Homework 1

BY NACHIKETA SAHOO

(Took help from Sajid in doing 1.3.4)

1 Conditional Probabilities

1.1 Prove that $X \perp Y|Z$ if and only if we can write P in the form: $P(X,Y,Z) = f(X,Y) \times g(Y,Z)$

Only-if:

$X \perp Y|Z \Rightarrow P(X,Y,Z) = P(Z)\,P(X|Z)\,P(Y|Z)$,
which can be written as $f(X,Z)\,g(Y,Z)$, where,

$$f(X,Z) = P(Z)\,P(X|Z)$$

$$g(Y,Z) = P(Y|Z)$$

If: (I'll assume that f and g are positive everywhere, hope that is not a big deal)

We have $P(X,Y|Z) = f(X,Z)\,g(Y,Z)$

$$\begin{aligned}\Rightarrow P(X,Y|Z) &= \frac{f(X,Z)}{P(Z)}\,g(Y,Z)\\ &= f'(X,Z)\,g(Y,Z)\end{aligned}$$

Integrating both the sides over X and Y we get,

$$\begin{aligned}1 &= \left(\int f'(X,Z)\,dX\right)\left(\int g(Y,Z)\,dZ\right)\\ &= I_{f'}(Z)\,I_g(Z), \text{say}\end{aligned}$$

So, $P(X,Y|Z)$ can be written as

$$\Rightarrow P(X,Y|Z) = \left(\frac{f'(X,Z)}{I_{f'}(Z)}\right)\left(\frac{g(Y,Z)}{I_g(Z)}\right)$$

For $Z = z$, each of these factors are valid p.d.f.s, because they are positive and they integrate to 1 over X and Y respectively.

Integrating $P(X,Y|Z)$ with respect to X, we have

$$P(Y|Z) = \frac{g(Y,Z)}{I_g(Z)}$$

Integrating $P(X,Y|Z)$ with respect to Y, we have

$$P(X|Z) = \frac{f'(X,Z)}{I_{f'}(Z)}$$

$$\begin{aligned}\Rightarrow P(X,Y|Z) &= P(X|Z)\,P(Y|Z)\\ \Rightarrow \quad & X \perp Y|Z\end{aligned}$$

1.2 Is it possible for every distribution P over (X ∪ Y ∪ Z) with the independence above, to be expressed as a product of a distribution over (X ∪ Z) and a distribution over (Y ∪ Z)?

Lets construct CPTs in the following manner over binary variables X, Y, Z.

$P(Z=0) = 0.3$

$P(X=0|Z=0)=0.7, P(X=0|Z=1)=0.3$
$P(Y=0|Z=0)=0.8, P(Y=0|Z=1)=0.4$
So, $X \perp Y|Z$ by construction.

$P(X=0,Y=0,Z=0)=P(Z=0)\,P(X=0|Z=0)\,P(Y=0|Z=0)=0.3\times 0.7\times 0.8=0.168$
$P(X=0,Z=0)=P(Z=0)\,P(X=0|Z=0)=0.21$
$P(Y=0,Z=0)=P(Z=0)\,P(X=0|Z=0)=0.24$
$P(X=0,Z=0)\times P(Y=0,Z=0)=0.0504$
So, we can see that $P(X=0,Y=0,Z=0)\neq P(X=0,Z=0)\times P(Y=0,Z=0)$
Therefore, when we write $P(X,Y,Z)=f(X,Z)\times g(Y,Z)$
we don't expect f and g to be p.d.f.s over (X,Z) and (Y,Z) respectively. But, they can be written as the conditional distribution of $X|Z$ by dividing by a suitable normalization constant.

