Probabilistic Models and Machine Learning

No. 7

Graphical Models

Hui Jiang

Department of Electrical Engineering and Computer Science
York University, Toronto, Canada

Outline

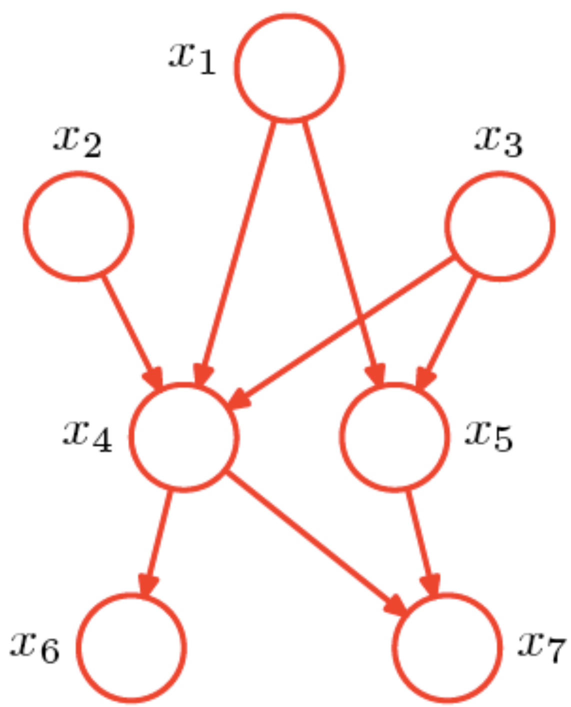
- **Graphical Model:** concepts
  - **Bayesian networks:** directed graphical models
  - **Markov random fields: indirected graphical models**
- **Bayesian Networks: Conditional independence**
- **Bayesian Networks: Learning**
- **Bayesian Networks: Inference**
  - **Exact Inference:** Sum-product algorithm; Max-sum algorithm
  - **Approximate Inference:** Loopy Belief Propagation; Variational Inference; Expectation Propagation; Monte Carlo Sampling
- **Markov random fields**

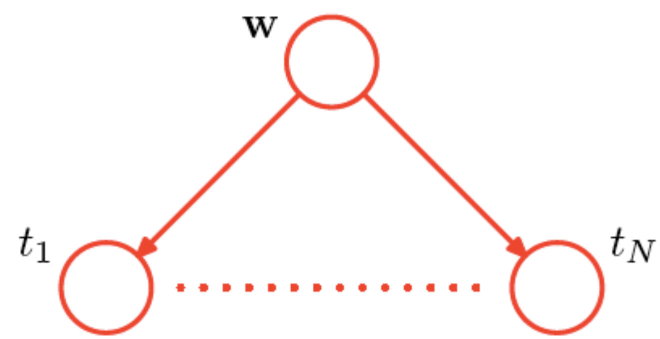
Graphical Model

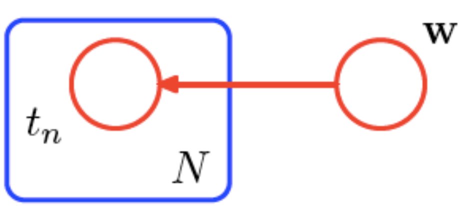
- Use a graph to represent joint distributions of random variables: $p(x, y, z)$
  - Nodes —> random variables (RV)
  - Linking —> dependency among RVs
- Graphical Models may imply conditional independence among RVs.
- Two types of graphical models
  - Directed links: Bayesian networks
  - Undirected links: Markov random field

Bayesian Networks (1)

- **Use a directed graph to represent joint distributions of random variables**
  - **Nodes —> random variables (RV)**
  - **Linking —> conditional distribution of children given the parents**

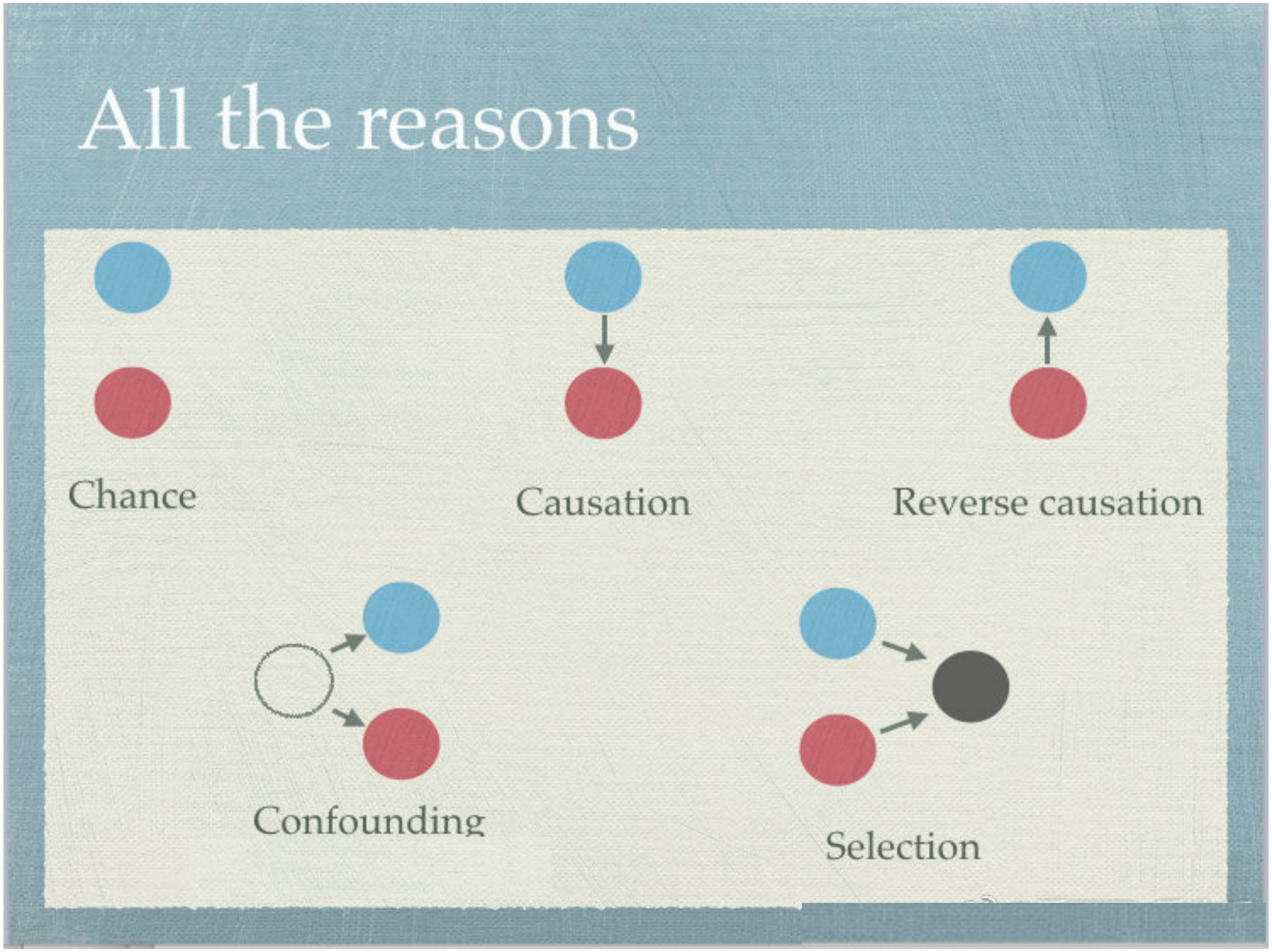






Bayesian Networks (1)

- **All relations in a Bayesian network:**



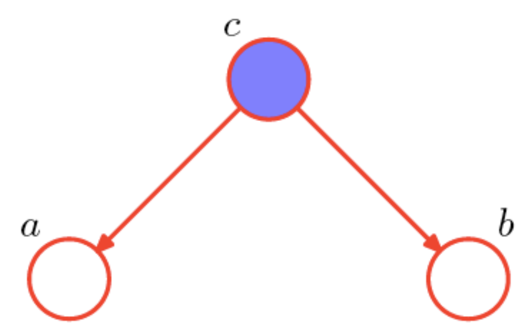
Cofounding variable

- **Cofounding variable: with a tail-to-tail connection**
  - **Unconditionally dependent:** $a \not\perp b \mid \emptyset \iff p(a,b) \neq p(a)p(b)$
  - **Conditionally independent:** $a \perp b \mid c \iff p(a,b|c) = p(a|c)p(b|c)$

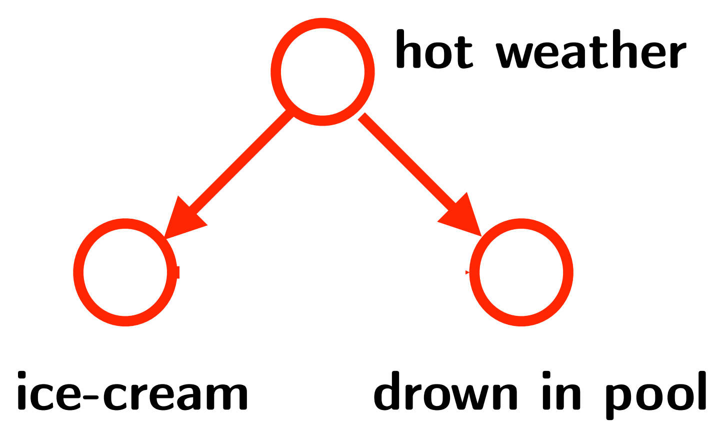
$$p(a,b,c) = p(c)p(a|c)p(b|c)$$

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

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



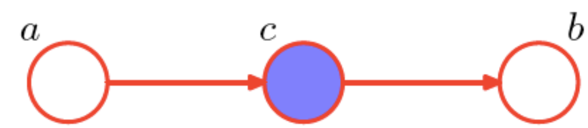
Example: eating ice cream $\Longrightarrow$ drowning in swimming pool



Chained variable

- **Chained variable: with a tail-to-head connection**
  - **Unconditionally dependent:** $a \not\perp b \mid \emptyset \iff p(a,b) \neq p(a)p(b)$
  - **Conditionally independent:** $a \perp b \mid c \iff p(a,b|c) = p(a|c)p(b|c)$

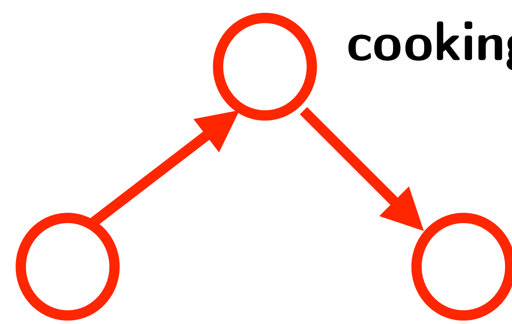
$$p(a,b,c) = p(a)p(c|a)p(b|c)$$



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

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

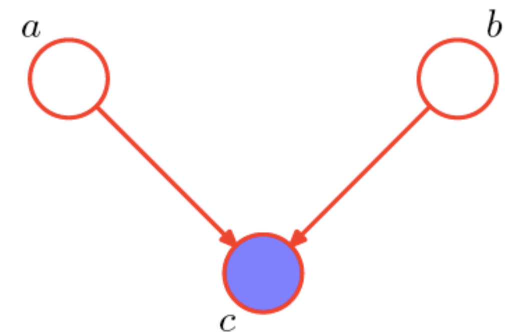
Example: hungry $\Longrightarrow$ cooking
$\Longrightarrow$ burning fingers



Selection variable

- **Selection variable: with a tail-to-tail connection**
  - **Unconditionally independent:** $a \perp b \mid \emptyset \iff p(a,b) = p(a)p(b)$
  - **Conditionally dependent:** $a \not\perp b \mid c \iff p(a,b|c) \neq p(a|c)p(b|c)$

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

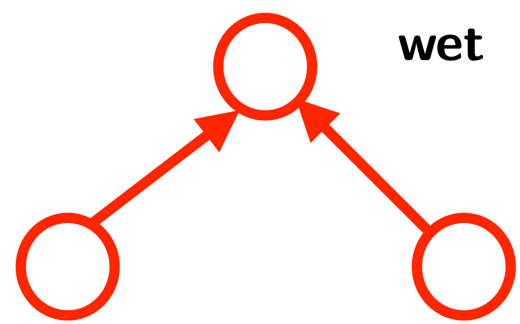


$$p(a,b) = \sum_c p(a)p(b)p(c|a,b) = p(a)p(b)\sum_c p(c|a,b) = p(a)p(b)$$

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

Example: raining $\Longrightarrow$ wet driveway

leaking water-pipe $\Longrightarrow$ wet driveway

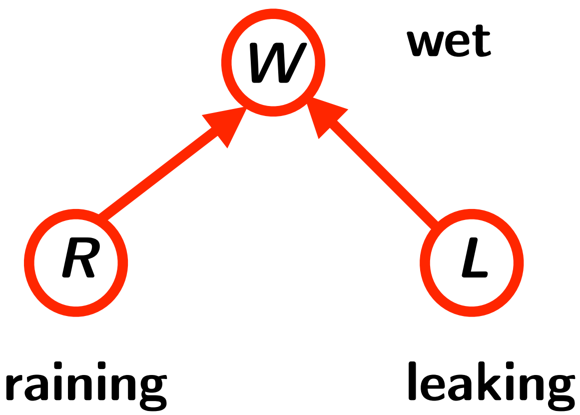


Seeing wet driveway, consider whether pipe is leaking:

Pr(leaking|wet) $>$ Pr(leaking | wet, raining)

Raining explain-away why driveway is wet

Example: how to explain-away



Pr(R=1) = 0.1 **Pr(L=1) = 0.01**

Pr(W=1 | R=1, L=1) = 0.90

Pr(W=1 | R=1, L=0) = 0.80

Pr(W=1 | R=0, L=1) = 0.50

Pr(W=1 | R=0, L=0) = 0.20

Observed: W=1

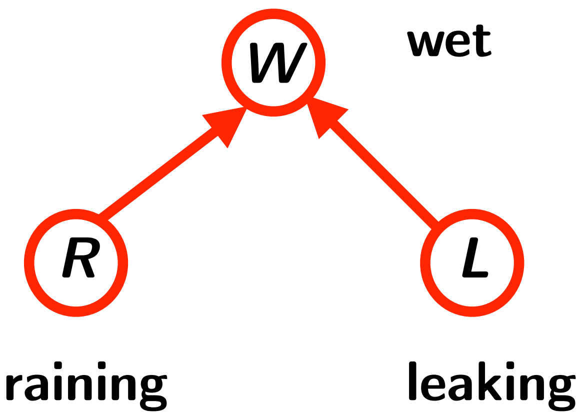
$$\Pr(L=1|W=1) = \frac{\Pr(W=1,L=1)}{\Pr(W=1)} = \frac{\Pr(W=1,L=1,R=1)+\Pr(W=1,L=1,R=0)}{\Pr(W=1,L=1,R=1)+\Pr(W=1,L=1,R=0)+\Pr(W=1,L=0,R=1)+\Pr(W=1,L=0,R=0)}$$

$$= \frac{0.1\times0.01\times0.9+0.9\times0.01\times0.5}{0.1\times0.01\times0.9+0.9\times0.01\times0.5+0.1\times0.99\times0.8+0.9\times0.99\times0.2} = 0.021$$

$$\Pr(L=1|W=1,R=1) = \frac{\Pr(W=1,L=1,R=1)}{\Pr(W=1,R=1)} = \frac{\Pr(W=1,L=1,R=1)}{\Pr(W=1,L=1,R=1)+\Pr(W=1,L=0,R=1)}$$

$$= \frac{0.1\times0.01\times0.9}{0.1\times0.01\times0.9+0.1\times0.99\times0.8} = 0.01125$$

Example: how to explain-away



Pr(R=1) = 0.1   Pr(L=1) = 0.01

Pr(W=1 | R=1, L=1) = 0.90
Pr(W=1 | R=1, L=0) = 0.80
Pr(W=1 | R=0, L=1) = 0.50
Pr(W=1 | R=0, L=0) = 0.20

Observed: W=1

$$\Pr(R=1|W=1) = \frac{\Pr(W{=}1,R{=}1)}{\Pr(W{=}1)} = \frac{\Pr(W{=}1,L{=}1,R{=}1)+\Pr(W{=}1,L{=}0,R{=}1)}{\Pr(W{=}1,L{=}1,R{=}1)+\Pr(W{=}1,L{=}0,R{=}1)+\Pr(W{=}1,L{=}1,R{=}0)+\Pr(W{=}1,L{=}0,R{=}0)}$$

$$= \frac{0.1\times0.01\times0.9+0.99\times0.1\times0.8}{0.1\times0.01\times0.9+0.99\times0.1\times0.8+0.01\times0.9\times0.5+0.99\times0.9\times0.2} = 0.3038$$

$$\Pr(R=1|W=1,L=1) = \frac{\Pr(W{=}1,L{=}1,R{=}1)}{\Pr(W{=}1,L{=}1)} = \frac{\Pr(W{=}1,L{=}1,R{=}1)}{\Pr(W{=}1,L{=}1,R{=}1)+\Pr(W{=}1,L{=}1,R{=}0)}$$

$$= \frac{0.1\times0.01\times0.9}{0.1\times0.01\times0.9+0.9\times0.1\times0.5} = 0.1667$$

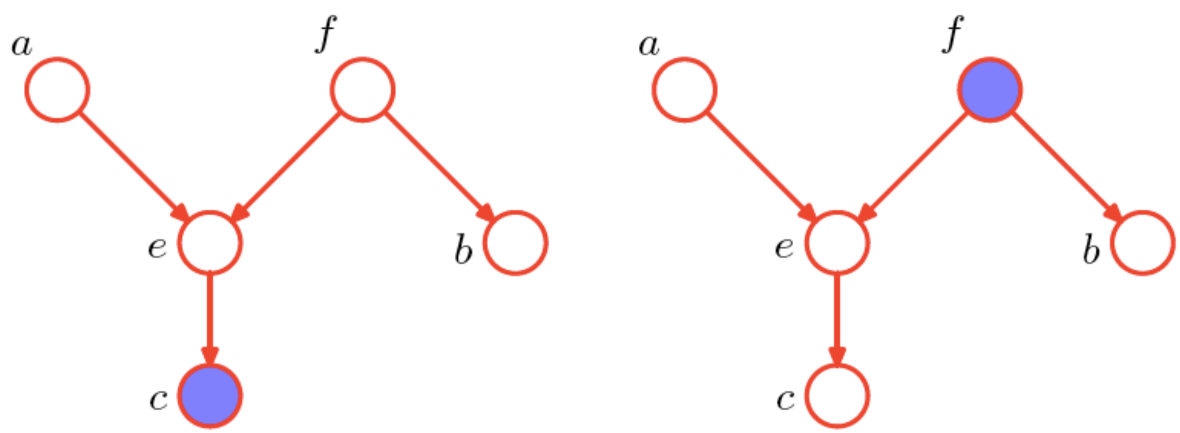
Bayesian Networks (3)

- **Conditional independence in Bayesian Networks in general**

- **D-separation rule:**

  Given any three disjoint subsets of variables: A, B, C

  $$A \perp B \mid C \quad \text{or} \quad A \not\perp B \mid C$$



Learning of Bayesian Networksn

- Model Estimation
  - Structure Learning (or manual specification)
  - Parameter Estimation
- Inference
- Popular Bayesian networks:
  - Mixture models
  - Hidden Markov models (HMM)
  - Latent Dirichlet Allocation (LDA)
  - Bayesian Learning

Learning of Bayesian Networks

- **Parameter Estimation: maximum likelihood estimation**

$$p_\theta(x, y, z, \cdots)$$

- **Full Data: fairly easy but rare**

$$\Big\{(x_1, y_1, z_1, \cdots), (x_2, y_2, z_2, \cdots), \cdots (x_i, y_i, z_i, \cdots), \cdots\Big\}$$

$$l(\theta) = \sum_i \ln p_\theta(x_i, y_i, z_i, \cdots)$$

- **Missing data: EM**

$$\Big\{(x_1, *, z_1, \cdots), (x_2, *, z_2, \cdots), \cdots (x_i, *, z_i, \cdots), \cdots\Big\}$$

$$l(\theta) = \sum_i \ln \sum_{y_i} p_\theta(x_i, y_i, z_i, \cdots)$$

Inference using Bayesian Networks

- **What is Inference?**

$$p_\theta(\underbrace{x_1, x_2, x_3}_{\text{observed } \mathbf{X}}, \underbrace{x_4, x_5, x_6}_{\text{interested } \mathbf{Y}}, \underbrace{x_7, x_8, \ldots}_{\text{missing } \mathbf{Z}})$$

