Overview

In this lecture

- Introduction to graphical models
- Basics of undirected models (called Markov networks)
- Basics of log-linear models
- Basics of inference in undirected models

Not in this lecture (but partly in subsequent lectures)

- Applications
- Directed models (called Bayesian networks)
- High coverage or in-depth discussion of inference
- Learning (neither structure nor parameters)
Outline

1. Introduction to Graphical Models
2. Preliminaries
3. Markov Networks
4. Log-Linear Models
5. Inference
6. Summary
Reasoning with uncertainty

• Goal: Automated reasoning system
  ▶ Take all available information
    (e.g., patient information: symptoms, test results, personal data)
  ▶ Reach conclusions
    (e.g., which diseases the patient has, which medication to give)

• Desiderata
  1. Separation of knowledge and reasoning
     ▶ Model-based representation of knowledge (ideal: declarative)
     ▶ General suite of reasoning algorithms, applicable to many domains
  2. Principled treatment of uncertainty
     ▶ Partially observed data
     ▶ Noisy observations
     ▶ Non-deterministic relationships

• Lots of applications
  ▶ Medical diagnosis, fault diagnosis, analysis of genetic and genomic data, communication and coding, analysis of marketing data, speech recognition, natural language understanding, segmenting and denoising images, social network analysis, ...
Probabilistic models

- Multiple interrelated aspects may relate to the reasoning task
  - Possible diseases
  - Hundreds of symptoms and diagnostic tests
  - Personal characteristics

1. Characterize data by a set of random variables
   - Flu (yes / no)
   - Hayfever (yes / no)
   - Season (Spring / Sommer / Autumn / Winter)
   - Congestion (yes / no)
   - MusclePain (yes / no)

   Variables and their possible values important design decision

2. Model dependencies by a *joint distribution*
   - Diseases, season, and symptoms are interdependent
   - Probabilistic models construct joint probability space
     - \( 2 \cdot 2 \cdot 4 \cdot 2 \cdot 2 \) outcomes (64 values, 63 non-redundant)
   - Given joint probability space, interesting questions can be answered

\[
P(\text{Flu} \mid \text{Season}=\text{Spring}, \text{Congestion}, \neg \text{MusclePain})
\]

Specifying a joint distribution is infeasible in general!
Probabilistic graphical models are a …

- Graph-based representation of **direct probabilistic interactions**
- Break-down of high-dimensional distributions into smaller **factors** (here: 63 vs. 17 non-redundant parameters)
- Compact representation of a set of **conditional independencies**

**Graph representation**

**Factorization**

\[
P(S, F, H, M, C) = P(S) \cdot P(F | S) \cdot P(H | S) \cdot P(C | F, H) \cdot P(M | F)
\]

**Independencies**

\[(F \perp H | S), (C \perp M, S | F, H), (M \perp H, C, S | F)\]
Main components

1. Representation
   ▶ Conciseness
     ▶ Variables tend to interact \textit{directly} only with very few others
     ▶ Natural and compact encoding as graphical model
   ▶ Transparency
     ▶ Models can be understood/evaluated by human experts

2. Inference
   ▶ Answer questions using the distribution as model of the world
   ▶ Work on graph structure
     → Orders of magnitude faster than working on joint probability

3. Learning
   ▶ Learn a model from data that captures past experience
   ▶ Involves learning parameters, sometimes also structure
     → Often better reflection of domain than hand-constructed models, sometimes surprising insights

Graphical models exploit locality structure, which (approximately) appears in many distributions arising in practice.
Outline

1. Introduction to Graphical Models

2. Preliminaries

3. Markov Networks

4. Log-Linear Models

5. Inference

6. Summary
Recap: conditional probability

- The **conditional probability** of $A$, given that $B$ is true, is defined as

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \quad \text{if } P(B) > 0$$

- Can be represented in a **conditional probability table** (CPT)

<table>
<thead>
<tr>
<th>Joint probabilities</th>
<th>CPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(X, Y)$</td>
<td>$P(X \mid Y)$</td>
</tr>
<tr>
<td>$X=H$</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.3</td>
</tr>
<tr>
<td>$X=T$</td>
<td>H</td>
</tr>
<tr>
<td>H</td>
<td>0.25</td>
</tr>
<tr>
<td>T</td>
<td>0.75</td>
</tr>
</tbody>
</table>

- Generally, for sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ of random variables,

$$P(\mathcal{X} \mid \mathcal{Y}, \mathcal{Z}) = \frac{P(\mathcal{X}, \mathcal{Y} \mid \mathcal{Z})}{P(\mathcal{Y} \mid \mathcal{Z})}$$
Recap: conditional independence

- Unconditional independence is rare
- $X$ and $Y$ are **conditionally independent** given $Z$ iff

\[ P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z) \]

- Denoted $X \perp Y \mid Z$
- Example
  - Event that it rains tomorrow ($X$)
  - Event that ground is wet ($Y$)
  - Event that it rains now ($Z$)
  - Then: $X \not\perp Y$ but $X \perp Y \mid Z$

