UVA CS 4501: Machine Learning

Lecture 6: Linear Regression Model with Regularizations

Dr. Yanjun Qi

University of Virginia

Department of
Computer Science

Where are we ? ➔ Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

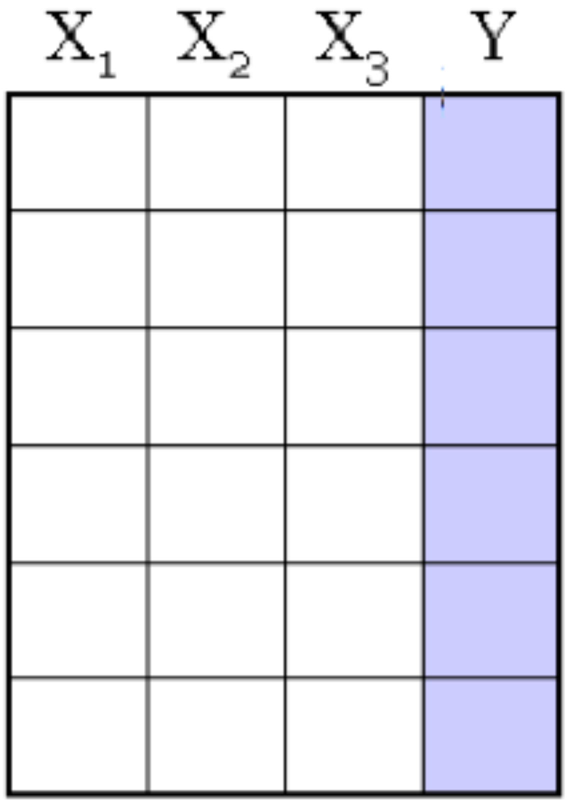
Today → Regression (supervised)

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

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

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
  - ✓ WHY and Influence of Regularization Parameter



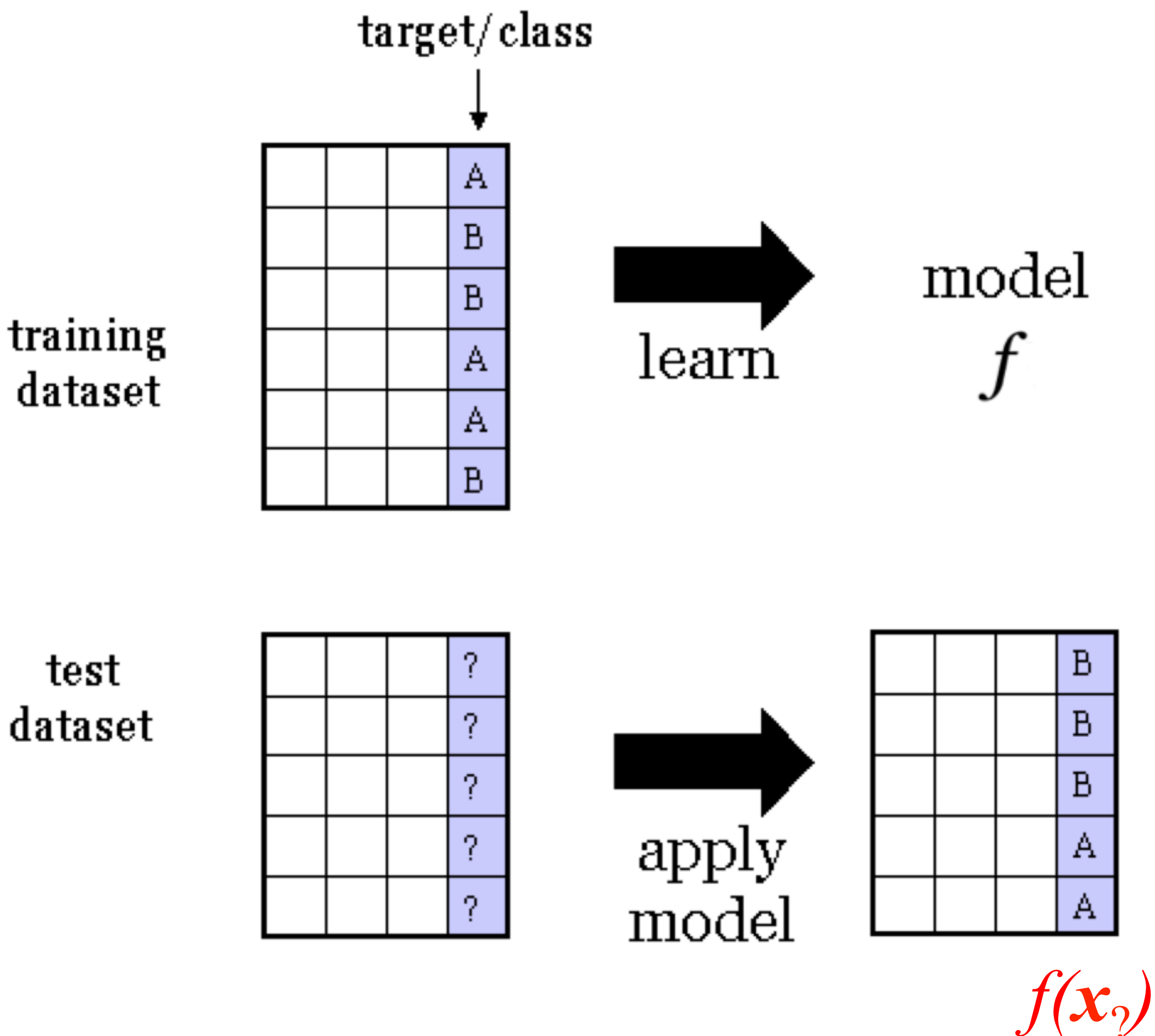
A Dataset for regression

$$f : \boxed{X} \longrightarrow \boxed{Y}$$

continuous valued variable

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/ predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

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:

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

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

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

→Add Regularization

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 rank(A). The following are some basic properties of the rank:

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \le \min(m, n)$ 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 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $\text{rank}(AB) \le \min(\text{rank}(A), \text{rank}(B))$. ①
- For $A, B \in \mathbb{R}^{m \times n}$, $\text{rank}(A + B) \le \text{rank}(A) + \text{rank}(B)$.

Page11 Of Handout

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

when $n < p$

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

singular / not invertible

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:

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

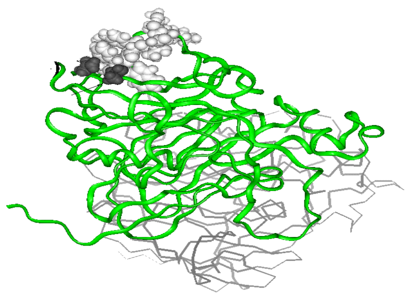
IV. Features

e.g. counts of a ngram in the text

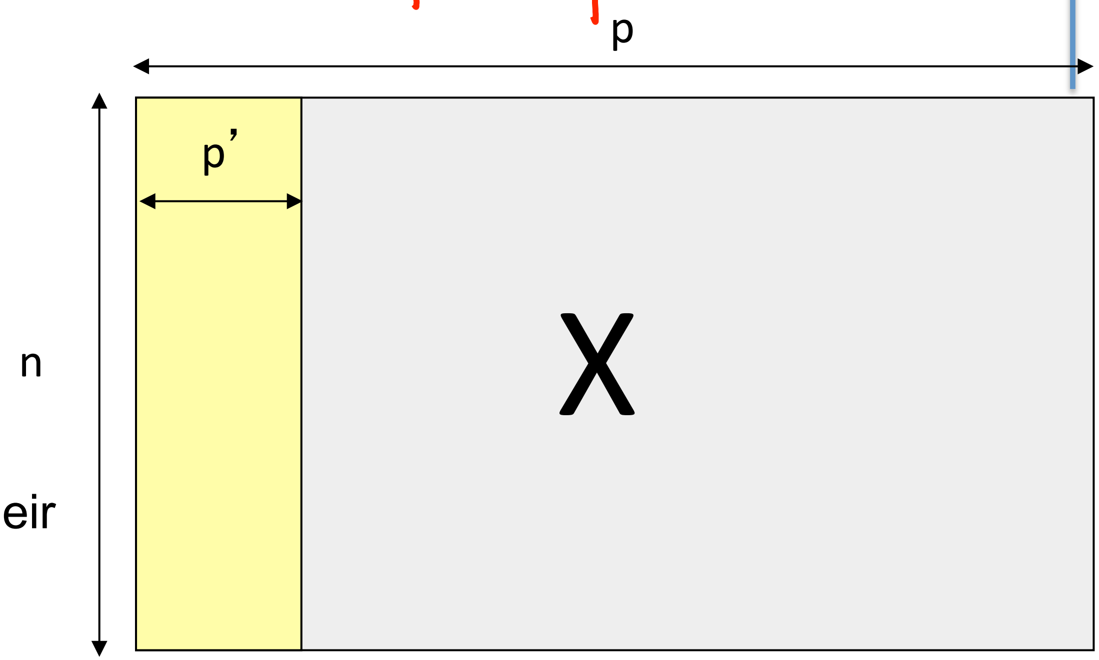
I	Lexical n-grams (1,2,3)
II	Part-of-speech n-grams (1,2,3)
III	Dependency relations (nsubj,advmod,...)
Meta	U.S. origin, running time, budget (log), # of opening screens, genre, MPAA rating, holiday release (summer, Christmas, Memorial day,...), star power (Oscar winners, high-grossing actors)

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

e.g., QSAR: Drug Screening



$p \to p'$ Select Featu



Binding to Thrombin (DuPont Pharmaceuticals)

- n: 2543 compounds tested for their ability to bind to a target site on thrombin (a human enzyme)

- **p: 139,351 binary features**, which describe three-dimensional properties of molecules.

Weston et al, Bioinformatics, 2002

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
  - ✓ Influence of Regularization Parameter

Review: Vector norms

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

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

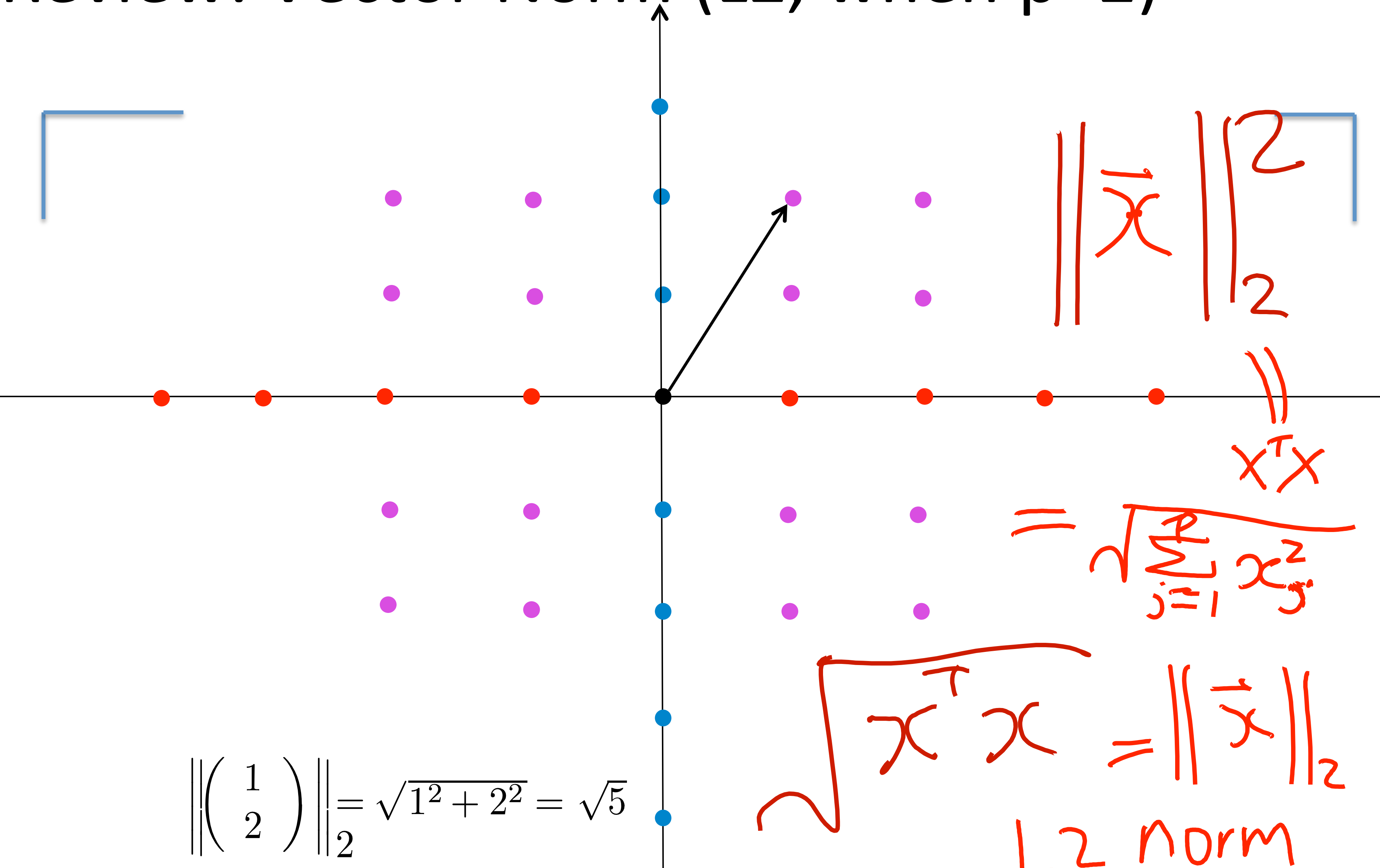
– Common norms: L_1, L_2 (Euclidean)

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

– $L_{infinity}$

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

Review: Vector Norm (L2, when p=2)



$$\|\vec{x}\|_2^2 = x^T x$$

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

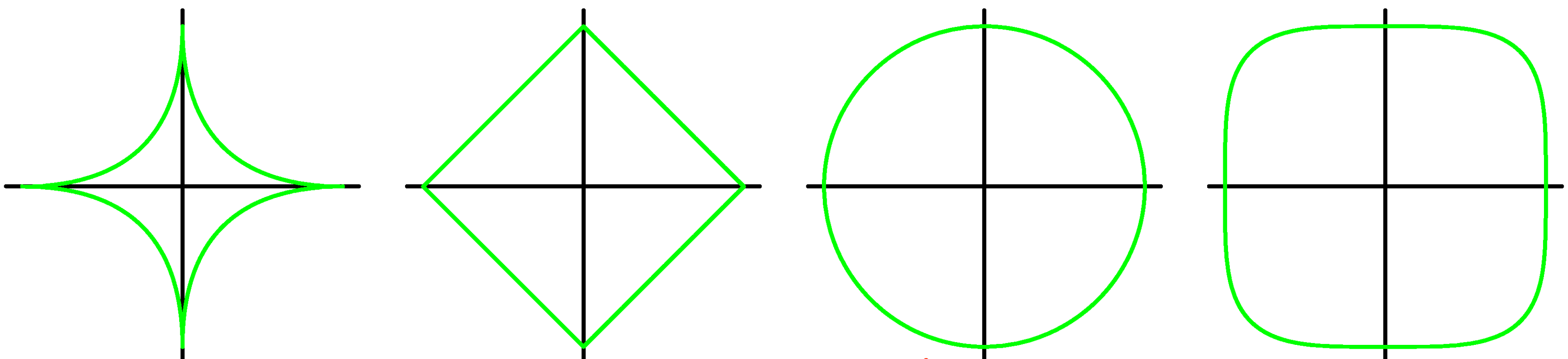
$$\sqrt{x^T x} = \|\vec{x}\|_2$$

L2 norm

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

Norms

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



Ridge Regression / L2 Regularization

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

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

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

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

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

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

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

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

$\Downarrow$

$X^T X$ 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

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

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

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

- As the solution from

HW2

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

to minimize, take derivative and set to zero

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

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

circle with radius s

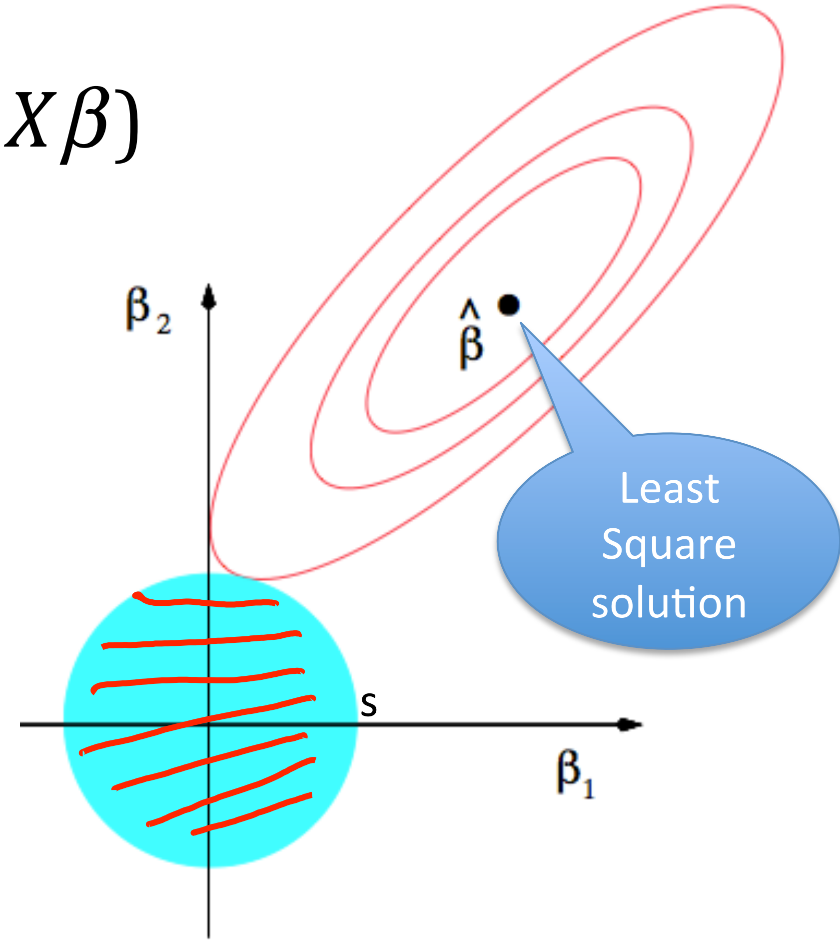
By convention, the bias/intercept term is typically not regularized.
Here we assume data has been centered … therefore no bias term

Objective Function's Contour lines from Ridge Regression

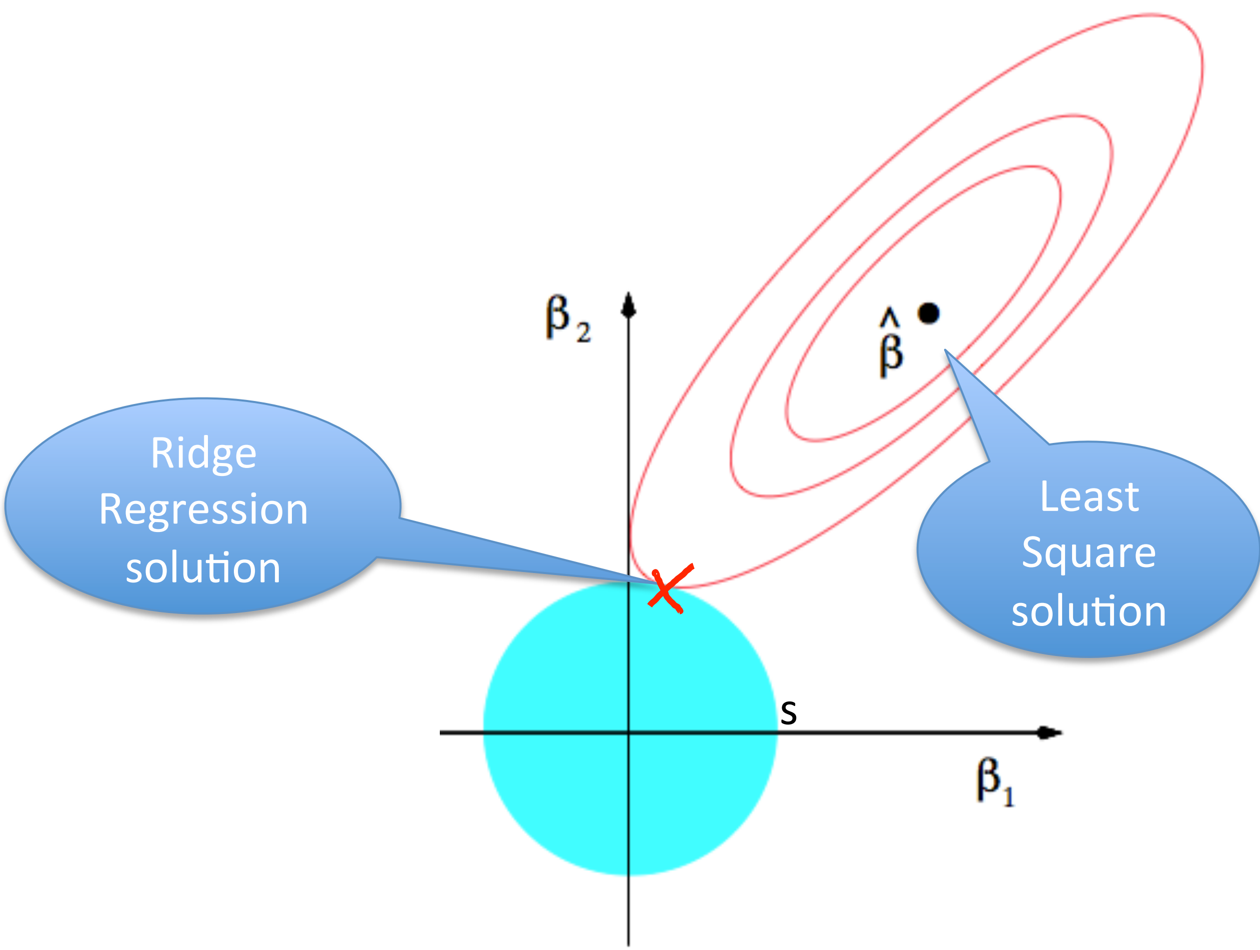
$$\hat{\beta}^{ridge} = \operatorname{argmin}(y - X\beta)^T (y - X\beta)$$

