Speech and Language Processing

Lecture 4

Bayesian networks and Bayesian inference

Information and Communications Engineering Course

Takahiro Shinozaki

Lecture Plan (Shinozaki's part)

I gives the first 6 lectures about speech recognition. Through these lectures, the backbone of the latest speech recognition techniques is explained.

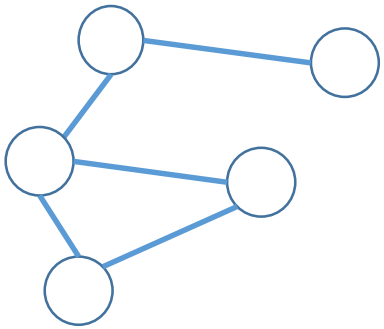
1. 10/4 (remote)
   Introduction and Preparation
2. 10/4 (remote)
   Probability Distributions, Markov Models, Samplings
3. 10/6 (remote)
   Maximum Likelihood Estimation and EM Algorithm
4. 10/6 (remote)
   Bayesian Networks and Bayesian Inference
5. 10/7 (remote)
   Neural networks
6. 10/7 (remote)
   Reinforcement Learning

Bayesian Network

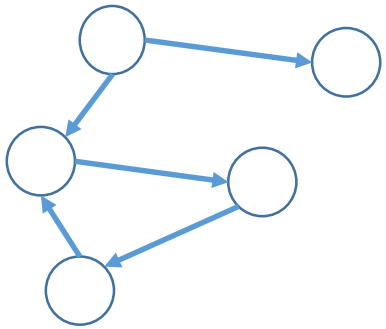
Graphs

- Undirected graph
  - A graph defined by nodes and undirected arcs
- Directed graph
  - A graph defined by nodes and a directed arcs
- Directed Acyclic Graph: DAG
  - Directed graph that does not contain a directed cycle

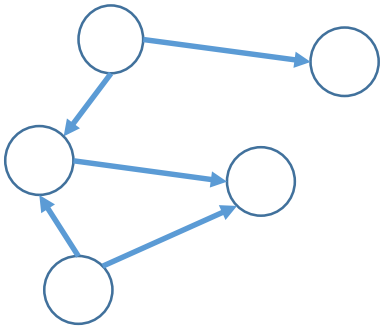
Examples:



Undirected graph



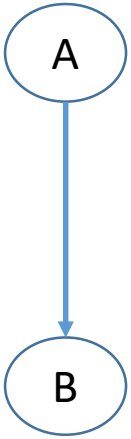
Directed graph
(Have a directed cycle)



Directed acyclic graph

Parent, Child, Ancestor, Descendant

Node A is a parent of node B



Node B is a child of node A

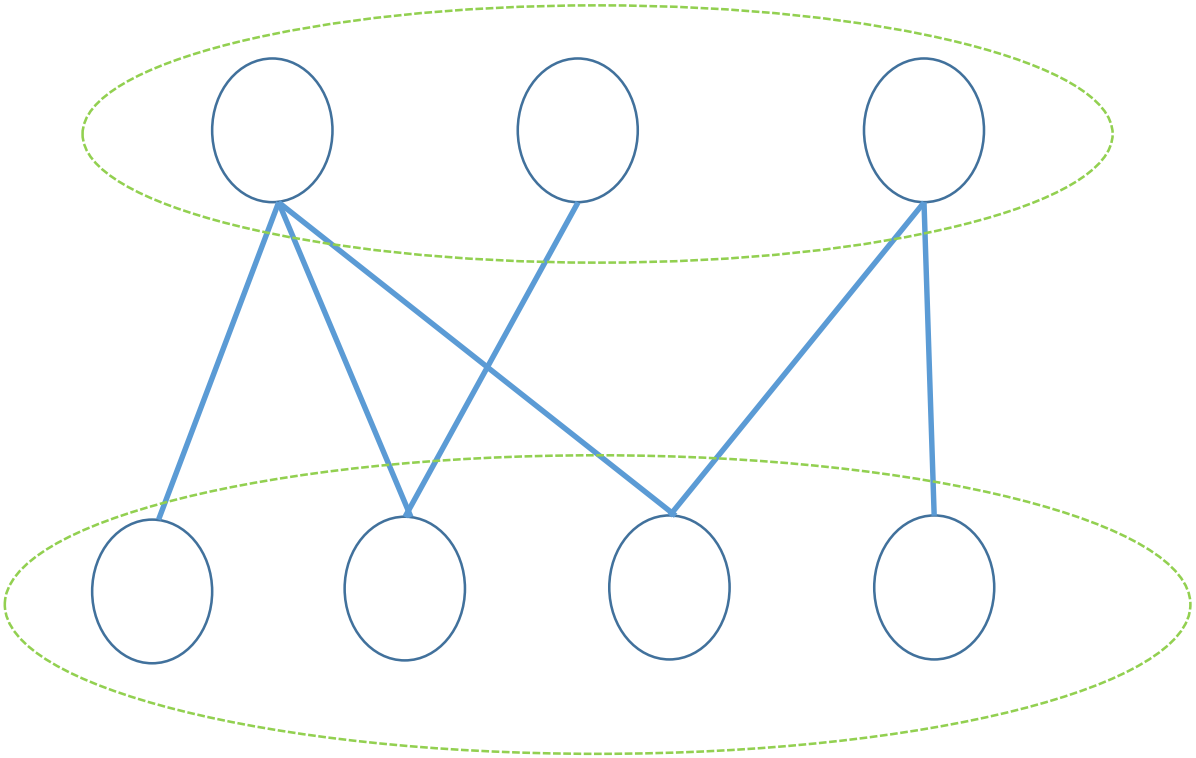
Node A, B, and C are ancestors of node D

Node B, C, and D are descendant of node A

Bipartite

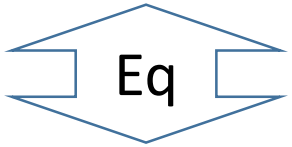
When nodes of a graph are separated to two groups and there is no arc inside the groups, it is called a bipartite

Example of Bipartite:

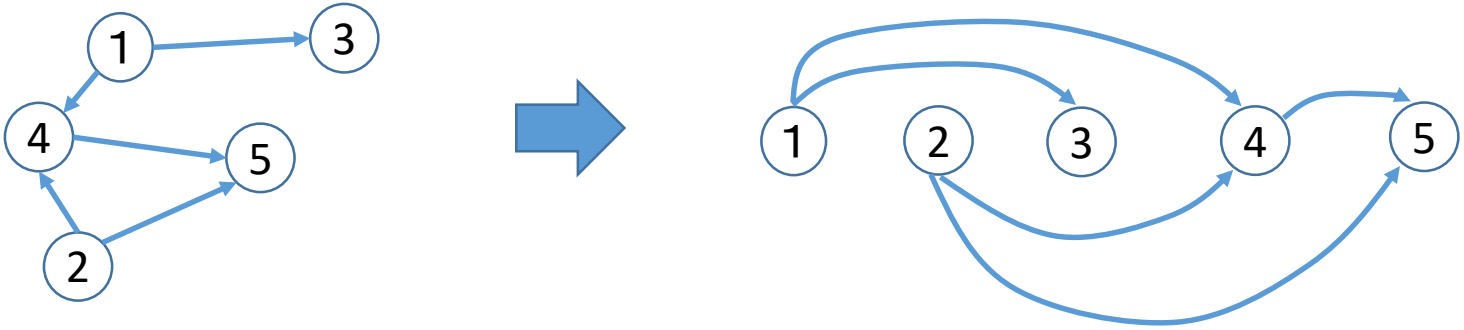


Directed Graph and Node Ordering

A directed graph is a DAG

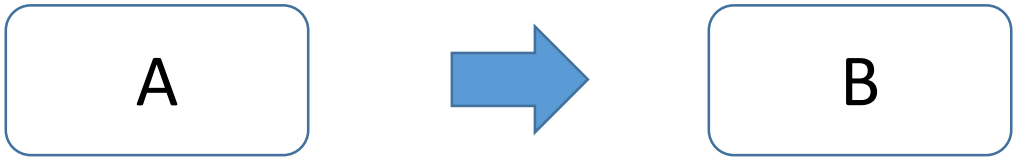


There is a ordering of nodes where all arcs face the same direction (=There is a numbering of nodes where all arcs go from a lower numbered to higher numbered nodes)

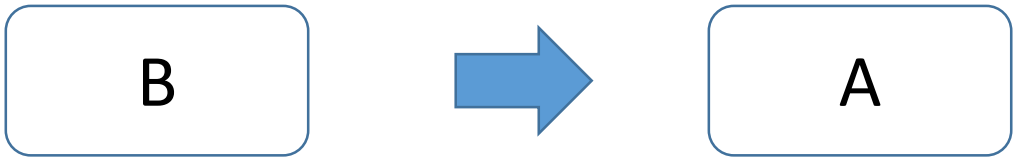


Outline of the Proof

- Statement A:
  - There is a ordering of nodes where all the arcs face the same direction
- Statement B:
  - A graph does not contain a directed cycle



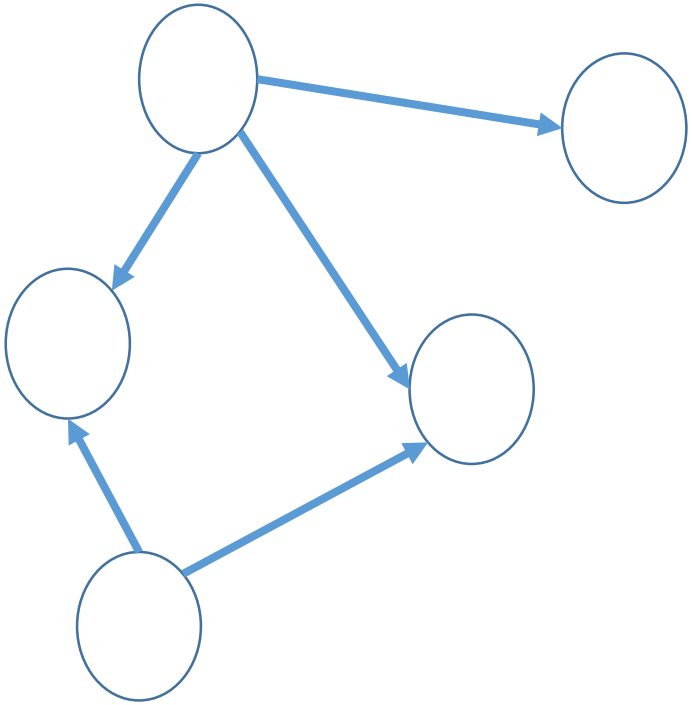
…Easy



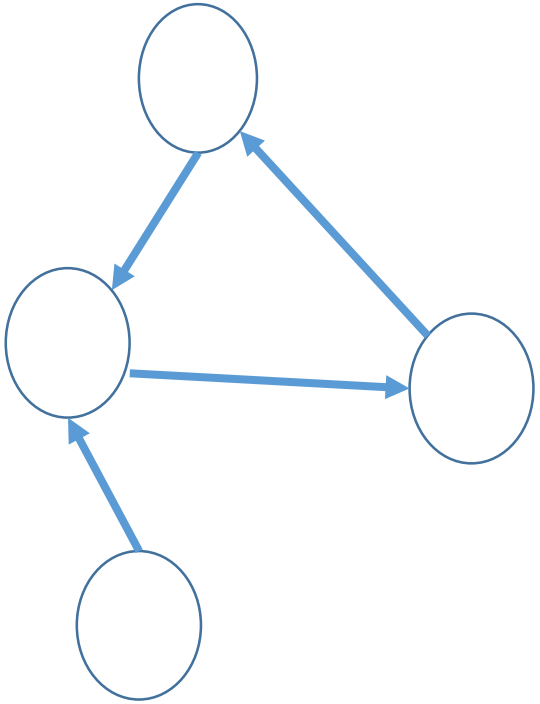
…Use lemma 3.1
(See appendix)

Exercise 4.1

- Is the directed graph a DAG?



Graph A

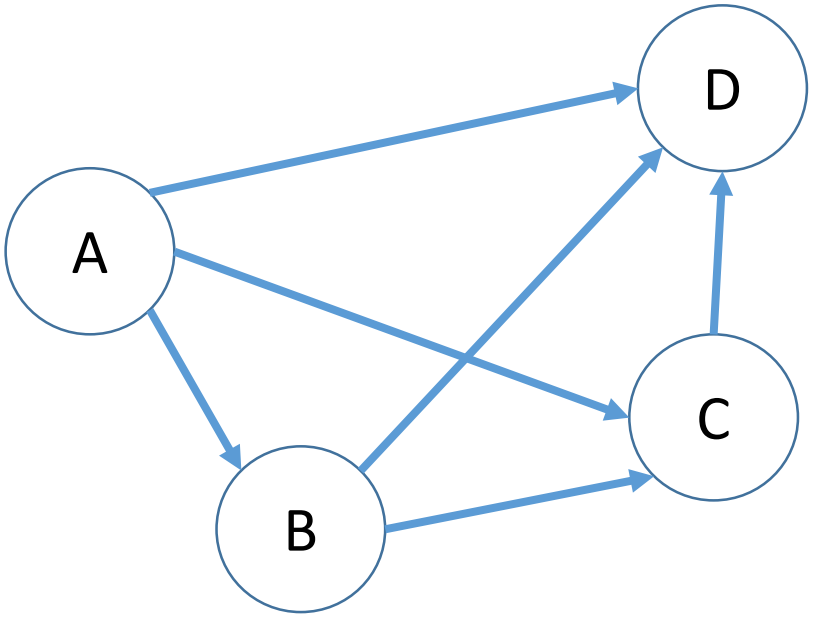


Graph B

Bayesian Network (BN)

BN : Probability theory + DAG

- Nodes represent random variables
- Directed arcs represent dependency

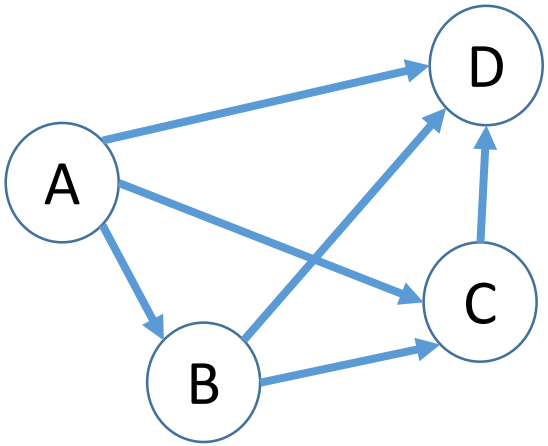


Decomposition of Joint Probability and BN

- By the product rule, arbitrary joint probability is decomposed to a product of conditional probabilities

$$P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$$

- A DAG is made by
  - Representing the variables as nodes
  - Connecting the nodes by directed arcs according the conditional probabilities



Conditional Independence and Arcs

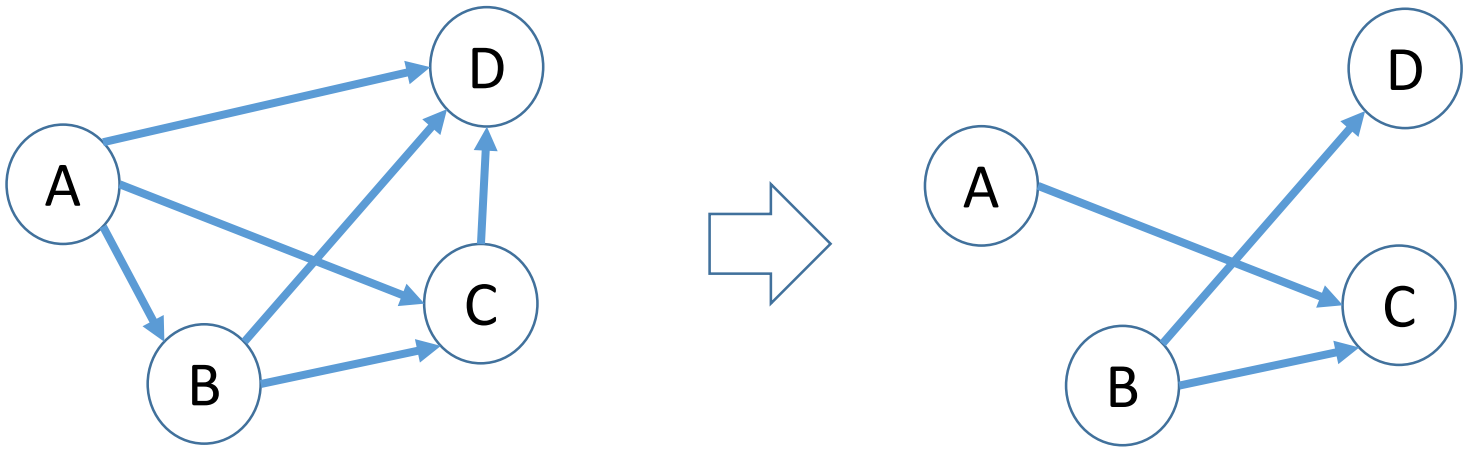
- Conditional independence is represented by **absence** of arcs

$$P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$$

$$P(B \mid A) = P(B)$$

$$P(D \mid A,B,C) = P(D \mid B)$$

$$P(A,B,C,D) = P(A)P(B)P(C \mid A,B)P(D \mid B)$$



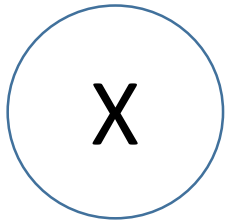
Exercise 4.2

- Represent the following joint probability by a BN

$$P(A,B,C,D,E) = P(A)P(B \mid A)P(C \mid B)P(D \mid B)P(E \mid D)$$

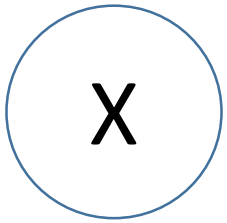
BN Representation of a Categorical Distribution

