

1. for convex  $f$  with  $L_2$  proximity term

$$w^{(t+1)} = \operatorname{arg\min}_w f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle + \frac{\lambda}{2} \|w - w^{(t)}\|^2$$

derive w.r.t.  $w$

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

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

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

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

$$\Rightarrow w^{(t+1)} = w^{(t)} - \frac{1}{\lambda} \nabla f(w^{(t)})$$

$$\Rightarrow w^* = w^{(t)} - \frac{1}{\lambda} \nabla f(w^{(t)})$$

This gives us the same form as the gradient descent update rule. Note that  $\frac{1}{\lambda} = n$ . This tells us that gradient descent minimizes the approximation at each step. The relationship tells us that as learning rate increases, regularization must decrease.

$$2. \text{ show } \sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2n} + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2$$

$$\text{Take LHS: } \sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle$$

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

$$= \underbrace{\sum_{t=1}^T \langle w^{(t)}, v_t \rangle}_{\textcircled{1}} - \underbrace{\langle w^*, \sum_{t=1}^T v_t \rangle}_{\textcircled{2}}$$

aside: consider update eq'n  
 $w^{(t+1)} = w^{(t)} - nV_T, w^{(1)} = 0$

$$w^2 = w^1 - nV_1 = -nV_1$$

$$w^3 = w^2 - nV_2 = -nV_1 - nV_2$$

$\vdots$

$$w^{(t+1)} = w^{(t)} - nV_t = -nV_1 - \dots - nV_t \\ = -n \sum_{i=1}^t V_i \quad \textcircled{3}$$

Take ① and ③:

$$\begin{aligned} \sum_{t=1}^T \langle w^{(t)}, v_t \rangle &= \sum_{t=1}^T \left\langle -n \sum_{i=1}^{t-1} V_i, v_t \right\rangle \\ &= -n \sum_{t=1}^T \left\langle \sum_{i=1}^{t-1} V_i, v_t \right\rangle \\ &= -n \left( \frac{\langle \sum_{t=1}^T V_t, \sum_{t=1}^T v_t \rangle}{2} - \frac{\sum_{t=1}^T \|v_t\|^2}{2} \right) \\ &= -\frac{n}{2} \left\| \sum_{t=1}^T V_t \right\|^2 + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2 \\ &= \frac{-1}{2n} \|w^{(t+1)}\|^2 + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2 \end{aligned}$$

Take ② and ③:

$$\text{since } w^{(t+1)} = -n \sum_{t=1}^T V_t \Rightarrow \sum_{t=1}^T V_t = -\frac{1}{n} w^{(t+1)}$$

$$\langle w^*, \sum_{t=1}^T v_t \rangle = \langle w^*, -\frac{1}{n} w^{(t+1)} \rangle$$

$$\therefore \text{LHS} = \frac{-1}{2n} \|w^{(t+1)}\|^2 + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2 + \frac{1}{n} \langle w^*, w^{(t+1)} \rangle$$

$$\text{RHS} = \frac{\|w^*\|^2}{2n} + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2 \quad \text{terms cancel out}$$

$$\text{We need to show } \underbrace{\frac{-1}{2n} \|w^{(t+1)}\|^2}_{\textcircled{4}} + \underbrace{\frac{1}{n} \langle w^*, w^{(t+1)} \rangle}_{\textcircled{5}} \leq \frac{\|w^*\|^2}{2n} \quad \textcircled{6}$$

since ④ is the negative product of a positive term ( $\|w^{(t+1)}\|^2$ ), ④  $\leq 0$

$$\text{take ⑤: } \frac{1}{n} \langle w^*, w^{(t+1)} \rangle = \frac{1}{2n} \langle w^*, w^{(t+1)} \rangle + \frac{1}{2n} \langle w^*, w^{(t+1)} \rangle$$

Note  $\langle w^*, w^{(t+1)} \rangle \leq 0$  and  $\frac{1}{2n} \langle w^*, w^{(t+1)} \rangle \leq \frac{1}{2n} \|w^*\|^2$

$$\therefore -\frac{1}{2n} \|w^{(t+1)}\|^2 + \frac{1}{2n} \langle w^*, w^{(t+1)} \rangle \leq 0$$

Thus our inequality is

$$\underbrace{-\frac{1}{2n} \|w^{(t+1)}\|^2}_{\text{this term } \leq 0} + \underbrace{\frac{1}{2n} \langle w^*, w^{(t+1)} \rangle + \frac{1}{2n} \langle w^*, w^{(t+1)} \rangle}_{\text{and the rest of the inequality is already proven}} \leq \frac{\|w^*\|^2}{2n}$$

so adding a value  $\leq 0$  holds the inequality  $\blacksquare$

3. For  $\bar{w} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$  we get  $f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$

by the definition of convexity:

$$f(w^*) \geq f(w^{(t)}) + \langle w^* - w^{(t)}, \nabla f(w^{(t)}) \rangle$$

$$\Rightarrow f(w^*) - f(w^{(t)}) \geq \langle w^* - w^{(t)}, \nabla f(w^{(t)}) \rangle$$

$$\Rightarrow f(w^{(t)}) - f(w^*) \leq \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

$$\Rightarrow \frac{1}{T} \sum_{t=1}^T [f(w^{(t)})] - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

Since  $f$  is convex, we know  $f(\bar{w}) = f\left(\frac{1}{T} \sum_{t=1}^T w^{(t)}\right) \leq \frac{1}{T} \sum_{t=1}^T f(w^{(t)})$

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

From Q2:  $\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2n} + \frac{n}{2} \sum_{t=1}^T \|v_t\|^2$

$$\Rightarrow f(\bar{w}) - f(w^*) \leq \frac{1}{T} \left[ \frac{\|w^*\|^2}{2n} + \frac{n}{2} \sum_{t=1}^T \|\nabla f(w^{(t)})\|^2 \right]$$

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

upper bound  $\|w^*\|$  with  $B$   
 upper bound  $\|\nabla f(w^{(t)})\|$  with  $\rho$

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

$$\Rightarrow f(\bar{w}) - f(w^*) \leq \frac{B^2}{2Tn} + \frac{nT\rho^2}{2T}$$

sub in  $n = \sqrt{\frac{B^2}{\rho^2 T}}$  and reduce

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

⋮

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

so convergence rate of  $w^*$  is  $O\left(\frac{1}{\sqrt{T}}\right)$  and  
 upper bound for  $f(\bar{w}) - f(w^*) \propto \frac{1}{\sqrt{T}}$

$$4. \text{ a) } g(x) = \underset{w}{\operatorname{argmin}} \left[ f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle + \frac{1}{2} (w - w^{(t)})^T H (w - w^{(t)}) \right]$$

take the derivative and set = 0

$$\nabla g(x) = \nabla f(w^{(t)}) + H(w - w^{(t)}) = 0$$

$$\Rightarrow w - w^{(t)} = -H^{-1} \nabla f(w^{(t)})$$

$$\Rightarrow w = w^{(t)} - H^{-1} \nabla f(w^{(t)})$$

↑ this is the form of Newton's update method

$H^{-1}$ , the inverse of the Hessian, shapes the gradient, giving emphasis on different directions. We can use the direction given by the Hessian to make the next iteration step.

$$4.b) \quad x \in \mathbb{R}^{50} \rightarrow h^{(1)} \in \mathbb{R}^{50} \rightarrow h^{(2)} \in \mathbb{R}^{50} \rightarrow s \in \mathbb{R}^{10}$$

$50 \times 50 + 50 \times 50 + 50 \times 50$  parameters in the MLP

$5500^2 \times 10$  dimensional Hessian

$$4.c) \nabla_w(w + \Delta w) = \nabla_w(w) + H\Delta w + O(\|\Delta w\|^2)$$

$$\text{let } \Delta w = rv$$

$$\Rightarrow H\Delta w = \nabla_w(w + \Delta w) - \nabla_w(w) - O(\|\Delta w\|^2)$$

$$\Rightarrow Hrv = \nabla_w(w + rv) - \nabla_w(w) - O(\|rv\|^2)$$

$$\Rightarrow Hv = \frac{1}{r} [\nabla_w(w + rv) - \nabla_w(w) + O(r^2)]$$

$$Hv = \frac{1}{r} [\nabla_w(w + rv) - \nabla_w(w)] + O(r)$$

$$4.d) R_v \{ f(w) \} = \frac{d}{dr} f(w+rv) = f'(w+rv) * (w+rv) dr = v f'(w+rv)$$

Proof of linearity:

$$R_v \{ cf(w) \} = \frac{d}{dr} cf(w+rv) = c \frac{d}{dr} f(w+rv) = cv f'(w+rv)$$

$$c R_v \{ f(w) \} = cv f'(w+rv)$$

$$\therefore R_v \{ cf(w) \} = c R_v \{ f(w) \} \quad \blacksquare$$

Proof of Product Rule:

$$\begin{aligned} R_v \{ f(w) g(w) \} &= \frac{d}{dr} (f(w+rv) g(w+rv)) \\ &= f(w+rv) \frac{d g(w+rv)}{dr} + g(w+rv) \frac{d f(w+rv)}{dr} \\ &= v f(w+rv) g'(w+rv) + v f'(w+rv) g(w+rv) \end{aligned}$$

$$R_v \{ f(w) \} g(w) + f(w) R_v \{ g(w) \} = v f'(w+rv) g(w+rv) + v f(w+rv) g'(w+rv)$$

$$\therefore R_v \{ f(w) g(w) \} = R_v \{ g(w) \} g(w) + f(w) R_v \{ g(w) \} \quad \blacksquare$$

$$4.e) H_v = \frac{1}{r} [\nabla_w(w+rv) - \nabla_w(w)] + O(r)$$

$$R_v \{ f(w) \} = \frac{d}{dr}(w+rv) \Big|_{r=0}$$

take the limit of  $H_v$  as  $r \rightarrow 0$ :

$$H_v = \lim_{r \rightarrow 0} \frac{\nabla_w(w+rv) - \nabla_w(w)}{r}$$

$$= \frac{d}{dr} \nabla_w(w+rv) \Big|_{r=0}$$

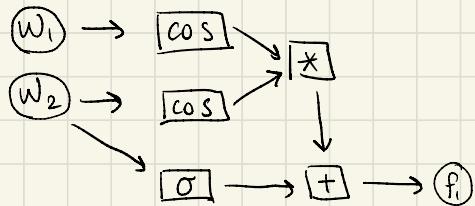
$$\Rightarrow H_v = R_v \{ \nabla_w(w) \} \blacksquare$$

4.f) We can apply the H<sub>v</sub> calculation in backpropagation of MLPs to derive  $R\{\text{backprop}\}$ . This allows us to avoid calculating the Hessian, and compute the gradient efficiently. With auto differentiation, we can use the implicit matrix products by using the forward-over-reverse method.

5. a)

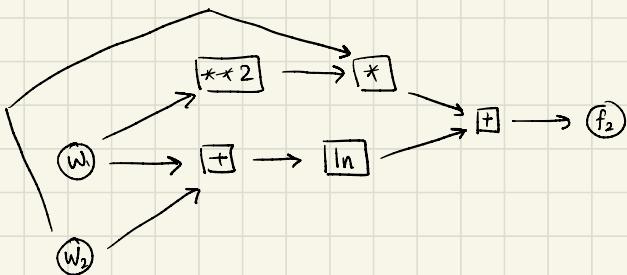
$$f_1(w_1, w_2) = \cos(w_1) \cos(w_2) + \sigma(w_2)$$

