10/36-702 Statistical Machine Learning Homework #2
Solutions

DUE: February 23, 2018

Problem 1 [10 pts.]
Consider the data \((X_1, Y_1), \ldots, (X_n, Y_n)\) where \(X_i \in \mathbb{R}\) and \(Y_i \in \mathbb{R}\). Inspired by the fact that \(E[Y|X=x] = \int y p(x,y) dy / p(x)\), define

\[
\hat{m}(x) = \int \frac{y \hat{p}(x,y) dy}{\hat{p}(x)}
\]

where

\[
\hat{p}(x) = \frac{1}{n} \sum_i \frac{1}{h} K \left( \frac{X_i - x}{h} \right)
\]

and

\[
\hat{p}(x,y) = \frac{1}{n} \sum_i \frac{1}{h^2} K \left( \frac{X_i - x}{h} \right) K \left( \frac{Y_i - y}{h} \right).
\]

Assume that \(\int K(u) du = 1\) and \(\int u K(u) du = 0\). Show that \(\hat{m}(x)\) is exactly the kernel regression estimator that we defined in class.

Solution.

\[
\frac{\int y \cdot \hat{p}(x,y) dy}{\hat{p}(x)} = \frac{1}{nh^2} \int y \frac{1}{n} \sum \frac{1}{h} K \left( \frac{x-X_i}{h} \right) K \left( \frac{y-Y_i}{h} \right) dy
\]

\[
= \frac{\sum K \left( \frac{x-X_i}{h} \right) \int y \frac{1}{h} K \left( \frac{y-Y_i}{h} \right) dy}{\sum K \left( \frac{x-X_i}{h} \right)}
\]

\[
= \frac{\sum K \left( \frac{x-X_i}{h} \right) Y_i}{\sum K \left( \frac{x-X_i}{h} \right)}
\]

\[
= \hat{m}(x).
\]
Problem 2 [15 pts.]

Suppose that \((X, Y)\) is bivariate Normal:

\[
\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left( \begin{pmatrix} \mu \\ \eta \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho \sigma \tau \\ \rho \sigma \tau & \tau^2 \end{pmatrix} \right).
\]

(a) (5 pts.) Show that \(m(x) = \mathbb{E}[Y|X = x] = \alpha + \beta x\) and find explicit expressions for \(\alpha\) and \(\beta\).

(b) (5 pts.) Find the maximum likelihood estimator \(\hat{m}(x) = \hat{\alpha} + \hat{\beta} x\).

(c) (5 pts.) Show that \(|\hat{m}(x) - m(x)|^2 = O_P(n^{-1})\).

Solution.

(a) Some simple calculations show

\[ Y|X = x \sim N \left( \eta + \frac{\tau}{\sigma} \rho(x - \mu), \left(1 - \rho^2\right) \tau^2 \right), \]

which gives

\[ \alpha = \eta - \frac{\tau \rho \mu}{\sigma} \quad \text{and} \quad \beta = \frac{\tau \rho}{\sigma}. \]

(b) Given a sample \((X_1, Y_1), \ldots, (X_n, Y_n)\), the MLEs for the bivariate normal parameters are

\[
\begin{align*}
\hat{\mu} &= \bar{X} \\
\hat{\eta} &= \bar{Y} \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
\hat{\tau}^2 &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \\
\hat{\text{Cov}}(X, Y) &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).
\end{align*}
\]

Note \(\beta = \frac{\tau \rho}{\sigma} = \frac{\tau \rho \sigma}{\sigma^2}\). Then by the equivariance property of the MLE,

\[
\hat{\beta} = \frac{\hat{\text{Cov}}(X, Y)}{\hat{\sigma}^2}
\]

and

\[ \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}. \]

Again by equivariance,

\[ \hat{m}(x) = \hat{\alpha} + \hat{\beta} x. \]
(c) \( \hat{m}(x) \) is an MLE and satisfies the regularity conditions for asymptotic normality. Therefore,

\[
\sqrt{n}(\hat{m}(x) - m(x)) \sim N(0, \Gamma^{-1}(m(x))),
\]

which implies

\[
|\hat{m}(x) - m(x)|^2 = O_p(n^{-1}).
\]
Problem 3 [20 pts.]

Let \( m(x) = \mathbb{E}[Y \mid X = x] \). Let \( X \in [0,1]^d \). Divide \([0,1]^d\) into cubes \( B_1, \ldots, B_N \) whose sides have length \( h \). Given data \((X_1,Y_1), \ldots, (X_n,Y_n)\) define

\[
\hat{m}(x) = \begin{cases} 
\frac{1}{n(x)} \sum_i Y_i \mathbb{1}(X_i \in B(x)) & \text{if } n(x) > 0 \\
0 & \text{if } n(x) = 0
\end{cases}
\]

where \( B(x) \) is the cube containing \( x \) and \( n(x) = \sum_i \mathbb{1}(X_i \in B(x)) \). Assume that

\[
|m(y) - m(x)| \leq L\|x - y\|_2
\]

for all \( x, y \) and assume that \( X \) has a uniform density on \([0,1]^d\). You may assume that \( \sup_x \text{Var}(Y \mid X = x) < \infty \).

(a) (10 pts.) Show that

\[
|\mathbb{E}[\hat{m}(x)] - m(x)| \leq C_1 h
\]

for some \( C_1 > 0 \). Also show that

\[
\text{Var}(\hat{m}(x)) \leq \frac{C_2}{n(x)}
\]

for some \( C_2 > 0 \).

(b) (10 pts.) Let \( h \equiv h_n = (C \log n/n)^{1/d} \). Show that, for \( C > 0 \) large enough, \( P(\min n_j = 0) \to 0 \) as \( n \to \infty \) where \( n_j \) is the number of observations in cube \( B_j \).

Solution.

Note: This problem was modified on Piazza to simplify calculations. In particular, we said to take the \( X_i \)'s to be fixed. The random \( X \) bounds can be obtained by taken the results below and applying the law of total expectation and law of total variance, respectively.

(a)

\[
|\mathbb{E}[\hat{m}(x)] - m(x)| = \left| \mathbb{E} \left[ \frac{1}{n(x)} \sum_i Y_i \mathbb{1}(X_i \in B(x)) \right] - m(x) \right|
\]

\[
= \left| \frac{1}{n(x)} \sum_i \left( \mathbb{E}[Y_i] - m(x) \right) \mathbb{1}(X_i \in B(x)) \right|
\]

\[
= \left| \frac{1}{n(x)} \sum_i \left( m(X_i) - m(x) \right) \mathbb{1}(X_i \in B(x)) \right|
\]

\[
\leq \frac{1}{n(x)} \sum_i \left| m(X_i) - m(x) \right| \mathbb{1}(X_i \in B(x))
\]

\[
\leq \frac{1}{n(x)} \sum_i L \sqrt{d} h \cdot \mathbb{1}(X_i \in B(x))
\]

\[
= L \sqrt{d} h
\]

Let \( \sup_x \text{Var}(Y \mid X = x) = M \).
\[
\text{Var}(\hat{m}(x)) = \text{Var}\left( \frac{1}{n(x)} \sum_i Y_i \mathbb{1}_{\{X_i \in B(x)\}} \right) \\
= \frac{1}{n^2(x)} \sum_i \text{Var}(Y_i) \mathbb{1}_{\{X_i \in B(x)\}} \\
\leq \frac{M}{n(x)}.
\]

(b)

\[
P(\min_j n_j = 0) = P\left( \bigcup_{j=1}^B \{n_j = 0\} \right) \\
\leq \sum_{j=1}^B P(n_j = 0) \\
= \sum_{j=1}^B n \prod_{i=1}^n (1 - P(X_i \in B_j)) \\
= \frac{1}{h^d} (1 - h^d)^n \\
= \frac{n}{C \log n} \left( 1 - \frac{C \log n}{n} \right)^n
\]

Take \( C = 1 \). Then

\[
\frac{n}{C \log n} \left( 1 - \frac{C \log n}{n} \right)^n < \frac{n}{C \log n} e^{-\frac{C \log n}{n}} \\
= \frac{n}{C \log n} n^{-C} \\
= \frac{1}{C \log n} \\
\to 0.
\]
Problem 4 [15 pts.]
Consider the RKHS problem
\[
\tilde{f} = \arg \min_{f \in \mathcal{H}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{1}
\]
for some Mercer kernel function \( K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \), with eigenexpansion
\[
K(x, z) = \sum_{i=1}^{\infty} \gamma_i \phi_i(x) \phi_i(z).
\]
Here, \( \phi_1, \phi_2, \ldots \) are the orthonormal eigenfunctions of \( K \). In this problem, you will prove that the above problem is equivalent to the finite dimensional one
\[
\hat{a} = \arg \min_{a \in \mathbb{R}^n} \|y - K \alpha\|_{2}^2 + \lambda \alpha^T K \alpha, \tag{2}
\]
where \( K \in \mathbb{R}^{n \times n} \) denotes the kernel matrix \( K_{ij} = K(x_i, x_j) \). Recall that the inner product \( \langle \cdot, \cdot \rangle_{\mathcal{H}} \) on \( \mathcal{H} \), between functions \( f = \sum_{i=1}^{\infty} c_i \phi_i \) and \( g = \sum_{i=1}^{\infty} d_i \phi_i \) can be written as
\[
\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} c_i d_i / \gamma_i
\]
where \( \gamma_1, \gamma_2, \ldots \) are the eigenvalues of the kernel \( K \). Also recall that the functions \( K(\cdot, x_i), i = 1, \ldots, n \) are called the representers of evaluation. Recall that
\[
\begin{align*}
\bullet & \quad \langle f, K(\cdot, x_i) \rangle_{\mathcal{H}} = f(x_i), \text{ for any function } f \in \mathcal{H} \\
\bullet & \quad \|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \text{ for any function } f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i).
\end{align*}
\]

(a) (5 pts.) Let \( f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \), and consider defining a function \( \tilde{f} = f + \rho \), where \( \rho \) is any function orthogonal to \( K(\cdot, x_i), i = 1, \ldots, n \). Using the properties of the representers, prove that \( \tilde{f}(x_i) = f(x_i) \) for all \( i = 1, \ldots, n \), and \( \|\tilde{f}\|_{\mathcal{H}}^2 \geq \|f\|_{\mathcal{H}}^2 \).

(b) (10 pts.) Conclude from part (a) that in the infinite-dimensional problem (1), we need only consider functions of the form \( f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \), and that this in turn reduces to (2).

Solution.

(a) Since \( f, \tilde{f} \in \mathcal{H}_K \), for all \( i = 1, \ldots, n \)
\[
\tilde{f}(x_i) = \langle \tilde{f}, K(\cdot, x_i) \rangle_{\mathcal{H}_K}
= \langle f, K(\cdot, x_i) \rangle_{\mathcal{H}_K} + \langle \rho, K(\cdot, x_i) \rangle_{\mathcal{H}_K}
= \langle f, K(\cdot, x_i) \rangle_{\mathcal{H}_K}
= f(x_i).
\]
Also, because
\[
\langle \rho, f \rangle_{\mathcal{H}_K} = \left( \rho, \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \right)_{\mathcal{H}_K}
= \sum_{i=1}^{n} \alpha_i \langle \rho, K(\cdot, x_i) \rangle_{\mathcal{H}_K}
= 0,
\]
we have,

\[
\| \tilde{f} \|_{\mathcal{H}_K}^2 = \langle f, f \rangle_{\mathcal{H}_K} + \langle \rho, \rho \rangle_{\mathcal{H}_K} + 2 \langle \rho, f \rangle_{\mathcal{H}_K}
\]

\[
= \| f \|_{\mathcal{H}_K}^2 + \| \rho \|_{\mathcal{H}_K}^2
\]

\[
\geq \| f \|_{\mathcal{H}_K}^2.
\]

(b) For any \( \tilde{f} \in \mathcal{H}_K \), let \( \bar{y} = (\tilde{f}(x_1), \ldots, \tilde{f}(x_n))^T \in \mathbb{R}^n \). Let \( f \in \mathcal{H}_K \) be \( f = \sum_{i=1}^n \alpha_i K(\cdot, x_i) \), where \( \alpha = K^{-1} \bar{y} \). Then

\[
\langle \tilde{f} - f, K(\cdot, x_i) \rangle_{\mathcal{H}_K} = \langle \tilde{f}, K(\cdot, x_i) \rangle_{\mathcal{H}_K} - \sum_{j=1}^n \alpha_j \langle K(\cdot, x_j), K(\cdot, x_i) \rangle_{\mathcal{H}_K}
\]

\[
= \tilde{f}(x_i) - \sum_{j=1}^n \alpha_j K(x_i, x_j)
\]

\[
= \tilde{f}(x_i) - [K(K^{-1} \bar{y})]_i
\]

\[
= \tilde{f}(x_i) - f(x_i)
\]

\[
= 0.
\]

Hence, \( \tilde{f} - f \perp K(\cdot, x_i) \) for all \( i = 1, \ldots, n \), and from (a), this implies \( \tilde{f}(x_i) = f(x_i) \) for all \( i = 1, \ldots, n \), and \( \| \tilde{f} \|_{\mathcal{H}_K}^2 \geq \| f \|_{\mathcal{H}_K}^2 \), where equality holds if and only if \( \tilde{f} = f \). Therefore,

\[
\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \| f \|_{\mathcal{H}_K}^2 \leq \sum_{i=1}^n (y_i - \tilde{f}(x_i))^2 + \lambda \| \tilde{f} \|_{\mathcal{H}_K}^2,
\]

where equality holds if and only if \( \tilde{f} = f \). Hence if \( \tilde{f} = \text{argmin}_{f \in \mathcal{H}_K} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \| f \|_{\mathcal{H}_K}^2 \), then \( \tilde{f} = \sum_{i=1}^n \alpha_i K(\cdot, x_i) \) with \( \alpha = K^{-1} \bar{y} \). So we only need to consider functions of the form \( f = \sum_{i=1}^n \alpha_i K(\cdot, x_i) \). By plugging in, we have

\[
\sum_{i=1}^n (y_i - f(x_i))^2 \lambda \| f \|_{\mathcal{H}_K}^2 = \sum_{i=1}^n \left( y_i \sum_{j=1}^n \alpha_j K(x_i, x_j) \right)^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j)
\]

\[
= \| y - K\alpha \|_2^2 + \lambda \alpha^T K \alpha.
\]
Problem 5 [15 pts.]

Let \( X = (X(1), \ldots, X(d)) \in \mathbb{R}^d \) and \( Y \in \mathbb{R} \). In the questions below, make any reasonable assumptions that you need but state your assumptions.

(a) (5 pts.) Prove that \( \mathbb{E}(Y - m(X))^2 \) is minimized by choosing \( m(x) = \mathbb{E}(Y | X = x) \).

(b) (5 pts.) Find the function \( m(x) \) that minimizes \( \mathbb{E}|Y - m(X)| \). (You can assume that the conditional cdf \( F(y | X = x) \) is continuous and strictly increasing, for every \( x \).)

(c) (5 pts.) Prove that \( \mathbb{E}(Y - \beta^T X)^2 \) is minimized by choosing \( \beta_\ast = B^{-1} \alpha \) where \( B = \mathbb{E}(XX^T) \) and \( \alpha = (\alpha_1, \ldots, \alpha_d) \) and \( \alpha_j = \mathbb{E}(YX(j)) \).

Solution.

(a) Let \( g(x) \) be any function of \( x \). Then
\[
\mathbb{E}(Y - g(X))^2 = \mathbb{E}(Y - m(X) + m(X) - g(X))^2 \\
= \mathbb{E}(Y - m(X))^2 + \mathbb{E}(m(X) - g(X))^2 + 2\mathbb{E}((Y - m(X))(m(X) - g(X))) \\
\geq \mathbb{E}(Y - m(X))^2 + 2\mathbb{E}((Y - m(X))(m(X) - g(X))) \\
= \mathbb{E}(Y - m(X))^2 + 2\mathbb{E}((Y|X) - m(X))(m(X) - g(X)) \\
= \mathbb{E}(Y - m(X))^2 + 2\mathbb{E}(m(X) - m(X))(m(X) - g(X)) \\
= \mathbb{E}(Y - m(X))^2
\]

(b) Let \( g(x) \) be any function of \( x \). Recall that
\[
\mathbb{E}[|Y - g(X)|] = \mathbb{E}\{\mathbb{E}[|Y - g(X)|] | X = x}\}
\]

The idea is to choose \( c \) such that \( \mathbb{E}[|Y - c | | X = x] \) is minimized. Now define:
\[
r(c) = \mathbb{E}[|Y - c | | X = x] = \int |y - c| p_{Y|X=x}(y) dy.
\]

The function \( h_y(c) = |y - c| \) is differentiable everywhere except when \( y = c \). Thus for \( c \neq y \)
\[
h'_y(c) = \begin{cases} 
1 & c > y \\
-1 & c < y \\
1(c > y) - 1(c < y). 
\end{cases}
\]

Since \( Y \) is continuous and has a density function, \( P(Y = c) = 0 \). So to minimize \( r(c) \) we can differentiate under the integral sign and set the derivative equal to 0 to obtain:
\[
r'(c) = \int h'_y(c)p_{Y|X=x}(y) dy = \int_{-\infty}^c p_{Y|X=x}(y) dy - \int_c^\infty p_{Y|X=x}(y) dy \\
= 2 \int_{-\infty}^c p_{Y|X=x}(y) dy - 1 = 0 \\
\iff \int_{-\infty}^c p_{Y|X=x}(y) dy = \frac{1}{2},
\]
so that \( c = m(x) \), which is the median of \( p_{Y|X=x}(y) \). It is a minimum since \( r'(c) < 0 \) for \( c < m(x) \) and \( r'(c) > 0 \) for \( c > m(x) \). Since \( m \) minimizes \( \mathbb{E}[|Y - c| \mid X = x] \) at every \( x \) for any \( g \) we get

\[
\mathbb{E}[|Y - g(X)| - |Y - m(X)||X = x] \geq 0
\]

which implies

\[
R(g) - R(m) = \mathbb{E}[|Y - g(X)| - |Y - m(X)|] = \mathbb{E}\{\mathbb{E}[|Y - g(X)| - Y - m(X)||X]\} \geq 0.
\]

(c) By setting the first derivative of the loss function equal to 0 we obtain:

\[
\frac{\partial R(\beta)}{\partial \beta} = 0
\]

\[
\Rightarrow \frac{\partial \mathbb{E}(Y - \beta^T X)^2}{\partial \beta} = 0
\]

\[
\Rightarrow \mathbb{E}[-2X(Y - \beta^T X)] = 0
\]

\[
\Rightarrow 2B\beta - 2\alpha = 0
\]

\[
\Rightarrow \beta_* = B^{-1}\alpha,
\]

where we can exchange the derivative and expectation by the dominated convergence theorem. The loss function \( R(\beta) \) is strictly convex so \( \beta_* \) is its unique minimum.
Problem 6 [25 pts.]

Suppose that \( Y_i = \beta^T X_i + \epsilon_i \) for \( i = 1, \ldots, n \). Let \( X \) be the \( n \times d \) design matrix. Assume that \( X^T X = I \) (the identity matrix). Let \( \hat{\beta} \) minimize

\[
\frac{1}{n} \sum_i (Y_i - \beta^T X_i)^2 + \lambda P(\beta).
\]

Find an explicit form for \( \hat{\beta} \) for three cases: (i) (10 pts.) \( P(\beta) = \|\beta\|_0 \), (ii) (10 pts.) \( P(\beta) = \|\beta\|_1 \) and (iii) (5 pts.) \( P(\beta) = \|\beta\|_2^2 \).

Solution.

(i) Note that

\[
\frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_0 = \frac{1}{2} y^T y - \beta^T X^T y + \frac{1}{2} \beta^T X^T X \beta + \lambda \|\beta\|_0
\]

\[
= \frac{1}{2} y^T X X^T y - \sum_{j=1}^{p} \beta_j (X^T y)_j + \frac{1}{2} \beta^T \beta + \lambda \sum_{j=1}^{p} \mathbb{1}(\beta_j \neq 0) + \frac{1}{2} y^T (I - XX^T) y
\]

\[
= \sum_{j=1}^{p} \left( \frac{1}{2} \beta^2_j - \beta_j (X^T y)_j^2 + \lambda \mathbb{1}(\beta_j \neq 0) \right) + \frac{1}{2} y^T (I - XX^T) y
\]

\[
= \sum_{j=1}^{p} \left( \frac{1}{2} (\beta_j - X_j^T y)^2 + \lambda \mathbb{1}(\beta_j \neq 0) \right) + \frac{1}{2} y^T (I - XX^T) y.
\]

Then

\[
\frac{1}{2} (\beta_j - X_j^T y)^2 + \lambda \mathbb{1}(\beta_j \neq 0) \geq \frac{1}{2} (X_j^T y)^2 \mathbb{1}(\beta_j = 0) + \lambda \mathbb{1}(\beta_j \neq 0)
\]

\[
\geq \min \left\{ \frac{1}{2} (X_j^T y)^2, \lambda \right\}
\]

and equality holds if and only if

\[
\beta_j = \begin{cases} 
0 & \text{if } \frac{1}{2} (X_j^T y)^2 < \lambda \\
0 \text{ or } X_j^T y & \text{if } \frac{1}{2} (X_j^T y)^2 = \lambda \\
X_j^T y & \text{if } \frac{1}{2} (X_j^T y)^2 > \lambda.
\end{cases}
\]

Hence

\[
\|y - X\beta\|_2^2 + \lambda \|\beta\|_0 = \sum_{j=1}^{p} \left( (\beta_j - X_j^T y)^2 + \lambda \mathbb{1}(\beta_j \neq 0) \right) + y^T (I - XX^T) y
\]

\[
\geq \sum_{j=1}^{p} \min \left\{ (X_j^T y)^2, \lambda \right\} + y^T (I - XX^T) y
\]

and equality holds if and only if

\[
\beta_j = \begin{cases} 
0 & \text{if } |X_j^T y| < \sqrt{\lambda} \\
0 \text{ or } X_j^T y & \text{if } |X_j^T y| = \sqrt{\lambda} \\
X_j^T y & \text{if } |X_j^T y| > \sqrt{\lambda}.
\end{cases}
\]
(ii) First write
\[
\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\| = \min_{\beta} (y - X\beta)^T (y - X\beta) + \lambda \sum_j |\beta_j| \\
= \min_{\beta} -2y^T X\beta + \beta^T X^T X\beta + \lambda \sum_j |\beta_j| \\
= \min_{\beta} -2 \sum_j (X^T y) \beta_j + \sum_j \beta_j^2 + \lambda \sum_j |\beta_j|.
\]

Now note the last line is simply equivalent to
\[
\min_{\beta_j} -2(X^T y) \beta_j + \beta_j^2 + \lambda |\beta_j| \\
\iff \min_{\beta_j} -2\hat{\beta}_j^{OLS} \beta_j + \beta_j^2 + \lambda |\beta_j|
\]
for all \(j = 1, \ldots, p\).

When \(\hat{\beta}_j^{OLS} \geq 0\), then \(\bar{\beta}_j \geq 0\) so
\[
-2\hat{\beta}_j^{OLS} \beta_j + \beta_j^2 + \lambda |\beta_j| = -2\hat{\beta}_j^{OLS} \beta_j + \beta_j^2 + \lambda \beta_j.
\]

Differentiating with respect to \(\beta_j\), setting equal to zero, and solving gives
\[
\bar{\beta}_j = (\hat{\beta}_j^{OLS} - \frac{\lambda}{2}) 1_{(\hat{\beta}_j^{OLS} \geq \frac{\lambda}{2})}.
\]

When \(\hat{\beta}_j^{OLS} \leq 0\), the analogous steps give
\[
\bar{\beta}_j = (\hat{\beta}_j^{OLS} + \frac{\lambda}{2}) 1_{(\hat{\beta}_j^{OLS} \leq -\frac{\lambda}{2})}.
\]

Putting them together gives
\[
\bar{\beta}_j = \begin{cases} 
\hat{\beta}_j^{OLS} - \frac{\lambda}{2} & \hat{\beta}_j^{OLS} \geq \frac{\lambda}{2} \\
0 & \hat{\beta}_j^{OLS} \in \left(-\frac{\lambda}{2}, \frac{\lambda}{2}\right) \\
\hat{\beta}_j^{OLS} + \frac{\lambda}{2} & \hat{\beta}_j^{OLS} \leq -\frac{\lambda}{2}.
\end{cases}
\]
(iii) Here the objective function is differentiable everywhere. Taking the gradient w.r.t. $\beta$ we have

$$\nabla_{\beta} \left( \frac{1}{n} \| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2 \right) = -\frac{2}{n} X^T(y - X\beta) + 2\lambda \beta.$$ 

Setting this equal to 0 and solving for $\beta$ gives

$$\hat{\beta} = (I + n\lambda I)^{-1} X^T y.$$ (3)

Since the objective is strictly convex, (3) is the unique solution.