Graphical Models

Lecture 9

Graphical Models (GM)

Probability Theory:

- Plays a central role in ML
- Can be expressed in two simple ways via the sum rule and the product rule
- Is algebraically manipulatable
- Becomes highly advantages for the analysis using diagrammatic reorientations of probability distributions – *probabilistic graphical models*

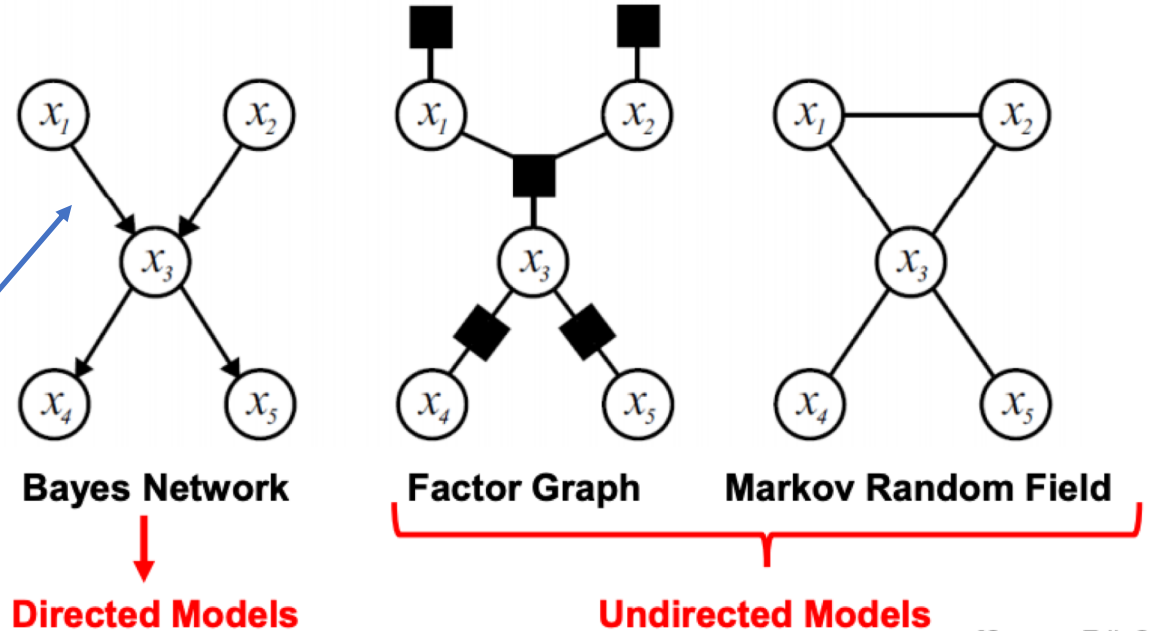
Probabilistic Graphical Model combines both probability and graph theory –

- Probabilities reason uncertainty
- Graphs model the dependency or correlation.
- Therefore, is useful to:
  1. To provide a simple visualization in structuring the model and in designing the new model.
  2. To insight the properties of model including conditional independence properties
  3. To express mathematical expressions for models requiring complex computations

GM

nodes (vertices): represents a random variable or a group of random variables

links (edges): Express probabilistic relationship between variables

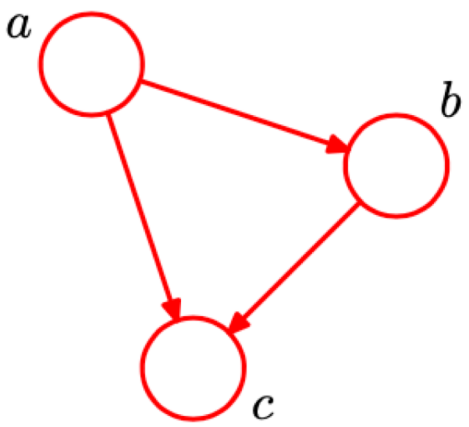


Directed graphical models – links of the graphs have a particular **directionality** indicated by arrows, (Bayesian networks (BN))

Undirected graphical models – links do not carry arrows (Markov random fields)

A variety of graphical models can represent the same probability distribution.

GM – Directed Graphical Models



Recall the probability chain rule – we can decompose any joint distribution as a product of conditionals

$$p(a,b,c) = p(c|a,b)p(b|a)p(a)$$

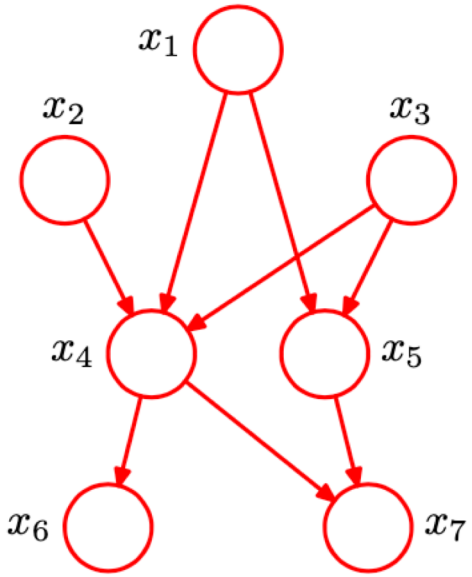
and valid for any ordering of the random variables (RVs)

$$p(a,b,c) = p(a)p(b|a)p(c|a,b)$$

For a collection of N RVs and any permutation ρ:

$$p(x_1,\dots,x_N) = p(x_{\rho(1)})\prod_{i=2}^{N} p(x_{\rho(i)}|x_{\rho(i-1)},\dots,x_{\rho(1)}) \quad (1)$$

GM – Directed Graphical Models



A graph $G(V, E)$ has a set of vertices V and a set of edges between RVs E.

A **directed graph** is a graph with edges $(s, t) \in E$ connecting parent vertex $s \in V$ to a child vertex $t \in V$.

Parents of vertex $t \in V$ are given by the set of nodes with edges pointing to t,

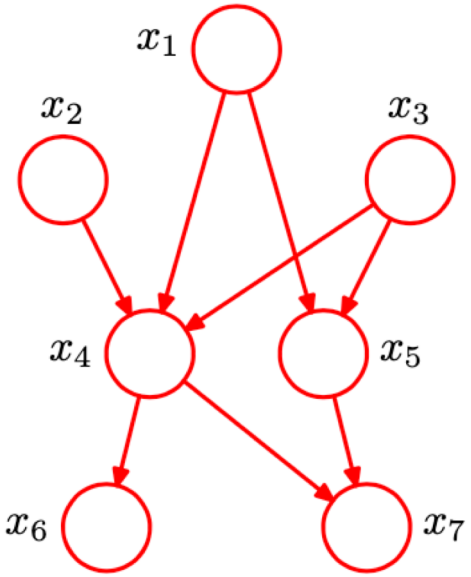
$$\mathrm{Pa}(t) = \{s : (s, t) \in E\}$$

Children of $t \in V$ are given by the set,

$$\mathrm{Ch}(t) = \{t : (t, k) \in E\}$$

Ancestors are parents of parents and **descendants** are children of children.

GM – Bayes Network



Model factors are normalized conditional distributions:

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{Pa}_k) \tag{2}$$

Directed acyclic graph (DAG) specifies factorized form of joint probability:

$$p(\boldsymbol{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

Note: Locally normalized factors yield **globally** normalized joint probability.

GM – Bayes Network

Consider the Bayesian polynomial regression model with

- The input data: $\boldsymbol{x} = (x_1, \ldots, x_N)^T$
- The RVs are the vector of coefficients: $\boldsymbol{w}$
- The observed data: $\boldsymbol{t} = (t_1, \ldots, t_N)^T$
- The noise variance: σ^2
- The hyperparameter for the precision of the Gaussian prior over $\boldsymbol{w}$: α

The joint distribution is

$$p(\boldsymbol{t}, \boldsymbol{w}) = p(\boldsymbol{w}) \prod_{n=1}^{N} p(t_n|\boldsymbol{w}) \quad (3)$$

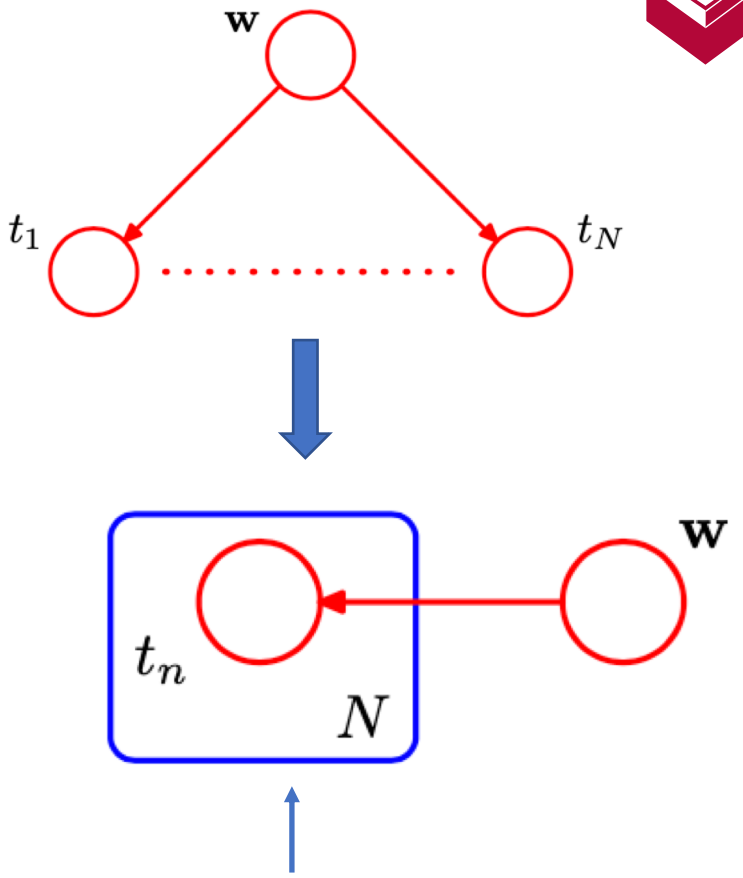
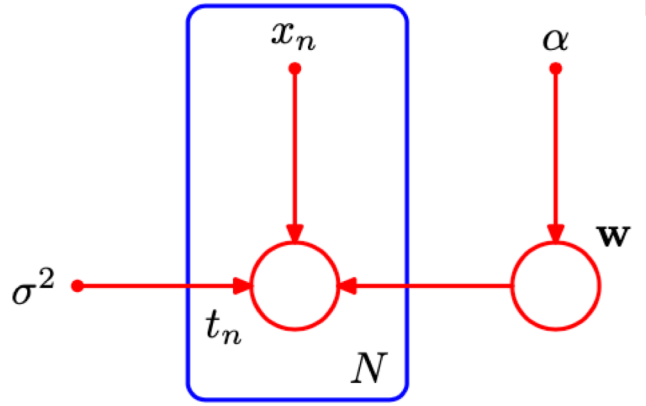


Plate – labelled with N indicating there are N nodes.

GM – Bayes Network

The explicit formation of Eq (3):

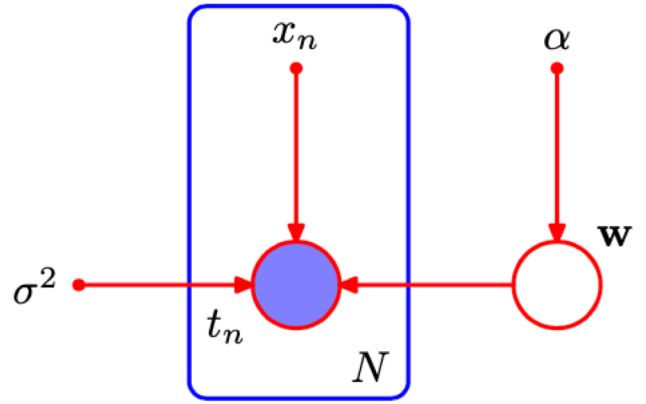
$$p(\boldsymbol{t}, \boldsymbol{w}|\boldsymbol{x}, \alpha, \sigma^2) = p(\boldsymbol{w}|\alpha) \prod_{n=1}^{N} p(t_n|\boldsymbol{w}, x_n, \sigma^2) \qquad (4)$$



When we apply a graphical model to a ML problem, we typically set some of the RVs to specific observed values.

In a GM, we denote such observed values by shading the corresponding nodes.

- $\boldsymbol{w}$ is not observed and is a latent variable (hidden variable)
- Observed values $\{t_n\}$ are the target values in the training set.



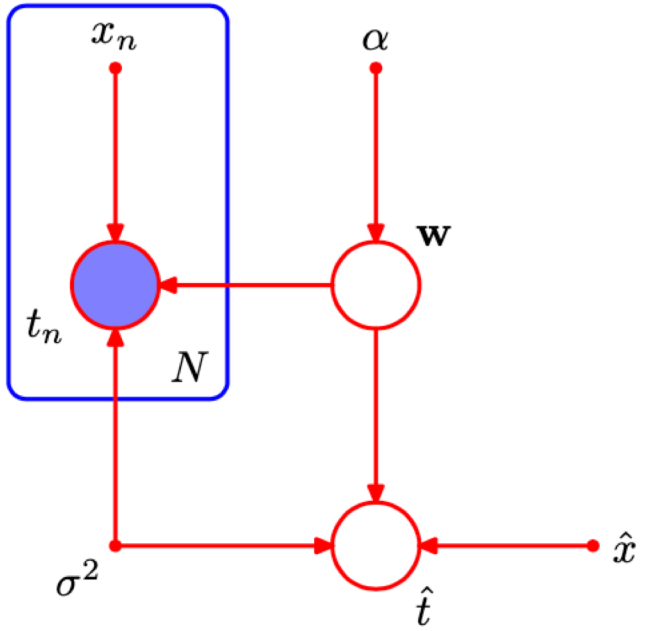
GM – Bayes Network

While w parameters are little direct interest, the ultimate goal is to make predictions for new inputs (test sets with $D(\hat{x}, \hat{t})$.

The joint distribution of all of the random variables for this model conditioned on the deterministic parameters is

$$p(\hat{t}, \boldsymbol{t}, \boldsymbol{w}|\hat{x}, \boldsymbol{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n, \boldsymbol{w}, \sigma^2)\right] p(\boldsymbol{w}|\alpha) p(\hat{t}|\hat{x}, \boldsymbol{w}, \sigma^2) \quad (5)$$

$$p(\hat{t}|\hat{x}, \boldsymbol{x}, \boldsymbol{t}, \alpha, \sigma^2) \propto \int p(\hat{t}, \boldsymbol{t}, \boldsymbol{w}|\hat{x}, \boldsymbol{x}, \alpha, \sigma^2) d\boldsymbol{w}$$



GM – Bayes Network

Linear Gaussian models as graphical models:

- An arbitrary DAG over D variables in which node i represents a single continuous random variable x_i having a GD.
- Then a linear combination of the state from its parent Pa_i is

$$p(x_i|\mathrm{Pa}_i) = \mathcal{N}\left(x_i \middle| \sum_{j\in\mathrm{Pa}_i} w_{ij}x_j + b_i, \sigma_i^2\right) \tag{6}$$

For all nodes,

$$\ln p(x) = \sum_{i=1}^{D} \ln p(x_i|\mathrm{Pa}_i) = -\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}\left(x_i - \sum_{j\in\mathrm{Pa_i}} w_{ij}x_j - b_i\right)^2 + \text{const} \tag{7}$$

GM – Bayes Network

The mean and covariance of the joint distribution can be determined recursively where each variable x_i has GD

$$x_i = \sum_{j \in \mathrm{Pa_i}} w_{ij} x_j + b_i + \sqrt{\sigma_i^2}\epsilon_i \tag{8}$$

where $\epsilon_i \sim \mathcal{N}(0,1)$ satisfying $\mathbb{E}(\epsilon_i) = 0$ & $\mathbb{E}(\epsilon_i, \epsilon_j) = I_{ij}$. Taking the expectation of Eq (8) is then

$$\mathbb{E}(x_i) = \sum_{j \in \mathrm{Pa_i}} w_{ij} \mathbb{E}(x_j) + b_i \tag{9}$$

and the covariance $\mathrm{cov}(x_i, x_j)$ is

$$cov(x_i, x_j) = \sum_{k \in \mathrm{Pa_j}} w_{jk} cov(x_i, x_k) + I_{ij}\sigma_j^2 \tag{10}$$

Extend the linear-Gaussian graphical model, the conditional distribution for node i can be expressed as

$$p(\boldsymbol{x_i}|\mathrm{Pa_i}) = \mathcal{N}\left(\boldsymbol{x_i}\middle|\sum_{j \in \mathrm{Pa}_i} \boldsymbol{W_{ij}}\boldsymbol{x_j} + \boldsymbol{b_i}, \boldsymbol{\Sigma_i}\right) \tag{11}$$

GM – Conditional Independence

Consider three variables a, b, and c, and suppose that the conditional distribution of a does not depend on value of b, so that

$$p(a|b,c) = p(a|c) \tag{12}$$

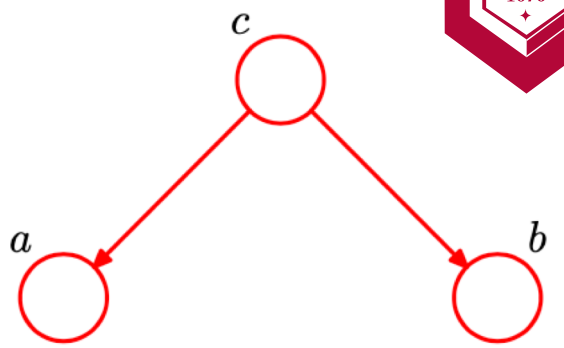
The join distribution of a and b conditioned on c is

$$p(a,b|c) = p(a|b,c)p(b|c) = p(a|c)p(b|c). \tag{13}$$

We say that **a is conditionally independent of b give c** and for the shorthand notation, we denote

$$a \perp b|c. \tag{14}$$

GM – Conditional Independence (tail-to-tail)



If two RVs a and b are conditionally independent give c, $a \perp b|c$, the DAG looks as the right graph and the corresponding joint distribution is

$$p(a,b,c) = p(a|c)p(b|c)p(c) \tag{15}$$

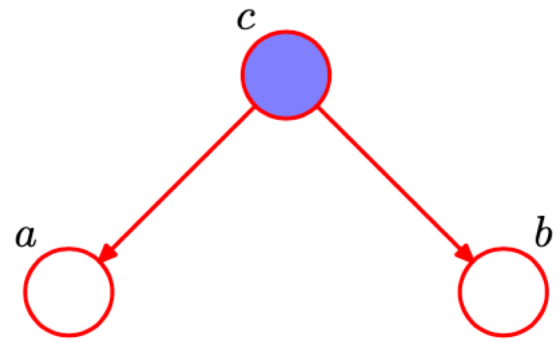
If none of the variables are observed, then we can test whether a and b are independent by marginalizing both side with respect to c:

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c) \tag{16}$$

and because this does not factorize into the product of $p(a)p(b)$, and so

$$a \not\perp b|\emptyset \tag{17}$$

meaning that the conditional independence property does not hold in general.



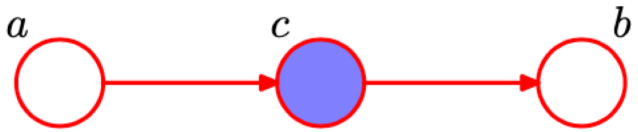
If we condition on the variable c, then

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = p(a|c)p(b|c) \tag{18}$$

and we obtain the condition independence property

$$a \perp b|c. \tag{19}$$

GM – Conditional Independence (head-to-tail)



The graph represents $p(a, b|c)$,

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} \quad (20)$$

Using Bayes' theorem, $p(c|a) = \frac{p(a|c)p(c)}{p(a)}$ and

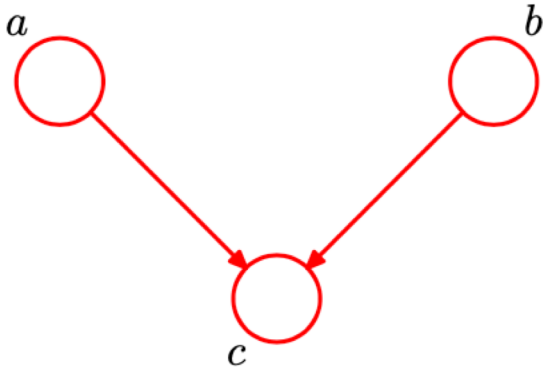
$$p(a, b|c) = \frac{p(a)p(a|c)p(c)p(b|c)}{p(a)p(c)} = p(a|c)p(b|c) \quad (21)$$

With conditioning c, we have independence!

$$a \perp b|c$$

If we do not condition c, the independence is the same as tail-to-tail case.

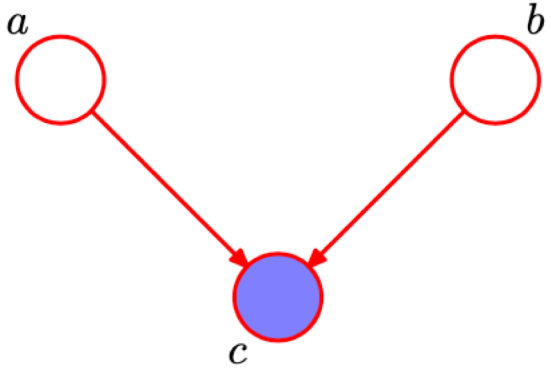
GM – Conditional Independence (head-to-head)



In this case,

$$p(a,b) = \sum_c p(a)p(b)p(c|a,b) = p(a)p(b) \tag{22}$$

$$\therefore a \perp b$$



In this case,

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \tag{23}$$

$$\neq p(a|c)p(b|c)$$

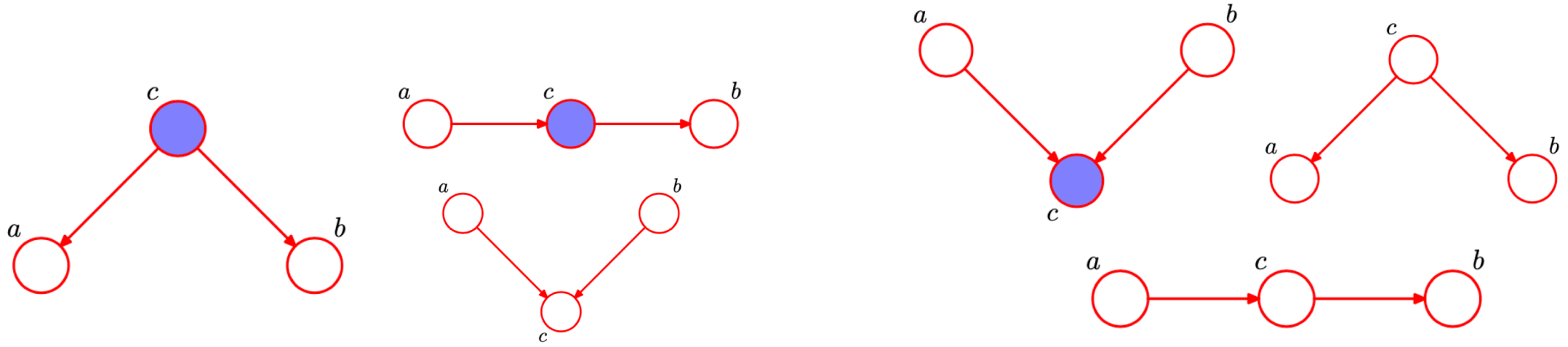
GM – Conditional Independence (d-separation)

- We have seen the analytical test of conditional independent via three cases.
- We also can examine directly reading from the graph without any performance of analytical manipulation via *d-separation.*

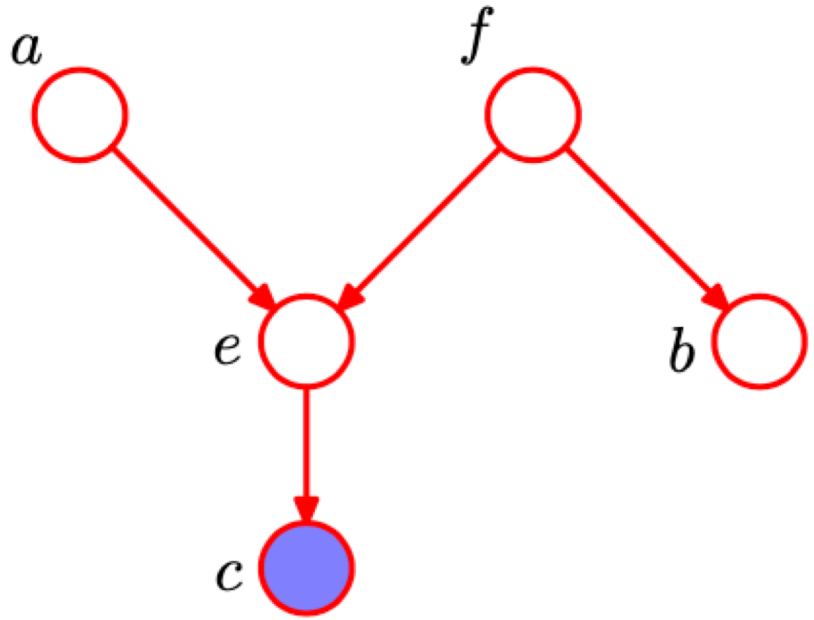
Consider a general directed graph, A, B, and C.
To ascertain that $A \perp B|C$, we check if all nodes from A to B are *blocked* – if it includes a node such that either

1. The arrows on the path meet either head-to-tail or tail-to-tail at the node in C
2. The arrows meet head-to-head at node and neither the node nor an of its descendants is in C.

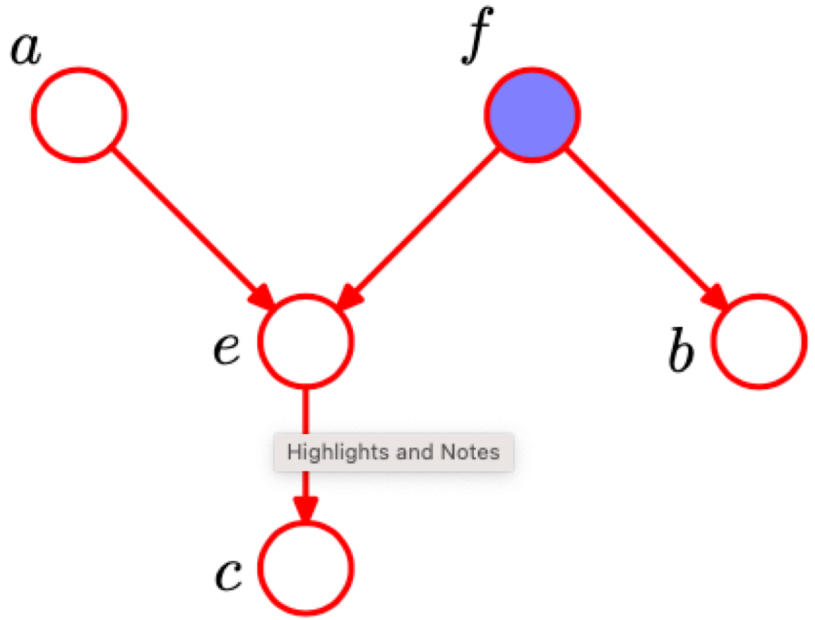
If all paths are blocked, then we say A is **d-separated** from B by C and the joint distribution over all of the variables in the graph will satisfy $A \perp B|C$.



GM – Conditional Independence (d-separation)



(a)



(b)

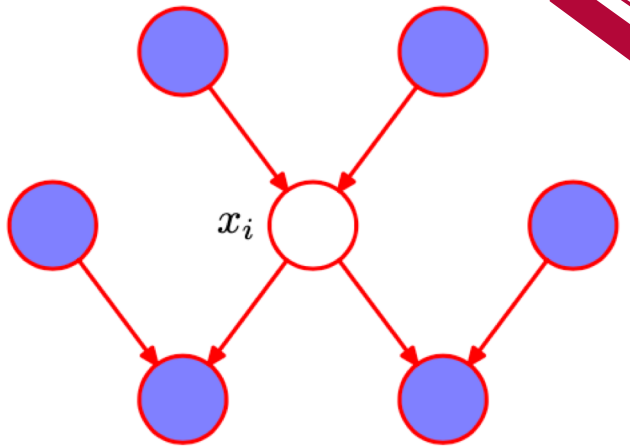
GM – Conditional Independence (Markov blanket)

A RV x_i with distribution $p(x_i)$ that is Markov w.r.t. graph $G = (V, E)$ has a **Markov blanket** given by:

$$Mb(x_i) = Pa(x) \cup Ch(x_i) \cup CoPa(x_i) \tag{23}$$

Mb of x_i comprises the set of parents, children, and co-parents of the node.

It has the property that the conditional distribution of x_i, conditioned on all the remaining variables in G, is dependent only on the variables in the Mb.



Why co-parents?

The observation of co-parent is required since the observations of the child nodes will not block paths to the co-parents.

Undirected Graph Models

A probability distribution over RVs $x = (x_1, \dots, x_d)$ can be written as a product of factors,

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c) \tag{24}$$

Where:

- $\mathcal{C}$ a collection of subsets of indices $\{1, \dots, d\}$
- $\psi(\cdot)$ nonnegative factors (or potential functions)
- Z the normalizer (or partition function)

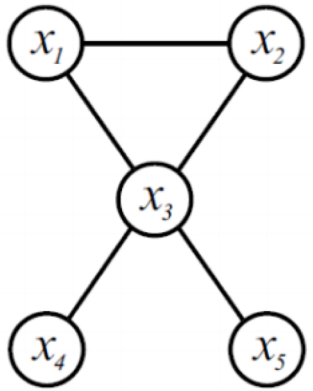
$$Z = \int \prod_{c \in \mathcal{C}} \psi_c(x_c)\, dx_c \tag{25}$$

Undirected graph $G = G(V, E)$ is a set of vertices and edges where edges are are specified irrespective of node ordering,

$$(s, t) \in E \Leftrightarrow (t, s) \in E$$

And distributions are typically specified with unknown normalizer,

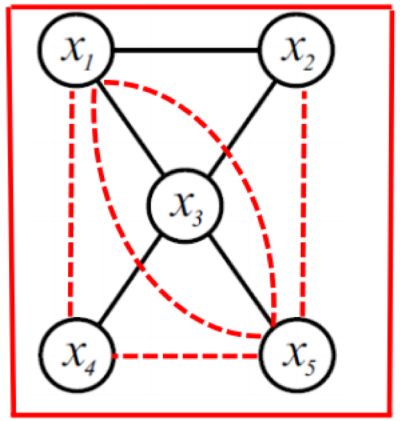
$$p(x) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c)\,. \tag{26}$$



Undirected Graphical Models – Markov Random Fields (MRFs)

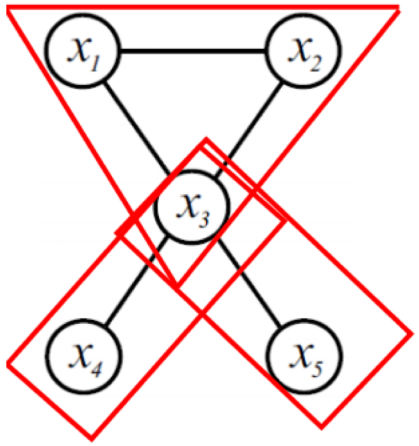
A factor $\psi_c(x_c)$ corresponds to a clique $c \in \mathcal{C}$ (fully connected subgraph) in the MRF.
An MRF does not imply a unique factorization:

Complete Graph



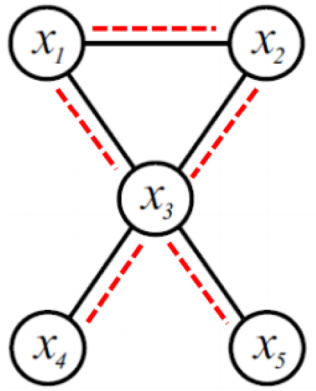
$\psi(x_1, \dots, x_5)$

Clique



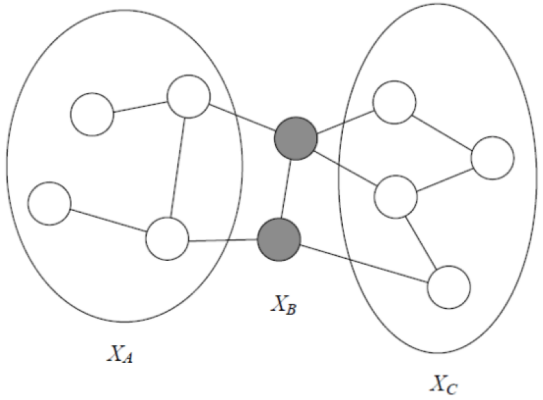
$\psi(x_1, x_2, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$

Pairwise Cliques



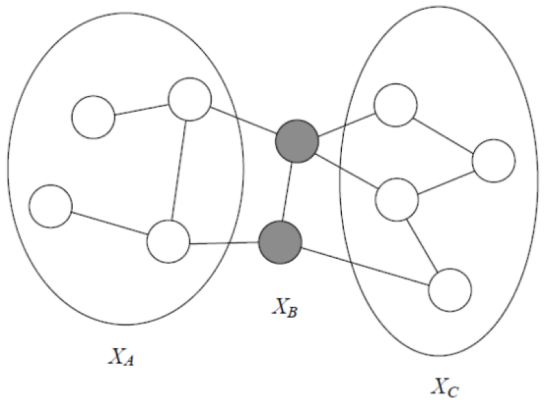
$\psi(x_1, x_2)\psi(x_2, x_3)\psi(x_1, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$

The factorization is valid if it satisfies the **Global Markov property**, defined by conditional independencies:



$p(x)$ is **globally Markov** w.r.t. G if $x_A \perp x_C | x_B$ in any separating set of G.

Undirected Graphical Models – Markov Random Fields (MRFs)



Global Markov Property:

- Set B separates A from C if all paths from A to C pass through B.
- By definition, distribution is Markov i.f.f. for any B separating A and C:
  - $p(x_A, x_C|x_B) = p(x_A|x_B)p(x_C|x_B)$
  - $p(x_A|x_B, x_C) = p(x_A|x_B)$
  - $p(x_C|x_B, x_A) = p(x_C|x_B)$

(26)

Local Markov Property:

- Given its neighbors, each node is independent of all other variables
  - $p(x_s|x_{V\setminus s}) = p(x_s|x_{\Gamma(s)})$
  - Where $\Gamma(s) = \{t \in V|(s,t) \in E\}$

(27)

- This local Markov property is a special case of the global Markov property

Hammersley-Clifford Theorem:

Let $\mathcal{C}$ denote the set of cliques of an undirected graph $\mathcal{G}$. A probability distribution defined as a normalized product of non-negative potential functions on those cliques is then always Markov with r.t. $\mathcal{G}$.

$$p(x) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c) \tag{28}$$

Conversely, any strictly positive density which is Markov with respect to $\mathcal{G}$ can be represented in this factored form.

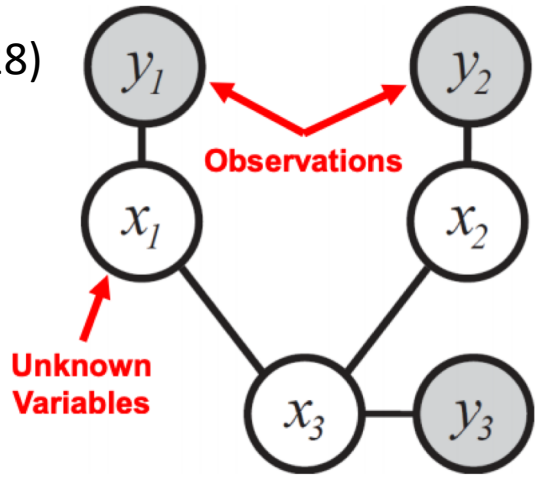
Undirected Graphical Models – Markov Random Fields (MRFs)

Often easier to specify and do inference on pairwise model:

$$p(x, y) \propto \prod_{s \in V} \psi_s(x_s, y) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t) \quad (28)$$

likelihood

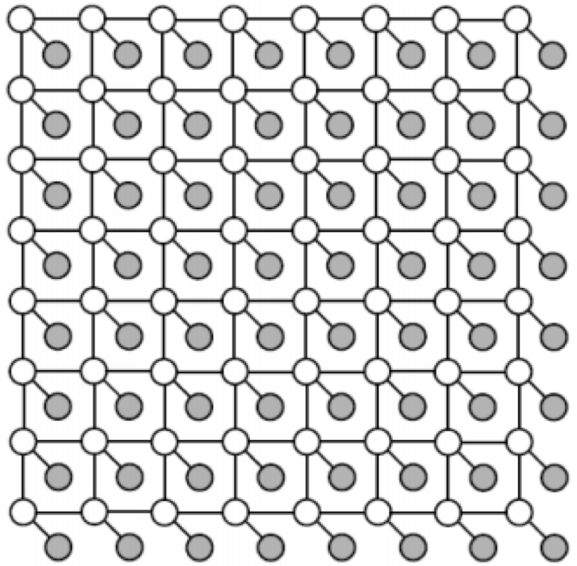
prior



Restricted class of MRFs:

- 2-node factor exists for every edge
- Explicit factorization of joint distribution
- High-order factors not always easily decomposed into pairwise terms

Undirected Graphical Models – Example



MRF for image:

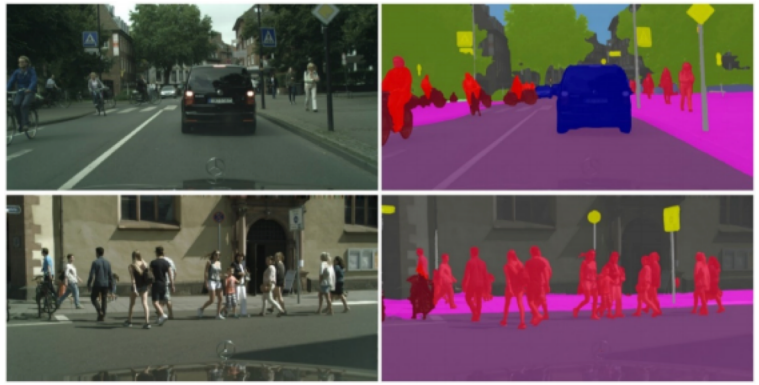
- x_i a binary variable denoting the state of pixel i in the unknown segmented noise-free image
- y_i the corresponding value of pixel i in the observed image

Pairwise MRF energy:

$$-\log p(x,y) = \log Z + \sum_i \psi_i(x_i, y_i) + \sum_{i,j} \psi_{ij}(x_i, x_j)$$

Low energy configurations = High probability

MAP (minimum energy) configuration =Piecewise constant regions



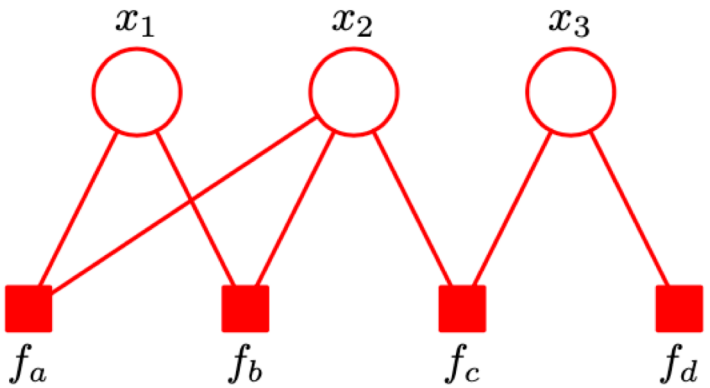
Factor Graphs

A hypergraph $\mathcal{H} = (V, F)$ where a hyperedge $f \in F$ is a subset of vertices $f \subset V$.

Factor node for each product term in the joint factorization:

$$p(\boldsymbol{x}) \propto \prod_s f_s(\boldsymbol{x_s})$$

where $\boldsymbol{x_s}$is the subset of variables in factor f_s. For example:



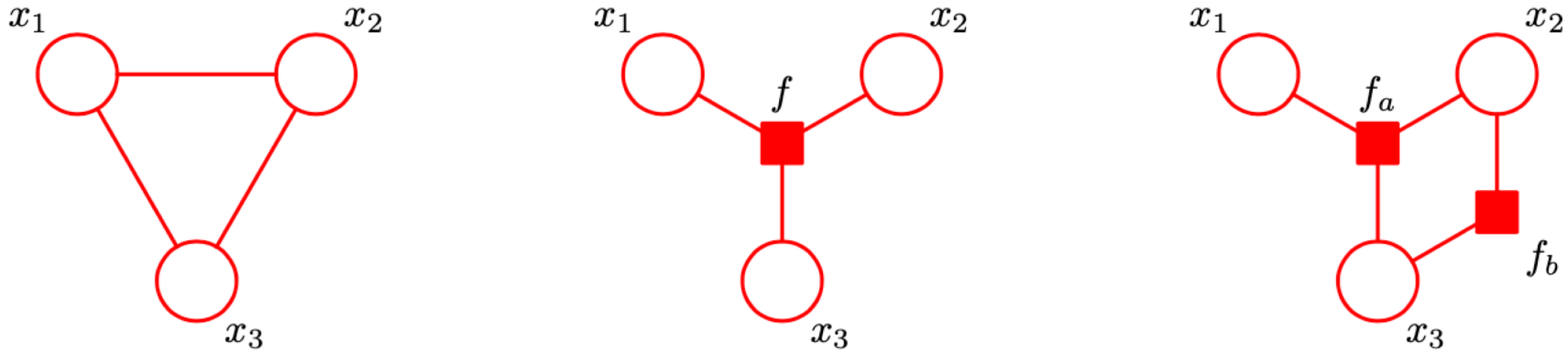
$$p(\boldsymbol{x}) = f_a(x_1, x_2) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_3)$$

Can be lumped together into the same clique potential $\psi(\cdot)$ in undirected graph.

Factor Graphs

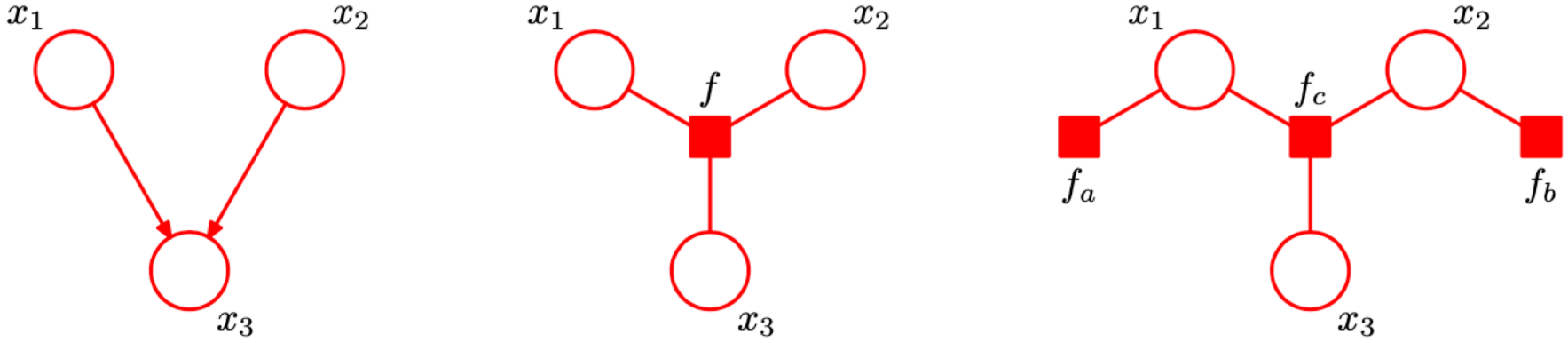
When a graph is converted into a factor graph, the result becomes a tree.

Undirected Graph



$$\psi(x_1, x_2, x_3) = f(x_1, x_2, x_3) = f_a(x_1, x_2, x_3) f_b(x_2, x_3)$$

Directed Graph



$$p(x_1)p(x_2)p(x_3|x_1, x_2) = f(x_1, x_2, x_3) = f_a(x_1) f_b(x_2) f_c(x_3, x_1, x_2)$$

Graphical Model - Evaluation

- **Sum-product**: to evaluate local marginal over nodes or subsets of nodes
- **Max-sum**: to find a set of variables with the largest probabilities and to find the value of that set.

Convert the original graph into a factor graph:

1. Can deal with both directed and undirected graphs in the same framework.
2. Can exploit the structure of graph effectively
    - Can obtain exact inference algorithm for marginal findings
    - Can share the commutations when there are several marginals.

Graphical Model – sum-product

Finding the marginal $p(x)$ for a particular variable node x:

- The sum of joint distribution over all variables except x:

$$p(x) = \sum_{\boldsymbol{x} \setminus x} p(\boldsymbol{x}) \tag{29}$$

where $\boldsymbol{x} \setminus x$ the set of variables in $\boldsymbol{x}$ with x omitted.

- Partition the factors in the joint distribution into groups in which each group associated with each of factor nodes that is a neighbor of the variable x:

$$p(\boldsymbol{x}) = \prod_{s \in ne(x)} F_s(x, X_s) \tag{30}$$

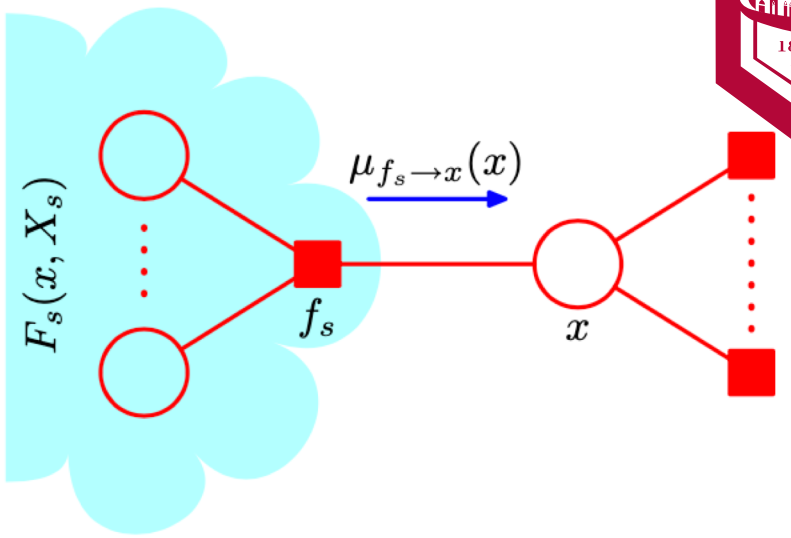
where:

- $ne(x)$: the set of neighbor factor nodes of x
- X_s the set of all variables in the subtree connected to x via f_s
- $F_s(x, X_s)$: the product of all the factors in the group associated with f_s

- The message from f_s to x:

$$\mu_{f_s \to x}(x) = \sum_{X_s} F_s(x, X_s) \tag{31}$$

$$p(x) = \prod_{s \in ne(x)} \left[\sum_{X_s} F_s(x, X_s) \right] = \prod_{s \in ne(x)} \mu_{f_s \to x}(x) \tag{32}$$



Graphical Model – sum-product

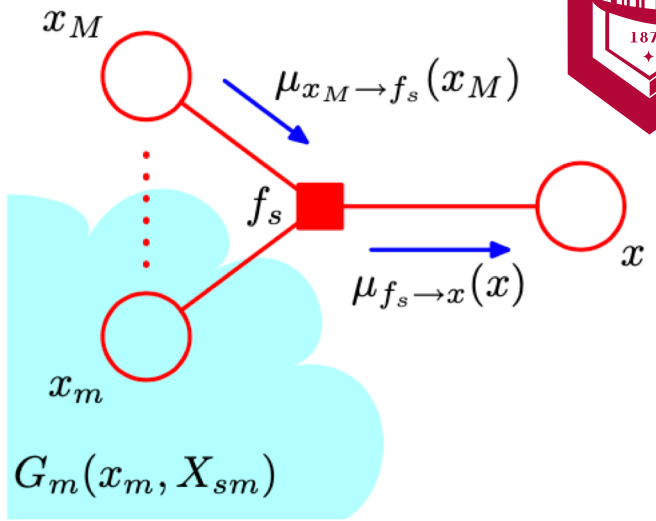
Each factor $F_s(x, X_s)$ can be described by a factor of sub-graph G_m with a variable x_m:

$$F_s(x, X_s) = f_s(x, x_1, \dots, x_M) G_1(x_1, X_{s1}) \cdots G_M(x_M, X_{sM})$$

The message $\mu_{f_s \to x}(x)$ becomes

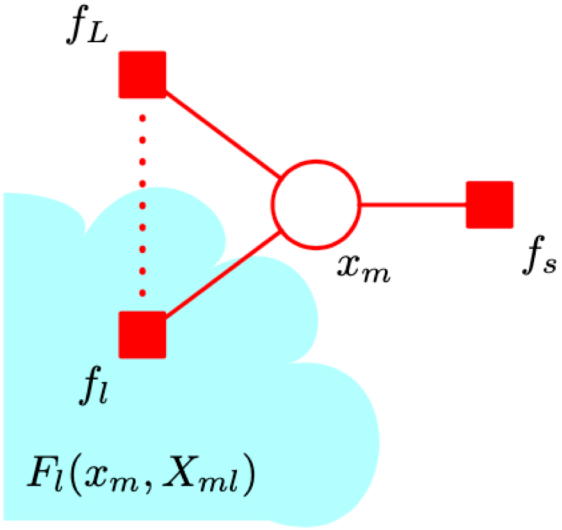
$$\mu_{f_s \to x}(x) = \sum_{x_1} \cdots \sum_{x_M} f(x, x_1, \dots, x_M) \prod_{m \in ne(f_s) \setminus x} \left[\sum_{X_{sm}} G_m(x_m, X_{sm}) \right] \quad (32)$$

$$\mu\_(x_m \to f_s(x_m)$$



If the messages go from variable nodes to factor nodes:

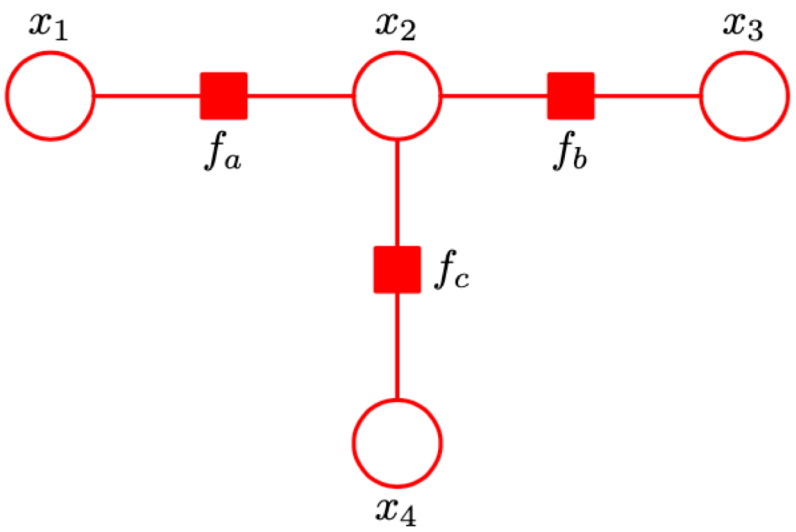
$$\mu_{x_m \to f_s}(x_m) = \prod_{l \in ne(x_m) \setminus f_s} \left[\sum_{X_{ml}} F_l(x_m, X_{ml}) \right] = \prod_{l \in ne(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$$



Using arguments, the marginal distributions $p(\boldsymbol{x}_s)$ associated with the sets of variables belonging to each of the factors is

$$p(x_s) = f_s(x_s) \prod_{i \in ne(f_s)} \mu_{(x_i \to f_s)}(x_i) \quad (33)$$

Graphical Model – sum-product



Let the root node be x_3 and x_1 and x_4 be the leaf nodes. Starting from the leaf nodes, the sequence of messages are

$$\begin{aligned}
\mu_{x_1 \to f_a}(x_1) &= 1 \\
\mu_{f_a \to x_2}(x_2) &= \sum_{x_1} f_a(x_1, x_2) \\
\mu_{x_4 \to f_c}(x_4) &= 1 \\
\mu_{f_c \to x_2}(x_2) &= \sum_{x_4} f_c(x_2, x_4) \\
\mu_{x_2 \to f_b}(x_2) &= \mu_{f_a \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \\
\mu_{f_b \to x_3}(x_3) &= \sum_{x_2} f_b(x_2, x_3)\mu_{x_2 \to f_b}.
\end{aligned} \tag{33}$$

Then we can propagate messages from the root node out to the leaf nodes:

$$\begin{aligned}
\mu_{x_3 \to f_b}(x_3) &= 1 \\
\mu_{f_b \to x_2}(x_2) &= \sum_{x_3} f_b(x_2, x_3) \\
\mu_{x_2 \to f_a}(x_2) &= \mu_{f_b \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \\
\mu_{f_a \to x_1}(x_1) &= \sum_{x_2} f_a(x_1, x_2)\mu_{x_2 \to f_a}(x_2) \\
\mu_{x_2 \to f_c}(x_2) &= \mu_{f_a \to x_2}(x_2)\mu_{f_b \to x_2}(x_2) \\
\mu_{f_c \to x_4}(x_4) &= \sum_{x_2} f_c(x_2, x_4)\mu_{x_2 \to f_c}(x_2).
\end{aligned} \tag{34}$$

Graphical Model – The max-sum

To find the latent variable values having high probability is to run the sum-product for each variable then find the value x_i^* that maximizes that marginal.

The set of values that jointly have the largest probability is

$$x^{\max} = \underset{x}{\operatorname{argmax}}\, p(\boldsymbol{x}) \tag{34}$$

for which the corresponding value of the joint probability will be given by

$$p(x^{\max}) = \max_{x} p(\boldsymbol{x}) = \max_{x_1} \cdots \max_{x_M} p(\boldsymbol{x}) \tag{35}$$

For many small probabilities, it is easier with log-scale because of the distributive property:

$$\max(a + b, a + c) = a + \max(b, c)$$

For example,

$$\mu_{f_s \to x}(x) = \sum_{x_1} \cdots \sum_{x_M} f_s(x, x_1, \ldots x_M) \prod_{m \in ne(f_s) \setminus x} \mu_{x_m \to f_s}(x_m) = \max_{x_1, \ldots, x_M} \left[\ln f_s(x, x_1, \ldots x_M) + \sum_{m \in ne(f_s)} \mu_{x_m \to f}(x_m) \right] \tag{36}$$

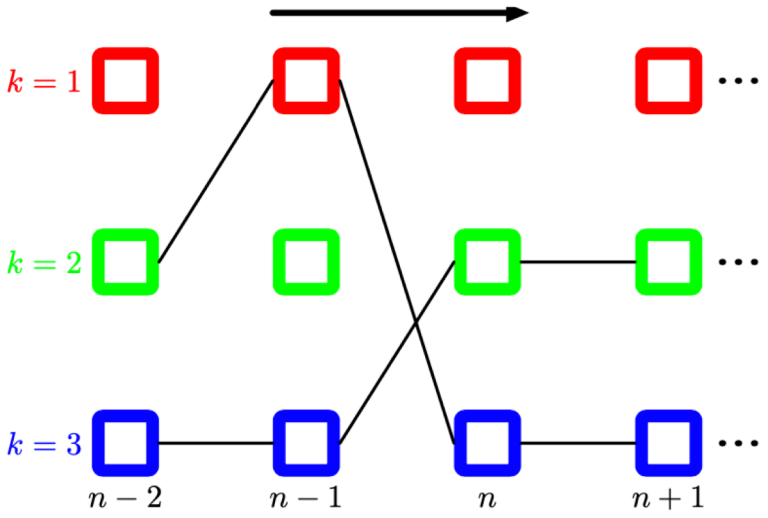
$$\mu_{x_m \to f_s}(x_m) = \sum_{l \in ne(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$$

Graphical Model – The max-sum

Using the messages, the vector $\boldsymbol{x}^{\max}$ and the corresponding joint distribution $p(\boldsymbol{x}^{\max})$ are

$$p^{\max} = \max_{x}\left[\sum_{s\in ne(x)} \mu_{f_s\to x}(x)\right] \tag{37}$$

$$x^{\max} = \operatorname*{argmax}_{x}\left[\sum_{s\in ne(x)} \mu_{f_s\to x}(x)\right]$$



Suppose we take node x_N to be the root node in which $x_N \in \{x_1, \dots x_N\}$ in each K states.

- If we propagate messages from the node x_n to x_{n+1},

$$\mu_{x_n\to f_{n,n+1}}(x_n) = \max_{x_{n-1}}\left[\ln f_{n-1,n}(x_{n-1}, x_n) + \mu_{x_{n-1}\to f_{n-1,n}}(x_n)\right]$$

- The most probable value for x_N is then

$$x_N^{\max} = \operatorname*{argmax}_{x}\left[\mu_{f_{N-1,N}\to x_N}(x_N)\right]$$

- The states of the previous variables that correspond to the same maximizing configuration by keeping track of which values of the variables give rise to the maximum state of each variable is

$$\phi(x_n) = \operatorname*{argmax}_{x}\left[\ln f_{n-1,n}(x_{n-1}, x_n) + \mu_{x_{n-1}\to f_{n-1,n}}(x_n)\right]$$

- We can then trace back to find the most probable state of x_{N-1}:

$$x_{N-1}^{\max} = \phi(x_N^{\max})$$

Graphical Models - Summary

Graphical models combine many ideas from different fields to allow an intuitive manipulation of high-dimensional problems and the corresponding multivariate probability distributions.

Markov Random Fields and Bayesian networks do not appear to be closely related, as they are so different in construction and interpretation. However, it can be shown that every dependency structure that can be expressed by a decomposable graph can be modelled both by a Markov network and a Bayesian network.