$$w = (1, 2)$$



$$f_1(w) = \cos(1) \cos(2) + \sigma(2) = 0.656$$

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



$$f_2(w) = \ln(1+2) + (1^2)(2) = 3.097$$

$$5.b) \quad \frac{df}{dw} = J = \begin{bmatrix} \frac{df_1}{dw_1} & \frac{df_1}{dw_2} \\ \frac{df_2}{dw_1} & \frac{df_2}{dw_2} \end{bmatrix} \quad \Delta w = 0.01$$

$$\frac{df_1}{dw_1} = \frac{f_1(1.01, 2) - f_1(1, 2)}{0.01} = 0.351$$

$$\frac{df_1}{dw_2} = \frac{f_1(1, 2.01) - f_1(1, 2)}{0.01} = -0.386$$

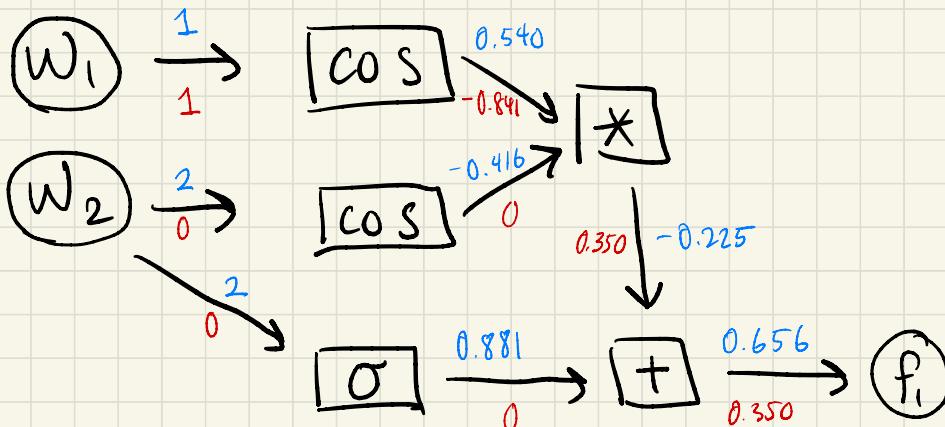
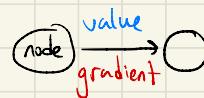
$$\frac{df_2}{dw_1} = \frac{f_2(1.01, 2) - f_2(1, 2)}{0.01} = 4.353$$

$$\frac{df_2}{dw_2} = \frac{f_2(1, 2.01) - f_2(1, 2)}{0.01} = 1.333$$

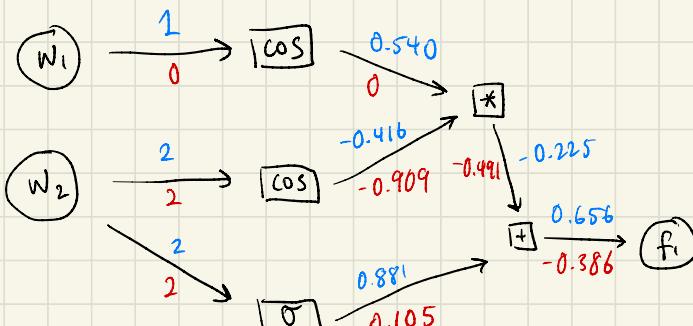
$$J = \begin{bmatrix} 0.351 & -0.386 \\ 4.353 & 1.333 \end{bmatrix}$$

5. c)

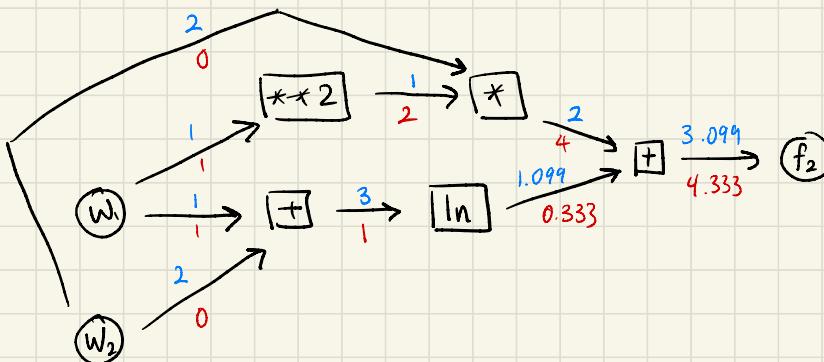
$$\text{Forward Mode AD } \frac{d}{dw_1} f_1(w_1, w_2)$$



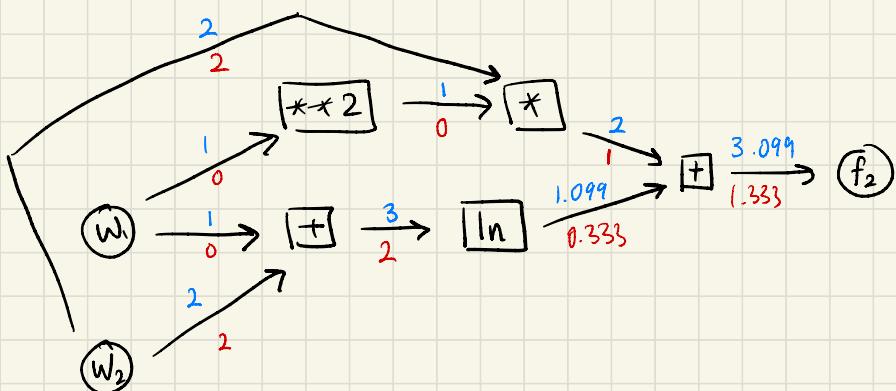
$$\text{Forward Mode AD } \frac{d}{dw_2} f_1(w_1, w_2)$$



$$\text{Forward Mode AD } \frac{d}{dw_1} f_2(w_1, w_2)$$

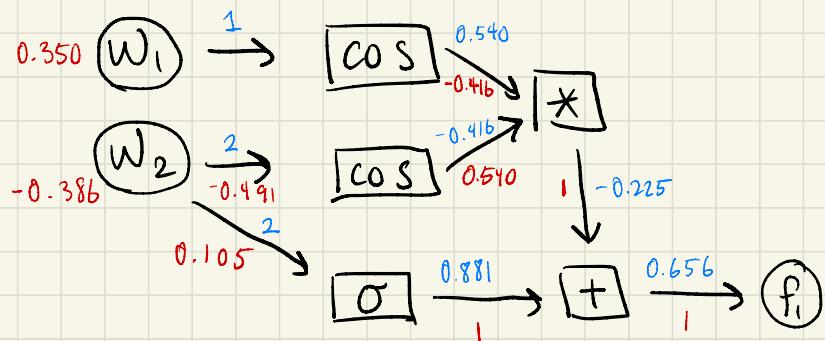


Forward Mode AD  $\frac{d}{dw_2} f_2(w_1, w_2)$

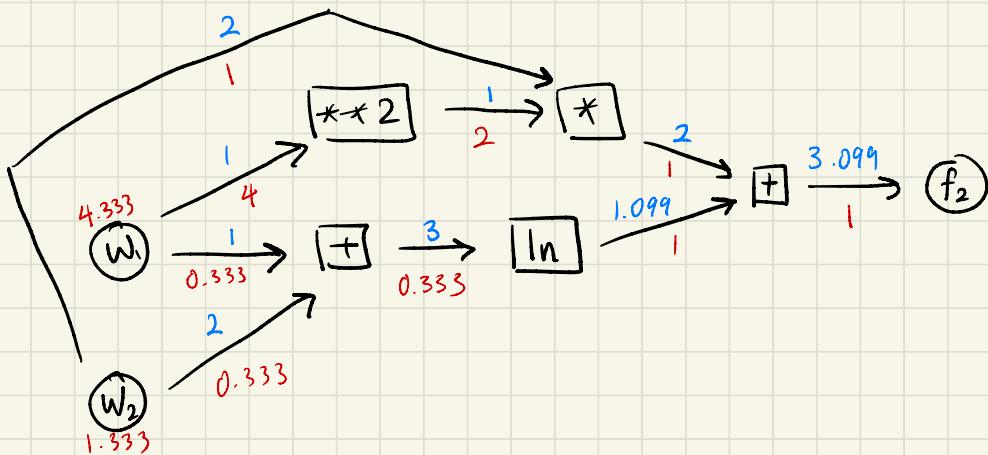


$$J = \begin{bmatrix} 0.333 & -0.333 \\ 1.333 & 1.333 \end{bmatrix}$$

5. d) Backward Mode AD  $\frac{d}{dw} f_1(w_1, w_2)$



Backward Mode AD  $\frac{d}{dw} f_2(w_1, w_2)$



$$J = \begin{bmatrix} 0.350 & -0.386 \\ 4.333 & 1.333 \end{bmatrix}$$

5.e) This is my 5<sup>th</sup>-6<sup>th</sup> time doing AD by hand  
and I'm thinking this is revenge of the machines

6. a)  $S$  is the circular right shift operator.

Let  $S^*$  be the circular left shift operator

$$Sx = \begin{bmatrix} 0 & & 1 \\ & \searrow & \\ 1 & & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} x_{n-1} \\ x_0 \\ \vdots \\ x_{n-2} \end{bmatrix}$$

$$S^*x = \begin{bmatrix} 0 & 1 & \\ & \searrow & \\ 1 & & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_0 \end{bmatrix}$$

From this it is clear that  $S^*$  is the transpose of  $S$ , thus also the adjoint  
Note both are circulant matrices

The actions of  $S$  and  $S^*$  can be expressed in terms of vector indices and  
modular arithmetic:

$$(Sx)_k = x_{k-1}, \quad (S^*x)_k = x_{k+1} \quad \text{where e.g. } (Sx)_0 = x_{n-1} \equiv_n x_{n-1}$$

The  $i,j^{\text{th}}$  entry of  $S$  can be expressed using the Kronecker delta:

$$(S)_{ij} = \delta_{i,j-1}$$

Let  $C_a$  be any circulant matrix where the entries of  $C_a$  are denote  $a_i$ :

$$(SC_a)_{ij} = \sum_e S_{ie} (C_a)_{ej} = \sum_e \delta_{i-e-1} a_{e-j} = \sum_e \delta_{(i-1)-e} a_{e-j} = a_{i-1-j}$$

$$(C_a S)_{ij} = \sum_e (C_a)_{ie} S_{ej} = \sum_e a_{i-e} \delta_{e-j-1} = \sum_e a_{i-e} \delta_{e-(j+1)} = a_{i-j-1} = a_{i-1-j}$$

We can see that  $SC_a = C_a S \quad \blacksquare$

6. b) Prove: Matrix  $M$  is circulant iff it commutes with the circular shift operator  $S$ .

$\Rightarrow$ : we proved the forward direction in part a)

$\Leftarrow$ : if  $M$  commutes with  $S$ , then  $M$  is circulant

$$SM = MS$$

We use the same Kronecker delta and index expression as in part a)

$$(SM)_{ij} = \sum_e S_{ij} (M)_{ej} = \sum_e \delta_{i-e-1} (M)_{ej} = (M)_{i-1,j}$$

$$(MS)_{ij} = \sum_e (M)_{ie} S_{ej} = \sum_e (M)_{ie} \delta_{e-j-1} = (M)_{i,j+1}$$

Since  $SM = MS$ :

$$\underbrace{(M)_{i-1,j}}_{\textcircled{1}} = \underbrace{(M)_{i,j+1}}_{\textcircled{2}} \iff \underbrace{(M)_{i-1,j-1}}_{\textcircled{2}} = \underbrace{(M)_{i,j}}_{\textcircled{3}} \iff \underbrace{(M)_{i,j}}_{\textcircled{3}} = \underbrace{(M)_{i+1,j+1}}_{\textcircled{3}}$$

Eq'n ③ implies that for any  $k$ ,  $(M)_{ij} = (M)_{i+k,j+k}$

$\Rightarrow$  Entries of  $M$  are constant along diagonals

take 1<sup>st</sup> column of  $M$  as  $m_i = (M)_{i,0}$

$$\Rightarrow (M)_{ij} = (M)_{i-j,j-j} = (M)_{i-j,0} = m_{i-j}$$

so all entries of  $M$  are obtained from the first column by circular shifts 

6. c) We can use circulant matrices to store weights in deep learning.  
Since circulant matrices have reduced number of independent parameters,  
we can do fast multiplication using it.

7. a) Any situation where digital hardware is reading continuous information, there could be errors in the readings. Such as reconstructing sounds from samples

$$7.b) \underset{z}{\operatorname{argmin}} E [L(z, y)]$$

$$L1 : \sum_{i=1}^n |z - y_i|$$

The minimum is found at median  $\{\vec{y}\}$

$$L_2 = \sum_{i=1}^n (z - y_i)^2$$

The minimum is found at the mean  
of  $\bar{y}$ :  $E_y\{\bar{y}\}$

$$7. d) D' = \{(x_i, y_i \pm \epsilon) | (x_i, y_i) \in D, \epsilon \sim N(0, \sigma^2)\}$$

Mean would stay the same

$$\mathbb{E}_y\{y\}$$

$$7. e) \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{(x,y)} [L(f_{\theta}(x), y)]$$

$$\mathbb{E}(x|Y) = \sum x \cdot P(x=x|y)$$

Law of joint distribution

$$P(A \cap B) = P(A) \cdot P(B|A)$$

$$\Rightarrow \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{(x,y)} [L(f_{\theta}(x), y)] = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_x [\mathbb{E}_{y|x} [L(f_{\theta}(x), y)]]$$

7.f)  $\underset{\theta}{\operatorname{argmin}} \sum_i L[f_\theta(\hat{x}_i, \hat{y}_i)]$  where both  $\hat{x}_i$  and  $\hat{y}_i$  are drawn from a corrupted distribution.

On the condition that the unobserved clean target  $\hat{y}_i$  s.t.  $E\{\hat{y}_i | \hat{x}_i\} = y_i$

From parts a-d, we know  $z = E_y \{y\}$  holds no matter the distribution from which  $y$ 's are drawn and the optimal network parameters of  $\underset{\theta}{\operatorname{argmin}} E_x \{E_{y|x} \{L(f_\theta(x), y)\}\}$  remain unchanged

7. g) Given infinite data, the soln to  $\operatorname{argmin}_{\theta} \sum_i L[f_{\theta}(\hat{x}_i), \hat{y}_i]$  is the same as  $\operatorname{argmin}_{\theta} \sum_i L(f_{\theta}(x_i), y_i)$ , which is the original given mapping between corrupt and clean observations

8. No.

Let  $w^{(1)} = 0$

$$\begin{aligned}f(w) &= \frac{1}{2}(w-2)^2 + \frac{1}{2}(w+1)^2 \\&= \frac{1}{2}(w^2 - 4w + 4) + \frac{1}{2}(w^2 + 2w + 1) \\&= w^2 - w + \frac{5}{2}\end{aligned}$$

$$f(w^{(1)}) = f(0) = \frac{5}{2}$$

Pick second term for next gradient descent

$$\begin{aligned}w^{(2)} &= w^{(1)} - \eta(w^{(1)} + 1) \\&= -\eta\end{aligned}$$

$$\begin{aligned}f(w^{(2)}) &= f(-\eta) \\&= \eta^2 + \eta + \frac{5}{2}\end{aligned}$$

$$\Rightarrow f(w^{(2)}) > f(w^{(1)})$$

$\Rightarrow$  SGD not guaranteed to decrease overall loss function in every iteration