
# Machine Learning

## AIML CLZG565

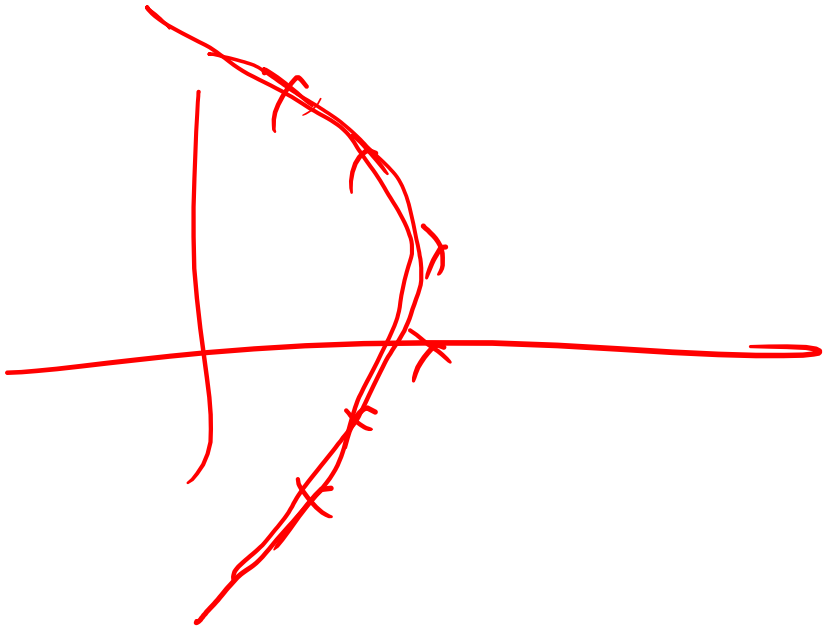
## M3 : Linear Models for Regression

Dr. Sugata Ghosal
sugata.ghosal@pilani.bits-pilani.ac.in

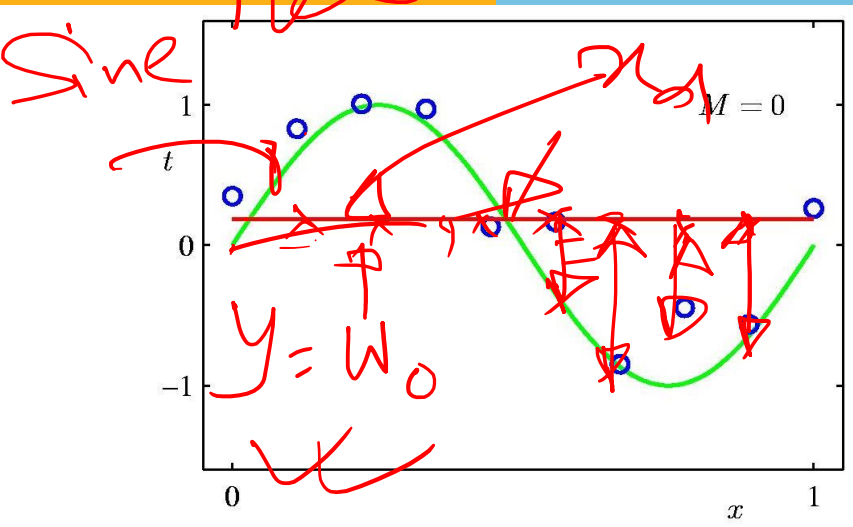
BITS Pilani
Pilani Campus

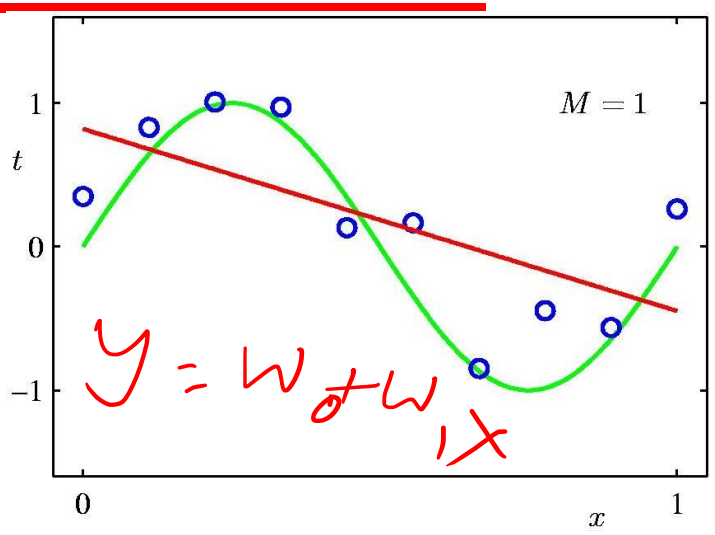
# Linear Regression

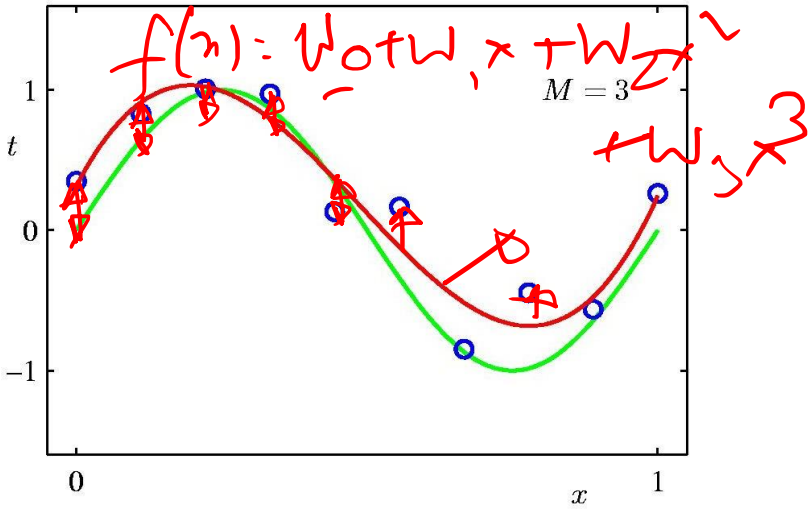
# Linear Basis Models

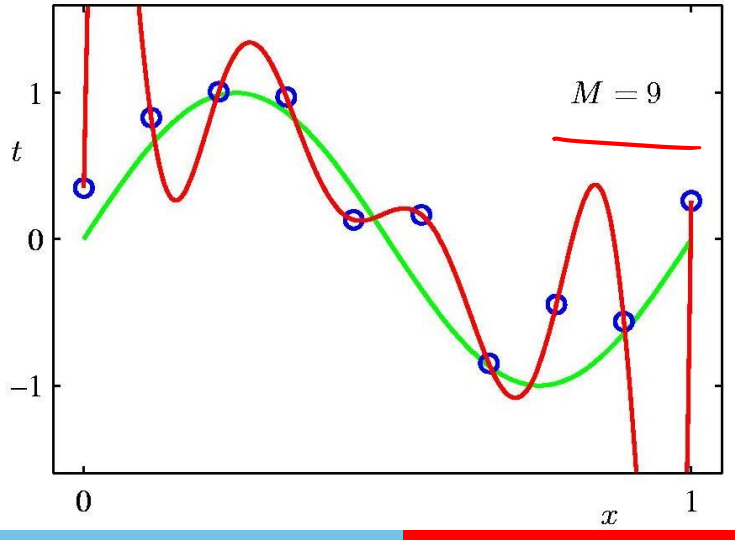


# Polynomial Regression



# Linear Basis Function Models

- The inputs **X** for linear regression can be:
  - Original quantitative inputs
  - Transformation of quantitative inputs
    - e.g. log, exp, square root, square, etc.
  - Polynomial transformation
    - example: $y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3$
  - Basis expansions
  - Dummy coding of categorical inputs
  - Interactions between variables
    - example: $x_3 = x_1 \cdot x_2$

This allows use of linear regression techniques to fit non-linear datasets.

| X No.of.Years of Experience (in Years) | $X^2$ | Y Salary Of the Employee (in Lakhs) |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 4 | 3 |
| 3 | 9 | 4 |
| 4 | 16 | 5 |
| 5 | 25 | 6 |

| X1 = Graduate | X2 = PostGraduate | X3 = Others | Y Salary Of the Employee |
|---|---|---|---|
| 0 | 0 | 1 | 2 |
| 1 | 0 | 0 | 3 |
| 0 | 0 | 1 | 4 |
| 0 | 1 | 0 | 5 |
| 1 | 0 | 0 | 6 |

# Linear Basis Function Models

| X No.of.Years of Experience (in Years) | $X^2$ | Y Salary Of the Employee (in Lakhs) |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 4 | 3 |
| 3 | 9 | 4 |
| 4 | 16 | 5 |
| 5 | 25 | 6 |

Example: an M-th order polynomial function of one dimensional feature x :

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j x^j$$

where $x^j$ = j-th power of x

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

$\Phi_j(x)$ are known as *basis functions*. Typically, $\Phi_0(x) = 1$, so that $w_0$ acts as a bias.

In the simplest case, we use linear basis functions : $\Phi_d(x) = x_d$.

They are called linear models because this function is linear in **w**.

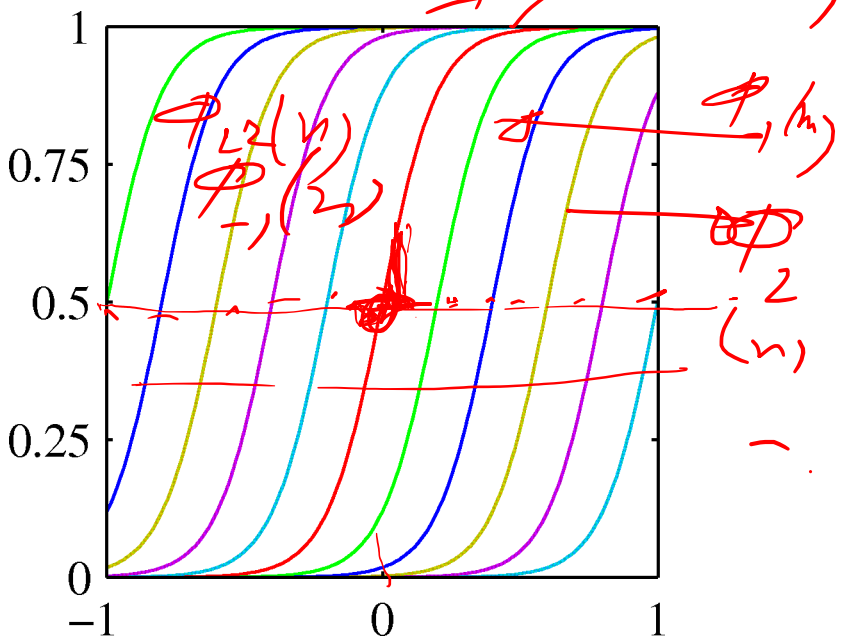
# Linear Basis Function Models - Examples

Sigmoidal basis functions:

where $$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$

Also these are local; a small change in $x$ only affect nearby basis functions. $\mu_j$ and $s$ control location and scale (slope).

