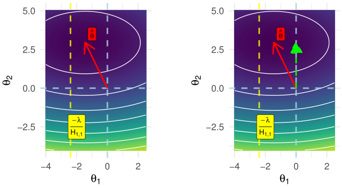
Introduction to Machine Learning

Geometric Analysis of L1-regularization



Learning goals

- Have a geometric understanding of L1-regularization
- Understand geometrically how L1-regularization induces sparsity

L1-REGULARIZATION

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

$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1$$

  and the (sub-)gradient is:

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

- Note that, unlike in the case of L2, the contribution of the L1 penalty to the gradient doesn't scale linearly with each θ_j.
- Let us now make (again) a quadratic Taylor approximation of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ around its minimizer $\hat{\boldsymbol{\theta}}$. To get a clean algebraic expression, we further assume the Hessian of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ is diagonal, i.e. $\boldsymbol{H} = \text{diag}([H_{1,1}, \ldots, H_{d,d}])$, where each $H_{j,j} \geq 0$.
- This assumption holds, for example, if the input features for a linear regression task have been decorrelated using PCA.

L1-REGULARIZATION

- If we plug this approximation into $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$, the result nicely decomposes into a sum over the parameters:

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

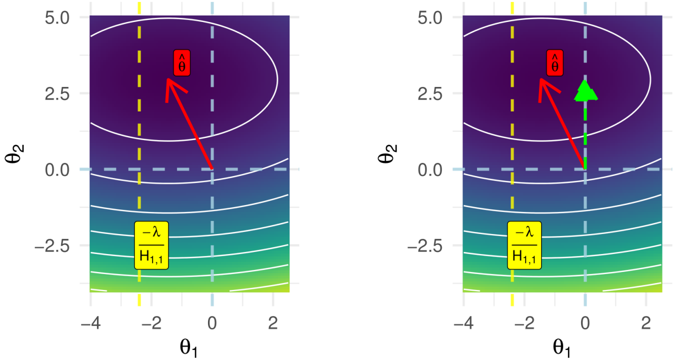
- We can minimize analyically:

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

- If $H_{j,j} = 0$ exactly, $\hat{\boldsymbol{\theta}}_{\text{Lasso},j} = 0$.

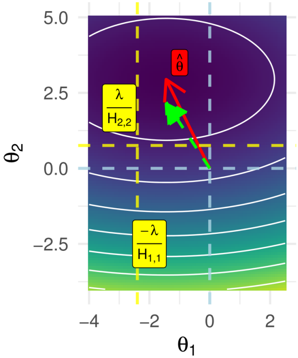
L1-REGULARIZATION

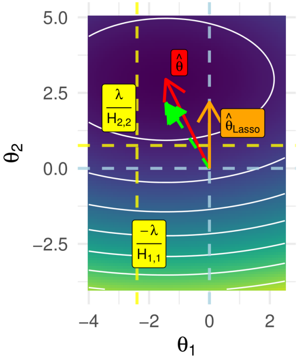
- If $0 < \hat{\theta}_j \leq \frac{\lambda}{H_{j,j}}$ or $0 > \hat{\theta}_j \geq -\frac{\lambda}{H_{j,j}}$, the optimal value of θ_j (for the regularized risk) is 0 because the contribution of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ to $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ is overwhelmed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



L1-REGULARIZATION

- 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 $\frac{\lambda}{H_{j,j}}$.





- Therefore, the L1 penalty induces sparsity in the parameter vector.