1. The *conditional entropy* $H(X \mid Y)$ is defined as the average entropy in the conditional distribution

$$H(X \mid Y) = \mathbb{E}[\mathbb{E}\{-\log f(X \mid Y) \mid Y\}] = \sum_y \left\{ \sum_x -f(x \mid y) \log f(x \mid y) \right\} f(y).$$

   (a) Use the information inequality to show that

$$H(X \mid Y) \leq H(X),$$

   i.e. the *entropy is always reduced by conditoning*

   (b) Show that

$$H(X, Y) = H(X \mid Y) + H(Y).$$

   (c) For three discrete random variables, show that

$$H(X, Y, Z) + H(Z) \leq H(X, Z) + H(Y, Z).$$

   (d) Show further that

$$X \perp\!\!\!\perp Y \mid Z \iff H(X, Y, Z) + H(Z) = H(X, Z) + H(Y, Z).$$

   Solutions:

   (a) Note first that

$$H(X \mid Y) = \sum_{x,y} -f(x, y) \log \frac{f(x, y)}{f(y)}$$

   and

$$H(X) = \sum_{x,y} -f(x, y) \log f(x).$$

   Then we get

$$H(X) - H(X \mid Y) = \sum_{x,y} f(x, y) \log \frac{f(x, y)}{f(x) f(y)} \geq 0.$$

   (b) Insert $f(x \mid y) = f(x, y)/f(y)$ in the definition and collect terms appropriately.

   (c) Using (a) on the conditional entropy and (b) yields

$$\begin{aligned} H(X, Y, Z) + H(Z) &= H(X \mid Y, Z) + H(Y, Z) + H(Z) \\ &\leq H(X \mid Z) + H(Y, Z) + H(Z) \\ &= H(X, Z) + H(Y, Z). \end{aligned}$$

   (d) The inequality used above:

$$H(X \mid Y, Z) \leq H(X \mid Z)$$

   is sharp unless $f(x \mid y, z) = f(x \mid z)$, i.e. $X \perp\!\!\!\perp Y \mid Z$.

2. Consider a directed graph $\mathcal{D} = (V, E)$ and assume given $k_v, v \in V$ with $k_v \geq 0$ and $\sum_{x_v \in \mathcal{X}_v} k_v(x_v \,|\, x_{\mathrm{pa}(v)}) = 1$. Define

$$p(x) = \prod_{v \in V} k_v(x_v \,|\, x_{\mathrm{pa}(v)}).$$

(a) Show that when $\mathcal{D}$ is acyclic, i.e. a DAG, this yields a well-defined probability distribution;
All we need to show is that

$$\sum_x p(x) = 1. \tag{1}$$

Induction after the number of variables $|V|$:
For $|V| = 1$ (1) is obviously true. Assume that (1) holds for $|V| \leq n$ and consider a DAG $\mathcal{D}$ with $|V| = n + 1$.
Since $\mathcal{D}$ is acyclic, it has a terminal vertex v^*. Since

$$\sum_x = \sum_{x_{V \setminus v^*}} \sum_{x_{v^*}}$$

we get

$$\begin{aligned}
\sum_x p(x) &= \sum_{x_{V \setminus v^*}} \sum_{x_{v^*}} \prod_{v \in V} k_v(x_v \,|\, x_{\mathrm{pa}(v)}) \\
&= \sum_{x_{V \setminus v^*}} \sum_{x_{v^*}} k_{v^*}(x_{v^*} \,|\, x_{\mathrm{pa}(v^*)}) \prod_{v \in V \setminus v^*} k_v(x_v \,|\, x_{\mathrm{pa}(v)}) \\
&= \sum_{x_{V \setminus v^*}} \prod_{v \in V \setminus v^*} k_v(x_v \,|\, x_{\mathrm{pa}(v)}) \sum_{x_{v^*}} k_{v^*}(x_{v^*} \,|\, x_{\mathrm{pa}(v^*)}) \\
&= \sum_{x_{V \setminus v^*}} \prod_{v \in V \setminus v^*} k_v(x_v \,|\, x_{\mathrm{pa}(v)}) = 1.
\end{aligned}$$

