DL HW1

Sruthi Sudhakar

September 9th 2020

1 Gradient Descent

1.

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

Lets say

$$F(\mathbf{w*}) = f(\mathbf{w}^t) + < \mathbf{w} - \mathbf{w}^t, \nabla f(\mathbf{w}^t) + \frac{\lambda}{2}(||\mathbf{w} - \mathbf{w}^t||)^2$$

We want to find **w*** where the value of F is the minimum. To do this we can set F'(**w***) = 0 and solve for **w***
First lets solve for $F'(\mathbf{w*})$

$$F'(\mathbf{w*}) = \frac{df(\mathbf{w}^t)}{d\mathbf{w*}} + \frac{d(\mathbf{w} - \mathbf{w}^t)}{d\mathbf{w*}} \cdot (\nabla f(\mathbf{w}^t)) + (\mathbf{w} - \mathbf{w}^t) \cdot \frac{d(\nabla f(\mathbf{w}^t))}{d\mathbf{w*}} + \frac{\lambda}{2} \frac{(d||\mathbf{w} - \mathbf{w}^t||)^2}{d\mathbf{w*}}$$

$$F'(\mathbf{w*}) = 0 + \nabla f(\mathbf{w}^t) + \frac{\lambda}{2}(2(\mathbf{w} - \mathbf{w}^t)$$

$$F'(\mathbf{w*}) = \nabla f(\mathbf{w}^t) + \lambda(\mathbf{w} - \mathbf{w}^t)$$

Now lets set $F'(\mathbf{w*}) = 0$ and solve for **w***

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

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

We can see the gradient descent update rule ($\mathbf{w}^{(t+1)} = \mathbf{w}^t - \eta \nabla f(\mathbf{w}^t)$) in this solution. This tells us that the gradient descent update rule *is* in-fact optimally minimizing the function $f(\mathbf{w})$ at each step along the way.
In gradient descent, we define η to be the step size that we take in the direction of the negative gradient. In the optimization problem, we define λ as the trade-off between the approximation being close to $\mathbf{w}^t$ and obtaining a lower bound.

We can see from this equation that the term η from the gradient descent update rule is equal to $\frac{1}{\lambda}$ from our l2 proximity term. Intuitively, we can say that the amount we step in the direction of the negative gradient should be enough so that we are still faithfully close to our starting point $\mathbf{w}^t$ while still achieving a minimum along our convex function.

2.

We want to prove that

$$\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t > \leq \frac{||\mathbf{w}^*||^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}||\mathbf{v}_t||^2$$

First lets apply some algebraic manipulations on the LHS of the inequality. By defn of inner product,

$$< \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t >= \frac{1}{\eta} < \mathbf{w}^t - \mathbf{w}^*, \eta\mathbf{v}_t > \tag{1}$$

We also know that

$$||x-y||^2 =< x-y, x-y >= ||x||^2 - 2 < x, y > +||y||^2$$

Therefore,

$$||\mathbf{w}^t - \mathbf{w}^* - \eta\mathbf{v}_t||^2 = ||\mathbf{w}^t - \mathbf{w}^*||^2 - 2 < \mathbf{w}^t - \mathbf{w}^*, \eta\mathbf{v}_t > +||\eta\mathbf{v}_t||^2 \tag{2}$$

So we can replace the LHS in (1) with (2).

$$< \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t >= \frac{1}{2\eta}(||\mathbf{w}^t - \mathbf{w}^*||^2 + ||\eta\mathbf{v}_t||^2 - ||\mathbf{w}^t - \mathbf{w}^* - \eta\mathbf{v}_t||^2) \tag{3}$$

By our gradient update rule $\mathbf{w}^{(t+1)} = \mathbf{w}^t - \eta\mathbf{v}_t)$, we can simplify this to

$$< \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t >= \frac{1}{2\eta}(||\mathbf{w}^t - \mathbf{w}^*||^2 + \eta^2||\mathbf{v}_t||^2 - ||\mathbf{w}^{t+1} - \mathbf{w}^*||^2) \tag{4}$$

