Sparse Classification RBM

Weidong Liang

Mathematical Formulation

$$\mathrm{E}(\mathrm{y},\mathbf{X},\mathbf{h}) = \mathrm{E}(\mathrm{y},\mathrm{x}_1,\dots,\mathrm{x}_\mathrm{C},\mathbf{h}) = -\sum_{1\le \mathrm{i}\le \mathrm{C}} \left(\mathbf{h}^\mathbf{T}\mathbf{W}^{(\mathrm{i})}\mathbf{e}_{\mathrm{x_i}} + \mathbf{b}^{(\mathrm{i})\mathbf{T}}\mathbf{e}_{\mathrm{x_i}}\right) - \mathbf{h}^\mathbf{T}\mathbf{c} - \mathrm{dy} - \mathbf{h}^\mathbf{T}\mathbf{U}\mathrm{y}$$

$$p(y,\boldsymbol{X},\boldsymbol{h}) = p(y,x_1,\dots,x_C,\boldsymbol{h}) = \frac{\exp(-E(y,x_1,\dots,x_C,\boldsymbol{h}))}{Z}$$

$$\mathrm{Z} = \sum_y \sum_{\boldsymbol{X}} \sum_{\boldsymbol{h}} p(y,\boldsymbol{X},\boldsymbol{h})$$

$$\boldsymbol{e_y} = (1_{i=y})_{1\le i\le C}$$

$$\mathrm{y} \in \{0,1\}$$

$$x_i \in \{1,\dots,k_i\}, 1 \le 1 \le C$$

$$h_i \in \{0,1\}, 1 \le \mathrm{i} \le H$$

Parameter Set: $\theta = \{\mathbf{W},\mathbf{b},\mathrm{d},\mathbf{c},\mathbf{U}\}$

Model Explanation

Y: click (y=1) or no click (y=0).

x_i: feature value of the i-th feature class (there are total of C feature classes).

h_i:i-th hidden unit value.

Derivation of Properties

$$\mathrm{p}(\mathbf{h}|\mathrm{y},\mathbf{X}) = \prod_j p(h_j|y,\boldsymbol{X})$$

$$\mathrm{p}(\mathbf{h}|\mathrm{y},\mathbf{X}) = \frac{p(y,\boldsymbol{X},\boldsymbol{h})}{p(y,\boldsymbol{X})} = \frac{p(y,\boldsymbol{X},\boldsymbol{h})}{\sum_{\mathbf{h}'} p(y,\boldsymbol{X},\boldsymbol{h}')} = \frac{\exp[-E(y,\boldsymbol{X},\boldsymbol{h})]}{\sum_{\mathbf{h}'}\exp[-E(y,\boldsymbol{X},\boldsymbol{h}')]}$$

$$= \frac{\exp\left\{\sum_{1\le \mathrm{i}\le \mathrm{C}}\left(\mathbf{h}^\mathbf{T}\mathbf{W}^{(\mathrm{i})}\mathbf{e}_{\mathrm{x_i}} + \mathbf{b}^{(\mathrm{i})\mathbf{T}}\mathbf{e}_{\mathrm{x_i}}\right) + \mathbf{h}^\mathbf{T}\mathbf{c} + \mathrm{dy} + \mathbf{h}^\mathbf{T}\mathbf{U}\mathrm{y}\right\}}{\sum_{\mathbf{h}'}\exp\left\{\sum_{1\le \mathrm{i}\le \mathrm{C}}\left(\mathbf{h}'^\mathbf{T}\mathbf{W}^{(\mathrm{i})}\mathbf{e}_{\mathrm{x_i}} + \mathbf{b}^{(\mathrm{i})\mathbf{T}}\mathbf{e}_{\mathrm{x_i}}\right) + \mathbf{h}'^\mathbf{T}\mathbf{c} + \mathrm{dy} + \mathbf{h}'^\mathbf{T}\mathbf{U}\mathrm{y}\right\}}$$

$$= \frac{\exp(dy)}{\exp(dy)}\frac{\exp\left(\sum_{1\le i\le C}\mathbf{b}^{(\mathrm{i})\mathbf{T}}\mathbf{e}_{\mathrm{x_i}}\right)}{\exp\left(\sum_{1\le i\le C}\mathbf{b}^{(\mathrm{i})\mathbf{T}}\mathbf{e}_{\mathrm{x_i}}\right)}\frac{\exp\left\{\mathbf{h}^T\left[\sum_{1\le i\le C}\mathbf{W}^{(\mathrm{i})}\mathbf{e}_{\mathrm{x_i}} + \mathbf{c} + \mathbf{U}y\right]\right\}}{\sum_{\mathbf{h}'}\exp\left\{\mathbf{h}'^T\left[\sum_{1\le i\le C}\mathbf{W}^{(\mathrm{i})}\mathbf{e}_{\mathrm{x_i}} + \mathbf{c} + \mathbf{U}y\right]\right\}}$$

$$= \frac{\exp\left\{\sum_{1\le j\le \mathrm{H}}\left[h_j\left(\sum_{1\le i\le C}W_{j,x_i}^{(i)} + c_i + U_i y\right)\right]\right\}}{\sum_{\mathbf{h}'}\exp\left\{\sum_{1\le j\le \mathrm{H}}\left[h'_j\left(\sum_{1\le i\le C}W_{j,x_i}^{(i)} + c_i + U_i y\right)\right]\right\}}$$

$$= \frac{\prod_{1\le j\le H} \exp[h_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}{\sum_{\mathbf{h}'} \prod_{1\le j\le H} \exp[h'_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}$$

$$= \frac{\prod_{1\le j\le H} \exp[h_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}{\sum_{h'_1} \dots \sum_{h'_H} \prod_{1\le j\le H} \exp[h'_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}$$

$$= \frac{\prod_{1\le j\le H} \exp[h_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}{\left\{\sum_{h'_1\in\{0,1\}} \exp[h'_1(\sum_{1\le i\le C} W_{1,x_i}^{(i)} + c_1 + U_1 y)]\right\} \dots \left\{\sum_{h'_H\in\{0,1\}} \exp[h'_H(\sum_{1\le i\le C} W_{H,x_i}^{(i)} + c_H + U_H y)]\right\}}$$

$$= \frac{\prod_{1\le j\le H} \exp[h_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}{\left\{1 + \exp[(\sum_{1\le i\le C} W_{1,x_i}^{(i)} + c_1 + U_1 y)]\right\} \dots \left\{1 + \exp[\sum_{1\le i\le C} W_{H,x_i}^{(i)} + c_H + U_H y]\right\}}$$

$$= \prod_{1\le j\le H} \frac{\exp[h_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}{1 + \exp[(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_j + U_j y)]}$$

$$= \prod_{1\le j\le H} p(h_j|y, \boldsymbol{X})$$

$$\mathrm{p}(h_j = 1 \middle| y, \boldsymbol{X}) = \sigma(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_j + U_j y)$$

$$\mathrm{p}(h_j = 1 \middle| y, \boldsymbol{X}) = \sum_{\boldsymbol{h}'\in\{\dots,h_j=1,\dots\}} p(\boldsymbol{h}|y, \boldsymbol{X})$$

$$= \sum_{\boldsymbol{h}'\in\{\dots,h_j=1,\dots\}} \prod_{1\le k\le H} \frac{\exp[h_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}{1 + \exp[(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_j + U_j y)]}$$

$$= \frac{\exp[h_j(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y)]}{1 + \exp[(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_j + U_j y)]} \prod_{1\le k\le H, k\ne j} \sum_{h_k\in\{0,1\}} \frac{\exp[h_k(\sum_{1\le i\le C} W_{k,x_i}^{(i)} + c_k + U_k y)]}{1 + \exp[(\sum_{1\le i\le C} W_{k,x_i}^{(i)} + c_k + U_k y)]}$$

$$= \frac{\exp[\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_i + U_i y]}{1 + \exp[(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_j + U_j y)]}$$

$$= \sigma(\sum_{1\le i\le C} W_{j,x_i}^{(i)} + c_j + U_j y)$$

$$\mathrm{p}(\mathrm{y}|\mathbf{X}) = \frac{\exp\{-F(y, X)\}}{\sum_{y''} \exp\{-F(y'', X)\}}$$

Note that:

$$\sum_{\mathbf{h}'} p(y, \mathbf{X}, \mathbf{h}') = \frac{1}{Z} \sum_{\mathbf{h}'} \exp(-E(y, \mathbf{X}, \mathbf{h}')$$

$$= \frac{1}{Z} \sum_{\mathbf{h}'} \exp\left\{\sum_{1\le \mathrm{i}\le \mathrm{C}} \left(\mathbf{h}'^{\mathbf{T}} \mathbf{W}^{(\mathrm{i})} \mathbf{e}_{\mathrm{x}_\mathrm{i}} + \mathbf{b}^{(\mathrm{i})\mathbf{T}} \mathbf{e}_{\mathrm{x}_\mathrm{i}}\right) + \mathbf{h}'^{\mathbf{T}} \mathbf{c} + \mathrm{dy} + \mathbf{h}'^{\mathbf{T}} \mathbf{U} \mathrm{y}\right\}$$

$$= \frac{1}{Z}\sum_{\mathbf{h}'}\{\exp(dy)\}\{\exp(\sum_{1\le i\le C}\mathbf{b}^{(i)T}\mathbf{e}_{x_i})\}\left\{\exp\left[\mathbf{h}'^{\mathbf{T}}\left(\sum_{1\le i\le C}\mathbf{W}^{(i)}\mathbf{e}_{x_i}+\mathbf{c}+\mathbf{U}y\right)\right]\right\}$$

$$= \frac{1}{Z}\{\exp(\sum_{1\le i\le C}\mathbf{b}^{(i)T}\mathbf{e}_{x_i})\}\{\exp(dy)\}\sum_{\mathbf{h}'}\left\{\exp\left[\mathbf{h}'^{\mathbf{T}}\left(\sum_{1\le i\le C}\mathbf{W}^{(i)}\mathbf{e}_{x_i}+\mathbf{c}+\mathbf{U}y\right)\right]\right\}$$

$$= \frac{1}{Z}\{\exp(\sum_{1\le i\le C}\mathbf{b}^{(i)T}\mathbf{e}_{x_i})\}\{\exp(dy)\}\prod_{1\le k\le H}\sum_{h_k\in\{0,1\}}\left\{\exp\left[h_k\left(\sum_{1\le i\le C}\boldsymbol{W}_{k,x_i}^{(i)}+c_k+\mathbf{U}_k y\right)\right]\right\}$$

$$= \frac{1}{Z}\{\exp(\sum_{1\le i\le C}\mathbf{b}^{(i)T}\mathbf{e}_{x_i})\}\{\exp(dy)\}\prod_{1\le k\le H}\left\{1+\exp\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y\right)\right\}$$

