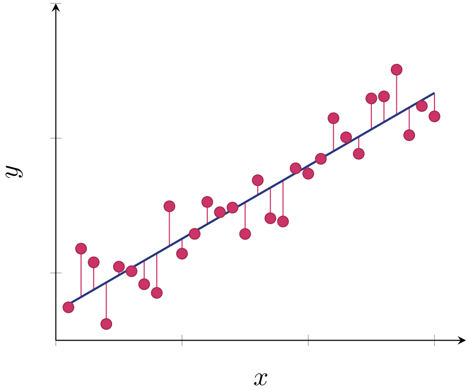
Regularization in Regression

Sibylle Hess

Technische Universiteit Eindhoven University of Technology

The Problem of Choosing the Right Regression Model

The Regression Optimization Problem



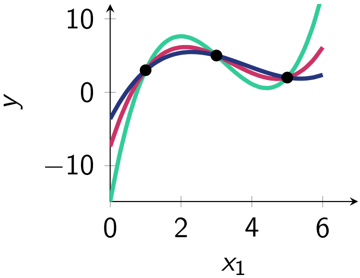
Regression Minimizers

The global minimizers of the regression problem are given by

$$\{\boldsymbol{\beta} \in \mathbb{R}^p \mid X^\top X\boldsymbol{\beta} = X^\top \mathbf{y}\}.$$

If the matrix $X^\top X$ is invertible, then there is only one minimizer:

$$\boldsymbol{\beta} = (X^\top X)^{-1} X^\top \mathbf{y}$$

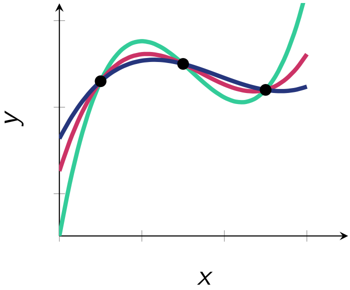


However, there also might be infinitely many local and global minimizers of $RSS(\boldsymbol{\beta})$.
Example: fit the function

$$f(x) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x^1 + \beta_0$$

to three observations

Toy Example: Regression with $p > n$



$$f(x) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0$$
$$= \phi(x)^\top \boldsymbol{\beta},$$

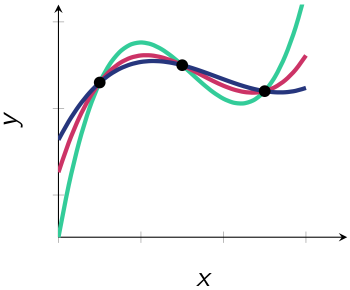
where $\phi(x)^\top = (1 \; x \; x^2 \; x^3)$

D	x_1	y
1	5	2
2	3	5
3	1	3

The design matrix is then given by

$$X = \begin{pmatrix} 1 & 5 & 25 & 125 \\ 1 & 3 & 9 & 27 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

Toy Example: Regression with $p > n$



$$X = \begin{pmatrix} 1 & 5 & 25 & 125 \\ 1 & 3 & 9 & 27 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

The global minimizers of the regression problem are given by

$$\{\beta \in \mathbb{R}^p \mid X^\top X\beta = X^\top \mathbf{y}\}.$$

However, the matrix $X^\top X$ is in this case not invertible.
How do we compute the global minimizers then?

Singular Value Decomposition

Theorem (SVD)

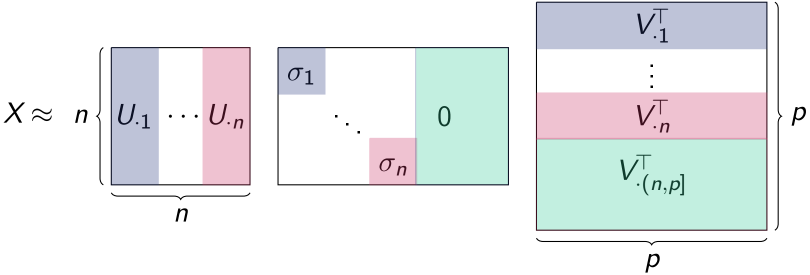
For every matrix $X \in \mathbb{R}^{n \times p}$ *there exist orthogonal matrices* $U \in \mathbb{R}^{n \times n}, V \in \mathbb{R}^{p \times p}$ *and* $\Sigma \in \mathbb{R}^{n \times p}$ *such that*

$$X = U \Sigma V^\top, \; \textit{where}$$

- $U^\top U = UU^\top = I_n, V^\top V = VV^\top = I_p$
- Σ *is a rectangular diagonal matrix,* $\Sigma_{11} \geq \ldots \geq \Sigma_{kk}$ *where* $k = \min\{n, p\}$

The column vectors $U_{\cdot s}$ and $V_{\cdot s}$ are called left and right singular vectors and the values $\sigma_i = \Sigma_{ii}$ are called singular values $(1 \leq i \leq l)$.

SVD Visualization for $p > n$



SVD Determines if a Matrix is Invertible

A $(n \times n)$ matrix $A = U\Sigma V^\top$ is invertible if all singular values are larger than zero. The inverse is given by

$$A^{-1} = V\Sigma^{-1}U^\top, \quad \text{where}$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \sigma_n \end{pmatrix}, \qquad \Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \frac{1}{\sigma_n} \end{pmatrix}$$

Using SVD to Obtain Solutions to the Regression Problem

The global minimizers $\boldsymbol{\beta}$ to the linear regression problem with design matrix X are given by

$$\{\boldsymbol{\beta} \in \mathbb{R}^p \mid X^\top X\boldsymbol{\beta} = X^\top \mathbf{y}\}.$$

Let $X = U\Sigma V^\top$ be the SVD of X, then we have

$$X^\top X\boldsymbol{\beta} = X^\top \mathbf{y} \quad \Leftrightarrow \quad \Sigma^\top \Sigma V^\top \boldsymbol{\beta} = \Sigma^\top U^\top \mathbf{y}$$

$\Sigma^\top\Sigma$ does not have an inverse if only $r < p$ singular values are nonzero.

Using SVD to Obtain Solutions to the Regression Problem

The global minimizers β to the linear regression problem with design matrix $X = U\Sigma V^\top$ are given by

$$\{\boldsymbol{\beta} \in \mathbb{R}^p \mid \Sigma^\top \Sigma V^\top \boldsymbol{\beta} = \Sigma^\top U^\top \mathbf{y}\}.$$

If only $r < p$ singular values are nonzero, we employ the pseudoinverse $(\Sigma^\top \Sigma)^+$ defined by

$$\Sigma^\top \Sigma = \left(\begin{array}{ccc|c} \sigma_1^2 & \dots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \dots & \sigma_r^2 & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right), \quad (\Sigma^\top \Sigma)^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1^2} & \dots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \dots & \frac{1}{\sigma_r^2} & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right)$$

The Set of all Regression Minimizers

If we have $r < p$ nonzero singular values, then we have infinitely many global optimizers

$$\beta = VA\Sigma^\top U^\top \mathbf{y}$$

where

$$A = \left(\begin{array}{ccc:c} \frac{1}{\sigma_1^2} & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \cdots & \frac{1}{\sigma_r^2} & \\ \hdashline & A_{r+1,1} \cdots & & A_{r+1,p} \\ & \vdots & & \vdots \\ & A_{p,\,1} \cdots & & A_{p,p} \end{array}\right) \in \mathbb{R}^{p\times p}$$

The Regression Minimizer by the Pseudo Inverse

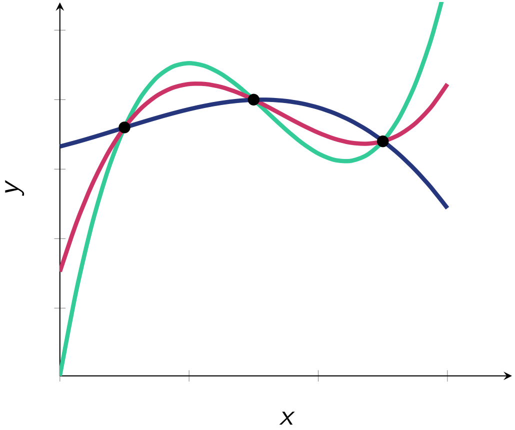
We define the regression solution derived by the pseudo inverse as

$$\beta_+ = V(\Sigma^\top \Sigma)^+ \Sigma^\top U^\top \mathbf{y}$$

where

$$(\Sigma^\top \Sigma)^+ = \left(\begin{array}{ccc:c} \frac{1}{\sigma_1^2} & \cdots & 0 & \\ \vdots & \ddots & \vdots & \mathbf{0} \\ 0 & \cdots & \frac{1}{\sigma_r^2} & \\ \hdashline & \mathbf{0} & & \mathbf{0} \end{array}\right) \in \mathbb{R}^{p \times p}$$

Toy Example: Regression for $p > n$



So, that's it, sometimes I just have to choose a regression function from infinitely many ones and roll with it?

Well, all regression minimizers are equal, but some minimizers are more equal than others.

Can I not just choose a more simple function class and circumvent the problem?

Not if $d > n$!

Example: Gene Expression Analysis

D	Gene 1	Gene 2	...	Gene 60,000	y: probability of survival
1	0.00	2.75		12.93	0.9
2	0.00	0.00		16.26	0.7
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
489	0.00	5.38		0.00	0.8

Even if we use a linear function class, we have a design matrix where $p = d = 60,000 \gg 489 = n$.

This introduces the problem of feature selection.

Feature Selection by Sparse Regression Vectors

The regression vector $\boldsymbol{\beta}$ encodes which features are relevant for prediction by nonnegative entries:

$$f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} = \sum_{i=1}^{p} \beta_i x_i = \sum_{i:\beta_k \neq 0} \beta_i x_i$$

The number of nonnegative entries is given by the L_0-'norm':

$$\|\boldsymbol{\beta}\|_0 = |\{i \mid \beta_i \neq 0\}|.$$

Be careful: The L_0-'norm' is not a real norm!

The Sparse Regression Task

Given a data matrix $D \in \mathbb{R}^{n \times d}$, a target vector $\mathbf{y} \in \mathbb{R}^n$, the design matrix $X \in \mathbb{R}^{n \times p}$, where $X_{i\cdot} = \phi(D_{i\cdot}^\top)^\top$ and the integer s.

Find the regression vector β, solving the following objective

$$\min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - X\beta\|^2 \qquad \text{s.t. } \|\beta\|_0 \leq s.$$

Return the predictor function $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \phi(\mathbf{x})^\top \beta$.

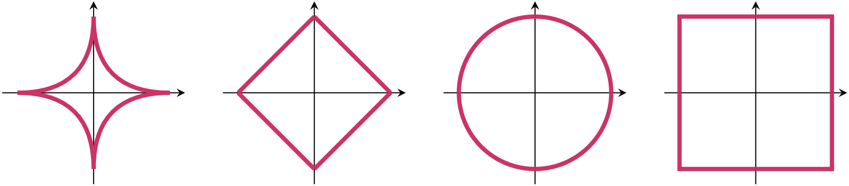
That's all nice and stuff, but how are we going to optimize that? The objective is not convex anymore.

Relax

The L_p-'norms'

The L_p-'norm' is defined for $p \in (0, \infty]$ as follows, and it is a real norm if $p \geq 1$:

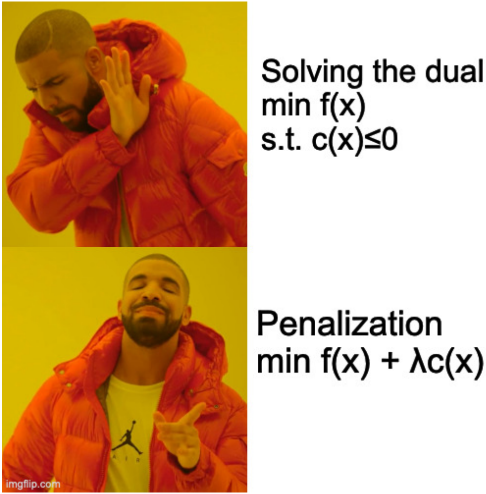
$$\|\mathbf{x}\|_p = \left(\sum_{k=1}^{d} |x_k|^p\right)^{1/p}$$



$p = \frac{1}{2}$ $p = 1$ $p = 2$ $p = \infty$

Okay, so we just take an L_p-norm for $p \geq 1$, then the sparse regression problem is convex.

But how do we optimize subject to the constraints?


Solving the dual
min f(x)
s.t. c(x)≤0
Penalization
min f(x) + λc(x)
imgflip.com

L_p-Norm Penalized Regression

L_p-Constrained Regression

Let $p \in [0, \infty]$, s>0, then the L_p-constrained regression is given as:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\beta\|^2 \qquad \text{s.t. } \|\beta\|_p \leq s$$

According to the theory of Lagrange multipliers, there exists a parameter $\lambda > 0$ such that the objective above is equivalent to

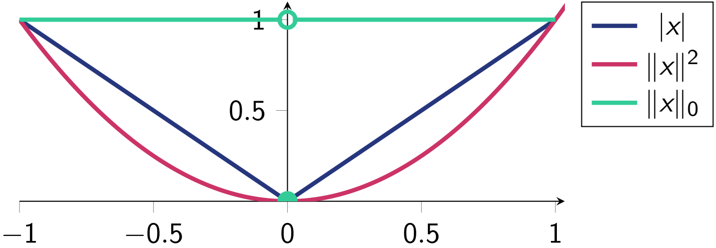
L_p-Penalized Regression

Let $p \in [0, \infty]$, $\lambda > 0$ then the L_p-penalized regression is given as:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\beta\|^2 + \lambda \|\beta\|_p$$

Analytical Properties of L_p-norms

norm	continuous	differentiable
$g(\mathbf{x}) = \|\mathbf{x}\|^2$	✓	✓
$g(\mathbf{x}) = \lvert\mathbf{x}\rvert$	✓	✗
$g(\mathbf{x}) = \|\mathbf{x}\|_0$	✗	✗



Let us start with a nice and smooth regularization term:

the squared L_2 norm.

Ridge Regression

Given a data matrix $D \in \mathbb{R}^{n \times d}$, a target vector $\mathbf{y} \in \mathbb{R}^n$, the design matrix $X \in \mathbb{R}^{n \times p}$, where $X_{i\cdot} = \phi(D_{i\cdot}^\top)^\top$ and a regularization weight $\lambda > 0$.

Find the regression vector $\boldsymbol{\beta}$, solving the following objective

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2.$$

Return the predictor function $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \phi(\mathbf{x})^\top \boldsymbol{\beta}$.

Minimizers of Ridge Regression

Ridge Regression Objective

$$\min_{\beta \in \mathbb{R}^p} RSS_{L_2}(\beta) = \|\mathbf{y} - X\beta\|^2 + \lambda\|\beta\|^2$$

The solution to ridge regression is given by the stationary points (RSS_{L_2} is convex as weighted sum of convex functions):

$$\begin{aligned} & \nabla_\beta RSS_{L_2}(\beta) = -2X^\top(\mathbf{y} - X\beta) + 2\lambda\beta = 0 \\ \Leftrightarrow \quad & (X^\top X + \lambda I)\beta = X^\top \mathbf{y} \end{aligned}$$

Is this now better?

Yes

Minimizers of Ridge Regression are Unique

The matrix $X^\top X + \lambda I$ is invertible for all $\lambda > 0$!
Let $X = U\Sigma V^\top$ be the singular value decomposition of X, then

$$X^\top X + \lambda I = V(\Sigma^\top \Sigma + \lambda I)V^\top$$

Hence, the uniquely defined global minimizer of Ridge Regression is given by

$$\beta_{L_2} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y}$$

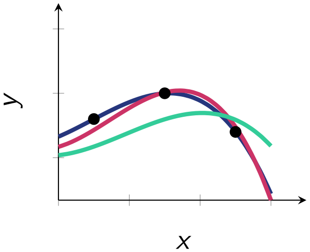
Ridge Regression and Regression Minimizers

Given the SVD of the design matrix $X = U\Sigma V^\top$, the ridge regression solution β_{L_2} with small regularization weight $\lambda > 0$ is similar to one of the global minimizers of regression:

$$\beta_{L_2} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{y} \quad = V(\Sigma^\top \Sigma + \lambda I)^{-1} \Sigma^\top U^\top \mathbf{y}$$
$$\approx VA\Sigma^\top U^\top \mathbf{y}$$

if $\lambda > 0$ is small and $A = \left(\begin{array}{ccc:ccc} \frac{1}{\sigma_1^2} & \cdots & 0 & & & \\ \vdots & \ddots & \vdots & & \mathbf{0} & \\ 0 & \cdots & \frac{1}{\sigma_r^2} & & & \\ \hdashline & & & \frac{1}{\lambda} & & \\ & \mathbf{0} & & & \ddots & \\ & & & & & \frac{1}{\lambda} \end{array}\right) \in \mathbb{R}^{p \times p}$

Toy Example: Ridge Regression for $p > n$



$$f(x_1) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x^1 + \beta_0$$

$$X = \begin{pmatrix} 1 & 5 & 25 & 125 \\ 1 & 3 & 9 & 27 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} 2 \\ 5 \\ 3 \end{pmatrix}$$

We obtain as a result for the regression parameters

$$\beta_+ = \begin{pmatrix} 1.6 \\ 1.2 \\ 0.3 \\ -0.1 \end{pmatrix}, \quad \beta_{L_2(\lambda=1)} = \begin{pmatrix} 0.8 \\ 0.8 \\ 0.7 \\ -0.2 \end{pmatrix}, \quad \beta_{L_2(\lambda=20)} = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.4 \\ -0.1 \end{pmatrix}$$

What if we choose a non-smooth regularization term?

Regularization with the L_1 norm.

Lasso Regression

Given a data matrix $D \in \mathbb{R}^{n \times d}$, a target vector $\mathbf{y} \in \mathbb{R}^n$, the design matrix $X \in \mathbb{R}^{n \times p}$, where $X_{i\cdot} = \phi(D_{i\cdot}^\top)^\top$ and a regularization weight $\lambda > 0$.

Find the regression vector $\boldsymbol{\beta}$, solving the following objective

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda |\boldsymbol{\beta}|.$$

Return the predictor function $f: \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \phi(\mathbf{x})^\top \boldsymbol{\beta}$.

Minimizers of Lasso

Lasso Objective

$$\min_{\beta \in \mathbb{R}^p} RSS_{L_1}(\beta) = \|\mathbf{y} - X\beta\|^2 + \lambda|\beta|$$

The L_1-norm has a subgradient:

$$\frac{\partial|\beta|}{\partial\beta_k} \in \begin{cases} \{1\}, & \text{if } \beta_k > 0 \\ \{-1\}, & \text{if } \beta_k < 0 \\ [-1, 1], & \text{if } \beta_k = 0 \end{cases}$$

Minimizers of objective functions which have a subgradient satisfy $\mathbf{0} \in \nabla f(\mathbf{x})$ (FONC for subgradients).
Gradients are a special case of subgradients.

How are we going to optimize the Lasso?

Solving for the stationary points of the subgradient $\nabla RSS_{L_1}(\beta) = 0$ is too complicated.

We could do subgradient descent but then we have to deal with step-sizes and additional difficulties of applying just the subgradient.

Luckily, the function is simple enough to derive the minimizers subject to one coordinate, enabling cooordinate descent.

Coordinate-Wise Minimizers of Lasso

The minimizer of Lasso subject to the coordinate β_k

$$\beta_k^* = \arg\min_{\beta_k \in \mathbb{R}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda|\boldsymbol{\beta}|$$

is given for $c_k = X_{\cdot k}^\top \mathbf{y} - \sum_{i \neq k} X_{\cdot k}^\top X_{\cdot i}\beta_i$ by

$$\beta_k^* = \begin{cases} \frac{1}{\|X_{\cdot k}\|^2}(c_k - \lambda) & \text{if } c_k > \lambda \\ \frac{1}{\|X_{\cdot k}\|^2}(c_k + \lambda) & \text{if } c_k < -\lambda \\ 0 & \text{if } -\lambda \leq c_k \leq \lambda. \end{cases}$$

FONC for subgradients $\mathbf{0} \in \frac{\partial}{\partial \beta_k} RSS_{L_1}$ yields the solutions to the coordinate-wise minimization problems.

