A: Warm-Up

**A1. Conditional Independence**

Let $X, Y$ be independent random variables taking the value 0 with probability $\frac{1}{2}$, and 1 otherwise. Now let $Z$ be a random variable with conditional distribution

$$P(Z = 1 \mid X = x, Y = y) = \begin{cases} \frac{3}{4} & \text{if } x = y \\ \frac{1}{4} & \text{if } x \neq y \end{cases}$$

with $P(Z = 0 \mid X = x, Y = y) = 1 - P(Z = 1 \mid X = x, Y = y)$.

(a) Find $P(X, Y \mid Z = 1)$.

By symmetry we can see that $P(Z = 1) = \frac{1}{2}$, so

$$P(X = x, Y = y \mid Z = 1) = 2P(X = x)P(Y = y)P(Z = 1 \mid X = x, Y = y)$$

$$= \frac{1}{4} + \frac{1}{2}I\{x = y\} = \frac{1}{2} + \frac{1}{2}I\{x = y\} \cdot \frac{1}{8}.$$ 

That is, $\frac{3}{8}$ if $x = y$, $\frac{1}{8}$ otherwise.

Importantly, we see that $X \perp \perp Y \mid Z$ even though $X \perp \perp Y$.

(b) Show that $X \perp \perp Z$ and $Y \perp \perp Z$ but that $X, Y \not\perp \perp Z$.

For the converse, it is easy to check that if Theorem 2.4 holds, plugging in its expression gives us the equality above.

**A2.** Show that $X \perp \perp Y \mid Z$ is equivalent to

$$p(x, y, z) \cdot p(x', y', z) = p(x', y, z) \cdot p(x, y', z)$$

for $Z$-almost all $x, x', y, y', z$.

Given $z$ with $p(z) > 0$, there exist $x', y'$ such that $p(x', y', z) > 0$; divide both sides by $p(x', y', z)$ and treat $x', y'$ as constants to as that $p$ factorizes as in Theorem 2.4(v). For the converse, it is easy to check that if Theorem 2.4(v) holds, plugging in its expression gives us the equality above.

B: Core Questions

**B1. Graphoids.**

By completing the implications of the theorem from lectures, show that

$$X \perp \perp Y, W \mid Z \iff X \perp \perp W \mid Z \text{ and } X \perp \perp Y \mid W, Z.$$ 

Property 3 is immediate from Theorem 2.4, since $X \perp \perp Y, W \mid Z$ implies that $p(x, y, w, z) = f(x, z)g(y, w, z)$, which implies $X \perp \perp W \mid Y, Z$. For property 4, the left hand sides suggest that

$$p(w \mid x, y, z) = p(w \mid y, z) \quad p(y \mid x, z) = p(y \mid z),$$
so multiplying these together gives \( p(w, y \mid x, z) = p(w, y \mid z) \) as required.

Show that, in general,

\[
X \perp Y \mid Z \text{ and } X \perp Z \mid Y \iff X \perp Y, Z.
\]

[Hint: \( X \perp Y \mid Y \) for any \( X, Y \).]

Choosing \( Y \) and \( Z \) to have a degenerate distribution and \( X \) correlated with them generally does the trick. A simple example: let \( P(X = Y = Z) = 1 \) with \( P(X = 0) = P(X = 1) = \frac{1}{2} \).

**B2. Some Strange Independences.**

(a) Let \((X_1, X_2, X_3)\) follow a multivariate Gaussian distribution with covariance matrix \( \Sigma \). Show that \( X_1 \perp \perp X_2 \mid X_3 \) if and only if

\[
\sigma_{33}\sigma_{12} - \sigma_{13}\sigma_{23} = 0.
\]

We know from lectures that the independence holds if and only if \( k_{12} = 0 \), where \( K = \Sigma^{-1} \) is the concentration matrix. The quantity above is the corresponding entry in the adjoint matrix for \( \Sigma \), and is therefore the same as \( k_{12} \) up to the (positive) determinant of \( \Sigma \). Hence they are zero at the same time.

(b) Deduce that for Gaussian random variables, \( X_1 \perp \perp X_2 \mid X_3 \) and \( X_1 \perp \perp X_2 \iff X_1 \perp \perp X_2, X_3 \) or \( X_2 \perp \perp X_1, X_3 \).

(c) Let \( X, Y \) be discrete random variables. Show that \( X \perp \perp Y \mid Z = z \) if and only if the matrix \( M_{xy} = (\pi_{xyz})_{x,y} \) has rank one.

[Recall that a matrix \( M \) has rank one if and only if it can be written as \( M = \alpha\beta^T \) for vectors \( \alpha, \beta \).]

We know that this is equivalent to \( p(x, y, z) = p(x, z)p(y \mid z) \), which implies that \( M_{xy} = (p(x, y, z))_{x,y} \) is the outer product of two vectors. Hence \( M_{xy} \) has rank 1. For the converse, if \( M_{xy} \) has rank 1 then it is the outer product of two vectors, so \( M_{xy} = \alpha_x\beta_y \) and, from lectures, this is equivalent to the original representation.

(d) Let \( A, B \) be \( a \times c \) and \( b \times c \) matrices each of rank \( c \). Show that \( AB^T \) also has rank \( c \).

Note that \( a, b \geq c \) in order for \( A \) and \( B \) to have rank \( c \). The column span of \( AB^T \) is the same as that of \( A \), so the rank is clearly at most \( c \). On the other hand since \( B^T \) has full column rank \( c \), by selecting the correct vector \( x \in \mathbb{R}^b \) it is clear that \( B^T x \) spans \( \mathbb{R}^c \). Hence \( AB^T x \) has the same column span as \( A \).

(e) Hence, or otherwise, show that for binary \( Z \) and finite discrete \( X, Y \) we have

\[
X \perp \perp Y \mid Z \text{ and } X \perp \perp Y \iff X \perp Y, Z \text{ or } Y \perp \perp X, Z.
\]

[Hint: show that if \( X \perp \perp Y \mid Z \), then \((\pi_{xy})_{xy} \) can be written as a product of matrices of rank 2].

We obtain

\[
\pi_{xy} = \pi_{xy0} + \pi_{xy1} = \alpha_0^0\beta_0^0 + \alpha_1^1\beta_1^1 = \left( \begin{array}{c} \alpha_0^0 \\ \alpha_1^1 \end{array} \right) \left( \begin{array}{c} \beta_0^0 \\ \beta_1^0 \end{array} \right);
\]
Hence the matrix $\pi_{xy+}$ can be written as

$$\pi_{xy+} = \begin{pmatrix} \alpha^0 & \alpha^1 \end{pmatrix} \begin{pmatrix} \beta^0 & \beta^1 \end{pmatrix}^T = AB^T.$$  

where $\alpha^0 = (\alpha_x^0)_x$ and $\beta^0 = (\beta_y^0)_y$ are vectors. Then $(\pi_{xy+})_{xy}$ is a product of matrices $A$ and $B$ with 2 columns, and so by the previous part has rank 2 unless either $A$ or $B$ has rank 1.

But $A$ being rank 1 means that $\alpha^0$ and $\alpha^1$ are proportional, i.e. $p(x \mid z = 0) = p(x \mid z = 1)$, so $X \perp Z$. Combining with $X \perp Y \mid Z$ gives $X \perp Y, Z$. If $B$ is rank 1 then we obtain $Y \perp X, Z$ similarly.

B3. Factorization and Conditional Independence.

Consider four binary variables $A, B, C, D$; let the support (i.e. the set of combinations whose probability is $> 0$) of these variables be:

$$(a, b, c, d) = (0, 0, 0, 0) \quad (1, 0, 0, 0) \quad (1, 1, 0, 0) \quad (1, 1, 1, 0)$$
$$(1, 1, 1, 1) \quad (0, 1, 1, 1) \quad (0, 0, 1, 1) \quad (0, 0, 0, 1).$$

(a) Show that $A \perp C \mid B, D$ and $B \perp D \mid A, C$ for any distribution with support in this set.

Note, for example, that if $B = D = 0$ we always have $C = 0$. Hence $C$ is a.s. constant on this set and thus trivially independent of $A$. Similar observations hold for other combinations of $B, D$ and also for the other conditional independence.

(b) Show that we cannot write the joint distribution in the form

$$P(A = a, B = b, C = c, D = d) = \psi_{ab}(a, b) \cdot \psi_{bc}(b, c) \cdot \psi_{cd}(c, d) \cdot \psi_{da}(d, a).$$

The key here is that for any pair of numbers in a factor (say $A, B$), every combination of their values is possible. It therefore follows that the factor $\psi_{ab}$ is non-zero for every $a, b$. But this means that $p$ is non-zero for every combination of values $a, b, c, d$, which contradicts the restriction on the support of $p$. 

