STA 4273H: Statistical Machine Learning

Russ Salakhutdinov

Department of Computer Science
Department of Statistical Sciences
rsalakhu@cs.toronto.edu
http://www.cs.utoronto.ca/~rsalakhu/

Lecture 4

Graphical Models

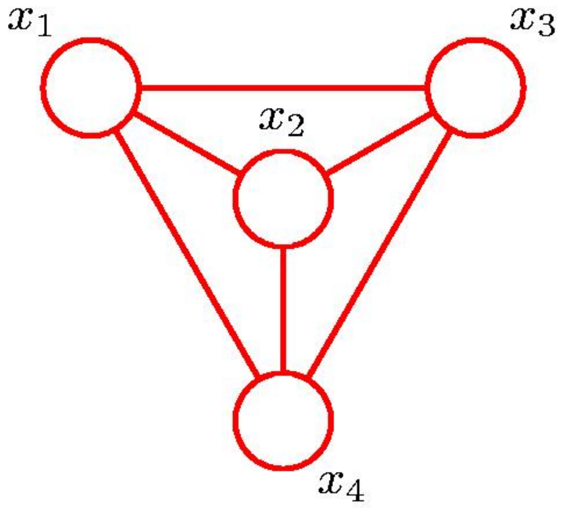
- Probabilistic graphical models provide a powerful framework for representing dependency structure between random variables.

- Graphical models offer several useful properties:

  - They provide a simple way to visualize the structure of a probabilistic model and can be used to motivate new models.
  - They provide various insights into the properties of the model, including conditional independence.
  - Complex computations (e.g. inference and learning in sophisticated models) can be expressed in terms of graphical manipulations.

Graphical Models

- A graph contains a set of nodes (vertices) connected by links (edges or arcs)



- In a probabilistic graphical model, each node represents a random variable, and links represent probabilistic dependencies between random variables.

- The graph specifies the way in which the joint distribution over all random variables decomposes into a product of factors, where each factor depends on a subset of the variables.

- Two types of graphical models:
  - Bayesian networks, also known as Directed Graphical Models (the links have a particular directionality indicated by the arrows)
  - Markov Random Fields, also known as Undirected Graphical Models (the links do not carry arrows and have no directional significance).

- Hybrid graphical models that combine directed and undirected graphical models, such as Deep Belief Networks.

Bayesian Networks

- Directed Graphs are useful for expressing causal relationships between random variables.

- Let us consider an arbitrary joint distribution $p(a, b, c)$ over three random variables a,b, and c.

- Note that at this point, we do not need to specify anything else about these variables (e.g. whether they are discrete or continuous).

- By application of the product rule of probability (twice), we get

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

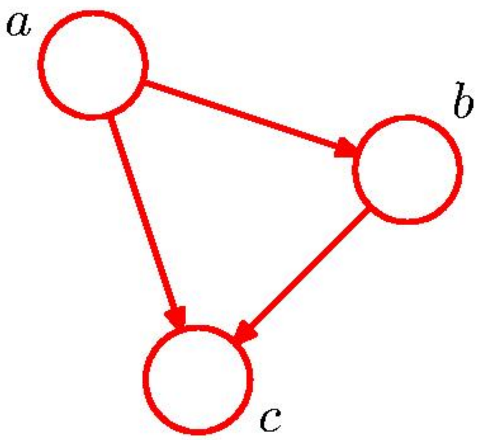
- This decomposition holds for any choice of the joint distribution.

Bayesian Networks

- By application of the product rule of probability (twice), we get

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

- Represent the joint distribution in terms of a simple graphical model:



- Introduce a node for each of the random variables.
- Associate each node with the corresponding conditional distribution in above equation.
- For each conditional distribution we add directed links to the graph from the nodes corresponding to the variables on which the distribution is conditioned.

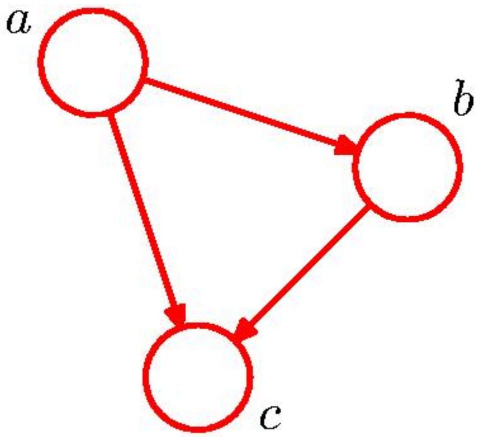
- Hence for the factor $p(c|a,b)$, there will be links from nodes a and b to node c.
- For the factor $p(a)$, there will be no incoming links.

Bayesian Networks

- By application of the product rule of probability (twice), we get

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

- If there is a link going from node a to node b, then we say that:



  - node a is a parent of node b.
  - node b is a child of node a.

- For the decomposition, we choose a specific ordering of the random variables: a,b,c.

- If we chose a different ordering, we would get a different graphical representation (we will come back to that point later).

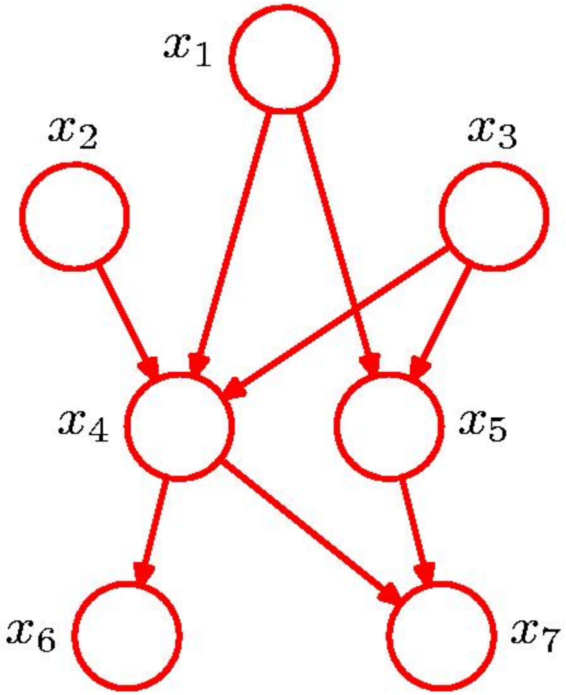
- The joint distribution over K variables factorizes:

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

- If each node has incoming links from all lower numbered nodes, then the graph is fully connected; there is a link between all pairs of nodes.

Bayesian Networks

- Absence of links conveys certain information about the properties of the class of distributions that the graph conveys.



- Note that this graph is not fully connected (e.g. there is no link from x_1 to x_2).

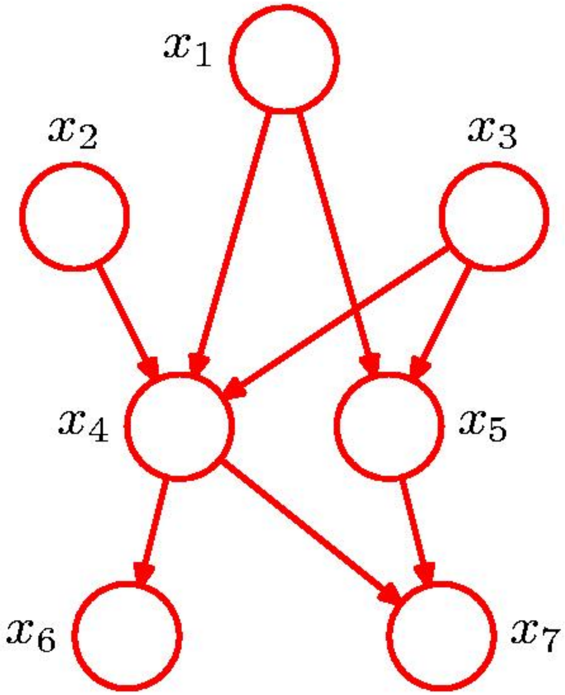
- The joint distribution over $x_1,\dots,x_7$ can be written as a product of a set of conditional distributions.

$$\begin{aligned} p(x_1, \ldots, x_7) \quad &= \quad p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ & \qquad p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5) \end{aligned}$$

- Note that according to the graph, x_5 will be conditioned only on x_1 and x_3.

Factorization Property

- The joint distribution defined by the graph is given by the product of a conditional distribution for each node conditioned on its parents:



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

where pa_k denotes a set of parents for the node x_k.

- This equation expresses a key factorization property of the joint distribution for a directed graphical model.

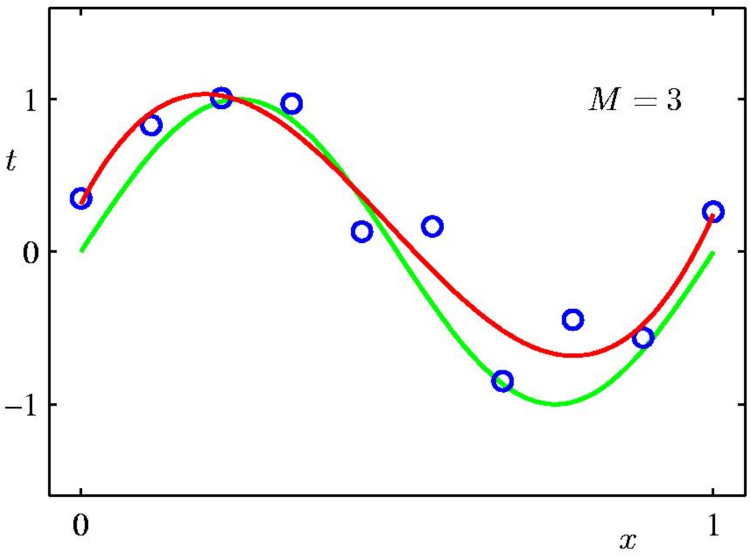
- Important restriction: There must be **no directed cycles!**

- Such graphs are also called directed acyclic graphs (DAGs).

Bayesian Curve Fitting

- As an example, remember Bayesian polynomial regression model:

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



- We are given inputs $\mathbf{X} = \{x_1, x_2, ..., x_N\}$ and target values $\mathbf{t} = [t_1, t_2, ..., t_N]^T$.

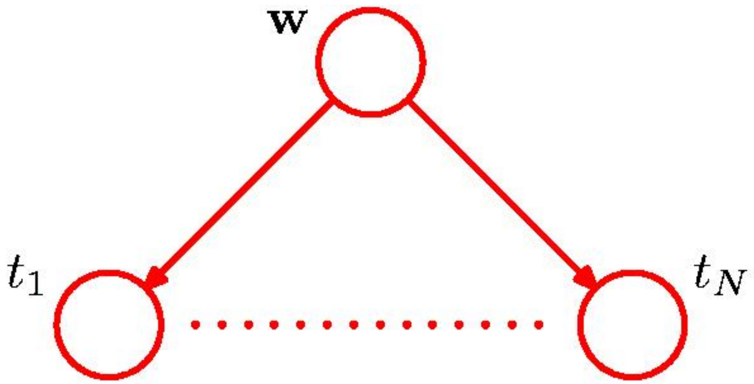
- Given the prior over parameters, the joint distribution is given by:

$$p(\mathbf{t}, \mathbf{w}|\mathbf{X}) = \underset{\nearrow \atop \text{Prior term}}{p(\mathbf{w})} \underbrace{\prod_{i=1}^{N} p(t_n|y(\mathbf{w}, x_n))}_{\text{Likelihood term}}.$$

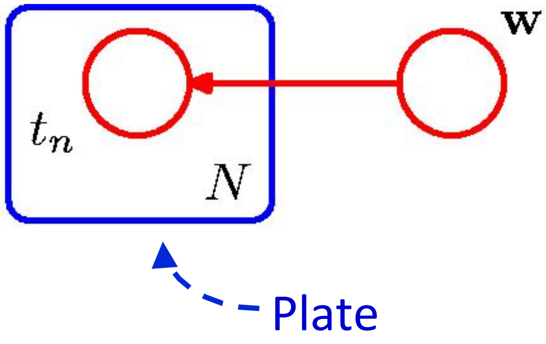
Graphical Representation

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

- This distribution can be represented as a graphical model.



- Same representation using plate notation.



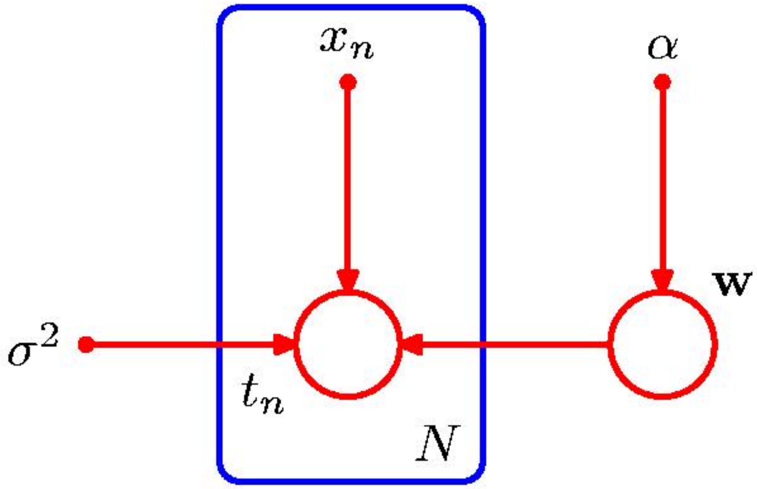
- Compact representation: we introduce a plate that represents N nodes of which only a single example t_n is shown explicitly.

- Note that w and $\mathbf{t} = [t_1, t_2, ..., t_N]^T$ represent random variables.

Graphical Representation

- It will often be useful to make the parameters of the model as well as random variables be explicit.

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



$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha \boldsymbol{I}),$$

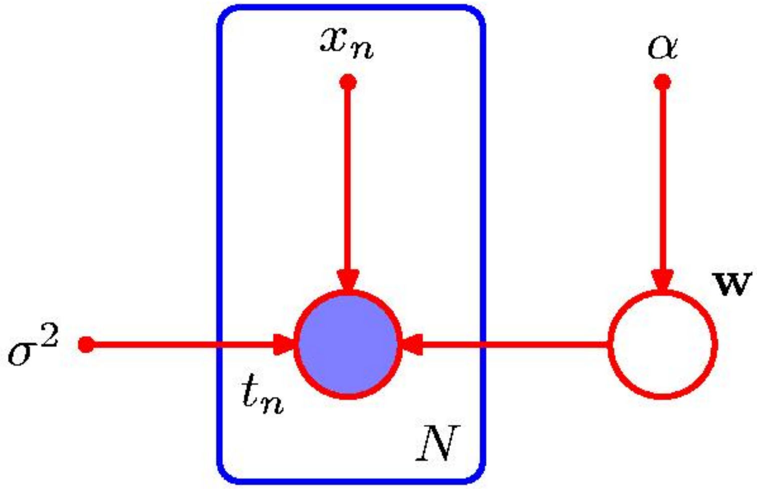
$$p(t_n|\mathbf{w}, x_n, \sigma^2) = \mathcal{N}(t_n|y(\mathbf{w}, x_n), \sigma^2),$$

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

- Random variables will be denoted by open circles and deterministic parameters will be denoted by smaller solid circles.

Graphical Representation

- When we apply a graphical model to a problem in machine learning, we will set some of the variables to specific observed values (e.g. condition on the data).



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

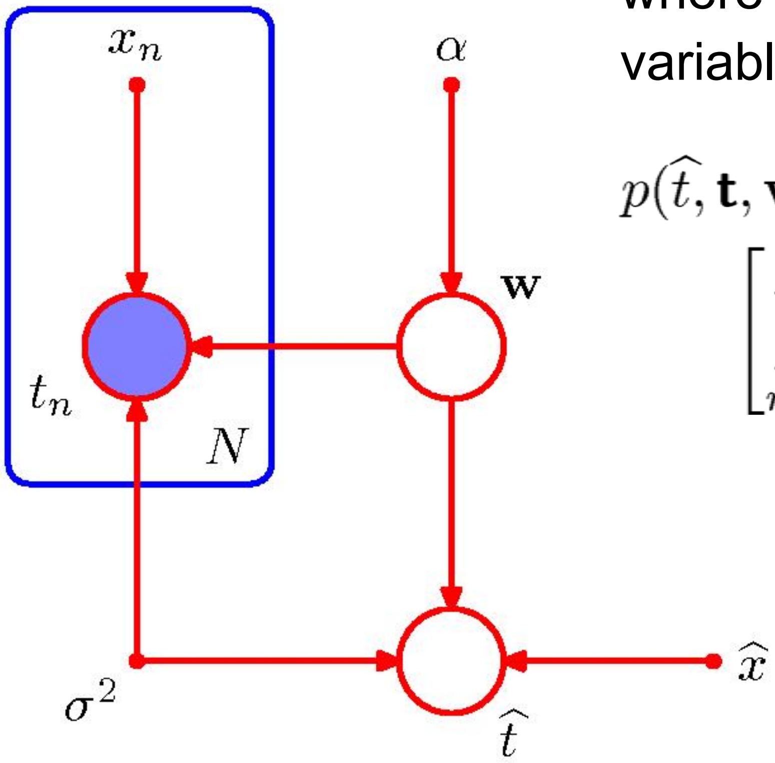
- For example, having observed the values of the targets $\{t_n\}$ on the training data, we wish to infer the posterior distribution over parameters w.

- In this example, we conditioned on observed data $\mathbf{t} = [t_1, t_2, ..., t_N]^T$ by shadowing the corresponding nodes.

Predictive Distribution

- We may also be interested in making predictions for a new input value $\hat{x}$.

$$p(\widehat{t}|\widehat{x}, \mathbf{x}, \mathbf{t}, \alpha, \sigma^2) \propto \int p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2)\,\mathrm{d}\mathbf{w}$$



where the joint distribution of all of the random variables is given by:

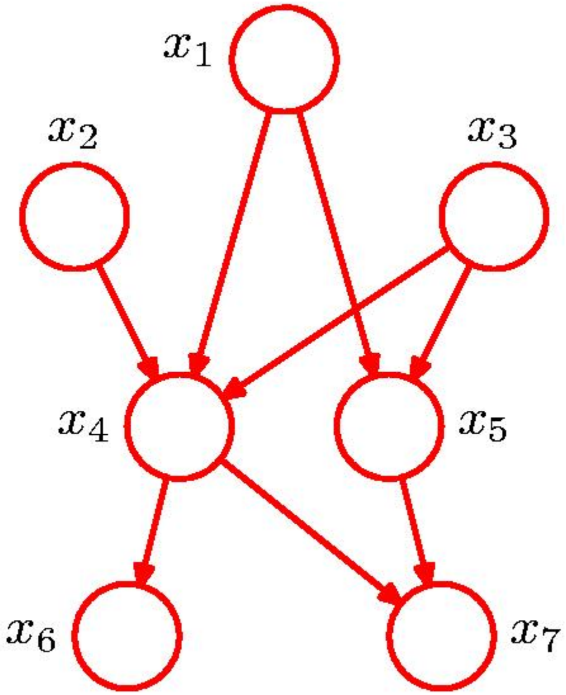
$$p(\widehat{t}, \mathbf{t}, \mathbf{w}|\widehat{x}, \mathbf{x}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\right] p(\mathbf{w}|\alpha)p(\widehat{t}|\widehat{x}, \mathbf{w}, \sigma^2)$$

- Here we are setting the random variables in t to the specific values observed in the data.

Ancestral Sampling

- Consider a joint distribution over K random variables $p(x_1, x_2, ..., x_K)$ that factorizes as:

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



- Our goal is draw a sample from this distribution.
- Start at the top and sample in order.

$$\hat{x}_1 \sim p(x_1)$$
$$\hat{x}_2 \sim p(x_2)$$
$$\hat{x}_3 \sim p(x_3)$$
$$\hat{x}_4 \sim p(x_4|\hat{x}_1, \hat{x}_2, \hat{x}_3)$$
$$\hat{x}_5 \sim p(x_5|\hat{x}_1, \hat{x}_3)$$

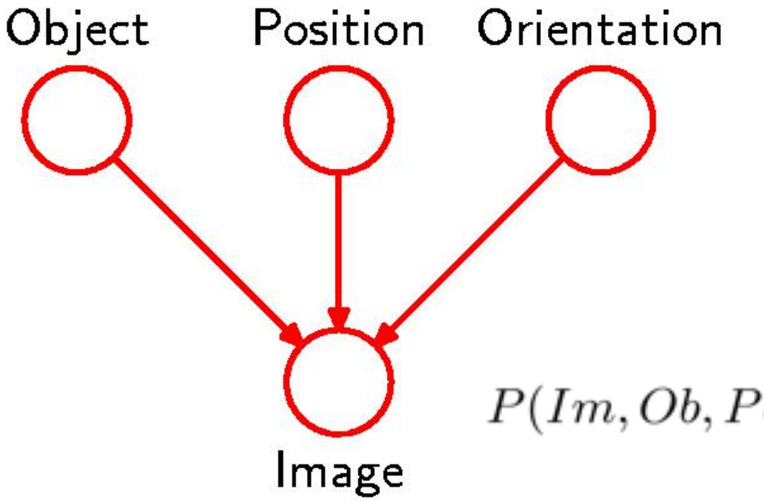
The parent variables are set to their sampled values

- To obtain a sample from the marginal distribution, e.g. $p(x_2, x_5)$, we sample from the full joint distribution, retain $\hat{x}_2, \hat{x}_5,$ and discard the remaining values.

Generative Models

- Higher-level nodes will typically represent latent (hidden) random variables.
- The primary role of the latent variables is to allow a complicated distribution over observed variables to be constructed from simpler (typically exponential family) conditional distributions.

Generative Model of an Image



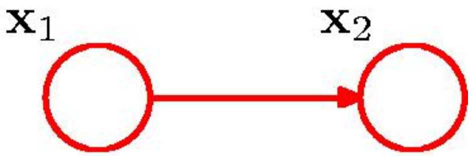
- Object identity, position, and orientation have independent prior probabilities.
- The image has a probability distribution that depends on the object identity, position, and orientation (likelihood function).

$$P(Im, Ob, Po, Or) = \underbrace{P(Im|Ob, Po, Or)}_{\text{Likelihood}}\underbrace{P(Ob)P(Po)P(Or)}_{\text{Prior}}$$

- The graphical model captures the causal process, by which the observed data was generated (hence the name generative models).

Discrete Variables

- We now examine the discrete random variables.
- Assume that we have two discrete random variables x_1 and x_2, each of which has K states.



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

- Using 1-of-K encoding, we denote the probability of observing both x_{1k}=1, x_{2l}=1 by the parameter μ_{kl}, where x_{1k} denotes the kth component of x_1 (similarly for x_2).

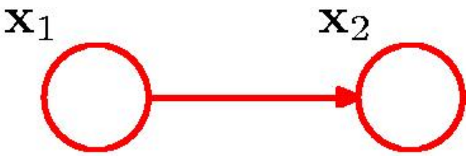
- This distribution is governed by K^2 - 1 parameters.

- The total number of parameters that must be specified for an arbitrary joint distribution over M random variables is K^M-1 (corresponds to a fully connected graph).

- Grows exponentially in the number of variables M!

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.

$\mathbf{x}_1$
$\mathbf{x}_2$

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

- We dropped the link between the nodes, so each variables is described by a separate multinomial distribution.

Discrete Variables

- In general:

  - Fully connected graphs have completely general distributions and have exponential K^M-1 number of parameters (too complex).

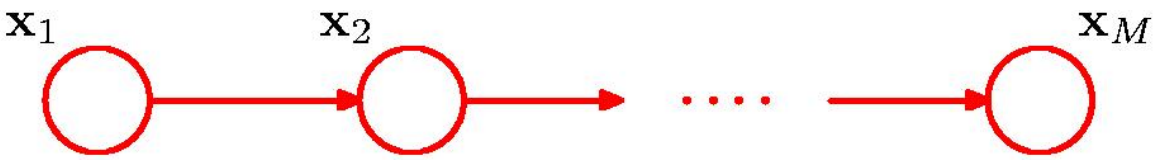
  - If there are no links, the joint distribution fully factorizes into the product of the marginals, and has M(K-1) parameters (too simple).

  - Graphs that have an intermediate level of connectivity allow for more general distributions compared to the fully factorized one, while requiring fewer parameters than the general joint distribution.

- Let us look at the example of the chain graph.

Chain Graph

- Consider an M-node Markov chain:



- The marginal distribution $p(\mathbf{x}_1)$ requires K-1 parameters.

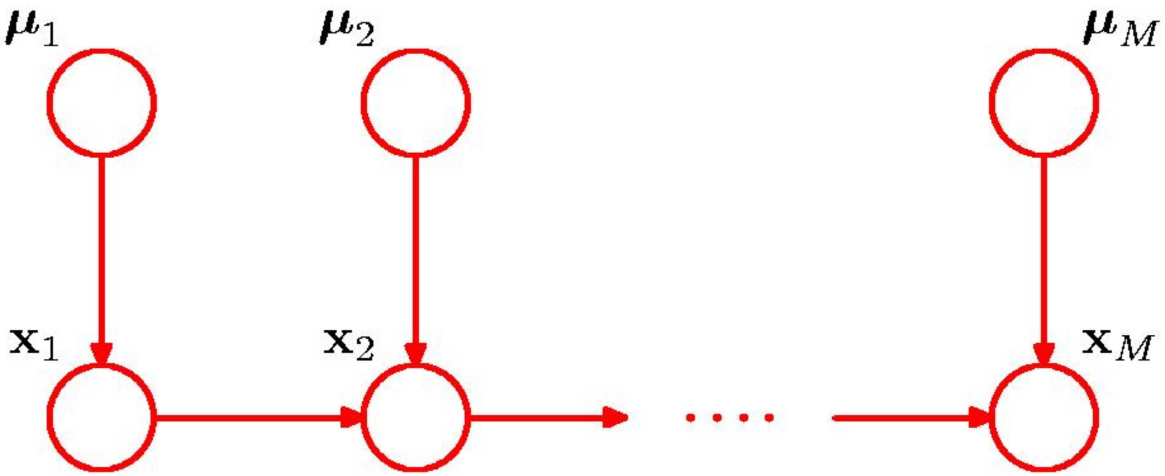
- The remaining conditional distributions $p(\mathbf{x}_i|\mathbf{x}_{i-1}), i = 2, ..., M$ require K(K-1) parameters.