Now lets add our summation back in

$$\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t >= \frac{1}{2\eta}\sum_{t=1}^{T}(||\mathbf{w}^t - \mathbf{w}^*||^2 + \eta^2||\mathbf{v}_t||^2 - ||\mathbf{w}^{t+1} - \mathbf{w}^*||^2)$$

$$\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t >= \frac{1}{2\eta}\sum_{t=1}^{T}(||\mathbf{w}^t - \mathbf{w}^*||^2 - ||\mathbf{w}^{t+1} - \mathbf{w}^*||^2) + \frac{1}{2\eta}\sum_{t=1}^{T}(\eta^2||\mathbf{v}_t||^2) =$$

$$\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t >= \frac{1}{2\eta}\sum_{t=1}^{T}(||\mathbf{w}^t - \mathbf{w}^*||^2 - ||\mathbf{w}^{t+1} - \mathbf{w}^*||^2) + \frac{\eta}{2}\sum_{t=1}^{T}(||\mathbf{v}_t||^2)$$

The summation of the first part of the RHS of the equation will reduce because terms will cancel out resulting in

$$\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t >= \frac{1}{2\eta}(||\mathbf{w}^1 - \mathbf{w}^*||^2 - ||\mathbf{w}^{T+1} - \mathbf{w}^*||^2) + \frac{\eta}{2}\sum_{t=1}^{T}(||\mathbf{v}_t||^2)$$

Now to form our inequality, observe that

$$||\mathbf{w}^{T+1} - \mathbf{w}^*||^2 \geq 0$$

because any real number squared can never achieve a negative value. Therefore, removing this term from the left side leaves us with something greater (since it was subtracted out) so

$$\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t > \leq \frac{1}{2\eta}(||\mathbf{w}^1 - \mathbf{w}^*||^2) + \frac{\eta}{2}\sum_{t=1}^{T}(||\mathbf{v}_t||^2)$$

Finally, since we are told that $\mathbf{w}^1 = 0$, we obtain

$$\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \mathbf{v}_t > \leq \frac{1}{2\eta}(||\mathbf{w}^*||^2) + \frac{\eta}{2}\sum_{t=1}^{T}(||\mathbf{v}_t||^2)$$

3.

For a convex function, Jensen's inequality states that $E[f(x)] \geq f(E[x])$ Therefore, we have that

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

Since $f(\mathbf{w}^*)$ is not affected by t, we can simplify this to

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

Furthermore, because of our convexity assumption, we know that

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T} < \mathbf{w}^t - \mathbf{w}^*, \nabla f(\mathbf{w}^t) >$$

Therefore, combining this inequality with what we proved in part 2 gives us

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{2T\eta}(||\mathbf{w}^*||^2) + \frac{\eta}{2T}\sum_{t=1}^{T}(||\nabla f(\mathbf{w}^t)||^2)$$

Now lets substitute the upperbounds in for $||\mathbf{w}^*||^2$ and $||\mathbf{v}_t||^2$ and $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$.

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{2T\sqrt{\frac{B^2}{\rho^2 T}}}(B^2) + \frac{\sqrt{\frac{B^2}{\rho^2 T}}}{2T}\sum_{t=1}^{T}(\rho^2)$$

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{1}{T}(\frac{B^2 p\sqrt{T}}{2B} + \frac{BT(\rho^2)}{2p\sqrt{T}})$$

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{3Bp}{\sqrt{T}}$$

Since 3, B, and ρ are constants, we know that

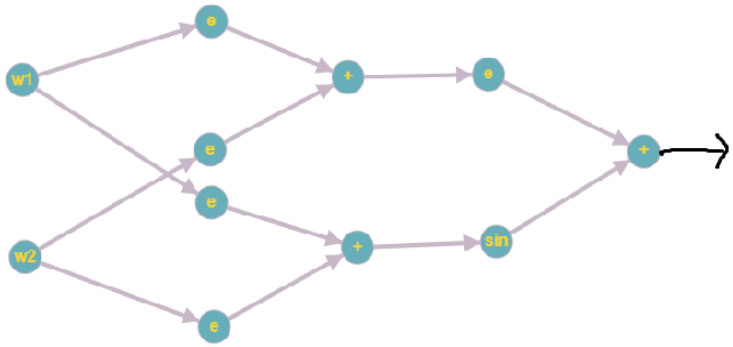
$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \propto \frac{1}{\sqrt{T}}$$

and therefore the convergence rate is $O(\frac{1}{\sqrt{T}})$

3 Automatic Differentiation

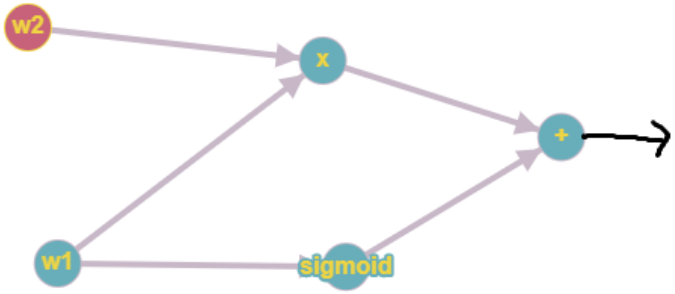
5.

a)



$$f_1(w_1, w_2) = e^{e^{w_1}+e^{2w_2}} + sin(e^{w_1} + e^{2w_2})$$

$$f_1(\mathbf{w}) = f_1(1, 2) = e^{e^1+e^{2*2}} + sin(e^1 + e^{2*2}) \approx 7.802e24$$



$$f_2(w_1, w_2) = w_1 * w_2 + \sigma(w_1)$$

$$f_2(\mathbf{w}) = f_2(1, 2) = 1 * 2 + \sigma(1) \approx 2.731$$

b)

i) First lets compute $\frac{\partial f_1}{\partial w_1}$

$$f_1(\mathbf{w} + (\Delta w, 0)) = f_1(1 + 0.01, 2) = 8.012e24$$

$$\frac{\partial f_1}{\partial w_1} = \frac{f_1(\mathbf{w} + (\Delta w, 0)) - f_1(\mathbf{w})}{\Delta w} = \frac{8.012e24 - 7.802e24}{0.01} = 2.162e25$$

ii) Now lets compute $\frac{\partial f_1}{\partial w_2}$

$$f_1(\mathbf{w} + (0, \Delta w)) = f_1(1, 2 + 0.01) = 2.351e25$$

$$\frac{\partial f_1}{\partial w_2} = \frac{f_1(\mathbf{w} + (0, \Delta w)) - f_1(\mathbf{w})}{\Delta w} = \frac{2.351e25 - 7.802e24}{0.01} = 1.571e27$$

iii) Now lets compute $\frac{\partial f_2}{\partial w_1}$

$$f_2(\mathbf{w} + (\Delta w, 0)) = f_2(1 + 0.01, 2) = 2.753$$

$$\frac{\partial f_2}{\partial w_1} = \frac{f_2(\mathbf{w} + (\Delta w, 0)) - f_2(\mathbf{w})}{\Delta w} = \frac{2.753 - 2.731}{0.01} = 2.2$$

iv) Now lets compute $\frac{\partial f_2}{\partial w_2}$

$$f_2(\mathbf{w} + (0, \Delta w)) = f_1(1, 2 + 0.01) = 2.741$$

$$\frac{\partial f_2}{\partial w_2} = \frac{f_2(\mathbf{w} + (0, \Delta w)) - f_2(\mathbf{w})}{\Delta w} = \frac{2.741 - 2.731}{0.01} = 1$$

c)

i) First lets compute $\frac{\partial f_1}{\partial w_1}$

Lets define several intermediate variables to simplify computations.
$w_1 = w_1 = 1$; $w_2 = w_2 = 2$; $w_3 = e^{w_1} = 2.718$; $w_4 = e^{2w_2} = 54.599$; $w_5 = w_3 + w_4 = 57.316$; $w_6 = e^{w_5} = 7.802e24$; $w_7 = w_3 = 2.718$; $w_8 = w_4 = 54.599$; $w_9 = w_5 = 57.316$; $w_{10} = sin(w_9) = 0.6945$; $f = w_{11} = w_{10} + w_6 = 7.802e24$;