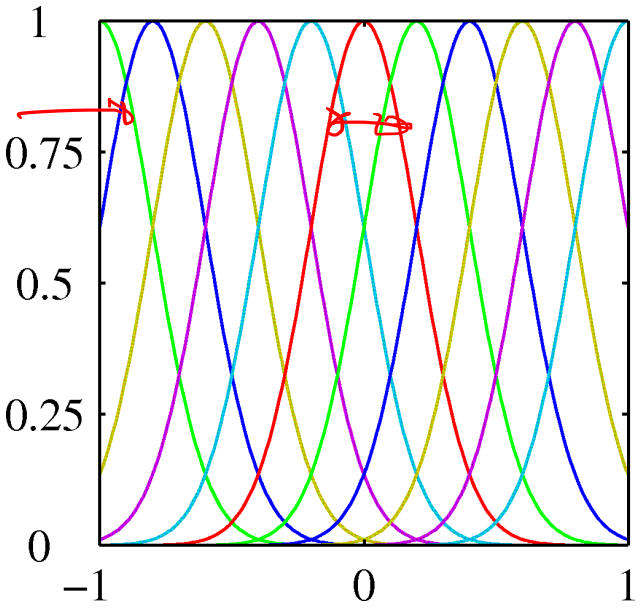


# Linear Basis Function Models - Examples

Gaussian basis functions:

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

These are local; a small change in $x$ only affect nearby basis functions. $\mu_j$ and $s$ control location and scale (width).

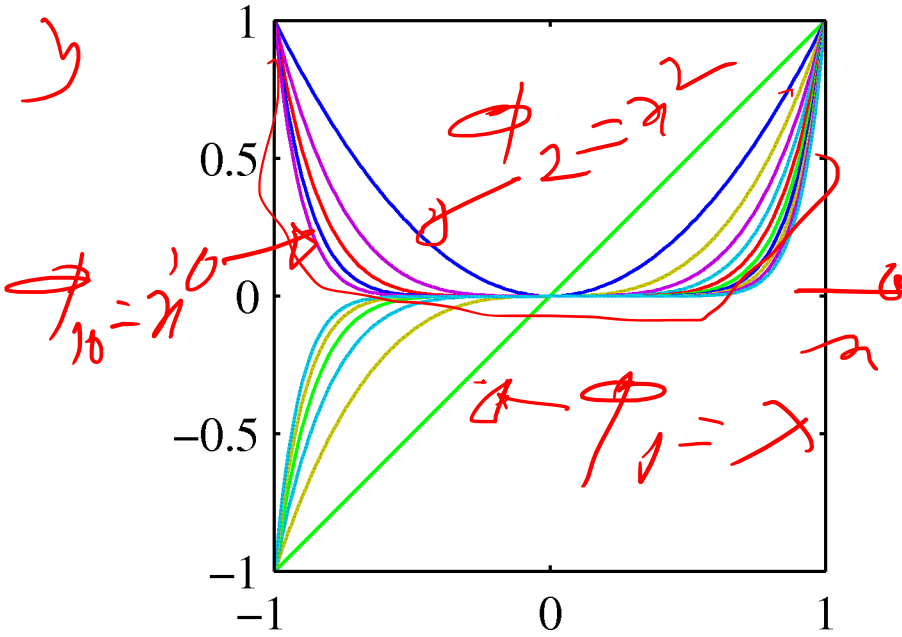


# Linear Basis Function Models - Examples

Polynomial basis functions:

$$\phi_j(x) = x^j.$$

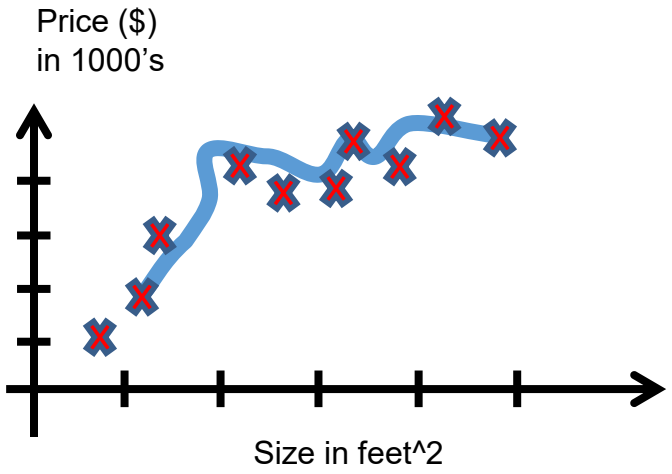
These are global; a small change in $x$ affect all basis functions.



# Notion of Bias - Variance

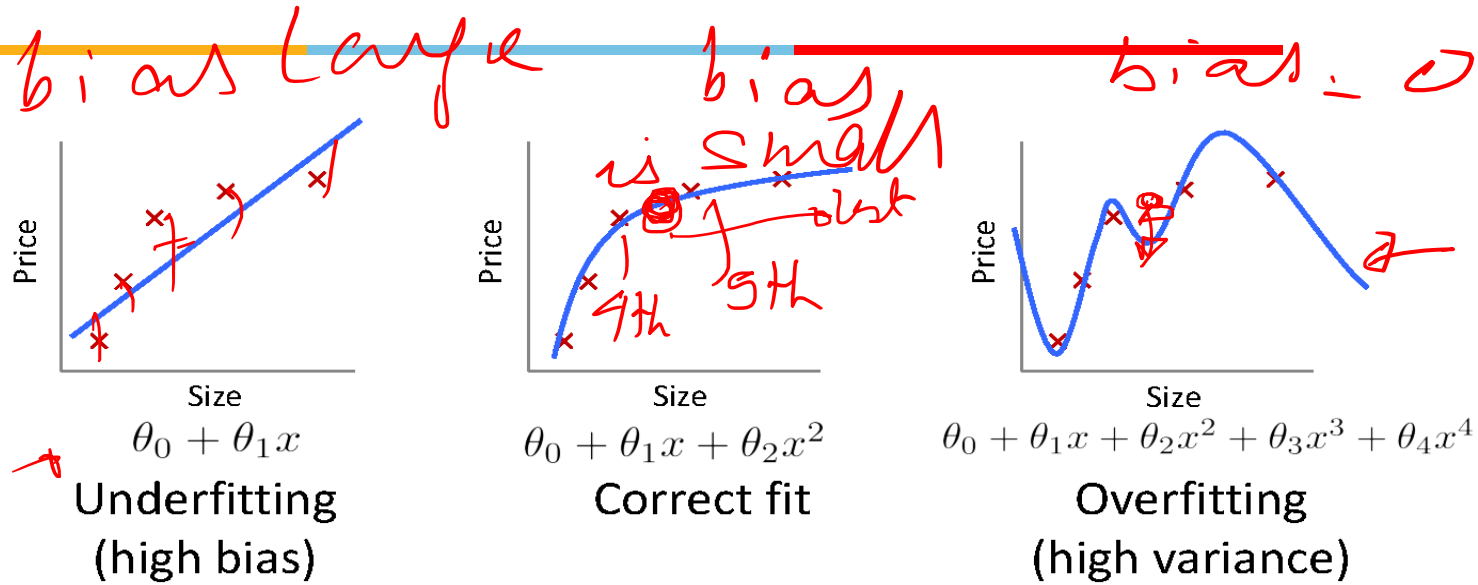
# Addressing overfitting

- $x_1 =$ size of house
- $x_2 =$ no. of bedrooms
- $x_3 =$ no. of floors
- $x_4 =$ age of house
- $x_5 =$ average income in neighborhood
- $x_6 =$ kitchen size
- $\vdots$
- $x_{100}$




Slide credit: Andrew Ng

# Quality of Fit



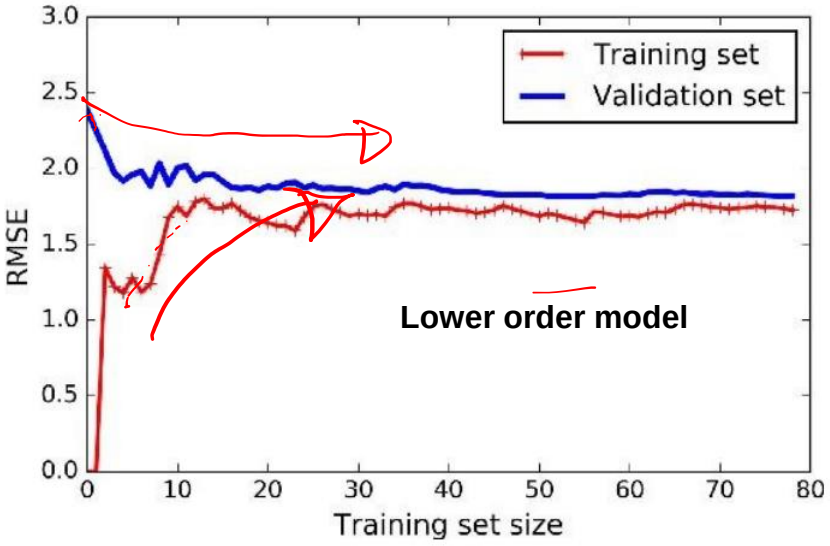

**Overfitting:**

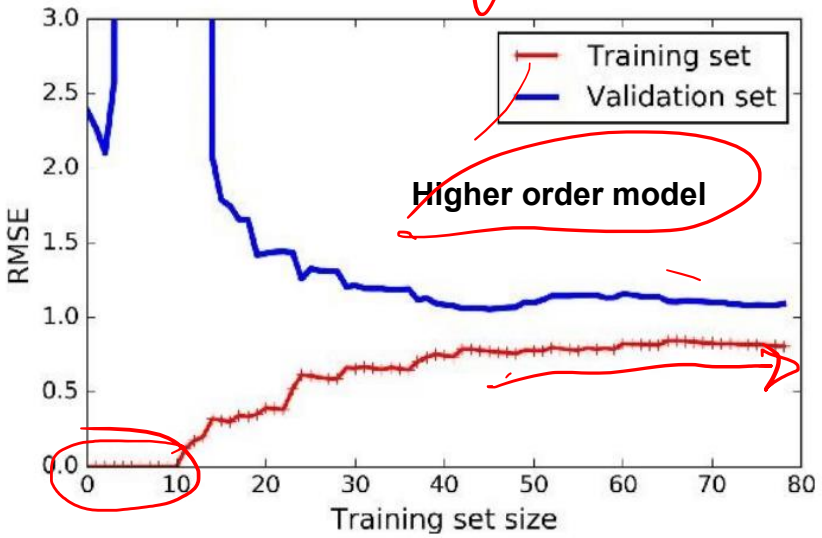
- The learned hypothesis may fit the training set very well ( $J(\boldsymbol{\theta}) \approx 0$ )
- ...but fails to generalize to new examples

Based on example by Andrew Ng

# Effect of Training Size on Over fitting

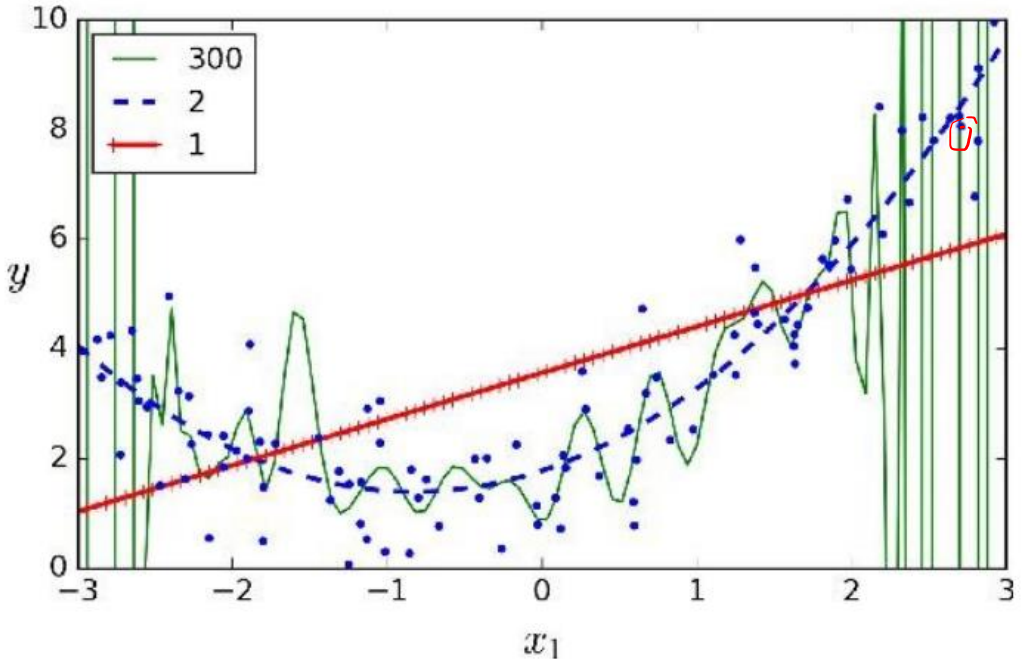
- Size of training dataset needs to be large to prevent when higher order model is used.



# Polynomial Fitting can lead to Over fitting

- Underlying target function is quadratic
- Linear model results in under fitting with large bias
- Polynomial of order 300 results in a large variance



# Regularization

# Regularization

- A method for automatically controlling the complexity of the learned hypothesis
- **Idea**: penalize for large values of $\theta_j$
  - Can incorporate into the cost function
  - Works well when we have a lot of features, each that contributes a bit to predicting the label
- Can also address overfitting by eliminating features (either manually or via model selection)

# Regularization

- Linear regression objective function

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

model fit to data | regularization

  - $\lambda$ is the regularization parameter ( $\lambda \geq 0$)
  - No regularization on $\theta_0$!

Ridge or $L_2$ regularization

# Understanding Regularization

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

- Note that $\sum_{j=1}^{d}\theta_j^2 = \|\boldsymbol{\theta}_{1:d}\|_2^2$
  - This is the magnitude of the feature coefficient vector!

- We can also think of this as:

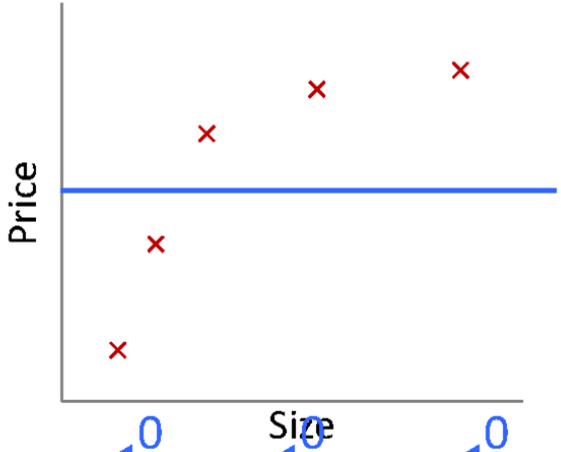
$$\sum_{j=1}^{d}(\theta_j - 0)^2 = \|\boldsymbol{\theta}_{1:d} - \vec{\mathbf{0}}\|_2^2$$

  - $L_2$ regularization pulls coefficients toward 0

# Understanding Regularization

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

- What happens if we set $\lambda$ to be huge (e.g., $10^{10}$)?




$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

0 0 0 0

Based on example by Andrew Ng

# Regularized Linear Regression

- Cost Function

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2$$

- Fit by solving $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

- Gradient update:

$\frac{\partial}{\partial \theta_0} J(\theta)$

$$\theta_0 \leftarrow \theta_0 - \alpha\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)$$

$\frac{\partial}{\partial \theta_j} J(\theta)$

$$\theta_j \leftarrow \theta_j - \alpha\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)x_j^{(i)} - \lambda\theta_j$$

- We can rewrite the gradient step as:

$$\theta_j \leftarrow \theta_j\left(1 - \alpha\lambda\right) - \alpha\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$$

# Lasso Regularization

L1 regularization

Feature selection

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{d}|\theta_j|$$

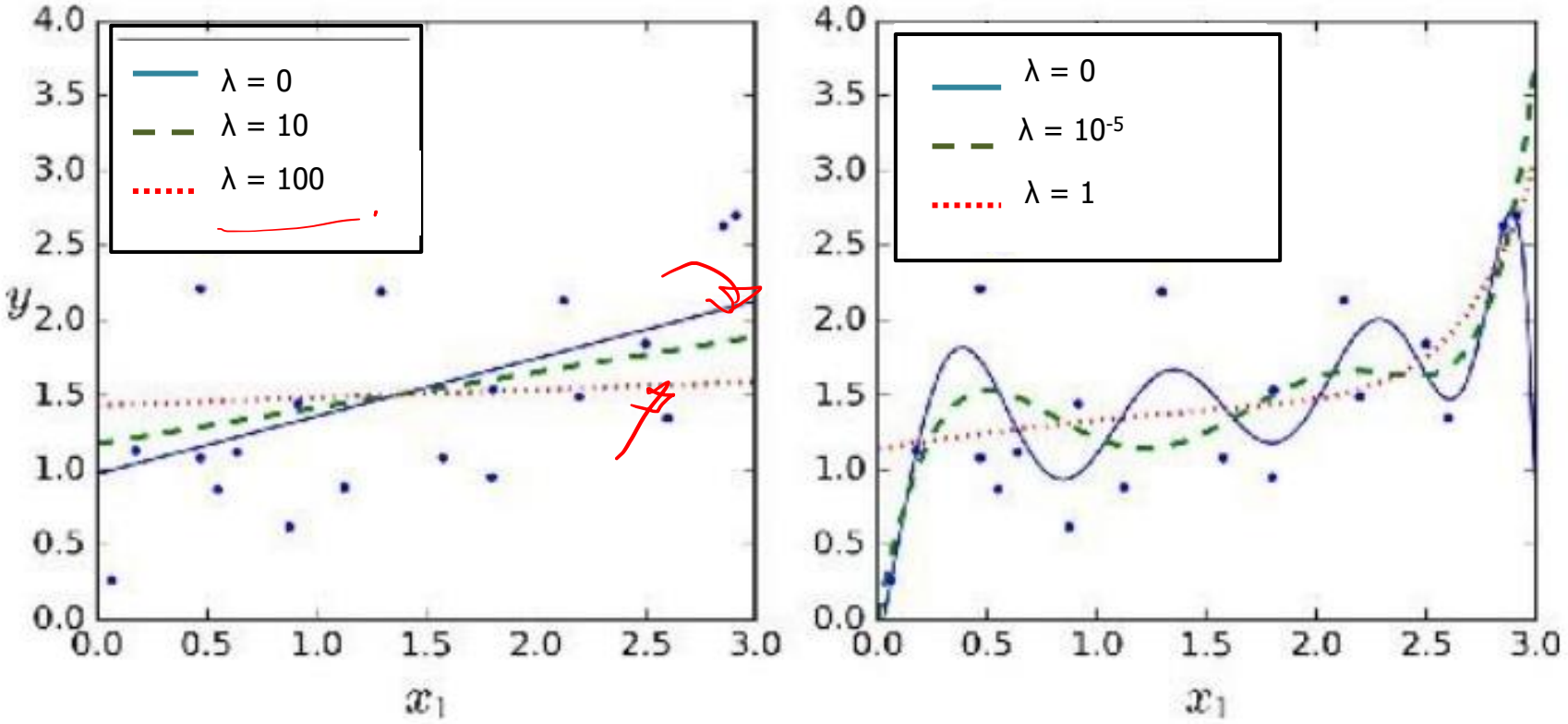
$$\theta_j = \theta_j - \frac{\alpha}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)\boldsymbol{x}_j^{(i)} - \alpha\lambda\,\text{sign}(\theta_j)$$

wt. assoc. with some feature $\to 0$

$-\alpha\lambda$ if $\theta_j > 0$

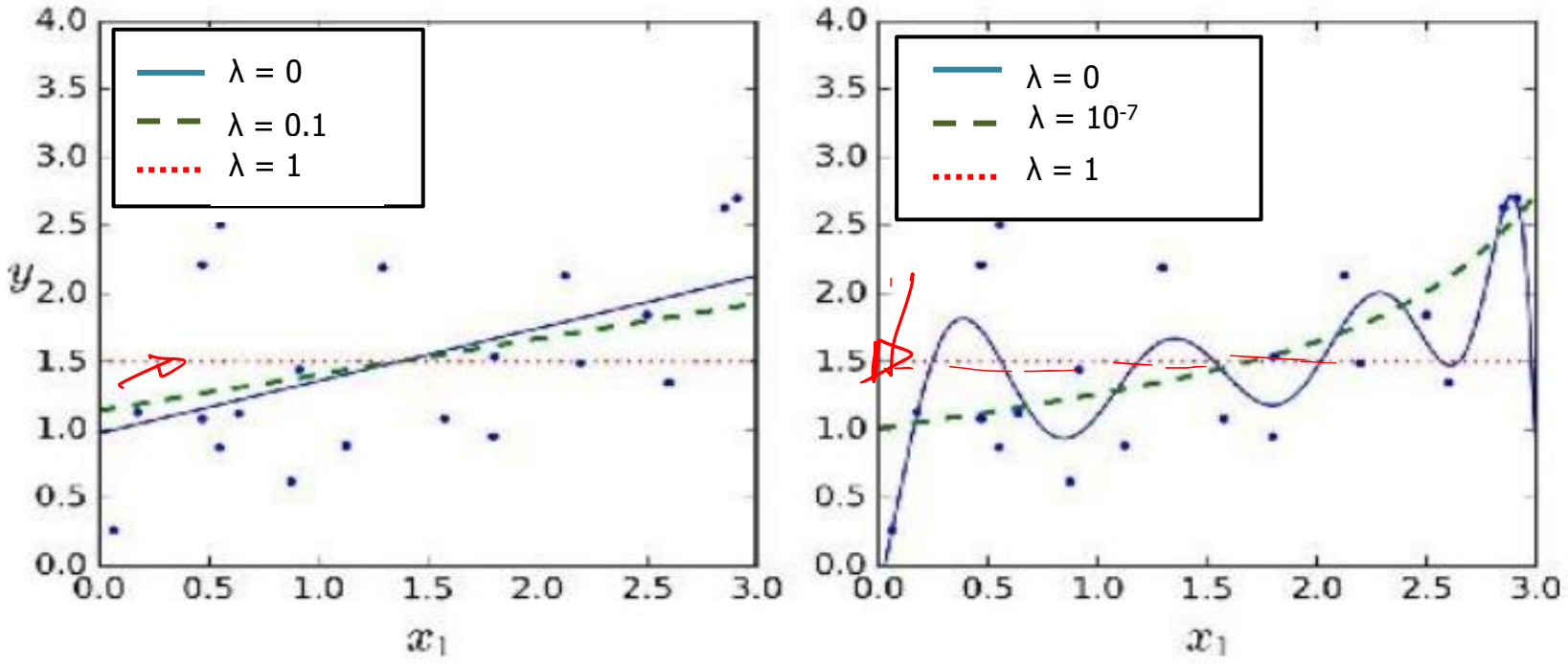
$+\alpha\lambda$ if $\theta_j < 0$

# Ridge Vs Lasso Regularization




Ridge

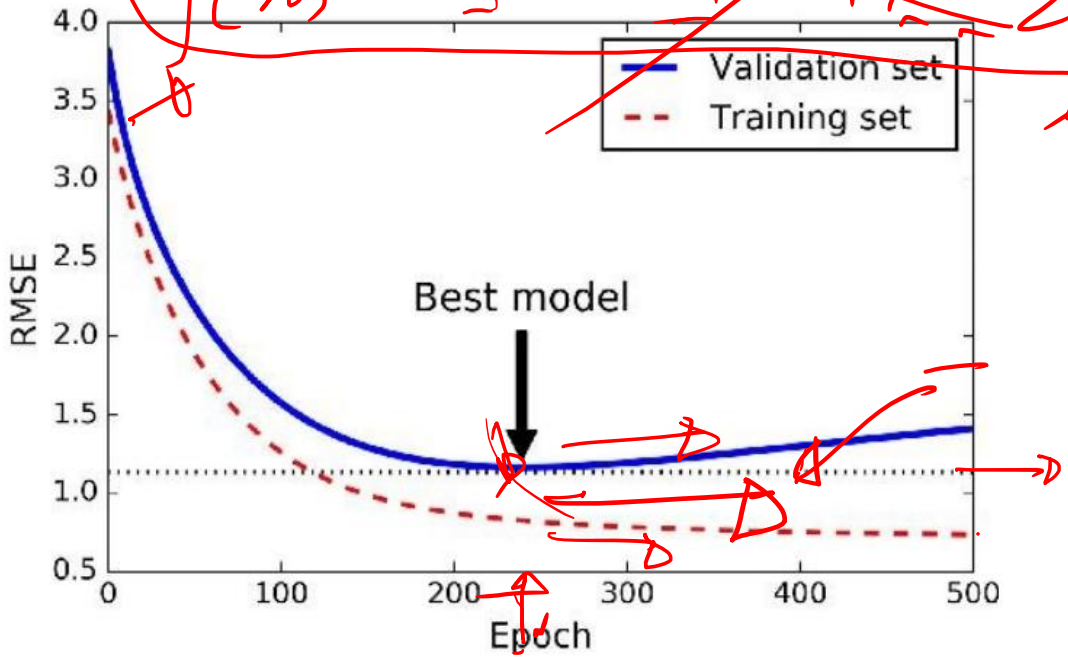
# Ridge Vs Lasso Regularization




Lasso

# Early Stopping

- Do Not Over train to prevent overfitting
- Stop training once error on the validation set starts showing an upward trend, even if the error on the training set keeps decreasing



# How Good is the Fit ?

- Coefficient of Determination

$$R^2 = \frac{SSM}{SST} = \frac{\sum_i [f(x_i) - \overline{y}]^2}{\sum_i [y_i - \overline{y}]^2}$$

- $R^2$ is related to correlation coefficient

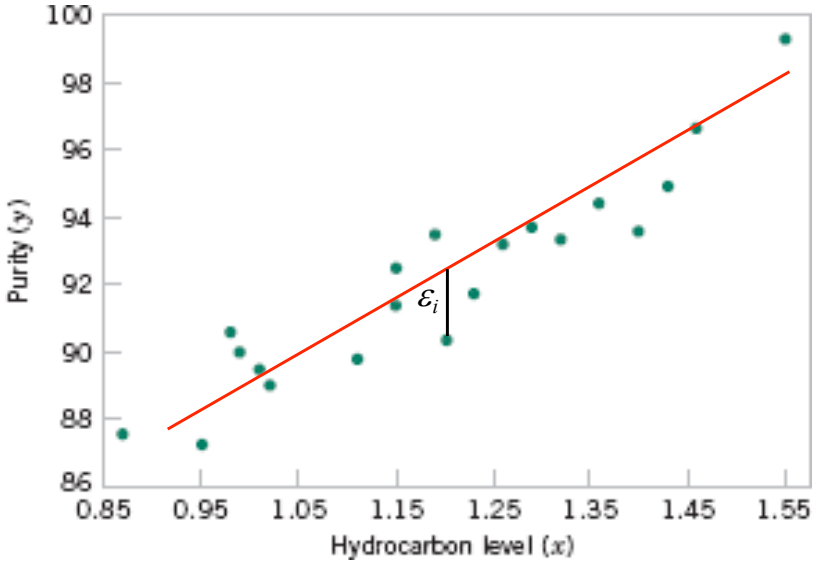
$$r = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}.$$

$$R^2 = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}}.$$

$$\text{Adjusted } R^2 = 1 - \left(\frac{N-1}{N-d}\right)(1-R^2),$$

# Simple linear regression

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to X by the following simple linear regression model:




Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,2,\ldots, n$$

Intercept

Slope

Random error

$$\varepsilon_i \sim N(0, \sigma^2)$$

where the slope and intercept of the line are called regression coefficients.

•The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

# Regression coefficients

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Fitted (estimated) regression model

Caveat: regression relationship are valid only for values of the regressor variable within the range the original data. Be careful with extrapolation.

# Estimation of variance

- Using the fitted model, we can estimate value of the response variable for given predictor

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Residuals: $r_i = y_i - \hat{y}_i$
- Our model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,\ldots,n$, $Var(\varepsilon_i) = \sigma^2$
- Unbiased estimator (MSE: Mean Square Error)

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

# Key Point

- the coefficients

$$\hat{\beta}_1 \text{ and } \hat{\beta}_0$$

are both calculated from data, and they are subject to error.

- if the true model is $y = \beta_1 x + \beta_0$ , $\hat{\beta}_1$ and $\hat{\beta}_0$ are point estimators for the true coefficients
- we can talk about the ``accuracy'' of $\hat{\beta}_1$ and $\hat{\beta}_0$

# Assessing linear regression model

- Test hypothesis about true slope and intercept

$$\beta_1 = ?, \qquad \beta_0 = ?$$

- Construct confidence intervals

$$\beta_1 \in \left[\hat{\beta}_1 - a, \hat{\beta}_1 + a\right] \quad \beta_0 \in \left[\hat{\beta}_0 - b, \hat{\beta}_0 + b\right] \quad \text{with probability } 1-\alpha$$

- Assume the errors are normally distributed

$$\varepsilon_i \sim \mathrm{N}\left(\ 0, \quad \sigma^2\ \right)$$

# Properties of Regression Estimators

| slope parameter $\beta_1$ | intercept parameter $\beta_0$ |
| --- | --- |
| $E(\hat{\beta}_1) = \beta_1$ | $E(\hat{\beta}_0) = \beta_0$ |
| $V(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$ | $V(\hat{\beta}_0) = \sigma^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]$ |

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

# Standard errors of coefficients



- We can replace $\sigma^2$ with its estimator $\hat{\sigma}^2$ ...

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

$$r_i = y_i - \hat{y}_i \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Using results from previous page, estimate the

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

# Hypothesis test in simple linear regression

- we wish to test the hypothesis whether the slope equals a constant $\beta_{1,0}$

$$H_0: \beta_1 = \beta_{1,0}$$
$$H_1: \beta \; != \; \beta_{1,0}$$

- e.g. relate **ads** to **sales**, we are interested in study whether or not increase a $ on ads will increase $ $\beta_{1,0}$ in sales?
- sale = $\beta_{1,0}$ ads + constant?

# A related and important question…

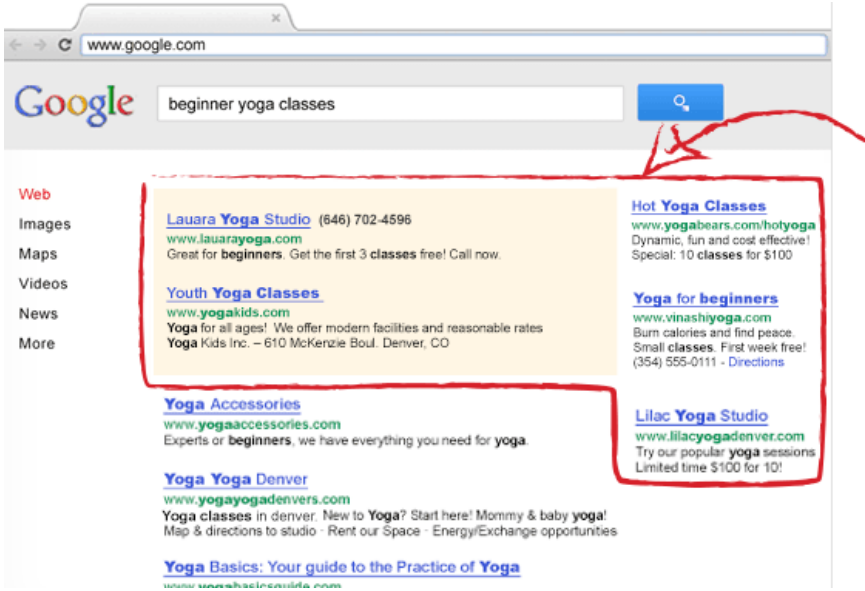
- whether or not the slope is zero ?

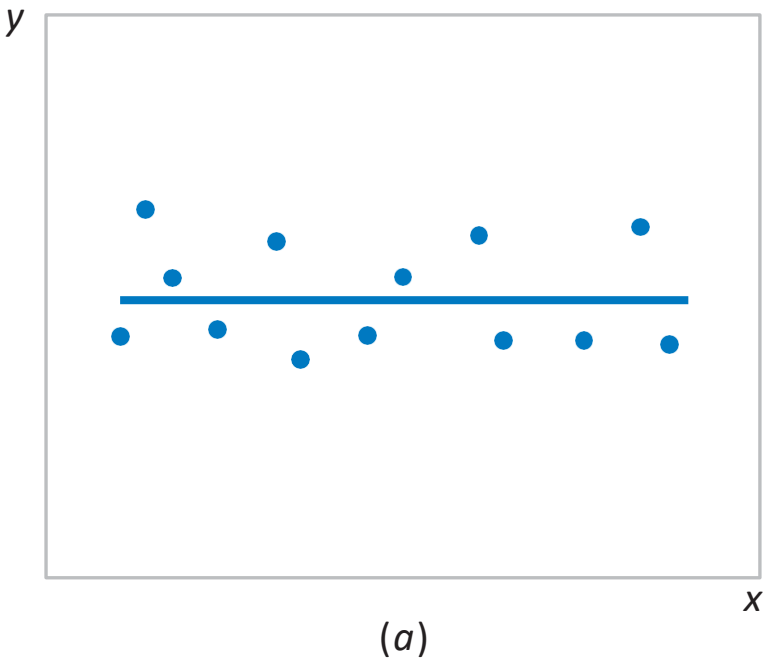
$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 != 0$$

**Significance of regression**

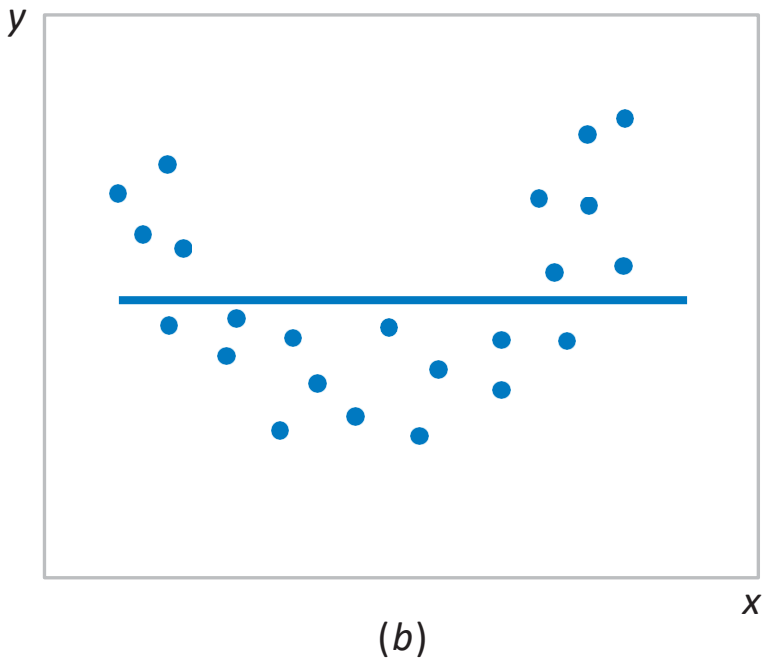
- if $\beta_1 = 0$, that means Y does not depend on X, i.e.,
- Y and X are **independent**
- In the advertisement example, does **ads** increase **sales**? or no effect?






(a)

- $H_0$ not rejected




(b)

- $H_0$ rejected

# Use t-test for slope



Under $H_0$

slope parameter $\beta_1$

$$E(\hat{\beta}_1) = \beta_{1,0}$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\hat{\beta}_1 \sim N\left(\beta_{1,0}, \sigma^2 / S_{xx}\right)$$

- Under $H_0$, test statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

~ t distribution with n-2 degree of freedom

- Reject $H_0$ if

$$|t_0| > t_{\alpha/2, n-2}$$

(two-sided test)

# Use t-test for intercept

- Use a similar form of test

$H_0$: $\beta_0 = \beta_{0,0}$

$H_1$: $\beta_0 \neq \beta_{0,0}$

- Test statistic $$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

Under $H_0$, $T_0$ ~ t distribution with n-2 degree of freedom

- Reject $H_0$ if $|t_0| > t_{\alpha/2,n-2}$

# Confidence interval

- we can obtain confidence interval estimates of slope and intercept
- width of confidence interval is a measure of the overall quality of the regression

true parameter

| slope | intercept |
| --- | --- |
| $T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}}$ | $T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$ |
| ~ t distribution with n-2 degree of freedom | ~ t distribution with n-2 degree of freedom |

# Confidence intervals

**a $100(1-\alpha)\%$ confidence interval on the slope $\beta_1$**

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

**a $100(1-\alpha)\%$ confidence interval on the intercept $\beta_0$**

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

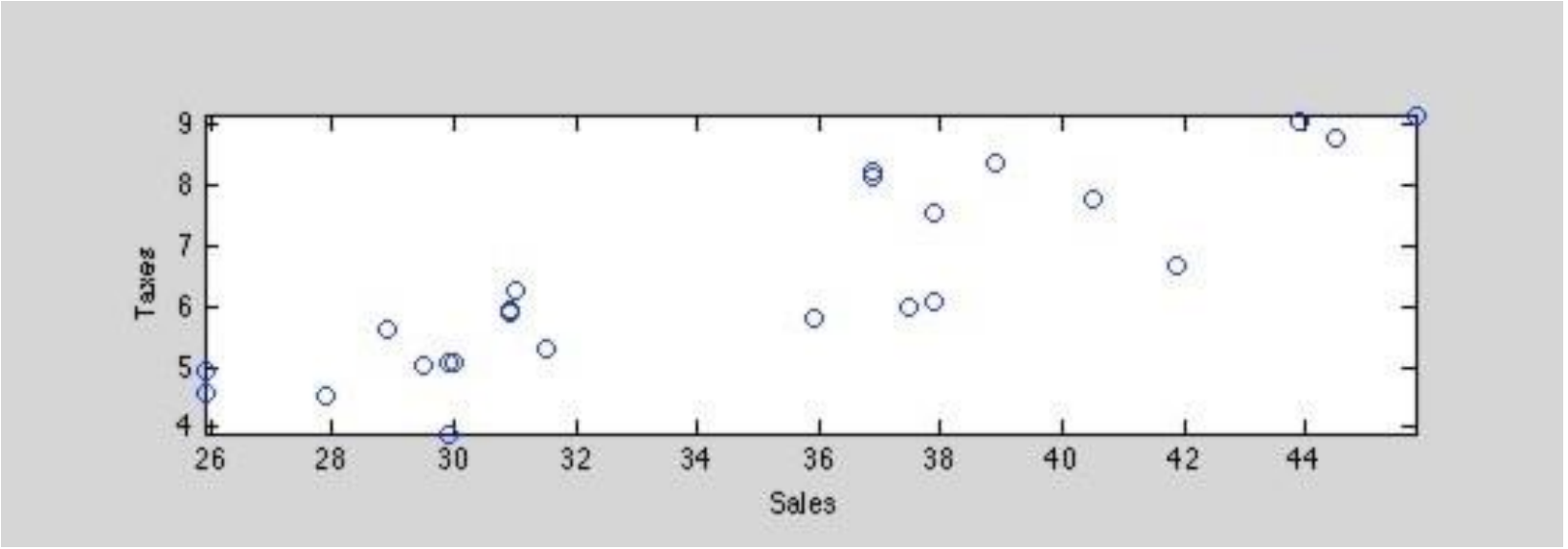
# Example: house selling price and annual taxes

| Sale Price/1000 | Taxes (Local, School), County)/1000 | Sale Price/1000 | Taxes (Local, School), County)/1000 |
|---|---|---|---|
| 25.9 | 4.9176 | 30.0 | 5.0500 |
| 29.5 | 5.0208 | 36.9 | 8.2464 |
| 27.9 | 4.5429 | 41.9 | 6.6969 |
| 25.9 | 4.5573 | 40.5 | 7.7841 |
| 29.9 | 5.0597 | 43.9 | 9.0384 |
| 29.9 | 3.8910 | 37.5 | 5.9894 |
| 30.9 | 5.8980 | 37.9 | 7.5422 |
| 28.9 | 5.6039 | 44.5 | 8.7951 |
| 35.9 | 5.8282 | 37.9 | 6.0831 |
| 31.5 | 5.3003 | 38.9 | 8.3607 |
| 31.0 | 6.2712 | 36.9 | 8.1400 |
| 30.9 | 5.9592 | 45.8 | 9.1416 |


Independent variable X: SalePrice

Dependent variable Y: Taxes

- qualitative analysis




Calculate correlation

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} = 0.8760$$

Independent variable Y: SalePrice

Dependent variable X: Taxes

$$n = 24 \qquad \bar{x} = 34.6125 \qquad \bar{y} = 6.4049$$

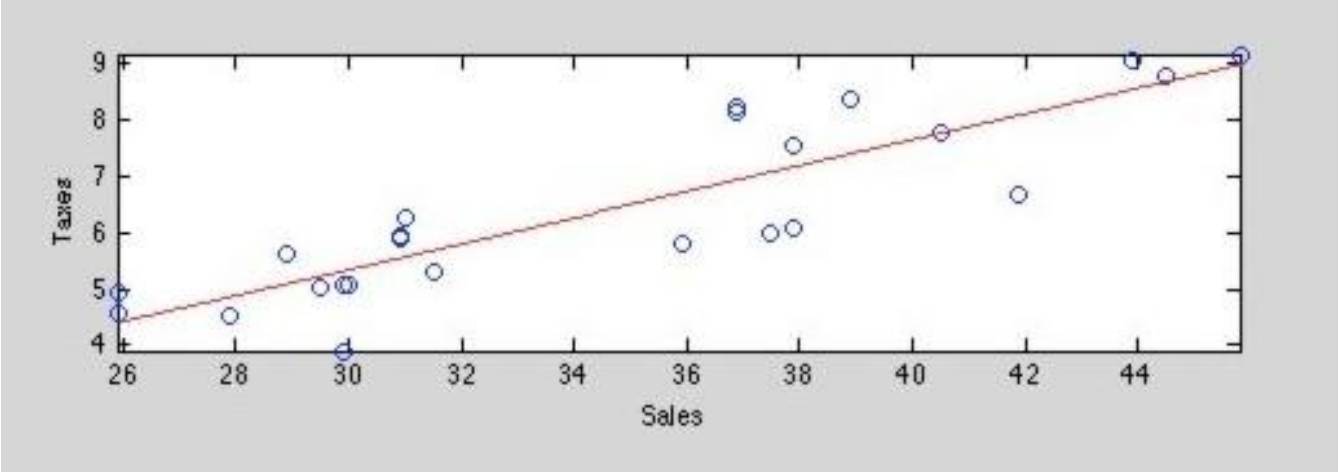
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 829.0462$$

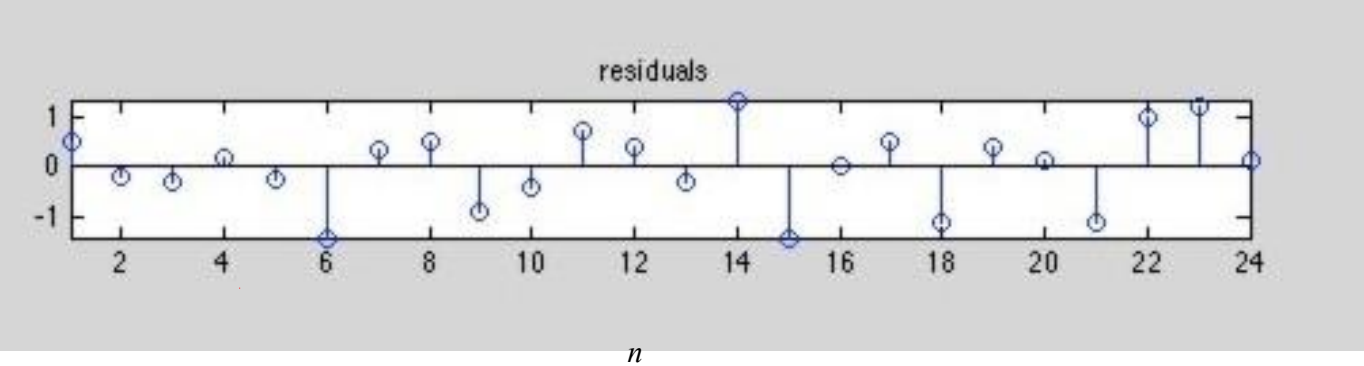
$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = 191.3612$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{191.3612}{829.0462} = 0.2308$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 6.4049 - 0.2308 \times 34.6125 = -1.5837$$

Fitted simple linear regression model $\hat{y} = -1.5837 + 0.2308x$




- residuals: $\hat{\sigma}^2 = MSE = \dfrac{\sum_{i=1}^{n} r_i^2}{n-2} = 0.6088$

- standard error of regression coefficients

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0.6088}{829.0462}} = 0.0271$$

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = \sqrt{0.6088\left[\frac{1}{24} + \frac{34.6125^2}{829.0462}\right]} = 0.9514$$

- test

Test $H_0$: $\beta_1 = 0$ using the $t$-test; use $\alpha = 0.005$

- calculate test statistics

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0.2308}{0.0271} = 8.5166$$

- threshold

$t_{\alpha/2,n-2} = t_{0.0025,22} = 3.119$

- value of test statistic is greater than threshold
- → reject $H_0$

- construct confidence interval for slope parameter

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$t_{\alpha/2,n-2} = t_{0.0025,22} = 3.119$$

