UVA CS 6316/4501 – Fall 2016 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
  - Ridge Regression
  - Lasso Regression
  - Elastic net

Review: Vector norms

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

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

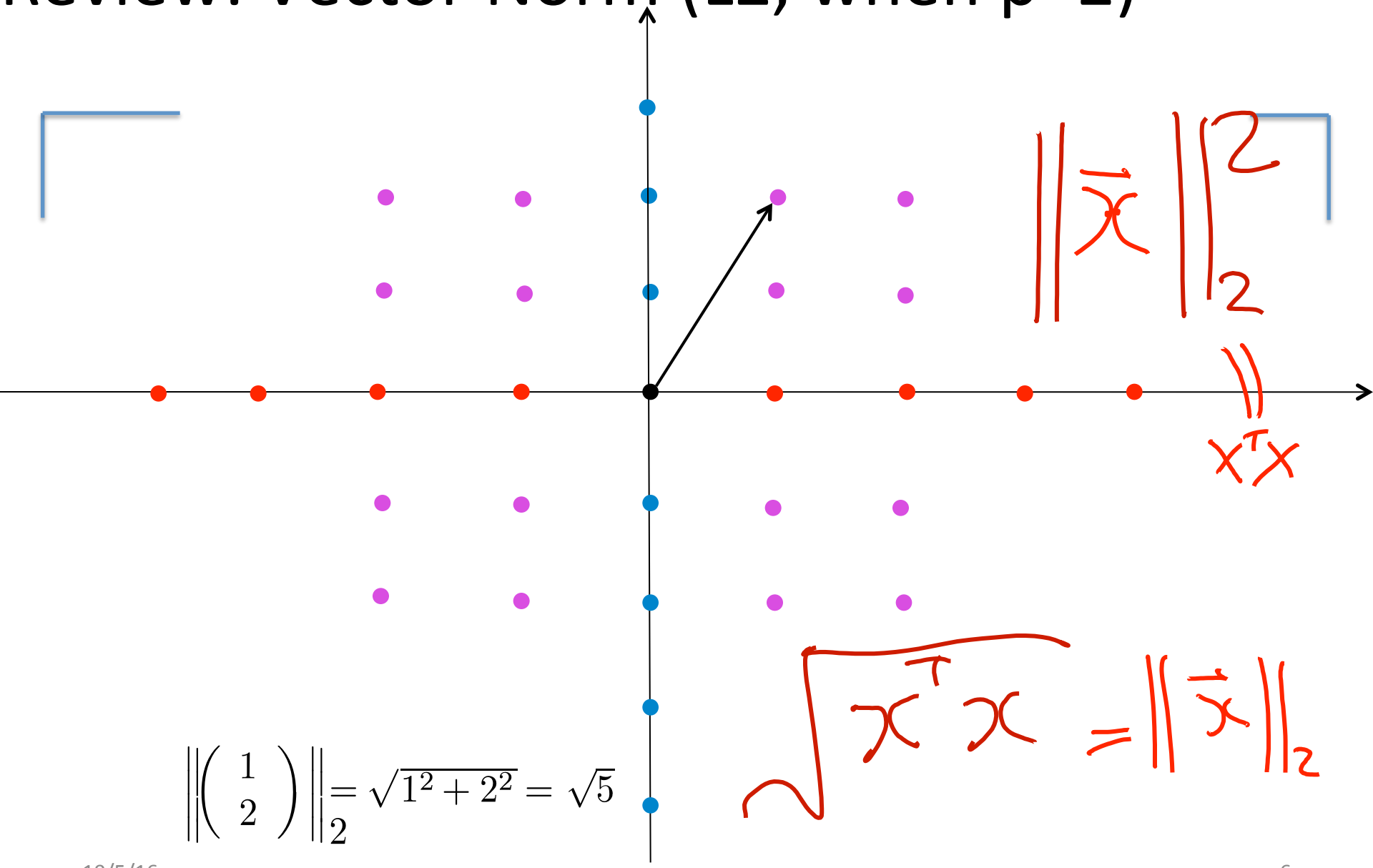
- Common norms: L_1, L_2 (Euclidean)

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

- $L_{infinity}$

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

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



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

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

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

Review: Normal equation for LR

- Write the cost function in matrix form:

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

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

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

$$\Rightarrow \quad X^T X\theta = X^T\vec{y}$$

The normal equations

$$\Downarrow$$

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

Assume that X^TX is invertible

Comments on the normal equation

- In most situations of practical interest, the number of data points N is larger than the dimensionality p of the input space and the matrix $\mathbf{X}$ is of full column rank. If this condition holds, then it is easy to verify that $X^T X$ is necessarily invertible.

- that $X^T X$ is invertible implies that it is positive definite (➔ SSE strong convex), thus the critical point we have found is a minimum.

- What if $\mathbf{X}$ has less than full column rank? → regularization (later).

Review: Page17 of Linear-Algebra Handout

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

- Likewise, a symmetric matrix $A \in \mathbb{S}^n$ is ***negative definite*** (ND), denoted $A \prec 0$ (or just $A < 0$) if for all non-zero $x \in \mathbb{R}^n$, $x^T Ax < 0$.

➔ One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible.

Finally, there is one type of positive definite matrix that comes up frequently, and so deserves some special mention. Given any matrix $A \in \mathbb{R}^{m \times n}$ (not necessarily symmetric or even square), the matrix $G = A^T A$ (sometimes called a ***Gram matrix***) is always positive semidefinite. Further, if $m \geq n$ (and we assume for convenience that A is full rank), then $G = A^T A$ is positive definite.

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 $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be ***full rank***. (2)
- 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 0f Handout

$X^T X$ is $p \times p$

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

$\min(n, p) = n$ when $n < p$

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

$X^T X$ singular / not invertible

Ridge Regression / L2

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

$$\hat{Y} = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Basic Model,

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

invertible

$\sum_j \beta_j^2 = \|\vec{\beta}\|_2^2$

HW2

- The ridge estimator is solution from

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

to minimize, take derivative and set to zero

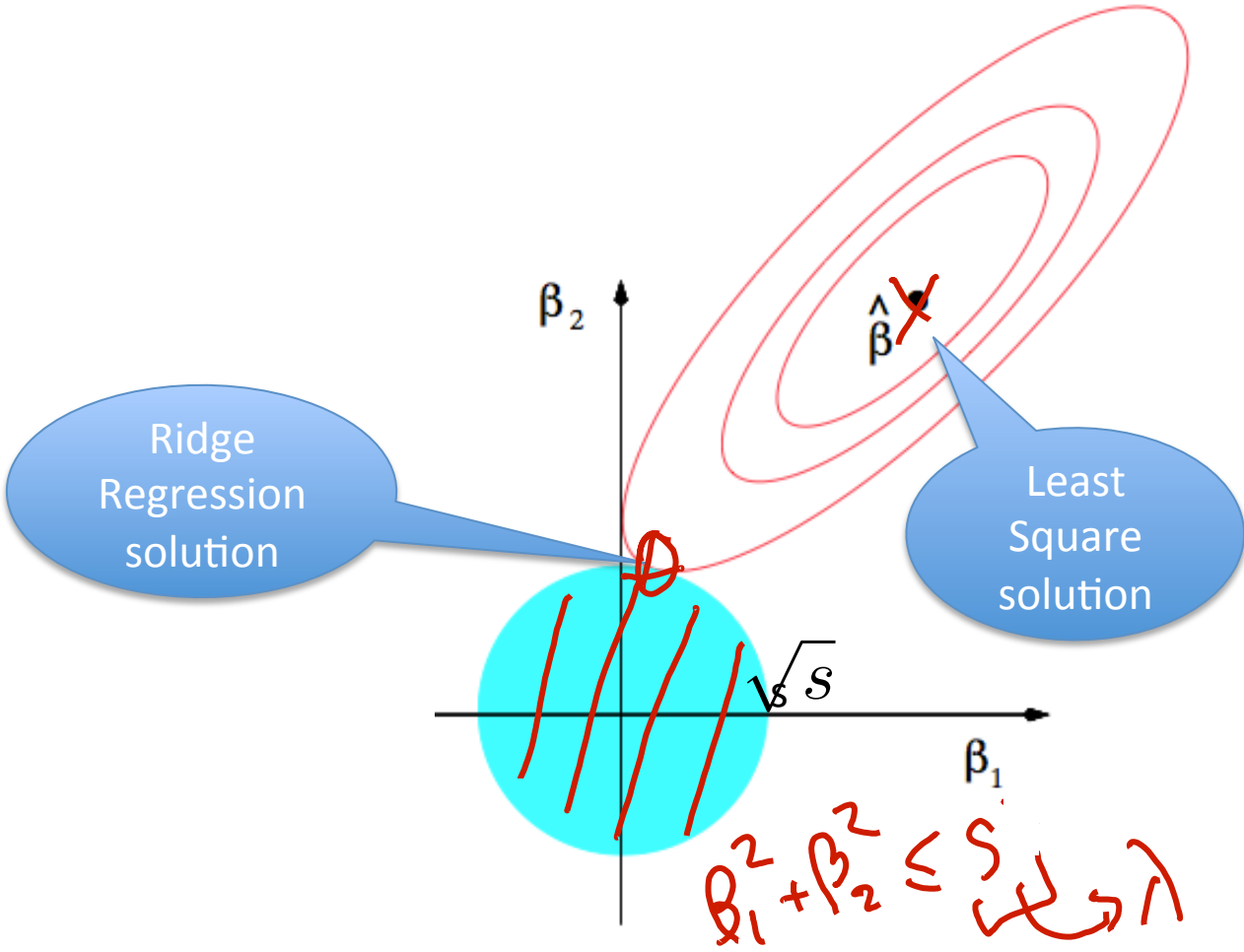
- Equivalently

$$\hat{\beta}^{ridge} = \arg\min(y - X\beta)^T (y - X\beta) \quad \text{subject to} \quad \sum_j \beta_j^2 \le s$$

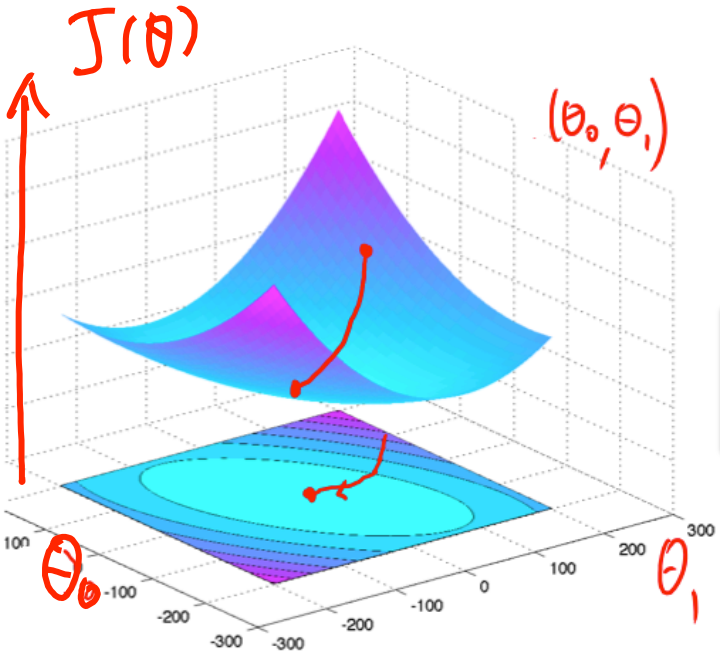
SSE

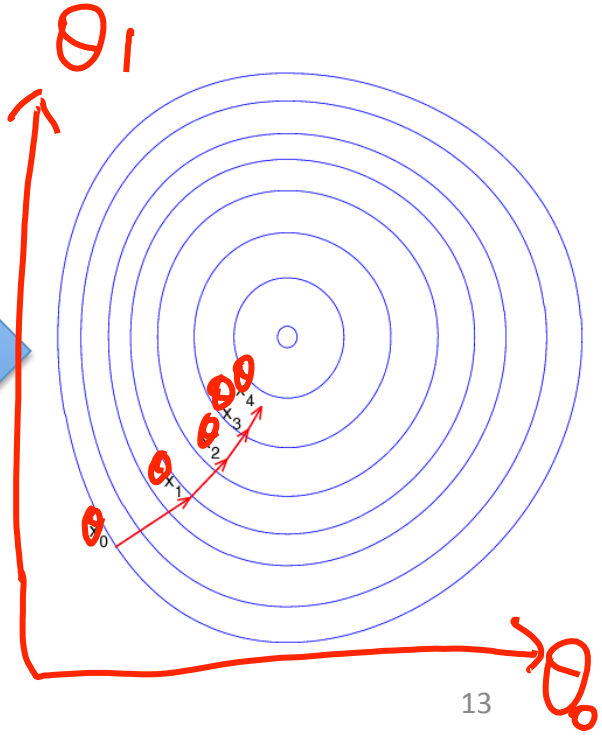
By convention, the bias/intercept term is typically not regularized. Here we assume data has been centered … therefore no bias term

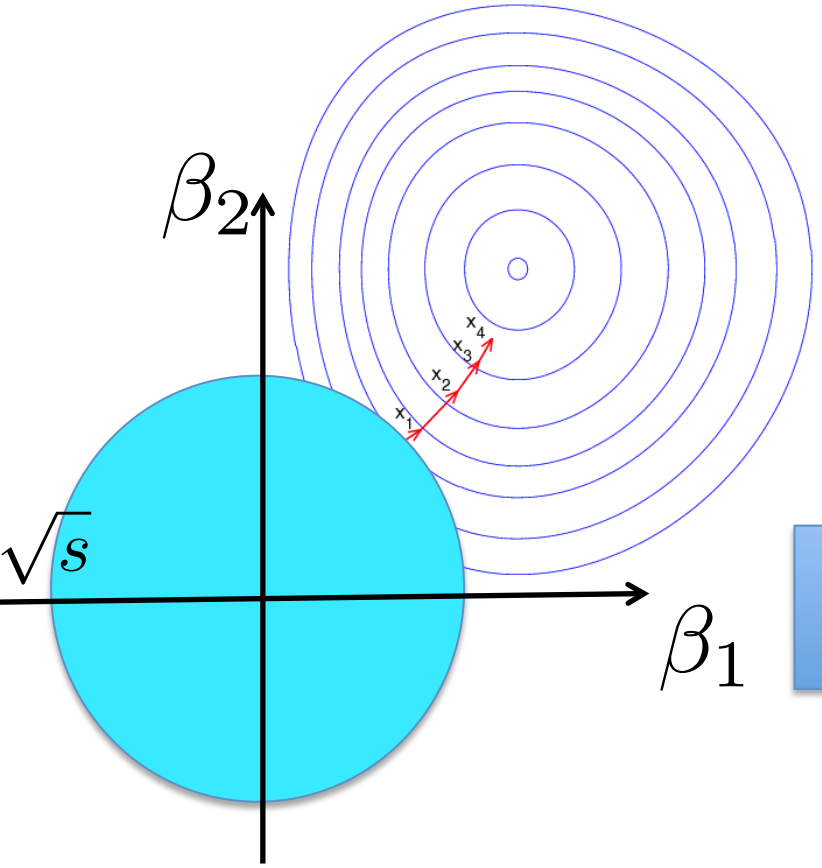
Objective Function's Contour lines from Ridge Regression

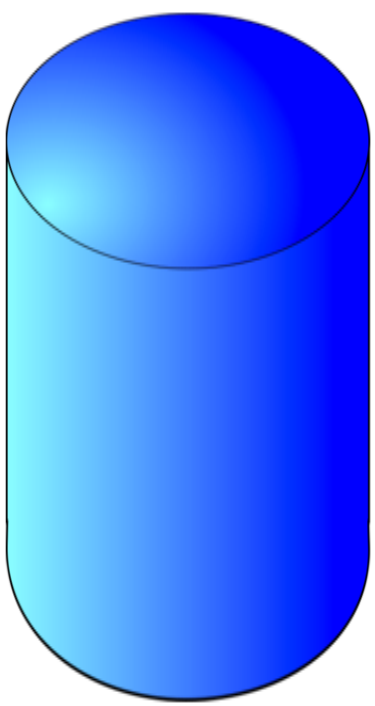


Review: from L3






β_2
$\sqrt{s}$
β_1
x1
x2
x3
x4



(1) Ridge Regression / L2

- The parameter $\lambda > 0$ penalizes β_j (next slide)
- Solution is $\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y$
- where I is the identity matrix. (next next slide S, λ)
- Note $\lambda = 0$ gives the least squares estimator;
- if $\lambda \to \infty$, then $\hat{\beta} \to 0$

Shrinkage ?

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

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

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

when $X^T X = I \Rightarrow \beta_{Rg} = \frac{1}{1+\lambda} X^T y = \frac{1}{1+\lambda} \beta_{OLS}$, $\lambda > 0$

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

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

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

Two forms of Ridge Regression

- Totally equivalent

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

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

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

When $X^T X = I$, $S = \sum_j (\beta_{Rg})_j^2 = \frac{1}{(1+\lambda)^2} \sum_j (\beta_{ols})_j^2$

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

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

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

Positive Definiteness

- One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible.

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

$$\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y$$

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

① $a^T X^T X a = (Xa)^T (Xa) = \|Xa\|_2^2 \geq 0$

(1×p, p×n, n×p, p×1; Xa: n×p, p×1 → n×1)

$\Downarrow$

$X^T X$ PSD

② $a^T (X^T X + \lambda I) a = a^T X^T X a + \lambda a^T I a$

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

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

PD → invertible

Today

- Linear Regression Model with Regularizations
  - Ridge Regression
  - Lasso Regression
  - Elastic net

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

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

$$\hat{\beta}^{lasso} = \arg\min (y - X\beta)^T (y - X\beta)$$

$$\text{subject to } \sum |\beta_j| \le s$$

L1 norm

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)

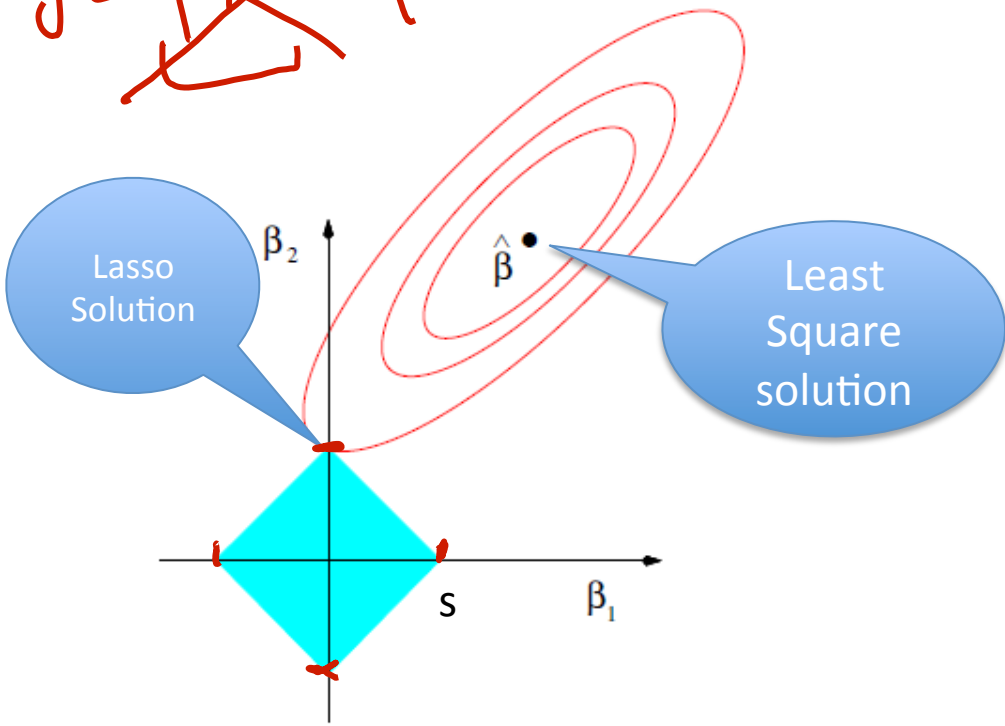
- Notice that ridge penalty $\sum \beta_j^2$ is replaced by $\sum \left|\beta_j\right|$

- 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 (least absolute shrinkage and selection

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

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



If data not centered, not-shrinking-the-bias-intercept-term-in-regression

Lasso Estimator

Ridge Regression

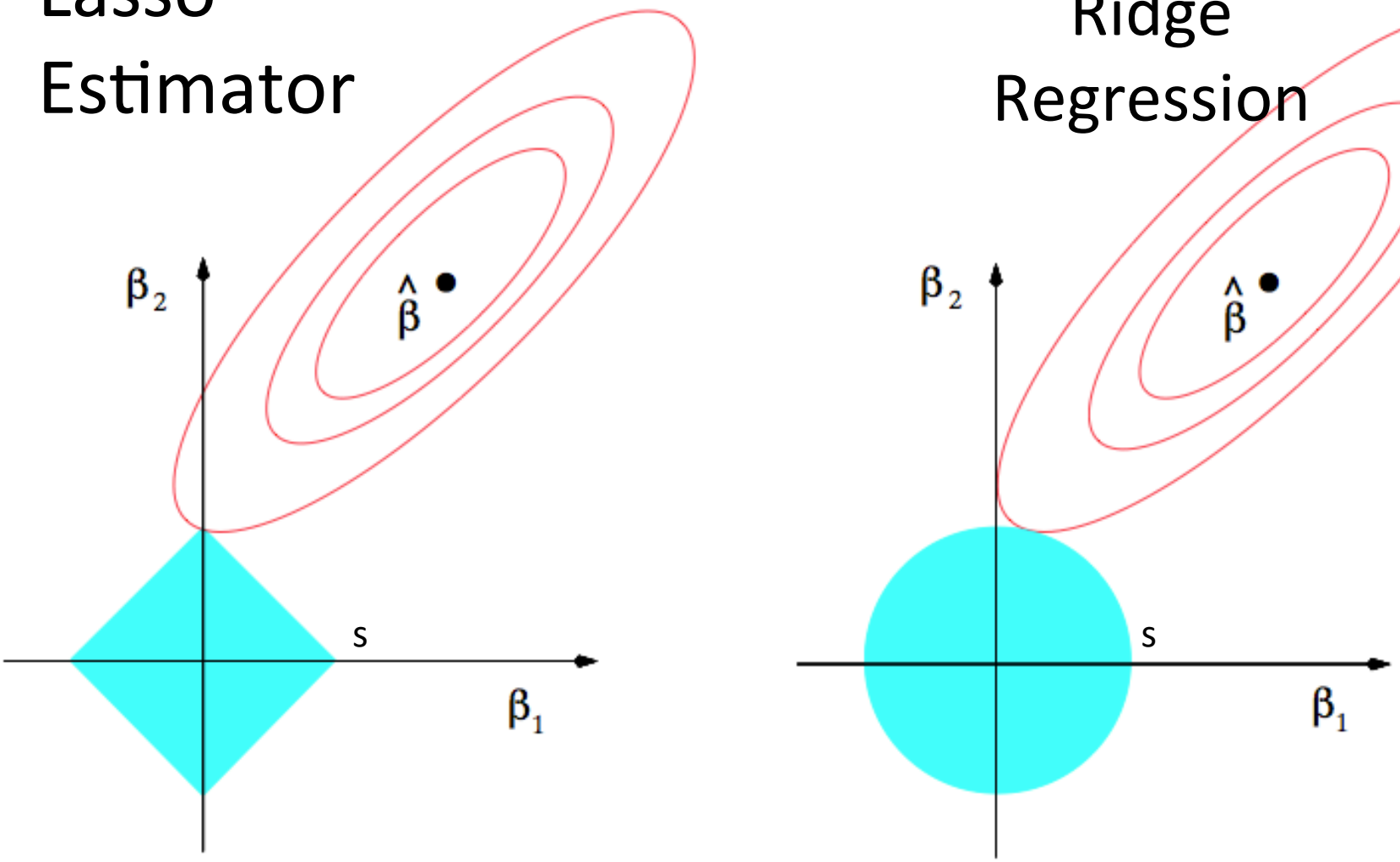


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are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

Today

- Linear Regression Model with Regularizations
  - Ridge Regression
  - Lasso Regression
  - Elastic net

(3) Hybrid of Ridge and Lasso

Elastic Net regularization

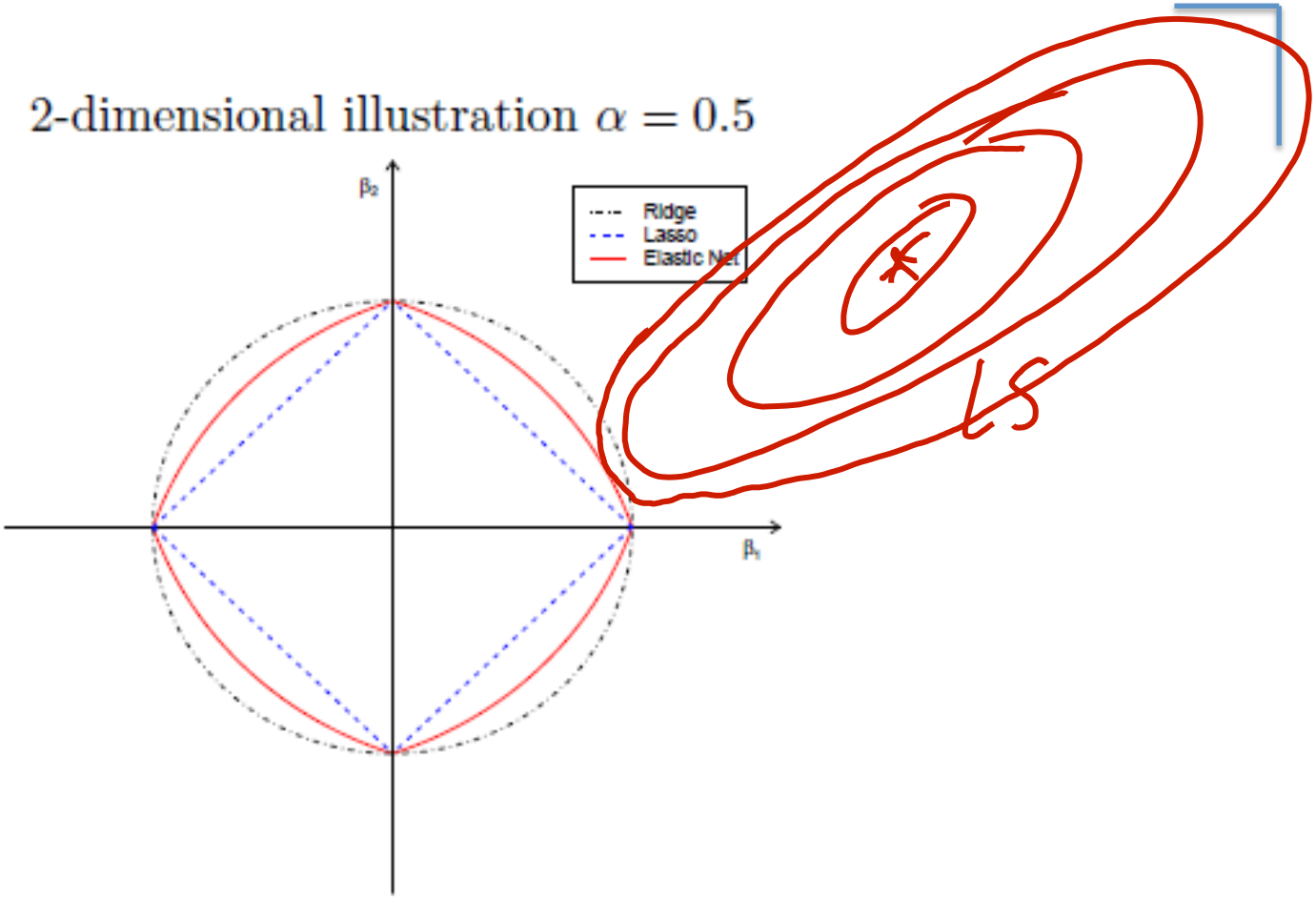
$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1$$

- The ℓ_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages *grouping effect*;
  - Stabilizes the ℓ_1 regularization path.

$\beta_j = 0$ Some $\{j\}$

Normally x and y have been centered, therefore no bias term needed in above !

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)

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

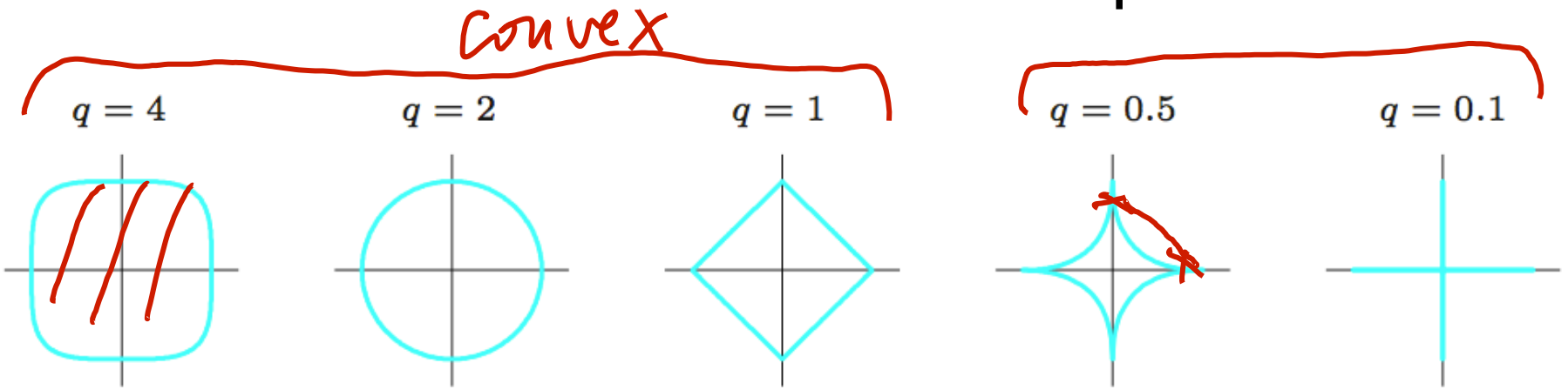
$$P(\boldsymbol{\beta}) = \sum_{j=1}^{p} \left(\tfrac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right)$$

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.

More: A family of shrinkage estimators

$$\beta = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$
$$\text{subject to } \sum |\beta_j|^q \le s$$

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



Here assume x and y have been centered (normally), therefore no bias term needed in above !

Summary:
Regularized multivariate linear regression

- Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$

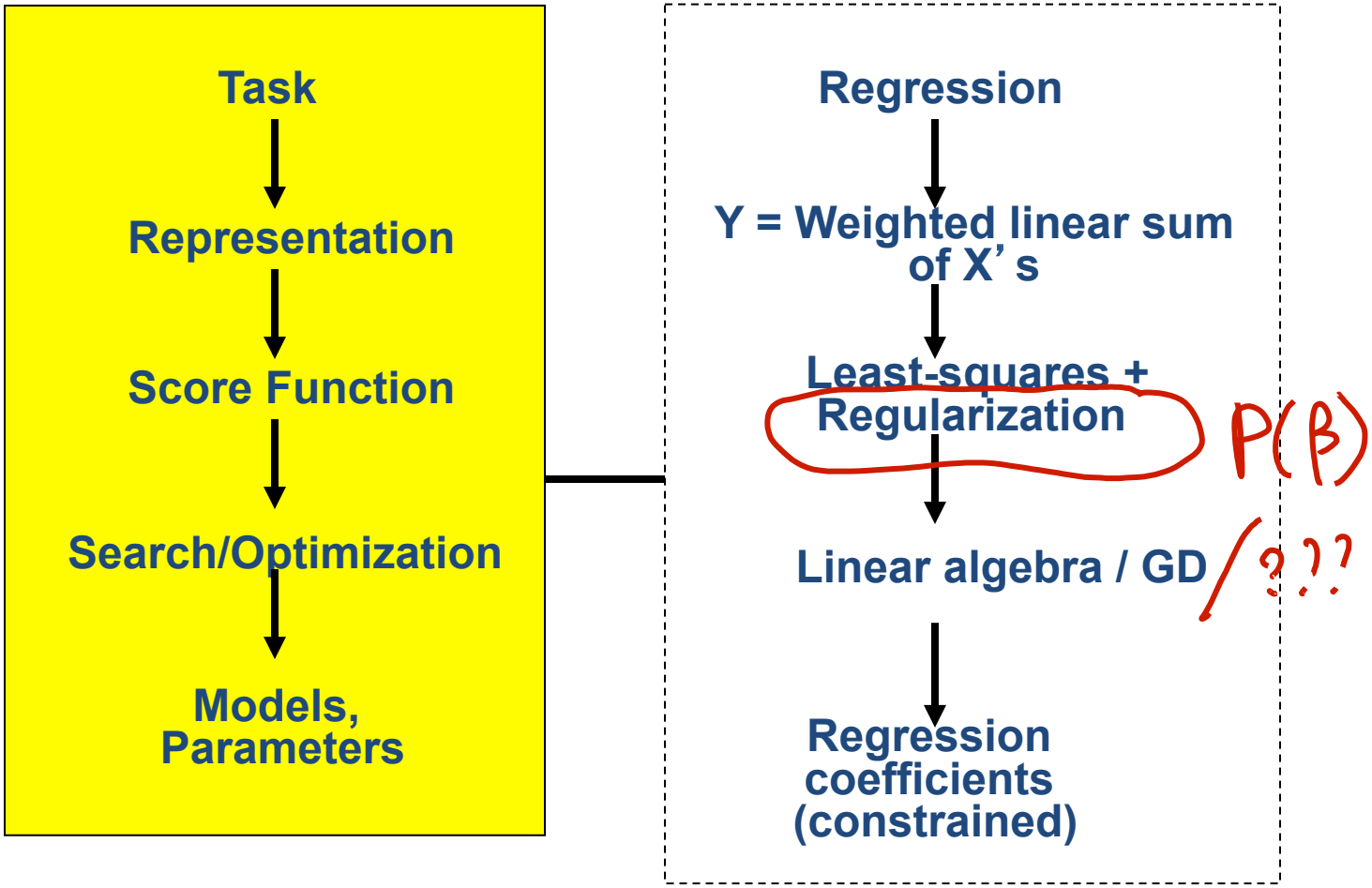
→ Penalty fun
→ Shrinkage term

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

30/54

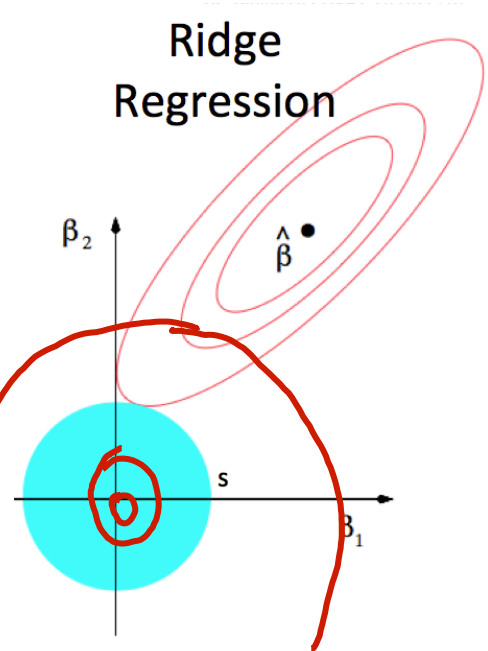
Error on data + Regularization

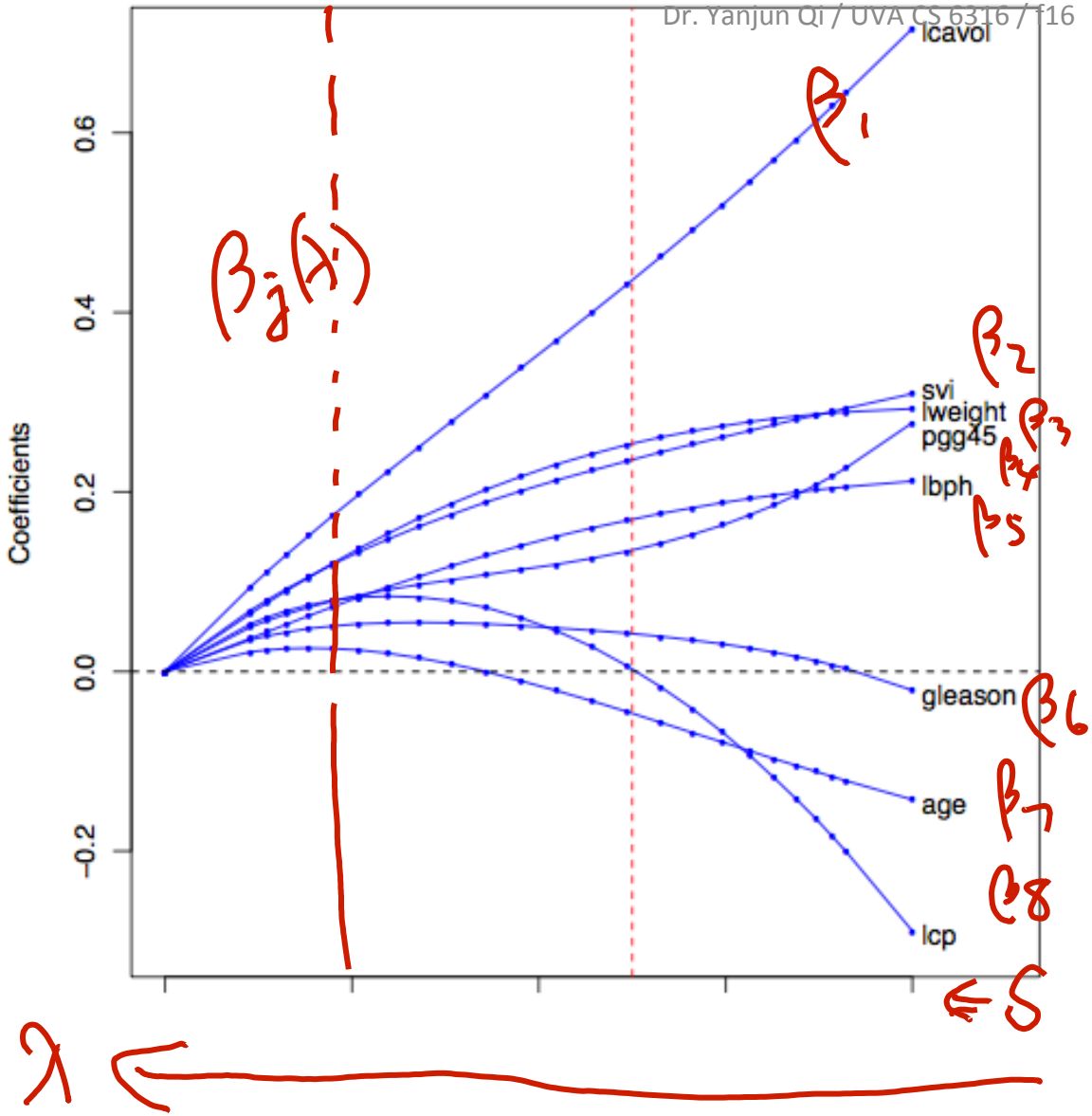
Regularized multivariate linear regression



