UVA CS 6316: Machine Learning

Lecture 6: Linear Regression Model with Regularizations

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Last: Multivariate Linear Regression with basis Expansion

Task: y

↓

Representation: x, f()

↓

Score Function: L()

↓

Search/Optimization : argmin()

↓

Models, Parameters :

Regression: y continuous

↓

Y = Weighted linear sum of (X basis expansion)

↓

Sum of Squared Error (Least Squared)

↓

Normal Equation / GD / SGD

↓

Regression coefficients

$$\hat{y} = \theta_0 + \sum_{j=1}^{m} \theta_j \varphi_j(x) = \varphi(x)^T \theta$$

Today: Regularized multivariate linear regression

Task	Regression
↓	↓
Representation	Y = Weighted linear sum of X's
↓	↓
Score Function	Least-squares + Regularization
↓	↓
Search/Optimization	Linear algebra for Ridge / sub-GD for Lasso & Elastic
↓	↓
Models, Parameters	Regression coefficients (regularized weights)

$$\min J(\beta) = \sum_{i=1}^{n}\left(Y-\hat{Y}\right)^2 + \lambda\left(\sum_{j=1}^{p}\beta_j^q\right)^{1/q}$$

We aim to make the learned model

- 1. Generalize Well

- 2. Computational Scalable and Efficient

- 3. Robust / Trustworthy / Interpretable
  - Especially for some domains, this is about trust!

Today

- Linear Regression Model with Regularizations

✓ Review: (Ordinary) Least squares: squared loss (Normal Equation)

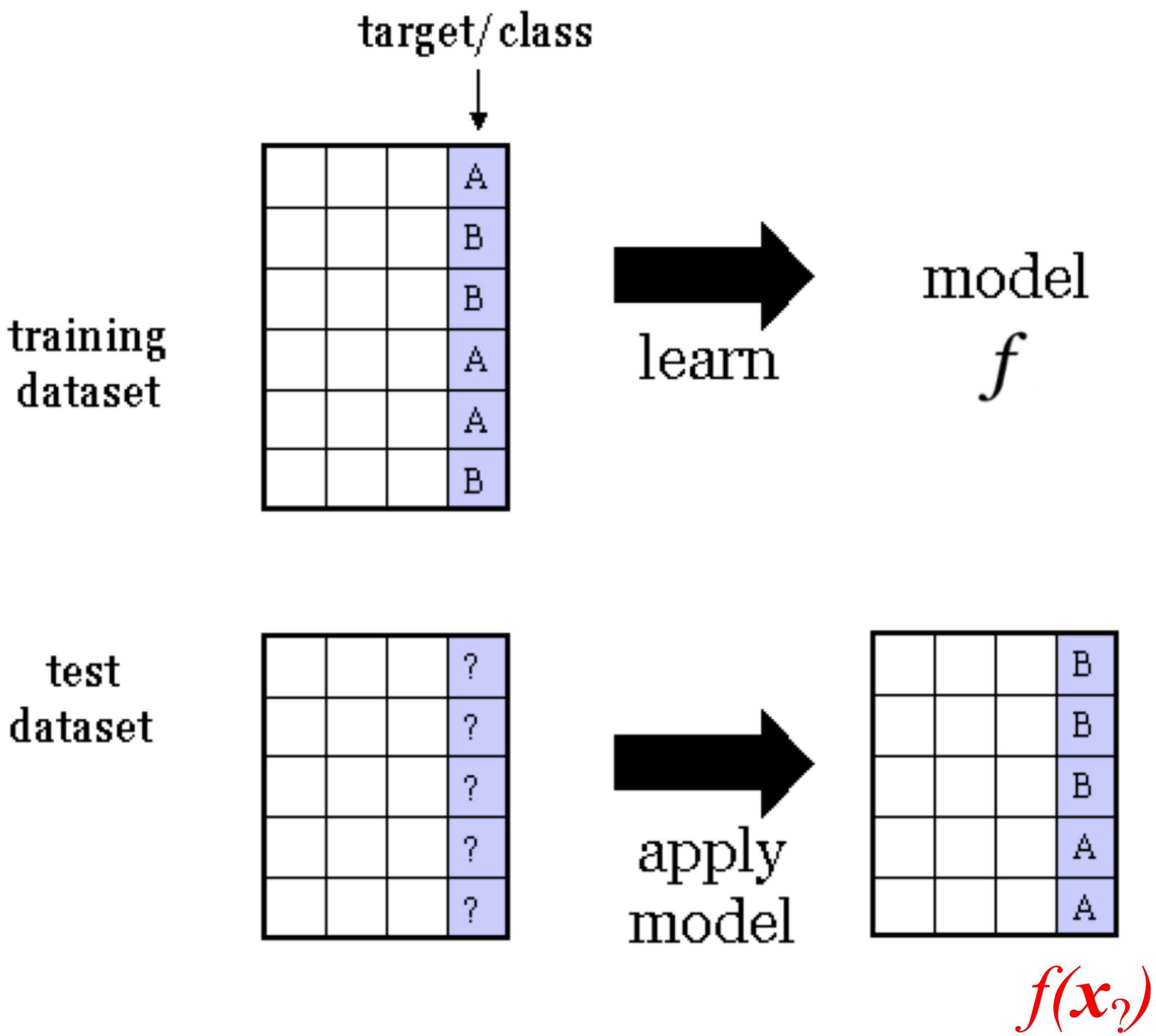
✓ Ridge regression: squared loss with L2 regularization

✓ Lasso regression: squared loss with L1 regularization

✓ Elastic regression: squared loss with L1 AND L2 regularization

✓ How to Choose Regularization Parameter

SUPERVISED Regression



Training dataset consists of **input-output** pairs

- When, target Y is a continuous target variable

Review: Normal equation for LR

- Write the cost function in matrix form:

$$J(\beta)=\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^T\beta-y_i)^2$$
$$=\frac{1}{2}\left(X\beta-\vec{y}\right)^T\left(X\beta-\vec{y}\right)$$
$$=\frac{1}{2}\left(\beta^T X^T X\beta-\beta^T X^T\vec{y}-\vec{y}^T X\beta+\vec{y}^T\vec{y}\right)$$

$$\mathbf{X}=\begin{bmatrix} -- & \mathbf{x}_1^T & -- \\ -- & \mathbf{x}_2^T & -- \\ \vdots & \vdots & \vdots \\ -- & \mathbf{x}_n^T & -- \end{bmatrix} \quad \mathbf{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

To minimize $J(\theta)$, take derivative and set to zero:

$$\Rightarrow \boxed{X^T X\beta = X^T\vec{y}}$$

The normal equations

$$\Downarrow$$

$$\beta^* = \left(X^T X\right)^{-1} X^T\vec{y}$$

Assume that $X^T X$ is invertible

Comments on the normal equation

What if $\mathbf{X}$ has less than full column rank?

→Not Invertible

$$\mathrm{rank}(X_{n\times p}) = \min(n, p)$$

when $p > n$

$$\mathrm{rank}(X) < p$$

$$\cancel{(X^T X)^{-1}}$$

$$\mathrm{rank}\Big(\underbrace{\underset{p\times n}{X^T}\ \underset{n\times p}{X}}_{p\times p}\Big) \le \min\big(r(X^T), r(X)\big) < p$$

For any matrix $A \in \mathbb{R}^{m\times n}$, it turns out that the column rank of A is equal to the row rank of A (though we will not prove this), and so both quantities are referred to collectively as the ***rank*** of A, denoted as $\text{rank}(A)$. The following are some basic properties of the rank:

- For $A \in \mathbb{R}^{m\times n}$, $\text{rank}(A) \leq \min(m, n)$ ② If $\text{rank}(A) = \min(m, n)$, then A is said to be ***full rank***.
- For $A \in \mathbb{R}^{m\times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$. ①
- For $A, B \in \mathbb{R}^{m\times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

Page11 Of Handout L2

$$\underbrace{X^T X}_{p\times p}$$

$$\text{rank}(X^T X) \leq \text{rank}(X) \leq \underbrace{\min(n, p)}_{n}$$

when $n < p$

$$\underbrace{\text{rank}(X^T X)}_{\Downarrow} < p$$

singular / not invertible

Today

- ❑ Linear Regression Model with Regularizations

✓Review: (Ordinary) Least squares: squared loss (Normal Equation)

✓Ridge regression: squared loss with L2 regularization

✓Lasso regression: squared loss with L1 regularization

✓Elastic regression: squared loss with L1 AND L2 regularization

✓How to Choose Regularization Parameter

Review: Vector norms

A norm of a vector ||x|| is informally a measure of the "length" of the vector.

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \qquad q = 1, 2, \ldots$$

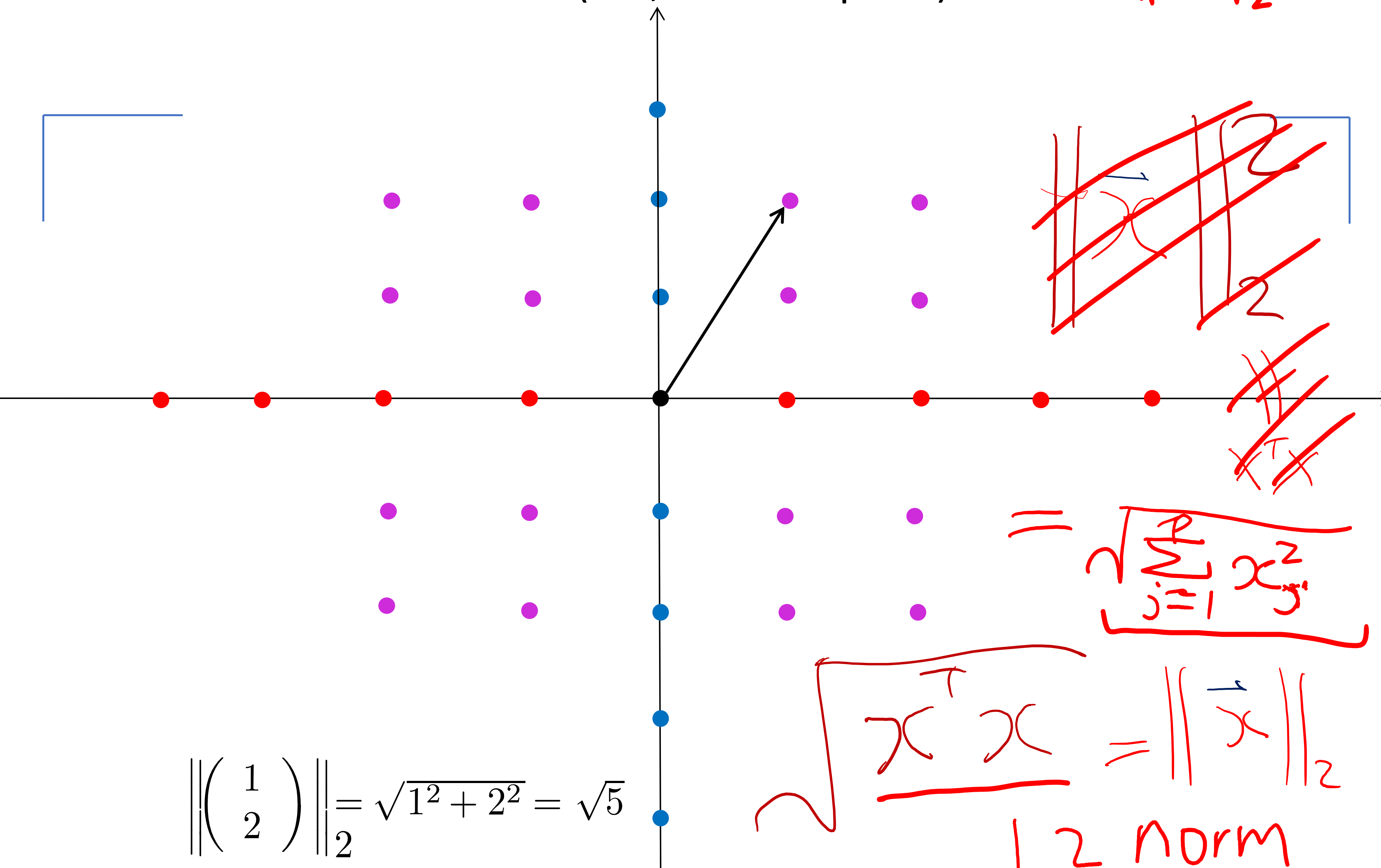
– Common norms: L_1, L_2 (Euclidean)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

– $L_{infinity}$

$$\|x\|_\infty = \max_i |x_i|$$

Review: Vector Norm (L2, when p=2) $\vec{x}^T x = \|x\|_2^2$



$\cancel{\|\vec{x}\|_2^2}$

$\cancel{x^T x}$

$= \underbrace{\sqrt{\sum_{j=1}^{p} x_j^2}}$

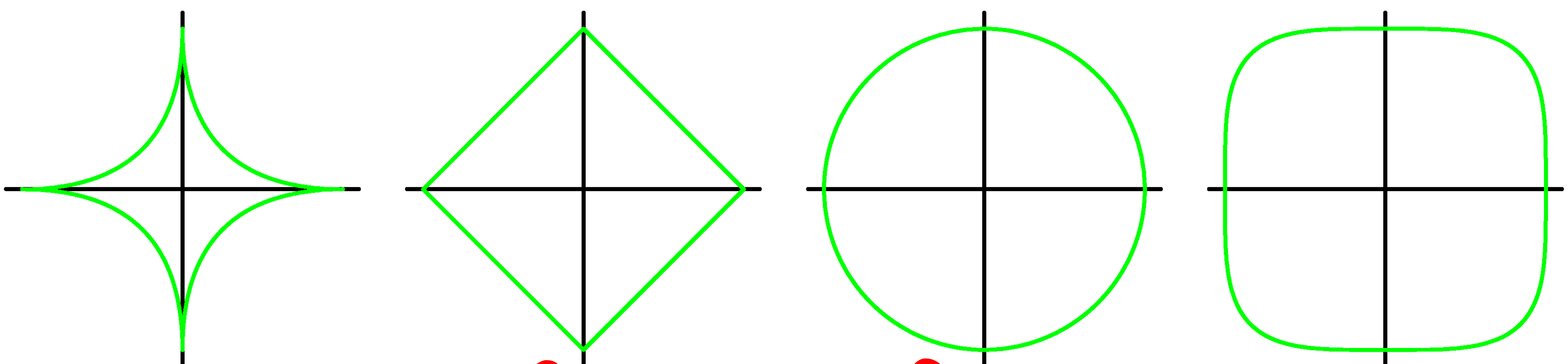
$\underline{\sqrt{x^T x}} = \|\vec{x}\|_2$

L2 norm

$$\left\| \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\|_2 = \sqrt{1^2 + 2^2} = \sqrt{5}$$

q Norms

$$\|x\|_q = \left(\sum_{i=1}^{n} |x_i|^q \right)^{1/q}$$



Ridge Regression / L2 Regularization

$$\hat{\beta}_{OLS} = \beta^* = \left(X^T X\right)^{-1} X^T \vec{y}$$

- If not invertible, a classical solution is to add a small positive element to diagonal

$\lambda > 0$

$$\beta^* = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

By convention, the bias/intercept term is typically not regularized.
Here we assume data has been centered … therefore no bias term

Extra: Positive Definite Matrix

- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive definite*** (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T A x > 0$. This is usually denoted $A \succ 0$ (or just $A > 0$), and often times the set of all positive definite matrices is denoted $\mathbb{S}^n_{++}$.
- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive semidefinite*** (PSD) if for all vectors $x^T A x \geq 0$. This is written $A \succeq 0$ (or just $A \geq 0$), and the set of all positive semidefinite matrices is often denoted $\mathbb{S}^n_+$.

One important property of positive definite matrices is that

➔ They are always full rank, and hence, invertible.

➔ Extra: See Proof at Page 17-18 of Linear-Algebra Handout positive definite (PD)

$$\forall a \neq 0 \quad \underbrace{a^T (X^T X + \lambda I) a}_{} > 0$$

$$= a^T X^T X a + \lambda a^T a$$

$$= \|Xa\|_2^2 + \lambda \|a\|_2^2 > 0$$

$$\beta^* = \underbrace{\left(X^T X + \lambda I\right)}^{-1} X^T \vec{y}$$

Extra: Positive Definite Matrix

$\forall \vec{a} \neq 0, \quad a^T A a \geq 0 \Rightarrow A \succeq 0$

① $\underset{1\times p}{a^T} \underset{p\times n}{X^T} \underset{n\times p}{X} \underset{p\times 1}{a} = \underbrace{(\underset{n\times p}{X} \underset{p\times 1}{a})}_{n\times 1}^T (Xa) = \|Xa\|_2^2 \geq 0$

$\Downarrow$

$X^T X$ PSD

[for any non-zero vector $a \in R^p$]

② $a^T \underbrace{\left(X^T X + \lambda I\right)}_{\text{PD} \Rightarrow \text{invertible}} a = a^T X^T X a + \lambda a^T I a$

$= \|Xa\|_2^2 + \lambda \|a\|_2^2 > 0$

$\lambda > 0, \; a \neq 0$

Ridge Regression / Squared Loss+L2

$$\beta^* = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

- As the solution from

HW2

$$\hat{\beta}^{ridge} = \text{argmin}(y - X\beta)^T (y - X\beta) + \lambda\beta^T\beta$$

to minimize, take derivative and set to zero

By convention, the bias/intercept term is typically not regularized.
Here we assume data has been centered … therefore no bias term

Ridge Regression / Squared Loss+L2

$\underset{n\times p}{X}\underset{p\times 1}{\beta} \rightarrow \underset{n\times 1}{\hat{y}}$

$$\beta^* = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

- As the solution from

HW2

$$\hat{\beta}^{ridge} = \operatorname{argmin} \overbrace{(y - X\beta)^T (y - X\beta)}^{\sum_{j=1}^{n} (y_j - \beta^T \vec{x}_j)^2} + \lambda \beta^T \beta$$

to minimize, take derivative and set to zero

By convention, the bias/intercept term is typically not regularized.
Here we assume data has been centered … therefore no bias term

Ridge Regression / Squared Loss+L2

$\underset{n\times p}{X}\underset{p\times 1}{\beta} \rightarrow \underset{n\times 1}{\hat{y}}$

$$\beta^* = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

- As the solution from

HW2

$$\hat{\beta}^{ridge} = \operatorname{argmin} \overbrace{(y - X\beta)^T (y - X\beta)}^{\sum_{j=1}^{n}(y_j - \beta^T \vec{x}_j)^2} + \lambda \beta^T \beta$$

to minimize, take derivative and set to zero

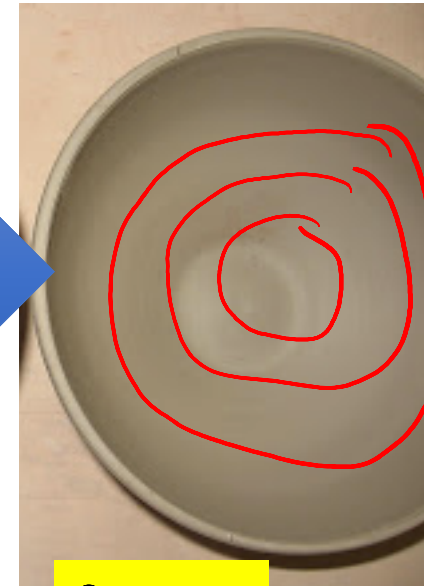
- Equivalently $\hat{\beta}^{ridge} = \operatorname{argmin} (y - X\beta)^T (y - X\beta)$

subject to $\sum_{j=\{1..p\}} \beta_j^2 \le s^2$

circle with radial s

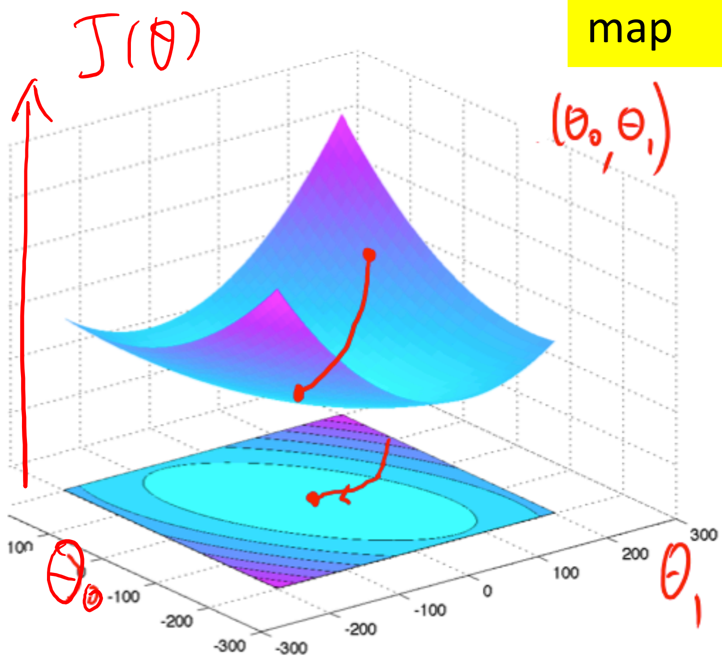
By convention, the bias/intercept term is typically not regularized.
Here we assume data has been centered … therefore no bias term

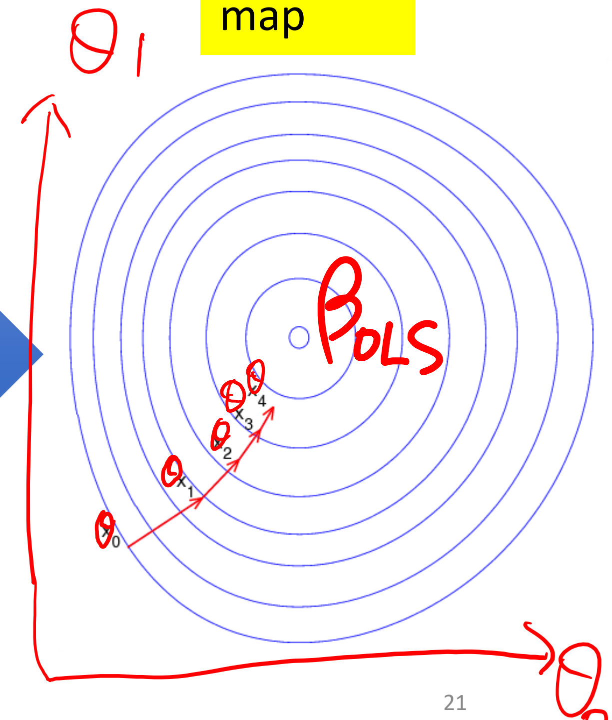
Review



Surface map

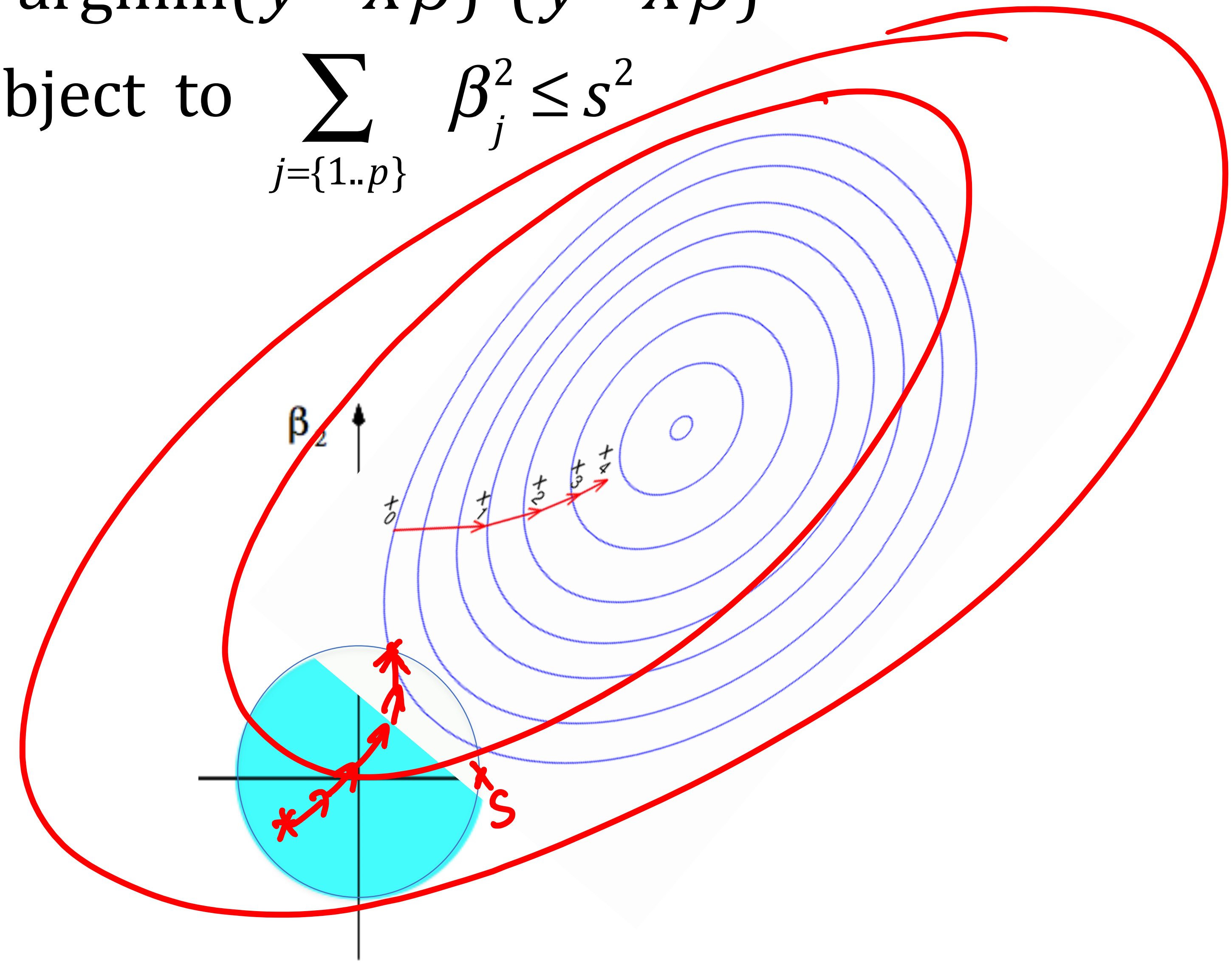
Contour map





$$\hat{\beta}^{ridge} = \text{argmin}(y - X\beta)^T(y - X\beta)$$

$$\text{subject to} \sum_{j=\{1..p\}} \beta_j^2 \le s^2$$

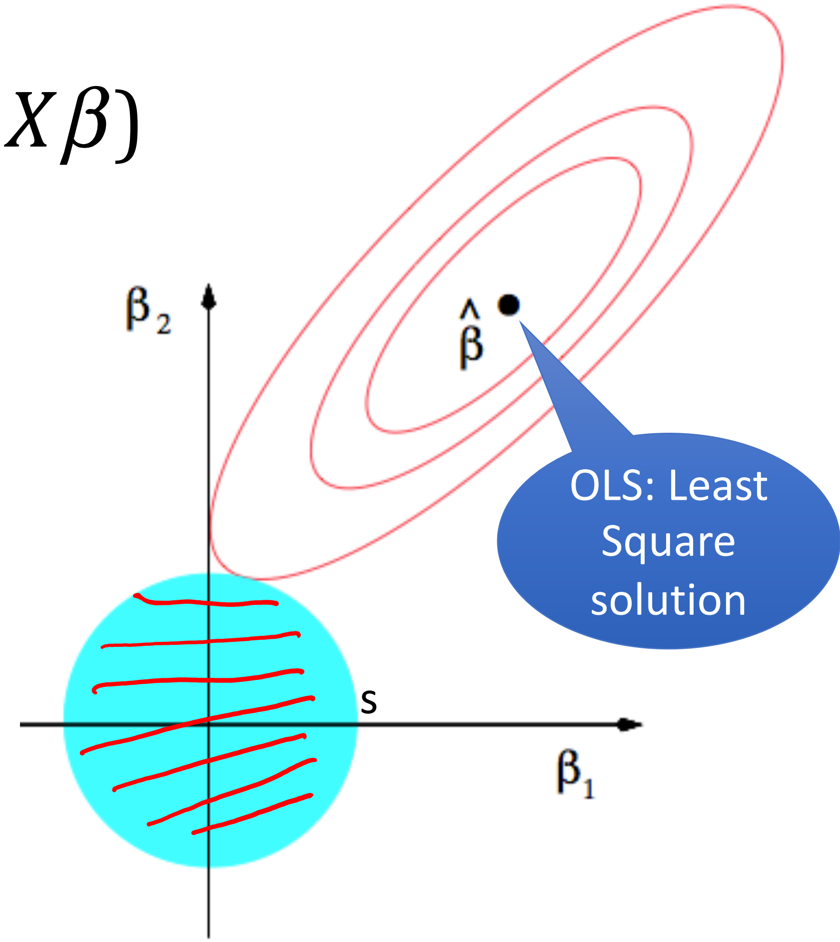


Objective Function's Contour lines from Ridge Regression

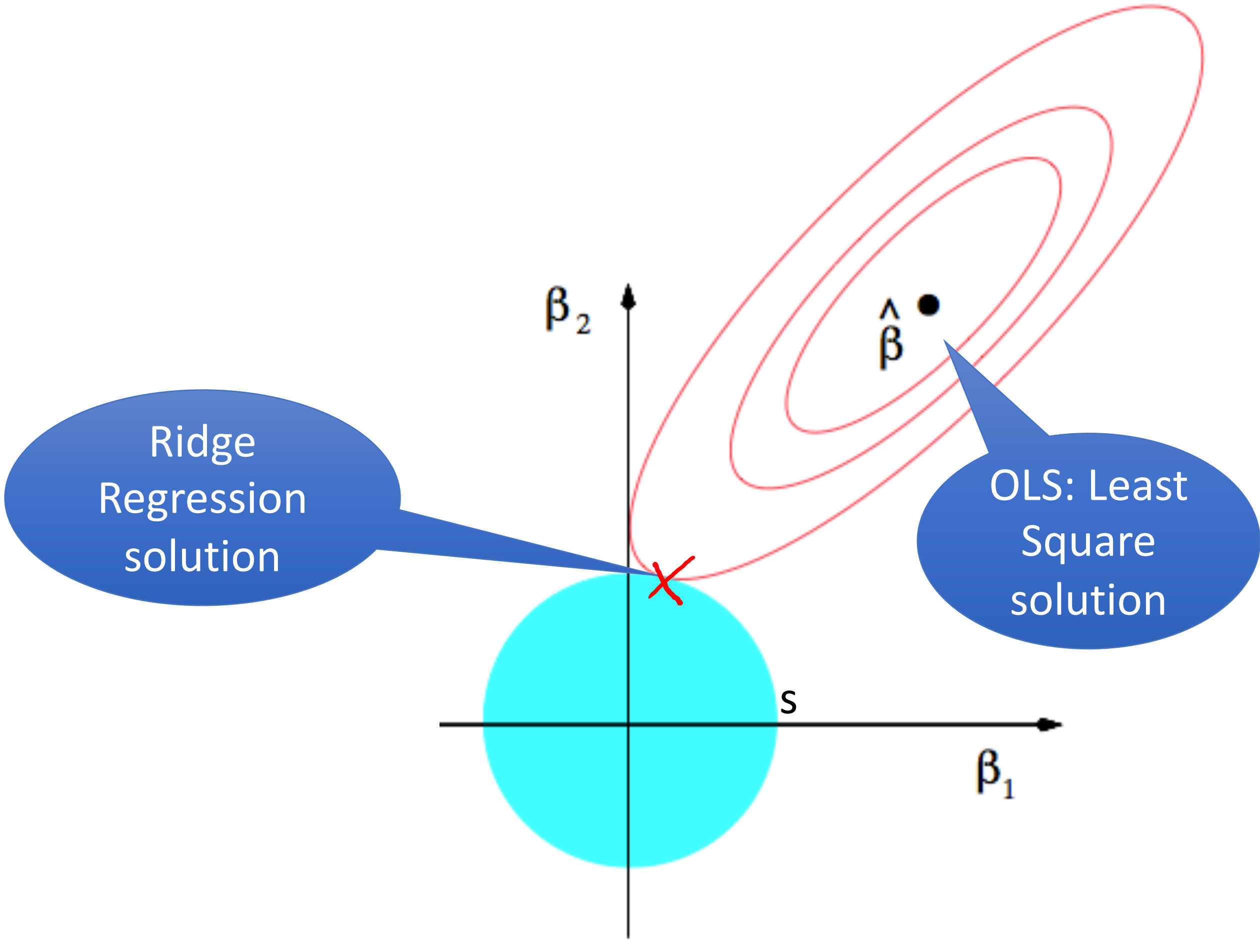
$$\hat{\beta}^{ridge} = \operatorname{argmin}(y - X\beta)^T (y - X\beta)$$

$$\text{subject to } \sum_{j=\{1..p\}} \beta_j^2 \le s^2$$

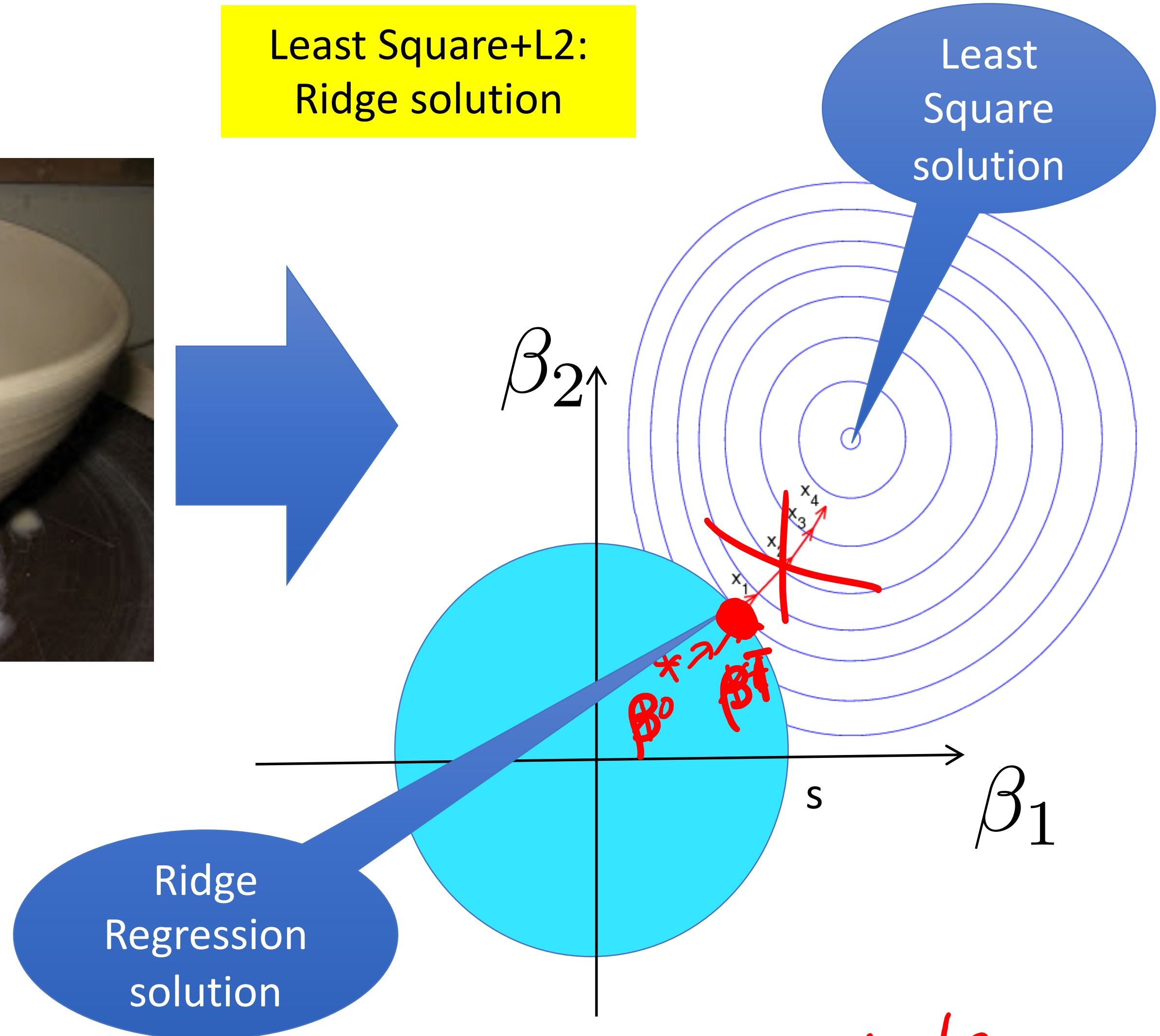
circle with radial s



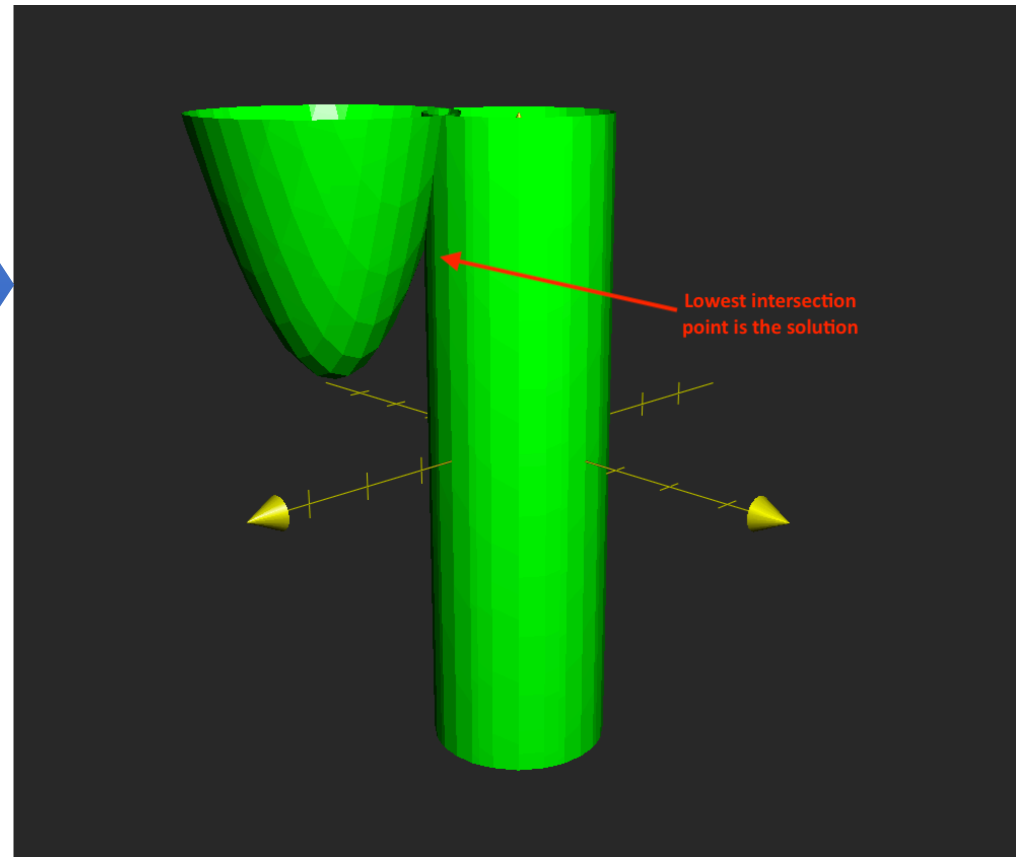
Objective Function's Contour lines from Ridge Regression



Least Square+L2: Ridge solution



must within the circle


Lowest intersection
point is the solution

Parameter Shrinkage

$\lambda > 0$

$$\beta_{OLS} = \left(X^T X\right)^{-1} X^T \vec{y}$$

when $X^TX = I$ $\Rightarrow$ $\beta_{OLS} = X^T y$

$\lambda > 0$

$(I + \lambda I)^{-1}$

$$\beta_{Rg} = \left(X^T X + \lambda I\right)^{-1} X^T \vec{y}$$

when $X^TX = I$ $\Rightarrow$ $\boxed{\beta_{Rg}} = \frac{1}{1+\lambda} X^T y = \boxed{\frac{1}{1+\lambda}} \beta_{OLS}$

When $X^T X = I \Rightarrow \beta_{Rg} = \frac{1}{1+\lambda} \beta_{OLS} \Rightarrow$ [Shrinkage]

When $X^T X$ general case, see advanced analysis @

Page65 of ESL book @ http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

Extra: two forms of Ridge Regression

- Totally equivalent

$$\begin{cases} ① \ \arg\min_{\theta} J(\beta) + \lambda \beta^T \beta \\ ② \ \arg\min_{\theta} J(\beta), \ \text{s.t. } \beta^T\beta \leq S^2 \end{cases}$$

Optimal solution β^*_{Rg} needs (necessary condition) $\lambda > 0$

$$\left[\lambda \left(\sum_j (\beta_{Rg})_j^2 - S^2 \right) = 0 \right] \Rightarrow S^2 = \sum_j (\beta_{Rg})_j^2$$

when $X^TX = I$, $\underbrace{S^2} = \sum_j (\beta_{Rg})_j^2 = \underbrace{\frac{1}{(1+\lambda)^2}} \sum_j (\beta_{OLS})_j^2$

$$\Rightarrow \underbrace{S^2 \propto \frac{1}{(1+\lambda)^2}}$$

$$\underbrace{\lambda} = \sqrt{\frac{\sum_j (\beta_{OLS})_j^2}{S^2}} - 1$$

http://stats.stackexchange.com/questions/190993/how-to-find-regression-coefficients-beta-in-ridge-regression

Ridge Regression: Squared Loss+L2

- $\lambda > 0$ penalizes each β_j

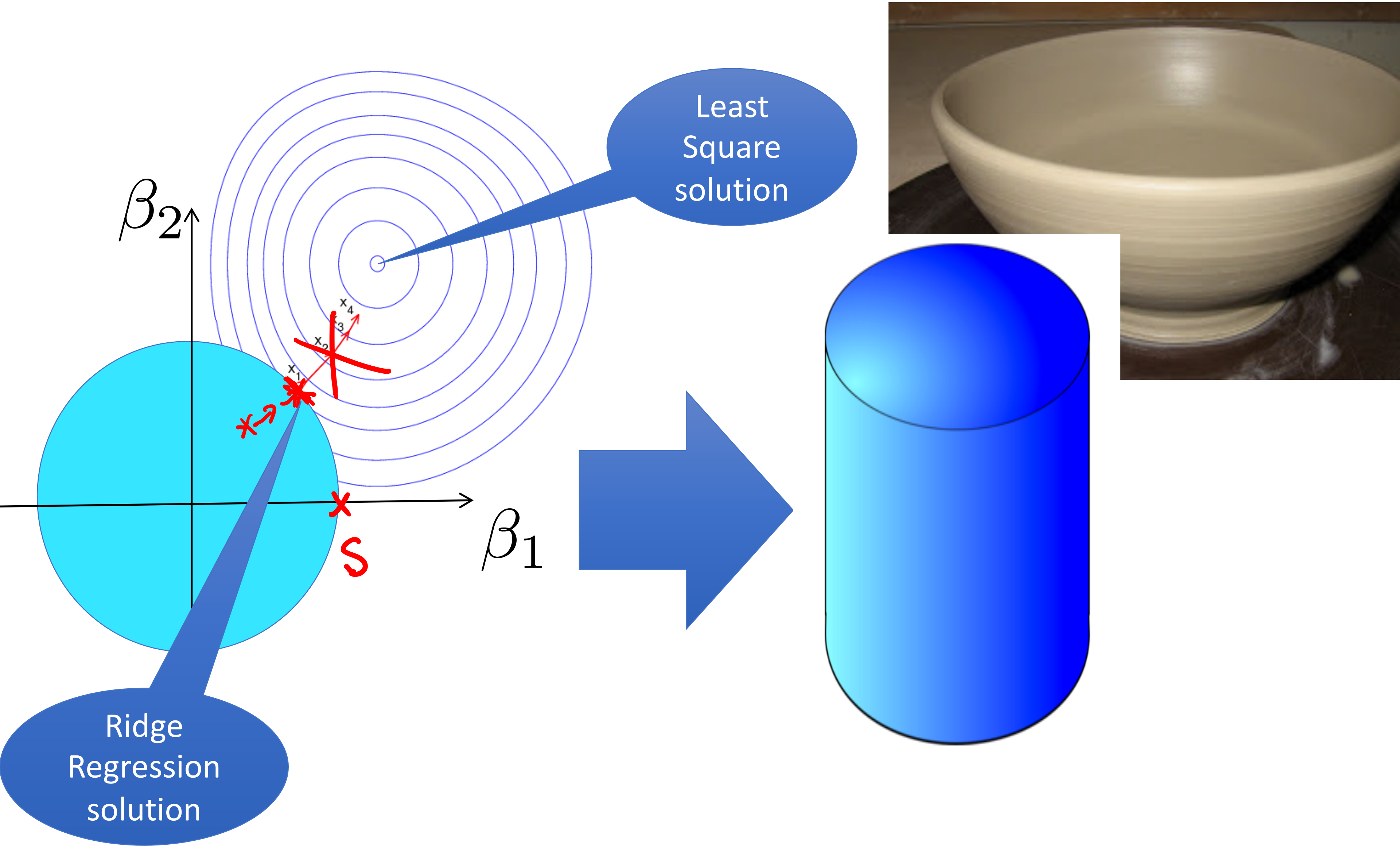
$$\frac{1}{1+\lambda} \beta_{OLS}$$

when $X^T X = I$

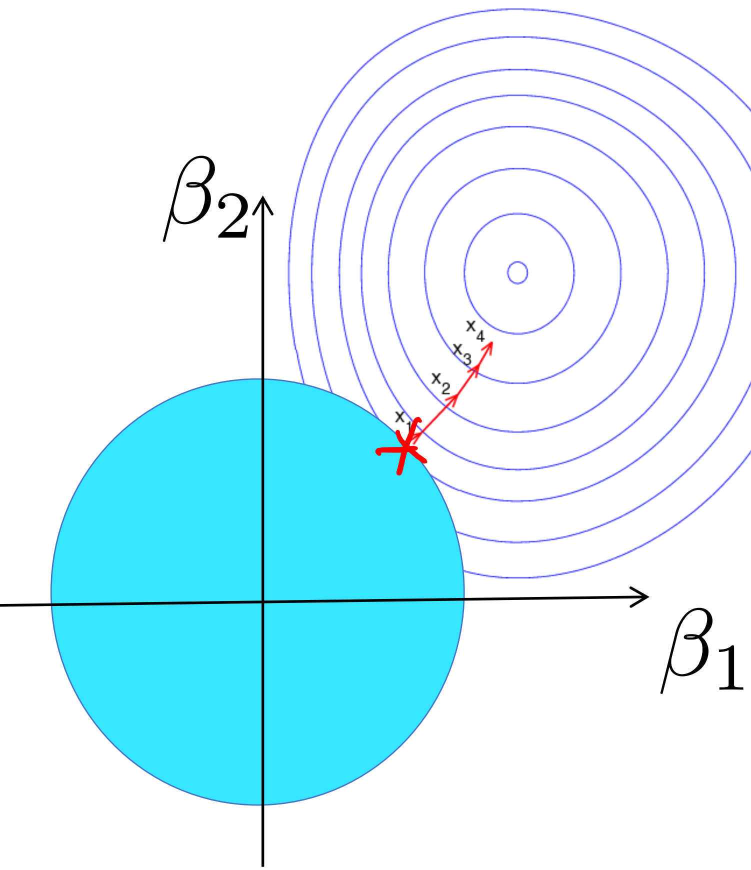
- if $\lambda = 0$ we get the least squares estimator;

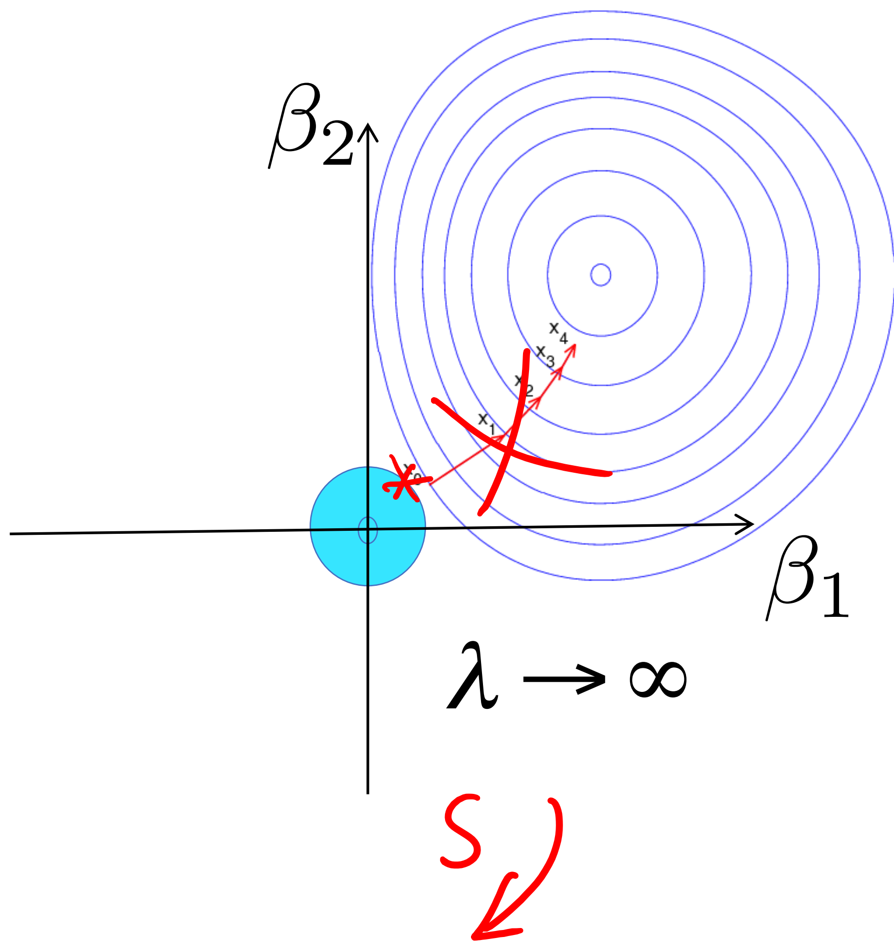
- if $\lambda \to \infty$, then β_j to zero

✓Influence of Regularization Parameter

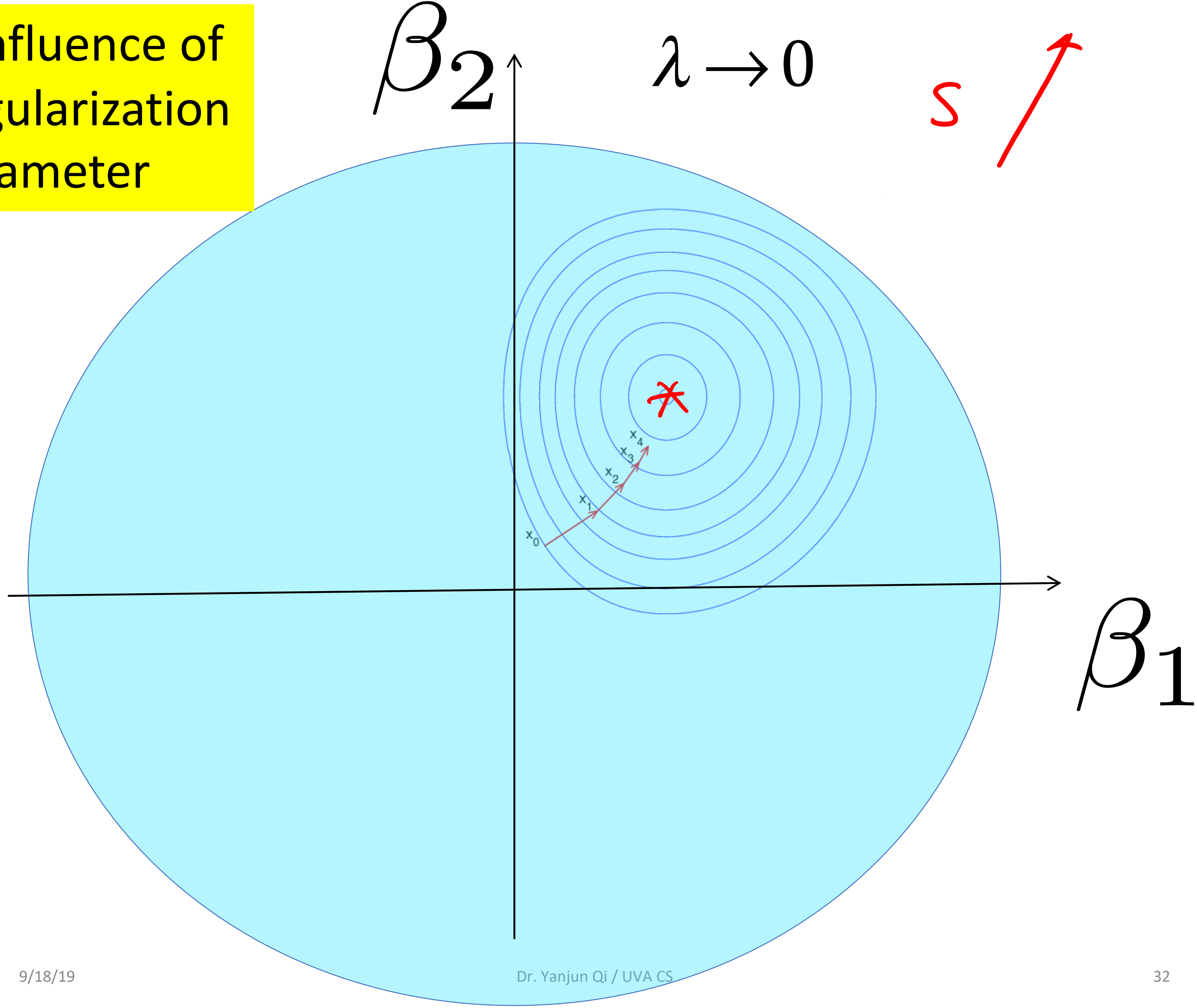


✓Influence of Regularization Parameter





✓Influence of Regularization Parameter



Today

- ❑ Linear Regression Model with Regularizations

✓Review: (Ordinary) Least squares: squared loss (Normal Equation)

✓Ridge regression: squared loss with L2 regularization

✓Lasso regression: squared loss with L1 regularization

✓Elastic regression: squared loss with L1 AND L2 regularization

✓How to Pick Regularization Parameter

(2) Lasso (least absolute shrinkage and selection operator) / Squared Loss+L1

- The lasso is a shrinkage method like ridge, but acts in a nonlinear manner on the outcome y.
- The lasso is defined by

$$\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

$$\hat{\beta}^{lasso} = \operatorname{argmin}(y - X\beta)^T(y - X\beta)$$

$$\text{subject to } \underbrace{\sum\left|\beta_j\right|}_{\text{L1 norm}} \le s$$

By convention, the bias/intercept term is typically not regularized.
Here we assume data has been centered … therefore no bias term

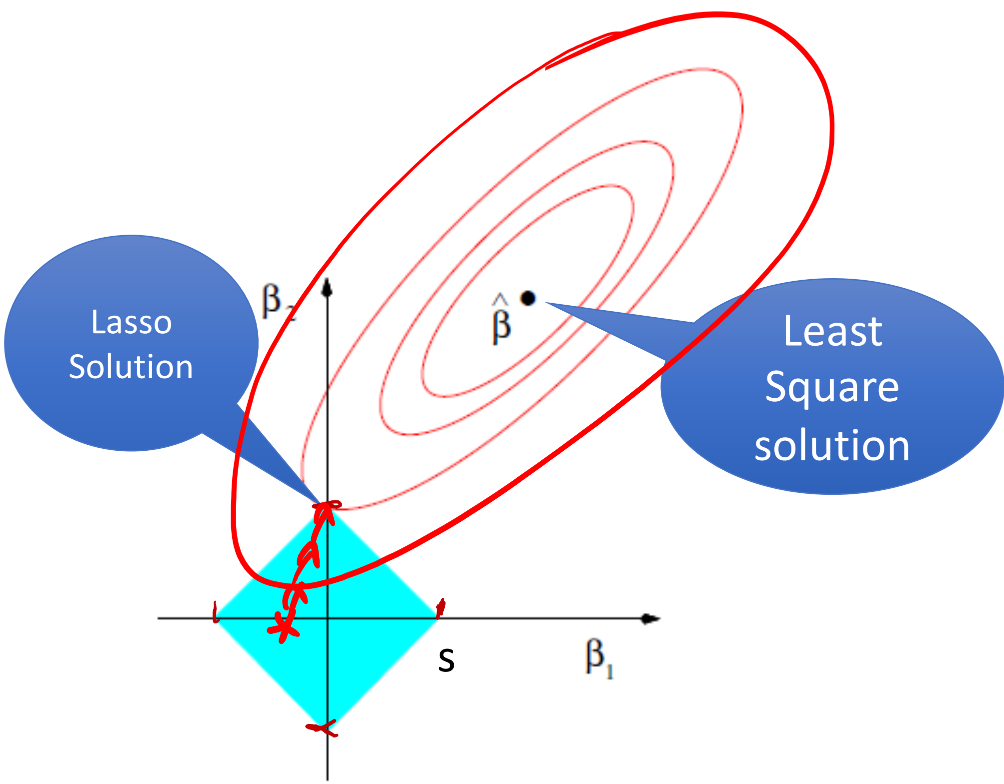
Lasso (least absolute shrinkage and selection operator)

push $\beta_j = 0$

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

$\beta^{lasso} = [\, 0, s, , 0, \,]^T$

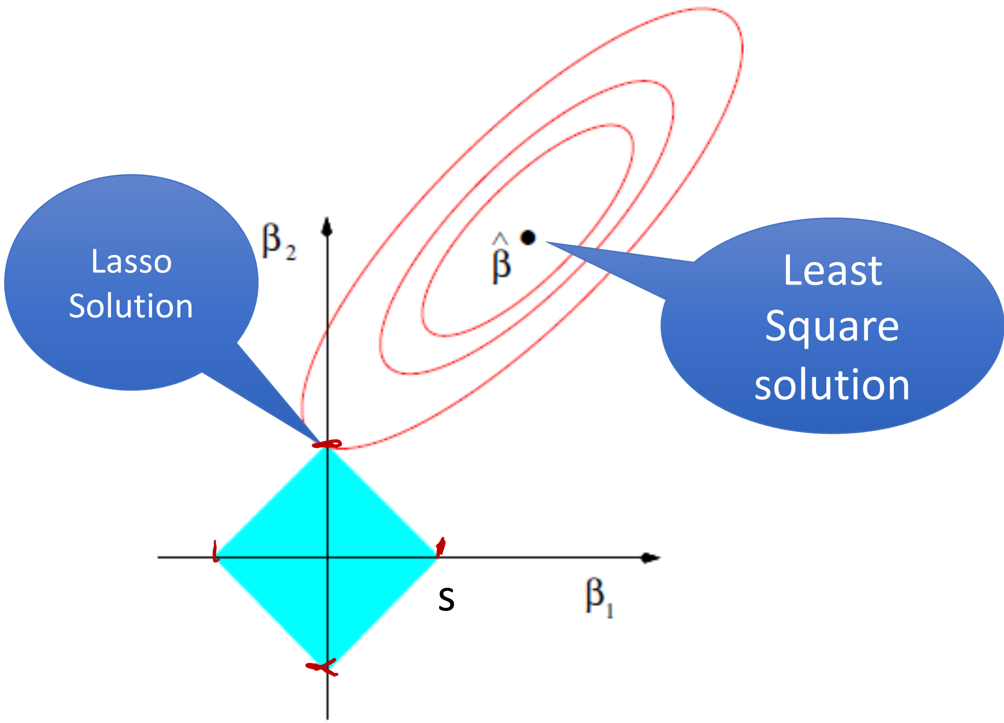
- Suppose in 2 dimension
- $\beta = (\beta_1, \beta_2)$
- $|\beta_1| + |\beta_2| = \text{const}$
- $|\beta_1| + |-\beta_2| = \text{const}$
- $|-\beta_1| + |\beta_2| = \text{const}$
- $|-\beta_1| + |-\beta_2| = \text{const}$