Hence:

$$p(y|\mathbf{X}) = \frac{p(y,\boldsymbol{X})}{p(\boldsymbol{X})} = \frac{\sum_{\mathbf{h}'}p(y,\mathbf{X},\mathbf{h}')}{\sum_{y''}\sum_{\mathbf{h}''}p(y'',\mathbf{X},\mathbf{h}'')}$$

$$= \frac{\{\exp(dy)\}\prod_{1\le k\le H}\left\{1+\exp\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y\right)\right\}}{\sum_{y''}\{\exp(dy'')\}\prod_{1\le k\le H}\left\{1+\exp\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y''\right)\right\}}$$

$$= \frac{\exp\left\{dy+\ln\left[\prod_{1\le k\le H}\left\{1+\exp\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y\right)\right\}\right]\right\}}{\sum_{y''}\exp\left\{dy''+\ln\left[\prod_{1\le k\le H}\left\{1+\exp\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y''\right)\right\}\right]\right\}}$$

$$= \frac{\exp\left\{dy+\sum_{1\le k\le H}\ln\left[1+\exp\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y\right)\right]\right\}}{\sum_{y''}\exp\left\{dy''+\sum_{1\le k\le H}\ln\left[1+\exp\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y''\right)\right]\right\}}$$

$$= \frac{\exp\left\{dy+\sum_{1\le k\le H}\mathrm{softplus}\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y\right)\right\}}{\sum_{y''}\exp\left\{dy''+\sum_{1\le k\le H}\mathrm{softplus}\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y''\right)\right\}}$$

$$= \frac{\exp\left\{dy+\sum_{1\le k\le H}\mathrm{softplus}\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k y\right)\right\}}{\exp\left\{\sum_{1\le k\le H}\mathrm{softplus}\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k\right)\right\}+\exp\left\{d+\sum_{1\le k\le H}\mathrm{softplus}\left(\sum_{1\le i\le C}W_{k,x_i}^{(i)}+c_k+\mathrm{U}_k\right)\right\}}$$

$$= \frac{\exp\{-F(y,X)\}}{\sum_{y''}\exp\{-F(y'',X)\}}$$

$$p(y|\mathbf{h}) = \frac{\exp\{dy+\mathbf{h}^{T}\mathbf{U}y\}}{1+\exp\{d+\mathbf{h}^{T}\mathbf{U}\}}$$

Note that

$$\sum_{X}p(y,\mathbf{X},\mathbf{h}) = \frac{1}{Z}\sum_{\mathbf{X}}\exp\left\{\sum_{1\le i\le C}\left(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(i)}+\mathbf{b}^{(i)T}\right)\mathbf{e}_{x_i}\right\}\exp\{\mathbf{h}^{T}\mathbf{c}\}\exp\{dy+\mathbf{h}^{T}\mathbf{U}y\}$$

$$= \frac{1}{Z}\exp\{\mathbf{h}^{T}\mathbf{c}\}\exp\{dy+\mathbf{h}^{T}\mathbf{U}y\}\sum_{\mathbf{X}}\exp\left\{\sum_{1\le i\le C}\left(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(i)}+\mathbf{b}^{(i)T}\right)\mathbf{e}_{x_i}\right\}$$

$$= \frac{1}{\mathrm{Z}} \exp\{\mathbf{h}^{\boldsymbol{T}}\mathbf{c}\} \exp\{\mathrm{dy} + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}\} \prod_{1 \le \mathrm{i} \le \mathrm{C}} \sum_{\mathrm{x_i}} \exp\{(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathbf{i})} + \mathbf{b}^{(\mathbf{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}}\}$$

$$= \frac{1}{\mathrm{Z}} \exp\{\mathbf{h}^{\boldsymbol{T}}\mathbf{c}\} \exp\{\mathrm{dy} + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}\} \mathrm{K}(\mathbf{h})$$

Hence:

$$\mathrm{p}(\mathrm{y}|\mathbf{h}) = \frac{\mathrm{p}(\mathrm{y}, \mathbf{h})}{\mathrm{p}(\mathbf{h})} = \frac{\sum_{\mathbf{X}'} \mathrm{p}(\mathrm{y}, \mathbf{X}', \mathbf{h})}{\sum_{\mathrm{y}''} \sum_{\mathbf{X}''} \mathrm{p}(\mathrm{y}'', \mathbf{X}'', \mathbf{h})}$$

$$= \frac{\exp\{\mathbf{h}^{\boldsymbol{T}}\mathbf{c}\} \exp\{\mathrm{dy} + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}\} \mathrm{K}(\mathbf{h})}{\sum_{\mathrm{y}''} \exp\{\mathbf{h}^{\boldsymbol{T}}\mathbf{c}\} \exp\{\mathrm{dy}'' + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}\} \mathrm{K}(\mathbf{h})} = \frac{\exp\{\mathrm{dy} + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}\}}{\sum_{\mathrm{y}''} \exp\{\mathrm{dy}'' + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}''\}} = \frac{\exp\{\mathrm{dy} + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}\}}{1 + \exp\{\mathrm{d} + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\}}$$

$$\mathrm{p}(\mathrm{y} = 1|\mathbf{h}) = \sigma(\mathrm{d} + \mathbf{h}^{\boldsymbol{T}}\mathbf{U})$$

$$\mathbf{p}(\mathbf{X}|\mathbf{h}) = \prod_{1 \le \mathrm{i} \le \mathrm{C}} p(x_i|\mathbf{h})$$

