EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE: August 24, 2023

RESPONSIBLE TEACHER: Dave Zachariah

NUMBER OF PROBLEMS: 5

AIDING MATERIAL: Calculator, mathematical handbooks

PRELIMINARY GRADES:	grade 3	23 points
	grade 4	33 points
	grade 5	43 points

Some general instructions and information:

- Your solutions should be given in English.
- Only write on one page of the paper.
- Write your exam code and a page number on all pages.
- Do not use a red pen.
- Use separate sheets of paper for the different problems (i.e. the numbered problems, 1–5).

With the exception of Problem 1, all your answers must be clearly motivated! A correct answer without a proper motivation will score zero points!

Good luck!

1. i) False. [x can be continuous as well]
   ii) True.
   iii) True.
   iv) False. [Typically the training error underestimates the test error.]
   v) True. [$f(x) = w^{\mathsf{T}}(Wx + b)$.]
   vi) False.
   vii) False.
   viii) True.
   ix) False.
   x) False.

2. a) i. $\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ with $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$.

ii. Predictions:
$\hat{y}_i = \theta_0 + \theta_1 \mathbf{x}_i$ or in vector form $\hat{\mathbf{y}} = X\boldsymbol{\theta}$.
Objective function:
$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ or $J(\boldsymbol{\theta}) = \frac{1}{n}||\mathbf{y} - X\boldsymbol{\theta}||_2^2$.

iii.

$$\begin{aligned} ||\mathbf{y} - X\boldsymbol{\theta}||_2^2 &= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top X\boldsymbol{\theta} - \boldsymbol{\theta}^\top X^\top \mathbf{y} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} \\ \frac{\partial}{\partial\boldsymbol{\theta}}||\mathbf{y} - X\boldsymbol{\theta}||_2^2 &= 0 - 2X^\top\mathbf{y} + 2X^\top X\boldsymbol{\theta} = 0 \\ 0 &= -X^\top\mathbf{y} + X^\top X\boldsymbol{\theta} \\ \boldsymbol{\theta} &= (X^\top X)^{-1}X^\top\mathbf{y} \end{aligned}$$

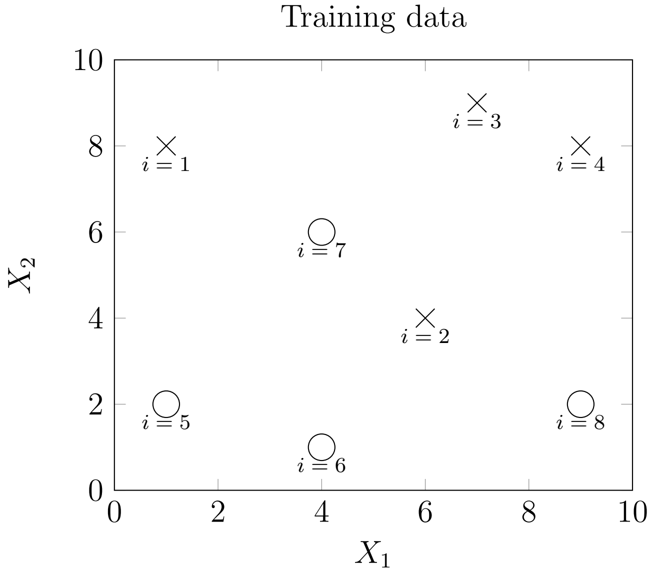
b) i. offset and slope of the linear function.

ii. the data has to be linear in $\mathbf{x}$ to capture the information. Otherwise, we have to change our model.