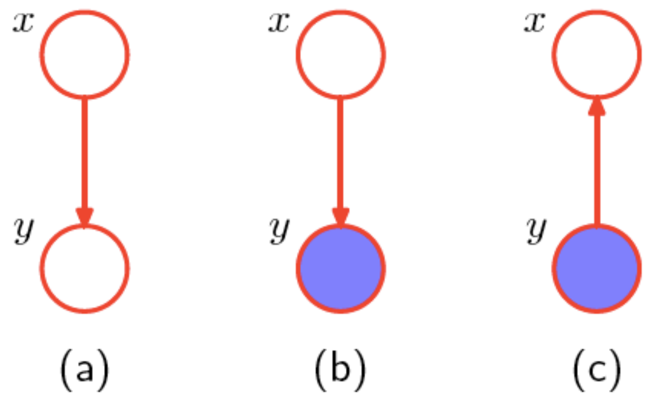
$$p_\theta(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \implies p_\theta(\mathbf{Y}|\mathbf{X})$$

$$p_\theta(\mathbf{Y}|\mathbf{X}) = \frac{p_\theta(\mathbf{X}, \mathbf{Y})}{p_\theta(\mathbf{X})} = \frac{\sum_{\mathbf{Z}} p_\theta(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{\sum_{\mathbf{Y}, \mathbf{Z}} p_\theta(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}$$

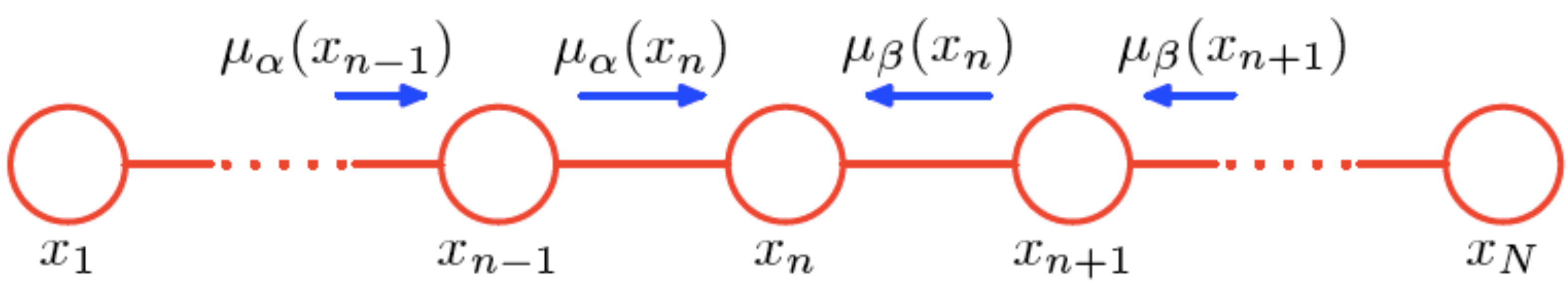
Exact Inference for Bayesian Networks

- Message Propagation for Inference

  - Two nodes —> Bayes' theorem



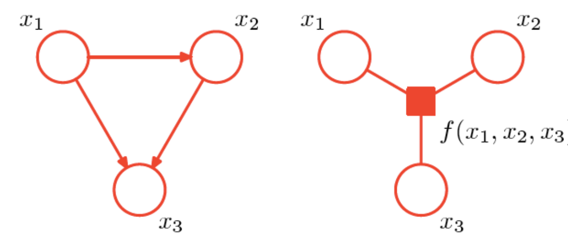
  - Inference on a chain



Exact Inference for Bayesian Networks

- Tree-structured Graphical models
  - Sum-Product (Max-sum) algorithm
- General Graphs
  - Junction tree algorithms
  - Computationally expensive

Sum-Product algorithm (1)

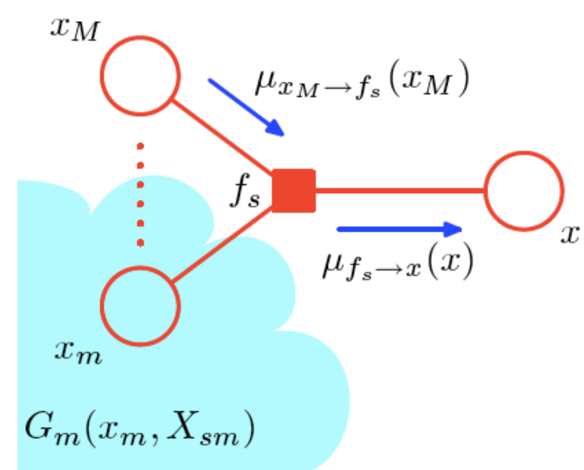


- Form factor graphs
- Message propagation
  - factor node —> variable node

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

  - variable node —> factor node

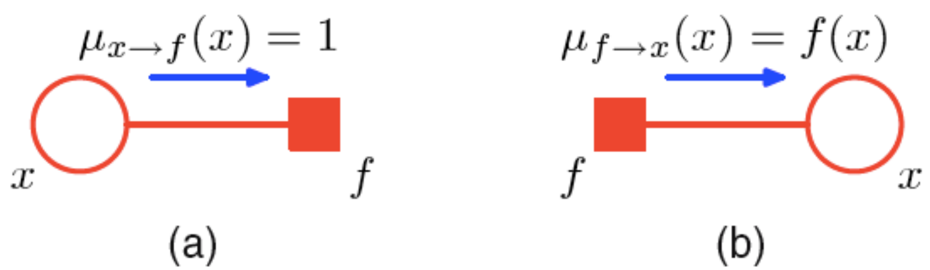
$$\mu_{x_m \to f_s}(x_m) = \prod_{l \in \mathrm{ne}(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$$

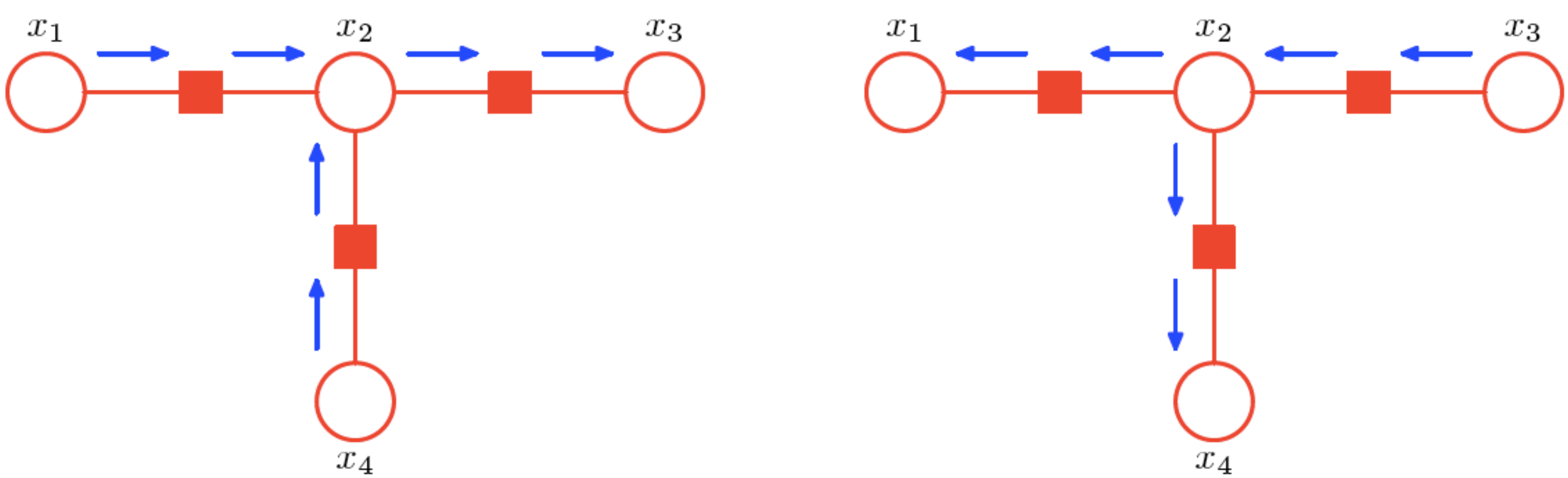


- Two-pass propagation
  - from leafs to root
  - from root to leafs

Sum-Product algorithm (2)