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

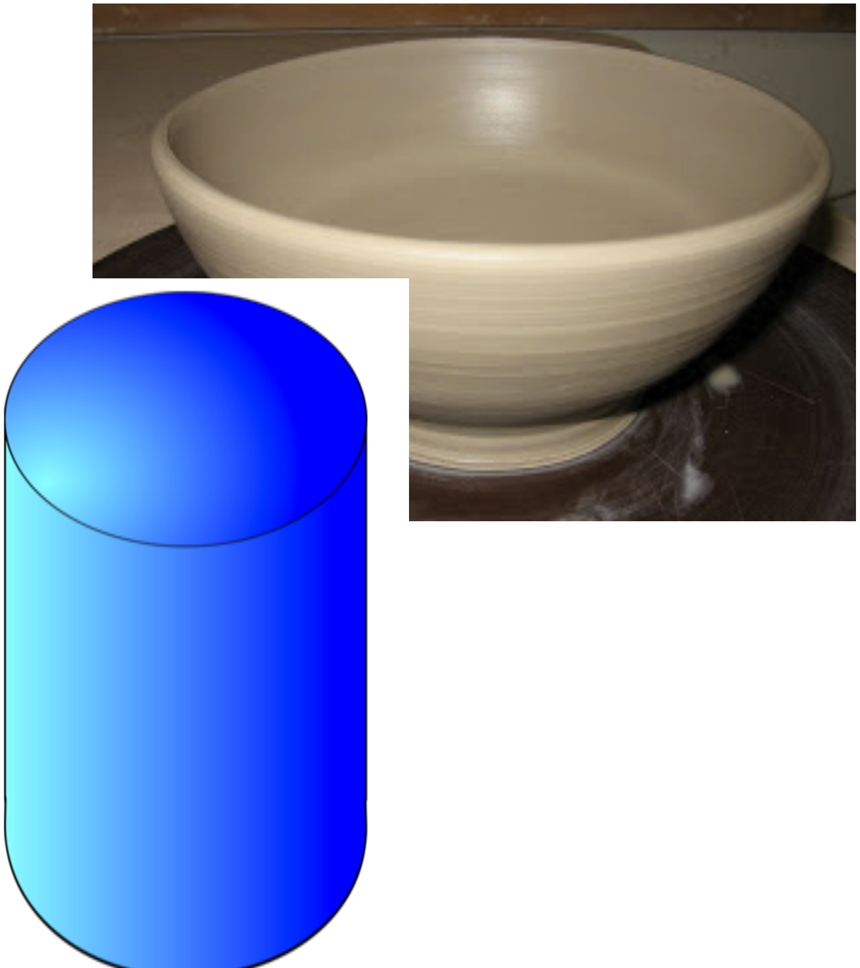
circle with radial s

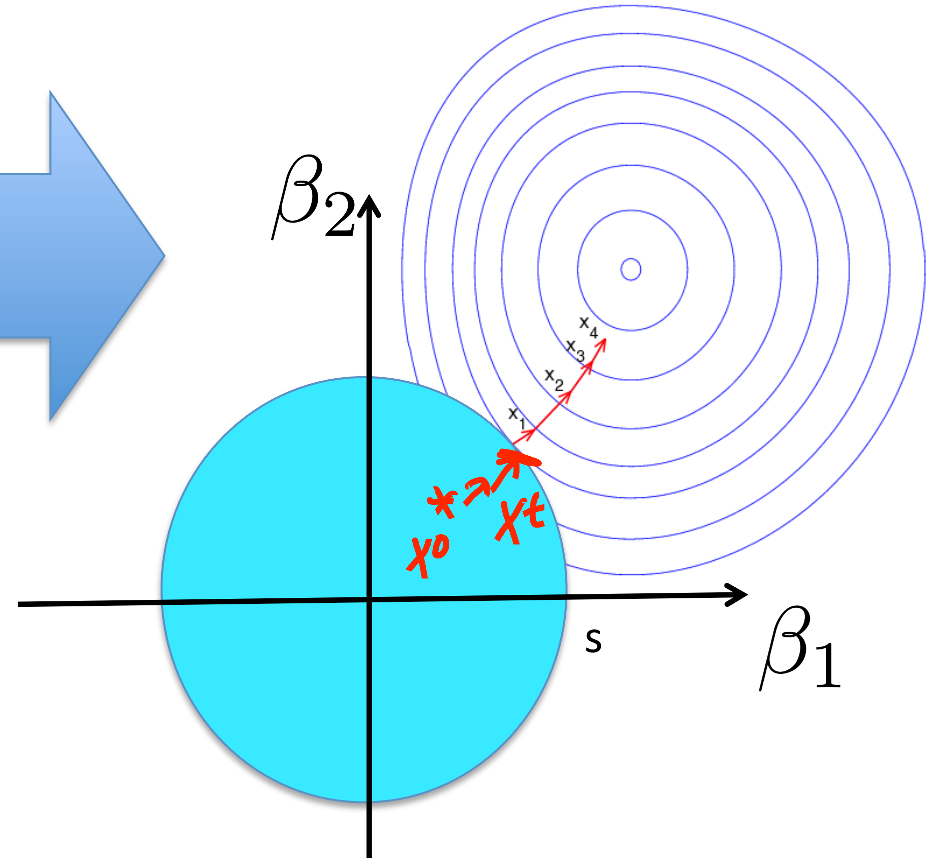


Objective Function's Contour lines from Ridge Regression



Least Square+L2:
Ridge solution





must within the circle

Ridge Regression / Squared Loss+L2

- λ > 0 penalizes each θ_j
  - See whiteboard

- if λ = 0 we get the least squares estimator;
- if $\lambda \to \infty$, then $\theta j \to 0$

Parameter Shrinkage

$\lambda > 0$

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

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

$\lambda > 0$

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

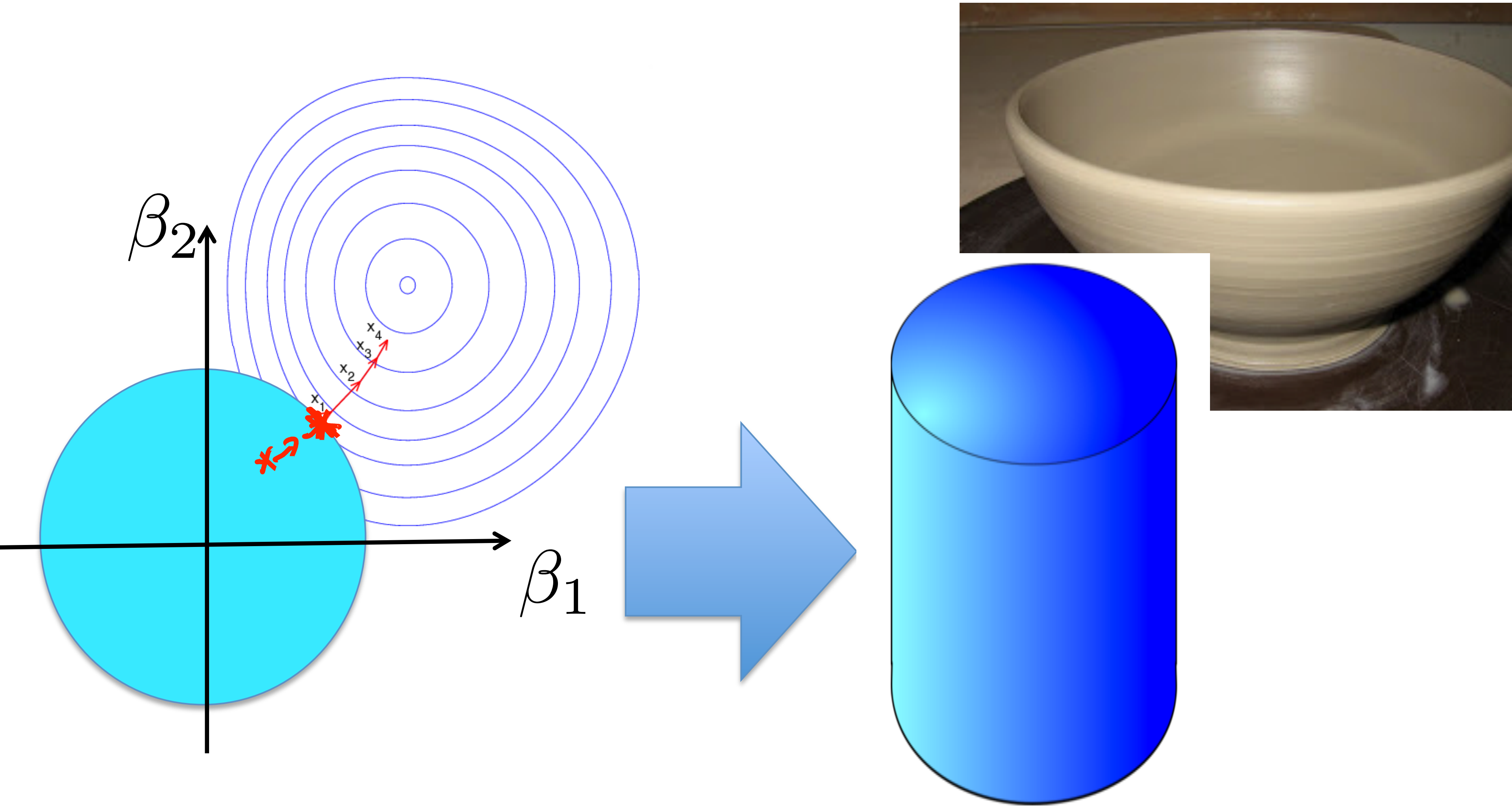
when $X^T X = I$ $\Rightarrow$ $\boxed{\beta_{Rg}} = \frac{1}{1+\lambda} X^T y = \boxed{\frac{1}{1+\lambda}} \beta_{OLS}$

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

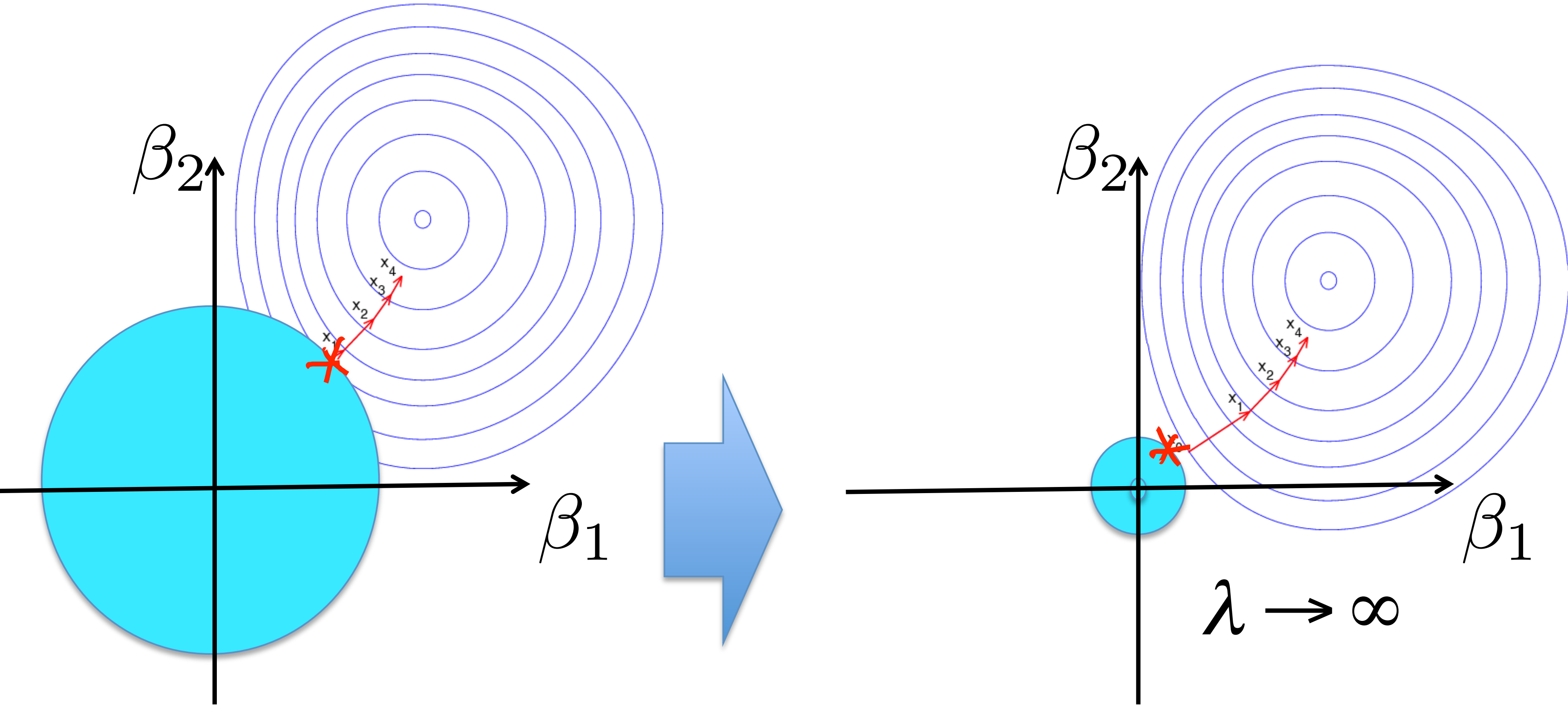
When $X^T X$ general case, see advanced analysis @

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

✓Influence of Regularization Parameter

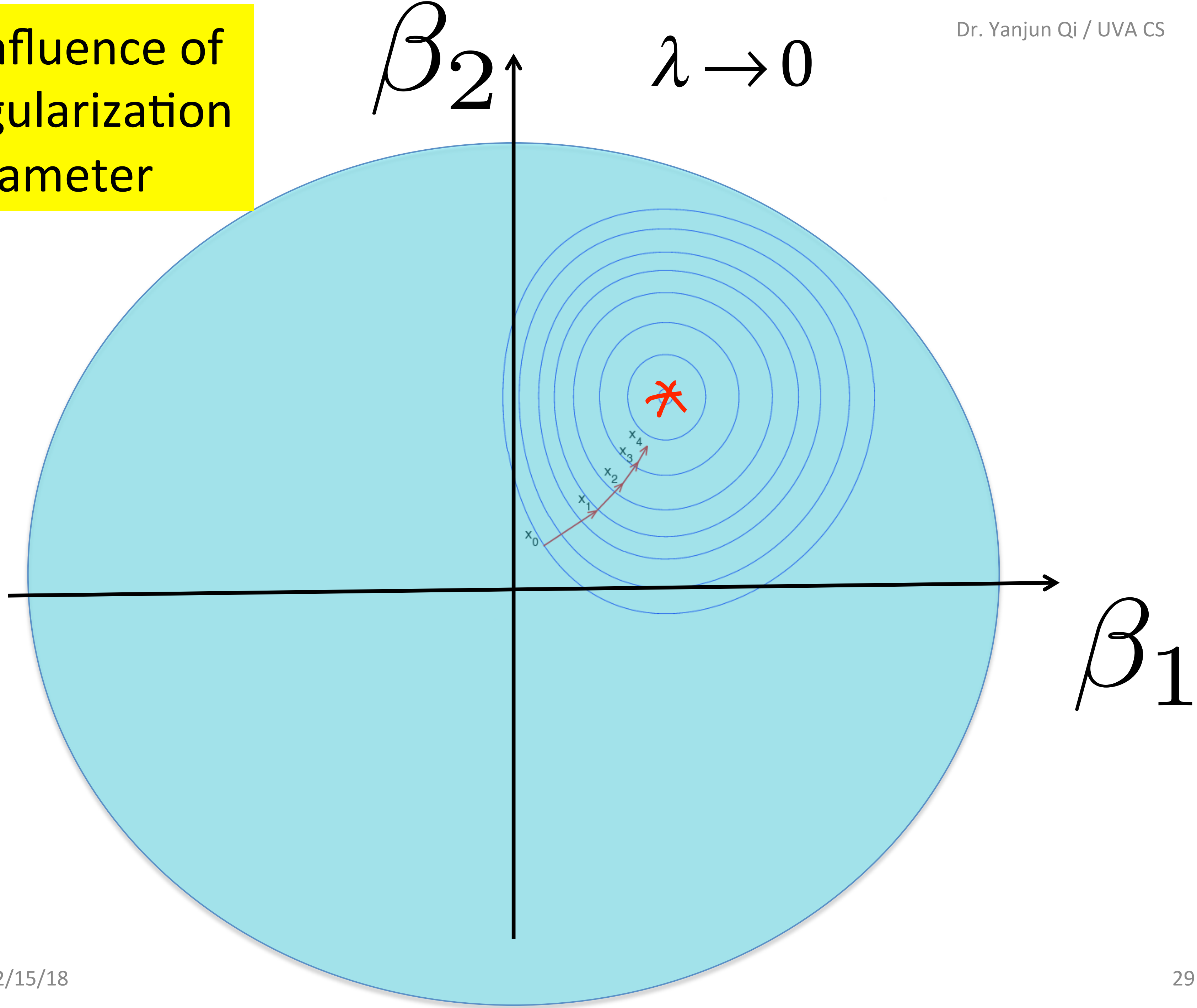


✓Influence of Regularization Parameter



✓Influence of Regularization Parameter

$\lambda \rightarrow 0$



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
- ✓ WHY and Influence of 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

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

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

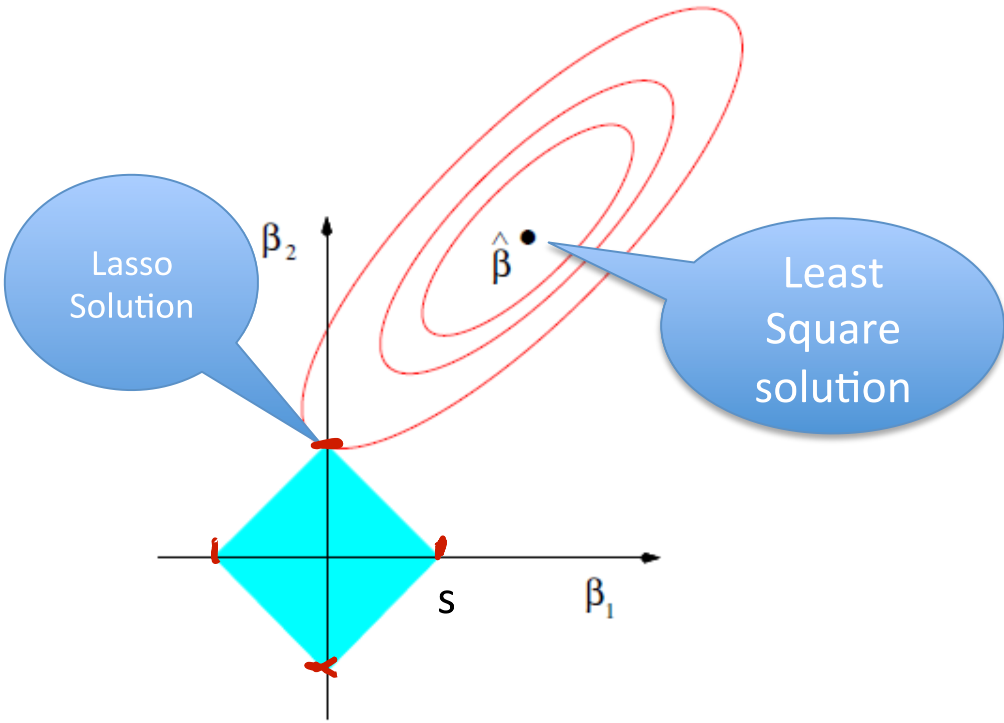
$$\text{subject to } \underbrace{\sum \left|\beta_j\right|}_{\text{L1 norm}} \leq s$$

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

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

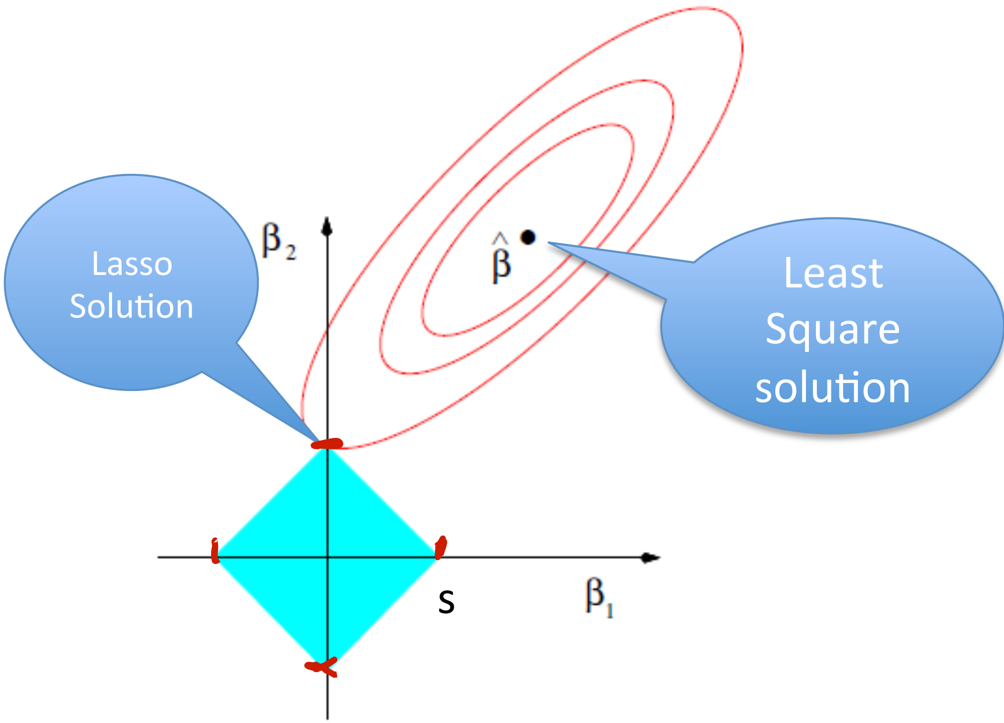
- Suppose in 2 dimension
- $\beta = (\beta_1, \beta_2)$
- $|\beta_1| + |\beta_2| = \text{const}$
- $|\beta_1| + |-\beta_2| = \text{const}$
- $|-\beta_1| + |\beta_2| = \text{const}$
- $|-\beta_1| + |-\beta_2| = \text{const}$

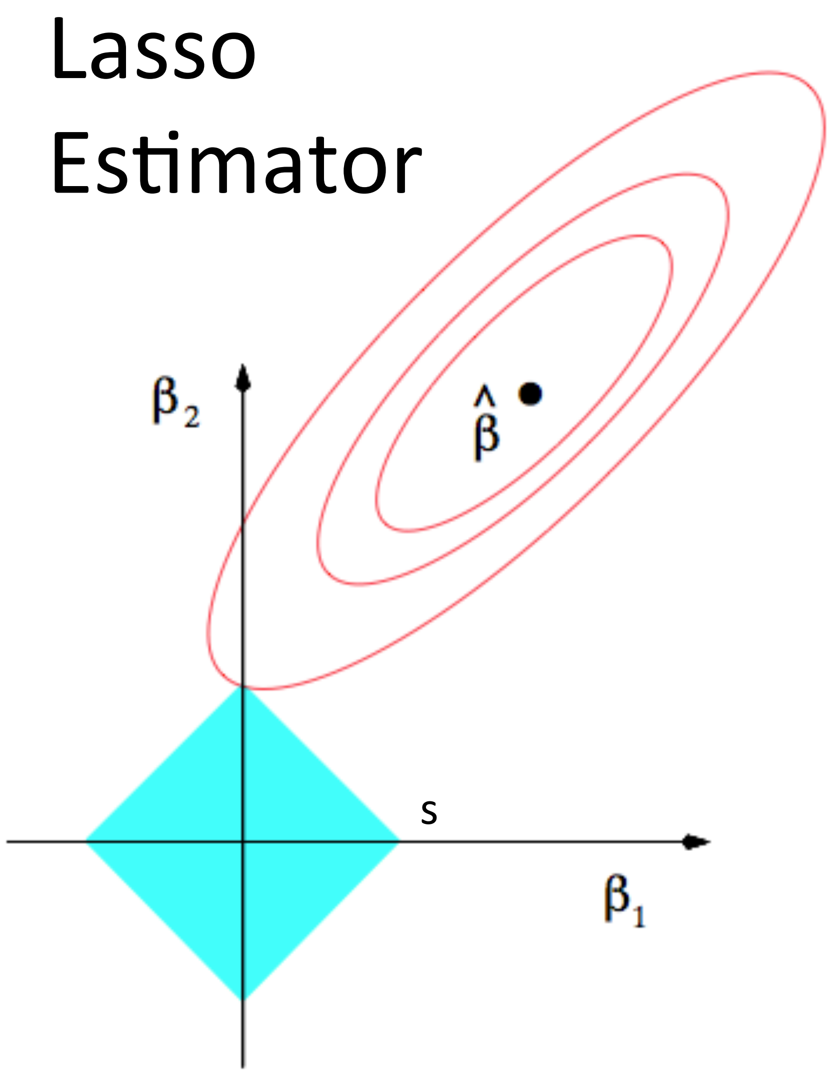


$$\hat{y} = \sum_{j=1}^{p} \beta_j x_j$$

when many β_j are zero $\Rightarrow$ select feature

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





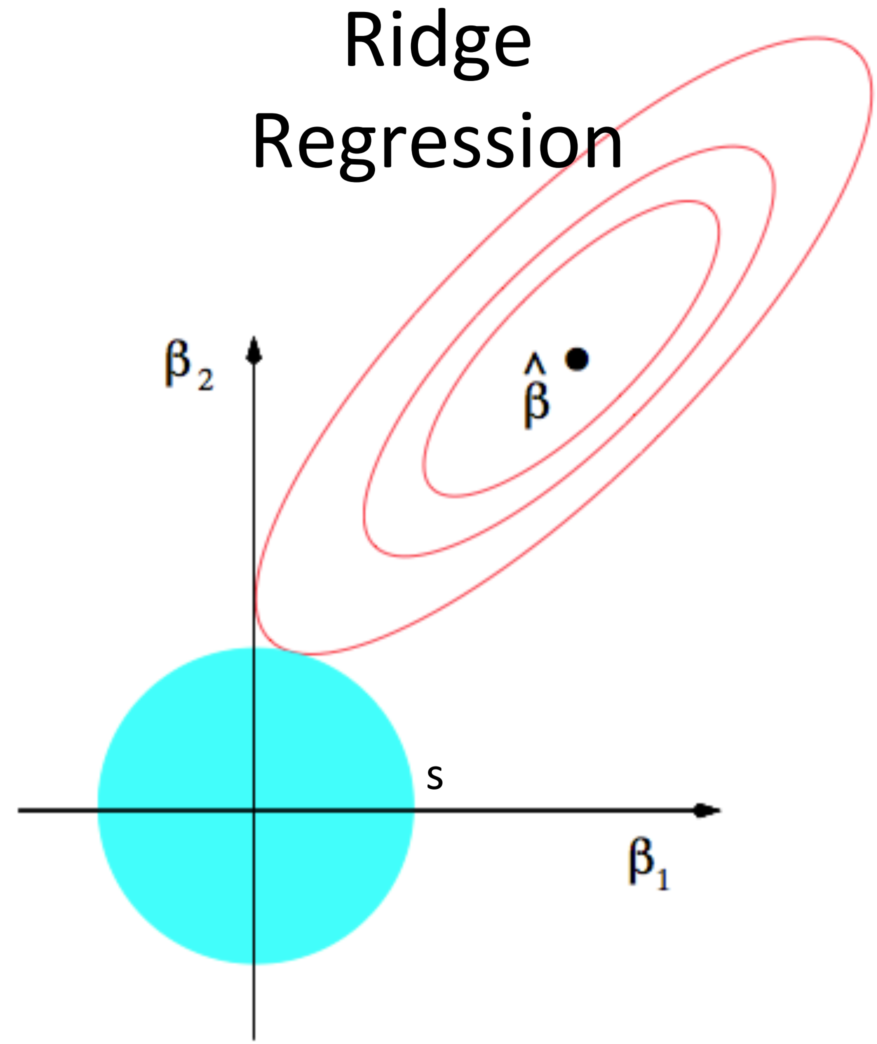


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \le t$ *and* $\beta_1^2 + \beta_2^2 \le t^2$*, respectively, while the red ellipses are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

- 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 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.

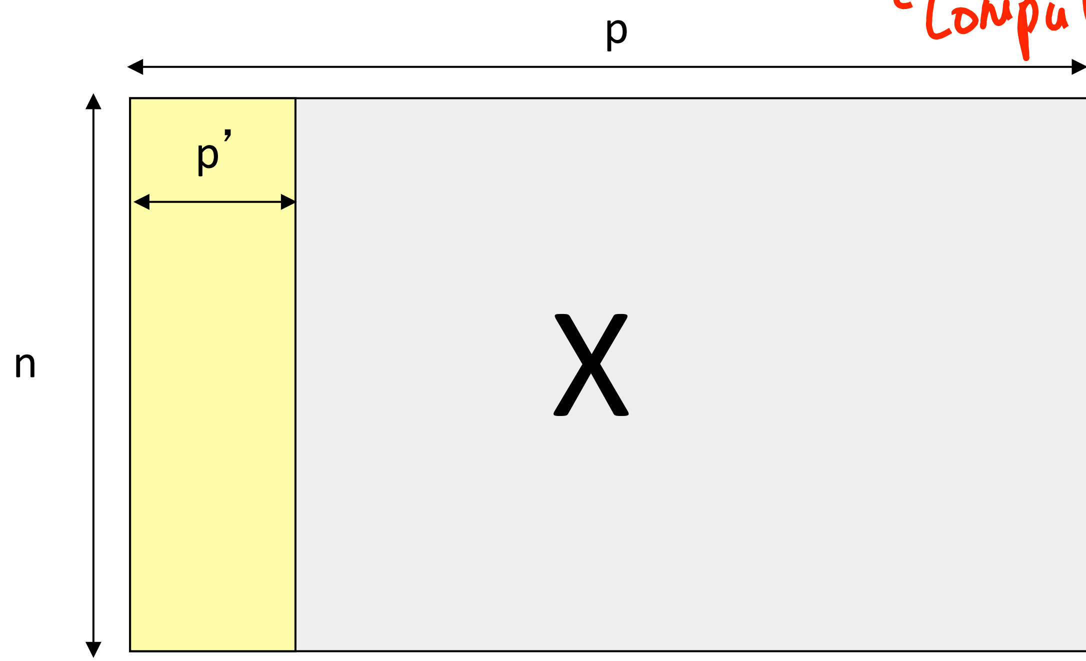
Disadvantage:

-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

Lasso: Implicit Feature Selection

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



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

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

Computationally,

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

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

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

$\Rightarrow$ choose to make $p \downarrow$ if we can ☆

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
- ✓ WHY and Influence of Regularization Parameter

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

- L1 part of the penalty generates a sparse model
- L2 part of the penalty:
  - 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:

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

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

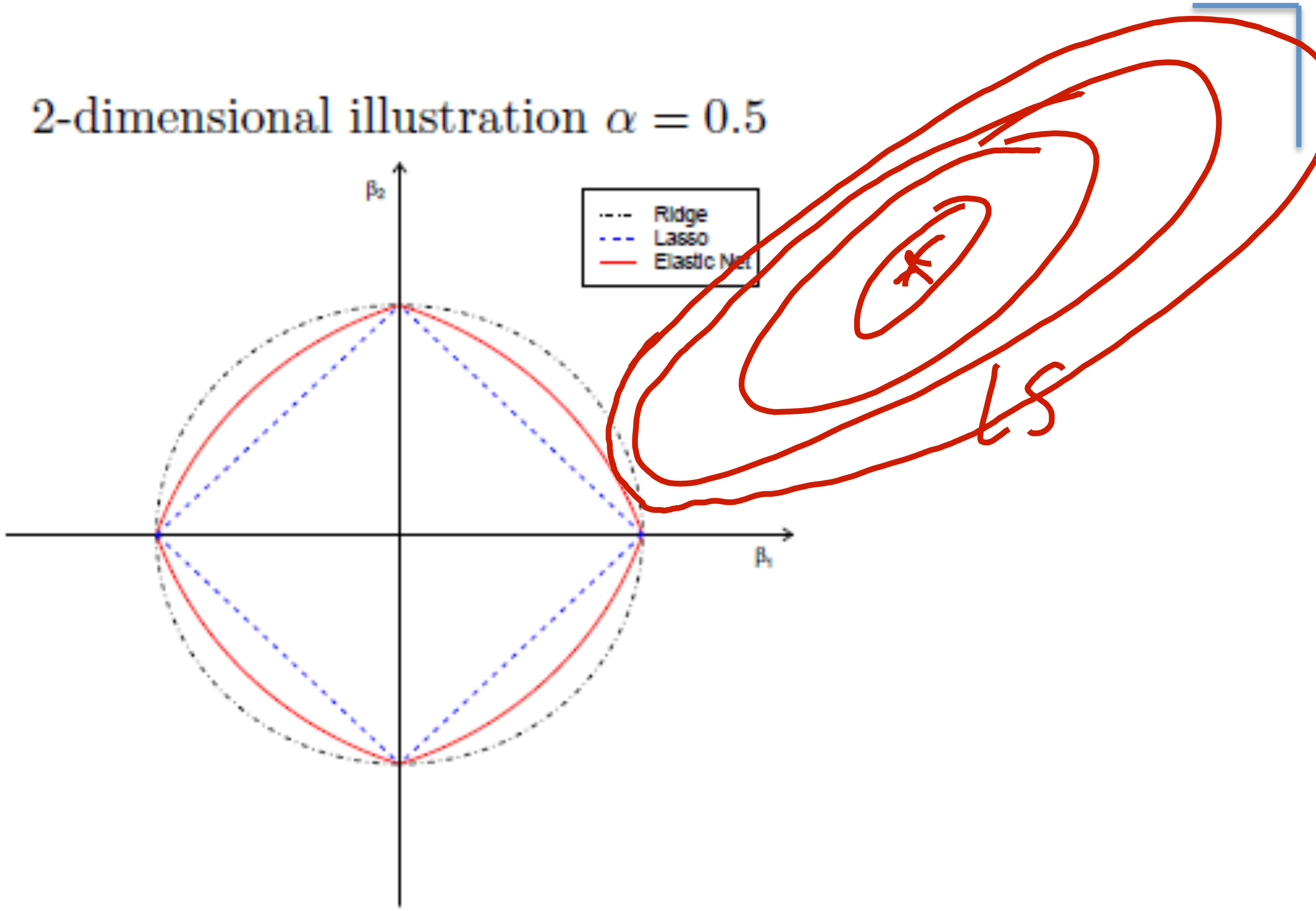
- The naive elastic net estimator is the minimizer of equation

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

- Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$ $\frac{\lambda_2}{\lambda_1+\lambda_2}$

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

Geometry of elastic net



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 y, based on features x extracted from the movie metadata and/or the text of the reviews.

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

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)**

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

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

✓ WHY and Influence of Regularization Parameter

Why to Use simpler models?

- Because:
  - Simpler to use (lower computational complexity)
  - Easier to train (needs less examples)
  - Less sensitive to noise
  - Easier to explain (more interpretable)
  - Generalizes better (lower variance - Occam's razor)

    --- More in future lectures!!!

Adapted from Dr. Christoph Eick Slides

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

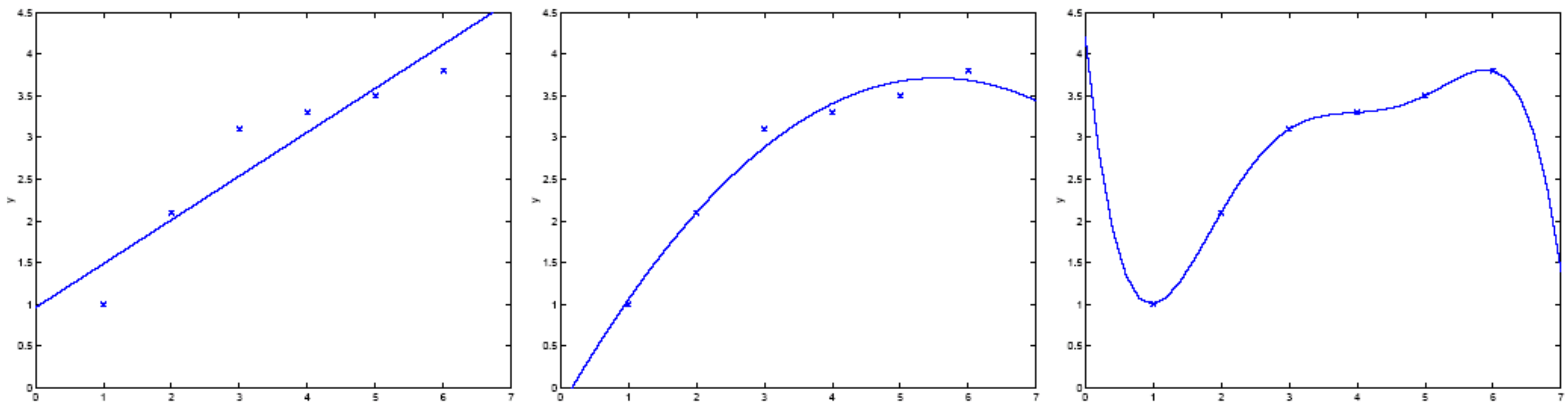
- Overfitting: when model is too complex and test errors are large although training errors are small.
  – After learning knowledge, model tends to learn "noise"

Issue: Overfitting and underfitting

Under fit

Looks good

Over fit



$y = \theta_0 + \theta_1 x$

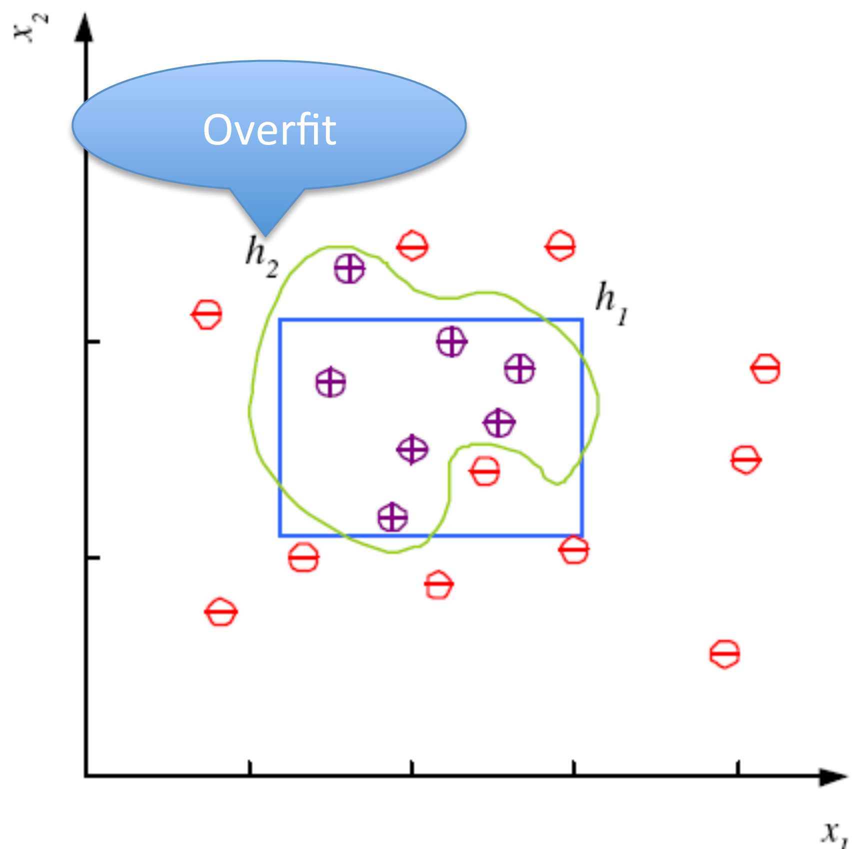
$y = \theta_0 + \theta_1 x + \theta_2 x^2$

$y = \sum_{j=0}^{5} \theta_j x^j$

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

K-fold Cross Validation !!!!

Overfitting and. Underfitting



Overfitting ----- 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.,

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

Common regularizers

$$\sum_j |\beta_j|$$

L2: Squared weights penalizes large values more
L1: Sum of weights will penalize small values more

$$\sum_j \beta_j^2$$

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

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: $\arg\min \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$
  - Ridge regression estimation: $\arg\min \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

53/54

Error on data + Regularization

More: A family of shrinkage estimators

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

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

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

convex

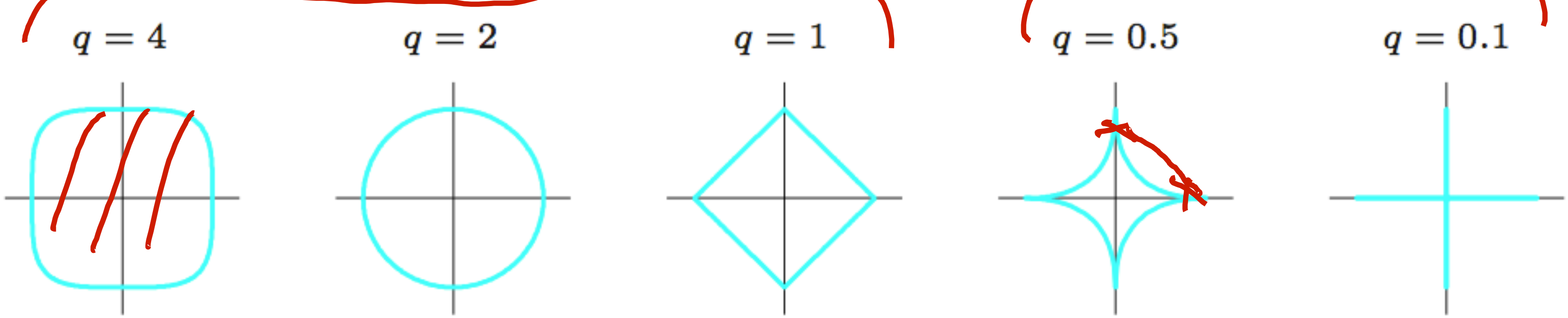
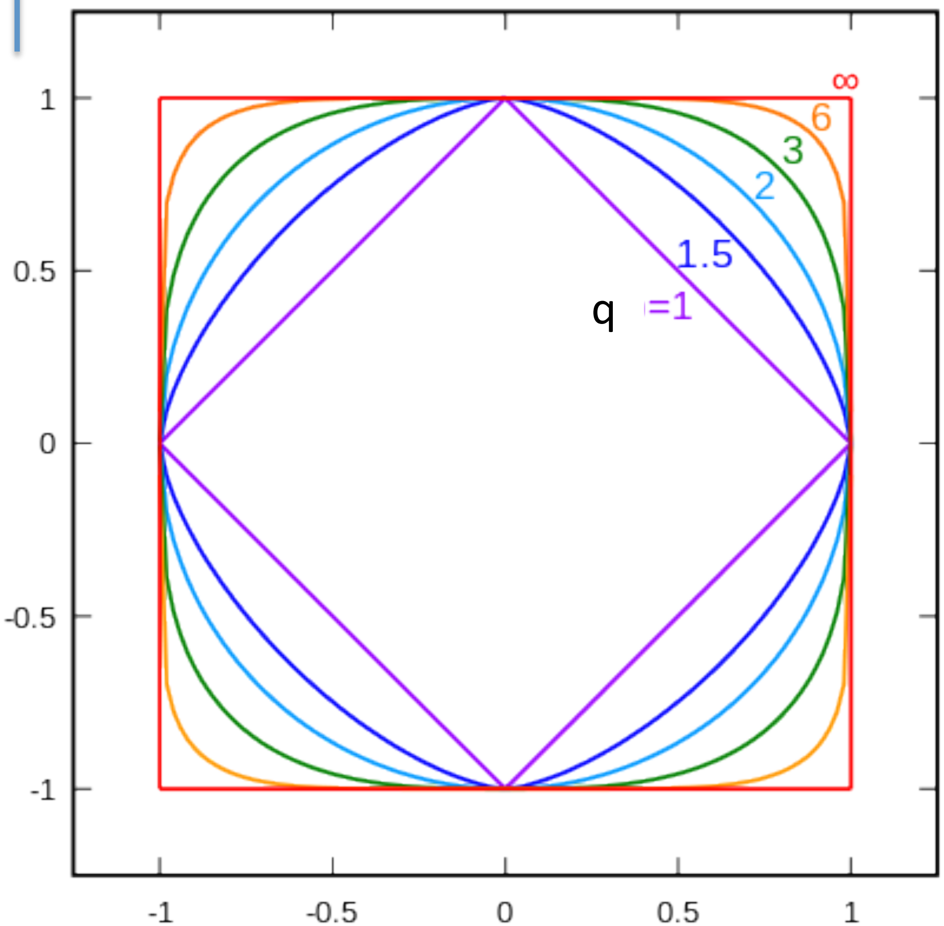


FIGURE 3.12. *Contours of constant value of $\sum_j |\beta_j|^q$ for given values of q.*

