H.W 4.

↳ Hardik Prajapati (2678294168)

1.

$X \sim Poiss(\lambda)$ $\qquad$ $p(x_i|\lambda) = \dfrac{\lambda^{x_i} e^{-\lambda}}{x_i!}$

$\lambda \sim \Gamma(2,1)$ $\qquad$ $p(\lambda) = \lambda e^{-\lambda}$

$$p(\lambda|D) \propto p(D|\lambda)\, p(\lambda)$$

Now, $p(D|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda)$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$= \frac{(e^{-\lambda})^n \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

$\Rightarrow$ Maximize $p(D|\lambda)p(\lambda)$

$$L = \frac{(e^{-\lambda})^n \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} (\lambda e^{-\lambda})$$

$$l = \log L = -\lambda n + \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log(x_i!) + \log \lambda - \lambda$$

$\Rightarrow \dfrac{\partial l}{\partial \lambda} = 0$

$\Rightarrow -n + \dfrac{\sum_{i=1}^{n} x_i}{\lambda} - 0 + \dfrac{1}{\lambda} - 1 = 0$

$$\Rightarrow \frac{1}{\lambda}\left\{\sum_{i=1}^{n} x_i + 1\right\} = n+1$$

$$\Rightarrow \hat{\lambda}_{MAP} = \frac{n\bar{x}+1}{n+1}$$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Now, $\left.\frac{\partial^2 \ell}{\partial \lambda}\right|_{\lambda=\hat{\lambda}_{MAP}} < 0$

$$\Rightarrow \left.\frac{\partial^2 \ell}{\partial \lambda}\right|_{\lambda=\hat{\lambda}_{MAP}} = \left.\left[-\frac{\sum_{i=1}^{n} x_i}{\lambda^2} - \frac{1}{\lambda^2}\right]\right|_{\lambda=\hat{\lambda}_{MAP}}$$

$$= \left.\left[-\frac{1}{\lambda^2}\left\{n\bar{x}+1\right\}\right]\right|_{\lambda=\hat{\lambda}_{MAP}}$$

$$= \frac{-\{n\bar{x}+1\}}{\left(\frac{n\bar{x}+1}{n+1}\right)^2}$$

$$= -\frac{(n+1)^2}{(n\bar{x}+1)}$$

$$< 0$$

$\Rightarrow \hat{\lambda}_{MAP} = \frac{n\bar{x}+1}{n+1}$ maximizes the posterior

2.

$$L(\lambda|D) = \prod_{i=1}^{n} p(x_i|\lambda)$$

$$\Rightarrow L(\lambda|D) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

$$\Rightarrow l = \text{Log } L = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

$$\frac{\partial l}{\partial \lambda} = 0$$

$$\Rightarrow \frac{n\bar{x}}{\lambda} - n = 0$$

$$\Rightarrow \hat{\lambda}_{MLE} = \bar{x}$$, where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Fisher information matrix:-

$$i(\lambda) = E_{x|\lambda}[I(\lambda)]$$

$$= -E_{x|\lambda}\left[\frac{\partial^2 l(\lambda|D)}{\partial^2\lambda \, \partial^2\lambda^T}\right]$$

$$= \sum_{i=1}^{n} -E_{x|\lambda}\left[-\frac{n\bar{x}}{\lambda^2}\right]$$

$$= \sum_{i=1}^{n} \frac{n\bar{x}}{\lambda^2}$$

$$\Rightarrow \ i(\lambda) = \frac{n^2 \bar{x}}{\lambda^2}$$

$$\Rightarrow \ \hat{\lambda}_n \xrightarrow{d} N\left(\lambda, [i(\lambda)]^{-1}\right) \quad , \text{as } n \to \infty$$

$$\Rightarrow \ \hat{\lambda}_n \xrightarrow{d} N\left(\lambda, \frac{\lambda^2}{n^2 \bar{x}}\right)$$

By central Limit theorem,

$$\hat{\lambda}_n \xrightarrow{d} N\left(\lambda, \frac{\lambda}{n}\right)$$

Q3.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$\Rightarrow \ Y = w^T x + \varepsilon$$

$$\Rightarrow \ Y \sim N(w^T x, \sigma^2)$$

$$\Rightarrow \ P(y \mid x, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - w^T x)^2}{2\sigma^2}}$$

$$\text{Log } L(w) = \log P(D \mid w) = \log P(Y \mid X, w)$$

$$\Rightarrow \ \log L(w) = \log \prod_{i=1}^{n} P(y_i \mid x_i, w)$$

$$\Rightarrow \log L(w) = \sum_{i=1}^{n} \log P(y_i | x_i, w)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}}$$

$$= \sum_{i=1}^{n} \left\{ \frac{-1}{2} \log(2\pi\sigma^2) - \frac{(y_i - w^T x_i)^2}{2\sigma^2} \right\}$$

Maximizing $\log L(w)$

$$\hat{w}_{MLE} = \arg\max_{w} \log L(w)$$

$$= \arg\max_{w} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

$$= \arg\min_{w} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

Now, for $\sigma = 1$ for each input, it's equivalent to the least-squares.

4.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad , \varepsilon \sim N(0, \sigma^2)$$

$$p(\beta_i) \sim N(0, \sigma^2/\lambda) \Rightarrow p(w) = \frac{1}{\left(2\pi \frac{\sigma^2}{\lambda}\right)^{p/2}} e^{-\frac{\lambda}{2}\frac{(w^T w)}{\sigma^2}}$$

$$\Rightarrow \log L(w) = \log P(D|w) + \log P(w)$$

$$\Rightarrow \log L(w) = \log \prod_{i=1}^{n} P(y_i | x_i, w) + \log P(w)$$

$$= \sum_{i=1}^{n} \log P(y_i | x_i, w) + \log P(w)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}}\right) + \log\left(\frac{1}{\left(2\pi\frac{\sigma^2}{\lambda}\right)^{p/2}} e^{-\frac{\lambda}{2}\frac{(w^T w)}{\sigma^2}}\right)$$

$$= \sum_{i=1}^{n}\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - w^T x_i)^2}{2\sigma^2}\right\} + \left\{-\frac{p}{2}\log\left(2\pi\frac{\sigma^2}{\lambda}\right) - \frac{\lambda(w^T w)}{2\sigma^2}\right\}$$

Maximizing $\log L(w)$

$$\hat{W}_{MAP} = \arg\max_w \log P(w|D)$$

$$= \arg\max_w \left\{\sum_{i=1}^{n} -\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{\lambda(w^T w)}{2\sigma^2}\right\}$$

$$= \arg\min_w \left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \frac{\lambda}{2\sigma^2} w^T w\right\}$$

5.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

$$\beta_i \sim \mathrm{Lap}\left(0, \frac{\sigma^2}{\lambda}\right) \Rightarrow p(w) = \frac{e^{-\frac{\sqrt{2\lambda}}{\sigma}|w|}}{\frac{2\sigma}{\sqrt{2\lambda}}} = \frac{\sqrt{\lambda}\, e^{-\frac{\sqrt{2\lambda}}{\sigma}|w|}}{\sqrt{2}\,\sigma}$$

$$2\beta^2 = \frac{\sigma^2}{\lambda}$$

$$\Rightarrow \beta = \frac{\sigma}{\sqrt{2\lambda}}$$

$$\log L(w) = \log P(D|w) + \log P(w)$$

$$\Rightarrow \log L(w) = \log \prod_{i=1}^{n} P(y_i | x_i, w) + \log P(w)$$

$$= \sum_{i=1}^{n} \log P(y_i | x_i, w) + \log P(w)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}}\right) + \log\left(\sqrt{\frac{\lambda}{2}}\, \frac{e^{-\frac{\sqrt{2\lambda}}{\sigma}|w|}}{\sigma}\right)$$

$$= \sum_{i=1}^{n}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) + \frac{1}{2}\log\frac{\lambda}{2} - \frac{\sqrt{2\lambda}|w|}{\sigma} - \log\sigma$$

Maximizing $\log L(w)$

$$\hat{W}_{MAP} = \arg\max_w \log P(w|D)$$

$$= \arg\max_w \left\{ \sum_{i=1}^{n}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) + \frac{1}{2}\log\frac{\lambda}{2} - \frac{\sqrt{2\lambda}}{\sigma}|w| - \log\sigma \right\}$$

$$= \arg\max_w \left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 - \frac{\sqrt{2\lambda}}{\sigma}|w| \right\}$$

Q6. a) $\hat{\beta}^R = (X^TX + \lambda I)^{-1} X^T y$

Now, $X = U\Sigma V^T$

$$\hat{y} = X\hat{\beta}^{Ridge}$$

$$= X(X^TX + \lambda I)^{-1} X^T y$$

$$= U\Sigma V^T (V\Sigma^T \underbrace{U^T U}_{I} \Sigma V^T + \lambda I)^{-1} V\Sigma^T U^T y$$

$$= U\Sigma V^T (V \underbrace{\Sigma^T \Sigma}_{\sigma^2 I} V^T + \lambda I)^{-1} V\Sigma^T U^T y$$

$$= U\Sigma V^T (V\sigma^2 I V^T + \lambda I)^{-1} V\Sigma^T U^T y$$

$$= U\Sigma V^T (\sigma^2 I V V^T + \lambda I)^{-1} V\Sigma^T U^T y$$

$$= U\Sigma V^T ((\sigma^2 + \lambda) I)^{-1} V\Sigma^T U^T y$$

$$= \frac{1}{(\sigma^2 + \lambda)} U\Sigma V^T V\Sigma^T U^T y$$

$$= \frac{U\Sigma\Sigma^T U^T y}{(\sigma^2 + \lambda)}$$

$$= \left(\frac{\sigma^2}{\sigma^2 + \lambda}\right) U U^T y$$

$$= \sum_{j=1}^{p} \vec{u}_j \left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right) \vec{u}_j^T \vec{y}$$

$\Rightarrow$ Can be seen that with smaller σ_j, $\left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right)$ will be $\uparrow$ higher & hence greater amount of shrinkage to respective $\vec{u}_j$.

Q6. b

$$tr\left[X(X^TX+\lambda I)^{-1}X^T\right]$$

$$= tr\left[U\Sigma V^T\left(V\Sigma^T U^T U\Sigma V^T + \lambda I\right)^{-1} V\Sigma^T U^T\right]$$

$$= tr\left[U\Sigma V^T\left(V\sigma^2 I V^T + \lambda I\right)^{-1} V\Sigma^T U^T\right]$$

$$= tr\left[U\Sigma V^T\left(\{\sigma^2+\lambda\} I\right)^{-1} V\Sigma^T U^T\right]$$

$$= tr\left[\frac{U\Sigma V^T V\Sigma^T U^T}{(\sigma^2+\lambda)}\right]$$

$$= tr\left[\frac{U\Sigma\Sigma^T U^T}{(\sigma^2+\lambda)}\right]$$

$$= tr\left[\left(\frac{\sigma^2}{\sigma^2+\lambda}\right) I\right]$$

$$= \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2+\lambda}$$