Restricted Boltzmann Machines for Collaborative Filtering

Ruslan Salakhutdinov Andriy Mnih Geoffrey Hinton

November 29, 2016

- Wikipedia:
  - In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

Collaborative Filtering

- Wikipedia:
  - In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

- Fundamental ideas:
  - If two items get similar rating patterns then they are probably similar

Collaborative Filtering

- Wikipedia:
  - In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

- Fundamental ideas:
  - If two items get similar rating patterns then they are probably similar
  - If two users rated items in a similar fashion, then they will probably give similar ratings to an unrated item

Collaborative Filtering

- Wikipedia:
  - In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

- Fundamental ideas:
  - If two items get similar rating patterns then they are probably similar
  - If two users rated items in a similar fashion, then they will probably give similar ratings to an unrated item
  - Properties of items are unknown

Collaborative Filtering

- Wikipedia:
  - In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).
- Fundamental ideas:
  - If two items get similar rating patterns then they are probably similar
  - If two users rated items in a similar fashion, then they will probably give similar ratings to an unrated item
  - Properties of items are unknown
- Applications:
  - Amazon (Customers Who Bought This Item Also Bought)

- Wikipedia:
  - In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).
- Fundamental ideas:
  - If two items get similar rating patterns then they are probably similar
  - If two users rated items in a similar fashion, then they will probably give similar ratings to an unrated item
  - Properties of items are unknown
- Applications:
  - Amazon (Customers Who Bought This Item Also Bought)
  - Netflix

Collaborative Filtering

- Wikipedia:
  - In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).
- Fundamental ideas:
  - If two items get similar rating patterns then they are probably similar
  - If two users rated items in a similar fashion, then they will probably give similar ratings to an unrated item
  - Properties of items are unknown
- Applications:
  - Amazon (Customers Who Bought This Item Also Bought)
  - Netflix
  - Spotify

The Task: Netflix Prize

- Basic statistics:
  - 480,189 users, <CustomerID>
  - 17,770 movies, <MovieID, YearOfRelease, Title>

The Task: Netflix Prize

- Basic statistics:
  - 480,189 users, <CustomerID>
  - 17,770 movies, <MovieID, YearOfRelease, Title>
- Training set:
  - 100,480,507 ratings, <CustomerID, MovieID, Rating, Date>
  - Rating is an integer that ranges from 1 to 5

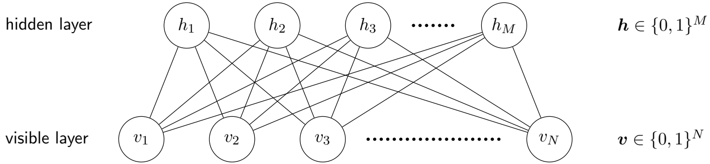
The Task: Netflix Prize

- Basic statistics:
  - 480,189 users, <CustomerID>
  - 17,770 movies, <MovieID, YearOfRelease, Title>
- Training set:
  - 100,480,507 ratings, <CustomerID, MovieID, Rating, Date>
  - Rating is an integer that ranges from 1 to 5
- Qualifying set:
  - 2,817,131 ratings, <CustomerID, MovieID, Date>

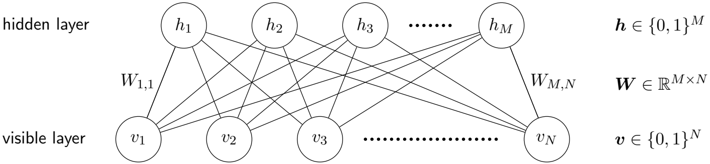
The Task: Netflix Prize

- Basic statistics:
  - 480,189 users, <CustomerID>
  - 17,770 movies, <MovieID, YearOfRelease, Title>
- Training set:
  - 100,480,507 ratings, <CustomerID, MovieID, Rating, Date>
  - Rating is an integer that ranges from 1 to 5
- Qualifying set:
  - 2,817,131 ratings, <CustomerID, MovieID, Date>
- Sparsity of ratings:
  - $\dfrac{100,480,507 + 2,817,131}{480,189 \times 17,770} = 0.0121 = 1.21\%$

Binary RBM

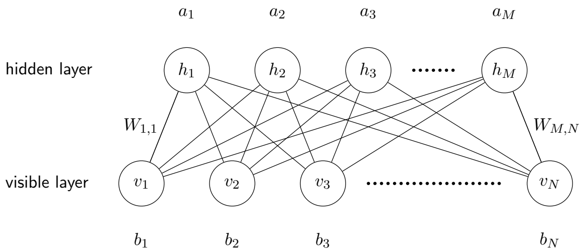

hidden layer
h_1
h_2
h_3
h_M
$\boldsymbol{h} \in \{0,1\}^M$
visible layer
v_1
v_2
v_3
v_N
$\boldsymbol{v} \in \{0,1\}^N$

Binary RBM



hidden layer
h_1
h_2
h_3
h_M
$\boldsymbol{h} \in \{0,1\}^M$
$W_{1,1}$
$W_{M,N}$
$\boldsymbol{W} \in \mathbb{R}^{M \times N}$
visible layer
v_1
v_2
v_3
v_N
$\boldsymbol{v} \in \{0,1\}^N$

Binary RBM



$\boldsymbol{a} \in \mathbb{R}^M$

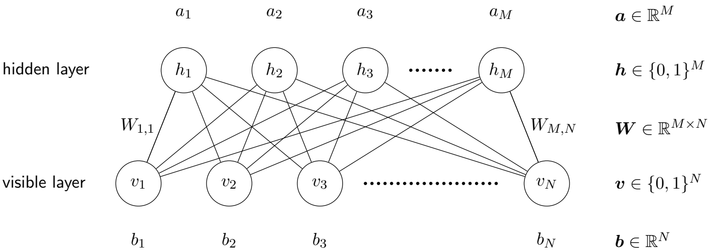
$\boldsymbol{h} \in \{0,1\}^M$

$\boldsymbol{W} \in \mathbb{R}^{M \times N}$

$\boldsymbol{v} \in \{0,1\}^N$

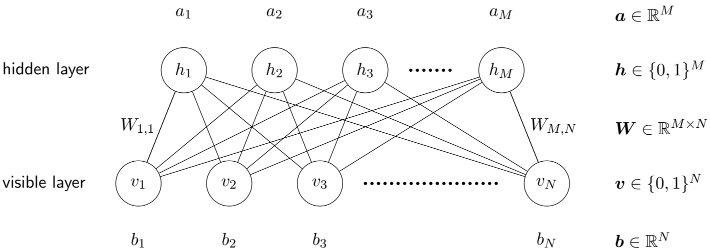
$\boldsymbol{b} \in \mathbb{R}^N$

Binary RBM



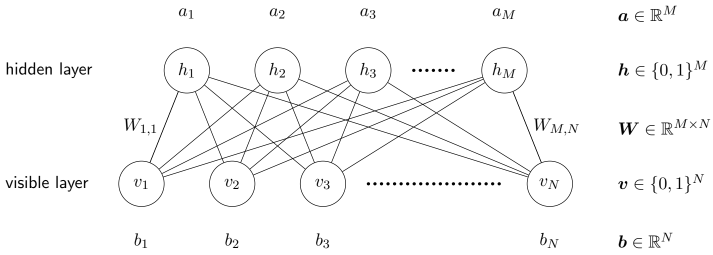
- $E(\boldsymbol{h}, \boldsymbol{v}) = -\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$

Binary RBM



- $E(\boldsymbol{h}, \boldsymbol{v}) = -\boldsymbol{a}^T \boldsymbol{h} - \boldsymbol{b}^T \boldsymbol{v} - \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}$
- $p(\boldsymbol{h}, \boldsymbol{v}) = \dfrac{1}{Z} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$

Binary RBM



- $E(\boldsymbol{h}, \boldsymbol{v}) = -\boldsymbol{a}^T \boldsymbol{h} - \boldsymbol{b}^T \boldsymbol{v} - \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}$
- $p(\boldsymbol{h}, \boldsymbol{v}) = \dfrac{1}{Z} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$
- $Z = \sum_{\boldsymbol{h}, \boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v})$$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{p(\boldsymbol{h}, \boldsymbol{v})}{p(\boldsymbol{v})}$$

Binary RBM, continued

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{p(\boldsymbol{h}, \boldsymbol{v})}{p(\boldsymbol{v})} \\
&= \frac{1}{p(\boldsymbol{v})} \frac{1}{Z} \exp(\boldsymbol{a}^T \boldsymbol{h} + \boldsymbol{b}^T \boldsymbol{v} + \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v})
\end{aligned}
$$

Binary RBM, continued

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{p(\boldsymbol{h}, \boldsymbol{v})}{p(\boldsymbol{v})} \\
&= \frac{1}{p(\boldsymbol{v})} \frac{1}{Z} \exp(\boldsymbol{a}^T \boldsymbol{h} + \boldsymbol{b}^T \boldsymbol{v} + \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}) \\
&= \frac{1}{p(\boldsymbol{v})} \frac{1}{Z} \exp(\boldsymbol{b}^T \boldsymbol{v}) \exp(\boldsymbol{a}^T \boldsymbol{h} + \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v})
\end{aligned}
$$

Binary RBM, continued

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{p(\boldsymbol{h},\boldsymbol{v})}{p(\boldsymbol{v})} \\
&= \frac{1}{p(\boldsymbol{v})}\frac{1}{Z}\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{b}^T\boldsymbol{v} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{p(\boldsymbol{v})}\frac{1}{Z}\exp(\boldsymbol{b}^T\boldsymbol{v})\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{Z'}\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v})
\end{aligned}
$$

Binary RBM, continued

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{p(\boldsymbol{h},\boldsymbol{v})}{p(\boldsymbol{v})} \\
&= \frac{1}{p(\boldsymbol{v})}\frac{1}{Z}\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{b}^T\boldsymbol{v} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{p(\boldsymbol{v})}\frac{1}{Z}\exp(\boldsymbol{b}^T\boldsymbol{v})\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{Z'}\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{Z'}\exp\left(\sum_{i=1}^{M} a_i h_i + \sum_{i=1}^{M} h_i \boldsymbol{W}_{i,:}\boldsymbol{v}\right)
\end{aligned}
$$

Binary RBM, continued

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{p(\boldsymbol{h}, \boldsymbol{v})}{p(\boldsymbol{v})} \\
&= \frac{1}{p(\boldsymbol{v})} \frac{1}{Z} \exp(\boldsymbol{a}^T \boldsymbol{h} + \boldsymbol{b}^T \boldsymbol{v} + \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}) \\
&= \frac{1}{p(\boldsymbol{v})} \frac{1}{Z} \exp(\boldsymbol{b}^T \boldsymbol{v}) \exp(\boldsymbol{a}^T \boldsymbol{h} + \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}) \\
&= \frac{1}{Z'} \exp(\boldsymbol{a}^T \boldsymbol{h} + \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}) \\
&= \frac{1}{Z'} \exp\left(\sum_{i=1}^{M} a_i h_i + \sum_{i=1}^{M} h_i \boldsymbol{W}_{i,:} \boldsymbol{v}\right) \\
&= \frac{1}{Z'} \exp\left(\sum_{i=1}^{M} h_i (a_i + \boldsymbol{W}_{i,:} \boldsymbol{v})\right)
\end{aligned}
$$

Binary RBM, continued

$$
\begin{aligned}
p(\boldsymbol{h}|\boldsymbol{v}) &= \frac{p(\boldsymbol{h},\boldsymbol{v})}{p(\boldsymbol{v})} \\
&= \frac{1}{p(\boldsymbol{v})}\frac{1}{Z}\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{b}^T\boldsymbol{v} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{p(\boldsymbol{v})}\frac{1}{Z}\exp(\boldsymbol{b}^T\boldsymbol{v})\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{Z'}\exp(\boldsymbol{a}^T\boldsymbol{h} + \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}) \\
&= \frac{1}{Z'}\exp\left(\sum_{i=1}^{M} a_i h_i + \sum_{i=1}^{M} h_i \boldsymbol{W}_{i,:}\boldsymbol{v}\right) \\
&= \frac{1}{Z'}\exp\left(\sum_{i=1}^{M} h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right) \\
&= \frac{1}{Z'}\prod_{i=1}^{M}\exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)
\end{aligned}
$$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{1}{Z'} \prod_{i=1}^{M} \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{1}{Z'} \prod_{i=1}^{M} \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$p(h_i|\boldsymbol{v}) \propto \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{1}{Z'} \prod_{i=1}^{M} \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$p(h_i|\boldsymbol{v}) \propto \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$p(h_i = 1|\boldsymbol{v}) = \frac{p(h_i = 1|\boldsymbol{v})}{p(h_i = 0|\boldsymbol{v}) + p(h_i = 1|\boldsymbol{v})}$$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{1}{Z'} \prod_{i=1}^{M} \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$p(h_i|\boldsymbol{v}) \propto \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$\begin{aligned} p(h_i = 1|\boldsymbol{v}) &= \frac{p(h_i = 1|\boldsymbol{v})}{p(h_i = 0|\boldsymbol{v}) + p(h_i = 1|\boldsymbol{v})} \\ &= \frac{\exp\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right)}{1 + \exp\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right)} \end{aligned}$$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{1}{Z'} \prod_{i=1}^{M} \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$p(h_i|\boldsymbol{v}) \propto \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$\begin{aligned} p(h_i = 1|\boldsymbol{v}) &= \frac{p(h_i = 1|\boldsymbol{v})}{p(h_i = 0|\boldsymbol{v}) + p(h_i = 1|\boldsymbol{v})} \\ &= \frac{\exp\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right)}{1 + \exp\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right)} \\ &= \mathsf{sigmoid}\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right) \end{aligned}$$

Binary RBM, continued

$$p(\boldsymbol{h}|\boldsymbol{v}) = \frac{1}{Z'} \prod_{i=1}^{M} \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$p(h_i|\boldsymbol{v}) \propto \exp\left(h_i(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})\right)$$

$$\begin{aligned} p(h_i = 1|\boldsymbol{v}) &= \frac{p(h_i = 1|\boldsymbol{v})}{p(h_i = 0|\boldsymbol{v}) + p(h_i = 1|\boldsymbol{v})} \\ &= \frac{\exp\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right)}{1 + \exp\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right)} \\ &= \mathsf{sigmoid}\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v}\right) \end{aligned}$$

$$p(v_j = 1|\boldsymbol{h}) = \mathsf{sigmoid}\left(b_j + \boldsymbol{h}^T\boldsymbol{W}_{:,j}\right)$$

Binary RBM Learning

- Binary RBM's parameters: $\theta = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$

Binary RBM Learning

- Binary RBM's parameters: $\theta = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$
- Maximum likelihood principle: $\mathrm{argmax}_\theta \, p(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(T)} | \theta)$

Binary RBM Learning

- Binary RBM's parameters: $\theta = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$
- Maximum likelihood principle: $\mathrm{argmax}_\theta \, p(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(T)} | \theta)$
- No closed-form solution

Binary RBM Learning

- Binary RBM's parameters: $\theta = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$
- Maximum likelihood principle: $\operatorname{argmax}_\theta p(\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \ldots, \boldsymbol{v}^{(T)} | \theta)$
- No closed-form solution
- Resort to gradient ascent

Binary RBM Learning, continued

$\ell(\theta)$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log p(\boldsymbol{v}^{(t)})$$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log p(\boldsymbol{v}^{(t)})$$
$$= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \boldsymbol{v}^{(t)})$$

Binary RBM Learning, continued

$$\begin{aligned}
\ell(\theta) &= \sum_{t=1}^{T} \log p(\boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \frac{1}{Z} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)
\end{aligned}$$

Binary RBM Learning, continued

$$
\begin{aligned}
\ell(\theta) &= \sum_{t=1}^{T} \log p(\boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \frac{1}{Z} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \\
&= \sum_{t=1}^{T} \log \frac{1}{Z} \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)
\end{aligned}
$$

Binary RBM Learning, continued

$$
\begin{aligned}
\ell(\theta) &= \sum_{t=1}^{T} \log p(\boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \frac{1}{Z} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \\
&= \sum_{t=1}^{T} \log \frac{1}{Z} \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log Z
\end{aligned}
$$

Binary RBM Learning, continued

$$
\begin{aligned}
\ell(\theta) &= \sum_{t=1}^{T} \log p(\boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \boldsymbol{v}^{(t)}) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \frac{1}{Z} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \\
&= \sum_{t=1}^{T} \log \frac{1}{Z} \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log Z \\
&= \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h}, \boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))
\end{aligned}
$$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\frac{\partial \ell(\theta)}{\partial \beta}$$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\frac{\partial \ell(\theta)}{\partial \beta} = \sum_{t=1}^{T} \frac{\partial \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta} - T \frac{\partial \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}$$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta} &= \sum_{t=1}^{T} \frac{\partial \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta} - T \frac{\partial \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta} \\
&= \sum_{t=1}^{T} \frac{\frac{\partial \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\frac{\partial \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}
\end{aligned}$$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta} &= \sum_{t=1}^{T} \frac{\partial \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta} - T \frac{\partial \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta} \\
&= \sum_{t=1}^{T} \frac{\frac{\partial \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\frac{\partial \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \frac{\partial \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\partial \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}
\end{aligned}$$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta} &= \sum_{t=1}^{T} \frac{\partial \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta} - T \frac{\partial \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta} \\
&= \sum_{t=1}^{T} \frac{\frac{\partial \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\frac{\partial \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \frac{\partial \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\partial \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \frac{\partial -E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v})) \frac{\partial -E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}
\end{aligned}$$

Binary RBM Learning, continued

$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta} &= \sum_{t=1}^{T} \frac{\partial \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta} - T \frac{\partial \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta} \\
&= \sum_{t=1}^{T} \frac{\frac{\partial \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\frac{\partial \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \frac{\partial \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\partial \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v})) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{h}} \frac{\exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta} - T \sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta}
\end{aligned}$$

$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta} &= \sum_{t=1}^{T} \frac{\partial \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta} - T \frac{\partial \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta} \\
&= \sum_{t=1}^{T} \frac{\frac{\partial \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\frac{\partial \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \frac{\partial \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\partial \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v})) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{h}} \frac{\exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta} - T \sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta} \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(t)}) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta} - T \sum_{\boldsymbol{h},\boldsymbol{v}} p(\boldsymbol{h}, \boldsymbol{v}) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta}
\end{aligned}$$

Binary RBM Learning, continued

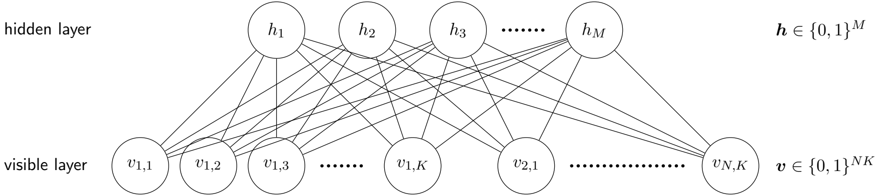
$$\ell(\theta) = \sum_{t=1}^{T} \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) - T \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))$$

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \beta} &= \sum_{t=1}^{T} \frac{\partial \log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\partial \beta} - T \frac{\partial \log \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\partial \beta} \\
&= \sum_{t=1}^{T} \frac{\frac{\partial \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h},\boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\frac{\partial \sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h},\boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \frac{\partial \exp\left(-E(\boldsymbol{h},\boldsymbol{v}^{(t)})\right)}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\partial \exp(-E(\boldsymbol{h},\boldsymbol{v}))}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \frac{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right) \frac{\partial -E(\boldsymbol{h},\boldsymbol{v}^{(t)})}{\partial \beta}}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} - T \frac{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v})) \frac{\partial -E(\boldsymbol{h},\boldsymbol{v})}{\partial \beta}}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{h}} \frac{\exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)}{\sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{h}, \boldsymbol{v}^{(t)})\right)} \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta} - T \sum_{\boldsymbol{h},\boldsymbol{v}} \frac{\exp(-E(\boldsymbol{h}, \boldsymbol{v}))}{\sum_{\boldsymbol{h},\boldsymbol{v}} \exp(-E(\boldsymbol{h}, \boldsymbol{v}))} \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta} \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(t)}) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta} - T \sum_{\boldsymbol{h},\boldsymbol{v}} p(\boldsymbol{h}, \boldsymbol{v}) \frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta} \\
&= \sum_{t=1}^{T} \mathbb{E}_{p(\boldsymbol{h}|\boldsymbol{v}^{(t)})} \left[\frac{\partial - E(\boldsymbol{h}, \boldsymbol{v}^{(t)})}{\partial \beta}\right] - T\, \mathbb{E}_{p(\boldsymbol{h},\boldsymbol{v})} \left[\frac{\partial - E(\boldsymbol{h}, \boldsymbol{v})}{\partial \beta}\right]
\end{aligned}$$

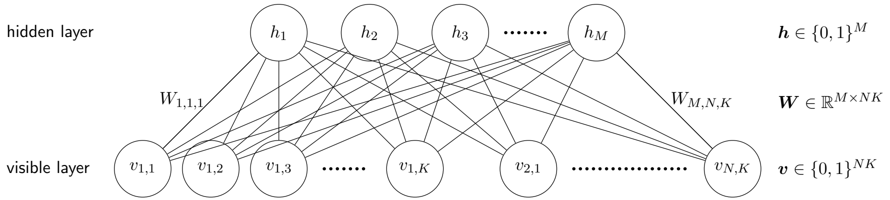
Binary RBM Summary

- A two layer RBM can be fully characterized by the following:
  - $E(\boldsymbol{h}, \boldsymbol{v})$
  - $p(\boldsymbol{h}, \boldsymbol{v})$
  - Z
  - $p(h_i = s|\boldsymbol{v})$
  - $p(v_j = t|\boldsymbol{h})$

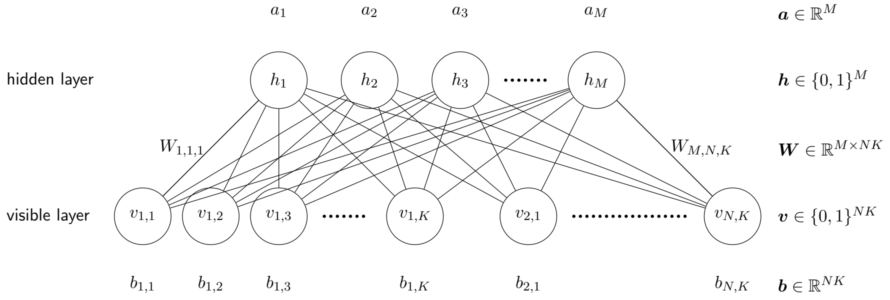
Towards RBM for CF, First Step: Make Visible Units K-nary



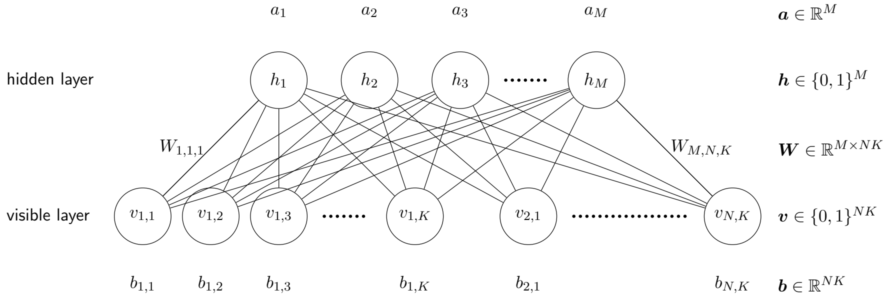
Towards RBM for CF, First Step: Make Visible Units K-nary



Towards RBM for CF, First Step: Make Visible Units K-nary



Towards RBM for CF, First Step: Make Visible Units K-nary



$v_{i,:}$	Valid Assignment?
00001	$\surd$
01000	$\surd$
11010	$\times$
00000	$\times$

Towards RBM for CF, First Step: Make Visible Units K-nary, continued

Binary	K-nary

Towards RBM for CF, First Step: Make Visible Units K-nary, continued

	Binary	K-nary	
$E(\boldsymbol{h}, \boldsymbol{v})$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	K-nary requires valid $\boldsymbol{v}$.

Towards RBM for CF, First Step: Make Visible Units K-nary, continued

	Binary	K-nary	
$E(\boldsymbol{h}, \boldsymbol{v})$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	K-nary requires valid $\boldsymbol{v}$.
$p(\boldsymbol{h}, \boldsymbol{v})$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.

Towards RBM for CF, First Step: Make Visible Units K-nary, continued

	Binary	K-nary	
$E(\boldsymbol{h}, \boldsymbol{v})$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	K-nary requires valid $\boldsymbol{v}$.
$p(\boldsymbol{h}, \boldsymbol{v})$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.
Z	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.

Towards RBM for CF, First Step: Make Visible Units K-nary, continued

	Binary	K-nary	
$E(\boldsymbol{h}, \boldsymbol{v})$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	K-nary requires valid $\boldsymbol{v}$.
$p(\boldsymbol{h}, \boldsymbol{v})$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.
Z	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.
$p(h_i = 1 \vert \boldsymbol{v})$	sigmoid $(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})$	sigmoid $(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})$	K-nary requires valid $\boldsymbol{v}$.

Towards RBM for CF, First Step: Make Visible Units K-nary, continued

	Binary	K-nary	
$E(\boldsymbol{h},\boldsymbol{v})$	$-\boldsymbol{a}^T\boldsymbol{h}-\boldsymbol{b}^T\boldsymbol{v}-\boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	$-\boldsymbol{a}^T\boldsymbol{h}-\boldsymbol{b}^T\boldsymbol{v}-\boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	K-nary requires valid $\boldsymbol{v}$.
$p(\boldsymbol{h},\boldsymbol{v})$	$\frac{1}{Z}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$	$\frac{1}{Z}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.
Z	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.
$p(h_i=1\vert\boldsymbol{v})$	sigmoid $(a_i+\boldsymbol{W}_{i,:}\boldsymbol{v})$	sigmoid $(a_i+\boldsymbol{W}_{i,:}\boldsymbol{v})$	K-nary requires valid $\boldsymbol{v}$.
$p(v_j=1\vert\boldsymbol{h})$	sigmoid $\left(b_j+\boldsymbol{h}^T\boldsymbol{W}_{:,j}\right)$		Undefined for K-nary.

Towards RBM for CF, First Step: Make Visible Units K-nary, continued

	Binary	K-nary	
$E(\boldsymbol{h}, \boldsymbol{v})$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	K-nary requires valid $\boldsymbol{v}$.
$p(\boldsymbol{h}, \boldsymbol{v})$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	$\frac{1}{Z}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.
Z	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h}, \boldsymbol{v}))$	K-nary requires valid $\boldsymbol{v}$.
$p(h_i = 1 \mid \boldsymbol{v})$	sigmoid $(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})$	sigmoid $(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})$	K-nary requires valid $\boldsymbol{v}$.
$p(v_j = 1 \mid \boldsymbol{h})$	sigmoid $\left(b_j + \boldsymbol{h}^T\boldsymbol{W}_{:,j}\right)$		Undefined for K-nary.
$p(v_{j,k} = 1 \mid \boldsymbol{h})$		$\dfrac{\exp\left(b_{j,k} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,k}\right)}{\sum_{l=1}^{K}\exp\left(b_{j,l} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,l}\right)}$	Undefined for Binary.

Towards RBM for CF, Second Step: Handle Missing Values

- How to handle missing values?

Towards RBM for CF, Second Step: Handle Missing Values

- How to handle missing values?
- Imputation
  - Fill in the blanks with some estimation based on available information.

Towards RBM for CF, Second Step: Handle Missing Values

- How to handle missing values?
- Imputation
  - Fill in the blanks with some estimation based on available information.
- Parameter Sharing
  - Use RBM with different numbers of visible units for different training/test cases.

Suppose $K = 3$ for the following case:

	m1	m2	m3
u1	1	?	?
u2	2	?	2
u3	1	3	?

Suppose $K = 3$ for the following case:

	m1	m2	m3
u1	1	?	?
u2	2	?	2
u3	1	3	?

	$\boldsymbol{v}$	$\widetilde{\boldsymbol{v}}$	Visible Units
u1	100 ??? ???	100	$v_{1,1}, v_{1,2}, v_{1,3}$
u2	010 ??? 010	010 010	$v_{1,1}, v_{1,2}, v_{1,3}, v_{3,1}, v_{3,2}, v_{3,3}$
u3	100 001 ???	100 001	$v_{1,1}, v_{1,2}, v_{1,3}, v_{2,1}, v_{2,2}, v_{2,3}$

Towards RBM for CF, Third Step: Make Predictions

$$p(v_{q,k} = 1|\widetilde{\boldsymbol{v}})$$

Towards RBM for CF, Third Step: Make Predictions

$$p(v_{q,k} = 1 | \widetilde{\boldsymbol{v}}) = \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k})$$

Towards RBM for CF, Third Step: Make Predictions

$$
\begin{aligned}
p(v_{q,k} = 1 | \widetilde{\boldsymbol{v}}) &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}) \\
&\propto \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}))
\end{aligned}
$$

Towards RBM for CF, Third Step: Make Predictions

$$
\begin{aligned}
p(v_{q,k} = 1 | \widetilde{\boldsymbol{v}}) &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}) \\
&\propto \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}))
\end{aligned}
$$

Can be computed in polynomial time.

Towards RBM for CF, Third Step: Make Predictions

$$
\begin{aligned}
p(v_{q,k} = 1 | \widetilde{\boldsymbol{v}}) &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}) \\
&\propto \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}))
\end{aligned}
$$

Can be computed in polynomial time.

$$
p(v_{q_1,k_1} = 1, v_{q_2,k_2} = 1, \ldots, v_{q_S,k_S} = 1 | \widetilde{\boldsymbol{v}})
$$

Towards RBM for CF, Third Step: Make Predictions

$$
\begin{aligned}
p(v_{q,k} = 1|\widetilde{\boldsymbol{v}}) &= \sum_{\boldsymbol{h}} p(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}) \\
&\propto \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{h}, \widetilde{\boldsymbol{v}}, v_{q,k}))
\end{aligned}
$$

Can be computed in polynomial time.

$$
p(v_{q_1,k_1} = 1, v_{q_2,k_2} = 1, \ldots, v_{q_S,k_S} = 1|\widetilde{\boldsymbol{v}})
$$

However, query of S movies on a user requires K^S evaluations.

$$\widehat{h_i} = p(h_i = 1|\widetilde{\boldsymbol{v}}) = \textsf{sigmoid}\left(a_i + \widetilde{\boldsymbol{W}}_{i,:}\widetilde{\boldsymbol{v}}\right)$$

$$\widehat{h}_i = p(h_i = 1|\widetilde{\boldsymbol{v}}) = \textsf{sigmoid}\left(a_i + \widetilde{\boldsymbol{W}}_{i,:}\widetilde{\boldsymbol{v}}\right)$$

$$p(v_{q,k} = 1|\widehat{\boldsymbol{h}}) = \frac{\exp\left(b_{q,k} + \widehat{\boldsymbol{h}}^T \boldsymbol{W}_{:,q,k}\right)}{\sum_{l=1}^{K} \exp\left(b_{q,l} + \widehat{\boldsymbol{h}}^T \boldsymbol{W}_{:,q,l}\right)}$$

$$\widehat{h}_i = p(h_i = 1|\widetilde{\boldsymbol{v}}) = \text{sigmoid}\left(a_i + \widetilde{\boldsymbol{W}}_{i,:}\widetilde{\boldsymbol{v}}\right)$$

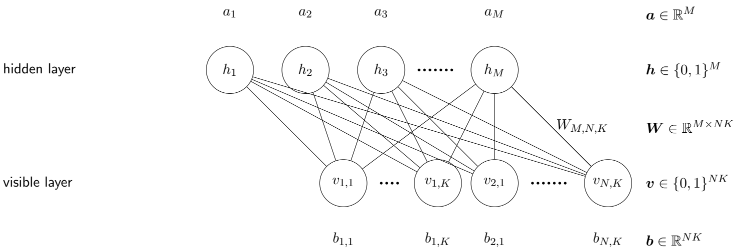
$$p(v_{q,k} = 1|\widehat{\boldsymbol{h}}) = \frac{\exp\left(b_{q,k} + \widehat{\boldsymbol{h}}^T \boldsymbol{W}_{:,q,k}\right)}{\sum_{l=1}^{K} \exp\left(b_{q,l} + \widehat{\boldsymbol{h}}^T \boldsymbol{W}_{:,q,l}\right)}$$

A lot faster but slightly less accurate.

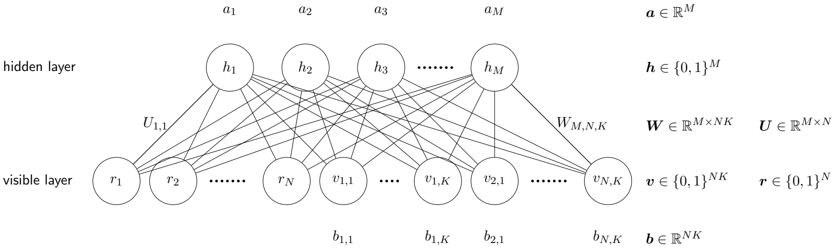
Extension 1: Gaussian Hidden Units

	K-nary	K-nary with Gaussian
$E(\boldsymbol{h},\boldsymbol{v})$	$-\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$	$\sum_{i=1}^{M} \frac{(h_i - a_i)^2}{2{\sigma_i}^2} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v}$
$p(\boldsymbol{h},\boldsymbol{v})$	$\frac{1}{Z}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$	$\frac{1}{Z}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$
Z	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$	$\sum_{\boldsymbol{h},\boldsymbol{v}}\exp(-E(\boldsymbol{h},\boldsymbol{v}))$
$p(v_{j,k}=1\|\boldsymbol{h})$	$\frac{\exp\left(b_{j,k} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,k}\right)}{\sum_{l=1}^{K}\exp\left(b_{j,l} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,l}\right)}$	$\frac{\exp\left(b_{j,k} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,k}\right)}{\sum_{l=1}^{K}\exp\left(b_{j,l} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,l}\right)}$
$p(h_i=1\|\boldsymbol{v})$	sigmoid $(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v})$	
$p(h_i=h\|\boldsymbol{v})$		$\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(h - a_i - \sigma_i\boldsymbol{W}_{i,:}\boldsymbol{v})^2}{2\sigma_i^2}\right)$

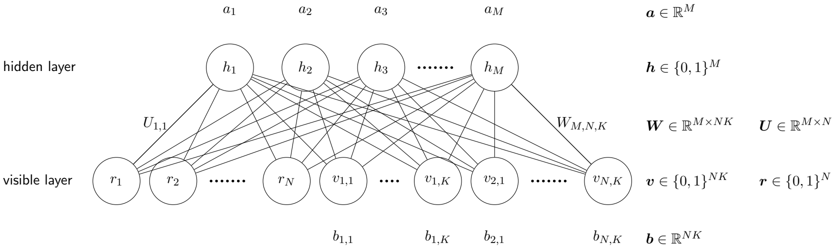
Extension 2: Condition On What Users Have Watched



Extension 2: Condition On What Users Have Watched



Extension 2: Condition On What Users Have Watched



- $E(\boldsymbol{h}, \boldsymbol{v}, \boldsymbol{r}) = -\boldsymbol{a}^T\boldsymbol{h} - \boldsymbol{b}^T\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v} - \boldsymbol{h}^T\boldsymbol{W}\boldsymbol{r}$
- $p(v_{j,k} = 1|\boldsymbol{h}) = \dfrac{\exp\left(b_{j,k} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,k}\right)}{\sum_{l=1}^{K} \exp\left(b_{j,l} + \boldsymbol{h}^T\boldsymbol{W}_{:,j,l}\right)}$
- $p(h_i = 1|\boldsymbol{v}, \boldsymbol{r}) = \text{sigmoid}\left(a_i + \boldsymbol{W}_{i,:}\boldsymbol{v} + \boldsymbol{U}_{i,:}\boldsymbol{r}\right)$

Extension 3: Factorize $\boldsymbol{W}$

- $\boldsymbol{W} \in \mathbb{R}^{M \times NK}$

Extension 3: Factorize $\boldsymbol{W}$

- $\boldsymbol{W} \in \mathbb{R}^{M \times NK}$
- $100 \times 17,770 \times 5 = 8,885,000$

Extension 3: Factorize $\boldsymbol{W}$

- $\boldsymbol{W} \in \mathbb{R}^{M \times NK}$
- $100 \times 17,770 \times 5 = 8,885,000$
- Factorize $\boldsymbol{W}$ into $\boldsymbol{PQ}$
- $\boldsymbol{P} \in \mathbb{R}^{M \times C}$, $\boldsymbol{Q} \in \mathbb{R}^{C \times NK}$
- $C \ll M$ and $C \ll NK$

Extension 3: Factorize $\boldsymbol{W}$

- $\boldsymbol{W} \in \mathbb{R}^{M \times NK}$
- $100 \times 17,770 \times 5 = 8,885,000$
- Factorize $\boldsymbol{W}$ into $\boldsymbol{PQ}$
- $\boldsymbol{P} \in \mathbb{R}^{M \times C}$, $\boldsymbol{Q} \in \mathbb{R}^{C \times NK}$
- $C \ll M$ and $C \ll NK$
- $100 \times 30 + 30 \times 17,770 \times 5 = 2,668,500$

Experimental Results

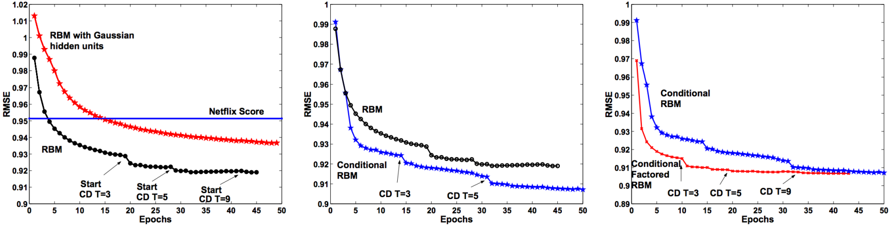


Figure 3. Performance of various models on the validation data. Left panel: RBM vs. RBM with Gaussian hidden units. Middle panel: RBM vs. conditional RBM. Right panel: conditional RBM vs. conditional factored RBM. The y-axis displays RMSE (root mean squared error), and the x-axis shows the number of epochs, or passes through the entire training dataset.

Questions?

Thank you!