Coordinate Descent for Lasso

function $\textsc{Lasso}(X, \lambda, \beta)$

  while not converged **do**

    for $k \in \{1, \ldots, p\}$ **do**

      $c_k \leftarrow X_{\cdot k}^\top \mathbf{y} - \sum_{i \neq k} X_{\cdot k}^\top X_{\cdot i} \beta_i$

      $$\beta_k \leftarrow \begin{cases} \frac{1}{\|X_{\cdot k}\|^2}(c_k - \lambda) & \text{if } c_k > \lambda \\ \frac{1}{\|X_{\cdot k}\|^2}(c_k + \lambda) & \text{if } c_k < -\lambda \\ 0 & \text{if } -\lambda \leq c_k \leq \lambda \end{cases}$$

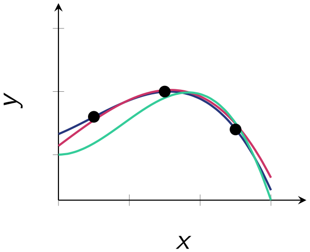
    end for

  end while

  return β

end function

Toy Example: Lasso for $p > n$



$$f(x) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x^1 + \beta_0$$

$$X = \begin{pmatrix} 1 & 5 & 25 & 125 \\ 1 & 3 & 9 & 27 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} 2 \\ 5 \\ 3 \end{pmatrix}$$

We obtain as a result for the regression parameters

$$\beta_+ = \begin{pmatrix} 1.6 \\ 1.2 \\ 0.3 \\ -0.1 \end{pmatrix}, \quad \beta_{L_1(\lambda=0.1)} = \begin{pmatrix} 0.7 \\ 2.1 \\ 0. \\ -0.07 \end{pmatrix}, \quad \beta_{L_1(\lambda=1)} = \begin{pmatrix} 0. \\ 0. \\ 1.1 \\ -0.2 \end{pmatrix}$$

Ok, but what is now better, L_1- or L_2-norm regularization?

L_1 vs. L_2 Regularization

The penalized Lasso and Ridge Regression objectives are equivalent to constrained optimization problems.

That is, for every $\lambda > 0$ there exists a radius $s > 0$ and vice versa, such that the following optimization problems are equivalent:

$$\min \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \qquad \text{s.t. } \boldsymbol{\beta} \in \mathbb{R}^p$$

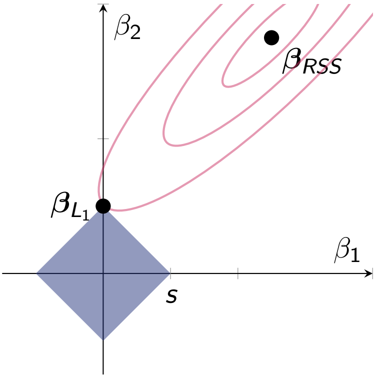
$$\min \|\mathbf{y} - X\boldsymbol{\beta}\|^2 \qquad \text{s.t. } \|\boldsymbol{\beta}\|^2 \leq s^2, \boldsymbol{\beta} \in \mathbb{R}^p$$

Similarly, for every $\lambda > 0$ there exists a radius $s > 0$ and vice versa, such that the following optimization problems are equivalent:

$$\min \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda|\boldsymbol{\beta}| \qquad \text{s.t. } \boldsymbol{\beta} \in \mathbb{R}^p$$

$$\min \|\mathbf{y} - X\boldsymbol{\beta}\|^2 \qquad \text{s.t. } |\boldsymbol{\beta}| \leq s, \boldsymbol{\beta} \in \mathbb{R}^p$$

$L1$-Regularization Tends to Sparser Solutions than $L2$



Summary L_1 vs. L_2 Regularization

Ridge Regression

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} RSS_{L_2}(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

Lasso

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} RSS_{L_1}(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\| + \lambda|\boldsymbol{\beta}|$$

1. The solution of Ridge Regression is computable very fast, analyically. The Ridge Regression minimizer is uniquely defined, but usually not sparse.
2. Lasso is optimized with coordinate descent, which is a theoretically well-founded optimization procedure. Lasso regression is more likely to return sparse regression vectors $\boldsymbol{\beta}$.