1. Collaborators

None

2. Activation Functions

As per Piazza note @162, this problem should be left empty (Instructors Krishanu Agarwal & Bowen Zuo)

3.1) Solve the optimization problem wrt w:

$$\underset{w}{\operatorname{argmin}} \underbrace{f(w^t) + \langle w - w^t, \nabla f(w^t) \rangle}_{\text{(I)}} + \underbrace{\frac{\lambda}{2} \|w - w^t\|^2}_{\text{(I)}}$$

↳ Convex problem. w^* reaches optimal when the gradient wrt $w = 0$

∴ Taking the derivatives wrt w:

(I) $\nabla f(w^t)$

(II) $\lambda \|w - w^t\|$

$$\Rightarrow \nabla f(w^t) + \lambda \|w - w^t\| = 0$$

↳ being minimized

$$\nabla f(w^t) + \lambda \|w^* - w^t\| = 0$$

$$\lambda \|w^* - w^t\| = -\nabla f(w^t)$$

$$\|w^* - w^t\| = -\frac{\nabla f(w^t)}{\lambda}$$

$$\boxed{w^* = w^t - \frac{\nabla f(w^t)}{\lambda}}$$

solution of the opt. problem

(III)

(III) Looks exactly like the GD update rule:

$w^{t+1} = w^t - \eta \nabla f(w^t)$, where η is a constant

Therefore, this tells me that if my loss is strictly convex, I will always achieve an optimal solution using norm 2 regularization when taking GD updates.

3.1 Continued

from the GD update rule: $w^{t+1} = w^t - \eta \nabla f(w^t)$

from (III): $w^* = w^t - \frac{\nabla f(w^t)}{\lambda}$

Assuming strictly convex loss function, we can say that w^* is the updated w, i.e., w^{t+1} as per my previous reasoning.

Therefore, $\eta = \frac{1}{\lambda}$, where η is the step size and λ is the penalization for the proximity term

This means that η & λ are inversely proportional, namely, if we want to heavily penalize the proximity trade-off, we need smaller step size; as well as low penalization calls for larger step size.

It is also true if we fix η and then adjust λ accordingly, but that is not what is most intuitive to me.

3.2) Prove: $\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^{\star}, \mathbf{v}_t\rangle \leq \frac{\|\mathbf{w}^{\star}\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\mathbf{v}_t\|^2$

Remember $v_t = \nabla f(w)$

$$\sum_{t=1}^{T}\langle w^t - w^*, v_t\rangle = \sum_{t=1}^{T}\langle w^t, v_t\rangle - \sum_{t=1}^{T}\langle w^*, v_t\rangle$$

w^t: open in GD

$$= \sum_{t=1}^{T}\langle -\eta\sum_{i} v_i, v_t\rangle - \langle w^*, -\frac{1}{\eta}w^{t+1}\rangle$$

$$= -\frac{\eta}{2}\Big\|\sum_{t=1}^{T} v_t\Big\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2 - \langle w^* - \frac{1}{\eta}w^{t+1}\rangle$$

$$= \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2 - \frac{1}{2\eta}\langle w^{t+1}, w^{t+1}\rangle - \langle w^* - \frac{1}{\eta}w^{t+1}\rangle$$

$$\leq \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|^2 + \frac{\|x^*\|^2}{2\eta}$$ ■

3.3) Show $f(\bar{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \nabla f(\mathbf{w}^{(t)})\rangle$ for $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

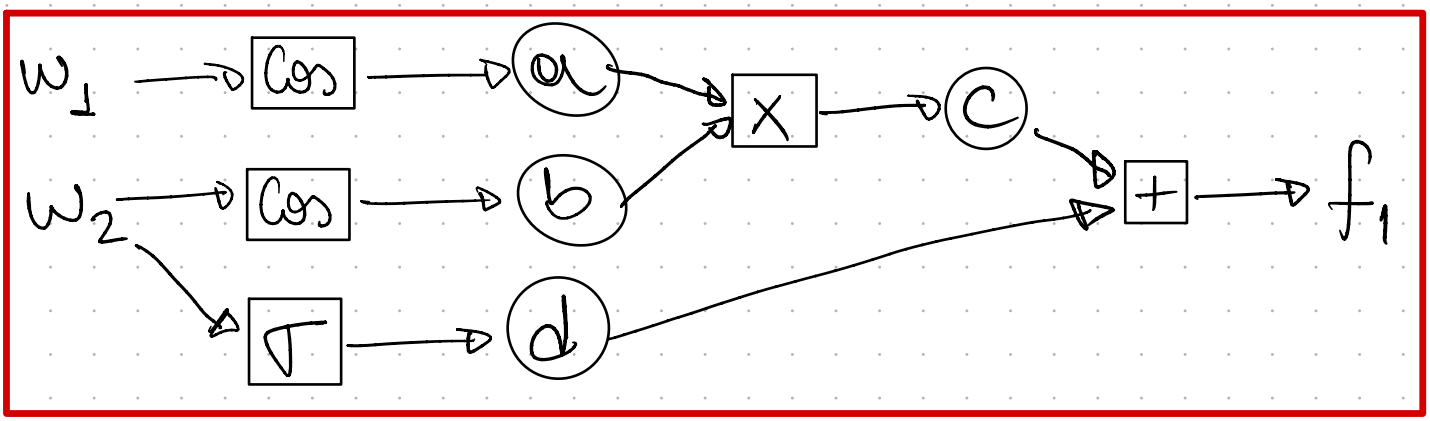
$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{T}\sum_{t=1}^{T} w^t\right) - f(w^*)$$

4. a)

$$f_1(w_1, w_2) = \cos(w_1)\cos(w_2) + \sigma(w_2)$$
$$f_2(w_1, w_2) = ln(w_1 + w_2) + w_1^2 w_2$$

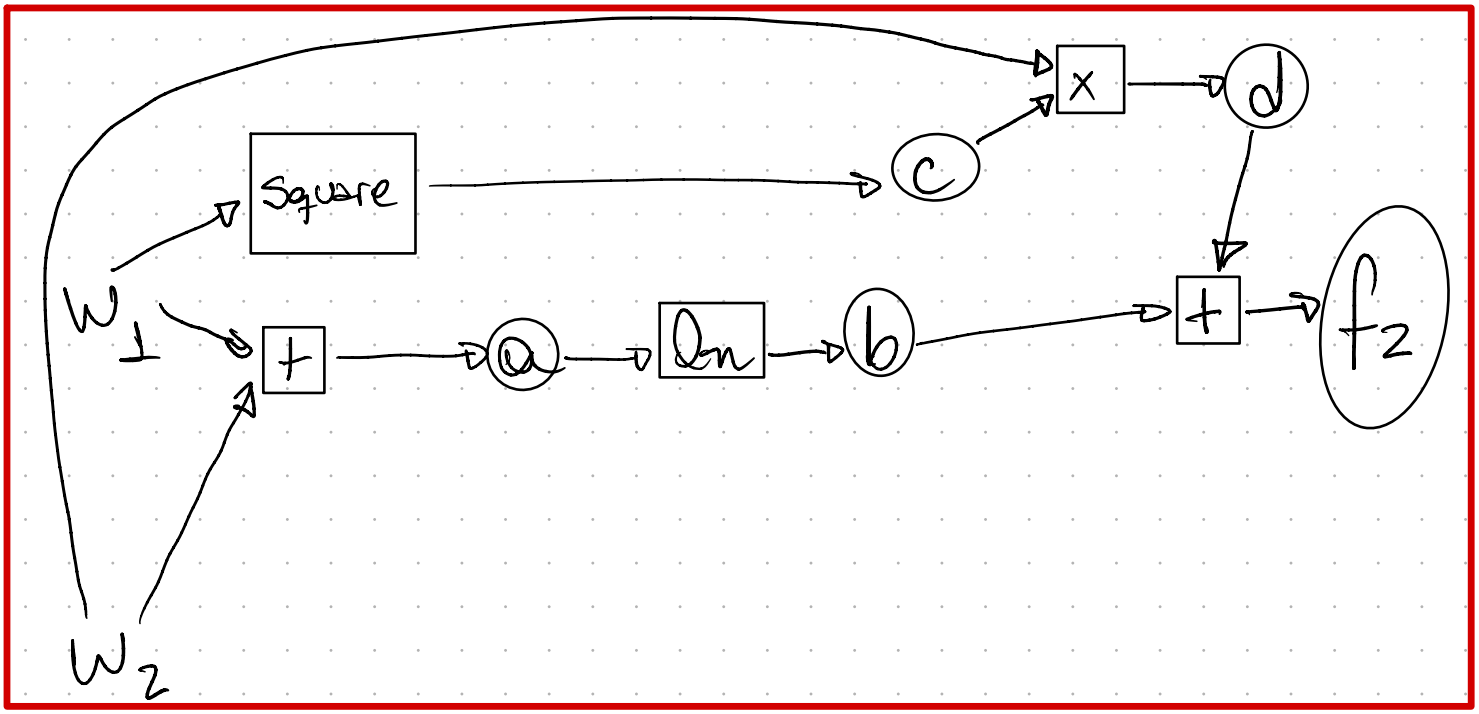
Computation graphs:

f_1: $a = \cos(w_1)$, $b = \cos(w_2)$, $c = a \cdot b$, $d = \sigma(w_2)$



@ $w = (1, 2)$: $f_1(1, 2) = \cos(1)\cos(2) + \sigma(2) \approx$ **0.655952**

f_2: $a = w_1 + w_2$, $b = \ln(a)$, $c = w_1^2$, $d = c \cdot w_2$



@ $w = (1, 2)$: $f_2(1, 2) = \ln(1 + 2) + 1^2 \cdot 2 \approx$ **3.098612**

4.b) $f_1(w_1, w_2) = \cos(w_1)\cos(w_2) + \sigma(w_2)$

$f_2(w_1, w_2) = ln(w_1 + w_2) + w_1^2 w_2$

@ $W = (1,2)$ w/ $\Delta W = 0.01$

f_1: $\frac{\partial f_1}{\partial w_1} = \frac{f_1(1.01,2) - f_1(1,2)}{0.01} = \frac{(0.659465 - 0.655952)}{0.01}$

$= 0.3513$

Similarly: $\frac{\partial f_1}{\partial w_2} = \frac{f_1(1,2.01) - f_1(1,2)}{0.01} = -0.3856$

f_2: $\frac{\partial f_2}{\partial w_1} = \frac{f_2(1.01,2) - f_2(1,2)}{0.01} = 4.3528$

$\frac{\partial f_2}{\partial w_2} = \frac{f_2(1,2.01) - f_2(1,2)}{0.01} = 1.3328$

