Assignment 1: CS 224d (Jerrick Hoang)

Due: Wednesday

1. Softmax is translational invariant $\rho(\boldsymbol{x}+c)_j = \frac{\exp(x_j+c)}{\sum_{k=1}^{\boldsymbol{K}} \exp(x_k+c)} = \frac{\exp(x_j)\exp(c)}{\exp(c)\sum_{k=1}^{\boldsymbol{K}} \exp(x_k)} = \frac{\exp(x_j)}{\sum_{k=1}^{\boldsymbol{K}} \exp(x_k)} = \rho(\boldsymbol{x})$

2. (a) Let $\rho(\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{x}}}$ Then $\frac{\partial \rho}{\partial \boldsymbol{x}} = \frac{-1}{(1+e^{-\boldsymbol{x}})^2}\frac{d(1+e^{-\boldsymbol{x}})}{d\boldsymbol{x}} = \frac{e^{-\boldsymbol{x}}}{(1+e^{-\boldsymbol{x}})^2} = (1-\rho(\boldsymbol{x})\rho(\boldsymbol{x})$

   (b) Let $h(z)_j = \frac{e^{z_j}}{\sum e^{z_i}}$. Let's derive $\frac{\partial h(z)_j}{\partial z_i}$. When $i=j$, must find $\frac{\partial h(z)_j}{\partial z_j} = \frac{\partial}{\partial z_j}\frac{e^{z_j}}{\sum e^{z_i}} = \frac{e^{z_j}(\sum e^{z_i}) - e^{2z_j}}{(\sum e^{z_i})^2} = h(z)_j(1-h(z)_j)$. When $i \neq j$, $\frac{\partial h(z)_j}{\partial z_i} = \frac{\partial}{\partial z_i}\frac{e^{z_j}}{\sum e^{z_i}} = \frac{-e^{z_j}e^{z_i}}{(\sum e^{z_i})^2} = -h(z)_i h(z)_j$.

   Now, $\frac{\partial CE(y,\hat{y})}{\partial \theta_j} = -\sum_i y_i \frac{\partial}{\partial \theta_j}\log h(\theta)_i = -\sum_i \frac{y_i}{h(\theta)_i}\frac{\partial}{\partial \theta_j}h(\theta)_i = -\sum_{i\neq j}\frac{y_i}{h(\theta)_i}h(\theta)_i h(\theta)_j - (h(\theta)_j(1-h(\theta)_j)) = -h(\theta)_j(\sum_{i\neq j} y_i + 1 - h(\theta)_j)$

   (c) $\frac{\partial \hat{y}}{\partial x} = \frac{\partial \hat{y}}{\partial h}\frac{dh}{dx} =$

   (d) $D_x * H + H * D_y$ weights

3. (a) $CE = -\sum_i y_i log(\frac{\exp(\hat{r}^\intercal w_i)}{\sum_j exp(\hat{r}^\intercal w_j)})$. So, $\frac{\partial CE}{\partial \hat{r}} = -\sum_i y_i \frac{\partial}{\partial \hat{r}}\log\frac{e^{\hat{r}^\intercal w_i}}{\sum_j e^{\hat{r}^\intercal w_j}} = -\sum_i y_i(\frac{\partial}{\partial \hat{r}}\log e^{\hat{r}^\intercal w_i} - \frac{\partial}{\partial \hat{r}}log(\sum_j e^{\hat{r}^\intercal w_j})) = -\sum y_i(w_i - \frac{w_j e^{\hat{r}^\intercal w_j}}{\sum_j e^{\hat{r}^\intercal w_j}}) = -\sum_i y_i(w_i - \sum_j w_j Pr(word_j|\hat{r}, w))$

   (b) $CE = -\sum_i y_i log(\frac{\exp(\hat{r}^\intercal w_i)}{\sum_j exp(\hat{r}^\intercal w_j)})$. So, $\frac{\partial CE}{\partial w_k} = -\sum_i y_i \frac{\partial}{\partial w_k}\log\frac{e^{\hat{r}^\intercal w_i}}{\sum_j e^{\hat{r}^\intercal w_j}} = -\sum_i y_i(\frac{\partial}{\partial w_k}\log e^{\hat{r}^\intercal w_i} - \frac{\partial}{\partial w_k}log(\sum_j e^{\hat{r}^\intercal w_j})) = -\sum y_i(1_{k=i}\hat{r} - \frac{\hat{r}e^{\hat{r}^\intercal w_k}}{\sum_j e^{\hat{r}^\intercal w_j}}) = -\sum_i y_i\hat{r}(1_{k=i} - Pr(word_k|\hat{r}, w))$