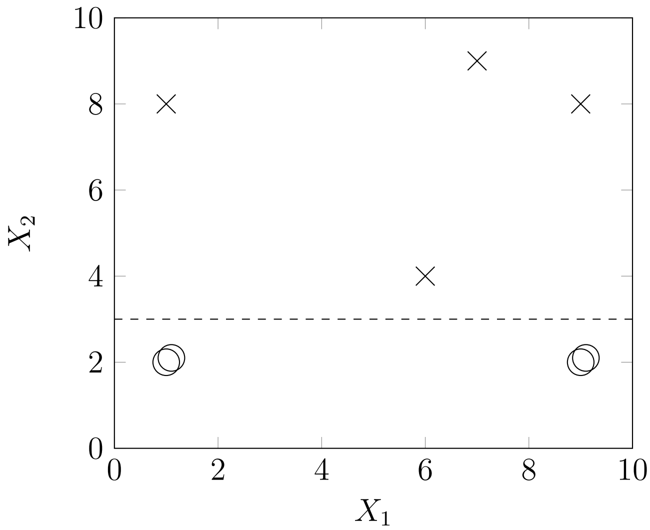
c) To select only a few features of $\mathbf{x}$ we want to use Lasso regularization with $R(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_1$ with fairly high strength λ. The reason is that the 1-norm suppresses some model parameters close to zero. Hence, some features are not selected for prediction.

d) Yes, you can use a one-layer neural network i.e., no hidden layer. Inputs are the features $\mathbf{x}$ and the single output is y. The layer weights are the parameters θ_1 and the bias is the parameter θ_0. There is no nonlinear activation function. Furthermore, the model is trained with the same MSE loss.

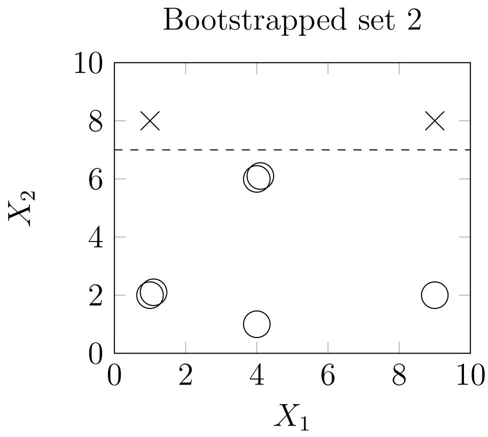
3. (a) The training data is illustrated in the following plot:

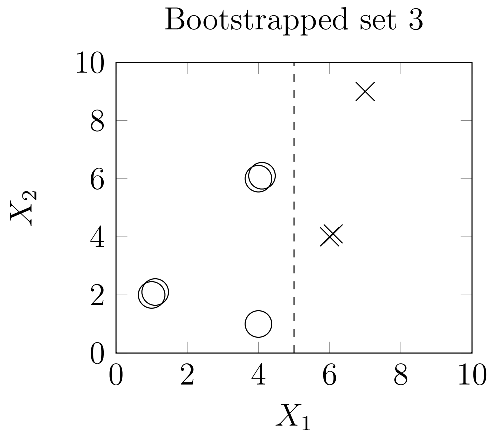


(b) Bootstraped dataset 1 looks as



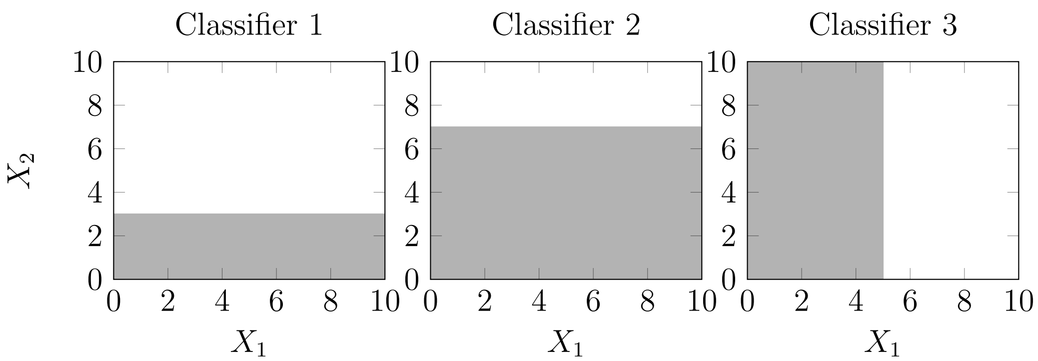
Only splits at $X_2 = r$ where $2 < r < 4$ gives zero missclassification error. We choose to split at $X_2 = 3$. Similar plots for bootstrapped dataset 2 and 3 are given below:



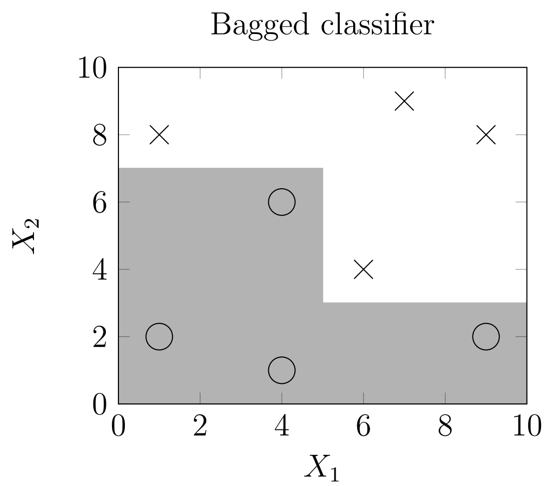


For the second dataset only splits at $X_2 = r$, $6 < r < 8$ gives zero misclassification error. We choose to split at $X_2 = 7$. For the third dataset only splits at $X_1 = r$, $4 < r < 6$ gives zero misclassification error. We choose to split at $X_1 = 5$.

(c) The decision boundary for each classifier becomes (gray: $Y = 0$ (circle), white: $Y = 1$ (cross))



which, with a majority vote, gives the final decision boundary



Note that in contrast to each ensemble member, the final bagged classifier manages to classify all data points correctly.

4. a)

$$\hat{\pi}_1 = \frac{n_1}{n} = \frac{4}{10} = 0.4,$$

where n is the total number of data points in the dataset and n_1 is the number of data points with label $y = 1$. Similarly, $\hat{\pi}_{-1} = 0.6$.
For the means $\hat{\mu}_1$ and $\hat{\mu}_{-1}$,

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n_1} \sum_{i:y_i=1} x_i \\ &= \frac{1}{4}(x_1 + x_2 + x_3 + x_4) \\ &= \frac{1}{4}(82 + 84 + 85 + 88) \\ &= 84.75,\end{aligned}$$

$$\begin{aligned}\hat{\mu}_{-1} &= \frac{1}{n_{-1}} \sum_{i:y_i=-1} x_i \\ &= \frac{1}{6}(x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}) \\ &= \frac{1}{6}(83 + 85 + 87 + 89 + 93 + 97) \\ &= 89.\end{aligned}$$

Since x is 1-dimensional, the covariance matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_{-1}$ are scalars, given by,

$$\begin{aligned}\hat{\Sigma}_1 &= \frac{1}{n_1 - 1} \sum_{i:y_i=1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^\mathsf{T} \\ &= \frac{1}{4 - 1} \sum_{i:y_i=1} (x_i - \hat{\mu}_1)^2 \\ &= \frac{1}{3}\big((x_1 - \hat{\mu}_1)^2 + (x_2 - \hat{\mu}_1)^2 + (x_3 - \hat{\mu}_1)^2 + (x_4 - \hat{\mu}_1)^2\big) \\ &= \frac{1}{3}\big((82 - 84.75)^2 + (84 - 84.75)^2 + (85 - 84.75)^2 + (88 - 84.75)^2\big) \\ &= 6.25,\end{aligned}$$

$$\begin{aligned}\hat{\Sigma}_{-1} &= \frac{1}{n_{-1} - 1} \sum_{i:y_i=-1} (x_i - \hat{\mu}_{-1})(x_i - \hat{\mu}_{-1})^\mathsf{T} \\ &= \frac{1}{6 - 1} \sum_{i:y_i=-1} (x_i - \hat{\mu}_{-1})^2 \\ &= \frac{1}{5}\big((x_5 - \hat{\mu}_{-1})^2 + \ldots + (x_{10} - \hat{\mu}_{-1})^2\big) \\ &= \frac{1}{5}\big((83 - 89)^2 + \ldots + (97 - 89)^2\big) \\ &= 27.2.\end{aligned}$$

b) Set $x_\star = 90$. The probability to finish the race in less than 3 hours is then given by $p(y=1|x_\star)$, which is computed according to,

$$
\begin{aligned}
p(y=1|x_\star) &= \frac{\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1)\cdot\hat{\pi}_1}{\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1)\cdot\hat{\pi}_1+\mathcal{N}(x_\star|\hat{\mu}_{-1},\hat{\Sigma}_{-1})\cdot\hat{\pi}_{-1}}\\
&= \frac{\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1)\cdot 0.4}{\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1)\cdot 0.4+\mathcal{N}(x_\star|\hat{\mu}_{-1},\hat{\Sigma}_{-1})\cdot 0.6}.
\end{aligned}
$$

Thus, we need to compute $\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1)$ and $\mathcal{N}(x_\star|\hat{\mu}_{-1},\hat{\Sigma}_{-1})$. Using the formula sheet (with $p=1$, since x is 1-dimensional), we get,

$$
\begin{aligned}
\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1) &= \frac{1}{\sqrt{2\pi\hat{\Sigma}_1}}\exp\Big(-\frac{1}{2\hat{\Sigma}_1}(x_\star-\hat{\mu}_1)^2\Big)\\
&= \frac{1}{\sqrt{2\pi\cdot 6.25}}\exp\Big(-\frac{1}{2\cdot 6.25}(90-84.75)^2\Big)\\
&= 0.017593\ldots,
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{N}(x_\star|\hat{\mu}_{-1},\hat{\Sigma}_{-1}) &= \frac{1}{\sqrt{2\pi\hat{\Sigma}_{-1}}}\exp\Big(-\frac{1}{2\hat{\Sigma}_{-1}}(x_\star-\hat{\mu}_{-1})^2\Big)\\
&= \frac{1}{\sqrt{2\pi\cdot 27.2}}\exp\Big(-\frac{1}{2\cdot 27.2}(90-89)^2\Big)\\
&= 0.075100\ldots.
\end{aligned}
$$

The probability $p(y=1|x_\star)$ is thus,

$$
\begin{aligned}
p(y=1|x_\star) &= \frac{\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1)\cdot 0.4}{\mathcal{N}(x_\star|\hat{\mu}_1,\hat{\Sigma}_1)\cdot 0.4+\mathcal{N}(x_\star|\hat{\mu}_{-1},\hat{\Sigma}_{-1})\cdot 0.6}\\
&= \frac{0.017593\ldots\cdot 0.4}{0.017593\ldots\cdot 0.4+0.075100\ldots\cdot 0.6}\\
&= 0.13508\ldots\\
&\approx 0.14.
\end{aligned}
$$

c) For a 90 minute split time, the probability is,

$$
\begin{aligned}
p(y=1|x=90;\hat{\theta}) &= \frac{e^{\hat{\theta}_0+\hat{\theta}_1\cdot 90}}{1+e^{\hat{\theta}_0+\hat{\theta}_1\cdot 90}}\\
&= \frac{e^{48.9-0.56\cdot 90}}{1+e^{48.9-0.56\cdot 90}}\\
&= 0.1824\ldots\\
&\approx 0.18.
\end{aligned}
$$

Similarly, the probability for a split time of 85 minutes is,

$$\begin{aligned} p(y=1|x=85;\hat{\theta}) &= \frac{e^{\hat{\theta}_0+\hat{\theta}_1\cdot 85}}{1+e^{\hat{\theta}_0+\hat{\theta}_1\cdot 85}} \\ &= 0.7858\ldots \\ &\approx 0.79. \end{aligned}$$

d) We want to find the value of x such that the probability equals 0.5. I.e., we need to solve the equation $p(y=1|x;\hat{\theta})=0.5$. Using the identity $\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}}$, we get,

$$\begin{aligned} p(y=1|x;\hat{\theta}) &= 0.5 \\ \frac{e^{\hat{\theta}_0+\hat{\theta}_1 x}}{1+e^{\hat{\theta}_0+\hat{\theta}_1 x}} &= \frac{1}{2} \\ \frac{1}{1+e^{-\hat{\theta}_0-\hat{\theta}_1 x}} &= \frac{1}{2} \\ 1 &= \frac{1}{2}(1+e^{-\hat{\theta}_0-\hat{\theta}_1 x}) \\ \frac{1}{2} &= \frac{1}{2}e^{-\hat{\theta}_0-\hat{\theta}_1 x} \\ 1 &= e^{-\hat{\theta}_0-\hat{\theta}_1 x} \\ 0 &= -\hat{\theta}_0-\hat{\theta}_1 x \\ \hat{\theta}_1 x &= -\hat{\theta}_0 \\ x &= -\frac{\hat{\theta}_0}{\hat{\theta}_1} \\ x &= \frac{48.9}{0.56} \\ x &= 87.3214\ldots \approx 87.3. \end{aligned}$$

(note that this will depend on the specific marathon course. Because the course in Stockholm has more hills in the second half of the race, a runner who runs the first half in 90 minutes [1.5 hours] will typically struggle to complete the race in less than 3 hours)

5. (a) The least squares estimate $\widehat{\boldsymbol{\theta}}_{\text{LS}}$ is found using the normal equations

   i.
$$\widehat{\boldsymbol{\theta}}_{\text{LS}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y} = \left(\begin{bmatrix}1 & 1\\ -2 & 1\end{bmatrix}\underbrace{\begin{bmatrix}1 & -2\\ 1 & 1\end{bmatrix}}_{X}\right)^{-1}\begin{bmatrix}1 & 1\\ -2 & 1\end{bmatrix}\underbrace{\begin{bmatrix}1\\ -2\end{bmatrix}}_{\mathbf{y}} = \begin{bmatrix}-1\\ -1\end{bmatrix}$$

   ii. The bias is the expected difference between $\widehat{f}(x_\star) = x_\star^\mathsf{T}\widehat{\theta}$ and $f_0(x_\star) = x_\star^\mathsf{T}\theta^*$ (where θ^* is the unknown parameter, and $x_\star$ is a column vector with one in its first position $\begin{bmatrix}1 & -1\end{bmatrix}^\mathsf{T}$),

$$\begin{aligned}\mathbb{E}[\widehat{f}(x_\star) - f(x_\star)] &= \mathbb{E}[x_\star^\mathsf{T}\widehat{\theta} - x_\star^\mathsf{T}\theta^*]\\ &= x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbb{E}[\mathbf{y}] - x_\star^\mathsf{T}\theta\\ &= \left\{\mathbb{E}[\mathbf{y}] = \mathbb{E}[X\theta^* + \boldsymbol{\epsilon}] = X\theta^* + \mathbb{E}[\boldsymbol{\epsilon}] = X\theta\right\} =\\ &= x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}X\theta^* - x_\star^\mathsf{T}\theta^* =\\ &= x_\star^\mathsf{T}\theta^* - x_\star^\mathsf{T}\theta^*\\ &= 0\end{aligned}$$

