FoML

24 Backpropagation

Dr. Konda Reddy Mopuri
Dept. of AI, IIT Hyderabad
July-Nov 2025

$$w^{t+1} = w^t - \eta \frac{\partial L}{\partial w^t}$$

Recap

- Gradient of a scalar valued function $f(\mathbf{x})$: $\mathbf{x} \rightarrow \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_D}\right)$

Recap

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- Gradient of a scalar valued function $f(\mathbf{x})$: $\mathbf{x} \rightarrow \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_D}\right)$
- Gradient of a vector valued function $\mathbf{f}(\mathbf{x})$ is called Jacobian:

$f: \mathbb{R}^n \rightarrow \mathbb{R}^m$

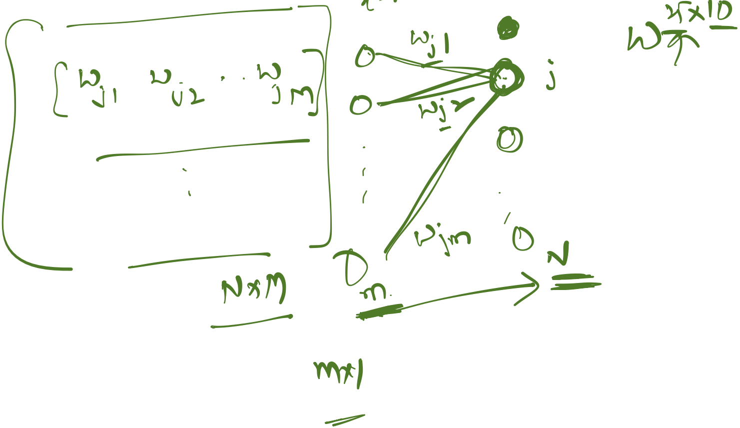
$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

$M \times N$

f_1, f_2, f_2, $\vdots$, f_m

1. w^l_{jk} is the weight connecting j^{th} neuron in l^{th} layer and k^{th} neuron in $(l-1)^{st}$ layer

$l-1$ 10 $\quad$ l 25

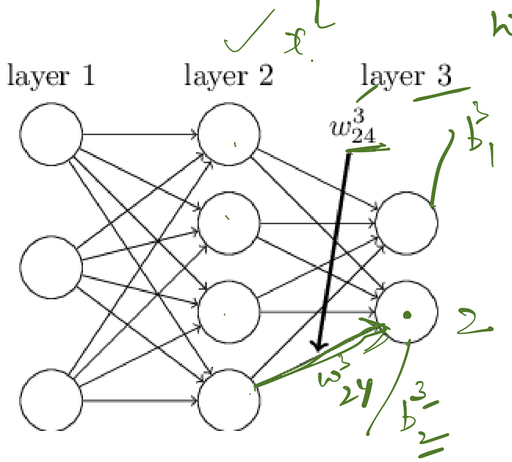


MLP: Some Notation

1. w^l_{jk} is the weight connecting j^{th} neuron in l^{th} layer and k^{th} neuron in $(l-1)^{st}$ layer

$W^3_{N \times M}$ $b_{N \times 1}$

$b^3 = [\ ]$

2. 

MLP: Some Notation

1. b_j^l is the bias of j^{th} neuron in l^{th} layer

MLP: Some Notation

1. b_j^l is the bias of j^{th} neuron in l^{th} layer
2. x_j^l is the activation (output) of j^{th} neuron in l^{th} layer

MLP: Some Notation

1. b^l_j is the bias of j^{th} neuron in l^{th} layer
2. x^l_j is the activation (output) of j^{th} neuron in l^{th} layer
3. $$x^l_j = \sigma(\sum_k w^l_{jk} x^{l-1}_k + b^l_j)$$

$$\sigma\left\{ \sum_{*} w^{l}_{j*} x^{l-1}_{*} + b^{l}_{j} \right\}$$

MLP: Some Notation

1. b_j^l is the bias of j^{th} neuron in l^{th} layer
2. x_j^l is the activation (output) of j^{th} neuron in l^{th} layer
3. x^l

$$x_j^l = \sigma(\sum_k w_{jk}^l x_k^{l-1} + b_j^l)$$

$\sigma(W^l x^{l-1} + b^l)$

$M \times N \quad N \times 1$

$M \times 1$

4. Vector of activations (or, biases) at a layer l is denoted by a bold-faced $\mathbf{x}^l$ (or $\mathbf{b}^l$) and W^l is the matrix of weights into layer l

$\sigma(M \times 1)$

MLP: Some Notation

1. s_j^l is the weighted input to j^{th} neuron in l^{th} layer

$$x^l = \sigma(\underline{s}^l)$$

MLP: Some Notation

1. s_j^l is the weighted input to j^{th} neuron in l^{th} layer
2. $s_j^l = \sum_k w_{jk}^l x_k^{l-1} + b_j^l$

MLP: Some Notation

1. s_j^l is the weighted input to j^{th} neuron in l^{th} layer
2. $s_j^l = \sum_k w_{jk}^l x_k^{l-1} + b_j^l$
3. $\mathbf{s}^l = W^l \mathbf{x}^{l-1} + \mathbf{b}^l$

$\underline{x}^l = \sigma(\underline{s}^l)$

MLP: Some Notation

1. s_j^l is the weighted input to j^{th} neuron in l^{th} layer
2. $s_j^l = \sum_k w_{jk}^l x_k^{l-1} + b_j^l$
3. $\mathbf{s}^l = W^l \mathbf{x}^{l-1} + \mathbf{b}^l$
4. σ is the activation function that applies element-wise

Gradient descent on MLP

- Loss is $\mathcal{L}(W, \mathbf{b}) = \sum_n l(f(x_n; W, \mathbf{b}), y_n) = \sum_n l(\mathbf{x}^L, y_n)$ (L is the number of layers in the MLP)

Gradient descent on MLP

- Loss is $\mathcal{L}(W, \mathbf{b}) = \sum_n l(f(x_n; W, \mathbf{b}), y_n) = \underline{\sum_n l(\mathbf{x}^L, y_n)}$ (L is the number of layers in the MLP)
- For applying Gradient descent, we need gradient of individual sample loss with respect to all the model parameters

$$l_n = l(f(x_n; W, \mathbf{b}), y_n)$$

$$\left\{ \frac{\partial l_n}{\partial W_{jk}^{(l)}} \text{ and } \frac{\partial l_n}{\partial \mathbf{b}_j^{(l)}} \right\} \text{ for all layers } l$$

$w^{t+1} = w^t - \eta$

Forward pass operation

$$x^{(0)} = x \xrightarrow{W^{(1)},\mathbf{b}^{(1)}} s^{(1)} \xrightarrow{\sigma} x^{(1)} \xrightarrow{W^{(2)},\mathbf{b}^{(2)}} s^{(2)} \ldots x^{(L-1)} \xrightarrow{W^{(L)},\mathbf{b}^{(L)}} s^{(L)} \xrightarrow{\sigma} x^{(L)} = f(x; W, \mathbf{b})$$

y

Formally, $x^{(0)} = x$, $f(x; W, \mathbf{b}) = x^{(L)}$

$$\forall l = 1, \ldots, L \quad \begin{cases} s^{(l)} & = W^{(l)} x^{(l-1)} + \mathbf{b}^{(l)} \\ x^{(l)} & = \sigma(s^{(l)}) \end{cases}$$

