1. 3.15
Consider the two networks in Figure 3.24. For each of these two networks, $G_a$ and $G_b$, determine whether there can be any other Bayesian network which is I-equivalent to it.

(a)

There is no graph $G_{a2}$ that is I-equivalent to $G_a$.

$G_1$ and $G_2$ are "I-Equivalent" if they imply the same set of conditional independencies. This can be broken down into two requirements:

1. the same skeleton (undirected graph)
2. the same set of immoralities

The following of conditional independencies hold for graph $G_a$:

(i) $A \perp B$
(ii) $A \perp D | C$
(iii) $C \perp D | C$

The unique undirected graph with the same skeleton as $G_a$ is shown below.
Graph $G_a$ has only one immorality, $A \rightarrow C \leftarrow B$. If the undirected graph has this immorality, it will be $G_a$, as changing the direction of any edge $AC$ or $AB$ would remove the immorality, and changing the direction of edge $CD$ would add a new immorality.

No graph I-equivalent to $G_a$ exists.

(b)

$G_b$ has no immoralities. The following graph, $G_{b2}$ has the same skeleton as $G_b$ & has no immoralities. It is found be reversing the direction of the edge $AB$. $G_{b2}$ is I-equivalent to $G_b$.

2. 4.15
Consider a logistic CPD where $X$ is a parent of $Y$.

(1) Assume that $Y$ is $m$-valued, and consider a multinomial logit function defined as:
\[ P(y^j \mid X) = \frac{\exp(w_{0,j} + w_{1,j}X)}{\sum_{j'=1}^{m} \exp(w_{0,j'} + w_{1,j'}X)} \]

This definition uses \(2m\) parameters. Provide an alternative parameterization for this CPD that uses only \((2m-2)\) parameters, and show how to convert from one to the other.

(2)

Now, assume that \(X\) is \(m\)-valued, and consider the parameterization of Equation 4.4:

\[ P(y^j \mid X) = \text{sigmoid}\left(w_0 + \sum_{j=1}^{m} [w_{1,j} \ast 1(X = x^j)]\right) \]

where \(1(X = x^j)\) is equal to 1 if \(X = x^j\) and equal to 0 otherwise.

Once again, this definition uses \(m+1\) parameters. Provide an alternative parameterization for this CDP that uses only \(m\) parameters, and show how to convert from one parameterization to the other.

(1)

We wish to subtract the final weights, \(w_{0,m}\) and \(w_{1,m}\), from all the other weights including the bias. The last weights will then be known to be zero, so we will have two fewer weights. To show that this does not change the distribution, multiply the original distribution by 1, or rather, multiply by: \(\exp(-[w_{0,m} + w_{1,m}]X)\) \(\exp(-[w_{0,m} + w_{1,m}]X)\).

Then distribute the denominator over each term in the sum.

\[ P(y^j \mid X) = \frac{\exp(w_{0,j} + w_{1,j}X)}{\sum_{j'=1}^{m} \exp(w_{0,j'} + w_{1,j'}X)} \ast 1 \]

\[ = \frac{\exp(w_{0,j} + w_{1,j}X)}{\sum_{j'=1}^{m} \exp(w_{0,j'} + w_{1,j'}X)} \ast \exp(-[w_{0,m} + w_{1,m}]X) \]

\[ = \sum_{j'=1}^{m} \exp(w_{0,j'} + w_{1,j'}X) \ast \exp(-[w_{0,m} + w_{1,m}]X) \]

The two factors that are multiplied together can be combined by summing their exponents. The four terms in the exponent can be rearrange into two: a constant and linear term.

\[ P(y^j \mid X) = \frac{\exp\left(\sum_{j'=1}^{m} \left[w_{0,j'} + w_{1,j'}X\right]\right)}{\sum_{j'=1}^{m} \exp\left(\sum_{j'=1}^{m} \left[w_{0,j'} + w_{1,j'}X\right]\right)} \]

\[ = \frac{\exp\left(\sum_{j'=1}^{m} \left[w_{0,j'} - w_{0,m} + w_{1,j'} - w_{1,m}\right]X\right)}{\sum_{j'=1}^{m} \exp\left(\sum_{j'=1}^{m} \left[w_{0,j'} - w_{0,m} + w_{1,j'} - w_{1,m}\right]X\right)} \]

The last \((m^{th})\) term can be pulled out of the sum in the denominator and considered separately. The exponent of this term is zero, so the term is 1.
\[ P(y^j \mid X) = \frac{\exp(w_{0,j} - w_{0,m} + [w_{i,j} - w_{1,m}]X)}{\exp(w_{0,m} - w_{0,m} + [w_{i,j} - w_{1,m}]X) + \sum_{j'=1}^{m-1}\exp(w_{0,j'} - w_{0,m} + [w_{i,j'} - w_{1,m}]X)} \]

\[ = \frac{\exp(w_{0,j} - w_{0,m} + [w_{i,j} - w_{1,m}]X)}{1 + \sum_{j'=1}^{m-1}\exp(w_{0,j'} - w_{0,m} + [w_{i,j'} - w_{1,m}]X)} \]

If we let \( W_{kj} = w_{kj} - w_{km} \) then this equation is of the form:

\[ P(y^j \mid X) = \frac{\exp(W_{0,j} + W_{1,j}X)}{1 + \sum_{j'=1}^{m-1}\exp(W_{0,j'} + W_{1,j'}X)} \]

There are now only \( 2^{(m-1)} \) \( W_{kj} \) terms, with \( k=[0,1] \) and \( j=[1,2,\ldots,m-1] \)

To transform back to a parameterization with \( 2m \) parameters, we keep each \( W_{kj} \) as is and let \( W_{km} = 0 \). If we wish to recover the original values of \( w_{kj} \) we must know the original value of \( w_{km} \) and let \( w_{kj} = W_{kj} + w_{km} \) for all \( kj \).

(2)
If \( X \) is the sole parent of \( Y \), \( X \) is \( m \)-valued, and \( Y \) is binary-valued, \( P(y^j \mid X) \) can be parameterized by Eq. 4.4:

\[ P(y^j \mid X) = \text{sigmoid} \left( w_0 + \sum_{j=1}^{m} [w_{1,j} \ast 1 \{ X = x^j \}] \right) \]

This uses \( m \) parameters for \( w_{1,j} \) and an \((m+1)\)th parameter \( w_0 \). A simple reparameterization can use only \( m \) parameters. Since the sum over \( m \) will contain only one nonzero term, we let \( W_{1,j} = w_{1,j} + w_0 \) for all \( j \) and we can rewrite \( P(y^j \mid X) \) as:

\[ P(y^j \mid X) = \text{sigmoid} \left( \sum_{j=1}^{m} [w_{1,j} \ast 1 \{ X = x^j \}] \right) \]

This has only \( m \) parameters. This can be transformed back into a parameterization with \( m+1 \) parameters by letting \( w_{1,j} = W_{1,j} \) and letting \( w_{0,j} = 0 \). To retrieve the original values of the \( m+1 \) parameter version, we must know the value of \( w_0 \), and let \( w_{1,j} = W_{1,j} - w_0 \) for all \( j \).
5.1 20120

(Proof by contradiction) Assume that $P$ factorizes over $H$:

$$P(X_1, X_2, X_3, X_4) = \frac{\pi_1(X_1, X_2)\pi_2(X_2, X_3)\pi_3(X_3, X_4)\pi_4(X_4, X_1)}{Z}$$

Proposition 1: $P(x_1^1, x_2^0, x_3^0, x_4^1) = 0$ implies that $\pi_1(x_1^1, x_2^1) = 0$ or $\pi_2(x_1^1, x_3^0) = 0$ or $\pi_3(x_3^0, x_4^1) = 0$ or $\pi_4(x_1^1, x_4^1) = 0$ this follows since $P$ factorizes over $H$. On the other hand we also have that:

$$P(x_1^1, x_2^1, x_3^0, x_4^0) = 1/8 \text{ implies that } \pi_1(x_1^1, x_2^1) \neq 0$$
$$P(x_1^1, x_2^0, x_3^0, x_4^0) = 1/8 \text{ implies that } \pi_2(x_2^1, x_3^0) \neq 0$$
$$P(x_1^0, x_2^0, x_3^0, x_4^1) = 1/8 \text{ implies that } \pi_3(x_3^0, x_4^1) \neq 0$$
$$P(x_1^1, x_2^1, x_3^1, x_4^1) = 1/8 \text{ implies that } \pi_4(x_1^1, x_4^1) \neq 0$$

which is a contradiction with the previous proposition.

4. 5.2 20120

In this exercise, you will prove that the modified energy functions $\varepsilon'_1[A,B]$ and $\varepsilon'_2[A,B]$ of Figure 5.9 result in precisely the same distribution as our original energy functions.

More generally, for any constants $\lambda^1$ and $\lambda^0$, we can redefine:

$$\varepsilon'_1[a,b^1] := \varepsilon_1[a,b^1] + \lambda^1$$

$$\varepsilon'_2[a,b^0] := \varepsilon_2[a,b^0] + \lambda^0$$
\( \varepsilon'_2[b^1,c] := \varepsilon_2[b^1,c] - \lambda^i \)

Show that the resulting energy function is equivalent.

The original and alternative energy functions are:

<table>
<thead>
<tr>
<th>(\varepsilon_1[A,B])</th>
<th>(\varepsilon_2[B,C])</th>
<th>(\varepsilon'_1[A,B])</th>
<th>(\varepsilon'_2[B,C])</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a^0) (b^0) (3.4)</td>
<td>(b^0) (c^0) (4.61)</td>
<td>(a^0) (b^0) (4.4)</td>
<td>(b^0) (c^0) (3.61)</td>
</tr>
<tr>
<td>(a^0) (b^1) (1.61)</td>
<td>(b^0) (c^1) (0)</td>
<td>(a^0) (b^1) (1.61)</td>
<td>(b^0) (c^1) (-1)</td>
</tr>
<tr>
<td>(a^1) (b^0) (2.3)</td>
<td>(b^1) (c^0) (0)</td>
<td>(a^1) (b^0) (1)</td>
<td>(b^1) (c^0) (0)</td>
</tr>
</tbody>
</table>

The joint probability distribution \(P\) is proportional to \(\exp(\Sigma_k \varepsilon_k)\), where the energy functions \(\varepsilon_k\) are defined for each clique. So long as \(\Sigma_k \varepsilon_k\) is the same, the individual terms \(\varepsilon_k\) can be modified as above without changing the distribution.

\(\varepsilon'_1 + \varepsilon'_2 = [\varepsilon_1 + \lambda] + [\varepsilon_2 - \lambda] = [\varepsilon_1 + \varepsilon_2] + [\lambda - \lambda] = [\varepsilon_1 + \varepsilon_2] + [0] = \varepsilon_1 + \varepsilon_2\)

5. 5.10

We define the following properties for a set of independencies:

**Strong Union:**

\((X \perp Y|Z) \rightarrow (X \perp Y|Z,W)\)

*Additional evidence W cannot induce dependence*

**Transitivity:**

\(\neg(X \perp A|Z) \& \neg(A \perp Y|Z) \rightarrow \neg(X \perp Y|Z)\) for all disjoint \(X,Y,Z,A\)

If \(X\) and \(Y\) are both correlated with \(A\), then they are also correlated with each other.

The contrapositive of this is also true: \((X \perp Y|Z) \rightarrow (X \perp A|Z)\) or \((A \perp Y|Z)\)

Prove that if \(I=I(H)\) for some markov network \(H\), then \(I\) satisfies the Strong Union and Transitive Properties

**Strong Union:** \((X \perp Y|Z) \rightarrow (X \perp Y|Z,W)\)

In a Markov Network, \((X \perp Y|Z)\) iff \(X\) and \(Y\) are have no active path between them, that is, if there is no series of edges from \(X\) to \(v_1\) to \(v_2\) to ... to \(v_N\) to \(Y\) going only to vertices \(V\) not in \(Z\).

If \((X \perp Y|Z)\), then there is no active path. Conditioning on additional variables \(W\) can only block paths, it cannot unblock them. Thus \((X \perp Y|Z) \rightarrow (X \perp Y|Z,W)\).

**Transitivity:** \(\neg(X \perp A|Z) \& \neg(A \perp Y|Z) \rightarrow \neg(X \perp Y|Z)\)

If \(\neg(X \perp A|Z)\) then there is some active path from \(X\) to \(A\) when conditioning on \(Z\).
If \( \neg(A \perp Y | Z) \) then there is some active path from A to Y when conditioning on Z.

In a markov network, if there is an active path from X to A and A to Y, then there is an active path from X to Y, going through A. Since A is not in Z (X,Y,A,Z are assumed disjoint).

Since there is an active path from X to Y when conditioning on Z, X and Y are not independent given Z.

Thus \( \neg(X \perp A | Z) \) & \( \neg(A \perp Y | Z) \) \( \Rightarrow \) \( \neg(X \perp Y | Z) \)