Chapter 7. Linear regression

7.1.
Because overfitting happens when the data size is not enough.

7.2.

$$\begin{aligned}
&\begin{pmatrix}\hat{\mathbf{y}}_1\\ \hat{\mathbf{y}}_2\end{pmatrix}=\begin{pmatrix}\hat{\mathbf{w}}_1^T\\ \hat{\mathbf{w}}_2^T\end{pmatrix}\begin{pmatrix}\phi_1(x)\\ \phi_2(x)\end{pmatrix}\\
&\mathbf{X}_1=\mathbf{X}_2=\mathbf{X}=\begin{pmatrix}1&0\\1&0\\1&0\\0&1\\0&1\\0&1\end{pmatrix},\mathbf{y}_1=\begin{pmatrix}-1\\-1\\-2\\1\\1\\2\end{pmatrix},\mathbf{y}_2=\begin{pmatrix}-1\\-2\\-1\\1\\2\\1\end{pmatrix}\\
&(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T=\begin{pmatrix}\frac{1}{3}&\frac{1}{3}&\frac{1}{3}&0&0&0\\0&0&0&\frac{1}{3}&\frac{1}{3}&\frac{1}{3}\end{pmatrix}\\
&\Rightarrow\hat{\mathbf{w}}_1=\begin{pmatrix}-\frac{4}{3}\\ \frac{4}{3}\end{pmatrix},\hat{\mathbf{w}}_2=\begin{pmatrix}-\frac{4}{3}\\ \frac{4}{3}\end{pmatrix}\\
&\Rightarrow\hat{\mathbf{W}}=\begin{pmatrix}-\frac{4}{3}&-\frac{4}{3}\\ \frac{4}{3}&\frac{4}{3}\end{pmatrix}.
\end{aligned}\tag{1}$$

7.3.

$$
\begin{aligned}
&J(\mathbf{w}, w_0) = (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n)^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n) + \lambda\mathbf{w}^T\mathbf{w} \\
&\frac{\partial J}{\partial \mathbf{w}} = 2(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n)^T \cdot \frac{\partial}{\partial \mathbf{w}}(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n) + 2\lambda(\mathbf{w}^T\frac{\partial \mathbf{w}}{\partial \mathbf{w}}) \\
&= -2(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n)^T\mathbf{X} + 2\lambda\mathbf{w}^T \\
&\frac{\partial J}{\partial w_0} = -(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n)^T\mathbf{1}_n - \mathbf{1}_n^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_n).
\end{aligned}
\tag{2}
$$

Since $\bar{\mathbf{x}} = 0$, we have $\mathbf{1}_n^T\mathbf{X} = \mathbf{0}$.

$$
\begin{aligned}
&\frac{\partial J}{\partial \mathbf{w}} = \mathbf{0} \Leftrightarrow (\mathbf{y} - \mathbf{X}\mathbf{w})^T\mathbf{X} = \lambda\mathbf{w}^T \\
&\Leftrightarrow \mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = \lambda\mathbf{w} \Leftrightarrow \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\mathbf{w} = \lambda\mathbf{w} \\
&\Leftrightarrow (\lambda\mathbf{I}_m + \mathbf{X}^T\mathbf{X})\mathbf{w} = \mathbf{X}^T\mathbf{y} \\
&\Rightarrow \hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_m)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
\tag{3}
$$

Since $\bar{\mathbf{x}} = 0$, we have $\sum_i x_{ij} = 0$ for all j.
Therefore, $\sum_i \sum_j x_{ij}w_j = \sum_j w_j \sum_i x_{ij} = 0$.

$$
\begin{aligned}
&\frac{\partial J}{\partial w_0} = 0 \Leftrightarrow \sum_i y_i - \sum_i(\sum_j x_{ij}w_j) - nw_0 = 0 \\
&\Leftrightarrow \sum_i y_i = nw_0 \\
&\Rightarrow \hat{w}_0 = \bar{y}.
\end{aligned}
\tag{4}
$$

7.4.

The log-likelihood is

$$
\begin{aligned}
&l(\mathbf{w}, \sigma^2) = -\frac{1}{2\sigma^2}\sum_i (y_i - \mathbf{w}^T\mathbf{x}_i)^2 - \frac{N}{2}\log(2\pi\sigma^2) \\
&\frac{\partial l}{\partial \sigma} = \frac{1}{\sigma^3}\sum_i (y_i - \mathbf{w}^T\mathbf{x}_i)^2 - \frac{N}{2}\frac{1}{2\pi\sigma^2}4\pi\sigma = 0 \\
&\Leftrightarrow \sigma^2 = \frac{\sum_i (y_i - \mathbf{w}^T\mathbf{x}_i)^2}{N} \\
&\frac{\partial l}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \hat{\mathbf{w}} \\
&\Rightarrow \hat{\sigma}^2 = \frac{1}{N}\sum_i (y_i - \mathbf{x}_i^T\hat{\mathbf{w}})^2.
\end{aligned} \tag{5}
$$

7.5.
Similar situation with Problem 7.3, but here we don't have $\bar{\mathbf{x}} = 0$ and there is no $\lambda\mathbf{w}^T\mathbf{w}$ term.

$$
\begin{aligned}
&\frac{\partial J}{\partial w_0} = 0 \Leftrightarrow \sum_i y_i - \sum_i (\sum_j x_{ij}w_j) - n\hat{w}_0 = 0 \\
&\Leftrightarrow \hat{w}_0 = \frac{1}{N}\sum_i y_i - \frac{1}{N}\sum_i\sum_j x_{ij}w_j = \frac{1}{N}\sum_i y_i - \frac{1}{N}\sum_i \mathbf{x}_i^T\mathbf{w} \\
&= \bar{y} - \bar{\mathbf{x}}^T\mathbf{w}.
\end{aligned} \tag{6}
$$

If we plug this result in the formula of $J(\mathbf{w}, w_0)$,

$$
\begin{aligned}
&(\mathbf{y} - \mathbf{X}\mathbf{w} - \hat{w}_0\mathbf{1}_n)^T(\mathbf{y} - \mathbf{X}\mathbf{w} - \hat{w}_0\mathbf{1}_n) = (\mathbf{y}_c - \mathbf{X}_c\mathbf{w})^T(\mathbf{y}_c - \mathbf{X}_c\mathbf{w}). \\
&\Rightarrow \frac{\partial J}{\partial \mathbf{w}} = 0 \Leftrightarrow (\mathbf{y}_c - \mathbf{X}_c\mathbf{w})^T\mathbf{X}_c = \mathbf{0} \\
&\Rightarrow \mathbf{X}_c^T\mathbf{y}_c = \mathbf{X}_c^T\mathbf{X}_c\mathbf{w} \\
&\Rightarrow \hat{\mathbf{w}} = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}_c.
\end{aligned} \tag{7}
$$

7.6.
Just setting $D = 1$ in the results from Problem 7.5 gives

$$\hat{w}_0 = \bar{y} - w_1\bar{x}, \hat{w}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

7.7.
a,

$$\hat{w}_1 = \frac{C_{xy}^{(n)}}{C_{xx}^{(n)}}.$$

b,

$$\hat{w}_0 = \bar{y}^{(n)} - \frac{C_{xy}^{(n)}}{C_{xx}^{(n)}}\bar{x}^{(n)}.$$

c.

$$\begin{aligned}&\bar{x}^{(n+1)} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}(\sum_{i=1}^{n} x_i + x_{n+1}) = \frac{1}{n+1}(n\bar{x}^{(n)} + x_{n+1})\\ &= \frac{n}{n+1}\bar{x}^{(n)} + \frac{1}{n+1}x_{n+1}.\end{aligned} \tag{8}$$

d.

$$
\begin{aligned}
C_{xy}^{(n+1)} &= \frac{1}{n+1}(\sum_{i=1}^{n+1}(x_i - \bar{x}^{(n+1)})(y_i - \bar{y}^{(n+1)})) \\
&= \frac{1}{n+1}(\sum_{i=1}^{n+1}(x_i y_i - y_i \bar{x}^{(n+1)} - x_i \bar{y}^{(n+1)} + \bar{x}^{(n+1)}\bar{y}^{(n+1)})) \\
&= \frac{1}{n+1}((\sum_{i=1}^{n+1} x_i y_i) - 2(n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)} + (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}) \\
&= \frac{1}{n+1}(\sum_{i=1}^{n+1} x_i y_i - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}) \\
&= \frac{1}{n+1}(\sum_{i=1}^{n} x_i y_i + x_{n+1}y_{n+1} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}) \\
&= \frac{1}{n+1}(\sum_{i=1}^{n}(x_i - \bar{x}^{(n)})(y_i - \bar{y}^{(n)}) - n\bar{x}^{(n)}\bar{y}^{(n)} + 2n\bar{x}^{(n)}\bar{y}^{(n)} \\
&+ x_{n+1}y_{n+1} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}) \\
&= \frac{1}{n+1}(x_{n+1}y_{n+1} + nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}).
\end{aligned}
\tag{9}
$$

7.8.
a. $\bar{w}_0 \approx -3.2564, \bar{w}_1 \approx -0.0426 \Rightarrow \bar{\sigma}^2 \approx 0.016975.$

b.

$$
\begin{aligned}
& p(w_0) = 1, p(w_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w_1^2}{2}} \\
& \Rightarrow p(\mathbf{w}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w_1^2}{2}} = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{V}_0)
\end{aligned}
\tag{10}
$$

where

$$\mathbf{w}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{V}_0 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{11}$$

c.
$p(\mathbf{w}|\mathcal{D}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N)$
where

$$\begin{aligned}
&\mathbf{V}_N = \sigma^2(\sigma^2\mathbf{V}_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1} \approx \begin{pmatrix} 0.1323 & -0.001 \\ -0.001 & 0.000011 \end{pmatrix}, \\
&\mathbf{w}_N = \begin{pmatrix} 0.1323 & -0.001 \\ -0.001 & 0.000011 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
&+ \frac{1}{0.016975} \begin{pmatrix} 0.1323 & -0.001 \\ -0.001 & 0.000011 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 94 & 96 & \cdots & 131 \end{pmatrix} \begin{pmatrix} 0.47 \\ 0.75 \\ \cdots \\ 2.23 \end{pmatrix} \\
&\approx \begin{pmatrix} -3.2555 \\ 0.04276 \end{pmatrix} \\
&\Rightarrow p(w_1|\mathcal{D}, \sigma^2) \approx \mathcal{N}(w_1|0.04276, 0.00001).
\end{aligned} \tag{12}$$

d.
The 95% credible interval is approximately $[0.042738, 0.042782]$.

7.9.
a.

$$\begin{aligned}
&\hat{\boldsymbol{\Sigma}}_{XX} = \frac{1}{n}\sum_i (\mathbf{x}_i - \mathbf{x})(\mathbf{x}_i - \mathbf{x})^T, \hat{\boldsymbol{\Sigma}}_{XY} = \frac{1}{n}\sum_i (\mathbf{y}_i - \mathbf{y})(\mathbf{x}_i - \mathbf{x})^T \\
&\hat{\boldsymbol{\mu}}_x = \frac{1}{n}\sum_i \mathbf{x}_i, \hat{\boldsymbol{\mu}}_y = \frac{1}{n}\sum_i \mathbf{y}_i \\
&p(y|\mathbf{x}) = p(y|\boldsymbol{\mu}_{y|\mathbf{x}}, \boldsymbol{\Sigma}_{y|\mathbf{x}})
\end{aligned} \tag{13}$$

Hence it suffices to compute $\hat{\boldsymbol{\mu}}_{y|\mathbf{x}} = \hat{\boldsymbol{\mu}}_y + \hat{\boldsymbol{\Sigma}}_{YX}\hat{\boldsymbol{\Sigma}}_{XX}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_x)$.

$$
\begin{aligned}
&\hat{\boldsymbol{\mu}}_{y|\mathbf{x}} = \bar{y} + (\sum_i (y_i - \bar{y})(\mathbf{x}_i - \bar{\mathbf{x}}))^T(\mathbf{X}_c^T\mathbf{X}_c)^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \\
&= \bar{y} + (\mathbf{X}_c^T\mathbf{y}_c)^T(\mathbf{X}_c^T\mathbf{X}_c)^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \\
&= \bar{y} + (((\mathbf{X}_c^T\mathbf{X}_c)^{-1})^T\mathbf{X}_c^T\mathbf{y}_c)^T(\mathbf{x} - \bar{\mathbf{x}}) \\
&= \bar{y} + ((\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}_c)^T(\mathbf{x} - \bar{\mathbf{x}}) = \bar{y} + \mathbf{w}^T(\mathbf{x} - \bar{\mathbf{x}}) \\
&= \bar{y} + \mathbf{w}^T\mathbf{x} - \mathbf{w}^T\bar{\mathbf{x}} = \bar{y} + \mathbf{x}^T\mathbf{w} - \bar{\mathbf{x}}^T\mathbf{w} \\
&= w_0 + \mathbf{w}^T\mathbf{x}.
\end{aligned}
\tag{14}
$$

7.10.

$$
\begin{aligned}
&p(\mathbf{w}, \sigma^2|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w}, \sigma^2)p(\mathbf{w}, \sigma^2) \\
&\propto \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N)\mathcal{N}(\mathbf{w}|\mathbf{0}, \sigma^2 g(\mathbf{X}^T\mathbf{X})^{-1})\mathrm{IG}(\sigma^2|0, 0) \\
&\propto (\sigma^2)^{-\frac{N}{2}}e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})}(\sigma^2)^{-\frac{D}{2}}e^{-\frac{1}{2\sigma^2}(\mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{g}\mathbf{w})}(\sigma^2)^{-1}.
\end{aligned}
\tag{15}
$$

$$
\begin{aligned}
&(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{g}\mathbf{w} \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{g}\mathbf{w} \\
&= \mathbf{w}^T\mathbf{V}_N^{-1}\mathbf{w} - \mathbf{w}^T(\frac{g+1}{g}\mathbf{X}^T\mathbf{X})(\frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&- (\frac{g}{g+1}\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})(\frac{g+1}{g}\mathbf{X}^T\mathbf{X})\mathbf{w} + \mathbf{y}^T\mathbf{y} \\
&= \mathbf{w}^T\mathbf{V}_N^{-1}\mathbf{w} - \mathbf{w}^T\mathbf{V}_N^{-1}\mathbf{w}_N - \mathbf{w}_N^T\mathbf{V}_N^{-1}\mathbf{w} + \mathbf{y}^T\mathbf{y} \\
&= (\mathbf{w} - \mathbf{w}_N)^T\mathbf{V}_N^{-1}(\mathbf{w} - \mathbf{w}_N) + \mathbf{y}^T\mathbf{y} - \mathbf{w}_N^T\mathbf{V}_N^{-1}\mathbf{w}_N.
\end{aligned}
\tag{16}
$$

$$
\begin{aligned}
&\Rightarrow p(\mathbf{w}, \sigma^2 | \mathcal{D}) \propto (\sigma^2)^{-\frac{D}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{w}-\mathbf{w}_N)^T \mathbf{V}_N^{-1}(\mathbf{w}-\mathbf{w}_N)} \\
&\cdot (\sigma^2)^{-\frac{N}{2}-1} e^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - \mathbf{w}_N^T \mathbf{V}_N^{-1}\mathbf{w}_N)} \\
&\propto \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \sigma^2 \mathbf{W}_N)\mathrm{IG}(\sigma^2 | \frac{N}{2}, \frac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{w}_N^T \mathbf{V}_N^{-1}\mathbf{w}_N))
\end{aligned}
\tag{17}
$$

To go further,

$$
\begin{aligned}
&\frac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{w}_N^T \mathbf{V}_N^{-1}\mathbf{w}_N) = \frac{1}{2}(\mathbf{y}^T\mathbf{y} - \frac{g}{g+1}\hat{\mathbf{w}}_{MLE}^T \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_{MLE}) \\
&= \frac{1}{2}((\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_{MLE})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_{MLE}) + \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&+ \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - \hat{\mathbf{w}}_{MLE}^T \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_{MLE} - \frac{g}{g+1}\hat{\mathbf{w}}_{MLE}^T \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_{MLE})
\end{aligned}
\tag{18}
$$

However,

$$
\begin{aligned}
&\hat{\mathbf{w}}_{MLE}^T \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_{MLE} = \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \\
&\Rightarrow \frac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{w}_N^T \mathbf{V}_N^{-1}\mathbf{w}_N) \\
&= \frac{1}{2}((\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_{MLE})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}_{MLE}) + \frac{1}{g+1}\hat{\mathbf{w}}_{MLE}^T \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_{MLE}) \\
&= \frac{s^2}{2} + \frac{1}{2(g+1)}\hat{\mathbf{w}}_{MLE}^T \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_{MLE} = b_N.
\end{aligned}
\tag{19}
$$

Therefore finally we can conclude

$$
p(\mathbf{w}, \sigma^2 | \mathcal{D}) = \mathrm{NIG}(\mathbf{w}, \sigma^2 | \mathbf{w}_N, \mathbf{V}_N, a_N, b_N)
$$

as desired.