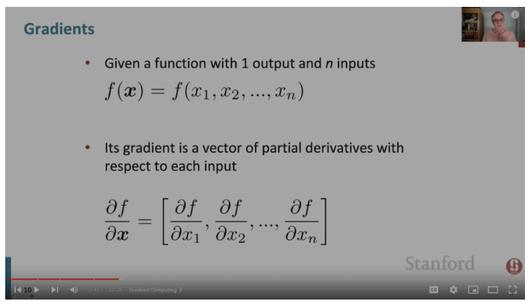
Notes

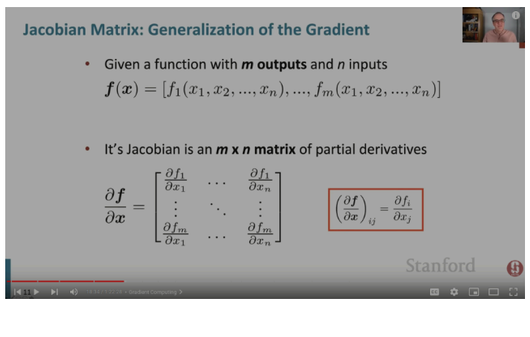
12 February 2024 08:21

Gradients:

Gradients with 1 variable functions:

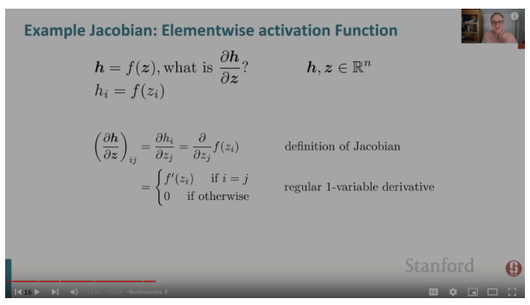


Gradients with Multi-variables functions:

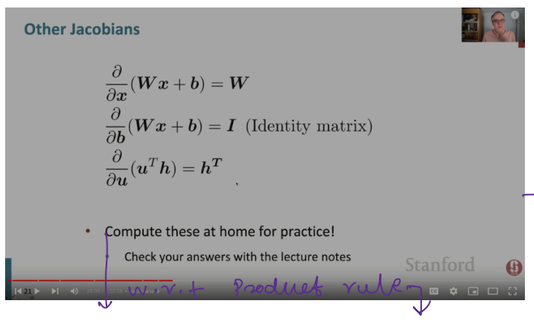


Some example Jacobians:

Activation Functions: e.g. Sigmoid(x)



Some Other Jacobians:



$\rightarrow \frac{\partial}{\partial x}(Wx+b) = \frac{\partial}{\partial x}(Wx) + \frac{\partial}{\partial x}(b) \Rightarrow \frac{\partial}{\partial f}(a+b) = \frac{\partial}{\partial f}(a) + \frac{\partial}{\partial f}(b)$

$= W + 0 \Rightarrow \frac{\partial}{\partial x}(const) = 0$

$\rightarrow \frac{\partial}{\partial b}(Wx+b) = 0 + 1 \Rightarrow$ similar to above

$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow$ Considering Jacobians.

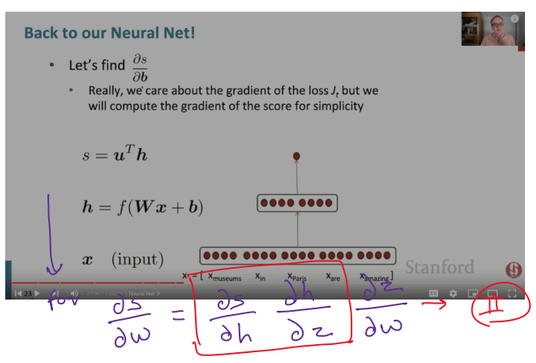
w.r.t. Product rule

$\frac{\partial}{\partial u}(u^T h) = \frac{\partial}{\partial u}(u^T)\cdot h + \frac{\partial}{\partial u}(h)\cdot u^T$

$= Is^T \cdot h + 0$

$= h^T$

Analytical Calculation of Gradients:



$\rightarrow$ simplifying $f(Wx+b)$ as $f(z)$

$\frac{\partial s}{\partial b} = \left[\frac{\partial s}{\partial h} \cdot \frac{\partial h}{\partial z}\right] \frac{\partial z}{\partial b} \rightarrow$ (I)

$= u^T \cdot \text{diag}(f'(z))$ Identity $\rightarrow \partial(Wx+b)$ w.r.t. b.