3
C: Optional

C1. Möbius Inversion

(a) Let \((\zeta_M)_{M \subseteq V}\) be a vector indexed by subsets, and let
\[
\eta_M = \sum_{Z \subseteq M} \zeta_Z, \quad \forall M \subseteq V.
\]
Show the Möbius inversion formula:
\[
\zeta_M = \sum_{Z \subseteq M} (-1)^{|M\setminus Z|} \eta_Z, \quad \forall M \subseteq V.
\]
Deduce that \(\eta_M = 0\) for all \(M\) if and only if \(\zeta_M = 0\) for all \(M\).

[Hint: any non-empty set \(A\) has the same number of even-sized subsets as odd-sized subsets.]

Using the definition,
\[
\sum_{Z: Z \subseteq M} (-1)^{|M\setminus Z|} \eta_Z = \sum_{Z: Z \subseteq M} (-1)^{|M\setminus Z|} \sum_{W: W \subseteq Z} \zeta_W
\]
\[= \sum_{Z, W: W \subseteq Z \subseteq M} (-1)^{|M\setminus Z|} \zeta_W
\]
\[= \sum_{W \subseteq M} \sum_{Z: Z \subseteq M \setminus W} (-1)^{|(M\setminus W)\setminus Z'|} \zeta_W
\]
\[= \sum_{W \subseteq M} (-1)^{|M \setminus W|} \zeta_W \sum_{Z: Z \subseteq M \setminus W} (-1)^{|Z'|}
\]

If \(W = M\) then the inner sum is 1. Otherwise, by the hint, there as as many even as odd sized subsets of \(M \setminus W\), so the sum is 0. The only non-zero term corresponds to \(\zeta_M\), which gives the result.

(b) Now let \(X_V\) be binary variables with joint distribution
\[
\log p(x_V) = \sum_{A \subseteq V} \lambda_A(x_A)
\]
using the identifiability constraints from lectures. Let \(a, b \in V\) and \(W = V \setminus \{a, b\}\). By considering
\[
\log p(x_a, x_b, x_W) + \log p(1, 1, x_W) - \log p(1, x_b, x_W) - \log p(x_a, 1, x_W)
\]
or otherwise, show that the joint distribution factorizes as \(p(x_V) = f(x_a, x_W)g(x_b, x_W)\) if and only if \(\lambda_{abD} = 0\) for all \(D \subseteq W\).

If \(\log p\) can be written that way then the result is obvious. Conversely if \(p\) factorizes in that way then
\[
\sum_{D \subseteq W} \{\lambda_{abD}(x_a, x_b, x_D) + \lambda_{abD}(1, 1, x_D) - \lambda_{abD}(1, x_b, x_D) - \lambda_{abD}(x_a, 1, x_D)\} = 0
\]
\[
\sum_{D \subseteq W} \lambda_{abD}(x_a, x_b, x_D) = 0
\]
Now, setting \(\zeta_D = \lambda_{abD}(x_a, x_b, x_D)\), we find that the expression above holds if and only if \(\zeta_D = 0\) for all \(D\), as required.
(c) Deduce that a positive distribution \( p(x_V) \) on binary variables is Markov with respect to an undirected graph \( G \) if and only if \( \lambda_A = 0 \) whenever \( A \) is not a complete set of vertices in \( G \).

The previous part shows that \( X_a \perp X_b \mid X_W \) if and only if \( \lambda_A = 0 \) for \( \{a, b\} \subseteq A \).

Extending this to the entire pairwise property shows that this must hold for every missing edge. Hence \( \lambda_A = 0 \) unless every pair of vertices in \( A \) is joined by an edge in \( G \); that is, unless \( A \) is complete.

(d) Extend the result to arbitrary discrete variables.

This is essentially the same as (b), but one must do it for every other level of \( X_V \).

C2. Conditional Expectation. [Involves some measure theory, though could be ‘proved’ without knowing it]. Given an integrable random variable \( X \) and two other random variables \( Y, Z \), we say that \( X \) is conditionally independent of \( Y \) given \( Z \) if for any integrable \( f(X) \) we have

\[
\mathbb{E}[f(X) \mid Y, Z] = \mathbb{E}[f(X) \mid Z] \quad \text{(a.s.)}
\]

Equivalently,

\[
\mathbb{E}[f(X, Z) \mid Y, Z] = \mathbb{E}[f(X, Z) \mid Z] \quad \text{(a.s.)}
\]

[Why is this equivalent?]

