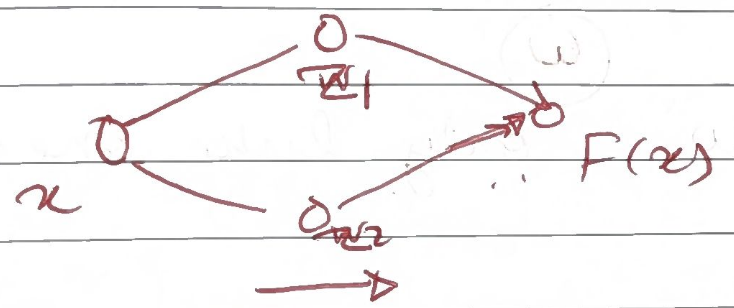
(1) How much time does it take to compute the "Hessian" of a network?

Say we have "n" nodes in the computational graph

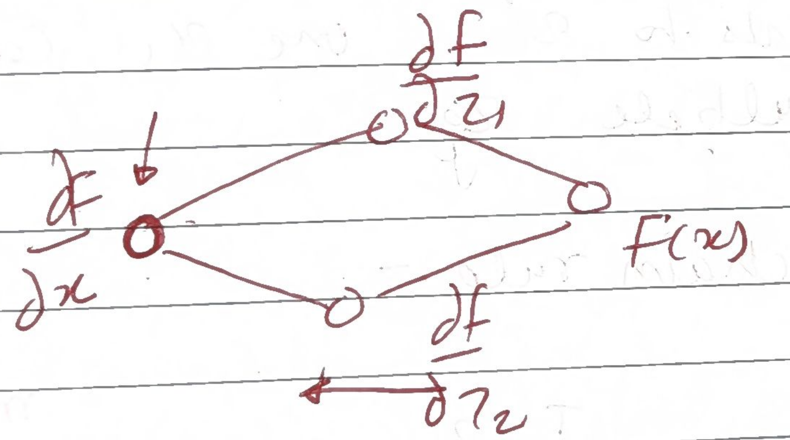
Forward pass $O(n)$ Backward pass $O(n)$

<u>Hessian</u> : Size of hessian $O(n^2)$

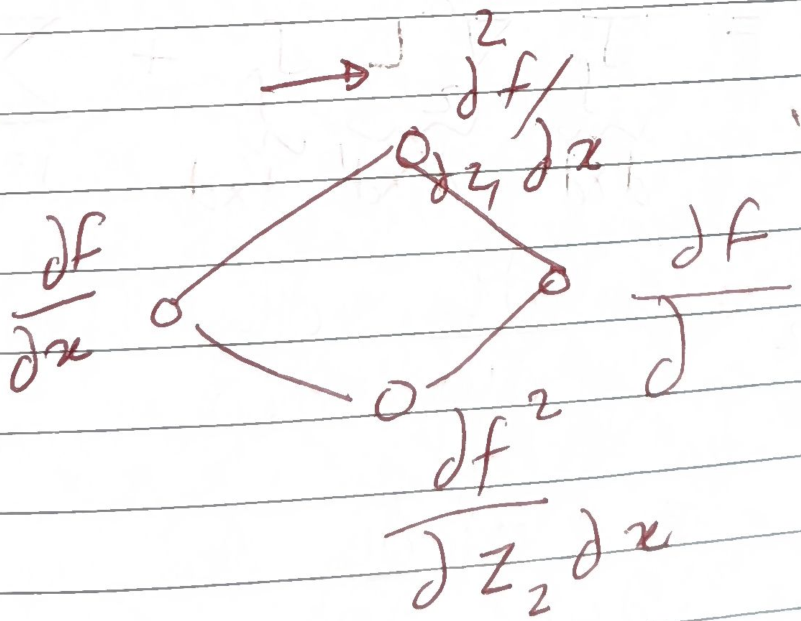
param $= O($Computational graph$)$ (n)



forward



backw



~~δf~~

"backward backward for each node of interest.

$O(n \times n) = O(n^2)$

Computation

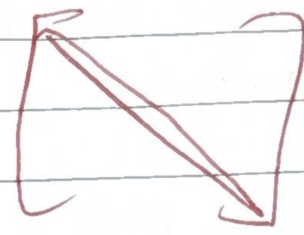
$O(n^2)$ is too large. || $O(N, n^2)$ Examples.

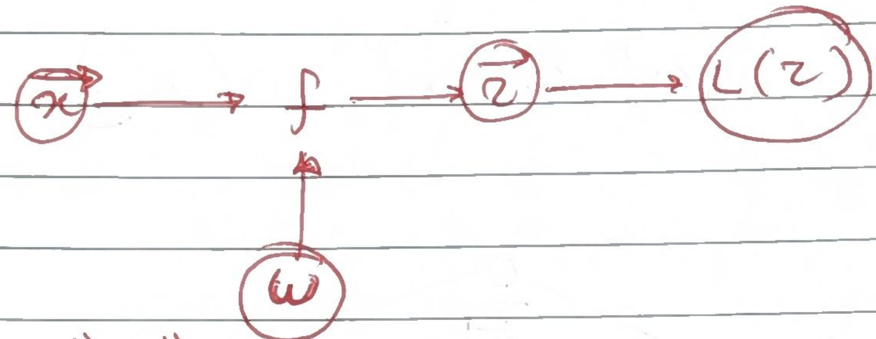
Page No.: paran
Date:

(2) Hessian diagonal

What if we care only about

$$\frac{\partial^2 L}{\partial w_i^2}$$

$H \rightarrow$ 



* Every "w" only links one of $\vec{x}$ to $\vec{z}$

* $\vec{x}$ links to $\vec{z}$: one x_i can ~~be~~ link to multiple z_j

Hessian chain rule:-

$$\underbrace{\nabla_x^2 L}_{d \times d'} = \underbrace{J_f^T}_{d \times d'} \underbrace{\nabla_z^2 L}_{d' \times d'} \underbrace{J_f}_{d' \times d} + \sum_{i=1}^{m} \underbrace{\frac{\partial L}{\partial z_i}}_{1} \underbrace{\nabla_x^2 z_i}_{d \times d}$$

$$\frac{\partial^2 L}{\partial w_i^2} = \frac{\partial L}{\partial w} \frac{\partial L}{\partial w}$$

$$= \frac{\partial}{\partial w} \overset{x_i}{\cancel{z_i}} \frac{\partial L}{\partial z_i}$$

$$= \frac{\partial}{\partial z}$$

$$\nabla_x^2 L = \underbrace{J_f^T \nabla_z^2 L \, J_f}_{\text{global curvature through the local gradient}} + \underbrace{\sum_{i=1}^{m} \frac{\partial L}{\partial z_i}}_{\text{global gradient}} \underbrace{\nabla_x^2 z_i}_{\text{local curvature}}$$

~~global curvature that~~

We only need local hessian's to be computed.

$\vec{x} \longrightarrow f \longrightarrow \vec{z} \longrightarrow L(\vec{z})$

$O(d^2)$ hessian

$O(d^3)$ hessian

$d \times d$

W

linear layer.

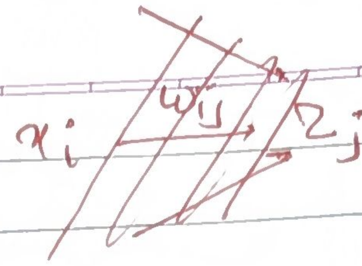
$O(n)$ Computation for hessian diagonal of all param

→ Can be acheived in $O($ single "non-traditional" backward pass.

+ W

$$\cancel{\frac{\partial^2 L}{\partial w_{ij}^2} = \frac{\partial}{\partial z_j} \frac{\partial}{\partial}}$$

W.r.t (w_{ij})



$x_i \rightarrow f \rightarrow z_j$, $\uparrow w$

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}} = \frac{\partial L}{\partial z_j} f'(w)$$

$$\frac{\partial^2 L}{\partial w_{ij}^2} = \frac{\partial}{\partial w_{ij}} \left(f'(x_i) \frac{\partial L}{\partial z_j} \right)$$

$$= f''(w_{ij}) \frac{\partial L}{\partial z_j} + f'(x) \cdot f'(x) \frac{\partial^2 L}{\partial z_j^2}$$

$$\boxed{\frac{\partial^2 L}{\partial w_{ij}^2} = f''(w_{ij}) \frac{\partial L}{\partial z_j} + \left(f'(x)\right)^2 \frac{\partial^2 L}{\partial z_j^2}}$$

OBD :- Single backward pass $\Rightarrow$ do OBD

$\rightarrow$ Hessian Approximation in a specific case of ℓ_2 loss

Let $o = F(W, x)$ $\qquad x \in R^{in}$

$\qquad\qquad\qquad\qquad o \in R^1$

t : target value

$$\text{loss} = \frac{1}{2} \sum_{1}^{n} \left(F(W, x) - t \right)^2$$

$$E = \frac{1}{2N} \sum_{i=1}^{N} (t - o)^2 \qquad o = f(\vec{w}, \vec{x})$$

$$\underbrace{\frac{\partial E}{\partial W}}_{n \times 1} = \frac{1}{2N} \sum_{i=1}^{N} 2(t - o) \cdot \frac{\partial}{\partial W} F(\vec{w}, \vec{x})$$

$$= \cancel{\frac{1}{2N}} \cancel{\sum_{i=1}^{N}}$$

$$\underset{n \times n}{\frac{\partial^2 E}{\partial W^2}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\underset{n \times 1}{\left(\frac{\partial}{\partial W} F(\vec{w}, \vec{x}) \right)} \underset{1 \times n}{\frac{\partial f(\vec{w}, \vec{x})}{\partial W}^T} + 2(t - o) \frac{\partial^2 f(\vec{w}, \vec{x})}{\partial W^2} \right]$$

If I assume $t \approx o$ can ignore second term

$$\underset{n \times n}{\frac{\partial^2 E}{\partial W^2}} = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\frac{\partial F}{\partial W} \frac{\partial F}{\partial W}^T}$$

Covariance matrix of every sample

N

Hessian can be approximated by a sample covariance of gradients.

Note this only holds when MSE is the loss.

Exercise: what if E has cross entropy as the loss?

Now

$$H = \frac{1}{N} \underbrace{\sum_{i}^{N} \underset{n \times 1}{X_i} \underset{1 \times n}{X_i^T}}$$

Is there something interesting about this?

Complexity of H: $O(Nn^2)$

What about H^{-1}? ~ generally $O(n^3)$

Does it help me with H^{-1} ~~computat~~ Computation?

Key: Rank 1 update

$$(A + uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

$H_0 \quad H_1 \quad H_2 \ldots \quad H_N$

$H_0^{-1} \quad H_1^{-1} \quad H_2^{-1} \ldots \quad H_N^{-1}$

$$H_n^{-1} = H_{n-1}^{-1} + \frac{H_{n-1}^{-1} u v^T H_{n-1}^{-1}}{1 + v^T H_{n-1}^{-1} u}$$

$$u = v = (\nabla F)_n$$

$H_0 = \lambda I$ λ is small.

Computation is insensitive to λ when small.

"λ" $\neq 0$ is equivalent ~~of~~ adding a regularisation term

$$E = \frac{1}{2N} \sum \left(F(x, \vec{w}) - t \right)^2 + \frac{1}{2} \lambda \|w\|_2^2$$