1. Consider $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^\top, \mathbf{y} = [y_1, ..., y_N]^\top$ with $\mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R}$ for $i \in \{1, ..., N\}$. For $\theta_0 \in \mathbb{R}, \theta \in \mathbb{R}^D$, define $J_1(\mathbf{X}, \mathbf{y}, \theta, \theta_0) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\theta - \theta_0 \mathbf{1}\|_2^2$. Then, denote $\hat{\theta} = \arg\min_\theta J_1(\mathbf{X}, \mathbf{y}, \theta, \theta_0)$, $\hat{\theta}_0 = \arg\min_{\theta_0} J_1(\mathbf{X}, \mathbf{y}, \theta, \theta_0)$.

(a) Let $\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i, \bar{y} = \frac{1}{N}\sum_i y_i$. Show that $\hat{\theta}_0 = \bar{y} - \bar{\mathbf{x}}^\top \theta$.

solution.

$$\begin{aligned}\frac{\partial J_1}{\partial \theta_0} = 0 &\iff \sum_i y_i - \sum_i (\sum_j x_{ij}\theta_j) - n\hat{\theta}_0 = 0 \\ &\iff \hat{\theta}_0 = \frac{1}{N}\sum_i y_i - \frac{1}{N}\sum_i x_{ij}\theta_j = \bar{y} - \bar{\mathbf{x}}^\top \theta\end{aligned}$$

10 point if the proof is correct
-1 point per minor mistake

(b) Show that

$$\hat{\theta} = (\mathbf{X}_c^\top \mathbf{X}_c)^{-1}\mathbf{X}_c^\top \mathbf{y}_c = \left[\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\right]^{-1}\left[\sum_{i=1}^N (y_i - \bar{y})(\mathbf{x}_i - \bar{\mathbf{x}})\right]$$

where $\mathbf{X}_c$ is the centered input matrix containing $\mathbf{x}_i^c = \mathbf{x}_i - \bar{\mathbf{x}}$ along its rows, and $\mathbf{y}_c = \mathbf{y} - \bar{\mathbf{y}}$ is the centered output vector.

solution. By plugging the result of 1-(a) into J_1,

$$\begin{aligned}&(\mathbf{y} - \mathbf{X}\theta - \hat{\theta}_0\mathbf{1})^\top(\mathbf{y} - \mathbf{X}\theta - \hat{\theta}_0\mathbf{1}) = (\mathbf{y}_c - \mathbf{X}_c\theta)^\top(\mathbf{y}_c - \mathbf{X}_c\theta) \\ &\Rightarrow \frac{\partial J_1}{\partial \theta} = 0 \Leftrightarrow (\mathbf{y}_c - \mathbf{X}_c\theta)^\top \mathbf{X}_c = \mathbf{0} \\ &\Rightarrow \mathbf{X}_c^\top \mathbf{y}_c = \mathbf{X}_c^\top \mathbf{X}_c \theta \\ &\Rightarrow \hat{\theta} = (\mathbf{X}_\mathbf{c}^\top \mathbf{X}_\mathbf{c})^{-1}\mathbf{X}_c^\top \mathbf{y}_c\end{aligned}$$

10 point if the proof is correct
-1 point per minor mistake

2. Consider the ridge regression problem, i.e.,

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} (y_i - \theta_0 - \sum_{j=1}^{D} x_{ij}\theta_j)^2 + \lambda \sum_{j=1}^{D} \theta_j^2$$

Show that this problem is equivalent to the following optimization:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} (y_i - \theta_0 - \sum_{j=1}^{D} (x_{ij} - \bar{x}_j)\theta_j)^2 + \lambda \sum_{j=1}^{D} \theta_j^2$$

where $\bar{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}$.

solution.

$$\begin{aligned}
&\sum_{i=1}^{N} (y_i - \theta_0 - \sum_{j=1}^{D} x_{ij}\theta_j)^2 + \lambda \sum_{j=1}^{D} \theta_j^2 \\
=&\sum_{i=1}^{N} \left(y_i - \theta_0 + \sum_{j=1}^{D} \bar{x}_j\theta_j + \sum_{j=1}^{D} (x_{ij} - \bar{x}_j)\theta_j \right)^2 + \lambda \sum_{j=1}^{D} \theta_j^2
\end{aligned}$$

Define centered values of θ as

$$\theta_0^c = \theta_0 + \sum_{j=1}^{D} \bar{x}_j\theta_j$$

$$\theta_j^c = \theta_j, \quad j = 1, 2, ..., D$$

then we can get

$$J_2 := \sum_{i=1}^{N} \left(y_i - \theta_0^c + \sum_{j=1}^{D} (x_{ij} - \bar{x}_j)\theta_j^c \right)^2 + \lambda \sum_{j=1}^{D} \theta_j^{c2}$$

10 point if the proof is correct
-1 point per minor mistake

3. Assume $y_i \sim \mathcal{N}(\theta_0 + x_i^\top \theta, \sigma^2), i = 1, 2, ..., N$ and the parameters $\theta_j, j = 1, ..., D$ are each distributed as $\mathcal{N}(0, \tau^2)$, independently of each other. Assuming σ^2 and τ^2 are known, and θ_0 is not governed by a prior, show that the (minus) log-posterior density of θ is proportional to $\sum_{i=1}^N (y_i - \theta_0 - \sum_{j=1}^D x_{ij}\theta_j)^2 + \lambda \sum_{j=1}^D \theta_j^2 + C$ where $\lambda = \sigma^2/\tau^2$ and some constant C which is not dependent on θ.

solution.
We have

$$p(\theta) = C_1 \exp\left(-\frac{\|\theta\|^2}{2\tau^2}\right)$$

$$p(\mathbf{y}|\theta) = C_2 \exp\left(-\frac{\|\mathbf{y} - X\theta\|^2}{2\sigma^2}\right)$$

for appropriate constants C_1, C_2. It follows that for a suitable constant C_3,

$$p(\theta|\mathbf{y}) = C_3 \exp\left(-\frac{\|\mathbf{y} - X\theta\|^2 + (\sigma^2/\tau^2)\|\theta\|^2}{2\sigma^2}\right)$$

Then, we get

$$\begin{aligned} -\log p(\theta|\mathbf{y}) &= \frac{\|\mathbf{y} - X\theta\|^2 + (\sigma^2/\tau^2)\|\theta\|^2}{2\sigma^2} - \log C_3 \\ &= \frac{1}{2\sigma^2}(\|\mathbf{y} - X\theta\|^2 + \lambda^2\|\theta\|^2 - 2\sigma^2 \log C_3) \end{aligned}$$

10 point if the proof is correct
-1 point per minor mistake

4. Consider a set of random variables $(x_1, y_1), ..., (x_N, y_N)$ (training data) where each (x_i, y_i) is drawn from the same distribution $\mathcal{D}$ independently. Then, consider the least squares estimate $\hat{\theta} = \arg\min_\theta R_{\mathtt{train}}(\theta)$ for the linear regression model, where $R_{\mathtt{train}}(\theta) = \frac{1}{N}\sum_{i=1}^N (y_i - \theta^\top x_i)^2$. Now, consider another set of random variables $(\tilde{x}_1, \tilde{y}_1), ..., (\tilde{x}_M, \tilde{y}_M)$ (test data) where each $(\tilde{x}_i, \tilde{y}_i)$ is also drawn from the same distribution $\mathcal{D}$ independently. Prove that

$$\mathbb{E}[R_{\mathtt{train}}(\hat{\theta})] \leq \mathbb{E}[R_{\mathtt{test}}(\hat{\theta})],$$

where $R_{\mathtt{test}}(\theta) = \frac{1}{M}\sum_{i=1}^M (\tilde{y}_i - \theta^\top \tilde{x}_i)^2$.

solution.
The expectation of the test term $\frac{1}{M}\sum_i \left(\tilde{y}_i - \theta^\top \tilde{x}_i\right)^2$ is equal to the expectation of $\left(\tilde{y}_1 - \theta^\top \tilde{x}_1\right)^2$, and is therefore independent of M.

Without loss of generality, we take $M = N$, then decrease the test expression on replacing $\hat{\theta}$ with a value of θ that minimizes the expression.

Now the expectation of the two terms are equal which proves the result.

(FYI: Note that we may have to use the Moore-Penrose pseudo-invese of $X^\top X$, if the rank of X is less than p.)

10 point if the proof is correct
-1 point per minor mistake