정규화 모델 (Regularization Model)

2025 Spring

머신러닝1

이 두 호

개요

- 정규화 모델 1
  - 정규화 모델 배경
  - 정규화 모델 개념
  - Ridge Regression (능형회귀)

- 정규화 모델 2
  - Lasso
  - Elastic Net
  - 여러가지 정규화 모델

What is a good model?

현재 데이터(training data)를 잘 설명하는 모델

⇒ **Explanatory modeling**

미래 데이터(testing data)에 대한 예측 성능이 좋은 모델

⇒ **Predictive modeling**

Good Explanatory Model

현재 데이터(training data)를 잘 설명하는 모델

= 학습오차(training error) 를 최소화하는 모델

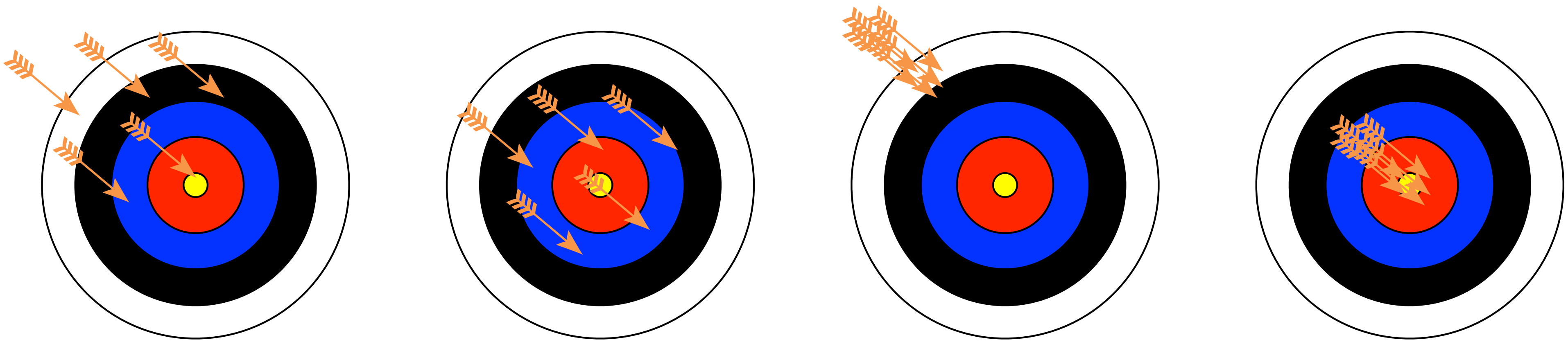
$$MSE_{(training)} = \frac{1}{n}\sum_{i=1}^{n}(Y - \hat{Y})^2$$

Good Predictive Model

$$E\left[MSE_{(testing)}\right] = E\left[(Y-\hat{Y})^2 \mid X\right]$$

$$= \sigma^2 + \left(E[\hat{Y}] - Y\right)^2 + E\left[\left(\hat{Y} - E[\hat{Y}]\right)^2\right]$$

$$= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}$$



Bias	High	Low	High	Low
Variance	High	High	Low	Low

Good Predictive Model

미래 데이터(testing data)에 대한 예측 성능이 좋은 모델

= 미래 데이터에 대한 expected error 가 낮은 모델

$$E\left[MSE\right] = \text{Error} + \text{Bias}^2 + \text{Variance}$$

- Expected MSE를 줄이려면 bias, variance, 혹은 둘다 낮춰야함
- 그렇지 못하다면 둘 중에 하나라도 작으면 좋음
- Bias가 증가되더라도 variance 감소폭이 더 크다면 expected MSE는 감소(예측성능 증가)

Ordinary Linear Regression Model

$$MSE = \sum_{i=1}^{n} \left\{ Y_i - (w_0 + w_1 x_{1i} + w_2 x_{2i} + \cdots + w_p x_{pi}) \right\}^2$$

$$\min_{w_0, \cdots, w_p \in \mathbb{R}} \quad MSE$$

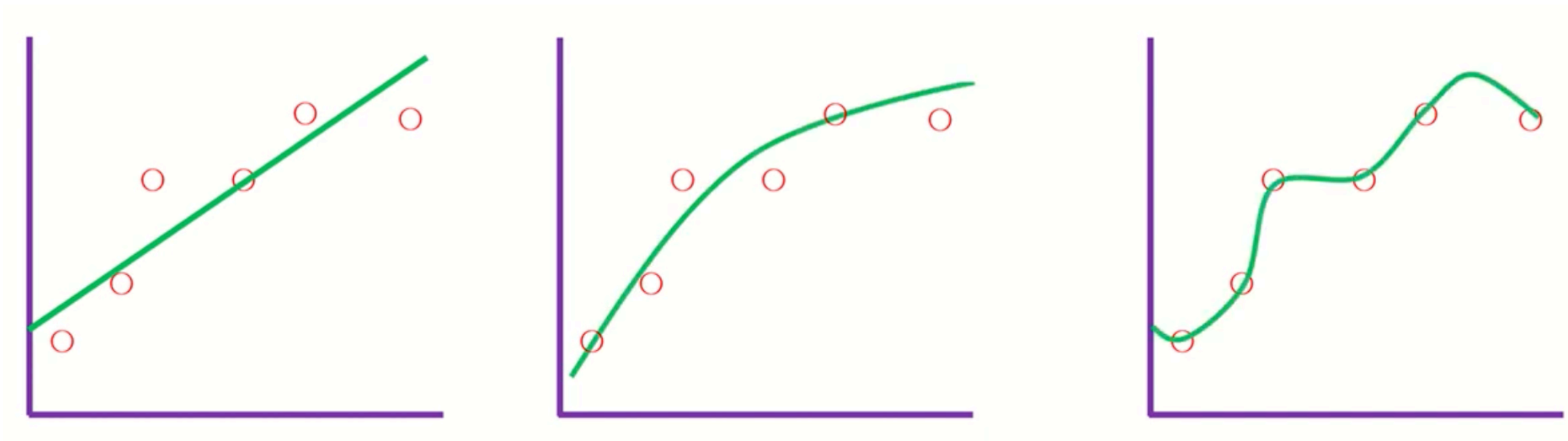
$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \quad \mathbf{w} = \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}}_{(1+p) \times 1}$$

회귀계수 $\mathbf{w}$ 에 대한 unbiased estimator 중 가장 분산이 작은 estimator (**B**est **L**inear and **U**nbiased **E**stimator: **BLUE**, Gauss-Markov Theorem)

