1 Energy-Based Models

A configuration (v, h) is made of an evident part v and a hiden part h, and let its energy be $E(v, h)$ with parameters θ. Then the probability of the configuration is

$$p(v, h) = \frac{1}{Z} e^{-E(v,h)}$$

Z is the normalization factor such that $\sum_{v,h} p(v, h) = 1$:

$$Z = \sum_{v,h} e^{-E(v,h)}$$

Let $\mathcal{D}$ be the dataset with $N = |\mathcal{D}|$, then the negative log likelihood of parameter θ is

$$\mathcal{L}(\mathcal{D}) = \frac{1}{N} \sum_{v \in \mathcal{D}} \log \sum_h p(v, h)$$

Define

$$p_{\mathcal{D}}(v) = \begin{cases} \frac{1}{N} & v \in \mathcal{D} \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathcal{L}(\mathcal{D}) = \sum_v p_{\mathcal{D}}(v) \log \sum_h p(v, h)$$

and the loss function is

$$\begin{aligned}
l(\mathcal{D}) &= -\sum_v p_{\mathcal{D}}(v) \log \sum_h p(v, h) \\
&= -\sum_v p_{\mathcal{D}}(v) \log \sum_h \frac{1}{Z} e^{-E(v,h)} \\
&= -\sum_v p_{\mathcal{D}}(v) \log \frac{1}{Z} \sum_h e^{-E(v,h)} \\
&= -\sum_v p_{\mathcal{D}}(v) \left\{ \log \sum_h e^{-E(v,h)} - \log Z \right\} \\
&= -\sum_v p_{\mathcal{D}}(v) \log \sum_h e^{-E(v,h)} - \sum_v p_{\mathcal{D}}(v) \log Z \\
&= -\sum_v p_{\mathcal{D}}(v) \log \sum_h e^{-E(v,h)} + \log Z
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial\theta} l(D) &= -\sum_v p_{\mathcal{D}}(v)\frac{\partial}{\partial\theta}\log\sum_h e^{-E(v,h)} + \frac{\partial}{\partial\theta}\log Z \\
&= -\sum_v p_{\mathcal{D}}(v)\frac{1}{\sum_h e^{-E(v,h)}}\sum_h \frac{\partial}{\partial\theta} e^{-E(v,h)} + \frac{1}{Z}\frac{\partial}{\partial\theta} Z \\
&= -\sum_v p_{\mathcal{D}}(v)\frac{1}{\sum_h e^{-E(v,h)}}\sum_h e^{-E(v,h)}\left[-\frac{\partial}{\partial\theta}E(v,h)\right] + \frac{1}{Z}\frac{\partial}{\partial\theta} Z \\
&= \sum_v p_{\mathcal{D}}(v)\frac{1}{\sum_h e^{-E(v,h)}}\sum_h e^{-E(v,h)}\frac{\partial}{\partial\theta}E(v,h) + \frac{1}{Z}\frac{\partial}{\partial\theta} Z \\
&= \sum_v p_{\mathcal{D}}(v)\frac{1}{\sum_h p(v,h)}\sum_h p(v,h)\frac{\partial}{\partial\theta}E(v,h) + \frac{1}{Z}\frac{\partial}{\partial\theta} Z \\
&= \sum_v p_{\mathcal{D}}(v)\frac{1}{p(v)}\sum_h p(v,h)\frac{\partial}{\partial\theta}E(v,h) + \frac{1}{Z}\frac{\partial}{\partial\theta} Z \\
&= \sum_v p_{\mathcal{D}}(v)\sum_h p(h|v)\frac{\partial}{\partial\theta}E(v,h) + \frac{1}{Z}\frac{\partial}{\partial\theta} Z \\
&= \sum_{v,h} p_{\mathcal{D}}(v)p(h|v)\frac{\partial}{\partial\theta}E(v,h) + \frac{1}{Z}\frac{\partial}{\partial\theta} Z \\
&= \sum_{v,h} p_{\mathcal{D}}(v)p(h|v)\frac{\partial}{\partial\theta}E(v,h) + \frac{1}{Z}\frac{\partial}{\partial\theta}\sum_{v,h} e^{-E(v,h)} \\
&= \sum_{v,h} p_{\mathcal{D}}(v)p(h|v)\frac{\partial}{\partial\theta}E(v,h) - \frac{1}{Z}\sum_{v,h} e^{-E(v,h)}\frac{\partial}{\partial\theta}E(v,h) \\
&= \sum_{v,h} p_{\mathcal{D}}(v)p(h|v)\frac{\partial}{\partial\theta}E(v,h) - \sum_{v,h} p(v,h)\frac{\partial}{\partial\theta}E(v,h)
\end{aligned}
$$

That is

$$
\frac{\partial}{\partial\theta} l(D) = \sum_{v,h} p_{\mathcal{D}}(v)p(h|v)\frac{\partial}{\partial\theta}E(v,h) - \sum_{v,h} p(v,h)\frac{\partial}{\partial\theta}E(v,h)
$$

2 Restricted Boltzmann Machine

2.1 Derivatives

$$
v_i, j_j \in \{0,1\}
$$

$$
E(v,h) = -\sum_i b_i v_i - \sum_j c_j h_j - \sum_{i,j} W_{ij} v_i h_j
$$

We have

$$\frac{\partial}{\partial b_i} E(v,h) = -v_i$$
$$\frac{\partial}{\partial c_j} E(v,h) = -h_j$$
$$\frac{\partial}{\partial W_{ij}} E(v,h) = -v_i h_j$$

