STAT 154: Homework 4 Solutions

Release date: **Thursday, Mar 7**

Due by: **11 PM, Friday, Mar 22**

Homework Overview

This homework revisits ordinary least squares and other regression methods. Some problems are from the **ESL book** 12th printing (The Elements of Statistical Learning) that is available at the following website:
https://web.stanford.edu/ hastie/ElemStatLearn/.)

1 True or False (8 pts)

Examine whether the following statements are true or false and *provide one line justification.* Linear model in the following statements refers to the linear model studied in class (see equation (1) for a concrete reference.)

(a) Under the linear model, the OLS estimator of the regression coefficients is unbiased.
Ans. True

(b) For the linear model, bias of the ridge regression increases and the variance decreases as we increase the regularization parameter λ.
Ans. True

(c) The LASSO, relative to least squares, is less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
Ans. False

(d) In LASSO, no matter how large you choose regularization λ, the estimator $\hat{\beta}(\lambda)$ will never be the vector 0.
Ans. False

(e) In LASSO, as λ increases, the ℓ_1-norm of the estimator $\hat{\beta}(\lambda)$ always decreases.
Ans. True. False is OK, if one argues that it should be non-increasing rather than decrease.

(f) Every eigenvalue of an idempotent matrix is always either zero or one. (Recall that A square matrix $M \in \mathbb{R}^{m \times m}$ is called *idempotent* if $M^2 = M$.)
Ans. True

(g) Let X be an $n \times d$ matrix of full rank. Let $H = X\left(X^\top X\right)^{-1} X^\top$. The matrix H is symmetrix, idempotent and PSD.
Ans. True

(h) Let $Q = \mathbb{I}_n - H$ where H is defined in the previous part. We have trace$(Q) = n$.
Ans. False

Fun fact: Note that for a feature matrix X with full column rank the matrix $H = X\left(X^\top X\right)^{-1} X^\top$ is called the *hat matrix* because in ordinary least squares, the predicted responses are given by $\hat{y} = Hy$, i.e., the matrix H adds a hat to y.

2 OLS theoretical properties (9 pts)

1. (2 pts) Consider a linear model where we observe the samples (x_i, y_i) for $i = 1, \ldots, n$, that are generated as follows

$$y_i = x_i^\top \beta^* + \epsilon, \tag{1}$$

where the error $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. **Show that the OLS estimate $\hat{\beta}$ on data $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^n$ satisfies**

$$\hat{\beta} \sim \mathcal{N}\left(\beta^*, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right).$$

Also conclude that $\mathbb{E}\left[\mathbf{X}\hat{\beta}\right] = \mathbf{X}\beta^*$.

Ans. Since OLS estimate gives $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$, plugging in the linear model, we have

$$\hat{\beta} = \beta^* + \left(X^\top X\right)^{-1} X^\top \boldsymbol{\epsilon}.$$

For the covariance, we have

$$\sigma^2 \left(X^\top X\right)^{-1} X^\top X \left(X^\top X\right)^{-1} = \sigma^2 \left(X^\top X\right)^{-1}.$$

$\mathbb{E}[\boldsymbol{\epsilon}] = 0$ **so $\hat{\beta}$ is unbiased. Hence we have unbiasedness of $\mathbf{X}\hat{\beta}$.**

2. (2 pts) In the Gaussian linear model described above, **show that** $\mathbb{E}[RSS] = \sigma^2(n-d)$.

Ans.

$$\begin{aligned} RSS &= \left\| \mathbf{X}\hat{\beta} - Y \right\|_2^2 \\ &= \left\| \left(\mathbb{I}_n - X\left(X^\top X\right)^{-1} X^\top\right) \boldsymbol{\epsilon} \right\|_2^2 \\ &= \mathbf{trace}\left(\left(\mathbb{I}_n - X\left(X^\top X\right)^{-1} X^\top\right) \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \left(\mathbb{I}_n - X\left(X^\top X\right)^{-1} X^\top\right)\right) \end{aligned}$$

Taking expectation, we have

$$\begin{aligned}
\mathbb{E}[RSS] &= \textbf{trace}\left(\left(\mathbb{I}_n - X\left(X^\top X\right)^{-1}X^\top\right)^2\right) \\
&\overset{(i)}{=} \textbf{trace}\left(\left(\mathbb{I}_n - X\left(X^\top X\right)^{-1}X^\top\right)\right) \\
&= n - \textbf{trace}\left(X\left(X^\top X\right)^{-1}X^\top\right) \\
&= n - \textbf{trace}\left(\left(X^\top X\right)^{-1}X^\top X\right) \\
&= n - \textbf{trace}\left(\mathbb{I}_d\right) \\
&= n - d.
\end{aligned}$$

Here inequality (i) uses the fact that the matrix $\left(\mathbb{I}_n - X\left(X^\top X\right)^{-1}X^\top\right)$ is a projection matrix (or one can also extend the square to verify).