In the third line we have used that v^* has no children, so $\prod_{v \in V \setminus v^*} k_v(x_v \,|\, x_{\mathrm{pa}(v)})$ does not involve x_{v^*} and can be taken out of the sum; in the fourth line we have used that $\sum_{x_{v^*}} k_{v^*}(x_{v^*} \,|\, x_{\mathrm{pa}(v^*)}) = 1$ by definition, as well as the inductive assumption yielding that $\sum_{x_{V \setminus v^*}} \prod_{v \in V \setminus v^*} k_v(x_v \,|\, x_{\mathrm{pa}(v)}) = 1$.

(b) Show that it holds that

$$p(x_v \,|\, x_{\mathrm{pa}(v)}) = k_v(x_v \,|\, x_{\mathrm{pa}(v)}). \tag{2}$$

Again we assume this to be true for $|V| \leq n$ and consider $\mathcal{D}$ with $|V| = n + 1$.
Since $\mathcal{D}$ is acyclic, it has a terminal vertex $v = v^*$. From the calculation just made we see that that

$$p(x_{V \setminus v^*}) = \prod_{v \in V \setminus v^*} k_v(x_v \,|\, x_{\mathrm{pa}(v)})$$

and hence

$$p(x_{v^*} \,|\, x_{V \setminus v^*}) = p(x)/p(x_{V \setminus v^*}) = k_{v^*}(x_{v^*} \,|\, x_{\mathrm{pa}(v^*)}).$$

Since the latter does only depends on x through $x_{v^* \cup \mathrm{pa}(v^*)}$, we have

$$v^* \perp\!\!\!\perp V \setminus v^* \mid \mathrm{pa}(v^*)$$

and thus

$$p(x_{v^*} \mid x_{V \setminus v^*}) = p(x_{v^*} \mid x_{\mathrm{pa}(v^*)}) = k_{v^*}(x_{v^*} \mid x_{\mathrm{pa}(v^*)}),$$

as desired.

(c) Give a counterexample in the case where $\mathcal{D}$ has directed cycles.
Consider for example $1 \to 2, 2 \to 3, 3 \to 1$, binary variables and

$$a(x) = k(x_1 \mid x_3)k(x_2 \mid x_1)k(x_3 \mid x_2),$$

where

$$k(x \mid y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{otherwise.} \end{cases}$$

Clearly $\sum_x k(x \mid y) = 1$ and $k(x \mid y) \geq 0$ but $a(x) = 0$ for all x so it does not define a probability distribution.

3. Consider a DAG $\mathcal{D}$ with arrows $1 \to 2, 2 \to 5, 2 \to 3, 5 \to 6, 4 \to 5, 4 \to 7, 5 \to 7$.

(a) Draw the DAG;

(b) List all conditional independence relations corresponding to the local, directed Markov property;

i. $1 \perp_{\mathcal{D}} 4$;
ii. $2 \perp_{\mathcal{D}} 4 \mid 1$;
iii. $3 \perp_{\mathcal{D}} 1, 4, 5, 6, 7 \mid 2$;
iv. $4 \perp_{\mathcal{D}} 1, 2, 3$;
v. $5 \perp_{\mathcal{D}} 1, 3 \mid 2, 4$;
vi. $6 \perp_{\mathcal{D}} 1, 2, 3, 4, 7 \mid 5$;
vii. $7 \perp_{\mathcal{D}} 1, 2, 3, 6 \mid 4, 5$.

(c) List all conditional independence relations corresponding to the ordered Markov property for the well-ordering induced by the given numbering;

i. $3 \perp_{\mathcal{D}} 1 \mid 2$;
ii. $4 \perp_{\mathcal{D}} 1, 2, 3$;
iii. $5 \perp_{\mathcal{D}} 1, 3 \mid 2, 4$;
iv. $6 \perp_{\mathcal{D}} 1, 2, 3, 4 \mid 5$;
v. $7 \perp_{\mathcal{D}} 1, 2, 3, 6 \mid 4, 5$.

