**Review: independence properties of DAGs**

- **Defn**: let $I_l(G)$ be the set of local independence properties encoded by DAG $G$, namely:
  \[\{X_i \perp \text{NonDescendants}(X_i)|\text{Parents}(X_i)\}\]

- **Defn**: A DAG $G$ is an I-map (independence-map) of $P$ if $I_l(G) \subseteq I(P)$.

- A fully connected DAG $G$ is an I-map for any distribution, since $I_l(G) = \emptyset \subseteq I(P)$ for any $P$.

- **Defn**: A DAG $G$ is a minimal I-map for $P$ if it is an I-map for $P$, and if the removal of even a single edge from $G$ renders it not an I-map.

- **To construct a minimal I-map**, Pick a node ordering, then let the parents of node $X_i$ be the minimal subset $U \subseteq \{X_1, \ldots, X_{i-1}\}$ s.t. $X_i \perp \{X_1, \ldots, X_{i-1}\} \setminus U|U$.

**A distribution may have several minimal I-maps**

- **Suppose the left DAG $G$ perfectly captures all and only the independence properties of some distribution $P$, i.e., $I(G) = I(P)$**.

- **Now consider a different node ordering**: $M, J, A, B, E$.

- **Consider adding parents to node $B$. Ancestors are $M, J, A$. We choose $A$ as smallest parent set since $B \perp_G \{M, J\}|A$.**
A distribution may have several minimal I-maps

- Order $B, E, A, J, M$
- Order $M, J, A, B, E$
- Order $M, J, E, B, A$

All represent exactly the same joint distribution, but some orderings are better in terms of
- Representation: easier to understand
- Inference: faster to compute $P(X_q|x_v)$
- Learning: fewer parameters

Soundness and completeness of d-separation

- Defn: $P$ factorizes over DAG $G$ if it can be represented as
  \[ P(X_1, \ldots, X_n) = \prod_i P(X_i | X_{\pi_i}) \]

- Thm 3.3.3 (soundness): If $P$ factorizes over $G$, then $I(H) \subseteq I(P)$.
- Thm 3.3.5 (completeness): If $dsep_G(X; Y | Z)$, then $X \not\in P Y | Z$ in some $P$ that factorizes over $G$.

Global Markov properties of DAGs

- $X$ is d-separated (directed-separated) from $Y$ given $Z$ if we can’t send a ball from any node in $X$ to any node in $Y$, where all nodes in $Z$ are shaded.

- Defn: $I(G) = \{ (X \perp Y | Z) : dsep_G(X; Y | Z) \}$

P-maps

- Defn: A DAG $G$ is a perfect map (P-map) for a distribution $P$ if $I(P) = I(G)$.
- Thm: not every distribution has a perfect map.

- Pf by counterexample. Suppose we have a model where $A \perp C|\{B, D\}$, and $B \perp D|\{A, C\}$. This cannot be represented by any Bayes net.

- e.g., BN1 wrongly says $B \perp D|A$, BN2 wrongly says $B \perp D$. 

\begin{align*}
\text{(a)} & \quad \text{(b)} \\
\text{(a)} & \quad \text{(b)}
\end{align*}
**Undirected Graphical Models**

- Graphs where nodes = random variables, and edges = correlation (direct dependence).
- Defn: Let \( H \) be an undirected graph. Then \( \text{sep}_H(A; C|B) \) iff all paths between \( A \) and \( C \) go through some nodes in \( B \) (simple graph separation).

**Parameterizing undirected graphical models**

- An undirected graph \( H \) specifies a family of distributions s.t., \( I(H) \subseteq I(P) \).
- To specify a particular distribution \( P \), we need to add parameters to the graph.
- For Bayes nets, we used conditional probability distributions (CPDs), \( P(X_i|X_{\pi_i}) \), where \( \sum_{X_i} P(X_i|X_{\pi_i}) = 1 \).
- For Markov nets, we use potential functions or factors defined on subsets of completely connected sets of nodes, where \( \psi_c(X_c) > 0 \).

