

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 :

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

↳ 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

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

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

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

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

solution of the opt. problem
III

III Looks exactly like the GD update rule:

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

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 $\eta \propto \lambda$ 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 η and then adjust λ accordingly, but that is not what is most intuitive to me.

3.2) Prove: $\sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2$ ④ Remember
 $\nabla f = \nabla f(\mathbf{w})$

$$\sum_{t=1}^T \langle \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t \rangle = \underbrace{\sum_{t=1}^T \langle \mathbf{w}^t, \mathbf{v}_t \rangle}_{\text{Open in } \mathbb{R}^D} - \underbrace{\sum_{t=1}^T \langle \mathbf{w}^*, \mathbf{v}_t \rangle}_{\nearrow}$$

$$= \sum_{t=1}^T \underbrace{\left\langle -\eta \sum_i \mathbf{v}_i, \mathbf{v}_t \right\rangle}_{-\eta \sum_i \mathbf{v}_i} - \langle \mathbf{w}^*, -\frac{1}{\eta} \mathbf{w}^{t+1} \rangle$$

$$= -\frac{\eta}{2} \underbrace{\left\| \sum_{t=1}^T \mathbf{v}_t \right\|^2}_{\sum_{t=1}^T \mathbf{v}_t} + \frac{\eta}{2} \underbrace{\sum_{t=1}^T \|\mathbf{v}_t\|^2}_{\sum_{t=1}^T \mathbf{v}_t} - \langle \mathbf{w}^* - \frac{1}{\eta} \mathbf{w}^{t+1} \rangle$$

$$= \frac{\eta}{2} \underbrace{\sum_{t=1}^T \|\mathbf{v}_t\|^2}_{\sum_{t=1}^T \mathbf{v}_t} - \frac{1}{2\eta} \underbrace{\langle \mathbf{w}^{t+1}, \mathbf{w}^{t+1} \rangle}_{\langle \mathbf{w}^* - \frac{1}{\eta} \mathbf{w}^{t+1}, \mathbf{w}^{t+1} \rangle} - \langle \mathbf{w}^* - \frac{1}{\eta} \mathbf{w}^{t+1} \rangle$$

$$\leq \frac{\eta}{2} \underbrace{\sum_{t=1}^T \|\mathbf{v}_t\|^2}_{\sum_{t=1}^T \mathbf{v}_t} + \frac{\|\mathbf{x}^*\|^2}{2\eta} \quad \blacksquare$$

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

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

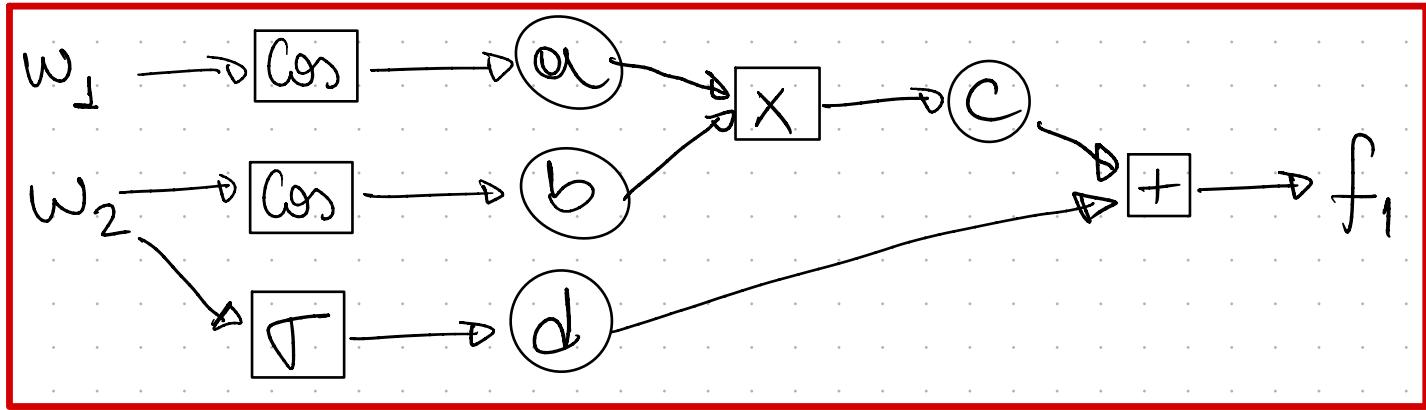
4.a)

