Backpropagation process

1 Forward pass

$$z^{[1]} = x^{[0]}W^{[1]} + b^{[1]} \tag{1}$$

$$x^{[1]} = relu\left(z^{[1]}\right) \tag{2}$$

$$z^{[2]} = s = x^{[1]}W^{[2]} + b^{[2]} \tag{3}$$

$$x^{[3]} = p = softmax\left(z^{[2]}\right) \tag{4}$$

$$\mathcal{L} = CE[y, p] + \lambda \sum_{i=0}^{i=1} W_i^2 \tag{5}$$

2 Backward pass

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\frac{y_i}{p_i} \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial s_k} = p_k - y_k \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial w_{pk}^{[2]}} = (p_k - y_k)x_p^{[1]} + 2\lambda w_{pk}^{[2]} \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial b_k^{[2]}} = (p_k - y_k) \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial x_t^{[1]}} = \sum_k (p_k - y_k)w_{tk}^{[2]} \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial z_t^{[1]}} = \sum_k (p_k - y_k)w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial w_{mt}^{[1]}} = \sum_k (p_k - y_k)w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} x_m^{[0]} + 2\lambda w_{mt}^{[1]} \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial b_t^{[1]}} = \sum_k (p_k - y_k)w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} \tag{13}$$

3 Backward pass proofs

Equation 6

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_i} &= \frac{\partial \{CE[y,p] + \lambda \sum_{i=0}^{i=1} W_i^2\}}{\partial p_i} && \text{applying Eq.5} \\
&= \frac{\partial CE[y,p]}{\partial p_i} + \frac{\partial \lambda \sum_{i=0}^{i=1} W_i^2}{\partial p_i} && \text{since } \frac{\partial \lambda \sum_{i=0}^{i=1} W_i^2}{\partial p_i} = 0 \\
&= \frac{\partial}{\partial p_i}\Big\{ - y_i \log(p_i) \Big\} \\
&= -\frac{y_i}{p_i}
\end{aligned}
$$

Equation 7

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial s_k} &= \sum_i \frac{\partial \mathcal{L}}{\partial p_i} \frac{\partial p_i}{\partial s_k} && \text{chain rule} \\
&= -\sum_i \frac{y_i}{p_i} \frac{\partial p_i}{\partial s_k} && \text{applying Eq. 6} \\
&= -\sum_i \frac{y_i}{p_i} \frac{\partial}{\partial s_k} \left\{ \frac{e^{s_i}}{\sum_l e^{s_l}} \right\} && \text{since } p_k = \frac{e^{s_k}}{\sum_l e^{s_l}} \\
&= -\sum_i \frac{y_i}{p_i} \left[(\delta_{i=k}(p_i(1-p_k))) - \delta_{i \neq k}(p_i p_k) \right] \\
&= -\sum_i \frac{y_i}{p_i} \left[p_i \left(\delta_{i=k}(1-p_k) \right) - \delta_{i \neq k} p_k) \right] \\
&= -\sum_i y_i (\delta_{i=k} - p_k) \\
&= -y_k + \sum_i y_i p_k && \text{since } \sum_i y_i \delta_{i=k} = y_k \\
&= p_k - y_k && \text{since } \sum_i y_i = 1
\end{aligned}
$$

Equation 8

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_{pk}^{[2]}} &= \frac{\partial \mathcal{L}}{\partial s_k} \frac{\partial s_k}{\partial w_{pk}^{[2]}} + \frac{\partial}{\partial w_{pk}^{[2]}} \left\{ \lambda \sum_{i=0}^{i=1} W_i^2 \right\} && \text{chain rule + regularization term} \\
&= (p_k - y_k) \frac{\partial s_k}{\partial w_{pk}^{[2]}} + 2\lambda w_{pk}^{[2]} && \text{applying Eq. 7} \\
&= (p_k - y_k) \frac{\partial}{\partial w_{pk}^{[2]}} \left\{ \sum_r x_r^{[1]} w_{rk}^{[2]} + b_k^{[2]} \right\} + 2\lambda w_{pk}^{[2]} && \text{since } s_k = \sum_r x_r^{[1]} w_{rk}^{[2]} + b_k^{[2]} \\
&= (p_k - y_k) x_p^{[1]} + 2\lambda w_{pk}^{[2]}
\end{aligned}
$$

