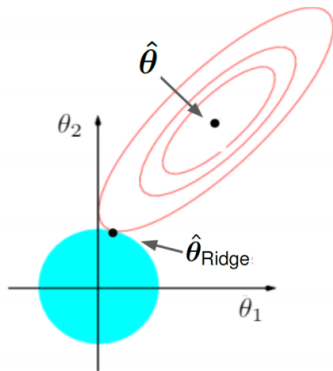


# Introduction to Machine Learning

## Lasso and Ridge Regression



Learning goals

- XXX
- XXX

# REGULARIZATION IN THE LINEAR MODEL

- Linear models can also overfit if we operate in a high-dimensional space with not that many observations.
- OLS usually require a full-rank design matrix.
- When features are highly correlated, the least-squares estimate becomes highly sensitive to random errors in the observed response, producing a large variance in the fit.
- We now add a complexity penalty to the loss:

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

- Intuitive to measure model complexity as deviation from the 0-origin, as the 0-model is empty and contains no effects. Models close to this either have few active features or only weak effects.
- So we measure  $J(\boldsymbol{\theta})$  through a vector norm. This shrinks coefficients closer 0, hence the term **shrinkage methods**.

# RIDGE REGRESSION

**Ridge regression** uses a simple  $L_2$  penalty:

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

Optimization is possible (as in the normal LM) in analytical form:

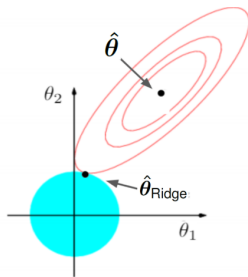
$$\hat{\boldsymbol{\theta}}_{\text{Ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Name comes from the fact that we add positive entries along the diagonal "ridge"  $\mathbf{X}^T \mathbf{X}$ .

# RIDGE REGRESSION

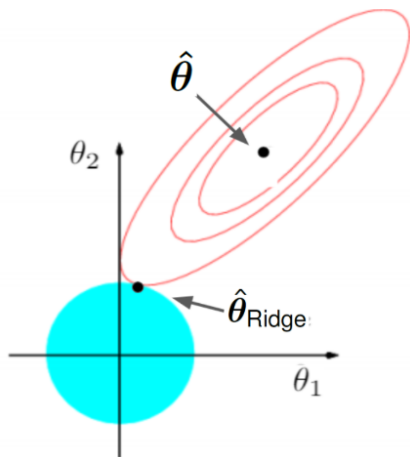
We understand the geometry of these 2 mixed components in our regularized risk objective much better, if we formulate the optimization as a constrained problem (see this a Lagrange multipliers in reverse).

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{i=1}^n \left( y^{(i)} - f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}) \right)^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_2^2 \leq t \end{aligned}$$



NB: Relationship between  $\lambda$  and  $t$  will be explained later.

# RIDGE REGRESSION



- We still optimize the  $\mathcal{R}_{\text{emp}}(\theta)$ , but cannot leave a ball around the origin.
- $\mathcal{R}_{\text{emp}}(\theta)$  grows monotonically if we move away from  $\hat{\theta}$ .
- Inside constraints perspective: From origin, jump from contour line to contour line (better) until you become infeasible, stop before.
- Outside constraints perspective: From  $\hat{\theta}$ , jump from contour line to contour line (worse) until you become feasible, stop then.
- So our new optimum will lie on the boundary of that ball.

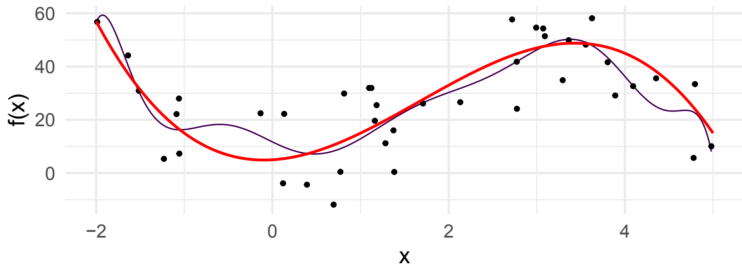
# EXAMPLE: POLYNOMIAL RIDGE REGRESSION

True (unknown) function is  $f(x) = 5 + 2x + 10x^2 - 2x^3 + \epsilon$  (in red).

Let us consider a  $d$ th-order polynomial

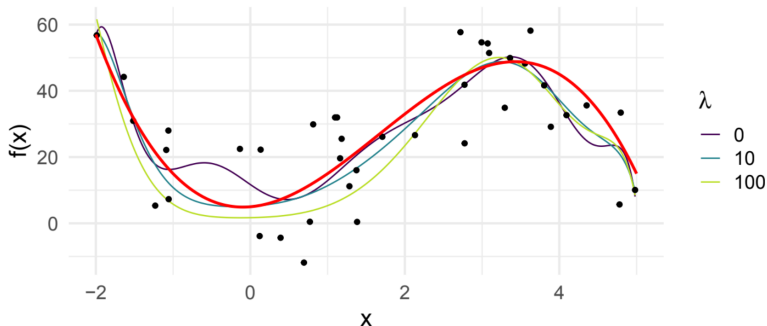
$$f(x) = \theta_0 + \theta_1 x + \dots + \theta_d x^d = \sum_{j=0}^d \theta_j x^j.$$

Using model complexity  $d = 10$  overfits:



# EXAMPLE: POLYNOMIAL RIDGE REGRESSION

With an L2 penalty we can now select "d too large" but regularize our model by shrinking its coefficients. Otherwise we have to optimize over the discrete  $d$ .



$\lambda$	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$	$\beta_7$	$\beta_8$	$\beta_9$	$\beta_{10}$
0.00	12.00	-16.00	4.80	23.00	-5.40	-9.30	4.20	0.53	-0.63	0.13	-0.01
10.00	5.20	1.30	3.70	0.69	1.90	-2.00	0.47	0.20	-0.14	0.03	-0.00
100.00	1.70	0.46	1.80	0.25	1.80	-0.94	0.34	-0.01	-0.06	0.02	-0.00

# LASSO REGRESSION

Another shrinkage method is the so-called **Lasso regression**, which uses an  $L_1$  penalty on  $\theta$ :

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

Note that optimization now becomes much harder.  $\mathcal{R}_{\text{reg}}(\theta)$  is still convex, but we have moved from an optimization problem with an analytical solution towards a non-differentiable problem.

Name: least absolute shrinkage and selection operator.



# LASSO REGRESSION

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

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