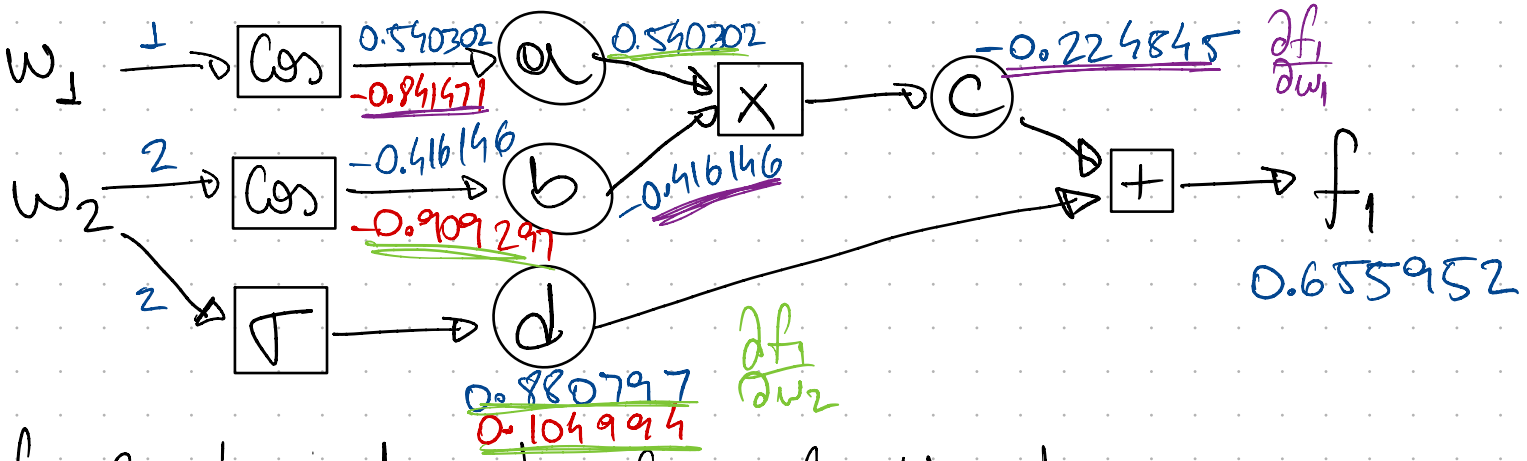
Concisely: $\frac{\partial \vec{f}}{\partial \vec{w}} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} \end{bmatrix} = \begin{bmatrix} 0.3513 & -0.3856 \\ 4.3528 & 1.3328 \end{bmatrix}$

4.c)

$$f_1(w_1, w_2) = \cos(w_1)\cos(w_2) + \sigma(w_2)$$
$$f_2(w_1, w_2) = ln(w_1 + w_2) + w_1^2 w_2$$

fwd:

$a = \cos(w_1)$, $b = \cos(w_2)$, $c = a \cdot b$, $d = \sigma(w_2)$



$f = c + d = ab + d = \cos w_1 \cos w_2 + d$

$\frac{\partial c}{\partial a} = b$; $\frac{\partial c}{\partial b} = a$; $\frac{\partial a}{\partial w_1} = -0.841471$; $\frac{\partial b}{\partial w_2} = -0.909297$

$\frac{\partial d}{\partial w_2} = (1 - \sigma(w_2))\sigma(w_2) = 0.104994$

fwd:

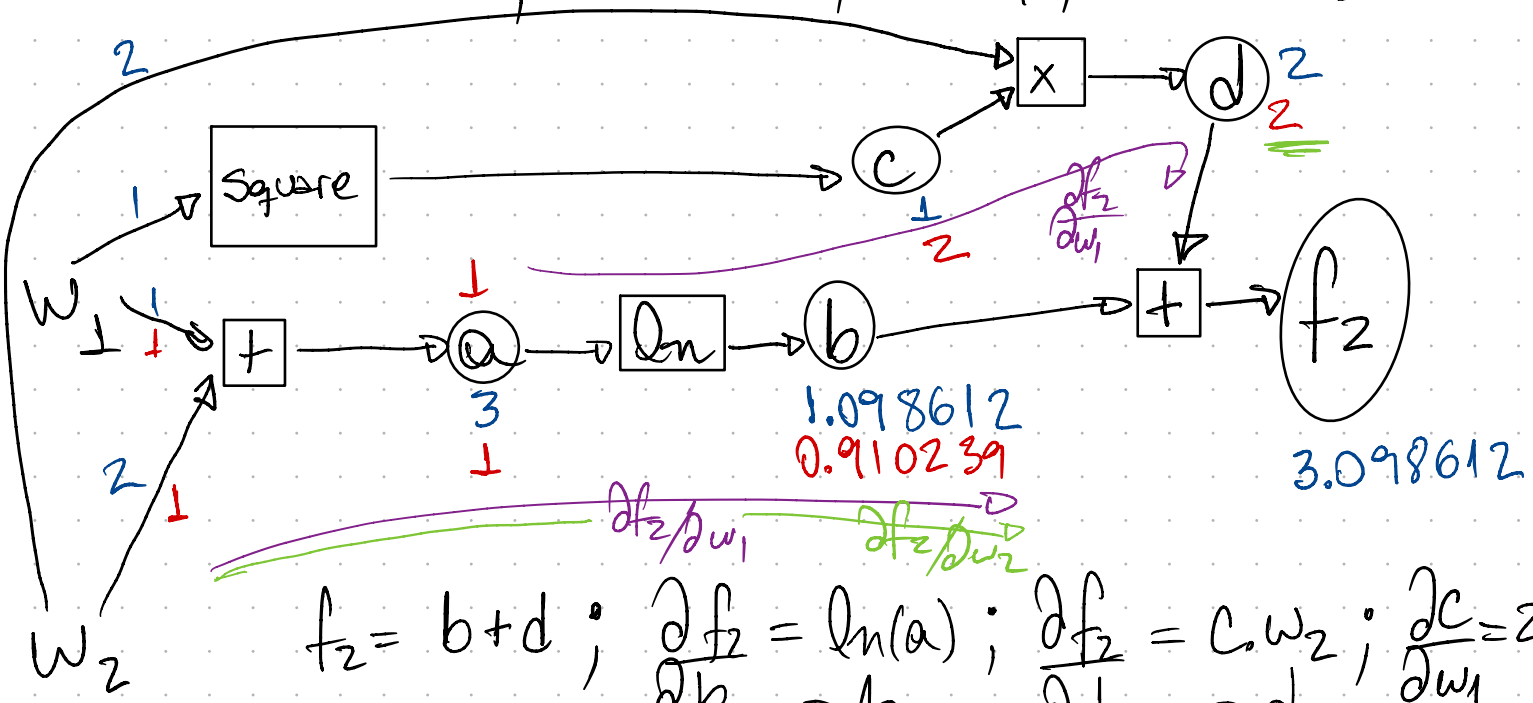
$$\frac{\partial f_1}{\partial w_1} = \frac{\partial a}{\partial w_1} \cdot \frac{\partial c}{\partial a} \cdot \frac{\partial f}{\partial c} = \boxed{-0.078735}$$

$$\frac{\partial f_1}{\partial w_2} = \frac{\partial b}{\partial w_2} \cdot \frac{\partial b}{\partial d} \cdot \frac{\partial c}{\partial b} \cdot \frac{\partial f}{\partial d} = \boxed{-0.045434}$$

4.c) continued

$$f_1(w_1, w_2) = \cos(w_1)\cos(w_2) + \sigma(w_2)$$
$$f_2(w_1, w_2) = ln(w_1 + w_2) + w_1^2 w_2$$

$a = w_1 + w_2$, $b = \ln(a)$, $c = w_1^2$, $d = c \cdot w_2$



$f_2 = b + d$; $\frac{\partial f_2}{\partial b} = 1$; $\frac{\partial f_2}{\partial d} = 1$; $\frac{\partial c}{\partial w_1} = 2w_1$

$\frac{\partial b}{\partial a} = \frac{1}{a}$; $\frac{\partial a}{\partial w_1} = 1$; $\frac{\partial a}{\partial w_2} = 1$; $\frac{\partial d}{\partial w_2} = c$; $\frac{\partial d}{\partial c} = w_2$

fwd:

$$\frac{\partial f_2}{\partial w_1} = \frac{\partial a}{\partial w_1} \cdot \frac{\partial b}{\partial a} \cdot \frac{\partial f}{\partial b} + \frac{\partial c}{\partial w_1} \cdot \frac{\partial d}{\partial c} \cdot \frac{\partial f}{\partial d} = 0.910239 + 4 = 4.910239$$

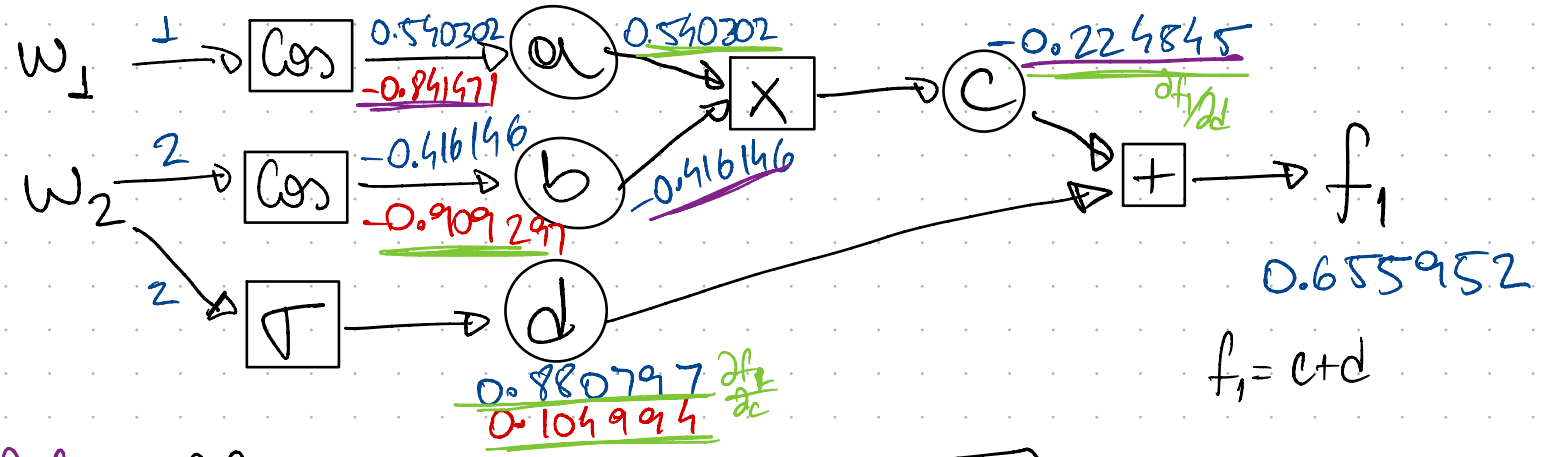
$$\frac{\partial f_2}{\partial w_2} = \frac{\partial a}{\partial w_2} \cdot \frac{\partial b}{\partial a} \cdot \frac{\partial f}{\partial b} + \frac{\partial d}{\partial w_2} \cdot \frac{\partial f}{\partial d} = 0.910239 + 2 = 2.910239$$

$$\nabla f = \begin{bmatrix} -0.078735 & -0.045434 \\ 4.910239 & 2.910239 \end{bmatrix}$$

4.d)

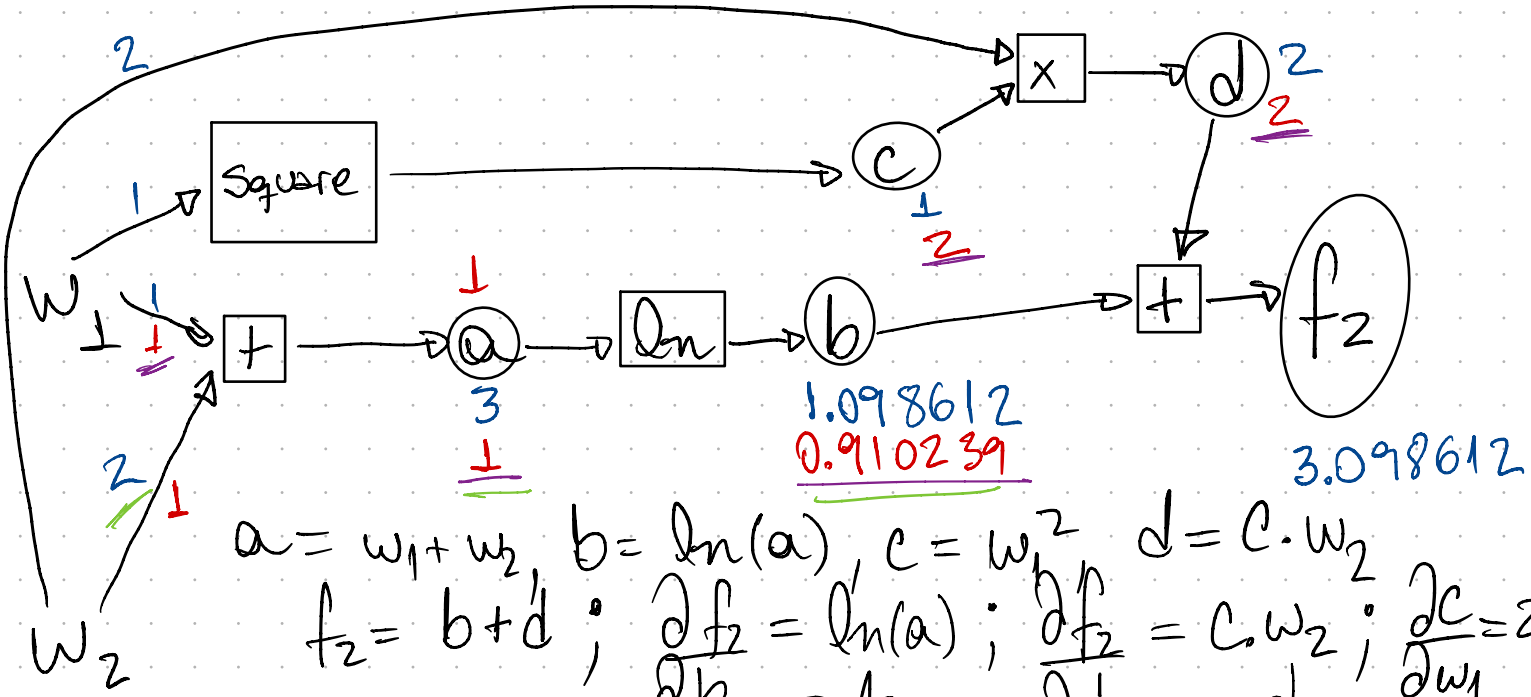
$$f_1(w_1, w_2) = \cos(w_1)\cos(w_2) + \sigma(w_2)$$
$$f_2(w_1, w_2) = ln(w_1 + w_2) + w_1^2 w_2$$

$a = \cos(w_1)$, $b = \cos(w_2)$, $c = a \cdot b$, $d = \sqrt{(w_2)}$



$$\frac{\partial f_1}{\partial w_1} = \frac{\partial f_1}{\partial c} \cdot \frac{\partial c}{\partial a} \cdot \frac{\partial a}{\partial w_1} = -0.078735$$

$$\frac{\partial f_1}{\partial w_2} = \underbrace{\frac{\partial f_1}{\partial d}}_{=c} \cdot \frac{\partial d}{\partial w_2} + \frac{\partial f_1}{\partial c} \cdot \frac{\partial c}{\partial b} \cdot \frac{\partial b}{\partial w_2} = -0.023607 - 0.432731 = -0.456338$$



