Supplementary Material for
"Supervised Restricted Boltzmann Machines"

Tu Dinh Nguyen, Dinh Phung, Viet Huynh, Trung Le
Center for Pattern Recognition and Data Analytics,
Deakin University, Australia.
{tu.nguyen, dinh.phung, viet.huynh, trung.l}@deakin.edu.au.

This document presents supplementary material to complement the manuscript entitled "*Supervised Restricted Boltzmann Machines*", accepted at the conference on Uncertainty in Artificial Intelligence (UAI), 2017. The first section provides full derivations of some inferences, followed by the parameter estimation algorithm of our proposed model. The last section presents our additional experimental results.

1 INFERENCE

1.1 Conditional probabilities

A unit is considered active if its value is one. Let $\mathbf{h}_{\neg k}$ denote the state of all hidden units except the k-th one. The conditional distribution of a single hidden unit given the visible units is:

$$\begin{aligned}
p\left(\mathrm{h}_k = 1 \mid \mathbf{v}\right) &= \frac{p\left(\mathrm{h}_k = 1, \mathbf{h}_{\neg k}, \mathbf{v}; \psi\right)}{p\left(\mathbf{h}_{\neg k}, \mathbf{v}; \psi\right)} \\
&= \frac{p\left(\mathrm{h}_k = 1, \mathbf{h}_{\neg k}, \mathbf{v}; \psi\right)}{p\left(\mathrm{h}_k = 0, \mathbf{h}_{\neg k}, \mathbf{v}; \psi\right) + p\left(\mathrm{h}_k = 1, \mathbf{h}_{\neg k}, \mathbf{v}; \psi\right)} \\
&= \frac{\exp\left[-E\left(\mathrm{h}_k = 1, \mathbf{h}_{\neg k}, \mathbf{v}; \psi\right)\right]}{\exp\left[-E\left(\mathrm{h}_k = 0, \mathbf{h}_{\neg k}, \mathbf{v}; \psi\right)\right] + \exp\left[-E\left(\mathrm{h}_k = 1, \mathbf{h}_{\neg k}, \mathbf{v}; \psi\right)\right]} \\
&= \frac{1}{1 + \exp\left[-\mathrm{b}_k - \mathbf{v}^\top \mathbf{w}_{\boldsymbol{\cdot} k}\right]} \\
&= \mathrm{sig}\left(\mathrm{b}_k + \mathbf{v}^\top \mathbf{w}_{\boldsymbol{\cdot} k}\right)
\end{aligned}$$

where we have substituted the joint distribution in Eq. (2) and the energy in Eq. (1) in the main paper into the second and third steps respectively. Recall that $\mathrm{sig}\left(x\right) = {}^{1}\!/_{1+e^{-x}}$ is the logistic sigmoid function. We have also used the property of sRBM that is the hidden units become conditionally independent when the visible units are observed and the label units are unobserved, thus $p\left(\mathrm{h}_k \mid \mathbf{v}\right) = p\left(\mathrm{h}_k \mid \mathbf{h}_{\neg k}, \mathbf{v}\right)$. Likewise, the conditional probability of a single hidden unit being inactive is:

$$p\left(\mathrm{h}_k = 0 \mid \mathbf{v}\right) = \mathrm{sig}\left(-\mathrm{b}_k - \mathbf{v}^\top \mathbf{w}_{\boldsymbol{\cdot} k}\right)$$

1.2 Gibbs sampling

In order to sample the hidden units, we specify the conditional distribution of a single hidden unit, that is:

$$\begin{aligned}
p\left(\mathrm{h}_k \mid \mathbf{h}_{\neg k}, \mathbf{v}, \mathrm{y}\right) &= \frac{p\left(\mathrm{h}_k, \mathbf{h}_{\neg k} \mid \mathbf{v}, \mathrm{y}\right)}{p\left(\mathbf{h}_{\neg k}, \mathbf{v}, \mathrm{y}\right)} \\
&\propto p\left(\mathrm{h}_k, \mathbf{h}_{\neg k} \mid \mathbf{v}, \mathrm{y}\right) \\
&= \frac{p\left(\mathrm{h}_k, \mathbf{h}_{\neg k}, \mathrm{y} \mid \mathbf{v}\right)}{p\left(\mathrm{y} \mid \mathbf{v}\right)} \\
&\propto p\left(\mathrm{h}_k, \mathbf{h}_{\neg k}, \mathrm{y} \mid \mathbf{v}\right) \\
&= p\left(\mathrm{y} \mid \mathrm{h}_k, \mathbf{h}_{\neg k}, \mathbf{v}\right) p\left(\mathrm{h}_k, \mathbf{h}_{\neg k} \mid \mathbf{v}\right) \\
&= p\left(\mathrm{y} \mid \mathbf{h}\right) p\left(\mathrm{h}_k \mid \mathbf{v}\right) p\left(\mathbf{h}_{\neg k} \mid \mathbf{v}\right) \\
&\propto p\left(\mathrm{y} \mid \mathbf{h}\right) p\left(\mathrm{h}_k \mid \mathbf{v}\right)
\end{aligned}$$

Note that $p\left(\mathbf{h}_{\neg k}, \mathbf{v}, \mathrm{y}\right)$, $p\left(\mathrm{y} \mid \mathbf{v}\right)$ and $p\left(\mathbf{h}_{\neg k} \mid \mathbf{v}\right)$ are constant since $\mathbf{h}_{\neg k}, \mathbf{v}$ and y are observed. Thus we can take the proportions as in the second, forth and last steps. As the hidden unit h_k is binary that takes the value of 0 or 1, its posterior distributions read:

$$\begin{aligned}
p\left(\mathrm{h}_k = 1 \mid \mathbf{h}_{\neg k}, \mathbf{v}, \mathrm{y}\right) &\propto p\left(\mathrm{y} \mid \mathrm{h}_k = 1, \mathbf{h}_{\neg k}\right) p\left(\mathrm{h}_k = 1 \mid \mathbf{v}\right) \\
p\left(\mathrm{h}_k = 0 \mid \mathbf{h}_{\neg k}, \mathbf{v}, \mathrm{y}\right) &\propto p\left(\mathrm{y} \mid \mathrm{h}_k = 0, \mathbf{h}_{\neg k}\right) p\left(\mathrm{h}_k = 0 \mid \mathbf{v}\right)
\end{aligned}$$

We now can perform Gibbs sampling by alternatively sampling a hidden unit given the states of all other variables.

1.3 Evidence lower bound (ELBO)

In sRBM, the log-likelihood of data is given by: $\log p\left(\mathbf{v}, \mathrm{y}; \psi, \boldsymbol{\theta}\right) = \log \sum_{\mathbf{h}} p\left(\mathbf{v}, \mathbf{h}, \mathrm{y}; \psi, \boldsymbol{\theta}\right)$. Let us derive a lower bound for the data log-likelihood. According to the Jensen's inequality, we have:

$$\begin{aligned}
\log \sum_{\mathbf{h}} p\left(\mathbf{v}, \mathbf{h}, \mathrm{y}; \psi, \boldsymbol{\theta}\right) &= \log \sum_{\mathbf{h}} q\left(\mathbf{h}\right) \frac{p\left(\mathbf{v}, \mathbf{h}, \mathrm{y}; \psi, \boldsymbol{\theta}\right)}{q\left(\mathbf{h}\right)} \\
&\geq \sum_{\mathbf{h}} q\left(\mathbf{h}\right) \log \frac{p\left(\mathbf{v}, \mathbf{h}, \mathrm{y}; \psi, \boldsymbol{\theta}\right)}{q\left(\mathbf{h}\right)} \\
&= \mathbb{E}_{q(\mathbf{h})}\left[\log p\left(\mathbf{v}, \mathbf{h}, \mathrm{y}; \psi, \boldsymbol{\theta}\right)\right] - \mathbb{E}_{q(\mathbf{h})}\left[\log q\left(\mathbf{h}\right)\right] \\
&= \mathcal{L}\left(\psi, \boldsymbol{\theta}\right)
\end{aligned}$$

for any arbitrary distribution $q\left(\mathbf{h}\right)$ and $\mathcal{L}\left(\psi, \boldsymbol{\theta}\right)$ is the evidence lower bound (ELBO) of the log-likelihood. We aim to approximate the posterior $p\left(\mathbf{h} \mid \mathbf{v}, \mathrm{y}; \psi, \boldsymbol{\theta}\right)$ by the distribution $q\left(\mathbf{h}; \boldsymbol{\mu}\right)$ with $\boldsymbol{\mu} = \left[\mu_1, \mu_2, ..., \mu_{\mathrm{K}}\right]^{\top}$ is the vector of variational parameters. Using a naive mean-field approximation, we choose a variational distribution that is fully factorized into K Bernoulli distributions as: $q\left(\mathbf{h}; \boldsymbol{\mu}\right) = \prod_{k=1}^{\mathrm{K}} q\left(\mathrm{h}_k; \mu_k\right)$ in which μ_k denotes the probability $q\left(\mathrm{h}_k = 1\right)$. The ELBO now reads:

$$\begin{aligned}
\mathcal{L}\left(\psi, \boldsymbol{\theta}, \boldsymbol{\mu}\right) &= \mathbb{E}_{q(\mathbf{h})}\left[\boldsymbol{\theta}^{\top}\boldsymbol{\phi}\left(\mathrm{y}, \mathbf{h}\right) - B\left(\boldsymbol{\theta}, \mathbf{h}\right) - E\left(\mathbf{v}, \mathbf{h}; \psi\right) - A\left(\psi\right)\right] \\
&\quad + \log t\left(\mathrm{y}\right) - \sum_{k=1}^{\mathrm{K}} \mathbb{E}_{q(\mathrm{h}_k)}\left[\log q\left(\mathrm{h}_k; \mu_k\right)\right] \\
&= \sum_{k=1}^{\mathrm{K}} \mathbb{E}_{q(\mathrm{h}_k)}\left[\theta_k \phi\left(\mathrm{y}, \mathrm{h}_k\right)\right] - \mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right)\right] - \sum_{k=1}^{\mathrm{K}} \mathbb{E}_{q(\mathrm{h}_k)}\left[E\left(\mathbf{v}, \mathrm{h}_k; \psi\right)\right] \\
&\quad - \sum_{k=1}^{\mathrm{K}}\left[\mu_k \log \mu_k + \left(1 - \mu_k\right)\log\left(1 - \mu_k\right)\right] - A\left(\psi\right) + \log t\left(\mathrm{y}\right) \\
&= \sum_{k=1}^{\mathrm{K}} \theta_k\left[\mu_k \phi\left(\mathrm{y}, \mathrm{h}_k = 1\right) + \left(1 - \mu_k\right)\phi\left(\mathrm{y}, \mathrm{h}_k = 0\right)\right] - \mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right)\right] \\
&\quad + \sum_{k=1}^{\mathrm{K}}\left[\mu_k\left(\sum_{n=1}^{\mathrm{N}} \mathrm{a}_n \mathrm{v}_n + \mathrm{b}_k + \sum_{n=1}^{\mathrm{N}} \mathrm{v}_n \mathrm{w}_{nk}\right) + \left(1 - \mu_k\right)\left(\sum_{n=1}^{\mathrm{N}} \mathrm{a}_n \mathrm{v}_n\right)\right] \\
&\quad - \sum_{k=1}^{\mathrm{K}}\left[\mu_k \log \mu_k + \left(1 - \mu_k\right)\log\left(1 - \mu_k\right)\right] - A\left(\psi\right) + \log t\left(\mathrm{y}\right) \\
&= \sum_{k=1}^{\mathrm{K}} \theta_k\left[\mu_k \phi\left(\mathrm{y}, \mathrm{h}_k = 1\right) + \left(1 - \mu_k\right)\phi\left(\mathrm{y}, \mathrm{h}_k = 0\right)\right] - \mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right)\right] \\
&\quad + \sum_{k=1}^{\mathrm{K}} \mu_k\left(\mathrm{b}_k + \sum_{n=1}^{\mathrm{N}} \mathrm{v}_n \mathrm{w}_{nk}\right) + \mathrm{K}\left(\sum_{n=1}^{\mathrm{N}} \mathrm{a}_n \mathrm{v}_n\right) + \log t\left(\mathrm{y}\right) \\
&\quad - \sum_{k=1}^{\mathrm{K}}\left[\mu_k \log \mu_k + \left(1 - \mu_k\right)\log\left(1 - \mu_k\right)\right] - A\left(\psi\right)
\end{aligned}$$

1.4 Expectation of log-partition function

We now present the full derivation of the second-order Taylor series approximation to approximate the expectation of log-partition function $\mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right)\right]$ as in Eq. (12). Recall that, given a twice differentiable function $f\left(\mathbf{x}\right)$ of M-dimensional vector $\mathbf{x} = \left[\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{M}\right]^{\top} \in \mathbb{R}^{\mathrm{N}}$, the second-order Taylor series for expectation approximation evaluated at the first moment $\boldsymbol{\pi} = \mathbb{E}\left[\mathbf{x}\right]$ is given by:

$$\begin{aligned}
\mathbb{E}\left[f\left(\mathbf{x}\right)\right] &\approx \mathbb{E}\left[f\left(\boldsymbol{\pi}\right) + f^{'}\left(\boldsymbol{\pi}\right)\left(\mathbf{x} - \boldsymbol{\pi}\right) + \frac{1}{2}\left(\mathbf{x} - \boldsymbol{\pi}\right)^{\top}\mathbf{H}\left[f\left(\boldsymbol{\pi}\right)\right]\left(\mathbf{x} - \boldsymbol{\pi}\right)\right] \\
&= f\left(\boldsymbol{\pi}\right) + \mathbb{E}\left[\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\pi}\right)^{\top}\mathbf{H}\left[f\left(\boldsymbol{\pi}\right)\right]\left(\mathbf{x} - \boldsymbol{\pi}\right)\right]
\end{aligned}$$

wherein $\mathbf{H}\left[f\left(\boldsymbol{\pi}\right)\right] \in \mathbb{R}^{\mathrm{M}\times\mathrm{M}}$ denotes the second derivative matrix called Hessian matrix of f evaluated at the mean $\boldsymbol{\mu}$ with $\mathrm{H}_{ij} = \partial_{\mathrm{x}_i}\partial_{\mathrm{x}_j} f\left(\mathbf{x}\right)_{|\mathbf{x}=\boldsymbol{\pi}}$. Note that we have used $\mathbb{E}\left[\mathbf{x} - \boldsymbol{\pi}\right] = 0$ in the last step.

Applying this approximation to the expectation of log-partition function $B\left(\boldsymbol{\theta}, \mathbf{h}\right)$, we obtain:

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right)\right] &= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \mathbb{E}_{q(\mathbf{h})}\left[\frac{1}{2}\left(\mathbf{h}-\boldsymbol{\mu}\right)^{\top} \mathbf{H}\left[B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right)\right]\left(\mathbf{h}-\boldsymbol{\mu}\right)\right] \\
&\overset{(a)}{=} B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\mathbb{E}_{q(\mathbf{h})}\left[\sum_{i=1}^{\mathrm{K}}\sum_{j=1}^{\mathrm{K}}\left(\mathrm{h}_i-\mu_i\right)\underbrace{\partial_{\mathrm{h}_i}\partial_{\mathrm{h}_j}B\left(\boldsymbol{\theta}, \mathbf{h}\right)_{|\mathbf{h}=\boldsymbol{\mu}}}_{\triangleq \mathrm{H}_{ij}}\left(\mathrm{h}_j-\mu_j\right)\right] \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\mathbb{E}_{q(\mathbf{h})}\left[\sum_{i=1}^{\mathrm{K}}\sum_{j=1}^{\mathrm{K}}\mathrm{H}_{ij}\left(\mathrm{h}_i-\mu_i\right)\left(\mathrm{h}_j-\mu_j\right)\right] \\
&\overset{(b)}{=} B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\sum_{i=1}^{\mathrm{K}}\sum_{j=1}^{\mathrm{K}}\mathrm{H}_{ij}\mathbb{E}_{q(\mathbf{h})}\left[\left(\mathrm{h}_i-\mu_i\right)\left(\mathrm{h}_j-\mu_j\right)\right] \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\sum_{i=1}^{\mathrm{K}}\sum_{j=1}^{\mathrm{K}}\mathrm{H}_{ij}\sum_{\mathbf{h}}q\left(\mathbf{h}\right)\left[\left(\mathrm{h}_i-\mu_i\right)\left(\mathrm{h}_j-\mu_j\right)\right] \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\sum_{i}\sum_{j\neq i}\mathrm{H}_{ij}\underbrace{\sum_{\mathbf{h}_{\neg\{i,j\}}}q\left(\mathbf{h}_{\neg\{i,j\}}\right)}_{=1}\sum_{\mathrm{h}_i,\mathrm{h}_j}q\left(\mathrm{h}_i\right)q\left(\mathrm{h}_j\right)\left[\left(\mathrm{h}_i-\mu_i\right)\left(\mathrm{h}_j-\mu_j\right)\right] \\
&\quad + \frac{1}{2}\sum_{i=1}^{\mathrm{K}}\mathrm{H}_{ii}\underbrace{\sum_{\mathbf{h}_{\neg i}}q\left(\mathbf{h}_{\neg i}\right)}_{=1}\sum_{\mathrm{h}_i}q\left(\mathrm{h}_i\right)\left[\left(\mathrm{h}_i-\mu_i\right)^2\right] \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\sum_{i}\sum_{j\neq i}\mathrm{H}_{ij}\sum_{\mathrm{h}_i}q\left(\mathrm{h}_i\right)\left(\mathrm{h}_i-\mu_i\right)\sum_{\mathrm{h}_j}q\left(\mathrm{h}_j\right)\left(\mathrm{h}_j-\mu_j\right) \\
&\quad + \frac{1}{2}\sum_{i=1}^{\mathrm{K}}\mathrm{H}_{ii}\sum_{\mathrm{h}_i}q\left(\mathrm{h}_i\right)\left[\left(\mathrm{h}_i-\mu_i\right)^2\right] \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\sum_{i=1}^{\mathrm{K}}\sum_{j=1}^{\mathrm{K}}\mathrm{H}_{ij}\underbrace{\mathbb{E}_{q(\mathrm{h}_i)}\left[\mathrm{h}_i-\mu_i\right]}_{=0}\underbrace{\mathbb{E}_{q(\mathrm{h}_j)}\left[\mathrm{h}_j-\mu_j\right]}_{=0} \\
&\quad + \frac{1}{2}\sum_{i=1}^{\mathrm{K}}\mathrm{H}_{ii}\mathbb{E}_{q(\mathrm{h}_i)}\left[\left(\mathrm{h}_i-\mu_i\right)^2\right] \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2}\sum_{i=1}^{\mathrm{K}}\mathrm{H}_{ii}\mu_i\left(1-\mu_i\right)
\end{aligned}
\tag{1}
$$

Note that we have denoted $\mathrm{H}_{ij} \triangleq \partial_{\mathrm{h}_i}\partial_{\mathrm{h}_j}B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right)$ that is the second-order derivative of $B\left(\boldsymbol{\theta}, \mathbf{h}\right)$ evaluated at $\mathbf{h} = \boldsymbol{\mu}$ in step (a), thus H_{ij} is constant w.r.t $\mathbf{h}$ and can be taken out from the expectation as in step (b). Now we have to specify H_{ii} – the element on the diagonal of Hessian matrix $\mathbf{H}$. This term depends on the form of log-partition function $B\left(\boldsymbol{\theta}, \mathbf{h}\right)$ of the density $p\left(\mathrm{y} \mid \mathbf{h}; \boldsymbol{\theta}\right)$.

1.4.1 Classification

For multiclass classification, assume that y follows a multinomial distribution with C possible outcomes parameterized by the *softmax* probability:

$$
\lambda_c = \frac{\exp\left(\boldsymbol{\theta}_{\bullet c}^{\top}\mathbf{h}\right)}{\sum_{t=1}^{\mathrm{C}}\exp\left(\boldsymbol{\theta}_{\bullet t}^{\top}\mathbf{h}\right)}
\tag{2}
$$

The conditional distribution and log-partition function can be written as follows:

$$
\begin{aligned}
p\left(\mathrm{y} \mid \mathbf{h}; \boldsymbol{\theta}\right) &= \exp\left\{ \mathbf{z}_{\mathrm{y}}^{\top} \boldsymbol{\theta}^{\top} \mathbf{h} - \log \sum_{c=1}^{\mathrm{C}} \exp\left(\boldsymbol{\theta}_{\bullet c}^{\top} \mathbf{h} \right) \right\} \\
B\left(\boldsymbol{\theta}, \mathbf{h}\right) &= \log \sum_{c=1}^{\mathrm{C}} \exp\left(\boldsymbol{\theta}_{\bullet c}^{\top} \mathbf{h} \right)
\end{aligned}
$$

where $\mathbf{z}_{\mathrm{y}}$ is random variable that has the one-hot representation of C-length vector with all zeros but one at y-th position and we have omitted the bias terms β for clarity of presentation. We reparameterize the set $\{\boldsymbol{\theta}, \mathbf{h}\}$ by the parameter $\boldsymbol{\delta}$ as a function $\boldsymbol{\delta} = \boldsymbol{\theta}^{\top} \mathbf{h}$ with $\boldsymbol{\delta} \in \mathbb{R}^{\mathrm{C} \times 1}$. Two functions now become:

$$
\begin{aligned}
p\left(\mathrm{y} \mid \boldsymbol{\delta}\right) &= \exp\left\{ \mathbf{z}_{\mathrm{y}}^{\top} \boldsymbol{\delta} - \log \sum_{c=1}^{\mathrm{C}} \exp\left(\delta_c\right) \right\} \\
B^{\star}\left(\boldsymbol{\delta}\right) &= \log \sum_{c=1}^{\mathrm{C}} \exp\left(\delta_c\right)
\end{aligned} \tag{3}
$$

wherein $B^{\star}\left(\boldsymbol{\delta}\right)$ denotes the new log-partition function.

Using the chain rule for computing the derivative, we obtain:

$$
\begin{aligned}
\frac{\partial^2 B\left(\boldsymbol{\theta}, \mathbf{h}\right)}{\partial_{\mathrm{h}_i} \partial_{\mathrm{h}_j}} &= \sum_{c=1}^{\mathrm{C}} \sum_{t=1}^{\mathrm{C}} \left(\frac{\partial \delta_c}{\partial \mathrm{h}_i} \right) \frac{\partial^2 B^{\star}\left(\boldsymbol{\delta}\right)}{\partial \delta_c \partial \delta_t} \left(\frac{\partial \delta_t}{\partial \mathrm{h}_j} \right) \\
&= \sum_{c=1}^{\mathrm{C}} \sum_{t=1}^{\mathrm{C}} \theta_{ic} \frac{\partial^2 B^{\star}\left(\boldsymbol{\delta}\right)}{\partial \delta_c \partial \delta_t} \theta_{jt} \\
&= \sum_{c=1}^{\mathrm{C}} \sum_{t=1}^{\mathrm{C}} \mathrm{J}_{ci} \mathrm{H}_{ct}^{\star} \mathrm{J}_{tj} \\
&= \mathbf{J}^{\top} \mathbf{H}^{\star} \mathbf{J}
\end{aligned} \tag{4}
$$

where $\mathrm{J}_{ci} = \frac{\partial \delta_c}{\partial \mathrm{h}_i} = \theta_{ic}$ is the element of Jacobian matrix $\mathbf{J} \in \mathbb{R}^{\mathrm{C} \times \mathrm{K}}$ of all first-order partial derivatives of vector-valued function $\boldsymbol{\delta} = \boldsymbol{\theta}^{\top} \mathbf{h}$, and $\mathrm{H}_{ct}^{\star} = \frac{\partial^2 B^{\star}(\boldsymbol{\delta})}{\partial \delta_c \partial \delta_t}$ is the element of Hessian matrix $\mathbf{H}^{\star} \in \mathbb{R}^{\mathrm{C} \times \mathrm{C}}$ of all second-order partial derivatives of function $B^{\star}\left(\boldsymbol{\delta}\right)$. We can see that the probability density in Eq. (3) follows an exponential family, thus the second derivative of log-partition function is equal to the variance, that is:

$$
\mathbf{H}^{\star} = \frac{\partial^2 B^{\star}\left(\boldsymbol{\delta}\right)}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}} = \mathrm{Cov}\left[\mathbf{z}\right] \tag{5}
$$

in which $\mathrm{Cov}\left[\mathbf{z}\right]$ is the covariance matrix of multinomial distribution with $\mathrm{Var}\left[\mathrm{z}_c\right] = \lambda_c\left(1 - \lambda_c\right)$ and $\mathrm{Cov}\left[\mathrm{z}_c, \mathrm{z}_t\right] = -\lambda_c \lambda_t$. From Eqs. (1,2,4,5), we have:

$$
\begin{aligned}
\mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right)\right] &= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i\left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \sum_{t=1}^{\mathrm{C}} \theta_{ic} \frac{\partial^2 B^{\star}\left(\boldsymbol{\delta}\right)}{\partial \delta_c \partial \delta_t} \theta_{it} \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i\left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}^2 \lambda_c\left(1 - \lambda_c\right) \\
&\quad - \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i\left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \sum_{t \neq c} \theta_{ic} \theta_{it} \lambda_c \lambda_t \\
&= B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i\left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}^2 \lambda_c \\
&\quad - \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i\left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \sum_{t=1}^{\mathrm{C}} \theta_{ic} \theta_{it} \lambda_c \lambda_t
\end{aligned} \tag{6}
$$

Taking the derivative of λ_c w.r.t to μ_k , we obtain:

$$\begin{aligned}
\nabla_{\mu_k} \lambda_c &= \lambda_c \nabla_{\mu_k} \log \lambda_c \\
&= \lambda_c \nabla_{\mu_k} \left[\sum_{i=1}^{\mathrm{K}} \theta_{ic}\mu_i - \log \sum_{t=1}^{\mathrm{C}} \exp \left(\sum_{i=1}^{\mathrm{K}} \theta_{it}\mu_i \right) \right] \\
&= \lambda_c \left[\theta_{kc} - \nabla_{\mu_k} B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) \right] \qquad (7) \\
&= \lambda_c \left[\theta_{kc} - \frac{\sum_{t=1}^{\mathrm{C}} \theta_{kt} e^{\boldsymbol{\theta}^{\top}_{\bullet t}\boldsymbol{\mu}}}{\sum_{l=1}^{\mathrm{C}} e^{\boldsymbol{\theta}^{\top}_{\bullet l}\boldsymbol{\mu}}} \right]
\end{aligned}$$

Let us denote $\alpha_{kc} = \nabla_{\mu_k} \lambda_c$. From Eq. (7), we have: $\nabla_{\mu_k} B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) = \theta_{kc} - {}^{\alpha_{kc}}/_{\lambda_c}$. The derivative of the expectation in Eq. (6) w.r.t μ_k can be computed as:

$$\begin{aligned}
\nabla_{\mu_k} \mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right) \right] &= \frac{\sum_{t=1}^{\mathrm{C}} \theta_{kt} e^{\boldsymbol{\theta}^{\top}_{\bullet t}\boldsymbol{\mu}}}{\sum_{l=1}^{\mathrm{C}} e^{\boldsymbol{\theta}^{\top}_{\bullet l}\boldsymbol{\mu}}} + \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i \left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}^2 \alpha_{kc} \\
&\quad + \frac{1}{2} \left(1 - 2\mu_k\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}^2 \lambda_c \\
&\quad - \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i \left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \sum_{t=1}^{\mathrm{C}} \theta_{ic}\theta_{it} \left(\alpha_{kc}\lambda_t + \alpha_{kt}\lambda_c \right) \\
&\quad - \frac{1}{2} \left(1 - 2\mu_k\right) \sum_{c=1}^{\mathrm{C}} \sum_{t=1}^{\mathrm{C}} \theta_{ic}\theta_{it}\lambda_c\lambda_t \\
&= \frac{\sum_{t=1}^{\mathrm{C}} \theta_{kt} e^{\boldsymbol{\theta}^{\top}_{\bullet t}\boldsymbol{\mu}}}{\sum_{l=1}^{\mathrm{C}} e^{\boldsymbol{\theta}^{\top}_{\bullet l}\boldsymbol{\mu}}} + \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i \left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}^2 \alpha_{kc} \\
&\quad + \frac{1}{2} \left(1 - 2\mu_k\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}^2 \lambda_c \\
&\quad - \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \mu_i \left(1 - \mu_i\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}\lambda_c \sum_{t=1}^{\mathrm{C}} \theta_{it}\alpha_{kt} \\
&\quad - \frac{1}{2} \left(1 - 2\mu_k\right) \sum_{c=1}^{\mathrm{C}} \theta_{ic}\lambda_c \sum_{t=1}^{\mathrm{C}} \theta_{it}\lambda_t \qquad (8)
\end{aligned}$$

1.4.2 Regression

For regression task, the outcome variable follows the following Gaussian distribution with the mean $\lambda = \boldsymbol{\theta}^{\top}\mathbf{h}$. The conditional distribution and log-partition function can be written as follows:

$$\begin{aligned}
p\left(\mathrm{y} \mid \mathbf{h}; \boldsymbol{\theta}\right) &= \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}\mathrm{y}^2 + \mathrm{y}\boldsymbol{\theta}^{\top}\mathbf{h} - \frac{1}{2}\left(\boldsymbol{\theta}^{\top}\mathbf{h}\right)^2 \right] \\
B\left(\boldsymbol{\theta}, \mathbf{h}\right) &= \frac{1}{2}\left(\boldsymbol{\theta}^{\top}\mathbf{h}\right)^2
\end{aligned}$$

wherein we have used unit standard deviation $\sigma = 1$ and omitted bias for clarity of presentation. The second derivative can be easily obtained:

$$\frac{\partial^2 B\left(\boldsymbol{\theta}, \mathbf{h}\right)}{\partial_{\mathrm{h}_i} \partial_{\mathrm{h}_i}} = \theta_i^2$$

The expectation in Eq. (1) now reads:

$$\mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta}, \mathbf{h}\right) \right] = B\left(\boldsymbol{\theta}, \boldsymbol{\mu}\right) + \frac{1}{2} \sum_{i=1}^{\mathrm{K}} \theta_i^2 \mu_i \left(1 - \mu_i\right)$$

Its derivative w.r.t μ_k is:

$$\begin{aligned}\nabla_{\mu_k}\mathbb{E}_{q(\mathbf{h})}\left[B\left(\boldsymbol{\theta},\mathbf{h}\right)\right] &= \nabla_{\mu_k}B\left(\boldsymbol{\theta},\mathbf{h}\right)+\frac{1}{2}\theta_k^2\left(1-2\mu_k\right)\\ &= \theta_k\left(\boldsymbol{\theta}^\top\boldsymbol{\mu}\right)+\frac{1}{2}\theta_k^2\left(1-2\mu_k\right)\end{aligned}$$

2 PARAMETER ESTIMATION

The derivatives of log-likelihood $\log p\left(\mathbf{v},\mathrm{y}\right)=\log\sum_{\mathbf{h}}p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)$ w.r.t parameters ψ and $\boldsymbol{\theta}$ in Eqs. (15, 16) in the manuscript can be derived as follows:

$$\begin{aligned}\frac{\partial\log p\left(\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{\partial\psi} &= \frac{\sum_{\mathbf{h}}\left[\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)-\partial_\psi A\left(\psi\right)\right]p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{\sum_{\mathbf{h}}p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}\\ &= \frac{\sum_{\mathbf{h}}\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{p\left(\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)}-\frac{\partial_\psi A\left(\psi\right)\sum_{\mathbf{h}}p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{p\left(\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)}\\ &= \sum_{\mathbf{h}}p\left(\mathbf{h}\mid\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)-\partial_\psi A\left(\psi\right)\\ &= \mathbb{E}_{p(\mathbf{h}|\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta})}\left[\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)\right]-\mathbb{E}_{p(\mathbf{v},\mathbf{h};\psi,\boldsymbol{\theta})}\left[\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)\right]\\ \frac{\partial\log p\left(\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} &= \frac{\sum_{\mathbf{h}}\left[\boldsymbol{\phi}\left(\mathrm{y},\mathbf{h}\right)-\partial_{\boldsymbol{\theta}}B\left(\boldsymbol{\theta},\mathbf{h}\right)\right]p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{p\left(\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)}\\ &= \mathbb{E}_{p(\mathbf{h}|\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta})}\left[\boldsymbol{\phi}\left(\mathrm{y},\mathbf{h}\right)\right]-\mathbb{E}_{p(\mathbf{h}|\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta})}\left[\partial_{\boldsymbol{\theta}}B\left(\boldsymbol{\theta},\mathbf{h}\right)\right]\end{aligned}$$

The pseudo-code of learning parameters for sRBM using CD-1 is described in Alg. 1.

Algorithm 1 Parameter update of sRBM using CD-1.

```
**Input:** Training data $(\mathbf{v},\mathrm{y})$, current parameters $\{\mathbf{a},\mathbf{b},\mathbf{W},\boldsymbol{\theta},\beta\}$, learning rate $\eta$.
Initialize $\boldsymbol{\mu}^0$ randomly.
Run mean-field updates in Eq. (14) in the main paper until convergence, obtain $\hat{\boldsymbol{\mu}}$.
$\mathbf{v}^0 \leftarrow \mathbf{v}$, $\mathrm{y}^0 \leftarrow \mathrm{y}$ and $\hat{\mathbf{h}}^0 \leftarrow \hat{\boldsymbol{\mu}}$.
$\mathbf{h}^{\langle 0\rangle} \sim p\left(\mathbf{h} \mid \mathbf{v}^{\langle 0\rangle},\mathrm{y}^{\langle 0\rangle}\right)$ with $p\left(\mathrm{h}_k=1 \mid \mathbf{v}^{\langle 0\rangle},\mathrm{y}^{\langle 0\rangle}\right)=\hat{\mu}_k$.
$\hat{\mathbf{v}}^{\langle 1\rangle}=p\left(\mathbf{v}\mid\mathbf{h}^{\langle 0\rangle}\right)$, $\mathbf{v}^{\langle 1\rangle}\sim p\left(\mathbf{v}\mid\mathbf{h}^{\langle 0\rangle}\right)$, $\mathbf{h}^{\langle 1\rangle}\sim p\left(\mathbf{h}\mid\mathbf{v}^{\langle 1\rangle}\right)$.
$\hat{\mathrm{a}}_n \leftarrow \mathrm{a}_n+\eta\left(\mathrm{v}_n^0-\hat{\mathrm{v}}_n^1\right)$ with $n=1,...,\mathrm{N}$.
$\hat{\mathrm{b}}_k \leftarrow \mathrm{b}_k+\eta\left(\hat{\mathrm{h}}_k^0-\mathrm{h}_k^1\right)$ with $k=1,...,\mathrm{K}$.
$\hat{\mathrm{w}}_{nk} \leftarrow \mathrm{w}_{nk}+\eta\left(\mathrm{v}_n^0\hat{\mathrm{h}}_k^0-\hat{\mathrm{v}}_n^1\mathrm{h}_k^1\right)$ with $n=1,...,\mathrm{N}$, $k=1,...,\mathrm{K}$.
$\hat{\theta} \leftarrow \theta+\eta\left[\partial_\theta\left[\boldsymbol{\theta}^\top\boldsymbol{\phi}\left(\mathrm{y}^0,\hat{\boldsymbol{\mu}}\right)\right]-\partial_\theta B\left(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}\right)\right]$
$\hat{\beta} \leftarrow \beta+\eta\left[\partial_\beta\left[\boldsymbol{\theta}^\top\boldsymbol{\phi}\left(\mathrm{y}^0,\hat{\boldsymbol{\mu}}\right)\right]-\partial_\beta B\left(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}\right)\right]$
**Output:** Updated parameters $\left\{\hat{\mathbf{a}},\hat{\mathbf{b}},\hat{\mathbf{W}},\hat{\boldsymbol{\theta}},\hat{\beta}\right\}$.
```

For semi-supervised RBM, the derivatives of log-likelihood $\log p\left(\mathbf{v}\right)=\log\sum_{\mathbf{h},\mathrm{y}}p\left(\mathbf{v},\mathbf{h},\mathrm{y}\right)$ w.r.t parameters ψ and $\boldsymbol{\theta}$ can be derived as follows:

$$\begin{aligned}\frac{\partial\log p\left(\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{\partial\psi} &= \frac{\sum_{\mathbf{h},\mathrm{y}}\left[\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)-\partial_\psi A\left(\psi\right)\right]p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{p\left(\mathbf{v};\psi,\boldsymbol{\theta}\right)}\\ &= \frac{\sum_{\mathbf{h},\mathrm{y}}\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{p\left(\mathbf{v};\psi,\boldsymbol{\theta}\right)}-\partial_\psi A\left(\psi\right)\\ &= \mathbb{E}_{p(\mathbf{h},\mathrm{y}|\mathbf{v};\psi,\boldsymbol{\theta})}\left[\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)\right]-\mathbb{E}_{p(\mathbf{v},\mathbf{h};\psi,\boldsymbol{\theta})}\left[\partial_\psi E\left(\mathbf{v},\mathbf{h};\psi\right)\right]\\ \frac{\partial\log p\left(\mathbf{v},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} &= \frac{\sum_{\mathbf{h},\mathrm{y}}\left[\boldsymbol{\phi}\left(\mathrm{y},\mathbf{h}\right)-\partial_{\boldsymbol{\theta}}B\left(\boldsymbol{\theta},\mathbf{h}\right)\right]p\left(\mathbf{v},\mathbf{h},\mathrm{y};\psi,\boldsymbol{\theta}\right)}{p\left(\mathbf{v};\psi,\boldsymbol{\theta}\right)}\\ &= \mathbb{E}_{p(\mathbf{h},\mathrm{y}|\mathbf{v};\psi,\boldsymbol{\theta})}\left[\boldsymbol{\phi}\left(\mathrm{y},\mathbf{h}\right)\right]-\mathbb{E}_{p(\mathbf{h},\mathrm{y}|\mathbf{v};\psi,\boldsymbol{\theta})}\left[\partial_{\boldsymbol{\theta}}B\left(\boldsymbol{\theta},\mathbf{h}\right)\right]\end{aligned}$$

3 EXPERIMENTAL RESULTS

We first investigate how the number of hidden units (K) of our method affects the predictive performance. Particularly, we vary K in the range of $\{500, 1000, 2000, 4000, 5000, 6000\}$, train the sRBM-1st and sRBM-2nd and record the classification errors on testing set for each value. Note that 500 hidden units for MNIST data are often used in prior literature on the standard RBM. Fig. 1 illustrates the results showing that the larger hidden layers yield better classification results, which is plausible since the hidden units need to capture both data and labels. This finding is also consistent with that of the ClassRBM, thus we ends up with using $6,000$ hidden units for MNIST and $1,000$ for 20 Newsgroups, that are identical to those used for ClassRBM [Larochelle et al., 2012].

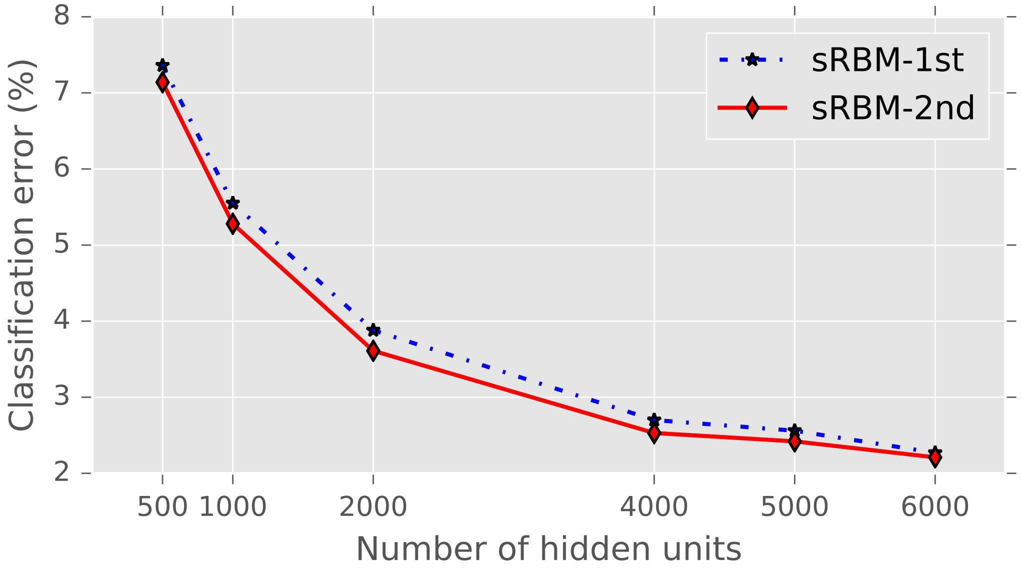

Figure 1: Classification results on MNIST dataset of two sRBM versions with different numbers of hidden units (K).

For document modeling task on 20 Newsgroups dataset, we analyze how our proposed model captures words that are coherent in a topic by examining the weight matrices $\boldsymbol{\theta}$ and $\mathbf{W}$. The entry of column $\mathbf{w}_{\bullet k}$ reflects the association strength of a particular word to the latent factor k, and $\boldsymbol{\theta}_{\bullet c}$ the strength of a latent factor to the topic c. We first specify top 100 hidden units with the largest weight for each topic c, then aggregate (by summing) the associated word-to-hidden weight vectors. This reveals the positive contribution of the words to each newsgroup via the hidden layer. Fig. 2 illustrates the top 8 words per topic, in descending order of their aggregated association strength, discovered by our model. The chart shows that the words under each feature are semantically related in a coherent way.

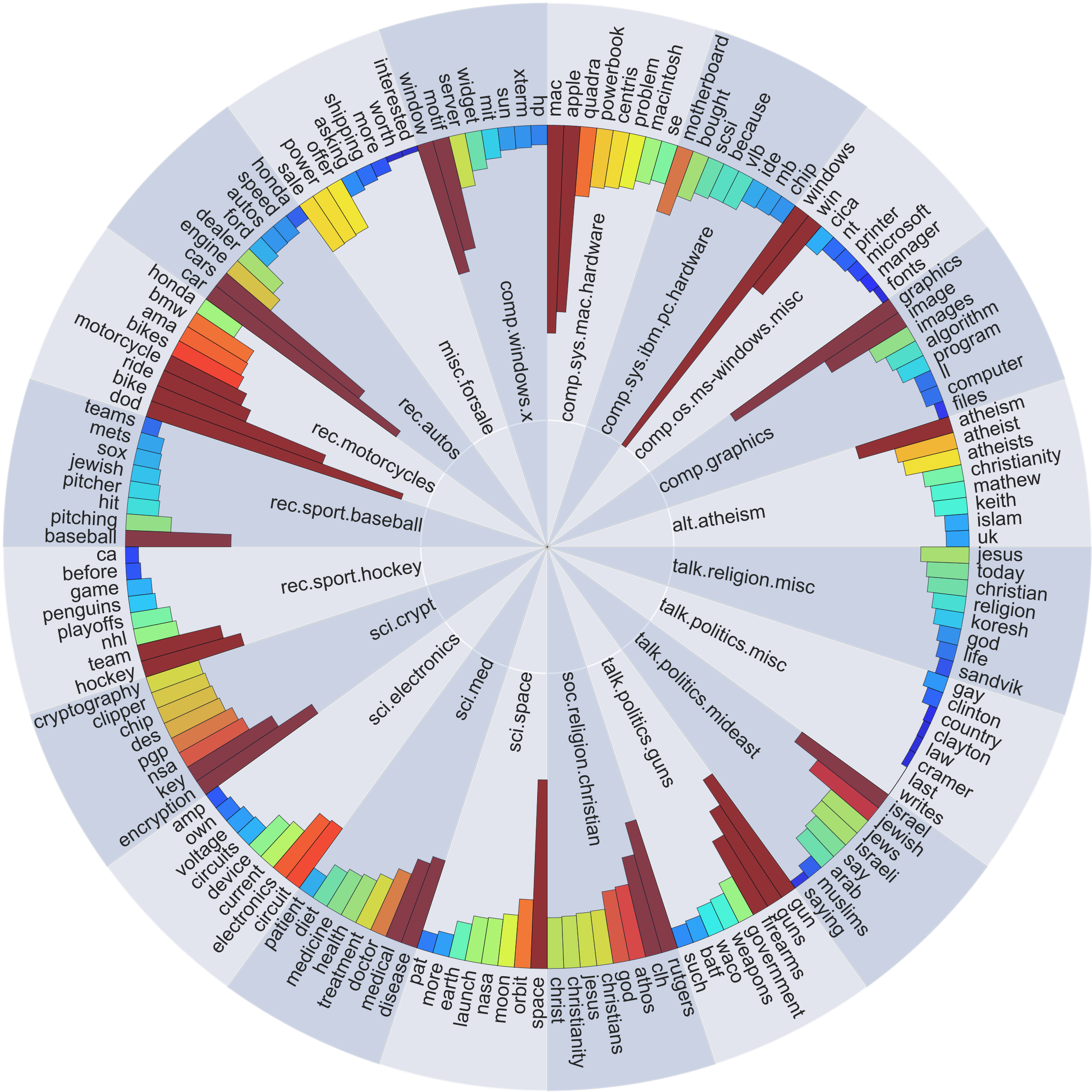

Figure 2: Topics associated with top 8 words discovered by sRBM from 20 Newsgroups dataset. The bar height and color relatively represent the aggregated association strength of a word. (Best viewed in colors).

References

Hugo Larochelle, Michael Mandel, Razvan Pascanu, and Yoshua Bengio. Learning algorithms for the classiffcation restricted boltzmann machine. *Journal of Machine Learning Research*, pages 643–669, 2012.