Question-1

(a)

(i) $\left\lfloor \frac{28+2(0)-3}{1} \right\rfloor + 1 = 26$

output shape would be [32, 128, 26, 26]

(ii) $\left\lfloor \frac{28+2-4}{2} \right\rfloor + 1 = 14$

output shape would be [32, 128, 14, 14]

(b) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$

(d) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

(e) Input shape [10, 1, 32, 32]

Layer a: [10, 16, 32, 32]

Layer b: [10, 32, 16, 16]

Layer c: [10, 128, 8, 8]

Layer d: [10, 16, 8, 8]

Layer e [10, 8, 16, 16]

Layer f [10, 1, 32, 32]

f) 10, 24, 32, 32

g) layer f => all ones with layer e all zeros

(h) Recursive Equation $\boxed{r_i = s_{i-1} \cdot r_{i-1} + (k_i - s_{i-1})}$

$r_a = 1$

$r_b = 3$

$r_c = 7$

$\boxed{r_d = 15}$ → so, receptive field is 15x15

Q2 (a) Learning Rate A is too high!

(b) place optim.zero-grad() before backward pass.

```
for images, labels in train-loader:
    optim.zero-grad()
    preds = model(images)

    loss = loss-fun(labels, preds)
    loss.backward()
    optim.step()
```

Q 3

Solution: C and D

Q4

	This	exam	is	a	midterm	exam
This		1				
exam			1			
is				1		
a					1	
midterm						1
exam						

Q7 Optimizers & their convergence :

ⓐ $\theta_{t+1} \leftarrow \theta_t - \alpha_t M_t \nabla f_t(\theta_t)$

stepsize

loss

recomputed over each epoch of training and just consists of a diagonal populated by the inverses of the square roots of the mean squared value for the gradients during the epoch for that specific coordinate

$n=1$

$$[1 \;\; 0.1 \;\; 0.01]\,\theta = 1$$

$\theta \downarrow \mathbb{R}^{3\times1}$

$$\theta_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$f_t(\theta) = \left(1 - [1 \;\; 0.1 \;\; 0.01]\,\theta\right)^2$$

ⓐ for $M_t = I$,

what vector θ would standard vanilla SGD converge to ?

→ least squares solution.

$$\nabla f_t(\theta) = 2\left(1 - [1 \;\; 0.1 \;\; 0.01]\,\theta\right)\begin{bmatrix} 1 \\ 0.1 \\ 0.01 \end{bmatrix} = 0.$$

$$= 2\begin{bmatrix} 1 \\ 0.1 \\ 0.01 \end{bmatrix} - 2\theta\left[1 + 0.1^2 + 0.01^2\right] = 0.$$

$$\cancel{2}\begin{bmatrix} 1 \\ 0.1 \\ 0.01 \end{bmatrix} = \cancel{2}\theta\left(1 + 0.1^2 + 0.01^2\right)$$

$$\theta = \begin{bmatrix} 1 \\ 0.1 \\ 0.01 \end{bmatrix} \cdot \left(\frac{1}{1 + 0.1^2 + 0.01^2}\right)$$

(b)

$$\theta_{t+1} \leftarrow \theta_t - \alpha_t \underline{M_t} \nabla f_t(\theta_t)$$

$$f_t(\theta) = (1 - [1, 0.1, 0.01]\theta)^2$$

~~$M_t \Rightarrow$~~

$$\nabla f_t(\theta_t) = \left(2(1 - [1 \;\; 0.1 \;\; 0.01]\theta)\begin{bmatrix} 1 \\ 0.1 \\ 0.01 \end{bmatrix}\right)$$

$$\nabla f_t(\theta_t) = \left(2\begin{bmatrix} 1 \\ 0.1 \\ 0.01 \end{bmatrix} - (1 + 0.1^2 + 0.01^2)\theta_t\right)$$

$$M_t = \frac{1}{\sqrt{\text{RMSE of gradients}}}$$

$$M_t = \begin{bmatrix} 1/\sqrt{1} & 0 & 0 \\ 0 & 1/\sqrt{0.1^2} & 0 \\ 0 & 0 & 1/\sqrt{0.01^2} \end{bmatrix}$$

multiply by $\nabla f_t(\theta_t)$

$$= 2(1 - [1 \;\; 0.1 \;\; 0.01]\theta \cdot \underset{3\times3}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{0.1^2} & 0 \\ 0 & 0 & 1/\sqrt{0.01^2} \end{bmatrix}} \underset{3\times1}{\begin{bmatrix} 1 \\ 0.1 \\ 0.01 \end{bmatrix}}$$

$$= 2(1 - [1 \;\; 0.1 \;\; 0.01]\theta)\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 0$$

$$\cancel{2}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \cancel{2}\theta(1 + 0.1 + 0.01)$$

$$\theta = \frac{1}{1.11}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

ⓒ $\mathcal{L}(\theta) = (\theta - \theta^*)^T \tilde{X}^T \tilde{X} (\theta - \theta^*)$

what should $\tilde{X}$ be?

Recall, $E[\mathcal{L}(\theta_{t+1}) | \theta_t] < (1-\rho)\mathcal{L}(\theta_t)$

$$
\begin{aligned}
\mathcal{L}(\theta) &= E[f_t(\theta)] \\
&= E[(y_I - x_I^T \theta)^2] \\
&= \sum_i^n p_i (y_i - x_i^T \theta)^2 \\
&= \sum_i p_i (x_i^T \theta^* - x_i^T \theta)^2 \\
&= \sum_i p_i (x_i^T (\theta^* - \theta))^2 \\
&= \sum_i p_i (\theta^* - \theta)^T x_i x_i^T (\theta^* - \theta) \\
&= (\theta^* - \theta)^T (\sum_i \sqrt{p_i}\, x_i)(\sqrt{p_i}\, x_i^T)(\theta^* - \theta) \\
&= (\theta^* - \theta)^T \tilde{X}^T \tilde{X} (\theta^* - \theta)
\end{aligned}
$$

where $\tilde{X} = \begin{bmatrix} \sqrt{p_1} \cdot x_1^T - \\ \sqrt{p_2}\ - x_2^T - \\ \vdots \\ \sqrt{p_n}\ - x_n^T - \end{bmatrix}$

(c) (3)

$[1 \;\; 0.1 \;\; 0.01]\,\theta = 1$

Rescaling => $[1 \;\; 1 \;\; 1]\,\bar{\theta} = 1$

solution => $\frac{1}{3}\begin{bmatrix}1\\1\\1\end{bmatrix}$

final answer => $\frac{1}{3}\begin{bmatrix}1\\100\\1000\end{bmatrix}$

$\longleftrightarrow$

(b) nonuniform sampling of training points for SGD:

$X\theta = y$

$\begin{bmatrix}\quad & \quad & \quad\end{bmatrix}_{n\times d} \quad \begin{bmatrix}\,\end{bmatrix}_{n\times 1}$

where $d > n$

full row rank

$\begin{bmatrix}p_1\\p_2\\\vdots\\p_n\end{bmatrix} \odot \begin{bmatrix}— x_1^T —\\— x_2^T —\\\vdots\\— x_n^T —\end{bmatrix}_{n\times d}$

constant term eventually cancels out so, we should get the same solution

$f_t(\theta) = (\,y_{i(t)} - x_{i(t)}^T\theta)^2$

where data point $i(t)$ is chosen with probability $p_i > 0$

where $\sum_{i=1}^{n} p_i = 1$

$\theta_0 = 0$

If SGD converges with a constant step size α, what solution θ^* must SGD converge to?

Convergence to min-norm solution

i.e., $\theta^* = X^T(XX^T)^{-1}Y$

(d)

$$\mathcal{L}(\theta_{t+1}) = \mathcal{L}(\theta_t) + A + B$$

$$E[A/\theta_t] = -4\alpha(\theta_t - \theta^*)^T \tilde{X} + \tilde{X} E[x_I x_I^T](\theta_t - \theta^*)$$

$$E[B/\theta_t] = 4\alpha^2(\theta_t - \theta^*)^T E[x_I x_I^T \tilde{X} + \tilde{X} x_I x_I^T](\theta_t - \theta^*)$$

$$= -4\alpha(\theta_t - \theta^*)^T \tilde{X}^T \tilde{X} \tilde{X}^T \cancel{\tilde{X}}(\theta_t - \theta^*)$$

$$\leq -4T_{min}\,\alpha(\theta_t - \theta^*)^T \tilde{X}^T \tilde{X}(\theta_t - \theta^*)$$

$$\leq -4T_{min}\,\alpha\,\mathcal{L}(\theta_t)$$

$$E[B/\theta_t] = 4\alpha^2(\theta_t - \theta^*)^T E[x_I x_I^T(\tilde{X}^T X)x_I x_I^T](\theta_t - \theta^*)$$

$$\leq T_{max}\,4\alpha^2(\theta_t - \theta^*)^T E[x_I x_I^T x_I x_I^T](\theta_t - \theta^*)$$

$$\leq T_{max}\,4\alpha^2((\max_i \|x_i\|^2)\alpha^2(\theta_t - \theta^*)(\tilde{X}^T\tilde{X})(\theta_t - \theta^*)$$

$$= T_{max}(\max_i \|x_i\|^2)\alpha^2\,\mathcal{L}(\theta_t)$$