Introduction to Machine Learning

Undirected Graphical Models

Barnabás Póczos

MACHINE LEARNING DEPARTMENT

Carnegie Mellon.
School of Computer Science

Credits

Many of these slides are taken from

Ruslan Salakhutdinov, Hugo Larochelle, & Eric Xing

- http://www.dmi.usherb.ca/~larocheh/neural_networks
- http://www.cs.cmu.edu/~rsalakhu/10707/
- http://www.cs.cmu.edu/~epxing/Class/10708/

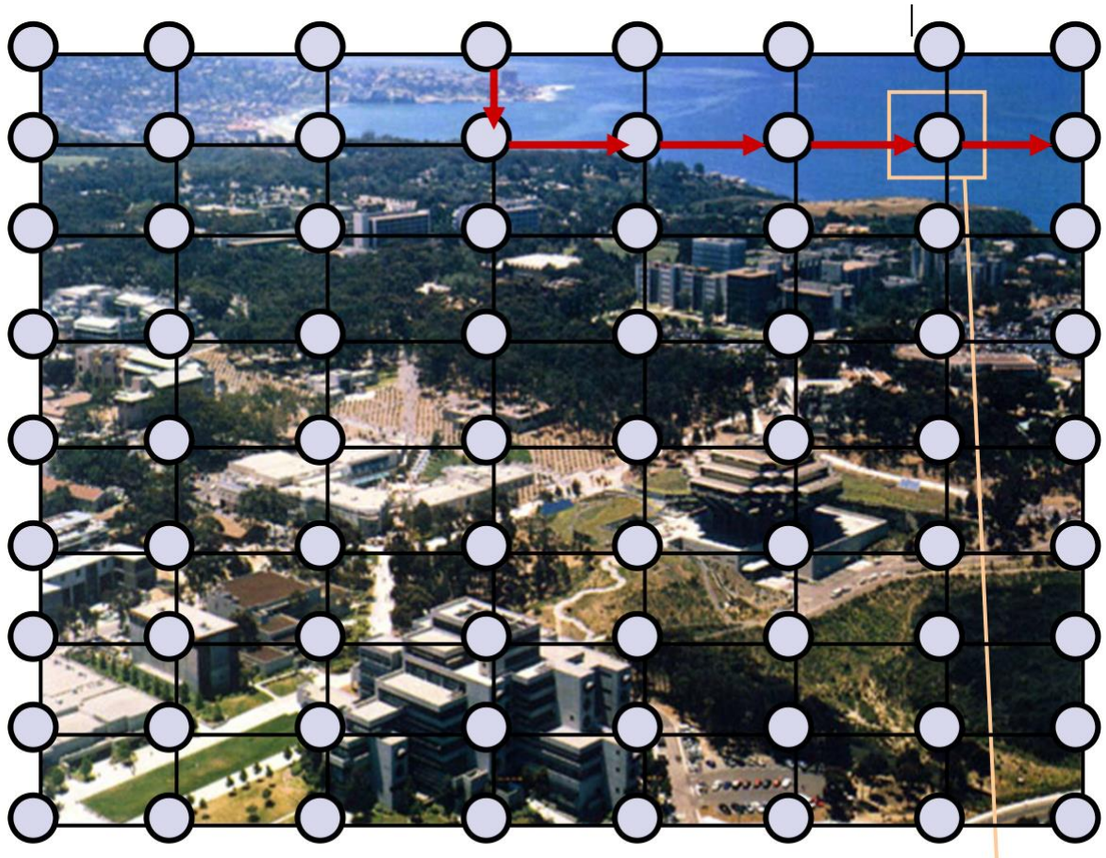
Reading material:

- http://www.cs.cmu.edu/~rsalakhu/papers/Russ_thesis.pdf
- Section 30.1 of Information Theory, Inference, and Learning Algorithms by David MacKay
- http://www.stat.cmu.edu/~larry/=sml/GraphicalModels.pdf

Undirected Graphical Models = Markov Random Fields

Probabilistic graphical models: a powerful framework for **representing dependency structure** between random variables.

Markov network (or undirected graphical model) is a set of random variables having a **dependency structure described by an undirected graph**.



Semantic labeling

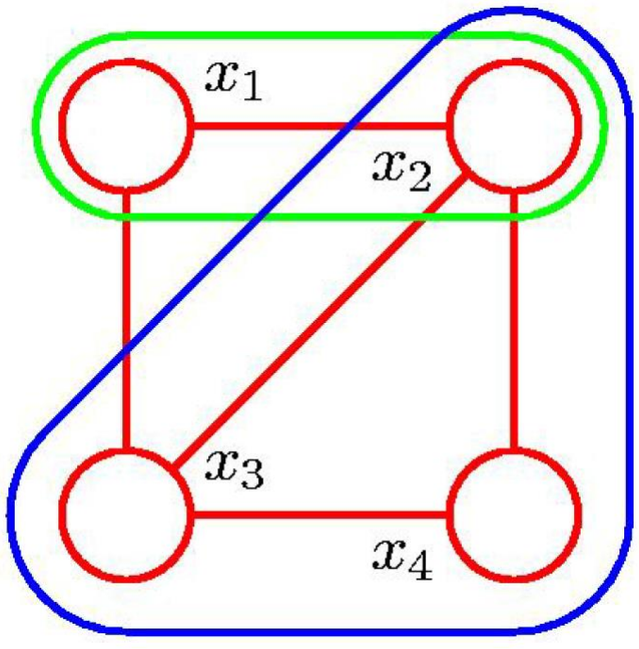
Cliques

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 two maximal cliques:

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

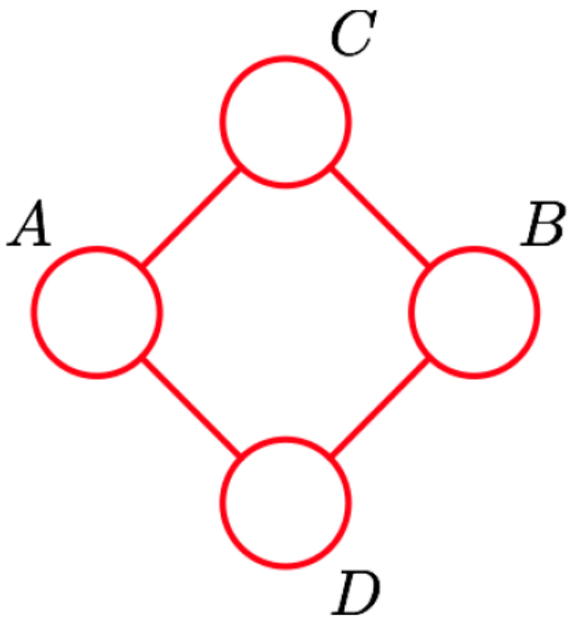


Other cliques:

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

Undirected Graphical Models = Markov Random Fields

Directed graphs are useful for expressing **causal relationships** between random variables, whereas **undirected graphs** are useful for expressing **dependencies** 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).

$$p(x_1, \ldots, x_d) = \frac{1}{\mathcal{Z}} \prod_{c \in \mathcal{C}} \phi_c(x_c), \qquad \mathcal{Z} = \sum_{x_1, \ldots, x_d} \prod_{c \in \mathcal{C}} \phi_c(x_c)$$

where the normalizing constant $\mathcal{Z}$ is called the partition function, and $\mathcal{C}$ is the set of maximal cliques.

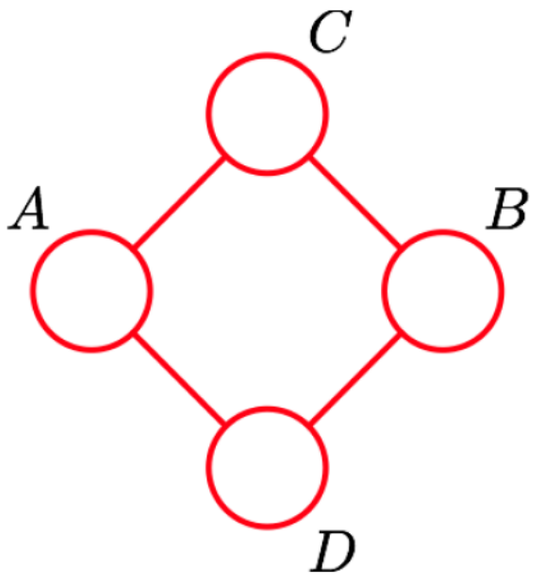
In this example, the joint distribution factorizes as:

$$p(A, B, C, D) = \frac{1}{\mathcal{Z}} \phi_{AC}(A, C) \phi_{CB}(C, B) \phi_{BD}(B, D) \phi_{AD}(A, D)$$

$$\text{where } \mathcal{Z} = \sum_{A,B,C,D} \phi_{AC}(A, C) \phi_{CB}(C, B) \phi_{BD}(B, D) \phi_{AD}(A, D)$$

$$\text{Maximal cliques: } \mathcal{C} = \{(A, C), (C, B), (B, D), (A, D)\}$$

Markov Random Fields (MRFs)



$$p(x_1, \ldots, x_d) = \frac{1}{\mathcal{Z}} \prod_{c \in \mathcal{C}} \phi_c(x_c), \qquad \mathcal{Z} = \sum_{x_1, \ldots, x_d} \prod_{c \in \mathcal{C}} \phi_c(x_c)$$

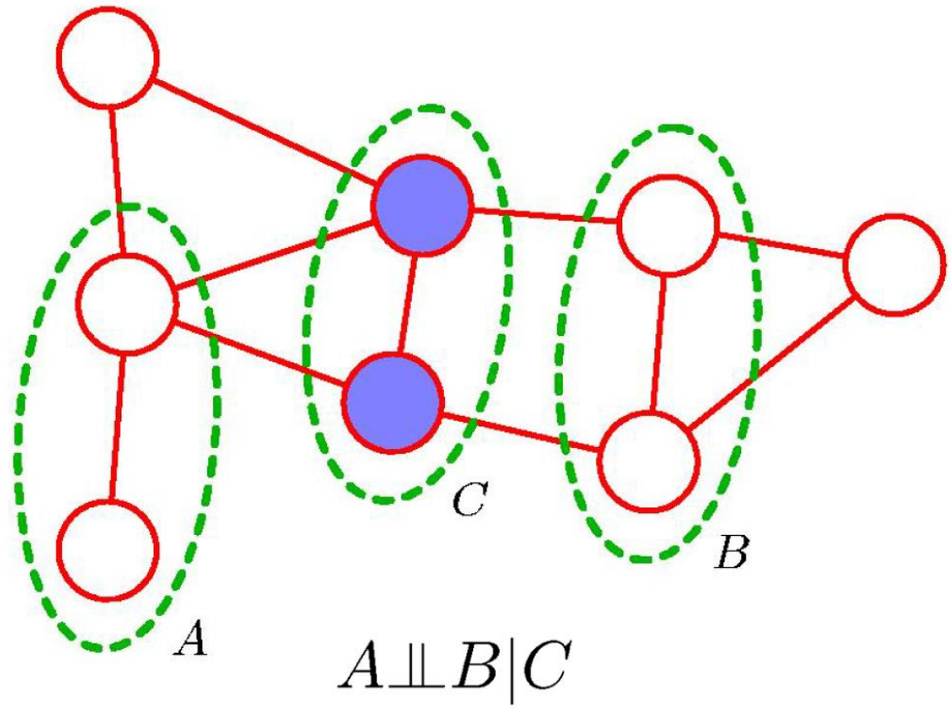
- Each potential function is a mapping from the joint configurations of random variables in a maximal clique to **non-negative real numbers**.
- The choice of potential functions is not restricted to having specific probabilistic interpretations.
- **Number of parameters:** $\sum_{c \in \mathcal{C}} 2^{|c|}$ (instead of $2^d - 1$)

$$\begin{aligned} p(x_1, \ldots, x_d) &= \frac{1}{\mathcal{Z}} \prod_{c \in \mathcal{C}} \phi_c(x_c) \\ &= \frac{1}{\mathcal{Z}} \exp\left(- \sum_{c \in \mathcal{C}} \log \frac{1}{\phi_c(x_c)} \right) \\ &= \frac{1}{\mathcal{Z}} \exp(- \sum_{c \in \mathcal{C}} E(x_c)) \end{aligned}$$

where *E(x)* is called an energy function.

Conditional Independence

Definition: [Global Markov Property] A probability distribution P for a random vector $X_1, \ldots, X_d$ satisfies the **global Markov property** with respect to an undirected graph G if for any disjoint vertex subsets A, B, and C such that C separates A and B, the random variables X_A and X_B are conditionally independent given X_C.



$A \perp\!\!\!\perp B | C$

It follows that the **undirected graphical structure represents conditional independence**:

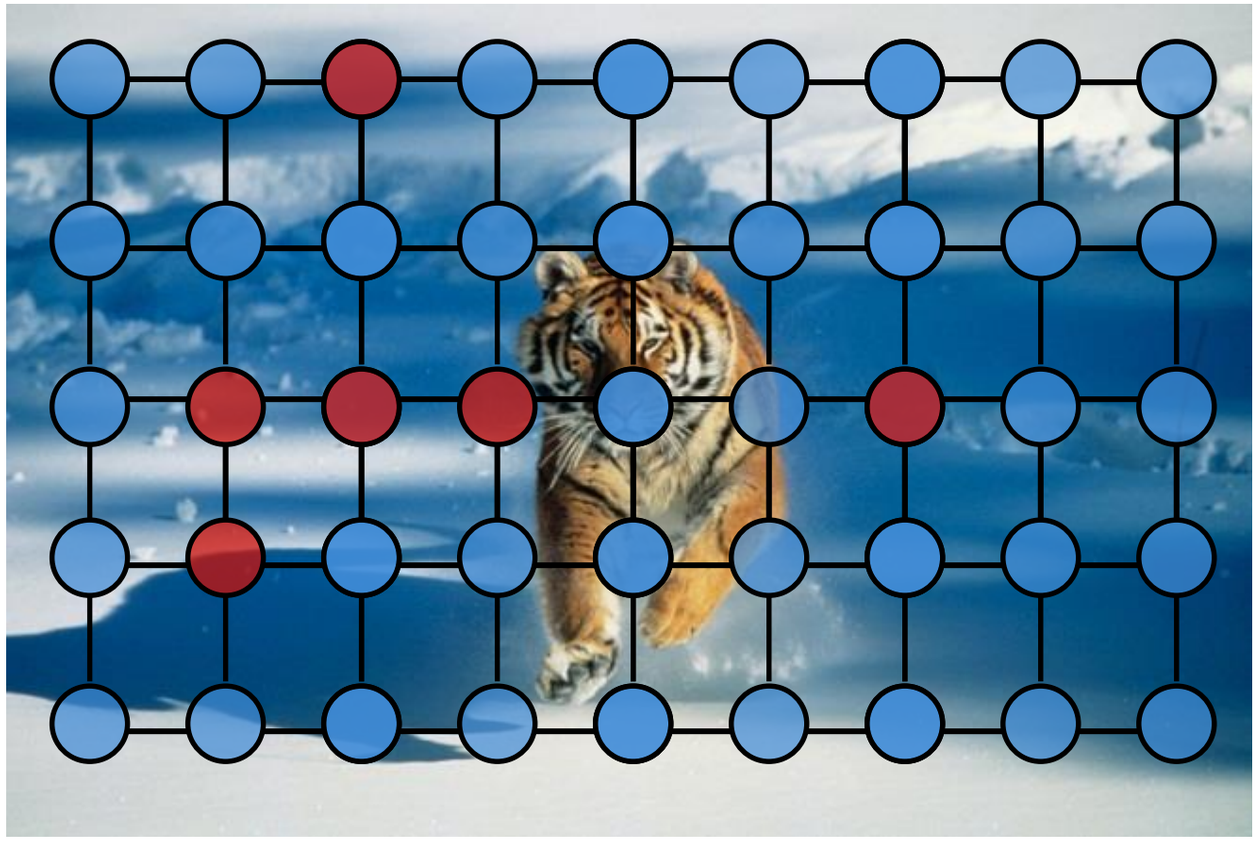
Theorem:

$p(x_1, \ldots, x_d) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c)$ satisfies the global Markov property. Hence their name is Markov Random Fields.

MRFs with Hidden Variables

For many interesting problems, we need to introduce **hidden or latent variables**.

- Our random variables will contain both visible and hidden variables x=(v,h)
- Computing the *Z* partition function is intractable
- Computing the summation over hidden variables is intractable
- Parameter learning is very challenging.

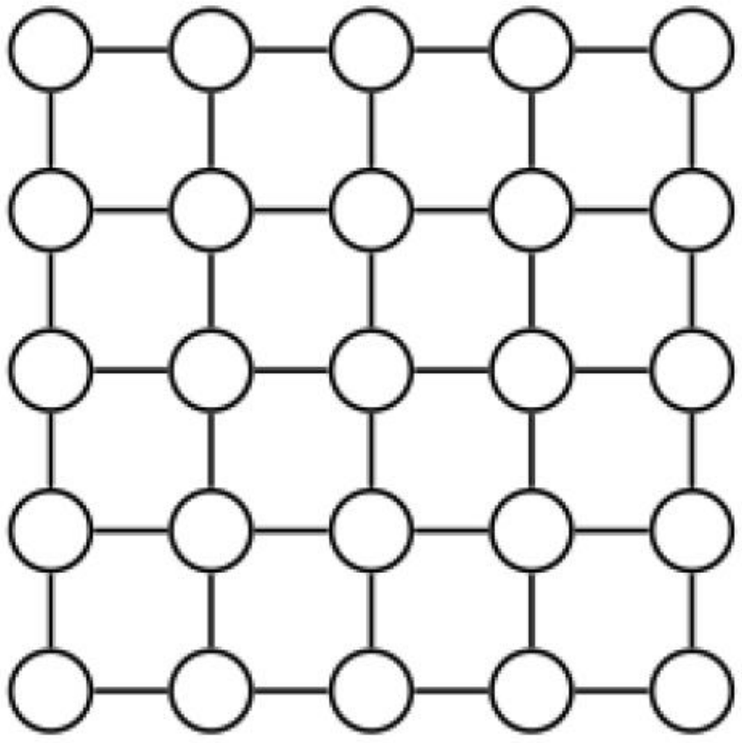


$$p(v) = \sum_h p(v,h)$$

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

Boltzmann Machines

Definition: [Boltzmann machines] MRFs with maximum click size two [pairwise (edge) potentials] on binary-valued nodes are called **Boltzmann machines**



The **joint probabilities** are 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 parameter θ_{ij} measures the **dependence of x_i on x_j, conditioned on the other nodes**.

Boltzmann Machines

Theorem: One can prove that the **conditional distribution of one node conditioned on the others** is given by the logistic function in Boltzmann Machines:

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

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

Proof:

$$p_\theta(x_1, \ldots, x_d) = \frac{1}{\mathcal{Z}(\theta)} \exp\left(\sum_{(i,j) \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \right)$$

$$p_\theta(x_1 | x_2, \ldots, x_d) = \frac{p_\theta(x_1, x_2, \ldots, x_d)}{p_\theta(x_2, \ldots, x_d)}$$

$$p_\theta(x_1 | x_2, \ldots, x_d) = \frac{p_\theta(x_1, x_2, \ldots, x_d)}{\sum_{x_1} p_\theta(x_1, x_2, \ldots, x_d)}$$

$$= \frac{\exp\left(\sum_{(i,j) \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \right)}{\sum_{x_1} \exp\left(\sum_{(i,j) \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \right)}$$

Boltzmann Machines

Proof [Continued]:

$$p_\theta(x_1|x_2,\ldots,x_d) =$$

$$= \frac{\exp\left(\sum_{(i,j)\in E} x_i x_j \theta_{ij} + \sum_{i\in V} x_i \theta_i\right)}{\sum_{x_1} \exp\left(\sum_{(i,j)\in E} x_i x_j \theta_{ij} + \sum_{i\in V} x_i \theta_i\right)}$$

$$- \frac{\exp\left(\sum_{(1,j)\in E} x_1 x_j \theta_{1j} + \sum_{(i\geq 2,j)\in E} x_i x_j \theta_{ij} + x_1\theta_1 + \sum_{i\geq 2\in V} x_i \theta_i\right)}{\sum_{x_1} \exp\left(\sum_{(1,j)\in E} x_1 x_j \theta_{1j} + \sum_{(i\geq 2,j)\in E} x_i x_j \theta_{ij} + x_1\theta_1 + \sum_{i\geq 2\in V} x_i \theta_i\right)}$$

$$= \frac{\exp\left(\sum_{(1,j)\in E} x_1 x_j \theta_{1j} + x_1\theta_1\right)}{\sum_{x_1} \exp\left(\sum_{(1,j)\in E} x_1 x_j \theta_{1j} + x_1\theta_1\right)}$$

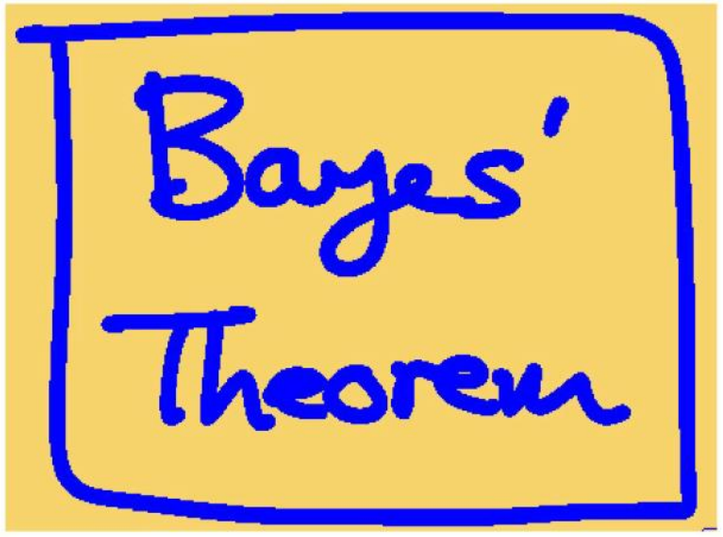
$$\Rightarrow p_\theta(x_1 = 1|x_2,\ldots,x_d) = \frac{\exp\left(\sum_{(1,j)\in E} x_j \theta_{1j} + \theta_1\right)}{\exp\left(\sum_{(1,j)\in E} x_j \theta_{1j} + \theta_1\right) + 1}$$

$$= \frac{1}{1 + \exp\left(-\sum_{(1,j)\in E} x_j \theta_{1j} - \theta_1\right)}$$

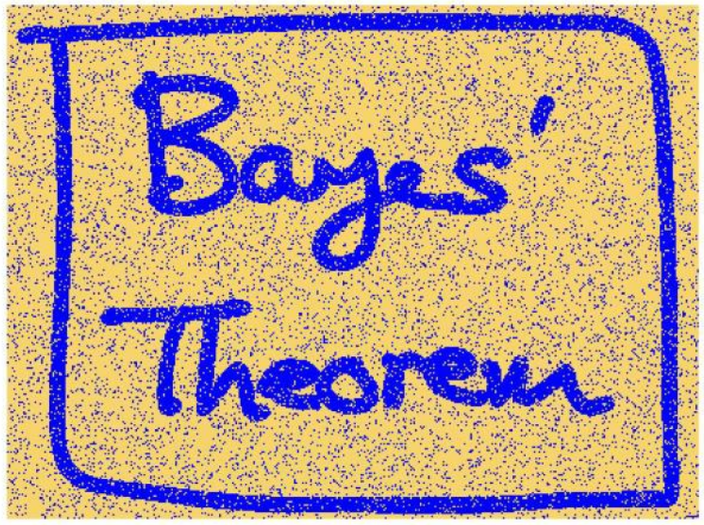
Q.E.D.

Example: Image Denoising

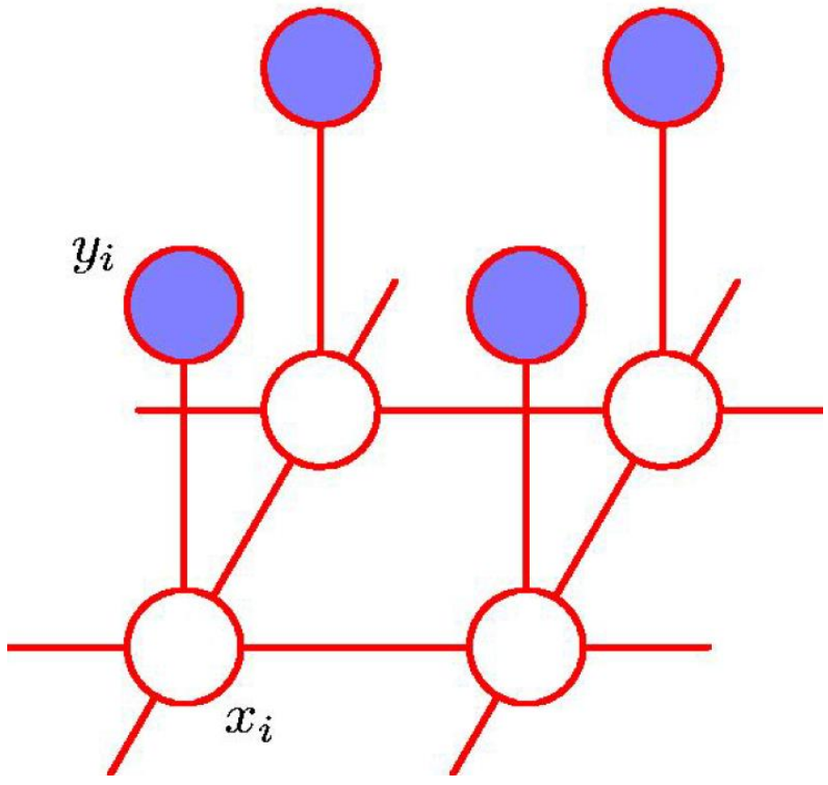
Let us look at the example of noise removal from a binary image.
The image is an array of {-1, +1} pixel values.



Original Image



Noisy Image



- We take the original noise-free image (x) and randomly flip the sign of pixels with a small probability. This process creates the noisy image (y)
- Our goal is to estimate the original image x from the noisy observations y.
- We model the joint distribution with

$$P(\mathbf{x}, \mathbf{y}) = \exp(-E(\mathbf{x}, \mathbf{y}))$$

$$\text{where } E(\mathbf{x}, \mathbf{y}) = \sum_i (y_i - x_i)^2 + \lambda \sum_{\{i,j\} \in E} (x_i - x_j)^2$$

Inference: Iterated Conditional Models

Goal: Using the observations $\mathbf{y}$ infer the unknown noise free pixels $\mathbf{x}$

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} P(\mathbf{x}, \mathbf{y})$$

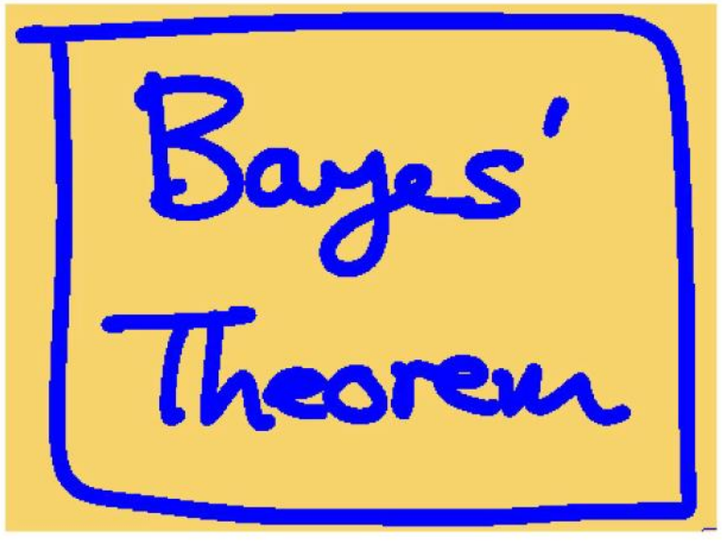
where $P(\mathbf{x}, \mathbf{y}) = \exp(-E(\mathbf{x}, \mathbf{y}))$

$$E(\mathbf{x}, \mathbf{y}) = \sum_i (y_i - x_i)^2 + \lambda \sum_{\{i,j\} \in E} (x_i - x_j)^2$$

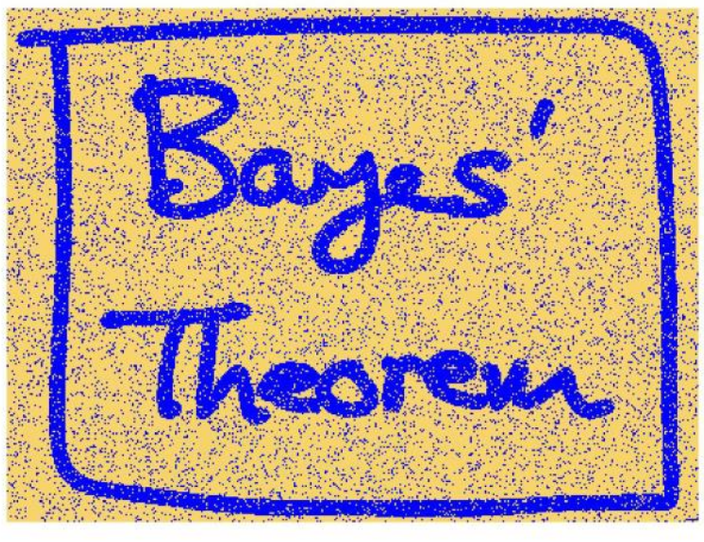
Solution: coordinate-wise gradient descent

- Iterated conditional modes: coordinate-wise gradient descent.

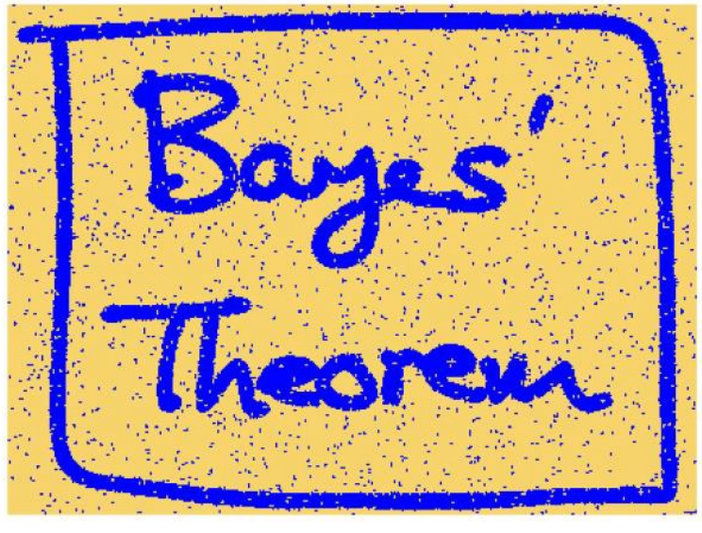
- Visit the unobserved nodes sequentially and set each x to whichever of its two values has the lowest energy.



Original Image



Noisy Image



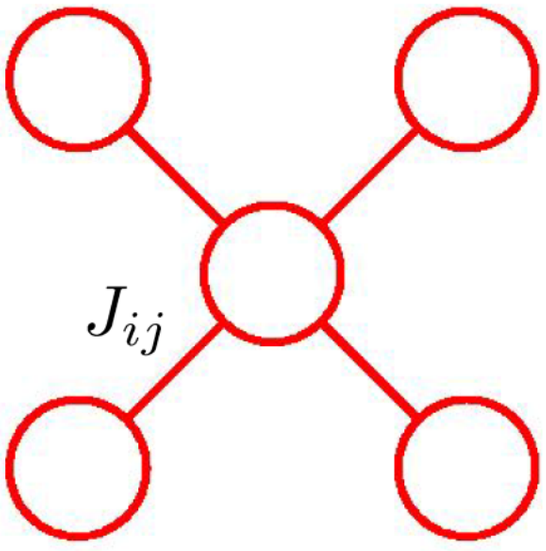
ICM

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

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

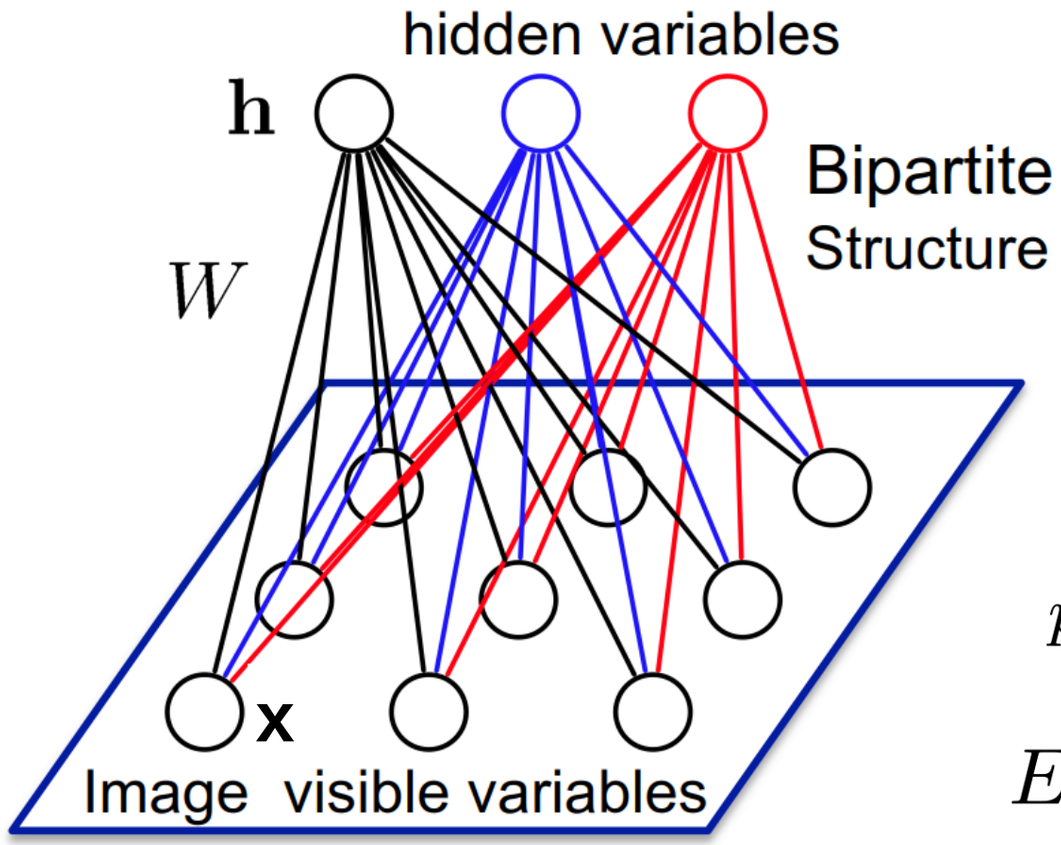
if $(i,j) \neq E,$ then $J_{ij} = 0.$

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

Restricted Boltzmann Machines

Restricted Boltzmann Machines

Restricted = no connections in the hidden layer + no connection in the visible layer



hidden variables: $\mathbf{h} \in \{0,1\}^{d_h}$

visible variables: $\mathbf{x} \in \{0,1\}^{d_x}$

parameters: $\theta = \{\mathbf{W}, \mathbf{c}, \mathbf{b}\}$

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

$$E(\mathbf{x}, \mathbf{h}; \theta) = -\mathbf{h}^T \mathbf{W} \mathbf{x} - \mathbf{c}^T \mathbf{x} - \mathbf{b}^T \mathbf{h}$$

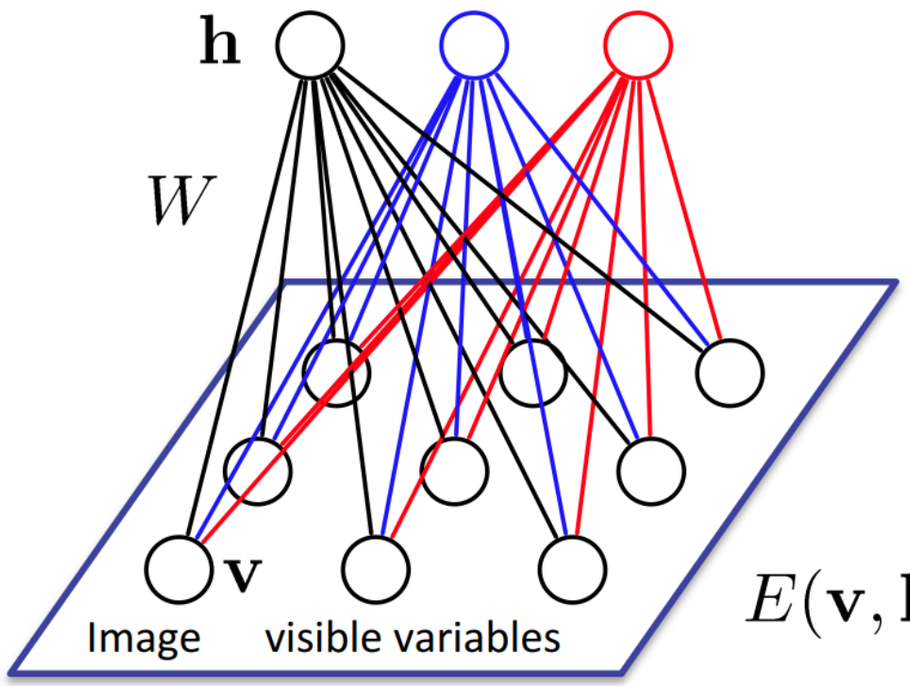
No $\mathbf{x}^T \Sigma \mathbf{x}$ or $\mathbf{h}^T \Sigma \mathbf{h}$ terms!

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

Partition function (intractable)

Gaussian-Bernoulli RBM

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:

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

[Quadratic in v linear in h]

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

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

Possible Tasks with RBM

Tasks:

RBM model: $p_\theta(\mathbf{x}, \mathbf{h})$

- Inference:

$$p_\theta(\mathbf{h}|\mathbf{x}) = ?$$

- Evaluate the likelihood function:

$$p_\theta(\mathbf{x}) = ?$$

- Sampling from RBM:

$$\tilde{\mathbf{x}}, \tilde{\mathbf{h}} \sim p_\theta(\mathbf{x}, \mathbf{h})$$

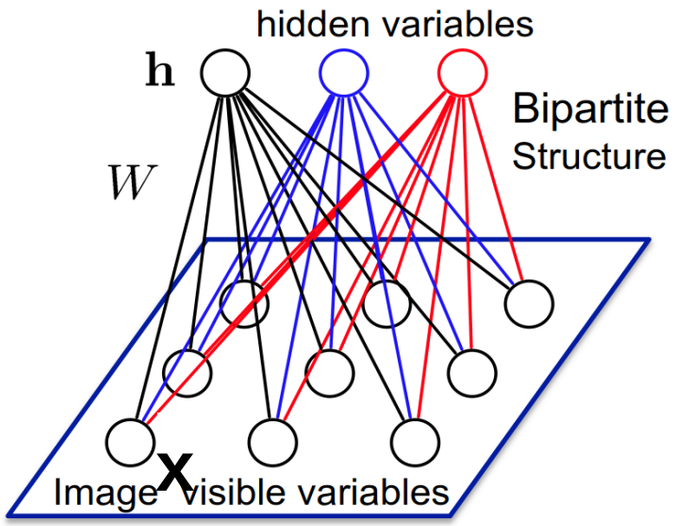
- Training RBM:

$$\max_\theta p_\theta(\mathbf{x}) = ?$$

Inference

Inference

Theorem: **Inference in RBM is simple: the conditional distributions are logistic functions**



$$p_\theta(\mathbf{h}|\mathbf{x}) = \prod_{j=1}^{d_h} p_\theta(h_j|\mathbf{x})$$

$$p_\theta(h_j = 1|\mathbf{x}) = \frac{1}{1 + \exp(-b_j - \mathbf{W}_j.\mathbf{x})}$$

$$= sigm(b_j + \mathbf{W}_j.\mathbf{x})$$

j^{th} row of $\mathbf{W}$

Similarly,

$$p_\theta(\mathbf{x}|\mathbf{h}) = \prod_{j=1}^{d_x} p_\theta(x_k|\mathbf{h})$$

$$p_\theta(x_k = 1|\mathbf{h}) = \frac{1}{1 + \exp(-c_k - h^T\mathbf{W}_{.k})}$$

$$= sigm(c_k + \mathbf{h}^T\mathbf{W}_{.k})$$

k^{th} column of $\mathbf{W}$

Proof:

$$p_\theta(\mathbf{h}|\mathbf{x}) = \frac{p_\theta(\mathbf{h},\mathbf{x})}{\sum_{\mathbf{h}'} p_\theta(\mathbf{h}',\mathbf{x})}$$

$$= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})/\mathcal{Z}}{\sum_{\mathbf{h}'\in\{0,1\}^{d_h}} \exp(\mathbf{h}'^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h}')/\mathcal{Z}}$$

$$= \frac{\exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{b}^T\mathbf{h})}{\sum_{\mathbf{h}'\in\{0,1\}^{d_h}} \exp(\mathbf{h}'^T\mathbf{W}\mathbf{x} + \mathbf{b}^T\mathbf{h}')}$$

$$= \frac{\exp(\sum_{j=1}^{d_h} h_j\mathbf{W}_{j.}\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}}\cdots\sum_{h'_{d_h}\in\{0,1\}} \exp(\sum_{j=1}^{d_h} h'_j\mathbf{W}_{j.}\mathbf{x} + b_j h'_j))}$$

$$= \frac{\prod_{j=1}^{d_h} \exp(h_j\mathbf{W}_{j.}\mathbf{x} + b_j h_j)}{\sum_{h'_1\in\{0,1\}}\cdots\sum_{h'_{d_h}\in\{0,1\}} \prod_{j=1}^{d_h} \exp(h'_j\mathbf{W}_{j.}\mathbf{x} + b_j h'_j))}$$

Proof [Continued]:

$$p_\theta(\mathbf{h}|\mathbf{x}) = \frac{\prod_{j=1}^{d_h} \exp(h_j \mathbf{W}_j.\mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_{d_h} \in \{0,1\}} \prod_{j=1}^{d_h} \exp(h'_j \mathbf{W}_j.\mathbf{x} + b_j h'_j))}$$

$$= \frac{\prod_{j=1}^{d_h} \exp(h_j \mathbf{W}_j.\mathbf{x} + b_j h_j)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 \mathbf{W}_1.\mathbf{x} + b_1 h'_1)\right) \cdots \left(\sum_{h'_{d_h} \in \{0,1\}} \exp(h'_{d_h} \mathbf{W}_{d_h}.\mathbf{x} + b_{d_h} h'_{d_h})\right)}$$

$$= \frac{\prod_{j=1}^{d_h} \exp(h_j \mathbf{W}_j.\mathbf{x} + b_j h_j)}{\prod_{j=1}^{d_h} \left(\sum_{h'_j \in \{0,1\}} \exp(h'_j \mathbf{W}_j.\mathbf{x} + b_j h'_j)\right)}$$

$$= \frac{\prod_{j=1}^{d_h} \exp(h_j \mathbf{W}_j.\mathbf{x} + b_j h_j)}{\prod_{j=1}^{d_h} \left(1 + \exp(\mathbf{W}_j.\mathbf{x} + b_j)\right)}$$

$$= \prod_{j=1}^{d_h} \frac{\exp(h_j \mathbf{W}_j.\mathbf{x} + b_j h_j)}{1 + \exp(\mathbf{W}_j.\mathbf{x} + b_j)}$$

$$= \prod_{j=1}^{d_h} p_\theta(h_j|\mathbf{x})$$

$$\begin{aligned} p(h_j = 1|\mathbf{x}) &= \frac{\exp(b_j + \mathbf{W}_j.\mathbf{x})}{1 + \exp(b_j + \mathbf{W}_j.\mathbf{x})} \\ &= \frac{1}{1 + \exp(-b_j - \mathbf{W}_j.\mathbf{x})} \\ &= \text{sigm}(b_j + \mathbf{W}_j.\mathbf{x}) \end{aligned}$$

Q.E.D.

Evaluating the Likelihood

Calculating the Likelihood of an RBM

Theorem: **Calculating the likelihood is simple in RBM (apart from the partition function)**

$$E(\mathbf{x}, \mathbf{h}; \theta) = -\mathbf{h}^T \mathbf{W} \mathbf{x} - \mathbf{c}^T \mathbf{x} - \mathbf{b}^T \mathbf{h}$$

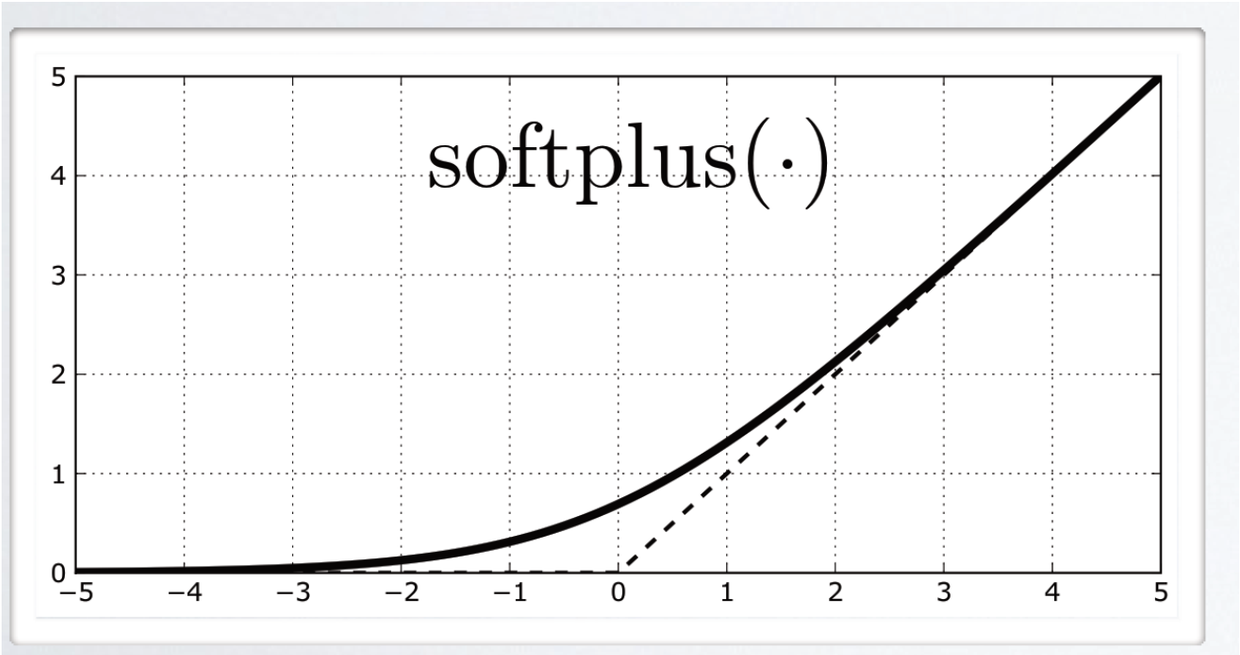
$$p_\theta(\mathbf{x}) = \sum_{\mathbf{h} \in \{0,1\}^{d_h}} p_\theta(\mathbf{x}, \mathbf{h}) = \sum_{\mathbf{h} \in \{0,1\}^{d_h}} \frac{1}{\mathcal{Z}(\theta)} \exp(-E(\mathbf{x}, \mathbf{h}; \theta))$$

$$= \exp\left(\mathbf{c}^T \mathbf{x} + \sum_{j=1}^{d_h} \log(1 + \exp(b_j + \mathbf{W}_j . \mathbf{x})) \right) / \mathcal{Z}(\theta)$$

$$= \exp\left(\mathbf{c}^T \mathbf{x} + \sum_{j=1}^{d_h} \text{softplus}(b_j + \mathbf{W}_j . \mathbf{x}) \right) / \mathcal{Z}(\theta)$$

$$= \exp\left(-F_\theta(\mathbf{x})\right) / \mathcal{Z}(\theta)$$

Free energy



Proof:

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{h}\in\{0,1\}^{d_h}} p_\theta(\mathbf{x},\mathbf{h}) = \sum_{\mathbf{h}\in\{0,1\}^{d_h}} \frac{1}{\mathcal{Z}} \exp(-E(\mathbf{x},\mathbf{h};\theta))$$

$$= \sum_{\mathbf{h}\in\{0,1\}^{d_h}} \frac{1}{\mathcal{Z}} \exp(\mathbf{h}^T\mathbf{W}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{b}^T\mathbf{h})$$

$$= \exp(\mathbf{c}^T\mathbf{x})\frac{1}{\mathcal{Z}} \sum_{h'_1\in\{0,1\}} \cdots \sum_{h_{d_h}\in\{0,1\}} \exp(\mathbf{h}^T\mathbf{W}\mathbf{x}+\mathbf{b}^T\mathbf{h})$$

$$= \exp(\mathbf{c}^T\mathbf{x})\frac{1}{\mathcal{Z}} \sum_{h'_1\in\{0,1\}} \cdots \sum_{h_{d_h}\in\{0,1\}} \exp(\sum_{j=1}^{d_h} h_j\mathbf{W}_j.\mathbf{x}+b_j h_j)$$

$$= \exp(\mathbf{c}^T\mathbf{x})\frac{1}{\mathcal{Z}} \left(\sum_{h_1\in\{0,1\}} \exp(h_1\mathbf{W}_1.\mathbf{x} + b_1 h_1)\right) \cdots \left(\sum_{h_{d_h}\in\{0,1\}} \exp(h_{d_h}\mathbf{W}_{d_h}.\mathbf{x} + b_{d_h} h_{d_h})\right)$$

$$= \exp(\mathbf{c}^T\mathbf{x})\frac{1}{\mathcal{Z}} (1 + \exp(\mathbf{W}_1.\mathbf{x} + b_1)) \cdots \left(1 + \exp(\mathbf{W}_{d_h}.\mathbf{x} + b_{d_h})\right)$$

$$= \exp(\mathbf{c}^T\mathbf{x})\frac{1}{\mathcal{Z}} \exp\left(\log(1 + \exp(\mathbf{W}_1.\mathbf{x} + b_1))\right) \cdots \exp\left(\log(1 + \exp(\mathbf{W}_{d_h}.\mathbf{x} + b_{d_h}))\right)$$

$$= \exp\left(\mathbf{c}^T\mathbf{x} + \sum_{j=1}^{d_h} \log(1 + \exp(b_j + \mathbf{W}_j.\mathbf{x}))\right)/\mathcal{Z}$$

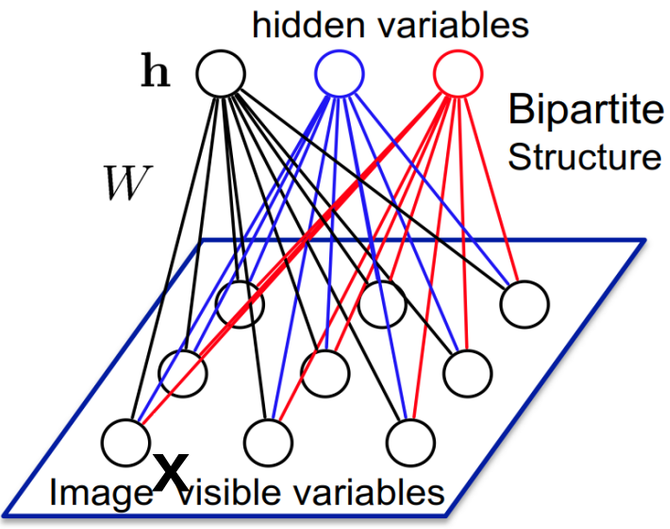
Q.E.D.

Sampling

Sampling from p(x,h) in RBM

Goal: Generate samples from $\tilde{\mathbf{x}}, \tilde{\mathbf{h}} \sim p_\theta(\mathbf{x}, \mathbf{h})$

Sampling is tricky… it is easier much in directed graphical models.
Here we will use **Gibbs sampling**.



$$p_\theta(\mathbf{h}|\mathbf{x}) = \prod_{i=1}^{d_h} p_\theta(h_j|\mathbf{x})$$

$$p_\theta(h_j = 1|\mathbf{x}) = \frac{1}{1 + \exp(-b_j - \mathbf{W}_{j.}\mathbf{x})}$$

$$= sigm(b_j + \mathbf{W}_{j.}\mathbf{x})$$

j^{th} row of $\mathbf{W}$

Similarly, $p_\theta(\mathbf{x}|\mathbf{h}) = \prod_{j=1}^{d_x} p_\theta(x_k|\mathbf{h})$

$$p_\theta(x_k = 1|\mathbf{h}) = \frac{1}{1 + \exp(-c_k - h^T\mathbf{W}_{.k})}$$

$$= sigm(c_k + \mathbf{h}^T\mathbf{W}_{.k})$$

k^{th} column of $\mathbf{W}$

Gibbs Sampling: The Problem

Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$

Let $p(x_1, \ldots, x_n) \geq 0$ be a non-normalized distribution ($\int p(x) \neq 1$, $p(x) \geq 0$), and let A be a complicated set.

Suppose that we can generate samples from

$$P(X_i = x | X_j = x_j, \forall j \neq i)$$

e.g. $P(X_3 = x_3 | X_1 = x_1, X_2 = x_2, X_4 = x_4, X_5 = x_5)$

Our goal is to generate samples from

$$f(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if } \mathbf{x} \notin A \\ \frac{p(\mathbf{x})}{p(\mathbf{x} \in A)} & \text{if } \mathbf{x} \in A \end{cases}$$

Gibbs Sampling: Pseudo Code

1. We are in $\mathbf{x} = (x_1, \ldots, x_n) \in A$

2. Draw a random state $i \in \{1, \ldots, n\}$ with prob. $1/n$.

3. Sample x from $x \sim P(X_i = x | X_j = x_j, \forall j \neq i)$.

4. Let $\mathbf{y} = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$

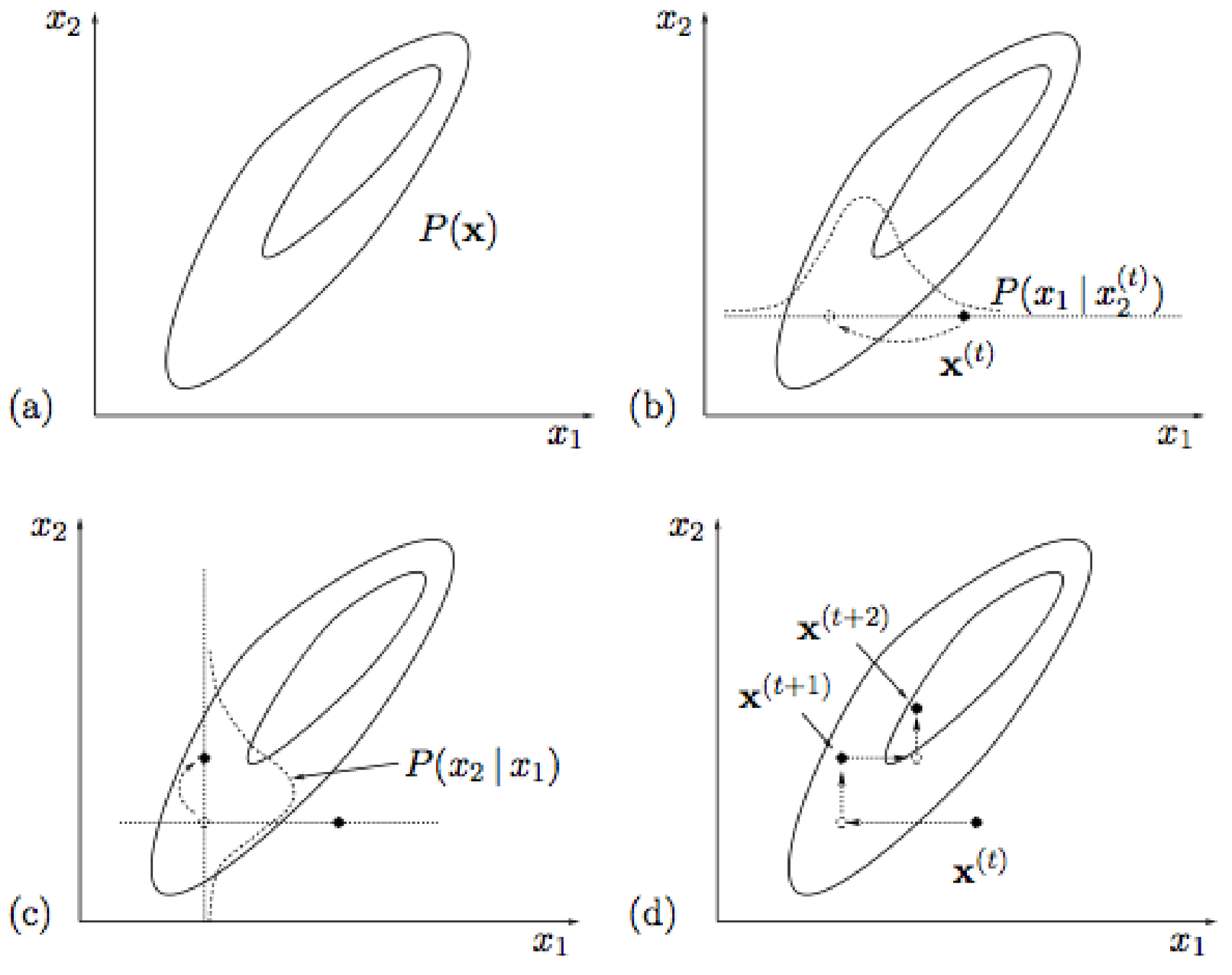
5. If

   $(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \in \mathbf{A} \Rightarrow x_i = x$, accept this new state

   $(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \notin \mathbf{A} \Rightarrow x_i$ stays in the old x_i

6. New sample point: $(x_1, \ldots, x_n)$. Go back to 2

Gibbs Sampling



Training

RBM Training

Training is complicated…

To train an RBM, we would like to minimize the negative log-likelihood function:

$$\min_{\theta} -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(\mathbf{x}^{(t)}) = ?$$

To solve this, we use stochastic gradient ascent:

Theorem:

$$\frac{\partial}{\partial\theta} -\log p_\theta(\mathbf{x}(t)) = \underbrace{\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}(t),\mathbf{h})}{\partial\theta}\middle|\mathbf{x}(t)\right]}_{\text{Positive phase}} - \underbrace{\mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]}_{\text{Negative phase (hard to computer)}}$$

Data-Dependent
Expectations w.r.t P(h|x)

Model: Expectation
w.r.t joint P(x,h)

RBM Training

Proof:

$$\mathcal{Z}(\theta) = \sum_{\mathbf{h},\mathbf{x}} \exp\left(\mathbf{h}^T \mathbf{W}\mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h})\right)$$

$$\frac{\partial}{\partial \theta} -\log p_\theta(\mathbf{x}(t)) = -\frac{\partial}{\partial \theta} \log \left(\sum_{\mathbf{h} \in \{0,1\}^{d_h}} \frac{1}{\mathcal{Z}(\theta)} \exp(\mathbf{h}^T \mathbf{W}\mathbf{x}(t) + \mathbf{c}^T \mathbf{x}(t) + \mathbf{b}^T \mathbf{h}) \right)$$

$$= -\frac{\partial}{\partial \theta} \log \left(\frac{1}{\mathcal{Z}(\theta)} \right) - \frac{\partial}{\partial \theta} \log \left(\sum_{\mathbf{h} \in \{0,1\}^{d_h}} \exp(\mathbf{h}^T \mathbf{W}\mathbf{x}(t) + \mathbf{c}^T \mathbf{x}(t) + \mathbf{b}^T \mathbf{h}) \right)$$

$$= \frac{\partial}{\partial \theta} \log \left(\sum_{\mathbf{h},\mathbf{x}} \exp\left(\mathbf{h}^T \mathbf{W}\mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h})\right) \right) - \frac{\partial}{\partial \theta} \log \left(\sum_{\mathbf{h} \in \{0,1\}^{d_h}} \exp(\mathbf{h}^T \mathbf{W}\mathbf{x}(t) + \mathbf{c}^T \mathbf{x}(t) + \mathbf{b}^T \mathbf{h}) \right)$$

RBM Training

Proof [Continued]:

$$= \frac{\partial}{\partial \theta} \log \left(\sum_{\mathbf{h},\mathbf{x}} \exp\left(\mathbf{h}^T \mathbf{W} \mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h}\right) \right) - \frac{\partial}{\partial \theta} \log \left(\sum_{\mathbf{h} \in \{0,1\}^{d_h}} \exp(\mathbf{h}^T \mathbf{W} \mathbf{x}(t) + \mathbf{c}^T \mathbf{x}(t) + \mathbf{b}^T \mathbf{h}) \right)$$

First term

Second term

First term:

$$\frac{\partial}{\partial \theta} \log \left(\sum_{\mathbf{h},\mathbf{x}} \exp\left(\mathbf{h}^T \mathbf{W} \mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h}\right) \right) = \frac{\sum_{\mathbf{h},\mathbf{x}} \exp\left(\mathbf{h}^T \mathbf{W} \mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h}\right)}{\sum_{\tilde{\mathbf{h}},\tilde{\mathbf{x}}} \exp\left(\tilde{\mathbf{h}}^T \mathbf{W} \tilde{\mathbf{x}} + \mathbf{c}^T \tilde{\mathbf{x}} + \mathbf{b}^T \tilde{\mathbf{h}}\right)} \frac{\partial (\mathbf{h}^T \mathbf{W} \mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h})}{\partial \theta}$$

Difficult to calculate the expectation

$$= \mathbb{E}_{\mathbf{x},\mathbf{h}} \left[\frac{\partial (\mathbf{h}^T \mathbf{W} \mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h})}{\partial \theta} \right]$$

$$= \mathbb{E}_{p_{model}} \left[\frac{\partial (\mathbf{h}^T \mathbf{W} \mathbf{x} + \mathbf{c}^T \mathbf{x} + \mathbf{b}^T \mathbf{h})}{\partial \theta} \right]$$

$$= -\mathbb{E}_{\mathbf{x},\mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right]$$ **Negative phase**

RBM Training

Proof [Continued]:

$$\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t) + \mathbf{c}^T\mathbf{x}(t) + \mathbf{b}^T\mathbf{h})}{\partial\theta} = ?$$

$$\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t) + \mathbf{c}^T\mathbf{x}(t) + \mathbf{b}^T\mathbf{h})}{\partial\mathbf{W}} = \mathbf{h}\mathbf{x}(t)^T$$

$$\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t) + \mathbf{c}^T\mathbf{x}(t) + \mathbf{b}^T\mathbf{h})}{\partial\mathbf{c}} = \mathbf{x}(t)$$

$$\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t) + \mathbf{c}^T\mathbf{x}(t) + \mathbf{b}^T\mathbf{h})}{\partial\mathbf{b}} = \mathbf{h}$$

RBM Training

Proof [Continued]:

$$= \frac{\partial}{\partial\theta}\log\left(\sum_{\mathbf{h},\mathbf{x}}\exp\left(\mathbf{h}^T\mathbf{W}\mathbf{x}+\mathbf{c}^T\mathbf{x}+\mathbf{b}^T\mathbf{h}\right)\right) - \frac{\partial}{\partial\theta}\log\left(\sum_{\mathbf{h}\in\{0,1\}^{d_h}}\exp(\mathbf{h}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\mathbf{h})\right)$$

First term **Second term**

Second term:

$$-\frac{\partial}{\partial\theta}\log\left(\sum_{\mathbf{h}\in\{0,1\}^{d_h}}\exp(\mathbf{h}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\mathbf{h})\right) =$$

$$= -\frac{\sum_{\mathbf{h}\in\{0,1\}^{d_h}}\exp(\mathbf{h}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\mathbf{h})}{\sum_{\tilde{\mathbf{h}}\in\{0,1\}^{d_h}}\exp(\tilde{\mathbf{h}}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\tilde{\mathbf{h}})}\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\mathbf{h})}{\partial\theta}$$

$$= -\mathbb{E}_{\mathbf{h}}\left[\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\mathbf{h})}{\partial\theta}\middle|\mathbf{x}(t)\right]$$

$$= -\mathbb{E}_{p_{data}}\left[\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\mathbf{h})}{\partial\theta}\right]$$

$$= \mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\middle|\mathbf{x}(t)\right]$$ **Positive phase**

The conditionals are independent logistic distributions

Q.E.D

RBM Training

Since $$\frac{\partial}{\partial\theta}-\log p_\theta(\mathbf{x}(t)) = \mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}(t),\mathbf{h})}{\partial\theta}\middle|\mathbf{x}(t)\right] - \mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]$$

We need to calculate $\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}(t),\mathbf{h})}{\partial\theta}\middle|\mathbf{x}(t)\right]$ and $\mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]$

The first term (positive phase) can be caclulated using $p_\theta(\mathbf{h}|\mathbf{x}(t))$ logistic distribution.

$$\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}(t),\mathbf{h})}{\partial W_{jk}}\middle|\mathbf{x}(t)\right] = -\mathbb{E}_{\mathbf{h}}\left[\frac{\partial(\mathbf{h}^T\mathbf{W}\mathbf{x}(t)+\mathbf{c}^T\mathbf{x}(t)+\mathbf{b}^T\mathbf{h})}{\partial W_{jk}}\middle|\mathbf{x}(t)\right] =$$

$$= -\mathbb{E}_{\mathbf{h}}\left[h_j x_k(t)\middle|\mathbf{x}(t)\right] = -\sum_{h_j\in\{0,1\}} h_j x_k(t) p_\theta(h_j|\mathbf{x}(t))$$

$$= -x_k(t)p_\theta(h_j=1|\mathbf{x}(t)) \quad = -x_k(t)sigm(b_j+\mathbf{W}_j.\mathbf{x}(t))$$

$$\Rightarrow \mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}(t),\mathbf{h})}{\partial\mathbf{W}}\middle|\mathbf{x}(t)\right] = -\mathbf{h}(\mathbf{x}(t))\mathbf{x}(t)^T \quad \text{where} \quad \mathbf{h}(\mathbf{x}) \overset{\text{def}}{=} \begin{pmatrix} p(h_1=1|\mathbf{x}) \\ \cdots \\ p(h_H=1|\mathbf{x}) \end{pmatrix}$$

$$= \text{sigm}(\mathbf{b}+\mathbf{W}\mathbf{x})$$

RBM Training

Since $$\frac{\partial}{\partial\theta}-\log p_\theta(\mathbf{x}(t)) = \mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}(t),\mathbf{h})}{\partial\theta}\middle|\mathbf{x}(t)\right] - \mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]$$

We need to calculate $\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x}(t),\mathbf{h})}{\partial\theta}\middle|\mathbf{x}(t)\right]$ and $\mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right]$

The second term is more tricky.

Approximate the expectations with a single sample:

$$\mathbb{E}_{\mathbf{x},\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\theta}\middle|\mathbf{x}\right]\right] \approx \mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\tilde{\mathbf{x}},\mathbf{h})}{\partial\theta}\middle|\tilde{\mathbf{x}}\right]$$

The missing $\tilde{\mathbf{x}}$ can be generated by Gibbs sampling.

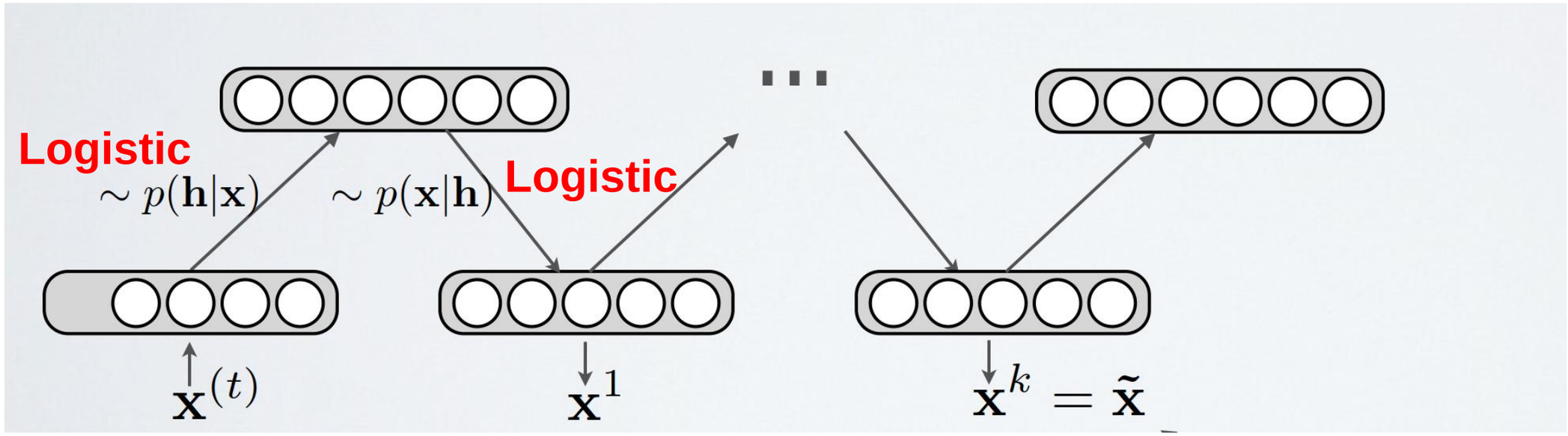
$$\Rightarrow \mathbb{E}_{\mathbf{h}}\left[\frac{\partial E(\tilde{\mathbf{x}},\mathbf{h})}{\partial\mathbf{W}}\middle|\tilde{\mathbf{x}}\right] = -\mathbf{h}(\tilde{\mathbf{x}})\tilde{\mathbf{x}}^T \quad \text{where} \quad \mathbf{h}(\mathbf{x}) \stackrel{\text{def}}{=} \begin{pmatrix} p(h_1=1|\mathbf{x}) \\ \cdots \\ p(h_H=1|\mathbf{x}) \end{pmatrix}$$

$$= \text{sigm}(\mathbf{b}+\mathbf{W}\mathbf{x})$$

RBM Training

The missing $\tilde{\mathbf{x}}$ can be generated by Gibbs sampling.

Start sampling a chain at $\mathbf{x}(t)$
Do k steps and generate $\tilde{\mathbf{x}}$.



CD-k (Contrastive Divergence) Pseudocode

I. For each training example $\mathbf{x}(t)$, generate a fake sample $\tilde{\mathbf{x}}$ using k steps of Gibbs sampling

II. Update the parameters

$$\begin{aligned}
\mathbf{W} &\Longleftarrow \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)})\, \mathbf{x}^{(t)\top} - \mathbf{h}(\tilde{\mathbf{x}})\, \tilde{\mathbf{x}}^{\top} \right) \\
\mathbf{b} &\Longleftarrow \mathbf{b} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}}) \right) \\
\mathbf{c} &\Longleftarrow \mathbf{c} + \alpha \left(\mathbf{x}^{(t)} - \tilde{\mathbf{x}} \right)
\end{aligned}$$

III. Go back to I until convergence.

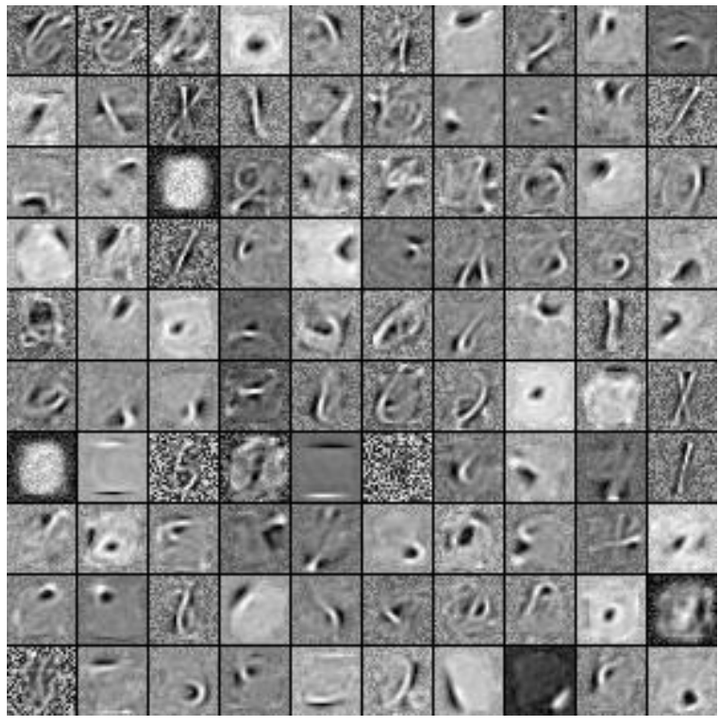
The bigger the k the better it is, but in practice $k = 1$ works well.

Results

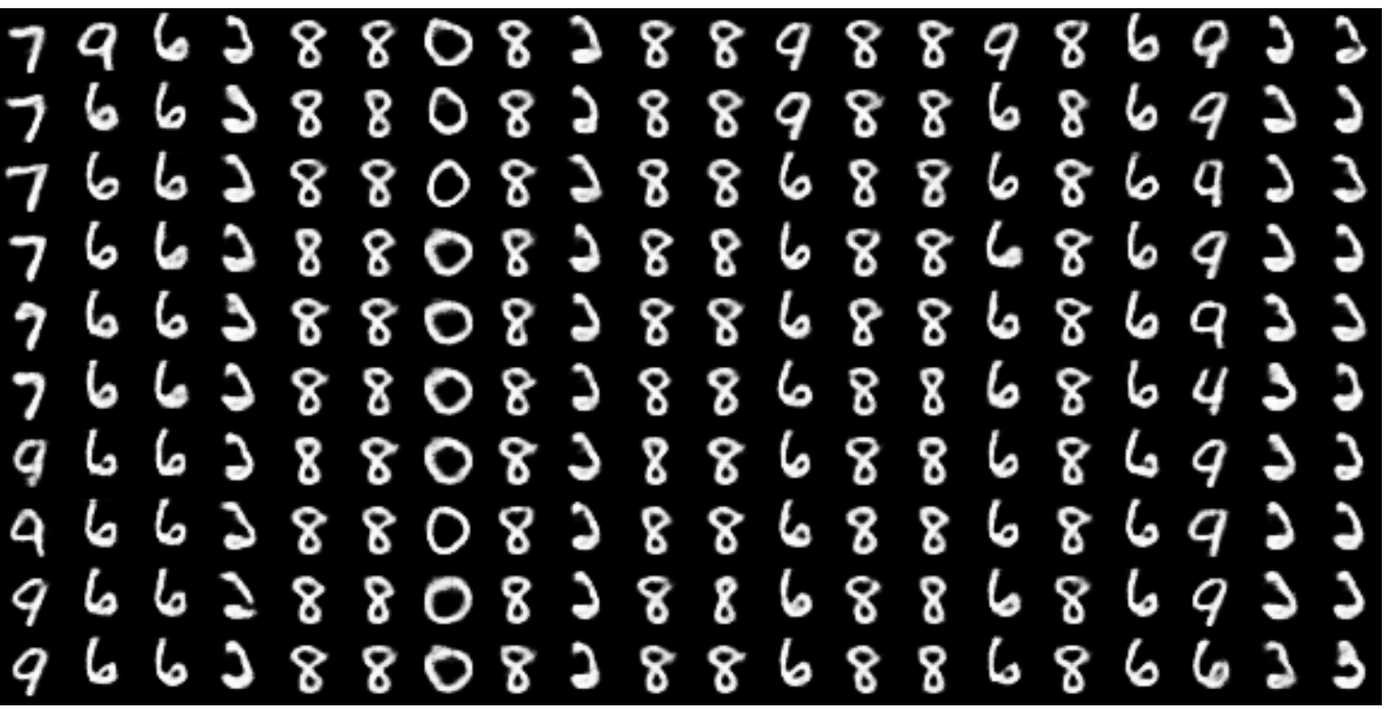
RBM Training Results

http://deeplearning.net/tutorial/rbm.html

Original images



Learned filters



Samples generated by the RBM after training.

Each row represents a mini-batch of negative particles (samples from independent Gibbs chains). 1000 steps of Gibbs sampling were taken between each of those rows.

Summary

Tasks:

RBM model: $p_\theta(\mathbf{x}, \mathbf{h})$

- Inference:

$$p_\theta(\mathbf{h}|\mathbf{x}) = ?$$

- Evaluate the likelihood function:

$$p_\theta(\mathbf{x}) = ?$$

- Sampling from RBM:

$$\tilde{\mathbf{x}}, \tilde{\mathbf{h}} \sim p_\theta(\mathbf{x}, \mathbf{h})$$

- Training RBM:

$$\max_\theta p_\theta(\mathbf{x}) = ?$$

Thanks for your Attention!

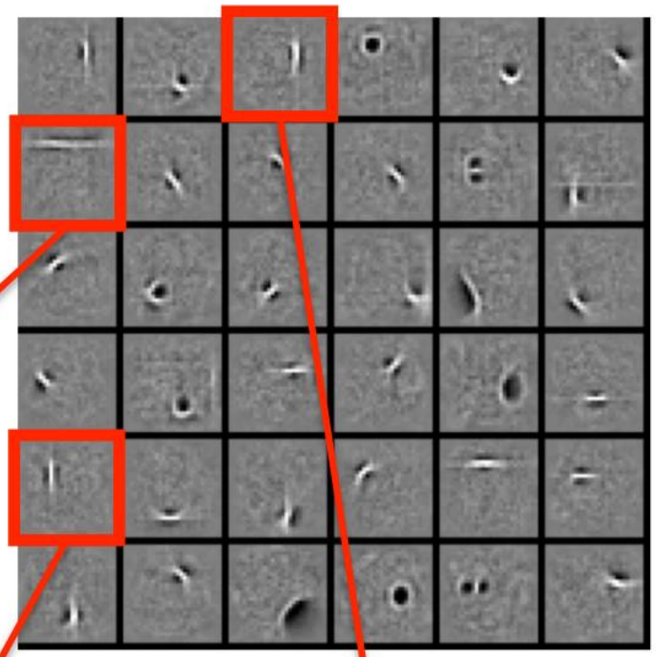
RBM Training Results

Observed Data

Subset of 25,000 characters

Learned W: “edges”

Subset of 1000 features



Most hidden variables are off

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

$$= \sigma\left(0.99 \times \quad + \quad 0.97 \times \quad + \quad 0.82 \times \quad \ldots.\right)$$

$$\sigma(x) = \frac{1}{1+\exp(-x)} \qquad p(v_k = 1|\mathbf{h}) = \sigma(c_k + \mathbf{h}^T \mathbf{W}_{.k})$$

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

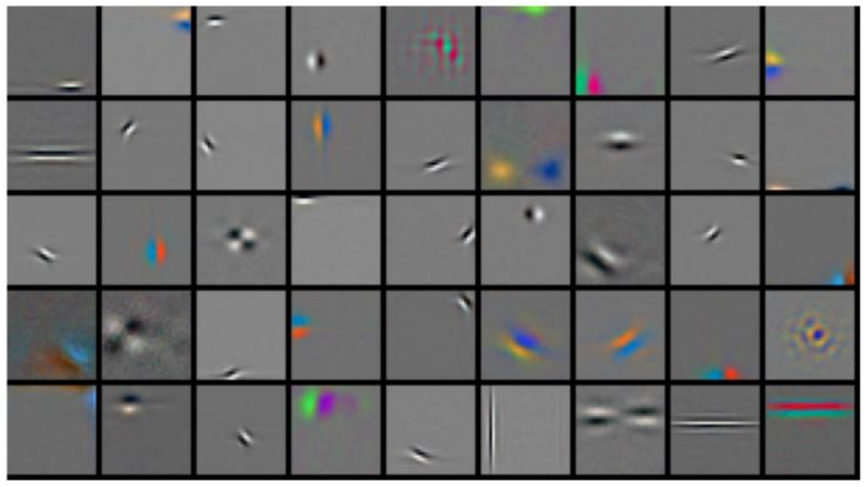
$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \mathbf{W}_j.\mathbf{v})$$

Gaussian-Bernoulli RBM Training Results

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	clinton	computer	trade	stock
russia	house	system	country	wall
moscow	president	product	import	street
yeltsin	bill	software	world	point
soviet	congress	develop	economy	dow

Each document (story) is represented with a bag of world coming from a multinomial distribution with parameters (h = topics). After training we can generate words from this topics.