Sparse regression

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

https://cims.nyu.edu/~cfgranda/pages/MTDS_spring20/index.html

Carlos Fernandez-Granda

Sparse regression

Linear regression is challenging when the number of features p is large

Solution: Select subset of features $\mathcal{I} \subset \{1, \ldots, p\}$, such that

$$y \approx \sum_{i \in \mathcal{I}} \beta[i] x[i]$$

Equivalently, find sparse coefficient vector $\beta \in \mathbb{R}^p$ such that

$$y \approx \langle x, \beta \rangle$$

Problem: How to promote sparsity?

Toy problem

Find t such that

$$v_t := \begin{bmatrix} t \\ t-1 \\ t-1 \end{bmatrix}$$

is sparse

Equivalently, find $\arg\min_t ||v_t||_0$

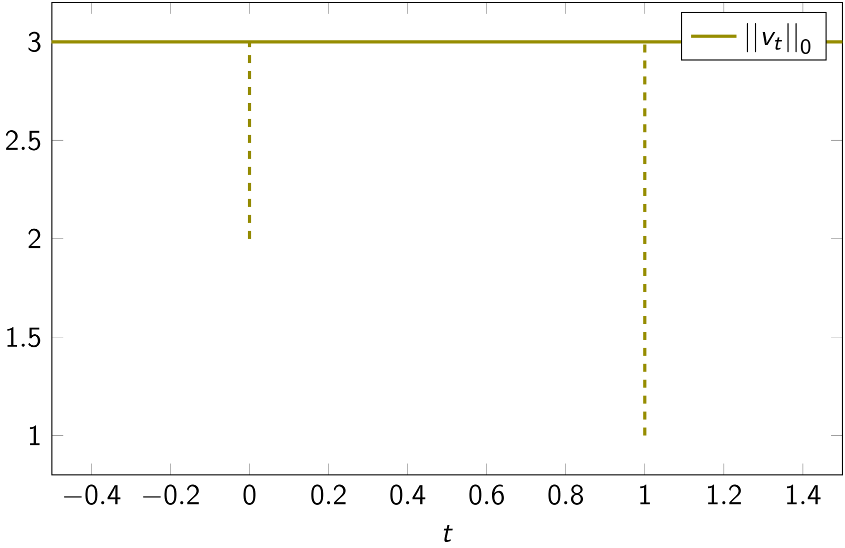
ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2x||_0 &= ||x||_0 \\ &\neq 2\,||x||_0 \end{aligned}$$

Toy problem

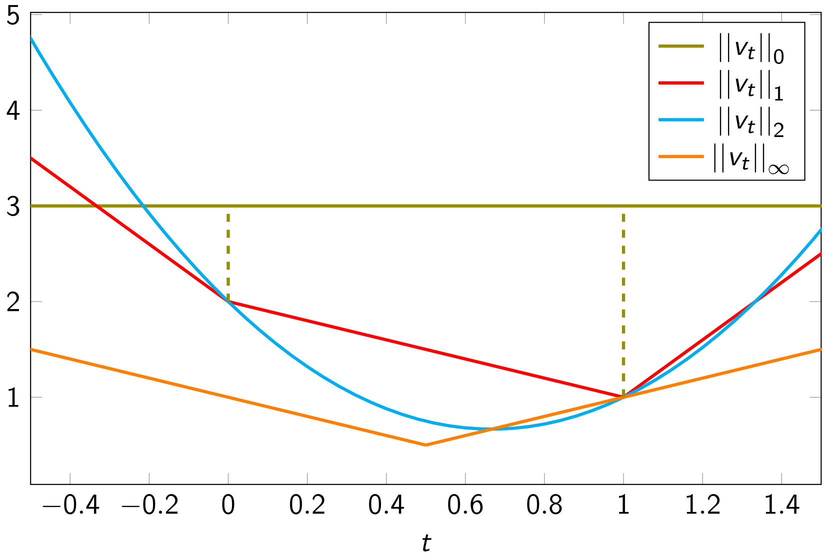


Alternative strategy

Minimize another norm

$$f(t) := ||v_t||$$

Toy problem



The lasso

Convexity

Subgradients

Analysis of the lasso estimator for a simple example

Sparse linear regression

Find a small subset of useful features

Model selection problem

Two objectives:

- Good fit to the data; $||X^T\beta - y||_2^2$ should be as small as possible
- Using a small number of features; β should be as sparse as possible

The lasso

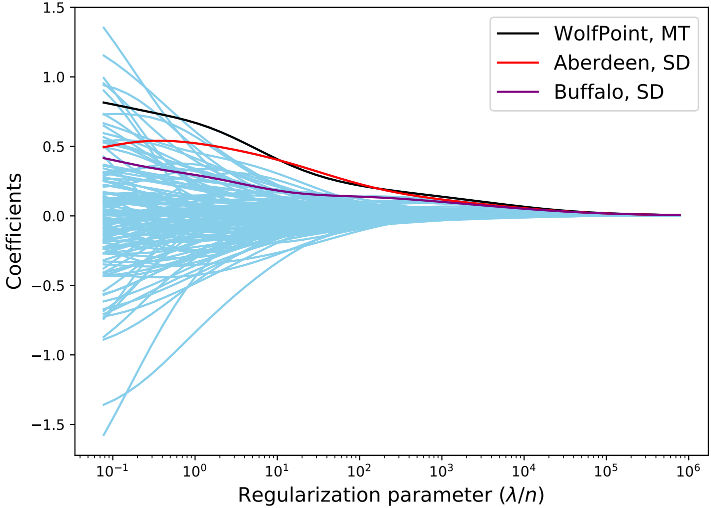
Uses ℓ_1-norm regularization to promote sparse coefficients

$$\beta_{\text{lasso}} := \arg\min_{\beta} \frac{1}{2} \left|\left| y - X^T \beta \right|\right|_2^2 + \lambda \left|\left|\beta\right|\right|_1$$

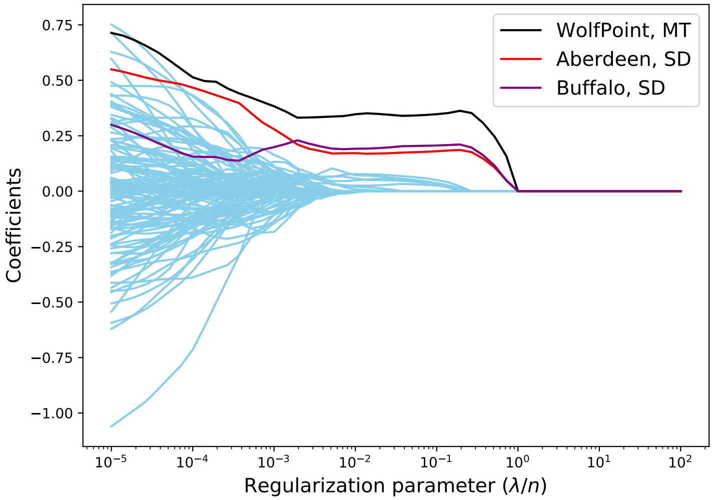
Temperature prediction via linear regression

- Dataset of hourly temperatures measured at weather stations all over the US
- Goal: Predict temperature in Jamestown (North Dakota) from other temperatures
- Response: Temperature in Jamestown
- Features: Temperatures in 133 other stations ($p = 133$) in 2015
- Test set: 10^3 measurements
- Additional test set: All measurements from 2016

Ridge regression $n := 135$

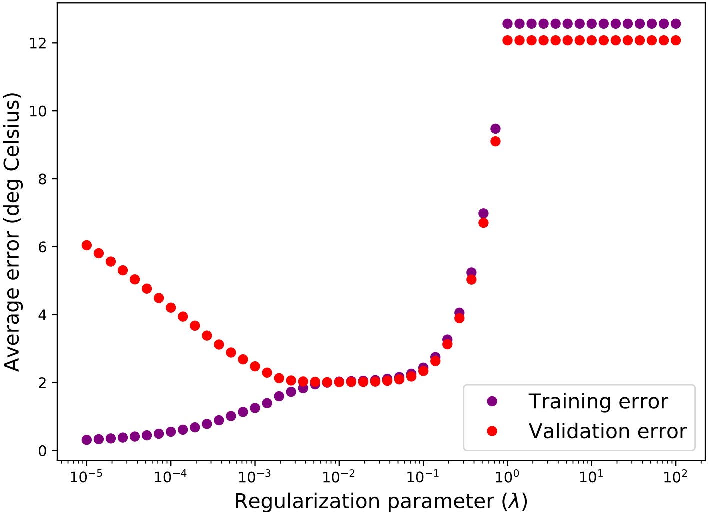


Lasso $n := 135$

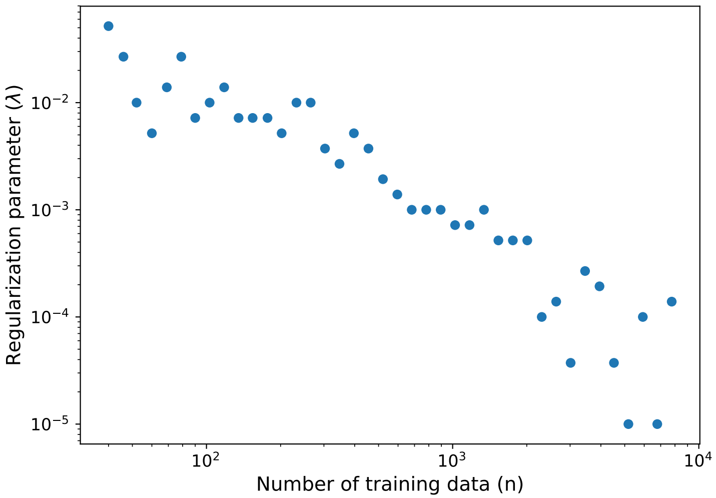


WolfPoint, MT
Aberdeen, SD
Buffalo, SD
Coefficients
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Regularization parameter (λ/n)

Lasso $n := 135$

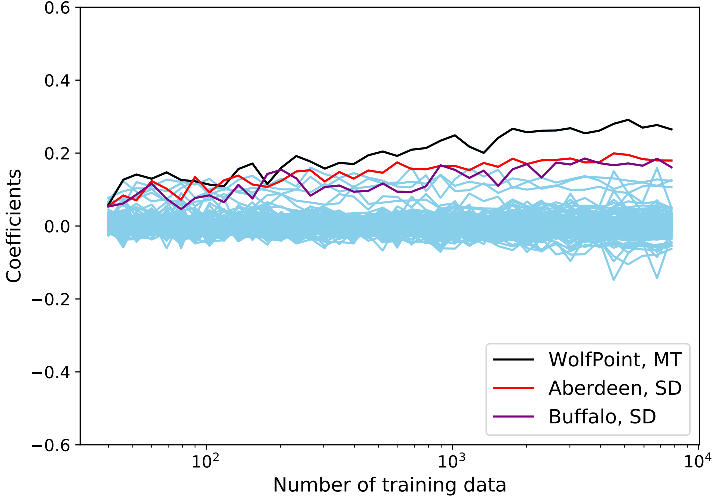

Average error (deg Celsius)
Regularization parameter (λ)
Training error
Validation error
0
2
4
6
8
10
12
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

Lasso $n := 135$

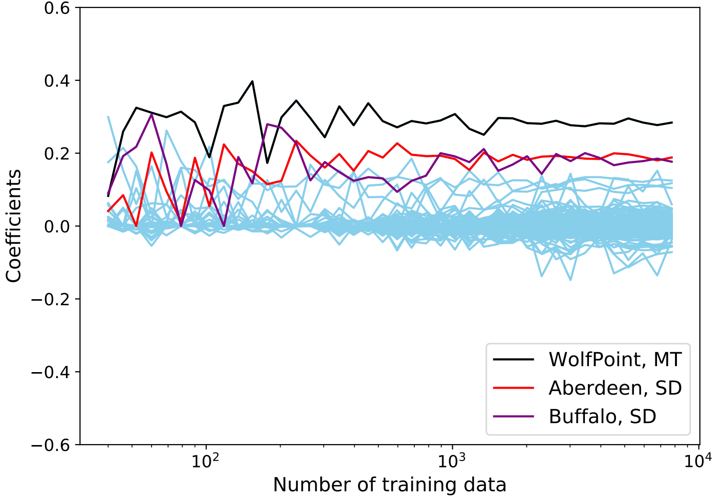


Regularization parameter (λ)
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{2}
10^{3}
10^{4}
Number of training data (n)

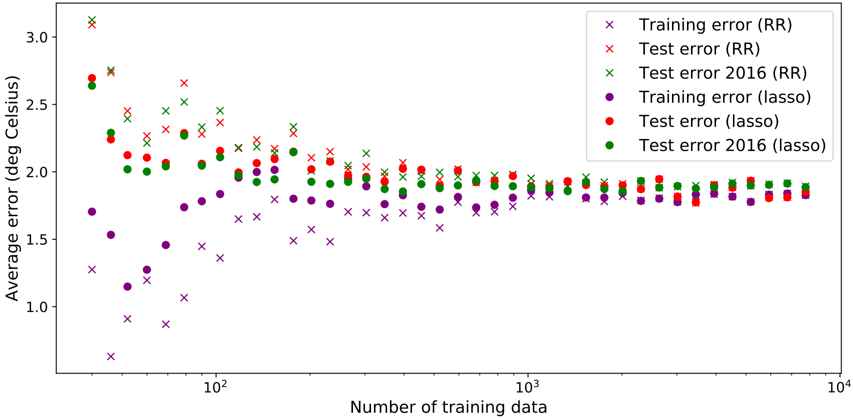
Ridge-regression coefficients



Lasso coefficients



Results



Training error (RR)
Test error (RR)
Test error 2016 (RR)
Training error (lasso)
Test error (lasso)
Test error 2016 (lasso)
Average error (deg Celsius)
Number of training data
3.0
2.5
2.0
1.5
1.0
10^2
10^3
10^4

The lasso

Convexity

Subgradients

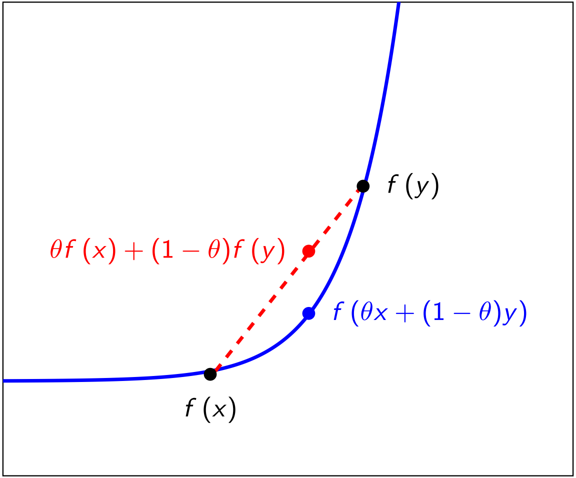
Analysis of the lasso estimator for a simple example

Convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\theta f(x) + (1 - \theta) f(y) \geq f(\theta x + (1 - \theta) y)$$

Convex functions



Strictly convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\theta f(x) + (1 - \theta) f(y) > f(\theta x + (1 - \theta) y)$$

Linear and quadratic functions

Linear functions are convex

$$f\left(\theta x + (1-\theta)\, y\right) = \theta f(x) + (1-\theta)\, f(y)$$

Positive definite quadratic forms are strictly convex

Norms are convex

For any $x, y \in \mathbb{R}^n$ and any $\theta \in (0,1)$

$$\begin{aligned} \|\theta x + (1-\theta)\, y\| &\leq \|\theta x\| + \|(1-\theta)\, y\| \\ &= \theta\, \|x\| + (1-\theta)\, \|y\| \end{aligned}$$

ℓ_0 "norm" is not convex

Let $x := \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $y := \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$, for any $\theta \in (0, 1)$

$$\left|\left|\theta x + (1 - \theta)\, y\right|\right|_0 = 2$$

$$\theta \left|\left|x\right|\right|_0 + (1 - \theta) \left|\left|y\right|\right|_0 = 1$$

Is the lasso cost function convex?

f strictly convex, g convex, $h := f + \lambda g$?

$$\begin{aligned} h(\theta x + (1-\theta)\, y) &= f\left(\theta x + (1-\theta)\, y\right) + \lambda g(\theta x + (1-\theta)\, y) \\ &< \theta f(x) + (1-\theta)\, f(y) + \lambda \theta g(x) + \lambda\,(1-\theta)\, g(y) \\ &= \theta h(x) + (1-\theta)\, h(y) \end{aligned}$$

Lasso cost function is convex

Sum of convex functions is convex

If at least one is strictly convex, then sum is strictly convex

Scaling by a positive factor preserves convexity

Lasso cost function is convex!

Local minima are global

Any local minimum of a convex function is also a global minimum

Strictly convex functions

Strictly convex functions have at most one global minimum

Proof: Assume two minima exist at $x \neq y$ with value $v_{\min}$

$$\begin{aligned} f\left(0.5x + 0.5y\right) &< 0.5f\left(x\right) + 0.5f\left(y\right) \\ &= v_{\min} \end{aligned}$$

The lasso

Convexity

Subgradients

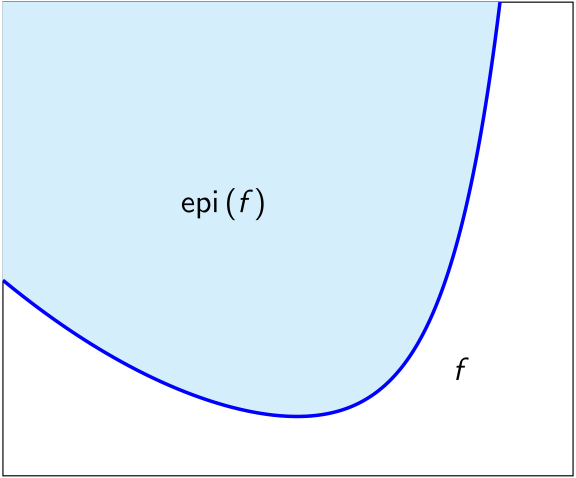
Analysis of the lasso estimator for a simple example

Epigraph

The epigraph of $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\operatorname{epi}(f) := \left\{ x \mid f\left(\begin{bmatrix} x[1] \\ \cdots \\ x[n] \end{bmatrix} \right) \leq x[n+1] \right\}$$

Epigraph

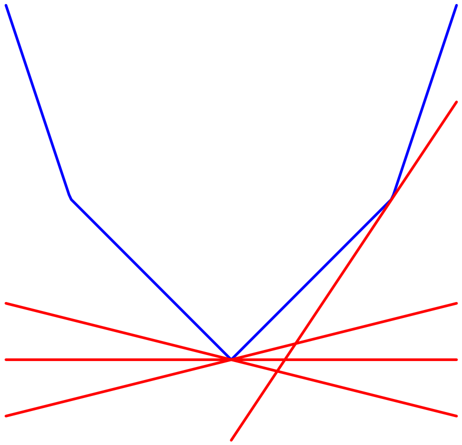


Supporting hyperplane

A hyperplane $\mathcal{H}$ is a supporting hyperplane of a set $\mathcal{S}$ at x if

- $\mathcal{H}$ and $\mathcal{S}$ intersect at x
- $\mathcal{S}$ is contained in one of the half-spaces bounded by $\mathcal{H}$

Supporting hyperplane



Subgradient

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if it has a supporting hyperplane at every point

It is strictly convex if and only for all $x \in \mathbb{R}^n$ it only intersects with the supporting hyperplane at one point

Subgradients

The subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is a vector $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x), \quad \text{for all } y \in \mathbb{R}^n$$

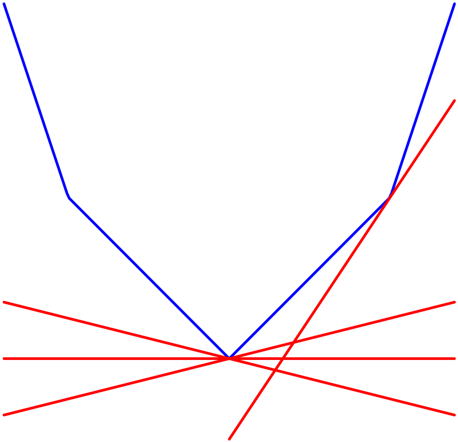
The hyperplane

$$\mathcal{H}_g := \left\{ y \mid y[n+1] = g^T \left(\begin{bmatrix} y[1] \\ \cdots \\ y[n] \end{bmatrix} \right) \right\}$$

is a supporting hyperplane of the epigraph at x

The set of all subgradients at x is called the subdifferential

Subgradients



Subgradient of differentiable function

If a function is differentiable, the only subgradient at each point is the gradient

Proof

Assume g is a subgradient at x, for any $\alpha \geq 0$

$$\begin{aligned}
f(x + \alpha\, e_i) &\geq f(x) + g^T \alpha\, e_i \\
&= f(x) + g[i]\, \alpha \\
f(x) &\leq f(x - \alpha\, e_i) + g^T \alpha\, e_i \\
&= f(x - \alpha\, e_i) + g[i]\, \alpha
\end{aligned}$$

Combining both inequalities

$$\frac{f(x) - f(x - \alpha\, e_i)}{\alpha} \leq g[i] \leq \frac{f(x + \alpha\, e_i) - f(x)}{\alpha}$$

Letting $\alpha \to 0$, implies $g[i] = \frac{\partial f(x)}{\partial x[i]}$

Optimality condition for nondifferentiable functions

x is a minimum of f if and only if the zero vector is a subgradient of f at x

$$\begin{aligned} f(y) &\geq f(x) + \vec{0}^T (y - x) \\ &= f(x) \end{aligned}$$

for all $y \in \mathbb{R}^n$

Under strict convexity the minimum is unique

Sum of subgradients

Let g_1 and g_2 be subgradients at $x \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$g := g_1 + g_2$ is a subgradient of $f := f_1 + f_2$ at x

<u>Proof:</u> For any $y \in \mathbb{R}^n$

$$\begin{aligned} f(y) &= f_1(y) + f_2(y) \\ &\geq f_1(x) + {g_1}^T (y - x) + f_2(y) + {g_2}^T (y - x) \\ &\geq f(x) + g^T (y - x) \end{aligned}$$

Subgradient of scaled function

Let g_1 be a subgradient at $x \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\alpha \geq 0$ $g_2 := \alpha g_1$ is a subgradient of $f_2 := \alpha f_1$ at x

Proof: For any $y \in \mathbb{R}^n$

$$\begin{aligned} f_2(y) &= \alpha f_1(y) \\ &\geq \alpha \left(f_1(x) + {g_1}^T (y - x) \right) \\ &\geq f_2(x) + {g_2}^T (y - x) \end{aligned}$$

Subdifferential of absolute value

At $x \neq 0$, $f(x) = |x|$ is differentiable, so $g = \operatorname{sign}(x)$

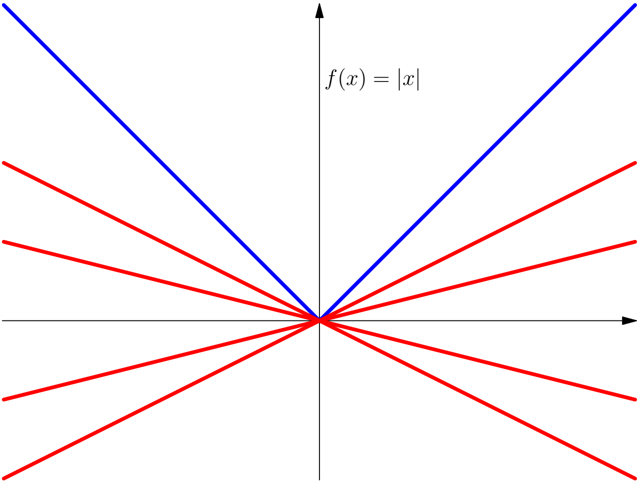
At $x = 0$, we need

$$f(0 + y) \geq f(0) + g(y - 0)$$

$$|y| \geq gy$$

Holds if and only if $|g| \leq 1$

Subdifferential of absolute value



Subdifferential of ℓ_1 norm

g is a subgradient of the ℓ_1 norm at $x \in \mathbb{R}^n$ if and only if

$$
\begin{aligned}
g[i] &= \operatorname{sign}(x[i]) && \text{if } x[i] \neq 0 \\
|g[i]| &\leq 1 && \text{if } x[i] = 0
\end{aligned}
$$

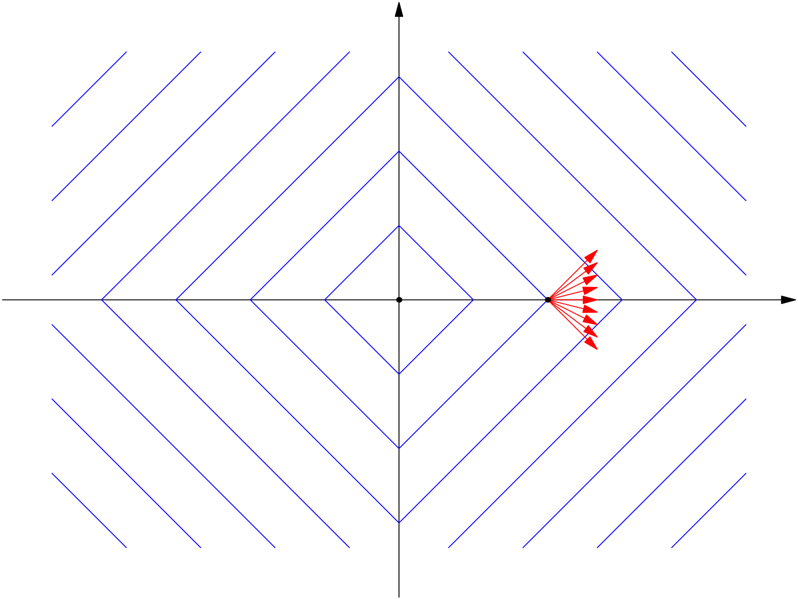
Proof (one direction)

Assume $g[i]$ is a subgradient of $|\cdot|$ at $|x[i]|$ for $1 \leq i \leq n$

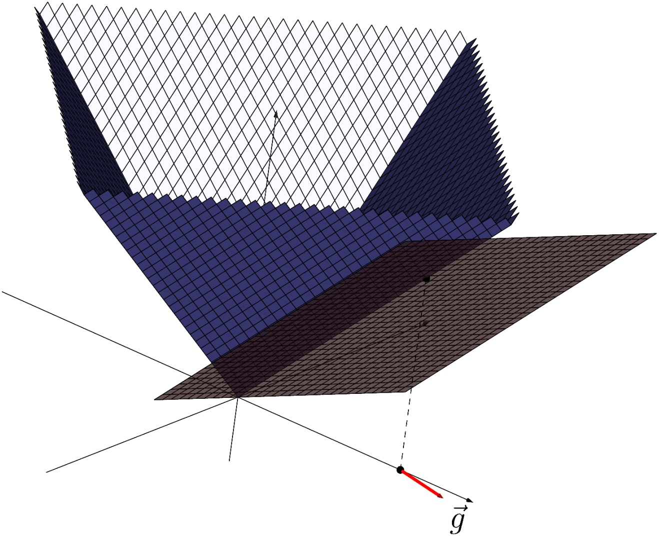
For any $y \in \mathbb{R}^n$

$$\begin{aligned}
||y||_1 &= \sum_{i=1}^{n} |y\,[i]| \\
&\geq \sum_{i=1}^{n} |x[i]| + g[i]\,(y\,[i] - x[i]) \\
&= ||x||_1 + g^{\,T}\,(y - x)
\end{aligned}$$

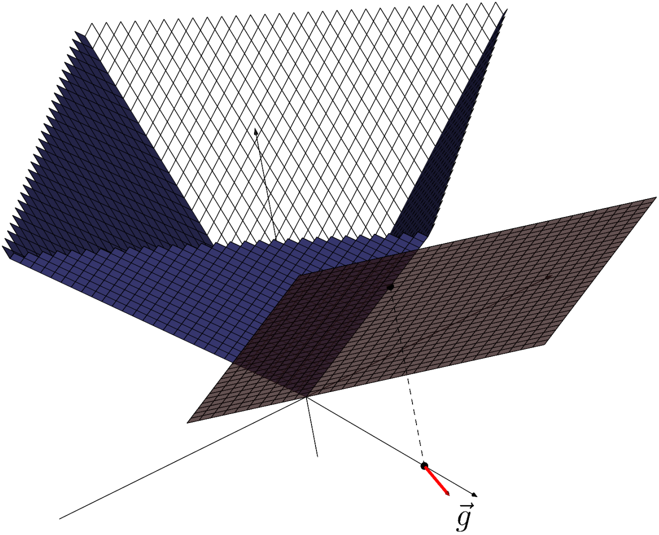
Subdifferential of ℓ_1 norm



Subdifferential of ℓ_1 norm



Subdifferential of ℓ_1 norm



The lasso

Convexity

Subgradients

Analysis of the lasso estimator for a simple example

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

Decomposition of lasso cost function

$$
\begin{aligned}
&\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2 + \lambda\,||\beta||_1 \\
&= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}}) + \lambda\,||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta
\end{aligned}
$$

Sparse regression with two features

One true feature

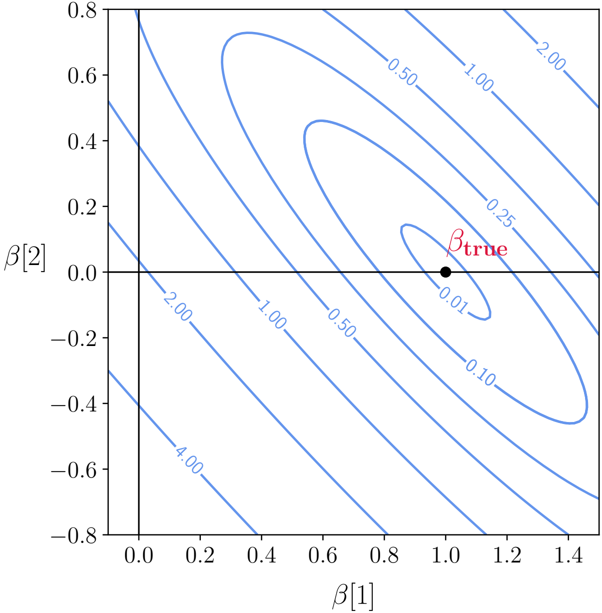
$$\tilde{y} := x_{\text{true}} + \tilde{z}$$

We fit a model using an additional feature

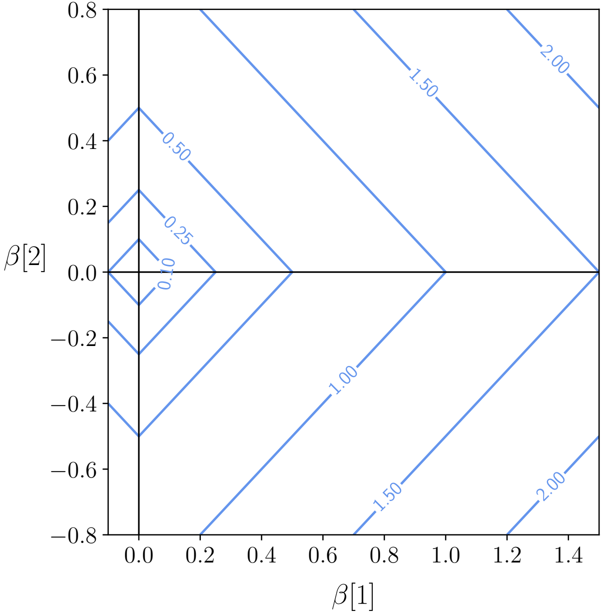
$$X := \begin{bmatrix} x_{\text{true}} & x_{\text{other}} \end{bmatrix}^T$$

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

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

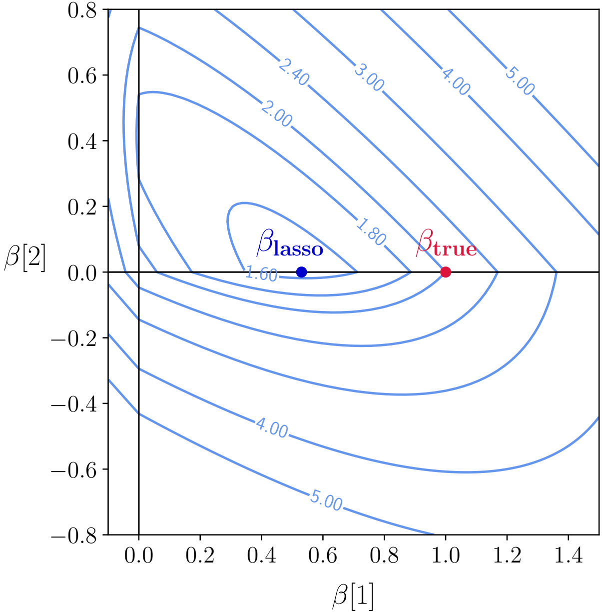


$||\beta||_1$

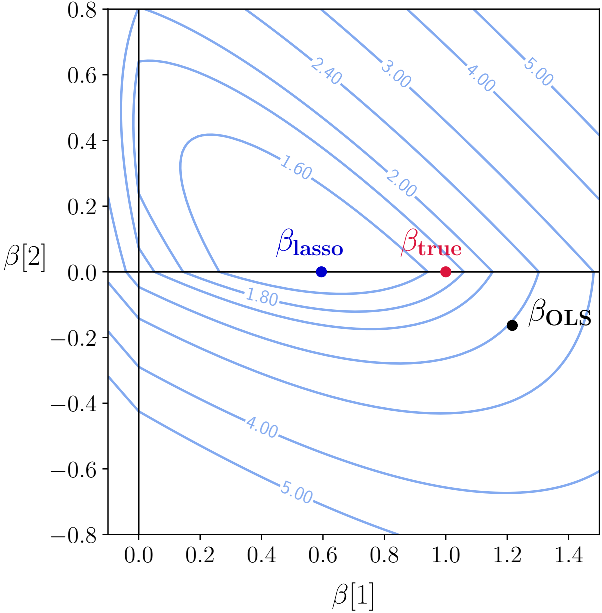


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
β[2]
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
β[1]
0.10
0.25
0.50
1.00
1.50
2.00
1.50
2.00

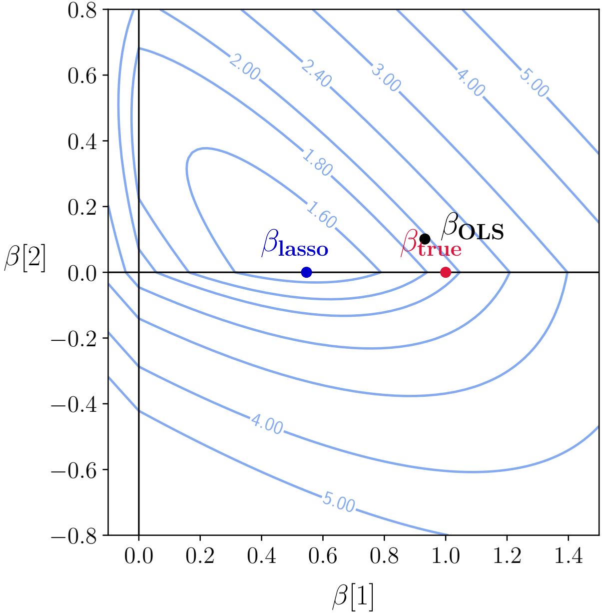
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \,||\beta||_1$$



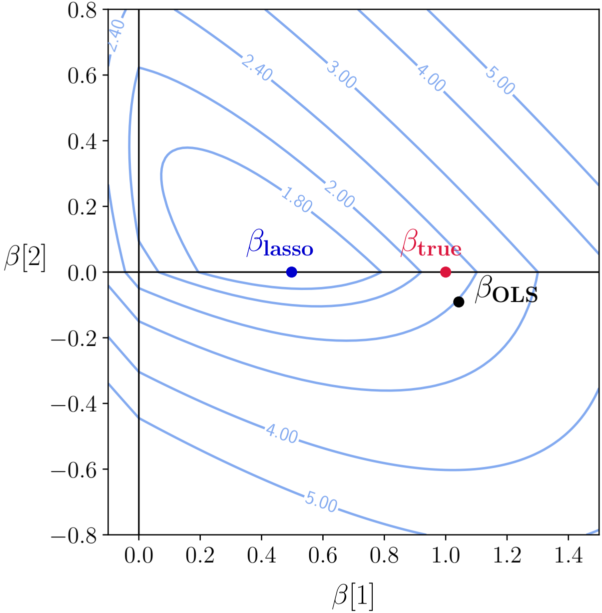
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \, ||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$



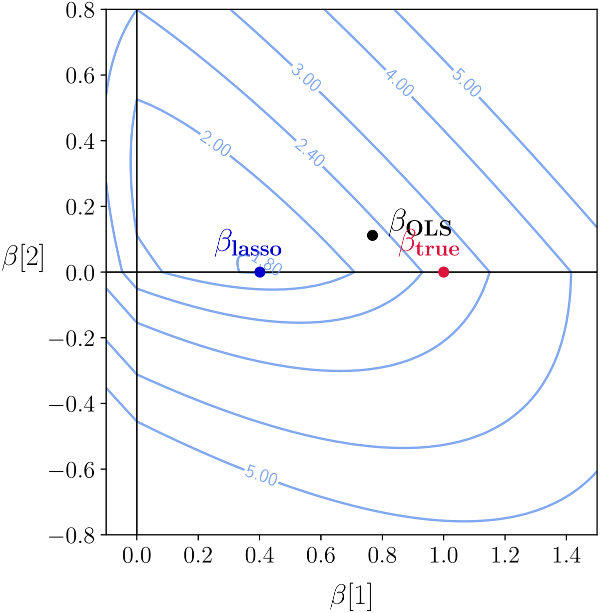
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \, ||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$



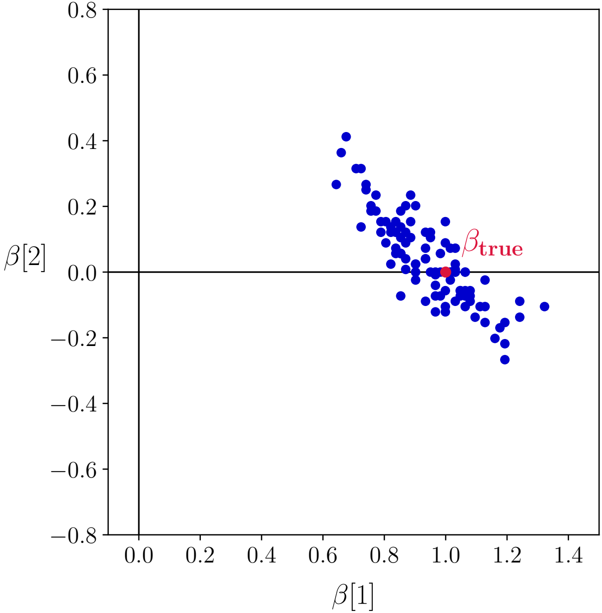
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \,||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta$$



