UNIVERSITY of DELAWARE

Modern Machine Learning
Lasso Regression

Kenneth E. Barner
Department of Electrical and Computer Engineering
University of Delaware

Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator)

- Penalizes the coefficients $\boldsymbol{\beta}$ using the ℓ_1 norm
- Formulation

$$\widehat{\boldsymbol{\beta}}_{lasso} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{(p+1)}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \qquad (*)$$

Lasso features:

- Sets some coefficients to zero, yielding a sparse solution (model selection)
- No closed form solution, but can be solved efficiently using convex optimization methods

Note: Lasso Regression can be equivalently formulated as

$$\widehat{\boldsymbol{\beta}}_{lasso} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{(p+1)}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \qquad (**)$$

$$\text{subject to} \|\boldsymbol{\beta}\|_1 < t$$

There is a one-to-one correspondence between λ in $(*)$ and t in $(**)$

Lasso Regression

Lasso & Ridge Comparison

Lasso constraint/penalty: $||\boldsymbol{\beta}||_1 \leq t$

- Lagrange multiplier: $\lambda||\boldsymbol{\beta}||_1$

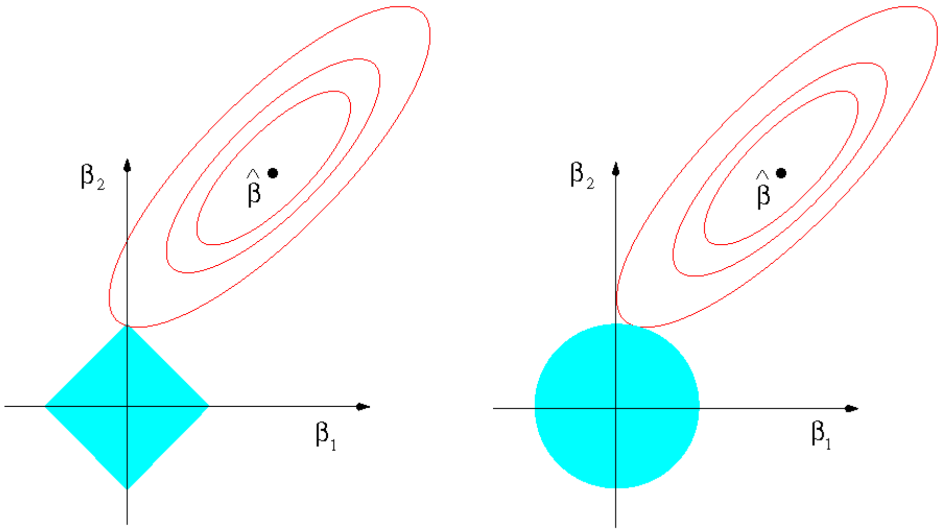
Ridge constraint/penalty: $||\boldsymbol{\beta}||_2 \leq t$

- Lagrange multiplier: $\lambda||\boldsymbol{\beta}||_2$

Thus

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{(p+1)}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_q \quad q \in \{1,2\}$$

Note: The solutions are the intersection between the constraint functions and the error (cost function) contours



Observation: the lasso constraint is diamond-shaped

- If the solution occurs at a corner, one coefficient is zero valued
- For higher dimensions, there are many "corners" forcing solutions with multiple zero coefficients

Shrinkage Constraint Comparison

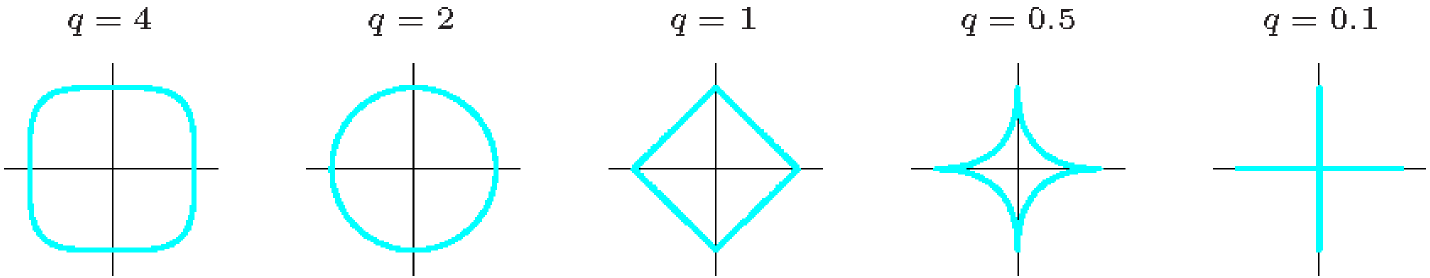
Consider the more generalized regression problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{(p+1)}}{\operatorname{argmin}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_q \qquad q \geq 0$$

The contours for $||\boldsymbol{\beta}||_q$ are plotted below for the 2D case.

Observations:

- The case $q = 1$ (Lasso) is the smallest q such that the constraint region is convex
- Decreasing q forces solutions to the coordinate axis (reducing the number of nonzero coefficients)



Lasso Regression

Shrinkage Comparison

Consider the case of $\mathbf{X}$ with orthogonal columns. The resulting shrinkage of the LS solution is:

$$\hat{\beta}_j/(1+\lambda) \qquad \text{[Ridge]}$$

$$\text{sign}(\hat{\beta}_j)\left(|\hat{\beta}_j| - \lambda\right)_+ \qquad \text{[Lasso]}$$

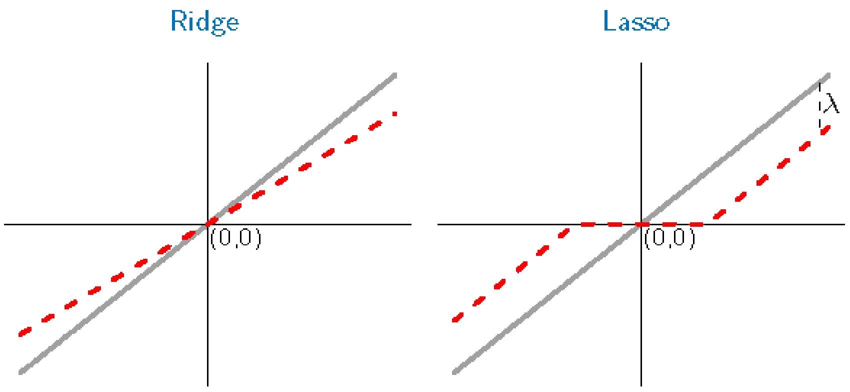
where $\hat{\beta}_j$ is element j of $\widehat{\boldsymbol{\beta}}_{LS}$, sign$(\cdot)$ denotes the sign of its argument (± 1) and

$$x_+ = \begin{cases} x \text{ if } x > 0 \\ 0 \text{ if } x \leq 0 \end{cases}$$

Note: results derived similarly to SVD analysis performed in Ridge Regression

Observations:

- Rigidly linearly scales (shrinks) coefficients
- Lasso translates (shrinks) each coefficient by a constant λ, truncating to 0. This is called soft thresholding



Example: solid 45° lines represent LS coefficient values and the dotted lines represent the resulting Ridge & Lasso coefficient values.

Optimization: No closed form Lasso solution exists, although the optimization problem is convex. One optimization approach is cyclic coordinate descent

Coordinate descent: based on the premise that minimizing a multivariate function $f(\boldsymbol{\beta})$ can be achieved by repeatedly minimizing along one coordinate direction at a time

$$\beta_1^{(k+1)} = \arg\min_{\beta} f\left(\beta, \beta_2^k, \beta_3^k, \dots, \beta_p^k\right)$$
$$\beta_2^{(k+1)} = \arg\min_{\beta} f\left(\beta_1^k, \beta, \beta_3^k, \dots, \beta_p^k\right)$$
$$\beta_3^{(k+1)} = \arg\min_{\beta} f\left(\beta_1^k, \beta_2^k, \beta, \dots, \beta_p^k\right)$$
$$\vdots$$
$$\beta_p^{(k+1)} = \arg\min_{\beta} f\left(\beta_1^k, \beta_2^k, \beta_3^k, \dots, \beta\right)$$

Consider the optimization along a single coordinate (with the $\frac{1}{2}$ included for convenience)

$$\min_{\beta_i} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1$$

$$\min_{\beta_i} \frac{1}{2} \sum_{l=1}^{n} \left(y_l - \sum_{m=1}^{p} x_{lm}\beta_m \right)^2 + \lambda \sum_{k=1}^{p} |\beta_k| \qquad (*)$$

Differentiating the first term in $(*)$ with respect to β_i:

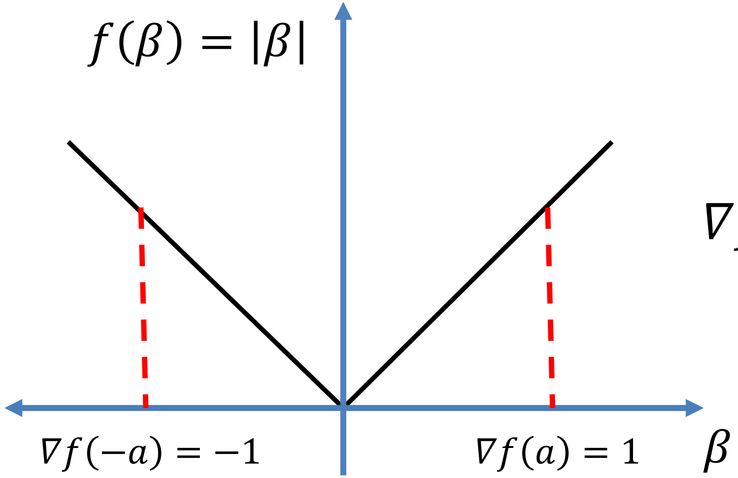
$$\begin{aligned} \frac{\partial}{\partial \beta_i} \frac{1}{2} \sum_{l=1}^{n} \left(y_l - \sum_{m=1}^{p} x_{lm}\beta_m \right)^2 &= \sum_{l=1}^{n} \left(y_l - \sum_{m=1}^{p} x_{lm}\beta_m \right)(-x_{li}) \\ &= \mathbf{X}_i^T (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) \\ &= \mathbf{X}_i^T (\mathbf{X}_{-i}\boldsymbol{\beta}_{-i} - \mathbf{y}) + \mathbf{X}_i^T \mathbf{X}_i \beta_i, \end{aligned}$$

where $\mathbf{X}_{-i}$ denotes matrix $\mathbf{X}$ excluding the ith column, and $\mathbf{X}_i$ is the ith column of matrix $\mathbf{X}$

Lasso Regression

Consider the second term in $(*)$. Note the non-differential term

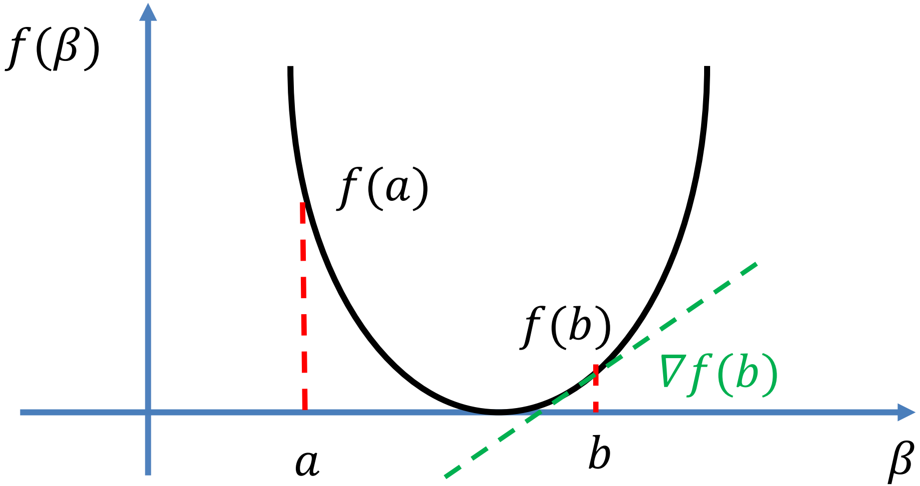
$$\frac{\partial}{\partial \beta}|\beta| =??$$



$\nabla f(0)$ is undetermined

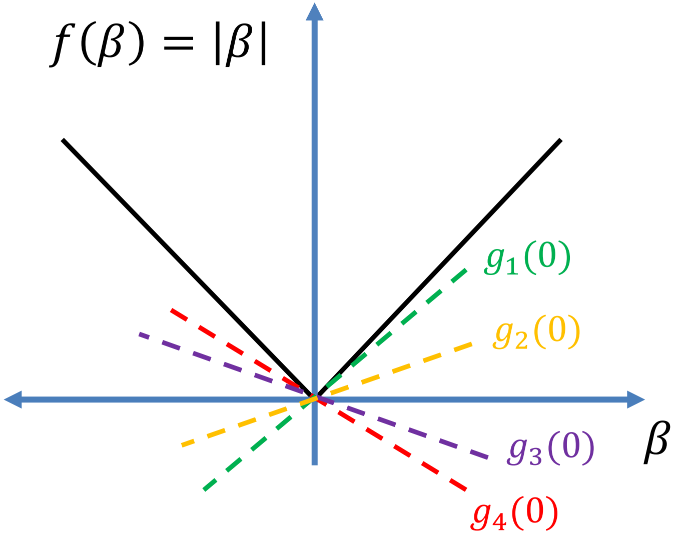
Convex function

For any convex function we have that $f(a) \geq f(b) + \nabla f(b)(a-b)$



Subgradient: Generalizes the concept to non-differentiable points

- g is a subgradient if $f(b) \geq f(a) + g(b-a)$
- $g(0)$: =Any plane that lower bounds the function $f(\beta)$ (it is a set)



$\partial f(\beta)$: = all subgradients of f at β

$$\partial f(\beta) = \begin{cases} -1 & \text{if } \beta < 0 \\ [-1,1] & \text{if } \beta = 0 \\ 1 & \text{if } \beta > 0 \end{cases}$$

Define

$$f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

and set

$$g_i = \frac{\partial}{\partial \beta_i}\ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \mathbf{X}_i^T(\mathbf{X}_{-i}\boldsymbol{\beta}_{-i} - \mathbf{y}) + \mathbf{X}_i^T\mathbf{X}_i\beta_i$$

Complete solution has 3 cases:

$$\partial f(\beta_i) = \begin{cases} g_i - \lambda & \text{if } \beta_i < 0 \\ [g_i - \lambda, g_i + \lambda] & \text{if } \beta_i = 0 \\ g_i + \lambda & \text{if } \beta_i > 0 \end{cases}$$

To find the optimal solution, set the derivative to 0 for all 3 cases

Case 1: $\beta_i < 0 \Longrightarrow g_i - \lambda = 0$

$$g_i - \lambda = 0 \Leftrightarrow \beta_i = \frac{\mathbf{X}_i^T(\mathbf{y} - \mathbf{X}_{-i}\boldsymbol{\beta}_{-i}) + \lambda}{\mathbf{X}_i^T\mathbf{X}_i} = \tilde{g}_i + \frac{\lambda}{\|\mathbf{X}_i\|^2}$$

where $\tilde{g}_i = \frac{\mathbf{X}_i^T(\mathbf{y} - \mathbf{X}_{-i}\boldsymbol{\beta}_{-i})}{\mathbf{X}_i^T\mathbf{X}_i}$

Case 2: $\beta_i > 0 \Longrightarrow g_i + \lambda = 0$

$$g_i + \lambda = 0 \Leftrightarrow \beta_i = \frac{\mathbf{X}_i^T(\mathbf{y} - \mathbf{X}_{-i}\boldsymbol{\beta}_{-i}) - \lambda}{\mathbf{X}_i^T\mathbf{X}_i} = \tilde{g}_i - \frac{\lambda}{\|\mathbf{X}_i\|^2}$$

Case 3: $\beta_i = 0 \Longrightarrow 0 \in [g_i - \lambda, g_i + \lambda]$ — need to prove conditions guaranteeing 0 is in this set

Consider the boundaries of Cases 1 & 2, and suppose: $-\frac{\lambda}{\|\mathbf{X}_i\|^2} < \tilde{g}_i < \frac{\lambda}{\|\mathbf{X}_i\|^2}$

$$-\frac{\lambda}{\|\mathbf{X}_i\|^2} < \tilde{g}_i < \frac{\lambda}{\|\mathbf{X}_i\|^2} \Leftrightarrow -\frac{\lambda}{\|\mathbf{X}_i\|^2} < \frac{\mathbf{X}_i^T(\mathbf{y} - \mathbf{X}_{-i}\boldsymbol{\beta}_{-i})}{\mathbf{X}_i^T\mathbf{X}_i} < \frac{\lambda}{\|\mathbf{X}_i\|^2}$$

$$\Leftrightarrow -\lambda < \mathbf{X}_i^T(\mathbf{y} - \mathbf{X}_{-i}\boldsymbol{\beta}_{-i}) < \lambda \qquad (*)$$

Recall $g_i = \mathbf{X}_i^T(\mathbf{X}_{-i}\boldsymbol{\beta}_{-i} - \mathbf{y}) + \mathbf{X}_i^T\mathbf{X}_i\beta_i$

Thus $\beta_i = 0 \Longrightarrow g_i = -\mathbf{X}_i^T(\mathbf{y} - \mathbf{X}_{-i}\boldsymbol{\beta}_{-i})$ and combining with $(*)$

$$-\lambda < -g_i < \lambda \Longrightarrow 0 \in [g_i - \lambda, g_i + \lambda]$$

Therefore we have shown that $0 \in \partial f(\beta_i)$ for $\beta_i = 0$ if $-\frac{\lambda}{\|\mathbf{X}_i\|^2} < \tilde{g}_i < \frac{\lambda}{\|\mathbf{X}_i\|^2}$

Combining the 3 cases yields the Lasso shrinkage optimization

$$\hat{\beta}_i = \begin{cases} \tilde{g}_i + \dfrac{\lambda}{\|\mathbf{X}_i\|^2} & \text{if } \tilde{g}_i < -\dfrac{\lambda}{\|\mathbf{X}_i\|^2} \\ 0 & \text{if } -\dfrac{\lambda}{\|\mathbf{X}_i\|^2} < \tilde{g}_i < \dfrac{\lambda}{\|\mathbf{X}_i\|^2} \\ \tilde{g}_i - \dfrac{\lambda}{\|\mathbf{X}_i\|^2} & \text{if } \tilde{g}_i > \dfrac{\lambda}{\|\mathbf{X}_i\|^2} \end{cases}$$

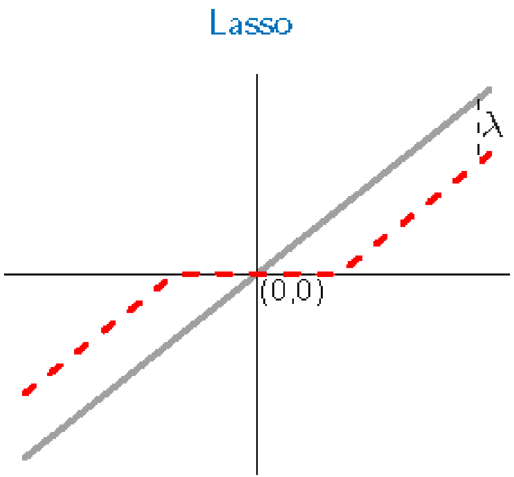
where $\tilde{g}_i = \frac{\mathbf{X}_i^T(\mathbf{y}-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i})}{\mathbf{X}_i^T\mathbf{X}_i}$

The solution can be expressed in terms of the soft thresholding operator

$$\hat{\beta}_i = \eta^S_{\lambda/\|\mathbf{X}_i\|^2}(\tilde{g}_i) = \eta^S_{\lambda/\|\mathbf{X}_i\|^2}\left(\frac{\mathbf{X}_i^T(\mathbf{y}-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i})}{\mathbf{X}_i^T\mathbf{X}_i}\right)$$

Lasso Regression

Thus for $\hat{\beta}_i = \eta^S_{\lambda/\|\mathbf{x}_i\|^2}(\tilde{g}_i)$ we use the Soft-Thresholding operator



$$\eta^S_\lambda(\beta) = \text{sign}(\beta)(|\beta| - \lambda)_+ = \begin{cases} \beta + \lambda & \text{if } \beta < -\lambda \\ 0 & \text{if } -\lambda < \beta < \lambda \\ \beta - \lambda & \text{if } \beta > \lambda \end{cases}$$

Final Result: Apply cyclic coordinate descent to obtain the lasso solution

$$\beta_1^{(k+1)} = \eta^S_{\lambda/\|\mathbf{x}_1\|^2}(\tilde{g}_1)$$
$$\beta_2^{(k+1)} = \eta^S_{\lambda/\|\mathbf{x}_2\|^2}(\tilde{g}_2)$$
$$\vdots$$
$$\beta_p^{(k+1)} = \eta^S_{\lambda/\|\mathbf{x}_p\|^2}(\tilde{g}_p),$$

where $\tilde{g}_i = \frac{\mathbf{x}_i^T(\mathbf{y}-\mathbf{X}_{-i}\boldsymbol{\beta}_{-i})}{\mathbf{x}_i^T\mathbf{x}_i}$

$\Longrightarrow$ cycle through the coordinates until convergence