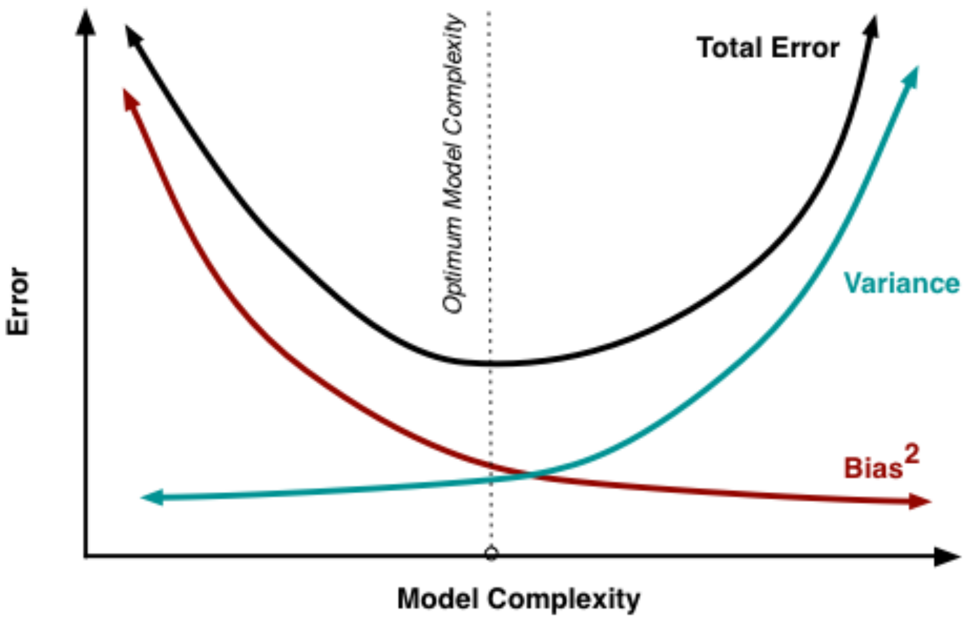
Unbiased 를 포기 하더라도 (예측) 분산을 획기적으로 줄일 수 없을까?

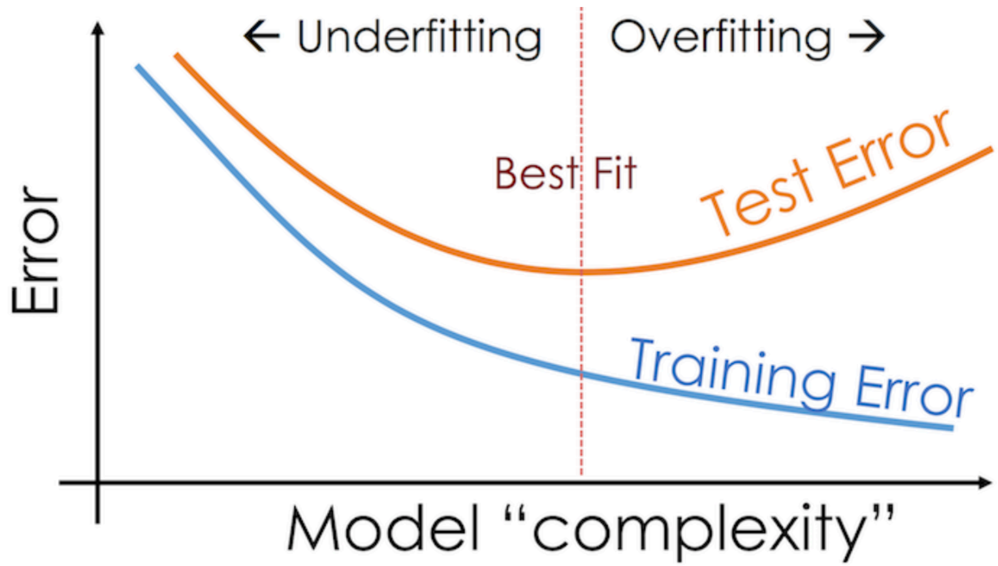
- Subset selection method는 전체 p 개의 입력변수 중 일부 k 개만을 사용하여 회귀계수 $\mathbf{w}$ 를 추정 하는 방법
- 전체 변수 중 일부만을 선택함으로써 bias가 증가할 수 있지만 variance는 감소함
  - Best subset selection
  - Forward stepwise selection
  - Backward stepwise selection
  - Least angle regression
  - Orthogonal matching pursuit

Regularization Concept

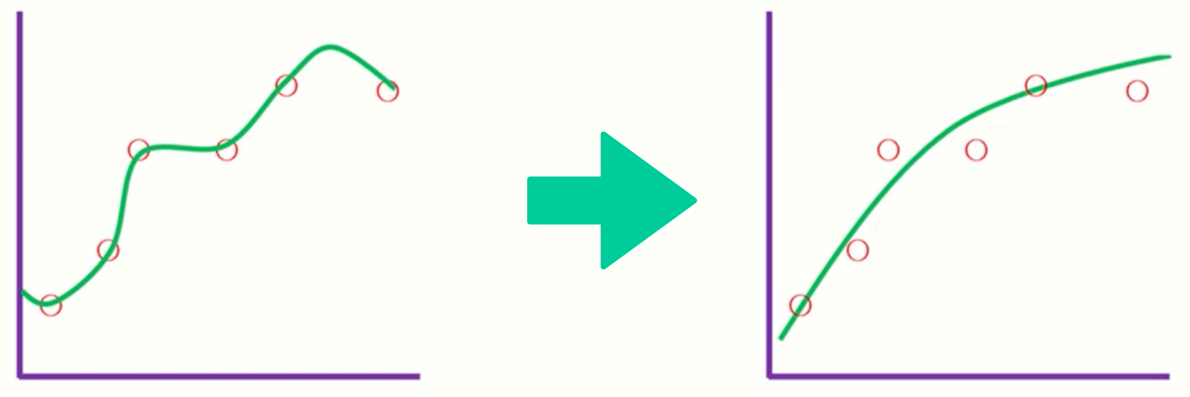


$w_0 + w_1 x$ $\qquad$ $w_0 + w_1 x + w_2 x^2$ $\qquad$ $w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$





Regularization Concept



$w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$

$w_0 + w_1 x + w_2 x^2$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 999999 w_3^2 + 999999 w_4^2$$

$$w_3 \rightarrow 0 \qquad w_4 \rightarrow 0$$

Doo Ho Lee, Dept. of AI & SW

Regularization Concept

$$w_1,\ w_2,\ \cdots,\ w_p$$

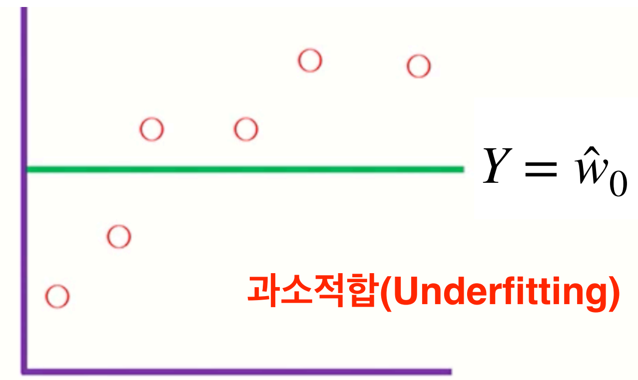
$$C(\mathbf{w}) = \min_{\mathbf{w}} \underbrace{\boxed{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}}_{\text{(1) Training error}} + \underbrace{\boxed{\lambda \sum_{j=1}^{p} w_j^2}}_{\text{(2) Generalization}}$$

λ: regularization parameter that controls the tradeoff between (1) and (2)

Regularization Concept

$$C(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} w_j^2$$

If λ is very big, then w_i approaches 0.



If λ is very small, then w_i may not 0.