X	1	2	…	K
$p(X)$	μ_1	μ_2	…	μ_K

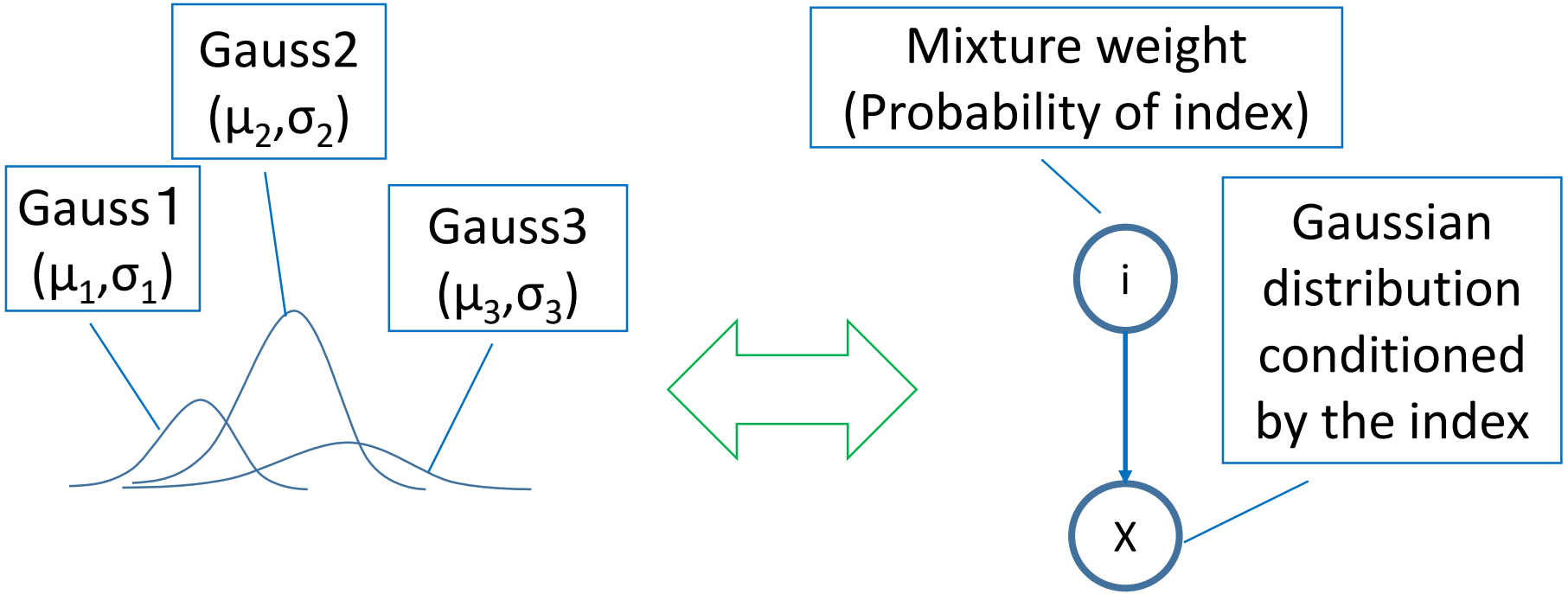


BN Representation of a Gaussian Distribution

$$N\left(X \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(X-\mu\right)^2\right\}$$



BN Representation of a GMM

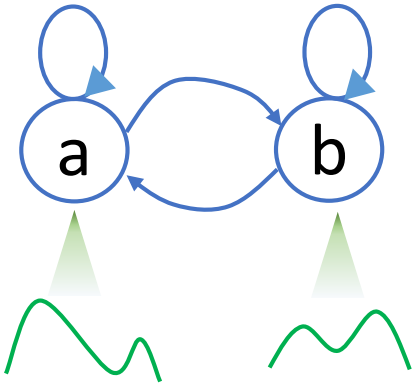


$$P(X) = \sum_i P(i) Gauss(X \mid \mu_i, \sigma_i)$$

$$= \sum_i P(i) P(X \mid i)$$

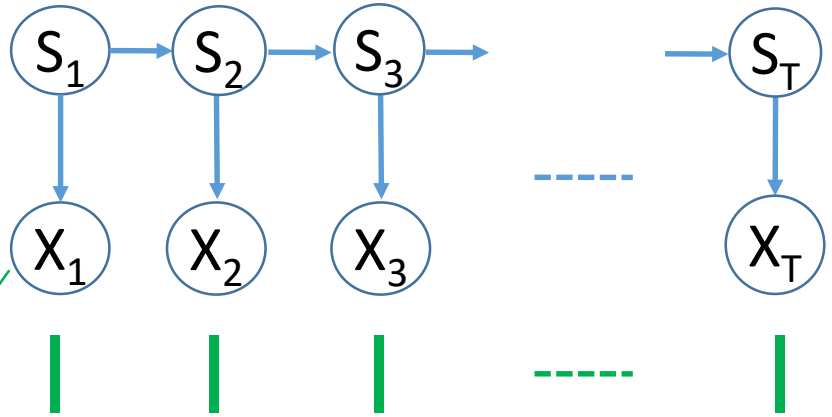
Bayesian Network Representation of a HMM

- The network has unrolled structure
- The length depends on the input sequence



HMM

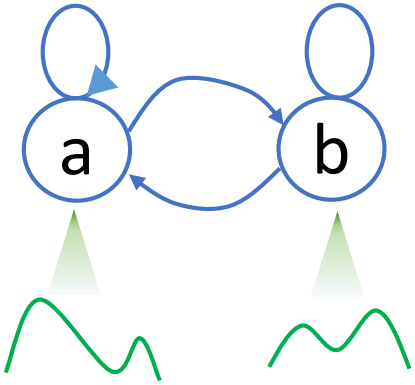
Transition Probability
e.g. $P(S_t=a|S_{t-1}=b)$



Emission distribution
conditioned by the state
e.g. $P(X_t|S_t=a)$

Bayesian network

Example of Alignment

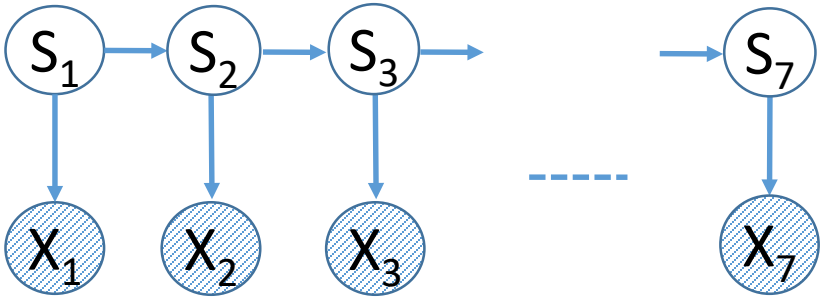


Feature sequence:

$x_1, x_2, x_3, x_4, x_5, x_6, x_7$

State sequence:

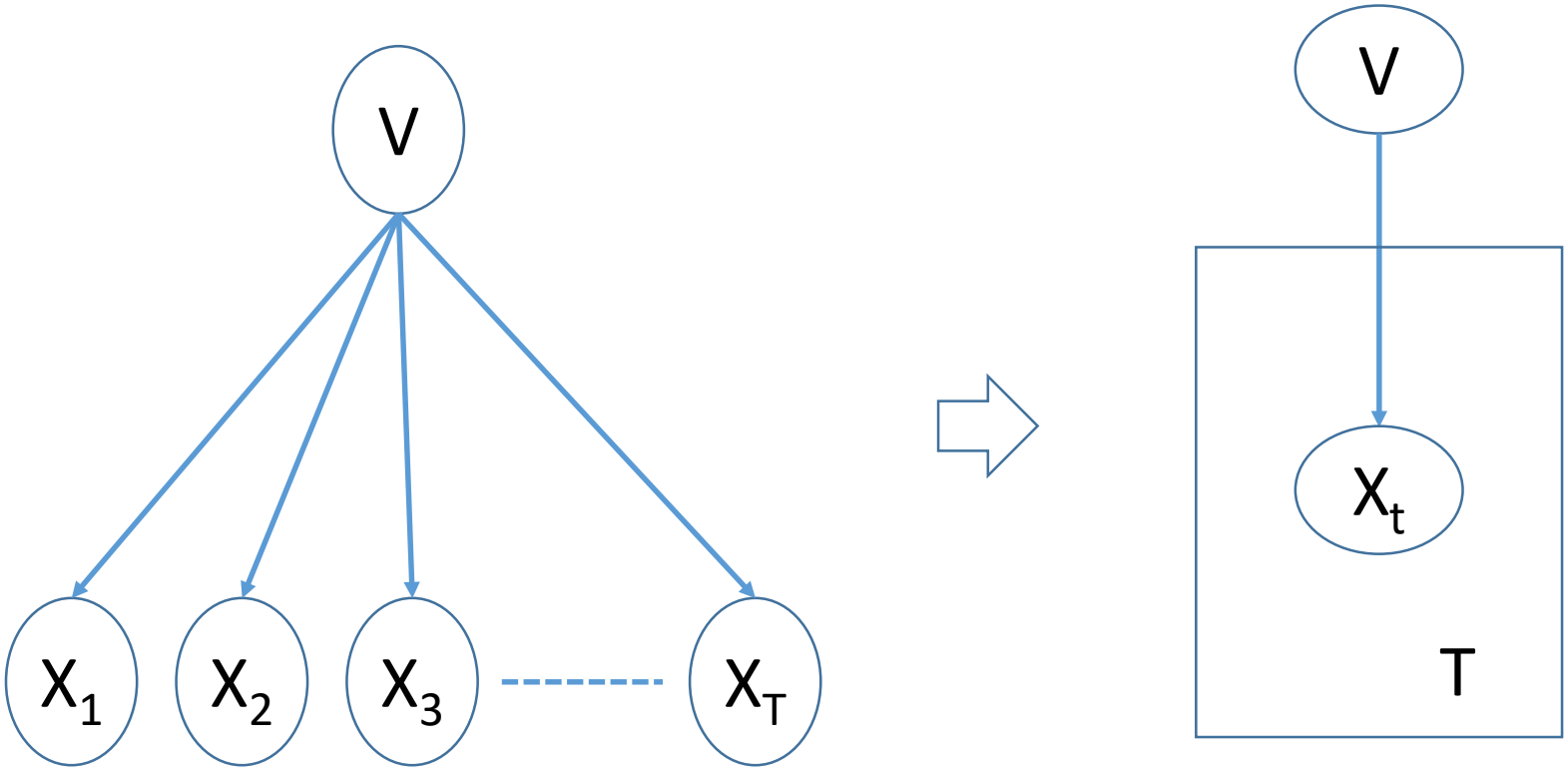
a,b,a,a,a,b,b



$X_1=x_1, X_2=x_2, X_3=x_3, X_4=x_4,$
$X_5=x_5, X_6=x_6, X_7=x_7$

S_1=a, S_2=b, S_3=a, S_4=a,
S_5=a, S_6=b, S_7=b

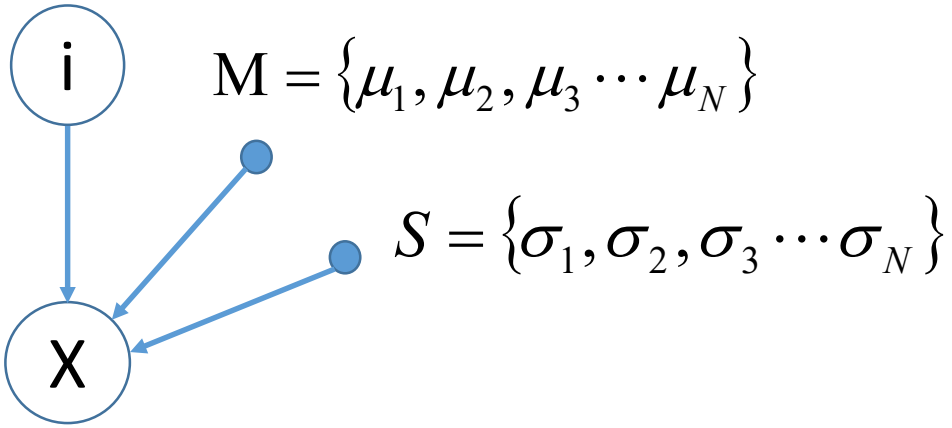
Representation of a Repeated Structure



Representation of Parameters

small circles represent parameters

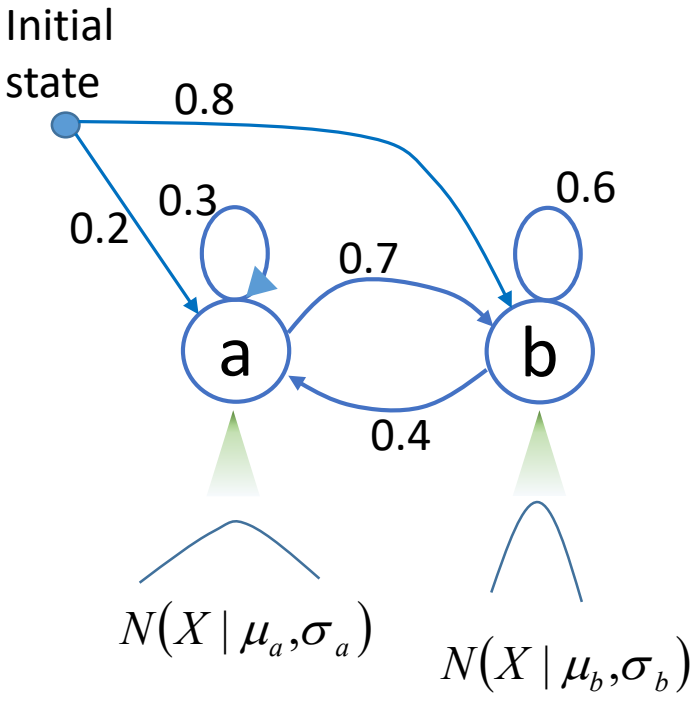
Example of GMM:



$$P(X) = \sum_{i=1}^{N} P(i) Gauss(X \mid \mu_i, \sigma_i)$$

Exercise 4.3

- Fill the blanks so that the following HMM and the BN become equivalent

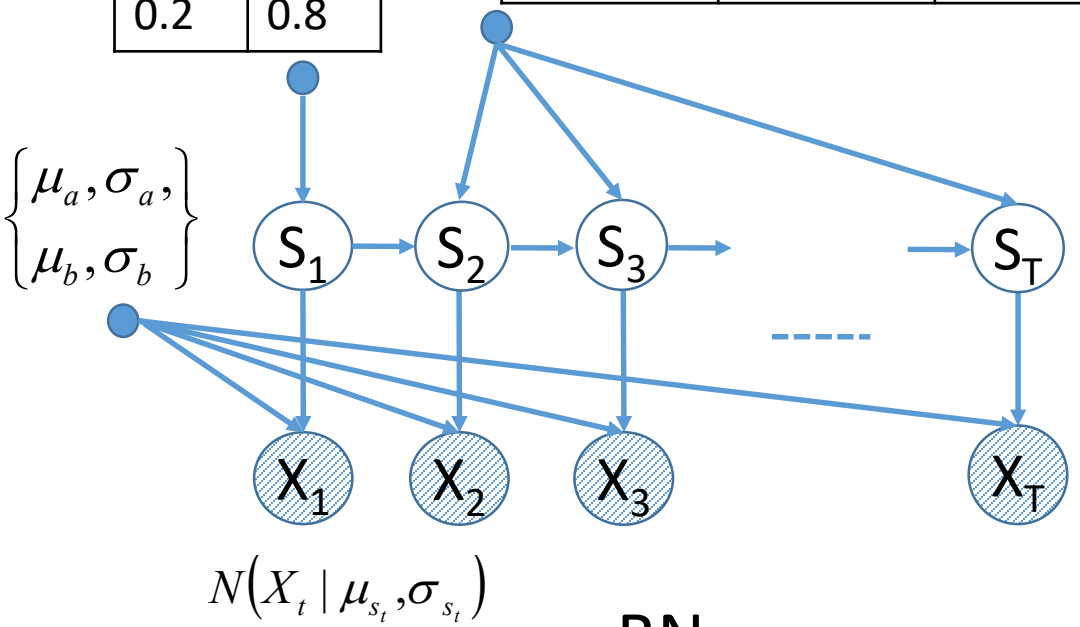


HMM

Initial state probability

P(a)	P(b)
0.2	0.8

$P(S_t \mid S_{t-1})$	S_t =a	S_t =b
S_{t-1} =a	(①)	(②)
S_{t-1} =b	(③)	0.6



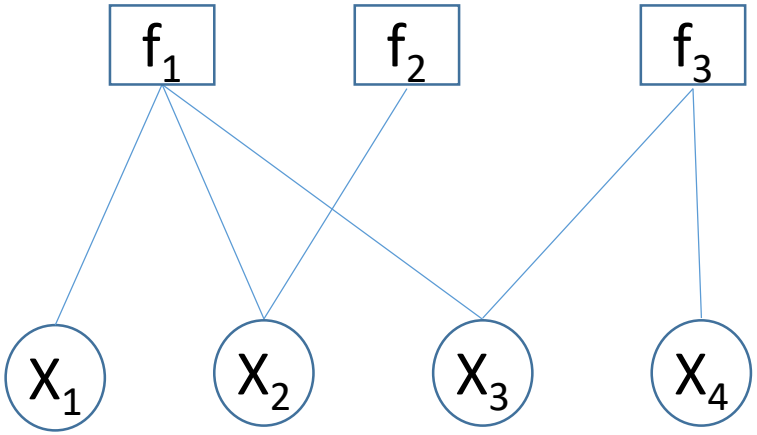
BN

Factor Graph

- A bipartite graph where one side of variables represent random variables and the others represent functions
- The arcs represent dependencies of the functions to the variables
- A factor graph defines a joint probability

$$P(X_1, X_2 \cdots X_N) = \prod_{s \in subsets\ of\ variables} f_s(X_s)$$

Example:



Factor nodes

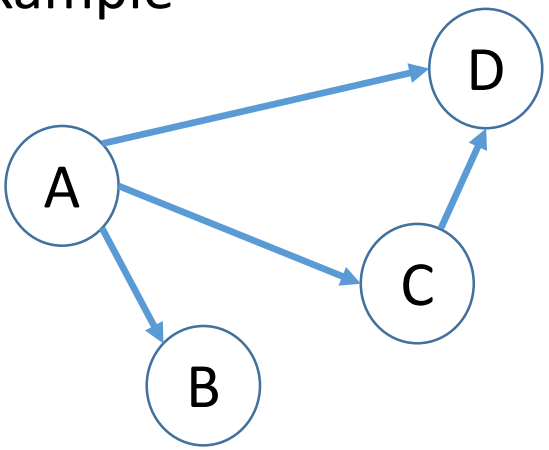
Variable nodes

$$P(X_1, X_2, X_3, X_4) = f_1(X_1, X_2, X_3) f_2(X_2) f_3(X_3, X_4)$$

Factor Graph Representation of Bayesian Network

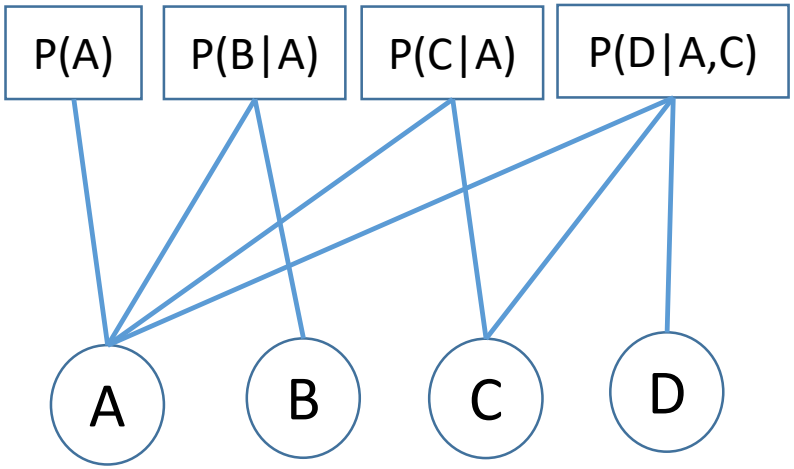
Each conditional probability can be regarded as a factor

Example



$$P(A)P(B \mid A)P(C \mid A)P(D \mid A,C)$$

Bayesian network



Factor graph

Probabilistic Inference

- Marginal and conditional probabilities are obtained from a joint probability by applying the sum and product rules

$$P(A) = \sum_{B,C} P(A,B,C)$$

$$P(A,B) = \sum_{C} P(A,B,C)$$

$$P(A,B,C)$$

$$P(B) = \sum_{A,C} P(A,B,C)$$

$$P(A,B \mid C) = \frac{P(A,B,C)}{\sum_{A,B} P(A,B,C)}$$

$$P(A \mid C) = \frac{\sum_{B} P(A,B,C)}{\sum_{A,B} P(A,B,C)}$$

Distribution Property and Computational Cost

- Product is distributive over addition

$$\sum_{x=1}^{N} af(x) = a\sum_{x=1}^{N} f(x)$$

Number of products : N
Number of summation : N-1

Number of products : 1
Number of summation : N-1

- The same property holds for sum and max, and product and max

$$\max_x [a + f(x)] = a + \max_x [f(x)]$$

$$\max_x [af(x)] = a \max_x [f(x)]$$

Computational Cost of Marginalization

Suppose A, B, C, and D take 1000 possible values

$$P(A) = \sum_{B=1}^{1000} \sum_{C=1}^{1000} \sum_{D=1}^{1000} P(A, B, C, D)$$

summation = $1000^3 = 10^9$

If the joint probability is decomposed to:

$$P(A, B, C, D) = P(A \mid C) P(B) P(C) P(D)$$

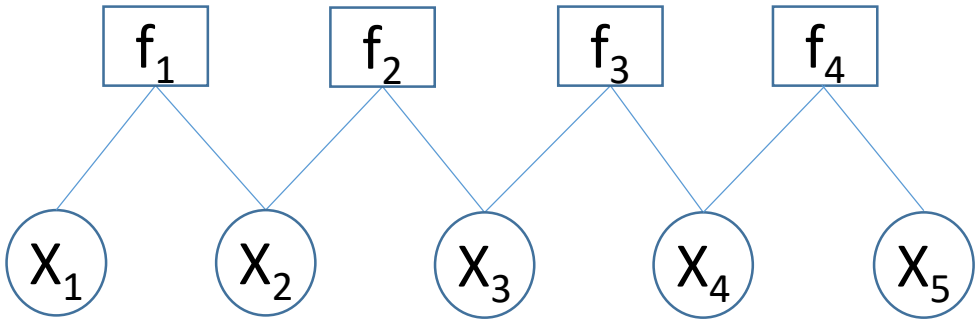
$$P(A) = \sum_{B=1}^{1000} \sum_{C=1}^{1000} \sum_{D=1}^{1000} P(A, B, C, D) = \left(\sum_{C=1}^{1000} P(A \mid C) P(C) \right) \left(\sum_{B=1}^{1000} P(B) \right) \left(\sum_{D=1}^{1000} P(D) \right)$$

summation = 3×10^3

⇒ Independence structure is important

When the Factor Graph is Linear

Suppose we want $P(X_3)$



$$P(X_3) = \sum_{X_1}\sum_{X_2}\sum_{X_4}\sum_{X_5} f_1(X_1,X_2) f_2(X_2,X_3) f_3(X_3,X_4) f_4(X_4,X_5)$$

$$= \left\{\sum_{X_1}\sum_{X_2} f_1(X_1,X_2) f_2(X_2,X_3)\right\}\left\{\sum_{X_4}\sum_{X_5} f_3(X_3,X_4) f_4(X_4,X_5)\right\}$$

$$= \left\{\sum_{X_2}\left\{f_2(X_2,X_3)\sum_{X_1}\{f_1(X_1,X_2)\cdot 1\}\right\}\right\}\left\{\sum_{X_4}\left\{f_3(X_3,X_4)\sum_{X_5}\{f_4(X_4,X_5)\cdot 1\}\right\}\right\}$$

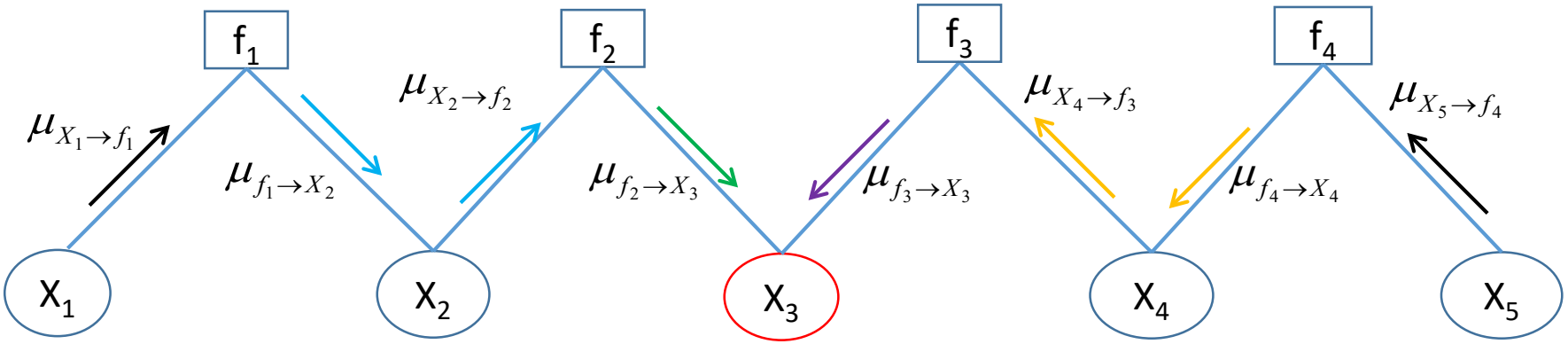
Message Passing View of the Inference

$$P(X_3)=\left\{\sum_{X_2}\left\{f_2(X_2,X_3)\sum_{X_1}\left\{f_1(X_1,X_2)\cdot 1\right\}\right\}\right\}\left\{\sum_{X_4}\left\{f_3(X_3,X_4)\sum_{X_5}\left\{f_4(X_4,X_5)\cdot 1\right\}\right\}\right\}$$

$\mu_{X_1\to f_1}$ $\mu_{X_5\to f_4}$

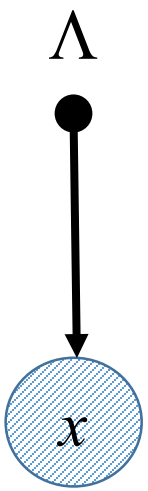
$\mu_{f_1\to X_2} = \mu_{X_2\to f_2}$ $\mu_{f_4\to X_4} = \mu_{X_4\to f_3}$

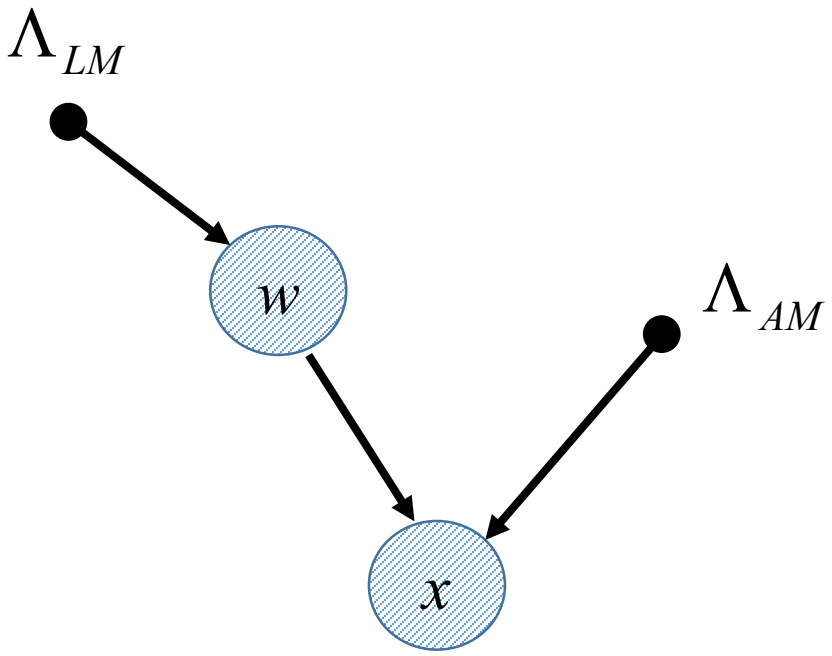
$\mu_{f_2\to X_3}$ $\mu_{f_3\to X_3}$



Bayesian Inference

Probabilistic Models and Their Parameters





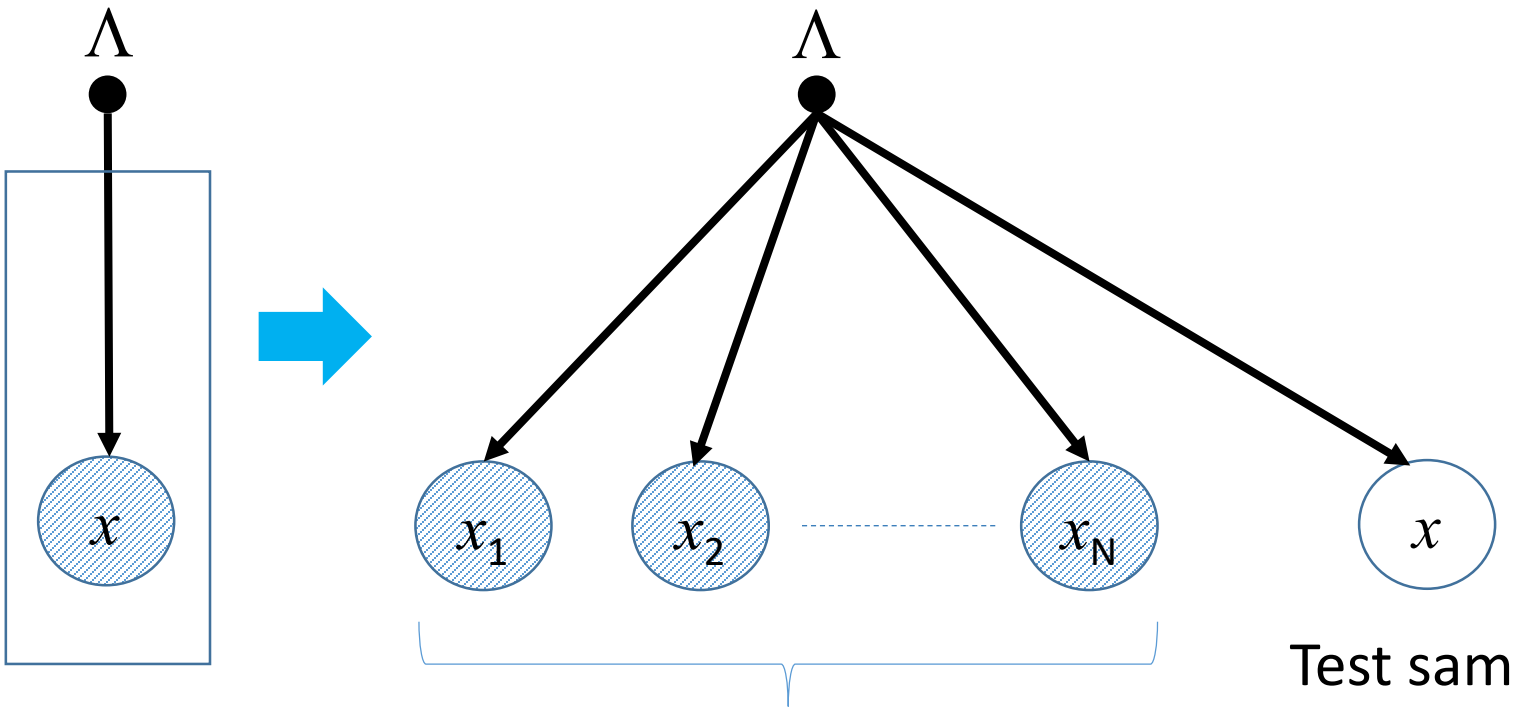
$$p(x \mid \Lambda)$$

$$p(x, w) = p(x \mid w, \Lambda_{AM}) p(w \mid \Lambda_{LM})$$

Gaussian distribution, Multinomial distribution, etc.

Speech model consisting of language and acoustic models

ML Training and Prediction



$$\Lambda^* = \arg\max_{\Lambda} p(D \mid \Lambda) = \arg\max_{\Lambda} \prod_n p(x_n \mid \Lambda)$$

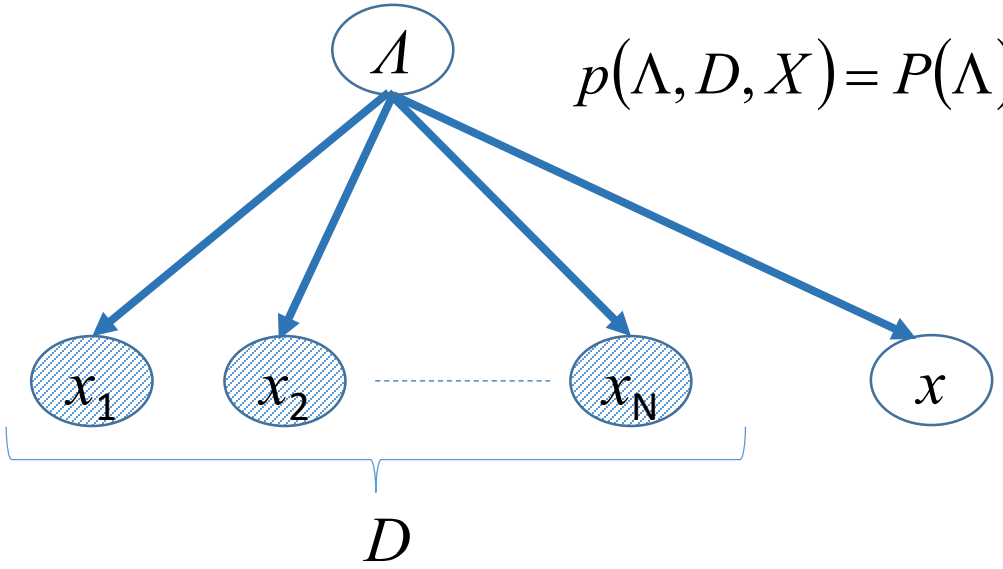
Maximum likelihood (ML) training

$$p(x \mid \Lambda^*)$$

Prediction

Bayesian Approach

- Treat parameters as random variables



$$p(\Lambda, D, X) = P(\Lambda)P(D \mid \Lambda)P(X \mid \Lambda)$$

Prediction of a new sample is formulated as an evaluation of conditional probability given a training set

$$p(x \mid D) = \frac{\int_{\Lambda} p(x, D, \Lambda)}{p(D)} = \int_{\Lambda} p(x \mid \Lambda)\frac{p(D \mid \Lambda)p(\Lambda)}{p(D)} = \int_{\Lambda} p(x \mid \Lambda)p(\Lambda \mid D)$$

