LASSO for Linear Models

Notation (i) Observations $y_1, y_2, y_3, \ldots, y_N \in \mathbb{R}$

(ii) predictors $x_1, x_2, x_3, \ldots, x_N \in \mathbb{R}^p$

(we can think of x_i as the i^{th} feature vector) row

(iii) regression weight

$$\beta = (\beta_1, \beta_2, \ldots, \beta_p) \in \mathbb{R}^p$$

The model: (linear model)

$$y_i \cong \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

Why take out β_0? Treat it as the offset. β_0 is not influenced by any feature.

Linear Least Squares

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right\} \qquad (1)$$

How to solve (1)?

rewrite as

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}}_{\beta}$$

If $\hat{\beta} = (X^T X)^{-1} X^T y$.

Limitation of least squares

(1) Prediction accuracy.
Usually low bias but high variance

* (2) Not always interpretable.
Not all predictors should be activated.

(3) The matrix X^TX is not always invertible
e.g. when X is rank deficient or has poor condition number.

→ how to make the inversion more stable?

* Regularization: (Regularized least square)

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \|\beta\|^2 \right\} \quad \text{—(2)}$$

This is equivalent to

$$\min_{\beta} \quad \|y - X\beta\|^2 + \lambda\|\beta\|^2$$

$$\Rightarrow \min_{\beta} \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \beta \right\|^2$$

$$\Rightarrow \quad \hat{\beta} = \left(X^TX + \lambda I\right)^{-1} X^T y .$$

However, interpretability remains an issue.

Remark: The regularized least squares is also called the ridge regression, taking the form

$$\min_{\beta_0,\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{——— (3)}$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \le t^2$$

Note that (2) and (3) are equivalent for appropriate choice of (λ, t) pair.

LASSO Estimator

The LASSO Estimator is defined as

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \|\beta\|_1 \right\} \quad \text{---- (4)}$$

The equivalent form of LASSO is

$$\min_{\beta_0, \beta_1} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right\} \quad \text{---------- (5)}$$

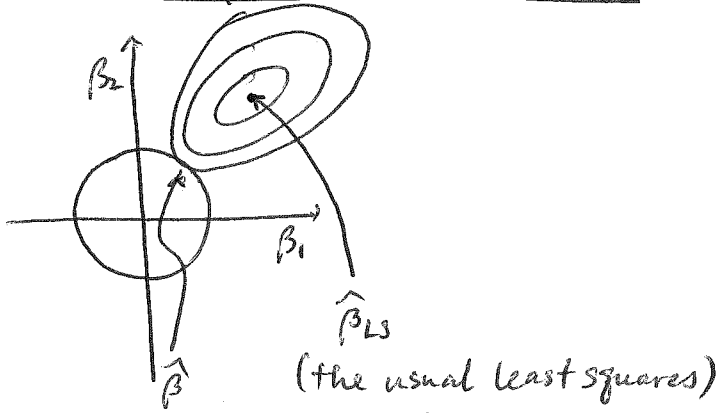
$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \le t.$$

Interpretation: t is the budget we put on the coefficients. Ideally, if we want to control the number of non-zeros, the we put

$$\sum_{j=1}^{p} \mathbb{I}\{\beta_j \neq 0\} \le t.$$

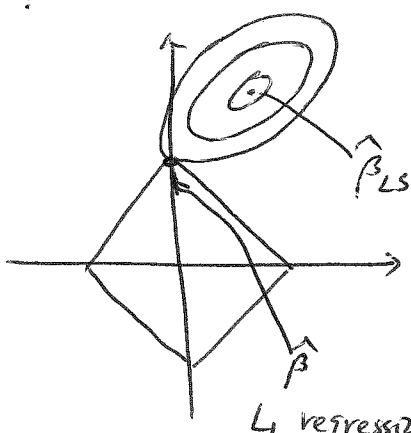
This constraint is equivalent to $\|\beta\|_0$. However $\|\beta\|_0$ is computationally not easy to solve. $\|\beta\|_1$, on the other hand, is convex.

Why can $\|\beta\|_1$ enhance sparsity?



(the usual least squares)

ridge regression solution



L1 regression solution (LASSO)

algorithms for finding its solutions. More details are given in Exercises (2.3) and (2.4).

