CSE 5523: Lecture Notes 20
Message Passing

Contents

Like in backpropagation, matrix chains can also define inference for Bayes nets.

20.1 Efficient inference in Bayes nets

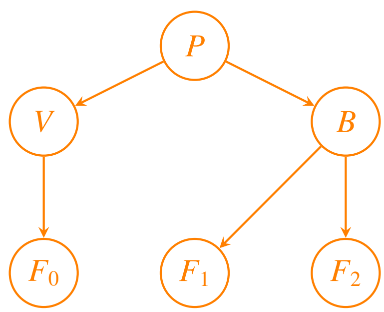
Bayes nets are models over joint probability spaces $\langle X_1 \times \ldots \times X_K, 2^{X_1 \times \cdots \times X_K}, \mathsf{P} \rangle$.

Using independence assumptions, each variable X_v conditions on subset $C_v \subseteq \{X_1, \ldots, X_{v-1}\}$.

(These conditioned-on variables get called 'parents', and the modeled variables are 'children'.)

Each X_v is then associated with a conditional probability matrix $\mathsf{P}(X_v \mid C_v) = \mathbf{M} \in \mathbb{R}^{(\prod_{X_u \in C_v} |X_u|) \times |X_v|}$.

For example, this network θ_{Sp} models phonemes, voicing, backness, and vowel formants:



There may be a lot of combinations of the variables in a Bayes net.

For example, a query on the variable b could be answered *inefficiently* using the full joint:

$$\mathsf{P}_{\theta_{Sp}}(b) = \sum_{p,v,f_0,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p, v, b, f_0, f_1, f_2)$$

$$\stackrel{\text{def}}{=} \sum_{p,v,f_0,f_1,f_2} \mathsf{P}_{\mathbf{M}_P}(p) \cdot \mathsf{P}_{\mathbf{M}_V}(v \mid p) \cdot \mathsf{P}_{\mathbf{M}_B}(b \mid p) \cdot \mathsf{P}_{\mathbf{M}_{F_0}}(f_0 \mid v) \cdot \mathsf{P}_{\mathbf{M}_{F_1}}(f_1 \mid b) \cdot \mathsf{P}_{\mathbf{M}_{F_2}}(f_2 \mid b)$$

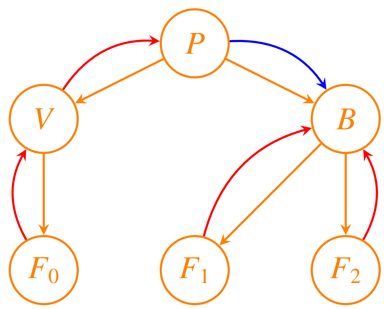
or *efficiently* by marginalizing as we go, storing marginals (probability tables) as 'messages':

$$\begin{aligned}
\mathsf{P}_{\theta_{Sp}}(b) &= \sum_{p,v,f_0,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p,v,b,f_0,f_1,f_2) \\
&\stackrel{\text{def}}{=} \sum_{p,v,f_0,f_1,f_2} \mathsf{P}_{\mathbf{M}_P}(p)\cdot\mathsf{P}_{\mathbf{M}_V}(v\,|\,p)\cdot\mathsf{P}_{\mathbf{M}_B}(b\,|\,p)\cdot\mathsf{P}_{\mathbf{M}_{F_0}}(f_0\,|\,v)\cdot\mathsf{P}_{\mathbf{M}_{F_1}}(f_1\,|\,b)\cdot\mathsf{P}_{\mathbf{M}_{F_2}}(f_2\,|\,b) \\
&\stackrel{\text{def}}{=} \sum_p \Big(\mathsf{P}(p)\cdot\Big(\sum_v \mathsf{P}(v\,|\,p)\cdot\Big(\sum_{f_0}\mathsf{P}(f_0\,|\,v)\Big)_v\Big)_p\Big)_p\cdot\mathsf{P}(b\,|\,p)\cdot\Big(\sum_{f_1}\mathsf{P}(f_1\,|\,b)\Big)_b\cdot\Big(\sum_{f_2}\mathsf{P}(f_2\,|\,b)\Big)_b
\end{aligned}$$

(Re-arrangement of terms just comes from distributing products over sums in the full joint.)

Blue parens show ***forward messages***: distributions over free modeled variables (subscripts).

Red parens show ***backward messages***: likelihood fns over free conditioned-on variables (subscr).



20.2 The message passing algorithm

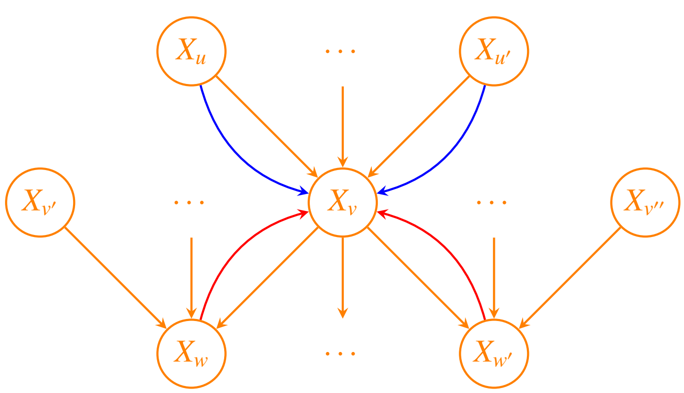
We can generally calculate probability distributions for query variables $\mathbf{y}_v^\top$ by passing messages.

These messages include:

- **forward messages** $\mathbf{f}_{v,w}^\top$: distributions over X_v given observations, and
- **backward messages** $\mathbf{b}_{v,u}$: likelihoods of observations given X_v.

(For simplicity, all probability tables $\mathbf{M}_v$ and forward messages $\mathbf{f}_{v,w}^\top$ have rows summing to one.)

So generally, in a graph that recursively repeats like this:



we have an equation for each modeled variable in terms of surrounding variables and observations:

$$\mathsf{P}(X_v, \text{obs}) = \sum_{x_1,\ldots,x_{v-1},x_{v+1},\ldots,x_V \in X_1,\ldots,X_{v-1},X_{v+1},\ldots,X_V} \prod_{v' \in \{1,\ldots,V\}} \mathsf{P}(X_{v'} \,|\, C_{v'})$$

$$= \sum_{x_u,\ldots,x_{u'} \in X_u,\ldots,X_{u'}} \overbrace{\mathsf{P}(x_u, \uparrow_v^u \text{ob}) \cdots \mathsf{P}(x_{u'}, \uparrow_v^{u'} \text{ob})}^{\text{forward messages}} \cdot \overbrace{\mathsf{P}(X_v \,|\, x_u, \ldots, x_{u'})}^{\text{modeled variable}} \cdot \overbrace{\mathsf{P}(\downarrow_w^v \text{ob} \,|\, X_v) \cdots \mathsf{P}(\downarrow_{w'}^v \text{ob} \,|\, X_v)}^{\text{backward messages}}$$

$$= \sum_{\times_{X_u \in C_v} X_u} \overbrace{\prod_{u \text{ s.t. } X_u \in C_v} \mathsf{P}(x_u, \uparrow_v^u \text{ob})}^{\text{forward messages}} \cdot \overbrace{\mathsf{P}(X_v \,|\, C_v)}^{\text{modeled variable}} \cdot \overbrace{\prod_{w \text{ s.t. } X_v \in C_w} \mathsf{P}(\downarrow_w^v \text{ob} \,|\, X_v)}^{\text{backward messages}}$$

where obs is all observations, and $\uparrow_v^u$ ob and $\downarrow_w^v$ ob are observations closer to X_u (or X_w) than X_v.

Forward message terms $\mathsf{P}(x_v, \uparrow_w^v \text{ob})$ are similar, but exclude destination in backward messages:

$$\mathsf{P}(x_v, \uparrow_w^v \text{ob}) = \sum_{\times_{X_u \in C_v} X_u} \overbrace{\prod_{u \text{ s.t. } X_u \in C_v} \mathsf{P}(x_u, \uparrow_v^u \text{ob})}^{\text{forward messages}} \cdot \overbrace{\mathsf{P}(X_v \,|\, C_v)}^{\text{modeled variable}} \cdot \overbrace{\prod_{w' \text{ s.t. } X_v \in C_{w'}, w' \neq w} \mathsf{P}(\downarrow_{w'}^v \text{ob} \,|\, X_v)}^{\text{backward messages}}$$

Backward message terms $\mathsf{P}(\downarrow_v^u \text{ob} \,|\, X_u)$ are similar, but exclude destination in forward messages:

$$\mathsf{P}(\downarrow_v^u \text{ob} \,|\, X_u) = \sum_{\times_{X'_u \in C_v - \{X_u\}} X_u} \overbrace{\prod_{u' \text{ s.t. } X'_u \in C_v, u' \neq u} \mathsf{P}(x_{u'}, \uparrow_v^{u'} \text{ob})}^{\text{forward messages}} \cdot \overbrace{\mathsf{P}(X_v \,|\, C_v)}^{\text{modeled variable}} \cdot \overbrace{\prod_{w \text{ s.t. } X_v \in C_w} \mathsf{P}(\downarrow_w^v \text{ob} \,|\, X_v)}^{\text{backward messages}}$$

This is equivalent to a marginal over all nuisance variables of the product of conditionals for all X_v.

20.3 Linear algebraic formulation

Forward messages $\mathbf{f}_{v,w}^\top$ multiply probabilities $\mathbf{M}_v$ by messages from parents $\mathbf{f}_{u,v}^\top$ and other kids $\mathbf{b}_{w',v}$:

$$\mathbf{f}_{v,w}^\top = \underbrace{\left(\bigotimes_{u \text{ s.t. } X_u \in C_v} \mathbf{f}_{u,v}^\top \right)}_{\text{joint of conditioned-on variables}} \mathbf{M}_v \bigodot_{w' \text{ s.t. } X_v \in C_{w'}, w' \neq w} \text{diag}(\mathbf{b}_{w',v})$$

…unless X_v is an observed variable, in which case $\mathbf{f}_{v,w}^\top = \delta_{X_v}^\top$.

Backward messages $\mathbf{b}_{v,u}$ multiply probabilities $\mathbf{M}_v$ by messages from kids $\mathbf{b}_{w,v}$ and other parents $\mathbf{f}_{u',v}^\top$:

$$\mathbf{b}_{v,u} = \underbrace{\left(\left(\bigotimes_{u' \text{ s.t. } X_{u'} \in C_v, u' < u} \mathbf{f}_{u',v}^\top \right) \otimes \text{diag}(\mathbf{1}) \otimes \left(\bigotimes_{u' \text{ s.t. } X_{u'} \in C_v, u' > u} \mathbf{f}_{u',v}^\top \right) \right)}_{\text{joint of conditioned-on variables, including } u} \mathbf{M}_v \bigodot_{w \text{ s.t. } X_v \in C_w} \mathbf{b}_{w,v}$$

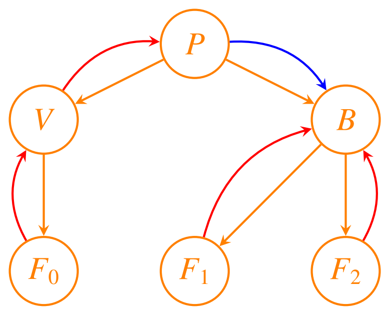
…unless X_v is observed, then $\mathbf{b}_{v,u} = \left(\left(\bigotimes_{u' \text{ s.t. } X_{u'} \in C_v, u' < u} \mathbf{f}_{u',v}^\top \right) \otimes \text{diag}(\mathbf{1}) \otimes \left(\bigotimes_{u' \text{ s.t. } X_{u'} \in C_v, u' > u} \mathbf{f}_{u',v}^\top \right) \right) \mathbf{M}_v \, \delta_{X_v}$.

The queried distribution is then the product of all forward and backward messages to that variable:

$$\mathbf{y}_v^\top = \underbrace{\left(\bigotimes_{X_u \in C_v} \mathbf{f}_{u,v}^\top\right)}_{\text{joint of conditioned-on variables}} \mathbf{M}_v \bigodot_{w \text{ s.t. } X_v \in C_w} \operatorname{diag}(\mathbf{b}_{w,v})$$

20.4 Example

For example, using the above network:



if we want to solve the following query (where variable F_0 is actually observed):

$$\mathsf{P}_{\theta_{Sp}}(b, f_0{=}12) = \sum_{p,v,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p, v, b, f_0{=}12, f_1, f_2)$$
$$\stackrel{\text{def}}{=} \sum_p \left({}_p\, \mathsf{P}(p) \cdot \left({}_p \sum_v \mathsf{P}(v \,|\, p) \cdot \mathsf{P}(f_0{=}12 \,|\, v)\right)\right) \cdot \mathsf{P}(b \,|\, p) \cdot \left({}_b \sum_{f_1} \mathsf{P}(f_1 \,|\, b)\right) \cdot \left({}_b \sum_{f_2} \mathsf{P}(f_2 \,|\, b)\right)$$

given the following models:

$\mathsf{P}_{\theta_P}(P) =$

/i/	/u/
.4	.6

$\mathsf{P}_{\theta_V}(V \,|\, P) =$

P	+	−
/i/	.8	.2
/u/	1	0

$\mathsf{P}_{\theta_{F_0}}(F_0 \,|\, V) =$

V	...	11	12	13	...
+	...	.04	.02	.01	...
−	...	.01	.01	.01	...

$\mathsf{P}_{\theta_B}(B \,|\, P) =$

P	+	−
/i/	0	1
/u/	.5	.5

we would generate the following messages:

from F_0 to V: $\mathsf{P}(F_0{=}12 \,|\, V) =$

V	12
+	.02
−	.01

from V to P: $\mathsf{P}(F_0{=}12 \mid P) =$

P	$F_0 = 12$
/i/	$\mathsf{P}_{\theta_{F_0}}(12 \mid +) \cdot \mathsf{P}_{\theta_V}(+ \mid /i/) + \mathsf{P}_{\theta_{F_0}}(12 \mid -) \cdot \mathsf{P}_{\theta_V}(- \mid /i/)$ $= .02 \cdot .8 + .01 \cdot .2 = .018$
/u/	$\mathsf{P}_{\theta_{F_0}}(12 \mid +) \cdot \mathsf{P}_{\theta_V}(+ \mid /u/) + \mathsf{P}_{\theta_{F_0}}(12 \mid -) \cdot \mathsf{P}_{\theta_V}(- \mid /u/)$ $= .02 \cdot 1 + .01 \cdot 0 = .020$

from P to B: $\mathsf{P}(P, F_0{=}12) =$

$P{=}/i/, F_0{=}12$	$P{=}/u/, F_0{=}12$
$\mathsf{P}_{\theta_P}(/i/) \cdot \mathsf{P}(F_0{=}12 \mid P{=}/i/)$ $= .4 \cdot .018 = .0072$	$\mathsf{P}_{\theta_P}(/u/) \cdot \mathsf{P}(F_0{=}12 \mid P{=}/u/)$ $= .6 \cdot .020 = .0120$

from F_1 to B: $\mathsf{P}(\text{any } F_1 \mid B) =$

B	any
+	1
−	1

from F_2 to B: $\mathsf{P}(\text{any } F_2 \mid B) =$

B	any
+	1
−	1

Product of model and three messages at B:

$\mathsf{P}(B, F_0{=}12) =$

$B{=}+, F_0{=}12$	$B{=}-, F_0{=}12$
$\mathsf{P}(P{=}/i/, F_0{=}12) \cdot \mathsf{P}_B(+ \mid /i/) \cdot 1 \cdot 1$ $+ \mathsf{P}(P{=}/u/, F_0{=}12) \cdot \mathsf{P}_B(+ \mid /u/) \cdot 1 \cdot 1$ $= .0072 \cdot 0 \cdot 1 \cdot 1 + .0120 \cdot .5 \cdot 1 \cdot 1 = .0060$	$\mathsf{P}(P{=}/i/, F_0{=}12) \cdot \mathsf{P}_B(- \mid /i/) \cdot 1 \cdot 1$ $+ \mathsf{P}(P{=}/u/, F_0{=}12) \cdot \mathsf{P}_B(- \mid /u/) \cdot 1 \cdot 1$ $= .0072 \cdot 1 \cdot 1 \cdot 1 + .0120 \cdot .5 \cdot 1 \cdot 1 = .0132$

Normalized:

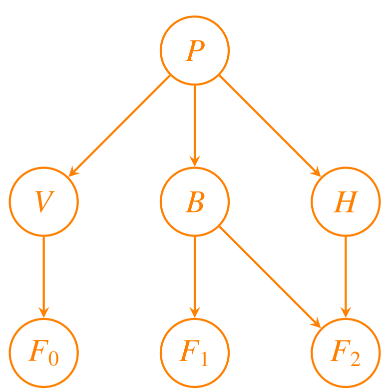
$\mathsf{P}(B \mid F_0{=}12) =$

$B{=}+$	$B{=}-$
$\frac{.0060}{.0060+.0132} = .3125$	$\frac{.0132}{.0060+.0132} = .6875$

20.5 Limits of message passing

Message passing degrades when the network is not singly connected.

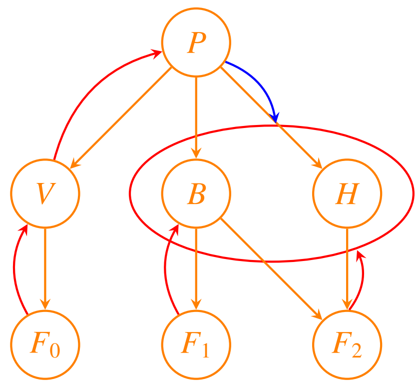
For example, adding a variable for height w. dependencies from P, to F_2, creates a 'diamond':



This means some marginals will have multiple free variables (which makes them larger):

$$
\begin{aligned}
\mathsf{P}_{\theta_{Sp}}(b) &= \sum_{p,v,h,f_0,f_1,f_2} \mathsf{P}_{\theta_{Sp}}(p,v,b,h,f_0,f_1,f_2) \\
&\overset{\text{def}}{=} \sum_{p} \Big(_{p} \mathsf{P}(p) \cdot \Big(_{p} \sum_{v} \mathsf{P}(v \mid p) \cdot \ldots \Big)\Big) \cdot \mathsf{P}(b \mid p) \cdot \Big(_{b} \sum_{f_1} \mathsf{P}(f_1 \mid b)\Big) \cdot \sum_{h} \mathsf{P}(h \mid p) \cdot \Big(_{b,h} \sum_{f_2} \mathsf{P}(f_2 \mid b,h)\Big)
\end{aligned}
$$

Graphically, messages must pass through ‘junctions’ of joint variables:

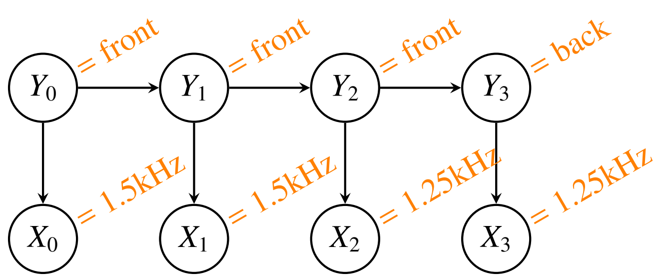


Well, they’re not full joints at least.

20.6 Hidden Markov models

Like neural nets, Bayes nets can be stationary too.

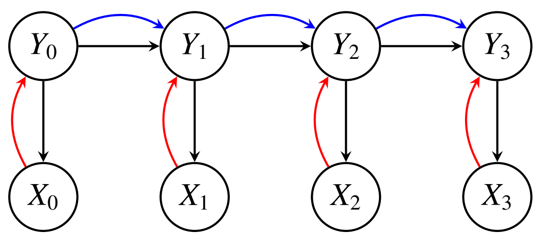
If they have a single interdependent sequence, they are called **hidden Markov models**.



Inference in these models uses message passing as well, but it’s called **filtering**:

$$
\begin{aligned}
\mathsf{P}_{\theta_{HMM}}(y_3) &= \sum_{x_0,x_1,x_2,x_3,y_0,y_1,y_2} \mathsf{P}_{\theta_{HMM}}(x_0,x_1,x_2,x_3,y_0,y_1,y_2,y_3) \\
&\overset{\text{def}}{=} \sum_{y_2} \Big(_{y_2} \sum_{y_1} \Big(_{y_1} \sum_{y_0} \Big(_{y_0} \mathsf{P}(y_0) \cdot \Big(_{y_0} \sum_{x_0} \mathsf{P}(x_0 \mid y_0)\Big)\Big) \cdot \mathsf{P}(y_1 \mid y_0) \cdot \Big(_{y_1} \sum_{x_1} \mathsf{P}(x_1 \mid y_1)\Big)\Big) \cdot \\
&\qquad \mathsf{P}(y_2 \mid y_1) \cdot \Big(_{y_2} \sum_{x_2} \mathsf{P}(x_2 \mid y_2)\Big)\Big) \cdot \mathsf{P}(y_3 \mid y_2) \cdot \Big(_{y_3} \sum_{x_3} \mathsf{P}(x_3 \mid y_3)\Big)
\end{aligned}
$$

Here it is, graphically:



and as a matrix chain, where $\mathbf{p} = \mathsf{P}(Y_0)$, $\mathbf{A} = \mathsf{P}(Y_t \mid Y_{t-1})$, $\mathbf{B} = \mathsf{P}(X_t \mid Y_t)$:

$$\mathsf{P}(Y_3, x_{0..3}) = \mathbf{p}^\top \mathrm{diag}(\mathbf{B}\ \delta_{x_0})\ \mathbf{A}\ \mathrm{diag}(\mathbf{B}\ \delta_{x_1})\ \mathbf{A}\ \mathrm{diag}(\mathbf{B}\ \delta_{x_2})\ \mathbf{A}\ \mathrm{diag}(\mathbf{B}\ \delta_{x_3})$$