Definitions of Terms

- A priori distribution of parameters

$$p(\Lambda)$$

- Probabilistic model

$$p(x \mid \Lambda)$$

- A posteriori distribution of parameters

$$p(\Lambda \mid D) = \frac{p(D \mid \Lambda) p(\Lambda)}{p(D)}$$

- Predictive distribution

$$p(x \mid D) = \int_{\Lambda} p(x \mid \Lambda) p(\Lambda \mid D)$$

Evaluation of A Posteriori Distribution

- Except for very simple models, how to evaluate the a posteriori distribution is a big issue since it requires integrations over many variables

$$p(\Lambda \mid D) = \frac{p(D \mid \Lambda)p(\Lambda)}{p(D)}$$

Approaches

- Analytical evaluation
  - Ideal, but only applicable for very simple models
  - For practical models, closed form solution is usually not obtained. Numerical integration is also not feasible when there are many variables
- Variational Bayes
  - Can be applied to large models if proper analytical approximation is introduced
- Sampling
  - Versatile, but requires very large computational cost

Conjugate Prior

- For some combinations of prior and probabilistic model, posterior takes the same functional form as the prior

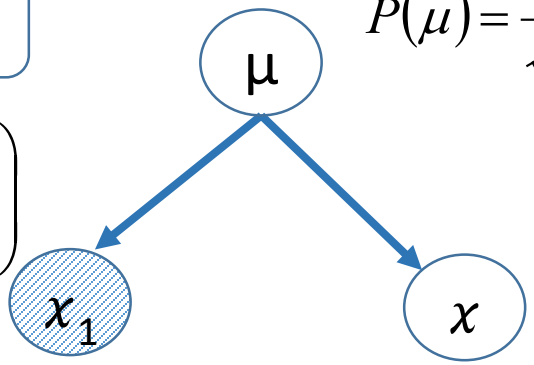
Probabilistic model	Conjugate prior
Binomial distribution	Beta distribution
Multinomial distribution	Dirichlet distribution
Gaussian distribution	Mean: Gaussian distribution Variance: Gamma distribution

Exercise 4.4

- Assumes a probabilistic model P($x|\mu$), a training sample x_1, and a prior distribution of a parameter P(μ) are given as follows.

Gaussian distribution with mean μ and variance 1

$$P(x \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$



$$P(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2}\right)$$

Gaussian distribution with mean 0 and variance 1

1) Estimate posterior distribution

$$P(\mu \mid x_1) = \frac{P(x_1 \mid \mu)P(\mu)}{P(x_1)}$$

Note: $\int_{-\infty}^{\infty} \exp(-cx^2)dx = \sqrt{\frac{\pi}{c}}$

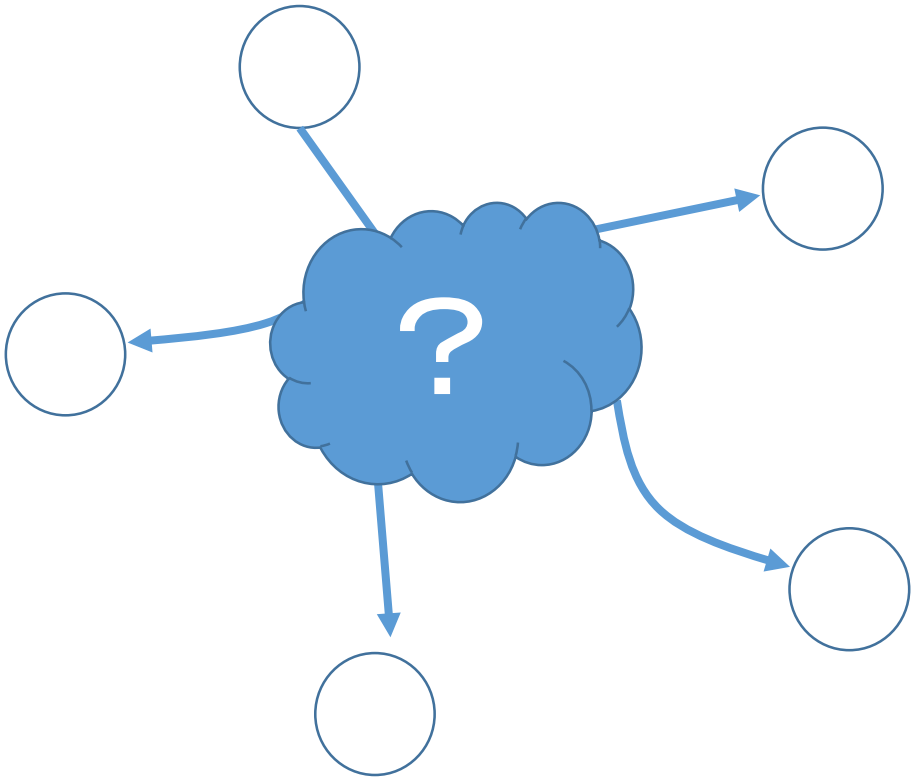
2) Estimate predictive distribution

$$P(x \mid x_1) = \int_{-\infty}^{\infty} P(x \mid \mu)P(\mu \mid x_1)d\mu$$

Appendix

Lemma 4.1

- If a graph does not contain a directed cycle, then there exist at least one node that has no incoming arc



Joint Probability Defined by BN

- Product of conditional probabilities associated with DAG always satisfy the sum-to-one constraint

Proof:

Since a Bayesian network is a DAG, with a proper ordering of the variables, the product has the following form

$$\prod_{i=1}^{N} P(X_i \mid C_i), \quad C_i \subseteq \{X_1, X_2, \cdots, X_{i-1}\}$$

That is, X_i does not appear in the conditional part of X_1 ,…, $X_{i\text{-}1}$. By thinking the summation of the following order, we have:

$$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_N}\prod_{i=1}^{N} P(X_i \mid C_i) = \sum_{X_1}\cdots\sum_{X_{N-1}} P(X_{N-1} \mid C_{N-1}) \sum_{X_N} P(X_N \mid C_N) = 1$$

Notation for Conditional Independence

- Let A,B, and C be disjoint sets of random variables. When the following equation holds, we say that A is independent of B given C, and denote it as $A \perp\!\!\!\perp B \mid C$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C) \quad \Leftrightarrow \quad A \perp\!\!\!\perp B \mid C$$

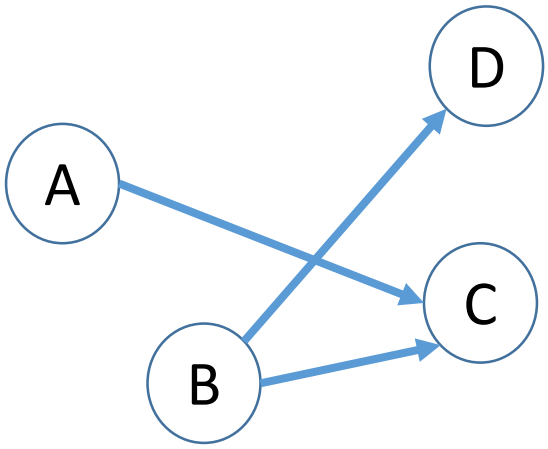
Note:

$$A \perp\!\!\!\perp B \mid C \quad \Leftrightarrow \quad P(A \mid B, C) = P(A \mid C)$$

$$\because P(A, B \mid C) = P(A \mid B, C)P(B \mid C)$$

Graph Structure and Conditional Independence

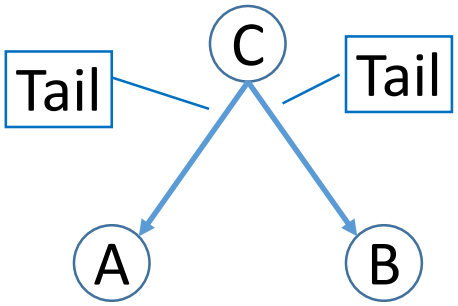
By investigating the graph structure, we can read relationships between random variables



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

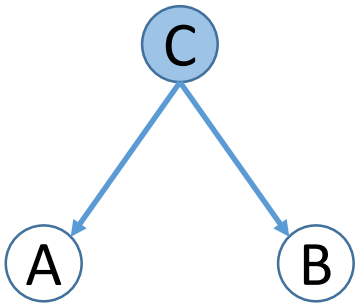
$A \perp\!\!\!\perp D \mid C,B$?