As an example of the lasso, let us consider the data given in Table 2.1, taken from Thomas (1990). The outcome is the total overall reported crime rate per

Table 2.1 *Crime data: Crime rate and five predictors, for $N = 50$ U.S. cities.*

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

one million residents in 50 U.S cities. There are five predictors: annual police funding in dollars per resident, percent of people 25 years and older with four years of high school, percent of 16- to 19-year olds not in high school and not high school graduates, percent of 18- to 24-year olds in college, and percent of people 25 years and older with at least four years of college. This small example is for illustration only, but helps to demonstrate the nature of the lasso solutions. Typically the lasso is most useful for much larger problems, including "wide" data for which $p \gg N$.

The left panel of Figure 2.1 shows the result of applying the lasso with the bound t varying from zero on the left, all the way to a large value on the right, where it has no effect. The horizontal axis has been scaled so that the maximal bound, corresponding to the least-squares estimates $\tilde{\beta}$, is one. We see that for much of the range of the bound, many of the estimates are exactly zero and hence the corresponding predictor(s) would be excluded from the model. Why does the lasso have this model selection property? It is due to the geometry that underlies the ℓ_1 constraint $\|\beta\|_1 \leq t$. To understand this better, the right panel shows the estimates from *ridge regression*, a technique that predates the lasso. It solves a criterion very similar to (2.3):

$$\begin{aligned} &\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \\ &\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t^2. \end{aligned} \tag{2.7}$$

The ridge profiles in the right panel have roughly the same shape as the lasso profiles, but are not equal to zero except at the left end. Figure 2.2 contrasts the two constraints used in the lasso and ridge regression. The residual sum

Remark:

The full name of LASSO is
(Least Absolute Shrinkage and Selection Operator)
by R. Tibshirani in 1996.

In the signal processing community, the same minimization is called the Basis Pursuit Denoise (BPDN).

Standardization

Without loss of generality we shall assume that the columns of X are normalized:

(1) $\frac{1}{N}\sum_{i=1}^{N} x_{ij} = 0$, i.e. column average $= 0$ ↓

(de-mean process)

(2) $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$, i.e. column variance $= 1$

(3) $\frac{1}{N}\sum_{i=1}^{N} y_i = 0$.

The Trajectory of LASSO

How does Ridge Regression (2) compared to LASSO (4)?

Crime data (Table 2.1 of HTW)

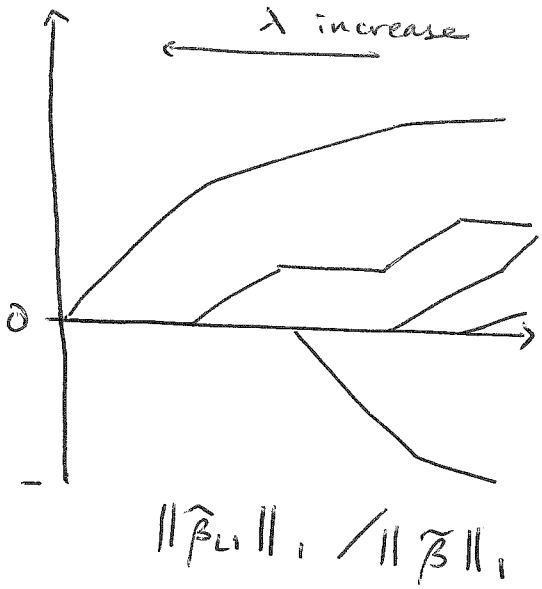
City	police funding	high school	college	...	crime rate
1					
2					
⋮					
50					

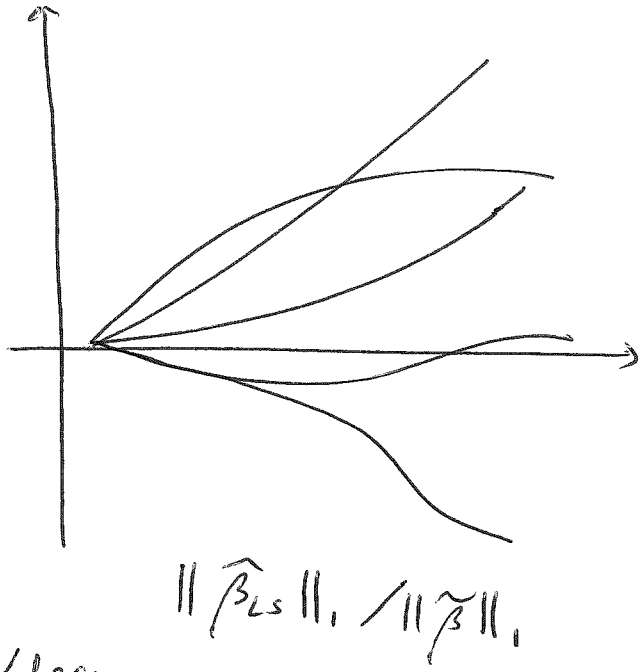
$\hat{\beta}_{LS}$: solution of ridge regression (2) (= the least squares replaced)

$\hat{\beta}_{L1}$: solution of LASSO (4)

$\tilde{\beta}$: solution of least square (1)

Note: ~~All these three~~ Both $\hat{\beta}_{LS}$ and $\hat{\beta}_{L1}$ solutions depend the parameter λ.



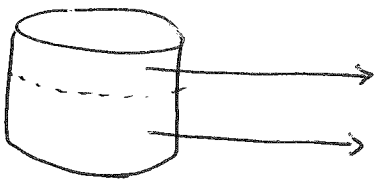


Observation: As λ increases/decreases, the # of non-zeros in ~~$\|\hat{\beta}\|_1$~~ $\hat{\beta}_{L1}$ changes.

large λ: more emphasis on $\lambda\|\beta\|_1$, so more sparse solution.

Cross-Validation

How to choose λ?



We have ground-truth labels for the testing. So we can calculate the mean square error

Idea: For every λ, compute the solution

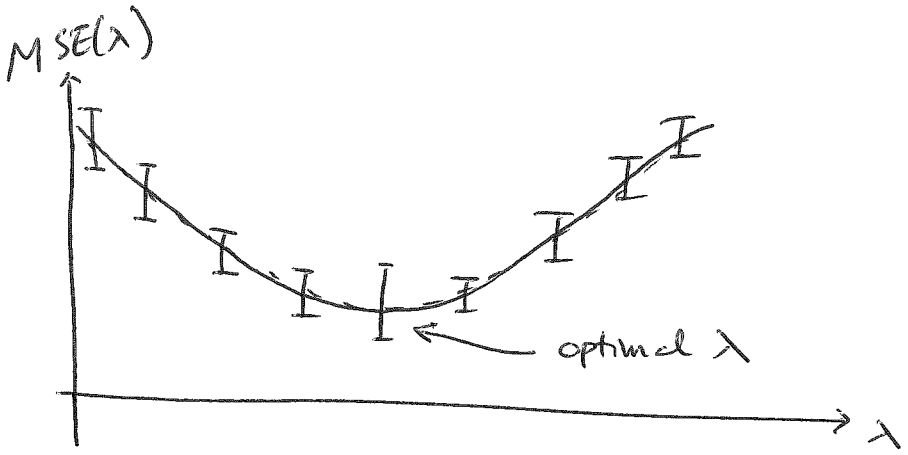
($y_i \in$ training (A))

$$\hat{\beta}_{L1}(\lambda) = \underset{\beta_0, \beta}{\operatorname{argmin}} \left\{ \frac{1}{2N_A} \sum_{i=1}^{N_A} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda\|\beta\|_1 \right\}$$

Then, calculate the MSE.

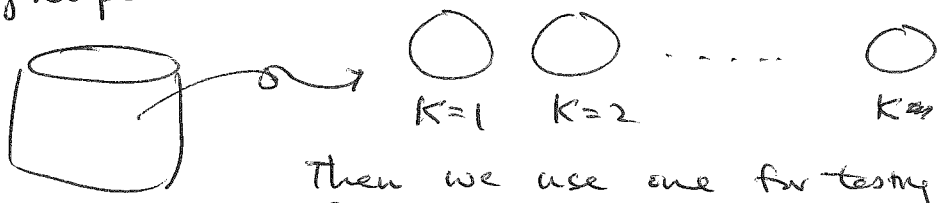
($y_i \in$ testing (B))

$$MSE(\lambda) = \frac{1}{N_B} \sum_{i=1}^{N_B} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j \right)^2$$



In principle, the MSE will reach a minimum somewhere. The corresponding λ is the optimal λ.

The standard error bars are computed by ~~switch~~ switching the training and testing data. Typically, for a data set containing N points we divide them into K groups



Then we use one for testing and $K-1$ for training. Repeat for K times by switching the testing group.

Solving the LASSO Equation

Consider a simpler problem: $\beta \in \mathbb{R}$, $x_{ij} = x_i$, i.e. no j.

$$\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - z_i \beta)^2 + \lambda |\beta| \right\} \qquad \text{——— (6)}$$

This is a 1D optimization in β.

Assumptions: (standardization)

(1) $\frac{1}{N} \sum_{i=1}^{N} z_i = 0$

(2) $\frac{1}{N} \sum_{i=1}^{N} z_i^2 = 1$

(3) $\frac{1}{N} \sum_{i=1}^{N} y_i = 0$

technically speaking $\frac{\partial}{\partial \beta}$ is undefined for $\beta = 0$.

$$\varphi(\beta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - z_i\beta)^2 + \lambda|\beta| \qquad \text{sign}(\beta) = \begin{cases} 1 & , \beta > 0 \\ \in [-1,1] & , \beta = 0 \\ -1 & , \beta < 0 \end{cases}$$

$$\frac{\partial}{\partial \beta}\varphi \overset{"="}{=} \frac{1}{N}\sum_{i=1}^{N}(y_i - z_i\beta)(-z_i) + \lambda\,\text{sign}(\beta) = 0$$

$$\Longrightarrow \quad \frac{1}{N}\sum_{i=1}^{N}(-z_i y_i) + \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} z_i^2\right)}_{1}\beta + \lambda\,\text{sign}(\beta) = 0$$

$$\Longrightarrow \quad \beta + \lambda\,\text{sign}(\beta) - \frac{1}{N}\langle \underline{z}, \underline{y}\rangle = 0 \qquad \text{——— (7)}$$

Case 1 : $\beta > 0$.

if $\beta > 0$, then (7) becomes

$$\beta + \lambda - \frac{1}{N}\langle \underline{z}, \underline{y}\rangle = 0$$

and so $\beta = \frac{1}{N}\langle \underline{z}, \underline{y}\rangle - \lambda > 0$ (it has to be > 0 because $\beta > 0$.)

So $\frac{1}{N}\langle \underline{z}, \underline{y}\rangle > \lambda$. ——— (8a)

Case 2 : $\beta < 0$,

if $\beta < 0$, then (7) becomes

$$\beta - \lambda - \frac{1}{N}\langle \underline{z}, \underline{y}\rangle = 0$$

this implies

$$\beta = \frac{1}{N}\langle \underline{z}, \underline{y}\rangle + \lambda < 0$$

and so $\frac{1}{N}\langle \underline{z}, \underline{y}\rangle < -\lambda$ ——— (8b)

Case 3 : $\beta = 0$. This happens when (8a) and (8b) do not happen, which is

$$\left|\frac{1}{N}\langle \underline{z}, \underline{y}\rangle\right| \leq \lambda.$$

Therefore, we have the solution:

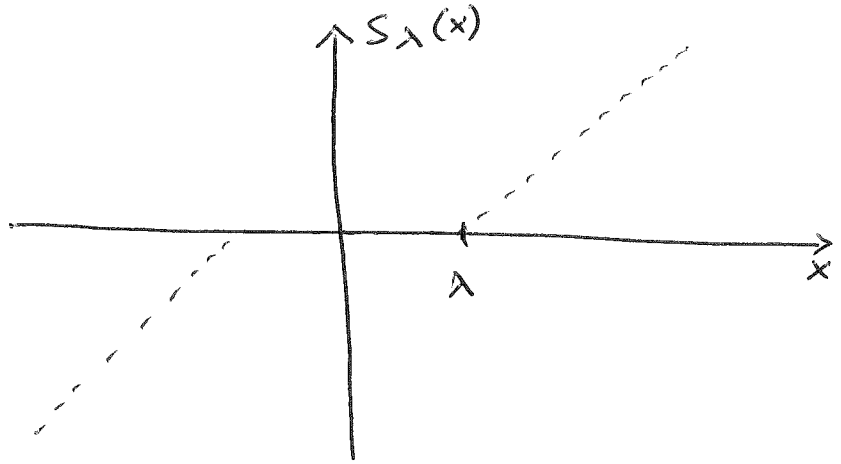
$$\hat{\beta} = \begin{cases} \frac{1}{N}\langle \underline{z}, \underline{y} \rangle - \lambda & , \text{ if } \frac{1}{N}\langle \underline{z}, \underline{y} \rangle > \lambda \\ 0 & , \text{ if } \left|\frac{1}{N}\langle \underline{z}, \underline{y} \rangle\right| \leq \lambda \\ \frac{1}{N}\langle \underline{z}, \underline{y} \rangle + \lambda & , \text{ if } \frac{1}{N}\langle \underline{z}, \underline{y} \rangle < -\lambda \end{cases}.$$

or we can write them compactly as

$$\hat{\beta} = \max\left(\left| \frac{1}{N}\langle \underline{z}, \underline{y} \rangle \right| - \lambda, 0 \right) \operatorname{sign}\left(\frac{1}{N}\langle \underline{z}, \underline{y} \rangle \right)$$

$$\stackrel{\text{def}}{=} S_\lambda\left(\frac{1}{N}\langle \underline{z}, \underline{y} \rangle \right).$$

↑ shrinkage function



Now, we can consider the general case:

$$\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

$$= \min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(\boxed{y_i - \sum_{k \neq j} x_{ik} \beta_k} - x_{ij} \beta_j \right)^2 + \lambda \sum_{k \neq j} |\beta_k| + \lambda |\beta_j| \right\}$$

($\boxed{\cdot}$ = $r_i^{(j)}$)

Why write in this form?
— to decouple the optimization into many 1D problems

$$= \min_{\beta_j} \left\{ \frac{1}{2N} \sum_{i=1}^{N} \left(r_i^{(j)} - x_{ij} \beta_j \right)^2 + \lambda |\beta_j| \right\}$$ ← repeat for $j = 1, 2, \ldots, p$

$$= S_\lambda\left(\frac{1}{N}\langle \underline{x}_j, \underline{r}^{(j)}\rangle\right), \qquad \underline{r}^{(j)} = \begin{bmatrix} r_1^{(j)} \\ \vdots \\ r_N^{(j)} \end{bmatrix} \qquad \underline{x}_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{Nj} \end{bmatrix}$$

So we have an algorithm:

$$\hat{\beta}_j \longleftarrow S_\lambda\left(\hat{\beta}_j + \frac{1}{N}\langle \underline{x}_j, \underline{r}\rangle\right)$$

$$\underline{r} = \begin{bmatrix} r_1 \\ \vdots \\ r_N \end{bmatrix}, \quad r_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$$

This algorithm is cyclic coordinate descent.

The algorithm works here because the optimization is convex

If $\underline{x}_j$ and $\underline{x}_k$ are orthogonal (e.g. Fourier/wavelet basis), then $\frac{1}{N}\langle \underline{x}_j, \underline{x}_k\rangle = 0$. In this case,

$$\frac{1}{N}\langle \underline{x}_j, \underline{r}^{(j)}\rangle = \frac{1}{N}\langle \underline{x}_j, \underline{y} - \sum_{k\neq j} \underline{x}_k \beta_k\rangle$$

$$= \frac{1}{N}\langle \underline{x}_j, \underline{y}\rangle - \frac{1}{N}\sum_{k\neq j}\beta_k\langle \underline{x}_j, \underline{x}_k\rangle$$

$$= \frac{1}{N}\langle \underline{x}_j, \underline{y}\rangle.$$

Then

$$\hat{\beta}_j = S_\lambda\left(\frac{1}{N}\langle \underline{x}_j, \underline{y}\rangle\right).$$

Uniqueness of LASSO solution

Let $\hat{\beta}$ and $\hat{\gamma}$ be two LASSO solutions of a common λ. Assume that they have the same optimal value c^*.

Then (1) $X\hat{\beta} = X\hat{\gamma}$ ← $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$

(2) $\|\hat{\beta}\|_1 = \|\hat{\gamma}\|_1$.

$\begin{bmatrix} | \\ \vdots \\ | \end{bmatrix}$

proof: if $X\hat{\beta} \neq X\hat{\gamma}$, then let $0 < \theta < 1$, and

(1) $$\hat{\alpha} = \theta\hat{\beta} + (1-\theta)\hat{\gamma}.$$

Then

$$X\hat{\alpha} = \theta X\hat{\beta} + (1-\theta)X\hat{\gamma}.$$

By convexity of $\|\cdot\|^2$ and $\|\cdot\|_1$, we have

$$\|y - X\hat{\alpha}\|^2 + \lambda\|\hat{\alpha}\|_1$$

$$\leq \theta\left[\|y - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|_1\right] + (1-\theta)\left[\|y - X\hat{\gamma}\|^2 + \lambda\|\hat{\gamma}\|_1\right]$$

$$= \theta c^* + (1-\theta)c^* = c^*. \quad \text{So } X\hat{\beta} = X\hat{\gamma}.$$

(2) if $X\hat{\beta} = X\hat{\gamma}$, then

$$\|y - X\hat{\beta}\|^2 = \|y - X\hat{\gamma}\|^2$$

Since both $\hat{\beta}$ and $\hat{\gamma}$ give c^*, it must be that

$$\|\hat{\beta}\|_1 = \|\hat{\gamma}\|_1.$$

Implication of these two results:

(i) Solutions are "unique" in the sense that they generate the same optimal value and have the same level of sparsity.

(ii) The solutions themselves do not need to be identical. Note that LASSO is convex but not <u>strictly convex</u>. So there could be multiple solutions giving the same global minimum.