- Total number of parameters: K-1 + (M-1)(K-1)K, which is quadratic in K and linear in the length M of the chain.

- This graphical model forms the basis of a simple Hidden Markov Model.

Adding Priors

- We can turn a graph over discrete random variables into a Bayesian model by introducing Dirichlet priors for the parameters.
- From a graphical model point of view, each node acquires an additional parent representing the Dirichlet distribution over parameters.

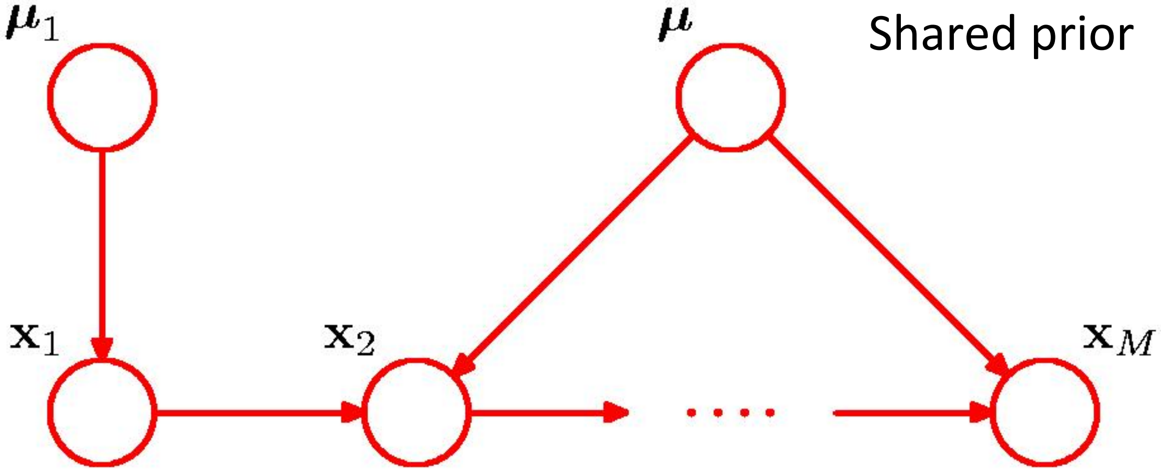


$$p\left(\{\mathbf{x}_m, \boldsymbol{\mu}_m\}\right) = p\left(\mathbf{x}_1 | \boldsymbol{\mu}_1\right) p\left(\boldsymbol{\mu}_1\right) \prod_{m=2}^{M} p\left(\mathbf{x}_m | \mathbf{x}_{m-1}, \boldsymbol{\mu}_m\right) p\left(\boldsymbol{\mu}_m\right)$$

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

Shared Prior

- We can further share the common prior over the parameters governing the conditional distributions.

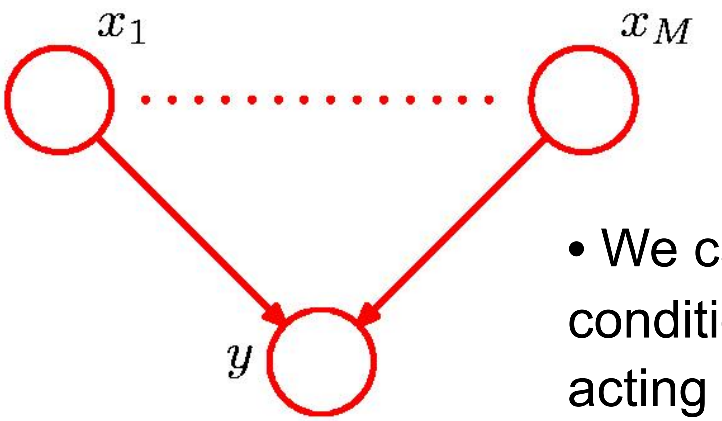


Shared prior

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

Parameterized Models

- We can use parameterized models to control exponential growth in the number of parameters.



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

- We can obtain a more parsimonious form of the conditional distribution by using a logistic function acting on a linear combination of the parent variables:

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

- This is a more restricted form of conditional distribution, but it requires only M+1 parameters (linear growth in the number of parameters).

Linear Gaussian Models

- So far we worked with joint probability distributions over a set of discrete random variables (expressed as nodes in directed acyclic graphs).
- We now show how a multivariate Gaussian distribution can be expressed as a directed graph corresponding to a linear Gaussian model.
- Consider an arbitrary acyclic graph over D random variables, in which each node represent a single continuous Gaussian distribution with its mean given by the linear function of the parents:

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

where w_{ij} and b_i are parameters governing the mean, and v_i is the variance.

Linear Gaussian Models

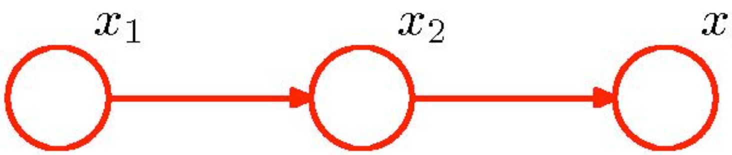
- The log of the joint distribution takes form:

$$\ln p(\mathbf{x}) = \sum_{i=1}^{D} \ln p(x_i|\mathrm{pa}_i) = -\sum_{i=1}^{D} \frac{1}{2v_i}\left(x_i - \sum_{j\in\mathrm{pa}_i} w_{ij}x_j - b_i\right)^2 + \mathrm{const},$$

where ‘const’ denotes terms independent of x.

- This is a quadratic function of x, and hence the joint distribution p(x) is a multivariate Gaussian.

- For example, consider a directed graph over three Gaussian variables with one missing link:



Computing the Mean

- We can determine the mean and covariance of the joint distribution. Remember:

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

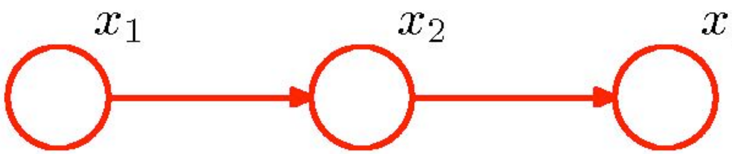
hence

$$x_i = \sum_{j\in\mathrm{pa}_i} w_{ij}x_j + b_i + \sqrt{v_i}\epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1),$$

so its expected value:

$$\mathbb{E}[x_i] = \sum_{j\in\mathrm{pa}_i} w_{ij}\mathbb{E}[x_j] + b_i.$$

- Hence we can find components: $\mathbb{E}[\mathbf{x}] = \big[\mathbb{E}[x_1], ..., \mathbb{E}[x_D]\big]$ by doing ancestral pass: start at the top and proceed in order (see example):



Computing the Covariance

- We can obtain the i,j element of the covariance matrix in the form of a recursion relation:

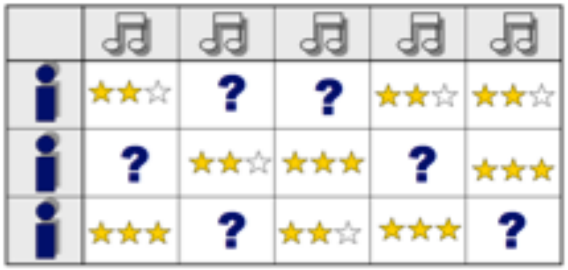
$$\begin{aligned}\mathrm{cov}[x_i, x_j] &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])\right] \\ &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i])\left(\sum_{k \in \mathrm{pa}_j} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_i}\epsilon_j\right)\right] \\ &= \sum_{k \in \mathrm{pa}_j} w_{jk}\mathrm{cov}[x_i, x_k] + I_{ij} v_j.\end{aligned}$$

- Consider two cases:

  - There are no links in the graph (graph is fully factorized), so that w_{ij}'s are zero. In this case: $\mathbb{E}[\mathbf{x}] = \left[b_1, ..., b_D\right]^T$, and the covariance is diagonal $\mathrm{diag}(v_1, ..., v_D)$. The joint distribution represents D independent univariate Gaussian distributions.

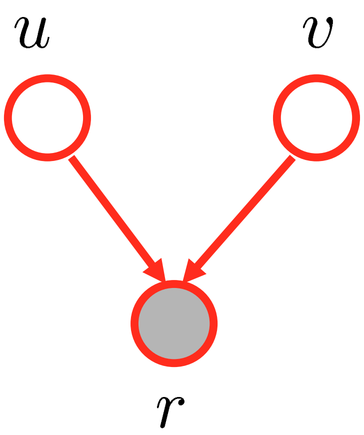
  - The graph is fully connected. The total number of parameters is D + D(D-1)/2. The covariance corresponds to a general symmetric covariance matrix.

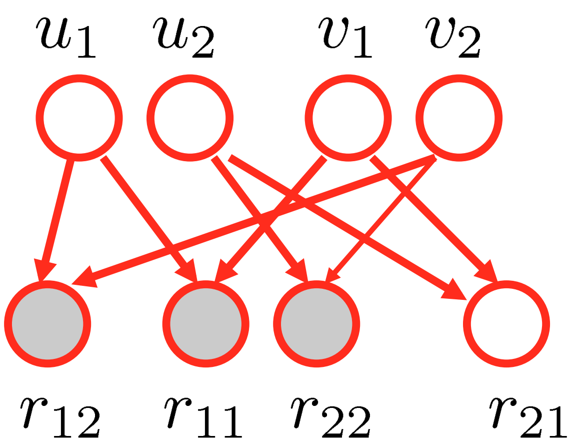
Bilinear Gaussian Model

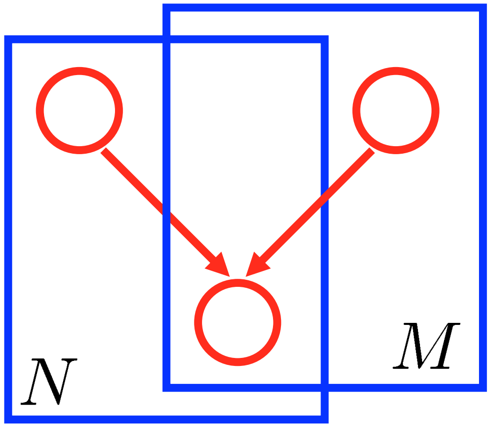


	♫	♫	♫	♫	♫
ℹ	★★☆	?	?	★★☆	★★☆
ℹ	?	★★☆	★★★	?	★★★
ℹ	★★★	?	★★☆	★★★	?

- Consider the following model:







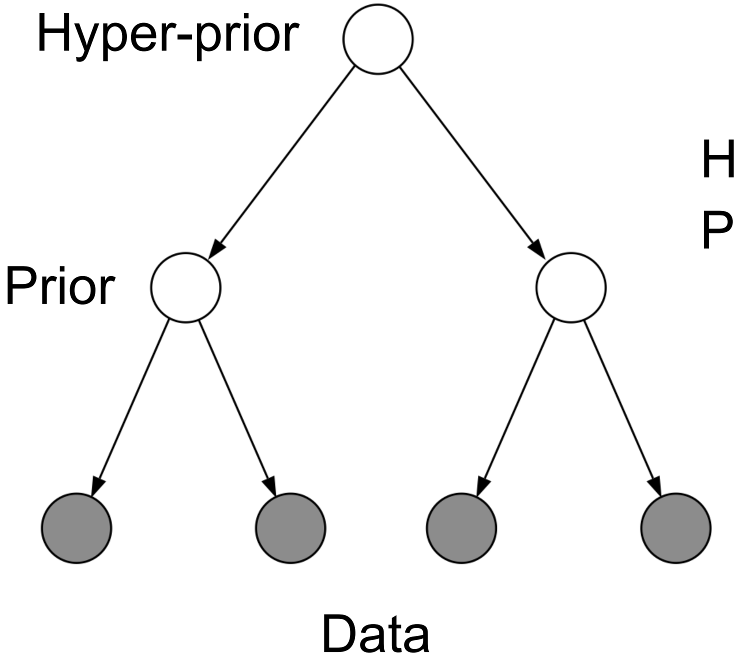
$$u \sim \mathcal{N}(0,1),$$
$$v \sim \mathcal{N}(0,1),$$
$$r \sim \mathcal{N}(uv,1).$$

Gaussian terms

$$u_i \sim \mathcal{N}(0,1), \;\; i = 1, ..., N$$
$$v_j \sim \mathcal{N}(0,1), \;\; j = 1, ..., M$$
$$r_{ij} \sim \mathcal{N}(u_i v_j, 1).$$

- The mean is given by the product of two Gaussians.

