Exercise Sheet 6

Machine Learning 2, SS16

June 6, 2016

Mario Tambos, 380599; Viktor Jeney, 348969; Sascha Huk, 321249; Jan Tinapp, 0380549

Exercise 1

(a)
Define

$$y := x^T W - b^T$$

$$\begin{aligned}
p_\theta(x) &= \sum_{h \in \{-1,0,1\}^N} p(x,h) \\
&= \sum_{h \in \{-1,0,1\}^N} \frac{1}{Z} \exp(yh + x^T a) \\
&= \frac{1}{Z} \exp(x^T a) \sum_{h \in \{-1,0,1\}^N} \exp(yh) \\
&= \frac{1}{Z} \exp(x^T a) \sum_{h \in \{-1,0,1\}^N} \exp(\sum_{i=1}^{N} y_i h_i) \\
&= \frac{1}{Z} \exp(x^T a) \sum_{h \in \{-1,0,1\}^N} \prod_{i=1}^{N} \exp(y_i h_i)
\end{aligned}$$

Because the expression $\exp(y_i h_i)$ only depends on the i'th component of h, we can rewrite the sum and product to get:

$$\begin{aligned}
p_\theta(x) &= \frac{1}{Z} \exp(x^T a) \prod_{i=1}^{N} \sum_{h \in \{-1,0,1\}} \exp(y_i h_i) \\
&= \frac{1}{Z} \exp(x^T a) \exp(\log(\prod_{i=1}^{N} \sum_{h \in \{-1,0,1\}} \exp(y_i h_i))) \\
&= \frac{1}{Z} \exp(x^T a) \exp(\sum_{i=1}^{N} \log(\sum_{h \in \{-1,0,1\}} \exp(y_i h_i))) \\
&= \frac{1}{Z} \exp(x^T a) \exp(\sum_{i=1}^{N} \log(1 + e^{y_i} + e^{-y_i})) \\
&= \frac{1}{Z} \exp(x^T a) \exp(\sum_{i=1}^{N} \log(1 + 2cosh(y_i))) \\
&= \frac{1}{Z} \exp(x^T a + \sum_{i=1}^{N} \log(1 + 2cosh(w_i x - b_i)))
\end{aligned}$$

$$= \frac{1}{Z} \exp(x^T a + \sum_{i=1}^{N} \log(2(\frac{1}{2} + cosh(w_i x - b_i))))$$
$$= \frac{1}{Z} \exp(x^T a + \sum_{i=1}^{N} \log(2) + \log(\frac{1}{2} + cosh(w_i x - b_i)))$$
$$= \frac{1}{Z} \exp(N \log(2) + x^T a + \sum_{i=1}^{N} \log(\frac{1}{2} + cosh(w_i x - b_i)))$$
$$= \frac{1}{Z} 2^N \exp(x^T a + \sum_{i=1}^{N} \log(\frac{1}{2} + cosh(w_i x - b_i)))$$

With
$$Z' := \frac{2^N}{Z}$$
the desired result follows.

(b)

First compute gradients of F:
$$\bigtriangledown_{a_i} F(x) = \bigtriangledown_{a_i} - a^T x = -x_i$$

$$\begin{aligned}\bigtriangledown_{b_j} F(x) &= -\sum_{k=1}^{N} \bigtriangledown_{b_j} \log(\frac{1}{2} + cosh(w_k x - b_k)) \\ &= -\bigtriangledown_{b_j} \log(\frac{1}{2} + cosh(w_j x - b_j)) \\ &= -\frac{1}{\frac{1}{2} + cosh(w_k x - b_k)} \bigtriangledown_{b_j} cosh(w_j x - b_j) \\ &= -\frac{1}{\frac{1}{2} + cosh(w_k x - b_k)} sinh(w_j x - b_j) \bigtriangledown_{b_j} b_j \\ &= -\frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)}\end{aligned}$$

$$\begin{aligned}\bigtriangledown_{w_{ij}} F(x) &= -\sum_{k=1}^{N} \bigtriangledown_{w_{ij}} \log(\frac{1}{2} + cosh(w_k x - b_k)) \\ &= -\bigtriangledown_{w_{ij}} \log(\frac{1}{2} + cosh(w_j x - b_j)) \\ &= -\frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} \bigtriangledown_{w_{ij}} (w_j^T x) \\ &= -\frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} x_i\end{aligned}$$

Now we can plug the gradients into the expectations:

To compute the expectations with respect to the empirical distribution, we consider observed data points $x^{(1)}, ..., x^{(n)}$

$$\begin{aligned}\bigtriangledown_{a_i} KL(\hat{p}||p_\theta) &= < -x_i >_{\hat{p}} - < -x_i >_{p_\theta} \\ &= \sum_{x \in \{0,1\}^d} x_i p_\theta(x) - < x_i >_{\hat{p}} \\ &= \sum_{x \in \{0,1\}^d} x_i p_\theta(x) - \frac{1}{n} \sum_{k=1}^{n} x_i^{(k)} \\ &= \frac{1}{Z'} \sum_{x \in \{0,1\}^d} x_i \exp(-F_\theta(x)) - \frac{1}{n} \sum_{k=1}^{n} x_i^{(k)}\end{aligned}$$

$$\bigtriangledown_{b_j} KL(\hat{p}||p_\theta) =< -\frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} >_{\hat{p}} - < -\frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} >_{p_\theta}$$
$$= \frac{1}{Z'} \sum_{x \in \{0,1\}^d} \frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} \exp(-F_\theta(x)) - \frac{1}{n} \sum_{k=1}^{n} \frac{sinh(w_j x^{(k)} - b_j)}{\frac{1}{2} + cosh(w_k x^{(k)} - b_k)}$$

$$\bigtriangledown_{w_{ij}} KL(\hat{p}||p_\theta) =< -\frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} x_i >_{\hat{p}} - < -\frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} x_i >_{p_\theta}$$
$$= \frac{1}{Z'} \sum_{x \in \{0,1\}^d} \frac{sinh(w_j x - b_j)}{\frac{1}{2} + cosh(w_k x - b_k)} x_i \exp(-F_\theta(x)) - \frac{1}{n} \sum_{k=1}^{n} \frac{sinh(w_j x^{(k)} - b_j)}{\frac{1}{2} + cosh(w_k x^{(k)} - b_k)} x_i^{(k)}$$

(c)
The independence model is $(x_i \perp\!\!\!\perp x_j \mid h)$ and $(h_i \perp\!\!\!\perp h_j \mid x)$ for all $i \neq j$:

$$p(h \mid x) \overset{\text{Model}}{=} \prod_{j=1}^{N} \Pr_j(h_j \mid x)$$

$$p(h \mid x) = \frac{p(h,x)}{p(x)} = \frac{\cancel{\frac{1}{Z} e^{x^\top a}} \prod_j e^{h_j w_j^\top x + h_j b_j}}{\cancel{\frac{1}{Z} e^{x^\top a}} \prod_j (e^{w_j^\top x + b_j} + e^{-w_j^\top x - b_j} + 1)} \overset{\text{Model}}{=} \frac{\prod_j \Pr_j(h_j, x)}{\prod_j \Pr_j(x)} = \prod_j \Pr_j(h_j \mid x)$$

$$\implies \Pr_j(h_j \mid x) = \frac{\Pr_j(h_j, x)}{\Pr_j(x)} = \frac{e^{h_j w_j^\top x + h_j b_j}}{\sum_{h_j} e^{h_j w_j^\top x + h_j b_j}} = \frac{e^{h_j w_j^\top x + h_j b_j}}{e^{w_j^\top x + b_j} + e^{-w_j^\top x - b_j} + 1}$$

This is kind of a softmax function.

$$\Pr_j(h_j = 1 \mid x) = \frac{e^{w_j^\top x + b_j}}{e^{w_j^\top x + b_j} + e^{-w_j^\top x - b_j} + 1} = \frac{1}{1 + \exp(-2w_j x - 2b_j) + \exp(-w_j x - b_j)}$$

$$\Pr_j(h_j = -1 \mid x) = \frac{e^{-w_j^\top x - b_j}}{e^{w_j^\top x + b_j} + e^{-w_j^\top x - b_j} + 1} = \frac{1}{\exp(2w_j x + 2b_j) + 1 + \exp(w_j x + b_j)}$$

$$\Pr_j(h_j = 0 \mid x) = \frac{1}{e^{w_j^\top x + b_j} + e^{-w_j^\top x - b_j} + 1} = \frac{1}{\exp(w_j x + b_j) + \exp(-w_j x - b_j) + 1}$$

$$p(x \mid h) \overset{\text{Model}}{=} \prod_{k=1}^{d} \Pr_k(x_k \mid h)$$

$$p(x \mid h) = \frac{p(x,h)}{p(h)} = \frac{\cancel{\frac{1}{Z} e^{h^\top b}} \prod_k e^{x_k a_k + x_k w_k^\top h}}{\cancel{\frac{1}{Z} e^{h^\top b}} \prod_k (1 + e^{a_k + w_k^\top h})} \overset{\text{Model}}{=} \frac{\prod_k \Pr_k(x_k, h)}{\prod_k \Pr_k(h)} = \prod_k \Pr_k(x_k \mid h)$$

$$\Pr(x_k = 1 \mid h) = \frac{e^{a_k + w_k^\top h}}{1 + e^{a_k + w_k^\top h}} = \frac{1}{e^{-a_k - w_k^\top h} + 1} = \text{sigm}(a_k + w_k^\top h)$$

$$\Pr(x_k = 0 \mid h) = \frac{1}{1 + e^{a_k + w_k^\top h}} = \text{sigm}(-a_k - w_k^\top h)$$