norms visualized

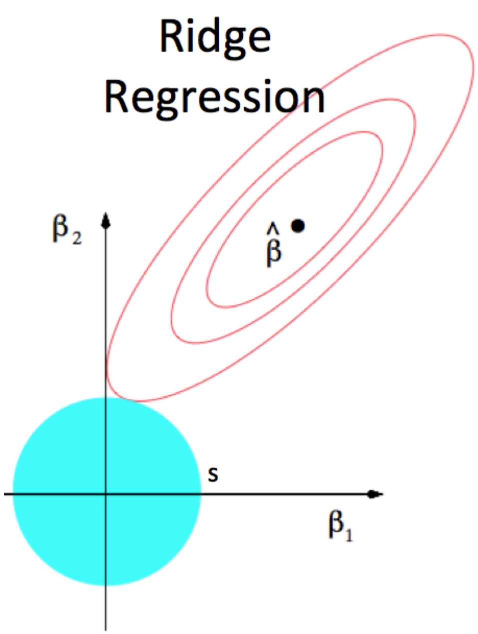


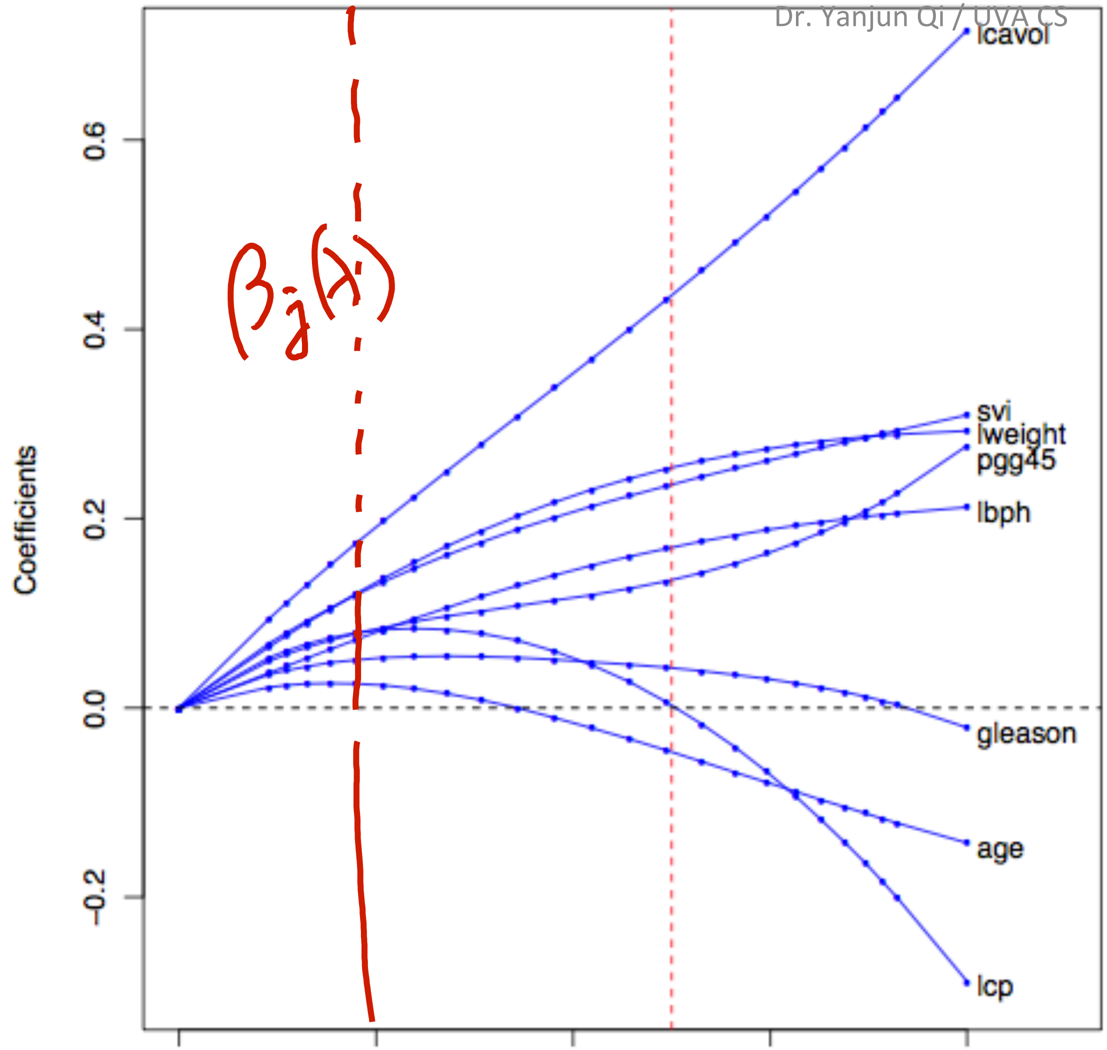
all p-norms penalize larger weights

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

q > 2 tends to push for similar weights

Regularization path of a Ridge Regression





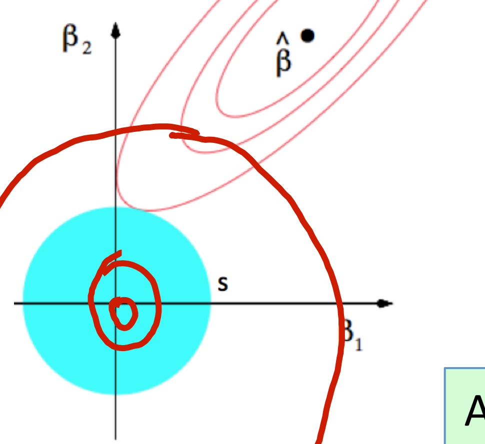
$\lambda \to \infty$ $\lambda = 0$

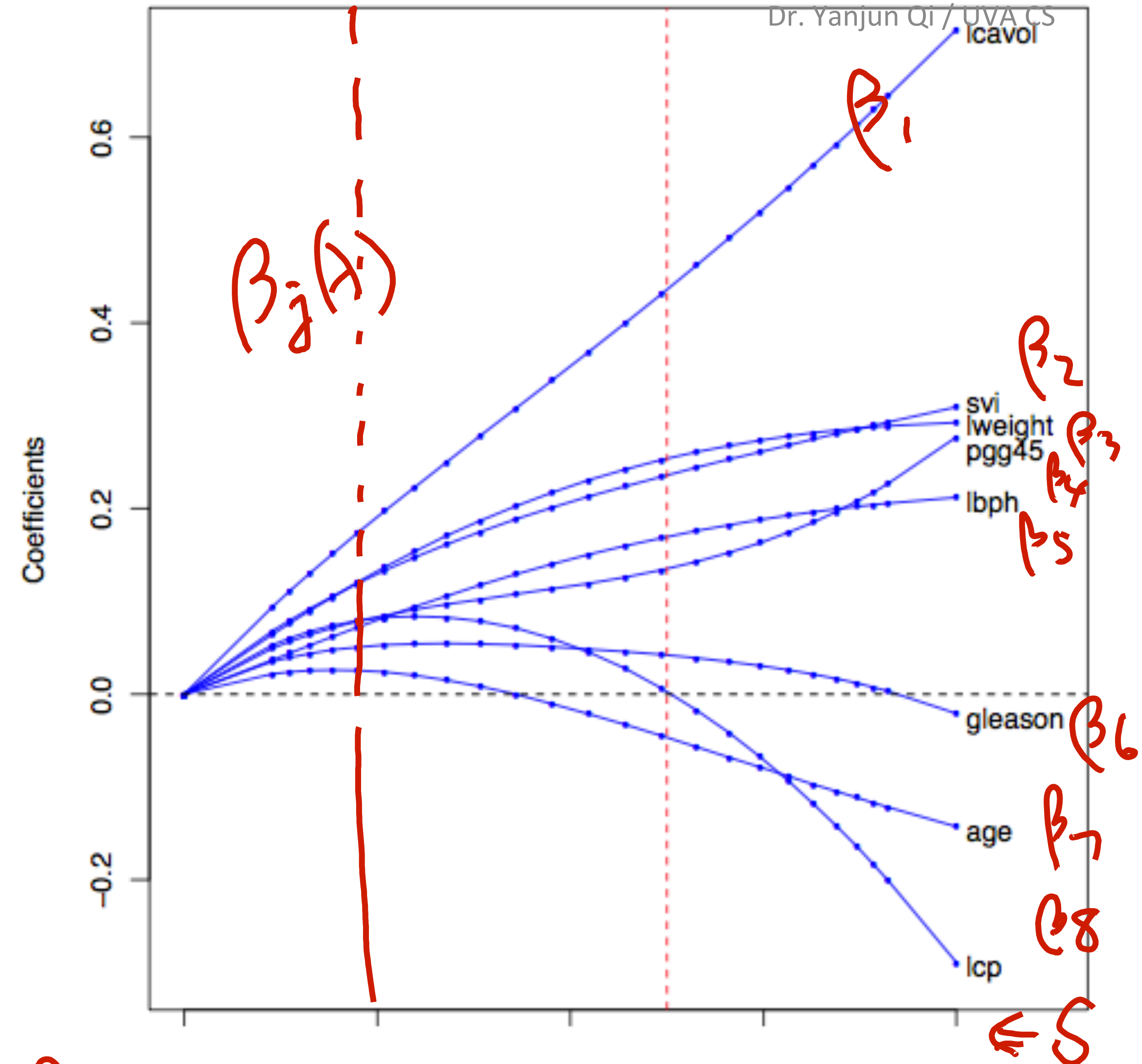
Regularization path of a Ridge Regression

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

Weight Decay

Ridge Regression



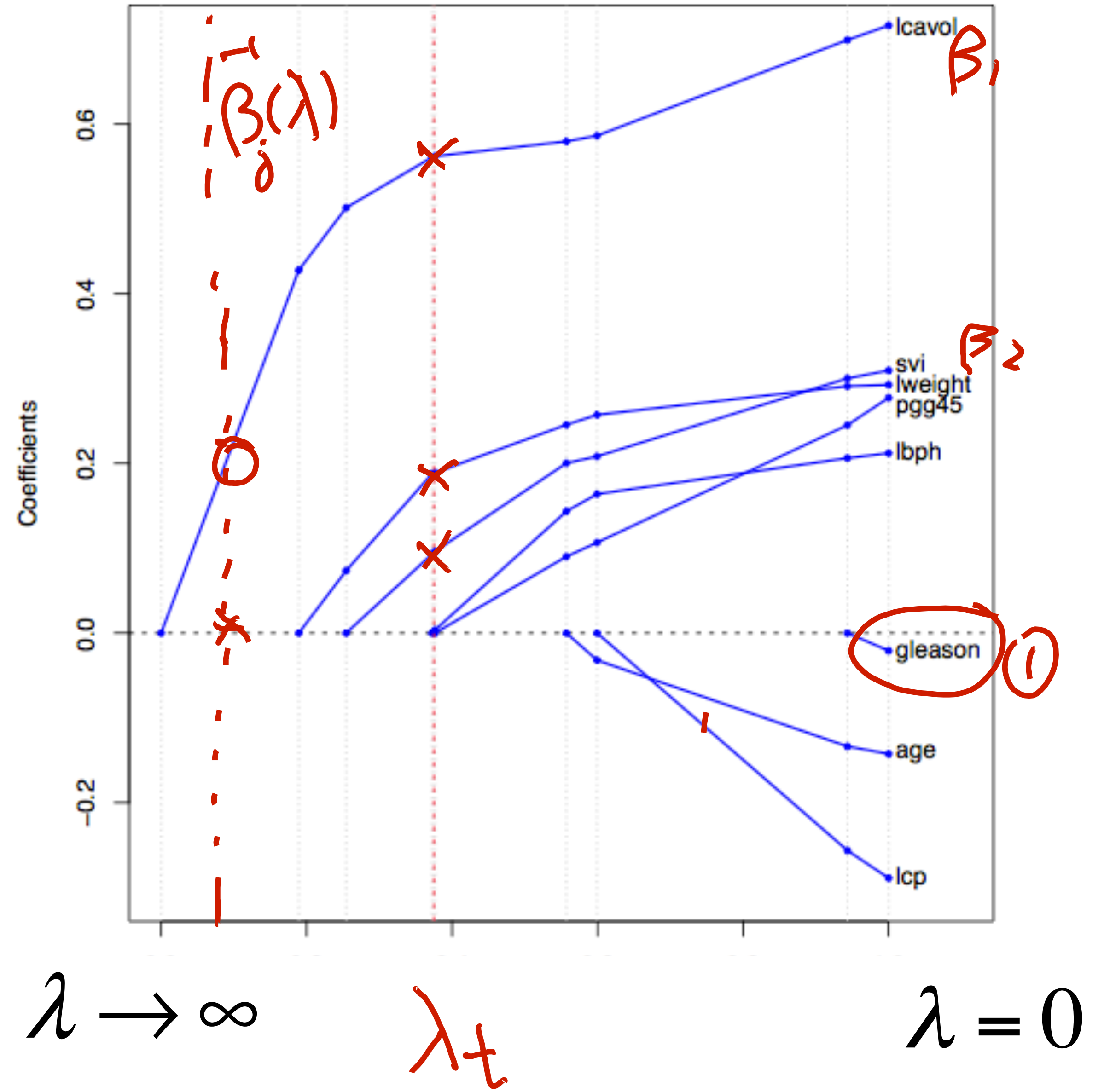


λ ←

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

An example with 8 features

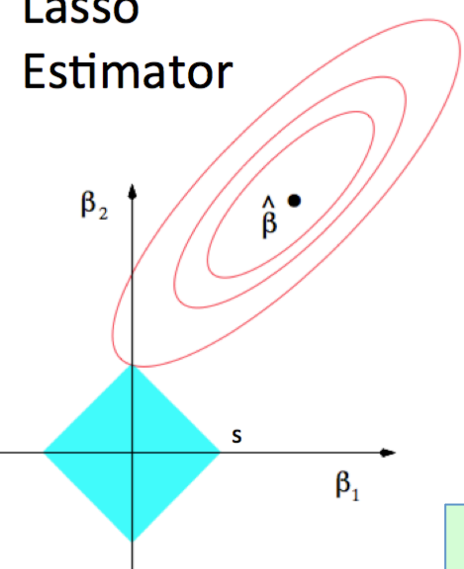
Regularization path of a Lasso Regression



$p=8$

when varying λ,
how β_j varies.

Lasso Estimator



An example with 8 features

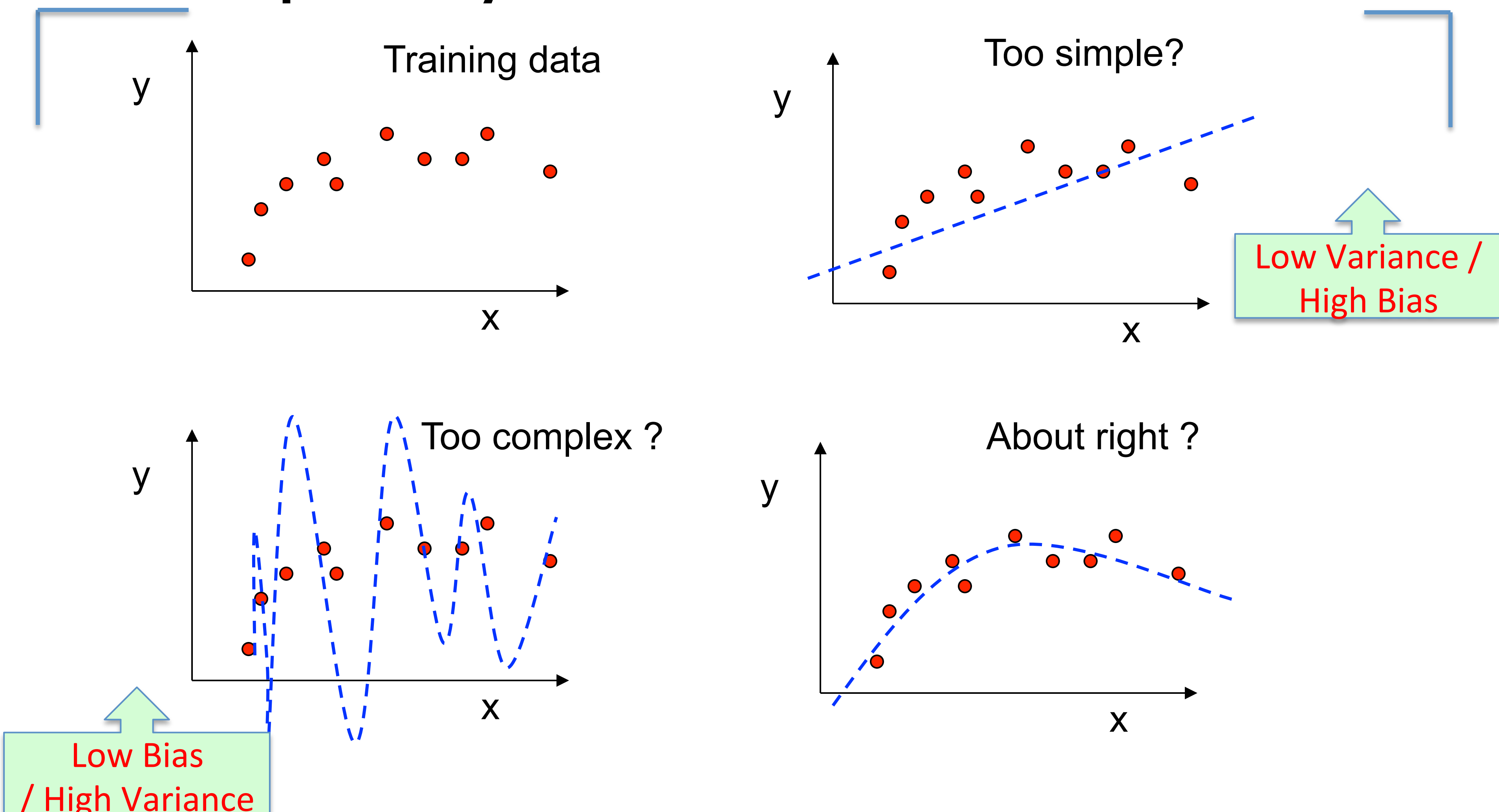
WHY and How to Select λ?

- 1. Generalization ability

  → k-folds CV to decide

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

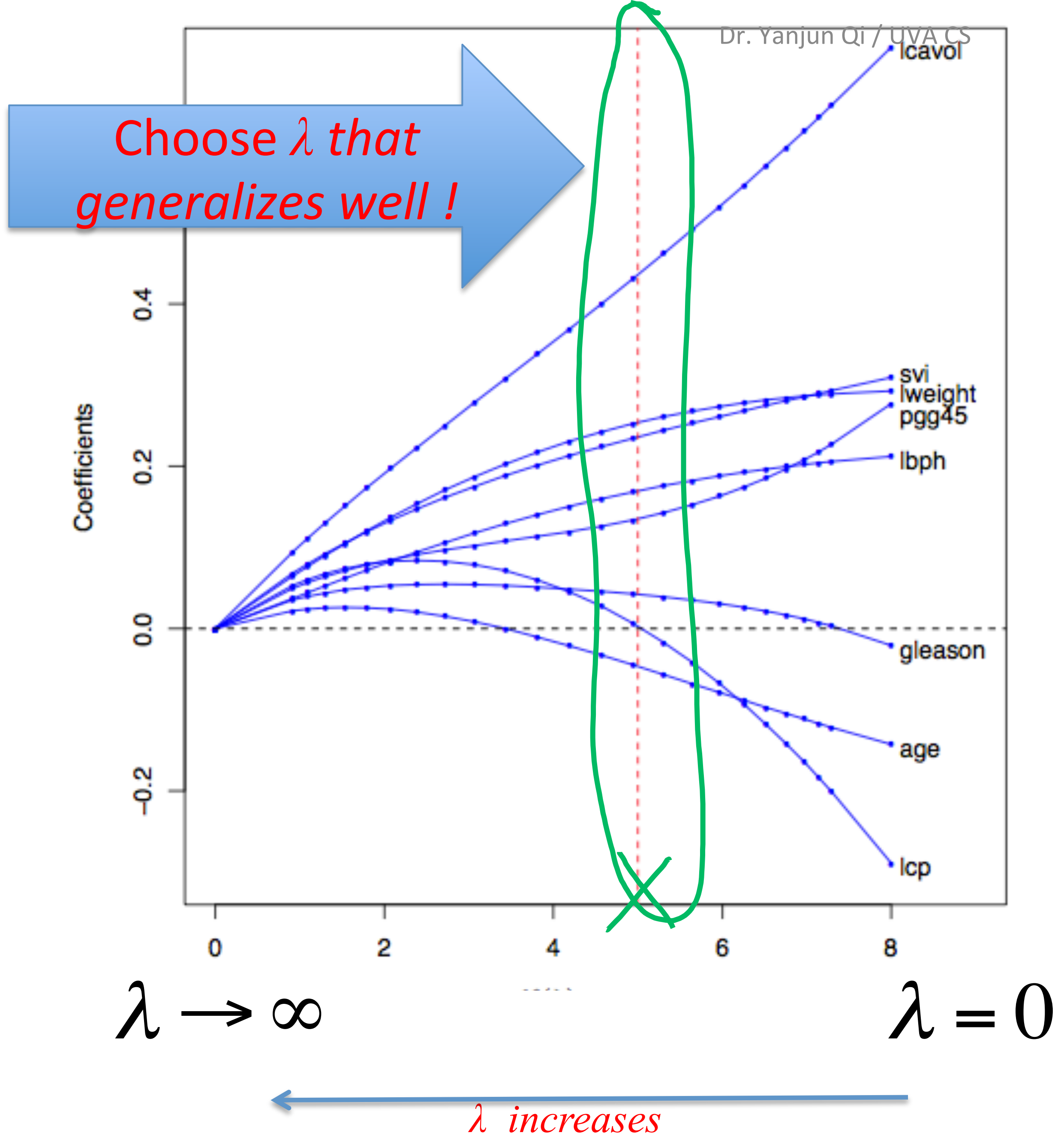
Regression: Complexity versus Goodness of Fit



What ultimately matters: ***GENERALIZATION***

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.*

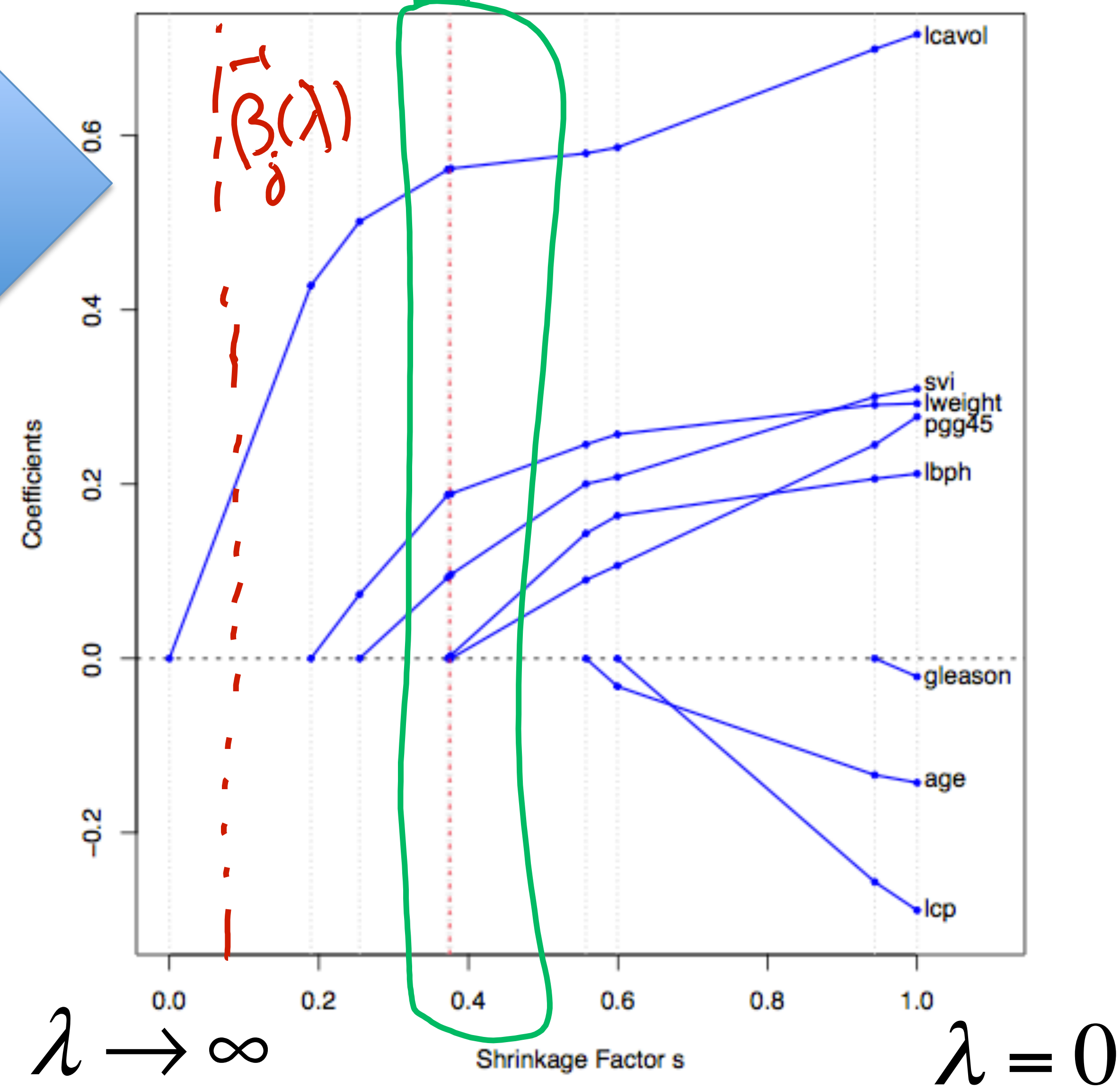


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for 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

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*

More

- Optimization of regularized regression:
  - 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