$$X = \begin{pmatrix} x_0^0 & \cdots & x_0^M \\ \vdots & & \vdots \\ x_N^0 & \cdots & x_N^M \end{pmatrix}, \quad t = \begin{pmatrix} t_0 \\ t_1 \\ \vdots \\ t_N \end{pmatrix}, \quad w = \begin{pmatrix} w_0 \\ \vdots \\ w_M \end{pmatrix}$$

We can see $E(w) = \frac{1}{2}(Xw - t)^T (Xw - t)$

In order to minimize we need to differentiate w.r.t w.

$$\frac{\partial E(w)}{\partial w} = \frac{1}{2}(w^T X^T X w - 2t^T X w + t^T t)$$

$$= X^T X w - X^T t \quad \text{(from Hw1)}$$

so $X^T X w = X^T t$ at minimum

For $X^T t$ consider the i^{th} component:

$$\begin{pmatrix} x_0^0 & \cdots & x_N^0 \\ \vdots & & \\ x_0^M & & x_N^M \end{pmatrix} \begin{pmatrix} t_0 \\ \vdots \\ t_N \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} x_n t_n \\ \vdots \\ \sum_{n=1}^{N} x_n^M t_N \end{pmatrix}$$

So the i^{th} component is $\sum_{n=1}^{N} (x_n)^i t_n = T_i$

2) $E_D(w) = \frac{1}{2}(Xw - t)^T (Xw - t)$

where $X =$

Now for $X^T X w$:

$$\begin{pmatrix} x_0^0 & \cdots & x_N^0 \\ \vdots & & \vdots \\ x_0^m & \cdots & x_N^m \end{pmatrix} \begin{pmatrix} x_0^0 & \cdots & x_0^m \\ \vdots & & \vdots \\ x_N^0 & \cdots & x_N^m \end{pmatrix} \begin{pmatrix} w_0 \\ \vdots \\ w_m \end{pmatrix}$$

We can see the i^{th} component will be obtained by summing the entries of the product of the i^{th} row of $X^T X$ and w, which is given by

$$\sum_{j=0}^{m} \left(\sum_{n=1}^{N} (x_n)^{i+j} \right) w_j = \sum_{j=0}^{m} A_{ij} w_j \text{ in our notation.}$$

Hence $\sum_{j=0}^{m} A_{ij} w_j = T_i$

2) For this question, consider the gradient.

$$\nabla E_D(W) = \nabla w \left(\frac{1}{2} \sum_{n=1}^{N} r_n (t_n - w^T \phi(x_n))^2 \right)$$

$$= \sum_{n=1}^{N} r_n (t_n - w^T \phi(x_n)) \cdot (-\phi(x_n))$$

$$= \sum_{n=1}^{N} r_n (\phi(x_n) \phi(x_n)^T w - t_n \phi(x_n))$$

Now taking this as equal to 0 we get:

$$\left(\sum_{n=1}^{N} r_n (\phi(x_n) \phi(x_n)^T w \right) = \sum_{n=1}^{N} r_n t_n \phi(x_n)$$

so $w^* = \left(\sum_{n=1}^{N} r_n \phi(x_n) \phi(x_n)^T \right)^{-1} \sum_{n=1}^{N} r_n t_n \phi(x_n)$

Now we need to confirm w^* is a min and not max.

3) For max of posterior:

We will consider $p(w|t, \alpha, \beta)$.

Using Bayes Theorem:

$$p(w|t, \alpha, \beta) \propto p(t|w, \beta) \times p(w|\alpha)$$

$$\ln p(t|w,\beta) = \ln\left(\prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{-\frac{\beta}{2}(y(x_n,w)-t_n)^2}\right)$$

$$= \ln\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} e^{-\frac{\beta}{2}\sum_{i=1}^{N}(y(x_n,w)-t_n)^2}$$

$$= -\frac{\beta}{2}\sum_{i=1}^{N}|y(x_n,w)-t_n|^2 + C$$

So $-\frac{\beta}{2}\sum_{i=1}^{N}(y(x_n,w)-t_n)^2 - \frac{\alpha}{2}w^T w$

The max of this is equivalent to the min of

$$\frac{\beta}{2}\sum_{i=1}^{N}(y(x_n,w)-t_n)^2 + \frac{\alpha}{2}w^T w.$$

MAP vs MLE:

Our MLE is just what we get from minimizing $\frac{B}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)^2$.

Hence, it doesn't consider previous information.

This means the MLE is more prone to overfitting

As $\alpha \to 0$, MAP $\Rightarrow$ MLE.