

# Bayesian Networks

Course of Signal Processing and Data Fusion

A.A. 2021-2022

# Probabilistic Graphical Models

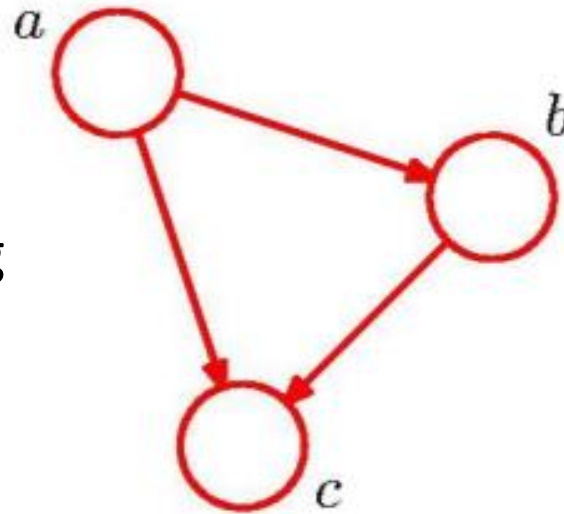
## Key Idea

- Represent the “World” as a set of **Random Variables**  $X_1, \dots, X_N$  and the **Joint Distribution**  $P_{X_1, \dots, X_N}(x_1, \dots, x_N)$
- Represent, graphically, the independence/dependence relations among Random Variables

# Bayesian Networks

## Directed Acyclic Graph (DAG)

- We have chosen a particular ordering
- Different ordering  $\rightarrow$  different decomposition  $\rightarrow$  different graphical representation



- Nodes: Random Variables
- Edges: Dependences
- No directed cycles

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

**Chain Rule**

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

# The benefits of structure

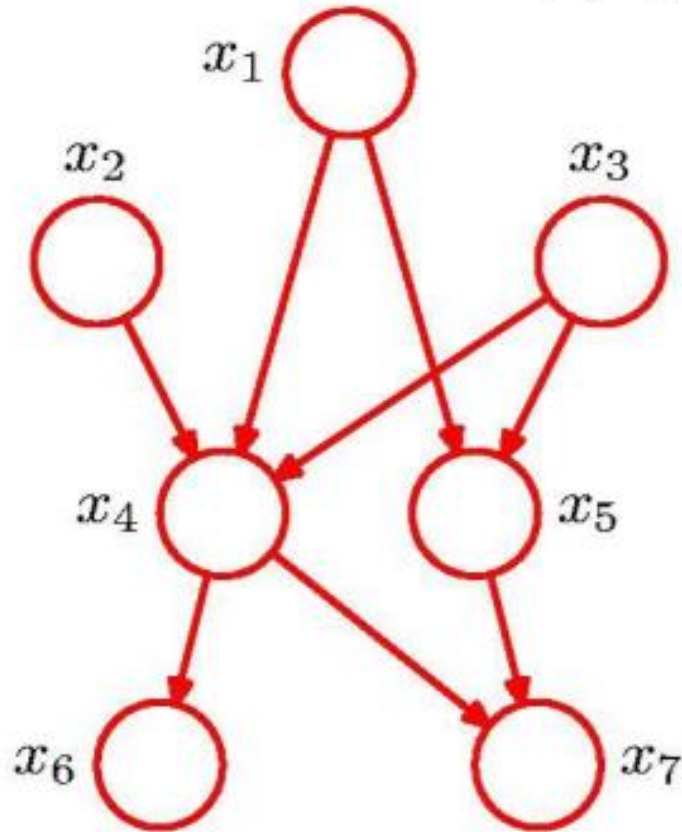
- Specify, *independently*, all entries of  $P_{X_1, \dots, X_N}(x_1, \dots, x_N)$ , with  $X_i$  binary variables, requires  $O(2^N)$  space
- Computing marginal of  $P_{X_i}(x_i)$  requires to sum over  $2^{(N-1)}$  states of other variables:

$$P_{X_i}(x_i) = \sum_{\sim x_i} P_{X_1, \dots, X_N}(x_1, \dots, x_N)$$

- The key idea is to specify which variables are independent of others, leading to a structured factorization of the joint probability distribution

# Bayesian Networks

$$p(x_1, \dots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$



***It is the absence of links that conveys interesting information***

General Factorization

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

# Wet Grass Example - 1

One morning Tracey leaves her house and realises that her grass is wet. Is it due to overnight rain or did she forget to turn off the sprinkler last night? Next she notices that the grass of her neighbour, Jack, is also wet. This *explains away* to some extent the possibility that her sprinkler was left on, and she concludes therefore that it has probably been raining.

A model of Tracey's world then corresponds to a probability distribution on the joint set of the variables of interest  $p(T, J, R, S)$  (the order of the variables is irrelevant).

$R \in \{0, 1\}$      $R = 1$  means that it has been raining, and 0 otherwise

$S \in \{0, 1\}$      $S = 1$  means that Tracey has forgotten to turn off the sprinkler, and 0 otherwise


$J \in \{0, 1\}$      $J = 1$  means that Jack's grass is wet, and 0 otherwise

$T \in \{0, 1\}$      $T = 1$  means that Tracey's Grass is wet, and 0 otherwise

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R|S)p(S) \end{aligned}$$

**Chain Rule**

# Wet Grass Example - 2

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R|S)p(S) \end{aligned}$$


8                      4                      2                      1

$p(T|J, R, S)$  requires us to specify  $2^3 = 8$  values – we need  $p(T = 1|J, R, S)$  for the 8 joint states of  $J, R, S$ .

The other value  $p(T = 0|J, R, S)$  is given by normalisation :  $p(T = 0|J, R, S) = 1 - p(T = 1|J, R, S)$

For a distribution of  $N$  binary variables, we need to specify  $2^N - 1$  values in range  $[0, 1]$

# Wet Grass Example - 3

We can make some conditional independence assumption

$$p(T|J, R, S) = p(T|R, S)$$

$$p(J|R, S) = p(J|R)$$

$$p(R|S) = p(R)$$



$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$

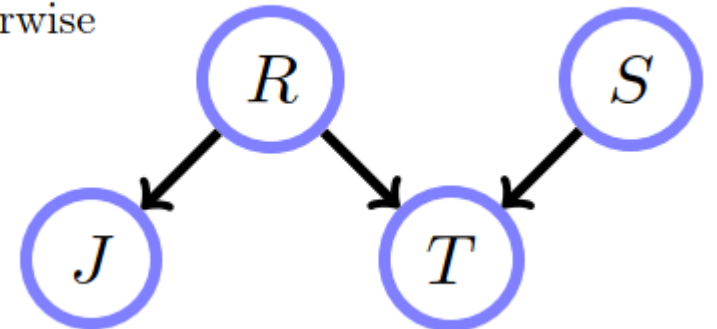
4                      2                      1                      1

$R \in \{0, 1\}$      $R = 1$  means that it has been raining, and 0 otherwise

$S \in \{0, 1\}$      $S = 1$  means that Tracey has forgotten to turn off the sprinkler, and 0 otherwise

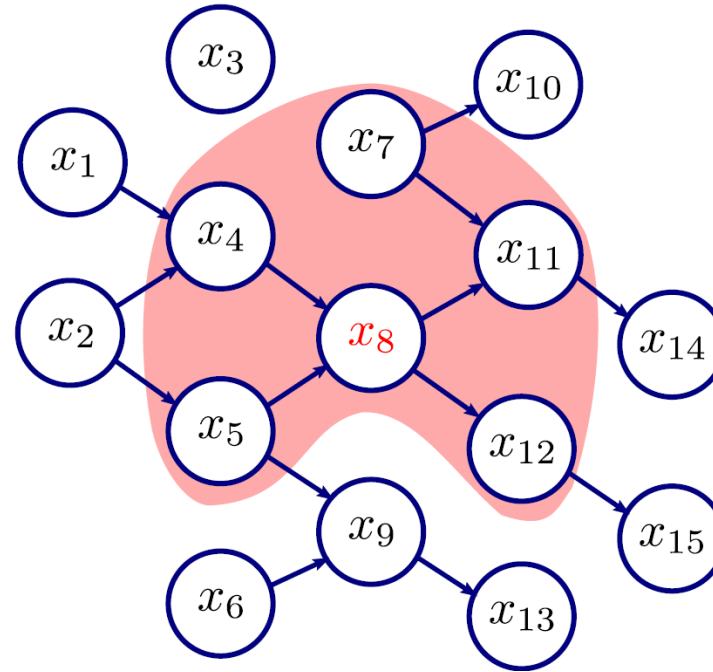
$J \in \{0, 1\}$      $J = 1$  means that Jack's grass is wet, and 0 otherwise

$T \in \{0, 1\}$      $T = 1$  means that Tracey's Grass is wet, and 0 otherwise





# Example Factorization



$$\begin{aligned} Pr(x_1 \dots x_{15}) = & Pr(x_1)Pr(x_2)Pr(x_3)Pr(x_4|x_1, x_2)Pr(x_5|x_2)Pr(x_6) \\ & Pr(x_7)Pr(x_8|x_4, x_5)Pr(x_9|x_5, x_6)Pr(x_{10}|x_7)Pr(x_{11}|x_7, x_8) \\ & Pr(x_{12}|x_8)Pr(x_{13}|x_9)Pr(x_{14}|x_{11})Pr(x_{15}|x_{12}). \end{aligned}$$

# Example

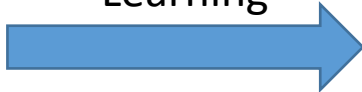
## Asbestos-Cancer-Smoke

$$p(A, S, C) = p(C|A, S) p(A) p(S)$$

Observed Data

A	S	C
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

Learning



$$p(C|A, S) = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$$

$$p(A) = \begin{bmatrix} \frac{3}{7} & \frac{4}{7} \end{bmatrix}$$

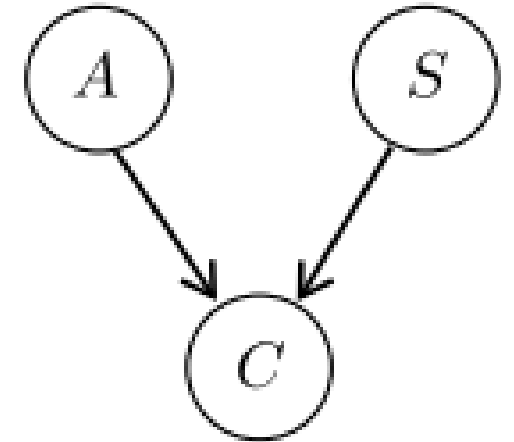
$$p(S) = \begin{bmatrix} \frac{3}{7} & \frac{4}{7} \end{bmatrix}$$

A,S	C	
	False	True
False, False	1	0
False, True	$\frac{1}{2}$	$\frac{1}{2}$
True, False	$\frac{1}{2}$	$\frac{1}{2}$
True, True	0	1

Inference



$$p(A = 1|C = 1, S = 0) = 1$$



# Single Random Variable - 1

- Consider a discrete single Random Variable  $X$
- Take values in a finite sample space  $\mathcal{X} = \{x^1, x^2, \dots, x^N\}$
- A generic value of the Random Variable:  $x$
- Random Variable distributed as  $\pi_x(x) : X \sim \pi_X(x)$

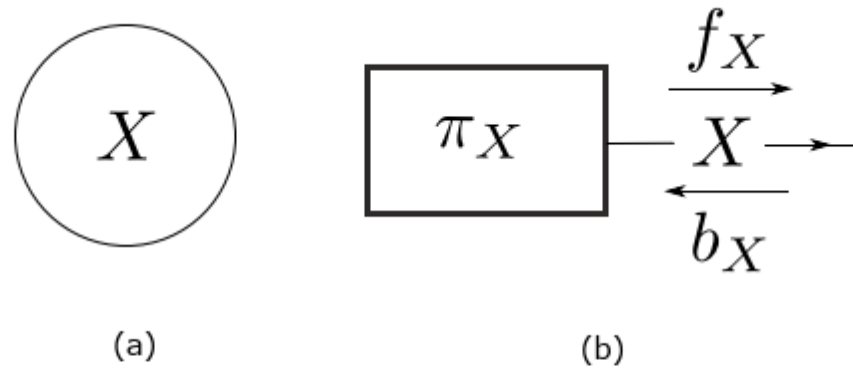


Figure 3.1: Graphical Representation of the Variable. (a) Bayesian representation (b) Factor Graph representation

# Single Random Variable - 2

- Functional Notation

$$f_X(x), b_X(x), p_X(x)$$

- Vector Notation

$$\mathbf{f}_X = \begin{bmatrix} f_X(x^1) \\ f_X(x^2) \\ \vdots \\ f_X(x^N) \end{bmatrix} \quad \mathbf{b}_X = \begin{bmatrix} b_X(x^1) \\ b_X(x^2) \\ \vdots \\ b_X(x^N) \end{bmatrix} \quad \mathbf{p}_X = \begin{bmatrix} p_X(x^1) \\ p_X(x^2) \\ \vdots \\ p_X(x^N) \end{bmatrix}$$

# Two Random Variables - 1

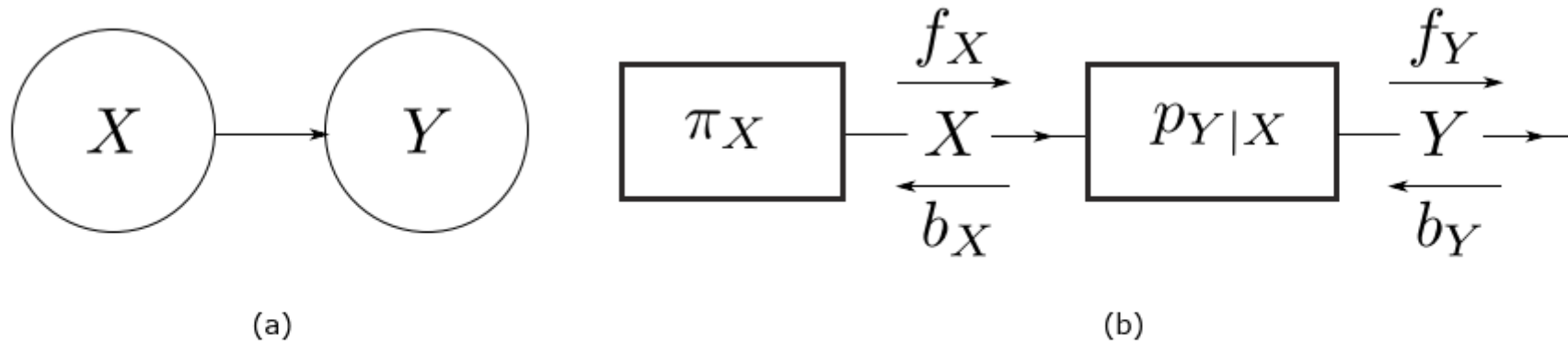


Figure 3.2: Graphical Representation of the relation between two random variables. (a) Bayesian Graph representation (b) Factor Graph representation

- Representation of the joint distribution  $p_{XY}(x, y)$  as:

$$p_{XY}(x, y) = p_{Y|X}(y|x)\pi_X(x)$$

# Two Random Variables - 2

- Functional Notation

$$p_{Y|X}(y|x)$$

- Vector-Matrix Notation

$$\mathbf{p}_{Y|X} = \begin{bmatrix} Pr(Y = y^1 | X = x^1) & Pr(Y = y^2 | X = x^1) & \dots & Pr(Y = y^{N_Y} | X = x^1) \\ Pr(Y = y^1 | X = x^2) & Pr(Y = y^2 | X = x^2) & \dots & Pr(Y = y^{N_Y} | X = x^2) \\ \vdots & \vdots & \vdots & \vdots \\ Pr(Y = y^1 | X = x^{N_X}) & Pr(Y = y^2 | X = x^{N_X}) & \dots & Pr(Y = y^{N_Y} | X = x^{N_X}) \end{bmatrix}$$

$$\sum_{y \in \mathcal{Y}} Pr\{Y = y | X = x^i\} = 1, \quad \forall i \in \{1, \dots, N_X\} \quad \textbf{Row-Stochastic}$$