Equation 9

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b_k^{[2]}} &= \frac{\partial \mathcal{L}}{\partial s_k}\frac{\partial s_k}{\partial b_k^{[2]}} && \text{chain rule} \\
&= (p_k - y_k)\frac{\partial s_k}{\partial b_k^{[2]}} && \text{applying Eq. 7} \\
&= (p_k - y_k)\frac{\partial}{\partial b_k^{[2]}}\left\{\sum_r x_r^{[1]} w_{rk}^{[2]} + b_k^{[2]}\right\} && \text{since } s_k = \sum_r x_r^{[1]} w_{rk}^{[2]} + b_k^{[2]} \\
&= (p_k - y_k)\frac{\partial}{\partial b_k^{[2]}}\left\{b_k^{[2]}\right\} && \text{since } \frac{\partial}{\partial b_k^{[2]}}\left\{\sum_r x_r^{[1]} w_{rk}^{[2]}\right\} = 0 \\
&= p_k - y_k
\end{aligned}
$$

Equation 10

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial x_t^{[1]}} &= \sum_k \frac{\partial \mathcal{L}}{\partial s_k}\frac{\partial s_k}{\partial x_t^{[1]}} && \text{chain rule} \\
&= \sum_k \frac{\partial \mathcal{L}}{\partial s_k}\frac{\partial}{\partial x_t^{[1]}}\left\{\sum_r x_r^{[1]} w_{rk}^{[2]} + b_k^{[2]}\right\} && \text{since } s_k = \sum_r x_r^{[1]} w_{rk}^{[2]} + b_k^{[2]} \\
&= \sum_k \frac{\partial \mathcal{L}}{\partial s_k} w_{tk}^{[2]} \\
&= \sum_k (p_k - y_k) w_{tk}^{[2]} && \text{applying Eq. 7}
\end{aligned}
$$

Equation 11

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial z_t^{[1]}} &= \frac{\partial \mathcal{L}}{\partial x_t^{[1]}}\frac{\partial x_t^{[1]}}{\partial z_t^{[1]}} && \text{chain rule} \\
&= \Big[\sum_k (p_k - y_k) w_{tk}^{[2]}\Big]\frac{\partial x_t^{[1]}}{\partial z_t^{[1]}} && \text{applying Eq. 10} \\
&= \Big[\sum_k (p_k - y_k) w_{tk}^{[2]}\Big]\frac{\partial}{\partial z_t^{[1]}}\left\{relu(z_t^{[1]})\right\} \\
&= \sum_k (p_k - y_k) w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\}
\end{aligned}
$$

Equation 12

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_{mt}^{[1]}} &= \frac{\partial \mathcal{L}}{\partial z_t^{[1]}} \frac{\partial z_t^{[1]}}{\partial w_{mt}^{[1]}} + \frac{\partial}{\partial w_{mt}^{[1]}} \left\{ \lambda \sum_{i=0}^{i=1} W_i^2 \right\} && \text{chain rule + regularization term} \\
&= \frac{\partial \mathcal{L}}{\partial z_t^{[1]}} \frac{\partial z_t^{[1]}}{\partial w_{mt}^{[1]}} + 2\lambda w_{mt}^{[1]} \\
&= \Big[\sum_k (p_k - y_k) w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} \Big] \frac{\partial z_t^{[1]}}{\partial w_{mt}^{[1]}} + 2\lambda w_{mt}^{[1]} && \text{applying Eq. 11} \\
&= \Big[\sum_k (p_k - y_k) w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} \Big] \frac{\partial}{\partial w_{mt}^{[1]}} \left\{ \sum_n x_n^{[0]} w_{nt}^{[1]} + b_t^{[1]} \right\} + 2\lambda w_{mt}^{[1]} && \text{since } z_t = \sum_n x_n^{[0]} w_{nt}^{[1]} + b_t^{[1]} \\
&= \sum_k (p_k - y_k) w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} x_m^{[0]} + 2\lambda w_{mt}^{[1]}
\end{aligned}
$$

Equation 13

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b_t^{[1]}} &= \frac{\partial \mathcal{L}}{\partial z_t^{[1]}} \frac{\partial z_t^{[1]}}{\partial b_t^{[1]}} && \text{chain rule} \\
&= \Big[\sum_k (p_k - y_k) w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} \Big] \frac{\partial z_t^{[1]}}{\partial b_t^{[1]}} && \text{since } z_t = \sum_n x_n^{[0]} w_{nt}^{[1]} + b_t^{[1]} \\
&= \Big[\sum_k (p_k - y_k) w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\} \Big] \frac{\partial}{\partial b_t^{[1]}} \left\{ \sum_n x_n^{[0]} w_{nt}^{[1]} + b_t^{[1]} \right\} && \text{since } z_t = \sum_n x_n^{[0]} w_{nt}^{[1]} + b_t^{[1]} \\
&= \sum_k (p_k - y_k) w_{tk}^{[2]} \mathbb{1}\{z_t^{[1]} \geq 0\}
\end{aligned}
$$