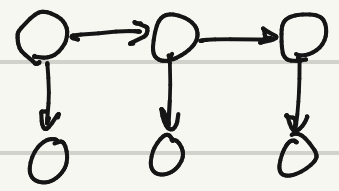
PRML Chapter 8 Graphical Model

Merit

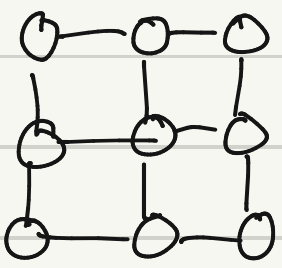
1) provide a simple way to visualize a probabilistic model
2) Insights into the properties of the model
3) complex computations can be expressed in terms of graphical manipulations.

Graphical Model



→ Bayesian Network.
useful for expressing causal relation between random variables

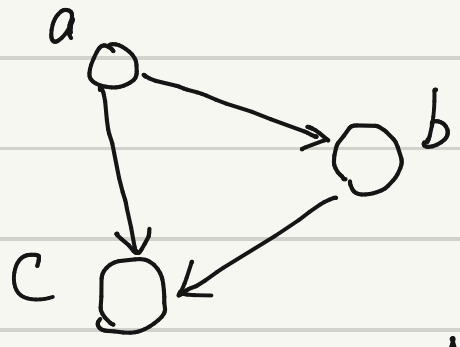
→ Markov Random Field
useful for expressing soft constraints between random variables.



8.1 Bayesian Networks

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

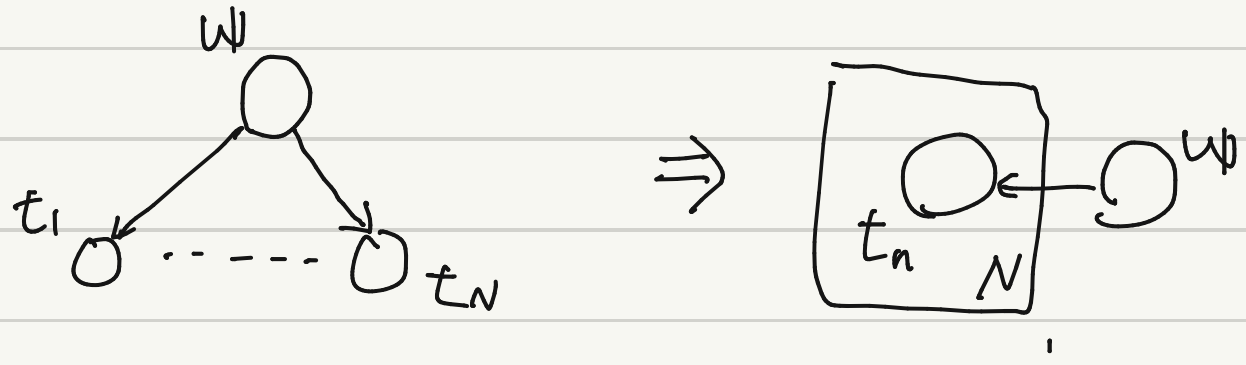
generalize



Absolutely, decomposition is not unique

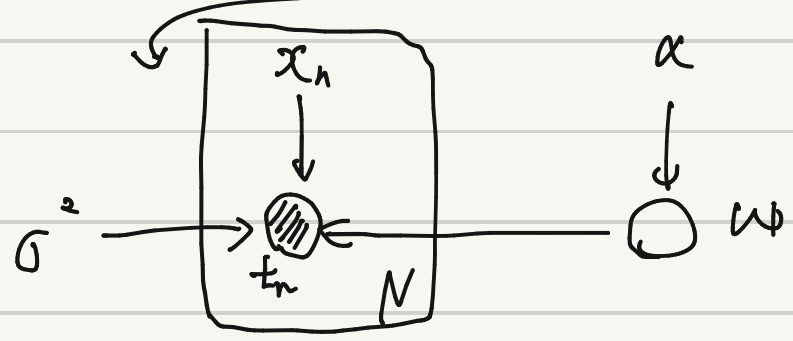
$$p(x_1, \cdots\cdots, x_k) = p(x_k | x_1, \cdots, x_{k-1}) \cdots\cdots p(x_2|x_1)\,p(x_1)$$

8.1.1 Polynomial regression



if we want stochastic variables to be explicit.

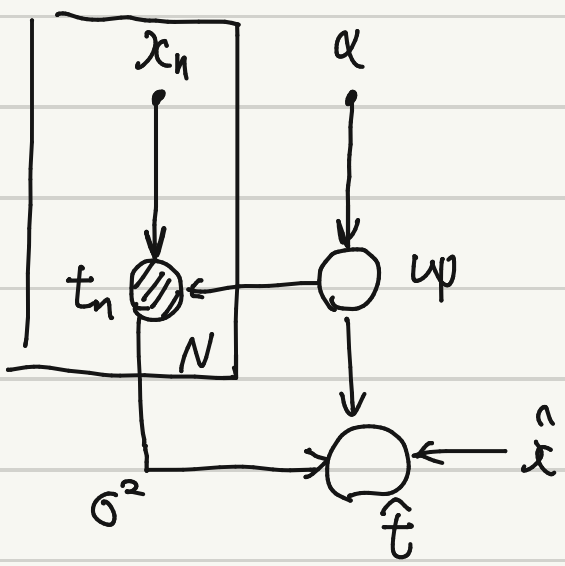
$$p(\mathbf{t}, \mathbf{w} \mid \mathbf{X}, \alpha, \sigma^2) = p(\mathbf{w} \mid \alpha) \prod_{n=1}^{N} p(t_n \mid \mathbf{w}, \mathbf{x}_n, \sigma^2)$$



: represents observed values.

