Bayesian Networks: Belief Propagation (Cont'd)

Huizhen Yu

janey.yu@cs.helsinki.fi
Dept. Computer Science, Univ. of Helsinki

Probabilistic Models, Spring, 2010

Outline

Belief Propagation

- Review and Examples
- Generalized Belief Propagation – Max-Product
- Applications to Loopy Graphs

Announcement: The last exercise will be posted online soon.

Outline

Belief Propagation

- Review and Examples
- Generalized Belief Propagation – Max-Product
- Applications to Loopy Graphs

Review of Last Lecture

We studied an algorithm for computing marginal posterior distributions:

- It works in singly connected networks, which are DAGs whose undirected versions are trees.
- It is suitable for parallel implementation.
- It is recursively derived by

  (i) dividing the total evidence in pieces, according to the independence structure represented by the DAG, and then

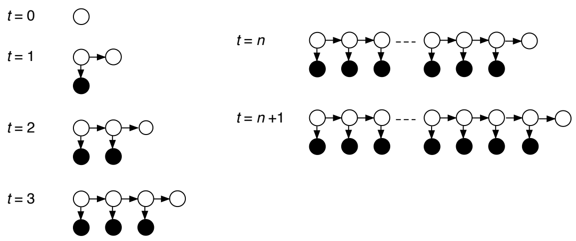
  (ii) incorporating evidence pieces in either the probability terms (π-messages) or the likelihood terms (conditional probability terms; λ-messages).

Queries answerable by the algorithm for a singly connected network:

- $P(X = x \mid \mathbf{e})$ for a single x;
- $P(X_v = x_v \mid \mathbf{e})$ for all x_v and $v \in V$;
- Most probable configurations, $\arg\max_x p(x \,\&\, \mathbf{e})$.

  This can be related to finding global optimal solutions by distributed local computation. (Details are given today.)

Practice: Belief Propagation for HMM



Observation variables (black) are instantiated; latent variables (white) are $X_1, X_2, \ldots$. The total evidence at time t is $\mathbf{e}_t$. How would you use message-passing to calculate

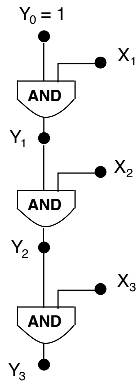
- $p(x_t \,|\, \mathbf{e}_t), \; \forall x_t$?

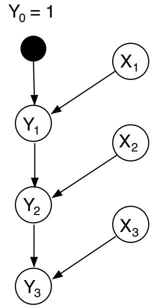
  (You'll obtain as a special case the so-called forward algorithm.)
- $p(x_{t+1} \,|\, \mathbf{e}_t), \; \forall x_{t+1}$? (This is a prediction problem.)
- $p(x_k \,|\, \mathbf{e}_t), \; \forall x_k, \; k < t$?

  (You'll obtain as a special case the so-called backward algorithm.)

A Fault-Detection Example

A logic circuit for fault detection and its Bayesian network (Pearl 1988):





$P(X_i = 1) = p_i,$
$P(X_i = 0) = 1 - p_i = q_i,$
$Y_i = Y_{i-1}$ AND X_i.

- $Y_0 = 1$ always.
- X_i is normal if $X_i = 1$, and faulty if $X_i = 0$.
- Normally all variables are on, and a failure occurs if $Y_3 = 0$.

Example: Belief Updating

Without observing any evidence, all the π-messages are prior probabilities:

$$\pi_{X_i,Y_i}(x_i) = \big[\, p_i \,,\, q_i \,\big], \quad i = 1,2,3; \quad \pi_{Y_0,Y_1}(y_0) = \big[\, 1 \,,\, 0 \,\big],$$
$$\pi_{Y_1,Y_2}(y_1) = \big[\, p_1 \,,\, q_1 \,\big], \quad \pi_{Y_2,Y_3}(y_2) = \big[\, p_1 p_2 \,,\, 1 - p_1 p_2 \,\big],$$

for $x_i = 1, 0$ and $y_i = 1, 0$.