- Regularization method 는 회귀계수 **w** 가 가질 수 있는 값에 제약조건을 부여하는 방법
- 제약조건에 의해 bias는 증가할 수 있지만 variance는 감소함

최소제곱법

$$\min_{w_1, w_2} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + w_2 x_{2i}) \right\}^2$$

정규화 방법

$$\min_{w_1, w_2} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + w_2 x_{2i}) \right\}^2$$

$$st$$

$$\boxed{w_1^2 + w_2^2 \leq 30}$$

w 값에 대한 제약조건 추가

Regularization Method

(w_1, w_2)	$w_1^2 + w_2^2$	MSE
(4, 5)	41	20
(3, 5)	34	23
(4, 4)	32	25
(2, 5)	27	27
(2, 4)	18	25
(2, 3)	13	29

$$w_1^2 + w_2^2 \leq 30$$

L2-norm regularization:

오차제곱합을 최소화하면서 회귀계수 **w** 의 L_2-norm을 제한

$$\mathbf{X} = \underbrace{\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}}_{n \times p} \qquad \mathbf{y} = \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{n \times 1} \qquad \mathbf{w} = \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}}_{p \times 1}$$

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2$$

s.t. $\quad w_1^2 + \cdots + w_p^2 \le t$

Doo Ho Lee, Dept. of AI & SW

Ridge Regression

Ridge Regression

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2$$

s.t. $\quad w_1^2 + \cdots + w_p^2 \leq t$

$\Updownarrow$ **Equivalent (Lagrangian multiplier)**

Ridge Regression

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda \sum_{j=1}^{p} w_j^2$$

MSE Contour

$$MSE(w_1, w_2) = \sum_{i=1}^{n} \{Y_i - (w_1 x_{1i} + w_2 x_{2i})\}^2$$

$$= \left(\sum_{i=1}^{n} x_{1i}^2\right) w_1^2 + \left(\sum_{i=1}^{n} x_{2i}^2\right) w_2^2 + \left(2\sum_{i=1}^{n} x_{1i} x_{2i}\right) w_1 w_2$$

$$-\left(2\sum_{i=1}^{n} Y_i x_{1i}\right) w_1 - \left(2\sum_{i=1}^{n} Y_i x_{2i}\right) w_2 + \sum_{i=1}^{n} Y_i^2$$

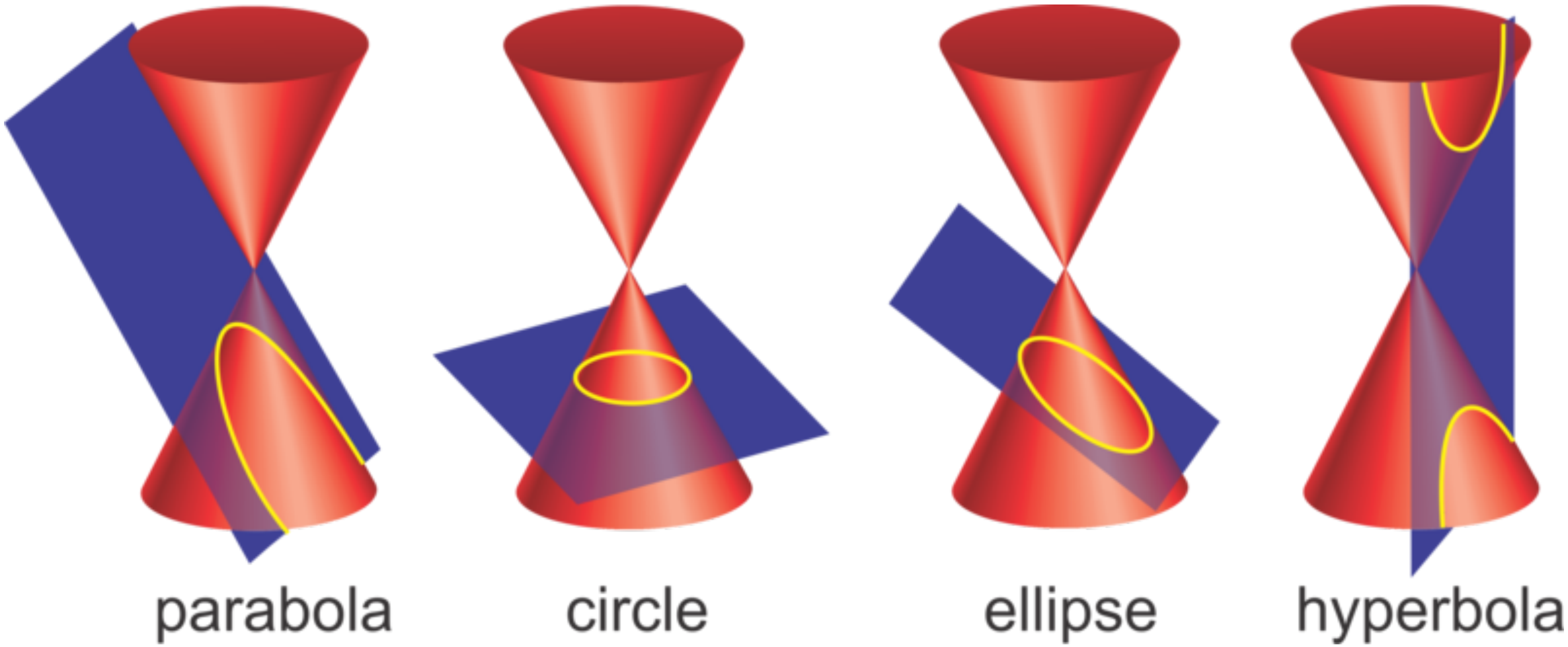
$$= \boxed{A w_1^2 + B w_1 w_2 + C w_2^2 + D w_1 + E w_2 + F}$$

Conic equation (2차원의 경우)

$$Aw_1^2 + Bw_1w_2 + Cw_2^2 + Dw_1 + Ew_2 + F = 0$$

Discriminant of conic equation (판별식): $B^2 - 4AC$

$B^2 - 4AC = 0 \quad \rightarrow$ **parabola (포물선)**
$B^2 - 4AC > 0 \quad \rightarrow$ **hyperbola (쌍곡선)**
$B^2 - 4AC < 0 \quad \rightarrow$ **ellipse (타원)**
$B = 0 \;\; and \;\; A = C \quad \rightarrow$ **circle (원)**



MSE Contour

$$MSE(w_1, w_2) = \left(\sum_{i=1}^{n} x_{1i}^2\right) w_1^2 + \left(\sum_{i=1}^{n} x_{2i}^2\right) w_2^2 + \left(2\sum_{i=1}^{n} x_{1i}x_{2i}\right) w_1 w_2 - \left(2\sum_{i=1}^{n} Y_i x_{1i}\right) w_1 - \left(2\sum_{i=1}^{n} Y_i x_{2i}\right) w_2 + \sum_{i=1}^{n} Y_i^2$$
$$= Aw_1^2 + Bw_1w_2 + Cw_2^2 + Dw_1 + Ew_2 + F$$

$$B^2 - 4AC = \left(2\sum_{i=1}^{n} x_{1i}x_{2i}\right)^2 - 4\sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2$$
$$= 4\left\{\left(\sum_{i=1}^{n} x_{1i}x_{2i}\right)^2 - \sum_{i=1}^{n} x_{1i}^2 \sum_{i=1}^{n} x_{2i}^2\right\} < 0$$

