(1) $R = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ $\quad p = [0.5 \quad 0.25]^T$

(a) $w_o = R^{-1} p$

$$= \frac{1}{1-0.25} \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix}$$

$$= \frac{1}{0.75} \begin{bmatrix} 0.5 - 0.5(0.25) \\ -0.5(0.5) + 0.25 \end{bmatrix} = \frac{1}{0.75} \begin{bmatrix} 0.375 \\ 0 \end{bmatrix}$$

$$\boxed{W_o = \begin{bmatrix} \frac{0.375}{0.75} \\ 0 \end{bmatrix}} \longrightarrow \text{Tap weights}$$

(b) $E_{out_{min}} = \sigma_d^2 - p^H R^{-1} p$

We know from (a) that $w_o = R^{-1} p$

$$\Rightarrow E_{out_{min}} = \sigma_d^2 - [0.5 \quad 0.25] \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$$

$$\Rightarrow \boxed{E_{out_{min}} = \sigma_d^2 - 0.25}$$

(c) We know $w_o = R^{-1} p$

using eigendecomposition $R = Q \Omega Q^H$

$$\Rightarrow w_o = (Q \Omega Q^H)^{-1} p$$

$$w_o = (Q^H \Omega^{-1} Q) p$$

If Ω is diagnol containing $\lambda_1, \lambda_2 \ldots \lambda_R$ eignen values

Then Ω^{-1} is diagnol containing $1/\lambda_1, 1/\lambda_2 \ldots 1/\lambda_R$ eigen values

$$\text{i.e } W_o = \begin{bmatrix} q_1^H \\ q_2^H \end{bmatrix} \begin{bmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{bmatrix} [q_1^* \quad q_2^*] [0.5 \quad 0.25]^T$$

$$\Rightarrow \quad w_0 = \sum_{i=1}^{2} \frac{1}{\lambda_i} q_i q_i^H \cdot p \quad \text{———} \quad (1)$$

To find eigen values

$$|R - \lambda I| = 0$$

$$\begin{vmatrix} 1-\lambda & 0.5 \\ 0.5 & 1-\lambda \end{vmatrix} = 0 \quad \Rightarrow (1-\lambda)^2 - 0.25 = 0$$

$$1 + \lambda^2 - 2\lambda - 0.25 = 0$$

$$\lambda^2 - 2\lambda + 0.75 = 0$$

$$\lambda_1 = 1.5 \qquad \lambda_2 = 0.5$$

$$\Rightarrow \quad (R - \lambda I) Q = 0$$

For $\lambda = 0.5$

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = 0 \quad \Rightarrow \quad q_1 = -q_2$$

$$q_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \qquad q_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

(1) becomes

$$\Rightarrow \quad w_0 = \frac{1}{1.5} \cdot \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} + \frac{1}{0.5} \cdot \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix}$$

$$\Rightarrow \quad w_0 = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \end{bmatrix}$$

$$\Rightarrow \quad \boxed{w_0 = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}}$$

(4) logistic function(θ) bounds output b/w 0 and 1

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$$

where $s = w^T x$ is the signal

larges signal $\Rightarrow$ $\theta(s)$ tends to be close to 1

smaller signal $\Rightarrow$ $\theta(s)$ " " " " " 0

$\therefore$ output b/w 0 & 1 it can be interpreted as probability

$$P[y|x] = \begin{cases} f(x) & \text{for } y = +1 \\ 1-f(x) & \text{for } y = -1 \end{cases}$$

$$\theta(-s) = 1 - \theta(s)$$

and $\theta(s) = \theta(s)$

Hence $P[y|x] = \theta(y\, w^T x)$

Likelihood of dataset $(x_1,y_1)(x_2,y_2) - \cdots - (x_N, y_N)$ is

$$\prod_{n=1}^{N} P[y_n|x_n] = \prod_{n=1}^{N} \theta(y_n w^T x_n)$$

using MLE

$$\hat{w} = \underset{w}{\mathrm{argmax}} \prod_{n=1}^{N} \theta(y_n w^T x_n)$$

maximizing $f(x)$, we can also maximize $\ln(f(x))$

$$\Rightarrow \hat{w} = \underset{w}{\mathrm{argmax}} \frac{1}{N} \ln\left(\prod_{n=1}^{N} \theta(y_n w^T x_n)\right)$$

or

$$\hat{w} = \underset{w}{\mathrm{argmin}} -\frac{1}{N} \ln\left(\prod_{n=1}^{N} \theta(y_n w^T x_n)\right)$$

$$\hat{w} = \underset{w}{\mathrm{argmin}} \frac{1}{N} \sum_{n=1}^{N} \ln\left(\frac{1}{\theta(y_n w^T x_n)}\right)$$

$\because \; \theta(s) = 1/1+e^{-s}$

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln\left(1 + e^{-y_n w^T x_n}\right)$$

Minimize $E_{in}(w)$ means move in ~~-ve~~ direction -ve of gradient of $E_{in}(w)$ i.e $-\nabla E_{in}(w)$

$$\Rightarrow \nabla_w E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}} \nabla_w \left(1+e^{-y_n w^T x_n}\right)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}} \cdot e^{-y_n w^T x_n} \cdot \nabla_w (-y_n w^T x_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \frac{e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}} (-y_n x_n)$$

dividing by $e^{-y_n w^T x_n}$

$$\nabla_w E_{in}(w) = -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n x_n}{1+e^{y_n w^T x_n}} \quad \text{——— } *$$

As $\theta(s) = \dfrac{1}{1+e^{-s}}$ ——— $**$

$$\frac{1}{1+e^{y_n w^T x_n}} = \theta(-y_n w^T x_n)$$ substitute

$$\Rightarrow \boxed{\nabla_w E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} -y_n x_n \, \theta(-y_n w^T x_n)}$$

Why missclassified contributes more?

$\Rightarrow$ consider ~~true~~ ~~y~~ a missclassified example with true ~~y = +1~~ $y = -1$ and signal $w^T x = +ve \Rightarrow y_n w^T x_n = -ve$

$\Rightarrow$ $1+e^{y_n w^T x_n}$ is small $\Rightarrow *$ is large i.e gradient is more -ve. Hence it means a missclassified example contributes more to gradient. Same can be argued for reverse case i.e $y = +1$ and signal $w^T x = -ve$

(5) $x_1, x_2, x_3 \ldots\ldots x_n$ are i.i.d Poisson Distribution

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, 2 \ldots\ldots \infty$$

use MLE i.e choose parameters which maximize or make the observed data most likely

Likelihood of data $D \rightarrow (x_1, y_1) \ldots\ldots (x_n\, y_n)$
or here $x_1, x_2 \ldots\ldots x_n$

$$f(k|\lambda) = \prod_{i=1}^{n} f(k_i|\lambda)$$

using MLE

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{n} f(k_i|\lambda)$$

$$= \underset{\lambda}{\operatorname{argmax}} \ln\left(\prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k!}\right)$$

$$= \sum_{i=1}^{n} \ln(\lambda^{k_i} e^{-\lambda}) - \sum_{i=1}^{n} \ln(k_i!)$$

$$\frac{\partial}{\partial\lambda}\left(\ln\lambda \sum_{i=1}^{n} k_i - n\lambda - \sum_{i=1}^{n} \ln(k_i!)\right) = 0$$

$$\Rightarrow \frac{1}{\lambda} \sum_{i=1}^{n} k_i - n = 0$$

$$\Rightarrow \boxed{\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} k_i} \longrightarrow \text{MLE of } \lambda$$