$a = w_1 + w_2$, $b = ln(a)$, $c = w_1^2$, $d = c \cdot w_2$

$f_2 = b + d$; $\frac{\partial f_2}{\partial b} = b$; $\frac{\partial f_2}{\partial d} = d$; $\frac{\partial c}{\partial w_1} = 2w_1$

$$\frac{\partial f_2}{\partial w_1} = \frac{\partial f_2}{\partial b} \cdot \frac{\partial b}{\partial a} \cdot \frac{\partial a}{\partial w_1} + \frac{\partial f_2}{\partial d} \cdot \frac{\partial d}{\partial c} \cdot \underbrace{\frac{\partial c}{\partial w_1}}_{=2w_1=2} = 1.820178 + 4 = 5.820478$$

$$\frac{\partial f_2}{\partial w_2} = \frac{\partial f_2}{\partial b} \cdot \frac{\partial b}{\partial a} \cdot \frac{\partial a}{\partial w_2} + \frac{\partial f_2}{\partial d} \cdot \frac{\partial d}{\partial c} \cdot \underbrace{\frac{\partial c}{\partial w_2}}_{=2 \cdot w_2 = 4} = 3.640956 + 4 = 7.820478$$

4.d) continued

$$\nabla f = \begin{bmatrix} -0.078735 & -0.456338 \\ 5.820478 & 7.820478 \end{bmatrix}$$

4.e) YES!

5.a) Prove: any circulant matrix is commutative with a shift matrix.

Let S be a shift matrix $\Rightarrow$ S is a circulant matrix (it's given that shift is a special case of circulants).

Let C be a circulant matrix

$$C = \sum_{i=0}^{n-1} c_i \cdot J_i$$ (from the definition of circulant matrices, Pado Zellini)

$$\therefore S = \sum_{j=0}^{n-1} s_j J_j$$

If $CS = SC$ (commutative property):

$$CS - SC = \sum_{i,j}^{n-1} c_i s_j (J_i J_j - J_j J_i) = 0 \checkmark$$

$\therefore$ Convolutions are commutative with shift operators (shift equivariants) ■

5.b) In the same line of thought, let B be a linear opperator that's not a convolution $\Rightarrow B \neq \sum_{i=0}^{n-1} b_i J_i$

Assume B and S are commutative:

$$BS - SB = 0$$

$$\sum_{i,j}^{n-1} b_i s_j (J_i J_j - J_j J_i) \neq 0 \quad \left(B \neq \sum_{i=0}^{n-1} b_i J_i\right)$$

Contradiction

∴ The statement is NOT Bidirectional, and Convolution is the only linear operator with shift equivariance ■

5.c) Convolutions are the only linear operations that is guaranteed to preserve the spatio-temporal aspects/nature of any input, therefore they make the most sense to be present in any DL architecture that deals with data that contains such structure/nature.

Of course other architectures can be tried, but nothing else (operations-wise) guarantees that the spatio-temporality will be kept/maintained.

6) No, SGD (stochastic gradient descent) is NOT guaranteed to decrease the OVERALL LOSS AT EVERY ITERATION.

Counter example: Overall Loss: $f(w) = \underbrace{\frac{1}{2}(w-2)^2}_{a} + \underbrace{\frac{1}{2}(w+1)^2}_{b}$

$f(w)$ is a parabola with 2 Real and 2 complex roots

Consider at $w^t = 0$. the algorithm has 2 choices:

1) pick a
2) pick b

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

If (2) $\Rightarrow$ $\nabla f^b(w) = \frac{\partial}{\partial w}\frac{1}{2}(w+1)^2 = \frac{\partial}{\partial w}\frac{1}{2}(w^2+2w+1)$

$$= \frac{1}{2}(2w+2) = w+1$$

$$w^{t+1} = w^t - \eta \nabla f^b(w)$$

$$w^{t+1} = 0 - \eta(w^t+1) = -\eta(0+1)$$

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

As $\eta > 0 \Rightarrow w^{t+1} > 0$

$$w^{t+1} > w^t$$

■