$$\hat{y} = \sum_{j=1}^{p} \beta_j x_j$$

when many β_j are zero $\Rightarrow$ select feature

- In the Figure, the solution has eliminated the role of x2, leading to sparsity



Lasso Estimator

Ridge Regression

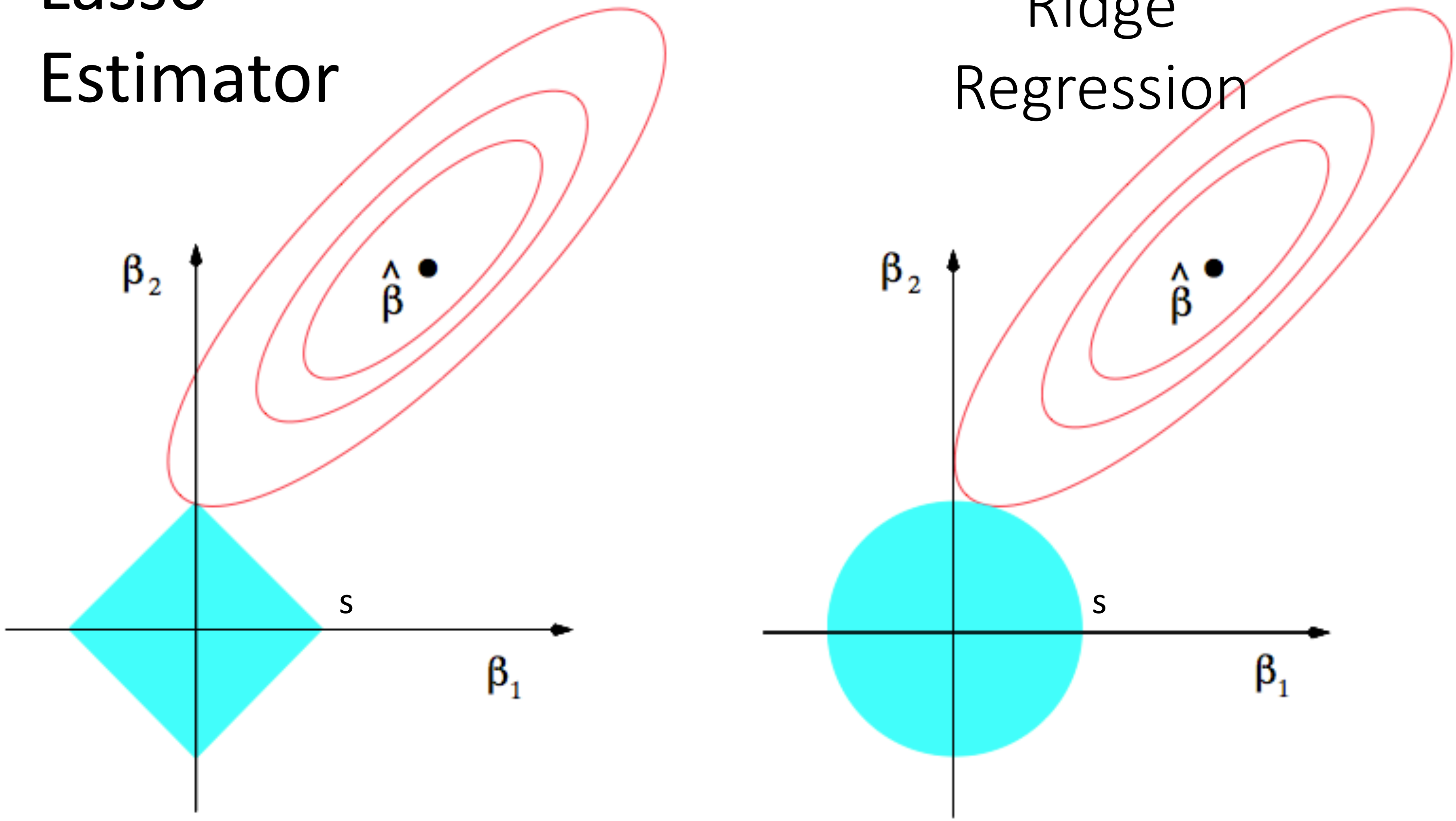


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

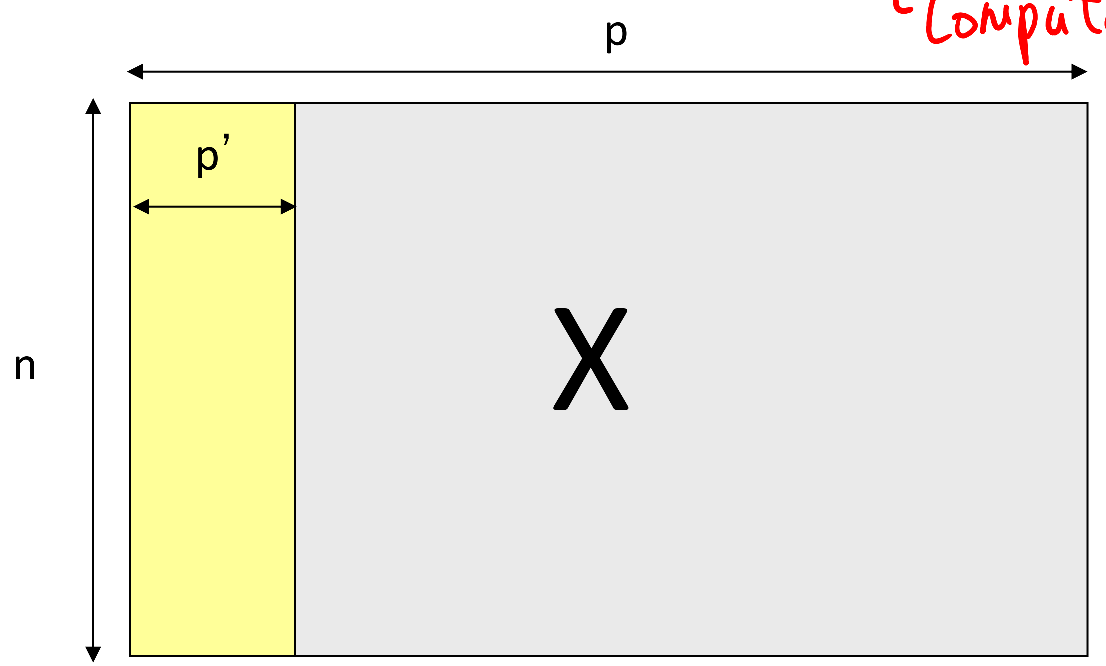
Lasso (least absolute shrinkage and selection operator)

- Notice that ridge penalty $\sum \beta_j^2$ is replaced by $\sum |\beta_j|$

- Due to the nature of the constraint, if tuning parameter is chosen small enough, then the lasso will set some coefficients exactly to zero.

Lasso: Implicit Feature Selection

$p \to p' \Rightarrow \begin{cases} \text{easy to understand} \\ \text{Computational efficient} \end{cases}$



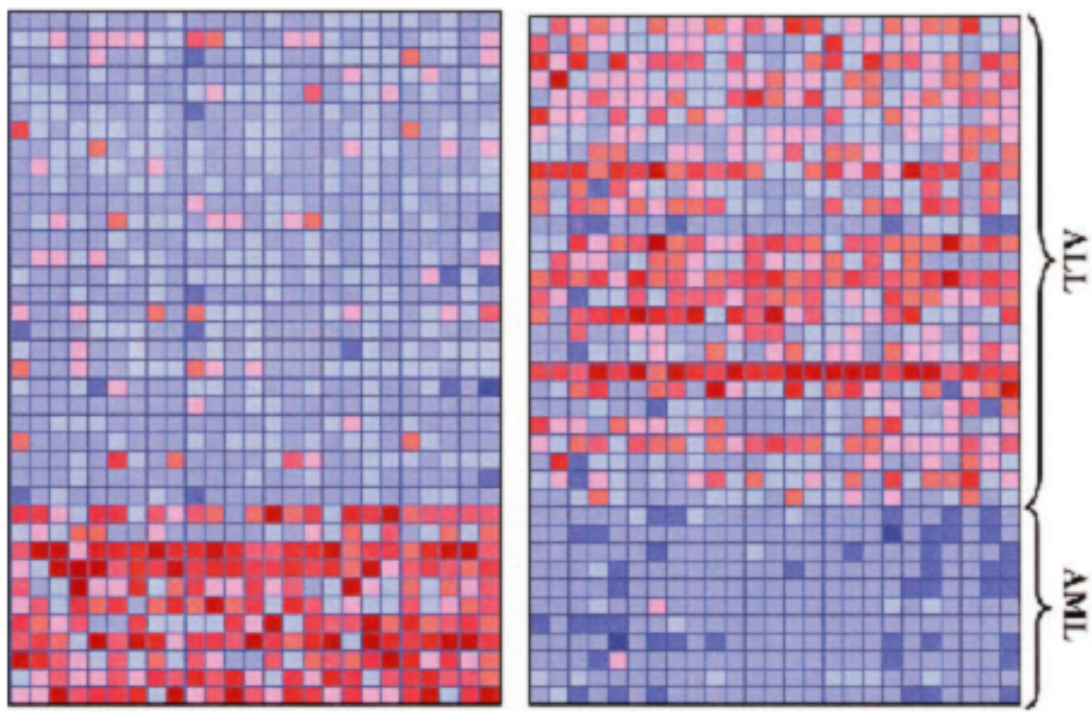
e.g., Leukemia Diagnosis

p' > 30,000

n < 100

-1

+1



Golub et al, Science Vol 286:15 Oct. 1999

$\{y_i\}$,

$$(X^T X + \lambda I)^{-1} X y$$

When $n<p$, $O(p^3)$

Computationally, choose to make $p \downarrow$ if we can

$$\Rightarrow \underset{p\times n}{X^T}\ \underset{n\times p}{X} : O(np^2)$$

$$\Rightarrow \underbrace{(X^T X + \lambda I)}_{p\times p}{}^{-1} : O(p^3)$$

$$\Rightarrow X y : O(np)$$

$\Rightarrow$

(operational mode $\underbrace{\underset{n\times p}{X}\ \underset{p\times 1}{\beta^*}}_{O(n'p)}$)

Today

- Linear Regression Model with Regularizations

✓ Review: (Ordinary) Least squares: squared loss (Normal Equation)

✓ Ridge regression: squared loss with L2 regularization

✓ Lasso regression: squared loss with L1 regularization

✓ Elastic regression: squared loss with L1 AND L2 regularization

✓ How to Pick Regularization Parameter

Lasso for when p>n

- Prediction accuracy and model interpretation are two important aspects of regression models.

- LASSO does shrinkage and variable selection simultaneously for better prediction and model interpretation.

Disadvantage:

-In p>n case, lasso selects at most n variable before it saturates

-If there is a group of variables among which the pairwise correlations are very high, then lasso select one from the group

(3) Hybrid of Ridge and Lasso : Elastic Net regularization

- L1 part of the penalty generates a sparse model
- L2 part of the penalty (extra):
  - Remove the limitation of the number of selected variables
  - Encouraging group effect
  - Stabilize the L1 regularization path

Naïve elastic net

- For any non negative fixed λ_1 and λ_2, naive elastic net criterion:

$$L(\lambda_1, \lambda_2, \beta) = |\mathbf{y} - \mathbf{X}\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1,$$

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2, \qquad |\beta|_1 = \sum_{j=1}^{p} |\beta_j|.$$

- The naive elastic net estimator is the minimizer of above equation

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\}.$$

Naïve elastic net

- For any non negative fixed λ_1 and $\lambda_{2,}$ naive elastic net criterion:

$$L(\lambda_1, \lambda_2, \beta) = |\mathbf{y} - \mathbf{X}\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1,$$

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2, \qquad |\beta|_1 = \sum_{j=1}^{p} |\beta_j|.$$

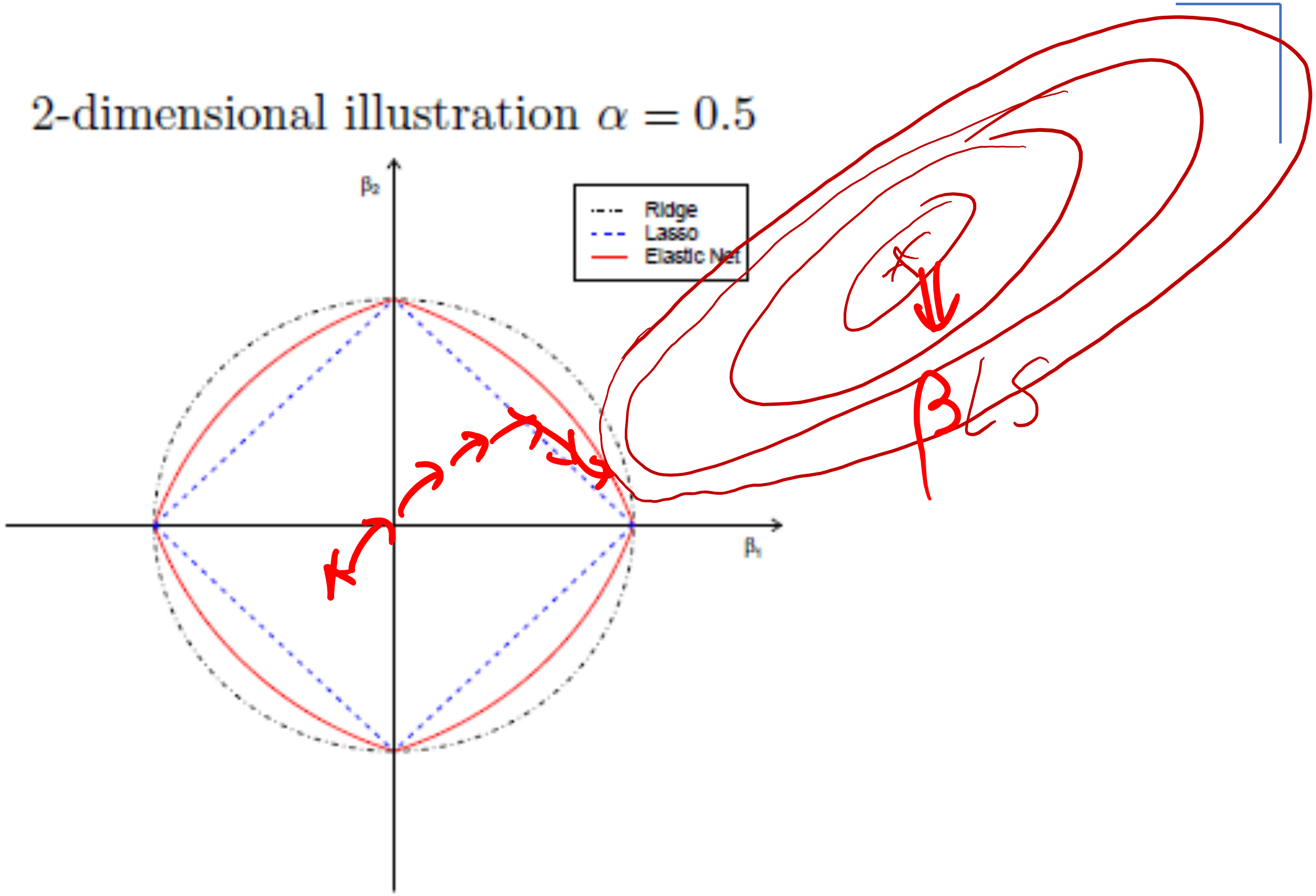
- The naive elastic net estimator is the minimizer of above

$$\hat{\beta} = \arg\min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\}.$$

- Equivalently: $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$

$$\hat{\beta} = \arg\min_{\beta} |\mathbf{y} - \mathbf{X}\beta|^2, \qquad \text{subject to } (1-\alpha)\,|\beta|_1 + \alpha\,|\beta|^2 \leqslant t \text{ for some } t.$$

Geometry of elastic net



e.g. A Practical Application of Regression Model

Movie Reviews and Revenues: An Experiment in Text Regression*

Mahesh Joshi   Dipanjan Das   Kevin Gimpel   Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{maheshj,dipanjan,kgimpel,nasmith}@cs.cmu.edu

Abstract

We consider the problem of predicting a movie's opening weekend revenue. Previous work on this problem has used metadata about a movie—e.g., its genre, MPAA rating, and cast—with very limited work making use of text *about* the movie. In this paper, we use the text of film critics' reviews from several sources to predict opening weekend revenue. We describe a new dataset pairing movie reviews with metadata and revenue data, and show that review text can substitute for metadata, and even improve over it, for prediction.

