20A_Q2 April 19th

III Automatic Differentiation

$$l(x) = f(g(h(x)))$$

$$\boxed{x} \xrightarrow{h} \bigcirc \xrightarrow{g} \bigcirc \xrightarrow{f} \bigcirc$$

(a) reverse mode calculate $\frac{\partial l}{\partial x}$ as follows,

at first forward and calc $l(x)$

Then backward from left to right

$$\frac{\partial l}{\partial x} = \underrightarrow{\frac{\partial l}{\partial f} \frac{\partial f}{\partial g} \frac{\partial g}{\partial h} \frac{\partial h}{\partial x}}$$

(b) forward mode evaluate each node and Jaccobian as follows

$$\frac{\partial l}{\partial x} = \underrightarrow{\frac{\partial h}{\partial x} \frac{\partial g}{\partial h} \frac{\partial f}{\partial g} \frac{\partial l}{\partial f}}$$ from left to right

(c) if $m \gg n$ which mode?

$x \in \mathbb{R}^n \quad l : \mathbb{R}^m \to \mathbb{R}^n$

forward mode is preferable

because its computational complexity

is $O(n|E| + n|V|)$

where $|E|$: num of edge.
$|V|$: num of vertex or cell.

Backward mode $O(m|E| + m|V|)$

20A_Q2

$\boxed{2}$ Exploding Gradients

(a) $$f^k(x) = \underbrace{(f \circ f \circ \cdots f)}_{k\text{ times}}(x) = W^k x$$

$$W^k x = \underbrace{(Q \Lambda Q^{-1}) \cdots (Q \Lambda Q^{-1})}_{k\text{ times}} \; x$$

$$= Q \Lambda^k Q^{-1} x \quad \in \mathbb{R}^n$$

$$= Q \begin{bmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{bmatrix} Q^{-1} x \in \mathbb{R}^n$$

$k \to \infty$

if $|\lambda_i| \; (1 \leq i \leq n) < 1$

$\rightarrow$ vanishing

$|\lambda_i| > 1$

$\rightarrow$ exploding

$\lambda_i = 1$ | $\lambda_i = -1$

$\rightarrow$ converge | flip. not converge.

$\boxed{2}$ (b)

$$\frac{\partial f^k}{\partial x} = \frac{\partial f_k}{\partial f_{k-1}} \frac{\partial f_{k-1}}{\partial f_{k-2}} \cdots \frac{\partial f_1}{\partial x}$$

$$x \xrightarrow{f_1} \circ \xrightarrow{f_2} \circ \xrightarrow{f_3} \circ \cdots \circ \xrightarrow{f_k} \circ$$

---

e.g. $f(x) = 3x+4$

$$\frac{\partial f^3}{\partial x} = \frac{\partial f(f(f(x)))}{\partial x}$$

$$\begin{aligned} f^3(x) &= f(f(f(x))) \\ &= f(f(3x+4)) \\ &= f(9x+12+4) = f(9x+16) \\ &= 27x+16 \end{aligned}$$

$$\frac{\partial f^3}{\partial x} = 27$$

---

e.g. $f(x) = 7x$

$$
\begin{aligned}
f^3(x) &= f(f(f(x))) \\
&= f(f(7x)) \\
&= f(49x) \\
&= 7^3 x.
\end{aligned}
\qquad \frac{\partial f^3}{\partial x} = 7^3
$$

$$
0 \xrightarrow{f} 1 \xrightarrow{f} 2 \xrightarrow{f} \cdots \xrightarrow{f} k
$$

$$
\frac{\partial f^k}{\partial x} = \frac{\partial k}{\partial 0} = \frac{\partial k}{\partial (k-1)} \frac{\partial (k-1)}{\partial (k-2)} \cdots \frac{\partial 1}{\partial 0}
$$

$$
= \prod_{i=1}^{k} w_i = W^k
$$

$W = w_i \quad i \ (1 \le i \le k)$

 (f is not linear)

$\boxed{2}$ ii

A. $\bar{x} + \delta u$ $\qquad u^T J = \lambda u^T$

$$J = \frac{\partial f}{\partial x}$$

$$\frac{\partial f}{\partial x} = J$$

$$(\bar{x} + \delta u)^T \frac{\partial f^k}{\partial x}(x) = (\bar{x} + \delta u)^T J^k$$

when we start from $\bar{x}^T$, $\quad \bar{x}^T \frac{\partial f^k}{\partial x} = \bar{x}^T J^k$

there is $\quad \delta u^T J^k = \delta \lambda u^T J^{k-1}$

$$= \delta \lambda^k u^T$$

B. Why $|\lambda_{max}|$ plays crucial role of RNN

from here, since small perturbation affects so much

if $\lambda_{max} > 1$

Accumulated gradient is affected by λ_{max} in above way.

[3] Second-Order

$$\hat{f}_{x_0}(x) \equiv f(x_0) + (x-x_0)^T g + \frac{1}{2}(x-x_0)^T H (x-x_0)$$

where $g \equiv \frac{\partial f}{\partial x}(x_0)$

$H \equiv \frac{\partial^2 f}{\partial x}(x_0)$

(a) $\hat{f}_{x_0}(x_0 - \varepsilon g) = f(x_0) - \varepsilon g^T g + \frac{\varepsilon^2}{2} g^T H g$

(b) $\hat{f}_{x_0}(x_0 - \varepsilon g) = \hat{f}_{x_0}(\varepsilon) = \frac{1}{2}\varepsilon^2 g^T H g - \varepsilon g^T g$

$+ f(x_0)$

$$\varepsilon^* = \underset{\varepsilon}{\operatorname{argmin}} \ \hat{f}_{x_0}(\varepsilon)$$

and $\uparrow$ is convex.

So $\varepsilon^* = \{\varepsilon \mid \frac{\partial}{\partial \varepsilon}\hat{f}_{x_0}(\varepsilon) = 0\}$

$\Leftrightarrow \quad \varepsilon \, g^T H g + g^T g = 0$

$\Leftrightarrow \quad \varepsilon = -(g^T H g)^{-1} g^T g$

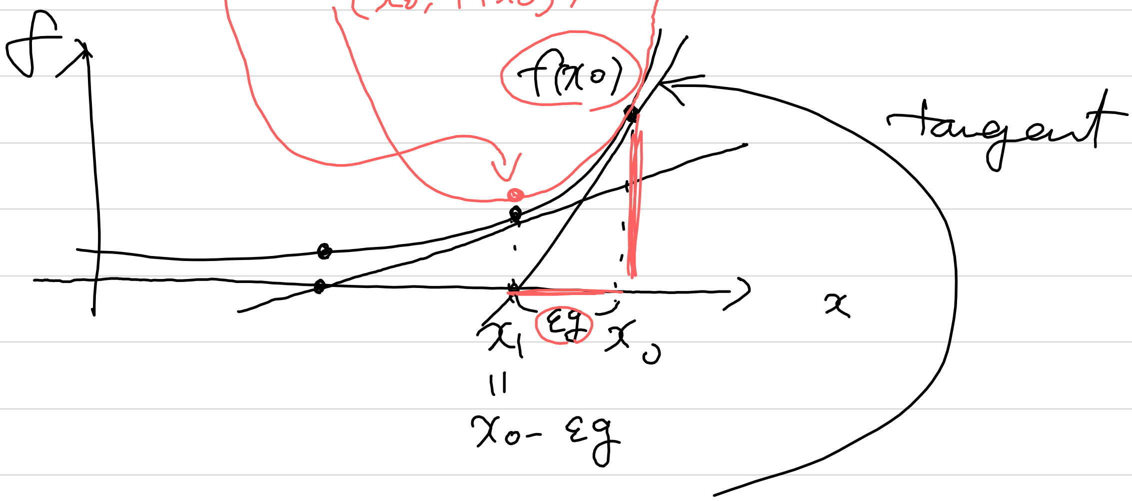
$= -H^{-1}$

20A_Q2 $\qquad\qquad\qquad\qquad f$

$(x_1, 0)$

$\boxed{3}$ b $\quad \hat{f}_{x_0}(x_0 - \varepsilon g) = f(x_0) + \varepsilon g^T g + \frac{\varepsilon^2}{2} g^T H g \quad \cdots *$

$(x_0, f(x_0))$



$$y = 0 = f(x_0) + f'(x_0)(x_1 - x_0)$$

$$\Leftrightarrow \quad x_1 - x_0 = -\frac{f(x_0)}{f'(x_0)}$$

$$x_1 = x_0 + \frac{f(x_0)}{f'(x_0)}$$

g is slope of tangent of f at x_0

H is curvature of f at x_0

Answer

$*$ is convex approximation

where $x = x_0 - \varepsilon g$

20A_Q2

$\boxed{3}$ (c)

$$\hat{f}_{x_0}(x) \equiv f(x_0) + (x-x_0)^T g + \frac{1}{2}(x-x_0)^T H (x-x_0)$$

$$\hat{f}_{x_0}(x_0+\delta) = f(x_0) + \delta^T g + \frac{1}{2}\delta^T H \delta$$

$$\hat{f}_{x_0}(\delta) = \quad ''$$

$$\delta_{opt} = \underset{\delta}{\operatorname{argmin}} \hat{f}_{x_0}(\delta)$$

$$\frac{\partial \hat{f}_{x_0}(\delta)}{\partial \delta} = H\delta + g = 0$$

$$\Leftrightarrow \quad \underline{\delta = -H^{-1} g}$$

Newton update direction.

$$\hat{f}_{x_0}(x) = f(x_0) - \varepsilon \underset{\substack{\Downarrow \\ 0}}{\delta^T g} + \frac{1}{2}\delta^T H \delta$$

$$= f(x_0) + \frac{\varepsilon^2}{2}\delta^T H \delta$$

if $H \preceq 0$ $\quad \varepsilon$ should be 0