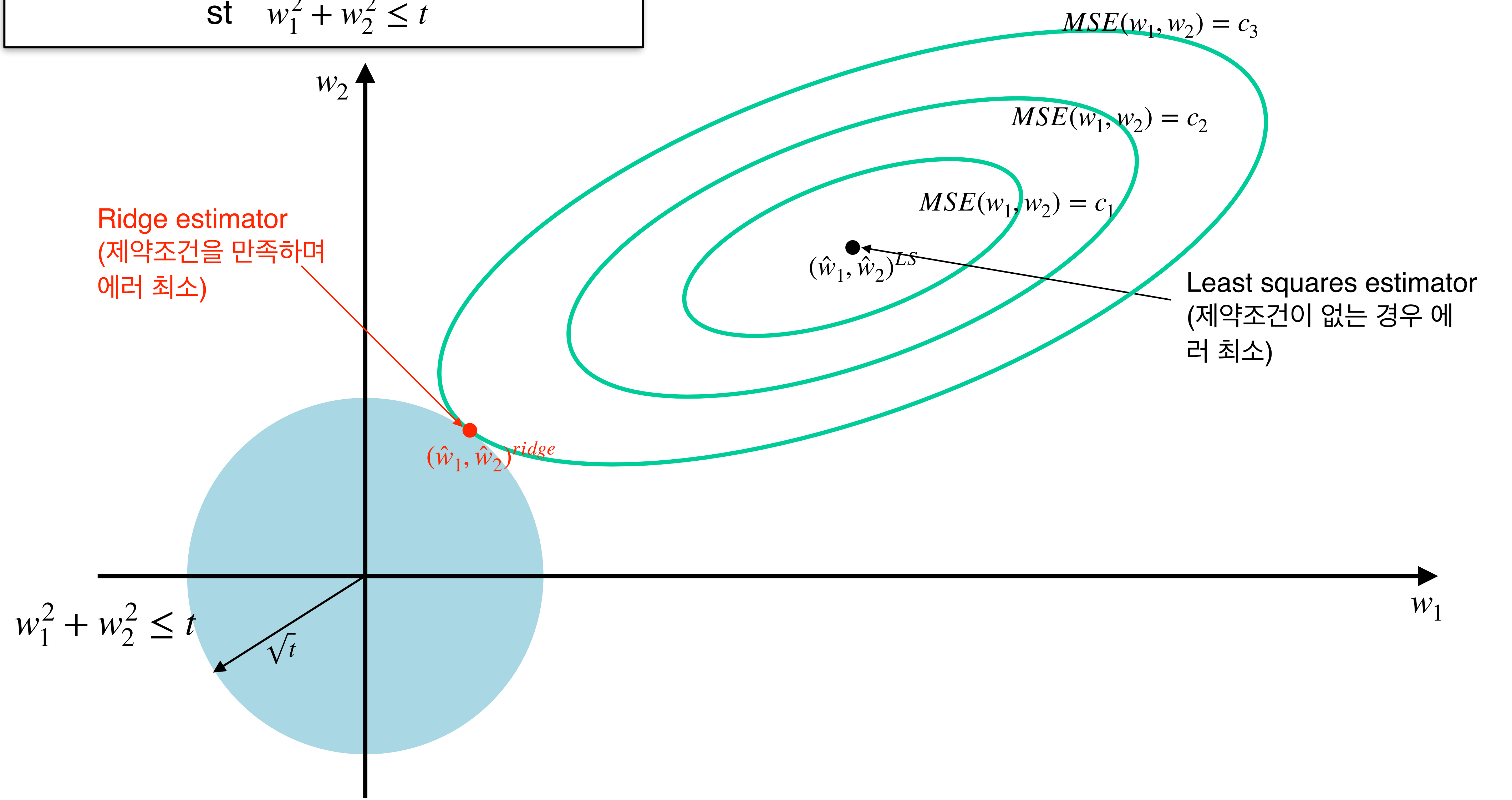
By Cauchy-Schwartz inequality

Ridge Regression

$$(\hat{w}_1, \hat{w}_2)^{ridge} = \arg\min_{w_1, w_2} \sum_{i=1}^{n} \left(Y_i - x_{1i}w_1 - x_{2i}w_2\right)^2$$

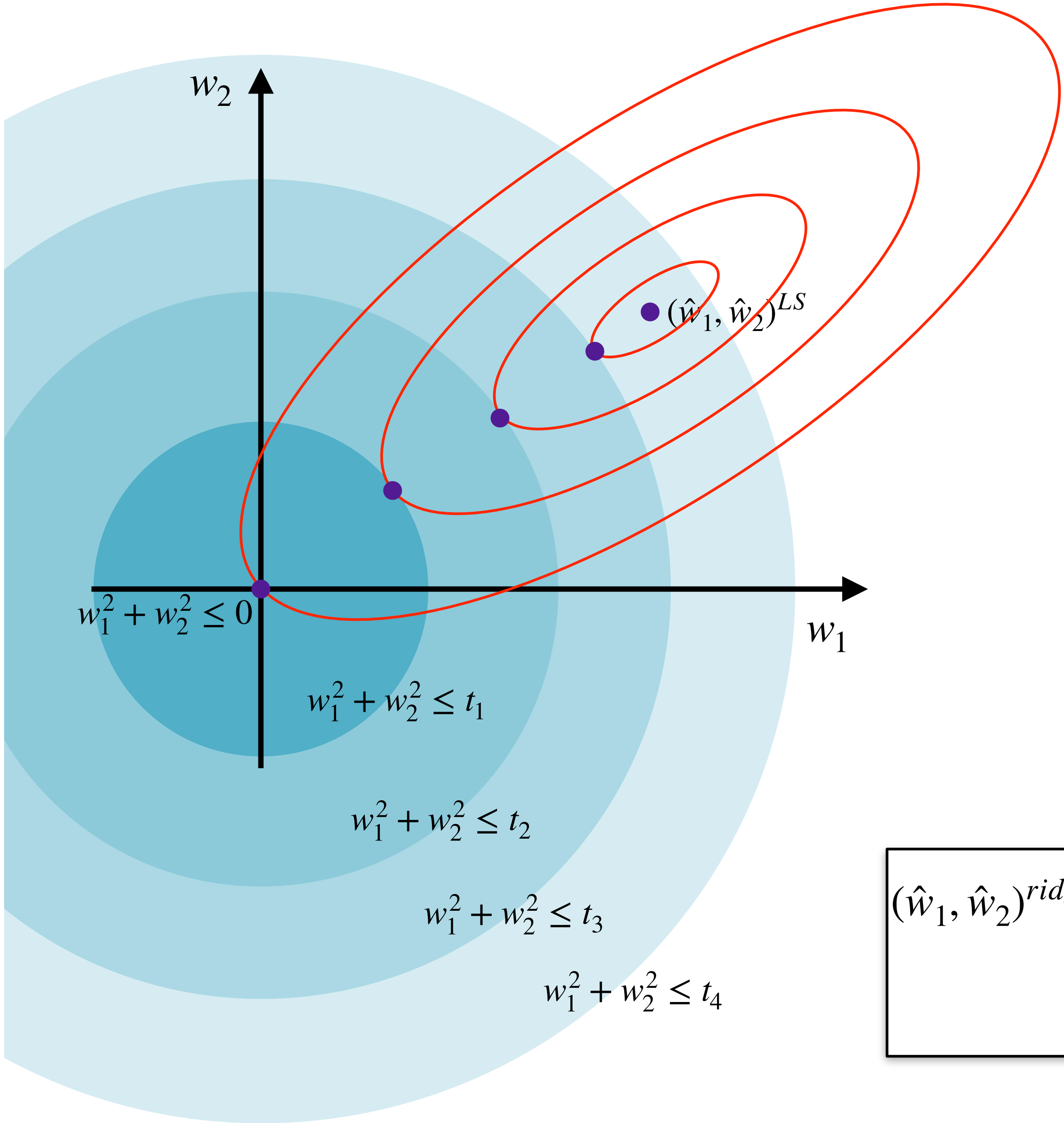
$$\text{st} \quad w_1^2 + w_2^2 \le t$$

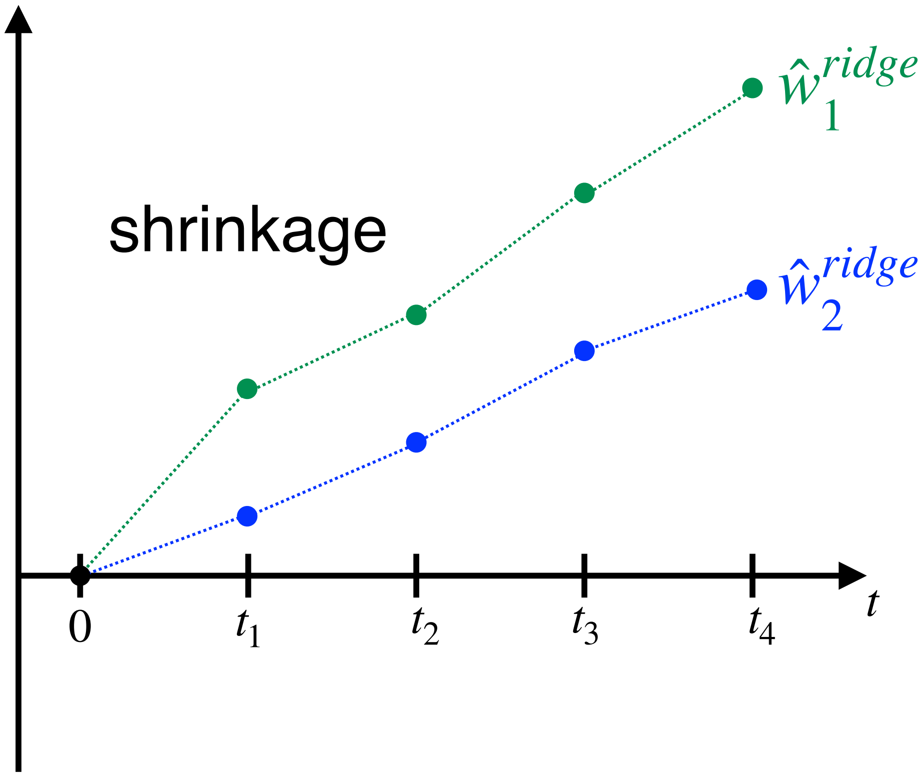
MSE contour
(중심에서 멀어질수록 에러 증가)



Ridge Solution Path

Solution path : t 값에 따른 $(\hat{w}_1, \hat{w}_2)^{ridge}$의 변화





$$(\hat{w}_1, \hat{w}_2)^{ridge} = \arg\min_{w_1, w_2} \sum_{i=1}^{n} \left(Y_i - x_{1i}w_1 - x_{2i}w_2\right)^2$$

$$\text{st} \quad w_1^2 + w_2^2 \leq t$$

Doo Ho Lee, Dept. of AI & SW

Least Squares Solutions

일반선형회귀분석의 회귀계수 **w**는 행렬 연산을 통해 구할 수 있음

$$
\mathbf{X} = \underset{n \times (1+p)}{\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}} \quad \mathbf{y} = \underset{n \times 1}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}} \quad \mathbf{w} = \underset{(1+p) \times 1}{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix}}
$$

$$
C(\mathbf{w})^{LS} = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w})
$$

$$
\frac{\partial C(\mathbf{w})^{LS}}{\partial \mathbf{w}} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{0}
$$

$$
\hat{\mathbf{w}}^{LS} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}
$$

Ridge Solutions

능형회귀분석의 회귀계수 **w**도 행렬 연산을 통해 구할 수 있음

$$\mathbf{X} = \underset{n \times p}{\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}} \quad \mathbf{y} = \underset{n \times 1}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}} \quad \mathbf{w} = \underset{p \times 1}{\begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}}$$

$$C(\mathbf{w})^{ridge} = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^\top \mathbf{w}$$

$$\frac{\partial C(\mathbf{w})^{ridge}}{\partial \mathbf{w}} = -2\mathbf{X}^\top \mathbf{y} + 2\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p\right)\mathbf{w} = \mathbf{0}$$

$$\hat{\mathbf{w}}^{ridge} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Least Absolute Shrinkage and Selection Operator

