Lecture 6: Exact Inference via Clique-Tree Algorithms

Liam Solus

KTH Royal Institute of Technology

solus@kth.se

15 September 2023
Graphical Models PhD Course
WASP

Graph structure gives us complexity bounds on VE.

Graph structure (specifically of chordal graphs) can also give us exact inference algorithms with nice computational benefits.

Example (Hidden Markov Model). Suppose we have a distribution Markov to the following graph:

$$\begin{tikzcd} X_1 \arrow[r] \arrow[d] & X_2 \arrow[r] \arrow[d] & X_3 \arrow[r] \arrow[d] & X_4 \arrow[r] \arrow[d] & X_5 \arrow[d] \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{tikzcd}$$

$X_i|X_{i-1} = x_{i-1} \sim \mathrm{U}(\{x_{i-1}/m, \ldots, (m-1)x_{i-1}/m\})$ $\quad m > 1$ fixed and $i = 1, \ldots, n$

$Y_i|X_i = x_i \sim \mathrm{Bin}(n, x_i)$ $\quad n > 0$ fixed and $i = 1, \ldots, n$

$[X_1, \ldots, X_5]^T$ is a joint prior for $\mathbf{Y} = [Y_1, \ldots, Y_5]^T$. $\quad$ model for sequences

We observe data $\mathbf{y} = [y_1, \ldots, y_5]^T$. Can we estimate the (marginal) posterior distributions $f_{X_i|\mathbf{Y}}(x_i|\mathbf{y})$?

$$\begin{tikzcd} X_1 \arrow[r] \arrow[d] & X_2 \arrow[r] \arrow[d] & X_3 \arrow[r] \arrow[d] & X_4 \arrow[r] \arrow[d] & X_5 \arrow[d] \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 \end{tikzcd}$$

One answer: Run VE. Need many runs to compute all of $f_{X_i|\mathbf{Y}}(x_i|\mathbf{y})$

Alternatively, we reuse τ_i's that we compute in one run in another run. This is the basic idea of the **Clique-Tree Algorithm**.

Start without evidence and assume we have estimated the conditional distributions: $X_i|X_{i-1} = x_{i-1}$ and $Y_i|X_i = x_i$.

Elimination order:

$\prec = (Y_5, X_5, Y_4, X_4, Y_3, X_3, Y_2, X_2, Y_1)$ get $f_{X_1}(x_1)$ with VE

i-th step creates factors: ψ_i and τ_i

$$\begin{aligned} \psi_1(x_5, y_5) &= f_{Y_5|X_5}(y_5|x_5) && \Rightarrow && \tau_1(x_5) \\ \psi_2(x_4, x_5) &= \tau_1(x_5) f_{X_5|X_4}(x_5|x_4) && \Rightarrow && \tau_2(x_4) \\ & && \vdots && \end{aligned}$$

$$
\begin{tikzcd}
X_1 \arrow[r] \arrow[d] & X_2 \arrow[r] \arrow[d] & X_3 \arrow[r] \arrow[d] & X_4 \arrow[r] \arrow[d] & X_5 \arrow[d] \\
Y_1 & Y_2 & Y_3 & Y_4 & Y_5
\end{tikzcd}
$$

τ_i is a **message** being passed from node i to its parent j.

The parent node then uses τ_i in the computation of its own factor ψ_{i-1} called its **potential**.

$$
\begin{array}{llcll}
\text{potential for } Y_5\text{:} & \psi_1(x_5, y_5) & \Rightarrow & \tau_1(x_5) & \text{message from } Y_5 \text{ to } X_5 \\
\text{potential for } X_5\text{:} & \psi_2(x_4, x_5) & \Rightarrow & \tau_2(x_4) & \text{message from } X_5 \text{ to } X_4
\end{array}
$$

Since the τ_i's are messages getting passed from one node i to another node j we denote them $m_{i \to j}$.

$$\begin{tikzcd}
X_1 \arrow[r] \arrow[d] & X_2 \arrow[r] \arrow[d] & X_3 \arrow[r] \arrow[d] & X_4 \arrow[r] \arrow[d] & X_5 \arrow[d] \\
Y_1 & Y_2 & Y_3 & Y_4 & Y_5
\end{tikzcd}$$

Messages are not really getting passed between **nodes**, but really **cliques** of nodes used in the potential that receives the message:

$$\begin{tikzcd}
X_1, Y_1 \arrow[r] & X_1, X_2 \arrow[r] \arrow[d] & X_2, X_3 \arrow[r] \arrow[d] & X_3, X_4 \arrow[r] \arrow[d] & X_4, X_5 \arrow[d] \\
 & X_2, Y_2 & X_3, Y_3 & X_4, Y_4 & X_5, Y_5
\end{tikzcd}$$

Cliques:

$$\mathcal{C} = \{C_1 = \{X_1, Y_1\},\ C_2 = \{X_1, X_2\},\ C_3 = \{X_2, X_3\},\ C_4 = \{X_3, X_4\},$$
$$C_5 = \{X_4, X_5\},\ C_6 = \{X_2, Y_2\},\ C_7 = \{X_3, Y_3\},\ C_8 = \{X_4, Y_4\},\ C_9 = \{X_5, Y_5\}\}.$$

Initial Factors:

$$\Phi = \{f(x_1), f(x_2|x_1), f(x_3|x_2), f(x_4|x_3), f(x_5|x_4),$$
$$f(y_1|x_1), f(y_2|x_2), f(y_3|x_3), f(y_4|x_4), f(y_5|x_5)\}.$$

$\alpha : \Phi \to \mathcal{C}$ such that

Scope$[\phi] \subseteq \alpha(C)$

$$\begin{tikzcd}
X_1, Y_1 \arrow[r] & X_1, X_2 \arrow[r] \arrow[d] & X_2, X_3 \arrow[r] \arrow[d] & X_3, X_4 \arrow[r] \arrow[d] & X_4, X_5 \arrow[d] \\
& X_2, Y_2 & X_3, Y_3 & X_4, Y_4 & X_5, Y_5
\end{tikzcd}$$

Call this graph $\mathcal{T}$

$\alpha : \Phi \longrightarrow \mathcal{C}$ defines the **initial potentials**: $\psi_i = \prod_{\phi:\alpha(\phi)=C_i} \phi$.

$$\begin{aligned}
\psi_1(x_1, y_1) &= f(x_1)f(y_1|x_1) && \text{initial potential for } C_1 = \{X_1, Y_1\} \\
\psi_2(x_1, x_2) &= f(x_2|x_1) && \text{initial potential for } C_2 = \{X_1, X_2\} \\
\psi_3(x_2, x_3) &= f(x_3|x_2) && \text{initial potential for } C_3 = \{X_2, X_3\} \\
\psi_4(x_3, x_4) &= f(x_4|x_3) && \text{initial potential for } C_4 = \{X_3, X_4\} \\
\psi_5(x_5, x_4) &= f(x_5|x_4) && \text{initial potential for } C_5 = \{X_4, X_5\} \\
\psi_6(x_2, y_2) &= f(y_2|x_2) && \text{initial potential for } C_6 = \{X_2, Y_2\} \\
\psi_7(x_2, y_2) &= f(y_3|x_3) && \text{initial potential for } C_7 = \{X_3, Y_3\} \\
\psi_8(x_2, y_2) &= f(y_4|x_4) && \text{initial potential for } C_8 = \{X_4, Y_4\} \\
\psi_9(x_2, y_2) &= f(y_5|x_5) && \text{initial potential for } C_9 = \{X_5, Y_5\}
\end{aligned}$$

$$
\begin{tikzcd}
X_1, Y_1 \arrow[r] & X_1, X_2 \arrow[r] \arrow[d] & X_2, X_3 \arrow[r] \arrow[d] & X_3, X_4 \arrow[r] \arrow[d] & X_4, X_5 \arrow[d] \\
 & X_2, Y_2 & X_3, Y_3 & X_4, Y_4 & X_5, Y_5
\end{tikzcd}
$$

$\mathcal{T}$ is a rooted tree with source node $C_1 = \{X_1, Y_1\}$:

Every node C_i in $\mathcal{T}$ has a unique parent.

VE according to any linear extension of this rooted tree will compute for us $f_{X_1,Y_1}(x_1, y_1)$.

Along the way we compute the **sum-product messages**:

$$
\delta_{i\to j}(S_{i,j}) = \sum_{C_i \setminus S_{i,j}} \psi_i(C_i) \prod_{k \in \mathrm{Ne}_{\mathcal{T}}(C_i)\setminus\{C_j\}} \delta_{k\to i}(S_{i,k}),
$$

where $S_{i,j} = C_i \cap C_j$ (a **separator set**).

The root clique C_r produces a factor called the **beliefs**:

$$\beta_r(C_r) = \psi_r(C_r) \prod_{k \in \mathrm{Ne}_{\mathcal{T}}(C_r)} \delta_{k \to r}(S_{r,k}).$$

In our example, we started with a DAG and did VE, so we have $r = 1$ and get

$$\beta_r(C_r) = \beta_r(x_1, y_1) = \psi_1(x_1, x_2)\delta_{2 \to 1}(x_1) = f_{X_1, Y_1}(x_1, y_1)$$

We have that:

$$\sum_{y_1 \in \mathrm{Val}[Y_1]} \beta_1(x_1, y_1) = f_{X_1}(x_1).$$

This is one of the marginals we wanted!

(essentially computed via VE in a fancy language)

$\mathcal{T}$ was used to compute $f_{X_1}(x_1)$:

$$\begin{tikzcd}
X_1, Y_1 \arrow[r] & X_1, X_2 \arrow[r] \arrow[d] & X_2, X_3 \arrow[r] \arrow[d] & X_3, X_4 \arrow[r] \arrow[d] & X_4, X_5 \arrow[d] \\
& X_2, Y_2 & X_3, Y_3 & X_4, Y_4 & X_5, Y_5
\end{tikzcd}$$

If we change the root node we get a different marginal. The following orientation $\mathcal{T}'$ can be used to compute $f_{X_2}(x_2)$:

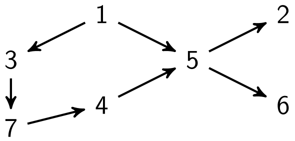
$$\begin{tikzcd}
X_1, Y_1 & X_1, X_2 \arrow[l] \arrow[r] \arrow[d] & X_2, X_3 \arrow[r] \arrow[d] & X_3, X_4 \arrow[r] \arrow[d] & X_4, X_5 \arrow[d] \\
& X_2, Y_2 & X_3, Y_3 & X_4, Y_4 & X_5, Y_5
\end{tikzcd}$$

Note that all the messages computed here are exactly the same except for now we compute $\delta_{1\to 2}(x_1)$ and the beliefs $\beta_2(x_1, x_2)$.

This means that we can reuse messages computed in our previous run for $\mathcal{T}$!

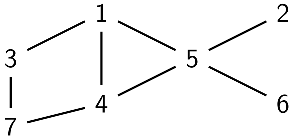
This algorithm works over any clique tree and hence is called the **clique-tree algorithm**.

Suppose we have a distribution $\mathbf{X} = [X_1, X_2, X_3, X_4, X_5, X_6, X_7]^T$ that is Markov to the following DAG $\mathcal{G}$:



$$f_{\mathbf{X}}(\mathbf{x}) = f(x_1)f(x_2|x_5)f(x_3|x_1)f(x_4|x_7)f(x_5|x_1, x_4)f(x_6|x_5)f(x_7|x_3).$$

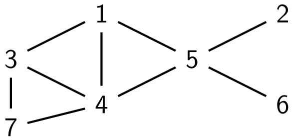
The distribution is also Markov to its **moralization** $\mathcal{G}^m$:



$$f_{\mathbf{X}}(\mathbf{x}) = h_1(x_1)h_2(x_2, x_5)h_3(x_1, x_3)h_4(x_4, x_7)h_5(x_1, x_4, x_5)h_6(x_5, x_6)h_7(x_3, x_7).$$

$$h_i = f(x_i|\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)})$$

Similarly, it is Markov to any **chordal cover (triangulation)** $\mathcal{G}^c$ of $\mathcal{G}^m$:

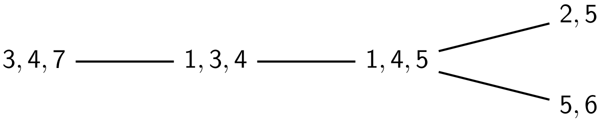


$$f_{\mathbf{X}}(\mathbf{x}) = \psi_{347}(x_3, x_4, x_7)\psi_{134}(x_1, x_3, x_4)\psi_{145}(x_1, x_4, x_5)\psi_{25}(x_2, x_5)\psi_{56}(x_5, x_6).$$

where

$$\begin{aligned}
\psi_{347}(x_3, x_4, x_7) &= h_4(x_4, x_7)h_7(x_3, x_7), \\
\psi_{134}(x_1, x_3, x_4) &= h_1(x_1)h_3(x_1, x_3), \\
\psi_{145}(x_1, x_4, x_5) &= h_5(x_1, x_4, x_5), \\
\psi_{25}(x_2, x_5) &= h_2(x_2, x_5), \\
\psi_{56}(x_5, x_6) &= h_6(x_5, x_6).
\end{aligned}$$

Since $\mathcal{G}^c$ is chordal we can find a clique-tree for $\mathcal{G}^c$, denoted $\mathcal{T}$:



$$\text{Cliques: } \mathcal{C} = \{C_1 = \{3,4,7\}, C_2 = \{1,3,4\}, C_3 = \{1,4,5\}, C_4 = \{2,5\}, C_5 = \{5,6\}\}.$$

$$\text{Initial potentials: } \Phi = \{\psi_{347}, \psi_{134}, \psi_{145}, \psi_{25}, \psi_{56}\}.$$

To compute the marginal distributions $f_{X_1}(x_1), f_{X_3}(x_3), f_{X_4}(x_4)$, compute the beliefs $\beta_2(x_1, x_3, x_4)$ using the orientation:

$$\begin{tikzcd}
 & & & 2,5 \\
3,4,7 & 1,3,4 \arrow[l] \arrow[r] & 1,4,5 \arrow[ur] \arrow[dr] & \\
 & & & 5,6
\end{tikzcd}$$

(three messages need to be computed)

Then to compute $f_{X_2}(x_2)$ and $f_{X_5}(x_5)$ use the orientation:

$$\begin{tikzcd}
 & & & 2,5 \arrow[dl] \\
3,4,7 & 1,3,4 \arrow[l] & 1,4,5 \arrow[l] \arrow[dr] & \\
 & & & 5,6
\end{tikzcd}$$

(two new messages need to be computed)

Then to compute $f_{X_6}(x_6)$ use the orientation:

$$
\begin{tikzcd}
 & & & 2,5 \\
3,4,7 & 1,3,4 \arrow[l] & 1,4,5 \arrow[l] \arrow[ur] & \\
 & & & 5,6 \arrow[ul]
\end{tikzcd}
$$

(one new message needs to be computed)

Finally, to compute $f_{X_7}(x_7)$ use the orientation:

$$
\begin{tikzcd}
 & & & 2,5 \\
3,4,7 \arrow[r] & 1,3,4 \arrow[r] & 1,4,5 \arrow[ur] \arrow[dr] & \\
 & & & 5,6
\end{tikzcd}
$$

(one new message needs to be computed)

7 messages computed in total.

Using multiple runs of VE to compute all seven marginals would mean computing 16 messages. 4 messages to compute each of:

$$
\begin{array}{ll}
f_{X_3,X_4,X_7}(x_3,x_4,x_7) & \qquad f_{X_5,X_6}(x_5,x_6) \\
f_{X_3,X_4,X_7}(x_1,x_4,x_5) & \qquad f_{X_2,X_5}(x_2,x_5)
\end{array}
$$

Final notes.

1. The equivalence of factorizations w.r.t. a graph and the markov properties for graphs give us multiple ways to interpret a graphical model (via factorization or conditional independence constraints).

2. Using these interpretations, we can learn a DAG (up to Markov equivalence) that represents relations in our data-generating distribution

3. After fitting the parameters for the DAG model according to the data we can use the DAG for inference

Final notes.

1. VE and the Clique-Tree algorithms are exact inference algorithms for computing marginals and posteriors exactly.

2. Combinatorics of the graph inform the complexity of these methods and speed up computations.

3. When the complexity of inference for the learned/constructed graph is too high we can consider **approximate inference algorithms**... To be continued in Module 2!