3. (2 pts) **ESL book Ex. 3.3 (a)** (part (b) is not needed). *Hint*: The notion unbiased linear estimator is explained in Section 3.2.2 around equation (3.19).
Ans. Take $\tilde{\theta} = c^\top y$ to be an unbiased linear estimator of $a^\top\beta^*$. That is, $\mathbb{E}\left[c^\top y\right] = a^\top\beta^*$. We have, for all $\beta^* \in \mathbb{R}^d$

$$c^\top X\beta^* = a^\top\beta^*.$$

Since the above equality is true for all β^*, we can choose β^* to be canonical vectors to make sure that each coordinate the following two vectors is identical and hence

$$c^\top X = a^\top.$$

Now we calculate the variance.

$$\textbf{Var}(c^\top y) = \sigma^2 c^\top c.$$

$$\textbf{Var}(a^\top\hat{\beta}) = \sigma^2 a^\top(X^\top X)^{-1}a = \sigma^2 c^\top X(X^\top X)^{-1}X^\top c.$$

We conclude by noting that $X(X^\top X)^{-1}X \preceq \mathbb{I}_n$.

4. (3 pts) **ESL book Ex. 2.9**. We expect a solution that **does not** use the explicit data generation process. *Partial credit is given if you use the explicit data generation process.*

Ans. Let (X, Y) be the training data and $(\tilde{X}, \tilde{Y})$ be the test data. According to OLS formula, we have

$$\hat{\beta} = \left(X^\top X\right)^{-1}X^\top Y.$$

In matrix notation, $R_{tr}(\beta) = \frac{1}{N}\|Y - X\beta\|_2^2$ and $R_{te}(\beta) = \frac{1}{M}\left\|\tilde{Y} - \tilde{X}\beta\right\|_2^2$. Note that

$$\mathbb{E}_{\tilde{X},\tilde{Y}|X,Y}\left[R_{te}(\hat{\beta})\right] = \mathbb{E}_{\tilde{x}_1,\tilde{y}_1|X,Y}\left(\tilde{y}_1 - \tilde{x}_1^\top\hat{\beta}\right)^2.$$

Hence M does not really matter. Without loss of generality, we can assume that $M = N$. Define

$$\tilde{\beta} = \left(\tilde{X}^\top\tilde{X}\right)^{-1}\tilde{X}^\top\tilde{Y}.$$

Observe that $\tilde{\beta}$ is the minimizer of R_{te}, we have

$$R_{te}(\tilde{\beta}) \le R_{te}(\hat{\beta}). \tag{2}$$

On the other hand, since (X, Y) and $(\tilde{X}, \tilde{Y})$ are generated from the same distribution, we have

$$\mathbb{E}_{X,Y}\left[R_{tr}(\hat{\beta})\right] = \mathbb{E}_{\tilde{X},\tilde{Y}}\left[R_{te}(\tilde{\beta})\right]. \tag{3}$$

Combining equation (2) and (3), we have

$$\mathbb{E}_{X,Y}\left[R_{tr}(\hat{\beta})\right] \le \mathbb{E}_{X,Y,\tilde{X},\tilde{Y}}\left[R_{te}(\hat{\beta})\right]$$

3 Theory of Ridge Regression (14 pts)

Given the feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and the responses $\mathbf{y} \in \mathbb{R}^n$, ridge regression solves the following penalized least squares problem:

$$\min_\theta \|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \lambda\|\theta\|_2^2 \tag{4}$$

Let $\mathbf{X}^\top\mathbf{X}$ have the following eigen-decomposition: $\mathbf{X}^\top\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ where $\mathbf{D}$ is a diagonal matrix with non-negative entries.

Let $\widehat{\theta}_\lambda$ denote the solution for the problem (4) above.

1. (1 pt) **Show that for any $\lambda > 0$, the solution $\widehat{\theta}_\lambda$ is unique and derive its expression.**
   Ans. $\widehat{\theta}_\lambda$ is unique because setting gradient of equation (4) has unique solution due to the fact that $X^\top X + \lambda\mathbb{I}_d$ is always invertible.

$$\widehat{\theta}_\lambda = \left(X^\top X + \lambda\mathbb{I}_d\right)^{-1}X^\top\mathbf{y}$$

2. (2 pts) For *all the following parts*, we assume the linear regression model: That is the data is generated as follows:

$$\mathbf{y} = \mathbf{X}\theta^* + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Using the previous part, **show that the distribution of the ridge-estimate is given as follows:**

$$\widehat{\theta}_\lambda \sim \mathcal{N}(\mathbf{W}_\lambda \theta^*, \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1})$$
$$\text{where} \quad \mathbf{W}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} \mathbf{X}^\top \mathbf{X}.$$

