Ridge Regression

LS loss fn ① + (A) + ℓ2 penalized mean loss (L)

find $w^* = \underset{\vec{w}}{\operatorname{argmin}} \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}'\|^2 = J(w)$

regularization / penalty term

don't count bias term

- X has fictitious dimension
- don't penalize bias term

Why regularization?

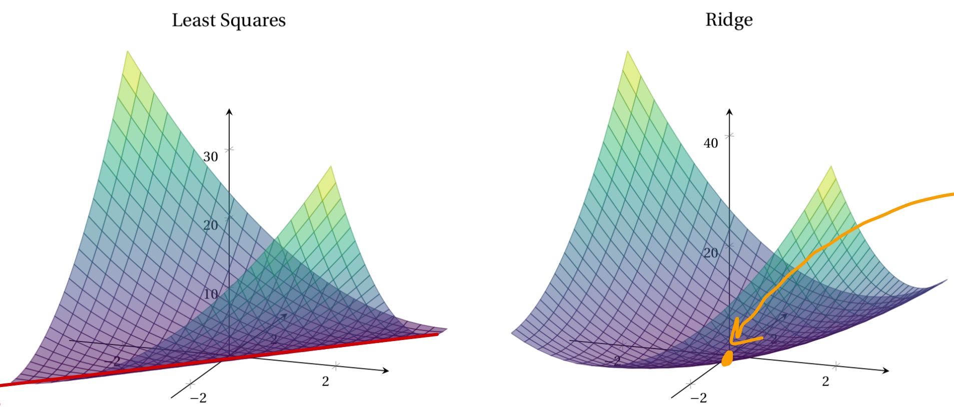
Shrinkage: make weights small

Why shrinkage?

- guarantees PD normal equations
  - ↳ always unique solution

e.g. when $d > n$,

unstable



Unique minimum

Figure 13.1: Least Squares Linear Regression vs. Ridge Regression Costs

"ill-posed" problem: not just 1 solution

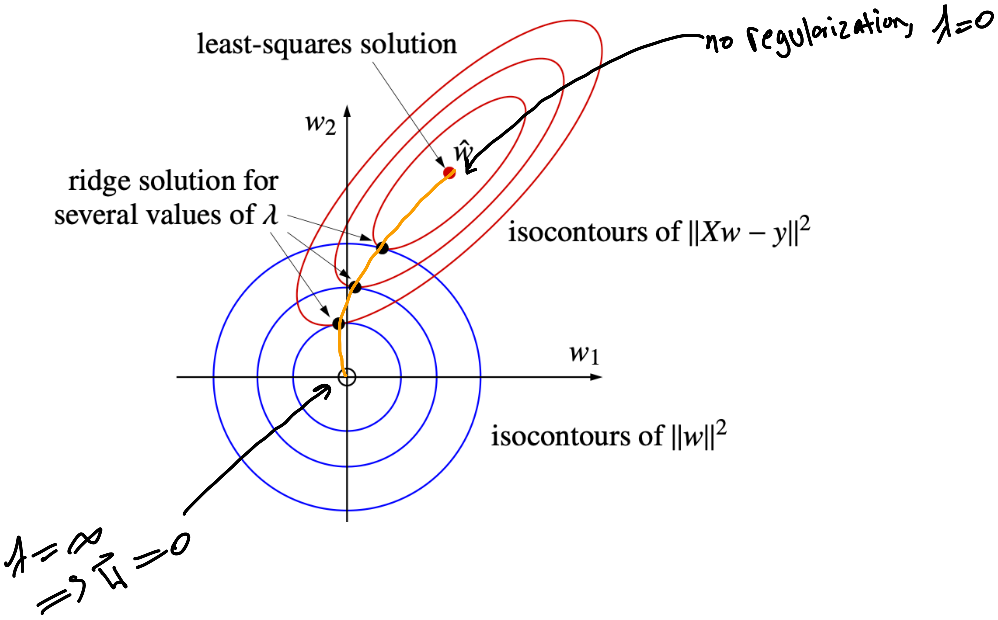
- regularization reduces overfitting
  - ↳ decreases variance

Why does this reduce variance?

Imagine: $500x_1 - 500x_2$ is line of best fit for well-separated pts, $y_i \in [0, 1]$

↳ large coeffs → instability
  - ↳ small changes in X → big change in y

↳ Solution: penalize large weights



Validation to find best λ

Set $\nabla J(w)$ to 0: last diag. entry is 0

$$(X^TX + \lambda I')\vec{w} = X^T y$$

$$\rightarrow \vec{w} = (X^TX + \lambda I')^{-1} X^T \vec{y}$$

$$\rightarrow h(\vec{z}) = \vec{w}^T \vec{z}$$

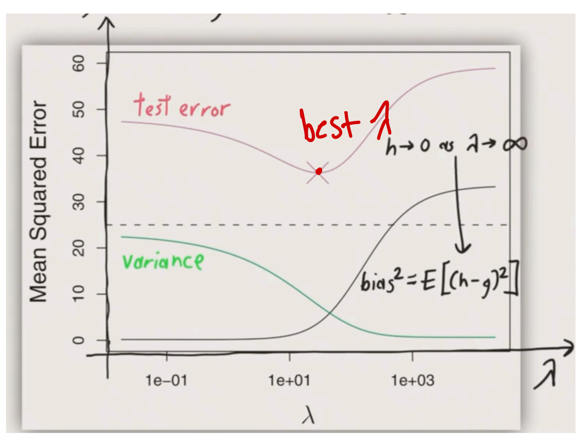
Increasing λ $\rightarrow$ more regularization $\rightarrow$ smaller $\|\vec{w}\|_2^2$ & vice versa

recall data model: $\vec{y} = X\vec{v} + \vec{\varepsilon}$, $\vec{\varepsilon}$ is noise

$$\mathrm{Var}(\text{ridge regr}) = \mathrm{Var}\left(\vec{z}^T (X^TX + \lambda I')^{-1} X^T \vec{e}\right)$$

$\lambda\uparrow$, Var$\downarrow$

As $\lambda \rightarrow \infty$, Var$(\cdot) \rightarrow 0$, bias $\uparrow$



- tune λ w/ validation or CV

- Ideally, features should be normalized to have same variance

alternatives: use asymmetric penalty by replacing I' w/ another diagonal matrix

*In polynomial regression, different weights need to be penalized differently

Bayesian Justification for Ridge Regression

- Assign a prior probability on w': $w' \sim N(0, \sigma^2)$

↳ true weights come from this distr.

- apply MLE to maximize posterior probability

Baye's thm: posterior $f(w \mid X, y) = \dfrac{f(y \mid X, w) \cdot \text{prior } f(w')}{f(y \mid X)} = \dfrac{\mathcal{L}(w) f(w')}{f(y \mid x)}$

w, y RVs, X is a constant

→ maximize log posterior: $\ln \mathcal{L}(w) + \ln f(w') - \text{constant}$

$$= -\text{const} \|X\vec{w} - \vec{y}\|_2^2 - \text{constant} \|w'\|^2$$

→ minimize $\|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|^2$

Feature Subset Selection

recall: all features increase variance, not all features reduce bias

Idea: identify which features we can get rid of

↳ set their weights to 0

↳ less overfitting, smaller test error

↳ Inference. Simpler models convey interpretable wisdom

- Useful in all classification & regression methods.
- Sometimes it's hard: different features can be redundant

Algo: Best Subset Selection. 1) Try all $2^d - 1$ nonempty subsets of features

2) choose best classifier w/ validation

- slow

Heuristics: 1) forward stepwise selection

- start w/ null model (0 features)
- repeatedly add best feature until validation errors start increasing (due to overfitting) instead of decreasing
- at each outer iteration, inner loop tries every feature & chooses best validation. Requires d^2 models trained (better than $O(2^d)$)

doesn't always pick best features

e.g. won't find best model w/ 2 features if neither feature yields best 1D model

Heuristic 2) backward step selection

best choice if most features are helpful

- start w/ all d features
- repeatedly remove features whose removal reduces validation error
- also $O(d^2)$

Lasso (Robert Tibshirani) "least absolute shrinkage & selection operator"

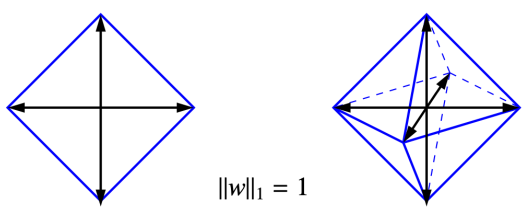
- regression w/ L1 penalty

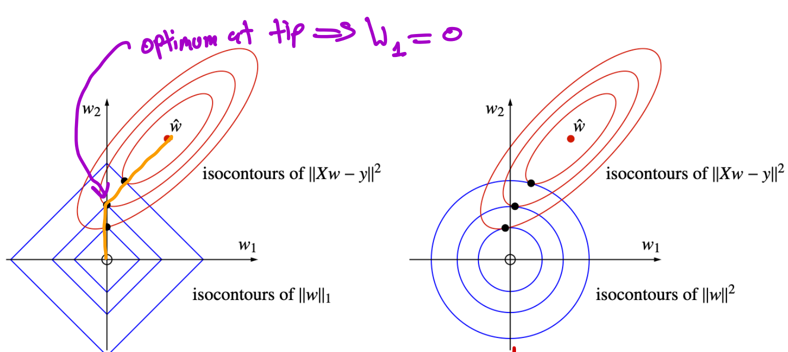
find $\vec{w}$ that minimizes: $J(w) = \|X\vec{w} - \vec{y}\|^2 - \lambda \|\vec{w}\|_1$

recall ridge regression: isosurfaces of $\|\vec{w}\|^2$ are hyperspheres

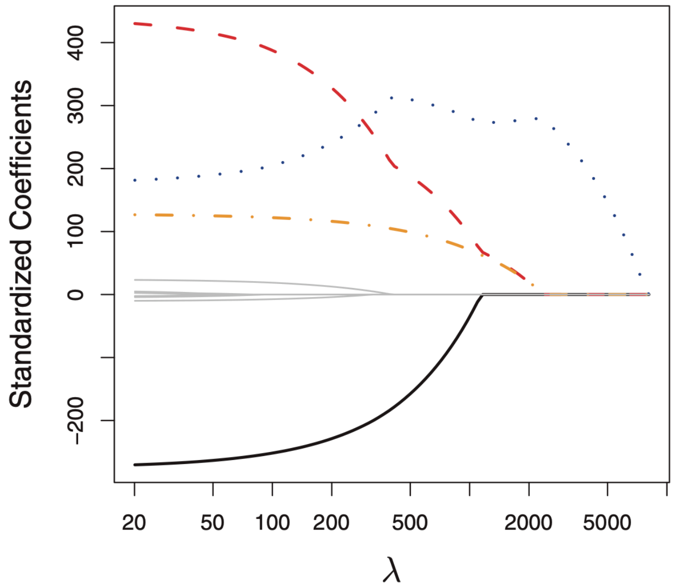
- those of $\ell 1$ are cross-polytopes
- unit cross-polytope is convex hull of unit coordinate vectors:

$3d \to 2^d$ sides





Lasso Ridge


Standardized Coefficients
400
300
200
100
0
−200
20
50
100
200
500
2000
5000
λ