- Generally, for sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ of random variables, \[ \mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z} \iff P(\mathcal{X}, \mathcal{Y} \mid \mathcal{Z}) = P(\mathcal{X} \mid \mathcal{Z})P(\mathcal{Y} \mid \mathcal{Z}) \]
Properties of conditional independence

- In general, $\mathcal{X} \perp \mathcal{Y}$ does not imply nor is implied by $\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}$

- The following relationships hold:

\[
(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}) \iff (\mathcal{Y} \perp \mathcal{X} \mid \mathcal{Z}) \quad \text{(symmetry)}
\]

\[
(\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) \implies (\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}) \quad \text{(decomposition)}
\]

\[
(\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) \implies (\mathcal{X} \perp \mathcal{Y} \mid \mathcal{W}, \mathcal{Z}) \quad \text{(weak union)}
\]

\[
(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}) \land (\mathcal{X} \perp \mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) \implies (\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) \quad \text{(contraction)}
\]

- For positive distributions and mutually disjoint sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{W}$:

\[
(\mathcal{X} \perp \mathcal{Y} \mid \mathcal{W}, \mathcal{Z}) \land (\mathcal{X} \perp \mathcal{W} \mid \mathcal{Y}, \mathcal{Z}) \implies (\mathcal{X} \perp \mathcal{Y}, \mathcal{W} \mid \mathcal{Z}) \quad \text{(intersection)}
\]
Example

- Conditional independencies
  - $F \perp H \mid S$
  - $C \perp M, S \mid F, H$
  - $M \perp H, C, S \mid F$

- By applying definitions and properties of conditional independence, implies factorization

\[
P(S, F, H, M, C) = P(S) P(F \mid S) P(H \mid S) P(M \mid F) P(C \mid F, H)
\]
Querying a distribution (1)

Consider a joint distribution on a set of variables $\mathcal{X}$.

- Let $\mathcal{E} \subseteq \mathcal{X}$ be a set of \textbf{evidence variables} that take values $\mathbf{e}$
- Then $\mathcal{W} = \mathcal{X} - \mathcal{E}$ is the set of \textbf{latent variables}
- Let $\mathcal{Y} \subseteq \mathcal{W}$ be a set of \textbf{query variables}
- Then $\mathcal{Z} = \mathcal{W} - \mathcal{Y}$ is the set of \textbf{non-query variables}

Example

- $\mathcal{X} = \{ \text{Season, Congestion, MusclePain, Flu, Hayfever} \}$
- $\mathcal{E} = \{ \text{Season, Congestion, MusclePain} \}$
- $\mathbf{e} = \{ \text{Spring, Yes, No} \}$
- $\mathcal{W} = \{ \text{Flu, Hayfever} \}$
- $\mathcal{Y} = \{ \text{Flu} \}$
- $\mathcal{Z} = \{ \text{Hayfever} \}$
Querying a distribution (2)

\[ e = \{ \text{Spring, Congestion, } \neg \text{MusclePain} \} \]

<table>
<thead>
<tr>
<th>( P(\mathcal{W} \mid e) )</th>
<th>Flu</th>
<th>( \neg )Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hayfever</td>
<td>5%</td>
<td>35%</td>
</tr>
<tr>
<td>( \neg )Hayfever</td>
<td>40%</td>
<td>20%</td>
</tr>
</tbody>
</table>

1. **Conditional probability query**
   - Compute the *posterior distribution* \( P(Y \mid e) \) of the query variables
   - E.g., \( P(\text{Flu} \mid e) \rightarrow \text{Yes (45%), No (55%)} \)

2. **MAP query**
   - Compute the most likely value of the latent variables
     \[
     \text{MAP}(\mathcal{W} \mid e) = \arg\max_w P(w \mid e) = \arg\max_w P(w, e)
     \]
   - E.g., MAP(\text{Flu, Hayfever} \mid e) \( \rightarrow \) Only flu

3. **Marginal MAP query**
   - Compute the most likely value of the query variables
     \[
     \text{MAP}(Y \mid e) = \arg\max_y P(y \mid e) = \arg\max_y \sum_z P(y, z, e)
     \]
   - E.g., MAP(\text{Flu} \mid e) \( \rightarrow \) No flu (!)
Querying graphical models

- Graphical models induce conditional independences
- Queries reason about dependencies between variables

Can we evaluate queries more efficiently given a graphical model and its associated independences?

- Yes, we can!

| $P(\mathcal{W} | \mathcal{E})$ | Flu | ¬Flu |
|-----------------------------|-----|------|
| Hayfever                   | 18% | 12%  |
| ¬Hayfever                  | 42% | 28%  |

Table known to satisfy $F \perp H | \mathcal{E}$

- E.g., monotonicity now holds:
  \[ \text{MAP}(\text{Flu}, \text{Hayfever} | \mathcal{E}) = (\text{MAP}(\text{Flu} | \mathcal{E}), \text{MAP}(\text{Hayfever} | \mathcal{E})) \]

- Independence properties help inference!
Outline

1. Introduction to Graphical Models
2. Preliminaries
3. Markov Networks
4. Log-Linear Models
5. Inference
6. Summary
Misconception example

- Alice, Bob, Charles, Debbie study in pairs for the HTML exam

- I misspoke in class, giving rise to a possible misconception
- Some students figured out the problem, others did not

Which of the students has the misconception?
- If A does not have the misconception, she may help B and D
  \[\rightarrow\] Students influence each other
- If A has the misconception, she may be helped by B and D
  \[\rightarrow\] Influence has no natural “direction”
- A does not study with C
  \[\rightarrow\] No *direct* influence between A and C
A **Markov network** is an undirected graph $\mathcal{H} = (\mathcal{X}, E)$, where $\mathcal{X}$ is a set of random variables and $E \subseteq \mathcal{X} \times \mathcal{X}$ is a set of edges.

\[ \mathcal{X} = \{ A, B, C, D \} \]

\[ E = \{ (A, B), (B, C), (C, D), (D, A) \} \]

We will see that Markov networks encode a set of conditional independence assumptions between its variables.
Local models

**Definition**

Let $\mathcal{D}$ be a set of random variables. A factor $\phi$ is a function from $\text{Val}(\mathcal{D}) \rightarrow \mathbb{R}^+$. The set $\mathcal{D}$ is called the **scope** of the factor and is denoted $\text{Scope}[\phi]$.

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$\phi_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a^0$</td>
<td>$b^0$</td>
<td>30</td>
</tr>
<tr>
<td>$a^0$</td>
<td>$b^1$</td>
<td>5</td>
</tr>
<tr>
<td>$a^1$</td>
<td>$b^0$</td>
<td>1</td>
</tr>
<tr>
<td>$a^1$</td>
<td>$b^1$</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$B$</th>
<th>$C$</th>
<th>$\phi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b^0$</td>
<td>$c^0$</td>
<td>100</td>
</tr>
<tr>
<td>$b^1$</td>
<td>$c^0$</td>
<td>1</td>
</tr>
<tr>
<td>$b^1$</td>
<td>$c^1$</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$C$</th>
<th>$D$</th>
<th>$\phi_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c^0$</td>
<td>$d^0$</td>
<td>1</td>
</tr>
<tr>
<td>$c^1$</td>
<td>$d^0$</td>
<td>100</td>
</tr>
<tr>
<td>$c^1$</td>
<td>$d^1$</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D$</th>
<th>$A$</th>
<th>$\phi_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d^0$</td>
<td>$a^0$</td>
<td>100</td>
</tr>
<tr>
<td>$d^0$</td>
<td>$a^1$</td>
<td>1</td>
</tr>
<tr>
<td>$d^1$</td>
<td>$a^0$</td>
<td>1</td>
</tr>
<tr>
<td>$d^1$</td>
<td>$a^1$</td>
<td>100</td>
</tr>
</tbody>
</table>

- Factors describe “compatibility” between values (not normalized)
- $\phi_1$: More “weight” when $A$ and $B$ agree than when they disagree
- $\phi_1$: More weight when $A$ and $B$ are both right than when both are wrong
- $\phi_1$: If they disagree, more weight when $A$ is right than when $B$ is right
Combining local models

Definition

Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be three disjoint sets of random variables and let $\phi_1(\mathcal{X}, \mathcal{Y})$ and $\phi_2(\mathcal{Y}, \mathcal{Z})$ be two factors. The factor product $\psi = \phi_1 \times \phi_2$ is given by the factor $\psi : \text{Val}(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) \rightarrow \mathbb{R}^+$ with

$$\psi(\mathcal{X}, \mathcal{Y}, \mathcal{Z}) = \phi_1(\mathcal{X}, \mathcal{Y}) \cdot \phi_2(\mathcal{Y}, \mathcal{Z}).$$
Factor products and the product rule

Recall the product rule of probability

\[ P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{Y}) P(\mathcal{X} | \mathcal{Y}). \]

<table>
<thead>
<tr>
<th>MusclePain</th>
<th>Flu</th>
<th>MusclePain</th>
<th>Flu, MusclePain</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>P</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>Yes</td>
<td>0.1</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
<td>0.9</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

- Set \( \phi_1(\text{MusclePain}) = P(\text{MusclePain}) \)
- Set \( \phi_2(\text{MusclePain}, \text{Flu}) = P(\text{Flu} | \text{MusclePain}) \)
- Set \( \psi(\text{MusclePain}, \text{Flu}) = P(\text{MusclePain}, \text{Flu}) \)
- Then \( \psi = \phi_1 \times \phi_2 \)

Factor products generalize the product rule of probability.
Gibbs distribution

**Definition**

A distribution $P_\Phi$ is a **Gibbs distribution** parameterized by a set of factors $\Phi = \{ \phi_1(X_1), \ldots, \phi_m(X_m) \}$ if it is defined by

$$
P_\Phi(X) = \frac{1}{Z} \tilde{P}_\Phi(X)
$$

$$
\tilde{P}_\Phi(X) = \phi_1(X_1) \times \phi_2(X_2) \times \cdots \times \phi_m(X_m)
$$

$$
Z = \sum_{x \in \text{Val}(\mathcal{X})} \tilde{P}_\Phi(x)
$$

Here, $\tilde{P}_\Phi(X)$ is an *unnormalized measure* and $Z$ a normalizing constant called the **partitioning function**.

- Factors *contribute* to the overall joint distribution
- Contributions from *all* factors determine joint distribution
- Gibbs distribution is **positive** if all factors have range $\mathbb{R}^+ \setminus \{0\}$
- $x$ is sometimes called a **possible world**

A set of factors defines a Gibbs distribution, i.e., a joint probability distribution over all variables.
Gibbs distribution for Misconception example

\[
\begin{align*}
\phi_1 & \\
\phi_2 & \\
\phi_3 & \\
\phi_4 & \\
\end{align*}
\]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>(\hat{P})</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_0)</td>
<td>(b_0)</td>
<td>(c_0)</td>
<td>(d_0)</td>
<td>300,000</td>
<td>0.04</td>
</tr>
<tr>
<td>(a_0)</td>
<td>(b_0)</td>
<td>(c_0)</td>
<td>(d_1)</td>
<td>300,000</td>
<td>0.04</td>
</tr>
<tr>
<td>(a_0)</td>
<td>(b_0)</td>
<td>(c_1)</td>
<td>(d_0)</td>
<td>300,000</td>
<td>0.04</td>
</tr>
<tr>
<td>(a_0)</td>
<td>(b_0)</td>
<td>(c_1)</td>
<td>(d_1)</td>
<td>30</td>
<td>(4.1 \cdot 10^{-6})</td>
</tr>
<tr>
<td>(a_0)</td>
<td>(b_1)</td>
<td>(c_0)</td>
<td>(d_0)</td>
<td>500</td>
<td>(6.9 \cdot 10^{-5})</td>
</tr>
<tr>
<td>(a_0)</td>
<td>(b_1)</td>
<td>(c_0)</td>
<td>(d_1)</td>
<td>500</td>
<td>(6.9 \cdot 10^{-5})</td>
</tr>
<tr>
<td>(a_0)</td>
<td>(b_1)</td>
<td>(c_1)</td>
<td>(d_0)</td>
<td>5,000,000</td>
<td>0.69</td>
</tr>
<tr>
<td>(a_0)</td>
<td>(b_1)</td>
<td>(c_1)</td>
<td>(d_1)</td>
<td>500</td>
<td>(6.9 \cdot 10^{-5})</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_0)</td>
<td>(c_0)</td>
<td>(d_0)</td>
<td>100</td>
<td>(1.4 \cdot 10^{-5})</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_0)</td>
<td>(c_0)</td>
<td>(d_1)</td>
<td>1,000,000</td>
<td>0.14</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_0)</td>
<td>(c_1)</td>
<td>(d_0)</td>
<td>100</td>
<td>(1.4 \cdot 10^{-5})</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_0)</td>
<td>(c_1)</td>
<td>(d_1)</td>
<td>100</td>
<td>(1.4 \cdot 10^{-5})</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_1)</td>
<td>(c_0)</td>
<td>(d_0)</td>
<td>10</td>
<td>(1.4 \cdot 10^{-6})</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_1)</td>
<td>(c_0)</td>
<td>(d_1)</td>
<td>100,000</td>
<td>0.014</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_1)</td>
<td>(c_1)</td>
<td>(d_0)</td>
<td>100,000</td>
<td>0.014</td>
</tr>
<tr>
<td>(a_1)</td>
<td>(b_1)</td>
<td>(c_1)</td>
<td>(d_1)</td>
<td>100,000</td>
<td>0.014</td>
</tr>
</tbody>
</table>

\(Z = 7,201,840\)
Factorization and factor graphs

Definition

A Gibbs distribution $P_\Phi$ with $\Phi = \{ \phi_1(X_1), \ldots, \phi_m(X_m) \}$ factorizes over a Markov network $\mathcal{H}$ if each $X_i$ is a complete subgraph of $\mathcal{H}$. The factors $\phi_i$ are often called *clique potentials*.

- $\Phi = \{ \phi_1(A, B), \phi_2(B, C), \phi_3(C, D), \phi_4(D, A) \}$
- $\Psi = \{ \psi_1(A, B, D), \psi_2(B, C) \}$
- $P_\Phi$ factorizes over both $\mathcal{H}_1$ and $\mathcal{H}_2$
- $P_\Psi$ factorizes over only $\mathcal{H}_2$
Active paths

Definition

Let \( X_1 \rightarrow \ldots \rightarrow X_k \) be a path in \( \mathcal{H} = (\mathcal{X}, E) \). Let \( Z \subseteq \mathcal{X} \) be a set of observed variables. The path \( X_1 \rightarrow \ldots \rightarrow X_k \) is **active** given \( Z \) if \( X_i \notin Z \) for \( 1 \leq i \leq k \).

Some active paths given \( A \):
- \( D \rightarrow C \)
- \( C \rightarrow B \)
- \( D \rightarrow C \rightarrow B \)

Some inactive paths given \( A \):
- \( D \rightarrow A \rightarrow B \)
- \( C \rightarrow D \rightarrow A \rightarrow B \)
Separation and independencies for Markov networks

Definition

We say that a set of nodes $Z$ separates $X$ and $Y$ in $H$, denoted $\text{sep}_H(X; Y \mid Z)$, if there is no active path given $Z$ between any node in $X$ and any node in $Y$. We associate with $H$ the following set of independencies:

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

- $\emptyset$ does not separate any nodes
- $\{ A \}$ does not separate any nodes
- $\{ A, C \}$ separates $\{ B \}$ and $\{ D \}$
- $\{ A, B, C \}$ does not separate any nodes

$$\mathcal{I}(H) = \{ (B \perp D \mid A, C), (D \perp B \mid A, C), (A \perp C \mid B, D), (C \perp A \mid B, D) \}$$
Gibbs distributions and Markov networks

- Let $P$ be a probability distribution. Define $\mathcal{I}(P)$ to be the set of independence assertions of the form $\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}$ that hold in $P$.
- A Markov network $\mathcal{H}$ is an **I-map** for $P$ if $\mathcal{I}(\mathcal{H}) \subseteq \mathcal{I}(P)$.

**Theorem (Soundness $\rightarrow$)**

*If $P$ is a Gibbs distribution that factorizes over $\mathcal{H}$, then $\mathcal{H}$ is an I-map for $P$.***

**Theorem (Hammersley-Clifford theorem, soundness $\leftarrow$)**

*Let $P$ be a positive distribution. If $\mathcal{H}$ is an I-map for $P$, then $P$ is a Gibbs distribution that factorizes over $\mathcal{H}$.***

**Theorem (Completeness)**

*If $\mathcal{X}$ and $\mathcal{Y}$ are not separated given $\mathcal{Z}$ in $\mathcal{H}$, then $\mathcal{X}$ and $\mathcal{Y}$ are dependent given $\mathcal{Z}$ for some distribution $P$ that factorizes over $\mathcal{H}$.***
Application: Image denoising

Original  Noisy image  Denoised image

Li and Huttenlocher, ECCV 2008
Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names.

- Local evidence often strong clue for label
- Long-range evidence (label consistency) helps when local evidence is insufficient
Outline

1. Introduction to Graphical Models
2. Preliminaries
3. Markov Networks
4. Log-Linear Models
5. Inference
6. Summary
Log-linear model

Definition

A positive distribution $P$ is a log-linear model over a Markov network $\mathcal{H}$ if it is associated with

- a set of features $\mathcal{F} = \{ f_1(\mathcal{X}_1), \ldots, f_m(\mathcal{X}_m) \}$, where each $\mathcal{X}_i$ is a complete subgraph in $\mathcal{H}$ and $f_i : \mathcal{X}_i \rightarrow \mathbb{R}$
- a set of weights $w_1, \ldots, w_m$

such that

$$P(\mathcal{X}) \propto \exp \left[ \sum_{i=1}^{m} w_i f_i(\mathcal{X}_i) \right].$$

The terms $\epsilon_i(\mathcal{X}_i) = -w_i f_i(\mathcal{X}_i)$ are called energy functions.

$log P(\mathcal{X})$ is a linear combination of the features. The linearity allows us to detect and eliminate redundancy in the features (using standard linear algebra techniques).
From factors to features

Let $\mathcal{X}$ be a subset of variables. An **indicator feature** is a function $f : \mathcal{X} \rightarrow \{0, 1\}$.

**Theorem**

*Every factor of a positive Gibbs distribution on discrete variables can be expressed in terms of a linear combination of weighted indicator features.*

**Proof (Boolean case).** Consider a factor $\phi(X_1, \ldots, X_k)$ on $k$ Boolean variables. Let $\Theta$ be the set of all *assignments* of values to $X_1, \ldots, X_k$. Set

\[
\begin{align*}
  w_\theta &= \ln \phi(X_1[\theta], \ldots, X_k[\theta]) \quad \text{(constants)} \\
  f_\theta(X_1, \ldots, X_k) &= \begin{cases} 
    1 & \text{if } X_1 = X_1[\theta], \ldots, X_k = X_k[\theta] \\
    0 & \text{otherwise}
  \end{cases} \quad \text{(indicator features)} \\
  \ln \phi(X_1, \ldots, X_k) &= \sum_{\theta \in \Theta} w_\theta f_\theta(X_1, \ldots, X_k) \quad \text{(decomposition)}
\end{align*}
\]
From factors to features (example)

Consider three friends with similar interests and let $A$, $B$, $C$ be Boolean variables that indicate whether each of the friends likes football.