- Two-pass propagation over the whole graph with initial messages
  - from leafs to root
  - from root to leafs





Approximate Inference Methods

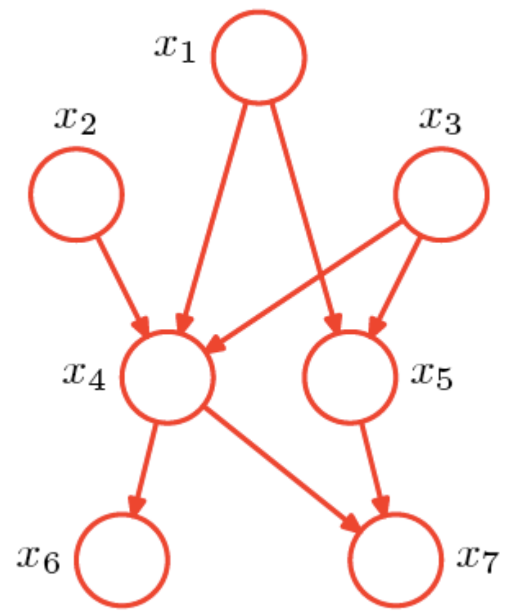
- Loopy Belief Propagation
- Variational Inference
- Expectation Propagation
- Mento Carlo Sampling

Variational Inference

- Approximate posterior distributions by the variational distributions
- Variational Distributions:
  - Factorized approximation based on mean field theory
- Variational message passing
  - updating is done via a local calculation on the graph
  - Applicable to large networks
- Example: Variational Bayes of Gaussian mixture models

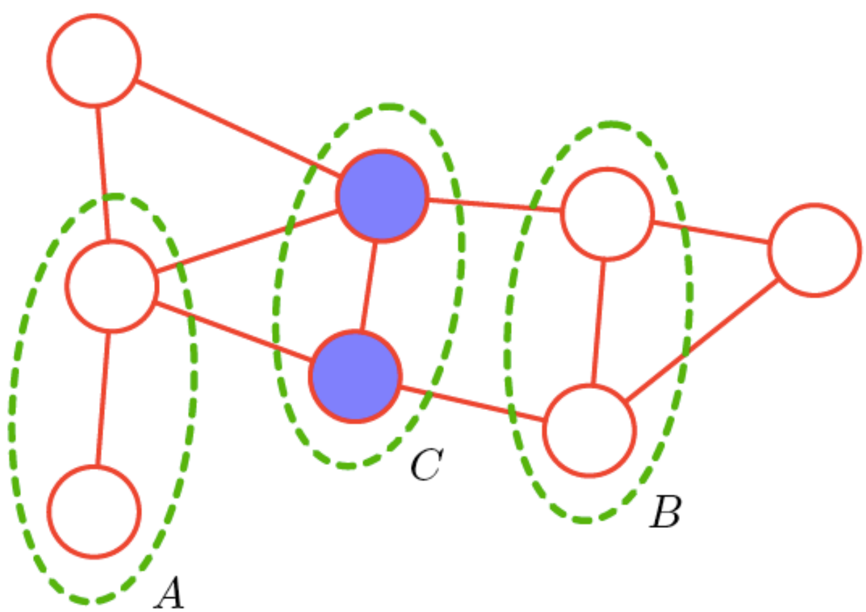
Mento Carlo Sampling

- **Estimate p(x2, x5, x6, x7 | x1, x3, x4)**



Markov Random Fields (1)

- Use an undirected graph to represent joint distributions of random variables
  - Nodes —> random variables (RV)
  - Linking —> conditional dependency
- Conditional independence == simple graph separation



Markov Random Fields (2)

- How to form the joint probability distribution?
  - Potential functions: defined over maximal cliques

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

  - Partition function: normalization constant

$$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C)$$