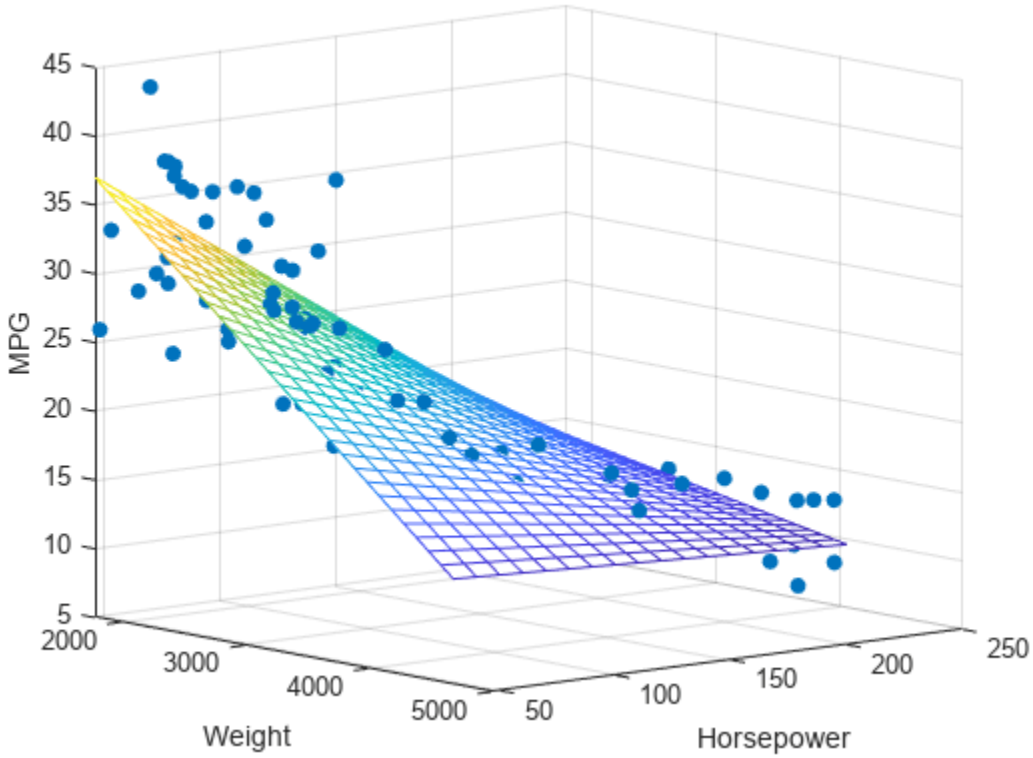
Linear Regression

Announcements

- HW 1 due today at 11:59 PM
- HW 2 posted today
- Python tutorial Friday evening – see announcement on Canvas

Regression

- Predict a continuous response variable y from a multi-dimensional feature vector $\boldsymbol{x}$



Boston Housing Dataset

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's

x

y

MSE

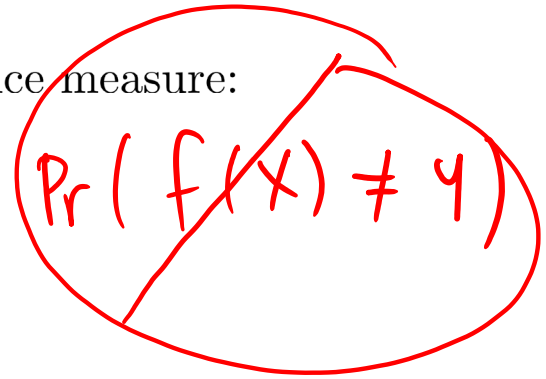
- Probabilistic setting: We have jointly distributed variables $(\boldsymbol{X}, Y)$ where

$$\boldsymbol{X} \in \mathbb{R}^d, \quad Y \in \mathbb{R}$$

and the goal is to predict Y from $\boldsymbol{X}$ using a function

$$f : \mathbb{R}^d \to \mathbb{R}.$$

- Performance measure:



mean squared error

$$R(f) = \mathbb{E}_{X,Y}\left[(Y - f(X))^2 \right]$$

Linear Regression

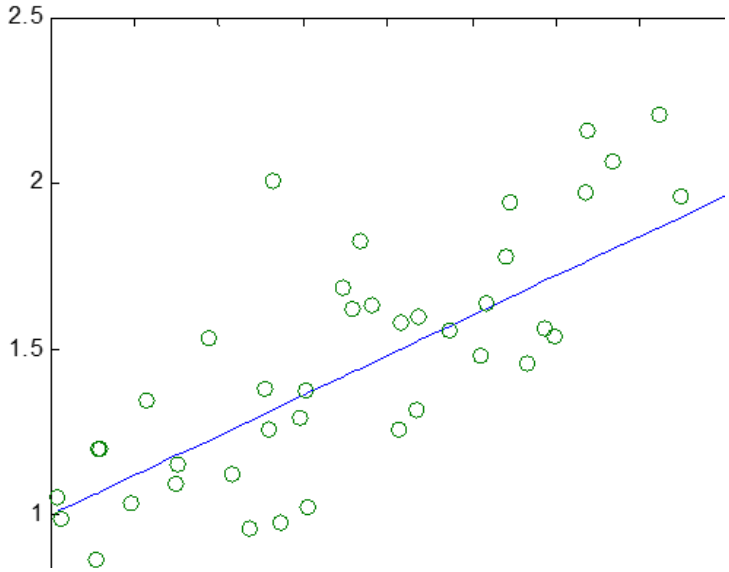
- In practice we don't have access to the joint distribution, but we do have training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$.
- Choose f to minimize the *empirical MSE*

$$\min_f \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \overset{\text{LLN}}{\approx} \mathbb{E}_{X,Y}\left[(Y - f(X))^2\right]$$

- To formulate a tractable optimization problem, we need to restrict f to belong to a *regression model*, i.e., a class of candidates for f.
- Today we'll focus on the *linear model*

$$f(x) = w^T x + b$$

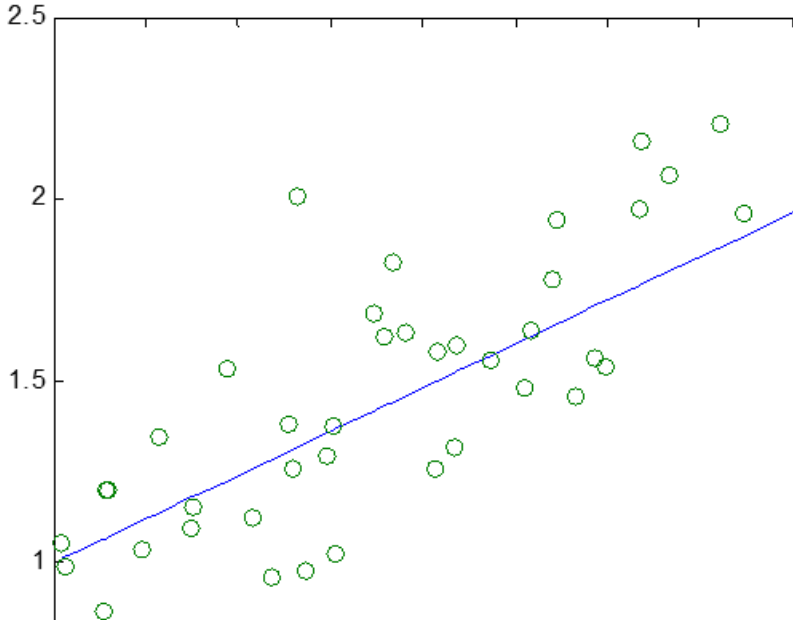
where $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$.



Least Squares Linear Regression

- *Least squares linear regression* solves

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i - b)^2$$



The method is also known as *ordinary least squares*.

- For greater generality, we can add a *regularization term*

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i - b)^2 + \lambda \|w\|^2$$

This method is known as *ridge regression*, and the term $\lambda \|\boldsymbol{w}\|^2$ is called the *ridge penalty*. $\lambda \geq 0$ is the *regularization parameter*.

