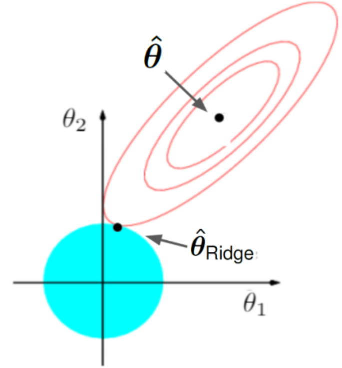
Introduction to Machine Learning

Lasso and Ridge Regression



Learning goals

- Know the regularized linear model
- Know Ridge regression (L2 penality)
- Know Lasso regression (L1 penality)

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 $J(\theta)$ through a vector norm. This shrinks coefficients closer 0, hence the term **shrinkage methods**.

RIDGE REGRESSION

Ridge regression uses a simple L_2 penalty:

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

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

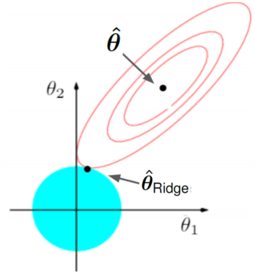
$$\hat{\boldsymbol{\theta}}_{\text{Ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y}$$

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

RIDGE REGRESSION

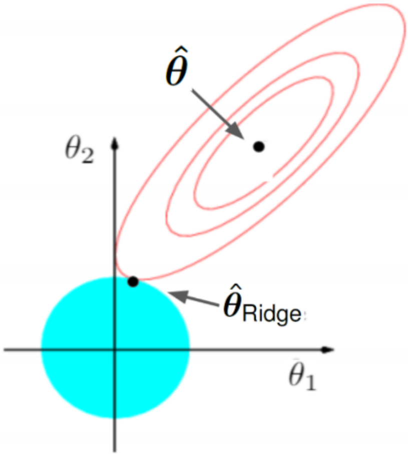
We understand the geometry of these 2 mixed components in our regularized risk objective much better, if we formulate the optimization as a constrained problem (see this a Lagrange multipliers in reverse).

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



NB: Relationship between λ and t will be explained later.

RIDGE REGRESSION



- We still optimize the $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$, but cannot leave a ball around the origin.
- $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ grows monotonically if we move away from $\hat{\boldsymbol{\theta}}$.
- Inside constraints perspective: From origin, jump from contour line to contour line (better) until you become infeasible, stop before.
- Outside constraints perspective: From $\hat{\boldsymbol{\theta}}$, 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.

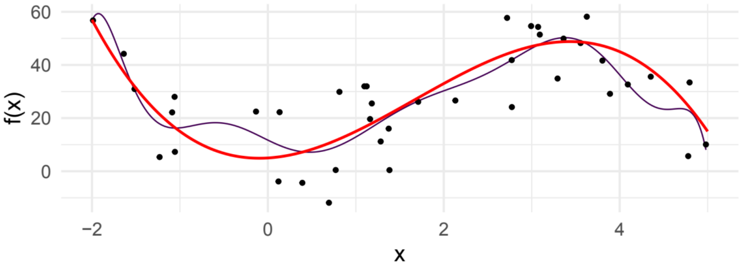
EXAMPLE: POLYNOMIAL RIDGE REGRESSION

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

Let us consider a dth-order polynomial

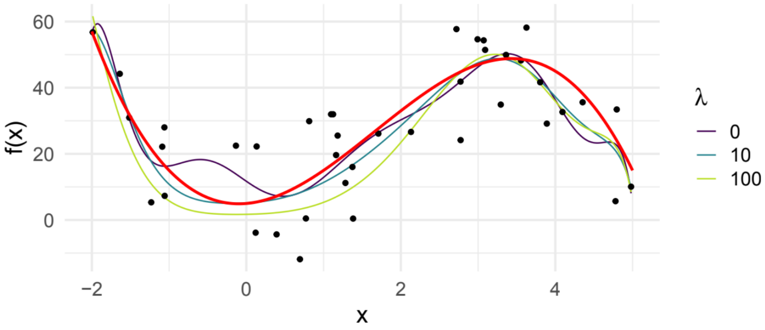
$$f(x) = \theta_0 + \theta_1 x + \cdots + \theta_d x^d = \sum_{j=0}^{d} \theta_j x^j.$$

Using model complexity $d = 10$ overfits:



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*.



λ	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{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 $\boldsymbol{\theta}$:

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

Note that optimization now becomes much harder. $\mathcal{R}_{\text{reg}}(\boldsymbol{\theta})$ 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.

$$\min_{\boldsymbol{\theta}} \quad \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2$$

$$\text{subject to:} \quad \|\boldsymbol{\theta}\|_1 \leq t$$