

10417/10617
Intermediate Deep Learning:
Fall2023

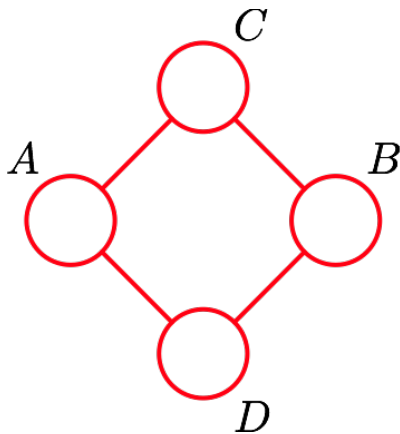
Russ Salakhutdinov

Machine Learning Department
rsalakhu@cs.cmu.edu

Graphical Models II

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).

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

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.

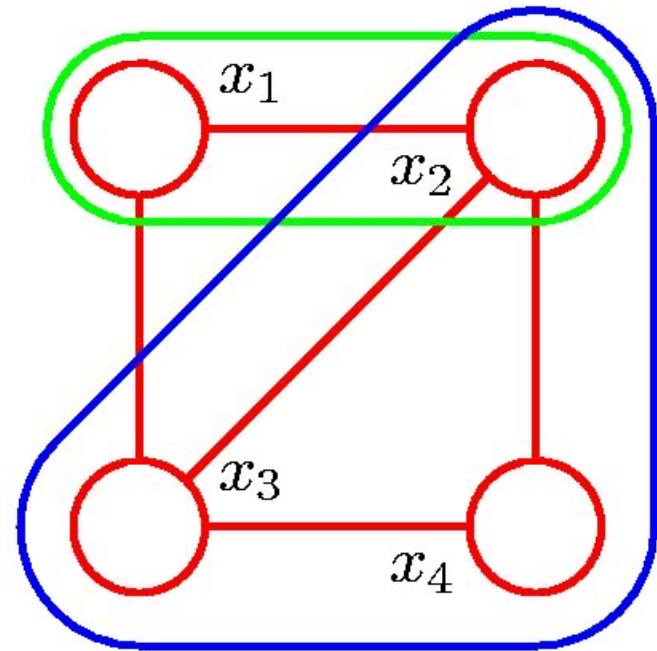
- This graph has 5 cliques:

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

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

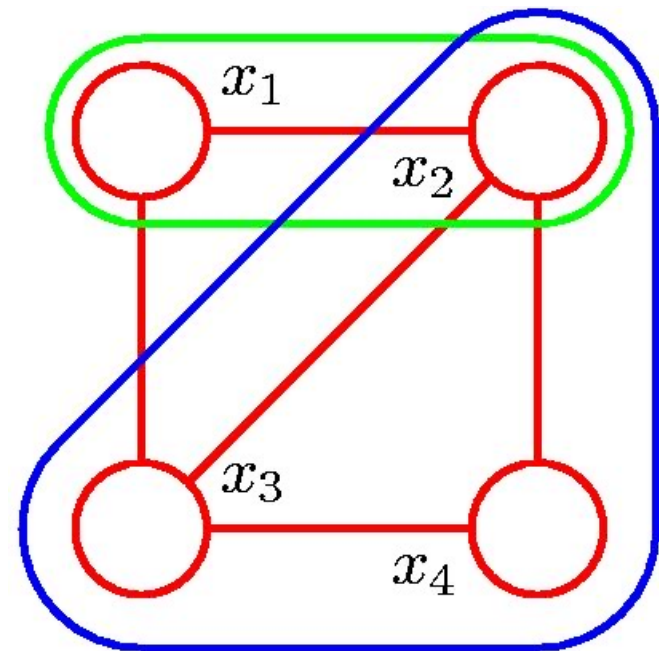
- Two maximal cliques:

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

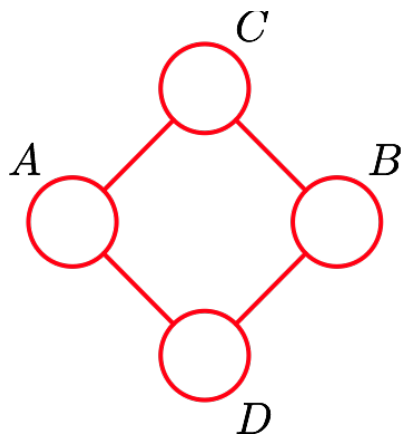


Using Cliques to Represent Subsets

- 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:

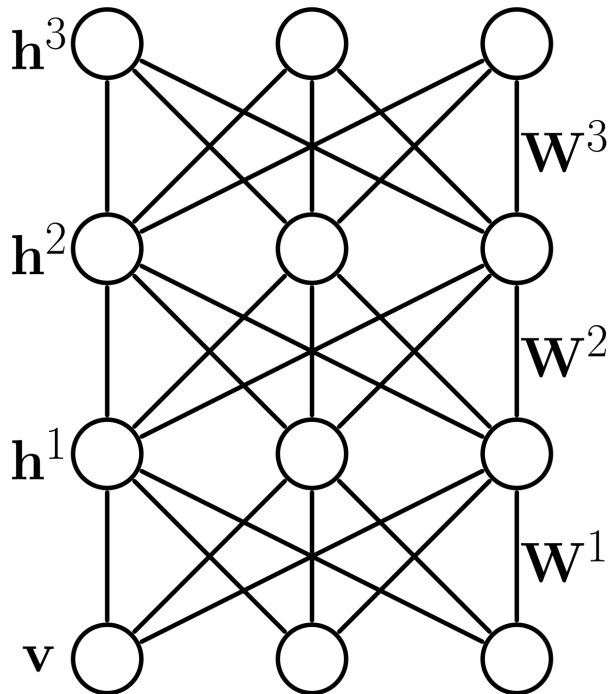
$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C) = \frac{1}{\mathcal{Z}} \exp\left(-\sum_C E(x_c)\right) = \frac{1}{\mathcal{Z}} \underbrace{\exp(-E(\mathbf{x}))}_{\text{Boltzmann distribution}}$$

where $E(\mathbf{x})$ is called an energy function.

Boltzmann distribution

MRFs with Hidden Variables

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



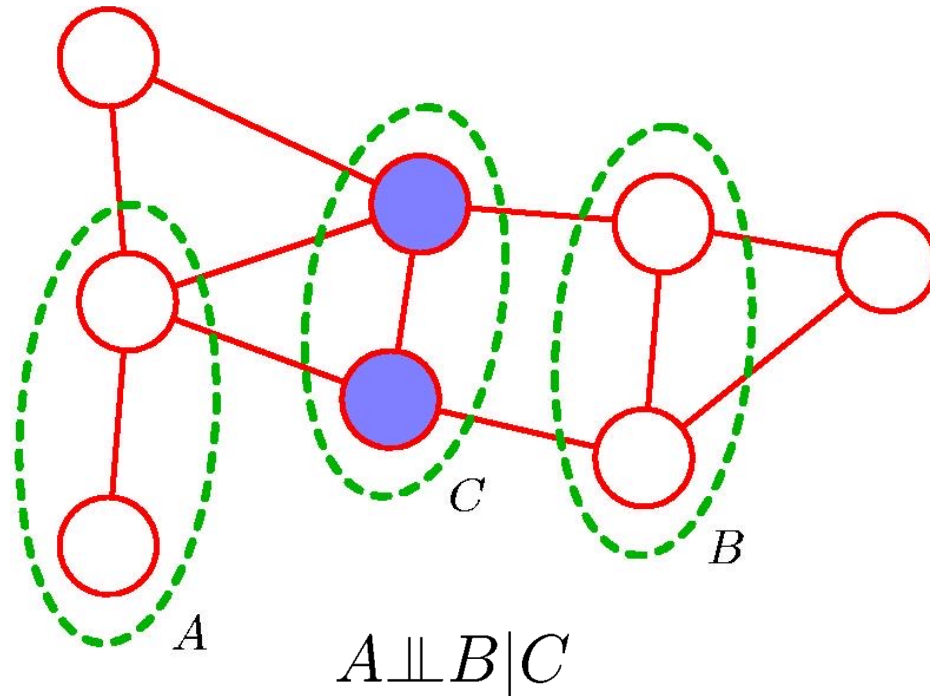
- Our random variables will contain both **visible and hidden** variables $x=(v,h)$.

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

- 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:

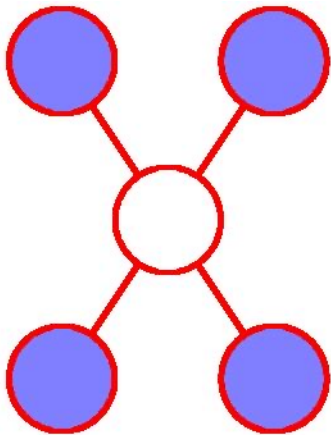


- Observation blocks a node.
- Two sets of nodes are conditionally independent if the observations block all paths between them.

Markov Blanket

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

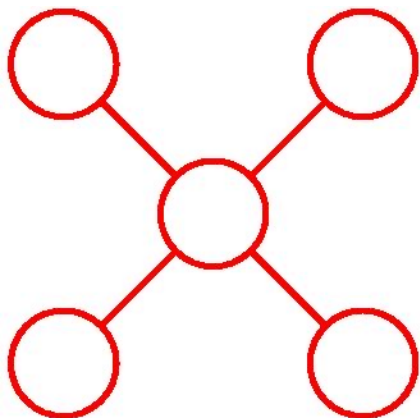
Markov Blanket



- 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

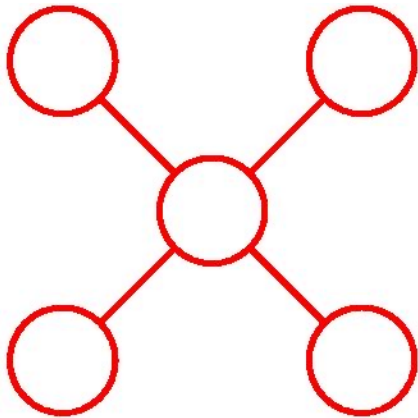
- Consider two sets of distributions:
 - The set of distributions consistent with the conditional independence relationships defined by the undirected graph.
 - The set of distributions consistent with the factorization defined by potential functions on maximal cliques of the graph.
- The **Hammersley-Clifford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



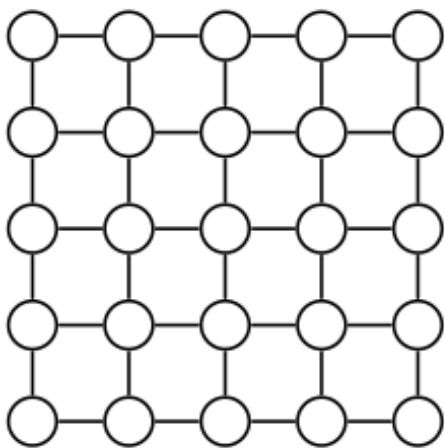
$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C) = \frac{1}{Z} \exp\left(-\sum_C E(x_C)\right)$$

- 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



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

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

- The conditional distribution is given by logistic:

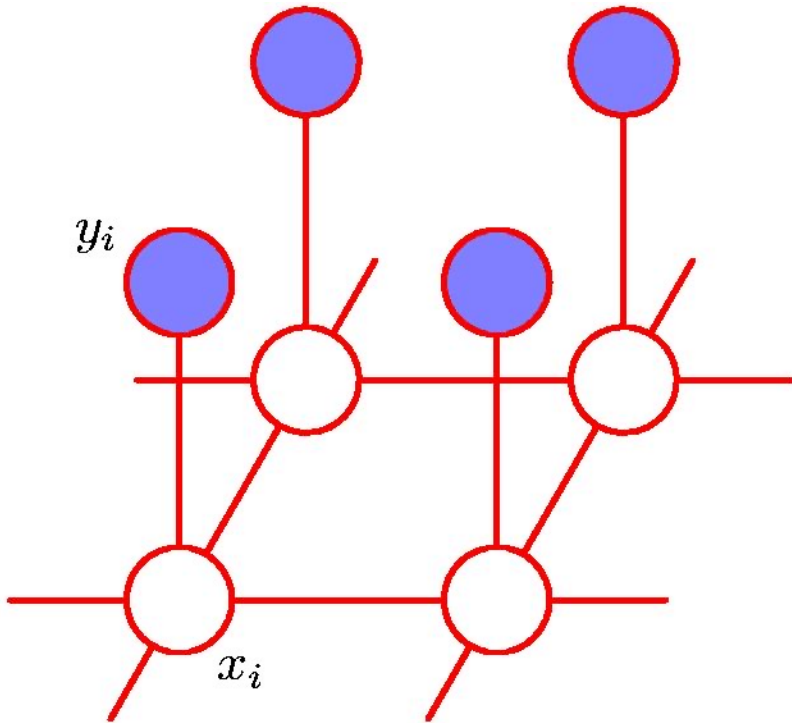
$$P_{\theta}(x_i = 1 | \mathbf{x}_{-i}) = \frac{1}{1 + \exp(-\theta_i - \sum_{ij \in E} x_j \theta_{ij})},$$
 where \mathbf{x}_{-i} denotes all nodes except for i .

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: $y_j \in \{-1, +1\}$, $i=1, \dots, D$.

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



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

Bias term

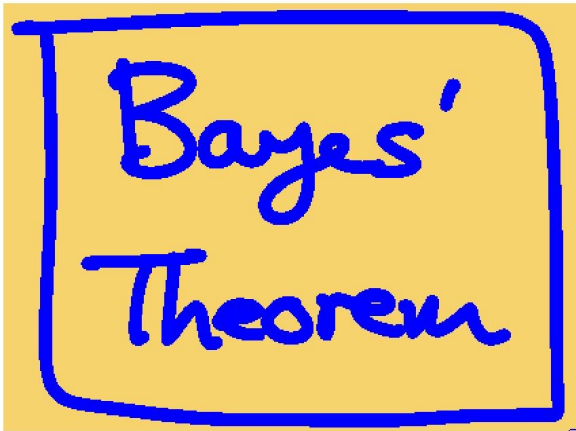
Neighboring pixels are likely to have the same sign

Noisy and clean pixels are likely to have the same sign

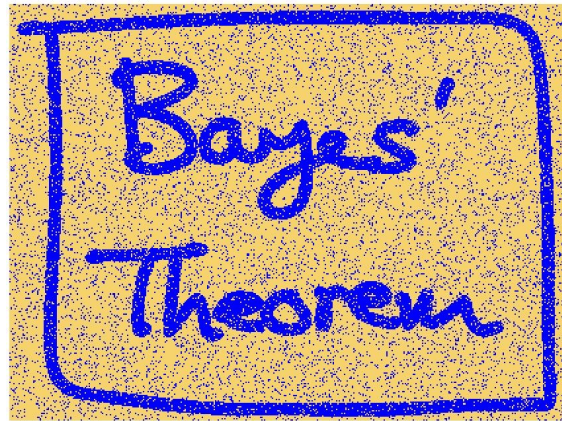
$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

Iterated Conditional Modes

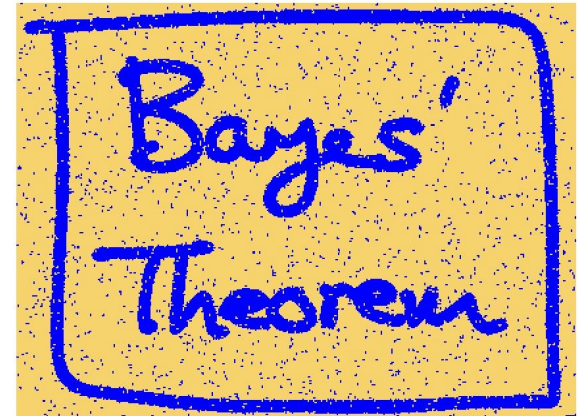
- **Iterated conditional modes:** coordinate-wise gradient descent.
- Visit the unobserved nodes sequentially and set each x to whichever of its two values has the lowest energy.
 - This only requires us to look at the Markov blanket, i.e. the connected nodes.
 - Markov blanket of a node is simply all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



Original Image



Noisy Image



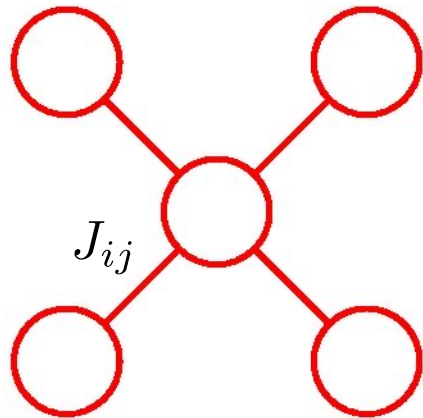
ICM 13

Gaussian MRFs

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

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

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



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

where

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

- The positive definite matrix J is known as the information matrix and is sparse with respect to the given graph:

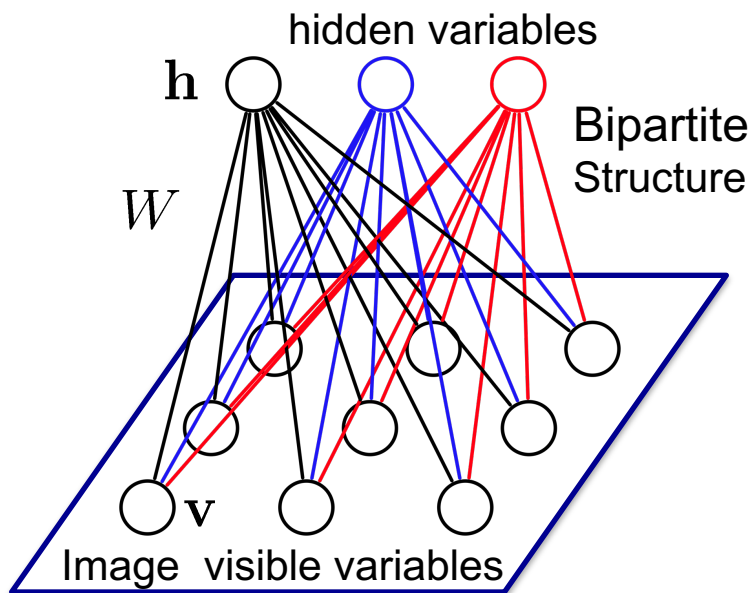
$$\mathbf{x}^T J \mathbf{x} = \sum_i J_{ii} x_i^2 + 2 \sum_{ij \in E} J_{ij} x_i x_j,$$

if $(i, j) \notin E$, 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 $\mathbf{x}=(\mathbf{v},\mathbf{h})$.



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.

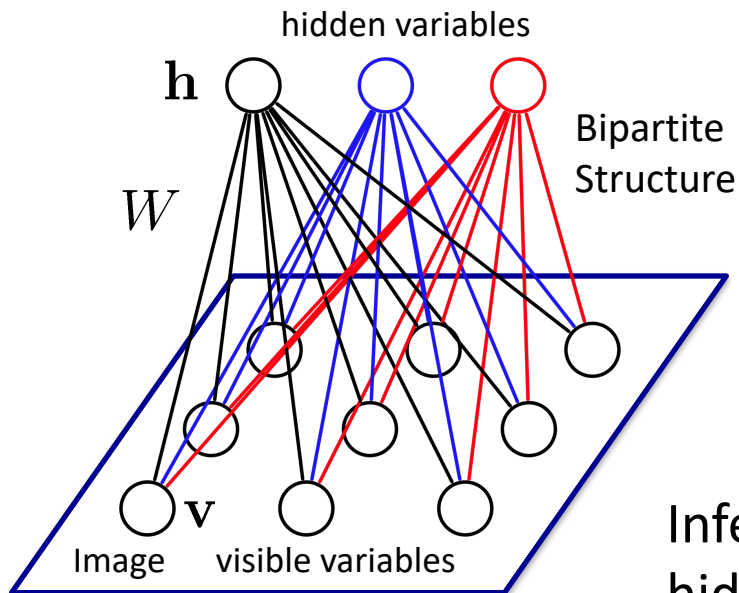
Probability of the joint configuration is given by the Boltzmann distribution:

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

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

partition function potential functions

Restricted Boltzmann Machines



Restricted: No interaction between hidden variables



Inferring the distribution over the hidden variables is easy:

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

Similarly: Factorizes: Easy to compute

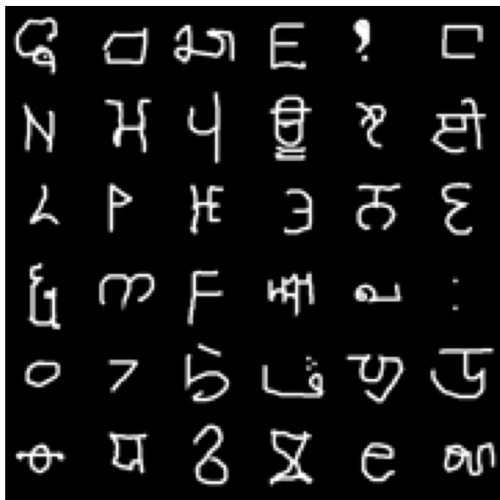
$$P(\mathbf{v}|\mathbf{h}) = \prod_i P(v_i|\mathbf{h}) \quad P(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(-\sum_j W_{ij}h_j - b_i)}$$

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

Restricted Boltzmann Machines

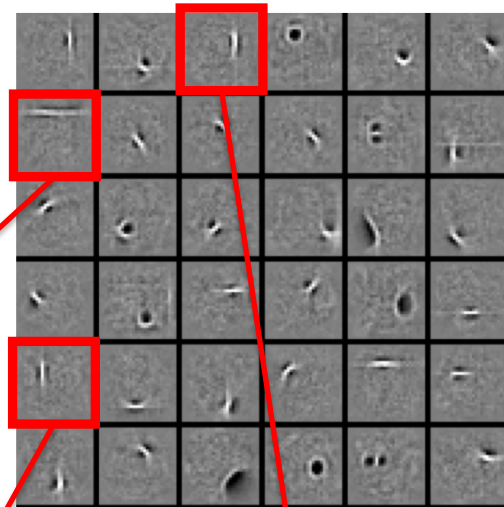
Observed Data

Subset of 25,000 characters



Learned W: "edges"

Subset of 1000 features



New Image:

$$p(h_7 = 1|v)$$

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

Most hidden variables are off



$$= \sigma \left(0.99 \times \text{[edge image]} + 0.97 \times \text{[edge image]} + 0.82 \times \text{[edge image]} + \dots \right)$$

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

Logistic Function: Suitable for modeling binary images

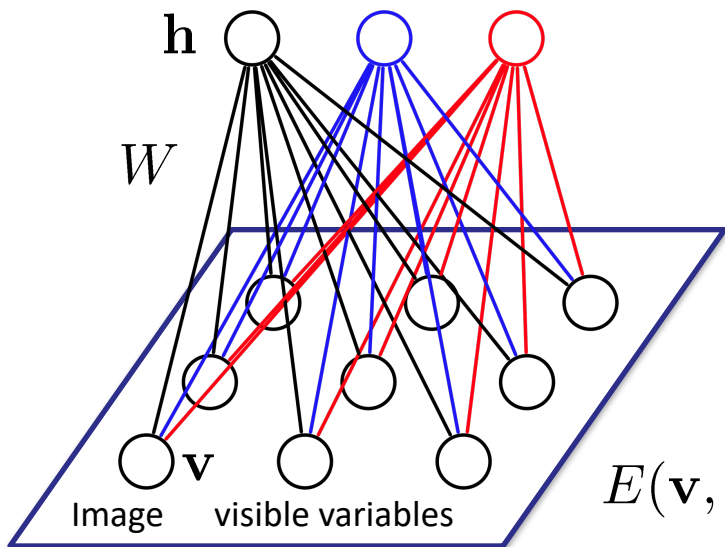
Represent:



$$\text{as } P(\mathbf{h}|\mathbf{v}) = [0, 0, 0.82, 0, 0, 0.99, 0, 0 \dots]$$

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:

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

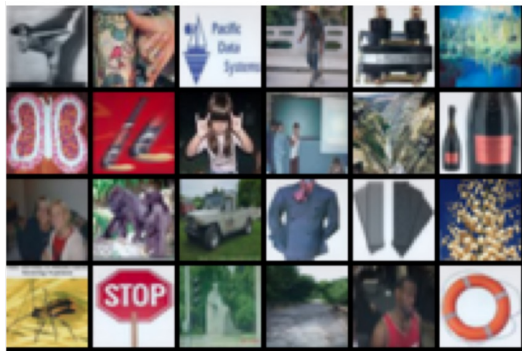
$$P(v_i = x | \mathbf{h}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - b_i - \sigma_i \sum_j W_{ij} h_j)^2}{2\sigma_i^2}\right) \quad \text{Gaussian}$$

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

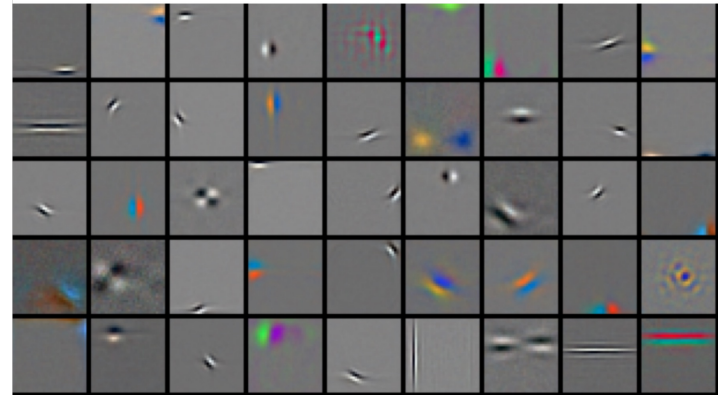
Gaussian-Bernoulli RBMs

Images: Gaussian-Bernoulli RBM

4 million unlabelled images



Learned features (out of 10,000)



Text: Multinomial-Bernoulli RBM



Reuters dataset:
804,414 unlabeled
newswire stories
Bag-of-Words



Learned features: "topics"

russian
russia
moscow
yeltsin
soviet

clinton
house
president
bill
congress

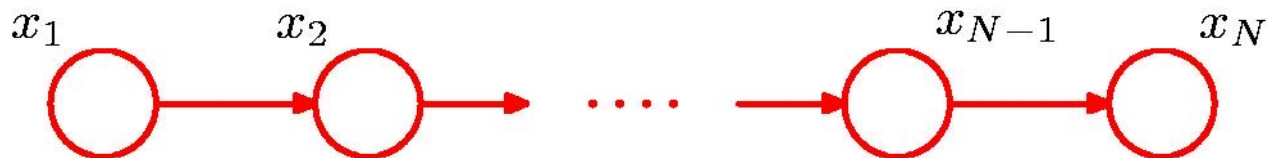
computer
system
product
software
develop

trade
country
import
world
economy

stock
wall
street
point
dow

Relation to Directed Graphs

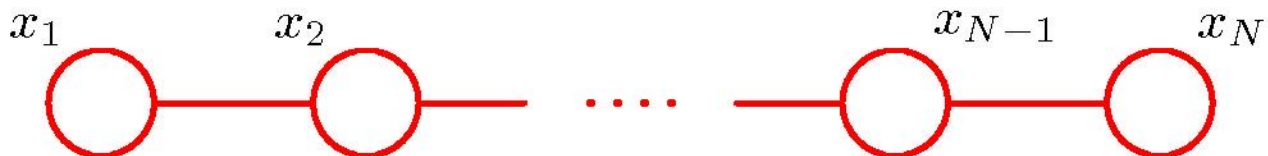
- Let us try to convert directed graph into an undirected graph:



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

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

Red double-headed arrows connect the underlined term in the first equation to $\psi_{1,2}(x_1, x_2)$ in the second equation. Similarly, red double-headed arrows connect $p(x_3|x_2)$ to $\psi_{2,3}(x_2, x_3)$ and $p(x_N|x_{N-1})$ to $\psi_{N-1,N}(x_{N-1}, x_N)$.

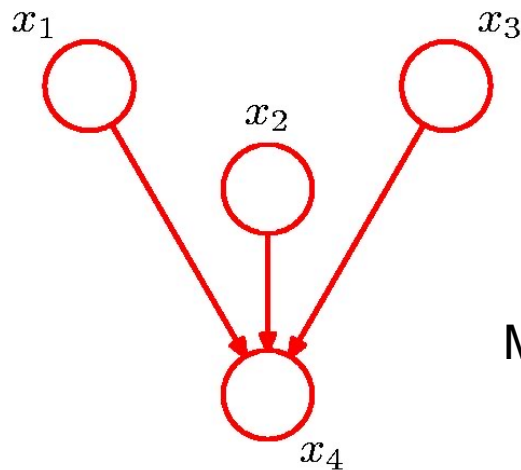


Directed vs. Undirected

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

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

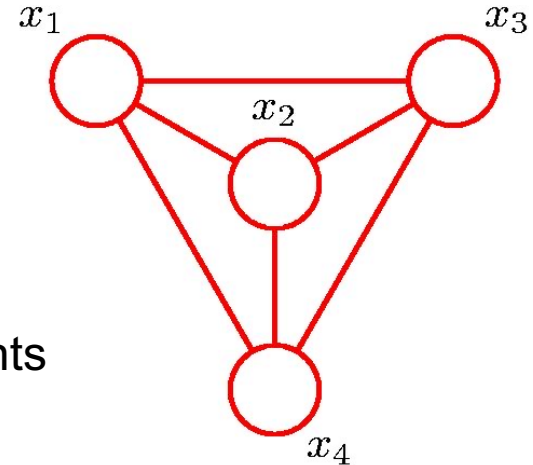
$$p(\mathbf{x}) = \frac{1}{Z} \psi(x_1, x_2, x_3, x_4)$$



need 4th
order clique



Moralize: Marry the parents



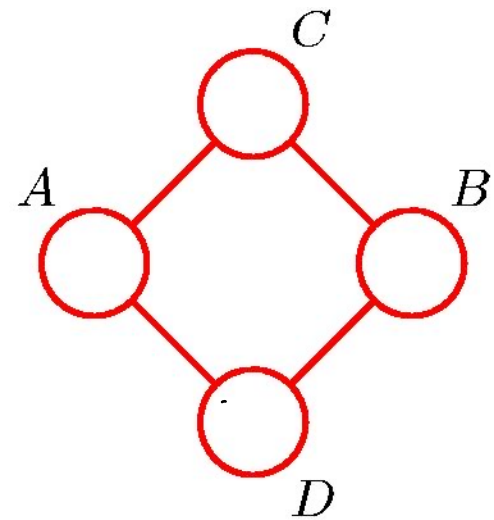
- All the parents of x_4 can interact to determine the distribution over x_4 .
- The directed graph represents independencies that the undirected graph cannot model.

- 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.



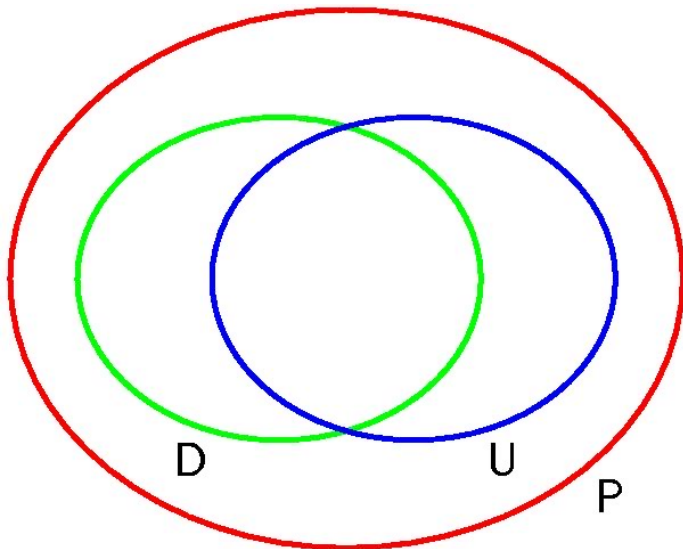
$$A \not\perp B \mid \emptyset$$

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

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

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.