Tail-To-Tail



In general, P(A,B) is not expressed as P(A)P(B). Therefor, $A \not\perp\!\!\!\perp B \mid \Phi$ does not hold. （Φ is an empty set）

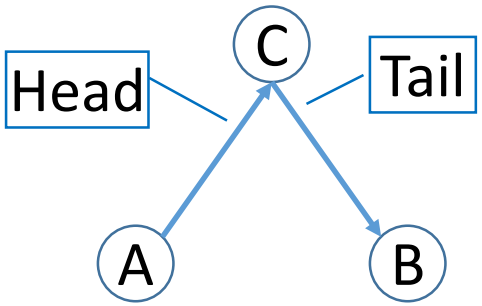
$$P(A,B) = \sum_C P(A,B,C) = \sum_C P(A \mid C)P(B \mid C)P(C)$$



P(A,B|C) is expressed as P(A|C)P(B|C). Therefore $A \perp\!\!\!\perp B \mid C$ holds.

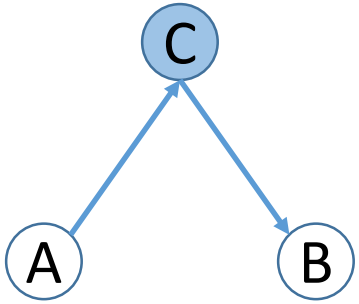
$$P(A,B \mid C) = \frac{P(A,B,C)}{P(C)} = \frac{P(A \mid C)P(B \mid C)P(C)}{P(C)}$$
$$= P(A \mid C)P(B \mid C)$$

Head-To-Tail



In general, P(A,B) is not expressed as P(A)P(B). Therefor, $A \perp\!\!\!\perp B \mid \Phi$ does not hold.

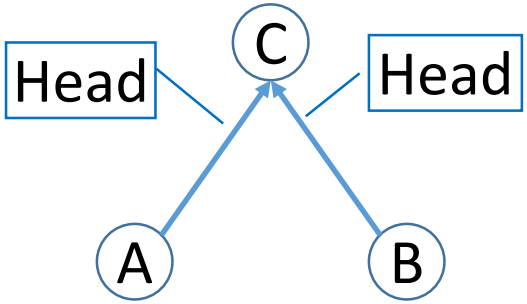
$$P(A,B) = \sum_C P(A,B,C) = \sum_C P(A)P(B \mid C)P(C \mid A)$$



P(A,B|C) is expressed as P(A|C)P(B|C). Therefore $A \perp\!\!\!\perp B \mid C$ holds.

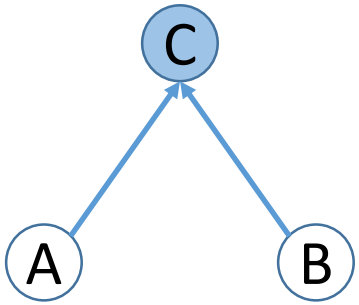
$$P(A,B \mid C) = \frac{P(A,B,C)}{P(C)} = \frac{(P(C \mid A)P(A))P(B \mid C)}{P(C)}$$
$$= P(A \mid C)P(B \mid C)$$

Head-To-Head



In general, P(A,B) is expressed as P(A)P(B).
Therefor, $A \perp\!\!\!\perp B \mid \Phi$ holds.

$$P(A,B) = \sum_C P(A,B,C) = \sum_C P(A)P(B)P(C \mid A,B) = P(A)P(B)$$

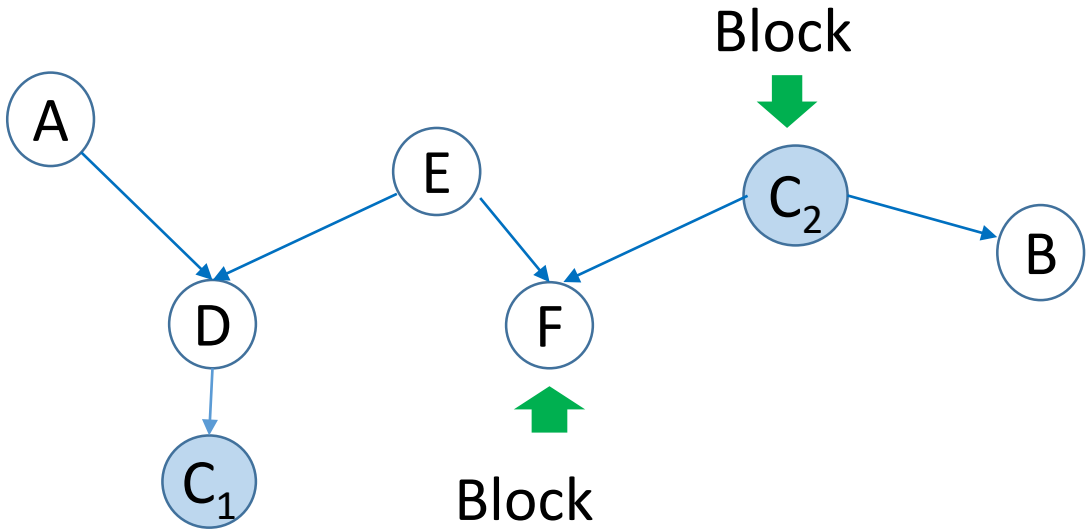


P(A,B|C) is not expressed as P(A|C)P(B|C).
Therefore $A \perp\!\!\!\perp B \mid C$ does not hold.

$$P(A,B \mid C) = \frac{P(A,B,C)}{P(C)} = \frac{P(A)P(B)P(C \mid A)}{P(C)}$$

Blocking a Path

- For a Bayesian network, let A and B be a node, and C be a set of nodes that does not include A and B. We say a path from A to B is blocked when either of the followings holds
  - On the path from A to B, there is a node in C and the connection of the arcs is tail-to-tail or head-to-tail
  - At one of the nodes on the path from A to B, the connection of the arcs is head-to-head. In addition, the node and its all descendants are not included in C



d-separation

For a Bayesian network, let A, B, and C be exclusive sets of nodes

- We say A is d-separated from B by C if all the paths starting from a node in A and ending at a node in B is blocked
- When A is d-separated from B by C, $A \perp\!\!\!\perp B \mid C$ holds for the joint probability defined by the Bayesian network (Pearl 1988)

Maximization of Joint Probability

- Obtained by replacing Σ in the sum-product algorithm with max
  → Max-product Algorithm

$$\langle X_1, X_2, \cdots, X_N \rangle = \underset{X_1, X_2, \cdots, X_N}{\arg\max}\ P(X_1, X_2, \cdots X_N)$$

Sum-Product Algorithm (for Tree)

- Message passing
  - Leaf nodes
    - Variable node: $\mu_{x\to f}(x)=1$
    - Factor node: $\mu_{f\to x}(x)=f(x)$
  - Variable node to factor node:

$$\mu_{x\to f}(x)=\prod_{i}\mu_{f_i\to x}(x)$$

  - Factor node to variable node:

$$\mu_{f\to x}(x)=\sum_{x_1}\cdots\sum_{x_M}f(x,x_1,\cdots,x_M)\prod_{m=1}^{M}\mu_{x_m\to f}(x_m)$$

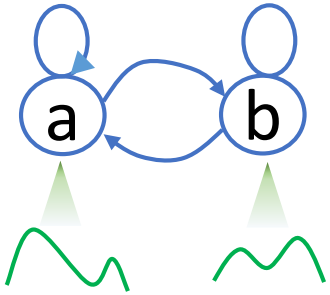
- Marginal probability

$$P(x)=\prod_{i}\mu_{f_i\to x}(x)$$

EM for HMM and Efficiency

$$Q(\Theta,\Theta_0)=\sum_K HMM(K\mid X,\Theta_0)\log HMM(K,X\mid\Theta) \quad (1)$$

$$K=\langle k_1,k_2,\cdots,k_T\rangle \quad X=\langle x_1,x_2,\cdots,x_T\rangle$$



- The summation is over state sequence K
- The number of the sequences is exponential to the length of input X
- 1 sec of feature sequence is 100 frames

Directly enumerating all the paths is impossible

Q-function (1) can be efficiently evaluated if posteriors $P(k_t=s\mid X,\Theta_0)$ and $P(k_{t-1}=s',k_t=s\mid X,\Theta_0)$ are obtained, where s' and s are HMM state ID

Use the Sum-Product algorithm to efficiently obtain the posteriors