Graphical Models

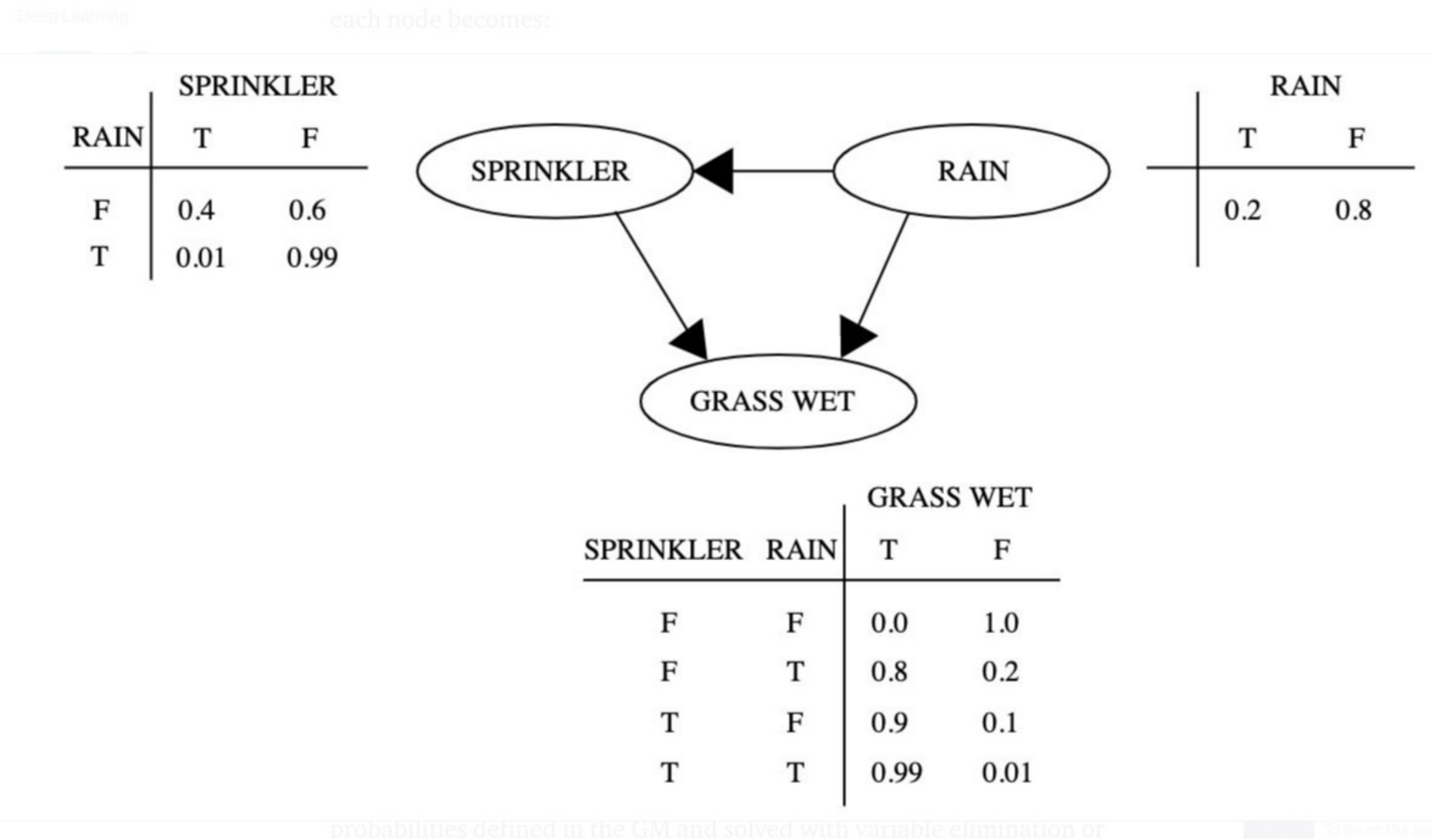
Lecture 10

10. Graphical Models

10.0 Motivation

Consider the relationship we find between R, S, and W.
What is the conditional probability of W=T when R=T?

- Solution: See the notebook.



each node becomes:

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

probabilities defined in the GM and solved with variable elimination or

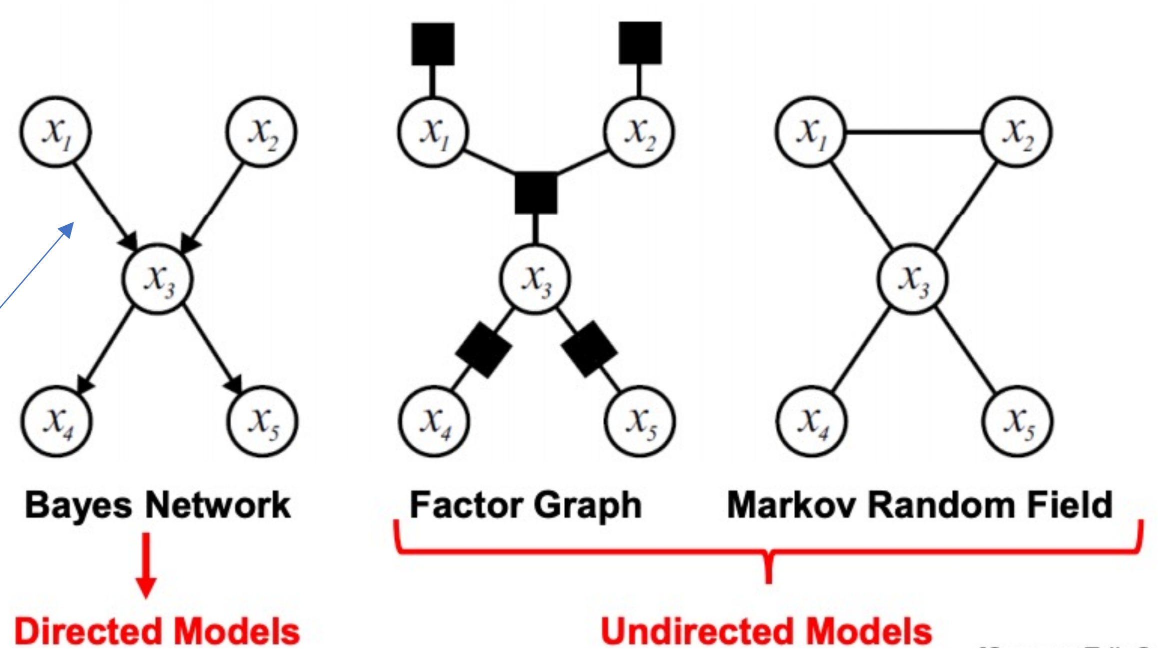
10.1 Introduction

- Probabilities play a central role in machine learning.
- Regardless of learning complexities, the repeat of sum and product rules can solve the problem algebraically.
- We can extend the analysis diagrammatically so-called **probabilistic graphical models**.
  - provide a simple way to visualize the structure of probabilistic model and propel to motivate new models.
  - insight into the properties of the model by inspecting the graph.
  - Can express complex computations in terms of graphical manipulations.

10.2 Terminology and Types

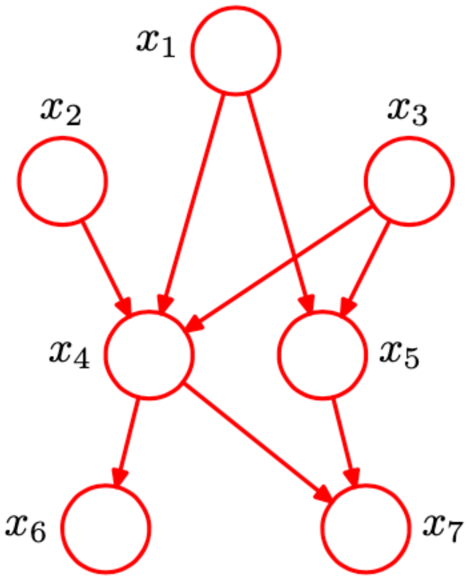
nodes (vertices): represents a random variable or a group of random variables

links (edges): Express probabilistic relationship between variables



- **Bayesian networks (directed graphical models)** have links with a particular direction indicated by arrows. Therefore, they express the relationships between RVs.
- **Markov random fields (undirected graphical models)** have links without directions and express soft constrains between RVs.
- **Factor graphs** convert both directed and undirected graphs into a different representation.

10.2 Terminology and Types



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,

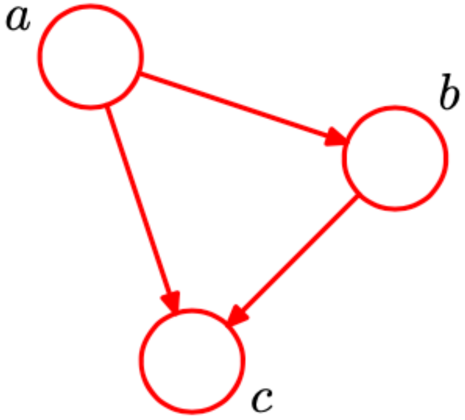
$$\mathrm{Pa}(t) = \{s: (s, t) \in E\}$$

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.

10.3 Bayesian Networks (Directed Graphs)



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:

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

Eq. 10 - 1

10.3 Bayesian Networks (Directed Graphs)

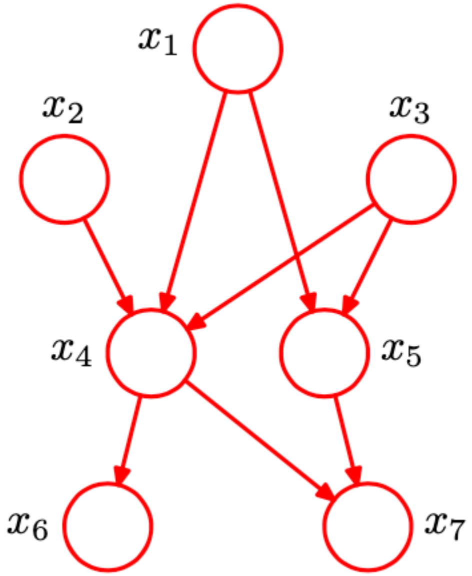


Fig 10.2

When a graph has a link between every pair of nodes, the graph is *fully connected*.

If the joint distribution has K many variables given by $p(x_1, \ldots, x_K)$, the product of conditional distributions can be formed as

$$p(x_1, \ldots, x_K) = p(x_K | x_1, \ldots, x_{K-1}) \cdots p(x_2 | x_1) p(x_1).$$

Eq. 10 - 2

Consider a case of absence of links as shown in Fig 10.2 where x_1 and x_2 are not connected as well as from x_3 to x_7.

The corresponding joint probability distribution in terms of the product of a set of conditional distribution can be written as

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5).$$

The general relationship between a given directed graph and the corresponding distribution over the variables can be formed by expressing the *factorization* properties of the point distribution as follow:

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

Eq. 10 - 3

where pa_k is the set of parents of $x_k$and $\boldsymbol{x} = \{x_1, \ldots, x_K\}$.

10.3.1. Graphical Representation

Consider the input data $\boldsymbol{x} = (x_1, \dots, x_N)^T$ and the target $\boldsymbol{t} = (t_1, \dots, t_N)^T$. The joint distribution can be formulated by the product of the prior and N conditional distributions

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

Eq. 10 - 4

and its graphical model is shown as Fig. 10.3.

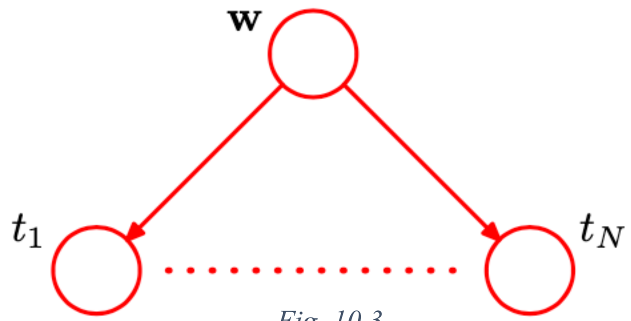


Fig. 10.3

For the simple representation, a single note t_n inside the *plate* that is labelled N – the number of nodes in the plate.

(3)

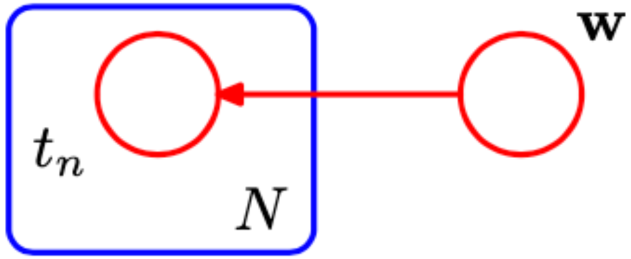


Fig. 10.4

Plate – labelled with N indicating there are N nodes.

10.3.1. Graphical Representation

Eq.10-4 can be expressed more explicitly with the noise variance σ^2 and the precision of the Gaussian prior over $\boldsymbol{w}$**,** α, as shown

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

Eq. 10 - 5

The deterministic parameters are denoted by the smaller circles and are outside of the plate as shown in Fig. 10.5.

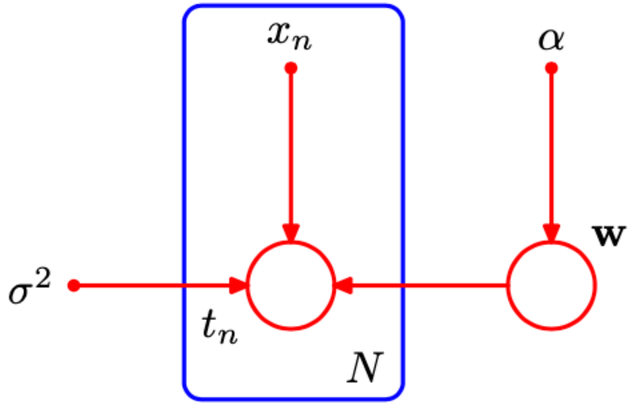


Fig. 10.5

For observed data (e.g., training data) can be presented with shading circles as shown in Fig. 10.6.

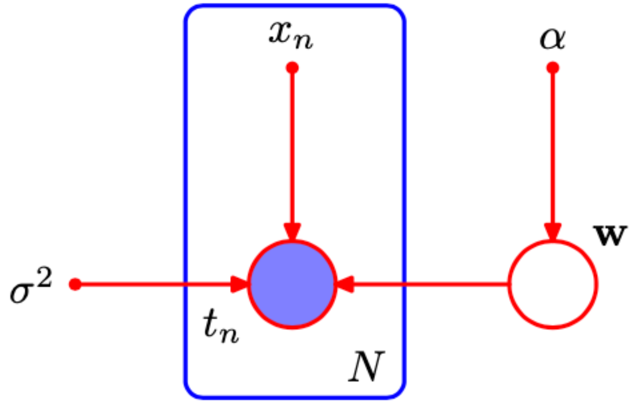


Fig. 10.6

10.3.1. Graphical Representation

For a given test data $\{\hat{x}, \hat{t}\}$, the joint distribution of all RVs can be written as

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

Eq. 10 - 6

The corresponding graphic model is shown as Fig. 10.7.

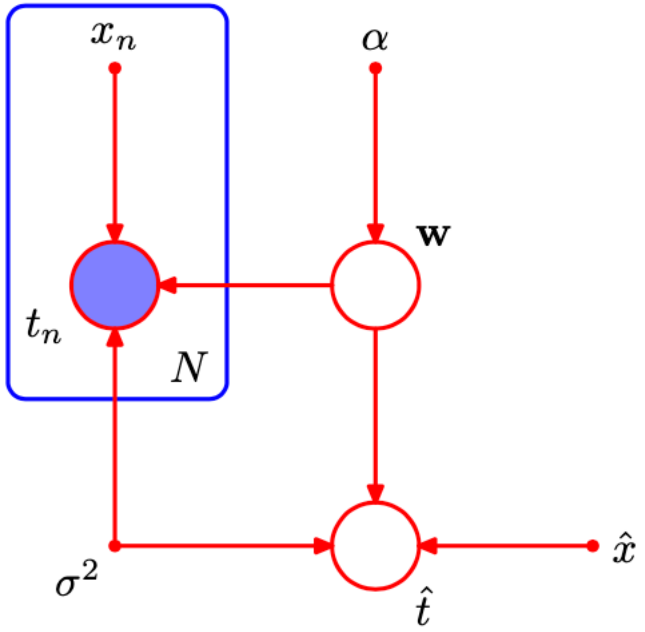


Fig. 10.7

10.3.2. Discrete Variables

Suppose there are two discrete variables $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ with K many states. The joint distribution is

$$p(x_1, x_2|\mu) = \prod_{k=1}^{K}\prod_{l=1}^{K}\mu_{kl}^{x_{1k}x_{2l}}.$$

Eq. 10 - 7

If $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are linked (e.g., $p(x_2|x_1)p(x_1)$),

- the marginal distribution over $\boldsymbol{x}_1$ is governed by $K-1$ parameters.
- the conditional distribution requires the specification of $K-1$ parameters for each of the K possible values of $\boldsymbol{x}_1$, $K(K-1)$.
- Then, the total number of parameters is $(K-1)+K(K-1)=K^2-1$.

If x_1 and x_2 are not linked and independent,

- the number of parameters governed by each is $K-1$.
- Then, the total number of parameters is $2(K-1)$.

10.3.3. Linear-Gaussian Models

Consider a directed graphed without cycles (DAG) over D variables.
Suppose each node i represents a single continuous RV x_i in Gaussian distribution.
A linear combination of the states of its parent node pa_i of i expresses the mean of distribution

$$p(x_i|\mathrm{pa}_i) = \mathcal{N}\left(x_i \middle| \sum_{j\in\mathrm{pa_i}} w_{ij}x_j + b_i, v_i\right)$$

Eq. 10 - 8

where v_i is the variance of the conditional distribution for x_i.

Taking log of the joint distribution over all nodes in the graph then be formed as

$$\ln p(\boldsymbol{x}) = \sum_{i=1}^{D} \ln p\left(x_i|\mathrm{pa_i}\right) = -\sum_{i=1}^{D} \frac{1}{2v_i}\left(x_i - \sum_{j\in\mathrm{pa_i}} w_{ij}x_j - b_i\right)^2 + \text{const.}$$

Eq. 10 - 9

10.3.3. Linear-Gaussian Models

Using the fact that each x_i in Eq.10-8 has conditional on the states of its parent, the alternative linear combination of x_i is

$$x_i = \sum_{j\in\mathrm{pa_i}} w_{ij}x_j + b_i + \sqrt{v_i}\epsilon_i$$

Eq. 10 - 10

where $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i, \epsilon_j] = I_{ij}$.

The mean and covariance of the joint distribution in recursion can determined as

$$\mathbb{E}[x_i] = \sum_{j\in\mathrm{pa_i}} w_{ij}\mathbb{E}[x_j] + b_i \,,$$

$$Type\ equation\ here.$$

$$\begin{aligned}\mathrm{cov}[x_i, x_j] &= \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] \\ &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])\left\{\sum_{k\in\mathrm{pa_i}} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_j}\epsilon_j\right\}\right] \\ &= \sum_{k\in\mathrm{pa_i}} w_{jk}\mathrm{cov}[x_i, x_k] + I_{ij}v_j\end{aligned}$$

Eq. 10 - 11

10.3.3. Linear-Gaussian Models

Using two relations in Eq.10-11, the mean and covariance of the joint distribution are

$$\boldsymbol{\mu} = (b_1, b_2 + w_{21}, b_3 + w_{32}b_2 + w_{32}w_{21}b_1, \ldots)^T,$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} v_1 & w_{21}v_1 & w_{32}w_{21}v_1 & \cdots \\ w_{21}v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) & \cdots \\ w_{32}w_{21}v_1 & w_{32}(v_2 + w_{21}^2 v_1) & v_3 + w_{32}^2(v_2 + w_{21}^2 v_1) & \cdots \\ \vdots & \vdots & \vdots & \cdots \end{pmatrix}.$$

Eq. 10 - 12

For multivariate Gaussian variables, the conditional distribution for node i can be expressed as

$$p(\boldsymbol{x}_i|\mathrm{pa}_\mathrm{i}) = \mathcal{N}\left(\boldsymbol{x}_i \middle| \sum_{j\in\mathrm{pa}_\mathrm{i}} \boldsymbol{W}_{ij}\boldsymbol{x}_j + \boldsymbol{b}_i, \boldsymbol{\Sigma}_i\right).$$

Eq. 10 - 13

10.4. Conditional Independence

Consider a case of three variables $\{a, b, c\}$. If a is conditional independent of b given c, $p(a, b|c) = p(a|c)p(b|c)$ as shown in Fig. 10.8.

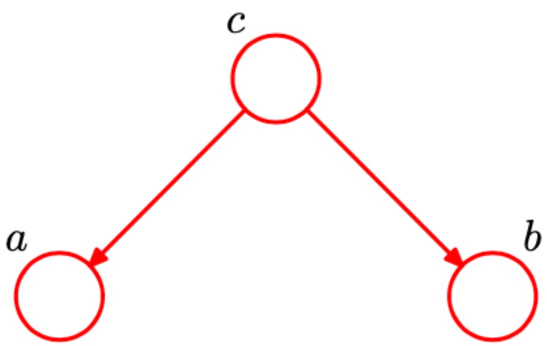


Fig. 10.8

In principle, any potential conditional independence property needs to be tested with sum and product rules repeatedly.
In graphical models, conditional independence properties of joint distribution can be read without any performance. This framework is called **d-separation** where *d* stands for ‘directed’.

10.4.1. Three Conditional Independence Graphical Representations

If a and b are conditioned on c,

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

as shown in Fig. 10-9, the conditional distribution of a and b, given c, is formed as

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

The node c is called **tail-to-tail** representation.

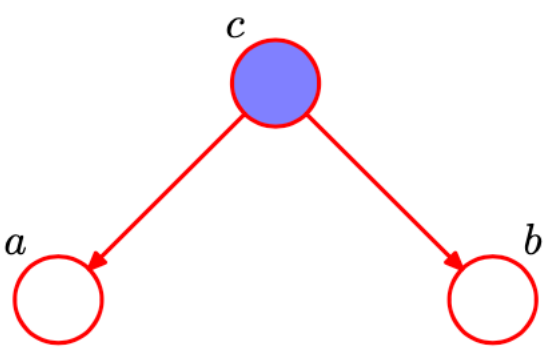


Fig. 10.9

Fig. 10.10, the node c is **head-to-tail** w.r.t. the path from node a to b.
If the joint distribution $p(a,b,c) = p(a)p(c|a)p(b|c)$, the conditional independency property can be obtained as

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

again using Bayes' theorem

$$p(c|a) = \frac{p(a|c)p(c)}{p(a)}.$$

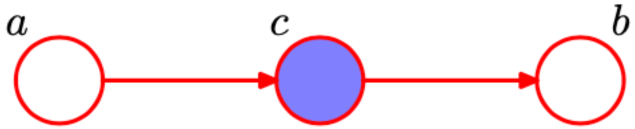


Fig. 10.10

If none of variables are observed, a case of Fig 10.11,

$$p(a,b) = p(a)\sum_c p(c|a)p(b|c) = p(a)p(b|a).$$

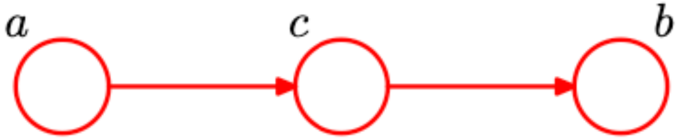


Fig. 10.11

10.4.1. Three Conditional Independence Graphical Representations

Fig. 10.12 shows the conditional independent joint distribution

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

the conditional distribution of a and b is then

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

where does not factorize into the product $p(a)p(b)$.

If none of variables are observed (Fig. 10.13), then $p(a,b)=p(a)p(b)$. These graphic representations are called **head-to-head**.

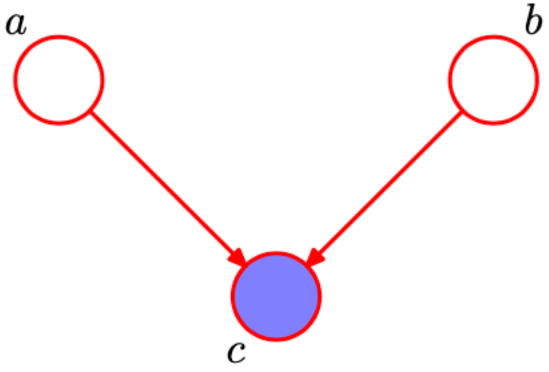


Fig. 10.12

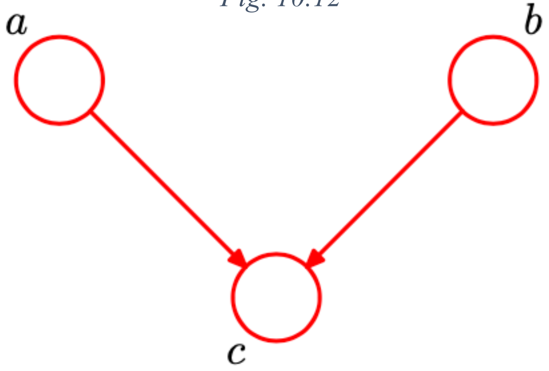


Fig. 10.13

10.4.1. Three Conditional Independence Graphical Representations

Consider three binary variables B, F, and G.

Suppose the prior probabilities are given as

$$p(B = 1) = 0.9$$
$$p(F = 1) = 0.9$$

and G probabilities are given by

$$p(G = 1|B = 1, F = 1) = 0.8$$
$$p(G = 1|B = 1, F = 0) = 0.2$$
$$p(G = 1|B = 0, F = 1) = 0.2$$
$$p(G = 1|B = 0, F = 0) = 0.1.$$

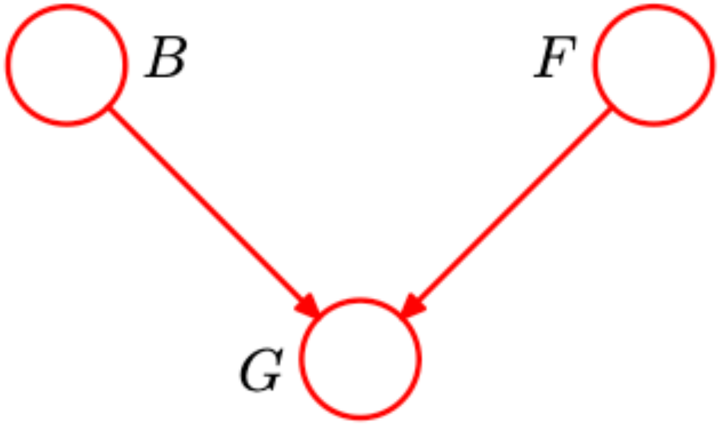


Fig. 10.14 a

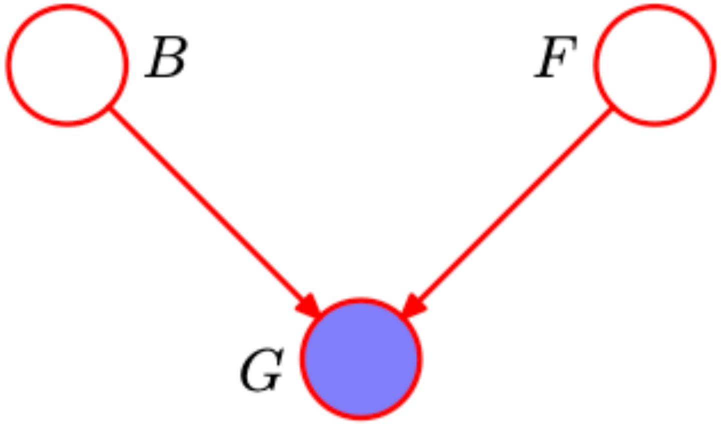


Fig. 10.14 b

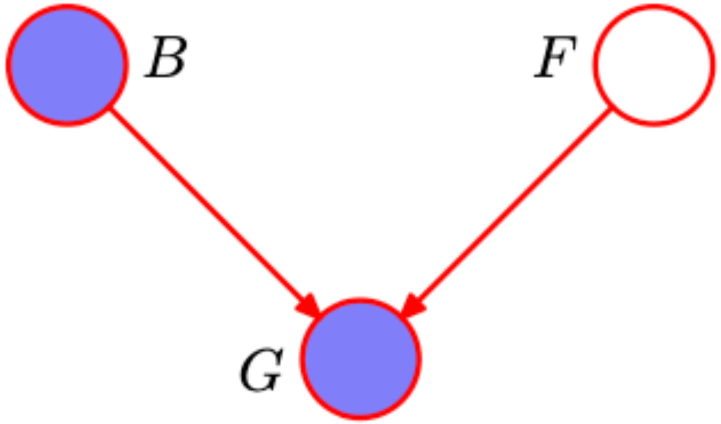


Fig. 10.14 c

10.4.1. Three Conditional Independence Graphical Representations

The probability of $p(F=0|G=0)$ as shown in Fig. 10.14b is

$$p(F=0|G=0)=\frac{p(G=0|F=0)p(F=0)}{p(G=0)}=0.81\cdot\frac{0.1}{0.315}=0.257.$$

The denominator is

$$\begin{aligned}p(G=0)=\sum_{B\in\{0,1\}}\sum_{F\in\{0,1\}}p(G=0|B,F)p(B)p(F)=0.315\\ =p(G=0|B=0,F=0)p(B=0)p(F=0)\\ +p(G=0|B=1,F=0)p(B=1)p(F=0)\\ +p(G=0|B=0,F=1)p(B=0)p(F=1)\\ +p(G=0|B=1,F=1)p(B=1)p(F=1)\\ =(0.9\cdot0.1\cdot0.1)+(0.8\cdot0.9\cdot0.1)+(0.8\cdot0.1\cdot0.9)+(0.2\cdot0.9\cdot0.9)\end{aligned}$$

and $p(G=0|F=0)$ is

$$\sum_{B\in\{0,1\}}p(G=0|B,F=0)p(B)=p(G=0|B=0,F=0)p(B=0)+p(G=0|B=1,F=0)p(B=1)$$

$$=(0.9\cdot0.1)+(0.8\cdot0.9)=0.81$$

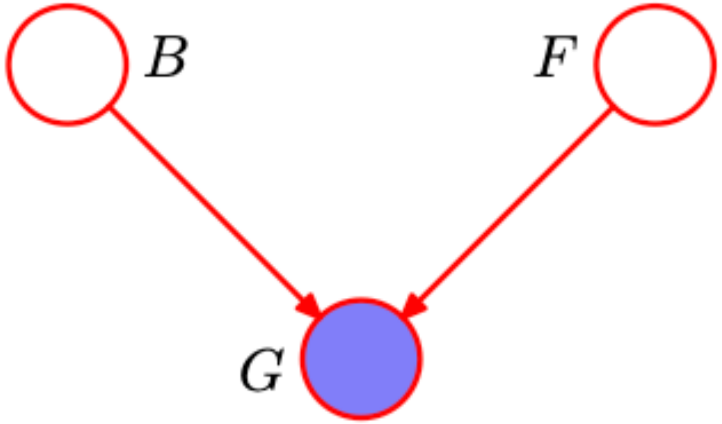


Fig. 10.14 b

Fig. 10.14 c

10.4.1. Three Conditional Independence Graphical Representations

The posterior of $p(F = 0|G = 0, B = 0)$ as shown in Fig.10.14c can be calculated as

$$= \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F\in\{0,1\}} p(G = 0|B = 0, F)p(F)} \cong 0.111.$$

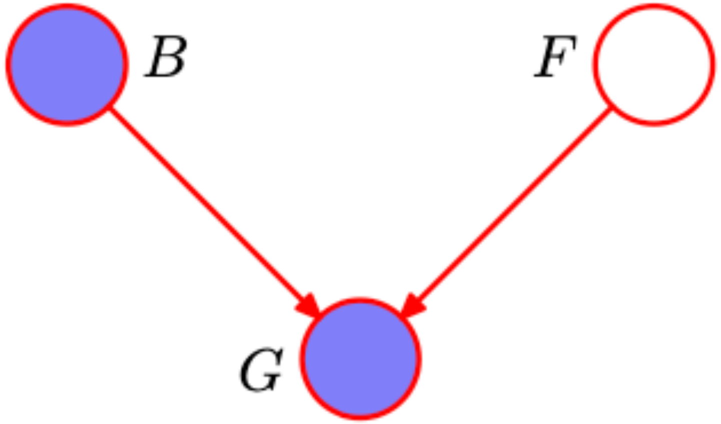


Fig. 10.14 c

$$p(G = 0|B = 0, F = 0)p(F = 0) = 0.9 \cdot 0.1 = 0.09$$

$$\sum_{F\in\{0,1\}} p(G = 0|B = 0, F)p(F)$$
$$= p(G = 0|B = 0, F = 0)p(F = 0) + p(G = 0|B = 0, F = 1)p(F = 1)$$
$$= (0.9 \cdot 0.1) + (0.8 \cdot 0.9) = 0.81$$

$$p(F = 0|G = 0, B = 0) = \frac{0.09}{0.81} = 0.11\overline{1}$$

10.4.2. d-separation

- Consider a general directed graph with three arbitrary nonintersecting sets of nodes $\{A, B, C\}$.
- Suppose we are to ascertain whether a particular conditional independence of A and B at given C. While all possible paths from any node in A to any node *in* B are considered, any of paths includes the following conditions are called *blocked*:
  1. the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in C, or
  2. the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in C.

- If all paths are blocked, then A is "**d-separated**" and joint distribution over all the variables in the graph satisfy the conditional independency of A from B at give C.

10.4.2.1 d-separation example 1: Graphic Interpretation

Fig. 10.15 shows the concept of d-separation.

Graph (a):

- a path $a \rightarrow b$ is not blocked by f : tail-to-tail and are not observed.
- e does not block the path: head-to-head and has c descendent.
- therefore, a and b are not independent given c.

Graph (b):

- a path $a \rightarrow b$ is blocked by f.
- a and b are independent given f.
- even e blocks the path, head-to-head and its descendent c is not in the set.

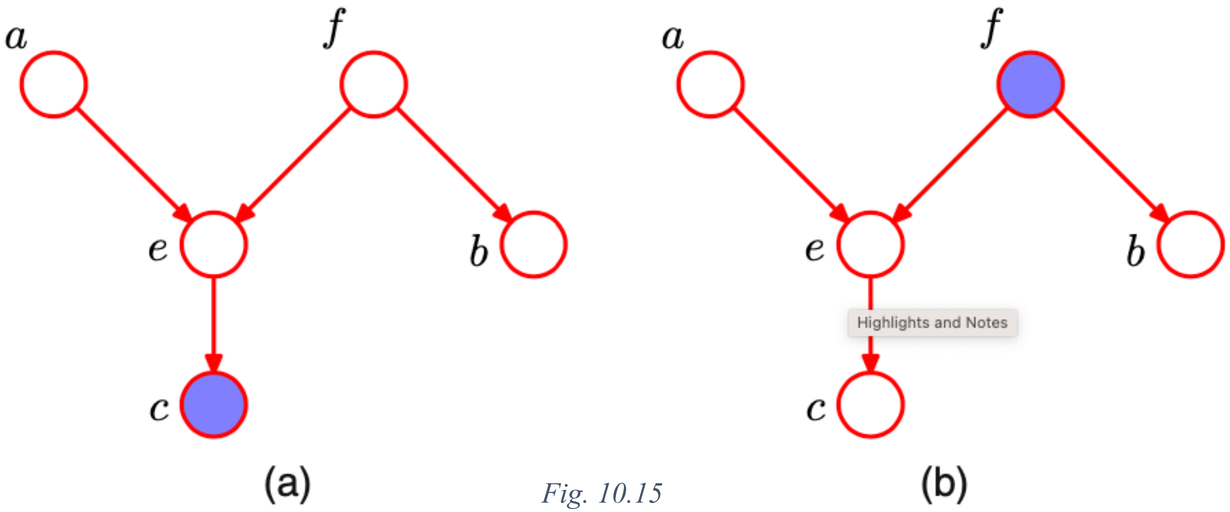


Fig. 10.15

10.4.2.1 d-separation example 2: the concept of i.i.d.

Fig. 10.16 represents a set of conditional distribution $p(\boldsymbol{x}|\mu)$.
Using d-separation, there is a unique tail-to-tail path $x_i \rightarrow x_{j\neq i}$ w.r.t. μ

$$p(x|\mu) = \prod_{n=1}^{N} p(x_n|\mu)\,.$$

For naïve Bayes classification model, Fig. 10.16 displays that the class label **z** blocks the path $x_i \rightarrow x_{j\neq i}$ and the assumption of input variables' distribution, that are independent to each other, are satisfied.

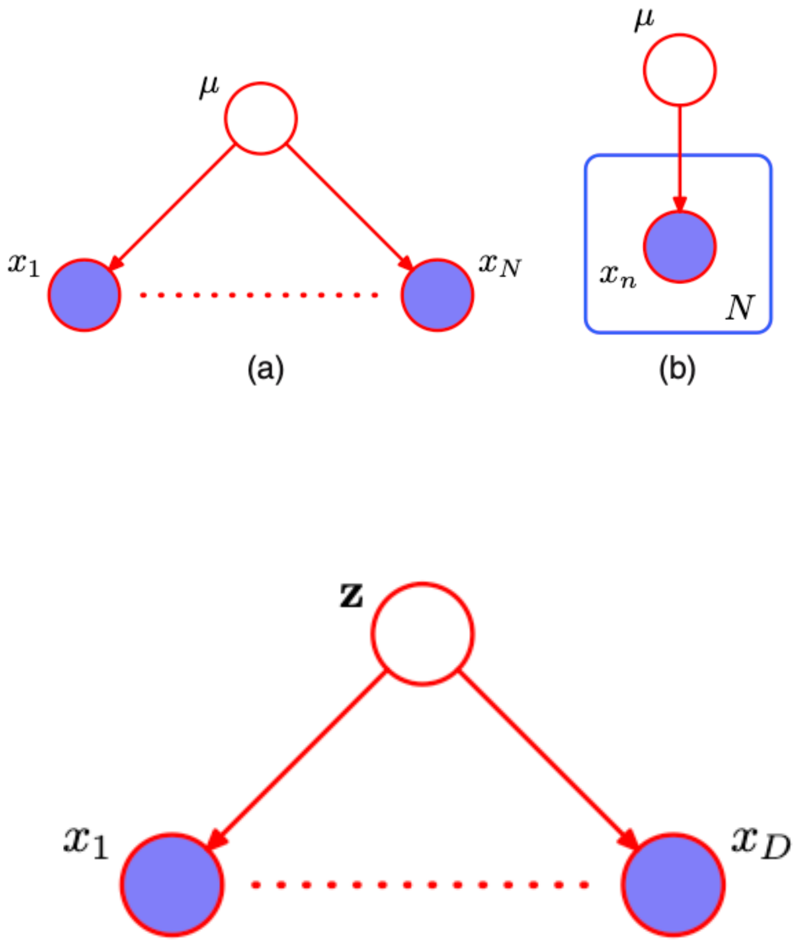


Fig. 10.16

10.4.2.1 d-separation example 3: Directed Factorization

- The two d-separation properties help to express a directed graph to represent the specific decomposition of a joint probability distribution into a product of conditional probabilities.
- A role of directed graph is like a filter as shown in Fig. 10.17.
  - It filters the set of all possible distribution $p(\boldsymbol{x})$ over the set of $\boldsymbol{x}$ and subset the distributions passed through.
  - These subsets are called **directed factorization** ($\mathcal{DF}$).

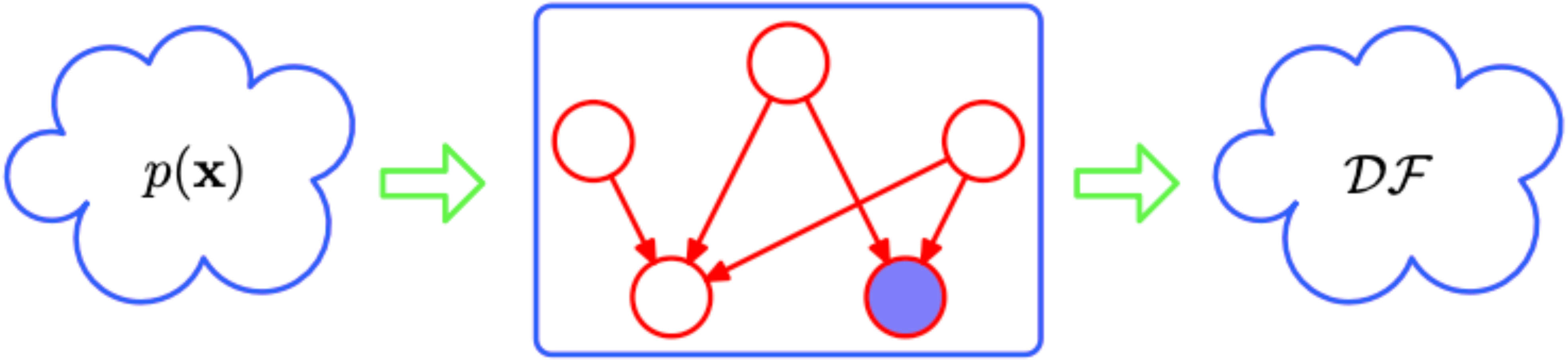


Fig. 10.17

10.4.2.4. Markov Blanket (Markov Boundary)

If a particular node with $\boldsymbol{x}_i$ conditioned on all the rest variables $\boldsymbol{x}_{j\neq i}$, the conditional distribution

$$p(\boldsymbol{x}_i|\boldsymbol{x}_{j\neq i}) = \frac{p(\boldsymbol{x}_1,\dots,\boldsymbol{x}_D)}{\int(p(\boldsymbol{x}_1,\dots,\boldsymbol{x}_D)\mathrm{d}\boldsymbol{x}_i} = \frac{\prod_k p(\boldsymbol{x}_k|\mathrm{pa_k})}{\prod_k(\boldsymbol{x}_k|\mathrm{pa_k})\mathrm{d}\boldsymbol{x}_i}$$

Eq. 10 - 14

shows that any $p(\boldsymbol{x}_k|\mathrm{pa_k})$ does not have any functional dependence on x_i and only the remaining factor will the conditional distribution $p(\boldsymbol{x}_i|\mathrm{pa_I})$ for x_i itself and any nodes $\boldsymbol{x}_k$ such that $\boldsymbol{x}_i$ is in the conditioning set of $p(\boldsymbol{x}_k|\mathrm{pa_k})$.

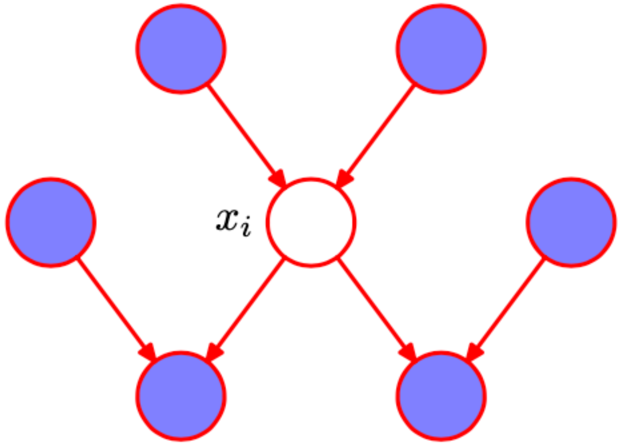


Fig. 10.18

In Eq.10-14, for which $\boldsymbol{x}_i$ is $\boldsymbol{x}_k.\mathrm{pa} = \boldsymbol{x}_i$,

- $p(\boldsymbol{x}_i|\mathrm{pa_i})$ is depends on $\boldsymbol{x}_i.\mathrm{pa}$.
- $p(\boldsymbol{x}_k|\mathrm{pa_k})$ is depends on children and co-parent of $\boldsymbol{x}_i$:
  - denotation: $x_i.\mathrm{Ch.}$ and $x_i.\mathrm{CoPa,}$ respectively
  - the observations of Ch nodes will not block paths to CoPa.

The set of nodes compromising the parent, the children, and the co-parent is called the **Markov blanket** as shown in Fig. 10.18.

10.5. Markov Random Fields (Undirected Graph)
10.5.1. Conditional Independence Properties

- Suppose there is an undirected graph with three nodes $\{A, B, C\}$ having the conditional independence property of A and B given C as shown in Fig. 10.19.

- If all nodes in C and links that connect to those nodes are removed, the conditional independence property can be proved if there are no paths connect from A to B.

- The Markov blanket takes the graph into a simple format because a node will be conditionally independent of all other nodes conditioned only on the neighboring nodes as shown in Fig. 10.20.

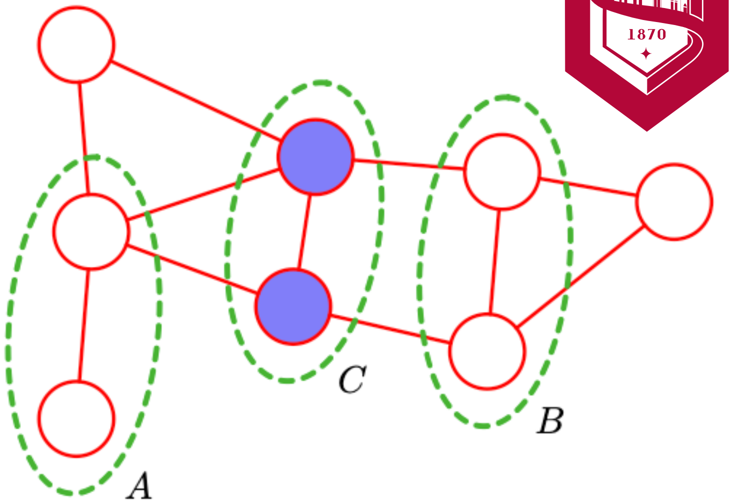


Fig. 10.19

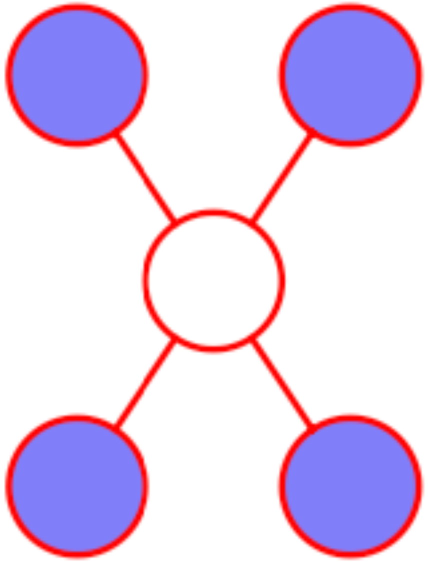


Fig. 10.20

10.5.2. Factorization Properties

If two nodes x_i and x_j are not connected by a link and conditionally independent, there is no direct path between the two nodes. This conditional independence property is expressed as

$$p(x_i, x_j|\boldsymbol{x}_{\{i,j\}}) = p(x_i|\boldsymbol{x}_{\setminus\{i,j\}})p(x_j|\boldsymbol{x}_{\setminus\{i,j\}})$$

where $\boldsymbol{x}_{\setminus\{i,j\}}$ denotes the set of x of all variables with x_i and x_j removed.

- A graphical concept, **clique**, is defined as a subset of nodes such that there is a link between all pairs of nodes in the subset. The set of nodes in a clique is fully connected.
- A **maximal clique** is a clique such that it is not possible to include any other nodes in the set without it ceasing to be a clique.
- See Fig. 10.21.

The joint distribution of the nodes in the clique, x_C, is expressed as

$$p(\boldsymbol{x}) = \frac{\prod_C \psi_C(\boldsymbol{x}_C)}{\sum_x \prod_C \psi_C(\boldsymbol{x}_C)} = \frac{1}{Z}\prod_C \psi_C(\boldsymbol{x}_C).$$

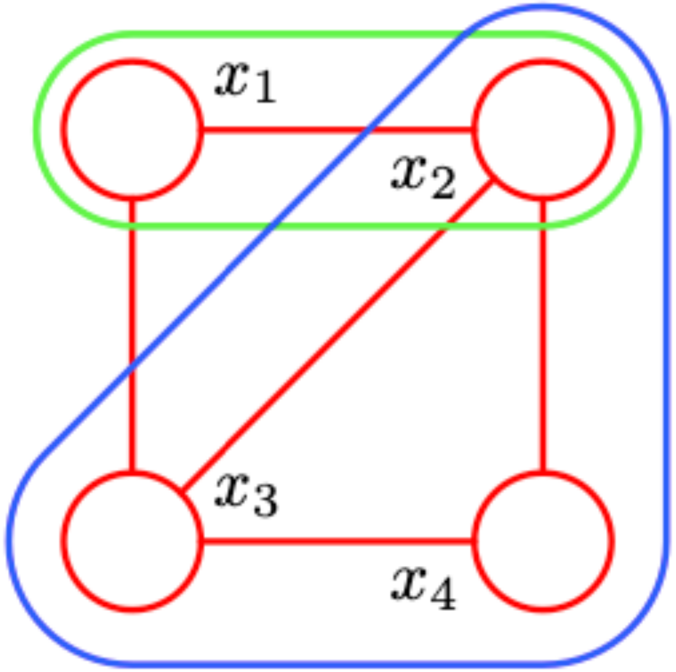


Fig. 10.21

10.6. Inference in Graphical Models

Suppose the joint distribution $p(x, y)$ over two variables x and y into a product of factors is $p(x, y) = p(x)p(y|x)$ as shown in Fig.10.22.

If the variable y is observed, the marginal distribution $p(x)$ is a prior over the latent variable x. To infer the corresponding posterior distribution over x. The sum and the product rules of probability are

$$p(y) = \sum_{x'} p(y|x')p(x')$$

Eq. 10 - 15

and

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}.$$

Eq. 10 - 16

The joint distribution is represented by the graph shown in Fig.10.22.

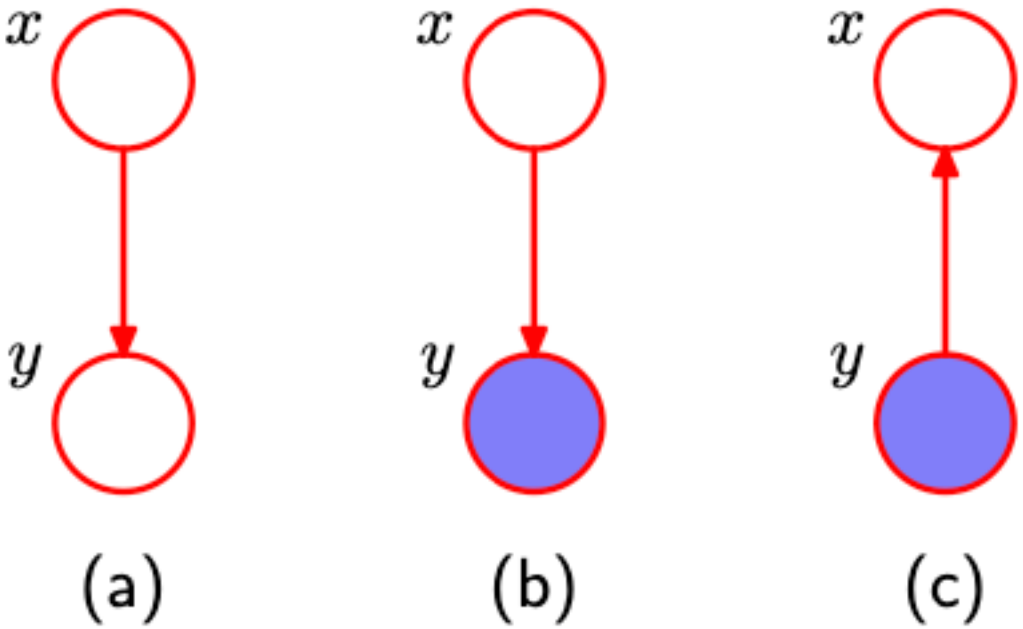


Fig. 10.22

10.6.1 Inference on a Chain

- The joint distribution for the graph takes the form

$$p(x) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\cdots\psi_{N-1,N}(x_{N-1,}x_N).$$

Eq. 10 - 17

  - If each node has K states, each $\psi_{n-1,n}(x_{n-1}, x_n)$ comprises $K\times K$ matrix and the joint distribution has $(N-1)K^2$ parameters.

- The marginal distribution $p(x_n)$ for a specific x_n that is along the chain is

$$p(x_n) = \sum_{x_1}\cdots\sum_{x_{n-1}}\sum_{x_{x+1}}\cdots\sum_{x_N} p(\boldsymbol{x})\,.$$

Eq. 10 - 18

- The summation in Eq.10-18 has K^N values.

10.6.1 Inference on a Chain

- For more efficient computation, substitute Eq.10-17 into Eq.10-18 and rearrange the order of summation and multiplication:
  - Since $\psi_{N-1,N}(x_{N-1}, x_N)$ is only depends on x_N, the summation is $\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)$ only involves the new function together only with $\psi_{N-2,N-1}(x_{N-2}, x_{N-1})$.
  - Therefore, the desired marginal is formed

$$p(x_n) = \frac{1}{Z}\left[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \cdots \left[\sum_{x_2} \psi_{2,3}(x_2, x_3) \left[\sum_{x_1} \psi_{1,2}(x_1, x_2)\right]\right] \cdots \right] \times$$
$$\left[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \cdots \left[\sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots \right] = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n).$$

Eq. 10 - 19

  - Eq.10-19 graph representation is shown as Fig. 10.23.
  - The interpretation can be as passing "messages" around the graph - $\mu_\alpha(x_n)$ as a forward passing message and $\mu_\beta(x_n)$ as a backward passing message.

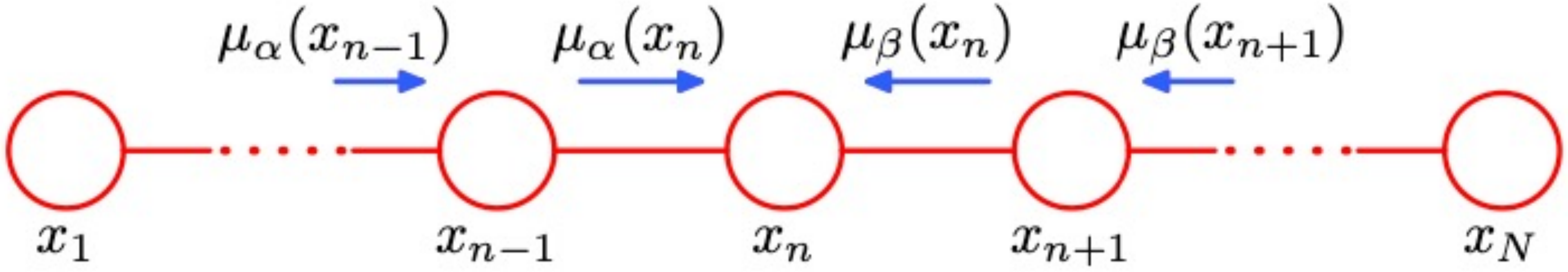


Fig. 10.23

10.6.2. Factor Graphs

- A **factor graph** adds a new factor node between subsets of variable nodes. It converts any graph into a tree format.

- The expression of joint distribution in the form of a product of factors is

$$p(\boldsymbol{x}) = \prod_{s} f_s(\boldsymbol{x}_s)$$

  where $\boldsymbol{x}_s$ is a subset of the variables.

- Factors are depicted by small squares as shown in Fig. 10.24
- The distribution of the example can be expressed as

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

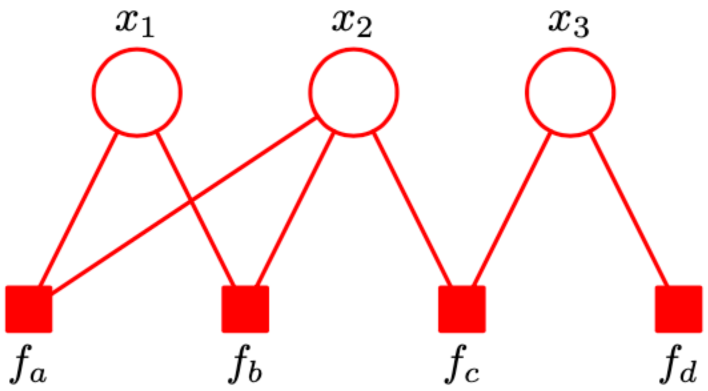


Fig. 10.24

10.6.2. Factor Graphs

- In an undirected graph, the product of factors can simply lumped together into the same clique potential when factors are defined over the same set of variables by creating an additional factor node corresponding to the maximal clique.

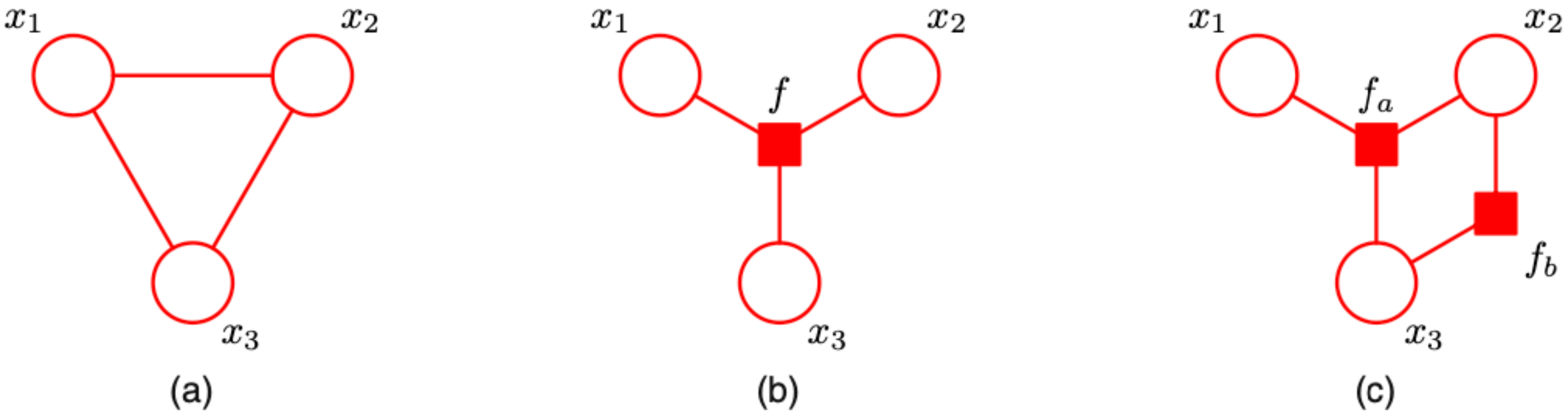
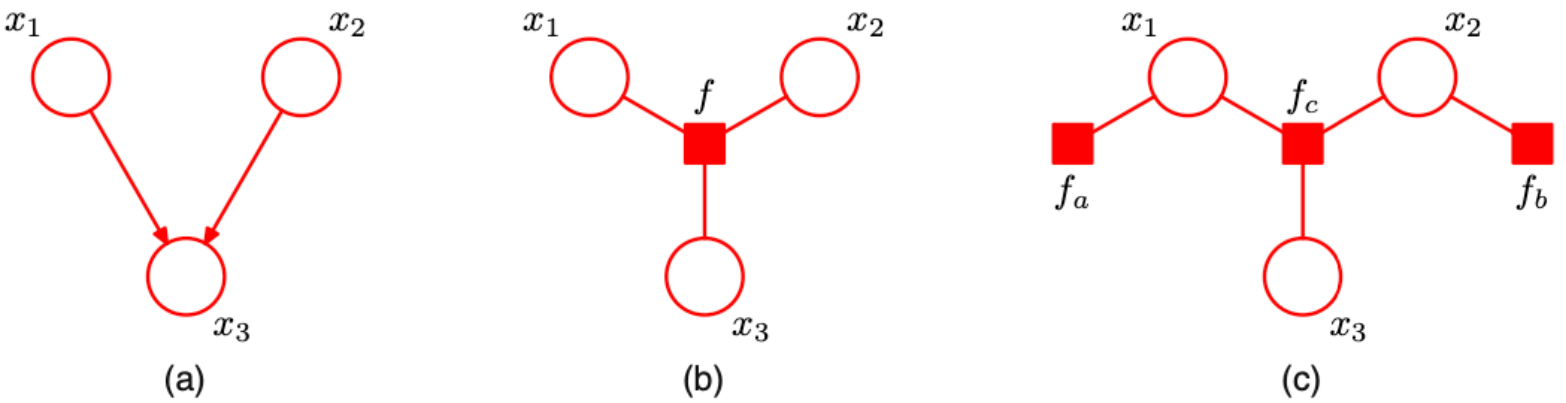


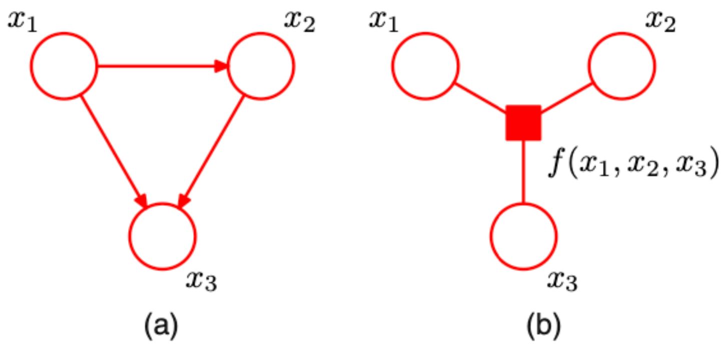
Fig. 10.25

10.6.2. Factor Graphs

- In an undirected graph, the product of factors can simply lumped together into the same clique potential when factors are defined over the same set of variables by creating an additional factor node corresponding to the maximal clique.
- In a directed graph, create the variable nodes in the factor graph corresponding to the nodes of the directed graph and then create factor nodes corresponding to the conditional distributions, and then add the appropriate links.



  - When there is a cycle in a direct graph, the links connecting parents of a node can be removed by defining the appropriate factor function.



10.6.3. 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.

10.6.3.1. The Sum-Product Algorithm

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

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

$$p(x) = \sum_{\mathbf{x}\backslash x} p(\mathbf{x})$$

where $\mathbf{x} \setminus x$ the set of variables in $\mathbf{x}$ 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:

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

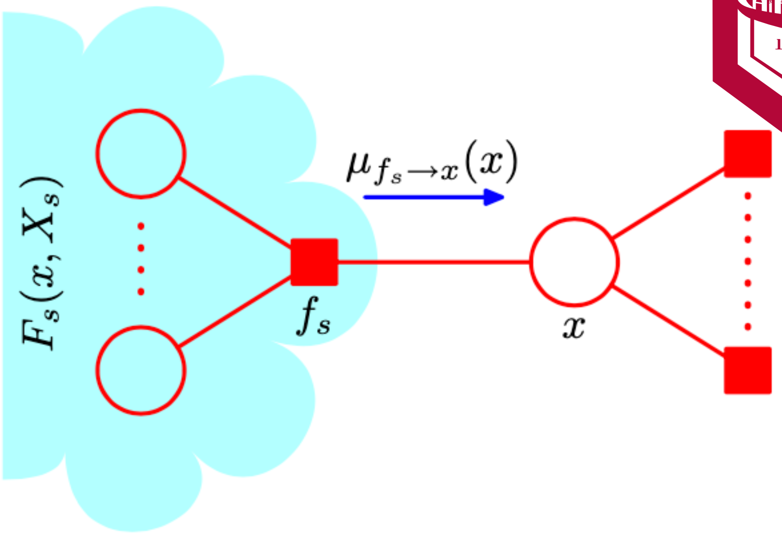
where:

  - ○ $ne(x)$: 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:

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

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



Eq. 10 - 20

10.6.3.1. The Sum-Product Algorithm

1. Evaluate the message sent by a factor node to a variable node along the link connecting them. Each factor $F_s(x, X_s)$ is described by a factor (sub-)graph and therefore, itself can be factorized:

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

Eq. 10 - 21

where $x_1, \dots, x_M$ are variables associated with f_s, see Fig. 10.25.
The substitution of Eq.10-21 to Eq.10-20 is

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

$$= \sum_{x_1} \cdots \sum_{x_M} f_s(x, x_1, \dots, x_M) \prod_{m \in ne(f_x) \setminus x} \mu_{x_m \to f_s}(x_m) \, .$$

Eq. 10 - 22

Note that a factor node can send a message to a variable node once it has received incoming messages from all other neighbor variable nodes.

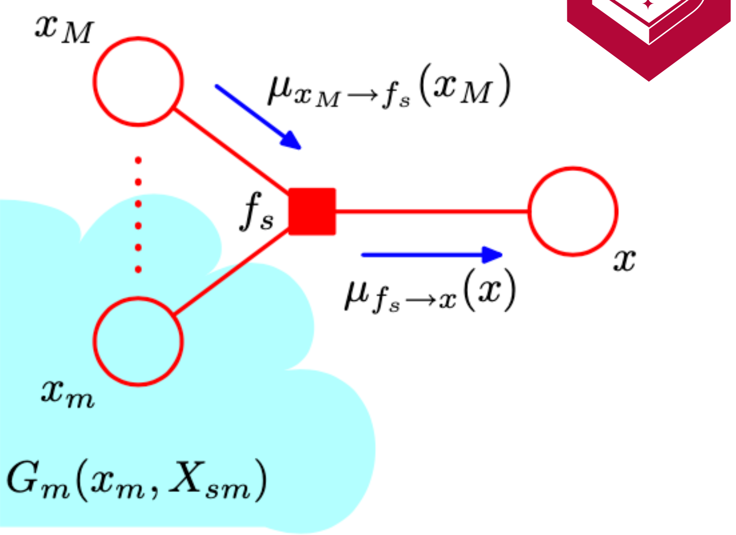


Fig. 10.25

10.6.3.1. The Sum-Product Algorithm

2. Derive to evaluate the messages from variable nodes to factor nodes. See note that $G_m(x_m, X_{sm})$ associated with node x_m is given by a product of $F_l(x_m, X_{ml})$ each associated with one of the factor node f_l that is linked to x_m (excluding node f_s),

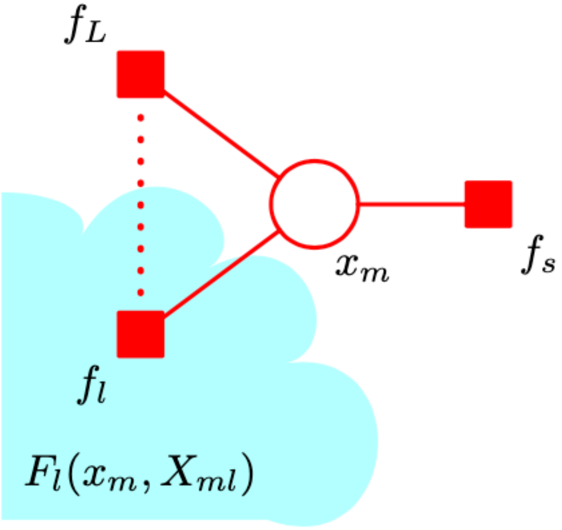
$$G_m(x_m, X_{sm}) = \prod_{l \in ne(x_m) \backslash f_s} F_l(x_m, X_{ml})\,.$$

Eq. 10 - 23

It leads to

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

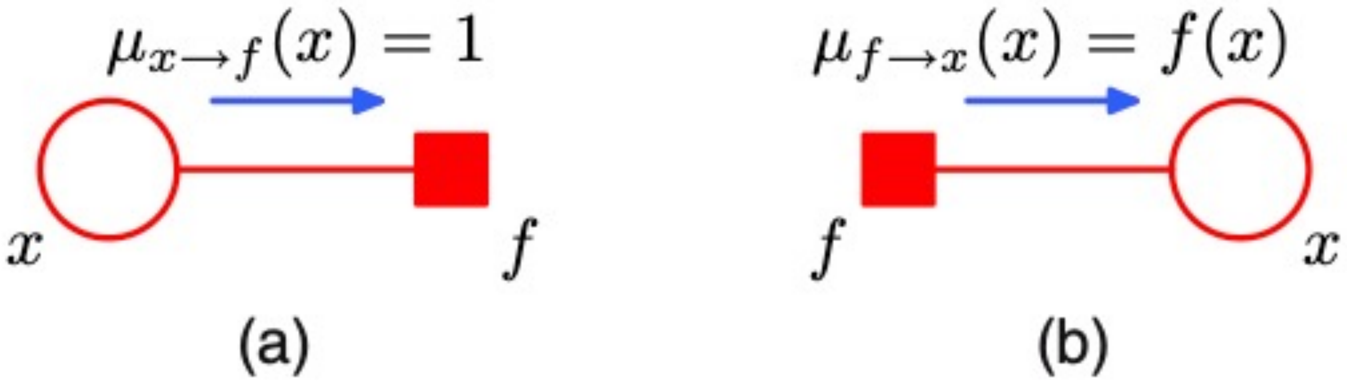
Eq. 10 - 24



10.6.3.1. The Sum-Product Algorithm

3. Calculate the marginal for x. Each of the incoming message of links arriving at x can be recursively commuted in terms of other messages. Two cases are shown in Figures.

- If a leaf node is a variable node, then the message sends along its one and only link is
$$\mu_{x\to f}(x) = 1.$$
- if the leaf node is a factor node, the message sent takes the form
$$\mu_{f\to x}(x) = f(x).$$



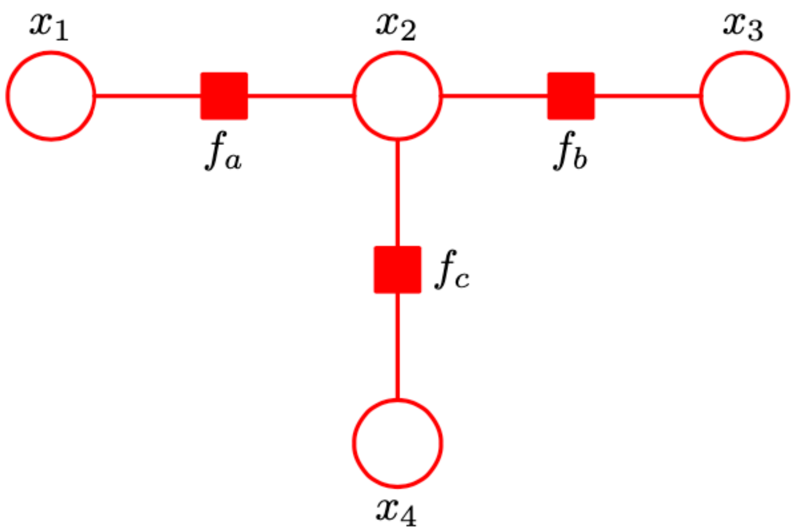
10.6.3.1. The Sum-Product Algorithm

4. Find $p(\boldsymbol{x}_s)$ associated with the sets of variables belonging to each of the factors. The marginal associated with a factor is given by the product of messages arriving at the factor node and the local factor at that node

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

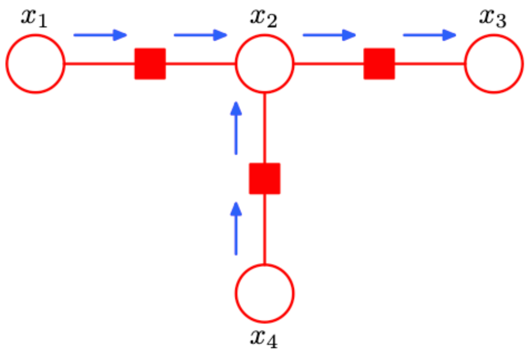
- Normalization Issue
  - Undirected Graph – Unknown normalization
    - Let the joint distribution $\tilde{p}(\boldsymbol{x})$ so $p(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})/Z$.
    - Run the sum-product algorithm to find the unnormalized marginals $\tilde{p}(x_i)$.
    - Then, normalize any one of the marginals.
  - Directed Graph
    - If the factor graph is derived, then the joint distribution is already normalized correctly and then run the sum-product algorithm.

10.6.3.1. The Sum-Product Algorithm



Suppose 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

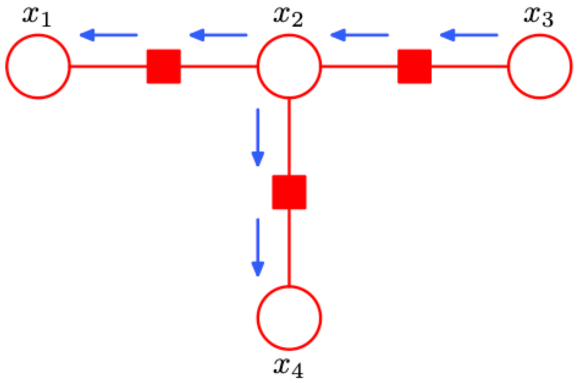
$$\begin{aligned}
\mu_{x_1 \to f_a}(x_1) &= 1 \\
\mu_{f_a \to x_2}(x_2) &= \sum_{x_1} f_a(x_1, x_2) \\
\mu_{x_4 \to f_c}(x_4) &= 1 \\
\mu_{f_c \to x_2}(x_2) &= \sum_{x_4} f_c(x_2, x_4) \\
\mu_{x_2 \to f_b}(x_2) &= \mu_{f_a \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \\
\mu_{f_b \to x_3}(x_3) &= \sum_{x_2} f_b(x_2,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



(a)

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

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



(b)

10.6.3.1. The Sum-Product Algorithm

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

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

- The marginal $p(x_2)$ is

$$
\begin{aligned}
\tilde{p}(x_2) &= \mu_{f_a \to x_2}(x_2)\mu_{f_b \to x_2}(x_2)\mu_{f_c \to x_2}(x_2) \\
&= \left[\sum_{x_1} f_a(x_1, x_2)\right]\left[\sum_{x_3} f_b(x_2, x_3)\right]\left[\sum_{x_4} f_c(x_2, x_4)\right] \\
&= \sum_{x_1}\sum_{x_3}\sum_{x_4} \tilde{p}(\boldsymbol{x}).
\end{aligned}
$$

10.6.3.2. The Max-Sum Algorithm

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

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

Eq. 10 - 24

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

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

Eq. 10 - 25

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,

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

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

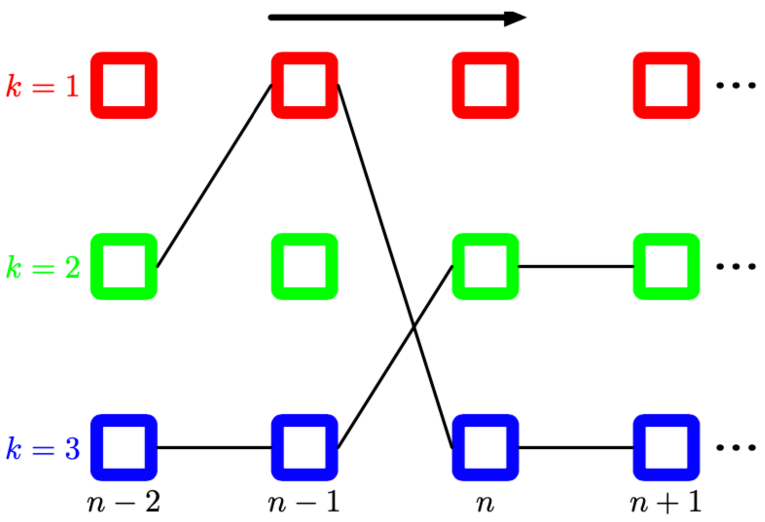
Eq. 10 - 26

10.6.3.2. The Max-Sum Algorithm

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

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

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



Suppose we take node x_N to be the root node in which $x_N \in \{x_1, \ldots x_N\}$ 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}:
$$x_{N-1}^{\max} = \phi(x_N^{\max})$$

10.7. Conclusion

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.