

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 α 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 $\lambda \geq 0$:

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

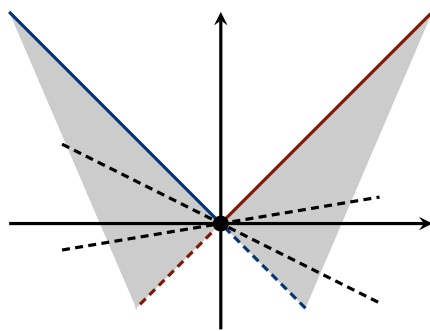
11.1.2 Solving Lasso

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

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

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

this result follows the fact the L-1 norm $\|\beta\|_1$ is piecewise linear:



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

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

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

11.1.3 Algorithm: from LARS to Lasso

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,

11.2 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

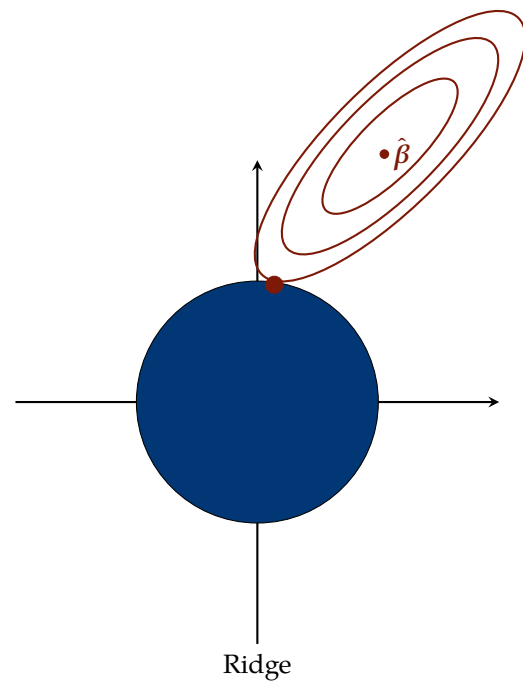
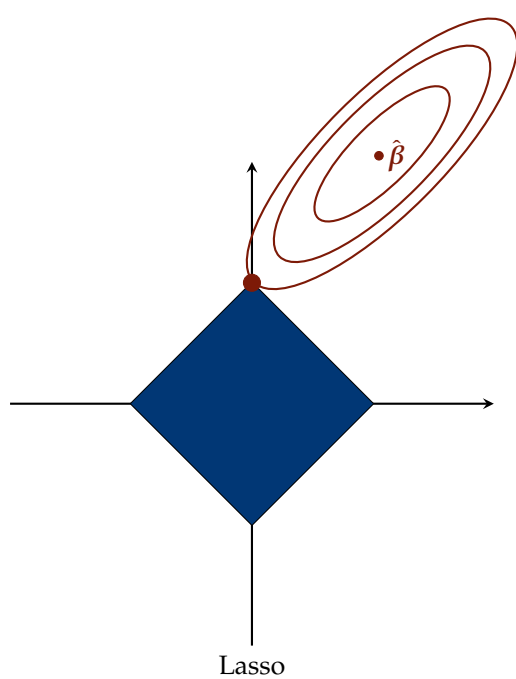
$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \right\}$$

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 **singularity** 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>.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.