Aaron Dardik

CS567 Problem Set #2

1.1) Given $q_i: 1 \le i \le K, C = \frac{1}{\sum q_i}$ we will denote Cq_i by p_i. We will let $[y = k] \stackrel{\text{def}}{=} 1\{y = k\}$. Now we must show that $p = \prod_{i=1}^{K} (p_i)^{[y=i]}$ is in the exponential family. Now, $p = e^{\log p} = e^{\log \prod_{i=1}^{K} p_i^{[y=i]}}$. By using the fact that the logarithm of a product is the sum of the logarithms the above expression simplifies to $e^{\sum_i \log\left(p_i^{[y=i]}\right)} = e^{\sum_i [y=i] \log p_i}$. Let $\eta = \begin{bmatrix} \log p_1 \\ \vdots \\ \log p_K \end{bmatrix}, t(y) = \begin{bmatrix} [y = 1] \\ \vdots \\ [y = K] \end{bmatrix}, b(y) = 1, a(\eta) = 0$ and this shows that the categorical distribution is in the exponential family.

1.2) $p(y|x) \sim$ Bernoulli $\Rightarrow p(y|x) = p^y(1-p)^{1-y} = e^{\log p^y(1-p)^{1-y}} = e^{y \log p + (1-y)\log(1-p)}$. This simplifies to $e^{y \log \frac{p}{1-p} + \log(1-p)}$. Now, $\eta = Wx = \log \frac{p}{1-p}, t(y) = y, b(y) = 1, a(\eta) = -\log(1-p)$. Now, $h(x; W) = E_y t(y) = 1 * e^{1 \log \frac{p}{1-p} + \log(1-p)} + 0 = p = p(x; W)$.

1.3) From part 1.1 we see that for the categorical distribution we have $\eta = \begin{bmatrix} \log p_1 \\ \vdots \\ \log p_K \end{bmatrix}, t(y) = \begin{bmatrix} [y = 1] \\ \vdots \\ [y = K] \end{bmatrix}, b(y) = 1, a(\eta) = 0$. From this we see that $p_i = e^{\eta_i}, \sum p_i = \sum e^{\eta_i} = 1$. By substituting back in and dividing we see that $p_i = \frac{e^{\eta_i}}{\sum_j e^{\eta_j}}$. Now, as $\eta = Wx \Rightarrow p(y = i|x) = {e^{(Wx)_i}}\big/{\sum_j e^{(Wx)_j}} = {e^{w_i^T x}}\big/{\sum_j e^{w_j^T x}}$. Now, we note that p_K can be expressed as a linear combination of $p_1 \dots p_{K-1}$ so we have that $E[t(y)|x] = E\begin{bmatrix} [y = 1] & \\ \vdots & |x \\ [y = K-1] & \end{bmatrix} = \begin{bmatrix} {e^{w_1^T x}}\big/{\sum e^{w_j^T x}} \\ \vdots \\ {e^{w_{K-1}^T x}}\big/{\sum e^{w_j^T x}} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_{K-1} \end{bmatrix}$ with p_K being a parameter expressed in terms of the first $p_1, \dots, p_{K-1}$ and not an independent variable itself.

1.4) $p(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = e^{\log \frac{\lambda^y e^{-\lambda}}{y!}} = e^{\log \lambda^y + \log e^{-\lambda} - \log y!} = \frac{1}{y!} * e^{y \log \lambda - \lambda}$. From this we conclude that the Poisson distribution is in the exponential family with $b(y) = \frac{1}{y!}, \eta = \log \lambda,\ t(y) = y, a(\eta) = \lambda$. Now, $\eta = w^T x = \log \lambda \Rightarrow e^{w^T x} = \lambda$. As the expected value of a Poisson distribution is λ we have $\widehat{y_i} = E[y|x; w] = \lambda = e^{w^T x}$.

2.1.1) For ease of notation we will denote our cost function by C instead of l. To find $\frac{\partial C}{\partial \boldsymbol{a}}$ it suffices to find $\frac{\partial C}{\partial a_i}, \forall i$. We will consider two cases: $i \ne y, i = y$. In the first case, we have $\frac{\partial C}{\partial a_i} = \frac{-1}{z_y} * \frac{\partial}{\partial a_i}(z_y)$. The latter derivative we will solve via the quotient rule to get ${0 - e^{a_y} e^{a_i}}\big/{(\sum e^{a_j})^2} = -z_y * z_i$. Combining these results we have $i \ne y \Rightarrow \frac{\partial C}{\partial a_i} = z_i$. In the latter case where $i = y$, we have $\frac{\partial C}{\partial a_y} =$

$(\sum e^{a_j}) * e^{a_y} - e^{a_y}e^{a_y} \Big/ (\sum e^{a_j})^2 = z_y - (z_y)^2$. Thus, letting $\boldsymbol{y} \in \mathbb{R}^K$ be the vector whose kth coordinate is 1 if $k = y$ and 0 otherwise, we have that $\frac{\partial C}{\partial \boldsymbol{a}} = \boldsymbol{z} - z_y^2\boldsymbol{y}$.

2.1.2) Now, $\frac{\partial C}{\partial \boldsymbol{W}^{(2)}} = \frac{\partial C}{\partial \boldsymbol{a}}\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{W}^{(2)}} = \frac{\partial C}{\partial \boldsymbol{a}}\boldsymbol{h}$. Additionally, $\frac{\partial C}{\partial \boldsymbol{b}^{(2)}} = \frac{\partial C}{\partial \boldsymbol{a}}\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}^{(2)}} = \frac{\partial C}{\partial \boldsymbol{a}}$.

2.1.3)