Let score $f_i = W_i^T X + b_i$ for i_{th} class and $softmax\ function\ S_i = \frac{e^{f_i}}{\sum_i^n e^{f_i}}$

derivation of $activation\ function\ (sigmoid)$:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \frac{0-\left(-e^{-x}\right)}{\left(1+e^{-x}\right)^2} = \frac{1+e^{-x}-1}{\left(1+e^{-x}\right)^2} = \frac{1+e^{-x}}{\left(1+e^{-x}\right)^2} - \frac{1}{\left(1+e^{-x}\right)^2} = \sigma(x) - \sigma(x)^2 = \sigma(x)(1-\sigma(x))$$

derivation of $loss\ fuction\ (softmax\ cross-entropy\ with\ L2\ norm\ of\ weight)$:

$$L = -\log(S_i) + \frac{1}{2}\lambda\sum W^2$$

Let $S_i = softmax(a_i)\ and\ L_{pre} = -\log(S_i)$, $L_{reg} = \frac{1}{2}\lambda\sum W^2$, N is the number of classification.

$$\frac{\partial L}{\partial W_i} = \frac{\partial L_{pre}}{\partial W_i} + \frac{\partial L_{reg}}{\partial W_i}$$

$$\frac{\partial L_{reg}}{\partial W_i} = \frac{\partial \frac{1}{2}\lambda\sum W^2}{\partial W_i} = \lambda W_i$$

For output unit, acccording to chain rule:

$$\frac{\partial L_{pre}}{\partial W_i} = \frac{\partial L_{pre}}{\partial S_y}\frac{\partial S_y}{\partial f_i}\frac{\partial f_i}{\partial W_i}$$

$$\frac{\partial L_{pre}}{\partial S_i} = \frac{\partial(-\log S_y)}{\partial S_y} = -\frac{1}{S_y}$$

$$\frac{\partial f_i}{\partial W_i} = \frac{\partial W_i^T X + b_i}{\partial W_i} = X$$

when $i = y$:

$$\frac{\partial S_y}{\partial f_i} = \frac{\partial\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right)}{\partial f_i} = \frac{e^{f_y}\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right) - e^{f_y}e^{f_i}}{\left(\sum_i^n e^{f_i}\right)^2} = \frac{e^{f_y}}{\sum_i^n e^{f_i}}\left(\frac{\sum_i^n e^{f_i} - e^{f_y}}{\sum_i^n e^{f_i}}\right) = S_y(1-S_i)$$

when $i \neq y$:

$$\frac{\partial S_y}{\partial f_i} = \frac{\partial\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right)}{\partial f_i} = \frac{0\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right) - e^{f_y}e^{f_i}}{\left(\sum_i^n e^{f_i}\right)^2} = \frac{-e^{f_y}e^{f_i}}{\left(\sum_i^n e^{f_i}\right)^2} = -S_y S_i$$

In summary:

$$\frac{\partial L}{\partial W_i} = \begin{cases} -\frac{1}{S_y} S_y(1-S_i)X + \lambda W_i = & (S_i - 1)X + \lambda W_i & if\ i = y \\ -\frac{1}{S_y}(-S_y S_i)X + \lambda W_i = & S_i X + \lambda W_i & if\ i \neq y \end{cases}$$

For hidden unit, acccording to chain rule:

$$\frac{\partial L_{pre}}{\partial W_i} = \frac{\partial L_{pre}}{\partial S_y}\frac{\partial S_y}{\partial f_i}\frac{\partial f_i}{\partial \sigma_i}\frac{\partial \sigma_i}{\partial f_i}\frac{\partial f_i}{\partial W_i}$$

$$\frac{\partial L_{pre}}{\partial S_i} = \frac{\partial(-\log S_y)}{\partial S_y} = -\frac{1}{S_y}$$

$$\frac{\partial f_i}{\partial \sigma_i} = \frac{\partial W_i^T \sigma_i + b_i}{\partial W_i} = \sigma_i$$

$$\frac{\partial \sigma_i}{\partial f_i} = \sigma_i(1-\sigma_i)$$

$$\frac{\partial f_i}{\partial W_i} = \frac{\partial W_i^T X + b_i}{\partial W_i} = X$$

when $i = y$:

$$\frac{\partial S_y}{\partial f_i} = \frac{\partial\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right)}{\partial f_i} = \frac{e^{f_y}\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right) - e^{f_y}e^{f_i}}{\left(\sum_i^n e^{f_i}\right)^2} = \frac{e^{f_y}}{\sum_i^n e^{f_i}}\left(\frac{\sum_i^n e^{f_i} - e^{f_y}}{\sum_i^n e^{f_i}}\right) = S_y(1-S_i)$$

when $i \neq y$:

$$\frac{\partial S_y}{\partial f_i} = \frac{\partial\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right)}{\partial f_i} = \frac{0\left(\frac{e^{f_y}}{\sum_i^n e^{f_i}}\right) - e^{f_y} e^{f_i}}{\left(\sum_i^n e^{f_i}\right)^2} = \frac{-e^{f_y} e^{f_i}}{\left(\sum_i^n e^{f_i}\right)^2} = -S_y S_i$$

In summary:

$$\frac{\partial L}{\partial W_i} = \begin{cases} -\frac{1}{S_y} S_y (1 - S_i) \sigma_i^2 (1 - \sigma_i) X + \lambda W_i = & (S_i - 1) \sigma_i^2 (1 - \sigma_i) X + \lambda W_i & if\ i = y \\ -\frac{1}{S_y} (-S_y S_i) \sigma_i^2 (1 - \sigma_i) X + \lambda W_i = & S_i \sigma_i^2 (1 - \sigma_i) X + \lambda W_i & if\ i \neq y \end{cases}$$