변수 선택 가능

L_1-norm regularization: 회귀계수 $\mathbf{w}$ 의 L_1-norm을 제한

Lasso

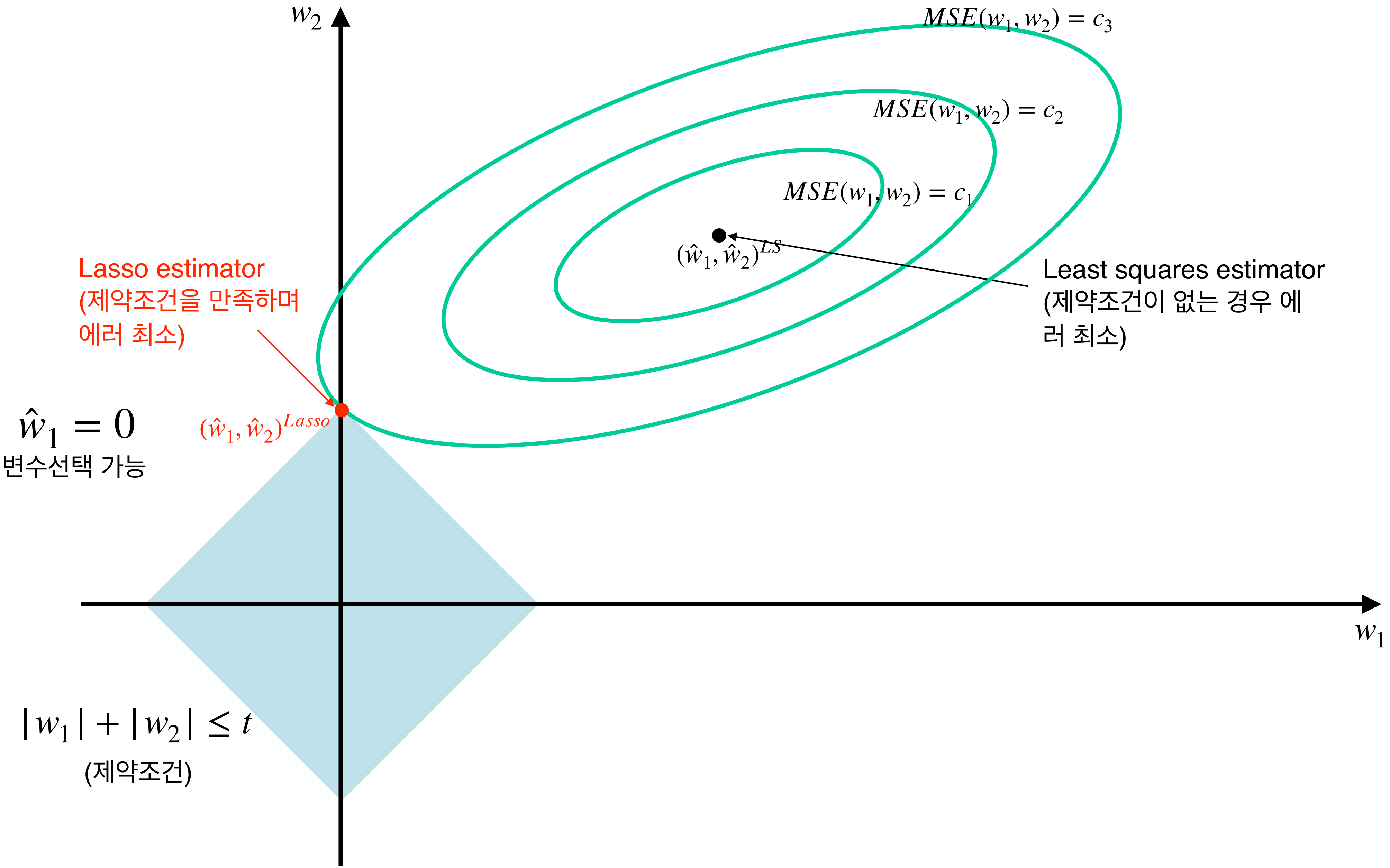
$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2$$

s.t. $|w_1| + \cdots + |w_p| \le t$

⇕ **Equivalent (Lagrangian multiplier)**

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda \sum_{j=1}^{p} |w_j|$$

Lasso Solution path



Lasso Solutions

$$\min \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda \sum_{j=1}^{p} w_j^2 \quad \Rightarrow \quad \hat{\mathbf{w}}^{ridge} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p \right)^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\min \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda \sum_{j=1}^{p} |w_j| \quad \Rightarrow \quad \hat{\mathbf{w}}^{lasso} = ?$$

- Ridge와 달리 Lasso는 closed form solution 을 구하는 것이 불가능 (L_1-norm 미분 불가능)
- Numerical optimization methods:
  - Quadratic programming techniques (1996, Tibshirani)
  - LARS algorithm (2004, Efron et al.)
  - Coordinate descent algorithm (2007, Friedman et al.)

Doo Ho Lee, Dept. of AI & SW

Lasso Parameter

Lasso

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda \sum_{j=1}^{p} |w_j|$$

$\lambda \to 0$: 최소제곱법

$\lambda \to \infty$: $\mathbf{w} \to \mathbf{0}$

λ 값을 어떻게 설정할 것인가?

몇 개의 변수를 선택할 것인가?

Doo Ho Lee, Dept. of AI & SW

Lasso Parameter

λ 값을 크게 잡으면?

적은 변수
간단한 모델
해석 쉬움
높은 학습 오차(training error 증가)
Underfitting 위험 증가

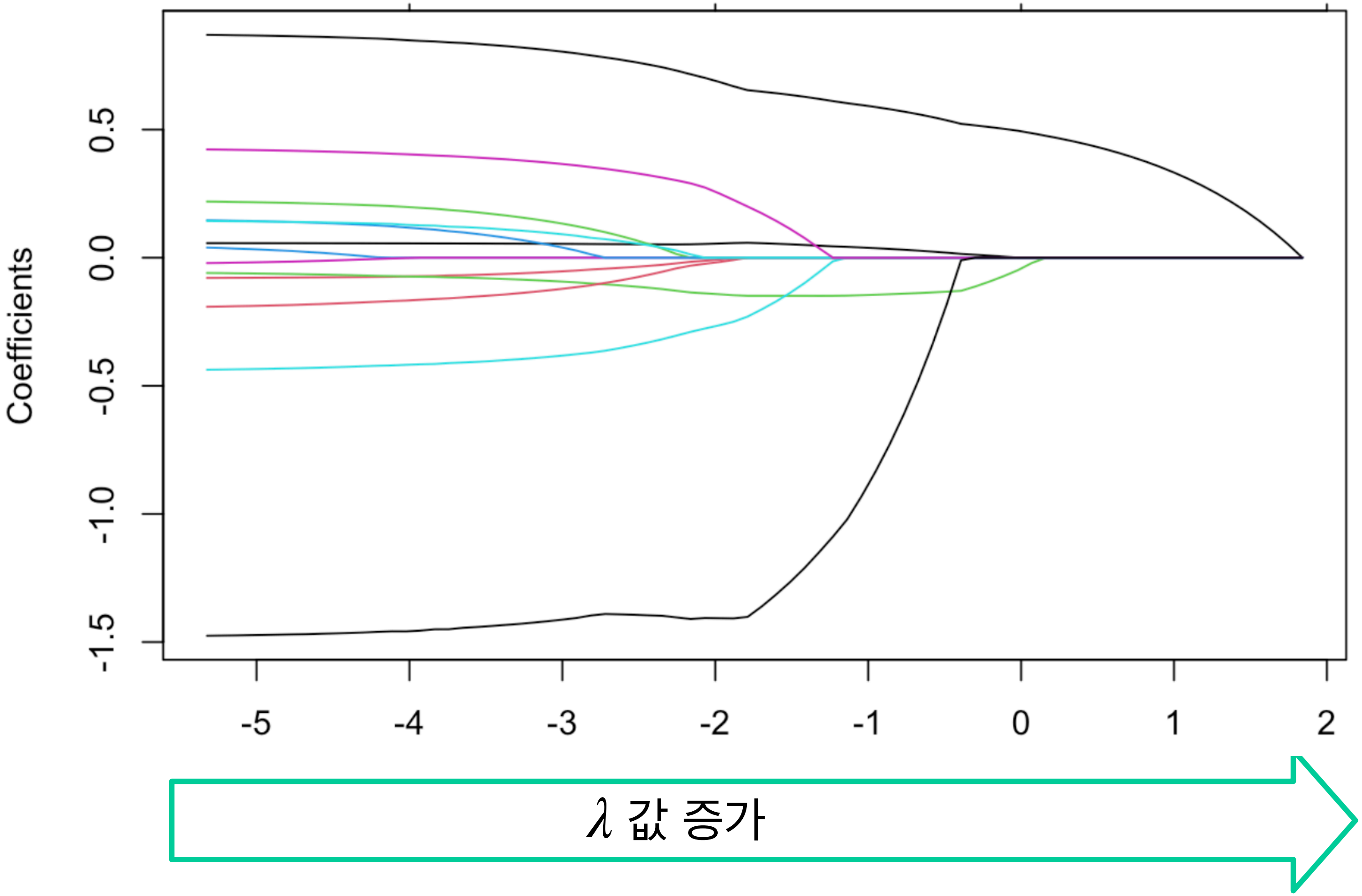
λ 값을 작게 잡으면?

