CS242: Probabilistic Graphical Models
Lecture 1B: Factor Graphs, Inference, & Variable Elimination

Professor Erik Sudderth
Brown University Computer Science
September 15, 2016

Some figures and materials courtesy David Barber,
Bayesian Reasoning and Machine Learning
http://www.cs.ucl.ac.uk/staff/d.barber/brml/
CS242: Lecture 1B Outline

- Factor graphs
- Loss Functions and Marginal Distributions
- Variable Elimination Algorithms
**Factor Graphs**

\( \mathcal{F} \rightarrow \) set of hyperedges linking subsets of nodes \( f \subseteq \mathcal{V} \)

\( \mathcal{V} \rightarrow \) set of \( N \) nodes or vertices, \( \{1, 2, \ldots, N\} \)

\[ p(x) = \frac{1}{Z} \prod_{f \in \mathcal{F}} \psi_f(x_f) \quad Z = \sum \prod_{x, f \in \mathcal{F}} \psi_f(x_f) \]

\( Z > 0 \rightarrow \) normalization constant (partition function)

\( \psi_f(x_f) \geq 0 \rightarrow \) arbitrary non-negative potential function

- In a hypergraph, the hyperedges link arbitrary subsets of nodes (not just pairs)
- Visualize by a bipartite graph, with square (usually black) nodes for hyperedges
- Motivation: factorization key to computation
- Motivation: avoid ambiguities of undirected graphical models
For a given undirected graph, there exist distributions with equivalent Markov properties, but different factorizations and different inference/learning complexities.

**Undirected Graphical Model**

\[ p(x) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \]

**Pairwise (edge) Potentials**

\[ p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c) \]

**Maximal Clique Potentials**

**An Alternative Factorization**

\[ p(x) = \frac{1}{Z} \prod_{f \in \mathcal{F}} \psi_f(x_f) \]

**Recommendation:**

*Use undirected graphs for pairwise MRFs, and factor graphs for higher-order potentials.*
Example: Nearest-Neighbor Spatial MRF

- Observed nodes: Features of 2D image (intensity, color, texture, ...)
- Hidden nodes: Property of 3D world (depth, motion, object category, ...)

Stereo Vision: Yamaguchi et al., ECCV 2012

Object Segmentation: Batra et al., ECCV 2012
Higher-Order MRF Potentials

- Observed nodes: Features of 2D image (intensity, color, texture, …)
- Hidden nodes: Property of 3D world (depth, motion, object category, …)

Kohli & Rother 2012
Low Density Parity Check (LDPC) Codes

Parity Check Factors

- Each variable node is binary, so $x_s \in \{0, 1\}$
- **Parity check factors** equal 1 if the sum of the connected bits is even, 0 if the sum is odd (invalid codewords are excluded)

Evidence (observation) Factors

- **Unary evidence factors** equal probability that each bit is a 0 or 1, given data. Assumes independent “noise” on each bit.

Data Bits

Parity Bits

Inference:

0 5 15 $\infty$
Directed Graphs as Factor Graphs

Directed Graphical Model

\[ p(x) = \prod_{i=1}^{N} p(x_i \mid x_{\Gamma(i)}) \]
Directed Graphs as Factor Graphs

**Directed Graphical Model**

\[
p(x) = \prod_{i=1}^{N} p(x_i \mid x_{\Gamma(i)})
\]

**Corresponding Factor Graph**

\[
p(x) = \prod_{i=1}^{N} \psi_i(x_i, x_{\Gamma(i)})
\]

- Associate one factor with each node, linking it to its parents and defined to equal the corresponding conditional distribution.
- Information lost: Directionality of conditional distributions, and fact that global partition function \(Z=1\).
Directed Graphs as Undirected Graphs

\[ p(x) = \prod_{i=1}^{N} \psi_i(x_i, x_{\Gamma(i)}) \quad p(x) = \prod_{i=1}^{N} p(x_i | x_{\Gamma(i)}) \]

- Retain all directed edges, but drop directionality
- Create *moral graph* by linking (“marrying”) pairs of nodes with a common child
- This undirected graph has a clique for conditional of each node given its parent

---

**Figure 2.4.** Three graphical representations of a distribution over five random variables (see §175).

(a) Directed graph \( G \) depicting a causal, generative process. (b) Factor graph expressing the factorization underlying \( G \). (c) A “moralized” undirected graph capturing the Markov structure of \( G \).
Directed Graphs and Evidence

Sally’s burglar **Alarm** is sounding. Has she been **Burgled**, or was the alarm triggered by an **Earthquake**? She turns the car **Radio** on for news of earthquakes.

