# 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:

$$\arg\min_{w} \underbrace{f(w^t) + \langle w-w^t, \nabla f(w^t)\rangle}_{\mathrm{I}} + \underbrace{\frac{\lambda}{2}\|w-w^t\|^2}_{\mathrm{II}}$$

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

$\therefore$ Taking the derivatives wrt $w$:

I $\nabla f(w^t)$

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

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

$\hookrightarrow$ 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 $\eta$ 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, $\eta = \frac{1}{\lambda}$, wher $\eta$ is the step size and $\lambda$ is the penalization for the proximity term

This means that $\eta$ & $\lambda$ 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.

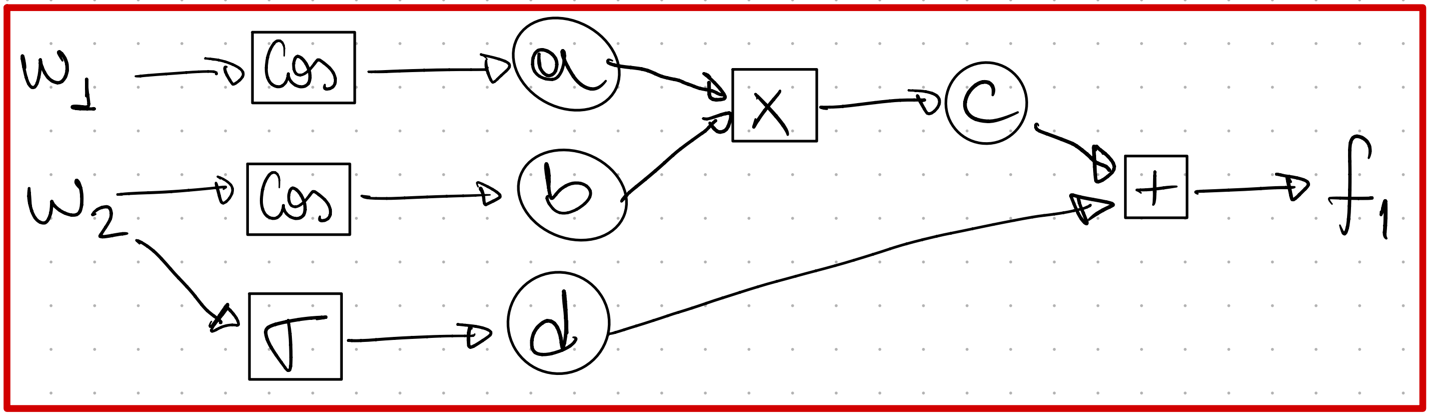
It is also true if we fix $\eta$ and then adjust $\lambda$ accordingly, but that is not what is most intuitive to me.

# 4. a)

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

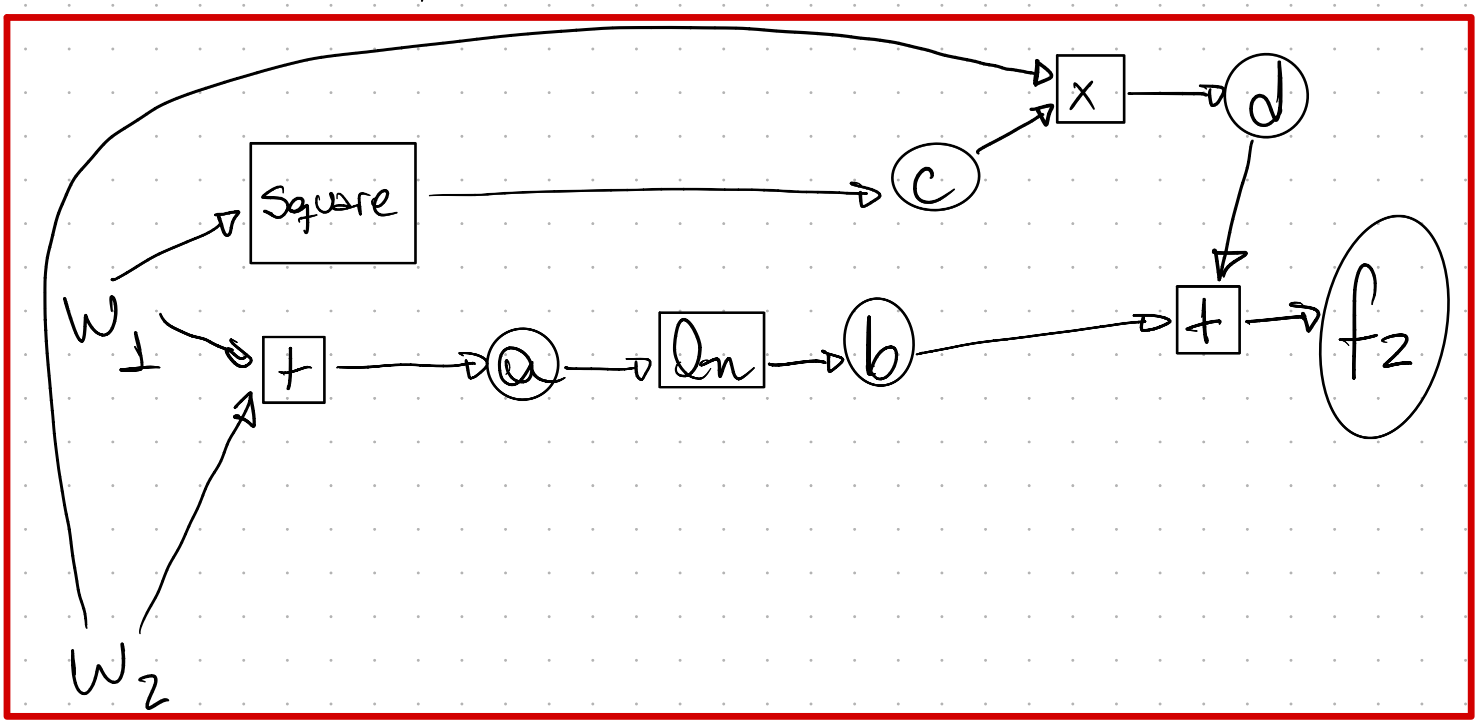
## Computation graphs:

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




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

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




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

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

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

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

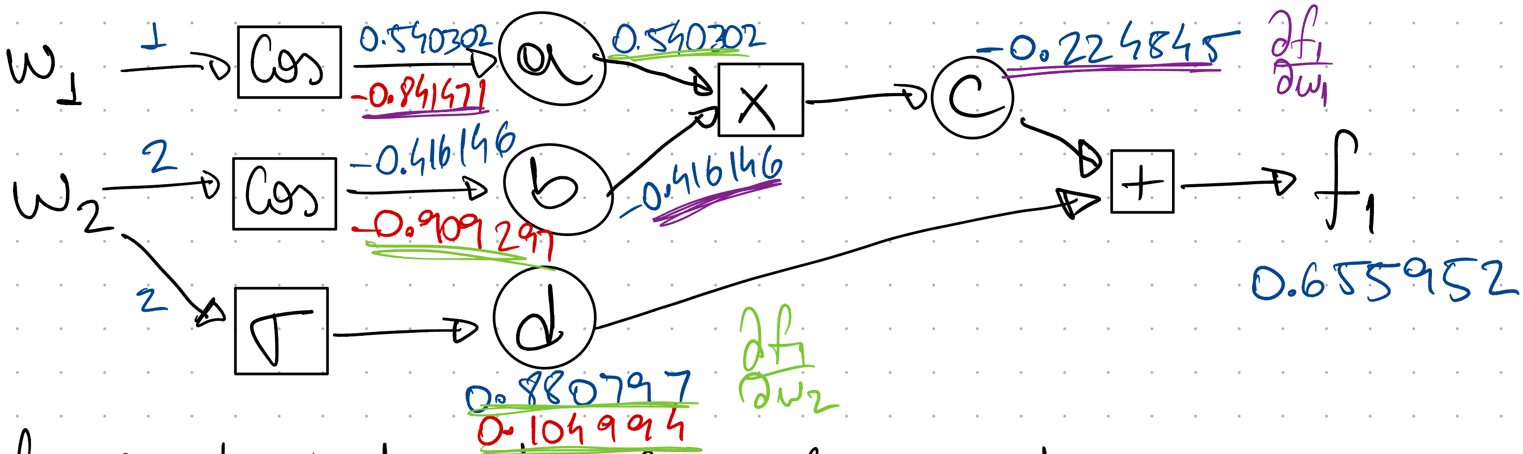
Concisely: $\frac{\partial \vec{f}}{\partial \vec{W}} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} \end{bmatrix} = \boxed{\begin{bmatrix} 0.3513 & -0.3856 \\ 4.3528 & 1.3328 \end{bmatrix}}$