Ans. Using the expression in the last question, plugging in the linear model, we have

$$\widehat{\theta}_\lambda = \left(X^\top X + \lambda \mathbb{I}_d\right)^{-1} X^\top \left(X\theta^* + \boldsymbol{\epsilon}\right).$$

Using the W_λ expression, we find that

$$\widehat{\theta}_\lambda = W_\lambda \theta^* + \left(X^\top X + \lambda \mathbb{I}_d\right)^{-1} X^\top \boldsymbol{\epsilon}.$$

$\widehat{\theta}_\lambda$ is Gaussian distributed. Then it is sufficient to determine its mean and variance.

$$\mathbb{E}[\widehat{\theta}_\lambda] = W_\lambda \theta^*,$$

and

$$\begin{aligned}\mathbf{Cov}[\widehat{\theta}_\lambda] &= \mathbb{E}\left[\left(X^\top X + \lambda \mathbb{I}_d\right)^{-1} X^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top X \left(X^\top X + \lambda \mathbb{I}_d\right)^{-1}\right] \\ &= \sigma^2 W_\lambda \left(X^\top X + \lambda \mathbb{I}_d\right)^{-1}\end{aligned}$$

3. (4 pts) Let $\mathbf{X}^\top \mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ with $\mathbf{D}_{ii} = d_i$. **Show the following:**

(a) The squared bias is given by

$$\textbf{squared bias} = \left\|\mathbb{E}[\widehat{\theta}_\lambda] - \theta^*\right\|_2^2 = \sum_{i=1}^{d} \frac{\lambda^2}{(d_i + \lambda)^2} (v_i)^2,$$

where $v = \mathbf{U}^\top \theta^*$.

Ans. Using the eigenvalue decomposition, we first obtain

$$(X^\top X + \lambda \mathbb{I}_d)^{-1} = U(D + \lambda \mathbb{I}_d)^{-1} U^\top.$$

Then we have

$$\begin{aligned} W_\lambda - \mathbb{I}_d &= (X^\top X + \lambda \mathbb{I}_d)^{-1} X^\top X - \mathbb{I}_d \\ &= (X^\top X + \lambda \mathbb{I}_d)^{-1} (-\lambda \mathbb{I}_d) \\ &= U (D + \lambda \mathbb{I}_d)^{-1} U^\top (-\lambda \mathbb{I}_d) \\ &= U \left[-\lambda (D + \lambda \mathbb{I}_d)^{-1} \right] U^\top. \end{aligned}$$

Using the expression of $\mathbb{E}[\widehat{\theta}_\lambda]$, we have

$$\begin{aligned} \textbf{squared bias} &= \left\| (W_\lambda - \mathbb{I}_d)\, \theta^* \right\|_2^2 \\ &= \theta^{*\top} U \left[-\lambda (D + \lambda \mathbb{I}_d)^{-1} \right] \left[-\lambda (D + \lambda \mathbb{I}_d)^{-1} \right] U^\top \theta^* \\ &= v^\top \left[-\lambda (D + \lambda \mathbb{I}_d)^{-1} \right] \left[-\lambda (D + \lambda \mathbb{I}_d)^{-1} \right] v \\ &= \sum_{i=1}^{d} \frac{\lambda^2}{(\lambda + d_i)^2} v_i^2. \end{aligned}$$

(b) The (scalar) variance is given by

$$\textbf{scalar-variance} = \mathbb{E} \left\| \widehat{\theta}_\lambda - \mathbb{E}[\widehat{\theta}_\lambda] \right\|_2^2 = \sigma^2 \sum_{i=1}^{d} \frac{d_i}{(d_i + \lambda)^2}.$$

Ans.

$$\begin{aligned} \textbf{scalar-variance} &= \textbf{trace} \left(\textbf{Cov}[\widehat{\theta}_\lambda] \right) \\ &= \sigma^2 \textbf{trace} \left(W_\lambda \left(X^\top X + \lambda \mathbb{I}_d \right)^{-1} \right) \\ &= \sigma^2 \textbf{trace} \left(U \left(D + \lambda \mathbb{I}_d \right)^{-1} D \left(D + \lambda \mathbb{I}_d \right)^{-1} U^\top \right) \\ &\overset{(i)}{=} \sigma^2 \textbf{trace} \left(\left(D + \lambda \mathbb{I}_d \right)^{-1} D \left(D + \lambda \mathbb{I}_d \right)^{-1} U^\top U \right) \\ &\overset{(ii)}{=} \sigma^2 \textbf{trace} \left(\left(D + \lambda \mathbb{I}_d \right)^{-1} D \left(D + \lambda \mathbb{I}_d \right)^{-1} \right) \\ &= \sigma^2 \sum_{i=1}^{d} \frac{d_i}{(d_i + \lambda)^2}. \end{aligned}$$

Here equality (i) uses the fact that $\textbf{trace}(AB) = \textbf{trace}(BA)$. Equality (i) uses U is an orthogonal matrix.

