Question 3.1: Consider the observation model

\[ y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2), \tag{3.1} \]

where \( f : \mathbb{R}^d \to \mathbb{R} \) is \( f(x_i) = \mathbb{E}[y_i | x_i] \). Let \( X \in \mathbb{R}^{n \times d}, n \geq d, \) be a full-rank data matrix with singular value decomposition \( X = U \Gamma V^T \), so \( U \in \mathbb{R}^{n \times d}, V \in \mathbb{R}^{d \times d}, \) and \( U^T U = I_d = V^T V = V V^T \).

We use the notation \( U = \begin{bmatrix} u_1 & \cdots & u_d \end{bmatrix} \) and \( V = \begin{bmatrix} v_1 & \cdots & v_d \end{bmatrix} \). We consider a principal-components ridge regression estimator

\[ \hat{\beta}_\lambda := \arg\min_b \left\{ \|Xb - y\|_2^2 + b^T \Lambda V^T b \right\}, \]

where \( \Lambda = \text{diag}(\lambda) = \text{diag}(\lambda_1, \ldots, \lambda_d) \) is a positive semidefinite matrix (so \( \lambda_j \geq 0 \) for all \( j \), i.e. the vector \( \lambda \in \mathbb{R}_+^d \)). Here the idea is that we can penalize the scale of \( \beta \) in the directions of the various principal components in a more intelligent way than naive ridge regression, perhaps shrinking components where we have less “information” more aggressively than others.

Define

\[ \hat{y}_\lambda = X \hat{\beta}_\lambda = H_\lambda y, \quad H_\lambda = X(X^T X + V \Lambda V^T)^{-1} X^T. \]

(a) Show that \( \hat{\beta}_\lambda = V \Gamma (\Gamma^2 + \Lambda)^{-1} U^T X \) and \( H_\lambda = U \Gamma (\Gamma^2 + \Lambda)^{-1} U^T \).

Recall the in-sample risk

\[ R_{in}(\hat{\beta}_\lambda) := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\hat{y}_i - f(x_i))^2] = \frac{1}{n} \mathbb{E} \left[ \|\hat{y}_\lambda - \mathbb{E}[y]\|_2^2 \right]. \]

(b) Show that the residual sum of squares satisfies

\[ \frac{1}{n} \mathbb{E} \left[ \|\hat{y}_\lambda - y\|_2^2 \right] = R_{in}(\hat{\beta}_\lambda) + \sigma^2 - \frac{2}{n} \sum_{i=1}^n \text{Cov}(\hat{y}_i, y_i). \]

Conclude that

\[ \frac{1}{n} \mathbb{E} \left[ \|\hat{y}_\lambda - y\|_2^2 \right] + \frac{2\sigma^2}{n} \text{tr}(H_\lambda) = R_{in}(\hat{\beta}_\lambda) + \sigma^2. \]

(Note: we have basically done this in class. It is not a trick question.)

By part (b), you have shown the fact (as proved in class) that residual sum of squares plus a trace is unbiased for in-sample risk, and so in particular if \( \hat{\sigma}^2 \) is an unbiased estimate for \( \sigma^2 \) in the model (3.1), then

\[ r(\lambda) := \frac{1}{n} \|\hat{y}_\lambda - y\|_2^2 + \frac{2\hat{\sigma}^2}{n} \text{tr}(H_\lambda) \]

is unbiased for \( R_{in}(\hat{\beta}_\lambda) + \sigma^2 \) (note that \( \lambda \in \mathbb{R}_+^d \) is a vector). We consider using this unbiased estimate to choose a \( \lambda \in \mathbb{R}^d \) to obtain—we hope—better predictions than a naive ridge regression estimate would provide. We shall assume that we have a reasonable estimate \( \hat{\sigma}^2 \).
(c) Let \( U_\perp \in \mathbb{R}^{n \times (n-d)} \) be an orthogonal basis for \( \text{span}(U)^\perp = \{ v \in \mathbb{R}^n \mid U^T v = 0 \} \), i.e., any matrix so that \( I_n = UU^T + U_\perp U_\perp^T \) and \( U_\perp^T U = 0 \). Show that
\[
 r(\lambda) = \frac{1}{n} \sum_{j=1}^d \left[ \frac{\lambda_j^2}{(\gamma_j^2 + \lambda_j)^2} \left( u_j^T y \right)^2 + 2\delta^2 \frac{\gamma_j^2}{\gamma_j^2 + \lambda_j} \right] + \frac{1}{n} \| U_\perp^T y \|_2^2.
\]

