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

Announcement: The last exercise will be posted online soon.

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 (\( \pi \)-messages) or the likelihood terms (conditional probability terms; \( \lambda \)-messages).

Queries answerable by the algorithm for a singly connected network:

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

This can be related to finding global optimal solutions by distributed local computation. (Details are given today.)
Observation variables (black) are instantiated; latent variables (white) are $X_1, X_2, \ldots$. The total evidence at time $t$ is $E_t$. How would you use message-passing to calculate

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

**Example: Belief Updating**

Without observing any evidence, all the $\pi$-messages are prior probabilities:

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

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

Suppose $e = \{X_1 = 1, Y_2 = 0\}$ is received. Then, $X_2$ updates its message to $Y_2$ and $Y_2$ updates its message to $Y_3$:

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

$\lambda$-messages starting from $Y_3$ upwards are given by:

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

So

- $P(X_3 = 0 | e) = \frac{q_1 p_3}{p_1 p_2 q_1 + q_1 p_2} = \frac{q_1}{p_1 q_2 + q_1} = \frac{q_1}{1 - p_1 p_2}$,
- $P(X_2 = 0 | e) = \frac{p_2 q_1}{p_2 q_3 + q_1 p_2} = \frac{q_1}{p_1 q_3 + q_1} = \frac{q_1}{1 - p_2 p_3}$.

**Example: Explanations based on Beliefs**

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

- $P(X_1 = 1 | e) = 0.672 > P(X_1 = 1 | e) = 0.328$,
- $P(X_1 = 0 | e) = 0.597 > P(X_1 = 1 | e) = 0.403$.

Is $h_1 = \{X_1 = 0, X_2 = 0\}$ the most probable explanation of $e$, however?

There are three possible explanations

- $h_1 = \{X_1 = 0, X_2 = 0\}$, $h_2 = \{X_1 = 0, X_3 = 1\}$, $h_3 = \{X_1 = 1, X_3 = 0\}$.

Direct calculation shows

- $P(h_1 | e) = \frac{q_1 q_3}{1 - p_1 p_3}$, $P(h_2 | e) = \frac{q_1 p_3}{1 - p_1 p_3}$, $P(h_3 | e) = \frac{p_1 q_3}{1 - p_1 p_3}$.

So, if $0.5 > q_1 > q_3$ then based on the evidence, $h_3$ is the most probable explanation, while $h_3$ is the least probable explanation.
We will similarly move certain maximization operations inside the products in
Evidence structure;
Max-Product.

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 $T$
- $e_T$: the partial evidence of $X_T$

Divide the total evidence $e$ in pieces:
- $e_T, u \in \text{pa}(v)$;
- $e_v$;
- $e_T, w \in \text{ch}(v)$.

We want to solve: $\max_x p(x \& e)$.

We obtain
\[
\max_{x \in \text{pa}(v)} p(x \& e_T) \cdot p(x \& e_v, \text{ch}(v)) = \sum_{x \in \text{ch}(v)} \max_{x \in \text{pa}(v)} p(x \& e_T) \cdot p(x \& e_v, \text{ch}(v)).
\]
Derivation of the Message Passing Algorithm

Thus we obtain

\[
\max_x p(x \& e) = \max_x p^*(x \& e)
\]

where

\[
p^*(x \& e) = \left( \max_v \left( \prod_{u \in \text{pa}(v)} p^*(x_u \& e_{Tu} \mid x_{Tu}) \cdot p(x_u \& e_v \mid x_{pa(v)}) \right) \cdot \prod_{w \in \text{ch}(v)} p^*(e_{Tw} \mid x_v) \right).
\]

(4)

If \( v \) can receive messages

- \( \pi_{u \to v} \) from all parents, where
  \( \pi_{u \to v}(x_u) = p^*(x_u \& e_{Tu} \mid x_{Tu}) \), \( \forall x_{Tu} \),

- \( \lambda^*_{u \to v} \) from all children, where
  \( \lambda^*_{u \to v}(x_u) = p^*(e_{Tu} \mid x_u) \), \( \forall x_u \),

then \( v \) can calculate its max-margin

\[
p^*(x_v \& e_v) \quad \forall x_v,
\]

and from which

\[
\max_{x_v} p^*(x_v \& e_v) = \max_x p(x \& e).
\]

Meansings of the Messages and Max-Margin

- \( p^*(x_u \& e_{Tu}) \): If \( X_u = x_u \), there exists some configuration of \( x_{Tu} \) which best explains the partial evidence \( e_{Tu} \) with this probability.

- \( p^*(e_{Tu} \mid x_u) \): If \( X_u = x_u \), there exists some configuration of \( x_{Tu} \) which best explains the partial evidence \( e_{Tu} \) conditional on \( X_u \) with this probability.

- \( p^*(x_v \& e) \): If \( X_v = x_v \), there exists some configuration of the rest of the variables which best explains the evidence \( e \) with this probability.

How to obtain \( x^* \in \arg \max_v p(x \& e) \)?

- If \( x^* \) is unique, then the solutions \( x^*_v \in \arg \max_{x_v} p^*(x_v \& e_v) \) 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 \( x^*_v \in \arg \max_{x_v} p^*(x_v \& e_v) \), \( v \) should record the corresponding best values \( x_{pa(v)} \) of the parents in the maximization problem defining \( p^*(x_v \& e_v) \) [Eq. (4)]:

\[
\max_{x_{pa(v)}} \prod_{u \in \text{pa}(v)} p^*(x_u \& e_{Tu}) \cdot p(x_u \& e_v \mid x_{pa(v)}).
\]

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 \( p^*(e_{Tu} \mid x_u) \) for all \( x_u \), based on the partial evidence \( e_{Tu} \) from the sub-polytree on \( v \)'s side with respect to \( u \):

\[
p^*(e_{Tu} \mid x_u) = \max_{x_{Tu}} p(x_{Tu} \& e_{Tu} \mid x_u).
\]

Indeed it is given by

\[
p^*(e_{Tu} \mid x_u) = \max_{x_{Tu}} \left\{ \left( \max_{x_{pa(v)} \mid (v)} p(x_u \& e_v \mid x_{pa(v)}) \cdot \prod_{u' \in \text{pa}(v) \setminus (u)} p^*(x_{u'} \& e_{Tu}) \right) \right. \\
\cdot \prod_{w \in \text{ch}(v)} p^*(e_{Tw} \mid x_v) \right\} \quad (5)
\]

\[
= \max_{x_{Tu}} \left\{ \left( \max_{x_{pa(v)} \mid (v)} p(x_u \mid x_{pa(v)}) \ell_v(x_v) \cdot \prod_{u' \in \text{pa}(v) \setminus (u)} \pi^*_{u' \to v}(x_{u'}) \right) \right. \\
\cdot \prod_{w \in \text{ch}(v)} \lambda^*_{u \to w}(x_u) \right\}.
\]

So this is the message \( \lambda^*_{u \to w}(x_u) \) that \( v \) needs to send to \( u \); it can be composed once \( v \) receives the messages from all the other linked nodes.

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

(For the details of derivation of Eq. (5), see slide 28.)
Max-Product Message Passing Algorithm Summary

Each node \( v \)
- sends to each \( u \) of its parents
  \[
  \lambda_{v,u}^*(x_u) = \max_{x_u} \left\{ \max_{x_{\text{pa}(v)} \setminus \{v\}} p(x_v | x_{\text{pa}(v)}) \ell_v(x_u) \cdot \prod_{u' \in \text{pa}(v) \setminus \{u\}} \pi_{u',v}^*(x_{u'}) \cdot \prod_{w \in \text{ch}(v)} \lambda_{w,v}^*(x_w) \right\}, \quad \forall x_u;
  \]
- sends to each \( w \) of its children
  \[
  \pi_{w,v}^*(x_w) = \max_{x_{\text{pa}(v)} \setminus \{w\}} p(x_v | x_{\text{pa}(v)}) \ell_v(x_w) \cdot \prod_{u \in \text{pa}(v)} \pi_{u,v}^*(x_u), \quad \forall x_v;
  \]
- when receiving all messages from parents and children, calculates
  \[
  \rho^*(x_v & e) = \left( \prod_{w \in \text{ch}(v)} \lambda_{w,v}^*(x_w) \right) \max_{x_{\text{pa}(v)}} \prod_{u \in \text{pa}(v)} \pi_{u,v}^*(x_u) \cdot p(x_v | x_{\text{pa}(v)}) \ell_v(x_v), \quad \forall x_v.
  \]

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

To obtain a \( x^* \in \arg \max_x \rho(x & e) \):
- If \( x^* \) is unique, then it is given by \( x^*_v \in \arg \max_{x_v} \rho^*(x_v & e) \) 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.

