Autograd

What AutoDiff Is Not

* Finite differences

→ One sided:

$$\frac{\partial}{\partial x_i} f(x_1, \dots, x_N) \approx \frac{f(x_1, \dots x_i + h, \dots) - f(x_1, \dots x_i, x_N)}{h}$$

→ Or Two sided

$$\frac{\partial}{\partial x_i} f(x_1, \dots, x_N) \approx \frac{f(x_1, \dots x_i + h, \dots) - f(x_1, \dots, x_i - h, \dots)}{2h}$$

* Challenges with finite differences

- Expensive: compute forward pass for each variable

- Numerically unstable

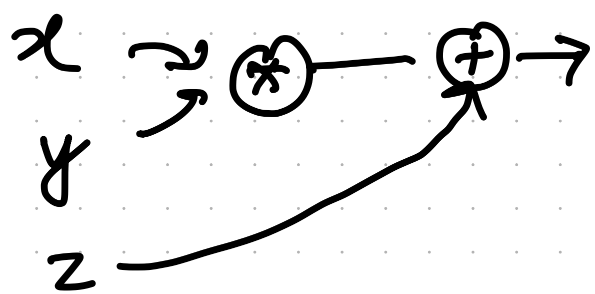
Computational Graphs

* Nodes : operations (+, *, ...)
* Edges : variables | Tensors

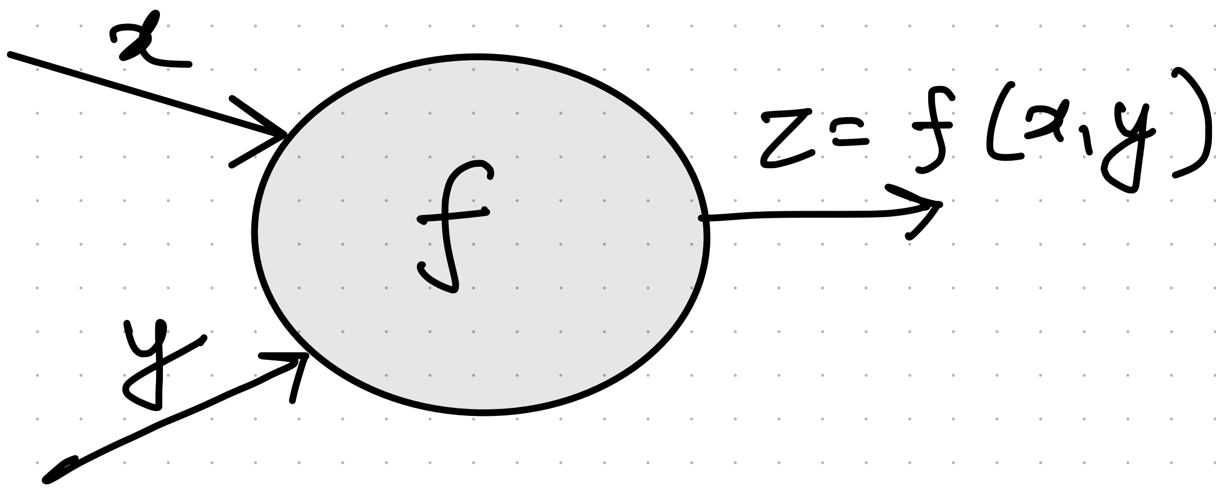
Computational Graphs

* Nodes : operations (+, *, ...)
* Edges : variables / Tensors (and data dependencies)

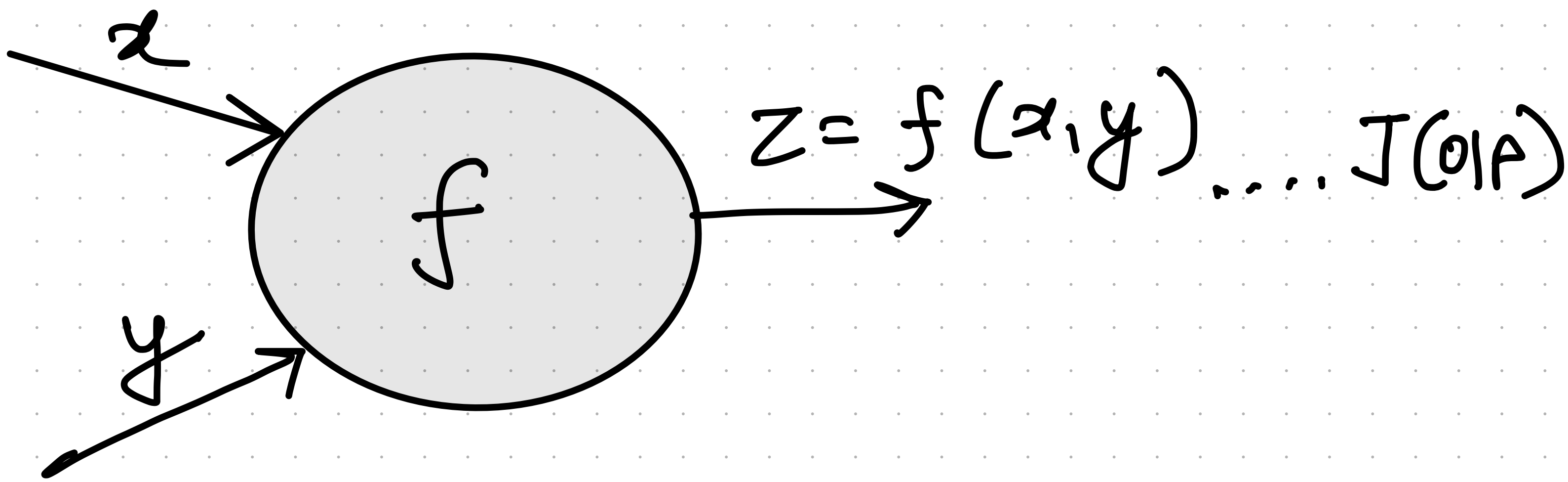
Example: $(x * y) + z$



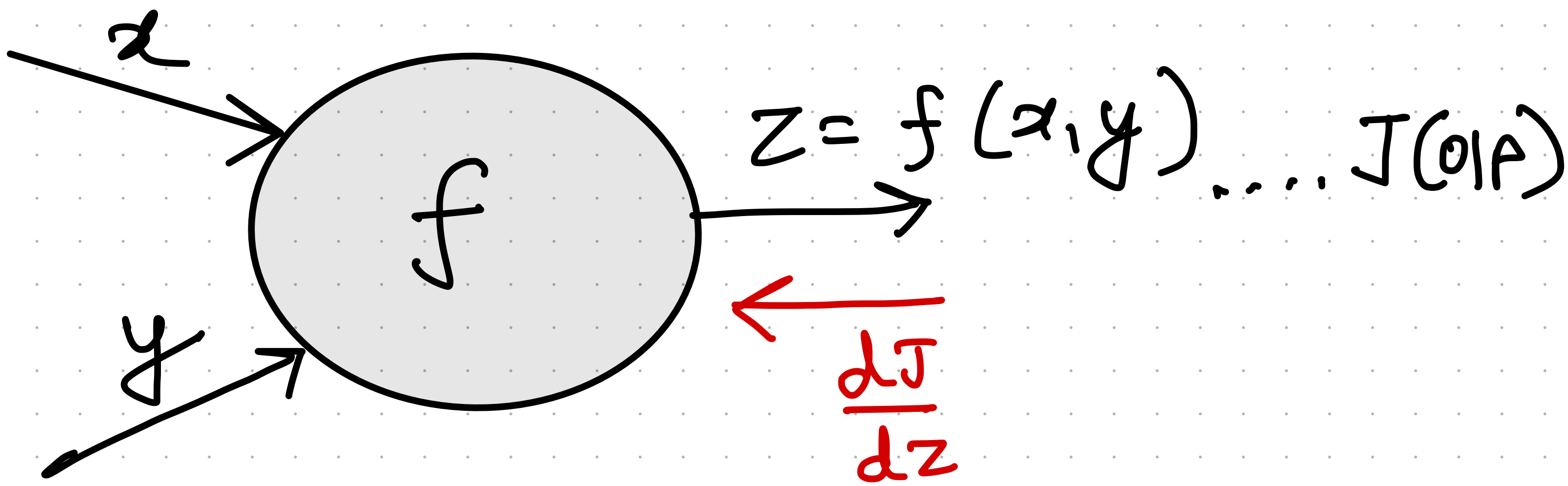
Back Prop Through Computational Graph



Back Prop Through Computational Graph



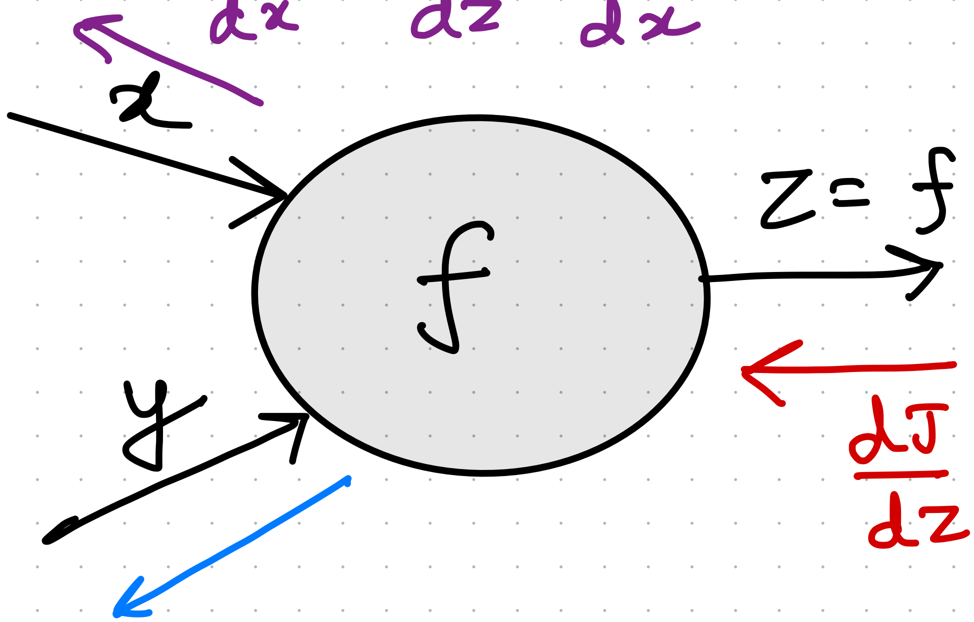
Back Prop Through Computational Graph



upstream gradient

Back Prop Through Computational Graph

$$\frac{dJ}{dx} = \frac{dJ}{dz} \cdot \frac{dz}{dx}$$



$z = f(x, y) \ldots J(O/P)$

$$\frac{dJ}{dz}$$

upstream gradient

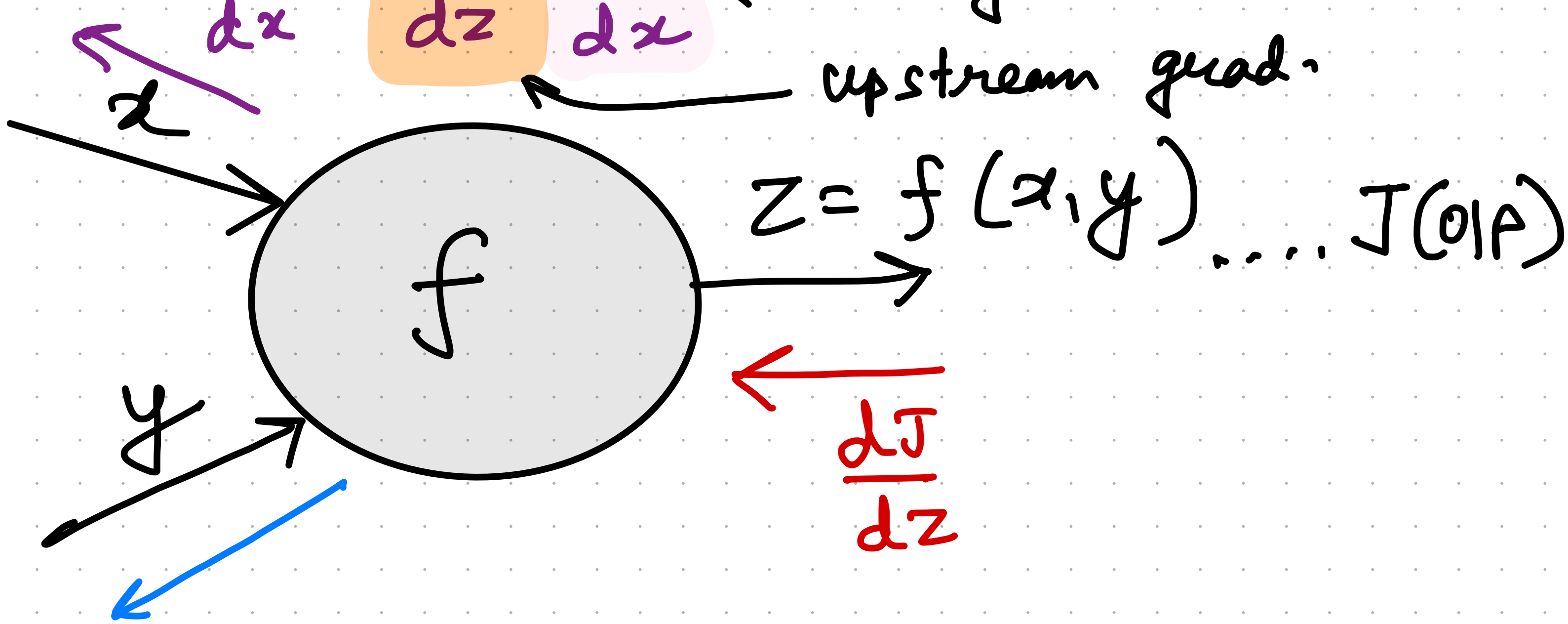
$$\frac{dJ}{dy} = \frac{dJ}{dz} \cdot \frac{dz}{dy}$$

Back Prop Through Computational Graph

$$\frac{dJ}{dx} = \frac{dJ}{dz} \cdot \frac{dz}{dx}$$

← Local grad.

upstream grad.



upstream gradient

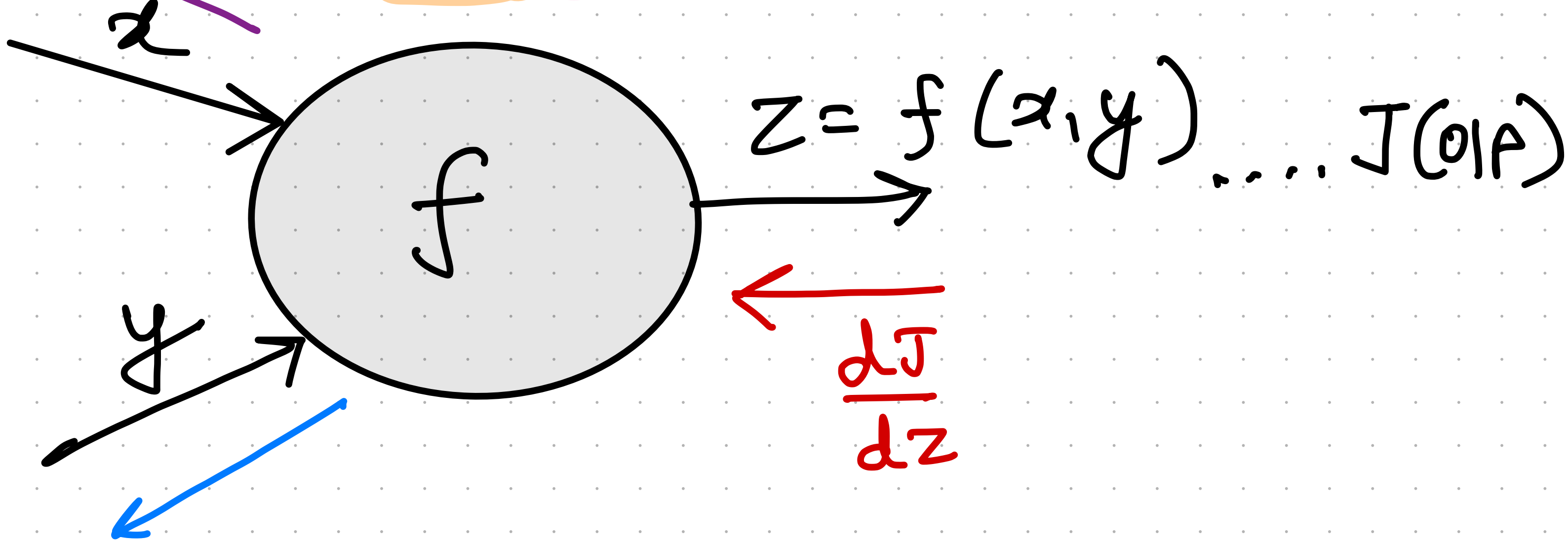
$$\frac{dJ}{dy} = \frac{dJ}{dz} \cdot \frac{dz}{dy}$$

upstream gradient

Local grad.

Back Prop Through Computational Graph

$$\frac{dJ}{dx} = \frac{dJ}{dz} \cdot \frac{dz}{dx}$$



$$\frac{dJ}{dy} = \frac{dJ}{dz} \cdot \frac{dz}{dy}$$

DOWNSTREAM GRADIENT

= UPSTREAM GRADIENT * LOCAL GRADIENT

$$\hat{y} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}$$

$$y = 1$$

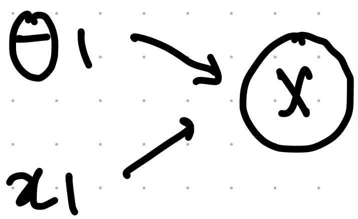
$$\text{LOSS} = -y\log\hat{y} - (1-y)\log(1-\hat{y})$$

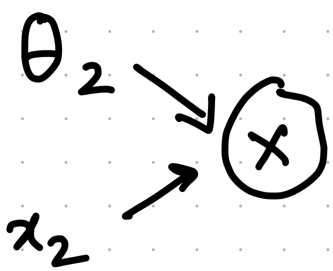
$$= -\log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

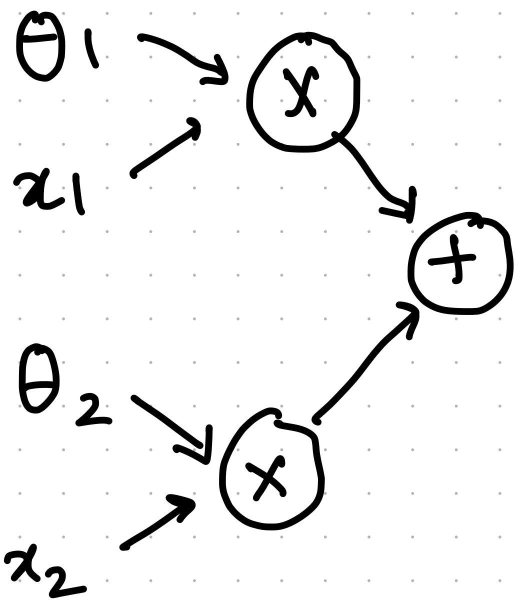
$$\begin{tikzcd}
\theta_1 \arrow[dr] & \\
& \times \\
x_1 \arrow[ur] &
\end{tikzcd}$$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

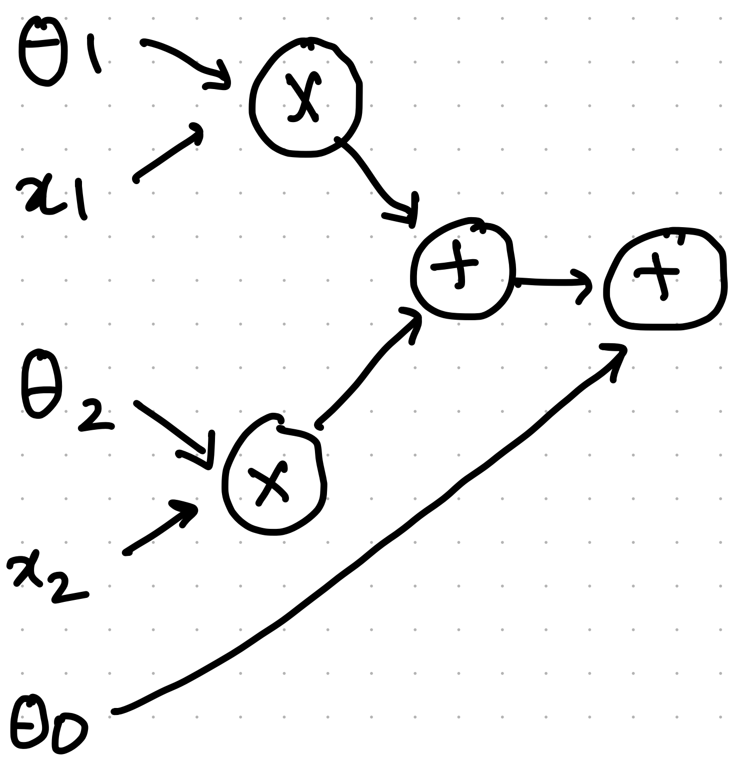




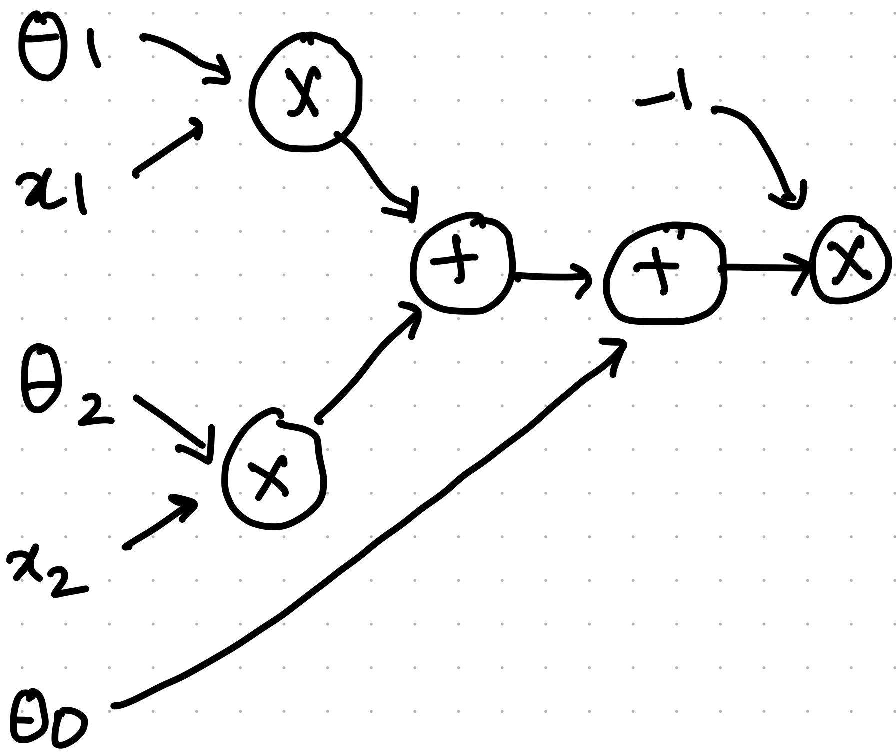
$$\mathrm{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



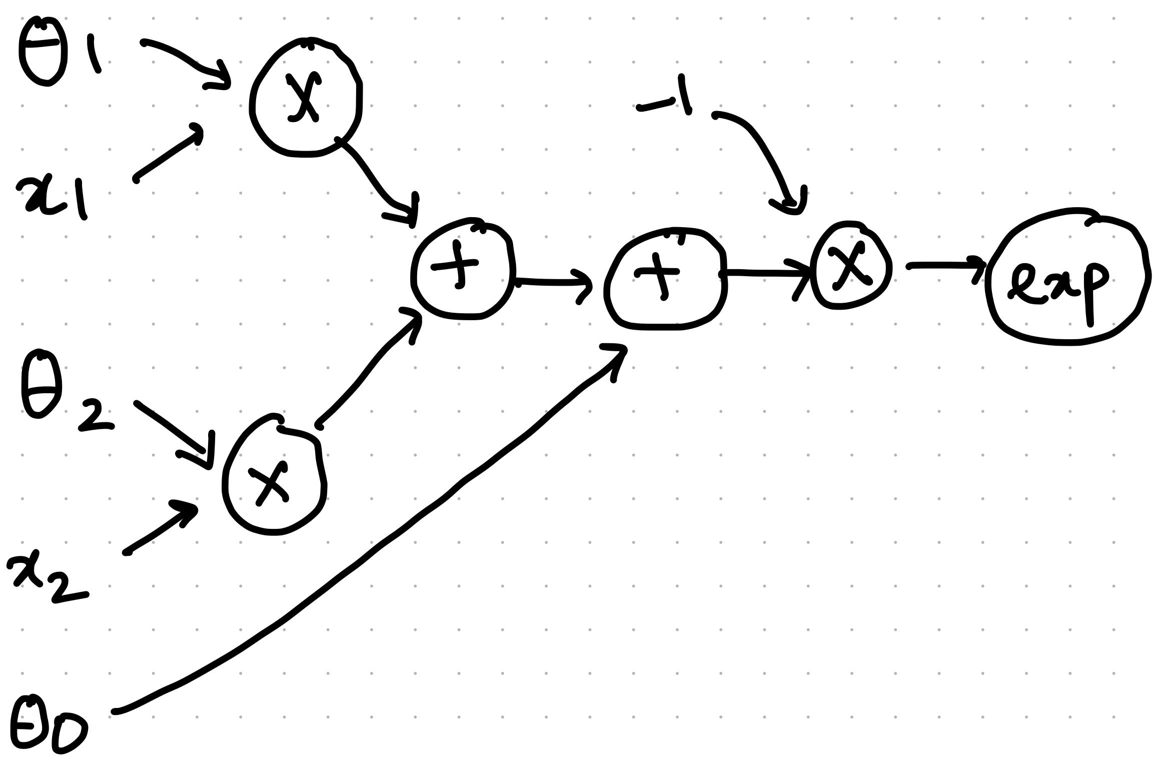
$$LOSS = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



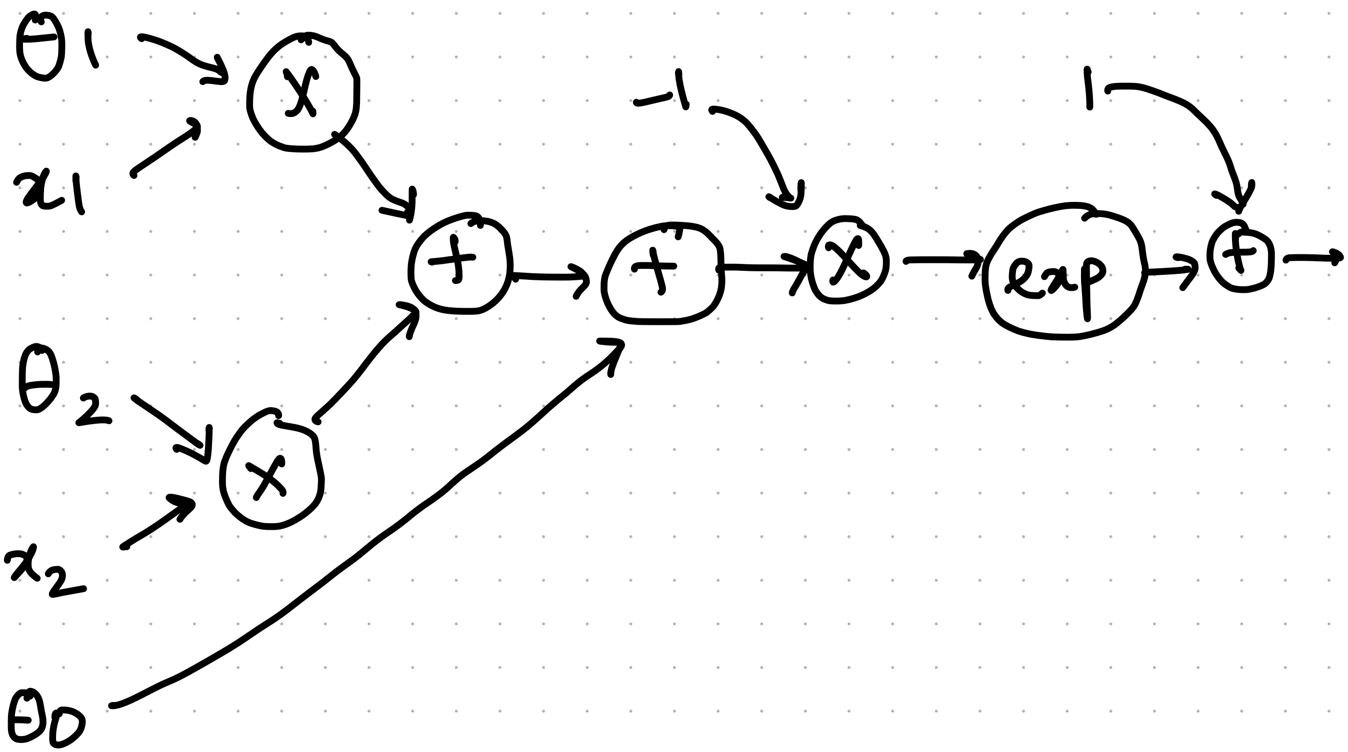
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



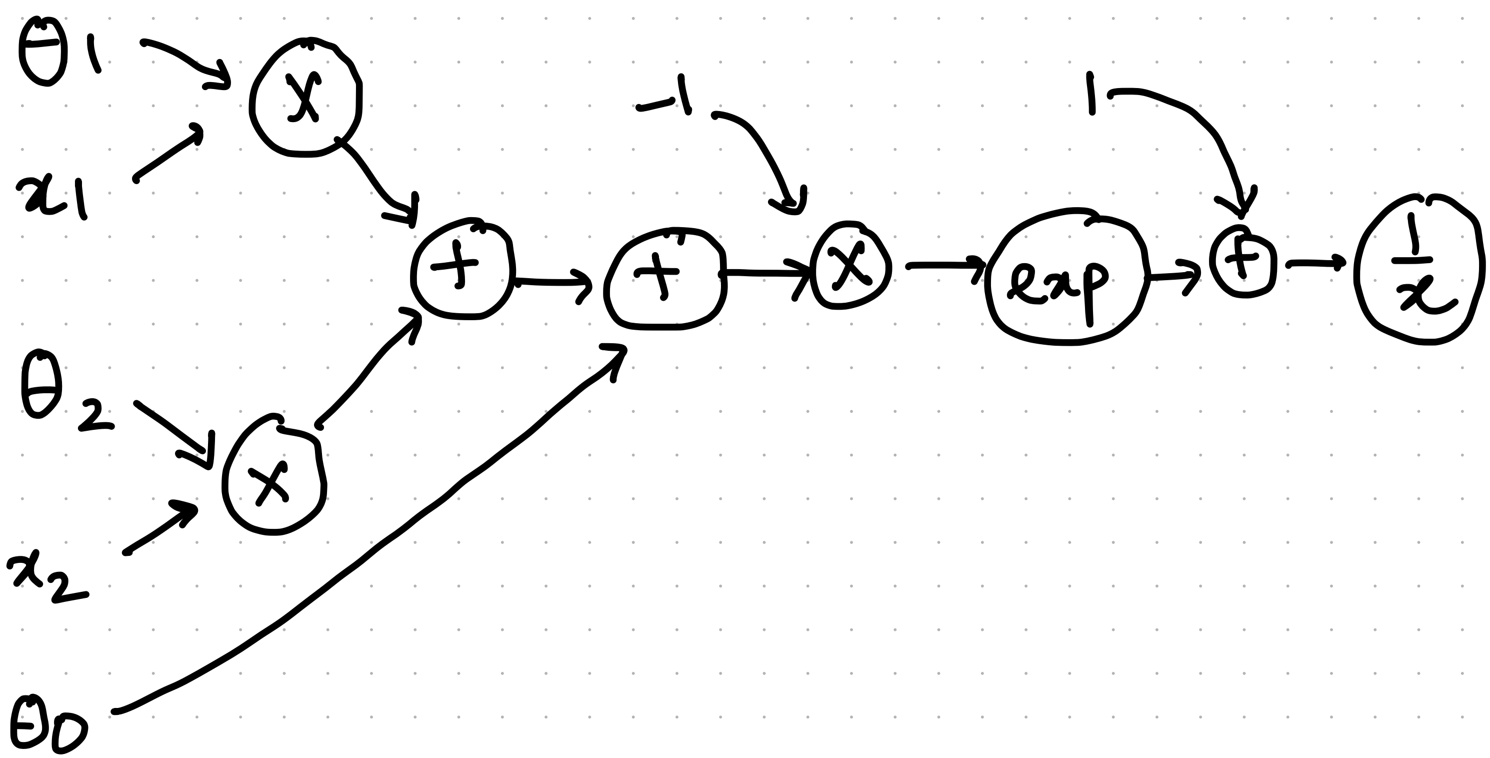
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



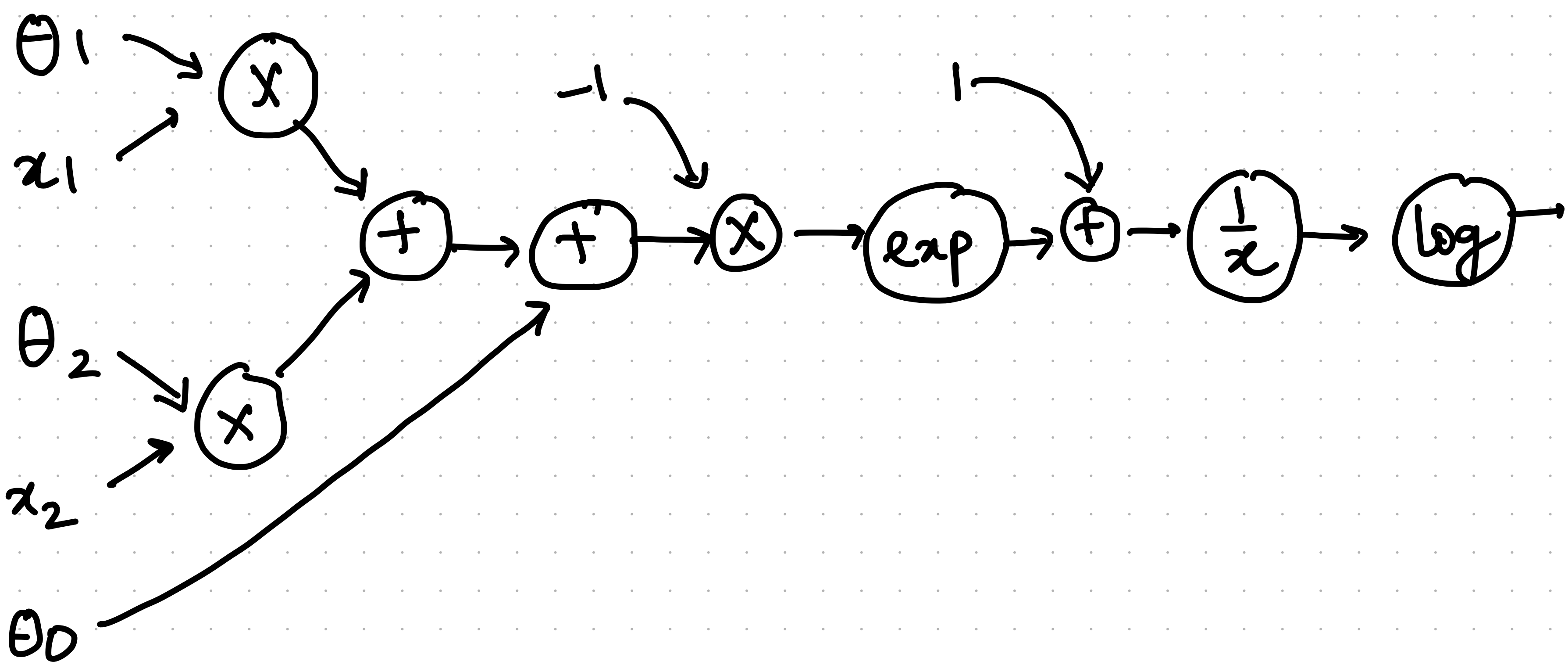
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}} \right)$$



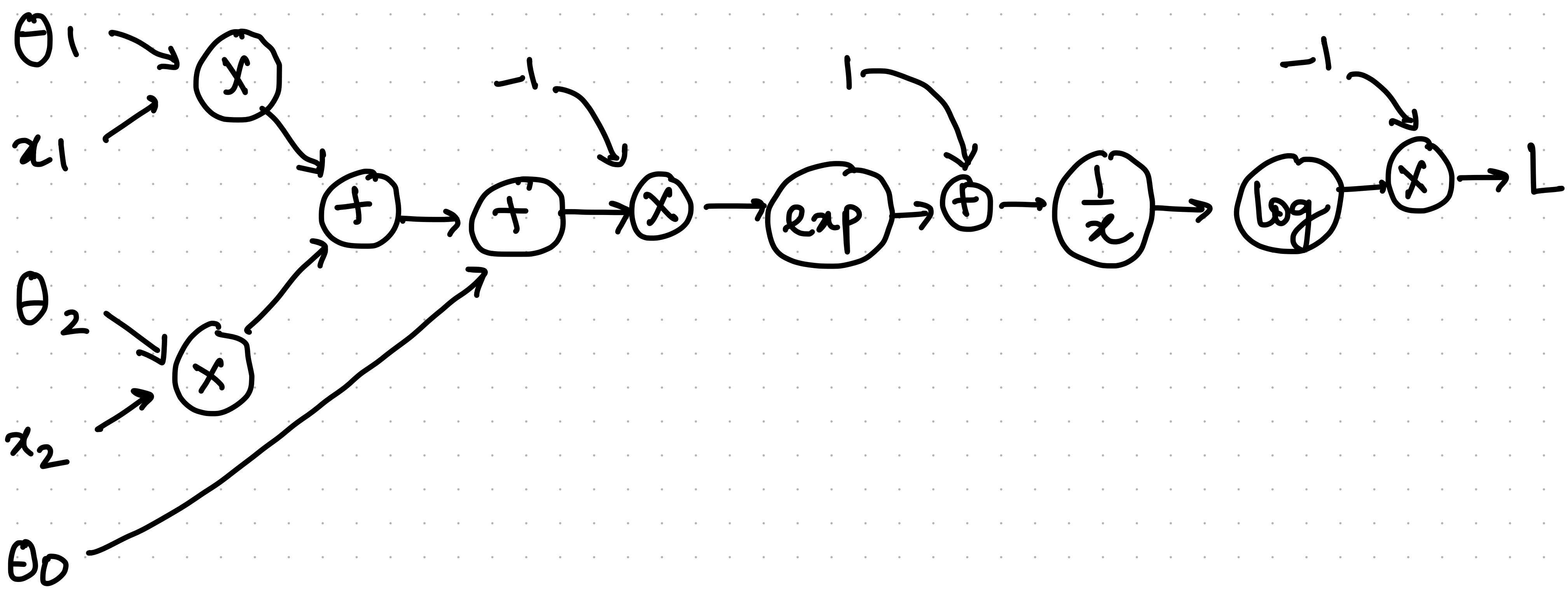
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



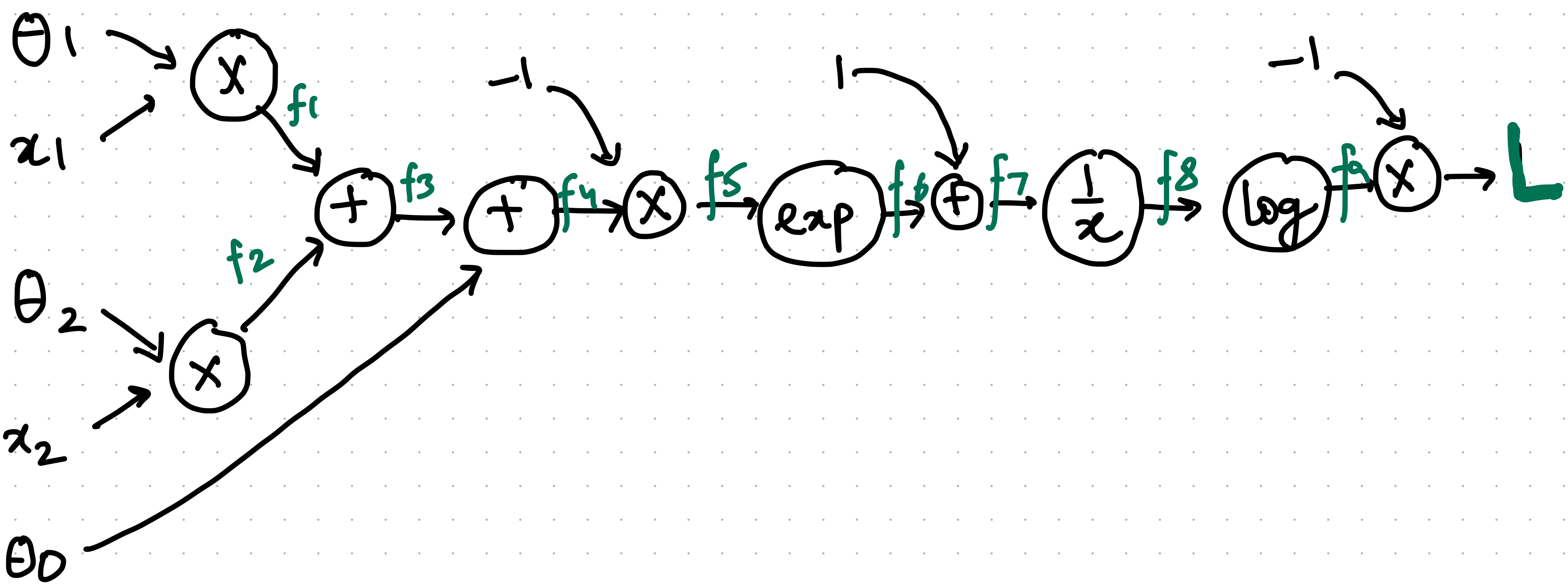
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



