② Back. Prop. Derivation.

Assume,
$$\boxed{J = \frac{1}{2m}\sum (y-a^L)^2}$$

① $\Delta a^L \sim \dfrac{\partial J}{\partial a^L} = \dfrac{1}{2m}\sum \dfrac{\partial}{\partial a^L}(y-a^L)^2$

$= \dfrac{1}{\cancel{2}m}\sum \cancel{2}\cdot(y-a^L)\dfrac{\cancel{\partial a^L}}{\cancel{\partial a^L}}$

$= \dfrac{1}{m}\sum (y-a^L) = \text{mean}(y-a^L)$

$= \text{L1 loss} \rightarrow (\text{scalar})$

② $\underset{(n^L,m)}{\Delta z^L} \sim \dfrac{\partial J}{\partial z^L} = \dfrac{\partial J}{\partial a^L}\cdot\dfrac{\partial a^L}{\partial z^L} = \Delta a^L \cdot g^{L'}(z^L)$

$= \underset{(\text{scalar})}{\Delta a^L} \cdot \underset{(n^L,m)}{a^{L'}} \rightarrow (n^L,m)$

③ $\underset{(n^{L-1},m)}{\Delta a^{L-1}} \sim \dfrac{\partial J}{\partial a^{L-1}} = \dfrac{\partial J}{\partial a^L}\cdot\dfrac{\partial a^L}{\partial z^L}\cdot\dfrac{\partial z^L}{\partial a^{L-1}}$

$= \underset{(n^L,m)}{\Delta z^L}\cdot\underset{(n^L,n^{L-1})}{w^L}$

$= \underset{(n^{L-1},n^L)}{w^{L^T}} \times \underset{(n^L,m)}{\Delta z^L} \rightarrow (n^{L-1},m)$

$\dfrac{\partial z^L}{\partial a^{L-1}} = \dfrac{\partial}{\partial a^{L-1}}(w^L a^{L-1} + b^L) = w^L$

④ $\underset{(n^L,n^{L-1})}{\Delta w^L} \sim \Delta a^L\cdot a^{L'}\cdot a^{L-1} = \underset{(n^L,m)}{\Delta z^L} \times \underset{(n^{L-1},m)^T}{a^{L-1^T}}$

$\Delta b^L \sim db^L = \underset{(n^L,m)}{\Delta z^L}$

$(\Delta z^L, a^{L-1})$

⑤ $\underset{(n^{L-1},m)}{\Delta z^{L-1}} \sim \dfrac{\partial J}{\partial z^{L-1}} = \dfrac{\partial J}{\partial a^L}\cdot\dfrac{\partial a^L}{\partial z^L}\cdot\dfrac{\partial z^L}{\partial a^{L-1}}\cdot\dfrac{\partial a^{L-1}}{\partial z^{L-1}} = \underset{(n^{L-1},m)}{w^{L^T}\Delta z^L}\cdot\underset{(n^{L-1},m)}{a^{L-1'}} \rightarrow (n^{L-1},m)$

⑥ $\Delta a^{L-2} \sim \dfrac{\partial J}{\partial a^{L-2}} = \Delta z^{L-1}\dfrac{\partial z^{L-1}}{\partial a^{L-2}} = \Delta z^{L-1}w^{L-1} = w^{L-1^T}\times\Delta z^{L-1}$

⑦ $\Delta w^{L-1} \sim dw^{L-1} = \Delta z^{L-1}\times a^{L-2^T}$

$\Delta b^{L-1} \sim \Delta z^{L-1}$

⋮

Backpropagation

Approach:

Variable	Dictionary
	y, a^L
Δa^L	Δa^L
Δz^L	$a^{L'}$
	Δz^L
Δb^L	a^{L-1}
Δw^L	⊗
	$\Delta z^L, w^L$
Δa^{L-1}	Δa^{L-1}
Δz^{L-1}	$a^{L-1'}$
	Δz^{L-1}
Δb^{L-1}	a^{L-2}
Δw^{L-1}	⊗

⋮

Gradient Dec.

$w^L = w^L - \eta\Delta w^L$

$b^L = b^L - \eta\Delta b^L$

$w^{L-1} = w^{L-1} - \eta\Delta w^{L-1}$

$b^{L-1} = b^{L-1} - \eta\Delta b^{L-1}$

⋮

loop.

F-prop

↓

B-prop

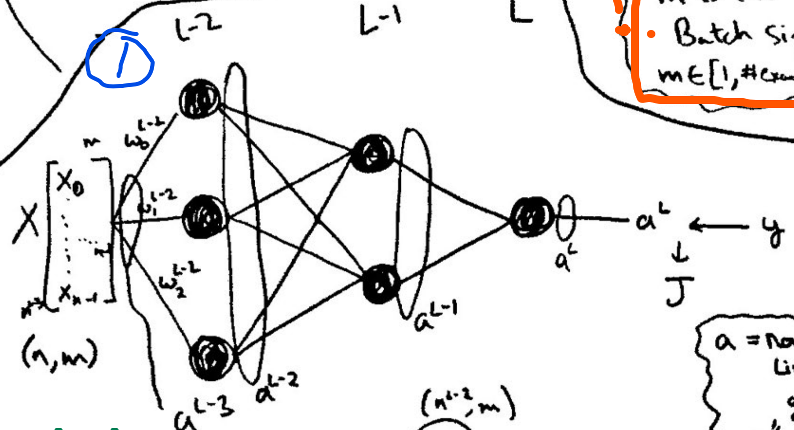
↓

Grad Dec

output Loss/Cost

m is the Batch size!

$m \in [1, \#\text{examples}]$



a = Non-Linearity

$g(z)$

$g'(z) = a'$

$a^{L-3} = X \rightarrow (n^{L-3}, m)$

$w^{L-2} \rightarrow (n^{L-2}, n^{L-3})$

$b^{L-2} \rightarrow (n^{L-2}, m)$

$z^{L-2} = \underset{(n^{L-2},n^{L-3})}{w^{[L-2]}}\underset{(n^{L-3},m)}{a^{[L-3]}} + \underset{(n^{L-2},m)}{b^{[L-2]}} \rightarrow (n^{L-2}, m)$

$a^{L-2} = g^{L-2}(z^{L-2}) \rightarrow (n^{L-2}, m)$

In general,

$dw^L \rightarrow w^L \rightarrow (n^L, n^{L-1})$

$db^L \rightarrow b^L \rightarrow (n^L, m)$

$dz^L \rightarrow z^L \rightarrow (n^L, m)$

$da^L \rightarrow a^L \rightarrow (n^L, m)$

$a^{L-1} \rightarrow (n^{L-1}, m)$

$a^L \rightarrow (n^L, m) \qquad y \rightarrow (n^L, m)$

$J \rightarrow (1, m) \Rightarrow \mathcal{J} \rightarrow \dfrac{\sum J}{m}$

for L2 loss,

$J = \dfrac{1}{2}(y-a^L)^2$

$\mathcal{J} = \dfrac{1}{2}\dfrac{1}{m}\sum(y-a^L)^2$

[redacted]

Agamdeep S. Chopra