Regularization path of a Ridge Regression

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





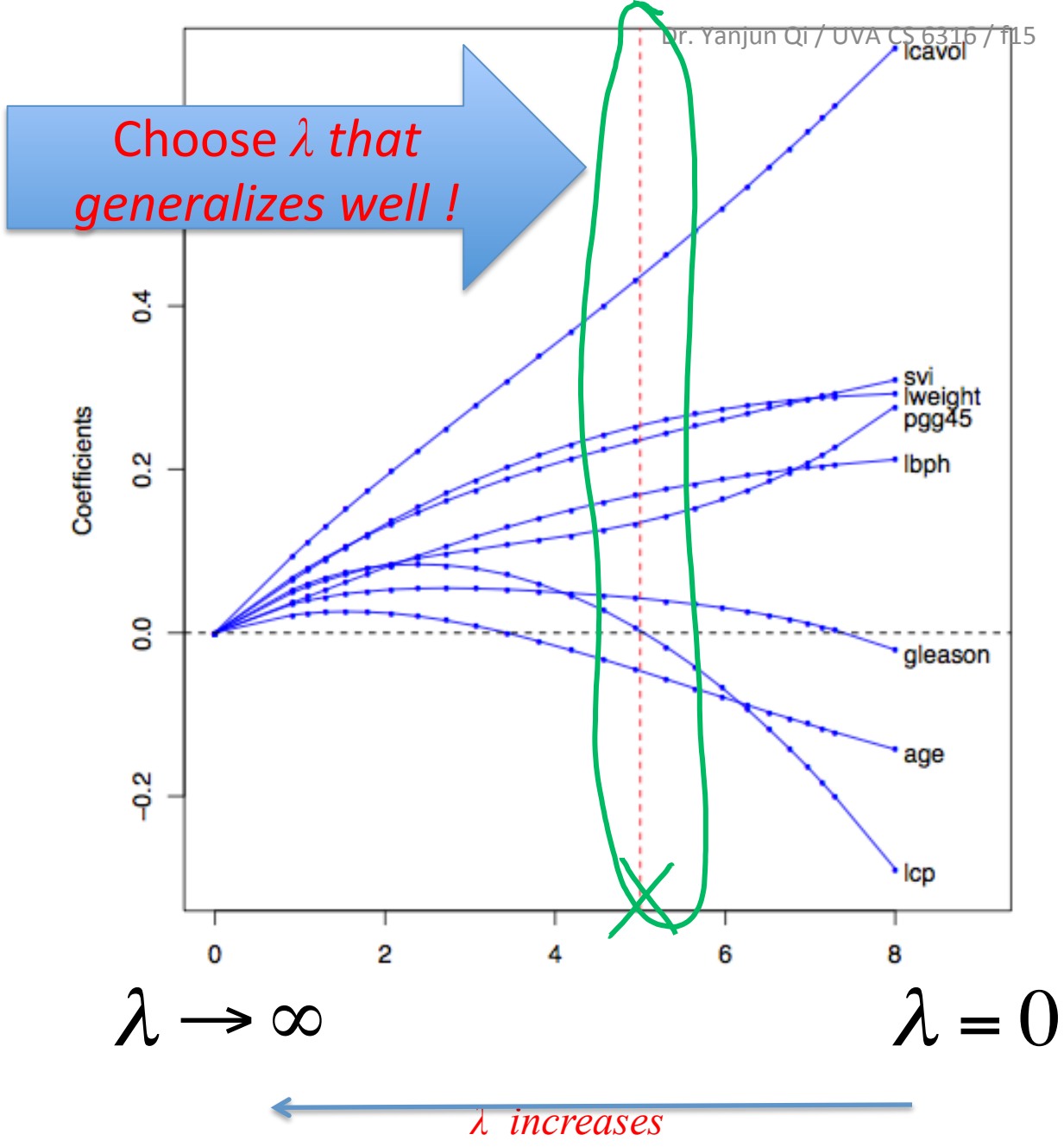
$\lambda \to \infty$

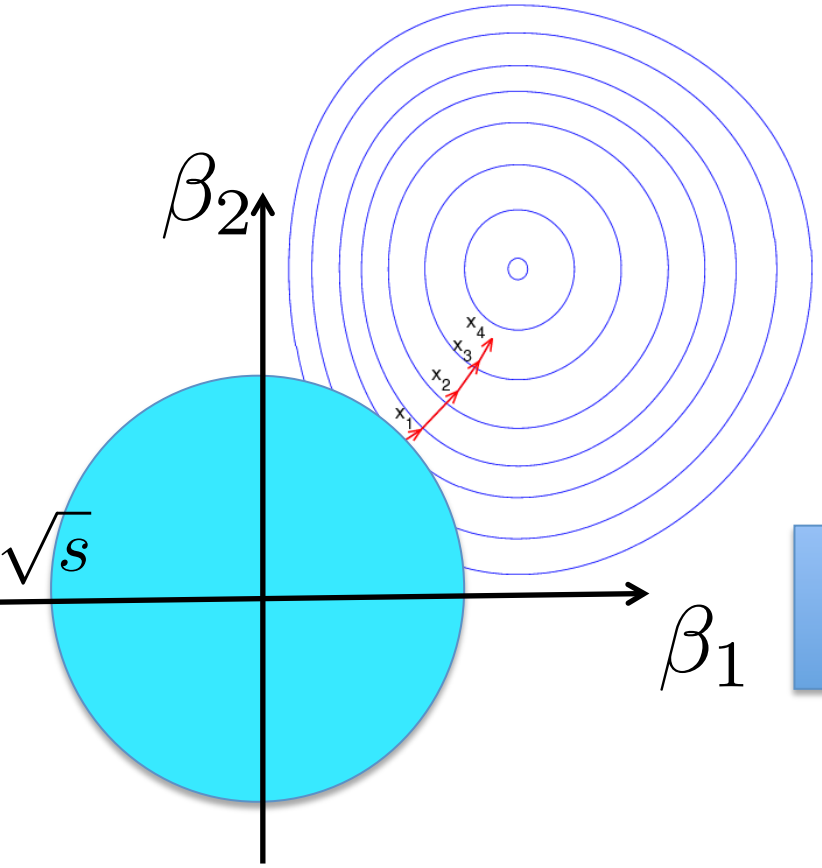
$\lambda = 0$

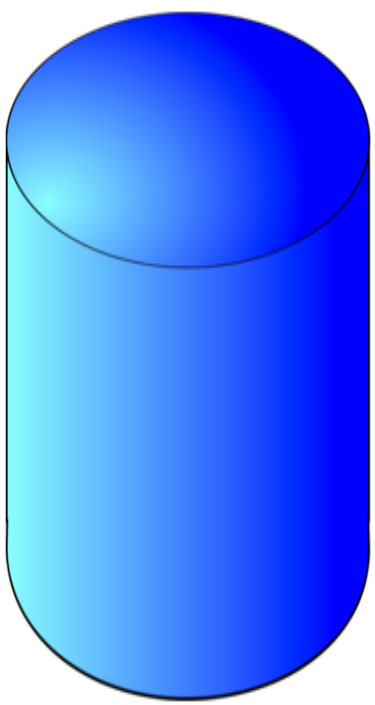
an example (ESL Fig3.8),

Ridge Regression

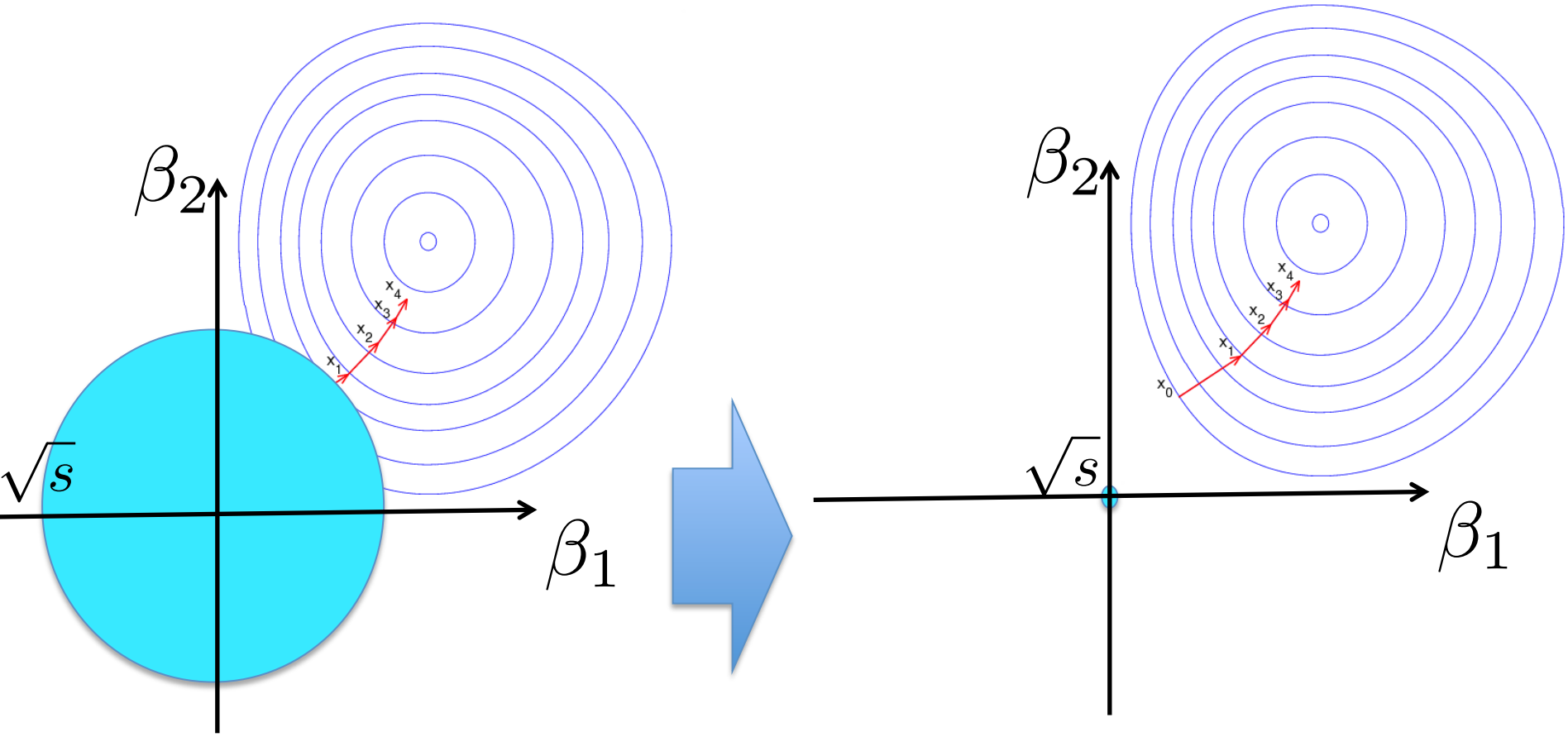
when varying λ, *how* β_j *varies.*

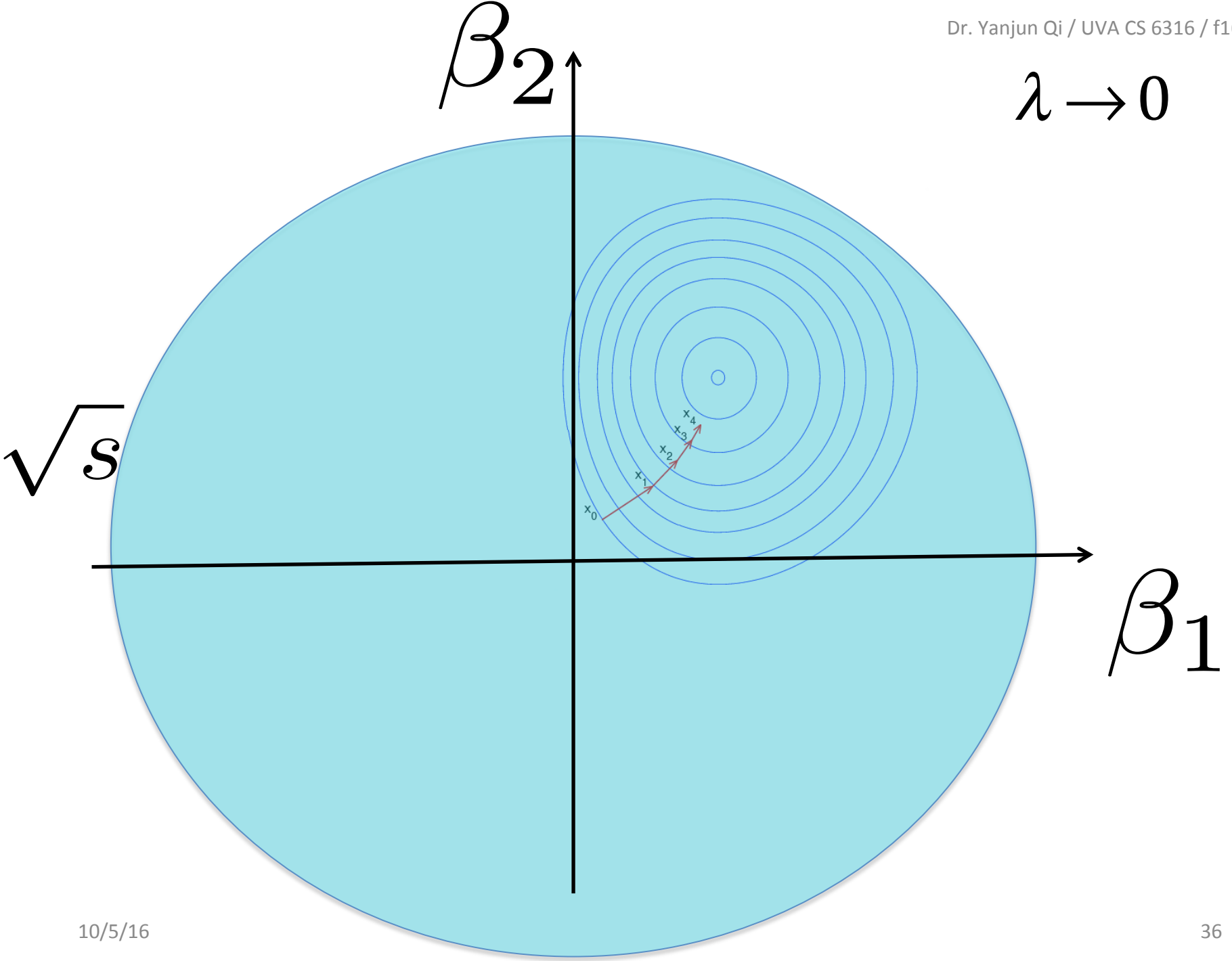



β_2
$\sqrt{s}$
β_1
x1
x2
x3
x4



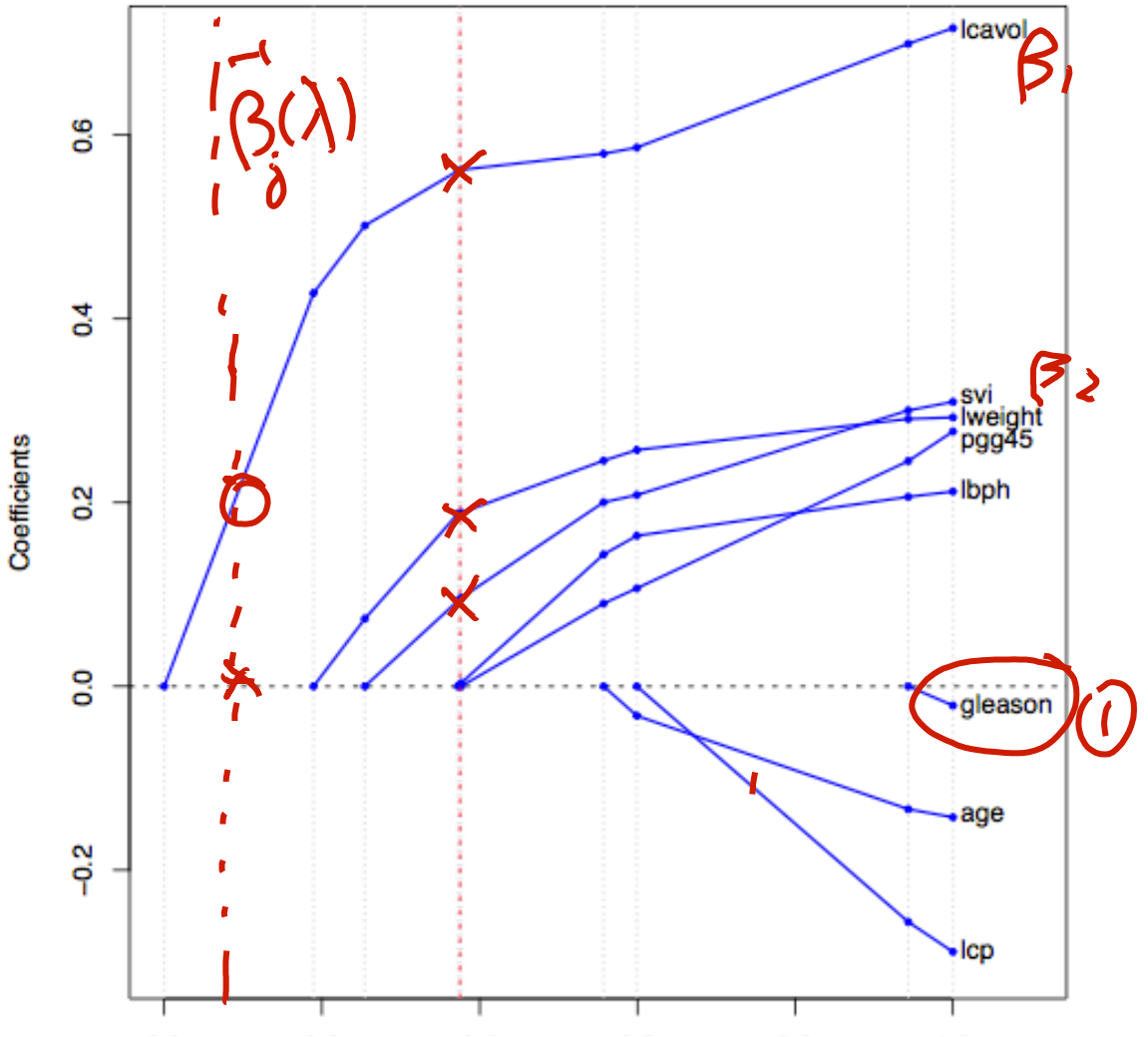
$$\lambda \to \infty$$



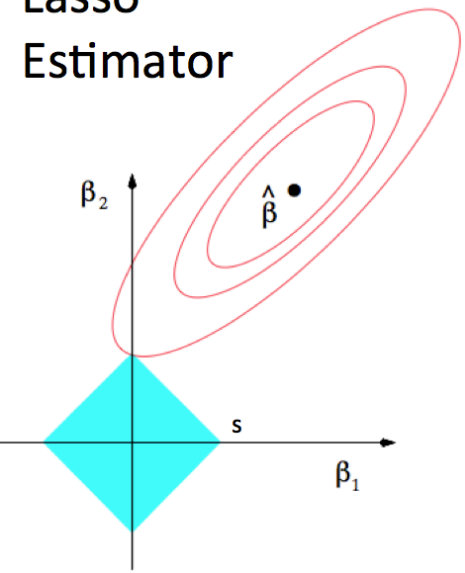

β_2
$\lambda \to 0$
$\sqrt{s}$
β_1
x_0
x_1
x_2
x_3
x_4