We have

$$\ln \phi(A, B, C) = \sum_{\theta} w_{\theta} f_{\theta}(A, B, C) = 2.3 \cdot f_{FFF}(A, B, C) + 2.3 \cdot f_{TTT}(A, B, C).$$

Even more compact: $\ln \phi(A, B, C) = 2.3 \cdot l_{ABC \lor \neg A \lor \neg B \lor \neg C}$
From Gibbs distribution to log-linear models

**Theorem**

*Every positive Gibbs distribution* $P$ *over* $\mathcal{H}$ *on Boolean variables* $X_1, \ldots, X_n$ *has a log-linear model over* $\mathcal{H}$ *with only indicator features and vice versa.*

**Proof.**

$$P(\mathcal{X}) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(\mathcal{X}_i)$$

$$= \frac{1}{Z} \exp \left[ \sum_{i=1}^{m} \ln \phi_i(\mathcal{X}_i) \right]$$

$$= \frac{1}{Z} \exp \left[ \sum_{i=1}^{m} \sum_{\theta \in \Theta X_i} w_{\theta} f_{i,\theta}(\mathcal{X}_i) \right].$$
Example: Markov Logic Networks (1)

Smoking and Quitting in Groups
Researchers studying a network of 12,067 people found that smokers and nonsmokers tended to cluster in groups of close friends and family members. As more people quit over the decades, remaining groups of smokers were increasingly pushed to the periphery of the social network.

1971 A sample of 1,000 people from the study includes many large groups of smokers.

2000 Nearly three decades later, groups of smokers tended to be smaller and more isolated.

Does Bob smoke?

MLNs learn correlations from structured data, then apply to new data.
Example: Markov Logic Networks (2)

<table>
<thead>
<tr>
<th>Name1</th>
<th>Name2</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna</td>
<td>Bob</td>
<td>Yes</td>
</tr>
<tr>
<td>Bob</td>
<td>Anna</td>
<td>Yes</td>
</tr>
<tr>
<td>Anna</td>
<td>Anna</td>
<td>Yes</td>
</tr>
<tr>
<td>Bob</td>
<td>Bob</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anna</td>
<td>No</td>
</tr>
</tbody>
</table>

1.5 \{ \text{Smoking causes cancer} \}
\forall x. \text{Smokes}(x) \implies \text{Cancer}(x)

1.1 \{ \text{Friends have similar smoking habits} \}
\forall x. \forall y. \text{Friends}(x, y) \implies (\text{Smokes}(x) \iff \text{Smokes}(y))

Inference result
- \( P(\text{Bob smokes}) = 84.6\% \)
- \( P(\text{Bob has cancer}) = 76.9\% \)

<table>
<thead>
<tr>
<th>S(B)</th>
<th>C(B)</th>
<th>#R1</th>
<th>#R2</th>
<th>( \sum w )</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>1</td>
<td>1</td>
<td>2.6</td>
<td>7.7%</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>1</td>
<td>1</td>
<td>2.6</td>
<td>7.7%</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>0</td>
<td>3</td>
<td>3.3</td>
<td>15.4%</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
<td>3</td>
<td>4.8</td>
<td>69.2%</td>
</tr>
</tbody>
</table>

Example is simplified; actual semantics are slightly different.

Richardson and Domingos, 2006.
Outline

1. Introduction to Graphical Models
2. Preliminaries
3. Markov Networks
4. Log-Linear Models
5. Inference
6. Summary
Inference in probabilistic graphical models

- Inference = answer queries

- Recall the queries of interest
  - \( e \) = evidence var’s, \( \mathcal{V} \) = query var’s, \( \mathcal{W} \) = latent var’s
  1. Conditional probability query: \( P(\mathcal{V} | e) \)
  2. MAP query: \( P(\mathcal{W} | e) \)
  3. Marginal MAP query: \( P(\mathcal{V} | e) \)

- These queries are often difficult to answer

- How difficult?
  - Let \( P_\Phi \) be a Gibbs distribution over variables \( \{X, X_1, \ldots, X_n\} \)
  - \( P_\Phi \)-decision problem: Is \( P_\Phi(X = x) > 0 \)?
  - \( P_\Phi \)-probability computation problem: What is \( P_\Phi(X = x) \)?

- Can show
  - \( P_\Phi \)-decision problem is NP-complete
  - \( P_\Phi \)-probability computation is \#P-hard
Complexity of inference in probabilistic graphical models

**Proof** (by reduction from 3-SAT and \#3-SAT).

- Take a 3-SAT formula $\Psi = C_1 \land C_2 \land \ldots \land C_m$ over variables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$
- Consider the following Gibbs distribution $P_\Phi$ over Booleans:

\[
\begin{align*}
X_1 & \quad X_2 & \quad X_3 & \quad X_4 & \quad \cdots & \quad X_n \\
C_1 & \quad C_2 & \quad C_3 & \quad C_4 & \quad \cdots & \quad C_m \\
A_1 & \quad A_2 & \quad A_3 & \quad A_4 & \quad \cdots & \quad A_m \\
\end{align*}
\]

- Set $\lor_i(C_i, \mathcal{X}_i) = 1$ if for assignment $\mathcal{X}_i$ the truth value of clause $C_i$ equals variable $C_i$, else 0; similarly the $\land$-factors
- $P_\Phi$ can be computed in polynomial time in the size of $\Psi$
- $P_\Phi(X = T) > 0$ if and only if $\Psi$ is satisfiable $\rightarrow$ NP-complete
- Now note that $P(X_i = T) = 1/2$ and $P(x) = 1/2^n$
- $P_\Phi(X = T) = \#\Psi/2^n \rightarrow \#P$-hard
Inference

• Many methods exist

• Exact inference
  ▶ Naive
  ▶ Variable elimination
  ▶ Clique trees
  ▶ Optimization methods
  ▶ ...

• Approximate inference
  ▶ Loopy belief propagation
  ▶ Variational inference
  ▶ Particle-based inference
  ▶ ...

• Here we briefly introduce the red algorithms for Boolean variables
Naive approach

Exponential in number of variables!

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>( \bar{P} )</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_0 )</td>
<td>( b_0 )</td>
<td>( c_0 )</td>
<td>( d_0 )</td>
<td>300,000</td>
<td>0.04</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>( b_0 )</td>
<td>( c_0 )</td>
<td>( d_1 )</td>
<td>300,000</td>
<td>0.04</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>( b_0 )</td>
<td>( c_1 )</td>
<td>( d_0 )</td>
<td>300,000</td>
<td>0.04</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>( b_0 )</td>
<td>( c_1 )</td>
<td>( d_1 )</td>
<td>30</td>
<td>4.1 \cdot 10^{-6}</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>( b_1 )</td>
<td>( c_0 )</td>
<td>( d_0 )</td>
<td>500</td>
<td>6.9 \cdot 10^{-5}</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>( b_1 )</td>
<td>( c_0 )</td>
<td>( d_1 )</td>
<td>500</td>
<td>6.9 \cdot 10^{-5}</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>( b_1 )</td>
<td>( c_1 )</td>
<td>( d_0 )</td>
<td>5,000,000</td>
<td>0.69</td>
</tr>
<tr>
<td>( a_0 )</td>
<td>( b_1 )</td>
<td>( c_1 )</td>
<td>( d_1 )</td>
<td>500</td>
<td>6.9 \cdot 10^{-5}</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_0 )</td>
<td>( c_0 )</td>
<td>( d_0 )</td>
<td>100</td>
<td>1.4 \cdot 10^{-5}</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_0 )</td>
<td>( c_0 )</td>
<td>( d_1 )</td>
<td>1,000,000</td>
<td>0.14</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_0 )</td>
<td>( c_1 )</td>
<td>( d_0 )</td>
<td>100</td>
<td>1.4 \cdot 10^{-5}</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_0 )</td>
<td>( c_1 )</td>
<td>( d_1 )</td>
<td>10</td>
<td>1.4 \cdot 10^{-6}</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_1 )</td>
<td>( c_0 )</td>
<td>( d_0 )</td>
<td>100,000</td>
<td>0.014</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_1 )</td>
<td>( c_0 )</td>
<td>( d_1 )</td>
<td>100,000</td>
<td>0.014</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_1 )</td>
<td>( c_1 )</td>
<td>( d_0 )</td>
<td>100,000</td>
<td>0.014</td>
</tr>
<tr>
<td>( a_1 )</td>
<td>( b_1 )</td>
<td>( c_1 )</td>
<td>( d_1 )</td>
<td>100,000</td>
<td>0.014</td>
</tr>
</tbody>
</table>

\[ Z = 7,201,840 \]
Variable elimination (idea)

Goal: Eliminate non-query variables from the graph. E.g., $B$:

$\Phi_1 \times \Phi_2$

$\Phi_{12}$

$B$ has been eliminated ("margin-alized out"). The resulting factor graph represents $P(A, C, D)$. 
Variable elimination (why it works)

Recall that

\[ P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B) \times \phi_2(B, C) \times \phi_3(C, D) \times \phi_4(D, A) \]

and thus

\[ P(A, C, D) = P(A, b^0, C, D) + P(A, b^1, C, D) \]

\[ = \frac{1}{Z} [\phi_1(A, b^0) \times \phi_2(b^0, C) \times \phi_3(C, D) \times \phi_4(D, A) \]

\[ + \phi_1(A, b^1) \times \phi_2(b^1, C) \times \phi_3(C, D) \times \phi_4(D, A)] \]

\[ = \frac{1}{Z} \left[ \left\{ \sum_{b \in \{b^0, b^1\}} \phi_1(A, b) \times \phi_2(b, C) \right\} \times \phi_3(C, D) \times \phi_4(D, A) \right] \]

\[ = \frac{1}{Z} [\phi_{12}(A, C) \times \phi_3(C, D) \times \phi_4(D, A)] \]
Variable elimination (remarks)

- Also called sum-product variable elimination
- Whenever we eliminate a variable $B$
  - We remove all factors connected to $B$
  - We introduce a single factor that is connected to the neighbors of $B$
  - If $B$ has $k$ Boolean neighbors, the new factor has $2^k$ rows
    → Potentially exponential blow-up

- Computational cost
  - Dominated by sizes of intermediate factors
  - Depends strongly on elimination ordering
  - NP-hard to find optimal ordering
  - Lots of useful heuristics exist
  - “Conditioning” can be used to avoid large factors for increased processing time

- Similar observations give rise to other algorithms, e.g., message passing in clique trees
Sampling methods

- Also called particle-based approximate inference
- Idea: Obtain samples from the distribution underlying the graphical model
- If samples were independent, we could count how often each variable is true/false → very good approximation
- But independent sampling is difficult in Markov networks
  - Easier to generate dependent samples
  - Goal is to minimize the dependencies
  - Generally, more samples needed than if samples were independent
  - If dependencies vanish between far-apart samples, approach is correct and converges
- Many techniques
  - Forward sampling (for directed models)
  - Likelihood weighting
  - Importance sampling
  - Gibbs sampling
  - Other Markov Chain Monte Carlo (MCMC) methods
  - Collapsed particles
Gibbs sampling (idea)

Gibbs sampling is a simple algorithm to sample from $P(X, Y)$. It is used when it is hard to sample from $P(X, Y)$, but easy to sample from $P(X \mid Y)$ and $P(Y \mid X)$.

1. Pick an initial point $(x_0, y_0)$
2. For $n = 1, 2, \ldots$
   2.1 Generate $x_n \sim P(X \mid Y = y_{n-1})$

\[ n = 1 \]
Gibbs sampling (idea)

Gibbs sampling is a simple algorithm to sample from $P(X, Y)$. It is used when it is hard to sample from $P(X, Y)$, but easy to sample from $P(X \mid Y)$ and $P(Y \mid X)$.

1. Pick an initial point $(x_0, y_0)$
2. For $n = 1, 2, \ldots$
   2.1 Generate $x_n \sim P(X \mid Y = y_{n-1})$
   2.2 Generate $y_n \sim P(Y \mid X = x_n)$

$n = 15$

$n = 500$
Gibbs sampling for Markov networks

Recall that

\[ P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B) \times \phi_2(B, C) \times \phi_3(C, D) \times \phi_4(D, A). \]

Sampling from \( P(A, B, C, D) \) is hard but sampling from

\[ P(A \mid B, C, D) = \frac{P(A, B, C, D)}{P(B, C, D)} \]

\[ = \frac{\frac{1}{Z} \left[ \phi_1(A, B) \times \phi_2(B, C) \times \phi_3(C, D) \times \phi_4(D, A) \right]}{\frac{1}{Z} \sum_{a \in \{ a^0, a^1 \}} \left[ \phi_1(a, B) \times \phi_2(B, C) \times \phi_3(C, D) \times \phi_4(D, a) \right]} \]

\[ = \frac{\phi_1(A, B) \times \phi_4(D, A)}{\sum_{a \in \{ a^0, a^1 \}} \phi_1(a, B) \times \phi_4(D, a)} \]

is easy. Only the factors connected to \( A \) remain.

When resampling a variable \( A \), we only have to look at the factors connected to \( A \), and thus only the subset of variables connected to \( A \). These variables are called the **Markov blanket** of \( A \).
Gibbs sampling for Markov networks (remarks)

- Variables are picked according to a **schedule**
  - sequential, random, ...
- An instance of the more general class of MCMC methods
  - Markov chains describe how the sampling process moves through the set of worlds
  - Irreducible if all worlds can be reached from all other worlds
  - Convergence speed depends on how fast the sampling process moves (**mixing time**)

<table>
<thead>
<tr>
<th>World</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gibbs sampling works well (fast mixing)</td>
<td></td>
</tr>
<tr>
<td>Gibbs sampling works reasonable (slow mixing)</td>
<td></td>
</tr>
<tr>
<td>Gibbs sampling does not work (not irreducible)</td>
<td></td>
</tr>
</tbody>
</table>

- MCMC methods can perform “bigger” steps than Gibbs sampling; they change multiple variables simultaneously
Outline

1. Introduction to Graphical Models
2. Preliminaries
3. Markov Networks
4. Log-Linear Models
5. Inference
6. Summary
Lessons learned

- Probabilistic graphical models
  - Describe and reason about probability distributions and independencies
  - Exploit locality structure (conditional independence)
  - Main components: representation, inference, learning

- Markov networks (undirected graphical models)
  - Markov networks encode a set of conditional independencies
  - Gibbs distribution are parameterized by set of factors
  - Both are equivalent for positive distributions

- Inference
  - \#P-hard in general
  - Exact methods for probability computation (e.g., variable elimination) often expensive, but they do work well on certain models
  - Particle-based inference is a common approximate method
  - Gibbs sampling is the simplest MCMC method
    → Change one variable at a time
Suggested reading

- Daphne Koller, Nir Friedman
  *Chapter 4: Undirected Graphical Models*
  *In Probabilistic Graphical Models: Principles and Techniques*
  The MIT Press, 2009