Restricted Boltzmann Machines

IFT 725 - Réseaux neuronaux
UNSUPERVISED LEARNING

Topics: unsupervised learning

• Unsupervised learning: only use the inputs $x^{(t)}$ for learning
  ‣ automatically extract meaningful features for your data
  ‣ leverage the availability of unlabeled data
  ‣ add a data-dependent regularizer to training ($-\log p(x^{(t)})$)

• 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: \[ E(x, h) = -h^T W x - c^T x - b^T h \]

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

Distribution: \[ p(x, h) = \exp(-E(x, h))/Z \]

partition function (intractable)
**Markov Network View**

**Topics:** Markov network (with vector nodes)

\[ p(x, h) = \frac{\exp(-E(x, h))}{Z} \]

\[ = \frac{\exp(h^T W x + c^T x + b^T h)}{Z} \]

\[ = \frac{\exp(h^T W x) \exp(c^T x) \exp(b^T h)}{Z} \]

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

\[ \det(X) = \det(X_{ij}) \]
**MARKOV NETWORK VIEW**

**Topics:** Markov network (with scalar nodes)

\[
p(x, h) = \frac{1}{Z} \prod_j \prod_k \exp(W_{j,k} h_j x_k) \prod_k \exp(c_k x_k) \prod_j \exp(b_j h_j)
\]

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

**Topics:** Markov network (with scalar nodes)

\[ p(x, h) = \frac{1}{Z} \prod_j \prod_k \exp(W_{j,k} h_j x_k) \prod_k \exp(c_k x_k) \prod_j \exp(b_j h_j) \]

- The scalar visualization is more informative of the structure within the vectors
**Topics:** Markov network (with scalar nodes)

\[
\begin{align*}
p(x, h) &= \frac{1}{Z} \prod_j \prod_k \exp(W_{j,k} h_j x_k) \\
&\quad \prod_k \exp(c_k x_k) \\
&\quad \prod_j \exp(b_j h_j)
\end{align*}
\]

- The scalar visualization is more informative of the structure within the vectors
**Topics:** factor graph of an RBM
**INFERENCE**

**Topics:** conditional distributions

\[ p(h|x) = \prod_j p(h_j|x) \]

\[ p(h_j = 1|x) = \frac{1}{1 + \exp(-(b_j + W_{j:.}x))} \]

\[ = \text{sigm}(b_j + W_{j:.}x) \]

\[ p(x|h) = \prod_k p(x_k|h) \]

\[ p(x_k = 1|h) = \frac{1}{1 + \exp(-(c_k + h^T W_{.k}))} \]

\[ = \text{sigm}(c_k + h^T W_{.k}) \]
\[ p(h|\mathbf{x}) = \frac{p(\mathbf{x}, h)}{\sum_{h'} p(\mathbf{x}, h')} \]
\[ p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')} \]

\[ = \frac{\exp(h^\top Wx + c^\top x + b^\top h)/Z}{\sum_{h' \in \{0,1\}^H} \exp(h'^\top Wx + c^\top x + b^\top h')/Z} \]
\[
p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')}
= \frac{\exp(h^\top Wx + c^\top x + b^\top h)/Z}{\sum_{h' \in \{0,1\}^H} \exp(h'^\top Wx + c^\top x + b^\top h')/Z}
\]
\[ p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')} \]

\[ = \frac{\exp(h^\top Wx + e^\top x + b^\top h)/Z}{\sum_{h' \in \{0, 1\}} \exp(h'^\top Wx + e^\top x + b^\top h')/Z} \]
\[ p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')} \]

\[ = \frac{\exp(h^\top W x + e^\top x + b^\top h)/Z}{\sum_{h' \in \{0, 1\}^H} \exp(h'^\top W x + e^\top x + b^\top h')/Z} \]

\[ = \frac{\exp(\sum_j h_j W_j . x + b_j h_j)}{\sum_{h'_1 \in \{0, 1\}} \cdots \sum_{h'_H \in \{0, 1\}} \exp(\sum_j h'_j W_j . x + b_j h'_j)} \]
\[
p(h|\mathbf{x}) = \frac{p(\mathbf{x}, h)}{\sum_{h'} p(\mathbf{x}, h')}
\]

\[
= \frac{\exp(\mathbf{h}^\top \mathbf{W} \mathbf{x} + \mathbf{e}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h})/\mathcal{Z}}{\sum_{h'\in\{0,1\}^H} \exp(h'^\top \mathbf{W} \mathbf{x} + \mathbf{e}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{h'})/\mathcal{Z}}
\]

\[
= \frac{\exp(\sum_j h_j \mathbf{W}_j \cdot \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \exp(\sum_j h'_j \mathbf{W}_j \cdot \mathbf{x} + b_j h'_j) \prod_j \exp(h_j \mathbf{W}_j \cdot \mathbf{x} + b_j h_j)}
\]

\[
= \frac{\prod_j \exp(h_j \mathbf{W}_j \cdot \mathbf{x} + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{W}_j \cdot \mathbf{x} + b_j h'_j)}
\]
\[ p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')} \]

\[ = \frac{\exp(h^T W x + \mathbf{e}^T x + b^T h)/Z}{\sum_{h' \in \{0,1\}^H} \exp(h'^T W x + \mathbf{e}^T x + b^T h')/Z} \]

\[ = \frac{\exp(\sum_j h_j \mathbf{w}_j \cdot x + b_j h_j)}{\sum_{h'_1 \in \{0,1\}} \cdots \sum_{h'_H \in \{0,1\}} \prod_j \exp(h'_j \mathbf{w}_j \cdot x + b'_j h'_j)} \]

\[ = \frac{\prod_j \exp(h_j \mathbf{w}_j \cdot x + b_j h_j)}{\left( \sum_{h'_1 \in \{0,1\}} \exp(h'_1 \mathbf{w}_1 \cdot x + b_1 h'_1) \right) \cdots \left( \sum_{h'_H \in \{0,1\}} \exp(h'_H \mathbf{w}_H \cdot x + b_H h'_H) \right)} \]
\[
\begin{align*}
p(h|x) &= \frac{p(x, h)}{\sum_{h'} p(x, h')} \\
&= \frac{\exp(h^\top W x + e^\top x + b^\top h)/Z}{\sum_{h'\in\{0,1\}^H} \exp(h'^\top W x + e^\top x + b^\top h')/Z} \\
&= \frac{\exp(\sum_j h_j W_j^\top x + b_j h_j)}{\sum_{h'_1\in\{0,1\}} \cdots \sum_{h'_H\in\{0,1\}} \exp(\sum_j h'_j W_j^\top x + b_j h'_j)} \\
&= \frac{\prod_j \exp(h_j W_j^\top x + b_j h_j)}{\prod_j \left(\sum_{h'_j\in\{0,1\}} \exp(h'_j W_j^\top x + b_j h'_j)\right)} \\
&= \prod_j \left(\sum_{h'_j\in\{0,1\}} \exp(h'_j W_j^\top x + b_j h'_j)\right) \\
&\quad \prod_j \exp(h_j W_j^\top x + b_j h_j)
\end{align*}
\]
\[ p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')} \]
\[ = \frac{\exp(h^\top W x + e^\top x + b^\top h)/Z}{\sum_{h'\in\{0,1\}^H} \exp(h'^\top W x + e^\top x + b^\top h')/Z} \]
\[ = \frac{\exp(\sum_j h_j W_j \cdot x + b_j h_j)}{\sum_{h'\in\{0,1\}^H_{(j)}} \exp(\sum_{h'_{(j)}} h'_{(j)} \sum_{k\neq j} W_{k, j} \cdot x + b_{j, k} h_{j, k})} \]
\[ = \frac{\prod_j \exp(h_j W_j \cdot x + b_j h_j)}{\prod_j \left( \sum_{h'_{(j)}} \exp(\sum_{h'_{(j)}} h'_{(j)} W_{(j,j)} \cdot x + b_{j, (j)} h_{j, (j)}) \right)} \]
\[ = \prod_j \frac{\exp(h_j W_j \cdot x + b_j h_j)}{\prod_j (1 + \exp(b_j + \sum_{j} W_{j,j} \cdot x))} \]
\[
p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')}
\]

\[
= \frac{\exp(h^\top Wx + e^\top x + b^\top h)/Z}{\sum_{h' \in \{0,1\}^H} \exp(h'^\top Wx + e^\top x + b^\top h')/Z}
\]

\[
= \frac{\exp(\sum_j h_j W_j . x + b_j h_j)}{\sum_{h' \in \{0,1\}^H} \prod_j \exp(h_j W_j . x + b_j h_j)}
\]

\[
= \prod_j \frac{\exp(h_j W_j . x + b_j h_j)}{\prod_j \left(\sum_{h'_j \in \{0,1\}} \exp(h'_j W_j . x + b_j h'_j)\right)}
\]

\[
= \prod_j \frac{\exp(h_j W_j . x + b_j h_j)}{\prod_j (1 + \exp(b_j + W_j . x))}
\]

\[
= \prod_j \frac{\exp(h_j W_j . x + b_j h_j)}{1 + \exp(b_j + W_j . x)}
\]
\[ p(h|x) = \frac{p(x, h)}{\sum_{h'} p(x, h')} \]

\[ = \frac{\exp(h^\top W x + e^\top x + b^\top h)/Z}{\sum_{h' \in \{0,1\}^H} \exp(h'^\top W x + e^\top x + b^\top h')/Z} \]

\[ = \frac{\exp(\sum_j h_j W_j \cdot x + b_j h_j)}{\sum_{h_1' \in \{0,1\}} \cdots \sum_{h_H' \in \{0,1\}} \exp(\sum_j h_j' W_j \cdot x + b_j h_j')} \]

\[ = \frac{\prod_j \exp(h_j W_j \cdot x + b_j h_j)}{\sum_{h_1' \in \{0,1\}} \cdots \sum_{h_H' \in \{0,1\}} \prod_j \exp(h_j' W_j \cdot x + b_j h_j')} \]

\[ = \frac{\prod_j \exp(h_j W_j \cdot x + b_j h_j)}{\prod_j \left( \sum_{h_j' \in \{0,1\}} \exp(h_j' W_j \cdot x + b_j h_j') \right)} \]

\[ = \prod_j \frac{\exp(h_j W_j \cdot x + b_j h_j)}{1 + \exp(b_j + W_j \cdot x)} \]

\[ = \prod_j p(h_j | x) \]
\[ p(h_j = 1 | x) \]
\[ p(h_j = 1 | x) = \frac{\exp(b_j + W_j \cdot x)}{1 + \exp(b_j + W_j \cdot x)} \]
\[ p(h_j = 1|\mathbf{x}) = \frac{\exp(b_j + \mathbf{W}_j \cdot \mathbf{x})}{1 + \exp(b_j + \mathbf{W}_j \cdot \mathbf{x})} = \frac{1}{1 + \exp(-b_j - \mathbf{W}_j \cdot \mathbf{x})} \]
\[ p(h_j = 1|x) = \frac{\exp(b_j + W_j \cdot x)}{1 + \exp(b_j + W_j \cdot x)} \]

\[ = \frac{1}{1 + \exp(-b_j - W_j \cdot x)} \]

\[ = \text{sigm}(b_j + W_j \cdot x) \]
LOCAL MARKOV PROPERTY

**Topics:** local Markov property

• In general, we have the following property:

\[
p(z_i | z_1, \ldots, z_v) = \frac{p(z_i, \text{Ne}(z_i))}{\sum_{z'_i} p(z'_i, \text{Ne}(z_i))} = \frac{\prod f \text{ involving } z_i \Psi_f (z_i, \text{Ne}(z_i))}{\sum_{z'_i} \prod f \text{ involving } z_i \Psi_f (z'_i, \text{Ne}(z_i))}
\]

- \( z_i \) 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
RESTRICTED BOLTZMANN MACHINE

Topics: free energy

• What about \( p(x) \)?

\[
p(x) = \sum_{h \in \{0,1\}^H} p(x, h) = \sum_{h \in \{0,1\}^H} \exp(-E(x, h))/Z
\]
\[
= \exp \left( c^\top x + \sum_{j=1}^H \log(1 + \exp(b_j + W_j.x)) \right) / Z
\]
\[
= \exp(-F(x))/Z
\]

department d'informatique

École de Sherbrooke

October 10, 2012

Hugo Larochelle
\( p(x) \)
\[ p(x) = \sum_{h \in \{0,1\}^H} \exp(h^T W x + c^T x + b^T h)/Z \]
\[ p(x) = \sum_{h \in \{0, 1\}^H} \exp(h^T W x + c^T x + b^T h)/Z \]

\[ = \exp(c^T x) \sum_{h_1 \in \{0, 1\}} \cdots \sum_{h_H \in \{0, 1\}} \exp \left( \sum_{j} h_j W_j x + b_j h_j \right) / Z \]
\[ p(x) = \sum_{h \in \{0, 1\}^H} \exp(h^T W x + c^T x + b^T h)/Z \]

\[ = \exp(c^T x) \sum_{h_1 \in \{0, 1\}} \cdots \sum_{h_H \in \{0, 1\}} \exp \left( \sum_j h_j W_j x + b_j h_j \right)/Z \]

\[ = \exp(c^T x) \left( \sum_{h_1 \in \{0, 1\}} \exp(h_1 W_1 x + b_1 h_1) \right) \cdots \left( \sum_{h_H \in \{0, 1\}} \exp(h_H W_H x + b_H h_H) \right)/Z \]

...
\[ p(x) = \sum_{h \in \{0,1\}^H} \exp(h^T W x + c^T x + b^T h) / Z \]

\[ = \exp(c^T x) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp\left( \sum_{j} h_j W_j x + b_j h_j \right) / Z \]

\[ = \exp(c^T x) \left( \sum_{h_1 \in \{0,1\}} \exp(h_1 W_1 x + b_1 h_1) \right) \cdots \left( \sum_{h_H \in \{0,1\}} \exp(h_H W_H x + b_H h_H) \right) / Z \]

\[ = \exp(c^T x) \left( 1 + \exp(b_1 + W_1 x) \right) \cdots \left( 1 + \exp(b_H + W_H x) \right) / Z \]
\[ p(x) = \sum_{h \in \{0,1\}^H} \exp(h^T W x + c^T x + b^T h) / Z \]

\[ = \exp(c^T x) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp \left( \sum_j h_j w_{j}.x + b_j h_j \right) / Z \]

\[ = \exp(c^T x) \left( \sum_{h_1 \in \{0,1\}} \exp(h_1 w_{1}.x + b_1 h_1) \right) \cdots \left( \sum_{h_H \in \{0,1\}} \exp(h_H w_{H}.x + b_H h_H) \right) / Z \]

\[ = \exp(c^T x) (1 + \exp(b_1 + W_{1}.x)) \cdots (1 + \exp(b_H + W_{H}.x)) / Z \]

\[ = \exp(c^T x) \exp(\log(1 + \exp(b_1 + W_{1}.x))) \cdots \exp(\log(1 + \exp(b_H + W_{H}.x))) / Z \]
\[ p(x) = \sum_{h \in \{0,1\}^H} \exp(h^T W x + c^T x + b^T h)/Z \]

\[ = \exp(c^T x) \sum_{h_1 \in \{0,1\}} \cdots \sum_{h_H \in \{0,1\}} \exp \left( \sum_{j} h_j W_j . x + b_j h_j \right) /Z \]

\[ = \exp(c^T x) \left( \sum_{h_1 \in \{0,1\}} \exp(h_1 W_1 . x + b_1 h_1) \right) \cdots \left( \sum_{h_H \in \{0,1\}} \exp(h_H W_H . x + b_H h_H) \right) /Z \]

\[ = \exp(c^T x) (1 + \exp(b_1 + W_1 . x)) \cdots (1 + \exp(b_H + W_H . x)) /Z \]

\[ = \exp(c^T x) \exp(\log(1 + \exp(b_1 + W_1 . x))) \cdots \exp(\log(1 + \exp(b_H + W_H . x))) /Z \]

\[ = \exp \left( c^T x + \sum_{j=1}^{H} \log(1 + \exp(b_j + W_j . x)) \right) /Z \]
RESTRICTED BOLTZMANN MACHINE

Topics: free energy

\[
p(x) = \exp \left( c^T x + \sum_{j=1}^{H} \log(1 + \exp(b_j + W_j.x)) \right) / Z
\]

\[
= \exp \left( c^T x + \sum_{j=1}^{H} \text{softplus}(b_j + W_j.x) \right) / Z
\]
RESTRICTED BOLTZMANN MACHINE

**Topics:** free energy

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

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

bias the prob of each \( \mathbf{x}_i \)
RESTRICTED BOLTZMANN MACHINE

**Topics:** free energy

\[
p(x) = \exp \left( c^\top x + \sum_{j=1}^{H} \log(1 + \exp(b_j + W_j \cdot x)) \right) / Z
\]

\[
= \exp \left( c^\top x + \sum_{j=1}^{H} \text{softplus}(b_j + W_j \cdot x) \right) / Z
\]

"feature" expected in \( x \)

bias the prob of each \( x_i \)
RESTRICTED BOLTZMANN MACHINE

**Topics:** free energy

\[
p(x) = \exp \left( c^\top x + \sum_{j=1}^{H} \log(1 + \exp(b_j + W_j^\top x)) \right) / Z
\]

\[
= \exp \left( c^\top x + \sum_{j=1}^{H} \text{softplus}(b_j + W_j^\top x) \right) / Z
\]

"feature" expected in \(x\)

bias of each feature

bias the prob of each \(x_i\)

**Softplus Function:** \(\text{softplus}(\cdot) = \log(1 + e^{\cdot})\)
TRAINING

Topics: training objective

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

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

• We’d like to proceed by stochastic gradient descent

\[
\frac{\partial - \log p(x^{(t)})}{\partial \theta} = E_h \left[ \frac{\partial E(x^{(t)}, h)}{\partial \theta} \bigg| x^{(t)} \right] - E_{x,h} \left[ \frac{\partial E(x, h)}{\partial \theta} \right]
\]

positive phase  negative phase
TRAINING

**Topics:** training objective

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

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

- We’d like to proceed by stochastic gradient descent

\[
\frac{\partial}{\partial \theta} - \log p(x^{(t)}) = E_h \left[ \frac{\partial E(x^{(t)}, h)}{\partial \theta} \bigg| x^{(t)} \right] - E_{x,h} \left[ \frac{\partial E(x, h)}{\partial \theta} \right]
\]

\text{positive phase} \quad \text{negative phase}
CONTRASTIVE DIVERGENCE (CD)
(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence, negative sample

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

\[
\tilde{h}^{(t)} = h^0 \quad \sim p(h|x) \quad \sim p(x|h) \quad \xrightarrow{\ldots} \quad h^k = \tilde{h}
\]

\[
\tilde{x} = x^{(t)} \quad \tilde{x}^1 \quad \ldots \quad \tilde{x}^k = \tilde{x}
\]

\[
\frac{1}{T} \sum_{t=1}^{T} \log p(x^{(t)}) = \frac{1}{T} \sum_{t=1}^{T} \log p(x^{(t)}) \quad \frac{\partial}{\partial \theta} \log p(x^{(t)}) = E_{h \sim p(h|x)} \left[ \frac{\partial}{\partial \theta} E_{x \sim p(x|h)} \log p(x|h) \right]
\]
CONTRASTIVE DIVERGENCE (CD)
(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence

\[
E_h \left[ \frac{\partial E(x^{(t)}, h)}{\partial \theta} \middle| x^{(t)} \right] \approx \frac{\partial E(x^{(t)}, \tilde{h}^{(t)})}{\partial \theta}
\]

\[
E_{x, h} \left[ \frac{\partial E(x, h)}{\partial \theta} \right] \approx \frac{\partial E(\tilde{x}, \tilde{h})}{\partial \theta}
\]
CONTRASTIVE DIVERGENCE (CD)
(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence

\[
\mathbb{E}_h \left[ \frac{\partial E(x^{(t)}, h)}{\partial \theta} \middle| x^{(t)} \right] \approx \frac{\partial E(x^{(t)}, \tilde{h}^{(t)})}{\partial \theta} \quad \mathbb{E}_{x, h} \left[ \frac{\partial E(x, h)}{\partial \theta} \right] \approx \frac{\partial E(\tilde{x}, \tilde{h})}{\partial \theta}
\]
**CONTRASTIVE DIVERGENCE (CD)**
(HINTON, NEURAL COMPUTATION, 2002)

**Topics:** contrastive divergence

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

\[ E_{x,h} \left[ \frac{\partial E(x, h)}{\partial \theta} \right] \approx \frac{\partial E(\tilde{x}, \tilde{h})}{\partial \theta} \]
CONTRASTIVE DIVERGENCE (CD)
(HINTON, NEURAL COMPUTATION, 2002)

Topics: contrastive divergence

\[
\begin{align*}
E_h \left[ \frac{\partial E(x^{(t)}, h)}{\partial \theta} \mid x^{(t)} \right] & \approx \frac{\partial E(x^{(t)}, \tilde{h}^{(t)})}{\partial \theta} \\
E_x, h \left[ \frac{\partial E(x, h)}{\partial \theta} \right] & \approx \frac{\partial E(\tilde{x}, \tilde{h})}{\partial \theta}
\end{align*}
\]
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
- We can actually ignore the samples of the hidden layer \( h \) and integrate them out, conditioned on a value of \( x \)
DERIVATION OF THE LEARNING RULE

**Topics:** contrastive divergence

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

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

\[
\nabla_w E(x, h) = -h x^T
\]
DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

• Derivation of $\mathbb{E}_h \left[ \frac{\partial E(x, h)}{\partial \theta} \mid x \right]$ for $\theta = W_{jk}$

\[
\mathbb{E}_h \left[ \frac{\partial E(x, h)}{\partial W_{jk}} \mid x \right] = \mathbb{E}_h \left[ -h_j x_k \mid x \right] = \sum_{h_j \in \{0,1\}} -h_j x_k p(h_j \mid x)
\]

\[
= -x_k p(h_j = 1 \mid x)
\]

\[
\mathbb{E}_h [\nabla_W E(x, h) \mid x] = -h(x) x^\top
\]

\[
h(x) \overset{\text{def}}{=} \begin{pmatrix} p(h_1 = 1 \mid x) \\ \vdots \\ p(h_H = 1 \mid x) \end{pmatrix}
\]

\[
= \text{sigm}(b + Wx)
\]
DERIVATION OF THE LEARNING RULE

Topics: contrastive divergence

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

\[
W \iff W - \alpha \left( \nabla_W - \log p(x^{(t)}) \right)
\]

\[
\iff W - \alpha \left( E_h \left[ \nabla_W E(x^{(t)}, h) \big| x^{(t)} \right] - E_{x,h} \left[ \nabla_W E(x, h) \right] \right)
\]

\[
\iff W - \alpha \left( E_h \left[ \nabla_W E(x^{(t)}, h) \big| x^{(t)} \right] - E_h \left[ \nabla_W E(\tilde{x}, h) \big| \tilde{x} \right] \right)
\]

\[
\iff W + \alpha \left( h(x^{(t)}) x^{(t)\top} - h(\tilde{x}) \tilde{x}^{\top} \right)
\]
CD-K: PSEUDOCODE

**Topics:** contrastive divergence

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

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

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

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

2. Go back to 1 until stopping criteria
PERSISTENT CD (PCD)
(TIELEMAN, ICML2008)

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 $\tilde{\mathbf{x}}$
PERSISTENT CD (PCD)
(TIELEMAN, ICML2008)

**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 \( \tilde{\mathbf{x}} \)

\[
\tilde{\mathbf{h}}^{(t)} = h^0
\]

\[
\sim p(h|x)
\]

\[
\sim p(x|h)
\]

\[
\mathbf{x}^1
\]

\[
\mathbf{x}^k = \tilde{\mathbf{x}}
\]

\[
h^k = \tilde{\mathbf{h}}
\]

\[
\text{negative sample}
\]
**PERSISTENT CD (PCD)**

*(TIELEMAN, ICML2008)*

**Topics:** persistent contrastive divergence

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

\[ \tilde{h}(t) = h^0 \leftarrow \ldots \leftarrow x \]

\[ x^1 \quad \sim p(x|h) \]

\[ \sim p(h|x) \]

$\tilde{x}$ comes from the previous iteration

\[ x^k = \tilde{x} \quad \text{negative sample} \]

\[ \vdots \]

\[ h^k = \tilde{h} \]
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 $\mathcal{Z}$ and see whether the (approximated) NLL decreases
    - On the Quantitative Analysis of Deep Belief Networks.
      Ruslan Salakhutdinov and Iain Murray, 2008
EXAMPLE OF DATA SET: MNIST
FILTERS
(LAROCHELLE ET AL., JMLR2009)

<table>
<thead>
<tr>
<th>FILTERS</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="FILTERS" /></td>
</tr>
</tbody>
</table>

- Filter 1
- Filter 2
- Filter 3
- Filter 4
- Filter 5
- Filter 6
- Filter 7
- Filter 8
- Filter 9
- Filter 10
- Filter 11
- Filter 12
- Filter 13
- Filter 14
- Filter 15
- Filter 16
- Filter 17
- Filter 18
- Filter 19
- Filter 20

![Figure 6: Display of the input weights of a random subsequence of the hidden units, learned by an RBM. The activation of units of the first hidden layer is obtained by an RBM weighted by the input image.](image)

![Figure 7: Input weights of a random subsequence of the hidden units, learned by an RBM. The activation of units of the first hidden layer is obtained by an RBM weighted by the input image.](image)
GAUSSIAN-BERNOULLI RBM

Topics: Gaussian-Bernoulli RBM

• Inputs \( x \) are unbounded reals
  ‣ add a quadratic term to the energy function
    \[
    E(x, h) = -h^T W x - c^T x - b^T h + \frac{1}{2} x^T x
    \]
  ‣ only thing that changes is that \( p(x|h) \) is now a Gaussian distribution
    with mean \( \mu = c + W^T h \) 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

• Designing RBMs for different types of data is a “hot” topic
FILTERS
(LAROCHELLE ET AL., JMLR2009)
BOLTZMANN MACHINE

**Topics:** Boltzmann machine

- The original Boltzmann machine has lateral connections in each layer

\[
E(x, h) = -h^T W x - c^T x - b^T h - \frac{1}{2} x^T V x - \frac{1}{2} h^T U h
\]

\[\text{when only one layer has lateral connection, it’s a semi-restricted Boltzmann machine}\]
Topics: Boltzmann machine

- The original Boltzmann machine has lateral connections in each layer

\[ E(x, h) = -h^T W x - c^T x - b^T h - \frac{1}{2} x^T V x - \frac{1}{2} h^T U h \]

- 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

\[ E(x, h) = -h^T W x - c^T x - b^T h \]

- when only one layer has lateral connection, it’s a semi-restricted Boltmann machine
CONCLUSION

• We saw what is a restricted Boltzmann machine and how to train it with Contrastive Divergence.

• It assumes the observations are binary (Bernoulli), but variations 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:
  ‣ and more (see course website)