1. Linear

param. $W = \overset{\#\text{ in}}{\overset{\#\text{ out}}{[\quad]}}$ $\quad b = \overset{1}{\overset{\#\text{ out}}{[\quad]}}$

input. $\overset{\text{batch\_size}}{\underset{\#\text{ in}}{[\quad]}}$

$$\text{output} = \overset{\text{batch\_size}}{[W]} \cdot \text{input} + [b, b, b \cdots b]$$

Compute Gradiant

@ w.r.t "w"

$$f(w) = W \cdot \overset{\text{input}}{x} + b \Rightarrow \text{grad.w} = \frac{dv\_output \cdot input^T}{batch\_size}$$

$\because \frac{\partial}{\partial w} f(w) = x^T$ recall $\frac{\partial}{\partial w} aw = a$, $\frac{\partial}{\partial w} wx = \frac{\partial}{\partial w}(x^T w^T)^T$

---

① w.r.t "b"

$$f(b) = w \cdot x + b \Rightarrow \frac{\partial}{\partial b} f = dv\_output \cdot \overset{1}{[1]}_{\#\, batch\_size}$$

$\frac{\partial}{\partial b} f(b) = 1$ $\Leftrightarrow$ sum(dv_output, 2)

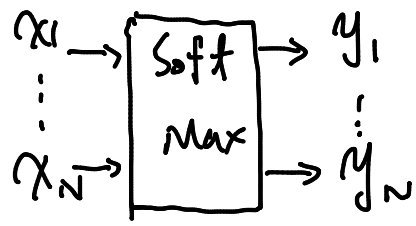
③ $dv\_input = \left(\frac{\partial f}{\partial x}\right)^T dv\_output = W^T \cdot dv\_output$

2. Softmax

$$y_i = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

Because Softmax is the last layer. so, it only has $\frac{\partial f(x)}{\partial x_i}$, or say the original dv_output.

Our network is



We know $\frac{\partial L}{\partial y_i}$, we want $\frac{\partial L}{\partial x_i}$

According to chain rule.

$$\frac{\partial L}{\partial x_i} = \sum_{K=1}^{N} \frac{\partial L}{\partial y_k} \cdot \frac{\partial y_k}{\partial x_i}$$

for $\frac{\partial}{\partial x_i} y_k$ let $\Gamma = \sum_{j=1}^{N} e^{x_j}$

* if $i \neq k$

$$\frac{\partial}{\partial x_i} y_k = \frac{\partial}{\partial x_i} \cdot \frac{e^{x_k}}{\sum_{j \neq i} e^{x_j} + e^{x_i}} = -\frac{e^{x_i} \cdot e^{x_k}}{\left(\sum_{j \neq i} e^{x_j} + e^{x_i}\right)^2} = -\frac{e^{x_i + x_k}}{\Gamma^2}$$

* if $i = k$

$$\frac{\partial}{\partial x_i} y_i = \frac{\partial}{\partial x_i} \frac{e^{x_i}}{\sum_{j \neq i} e^{x_i} + e^{x_i}} = \frac{e^{x_i}\Gamma - e^{x_i} \cdot e^{x_i}}{\Gamma^2} = \frac{e^{x_i}}{\Gamma} - \frac{e^{x_i + x_k}}{\Gamma^2}$$

$$\therefore \frac{\partial L}{\partial x_i} = \sum_{k=1}^{N} \frac{\partial L}{\partial y_k} \frac{\partial y_k}{\partial x_i}$$

$$= \sum_{k \neq i} \frac{\partial L}{\partial y_k} \cdot \left(-\frac{e^{x_i + x_k}}{\Gamma^2}\right) + \frac{\partial L}{\partial y_i} \cdot \left(\frac{e^{x_i}}{\Gamma} - \frac{e^{x_i + x_i}}{\Gamma^2}\right)$$

$$= \sum_{k=1}^{N} \frac{\partial L}{\partial y_k} \left(-\frac{e^{x_i}}{\Gamma} \cdot \frac{e^{x_k}}{\Gamma}\right) + \frac{\partial L}{\partial y_i} \frac{e^{x_i}}{\Gamma}$$

3. Cross entropy

$$f(\underline{X}) = -\frac{\sum_{t \in T} \log x_t}{N}$$

T is the set of target label, or ground truth

N is batch size.

Also, we don't have $\frac{\partial}{\partial w} f(\underline{X})$

$$\frac{\partial}{\partial x_t} f(\underline{X}) = \frac{\partial}{\partial x_t} -\frac{\sum_{t \in T} \log x_t}{N} = -\frac{1}{x_t \cdot N}$$

4. ReLU

$$\frac{\partial}{\partial x} L = \frac{\partial L}{\partial y} \cdot \mathbb{1}(x \geq 0)$$