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

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

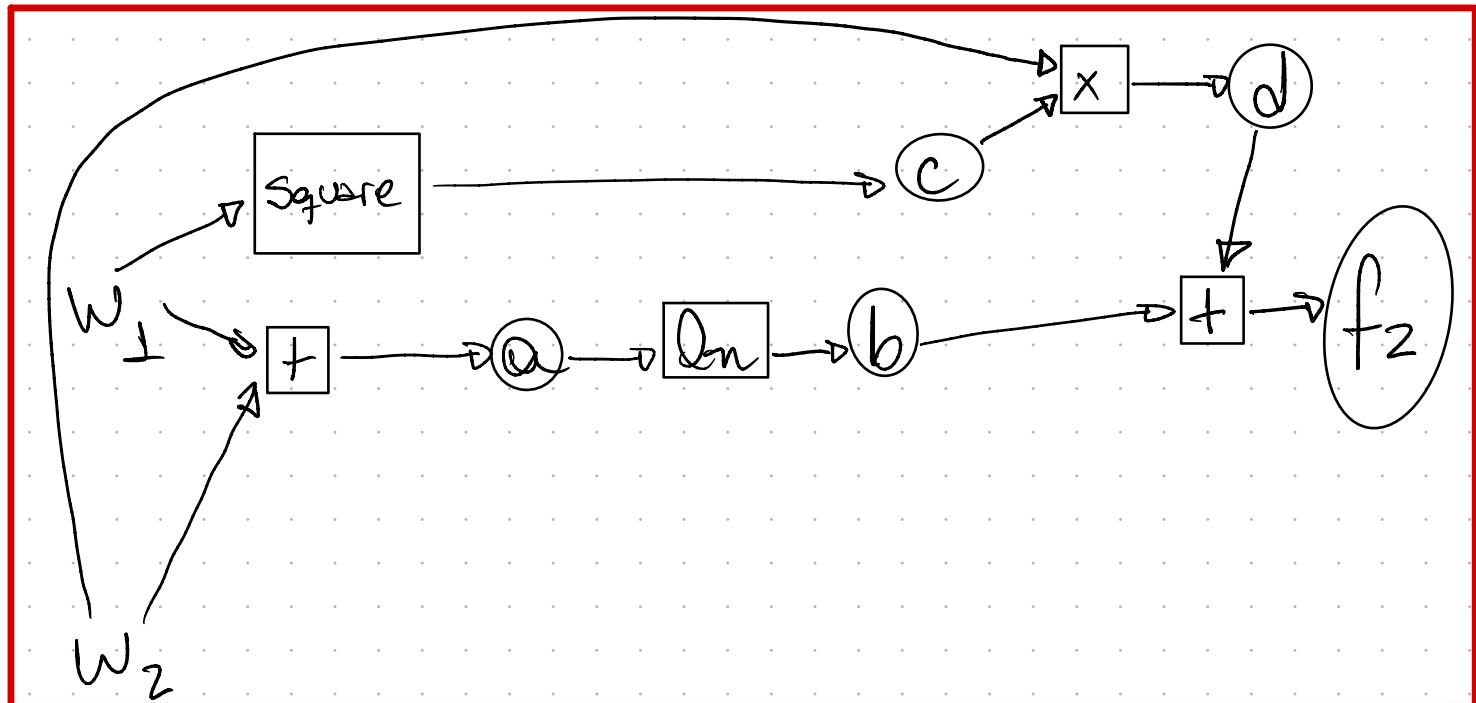
Computation graphs:

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



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

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



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

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

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

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

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

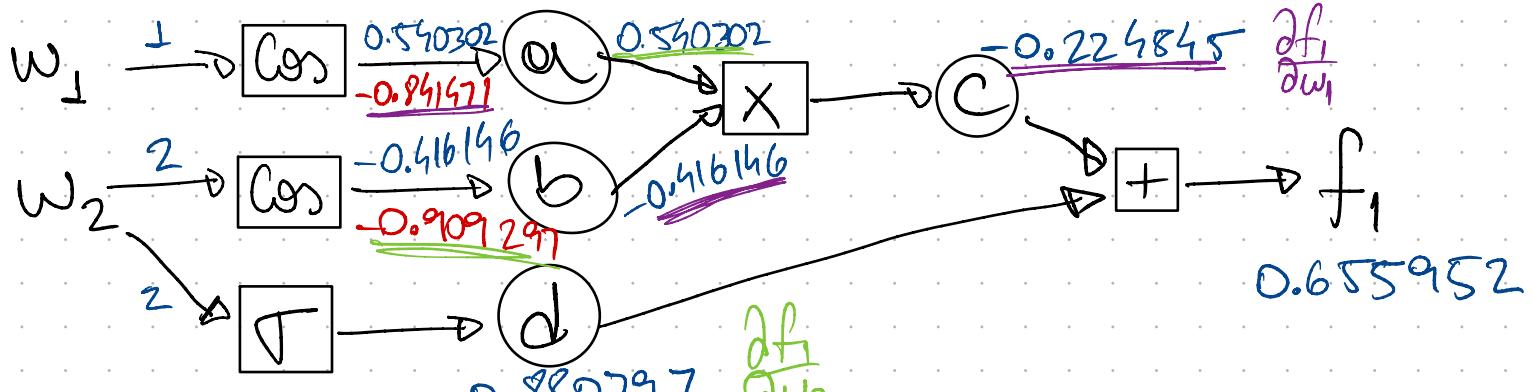
4.c)