Consider the following alternative statements.

A. \( \mathbb{E}[f(X, Z)g(Y, Z)] = \mathbb{E}[\mathbb{E}[f(X, Z) \mid Z]\mathbb{E}[g(Y, Z) \mid Z]] \) for all integrable \( f, g \).

B. \( \mathbb{E}[f(X, Z)g(Y, Z) \mid Z] = \mathbb{E}[f(X, Z) \mid Z]\mathbb{E}[g(Y, Z) \mid Z] \) a.s. for all integrable \( f, g \).

Show that A and B are equivalent to one another, and also to the definition of conditional independence.

[Hint: you will need the tower property: \( \mathbb{E}[X \mid Y] = \mathbb{E}[\mathbb{E}[X \mid Y, Z] \mid Y] \) holds for any \( Y, Z \) and integrable \( X \), and ‘taking out what is known’: \( \mathbb{E}[f(X)g(Z) \mid Z] = g(Z)\mathbb{E}[f(X) \mid Z].\)]

For \( CI \implies A \), first assume without loss of generality that \( \mathbb{E}[f(X, Z) \mid Z] = 0 \) (else replace \( f \) with \( \tilde{f} = f - \mathbb{E}[f \mid Z] \)). Then we have:

\[
\mathbb{E}[f(X, Z)g(Y, Z)] = \mathbb{E}[\mathbb{E}[f(X, Z)g(Y, Z) \mid Y, Z] \mid Z]
\]

\[
= \mathbb{E}[\mathbb{E}[f(X, Z) \mid Y, Z]g(Y, Z) \mid Z] \quad \text{(g is \( Y, Z \)-measurable)}
\]

\[
= \mathbb{E}[\mathbb{E}[f(X, Z) \mid Z]g(Y, Z) \mid Z] \quad \text{(by conditional independence)}
\]

\[
= 0 \quad \text{(by assumption)}.
\]

as required.

For \( A \implies B \), define \( h(Z) = \mathbb{E}[f(X, Z)g(Y, Z) \mid Z] \), which we must show is almost surely zero. Then

\[
h(Z) = h(Z)1_{\{h(Z) \geq 0\}} + h(Z)1_{\{h(Z) \leq 0\}}
\]

\[
= \mathbb{E}[f(X, Z)g(Y, Z)1_{\{h(Z) \geq 0\}} \mid Z] + \mathbb{E}[f(X, Z)g(Y, Z)1_{\{h(Z) \leq 0\}} \mid Z]
\]

\[
\equiv h^+(X) + h^-(Z).
\]
Note that \( g(Y, Z) \mathbbm{1}_{\{h(Z) \geq 0\}} \) is smaller in magnitude than \( g \), and hence is also integrable. Now using the factorization,

\[
\mathbb{E}h^+(X) = \mathbb{E}[f(X, Z)g(Y, Z) \mathbbm{1}_{\{h(Z) \geq 0\}}] \\
= \mathbb{E}[\mathbb{E}[f(X, Z) | Z] \cdot \mathbb{E}g(Y, Z) \mathbbm{1}_{\{h(Z) \geq 0\}} | Z] \\
= 0.
\]

But since \( h^+(Z) \geq 0 \), this means \( h^+(Z) = 0 \) almost surely. Similarly for \( h^- \) and therefore \( h \).

Finally consider \( B \implies CI \). Let \( g(Y, Z) = \mathbbm{1}_{\{\mathbb{E}[f|Y,Z] \geq \mathbb{E}[f|Z]\}} \). Then, using \( B \),

\[
\mathbb{E}[f | Z]\mathbb{E}[g | Z] = \mathbb{E}[fg | Z] = \mathbb{E}[\mathbb{E}[f | Y, Z]g | Z] \geq \mathbb{E}[\mathbb{E}[f | Z]g | Z] = \mathbb{E}[f | Z]\mathbb{E}[g | Z].
\]

This shows that

\[
\mathbb{E}[(\mathbb{E}[f | Y, Z] - \mathbb{E}[f | Z])g | Z] = 0 \quad \text{a.s.}
\]

since the integrand is non-negative, this shows that it is zero almost surely; a similar approach with \( g \) replaced by \( 1 - g \) gives us the result.