**CS281A/Stat241A: Statistical Learning Theory**

# Lecture IV (9/9/04)

*Lecturer: Michael I. Jordan* *Scribes: Yao-Jung Wen and Nathan Kitchen*

## 1 Review of Last Time

### 1.1 Directed Graphical Models

The joint probability of the random variables associated with a directed graphical model are factorized as follows ($\pi_i$ is the set of parents of node $i$):

$$p(x) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})$$

### 1.2 Undirected Graphical Models

The joint probability of the random variables associated with an undirected graphical model are factorized as follows ($\mathcal{C}$ is the set of cliques associated with the undirected graph; $\psi_{C_i}$ is a potential function for the elements of clique $C_i$):

$$p(x) = \frac{1}{Z(\psi)} \prod_{C_i \in \mathcal{C}} \psi_{C_i}(x_{C_i})$$

which, if all potential functions $\psi_{C_i}$ are strictly positive,

$$= \exp\left\{ \sum_{C_i \in \mathcal{C}} \underbrace{\log \psi_{C_i}(x_{C_i})}_{\theta_i T_i(X)} \underbrace{- \log Z(\psi)}_{-A(\theta)} \right\}$$

In some cases, this can be expressed as

$$= \exp\left\{ \sum_i \theta_i T_i(X) - A(\theta) \right\},$$

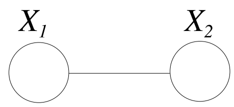
The distributions which can be expressed in this last form are known as the exponential family of distributions. Many common distributions belong to this family.

Common names:

| | |
|---|---|
| $Z$ | partition function |
| $\log Z$ | cumulant generation function |
| $-\log Z$ | free energy (especially in statistical physics) |

### 1.3 Maximal Cliques

Consider the following graph:




We could use the set of cliques $\mathcal{C} = \{\{X_1, X_2\}, \{X_1\}, \{X_2\}\}$, and define the joint probability

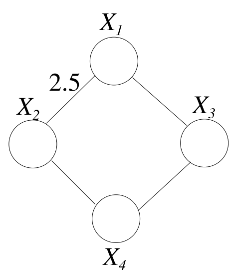
$$p(x_1, x_2) = \frac{1}{Z} \underbrace{\psi_{X_1 X_2}(x_1, x_2)\psi_{X_1}(x_1)\psi_{X_2}(x_2)}_{\psi'_{X_1 X_2}(x_1, x_2)}.$$

However, we could also use the set of maximal cliques $\mathcal{C} = \{\{X_1, X_2\}\}$, and represent the joint probability as

$$p(x_1, x_2) = \frac{1}{Z} \psi'_{X_1 X_2}(x_1, x_2).$$

Proofs are usually simpler when we parameterize with maximal cliques. However, some applications are more convenient with non-maximal cliques.

**Note**: $\psi(x_1, x_2)$ is in general unrelated to $p(x_1, x_2)$. For example:




| $X_1$ \ $X_2$ | 0 | 1 |
|---|---|---|
| 0 | 4.0 | 0.2 |
| 1 | 0.2 | 2.5 |

$\psi_{X_1 X_2}(x_1, x_2)$ is a local function defining the relationship between $X_1$ and $X_2$, which quantitatively shows that $X_2$ is likely to be 0 when $X_1$ is 0 and also likely to be 1 if $X_1$ is 1. $\psi_{X_1 X_2}(x_1, x_2)$ expressed a local relationship between $X_1$ and $X_2$. Local effects can be swamped out by global effects (such as stronger, opposing influences on $X_1$ and $X_2$). The global effects are captured by the marginal probability $p(x_1, x_2)$.

## 2 Types of Problems

We will look at three different kinds of problems.

### 2.1 Computing Marginal Probabilities ("Inference")

Given $p(x_V)$, compute $p(x_A)$, where $A \subset V$.
Example:
$p(x_V) = p(x_1, x_2, x_3, x_4, x_5, x_6)$
Compute $p(x_1)$.

### 2.2 Computing Conditional Probabilities

Given $p(x_V)$
Compute $p(x_A|x_B)$, where $A, B \subset V$.

$$p(x_A|x_B) = \frac{p(x_A, x_B)}{p(x_B)}$$

We can compute conditional probabilities using marginal probabilities.

### 2.3 Computing Maximal Probability Configuration

The goal is to find values for $x_V$ such that $p(x_V)$ is maximized. Given $p(x_V)$
Compute $\sup_{x_V} p(x_V)$, $\text{argmax}_{x_V} p(x_V)$

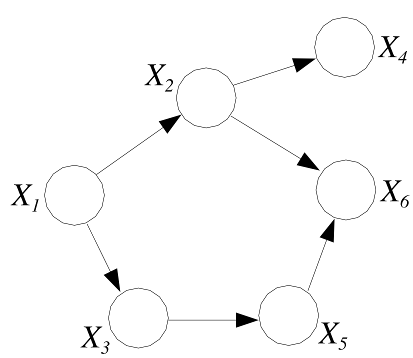
## 3 "Exact" Algorithms (Symbolic Algorithms)

In exact algorithms, the control flow is the same independent of the parameters (although dependent on the graph structure), and non-random, meaning they will produce the same results each time for identical input. Numeric algorithms are an alternative. We will talk about three exact algorithms:

**Elimination:** gives one marginal probability per run
**Sum-Product:** gives all marginal probabilities in one run; exact only for trees, not arbitrary graphs
**Junction-Tree:** gives all marginal probabilities; exact for arbitrary graphs, but may be computationally expensive. However, runtime can be determined beforehand for a particular graphical structure.

### 3.1 An Example

Consider the directed graphic model,




Suppose we want to compute the marginal $p(x_1, \ldots, x_5)$.

$$p(x_1) = \sum_{x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)$$

If we compute this with the summation at the front, the complexity is $O(k^6)$ for a single choice of values of $X_1, \ldots, X_5$.. On the other hand, if we move the summation as far as possible inward in the factorized joint probability, the complexity of this sum drops to $O(k^3)$:

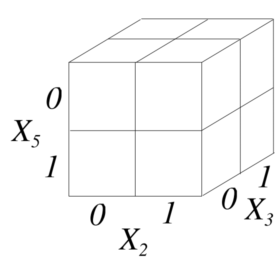
$$p(x_1) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)\sum_{x_6} p(x_6|x_2,x_5)$$

The Elimination algorithm distributes summations in this way. The intermediate results of the inner summations are labeled as $m_i(x_{S_i})$, where $S_i$ is the set of variables that appear in the summand of the summation over $X_i$. For example:

$$\begin{aligned}
p(x_1,x_2) = &\sum_{x_3\ldots x_6} p(x_1,x_2,x_3,x_4,x_5,x_6)\\
=&\sum_{x_3}\sum_{x_4}\sum_{x_5}\sum_{x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2,x_5)\\
=&\sum_{x_3}\sum_{x_4}\sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)\underbrace{\sum_{x_6} p(x_6|x_2,x_5)}_{m_6(x_2,x_5)}\\
=&\sum_{x_3}\sum_{x_4}\sum_{x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)m_6(x_2,x_5)\\
=&\sum_{x_3}\sum_{x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)\underbrace{\sum_{x_5} p(x_5|x_3)m_6(x_2,x_5)}_{m_5(x_2,x_3)}\\
=&\sum_{x_3}\sum_{x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)m_5(x_2,x_3)\\
=&\sum_{x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)m_5(x_2,x_3)\underbrace{\sum_{x_4} p(x_4|x_2)}_{m_4(x_2)}\\
=&\sum_{x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)m_5(x_2,x_3)m_4(x_2)\\
=&p(x_1)p(x_2|x_1)m_4(x_2)\underbrace{\sum_{x_3} p(x_3|x_1)m_5(x_2,x_3)}_{m_3(x_1,x_2)}
\end{aligned}$$

