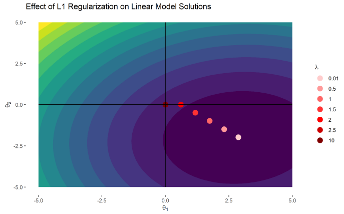
Introduction to Machine Learning

Regularization
Lasso Regression



Learning goals

- Lasso regression / $L1$ penalty
- Know that lasso selects features
- Support recovery

LASSO REGRESSION

Another shrinkage method is the so-called **lasso regression** (least absolute shrinkage and selection operator), which uses an $L1$ penalty on $\boldsymbol{\theta}$:

$$
\begin{aligned}
\hat{\theta}_{\text{lasso}} &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)} \right)^2 + \lambda \sum_{j=1}^{p} |\theta_j| \\
&= \arg\min_{\boldsymbol{\theta}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta} \right)^\top \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta} \right) + \lambda \|\boldsymbol{\theta}\|_1
\end{aligned}
$$

Optimization is much harder now. $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is still convex, but in general there is no analytical solution and it is non-differentiable.

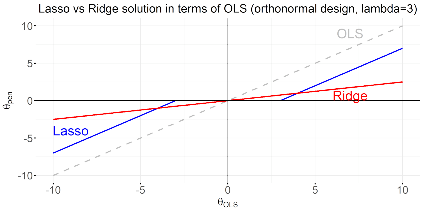
$L1$ AND $L2$ REG. WITH ORTHONORMAL DESIGN

For special case of orthonormal design $\mathbf{X}^\top\mathbf{X} = \boldsymbol{I}$ we can derive a closed-form solution in terms of $\hat{\theta}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{y}$:

$$\hat{\theta}_{\text{lasso}} = \text{sign}(\hat{\theta}_{\text{OLS}})(|\hat{\theta}_{\text{OLS}}| - \lambda)_+ \quad \text{(sparsity)}$$

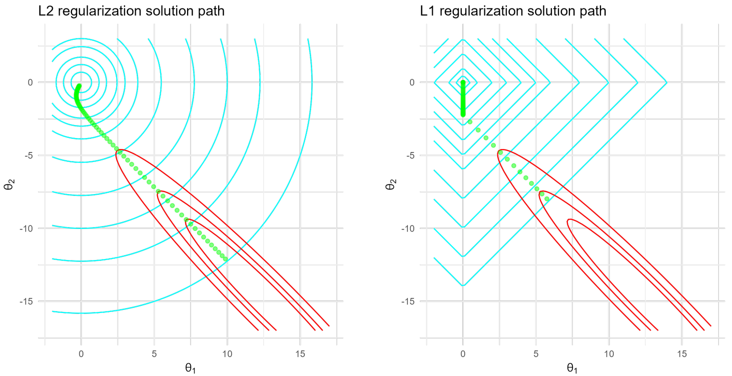
Function $S(\theta, \lambda) := \text{sign}(\theta)(|\theta| - \lambda)_+$ is called **soft thresholding** operator:
For $|\theta| \leq \lambda$ it returns 0, whereas params $|\theta| > \lambda$ are shrunken toward 0 by λ.
Comparing this to $\hat{\theta}_{\text{Ridge}}$ under orthonormal design:

$$\hat{\theta}_{\text{Ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y} = ((1+\lambda)\boldsymbol{I})^{-1}\hat{\theta}_{\text{OLS}} = \frac{\hat{\theta}_{\text{OLS}}}{1+\lambda} \quad \text{(no sparsity)}$$



COMPARING SOLUTION PATHS FOR $L1/L2$

- Ridge results in smooth solution path with non-sparse params
- Lasso induces sparsity, but only for large enough λ



SUPPORT RECOVERY OF LASSO ▸ Zhao and Yu 2006

When can lasso select true support of $\boldsymbol{\theta}$, i.e., only the non-zero parameters? Can be formalized as sign-consistency:

$$\mathbb{P}\big(\text{sign}(\hat{\theta}) = \text{sign}(\boldsymbol{\theta})\big) \to 1 \text{ as } n \to \infty \quad (\text{where } \text{sign}(0) := 0)$$

Suppose the true DGP given a partition into subvectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{X}_2\boldsymbol{\theta}_2 + \varepsilon \text{ with } \varepsilon \sim (0, \sigma^2 \boldsymbol{I})$$

and only $\boldsymbol{\theta}_1$ is non-zero. Let $\mathbf{X}_1$ denote the $n \times q$ matrix with the relevant features and $\mathbf{X}_2$ the matrix of noise features. It can be shown that $\hat{\theta}_{lasso}$ is sign consistent under an **irrepresentable condition**:

$$|(\mathbf{X}_2^\top \mathbf{X}_1)(\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \text{sign}(\boldsymbol{\theta}_1)| < \mathbf{1} \text{ (element-wise)}$$

In fact, lasso can only be sign-consistent if this condition holds.

Intuitively, the irrelevant variables in $\mathbf{X}_2$ must not be too correlated with (or *representable* by) the informative features ▸ Meinshausen and Yu 2009