Now lets go in order to find the derivatives of each w_i for $i = 1...w_11$ which will continuously build on top of each other.
$\partial w_1/w_1 = 1$, $\partial w_1/w_2 = 0$;
$\partial w_2/w_1 = 0$, $\partial w_2/w_2 = 1$;
$\partial w_3/w_1 = e^{w_1}$, $\partial w_3/w_2 = 0$;
$\partial w_4/w_1 = 0$, $\partial w_4/w_2 = 2e^{2w_2}$;
$\partial w_5/w_1 = e^{w_1}$, $\partial w_5/w_2 = 2e^{2w_2}$;
$\partial w_6/w_1 = e^{w_1}e^{w_5}$, $\partial w_6/w_2 = 2e^{2w_2}e^{w_5}$;
$\partial w_7/w_1 = e^{w_1}$, $\partial w_7/w_2 = 0$;
$\partial w_8/w_1 = 0$, $\partial w_8/w_2 = 2e^{2w_2}$;
$\partial w_9/w_1 = e^{w_1}$, $\partial w_9/w_2 = 2e^{2w_2}$;
$\partial w_{10}/w_1 = cos(w_9)e^{w_1}$, $\partial w_{10}/w_2 = 2e^{2w_2}cos(w_9)$;
$\partial f/w_1 = \partial w_{11}/w_1 = cos(w_9)e^{w_1} + e^{w_1}e^{w_5}$,
$\partial f/w_2 = \partial w_{11}/w_2 = 2e^{2w_2}cos(w_9) + 2e^{2w_2}e^{w_5}$;

So finally by substituting in w_1 and w_2 we get

$$\partial f(1,2)/w_1 = cos(w_9)e^{w_1} + e^{w_1}e^{w_5} = cos(57.316)e^1 + e^1e^{57.316} = 2.12e25$$

ii) Now lets compute $\frac{\partial f_1}{\partial w_2}$

Using our calculations above, we just substitute and get

$$\partial f(1,2)/w_2 = 2e^{2w_2}cos(w_9)+2e^{2w_2}e^{w_5} = 2e^{2*2}cos(57.316)+2e^{2*2}e^{57.316} = 8.516e26$$

iii) Now lets compute $\frac{\partial f_2}{\partial w_1}$

$\partial w_1/w_1 = 1$, $\partial w_2/w_1 = 0$, $\partial x_1/w_1 = w_2$, $\partial x_2/w_1 = \sigma(w_1)(1-\sigma(w_1))$ $\partial x_3/w_1 = \partial f_2/w_1 = \partial x_1/w_1 + \partial x_2/w_1 = w_2 + \sigma(w_1)(1 - \sigma(w_1))$

$$\partial f_2(1,2)/w_1 = 2 + \sigma(1)(1 - \sigma(1)) = 2.2$$

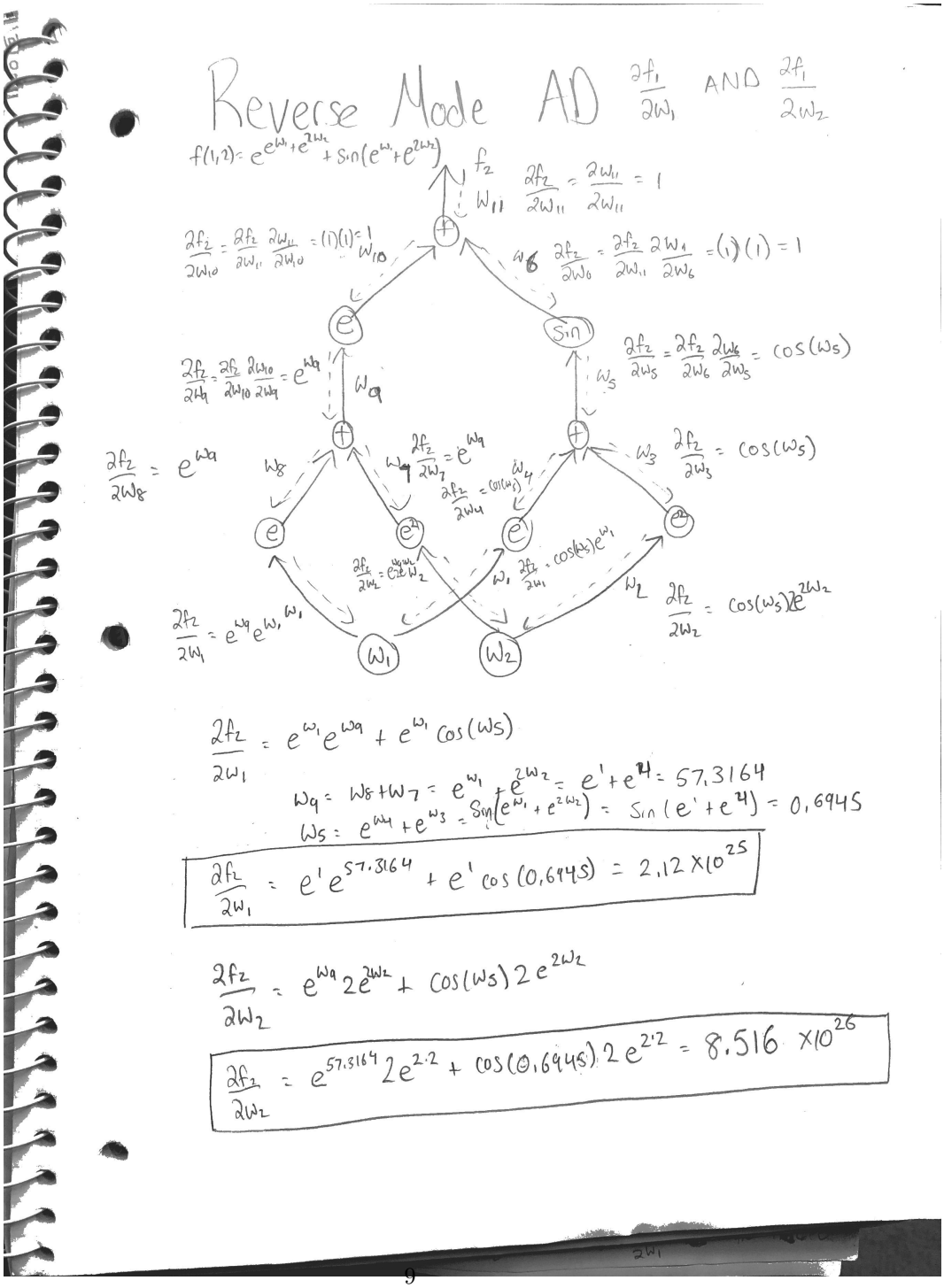
iii) Now lets compute $\frac{\partial f_2}{\partial w_2}$