Joint distribution specified by graph and eight parameters:

\[
p(B = 1) = 0.01 \\
p(A|B, E) \\
p(E = 1) = 0.000001 \\
p(R|E)
\]

<table>
<thead>
<tr>
<th>Alarm = 1</th>
<th>Burglar</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.99</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Radio = 1</th>
<th>Earthquake</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Explaining Away:** **B** and **E** are independent causes that become dependent given **A**

\[
p(B = 1|A = 1) = \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)} \approx 0.990002
\]

\[
p(B = 1 \mid A = 1, R = 1) \approx 0.0101
\]
Directed Graphs and Evidence

Incorrect Transformation: Create undirected graph by dropping edge orientation.

Falsely implies that given $A$, $B$ and $E$ are independent.

Explaining Away: $B$ and $E$ are independent causes that become dependent given $A$

$$p(B = 1 | A = 1) = \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)} \approx 0.990002$$

$$p(B = 1 | A = 1, R = 1) \approx 0.0101$$
**Directed Graphs and Evidence**

**Correct Transformation:** Create factor graph linking each node to its parent(s), or an undirected moral graph linking parents with common children.

![Directed Graphs](image)

**Explaining Away:** $B$ and $E$ are independent causes that become dependent given $A$

$$p(B = 1|A = 1) = \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)} \approx 0.990002$$

$$p(B = 1 \mid A = 1, R = 1) \approx 0.0101$$
Directed Factor Graphs?

Chain Graphs: Another hybrid of directed & undirected graphical models (see readings)

A rarely-used proposal by Brendan Frey, UAI 2003
Probabilistic Graphical Models

Directed Graphical Model (Bayesian Network)

Markov Properties (coming later)

Factorization

Factor Graph (Hypergraph)

A more detailed view of factorization, useful for algorithms

Undirected Graphical Model (Markov Random Field)

Markov Properties

Factorization
CS242: Lecture 1B Outline

- Factor graphs
- Loss Functions and Marginal Distributions
- Variable Elimination Algorithms
Inference with Two Variables

\[
p(x, y) = p(x)p(y \mid x)
\]

**Prediction via Marginalization:**

\[
p(y) = \sum_x p(x)p(y \mid x)
\]

**Inference via Table Lookup:**

\[
p(y \mid x = \bar{x})
\]

**Inference via Bayes Rule:**

\[
p(x \mid y = \bar{y}) = \frac{p(\bar{y} \mid x)p(x)}{p(\bar{y})}
\]
Marginal & Conditional Distributions

\[ p(x, y)^x = \sum_{z \in Z} p(x, y, z) \]

\[ p(x) = \sum_{y \in Y} p(x, y) \]

\[ p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)} \]

- **Target:** Subset of variables (nodes) whose values are of interest.
- **Condition:** Whatever other variables (nodes) we have observed.
- **Marginalize:** Everything that’s not a target or observation.

*This is a finite sum, but the number of terms may be exponentially large.*
Inference & Decision Theory

\( y \in \mathcal{Y} \rightarrow \) observed data, can take values in any space

\( x \in \mathcal{X} \rightarrow \) unknown or hidden variables (assume discrete for now)

\( A = \mathcal{X} \rightarrow \) action is to estimate the hidden variables

\( L(x, a) \rightarrow \) table giving loss for all possible mistakes

- The **posterior expected loss** of taking action \( a \) is

\[
\rho(a \mid y) = \mathbb{E}[L(x, a) \mid y] = \sum_{x \in \mathcal{X}} L(x, a) p(x \mid y)
\]

- The optimal **Bayes decision rule** is then

\[
\delta(y) = \arg \min_{a \in \mathcal{X}} \rho(a \mid y)
\]

- **Challenge:** For graphical models, the decision space is exponentially large
Decomposable Loss Functions

- Estimate hidden variables $x$ from observations $y$ via \textit{posterior expected loss}:
  \[ \rho(a \mid y) = \mathbb{E}[L(x,a) \mid y] = \sum_{x \in \mathcal{X}} L(x,a)p(x \mid y) \]

- Many practical losses decompose into a term for each variable node:
  \[ L(x,a) = \sum_{i=1}^{N} L_i(x_i,a_i) \quad \rho(a \mid y) = \sum_{i=1}^{N} \sum_{x_i} L(x_i,a_i)p(x_i \mid y) \]

- \textbf{Hamming Loss}: For how many nodes is our estimate incorrect?
  \[ L(x,a) = \sum_{i=1}^{N} \mathbb{I}(x_i \neq a_i) \]

- \textbf{Quadratic Loss}: Expected squared distance minimized by posterior mean
  \[ L(x,a) = \sum_{i=1}^{N} (x_i - a_i)^2 \]

- For any such loss, sufficient to find \textit{posterior marginals}:
  \[ p(x_i \mid y) \]
CS242: Lecture 1B Outline

- Factor graphs
- Loss Functions and Marginal Distributions
- Variable Elimination Algorithms
Distributive Law for Functions

- The classical distributive law of addition:

\[ a \cdot b + a \cdot c = a(b+c) \]

Applying the distributive law to functions of discrete variables:

\[ f(x) = \sum_y \sum_z g(x, y) h(x, z) = \left[ \sum_y g(x, y) \right] \cdot \left[ \sum_z h(x, z) \right] \]

Assuming \( y \) and \( z \) each take on \( N \) possible values, for any fixed \( x \):

- \( O(N^2) \) operations
- \( O(N) \) operations

If \( x \) also takes on \( N \) possible values, computing \( f(x) \) for all \( x \) has cost:

- \( O(N^3) \) operations
- \( O(N^2) \) operations
Example: Inference in a Directed Graph

- \( N=6 \) discrete variables, each taking one of \( K \) states

\[
p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(x_6 \mid x_2, x_5)
\]
Example: Inference in a Directed Graph

- $N=6$ discrete variables, each taking one of $K$ states
- **Goal:** Compute the *marginal* distribution of $x_1$, given *observations* of $x_4$ and $x_6$

$$p(x) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(x_6 \mid x_2, x_5)$$

$$p(x_1 \mid x_4 = \bar{x}_4, x_6 = \bar{x}_6) \propto \sum_{x_2} \sum_{x_3} \sum_{x_5} p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(\bar{x}_4 \mid x_2)p(x_5 \mid x_3)p(\bar{x}_6 \mid x_2, x_5)$$

$$p(x_1 \mid x_4 = \bar{x}_4, x_6 = \bar{x}_6) \propto \sum_{x_5} \sum_{x_3} \sum_{x_2} p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(\bar{x}_4 \mid x_2)p(x_5 \mid x_3)p(\bar{x}_6 \mid x_2, x_5)$$

Number of arithmetic operations to naively compute this sum: $\mathcal{O}(NK^4)$

*Exponential in number of unobserved variables!*
Conversion to Factor Graph Form

\begin{align*}
p(x) &= p(x_1)p(x_2 | x_1)p(x_3 | x_1)p(x_4 | x_2)p(x_5 | x_3)p(x_6 | x_2, x_5) \\
p(x) &= \phi_1(x_1)\phi_2(x_2, x_1)\phi_3(x_3, x_1)\phi_4(x_4, x_2)\phi_5(x_5, x_3)\phi_6(x_6, x_2, x_5) \\
p(x_1, x_2, x_3, x_5 | \bar{x}_4, \bar{x}_6) &\propto \phi_1(x_1)\phi_2(x_2, x_1)\phi_3(x_3, x_1)m_4(x_2)\phi_5(x_5, x_3)m_6(x_2, x_5)
\end{align*}

1. Convert directed graph to corresponding factor graph. 
   \textit{Algorithm is equally valid for factor graphs derived from undirected MRFs.}
2. Replace observed variables by lower-order \textit{evidence factors.}
Iterative Variable Elimination I

3. Pick some unobserved variable for elimination.
   Replace all factors which depend on that variable by a single, equivalent factor linking the eliminated variable’s neighbors. Repeat recursively.