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?
Example (Pearl, 1988): Instantiating variable $X_3$ renders the network singly connected.

Modified from McEliece et al., 1998:

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 $G_{\text{An}(\bigcup_{u \in \text{pa}(v)} T_{vu})}$ 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 $\bigcup_{u \in U_1} T_{vu}$ and $\bigcup_{u \in U_2} T_{vu}$ are disconnected, implying that

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

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

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

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

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

Combining this with the preceding equation, we have

$$p(x_{T_{vu}}, \ldots, x_{T_{vu}}, x_v) = \prod_{u \in \text{pa}(v)} p(x_{T_{vu}}) \cdot p(x_v | x_{\text{pa}(v)}).$$
Details of Derivation for Eq. (1)

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

$$\{X_{T_{wv}}, w \in \text{ch}(v)\} \perp \{X_{T_{wv}}, u \in \text{pa}(v)\} \mid X_v.$$ 

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

$$p(x) = \prod_{w \in \text{pa}(v)} p(x_{T_{wv}}) \cdot p(x_v \mid x_{\text{pa}(v)}) \cdot \prod_{w \in \text{ch}(v)} p(x_{T_{wv}} \mid x_v).$$

Details of Derivation for Eq. (2)

Recall that the total evidence $e$ has a factor form:

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

For a given node $v$, we can also express $e$ in terms of the pieces of evidence, $e_v, e_{T_{wv}}, u \in \text{pa}(v)$ and $e_{T_{wv}}, w \in \text{ch}(v)$ as

$$e(x) = \left( \prod_{u \in \text{pa}(v)} e_{T_{wu}}(x_{T_{wu}}) \right) \cdot e_v(x_v) \cdot \prod_{w \in \text{ch}(v)} e_{T_{wv}}(x_{T_{wv}}),$$

where

$$e_{T_{wv}}(x_{T_{wv}}) = \prod_{v' \in T_{wv}} \ell_{v'}(x_{v'}), \quad e_v(x_v) = \ell_v(x_v), \quad e_{T_{wu}}(x_{T_{wu}}) = \prod_{v' \in T_{wu}} \ell_{v'}(x_{v'}).$$

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

$$p(x) = \prod_{w \in \text{pa}(v)} p(x_{T_{wv}}) \cdot p(x_v \mid x_{\text{pa}(v)}) \cdot \prod_{w \in \text{ch}(v)} p(x_{T_{wv}} \mid x_v),$$

to obtain

$$p(x) \cdot e(x) = \prod_{u \in \text{pa}(v)} p(x_{T_{wu}} \mid x_v) e_{T_{wu}}(x_{T_{wu}}) \cdot p(x_v \mid x_{\text{pa}(v)}) e_v(x_v) \cdot \prod_{w \in \text{ch}(v)} p(x_{T_{wv}} \mid x_v) e_{T_{wv}}(x_{T_{wv}}).$$

Details of Derivation for Eq. (5)

We derive the expression for $p^*(e_{T_{wv}} \mid x_v)$. Similar to the derivation of Eqs. (1)-(2),

$$p(x v \nu^* w & e_{T_{wu}} \mid x_v) = \prod_{w' \in \text{pa}(v) \setminus \{w\}} \cdot \prod_{w \in \text{ch}(v)} p(x_{T_{wv}} \mid x_v).$$

Also,

$$\max_{x_{T_{wu}}} \Rightarrow \max_{x_v \mid x_{\text{pa}(v)}} \max_{x_{T_{wu}} \mid (v')} \max_{x_{T_{wu}} \mid (v')} \cdot \max_{w \in \text{ch}(v)}.$$

Moving certain maximization operations inside the products, we obtain

$$p^*(e_{T_{wv}} \mid x_v) = \max_{x_{\text{pa}(v)}} \max_{x_{T_{wu}} \mid (v')} \cdot \max_{w \in \text{ch}(v)}.$$

By the definitions of messages in slide 13, this is

$$p^*(e_{T_{wv}} \mid x_v) = \max_{x_{\text{pa}(v)}} \max_{x_{T_{wu}} \mid (v')} \cdot \max_{w \in \text{ch}(v)}.$$