67800 | Probabilistic Methods in Artificial Intelligence | Ex 1

Guy Lutsker 207029448

Part I

Theoretical Part Answers

Question 1 - Xor Distributions

Consider the following distribution over 3 binary variables X, Y, Z:

$$p(x, y, z) \quad = \begin{cases} 1/12 & x \oplus y \oplus z = 0 \\ 1/6 & x \oplus y \oplus z = 1 \end{cases}$$

where $\oplus$ denotes a XOR function.

Show that there is no BN graph structure $\mathcal{G}$ such that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p)$

Hint: Start by testing marginal independecies (ones of the form $X \perp\!\!\!\perp Y$)

Solution :

Let us assume by way of contradiction that for the given distribution function p, there indeed exists a Bayesian Network (BN) graph $\mathcal{G}$ such that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(p)$.

Firstly, lets take the hints advice, and test the marginal distributions. Notice that our three random variables X, Y, Z are binary, and so, WLOG if we calculate $p(X = 0)$ its enough for calculating the marginal distribution of X since $p(X = 0) + p(X = 1) = 1$. With that in mind, lets proceed:

$$p(X=0) \overset{\text{Summing out } X}{=} \sum_{y \in Val(Y), z \in Val(Z)} p(0, y, z) \overset{Val(Y)=Val(Z)=\{0,1\}}{=} \sum_{y \in \{0,1\}} \sum_{z \in \{0,1\}} p(0, y, z) \overset{\text{Opening}}{=}$$

$$p(0,0,0) + p(0,0,1) + p(0,1,0) + p(0,1,1) \overset{\text{Rearange in xor grouping}}{=}$$

$$\sum_{y,z \in \{(0,0),(1,1)\}} p(0, y, z | 0 \oplus y \oplus z = 0) + \sum_{y,z \in \{(0,1),(0,1)\}} p(0, y, z | 0 \oplus y \oplus z = 1) =$$

$$= \sum_{y,z \in \{(0,0),(1,1)\}} \frac{1}{12} + \sum_{y,z \in \{(0,1),(0,1)\}} \frac{1}{6} \overset{\text{2 components in each sum}}{=} \frac{2}{12} + \frac{2}{6} = \frac{1}{2}$$

$$\Rightarrow p(X = 1) = \frac{1}{2}$$

As we can see the case of looking at the distribution of X is not special, and from the symmetry of the xor operation we can deduce that the marginal distributions of X, Y, Z are identical, and so we have done the calculation for all of them WLOG: $P(X) = P(Y) = P(Z)$.

Lets us now continue with the hint, and look at the dependency (or lack there of) of say, X, Y:

$$p(x|y) \overset{\text{def}}{=} \frac{p(x, y)}{p(y)} \overset{\text{Summing out } Z}{=} \frac{\sum_{z \in Val(Z)} p(x, y, z)}{p(y)} \overset{\text{There are only two options: Either we get xor=0, or xor=1}}{=}$$

$$\frac{p(x,y,z|x\oplus y\oplus z=0)+p(x,y,z|x\oplus y\oplus z=1)}{p(y)}=\frac{\frac{1}{12}+\frac{1}{6}}{p(y)}=\frac{\frac{1}{4}}{p(y)}=\frac{\frac{1}{2}\cdot\frac{1}{2}}{p(y)}=\frac{p(x)\cdot\cancel{p(y)}}{\cancel{p(y)}}=p(x)$$

Meaning that $p(x|y)=p(x)$, and this implies that y gives us no new information. This also indicates that $X\perp Y$. And again from the symmetry of the xor operation we get that all three random variables X,Y,Z are independent.

Now, armed we our new knowledge ($P(X)=P(Y)=P(Z)=\frac{1}{2}$ and that $X\perp Y\perp Z$), we can get back to our initial assumption about the existence of $\mathcal{G}$.

$$\text{It is given that } p \text{ holds that : } p(x,y,z|x\oplus y\oplus z=0)=\frac{1}{12}\cdot\#\{x\oplus y\oplus z=0\}=\frac{4}{12}=\frac{1}{3}.$$

Let us calculate the same proposition in the distribution $\mathcal{G}$ spaces: p_b which is defined by $p_b(x_1,...,x_n)=\prod_i p(x_i|x_{pa(i)})$

$$p_b(x,y,z|x\oplus y\oplus z=0)\overset{\text{According to xor truth table}}{=}$$

$$p_b(0,0,0)+p_b(0,1,1)+p_b(1,0,1)+p_b(1,1,0)\overset{p_b(x_1,...,x_n)=\prod_i p(x_i|x_{pa(i)})}{=}$$

$$\prod_{a\in\{X,Y,Z\}}p(a=0|x_{pa(i)})+\prod_{\substack{i\in\{0,1,1\}\\a\in\{X,Y,Z\}}}p(a=i|x_{pa(i)})+\prod_{\substack{i\in\{1,0,1\}\\a\in\{X,Y,Z\}}}p(a=i|x_{pa(i)})+\prod_{\substack{i\in\{1,1,0\}\\a\in\{X,Y,Z\}}}p(a=i|x_{pa(i)})\overset{X\perp Y\perp Z}{=}$$

$$\prod_{a\in\{X,Y,Z\}}p(a=0)+\prod_{a\in\{X,Y,Z\},i\in\{0,1,1\}}p(a=i)+\prod_{a\in\{X,Y,Z\},i\in\{1,0,1\}}p(a=i)+\prod_{a\in\{X,Y,Z\},i\in\{1,1,0\}}p(a=i)\overset{\text{Symmetry}}{=}$$

$$(\frac{1}{2})^3+(\frac{1}{2})^3+(\frac{1}{2})^3+(\frac{1}{2})^3=\frac{1}{2}$$

And so we get that $p(x,y,z|x\oplus y\oplus z=0)=\frac{1}{3}\neq\frac{1}{2}=p_b(x,y,z|x\oplus y\oplus z=0)\Rightarrow p_b\neq p$.
Which implies that our assumption that $\exists\mathcal{G}$ such that $\mathcal{I}(\mathcal{G})=\mathcal{I}(p)$ is wrong.

Question 2 - Importance of Being Acyclic

Show that if we relax the requirement that the BN graph has no cycles, it is no longer guaranteed that $p_{\mathcal{B}}(X_1,\ldots,X_n)=\prod_{i=1}^n p_i\left(X_i\mid \mathrm{Pa}_{X_i}^{\mathcal{G}}\right)$ is a valid probability distribution.
Specifically, give an example of a BN $\mathcal{B}=\langle\mathcal{G},p_{\mathcal{B}}\rangle$, where $\mathcal{G}$ has cycles, and show that $\sum\limits_{x_1,\ldots,x_n}p_{\mathcal{B}}(x_1,\ldots,x_n)\neq 1$

Solution :

Let us define a BN $\mathcal{B}=\langle\mathcal{G},p_{\mathcal{B}}\rangle$ where $\mathcal{G}$ has cycles with 2 binary random variables X_1,X_2:

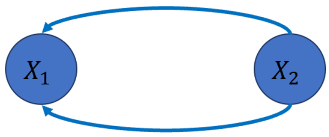


Figure 1: The graph of $\mathcal{G}$

Set $p(X_i = 0) = 0.3, p(X_i = 1) = 0.7$.

