(Q1.1.1)

assume dimensions:

$w_t = d \times 1$

$x_i = d \times 1$

Then,

$$\mathcal{L}(x_i, w_t) = \| w_t^T x_i - t_i \|_2^2$$

$$\nabla_{w_t} \mathcal{L}(x_i, w_t) = \underbrace{2(w_t^T x_i - t_i)}_{\in \mathbb{R}} x_i \quad , \text{ and } x_i \text{ is in rowspace of } X \quad (\star)$$

Assuming that we start from zero weight initialization

$w_0 = 0$, which is a linear combination of X,

the next weight updates are always in the row span of X,

since

$$w_{t+1} = \underbrace{w_t}_{\in \text{Span}\{X\}} - \frac{\eta}{b} \sum_{x_j \in B} \underbrace{\nabla_{w_t} \mathcal{L}(x_j, w_t)}_{\in \text{span}\{X\} \text{ by } (\star)}$$

is $\in \text{Span}\{X\}$ by induction.

So we can represent $\hat{w} = X^T a$ for some $a \in \mathbb{R}^n$, i.e. $\hat{w} \in \text{Span}(X)$.

Next, want to show that $\hat{w} = X^T a$ is indeed the minimum norm solution. This would be sufficient to show that $\hat{w} = w^*$ from gradient descent since w^* is the unique min. norm solution.

take any solution w', we have

$$(\hat{w} - w')^T \hat{w} = (\hat{w} - w')^T X^T a = \left(X(\hat{w} - w')\right)^T a = \left(X\hat{w} - Xw'\right)^T a = (t - t)^T a = 0$$

since $\hat{w}$ and w' are both solutions.

which means that $\hat{w} - w'$ is orthogonal to $\hat{w}$.

Then by the Pythagorean theorem,

$$\|w'\|_2^2 = \|\hat{w} - w' - \hat{w}\|_2^2 = \|\hat{w} - w'\|_2^2 + \|\hat{w}\|_2^2 \geq \|\hat{w}\|_2^2$$

(since they are orthogonal)

so $\hat{w}$ is the min norm solution.

and so $\hat{w} = w^*$ from gradient descent.

Q1.2.1)

Claim: RMSProp does not always obtain the minimum norm solution.

Counterexample:

It is sufficient to find a counterexample where RMSprop converges after one iteration at a $\hat{w}$ that does not have the minimum norm.

take $x_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $w_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $t = [2]$,
and hyperparameters β, ε, η.

Then $\nabla_{w,t=0} \mathcal{L} = 2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}^T \begin{bmatrix} 2 \\ 1 \end{bmatrix} - 2\right)\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} -8 \\ -4 \end{bmatrix}$

and $w_{t=1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{\eta}{\sqrt{\beta(0) + (1-\beta)(-8)^2}} \cdot (-8) \\ \frac{\eta}{\sqrt{\beta(0) + (1-\beta)(-4)^2}} \cdot (-4) \end{bmatrix} = \begin{bmatrix} \frac{\eta}{\sqrt{1-\beta}} \\ \frac{\eta}{\sqrt{1-\beta}} \end{bmatrix}$ assuming $\varepsilon = 0$

$\Rightarrow \nabla_{w,t=1} \mathcal{L} = 2\left(\begin{bmatrix} \frac{\eta}{\sqrt{1-\beta}} \\ \frac{\eta}{\sqrt{1-\beta}} \end{bmatrix}^T \begin{bmatrix} 2 \\ 1 \end{bmatrix} - 2\right)\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \left(\frac{4\eta}{\sqrt{1-\beta}} + \frac{2\eta}{\sqrt{1-\beta}} - 4\right)\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

assuming convergence here $\Rightarrow$ setting $\nabla_{w,t=1} \mathcal{L} = 0$.

$\Rightarrow \left(\frac{4\eta}{\sqrt{1-\beta}} + \frac{2\eta}{\sqrt{1-\beta}} - 4\right)\begin{bmatrix} 2 \\ 1 \end{bmatrix} \overset{\text{set}}{=} 0$

$\Rightarrow \frac{4\eta}{\sqrt{1-\beta}} + \frac{2\eta}{\sqrt{1-\beta}} - 4 = 0$

$\Rightarrow \frac{4\eta}{\sqrt{1-\beta}} + \frac{2\eta}{\sqrt{1-\beta}} = 4$

$\Rightarrow \frac{6\eta}{\sqrt{1-\beta}} = 4$

$\Rightarrow \eta = \frac{4\sqrt{1-\beta}}{6}$

So if we take $\beta = 0.99$, and $\eta = \frac{1}{15}$

then $w_{t=1} = \begin{bmatrix} \frac{\frac{1}{15}}{\sqrt{1-0.99}} \\ \frac{\frac{1}{15}}{\sqrt{1-0.99}} \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$, and $\nabla_{w,t=1} \mathcal{L} = 0$

$\Rightarrow w_{t=2} = w_{t=1}$ so we have reached convergence with $\hat{w} = \begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}$.

However, this is not the minimum norm weights since

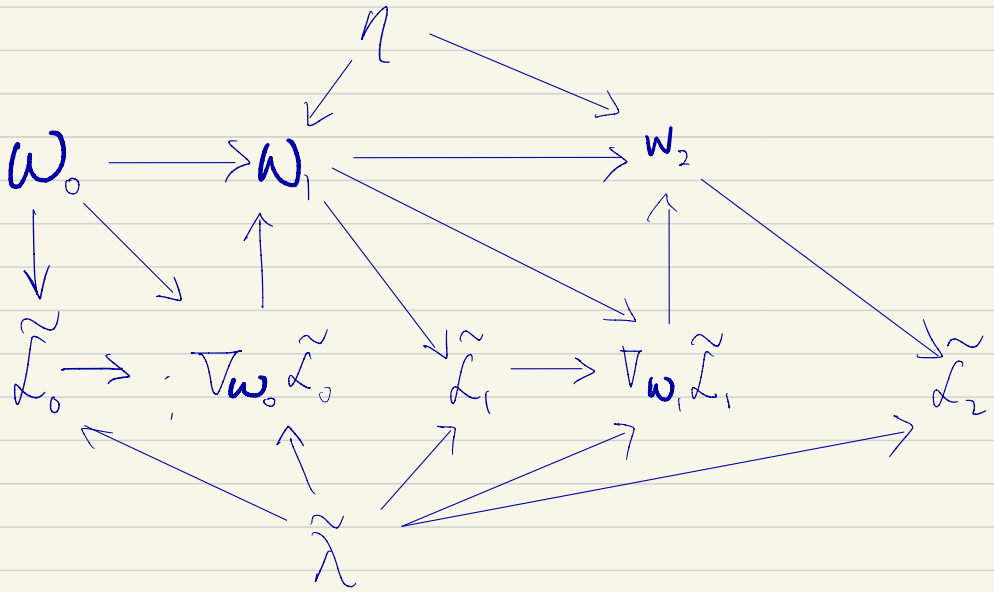
the norm of the gradient descent solution $\|w^*\|_2^2 = \left\|\begin{bmatrix} \frac{4}{5} \\ \frac{2}{5} \end{bmatrix}\right\|_2^2 = \frac{4}{5} < \|\hat{w}\|_2^2 = \left\|\begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}\right\|_2^2 = \frac{8}{9}$.

This is because the gradient updates for RMSprop do not necessarily lie in the span of X.

2.1.1)

$$\mathcal{L} = \frac{1}{n}\|X\hat{w} - t\|_2^2 + \tilde{\lambda}\|\hat{w}\|_2^2$$

$$\frac{d\mathcal{L}}{d\hat{w}} = \frac{2}{n}X^T(X\hat{w} - t) + 2\tilde{\lambda}\hat{w} = \frac{2}{n}(X^TX\hat{w} - X^Tt) + 2\tilde{\lambda}\hat{w}$$



2.1.2)

forward : $O(1)$

since each iteration only needs to access results from the previous iteration.

backward : $O(t)$

since each iteration, $\frac{\partial \mathcal{L}_t}{\partial \eta}$ requires information from all the iterations prior (ex, needs $w_1, w_2 \dots w_t$).

2.2.1)

$$\mathcal{L} = \frac{1}{n} \| X\hat{w} - t \|_2^2$$

$$w_1 = w_0 - \eta \nabla_{w_0} \tilde{\mathcal{L}}_0$$

$$= w_0 - \eta \left(\frac{2}{n} X^T (X w_0 - t) \right)$$

$$= w_0 - \frac{2\eta}{n} X^T a \qquad \text{, where } a = X w_0 - t$$

$$\mathcal{L}_1 = \frac{1}{n} \| X w_1 - t \|_2^2$$

$$= \frac{1}{n} \| X \left(w_0 - \frac{2\eta}{n} X^T a \right) - t \|_2^2$$

$$= \frac{1}{n} \| X w_0 - \frac{2\eta}{n} X X^T a - t \|_2^2$$

$$= \frac{1}{n} \| X w_0 - t - \frac{2\eta}{n} X X^T a \|_2^2$$

$$= \frac{1}{n} \| a - \frac{2\eta}{n} X X^T a \|_2^2$$

$$= \frac{1}{n} \left(a - \frac{2\eta}{n} X X^T a \right)^T \left(a - \frac{2\eta}{n} X X^T a \right)$$

$$= \frac{1}{n} \left(a^T - \left(\frac{2\eta}{n} X X^T a \right)^T \right) \left(a - \frac{2\eta}{n} X X^T a \right)$$

$$= \frac{1}{n} \left(a^T - \frac{2\eta}{n} a^T X X^T \right) \left(a - \frac{2\eta}{n} X X^T a \right)$$

$$= \frac{1}{n} a^T \left(I - \frac{2\eta}{n} X X^T \right) \left(I - \frac{2\eta}{n} X X^T \right) a$$

$$= \frac{1}{n} a^T \left(I - \frac{2\eta}{n} X X^T \right)^2 a$$

2.2.3)

$$\frac{\partial \mathcal{L}_1}{\eta} = \frac{\partial}{\eta}\left(\frac{1}{n} a^T \left(I - \frac{2\eta}{n} X X^T\right)^2 a\right)$$

$$= \frac{2}{n} a^T \left(I - \frac{2\eta}{n} X X^T\right)\left(-\frac{2}{n} X X^T\right) a$$

$$= \frac{4}{n^2} a^T \left(I - \frac{2\eta}{n} X X^T\right)\left(-X X^T\right) a$$

$$\overset{\text{set}}{=} 0$$

$$\Rightarrow a^T \left(I - \frac{2\eta^*}{n} X X^T\right)\left(-X X^T\right) a = 0$$

$$\Rightarrow a^T \left(-X X^T\right) a + a^T \left(-\frac{2\eta^*}{n} X X^T\right)\left(-X X^T\right) a = 0$$

$$\Rightarrow -a^T \left(X X^T\right) a + a^T \left(\frac{2\eta^*}{n} X X^T\right)\left(X X^T\right) a = 0$$

$$\Rightarrow a^T \left(\frac{2\eta^*}{n} X X^T\right)\left(X X^T\right) a = a^T \left(X X^T\right) a$$

$$\Rightarrow \frac{2\eta^*}{n} a^T X X^T X X^T a = a^T X X^T a$$

$$\Rightarrow \frac{2\eta^*}{n} \left\| X X^T a \right\|_2^2 = \left\| X^T a \right\|_2^2$$

$$\Rightarrow \frac{2\eta^*}{n} = \frac{\left\| X^T a \right\|_2^2}{\left\| X X^T a \right\|^2} \quad \text{(element wise division)}$$

$$\Rightarrow \eta^* = \frac{n}{2} \cdot \frac{\left\| X^T a \right\|_2^2}{\left\| X X^T a \right\|^2} \quad \text{, where } a = X w_0 - t$$

2.3.1)

$$\tilde{\mathcal{L}} = \frac{1}{n}\|Xw-t\|_2^2 + \tilde{\lambda}\|w\|_2^2 \quad , \quad \mathcal{L} = \frac{1}{n}\|Xw-t\|_2^2$$

$$\Rightarrow \frac{\partial \tilde{\mathcal{L}}}{\partial w} = \frac{2}{n}X^T(Xw-t) + 2\tilde{\lambda}w \quad , \quad \frac{\partial \mathcal{L}}{\partial w} = \frac{2}{n}X^T(Xw-t)$$