OLS Derivation

$$\theta = \begin{bmatrix} b \\ w \end{bmatrix} \in \mathbb{R}^{d+1}$$

Assume $\lambda = 0$ (*ordinary least squares*). Then

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b)^2$$

$$= \frac{1}{n}\|y - X\theta\|^2$$

$$= \frac{1}{n}(y - X\theta)^T(y - X\theta)$$

$$= \frac{1}{n}(y^Ty - 2y^TX\theta + \theta^TX^TX\theta)$$

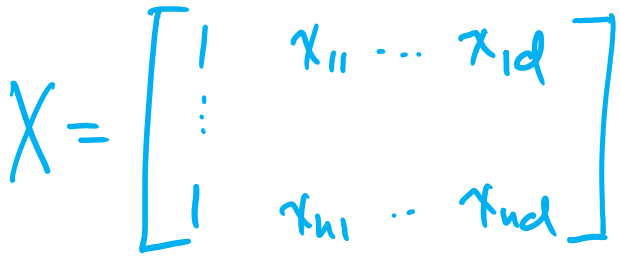
$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$$

$z^TX^TXz = \|Xz\|^2 \geq 0$

$(AB)^T = B^TA^T$ $\quad\quad$ $= (Xz)^T(Xz)$

Poll



Warning: Overloaded notation ahead.

Assume $\boldsymbol{X}$ has full column rank. Then the solution of

$$\min_{\boldsymbol{\theta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$$

is

(A) $\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}$

(B) $\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^{-1}\boldsymbol{y}$

(C) $\widehat{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

(D) $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ ✓

OLS Derivation

Fact: X has full col. rank $\Leftrightarrow$ X^TX is invertible

$$\min_{\theta} \quad y^Ty - 2y^TX\theta + \theta^TX^TX\theta$$

$$\nabla_\theta(\downarrow) = -2X^Ty + 2X^TX\theta = 0$$

$$\Rightarrow \hat{\theta} = \underbrace{(X^TX)^{-1}X^T}_{\text{pseudoinverse of } X}y \quad \text{is a critical pt.}$$

Hessian $= 2X^TX$ is PD $\Rightarrow$ strictly convex $\Rightarrow$ $\hat{\theta}$ is the unique global min

Poll

What happens in ordinary least squares when $d \geq n$? Select the best answer.

(A) $\boldsymbol{X}$ does not have full column rank ✓

(B) $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible ✓

(C) The OLS problem has infinitely many minimizers ✓

(D) All of the above ✓

$\Updownarrow$

$d+1 > n$

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ & & & \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

Ridge Regression Derivation

- Optimization problem:

$$\min_{\boldsymbol{w},b} \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b)^2 + \lambda\|\boldsymbol{w}\|^2$$

$\bar{y} - w^T\bar{x}$

- First step: eliminate b

$$\min_{w}\left[\min_{b} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i - b)^2 + \lambda\|w\|^2\right]$$

$$\frac{\partial}{\partial b}(\ ) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - w^T x_i - b) = 0$$

$$\Rightarrow \hat{b} = \hat{b}(w) = \bar{y} - w^T\bar{x}, \quad \bar{y} = \frac{1}{n}\sum y_i \in \mathbb{R}$$

$$\bar{x} = \frac{1}{n}\sum x_i \in \mathbb{R}^d$$

Ridge Regression Derivation

$$\min_w \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w^T x_i - (\bar{y} - w^T\bar{x})\right)^2 + \lambda\|w\|^2$$

$$\min_w \frac{1}{n}\sum_i \left((y_i-\bar{y}) - w^T(x_i-\bar{x})\right)^2 + \lambda\|w\|^2$$

$$\min_w \frac{1}{n}\sum_i \left(\tilde{y}_i - w^T\tilde{x}_i\right)^2 + \lambda\|w\|^2$$

$$\tilde{y}_i = y_i - \bar{y}$$
$$\tilde{x}_i = x_i - \bar{x}$$

$$\frac{1}{n}\|\tilde{y} - \tilde{X}w\|^2 + \lambda\|w\|^2$$

$$\tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix}, \tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix}$$

Ridge Regression Derivation

- The regularized least-squares (i.e., ridge regression) objective function can be written (after eliminating b)

$$J(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T \boldsymbol{A}\boldsymbol{w} + \boldsymbol{r}^T\boldsymbol{w} + c,$$

where $\boldsymbol{A} = 2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + n\lambda\boldsymbol{I})$, $\boldsymbol{r} = -2\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}$, and $c = \tilde{\boldsymbol{y}}^T\tilde{\boldsymbol{y}}$ where

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix} \qquad \tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{\boldsymbol{x}}_{11} & \cdots & \tilde{\boldsymbol{x}}_{1d} \\ \vdots & \ddots & \vdots \\ \tilde{\boldsymbol{x}}_{n1} & \cdots & \tilde{\boldsymbol{x}}_{nd} \end{bmatrix},$$

$\tilde{y}_i = y_i - \bar{y}$ and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$.

If $\lambda > 0$, then A is PD:

- Solution:

$$\widehat{\boldsymbol{w}} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} + \lambda n\boldsymbol{I})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{y}}$$

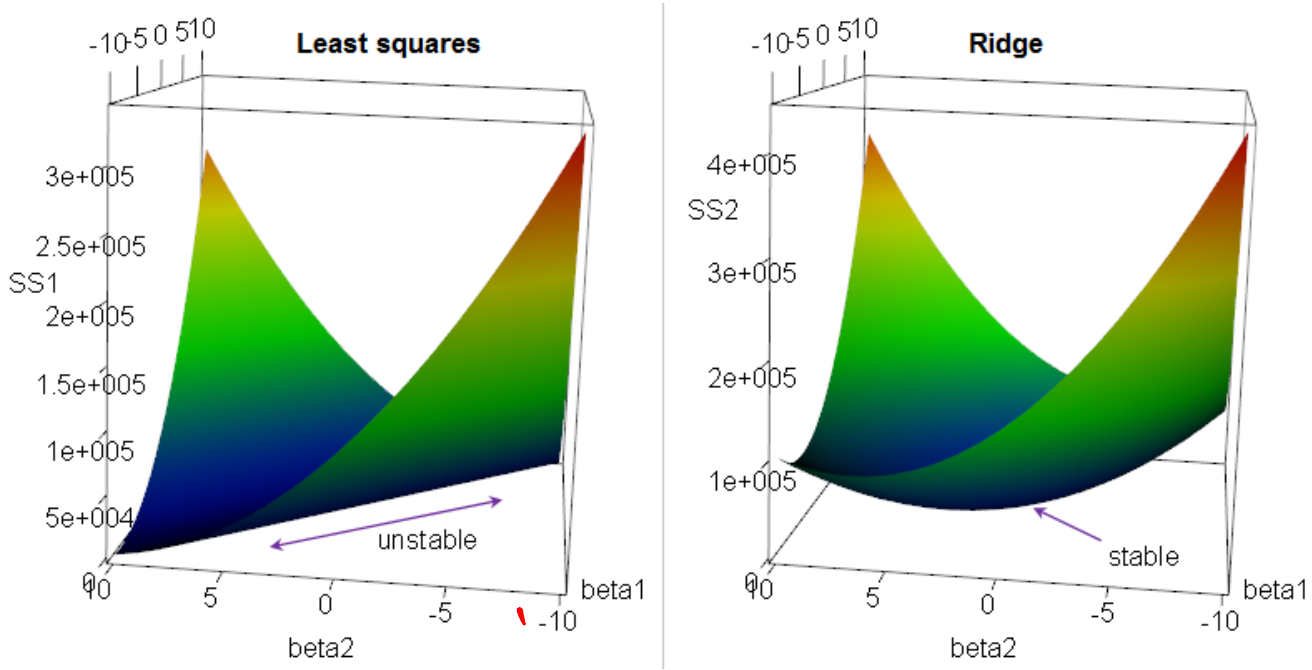
- Then $\widehat{b} = \bar{y} - \widehat{\boldsymbol{w}}^T\bar{\boldsymbol{x}}$.

If $z \neq 0$, $z^T A z = z^T(\tilde{X}^T\tilde{X} + n\lambda I)z$

$= \|\tilde{X}z\|^2 + n\lambda\|z\|^2 > 0$

OLS vs RR

- Ridge penalty: makes objective strictly convex, solution unique



https://i.stack.imgur.com/1d8XV.png

The Lasso

- ℓ_p norm $p \geq 1$

$$\|w\|_p = \left(\sum_{j=1}^{d} |w_j|^p\right)^{1/p} \qquad w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$p=2$: $\|w\|_2$ = Euclidean norm

$p=1$: $\|w\|_1 = \sum_{j=1}^{d} |w_j|$

- The lasso is an alternative to ridge regression:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - w^T x_i - b\right)^2 + \lambda\|w\|_1$$

- Motivation? Leads to sparse w (many entries are 0)

$\Rightarrow$ feature selection

The Lasso

- The contours of the sum of squared errors are elliptical:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i - b)^2 = \|\tilde{y} - \tilde{X}w\|_2^2$$

$$= (\tilde{y} - \tilde{X}w)^T(\tilde{y} - \tilde{X}w) = \tilde{y}^T\tilde{y} - 2\tilde{y}^T\tilde{X}w + w^T\tilde{X}^T\tilde{X}w$$

$$= \frac{1}{2}w^T A w + r^T w + c$$

$$= \frac{1}{2}(w - w^*)^T A (w - w^*) + c'$$

→ elliptical contours

$$w^* = -A^{-1}r$$

$$c' = c - r^T A^{-1} r$$

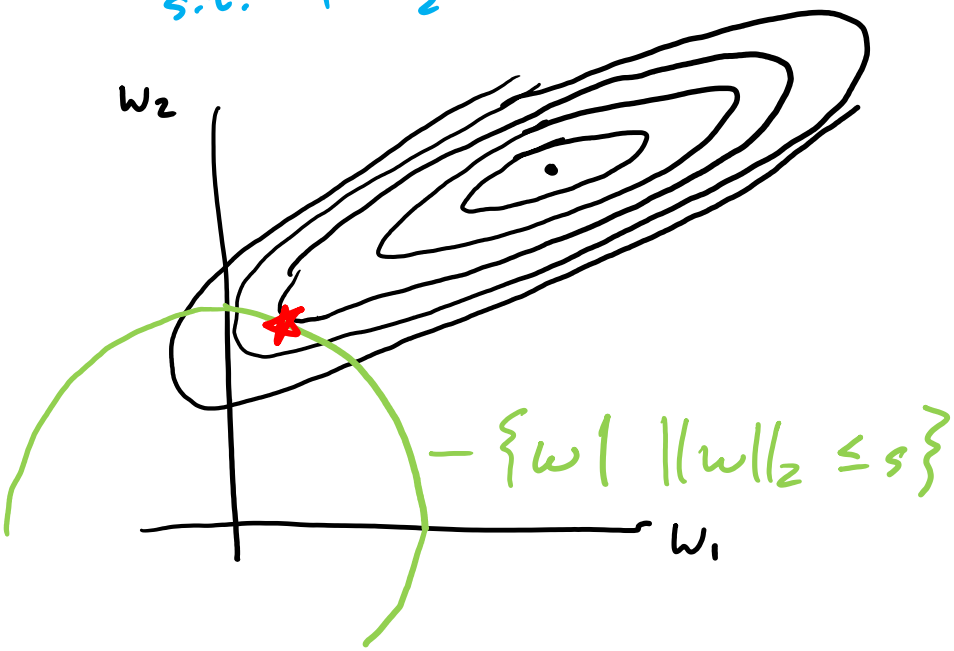
The Lasso

- Why does the lasso yield a sparse solution?

$$\min_w \|\tilde{y} - \tilde{X}w\|_2^2 + n\lambda\|w\|_2^2$$

$$\Updownarrow$$

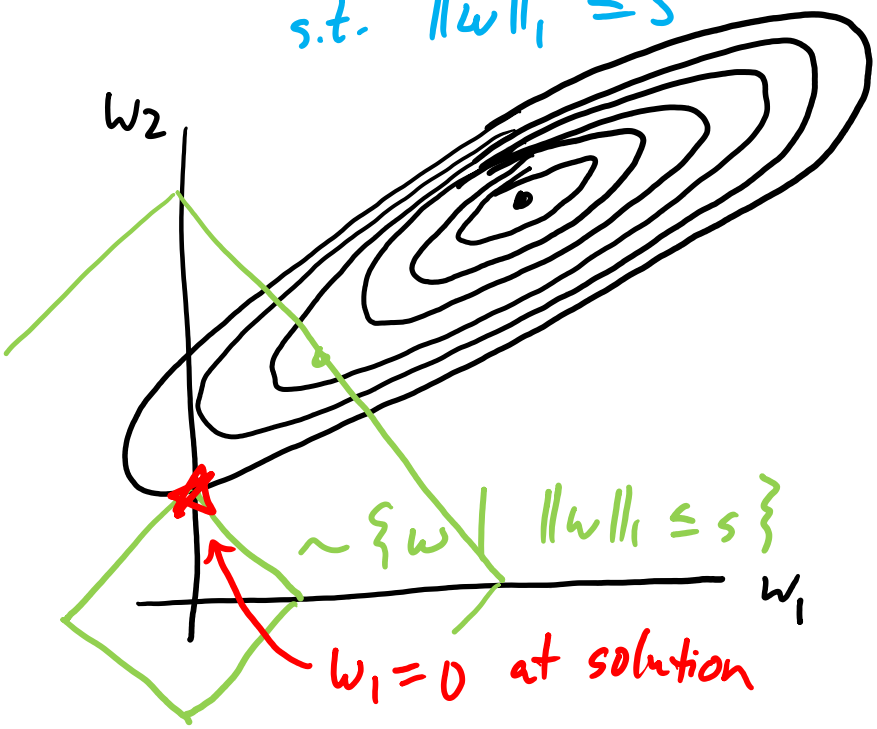
$$\min_w \|\tilde{y} - \tilde{X}w\|_2^2 \quad \text{s.t.} \quad \|w\|_2 \le s$$



$$\min_w \|\tilde{y} - \tilde{X}w\|_2^2 + n\lambda\|w\|_1$$

$$\Updownarrow$$

$$\min_w \|\tilde{y} - \tilde{X}w\|_2^2 \quad \text{s.t.} \quad \|w\|_1 \le s$$



Poll $\min_{w,b} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i - b)^2 + \lambda\|w\|_1$

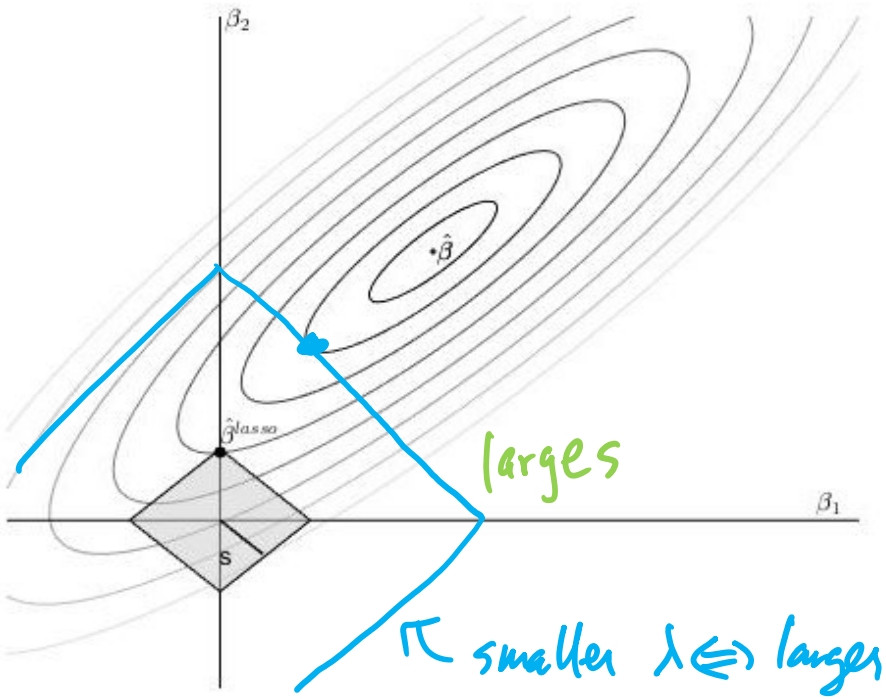
- As λ increases, the sparsity of the minimizer $\boldsymbol{w}$

✓ (A) Increases (fewer nonzero entries)

(B) Decreases (more nonzero entries)

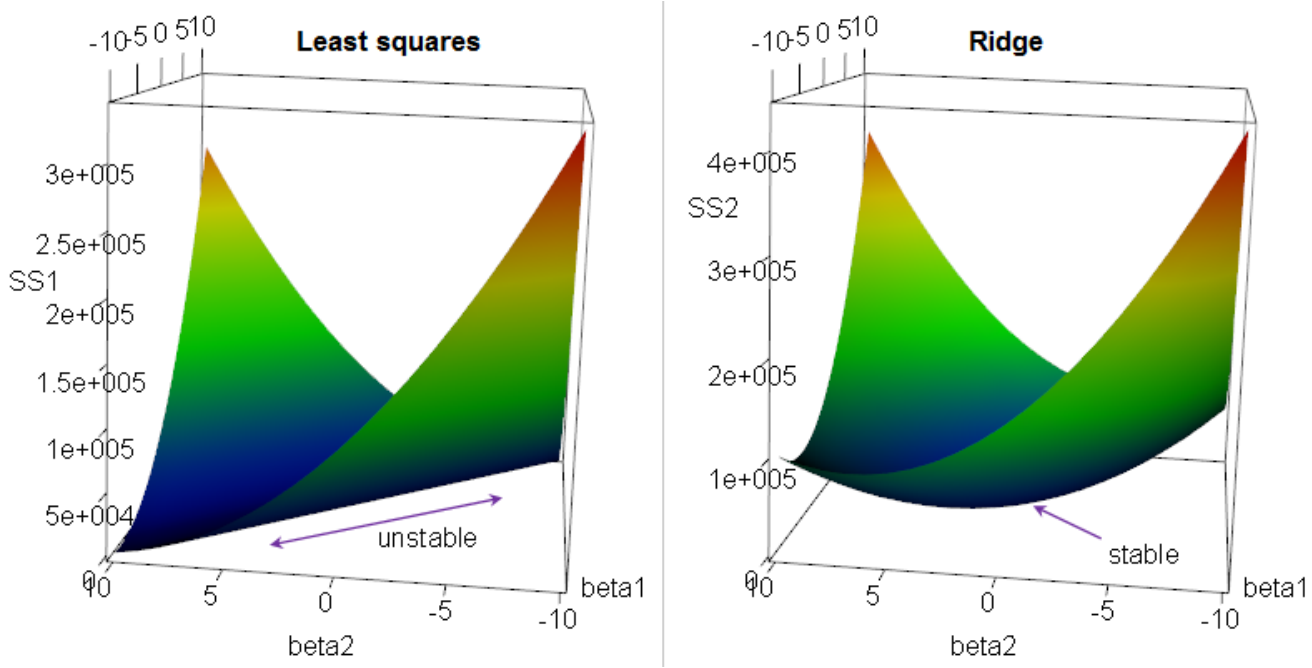
1) Larger λ makes $\|w\|_1$ smaller which leads to more 0 entries

2) Smaller λ places more emphasis on the SSE and hence a larger s.



Concluding Thoughts

- Least squares: exact solution possible, but may be infeasible for large d
- Ridge penalty: makes objective strictly convex, solution unique
- Lecture notes: robust regression
- Linear regression is the basis for several nonlinear regression methods
- Two lectures from now: computation and optimization for RR/lasso



https://i.stack.imgur.com/1d8XV.png