**UGMs also called Markov Random Fields (MRFs) or Markov Networks.**

---

**Cliquets**

- Defn: a complete subgraph is a fully interconnected set of nodes.
- Defn: a (maximal) clique \( C \) is a complete subgraph s.t. any superset \( C' \supset C \) is not complete.
- Defn: a sub-clique is a not-necessarily-maximal clique.

- Example: max-cliques = \( \{A, B, D\}, \{B, C, D\} \), sub-cliques = edges = \( \{A, D\}, \{A, B\}, \ldots \)

---

**Undirected graphical models**

- Defn: an undirected graphical model representing a distribution \( P(X_1, \ldots, X_n) \) is an undirected graph \( H \), and a set of positive potential functions \( \psi_c \) associated with sub-cliques of \( H \), s.t.

\[
P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)
\]

where \( Z \) is the partition function:

\[
Z = \sum_{x_1, \ldots, x_n} \prod_{c \in C} \psi_c(x_c)
\]

- Defn: if \( H \) is a UGM for \( P \), we say that \( P \) factorizes over \( H \), or that \( P \) is a Gibbs distribution over \( H \).
Example of UGM - max cliques

\[ P(x_{1:4}) = \frac{1}{Z} \psi_{124}(x_{124}) \times \psi_{234}(x_{234}) \]
\[ Z = \sum_{x_1, x_2, x_3, x_4} \psi_{124}(x_{124}) \times \psi_{234}(x_{234}) \]

- We can represent \( P(X_{1:4}) \) as two 3D tables instead of one 4D table.

---

Example of UGM - subcliques

\[ P(x_{1:4}) = \frac{1}{Z} \prod_{<ij>} \psi_{ij}(x_{ij}) \]
\[ = \frac{1}{Z} \psi_{12}(x_{12}) \psi_{14}(x_{14}) \psi_{23}(x_{23}) \psi_{24}(x_{24}) \psi_{34}(x_{34}) \]
\[ Z = \sum_{x_1, x_2, x_3, x_4 <ij>} \prod_{<ij>} \psi_{ij}(x_{ij}) \]

- We can represent \( P(X_{1:4}) \) as five 2D tables instead of one 4D table.

---

Max cliques vs sub cliques

- Max clique version
  \[ P(X_{1:4}) = \frac{1}{Z} \psi_{1234}(X_{1234}) \]

- Sub clique version
  \[ P(X_{1:4}) = \frac{1}{Z} \prod_{<ij>} \psi_{ij}(x_i, x_j) \]
  \[ = \frac{1}{Z} \psi_{12}(x_{12}) \psi_{13}(x_{13}) \psi_{14}(x_{14}) \psi_{23}(x_{23}) \psi_{24}(x_{24}) \psi_{34}(x_{34}) \]

---

Interpretation of Clique Potentials

- The model implies \( x \perp z \mid y \)
  \[ p(x, y, z) = p(y)p(x|y)p(z|y) \]

- We can write this as:
  \[ p(x, y, z) = p(x, y)p(z|y) = \psi_{xy}(x, y)\psi_{yz}(y, z) \]
  \[ p(x, y, z) = p(x|y)p(z, y) = \psi_{xy}(x, y)\psi_{yz}(y, z) \]

  cannot have all potentials be marginals
  cannot have all potentials be conditionals

- The positive clique potentials can only be thought of as general “compatibility”, “goodness” or “happiness” functions over their variables, but not as probability distributions.
**Boltzmann Distributions/ log-linear models**

- We often represent the clique potentials using their logs:
  \[ \psi_C(x_C) = \exp\{-H_C(x_C)\} \]
  for arbitrary real valued “energy” functions \( H_C(x_C) \).
  The negative sign is a standard convention.
- This gives the joint a nice additive structure:
  \[ P(X) = \frac{1}{Z} \exp\{-\sum_C H_C(x_C)\} = \frac{1}{Z} \exp\{-H(X)\} \]
  where the sum in the exponent is called the “free energy”:
  \[ H(X) = \sum_C H_C(x_C) \]
- In physics, this is called the “Boltzmann distribution”.
- In statistics, this is called a log-linear model.

