Causal Inference for Health Data (Winter 2026)
STATS C160/C260
Homework 1 Solutions

Exercise 1. Basic probabilities

Seventy percent of cancer cases in a certain population are diagnosed in an early stage. Of those diagnosed early, 60% of the patients went to routine consultations twice a year, whereas 90% of the patients diagnosed late did not.

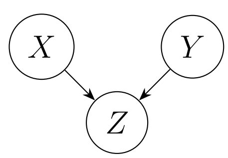
(a) Suppose a certain person has developed cancer and goes to routine consultations. What is the probability that cancer will be diagnosed early?

Let E be being diagnosed early and C going to consultation regularly. From the problem we have $P(C = 1 \mid E = 1) = 0.6, P(E = 0 \mid E = 0) = 0.9, P(E = 1) = 0.7$

$$\begin{aligned} P(E = 1 \mid C = 1) &= \frac{P(E = 1, C = 1)}{P(C = 1)} = \frac{P(C = 1 \mid E = 1)P(E = 1)}{P(C = 1)} \\ &= \frac{P(C = 1 \mid E = 1)P(E = 1)}{P(C = 1 \mid E = 1)P(E = 1) + P(C = 1 \mid E = 0)P(E = 0)} \\ &= \frac{(0.6)(0.7)}{(0.6)(0.7) + (1 - 0.9)(0.3)} = \frac{14}{15} \approx 0.933 \end{aligned}$$

(b) Construct a probability distribution over three random variables X, Y, Z such that $(X \perp\!\!\!\perp Y)$ but $(X \perp\!\!\!\perp Y \mid Z)$ does not hold. You can either describe the full joint distribution or their conditionals.

Create a distribution that factorizes according to the following graph



Pick parameters for the distributions $P(X), P(Y)$ and $P(Z \mid X, Y)$. As long as the parameters for the latter ($P(Z = 1 \mid x, y)$ for all x and y) do not repeat, the independence relationships required will hold. All that is left is to show that $P(x \mid y) = P(x)P(y)$ and $P(x \mid y, z) \neq P(x \mid z)$ by computing those quantities with the chosen parametrization.

Exercise 2. Estimation and Independence Relations

Consider random variables X_1, X_2, and Y and assume our goal is to compute the query $Q = P(y \mid x_1, x_2)$. We do not have any prior information about the conditional independence relations among these variables.

(a) For every one of the following sets of distributions, show how to compute the query Q based on them, or explain why this is not possible:

1. $P(x_1, x_2), P(y), P(x_1 \mid y)$, and $P(x_2 \mid y)$
2. $P(x_1, x_2), P(y)$, and $P(x_1, x_2 \mid y)$
3. $P(x_1 \mid y), P(x_2 \mid y)$, and $P(y)$
4. $P(x_1), P(x_2)$, and $P(x_1, x_2 \mid y)$
5. $P(x_1), P(x_2), P(x_1 \mid y)$, and $P(x_2 \mid y)$

First consider the alternative form of the query:

$$P(y \mid x_1, x_2) = \frac{P(y, x_1, x_2)}{P(x_1, x_2)} = \frac{P(y)P(x_1, x_2 \mid y)}{P(x_1, x_2)}$$

- No, the term $P(x_1, x_2 \mid y)$ is required and is not estimable from the given distributions.
- Yes, this case provides all the distributions required as in the second equality above.
- No, the term $P(x_1, x_2)$ is not estimable from the given distributions.
- No, the denominator in the equalities above is not estimable, and $P(y)$ cannot be obtained either.
- No, neither the numerator or denominator of the expressions above is estimable from the input.

(b) Suppose we learned that $(X_1 \perp\!\!\!\perp X_2 \mid Y)$ holds in P. Now, which of the sets before are sufficient to compute the query Q? Show how or explain why it is not possible.

