Analysis of the Lasso

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

Carlos Fernandez-Granda

Prerequisites

Sparse regression via the lasso

Convexity

Subgradients

Additive model

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

Goal: Gain intuition about why the lasso promotes sparse solutions

Sparse regression with two features

One true feature

$$\tilde{y} := x_{\text{true}} + \tilde{z}$$

We fit a model using an additional feature

$$X := \begin{bmatrix} x_{\text{true}} & x_{\text{other}} \end{bmatrix}^T$$

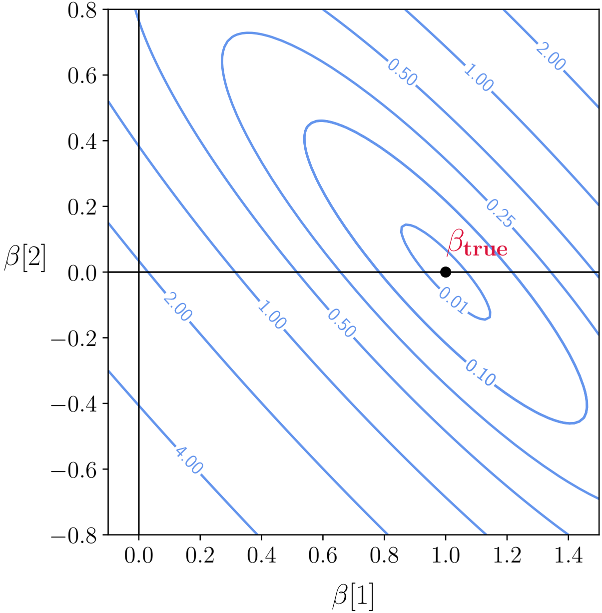
$$\beta_{\text{true}} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Decomposition of lasso cost function

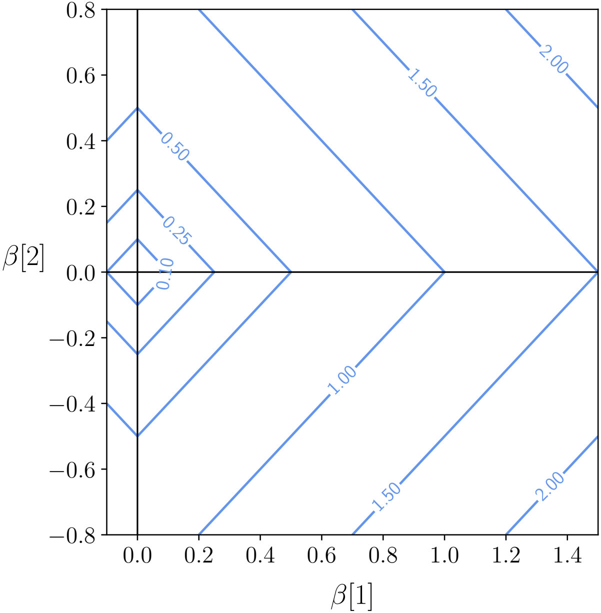
$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2 + \lambda\,||\beta||_1$$

$$= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda\,||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$

$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}})$

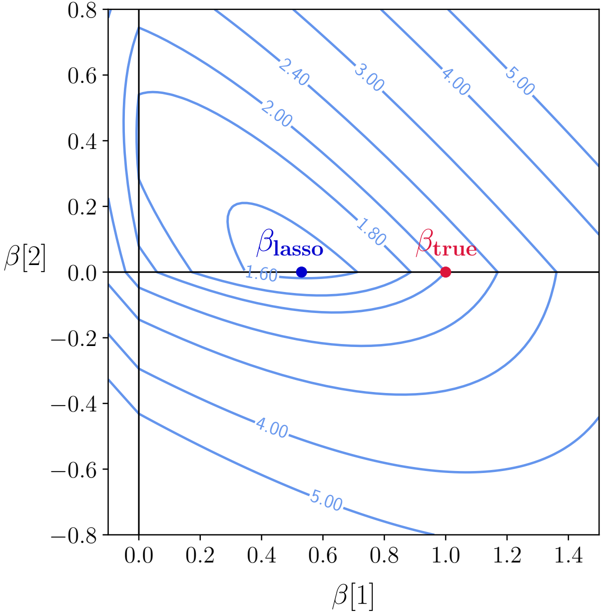


$||\beta||_1$

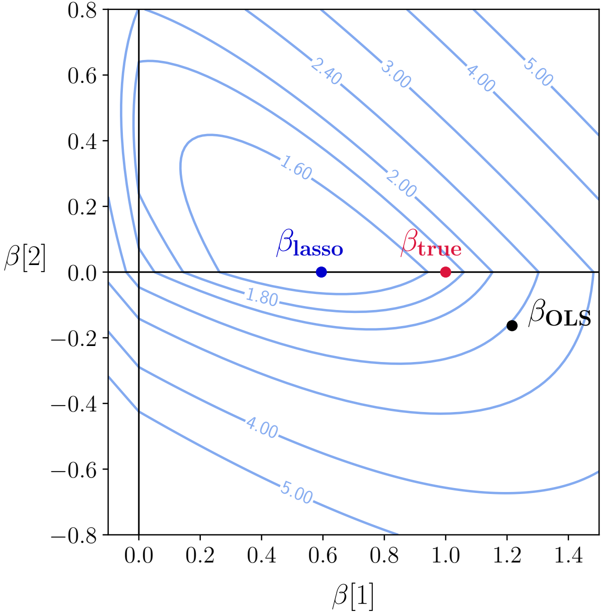


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
β[2]
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
β[1]
0.10
0.25
0.50
1.00
1.50
2.00
1.50
2.00

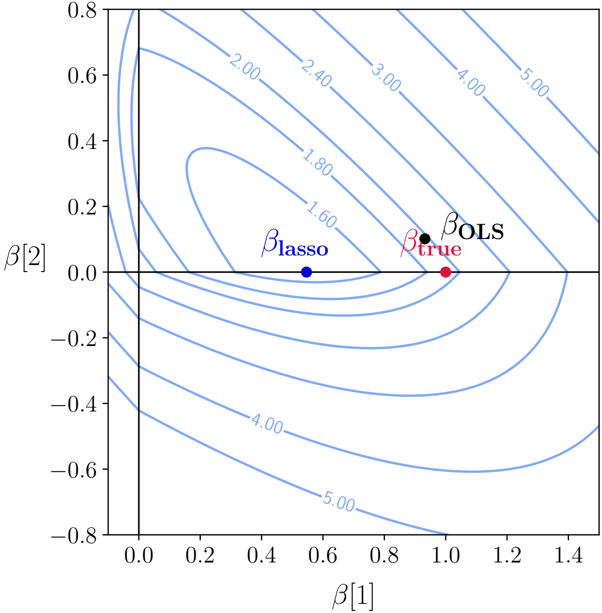
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda\, ||\beta||_1$$



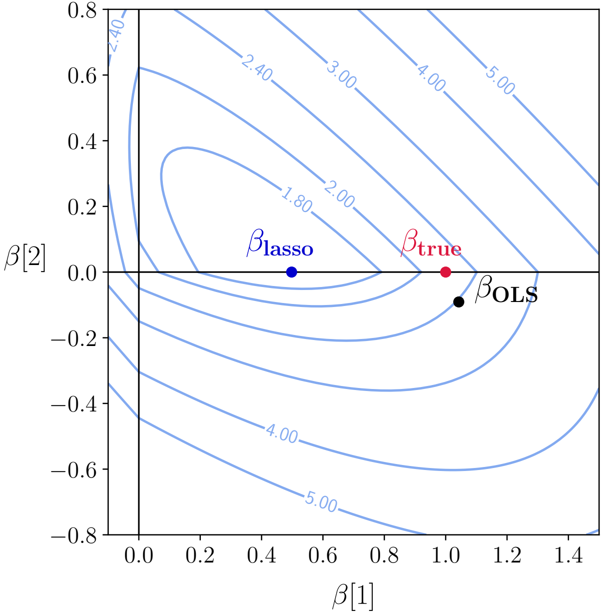
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \, ||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$



