18.1 Towards Inference

Recall the directed factorization (DF) property of a DGM and the factorization (F) property of a UGM.

\[ p(x) = \prod_v p(x_v|x_{pa(v)}) \quad \text{(DF)} \]

\[ p(x) = \prod_{c \in \mathcal{C}} \psi_c(x) \quad \text{(F)} \]

where \( \psi_c \)'s represent the clique potentials. The sequence of steps to go from a DGM to a UGM are

- Initialize all clique potentials to unity; i.e, set \( \psi_c(x) = 1 \), \( \forall c \)
- For each \( p(x_v|x_{pa(v)}) \), choose a clique \( c \) such that \( x_v \cup x_{pa(v)} \) is contained within the clique. Update the clique potential as follows -

\[ \psi_c^{(u)}(x) = \psi_c(x) p(x_v|x_{pa(v)}) \quad \text{(18.1)} \]

where \( \psi_c^{(u)}(x) \) is the updated version of \( \psi_c(x) \).

Let’s now work out an example to illustrate how the clique potentials can be used to represent the joint probability distribution on a graph.

![Six node DAG](image.png)
After we moralized and triangulate, the cliques of the above graph might be \((ABC)\), \((CDE)\), and \((DEF)\) and the associated clique potentials are -

\[
\begin{align*}
\psi_{ABC} &= p(A)p(B)p(C|A, B) \\
\psi_{CDE} &= p(D|C)p(E|C) \\
\psi_{DEF} &= p(F|D, E)
\end{align*}
\]

From the graph, the joint distribution of the nodes can be written as

\[
p(ABCDE) = p(A)p(B)p(C|A, B)p(D|C)p(E|C)p(F|D, E) = \psi_{ABC}\psi_{CDE}\psi_{DEF}
\]

In general, clique potentials are not unique. Introducing some scaling constants, we can modify the clique potentials such that the global distribution \(p(x)\) is maintained. That is, we can write

\[
p(x) = \prod_c \psi_c(x) = \prod_c \psi'_c(x) \tag{18.2}
\]

where \(\psi'_{c_1}(x) = \alpha \psi_{c_1}(x)\) and \(\psi'_{c_2}(x) = \frac{1}{\alpha} \psi_{c_2}(x)\). In fact, we could allow the clique functions to be arbitrary positive values, and as long as we normalize to ensure that the probability sums to one, we can have:

\[
p(x) = \frac{1}{Z} \prod_c \psi_c(x) \tag{18.3}
\]

where

\[
Z = \int \prod_c \psi_c(x) dx
\]

Note that the normalization constant could be absorbed by one of the clique potential functions. The main point is that the clique potentials can, in the general case, be arbitrary non-negative functions.

Before proceeding any further, let us recall the reasons why moralization is valid. Moralization on a DGM is valid because it does not violate the semantics of the original graph and makes only fewer conditional independence statements. Fewer independence statements corresponds to a larger family of probability distributions, and if we solve the inference problem for this larger family, we have solved it for the more specific family as well. Moralization is necessary so that the undirected model does not violate any of the conditional independence statements in the original graph.

### 18.2 Introducing evidence

We now consider the problem of conditioning, or “introducing evidence”. We suppose that the nodes are partitioned into subsets \(H\) and \(E\), and that the random vector \(X_E\) is observed to take on a specific value. The problem that we discuss in this section is that representing the conditional probability \(p(x_H|x_E)\). Once we have decided on such a representation, the inferential problem of computing marginals under this probability—the conditional probabilities of subsets of the nodes \(X_H\) – will be no different in principle from the calculation of marginal probabilities under the overall joint \(p(x)\).

Before proceeding with this section, let us first define the types of evidence we might encounter. Evidence might be of three types, as described below

- \(E = e\) (i.e., some set of values assigned to \(E\))
• $E = \{ e_1 \text{ or } e_2 \text{ or } \cdots e_n \}$ or

• $E \notin \{ e_1 \text{ or } e_2 \text{ or } \cdots e_n \}$

Note that these types are not exclusive of each other. And we are not dealing with the type of $E_2/E_1 = \alpha$ evidence.

For concreteness, we will concentrate on the first type here. Let $V$ be a set of nodes such that

$$V = E \cup H,$$

where $E$ refers to the set of nodes for which we have evidence, and $H$ refers to the hidden nodes. Thus

$$p(V) = p(H, E) = \frac{\prod_c \psi_c(H, E)}{Z} \quad (18.4)$$

The probability $p(H|E)$ can be computed as

$$p(H|E = e) = \frac{p(H, E = e)}{p(E = e)} = \frac{\prod_c \psi_c(H, E = e)/Z}{\sum_H \prod_c \psi_c(H, E = e)/Z} = \frac{\prod_c \psi_c(H, E = e)}{\sum_H \prod_c \psi_c(H, E = e)} \quad (18.5)$$

Thus, $\prod_c \psi_c(H, E = e)$ is an un-normalized version of both the joint distribution with evidence $p(H, E = e)$ and of the conditional of the hidden variables given the evidence $p(H|E = e)$. To get either, we just do the appropriate normalization. Furthermore, note that if the cliques potentials represented true marginal distributions over their variables (i.e., if we had that $\psi_c(H, E = e) = p(H \cap c, E \cap c)$), then it would be easy to do per-clique normalizatoin to get $p(H \cap c|e \cap c)$.

The next question is, how do we compute $p(H|E = e)$? We do this by normalization, and there are two potential ways of doing this:

• By summing over $H$ the product of the clique potential functions (i.e, $\sum_H \prod_c \psi_c(H, E = e)$) or

• By summing over $E \& H$ the modified clique potential functions such that all entries which are inconsistent with the evidence are set to zero, i.e.

$$\sum \prod \psi_c(H, E = e)\delta_{E=e} = \sum \prod \psi_c(H, E = e)\delta_{E\cap C = \{e\cap c\}}$$

$$= \sum \prod \psi_c'(H, E = e)$$

We will use the second way. This means that to incorporate evidence $E = e$, it is sufficient to change the potential functions to be zero for configurations that are inconsistent with $E = e$. Of course, the clique potentials will not in that case be marginal distributions (as described above), but we will deal with that issue soon.

The following example will help to set matters straight.

![Figure 18.2: 2 node DAG](image)

We will assume that both the nodes are binary, $p(A = 1) = 0.8$, $p(A = 0) = 0.2$, $p(B = 1|A = 1) = 0.7$ and $p(B = 1|A = 0) = 0.4$. Therefore, $\psi_{AB} = p(A)p(B|A) = \begin{bmatrix} 0.12 & 0.08 \\ 0.24 & 0.56 \end{bmatrix}$

If the evidence is $B = 1$, we can zero out the first column. We then have $\psi_{AB} = \begin{bmatrix} 0 & 0.08 \\ 0 & 0.56 \end{bmatrix}$
After normalizing, $\psi_{AB} = \begin{bmatrix} 0 & 0.125 \\ 0 & 0.875 \end{bmatrix} = p(A|B = 1)$

To summarize, incorporating evidence means we zero out entries in the clique potentials that are inconsistent with the evidence.

### 18.3 Clique potentials as Marginals

Using the Hammersley-Clifford theorem we proved a few lectures ago, if we have a distribution that is positive and that obeys the global Markov property, we can represent this distribution using the undirected factorization property of a graph, meaning that we can make the following statement: $p(x) = \prod_c \psi_c(x)$. This says that the joint distribution can be factored as a product of the clique potentials.

But, in the general case (when the distribution might not be positive), we might not have this factorization. Furthermore, clique potentials might not be able to represent marginal probabilities over the corresponding clique variables when we use the above factorization form (this is desired as we will see). This is illustrated in the following example of a 3-state Markov chain.

![Figure 18.3: 3 state MC](image)

The cliques in the above graph are $(AB)$ and $(BC)$. The joint probability distribution of $A, B, C$ can be written as

$$p(A, B, C) = p(A)p(B|A)p(C|B) = p(A, B)p(C|B) = \psi_{AB}\psi_{BC} \neq p(A, B)p(B, C)$$

Thus, we cannot represent the joint over $A, B, C$ as the product over two functions only of $(AB)$ and $(BC)$ respectively.

However, as we say 2 lectures ago, if the graph is decomposable, we can write it as:

$$p(x) = \frac{\text{product of clique marginals}}{\text{product of separator marginals}} = \frac{\prod_{c \in \mathcal{C}} p(x_c)}{\prod_{s \in \mathcal{S}} p(x_s)} = \frac{\text{product of clique potentials}}{\text{product of separator potentials}} = \frac{\prod_c \psi_c(x)}{\prod_s \phi_s(x)}$$

Note again that the potential functions can be arbitrarily scaled if we maintain the global probability values $p(x)$. This shows that it is possible to have the clique potentials equal to the marginals in this representation, i.e., we can have $\psi_c(x_c) = p(x_c)$ and $\phi_s(x_s) = p(x_s)$.

As we will see, the JT algorithm will adjust the potential functions so that the are real marginals, starting from any initial valid configuration. This means that the above will hold for the current clique potentials.
18.3.1 Why do we want marginals?

If local clique potentials are equal to the marginal distributions over the corresponding variables, then it is very easy to calculate the probabilities of individual nodes, i.e., if \( \psi_c(x_c) = p(x_c) \), then it is easy to calculate

\[
p(x_{c'}) = \sum_{c \subseteq c'} \psi_c(x_c)
\]

where \( c' \subseteq c \).