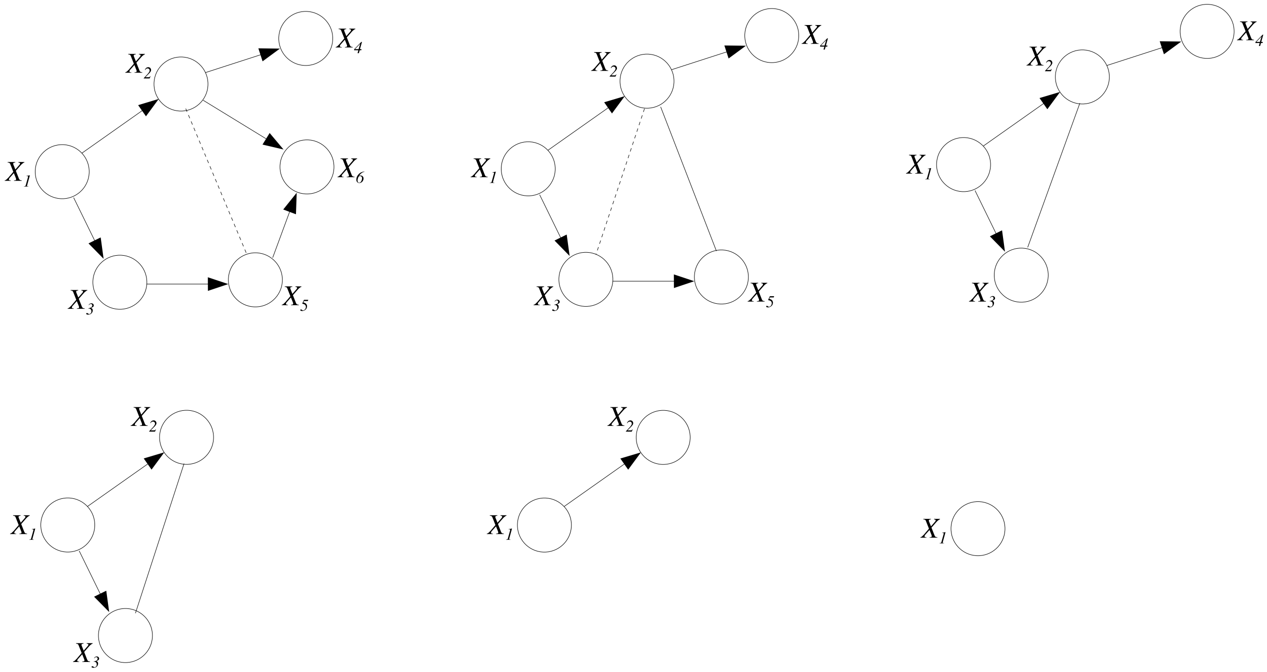
The $m_i(x_{S_i})$ are called *messages*. The complexity of Elimination is determined by the maximum message size.

For example, to produce $m_5$, we have to sum over a matrix that looks like (for binary variables):

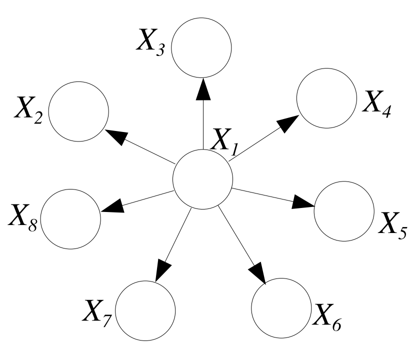



The Elimination algorithm is related to graph elimination. If we eliminate nodes in the same order as we sum over variables, the size of the largest elimination clique is the same as the maximum number of variables involved in a summation.