$$p(X_i = 0, X_j = 0) = 0.1$$

$$p(X_i = 1, X_j = 0) = p(X_i = 0, X_j = 1) = 0.2$$

$$p(X_i = 1, X_j = 1) = 0.5$$

And :

$$p(X_i = 0|pa(X_i) = 0) = \frac{p(X_i = 0, pa(X_i) = 0)}{p(pa(X_i) = 0)} = \frac{0.1}{0.3} = \frac{1}{3}$$
$$p(X_i = 0|pa(X_i) = 1) = \frac{p(X_i = 0, pa(X_i) = 1)}{p(pa(X_i) = 1)} = \frac{0.2}{0.7} = \frac{2}{7}$$
$$p(X_i = 1|pa(X_i) = 0) = \frac{p(X_i = 1, pa(X_i) = 0)}{p(pa(X_i) = 0)} = \frac{0.2}{0.3} = \frac{2}{3}$$
$$p(X_i = 1|pa(X_i) = 1) = \frac{p(X_i = 1, pa(X_i) = 1)}{p(pa(X_i) = 1)} = \frac{0.5}{0.7} = \frac{5}{7}$$

Let us calculate the joint probability distribution $p_{\mathcal{B}}$:

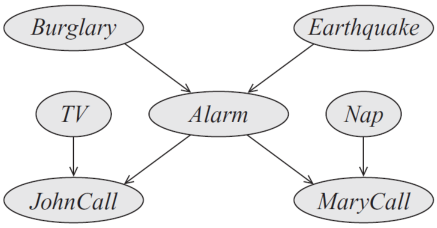
$$p_{\mathcal{B}}(X_1 = 0, X_2 = 0) = p(X_1 = 0|X_2 = 0) \cdot p(X_2 = 0|X_1 = 0) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$$
$$p_{\mathcal{B}}(X_1 = 0, X_2 = 1) = p(X_1 = 0|X_2 = 1) \cdot p(X_2 = 1|X_1 = 0) = \frac{2}{7} \cdot \frac{2}{3} = \frac{4}{21}$$
$$p_{\mathcal{B}}(X_1 = 1, X_2 = 0) = p(X_1 = 1|X_2 = 0) \cdot p(X_2 = 0|X_1 = 1) = \frac{2}{7} \cdot \frac{2}{3} = \frac{4}{21}$$
$$p_{\mathcal{B}}(X_1 = 1, X_2 = 1) = p(X_1 = 1|X_2 = 1) \cdot p(X_2 = 1|X_1 = 1) = \frac{5}{7} \cdot \frac{5}{7} = \frac{25}{49}$$

And in summation:

$$\sum_{X_1, X_2} p_{\mathcal{B}}(X_1, X_2) = 1.00226 \neq 1$$

Which means this is not a valid probability distribution, as required.

Question 3 - Removing a Variable From a Bayesian Network



1. Consider the Burglary Alarm network given above (It's a bit different than the example we saw in class - create a suitable story if you wish). Construct a Bayesian network over all nodes except the Alarm node that is a minimal I-Map for the marginal distribution over the remaining variables (B, E, N, T, J, M).
   Specifically, construct a new graph over the remaining variables which models all of the dependencies (active paths) from the original network in which A is unobserved (it's an I-map) and in which no edge can be removed without creating new independencies (it's minimal).

Solution :

The solution is as follows:

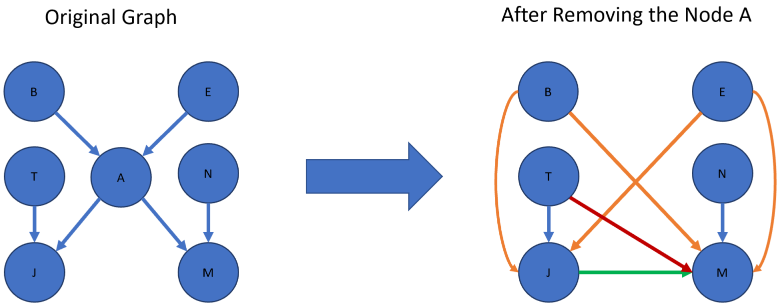


Figure 2: BN after removing "A"

Explanation: Firsly since we remove A we need to assume that A is indeed unobserved. Next, our goal here is to to preserve all independencies in the original graph:

$$B \perp E, T, N$$

$$E \perp B, T, N$$

$$T \perp B, E, A, N, M$$

$$N \perp B, E, A, T, J$$

$$A \perp T, N | B, E$$

$$J \perp B, E, N, M | T, A$$

$$M \perp B, E, T, J | A, N$$

Firstly, nodes that are unrelated to A are untouched and therefore we leave them be (Blue edges). The most obvious connection we see is that with A gone, there exist active paths between A's parents and children. And so we need to connect A's parents - B, E to A's children - J, M (Orange edges).

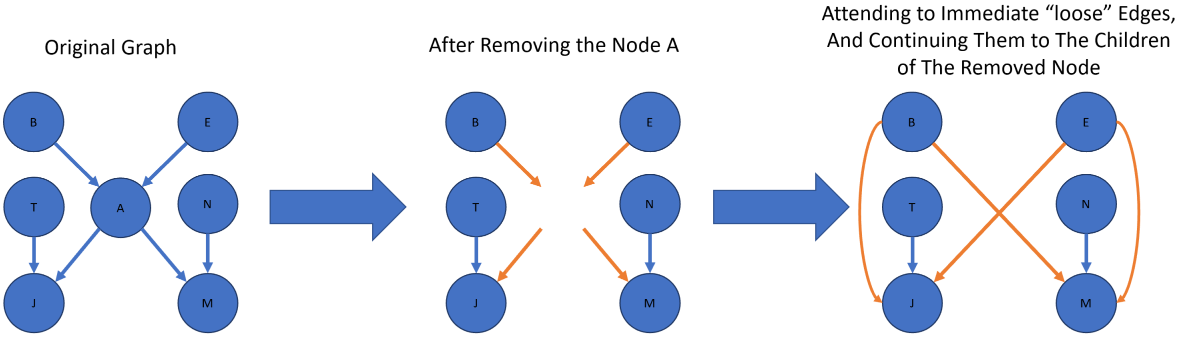
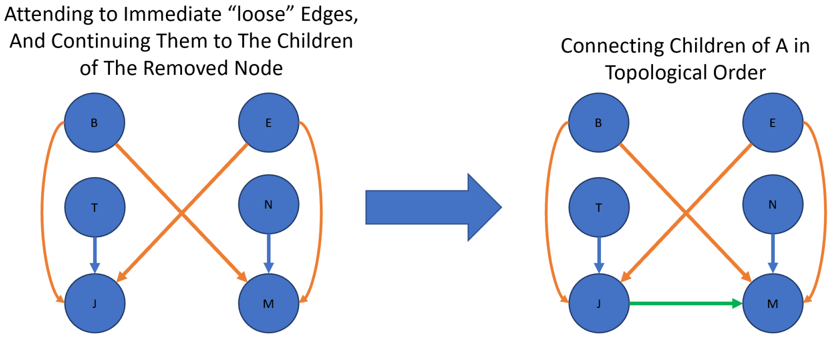

Original Graph
After Removing the Node A
Attending to Immediate "loose" Edges, And Continuing Them to The Children of The Removed Node
B
E
T
A
N
J
M
B
E
T
N
J
M
B
E
T
N
J
M

Figure 3: Step 1: Connecting The Parents of A to its Children (Orange Edges)

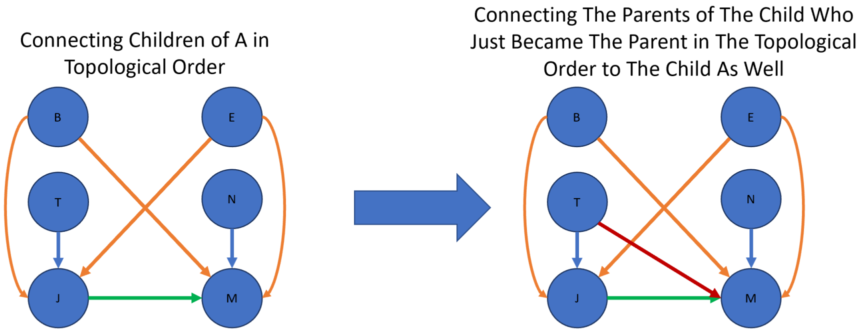
In addition we know that all children of A will now have an active path between them (called in book “common cause trail) we must draw an edge between them (the order does not matter, and we could use the topological ordering of the graph to decide the direction of the edge), And so we draw an edge from J to M (Green edge).


Attending to Immediate “loose” Edges, And Continuing Them to The Children of The Removed Node

Connecting Children of A in Topological Order

Figure 4: Step 2: Connecting Children of A in Topological Order (Green Edge)

Finally we need to attend to the v structure that was between $T \to J \leftarrow A$. We now see that there is an active path between T, M, since there might be an “explaining away” type of relationship here (if T happened, and it caused J, it means that our inference of the existence of A has less meaning, and we should update our probability of M happening). And so we need to connect an edge from T to M (Red edge).


Connecting Children of A in Topological Order

Connecting The Parents of The Child Who Just Became The Parent in The Topological Order to The Child As Well

Figure 5: Step 3: Connecting The Parents of The Child Who Just Became The Parent in The Topological Order to The Child As Well (Red Edge)

As we can there are 3 classes of new edges we need to add (color coded in my answer), and we will generalize this form in the next question. Note: since we have chosen the edge from J to M arbitrarily, we could have had an edge from M to J and then we would have also had N to J from the same logic.

2. Generalize the procedure you used above to an arbitrary network. More precisely, assume we are given a BN $\mathcal{B}$, an ordering $X_1, \ldots, X_n$ that is consistent with the topological ordering of the variables in $\mathcal{B}$, and a node X_i to be removed. Specify a network $\mathcal{B}'$ which is consistent with this ordering and which is a minimal I-Map of $p_{\mathcal{B}}(X_{-i}) = p_{\mathcal{B}}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Your answer should be an explicit specification of the set of parents for each variable in $\mathcal{B}'$.

Solution :

Suppose a graph with a topological order on the variables: $V = \{X_i\}_{i=1}^n$ and suppose we want to remove the variable $R = X_j \ s.t \ j \in [n]$. Based on my answer to question 3.1, intuitively we need to keep all of the vertices that have to relation to R the same, and do the following 3 operation:

1. Connect the all of the parents of R to all of the children (directly connected to R) - meaning $\forall p \in pa(R), c \in V \ s.t \ pa(c) = R$ connect $p \to c$.

2. Connect all of the children of R together based on the topological order - meaning we will connect the children $c_1, ...c_m$ such that $\forall k, l\ c_k \to c_l \Leftrightarrow k < l$.
3. For each pair of children c_1, c_2 that were connected in step 2 such that $c_1 \to c_2$ we need to connect the parents of c_1 to c_2 - $\forall p \in pa(c_1)$, connect $p \to c_2$.

In conclusion the set of parents is:

For convinience sake let us denote:

$$A = \overbrace{pa(pa(R))}^{\text{Connecting to the parents of the removed node}}$$

$$B = \overbrace{\{c_j | pa(c_j) \ni R,\ j < i\}}^{\text{Connecting to my siblings in topological order}}$$

$$C = \overbrace{\{pa(c_j) | pa(c_j) \ni R,\ j < i\}}^{\text{Connecting to the paretns of those siblings im connected to in topological order}}$$

$$\forall x \in V :$$

$$pa(x_i) = \begin{cases} A \cup B \cup C & pa(x_i) \ni R \\ \overbrace{pa(x_i)}^{\text{If I'm not related to the removed node, do not change anything}} & pa(x_i) \not\ni R \end{cases}$$

Question 4 - Towards Inference in Bayesian Networks

Suppose you have a Bayes net over variables $X_1, \ldots, X_n$ and all variables except X_i are observed. Using the chain rule and the network's conditional independence assumptions, find an efficient way to compute $p(X_i \mid \boldsymbol{x}_{-i}) = p(X_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.
In particular, your calculation should not require evaluating the full joint distribution.

Solution :

Let us assume a topological ordering on our variables, now:

$$p(X_i|x_{-i}) \overset{\text{Def}}{=} \frac{p(x_1, ..., x_n)}{p(x_{-i})} \overset{BN}{=} \frac{\prod_j p(x_j|pa(j))}{p(x_{-i})} \overset{\text{Go over all possible assignments of } X_i}{=}$$

$$\frac{\prod_j p(x_j|pa(j))}{\sum_{x_i} p(X_i = x_i, x_{-i})} \overset{BN}{=} \frac{\prod_j p(x_j|pa(j))}{\sum_{x_i} \prod_{j=1}^{n} p(x_j|pa(j))}$$

Notice that X_i cant be a parent of nodes labled witha lower topological order, and so:

$$= \frac{\prod_j p(x_j|pa(j))}{\sum_{x_i} \prod_{j=i}^{n} p(x_j|pa(j))} = \frac{p(X_i|pa(i)) \cdot \prod_{j:X_i \in pa(j)} p(X_j|pa(j))}{\sum_{x_i} p(X_i|pa(i)) \cdot \prod_{j:X_i \in pa(j)} p(X_j|pa(j))}$$

Question 5 -

In this question, we'll investigate when two BN graphs encode the same set of independencies

Definition. We say that two BN graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are *I-Equivalent* if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$.

Definition. Given a directed graph $\mathcal{G}$, we define its *skeleton* as the undirected graph that results from removing all arrows in $\mathcal{G}$.

1. Show that having the same skeleton is insufficient for I-equivalence. i.e. give a counter example of two graphs with the same skeleton but a different set of independencies.
2. Prove that if two BN graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same skeleton **and** the same set of v-structures then they are I-equivalent. In other words, you can flip the directionality of any edge that does not participate in a v-structure and this will not affect the graph's encoded independencies.
3. Show that the converse to (2) doesn't hold. i.e. find two two BN graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ which are I-equivalent, but which do not have the same skeleton and set of v-structures.

1

$$Soluion:$$

Counter example to the statement:

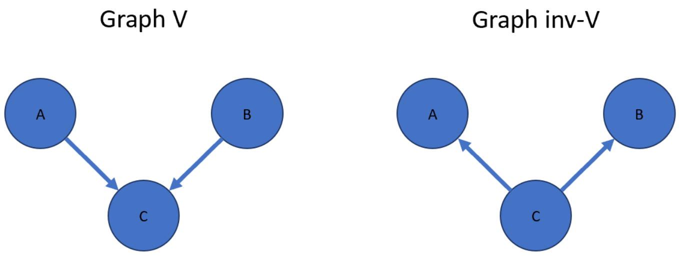


Figure 6: Graph V, Graph $inv - V$ have the same skeleton

As we can see graph V has a dependency such that $P \not\models A \perp B|C$, $B \perp A|C$, because of the nature if the v-structure.

While the independencies of graph $inv - V$ are:

$$A \perp B|C$$

$$B \perp A|C$$

2

$$Soluion:$$

Let $\mathcal{G}_1, \mathcal{G}_2$ be 2 BN graphs, such that they have the same skeleton and the same set of v-structures, then we need to prove that $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$.

We will show an equivalence of being in an active path in $\mathcal{G}_1$ and in $\mathcal{G}_2$.

Firstly let us observe that since the two graphs have the same skeleton, and so any path that exists in $\mathcal{G}_1$ also exists in $\mathcal{G}_2$. This will help us with the following lemma:

$$\forall \text{ active path } X_1 - ... - X_n \in \mathcal{G}_1, \text{ given observed RN's } Z,\ X_1\text{-...-}X_n \in \mathcal{G}_2$$

$$Proof :$$

WLOG Let $W = X_1 - ... - X_n$ be an active path in $\mathcal{G}_1$, and let $i \in [2, k-1]$ such that X_i is a node somewhere in the middle of the active path.

We can split the space into 2 options either X_i is the "sink" of a v-structure, or it isn't:

1. X_i is indeed the "sink" of a v-structure, meaning $\exists j, k$ such that $X_j \rightarrow X_i \leftarrow X_k$. Because W be an active path in $\mathcal{G}_1$ we can also deduce that either $X_i \in Z$ or $\exists c \in \overbrace{\{X_m | \exists I = (i = I_1, ..., I_s = m) \subset [k] \text{ s.t } \exists X_{I_1} \rightarrow ... \rightarrow X_{I_s} \in \mathcal{G}_1\}}^{\text{Descendants of } X_i}$ s.t $c \in Z$ (in layman's terms it means that either X_i is in the observable set Z or one of its descendants is). And again we have two options:

   - If $X_i \in Z$: We know that both $\mathcal{G}_1, \mathcal{G}_2$ have the same skeleton, and we have also inferred that any path that exists in $\mathcal{G}_1$ also exists in $\mathcal{G}_2$, in particular this active v-structure we have in $\mathcal{G}_1$. This means that this path will also be active in $\mathcal{G}_2$.
   - If one of its descendants of X_i is in Z: We can see that this node c must also be a descendant of X_i in $\mathcal{G}_2$, since we assumed that $\mathcal{G}_1, \mathcal{G}_2$ have the same skeleton, and so c must also be a descendant of X_i in $\mathcal{G}_2$. And this also means that this path will also be active in $\mathcal{G}_2$.

   Meaning that in each case having a v-structure will lead to an active path in both graphs.

2. X_i is not the "sink" of a v-structure, and we know that W is active, it immediately implies that $X_i \notin Z$. This is since we assumed that the two graphs have the same v-stuctures, and since X_i is not the "sink" of a v-structure in $\mathcal{G}_1$ it follows that it isn't a part of a v-structure in $\mathcal{G}_2$.

3

$$Soluion :$$

As we seen by now, having the same skeleton is necessary for being I-equivalent, but not sufficient, ans so we will construct a two graphs $\mathcal{G}_1, \mathcal{G}_2$ that have the same skeleton but different set of v-structures, such that $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$:

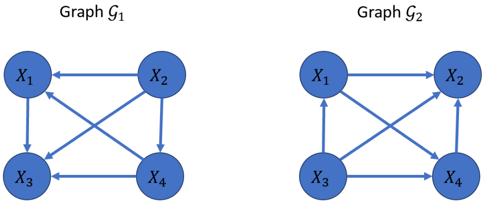


Figure 7: Graphs $\mathcal{G}_1, \mathcal{G}_2$

As we can see all nodes are connected and so $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$. And yet they do not have the same set of v-stuctures.

Part II
Practical Part Answers

Warmup Question 1: How many degrees of freedom does the joint have? i.e. how many parameters would you need to specify an arbitrary probability distribution over all possible 28 × 28 binary images?

As we have learned in class a joint probability function with n parameters, has $2^n - 1$ degrees of freedom. In our case we have 784 variables, and so we would have to specify $2^{784} - 1$ variables.

Warmup Question 2: How many degrees of freedom does the BN in fig. 1 have?

As we have seen in the recitation $deg(X|Y) = \sum_y deg(X|y) = (|Val(X) - 1| \cdot |Val(Y)|$. And so in our case we have :

$$deg(p_{\mathcal{B}}(X)) = \sum_{i=1}^{784} (|Val(X_i| - 1) \cdot |Val(Z_1)| \cdot |Val(Z_2)| = 784 \cdot 25 \cdot 25 = 490,000$$