Suppose $\mathbf{e} : \{X_2 = 1, Y_3 = 0\}$ is received. Then, X_2 updates its message to Y_2 and Y_2 updates its message to Y_3:

$$\pi_{X_2,Y_2}(x_2) = \big[\, p_2 \,,\, 0 \,\big], \quad \pi_{Y_2,Y_3}(y_2) = \big[\, p_1 p_2 \,,\, q_1 p_2 \,\big].$$

λ-messages starting from Y_3 upwards are given by:

$$\lambda_{Y_3,X_3}(x_3) = \big[\, p_2 q_1 \,,\, p_2 \,\big], \qquad \lambda_{Y_3,Y_2}(y_2) = \big[\, q_3 \,,\, 1 \,\big];$$
$$\lambda_{Y_2,X_2}(x_2) = \big[\, p_1 q_3 + q_1 \,,\, p_1 + q_1 q_3 \,\big], \qquad \lambda_{Y_2,Y_1}(y_1) = \big[\, p_2 q_3 \,,\, p_2 \,\big];$$
$$\lambda_{Y_1,X_1}(x_1) = \big[\, p_2 q_3 \,,\, p_2 \,\big].$$

So

$$P(X_3 = 0 \,|\, \mathbf{e}) = \frac{q_3 p_2}{p_3 p_2 q_1 + q_3 p_2} = \frac{q_3}{p_3 q_1 + q_3} = \frac{q_3}{1 - p_1 p_3},$$
$$P(X_1 = 0 \,|\, \mathbf{e}) = \frac{q_1 p_2}{p_1 p_2 q_3 + q_1 p_2} = \frac{q_1}{p_1 q_3 + q_1} = \frac{q_1}{1 - p_1 p_3}.$$

Example: Explanations based on Beliefs

If $q_1 = 0.45$ and $q_3 = 0.4$, we obtain

$$P(X_1 = 0 \,|\, \mathbf{e}) = 0.672 > P(X_1 = 1 \,|\, \mathbf{e}) = 0.328,$$
$$P(X_3 = 0 \,|\, \mathbf{e}) = 0.597 > P(X_3 = 1 \,|\, \mathbf{e}) = 0.403.$$

Is $I_1 = \{X_1 = 0, X_3 = 0\}$ the most probable explanation of $\mathbf{e}$, however?

There are three possible explanations

$$I_1 = \{X_1 = 0, X_3 = 0\}, \quad I_2 = \{X_1 = 0, X_3 = 1\}, \quad I_3 = \{X_1 = 1, X_3 = 0\}.$$

Direct calculation shows

$$P(I_1 \,|\, \mathbf{e}) = \frac{q_1 q_3}{1 - p_1 p_3}, \quad P(I_2 \,|\, \mathbf{e}) = \frac{q_1 p_3}{1 - p_1 p_3}, \quad P(I_3 \,|\, \mathbf{e}) = \frac{p_1 q_3}{1 - p_1 p_3}.$$

So, if $0.5 > q_1 > q_2 > q_3$, then based on the evidence, I_2 is the most probable explanation, while I_1 is the *least* probable explanation.

Outline

Recall Notation for Singly Connected Networks

Consider a vertex v.

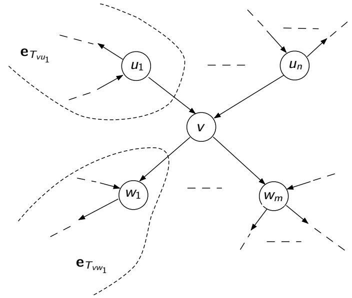
- $\text{pa}(v) = \{u_1, \ldots, u_n\}$, $\text{ch}(v) = \{w_1, \ldots, w_m\}$;
- $T_{vu},\ u \in \text{pa}(v)$: the sub-polytree containing the parent u, resulting from removing the edge (u, v);
- $T_{vw},\ w \in \text{ch}(v)$: the sub-polytree containing the child w, resulting from removing the edge (v, w).

For a sub-polytree T, denote

- X_T: the variables associated with nodes in T
- $\mathbf{e}_T$: the partial evidence of X_T

Divide the total evidence $\mathbf{e}$ in pieces:

- $\mathbf{e}_{T_{vu}}, u \in \text{pa}(v)$;
- $\mathbf{e}_v$;
- $\mathbf{e}_{T_{vw}}, w \in \text{ch}(v)$.



We want to solve: $\max_x p(x \,\&\, \mathbf{e})$.

Derivation of the Message Passing Algorithm

Evidence structure: We can express the joint distribution $P(X)$ as

$$p(x) = \prod_{u \in \text{pa}(v)} p(x_{T_{vu}}) \cdot p(x_v \,|\, x_{\text{pa}(v)}) \cdot \prod_{w \in \text{ch}(v)} p(x_{T_{vw}} \,|\, x_v). \tag{1}$$

We then enter the evidence $\mathbf{e}$ (put each piece in a proper term) to obtain

$$p(x \,\&\, \mathbf{e}) = \prod_{u \in \text{pa}(v)} p(x_{T_{vu}} \,\&\, \mathbf{e}_{T_{vu}}) \cdot p(x_v \,\&\, \mathbf{e}_v \,|\, x_{\text{pa}(v)}) \cdot \prod_{w \in \text{ch}(v)} p(x_{T_{vw}} \,\&\, \mathbf{e}_{T_{vw}} \,|\, x_v). \tag{2}$$

(For a detailed derivation of Eqs. (1) and (2), see slides 24-27.)

Max-Product: To solve $\max_x p(x \,\&\, \mathbf{e})$, we consider maximizing with respect to groups of variables in the following order:

$$\max_x \;\Leftrightarrow\; \max_{x_v} \max_{x_{\text{pa}(v)}} \max_{x_{T_{vu_1} \setminus \{u_1\}}} \cdots \max_{x_{T_{vu_n} \setminus \{u_n\}}} \max_{x_{T_{vw_1}}} \cdots \max_{x_{T_{vw_m}}},$$

where $T_{vu} \setminus \{u\}$ denotes the set of nodes in the sub-polytree T_{vu} except for $\{u\}$.

Notice that for any two functions $f_1(x), f_2(x, y)$, we have the identity

$$\max_{x,y} \{f_1(x) f_2(x, y)\} = \max_x \Big\{ f_1(x) \cdot \big(\max_y f_2(x, y)\big) \Big\}.$$

We will similarly move certain maximization operations inside the products in Eq. (2) to obtain a desirable factor form of $\max_x p(x \,\&\, \mathbf{e})$.

Derivation of the Message Passing Algorithm

Consider first the maximization with respect to $x_{T_{vw}}, w \in \text{ch}(v)$. We have

$$\max_{x_{T_{vw_1}}} \cdots \max_{x_{T_{vw_m}}} p(x \,\&\, \mathbf{e}) = \Big(\prod_{u \in \text{pa}(v)} p(x_{T_{vu}} \,\&\, \mathbf{e}_{T_{vu}}) \Big) \cdot p(x_v \,\&\, \mathbf{e}_v \,|\, x_{\text{pa}(v)}) \cdot \prod_{w \in \text{ch}(v)} \max_{x_{T_{vw}}} p(x_{T_{vw}} \,\&\, \mathbf{e}_{T_{vw}} \,|\, x_v).$$

Maximizing the above expression with respect to $x_{T_{vu_1} \setminus \{u_1\}}, \ldots, x_{T_{vu_n} \setminus \{u_n\}}$, we obtain

$$\Big(\prod_{u \in \text{pa}(v)} \max_{x_{T_{vu} \setminus \{u\}}} p(x_{T_{vu}} \,\&\, \mathbf{e}_{T_{vu}}) \Big) \cdot p(x_v \,\&\, \mathbf{e}_v \,|\, x_{\text{pa}(v)}) \cdot \prod_{w \in \text{ch}(v)} \max_{x_{T_{vw}}} p(x_{T_{vw}} \,\&\, \mathbf{e}_{T_{vw}} \,|\, x_v).$$

Define

$$p^*(x_u \,\&\, \mathbf{e}_{T_{vu}}) = \max_{x_{T_{vu} \setminus \{u\}}} p(x_{T_{vu}} \,\&\, \mathbf{e}_{T_{vu}}), \qquad p^*(\mathbf{e}_{T_{vw}} \,|\, x_v) = \max_{x_{T_{vw}}} p(x_{T_{vw}}, \mathbf{e}_{T_{vw}} \,|\, x_v). \tag{3}$$

We obtain

$$\max_x p(x \,\&\, \mathbf{e}) = \max_{x_v} \Big(\max_{x_{\text{pa}(v)}} \prod_{u \in \text{pa}(v)} p^*(x_u \,\&\, \mathbf{e}_{T_{vu}}) \cdot p(x_v \,\&\, \mathbf{e}_v \,|\, x_{\text{pa}(v)}) \Big) \cdot \prod_{w \in \text{ch}(v)} p^*(\mathbf{e}_{T_{vw}} \,|\, x_v).$$

We will call the expression inside '$\max x_v$' the max-margin of X_v, denoted $p^*(x_v \,\&\, \mathbf{e})$.

Derivation of the Message Passing Algorithm

Thus we obtain

$$\max_x p(x \,\&\, \mathbf{e}) = \max_{x_v} p^*(x_v \,\&\, \mathbf{e})$$

where

$$p^*(x_v \,\&\, \mathbf{e}) = \Big(\max_{x_{\mathrm{pa}(v)}} \prod_{u \in \mathrm{pa}(v)} p^*(x_u \,\&\, \mathbf{e}_{T_{vu}}) \cdot p(x_v \,\&\, \mathbf{e}_v \,|\, x_{\mathrm{pa}(v)}) \Big) \cdot \prod_{w \in \mathrm{ch}(v)} p^*(\mathbf{e}_{T_{vw}} \,|\, x_v). \tag{4}$$

If v can receive messages

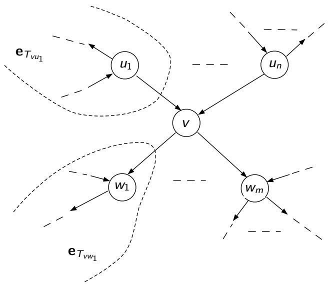
- $\pi^*_{u,v}$ from all parents, where
  $$\pi^*_{u,v}(x_u) = p^*(x_u \,\&\, \mathbf{e}_{T_{vu}}), \quad \forall x_u,$$
- $\lambda^*_{w,v}$ from all children, where
  $$\lambda^*_{w,v}(x_v) = p^*(\mathbf{e}_{T_{vw}} \,|\, x_v), \quad \forall x_v,$$

then v can calculate its max-margin

$$p^*(x_v \,\&\, \mathbf{e}), \quad \forall x_v,$$

and from which

$$\max_{x_v} p^*(x_v \,\&\, \mathbf{e}) = \max_x p(x \,\&\, \mathbf{e}).$$



Meanings of the Messages and Max-Margin

- $p^*(x_u \,\&\, \mathbf{e}_{T_{vu}})$: If $X_u = x_u$, there exists some configuration of $x_{T_{vu}}$ which best explains the partial evidence $\mathbf{e}_{T_{vu}}$ with this probability.
- $p^*(\mathbf{e}_{T_{vw}} \,|\, x_v)$: If $X_v = x_v$, there exists some configuration of $x_{T_{vw}}$ which best explains the partial evidence $\mathbf{e}_{T_{vw}}$ conditional on X_v, with this probability.
- $p^*(x_v \,\&\, \mathbf{e})$: If $X_v = x_v$, there exists some configuration of the rest of the variables which best explains the evidence $\mathbf{e}$ with this probability.

How to obtain $x^* \in \arg\max_x p(x \,\&\, \mathbf{e})$?

- If x^* is unique, then the solutions $x^*_v \in \arg\max_{x_v} p^*(x_v \,\&\, \mathbf{e})$ for all v form the global optimal solution (best explanation) x^*.
- If x^* is not unique, then we will need to trace out a solution from some node v. This shows that for each $x^*_v \in \arg\max_{x_v} p^*(x_v \,\&\, \mathbf{e})$, v should record the corresponding best values $x^*_{\mathrm{pa}(v)}$ of the parents in the maximization problem defining $p^*(x_v \,\&\, \mathbf{e})$ [Eq. (4)]:
  $$\max_{x_{\mathrm{pa}(v)}} \prod_{u \in \mathrm{pa}(v)} p^*(x_u \,\&\, \mathbf{e}_{T_{vu}}) \cdot p(x^*_v \,\&\, \mathbf{e}_v \,|\, x_{\mathrm{pa}(v)}).$$

Derivation of the Message Passing Algorithm

Now we only need to check if v can compose messages for its parents and children to calculate their max-margins.

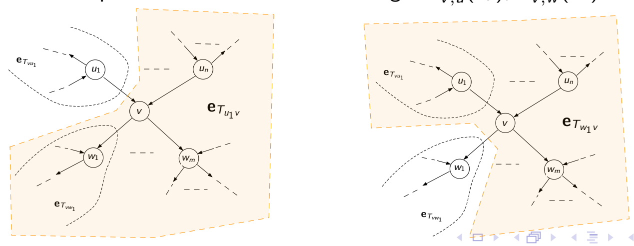
- A parent u needs $p^*(\mathbf{e}_{T_{uv}} \,|\, x_u)$ for all x_u based on the partial evidence $\mathbf{e}_{T_{uv}}$ from the sub-polytree on v's side with respect to u:
  $$p^*(\mathbf{e}_{T_{uv}} \,|\, x_u) = \max_{x_{T_{uv}}} p(x_{T_{uv}} \,\&\, \mathbf{e}_{T_{uv}} \,|\, x_u).$$
  Indeed it is given by
  $$p^*(\mathbf{e}_{T_{uv}} \,|\, x_u) = \max_{x_v} \Big\{ \Big(\max_{x_{\mathrm{pa}(v)\setminus\{u\}}} p(x_v \,\&\, \mathbf{e}_v \,|\, x_{\mathrm{pa}(v)}) \cdot \prod_{u' \in \mathrm{pa}(v)\setminus\{u\}} p^*(x_{u'} \,\&\, \mathbf{e}_{T_{vu'}}) \Big) \cdot \prod_{w \in \mathrm{ch}(v)} p^*(\mathbf{e}_{T_{vw}} \,|\, x_v) \Big\} \tag{5}$$
  $$= \max_{x_v} \Big\{ \Big(\max_{x_{\mathrm{pa}(v)\setminus\{u\}}} p(x_v \,|\, x_{\mathrm{pa}(v)})\, \ell_v(x_v) \cdot \prod_{u' \in \mathrm{pa}(v)\setminus\{u\}} \pi^*_{u',v}(x_{u'}) \Big) \cdot \prod_{w \in \mathrm{ch}(v)} \lambda^*_{w,v}(x_v) \Big\}.$$
  So this is the message $\lambda^*_{v,u}(x_u)$ that v needs to send to u; it can be composed once v receives the messages from all the other linked nodes.

(For the details of derivation of Eq. (5), see slide 28.)

Derivation of the Message Passing Algorithm

- A child w needs $p^*(x_v \,\&\, \mathbf{e}_{T_{wv}})$ for all x_v, which incorporates the partial evidence $\mathbf{e}_{T_{wv}}$ from the sub-polytree on v's side with respect to w:
  $$p^*(x_v \,\&\, \mathbf{e}_{T_{wv}}) = \max_{x_{T_{wv}\setminus\{v\}}} p(x_{T_{wv}} \,\&\, \mathbf{e}_{T_{wv}}).$$
  By a similar calculation as in the previous slides, one can show that
  $$p^*(x_v \,\&\, \mathbf{e}_{T_{wv}}) = \Big(\max_{x_{\mathrm{pa}(v)}} p(x_v \,|\, x_{\mathrm{pa}(v)})\, \ell_v(x_v) \cdot \prod_{u \in \mathrm{pa}(v)} \pi^*_{u,v}(x_u) \Big) \cdot \prod_{w' \in \mathrm{ch}(v)\setminus\{w\}} \lambda^*_{w',v}(x_v).$$
  So this is the message $\pi^*_{v,w}(x_v)$ that v needs to send to w; it can be composed once v receives the messages from all the other linked nodes.

Illustration of the partial evidence that the messages $\lambda^*_{v,u}(x_u)$, $\pi^*_{v,w}(x_v)$ carry:



Max-Product Message Passing Algorithm Summary

Each node v

- sends to each u of its parents

$$\lambda^*_{v,u}(x_u) = \max_{x_v} \Big\{ \max_{x_{\text{pa}(v)\setminus\{u\}}} p(x_v \,|\, x_{\text{pa}(v)})\, \ell_v(x_v) \cdot \prod_{u' \in \text{pa}(v)\setminus\{u\}} \pi^*_{u',v}(x_{u'}) \cdot \prod_{w \in \text{ch}(v)} \lambda^*_{w,v}(x_v) \Big\}, \qquad \forall x_u;$$

- sends to each w of its children

$$\pi^*_{v,w}(x_v) = \prod_{w' \in \text{ch}(v)\setminus\{w\}} \lambda^*_{w',v}(x_v) \cdot \max_{x_{\text{pa}(v)}} p(x_v \,|\, x_{\text{pa}(v)})\, \ell_v(x_v) \cdot \prod_{u \in \text{pa}(v)} \pi^*_{u,v}(x_u), \quad \forall x_v;$$

- when receiving all messages from parents and children, calculates

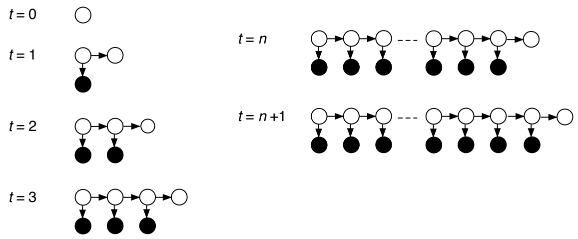
$$p^*(x_v \,\&\, \mathbf{e}) = \Big(\prod_{w \in \text{ch}(v)} \lambda^*_{w,v}(x_v) \Big) \cdot \max_{x_{\text{pa}(v)}} \prod_{u \in \text{pa}(v)} \pi^*_{u,v}(x_u) \cdot p(x_v \,|\, x_{\text{pa}(v)})\, \ell_v(x_v), \quad \forall x_v.$$

This is identical to the algorithm in the last lecture, with maximization replacing the summation.

To obtain a $x^* \in \arg\max_x p(x \,\&\, \mathbf{e})$:

- If x^* is unique, then it is given by $x^*_v \in \arg\max_{x_v} p^*(x_v \,\&\, \mathbf{e})$ for all v.
- If x^* is not unique, we can start from any node v, fix x^*_v and then trace out the solutions at other nodes.

HMM Example

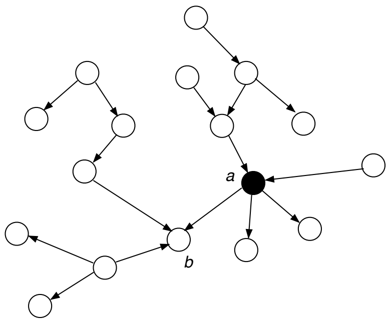


How would you use message-passing to calculate

- $\max_x p(x_1, \ldots, x_t \,|\, \mathbf{e}_t)$?

  (You'll obtain as a special case the Viterbi algorithm.)

Discussion on Differences between Algorithms



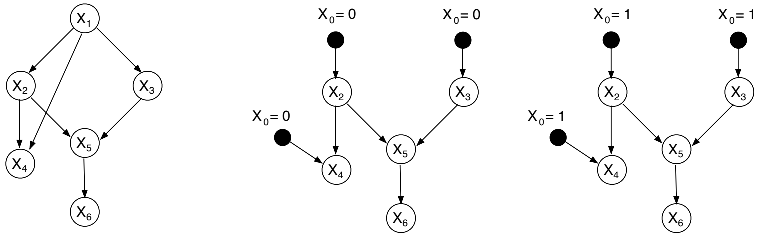
Node a is instantiated. Node b never receives any evidence. New pieces of evidence arrive to other nodes.

- Does a need to update messages to all the linked nodes for belief updating? for finding the most probable configuration?
- Does b need to update messages to all the linked nodes for belief updating? for finding the most probable configuration?

Outline

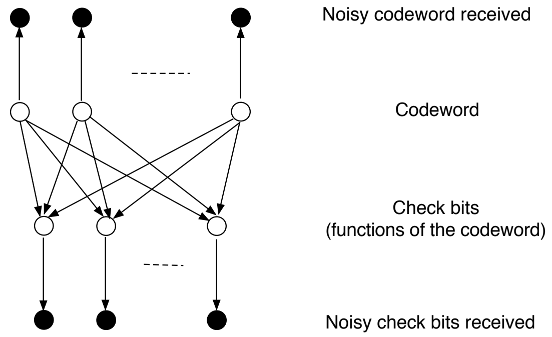
Illustration of Conditioning

Example (Pearl, 1988): Instantiating variable X_1 renders the network singly connected.



Turbo Decoding Example

Modified from McEliece et al., 1998:



Further Reading

1. Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, 1988. Chap. 5.

Details of Derivation for Eq. (1)

1. First we argue that $X_{T_{vu}}, u \in \text{pa}(v)$ are mutually independent. Abusing notation, for a sub-polytree T, we use T also for the set of nodes in T. Since G is singly connected, the subgraph $G_{\text{An}\left(\cup_{u\in\text{pa}(v)} T_{vu}\right)}$ consists of $n = |\text{pa}(v)|$ disconnected components, $T_{vu}, u \in \text{pa}(v)$. For any two disjoint subsets $U_1, U_2 \subseteq \text{pa}(v)$, the set of nodes $\cup_{u\in U_1} T_{vu}$ and $\cup_{u\in U_2} T_{vu}$ are disconnected, implying that

$$X_{\cup_{u\in U_1} T_{vu}} \perp X_{\cup_{u\in U_2} T_{vu}}$$

for any disjoint subsets U_1, U_2. This shows that $X_{T_{vu}}, u \in \text{pa}(v)$ are mutually independent, so

$$p(x_{T_{vu_1}}, \ldots, x_{T_{vu_n}}) = \prod_{u\in\text{pa}(v)} p(x_{T_{vu}}).$$

2. Next, choosing any well-ordering such that all the nodes in $T_{vu}, u \in \text{pa}(v)$ have smaller numbers than v, we can argue by (DO) that

$$p(x_v \,|\, x_{T_{vu_1}}, \ldots, x_{T_{vu_n}}) = p(x_v \,|\, x_{\text{pa}(v)}).$$

Combining this with the preceding equation, we have

$$p(x_{T_{vu_1}}, \ldots, x_{T_{vu_n}}, x_v) = \prod_{u\in\text{pa}(v)} p(x_{T_{vu}}) \cdot p(x_v \,|\, x_{\text{pa}(v)}).$$

Details of Derivation for Eq. (1)

3. Finally, we consider $X_{T_{vw}}, w \in \mathrm{ch}(v)$. Since G is singly connected, from G^m we see that v separates nodes in $T_{vw}, w \in \mathrm{ch}(v)$ from nodes in $T_{vu}, u \in \mathrm{pa}(v)$. Therefore,

$$\{X_{T_{vw}}, w \in \mathrm{ch}(v)\} \perp \{X_{T_{vu}}, u \in \mathrm{pa}(v)\} \mid X_v.$$

Furthermore, removing the node v, the subgraph of G^m induced by $T_{vw}, w \in \mathrm{ch}(v)$ is disconnected and has $m = |\mathrm{ch}(v)|$ components, each corresponding to a T_{vw}. So arguing as in the first step, we have that given X_v, the variables $X_{T_{vw}}, w \in \mathrm{ch}(v)$ are mutually independent. This gives us Eq. (1):

$$p(x) = \prod_{u \in \mathrm{pa}(v)} p(x_{T_{vu}}) \cdot p(x_v \mid x_{\mathrm{pa}(v)}) \cdot \prod_{w \in \mathrm{ch}(v)} p(x_{T_{vw}} \mid x_v).$$

Details of Derivation for Eq. (2)

Recall that the total evidence $\mathbf{e}$ has a factor form:

$$\mathbf{e}(x) = \prod_{v \in V} \ell_v(x_v).$$

For a given node v, we can also express $\mathbf{e}$ in terms of the pieces of evidence, $\mathbf{e}_v$, $\mathbf{e}_{T_{vu}}, u \in \mathrm{pa}(v)$ and $\mathbf{e}_{T_{vw}}, w \in \mathrm{ch}(v)$ as

$$\mathbf{e}(x) = \Big(\prod_{u \in \mathrm{pa}(v)} \mathbf{e}_{T_{vu}}(x_{T_{vu}}) \Big) \cdot \mathbf{e}_v(x_v) \cdot \prod_{w \in \mathrm{ch}(v)} \mathbf{e}_{T_{vw}}(x_{T_{vw}}),$$

where

$$\mathbf{e}_{T_{vu}}(x_{T_{vu}}) = \prod_{v' \in T_{vu}} \ell_{v'}(x_{v'}), \quad \mathbf{e}_v(x_v) = \ell_v(x_v), \quad \mathbf{e}_{T_{vw}}(x_{T_{vw}}) = \prod_{v' \in T_{vw}} \ell_{v'}(x_{v'}).$$

We now combine each piece of evidence with the respective term in $p(x)$, which by Eq. (1) is

$$p(x) = \prod_{u \in \mathrm{pa}(v)} p(x_{T_{vu}}) \cdot p(x_v \mid x_{\mathrm{pa}(v)}) \cdot \prod_{w \in \mathrm{ch}(v)} p(x_{T_{vw}} \mid x_v),$$

to obtain

$$p(x) \cdot \mathbf{e}(x) = \prod_{u \in \mathrm{pa}(v)} p(x_{T_{vu}})\, \mathbf{e}_{T_{vu}}(x_{T_{vu}}) \cdot p(x_v \mid x_{\mathrm{pa}(v)})\, \mathbf{e}_v(x_v) \cdot \prod_{w \in \mathrm{ch}(v)} p(x_{T_{vw}} \mid x_v) \mathbf{e}_{T_{vw}}(x_{T_{vw}}).$$

Details of Derivation for Eq. (2)

Using short-hand notation for probabilities of events (defined in Lec. 9), we have

$$p(x) \cdot \mathbf{e}(x) = p(x \,\&\, \mathbf{e}),$$
$$p(x_v \mid x_{\mathrm{pa}(v)}) \cdot \mathbf{e}_v(x_v) = p(x_v \,\&\, \mathbf{e}_v \mid x_{\mathrm{pa}(v)}),$$
$$p(x_{T_{vu}}) \cdot \mathbf{e}_{T_{vu}}(x_{T_{vu}}) = p(x_{T_{vu}} \,\&\, \mathbf{e}_{T_{vu}}),$$
$$p(x_{T_{vw}} \mid x_v) \cdot \mathbf{e}_{T_{vw}}(x_{T_{vw}}) = p(x_{T_{vw}} \,\&\, \mathbf{e}_{T_{vw}} \mid x_v).$$

So, we may write $P(X = x, \mathbf{e}) = p(x) \cdot \mathbf{e}(x)$ as

$$p(x \,\&\, \mathbf{e}) = \prod_{u \in \mathrm{pa}(v)} p(x_{T_{vu}} \,\&\, \mathbf{e}_{T_{vu}}) \cdot p(x_v \,\&\, \mathbf{e}_v \mid x_{\mathrm{pa}(v)}) \cdot \prod_{w \in \mathrm{ch}(v)} p(x_{T_{vw}} \,\&\, \mathbf{e}_{T_{vw}} \mid x_v),$$

which is Eq. (2).

Details of Derivation for Eq. (5)

We derive the expression for $p^*(\mathbf{e}_{T_{uv}} \mid x_u)$. Similar to the derivation of Eqs. (1)-(2),

$$p(x_{T_{uv}} \,\&\, \mathbf{e}_{T_{uv}} \mid x_u) = \prod_{u' \in \mathrm{pa}(v) \setminus \{u\}} p(x_{T_{vu'}} \,\&\, \mathbf{e}_{T_{vu'}}) \cdot p(x_v \,\&\, \mathbf{e}_v \mid x_{\mathrm{pa}(v)}) \cdot \prod_{w \in \mathrm{ch}(v)} p(x_{T_{vw}} \,\&\, \mathbf{e}_{T_{vw}} \mid x_v).$$

Also,

$$\max_{x_{T_{uv}}} \quad \Leftrightarrow \quad \max_{x_v} \max_{x_{\mathrm{pa}(v) \setminus \{u\}}} \max_{\substack{x_{T_{vu'}} \setminus \{u'\} \\ u' \in \mathrm{pa}(v)}} \max_{\substack{x_{T_{vw}} \\ w \in \mathrm{ch}(v)}},$$

Moving certain maximization operations inside the products, we obtain

$$p^*(\mathbf{e}_{T_{uv}} \mid x_u) = \max_{x_v} \max_{x_{\mathrm{pa}(v) \setminus \{u\}}} p(x_v \,\&\, \mathbf{e}_v \mid x_{\mathrm{pa}(v)}) \cdot \prod_{u' \in \mathrm{pa}(v) \setminus \{u\}} p^*(x_{u'} \,\&\, \mathbf{e}_{T_{vu'}}) \cdot \prod_{w \in \mathrm{ch}(v)} p^*(\mathbf{e}_{T_{vw}} \mid x_v).$$

By the definitions of messages in slide 13, this is

$$p^*(\mathbf{e}_{T_{uv}} \mid x_u) = \max_{x_v} \Big(\max_{x_{\mathrm{pa}(v) \setminus \{u\}}} p\big(x_v \mid x_{\mathrm{pa}(v)}\big)\, \ell_v(x_v) \cdot \prod_{u' \in \mathrm{pa}(v) \setminus \{u\}} \pi^*_{u',v}(x_{u'}) \Big) \cdot \prod_{w \in \mathrm{ch}(v)} \lambda^*_{w,v}(x_v).$$