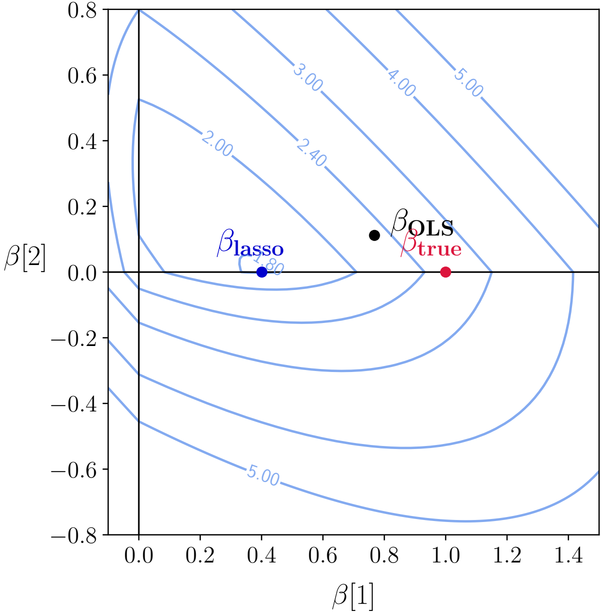
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \,||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$



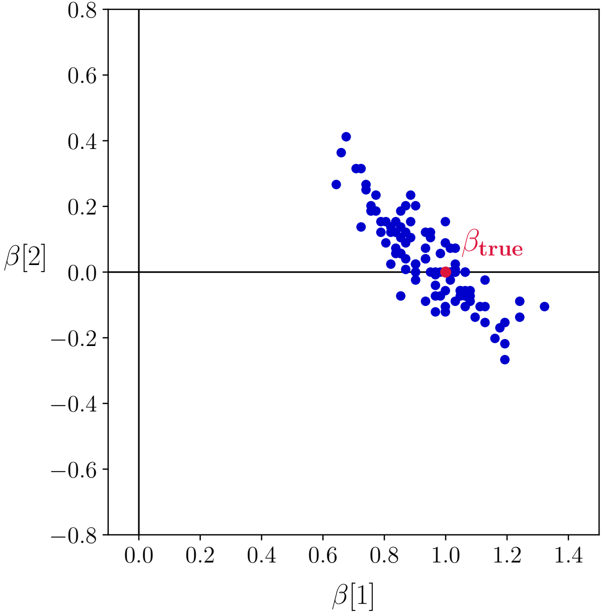
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \|\beta\|_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$



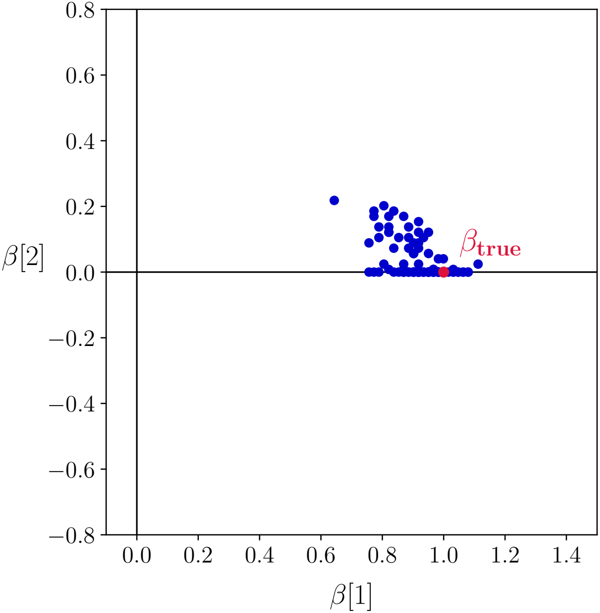
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \, ||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$



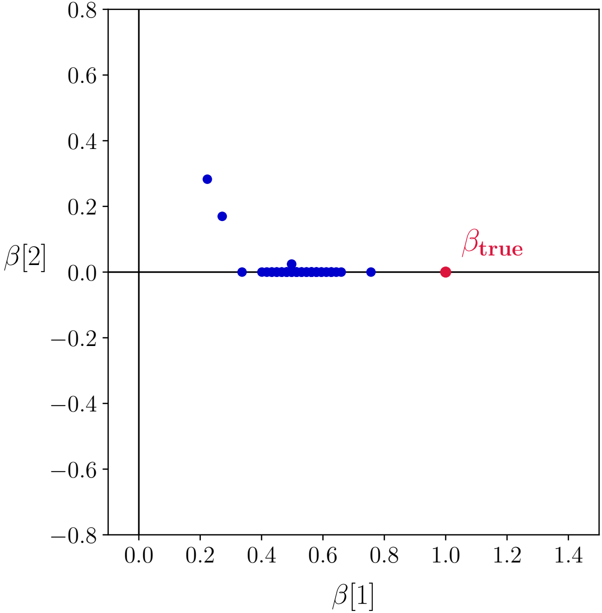
$\lambda = 0.02$



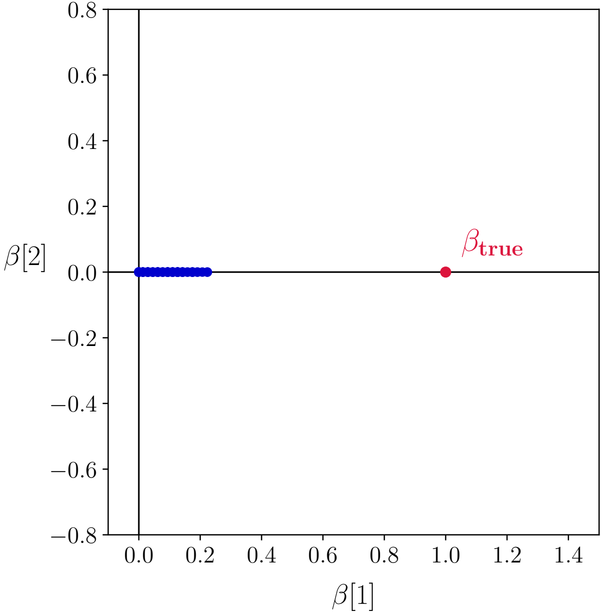
$\lambda = 0.2$



$\lambda = 2$



$\lambda = 4$



$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Sparse regression with two features

Feature vectors and noise are fixed n-dimensional vectors

$$y := x_{\text{true}} + z$$

We fit a model using an additional feature

$$X := \begin{bmatrix} x_{\text{true}} & x_{\text{other}} \end{bmatrix}^T$$

$$\beta_{\text{true}} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$||x_{\text{true}}||_2 = ||x_{\text{other}}||_2 = 1$$

Sparse regression with two features

If λ satisfies

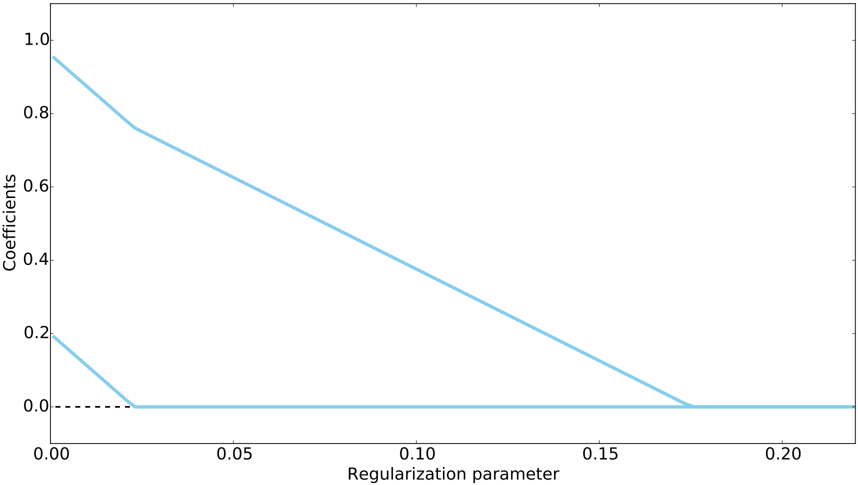
$$\frac{\left|x_{\text{other}}^T z - \rho x_{\text{true}}^T z\right|}{1 - |\rho|} \leq \lambda \leq 1 + x_{\text{true}}^T z$$

then the lasso coefficient estimate equals

$$\beta_{\text{lasso}} = \begin{bmatrix} 1 + x_{\text{true}}^T z - \lambda \\ 0 \end{bmatrix}$$

where $\rho := x_{\text{true}}^T x_{\text{other}}$

Lasso coefficients



Analyzing the lasso

How do we prove this?

No closed-form solution!

Show that zero is a subgradient of lasso cost function at β_{lasso}

Subgradients of lasso cost function

Gradient of $\frac{1}{2} \left|\left|X^T \beta - y\right|\right|_2^2$ at β_{lasso}:

$$X \left(X^T \beta_{\text{lasso}} - y \right)$$

Subgradient of ℓ_1 norm at β_{lasso} if only first entry is nonzero and positive:

$$g_{\ell_1} := \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \qquad |\gamma| \leq 1$$

Subgradient of lasso cost function at β_{lasso} if only first entry is nonzero and positive:

$$g_{\text{lasso}} := X \left(X^T \beta_{\text{lasso}} - y \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \qquad |\gamma| \leq 1$$

Subgradients of lasso cost function

$$
\begin{aligned}
g_{\text{lasso}} &:= X\left(X^T\beta_{\text{lasso}} - y\right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= X\left(\beta_{\text{lasso}}[1]x_{\text{true}} - x_{\text{true}} - z\right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= \begin{bmatrix} x_{\text{true}}^T\left((\beta_{\text{lasso}}[1]-1)x_{\text{true}} - z\right) + \lambda \\ x_{\text{other}}^T\left((\beta_{\text{lasso}}[1]-1)x_{\text{true}} - z\right) + \lambda\gamma \end{bmatrix} \\
&= \begin{bmatrix} \beta_{\text{lasso}}[1] - 1 - x_{\text{true}}^T z + \lambda \\ \rho(\beta_{\text{lasso}}[1]-1) - x_{\text{other}}^T z + \lambda\gamma \end{bmatrix}
\end{aligned}
$$

Is zero a valid subgradient?

Setting $g_{\text{lasso}} = 0$

$$\beta_{\text{lasso}}[1] = 1 - \lambda + x_{\text{true}}^T z$$

$$\gamma = \frac{\rho + x_{\text{other}}^T z - \rho \beta_{\text{lasso}}[1]}{\lambda}$$

$$= \frac{x_{\text{other}}^T z - \rho x_{\text{true}}^T z}{\lambda} + \rho$$

We need $\beta_{\text{lasso}}[1] \geq 0$

$$\lambda \leq 1 + x_{\text{true}}^T z$$

We need $|\gamma| \leq 1$

$$\frac{\left| x_{\text{other}}^T z - \rho x_{\text{true}}^T z \right|}{1 - |\rho|} \leq \lambda$$

What have we learned?

How to analyze nondifferentiable convex cost functions using subgradients

Why the lasso works (for a very simple example)