From the given independence the following identity follows:

$$P(x_1, x_2 \mid y) = P(x_1 \mid y)P(x_2 \mid y)$$

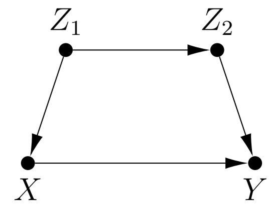
$$P(y \mid x_1, x_2) = \frac{P(y, x_1, x_2)}{P(x_1, x_2)} = \frac{P(y)P(x_1, x_2 \mid y)}{P(x_1, x_2)}$$

- Yes, now $P(x_1, x_2 \mid y)$ is available as well as the other terms in the second equality.
- Yes, same as before.
- Yes, compute $P(x_1, x_2) = \sum_Y P(x_1, x_2, y) = \sum_Y P(x_1, x_2 \mid y)P(y)$ and $P(x_1, x_2 \mid y)$ as shown above. Then, all terms in the second expression from the previous question are estimable.
- No, $P(y)$ is needed and cannot be obtained from the given distributions.
- No, after considering $P(x_1, x_2 \mid y) = P(x_1 \mid y)P(x_2 \mid y)$ this is equivalent to the previous item and the same reason applies.

Exercise 3. Query Estimation

Consider the following graphical model $\mathcal{G}$ below:

Suppose we want to compute the query $Q = \sum_{Z_1} P(y \mid x, z_1)P(z_1)$.



(a) Is $Q = P(y \mid x)$ in $\mathcal{G}$? Justify your answer.

No. Consider $P(y \mid x)$ after conditioning on Z_1:

$$P(y \mid x) = \sum_{Z_1} P(y \mid x, z_1) P(z_1 \mid x)$$

Equality holds when $P(z_1) = P(z_1 \mid x)$ which is not guaranteed by the independences readable in the model.

(b) Suppose we have access to the marginal distribution $P(X, Y, Z_2)$. Is it possible to estimate Q? If so, show how to do it. Otherwise, explain why that is not the case.

$$\begin{aligned}
Q &= \sum_{Z_1} P(y \mid x, z_1) P(z_1) \\
&= \sum_{Z_1, Z_2} P(y \mid x, z_1, z_2) P(z_2 \mid x, z_1) P(z_1) && \text{Condition on } Z_2 \\
&= \sum_{Z_1, Z_2} P(y \mid x, z_2) P(z_2 \mid x, z_1) P(z_1) && (Y \perp\!\!\!\perp Z_1 \mid Z_2, X) \\
&= \sum_{Z_1, Z_2} P(y \mid x, z_2) P(z_2 \mid z_1) P(z_1) && (Z_2 \perp\!\!\!\perp X \mid Z_1) \\
&= \sum_{Z_2} P(y \mid x, z_2) \sum_{Z_1} P(z_2, z_1) && \text{Chain rule and move summation in} \\
&= \sum_{Z_2} P(y \mid x, z_2) P(z_2) && \text{Sum out } Z_1
\end{aligned}$$

The last expression is equal to Q and is estimable from the available data.

Exercise 4. Specifying Structural Causal Models

The following is a description of a clinical decision support (CDS) alert system used in a hospital:

- The delivery channel location of the alert (L) could be via the clinician's *phone* or via the *EHR system* and is chosen at random every time with equal likelihood.
- Based on the clinician's age group (A), either a *text-only* alert or a *text+image* alert is used as the alert modality (M). The age group is either '*less than 40*' or '*40 or more*'.

- The clinician will see (S) the alert with the following probabilities:
  1. If $L = \textit{phone}$ with probability 1/3,
  2. if $L = \textit{EHR system}$, $A =$'*less than 40*' with probability 1/5, and
  3. if $L = \textit{EHR system}$, $A =$'*40 or more*' with probability 1/6.
- The clinician will judge the alert to be clinically relevant (I) 60% of the time if the modality is *text+image*, or 40% of the time if the modality is *text-only*.
- The clinician has permission to administer the drug (D) in 40% of cases.
- The clinician will administer the drug (C) if the alert is seen, judged relevant, and the clinician has permission to administer the drug.

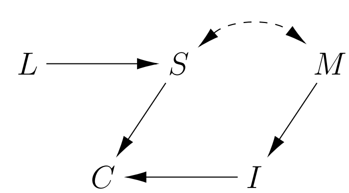
Variables L, S, M, I and C are observable, while A and D are unobservable (other unobservables, which are not mentioned, may be present as well).

(a) Specify a structural causal model $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ that captures this setting. Make a reasonable choice for the distribution of A and the way M is determined.

$$\mathbf{V} = \{L, S, M, I, C\},$$
$$\mathbf{U} = \{A, D, U_l, U_m, U_{s1}, U_{s2}, U_{s3}, U_{i1}, U_{i2}\},$$
$$\mathcal{F} = \begin{cases} f_L = U_l \\ f_M = A \oplus U_m \\ f_S = (L \wedge U_{s1}) \vee (\neg L \wedge ((A \wedge U_{s2}) \vee (\neg A \wedge U_{s3}))) \\ f_I = (M \wedge U_{i1}) \vee (\neg M \wedge U_{i2}) \\ f_C = S \wedge I \wedge D \end{cases},$$
$$P(\mathbf{u}) = \{P(D = 1) = 0.4, P(A = 1) = 0.2, P(U_l = 1) = 1/2, P(U_m = 1) = 1/4,$$
$$P(U_{s1} = 1) = 1/3, P(U_{s2} = 1) = 1/5, P(U_{s3} = 1) = 1/6,$$
$$P(U_{i1} = 1) = 0.6, P(U_{i2} = 1) = 0.4\};$$

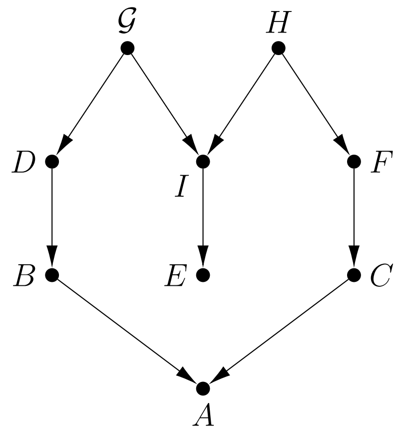
$L = 1$ represents '*phone*' and $L = 0$ represents '*EHR system*'; $A = 1$ represents '*40 or more*' and $A = 0$ represents '*less than 40*'; $S = 1$ represents '*alert seen*'; $I = 1$ represents '*alert judged clinically relevant*'; $D = 1$ represents '*has permission to administer the drug*'; $C = 1$ represents '*drug administered*'.

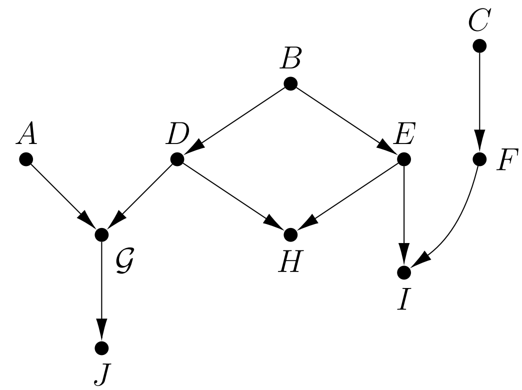
(b) Draw the causal diagram corresponding to the given SCM.



Exercise 5. d-Connectedness

Consider the two graphs below, $\mathcal{G}$ (left) and $\mathcal{G}'$ (right).





(a) List the variables that are d-connected to A given $\{B\}$ in $\mathcal{G}$.

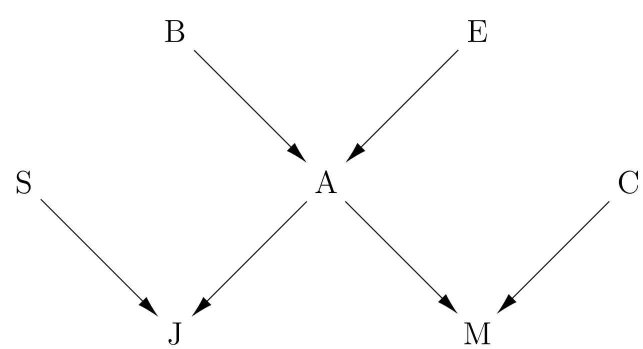
$\{C, F, H, I, E\}$

(b) List the variables that are d-connected to A given $\{J\}$ in $\mathcal{G}'$.

$\{G, D, B, E, H, I\}$

Exercise 6. d-Separation

Consider the following graphical model,



and the conditional probability tables:

P(B=1)	P(E=1)	P(S=1)	P(C=1)
0.02	0.01	0.80	0.15

B	E	$P(A=1\|BE)$	A	C	$P(M=1\|AC)$	A	S	$P(J=1\|AS)$
0	0	0.01	0	0	0.05	0	0	0
0	1	0.3	0	1	0	0	1	0
1	0	0.9	1	0	0.85	1	0	0.97
1	1	0.98	1	1	0.15	1	1	0.1

(a) List all d-separation statements that hold assuming that $J = 1$.

$(S \perp\!\!\!\perp B, E \mid J, A)$,
$(C \perp\!\!\!\perp B, E, A, S \mid J)$,
$(M \perp\!\!\!\perp B, E, S \mid J, A, C)$
Any other d-separation statements (with J observed) can be derived from those above using the graphoid axioms.