1.3

1.3.1 $X\perp Y,W|Z \Rightarrow X\perp Y|Z$

Proof:

$$\begin{aligned} P(X,Y,W|Z) &= P(X|Z)\,P(Y,W|Z)\\ \Rightarrow \sum_W P(X,Y,W|Z) &= P(X|Z)\sum_W P(Y,W|Z)\\ \Rightarrow P(X,Y|Z) &= P(X|Z)\,P(Y|Z)\\ \Rightarrow & \quad X\perp Y|Z \end{aligned}$$

1.3.2 $X\perp Y|Z$ and $X,Y\perp W|Z \Rightarrow X\perp W|Z$

Proof:

$$X,Y\perp W|Z \Rightarrow X\perp W|Z\text{ , from 1.3.1}$$

1.3.3 $X\perp Y,W|Z$ and $Y\perp W|Z \Rightarrow X,W\perp Y|Z$

Proof:

$$\begin{aligned} & X\perp Y,W|Z\\ \Rightarrow P(X,Y,W|Z) &= P(X|Z)\,P(Y,W|Z)\\ &= P(X|Z)\,P(Y|Z)\,P(W|Z)\text{ because } Y\perp W|Z\\ &= P(X|Z)\,P(W|Z)\,P(Y|Z)\\ &= P(X,W|Z)\,P(Y|Z)\ \text{ as, } X\perp Y,W|Z \Rightarrow X\perp W|Z\\ \Rightarrow P(X,Y,W|Z) &= P(X,W|Z)\,P(Y|Z)\\ &\Rightarrow X,W\perp Y|Z \end{aligned}$$

1.3.4 $X\perp Y|Z$ and $X\perp Y|W \Rightarrow X\perp Y|Z,W$

It is not implied.

The idea here is when we add the probabilities for X and Y over all values of W while keeping Z constant there might be some effect of Y on X that cancels out over different values of W and X and Y become stochastically independent. In which case for a given value of

I'll illustrate this with an example.

Lets define the binary variables X,Y,Z,W and V as:

$$\begin{aligned} V &= Z\,\mathrm{xor}\,W\\ Y &= X\,\mathrm{xor}\,V \end{aligned}$$

and let the TRUTH probability of Z,W and X be 0.5.

The joint distribution of X,Y,Z,W is

Z	W	X	Y	$P(X,Y,W,Z)$
1	1	1	1	$\frac{1}{16}$
1	1	1	1	$\frac{1}{16}$
1	1	0	0	$\frac{1}{16}$
1	1	0	0	$\frac{1}{16}$
1	0	1	0	$\frac{1}{16}$
1	0	1	0	$\frac{1}{16}$
1	0	0	1	$\frac{1}{16}$
1	0	0	1	$\frac{1}{16}$
0	1	1	0	$\frac{1}{16}$
0	1	1	0	$\frac{1}{16}$
0	1	0	1	$\frac{1}{16}$
0	1	0	1	$\frac{1}{16}$
0	0	1	1	$\frac{1}{16}$
0	0	1	1	$\frac{1}{16}$
0	0	0	0	$\frac{1}{16}$
0	0	0	0	$\frac{1}{16}$

Table 1. Joint probability distribution of X, Y, Z, W

For any value of Z we have X and Y independent of each other.
For any value of W we have X and Y independent of each other.
But, for a pair of Z and W, such as (1,1), X is correlated to Y.
So, we have an example where $X\perp Y|Z$, $X\perp Y|W$,
but $X\perp Y|Z,W$ is not true.

1.4

Let X_1 and X_2 be independent binary random variables. Lets define a third binary random variable X_3 as

$$X_3 = X_1 \operatorname{xor} X_2$$

X_3	X_1	X_2	$P(X_1,X_2,X_3)$
0	1	1	0.25
1	1	0	0.25
1	0	1	0.25
0	0	1	0.25

Table 2. XOR table

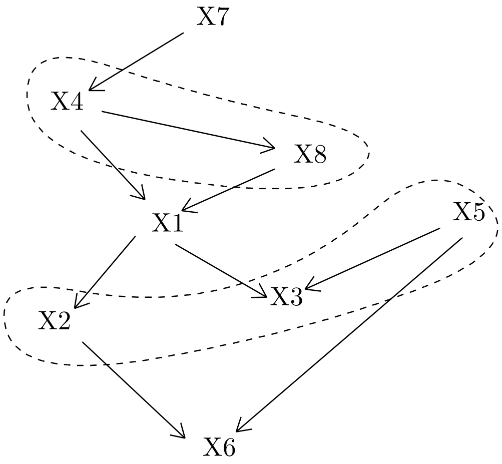
So, any two of X_3, X_1 and X_2 are independent. But, for a given value of one of the X's other two are dependent.

