\{A,B\}} \tag{5}$$

Next, we consider the MN. Here, too, adjacent nodes cannot be separated, so the possibilities are reduced to the same pairs as with the BNs:

$$(A, D) : (A \perp\!\!\!\perp D \mid B, C),\ (A \perp\!\!\!\perp D \mid B, C, E) \tag{6}$$
$$(A, E) : (A \perp\!\!\!\perp E \mid B, C),\ (A \perp\!\!\!\perp E \mid D, S) \quad \text{with } S \text{ any element of the powerset } \mathcal{P}^{\{B,C\}} \tag{7}$$
$$(B, C) : (B \perp\!\!\!\perp C \mid A, D),\ (B \perp\!\!\!\perp C \mid A, D, E) \tag{8}$$
$$(B, E) : (B \perp\!\!\!\perp E \mid D, S), \quad \text{with } S \text{ any element of the powerset } \mathcal{P}^{\{A,C\}} \tag{9}$$
$$(C, E) : (C \perp\!\!\!\perp E \mid D, S), \quad \text{with } S \text{ any element of the powerset } \mathcal{P}^{\{A,B\}} \tag{10}$$

Finally, comparing the two sets above, we find:

$$\text{in BN but not in MN} : (B \perp\!\!\!\perp C \mid A) \tag{11}$$
$$\text{in MN but not in BN} : (B \perp\!\!\!\perp C \mid A, D), \quad (B \perp\!\!\!\perp C \mid A, D, E). \tag{12}$$
$$\tag{13}$$

Check here for determining marginal independence: http://web.mit.edu/jmn/www/6.034/d-separation.pdf. In general, nothing is marginally independent if there is a direct path going from one variable to the other.

2 Inference and Learning

(a) Consider a model of diseases and symptoms. $s_i \in \{0, 1\}$ is a binary random variable indicating whether the patient is showing the i-th symptom and $d_j \in \{0, 1\}$ is a binary random variable indicating whether the patient has j-th disease. A model for this is given by

$$p(\mathbf{s}, \mathbf{d}) = \frac{1}{Z} \exp(\mathbf{s}^T \mathbf{W} \mathbf{d} + \mathbf{a}^T \mathbf{s} + \mathbf{b}^T \mathbf{d}) \tag{14}$$

where Z is the normalization constant and $\mathbf{W}, \mathbf{a}, \mathbf{b}$ are the parameters of the model.
(i) Question Draw a Markov Network for this model.
Answer:
The probability $p(\mathbf{a}, \mathbf{d})$ can be written as follows:

$$p(\mathbf{s}, \mathbf{d}) = \frac{1}{Z} \exp(\mathbf{s}^T \mathbf{W} \mathbf{d} + \mathbf{a}^T \mathbf{s} + \mathbf{b}^T \mathbf{d}) \tag{15}$$

$$= \frac{1}{Z} \prod_{ik} \exp(s_i W_{ik} d_k) \prod_r \exp(a_r s_r) \prod_j \exp(b_j d_j) \tag{16}$$

Thus we obtain the following MN:

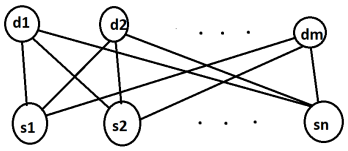

Figure 5

There is an edge between s_i and d_k if and only if $W_{ik} \neq 0$.
(ii) Question Derive the expression for $p(\mathbf{s}|\mathbf{d})$. It this distribution factorised?
Answer:
The fact that $p(\mathbf{s}|\mathbf{d})$ factors in $\mathbf{s}$ factors in $\mathbf{s}$ follows from the MN. Conditioning on $\mathbf{d}$ completely separates this bipartite graph, so all the d_i are independent, i.e., the pdf factorises. We have that

$$p(\mathbf{s}|\mathbf{d}) = \frac{p(\mathbf{s}, \mathbf{d})}{p(\mathbf{d})} \tag{17}$$

$$\propto \exp\left(\mathbf{s}^T \mathbf{W} \mathbf{d} + \mathbf{a}^T \mathbf{s}\right) \tag{18}$$

$$= \exp\left(\mathbf{s}^T (\mathbf{W} \mathbf{d} + \mathbf{a})_i\right) \tag{19}$$

$$= \prod_i \exp\left(s_i (\mathbf{W} \mathbf{d} + \mathbf{a})_i\right) \tag{20}$$

$$= \prod_i f_i(s_i) \tag{21}$$

This confirms that $p(\mathbf{s}|\mathbf{d})$ factorises.
(iii) Question There is a set of N patients, each with a patient record $(\mathbf{s}^n, \mathbf{d}^n)$. Suppose, that you want to learn the parameters of the model by maximum likelihood. Derive the expression for the log-likelihood L, assuming the patient records are i.i.d.
Answer:
Let $\mathcal{N} = \{(\mathbf{s}^n, \mathbf{d}^n, n = 1, ..., N\}$. Then

$$p(\mathcal{N}) = \prod_n p(s^n, d^n)$$

.
We then have that the log-likelihood L is defined as

$$L(\mathbf{W}, \mathbf{a}, \mathbf{b}) = \log \prod_{n=1}^{N} p(\mathbf{s}^n, \mathbf{d}^n | \mathbf{W}, \mathbf{a}, \mathbf{b}) \tag{22}$$

So

$$L(\mathbf{W},\mathbf{a},\mathbf{b}) = \log \prod_{n=1}^{N} \frac{1}{Z(\mathbf{W},\mathbf{a},\mathbf{b})} \exp\left((\mathbf{s}^n)^T \mathbf{W}\mathbf{d}^n + \mathbf{a}^T\mathbf{s}^n + \mathbf{b}^T\mathbf{d}^n\right) \tag{23}$$

$$= -N\log(Z(\mathbf{W},\mathbf{a},\mathbf{b})) + \sum_{n=1}^{N}\left((\mathbf{s}^\mathbf{n})^T \mathbf{W}\mathbf{d}^n + \mathbf{a}^T\mathbf{s}^n + \mathbf{b}^T\mathbf{d}^n\right) \tag{24}$$

(iv) **Question** Derive the expressions $\frac{\partial L}{\partial W_{ij}}$, $\frac{\partial L}{\partial a_i}$, $\frac{\partial L}{\partial b_i}$
Answer:

$$\frac{\partial L}{\partial W_{ij}} = \frac{\partial - N\log Z(\mathbf{W},\mathbf{a},\mathbf{b})}{\partial W_{ij}} + \frac{\partial}{\partial W_{ij}}\sum_{n=1}^{N}\left((\mathbf{s}^\mathbf{n})^T \mathbf{W}\mathbf{d}^n + \mathbf{a}^T\mathbf{s}^n + \mathbf{b}^T\mathbf{d}^n\right) \tag{25}$$

$$= \frac{-N}{Z(\mathbf{W},\mathbf{a},\mathbf{b})}\frac{\partial Z(\mathbf{W},\mathbf{a},\mathbf{b})}{\partial W_{ij}} + \frac{\partial}{\partial W_{ij}}\sum_{n=1}^{N}\sum_{ij} s_i^n W_{ij} d_j^n + ... \tag{26}$$

$$= \frac{-N}{Z(\mathbf{W},\mathbf{a},\mathbf{b})}\frac{\partial Z(\mathbf{W},\mathbf{a},\mathbf{b})}{\partial W_{ij}} + \sum_{n=1}^{N}\sum_{ij} s_i^n d_j^n \tag{27}$$

For $\frac{\partial L}{\partial a_i}$:

$$\frac{\partial L}{\partial a_i} = \frac{-N}{Z(\mathbf{W},\mathbf{a},\mathbf{b})}\frac{\partial Z(\mathbf{W},\mathbf{a},\mathbf{b})}{\partial a_i} + \sum_{n=1}^{N}\sum_{i} s_i^n \tag{28}$$

For $\frac{\partial L}{\partial b_i}$:

$$\frac{\partial L}{\partial b_i} = \frac{-N}{Z(\mathbf{W},\mathbf{a},\mathbf{b})}\frac{\partial Z(\mathbf{W},\mathbf{a},\mathbf{b})}{\partial b_i} + \sum_{n=1}^{N}\sum_{j} d_j^n \tag{29}$$

(v) **Question** Could these derivatives be computed efficiently in the general case? Justify your answer.
Answer:
It is not possible to compute these derivatives efficiently for an arbitrary interaction matrix $\mathbf{W}$ since the derivative of $\log(Z)$ would be computationally intractable and therefore the derivatives need to be approximated.
(b) Consider Hiddern Markov Model with hidden states $h_{1:T} = \{h_1, ..., h_T\}$ and observed states $v_{1:T} = \{v_1, ..., v_T\}$.
(i) **Question** When the sequence of outcomes $v_{1:T}$ is observed, it induces the distribution on the hidden states $p_v(h_{1:T}) = p(h_{1:T}|v_{1:T})$. what is the graphical model of this distribution? Based on the graphical model, is h_1 independent of h_T in this distribution?
Answer:
We have that $p(h_1, ..., h_n|v_{1:n})$ forms a first order Markov chain. The simplest way to show this is to notice that the undirected graph for the hidden Markov model is the same as the DAG but with the arrows removed as there are no colliders in the DAG. Moreover, conditioning corresponds to removing nodes from an undirected graph. This leaves us with a chain that connects the h_i.
By graph separation, we see that $p(h_1, ..., h_n|v_{1:n})$ forms a first-rder Markov chain so that e.g. $h_{1:t-1}$ is independent of $h_{t+1:n}$ given h_t (past independent from the future given the present).

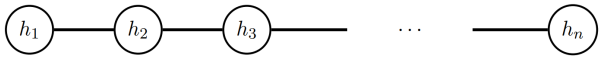

Figure 6

(ii) Question Suppose you want to sample from $p_v(h_{1:T})$. Is it possible to efficiently sample from it? If it is possible, how would you do it? If not, explain why.

Answer:

It is possible to sample it via Ancestral Sampling. To do this, we start with the lowest-numbe rnode and raw a sample from the distribution $p(h_1)$, which we call $\hat{h_1}$. We then work through each of nodes in order, so that for node n we draw a sample from the conditional distribution $p(x_m|pa_n)$ in which the parent variables have been set to their sampled values.

(iii) Question Let $v_t \in \{0,1,2\}, h_t \in \{0,1\}$ and the parameters of the model are as follows. $p(h_1 = 1) = 0.5$, the transition matrix $T = \begin{pmatrix} 0.5 & 0.8 \\ 0.5 & 0.2 \end{pmatrix}$ and the emission matrix $\begin{pmatrix} 0.3 & 0.5 \\ 0.6 & 0 \\ 0.1 & 0.5 \end{pmatrix}$. Suppose that you observe the sequence of outcomes $v_{1:10} = [0,1,0,2,0,2,1,0,2,0]$. Is h_1 independent of h_{10} in $p(h_{1:10}|v_{1:10})$ given this particular sequence?

Answer:

We notice in the emission matrix that $p(v_t = 1|h_t = 1) = 0$, so it is certain that in this particular sequence, $h_2 = 0, h_7 = 0$. Since the distribution $p(h_{1:10}|v_{1:10})$ is a Markov chain, $h_{1:t-1} \perp\!\!\!\perp h_{t+1:n}|h_t$, so we have that in this particular sequence, we $p(h_{1:10}|v_{1:10}) = p(h_{1:10}|v_{1:10}, h_2, h_7)$, making h_1 and h_{10} independent.

(iv) Question Compute the filtering distributions $p(h_8|v_{1:8})$ and $p(h_9|v_{1:9})$ (Hint: you do not need to run the full Forward algorithm here).

Answer:

By the previous question we know that $p(h_8|v_{1:10}, h_7, h_2) = p(h_8|h_7 = 0)$. Hence, by the transition matrix, $p(h_8|h_7 = 0) = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ and for $p(h_9|v_{1:10})$:

$$p(h_9|v_{1:10}) \propto p(v_9|h_9) \sum_{h_8} p(h_9|h_8) p(h_8|v_{1:8}) \tag{30}$$

$$= \begin{pmatrix} p(v_9 = 2|h_9 = 0)\left(\frac{1}{2}p(h_9 = 0|h_8 = 0) + \frac{1}{2}p(h_9 = 0|h_8 = 1)\right) \\ p(v_9 = 2|h_9 = 1)\,)\frac{1}{2}p(h_9 = 1|h_8 = 0) + \frac{1}{2}p(h_9 = 1|h_8 = 1)) \end{pmatrix} \tag{31}$$

$$= \begin{pmatrix} 0.1(\frac{1}{2}\frac{1}{2} + 0.8\frac{1}{2}) \\ \frac{1}{2}(\frac{1}{2}\frac{1}{2} + 0.2\frac{1}{2}) \end{pmatrix} = \begin{pmatrix} 0.065 \\ 0.175 \end{pmatrix} \tag{32}$$

After normalizing it becomes $0.27, 0.73$

3 Variational inference and sampling

(a) Question Explain what KL-divergence is. Suppose $P(X) = (1/4, 1/4, 1/4, 1/4), Q(X) = (1/2, 1/4, 1/8, 1/8), P(Y = i, X = j) = P_{ij}$ and $Q(Y = i|X = j) = Q_{ij}$ where:

$$P = \begin{pmatrix} 0.25 & 0.5 & 0 & 0.2 \\ 0.25 & 0 & 0.5 & 0.3 \\ 0.25 & 0 & 0 & 0.1 \\ 0.25 & 0.5 & 0.5 & 0.4 \end{pmatrix}, Q = \begin{pmatrix} 0.1 & 0 & 1 & \\ 0 & 0 & 0.5 & 0 \\ 0.9 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

Compute $KL(P(Y)\|Q(Y))$ and $KL(Q(Y)\|(P(Y))$.

Answer:

KL Divergence is a type of statistical distance (not a metric): a measure of how one probability distribution p is different from a second, reference probability distribution q. Let $p(x)$ and $q(x)$ be two probability distributions. Then $KL(p\|q) = \sum_x p(x) \ln\left(\frac{p(x)}{q(x)}\right)$

We first compute $P(Y)$ and $Q(Y)$:
$P(Y) = \sum_x p(y|x)p(x)$, hence

$$p(Y=1) = \sum_x p(y=1|x=1)p(x=1) = \frac{1}{4}(0.25+0.5+0+0.2) = 0.2375 \tag{33}$$

$$P(Y=2) = \frac{1}{4}(0.25+0+0.5+0.3) = 0.2625 \tag{34}$$

$$p(Y=3) = \frac{1}{4}(0.25+0+0+0.1) = 0.0875 \tag{35}$$

$$P(Y=4) = \frac{1}{4}(0.25+0.5+0.5+0.4) = 0.4125 \tag{36}$$

Similarly for $Q(Y)$:

$$\begin{pmatrix} 0.175 \\ 0.0625 \\ 0.575 \\ 0.1875 \end{pmatrix}$$

Hence,

$$KL(P(Y)\|Q(Y)) = \sum_y p(y) \ln \frac{p(y)}{q(y)} \tag{37}$$

$$= 0.2375 \ln(0.2375/0.175) + 0.2625 \ln(0.2625/0.0625) + 0.0875 \ln(0.0875/0.575) + 0.4125 \ln(0.4125/0.1 \tag{38}$$

$$= 0.61 \tag{39}$$

Furthermore,

$$KL(Q(Y)\|P(Y)) = \sum_y q(y) \ln \frac{q(y)}{p(y)} \tag{40}$$

$$= 0.175 \ln(0.175/0.2375) + 0.0625 \ln(0.0625/0.2625) + 0.575 \ln(0.575/0.0875) + 0.1875 \log(0.0.1875/0.4 \tag{41}$$

$$= 0.79 \tag{42}$$

(b) Question

Consider Markov Networks on a rectangular M-by-N lattice with each X_i a binary random variable and

$$P(X) \propto \prod_{i \sim j} P_{ij}(X_i, X_j)$$

where each $P_{ij}(X_i, X_j) = ^{-I(X_i \neq X_j)}$ and $\prod_{i \sim j}$ indicates the product over all the edges. Suppose you want to use rejection sampling for trying to get samples from this distribution. Is this a good idea?

Compute the average acceptance rates for the following pairs of M and N

(i) $M = 2, N = 1$

(ii) $M = 2, N = 2$

Answer:

Using rejection sampling for trying to get samples from this distribution is not a good idea because finding an appropriate distribution q is hard when the lattice increases in size. A better way would be to use Gibbs sampling because it exploits the conditional independence structure in Markov Networks and yields proposal distributions which are easy to sample from.

(i) We have that $P(X_1, X_2) = \frac{\exp(-I(X_1 \neq X_2)}{2(1+e^{-1}}$. We have that $\max_{x_1,x_2} p(x_1, x_2) = 0.366$. We need to find $c > 0$ and $q(x_1, x_2)$ s.t. $p(x_1, x_2) < cq(x_1, x_2), \forall x_1, x_2$

And then we would get the average acceptance rate which is $\frac{1}{c}$

Let's e.g. choose $q(x_1, x_2) = 1/4, \forall(x_1, x_2)$ and $c\frac{0.365}{0.25} = 1.464$. Then:

$$p(x_1, x_2) < 1.464q(x_1, x_2), \forall(x_1, x_2)$$

Hence the average acceptance rate is 68.3%

(ii) We have that $p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{i\sim j} \exp(-I(x_i \neq x_j))$

Let's choose a uniform $Q, \forall(x_1, x_2, x_3, x_4)$ s.t. $Q(x_1, x_2, x_3, x_4) = \frac{1}{2^4}$

Then we have that $c = \frac{\max_{x_1,x_2,x_3,x_4} p(x_1,x_2,x_3,x_4)}{\frac{1}{2^4}}$

Then the average acceptance rate is $\frac{Z}{c}$

(c) Question Explain the procedure of MCMC sampling and what is meant by a detailed balance condition. Is detailed balance a necessary condition for a sampler? Try to come up with a distribution P and a valid MCMC sampler for which the detailed balance condition does not hold.

Answer:

Suppose we want to sample from distribution $p(x)$. The procedure is to build a Markov chain such that $p(x)$ is the stationary distribution $p_\infty(x)$ of the chain:

$$p_\infty(x') = \sum_x p_\infty(x) T(x \to x')$$

Then in the long run sample from the chain will be sample from $p(x)$. We draw the first sample from distribution $p_0(x)$ and then use proposal dsitribution $T(x \to x') = p(x_t = x' | x_{t-1} = x)$.

The detailed balance condition is a condition on $p(x)$ being stationary with

$$p(x')T(x' \to x)p(x)T(x \to x')$$

The detailed balance condition is sufficient, however it is not necessary.

Consider the following example:

(x_1, x_2) are distributed according to:

$$p(x_1 = 0, x_2 = 0) = 1/3 \tag{43}$$
$$p(x_1 = 1, x_2 = 0) = 0 \tag{44}$$
$$p(x_1 = 0, x_2 = 1) = 1/3 \tag{45}$$
$$p(x_1 = 1, x_2 = 1) = 1/3 \tag{46}$$

Now, assume that we want to construct a Gibbs sampler so that x_1 is sampled first. Moving from $(x_1 = 0, x_2 = 0)$ to $(x_1' = 1, x_2' = 1)$ in just 1 move is impossible since $p(x_1' = 1 | x_2 = 0) \propto p(x_1' = 1 | x_2 = 0) = 0$.

However, $p(x' \to x) = p(x_1 = 0 | x_2' = 1) p(x_2 = 0 | x_1 = 0, x_2' = 1) > 0$

Hence, $p(x')T(x' \to x) \neq p(x)T(x \to x')$ and so detailed balance does not hold.

4 2021 Exam

4.1 1. Independence in Graphical Models

(a) Suppose X_1 and X_2