Review: independence properties of DAGs

- Defn: let $I(G)$ be the set of local independence properties encoded by DAG $G$, namely:
  \[
  \{ X_i \perp \text{NonDescendants}(X_i) \mid \text{Parents}(X_i) \}
  \]
- Defn: A DAG $G$ is an l-map (independence-map) of $P$ if $I(G) \subseteq I(P)$
- A fully connected DAG $G$ is an l-map for any distribution, since $I(G) = \emptyset \subseteq I(P)$ for any $P$.
- Defn: A DAG $G$ is a minimal l-map for $P$ if it is an l-map for $P$, and if the removal of even a single edge from $G$ renders it not an l-map.
- A distribution may have several minimal l-maps
  - Each corresponding to a specific node-ordering
Global Markov properties of DAGs

- $X$ is **d-separated** (directed-separated) from $Z$ given $Y$ if we can't send a ball from any node in $X$ to any node in $Z$ using the "Bayes-ball" algorithm illustrated below:

  - Defn: $\mathcal{I}(G) = \text{all independence properties that correspond to d-separation}:
    \[ \mathcal{I}(G) = \{X \perp Z | Y\} : \text{dsep}_e(X, Z | Y) \]
  - D-separation is sound and complete (Chap 3, Koller & Friedman)

P-maps

- Defn: A DAG $G$ is a **perfect map** (P-map) for a distribution $P$ if $\mathcal{I}(P) = \mathcal{I}(G)$.
- Thm: not every distribution has a perfect map as DAG.
  - Pf by counterexample. Suppose we have a model where $A \perp C \mid \{B, D\}$, and $B \perp D \mid \{A, C\}$. This cannot be represented by any Bayes net.
  - e.g., BN1 wrongly says $B \perp D \mid A$, BN2 wrongly says $B \perp D$. 

\[ \begin{align*}
  & \text{BN1} & \text{BN2} & \text{MRF} \\
  D & \searrow & D & D \searrow \\
  C & \downarrow & C & C \downarrow \\
  A & \nwarrow & B & A \nwarrow \\
  B & \nearrow & D & B \nearrow \\
\end{align*} \]
Undirected graphical models

- Pairwise (non-causal) relationships
- Can write down model, and score specific configurations of the graph, but no explicit way to generate samples
- Contingency constrains on node configurations

Canonical examples

- The grid model
  - Naturally arises in image processing, lattice physics, etc.
  - Each node may represent a single "pixel", or an atom
    - The states of adjacent or nearby nodes are "coupled" due to pattern continuity or electro-magnetic force, etc.
    - Most likely joint-configurations usually correspond to a "low-energy" state
Social networks

The Social Structure of “Countryside” School District

Ignoring the arrows, this is a "relational network" among people

Protein interaction networks
Modeling Go

This is the middle position of a Go game. Overlaid is the estimate for the probability of becoming black or white for every intersection. Large squares mean the probability is higher.

Information retrieval
Semantics of Undirected Graphs

- Let $H$ be an undirected graph:

- $B$ separates $A$ and $C$ if every path from a node in $A$ to a node in $C$ passes through a node in $B$. $\text{sep}_H(A; C | B)$

- A probability distribution satisfies the **global Markov property** if for any disjoint $A$, $B$, $C$, such that $B$ separates $A$ and $C$, $A$ is independent of $C$ given $B$: $\text{I}(H) = \{ A \perp C | B : \text{sep}_H(A; C | B) \}$

Undirected Graphical Models

- Defn: an undirected graphical model represents a distribution $P(X_1, \ldots, X_n)$ defined by an undirected graph $H$, and a set of positive **potential functions** $\psi_c$ associated with 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 known as the partition function:

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

- Also known as Markov Random Fields, Markov networks …

- The **potential function** can be understood as an contingency function of its arguments assigning “pre-probabilistic” score of their joint configuration.
Cliques

- For $G = (V, E)$, a complete subgraph (clique) is a subgraph $G' = (V', E')$ such that nodes in $V'$ are fully interconnected.
- A (maximal) clique is a complete subgraph s.t. any superset $V'' \supset V'$ is not complete.
- A sub-clique is a not-necessarily-maximal clique.

Example:
- max-cliques = \{A, B, D\}, \{B, C, D\},
- sub-cliques = \{A, B\}, \{C, D\}, ... all edges and singletons

Example UGM – using max cliques

- For discrete nodes, we can represent $P(X_1:4)$ as two 3D tables instead of one 4D table.
Example UGM – using subcliques

For discrete nodes, we can represent $P(X_1:4)$ as 5 2D tables instead of one 4D table.

Interpretation of Clique Potentials

The model implies $X_1 \perp Z | Y$. This independence statement implies (by definition) that the joint must factorize as:

$$p(x, y, z) = p(y)p(x | y)p(z | y)$$

We can write this as:

$$p(x, y, z) = p(x, y)p(z | y), \text{ but } p(x, y, z) = p(x | y)p(z, y)$$

- 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.
Exponential Form

- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential $\psi_c(x_c)$ in an unconstrained form using a real-value "energy" function $\phi_c(x_c)$:

$$\psi_c(x_c) = \exp[-\phi_c(x_c)]$$

For convenience, we will call $\phi_c(x_c)$ a potential when no confusion arises from the context.

- This gives the joint a nice additive structure

$$p(x) = \frac{1}{Z} \exp \left\{ - \sum_{c \in C} \phi_c(x_c) \right\} = \frac{1}{Z} \exp \{ -H(x) \}$$

where the sum in the exponent is called the "free energy":

$$H(x) = \sum_{c \in C} \phi_c(x_c)$$

- 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

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} \exp \left\{ \sum_{i,j} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C \right\}$$

$$H(x) = \sum_{i,j} (x_i - \mu) \theta_{ij} (x_j - \mu) = (x - \mu)^T \Theta (x - \mu)$$
Example: Ising (spin-glass) models

- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors.

- Same as sparse Boltzmann machine, where $\Theta_{ij} \neq 0$ 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 over continuous nodes.
- The overall energy has the form

$$H(x) = \sum_{ij} (x_i - \mu) \Theta_{ij} (x_j - \mu) = (x - \mu)^T \Theta (x - \mu)$$

where $\mu$ is the mean and $\Theta$ is the inverse covariance (precision) matrix.

- Also known as Gaussian graphical model (GGM), same as Boltzmann machine except $x_i \in \mathbb{R}$.
Sparse precision vs. sparse covariance in GGM

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 global Markov properties of a UG $H$ are

$$I(H) = \{X \perp Z \mid Y \} : \text{sep}_H(X; Z \mid Y)$$

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 $\mathcal{I}(H) \subseteq \mathcal{I}(P)$, i.e., $P$ entails $\mathcal{I}(H)$.
- Defn: $P$ is a Gibbs distribution over $H$ if it can be represented as
  \[ P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \]
  \[ \mathcal{I}(H) \subseteq \mathcal{I}(P), \]
- Thm 5.4.2 (soundness): If $P$ is a Gibbs distribution over $H$, then $H$ is an I-map of $P$.
- Thm 5.4.5 (completeness): If $\neg \text{sep}_P(X; Z | Y)$, then $X \perp Z | Y$ in some $P$ that factorizes over $H$.

Local and global Markov properties

- 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 $H = (\mathcal{V}, E)$ are
  \[ I_j(H) = \{X \perp Y | \mathcal{V} \setminus \{X,Y\} : \{X,Y\} \in E\} \]
- e.g., $X_1 \perp X_5 | \{X_2, X_3, X_4\}$

1 2 3 4 5
Local Markov properties

- A distribution has the local Markov property w.r.t. a graph $\mathcal{H}=(V,E)$ if the conditional distribution of variable given its neighbors is independent of the remaining nodes.

  $$I_c(\mathcal{H}) = \{X \perp V \setminus (X \cup \mathcal{N}_c(X)) | \mathcal{N}_c(X) : X \in V\}$$

- **Theorem** (Hammersley-Clifford): If the distribution is strictly positive and satisfies the local Markov property, then it factorizes with respect to the graph.

- $\mathcal{N}_c(X)$ is also called the Markov blanket of $X$.

Relationship between local and global Markov properties

- Thm 5.5.3. If $P \models I_c(\mathcal{H})$ then $P \models I_p(\mathcal{H})$.
- Thm 5.5.4. If $P \models I(\mathcal{H})$ then $P \models I_c(\mathcal{H})$.
- Thm 5.5.5. If $P > 0$ and $P \models I_p(\mathcal{H})$, then $P \models I(\mathcal{H})$.
  - Pf sketch: $p(a,b|c,d)=p(a|c,d)p(b|c,d)$ and $d$ separate $b$ from $\{a,c\}$
    $$p(a,b|c,d)p(c|d)=p(a|c,d)p(b|c,d)p(c|d)=p(a,c|d)p(b|d)$$
- Corollary 5.5.6: If $P > 0$, then $I_i = I_p = I$.
- If $\exists x: P(x) = 0$, then we can construct an example (using deterministic potentials) where $I_i \not\models I$ or $I_p \not\models I$. 

Given more info

Given less info

Given more info
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.
    \[ P \models (X \perp Y | V \setminus \{X, Y\}) \]
  - Local method: add edges between $X$ and all $Y \in MBP(X)$, where $MBP(X)$ is the minimal set of nodes $U$ s.t.
    \[ P \models (X \perp U \setminus \{X\} \cup \{U\} | Y) \]
- Thm 5.5.11/12: both methods induce the unique minimal I-map.
- If $\exists x$ s.t. $P(x) = 0$, 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
  \[ \text{sep}_H(X; Z | Y) \iff P \models (X \perp Z | Y) \]
- Thm: not every distribution has a perfect map as UGM.
  - Pf by counterexample. No undirected network can capture all and only the independencies encoded in a v-structure $X \Rightarrow Z \Leftarrow Y$. 

\[ \text{Diagram: } \begin{tikzpicture}
  \node (X) at (0,0) {$X$};
  \node (Z) at (2,0) {$Z$};
  \node (Y) at (1,1) {$Y$};
  \node (D) at (-1,1) {$D$};
  \node (U) at (1,-1) {$U$};
  \node (P) at (3,0) {$P$};
  \draw (X) -- (Z);
  \draw (X) -- (Y);
  \draw (Z) -- (Y);
  \end{tikzpicture} \]
The expressive power of UGM

- Can we always convert directed ↔ undirected?
- No.

No directed model can represent these and only these independencies.

\[ X \perp Y \mid \{W, Z\} \]
\[ W \perp Z \mid \{X, Y\} \]

No undirected model can represent these and only these independencies.

\[ X \perp Y \]

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

- The moral graph $\mathcal{G}(\mathcal{G})$ 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.

- To turn a BN into a MRF, we assign each CPD to one of the clique potentials that contains it.