## 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:

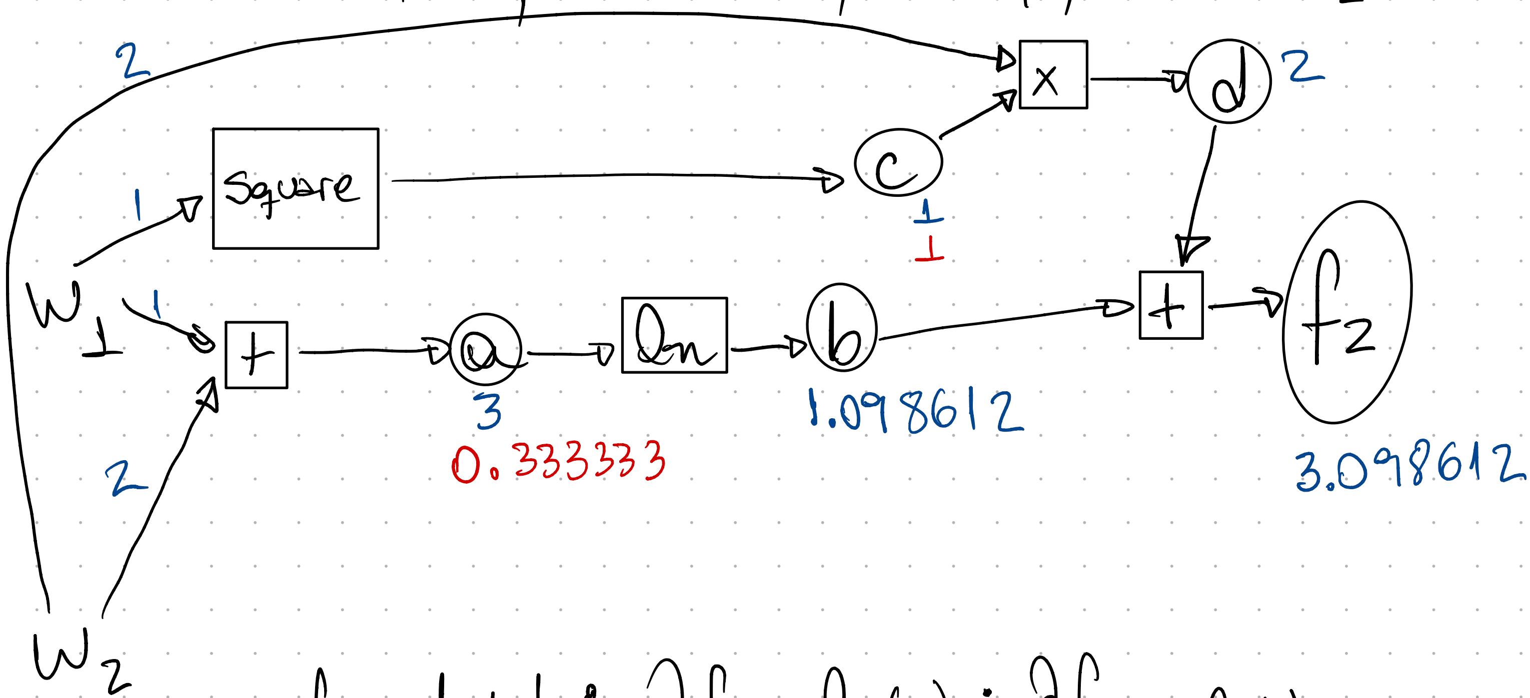
$$\frac{\partial f_1}{\partial w_1} = \frac{\partial a}{\partial w_1} \cdot \frac{\partial c}{\partial a} \cdot \frac{\partial f}{\partial c} = -0.078735$$

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

# 4.c) continued

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

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




$f_2 = b + d$; $\frac{\partial f_2}{\partial b} = \ln(a) = b$; $\frac{\partial f_2}{\partial d} = c \cdot w_2 = d$

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

fwd:

$$\frac{\partial f_2}{\partial w_1} = \frac{\partial a}{\partial w_1} \cdot \frac{\partial b}{\partial a} \cdot \frac{\partial d}{\partial c} \cdot \frac{\partial c}{\partial b}$$