$\Rightarrow$ ① L2 regularization : $w_1 = w_0 - \eta \nabla_{w_0} \tilde{\mathcal{L}}_0$

$$= w_0 - \eta\left(\frac{2}{n}X^T(Xw_0-t) + 2\tilde{\lambda}w_0\right)$$

$$= w_0 - \eta\left(\frac{2}{n}X^T a + 2\tilde{\lambda}w_0\right) \quad \text{, where } a = Xw_0 - t$$

and ② weight decay : $w_1 = (1-\lambda)w_0 - \eta \nabla_{w_0}\mathcal{L}_0$

$$= (1-\lambda)w_0 - \eta\left(\frac{2}{n}X^T(Xw_0-t)\right)$$

$$= (1-\lambda)w_0 - \frac{2\eta}{n}X^T a \quad \text{, where } a = Xw_0 - t$$

2.3.2)

set $w_0 - \eta\left(\frac{2}{n}X^T a + 2\tilde{\lambda}w_0\right) = (1-\lambda)w_0 - \frac{2\eta}{n}X^T a$

$$\Rightarrow w_0 - \frac{2\eta}{n}X^T a - 2\eta\tilde{\lambda}w_0 = w_0 - \lambda w_0 - \frac{2\eta}{n}X^T a$$

$$\Rightarrow 2\eta\tilde{\lambda}w_0 = \lambda w_0$$

$$\Rightarrow 2\eta\tilde{\lambda} = \lambda$$

$$\Rightarrow \boxed{\tilde{\lambda} = \frac{\lambda}{2\eta}}$$

Q3.1

$$I = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad J = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\Rightarrow I * J = \begin{bmatrix} 0 & -1 & -2 & -3 & -2 \\ -2 & -3 & -3 & -2 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ 2 & 2 & 2 & 1 & 1 \\ 1 & 2 & 3 & 2 & 1 \end{bmatrix}$$

This filter detects horizontal edges. Negative values in output means original image is dark (small values) above and light (large values) below, while postive values in output means original image is light (large values) above and dark (small values) below.

Q3.2) assuming no bias,
assuming that we don't count input image as # neurons.

CNN

neurons : $\underbrace{(32)(32)}_{C_1} + \underbrace{(16)(16)}_{P_1} + \underbrace{(16)(16)}_{C_2} + \underbrace{(8)(8)}_{P_2} + \underbrace{(8)(8)}_{C_3} = \boxed{1664}$

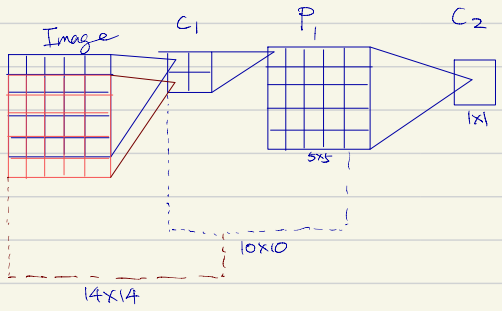
trainable parameters : $\underbrace{(3)(3)}_{k_1} + \underbrace{(3)(3)}_{k_2} + \underbrace{(3)(3)}_{k_3} = \boxed{27}$

FCNN :

neurons : $\underbrace{(32)(32)}_{FC_1} + \underbrace{(16)(16)}_{P_1} + \underbrace{(16)(16)}_{FC_2} + \underbrace{(8)(8)}_{P_2} + \underbrace{(8)(8)}_{FC_3} = \boxed{1664}$

trainable parameters : $\underbrace{(32)(32)(32)(32)}_{W_1} + \underbrace{(16)(16)(16)(16)}_{W_2} + \underbrace{(8)(8)(8)(8)}_{W_3} = \boxed{1,118,208}$

(Q3.3)



Assuming a filter size of 5x5 for the convolution layers, then:

→ every neuron after the second convolution layer has a receptive field of a 5x5 patch of neurons from the result of the first pooling layer.

→ every neuron from the result of the first pooling layer has a receptive field of a 2x2 patch of neurons from the result of the first convolution layer.

↳ since there are 5x5 neurons from the result of the first pooling layer to account for, the receptive field for the result of the first convolution layer must be a 10x10 patch of neurons. (since pooling has stride=2, no overlap)

→ the receptive field of the 10x10 patch of neurons on the result of the first convolution layer must then have a receptive field of a 14x14 neuron patch since the convolution has filter size 5x5 and stride 1.

since $10 = (W-5)+1 \Rightarrow W = 14$

⇒ So in conclusion, the receptive field of a neuron after the second convolution layer is **14 x 14**.

Other things that can affect the size of the receptive field:

① the stride of the max pooling layer
② the stride of the convolution filter in convolution layers