Note that here we are considering the scalar versions of the bias and variance defined in class by taking the norms of the corresponding quantities.

4. (2 pts) **What is the value of the squared-bias and variance at $\lambda = 0$ and as $\lambda \to \infty$. Do you see a trade-off in the squared-bias and variance change as λ increases?**
Ans.

- $\lambda = 0$: **squared bias** $= 0$ **and scalar-variance** $= \sigma^2 \sum_{i=1}^{d} \frac{1}{d_i}$.
- $\lambda = \infty$: **squared bias** $= \|v\|_2^2$ **and scalar-variance** $= 0$.

Squared bias is increasing with λ, while scalar-variance is decreasing with λ. There will be a trade-off.

5. (2 pts) Recall the definition of the moment matrix $\mathbf{M}$ from class:

$$\mathbf{M}(\lambda) = \mathbb{E}[(\widehat{\theta}_\lambda - \theta^*)(\widehat{\theta}_\lambda - \theta^*)^\top].$$

Recall the mean-squared error

$$\mathrm{MSE}(\widehat{\theta}_\lambda) := \mathbb{E}[\left\|\widehat{\theta}_\lambda - \theta^*\right\|_2^2].$$

Show that $\mathbb{E}[\left\|\widehat{\theta}_\lambda - \theta^*\right\|_2^2] = \mathrm{trace}(\mathbf{M}(\lambda))$**. Moreover, show that**

$$\textbf{MSE}(\widehat{\theta}_\lambda) = \textbf{squared-bias} + \textbf{scalar-variance}$$

Ans. $\mathbb{E}[\left\|\widehat{\theta}_\lambda - \theta^*\right\|_2^2] = \mathrm{trace}(\mathbf{M}(\lambda))$ **is just a result of** $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ **and both trace and** $\mathbb{E}$ **are linear. The usual trick to derive the bias-variance trade-off is to add and substract the expectation term.**

$$\begin{aligned}
\textbf{MSE}(\widehat{\theta}_\lambda) &= \mathbb{E}[\left\|\widehat{\theta}_\lambda - \theta^*\right\|_2^2] \\
&= \mathbb{E}[\left\|\widehat{\theta}_\lambda - \mathbb{E}[\widehat{\theta}_\lambda] + \mathbb{E}[\widehat{\theta}_\lambda] - \theta^*\right\|_2^2] \\
&\overset{(i)}{=} \mathbb{E}[\left\|\widehat{\theta}_\lambda - \mathbb{E}[\widehat{\theta}_\lambda]\right\|_2^2] + \left\|\mathbb{E}[\widehat{\theta}_\lambda] - \theta^*\right\|_2^2 \\
&= \textbf{scalar-variance} + \textbf{squared-bias}
\end{aligned}$$

The cross-term does not appear in step (i) because $\mathbb{E}[\widehat{\theta}_\lambda - \mathbb{E}[\widehat{\theta}_\lambda]] = 0$.

6. (3 pts) Recall that when $\mathbf{X}^\top\mathbf{X}$ is invertible, the OLS-estimator is unbiased. For this case, **show that its mean squared error satisfies** $\textbf{MSE}(\theta^{\mathrm{OLS}}) = \mathrm{trace}(\mathbf{M}(0))$**. Furthermore, show that there exists a range of** $\lambda > 0$ **for which**

$$\mathrm{MSE}(\widehat{\theta}_\lambda) < \mathrm{MSE}(\theta^{\mathrm{OLS}}).$$

Conclude that there always exists a range of $\lambda > 0$, for which the MSE is smaller for ridge regression when compared to OLS in the Gaussian linear model.

Ans. OLS can be seen a special case of ridge by taking $\lambda = 0$. Thus $\textbf{MSE}(\theta^{\mathrm{OLS}}) = \mathrm{trace}(\mathbf{M}(0))$ from previous question.

Since the eigenvalue decomposition of $X^\top X$ in question 3.3 will always exist, we can use the results in question 3.3.

$$\textbf{MSE}(\widehat{\theta}_\lambda) = \sum_{i=1}^{d} \frac{\lambda^2}{(\lambda + d_i)^2} v_i^2 + \sigma^2 \sum_{i=1}^{d} \frac{d_i}{(\lambda + d_i)^2}.$$

Taking derivative with respect to λ, we have

$$\frac{d\textbf{MSE}(\widehat{\theta}_\lambda)}{d\lambda} = \sum_{i=1}^{d} \frac{2d_i(\lambda v_i^2 - \sigma^2)}{(\lambda + d_i)^3}.$$

Taking $\lambda = 0$, we have

$$\frac{d\textbf{MSE}(\widehat{\theta}_\lambda)}{d\lambda}|_{\lambda=0} < 0.$$

Thus $\textbf{MSE}(\widehat{\theta}_\lambda)$ is a decreasing function around 0. There always exists a range of λ where MSE is smaller for ridge.

4 Gradient descent for simple functions (8 pts)

A function $f : \mathbb{R}^d \to \mathbb{R}$ is called *convex*, if

$$\forall x \in \mathbb{R}^d, y \in \mathbb{R}^d, \lambda \in [0, 1], f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

Recall Taylor's theorem for twice differentiable functions of vectors, which holds for all $x, y \in \mathbb{R}^d$:

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(\widetilde{x})(y - x),$$

for some $\widetilde{x}$. One can show that a *twice differentiable function* f is convex if and only if $\nabla^2 f(x)$ is positive semidefinite for all $x \in \mathbb{R}^d$. You can take this fact for granted.[1]

1. (2 pts) Let $L \geq 0$. Consider the function of one variable $f(x) = \frac{L}{2}x^2$. **Show that it is convex. Derive the gradient descent update where we use a step-size of γ and start at some point $x^{(0)} \neq 0$.**
   Ans. From the above discussion, it suffices to show that $\nabla^2 f(x)$ is PSD for all x. Since we are in one dimension, $\nabla^2 f(x) = L \geq 0$.
   Ans. The gradient is given by $f'(x) = Lx$. The gradient descent update is therefore given by

$$\begin{aligned} x^{(i+1)} &= x^{(i)} - \gamma \nabla f(x^{(i)}) \\ &= (1 - \gamma L) x^{(i)}. \end{aligned}$$

[1] This problem draws inspiration from the class CS 189.

2. (3 pts) **What does the behavior of the updates look like for the above setting and the choices $\gamma \in \{1/L, 2/L\}$? What happens when we use a step size $\gamma \in \left[0, \frac{2}{L}\right)$ such that $\gamma \neq 1/L$? For this step size, how many steps does it take for gradient descent updates to converge to within ϵ of the optimum (as a function of the tuple $(\gamma, L, |x^{(0)}|, \epsilon)$)?**
Ans. Plugging the choice $\gamma = 1/L$ into the update, we see that $x^{(1)} = 0$, and so we reach the optimal solution in just one step. On the other hand, the choice $\gamma = 2/L$ yields $x^{(1)} = -x^{(0)}$, and so we oscillate between the points $x^{(0)}$ and $-x^{(0)}$ forever.
Ans. Iterating the gradient descent update, we have

$$x^{(i)} = (1 - \gamma L)^i x^{(0)}.$$

We know that the optimum is at 0, and so we would like $|x^{(i)}| \leq \epsilon$. Thus, we need

$$(1 - \gamma L)^i \leq \epsilon / |x^{(0)}|.$$

Simplifying, we see that setting $i \geq \frac{\log \frac{\epsilon}{|x^{(0)}|}}{\log(1-\gamma L)}$ suffices.
A better way to see the scaling of the problem is to use the inequality $1 - t \leq e^{-t}$, which holds for all scalar t. Thus, it suffices to have

$$\left(e^{-\gamma L}\right)^i \leq \epsilon / |x^{(0)}|,$$

and simplifying yields that $i \geq \frac{1}{\gamma L} \log(|x^{(0)}/\epsilon)$ is sufficient.

3. (1 pt) **How do your answers in the previous part change if $f(x) = \frac{L}{2}(x - c)^2$ for some constant c?**

Ans. The gradient changes, but the behavior of the algorithm does not since the optimum also changes to being at $x = c$.

4. (2 pts) Let $L \geq m \geq 0$. Now consider the function of two variables $f(x) = \frac{L}{2}x_1^2 + \frac{m}{2}x_2^2$. **Show that the function is convex by computing its Hessian $\nabla^2 f(x)$. Derive closed form expressions for the iterations if we start at the point (a, b), and run gradient descent with step-size γ.** Start by writing out the result of the first iteration as $A \begin{bmatrix} a \\ b \end{bmatrix}$ for some matrix A.
Ans. In order to compute the Hessian, we can compute the "gradient" of the gradient. We have $\frac{\partial f}{\partial x_1} = Lx_1$, and $\frac{\partial f}{\partial x_2} = Lx_2$. Differentiating once more, we have

$$\frac{\partial^2 f}{(\partial x_1)^2} = L$$

$$\frac{\partial^2 f}{(\partial x_2)^2} = m$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = 0$$

$$\frac{\partial^2 f}{\partial x_2 \partial x_1} = 0.$$

Thus, the Hessian is given by the diagonal matrix $\begin{bmatrix} L & 0 \\ 0 & m \end{bmatrix}$, **which is clearly positive semidefinite.**
Ans. As we derived above, the gradient of the function is given by

$$\nabla f(x) = \begin{bmatrix} Lx_1 \\ mx_2 \end{bmatrix},$$

and so the first iterate is given by

$$x_1^{(1)} = (1 - \gamma L)x_1^{(0)}$$
$$x_2^{(1)} = (1 - \gamma m)x_2^{(0)}.$$

Writing this in matrix form, we have

$$x^{(1)} = \begin{bmatrix} (1 - \gamma L) & 0 \\ 0 & (1 - \gamma m) \end{bmatrix} x^{(0)}.$$

Denoting the matrix by A, the ith iterate therefore takes the form $x^{(i)} = A^i x^{(0)}$.

