Inference with message passing

COMS 4721 Spring 2018
Computing (conditional) probabilities
**Problem:** Given $X_E = x_E$ (the “evidence”), what is the marginal conditional distribution of $x_f$?
**Problem:** Given $X_E = x_E$ (the “evidence”), what is the marginal conditional distribution of $x_f$?

**Applications:**

- Compute optimal predictions (e.g., conditional mean)
- General probabilistic reasoning
- Etc.
Recall shorthand notation

\[ X = (X_1, \ldots, X_n), \quad x = (x_1, \ldots, x_n) \]
Recall shorthand notation

- $X = (X_1, \ldots, X_n), \ x = (x_1, \ldots, x_n)$
- $p(x) = p(x_1, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n)$
Recall shorthand notation

- $\bm{X} = (X_1, \ldots, X_n)$, $\bm{x} = (x_1, \ldots, x_n)$
- $p(\bm{x}) = p(x_1, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n)$
- $p(x_v) = \Pr(X_v = x_v)$
Recall shorthand notation

- \( \mathbf{X} = (X_1, \ldots, X_n), \mathbf{x} = (x_1, \ldots, x_n) \)
- \( p(\mathbf{x}) = p(x_1, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n) \)
- \( p(x_v) = \Pr(X_v = x_v) \)
- For \( S \subseteq \mathcal{V} \),

\[
\mathbf{X}_S = (X_v : v \in S), \quad \mathbf{x}_S = (x_v : v \in S)
\]
Recall shorthand notation

- \( \mathbf{X} = (X_1, \ldots, X_n) \), \( \mathbf{x} = (x_1, \ldots, x_n) \)
- \( p(\mathbf{x}) = p(x_1, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n) \)
- \( p(x_v) = \Pr(X_v = x_v) \)
- For \( S \subseteq \mathcal{V} \),
  \[
  \mathbf{X}_S = (X_v : v \in S), \quad \mathbf{x}_S = (x_v : v \in S)
  \]
- For \( S, T \subseteq \mathcal{V} \),
  \[
  p(\mathbf{x}_S \mid \mathbf{x}_T) = \Pr(\mathbf{X}_S = \mathbf{x}_S \mid \mathbf{X}_T = \mathbf{x}_T)
  \]
Problem: compute $p(x_1 \mid \tilde{x}_6)$. (Assume domain of each $X_i$ is $\{0, 1\}$.)

I.e., compute the conditional distribution of $X_1$ given $X_6 = \tilde{x}_6$. 
Problem: compute \( p(x_1 \mid \tilde{x}_6) \). (Assume domain of each \( X_i \) is \( \{0, 1\} \).)

I.e., compute the conditional distribution of \( X_1 \) given \( X_6 = \tilde{x}_6 \).

By definition:

\[
p(x_1 \mid \tilde{x}_6) = \frac{p(x_1, \tilde{x}_6)}{p(\tilde{x}_6)}
\]

so separately compute numerator \( p(x_1, \tilde{x}_6) \) and denominator \( p(\tilde{x}_6) \).
Example (setup)

- Parameters of a distribution from the model:
  
  \[ p(x_1), p(x_2|x_1), p(x_3|x_1), p(x_4|x_2), p(x_5|x_3), p(x_6|x_2, x_5). \]
Example (setup)

- Parameters of a distribution from the model:

\[ p(x_1), p(x_2|x_1), p(x_3|x_1), p(x_4|x_2), p(x_5|x_3), p(x_6|x_2, x_5). \]

- E.g., \( p(x_6|x_2, x_5) \) is really a table of four values:

\[
\begin{bmatrix}
\Pr(X_6 = 1 \mid X_2 = 0, X_5 = 0) & \Pr(X_6 = 1 \mid X_2 = 0, X_5 = 1) \\
\Pr(X_6 = 1 \mid X_2 = 1, X_5 = 0) & \Pr(X_6 = 1 \mid X_2 = 1, X_5 = 1)
\end{bmatrix}.
\]

(We know \( \Pr(X_6 = 0 \mid X_2 = 0, X_5 = 0) = 1 - \Pr(X_6 = 1 \mid X_2 = 0, X_5 = 0) \).)
Example (setup)

- Parameters of a distribution from the model:
  \[ p(x_1), p(x_2|x_1), p(x_3|x_1), p(x_4|x_2), p(x_5|x_3), p(x_6|x_2, x_5). \]

- E.g., \( p(x_6|x_2, x_5) \) is really a **table of four values**:
  \[
  \begin{bmatrix}
  \Pr(X_6 = 1 \mid X_2 = 0, X_5 = 0) & \Pr(X_6 = 1 \mid X_2 = 0, X_5 = 1) \\
  \Pr(X_6 = 1 \mid X_2 = 1, X_5 = 0) & \Pr(X_6 = 1 \mid X_2 = 1, X_5 = 1)
  \end{bmatrix}.
  \]

  (We know \( \Pr(X_6 = 0 \mid X_2 = 0, X_5 = 0) = 1 - \Pr(X_6 = 1 \mid X_2 = 0, X_5 = 0) \).)

