$$L = \sum_{j=1}^{n} \log P(y_j^* \mid x_j)$$

Log Laws:

$$\log \frac{m}{n} = \log m - \log n$$

jth example in training data

$$= \sum_{j=1}^{n} \log \frac{e^{w^T f(x_j, y_j^*)}}{\sum_y e^{w^T f(x_j, y)}}$$

$$= \sum_{j=1}^{n} \left(\log\left(e^{w^T f(x_j, y_j^*)}\right) - \log\left(\sum_y e^{w^T f(x_j, y)}\right) \right)$$

$$= \sum_{j=1}^{n} \left(w^T f(x_j, y_j^*) - \log\left(\sum_y e^{w^T f(x_j, y)}\right) \right)$$

$$L(x_j, y_j^*) = \underbrace{\boxed{W^T f(x_j, y_j^*)}}_{F} - \log \underbrace{\boxed{\sum_y e^{W^T f(x_j, y)}}}_{E}$$

$$\frac{\partial L}{\partial w_i} = \frac{\partial F}{\partial w_i} - \frac{\partial \log E}{\partial w_i} = \frac{\partial F}{\partial w_i} - \frac{1}{E} \cdot \frac{\partial E}{\partial w_i}$$

i'th feature/weight

$$= f_i(x_j, y_j^*) - \frac{\sum_y f_i(x_j, y^*) \cdot e^{W^T f(x_j, y)}}{\sum_y e^{W^T f(x_j, y)}}$$

(1)

$$F = W^T f(x_j, y_j^*) \qquad W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_i \\ \vdots \end{bmatrix} \qquad f(x_j, y_j^*) = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_i \\ \vdots \end{bmatrix}$$

$$\underline{\underline{\frac{\partial F}{\partial w_i}}} = f_i(x_j, y_j^*)$$

(2)

$$E = \sum_y e^F$$

$$\frac{\partial E}{\partial w_i} = \sum_y \frac{\partial e^F}{\partial w_i} = \sum_y \boxed{\frac{\partial e^F}{\partial F}} \cdot \frac{\partial F}{\partial w_i} = \sum_y \boxed{e^F} \cdot \underline{\underline{\frac{\partial F}{\partial w_i}}}$$

$$= \sum_y e^{W^T f(x_j, y)} \cdot f_i(x_j, y_j^*)$$

Chain Rule:

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

$$f(x) = e^x \rightarrow \frac{\partial f}{\partial x} = e^x$$

$$f(x) = \log(x) \rightarrow \frac{\partial f}{\partial x} = \frac{1}{x}$$

$$f(x) = c + x \rightarrow \frac{\partial f}{\partial x} = 1$$