2 Graph Independencies

2.1

A must include the following classes of variables.

1. Direct parents of $X1$,
2. Direct children of $X2$, and
3. Parents of nodes in step 2, other than $X1$. This is necessary to remove V-structure dependencies.


X7
X4
X8
X1
X5
X3
X2
X6

By numbering the nodes in the topological order and pushing the sum over the node with largest index in, we can reduce the above sum to:

$$P(X_i, A) = \prod_{\forall j: X_j \in \{\{X_i\} \cup A\}}^{n} P\big(X_j | \mathrm{Pa}_{X_j}\big)$$

This is possible because we shall always sum over a variable that occurs random variable part of the conditional probability factors—not on the conditional part. Therefore, all such sum will reduce to 1.

If we had followed some different order for pushing the $\sum$ to the right, then we would have risked getting the variable we are summing over on the conditional part of some other CPD factor. This will not reduce to 1 and we'll be left with big joint distributions in the sum.

Since, X_i is independent of the rest of the variables given A, we can write

$$\begin{aligned} P(X_i | x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) &= P(X_i | A) \\ &= \frac{P(X_i, A)}{P(A)} \\ &= \frac{\prod_{j: X_j \in \{\{X_i\} \cup A\}}^{n} P(X_j | \mathrm{Pa}_{X_j})}{\sum_{X_i} \prod_{j: X_j \in \{\{X_i\} \cup A\}}^{n} P\big(X_j | \mathrm{Pa}_{X_j}\big)} \end{aligned}$$

If the maximum branching factor of this graph is F, the size of A is of the order $O\big(F^2\big)$ since we are including parents of children of X_i.

Lets say X_i can take k different values, then the complexity of this computation is of the order $O\big(k \cdot F^2\big)$.

3 Factorization

Q. Show that if P factorizes according to G, then G is an I-map of P

Lets call the set of nodes that are non-descendants of any node X_i as Nd_{X_i} and the set of nodes that are descendants D_{X_i}.

Given P factorizes according to G we have to prove that

$$P(X_i | \mathrm{Pa}_{X_i}, \mathrm{Nd}_{X_i}) = P(X_i | \mathrm{Pa}_{X_i}), \forall i$$

Let's start with the joint of χ, the set of all variables.

$$\begin{aligned} P(\chi) &= P(X_i, \mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i}, D_{X_i}) \\ &= P(X_i | \mathrm{Pa}_{X_i}) \left(\prod_{j: X_j \in \mathrm{Pa}_{X_i}} P(X_j | \mathrm{Pa}_{X_j}) \right) \\ &\quad \left(\prod_{j: X_j \in \mathrm{Nd}_{X_i}} P(X_j | \mathrm{Pa}_{X_j}) \right) \\ &\quad \left(\prod_{j: X_j \in D_{X_i}} P(X_j | \mathrm{Pa}_{X_j}) \right) \\ \Rightarrow \sum_{D_{X_i}} P(X_i, \mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i}, D_{X_i}) &= P(X_i | \mathrm{Pa}_{X_i}) \left(\prod_{j: X_j \in \mathrm{Pa}_{X_i}} P(X_j | \mathrm{Pa}_{X_j}) \right) \\ &\quad \left(\prod_{j: X_j \in \mathrm{Nd}_{X_i}} P(X_j | \mathrm{Pa}_{X_j}) \right) \\ &\quad \sum_{D_{X_i}} \left(\prod_{j: X_j \in D_{X_i}} P(X_j | \mathrm{Pa}_{X_j}) \right) \end{aligned}$$