5 High dimensional regression (9 pts, readable code snippet required in the write-up)

Suppose we have a design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ where $d > n$, i.e., we have many more predictor variables than observations. Moreover, we have quantitative response vector $\mathbf{y} \in \mathbb{R}^n$, and we plan to fit a linear regression model.

1. (2 pts) **Explain why the ordinary least squares solution for $(\mathbf{X}, \mathbf{y})$ is not unique. What can you say about the residuals of any OLS solution?**
Ans. When $d > n$, X is not full column rank (rank d). It means that there exits $\theta_1 \neq 0$ satisfying $X\theta_1 = 0$. If $\hat{\beta}$ is OLS one solution, then $\hat{\beta} + \theta_1$ is another solution. So the OLS solutions are not unique.
If in addition, X is full row rank (rank n), then there always exists $\tilde{\beta}$ such that

$$X\tilde{\beta} = Y.$$

Then the residuals is the vector zero. It is wrong to say that the residuals is zero vector without assuming X is full row rank (rank n).

2. (1 pt) **Is the ridge regression solution unique? Why or why not?**
Ans. When $\lambda > 0$, the ridge regression solution is unique. We show it by

contradiction. Assume that we have two distinct ridge solutions $\hat{\beta}_1$ and $\hat{\beta}_2$, then we can construct a new estimate $\hat{\beta}_{\text{new}} = (\hat{\beta}_1 + \hat{\beta}_2)/2$. $\hat{\beta}_{\text{new}}$ satisfies that

$$\left\|X\hat{\beta}_{\text{new}} - \mathbf{y}\right\|_2^2 = \left\|X\hat{\beta}_1 - \mathbf{y}\right\|_2^2 \text{ and } \left\|\hat{\beta}_{\text{new}}\right\|_2^2 < \frac{\left\|\hat{\beta}_1\right\|_2^2 + \left\|\hat{\beta}_1\right\|_2^2}{2}.$$

Hence the new estimate $\hat{\beta}_{\text{new}}$ achieves a strictly smaller ridge objective than that of $\hat{\beta}_1$. This is contradictory to the fact that $\hat{\beta}_1$ is the minimizer. Alternatively, first showing the ridge objective is strongly convex, then inoke paper/result or show by themselves about strongly convex objective has only one minimum is OK.

3. (1 pt) Suppose you compute a series of ridge solutions $\widehat{\beta}(\lambda)$ for $\mathbf{X}$ and $\mathbf{y}$, letting λ get monotonically smaller. **What can you say about the limiting ridge solution as $\lambda \downarrow 0$?**
   Ans. It becomes one OLS solution.

4. (1 pt) (Code) Fix $n = 1000, d = 5000$. Generate a design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with each entry drawn i.i.d. from $\mathcal{N}(0, 1)$, an error vector $\boldsymbol{\epsilon} \in \mathbb{R}^n$ with each entry drawn i.i.d. from $\mathcal{N}(0, 0.25)$ and a response vector $\mathbf{y} \in \mathbb{R}^n$, satisfying

$$\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\epsilon},$$

   where $\beta^* = (\underbrace{1, \ldots, 1}_{15}, 0, \ldots, 0)^\top$ having 15 non-zero entries. **Show your code for this part in the write-up. *Do not* display any data.**

```
library(ggplot2)
library(glmnet)
set.seed(123456)
n = 1000
d = 5000
X = matrix(rnorm(n * d), nrow = n)
error = rnorm(n, mean = 0, sd = 0.5)
betastar <- as.vector(rep(0, d))
betastar[1:15] = 1.
Y = X %*% betastar + error
```

5. (1 pt) (Code) Split the n samples into training (size $4n/5$) and test (size $n/5$) sets. **Show your code for this part in the write-up**.

```
test_size <- n / 5
test_ind <- sample(n, size = test_size)
Xtest <- X[test_ind, ]
Ytest <- Y[test_ind]
Xtrain <- X[-test_ind, ]
Ytrain <- Y[-test_ind]
```

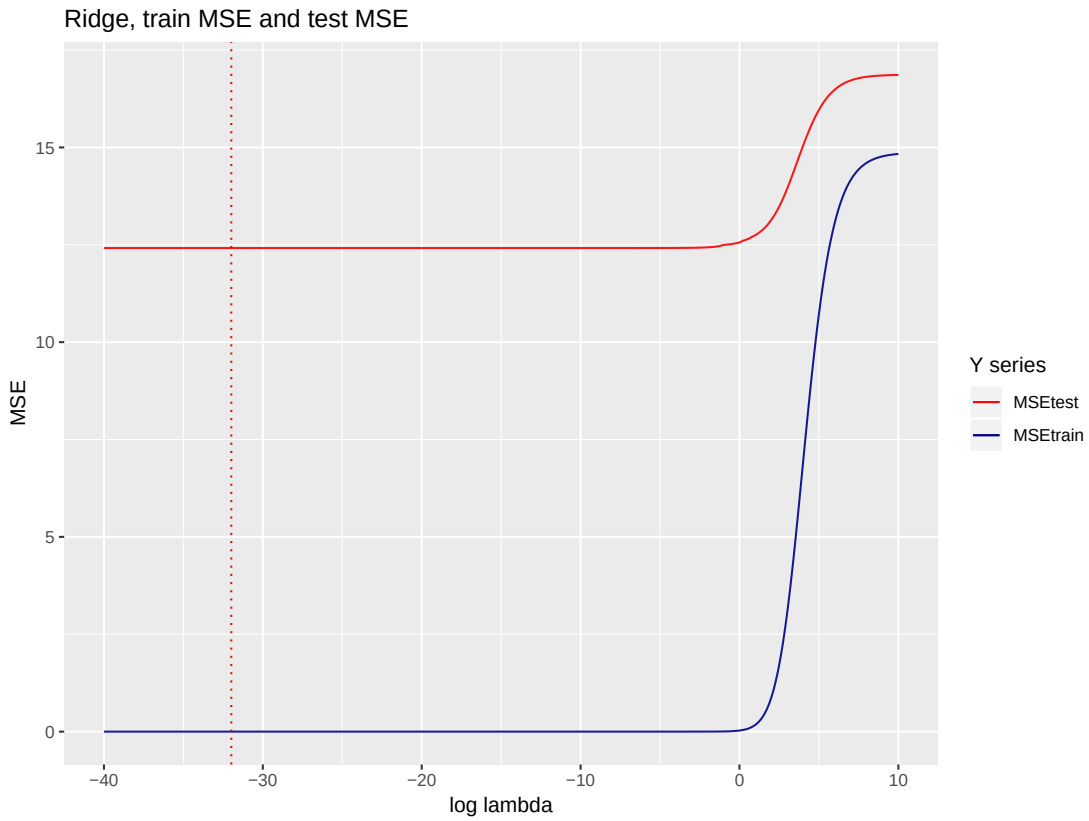
6. (1 pt) (Code) Using the **glmnet** package, fit ridge regression on the training data. Plot the training MSE vs lambda and test MSE vs lambda in the same plot with two different colors. We expect you to choose a large range of lambdas so that the test MSE is not monotone. **Show the plots and the code for this part in the write-up.**

```
lambdasRidge <- exp(seq(10, -40, length = 500))
fitRidge <- glmnet(Xtrain, Ytrain, lambda = lambdasRidge, family="gaussian",
                   alpha=0, standardize = F, intercept = F)
YhatTrainRidge <- predict(fitRidge, Xtrain)
YhatTestRidge <- predict(fitRidge, Xtest)

MSEtrainRidge <- rep(NA, length(lambdasRidge))
MSEtestRidge <- rep(NA, length(lambdasRidge))

for (i in 1:length(lambdasRidge)) {
  MSEtrainRidge[i] <- mean((Ytrain - YhatTrainRidge[,i])^2)
  MSEtestRidge[i] <- mean((Ytest - YhatTestRidge[,i])^2)
}

ggplot() + geom_line(aes(x=log(lambdasRidge), y=MSEtrainRidge, color="MSEtrain"),
            alpha=0.9) +
  geom_line(aes(x=log(lambdasRidge), y=MSEtestRidge, color="MSEtest"),
            alpha=0.9) +
  geom_vline(xintercept=log(lambdasRidge)[which.min(MSEtestRidge)],
             linetype="dotted", color="red") +
  scale_color_manual(values = c(
    'MSEtrain' = 'darkblue',
    'MSEtest' = 'red')) +
  labs(color = 'Y series') +
  labs(x = "log lambda", y = "MSE", title = "Ridge, train MSE and test MSE")
```



7. (1 pt) (Code) Using the **glmnet** package, fit LASSO regression on the training data. Plot the training MSE vs lambda and test MSE vs lambda in the same plot with two different colors. We expect you to choose a large range of lambdas so that the test MSE is not monotone. **Show the plots and the code for this part in the write-up.**