Proceedings of HLT '2010 Human Language Technologies:

I. The Story in Short

- Use metadata and critics' reviews to predict opening weekend revenues of movies
- Feature analysis shows what aspects of reviews predict box office success

$n = 1{,}718$

II. Data

- 1718 Movies, released 2005-2009
- Metadata (genre, rating, running time, actors, director, etc.): *www.metacritic.com*
- Critics' reviews (~7K): Austin Chronicle, Boston Globe, Entertainment Weekly, LA Times, NY Times, Variety, Village Voice
- Opening weekend revenues and number of opening screens: *www.the-numbers.com*

e.g., Movie Reviews and Revenues: An Experiment in Text Regression, Proceedings of HLT '10 (1.7k n / >3k features)

IV. Features

e.g. counts of a ngram in the text

I	Lexical n-grams (1,2,3)
II	Part-of-speech n-grams (1,2,3)
III	Dependency relations (nsubj,advmod,...)
Meta	U.S. origin, running time, budget (log), # of opening screens, genre, MPAA rating, holiday release (summer, Christmas, Memorial day,...), star power (Oscar winners, high-grossing actors)

$n \approx 1700$ / $p > 30{,}000$

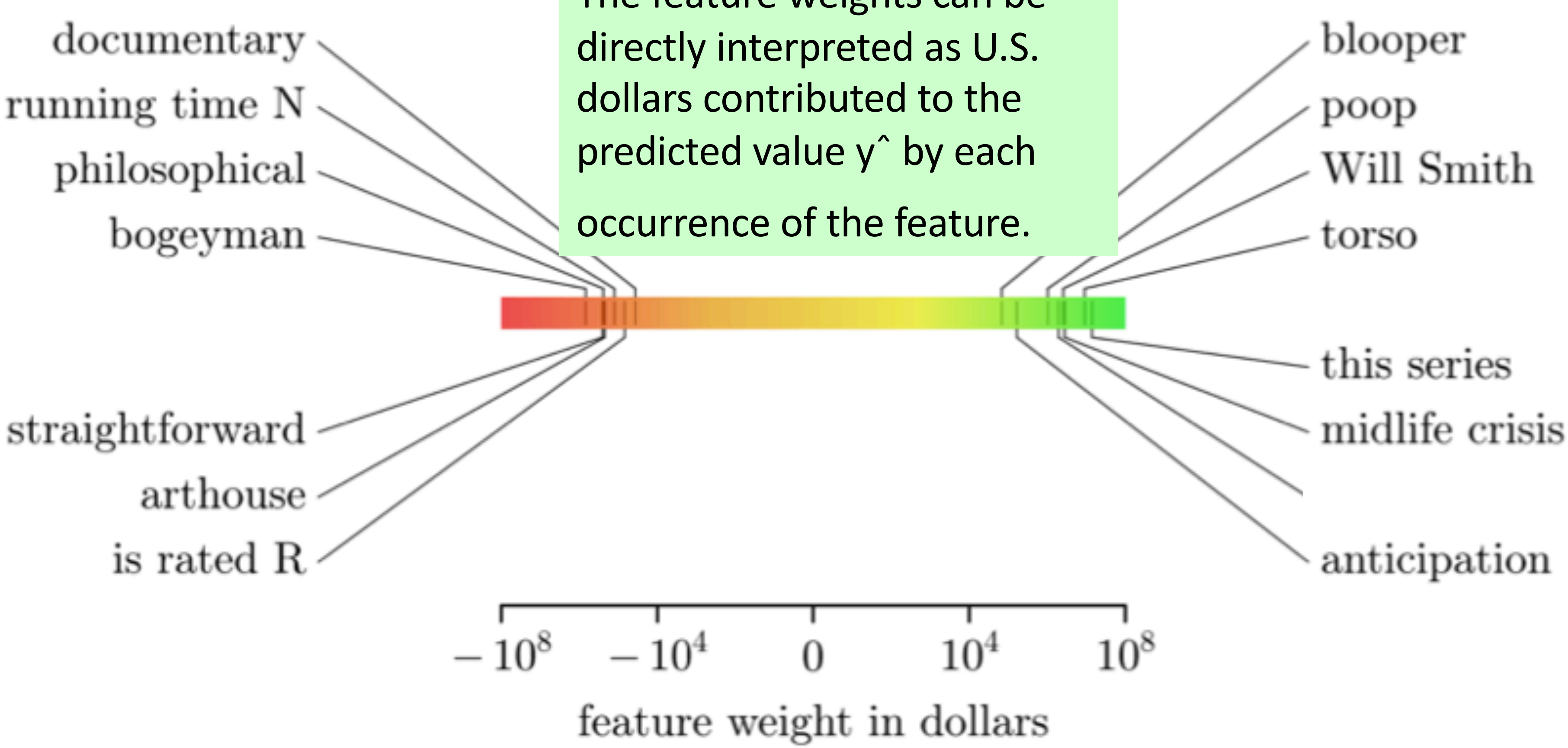
A REAL APPLICATION: Movie Reviews and meta to Revenues

VIII. Get the Data!

www.ark.cs.cmu.edu/movie$-data

V. What May Have Brought You to movies

The feature weights can be directly interpreted as U.S. dollars contributed to the predicted value y^ by each occurrence of the feature.



Movie Reviews and Revenues: An Experiment in Text Regression, Proceedings of HLT '10 Human Language Technologies:

III. Model

- Linear regression with the elastic net (Zou and Hastie, 2005)

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta})} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}) \right)^2 + \lambda P(\boldsymbol{\beta})$$

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\tfrac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right)$$

Use linear regression to directly predict the opening weekend gross earnings, denoted as y, based on features x extracted from the movie metadata and/or the text of the reviews.

	Feature	Weight ($M)
rating	pg	+0.085
	New York Times: adult	-0.236
	New York Times: rate_r	-0.364
sequels	this_series	+13.925
	LA Times: the_franchise	+5.112
	Variety: the_sequel	+4.224
people	*Boston Globe*: will_smith	+2.560
	Variety: brittany	+1.128
	^_producer_brian	+0.486
genre	*Variety*: testosterone	+1.945
	Ent. Weekly: comedy_for	+1.143
	Variety: a_horror	+0.595
	documentary	-0.037
	independent	-0.127
sentiment	*Boston Globe*: best_parts_of	+1.462
	Boston Globe: smart_enough	+1.449
	LA Times: a_good_thing	+1.117
	shame_$	-0.098
	bogeyman	-0.689
plot	*Variety*: torso	+9.054
	vehicle_in	+5.827
	superhero_$	+2.020

Table 3: Highly weighted features categorized manually. ^ and $ denote sentence boundaries.

An example of how real applications use the elastic net and its weights!

Here, the features are from the text-only model annotated in Table 2.

The feature weights can be directly interpreted as U.S. dollars contributed to the predicted value by each occurrence of the feature.

Sentiment-related text features are not as prominent as might be expected, and their overall proportion in the set of features with non-zero weights is quite small (estimated in preliminary trials at less than 15%). Phrases that refer to metadata are the more highly weighted and frequent ones.

	Features	Site	Total MAE ($M)	Total r	Per Screen MAE ($K)	Per Screen r
	Predict mean		11.672	–	6.862	–
	Predict median		10.521	–	6.642	–
meta	Best		5.983	0.722	6.540	0.272
text		–	8.013	0.743	6.509	0.222
	I	+	7.722	0.781	6.071	0.466
	see Tab. 3	B	7.627	0.793	6.060	0.411
		–	8.060	0.743	6.542	0.233
	I ∪ II	+	**7.420**	0.761	6.240	0.398
		B	7.447	0.778	6.299	0.363
		–	8.005	0.744	6.505	0.223
	I ∪ III	+	7.721	0.785	6.013	**0.473**
		B	7.595	**0.796**	†**6.010**	0.421
meta ∪ text		–	5.921	**0.819**	6.509	0.222
	I	+	5.757	0.810	6.063	0.470
		B	5.750	**0.819**	6.052	0.414
		–	5.952	0.818	6.542	0.233
	I ∪ II	+	5.752	0.800	6.230	0.400
		B	5.740	**0.819**	6.276	0.358
		–	5.921	**0.819**	6.505	0.223
	I ∪ III	+	**5.738**	0.812	6.003	**0.477**
		B	5.750	**0.819**	†**5.998**	0.423

Table 2: Test-set performance for various models, measured using mean absolute error (MAE) and Pearson's correlation (r), for two prediction tasks.

I. n-grams. We considered unigrams, bigrams, and trigrams. A 25-word stoplist was used; bigrams and trigrams were only filtered if all words were stopwords.

II. Part-of-speech n-grams. As with words, we added unigrams, bigrams, and trigrams. Tags were obtained from the Stanford part-of-speech tagger (Toutanova and Manning, 2000).

III. Dependency relations. We used the Stanford parser (Klein and Manning, 2003) to parse the critic reviews and extract syntactic dependencies. The dependency relation features consist of just the relation part of a dependency triple ⟨relation, head word, modifier word⟩.

A combination of the meta and text features **achieves the best performance both in terms of MAE and pearson r.**

We consider three ways to combine the collection of reviews for a given movie. The first ("–") simply concatenates all of a movie's reviews into a single document before extracting features. The second ("+") conjoins each feature with the source site (e.g., *New York Times*) from whose review it was extracted. A third version (denoted "B") combines both the site-agnostic and site-specific features.

More Ways for Measuring Regression Predictions: Correlation Coefficient

- Pearson correlation coefficient

$$r(x,y)=\frac{\sum_{i=1}^{m}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\overline{x})^2\times\sum_{i=1}^{m}(y_i-\overline{y})^2}}$$

$$\text{where } \overline{x}=\tfrac{1}{m}\sum_{i=1}^{m}x_i \text{ and } \overline{y}=\tfrac{1}{m}\sum_{i=1}^{m}y_i.$$

$$|r(x,y)|\le 1$$

- Measuring the linear **correlation** between two sequences, x and y,
- giving a value between +1 and −1 inclusive, where 1 is total positive **correlation**, 0 is no **correlation**, and −1 is total negative **correlation**.

- For regression: $r(\vec{y}_{\text{predicted}},\vec{y}_{\text{known}})$

Advantage of Elastic net (Extra)

$p >> n$

- Native Elastic set can be converted to lasso with augmented data form $\Rightarrow X_{n \times p}$ (when $n << p$)
- In the augmented formulation, $\Rightarrow X^*_{(n+p) \times p}$
  - sample size n+p and X^* has rank p
  - ➔ can potentially select all the predictors
- Naïve elastic net can perform automatic variable selection like lasso

Summary: Regularized multivariate linear regression

- Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$
  - LR estimation: $\arg\min \sum \left(Y - \hat{Y} \right)^2$
  - LASSO estimation: $\arg\min \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \boxed{\lambda \sum_{j=1}^{p} \left| \beta_j \right|}$
  - Ridge regression estimation: $\arg\min \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2}$

57/54

Error on data + Regularization

Regularized multivariate linear regression

Task → Representation → Score Function → Search/Optimization → Models, Parameters

Regression → Y = Weighted linear sum of X' s → Least-squares + Regularization → Linear algebra for Ridge / sub-GD for Lasso & Elastic → Regression coefficients (regularized weights)

$$\min J(\beta) = \sum_{i=1}^{n}\left(Y-\hat{Y}\right)^2 + \lambda\left(\sum_{j=1}^{p}\beta_j^q\right)^{1/q}$$

More: A family of shrinkage estimators

$$\beta = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

$$\text{subject to} \quad \sum \left| \beta_j \right|^q \leq s$$

- for q >=0, contours of constant value of $\sum_j \left|\beta_j\right|^q$ are shown for the case of two inputs.

convex

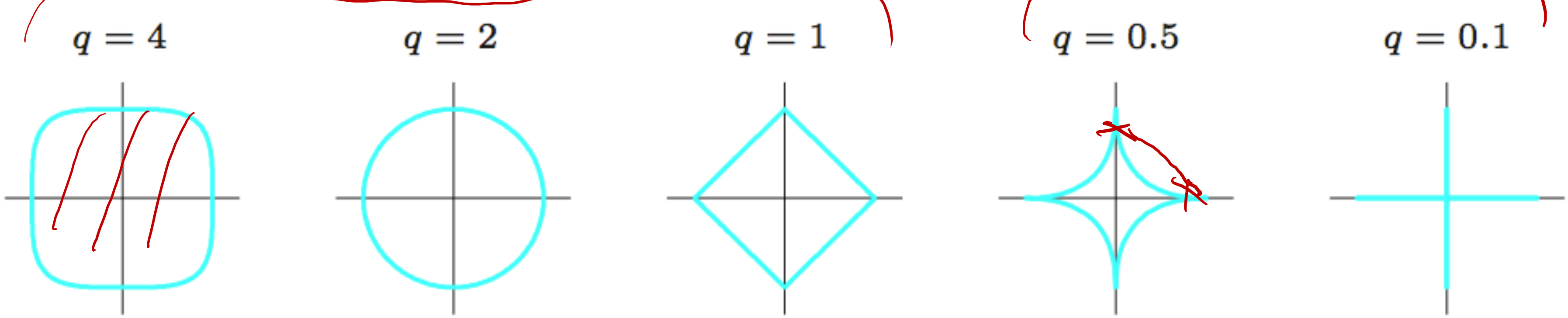
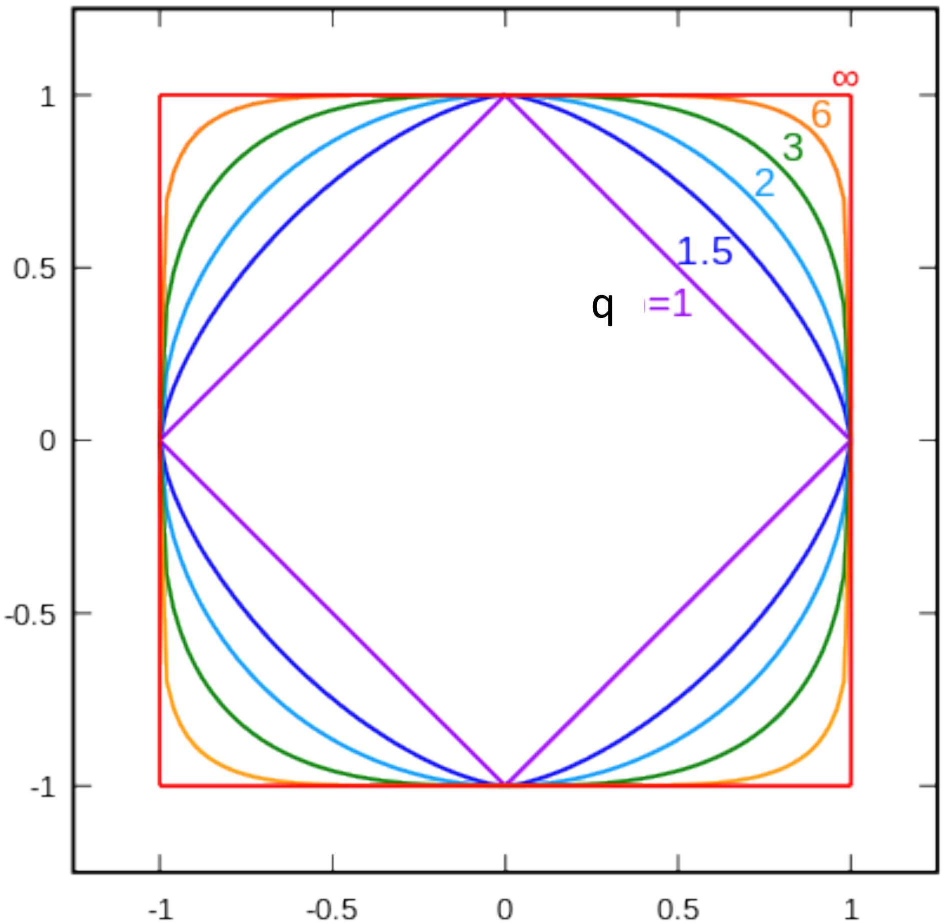


FIGURE 3.12. *Contours of constant value of $\sum_j |\beta_j|^q$ for given values of q.*

$X_{n\times p} \longrightarrow \left(\begin{bmatrix} X_{n\times p} \\ \sqrt{\lambda_2}\, I_{p\times p} \end{bmatrix}, \begin{bmatrix} Y \\ 0 \end{bmatrix} \right)$ ① $\quad X^{*}_{\underbrace{(n+p)}\times p}$

② group L1 norm

norms visualized



all p-norms penalize larger weights

q < 2 tends to create sparse (i.e. lots of 0 weights)

q > 2 tends to push for similar weights

We aim to make the learned model

- 1. Generalize Well

$p \rightarrow p'$

reduce model variance

- 2. Computationally Scalable and Efficient

$\underline{n}_{ts} \times p'$

- 3. Robust / Trustworthy / Interpretable
  - Especially for some domains, this is about trust!

Today

- Linear Regression Model with Regularizations

✓Review: (Ordinary) Least squares: squared loss (Normal Equation)

✓Ridge regression: squared loss with L2 regularization

✓Lasso regression: squared loss with L1 regularization

✓Elastic regression: squared loss with L1 AND L2 regularization

✓How to pick Regularization Parameter

Regularized multivariate linear regression

Task

Representation

Score Function

Search/Optimization

Models, Parameters

Regression

Y = Weighted linear sum of X’s

Least-squares + Regularization

Linear algebra for Ridge / sub-GD for Lasso & Elastic

Regression coefficients (regularized weights)

$$\min J(\beta) = \sum_{i=1}^{n}\left(Y - \hat{Y}\right)^2 + \lambda\left(\sum_{j=1}^{p}\beta_j^q\right)^{1/q}$$

Common regularizers

$x_1 = x_2$

$\Rightarrow \beta_1 x_1 + \beta_2 x_2$

$\beta_1 + \beta_2 = 0$

L2: Squared weights penalizes large values more

L1: Sum of weights will penalize small values more

$$\sum_j \left|\beta_j\right|$$

$$\sum_j \beta_j^2$$

Generally, we don't want huge weights

If weights are large, a small change in a feature can result in a large change in the prediction

Might also prefer weights of 0 for features that aren't useful

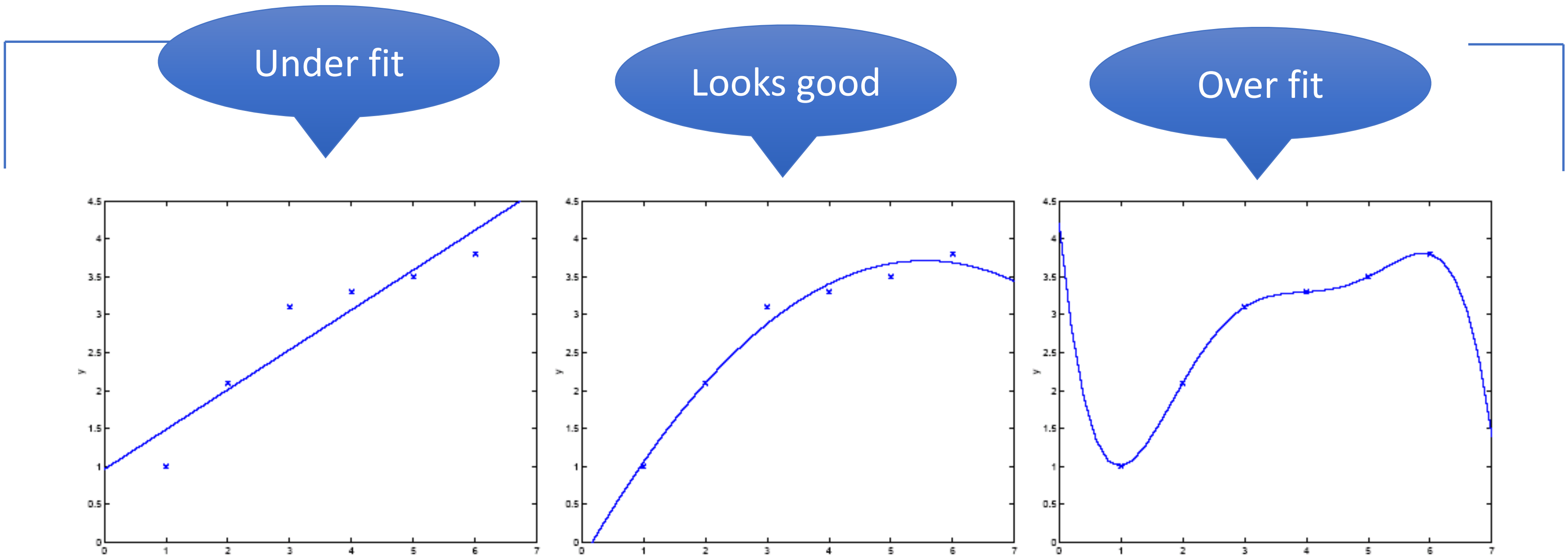
Model Selection & Generalization

- Generalisation: learn function / hypothesis from past data in order to “explain”, “predict”, “model” or “control” new data examples

- Underfitting: when model is too simple, both training and test errors are large

- Overfitting: when model is too complex and test errors are large although training errors are small.
  - After learning knowledge, model tends to learn “noise”

Issue: Overfitting and underfitting



$$y = \theta_0 + \theta_1 x \qquad y = \theta_0 + \theta_1 x + \theta_2 x^2 \qquad y = \sum_{j=0}^{5} \theta_j x^j$$

Generalisation: learn function / hypothesis from past data in order to "explain", "predict", "model" or "control" new data examples

K-fold Cross Validation !!!!

Overfitting: Handled by Regularization

A regularizer is an additional criteria to the loss function to make sure that we don’t overfit

It’s called a regularizer since it tries to keep the parameters more normal/regular

It is a bias on the model forces the learning to prefer certain types of weights over others, e.g.,

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \beta^T \beta$$

WHY and How to Select λ?

- 1. Generalization ability

  ➔ k-folds CV to decide

- 2. Control the bias and Variance of the model (details in future lectures)

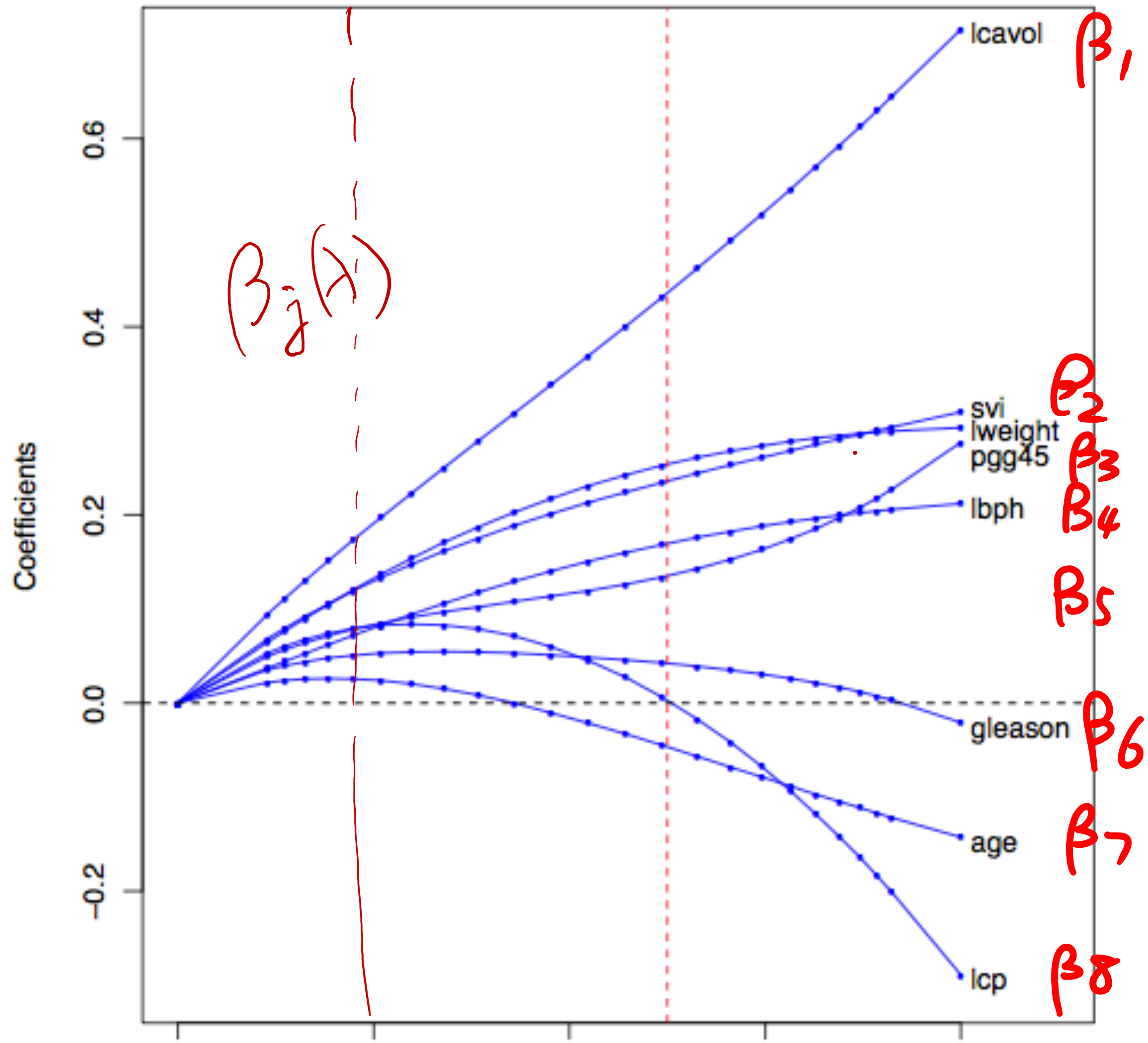
L2: Squared weights penalizes large values more

L1: Sum of weights will penalize small values more

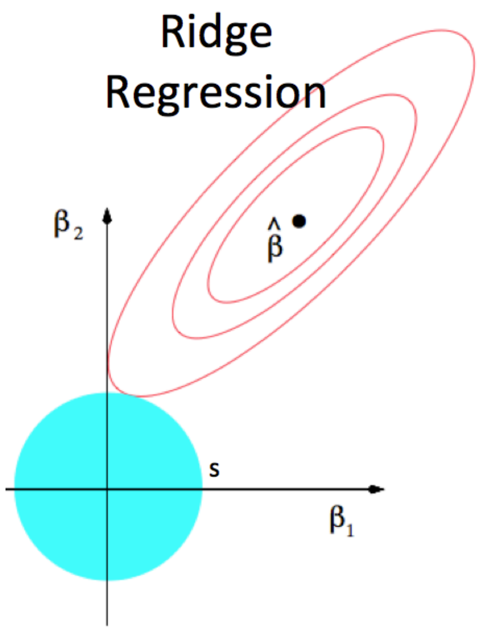
$$\sum_j \left|\beta_j\right|$$

$$\sum_j \beta_j^2$$

Regularization path of a Ridge Regression



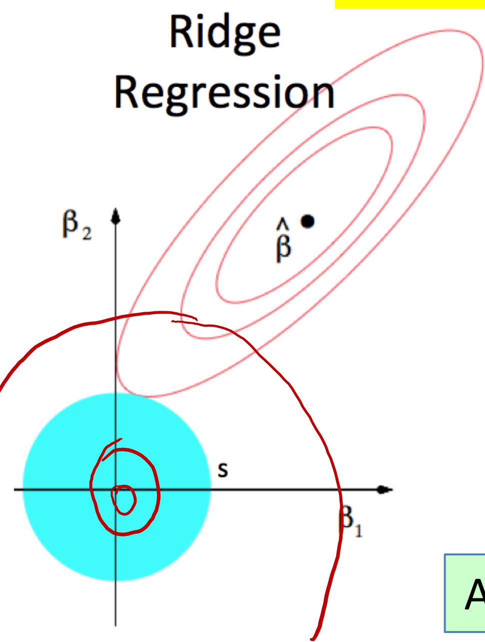
$\lambda \to \infty$ $\qquad\qquad$ $\lambda = 0$