많은 변수
복잡한 모델
해석 어려움
낮은 학습 오차
Overfitting 위험 증가

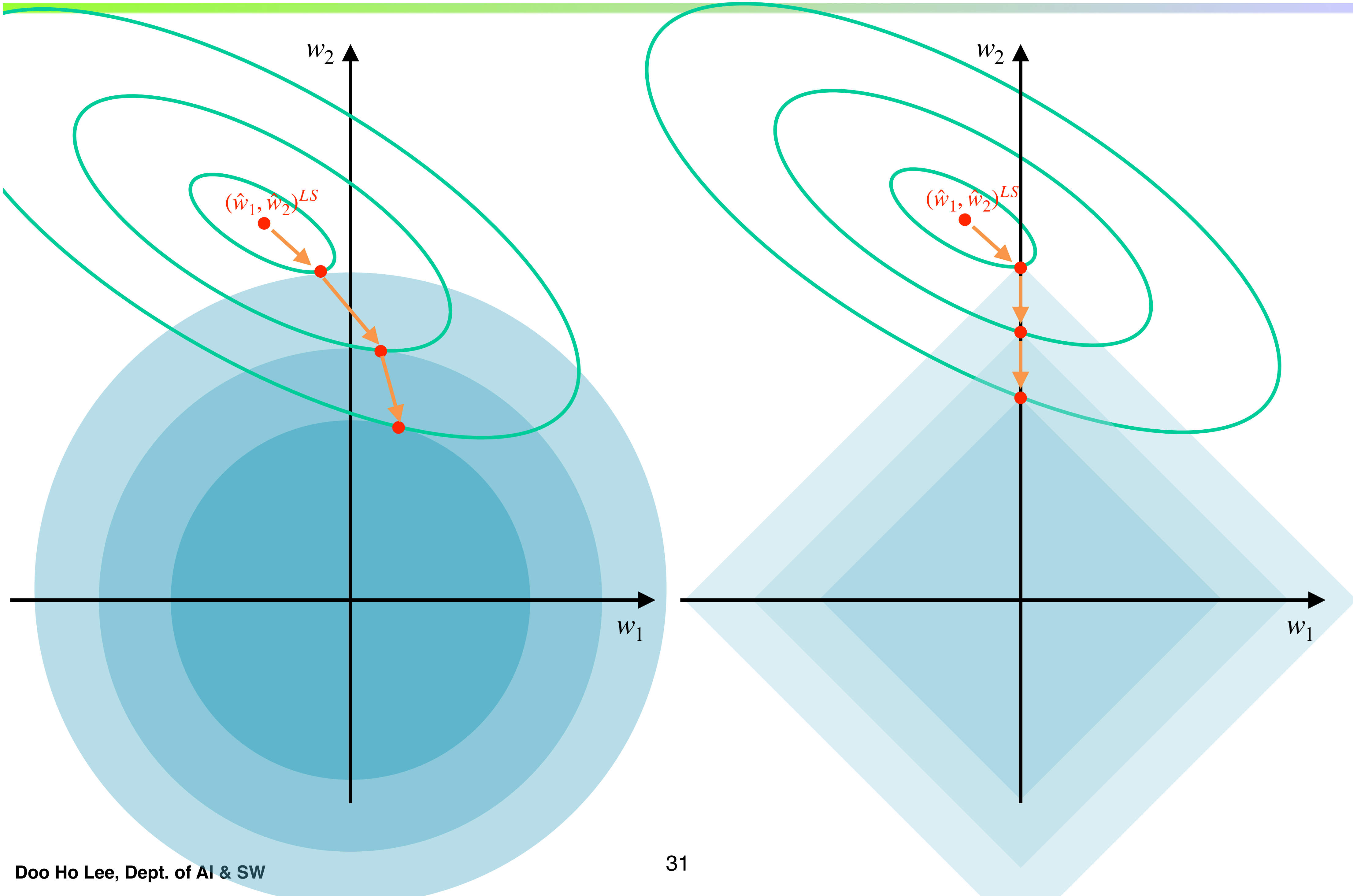
Doo Ho Lee, Dept. of AI & SW

Lasso parameter

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda \sum_{j=1}^{p} |w_j|$$



Solution Paths of Ridge and Lasso



Ridge vs Lasso

Ridge	Lasso
L_1-norm regularization	L_2-norm regularization
변수 선택 불가능	변수 선택 가능
Closed form solution 존재	Closed form solution 존재하지 않음 (수치해석적 방법 이용)
입력변수 간 상관관계가 높은 상황에서 좋은 예측 성능	입력변수 간 상관관계가 높은 상황에서 ridge에 비해 상대적으로 예측성능이 떨어짐

Elastic Net

- Elastic Net = Ridge + Lasso (L_1- and L_2- norm regularization)
- Elastic Net은 상관관계가 큰 입력변수를 동시에 선택/배제하는 특성

Elastic Net

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2$$

$$\text{s.t.} \quad \alpha \sum_{j=1}^{p} |w_j| + (1-\alpha) \sum_{j=1}^{p} w_j^2 \leq t$$