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

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

fwd.:

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



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

$$\frac{\partial f}{\partial a} = b; \frac{\partial f}{\partial b} = a; \frac{\partial a}{\partial w_1} = -0.841471; \frac{\partial b}{\partial w_2} = -0.9091297$$

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

fwd.:

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

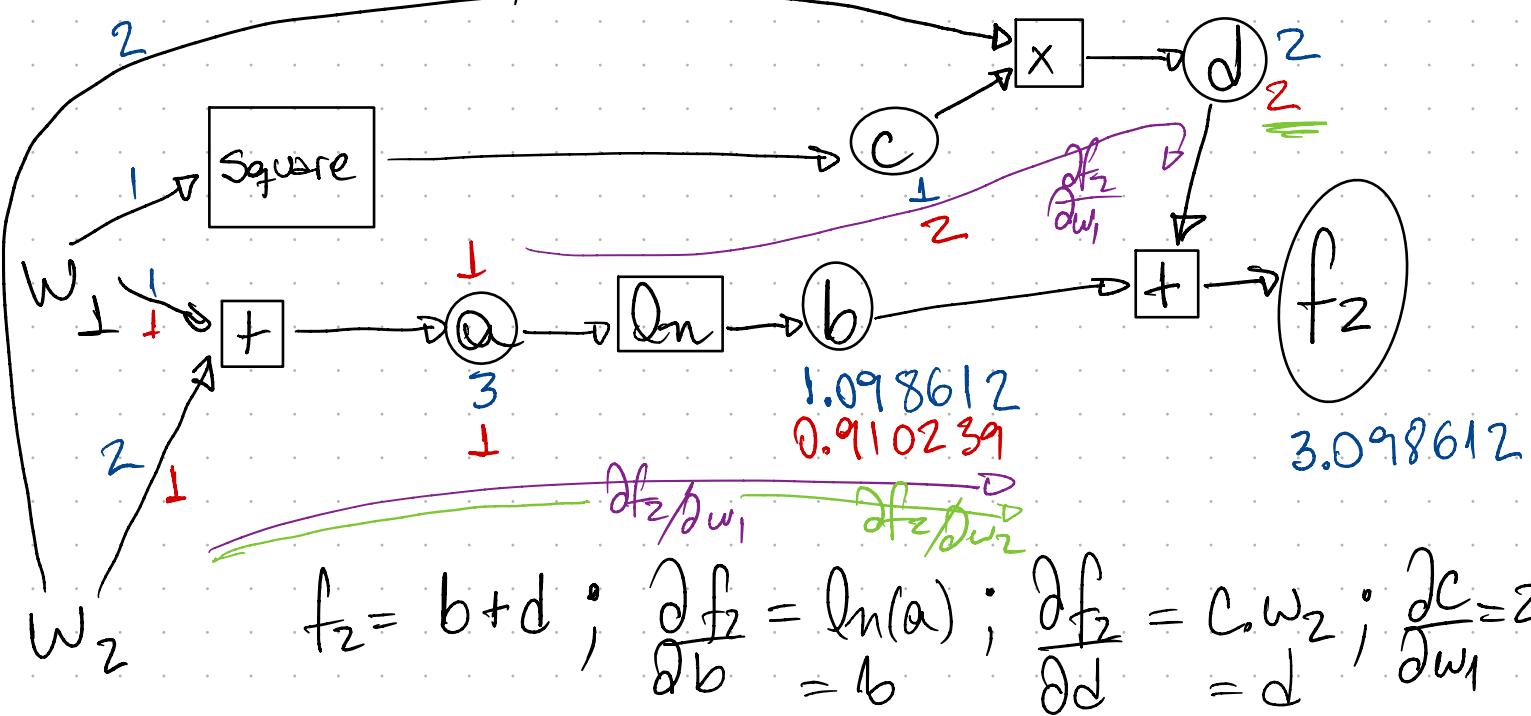
$$\frac{\partial f_1}{\partial w_2} = \frac{\partial b}{\partial w_2} \cdot \frac{\partial c}{\partial b} \cdot \frac{\partial f}{\partial b} \cdot \frac{\partial f}{\partial d} = \boxed{-0.045439}$$