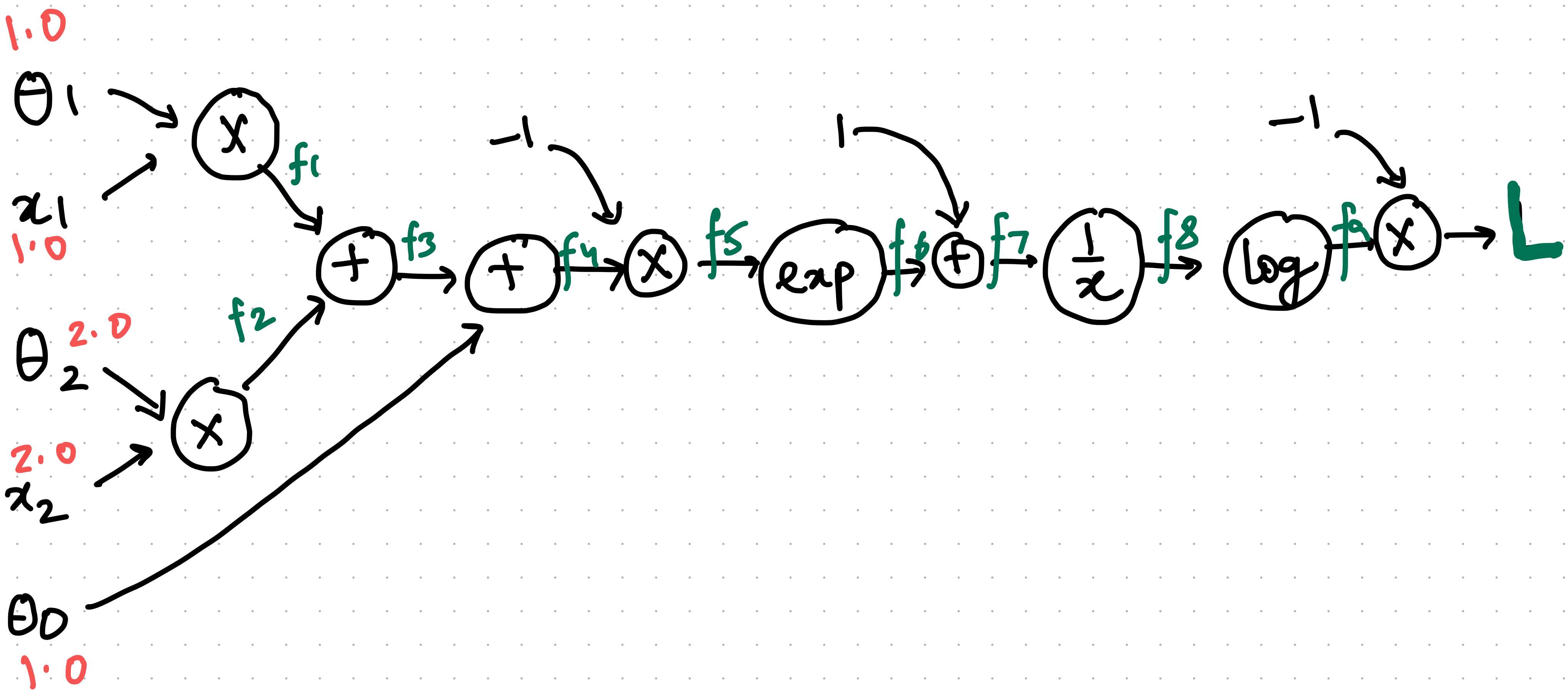
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



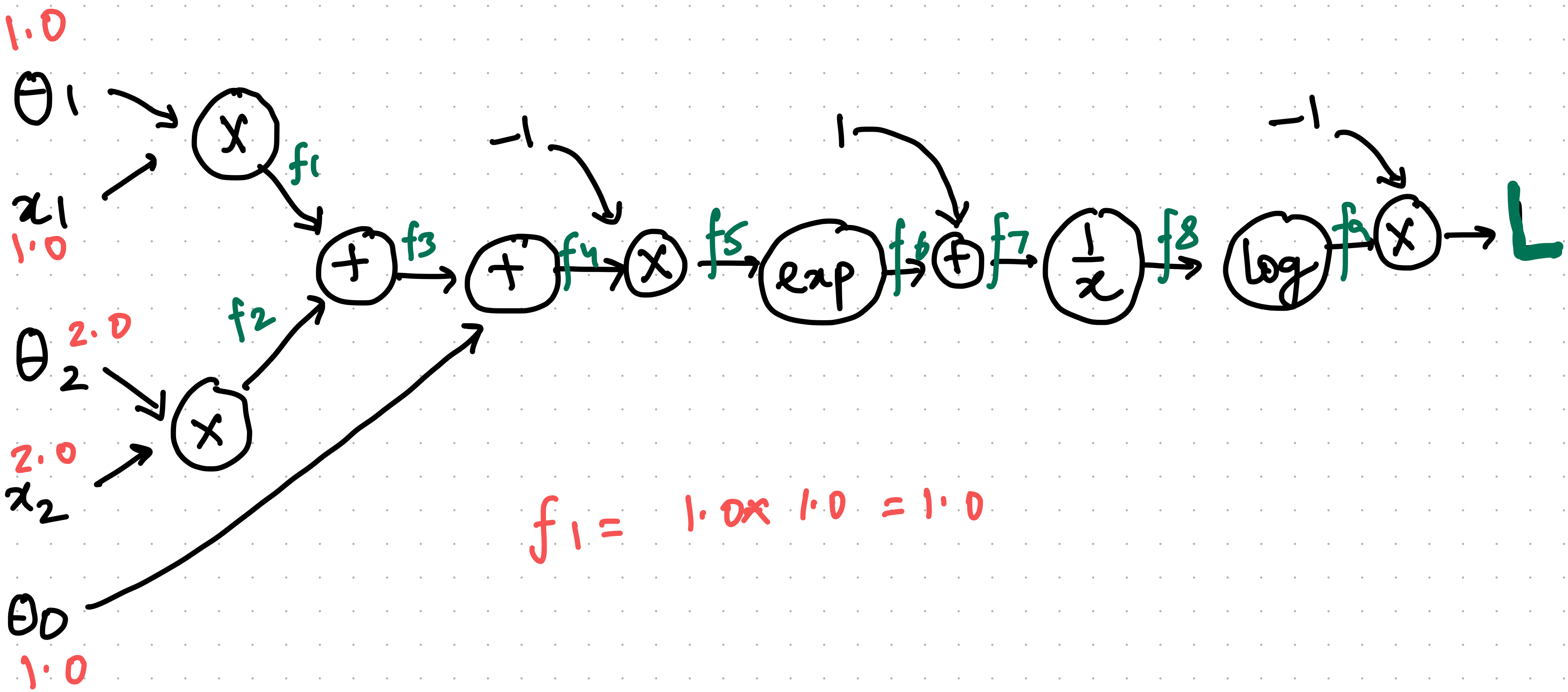
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

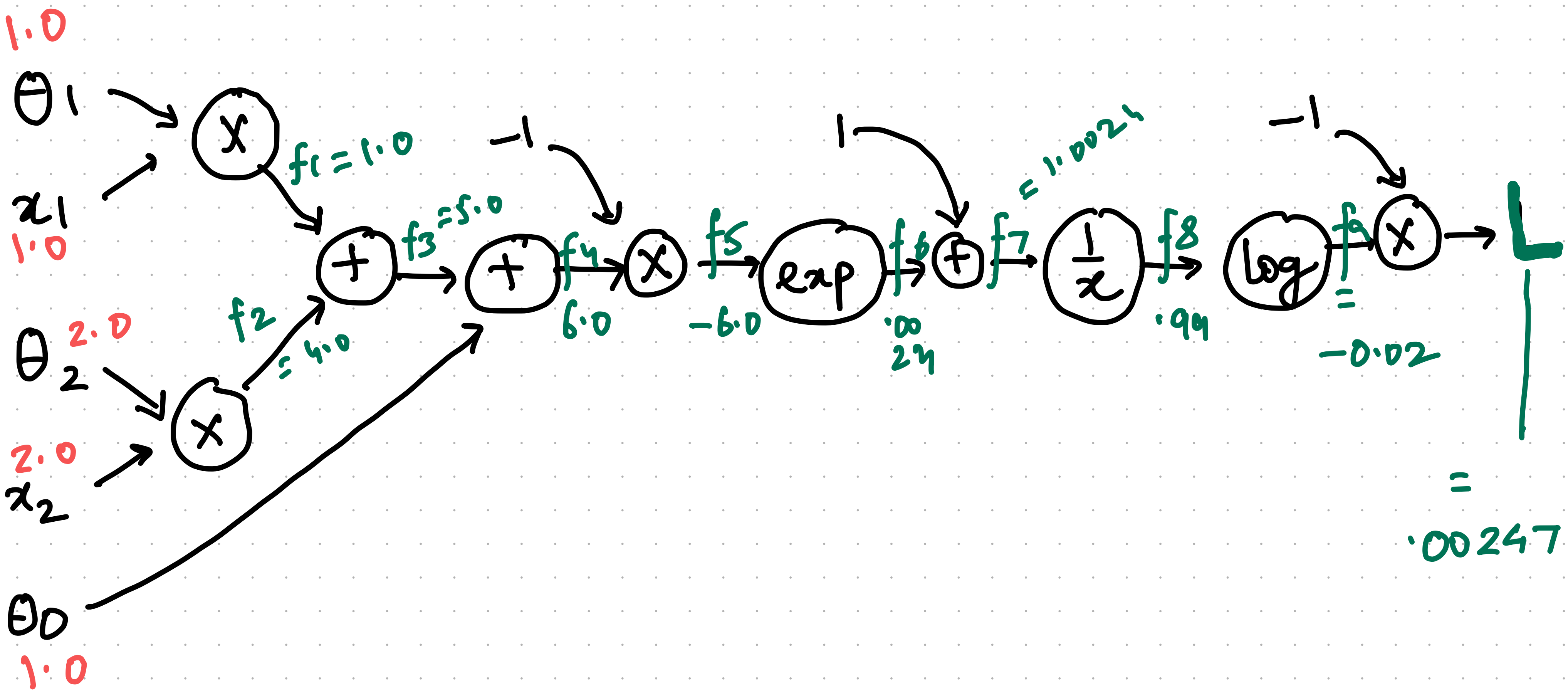


$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

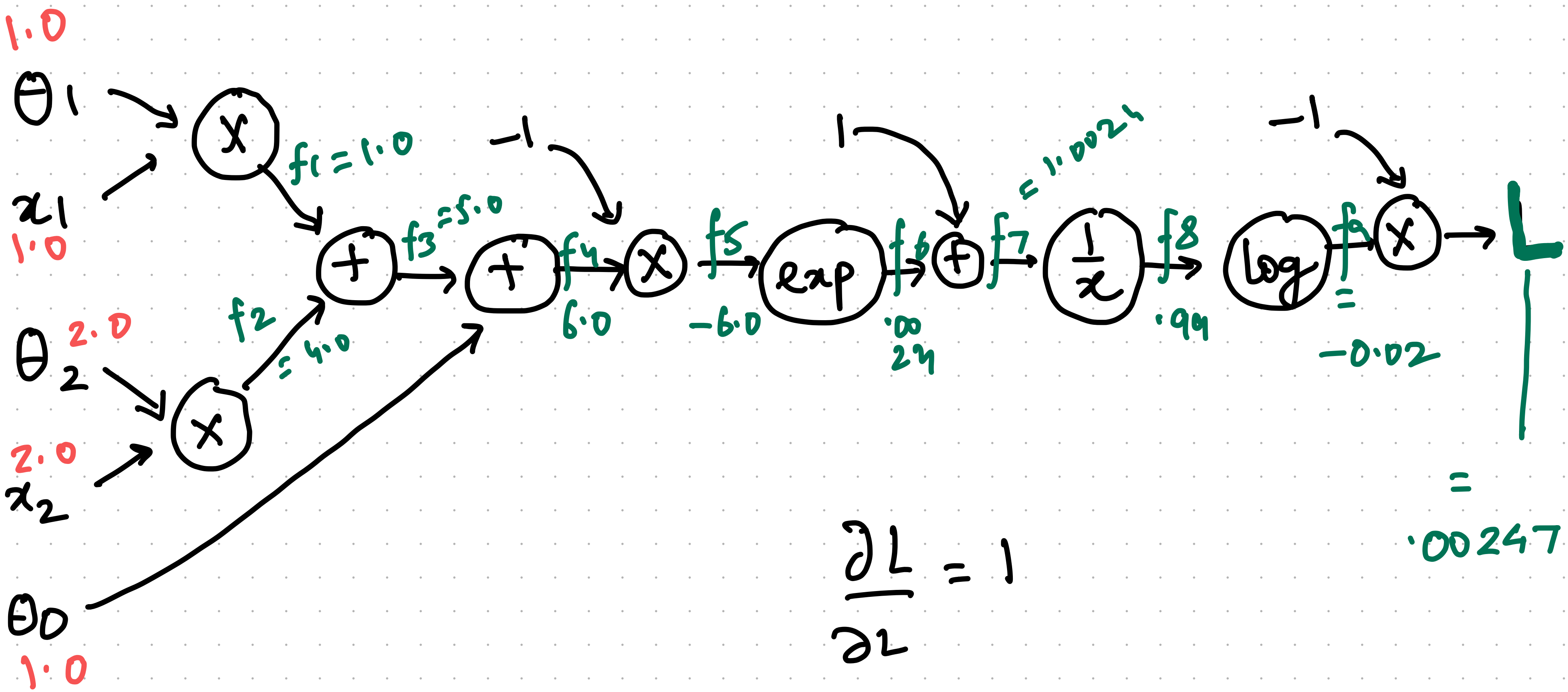


$f_1 = 1.0 \times 1.0 = 1.0$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

