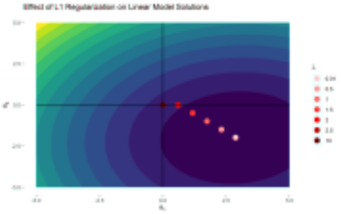
Introduction to Machine Learning

Regularization
Lasso Regression

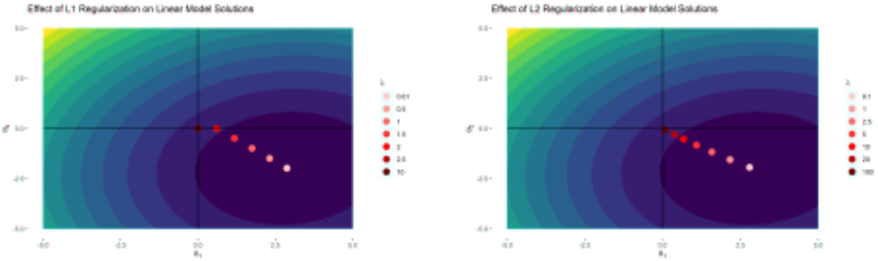


Learning goals

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

LASSO REGRESSION / 2

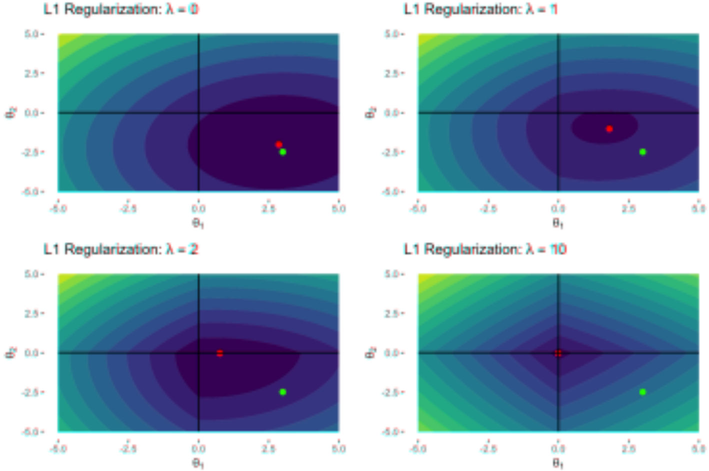
Let $y = 3x_1 - 2x_2 + \epsilon, \epsilon \sim N(0, 1)$. The true minimizer is $\theta^* = (3, -2)^T$. LHS = $L1$ regularization; RHS = $L2$



With increasing regularization, $\hat{\theta}_{lasso}$ is pulled back to the origin, but takes a different “route”. θ_2 eventually becomes 0!

LASSO REGRESSION / 3

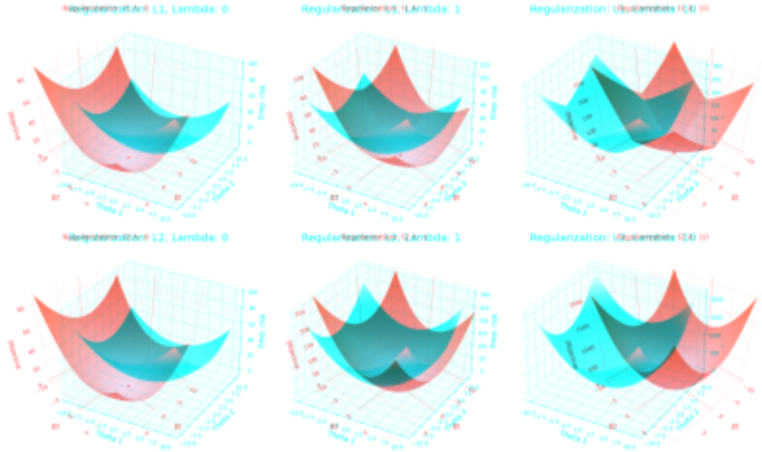
Contours of regularized objective for different λ values.



Green = true minimizer of the unreg.objective and red = lasso solution.

LASSO REGRESSION / 4

Regularized empirical risk $\mathcal{R}_{\text{reg}}(\theta_1, \theta_2)$ using squared loss for $\lambda \uparrow$. $L1$ penalty makes non-smooth kinks at coordinate axes more pronounced, while $L2$ penalty warps $\mathcal{R}_{\text{reg}}$ toward a "basin" (elliptic paraboloid).

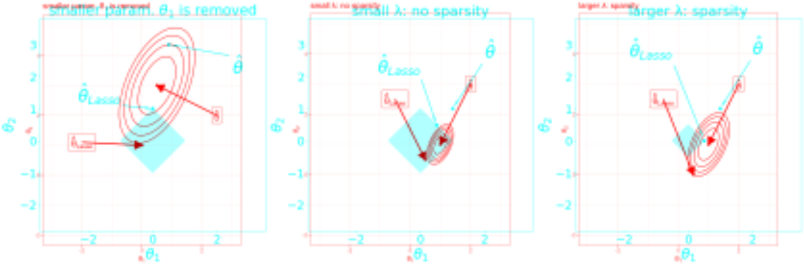


LASSO REGRESSION / 5

We can also rewrite this as a constrained optimization problem. The penalty results in the constrained region to look like a diamond shape.

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 \text{ subject to: } \|\boldsymbol{\theta}\|_1 \leq t$$

The kinks in $L1$ enforce sparse solutions because “the loss contours first hit the sharp corners of the constraint” at coordinate axes where (some) entries are zero.



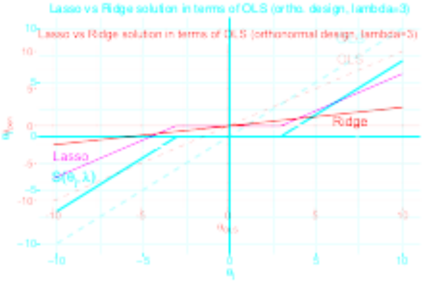
$L1$ AND $L2$ REG. WITH ORTHONORMAL DESIGN

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