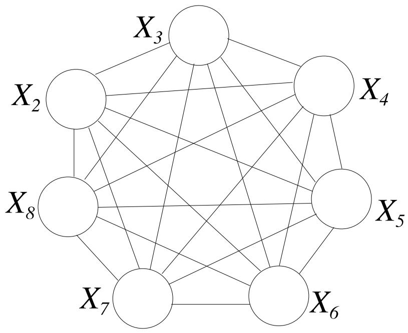
The graph elimination sequence corresponding to the above calculation is as following (note how parents of each eliminated node are connected with undirected edges at each step).

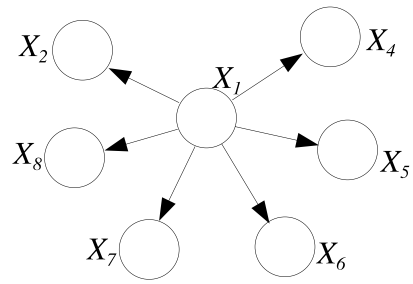



**Note**: Elimination order is critical!
Example: Look at the elimination of a star-shaped graph:




If $X_1$ is eliminated first, then all the other notes will have to be connected as shown in the left-hand side of the following figure. The right-hand side figure illustrates that if $X_3$ is eliminated first, then no additional connection is required.

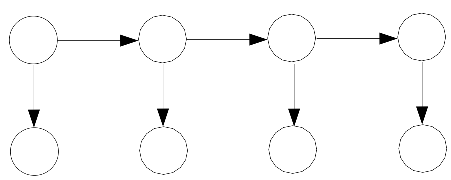


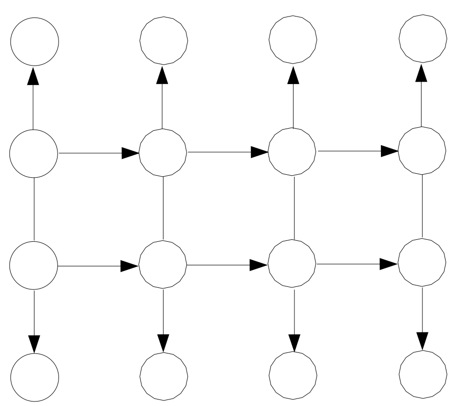

The best possible complexity of the elimination algorithm is described by the *treewidth*.

**Definition 1** ***Treewidth***
*= (Minimum of the maximal cliques created during graph elimination) - 1*

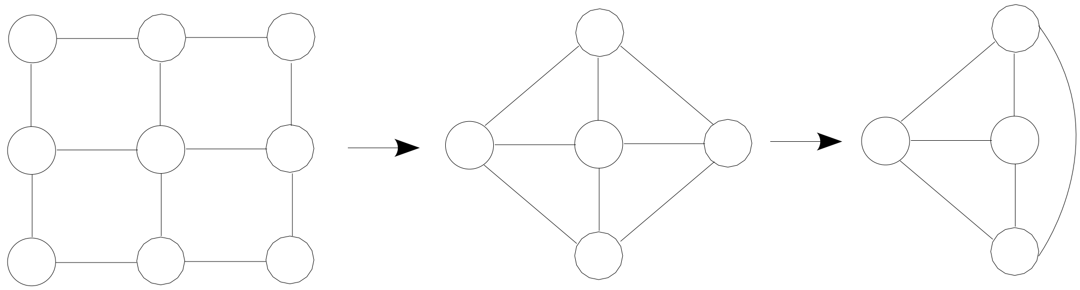
Examples:
Markov chain model (Treewidth = 1):



Sensor fusion model (Treewidth = 2):