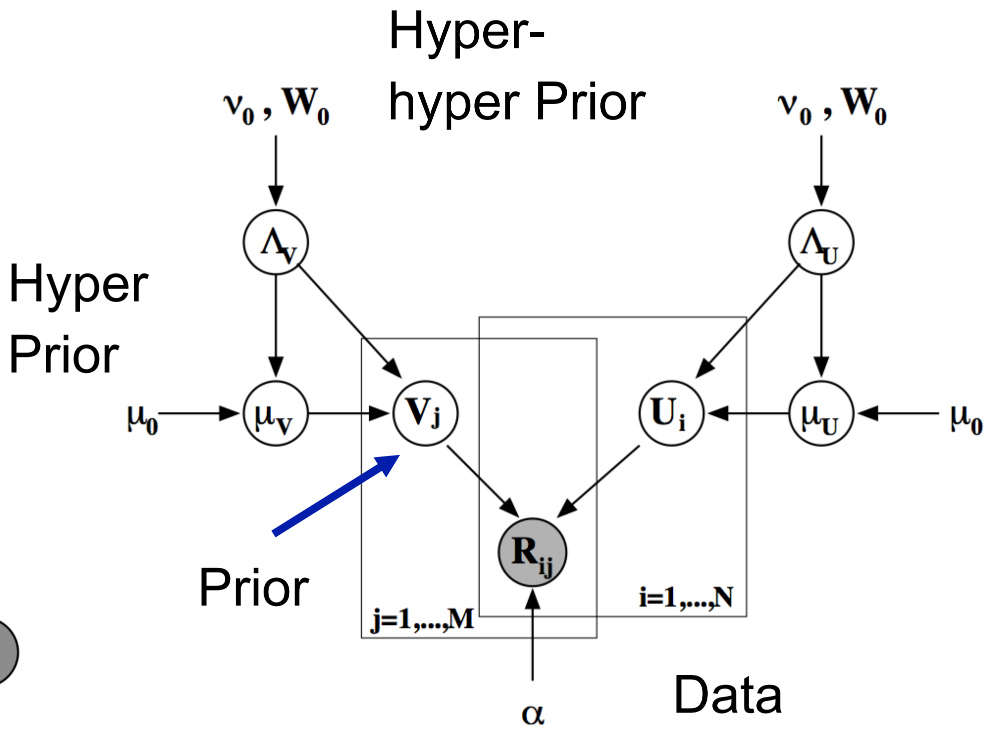
Hierarchical Models



Hyper-prior

Prior

Data



Hyper-
hyper Prior

Hyper
Prior

Prior

Data

Conditional Independence

- We now look at the concept of conditional independence.
- a is independent of b given c:

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

- Equivalently:

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

- We will use the notation:

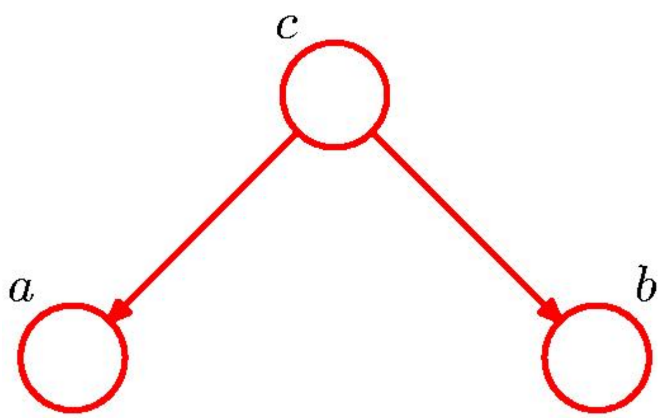
$$a \perp\!\!\!\perp b \mid c$$

- An important feature of graphical models is that conditional independence properties of the joint distribution can be read directly from the graph without performing any analytical manipulations

- The general framework for achieving this is called d-separation, where d stands for ‘directed’ (Pearl 1988).

Example 1: Tail-to-Tail Node

- The joint distribution over three variables can be written:



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

- If none of the variables are observed, we can examine whether a and b are independent:

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

- In general, this does not factorize into the product $p(a,b) = p(a)p(b)$.

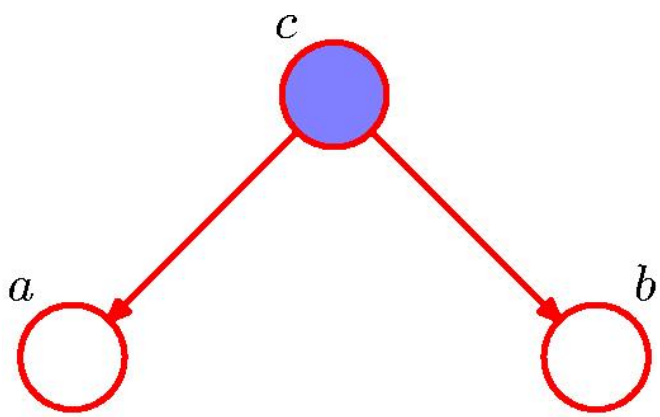
$$a \not\perp\!\!\!\perp b \mid \emptyset$$

- a and b have a common cause.

- The node c is said to be tail-to-tail node with respect to this path (the node is connected to the tails of the two arrows).

Example 1: Tail-to-Tail Node

- Suppose we condition on the variable c:



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

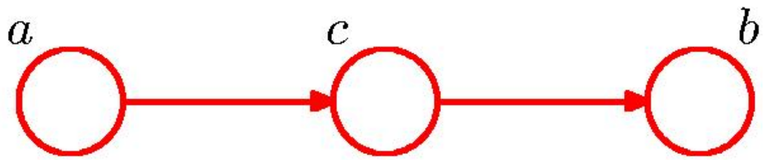
- We obtain conditional independence property:

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

- Once c has been observed, a and b can no longer have any effect on each other. They become independent.

Example 2: Head-to-Tail Node

- The joint distribution over three variables can be written:



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

- If none of the variables are observed, we can examine whether a and b are independent:

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

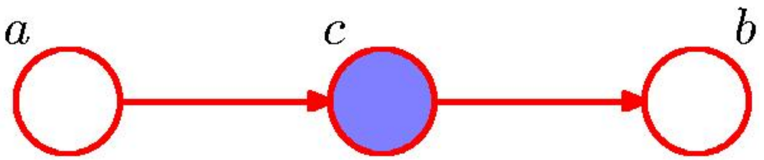
$$a \not\perp\!\!\!\perp b \mid \emptyset$$

- If c is not observed, a can influence c, and c can influence b.

- The node c is said to be head-to-tail node with respect to the path from node a to node b.

Example 2: Head-to-Tail Node

- Suppose we condition on the variable c:



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

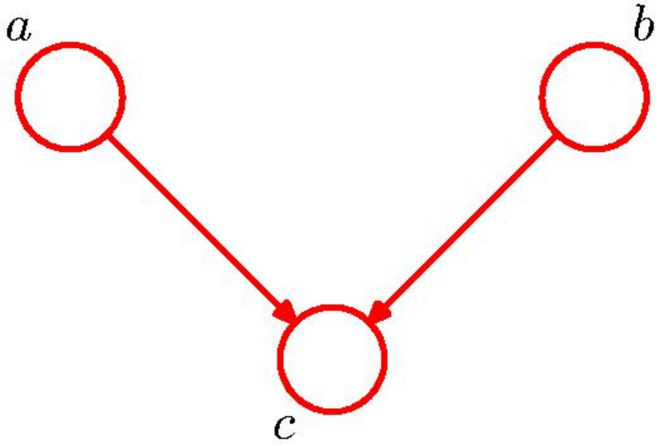
- We obtain conditional independence property:

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

- If c is observed, the value of a can no longer influence b.

Example 3: Head-to-Head Node

- The joint distribution over three variables can be written:



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

- If none of the variables are observed, we can examine whether a and b are independent:

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

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

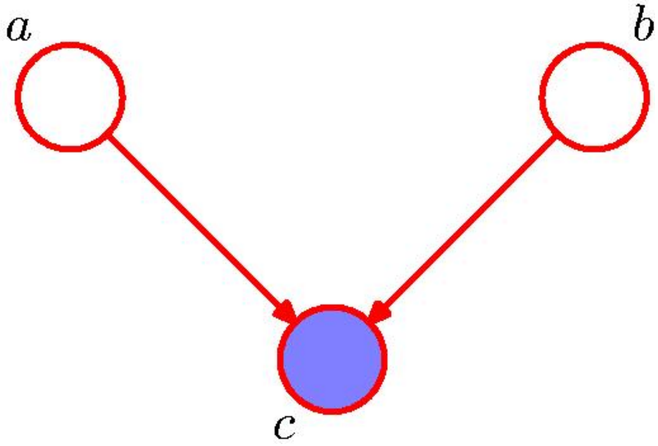
- Opposite to Example 1.

- An unobserved descendant has no effect.

- The node c is said to be head-to-head node with respect to the path from a to b (because it connects to the heads of two arrows).

Example 3: Head-to-Head Node

- Suppose we condition on the variable c:



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

- In general, this does not factorize into the product.

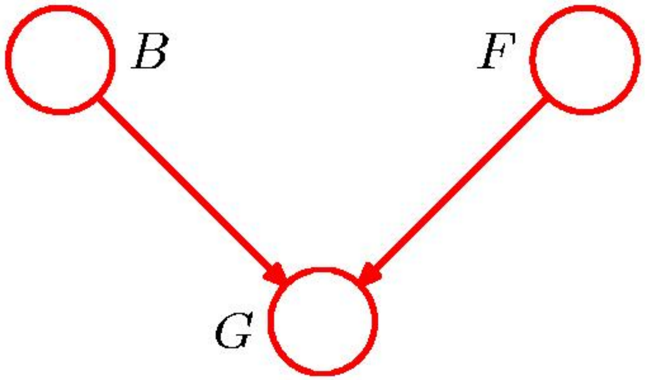
$$a \not\perp\!\!\!\perp b \mid c$$

- Opposite to Example 1.

- If the descendant (or any of its descendants) is observed, its value has implications for both a and b,

Fuel Example

- Consider the following example over three binary random variables:



$$\begin{aligned} p(B=1) &= 0.9 \\ p(F=1) &= 0.9 \end{aligned}$$

and hence

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

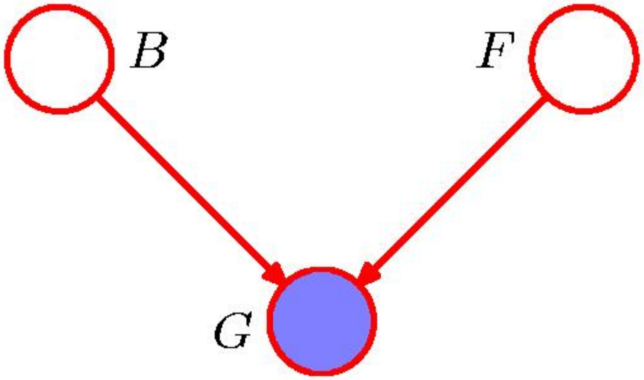
B = Battery (0=dead, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading
(0=empty, 1=full)

$$\begin{aligned} p(G=1|B=1,F=1) &= 0.8 \\ p(G=1|B=1,F=0) &= 0.2 \\ p(G=1|B=0,F=1) &= 0.2 \\ p(G=1|B=0,F=0) &= 0.1 \end{aligned}$$

Fuel Example

- Suppose that we observe that the Fuel Gauge Reading is empty G = 0.

$$\begin{aligned} p(F=0|G=0) &= \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \\ &\simeq 0.257 \end{aligned}$$



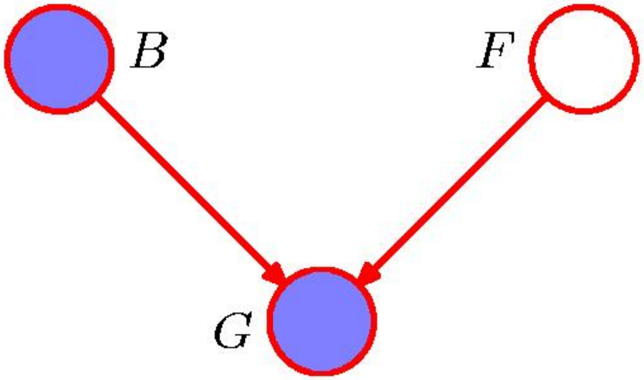
- Probability of an empty tank increased by observing G = 0.

B = Battery (0=dead, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading
(0=empty, 1=full)

Explaining Away

- If we observe that the Fuel Gauge Reading is empty G = 0 and that the battery is dead B=0.

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

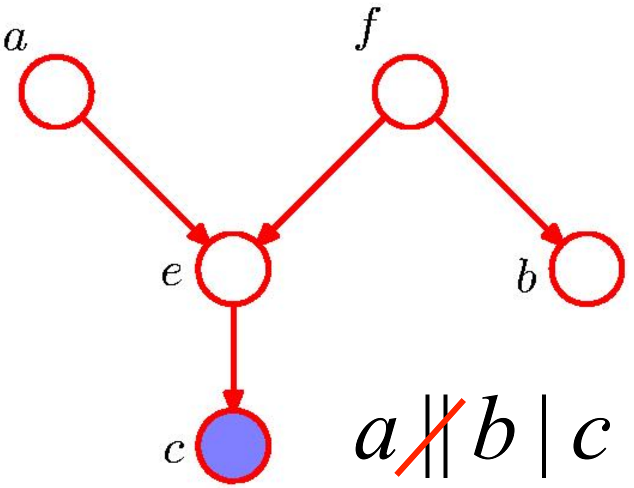


- Probability of an empty tank F=0 is reduced by observing that the battery is dead B = 0.

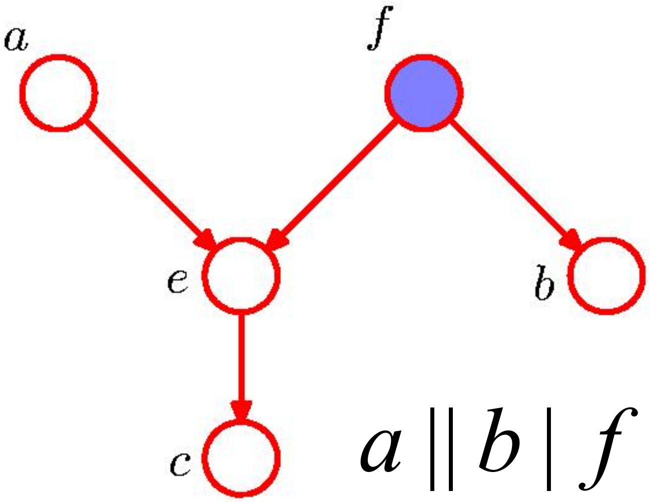
- If we observe that the fuel gauge reading is empty, you assume that one of the causes happen (either the battery is dead or the fuel tank is empty).
- One cause removes `explains away' the need for the other cause.

B = Battery (0=dead, 1=fully charged)
F = Fuel Tank (0=empty, 1=full)
G = Fuel Gauge Reading (0=empty, 1=full)

D-separation



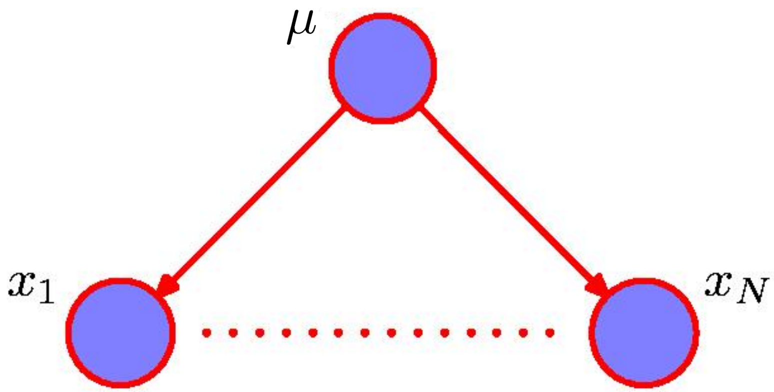
$a \cancel{\|} b \mid c$



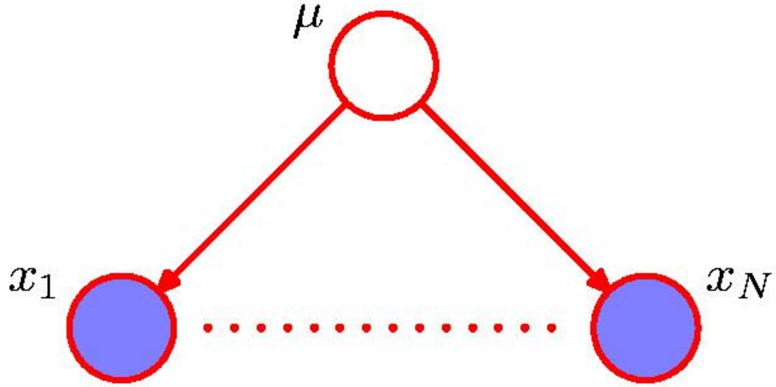
$a \| b \mid f$

- a is independent of b if and only if all paths connecting a and b are blocked.
- head-to-tail and tail-to-tail nodes are blocked when observed.
- head-to-head nodes are blocked when the node and all its descendants are unobserved.
- For example (on top), the path from a to b is not blocked by f because it is tail-to-tail node and it is unobserved.
- But conditioned on f, a and b become independent.

D-separation and i.i.d data



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



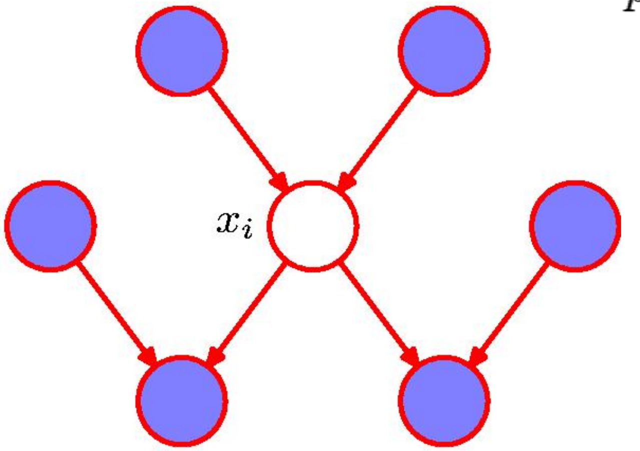
$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu)\,\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n)$$

- Another example of conditional independence and d-separation is provided by the concept of independent and identically distributed data.
- Consider the problem of finding the posterior distribution over mean μ in Bayesian linear regression model.
- Suppose that we condition on μ and consider the joint over observed variables.
- Using d-separation, note that there is unique path from x_i to any other x_j, and this path is head-to-head with respect to μ.
- If we integrate out μ, the observations are no longer independent.

Markov Blanket in Directed Models

- The Markov blanket of a node is the minimal set of nodes that must be observed to make this node independent of all other nodes

- In a directed model, the Markov blanket includes parents, children and co-parents (i.e. all the parents of the node's children) due to explaining away.

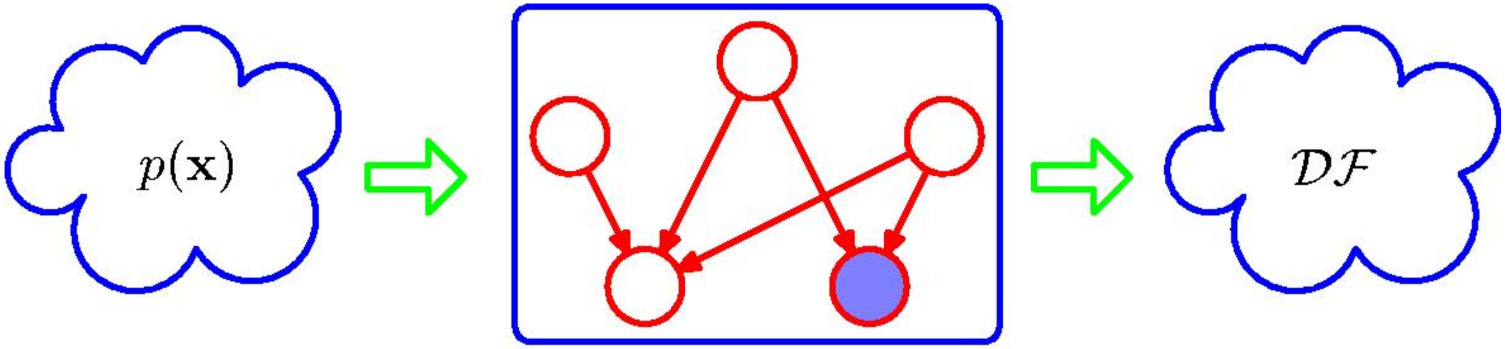


$$p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_M)}{\int p(\mathbf{x}_1,\ldots,\mathbf{x}_M)\,\mathrm{d}\mathbf{x}_i} = \frac{\prod_k p(\mathbf{x}_k|\mathrm{pa}_k)}{\int \prod_k p(\mathbf{x}_k|\mathrm{pa}_k)\,\mathrm{d}\mathbf{x}_i}$$

Factors independent of x_i cancel between numerator and denominator.

Directed Graphs as Distribution Filters

- We can view the graphical model as a filter.



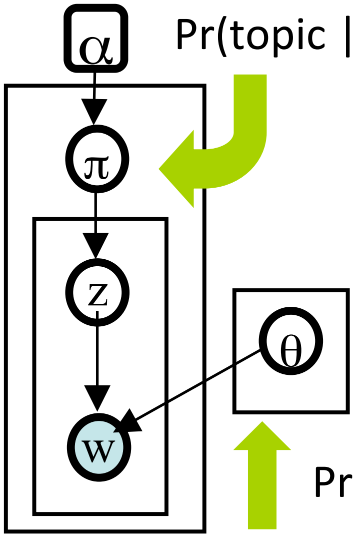
• The joint probability distribution p(x) is allowed through the filter if and only if it satisfies the factorization property.

• Note: The fully connected graph exhibits no conditional independence properties at all.

• The fully disconnected graph (no links) corresponds to a joint distribution that factorizes into the product of marginal distributions.

Popular Models

Latent Dirichlet Allocation

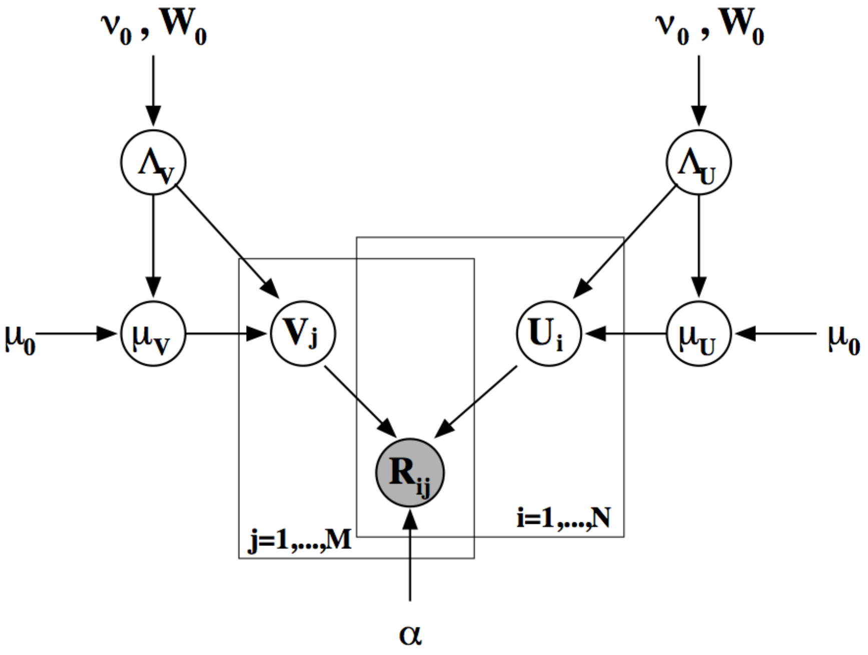


Pr(topic | doc)

Pr(word | topic)

- One of the popular models for modeling word count vectors. We will see this model later.

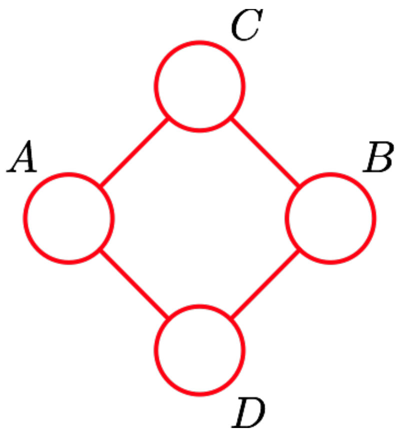
Bayesian Probabilistic Matrix Factorization



- One of the popular models for collaborative filtering applications.

Undirected Graphical Models

Directed graphs are useful for expressing causal relationships between random variables, whereas undirected graphs are useful for expressing soft constraints between random variables



- The joint distribution defined by the graph is given by the product of non-negative potential functions over the maximal cliques (connected subset of nodes).

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C) \qquad \mathcal{Z} = \sum_{\mathbf{x}} \prod_C \phi_C(x_C)$$

where the normalizing constant $\mathcal{Z}$ is called a partition function.

- For example, the joint distribution factorizes:

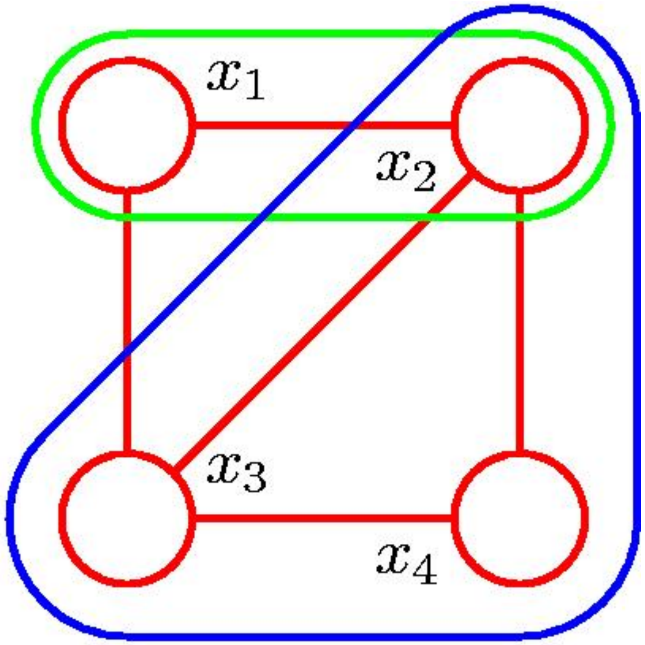
$$p(A,B,C,D) = \frac{1}{\mathcal{Z}} \phi(A,C)\phi(C,B)\phi(B,D)\phi(A,D)$$

- Let us look at the definition of cliques.

Cliques

- The subsets that are used to define the potential functions are represented by maximal cliques in the undirected graph.

- Clique: a subset of nodes such that there exists a link between all pairs of nodes in a subset.



- Maximal Clique: a clique such that it is not possible to include any other nodes in the set without it ceasing to be a clique.

- This graph has 5 cliques:

$$\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\},$$
$$\{x_4, x_2\}, \{x_1, x_3\}.$$

- Two maximal cliques:

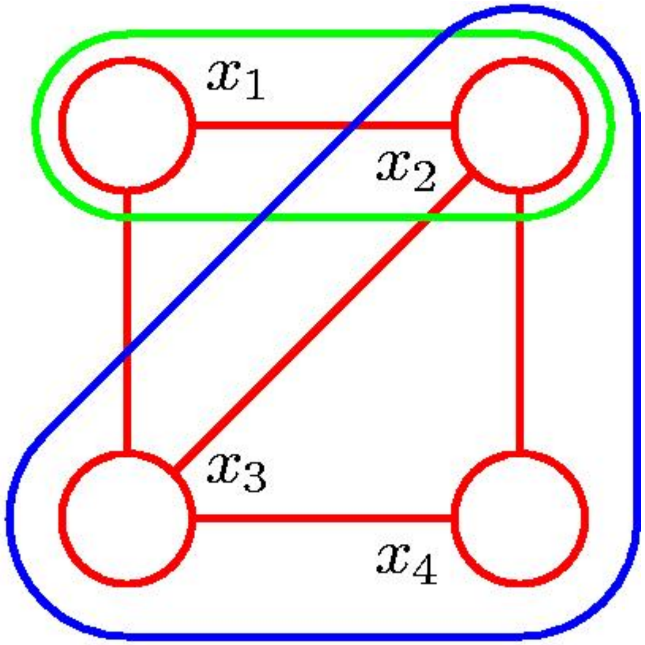
$$\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}.$$

Using Cliques to Represent Subsets

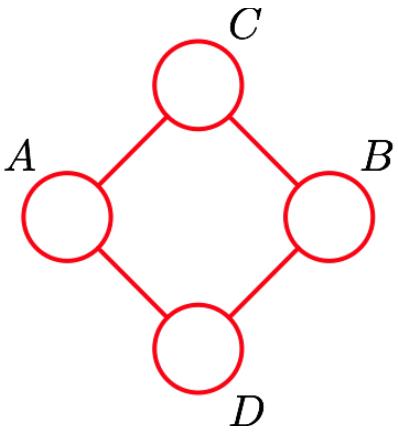
- If the potential functions only involve two nodes, an undirected graph has a nice representation.

- If the potential functions involve more than two nodes, using a different factor graph representation is much more useful.

- For now, let us consider only potential functions that are defined over two nodes.



Markov Random Fields (MRFs)



$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C)$$

- Each potential function is a mapping from the joint configurations of random variables in a clique to non-negative real numbers.
- The choice of potential functions is not restricted to having specific probabilistic interpretations.

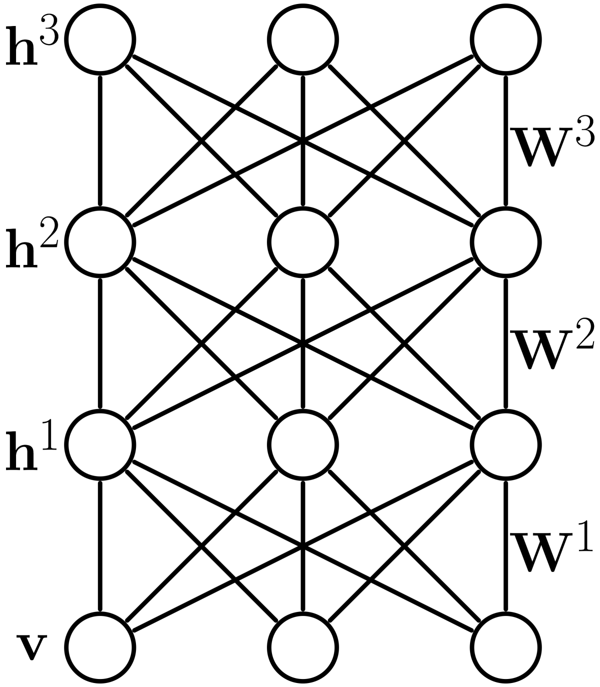
Potential functions are often represented as exponentials:

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C) = \frac{1}{\mathcal{Z}} \exp(-\sum_C E(x_c)) = \underbrace{\frac{1}{\mathcal{Z}} \exp(-E(\mathbf{x}))}_{\text{Boltzmann distribution}}$$

where E(x) is called an energy function.

MRFs with Hidden Variables

For many interesting real-world problems, we need to introduce hidden or latent variables.



- Our random variables will contain both visible and hidden variables x=(v,h).

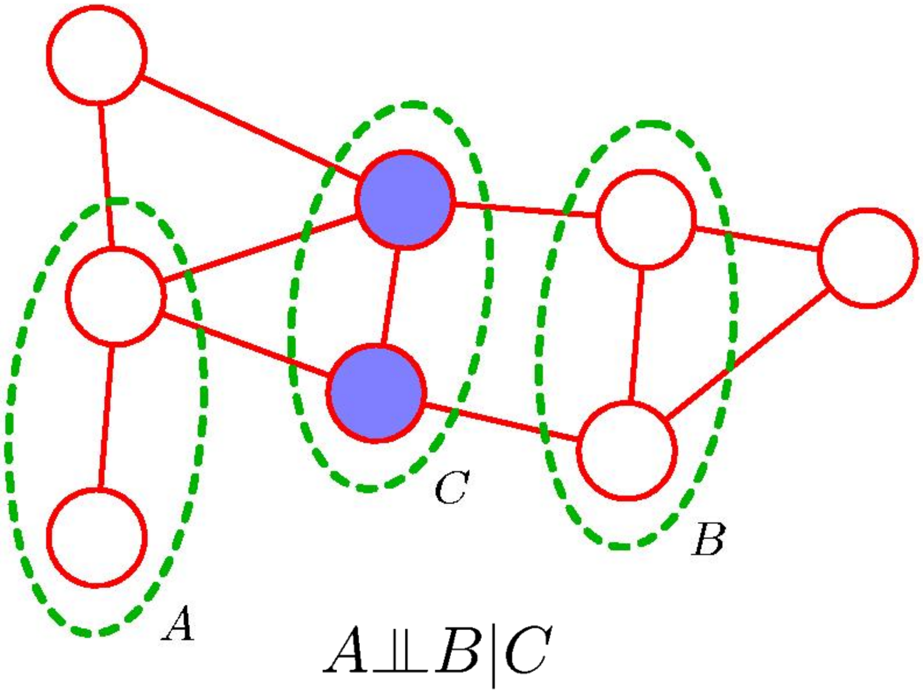
$$p(\mathbf{v}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$$

- In general, computing both partition function and summation over hidden variables will be intractable, except for special cases.

- Parameter learning becomes a very challenging task.

Conditional Independence

- Conditional Independence is easier compared to directed models:

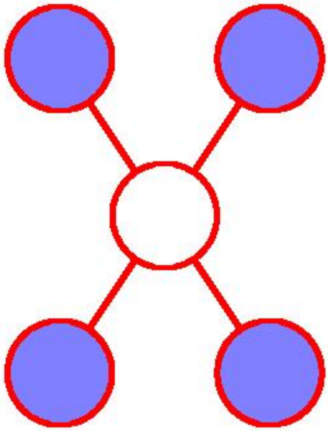


- Observation blocks a node.
- Two sets of nodes are conditionally independent if the observations block all paths between them.

Markov Blanket

- The Markov blanket of a node is simply all of the directly connected nodes.

Markov Blanket



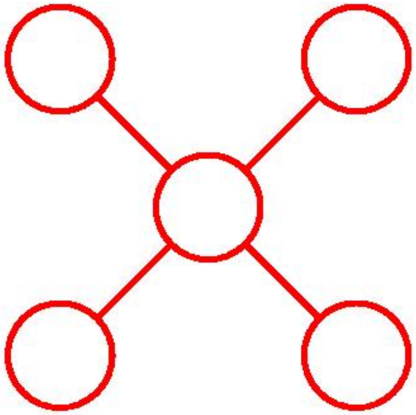
- This is simpler than in directed models, since there is no explaining away.

- The conditional distribution of x_i conditioned on all the variables in the graph is dependent only on the variables in the Markov blanket.

Conditional Independence and Factorization

- Consider two sets of distributions:
  - The set of distributions consistent with the conditional independence relationships defined by the undirected graph.
  - The set of distributions consistent with the factorization defined by potential functions on maximal cliques of the graph.

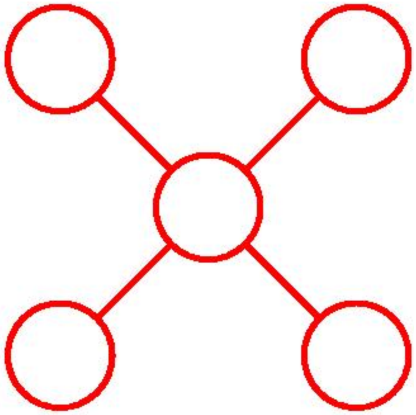
- The Hammersley-Clifford theorem states that these two sets of distributions are the same.



$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C)$$

Interpreting Potentials

- In contrast to directed graphs, the potential functions **do not have a specific probabilistic interpretation**.



$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C) = \frac{1}{\mathcal{Z}} \exp(-\sum_C E(x_c))$$

- This gives us greater flexibility in choosing the potential functions.

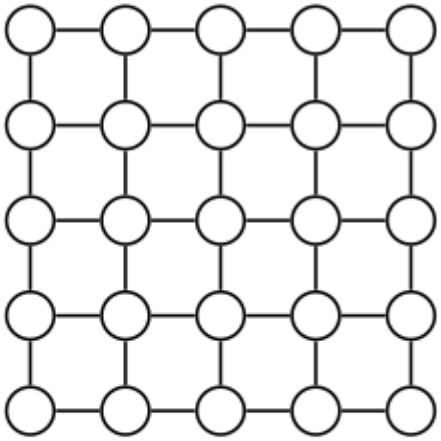
- We can view the potential function as expressing which configuration of the local variables are preferred to others.

- Global configurations with relatively high probabilities are those that find a good balance in satisfying the (possibly conflicting) influences of the clique potentials.

- So far we did not specify the nature of random variables, discrete or continuous.

Discrete MRFs

- MRFs with all discrete variables are widely used in many applications.
- MRFs with binary variables are sometimes called Ising models in statistical mechanics, and Boltzmann machines in machine learning literature.



- Denoting the binary valued variable at node j by $x_j \in \{0,1\}$, the Ising model for the joint probabilities is given by:

$$P_\theta(\mathbf{x}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big(\sum_{ij \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i\Big)$$

- The conditional distribution is given by logistic:

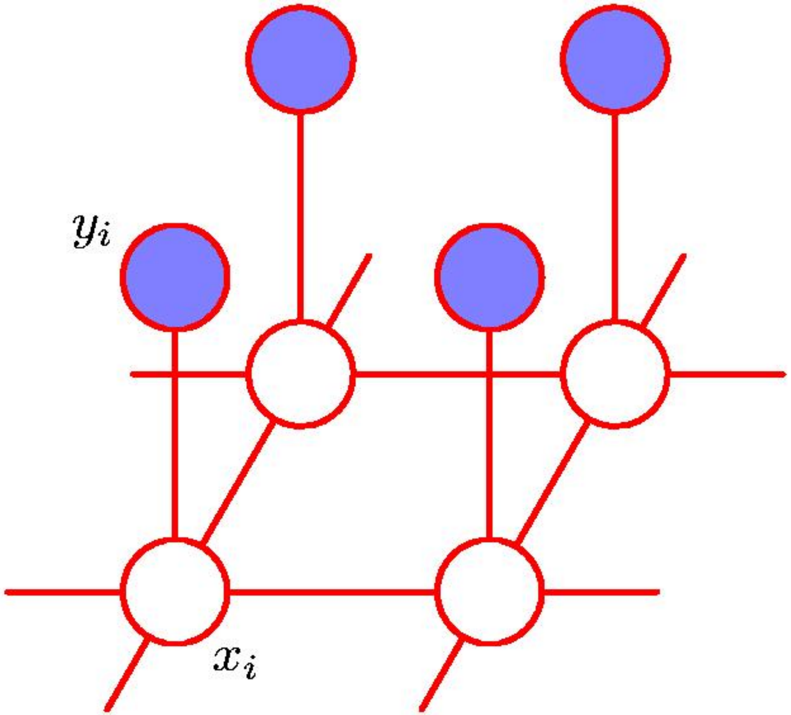
$$P_\theta(x_i = 1 | \mathbf{x}_{-i}) = \frac{1}{1 + \exp(-\theta_i - \sum_{ij \in E} x_j \theta_{ij})},$$

where x_{-i} denotes all nodes except for i.

Hence the parameter θ_{ij} measures the dependence of x_i on x_j, conditional on the other nodes.

Example: Image Denoising

- Let us look at the example of noise removal from a binary image.
- Let the observed noisy image be described by an array of binary pixel values: $y_j \in \{-1, +1\}$, i=1,…,D.



- We take a noise-free image $x_j \in \{-1, +1\}$, and randomly flip the sign of pixels with some small probability.

Bias term

Neighboring pixels are likely to have the same sign

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

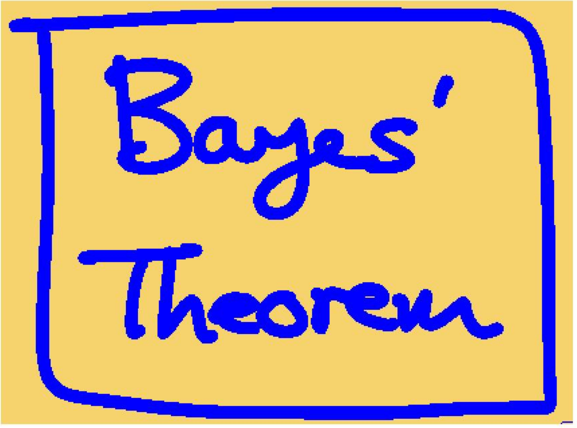
Noisy and clean pixels are likely to have the same sign

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

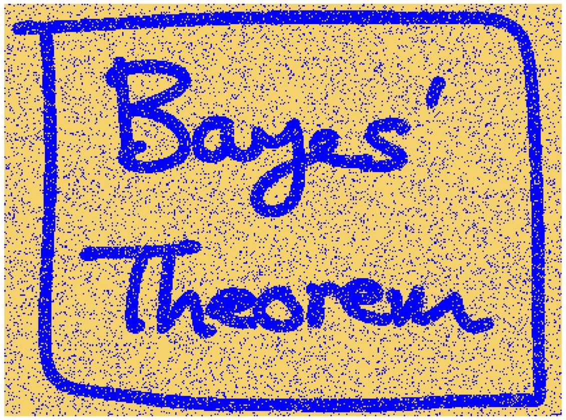
Iterated Conditional Modes

- Iterated conditional modes: coordinate-wise gradient descent.

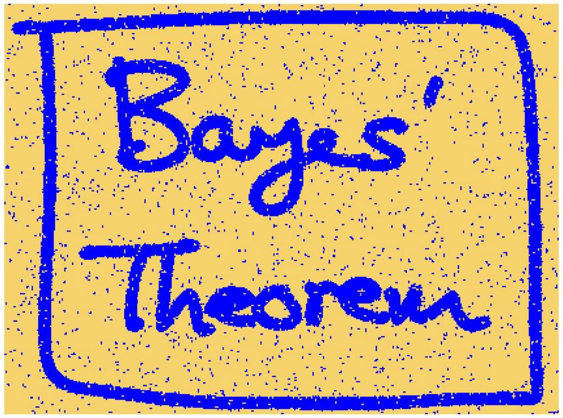
- Visit the unobserved nodes sequentially and set each x to whichever of its two values has the lowest energy.
  - This only requires us to look at the Markov blanket, i.e. the connected nodes.
  - Markov blanket of a node is simply all of the directly connected nodes.



Original Image



Noisy Image



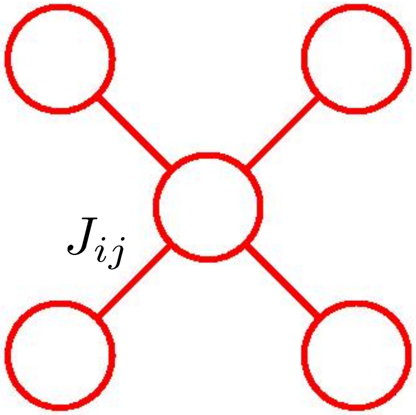
ICM

Gaussian MRFs

- We assume that the observations have a multivariate Gaussian distribution with mean μ and covariance matrix Σ.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Since the Gaussian distribution represents at most second-order relationships, it automatically encodes a pairwise MRF. We rewrite:



$$P(\mathbf{x}) = \frac{1}{\mathcal{Z}} \exp(-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \mathbf{g}^T \mathbf{x}),$$

where $$J = \Sigma^{-1}, \qquad \mu = J^{-1}\mathbf{g}.$$

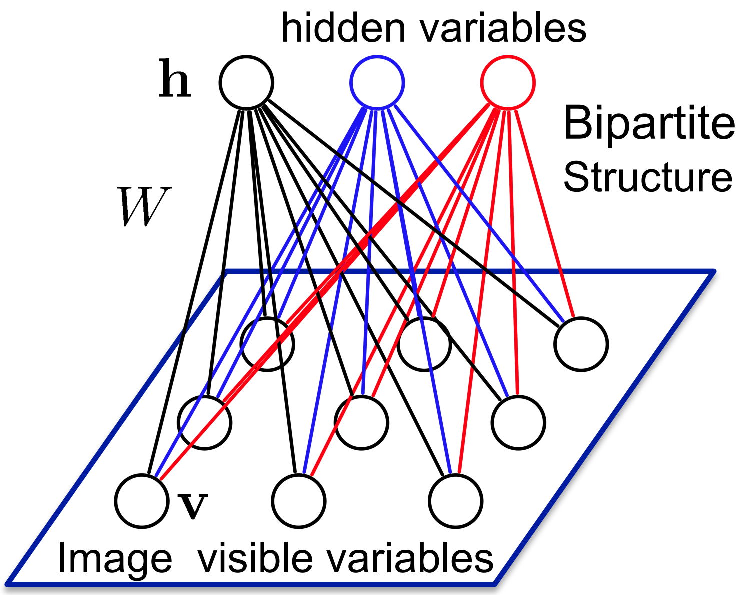
- The positive definite matrix J is known as the information matrix and is sparse with respect to the given graph: $\mathbf{x}^T J \mathbf{x} = \sum_i J_{ii} x_i^2 + 2 \sum_{ij \in E} J_{ij} x_i x_j,$

$$\text{if } (i,j) \neq E, \text{ then } J_{ij} = 0.$$

- The information matrix is sparse, but the covariance matrix is not sparse.

Restricted Boltzmann Machines

- For many real-world problems, we need to introduce hidden variables.
- Our random variables will contain visible and hidden variables x=(v,h).



Stochastic binary visible variables $\mathbf{v} \in \{0,1\}^D$ are connected to stochastic binary hidden variables $\mathbf{h} \in \{0,1\}^F$.

The energy of the joint configuration:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{ij} W_{ij} v_i h_j - \sum_i b_i v_i - \sum_j a_j h_j$$

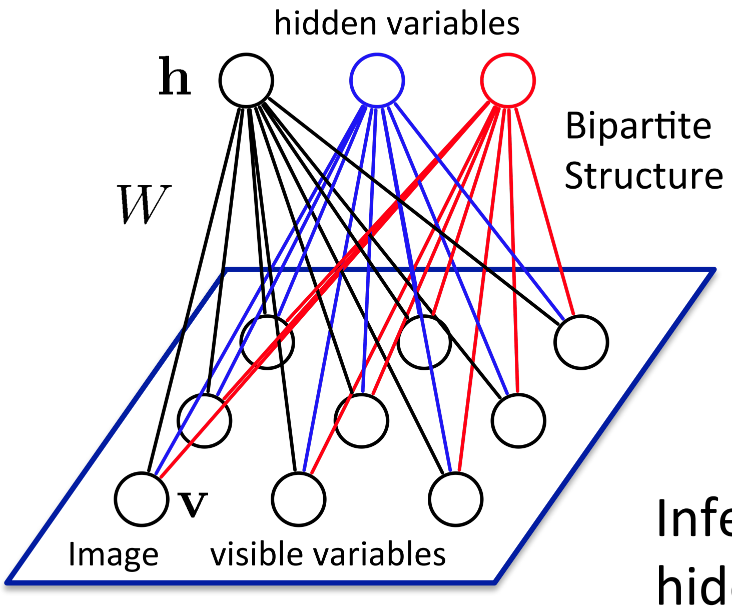
$\theta = \{W, a, b\}$ model parameters.

Probability of the joint configuration is given by the Boltzmann distribution:

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big) = \underbrace{\frac{1}{\mathcal{Z}(\theta)}}_{\text{partition function}} \prod_{ij} \underbrace{e^{W_{ij} v_i h_j}}_{\text{potential functions}} \prod_i e^{b_i v_i} \prod_j e^{a_j h_j}$$

$$\mathcal{Z}(\theta) = \sum_{\mathbf{h},\mathbf{v}} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big)$$

Restricted Boltzmann Machines



Restricted: No interaction between hidden variables

Inferring the distribution over the hidden variables is easy:

$$\underbrace{P(\mathbf{h}|\mathbf{v}) = \prod_j P(h_j|\mathbf{v})}_{\text{Factorizes: Easy to compute}} \quad P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp(-\sum_i W_{ij} v_i - a_j)}$$

Similarly:

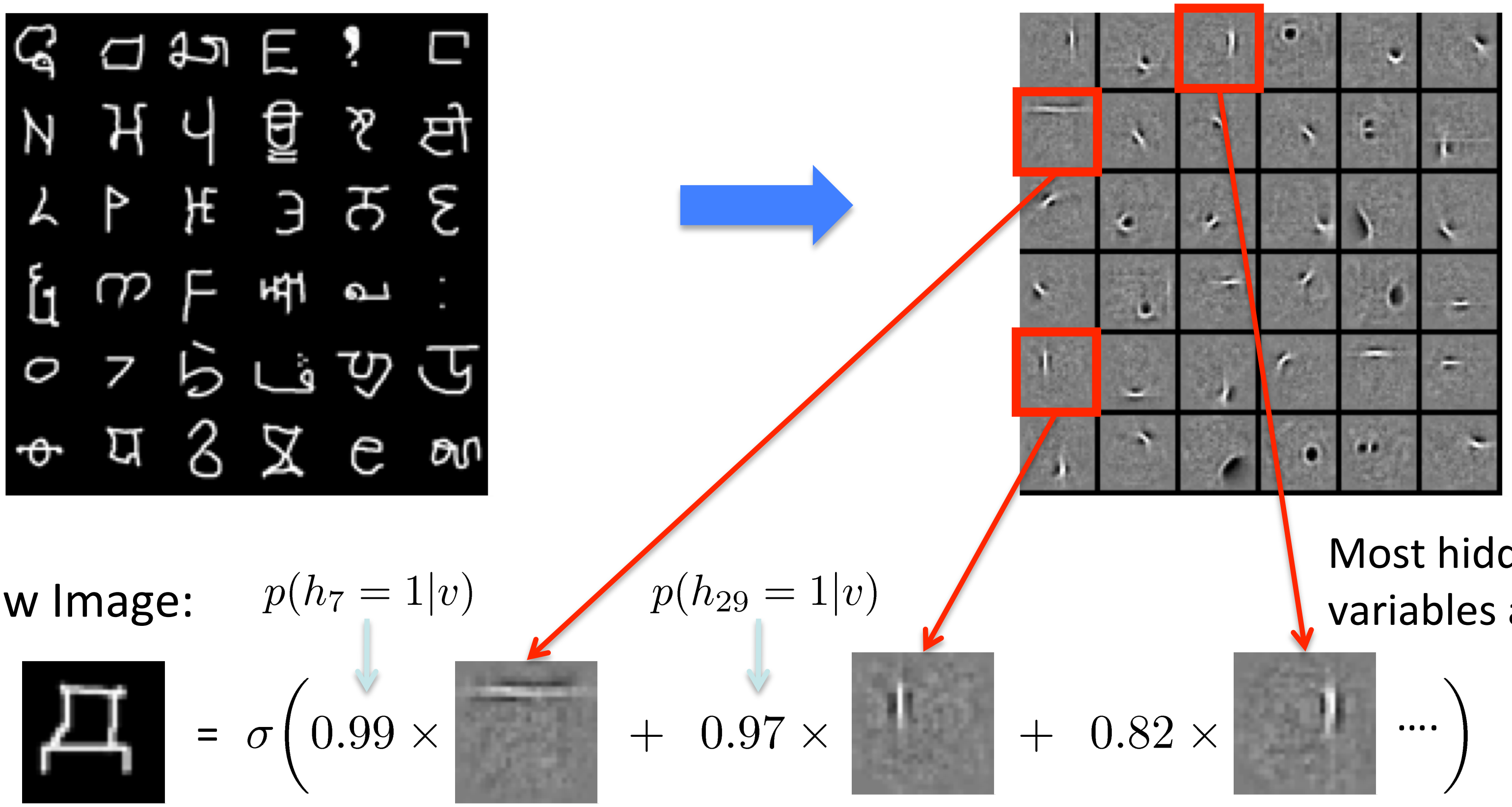
$$P(\mathbf{v}|\mathbf{h}) = \prod_i P(v_i|\mathbf{h}) \quad P(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(-\sum_j W_{ij} h_j - b_i)}$$

Markov random fields, Boltzmann machines, log-linear models.

Restricted Boltzmann Machines

Observed Data
Subset of 25,000 characters

Learned W: “edges”
Subset of 1000 features



Most hidden variables are off

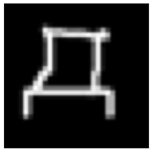
New Image:

$p(h_7 = 1|v)$ $p(h_{29} = 1|v)$

$$= \sigma\left(0.99 \times \quad + \; 0.97 \times \quad + \; 0.82 \times \quad \cdots\right)$$

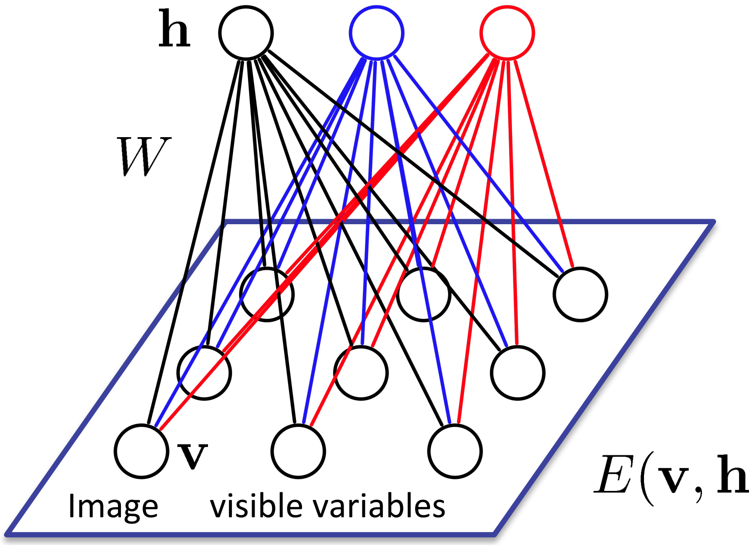
$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

Logistic Function: Suitable for modeling binary images

Represent:  as $P(\mathbf{h}|\mathbf{v}) = [0,\ 0,\ 0.82,\ 0,\ 0,\ 0.99,\ 0,\ 0\ \ldots\ ]$

Gaussian-Bernoulli RBMs

Gaussian-Bernoulli RBM:



$$P_\theta(\mathbf{v}, \mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}; \theta))$$

Define energy functions for various data modalities:

$$E(\mathbf{v}, \mathbf{h}; \theta) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{ij} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_j a_j h_j$$

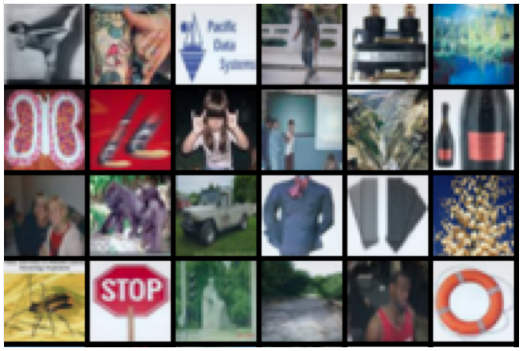
$$P(v_i = x|\mathbf{h}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - b_i - \sigma_i \sum_j W_{ij} h_j)^2}{2\sigma_i^2}\right)$$ Gaussian

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp(-\sum_i W_{ij} \frac{v_i}{\sigma_i} - a_j)}$$ Bernoulli

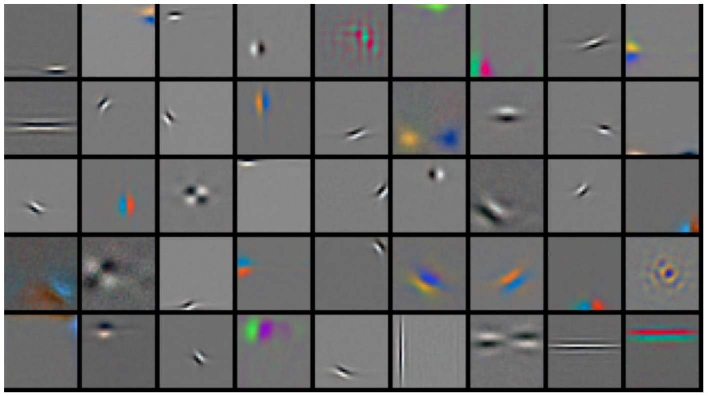
Gaussian-Bernoulli RBMs

Images: Gaussian-Bernoulli RBM

4 million unlabelled images



Learned features (out of 10,000)



Text: Multinomial-Bernoulli RBM

WIKIPEDIA
The Free Encyclopedia

REUTERS

AP Associated Press

Reuters dataset:
804,414 unlabeled
newswire stories

Bag-of-Words

Learned features: ``topics''

russian
russia
moscow
yeltsin
soviet

clinton
house
president
bill
congress

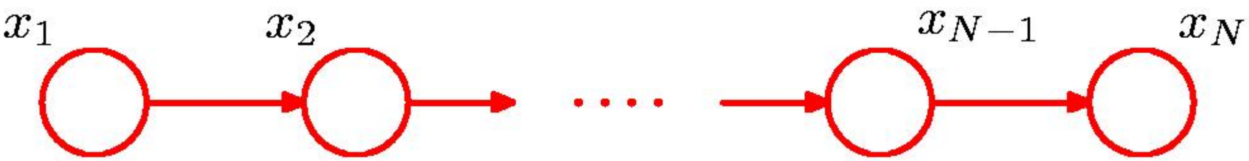
computer
system
product
software
develop

trade
country
import
world
economy

stock
wall
street
point
dow

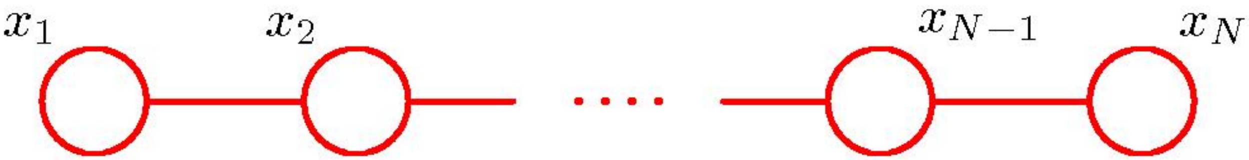
Relation to Directed Graphs

- Let us try to convert directed graph into an undirected graph:



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

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



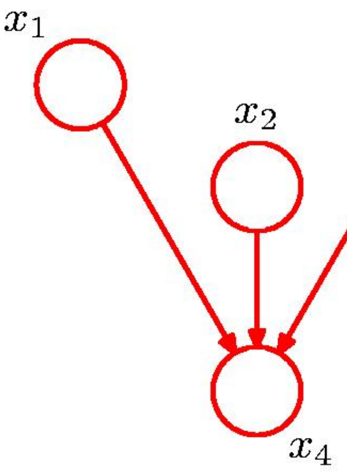
Directed vs. Undirected

- Directed Graphs can be more precise about independencies than undirected graphs.

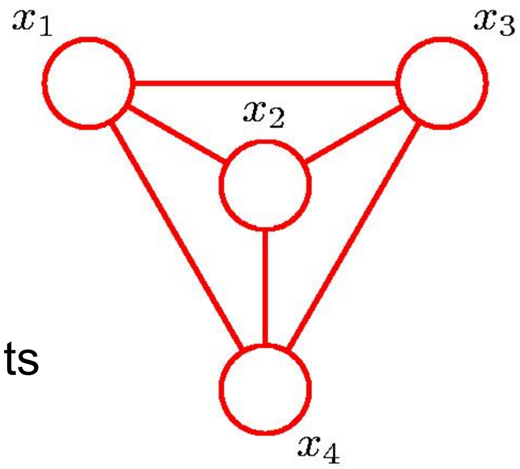
$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}}\psi(x_1, x_2, x_3, x_4)$$

need 4th order clique



Moralize: Marry the parents



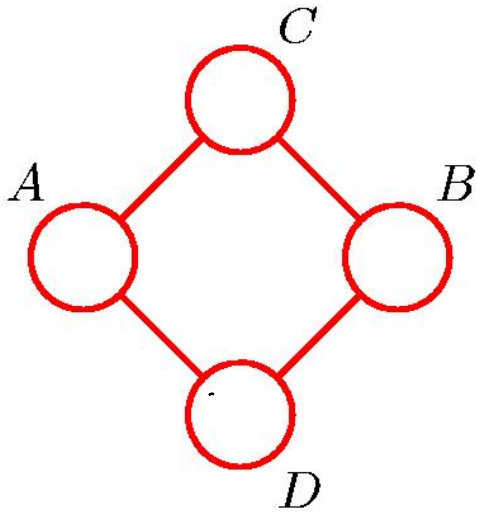
- All the parents of x_4 can interact to determine the distribution over x_4.
- The directed graph represents independencies that the undirected graph cannot model.

- To represent the high-order interaction in the directed graph, the undirected graph needs a fourth-order clique.
- This fully connected graph exhibits no conditional independence properties

Undirected vs. Directed

- Undirected Graphs can be more precise about independencies than directed graphs

  - There is no directed graph over four variables that represents the same set of conditional independence properties.



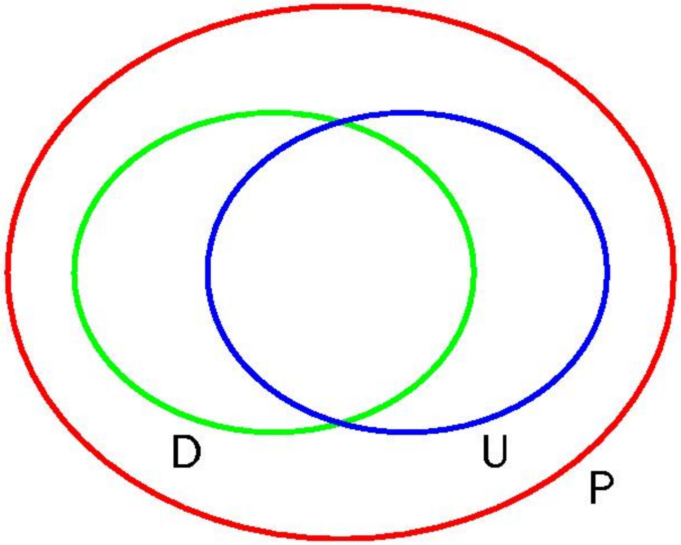
$$A \not\perp\!\!\!\perp B \mid \emptyset$$

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

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

Directed vs. Undirected

- If every conditional independence property of the distribution is reflected in the graph and vice versa, then the graph is a perfect map for that distribution.



- Venn diagram:
  - The set of all distributions P over a given set of random variables.
  - The set of distributions D that can be represented as a perfect map using directed graph.
  - The set of distributions U that can be represented as a perfect map using undirected graph.

- We can extend the framework to graphs that include both directed and undirected graphs.