The factor $\sum_{D_{X_i}} \left(\prod_{j:X_j \in D_{X_i}} P(X_j|\mathrm{Pa}_{X_j}) \right)$ reduces to 1. This can be observed by numbering the nodes in their topological order and repeatedly pushing the $\sum$ over to the factor with the highest numbered variable on it. This term reduces to 1, since it is a sum of conditional probabilities of the variable. This gives us

$$\begin{aligned} P(X_i, \mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i}) &= P(X_i|\mathrm{Pa}_{X_i}) \left(\prod_{j:X_j \in \mathrm{Pa}_{X_i}} P(X_j|\mathrm{Pa}_{X_j}) \right) \\ & \left(\prod_{j:X_j \in \mathrm{Nd}_{X_i}} P(X_j|\mathrm{Pa}_{X_j}) \right) \end{aligned} \tag{1}$$

By summing over X_i we get

$$\begin{aligned} P(\mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i}) &= \sum_{X_i} P(X_i|\mathrm{Pa}_{X_i}) \left(\prod_{j:X_j \in \mathrm{Pa}_{X_i}} P(X_j|\mathrm{Pa}_{X_j}) \right) \\ & \left(\prod_{j:X_j \in \mathrm{Nd}_{X_i}} P(X_j|\mathrm{Pa}_{X_j}) \right) \end{aligned}$$

Since, X_i can not occur in the conditional part of any of the parent nodes or the non-descendants, we can move the factor to the right and sum it to 1. This gives us

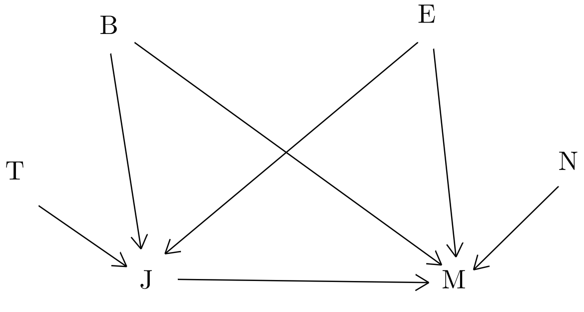
$$P(\mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i}) = \left(\prod_{j:X_j \in \mathrm{Pa}_{X_i}} P(X_j|\mathrm{Pa}_{X_j}) \right) \left(\prod_{j:X_j \in \mathrm{Nd}_{X_i}} P(X_j|\mathrm{Pa}_{X_j}) \right) \tag{2}$$

Dividing (1) by (2) we get

$$\begin{aligned} \frac{P(X_i, \mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i})}{P(\mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i})} &= P(X_i|\mathrm{Pa}_{X_i}) \\ \Rightarrow P(X_i|\mathrm{Nd}_{X_i}, \mathrm{Pa}_{X_i}) &= P(X_i|\mathrm{Pa}_{X_i}), \text{proved.} \end{aligned}$$

4 Marginalization

4.1



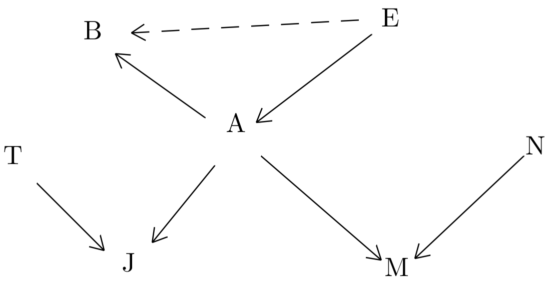
4.2

1. Join all pairs of nodes that are neighbours of X_l in G and they have a path between them in G that goes through X_l.

2. Delete X_l and edges containing X_l from G

5

5.1



When we reversed $\overrightarrow{BA}$ to $\overrightarrow{AB}$, we introduced the conditional independence $B \perp E | A$ which was not there in the original graph. In the original graph B and E were dependent given A because of the V-structure. To keep the new graph still an I-map of the original distribution we need to remove this independence. This can be done by adding an arrow from E to B as shown in the graph (Not from B to E, because we don't want a cycle).

Note that we lost an independence from the original graph: $B \perp E$. But, that is O.K. since the remaining conditional independencies are a subset of the conditional independencies satisfied by the original distribution.

5.2

1. Reverse the direction of the arc $\overrightarrow{XY}$ to $\overrightarrow{YX}$.
2. If there were V-structures with Y as the center and X as one of the parents, then join the other parents to X using arrows pointing to X.

