Home Assignment 4 Solutions

STAT 151A Linear Modelling: Theory and Applications

Ajay Sharma — Spring 2025

Problem 1: Alternative Derivation of Ridge Regression
Recall the ridge regression problem for regularization parameter $\lambda > 0$, with

$$\hat{\beta}_\lambda := \underset{\beta \in \mathbb{R}^d}{\arg\min} \ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where $(X, Y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$.

(a) *Closed Form for $\hat{\beta}_\lambda$ & Computability*
Define the function $\mathcal{L}(\beta) := \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$. Computing the gradient gives

$$\begin{aligned}\nabla_\beta \mathcal{L}(\beta) &= \nabla_\beta \left\{(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta\right\} \\ &= \nabla_\beta \left\{-2(X^TY)^T\beta + \beta^TX^TX\beta + \lambda\beta^T\beta\right\} \\ &= -2X^TY + 2X^TX\beta + 2\lambda\beta.\end{aligned}$$

To find the stationary point, simply set $\nabla_\beta \mathcal{L}(\beta) = 0 \Rightarrow \hat{\beta}_\lambda = (X^TX + \lambda I_d)^{-1}X^TY$. We should note that since $\lambda > 0$, $M := X^TX + \lambda I_d$ is always positive-definite and in particular the regularization term λI_d ensures invertibility (simply a shifting of eigen-values). Thus $\hat{\beta}_\lambda$ is computable.

(b) *Augmented Matrices*
Let us consider the augmented matrix defined by

$$\tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda} I_d \end{pmatrix} \in \mathbb{R}^{(n+d)\times d}.$$

Our goal is to find $\tilde{Y} \in \mathbb{R}^{n\times d}$ such that

$$\|\tilde{Y} - \tilde{X}\beta\|_2^2 = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \quad \text{for any } \beta \in \mathbb{R}^d.$$

Write

$$\left\|\tilde{Y} - \tilde{X}\beta\right\|_2^2 = \left\|\begin{pmatrix} Y \\ z \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda} I_d \end{pmatrix}\beta\right\|_2^2 = \left\|\begin{pmatrix} Y - X\beta \\ z - \sqrt{\lambda}\beta \end{pmatrix}\right\|_2^2.$$

Then equating this with $\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$ gives

$$\|Y - X\beta\|_2^2 + \|z - \sqrt{\lambda}\beta\|_2^2 = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

And so we obtain $\|z - \sqrt{\lambda}\beta\|_2^2 = \|z\|_2^2 - 2\sqrt{\lambda}z^T\beta + \lambda\|\beta\|_2^2 = \lambda\|\beta\|_2^2 \Rightarrow z = 0$. Therefore we have that $\tilde{Y} = (Y, 0)^T$.

(c) *Uniqueness of Solutions*
Suppose $\tilde{\beta}$ is the OLS solution from augmented data $(\tilde{X}, \tilde{Y})$. Our goal is to show that $\tilde{\beta} = \hat{\beta}_\lambda$.

Note that if $\tilde{\beta}$ satisfies the ridge regression problem, then we may write

$$\tilde{\beta} = (\tilde{X}^T\tilde{X} + \lambda I_d)^{-1}\tilde{X}^T\tilde{Y} = \hat{\beta}_\lambda = (X^TX + \lambda I_d)^{-1}X^TY.$$

Therefore we must have that $\tilde{X} = X$ and $\tilde{Y} = Y$ and so the solution is unique.

(d) *Two Equivalent Forms*
Having established in part (c) that the solution to the ridge regression problem is unique, we want to show that the two forms of the solutions

$$\hat{\beta}_\lambda = \hat{\beta}^*_\lambda := X^T(XX^T + \lambda I_n)^{-1}Y.$$

To accomplish this task, we'll use the Woodbury Identity Formula

$$(A + UCV)^{-1} = A^{-1} + A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

In our case, begin with the LHS and simplify the expression $\hat{\beta}_\lambda$ to the RHS $\hat{\beta}^*_\lambda$. Matching terms in the inverse function gives

- $A = \lambda I_d \in \mathbb{R}^{d\times d}$
- $U = X^T \in \mathbb{R}^{d\times n}$
- $C = I_n \in \mathbb{R}^{n\times n}$
- $V = X \in \mathbb{R}^{n\times d}$

Then we have

$$\begin{aligned}
\hat{\beta}_\lambda &= (X^TX + \lambda I_d)^{-1}X^TY \\
&= \left[(\lambda I_d)^{-1} + (\lambda I_d)^{-1}X^T\left(I_n^{-1} + X(\lambda I_d)^{-1}X^T\right)^{-1}X(\lambda I_d)^{-1}\right]X^TY \\
&= \left[\lambda^{-1}I_d - \lambda^{-2}X^T\left(\lambda^{-1}(\lambda I_n + XX^T)\right)^{-1}X\right]X^TY \\
&= \lambda^{-1}X^TY - \lambda^{-1}X^T(XX^T + \lambda I_n)^{-1}XX^TY \\
&= X^T\left[\lambda^{-1}I_n - \lambda^{-1}(XX^T + \lambda I_n)^{-1}XX^T\right]Y \\
&= X^T(XX^T + \lambda I_n)^{-1}Y \\
&= \hat{\beta}^*_\lambda
\end{aligned}$$

Indeed these solutions to the ridge problem are equivalent. When $n \ll d$, there are more features than observations and so in the expression $M := X^TX + \lambda I_d \in \mathbb{R}^{d\times d}$ is very large and thus inverting the matrix becomes computationally expensive. On the other hand, the expression $M^* := XX^T + \lambda I_n \in \mathbb{R}^{n\times n}$ becomes small and is computationally more efficient and cheaper.

Problem 2: Revisit Min-Norm Solution

For data $(X, Y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ assume $n < d$ and that $\text{rank}(X) = n$. We have seen in the lecture (before the midterm) that in the over-parametrized regime, the linear regression has infinitely many solutions, and thus we focus on special solution called the min-norm solution

$$\hat{\beta}_{\min} := \underset{\beta\in\mathbb{R}^d: X\beta=Y}{\arg\min} \|\beta\|_2^2.$$

(a) *Closed Form Solution $\hat{\beta}_{min}$*
Define the Lagrangian $\mathcal{L}(\beta, \lambda) := \frac{1}{2}\|\beta\|_2^2 - \lambda^T(Y - X\beta)$. Then we can find solve the system of equations by taking the gradient $\nabla := (\frac{\partial}{\partial\beta}, \frac{\partial}{\partial\lambda})$. In particular

- $\nabla_\beta\mathcal{L}(\beta, \lambda) = \beta - X^T\lambda = 0 \Rightarrow \beta = X^T\lambda$
- $\nabla_\lambda\mathcal{L}(\beta, \lambda) = Y - X\beta \Rightarrow Y = X\beta$

Then substituting the first into the second gives $Y = X(X^T\lambda) \Rightarrow \lambda = (XX^T)^{-1}Y$. Therefore $\hat{\beta}_{\min} = X^T(XX^T)^{-1}Y$.

(b) *Convergence of Ridge Estimator*
From exercise 1, we know the ridge solution is given by $\hat{\beta}_\lambda := X^T(XX^T + \lambda I_n)^{-1}Y$. We want to show $\hat{\beta}_\lambda \to \hat{\beta}_{\min}$ as $\lambda \to 0$. To show this convergence, we'll use the operator norm definition with

$$\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2 \Rightarrow \|Ax\|_2 \le \|A\|_2 \cdot \|x\|_2.$$

In particular, define $M = XX^T, A_\lambda = (M + \lambda I_n)^{-1}, A = M^{-1}$. We'll show that

- $\|A_\lambda - A\|_2 \to 0$ and
- $\|A_\lambda Y - AY\|_2 \le \|A_\lambda - A\|_2 \cdot \|Y\|_2$.

As for the first point, notice that $M \ge 0$ and since X has full (row) rank, then $\lim_{\lambda\to 0} A_\lambda = A$, which also holds under the operator norm. We then have

$$\|A_\lambda - A\|_2 = \sup_{\|x\|_2=1} \|(A_\lambda - A)x\| \to 0$$

by continuity of inverse functions in operator norm for PSD matrices. As for the second point, we have we have that $\|A_\lambda Y - AY\| \le \|A_\lambda - A\| \cdot \|Y\| \to 0$ using the previous result. Therefore, we have shown $\lim_{\lambda\to 0} \hat{\beta}_\lambda = \hat{\beta_{\min}}$.

(c) *Application: Gradient Descent*
Write the general gradient descent update rule as

$$\beta^{(t+1)} = \beta^{(t)} - \eta\nabla J(\beta^{(t)}), \beta^{(0)} = 0, \quad \text{for } t \in [n], \beta^{(t)} \in \mathbb{R}^d.$$

We want to show $\beta^{(t)} \in \text{Col}(X^T)$. Define the loss function $J(\beta^{(t)}) = \frac{1}{2}\|Y - X\beta^{(t)}\|_2^2$. Then computing the gradient gives

$$\begin{aligned}
\nabla_{\beta^{(t)}} J(\beta^{(t)}) &= \nabla_{\beta^{(t)}} \left\{ \frac{1}{2}(Y - X\beta^{(t)})^T(Y - X\beta^{(t)}) \right\} \\
&= \nabla_{\beta^{(t)}} \left\{ \frac{1}{2}\left(Y^TY - 2{\beta^{(t)}}^T X^TY + {\beta^{(t)}}^T X^TX\beta^{(t)} \right) \right\} \\
&= X^T(X\beta^{(t)} - Y).
\end{aligned}$$

In other words, we can write $\beta^{(t+1)} = \beta^{(t)} - \eta X^T(X\beta^{(t)} - Y)$. To proceed, we'll use the principal of induction. In particular

- Base Case: $\beta^{(0)} = 0 \in \text{Col}(X^T)$ since 0 is a member of every subspace.
- Inductive Hypothesis: Suppose $\beta^{(t)} \in \text{Col}(X^T)$ holds for each $t \in [n]$.
- Inductive Step: In the expression $\beta^{(t+1)} = \beta^{(t)} - \eta X^T(X\beta^{(t)} - Y)$, we already know by our inductive hypothesis that $\beta^{(t)} \in \text{Col}(X^T)$. Furthermore, we know that $M := X\beta^{(t)} - Y \in \mathbb{R}^n$. Therefore $X^T M \in \text{Col}(X^T)$ as it is simply a linear map and the result trivially follows. So we conclude $\beta^{(t+1)} \in \text{Col}(X^T)$ as claimed.

(d) *Additional Properties of Solution*
We want to show that $\{\beta \in \text{Col}(X^T) : X\beta = Y\} = \{\hat{\beta}_{\min}\}$. To do this, we'll first show uniqueness. In particular, suppose there exists $\beta_1, \beta_2 \in \text{Col}(X^T)$. Note also that $\beta_1 - \beta_2 \in \text{Col}(X^T)$.

So $X\beta_1 = Y = X\beta_2$. Then we can write $X(\beta_1 - \beta_2) = 0 \Rightarrow \beta_1 - \beta_2 \in \text{Null}(X)$. But since $\text{Null}(X) = \text{Col}(X^T)^\perp, \beta_1 - \beta_2 \in \text{Col}(X^T) \cup \text{Col}(X^T)^\perp$. By definition of orthogonal complement, we have that $\beta_1 - \beta_2 = 0 \Rightarrow \beta_1 = \beta_2$, which proves uniqueness. Having established this and using previous facts, we know that $\hat{\beta}_{\min} = X^T(XX^T)^{-1}Y \in \text{Col}(X^T)$ with $X\hat{\beta}_{\min} = Y$, so $\{\beta \in \text{Col}(X^T) : X\beta = Y\} = \{\hat{\beta}_{\min}\}$ as claimed.

Problem 3: LASSO and Soft-Thresholding in Orthonormal X
Given data $(X, Y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ such that $X^TX = XX^T = I_n$. Recall the lasso estimator for regularization parameter $\lambda > 0$ is given by

$$\hat{\beta}_\lambda := \underset{\beta\in\mathbb{R}^d}{\arg\min}\ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

(a) *OLS Solution*
We know the general solution $\hat{\beta}_{\text{OLS}} = (X^TX)^{-1}X^TY = X^TY$, since X is orthogonal.

(b) *Equivalent Formulation of LASSO Problem*
We want to show

$$\hat{\beta}_\lambda := \underset{\beta\in\mathbb{R}^d}{\arg\min}\ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 = \underset{\beta\in\mathbb{R}^d}{\arg\min}\ \|\hat{\beta}_{\text{OLS}} - \beta\|_2^2 + \lambda\|\beta\|_1.$$

To do this we'll simplify the LHS and write it as an equivalent expression. In particular, we know

$$\begin{aligned}
\|Y - X\beta\|_2^2 &= \langle Y - X\beta,\ Y - X\beta\rangle \\
&= \left\langle (Y - X\hat{\beta}_{\text{OLS}}) + X(\hat{\beta}_{\text{OLS}} - \beta),\ (Y - X\hat{\beta}_{\text{OLS}}) + X(\hat{\beta}_{\text{OLS}} - \beta)\right\rangle \\
&= \|Y - X\hat{\beta}_{\text{OLS}}\|_2^2 + 2\left\langle Y - X\hat{\beta}_{\text{OLS}},\ X(\hat{\beta}_{\text{OLS}} - \beta)\right\rangle + \|X(\hat{\beta}_{\text{OLS}} - \beta)\|_2^2 \\
&= \|Y - X\hat{\beta}_{\text{OLS}}\|_2^2 + \|X(\hat{\beta}_{\text{OLS}} - \beta)\|_2^2 \\
&= \text{const} + (\hat{\beta}_{\text{OLS}} - \beta)^TX^TX(\hat{\beta}_{\text{OLS}} - \beta) \\
&= \text{const} + \|\hat{\beta}_{\text{OLS}} - \beta\|_2^2.
\end{aligned}$$

Note that in the fourth equality, define the residual vector $\hat{e} := Y - X\hat{\beta}_{\text{OLS}}$. Thus the expression $\langle \hat{e}, X(\hat{\beta}_{\text{OLS}} - \beta)\rangle = 0$ simplifies since $\hat{e} \perp \text{Col}(X)$. Hence we obtain

$$\underset{\beta\in\mathbb{R}^d}{\arg\min}\ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 = \underset{\beta\in\mathbb{R}^d}{\arg\min}\ \|\hat{\beta}_{\text{OLS}} - \beta\|_2^2 + \lambda\|\beta\|_1,$$

as claimed.

(c) *Coordinate-wise Minimization*
Define the objective function

$$\mathcal{L}(\beta) = \|\hat{\beta}_{\text{OLS}} - \beta\|_2^2 + \lambda\|\beta\|_1 = \sum_{j=1}^{d}\{(\hat{\beta}_{\text{OLS}} - \beta)^2 + \lambda|\beta_j|\}.$$

Now consider the coordinate-wise objective function

$$\mathcal{L}_j(\beta) = \begin{cases} (\hat{\beta}^j_{\text{OLS}} - \beta_j)^2 + \lambda\beta_j, & j > 0 \\ (\hat{\beta}^j_{\text{OLS}} - \beta_j)^2 - \lambda\beta_j, & j < 0 \\ (\hat{\beta}^j_{\text{OLS}} - \beta_j)^2, & j = 0. \end{cases}$$

Then

$$\frac{\partial(\mathcal{L}_j(\beta_j))}{\partial\beta} = \begin{cases} -2(\hat{\beta}^j_{\text{OLS}} - \beta_j) + \lambda, & j > 0 \\ -2(\hat{\beta}^j_{\text{OLS}} - \beta_j) - \lambda, & j < 0 \\ -2(\hat{\beta}^j_{\text{OLS}} - \beta_j), & j = 0. \end{cases}$$

To find the stationary points set $\frac{\partial(\mathcal{L}_j(\beta_j))}{\partial\beta} = 0$. Then we obtain the critical points (minimum) as follows

$$\hat{\beta}^j_\lambda = \begin{cases} \hat{\beta}^j_{\text{OLS}} - \lambda/2, & \hat{\beta}^j_{\text{OLS}} > \lambda/2 \\ \hat{\beta}^j_{\text{OLS}} + \lambda/2, & \hat{\beta}^j_{\text{OLS}} < -\lambda/2 \\ 0, & |\hat{\beta}^j_{\text{OLS}}| \leq \lambda/2 \end{cases},$$

which we can write more succinctly as $\hat{\beta}^j_\lambda = \text{sign}(\hat{\beta}^j_{\text{OLS}}) \cdot \max(|\hat{\beta}^j_{\text{OLS}}| - \lambda/2, 0)$.

(d) *LASSO Computational Exercise*
We wish to compute the LASSO solution when $\lambda = 3, \hat{\beta}_{\text{OLS}} = (4, 3, -2, 1)^T$. Making use of our expression in part (c), we'll perform coordinate-wise computations. So

- $\hat{\beta}^1_\lambda = \text{sign}(4) \cdot \max(|4| - 3/2, 0) = 5/2.$
- $\hat{\beta}^2_\lambda = \text{sign}(3) \cdot \max(|3| - 3/2, 0) = 3/2.$
- $\hat{\beta}^3_\lambda = \text{sign}(-2) \cdot \max(|-2| - 3/2, 0) = -1/2.$
- $\hat{\beta}^4_\lambda = \text{sign}(1) \cdot \max(|1| - 3/2, 0) = 0.$

Therefore we obtain $\hat{\beta}^\lambda_{\text{LASSO}} = (5/2, 3/2, -1/2, 0)^T$.

Problem 4: Elastic Net

The elastic net is a regularized regression method that linearly combines the ridge and lasso penalty. In particular, define

$$\hat{\beta}_{\text{EN}} := \underset{\beta \in \mathbb{R}^d}{\arg\min} \ \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \quad \lambda_1, \lambda_2 \geq 0.$$

Suppose we have the data $(X, Y) \in \mathbb{R}^{n \times 2} \times \mathbb{R}^n$. Let $x_1, x_2 \in \mathbb{R}^n$ be the columns of X. Write $\hat{\beta}_{\text{EN}} = (\hat{\beta}^1_{\text{EN}}, \hat{\beta}^2_{\text{EN}})^T$.

(a) *Showing LASSO Solution is Not Unique*

Consider the case when $x_1 = x_2 = x \in \mathbb{R}^n$ and define the loss function

$$\mathcal{L}_1(\beta) = \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

Intuitively, the answer is no. The solutions are not unique since X becomes rank deficient (columns of X are linearly dependent) and essentially reduces to finding $\min\{|\beta_1| + |\beta_2| : (\beta_1, \beta_2)^T \in \mathbb{R}^2, \beta_1 + \beta_2 = \hat{\beta}^*_{\text{LASSO}}\}$, which admits infinitely many solutions. Our goal is to show more rigorously that $(\hat{\beta}_1, \hat{\beta}_2)$ are distinct solutions, both satisfying $\underset{\beta \in \mathbb{R}^d}{\arg\min}\ \mathcal{L}_1(\beta)$.

Under these assumptions, we can write $X = (x, x) = x \cdot \mathbf{1}$, for $\mathbf{1}^T = (1, 1) \in \mathbb{R}^2$. Furthermore, define $\beta = (\beta_1, \beta_2)^T \in \mathbb{R}^2$. Then we can simply write the objective (loss) function as

$$\mathcal{L}_1(\beta_\alpha) = \|Y - \alpha x\|_2^2 + \lambda |\alpha|,$$

since

$$Y - X\beta = Y - (\beta_1 x + \beta_2 x) = Y - (\beta_1 + \beta_2)x = Y - \alpha x,$$

for some constant $\alpha = \beta \cdot \mathbf{1}^T = \beta_1 + \beta_2 \in \mathbb{R}$. Then using the formula we derived in exercise 3, we can write the solution to this LASSO problem as

$$\hat{\beta}^*_{\text{LASSO}} = \underset{\beta \in \mathbb{R}^2}{\arg\min}\ \mathcal{L}_1(\beta_\alpha) = \text{sign}(v) \cdot \max\left(|v| - \frac{\lambda}{2\|x\|_2^2},\ 0\right), \quad \text{where} \quad v = \frac{x^T Y}{\|x\|_2^2}.$$

Having found this solution, we can construct infinitely many pairs $\beta \in \mathbb{R}^2$ satisfying $\beta_1 + \beta_2 = \hat{\beta}^*_{\text{LASSO}}$. For instance, choose $\beta_1 = (\hat{\beta}^*_{\text{LASSO}}, 0)^T, \beta_2 = (0, \hat{\beta}^*_{\text{LASSO}})^T$. Then observe $X\beta_{j \in [2]} = \hat{\beta}^*_{\text{LASSO}} x$, with both $\|Y - X\beta_{j \in [2]}\|_2^2 = \|Y - \hat{\beta}^*_{\text{LASSO}} x\|_2^2$ and $\|\beta_{j \in [2]}\|_1 = |\hat{\beta}^*_{\text{LASSO}}|$. In other words, we have found two distinct solutions to this minimization lasso problem showing the solution is not unique.

(b) *Showing Ridge Solution is Unique*

Consider the case $x_1 = x_2 = x \in \mathbb{R}^n$ and define the loss function

$$\mathcal{L}_2(\beta) = \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

Then under these assumptions, write $X = (x, x) = x \cdot \mathbf{1}$, for $\mathbf{1}^T = (1, 1) \in \mathbb{R}^2$. Furthermore, define $\beta = (\beta_1, \beta_2)^T \in \mathbb{R}^2$. Therefore simplify the objective (loss) function as $\mathcal{L}_2(\beta_\alpha) = \|Y - \alpha x\|_2^2 + \lambda(\beta_1^2 + \beta_2^2), \quad \alpha = \beta \cdot \mathbf{1}^T = \beta_1 + \beta_2 \in \mathbb{R}$.

Note that since $\|Y - \alpha x\|_2^2$ is fixed, the problem reduces to finding $\min_{\alpha \in \mathbb{R}} \lambda(\beta_1^2 + \beta_2^2)$. The solution $\hat{\beta}_{j \in [2]} = \alpha/2$ is unique (and obvious) to this minimization problem.

(c) *Properties of Elastic Net I*
We have demonstrated separately that both penalties $l_{1,2}$ have different types of solutions. In particular, the l_1 penalty has infinitely many solutions to the norm-constraint minimization problem (partly the function is not strictly-convex) but under the l_2 penalty, we do have a strictly-convex function and therefore has a unique minimizer. We know for sure that a strictly convex function is guaranteed to have a minimum and so the result (elastic net) hinges on this fact. Let's use the definition of convexity to demonstrate.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and $g : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex function. Define $h(x) = f(x) + g(x)$. We aim to show that h is strictly convex. By definition, we can write

- f is convex if for all $x, y \in \mathbb{R}^d$, and $\theta \in [0, 1]$,

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

- g is strictly convex if for all $x \neq y \in \mathbb{R}^d$, and $\theta \in (0, 1)$,

$$g(\theta x + (1-\theta)y) < \theta g(x) + (1-\theta)g(y).$$

Adding these two inequalities, we obtain

$$f(\theta x + (1-\theta)y) + g(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y) + \theta g(x) + (1-\theta)g(y).$$

Therefore

$$h(\theta x + (1-\theta)y) < \theta h(x) + (1-\theta)h(y),$$

is strictly convex as claimed (guaranteeing a unique minimizer). Furthermore as the l_2 penalty dominates (dictates the type of solution), the elastic net prefers $\hat{\beta}^1_{\text{EN}} = \hat{\beta}^2_{\text{EN}}$ due to the symmetry of the solution, as derived in part (b) which has some kind of equal splitting (and properties of strictly convex functions mentioned previously).

(d) *Properties of Elastic Net II*
Suppose now $x_1 \approx x_2$ (they are not equal but very similar). In other words, they are highly correlated. The elastic net produce better prediction or more stable models than lasso in this case since it tends to assign similar weights to both features instead of arbitrarily selecting one, as lasso does. In fact we have shown in part (a) that the lasso solution has infinitely many solutions (therefore not strictly convex) and encourages sparsity (certain coefficients to be 0) so it picks one of the weights, say x_1 at random at sets the other weight to 0. On the other hand, the elastic net leads to more stable models and better predictive performance in the presence of collinearity due to construction and symmetry (distributing the weights among predictors), as explored in part (b) and (c).