$\partial w_1/w_2 = 0$, $\partial w_2/w_2 = 1$, $\partial x_1/w_2 = w_1$, $\partial x_2/w_2 = 0$ $\partial x_3/w_2 = \partial f_2/w_2 = \partial x_1/w_2 + \partial x_2/w_2 = w_1 + 0$

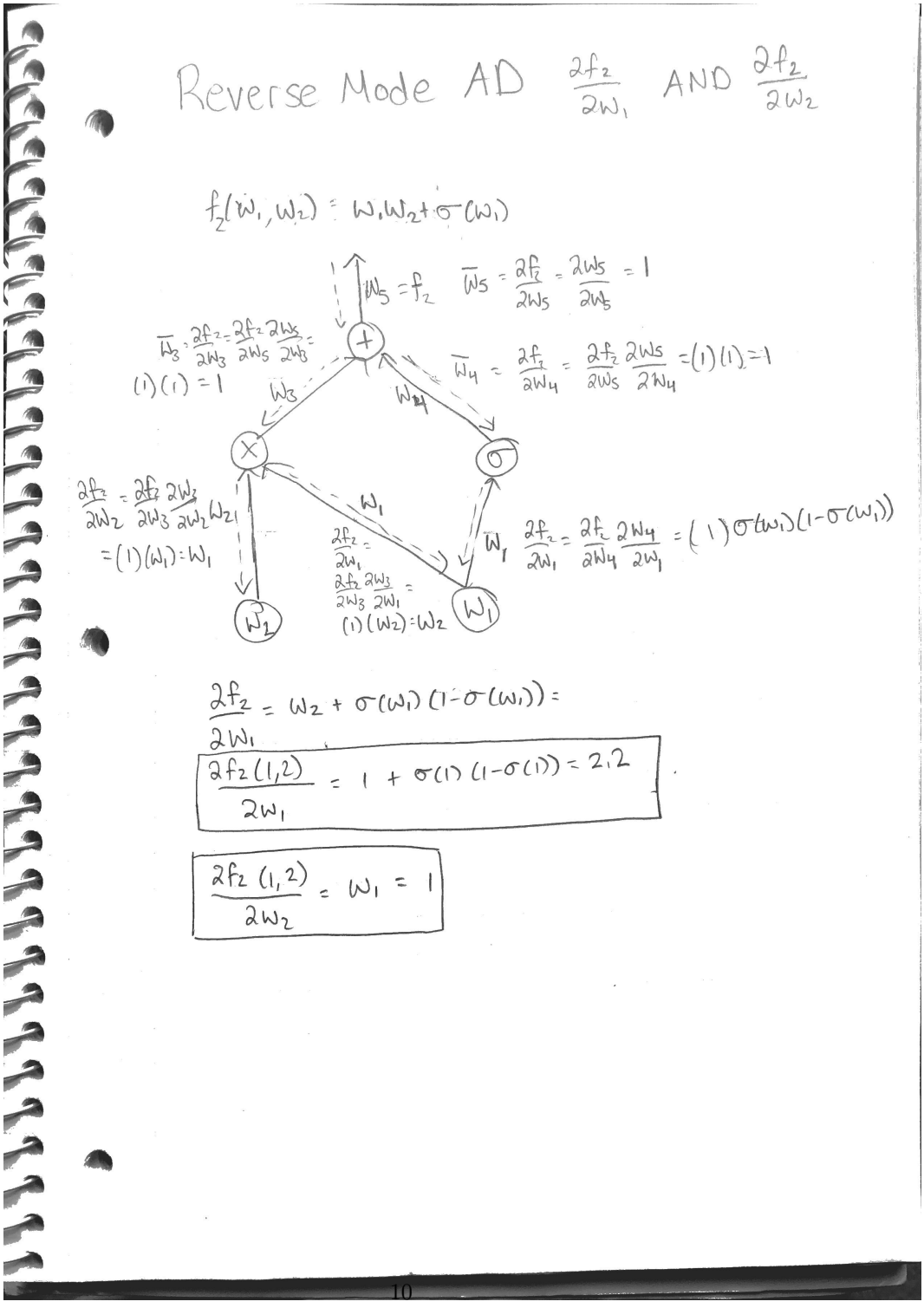
$$\partial f_2(1,2)/w_2 = 1$$

d)

i and ii


Reverse Mode AD $\frac{\partial f_1}{\partial w_1}$ AND $\frac{\partial f_1}{\partial w_2}$
$f(1,2) = e^{e^{w_1} + e^{2w_2}} + \sin(e^{w_1} + e^{2w_2})$
f_2
w_{11}
$\frac{\partial f_2}{\partial w_{11}} = \frac{\partial w_{11}}{\partial w_{11}} = 1$
$\frac{\partial f_2}{\partial w_{10}} = \frac{\partial f_2}{\partial w_{11}} \frac{\partial w_{11}}{\partial w_{10}} = (1)(1) = 1$
w_{10}
w_6
$\frac{\partial f_2}{\partial w_6} = \frac{\partial f_2}{\partial w_{11}} \frac{\partial w_{11}}{\partial w_6} = (1)(1) = 1$
e
$\sin$
$\frac{\partial f_2}{\partial w_9} = \frac{\partial f_2}{\partial w_{10}} \frac{\partial w_{10}}{\partial w_9} = e^{w_9}$
w_9
w_5
$\frac{\partial f_2}{\partial w_5} = \frac{\partial f_2}{\partial w_6} \frac{\partial w_6}{\partial w_5} = \cos(w_5)$