\[ p(x_1, x_2, x_3, x_5 \mid \bar{x}_4, \bar{x}_6) \propto \phi_1(x_1)\phi_2(x_2, x_1)\phi_3(x_3, x_1)m_4(x_2)\phi_5(x_5, x_3)m_6(x_2, x_5) \]

\[ p(x_1, x_2, x_3 \mid \bar{x}_4, \bar{x}_6) \propto \sum_{x_5} \phi_1(x_1)\phi_2(x_2, x_1)\phi_3(x_3, x_1)m_4(x_2)\phi_5(x_5, x_3)m_6(x_2, x_5) \]

\[ p(x_1, x_2, x_3 \mid \bar{x}_4, \bar{x}_6) \propto \phi_1(x_1)\phi_2(x_2, x_1)\phi_3(x_3, x_1)m_4(x_2)m_5(x_2, x_3) \]

\[ m_5(x_2, x_3) = \sum_{x_5} \phi_5(x_5, x_3)m_6(x_2, x_5) \]

\[ \mathcal{O}(K^3) \] operations (matrix-matrix product)
3. Pick some unobserved variable for elimination. Replace all factors which depend on that variable by a single, equivalent factor linking the eliminated variable’s neighbors. Repeat recursively.
3. Pick some unobserved variable for elimination. Replace all factors which depend on that variable by a single, equivalent factor linking the eliminated variable’s neighbors. Repeat recursively.

\[
p(x_1, x_2 \mid \bar{x}_4, \bar{x}_6) \propto \phi_1(x_1)\phi_2(x_2, x_1)m_4(x_2)m_3(x_2, x_1)
\]

\[
p(x_1 \mid \bar{x}_4, \bar{x}_6) \propto \sum_{x_2} \phi_1(x_1)\phi_2(x_2, x_1)m_4(x_2)m_3(x_2, x_1)
\]

\[
p(x_1 \mid \bar{x}_4, \bar{x}_6) = Z_1^{-1}\phi_1(x_1)m_2(x_1), \quad Z_1 = \sum_{x_1} \phi_1(x_1)m_2(x_1)
\]
A General Graph Elimination Algorithm

Algebraic Marginalization Operations

- Marginalize out the variable associated with node being eliminated
- Compute a new potential function (factor) involving all other variables which are linked to the just-marginalized variable

Graph Manipulation Operations

- Remove, or eliminate, a single node from the graph
- Remove any factors linked to that node, and replace them by a single factor joining the neighbors of the just-removed node
- Alternative view based on undirected graphs: Add edges between all pairs of nodes neighboring the just-removed node, forming a clique

A Graph Elimination Algorithm

- Convert model to factor graph, encoding observed nodes as evidence factors
- Choose an elimination ordering (query node must be last)
- Iterate: Remove node, replace old factors by new message factor to neighbors
• **Elimination cliques**: Sets of neighbors of eliminated nodes
• **Marginalization cost**: Exponential in number of variables in each elimination clique (dominated by largest clique). Similar storage cost.
• **Treewidth of graph**: Over all possible elimination orderings, the smallest possible max-elimination-clique size, minus one
• **NP-Hard**: Finding the best elimination ordering for an arbitrary input graph (but heuristic approximation algorithms can be effective)
• **Limitation**: Lack of shared computation for finding multiple marginals
Treewidth and Elimination Order

$p(x_1) = ?$

$x_0$

$x_1$

$x_6$

$x_2$

$x_5$

$x_3$

$x_4$

*Treewidth = 1*

Where should $x_0$ be placed in the elimination order?
Treewidth and Elimination Order

$p(x_1) = ?$

$\text{Treewidth} = 1$

Does the elimination order matter for this graph?
Treewidth and Elimination Order

\[ p(x_1) = ? \]

\[
\begin{array}{ccc}
  x_1 & \rightarrow & x_2 \\
  \downarrow & & \downarrow \\
  x_4 & \rightarrow & x_5 \\
  \uparrow & & \uparrow \\
  x_3 & \rightarrow & x_6
\end{array}
\]

**Treewidth = 2**

What elimination order minimizes the maximal clique?
Efficient Elimination Orders may not Exist

For a square grid with $N$ total nodes, treewidth is provably $O(N^{0.5})$