Graphical Models and Message passing

Yunshu Liu

ASPITRG Research Group

2013-07-16

References:

[1]. Steffen Lauritzen, *Graphical Models*, Oxford University Press, 1996

[2]. Michael I. Jordan, *Graphical models*, Statistical Science, Vol.19, No. 1. Feb., 2004

[3]. Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc. 2006

[4]. Daphne Koller and Nir Friedman, *Probabilistic Graphical Models - Principles and Techniques*, The MIT Press, 2009

[5].Kevin P. Murphy, *Machine Learning - A Probabilistic Perspective*, The MIT Press, 2012

Outline

Outline

- Preliminaries on Graphical Models
- Message passing and sum-product algorithm

Preliminaries on Graphical Models

Motivation: Curse of dimensionality

- Matroid enumeration
- Polyhedron computation
- Entropic Region
- Machine Learning: computing likelihoods, marginal probabilities $\cdots$

Graphical Models:

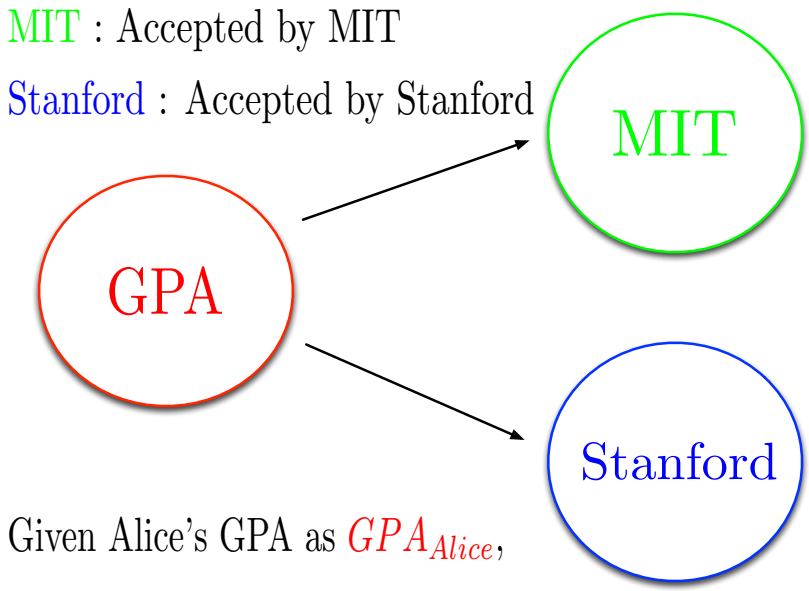
Help capturing complex dependencies among random variables, building large-scale statistical models and designing efficient algorithms for Inference.

Preliminaries on Graphical Models

Definition of Graphical Models:

A graphical model is a probabilistic model for which a graph denotes the conditional dependence structure between random variables.

Example:
Suppose MIT and Stanford accepted undergraduate students only based on GPA



Given Alice's GPA as GPA_{Alice},

$$\mathbb{P}(MIT|Stanford, GPA_{Alice}) = \mathbb{P}(MIT|GPA_{Alice})$$

We say MIT is conditionally independent of Stanford given GPA_{Alice}

Sometimes use symbol $(MIT \perp Stanford|GPA_{Alice})$

Bayesian networks: directed graphical model

Bayesian networks

A Bayesian network consists of a collection of probability distributions P over $\mathbf{x} = \{x_1, \cdots, x_K\}$ that factorize over a directed acyclic graph(DAG) in the following way:

$$p(\mathbf{x}) = p(x_1, \cdots, x_K) = \prod_{k \in K} p(x_k | pa_k)$$

where pa_k is the direct parents nodes of x_k.

Alias of Bayesian networks:

- probabilistic directed graphical model: via directed acyclic graph(DAG)
- belief networks
- causal networks: directed arrows represent causal realtions

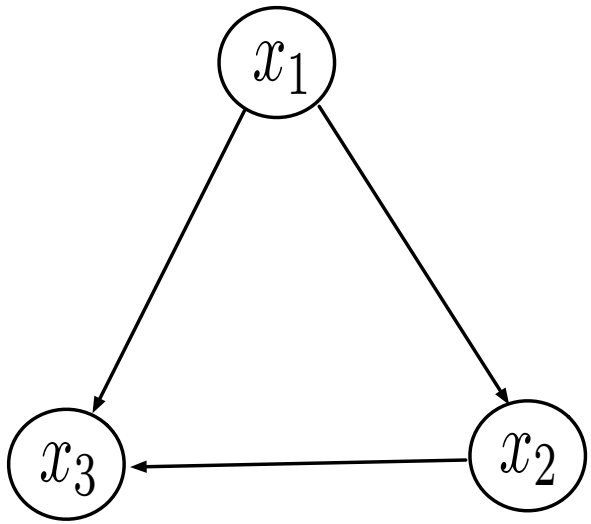
Bayesian networks: directed graphical model

Examples of Bayesian networks

Consider an arbitrary joint distribution $p(\mathbf{x}) = p(x_1, x_2, x_3)$ over three variables, we can write:

$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_3|x_1, x_2)p(x_1, x_2) \\ &= p(x_3|x_1, x_2)p(x_2|x_1)p(x_1) \end{aligned}$$

which can be expressed in the following directed graph:



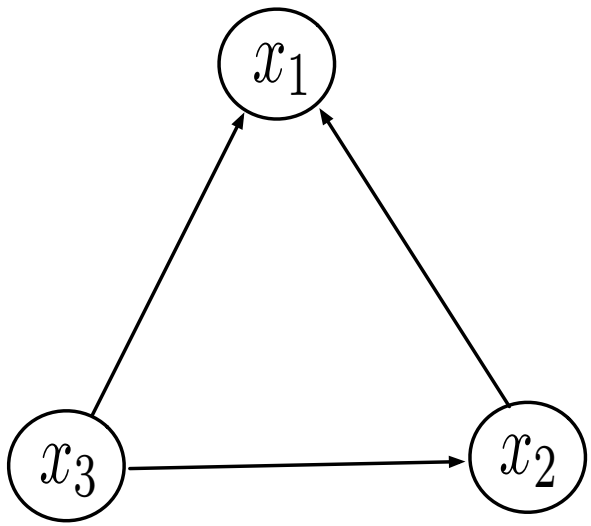
Bayesian networks: directed graphical model

Examples

Similarly, if we change the order of x_1, x_2 and x_3(same as consider all permutations of them), we can express $p(x_1, x_2, x_3)$ in five other different ways, for example:

$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_1|x_2, x_3)p(x_2, x_3) \\ &= p(x_1|x_2, x_3)p(x_2|x_3)p(x_3) \end{aligned}$$

which corresponding to the following directed graph:



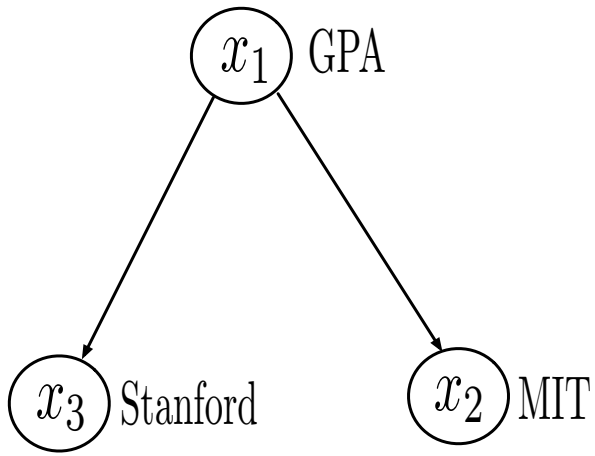
Bayesian networks: directed graphical model

Examples

Recall the previous example about how MIT and Stanford accept undergraduate students, if we assign x_1 to "GPA", x_2 to "accepted by MIT" and x_3 to "accepted by Stanford", then since $p(x_3|x_1, x_2) = p(x_3|x_1)$ we have

$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_3|x_1, x_2)p(x_2|x_1)p(x_1) \\ &= p(x_3|x_1)p(x_2|x_1)p(x_1) \end{aligned}$$

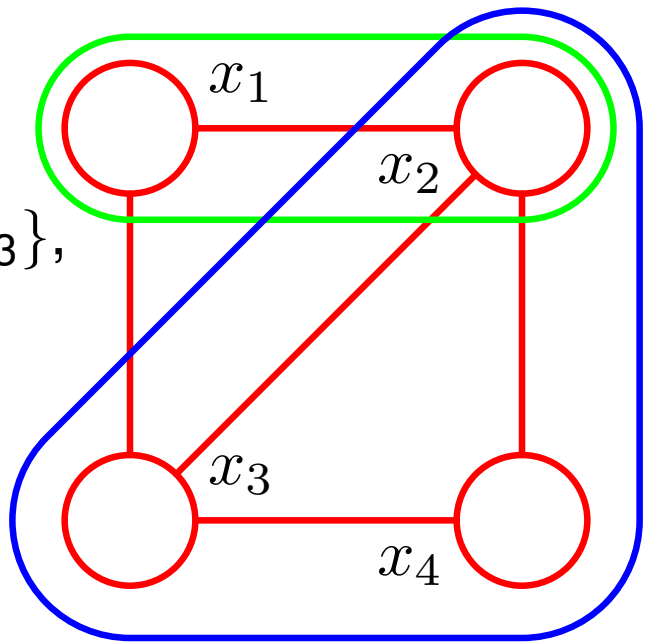
which corresponding to the following directed graph:



Markov random fields: undirected graphical model

In the undirected case, the probability distribution factorizes according to functions defined on the clique of the graph.
A clique is a subset of nodes in a graph such that there exist a link between all pairs of nodes in the subset.
A maximal clique is a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.

Example of cliques:
$\{x_1,x_2\},\{x_2,x_3\},\{x_3,x_4\},\{x_2,x_4\},\{x_1,x_3\},$
$\{x_1,x_2,x_3\}, \{x_2,x_3,x_4\}$
Maximal cliques:
$\{x_1,x_2,x_3\}, \{x_2,x_3,x_4\}$



Markov random fields: undirected graphical model

Markov random fields: Definition

Denote C as a clique, $\mathbf{x}_C$ the set of variables in clique C and $\psi_C(\mathbf{x}_C)$ a nonnegative potential function associated with clique C. Then a Markov random field is a collection of distributions that factorize as a product of potential functions $\psi_C(\mathbf{x}_C)$ over the **maximal cliques** of the graph:

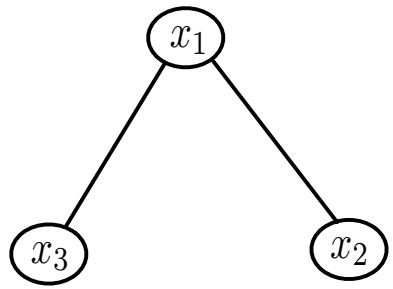
$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C)$$

where normalization constant $Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C)$ sometimes called the partition function.

Markov random fields: undirected graphical model

Factorization of undirected graphs

Question: how to write the joint distribution for this undirected graph?



$(2 \perp 3|1)$ hold

Answer:

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)$$

where $\psi_{12}(x_1, x_2)$ and $\psi_{13}(x_1, x_3)$ are the potential functions and Z is the partition function that make sure $p(\mathbf{x})$ satisfy the conditions to be a probability distribution.

Markov random fields: undirected graphical model

Markov property

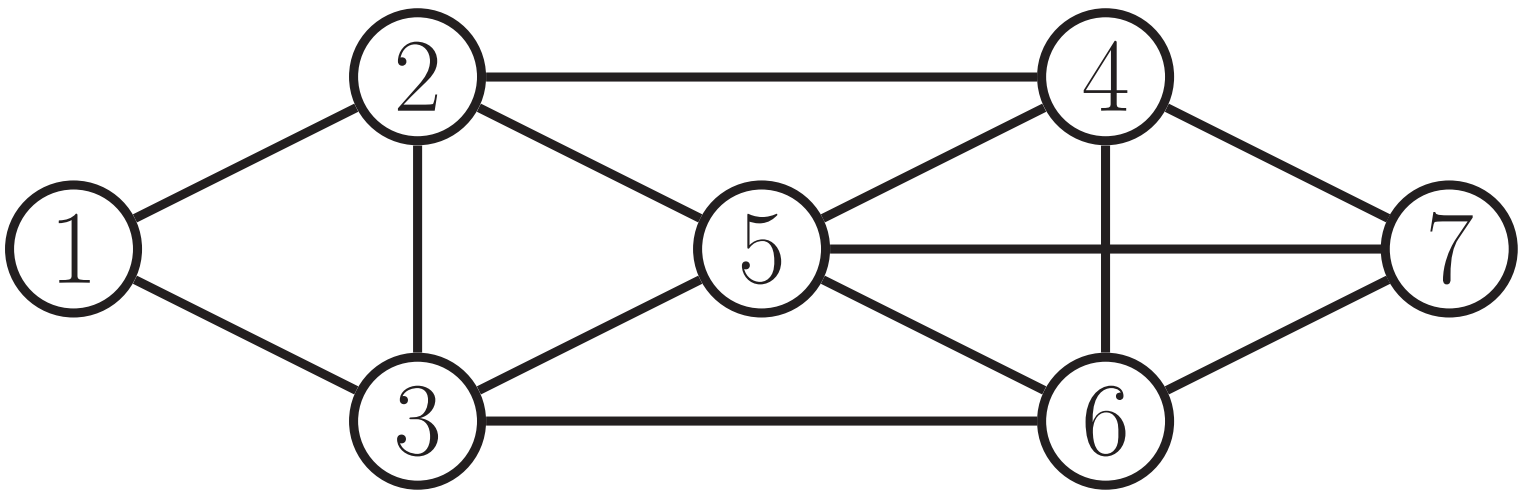
Given an undirected graph G = (V,E), a set of random variables $X = (X_a)_{a \in V}$ indexed by V , we have the following Markov properties:

- **Pairwise Markov property**: Any two non-adjacent variables are conditionally independent given all other variables: $X_a \perp X_b | X_{V \setminus \{a,b\}}$ if $\{a, b\} \notin E$
- **Local Markov property**: A variable is conditionally independent of all other variables given its neighbors: $X_a \perp X_{V \setminus \{nb(a) \cup a\}} | X_{nb(a)}$
  where nb(a) is the neighbors of node a.
- **Global Markov property**: Any two subsets of variables are conditionally independent given a separating subset: $X_A \perp X_B | X_S$, where every path from a node in A to a node in B passes through S(means when we remove all the nodes in S, there are no paths connecting any nodes in A to any nodes in B).

Markov random fields: undirected graphical model

Examples of Markov properties

Pairwise Markov property: $(1 \perp 7|23456)$, $(3 \perp 4|12567)$
Local Markov property: $(1 \perp 4567|23)$, $(4 \perp 13|2567)$
Global Markov property: $(1 \perp 67|345)$, $(12 \perp 67|345)$



Markov random fields: undirected graphical model

Relationship between different Markov properties and factorization property

(F): Factorization property; (G): Global Markov property;
(L): Local Markov property; (P):Pairwise Markov property

$$(F) \Rightarrow (G) \Rightarrow (L) \Rightarrow (P)$$

if assuming strictly positive $p(\cdot)$

$$(P) \Rightarrow (F)$$

which give us the Hammersley-Clifford theorem.

Markov random fields: undirected graphical model

The Hammersley-Clifford theorem(see Koller and Friedman 2009, p131 for proof)

Consider graph G, for strictly positive $p(\cdot)$, the following Markov property and Factorization property are equivalent:
Markov property: Any two subsets of variables are conditionally independent given a separating subset $(X_A, X_B|X_S)$ where every path from a node in A to a node in B passes through S.
Factorization property: The distribution p factorizes according to G if it can be expressed as a product of potential functions over maximal cliques.

Markov random fields: undirected graphical model

Example: Gaussian Markov random fields

A multivariate normal distribution forms a Markov random field w.r.t. a graph $G = (V, E)$ if the missing edges correspond to zeros on the concentration matrix(the inverse convariance matrix)

Consider $\mathbf{x} = \{x_1, \cdots, x_K\} \sim \mathcal{N}(0, \Sigma)$ with Σ regular and $K = \Sigma^{-1}$. The concentration matrix of the conditional distribution of (x_i, x_j) given $x_{\setminus\{i,j\}}$ is

$$K_{\{i,j\}} = \begin{pmatrix} k_{ii} & k_{ij} \\ k_{ji} & k_{jj} \end{pmatrix}$$

Hence

$$x_i \perp x_j | x_{\setminus ij} \Leftrightarrow k_{ij} = 0 \Leftrightarrow \{i, j\} \notin E$$

Markov random fields: undirected graphical model

Example: Gaussian Markov random fields

The joint distribution of gaussian markov random fields can be factorizes as:

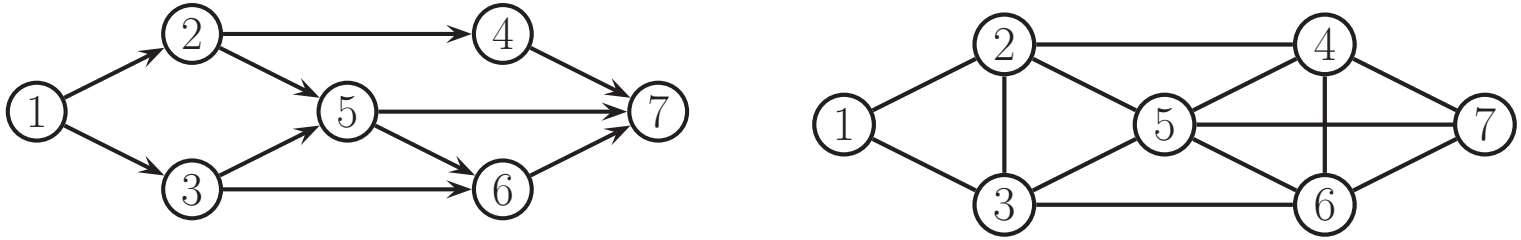
$$\log p(x) = const - \frac{1}{2}\sum_{i} k_{ii} x_i^2 - \sum_{\{i,j\}\in E} k_{ij} x_i x_j$$

The zero entries in K are called structural zeros since they represent the absent edges in the undirected graph.

Directed graph and Undirected graph

Moral graph: from directed graph to undirected graph

Moralization: ”marrying the parents”, so to speak, force the parents of a node to form a completed sub-graph.



(a) (b)

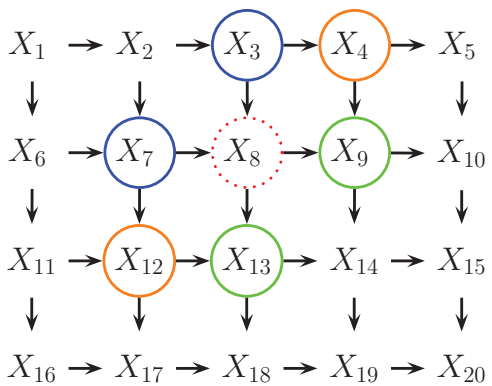
Motivation: Why not simply convert directed graph to undirected graph?
Works fine for Markov chain, not work for when there are nodes that have more than one parents(previous example).

Directed graph and Undirected graph

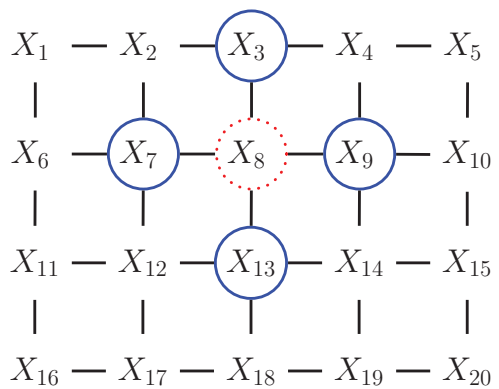
Markov blanket

In directed graph, a Markov blanket of a given node x_i is the set of nodes comprising the parents, the children and the co-parents(other parents of its children) of x_i.

In undirected graph, a Markov blanket of a given node x_i is the set of its neighboring nodes.



(a)



(b)

Directed graph and Undirected graph

Markov blanket In directed graph

Let $\mathbf{x} = \{x_1, \cdots, x_K\}$

$$\begin{aligned} p(x_i|x_{\setminus i}) &= \frac{p(x_1, \cdots, x_K)}{\sum_{x_i} p(x_1, \cdots, x_K)} \\ &= \frac{\prod_k p(x_k|pa_k)}{\sum_{x_i} \prod_k p(x_k|pa_k)} \\ &= f(x_i, mb(x_i)) \end{aligned}$$

where pa_k is the parents nodes of x_k, $mb(x_i)$ is the markov blanket of node x_i.

Example: $p(X_8|X_{\setminus 8}) = \frac{p(X_8|X_3,X_7)\ p(X_9|X_4,X_8)p(X_{13}|X_8,X_{12})}{\sum_{X_8} p(X_8|X_3,X_7)p(X_9|X_4,X_8)p(X_{13}|X_8,X_{12})}$

Directed graph and Undirected graph

Markov blanket In undirected graph

Let $\mathbf{x} = \{x_1, \cdots, x_K\}$,

$$p(x_i|x_{\setminus i}) = p(x_i|mb(x_i))$$

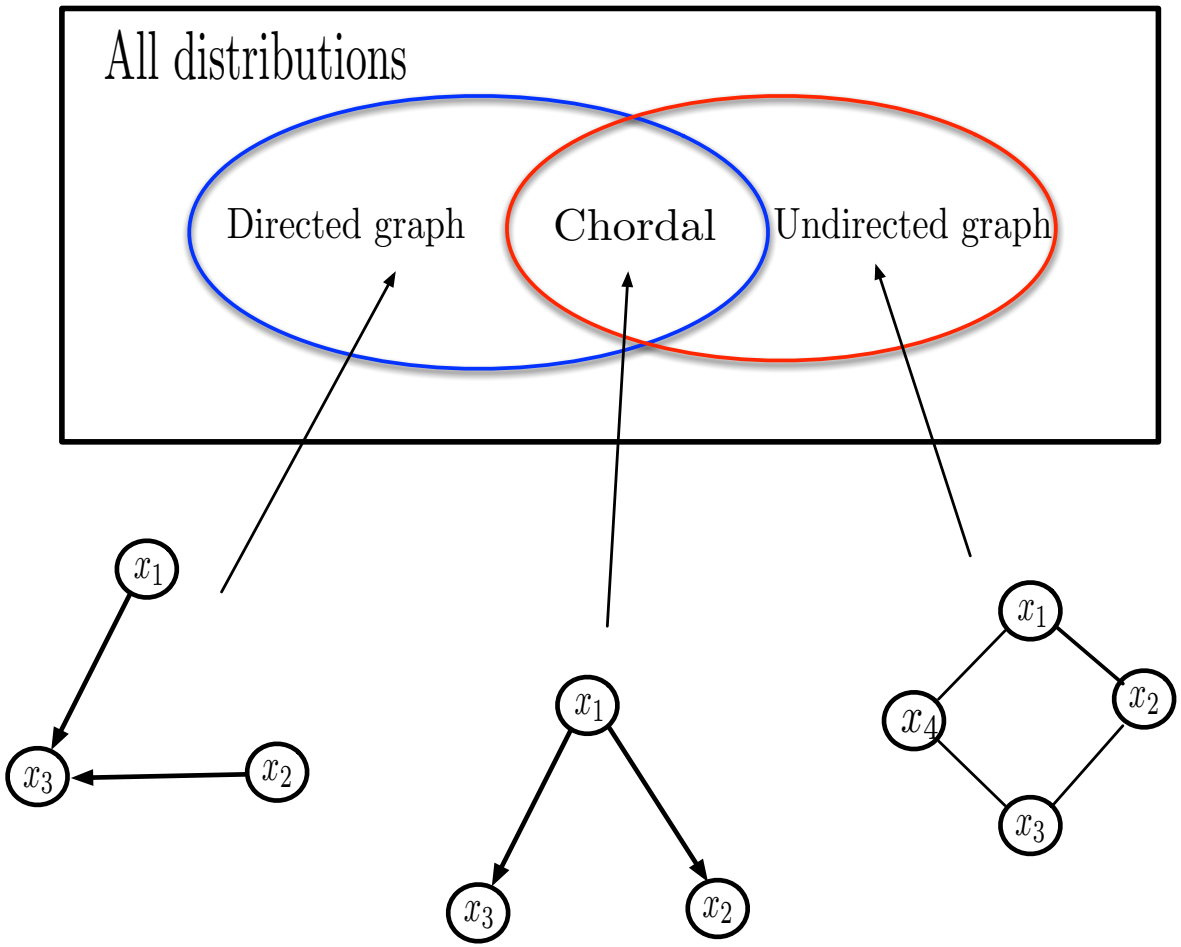
where $mb(x_i)$ is the markov blanket of node x_i.
Example: $p(X_8|X_{\setminus 8}) = p(X_8|X_3, X_7, X_9, X_{13})$

Notes

The Markov blanket of a directed graph is the Markov blanket(neighbours) of its moral graph.

Directed graph and Undirected graph

Directed graph VS. Undirected graph



What is Chordal?

Directed graph and Undirected graph

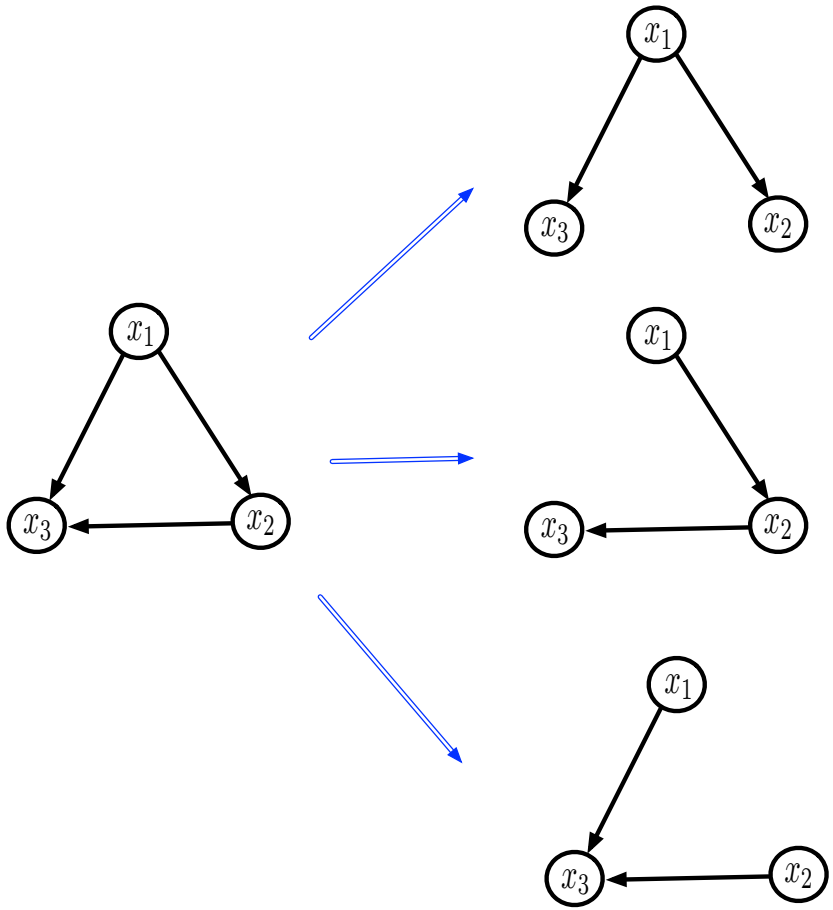
What is Chordal?

Chordal or decomposable: If we collapse together all the variables in each maximal clique to make the ”mega-variables”, the resulting graph will be a tree.

Why Chordal graphs are the intersection of directed graph and undirected graph?
Hint: *Moral graph*

Directed graph and Undirected graph

Directed Graph:Dropping edges



$(2 \perp 3|1)$

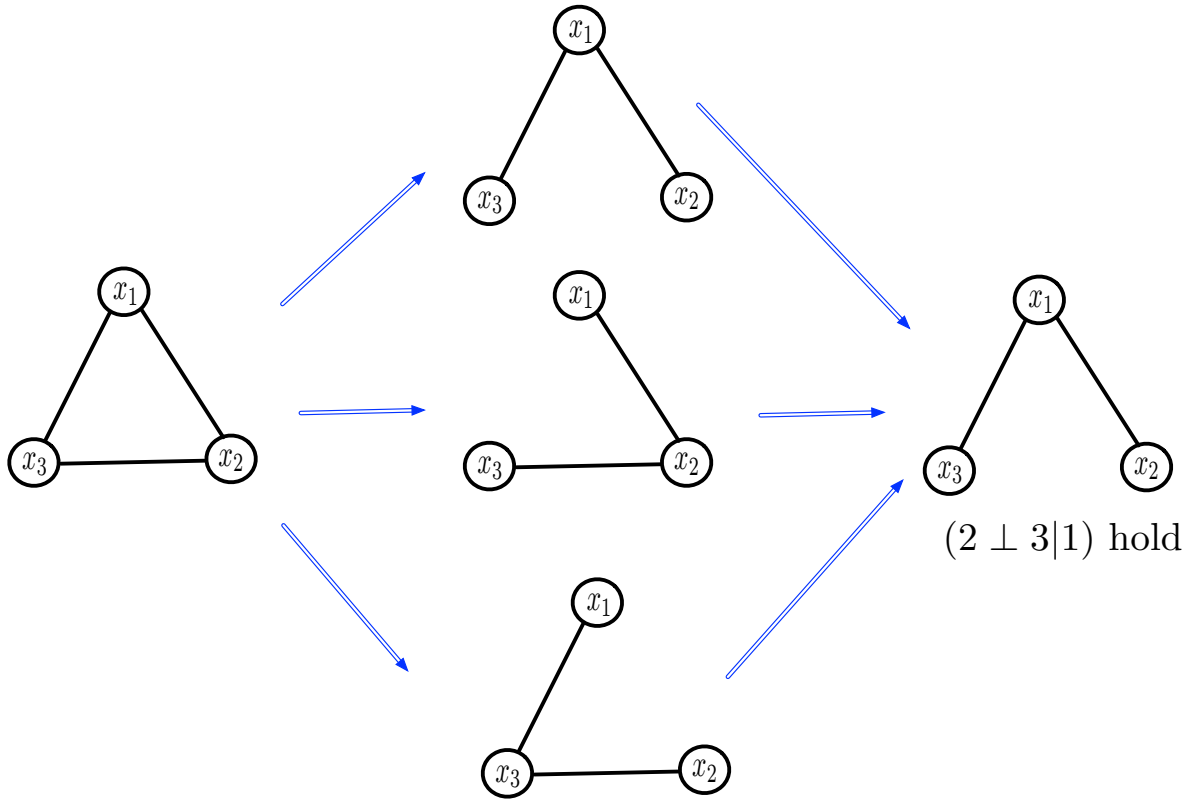
$(1 \perp 3|2)$

$(1 \perp 2|\emptyset)$

But not $(1 \perp 2|3)$

Directed graph and Undirected graph

Undirected Graphs: remove arrows



Directed graph and Undirected graph

Directed graph and Undirected graph

Example of 4 atom distribution:

			x		
x_4	x_3	x_2	x_1	$p(\mathbf{x})$	When $I(x_3; x_4) = 0$
0	0	0	0	a	$\alpha\beta$
0	1	0	1	b-a	$\alpha(1-\beta)$
1	0	0	1	c-a	$(1-\alpha)\beta$
1	1	1	1	1+a-b-c	$(1-\alpha)(1-\beta)$

Satisfy following Conditional independence relations:

$(1,2|3), (1,2|4)$
$(1,2|34)$, $(1|34)$, $(2|34)$
$(1|234)$, $(2|134)$, $(3|124)$, $(4|123)$
$(3,4|\emptyset)$

Outline

- Preliminaries on Graphical Models
- Message passing and sum-product algorithm

Factor graphs

Common thing about directed graphical model and undirected graphical model

Both allow global function to be expressed as product of factors over subsets of variables.

Factor graphs: Definition

A factor graph is a bipartite(bigraph) that expresses the structure of the factorization (1).

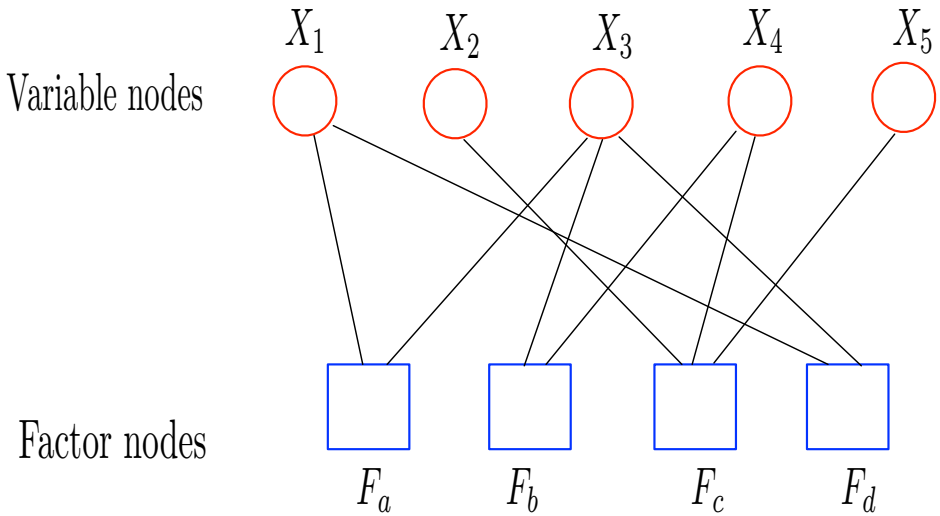
$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s) \tag{1}$$

where $\mathbf{x}_s$ denotes a subset of the variables $\mathbf{x}$.
In a factor graph, the random variables usually are round nodes and the factors are square nodes. There is an edge between the factor node $f_s(\mathbf{x}_s)$ and the variable node x_k if and only if $x_i \in \mathbf{x}_s$.

Factor graphs

Example of factor graph:

Factor graph: connection only between variable nodes and factor nodes.

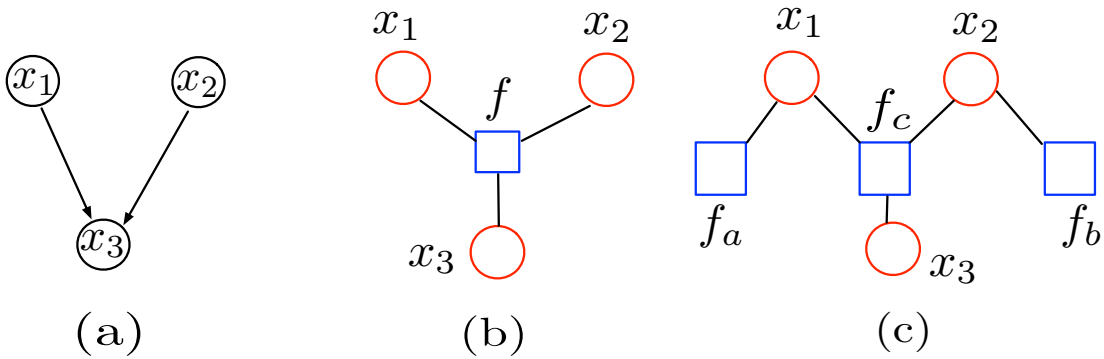


$$p(\mathbf{x}) = f_a(x_1, x_3) f_b(x_3, x_4) f_c(x_2, x_4, x_5) f_d(x_1, x_3)$$

Note: The same subset of variables $\mathbf{x}_s$ can be repeated multiple times, for example: $\mathbf{x}_s = \{x_1, x_3\}$ in the above factor graph.

Factor graphs

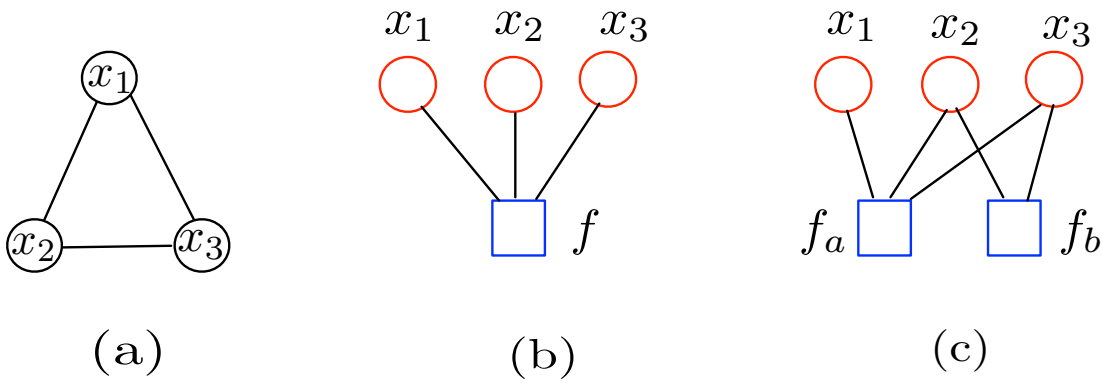
From directed graph to factor graph:



(a): An directed graph with factorization $p(x_1)p(x_2)p(x_3|x_1, x_2)$.
(b): A factor graph with factor
$f(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$.
(c): A different factor graph with factors $f_a(x_1)f_b(x_2)f_c(x_1, x_2, x_3)$ such that $f_a(x_1) = p(x_1)$, $f_b(x_2) = p(x_2)$ and
$f_c(x_1, x_2, x_3) = p(x_3|x_1, x_2)$.

Factor graphs

From undirected graph to factor graph: difference between $\frac{1}{Z}\prod_C \psi_C(\mathbf{x}_C)$ and $\prod_x f_s(\mathbf{x}_s)$



(a): An undirected graph with a single clique potential $\psi(x_1, x_2, x_3)$.
(b): A factor graph with factor $f(x_1, x_2, x_3) = \psi(x_1, x_2, x_3)$.
(c): A different factor graph with factors
$f_a(x_1, x_2, x_3) f_b(x_1, x_2) = \psi(x_1, x_2, x_3)$
Note: When clique is large, further factorize is usually very useful.

Factor graphs

Why introduce Factor graphs?

- Both directed graph and undirected graph can be converted to factor graphs.
- Factor graphs allows us to be more explicit about the details of the factorization.
- When connect factors to exponential families, the product of factors becomes sum of parameters under single exp.
- Multiple different factor graphs can represent the same directed or undirected graph, we can find factor graphs that are better for our inference.

Sum-product algorithms

Compute marginal probability $p(x_i)$

$$p(x_i) = \sum_{\mathbf{x}_{\setminus x_i}} p(\mathbf{x})$$

What if p($\mathbf{x}$) = $\frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C)$ or $\prod_x f_s(\mathbf{x}_s)$?

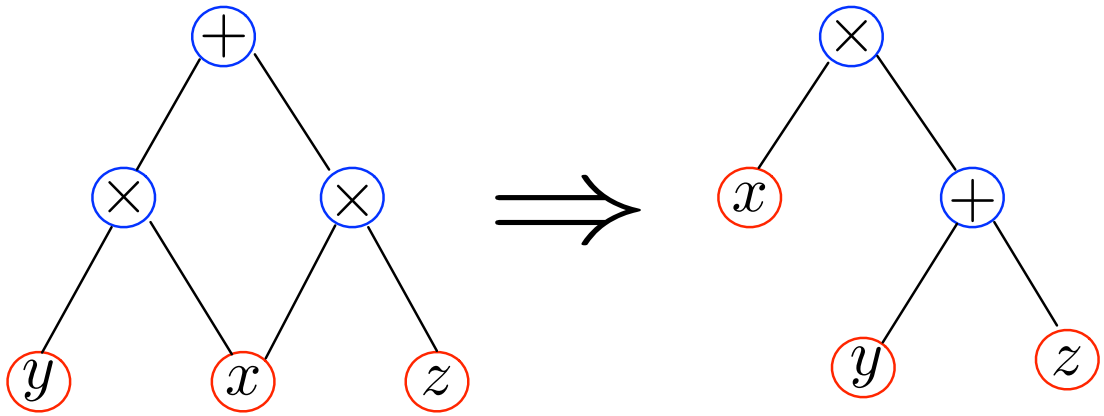
Can we design more efficient algorithms by exploring the conditional independence structure of the model?

Sum-product algorithms

Yes, we can.

The key idea of Sum-product algorithm

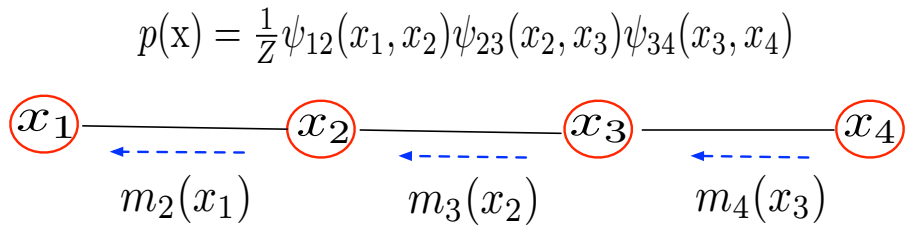
$$xy + xz = x(y + z)$$



Expression tree of xy + xz and x(y+z), where the leaf nodes represent variables and internal nodes represent arithmetic operators(addition, multiplication, negation, etc.).

Sum-product algorithms

Example of computing $p(x_1) = \sum_{\mathbf{x} \setminus x_1} p(\mathbf{x})$ on an undirected chain

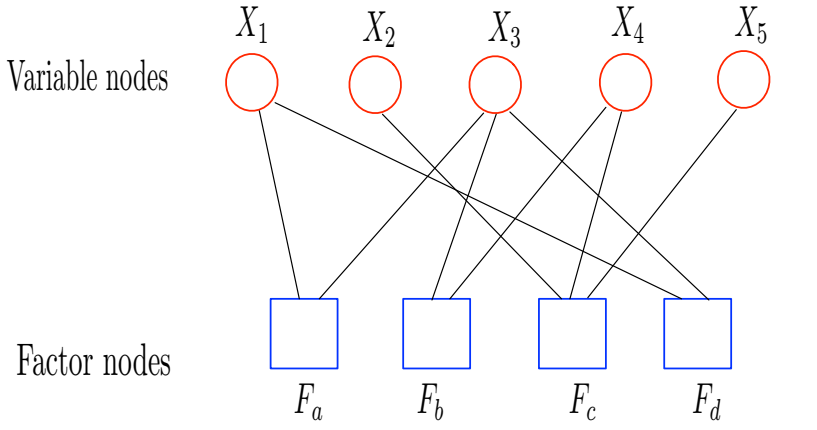


$$
\begin{aligned}
p(x_1) &= \frac{1}{Z} \sum_{x_2} \sum_{x_3} \sum_{x_4} \psi_{12}(x_1, x_2)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4) \\
&= \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{23}(x_2, x_3) \sum_{x_4} \psi_{34}(x_3, x_4) \\
&= \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{23}(x_2, x_3) m_4(x_3) \\
&= \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) m_3(x_2) = \frac{1}{Z} m_2(x_1)
\end{aligned}
$$

Sum-product algorithms

Example of computing $p(x_1) = \sum_{\mathbf{x}\setminus x_1} p(\mathbf{x})$

Factor graph: connection only between variable nodes and factor nodes.



$$p(\mathbf{x}) = f_a(x_1, x_3) f_b(x_3, x_4) f_c(x_2, x_4, x_5) f_d(x_1, x_3)$$

$$p(x_1) = \sum_{\mathbf{x}\setminus x_1} p(\mathbf{x}) = \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} p(\mathbf{x})$$

$$= \sum_{x_2, x_3} f_a(x_1, x_3) f_d(x_1, x_3) \sum_{x_4} f_b(x_3, x_4) \sum_{x_5} f_c(x_2, x_4, x_5)$$

Sum-product algorithms

Sum-product algorithm

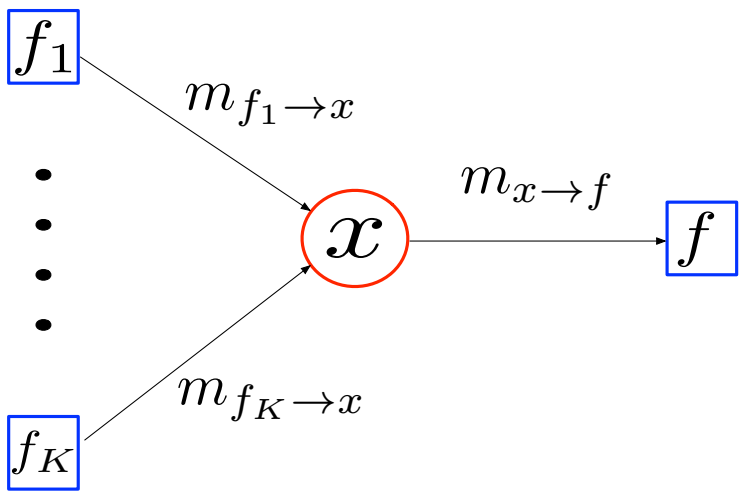
The sum-product algorithm involves using a message passing scheme to explore the conditional independence and use these properties to change the order of sum and product.

Sum-product algorithms

Message passing on a factor graph:

Variable x to factor f:

$$m_{x \to f} = \prod_{k \in nb(x) \setminus f} m_{f_k \to x}(x)$$



$nb(x)$ denote the neighboring factor nodes of x

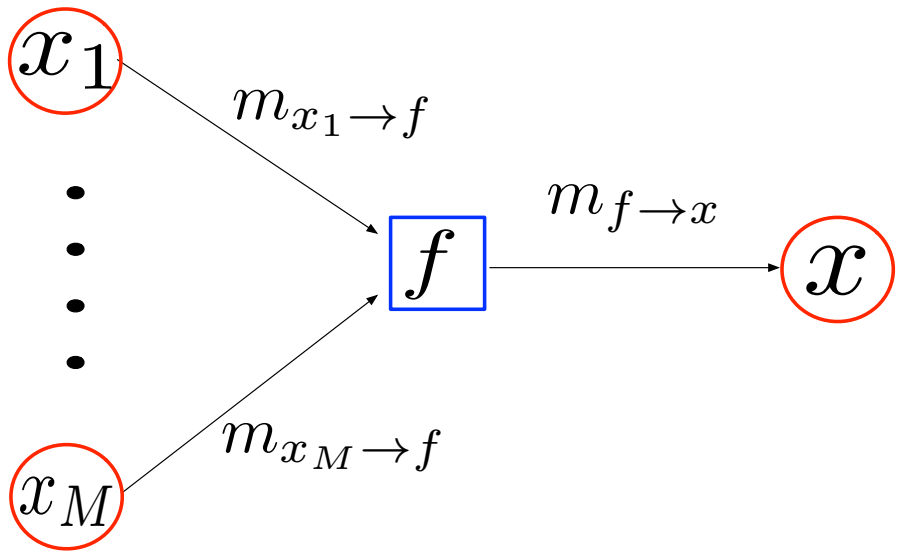
If $nb(x) \setminus f = \emptyset$ then $m_{x \to f} = 1$.

Sum-product algorithms

Message passing on a factor graph:

Factor f to variable x:

$$m_{f\to x} = \sum_{x_1,\cdots x_M} f(x, x_1, \cdots, x_M) \prod_{m\in nb(f)\backslash x} m_{x_m\to f}(x_m)$$

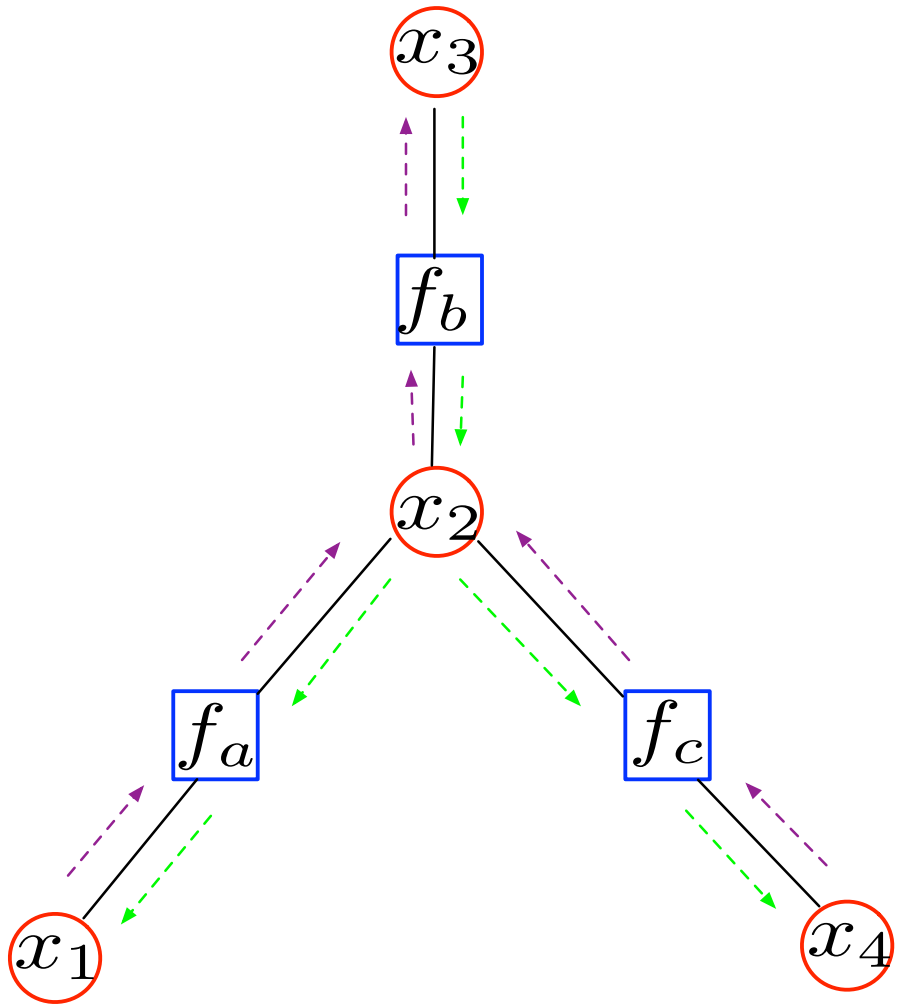


$nb(f)$ denote the neighboring variable nodes of f

If $nb(f)\backslash x = \emptyset$ then $m_{f\to x} = f(x)$.

Sum-product algorithms

Message passing on a tree structured factor graph



Sum-product algorithms

Message passing on a tree structured factor graph

Down - up:

$$
\begin{aligned}
m_{x_1 \to f_a}(x_1) &= 1 \\
m_{f_a \to x_2}(x_2) &= \sum_{x_1} f_a(x_1, x_2) \\
m_{x_4 \to f_c}(x_4) &= 1 \\
m_{f_c \to x_2}(x_2) &= \sum_{x_4} f_c(x_2, x_4) \\
m_{x_2 \to f_b}(x_2) &= m_{f_a \to x_2}(x_2) m_{f_c \to x_2}(x_2) \\
m_{f_b \to x_3}(x_3) &= \sum_{x_2} f_b(x_2, x_3) m_{x_2 \to f_b}
\end{aligned}
$$

Sum-product algorithms

Message passing on a tree structured factor graph

Up - down:

$$
\begin{aligned}
m_{x_3 \to f_b}(x_3) &= 1 \\
m_{f_b \to x_2}(x_2) &= \sum_{x_3} f_b(x_2, x_3) \\
m_{x_2 \to f_a}(x_2) &= m_{f_b \to x_2}(x_2) m_{f_c \to x_2}(x_2) \\
m_{f_a \to x_1}(x_1) &= \sum_{x_2} f_a(x_1, x_2) m_{x_2 \to f_a}(x_2) \\
m_{x_2 \to f_c}(x_2) &= m_{f_a \to x_2}(x_2) m_{f_b \to x_2}(x_2) \\
m_{f_c \to x_4}(x_4) &= \sum_{x_2} f_c(x_2 . x_4) m_{x_2 \to f_c}(x_2)
\end{aligned}
$$

Sum-product algorithms

Message passing on a tree structured factor graph

Calculating $p(x_2)$

$$
\begin{aligned}
p(x_2) &= \frac{1}{Z} m_{f_a \to x_2}(x_2) m_{f_b \to x_2}(x_2) m_{f_c \to x_2}(x_2) \\
&= \frac{1}{Z} \Big\{ \sum_{x_1} f_a(x_1, x_2) \Big\} \Big\{ \sum_{x_3} f_b(x_2, x_3) \Big\} \Big\{ \sum_{x_4} f_c(x_2, x_4) \Big\} \\
&= \frac{1}{Z} \sum_{x_1} \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \\
&= \sum_{x_1} \sum_{x_3} \sum_{x_4} p(\mathbf{x})
\end{aligned}
$$

Junction tree algorithm

Junction tree algorithm: message passing on general graphs

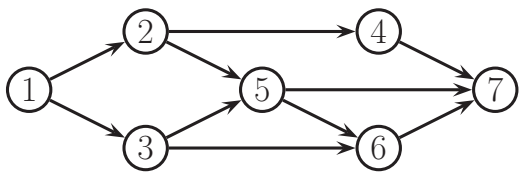
- 1. Moralize to a undirected graph if starting with a directed graph
- 2. Triangulate the graph: to make the graph chordal
- 3. Build the junction tree
- 4. Run message passing algorithm on the junction tree

Junction tree algorithm

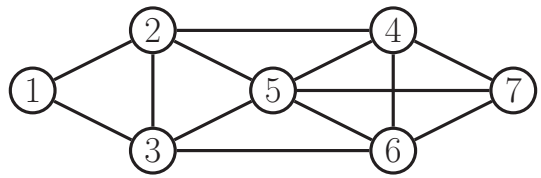
Junction tree algorithm

- 1. Moralize to a undirected graph if starting with a directed graph

Moralization: ”marrying the parents”, so to speak, force the parents of a node to form a completed sub-graph.



(a)



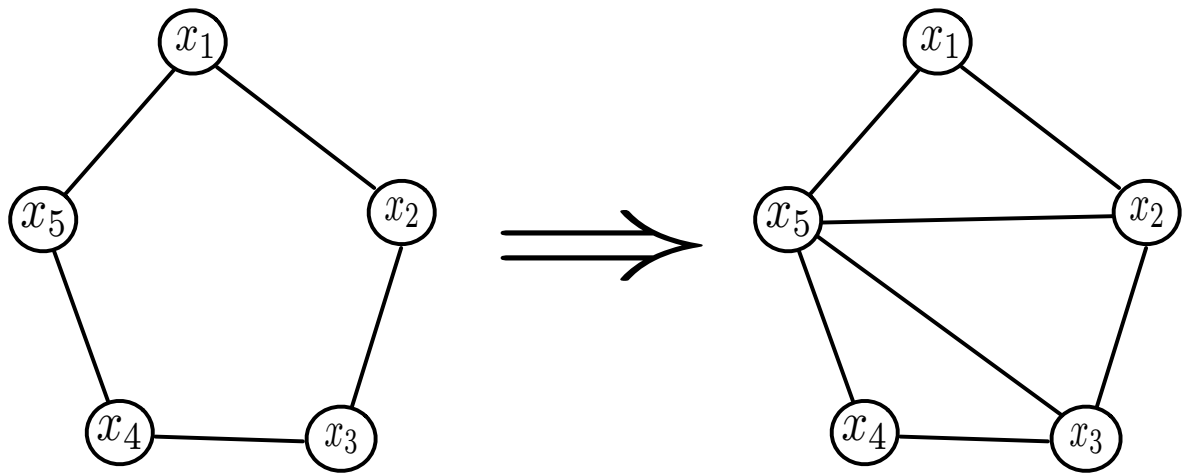
(b)

Junction tree algorithm

Junction tree algorithm

- 2. Triangulate the graph: to make the graph chordal

Finding chord-less cycles containing four or more nodes and adding extra links to eliminate such chord-less cycles.

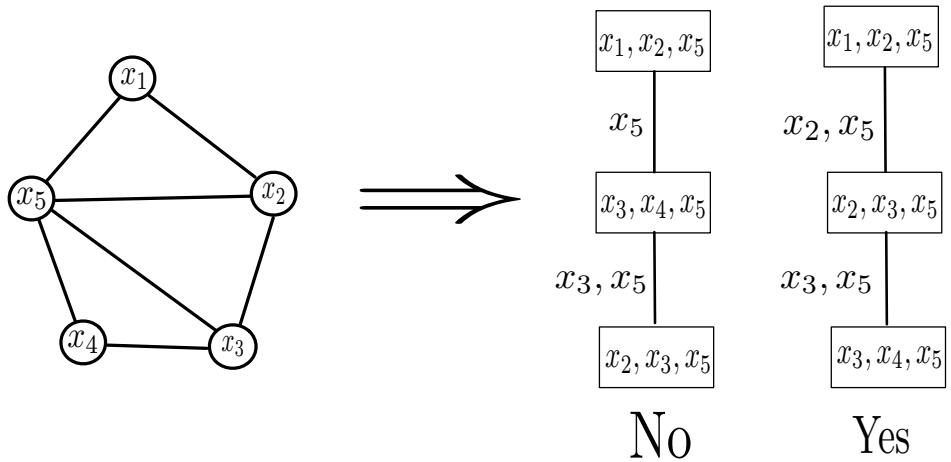


Junction tree algorithm

Junction tree algorithm

- 3. Build the junction tree

The nodes of a junction tree are the maximal cliques of the triangulated graph, The links connect pairs of nodes(maximal cliques) that have variables in common.
The junction tree is build so that the summation of the edge-weight in the tree is maximal, where the weight of an edge is the number of variables shared by the two maximal cliques it connected.



Junction tree algorithm

Junction tree algorithm

- 4. Run message passing algorithm on the junction tree

The message passing algorithm on the junction tree is similarly to the sum-product algorithm on a tree, the result give the joint distribution of each maximal clique of the triangulated graph.

The computational complexity of the junction tree algorithm is determined by the size of the largest clique, which is lower bounded by the treewidth of the graph,where treewidth is one less than the size of the largest clique.

Thanks!

Questions!