By reversing an arc if we are removing some V-structures, then, we are creating extra independencies that were not in the original distribution, e.g., if $X \rightarrow Y \leftarrow Z$ was one of the V-structures and we reversed $X \rightarrow Y$ to $X \leftarrow Y$, then now $X \perp Z | Y$ in the graph segment $X \leftarrow Y \leftarrow Z$. This extra independence might cause the graph to not be an I-map anymore. We can mitigate this by drawing an arrow from Z to X, which removes the independencies.

5.3 Are we guaranteed to get back the original graph by reversing the same arc twice?

No!

Notice that we might have lost independencies while doing arc reversal (illustrated in 5.1). When we reverse the same arc in G' we can't remove the extra edges we added while coming from G to G', because that would require making more independence assumptions than there are in G' (dropping edges makes more independence assumptions).

So, no, we are not guaranteed to get back the original G. Independencies might have been lost on the way and we might have been left with more edges than what we had in G.

6 Inference

6.1 $P(\mathrm{WD}=T|G=T)$

$$P(\mathrm{WD}=T|G=T) \;\propto\; P(\mathrm{WD}=T,G=T) =$$

$$\begin{aligned}
&= \sum_{\mathrm{TF},\mathrm{FF},M,P,K} P(\mathrm{TF})\,P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\,P(K|G=T)\,P(M|G=T)\,P(P|M)\\
&\quad P(\mathrm{WD}=T|P,M)\\
&= \sum_{\mathrm{TF},\mathrm{FF},M,P} P(\mathrm{TF})\,P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\,P(M|G=T)\,P(P|M)\,P(\mathrm{WD}=T|P,M)\\
&\quad \sum_K P(K|G=T)\\
&= \sum_{\mathrm{TF}} P(\mathrm{TF}) \sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF}) \sum_M P(M|G=T) \sum_P P(P|M)\,P(\mathrm{WD}=T|P,M)\\
&= \left(\sum_{\mathrm{TF}} P(\mathrm{TF}) \sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\right)\times\\
&\quad \left(\sum_M P(M|G=T) \sum_P P(P|M)\,P(\mathrm{WD}=T|P,M)\right)
\end{aligned}$$

Note that

$$\begin{aligned}
g_1(M,\mathrm{WD}) &= \sum_P P(P|M)\,P(\mathrm{WD}|P,M)\\
g_2(\mathrm{WD},G=T) &= \sum_M P(M|G=T)\,g_1(M,\mathrm{WD})\\
g_3(\mathrm{TF},G=T) &= \sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\\
g_4(G=T) &= \sum_{\mathrm{TF}} P(\mathrm{TF})\,g_3(\mathrm{TF},G=T)
\end{aligned}$$

The reason I am writing $G=T$ is to indicate that for this inference we need to evaluate these functions only for $G=T$. However, as we shall see we need to evaluate these functions for $\mathrm{WD}=T$ and $\mathrm{WD}=F$ in order to determine the normalizing factor for computing $P(\mathrm{WD}|G=T)$.

The number of computation needed here is 40 instead of 386 that we would have had to do if we had constructed the complete joint from the conditional probability tables.

$$P(\mathrm{WD}=T|G=T) \;=\; \frac{P(\mathrm{WD}=T,G=T)}{P(\mathrm{WD}=F|G=T)+P(\mathrm{WD}=T,G=T)}$$

The value of the probability turns out to be

$$P(\mathrm{WD}=T|G=T) \;=\; 0.426$$

6.2 $P(\mathrm{WD}=T|\mathrm{FF}=T)$

$$\begin{aligned}
&P(\mathrm{WD}=T|\mathrm{FF}=T)\\
\propto\;& P(\mathrm{WD}=T,\mathrm{FF}=T)\\
=\;& \sum_{\mathrm{TF},\mathrm{FF},M,P,K} P(\mathrm{TF})\,P(\mathrm{FF}=T|\mathrm{TF})\,P(G|\mathrm{TF},\mathrm{FF}=T)\,P(K|G)\,P(M|G)\,P(P|M)\\
&P(\mathrm{WD}=T|P,M)\\
=\;& \sum_{\mathrm{TF},G,M,P} P(\mathrm{TF})\,P(\mathrm{FF}=T|\mathrm{TF})\,P(G|\mathrm{TF},\mathrm{FF}=T)\,P(M|G)\,P(P|M)\,P(\mathrm{WD}=T|P,M)\\
&\sum_K P(K|G)\\
=\;& \sum_{\mathrm{TF}} P(\mathrm{TF})\,P(\mathrm{FF}=T|\mathrm{TF}) \sum_G P(G|\mathrm{TF},\mathrm{FF}=T) \sum_M P(M|G) \sum_P P(P|M)\,P(\mathrm{WD}=T|P,M)
\end{aligned}$$

Here also, by computing the functions from right to left as illustrated in 6.1, we'll save in computation. I'll not find out the exact number of operations involved here as it is not required in the assignment, but, it can be observed that the distributive property of algebra saves us computation here as well.

Upon calculation we arrive at the probability

$$\begin{aligned} P(\mathrm{WD}=T|\mathrm{FF}=T) &= \frac{P(\mathrm{WD}=T|\mathrm{FF}=T)}{P(\mathrm{WD}=F,\mathrm{FF}=T)+P(\mathrm{WD}=T,\mathrm{FF}=T)} \\ &= 0.402 \end{aligned}$$

6.3 $P(G)$

$$\begin{aligned} P(G=T) &= \sum_{\mathrm{TF},\mathrm{FF},M,P,K,\mathrm{WD}} P(\mathrm{TF})\,P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF}) \\ &\quad P(K|G=T)\,P(M|G=T)\,P(P|M)\,P(\mathrm{WD}|P,M) \\ &= \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\sum_{M} P(M|G=T)\sum_{P} P(P|M) \\ &\quad \sum_{\mathrm{WD}} P(\mathrm{WD}|P,M)\sum_{K} P(K|G=T) \\ &= \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\sum_{M} P(M|G=T)\sum_{P} P(P|M) \\ &\quad \sum_{\mathrm{WD}} P(\mathrm{WD}|P,M)\times 1 \\ &= \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\sum_{M} P(M|G=T)\sum_{P} P(P|M)\times 1\times 1 \\ &= \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\sum_{M} P(M|G=T)\times 1\times 1\times 1 \\ &= \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF})\times 1\times 1\times 1\times 1 \\ &= \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\,P(G=T|\mathrm{TF},\mathrm{FF}) \\ &= 0.392 \end{aligned}$$

6.4 $P(M=T|K=T)$

$$\begin{aligned} & P(M|K=T) \\ \propto\ & P(M,K=T) \\ =\ & \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\sum_{G} P(G|\mathrm{TF},\mathrm{FF}) \\ & P(K=T|G)\,P(M|G)\sum_{P} P(P|T)\sum_{\mathrm{WD}} P(\mathrm{WD}|P,M) \\ =\ & \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\sum_{G} P(G|\mathrm{TF},\mathrm{FF}) \\ & P(K=T|G)\,P(M|G)\sum_{P} P(P|M)\times 1 \\ =\ & \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\sum_{G} P(G|\mathrm{TF},\mathrm{FF}) \\ & P(K=T|G)\,P(M|G)\times 1\times 1 \\ =\ & \sum_{\mathrm{TF}} P(\mathrm{TF})\sum_{\mathrm{FF}} P(\mathrm{FF}|\mathrm{TF})\sum_{G} P(G|\mathrm{TF},\mathrm{FF})\,P(K=T|G)\,P(M|G) \end{aligned}$$

Constructing and computing the functions from right to left we'll save on computation. The conditional probability is

$$\begin{aligned} P(M=T|K=T) &= \frac{P(M=T,K=T)}{P(M=F,K=T)+P(M=T,K=T)} \\ &= 0.542 \end{aligned}$$

7 Conditional Probabilities in Variable Elimination

7.1

$$W \;=\; \bigcup_{i=1}^{k} \Big(\{Y_i\} \cup \mathrm{Pa}_{Y_i}\Big)$$

Lets construct a graph B' by creating disconnected nodes $W - \cup_{i=1}^{k} Y_i$.

Then complete the graph by adding Y_i and making Pa_{Y_i} its parents, for $1 < i < k$.

$$\begin{aligned} P(W) \;&=\; \left(\prod_{i=1}^{k} P(Y_i|\mathrm{Pa}_{Y_i})\right)\left(\prod_{j:Y_j\in\{W-\cup_{i=1}^{k}Y_i\}} P(Y_j)\right) \\ &=\; \tau(W)\left(\prod_{j:Y_j\in\{W-\cup_{i=1}^{k}Y_i\}} P(Y_j)\right) \qquad (3) \end{aligned}$$

It can be shown that $\Big(\prod_{j:Y_j\in\{W-\cup_{i=1}^{k}Y_i\}} P(Y_j)\Big)$ is indeed equal to $P\big(\{W - \cup_{i=1}^{k} Y_i\}\big)$, by marginalizing $\cup_{i=1}^{k} Y_i$ out from $P(W)$. To do this:

1. Order the nodes in W in topological order.
2. sum over the variable in $\cup_{i=1}^{k} Y_i$ with the highest index, this will reduce to 1.

By repeating 2, we can reduce the factor $\prod_{i=1}^{k} P(Y_i|\mathrm{Pa}_{Y_i})$ to 1 and left with

$$P\big(\{W - \cup_{i=1}^{k} Y_i\}\big) \;=\; \left(\prod_{j:Y_j\in\{W-\cup_{i=1}^{k}Y_i\}} P(Y_j)\right)$$

Hence from (3)

$$\begin{aligned} \tau(W) \;&=\; \frac{P(W)}{P\big(\{W - \cup_{i=1}^{k} Y_i\}\big)} \\ &=\; P\big(\cup_{i=1}^{k} Y_i | W - \cup_{i=1}^{k} Y_i\big) \end{aligned}$$

7.2

Variable elimination is a process of summing out some variables from a joint distribution. When a joint distribution is factorized according to a graph, this process looks something like

$$\sum_{(X_1,\ldots,X_j)} P(X_1, X_2, X_3 \ldots X_n) \;=\; \sum_{(X_1,\ldots,X_j)} \prod_{i=1}^{n} P(X_i|\mathrm{Pa}_{X_i})$$

Here we are eliminating j variables and without any loss of generality we can number them from 1 to j. Lets say there are k factors that contain some of $(X_1, \ldots, X_j)$. Then the intermediate factor generated after summing out $(X_1, \ldots, X_j)$ will be of the form $g_j(X_{j+1}, \ldots, X_k)$ and it is generated solely out of $\prod_{i=1}^{k} P(X_i|\mathrm{Pa}_{X_i})$ by marginalizing out $(X_1, \ldots, X_j)$.

We called $\prod_{i=1}^{k} P(X_i|\mathrm{Pa}_{X_i})$ as $\tau(W)$ in 7.1, where, $W = \{X_1, \ldots, X_k\}$ and showed that

$$\tau(X_1, \ldots, X_k) \;=\; P\big(X_1, \ldots, X_k | W - \cup_{i=1}^{k} X_i\big)$$

As the product is already a conditional probability, we can marginalize $X_1, \ldots, X_j$ from it obtaining

$$\begin{aligned} g_j(X_{j+1}, \ldots, X_k) \;&=\; \sum_{(X_1,\ldots,X_j)} \prod_{i=1}^{k} P(X_i|\mathrm{Pa}_{X_i}) \\ &=\; \sum_{(X_1,\ldots,X_j)} \tau(X_1, \ldots, X_k) \\ &=\; \sum_{(X_1,\ldots,X_j)} P\big(X_1, \ldots, X_k | W - \cup_{i=1}^{k} X_i\big) \\ &=\; P\big(X_{j+1}, \ldots, X_k | W - \cup_{i=1}^{k} X_i\big) \end{aligned}$$

which is a conditional probability.