**Example: Boltzmann machines**

- A fully connected graph with pairwise (edge) potentials on binary-valued nodes (for \( x_i \in \{-1, +1\} \) or \( x_i \in \{0, 1\} \)) is called a Boltzmann machine.
  \[ P(X_{1:4}) = \frac{1}{Z} \prod_{<ij>} \psi_{ij}(x_i, x_j) \]
  where \( \psi_{ij}(x_i, x_j) = \exp(-H_{ij}(x_i, x_j)) \), and
  \[ H(x_i, x_j) = (x_i - \mu_i)V_{ij}(x_j - \mu_j) \]
- Hence overall energy has form
  \[ H(x) = \sum_{ij} V_{ij}x_ix_j + \sum_i \alpha_ix_i + C \]

**Example: Ising (spin-glass) models**

- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbours.
- Same as sparse Boltzmann machine, where \( V_{ij} \neq 0 \) iff \( i, j \) are neighbors.
- e.g., nodes are pixels, potential function encourages nearby pixels to have similar intensities.
- Potts model = multi-state Ising model.

**Example: multivariate Gaussian Distribution**

- A Gaussian distribution can be represented by a fully connected graph with pairwise (edge) potentials of the form
  \[ H(x) = \sum_{ij} (x_i - \mu_i)V_{ij}(x_j - \mu_j) \]
  where \( \mu \) is the mean and \( V \) is the inverse covariance (precision) matrix, since
  \[ P(x_{1:n}) = \frac{1}{Z} e^{-H(x)} \]
- Same as Boltzmann machine except \( x_i \in R \).
Sparse graph $\equiv$ zeros in precision matrix

- $V_{ij} = 0$ iff no edge between $X_i$ and $X_j$.
- Chain structured graph $\equiv$ block diagonal precision matrix

\[ V = \Sigma^{-1} = \begin{pmatrix}
  \cdot & 0 & 0 & 0 \\
  \cdot & \cdot & 0 & 0 \\
  0 & \cdot & \cdot & \cdot \\
  0 & 0 & 0 & \cdot 
\end{pmatrix} \]

Graphs and distributions

- Let us return to the question of what kinds of distributions can be represented by undirected graphs (ignoring the details of the particular parameterization).
- Defn: the {\it global Markov properties} of a UG $H$ are

\[ I(H) = \{(X \perp Y|Z) : sep_H(X;Y|Z)\} \]

- Is this definition sound and complete?

Soundness and completeness of global Markov property

- Defn: An UG $H$ is an I-map for a distribution $P$ if $I(H) \subseteq I(P)$, i.e., $P \models I(H)$.
- Defn: $P$ is a Gibbs distribution over $H$ if it can be represented as

\[ P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{c \in C(H)} \psi_c(x_c) \]

- Thm 5.4.2 (soundness): If $P$ is a Gibbs distribution over $H$, then $H$ is an I-map of $P$.
- Thm 5.4.3 (Hammersley-Clifford): Let $P$ be a positive distribution (i.e., $\forall x . P(x) > 0$). If $H$ is an I-map for $P$, then $P$ can be represented as a Gibbs distribution over $H$.
- Thm 5.4.5 (completeness): If $\neg sep_H(X;Y|Z)$, then $X \not\in P Y|Z$ in some $P$ that factorizes over $H$.
For directed graphs, we defined I-maps in terms of local Markov properties, and derived global independence.

For undirected graphs, we defined I-maps in terms of global Markov properties, and will now derive local independence.

Defn: The pairwise markov independencies associated with $UG_H = (V, E)$ are

$$I_p(H) = \{(X \perp Y | V \setminus \{X, Y\}) : \{X, Y\} \notin E\}$$

e.g., $X_1 \perp X_5 | \{X_2, X_3, X_4\}$

Local Markov properties

Defn: The local markov independencies associated with $UG_H = (V, E)$ are

$$I_l(H) = \{(X \perp V \setminus \{X\} | N_H(X)|N_H(X)) : X \in V\}$$

where $N_H(X)$ are the neighbors
e.g., $X_1 \perp \{X_3, X_4, X_5\} | X_2$

$N_H(X)$ is also called the Markov blanket of $X$.

Relationship between local and global Markov properties

Thm 5.5.3. If $P \models I_l(H)$ then $P \models I_p(H)$.
Thm 5.5.4. If $P \models I(H)$ then $P \models I_l(H)$.
Thm 5.5.5. If $P > 0$ and $P \models I_p(H)$, then $P \models I(H)$.
Corollary 5.5.6: If $P > 0$, then $I_l = I_p = I$.

If $\exists x. P(x) = 0$, then we can construct an example (using deterministic potentials) where $I_p \not\models I_l$ or $I_l \not\models I$.

I-maps for undirected graphs

Defn: A Markov network $H$ is a minimal I-map for $P$ if it is an I-map, and if the removal of any edge from $H$ renders it not an I-map.

How can we construct a minimal I-map from a positive distribution $P$?

Pairwise method: add edges between all pairs $X,Y$ s.t.

$$P \not\models (X \perp Y | V \setminus \{X, Y\})$$

Local method: add edges between $X$ and all $Y \in MB_p(X)$, where $MB_p(X)$ is the minimal set of nodes $U$ s.t.

$$P \models (X \perp V \setminus \{X\} \setminus \{U\})$$

Thm 5.5.11/12: both methods induce the unique minimal I-map.

If $\exists x. P(x) = 0$, then we can construct an example where either method fails to induce an I-map.
Perfect Maps

- Defn: A Markov network $H$ is a **perfect map** for $P$ if for any $X, Y, Z$ we have that
  \[ \text{sep}_H(X; Y | Z) \iff P \models (X \perp Y | Z) \]
- Thm: not every distribution has a perfect map.
- Pf by counterexample. No undirected network can capture all and only the independencies encoded in a v-structure $X \rightarrow Z \leftarrow Y$.

Expressive Power

- Can we always convert directed $\leftrightarrow$ undirected?
  - No.

Converting Bayes nets to Markov nets

- Defn: A Markov net $H$ is an I-map for a Bayes net $G$ if $I(H) \subseteq I(G)$.
- We can construct a minimal I-map for a BN by finding the minimal Markov blanket for each node.
- We need to block all active paths coming into node $X$, from parents, children, and co-parents; so connect them all to $X$.

Moralization

- Defn: the moral graph $H(G)$ of a DAG is constructed by adding undirected edges between any pair of disconnected (“unmarried”) nodes $X, Y$ that are parents of a child $Z$, and then dropping all remaining arrows.
Moralization

- Thm 5.7.5: The moral graph $H(G)$ is the minimal I-map for Bayes net $G$.
- Pf: moralization loses conditional independence information, and hence is conservative; hence $H(G)$ is an I-map of $G$. Moralization only introduces where needed to make the semantics of simple separation capture d-separation, hence minimal.

Bayes net to Markov net

- We assign each CPD to one of the clique potentials that contains it, e.g.

$$P(U, X, Y, Z) = \frac{1}{Z} \psi(U, X) \times \psi(X, Y, Z)$$
$$= \frac{1}{1} P(U)P(X|U) \times P(Y)P(Z|X,Y)$$
$$= P(X,U) \times P(Z|X,Y)P(Y)$$

Alternative to d-separation

- Thm 5.7.7. Let $X, Y, Z$ be 3 disjoint sets of nodes in DAG $G$. Let $U = X \cup Y \cup Z$, let $G^+[U]$ be the induced DAG over $\text{Ancestors}(U)$, and let $H' = \text{moralize}(G^+[U])$ be the moralized ancestral subgraph. Then $dsep_G(X; Y|Z) \iff sep_{H'}(X; Y|Z)$.

- Example: $dsep_G(Z_1; U_1|Y_1)$?