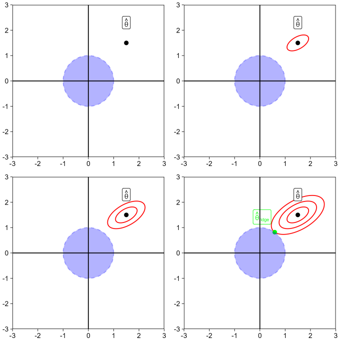
Introduction to Machine Learning

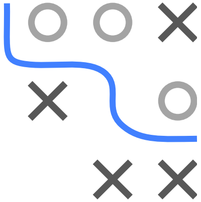
Regularization
Ridge Regression



Learning goals

- Regularized linear model
- Ridge regression / *L*2 penalty
- Understand parameter shrinkage
- Understand correspondence to constrained optimization

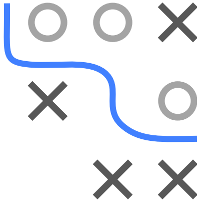
REGULARIZATION IN LM



- Can also overfit if p large and n small(er)
- OLS estimator requires full-rank design matrix
- For highly correlated features, OLS becomes sensitive to random errors in response, results in large variance in fit
- We now add a complexity penalty to the loss:

$$\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)} \right)^2 + \lambda \cdot J(\boldsymbol{\theta}).$$

RIDGE REGRESSION / L2 PENALTY



Intuitive measure of model complexity is deviation from 0-origin. So we measure $J(\boldsymbol{\theta})$ through a vector norm, shrinking coeffs closer to 0.

$$\begin{aligned}\hat{\theta}_{\text{ridge}} &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)}\right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \\ &= \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2\end{aligned}$$

Can still analytically solve this:

$$\hat{\theta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Name: We add pos. entries along the diagonal "ridge" of $\mathbf{X}^T\mathbf{X}$

EXAMPLE: POLYNOMIAL RIDGE REGRESSION

Consider $y = f(x) + \epsilon$ where the true (unknown) function is $f(x) = 5 + 2x + 10x^2 - 2x^3$ (in red).

Let's use a dth-order polynomial

$$f(x) = \theta_0 + \theta_1 x + \cdots + \theta_d x^d = \sum_{j=0}^{d} \theta_j x^j.$$

Using model complexity $d = 10$ overfits: