Undirected Graphical Models

Bayesian Network Representation

- Directed acyclic graph structure
  - Conditional parameterization
  - Independencies in graphs
  - From distribution to BN graphs

- Conditional probability distributions (CPDs)
  - Table
  - Deterministic
  - Context-specific (Tree, Rule CPDs)
  - Independence of causal influence (Noisy OR, GLMs)
  - Continuous variables
  - Hybrid models
The *Misconception* Example

- Four students get together in pairs to work on HWs: Alice, Bob, Charles, Debbie
- Only the following pairs meet: (A&B), (B&C), (C&D), (D&A)
- Let’s say that the prof accidentally misspoke in class
  - Each student may subsequently have figured out the problem.
  - In subsequent study pairs, they may transmit this newfound understanding to their partners.
- Consider 4 binary random variables
  - A, B, C, D: whether the student has the misconception or not.
- Independence assumptions?
- Can we find the P-map for these?

Reminder: Perfect Maps

- G is a perfect map (P-map) for P if I(P) = I(G)
- Does every distribution have a P-map?
  - No: some structures cannot be represented in a BN
    - *Independencies in P*: (A ⊥ D | B, C) and (B ⊥ C | A, D)

(B ⊥ C | A,D) does not hold  (A ⊥ D) also holds
Representing Dependencies

- \((A \perp D \mid B,C)\) and \((B \perp C \mid A,D)\)
  - Cannot be modeled with a Bayesian network.
  - Can be modeled with an undirected graphical models (Markov networks).

Undirected Graphical Models (Informal)

- **Nodes** correspond to random variables
- **Edges** correspond to direct probabilistic interaction
  - An interaction not mediated by any other variables in the network.

- How to **parameterize**?
  - Local factor models are attached to sets of nodes
    - Factor elements are positive
    - Do not have to sum to 1
    - Represent affinities, compatibilities

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>D</th>
<th>π(_{A,C})</th>
</tr>
</thead>
<tbody>
<tr>
<td>a(^0)</td>
<td>d(^0)</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>a(^0)</td>
<td>d(^1)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>a(^1)</td>
<td>d(^0)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>a(^1)</td>
<td>d(^1)</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>π(_{A,B})</th>
</tr>
</thead>
<tbody>
<tr>
<td>a(^0)</td>
<td>b(^0)</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>a(^0)</td>
<td>b(^1)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>a(^1)</td>
<td>b(^0)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>a(^1)</td>
<td>b(^1)</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
<th>π(_{C,D})</th>
</tr>
</thead>
<tbody>
<tr>
<td>c(^0)</td>
<td>d(^0)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>c(^0)</td>
<td>d(^1)</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>c(^1)</td>
<td>d(^0)</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>c(^1)</td>
<td>d(^1)</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>C</th>
<th>π(_{B,C})</th>
</tr>
</thead>
<tbody>
<tr>
<td>b(^0)</td>
<td>c(^0)</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>b(^0)</td>
<td>c(^1)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>b(^1)</td>
<td>c(^0)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>b(^1)</td>
<td>c(^1)</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>
Undirected Graphical Models (Informal)

- Represents joint distribution
  - Unnormalized factor
    \[ F(a,b,c,d) = \pi_1[a,b]\pi_2[a,c]\pi_3[b,d]\pi_4[c,d] \]
  - Probability
    \[ P(a,b,c,d) = \frac{1}{Z}\pi_1[a,b]\pi_2[a,c]\pi_3[b,d]\pi_4[c,d] \]
  - Partition function
    \[ Z = \sum_{a,b,c,d} \pi_1[a,b]\pi_2[a,c]\pi_3[b,d]\pi_4[c,d] \]

- As undirected graphical models represent joint distributions, they can be used for answering queries.

Undirected Graphical Models Blurb

- Useful when edge directionality cannot be assigned
- Simpler interpretation of structure
  - Simpler inference
  - Simpler independency structure
- Harder to learn parameters/structures
- We will also see models with combined directed and undirected edges
- Markov networks
Markov Network Structure

- Undirected graph \( H \)
  - Nodes \( X_1, \ldots, X_n \) represent random variables

- \( H \) encodes independence assumptions
  - A path \( X_1-X_2-\ldots-X_k \) is active if none of the \( X_i \) variables along the path are observed
  - \( X \) and \( Y \) are separated in \( H \) given \( Z \) if there is no active path between any node \( x \in X \) and any node \( y \in Y \) given \( Z \)
    - Denoted \( \text{sep}_H(X; Y \mid Z) \)

\[
\begin{align*}
D \perp \{A,C\} & \mid B & \iff & \text{Global independence associated with } H: \\
I(H) &= \{ (X \perp Y \mid Z) : \text{sep}_H(X; Y \mid Z) \}
\end{align*}
\]

Relationship with Bayesian Network

- Bayesian network
  - Local independencies \( \rightarrow \) Independence by d-separation (global)
- Markov network
  - Global independencies \( \rightarrow \) Local independencies

- Can all independencies encoded by Markov networks be encoded by Bayesian networks?
  - No, counter example – \((A \perp B \mid C,D)\) and \((C \perp D \mid A,B)\)

- Can all independencies encoded by Bayesian networks be encoded by Markov networks?
  - No, immoral v-structures (explaining away)

- Markov networks encode monotonic independencies
  - If \( \text{sep}_H(X; Y \mid Z) \) and \( Z \perp Z' \) then \( \text{sep}_H(X; Y \mid Z') \)
Markov Network Factors

- A **factor** is a function from value assignments of a set of random variables $\mathbf{D}$ to real positive numbers $\Re^+$
  - The set of variables $\mathbf{D}$ is the **scope** of the factor

- Factors generalize the notion of CPDs
  - Every CPD is a factor (with additional constraints)

Factors and Joint Distribution

- Can we represent any joint distribution by using only factors that are defined on edges?
  - **No!** Compare # of parameters
  - Example: $n$ binary RVs
    - Joint distribution has $2^n-1$ independent parameters
    - Markov network with edge factors has $4 \binom{n}{2}$ parameters

**Needed:** $2^7 - 1 = 127!$

**Edge parameters:** $4 \cdot \binom{7}{2} = 84$

- Factors introduce constraints on joint distribution
Factors and Graph Structure

- Are there constraints imposed on the network structure \( H \) by a factor whose scope is \( D \)?
  - Hint 1: think of the independencies that must be satisfied
  - Hint 2: generalize from the basic case of \(|D| = 2\)

The induced subgraph over \( D \) must be a clique (fully connected)

Why? otherwise two unconnected variables may be independent by blocking the active path between them, contradicting the direct dependency between them in the factor over \( D \).

Maximal cliques
- \( \{A,B\} \)
- \( \{B,C\} \)
- \( \{C,D\} \)
- \( \{A,D\} \)

Markov Network Factors: Examples
Markov Network Distribution

- A distribution \( P \) factorizes over \( H \) if it has:
  - A set of subsets \( D_1, \ldots, D_m \) where each \( D_i \) is a complete (fully connected) subgraph in \( H \)
  - Factors \( \pi_1[D_1], \ldots, \pi_m[D_m] \) such that

\[
P(X_1, \ldots, X_n) = \frac{1}{Z} \prod f(X_1, \ldots, X_n) = \frac{1}{Z} \prod \pi_i[D_i]
\]

where un-normalized factor: \( f(X_1, \ldots, X_n) = \prod \pi_i[D_i] \)

- \( Z \) is called the partition function
- \( P \) is also called a Gibbs distribution over \( H \)

Pairwise Markov Networks

- A pairwise Markov network over a graph \( H \) has:
  - A set of node potentials \( \{\pi[X_i]: i=1, \ldots, n\} \)
  - A set of edge potentials \( \{\pi[X_i,X_j]: X_i,X_j \in H\} \)

- Example:
Logarithmic Representation

- We represent energy potentials by applying a log transformation to the original potentials
  - $\pi[D] = \exp(-\varepsilon[D])$ where $\varepsilon[D] = -\ln \pi[D]$
- Any Markov network parameterized with factors can be converted to a logarithmic representation
- The log-transformed potentials can take on any real value
- The joint distribution decomposes as

\[
P(X_1, \ldots, X_n) = \frac{1}{Z} \exp \left[ -\sum_{i=1}^{m} \varepsilon_i[D_i] \right]
\]

\[\text{Log } P(\mathbf{X}) \text{ is a linear function.}\]

I-Maps and Factorization

- Independency mappings (I-map)
  - $I(P)$ – set of independencies $(X \perp Y | Z)$ in $P$
  - I-map – independencies by a graph is a subset of $I(P)$

- Bayesian Networks
  - Factorization and reverse factorization theorems
    - $G$ is an I-map of $P$ iff $P$ factorizes as $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i))$

- Markov Networks
  - Factorization and reverse factorization theorems
    - $H$ is an I-map of $P$ iff $P$ factorizes as $P(X_1, \ldots, X_n) = \frac{1}{Z} \prod \pi[D_i]$
Reverse Factorization

- \( P(X_1, \ldots, X_n) = \frac{1}{Z} \prod \pi_i[D_i] \Rightarrow H \) is an I-map of \( P \)

- **Proof:**
  - Let \( X, Y, W \) be any three disjoint sets of variables such that \( W \) separates \( X \) and \( Y \) in \( H \)
  - We need to show \( (X \perp Y | W) \in I(P) \)

- **Case 1:** \( X \cup Y \cup W = U \) (all variables)
  - As \( W \) separates \( X \) and \( Y \) there are no direct edges between \( X \) and \( Y \)
  - any clique in \( H \) is fully contained in \( X \cup W \) or \( Y \cup W \)
  - Let \( I_X \) be subcliques in \( X \cup W \) and \( I_Y \) be subcliques in \( Y \cup W \) (not in \( X \cup W \))

\[
\prod_{\pi_i[D_i], \pi_i[D_j]} = \frac{1}{Z} f(X, W)g(Y, W) \\
\Rightarrow (X \perp Y | W) \in I(P)
\]
Factorization

- If H is an I-map of P then \( P(X_1, \ldots, X_n) = \frac{1}{Z} \prod \pi_i[D_i] \)

- Holds only for positive distributions P
  - Hammerly-Clifford theorem

- Defer proof

Relationship with Bayesian Network

- Bayesian Networks
  - **Semantics** defined via local independencies \( I_L(G) \).
  - Global independencies induced by d-separation
  - Local and global independencies equivalent since one implies the other

- Markov Networks
  - **Semantics** defined via global separation property \( I(H) \)
  - Can we define the induced local independencies?
    - We show two definitions (call them "Local Markov assumptions")
    - All three definitions (global and two local) are equivalent only for positive distributions P
Pairwise Independencies

- Every pair of disconnected nodes are separated given all other nodes in the network

Formally: $I_p(H) = \{ (X \perp Y | U - \{X,Y\}) : X \neq Y \notin H \}$

Example:
- $(A \perp D | B,C,E)$
- $(B \perp C | A,D,E)$
- $(D \perp E | A,B,C)$

Local Independencies

- Every node is independent of all other nodes given its immediate neighboring nodes in the network
  Markov blank of X, $MB_H(X)$

Formally: $I_L(H) = \{ (X \perp U-\{X\}-MB_H(X) | MB_H(X)) : X \in H \}$

Example:
- $(A \perp D | B,C,E)$
- $(B \perp C | A,D,E)$
- $(C \perp B | A,D,E)$
- $(D \perp E,A | B,C)$
- $(E \perp D | A,B,C)$
Relationship Between Properties

- Let $I(H)$ be the **global separation** independencies
- Let $I_L(H)$ be the **local (Markov blanket)** independencies
- Let $I_p(H)$ be the **pairwise** independencies

- For any distribution $P$:
  - $I_p(H) \rightarrow I(H)$
    - The assertion in $I(H)$, that a node is independent of all other nodes given its neighbors, is part of the separation independencies since there is no active path between a node and its non-neighbors given its neighbors
  - $I_L(H) \rightarrow I_p(H)$
    - Follows from the monotonicity of independencies in Markov networks (if $(X \perp Y|Z)$ and $Z \subseteq Z'$ then $(X \perp Y|Z')$)

Proof relies on intersection property for probabilities $(X \perp Y|Z,W)$ and $(X \perp W|Z,Y) \rightarrow (X \perp Y,W|Z)$ which holds in general only for positive distributions

- Details on the textbook

Thus, for positive distributions
- $I(H) \leftrightarrow I_L(H) \leftrightarrow I_p(H)$

- How about a non-positive distribution?
The Need for Positive Distribution

- Let $P$ satisfy
  - $A$ is uniformly distributed
  - $A=B=C$

- $P$ satisfies $I_P(H)$
  - $(B \perp C|A)$, $(A \perp C|B)$
    (since each variable determines all others)

- $P$ does not satisfy $I_L(H)$
  - $(C \perp A,B)$ needs to hold according to $I_L(H)$ but does not hold in the distribution

Constructing Markov Network for $P$

- **Goal**: Given a distribution, we want to construct a Markov network which is an I-map of $P$

- Complete (fully connected) graphs will satisfy but are not interesting

- Minimal I-maps: A graph $G$ is a minimal I-Map for $P$ if:
  - $G$ is an I-map for $P$
  - Removing any edge from $G$ renders it not an I-map

- **Goal**: construct a graph which is a minimal I-map of $P$
Constructing Markov Network for $P$

- If $P$ is a positive distribution, then $I(H) \leftrightarrow I_L(H) \leftrightarrow I_P(H)$
- Thus, sufficient to construct a network that satisfies $I_P(H)$

- Construction algorithm
  - For every $(X,Y)$ add edge if $(X \perp Y | U - \{X,Y\})$ does not hold in $P$

- Theorem: network is minimal and unique $I$-map
  - Proof:
    - $I$-map follows since $I_P(H)$ by construction and $I(H)$ by equivalence
    - Minimality follows since deleting an edge implies $(X \perp Y | U - \{X,Y\})$
      But, we know by construction that this does not hold in $P$ since we added the edge in the construction process
    - Uniqueness follows since any other $I$-map has at least these edges and to be minimal cannot have additional edges

---

Constructing Markov Network for $P$

- If $P$ is a positive distribution then
  $$I(H) \leftrightarrow I_L(H) \leftrightarrow I_P(H)$$
- Thus, sufficient to construct a network that satisfies $I_L(H)$

- Construction algorithm
  - Connect each $X$ to every node in the minimal set $Y$ s.t.:
    $$\{(X \perp U - \{X\} - Y | Y) : X \in H\}$$

- Theorem: network is minimal and unique $I$-map
Markov Network Parameterization

- Markov networks have too many degrees of freedom
  - A clique over n binary variables has $2^n$ parameters but the joint has only $2^n-1$ parameters
  - The network A—B—C has clique \{A,B\} and \{B,C\}
    - Both capture information on B which we can choose where we want to encode (in which clique)
    - We can add/subtract between the cliques
  - We can come up with infinitely many sets of factor values that lead to the same distribution

- Need: conventions for avoiding ambiguity in parameterization
  - Can be done using a canonical parameterization (see K&F 4.4.2.1)

Factor Graphs

- From the Markov network structure we do not know whether parameterization involves maximal cliques or edge potentials
  - Example: fully connected graph may have pairwise potentials or one large (exponential) potential over all nodes

- Solution: Factor Graphs
  - Undirected graph
  - Two types of nodes
    - Variable nodes
    - Factor nodes
  - Parameterization
    - Each factor node is associated with exactly one factor
    - Scope of factor are all neighbor variables of the factor node
**Factor Graphs**

- Example
  - Exponential (joint) parameterization
  - Pairwise parameterization

![Factor graphs diagram]

- **Markov network**
  - Factor graph for joint parameterization
  - Factor graph for pairwise parameterization

**Local Structure**

- Factor graphs still encode complete tables

- A feature $\phi[D]$ on variables $D$ is an indicator function that for some $y \in D$:

  $\phi[D] = \begin{cases} 
  1 & \text{when } x = w \\
  0 & \text{otherwise}
  \end{cases}$

- A distribution $P$ is a log-linear model over $H$ if it has
  - Features $\phi_1[D_1], \ldots, \phi_k[D_k]$ where each $D_i$ is a complete subgraph in $H$
  - A set of weights $w_1, \ldots, w_k$ such that

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \exp\left[-\sum_{i=1}^{k} w_i \phi[D_i]\right]$$
Feature Representation

- Several features can be defined on one clique
  → any factor can be represented by features, where in the most general case we define a feature and weight for each entry in the factor

- Log-linear model is more compact for many distributions especially with large domain variables

- Representation is intuitive and modular
  - Features can be modularly added between any interacting sets of variables

Markov Network Parameterizations

- Choice 1: Markov network
  - Product over potentials
  - Right representation for discussing independence queries

- Choice 2: Factor graph
  - Product over graphs
  - Useful for inference (later)

- Choice 3: Log-linear model
  - Product over feature weights
  - Useful for discussing parameterizations
  - Useful for representing context specific structures

- All parameterizations are interchangeable
Domain Application: Vision

- The image segmentation problem
  - Task: Partition an image into distinct parts of the scene
  - Example: separate water, sky, background

Markov Network for Segmentation

- Grid structured Markov network
- Random variable $X_i$ corresponds to pixel $i$
  - Domain is $\{1,...,K\}$
  - Value represents region assignment to pixel $i$
- Neighboring pixels are connected in the network
Markov Network for Segmentation

- Appearance distribution
  - $w^k_i$ - extent to which pixel $i$ “fits” region $k$ (e.g., difference from typical pixel for region $k$)
  - Introduce node potential $\exp(-w^k_i \mathbf{1}{X_i=k})$

- Edge potentials
  - Encodes contiguity preference by edge potential $\exp(\lambda \mathbf{1}{X_i=X_j})$ for $\lambda > 0$

- Solution: inference
  - Find most likely assignment to $X_i$ variables
Example Results

Result of segmentation using node potentials alone, so that each pixel is classified independently.

Result of segmentation using a pairwise Markov network encoding interactions between adjacent pixels.

Summary: Markov Network Representation

- Independencies in graph $H$
  - Global independencies $I(H) = \{ (X \perp Y | Z) : \text{sep}_H(X;Y|Z) \}$
  - Local independencies $L(H) = \{ (X \perp \text{U} - \{X\} - \text{MB}_H(X) | \text{MB}_H(X)) : X \in H \}$
  - Pairwise independencies $P(H) = \{ (X \perp Y | \text{U} - \{X,Y\}) : X - Y \notin H \}$
  - For any positive distribution $P$, they are equivalent.
- (Reverse) factorization theorem: I-map $\leftrightarrow$ factorization
- Markov network factors
  - Has to encompass cliques
  - Maximal cliques, edge factors
- Log-linear model
  - Features instead of factors
- Pairwise Markov network
  - Node/edge potentials
  - Application in vision (image segmentation)
- What next?
  - Constructing Markov networks from Bayesian networks
  - Hybrid models (e.g. Conditional Random Fields)