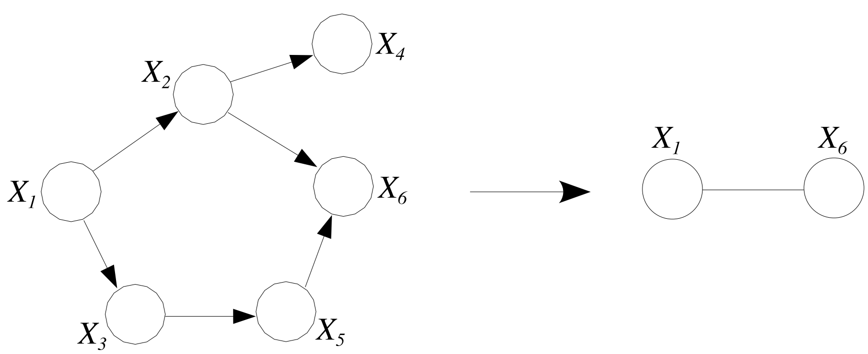


Grid graphical model (Treewidth= 3 for this grid, $N$ for $N \times N$ grid):



**Graph elimination pseudo code**:
ELIMINATE(G,F)
    INIT(G,F)
    UPDATE(G)
    NORMALIZE(F)
INIT(G,F)
    choose an ordering $I$ such that F appears last
    for each node $X_i$
        place $p(x_i|x_{\pi_i})$ on the "active list"
UPDATE(G)
    for each i in I
        find all factors on the list that refer to $X_i$ and remove from the list
        let $\phi_i(x_{T_i})$ denote the product of these factors
        let $m_i(x_{s_i}) = \sum_{x_i} \phi_i(x_{T_i})$
        put $m_i(x_{s_i})$ on the active list
NORMALIZE(F)
    $p(x_F) = \frac{\phi_F(x_F)}{\sum_{x_F} \phi_F(x_F)}$

Calculate $p(x_1, x_6)$ for the following graph using elimination:



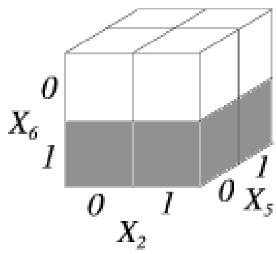

$$p(x_1, x_6) \propto m_2(x_1, x_6),$$

and so we can obtain the conditional probability $p(x_1|x_6)$,

$$p(x_1|x_6) = \frac{p(x_1, x_6)}{p(x_6)},$$

where $p(x_6)$ can easily be obtained by doing an additional elimination of $X_1$ to $p(x_1, x_6)$, i.e. summing over $x_1$ on $p(x_1, x_6)$.

$p(x_1|\overline{x}_6)$: Conditional probability conditioning on a specific value of $X_6$, i.e. $\overline{x}_6$.
Example: $p(\overline{x}_6|x_2, x_5)$, if $\overline{x}_6$ is $X_6 = 1$, then we only care about the probabilities at the lower part of the cube as shown.




**Definition 2** ***Evidence Potentials***

$$\delta(x_6, \overline{x}_6) = \begin{cases} 1, & \textit{if } x_6 = \overline{x}_6 \\ 0, & \textit{otherwise.} \end{cases}$$

$p(x_1, x_2, x_3, x_4, x_5, \overline{x}_6) = \sum_{x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)\delta(x_6, \overline{x}_6)$

$$\begin{aligned}
\sum_{x_1 \dots x_6} p(x_1, x_2, x_3, x_4, x_5, \overline{x}_6) &= \sum_{x_1 \dots x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3) \sum_{x_6} p(x_6|x_2, x_5)\delta(x_6, \overline{x}_6) \\
&= \sum_{x_1 \dots x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)m_6(x_2, x_5) \\
&\vdots \\
&= p(\overline{x}_6)
\end{aligned}$$

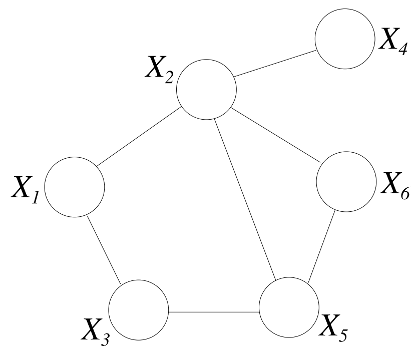
As another example:

$$\frac{\sum_{x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)\delta(x_6, \overline{x}_6)}{\sum_{x_1 \dots x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)\delta(x_6, \overline{x}_6)} = \frac{p(x_1, x_2, x_3, x_4, x_5, \overline{x}_6)}{p(\overline{x}_6)} = p(x_1, x_2, x_3, x_4, x_5|\overline{x}_6)$$

Note: $m_6(x_2, x_5)$ in above equation does not sum to 1. More generally, the results (as noted in the algorithm) must be normalized to reflect probabilities.

# 4 Marginalization for Undirected Graphs

Consider the undirected graphical model:




Elimination for undirected graphs works similarly to directed graphs:

$$
\begin{aligned}
p(x_1) &= \sum_{x_2 \dots x_6} \frac{1}{Z} \prod_{Ci \in \mathcal{C}} \psi_{C_i}(x_{C_i}) \\
&= \sum_{x_2 \dots x_5} \frac{1}{Z} \psi(x_1, x_2)\psi(x_2, x_4)\psi(x_1, x_3)\psi(x_3, x_5) \underbrace{\sum_{x_6} \psi(x_2, x_5, x_6)}_{m_6(x_2, x_5)} \\
&= \sum_{x_2 \dots x_5} \frac{1}{Z} \psi(x_1, x_2)\psi(x_2, x_4)\psi(x_1, x_3)\psi(x_3, x_5) m_6(x_2, x_5) \\
&= \sum_{x_2 \dots x_4} \frac{1}{Z} \psi(x_1, x_2)\psi(x_2, x_4)\psi(x_1, x_3) \underbrace{\sum_{x_5} \psi(x_3, x_5) m_6(x_2, x_5)}_{m_5(x_2, x_3)} \\
&= \sum_{x_2 \dots x_4} \frac{1}{Z} \psi(x_1, x_2)\psi(x_2, x_4)\psi(x_1, x_3) m_5(x_2, x_3) \\
&= \sum_{x_2, x_3} \frac{1}{Z} \psi(x_1, x_2)\psi(x_1, x_3) m_5(x_2, x_3) \underbrace{\sum_{x_4} \psi(x_2, x_4)}_{m_4(x_2)} \\
&= \sum_{x_2, x_3} \frac{1}{Z} \psi(x_1, x_2)\psi(x_1, x_3) m_5(x_2, x_3) m_4(x_2) \\
&= \sum_{x_2} \frac{1}{Z} \psi(x_1, x_2) m_4(x_2) \underbrace{\sum_{x_3} \psi(x_1, x_3) m_5(x_2, x_3)}_{m_3(x_1, x_2)} \\
&= \sum_{x_2} \frac{1}{Z} \psi(x_1, x_2) m_4(x_2) m_3(x_1, x_2) \\
&= \frac{1}{Z} \underbrace{\sum_{x_2} \psi(x_1, x_2) m_4(x_2) m_3(x_1, x_2)}_{m_2(x_1)} \\
&= \frac{1}{Z} m_2(x_1).
\end{aligned}
$$

But we don't want to compute $Z$ beforehand, because it requires a summation over all our variables. So we use:

$$p(x_1) \propto \sum_{x_2 \ldots x_6} \prod_{Ci \in \mathcal{C}} \psi_{C_i}(x_{C_i}) = m_2(x_1)$$

Finally, we can compute $Z$ from our result $m_2(x_1)$:

$$\sum_{x_1} m_2(x_1) = \sum_{x_1} \sum_{x_2, \ldots, x_6} \prod_{Ci \in \mathcal{C}} \psi_{C_i}(x_{C_i}) = Z$$

The elimination sequence corresponding to the above calculation is: