기계학습 (Machine Learning)

L10
- Regularization

한밭대학교

정보통신공학과

최 해 철

ToC

◆ Overfitting & Underfitting

◆ Bias and Variance

◆ Regularization by Weight Penalty

References

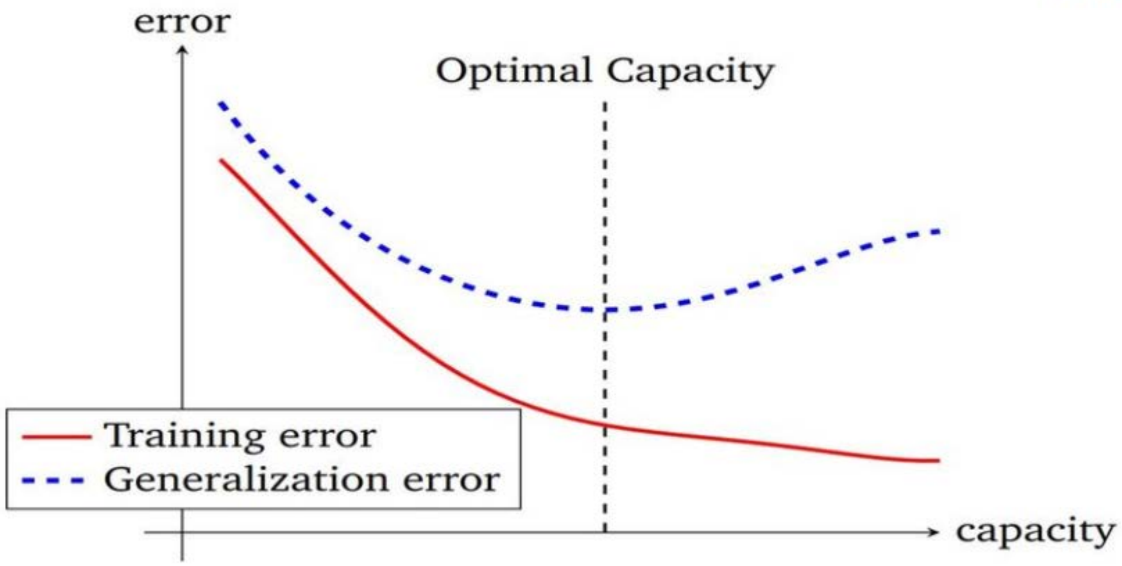
- *기계 학습 "3장 다층 퍼셉트론"* by 오일석, *패턴 인식* by 오일석
- 단단한 머신러닝 by 조우쯔와

1. Overfitting & Underfitting

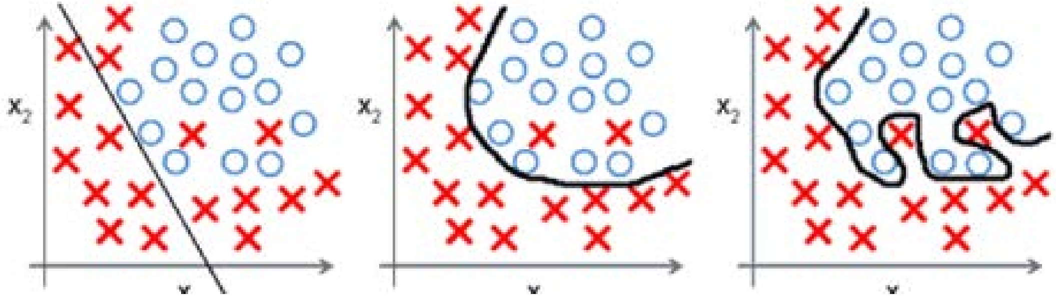
Generalization

◆ 일반화

● 모델이 학습 데이터에 대해 학습한 후, **이전에 본 적이 없는** 새로운 데이터**에 대해 정확하게 예측할 수 있는 능력**



• *Example*: in Classification...



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$

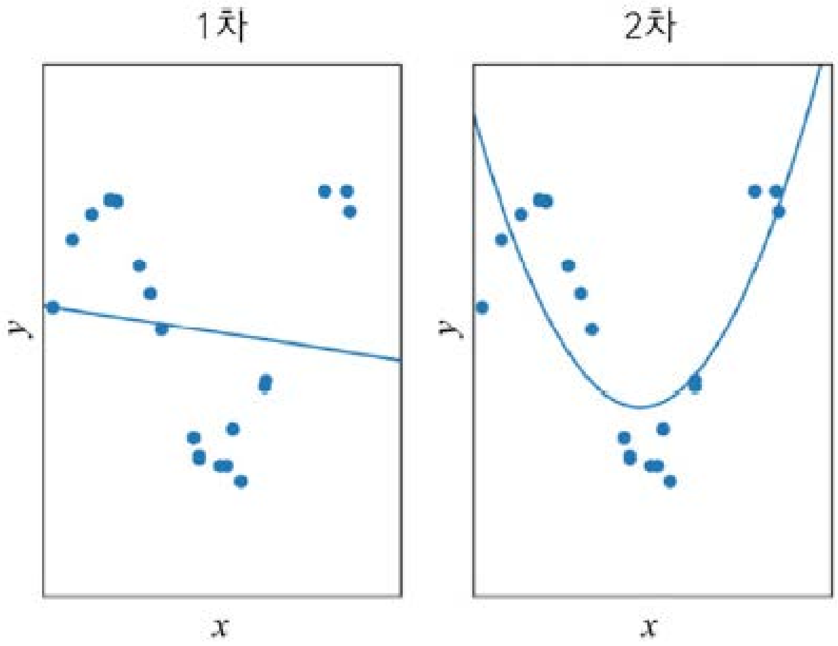
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2^2 + \cdots)$$

Underfitting

◆ 과소적합과 훈련 오차

- '모델의 용량이 너무 작아' or '훈련집합이 너무 작아' 오차가 클 수밖에 없는 현상
  - 예) 아래 그림의 선형(1차 다항식) 또는 2차 다항식 모델을 사용한 경우

$$h_\theta(x) = g(\theta_0 + \theta_1 x)$$

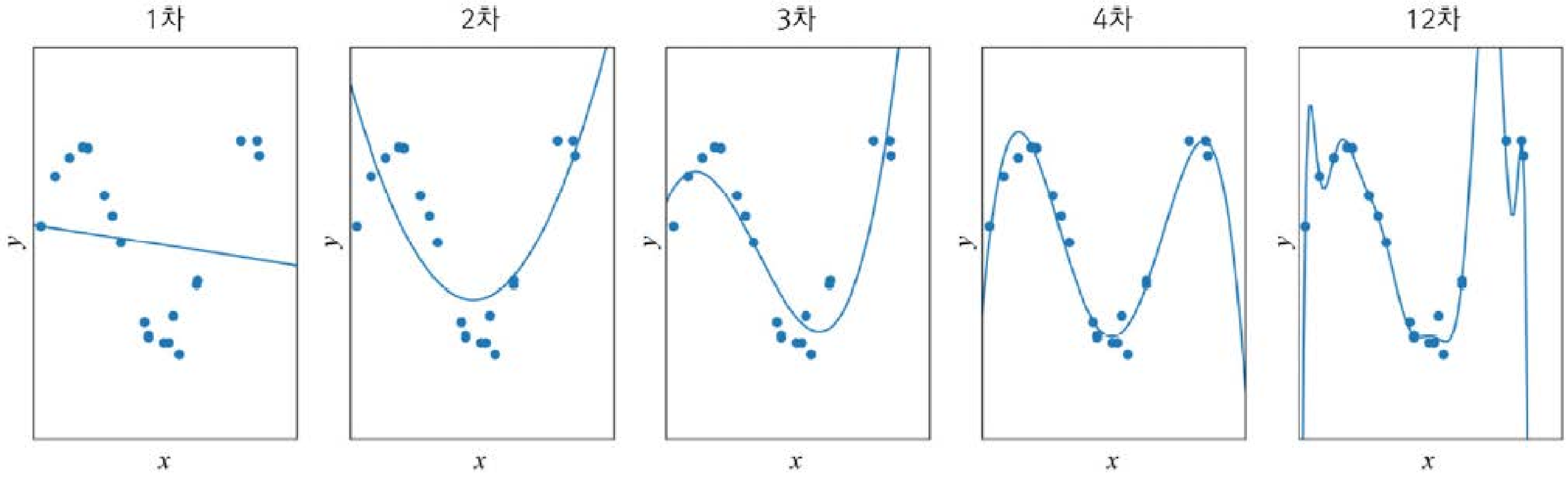


$$h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x^2)$$

Underfitting

◆ Underfitting 방지

- 비선형 모델 등과 같이 **용량이 더 큰 모델**을 사용
- 충분한 훈련 집합을 활용
  - 예) 아래 그림의 3차, 4차, 12차 다항식 모델의 경우

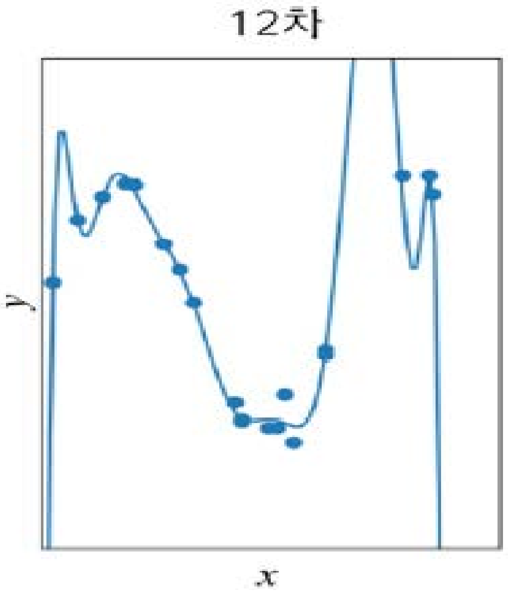


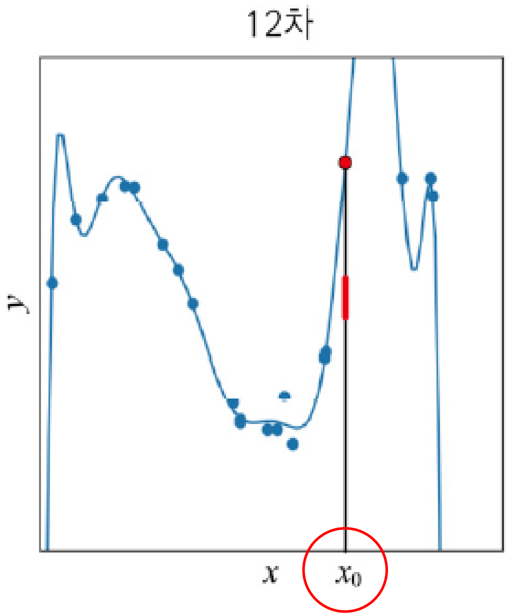
$$h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x^2 + \cdots \theta_{11} x^{11} + \theta_{12} x^{12})$$

Overfitting

◆ 과적합 (overfitting)과 예측(시험) 오차

- 12차 다항식 곡선을 채택한다면 훈련집합에 대해 거의 완벽하게 근사화함
- 하지만 '새로운' 데이터를 예측한다면 큰 문제 발생 (빨간 점)
- 이유는 '용량이 너무 크기' 때문에 학습 과정에서 잡음까지 수용 → 과잉적합 현상





$$\hat{y} = f(x) + \textbf{noise}$$

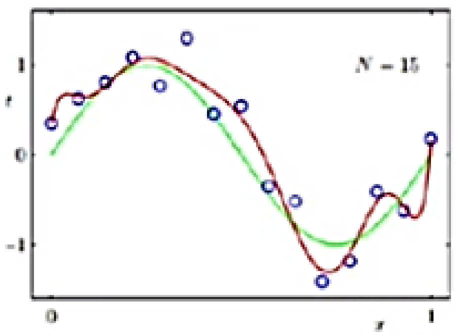
Causes of Overfitting – 1. Data

◆ Insufficient # of Training Examples

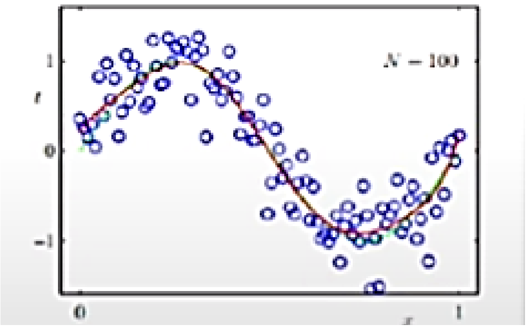
- the training set may be **too** sparse or cannot represent the full variety of the data

데이터가 듬성듬성 있다 (데이터 부족할 때)

N=15



N=100



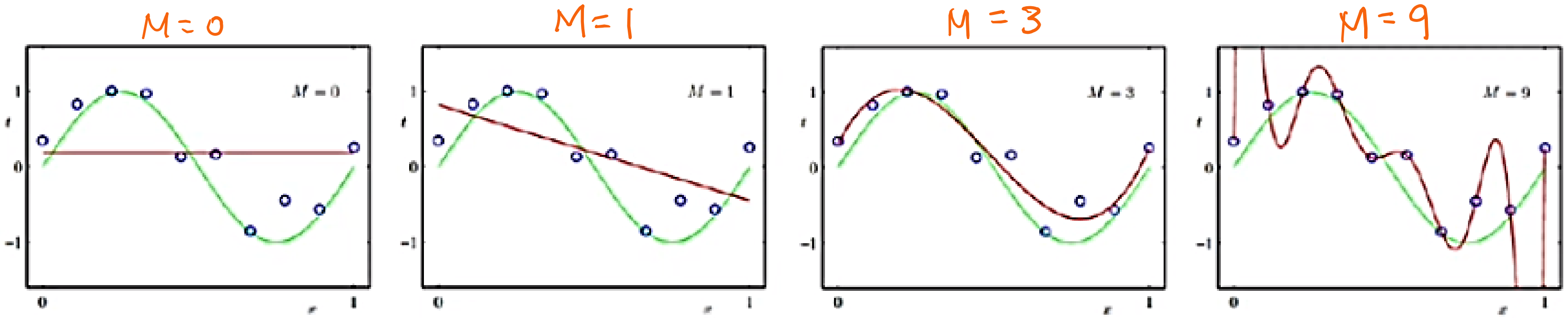
N : # of traing examples

- **해결책**: **충분히 많은 Training Data** 사용
  - Cf.) 데이터 증대(Data Augmentation) 기법 등을 통해 기존 Training Data 증대 가능

Causes of Overfitting – 2. Model

◆ Too Large # of Parameters (Model Capacity)

- the model is relatively too flexible for the dataset
- the resulting parameters tend to have large values



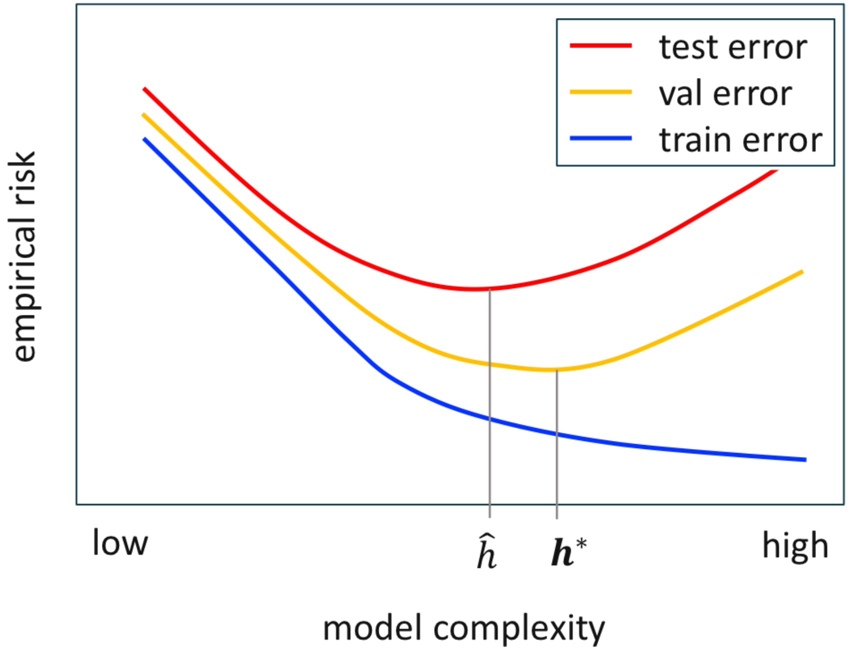
	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

3차 예, $h_\theta(x) = g(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$

Causes of Overfitting – 2. Model (cont'd)

◆ 해결책 1: 검증집합 을 이용한 모델 선택

- 훈련집합과 테스트집합과 다른 별도의 **검증집합**을 준비한다.
- **모델집합**에 속한 각각의 모델에 대해 **훈련집합**으로 학습시킨다. (훈련 성능)
  - 앞의 예에서는 서로 다른 차수의 다항식의 집합(서로 다른 용량)이 모델집합인 셈
- **검증집합**에 대해 최고의 성능을 보인 모델을 선택한다. (검증 성능) → *Overfitting 방지*



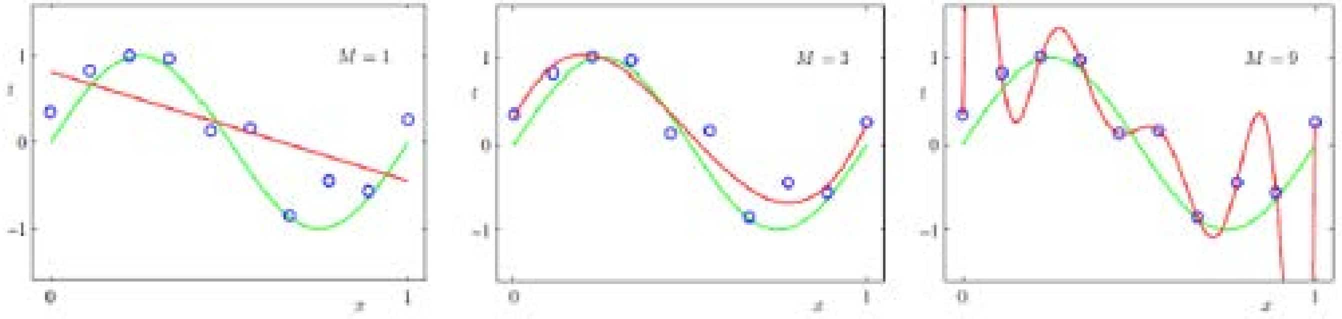
h^* : model with lowest validation error
$\hat{h}$: model with lowest test error

Overfitting

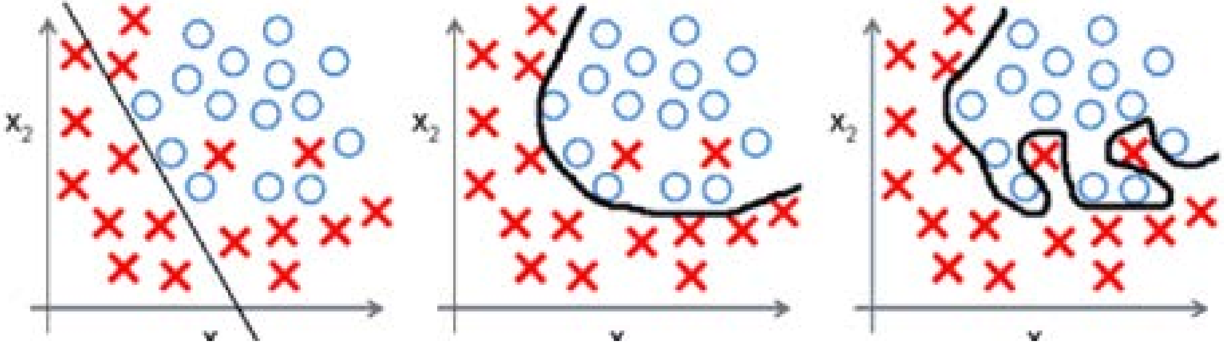
◆ 해결책 2: 적당한 용량 의 모델을 선택

- Model selection, model evaluation 작업을 수행

  - *Example*: in Regression...



  - *Example*: in Classification...



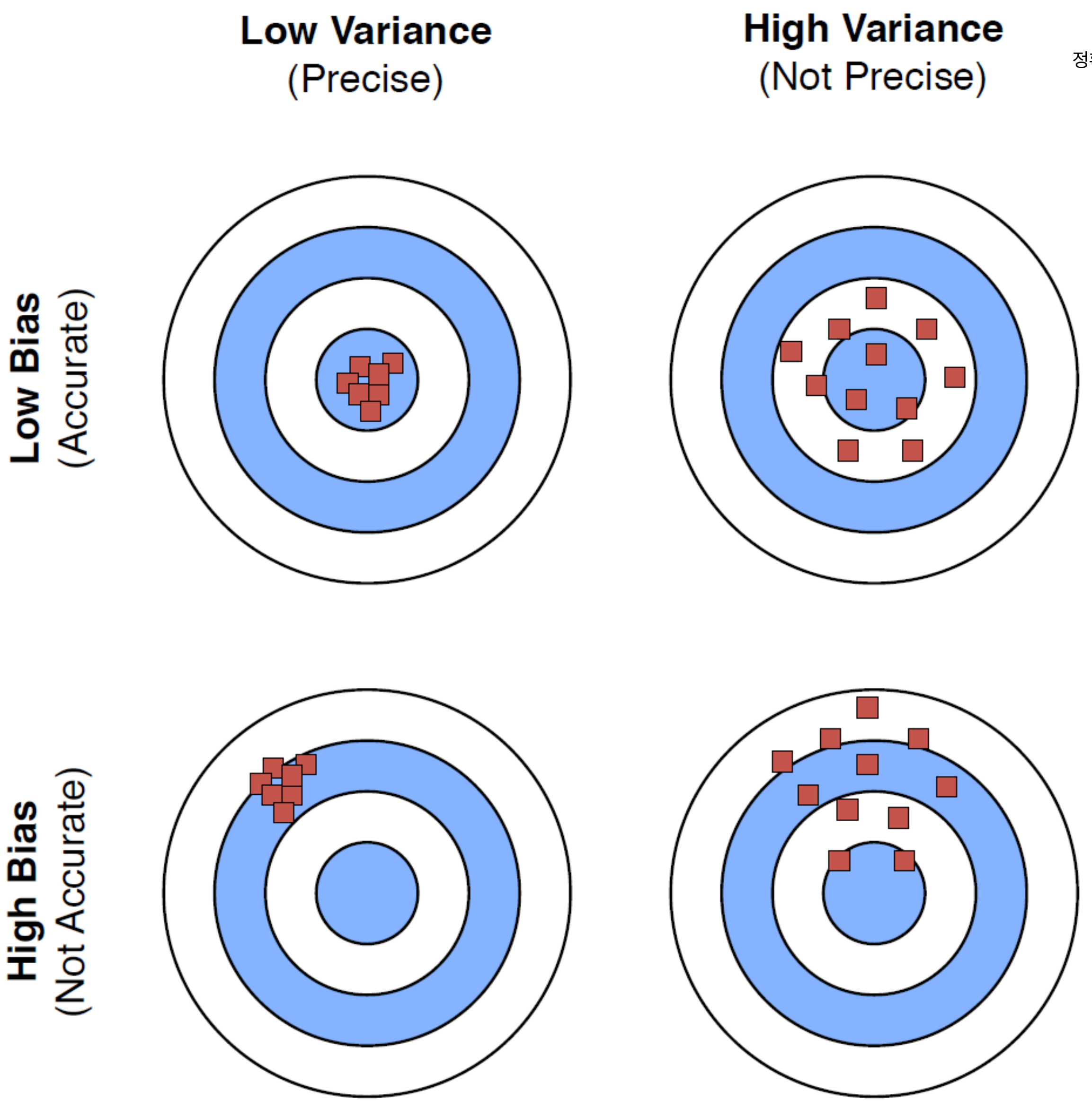
Causes of Overfitting – 2. Model (cont'd)

◆ 해결책 3: 규제

- **용량이 충분히 큰 모델 + 다양한 규제(*Regularization*) 기법을 적용**
  - 예시: Weight Penalty, Drop-out…
  - Overfitting을 방지하기 위한 기술을 통칭하여 '규제'라고 부르기도 함
  - **Regularization Parameter들은 *Validation Set* 을 이용하여 결정 가능 (model selection)**

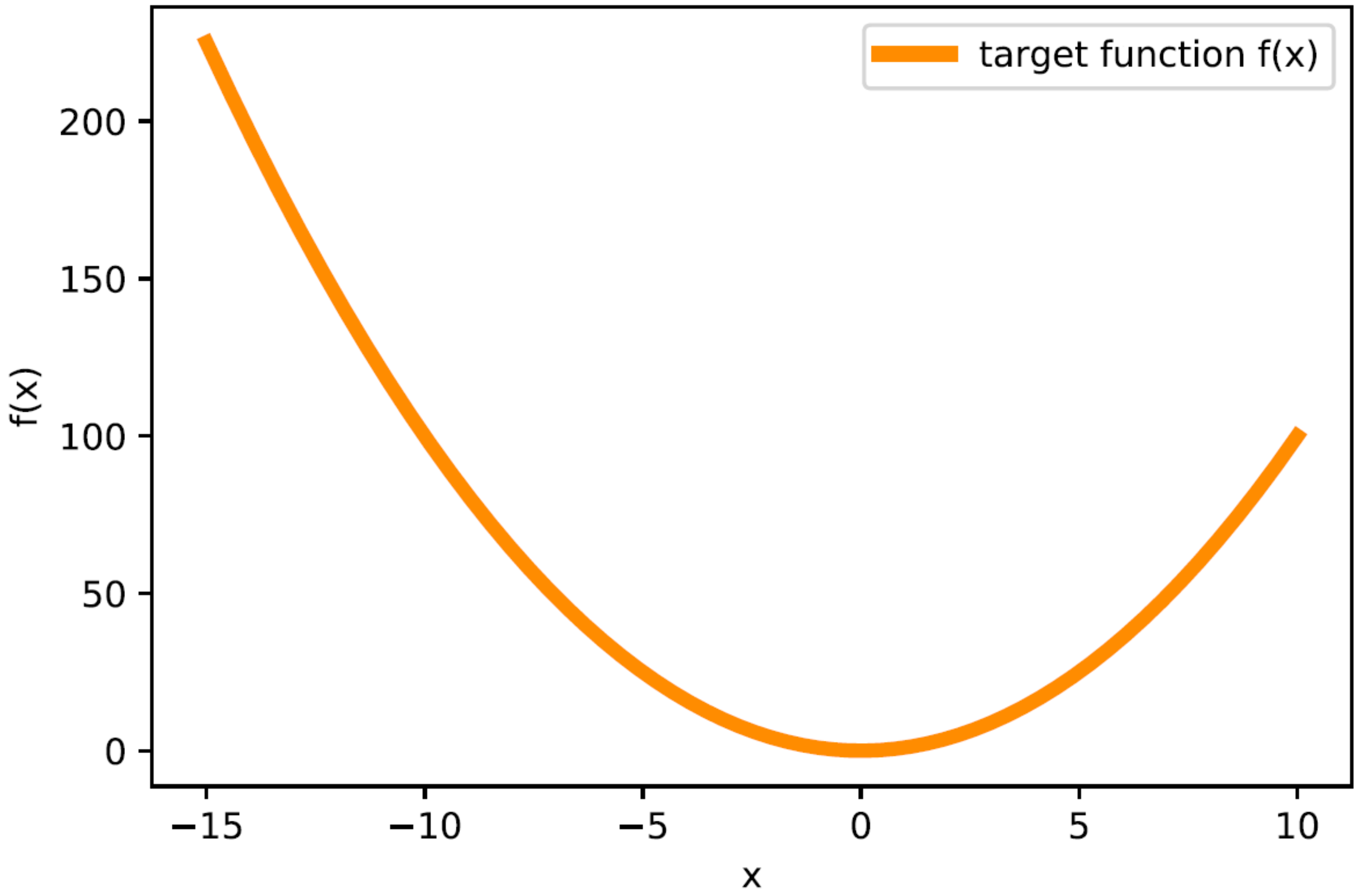
2. Bias and Variance

Bias-Variance 직관

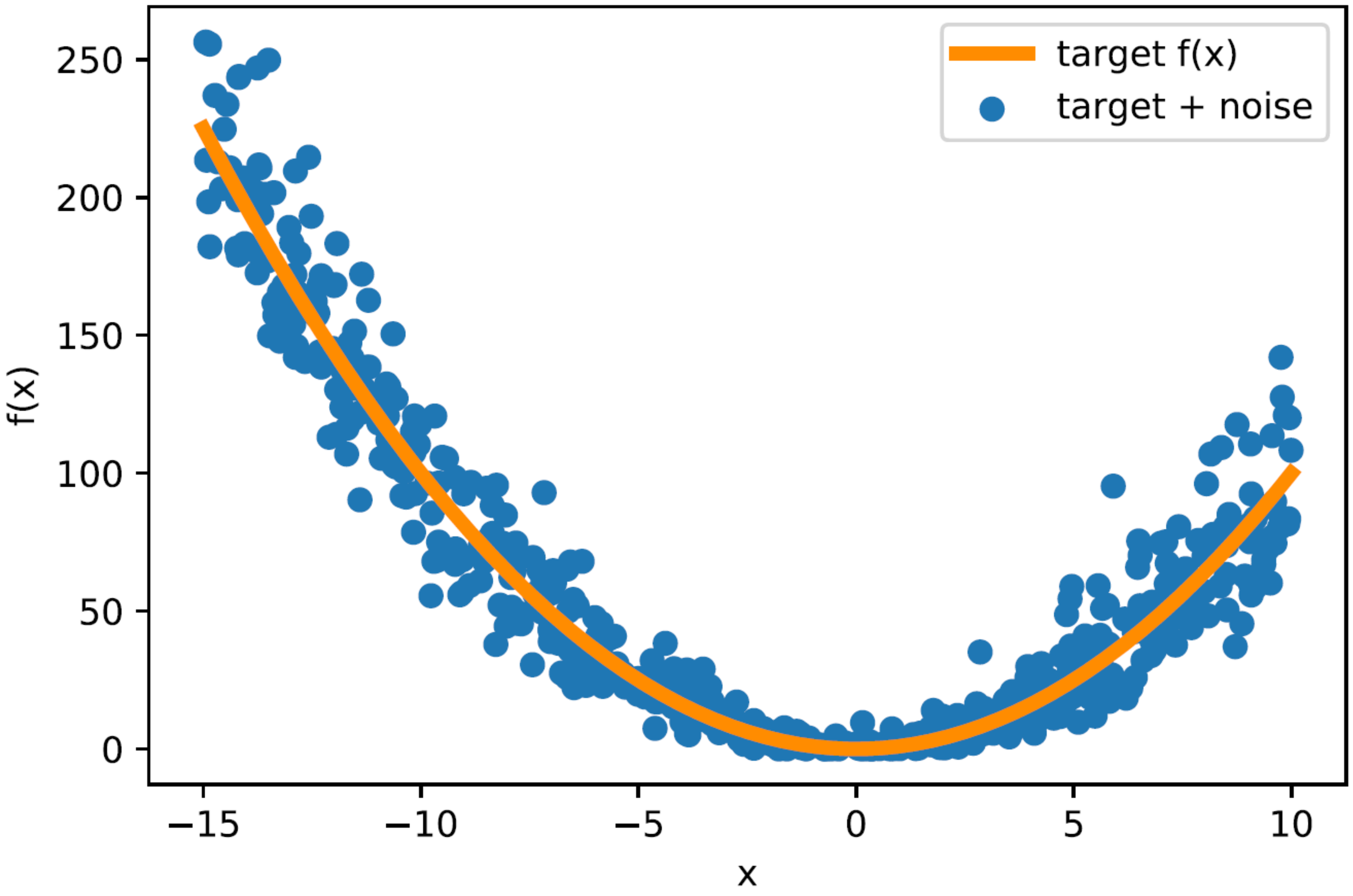


정확도와 정밀도의 차이

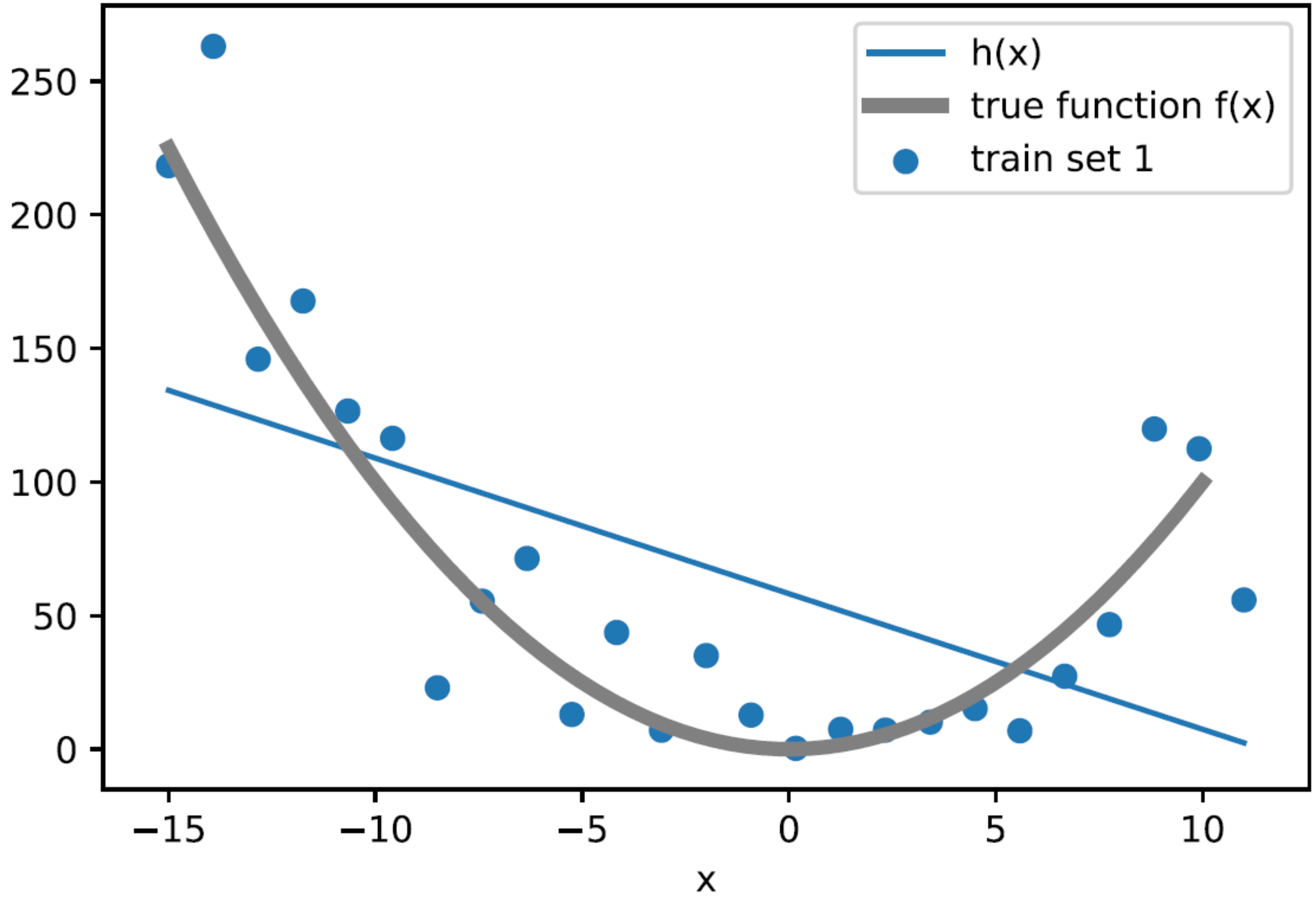
Bias-Variance 직관



Bias-Variance 직관

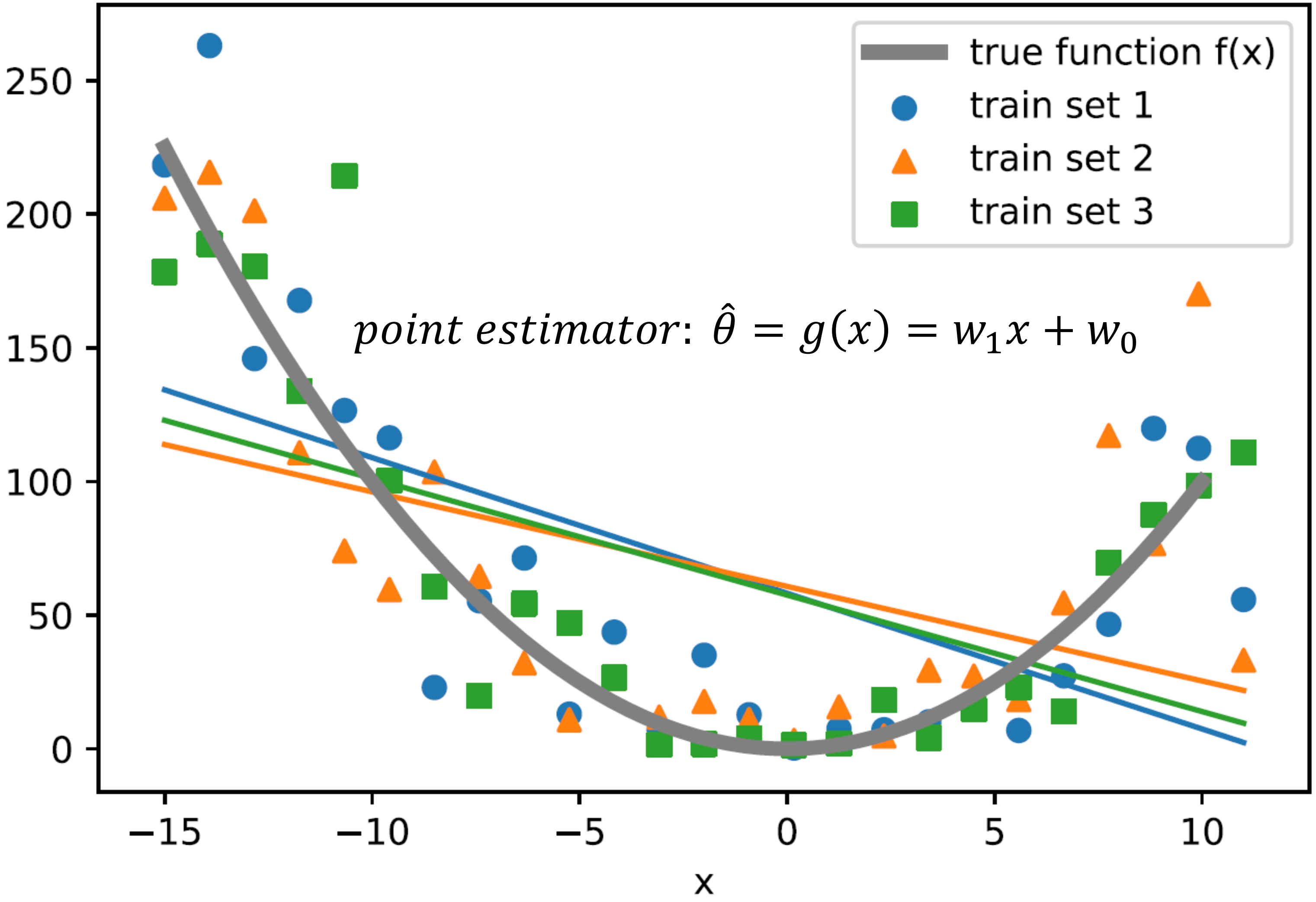


Bias-Variance 직관

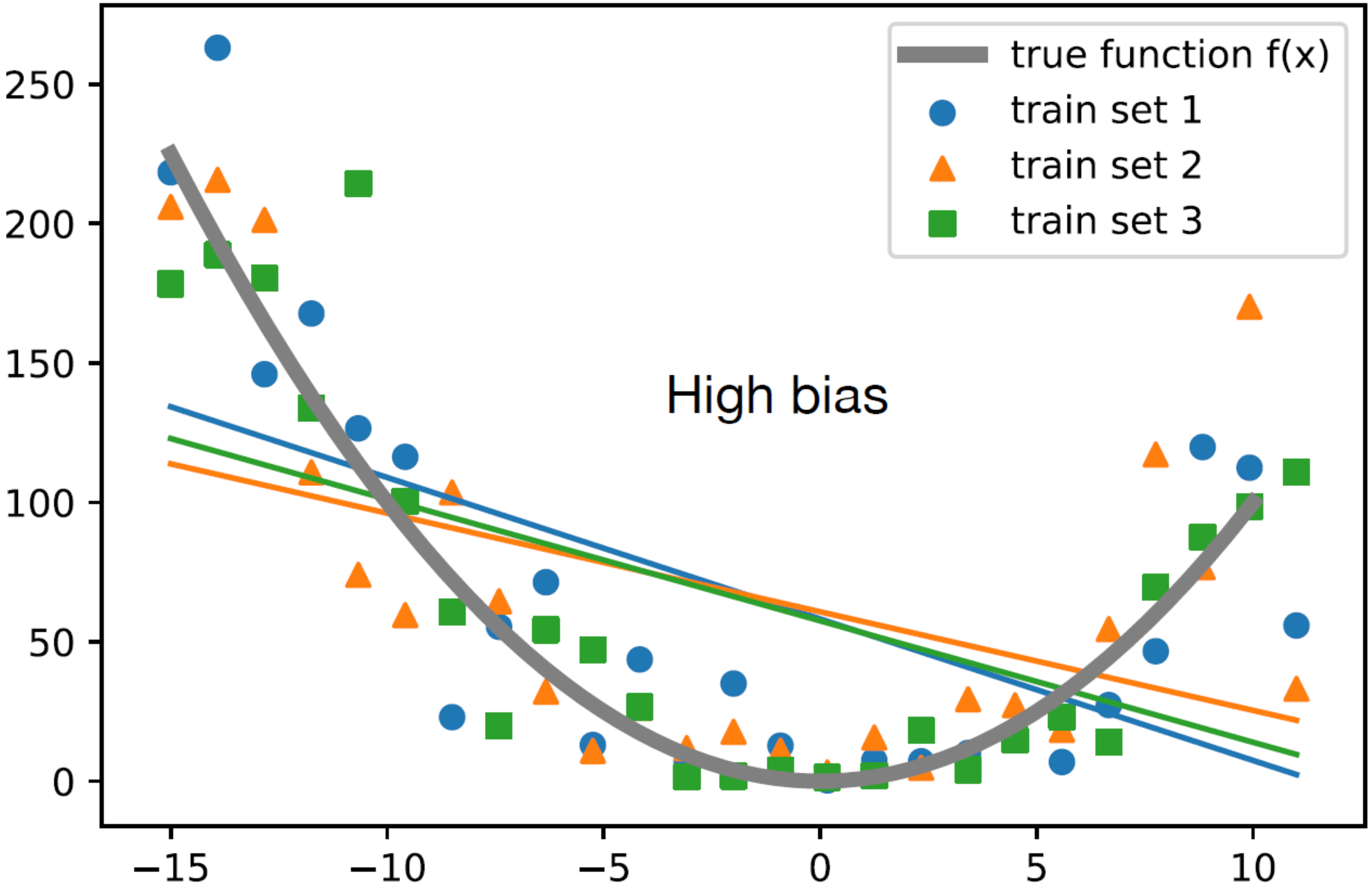


Bias-Variance 직관

Bias-Variance 직관



Bias-Variance 직관

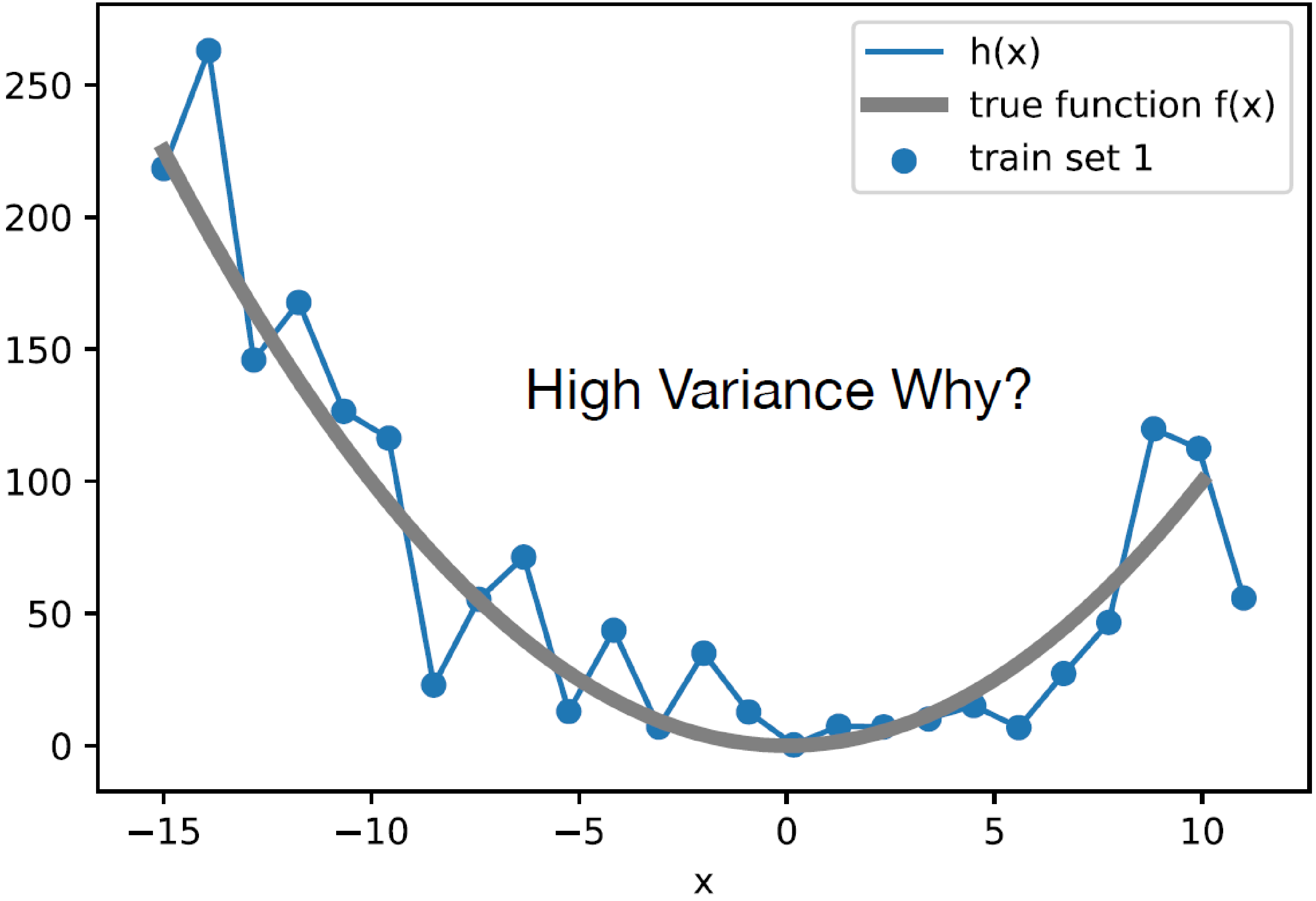


(There are points where the bias is zero ...)

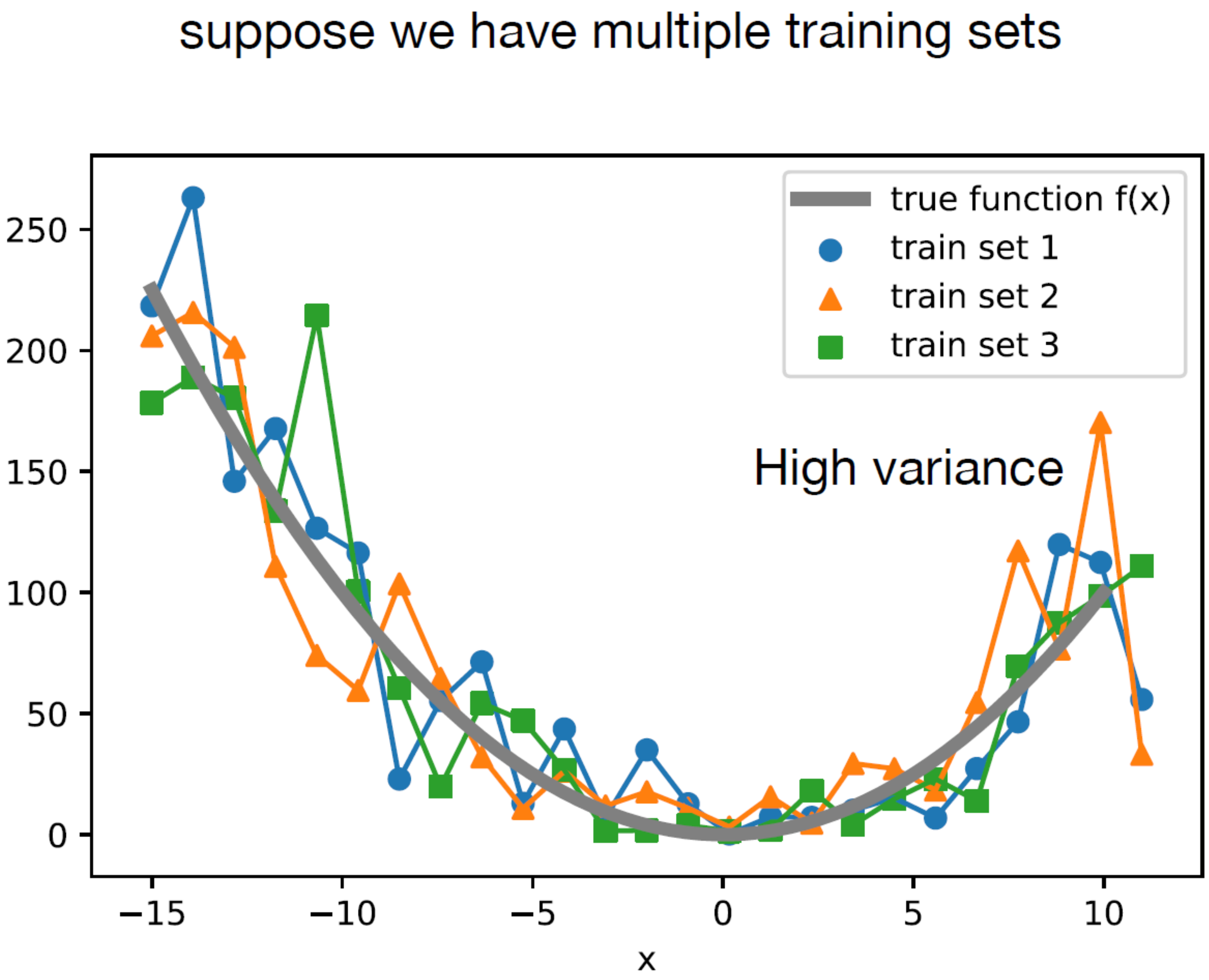
V ↓

B ↑

Bias-Variance 직관



Bias-Variance 직관



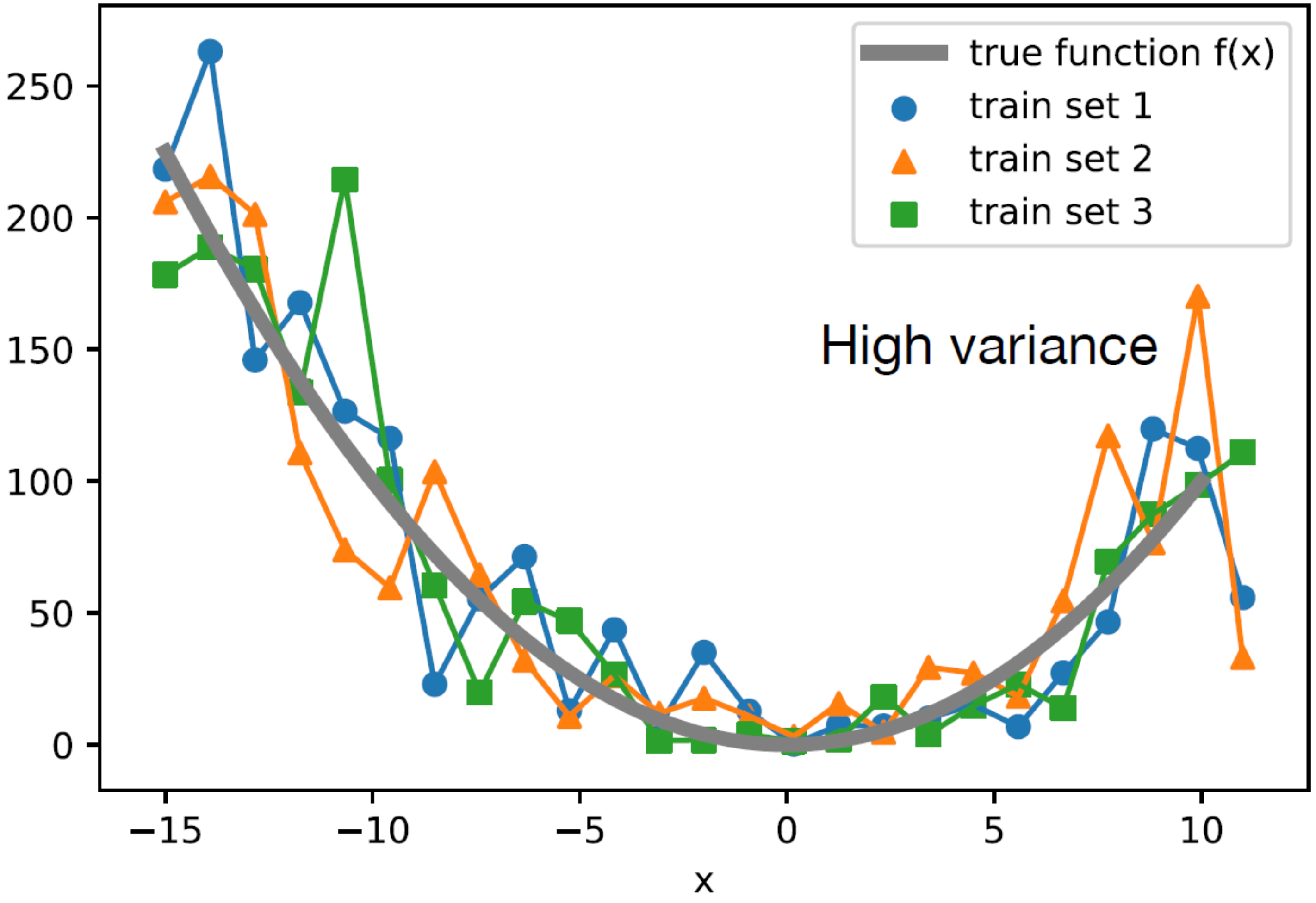
V ↑

B ↓

Bias-Variance 직관

평균의 취하면?

suppose we have multiple training sets



Bias-Variance 정의

◆ A point estimator $\hat{\theta}$ of some parameter or function θ