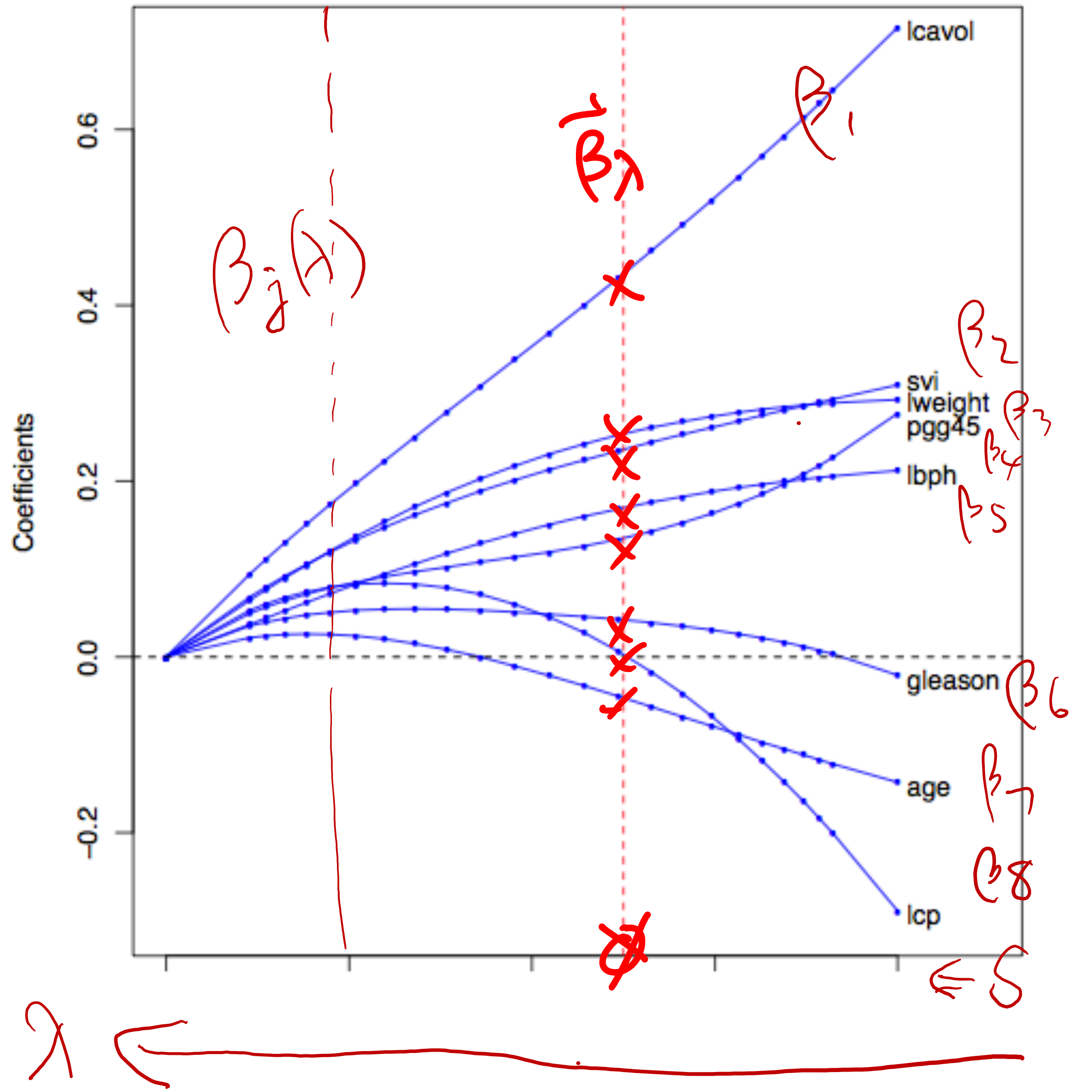


Regularization path of a Ridge Regression

when $X'X = I \Rightarrow \frac{1}{1+\lambda}\beta_{OLS}$

Weight Decay



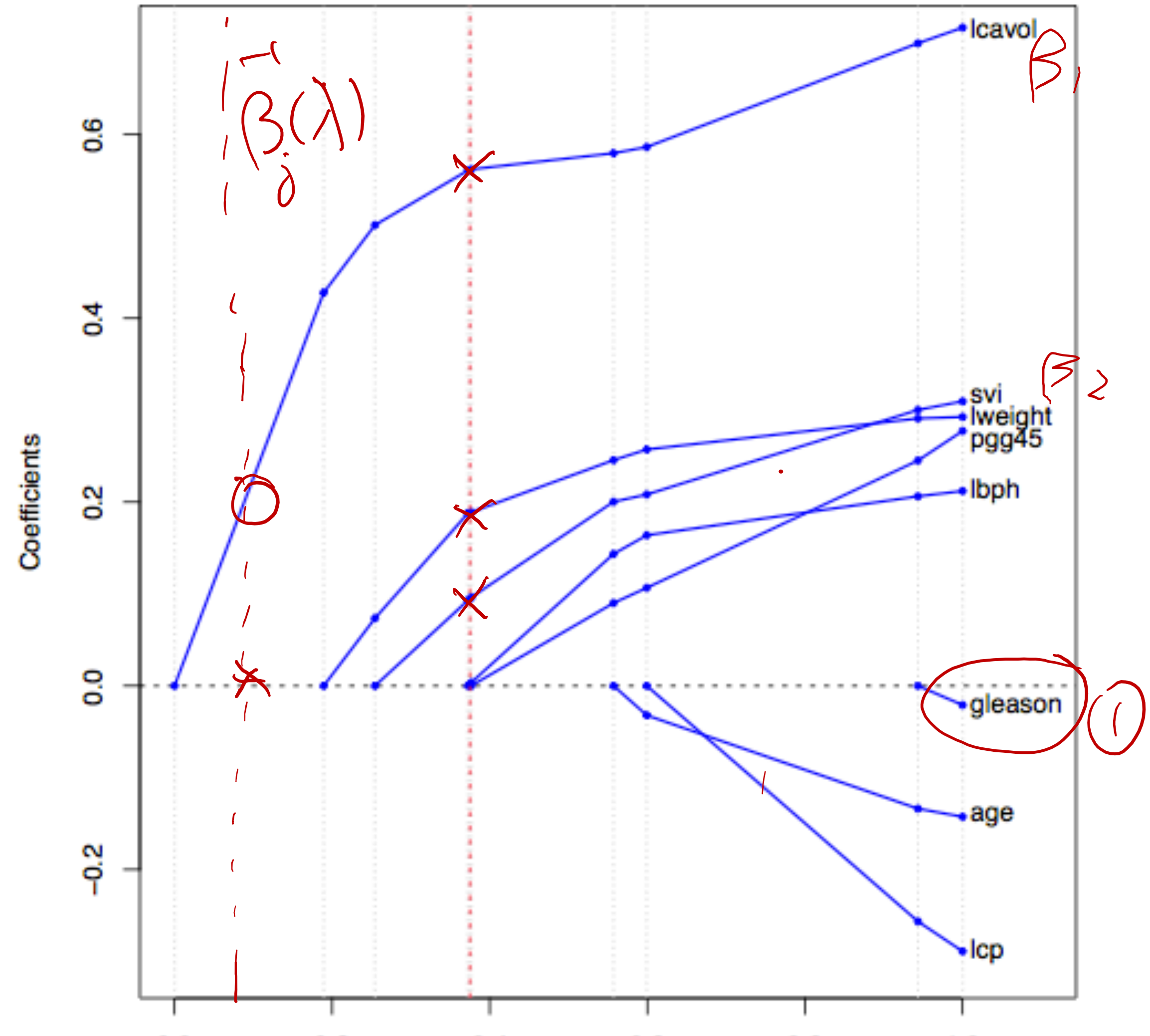


$\lambda \to \infty$ $\lambda = 0$

An example with 8 features

Regularization path of a Lasso Regression

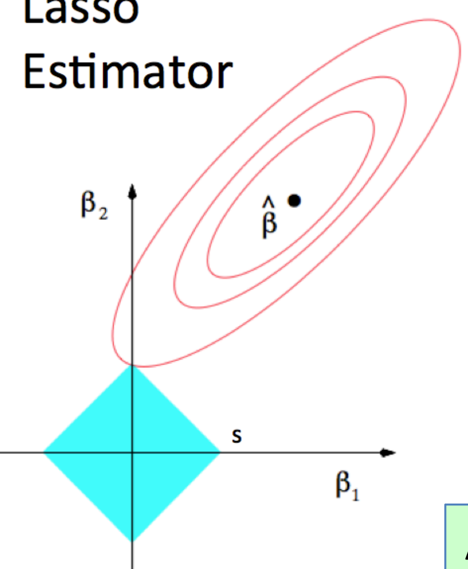
when varying λ,
how β_j varies.



$\lambda \rightarrow \infty$ λ_t $\lambda = 0$

$p=8$

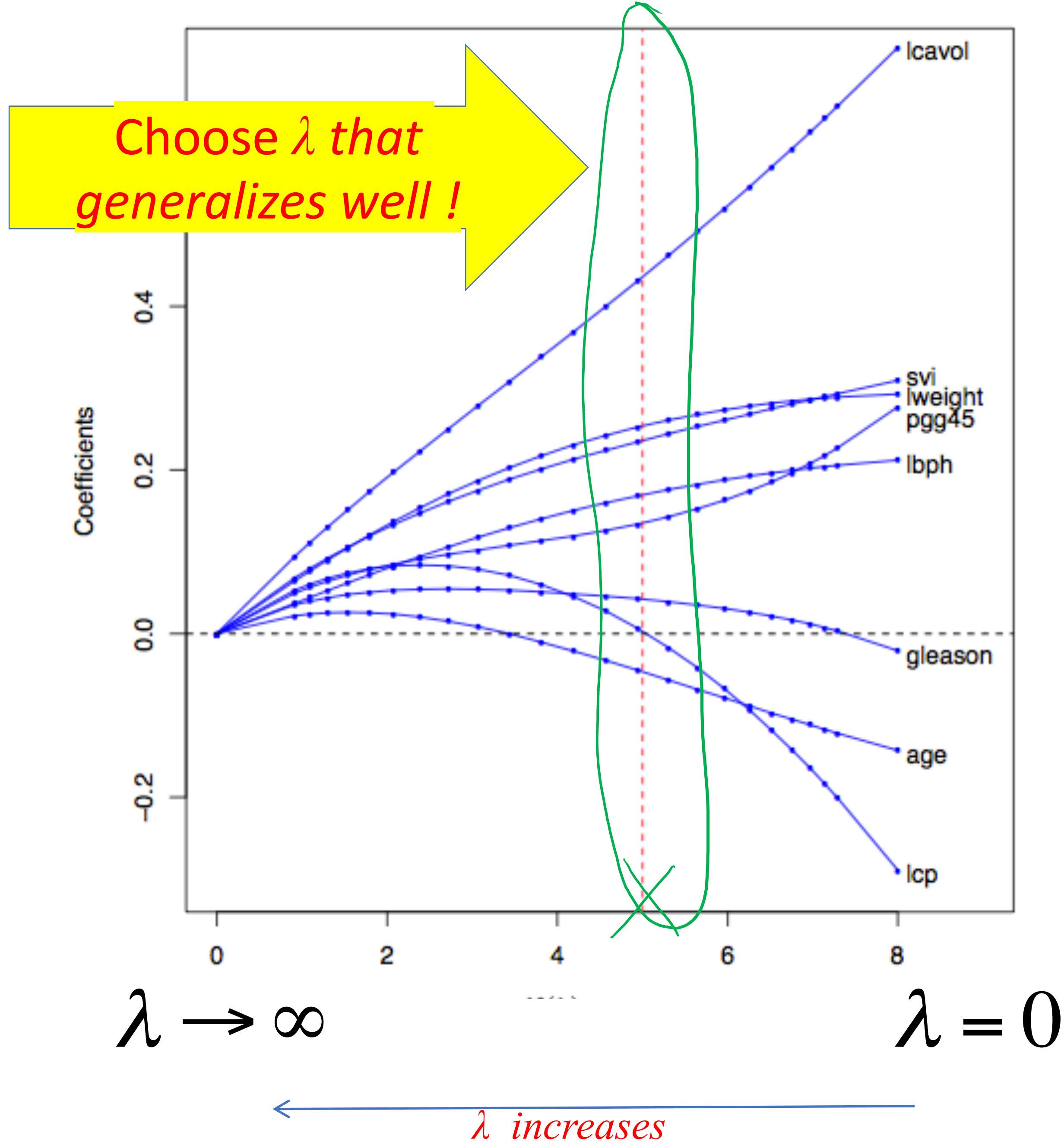
Lasso Estimator



An example with 8 features

An example of Ridge Regression

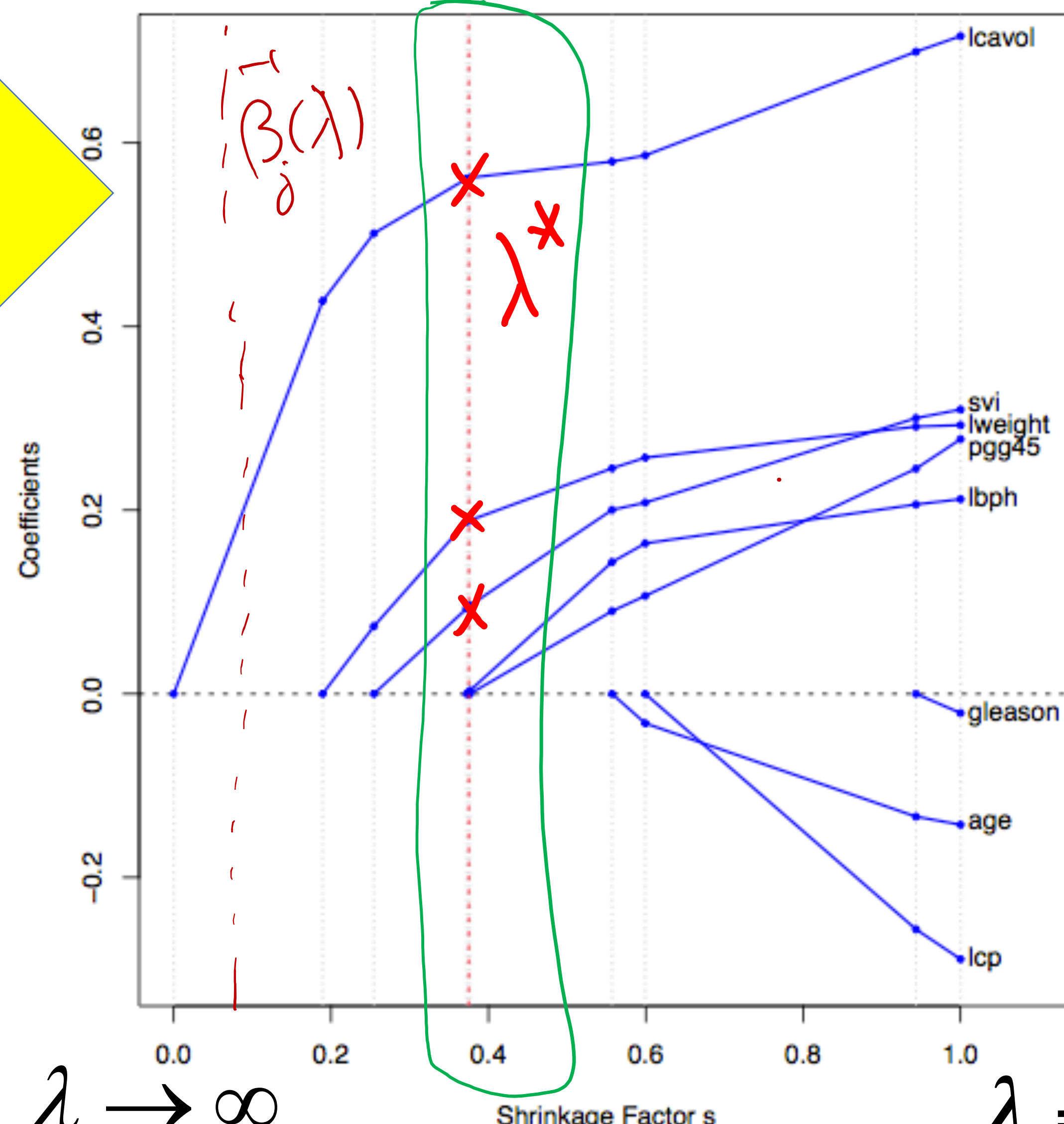
when varying λ, *how* β_j *varies.*



An example with 8 features

Choose λ *that generalizes well !*

when varying λ,
how β_j *varies.*



$\lambda \to \infty$ $\lambda = 0$

FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus* $s = t/\sum_1^p |\hat{\beta}_j|$. *A vertical line is drawn at* $s = 0.36$, *the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

An example with 8 features

Today Recap

- ❑ Linear Regression Model with Regularizations

✓Review: (Ordinary) Least squares: squared loss (Normal Equation)

✓Ridge regression: squared loss with L2 regularization

✓Lasso regression: squared loss with L1 regularization

✓Elastic regression: squared loss with L1 AND L2 regularization

✓Influence of Regularization Parameter

Regression (supervised)

- Four ways to train / perform optimization for linear regression models
  - Normal Equation
  - Gradient Descent (GD)
  - Stochastic GD
  - Newton's method

variations of $\underset{\theta}{\text{argmin}}\, L(\theta)$

- Supervised regression models
  - Linear regression (LR)
  - LR with non-linear basis functions
  - Locally weighted LR
  - LR with Regularizations

variations of $f(x)$

variations of $L(\theta)$

Extra More

- Optimization of regularized regressions:
  - See L6-extra slide
- Relation between λ *and* s
  - See L6-extra slide
- Why Elastic Net has a few nice properties
  - See L6-extra slide

References

- ❑ Big thanks to Prof. Eric Xing @ CMU for allowing me to reuse some of his slides
- ❑ Prof. Nando de Freitas's tutorial slide
- ❑ **Regularization and variable selection via the elastic net,** Hui Zou and Trevor Hastie, *Stanford University, USA*
- ❑ *ESL book: Elements of Statistical Learning*