

## Topic 11: Lasso And Beyond: Convex Learning

by Sai Zhang

Key points:

Disclaimer:

### 11.1 Lasso

Lasso (Least absolute Shrinkage and Selection Operator), proposed by Tibshirani (1996), aims to minimize the **SSR (sum of residual squares)** subject to the **L1-norm (sum of the absolute value)** of the coefficients being less than a constant.

#### 11.1.1 Set up

For data  $(\mathbf{x}_i, y_i)_{i=1}^n$ , where

- $y_i$  is the outcome for individual  $i$
- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  is the  $p \times 1$  vector of predictors

Then the Lasso estimator  $(\hat{\alpha}, \hat{\beta})$  is defined as

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta} \left\{ \sum_{i=1}^n \left( y_i - \alpha - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} \quad \text{s.t.} \quad \sum_{j=1}^p |\beta_j| \leq t$$

for the  $n \times 1$  response vector  $\mathbf{y} = (y_1, \dots, y_n)'$ , the  $n \times p$  design matrix  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$  where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  is a  $p \times 1$  vector. Here  $\hat{\alpha} = \bar{y}$ , w.l.o.g., let  $\bar{y} = 0$  and omit  $\alpha$  for simplicity.

In matrix form, we have

- constrained form:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \right\} \quad \text{s.t.} \quad \|\beta\|_1 \leq t$$

- unconstrained form:

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

where the regularization parameter  $\lambda \geq 0$ :

- $\lambda \rightarrow \infty$ :  $\hat{\beta}_{lasso} = \mathbf{0}$
- $\lambda = 0$ :  $\hat{\beta}_{lasso} \rightarrow \hat{\beta}_{OLS}$

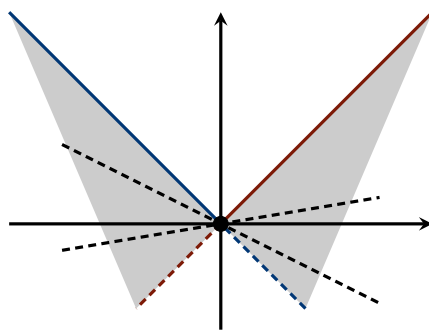
### 11.1.2 Solving Lasso

Lasso is essentially a quadratic optimization problem. Hence, the solution is given by taking the derivative (of the unconstrained question) and set it equal to 0

$$\frac{d}{d\beta} \left( \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right) = 0$$

$\xrightarrow{\text{KKT condition}}$ 
 $\frac{1}{n} \underbrace{X'}_{p \times n} \underbrace{(y - X\beta)}_{= \epsilon, n \times 1} = \lambda \begin{cases} \text{sign}(\beta_j), & \beta_j \neq 0 \\ [-1, 1], & \beta_j = 0 \end{cases}$

this result follows the fact the L-1 norm  $\|\beta\|$  is piecewise linear (convex)<sup>1</sup>:



L1-norm (1-dimension)

For each component of the vector of the L-1 norm

$f(\beta_j) = |\beta_j|$ , we have:

-  $\beta_j > 0$ :  $f'(\beta_j) = 1$

-  $\beta_j < 0$ :  $f'(\beta_j) = -1$

-  $\beta_j = 0$ :  $df \in [-1, 1]$  (shaded area)  
which gives the results stated above.

Take another look at this result

#### Proposition 11.1.1: Lasso Parameter Selection Rule

$$\frac{1}{n} X' (y - X\beta) = \frac{1}{n} X' \epsilon = \lambda \begin{cases} \text{sign}(\beta_j), & \beta_j \neq 0 \\ [-1, 1], & \beta_j = 0 \end{cases}$$

which gives a parameter selection criterion: for  $\beta_j \neq 0$ ,  $\text{sign}(\beta_j)$  **must agree** with  $\text{Corr}(x_j, \epsilon)$ , the correlation between the  $j$ -th variable  $x_j$  and (full-model) residuals  $\epsilon = y - X\beta$ .

### 11.1.3 Algorithm: LARS

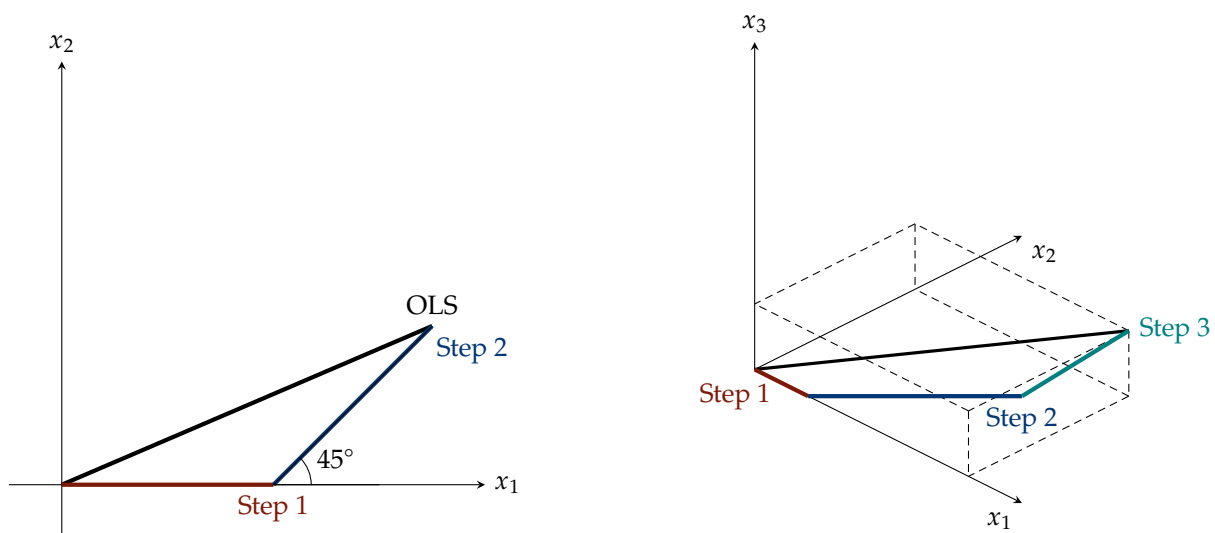
Mathematically, Lasso is quite intuitive, but computationally, it can be quite consuming. Efron et al. (2004) propose an algorithm that takes steps from a all-0 model to the biggest model (OLS), that is, **Least Angle Regression (LARS)**.

#### Intuition

The basic intuition of LARS is quite straight-forward: covariates are considered from the **highest** correlation with  $y$  (*smallest* angle from  $y$ ) to the **least** correlated one (*largest* angle from  $y$ ) (illustrated below).

And the steps of the LARS algorithm are

<sup>1</sup>KKT condition gives the analytical optimization rule for **convex** function.



- 1 start with the null model  $\hat{\beta} = \mathbf{0}$ :  $\hat{\mu} = \mathbf{X}'\mathbf{0} = \mathbf{0}$
- 2 calculate residual vector  $\mathbf{r} = \mathbf{y} - \hat{\mu}$
- 3 determine the correlation vector between  $\mathbf{r}$  and each parameter  $\mathbf{x}_j, \forall j = 1, \dots, p$ :  $\mathbf{X}'\mathbf{r}$
- 4 pick the largest correlation  $\mathbf{x}_{\text{step1},1}^*$ , increase its  $\hat{\beta}$  to the point where its correlation with  $\mathbf{r}$  will be **equal** with that of another parameter  $\mathbf{x}_{\text{step1},2}^*$
- 5 next, increase the  $\hat{\beta}$  for both  $\mathbf{x}_{\text{step1},1}^*, \mathbf{x}_{\text{step1},2}^*$  in an **equiangular** direction between these two, until a third parameter becomes equally important

And keep looping this way, until all the predictors enter the model and eventually  $\mathbf{X}'\mathbf{r} = \mathbf{0}$

### Properties of LARS

LARS has several properties:

- geometrically travels in the direction of **equal** angle to all active covariates
- assume all covariates are independent
- computationally quick: only take  $m$  steps, where  $m$  is the number of parameters being considered

And it is in between 2 classic model-selection methods: **Forward Selection** and **Stagewise Selection**:

- **Forward Selection**

- for  $\mathbf{y}$ , select the most correlated  $\mathbf{x}_{j_1}$
- regress  $\mathbf{x}_{j_1}$  on  $\mathbf{y}$ , get the residuals
- select the most correlated  $\mathbf{x}_{j_2}$  with the residual of  $\mathbf{y}$  net of  $\mathbf{x}_{j_1}$

looping this, for a  $k$ -parameter linear model, it takes  $k$  steps. Forward Selection is an aggressive fitting technique, can be overly greedy (some important predictors may be eliminated due to correlation with already selected variables).

- **Forward Stagewise**

- also begin with  $\hat{\mu} = \mathbf{0}$
- for a current Stagewise estimate  $\hat{\mu}$ , the current residual vector is then  $\mathbf{y} - \hat{\mu}$ , its correlation with  $\mathbf{X}$  is then  $\mathbf{X}'(\mathbf{y} - \hat{\mu}) \equiv \hat{\mathbf{c}}$

- next, heavily computational, go in the direction of the greatest current correlation, but by only a **small** step

$$\hat{j} = \arg \max |\hat{c}_j|, \hat{\mu} \rightarrow \hat{\mu} + \epsilon \cdot \text{sign}(\hat{c}_{\hat{j}}) \cdot \mathbf{x}_{\hat{j}}$$

here,  $\epsilon$  is a **small** constant, hence avoiding the greediness of Forward Selection, at a cost of computational efficiency<sup>2</sup>.

LARS avoids the over-greediness of Forward Selection and computational heaviness of Forward Stagewise.

### 11.1.4 From LARS to Lasso

The Lasso algorithm is built upon LARS, with the constraint from the mathematical condition of Proposition 11.1.1:  $\text{sign}(\beta_j)$  **must agree** with  $\text{Corr}(\mathbf{x}_j, \epsilon)$ .

#### Theorem 11.1.2: Lasso Modification Condition

If  $\tilde{\gamma} < \hat{\gamma}$ , stop the ongoing LARS step at  $\gamma = \tilde{\gamma}$  and remove  $j$  from the calculation of the next equiangular direction, where

- the path at any LARS step is

$$\beta(\gamma), \beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j$$

$\hat{d}_j$  specifies the **direction** to take the  $j$ -th component,  $\gamma$  is **how far** to travel in the direction of  $\hat{d}_j$  before adding in a new covariate

- $\hat{\gamma}$  represents the smallest **positive** value of  $\gamma$  s.t. some new covariate joins the active set (the set of covariates used on path)
- $\tilde{\gamma}$  represents the first time  $\beta_j(\gamma)$  **changes signs**.

The key point of 11.1.2 is that Lasso does **NOT** allow the  $\hat{\beta}_j$  to change signs, if it changes sign, it will be subtracted from the active set. Now, from this point of view, we can compare the 3 algorithms:

|           |  |
|-----------|--|
| LARS      | no sign restrictions   |
| Lasso     | $\hat{\beta}_j$ agrees in sign with $\hat{c}_j$  |
| Stagewise | successive differences of $\hat{\beta}_j$ agree in sign with the current correlation $\hat{c} = \mathbf{x}'_j(\mathbf{y} - \hat{\mu})$ |

Again, LARS requires the least steps but is most greedy, Stagewise is computationally consuming but robust. Lasso is in between.

## 11.2 Consistency of Lasso

Next, we want to establish the consistency of Lasso, by showing that Lasso selects exactly the relevant covariates asymptotically. We do this in 2 steps:

- show that Lasso at least captures all the relevant covariates
- asymptotically, under some conditions, Lasso selects exactly all the relevant covariates, not more

<sup>2</sup>Forward Selection is essentially choosing  $\epsilon = |\hat{c}_{\hat{j}}|$

### 11.2.1 Overestimation

First, Lasso tends to select a superset of the relevant covariates.

Define the true relevant set Lasso selection estimation  $\hat{S}_0$  aim to select as

$$S_0 = \{j : \beta_j^0 \neq 0, j = 1, \dots, p\}$$

and for some  $C > 0$ , define the relevant set w.r.t.  $C$  as

$$S_0^{\text{relevant}(C)} = \{j : |\beta_j^0| \geq C, j = 1, \dots, p\}$$

then we have

#### Theorem 11.2.1: Lasso Overestimation Condition

$\forall 0 < C < \infty$

$$\mathbb{P} \left[ \hat{S}_0(\lambda) \supset S_0^{\text{relevant}(C)} \right] \xrightarrow{n \rightarrow \infty} 1$$

### Consistency

The consistency of Lasso is established by [Meinshausen and Bühlmann \(2006\)](#) as

#### Theorem 11.2.2: Consistency of Lasso

For a suitable  $\lambda = \lambda_n \gg \sqrt{s_0 \log(p)/n}$ , Lasso is consistent, i.e.

$$\mathbb{P} \left[ \hat{S}(\lambda) = S_0 \right] \xrightarrow{n \rightarrow \infty} 1$$

if and only if it satisfies the 2 properties:

- $\beta$ -min condition (unselected coefficients non-trivial):  $\inf_{j \in S_0^c} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}$
- **irrepresentable condition**:  $\mathbf{X}$  should **NOT** exhibit too strong a degree of linear dependence w.r.t. the selected covariates

**discussion on the irrepresentable condition** denote  $\hat{\Sigma} = n^{-1} \mathbf{X} \mathbf{X}'$ , and let the active set  $S_0 = \{j : \beta_j^0 \neq 0\} = \{1, \dots, s_0\}$  consists of the first  $s_0$  variables, let

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{1,1} & \hat{\Sigma}_{1,2} \\ \hat{\Sigma}_{2,1} & \hat{\Sigma}_{2,2} \end{pmatrix}$$

where  $\hat{\Sigma}_{1,1}$  is a  $s_0 \times s_0$  var-cov matrix of the active variables  $\mathbf{X}_1$ ,  $\hat{\Sigma}_{2,2}$  is a  $(p - s_0) \times (p - s_0)$  cov-var matrix of the other variables  $\mathbf{X}_2$ <sup>3</sup>, then we for a Lasso estimation  $\hat{\beta}$  that assign non-zero coefficients **only** to  $\mathbf{X}_1$ , we can, following the Lasso result in proposition 11.1.1, have

$$\begin{cases} \frac{1}{n} \mathbf{X}_1' (\mathbf{y} - \mathbf{X} \hat{\beta}) &= \lambda \text{sign}(\hat{\beta}_1) \\ \left\| \frac{1}{n} \mathbf{X}_2' (\mathbf{y} - \mathbf{X} \hat{\beta}) \right\|_{\infty} &\leq \lambda \end{cases}$$

<sup>3</sup>Here,  $\mathbf{X}_1$  is  $n \times s_0$ ,  $\mathbf{X}_2$  is  $n \times p - s_0$

$\|\cdot\|$  is the component-wise max-norm

### 11.2.2 Oracle

Next, we want show Lasso has the oracle procedure, which gives the consistency.

#### Definition 11.2.3: Oracle Property

For a fitting procedure  $\delta$ , and the estimation  $\hat{\beta}(\delta)$ , then if  $\delta$  is an oracle procedure if  $\hat{\beta}(\delta)$  asymptotically has the following properties

- **consistency** (identifying right subset model):  $\{j : \hat{\beta}_j \neq 0\} = S_0$
- **optimal estimation rate** (asymptotically normal):  $\sqrt{n} \left( \beta(\delta)_{S_0} - \beta_{S_0}^0 \right) \xrightarrow{d} \mathcal{N}(0, \Sigma_0)$ , where  $\Sigma_0$  is the true subset covariance matrix

## 11.3 Variants of Lasso

### 11.3.1 Other Variants

There are also some other useful variants of Lasso

- **Positive Lasso**: Constrains the  $\hat{\beta}_j$  to enter the prediction equation in their **defined** directions, non-negative here

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \right\} \quad \text{s.t. } \|\beta\|_1 \leq t \text{ and } \beta_j > 0, \forall j$$

- **LARS-OLS hybrid**: Use the covariates selected by LARS, but use  $\hat{\beta}$  from the OLS model
- **Main effects first**:
  - Step 1: run LARS for a model, considering **only** main effects
  - Step 2: run LARS again, with the chosen main effects, and **all possible interactions** between them
- **Backward Lasso**: start from the **full** OLS model, and eliminate covariates **backwards** (by the order of correlation going 0 the earliest)

## 11.4 Penalized Least Square Estimation

Lasso is one special class of Penalized Least Square (PLS) Estimation. For the linear regression model  $\mathbf{y} = \mathbf{X}\beta + \epsilon$ , if  $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ , we have PLS as

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\}$$

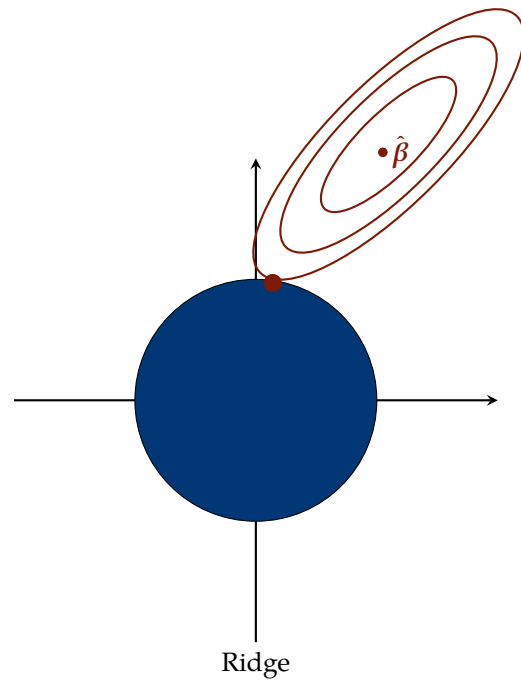
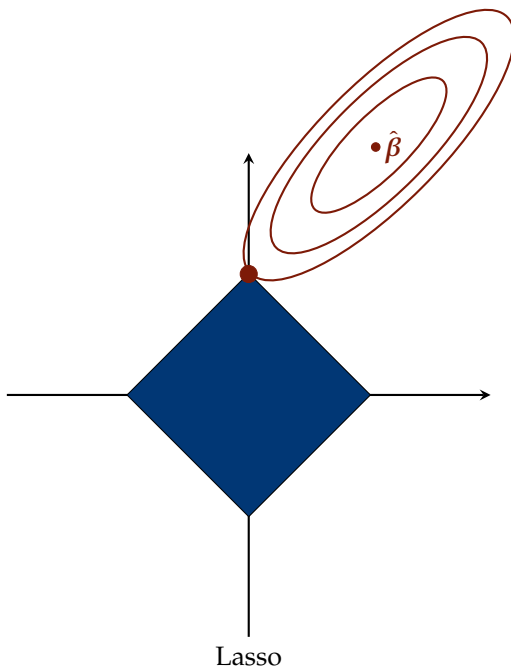
where  $p_\lambda(\cdot)$  is a penalty function indexed by the regularization parameter  $\lambda \geq 0$ . **Antoniadis and Fan (2001)** showed that the PLS estimator  $\hat{\beta}$  has the following properties:

- **sparsity**: if  $\min_{t \geq 0} \{t + p'_\lambda(t)\} > 0$
- **approximate unbiasedness**: if  $p'_\lambda(t) = 0$  for  $t$  large enough

- **continuity**: iff  $\arg \min_{t \geq 0} \{t + p'_\lambda(t)\} = 0$

In general

- the **sigularity** of penalty function at the origin,  $p'_\lambda(0_+) > 0$  is needed for generating **sparsity** in variable selection
- the **concavity** is needed to reduce the bias



## References

- Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001.
- Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. doi: 10.1214/0090536040000000067. URL <https://doi.org/10.1214/0090536040000000067>.
- Nicolai Meinshausen and Peter Bühlmann. Variable selection and high-dimensional graphs with the lasso. *Ann Stat*, 34:1436–1462, 2006.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.