4.c) Continued

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

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

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



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

fwd:

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

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

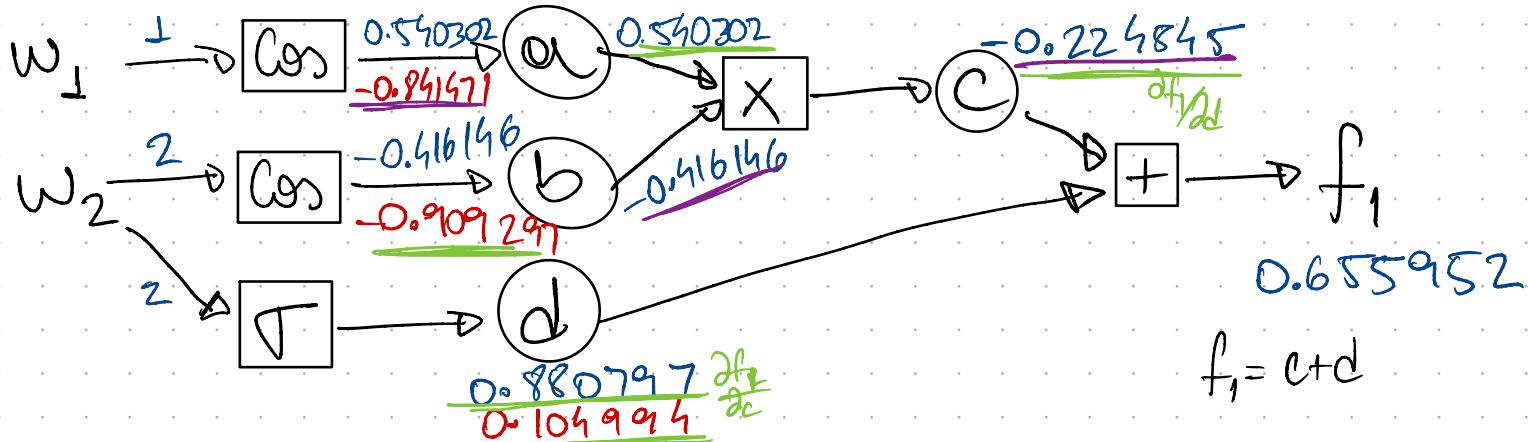
$$\nabla f = \begin{bmatrix} -0.078735 & -0.045439 \\ 4.910239 & 2.910239 \end{bmatrix}$$

4.d)

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

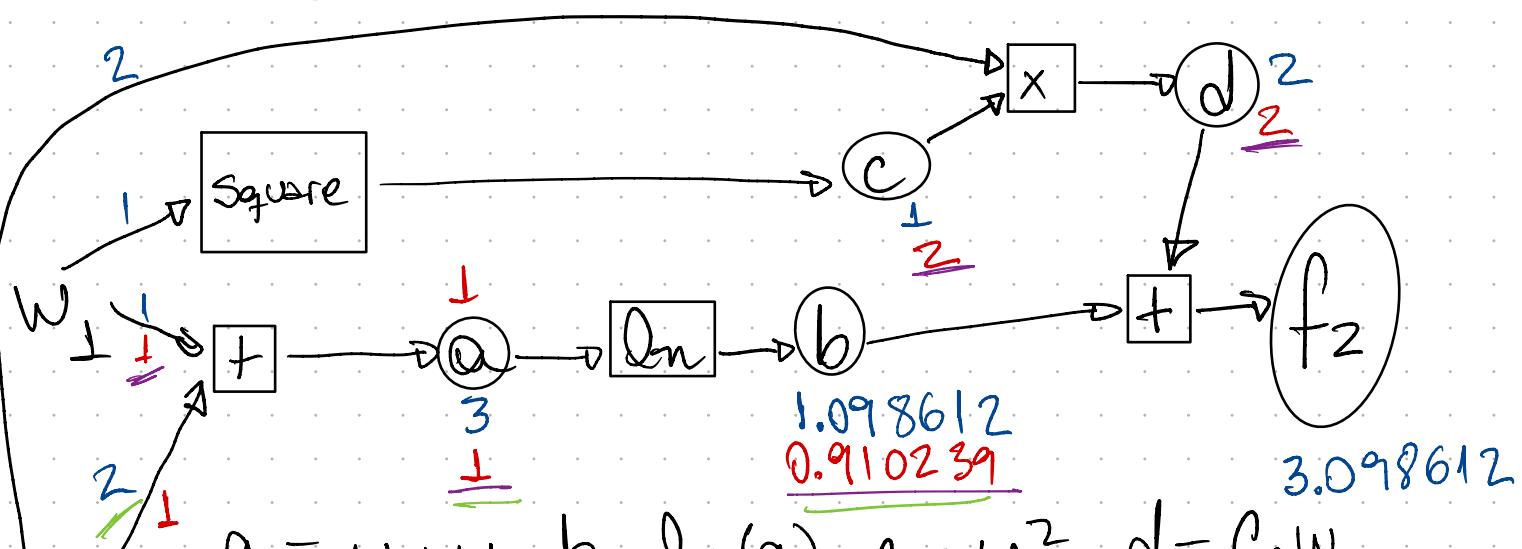
$$f_2(w_1, w_2) = \ln(w_1 + w_2) + w_1^2 w_2$$

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



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

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



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

$$f_2 = b + d ; \frac{\partial f_2}{\partial b} = \ln(a) ; \frac{\partial f_2}{\partial d} = c \cdot w_2 ; \frac{\partial c}{\partial w_1} = 2w_1$$

$$\frac{\partial f_2}{\partial w_1} = \frac{\partial f_2}{\partial b} \cdot \frac{\partial b}{\partial a} \cdot \frac{\partial a}{\partial w_1} + \frac{\partial f_2}{\partial d} \cdot \frac{\partial d}{\partial c} \cdot \frac{\partial c}{\partial w_1} = 1.820478 + \frac{4}{2w_1} = 5.820478$$

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

4.d) continued

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

4.e) YES!

5.2) 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

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

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

If $CS = SC$ (commutative property):

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

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

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

Assume B and S are commutative:

$$BS - SB = 0$$

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

Contradiction

\therefore The statement is NOT Bidirectional, and Convolution is the only linear operator with shift equivariance



S.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.

Counterexample: 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 $w=0$. The algorithm has 2 choices:

- 1) pick a
- 2) pick b

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

$$\begin{aligned} \text{If (2)} \Rightarrow \nabla f^b(w) &= \frac{\partial}{\partial w} \frac{1}{2}(w+1)^2 = \frac{\partial}{\partial w} \frac{1}{2}(w^2 + 2w + 1) \\ &= \frac{1}{2}(2w + 2) = w + 1 \end{aligned}$$

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

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

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

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

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



CS 4644/7643: Deep Learning
Spring 2024
HW2 Solutions

Arthur Scaquetti do Nascimento

Due: 11:59pm, Feb 19, 2024

7.a)

- Key contributions: This work clearly lays out a well-defined framework for gauging the effective capacity of machine learning models. Even though the authors do not address the shortcomings of Deep Learning that they found, they do state that engineering a (deep) network to perform optimization is not a hard task despite of the models not being able to generalize. Moreover, the authors state that this is an indication that the very reasons that make optimization easy in practice must not be why generalization works.
- Strengths:
 - The design of an evaluation framework for "measuring" the effective capacity of ML models (i.e., looking at randomization tests, explicit and implicit regularization);
 - The very questioning tone, leading to 'how can we trust that blackbox models are generalizing other than looking at test and validation curves?'. That point is very important, because it brings to light that sometimes we are looking at the wrong metrics, even though those metrics have been extensively used in the field.
 - Bringing to light implicit regularization, which is often overseen, and was particularly unknown to me.
- Weaknesses: The main thing that unsettles me about this type of paper is that they do not propose an answer, or designing guidelines such as a structured methodology in which it is guaranteed to only optimize when generalization happens. I know, this is basically solving the whole blackbox problem at once, but it bugs me when a work is published "just" to point out flaws. However, I think this work would have been a dozen times stronger if it had some design guidelines (even if faint), or proposed a different metric to look at.

7.b) Personal takeaways:

I had never thought of the impacts of implicit regularization, and will definitely keep an eye out for that. I also had never considered that Deep Learning models could present such good performances using completely corrupted datasets – this is potentially one of the things I will carry the most from reading this work: when in doubt if a model is actually good, I will try it out on a noisy and/or randomly labeled dataset as a sanity check.

However, other than that, I do not think that I gained that much insight from this work. Unfortunately, we still need to rely on blackboxes for the sake of solving certain problems. Hopefully XAI helps us with that.