Chain rule of differential calculus

- Core concept of backpropagation

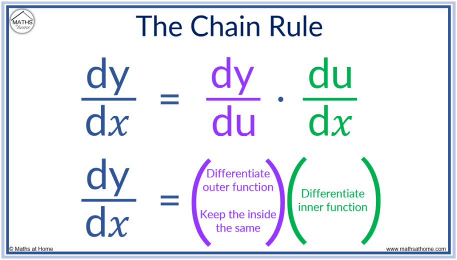
Chain rule of differential calculus

- Core concept of backpropagation
-
$$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$

Chain rule of differential calculus

- Core concept of backpropagation
- $$(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$$
- $$\frac{\partial}{\partial x} f(g(x)) = \left.\frac{\partial f(a)}{\partial a}\right|_{a=g(x)} \cdot \frac{\partial g(x)}{\partial x}$$

Chain rule of differential calculus



Chain rule of differential calculus

- For any nested function $y = f(g(x))$

Chain rule of differential calculus

- For any nested function $y = f(g(x))$
- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$

Chain rule of differential calculus

- For any nested function $y = f(g(x))$
- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$
- $\Delta y = \frac{dy}{dx} \Delta x$

Chain rule of differential calculus

- For any nested function $y = f(g(x))$
- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$
- $\Delta y = \frac{dy}{dx} \Delta x$
- $z = g(x) \rightarrow \Delta z = \frac{dg(x)}{dx} \Delta x$

Chain rule of differential calculus

- For any nested function $y = f(g(x))$
- $\frac{dy}{dx} = \frac{\partial f}{\partial g(x)} \frac{dg(x)}{dx}$
- $\Delta y = \frac{dy}{dx} \Delta x$
- $z = g(x) \rightarrow \Delta z = \frac{dg(x)}{dx} \Delta x$
- $y = f(z) \rightarrow \Delta y = \frac{df}{dz} \Delta z = \frac{df}{dz} \frac{dg(x)}{dx} \Delta x = \frac{df}{dg(x)} \boxed{\frac{dg(x)}{dx} \Delta x}$

Δg

Distributed Chain rule of differential calculus

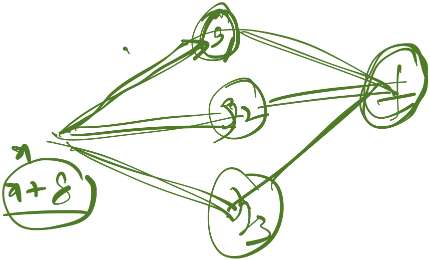
1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$

x

$f(x, y)$

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$
2. $\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$



Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$
2. $\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$
3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$
2. $\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$
3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$
4. $\underline{\Delta y} = \underline{\frac{\partial f}{\partial z_1} \Delta z_1} + \underline{\frac{\partial f}{\partial z_2} \Delta z_2} + \ldots + \underline{\frac{\partial f}{\partial z_M} \Delta z_M}$

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$
2. $\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$
3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$
4. $\Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \ldots + \frac{\partial f}{\partial z_M} \Delta z_M$
5. $\Delta y = \frac{\partial f}{\partial z_1} \frac{dz_1}{dx} \Delta x + \frac{\partial f}{\partial z_2} \frac{dz_2}{dx} \Delta x + \ldots + \frac{\partial f}{\partial z_M} \frac{dz_M}{dx} \Delta x$

Distributed Chain rule of differential calculus

1. $y = f(g_1(x), g_2(x), \ldots, g_M(x))$

2. $\frac{dy}{dx} = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx}$

3. Let $g_i(x) = z_i \rightarrow y = f(z_1, z_2, \ldots, z_M)$

4. $\Delta y = \frac{\partial f}{\partial z_1} \Delta z_1 + \frac{\partial f}{\partial z_2} \Delta z_2 + \ldots + \frac{\partial f}{\partial z_M} \Delta z_M$

5. $\Delta y = \frac{\partial f}{\partial z_1} \frac{dz_1}{dx} \Delta x + \frac{\partial f}{\partial z_2} \frac{dz_2}{dx} \Delta x + \ldots + \frac{\partial f}{\partial z_M} \frac{dz_M}{dx} \Delta x$

6. $\Delta y = \frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} \Delta x + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} \Delta x + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \Delta x$

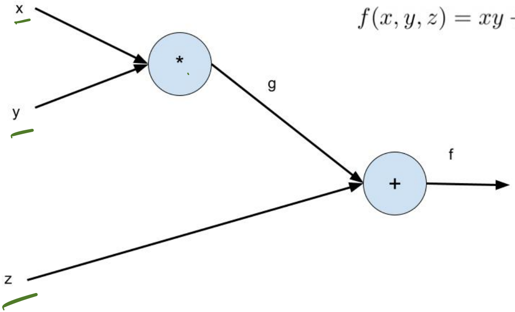
7. $\Delta y = \left(\frac{\partial f}{\partial g_1(x)} \frac{dg_1(x)}{dx} + \frac{\partial f}{\partial g_2(x)} \frac{dg_2(x)}{dx} + \ldots + \frac{\partial f}{\partial g_M(x)} \frac{dg_M(x)}{dx} \right) \Delta x$

Chain rule of differential calculus

1. $f(x) = e^{sin(x^2)}$, let's find $\frac{\partial f}{\partial x}$

$$f'(x) = e^{\sin x^2} \cdot \cos x^2 \cdot 2x$$

Chain rule of differential calculus



$$g = xy$$

$$f = g + z$$

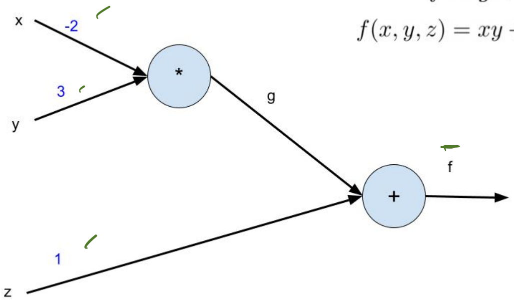
$$f(x, y, z) = xy + z$$

Chain rule of differential calculus

$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$

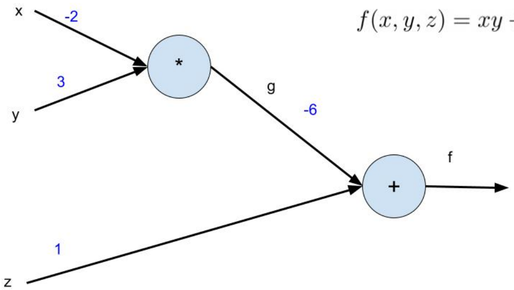


Chain rule of differential calculus

$$g = xy$$

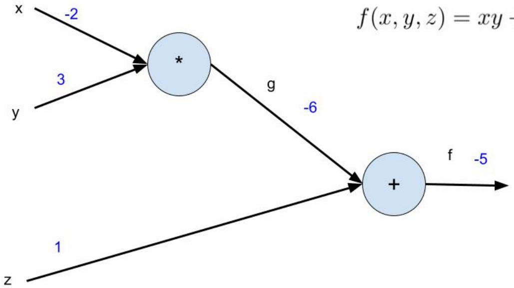
$$f = g + z$$

$$f(x, y, z) = xy + z$$



Chain rule of differential calculus

$$g = xy$$
$$f = g + z$$
$$f(x, y, z) = xy + z$$

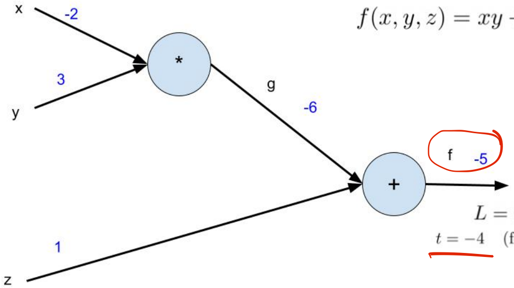


Chain rule of differential calculus

$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$



$$L = (f - t)^2$$

$t = -4$ (for this input)

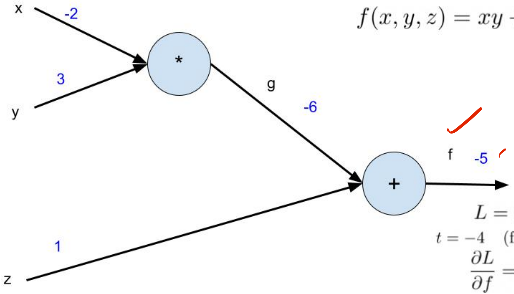
Chain rule of differential calculus

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$



$$L = (f - t)^2$$

$t = -4$ (for this input)

$$\frac{\partial L}{\partial f} = 2(f - t)$$

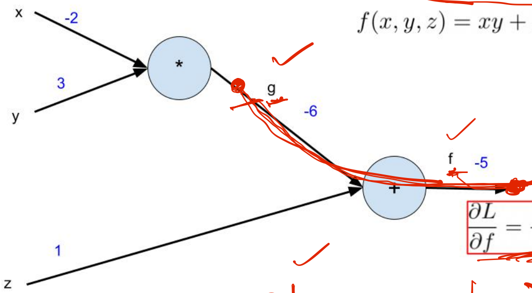
Chain rule of differential calculus

$$g = xy$$

$$\boxed{f = g + z}$$

$$f(x, y, z) = xy + z$$

$$\frac{\partial f}{\partial g} = 1$$



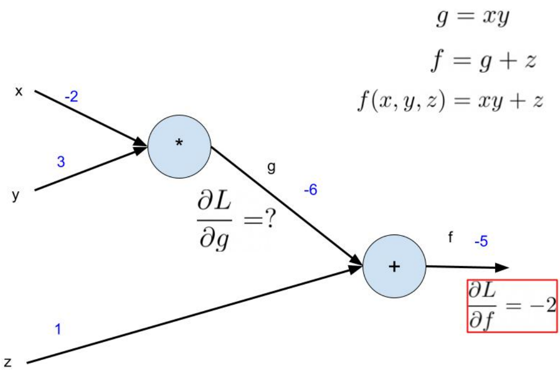
$$L = (f - t)^2$$

$$\boxed{\frac{\partial L}{\partial f} = -2}$$

$$\frac{\partial L}{\partial (g + \Delta)} = \frac{\partial f}{\partial g} \cdot \frac{\partial L}{\partial f}$$

$$1 \cdot -2$$

Chain rule of differential calculus

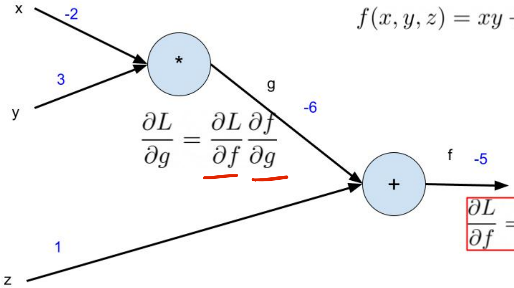


Chain rule of differential calculus

$$g = xy$$

$$f = g + z$$

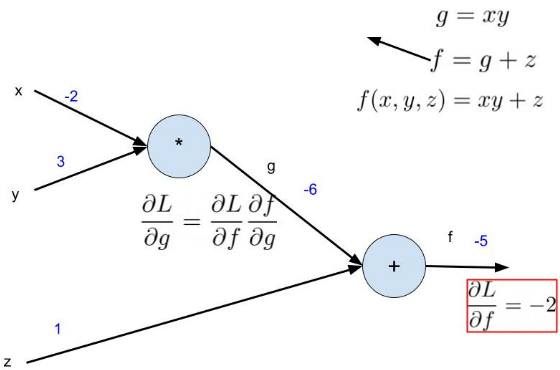
$$f(x, y, z) = xy + z$$



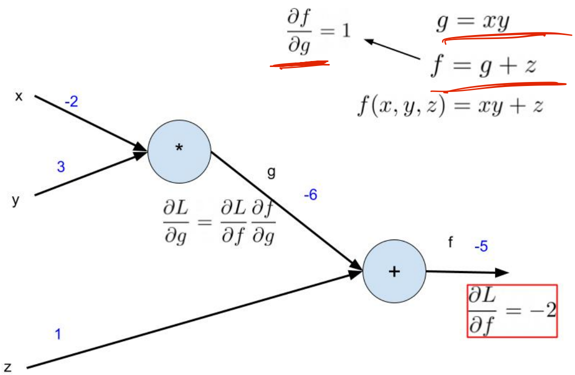
$$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial f} \frac{\partial f}{\partial g}$$

$$\boxed{\frac{\partial L}{\partial f} = -2}$$

Chain rule of differential calculus



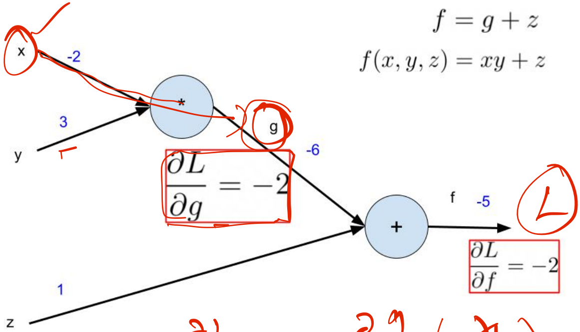
Chain rule of differential calculus



Chain rule of differential calculus

$$g = xy$$
$$f = g + z$$
$$f(x, y, z) = xy + z$$

$$\frac{\partial g}{\partial x} = y$$

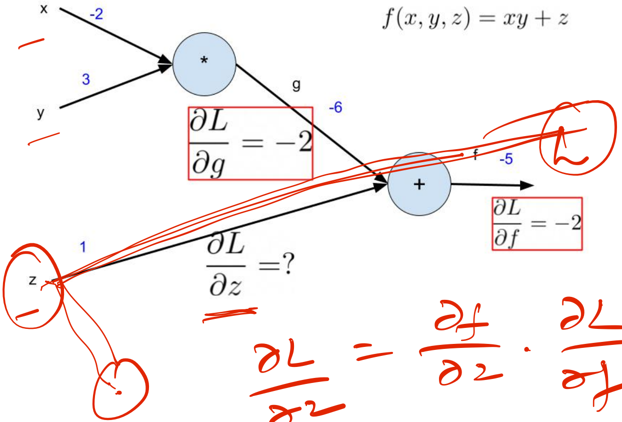


$$\frac{\partial L}{\partial x} = \frac{\partial g}{\partial x}\left(\frac{\partial L}{\partial g}\right)$$
$$y \cdot -2 = -6$$

$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$



$$\frac{\partial L}{\partial z} = \frac{\partial f}{\partial z} \cdot \frac{\partial L}{\partial f}$$

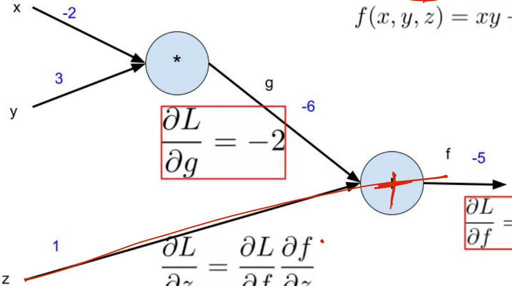
Chain rule of differential calculus

$$g = xy$$

$$\boxed{f = g + z}$$

$$\frac{\partial f}{\partial z} = 1$$

$$f(x, y, z) = xy + z$$



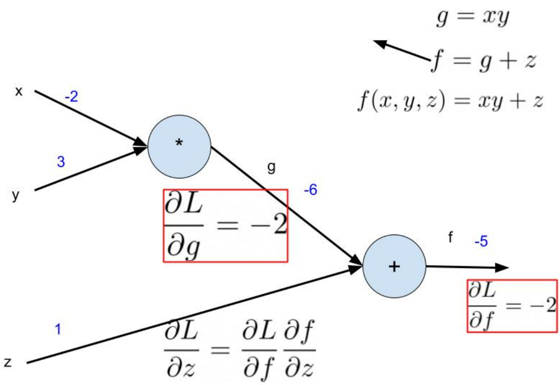
$$\boxed{\frac{\partial L}{\partial g} = -2}$$

$$\boxed{\frac{\partial L}{\partial f} = -2}$$

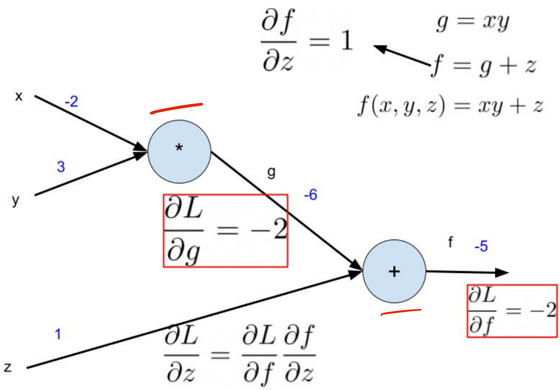
$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial z}$$

$$-2 \cdot 1$$

Chain rule of differential calculus



Chain rule of differential calculus

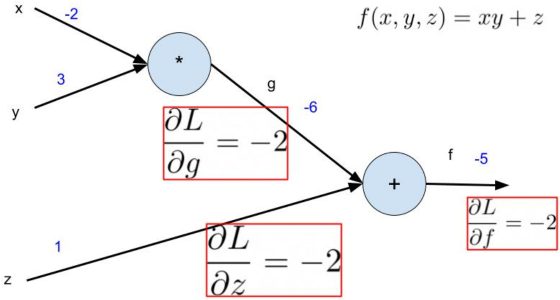


Chain rule of differential calculus

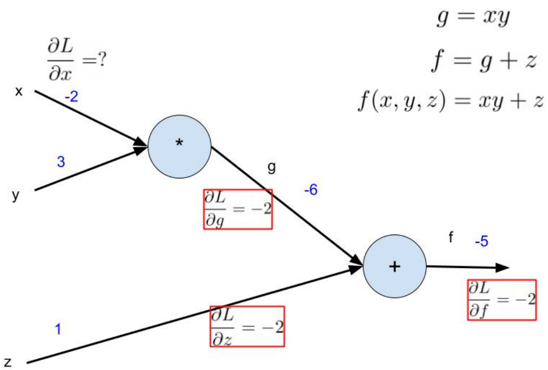
$$g = xy$$

$$f = g + z$$

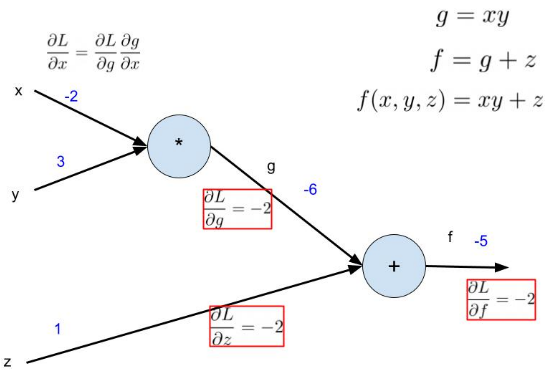
$$f(x, y, z) = xy + z$$



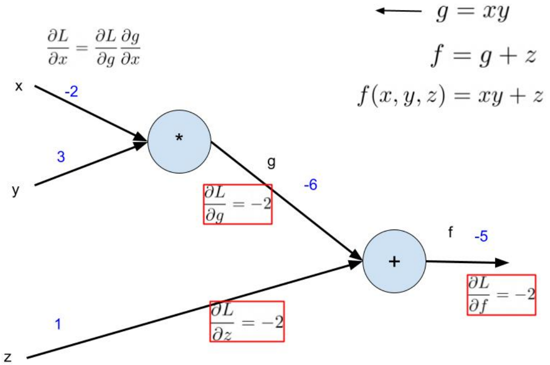
Chain rule of differential calculus



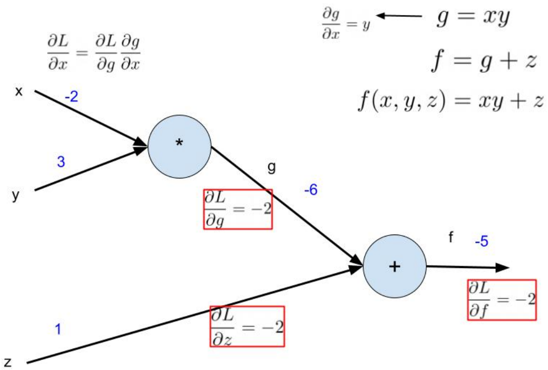
Chain rule of differential calculus



Chain rule of differential calculus



Chain rule of differential calculus

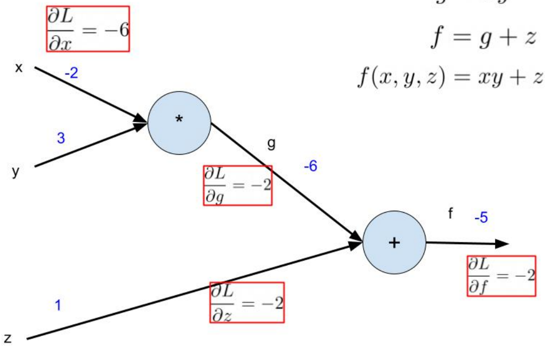


Chain rule of differential calculus

$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$

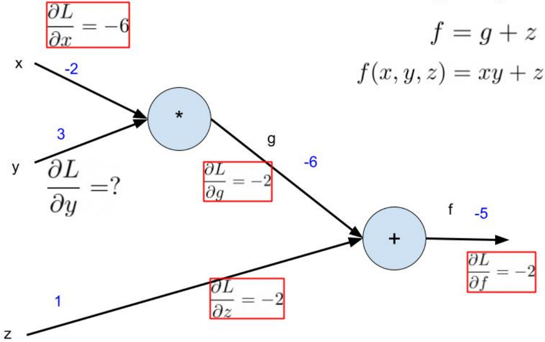


Chain rule of differential calculus

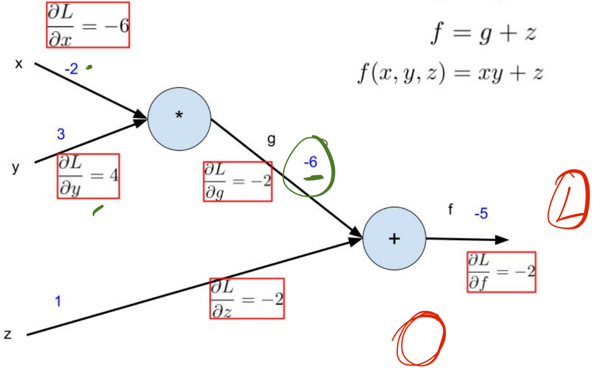
$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$



Chain rule of differential calculus

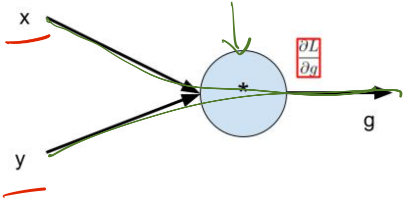


$$g = xy$$

$$f = g + z$$

$$f(x, y, z) = xy + z$$

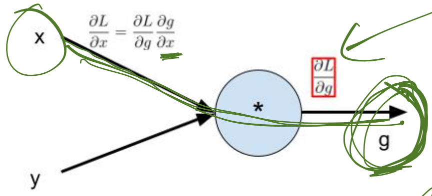
Gradient Flow



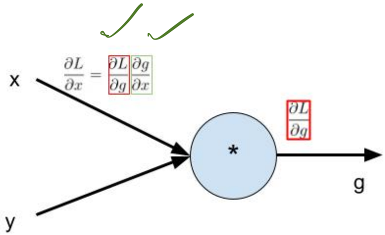
$$\frac{\partial L}{\partial x} = \frac{\partial g}{\partial x} \cdot \frac{\partial L}{\partial g}$$

$$\frac{\partial L}{\partial y} = \frac{\partial g}{\partial y} \frac{\partial L}{\partial g}$$

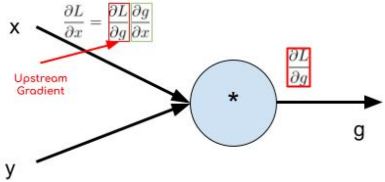
Gradient Flow



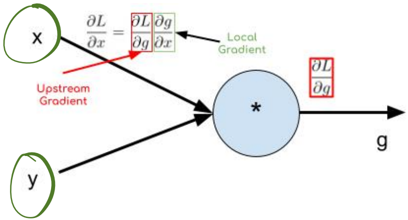
Gradient Flow



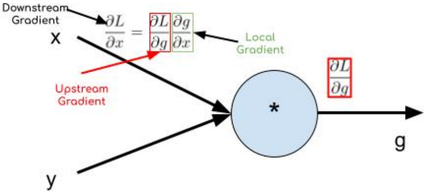
Gradient Flow



Gradient Flow

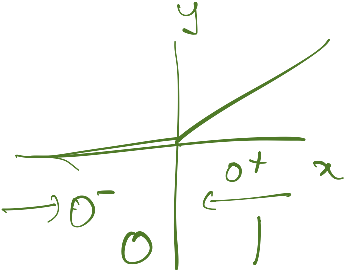


Gradient Flow



$y = ReLU(x)$

$\frac{\partial y}{\partial x} =$



Chain rule of differential calculus for an MLP

$f_N(f_{N-1}(f_{N-2}(\cdots f_1(x))))$

- $$J_{f_N \circ f_{N-1} \circ \ldots f_1(x)} = J_{f_N(f_{N-1}(\ldots f_1(x)))} \cdot J_{f_{N-1}(f_{N-2}(\ldots f_1(x)))} \cdot \cdots \cdot J_{f_2(f_1(x))} \cdot J_{f_1(x)}$$

  $J_{f(x)}$ is Jacobian of f computed at x.

Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$

Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $x_i^{(l)} = \sigma(s_i^{(l)})$

Consider a specific Layer

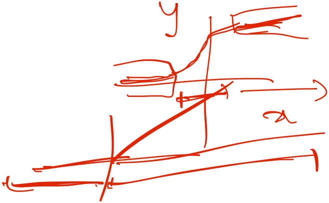
- $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$ $\mathcal{L}$
- $x_i^{(l)} = \sigma(s_i^{(l)})$
- Since $s^{(l)}$ influences loss $\mathcal{L}$ through only $x^{(l)}$,

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \boxed{\frac{\partial x_i^{(l)}}{\partial s_i^{(l)}}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$$

$x^l = \sigma(W^l x^{l-1} + b^l)$

$\frac{\partial L}{\partial x^{l}}$

$w^{t+1} = w^t - \eta$



Consider a specific Layer

- $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$ $= L$
- $x_i^{(l)} = \sigma(s_i^{(l)})$
- Since $s^{(l)}$ influences loss $\mathcal{L}$ through only $x^{(l)}$,

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$$

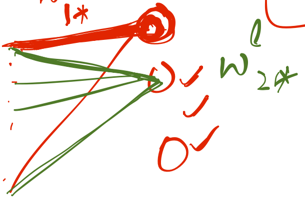
- $\boxed{s_i^{(l)} = \Sigma_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}}$

W^l

a^L

$\#l \quad X \#l-1$

10



a_i^l / s_i^l

W^l_{1*} W^l_{2*}

$a^{l-1} \to W^l \to a^l$, b^l

$\frac{\partial L}{\partial w^l}$ $\qquad$ $\frac{\partial L}{\partial b^l}$

Consider a specific Layer

$\frac{\partial L}{\partial w}$

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $x_i^{(l)} = \sigma(s_i^{(l)})$
- Since $s^{(l)}$ influences loss $\mathcal{L}$ through only $x^{(l)}$,

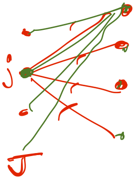
$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$$

$\frac{\partial L}{\partial s} \frac{\partial s}{\partial w}$

- $$s_i^{(l)} = \Sigma_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$$

a^{l-1}

- Since $x^{(l-1)}$ influences the loss $\mathcal{L}$ only through $s^{(l)}$,



$$\frac{\partial \ell}{\partial x_j^{(l-1)}} = \Sigma_i \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial x_j^{(l-1)}} = \Sigma_i \frac{\partial \ell}{\partial s_i^{(l)}} W_{i,j}^{(l)}$$

$w_{i,j}$

IJ

$W^{IJ} =$

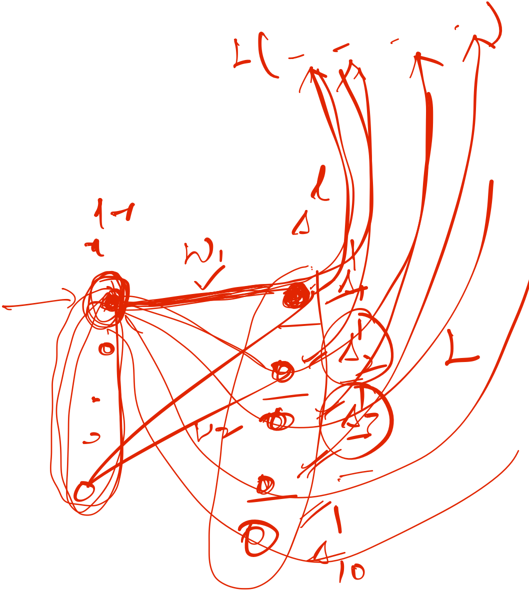
W

$i \in (1, I)$

We need gradients wrt parameters W and b

- $x^{(l-1)} \xrightarrow{\overline{W^{(l)}},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$

$$\frac{\partial L}{\partial \Delta_i^l}$$

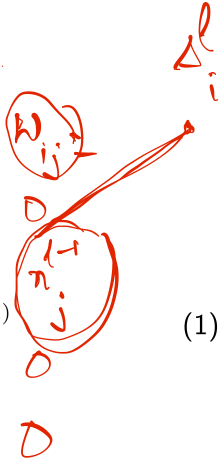


We need gradients wrt parameters W and b

- $x^{(l-1)} \xrightarrow{W^{(l)},\mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \Sigma_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,

We need gradients wrt parameters W and b

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \Sigma_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,
- $$\frac{\partial \ell}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \tag{1}$$



We need gradients wrt parameters W and b

- $x^{(l-1)} \xrightarrow{W^{(l)}, \mathbf{b}^{(l)}} s^{(l)} \xrightarrow{\sigma} x^{(l)}$
- $W_{i,j}^{(l)}$ and $\mathbf{b}^{(l)}$ influence the loss through $s^{(l)}$ via $s_i^{(l)} = \Sigma_j W_{i,j}^{(l)} x_j^{(l-1)} + b_i^{(l)}$,
- $$\frac{\partial \ell}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial W_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \tag{1}$$
- $$\frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \frac{\partial s_i^{(l)}}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} \tag{2}$$

Summary of Backprop

- From the definition of loss, obtain $\dfrac{\partial l}{\partial x_i^{(l)}}$

Summary of Backprop

- From the definition of loss, obtain $\frac{\partial l}{\partial x_i^{(l)}}$
- Recursively compute the loss derivatives wrt the activations $x^{(l)}$

$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)})$ and $\frac{\partial \ell}{\partial x_j^{(l-1)}} = \Sigma_i \frac{\partial \ell}{\partial s_i^{(l)}} w_{i,j}^{(l)}$

Summary of Backprop

- From the definition of loss, obtain $\frac{\partial l}{\partial x_i^{(l)}}$
- Recursively compute the loss derivatives wrt the activations

$$\frac{\partial \ell}{\partial s_i^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \sigma'(s_i^{(l)}) \text{ and } \frac{\partial \ell}{\partial x_j^{(l-1)}} = \Sigma_i \frac{\partial \ell}{\partial s_i^{(l)}} w_{i,j}^{(l)}$$

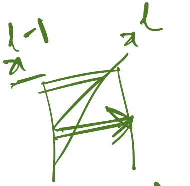
- Then wrt the parameters

$$\frac{\partial \ell}{\partial w_{i,j}^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}} x_j^{(l-1)} \text{ and } \frac{\partial \ell}{\partial b_i^{(l)}} = \frac{\partial \ell}{\partial s_i^{(l)}}$$

Jocobian in Tensorial form

- $\psi : \mathcal{R}^N \rightarrow \mathcal{R}^M$ then $\left[\frac{\partial \psi}{\partial x}\right] = \begin{bmatrix} \frac{\partial \psi_1}{\partial x_1} & \cdots & \frac{\partial \psi_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_M}{\partial x_1} & \cdots & \frac{\partial \psi_M}{\partial x_N} \end{bmatrix}$

Jocobian in Tensorial form



N M

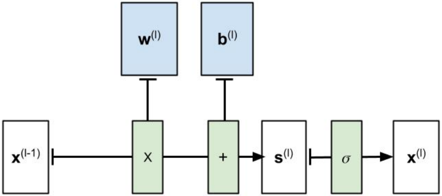
- $\psi : \mathcal{R}^N \to \mathcal{R}^M$ then $\left[\frac{\partial \psi}{\partial x}\right] = \begin{bmatrix} \frac{\partial \psi_1}{\partial x_1} & \cdots & \frac{\partial \psi_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_M}{\partial x_1} & \cdots & \frac{\partial \psi_M}{\partial x_N} \end{bmatrix}$

$W^{m \times N}$

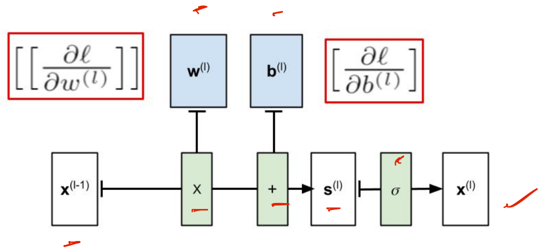
- $\psi : \mathcal{R}^{N \times M} \to \mathcal{R}$ then $\left[\left[\frac{\partial \psi}{\partial x}\right]\right] = \begin{bmatrix} \frac{\partial \psi}{\partial w_{1,1}} & \cdots & \frac{\partial \psi}{\partial w_{1,M}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi}{\partial w_{N,1}} & \cdots & \frac{\partial \psi}{\partial w_{N,M}} \end{bmatrix}$

loss

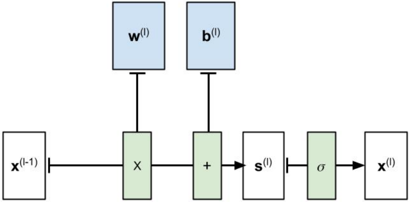
Forward Pass



Goal of Backward Pass

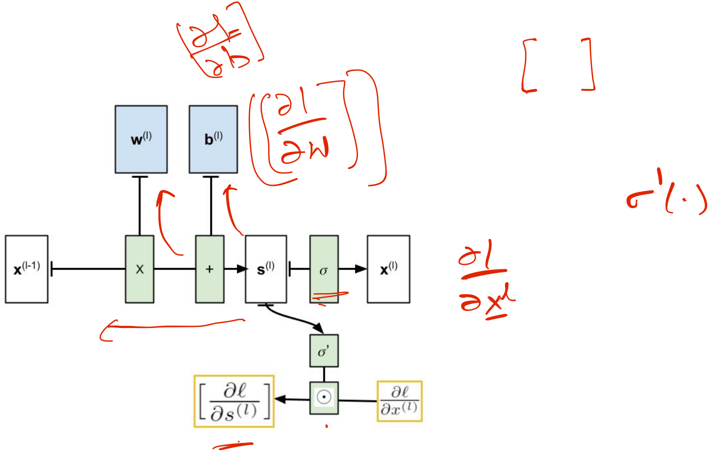


Begin from succeeding layer

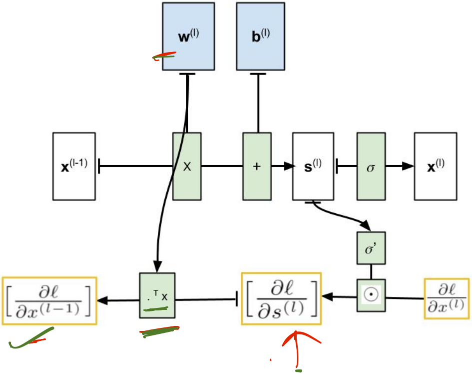


$$\boxed{\frac{\partial \ell}{\partial x^{(l)}}}$$

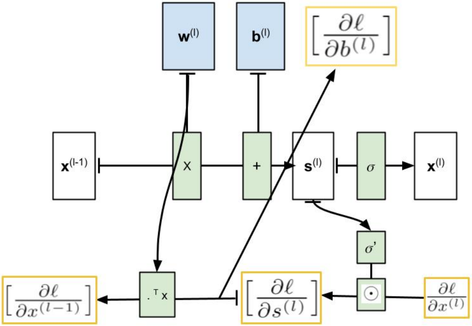
Begin from succeeding layer



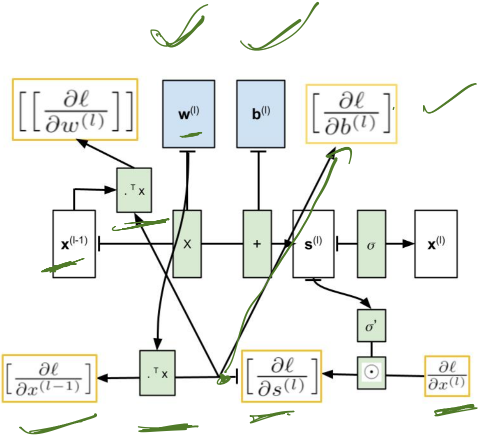
Begin from succeeding layer



Begin from succeeding layer



Begin from succeeding layer



Update the parameters

- $W^{(l)} = W^{(l)} - \eta \left[\left[\frac{\partial \ell}{\partial w^{(l)}} \right] \right]$ and $\mathbf{b}^{(l)} = \mathbf{b}^{(l)} - \eta \left[\frac{\partial \ell}{\partial b^{(l)}} \right]$

Observations

- BP is basically simple: applying chain rule iteratively

Observations

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)

Observations

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations

Observations

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations
- Nonlinearities go into simple element wise operations

Observations

- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations
- Nonlinearities go into simple element wise operations
- BP Needs all the intermediate layer results to be in memory

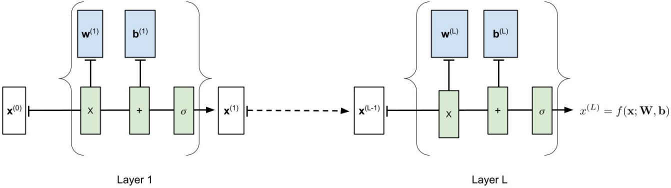
Observations

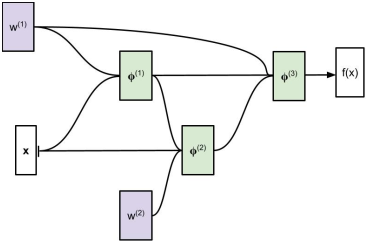
- BP is basically simple: applying chain rule iteratively
- It can be expressed in tensorial form (similar to the forward pass)
- Heavy computations are with the linear operations
- Nonlinearities go into simple element wise operations
- BP Needs all the intermediate layer results to be in memory
- Takes twice the computations of forward pass

$$\frac{\partial l}{\partial w^l} = \boxed{x^{(l-1)}} \frac{\partial l}{\partial x^l}$$

Beyond MLP

- We can generalize MLP

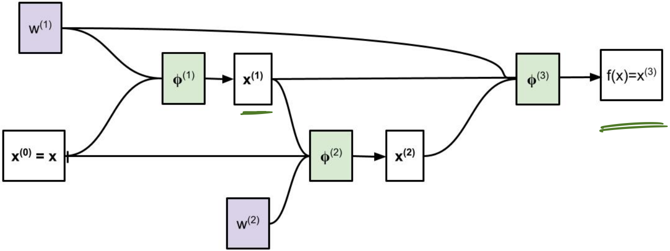




To an arbitrary Directed Acyclic Graph (DAG)

Forward pass in the computational graph

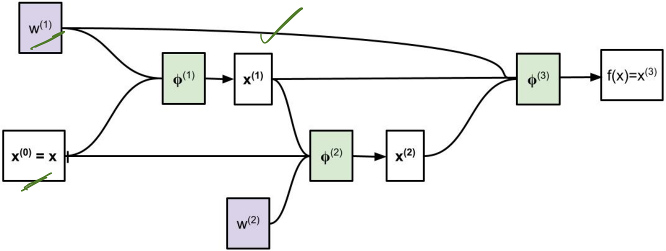
భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad



$\mathcal{L}$

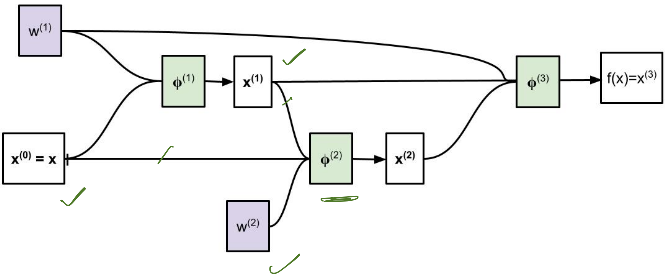
- $x^{(0)} = x$

Forward pass in the computational graph



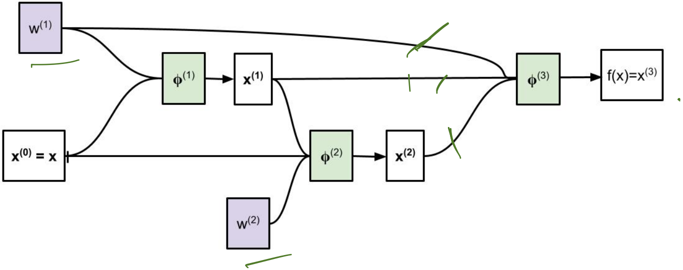
- $x^{(0)} = x$
- $x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)})$

Forward pass in the computational graph



- $x^{(0)} = x$
- $x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)})$
- $x^{(2)} = \phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)})$

Forward pass in the computational graph



- $x^{(0)} = x$
- $x^{(1)} = \phi^{(1)}(x^{(0)}; w^{(1)})$
- $x^{(2)} = \phi^{(2)}(x^{(0)}, x^{(1)}; w^{(2)})$
- $f(x) = x^{(3)} = \phi^{(3)}(x^{(1)}, x^{(2)}; w^{(1)})$

Notation: Jacobian of a general transformation

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

- $$\text{if } (a_1 \dots a_Q) = \phi(b_1 \dots b_R) \text{ then we use the notation} \tag{3}$$

$$\left[\frac{\partial a}{\partial b}\right] = J_\phi^T = \begin{bmatrix} \frac{\partial a_1}{\partial b_1} & \cdots & \frac{\partial a_Q}{\partial b_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_1}{\partial b_R} & \cdots & \frac{\partial a_Q}{\partial b_R} \end{bmatrix} \tag{4}$$

Notation: Jacobian of a general transformation

- $$\text{if } (a_1 \dots a_Q) = \phi(b_1 \dots b_R) \text{ then we use the notation} \tag{3}$$

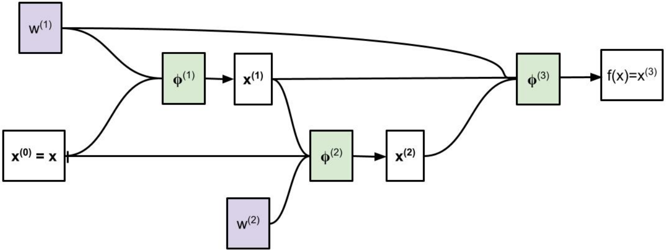
$$\left[\frac{\partial a}{\partial b}\right] = J_\phi^T = \begin{bmatrix} \frac{\partial a_1}{\partial b_1} & \cdots & \frac{\partial a_Q}{\partial b_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_1}{\partial b_R} & \cdots & \frac{\partial a_Q}{\partial b_R} \end{bmatrix} \tag{4}$$

$x^l = \phi(x^{l-1}, w^l)$

- $$\text{if } (a_1 \dots a_Q) = \phi(b_1 \dots b_R; c_1 \dots c_S) \text{ then we use the notation} \tag{5}$$

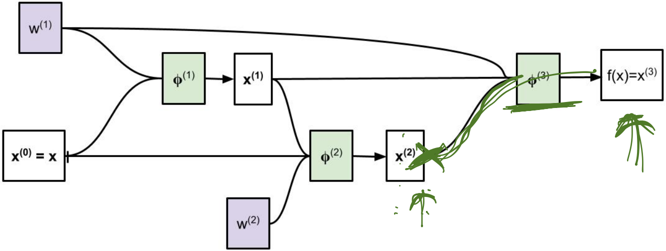
$$\left[\frac{\partial a}{\partial c}\right] = J_{\phi|c}^T = \begin{bmatrix} \frac{\partial a_1}{\partial c_1} & \cdots & \frac{\partial a_Q}{\partial c_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_1}{\partial C_S} & \cdots & \frac{\partial a_Q}{\partial c_S} \end{bmatrix} \tag{6}$$

Backward pass



- From the loss equation, we can compute $\left[\frac{\partial \ell}{\partial x^{(3)}}\right]$

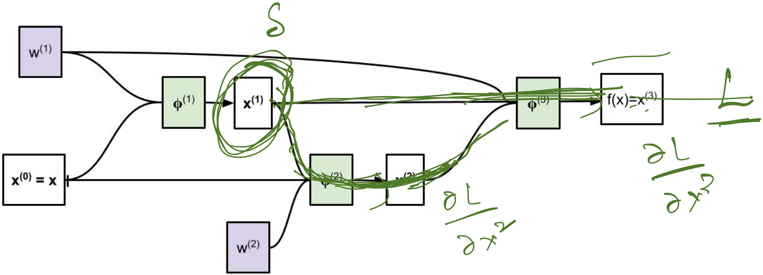
Backward pass



- From the loss equation, we can compute $\left[\frac{\partial \ell}{\partial x^{(3)}}\right]$
- $$\left[\frac{\partial \ell}{\partial x^{(2)}}\right] = \left[\frac{\partial x^{(3)}}{\partial x^{(2)}}\right]\left[\frac{\partial \ell}{\partial x^{(3)}}\right] = J^T_{\phi^{(3)}|x^{(2)}} \cdot \left[\frac{\partial \ell}{\partial x^{(3)}}\right]$$

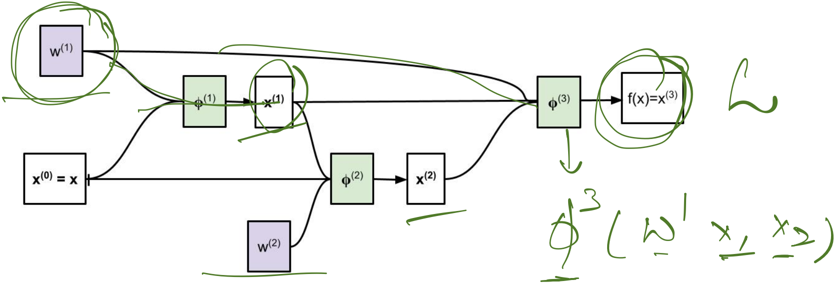
$\sigma' \cdot \frac{\partial l}{\partial a'}$

Backward pass



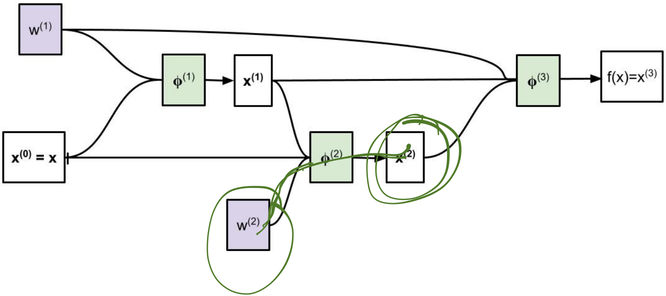
- From the loss equation, we can compute $\left[\frac{\partial \ell}{\partial x^{(3)}}\right]$
- $$\left[\frac{\partial \ell}{\partial x^{(2)}}\right] = \left[\frac{\partial x^{(3)}}{\partial x^{(2)}}\right]\left[\frac{\partial \ell}{\partial x^{(3)}}\right] = J^T_{\phi^{(3)}|x^{(2)}}\left[\frac{\partial \ell}{\partial x^{(3)}}\right]$$
- $$\begin{aligned}\left[\frac{\partial \ell}{\partial x^{(1)}}\right] &= \left[\frac{\partial x^{(3)}}{\partial x^{(1)}}\right]\left[\frac{\partial \ell}{\partial x^{(3)}}\right] + \left[\frac{\partial x^{(2)}}{\partial x^{(1)}}\right]\left[\frac{\partial \ell}{\partial x^{(2)}}\right] \\ &= J^T_{\phi^{(3)}|x^{(1)}}\left[\frac{\partial \ell}{\partial x^{(3)}}\right] + J^T_{\phi^{(2)}|x^{(1)}}\left[\frac{\partial \ell}{\partial x^{(2)}}\right]\end{aligned}$$

Backward pass



- $$\left[\frac{\partial \ell}{\partial w^{(1)}}\right] = \left[\frac{\partial x^{(3)}}{\partial w^{(1)}}\right]\left[\frac{\partial \ell}{\partial x^{(3)}}\right] + \left[\frac{\partial x^{(1)}}{\partial w^{(1)}}\right]\left[\frac{\partial \ell}{\partial x^{(1)}}\right]$$
$$= J^T_{\phi^{(3)}|w^{(1)}}\left[\frac{\partial \ell}{\partial x^{(3)}}\right] + J^T_{\phi^{(1)}|w^{(1)}}\left[\frac{\partial \ell}{\partial x^{(1)}}\right]$$

Backward pass



- $$\left[\frac{\partial \ell}{\partial w^{(1)}}\right] = \left[\frac{\partial x^{(3)}}{\partial w^{(1)}}\right]\left[\frac{\partial \ell}{\partial x^{(3)}}\right] + \left[\frac{\partial x^{(1)}}{\partial w^{(1)}}\right]\left[\frac{\partial \ell}{\partial x^{(1)}}\right]$$
$$= J^T_{\phi^{(3)}|w^{(1)}}\left[\frac{\partial \ell}{\partial x^{(3)}}\right] + J^T_{\phi^{(1)}|w^{(1)}}\left[\frac{\partial \ell}{\partial x^{(1)}}\right]$$

- $$\left[\frac{\partial \ell}{\partial w^{(2)}}\right] = \left[\frac{\partial x^{(2)}}{\partial w^{(2)}}\right]\left[\frac{\partial \ell}{\partial x^{(2)}}\right] = J^T_{\phi^{(2)}|w^{(2)}}\left[\frac{\partial \ell}{\partial x^{(2)}}\right]$$