   To compute the covariance, we start with

$$\begin{aligned}\mathbb{E}[f(x_\star;\widehat{\boldsymbol{\theta}})] &= \mathbb{E}[x_\star^\mathsf{T}\widehat{\theta}]\\ &= \mathbb{E}[x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}]\\ &= \mathbb{E}[x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}(X\theta + \boldsymbol{\epsilon})]\\ &= x_\star^\mathsf{T}\underbrace{(X^\mathsf{T}X)^{-1}X^\mathsf{T}X}_{I}\theta^* + x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\underbrace{\mathbb{E}[\boldsymbol{\epsilon}]}_{0}\\ &= x_\star^\mathsf{T}\theta^*,\end{aligned}$$

   which we can insert into the definition of the variance

$$\begin{aligned}\mathbb{E}[(f(x_\star;\widehat{\boldsymbol{\theta}} - \mathbb{E}[f(x_\star;\widehat{\boldsymbol{\theta}})])^2] &= \mathbb{E}[(x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}(X\theta+\epsilon) - x_\star^\mathsf{T}\theta)^2] =\\ &= \mathbb{E}[(\underbrace{x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}(X^\mathsf{T}X)\theta - x_\star^\mathsf{T}\theta}_{0} + x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\epsilon)^2] =\\ &= x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\underbrace{\mathbb{E}[\epsilon\epsilon^\mathsf{T}]}_{I\sigma^2}\left(x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\right)^\mathsf{T} =\\ &= x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}X\underbrace{\left((X^\mathsf{T}X)^{-1}\right)^\mathsf{T}}_{(X^\mathsf{T}X)^{-1}}x_\star\sigma^2 =\\ &= x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}x_\star\sigma^2,\end{aligned}$$

   which, when inserting numbers, gives

$$x_\star^\mathsf{T}(X^\mathsf{T}X)^{-1}x_\star\sigma^2 = \begin{bmatrix}1 & -1\end{bmatrix}\left(\begin{bmatrix}1 & 1\\ -2 & 1\end{bmatrix}\begin{bmatrix}1 & -2\\ 1 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}1\\ -1\end{bmatrix}1^2 = \frac{5}{9}$$

iii. In general, we can derive that (see exercise 5.3 for details)

$$
\begin{aligned}
\text{Cov}\left[\widehat{\boldsymbol{\theta}}_{\text{LS}}\right] &= \text{Cov}\left[(X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y}\right] \\
&= (X^\mathsf{T}X)^{-1}X^\mathsf{T}\text{Cov}\left[\mathbf{y}\right]\left((X^\mathsf{T}X)^{-1}X^\mathsf{T}\right)^\mathsf{T} \\
&= (X^\mathsf{T}X)^{-1}X^\mathsf{T}I\sigma^2\left((X^\mathsf{T}X)^{-1}X^\mathsf{T}\right)^\mathsf{T} \\
&= (X^\mathsf{T}X)^{-1}X^\mathsf{T}I\sigma^2 X\left((X^\mathsf{T}X)^{-1}\right)^\mathsf{T} \\
&= (X^\mathsf{T}X)^{-1}X^\mathsf{T}X\left((X^\mathsf{T}X)^{-1}\right)^\mathsf{T} I\sigma^2 \\
&= (X^\mathsf{T}X)^{-1}X^\mathsf{T}X(X^\mathsf{T}X)^{-1}\sigma^2 \\
&= (X^\mathsf{T}X)^{-1}\sigma^2,
\end{aligned}
$$

which, with X as above and $\sigma^2 = 1^2$, gives

$$
\text{Cov}\left[\widehat{\boldsymbol{\theta}}_{\text{LS}}\right] = \frac{1}{9}\begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}
$$

iv. The ridge regression estimate $\widehat{\boldsymbol{\theta}}_{\text{RR}}$ is computed as

$$
\widehat{\boldsymbol{\theta}}_{\text{RR}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\mathbf{y} =
$$

$$
\left(\begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix} + \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)^{-1}\begin{bmatrix} 1 & 1 \\ -2 & 1 \end{bmatrix}\underbrace{\begin{bmatrix} 1 \\ -2 \end{bmatrix}}_{\mathbf{y}} = -\frac{1}{\lambda^2 + 7\lambda + 9}\begin{bmatrix} \lambda + 9 \\ 4\lambda + 9 \end{bmatrix}
$$