## Two Random Variables - 3

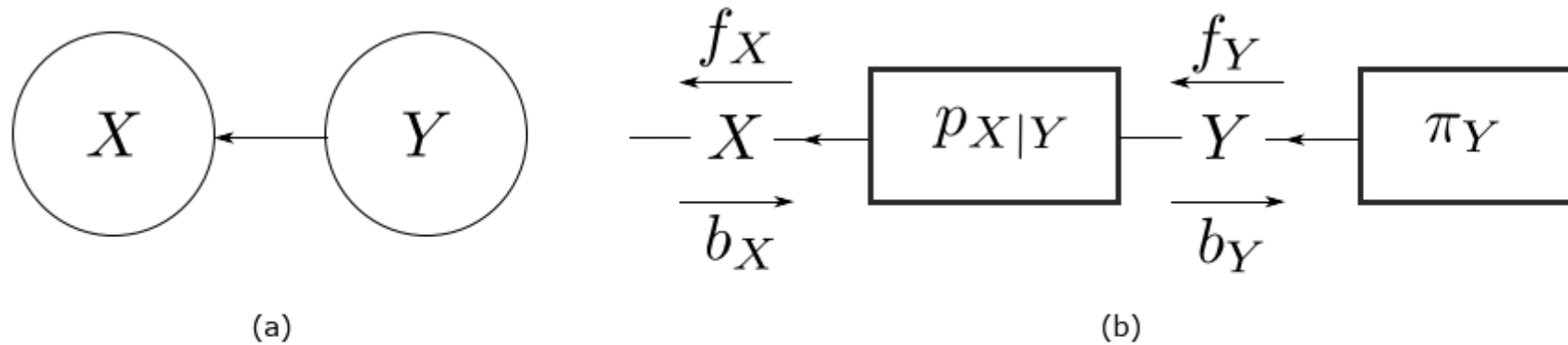


Figure 3.4: Graphical Representation of the relation between two random variables. (a) Bayesian representation (b) Factor Graph representation

- Representation of the joint distribution  $p_{XY}(x, y)$  as:

$$p_{XY}(x, y) = p_{X|Y}(x|y)\pi_Y(y)$$

# Two Random Variables - 4

- The choice of the graph and its directionality depends on the problem under study and computational convenience
- Evidence on a variable (i.e.  $X = \bar{x}$ ) is depicted as a “shaded” node

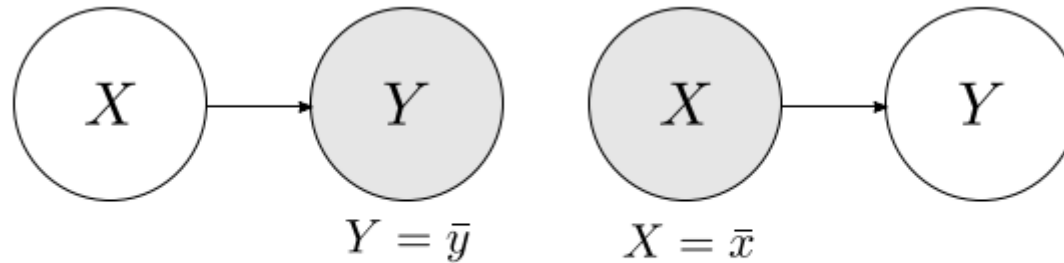


Figure 3.5: Evidence on variables

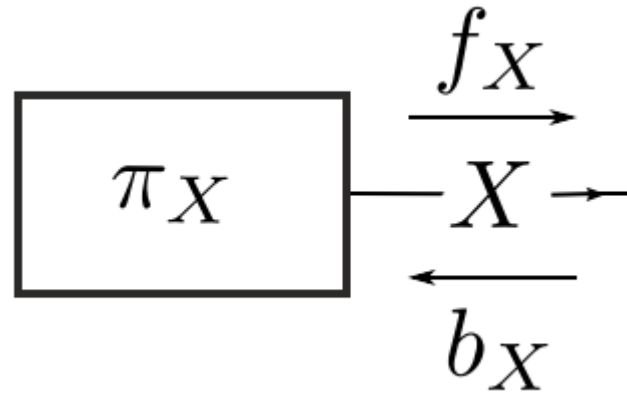


# Variable

$$\begin{array}{c} \xrightarrow{f_X} \\ \text{---} X \text{---} \\ \xleftarrow{b_X} \end{array}$$

$$p_X(x|evidence) = f_X(x)b_X(x)$$

# Source Block



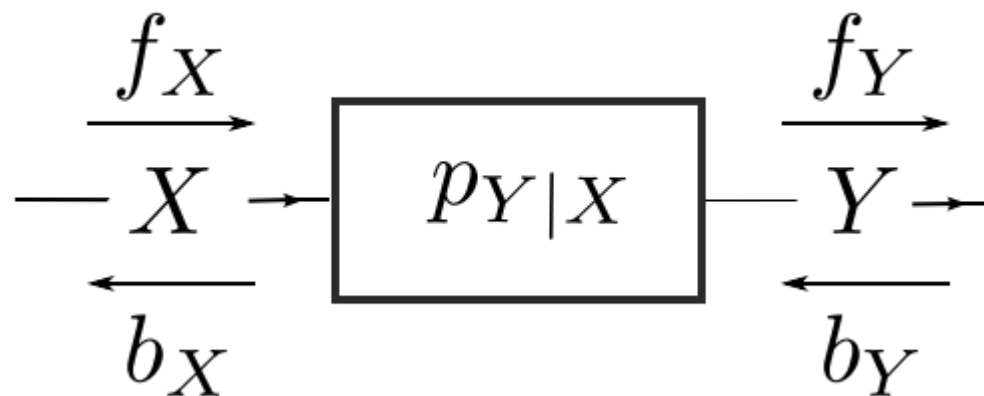
Functional notation:

$$\pi_X(x), \quad f_X(x) = \pi_X(x)$$

Vector notation:

$$\boldsymbol{\pi}_X, \quad \boldsymbol{f}_X = \boldsymbol{\pi}_X$$

# SISO Block



**Forward Flow**

In functional notation:

$$f_Y(y) = \sum_{x \in \mathcal{X}} p_{Y|X}(y|x) f_X(x)$$

In vector-matrix notation:

$$\mathbf{f}_Y = \mathbf{P}_{Y|X}^T \mathbf{f}_X$$

**Backward Flow**

In functional notation:

$$b_X(x) = \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) b_Y(y)$$

In vector-matrix notation:

$$\mathbf{b}_X = \mathbf{P}_{Y|X} \mathbf{b}_Y$$

Examples with 2 Variables

# Diverter

- Diverter or Equality constraint

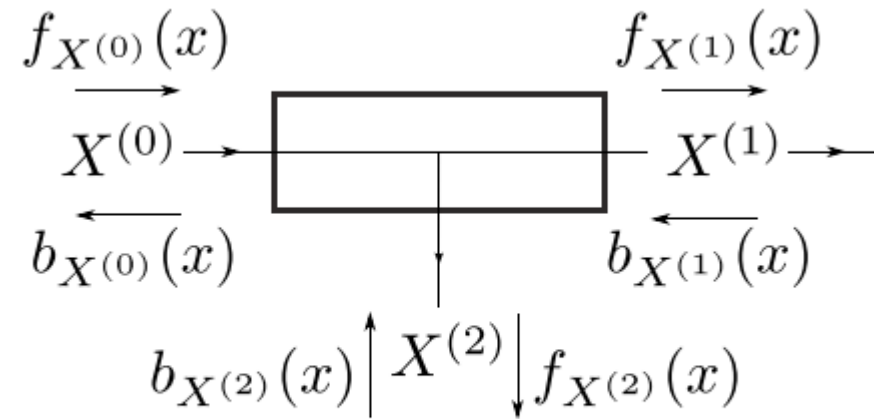


Figure 3.10: Diverter

$$\begin{aligned}
 f_{X^{(1)}}(x) &\propto f_{X^{(0)}}(x)b_{X^{(2)}}(x) \\
 b_{X^{(0)}}(x) &\propto b_{X^{(2)}}(x)b_{X^{(1)}}(x) \\
 f_{X^{(2)}}(x) &\propto b_{X^{(1)}}(x)f_{X^{(0)}}(x)
 \end{aligned}$$

Examples with 2 Variables + Diverter

# Smooth Evidence

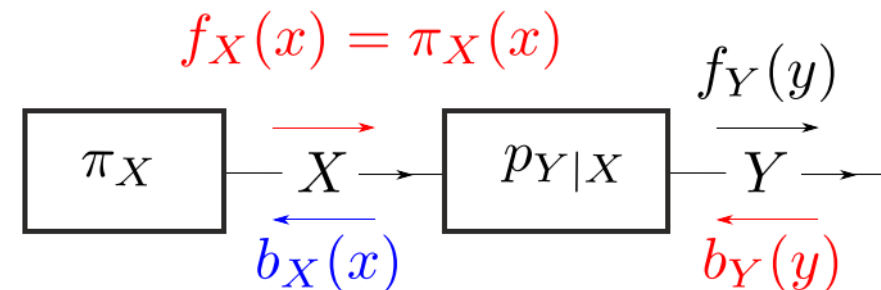
- In many applications we mainly have available only approximate inference about a variable
- We can compute the average posterior

$$\sum_{y \in \mathcal{Y}} p_X(x|Y=y) b_Y(y) = \mathbb{E}_{Y \sim b_Y(y)} [p_X(x|Y=y)]$$



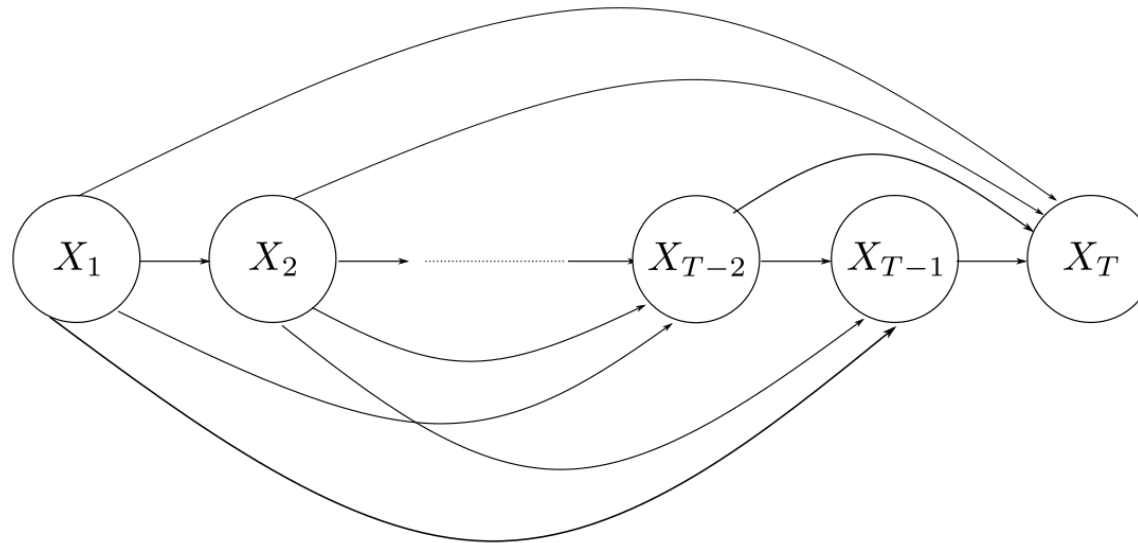
$$b_X(x) \propto \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \underbrace{b_Y(y)}_{\text{evidence}}$$

$$p_X(x) \propto f_X(x) b_X(x) = \pi_X(x) \underbrace{b_X(x)}_{\substack{\text{summary} \\ \text{of evidence}}}$$



# More Random Variables

- Consider  $T$  random variables  $X_1, \dots, X_T$



- The joint distribution can always be factorized using the *chain rule* (we have used just one of conditioning orders):

$$p_{X_1 X_2 \dots X_T}(x_1, x_2, \dots, x_T) = p_{X_T | X_{T-1} \dots X_1}(x_T | x_{T-1}, \dots, x_1) p_{X_{T-1} | X_{T-2} \dots X_1}(x_{T-1} | x_{T-2}, \dots, x_1) \\ \dots p_{X_3 | X_2 X_1}(x_3 | x_2, x_1) p_{X_2 | X_1}(x_2 | x_1) p_{X_1}(x_1)$$



# Markov Chain - 1

- Markov chain is a system in which we are given a time order for the variables

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \dots \rightarrow X_T$$

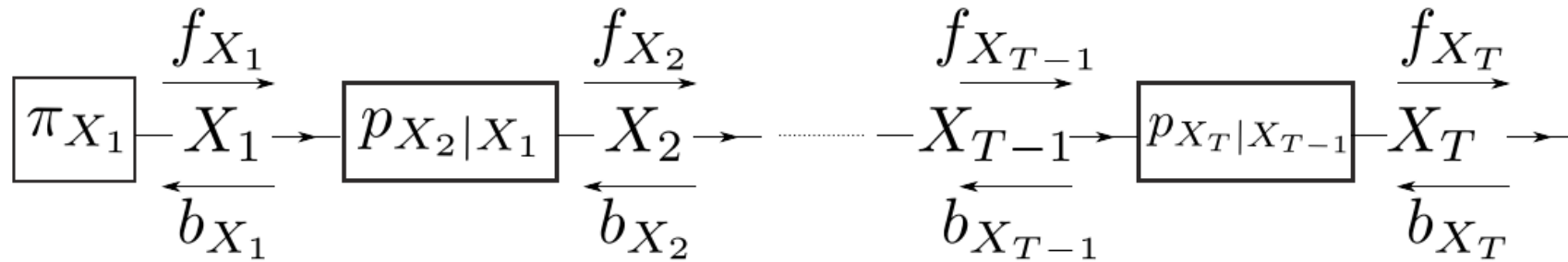
- Conditioning can be limited to the previous variable

$$p_{X_1 X_2 \dots X_T}(x_1, x_2, \dots, x_T) = p_{X_T | X_{T-1}}(x_T | x_{T-1}) p_{X_{T-1} | X_{T-2}}(x_{T-1} | x_{T-2}) \\ \dots p_{X_3 | X_2}(x_3 | x_2) p_{X_2 | X_1}(x_2 | x_1) p_{X_1}(x_1)$$

# Markov Chain - 2



(a)



(b)

Figure 3.12: Markov Chain. (a) Bayesian representation (b) Factor Graph representation

$$p_{X_1 X_2 \dots X_T}(x_1, x_2, \dots, x_T) = p_{X_T|X_{T-1}}(x_T|x_{T-1})p_{X_{T-1}|X_{T-2}}(x_{T-1}|x_{T-2}) \\ \dots p_{X_3|X_2}(x_3|x_2)p_{X_2|X_1}(x_2|x_1)p_{X_1}(x_1)$$

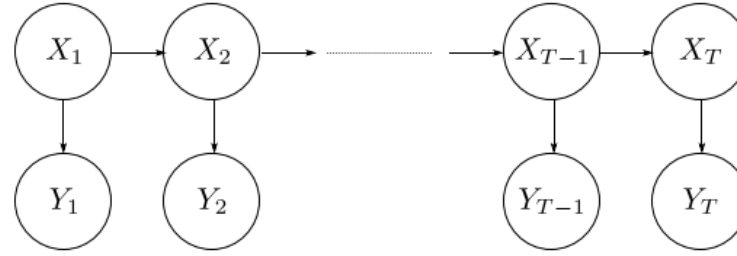
# Markov Chain - 3

- Total characterization requires knowledge of the distributions:

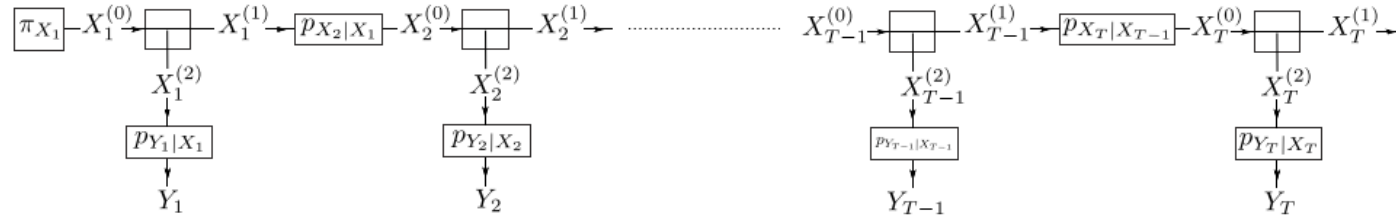
$$\{\pi_{X_1}, p_{X_2|X_1}, p_{X_3|X_2}, \dots, p_{X_T|X_{T-1}}\}$$

- *Time invariant* if:  $p_{X_2|X_1}(\xi, \mu) = p_{X_3|X_2}(\xi, \mu) = \dots = p_{X_T|X_{T-1}}(\xi, \mu)$
- If time invariant,  $p_{X_t|X_{t-1}}(x_t|x_{t-1})$  and  $\pi_{X_1}(x_1)$  are sufficient to characterize the model

# Hidden Markov Model - 1



(a)



(b)

Figure 3.22: Hidden Markov Model: (a) Bayesian Graph; (b) Factor Graph

$$\begin{aligned}
 p_{X_1 \dots X_T Y_1 \dots Y_T}(x_1, \dots, x_T, y_1, \dots, y_T) = & p_{X_T|X_{T-1}}(x_T|x_{T-1})p_{X_{T-1}|X_{T-2}}(x_{T-1}|x_{T-2}) \dots \\
 & p_{X_3|X_2}(x_3|x_2)p_{X_2|X_1}(x_2|x_1)p_{X_1}(x_1) \\
 & p_{Y_1|X_1}(y_1|x_1)p_{Y_2|X_2}(y_2|x_2) \dots p_{Y_T|X_T}(y_T|x_T)
 \end{aligned} \tag{3.38}$$

or more synthetically:

$$p_{X_1 \dots X_T Y_1 \dots Y_T}(x_1, \dots, x_T, y_1, \dots, y_T) = p_{X_1}(x_1) \prod_{t=2}^T p_{X_t|X_{t-1}}(x_t|x_{t-1}) \prod_{t=1}^T p_{Y_t|X_t}(y_t|x_t) \tag{3.39}$$

# Hidden Markov Model - 2

- This model is often used to represent non-observable states, assuming that variables  $Y_1, \dots, Y_N$  are the observables.
- HMM are very popular in a very large number of applications that go from the text classification to tracking

Examples on Markov Chain with 4 variables

# Latent Variable Model - 1

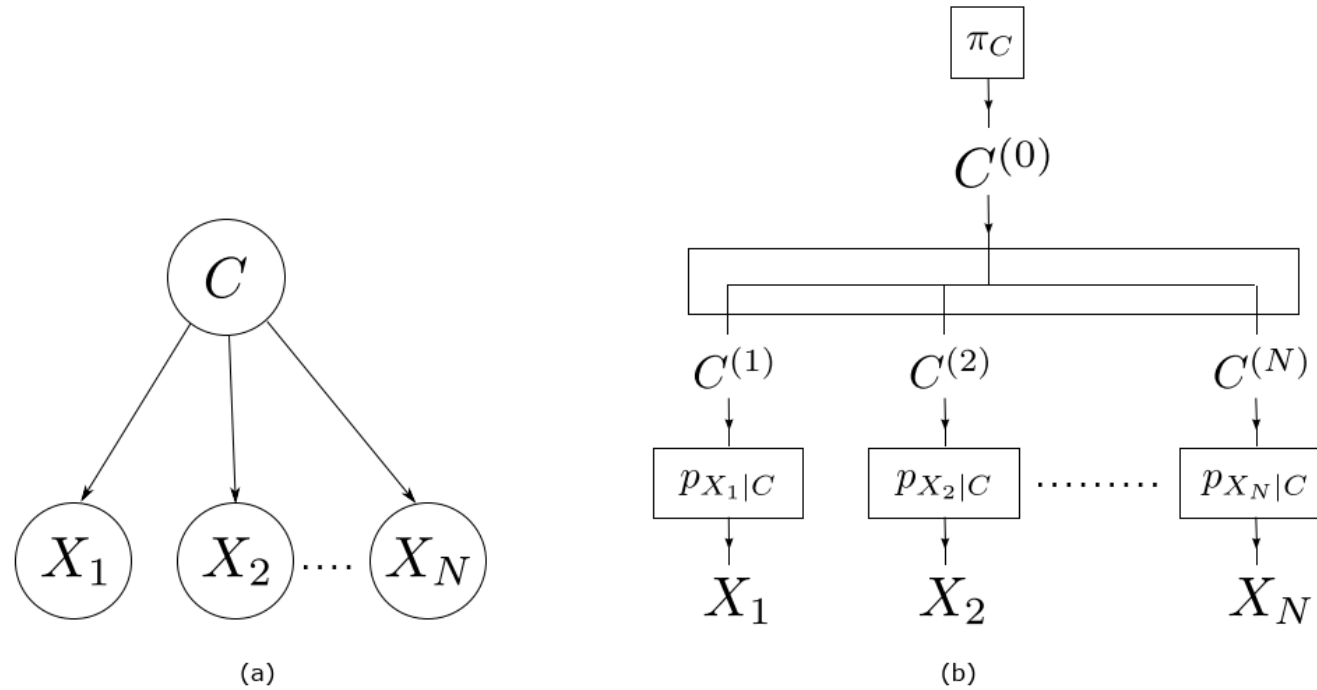


Figure 3.15: Graphical representation of Latent Variable Model as (a) Bayesian Network (b) Factor Graph

$$p_{X_1 X_2 \dots X_N C}(x_1, x_2, \dots, x_N, c) = p_{X_1|C}(x_1|c) p_{X_2|C}(x_2|c) \dots p_{X_N|C}(x_N|c) p_C(c)$$

$X_1, X_2, \dots, X_N$  are *conditionally independent given  $C$* .

Examples on LVM + Sensor Fusion on Matlab



# Variables with more than one parent - 1

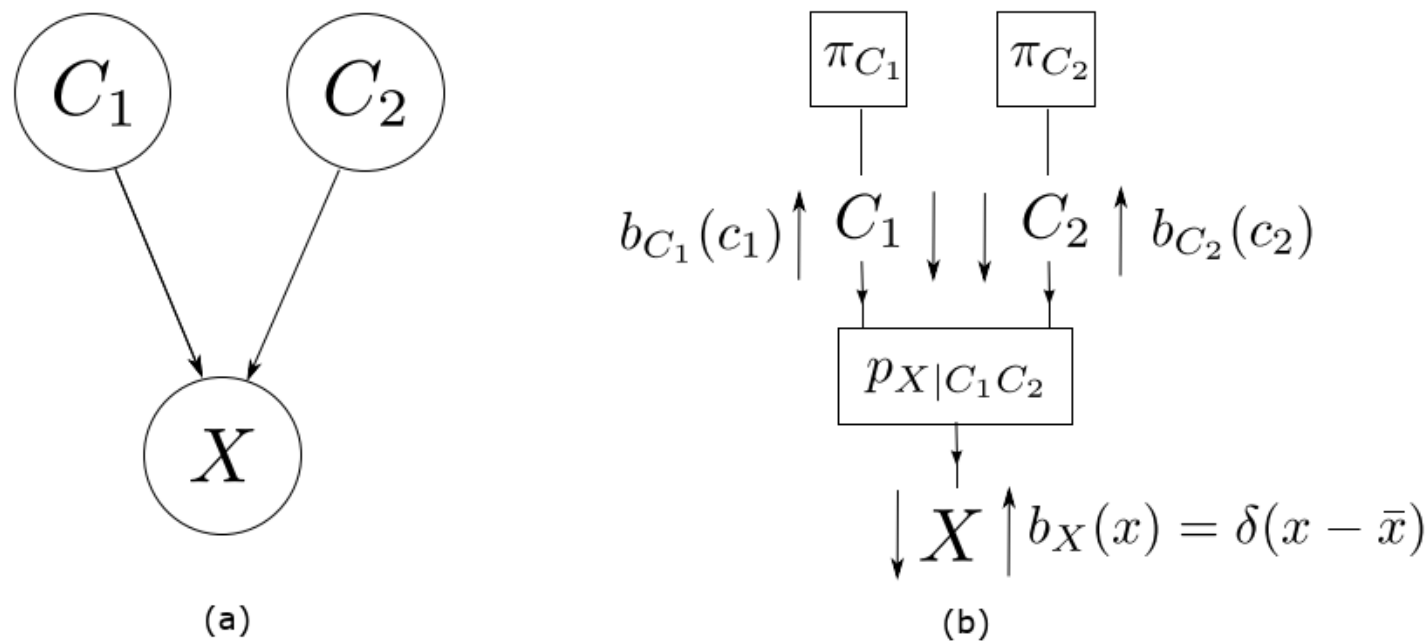


Figure 3.19: Married Variables: (a) Bayesian Graph, (b) Factor Graph

$$p_{XC_1C_2}(x, c_1, c_2) = p_{X|C_1C_2}(x|c_1, c_2)p_{C_1}(c_1)p_{C_2}(c_2)$$

# Variables with more than one parent - 2

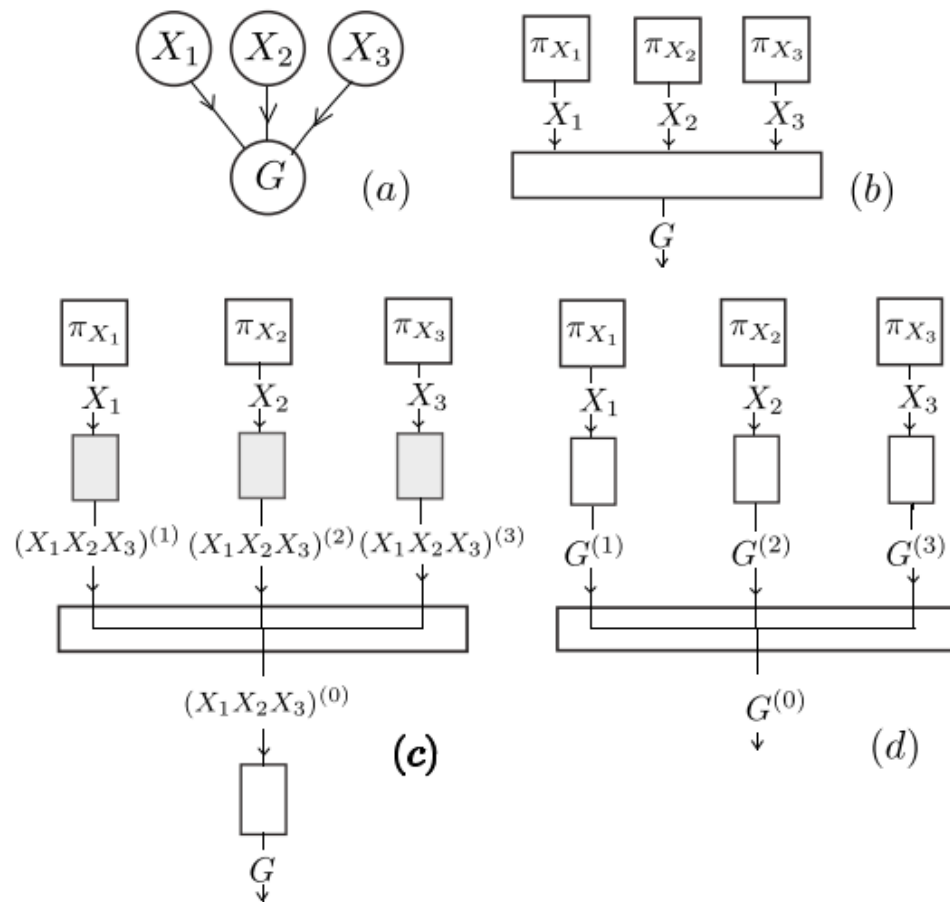


Fig. 3. (a) Bayesian-directed graph for three parents with a single child. (b) Equivalent FGn. (c) Equivalent FGn with the inclusion of the product-space variable. (d) Simplified FGn that is not necessarily equivalent to the graph shown in (a).

# Derivation of matrix of Shaded Blocks

Examples of Burglary - Earthquake

# Example: Los Angeles Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary, Earthquake, Alarm, JohnCalls, MaryCalls*
- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call



# Example: Burglar Alarm