Regularization path of a Lasso Regression

when varying λ, *how* β_j *varies.*

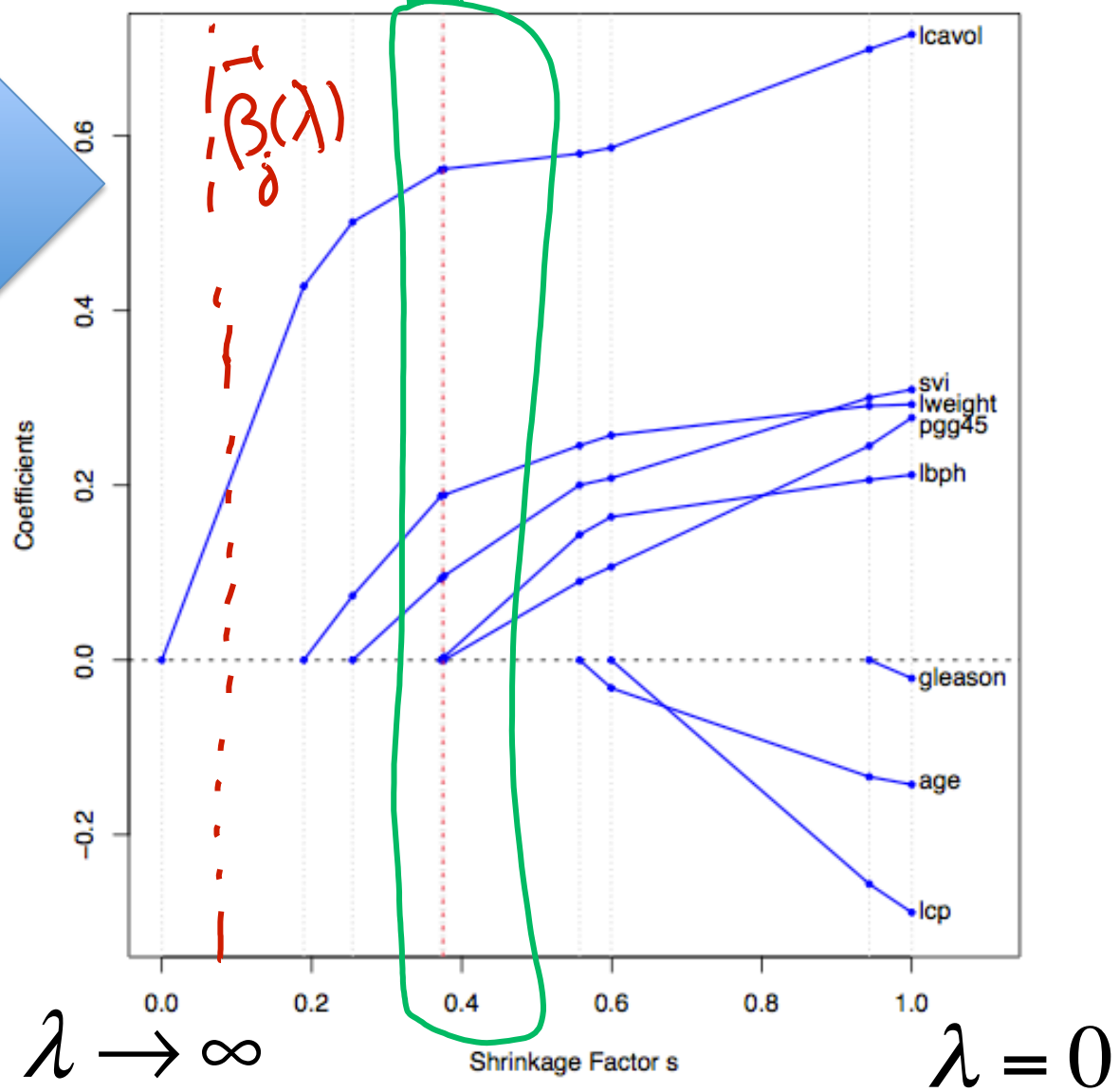


Lasso Estimator



Choose λ *that generalizes well !*

an example (ESL Fig3.10),



$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta$$

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

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.*

Lasso 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

Bias/Intercept Term is not Shrinked

- If the data is not centered, there exists bias term
  - http://stats.stackexchange.com/questions/86991/reason-for-not-shrinking-the-bias-intercept-term-in-regression

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

For ridge, in implementation, just set the bias corresponding entry as 0 in the I-matrix

- We normally assume we centered x and y. If this is true, no need to have bias term, e.g., for lasso,

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \|\beta\|_1$$

for ridge $+\lambda \|\beta\|_2^2$

Today Recap

- Linear Regression Model with Regularizations
  - Ridge Regression
    - why invertible (next class)
  - Lasso Regression
    - Extra: how to perform training (next class)
  - Elastic net
    - Extra: how to perform training (next class)

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*