$\frac{\partial f_2}{\partial w_8} = e^{w_8}$
w_8
w_7
$\frac{\partial f_2}{\partial w_7} = e^{w_9}$
$\frac{\partial f_2}{\partial w_4} = \cos(w_4)$
w_4
w_3
$\frac{\partial f_2}{\partial w_3} = \cos(w_5)$
e
e^2
e
e
$\frac{\partial f_2}{\partial w_2} = e^{2w_2} w_2$
w_1
$\frac{\partial f_2}{\partial w_1} = \cos(w_5) e^{w_1}$
w_2
$\frac{\partial f_2}{\partial w_1} = e^{w_9} e^{w_1} w_1$
$\frac{\partial f_2}{\partial w_2} = \cos(w_5) 2e^{2w_2}$
w_1
w_2
$\frac{\partial f_2}{\partial w_1} = e^{w_1} e^{w_9} + e^{w_1} \cos(w_5)$
$w_9 = w_8 + w_7 = e^{w_1} + e^{2w_2} = e^1 + e^4 = 57.3164$
$w_5 = e^{w_4} + e^{w_3} = \sin(e^{w_1} + e^{2w_2}) = \sin(e^1 + e^4) = 0.6945$
$\frac{\partial f_2}{\partial w_1} = e^1 e^{57.3164} + e^1 \cos(0.6945) = 2.12 \times 10^{25}$
$\frac{\partial f_2}{\partial w_2} = e^{w_9} 2e^{2w_2} + \cos(w_5) 2e^{2w_2}$
$\frac{\partial f_2}{\partial w_2} = e^{57.3164} 2e^{2 \cdot 2} + \cos(0.6945) 2e^{2 \cdot 2} = 8.516 \times 10^{26}$

iii and iv


Reverse Mode AD $\frac{2f_2}{2w_1}$ AND $\frac{2f_2}{2w_2}$
$f_2(w_1, w_2) = w_1 w_2 + \sigma(w_1)$
$w_5 = f_2$
$\bar{w}_5 = \frac{2f_2}{2w_5} = \frac{2w_5}{2w_5} = 1$
$\bar{w}_3 = \frac{2f_2}{2w_3} = \frac{2f_2}{2w_5}\frac{2w_5}{2w_3} = (1)(1) = 1$
$\bar{w}_4 = \frac{2f_2}{2w_4} = \frac{2f_2}{2w_5}\frac{2w_5}{2w_4} = (1)(1) = 1$
$\frac{2f_2}{2w_2} = \frac{2f_2}{2w_3}\frac{2w_3}{2w_2}w_2 = (1)(w_1) = w_1$
$\bar{w}_1 \frac{2f_2}{2w_1} = \frac{2f_2}{2w_4}\frac{2w_4}{2w_1} = (1)\sigma(w_1)(1-\sigma(w_1))$
$\frac{2f_2}{2w_1} = \frac{2f_2}{2w_3}\frac{2w_3}{2w_1} = (1)(w_2) = w_2$
w_1
w_2
$\frac{2f_2}{2w_1} = w_2 + \sigma(w_1)(1-\sigma(w_1)) =$
$\frac{2f_2(1,2)}{2w_1} = 1 + \sigma(1)(1-\sigma(1)) = 2.2$
$\frac{2f_2(1,2)}{2w_2} = w_1 = 1$

e)

4 Convolutions

6.

a)

To prove that S and C_a commute, lets show that $SC_a = C_aS$ which means that $(SC_a)_{ij} = (C_aS)_{ij}$ for any row i and col j.

We know that since S is the shift matrix, the value at each row i and col j is only equal to 1 when i-j-1=0. Therefore, we can say that$S_{ij} = \delta_{i-j-1}$.
Furthermore, since we have the additional property that $a_i = a_{i+n}$, we know that for any element a_{ij} in the matrix C_a, $a_{ij} = a_{i-j}$

$$(SC_a)_{ij} = \sum_{b=0}^{n-1} S_{ib}(C_a)_{bj} = \sum_{b=0}^{n-1} \delta_{i-b-1}a_{b-j}$$

The expression δ_{i-b-1} is only equal to 1 when i-b-1=0, so b = i-1. This gives us

$$(SC_a)_{ij} = i - 1 - j$$

$$(C_aS)_{ij} = \sum_{b=0}^{n-1}(C_a)_{ib}S_{bj} = \sum_{b=0}^{n-1} a_{i-b}\delta_{b-j-1}$$

The expression δ_{b-j-1} is only equal to 1 when b-j-1=0, so b = j+1. This gives us

$$(SC_a)_{ij} = i - j - 1$$

Since we have proved that $(SC_a)_{ij} = (SC_a)_{ij}$, we have shown that the shift matrix and the circulant matrix are commutative.
*NOTE: This proof works even if our circulant matrix took the form of $(C_a)^T$ where C_a is the circulant matrix form given in the hw. This is because our assumption of $a_i = a_{i+n}$ is generalized to all types of circulant matrices so the computations do not change.

b)

Lets prove the bidirectional implication that A matrix is circular convolution IF and only IF it is a linear operation with shift invariance

IF there is a circular convolution => it is a linear operation with shift invariance

In 6a, we proved that f(Sx)=Sf(x) where f(x)=C_ax. This proves one direction that states that if C_a is a circulant convolution matrix, then it is shift invariant.

