1. Derivative of Softmax Function

$$t \neq j \qquad {}^{r}a_j^{(L)} = soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right) = \frac{e^{{}^{r}z_j^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} \qquad soft\left({}^{r}\mathbf{z}_{\neq\mathbf{t}}^{(\mathbf{L})}, j\right) = \frac{e^{{}^{r}z_j^{(L)}}}{\sum_{s\neq t} e^{{}^{r}z_s^{(L)}}}$$

$$ln\, soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right) = ln\frac{e^{{}^{r}z_j^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = {}^{r}z_j^{(L)} - ln\sum_s e^{{}^{r}z_s^{(L)}}$$

$$ln\left(1 - soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right)\right) = ln\frac{\sum_{s\neq j} e^{{}^{r}z_s^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = ln\sum_{s\neq j} e^{{}^{r}z_s^{(L)}} - ln\sum_s e^{{}^{r}z_s^{(L)}}$$

$$ln\, soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, t\right) = ln\frac{e^{{}^{r}z_t^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = {}^{r}z_t^{(L)} - ln\sum_s e^{{}^{r}z_s^{(L)}}$$

$$ln\left(1 - soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, t\right)\right) = ln\frac{\sum_{s\neq t} e^{{}^{r}z_s^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = ln\sum_{s\neq t} e^{{}^{r}z_s^{(L)}} - ln\sum_s e^{{}^{r}z_s^{(L)}}$$

$$\frac{\partial}{\partial\, {}^{r}z_j^{(L)}} ln\, soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right) = \frac{\partial}{\partial\, {}^{r}z_j^{(L)}}\left({}^{r}z_j^{(L)} - ln\sum_s e^{{}^{r}z_s^{(L)}}\right) = 1 - \frac{e^{{}^{r}z_j^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = 1 - soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right)$$

$$\frac{\partial}{\partial\, {}^{r}z_j^{(L)}} ln\left(1 - soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right)\right) = \frac{\partial}{\partial\, {}^{r}z_j^{(L)}}\left(ln\sum_{s\neq j} e^{{}^{r}z_s^{(L)}} - ln\sum_s e^{{}^{r}z_s^{(L)}}\right) = -\frac{e^{{}^{r}z_j^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = -soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right)$$

$$\frac{\partial}{\partial\, {}^{r}z_j^{(L)}} ln\, soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, t\right) = \frac{\partial}{\partial\, {}^{r}z_j^{(L)}}\left({}^{r}z_t^{(L)} - ln\sum_s e^{{}^{r}z_s^{(L)}}\right) = -\frac{e^{{}^{r}z_j^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = -soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right)$$

$$\frac{\partial}{\partial\, {}^{r}z_j^{(L)}} ln\left(1 - soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, t\right)\right) = \frac{\partial}{\partial\, {}^{r}z_j^{(L)}}\left(ln\sum_{s\neq t} e^{{}^{r}z_s^{(L)}} - ln\sum_s e^{{}^{r}z_s^{(L)}}\right) = \frac{e^{{}^{r}z_j^{(L)}}}{\sum_{s\neq t} e^{{}^{r}z_s^{(L)}}} - \frac{e^{{}^{r}z_j^{(L)}}}{\sum_s e^{{}^{r}z_s^{(L)}}} = soft\left({}^{r}\mathbf{z}_{\neq\mathbf{t}}^{(\mathbf{L})}, j\right) - soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, j\right)$$

2. Derivative of Softmax Cross Entropy Function

$$C = -\sum_{t=1}^{m^{(L)}} \sum_{r} \left({}^{r}y_t \, ln \, soft\left({}^{r}\mathbf{z}^{(L)}, t \right) + (1 - {}^{r}y_t) \, ln\left(1 - soft\left({}^{r}\mathbf{z}^{(L)}, t\right)\right)\right)$$

$$= -\sum_{r} \left({}^{r}y_j \, ln \, soft\left({}^{r}\mathbf{z}^{(L)}, j \right) + \left(1 - {}^{r}y_j\right) ln\left(1 - soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right)\right) - \sum_{\substack{t=1 \\ t \neq j}}^{m^{(L)}} \sum_{r} \left({}^{r}y_t \, ln \, soft\left({}^{r}\mathbf{z}^{(L)}, t \right) + (1 - {}^{r}y_t) \, ln\left(1 - soft\left({}^{r}\mathbf{z}^{(L)}, t\right)\right)\right)$$

$$\frac{\partial C}{\partial \, {}^{r}z_j^{(L)}} = -\left({}^{r}y_j\left(1 - soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right) + \left(1 - {}^{r}y_j\right)\left(- soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right)\right) - \sum_{\substack{t=1 \\ t \neq j}}^{m^{(L)}} \left({}^{r}y_t\left(- soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right) + (1 - {}^{r}y_t)\left(soft\left({}^{r}\mathbf{z}_{\neq t}^{(L)}, j\right) - soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right)\right)$$

$$= -\left({}^{r}y_j - soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right) - \sum_{\substack{t=1 \\ t \neq j}}^{m^{(L)}} \left((1 - {}^{r}y_t) \, soft\left({}^{r}\mathbf{z}_{\neq t}^{(L)}, j\right) - soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right)$$

$$= -\left({}^{r}y_j - soft\left({}^{r}\mathbf{z}^{(L)}, j\right)\right) - \sum_{\substack{t=1 \\ t \neq j}}^{m^{(L)}} (1 - {}^{r}y_t) \, soft\left({}^{r}\mathbf{z}_{\neq t}^{(L)}, j\right) + \sum_{\substack{t=1 \\ t \neq j}}^{m^{(L)}} soft\left({}^{r}\mathbf{z}^{(L)}, j\right)$$

$$= soft\left({}^{r}\mathbf{z}^{(L)}, j\right) - {}^{r}y_j + \left(m^{(L)} - 1\right) soft\left({}^{r}\mathbf{z}^{(L)}, j\right) - \sum_{\substack{t=1 \\ t \neq j}}^{m^{(L)}} (1 - {}^{r}y_t) \, soft\left({}^{r}\mathbf{z}_{\neq t}^{(L)}, j\right)$$

$$= m^{(L)} \, soft\left({}^{r}\mathbf{z}^{(L)}, j\right) - {}^{r}y_j - \sum_{\substack{t=1 \\ t \neq j}}^{m^{(L)}} (1 - {}^{r}y_t) \, soft\left({}^{r}\mathbf{z}_{\neq t}^{(L)}, j\right)$$

3. Backpropagation

$$C = -\sum_{t=1}^{m^{(L)}} \sum_{r} \left({}^{r}y_t \, ln\, soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, t\right) + \left(1 - {}^{r}y_t\right) ln\left(1 - soft\left({}^{r}\mathbf{z}^{(\mathbf{L})}, t\right)\right)\right)$$

$${}^{r}z_j^{(L)} = b_j^{(L)} + \sum_i w_{i,j}^{(L)} \, {}^{r}a_i^{(L-1)}$$

$$\frac{\partial C}{\partial b_j^{(L)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}z_j^{(L)}} \frac{\partial\, {}^{r}z_j^{(L)}}{\partial b_j^{(L)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}z_j^{(L)}}$$

$$\frac{\partial C}{\partial w_{i,j}^{(L)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}z_j^{(L)}} \frac{\partial\, {}^{r}z_j^{(L)}}{\partial w_{i,j}^{(L)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}z_j^{(L)}} \; {}^{r}a_i^{(L-1)}$$

$$\frac{\partial C}{\partial\, {}^{r}a_i^{(L-1)}} = \sum_j \frac{\partial C}{\partial\, {}^{r}z_j^{(L)}} \frac{\partial\, {}^{r}z_j^{(L)}}{\partial\, {}^{r}a_i^{(L-1)}} = \sum_j \frac{\partial C}{\partial\, {}^{r}z_j^{(L)}} w_{i,j}^{(L)}$$

$${}^{r}a_j^{(l)} = act \; {}^{r}z_j^{(l)} \qquad {}^{r}z_j^{(l)} = b_j^{(l)} + \sum_i w_{i,j}^{(l)} \, {}^{r}a_i^{(l-1)}$$

$$\frac{\partial C}{\partial b_j^{(l)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(l)}} \frac{\partial\, {}^{r}a_j^{(l)}}{\partial\, {}^{r}z_j^{(l)}} \frac{\partial\, {}^{r}z_j^{(l)}}{\partial b_j^{(l)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(l)}} \frac{\partial\, {}^{r}a_j^{(l)}}{\partial\, {}^{r}z_j^{(l)}}$$

$$\frac{\partial C}{\partial w_{i,j}^{(l)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(l)}} \frac{\partial\, {}^{r}a_j^{(l)}}{\partial\, {}^{r}z_j^{(l)}} \frac{\partial\, {}^{r}z_j^{(l)}}{\partial w_{i,j}^{(l)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(l)}} \frac{\partial\, {}^{r}a_j^{(l)}}{\partial\, {}^{r}z_j^{(l)}} \; {}^{r}a_i^{(l-1)}$$

$$\frac{\partial C}{\partial\, {}^{r}a_i^{(l-1)}} = \sum_j \frac{\partial C}{\partial\, {}^{r}a_j^{(l)}} \frac{\partial\, {}^{r}a_j^{(l)}}{\partial\, {}^{r}z_j^{(l)}} \frac{\partial\, {}^{r}z_j^{(l)}}{\partial\, {}^{r}a_i^{(l-1)}} = \sum_j \frac{\partial C}{\partial\, {}^{r}a_j^{(l)}} \frac{\partial\, {}^{r}a_j^{(l)}}{\partial\, {}^{r}z_j^{(l)}} w_{i,j}^{(l)}$$

$${}^{r}a_j^{(1)} = act \; {}^{r}z_j^{(1)} \qquad {}^{r}z_j^{(1)} = b_j^{(1)} + \sum_i w_{i,j}^{(1)} \, {}^{r}x_i$$

$$\frac{\partial C}{\partial b_j^{(1)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(1)}} \frac{\partial\, {}^{r}a_j^{(1)}}{\partial\, {}^{r}z_j^{(1)}} \frac{\partial\, {}^{r}z_j^{(1)}}{\partial b_j^{(1)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(1)}} \frac{\partial\, {}^{r}a_j^{(1)}}{\partial\, {}^{r}z_j^{(1)}}$$

$$\frac{\partial C}{\partial w_{i,j}^{(1)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(1)}} \frac{\partial\, {}^{r}a_j^{(1)}}{\partial\, {}^{r}z_j^{(1)}} \frac{\partial\, {}^{r}z_j^{(1)}}{\partial w_{i,j}^{(1)}} = \sum_r \frac{\partial C}{\partial\, {}^{r}a_j^{(1)}} \frac{\partial\, {}^{r}a_j^{(1)}}{\partial\, {}^{r}z_j^{(1)}} \; {}^{r}x_i$$