(b) Compute the given probabilities using the graph and the probability tables:

(i) $P(M = 1)$

To solve this, we first compute $P(A)$.

$$\begin{aligned} P(A = 1) = \sum_{b,e} P(A = 1|b, e)P(b)P(e) &= 0.02 \times 0.01 \times 0.98 \\ &+ 0.02 \times 0.99 \times 0.9 \\ &+ 0.98 \times 0.01 \times 0.3 \\ &+ 0.98 \times 0.99 \times 0.01 \\ &= 0.03 \end{aligned}$$

Next, we compute $P(M)$

$$\begin{aligned} P(M = 1) = \sum_{a,c} P(M = 1|a, c)P(a)P(c) &= 0.05 \times 0.97 \times 0.85 \\ &+ 0.85 \times 0.03 \times 0.85 \\ &+ 0.15 \times 0.03 \times 0.15 \\ &= 0.064 \end{aligned}$$

(ii) $P(J = 1|C = 0)$

J is independent of C, so this is just $P(J = 1)$

$$\begin{aligned} P(J = 1) &= \sum_{s,a} P(J = 1 \mid a, s)P(a, s) \\ &= \sum_{s,a} P(J = 1 \mid a, s)P(a)P(s) \\ &= 0 + 0 + (0.97)(0.03)(0.2) + (0.1)(0.03)(0.8) \\ &= 0.00822 \end{aligned}$$

(iii) $P(E = 1|M = 1, B = 0)$

Let's consider $P(e, M=1, B=0)$ first:

$$\begin{aligned}
P(e, M=1, B=0) &= \sum_{a,c} P(e, a, c, M=1, B=0) \\
&= \sum_{a,c} P(e)P(B=0)P(a \mid e, B=0)P(M=1 \mid a, c)P(c) \\
&= P(e)P(B=0) \sum_{a} P(a \mid e, B=0) \sum_{c} P(M=1 \mid a, c)P(c)
\end{aligned}$$

First compute $P(M=1 \mid a)$:

$$\begin{aligned}
P(M=1 \mid A=0) &= (0.05)(0.85) + (0)(0.15) = 0.0425 \\
P(M=1 \mid A=1) &= (0.85)(0.85) + (0.15)(0.15) = 0.745
\end{aligned}$$

For $E=0$ and $E=1$,

$$\begin{aligned}
P(E=0, M=1, B=0) &= (0.99)(0.98)[(0.99)(0.0425) + (0.01)(0.745)] = 0.048049155 \\
P(E=1, M=1, B=0) &= (0.01)(0.98)[(0.7)(0.0425) + (0.3)(0.745)] = 0.00248185.
\end{aligned}$$

Then,

$$\begin{aligned}
P(E=1 \mid M=1, B=0) &= \frac{P(E=1, M=1, B=0)}{P(E=0, M=1, B=0) + P(E=1, M=1, B=0)} \\
&= 0.04911538965.
\end{aligned}$$

(iv) $P(M=1|B=1, J=0)$

First consider $P(m, B=1, J=0)$, then

$$\begin{aligned}
&P(m, B=1, J=0) \\
&= \sum_{e,s,a,c} P(B=1, e, s, a, c, J=0, m) \\
&= \sum_{e,s,a,c} P(B=1)P(e)P(a \mid B=1, e)P(s)P(c)P(J=0 \mid a, s)P(m \mid a, c) \\
&= P(B=1) \sum_{a} \left(\sum_{e} P(a \mid B=1, e)P(e) \right) \left(\sum_{c} P(m \mid a, c)P(c) \right) \left(\sum_{s} P(J=0 \mid a, s)P(s) \right).
\end{aligned}$$

It is convenient to first compute the distributions in parenthesis:

$$\begin{aligned}
P(J=0 \mid A=0) &= (1)(0.2) + (1)(0.8) = 1 \\
P(J=0 \mid A=1) &= (0.03)(0.2) + (0.9)(0.8) = 0.726
\end{aligned}$$

$$\begin{aligned}
P(M=0 \mid A=0) &= (0.95)(0.85) + (1)(0.15) = 0.9575 \\
P(M=0 \mid A=1) &= (0.15)(0.85) + (0.85)(0.15) = 0.255 \\
P(M=1 \mid A=0) &= (0.05)(0.85) + (0)(0.15) = 0.0425 \\
P(M=1 \mid A=1) &= (0.85)(0.85) + (0.15)(0.15) = 0.745
\end{aligned}$$

$$\begin{aligned}
P(A=0 \mid B=1) &= (0.1)(0.99) + (0.02)(0.01) = 0.0992 \\
P(A=1 \mid B=1) &= (0.9)(0.99) + (0.98)(0.01) = 0.90080.
\end{aligned}$$

Back to the joint distribution:

$$P(m, B=1, J=0) = P(B=1) \sum_a P(a \mid B=1) P(m \mid a) P(J=0 \mid a)$$

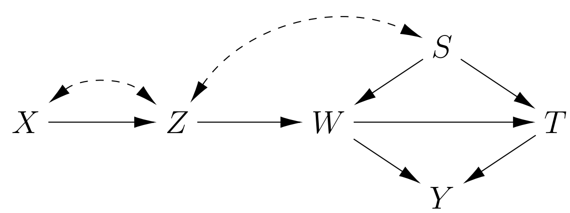
$$\begin{aligned} P(M=0, B=1, J=0) &= (0.02)[(0.0992)(0.9575)(1) + (0.90080)(0.255)(0.726)] \\ &= 0.00523498208 \\ P(M=1, B=1, J=0) &= (0.02)[(0.0992)(0.0425)(1) + (0.90080)(0.745)(0.726)] \\ &= 0.00982863392. \end{aligned}$$

Finally,

$$\begin{aligned} P(M=1 \mid B=1, J=0) &= \frac{P(M=1, B=1, J=0)}{P(M=0, B=1, J=0) + P(M=1, B=1, J=0)} \\ &= 0.6524750711. \end{aligned}$$

Exercise 7. d-Separating Sets

Let $\mathcal{G}$ (shown below) be the causal diagram of some unknown model $\mathcal{M}$ and let P be $\mathcal{M}$'s observational distribution.



(a) Find a minimal set $\mathbf{A}$ (if it exists) that d-separates X and W.

$\mathbf{A} = \{Z, S\}$

(b) Find a minimal set $\mathbf{A}$ (if it exists) that d-separates X and S.

$\mathbf{A} = \emptyset$

(c) Find **all** minimal set $\mathbf{A}$ (if any exists) that d-separate Z and Y.

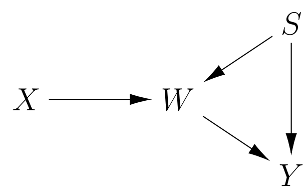
$\mathbf{A} = \{W, S\}, \{W, T\}$

(d) [Harder] Draw a graph $\mathcal{G}'$ over the variables $\{X, S, W, Y\}$ such that

- $\mathcal{G}'$ has exactly the same independence as $\mathcal{G}$ with respect to $P(X, S, W, Y)$, and
- $\mathcal{G}'$ has the minimum number of edges while satisfying the constraint in the previous bullet.

Hint: In the first step, determine which graph is obtained if Z and T are unobserved, and list the independences implied by this graph. In the second step, consider which of the edges can be removed without affecting any of the independences.

The independences that we can derive from $\mathcal{G}$ involving $\{X, S, W, Y\}$ is just $(S \perp\!\!\!\perp X)$ and $X \perp\!\!\!\perp Y \mid S, W$.



Exercise 8. [Optional] d-Separation Theory

Let $\mathcal{G}$ be a causal diagram (may include bidirected arrows) over a set of variables $\mathbf{V}$ and let $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be disjoint sets of variables.

(a) Prove that there exists a set $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ such that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$ if and only if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_0})_{\mathcal{G}}$, where $\mathbf{Z_0} = An(\mathbf{X} \cup \mathbf{Y})_{\mathcal{G}} \setminus (\mathbf{X} \cup \mathbf{Y})$.[1]

Hint: Keep in mind that this proof has two directions, one of which is quite simple. For the other, a possible strategy is to assume the existence of $\mathbf{Z}$ *and proceed by contradiction. Along the proof it may be able to fix any particular path* $\overline{p}$ *between some* $X \in \mathbf{X}$ *and* $Y \in \mathbf{Y}$ *and consider the types of triplets that it may contain.*

Proof. $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_0})_{\mathcal{G}} \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$: Trivially $\mathbf{Z} = \mathbf{Z_0}$ is sufficient to argue this direction.

$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}} \implies (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_0})_{\mathcal{G}}$: We will show this by contradiction. Suppose there exists such $\mathbf{Z}$ but $\mathbf{Z_0}$ is not a separator. This implies $\mathbf{Z}$ and $\mathbf{Z_0}$ are different, which could happen in two ways:

$$\mathbf{Z} \setminus \mathbf{Z_0} \neq \emptyset, \text{ or } \mathbf{Z_0} \setminus \mathbf{Z} \neq \emptyset. \tag{1}$$

For succinctness let $S(\mathbf{Z}) := (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$. Then $S(\mathbf{Z}) \wedge \neg S(\mathbf{Z_0})$ implies there exists a path $\overline{p}$ between some $X \in \mathbf{X}$ and some $Y \in \mathbf{Y}$ that is inactive given $\mathbf{Z}$ but active given $\mathbf{Z_0}$.

(1) This could happen if $\exists W \in (\mathbf{Z} \setminus \mathbf{Z_0}) \cap \overline{p}$ that blocks a triplet in $\overline{p}$. Note that by definition of $\mathbf{Z_0}$, W is not an ancestor of X or Y. Now for W to block a triplet in $\overline{p}$ this should look like $\rightarrow W \rightarrow$, $\leftarrow W \leftarrow$ or $\leftarrow W \rightarrow$. While moving from W towards X or Y using a directed arrow in $\overline{p}$, we must find a collider W' at some point, otherwise we would have $W \rightarrow \cdots \rightarrow X$ or $W \rightarrow \cdots \rightarrow Y$, which contradict the fact that W is not an ancestor of X and Y. It follows that W' or any of its descendants is also not an ancestor of X and Y; therefore, they are not in $\mathbf{Z_0}$ and $\overline{p}$ is blocked given $\mathbf{Z_0}$ since W' is not observed.

[1] The set $An(X)_{\mathcal{G}}$ is defined as $\{V \in \mathbf{V} \mid \exists$ a path (possibly of zero length) $V \rightarrow \cdots \rightarrow X$ in $\mathcal{G}\}$. Then $An(\mathbf{X})_{\mathcal{G}} = \bigcup_{X \in \mathbf{X}} An(X)_{\mathcal{G}}$. Notice that $An(X)_{\mathcal{G}}$ includes X.

(2) The other possibility is that $\exists T \in (\mathbf{Z_0} \setminus \mathbf{Z})$ that activated a triplet in $\overline{p}$ that was inactive when only $\mathbf{Z}$ is conditioned on. Since $T \in \mathbf{Z_0}$, T is an ancestor of $\mathbf{X}$ or $\mathbf{Y}$ and there exists a path $\overline{q}$ that looks like $T \to \cdots \to R$, with $R \in \mathbf{X} \cup \mathbf{Y}$. If we join the portion of $\overline{p}$ that goes from X (Y, if $R \in X$) to some $Y' \in \mathbf{Y}$ ($X' \in \mathbf{X}$, if $R \in X$) with the path $\overline{q}$, we get a path from X to Y' that is active from X to T and unless the portion $\overline{q}$ is also blocked, contradicts $S(\mathbf{Z})$. For $\overline{q}$ to be blocked given $\mathbf{Z}$ some descendant of T needs to be in $\mathbf{Z}$. But then, T would also be an active collider given $\mathbf{Z}$ and the aforementioned triplet in $\overline{p}$ is active, a contradiction.

In any case we either get that $\mathbf{Z_0}$ is also a separator or reach a contradiction, therefore, this direction holds. □

(b) Let $\mathbf{R} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ and $\mathbf{I} \subseteq \mathbf{R}$. Prove that there exists a set $\mathbf{Z}$, $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ such that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$ if and only if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_0})_{\mathcal{G}}$, where $\mathbf{Z_0} = An(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I})_{\mathcal{G}} \cap \mathbf{R}$.

Similar to (i), there are two cases. Since both $\mathbf{Z}$ and $\mathbf{Z_0}$ have to be subsets of $\mathbf{R}$, case (1) can be argued the same way.

For case (2) we need to consider the T may also be an ancestor of $\mathbf{I}$. In that case we can argue that $\mathbf{Z} \supseteq \mathbf{I}$, therefore T would also be an active collider given $\mathbf{Z}$ (since $\mathbf{Z}$ contains $\mathbf{I}$ and $\mathbf{I}$ contains descendants of T), reaching once again a contradiction.