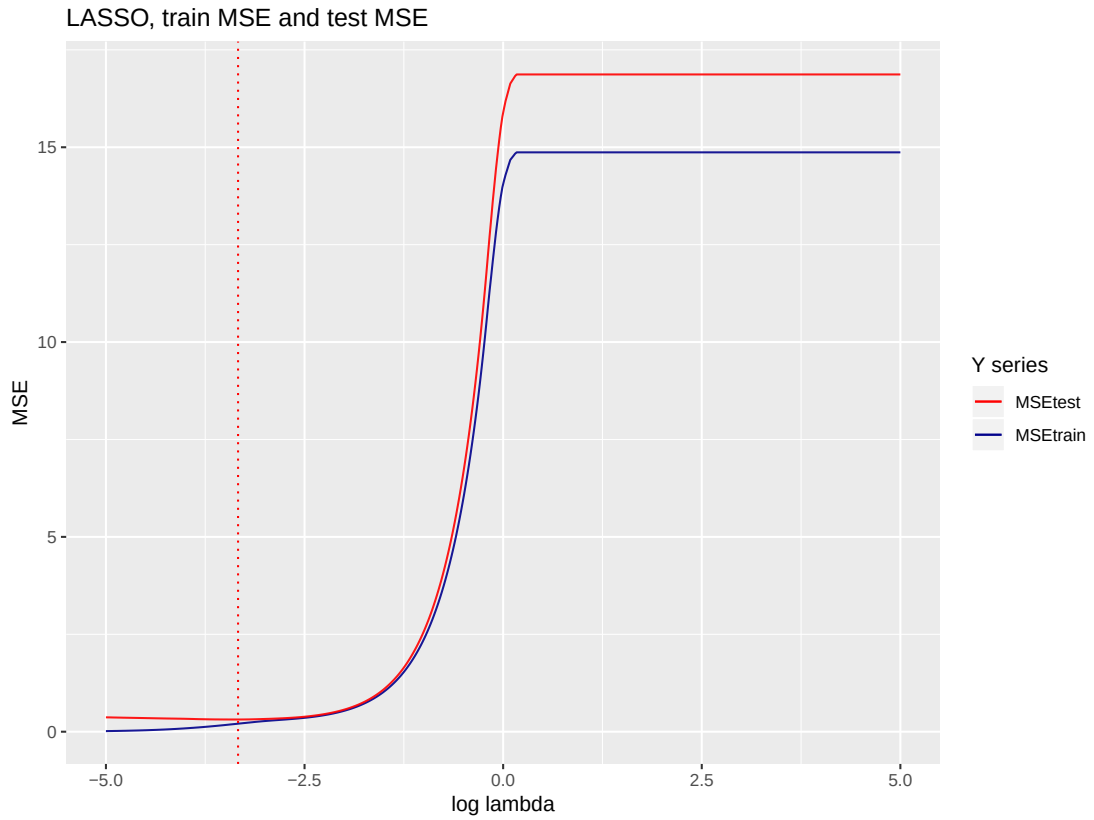
```
lambdasLASSO <- exp(seq(5, -5, length = 500))
fitLASSO <- glmnet(Xtrain, Ytrain, lambda = lambdasLASSO, family="gaussian",
                   alpha=1, standardize = F, intercept = F)
YhatTrainLASSO <- predict(fitLASSO, Xtrain)
YhatTestLASSO <- predict(fitLASSO, Xtest)

MSEtrainLASSO <- rep(NA, length(lambdasLASSO))
MSEtestLASSO <- rep(NA, length(lambdasLASSO))

for (i in 1:length(lambdasLASSO)) {
  MSEtrainLASSO[i] <- mean((Ytrain - YhatTrainLASSO[,i])^2)
  MSEtestLASSO[i] <- mean((Ytest - YhatTestLASSO[,i])^2)
}

ggplot() + geom_line(aes(x=log(lambdasLASSO), y=MSEtrainLASSO, color="MSEtrain"),
            alpha=0.9) +
```

```
  geom_line(aes(x=log(lambdasLASSO), y=MSEtestLASSO, color="MSEtest"),
            alpha=0.9) +
  geom_vline(xintercept=log(lambdasLASSO)[which.min(MSEtestLASSO)],
             linetype="dotted", color="red") +
  scale_color_manual(values = c(
    'MSEtrain' = 'darkblue',
    'MSEtest' = 'red')) +
  labs(color = 'Y series') +
  labs(x = "log lambda", y = "MSE", title = "LASSO, train MSE and test MSE")
```



8. (1 pt) **Which model ridge or LASSO and what regularization parameter has the smallest test MSE?**
   Ans.

   - **Ridge:** $\lambda = 1.29e-14$.
   - **LASSO:** $\lambda = 3.56e-2$.

```
lambdasRidge[which.min(MSEtestRidge)]

## [1] 1.286883e-14
```

```
lambdasLASSO[which.min(MSEtestLASSO)]

## [1] 0.03555504
```