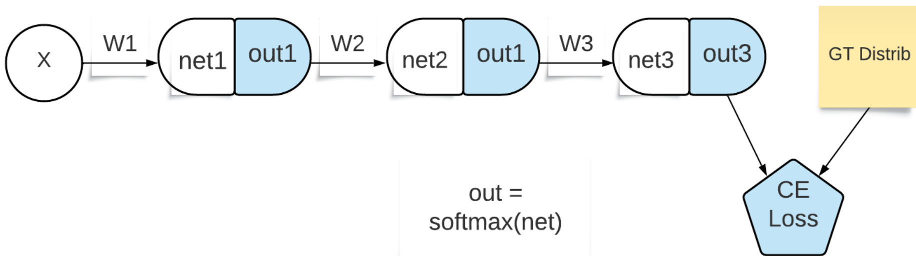
Symbols



For simplicity:

- Let net3 = z
- Let out3 = y'   = softmax(z)
- Let y = ground truth   (always constants in below derivatives)
- The sum of y = 1
- The sum of y' = 1

Recall that

The **cross-entropy** loss for a classification task is defined as follows for a **single** sample:

$$L(y, \hat{y}) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

The softmax derivative is 2 cases

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1 - \hat{y}_i) & \text{if } i = j \\ -\hat{y}_i \hat{y}_j & \text{if } i \neq j \end{cases}$$

The derivative of the log function

Derivative of Common Logarithm:

$$\frac{d}{dx}\log_a(x) = \frac{1}{x\ln(a)}$$

Derivative of Natural Logarithm:

$$\frac{d}{dx}\ln(x) = \frac{1}{x}$$

The Solution

We'll apply the chain rule for differentiation:

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{K} \frac{\partial L}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_j}$$

- The first term is direct based on log derivative
- The second is the softmax derivative
- We put together
  - Break the 2 cases
  - Simplify
  - Join

For the first term

$$L(y, \hat{y}) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

$$\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

Observe

- All the terms in the summation are canceled except the relevant one: -yi log yi’
    - -yi is just a constant
    - Then just apply the derivative of the log yi’ = 1 / yi’

Put together

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{K} \left(-\frac{y_i}{\hat{y}_i} \right) \frac{\partial \hat{y}_i}{\partial z_j}$$

Substitute the softmax derivative as 2 cases

- Jth term + sum of non Jth terms

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1-\hat{y}_i) & \text{if } i = j \\ -\hat{y}_i\hat{y}_j & \text{if } i \neq j \end{cases}$$

$$\frac{\partial L}{\partial z_j} = -y_j(1-\hat{y}_j) + \sum_{i \neq j} y_i \hat{y}_j$$

- First term: Distribute
- Second term: Get yj’ out the sum as it is constant (index on i not j)

$$\frac{\partial L}{\partial z_j} = -y_j + y_j\hat{y}_j + \hat{y}_j \sum_{i \neq j} y_i$$

- The shared symbol of the last 2 terms is yj’
- So merge them into yj’ * Sum over ALL indices yi
    - But this is sum of probability, so = 1
    - Then the last 2 terms = yj’

- So in total -yj + yj’

$$\frac{\partial L}{\partial z_j} = \hat{y}_j - y_j$$