Bias-Variance of the Squared Error

$$\mathrm{Bias}[\hat{\theta}] = E[\underset{\text{추정치}}{\underline{\hat{\theta}}}] - \underset{\text{정답}}{\underline{\theta}}$$

$$\mathrm{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

$$\mathrm{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

"ML Notation" for Squared Error Loss

$y = f(x)$ target ← For simplicity, we ignore the noise term

$\hat{y} = \hat{f}(x) = h(x)$ prediction

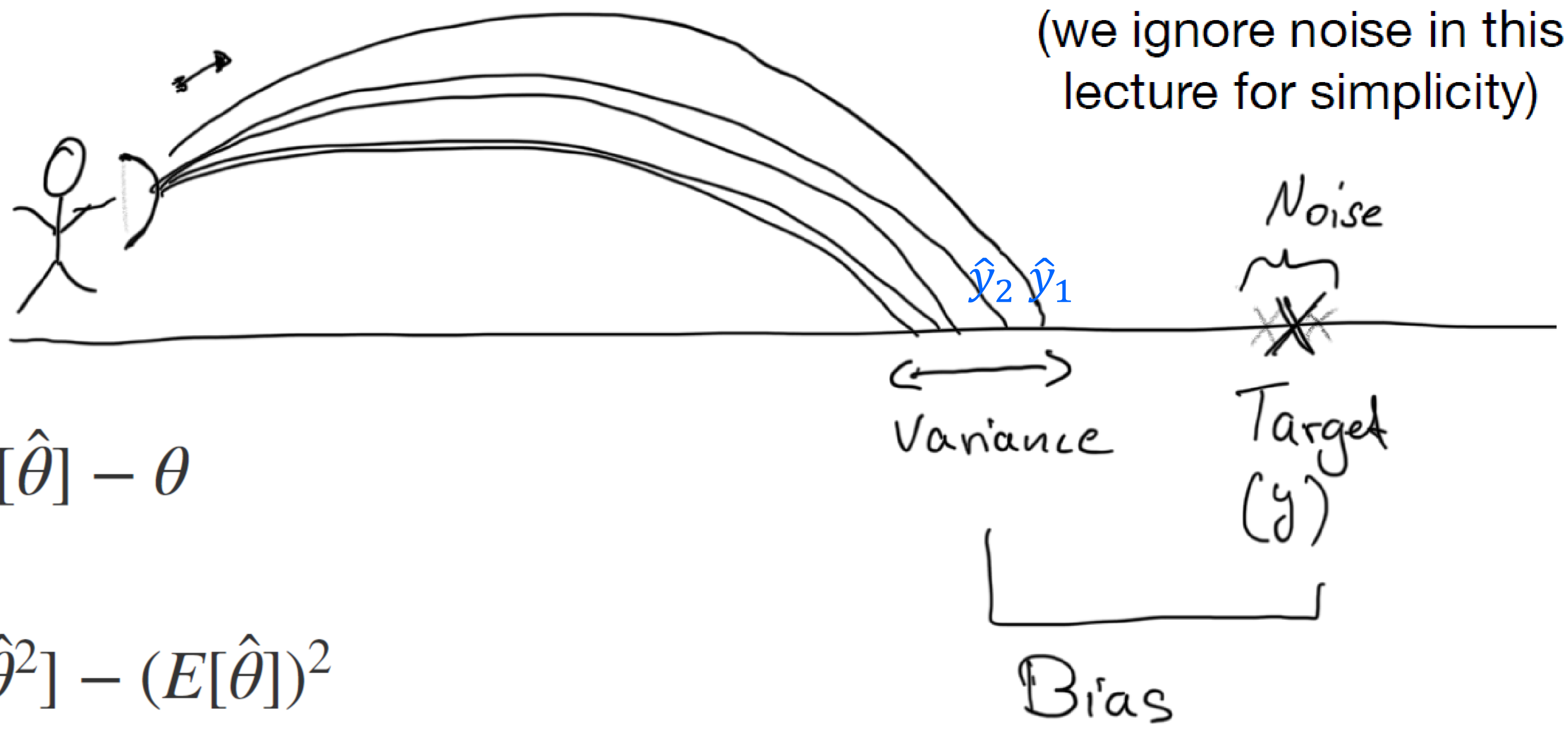
$S = (y - \hat{y})^2$ squared error

- Expectation: 확률 변수의 평균값 또는 평균적인 결과를 나타내는 개념
- The **expectation** is over the training data, i.e, the average estimator from different training samples

Bias-Variance 정의

Intuition

(we ignore noise in this lecture for simplicity)

