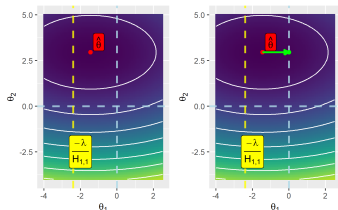


Introduction to Machine Learning

Regularization

Geometry of L1 Regularization



Learning goals

- Approximate transformation of unregularized minimizer to regularized
- Soft-Thresholding

L1-REGULARIZATION

- The L1-regularized risk of a model $f(\mathbf{x} | \boldsymbol{\theta})$ is

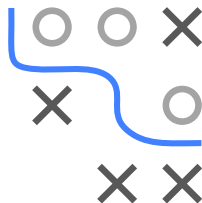
$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \sum_j \lambda |\theta_j|$$

and the (sub-)gradient is:

$$\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \text{sign}(\boldsymbol{\theta})$$

- Unlike in L_2 , contribution to grad. doesn't scale with θ_j elements.
- Again: quadratic Taylor approximation of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ around its minimizer $\hat{\boldsymbol{\theta}}$, then regularize:

$$\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \mathbf{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \sum_j \lambda |\theta_j|$$



L1-REGULARIZATION / 2

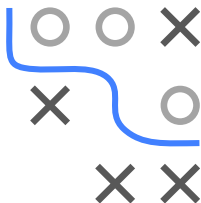
- To cheat and simplify, we assume the \mathbf{H} is diagonal, with $H_{j,j} \geq 0$
- Now $\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta})$ decomposes into sum over params θ_j (separable!):

$$\tilde{\mathcal{R}}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\hat{\boldsymbol{\theta}}) + \sum_j \left[\frac{1}{2} H_{j,j} (\theta_j - \hat{\theta}_j)^2 \right] + \sum_j \lambda |\theta_j|$$

- We can minimize analytically:

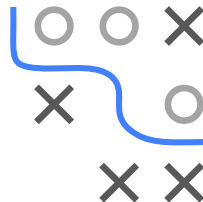
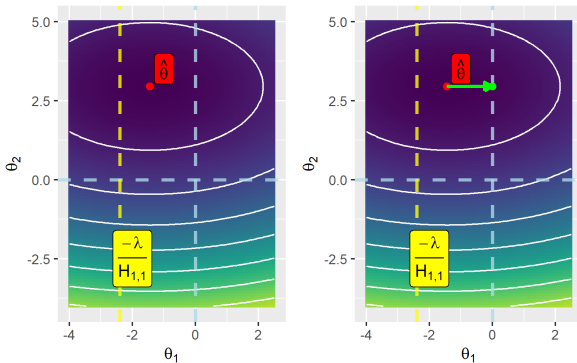
$$\begin{aligned} \hat{\theta}_{\text{lasso},j} &= \text{sign}(\hat{\theta}_j) \max \left\{ |\hat{\theta}_j| - \frac{\lambda}{H_{j,j}}, 0 \right\} \\ &= \begin{cases} \hat{\theta}_j + \frac{\lambda}{H_{j,j}} & , \text{ if } \hat{\theta}_j < -\frac{\lambda}{H_{j,j}} \\ 0 & , \text{ if } \hat{\theta}_j \in \left[-\frac{\lambda}{H_{j,j}}, \frac{\lambda}{H_{j,j}} \right] \\ \hat{\theta}_j - \frac{\lambda}{H_{j,j}} & , \text{ if } \hat{\theta}_j > \frac{\lambda}{H_{j,j}} \end{cases} \end{aligned}$$

- Shows how lasso (approx) transforms the normal minimizer
- If $H_{j,j} = 0$ exactly, $\hat{\theta}_{\text{lasso},j} = 0$



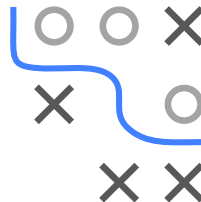
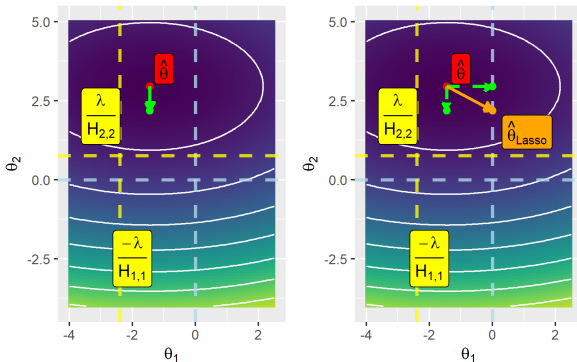
L1-REGULARIZATION / 3

- If $0 < \hat{\theta}_j \leq \frac{\lambda}{H_{j,j}}$ or $0 > \hat{\theta}_j \geq -\frac{\lambda}{H_{j,j}}$, the optimal value of θ_j (for the regularized risk) is 0 because the contribution of $\mathcal{R}_{\text{emp}}(\theta)$ to $\mathcal{R}_{\text{reg}}(\theta)$ is overwhelmed by the L1 penalty, which forces it to be 0.



L1-REGULARIZATION / 4

- If $0 < \frac{\lambda}{H_{j,j}} < \hat{\theta}_j$ or $0 > -\frac{\lambda}{H_{j,j}} > \hat{\theta}_j$, the L1 penalty shifts the optimal value of θ_j toward 0 by the amount $\frac{\lambda}{H_{j,j}}$.



- Yellow dotted lines are limits from soft-thresholding
- Therefore, the L1 penalty induces sparsity in the parameter vector.