Question ①

1 (a)

$$y = f(x) = w_0 + \sum_{k=1}^{K} w_k b_k(x) = b^T(x)\, w$$

$$b(x) = \begin{bmatrix} 1 \\ w_1(x) \\ \vdots \\ w_K(x) \end{bmatrix} \qquad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$$

(b)

$$E(\tilde{w}) = \sum_{i=1}^{N} (y_i - b^T(x_i)\, w)^2$$

$$= \| y - Bw \|^2$$

where $B = \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \cdots & b_K(x_1) \\ \vdots & & & & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \cdots & b_K(x_N) \end{bmatrix}$ $\quad w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_K \end{bmatrix}$ $\quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

Therefore, we can know that

$$E(\tilde{w}) = \|y - Bw\|^2 = (y - Bw)^2 = (y - Bw)^T (y - Bw)$$
$$= y^T (y - Bw) + (Bw)^T (Bw) - (Bw)^T y$$
$$= y^T y - y^T Bw - w^T B^T y + (Bw)^T Bw$$
$$= y^T y - 2y^T Bw + (Bw)^T Bw = y^T y - 2y^T Bw + w^T B^T Bw$$
$$= y^T y - 2w^T B^T y + w^T B^T Bw$$

$\therefore$ The gradient of objective function is

$$\nabla E = \frac{\partial E(\tilde{w})}{\partial w} = -2B^T y + (B^T B + B^T B) w = -2B^T y + 2B^T Bw.$$

(c)

To solve for the optimal weight vector w

$$\nabla E = 0$$

$$\Rightarrow \; -2B^T y + 2B^T Bw = 0$$

$$B^T Bw = \cancel{2} B^T y$$

$$w = (B^T B)^{-1} B^T y$$

Question ②

(a) In part (c), the gradient we get is

$$\nabla E = -2B^T y + 2B^T B W = 0$$

$$\Rightarrow \quad B^T B W = B^T y$$

The solution of w will not be unique only when both $B^T B = B^T y = 0$

$$B = \begin{bmatrix} 1 & b_1(x_1) & b_2(x_1) & \cdots & b_K(x_1) \\ \vdots & & & & \vdots \\ 1 & b_1(x_N) & b_2(x_N) & \cdots & b_K(x_N) \end{bmatrix} \qquad B^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ b_1(x_1) & & \cdots & b_1(x_N) \\ b_2(x_1) & & & \vdots \\ \vdots & & & \vdots \\ b_K(x_1) & & \cdots & b_K(x_N) \end{bmatrix}$$

A simple regression that make the weight not unique is that.

$b_K(x_1), b_K(x_2) \cdots b_K(x_N) = \sqrt{-1}$, All $b_j(x_i) = 0$. $1 \le j \le K-1$ $i \in [1, N]$ i, j

$$y = \underbrace{[0, 0, 0 \cdots 0]}_{N \text{ number of } 0\text{s}}.$$

In this case, each row of B^T * each column of B

$$= [1, 0 \cdots 0, \sqrt{-1}] * \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \sqrt{-1} \end{bmatrix} = 0$$

$\therefore$ $B^T B$ = zero matrix, also since $y = [0, 0, \cdots 0]$ $\therefore$ $B^T y = 0$.

$\therefore$ In this situation W is not unique.

(b)

Q3

(a)

$\because \quad y = f(x) + \varepsilon$

$\quad\quad = b(x)^T w + \varepsilon$

where

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{bmatrix} \quad b(x) = \begin{bmatrix} 1 \\ b_1(x) \\ \vdots \\ b_k(x) \end{bmatrix} \quad \varepsilon \sim N(0, \sigma^2)$$

$\because$ y given x follows a Gaussian distribution with mean $f(x)$ and variance σ^2

$$\therefore \quad p(y|x_i, w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - w^T b(x))^2}$$

$\therefore$ The likelihood is

$$L(w) = \prod_{i=1}^{N} p(y_i|x_i, w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - w^T b(x))^2}$$

The maximum likelihood is

$$\arg\max L(w) = \arg\max \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - w^T b(x))^2}$$

(b)

The negative log likelihood is

$$L'(w) = -\log\left(\prod_{i=1}^{N} p(y_i|x_i, w)\right)$$

$$= -\log\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - w^T b(x))^2}\right)$$

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

$$= \sum_{i=1}^{N} \frac{1}{2\sigma^2}(y_i - w^T b(x))^2 - \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma}$$

The minimum of our negative log likelihood is

$$\arg\min L'(w) = \arg\min \sum_{i=1}^{N} (y_i - w^T b(x))^2$$

Thus, Which got the same thing that we have for LS objective in Q

Q3(c)

$\because$ $w \sim N(0, \alpha^{-1}I)$, therefore our posterior is

$$p(w|x_i, y_i) = p(y_i|x_i, w)\,p(w)$$
$$= \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(y_i - w^T\phi(x))^2} \cdot \frac{1}{\sqrt{(2\pi)^d |\alpha^{-1}I|}} e^{-\frac{1}{2}(w-0)^T(\alpha^{-1}I)^{-1}(w-0)}$$
$$= \frac{1}{\sqrt{2\pi}\,\sigma \sqrt{(2\pi)^d|\alpha^{-1}I|}} e^{-\frac{1}{2\sigma^2}(y_i - w^T\phi(x))^2 - \frac{\alpha}{2}w^Tw)}$$

$$w_{MAP} = \arg\min\,(-\log p(w|x_i, y_i))$$
$$= \arg\min \sum_{i=1}^{N}(y_i - w^T\phi(x))^2 + \lambda\|w\|^2$$
where $\lambda = \alpha\sigma^2$

(d)

~~When~~ after minimizing the negative log posterior.
We got the same thing as the regularied LS estimate.

(e) If model parameter follows a uniform distribution.

$$w_{MAP} = \arg\min -\log p(w|x_i, y_i)$$
$$= \arg\min -\left(\sum_{i=1}^{N}\log p(y_i|x_i, w) + \underbrace{\sum_{i=1}^{N}\log p(w)}_{\text{constant.}}\right)$$
$$= \arg\min -\sum_{i=1}^{N}\log p(y_i|x_i, w)$$
$$= w_{ML}$$

$\therefore$ in this case, $w_{MAP} = w_{ML}$