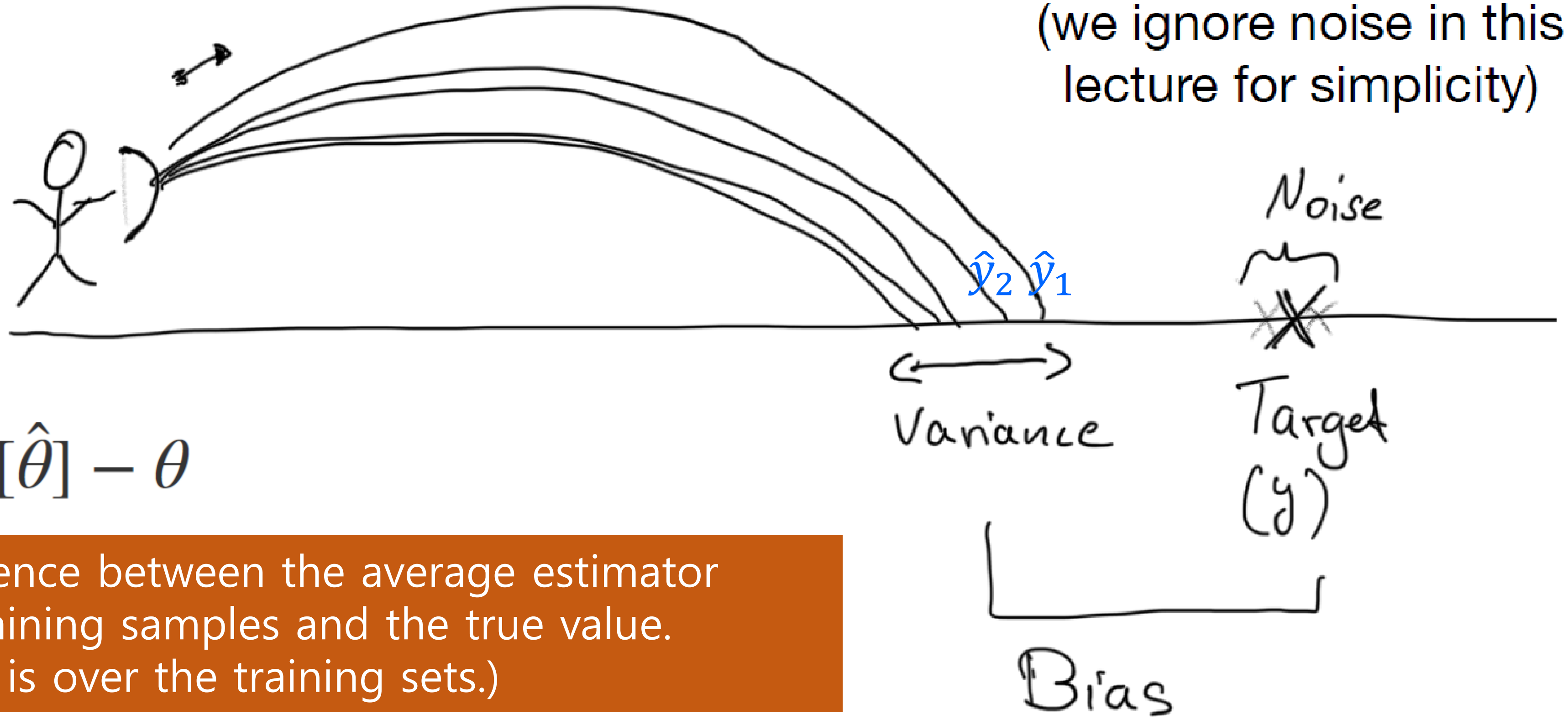


$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

Intuition

(we ignore noise in this lecture for simplicity)

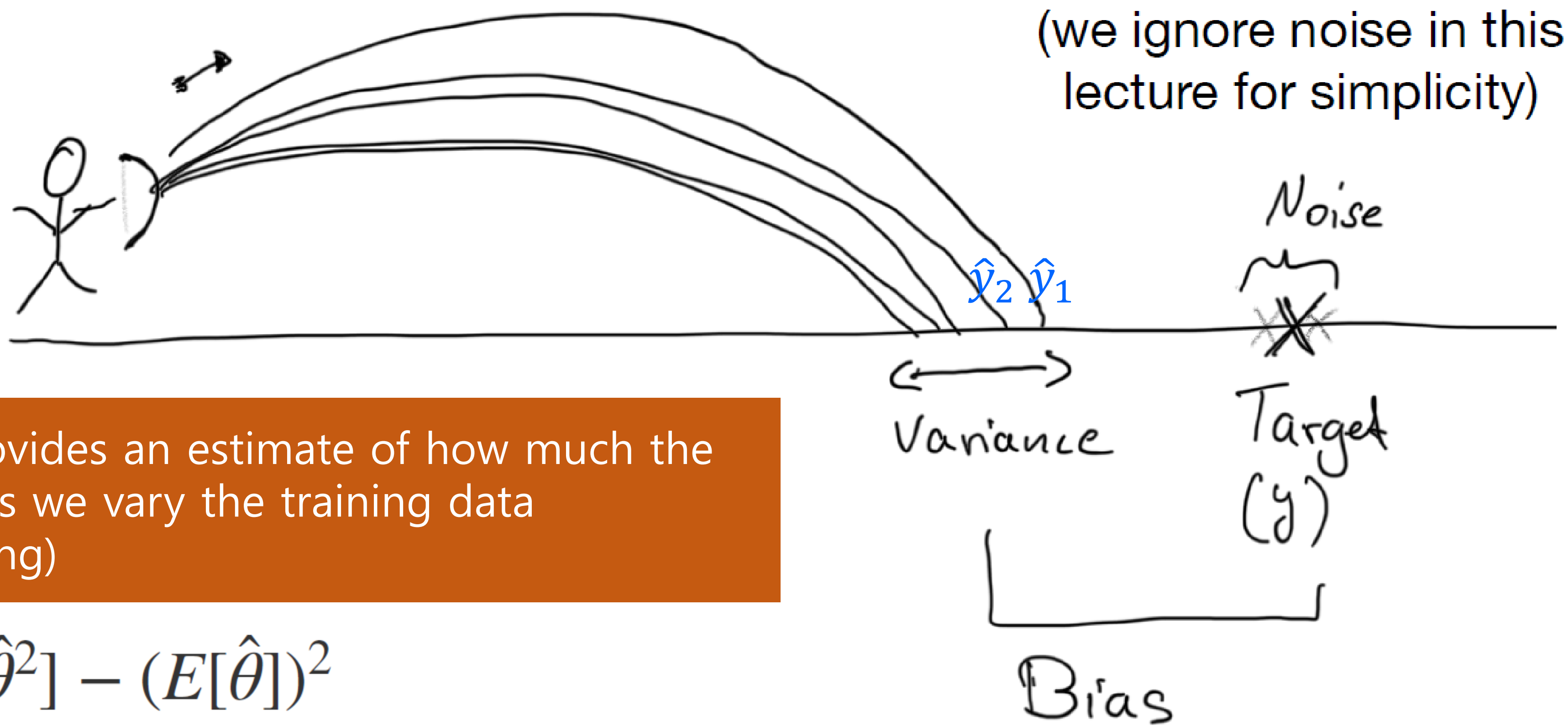


$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

Bias is the difference between the average estimator from different training samples and the true value. (The expectation is over the training sets.)

Bias-Variance 정의

Intuition



The **variance** provides an estimate of how much the estimate varies as we vary the training data (e.g. by resampling)

$$\mathrm{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

Bias-Variance Decomposition

Loss = Bias + Variance + Noise

Bias-Variance of the Squared Error

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

$y = f(x)$ target

$\hat{y} = \hat{f}(x) = h(x)$ prediction

$S = (y - \hat{y})^2$ squared error

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})\end{aligned}$$

Bias-Variance of the Squared Error

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

$y = f(x)$ target

$\hat{y} = \hat{f}(x) = h(x)$ prediction

$S = (y - \hat{y})^2$ squared error

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + 2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})\end{aligned}$$

$$E[S] = E\left[(y - \hat{y})^2\right]$$

Bias-Variance of the Squared Error

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

$$\text{Var}[\hat{\theta}] = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$$

$$\text{Var}[\hat{\theta}] = E\left[(E[\hat{\theta}] - \hat{\theta})^2\right]$$

$y = f(x)$ target

$\hat{y} = \hat{f}(x) = h(x)$ prediction

$S = (y - \hat{y})^2$ squared error

$E[\,] = 0$

$$\begin{aligned}(y - \hat{y})^2 &= (y - E[\hat{y}] + E[\hat{y}] - \hat{y})^2 \\ &= (y - E[\hat{y}])^2 + (E[\hat{y}] - \hat{y})^2 + \boxed{2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})}\end{aligned}$$

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

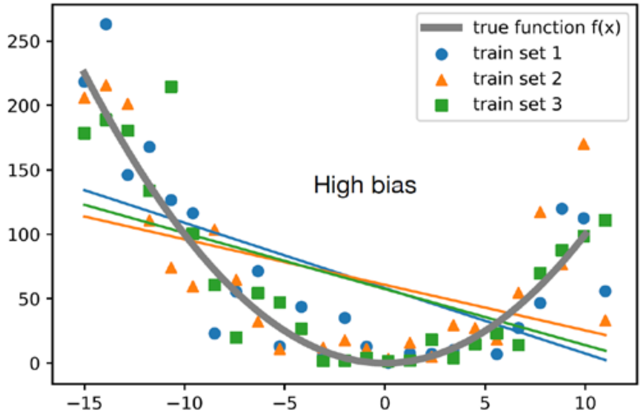
$=$ Bias^2 _ Var

Bias-Variance of the Squared Error

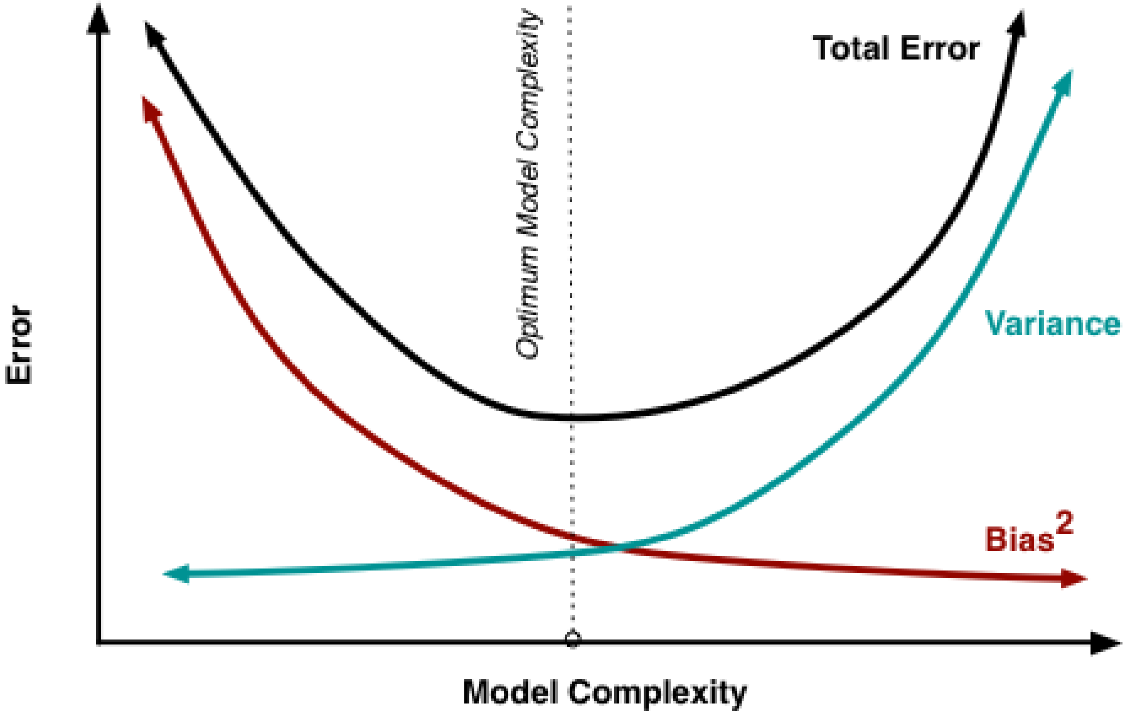
$$
\begin{aligned}
E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] &= 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] \\
&= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})] \\
&= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}]) \\
&= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}]) \\
&= 0
\end{aligned}
$$

Bias-Variance Tradeoff

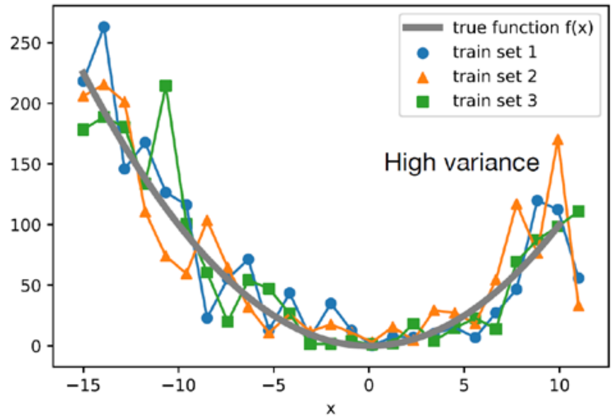
◆ 모델의 복잡도 관점에서 봤을 때 분산과 편향이 트레이드 오프(trade-off) 관계



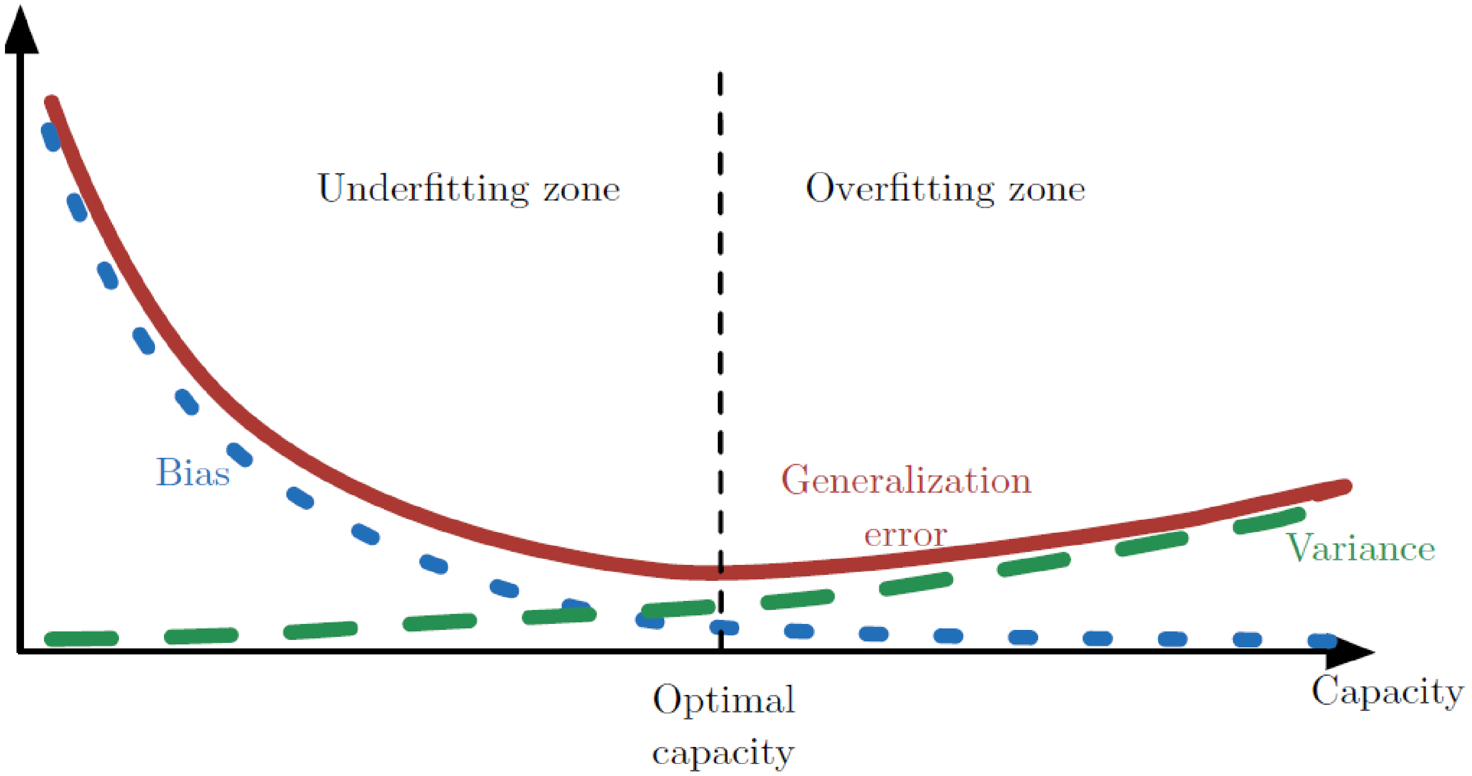
(There are points where the bias is zero ...)

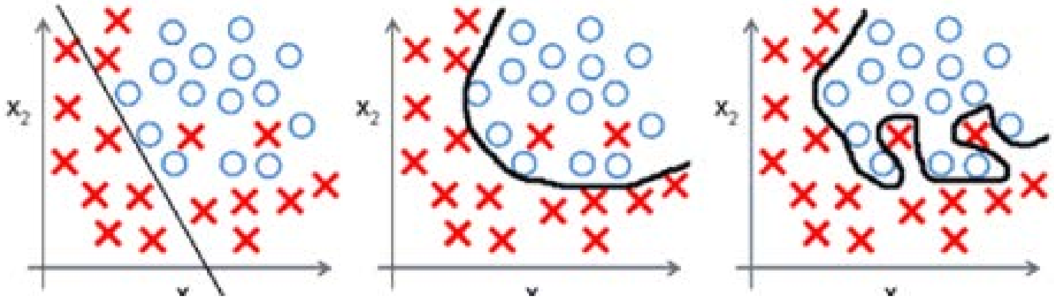


suppose we have multiple training sets



Generalization Error





3. Regularization by Weight Penalty

Regularization (규제)

◆ 『Deep Learning』 책의 규제 정의

- "...any modification we make to a learning algorithm that is intended to *reduce its generalization error* ..."
  (일반화 오류를 줄이려는 의도를 가지고 학습 알고리즘을 수정하는 방법 모두)

◆ 규제는 오래 전부터 수학과 통계학에서 연구해온 주제

- **모델 용량에 비해 데이터가 부족한 경우**의 불량 문제를ill-posed problem 푸는 데 사용
- 현대 기계학습도 규제를 널리 사용

◆ 명시적 규제와 암시적 규제

- **명시적 규제**: 가중치 감쇠나 드롭아웃처럼 목적함수나 신경망 구조를 직접 수정하는 방식
- **암시적 규제**: 조기 멈춤, 데이터 증대, 잡음 추가, 앙상블처럼 간접적으로 영향을 미치는 방식

Regularization by Weight Penalty (가중치 감쇠)

◆ Regularized Cost Function

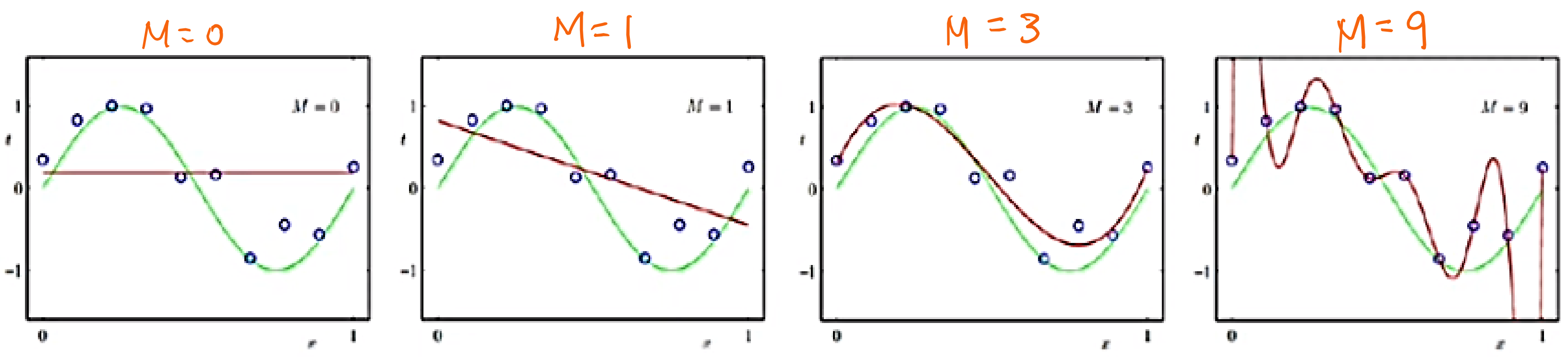
$$\underbrace{J_{regularized}(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})}_{\text{목적함수}} + \lambda \underbrace{R(\boldsymbol{\Theta})}_{\text{규제 항}}$$

- **규제항**은 훈련집합과 무관하며, 데이터 생성 과정에 내재한 **사전 지식**에 해당

Regularization by Weight Penalty (가중치 감쇠)

$$\underbrace{J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{목적함수}} + \lambda\,\underbrace{R(\boldsymbol{\Theta})}_{\text{규제 항}}$$

- ◆ 규제항 $R(\Theta)$로 무엇을 사용할 것인가? → 가중치 감쇠 (가중치 벌칙)
  - 큰 가중치(Θ)에 벌칙을 가해 작은 가중치를 유지 → 최종해를 원점 가까이 당기는 효과
  - L2 norm 사용: $R(\Theta) = \|\Theta\|_2^2$
  - L1 norm 사용: $R(\Theta) = \|\Theta\|_1$
  - 가중치 감쇠는 모델의 구조적 용량을 충분히 크게 하고 모델의 수치적 용량을 제한 하는 규제 기법



	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Regularization – L2 Norm,

◆ Regularized Cost & Gradient

$$\underbrace{J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{목적함수}} + \lambda\,\underbrace{\|\boldsymbol{\Theta}\|_2^2}_{\text{규제 항}}$$

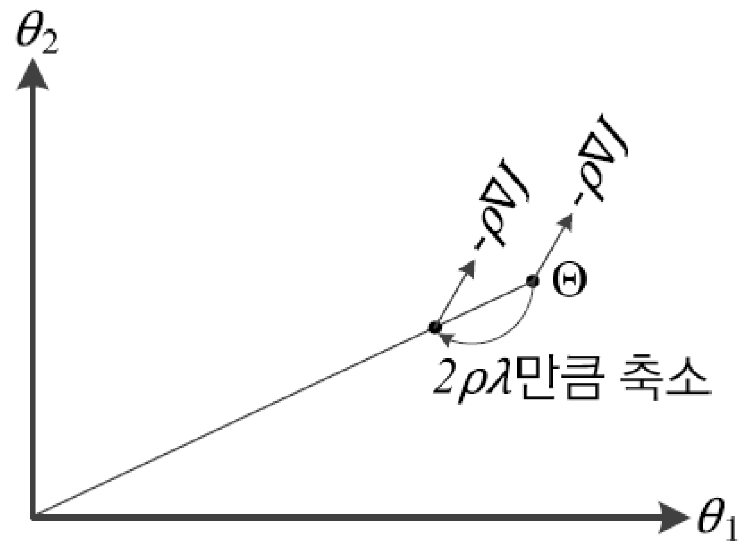
$$\nabla J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) = \nabla J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) + 2\lambda\boldsymbol{\Theta}$$

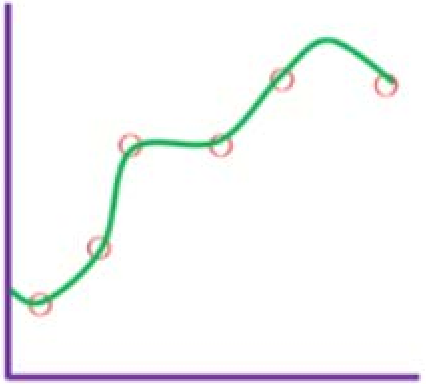
◆ Parameter Update

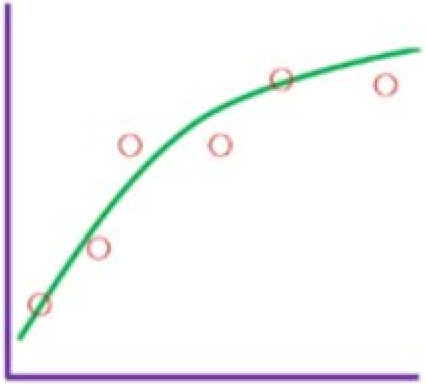
$$\begin{aligned}\boldsymbol{\Theta} &= \boldsymbol{\Theta} - \rho\nabla J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) \\ &= \boldsymbol{\Theta} - \rho(\nabla J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) + 2\lambda\boldsymbol{\Theta}) \\ &= (1-2\rho\lambda)\boldsymbol{\Theta} - \rho\nabla J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})\end{aligned}$$

- L2 규제는 $\boldsymbol{\Theta}$를 $2\rho\lambda$의 비율로 줄인 후 업데이트 하는 셈
  - 즉, 가중치 감소 정도가 현재 가중차 크기에 비례함 함

Weight decay (가중치 감쇠)







$$\beta_0 + \beta_1 x + \beta_2 x^2 + \cancel{\beta_3} x^3 + \cancel{\beta_4} x^4$$

$$\beta_0 + \beta_1 x + \beta_2 x^2$$

$$\underbrace{J_{regularized}(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})}_{\text{목적함수}} + \lambda \underbrace{\|\boldsymbol{\Theta}\|_2^2}_{\text{규제 항}}$$

$$\min_{\beta} \sum_{i=1} (y_i - \hat{y}_i)^2 + 5000{\beta_3}^2 + 5000{\beta_4}^2 \Rightarrow \beta_3 \approx 0 \qquad \beta_4 \approx 0$$

Regularization – L1 Norm

◆ Regularized Cost & Gradient

$$\underbrace{J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{목적함수}} + \lambda \underbrace{\|\boldsymbol{\Theta}\|_1}_{\text{규제 항}}$$

$$\nabla J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) = \nabla J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) + \lambda\mathbf{sign}(\boldsymbol{\Theta})$$

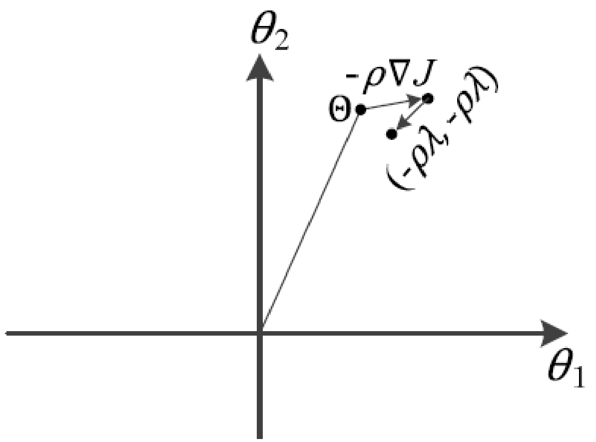
$\mathbf{sign}(\boldsymbol{\Theta})$: $\boldsymbol{\Theta}$의 부호 벡터 (1, -1)

◆ Parameter Update

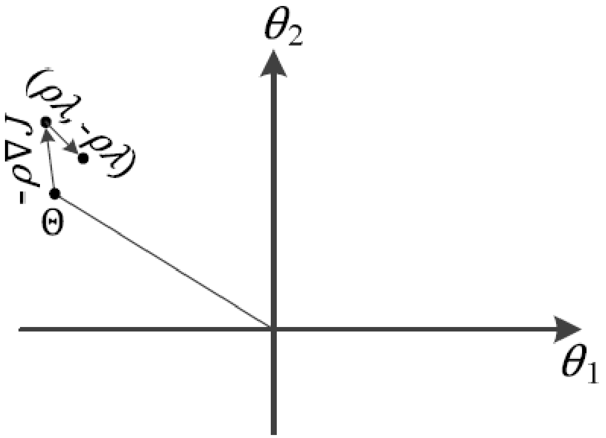
$$\begin{aligned}\boldsymbol{\Theta} &= \boldsymbol{\Theta} - \rho\nabla J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) \\ &= \boldsymbol{\Theta} - \rho(\nabla J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) + \lambda\mathbf{sign}(\boldsymbol{\Theta})) \\ &= \boldsymbol{\Theta} - \rho\nabla J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y}) - \rho\lambda\mathbf{sign}(\boldsymbol{\Theta})\end{aligned}$$

원점 방향으로

- L1 규제는 $\boldsymbol{\Theta}$를 $\rho\lambda$(고정값)만큼 줄인 후 업데이트 하는 셈
- L1 규제의 희소성(Sparse) 효과: 0이 되는 가중치가 많이 발생
  - 선형 회귀에 적용하면 특징 선택 효과



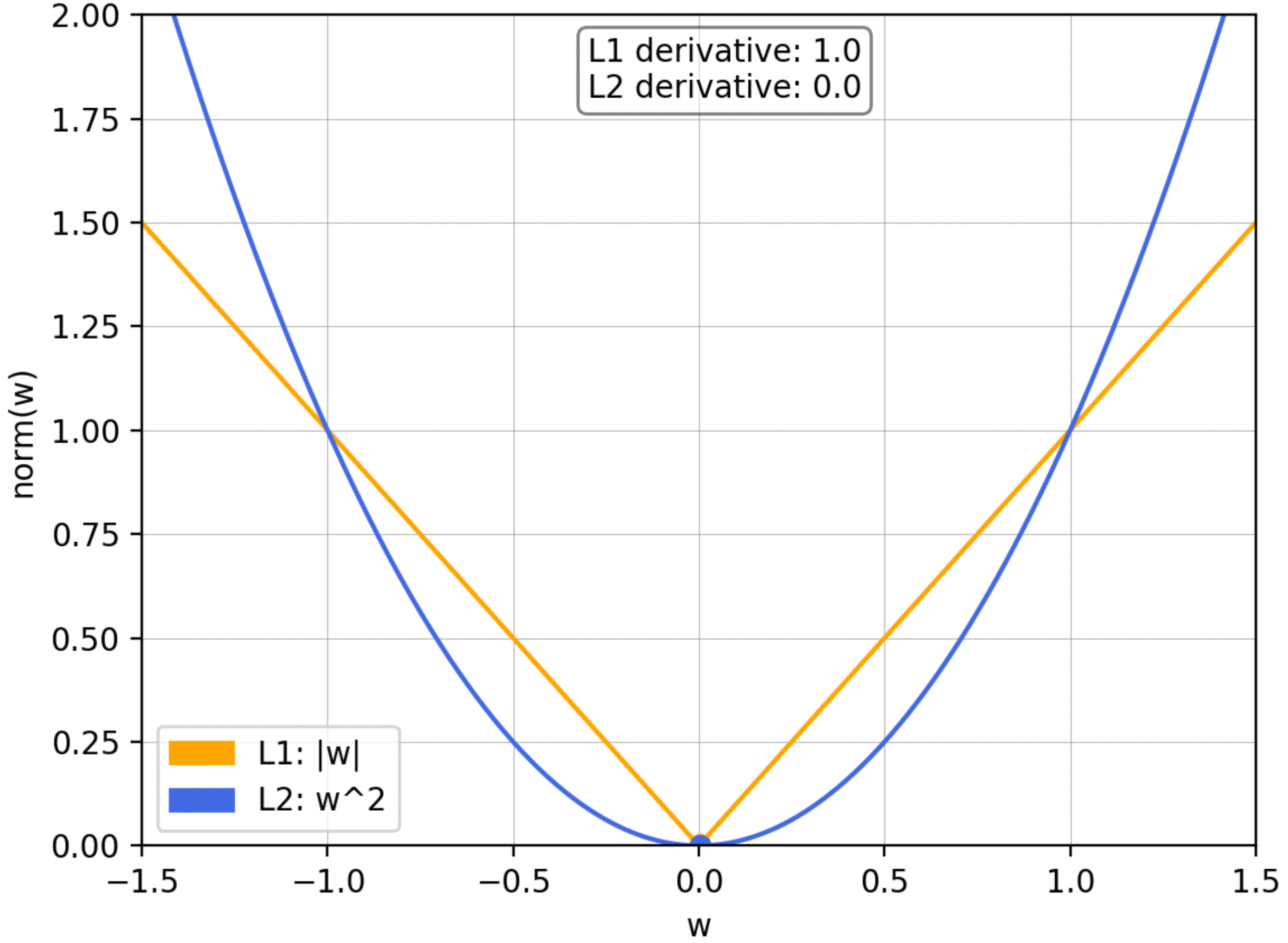
(a) $\mathbf{sign}(\boldsymbol{\Theta}) = (1,1)^{\mathrm{T}}$인 경우



(b) $\mathbf{sign}(\boldsymbol{\Theta}) = (-1,1)^{\mathrm{T}}$인 경우

Regularization – L1 norm vs. L2 norm

L1 norm이 0이 되는 가중치가 많이 발생하는 이유 :
1. 업데이트 속도차이 L1은 고정적인 비율로 갱신 L2는 가중치 크기에 비례
2.



◆ **L1 norm** updae

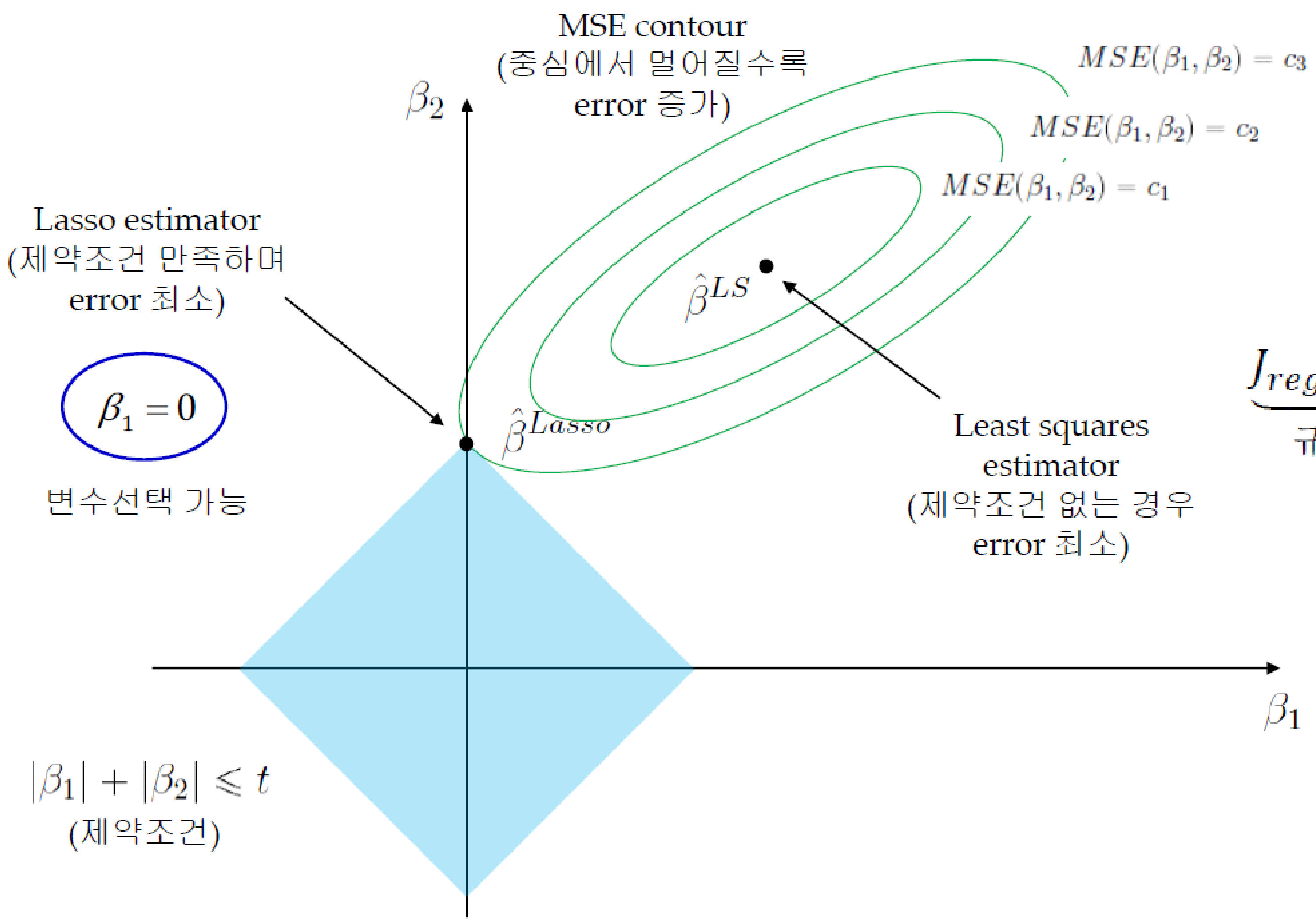
$$\boldsymbol{\Theta} = \boldsymbol{\Theta} - \rho \nabla J_{regularized}(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})$$
$$= \boldsymbol{\Theta} - \rho(\nabla J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y}) + \lambda \mathbf{sign}(\boldsymbol{\Theta}))$$
$$= \boldsymbol{\Theta} - \rho \nabla J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y}) - \rho\lambda \mathbf{sign}(\boldsymbol{\Theta})$$

◆ **L2 norm** update

$$\boldsymbol{\Theta} = \boldsymbol{\Theta} - \rho \nabla J_{regularized}(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})$$
$$= \boldsymbol{\Theta} - \rho(\nabla J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y}) + 2\lambda \boldsymbol{\Theta})$$
$$= (1 - 2\rho\lambda)\boldsymbol{\Theta} - \rho \nabla J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})$$

https://miro.medium.com/v2/resize:fit:720/format:webp/1*bM2txQ6caL4AKiN19oH5bQ.gif

Lasso Regression의 기하학적 이해



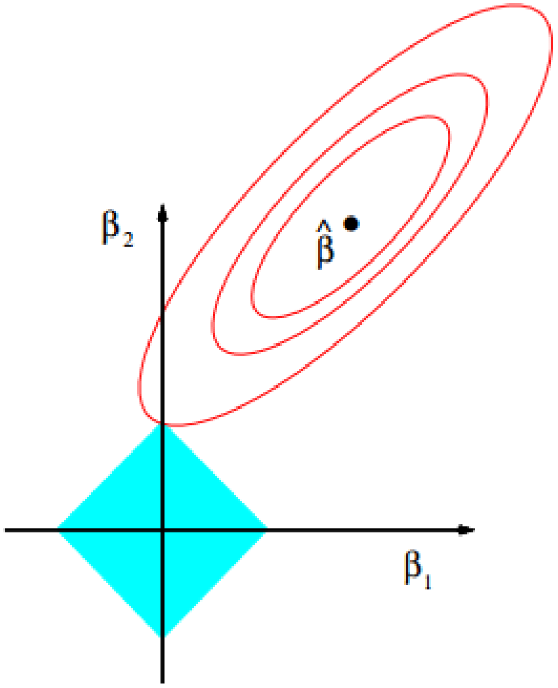
$$\underbrace{J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{목적함수}} + \lambda \underbrace{\|\boldsymbol{\Theta}\|_1}_{\text{규제 항}}$$

https://ratsgo.github.io/machine%20learning/2017/05/22/RLR/

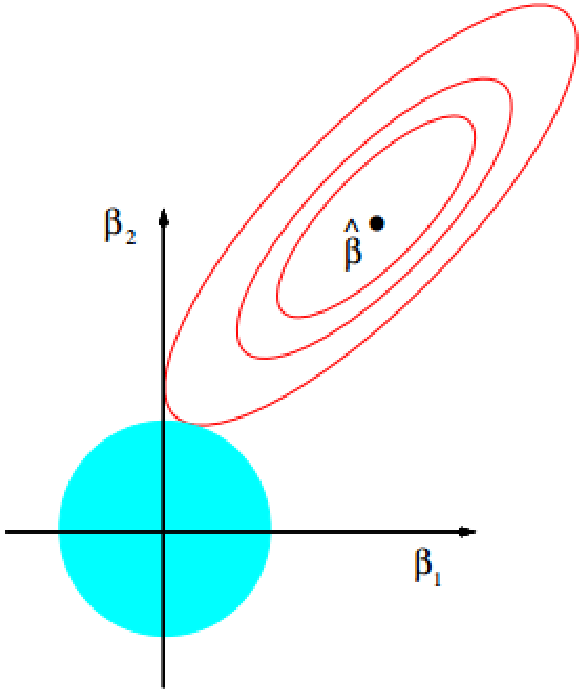
Regularization – L1 norm vs. L2 norm

◆ Lasso (L1) vs. Ridge (L2) Regression

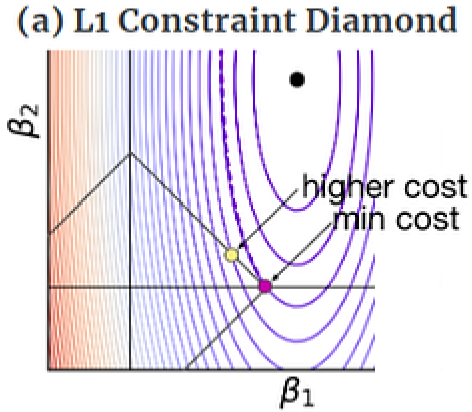
$$\underbrace{J_{regularized}(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})}_{\text{목적함수}} + \lambda \underbrace{R(\boldsymbol{\Theta})}_{\text{규제 항}}$$



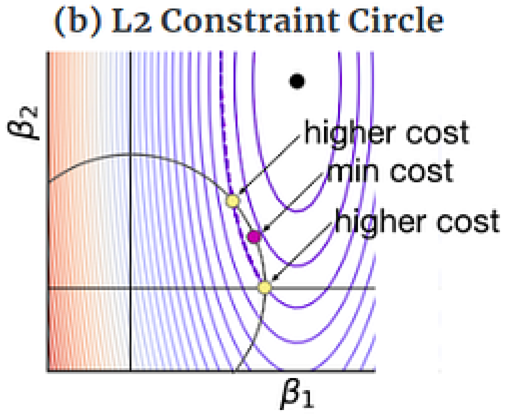
$|\beta_1| + |\beta_2| < t$



$\beta_1^2 + \beta_2^2 < t$



0이 되는 가중치가 많이 발생

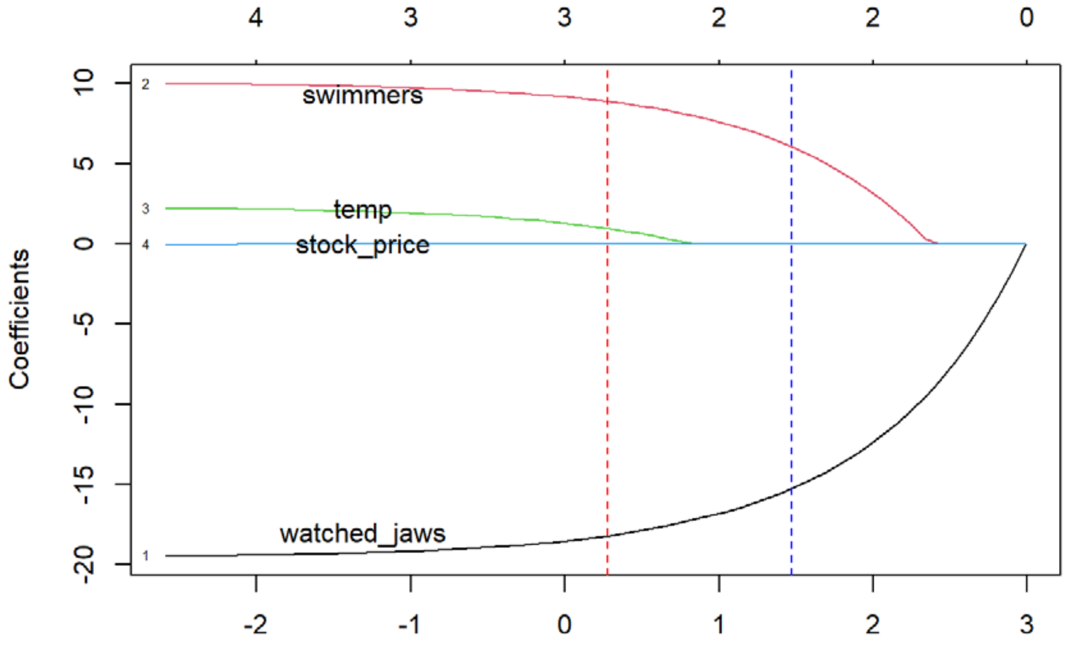


https://medium.com/@mukulranjan/how-does-lasso-regression-l1-encourage-zero-coefficients-but-not-the-l2-20e4893cba5d

Regularization – L1 norm vs. L2 norm

◆ Lasso (L1) vs. Ridge (L2) Regression

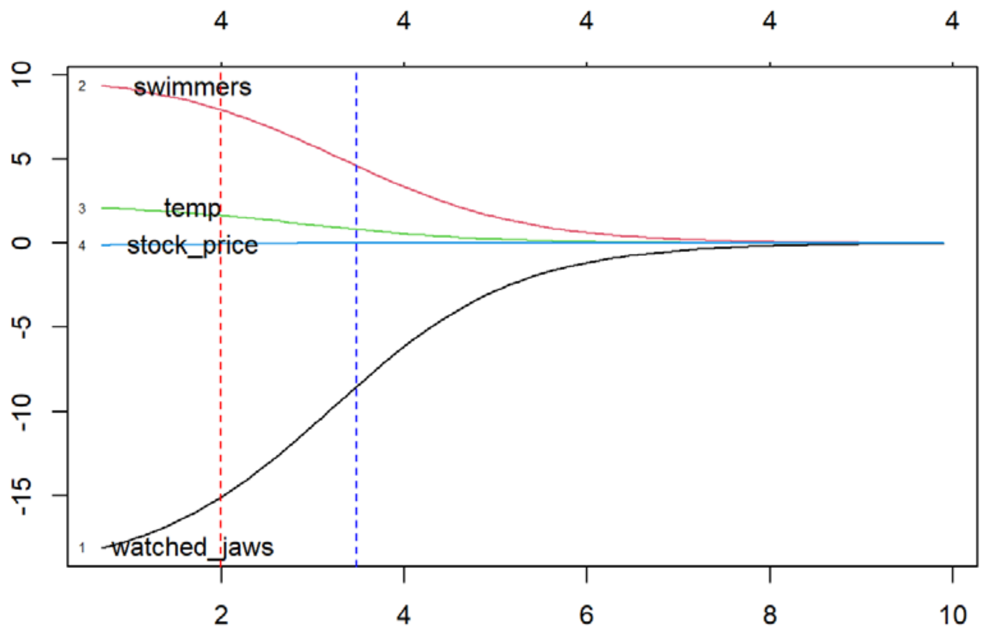
$$\underbrace{J_{regularized}(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{규제를 적용한 목적함수}} = \underbrace{J(\boldsymbol{\Theta};\mathbb{X},\mathbb{Y})}_{\text{목적함수}} + \lambda \underbrace{R(\boldsymbol{\Theta})}_{\text{규제 항}}$$



0이 되는 가중치가 많이 발생

학습속도가 빠름 정확도는 낮은편

Lasso (L1 norm)

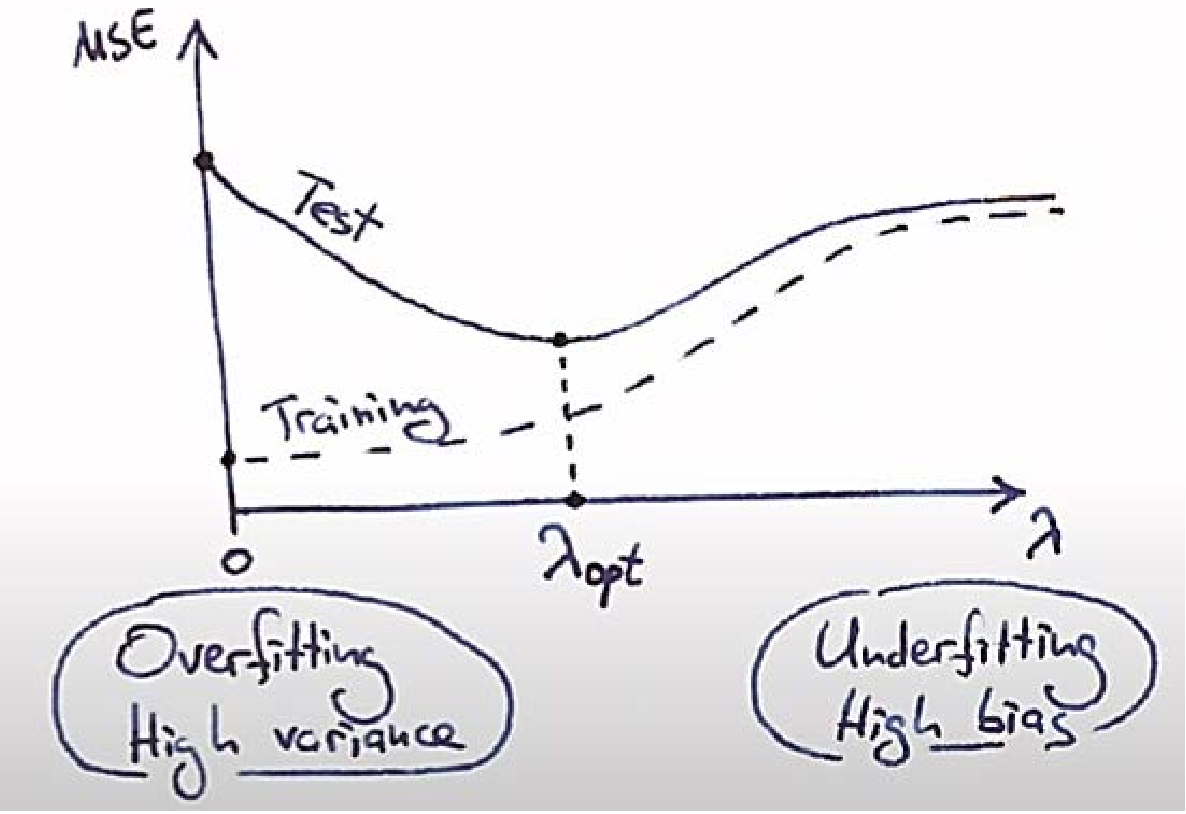


Ridge (L2 norm)

https://www.atakanekiz.com/en/technical/understanding-lasso-and-ridge-regression/

Regularization – Selecting Lambda

- Test Error가 가장 작게 되는 λ가 최적
  - 그러나 학습시에는 test set에 접근할 수 없으므로, validation set을 이용하여 최적의 λ를 선택함



https://deep-learning-study.tistory.com/657

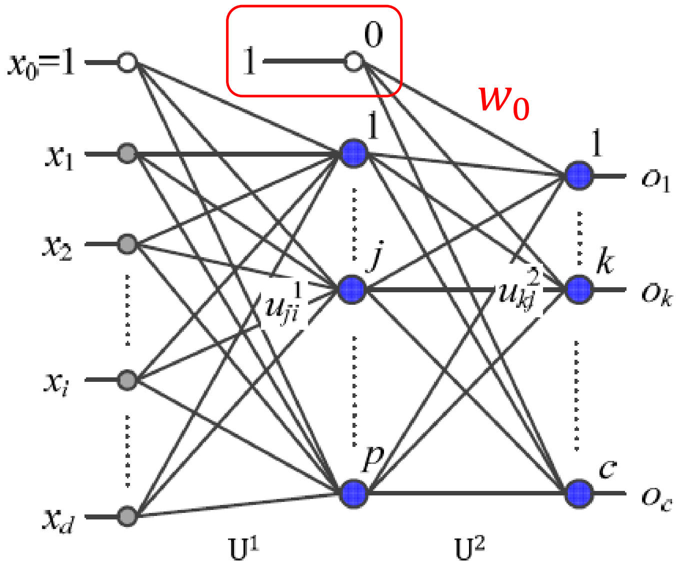
Regularization – Do Not Penalize Bias!

- For Centered Dataset (when both x and y have zero mean)
  - No problem even if we have zero bias (i.e., $w_0 = 0$).

$$J(\boldsymbol{\Theta}) = \frac{1}{n}\left\|\mathbf{y} - \mathbf{x}^{\mathrm{T}}\mathbf{w}\right\|_2^2 + \lambda\|\mathbf{w}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w_0 - \sum_{j=1}^{d} w_j x_{ij}\right)^2 + \lambda\sum_{j=\mathbf{0}}^{d} w_j^2$$

- For Non-centered Dataset (the general case)
  - Penalizing bias often leads to bad performance.
  - Thus we need to ***exclude the bias (w_0) from the regularization term***:

$$J(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - w_0 - \sum_{j=1}^{d} w_j x_{ij}\right)^2 + \lambda\sum_{j=\mathbf{1}}^{d} w_j$$



Regularization – Example: Linear Regression

■ 선형 회귀에 적용

- 선형 회귀는 훈련집합 $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, $\mathbb{Y} = \{y_1, y_2, \cdots, y_n\}$이 주어지면, 식 (5.24)를 풀어 $\mathbf{w} = (w_1, w_2, \cdots, w_d)^\mathrm{T}$를 구하는 문제. 이때 $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^\mathrm{T}$

$$w_1 x_{i1} + w_2 x_{i2} \cdots + w_d x_{id} = \mathbf{x}_i^\mathrm{T}\mathbf{w} = y_i, \qquad i = 1,2,\cdots,n \tag{5.24}$$

- 식 (5.24)를 행렬식으로 바꿔 쓰면,

$$\mathbf{Xw} = \mathbf{y} \tag{5.25}$$

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\mathrm{T} \\ \mathbf{x}_2^\mathrm{T} \\ \vdots \\ \mathbf{x}_n^\mathrm{T} \end{bmatrix}$$

- 가중치 감쇠를 적용한 목적함수

$$J_{regularized}(\mathbf{w}) = \frac{1}{2}\|\mathbf{Xw} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 = \frac{1}{2}(\mathbf{Xw} - \mathbf{y})^\mathrm{T}(\mathbf{Xw} - \mathbf{y}) + \lambda\|\mathbf{w}\|_2^2 \tag{5.27}$$

Regularization – Example: Linear Regression (cont'd)

- 식 (5.27)을 미분하여 0으로 놓으면,

$$\frac{\partial J_{regularized}}{\partial \mathbf{w}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\lambda\mathbf{w} = \mathbf{0} \quad \Longrightarrow \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\lambda\mathbf{I})\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (5.28)$$

- 식 (5.28)을 정리하면,

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (5.29)$$

  - 공분산 행렬 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$의 대각 요소가 2λ만큼씩 증가 → 역행렬을 곱하므로 가중치를 축소하여 원점으로 당기는 효과 ([그림 5-21])

- 예측 단계에서는.

$$y = \mathbf{x}^{\mathrm{T}}\hat{\mathbf{w}} \qquad (5.30)$$

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \|\boldsymbol{X}\mathbf{w} - \mathbf{y}\|_2^2$$ 풀이

$$\|\boldsymbol{X}\mathbf{w} - \mathbf{y}\|_2^2 = (\boldsymbol{X}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\boldsymbol{X}\mathbf{w} - \mathbf{y}) \overset{①}{=} \left(\mathbf{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}} - \mathbf{y}^{\mathrm{T}}\right)(\boldsymbol{X}\mathbf{w} - \mathbf{y})$$
$$= \mathbf{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\mathbf{w} - \mathbf{y}^{\mathrm{T}}\boldsymbol{X}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}$$
$$\overset{②}{=} \mathbf{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\mathbf{w} - 2\mathbf{y}^{\mathrm{T}}\boldsymbol{X}\mathbf{w} + \mathbf{y}^{\mathrm{T}}\mathbf{y}$$

$$\frac{\partial}{\partial \mathbf{w}} \|\boldsymbol{X}\mathbf{w} - \mathbf{y}\|_2^2 = \frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\mathbf{w}) - \frac{\partial}{\partial \mathbf{w}}(2\mathbf{y}^{\mathrm{T}}\boldsymbol{X}\mathbf{w}) + \frac{\partial}{\partial \mathbf{w}}(\mathbf{y}^{\mathrm{T}}\mathbf{y})$$
$$\overset{③}{=} 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\mathbf{w} - 2\boldsymbol{X}^{\mathrm{T}}\mathbf{y} + 0$$
$$= 2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\mathbf{w} - \mathbf{y})$$

① $(\mathbf{A} - \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} - \mathbf{B}^{\mathrm{T}}$

$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$

② $\mathbf{x}^{\mathrm{T}}\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{x}$

$\mathbf{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\mathbf{y} = (\mathbf{X}\mathbf{w})^{\mathrm{T}}\mathbf{y} = \mathbf{y}^{\mathrm{T}}(\mathbf{X}\mathbf{w})$

③ $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{A}\mathbf{x}) = \left(\mathbf{A} + \mathbf{A}^{\mathrm{T}}\right)\mathbf{x}$

$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$: symmetric

$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{\mathrm{T}}$

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\mathbf{w}) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{\mathrm{T}}\right)\mathbf{w}$$
$$= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)\mathbf{w}$$
$$= 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\mathbf{w}$$

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{y}^{\mathrm{T}}\mathbf{x}) = \mathbf{y}$$

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}) = \frac{\partial}{\partial \mathbf{w}}\left((\mathbf{X}^{\mathrm{T}}\mathbf{y})^{\mathrm{T}}\mathbf{w}\right) = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$\mathbf{y}^{\mathrm{T}}\mathbf{X}$: a row vector

Regularization – Example: Linear Regression (cont'd)

예제 5-1 **리지 회귀**

훈련집합 $\mathbb{X} = \{\mathbf{x}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}2\\3\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}3\\3\end{pmatrix}\}$, $\mathbb{Y} = \{y_1 = 3.0,\ y_2 = 7.0, y_3 = 8.8\}$이 주어졌다고 가정하자. 특징 벡터가 2차원이므로 d=2이고 샘플이 3개이므로 n=3이다. 훈련집합으로 설계행렬 **X**와 레이블 행렬 **y**를 다음과 같이 쓸 수 있다.

$$\mathbf{X} = \begin{pmatrix}1 & 1\\2 & 3\\3 & 3\end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix}3.0\\7.0\\8.8\end{pmatrix}$$

이 값들을 식 (5.29)에 대입하여 다음과 같이 $\hat{\mathbf{w}}$을 구할 수 있다. 이때 $\lambda = 0.25$라 가정하자.

$$\hat{\mathbf{w}} = \left(\begin{pmatrix}1 & 2 & 3\\1 & 3 & 3\end{pmatrix}\begin{pmatrix}1 & 1\\2 & 3\\3 & 3\end{pmatrix} + \begin{pmatrix}0.5 & 0\\0 & 0.5\end{pmatrix}\right)^{-1}\begin{pmatrix}1 & 2 & 3\\1 & 3 & 3\end{pmatrix}\begin{pmatrix}3.0\\7.0\\8.8\end{pmatrix} = \begin{pmatrix}1.4916\\1.3607\end{pmatrix}$$

따라서 하이퍼 평면은 $y = 1.4916x_1 + 1.3607x_2$이다. 새로운 샘플로 $\mathbf{x} = (5 \quad 4)^{\mathrm{T}}$가 입력되면 식 (5.30)을 이용하여 12.9009를 예측한다.

감사합니다.