Note that:

$$\sum_{y'} p(y', \mathbf{X}, \mathbf{h}) = \frac{1}{Z} \sum_{y'} \exp\left\{ \sum_{1 \le \mathrm{i} \le \mathrm{C}} \left(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})}\mathbf{e}_{\mathrm{x_i}} + \mathbf{b}^{(\mathrm{i})\mathbf{T}}\mathbf{e}_{\mathrm{x_i}}\right) + \mathbf{h}^{\boldsymbol{T}}\mathbf{c} + \mathrm{dy}' + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}' \right\}$$

$$= \frac{1}{Z} \sum_{y'} \exp\left\{ \sum_{1 \le \mathrm{i} \le \mathrm{C}} (\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})} + \mathbf{b}^{(\mathrm{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}} \right\} \exp\{\mathbf{h}^{\boldsymbol{T}}\mathbf{c}\} \exp\{\mathrm{dy}' + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}'\}$$

$$= \frac{1}{Z} \left[\exp\left\{ \sum_{1 \le \mathrm{i} \le \mathrm{C}} (\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})} + \mathbf{b}^{(\mathrm{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}} \right\} \exp\{\mathbf{h}^{\boldsymbol{T}}\mathbf{c}\} \right] \sum_{y'} \exp\{\mathrm{dy}' + \mathbf{h}^{\boldsymbol{T}}\mathbf{U}\mathrm{y}'\}$$

$$= \frac{1}{Z} [1 + \exp(d + \mathbf{h}^{\boldsymbol{T}}\mathbf{U})] \exp\{\mathbf{h}^{\boldsymbol{T}}\mathbf{c}\} \exp\left\{ \sum_{1 \le \mathrm{i} \le \mathrm{C}} (\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})} + \mathbf{b}^{(\mathrm{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}} \right\}$$

Hence:

$$\mathrm{p}(\mathbf{X}|\mathbf{h}) = \frac{p(\mathbf{X}, \mathbf{h})}{p(\mathbf{h})} = \frac{\sum_{y'} p(y', \mathbf{X}, \mathbf{h})}{\sum_{X''} \sum_{y''} p(y'', \mathbf{X}'', \mathbf{h})}$$

$$= \frac{\exp\{\sum_{1 \le \mathrm{i} \le \mathrm{C}} (\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})} + \mathbf{b}^{(\mathrm{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}}\}}{\sum_{X''} \exp\left\{\sum_{1 \le \mathrm{i} \le \mathrm{C}} (\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})} + \mathbf{b}^{(\mathrm{i})\mathbf{T}})\mathbf{e}_{x_i''}\right\}}$$

$$= \frac{\prod_{1 \le \mathrm{i} \le \mathrm{C}} \exp\{(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})} + \mathbf{b}^{(\mathrm{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}}\}}{\prod_{1 \le \mathrm{i} \le \mathrm{C}} \sum_{x_i'' \in \{1, \dots, C_i\}} \exp\{(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})} + \mathbf{b}^{(\mathrm{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}}\}}$$

$$= \prod_{1 \le \mathrm{i} \le \mathrm{C}} p(x_i|h)$$

$$\mathrm{p}(x_i = k|\mathbf{h}) = \frac{\exp(\sum_{1\le j\le H} h_j W_{j,k}^{(i)})}{\sum_{1\le \mathrm{q}\le \mathrm{C_i}} \exp(\sum_{1\le j\le H} h_j W_{j,q}^{(i)})}$$

$$\mathrm{p}(x_i = k|\mathbf{h}) = \frac{\exp\{(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathbf{i})} + \mathbf{b}^{(\mathbf{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}}\}}{\sum_{x_i''\in\{1,\ldots,C_i\}} \exp\{(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathbf{i})} + \mathbf{b}^{(\mathbf{i})\mathbf{T}})\mathbf{e}_{\mathrm{x_i}}\}}$$

$$= \frac{\exp(\sum_{1\le j\le H} h_j W_{j,k}^{(i)})}{\sum_{1\le \mathrm{q}\le \mathrm{C_i}} \exp(\sum_{1\le j\le H} h_j W_{j,q}^{(i)})}$$

Derivation of Learning Algorithm

Generative Learning

$$\frac{\partial \ln p(y, \mathbf{X})}{\partial\theta} = -\mathbb{E}_{p(\mathbf{h}|y,\mathbf{X})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right] + \mathbb{E}_{p(y,\mathbf{X},\mathbf{h})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right]$$

$$\frac{\partial \ln p(y, \mathbf{X})}{\partial\theta} = \frac{\partial}{\partial\theta}\left\{\ln \sum_{\mathbf{h}'} \exp(-E(y, \mathbf{X}, \mathbf{h}') - \ln Z\right\}$$

$$= -\sum_{\mathbf{h}'}\left[\frac{exp(-E(y, \mathbf{X}, \mathbf{h}')}{\sum_{\mathbf{h}''} exp(-E(y, \mathbf{X}, \mathbf{h}''))}\frac{\partial E(y, \mathbf{X}, \mathbf{h}')}{\partial\theta}\right] + \sum_{y',\mathbf{X},\mathbf{h}'}\left[\frac{\exp(-E(y', \mathbf{X}', \mathbf{h}')}{Z}\frac{\partial E(y', \mathbf{X}', \mathbf{h}')}{\partial\theta}\right]$$

$$= -\sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})\frac{\partial E(y, \mathbf{X}, \mathbf{h}')}{\partial\theta} + \sum_{y',\mathbf{X}',\mathbf{h}'} p(y', \mathbf{X}', \mathbf{h}')\frac{\partial E(y', \mathbf{X}', \mathbf{h}')}{\partial\theta}$$

Discriminative Learning

$$\frac{\partial \ln p(y|\boldsymbol{X})}{\partial\theta} = -\mathbb{E}_{p(\mathbf{h}|y,\mathbf{X})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right] + \mathbb{E}_{p(y,\mathbf{h}|\mathbf{X})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right]$$

$$\frac{\partial \ln p(y|\boldsymbol{X})}{\partial\theta} = \frac{\partial}{\partial\theta}\ln\left[\frac{\frac{1}{Z}\sum_{\mathbf{h}'} p(y, \mathbf{X}, \mathbf{h}')}{\frac{1}{Z}\sum_{y'}\sum_{\mathbf{h}'} p(y', \mathbf{X}, \mathbf{h}')}\right]$$

$$= \frac{\partial}{\partial\theta}\left\{\ln\left[\sum_{\mathbf{h}'} p(y, \mathbf{X}, \mathbf{h}')\right] - \ln\left[\sum_{y'}\sum_{\mathbf{h}'} p(y', \mathbf{X}, \mathbf{h}')\right]\right\}$$

$$= -\sum_{\mathbf{h}'}\left\{\frac{p(y, \mathbf{X}, \mathbf{h}')}{\sum_{h''} p(y, \mathbf{X}, \mathbf{h}'')}\frac{\partial E(y, \mathbf{X}, \mathbf{h}')}{\partial\theta}\right\} + \sum_{y'}\sum_{\mathbf{h}'}\left\{\frac{p(y',\ \mathbf{X}, \mathbf{h}')}{\sum_{y''}\sum_{h''} p(y'', \mathbf{X}, \mathbf{h}'')}\frac{\partial E(y', \mathbf{X}, \mathbf{h}')}{\partial\theta}\right\}$$

$$= -\sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})\frac{\partial E(y, \mathbf{X}, \mathbf{h}')}{\partial\theta} + \sum_{y'}\sum_{\mathbf{h}'} p(y', \mathbf{h}'|\mathbf{X})\frac{\partial E(y', \mathbf{X}, \mathbf{h}')}{\partial\theta}$$

Hybrid Learning

$$\frac{\ln Hybrid(\alpha, y, \mathbf{X})}{\partial\theta} = \frac{\partial \ln\ p(y|\mathbf{X})}{\partial\theta} + \alpha\frac{\partial \ln p(y, \mathbf{X})}{\partial\theta}$$

$$= -(1+\alpha)\mathbb{E}_{p(\mathbf{h}|y,\mathbf{X})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right] + \mathbb{E}_{p(y,\mathbf{h}|\mathbf{X})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right] + \alpha\mathbb{E}_{p(y,\mathbf{X},\mathbf{h})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right]$$

$$= -(1+\alpha)\mathrm{A}(\mathrm{y}, \mathbf{X}, \mathbf{h}, \theta) + \mathrm{B}(\mathrm{y}, \mathbf{X}, \mathbf{h}, \theta) + \alpha C(\mathrm{y}, \mathbf{X}, \mathbf{h}, \theta)$$

Derivatives of Gradients

$$\mathrm{E}(\mathrm{y}, \mathbf{X}, \mathbf{h}) = \mathrm{E}(\mathrm{y}, \mathrm{x}_1, \dots, \mathrm{x}_C, \mathbf{h}) = -\sum_{1 \le \mathrm{i} \le \mathrm{C}} \left(\mathbf{h}^{\mathbf{T}}\mathbf{W}^{(\mathrm{i})}\mathbf{e}_{\mathrm{x}_\mathrm{i}} + \mathbf{b}^{(\mathrm{i})\mathbf{T}}\mathbf{e}_{\mathrm{x}_\mathrm{i}}\right) - \mathbf{h}^{\mathbf{T}}\mathbf{c} - \mathrm{dy} - \mathbf{h}^{\mathbf{T}}\mathbf{U}\mathrm{y}$$

Parameter Set: $\theta = \{\mathbf{W}, \mathbf{b}, \mathrm{d}, \mathbf{c}, \mathbf{U}\}$

$$\frac{\partial \mathrm{E}(\mathrm{y}, \mathbf{X}, \mathbf{h})}{\partial W_{j,k}^{(i)}} = -h_j 1_{(x_i=k)}$$

$$\frac{\partial \mathrm{E}(\mathrm{y}, \mathbf{X}, \mathbf{h})}{\partial b_j^i} = -1_{(x_i=j)}$$

$$\frac{\partial \mathrm{E}(\mathrm{y}, \mathbf{X}, \mathbf{h})}{\partial \mathrm{d}} = -y$$

$$\frac{\partial \mathrm{E}(\mathrm{y}, \mathbf{X}, \mathbf{h})}{\partial c_i} = -h_i$$

$$\frac{\partial \mathrm{E}(\mathrm{y}, \mathbf{X}, \mathbf{h})}{\partial U_i} = -h_i y$$

Parameter Updates

$$A(y, \mathbf{X}, \mathbf{h}, \boldsymbol{\theta})$$

$$A\left(y, \mathbf{X}, \mathbf{h}, W_{j,k}^{(i)}\right) = -p(h_j = 1|y, X)1_{(x_i=k)}$$

$$A(y, \mathbf{X}, \mathbf{h}, b_k^i) = -1_{(x_i=k)}$$

$$A(y, \mathbf{X}, \mathbf{h}, d) = -y$$

$$A(y, \mathbf{X}, \mathbf{h}, c_j) = -p(h_j = 1|y, X)$$

$$A(y, \mathbf{X}, \mathbf{h}, U_j) = -p(h_j = 1|y, X)y$$

$$A(y, \mathbf{X}, \mathbf{h}, \boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|y,\mathbf{X})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial\theta}\right] = \sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})\frac{\partial E(y, \mathbf{X}, \mathbf{h}')}{\partial\theta}$$

$$A\left(y, \mathbf{X}, \mathbf{h}, W_{j,k}^{(i)}\right) = -\sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})h_j 1_{(x_i=k)} = -p\left(h_j = 1\middle|y, X\right)1_{(x_i=k)} \prod_{q,q\neq j}\sum_{h_q \in \{0,1\}} p\left(h_q\middle|y, X\right)$$

$$= -p(h_j = 1|y, X)1_{(x_i=k)}$$

$$A(y, \mathbf{X}, \mathbf{h}, b_k^i) = -\sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})1_{(x_i=k)} = -1_{(x_i=k)}$$

$$A(y, \mathbf{X}, \mathbf{h}, d) = -\sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})y = -y$$

$$A(y, \mathbf{X}, \mathbf{h}, c_j) = -\sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})h_j = -p(h_j = 1|y, X)$$

$$A(y, \mathbf{X}, \mathbf{h}, U_j) = -\sum_{\mathbf{h}'} p(\mathbf{h}'|y, \mathbf{X})h_j y = -p(h_j = 1|y, X)y$$

$B(y, \mathbf{X}, \mathbf{h}, \boldsymbol{\theta})$

$$B\left(y, \mathbf{X}, \mathbf{h}, W_{j,k}^{(i)}\right) = -1_{(x_i=k)} \sum_{y'} p(y'|\mathbf{X})p(h_j = 1|y', \mathbf{X})$$

$$B(y, \mathbf{X}, \mathbf{h}, b_k^i) = -1_{(x_i=k)}$$

$$B(y, \mathbf{X}, \mathbf{h}, d) = -\sum_{y'} p(y|\mathbf{X})y'$$

$$B(y, \mathbf{X}, \mathbf{h}, c_j) = -\sum_{y'} p(y'|\mathbf{X})p(h_j = 1|y', \mathbf{X})$$

$$B(y, \mathbf{X}, \mathbf{h}, U_j) = -\mathrm{p}(\mathrm{y} = 1|\mathbf{X})\mathrm{p}(h_j = 1|y = 1, \mathbf{X})$$

$$B(y, \mathbf{X}, \mathbf{h}, \boldsymbol{\theta}) = \mathbb{E}_{p(y,\mathbf{h}|\mathbf{X})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial \theta}\right] = \sum_{y'}\sum_{\mathbf{h}'} p(y', \mathbf{h}'|\mathbf{X})\frac{\partial E(y', \mathbf{X}, \mathbf{h}')}{\partial \theta}$$

$$B\left(y, \mathbf{X}, \mathbf{h}, W_{j,k}^{(i)}\right) = -\sum_{y'}\sum_{\mathbf{h}'} p(y', \mathbf{h}'|\mathbf{X})\, h_j 1_{(x_i=k)} = -1_{(x_i=k)} \sum_{y'} p(y|\mathbf{X}) \sum_{\mathbf{h}'} p(\mathbf{h}'|\mathrm{y}, \mathbf{X})h_j$$
$$= -1_{(x_i=k)} \sum_{y'} p(y'|\mathbf{X})p(h_j = 1|y', \mathbf{X})$$

$$B(y, \mathbf{X}, \mathbf{h}, b_k^i) = -\sum_{y'}\sum_{\mathbf{h}'} p(y', \mathbf{h}'|\mathbf{X})\, h_j 1_{(x_i=k)} = -1_{(x_i=k)}$$

$$B(y, \mathbf{X}, \mathbf{h}, d) = -\sum_{y'}\sum_{\mathbf{h}'} p(y', \mathbf{h}'|\mathbf{X})y = -\sum_{y'} p(y'|\mathbf{X})y' \sum_{\mathbf{h}'} p(\mathbf{h}'|\mathrm{y}', \mathbf{X}) = -\sum_{y'} p(y|\mathbf{X})y'$$

$$B(y, \mathbf{X}, \mathbf{h}, c_j) = -\sum_{y'} p(y|\mathbf{X}) \sum_{\mathbf{h}'} p(\mathbf{h}'|\mathrm{y}, \mathbf{X})h_j = -\sum_{y'} p(y'|\mathbf{X})p(h_j = 1|y', \mathbf{X})$$

$$B(y, \mathbf{X}, \mathbf{h}, U_j) = -\sum_{y'} y' p(y'|\mathbf{X}) \sum_{\mathbf{h}'} p(\mathbf{h}'|\mathrm{y}', \mathbf{X})h_j = -\mathrm{p}(\mathrm{y} = 1|\mathbf{X})\mathrm{p}(h_j = 1|y = 1, \mathbf{X})$$

$C(y, \mathbf{X}, \mathbf{h}, \boldsymbol{\theta})$ Using CD-k Approximation

$$C\left(y, \mathbf{X}, \mathbf{h}, W_{j,k}^{(i)}\right) = -h_j 1_{(x_i=k)}$$

$$C(y, \mathbf{X}, \mathbf{h}, b_k^i) = -1_{(\widehat{x_i}=k)}$$

$$C(y, \mathbf{X}, \mathbf{h}, d) = -\hat{\mathrm{y}}$$

$$C(y, \mathbf{X}, \mathbf{h}, c_j) = -\widehat{h_j}$$

$$C(y, \mathbf{X}, \mathbf{h}, U_j) = -\widehat{h_j}\hat{\mathrm{y}}$$

$$C(y, \mathbf{X}, \mathbf{h}, \boldsymbol{\theta}) = \mathbb{E}_{p(y,\mathbf{X},\mathbf{h})}\left[\frac{\partial E(y, \mathbf{X}, \mathbf{h})}{\partial \theta}\right] = \sum_{y',\mathbf{X}',\mathbf{h}'} p(y', \mathbf{X}', \mathbf{h}')\frac{\partial E(y', \mathbf{X}', \mathbf{h}')}{\partial \theta} \approx \frac{\partial E(\hat{\mathrm{y}}, \widehat{\mathbf{X}}, \hat{\mathbf{h}})}{\partial \theta}$$

$$C\left(y, \mathbf{X}, \mathbf{h}, W_{j,k}^{(i)}\right) = -\widehat{h_j} 1_{(\widehat{x_i}=k)}$$

$$C(y, \mathbf{X}, \mathbf{h}, b_k^i) = -1_{(\widehat{x_i}=k)}$$

$$C(y, \mathbf{X}, \mathbf{h}, d) = -\hat{y}$$
$$C(y, \mathbf{X}, \mathbf{h}, c_j) = -\widehat{h_j}$$
$$C(y, \mathbf{X}, \mathbf{h}, U_j) = -\widehat{h_j}\hat{y}$$

Combine All

$$\frac{\ln Hybrid(\alpha, y, \mathbf{X})}{\partial \theta} = -(1+\alpha)\mathrm{A}(\mathrm{y}, \mathbf{X}, \mathbf{h}, \theta) + \mathrm{B}(\mathrm{y}, \mathbf{X}, \mathbf{h}, \theta) + \alpha C(\mathrm{y}, \mathbf{X}, \mathbf{h}, \theta)$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial W_{j,k}^{(i)}} = 1_{(x_i=k)} \left\{ (1+\alpha)\mathrm{p}\left(h_j = 1 \middle| y, X\right) - \sum_{y'} \mathrm{p}(y'|\mathbf{X})\mathrm{p}\left(h_j = 1 \middle| y', \mathbf{X}\right) - \alpha\widehat{h_j} \right\}$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial b_j^i} = \alpha\left[1_{(x_i=k)} - 1_{(\hat{x}_i=k)}\right]$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial \mathrm{d}} = (1+\alpha)y - \mathrm{p}(\mathrm{y}=1|\mathbf{X}) - \alpha\hat{y}$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial c_i} = (1+\alpha)\mathrm{p}(h_j = 1|y, X) - \sum_{y'} \mathrm{p}(y'|\mathbf{X})\mathrm{p}(h_j = 1|y', \mathbf{X}) - \alpha\widehat{h_j}$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial U_i} = (1+\alpha)\mathrm{p}\left(h_j = 1 \middle| y, \mathbf{X}\right)y - \mathrm{p}(\mathrm{y}=1|\mathbf{X})\mathrm{p}\left(h_j = 1 \middle| y = 1, \mathbf{X}\right) - \alpha\widehat{h_j}\hat{y}$$

CD-K Updates

$X^{(0)} \leftarrow X, y^{(0)} \leftarrow y$

For t = 0, …, k – 1 do

For i = 1, …, H do sample $h_i^{(t)} \sim p(h_i = 1|y^{(t)}, X^{(t)})$

For i = 1, …, C do sample $x_i^{(t+1)} \sim p(x_i|y^{(t)}, h^{(t)})$

Do Sample $y^{(t+1)} \sim p(y = 1|h^{(t)})$

For i = 1,…H do $h_i^{(k)} \leftarrow p(h_i = 1|y^{(k)}, X^{(k)})$

Maximum Likelihood and the Delta Rule

- Maximum Likelihood

$$\ln L(\vec{\theta}|\vec{S}) = \ln \prod_{i=1}^{l} p(\overrightarrow{v^i}|\vec{\theta}) = \sum_{i=1}^{l} \ln p(\overrightarrow{v^i}|\vec{\theta})$$

- Mini-batch Gradient Ascent (Delta Rule)

$$\overrightarrow{\theta^{t+1}} = \overrightarrow{\theta^t} + \underbrace{\eta \frac{\partial}{\partial \overrightarrow{\theta^t}} \left[\sum_{i=1}^{N} \ln L(\overrightarrow{\theta^t}|\overrightarrow{v^i}) \right] - \lambda\overrightarrow{\theta^t} + \nu\Delta\overrightarrow{\theta^{t-1}}}_{:=\Delta\overrightarrow{\theta^t}}$$

Learning Algorithm

//CD-k, only if $\alpha \neq 0$

$X^{(0)} \leftarrow X, y^{(0)} \leftarrow y$

For t = 0, ..., k – 1 do

For j = 1, ..., H do sample $h_j^{(t)} \sim p\left(h_j = 1 \middle| y^{(t)}, \mathbf{X}^{(\mathbf{t})}\right) = \sigma(\sum_{1 \le i \le C} W^{(i)}_{j, x_i^{(t)}} + c_j + U_j y^{(t)})$[O(C+H)]

For i = 1, ..., C do sample $\boldsymbol{x_i^{(t+1)}} \sim \boldsymbol{p(x_i = k|}\mathbf{h}^{(\mathbf{t})}\boldsymbol{)} = \frac{\mathbf{exp}(\sum_{1 \le j \le H} \boldsymbol{h_j^{(t)} W_{j,k}^{(i)}})}{\sum_{1 \le q \le C_i} \mathbf{exp}(\sum_{1 \le j \le H} \boldsymbol{h_j^{(t)} W_{j,q}^{(i)}})}$(Computation Intensive! O(**H*V**))

Do Sample $y^{(t+1)} \sim p\left(y = 1 \middle| \mathbf{h}^{(\mathbf{t})}\right) = \sigma(\mathrm{d} + \mathbf{h}^{(\mathbf{t})\boldsymbol{T}}\mathbf{U})$[O(H)]

For i = 1,...H do $h_i^{(k)} \leftarrow p(h_i = 1 | y^{(k)}, X^{(k)}) = \sigma(\sum_{1 \le i \le C} W^{(i)}_{j, x_i^{(k)}} + c_j + U_j y^{(k)})$ [O(H)]

//Gradient Calculation

$\hat{y} \leftarrow y^{(k)}, \hat{X} \leftarrow X^{(k)}, \hat{h} \leftarrow h^{(k)}$

$$\mathbf{p}\left(\boldsymbol{h_j} = \mathbf{1} \middle| \boldsymbol{y}, \boldsymbol{X}\right) = \boldsymbol{\sigma}\left(\sum_{1 \le i \le C} \boldsymbol{W_{j,x_i}^{(i)}} + \boldsymbol{c_j} + \boldsymbol{U_j y}\right)$$

$$\mathbf{p}(\mathbf{y} = \mathbf{1}|\mathbf{X}) = \frac{\exp\left\{dy + \sum_{1 \le k \le H} \operatorname{softplus}\left(\sum_{1 \le \mathrm{i} \le \mathrm{C}} \boldsymbol{W_{k,x_i}^{(i)}} + c_k + \mathrm{U_k y}\right)\right\}}{\exp\left\{\sum_{1 \le k \le H} \operatorname{softplus}\left(\sum_{1 \le \mathrm{i} \le \mathrm{C}} \boldsymbol{W_{k,x_i}^{(i)}} + c_k\right)\right\} + \exp\left\{d + \sum_{1 \le k \le H} \operatorname{softplus}\left(\sum_{1 \le \mathrm{i} \le \mathrm{C}} \boldsymbol{W_{k,x_i}^{(i)}} + c_k + \mathrm{U_k}\right)\right\}}$$

[O(H*C)]

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial W_{j,k}^{(i)}} = \mathbf{1}_{(\boldsymbol{x_i = k})} \left\{ (1 + \alpha)\mathbf{p}\left(\boldsymbol{h_j} = \mathbf{1} \middle| \boldsymbol{y}, \boldsymbol{X}\right) - \sum_{y'} \mathbf{p}(\boldsymbol{y'}|\mathbf{X})\mathbf{p}\left(\boldsymbol{h_j} = \mathbf{1} \middle| \boldsymbol{y'}, \mathbf{X}\right) - \alpha \widehat{\boldsymbol{h_j}} \right\}$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial b_j^i} = \alpha\left[1_{(x_i = k)} - 1_{(\hat{x}_i = k)}\right]$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial \mathrm{d}} = (1 + \alpha)y - \mathbf{p}(\mathbf{y} = \mathbf{1}|\mathbf{X}) - \alpha \hat{\mathbf{y}}$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial c_i} = (1 + \alpha)\mathbf{p}\left(\boldsymbol{h_j} = \mathbf{1} \middle| \boldsymbol{y}, \mathbf{X}\right) - \sum_{y'} \mathbf{p}(\boldsymbol{y'}|\mathbf{X})\mathbf{p}\left(\boldsymbol{h_j} = \mathbf{1} \middle| \boldsymbol{y'}, \mathbf{X}\right) - \alpha \widehat{\boldsymbol{h_j}}$$

$$\frac{\partial \ln Hybrid(\alpha, y, \mathbf{X})}{\partial U_i} = (1 + \alpha)\mathbf{p}\left(\boldsymbol{h_j} = \mathbf{1} \middle| \boldsymbol{y}, \mathbf{X}\right)y - \mathbf{p}(\mathbf{y} = \mathbf{1}|\mathbf{X})\mathbf{p}\left(\boldsymbol{h_j} = \mathbf{1} \middle| \boldsymbol{y} = \mathbf{1}, \mathbf{X}\right) - \alpha \widehat{\boldsymbol{h_j}}\hat{\mathbf{y}}$$

//Parameter Update

$$\overrightarrow{\theta^{t+1}} = \overrightarrow{\theta^t} + \underbrace{\eta \frac{\partial}{\partial \overrightarrow{\theta^t}} \left[\sum_{i=1}^{N} \ln L(\overrightarrow{\theta^t} | \overrightarrow{v^i}) \right] - \lambda \overrightarrow{\theta^t} + \nu \Delta \overrightarrow{\theta^{t-1}}}_{:= \Delta \overrightarrow{\theta^t}}$$

[Note, storing two sets of parameters would be memory hungry, store the differences only??]

Note:

Efficient calculation of the following is the key to the performance of CD-k!

Naïve implementation requires O(|X|*|H|)

$$x_i^{(t+1)} \sim p\left(x_i = k \middle| \mathbf{h}^{(t)}\right) = \frac{\exp\left(\sum_{1\le j\le H} h_j^{(t)} W_{j,k}^{(i)}\right)}{\sum_{1\le q\le C_i} \exp\left(\sum_{1\le j\le H} h_j^{(t)} W_{j,q}^{(i)}\right)}, 1 \le \mathrm{k} \le C_i$$

Use Mini-batch + Cache Strategy??

Nonexact sampling??MCMC???Importance Sampling, rejection sampling etc to avoid the normalization constant? ????

What else??

SparseClassRBM Verses Logistic Regression

Note that in logisitc regression, we have:

$$\mathrm{p(y|X)} = \frac{\exp\{\mathrm{y(W^T x + c)}\}}{1 + \exp\{\mathrm{W^T x + c}\}}$$

And for SparseClassRBM, we have

$$\mathrm{p(y|X)} = \frac{\exp\{\mathrm{dy}\} \prod_{1\le k\le H} \sum_{h_k\in\{0,1\}} \exp\{h_k[\sum_{1\le i\le C} W_{k,x_i}^{(i)} + c_k + U_k y]\}}{\sum_{y'} \exp\{\mathrm{dy'}\} \prod_{1\le k\le H} \sum_{h_k\in\{0,1\}} \exp\{h_k[\sum_{1\le i\le C} W_{k,x_i}^{(i)} + c_k + U_k y']\}}$$

Setting H=1, $h_1 = 1$,d=1,U = 0, we have

$$\mathrm{p(y|X)} = \frac{\exp\{\mathrm{y}\left(\sum_{1\le i\le C} W_{x_i}^{(i)} + c\right)\}}{1 + \exp\{\sum_{1\le i\le C} W_{x_i}^{(i)} + c\}}$$

which is a form of logistic regression.

Therefore, we can view logistic regression as a special form of RBM with less variables.