

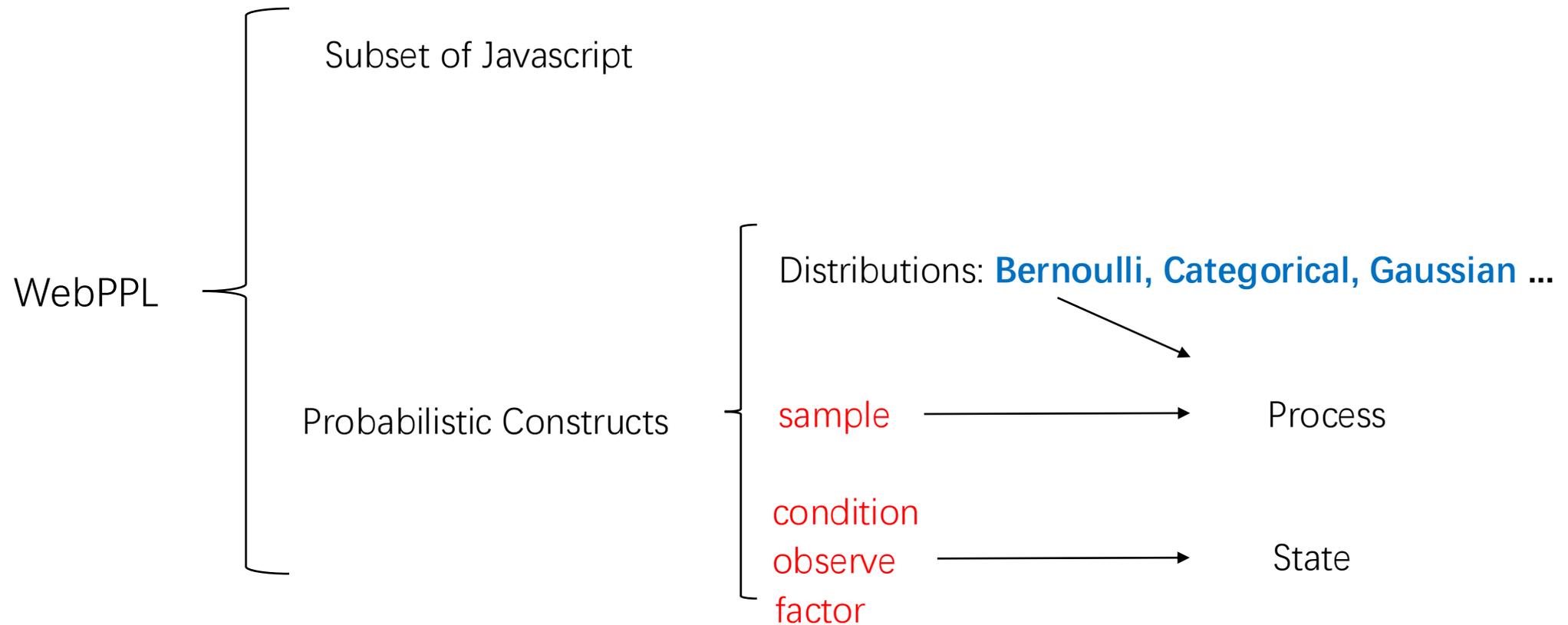
Probabilistic Graphical Models

Xin Zhang

Peking University

Adapted from the slides of “Pattern Recognition and Machine Learning” Chapter 8

Recap of Last Lecture - WebPPL



Recap of Last Lecture - Applications

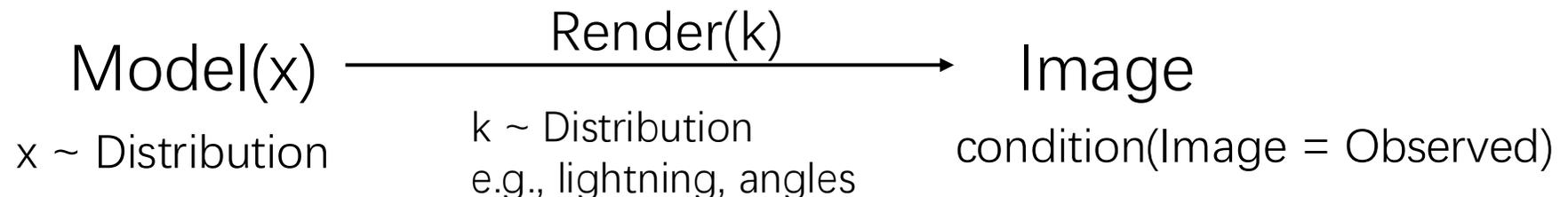
- Bayesian learning models

$$\operatorname{argmax}_{\omega} P(D|\omega) \quad \longrightarrow \quad \operatorname{argmax}_{\omega} P(D|\omega) * P(\omega)$$

- Optimal experiment design

$$\operatorname{argmax}_X \mathbf{E}_{p(X,Y)} (D_{KL}(m | x = X, y = Y || m))$$

- Inverse graphics



Is the following statement correct?

- The Bayesian way to do linear regression is strictly more powerful than the conventional way to do linear regression.
- Yes.

Is the following statement correct?

- In a Bayesian learning model, the more training data there is, the less the prediction results will be affected by the prior distribution of the parameters.
- Yes.

Is the following statement correct?

- When using a Bayesian model, one should always use the most likely result in the prediction distribution.
- No. Sometimes expectations are better.

Is the following statement correct?

- Given two distributions A, B , we have

$$D_{\text{KL}}(A \parallel B) = D_{\text{KL}}(B \parallel A) .$$

- No.

Is the following statement correct?

- The goal of the optimal experiment design is to choose an experiment whose expected result (i.e., output value) is the highest among all experiments.
- No.

What are the applications of inverse graphics?

1. Scene understanding.
 2. Data generation.
 3. Both.
- 3.

Why do we need graphical models?

- How would you represent a probability distribution, so you can
 - Visualize and design a model.
 - Gain insights about relationships between random variables.
 - Do complex inferences.

Naïve Method

A and B are Bernoulli random variables.

	A= True	A= False
B= True	0.25	0.25
B = False	0.25	0.25

Naïve Method

A and B are Bernoulli random variables.

	A= True	A= False
B= True	0.25	0.25
B = False	0.25	0.25

What questions can we ask?

Probabilistic Inference Problems

- Marginal inference:

- Let X be the set of random variables, Y be a subset of it, $Z = X/Y$ then marginal inference is to compute

$$P(Y = V_Y) = \sum_{V_{Z_i}} P(Y = V_Y, Z = V_{Z_i})$$

- Conditional inference:

- Let X be the set of random variables, Y and W be subsets of it then conditional inference is to compute

$$P(Y = V_Y | W = V_W)$$

Probabilistic Inference in Table Method

	A= True	A= False
B= True	0.25	0.25
B = False	0.25	0.25

$$P(A = \text{True}) = P(A = \text{True}, B = \text{False}) + P(A = \text{True}, B = \text{True})$$

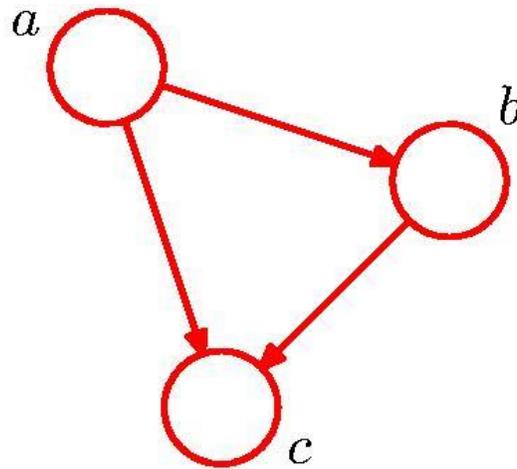
Probabilistic Inference in Table Method

	A= True	A= False
B= True	0.25	0.25
B = False	0.25	0.25

$$P(A = \text{True} \mid B = \text{True}) = \frac{P(A = \text{True}, B = \text{True})}{P(A = \text{True}, B = \text{True}) + P(A = \text{False}, B = \text{True})}$$

Bayesian Networks

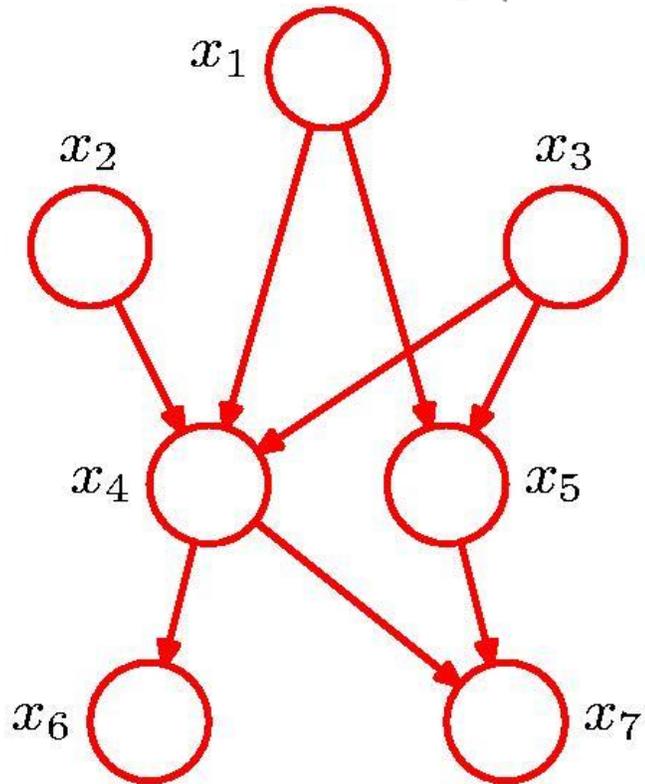
- Directed Acyclic Graph (DAG)



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

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

Bayesian Networks

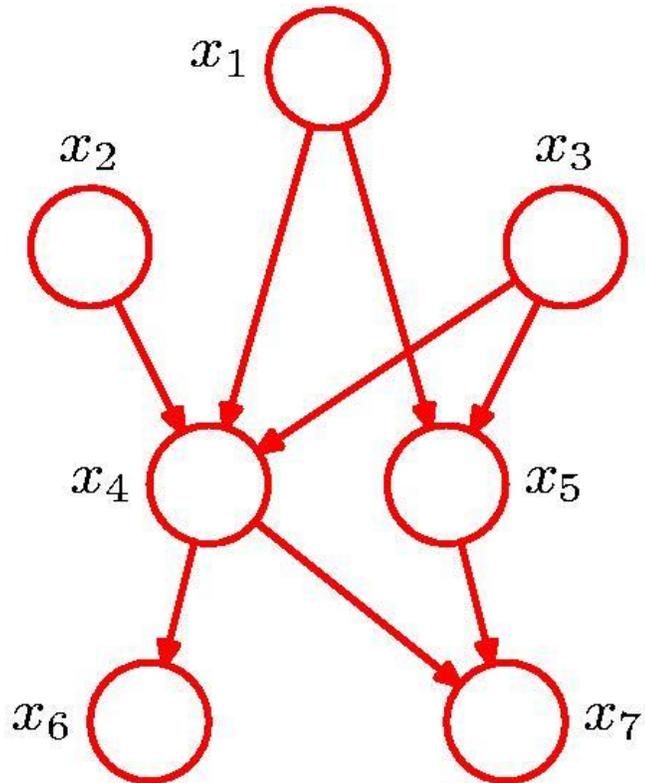


$$p(x_1, \dots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

General Factorization

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

Bayesian Networks



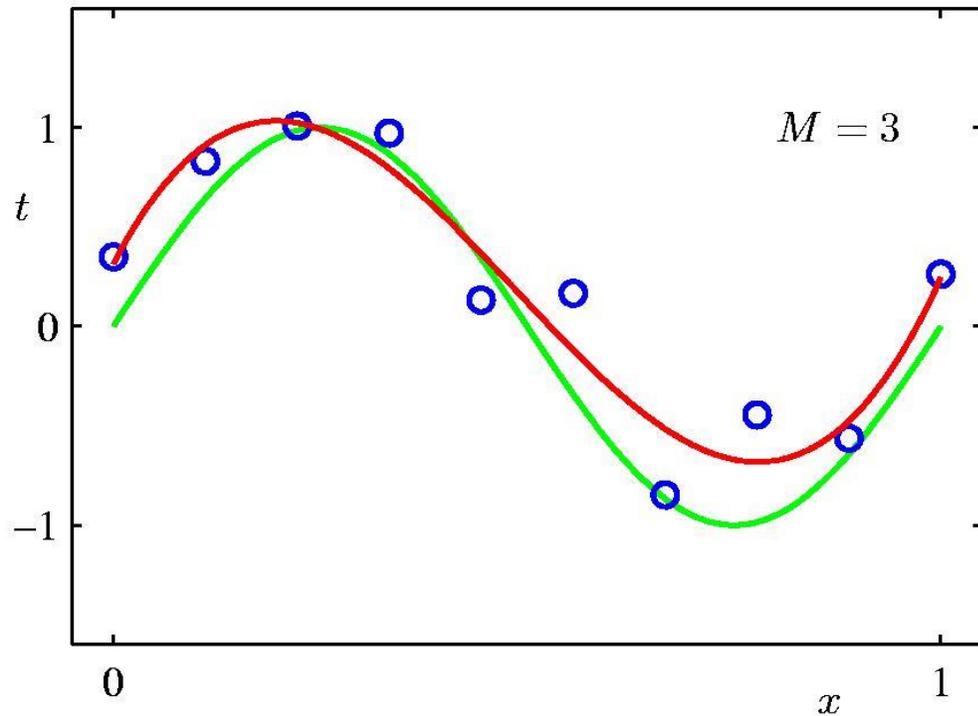
Are x_1 and x_2 independent?

What about x_4 and x_5 ?

What about x_4 and x_5 when x_1 is fixed?

We will talk about dependence later!

Example Application: Bayesian Curve Fitting



Polynomial

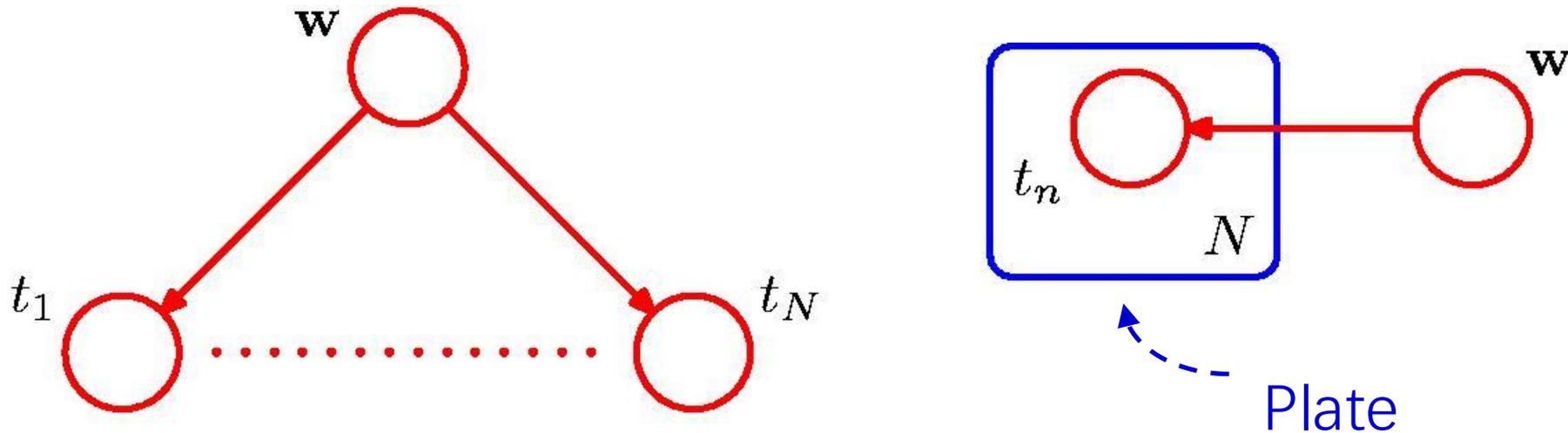
$$y(x, \mathbf{w}) = \sum_{j=0}^M w_j x^j$$

\mathbf{x} is the set of training inputs
while \mathbf{t} is their predictions.

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

Example Application: Bayesian Curve Fitting

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



Example Application: Bayesian Curve Fitting

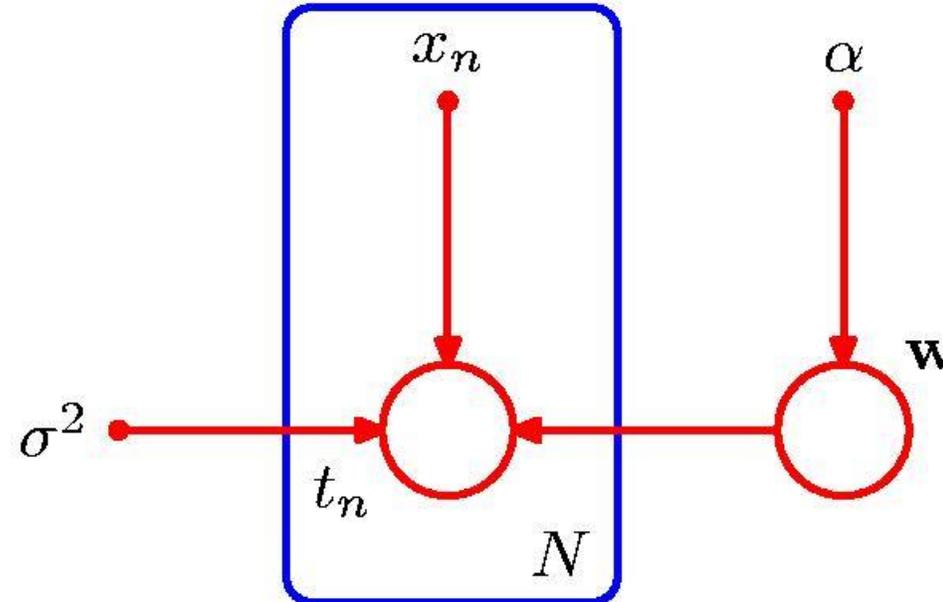
- Input variables and explicit hyperparameters

- α is the parameter of the parameter. For example:

$$w_i \sim N(\alpha, 1)$$

- σ^2 is the variance of the gaussian noise in training.

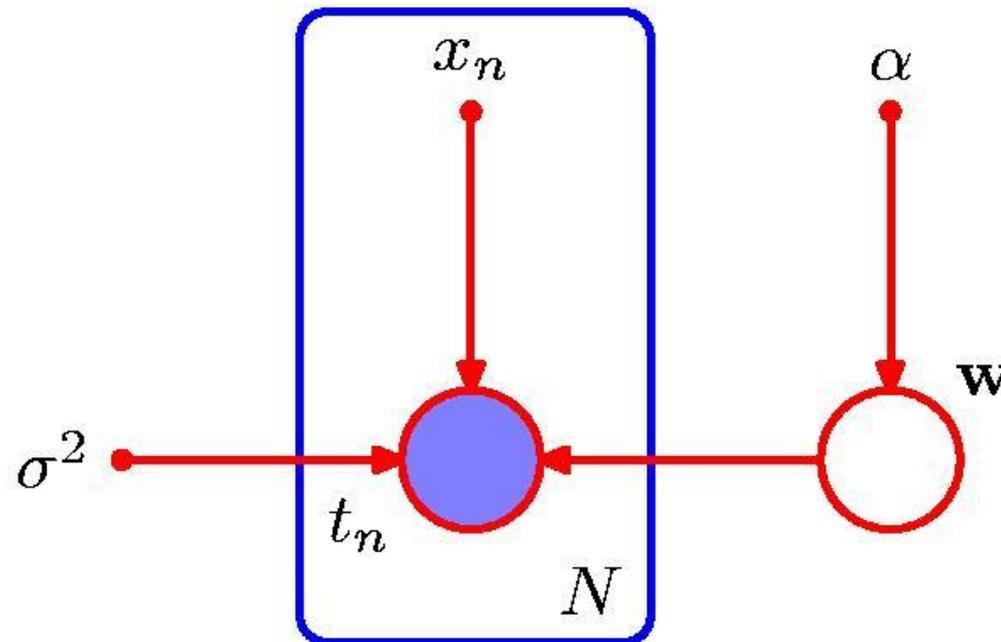
$$p(\mathbf{t}, \mathbf{w} | \mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w} | \alpha) \prod_{n=1}^N p(t_n | \mathbf{w}, x_n, \sigma^2).$$



Bayesian Curve Fitting — Learning

- Condition on data

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

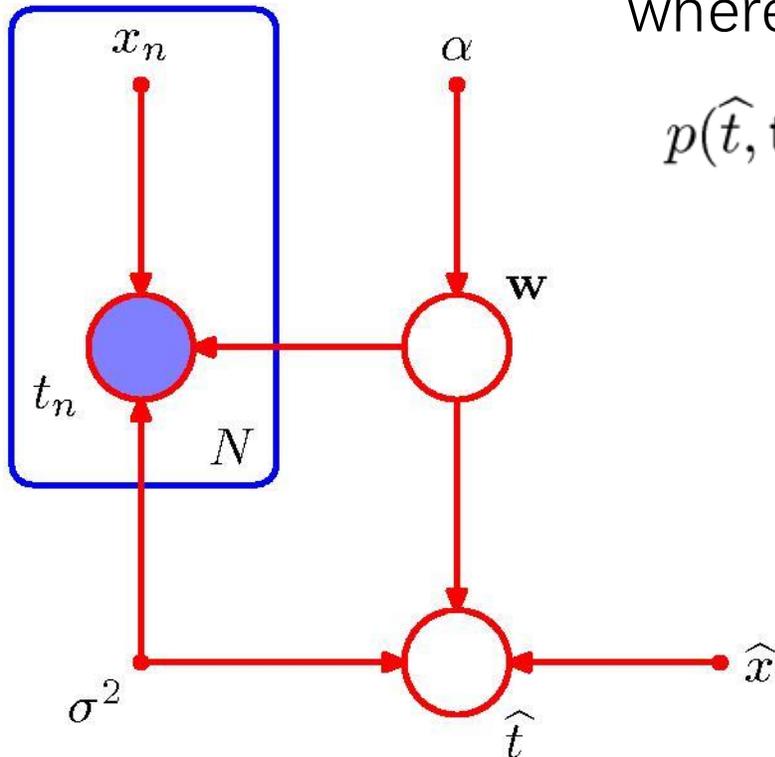


Bayesian Curve Fitting — Prediction

Predictive distribution: $p(\hat{t}|\hat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\hat{t}, \mathbf{t}, \mathbf{w}|\hat{x}, \mathbf{x}, \alpha, \sigma^2) d\mathbf{w}$

where

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

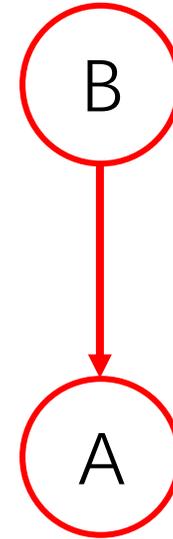
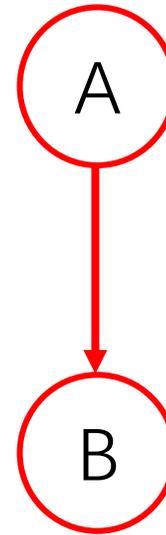


Which model is correct?

A: whether the school bus has a crash

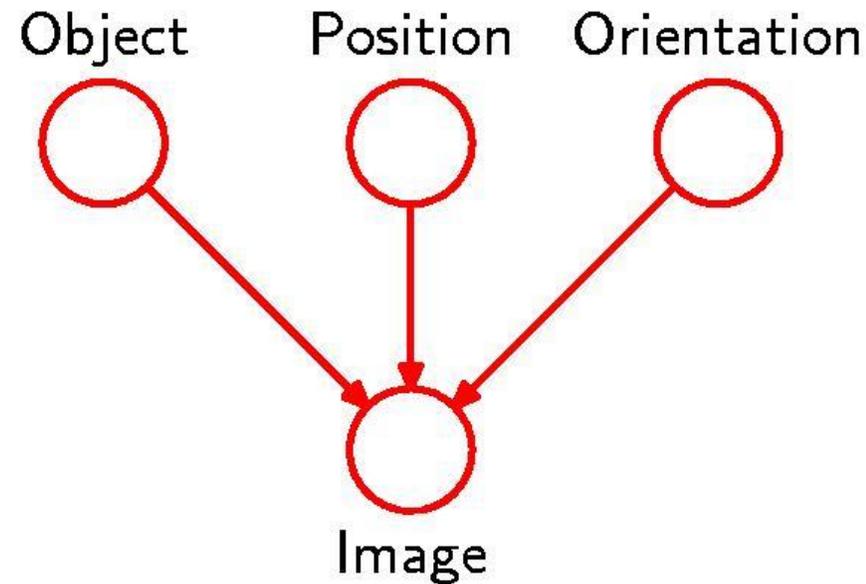
B: whether the teacher is late for the class

	A= True	A= False
B= True	0.09	0.09
B = False	0.01	0.81



Generative Models

- Causal process for generating images



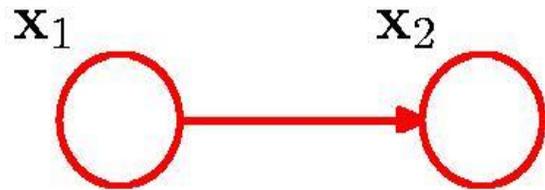
We will talk about causality in a later lecture!

Two Special Cases

- Discrete variables
- Gaussian variables

Discrete Variables

- General joint distribution: $K^2 - 1$ parameters



$$p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}) = \prod_{k=1}^K \prod_{l=1}^K \mu_{kl}^{x_{1k} x_{2l}}$$

- Independent joint distribution: $2(K - 1)$ parameters



$$\hat{p}(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\mu}) = \prod_{k=1}^K \mu_{1k}^{x_{1k}} \prod_{l=1}^K \mu_{2l}^{x_{2l}}$$

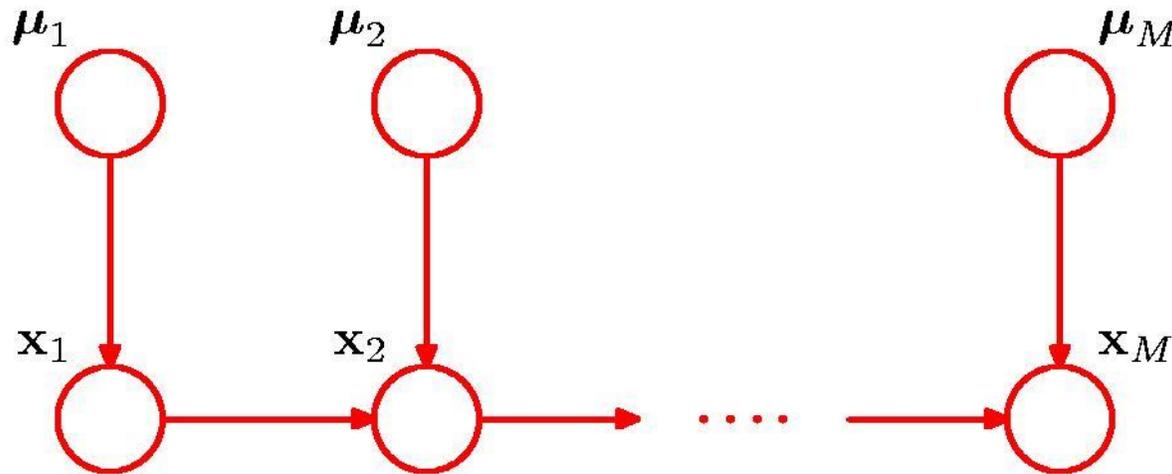
Discrete Variables

General joint distribution over M variables:
 $K^M - 1$ parameters

M -node Markov chain: $K - 1 + (M - 1) K(K - 1)$
parameters



Discrete Variables: Bayesian Parameters



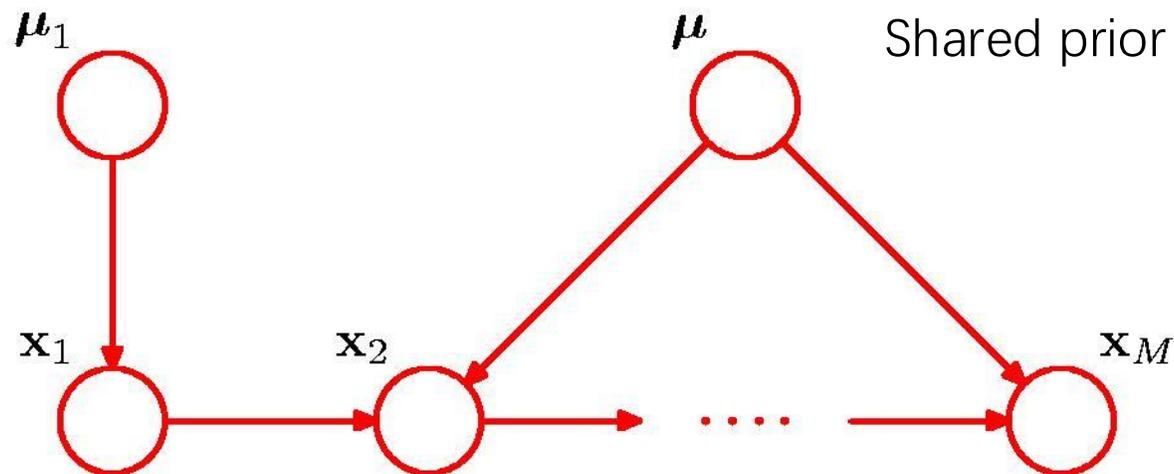
$$p(\{\mathbf{x}_m, \boldsymbol{\mu}_m\}) = p(\mathbf{x}_1 | \boldsymbol{\mu}_1) p(\boldsymbol{\mu}_1) \prod_{m=2}^M p(\mathbf{x}_m | \mathbf{x}_{m-1}, \boldsymbol{\mu}_m) p(\boldsymbol{\mu}_m)$$

$$p(\boldsymbol{\mu}_m) = \text{Dir}(\boldsymbol{\mu}_m | \boldsymbol{\alpha}_m)$$

Discrete Variables: Bayesian Parameters

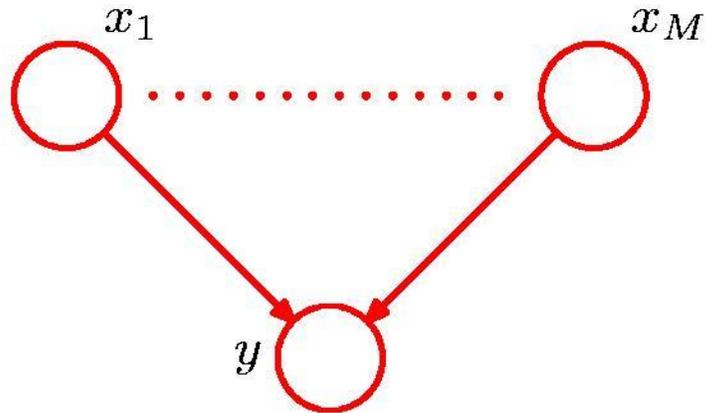
- Why are Dirichlet distributions used?
 - They are conjugate priors for categorical and binomial distributions.
- Further reading: <https://towardsdatascience.com/dirichlet-distribution-a82ab942a879>

Discrete Variables: Bayesian Parameters



$$p(\{x_m\}, \mu_1, \mu) = p(x_1 | \mu_1) p(\mu_1) \prod_{m=2}^M p(x_m | x_{m-1}, \mu) p(\mu)$$

Parameterized Conditional Distributions



If x_1, \dots, x_M are discrete,
K-state variables,
 $p(y = 1 | x_1, \dots, x_M)$ in
general has $O(K^M)$
parameters.

The parameterized form

$$p(y = 1 | x_1, \dots, x_M) = \sigma \left(w_0 + \sum_{i=1}^M w_i x_i \right) = \sigma(\mathbf{w}^T \mathbf{x})$$

requires only $M + 1$ parameters

Linear-Gaussian Models

- Directed Graph

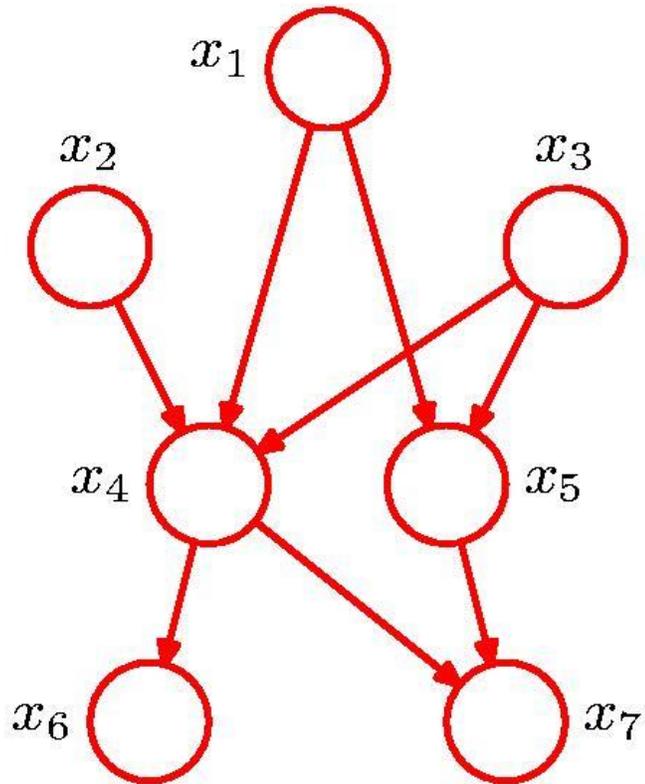
$$p(x_i | \text{pa}_i) = \mathcal{N} \left(x_i \mid \sum_{j \in \text{pa}_i} w_{ij} x_j + b_i, v_i \right)$$

Each node is Gaussian, the mean is a linear function of the parents.

- Vector-valued Gaussian Nodes

$$p(\mathbf{x}_i | \text{pa}_i) = \mathcal{N} \left(\mathbf{x}_i \mid \sum_{j \in \text{pa}_i} \mathbf{W}_{ij} \mathbf{x}_j + \mathbf{b}_i, \mathbf{\Sigma}_i \right)$$

Recall This Graph



Are x_1 and x_2 independent?

What about x_4 and x_5 ?

What about x_4 and x_5 when x_1 is fixed?

We will talk about dependence now!

Conditional Independence

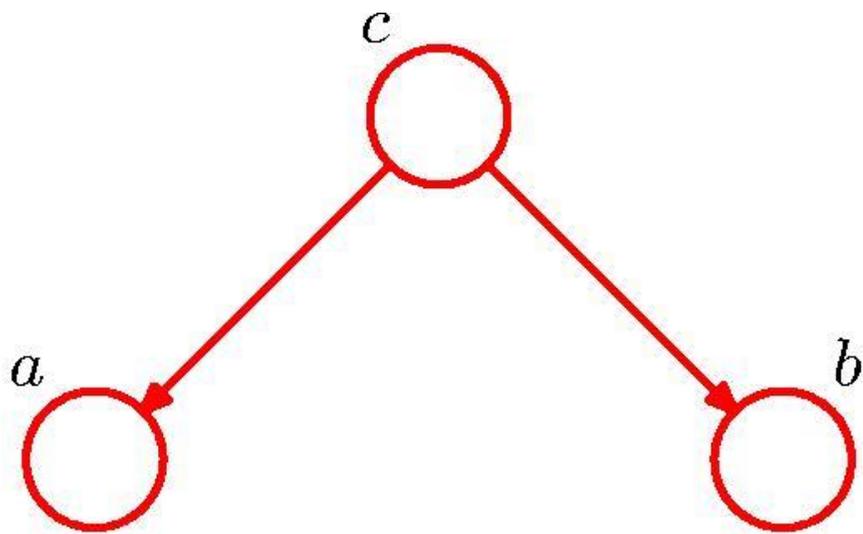
- a is independent of b given c

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

- Equivalently
$$\begin{aligned} p(a, b|c) &= p(a|b, c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}$$

- Notation
$$a \perp\!\!\!\perp b \mid c$$

Conditional Independence: Example 1

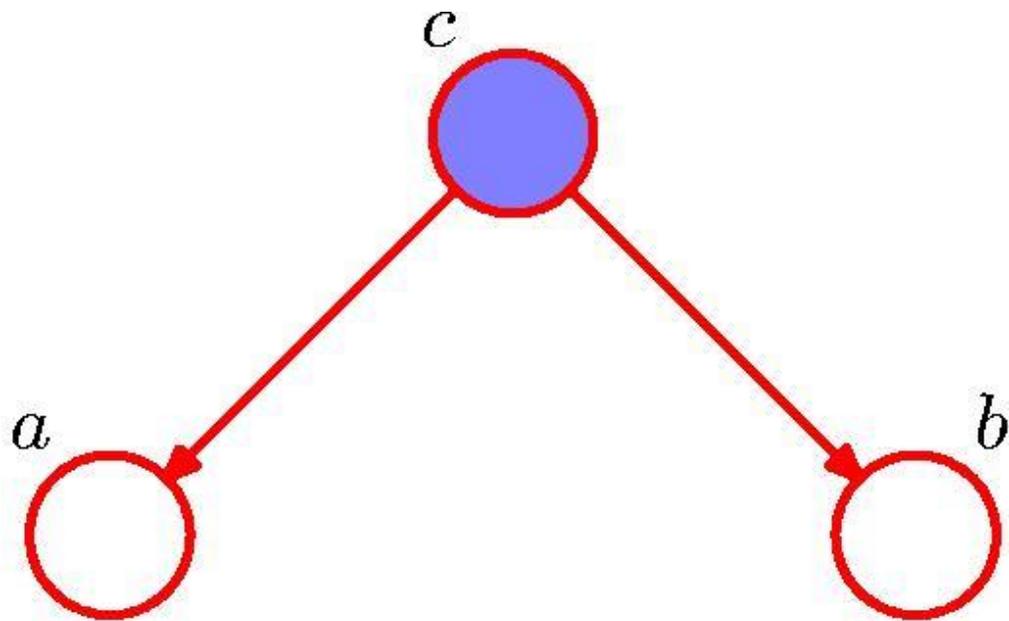


$$p(a, b, c) = p(a|c)p(b|c)p(c)$$

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

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

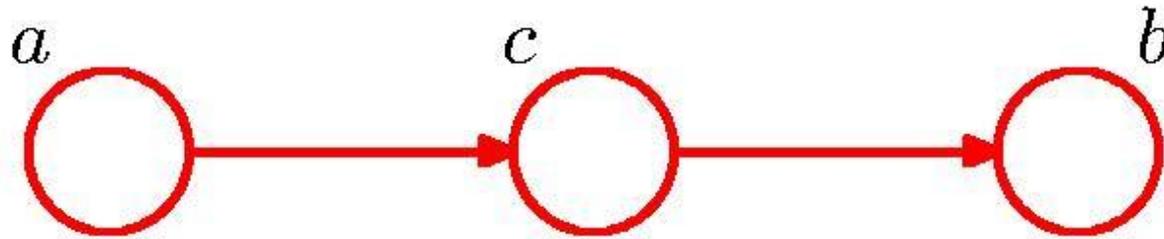
Conditional Independence: Example 1



$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

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

Conditional Independence: Example 2

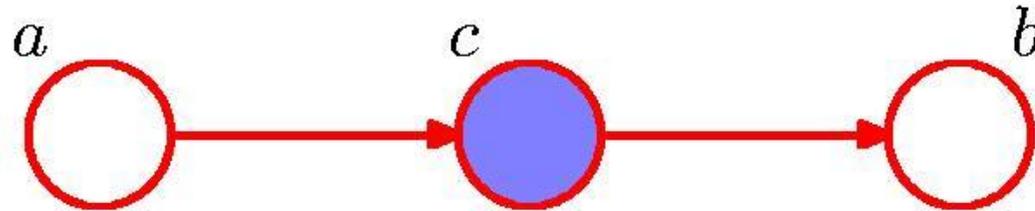


$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

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

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

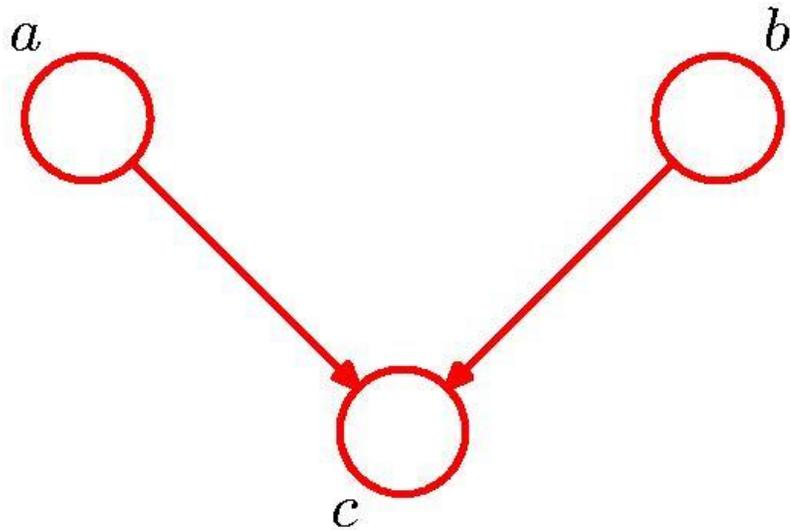
Conditional Independence: Example 2



$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

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

Conditional Independence: Example 3



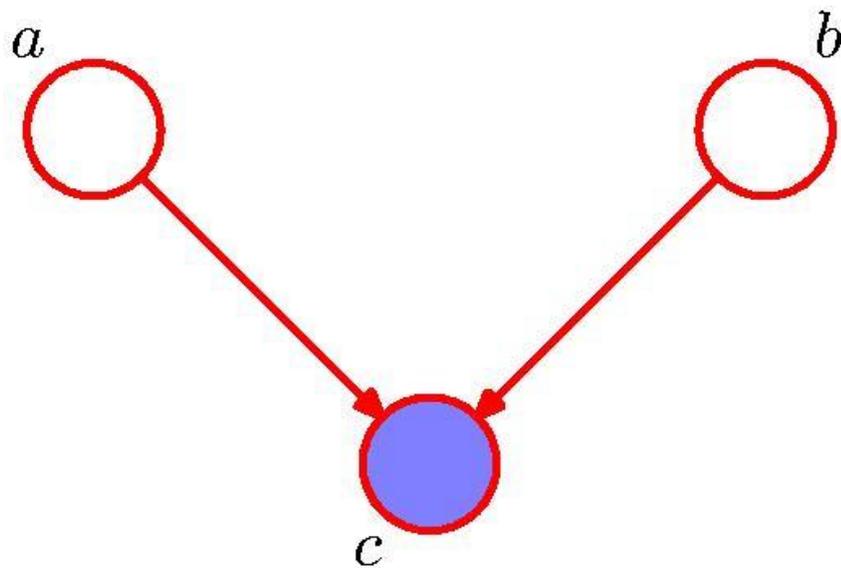
$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

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

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

- Note: this is the opposite of Example 1, with c unobserved.

Conditional Independence: Example 3



$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(b)p(c|a, b)}{p(c)} \end{aligned}$$

$$a \not\perp b \mid c$$

Note: this is the opposite of Example 1, with c observed.