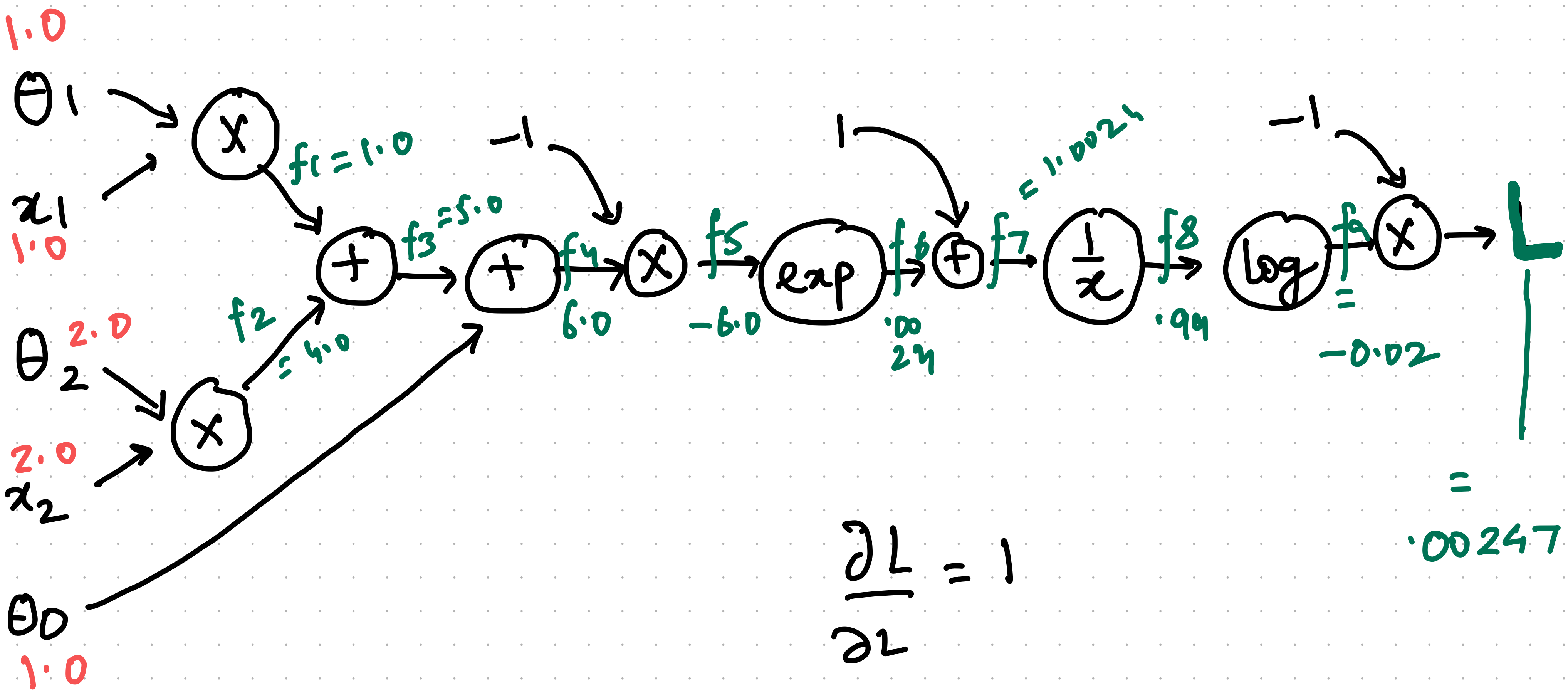


$$LOSS = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



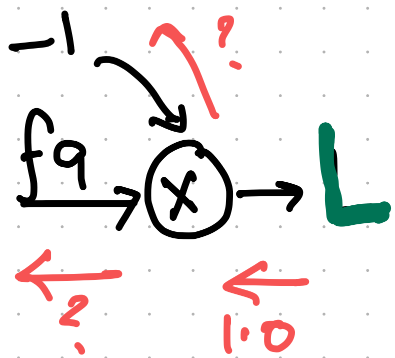
$$\frac{\partial L}{\partial L} = 1$$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



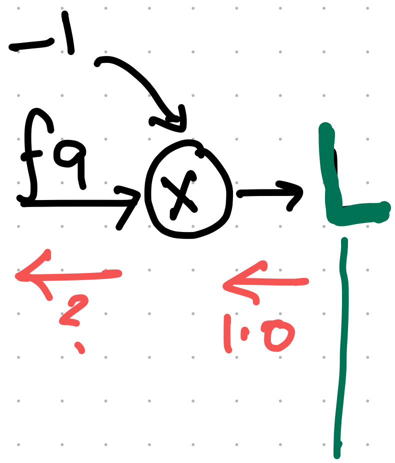
$$\frac{\partial L}{\partial L} = 1$$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



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

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



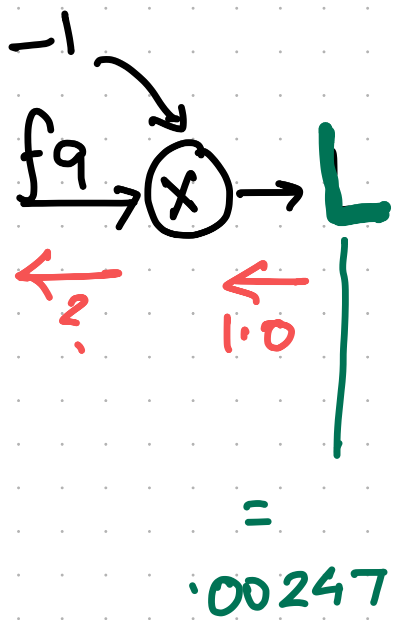
=

·00247

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

upstream gradiant = 1.0

$$\text{LOSS} = -1*\log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



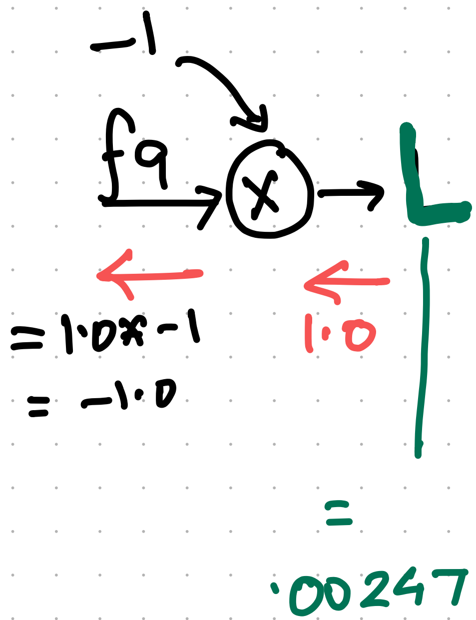
$$\frac{\partial L}{\partial L} = 1$$

upstream gradiant = 1.0

$$L = fq * -1$$

$$\frac{\partial L}{\partial fq} = -1 \qquad \text{LOCAL GRADIENT} = -1$$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



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

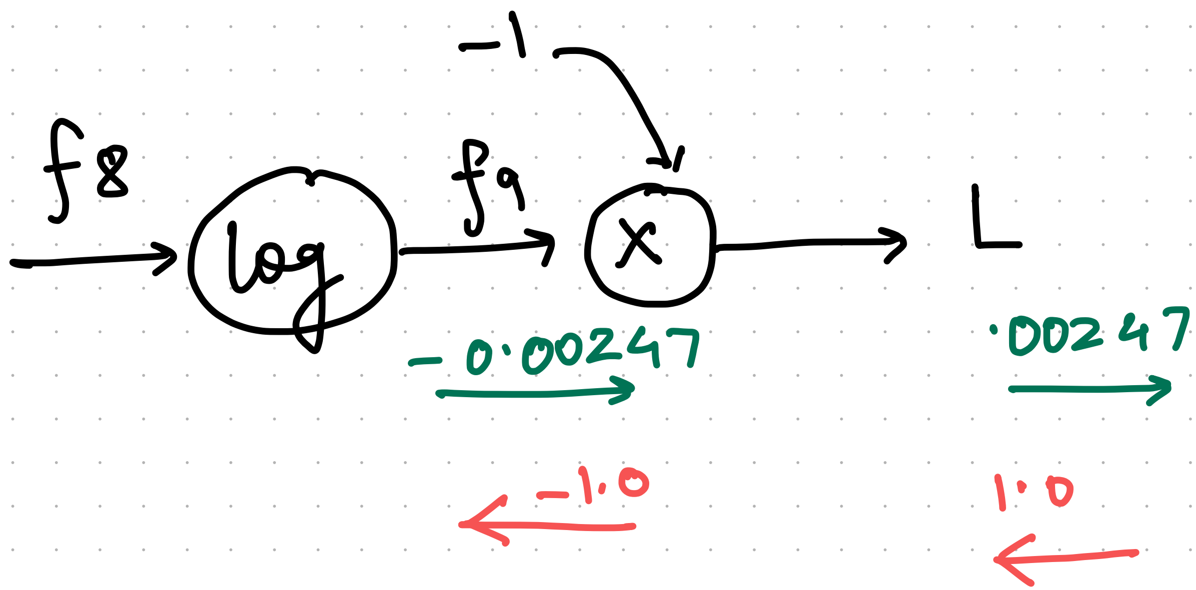
upstream gradient = 1.0

$$L = fq * -1$$

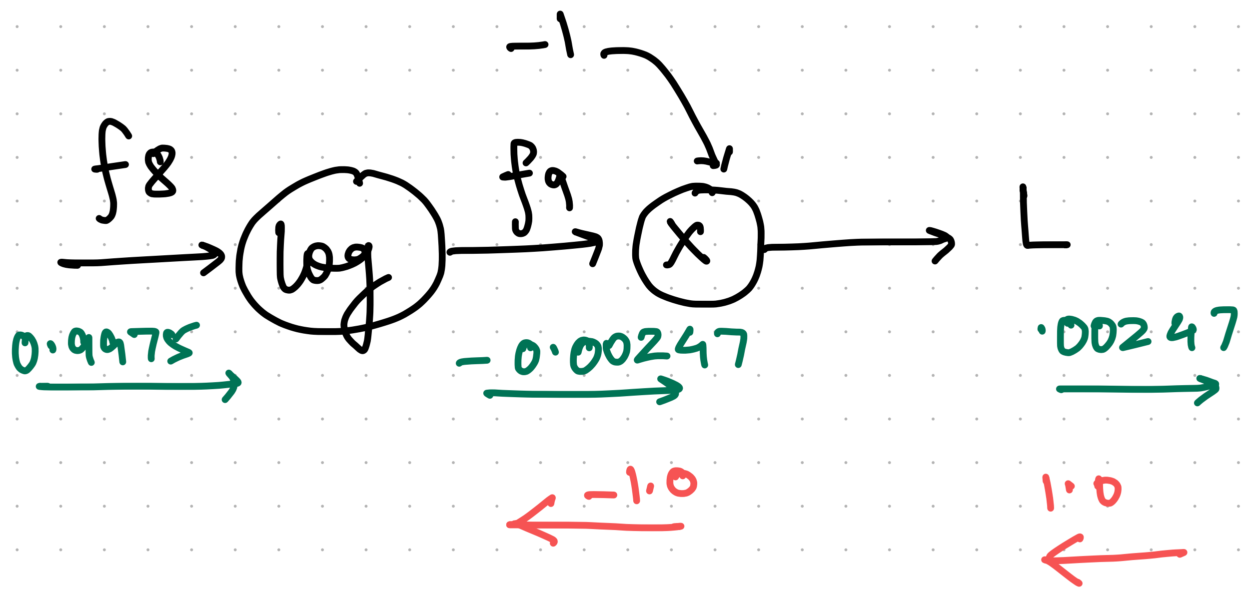
$$\frac{\partial L}{\partial fq} = -1$$

LOCAL GRADIENT = -1

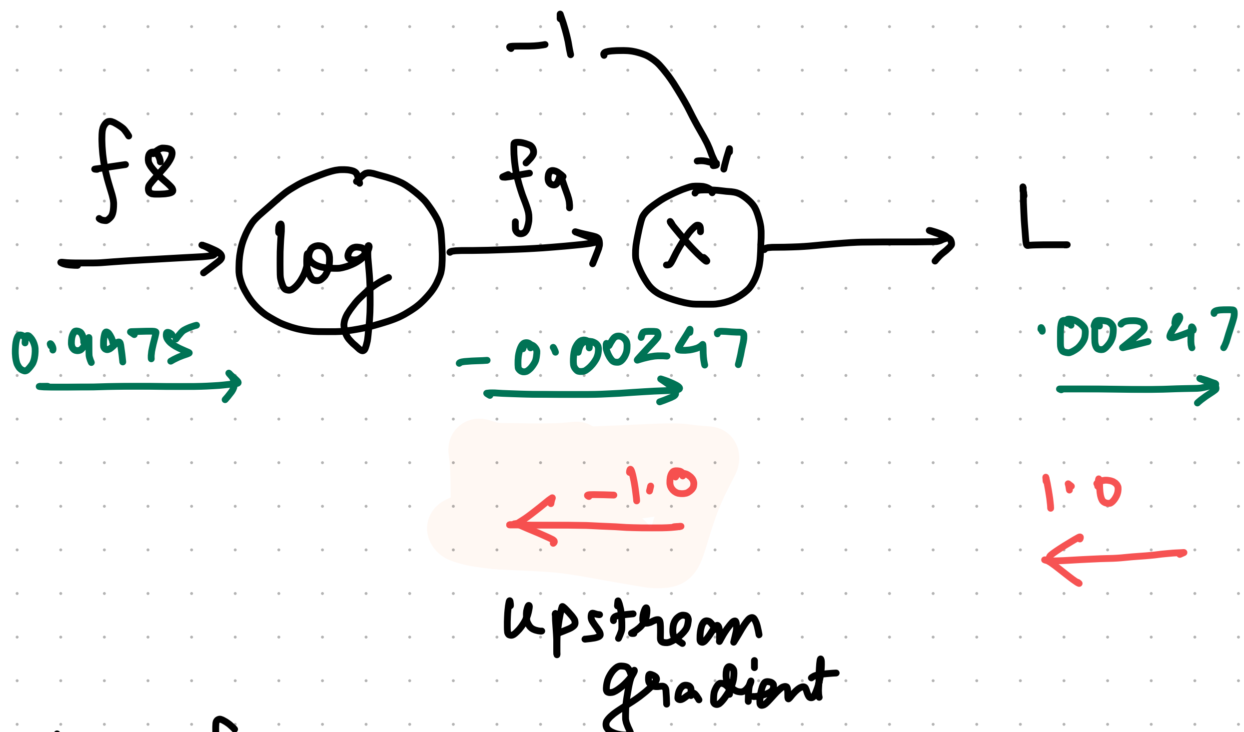
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



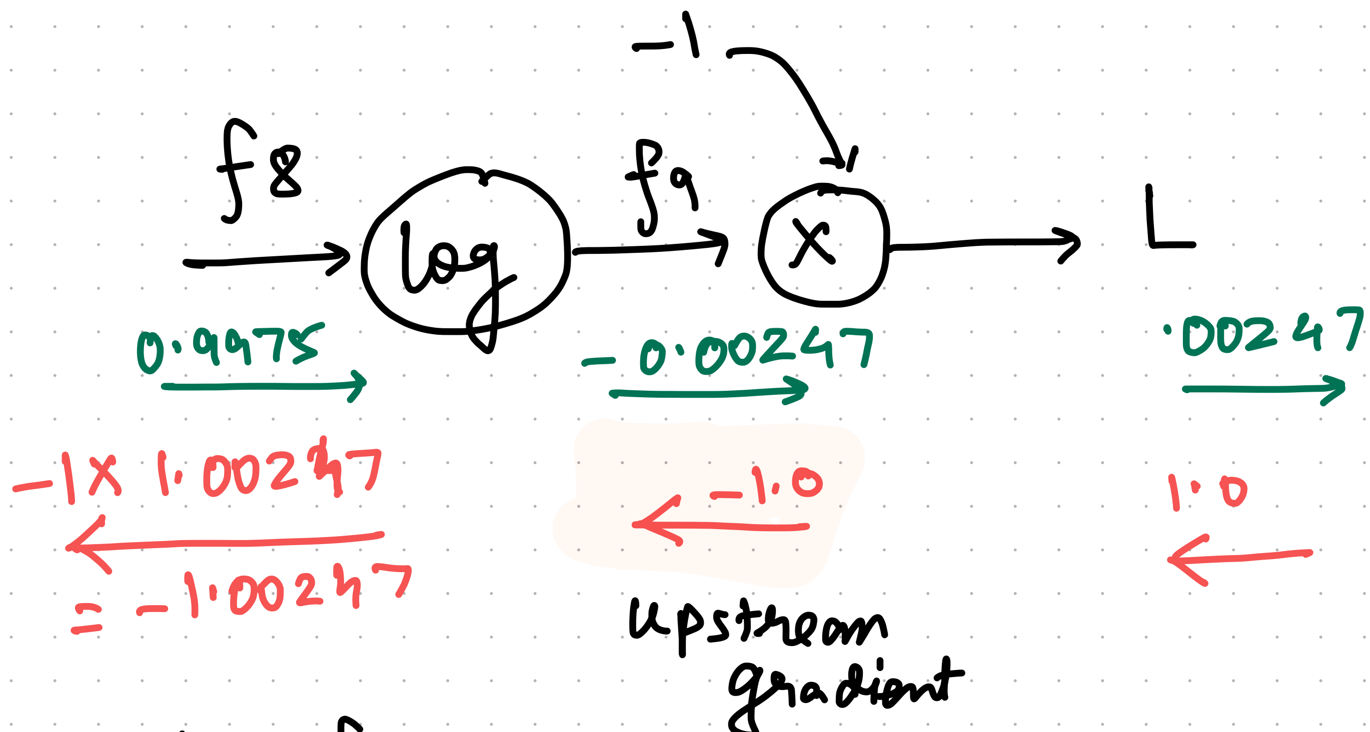
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



$$f_9 = \log f_8$$

$$\frac{\partial f_9}{\partial f_8} = \frac{1}{f_8} = \frac{1}{.9975} = 1.00247 = \text{local gradient}$$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



$$f9 = \log f8$$

$$\frac{\partial f9}{\partial f8} = \frac{1}{f8} = \frac{1}{.9975} = 1.00247 = \text{local gradient}$$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

$$f_7 \xrightarrow[1.00247]{} \boxed{\frac{1}{x}} \xrightarrow[0.9975]{} f_8 \;\cdots\cdots$$

$$\xleftarrow{\;-1.00247\;}$$

$$\text{Loss} = -1*\log\left(\frac{1}{1+e^{-(\theta_0+\theta_1x_1+\theta_2x_2)}}\right)$$

$$f_7 \xrightarrow{\quad} \frac{1}{x} \xrightarrow{\quad} f_8 \;\cdots\cdots$$

$\xrightarrow{\quad}$ 1.00247

0.9975 $\xrightarrow{\quad}$

$\xleftarrow{\quad}$ −1.00247 upstream gradient

$$f_8 = \frac{1}{f_7} \qquad \frac{\partial f_8}{\partial f_7} = \frac{-1}{f_7^2} = -0.9951$$

$=$ Local gradient

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

$$f_7 \longrightarrow \frac{1}{x} \longrightarrow f_8 \cdots\cdots$$

$\xrightarrow{\quad}$ 1.00247

0.9975 $\xrightarrow{\quad}$

$\xleftarrow{\quad}$

$-0.9951 * -1.00247$
$= 0.9975$

$\xleftarrow{\quad}$

-1.00247

upstream gradient

$$f_8 = \frac{1}{f_7} \qquad \frac{\partial f_8}{\partial f_7} = \frac{-1}{f_7^{\,2}} = -0.9951$$

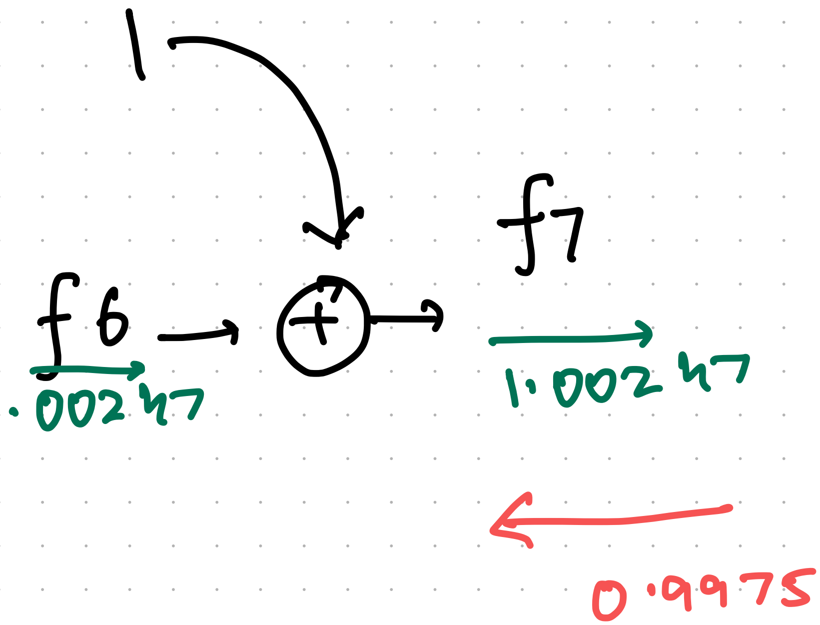
$=$ Local gradient

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

$$\begin{tikzcd} 1 \arrow[dr, bend left] & & \\ f6 \arrow[r, ".00247"'] & + \arrow[r, "1.00247"'] & f7 \end{tikzcd}$$

$$\xleftarrow{\quad 0.9975 \quad}$$

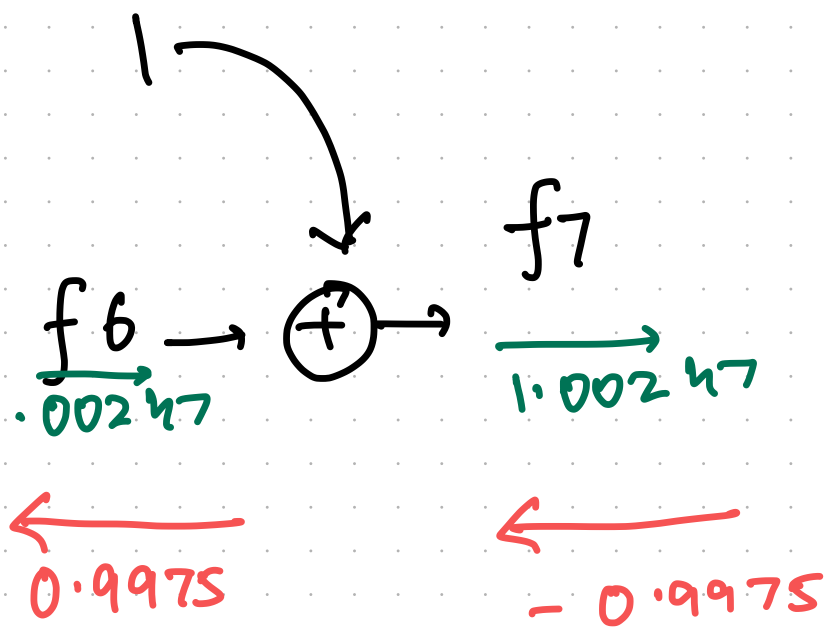
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



upstream grad. = 0.9975

local grad. = 1

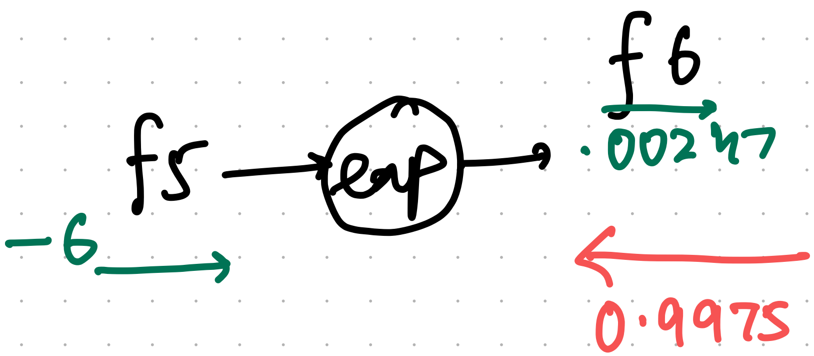
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



upstream grad. = 0.9975

local grad. = 1

$$\text{Loss} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}} \right)$$



upstream grad. = 0.9975

local grad. $= \dfrac{\partial f6}{\partial fs} = \dfrac{\partial}{\partial fs} e^{fs} = e^{fs} = e^{-6} = 0.0025$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$

$$f_5 \longrightarrow \text{exp} \longrightarrow f_6$$

f_5: -6 f_6: $.00247$

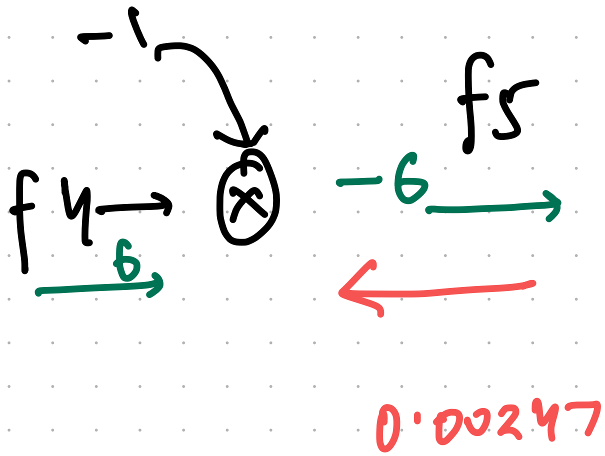
$\leftarrow 0.9975$

$\leftarrow .0025 * .9975 = 0.00247$

upstream grad. $= 0.9975$

local grad. $= \dfrac{\partial f_6}{\partial f_5} = \dfrac{\partial e^{f_5}}{\partial f_5} = e^{f_5} = e^{-6} = 0.0025$

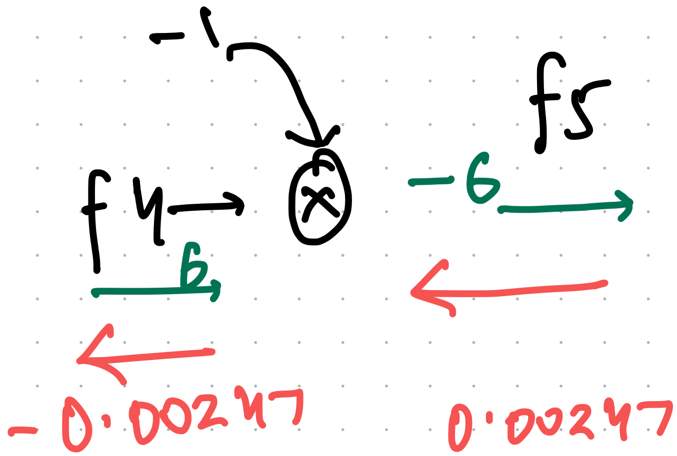
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



upstream grad. = 0.00247

local grad. = -1

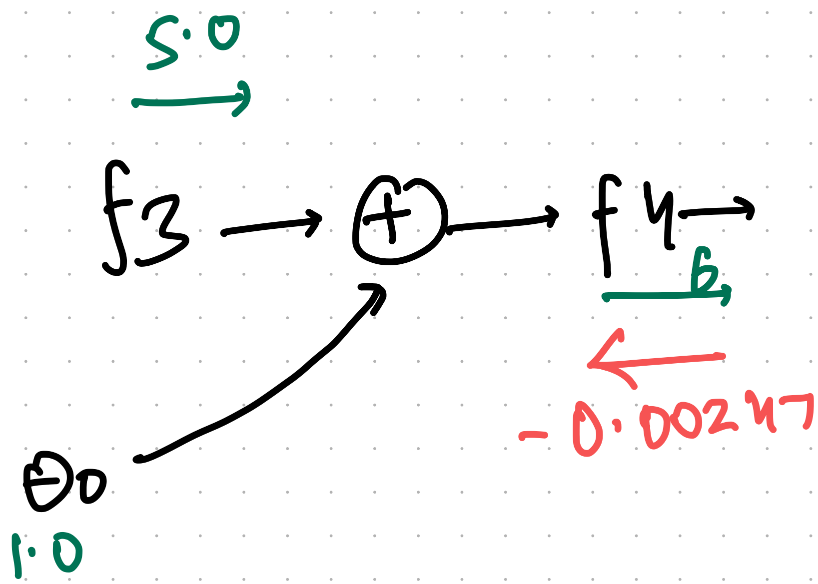
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



upstream grad. = 0.00247

local grad. = -1

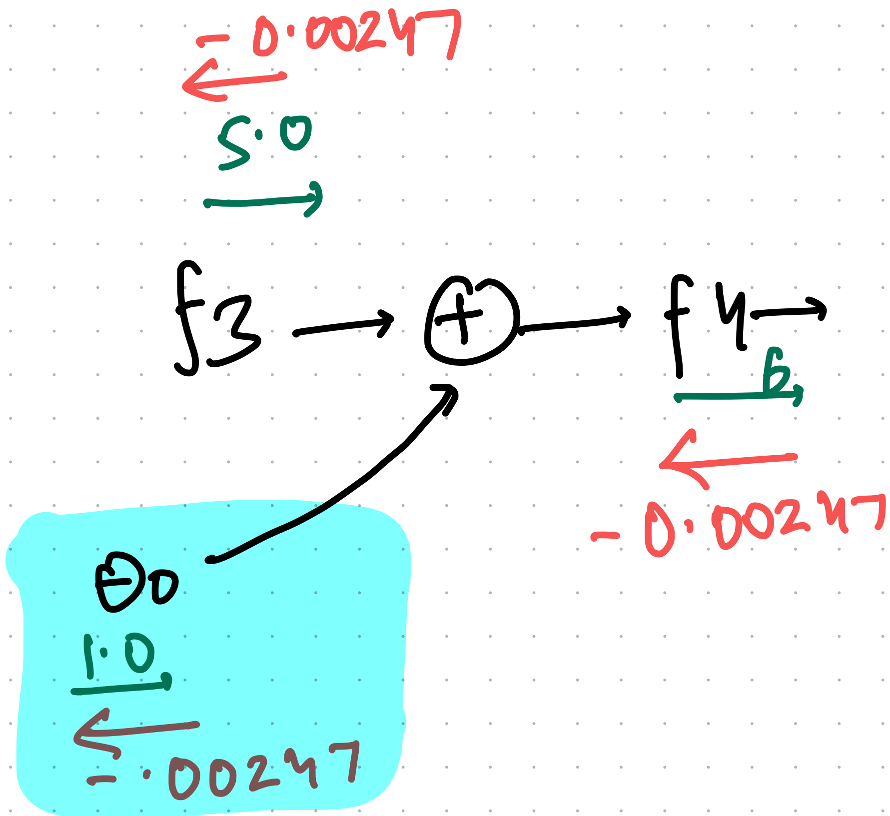
$$\text{LOSS} = -1*\log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



upstream grad. $= -0.00247$

local grad. $(\theta_0) = \frac{\partial f_4}{\partial \theta_0} = 1$; local grad for $f_3 = 1$

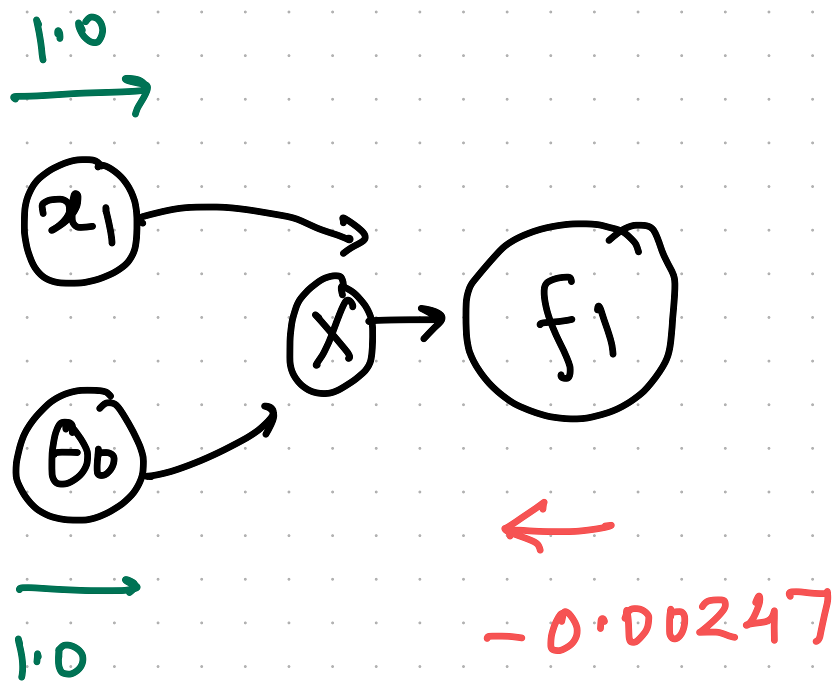
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



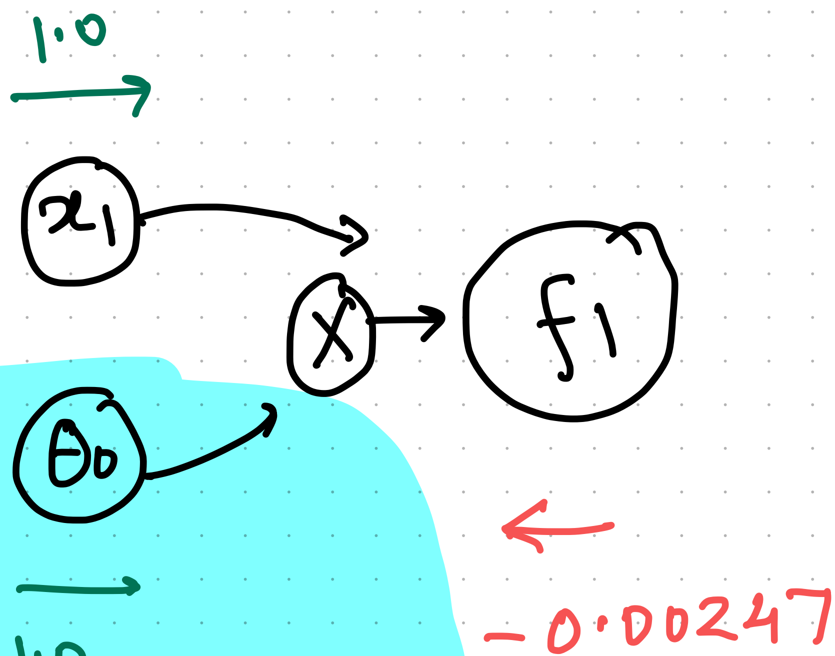
upstream grad. = −0.00247

local grad. $(\theta_0) = \frac{\partial f_4}{\partial \theta_0} = 1$; local grad for $f_3 = 1$

$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



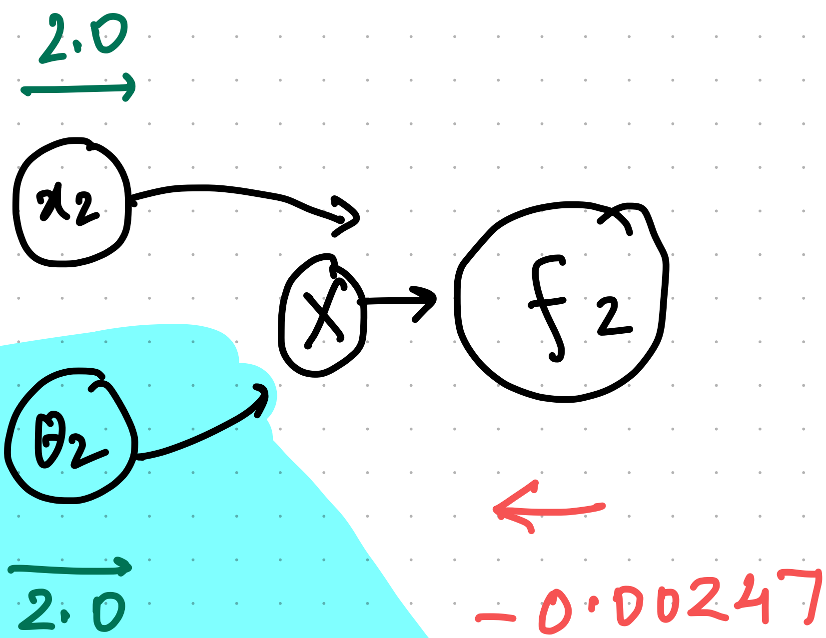
$$\text{LOSS} = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



$-0.00247 * x_1$
(1.0)

$= -.00247$

$$LOSS = -1 * \log\left(\frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2)}}\right)$$



$= -.00247 * x_2$ (2.0)

$= -.0049$

What autodiff library needs to know

(i) $f = a * b$; $\frac{\partial f}{\partial a} = b$; $\frac{\partial f}{\partial b} = a$

(ii) $f = a + b$; $\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b} = 1$

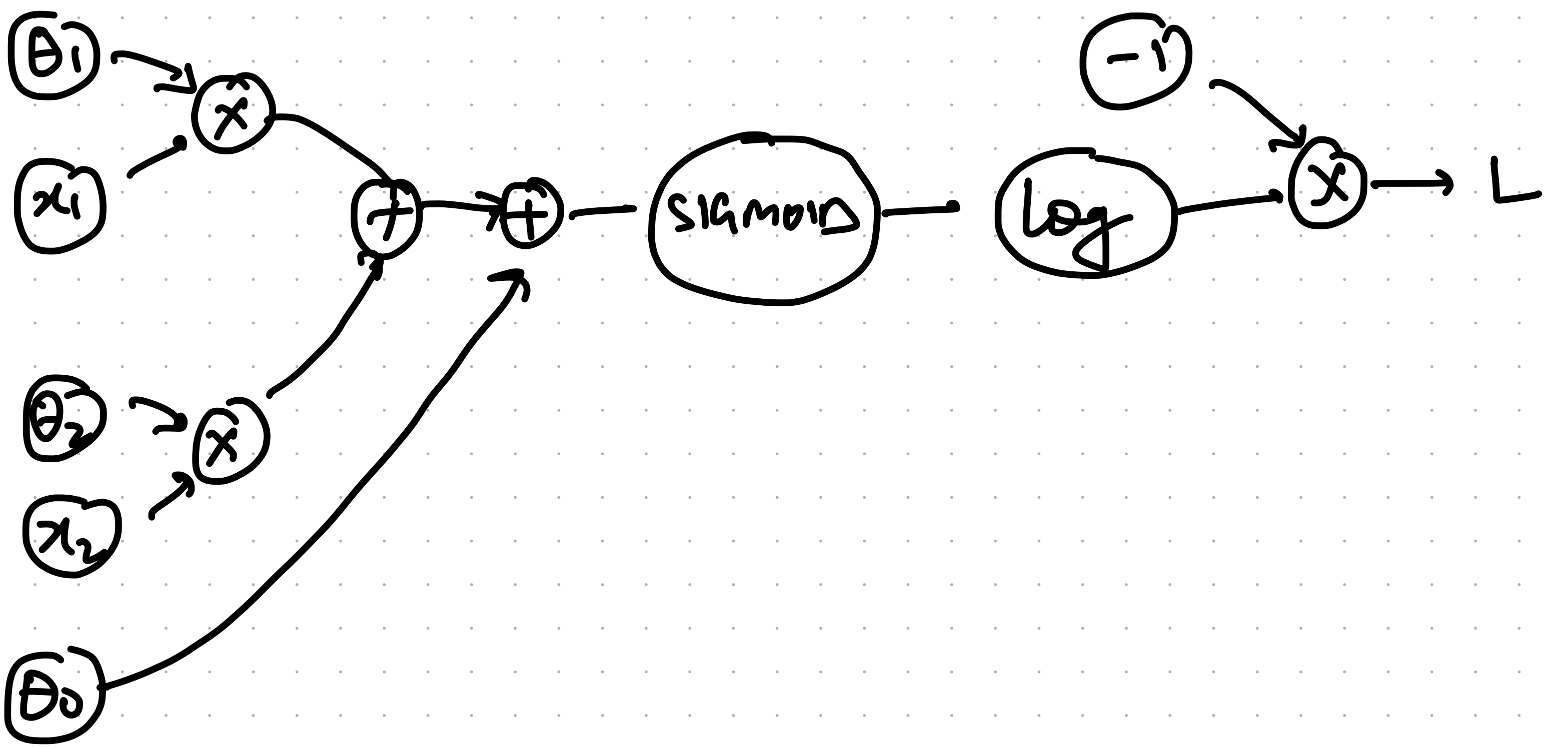
(iii) $f = e^{a}$; $\frac{\partial f}{\partial a} = e^{a}$

(iv) $f = \frac{1}{a}$; $\frac{\partial f}{\partial a} = -1/a^{2}$

$\vdots$

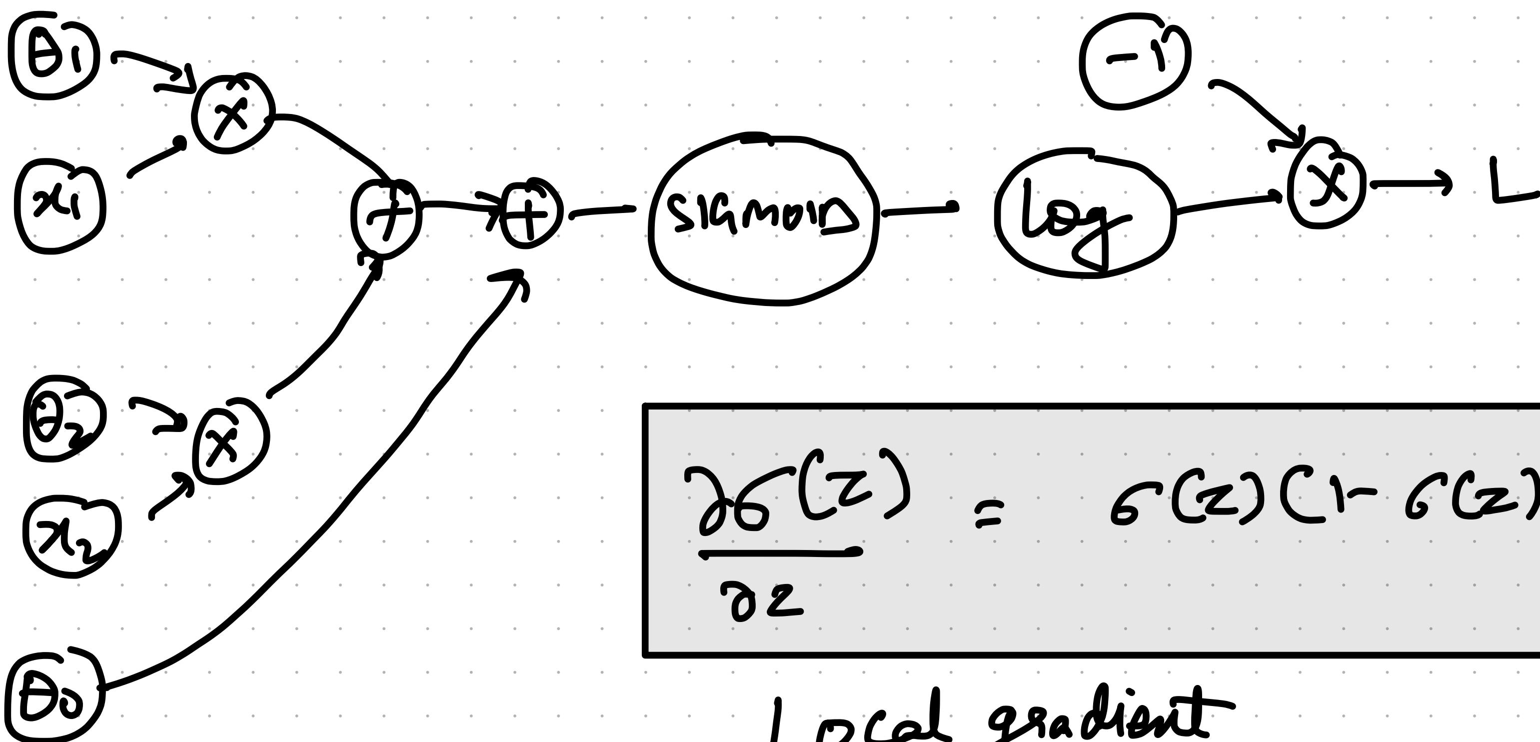
Simplifying computat^n graph

$$L = -1 * \log(\text{SIGMOID}(\theta_0 + \theta_1 x_1 + \theta_2 x))$$



* Simplifying computat^n graph

$$L = -1 * \log(\text{SIGMOID}(\theta_0 + \theta_1 x_1 + \theta_2 x))$$



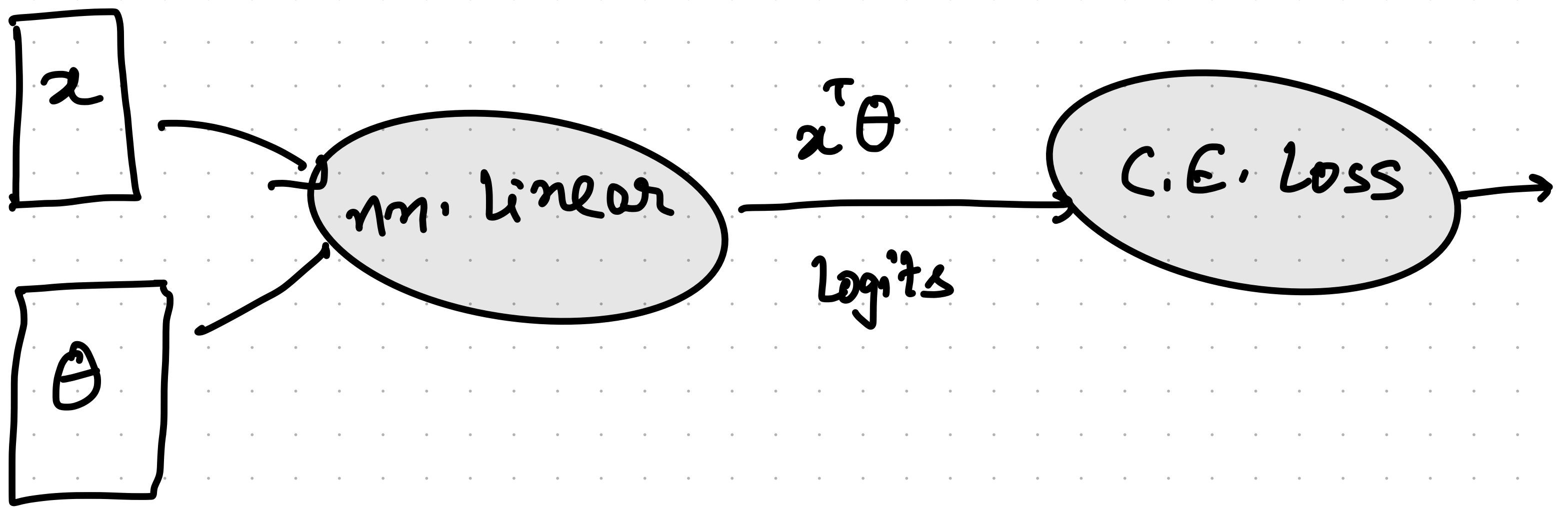
$$\boxed{\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z))}$$

Local gradient

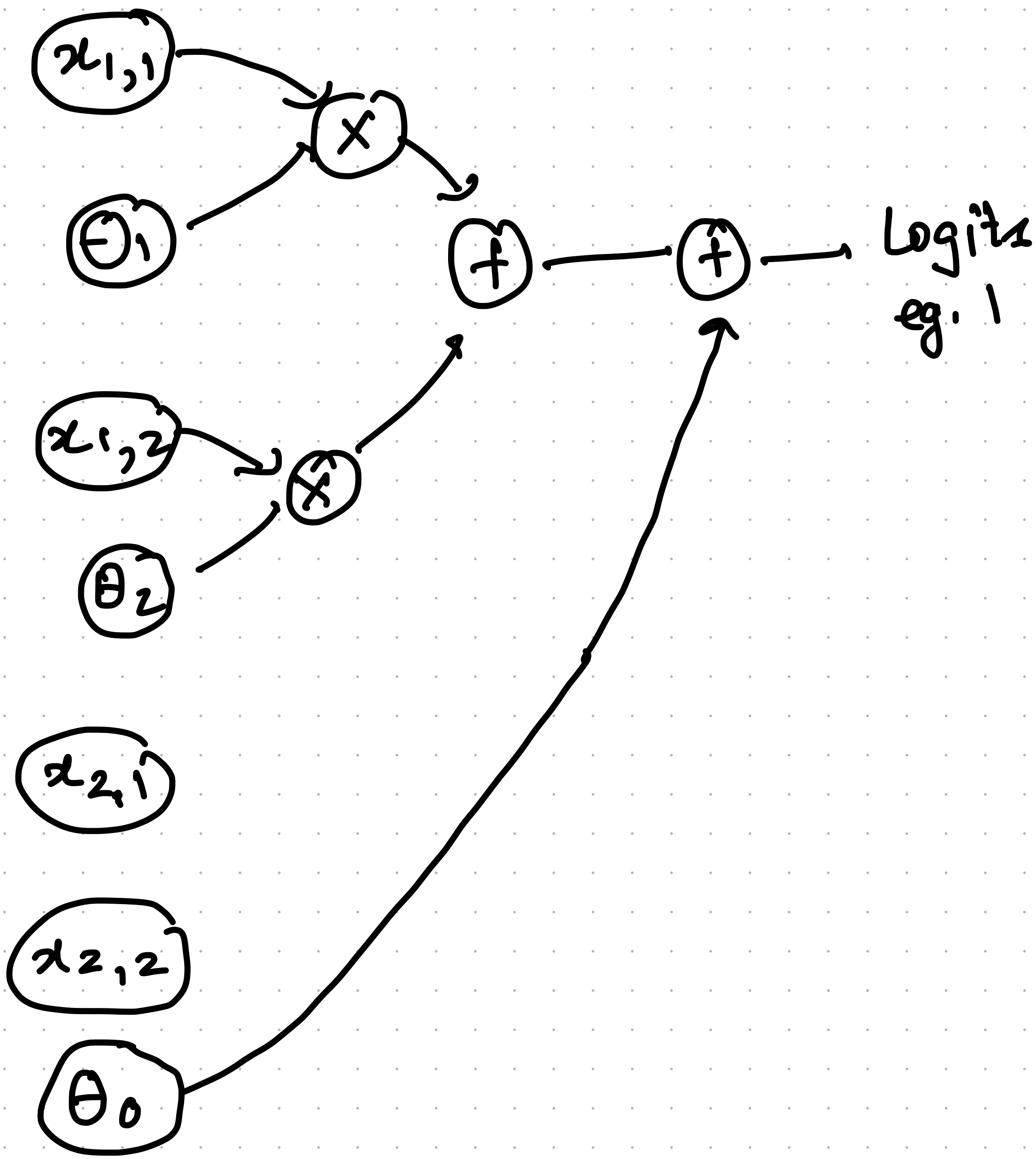
Exercise: show you get same answer as before

* Simplifying computat^n graph

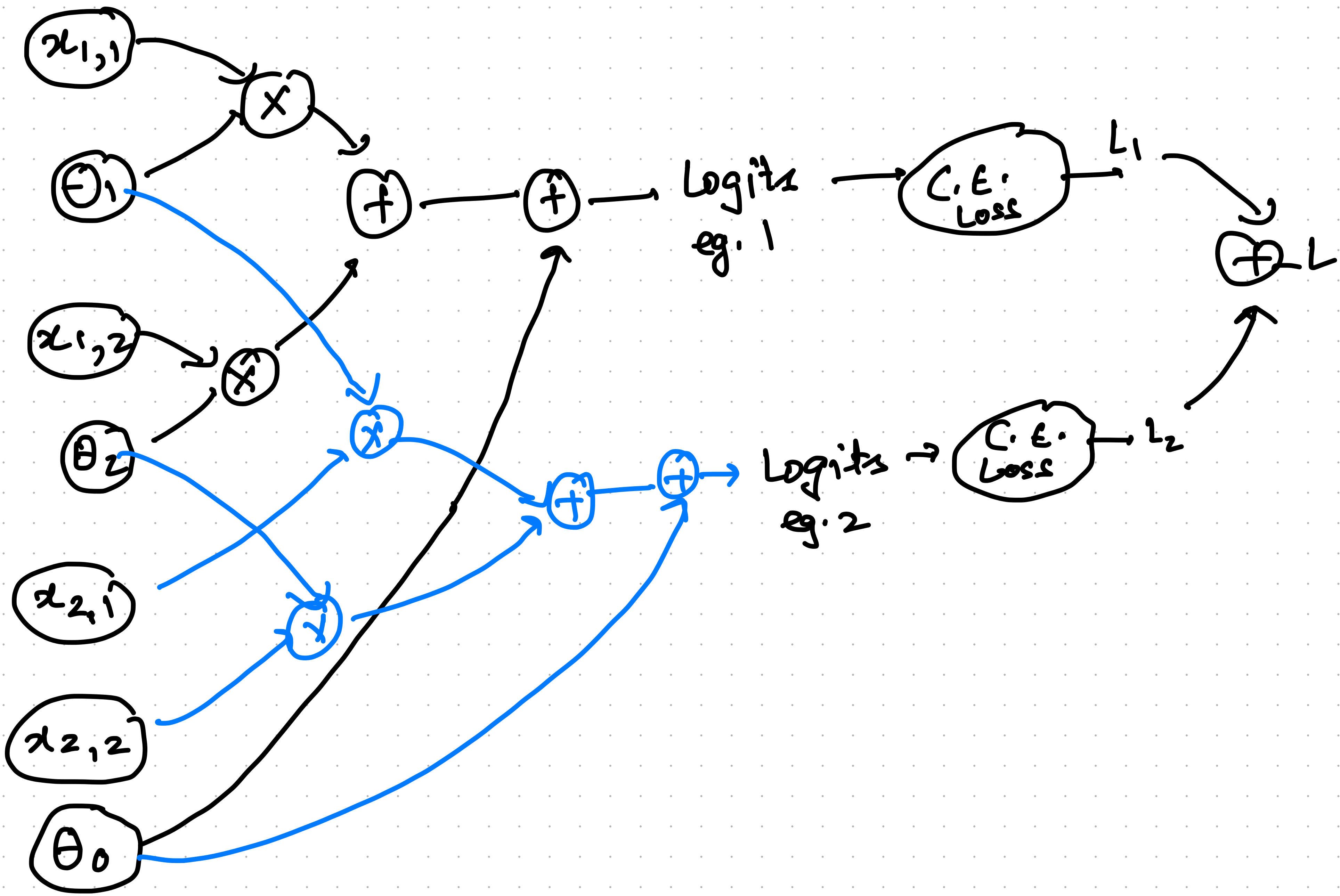
$$L = -1 * \log(\text{SIGMOID}(\theta_0 + \theta_1 x_1 + \theta_2 x))$$



* Training over N-examples



* Training over N- examples



* Training over N- examples

Chain Rule for One Independent Variable

Suppose that $x = g(t)$ and $y = h(t)$ are differentiable functions of t and $z = f(x, y)$ is a differentiable function of x and y. Then $z = f(x(t), y(t))$ is a differentiable function of t and

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial z}{\partial y} \cdot \frac{dy}{dt},$$

where the ordinary derivatives are evaluated at t and the partial derivatives are evaluated at (x, y).

* Training over N- examples

Chain Rule for One Independent Variable

Suppose that $x = g(t)$ and $y = h(t)$ are differentiable functions of t and $z = f(x, y)$ is a differentiable function of x and y. Then $z = f(x(t), y(t))$ is a differentiable function of t and

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial z}{\partial y} \cdot \frac{dy}{dt},$$

where the ordinary derivatives are evaluated at t and the partial derivatives are evaluated at (x, y).

$$L = L_1 + L_2$$

$$L_1 = x_1 \theta$$

$$L_2 = x_2 \theta$$

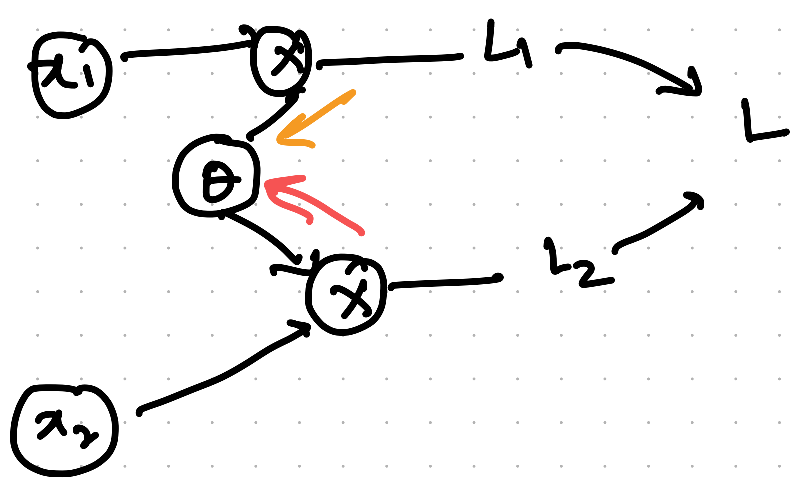
* Training over N- examples

> **Chain Rule for One Independent Variable**
>
> Suppose that $x = g(t)$ and $y = h(t)$ are differentiable functions of t and $z = f(x, y)$ is a differentiable function of x and y. Then $z = f(x(t), y(t))$ is a differentiable function of t and
>
> $$\frac{dz}{dt} = \frac{\partial z}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial z}{\partial y} \cdot \frac{dy}{dt},$$
>
> where the ordinary derivatives are evaluated at t and the partial derivatives are evaluated at (x, y).

$$L = L_1 + L_2$$

$$L_1 = x_1 \theta$$

$$L_2 = x_2 \theta$$



$$\frac{\partial L}{\partial \theta} = \rightarrow + \rightarrow$$

Addition of all incoming gradients