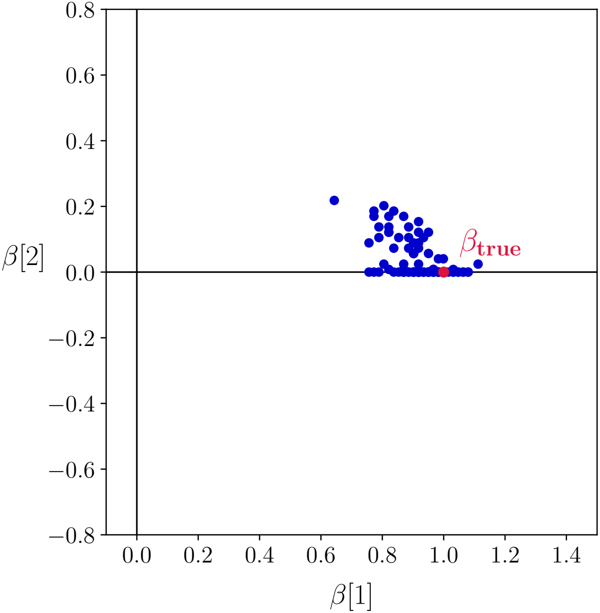
$$(\beta-\beta_{\text{true}})^T XX^T(\beta-\beta_{\text{true}})+\lambda\,||\beta||_1-2\tilde{z}_{\text{train}}^T X^T\beta$$



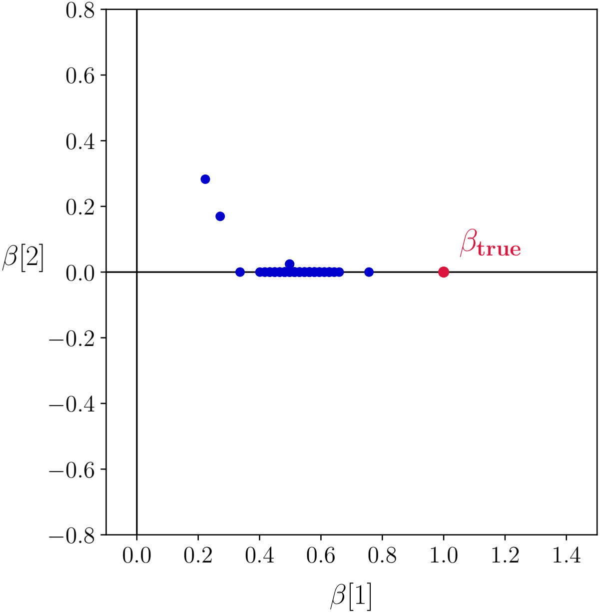
$\lambda = 0.02$



$\lambda = 0.2$

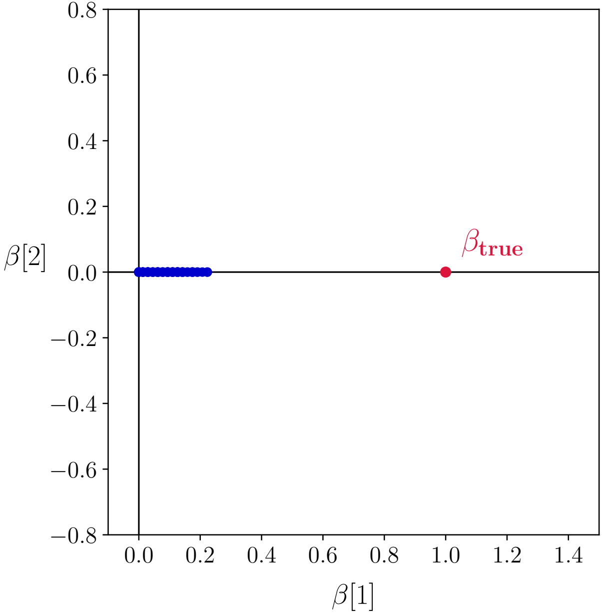


$\lambda = 2$



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

$\lambda = 4$



$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
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

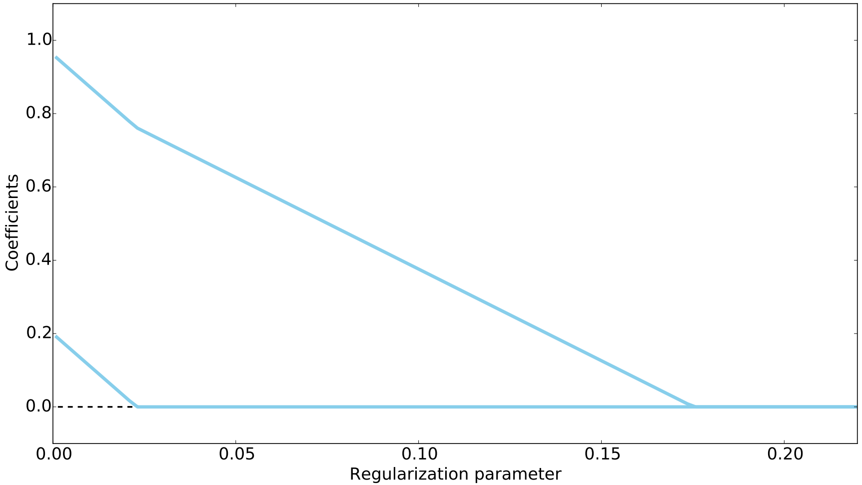
$$\frac{\left|x_{\text{other}}^T z - \rho x_{\text{true}}^T z\right|}{1 - |\rho|} \leq \lambda \leq 1 + x_{\text{true}}^T z$$

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 there is a horizontal supporting hyperplane at β_{lasso}

Equivalently, zero is 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 $|\gamma| \leq 1$

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