“Am I out of fuel?”

$$p(G = 1|B = 1, F = 1) = 0.8$$

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

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

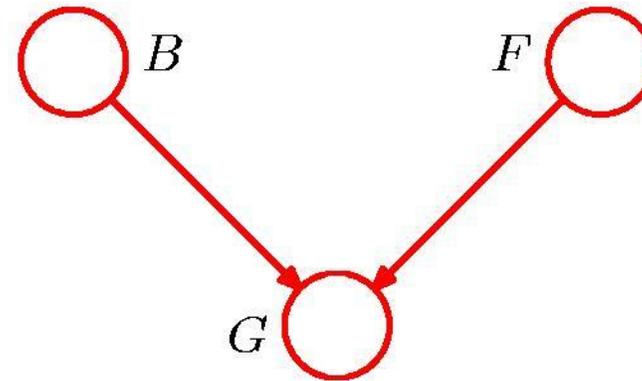
$$p(G = 1|B = 0, F = 0) = 0.1$$

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

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

and hence

$$p(F = 0) = 0.1$$



B = Battery (0=flat, 1=fully charged)

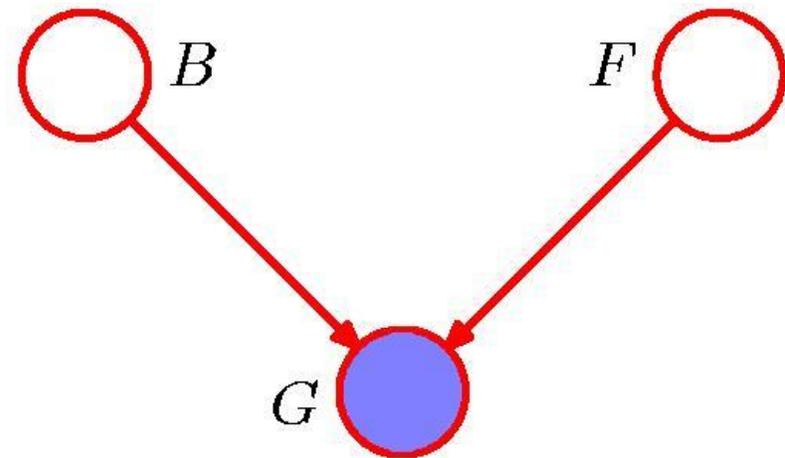
F = Fuel Tank (0=empty, 1=full)

G = Fuel Gauge Reading
(0=empty, 1=full)

“Am I out of fuel?”

$$p(F = 0 | G = 0) = \frac{p(G = 0 | F = 0)p(F = 0)}{p(G = 0)}$$

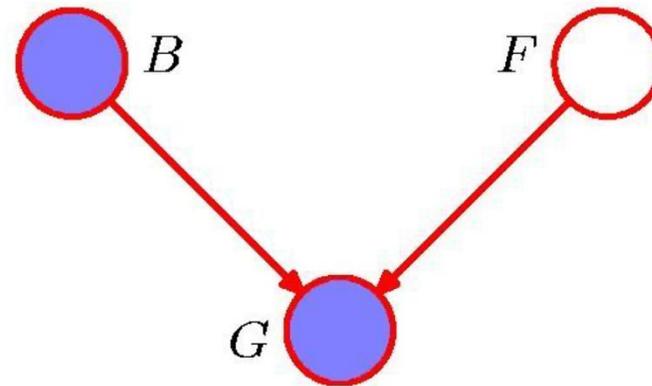
$$\simeq 0.257$$



Probability of an empty tank increased by observing $G = 0$.

What if now we also know the battery is flat?

“Am I out of fuel?”



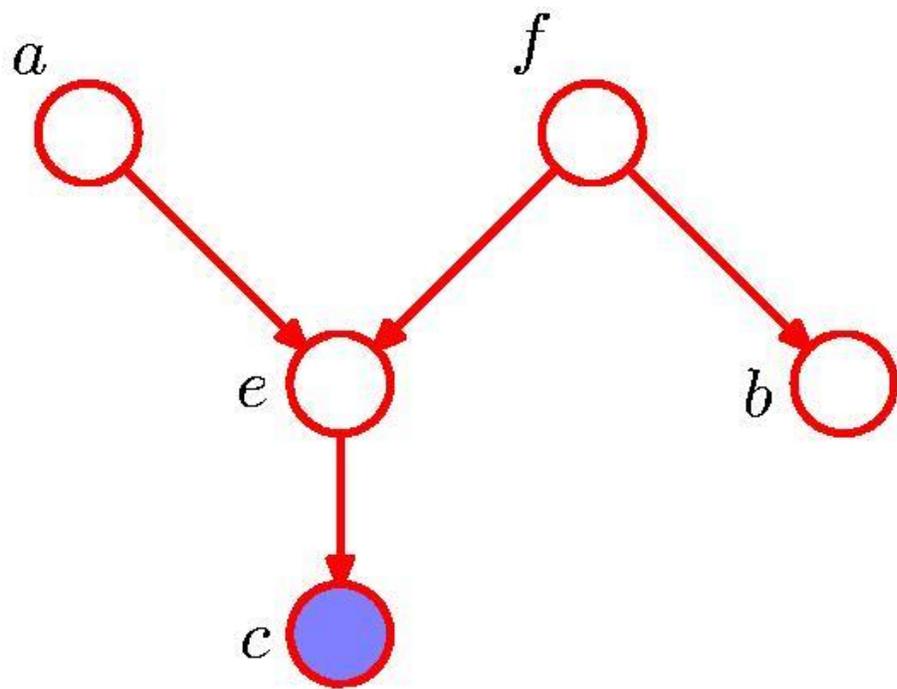
$$\begin{aligned} p(F = 0 | G = 0, B = 0) &= \frac{p(G = 0 | B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0,1\}} p(G = 0 | B = 0, F)p(F)} \\ &\simeq 0.111 \end{aligned}$$

Probability of an empty tank reduced by observing $B = 0$.
This referred to as “explaining away”.

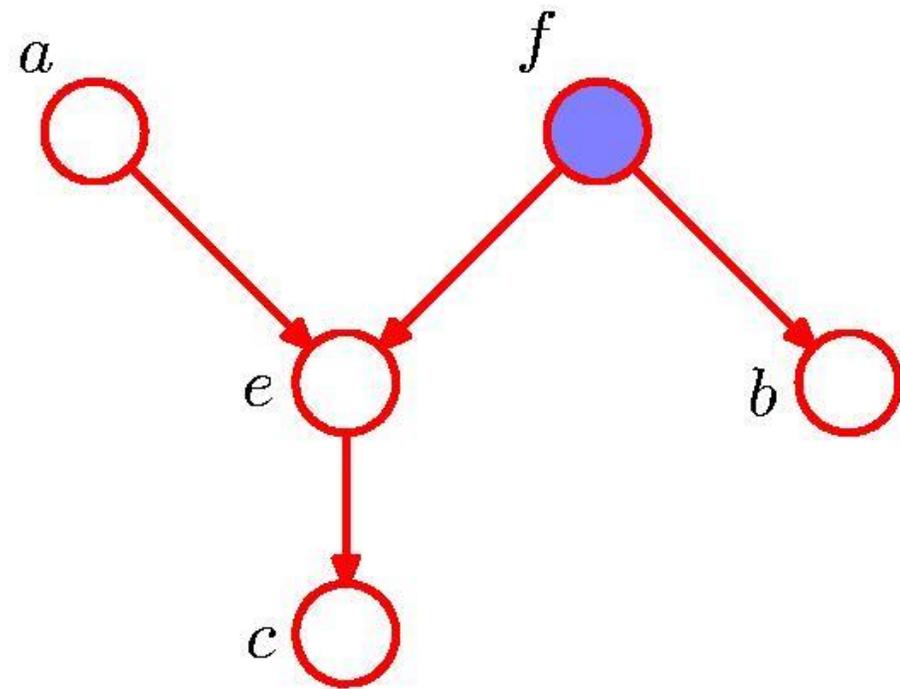
D-separation

- **A**, **B**, and **C** are non-intersecting subsets of nodes in a directed graph.
- A path from **A** to **B** is blocked if it contains a node such that either
 - a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set **C**, or
 - b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set **C**.
- If all paths from **A** to **B** are blocked, **A** is said to be d-separated from **B** by **C**.
- If **A** is d-separated from **B** by **C**, the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$.

D-separation: Example

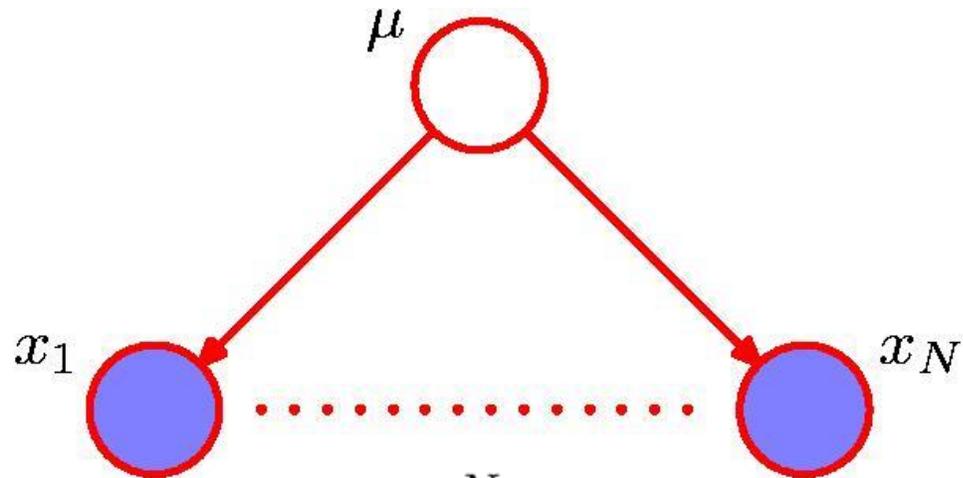


$$a \not\perp\!\!\!\perp b \mid c$$



$$a \perp\!\!\!\perp b \mid f$$

D-separation: I.I.D. Data



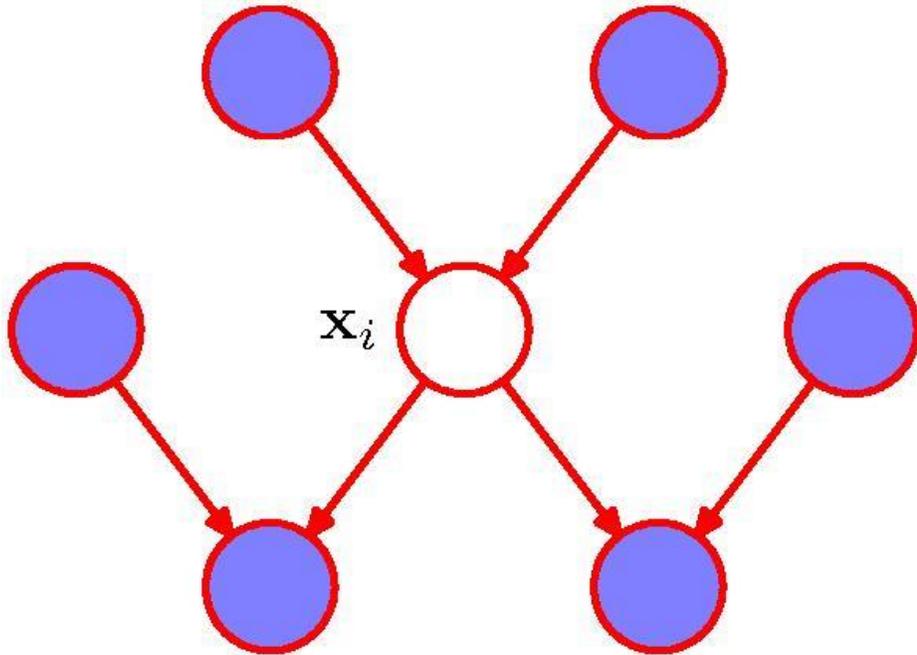
$$p(\mathcal{D}|\mu) = \prod_{n=1}^N p(x_n|\mu)$$

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu) d\mu \neq \prod_{n=1}^N p(x_n)$$

Question

- What can D-separation be used for?

The Markov Blanket



$$\begin{aligned}
 p(\mathbf{x}_i | \mathbf{x}_{\{j \neq i\}}) &= \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_M)}{\int p(\mathbf{x}_1, \dots, \mathbf{x}_M) d\mathbf{x}_i} \\
 &= \frac{\prod_k p(\mathbf{x}_k | \text{pa}_k)}{\int \prod_k p(\mathbf{x}_k | \text{pa}_k) d\mathbf{x}_i}
 \end{aligned}$$

Factors independent of \mathbf{x}_i cancel between numerator and denominator.

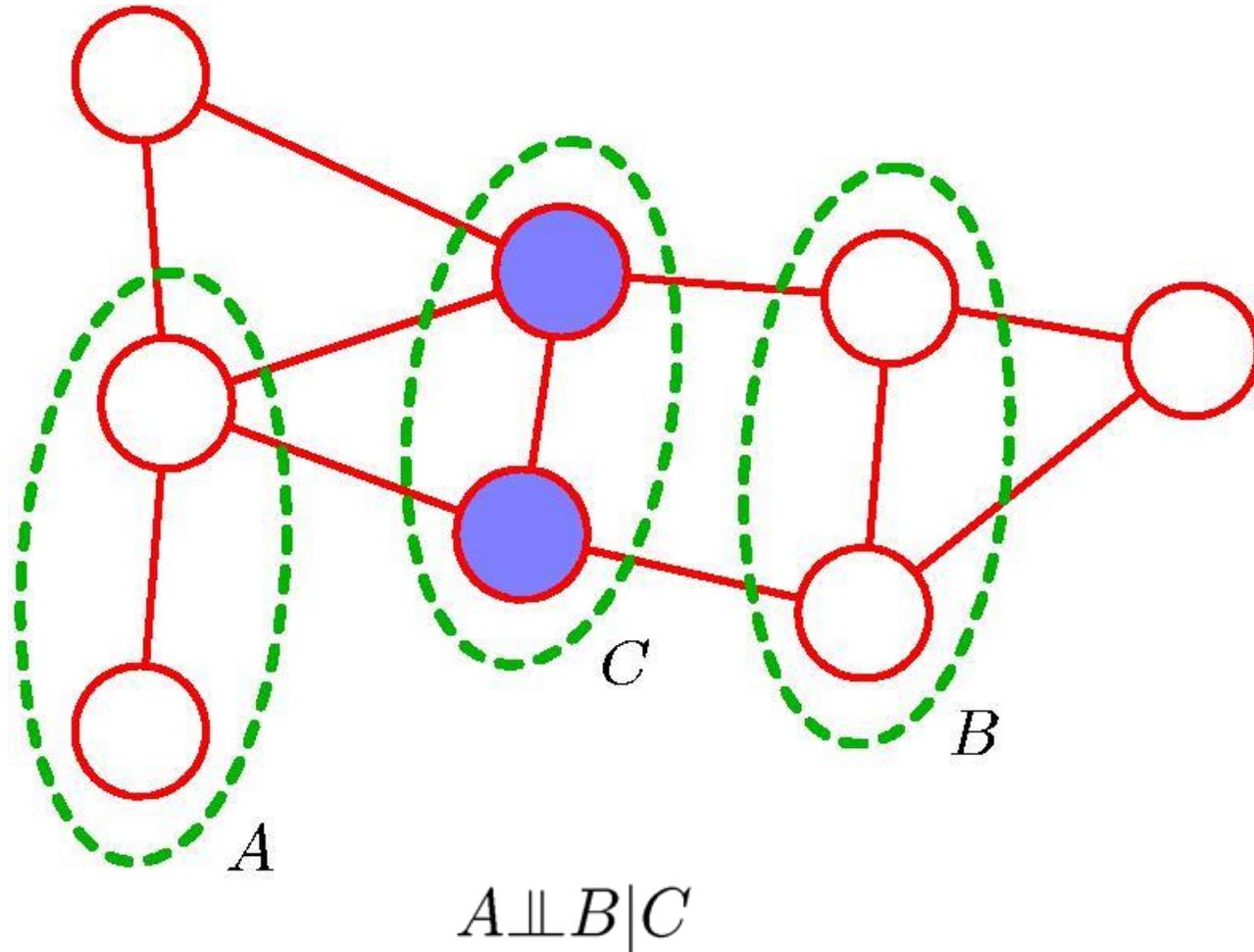
Bayesian Networks: Summary

- Directed
- Factorizations of conditional probabilities
- Reason about the relationships between different variables using conditional independence

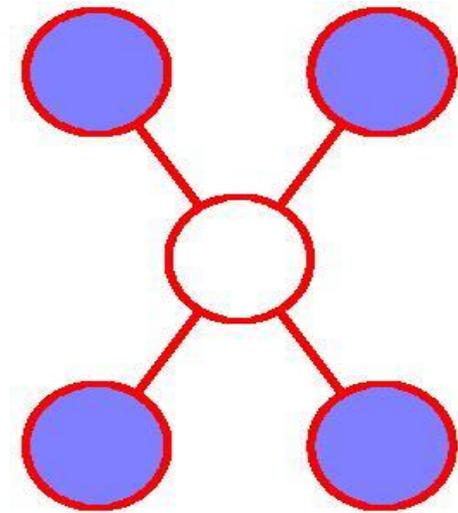
Markov Random Fields

- Undirected
- Markov networks
- One motivation: reasoning about conditional independence is subtle in Bayesian networks. Can we have something simpler?

Markov Random Fields



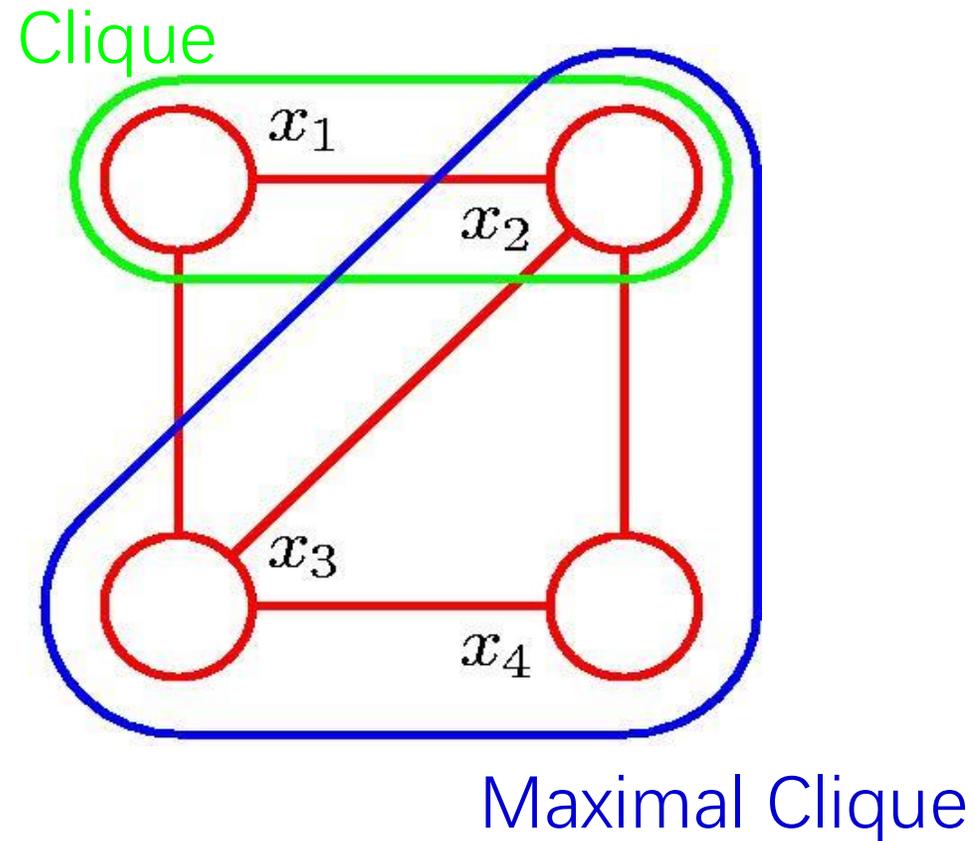
Markov Blanket



Markov Random Fields: Intuitions

- If x and y are not directly connected, then they should be independent conditioning on the other variables
- $P(x, y | V / \{x, y\}) = P(x | V / \{x, y\}) * P(y | V / \{x, y\})$
- x and y should not appear in the same factor
- We should put nodes that are directly connected in the same factor

Cliques and Maximal Cliques



Joint Distribution

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

- where $\psi_C(\mathbf{x}_C)$ is the potential over maximal clique C and

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

- is the normalization coefficient; note: M K -state variables $\rightarrow K^M$ terms in Z .
- In general, we only require potentials to be positive. One example: Energies and the Boltzmann distribution

$$\psi_C(\mathbf{x}_C) = \exp \{-E(\mathbf{x}_C)\}$$

Factorization and Conditional Independence

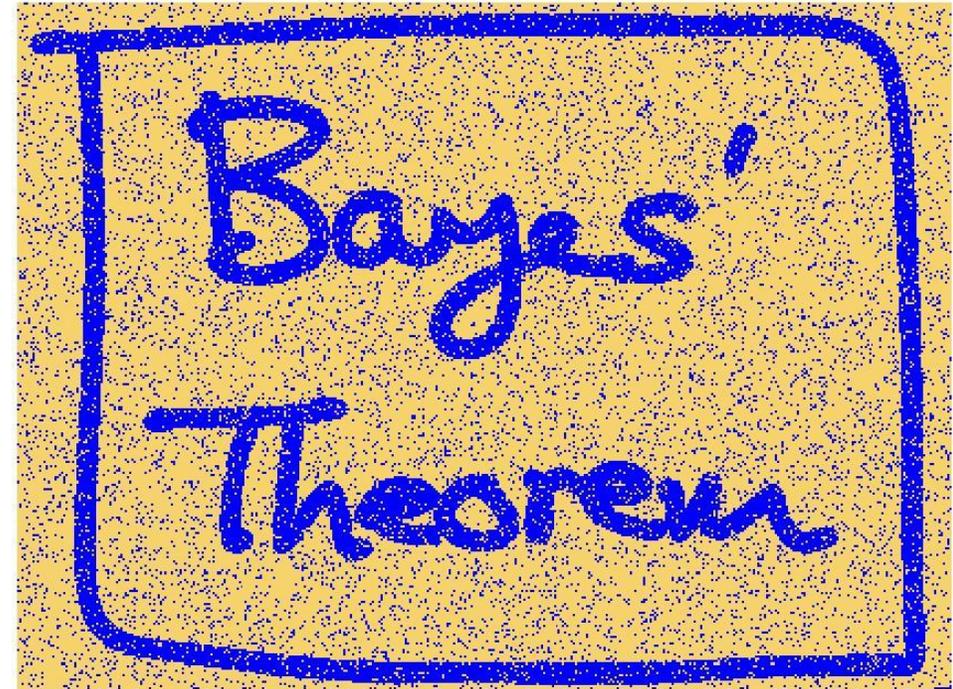
- Given a graph (potential function unknown), let UI be the distributions whose conditional independence fits the graph
- Let UF be the subset of UI that can be expressed in the factorization form
- We have $UF = UI$: the Hammersley-Clifford theorem (Clifford, 1990)

Illustration: Image De-Noising



Original Image

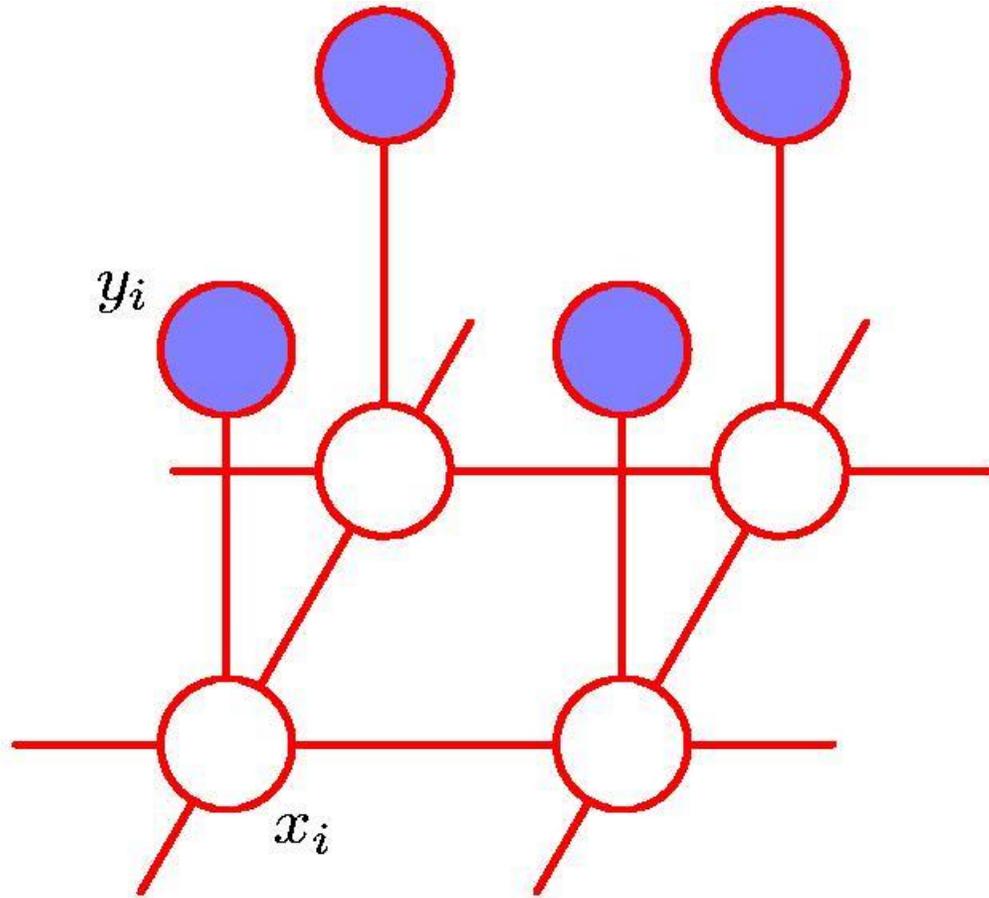
$$x_i \in \{-1, 1\}$$



Noisy Image

$$y_j \in \{-1, 1\}$$

Illustration: Image De-Noising



$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

Special Case: Conditional Random Field

- There two sets of variables X and Y
- The conditional distribution $Y | X$ forms a Markov Random Field
- By observing Y , predict X
- Example: text segmentation: X : text, Y : segments

Summary

- Bayesian networks
 - Directed
 - Factorization of conditional probabilities
 - Conditional independence: D-separation
- Markov random fields
 - Undirected
 - Factorization over maximum cliques