If the nodes we are interested in live in the same cliques, i.e., \( \psi_c(H \cap C, E \cap C) = P(H \cap C, E = e) \), where \( H = H_1, H_2, \ldots, \) then

\[
P(H_3|E = e) = \frac{P(H_3, E = e)}{P(E = e)} = \frac{\sum_{c \setminus (H_3)} P(C, E = e)}{\sum_{c'} P(C, E = e)}
\]

If the nodes we are interested in live in different cliques of a junction tree, say \( c_1 \) and \( c_2 \) and we wish to compute \( p(x_{c_1}, x_{c_2}) \) where \( c'_1 \subseteq c_1 \) and \( c'_2 \subseteq c_2 \), then can also easily be computed as

\[
p(x_{c_1'}, x_{c_2'}) = p(x_{c_1'}|x_{c_2'})p(x_{c_2'})
\]

\[
= p(x_{c_1'}|x_{s'})p(x_{c_2'})
\]

\[
= p(x_{c_1'}, x_{s'})p(x_{c_2'})/p(x_{s'})
\]

\[
= p(x_{c_1'})p(x_{c_2'})/p(x_{s'})
\]

where

\[
p(x_{c_1'}) = \sum_{c_1 \setminus c_1'} \psi_{c_1}(x_{c_1})
\]

and

\[
p(x_{c_2'}) = \sum_{c_2 \setminus c_2'} \psi_{c_2}(x_{c_2})
\]

and

\[
p(x_{s'}) = \sum_{c_1 \setminus s'} \psi_{c_1}(x_{c_1}) = \sum_{c_2 \setminus s} \psi_{c_2}(x_{c_2})
\]

and where \( s' = c'_1 \cap c'_2 \) is the separator. The above equations follow because of the conditional independence properties of the underlying graph. We have that \( p(x_{c_1'}, x_{s'}) = p(x_{c_1'}) \) because \( s' \subseteq c_1' \).

So, our goal is to get the potentials as marginals as that makes things much easier. But the potentials are not always marginals, there are several reasons why the clique potentials might not be marginals:

1. When we incorporate evidence (by zeroing out entries in clique potentials), the potentials no longer contain the right marginals of the form \( \psi_c(H, E = e) \). \( P(H, E = e) = \prod_c \psi_c(H, E=e) \prod_c \psi_c(H, E=e) \) Some clique potential functions will be unchanged, i.e., those with \( C \cap E = \emptyset \). Those will not be able to produce the quantities \( p(C|E = e) \) by normalizing but this is something that we will want as it will not in general be the case that \( C \perp| \, E \) in the graph (so changing \( E \) should effect the conditional distribution). In this case, normalization would get some marginal but it would be the wrong marginal.

2. Clique potentials might have been initialized by conditional probabilities from the directed graph, in which case they are not marginals. For example, we had the example from before where \( \phi_{D,E,F} = p(F|D,E) \) which is not a marginal over \( D,E,F \).
3. Raw clique potentials: might specify clique potentials arbitrarily at first (e.g., a Markov random field on an image), so clique potentials represent some unnormalized values, where the normalization constant $Z$ could be distributed amongst the potentials in an arbitrary way. For example,

$$
\psi_c(x) = e^{\sum_i \lambda_i f_i(x)}
$$

where

$$
f_i(x) = \begin{cases}
3 : & x^{(1)}_c = x^{(2)}_c \\
0 : & \text{else}
\end{cases}
$$

So, for now assume we start with

$$
P(H, E = e) = \frac{\prod_c \psi_c(H, E = e)}{\prod_s \psi_s(H, E = e)}
$$

where $\psi_s(H, E = e) = 1$ for now.

The key thing is this: For $\psi_c$, the clique potentials to be marginals, they at least must agree on the nodes they have in common with each other. Agree in this case means “be consistent with”, i.e., They must give the same marginals over their intersection.

If $c_1$ and $c_2$ are connected in a junction tree through a separator set $S$, local consistency means that

$$
\sum_{c_1 \setminus S} \psi_{c_1}(x_{c_1}) = \sum_{c_2 \setminus S} \psi_{c_2}(x_{c_2}) = \phi_s(x_s)
$$

Running the JT algorithm will give us that all the clique potentials are marginals and therefore are consistent. But this algorithm does this in a way that we only need to ensure it locally. That is we can prove that it will be true globally if it is true locally. But first, we need to see how to ensure local consistency.

### 18.4 Local consistency: How to achieve.

Suppose we have two cliques in the JT $V$ and $W$ that are connected with separator set $S$, so that $S = V \cap W \neq \emptyset$, as shown in the following figure.