IF there is a linear operation with shift invariance => it is the circular convolution

Here, we get to assume that f(Sx)=Sf(x). The question is how do we show that C_a in f(x)=C_ax is the circulant convolution? Lets define $C_a = A$ to be a random matrix with column vectors = $[a_0, ..., a_{n-1}]$. We know that S is the shift matrix as shown above in the homework problem. Lets define the x vector with values $x = [x_0, ..., x_{n-1}]$.
Lets first look at ASx. First compute Sx. Since S is the shift matrix, we know that Sx = $[x_{n-1}, x_0, x_1, ..., x_{n-3}, x_{n-2}]$. Now lets apply A to this vector Sx.

$$[a_0, ..., a_{n-1}][x_{n-1}, x_0, x_1, ..., x_{n-3}, x_{n-2}] = [a_0x_{n-1}+a_1x_0+a_2x_1+...+a_{n-2}x_{n-3}+a_{n-1}x_{n-2}]$$

Now lets look at SAx. First compute Ax.

$$[a_0, ..., a_{n-1}][x_0, ..., x_{n-1}] = a_0x_0 + a_1x_1 + a_2x_2 + ... + a_{n-2}x_{n-2} + a_{n-1}, x_{n-1}$$

Now lets apply S to the vector Ax.

$$S[a_0x_0+a_1x_1+a_2x_2+...+a_{n-2}x_{n-2}+a_{n-1}, x_{n-1}] = [Sa_0x_0+Sa_1x_1+Sa_2x_2+...+Sa_{n-2}x_{n-2}+Sa_{n-1}, x_{n-1}]$$

Now since ASx = SAx, these two calculations are equivalent.

$$[a_0x_{n-1}+a_1x_0+a_2x_1+...+a_{n-2}x_{n-3}+a_{n-1}x_{n-2}] = [Sa_0x_0+Sa_1x_1+Sa_2x_2+...+Sa_{n-2}x_{n-2}+Sa_{n-1}, x_{n-1}]$$

This means that Sa_ix_i must be equivalent to $a_{(i+1)modn}x_i$. This means that $Sa_i = a_{(i+1)modn}$. If this is true, then we know that $a_{(i+1)modn}$ is simply the vector whose column is made up of the values in the column a_i with a circular shift. Since each subsequent column of A is obtained by a circular shift of the previous column, we know that this is a circulant convolution matrix.

c)

This tells us that deep learning architecture for spatio-temporal data like images and videos are only possible through convolutions IF you want the model to be shift invariant. This is because as we proved, convolutions are are the only shift invariant linear operators and therefore, there is no other linear operators that would allow a model to generalize to small shifts in input images/videos. Convolutions are thus ideal for finding features that are encoded in spatial processing and being invariant to location of the features relative to a standard xyz reference frame.

5 Paper Review

7.

The central theme of this paper was trying to make deep learning scientist think about how regularization really affects generalization and what truly causes a deep neural network to generalize well to the held out test data. The authors ran an interesting experiment where they randomized the training labels such that they no longer corresponded to the correct data points and noticed that a model could still achieve a training loss of zero though there is no relationship between the data and the labels! They also tried an experiment where the input images were just noise and noticed the network again achieved zero training error. With such experiments, they concluded a that neural networks have the capability to memorize an entire dataset, and that architecture does indeed affect regularization. To figure out what part of the architecture affects the generalization of the model, the authors ran several studies with different regularization methods but the ultimatum was that there is no clear answer to this question and this subject requires further analysis.

8.

First, I realized that regularization is not the be all end all for ensuring a network is generalizeable. This paper showed that many times, a regularized network preforms well in unseen test data but the regularized network with the same hyper parameters has very little difference. This shows that there is a high chance the regularization may not be contributing anything to the outputs. Secondly, optimization's connection to generalization is not necessarily what we think it is. We add a regularization term so that the optimizer can converge but if this regularization term is not affecting the model's performance, then why is it able to optimize so quickly? This topic needs to be explored in further detail. Furthermore, a future discussion that could be brought up is how there is a possibility that large deep learning network architectures are actually just memorizing the training data in a 'smart' way but are still generalizing well to unseen data. There is no way we have a confirmation that this does not happen so I wonder if this is a possibility?