(d) Find the ancestral sets generated by the following subsets:

i. $\{5\}$;
ii. $\{2, 7\}$;
iii. $\{4, 6\}$;

$\mathrm{An}(\{5\}) = \{1, 2, 4, 5\}$; $\mathrm{An}(\{2, 7\}) = \{1, 2, 4, 5, 7\}$; $\mathrm{An}(\{4, 6\}) = \{1, 2, 4, 5, 6\}$.

(e) Which of the following separation statements are true? For those that are not true, identify an active trail.

i. $2 \perp_{\mathcal{D}} 4 \,|\, 5$;
ii. $2 \perp_{\mathcal{D}} 7 \,|\, 5$,
iii. $1 \perp_{\mathcal{D}} 7 \,|\, 5, 6$;
iv. $1 \perp_{\mathcal{D}} 4 \,|\, 6$;

All four statements are false. In i., the trail $2 \to 5 \leftarrow 4$ is active because 5 is a collider and observed. In ii., $2 \to 5 \leftarrow 4 \leftarrow 7$ is active as 5 is an observed collider. In iii., $1 \to 2 \to 5 \leftarrow 4 \leftarrow 7$ is active as 5 is an observed collider. In iv., the trail $1 \to 2 \to 5 \leftarrow 4$ is active because arrows 5 is a collider and an ancestor to 6 which is observed.

4. Let P be a distribution which factorizes over the DAG $\mathcal{D}$ and let $G(P)$ be its dependence graph. Show that $G(P) \subseteq \mathcal{D}^m$.

Since P factorizes over $\mathcal{D}$ it also factorizes over $\mathcal{D}^m$. Hence, it satisfies the undirected global Markov property w.r.t. $\mathcal{D}^m$. This again implies that it satisfies the pairwise Markov property w.r.t. $\mathcal{D}^m$. Thus, any edge missing in $\mathcal{D}^m$ corresponds to pairwise conditional independence given the remaining and thus we must have $G(P) \subseteq \mathcal{D}^m$, as requested.

5. A DAG $\mathcal{D}$ is said to be *perfect* if all parents are married, i.e. if it holds that

$$\alpha, \beta \in \mathrm{pa}(\gamma) \Rightarrow \alpha \to \beta \text{ or } \beta \to \alpha.$$

(a) Show that a perfect DAG $\mathcal{D}$ is Markov equivalent to its *skeleton* i.e. the undirected graph obtained by ignoring directions on all arrows;
Use induction after the number of vertices in $\mathcal{D}$
Consider a triplet A, B, S and let v^* be a terminal vertex outside $A \cup B \cup S$. If this does not exist, we have $\mathrm{An}(A \cup B \cup S) = V$ and the moralisation criterion gives $A \perp_{\mathcal{D}} B \,|\, S \iff A \perp_{\mathcal{G}} B \,|\, S$ where $\mathcal{G}$ is the skeleton of $\mathcal{D}$. Else remove v^* and use the induction hypothesis.

(b) Show the converse, i.e. that if $\mathcal{D}$ is Markov equivalent to its skeleton, then $\mathcal{D}$ is perfect.
Again use induction. So we only have to consider the case where a terminal vertex v^* has unmarried parents α and β. These are then separated in the skeleton by $V \setminus \{\alpha, \beta\}$ but not d-separated as the trail $\alpha \to v^* \leftarrow \beta$ is active.

(c) Show that the skeleton of a perfect $\mathcal{D}$ is chordal.
Induction again. Take v^* and its parents. Then $V \setminus v^*$ and $v^* \cup \mathrm{pa}(v^*)$ form a decomposition of the skeleton and the latter component is a clique. The subgraph of $V \setminus v^*$ is the skeleton of a DAG of smaller cardinality and must be chordal by induction. Chordality follows.

(d) Show that the edges of a chordal graph $\mathcal{G}$ can be directed to create a Markov equivalent DAG. *Hint:* Exploit the existence of a perfect numbering of a chordal graph.
Follows directly by using the hint and directing arrows from low to high.