$\rightarrow$ activation fun Jacobian

$= u^T \cdot f'(z)$

$\frac{\partial s}{\partial w} = \left[\frac{\partial s}{\partial h} \frac{\partial h}{\partial z}\right] \frac{\partial z}{\partial w} \rightarrow$ (II)

eg. I & II share same computations. so to avoid duplications.

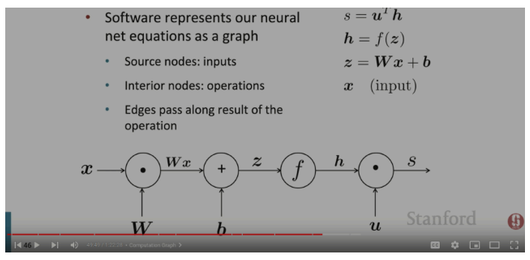
$\delta = \frac{\partial s}{\partial h} \cdot \frac{\partial h}{\partial z}$

$\frac{\partial s}{\partial b} = \delta \cdot \frac{\partial z}{\partial b}$ &

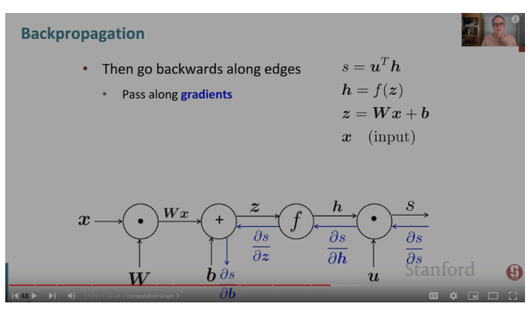
$\frac{\partial s}{\partial w} = \delta \cdot \frac{\partial z}{\partial w}$.

Backpropagation

Starting with Forward pass: Passing inputs through the computational graph from left to right

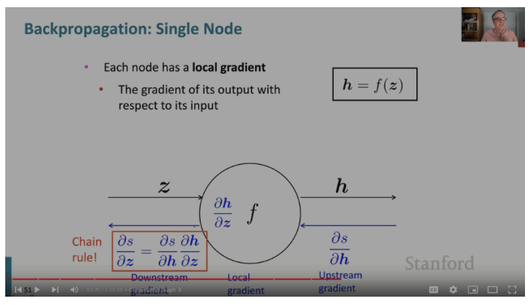


Backward pass: Passing gradients from right to left



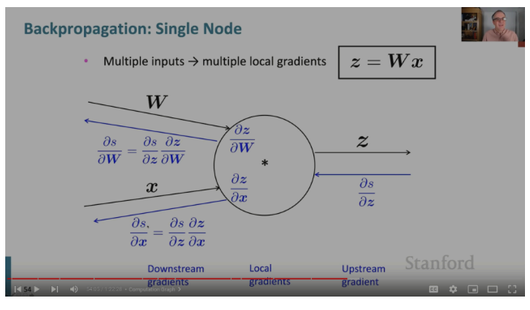
Example of single node with single input:

Upstream gradient: gradient of final output/error wrt. Node's output
Local gradient: gradient of node's output wrt. node's input
Downstream gradient: gradient of final output/ error wrt. node's input

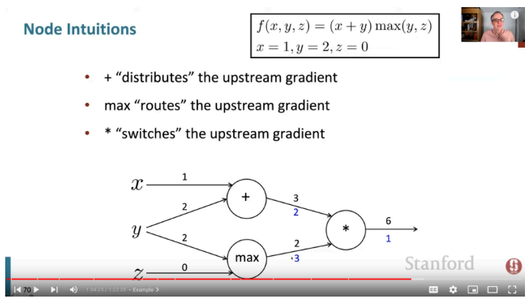


Example of single node with multiple inputs:

Only difference is in local gradients step. For each input of a node, we calculate a local gradient.



Impact of Node Types on Gradients



Manual Gradient Checking: