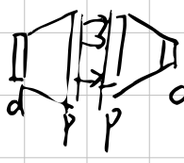
Automatic differention:

2-layer-NN:

$$a \to \underbrace{\sqcup a \mapsto \sigma(\sqcup a) \to V\sigma(\sqcup a)}_{g(a,\theta)}$$

$\theta = (U, V)$

$\sigma(z) = (\sigma(z_i))_{i=1}^n$ $\quad \sigma(s) = \frac{e^s}{1+e^s}$ $\quad \sigma(s) = \max(s, 0)$



$$U = \begin{matrix} u_1 \\ u_2 \\ \vdots \end{matrix} \qquad V = v_1 v_2 \cdots$$

$$g(a,\theta) = V\sigma(Ua) = \sum_{i=1}^{p} \underbrace{v_i}_{\text{"a neuron"}} \cdot \sigma(\langle u_i, a \rangle)$$

$$\min_{\theta=U,V} \sum_{i=1}^{n} (g(a_i,\theta) - y_i)^2$$

$$\min_{U,V} \sum_{i=1}^{n} (\underbrace{V}_{\text{easy}} \underbrace{\sigma(\boxed{U} a_i)}_{b_i} - y_i)^2 \quad \text{\\} \; f(U,V)$$

Exercise: Compute $\nabla_U f$ $\nabla_V f$

Q°: $f: x \in \mathbb{R}^d \to f(x) \in \mathbb{R}$

Hyp: apply $f(x)$ cost k

$\nabla f: x \in \mathbb{R}$

$\nabla f: x \in \mathbb{R}^d \to \nabla f(x) \in \mathbb{R}^d$

Q°: How many operations to compute $\nabla f(x)$

$\nabla f(x) \approx \left(\frac{f(x+\varepsilon\delta_1)-f(x)}{\varepsilon}, \frac{f(x+\varepsilon\delta_2)-f(x)}{\varepsilon}, \ldots, \right) \quad \delta_1 \ldots \delta_d$

$\hookrightarrow (d+1)k$ operations

Thm: cost of $\nabla f(x) \sim 3k$

+ algo $\quad Algo(f) \to Algo(\nabla f)$

$\qquad\qquad k \qquad\qquad 3k$

$f(x) = f_k \circ f_{k-1} \circ \cdots \circ f_0(x)$

$\nabla f(x) = \nabla f_k(x_k) \times \nabla f_{k-1}(x_{k-1}) \times \cdots \times \nabla f_0(x_0 = x)$

$x_0 = x$

$x_{k+1} = f_k(x_k) \qquad x_k \in \mathbb{R}^{d_k}$

$\nabla f(x) = (A_k \times (A_{k-1} \times (A_{k-2}))) \times A_0$

Forward method: $A_k(A_3(A_2(\underbrace{A_1(A_0 \cdots}_{n^3}$

Backward $\quad (\underbrace{(A_k A_{k-1})}_{n^2} A_{k-2}) \cdots A_0$

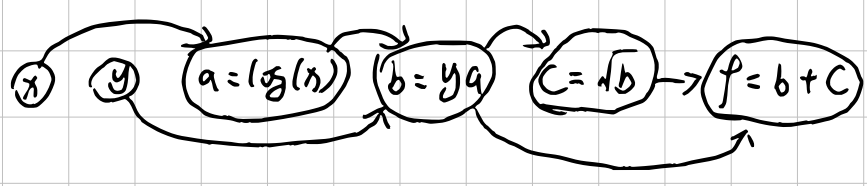
Ex. $d_k = 1 \quad d_{k-1} = d_{k-2} = \cdots = d$ $\quad$ forward $n^2 + (k-1)n^3$ $\quad$ backward kn^2

Computational graph

DAG : directed acyclic graph

Algo : $x_k = f_k(x_{k-1}, x_{k-2}, \dots x_1)$ exam

$f(x,y) = y\log(x) + \sqrt{y\log(x)}$

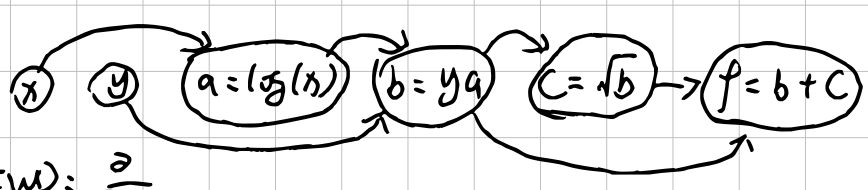


Forward = classical chain rule

$x_k = f_k((x_i)_{i \in Father})$

Thm : $\frac{\partial x_k}{\partial x_1} = \sum_{l \in Father(k)} \left[\frac{\partial x_k}{\partial x_l}\right] \times \frac{\partial x_l}{\partial x_1} \equiv \frac{\partial f_k(\dots)}{\partial x_l}$

Initia : $\frac{\partial x_1}{\partial x_1} = Id_{1 \times d_1}$ $\frac{\partial x_2}{\partial x_1} = 0$ ....



FWD : $\frac{\partial}{\partial x}$

$\frac{\partial x}{\partial x} = 1 \quad \frac{\partial y}{\partial x} = 0$

$\frac{\partial a}{\partial x} = \left[\frac{\partial a}{\partial x}\right] \times \frac{\partial x}{\partial x} = \frac{1}{x} \times \frac{\partial x}{\partial x}$

$\frac{\partial b}{\partial x} = \frac{\partial b}{\partial y}\frac{\partial y}{\partial x} + \frac{\partial b}{\partial a}\frac{\partial a}{\partial x} = a\frac{\partial y}{\partial x} + y\frac{\partial a}{\partial x}$

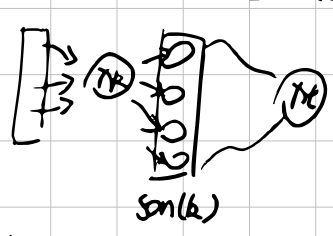
$\frac{\partial c}{\partial x} = \frac{\partial c}{\partial b}\frac{\partial b}{\partial x} = \frac{1}{2\sqrt{b}} \times \frac{\partial b}{\partial x}$

$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial b}\frac{\partial b}{\partial x} + \frac{\partial f}{\partial c}\frac{\partial c}{\partial x} = 1\frac{\partial b}{\partial x} + 1\frac{\partial c}{\partial x}$

$\frac{\partial}{\partial y}$ ....

BWD $\frac{\partial x_k}{\partial x_k}$ ← current, ← last

FWD $\frac{\partial x_k}{\partial x_1}$ ← current, ← 1st



Thm : $\frac{\partial x_t}{\partial x_k} \overset{*}{=} \sum_{l \in son(k)} \frac{\partial x_t}{\partial x_l}\left[\frac{\partial x_l}{\partial x_k}\right]$

$\frac{\partial x_t}{\partial x_t} = Id$

In practice : $x_t \in \mathbb{R}$ (last) $f = x_t$

$\frac{\partial f}{\partial x_k} = \square\!\square = [\nabla_{x_k} f]^T$

$\nabla_{x_k} f \overset{**}{=} \sum_{l \in son(k)} \left[\frac{\partial f_l}{\partial x_k}\right]^T (\nabla_{x_l} f)$

$*$ $\square\!\square$ } Jacobian

$**$ $\square^T \square$ } vector product

Summary :

Algo ($x_1 - x_b$)

For $k = s+1 \dots t$

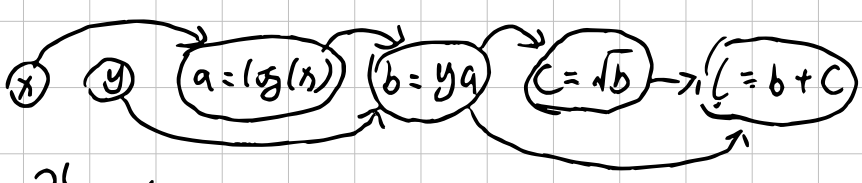
  $x_k = f_k(x_1 - x_{k-1})$

  [store x_k]

Backprop ($x_1 - x_s$) $\nabla_{x_t} l = 1$

For $k = t-1, t-2, \dots 1$

$\nabla_{x_k} l = \sum_{l \in son(k)} \left[\frac{\partial f_l}{\partial x_k}\right]^T (\nabla_{x_l} l)$



$\frac{\partial l}{\partial l} = 1$

$\frac{\partial l}{\partial c} = \frac{\partial l}{\partial l} \times \left[\frac{\partial l}{\partial c}\right] = \frac{\partial l}{\partial l} \times 1$

$\frac{\partial l}{\partial b} = \frac{\partial l}{\partial c}\left[\frac{\partial c}{\partial b}\right] + \frac{\partial l}{\partial l}\left[\frac{\partial l}{\partial b}\right] = \frac{\partial l}{\partial c} \times \frac{1}{2\sqrt{b}} + \frac{\partial l}{\partial l} 1$

$\frac{\partial l}{\partial a} = \frac{\partial l}{\partial b}\left[\frac{\partial b}{\partial a}\right] = \frac{\partial l}{\partial b} y$

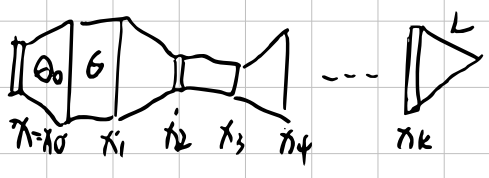
$\frac{\partial l}{\partial y} = \frac{\partial l}{\partial b}\left[\frac{\partial b}{\partial y}\right] = \frac{\partial l}{\partial b} a$

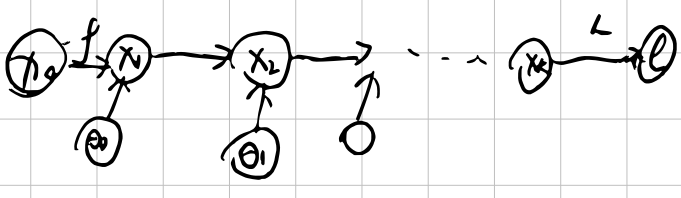
$\frac{\partial l}{\partial x} = \frac{\partial l}{\partial a}\left[\frac{\partial a}{\partial x}\right] = \frac{\partial l}{\partial a}\frac{1}{x}$

MLP: x_0 input (ex. image)

$k = 0 \dots k-1 \quad x_{k+1} = \sigma(\theta_k x_k)$

$f(x, \theta) = L(x_k)$





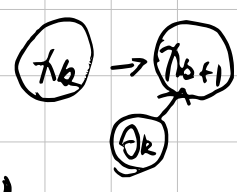
$f_1 = f_2 = \dots = f_k$ (layers)

$x_{k+1} = f_{k+1}(x_k, \theta_k) = \sigma(\theta_k \times x_k)$

Back-prop $\nabla_{x_k} f$

$\nabla_{\theta_k} f$

INIT: $\nabla_{x_k} \ell = \nabla L(x_k)$



Recursion: $\nabla_{x_k} \ell = \left[\frac{\partial f_k}{\partial x_k}\right]^T (\nabla_{x_{k+1}} \ell)$

$\nabla_{\theta_k} \ell = \left[\frac{\partial f_k}{\partial \theta_k}\right]^T (\nabla_{x_{k+1}} f)$

$f: (x, \theta) \longmapsto \sigma(\theta \times x) \in \mathbb{R}^{d' \times N}$ (setting)

$x: \mathbb{R}^{d \times N}$, $\theta: \mathbb{R}^{d' \times d}$, $\theta \times x: d' \times N$

What is implementation $\frac{\partial f}{\partial x}(x, \theta)^T[z] = ??$

$\frac{\partial f}{\partial \theta}(x, \theta)^T[z] = ??$

$\frac{\partial f}{\partial x}(x, \theta): \mathbb{R}^{d \times N} \to \mathbb{R}^{d' \times N}$ linear

$\frac{\partial f}{\partial x}(x, \theta)^T: z \in \mathbb{R}^{d' \times N} \to \mathbb{R}^{d \times N}$

$\frac{\partial f}{\partial \theta}(x, \theta): \mathbb{R}^{d' \times d} \to \mathbb{R}^{d' \times N}$

$\frac{\partial f}{\partial \theta}(x, \theta)^T: z \in \mathbb{R}^{d' \times N} \to \mathbb{R}^{d' \times d}$

$f(x, \theta) = \sigma(\theta x)$

$f(x + \varepsilon T, \theta) = \sigma(\theta x + \varepsilon \theta T)$

$\sigma(w + \varepsilon \gamma) = \begin{pmatrix} \sigma(w_1 + \varepsilon \gamma_1) \\ \vdots \\ \sigma(w_d + \varepsilon \gamma_d) \end{pmatrix}$

$= \begin{pmatrix} \sigma(w_1) + \varepsilon \sigma'(w_1)\gamma_1 + \varepsilon^2 \cdot (\dots) \\ \vdots \\ \sigma(w_d) + \varepsilon \sigma'(w_d)\gamma_d + \varepsilon^2 \cdot (\dots) \end{pmatrix} + o(\varepsilon)$

$= \sigma(w) + \varepsilon \underbrace{\sigma'(w)}_{\text{diag}(\sigma'(w))} \gamma + o(\varepsilon)$

$\gamma \times (\sigma'(w) \cdot \gamma)$

Recap: $\partial \sigma(w) = \text{diag}(\sigma'(w))$

Ex. $\text{Relu}(s) =$

$f(x, \theta) = \sigma(\theta x)$

$f(x + \varepsilon T, \theta) = \sigma(\theta x + \varepsilon \theta T)$

$= \sigma(\theta x) + \varepsilon \, \text{diag}(\sigma'(\theta x)) \times (\theta T) + o(\varepsilon)$

↑ Jacob. of σ

$\frac{\partial f}{\partial x}(x, \theta)[T] = \text{diag}(\sigma'(\theta x)) \times \theta T$

Remember: $\langle A[T], z \rangle = \langle T, A^T[z] \rangle$

$\langle \frac{\partial f}{\partial x}(x, \theta)[T], z \rangle = \langle \text{diag}(\sigma'(\theta x)) \times \theta T, z \rangle$

$= \langle T, \underbrace{\theta^T \text{diag}(\sigma'(\theta x)) z}_{\text{adjoint}} \rangle$

Concl° 1: $\frac{\partial f}{\partial x}(x, \theta)^T[z] = \theta^T \text{diag}(\sigma'(\theta x)) z$

on θ: $f(x, \theta + \varepsilon Z) = \sigma((\theta + \varepsilon Z)x)$

$= \sigma(\theta x + \varepsilon Z x) + \sigma(\theta x) + \varepsilon \, \text{diag}(\sigma'(\theta x)) Z x + o(\varepsilon)$

$\frac{\partial f}{\partial \theta}(x, \theta)[z] = \text{diag}(\sigma'(\theta x)) \times Z x$

$\langle \frac{\partial f}{\partial \theta}(x, \theta)[z], T \rangle = \langle \text{diag}(\dots) Z x, T \rangle$

$= \langle z, \text{diag}(\dots)^T x^T \rangle$

Concl° #2: $\frac{\partial f}{\partial \theta}(x, \theta)^T[z] = \text{diag}(\sigma'(\theta x)) \, z \, x^T$

$z \in \mathbb{R}^{d' \times N}$, $x^T \in \mathbb{R}^{N \times d}$, product $d' \times d$

$\theta_k = (U_k, V_k)$

$x_{k+1} = x_k + U_k \sigma(V_k x_k)$

Residual