![Figure 18.4: The basic data structures underlying the flow of information between cliques $V$ and $W.](image)

The cliques $V$ and $W$ have potentials $\psi_V$ and $\psi_W$, and we also endow $S$ with a potential $\phi_S$ that we initialize to unity. The basic operation of the junction tree algorithm is an exchange of information between $V$ and $W$, with $S$ serving as a conduit for the flow of information.

Suppose that we start off with the potentials inconsistent, so that

$$
\sum_{V \setminus S} \psi_V \neq \sum_{W \setminus S} \psi_W
$$

and $\phi_S = 1$. We also have that $p(H, E = e) = \psi_V \psi_W / \phi_S$. 
There are the operations. First we update $W$ based on $V$, where the asterisk means “updated value of”:

$$\phi_S^* = \sum_{V \setminus S} \psi_V$$

This marginalizes the potential $\psi_V$ with respect to $S$, storing the result in the separator potential.

Second, we re-scale the potential on $W$ by multiplying an “update factor”, that is the ratio of the new separator potential to its old value.

$$\phi_W^* = \frac{\phi_S^*}{\phi_S} \psi_W$$

This update has an important invariant: the joint distribution $p(H, E = e)$. Note that $\psi_V$ is unchanged during the update, defining $\psi_V^* = \psi_V$, we have

$$\psi_V^* \psi_W^* = \frac{\psi_V \psi_W \phi_S^*}{\phi_S \phi_S^*} = \frac{\phi_S^*}{\phi_S}$$

but we have not achieved consistency yet as

$$\sum_{V \setminus S} \psi_V^* = \sum_{V \setminus S} \psi_V = \phi_S^* \neq \sum_{W \setminus S} \psi_W = \frac{\psi_S^*}{\psi_S} \sum_{W \setminus S} \psi_W$$

since $\phi_S \neq \sum_{W \setminus S} \psi_W$.

We now pass information from $W$ back to $V$, using the same update rule. In particular,

$$\phi_S^{**} = \sum_{W \setminus S} \psi_W$$

$$\psi_V^{**} = \frac{\phi_S^{**}}{\phi_S} \psi_V$$

And $\psi_V^*$ is unchanged during this update, we define $\psi_W^{**} = \psi_W$.

There is another important property that characterizes the pair of updates. In particular, the potentials $\psi_W^{**}$ and $\psi_V^{**}$ are consistent with respect to their intersection $S$; that is, they have the same marginals. This is easily verified:

$$\sum_{V \setminus S} \psi_V^{**} = \sum_{V \setminus S} \frac{\phi_S^{**}}{\phi_S} \psi_V^*$$

$$= \frac{\phi_S^{**}}{\phi_S} \sum_{V \setminus S} \psi_V^*$$

$$= \frac{\phi_S^{**}}{\phi_S} \phi_S^*$$

$$= \phi_S^*$$

$$= \sum_{W \setminus S} \psi_W^{**}$$

Inspecting this derivation, we can see the key steps for achieving consistency. In the forward pass, from $V$ to $W$, the algorithm stores the marginal of the $V$ potential in the separator potential. In the backward pass, from $W$ to $V$, the algorithm divides the $V$ potential by its stored marginal and multiplies the result by the new marginals $\phi_S^{**}$. This latter marginal is the marginal of the $W$ potential. The rescaling equation essentially substitutes one marginal for another, thus making the two clique potentials consistent. This is achieved in the context of a symmetric algorithm that passes information in both directions, and leaves the joint probability distribution invariant.

Consider for example the Markov chain in the following figure.
The corresponding JT is

\[ AB \overset{\phi_B^*}{\longrightarrow} B \overset{\psi_{BC}^*}{\longrightarrow} BC \]

Figure 18.6: JT of 3 state MC

Initially, the clique potential on A,B is \( \psi_{AB} = P(A, B) \), ad the clique potential on B,C is \( \psi_{BC} = P(C|B) \). We choose \( \phi_B = 1 \)

Message passing in the forward direction:

\[
\phi_B^* = \sum_{(A,B) \setminus B} P(A, B) = \sum_A P(A, B) = P(B)
\]

\[
\psi_{BC}^* = \frac{\phi_B^*}{\phi_B} P(C|B) = \frac{P(B)}{1} P(C|B) = P(C, B)
\]

And we see that the clique potentials have become marginal probabilities. In general, we need the finish the backward pass to get the marginal.

Passing backward:

\[
\phi_B^{**} = \sum_C \psi_{BC}^{**} = \sum_C P(C, B) = P(B)
\]

\[
\psi_{AB}^{**} = \frac{\phi_B^{**}}{\phi_B} \psi_{AB}^* = \frac{P(B)}{P(B)} P(A, B) = P(A, B)
\]

\[
\Rightarrow P(A, B, C) = \frac{\psi_{AB}^{**} \psi_{BC}^{**}}{\phi_B^{**}} = \frac{P(A, B) P(B, C)}{P(B)}
\]

Local consistency is achieved.