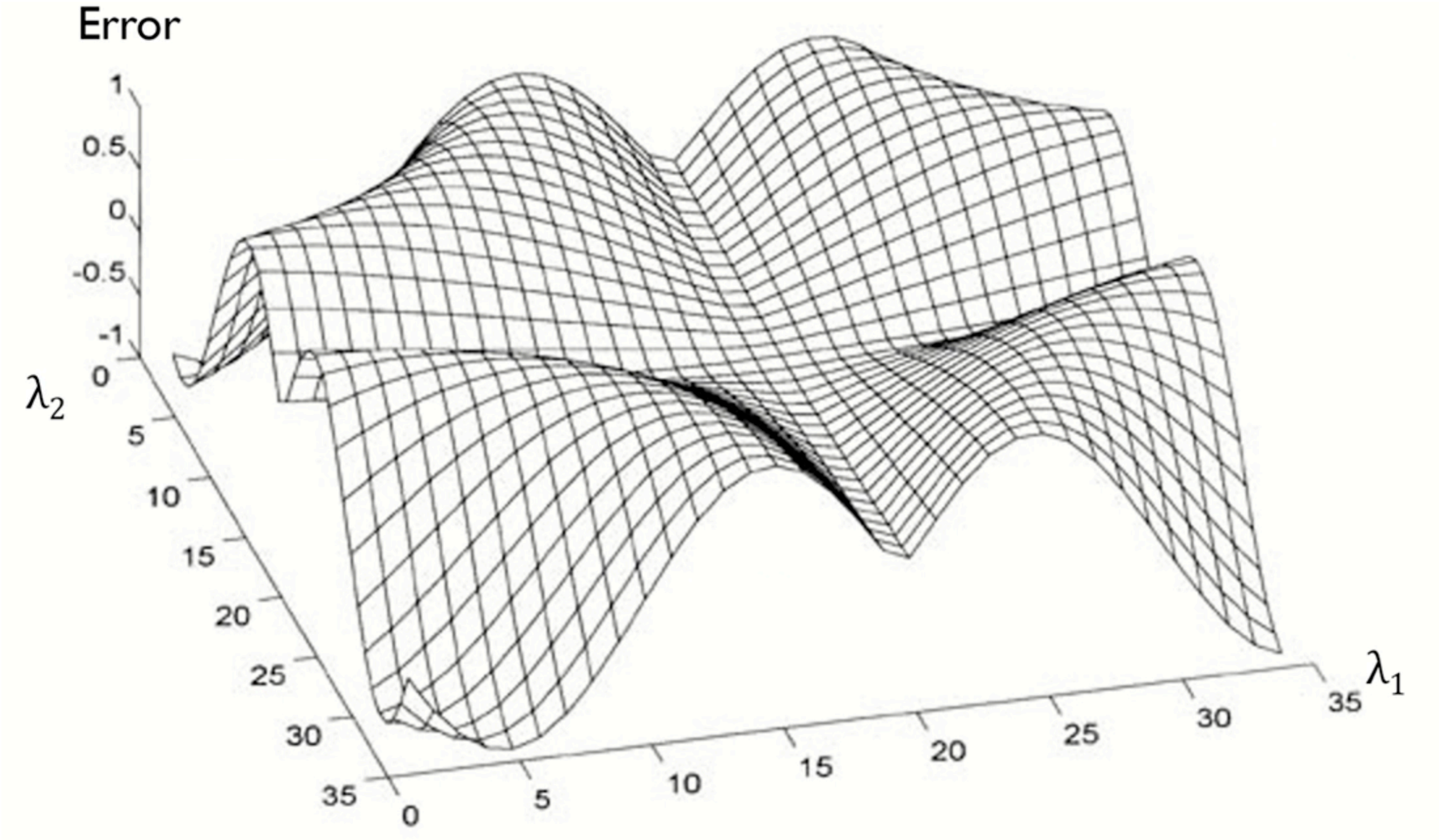
⇕ **Equivalent (Lagrangian multiplier)**

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda \left(\alpha \sum_{j=1}^{p} |w_j| + (1-\alpha) \sum_{j=1}^{p} w_j^2 \right)$$

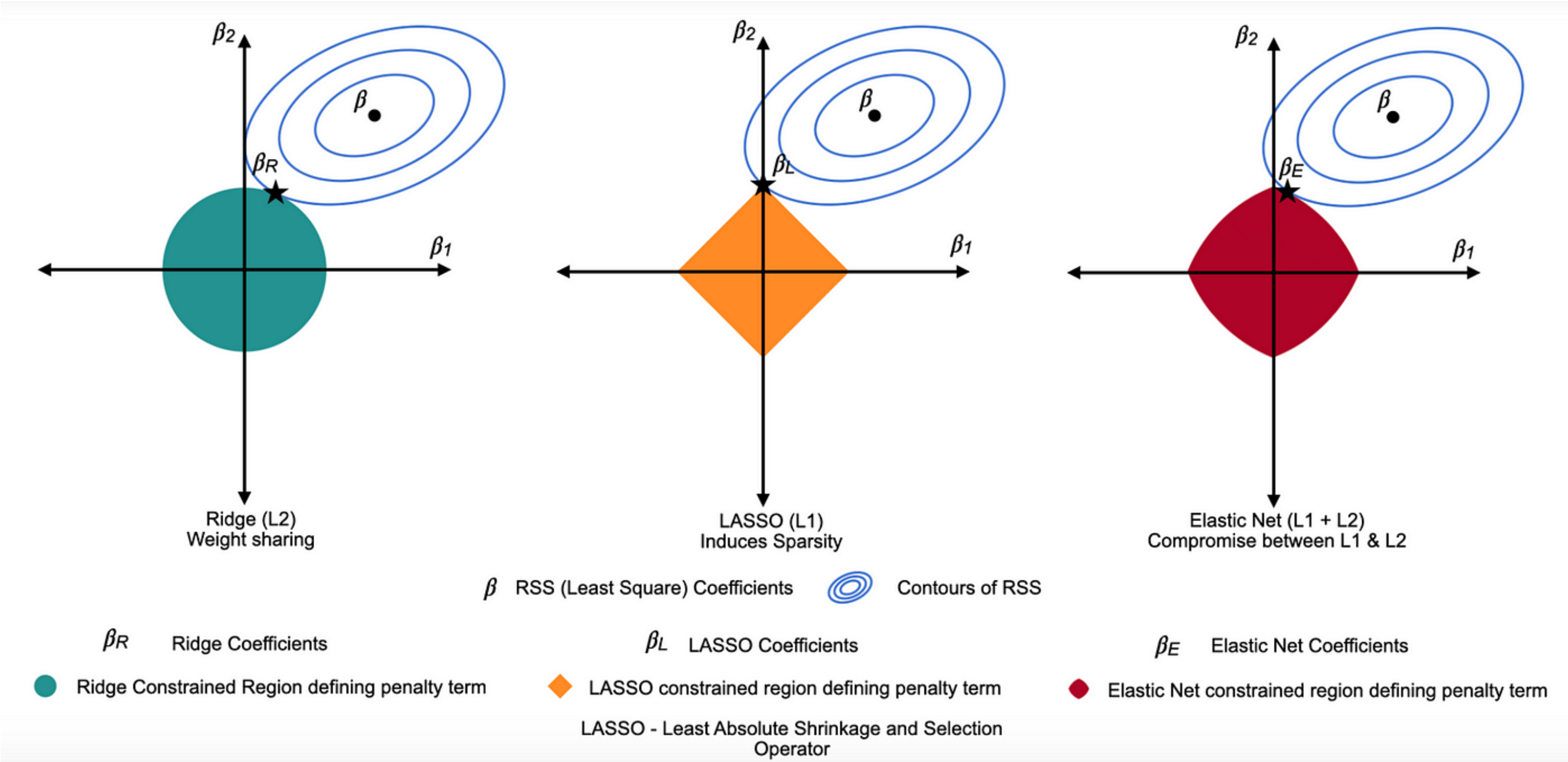
Elastic Net Parameters

$$\min_{w_1,\cdots,w_p} \sum_{i=1}^{n} \left\{ Y_i - (w_1 x_{1i} + \cdots + w_p x_{pi}) \right\}^2 + \lambda_1 \sum_{j=1}^{p} |w_j| + \lambda_2 \sum_{j=1}^{p} w_j^2$$

- 일정 범위 내로 λ_1과 λ_2를 조정하여 오차가 가장 작은 결과를 보이는 λ_1과 λ_2 값을 선정함



Ridge vs. Lasso vs. Elastic Net



Doo Ho Lee, Dept. of AI & SW