Conclude that
\[
\frac{\partial}{\partial \lambda_j} r(\lambda) \cdot \frac{n}{2} = \frac{\lambda_j}{\gamma_j^2 + \lambda_j} \left[ \frac{\gamma_j^2}{(\gamma_j^2 + \lambda_j)^2} \right] (u_j^T y)^2 - \frac{\gamma_j^2}{\gamma_j^2 + \lambda_j} \delta^2
\]
\[
= \frac{\gamma_j^2}{(\gamma_j^2 + \lambda_j)^2} \left[ \frac{\lambda_j}{\gamma_j^2 + \lambda_j} (u_j^T y)^2 - \delta^2 \right].
\]

(d) We now consider minimizing \( r(\lambda) \) over \( \lambda \geq 0 \), i.e., elementwise \( \lambda_j \geq 0 \). Using the result of part (c), argue that \( \frac{\partial}{\partial \lambda_j} r(\lambda) < 0 \) for all \( \lambda \geq 0 \) whenever \( \delta^2 \geq (u_j^T y)^2 \). Then conclude that the \( \lambda^* \) minimizing \( r(\lambda) \) satisfies
\[
\lambda_j^* = \begin{cases} 
+\infty & \text{if } \delta^2 \geq (u_j^T y)^2 \\
\frac{\delta^2 \gamma_j^2}{(u_j^T y)^2 - \delta^2} & \text{otherwise.}
\end{cases} \tag{3.2}
\]

Interpret this in one or two sentences.

(e) Now you will perform some comparisons between your “tuned” principal components ridge regression with penalties \( \lambda^* \in \mathbb{R}_+^d \) as in (3.2) and a standard ridge estimator. Using the data in \texttt{lprostate.dat}, you will run ridge (or this optimized ridge) regression with \texttt{lpsa} as a response, performing the following experiment 25 times:

i. Load the data in \texttt{lprostate.dat}: split the data into a training set \((X_{\text{train}}, y_{\text{train}})\) containing a random .6 proportion of the data and test set \((X_{\text{test}}, y_{\text{test}})\) containing the remaining data.

ii. Standardize the data so that \( y_{\text{train}} \) is mean zero and so each column of data matrix \( X_{\text{train}} \) has mean zero and variance 1. Apply the same normalization to \( X_{\text{test}} \) and \( y_{\text{test}} \). (Note that you should use the transformation you apply to the training data.)

iii. Fit the “optimal” ridge estimator \( \hat{\beta}_{\lambda^*} \) using the \( \lambda^* \) in Eq. (3.2) and compute the hold-out risk \( \hat{\tau} = \frac{1}{n_{\text{test}}} \| X_{\text{test}} \hat{\beta}_{\lambda^*} - y_{\text{test}} \|_2^2 \), where \( n_{\text{test}} \) is the sample size of the test data. You should use the usual estimate \( \hat{\sigma}^2 = \frac{1}{n_{\text{train}} - d} \min_b \| X_{\text{train}} b - y_{\text{train}} \|_2^2 \).

iv. For each \( \tau = 10^{j/10}, i = -10, -9, \ldots, 20 \), fit a standard ridge regression estimate
\[
\hat{\beta}_\tau = (X_{\text{train}}^T X_{\text{train}} + \tau I)^{-1} X_{\text{train}}^T y_{\text{train}}
\]

and compute the hold-out risk \( \hat{\tau}_\tau = \frac{1}{n_{\text{test}}} \| X_{\text{test}} \hat{\beta}_{\tau} - y_{\text{test}} \|_2^2 \).

Then plot the average gap \( \hat{\tau}_\tau - \hat{\tau} \) over the 25 experiments as a function of \( \tau \), where your horizontal axis should correspond to \( \tau \) on a logarithmic scale. We have written code in the file \texttt{ridge-prostate-dataprep.*} that will perform steps i–ii for you.

Give one potential explanation for what you see in a sentence or two.