purpose : predict $\hat{t}$

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



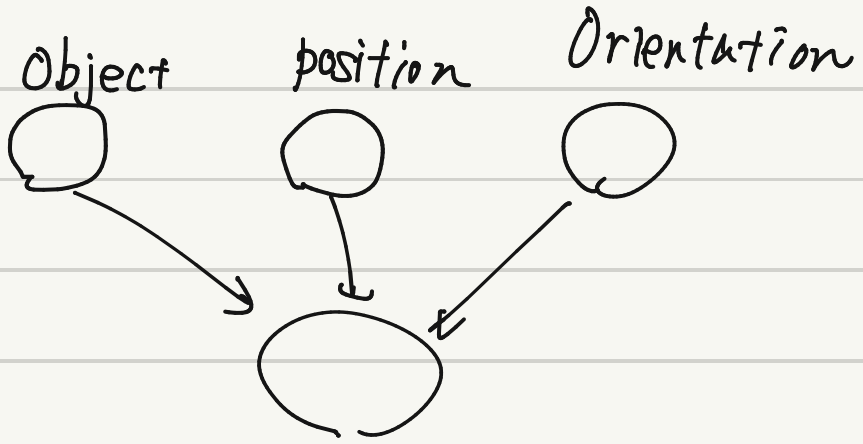
8.1.2 Generative models

Sometimes, we want to draw samples from given probability distribution

↳ (discuss later in Chapter 11)

- Ancestral sampling

  ex) images object creation



- The graphical model : represents the causal process

  ↓

  called "generative models"

8.1.3 Discrete variables

- We want to choose models so that parents and childs are conjugate.

- probability distribution for a single discrete variable x having K possible states

$$p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

(governed by μ) $\left(\sum_k \mu_k = 1\right)$

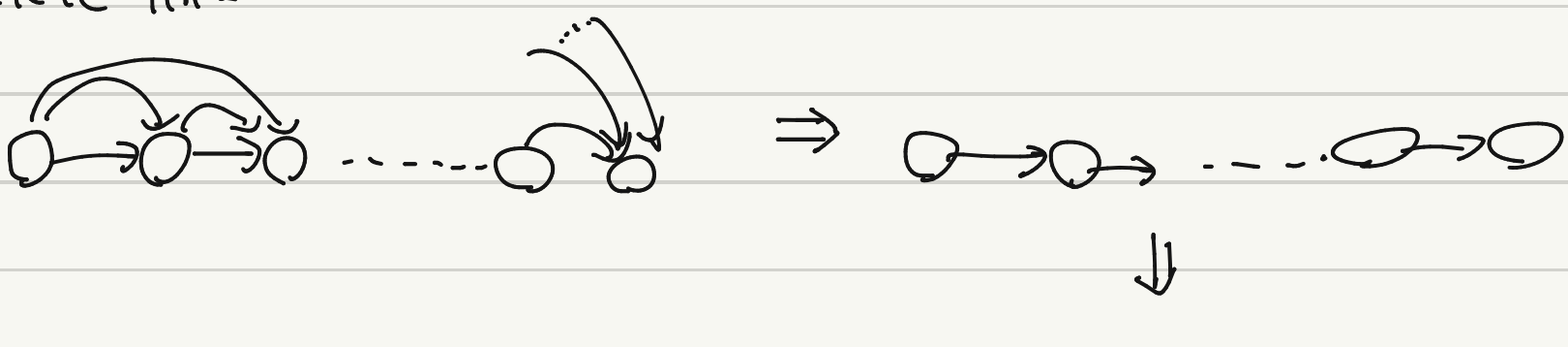
- Suppose that we have M discrete variable, each of which has K states

parameters : $K^M - 1$

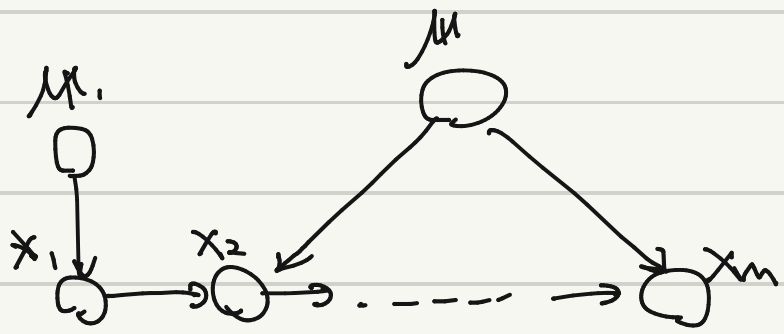
if they were independent, $M(K-1)$ want to reduce.

- How to reduce the number of independent parameters

1. delete link

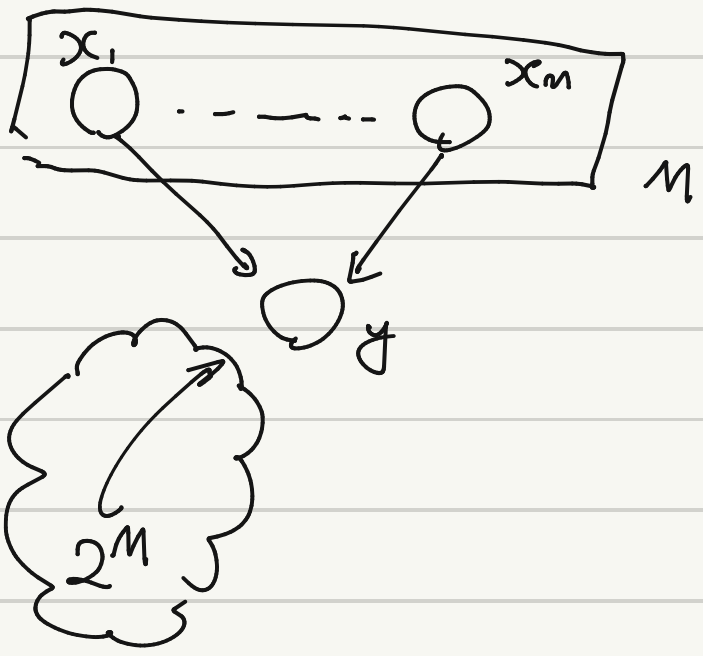


2. Sharing μ.



(share parameters)

3. use parameterized models



2^M

$p(y|x_1 \cdots x_M) = \sigma\left(w_0 \sum_{i=1}^{M} a_i x_i\right)$

$= \sigma(w^T x)$

$M+1$

(w is $(M+1)$ dimensional vector)

8.1.4 Linear-Gaussian models

- how a multivariate Gaussian can be expressed as a graphical model which corresponds to a linear-Gaussian model

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

pa_i : parent node

$$\ln p(\mathbf{x}) = \sum_{i=1}^{D} \ln p(x_i \mid pa_i)$$

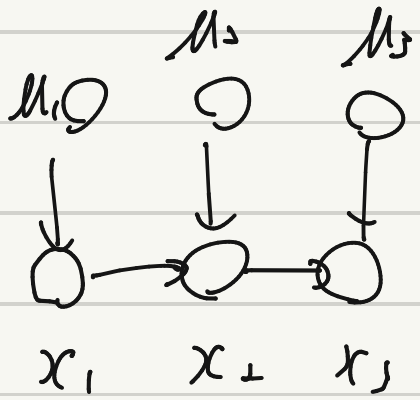
$$= -\sum_{i} \frac{1}{2v_i}\left(x_i - \sum_{j \in pa_i} w_{ij} x_j - b_i\right)^2 + C$$

quadratic function

$\Rightarrow$ $p(x)$ is a multivariate Gaussian.

★

$$\begin{cases} E[x_i] = \sum_{j \in pa_i} w_{ij} E[x_j] + b_i \\ \operatorname{cov}[x_i, x_j] = \sum_{k \in pa_i} w_{jk} \operatorname{cov}[x_i, x_k] + I_{ij} v_j \end{cases}$$



8.2 Conditional Independence

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

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

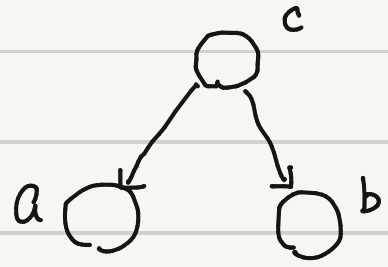
a is conditionally independent of b given c

$a \perp\!\!\!\perp b \mid c \qquad p(a|b,c) = p(a|c)$

- How to examine independency?
  - ↳ judge from the shape of graph.

8.2.1 Three example graphs

1. tail-to-tail



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

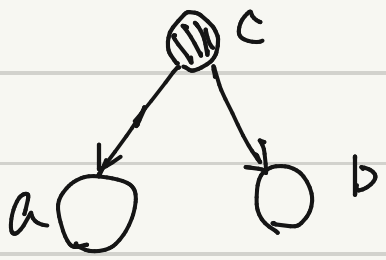
↓ marginalize with respect to c

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

↳ can't factorize into $p(a)\cdot p(b)$

↳ $a \not\perp\!\!\!\perp b \mid \emptyset$

dependence



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

↓ condition on the variable c

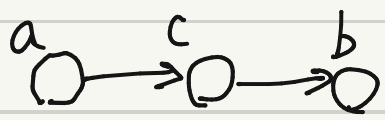
$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = p(a|c)\,p(b|c)$

↳ obtain $a \perp\!\!\!\perp b \mid c$

conditional independence

(if c is observed ...)

2. head-to-tail

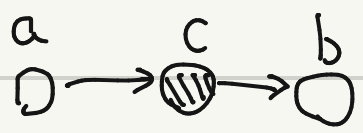


$$p(a,b,c) = p(a)\,p(c|a)\,p(b|c)$$

↓ marginalize

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

↳ $a \perp\!\!\!\perp b \mid \emptyset$



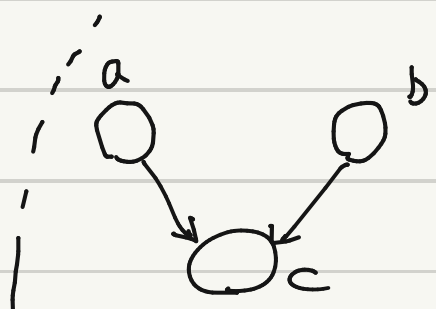
condition on the variable c

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

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

↳ $a \perp\!\!\!\perp b \mid c$

3. head-to-head



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

↓ marginalize

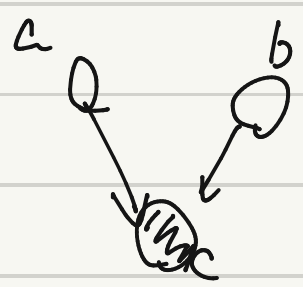
$p(a,b) = p(a)\,p(b)$ → $a \perp\!\!\!\perp b \mid \emptyset$ independent

condition on the variable c

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

↳ $a \not\perp\!\!\!\perp b \mid c$

※ explaining way



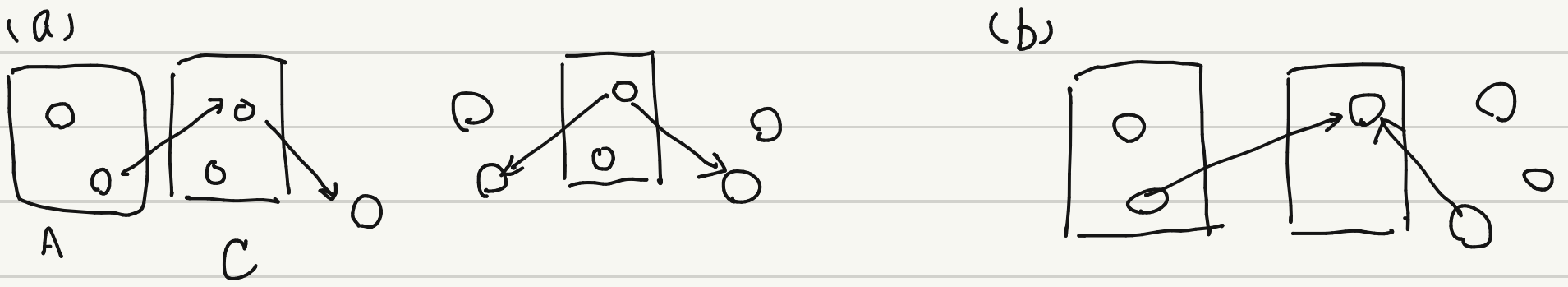
if b is observed, probability of a is changed

8.2.2 D-separation

Any path is said to be blocked if it includes a node such that

(a) the arrows on the path meet either head-to tail or tail-to-tail at the node, and node is in C

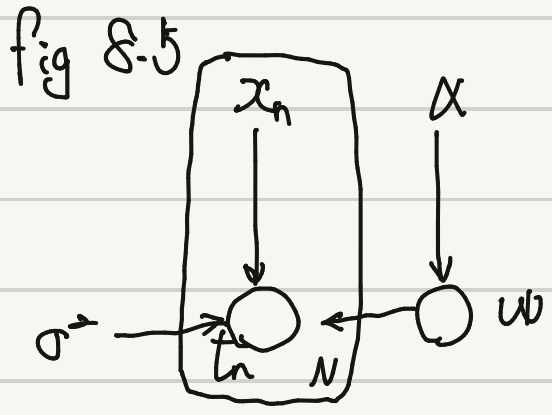
(b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in C



If All paths are blocked, A is said to be d-separated from B by C.

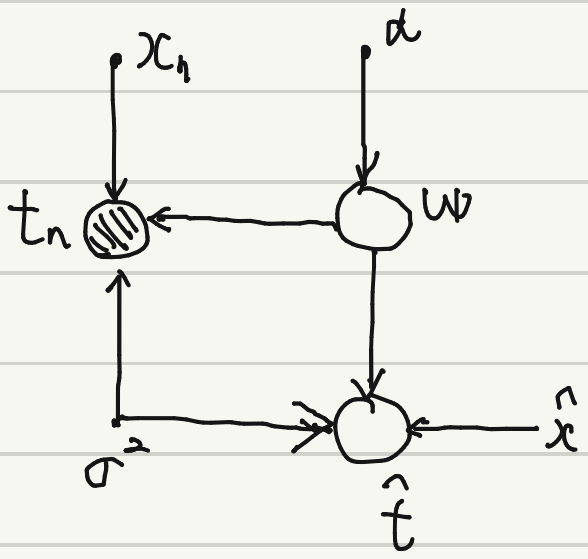
$$A \perp\!\!\!\perp B \mid C$$

fig 8.5



α and σ^2 behave as observed nodes.
are tail to tail.
→ play no role in d-separation.

fig 8.7

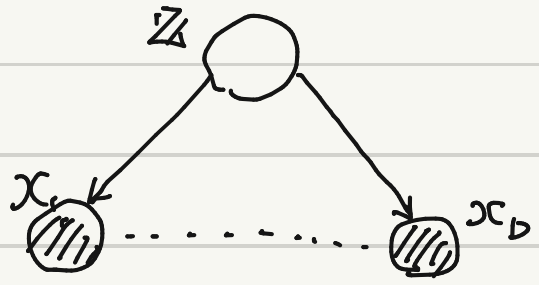


w : tail-to-tail

$\hat{t} \perp\!\!\!\perp t_n \mid w$

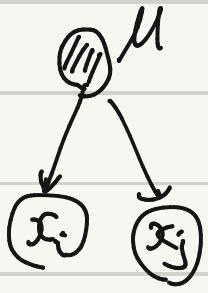
↳ conditioned on the w, $\hat{t}$ is independent of t_n

Naive Bayes



conditioned on z, x are assumed to be independent.

The graph - represents a specific decomposition of a joint probability distribution into a product of conditional probabilities
- expresses a set of conditional independence statements obtained through d-separation criterion.

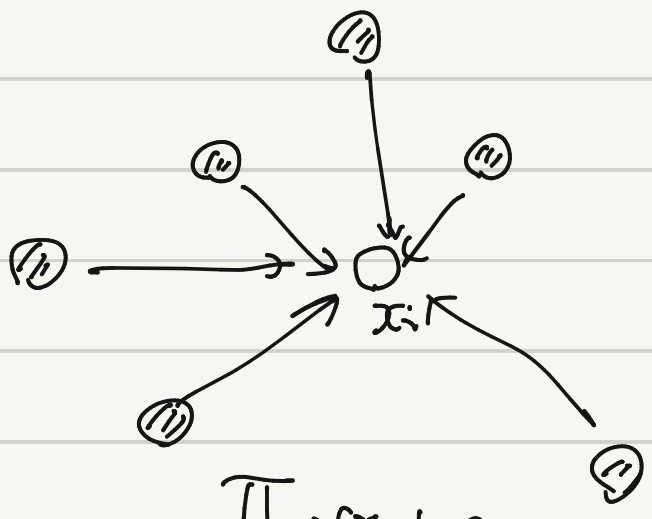


- μ is observed (we already knew μ (probablistic distribation))

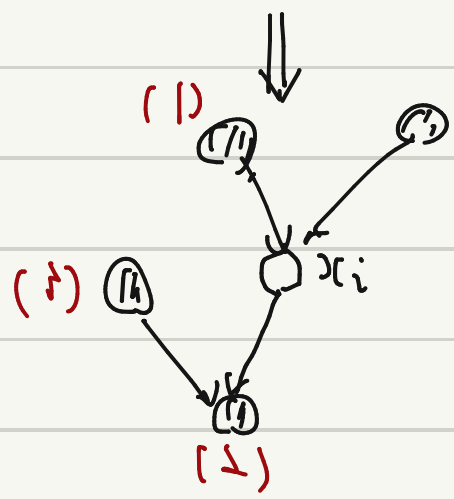
x_i and x_j are i.i.d (same probablistic distribution but independent)

Markov blanket

consider the situation where every node except x_i was observed



$$p(x_i \mid x_{\{j \neq i\}}) = \frac{\prod_k p(x_k \mid pa_k)}{\int \prod_k p(x_k \mid pa_k)\, dx_i}$$

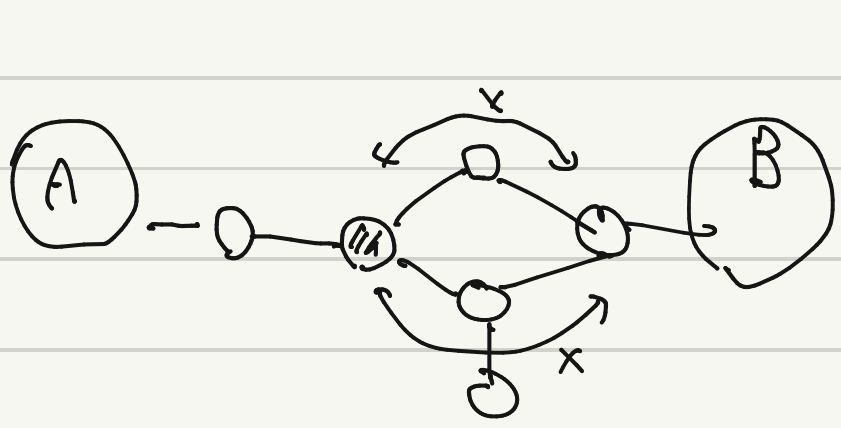


(1) ~ (3) are depend on x_i

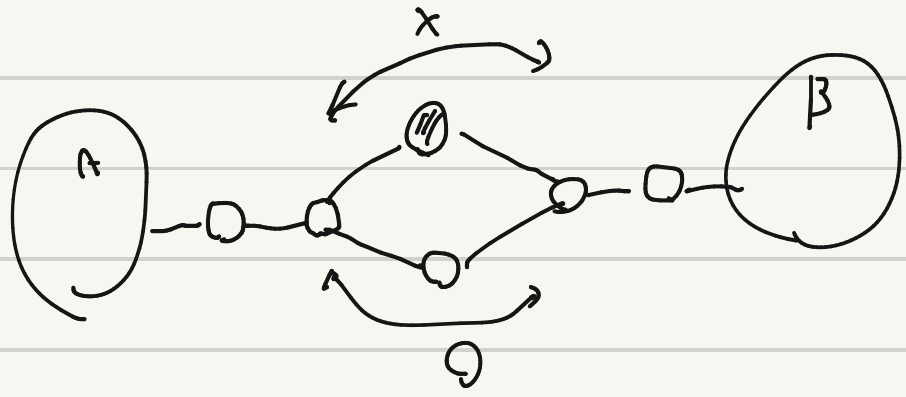
{(1)} + {(2)} + {(3)} + {x_i} = Markov blanket

8.3 Markov Random Fields

8.3.1 Conditional independence properties

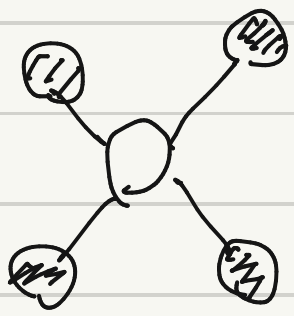


$A \perp\!\!\!\perp B \mid C$



$A \not\perp\!\!\!\perp B \mid C$

Markov blanket



8.3.2 Factorization properties

How to express the joint distribution $p(\mathbf{x})$ as a product of functions defined over sets of variables

Conditional independence properties can be expressed as $p(x_i, x_j \mid \mathbf{x}_{\setminus\{i,j\}}) = p(x_i \mid \mathbf{x}_{\setminus\{i,j\}}) \cdot p(x_j \mid \mathbf{x}_{\setminus\{i,j\}})$

let x_i and x_j not appear in the same factor

clique

a subset of the nodes in a graph such that there exists a link between all pairs of nodes in the subset.

Let C be a clique.

$$p(\mathbf{x}) = \frac{1}{\sum} \prod_{C} \psi_C(\mathbf{x}_C)$$

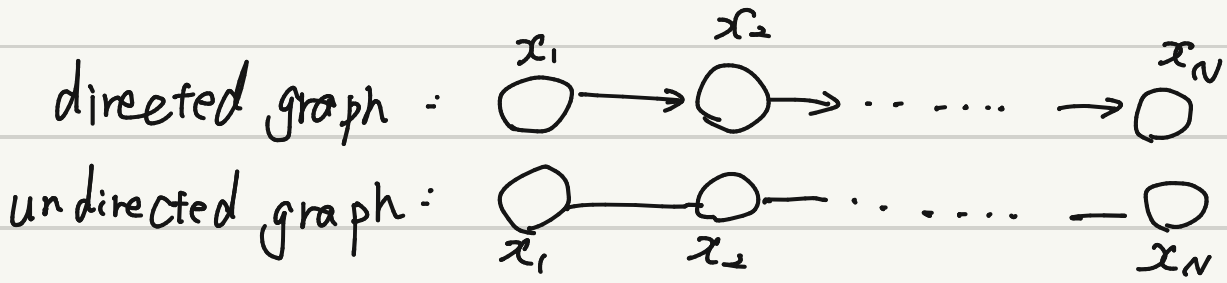
↖ hard to calculate

How can we define ψ_C?

↳ viewing potential function as expressing which configurations of the local variables are preferred to others.

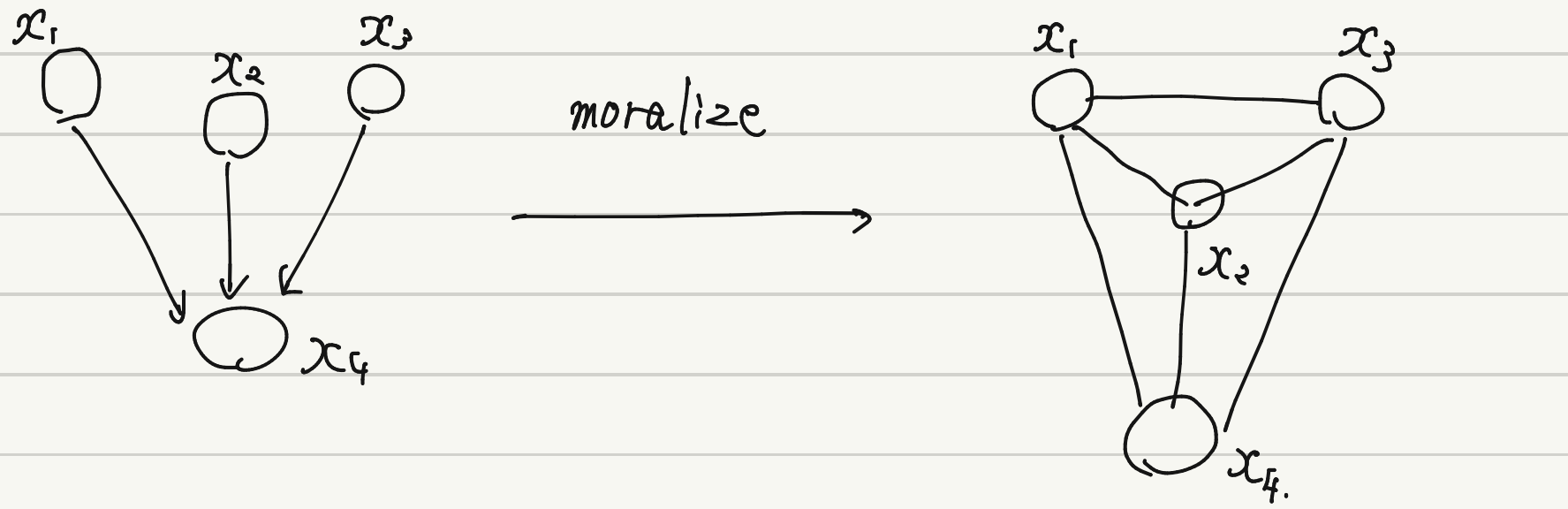
$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$ (cause $\forall \mathbf{x}_C, \psi_C(\mathbf{x}_C) > 0$)

8.3.4 Relation to directed graph



$$p(\mathbf{x}) = p(x_1)\,p(x_2|x_1)\,p(x_3|x_2)\cdots\cdots p(x_N|x_{N-1})$$

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

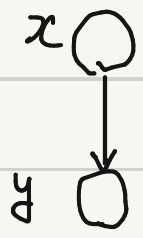


How to convert directed graph into undirected graph.

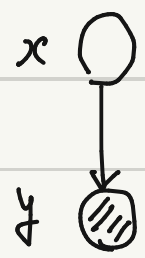
1) add additional undirected links between all pairs of parents
2) drop the arrows
3) initialize all of the clique potentials to 1
4) multiply each conditional distribution factor into one of the clique potentials.

8.4 Inference in Graphical Models

Bayes' theorem



$p(x,y) = p(y|x)\,p(x)$



y is observed

direction change

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

8.4.1 Inference on a chain

Consider chain of nodes, which will lay the foundation for general graphs.

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

Consider the inference problem of finding marginal distribution $p(x_n)$

$$p(x_n) = \sum_{x_1}\cdots\sum_{x_{n-1}}\sum_{x_{n+1}}\cdots\sum_{x_N} p(X). \quad \cdots ②$$

↳ naive implementation : hard to do

↳ graphical model gives us efficient algorithm.

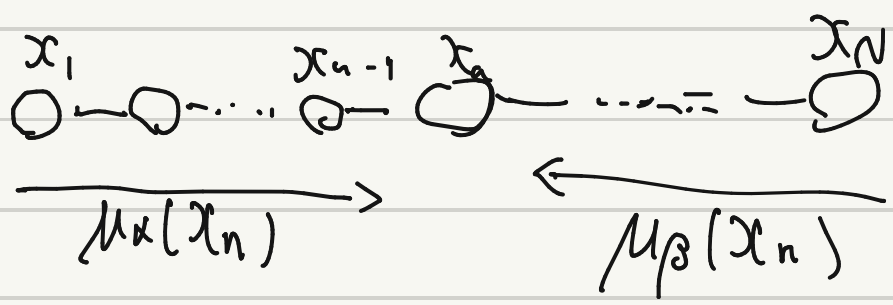
from ① and ②,

$$p(x_n) = \frac{1}{Z}\left[\sum_{x_{n-1}}\psi_{n-1,n}(x_{n-1},x_n)\cdots\left[\sum_{x_2}\psi_{2,3}(x_2,x_3)\left[\sum_{x_1}\psi_{1,2}(x_1,x_2)\right]\cdots\right]\right]$$

$\mu_\alpha(x_n)$

$$\left[\sum_{x_{n+1}}\psi_{n,n+1}(x_n,x_{n+1})\cdots\left[\sum_{x_N}\psi_{N-1,N}(x_{N-1},x_N)\right]\cdots\right]$$

$\mu_\beta(x_n)$



interpret $\mu_\alpha(x_n)$ as a message x_{n-1} to x_n.

$$\mu_\alpha(x_n) = \sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \cdot \mu_\alpha[x_{n-1}]$$

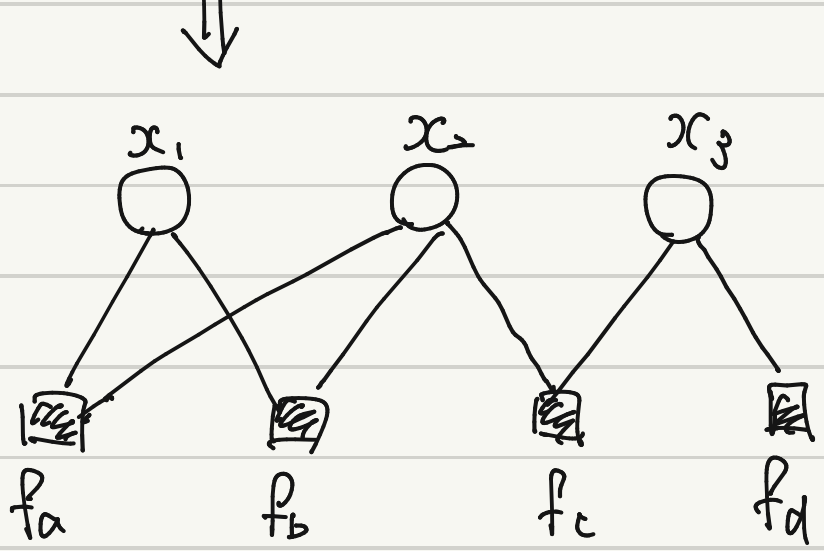
↑ inductive!! → O(K)

message passing can also be applied to tree structure.

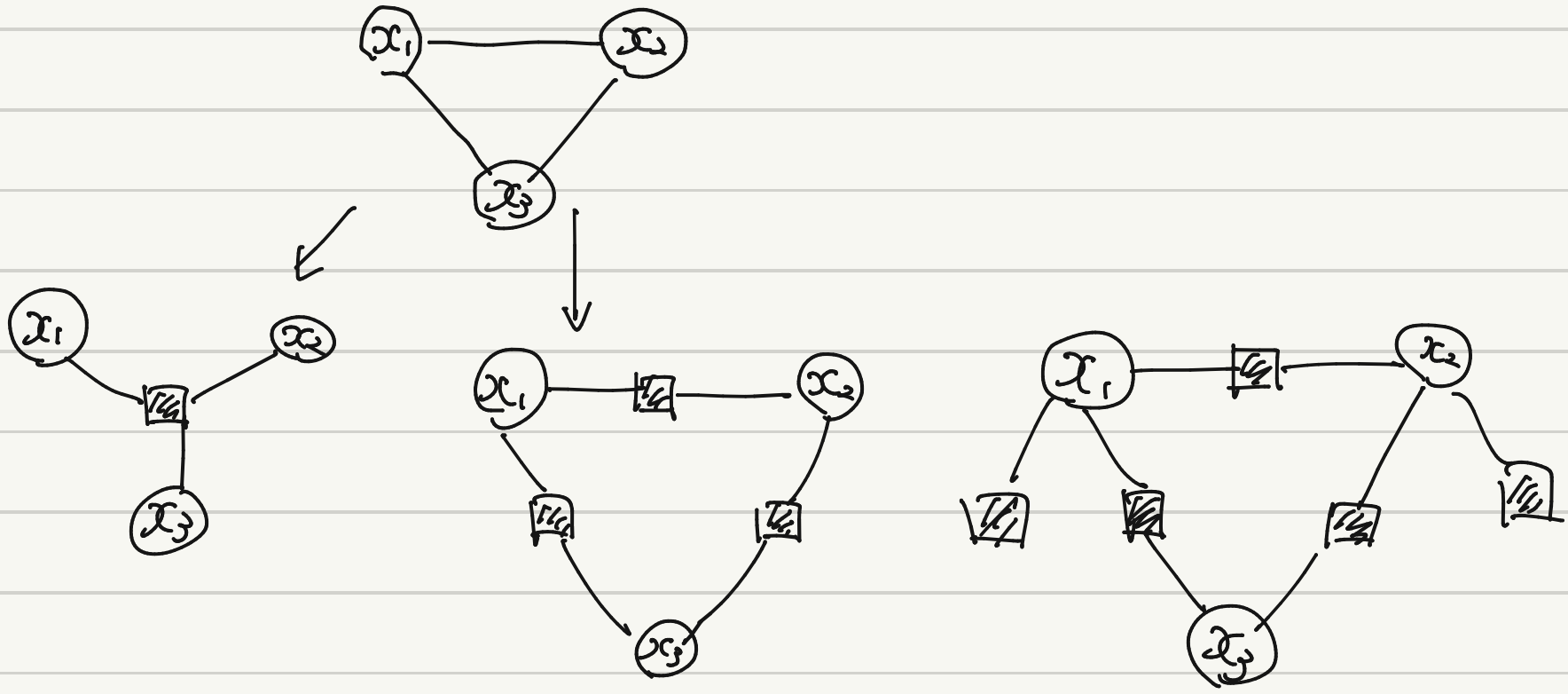
8.4.3 Factor graphs

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

⇓



- factor graph is not unique



- but there is an algorithm that convert undirected graph into factor graph.

8.4.4 The sum-product algorithm

Assume that original graph is an undirected tree or directed tree or polytree.

↳ corresponding factor graph has tree structure.

Goal: (1) to obtain an efficient, exact inference algorithm for finding marginals

(2) in situations where several marginals are required to allow computations to be shared efficiently

• How to evaluate marginal $p(x)$

1. view the variable node x as the root of the factor graph.

2. initiate message at the leaves of graph using

$$\mu_{x \to f(x)} = 1 \quad , \quad \mu_{f \to x}(x) = f(x)$$

3. the messages are passed through

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

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

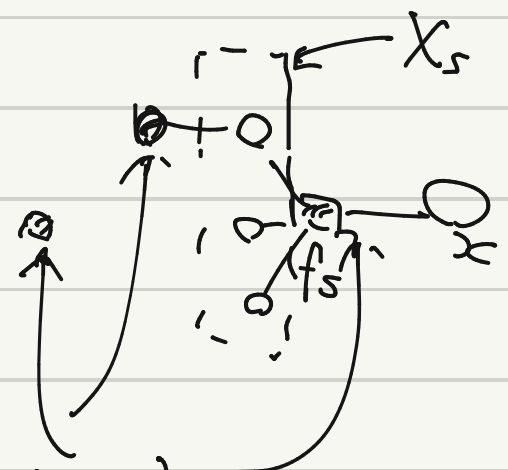
applied recursively until messages have been progated along every link, and the root node has receiveel messages from all of their neighbors.

4. $p(x) = \prod_{s \in ne(x)} \mu_{f_s \to x}(x)$

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

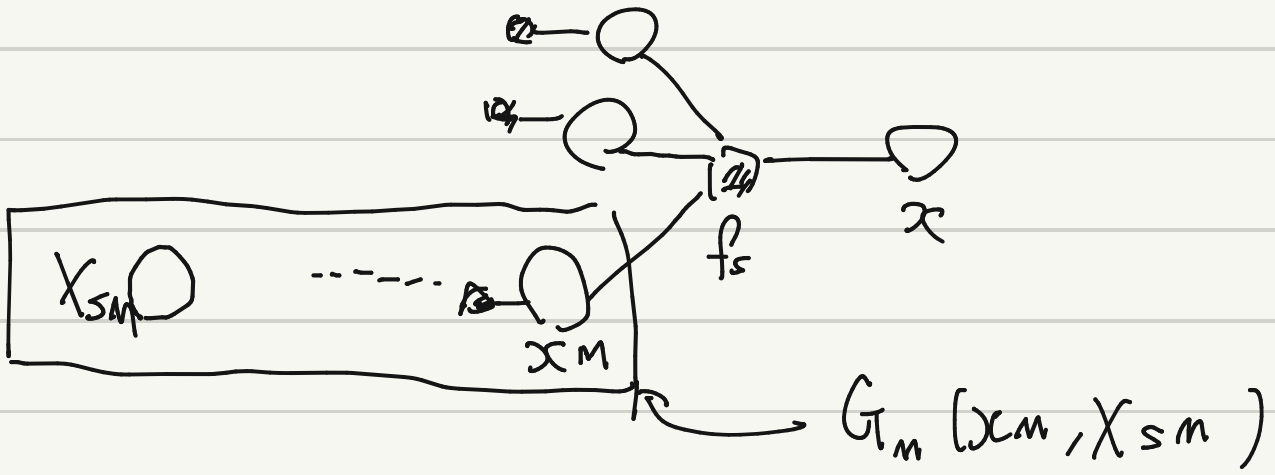
↓

$\prod$ $\sum$ F ~



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

F_s : multiply all factor that has relation with f_s



1.

<base_case> | <step case> ...

$\mu_{x \to f}(x) = 1$ | $\mu_{f_s \to x}(x)$

$\mu_{f \to x}(x) = f(x)$ | $\mu_{x_m \to f_s}(x_m)$

2.

$\Rightarrow F_s(x, X_s)$

$= f_s(x, x_1, \cdots, x_m) G_1(x_1, X_{s1})$

$\cdots \cdot G_m(x_m, X_{sm})$

3.

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

$p(x) = \sum_{X \setminus x} p(X)$

8.4.5 max-sum algorithm

in order { - to find a setting of variables that has the largest prob~
- to find the value of that probability.

$$\max_{\mathbf{x}} p(\mathbf{x}) = \max_{x_1} \cdots \max_{x_M} p(\mathbf{x})$$

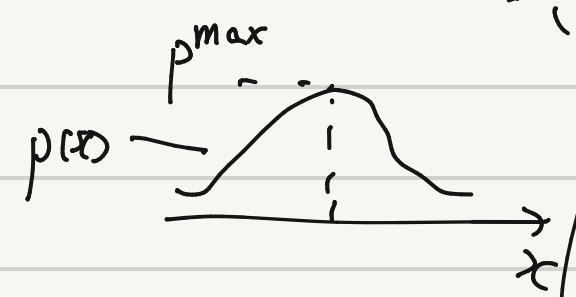
consider the chain nodes.

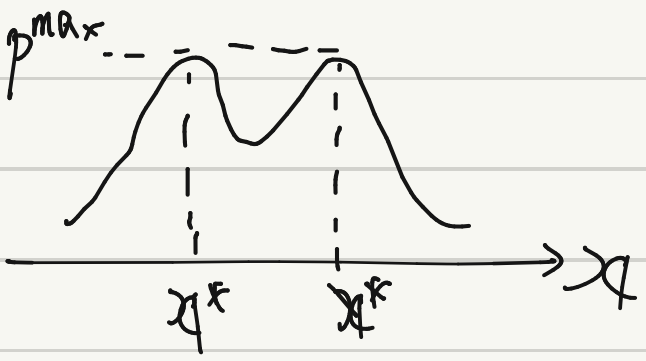
$$\max_{\mathbf{x}} p(\mathbf{x}) = \frac{1}{Z} \max_{x_1} \cdots \max_{x_N} \left[\psi_{1,2}(x_1, x_2) \cdots \psi_{N-1,N}(x_{N-1}, x_N)\right]$$

$$= \frac{1}{Z} \max_{x_1} \left[\max_{x_2} \left[\psi_{1,2}(x_1, x_2) \left[\cdots \max \psi_{N-1,N}(x_{N-1}, x_N)\right] \cdots\right]\right]$$

→ take same process as 8.4.4

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





p^{max} is max value of $p(\mathbf{x})$

$\mathbf{x}^*(x_1^*, x_2^*, \cdots\cdots, x_N^*)$

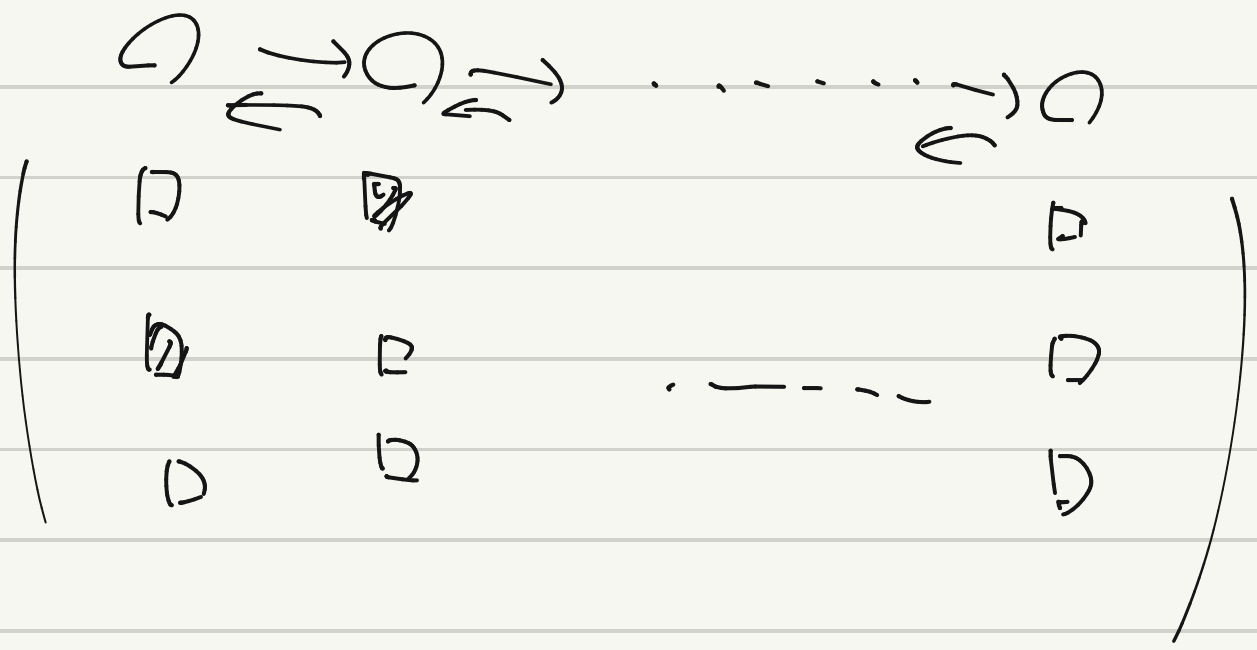
root

↳ memorize (Trellis diagram)

every node except root is not sure.

&

$\mathbf{x}^*$ may not be unique.



$$x_{n-1}^{max} = \phi(x_n^{max}) \quad : \text{backtrack.}$$

$$x_N^{max} = \arg\max \left[\mu_{f_{N-1,N} \to x_N}(x_N) \right]$$

memorize ϕ:

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

↑ x_{n-1} such that maximize a message