- We’ll compute \( p(x_1, \tilde{x}_6) \) and \( p(\tilde{x}_6) \) by **marginalizing out** variables.
Example (setup)

- Parameters of a distribution from the model:

\[ p(x_1), p(x_2|x_1), p(x_3|x_1), p(x_4|x_2), p(x_5|x_3), p(x_6|x_2, x_5). \]

- E.g., \( p(x_6|x_2, x_5) \) is really a **table of four values**:

\[
\begin{bmatrix}
\Pr(X_6 = 1 \mid X_2 = 0, X_5 = 0) & \Pr(X_6 = 1 \mid X_2 = 0, X_5 = 1) \\
\Pr(X_6 = 1 \mid X_2 = 1, X_5 = 0) & \Pr(X_6 = 1 \mid X_2 = 1, X_5 = 1)
\end{bmatrix}.
\]

(We know \( \Pr(X_6 = 0 \mid X_2 = 0, X_5 = 0) = 1 - \Pr(X_6 = 1 \mid X_2 = 0, X_5 = 0). \))

- We’ll compute \( p(x_1, \tilde{x}_6) \) and \( p(\tilde{x}_6) \) by **marginalizing out** variables.

- **Algebraic trick:**

\[
p(x_1, \tilde{x}_6) = \sum_{x_6} p(x_1, x_6) \cdot \delta_{\tilde{x}_6}(x_6)
\]

where \( \delta_{\tilde{x}_6}(x_6) \) is the **evidence potential** (a delta function)

\[
\delta_{\tilde{x}_6}(x_6) := \begin{cases} 
1 & \text{if } x_6 = \tilde{x}_6, \\
0 & \text{otherwise}.
\end{cases}
\]
Example (continued)

Numerator, in factorized form, including evidence potential:

\[ p(x_1, \bar{x}_6) = \sum_{x_2, x_3, x_4, x_5, x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)\delta_{\bar{x}_6}(x_6). \]
Numerator, in factorized form, including evidence potential:

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5, x_6} p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2) p(x_5|x_3) p(x_6|x_2, x_5) \delta_{\tilde{x}_6}(x_6). \]

First “marginalize out” \( x_6 \):

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5} p(x_1) p(x_2|x_1) p(x_3|x_1) p(x_4|x_2) p(x_5|x_3) \sum_{x_6} p(x_6|x_2, x_5) \delta_{\tilde{x}_6}(x_6). \]
Example (continued)

Numerator, in factorized form, including evidence potential:

\[ p(x_1, \bar{x}_6) = \sum_{x_2, x_3, x_4, x_5, x_6} p(x_1) p(x_2 | x_1) p(x_3 | x_1) p(x_4 | x_2) p(x_5 | x_3) p(x_6 | x_2, x_5) \delta_{\bar{x}_6}(x_6). \]

First “marginalize out” \( x_6 \):

\[ p(x_1, \bar{x}_6) = \sum_{x_2, x_3, x_4, x_5} p(x_1) p(x_2 | x_1) p(x_3 | x_1) p(x_4 | x_2) p(x_5 | x_3) \sum_{x_6} p(x_6 | x_2, x_5) \delta_{\bar{x}_6}(x_6). \]

Let \( m_6(x_2, x_5) := \sum_{x_6} p(x_6 | x_2, x_5) \delta_{\bar{x}_6}(x_6) \), a new function of \( x_2 \) and \( x_5 \).
Example (continued)

Numerator, in factorized form, including evidence potential:

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5, x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6). \]

First “marginalize out” \( x_6 \):

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3) \sum_{x_6} p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6). \]

Let \( m_6(x_2, x_5) := \sum_{x_6} p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6) \), a new function of \( x_2 \) and \( x_5 \).

Now “marginalize out” \( x_5 \):

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2) \sum_{x_5} p(x_5|x_3)m_6(x_2, x_5). \]
Example (continued)

Numerator, in factorized form, including evidence potential:

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5, x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6). \]

First “marginalize out” \( x_6 \):

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3) \sum_{x_6} p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6). \]

Let \( m_6(x_2, x_5) := \sum_{x_6} p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6) \), a new function of \( x_2 \) and \( x_5 \).

Now “marginalize out” \( x_5 \):

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2) \sum_{x_5} p(x_5|x_3)m_6(x_2, x_5). \]

Let \( m_5(x_2, x_3) := \sum_{x_5} p(x_5|x_3)m_6(x_2, x_5) \), a new function of \( x_2 \) and \( x_3 \).
Example (continued)

Numerator, in factorized form, including evidence potential:

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5, x_6} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6). \]

First "marginalize out" \( x_6 \):

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4, x_5} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3) \sum_{x_6} p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6). \]

Let \( m_6(x_2, x_5) := \sum_{x_6} p(x_6|x_2, x_5)\delta_{\tilde{x}_6}(x_6) \), a new function of \( x_2 \) and \( x_5 \).

Now "marginalize out" \( x_5 \):

\[ p(x_1, \tilde{x}_6) = \sum_{x_2, x_3, x_4} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2) \sum_{x_5} p(x_5|x_3)m_6(x_2, x_5). \]

Let \( m_5(x_2, x_3) := \sum_{x_5} p(x_5|x_3)m_6(x_2, x_5) \), a new function of \( x_2 \) and \( x_3 \).

\ldots
Example (conclusion)

Continue eliminating variables one-by-one (in order 6,5,4,3,2):

\[
p(x_1, \tilde{x}_6) = \sum_{x_2, x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)m_5(x_2, x_3) \sum_{x_4} p(x_4|x_2)
\]

= \sum_{x_2} p(x_1)p(x_2|x_1)m_4(x_2) \sum_{x_3} p(x_3|x_1)m_5(x_2, x_3)

= p(x_1) \sum_{x_2} p(x_2|x_1)m_3(x_1, x_2)m_4(x_2)

= p(x_1)m_2(x_1).

Eliminate \(x_1\) to get \(p(\tilde{x}_6)\):

\[
p(\tilde{x}_6) = \sum_{x_2} p(x_1)p(x_2|x_1)m_2(x_1).
\]
Continue eliminating variables one-by-one (in order 6,5,4,3,2):

\[
p(x_1, \tilde{x}_6) = \sum_{x_2, x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)m_5(x_2, x_3) \sum_{x_4} p(x_4|x_2)
\]

\[
= \sum_{x_2} p(x_1)p(x_2|x_1)m_4(x_2) \sum_{x_3} p(x_3|x_1)m_5(x_2, x_3)
\]

\[
= p(x_1) \sum_{x_2} p(x_2|x_1)m_3(x_1, x_2)m_4(x_2)
\]

\[
= p(x_1)m_2(x_1).
\]

Eliminate \( x_1 \) to get \( p(\tilde{x}_6) \):

\[
p(\tilde{x}_6) = \sum_{x_1} p(x_1)m_2(x_1).
\]
Continue eliminating variables one-by-one (in order 6,5,4,3,2):

\[
p(x_1, \tilde{x}_6) = \sum_{x_2, x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)m_5(x_2, x_3) \sum_{x_4} p(x_4|x_2)
\]

\[
= \sum_{x_2} p(x_1)p(x_2|x_1)m_4(x_2) \sum_{x_3} p(x_3|x_1)m_5(x_2, x_3)
\]

\[
= p(x_1) \sum_{x_2} p(x_2|x_1)m_3(x_1, x_2)m_4(x_2)
\]

\[
= p(x_1)m_2(x_1).
\]

Eliminate \(x_1\) to get \(p(\tilde{x}_6)\):

\[
p(\tilde{x}_6) = \sum_{x_1} p(x_1)m_2(x_1).
\]

Finally, compute \(p(x_1 | \tilde{x}_6) = \frac{p(x_1, \tilde{x}_6)}{p(\tilde{x}_6)}\).
While eliminating variables one-by-one, we keep track of \textit{messages}: $m_6(x_2, x_5)$, $m_5(x_2, x_3)$, etc.
Example (details)

While eliminating variables one-by-one, we keep track of messages: $m_6(x_2, x_5)$, $m_5(x_2, x_3)$, etc.

- Message $m_5(x_2, x_3)$ is really a table of four values:

$$
\begin{bmatrix}
    m_5(0, 0) & m_5(0, 1) \\
    m_5(1, 0) & m_5(1, 1)
\end{bmatrix}.
$$
While eliminating variables one-by-one, we keep track of messages: 
\( m_6(x_2, x_5), m_5(x_2, x_3), \) etc.

- Message \( m_5(x_2, x_3) \) is really a table of four values:

\[
\begin{bmatrix}
    m_5(0, 0) & m_5(0, 1) \\
    m_5(1, 0) & m_5(1, 1)
\end{bmatrix}.
\]

- Need to compute all four values of \( m_5(x_2, x_3) \).
Example (details)

While eliminating variables one-by-one, we keep track of messages: $m_6(x_2, x_5), m_5(x_2, x_3)$, etc.

- Message $m_5(x_2, x_3)$ is really a table of four values:

\[
\begin{bmatrix}
m_5(0, 0) & m_5(0, 1) \\
m_5(1, 0) & m_5(1, 1)
\end{bmatrix}.
\]

- Need to compute all four values of $m_5(x_2, x_3)$.
- Recall:

\[
m_5(x_2, x_3) := \sum_{x_5} p(x_5|x_3)m_6(x_2, x_5).
\]
While eliminating variables one-by-one, we keep track of messages: \( m_6(x_2, x_5), m_5(x_2, x_3) \), etc.

- **Message** \( m_5(x_2, x_3) \) is really a table of four values:

  \[
  \begin{bmatrix}
  m_5(0, 0) & m_5(0, 1) \\
  m_5(1, 0) & m_5(1, 1)
  \end{bmatrix}.
  \]

- **Need to compute all four values of** \( m_5(x_2, x_3) \).

- **Recall**:

  \[
  m_5(x_2, x_3) := \sum_{x_5} p(x_5 | x_3) m_6(x_2, x_5).
  \]

- **Compute each value by looking up in tables for** \( p(x_5 | x_3) \) and \( m_6(x_2, x_5) \) **and summing over possible values of** \( x_5 \).
This inference method is called *variable elimination*. 

Message of size $2^4 = 16$. 

How to figure out best ordering of variables (to minimize "size" of largest message)? Computationally intractable in general.
This inference method is called *variable elimination*.

Complexity of variable elimination comes from “size” of messages.
This inference method is called \textit{variable elimination}.

Complexity of variable elimination comes from “size” of messages.

Eliminating variables in different order may require larger messages!
This inference method is called **variable elimination**.

- Complexity of variable elimination comes from “size” of messages.
- Eliminating variables in different order may require larger messages!
  - E.g., if we tried to first eliminate $x_2$:

\[
p(x_1, \tilde{x}_6) = \sum_{x_3, x_4, x_5, x_6} p(x_1)p(x_3|x_1)p(x_5|x_3)\delta_{\tilde{x}_6}(x_6) \sum_{x_2} p(x_2|x_1)p(x_4|x_2)p(x_6|x_2, x_5). \]

\[
= m_2(x_1, x_4, x_5, x_6)
\]

Message of size $2^4 = 16$. 
This inference method is called **variable elimination**.

Complexity of variable elimination comes from “size” of messages.

Eliminating variables in different order may require larger messages!

E.g., if we tried to first eliminate $x_2$:

$$p(x_1, \bar{x}_6) = \sum_{x_3, x_4, x_5, x_6} p(x_1)p(x_3|x_1)p(x_5|x_3)\delta_{\bar{x}_6}(x_6) \sum_{x_2} p(x_2|x_1)p(x_4|x_2)p(x_6|x_2, x_5).$$

Message of size $2^4 = 16$.

How to figure out best ordering of variables (to minimize “size” of largest message)?
This inference method is called variable elimination.

Complexity of variable elimination comes from “size” of messages.

Eliminating variables in different order may require larger messages!

E.g., if we tried to first eliminate $x_2$:

\[
p(x_1, \tilde{x}_6) = \sum_{x_3, x_4, x_5, x_6} p(x_1)p(x_3|x_1)p(x_5|x_3)\delta_{\tilde{x}_6}(x_6)\sum_{x_2} p(x_2|x_1)p(x_4|x_2)p(x_6|x_2, x_5) .
\]

Message of size $2^4 = 16$.

How to figure out best ordering of variables (to minimize “size” of largest message)?

Computationally intractable in general.
Message passing on directed trees
Directed tree models

For rest of lecture, we restrict attention to directed trees, i.e., DAGs with a single source vertex and all other vertices have a single parent.

(Will also work for directed forests like IID model.)

\[
p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2) \cdot p(x_4|x_2)p(x_5|x_1)p(x_6|x_5)
\]
Directed tree models

For rest of lecture, we restrict attention to directed trees, i.e., DAGs with a single source vertex and all other vertices have a single parent.

(Will also work for directed forests like IID model.)

\[
p(\mathbf{x}) = p(x_1)p(x_2 | x_1)p(x_3 | x_2) \\
p(x_4 | x_2)p(x_5 | x_1)p(x_6 | x_5)
\]

In directed trees, d-separation is determined by undirected connectivity.

I.e., \( X_A \perp \!\!\!\!\perp X_B | X_E \) if, upon removing vertices in \( E \), there is no undirected path in \( G \) between \( A \) and \( B \).
Example: caterpillar tree

Caterpillar tree (also known as Hidden Markov Model).
Example: caterpillar tree

Caterpillar tree (also known as Hidden Markov Model).
Define *vertex potentials* and *edge potentials*:

- $\phi(x_r) = p(x_r)$ for root vertex $r$.
- $\phi(x_v) = 1$ for all non-root vertices $v$.
- $\phi(x_u, x_v) = p(x_v | x_u)$ for edge $(u, v) \in E$.

Remember edge orientation only to define

\[ \phi(x_u, x_v) = \phi(x_v, x_u) = p(x_v | x_u). \]

(Notation rules for $\phi$ are analogous to those for $p$.)
Potential function representation

Define *vertex potentials* and *edge potentials*:

- $\phi(x_r) = p(x_r)$ for root vertex $r$.
- $\phi(x_v) = 1$ for all non-root vertices $v$.
- $\phi(x_u, x_v) = p(x_v | x_u)$ for edge $(u, v) \in \mathcal{E}$.

Remember edge orientation only to define $\phi(x_u, x_v) = \phi(x_v, x_u) = p(x_v | x_u)$.

(Notation rules for $\phi$ are analogous to those for $p$.)

Then

$$
\Pr(X = \mathbf{x}) = \prod_{v \in V} \phi(x_v) \prod_{(u,v) \in \mathcal{E}} \phi(x_u, x_v).
$$

(Product of vertex and edge potentials.)
Define *vertex potentials* and *edge potentials*:

- \( \phi(x_r) = p(x_r) \) for root vertex \( r \).
- \( \phi(x_v) = 1 \) for all non-root vertices \( v \).
- \( \phi(x_u, x_v) = p(x_v \mid x_u) \) for edge \((u, v) \in E\).

Remember edge orientation only to define
\[
\phi(x_u, x_v) = \phi(x_v, x_u) = p(x_v \mid x_u).
\]

(Notation rules for \( \phi \) are analogous to those for \( p \).)

Then
\[
\Pr(X = x) = \prod_{v \in V} \phi(x_v) \prod_{(u,v) \in E} \phi(x_u, x_v).
\]

(Product of vertex and edge potentials.)

Now root and non-root vertices are treated the same way!
Define *vertex potentials* and *edge potentials*:

- $\phi(x_r) = p(x_r)$ for root vertex $r$.
- $\phi(x_v) = 1$ for all non-root vertices $v$.
- $\phi(x_u, x_v) = p(x_v | x_u)$ for edge $(u, v) \in \mathcal{E}$.

Remember edge orientation only to define

$x_u, x_v) = \phi(x_v, x_u) = p(x_v | x_u)$.

(Notation rules for $\phi$ are analogous to those for $p$.)

Then

$$\Pr(X = x) = \prod_{v \in V} \phi(x_v) \prod_{(u,v) \in \mathcal{E}} \phi(x_u, x_v).$$

(Product of vertex and edge potentials.)

Now root and non-root vertices are treated the same way!

Can forget about edge orientations after defining potentials.
Suppose we want to condition on $X_E = \tilde{x}_E = (\tilde{x}_v : v \in E)$.

Define new vertex potentials that incorporate “evidence”:

1. $\phi^E(x_v) = \phi(x_v) \cdot \delta_{\tilde{x}_v}(x_v)$ for $v \in E$
2. $\phi^E(x_v) = \phi(x_v)$ for $v \notin E$

Then $\Pr(X = x | X_E = \tilde{x}_E) = 1 \frac{\prod_v \phi^E(x_v) \prod_{(u,v) \in E} \phi(x_u,x_v)}{Z_E}$

where $Z_E$ is the normalizing factor $Z_E := \sum_x \frac{\prod_v \phi^E(x_v) \prod_{(u,v) \in E} \phi(x_u,x_v)}{p(\tilde{x}_E)}$. 

Root and non-root vertices still treated the same way!
Incorporating evidence

Suppose we want to condition on \( X_E = \tilde{x}_E = (\tilde{x}_v : v \in E) \).

Define new vertex potentials that incorporate “evidence”:

\[
\begin{align*}
\phi^E(x_v) &= \phi(x_v) \cdot \delta_{\tilde{x}_v}(x_v) \text{ for } v \in E \\
\phi^E(x_v) &= \phi(x_v) \text{ for } v \notin E
\end{align*}
\]

Then

\[
\Pr(X = x \mid X_E = \tilde{x}_E) = \frac{1}{Z^E} \prod_{v \in V} \phi^E(x_v) \prod_{(u,v) \in E} \phi(x_u, x_v)
\]

where \( Z^E \) is the normalizing factor

\[
Z^E := \sum_x \left( \prod_{v \in V} \phi^E(x_v) \prod_{(u,v) \in E} \phi(x_u, x_v) \right) = p(\tilde{x}_E).
\]
Incorporating evidence

Suppose we want to condition on $X_E = \tilde{x}_E = (\tilde{x}_v : v \in E)$.

Define new vertex potentials that incorporate “evidence”:

- $\phi^E(x_v) = \phi(x_v) \cdot \delta_{\tilde{x}_v}(x_v)$ for $v \in E$
- $\phi^E(x_v) = \phi(x_v)$ for $v \notin E$

Then

$$
\Pr(X = x \mid X_E = \tilde{x}_E) = \frac{1}{Z^E} \prod_{v \in V} \phi^E(x_v) \prod_{(u,v) \in E} \phi(x_u, x_v)
$$

where $Z^E$ is the normalizing factor

$$
Z^E := \sum_x \left( \prod_{v \in V} \phi^E(x_v) \prod_{(u,v) \in E} \phi(x_u, x_v) \right) = p(\tilde{x}_E).
$$

Root and non-root vertices still treated the same way!
Suppose you want to compute $p(x_r \mid \tilde{x}_E)$. 

Claim: This becomes very nice on trees.
Computing marginal conditional probabilities

Suppose you want to compute $p(x_r | \tilde{x}_E)$.

1. Marginalize out all variables except $x_r$, to get $p(x_i, \tilde{x}_E)$:

$$p(x_r, \tilde{x}_E) = \sum_{x_i: i \neq r} \left( \prod_{v \in V} \phi^E(x_v) \prod_{(u,v) \in E} \phi(x_u, x_v) \right).$$

2. Now also marginalize out $x_r$ to get normalizer:

$$Z^E = p(\tilde{x}_E) = \sum_{x_r} p(x_r, \tilde{x}_E).$$

3. Divide these quantities to get $p(x_r | \tilde{x}_E)$:

$$p(x_r | \tilde{x}_E) = \frac{p(x_r, \tilde{x}_E)}{p(\tilde{x}_E)}.$$
Computing marginal conditional probabilities

Suppose you want to compute \( p(x_r \mid \tilde{x}_E) \).

1. Marginalize out all variables except \( x_r \), to get \( p(x_i, \tilde{x}_E) \):

\[
p(x_r, \tilde{x}_E) = \sum_{x_i : i \neq r} \left( \prod_{v \in \mathcal{V}} \phi^E(x_v) \prod_{(u,v) \in \mathcal{E}} \phi(x_u, x_v) \right).
\]

2. Now also marginalize out \( x_r \) to get normalizer:

\[
Z^E = p(\tilde{x}_E) = \sum_{x_r} p(x_r, \tilde{x}_E).
\]

3. Divide these quantities to get \( p(x_r \mid \tilde{x}_E) \):

\[
p(x_r \mid \tilde{x}_E) = \frac{p(x_r, \tilde{x}_E)}{p(\tilde{x}_E)}.
\]

**Claim**: This becomes very nice on trees.
Messages

Pick any vertex \( r \) to be the root, and consider eliminating variables by always picking a leaf to eliminate next, until only \( r \) remains.
Pick any vertex \( r \) to be the root, and consider eliminating variables by always picking a leaf to eliminate next, until only \( r \) remains.

- Suppose we’re about to eliminate \( x_j \), and parent of \( x_j \) is \( x_i \).
Pick any vertex $r$ to be the root, and consider eliminating variables by always picking a leaf to eliminate next, until only $r$ remains.

- Suppose we’re about to eliminate $x_j$, and parent of $x_j$ is $x_i$.

- (Already eliminated vertices in subtree rooted at $x_j$.)
Pick any vertex $r$ to be the root, and consider eliminating variables by always picking a leaf to eliminate next, until only $r$ remains.

- Suppose we’re about to eliminate $x_j$, and parent of $x_j$ is $x_i$.
- (Already eliminated vertices in subtree rooted at $x_j$.)
- Which remaining potentials involve $x_j$?
Messages

Pick any vertex \( r \) to be the root, and consider eliminating variables by always picking a leaf to eliminate next, until only \( r \) remains.

- Suppose we’re about to eliminate \( x_j \), and parent of \( x_j \) is \( x_i \).
- (Already eliminated vertices in subtree rooted at \( x_j \).)
- Which remaining potentials involve \( x_j \)?
- Vertex potential \( \phi^E(x_j) \), of course, and edge potential \( \phi(x_i, x_j) \).
Pick any vertex $r$ to be the root, and consider eliminating variables by always picking a leaf to eliminate next, until only $r$ remains.

- Suppose we’re about to eliminate $x_j$, and parent of $x_j$ is $x_i$.

- (Already eliminated vertices in subtree rooted at $x_j$.)

- Which remaining potentials involve $x_j$?

- Vertex potential $\phi^E(x_j)$, of course, and edge potential $\phi(x_i, x_j)$.

- No other edge potentials involve $x_j$, because $\mathcal{G}$ is a tree!
Messages

Pick any vertex $r$ to be the root, and consider eliminating variables by always picking a leaf to eliminate next, until only $r$ remains.

- Suppose we’re about to eliminate $x_j$, and parent of $x_j$ is $x_i$.

- (Already eliminated vertices in subtree rooted at $x_j$.)

- Which remaining potentials involve $x_j$?

- Vertex potential $\phi^E(x_j)$, of course, and edge potential $\phi(x_i, x_j)$.

- No other edge potentials involve $x_j$, because $G$ is a tree!

- Message created upon eliminating $x_j$ is only function of $x_i$: $m_{j\rightarrow i}(x_i)$. 
Messages (recursively)

Recursively, can see that

\[ m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right). \]

where \( \text{neigh}(j) \) are the neighbors of vertex \( j \), so \( \text{neigh}(j) \setminus \{i\} \) are neighbors of \( j \) other than \( i \).
Recursive, can see that

\[
m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right).
\]

where \(\text{neigh}(j)\) are the neighbors of vertex \(j\), so \(\text{neigh}(j) \setminus \{i\}\) are neighbors of \(j\) other than \(i\).

Back to marginal conditional probability:

\[
p(x_r | \tilde{x}_E) = \frac{1}{Z^E} \cdot \phi^E(x_r) \prod_{s \in \text{neigh}(r)} m_{s \rightarrow r}(x_r).
\]
Messages (recursively)

Recursively, can see that

\[ m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right) . \]

where \( \text{neigh}(j) \) are the neighbors of vertex \( j \), so \( \text{neigh}(j) \setminus \{i\} \) are neighbors of \( j \) other than \( i \).

Back to marginal conditional probability:

\[ p(x_r \mid \tilde{x}_E) = \frac{1}{Z^E} \cdot \phi^E(x_r) \prod_{s \in \text{neigh}(r)} m_{s \rightarrow r}(x_r) . \]

But it didn’t matter which vertex was designated the root!
Exchanging messages

Can define messages \( m_{j \rightarrow i}(x_i) \) without caring about how tree is rooted:

\[
m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j)\phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right).
\]
Exchanging messages

Can define messages $m_{j\rightarrow i}(x_i)$ without caring about how tree is rooted:

$$m_{j\rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k\rightarrow j}(x_j) \right).$$
Exchanging messages

Can define messages $m_{j \rightarrow i}(x_i)$ without caring about how tree is rooted:

$$m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right).$$

Even if edge $(i, j) \in E$ is oriented $i \rightarrow j$, can also define message $m_{j \rightarrow i}(x_i)$.
Exchanging messages

Can define messages $m_{j \rightarrow i}(x_i)$ without caring about how tree is rooted:

$$m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^{E}(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right).$$

Even if edge $(i, j) \in \mathcal{E}$ is oriented $i \rightarrow j$, can also define message $m_{j \rightarrow i}(x_i)$.

When can we compute $m_{j \rightarrow i}(x_i)$?
Can define messages \( m_{j \rightarrow i}(x_i) \) without caring about how tree is rooted:

\[
m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right).
\]

Even if edge \((i, j) \in \mathcal{E}\) is oriented \(i \rightarrow j\), can also define message \( m_{j \rightarrow i}(x_i) \).

When can we compute \( m_{j \rightarrow i}(x_i) \)?

**Answer:** when \( m_{k \rightarrow j}(x_j) \) have been computed for all \( k \in \text{neigh}(j) \setminus \{i\} \).
Can (re-)use **same messages** to compute $p(x_v | \tilde{x}_E)$ for all $v \in \mathcal{V}$:

$$p(x_v | \tilde{x}_E) = \frac{1}{Z_E} \cdot \phi^E(x_v) \prod_{u \in \text{neigh}(v)} m_{u \rightarrow v}(x_v),$$

where

$$m_{j \rightarrow i}(x_i) = \sum_{x_j} \left( \phi^E(x_j) \phi(x_i, x_j) \prod_{k \in \text{neigh}(j) \setminus \{i\}} m_{k \rightarrow j}(x_j) \right).$$
Application: phylogenetic tree

$X_v$ is nucleotide (A,C,G,T) at particular location in DNA for species $v$. 
Application: phylogenetic tree

$X_v$ is nucleotide (A,C,G,T) at particular location in DNA for species $v$.

What is $p(x_r \mid x_{\text{extant species}})$?
Application: phylogenetic tree

$X_v$ is nucleotide (A,C,G,T) at particular location in DNA for species $v$.

What is $p(x_r \mid x_{\text{extant species}})$?

What is $p(x_{\text{t-rex}} \mid x_{\text{extant species}})$?
Application: phylogenetic tree

$X_v$ is nucleotide (A,C,G,T) at particular location in DNA for species $v$.

What is $p(x_r \mid \text{extant species})$?

What is $p(x_{t-rex} \mid \text{extant species})$?

Etc.
Key takeaways

Inference for directed trees is **easy**.

1. Potential function view lets us treat all vertices the same.
2. Incorporating evidence only modifies vertex potentials.
3. Message passing (a.k.a. *sum-product* algorithm) only creates “small” messages $m_{j\rightarrow i}(x_i)$.
4. Can reuse messages to compute $p(x_v \mid \tilde{x}_E)$ for all $v \in \mathcal{E}$. 
Hidden Markov models
Caterpillar tree (also known as Hidden Markov Model).
Caterpillar tree

Caterpillar tree (also known as Hidden Markov Model).

Example from biology:

- \( X_t \) = nucleotide at position \( t \) in DNA.
- \( H_t = 1 \{ \text{position } t \text{ is in coding region} \} \).
Example from biology:

- $X_t = \text{nucleotide at position } t \text{ in DNA.}$
- $H_t = 1\{\text{position } t \text{ is in coding region}\}.$

Typical inference question:

What is the probability that $H_{300} = 1$ (300th position is in coding region) given $X = \tilde{x}$ (human DNA sequence)? $p(h_{300} | \tilde{x})$
Generative story

- Generate **hidden state sequence** $H_1, \ldots, H_T$.
- Given $H_1, \ldots, H_T$, generate **observation sequence** $X_1, \ldots, X_T$. 
Generative story

- Generate hidden state sequence $H_1, \ldots, H_T$.
- Given $H_1, \ldots, H_T$, generate observation sequence $X_1, \ldots, X_T$.

Alternatively:

- Generate $H_1$.
- Then generate $X_1$.
- Then generate $H_2$.
- Then generate $X_2$.
- Etc.
Parameters

\[ p(x_1, \ldots, x_T, h_1, \ldots, h_T) = p(h_1) \cdot \prod_{t=2}^{T} p(h_t \mid h_{t-1}) \cdot \prod_{t=1}^{T} p(x_t \mid h_t). \]

- **initial state distribution** \( p(h_1) \)
- **transition probabilities** \( p(h_t \mid h_{t-1}) \)
- **observation probabilities** \( p(x_t \mid h_t) \)
Message passing

Typical inference problem with HMMs:

Fix a particular distribution $p$ from the HMM.

Conditioned on $(X_1, \ldots, X_T) = (\tilde{x}_1, \ldots, \tilde{x}_T)$, what is $p(h_t \mid \tilde{x}_1, \ldots, \tilde{x}_T)$?
Typical inference problem with HMMs:

Fix a particular distribution $p$ from the HMM.

Conditioned on $(X_1, \ldots, X_T) = (\tilde{x}_1, \ldots, \tilde{x}_T)$, what is $p(h_t | \tilde{x}_1, \ldots, \tilde{x}_T)$?

- Predict the future: $p(h_{T+1} | \tilde{x}_1, \ldots, \tilde{x}_T)$ and $p(x_{T+1} | \tilde{x}_1, \ldots, \tilde{x}_T)$
- Predict the past: $p(h_t | \tilde{x}_1, \ldots, \tilde{x}_T)$ for $t \leq T$. 

Since HMM is a directed tree, can use message passing to efficiently compute all conditional probability queries. Also called forward-backward algorithm in context of HMMs.
Message passing

Typical inference problem with HMMs:

Fix a particular distribution $p$ from the HMM.

Conditioned on $(X_1, \ldots, X_T) = (\tilde{x}_1, \ldots, \tilde{x}_T)$, what is $p(h_t \mid \tilde{x}_1, \ldots, \tilde{x}_T)$?

- Predict the future: $p(h_{T+1} \mid \tilde{x}_1, \ldots, \tilde{x}_T)$ and $p(x_{T+1} \mid \tilde{x}_1, \ldots, \tilde{x}_T)$
- Predict the past: $p(h_t \mid \tilde{x}_1, \ldots, \tilde{x}_T)$ for $t \leq T$.

Since HMM is a directed tree, can use message passing to efficiently compute all conditional probability queries.

Also called *forward-backward algorithm* in context of HMMs.
Parameters (again)

Typically assume \( p(x_t \mid h_t) \) and \( p(h_t \mid h_{t-1}) \) are time-homogeneous: i.e., do not depend on \( t \).
Parameters (again)

Typically assume \( p(x_t | h_t) \) and \( p(h_t | h_{t-1}) \) are time-homogeneous: i.e., do not depend on \( t \).

**Example:**

- \( X_t \) are \{1, \ldots, d\}-valued random variables.
- \( H_t \) are \{1, \ldots, k\}-valued random variables.
Parameters (again)

Typically assume $p(x_t \mid h_t)$ and $p(h_t \mid h_{t-1})$ are time-homogeneous: i.e., do not depend on $t$.

**Example:**

- $X_t$ are $\{1, \ldots, d\}$-valued random variables.
- $H_t$ are $\{1, \ldots, k\}$-valued random variables.

**Time-homogeneous parameters:**

- $p(h_t \mid h_{t-1})$: $k \times (k - 1)$ parameters.
  - $(A_{i,j})$ with $\sum_{j=1}^{k} A_{i,j} = 1$ for all $i$.
- $p(x_t \mid h_t)$: $k \times (d - 1)$ parameters.
  - $(B_{i,j})$ with $\sum_{j=1}^{d} B_{i,j} = 1$ for all $i$. 
Parameters (again)

Typically assume $p(x_t \mid h_t)$ and $p(h_t \mid h_{t-1})$ are \textit{time-homogeneous}: i.e., do not depend on $t$.

\textbf{Example}:

\begin{itemize}
  \item $X_t$ are $\{1, \ldots, d\}$-valued random variables.
  \item $H_t$ are $\{1, \ldots, k\}$-valued random variables.
\end{itemize}

\textbf{Time-homogeneous parameters}:

\begin{itemize}
  \item $p(h_t \mid h_{t-1})$: $k \times (k - 1)$ parameters.
    \begin{itemize}
      \item $(A_{i,j})$ with $\sum_{j=1}^{k} A_{i,j} = 1$ for all $i$.
    \end{itemize}
  \item $p(x_t \mid h_t)$: $k \times (d - 1)$ parameters.
    \begin{itemize}
      \item $(B_{i,j})$ with $\sum_{j=1}^{d} B_{i,j} = 1$ for all $i$.
    \end{itemize}
\end{itemize}

What about $p(h_1)$?
Parameter estimation

Given data $\tilde{x}_1, \ldots, \tilde{x}_T, \tilde{h}_1, \ldots, \tilde{h}_T$, how to estimate $A$ and $B$?
Parameter estimation

Given data \( \tilde{x}_1, \ldots, \tilde{x}_T, \tilde{h}_1, \ldots, \tilde{h}_T \), how to estimate \( A \) and \( B \)?

**Maximum likelihood estimation:**

\[
\hat{A}_{i,j} = \frac{1}{T-1} \sum_{t=1}^{T-1} 1\{\tilde{h}_t = i, \tilde{h}_{t+1} = j\},
\]

\[
\hat{B}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} 1\{\tilde{h}_t = i, \tilde{x}_t = j\}.
\]
Parameter estimation

Given data $\tilde{x}_1, \ldots, \tilde{x}_T, \tilde{h}_1, \ldots, \tilde{h}_T$, how to estimate $A$ and $B$?

**Maximum likelihood estimation:**

\[
\hat{A}_{i,j} = \frac{1}{T-1} \sum_{t=1}^{T-1} 1\{\tilde{h}_t = i, \tilde{h}_{t+1} = j\},
\]

\[
\hat{B}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} 1\{\tilde{h}_t = i, \tilde{x}_t = j\}.
\]

If only given $\tilde{x}_1, \ldots, \tilde{x}_T$, estimation problem becomes much more difficult! (We’ll discuss this another time.)
1. Efficient inference for HMMs via message passing (forward and backward messages).
2. If both “hidden state sequence” and “observation sequence” are observed, can estimate time-homogeneous parameters via MLE.