3 Conditional Probability

$$\begin{aligned} p(v,h) &= \frac{1}{Z} \exp\left(\sum_i b_i v_i + \sum_j c_j h_j + \sum_{i,j} W_{ij} v_i h_j \right) \\ &= \frac{1}{Z} \prod_i e^{b_i v_i} \prod_j e^{c_j h_j} \prod_{i,j} e^{W_{ij} v_i h_j} \end{aligned}$$

$$\begin{aligned} p(v|h) &= \prod_i p(v_i|h) \\ p(h|v) &= \prod_j p(h_j|v) \end{aligned}$$

Let $h = \{h_j\} \cup h'_j$,

$$\begin{aligned} p(h_j|v) &= \sum_{h'_j} p(h_j, h'_j|v) \\ &= \sum_{h'_j} p(h_j|v) p(h'_j|v) \\ &= \frac{\sum_{h'} p(h_j, h', v)}{\sum_{h_j, h'_j} p(h_j, h', v)} \\ &= \frac{\sum_{h'} \prod_i e^{b_i v_i} \prod_j e^{c_j h_j} \prod_{i,j} e^{W_{ij} v_i h_j}}{\sum_h \prod_i e^{b_i v_i} \prod_j e^{c_j h_j} \prod_{i,j} e^{W_{ij} v_i h_j}} \\ &= \frac{\exp\{(c_j + \sum_i W_{ij} v_i) h_j\}}{\sum_{h_j} \exp\{(c_j + \sum_i W_{ij} v_i) h_j\}} \end{aligned}$$

$$\begin{aligned} p(h_j = 1|v) &= \mathrm{sigm}(c_j + \sum_i W_{ij} v_i) \\ p(v_i = 1|h) &= \mathrm{sigm}(b_i + \sum_j W_{ij} h_j) \end{aligned}$$

3.1 Combined

$$
\begin{aligned}
\frac{\partial}{\partial W_{ij}} l(D) &= -\sum_{v,h} p_{\mathcal{D}}(v) p(h|v) v_i h_j + \sum_{v,h} p(v) p(h|v) v_i h_j \\
&= -\sum_{v,h} p_{\mathcal{D}}(v) p(h_j = 1|v) v_i + \sum_{v,h} p(v) p(h_j = 1|v) v_i \\
&= -\sum_{v} p_{\mathcal{D}}(v) \mathrm{sigm}(\dots) v_i + \sum_{v} p(v) \mathrm{sigm}(\dots) v_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial b_i} l(D) &= -\sum_{v,h} p_{\mathcal{D}}(v) p(h|v) v_i + \sum_{v,h} p(v) p(h|v) v_i \\
&= -\sum_{v} p_{\mathcal{D}}(v) v_i + \sum_{v} p(v) v_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial c_j} l(D) &= -\sum_{v,h} p_{\mathcal{D}}(v) p(h|v) h_i + \sum_{v,h} p(v) p(h|v) h_i \\
&= -\sum_{v,h} p_{\mathcal{D}}(v) p(h_j = 1|v) + \sum_{v,h} p(v) p(h_j = 1|v) \\
&= -\sum_{v} p_{\mathcal{D}}(v) \mathrm{sigm}(\dots) + \sum_{v} p(v) \mathrm{sigm}(\dots)
\end{aligned}
$$

3.2 Vector Forms

$$
\begin{aligned}
\frac{\partial}{\partial b} l(D) &= -\sum_{v} p_{\mathcal{D}}(v) v + \sum_{v} p(v) v \\
\frac{\partial}{\partial c} l(D) &= -\sum_{v} p_{\mathcal{D}}(v) \mathrm{sigm}(c + v'W) + \sum_{v} p(v) \mathrm{sigm}(c + v'W) \\
\frac{\partial}{\partial W} l(D) &= -\sum_{v} p_{\mathcal{D}}(v) v \times \mathrm{sigm}(c + v'W)' + \sum_{v} p(v) v \times \mathrm{sigm}(c + v'W)'
\end{aligned}
$$

Or

$$
\begin{aligned}
\frac{\partial}{\partial b} l(D) &= -\frac{1}{N} \sum_{v \in \mathrm{data}} v + \frac{1}{N_s} \sum_{v \in \mathrm{sample}} v \\
&= -\langle v \rangle_{\mathrm{data}} + \langle v \rangle_{\mathrm{sample}} \\
\frac{\partial}{\partial c} l(D) &= -\frac{1}{N} \sum_{v \in \mathrm{data}} \mathrm{sigm}(c + v'W) + \frac{1}{N_s} \sum_{v \in \mathrm{sample}} \mathrm{sigm}(c + v'W) \\
&= -\langle \mathrm{sigm}(c + v'W) \rangle_{\mathrm{data}} + \langle \mathrm{sigm}(c + v'W) \rangle_{\mathrm{sample}} \\
\frac{\partial}{\partial W} l(D) &= -\frac{1}{N} \sum_{v \in \mathrm{data}} v \times \mathrm{sigm}(c + v'W)' + \frac{1}{N_s} \sum_{v \in \mathrm{sample}} v \times \mathrm{sigm}(c + v'W)'
\end{aligned}
$$