

Neural networks

Restricted Boltzmann machine - definition

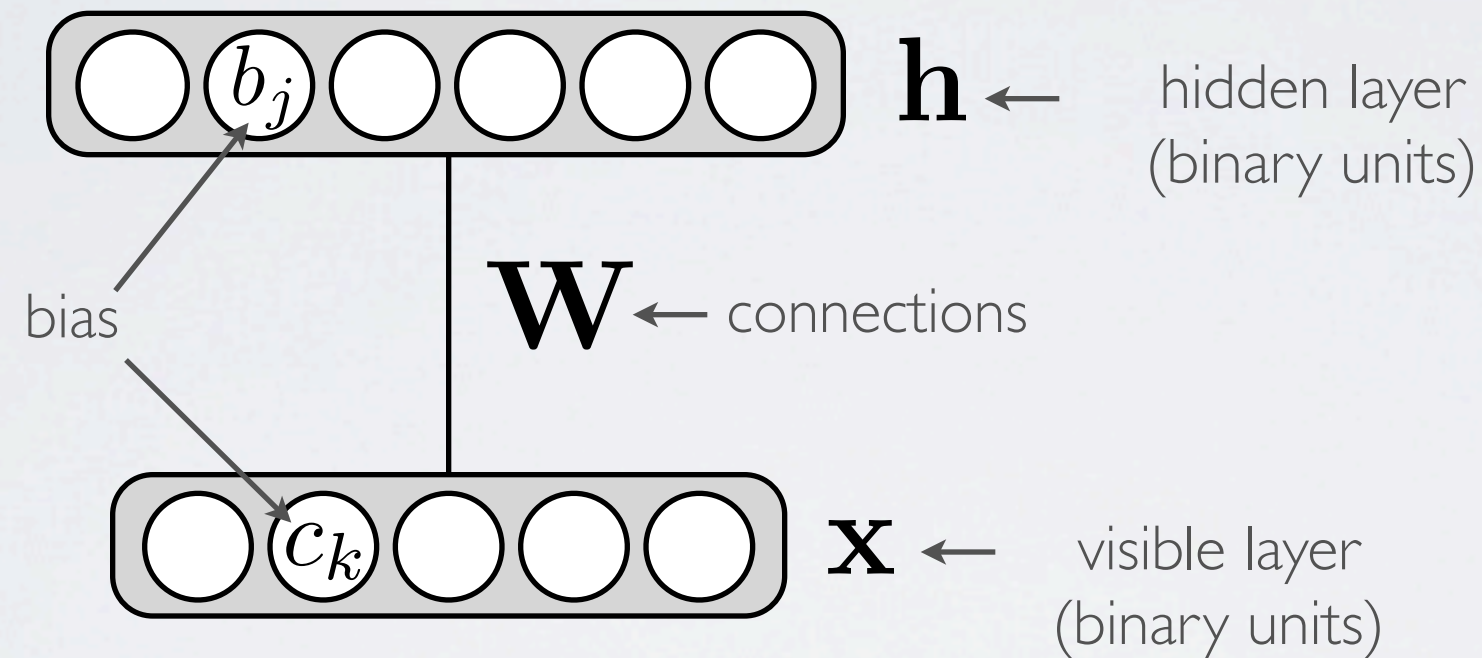
UNSUPERVISED LEARNING

Topics: unsupervised learning

- Unsupervised learning: only use the inputs $\mathbf{x}^{(t)}$ for learning
 - automatically extract meaningful features for your data
 - leverage the availability of unlabeled data
 - add a data-dependent regularizer to training ($-\log p(\mathbf{x}^{(t)})$)
- We will see 3 neural networks for unsupervised learning
 - **restricted Boltzmann machines**
 - autoencoders
 - sparse coding model

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h}$$

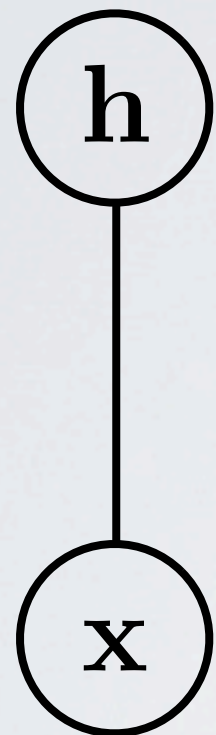
$$= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$$

Distribution: $p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h})) / Z$

partition function
(intractable)

MARKOV NETWORK VIEW

Topics: Markov network (with vector nodes)

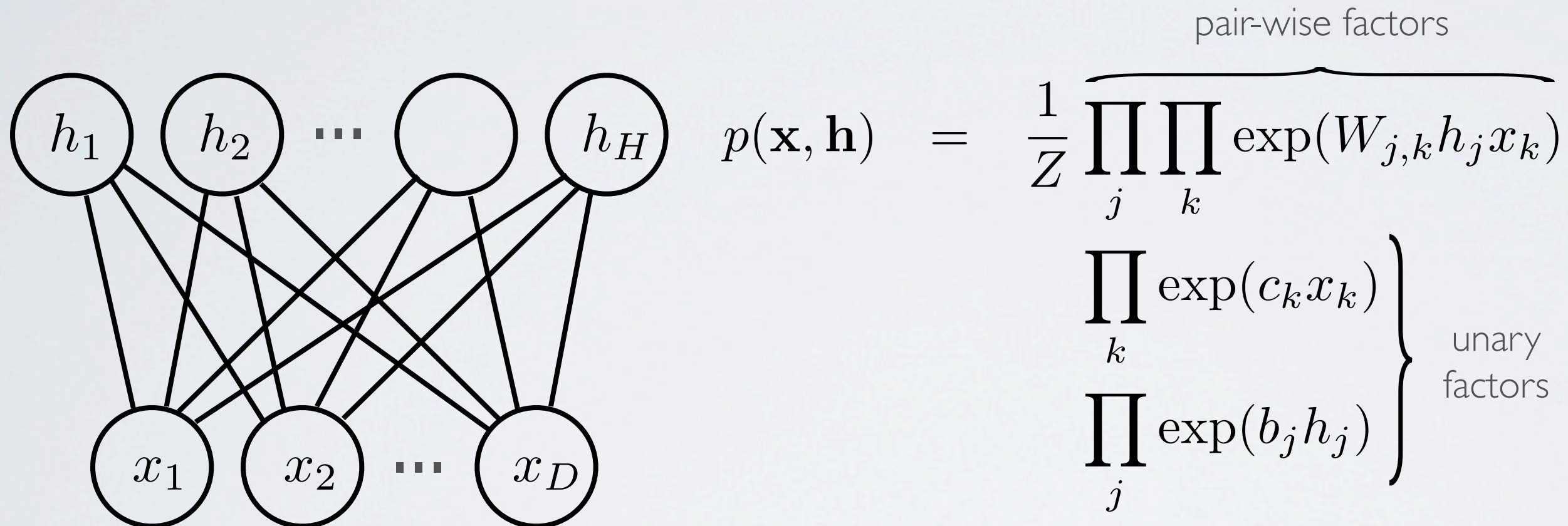


$$\begin{aligned}
 p(\mathbf{x}, \mathbf{h}) &= \exp(-E(\mathbf{x}, \mathbf{h}))/Z \\
 &= \exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h})/Z \\
 &= \underbrace{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x}) \exp(\mathbf{c}^\top \mathbf{x}) \exp(\mathbf{b}^\top \mathbf{h})}_{\text{factors}}/Z
 \end{aligned}$$

- The notation based on an energy function is simply an alternative to the representation as the product of factors

MARKOV NETWORK VIEW

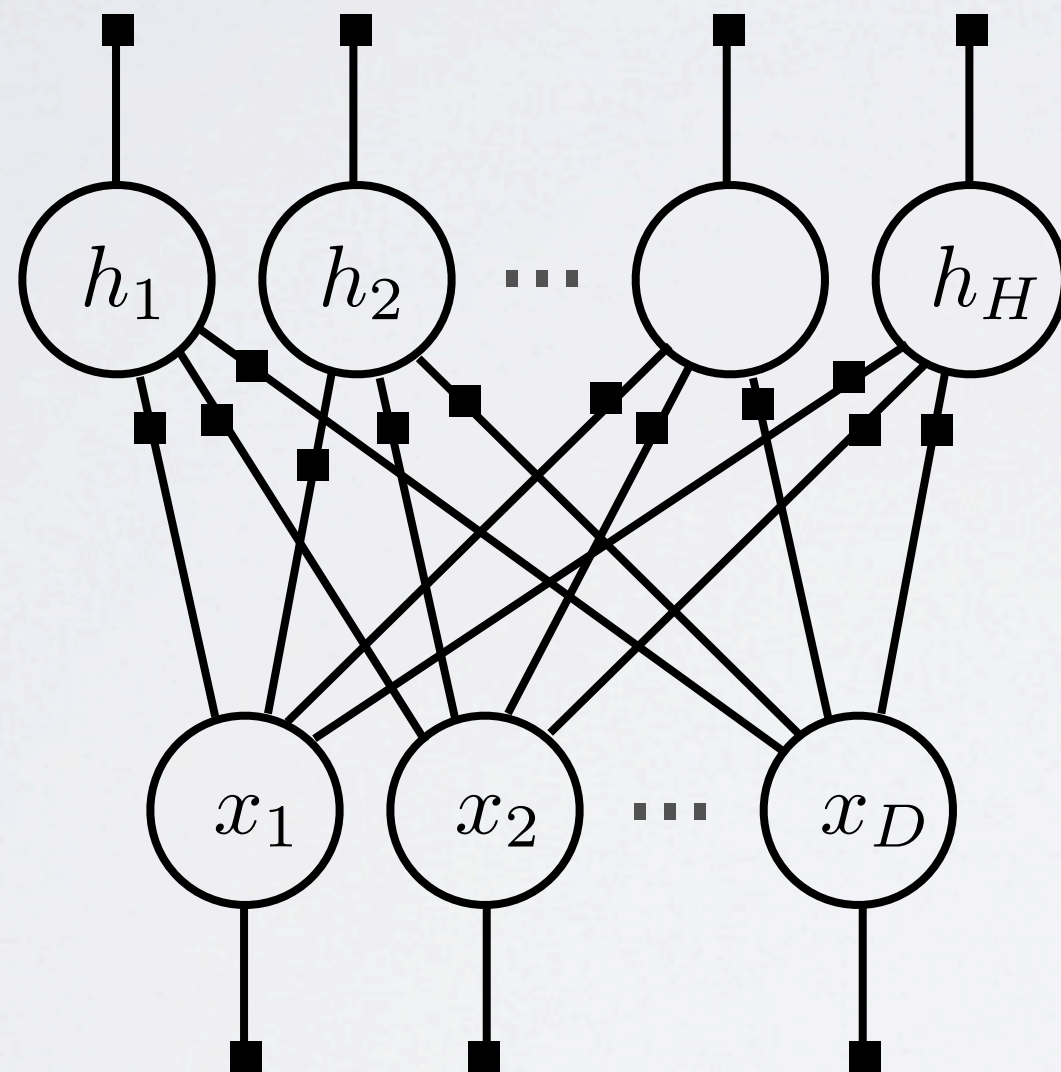
Topics: Markov network (with scalar nodes)



- The scalar visualization is more informative of the structure within the vectors

FACTOR GRAPH VIEW

Topics: factor graph of an RBM

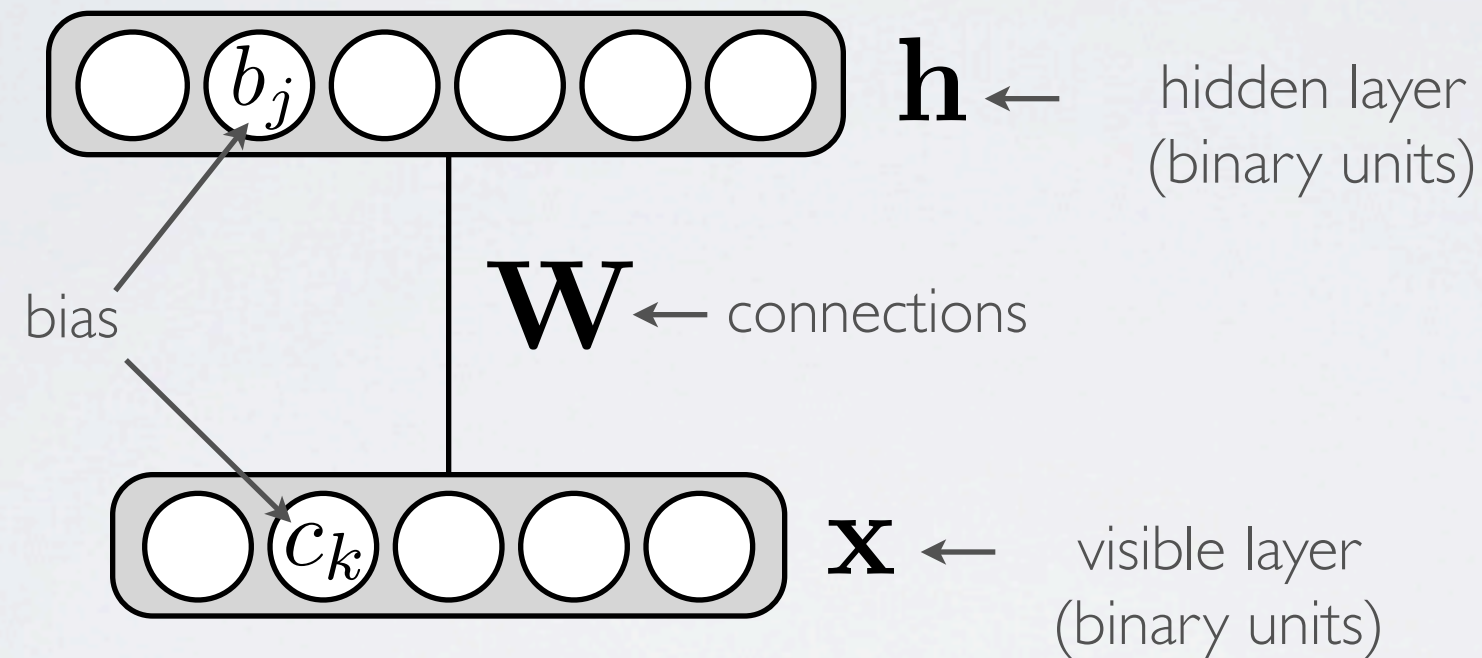


Neural networks

Restricted Boltzmann machine - inference

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h}$$

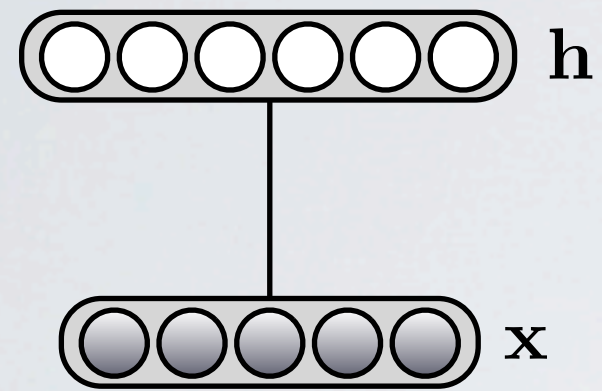
$$= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$$

Distribution: $p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h})) / Z$

← partition function (intractable)

INFERENCE

Topics: conditional distributions

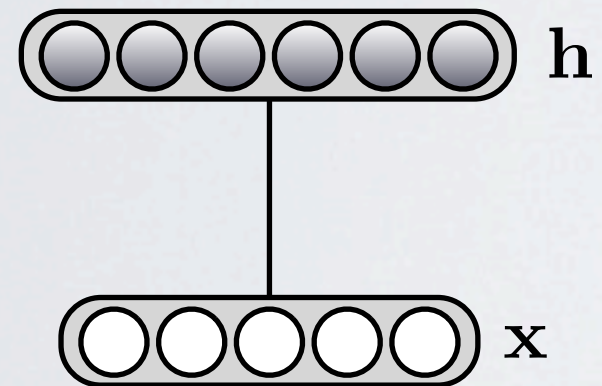


$$p(\mathbf{h}|\mathbf{x}) = \prod_j p(h_j|\mathbf{x})$$

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

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

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



$$p(\mathbf{x}|\mathbf{h}) = \prod_k p(x_k|\mathbf{h})$$

$$p(x_k = 1|\mathbf{h}) = \frac{1}{1 + \exp(-(c_k + \mathbf{h}^\top \mathbf{W}_{\cdot k}))}$$

$$= \text{sigm}(c_k + \mathbf{h}^\top \mathbf{W}_{\cdot k})$$

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

$$p(\mathbf{h}|\mathbf{x})$$

$$p(\mathbf{h}|\mathbf{x}) = p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}')$$

$$\begin{aligned}
 p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
 &= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}') / Z}
 \end{aligned}$$

4

$$\begin{aligned}
 p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
 &= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}') / Z}
 \end{aligned}$$

4

$$\begin{aligned}
 p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
 &= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / \cancel{Z}}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / \cancel{Z}}
 \end{aligned}$$

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / \cancel{Z}}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / \cancel{Z}} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / \cancel{Z}}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / \cancel{Z}} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / \cancel{Z}}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / \cancel{Z}} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h'_1) \right) \cdots \left(\sum_{h'_H \in \{0,1\}} \exp(h'_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h'_H) \right)}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / \cancel{Z}}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / \cancel{Z}} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h'_1) \right) \cdots \left(\sum_{h'_H \in \{0,1\}} \exp(h'_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h'_H) \right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h'_j \in \{0,1\}} \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j) \right)}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / \cancel{Z}}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / \cancel{Z}} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h'_1) \right) \cdots \left(\sum_{h'_H \in \{0,1\}} \exp(h'_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h'_H) \right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h'_j \in \{0,1\}} \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j) \right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\prod_j (1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x}))}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / Z}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / Z} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h'_1) \right) \cdots \left(\sum_{h'_H \in \{0,1\}} \exp(h'_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h'_H) \right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h'_j \in \{0,1\}} \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j) \right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\prod_j (1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x}))} \\
&= \prod_j \frac{\exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x})}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{h}|\mathbf{x}) &= p(\mathbf{x}, \mathbf{h}) / \sum_{\mathbf{h}'} p(\mathbf{x}, \mathbf{h}') \\
&= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}) / \cancel{Z}}{\sum_{\mathbf{h}' \in \{0,1\}^H} \exp(\mathbf{h}'^\top \mathbf{W} \mathbf{x} + \cancel{\mathbf{c}^\top \mathbf{x}} + \mathbf{b}^\top \mathbf{h}') / \cancel{Z}} \\
&= \frac{\exp(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\left(\sum_{h'_1 \in \{0,1\}} \exp(h'_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h'_1) \right) \cdots \left(\sum_{h'_H \in \{0,1\}} \exp(h'_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h'_H) \right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\prod_j \left(\sum_{h'_j \in \{0,1\}} \exp(h'_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h'_j) \right)} \\
&= \frac{\prod_j \exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{\prod_j (1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x}))} \\
&= \prod_j \frac{\exp(h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j)}{1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x})} \\
&= \prod_j p(h_j | \mathbf{x})
\end{aligned}$$

$$p(h_j = 1|\mathbf{x})$$

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

$$\begin{aligned}
 p(h_j = 1|\mathbf{x}) &= \frac{\exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x})}{1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x})} \\
 &= \frac{1}{1 + \exp(-b_j - \mathbf{W}_{j \cdot} \mathbf{x})}
 \end{aligned}$$

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

LOCAL MARKOV PROPERTY

Topics: local Markov property

- In general, we have the following property:

$$\begin{aligned}
 p(z_i | z_1, \dots, z_V) &= p(z_i | \text{Ne}(z_i)) \\
 &= \frac{p(z_i, \text{Ne}(z_i))}{\sum_{z'_i} p(z'_i, \text{Ne}(z_i))} \\
 &= \frac{\prod_{\substack{f \text{ involving } z_i \\ \text{and any } \text{Ne}(z_i)}} \Psi_f(z_i, \text{Ne}(z_i))}{\sum_{z'_i} \prod_{\substack{f \text{ involving } z_i \\ \text{and any } \text{Ne}(z_i)}} \Psi_f(z'_i, \text{Ne}(z_i))}
 \end{aligned}$$

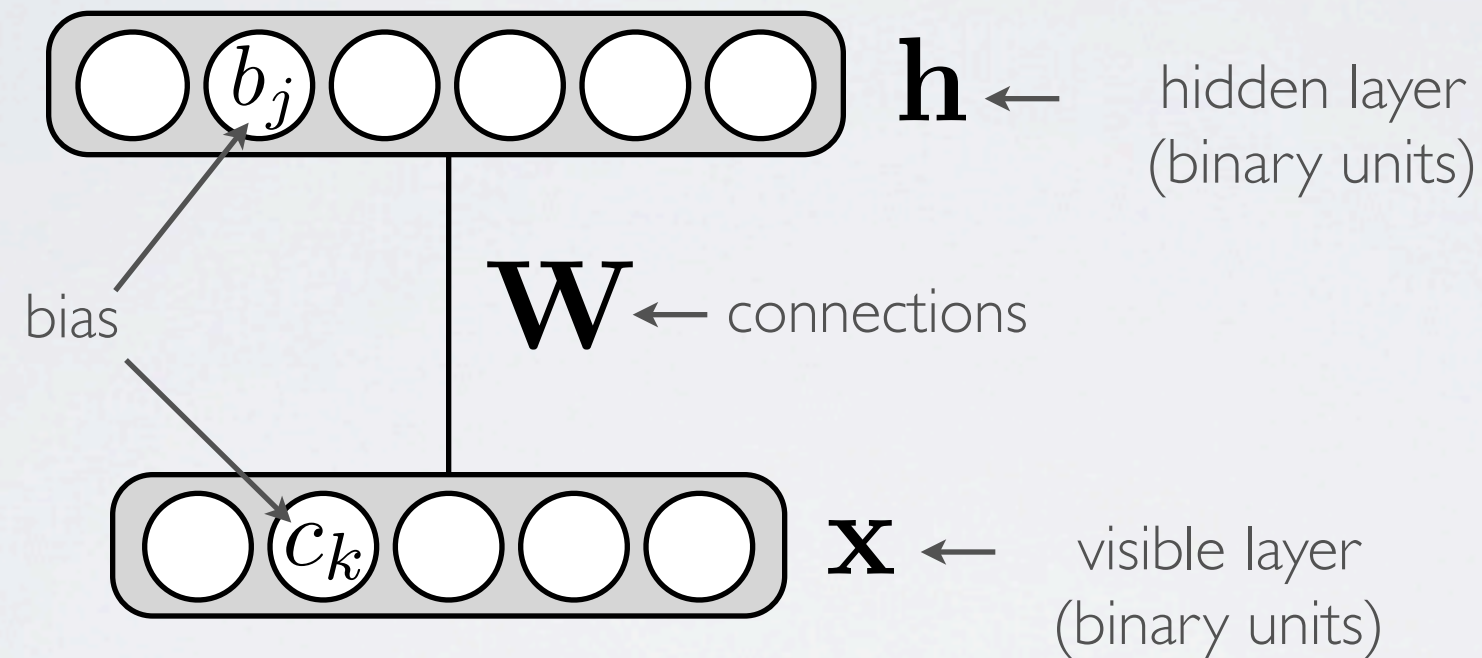
- ▶ z_i is any variable in the Markov network (x_k or h_j in an RBM)
- ▶ $\text{Ne}(z_i)$ are the neighbors of z_i in the Markov network

Neural networks

Restricted Boltzmann machine - free energy

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h}$$

$$= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$$

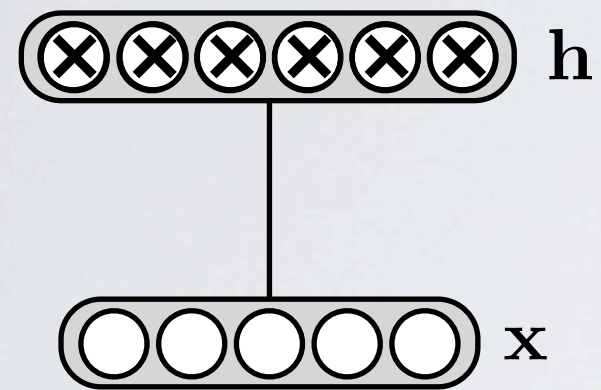
Distribution: $p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h})) / Z$

← partition function
(intractable)

FREE ENERGY

Topics: free energy

- What about $p(\mathbf{x})$?



$$\begin{aligned}
 p(\mathbf{x}) &= \sum_{\mathbf{h} \in \{0,1\}^H} p(\mathbf{x}, \mathbf{h}) = \sum_{\mathbf{h} \in \{0,1\}^H} \exp(-E(\mathbf{x}, \mathbf{h})) / Z \\
 &= \exp \left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \log(1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x})) \right) / Z \\
 &= \exp(-F(\mathbf{x})) / Z
 \end{aligned}$$

free energy

$p(\mathbf{x})$

$$p(\mathbf{x}) = \sum_{\mathbf{h} \in \{0,1\}^H} \exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z$$

$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{h} \in \{0,1\}^H} \exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp \left(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j \right) / Z
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{h} \in \{0,1\}^H} \exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp \left(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j \right) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \left(\sum_{h_1 \in \{0,1\}} \exp(h_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h_1) \right) \cdots \left(\sum_{h_H \in \{0,1\}} \exp(h_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h_H) \right) / Z
\end{aligned}$$

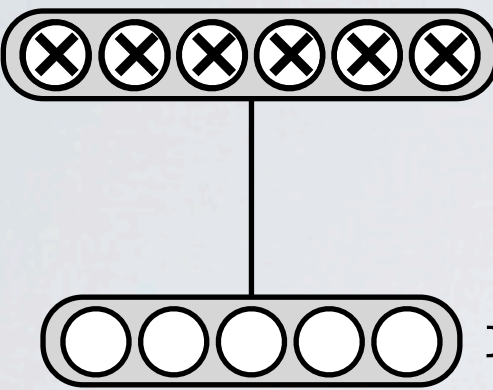
$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{h} \in \{0,1\}^H} \exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp \left(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j \right) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \left(\sum_{h_1 \in \{0,1\}} \exp(h_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h_1) \right) \cdots \left(\sum_{h_H \in \{0,1\}} \exp(h_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h_H) \right) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) (1 + \exp(b_1 + \mathbf{W}_{1 \cdot} \mathbf{x})) \cdots (1 + \exp(b_H + \mathbf{W}_{H \cdot} \mathbf{x})) / Z
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{h} \in \{0,1\}^H} \exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp \left(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j \right) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \left(\sum_{h_1 \in \{0,1\}} \exp(h_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h_1) \right) \cdots \left(\sum_{h_H \in \{0,1\}} \exp(h_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h_H) \right) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) (1 + \exp(b_1 + \mathbf{W}_{1 \cdot} \mathbf{x})) \cdots (1 + \exp(b_H + \mathbf{W}_{H \cdot} \mathbf{x})) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \exp(\log(1 + \exp(b_1 + \mathbf{W}_{1 \cdot} \mathbf{x}))) \cdots \exp(\log(1 + \exp(b_H + \mathbf{W}_{H \cdot} \mathbf{x}))) / Z
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{h} \in \{0,1\}^H} \exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h}) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp \left(\sum_j h_j \mathbf{W}_{j \cdot} \mathbf{x} + b_j h_j \right) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \left(\sum_{h_1 \in \{0,1\}} \exp(h_1 \mathbf{W}_{1 \cdot} \mathbf{x} + b_1 h_1) \right) \cdots \left(\sum_{h_H \in \{0,1\}} \exp(h_H \mathbf{W}_{H \cdot} \mathbf{x} + b_H h_H) \right) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) (1 + \exp(b_1 + \mathbf{W}_{1 \cdot} \mathbf{x})) \cdots (1 + \exp(b_H + \mathbf{W}_{H \cdot} \mathbf{x})) / Z \\
&= \exp(\mathbf{c}^\top \mathbf{x}) \exp(\log(1 + \exp(b_1 + \mathbf{W}_{1 \cdot} \mathbf{x}))) \cdots \exp(\log(1 + \exp(b_H + \mathbf{W}_{H \cdot} \mathbf{x}))) / Z \\
&= \exp \left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \log(1 + \exp(b_j + \mathbf{W}_{j \cdot} \mathbf{x})) \right) / Z
\end{aligned}$$

RESTRICTED BOLTZMANN MACHINE

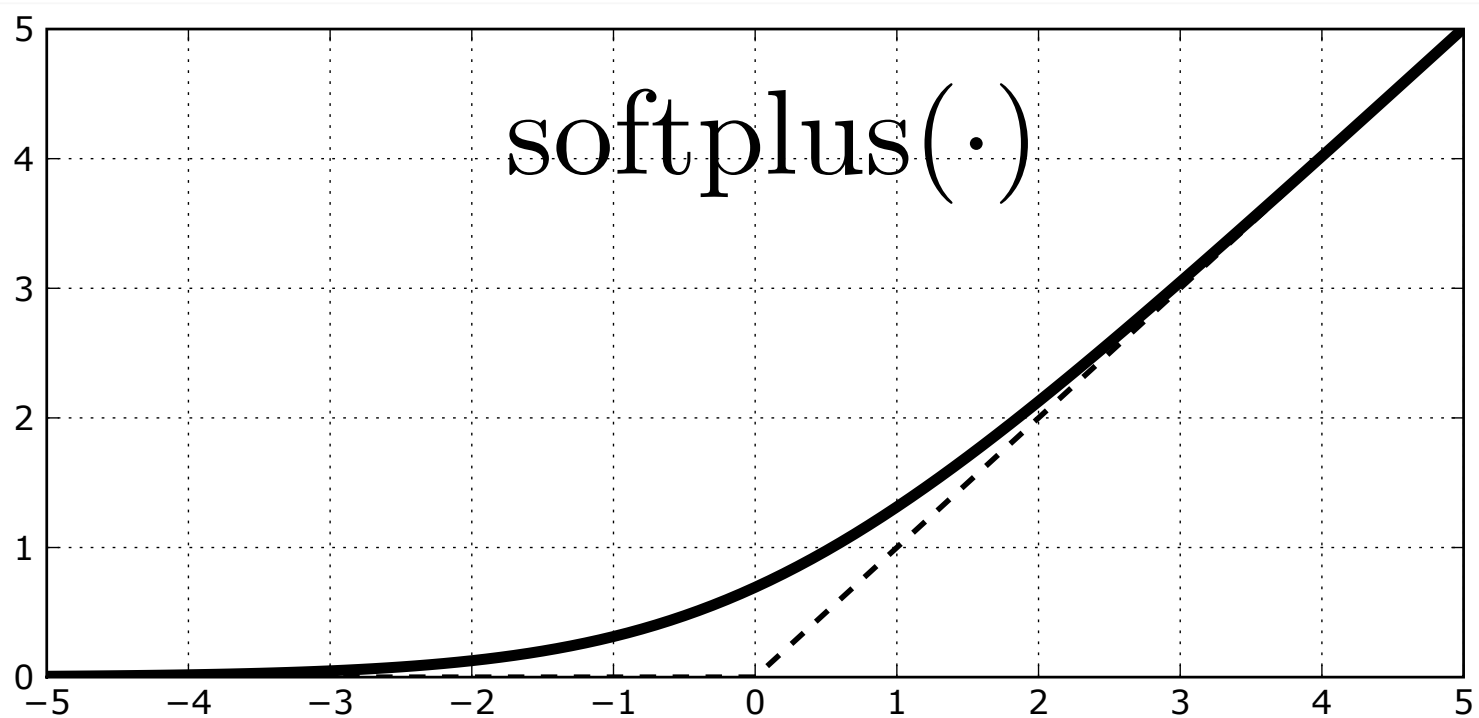
Topics: free energy



$$p(\mathbf{x}) = \frac{\exp\left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \log(1 + \exp(b_j + \mathbf{W}_j \cdot \mathbf{x}))\right)}{Z}$$

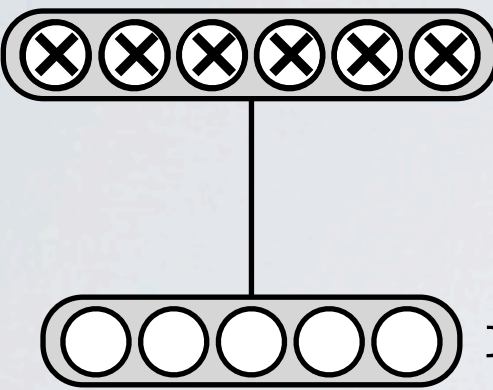
$$= \frac{\exp\left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \text{softplus}(b_j + \mathbf{W}_j \cdot \mathbf{x})\right)}{Z}$$

softplus(\cdot)



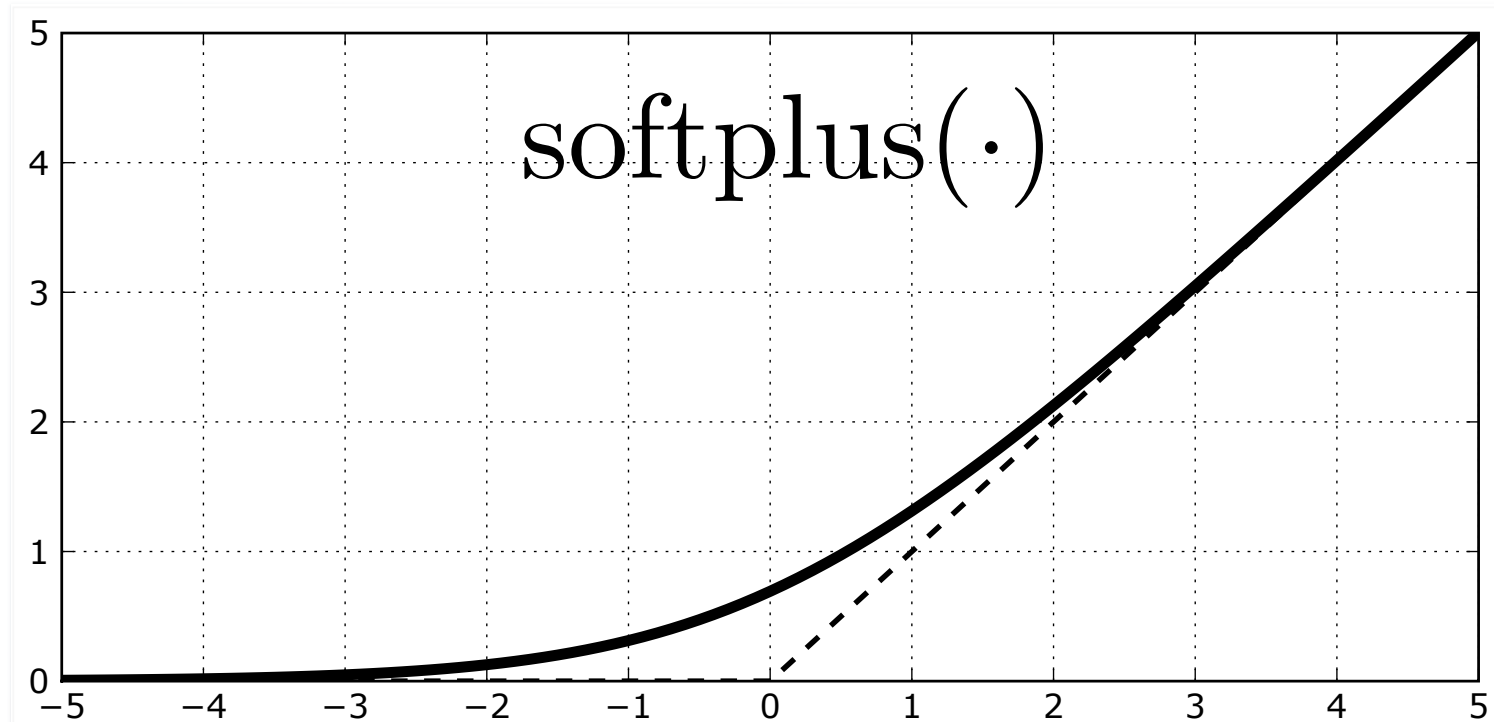
RESTRICTED BOLTZMANN MACHINE

Topics: free energy



$$p(\mathbf{x}) = \frac{\exp\left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \log(1 + \exp(b_j + \mathbf{W}_j \cdot \mathbf{x}))\right)}{Z}$$

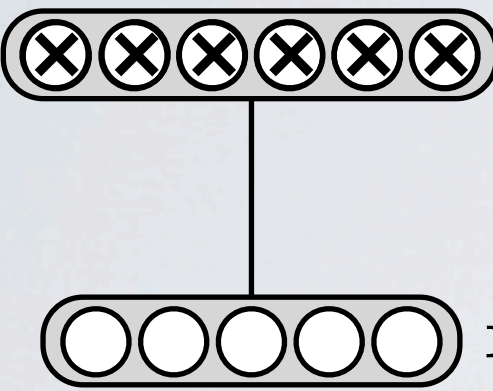
$$= \frac{\exp\left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \text{softplus}(b_j + \mathbf{W}_j \cdot \mathbf{x})\right)}{Z}$$



bias the prob of each x_i

RESTRICTED BOLTZMANN MACHINE

Topics: free energy



\mathbf{h}

\mathbf{x}

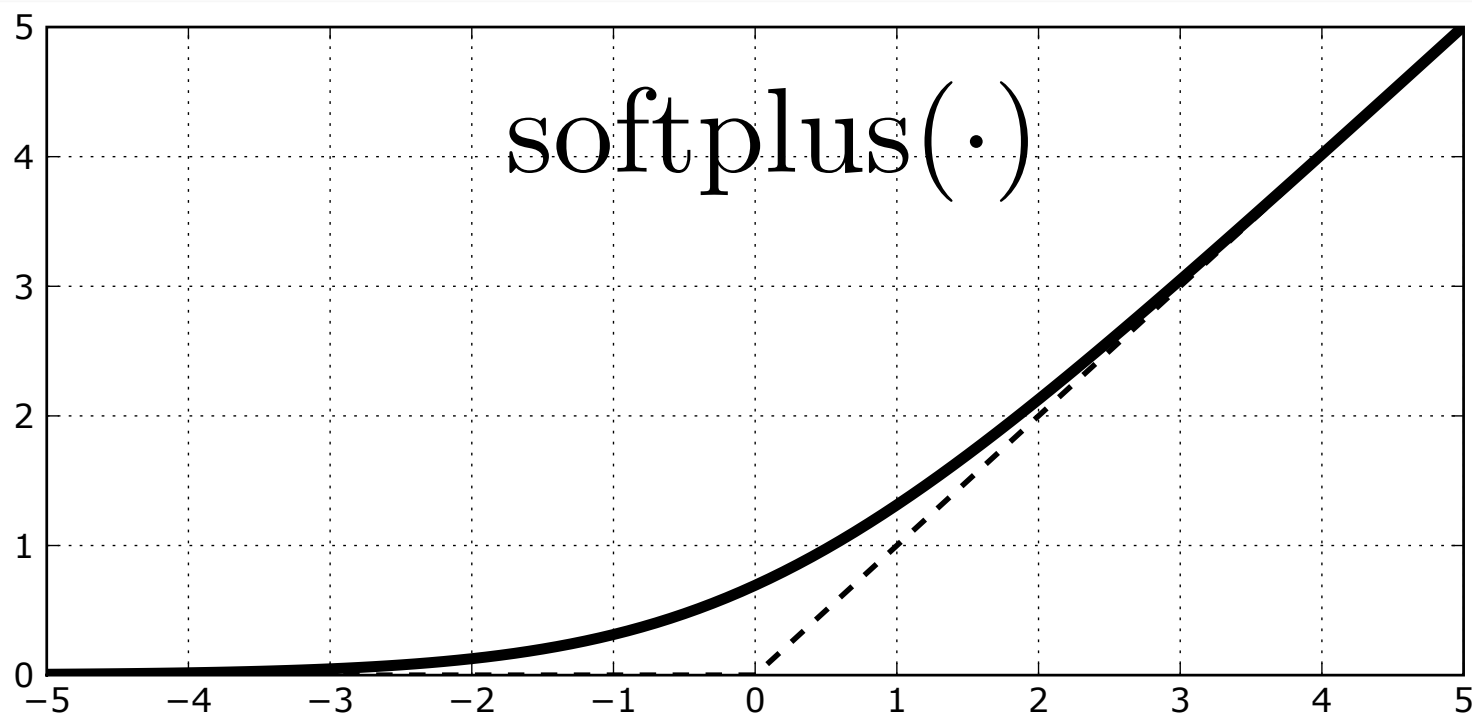
$$p(\mathbf{x}) = \exp \left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \log(1 + \exp(b_j + \mathbf{W}_j \cdot \mathbf{x})) \right) / Z$$

$$= \exp \left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \text{softplus}(b_j + \mathbf{W}_j \cdot \mathbf{x}) \right) / Z$$

“feature” expected in \mathbf{x}

bias the prob of each x_i

softplus(\cdot)



RESTRICTED BOLTZMANN MACHINE

Topics: free energy



\mathbf{h}

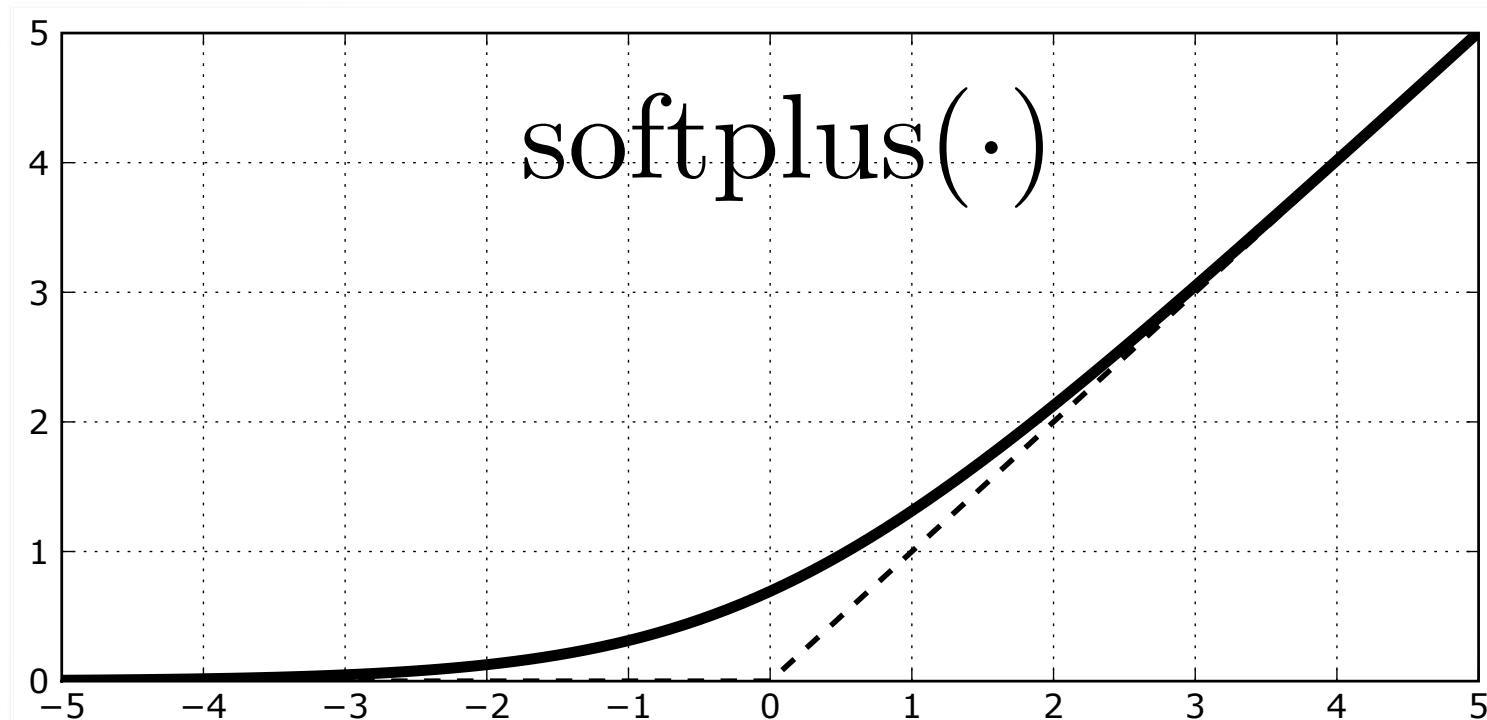
$p(\mathbf{x})$

$$= \exp \left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \log(1 + \exp(b_j + \mathbf{W}_j \cdot \mathbf{x})) \right) / Z$$



\mathbf{x}

$$= \exp \left(\mathbf{c}^\top \mathbf{x} + \sum_{j=1}^H \text{softplus}(b_j + \mathbf{W}_j \cdot \mathbf{x}) \right) / Z$$



“feature” expected in \mathbf{x}

bias of each feature

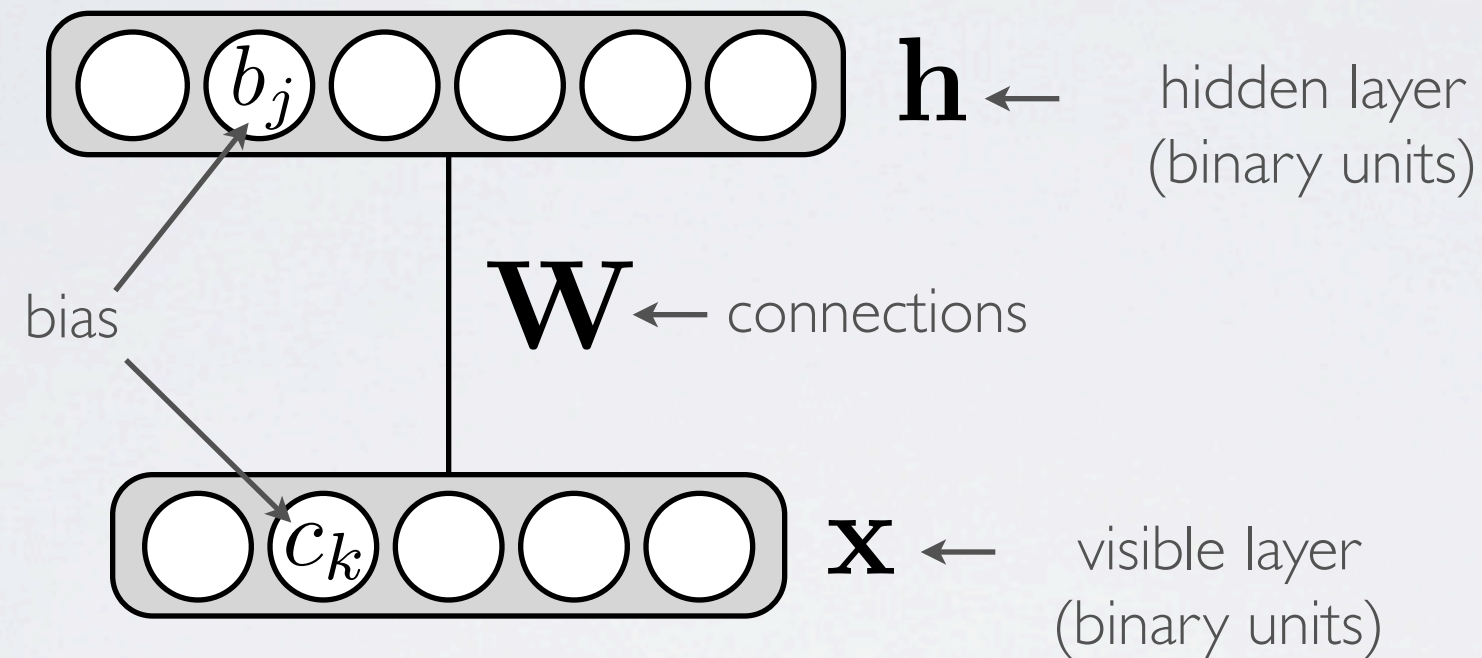
bias the prob of each x_i

Neural networks

Restricted Boltzmann machine - contrastive divergence

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h}$$

$$= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$$

Distribution: $p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h})) / Z$

← partition function
(intractable)

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T} \sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T} \sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \middle| \mathbf{x}^{(t)} \right]}_{\text{positive phase}} - \underbrace{\mathbb{E}_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right]}_{\text{negative phase}}$$

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T} \sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T} \sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \middle| \mathbf{x}^{(t)} \right]}_{\text{positive phase}} - \underbrace{\mathbb{E}_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right]}_{\text{negative phase}}$$

hard to
compute
↙

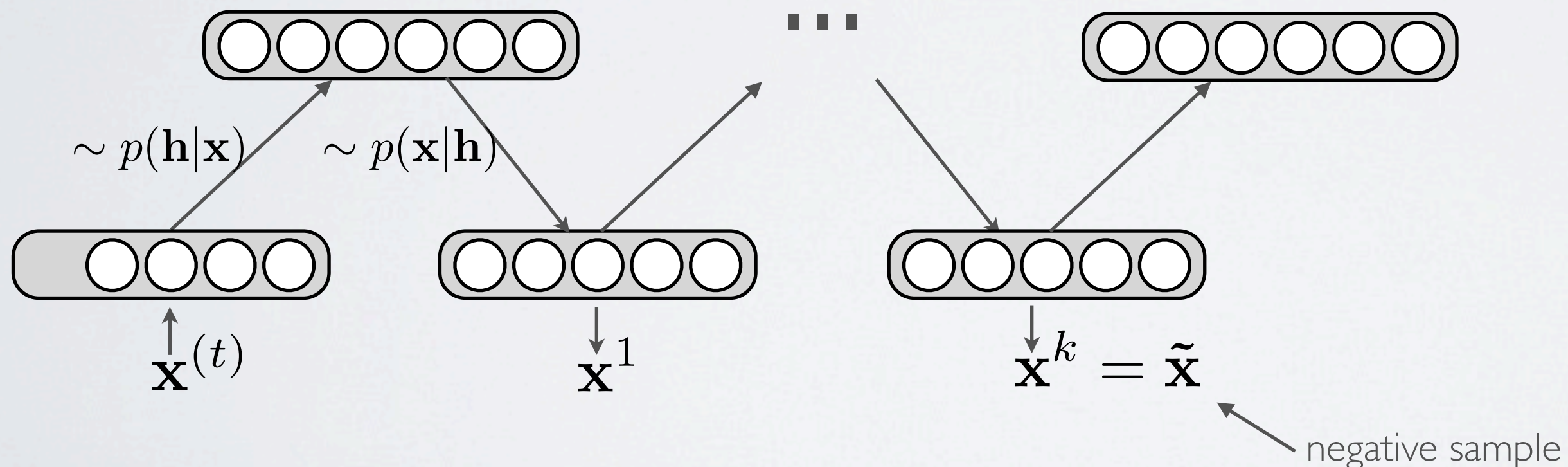
CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

• Idea:

1. replace the expectation by a point estimate at $\tilde{\mathbf{x}}$
2. obtain the point $\tilde{\mathbf{x}}$ by Gibbs sampling
3. start sampling chain at $\mathbf{x}^{(t)}$

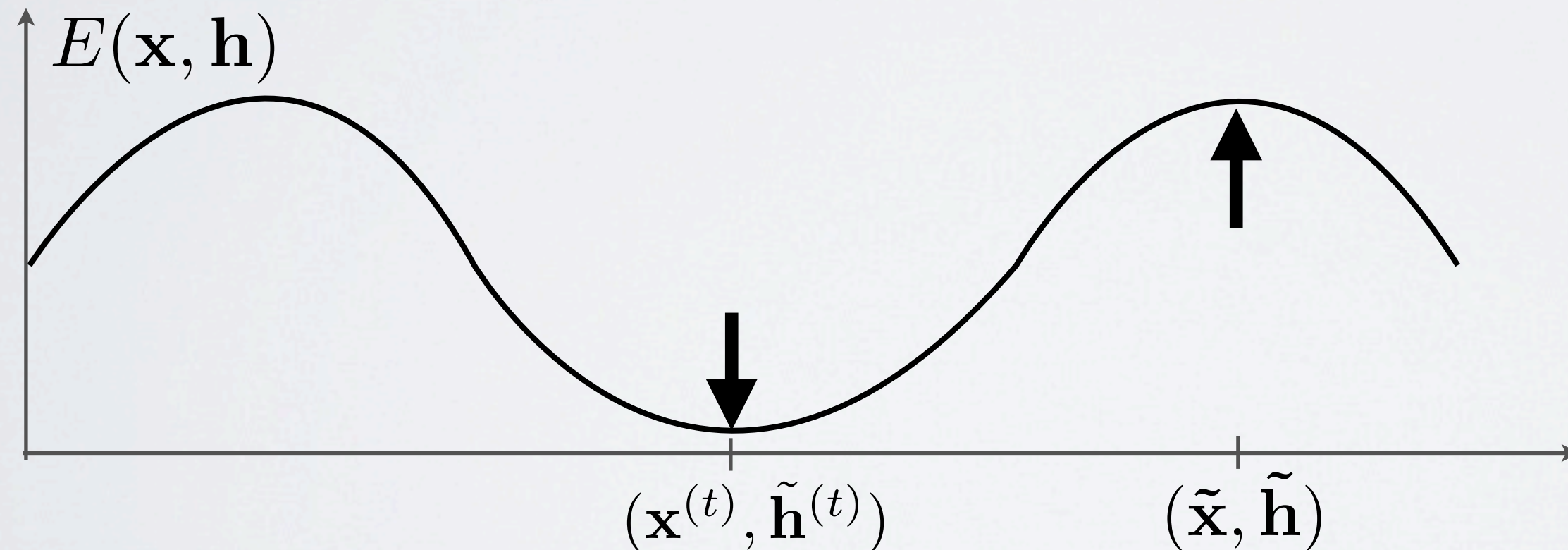


CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$\mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \mid \mathbf{x}^{(t)} \right] \approx \frac{\partial E(\mathbf{x}^{(t)}, \tilde{\mathbf{h}}^{(t)})}{\partial \theta} \quad \mathbb{E}_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right] \approx \frac{\partial E(\tilde{\mathbf{x}}, \tilde{\mathbf{h}})}{\partial \theta}$$

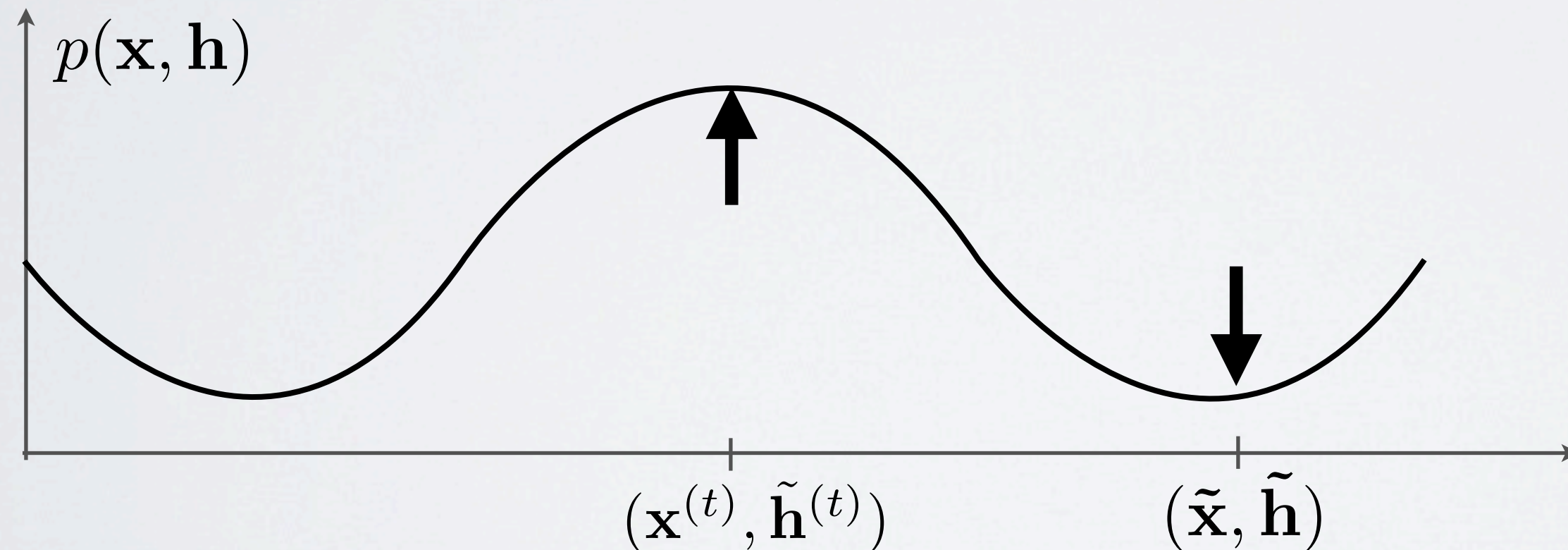


CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$\mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \mid \mathbf{x}^{(t)} \right] \approx \frac{\partial E(\mathbf{x}^{(t)}, \tilde{\mathbf{h}}^{(t)})}{\partial \theta} \quad \mathbb{E}_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right] \approx \frac{\partial E(\tilde{\mathbf{x}}, \tilde{\mathbf{h}})}{\partial \theta}$$

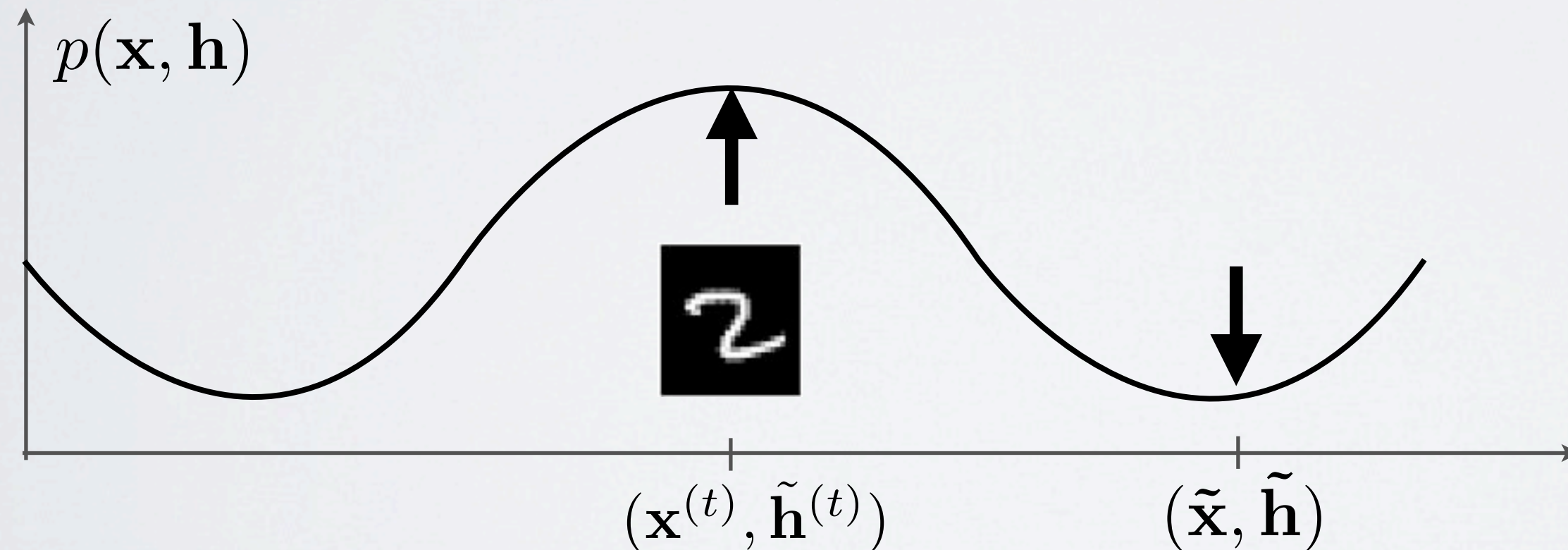


CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$E_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \mid \mathbf{x}^{(t)} \right] \approx \frac{\partial E(\mathbf{x}^{(t)}, \tilde{\mathbf{h}}^{(t)})}{\partial \theta} \quad E_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right] \approx \frac{\partial E(\tilde{\mathbf{x}}, \tilde{\mathbf{h}})}{\partial \theta}$$

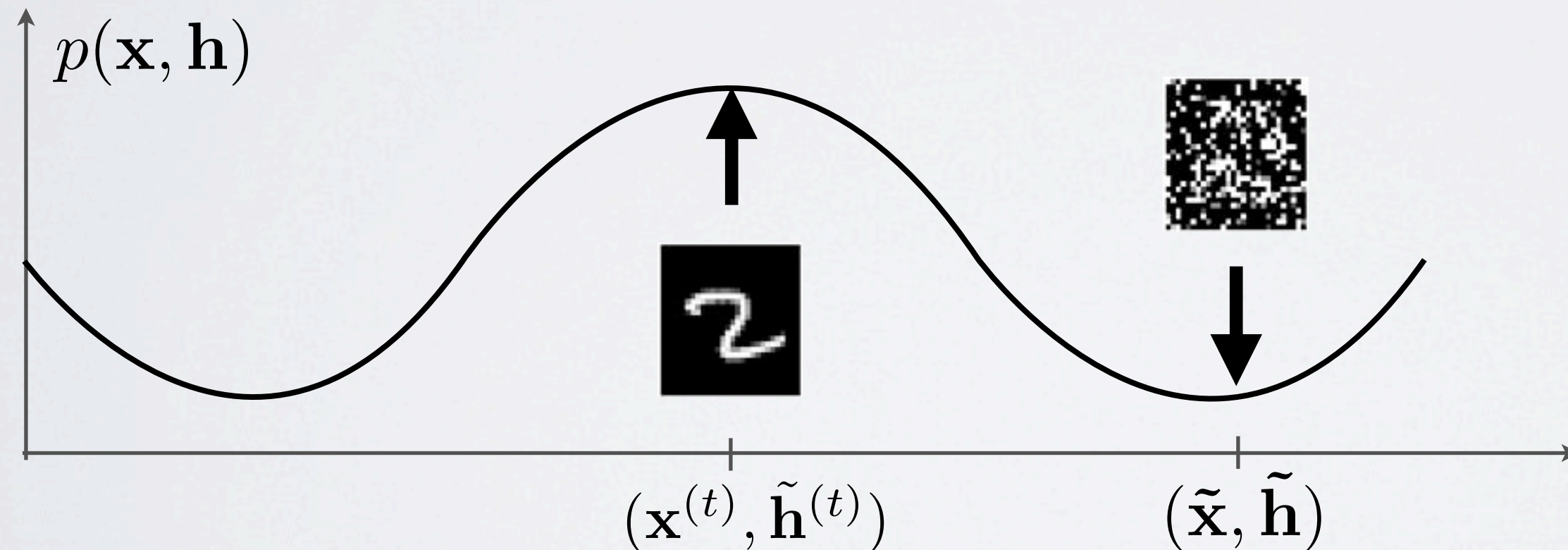


CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

$$E_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \mid \mathbf{x}^{(t)} \right] \approx \frac{\partial E(\mathbf{x}^{(t)}, \tilde{\mathbf{h}}^{(t)})}{\partial \theta} \quad E_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right] \approx \frac{\partial E(\tilde{\mathbf{x}}, \tilde{\mathbf{h}})}{\partial \theta}$$



Neural networks

Restricted Boltzmann machine - contrastive divergence (parameter update)

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T} \sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T} \sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \middle| \mathbf{x}^{(t)} \right]}_{\text{positive phase}} - \underbrace{\mathbb{E}_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right]}_{\text{negative phase}}$$

TRAINING

Topics: training objective

- To train an RBM, we'd like to minimize the average negative log-likelihood (NLL)

$$\frac{1}{T} \sum_t l(f(\mathbf{x}^{(t)})) = \frac{1}{T} \sum_t -\log p(\mathbf{x}^{(t)})$$

- We'd like to proceed by stochastic gradient descent

$$\frac{\partial -\log p(\mathbf{x}^{(t)})}{\partial \theta} = \underbrace{\mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}^{(t)}, \mathbf{h})}{\partial \theta} \middle| \mathbf{x}^{(t)} \right]}_{\text{positive phase}} - \underbrace{\mathbb{E}_{\mathbf{x}, \mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \right]}_{\text{negative phase}}$$

hard to
compute
↙

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Derivation of $\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta}$ for $\theta = W_{jk}$

$$\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial W_{jk}} = \frac{\partial}{\partial W_{jk}} \left(- \sum_{jk} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \right)$$

$$= - \frac{\partial}{\partial W_{jk}} \sum_{jk} W_{jk} h_j x_k$$

$$= -h_j x_k$$

$$\nabla_{\mathbf{W}} E(\mathbf{x}, \mathbf{h}) = -\mathbf{h} \mathbf{x}^\top$$

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Derivation of $\mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial \theta} \middle| \mathbf{x} \right]$ for $\theta = W_{jk}$

$$\begin{aligned} \mathbb{E}_{\mathbf{h}} \left[\frac{\partial E(\mathbf{x}, \mathbf{h})}{\partial W_{jk}} \middle| \mathbf{x} \right] &= \mathbb{E}_{\mathbf{h}} \left[-h_j x_k \middle| \mathbf{x} \right] = \sum_{h_j \in \{0,1\}} -h_j x_k p(h_j | \mathbf{x}) \\ &= -x_k p(h_j = 1 | \mathbf{x}) \end{aligned}$$

$$\mathbb{E}_{\mathbf{h}} [\nabla_{\mathbf{W}} E(\mathbf{x}, \mathbf{h}) | \mathbf{x}] = -\mathbf{h}(\mathbf{x}) \mathbf{x}^\top$$

$$\begin{aligned} \mathbf{h}(\mathbf{x}) &\stackrel{\text{def}}{=} \begin{pmatrix} p(h_1=1|\mathbf{x}) \\ \vdots \\ p(h_H=1|\mathbf{x}) \end{pmatrix} \\ &= \text{sigm}(\mathbf{b} + \mathbf{W}\mathbf{x}) \end{aligned}$$

DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

- Given $\mathbf{x}^{(t)}$ and $\tilde{\mathbf{x}}$ the learning rule for $\theta = \mathbf{W}$ becomes

$$\begin{aligned}
 \mathbf{W} &\Leftarrow \mathbf{W} - \alpha \left(\nabla_{\mathbf{W}} - \log p(\mathbf{x}^{(t)}) \right) \\
 &\Leftarrow \mathbf{W} - \alpha \left(\mathbb{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \mid \mathbf{x}^{(t)} \right] - \mathbb{E}_{\mathbf{x}, \mathbf{h}} [\nabla_{\mathbf{W}} E(\mathbf{x}, \mathbf{h})] \right) \\
 &\Leftarrow \mathbf{W} - \alpha \left(\mathbb{E}_{\mathbf{h}} \left[\nabla_{\mathbf{W}} E(\mathbf{x}^{(t)}, \mathbf{h}) \mid \mathbf{x}^{(t)} \right] - \mathbb{E}_{\mathbf{h}} [\nabla_{\mathbf{W}} E(\tilde{\mathbf{x}}, \mathbf{h}) \mid \tilde{\mathbf{x}}] \right) \\
 &\Leftarrow \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) \mathbf{x}^{(t)\top} - \mathbf{h}(\tilde{\mathbf{x}}) \tilde{\mathbf{x}}^\top \right)
 \end{aligned}$$

CD-K: PSEUDOCODE

Topics: contrastive divergence

- I. For each training example $\mathbf{x}^{(t)}$
 - i. generate a negative sample $\tilde{\mathbf{x}}$ using k steps of Gibbs sampling, starting at $\mathbf{x}^{(t)}$
 - ii. update parameters

$$\mathbf{W} \leftarrow \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) \mathbf{x}^{(t)\top} - \mathbf{h}(\tilde{\mathbf{x}}) \tilde{\mathbf{x}}^\top \right)$$

$$\mathbf{b} \leftarrow \mathbf{b} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}}) \right)$$

$$\mathbf{c} \leftarrow \mathbf{c} + \alpha \left(\mathbf{x}^{(t)} - \tilde{\mathbf{x}} \right)$$

2. Go back to I until stopping criteria

CONTRASTIVE DIVERGENCE (CD)

(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence

- CD-k: contrastive divergence with k iterations of Gibbs sampling
- In general, the bigger k is, the less **biased** the estimate of the gradient will be
- In practice, $k=1$ works well for pre-training

Neural networks

Restricted Boltzmann machine - persistent CD

CD-K: PSEUDOCODE

Topics: contrastive divergence

- I. For each training example $\mathbf{x}^{(t)}$
 - i. generate a negative sample $\tilde{\mathbf{x}}$ using k steps of Gibbs sampling, starting at $\mathbf{x}^{(t)}$
 - ii. update parameters

$$\mathbf{W} \leftarrow \mathbf{W} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) \mathbf{x}^{(t)\top} - \mathbf{h}(\tilde{\mathbf{x}}) \tilde{\mathbf{x}}^\top \right)$$

$$\mathbf{b} \leftarrow \mathbf{b} + \alpha \left(\mathbf{h}(\mathbf{x}^{(t)}) - \mathbf{h}(\tilde{\mathbf{x}}) \right)$$

$$\mathbf{c} \leftarrow \mathbf{c} + \alpha \left(\mathbf{x}^{(t)} - \tilde{\mathbf{x}} \right)$$

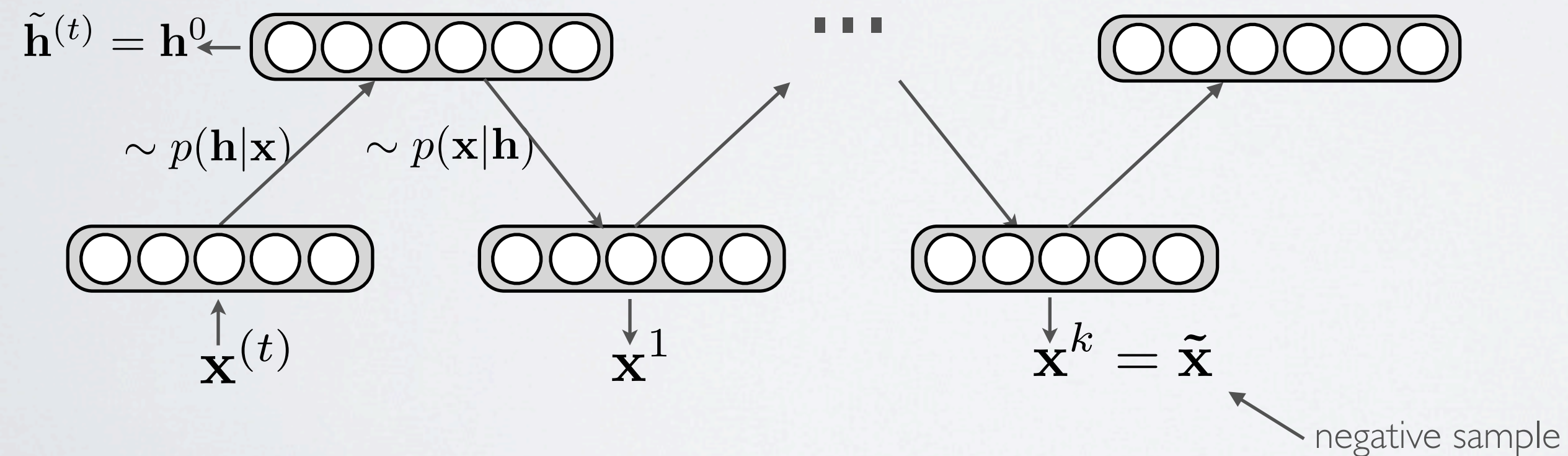
2. Go back to I until stopping criteria

PERSISTENT CD (PCD)

(TIELEMAN, ICML 2008)

Topics: persistent contrastive divergence

- Idea: instead of initializing the chain to $\mathbf{x}^{(t)}$, initialize the chain to the negative sample of the last iteration

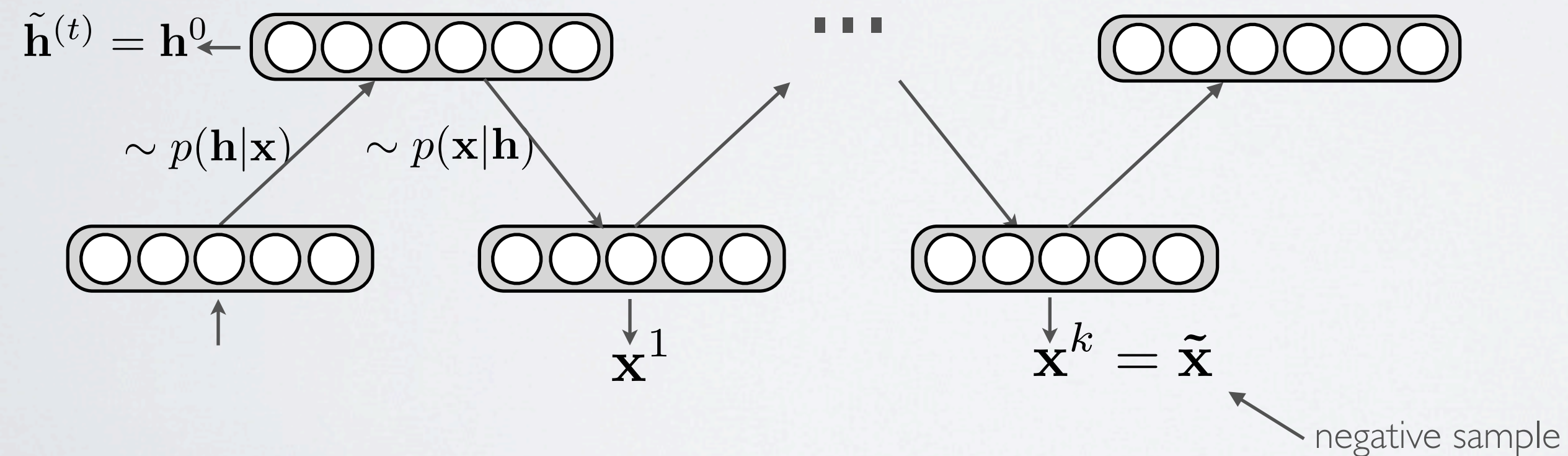


PERSISTENT CD (PCD)

(TIELEMAN, ICML 2008)

Topics: persistent contrastive divergence

- Idea: instead of initializing the chain to $\mathbf{x}^{(t)}$, initialize the chain to the negative sample of the last iteration

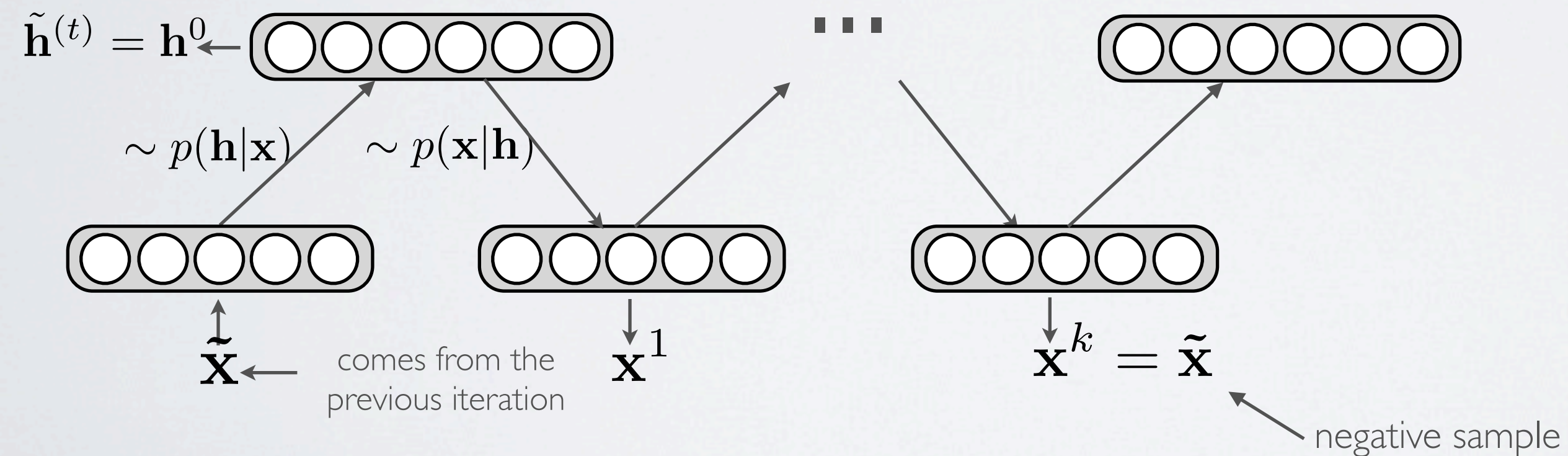


PERSISTENT CD (PCD)

(TIELEMAN, ICML 2008)

Topics: persistent contrastive divergence

- Idea: instead of initializing the chain to $\mathbf{x}^{(t)}$, initialize the chain to the negative sample of the last iteration

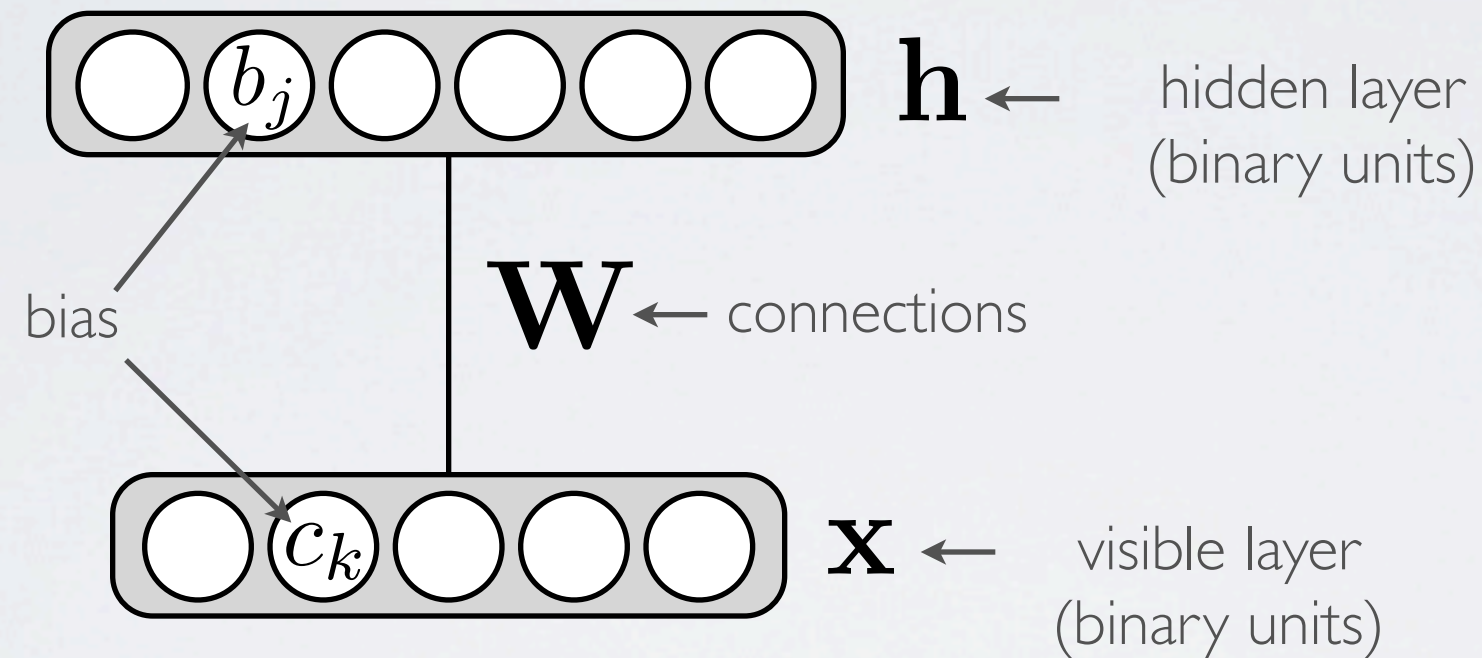


Neural networks

Restricted Boltzmann machine - example

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h}$$

$$= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$$

Distribution: $p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h})) / Z$

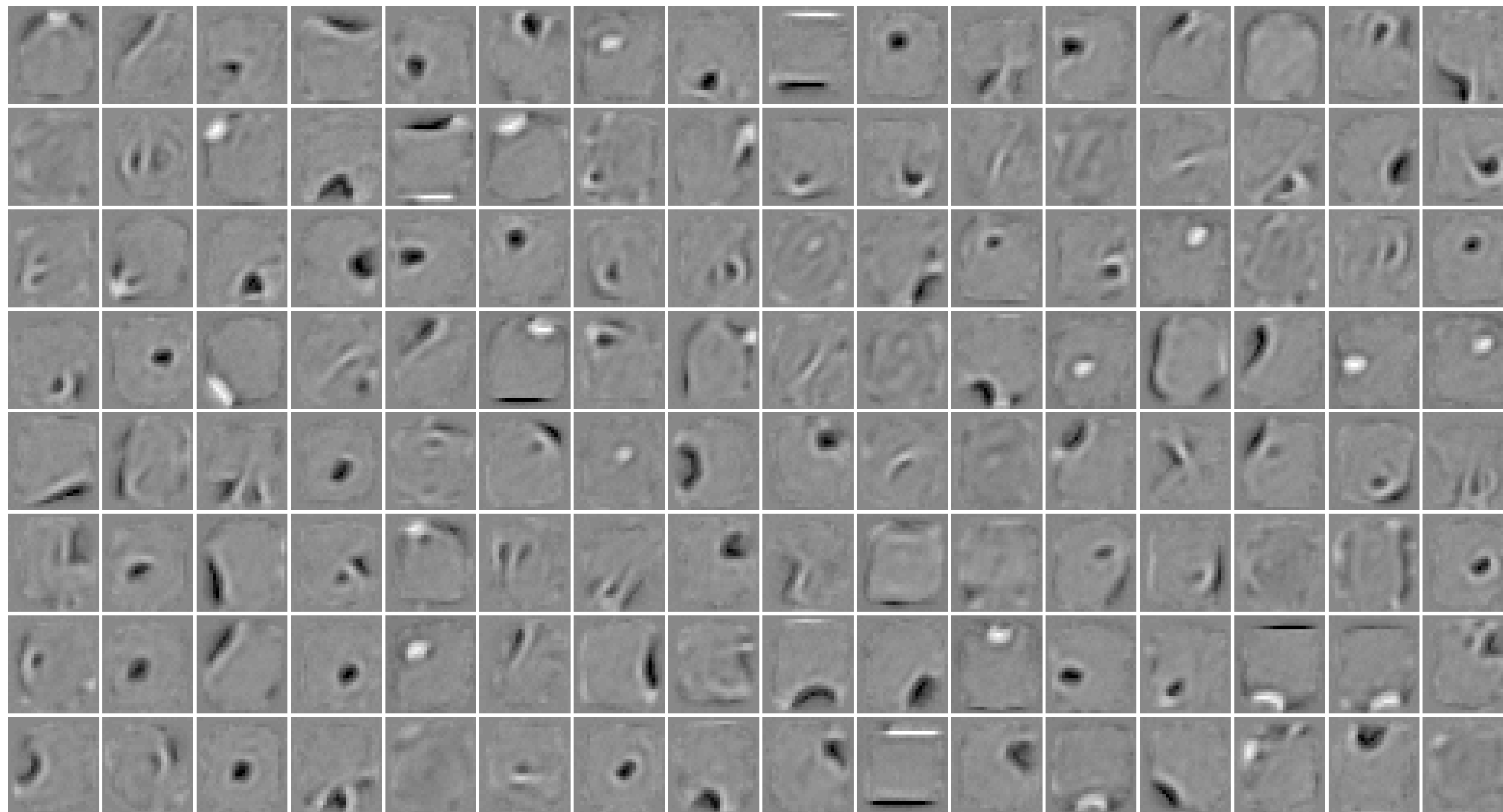
partition function
(intractable)

EXAMPLE OF DATA SET: MNIST

3	8	6	9	6	4	5	3	8	4	5	2	3	8	4	8
1	5	0	5	9	7	4	1	0	3	0	6	2	9	9	4
1	3	6	8	0	7	7	6	8	9	0	3	8	3	7	7
8	4	4	1	2	9	8	1	1	0	6	6	5	0	1	1
7	2	7	3	1	4	0	5	0	6	8	7	6	8	9	9
4	0	6	1	9	2	6	3	7	4	4	5	6	6	1	7
2	8	6	9	7	0	9	1	6	2	8	3	6	4	9	5
8	6	8	7	8	8	6	9	1	7	6	0	9	6	7	0

FILTERS

(LAROCHELLE ET AL., JMLR2009)



DEBUGGING

Topics: stochastic reconstruction, filters

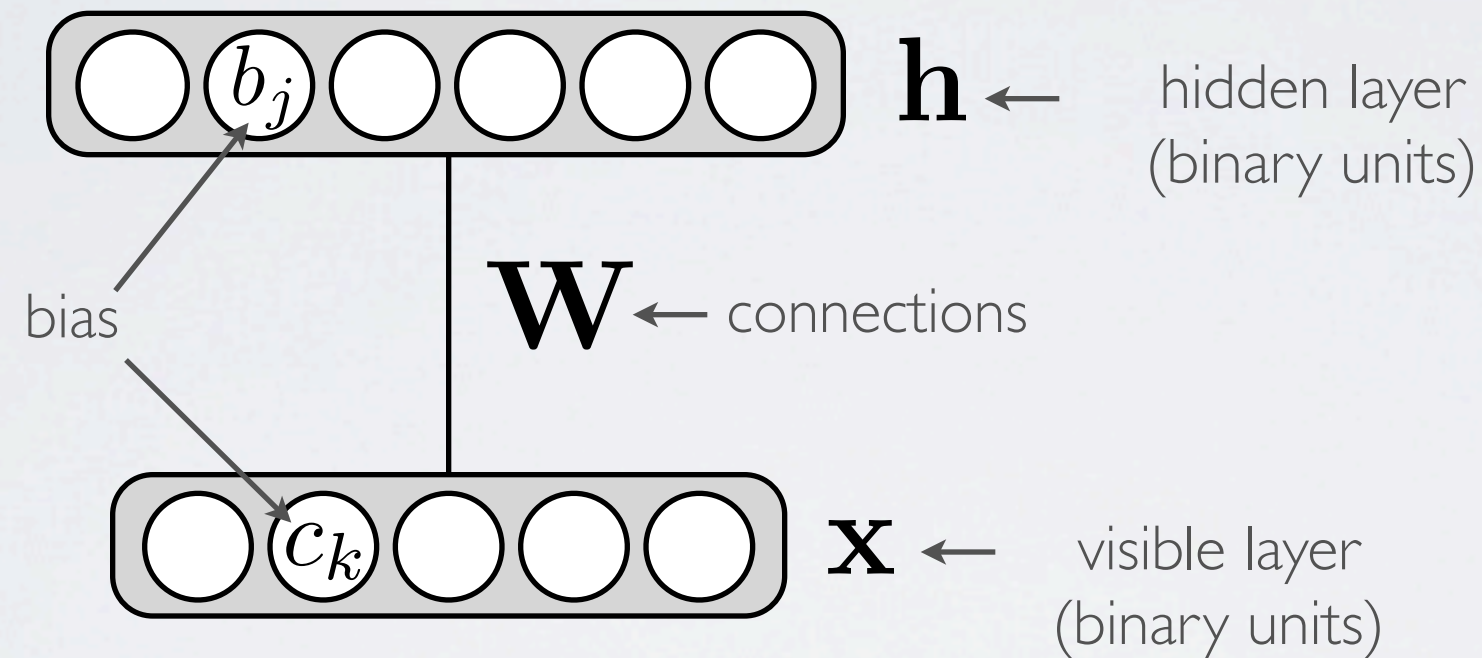
- Unfortunately, we can't debug with a comparison with finite difference
- We instead rely on approximate “tricks”
 - ▶ we plot the average stochastic reconstruction $\|\mathbf{x}^{(t)} - \tilde{\mathbf{x}}\|^2$ and see if it tends to decrease:
 - ▶ for inputs that correspond to image, we visualize the connection coming into each hidden unit as if it was an image
 - gives an idea of the type of visual feature each hidden unit detects
 - ▶ we can also try to approximate the partition function Z and see whether the (approximated) NLL decreases
 - On the Quantitative Analysis of Deep Belief Networks.
Ruslan Salakhutdinov and Iain Murray, 2008

Neural networks

Restricted Boltzmann machine - extensions

RESTRICTED BOLTZMANN MACHINE

Topics: RBM, visible layer, hidden layer, energy function



Energy function:
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h}$$

$$= -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$$

Distribution: $p(\mathbf{x}, \mathbf{h}) = \exp(-E(\mathbf{x}, \mathbf{h})) / Z$

← partition function (intractable)

GAUSSIAN-BERNOULLI RBM

Topics: Gaussian-Bernoulli RBM

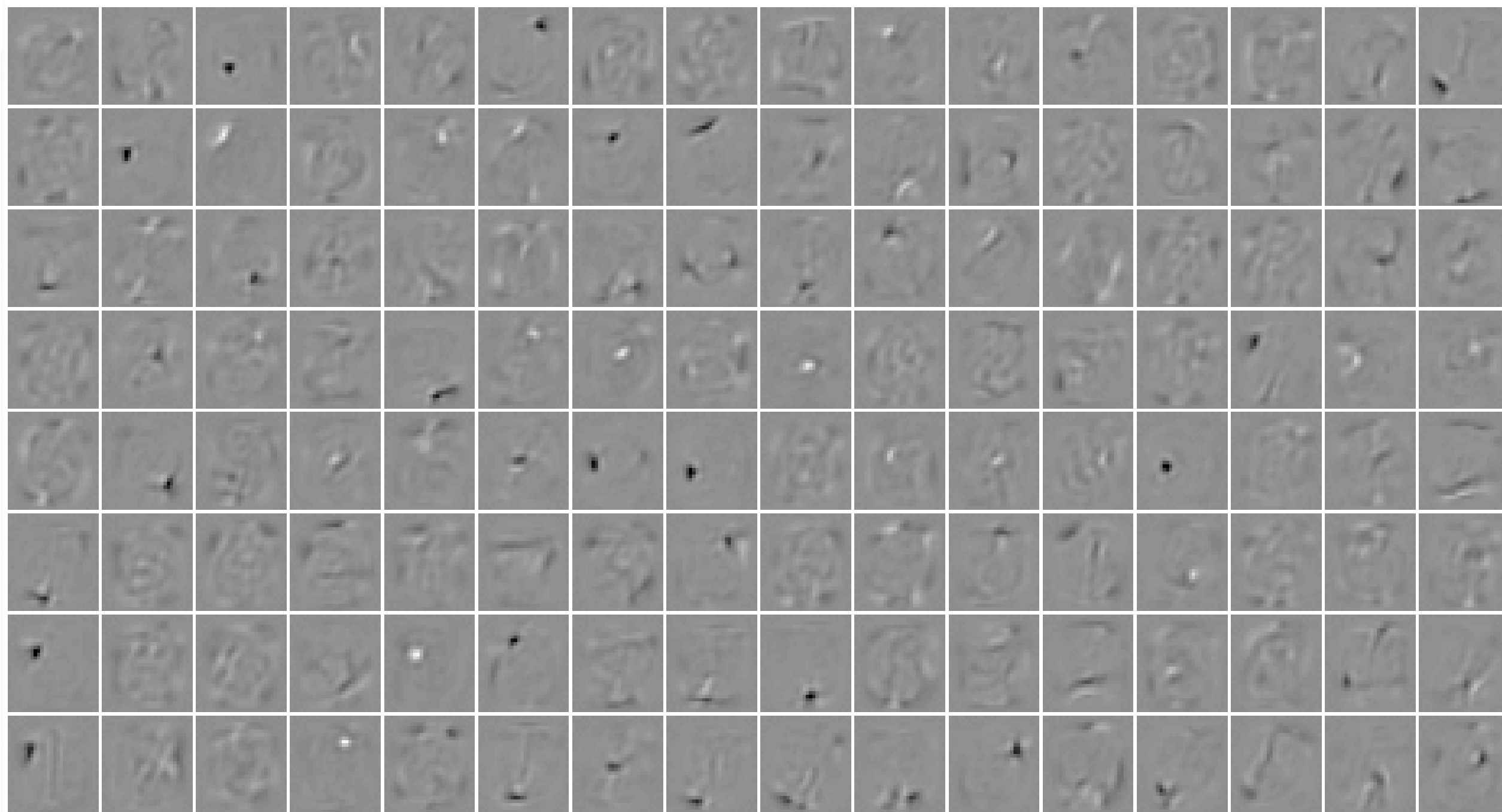
- Inputs \mathbf{x} are unbounded reals
 - add a quadratic term to the energy function

$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h} + \frac{1}{2} \mathbf{x}^\top \mathbf{x}$$

- only thing that changes is that $p(\mathbf{x}|\mathbf{h})$ is now a Gaussian distribution with mean $\boldsymbol{\mu} = \mathbf{c} + \mathbf{W}^\top \mathbf{h}$ and identity covariance matrix
- recommended to normalize the training set by
 - subtracting the mean of each input
 - dividing each input x_k by the training set standard deviation
- should use a smaller learning rate than in the regular RBM

FILTERS

(LAROCHELLE ET AL., JMLR2009)



OTHER TYPES OF OBSERVATIONS

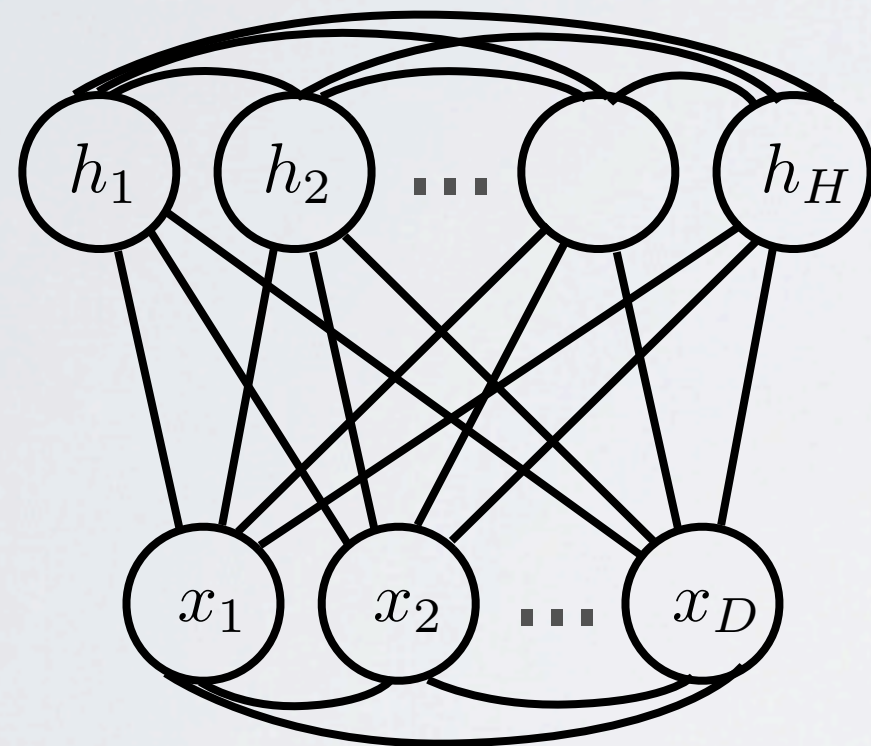
Topics: extensions to other observations

- Extensions support other types:
 - ▶ real-valued: Gaussian-Bernoulli RBM
 - ▶ Binomial observations:
 - Rate-coded Restricted Boltzmann Machines for Face Recognition.
Yee Whye Teh and Geoffrey Hinton, 2001
 - ▶ Multinomial observations:
 - Replicated Softmax: an Undirected Topic Model.
Ruslan Salakhutdinov and Geoffrey Hinton, 2009
 - Training Restricted Boltzmann Machines on Word Observations.
George Dahl, Ryan Adam and Hugo Larochelle, 2012
 - ▶ and more (see course website)

BOLTZMANN MACHINE

Topics: Boltzmann machine

- The original Boltzmann machine has lateral connections in each layer



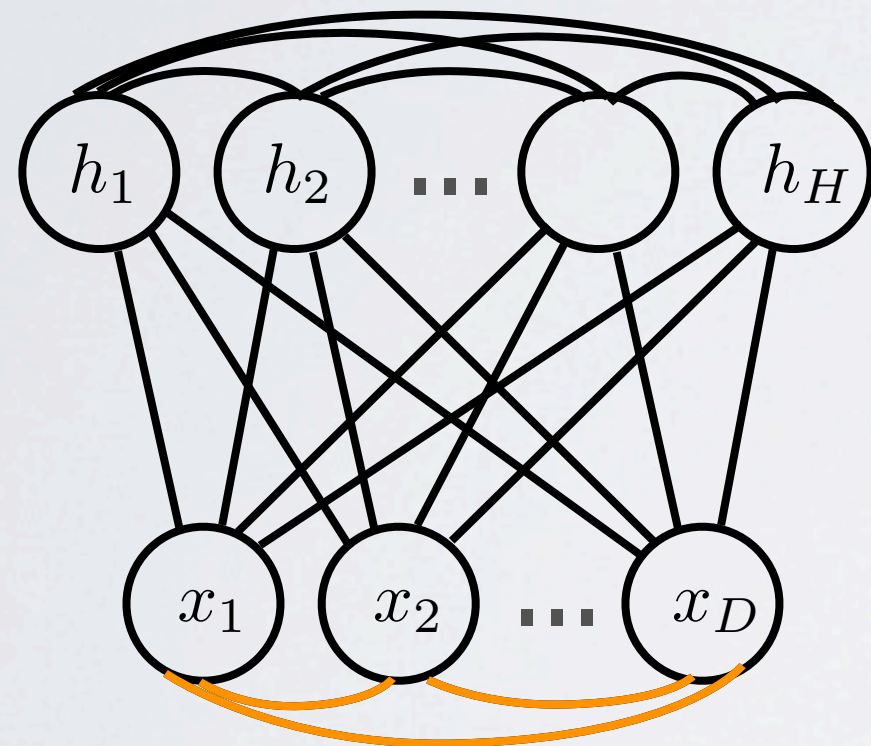
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h} - \frac{1}{2} \mathbf{x}^\top \mathbf{V} \mathbf{x} - \frac{1}{2} \mathbf{h}^\top \mathbf{U} \mathbf{h}$$

- ▶ when only one layer has lateral connection, it's a semi-restricted Boltzmann machine

BOLTZMANN MACHINE

Topics: Boltzmann machine

- The original Boltzmann machine has lateral connections in each layer



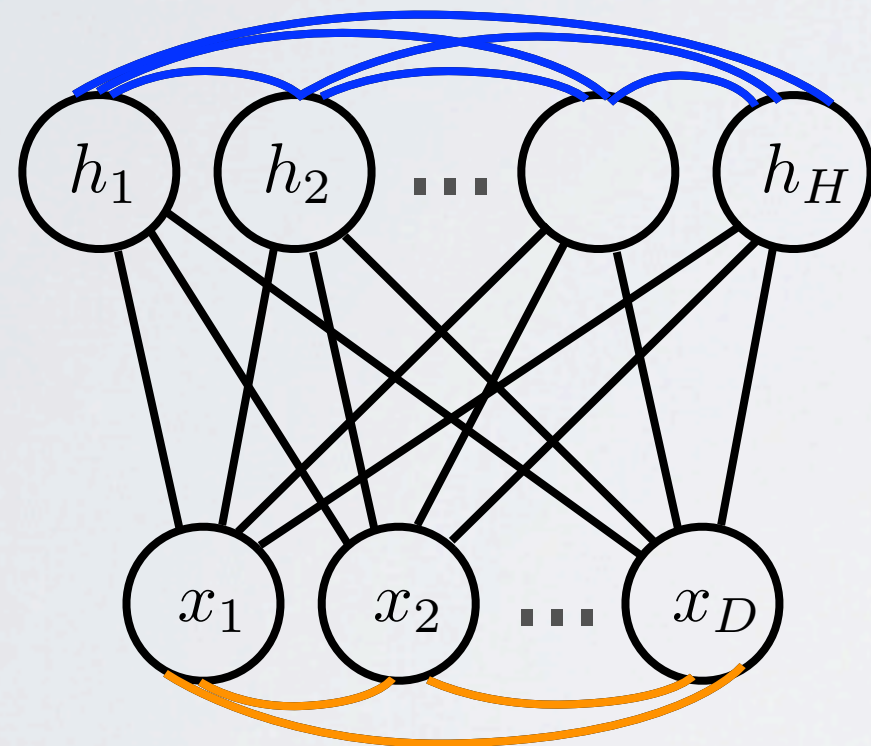
$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h} \\ - \frac{1}{2} \mathbf{x}^\top \mathbf{V} \mathbf{x} - \frac{1}{2} \mathbf{h}^\top \mathbf{U} \mathbf{h}$$

- when only one layer has lateral connection, it's a semi-restricted Boltzmann machine

BOLTZMANN MACHINE

Topics: Boltzmann machine

- The original Boltzmann machine has lateral connections in each layer



$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W} \mathbf{x} - \mathbf{c}^\top \mathbf{x} - \mathbf{b}^\top \mathbf{h} \\ - \frac{1}{2} \mathbf{x}^\top \mathbf{V} \mathbf{x} - \frac{1}{2} \mathbf{h}^\top \mathbf{U} \mathbf{h}$$

- when only one layer has lateral connection, it's a semi-restricted Boltzmann machine