

### QUESTION 3

ABHAY SHANKAR K: CS21BTECH11001 & KARTHEEK TAMMANA: CS21BTECH11028

- (I) **Question:** Derive the expression of likelihood and prior for a heteroscedastic setting for a single data point with input  $\mathbf{x}_n$  and output  $t_n$ .

**Solution:**

- (a) Consider the following formula for the target variable:

$$t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon(\mathbf{x})$$

where  $\epsilon$  is a zero mean Gaussian random variable with precision  $\beta(\mathbf{x})$  and  $\phi$  is a deterministic function. Due to heteroscedasticity, the Gaussian noise is dependent on the input  $\mathbf{x}$ .

Due to the properties of Gaussian distribution,  $t$  is also normal, with its distribution i.e. the likelihood function given by:

$$p(t_n | \mathbf{x}_n, \mathbf{w}, \beta) = \sqrt{\frac{\beta_n}{2\pi}} \exp \left\{ -\frac{\beta_n}{2} (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 \right\} \quad (1)$$

So,

$$p(t_n | \mathbf{x}_n, \mathbf{w}, \beta) = \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$$

- (b) We may assume a Gaussian prior for  $\mathbf{w}$ , with arbitrary mean  $\mathbf{m}_0$  and covariance matrix  $\mathbf{S}_0$  in which case the prior is given by

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)$$

- (II) **Question:** Provide the expression for the objective function that you will consider for the ML and MAP estimation of the parameters considering a data set of size  $N$ .

**Solution:**

- (a) To express this more succinctly, we define the design matrix

$$\Phi = \begin{pmatrix} \phi(\mathbf{x}_1)^T \\ \vdots \\ \phi(\mathbf{x}_n)^T \end{pmatrix}$$

the data set  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  and the diagonal weighing matrix  $\mathbf{R}$  with  $R_{ii} = \beta_i$ .

Thus, the objective function for ML estimation is given by:

$$\begin{aligned} p(\mathbf{t} | \mathbf{X}, \mathbf{w}, \beta) &= \prod_{n=1}^N p(t_n | \mathbf{x}_n, \mathbf{w}, \beta) \\ &= \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta_n^{-1}) \\ &= \exp \left\{ -\frac{1}{2} \sum_{n=1}^N \beta_n (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 \right\} \\ &= \exp \left\{ -\frac{1}{2} (\mathbf{t} - \Phi^T \mathbf{w})^T \mathbf{R} (\mathbf{t} - \Phi^T \mathbf{w}) \right\} \\ &= \mathcal{N}(\mathbf{t} | \Phi \mathbf{w}, \mathbf{R}^{-1}) \end{aligned} \quad (2)$$

(b) Given

$$p(\mathbf{w}) = \mathcal{N}(w|\mathbf{m}_0, \mathbf{S}_0)$$

and

$$p(\mathbf{t}|\mathbf{w}) = \mathcal{N}(\Phi\mathbf{w}, \mathbf{R}^{-1})$$

Let  $\mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \mathbf{t} \end{pmatrix}$  as follows:

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{w}) + \ln p(\mathbf{t}|\mathbf{w}) \\ &= -\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \\ &\quad - \frac{1}{2}(\mathbf{t} - \Phi\mathbf{w})^T \mathbf{R}(\mathbf{t} - \Phi\mathbf{w}) + \text{constant} \\ &= -\frac{1}{2} [\mathbf{w}^T \mathbf{S}_0^{-1} \mathbf{w} - 2\mathbf{w}^T \mathbf{S}_0^{-1} \mathbf{m}_0] \\ &\quad - \frac{1}{2} [\mathbf{t}^T \mathbf{R} \mathbf{t} - 2\mathbf{t}^T \mathbf{R} \Phi \mathbf{w} + \mathbf{w}^T \Phi^T \mathbf{R} \Phi \mathbf{w}] + \text{constant} \end{aligned} \tag{3}$$

where we have removed terms which are constant with respect to  $\mathbf{w}$  and  $\mathbf{t}$ .

Furthermore, due to linearity of expectation, we have

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \mathbb{E}[\mathbf{w}] \\ \mathbb{E}[\mathbf{t}] \end{pmatrix}$$

from which

$$\text{cov}[\mathbf{z}] = \begin{pmatrix} \text{var}[\mathbf{w}] & \text{cov}[\mathbf{w}, \mathbf{t}] \\ \text{cov}[\mathbf{t}, \mathbf{w}] & \text{var}[\mathbf{t}] \end{pmatrix}$$

From (3), it is clear that  $p(\mathbf{z})$  is a Gaussian distribution. Now we complete the square.

To find the covariance of  $\mathbf{w}|\mathbf{t}$ , we consider the single term of second order in  $\mathbf{w}$  from (3):

$$\frac{1}{2} \mathbf{w}^T \Sigma^{-1} \mathbf{w} = \frac{1}{2} \mathbf{w}^T (\mathbf{S}_0^{-1} + \Phi^T \mathbf{R} \Phi) \mathbf{w}$$

We treat  $\mathbf{t}$  as a constant.

Thus, the covariance is given by

$$\Sigma = (\mathbf{S}_0^{-1} + \Phi^T \mathbf{R} \Phi)^{-1}$$

Similarly, we may obtain  $\boldsymbol{\mu}$  using the terms of (3) of first order in  $\mathbf{w}$ . We have

$$\mathbf{w}^T \Sigma^{-1} \boldsymbol{\mu} = \mathbf{w}^T \mathbf{S}_0^{-1} \mathbf{m}_0 + \mathbf{w}^T \Phi^T \mathbf{R} \mathbf{t}$$

which yields

$$\boldsymbol{\mu} = \Sigma (\mathbf{S}_0^{-1} \mathbf{m}_0 + \Phi^T \mathbf{R} \mathbf{t})$$

Thus, the MAP objective function is

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \Sigma)$$

(III) **Question:** Show

$$E_{\mathcal{D}} = \sum_{n=1}^N r_n \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2$$

and find  $\mathbf{w}$  that minimizes  $E_{\mathcal{D}}$ . **Solution:**

(a) Taking logarithm of (1)

$$-\ln p(\mathcal{D}|\mathbf{w}) = \frac{N}{2} \ln \beta_n - \frac{N}{2} \ln(2\pi) + E_{\mathcal{D}}$$

with

$$E_{\mathcal{D}} = \frac{1}{2} \sum_{n=1}^N \beta_n \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 \quad (4)$$

Setting  $r_n = \frac{\beta_n}{2}$ , we obtain the desired equation.

(b) Now, we may obtain the  $\mathbf{w}$  that minimizes  $E_{\mathcal{D}}$  by differentiating (4) and setting the derivative to zero like so:

$$\sum_{n=1}^N \beta_n \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\} \phi(\mathbf{x}_n)^T = 0$$

Whence

$$\sum_{n=1}^N t_n \beta_n \phi(\mathbf{x}_n)^T = \mathbf{w}^T \sum_{n=1}^N \beta_n \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T$$

Using the matrix  $\mathbf{R}$  defined as before, taking the transpose of both sides yields

$$\begin{aligned} \Phi^T \mathbf{R} \mathbf{t} &= (\Phi^T \mathbf{R} \Phi) \mathbf{w} \\ \implies \mathbf{w}_{\text{ML}} &= (\Phi^T \mathbf{R} \Phi)^{-1} \Phi^T \mathbf{R} \mathbf{t} \end{aligned} \quad (5)$$