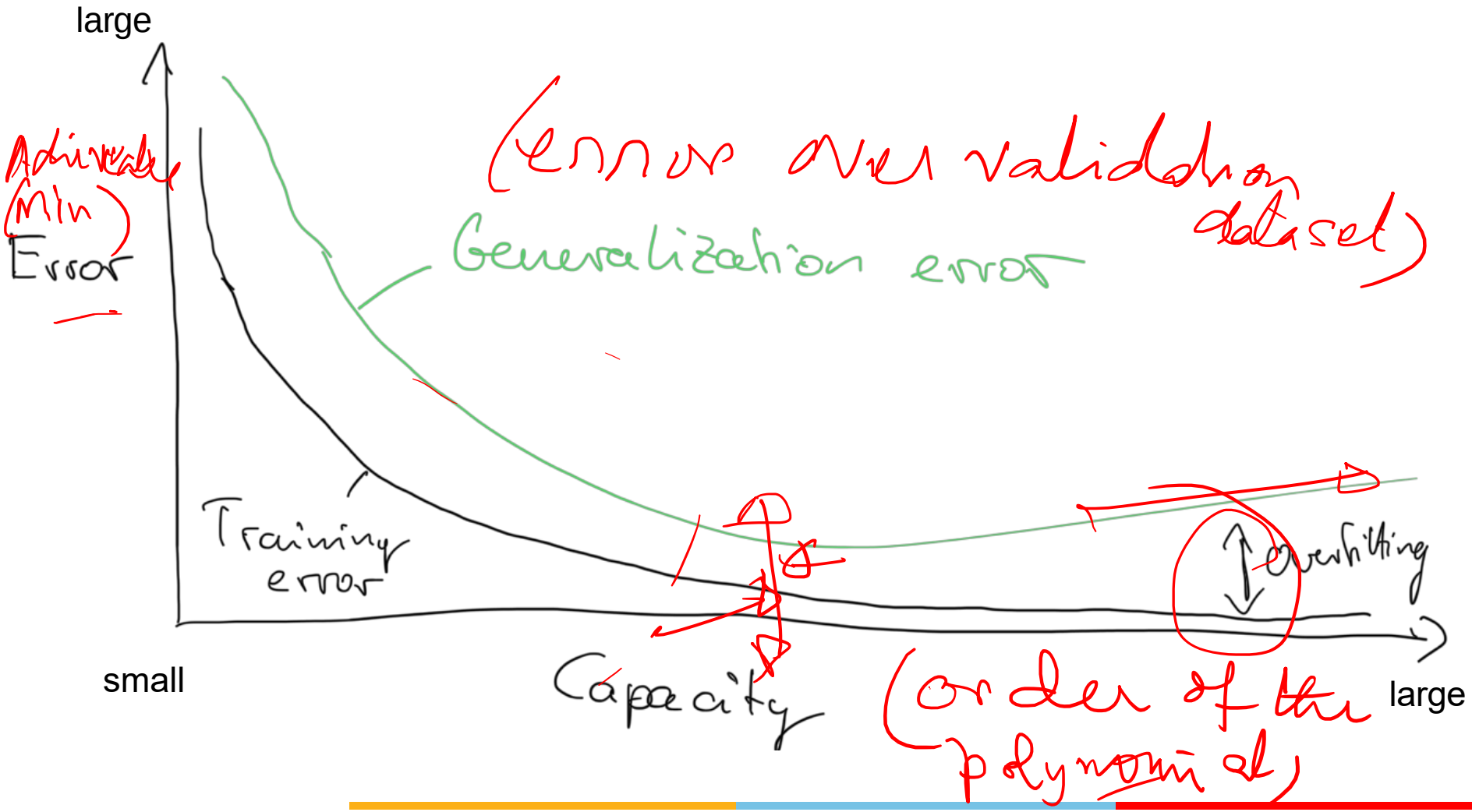
$$0.2308 - 3.119 \times 0.0271 \le \beta_1 \le 0.2308 + 3.119 \times 0.0271$$

$$0.14631 \le \beta_1 \le 0.3153$$

# Bias-Variance Decomposition and Bias-Variance Trade-off

**(and how it related to overfitting and underfitting)**

# Overfitting and Underfitting



# Bias-Variance Decomposition

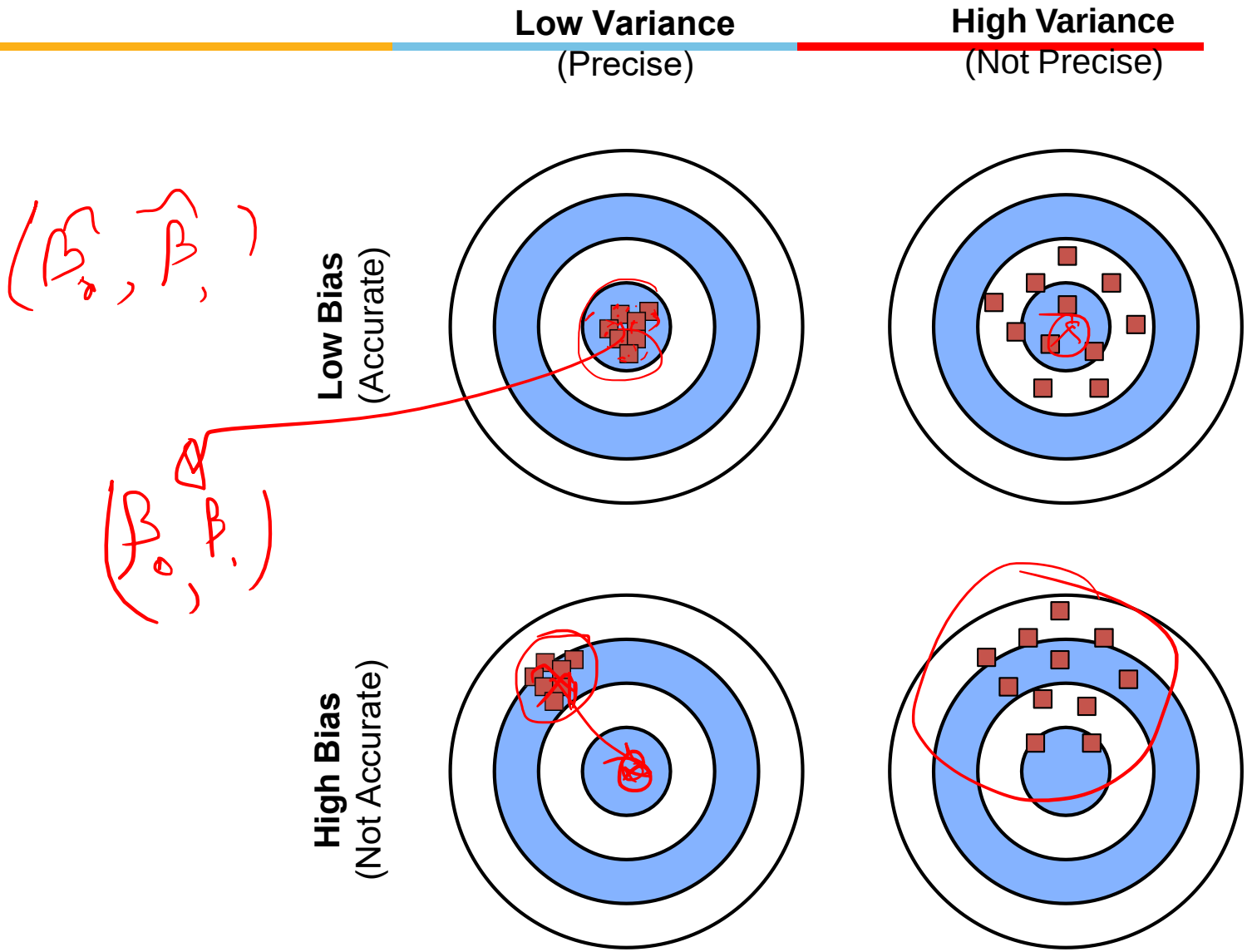


- Decomposition of the loss into bias and variance help us understand learning algorithms, concepts are related to underfitting and overfitting

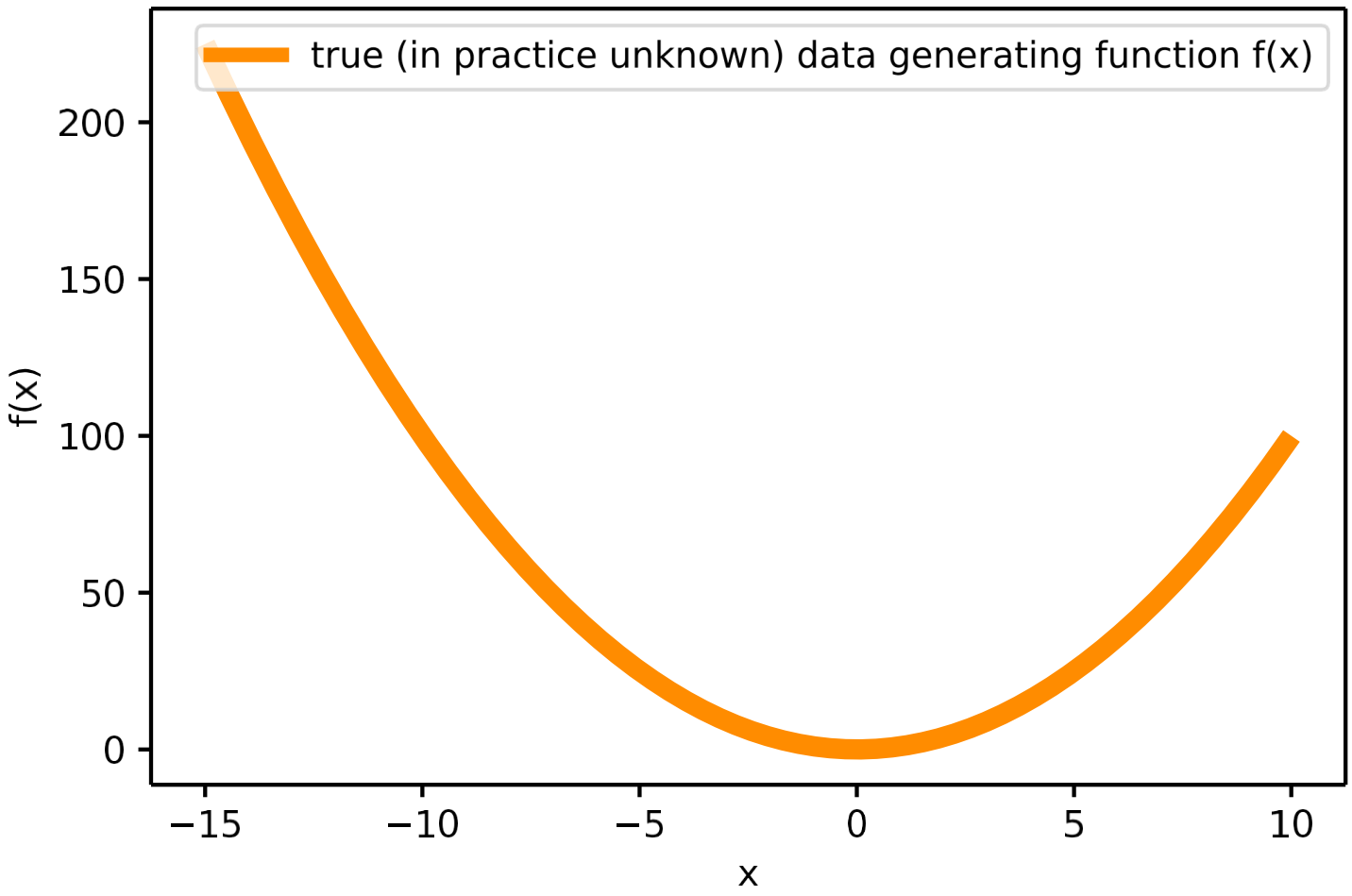
- Helps explain why ensemble methods might perform better than single models

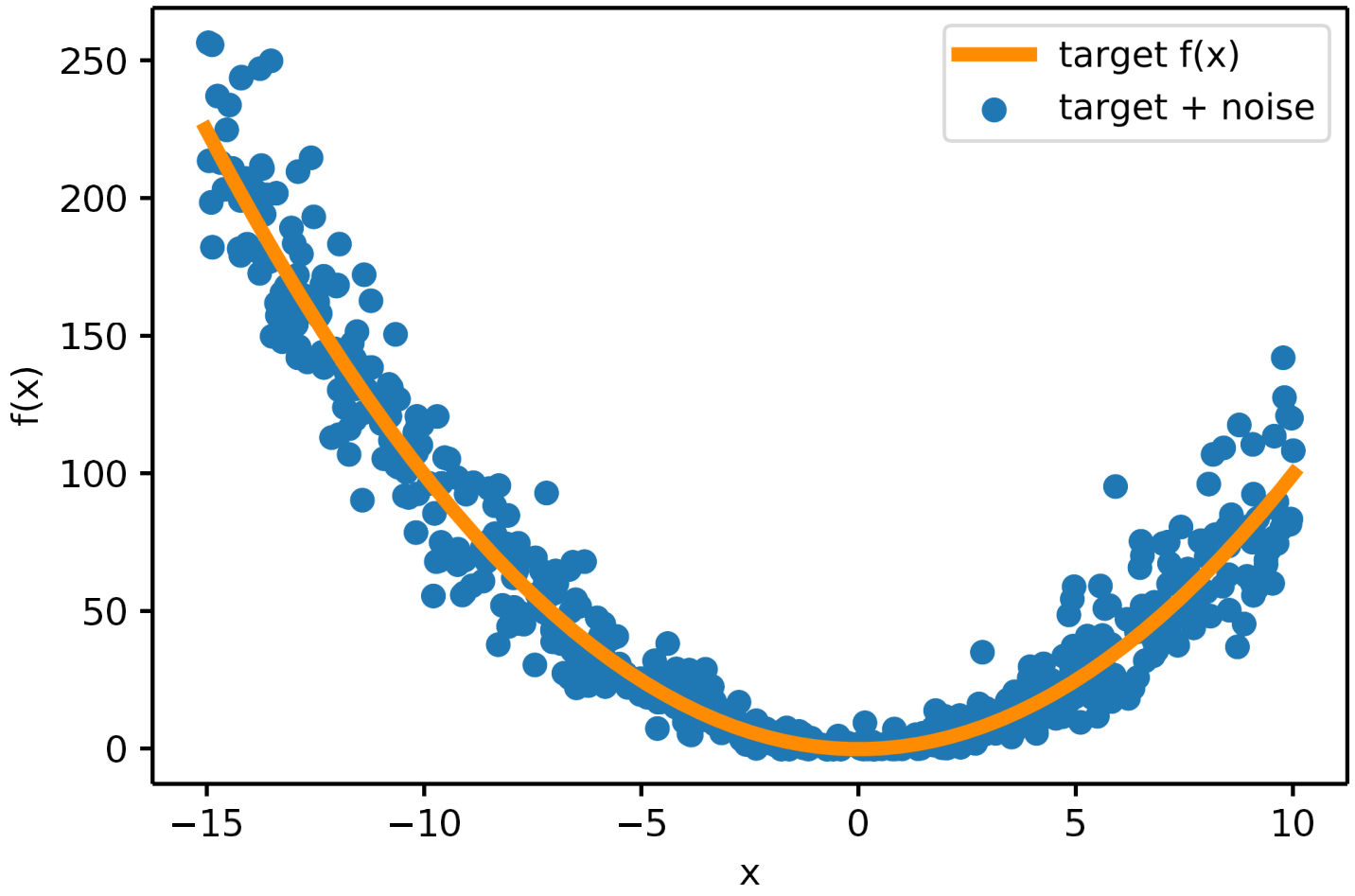
# Bias-Variance Intuition



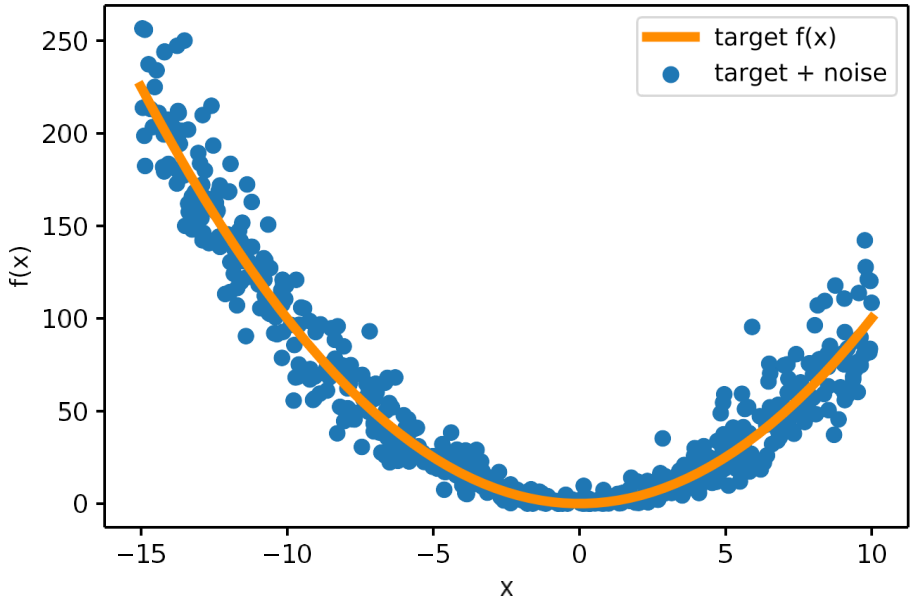


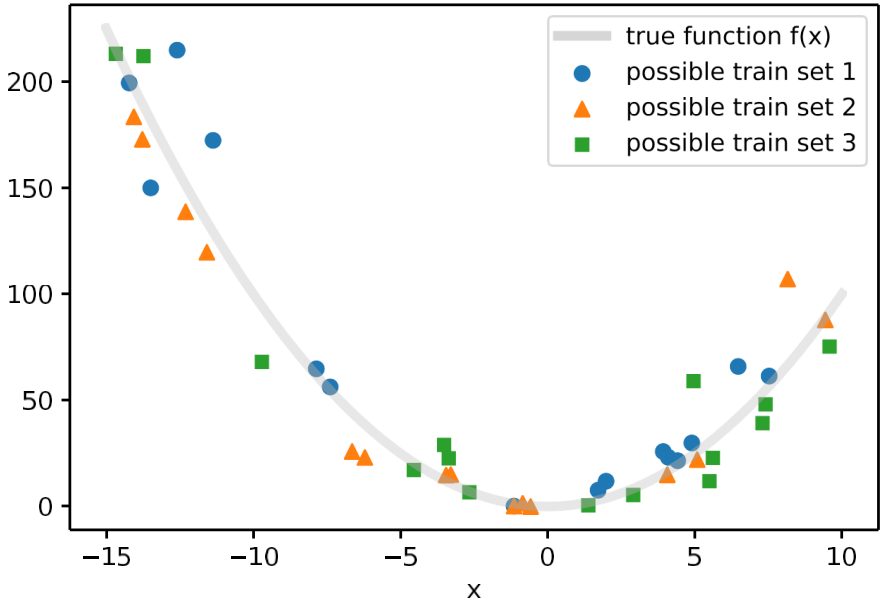
# Bias-Variance Intuition



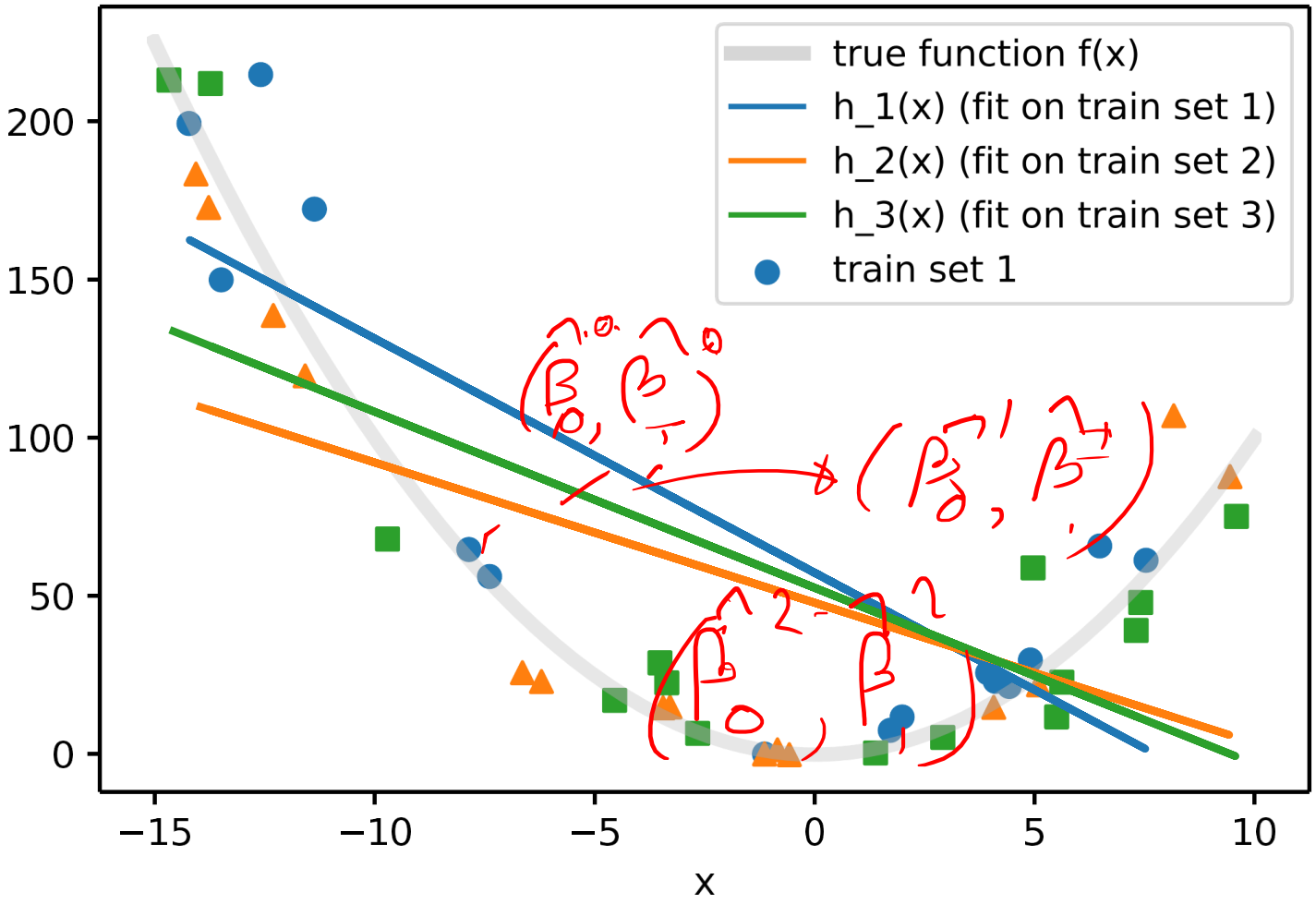

target f(x)
target + noise
250
200
150
100
50
0
f(x)
−15
−10
−5
0
5
10
x

# Bias-Variance Intuition



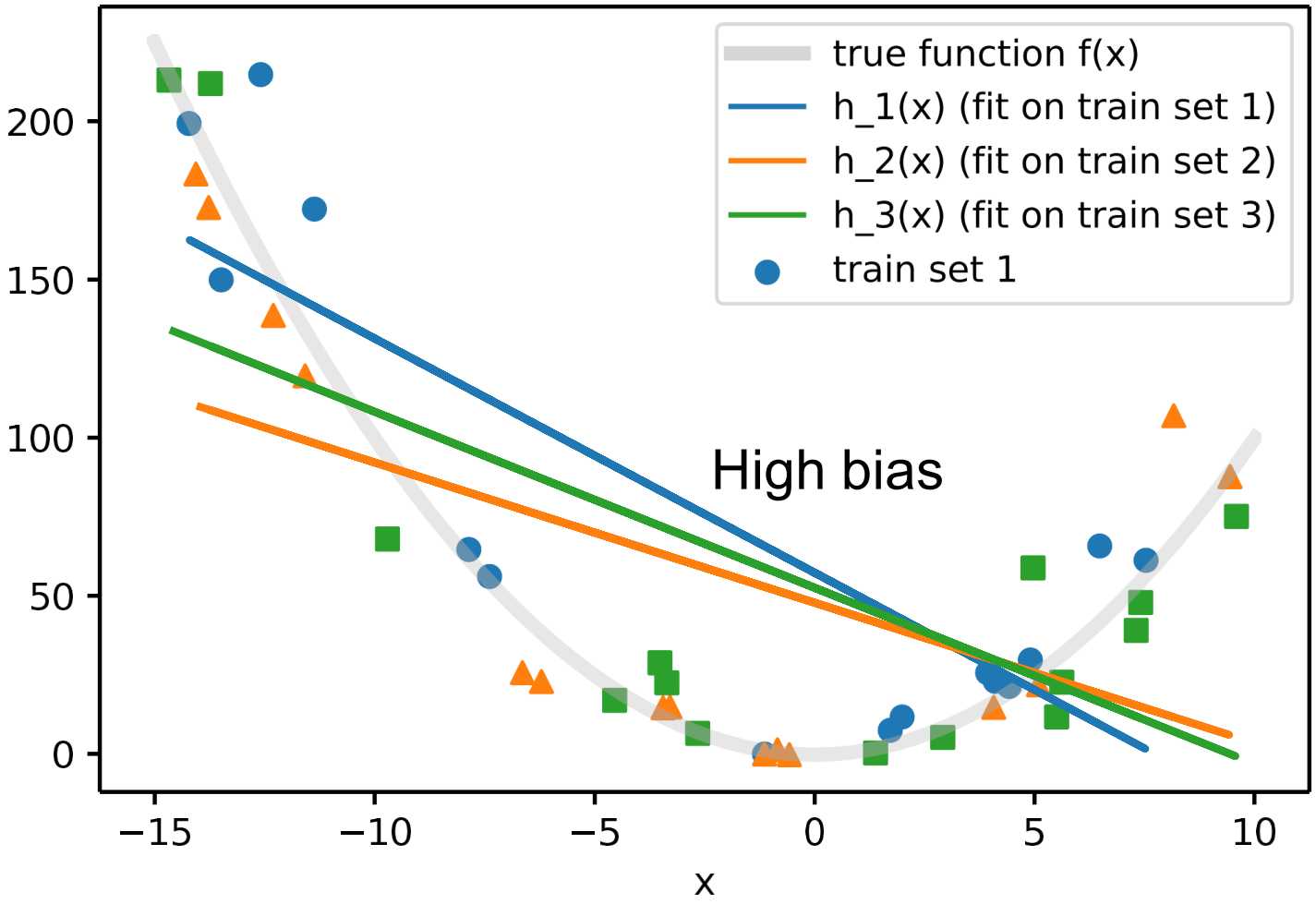



# Bias-Variance Intuition




linear regression models

# Bias-Variance Intuition




linear regression models

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

# Terminology

Point estimator $\hat{\vartheta}$ of some parameter

(could also be a function, e.g., the hypothesis is an estimator of some target function)

# Terminology



Point estimator $\hat{\vartheta}$ of some parameter

(could also be a function, e.g., the hypothesis is an estimator of some target function)

$E(\hat{\theta})$ $\hat{\theta}^1 + \hat{\theta}^2 + \hat{\theta}^3$

$\hat{\theta}^1 + \hat{\theta}^2 + \hat{\theta}^3$

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

$\vartheta$ = true value

# Terminology



## General Definition

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

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

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

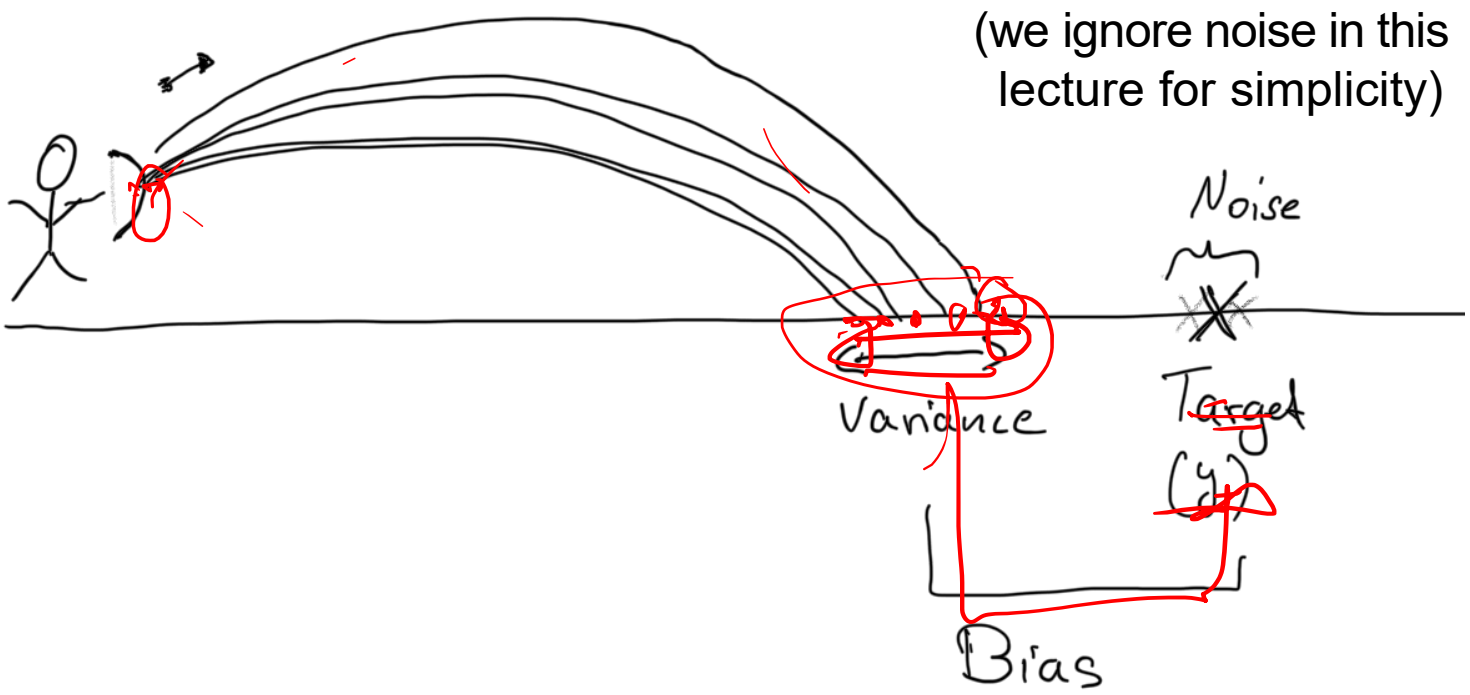
# Terminology



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

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

## Intuition



# Terminology

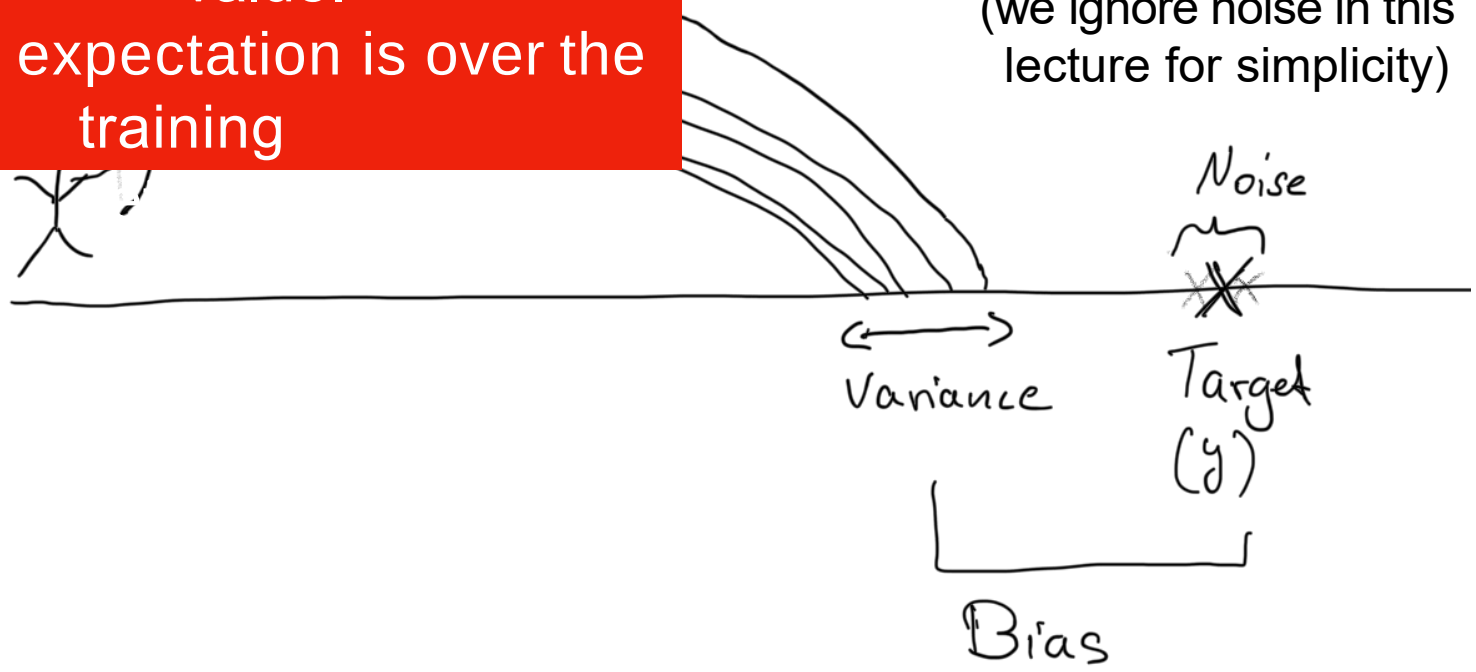


$$\text{Bias}[\hat{\vartheta}] = E[\hat{\vartheta}] - \vartheta$$

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

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

(we ignore noise in this lecture for simplicity)



# Terminology

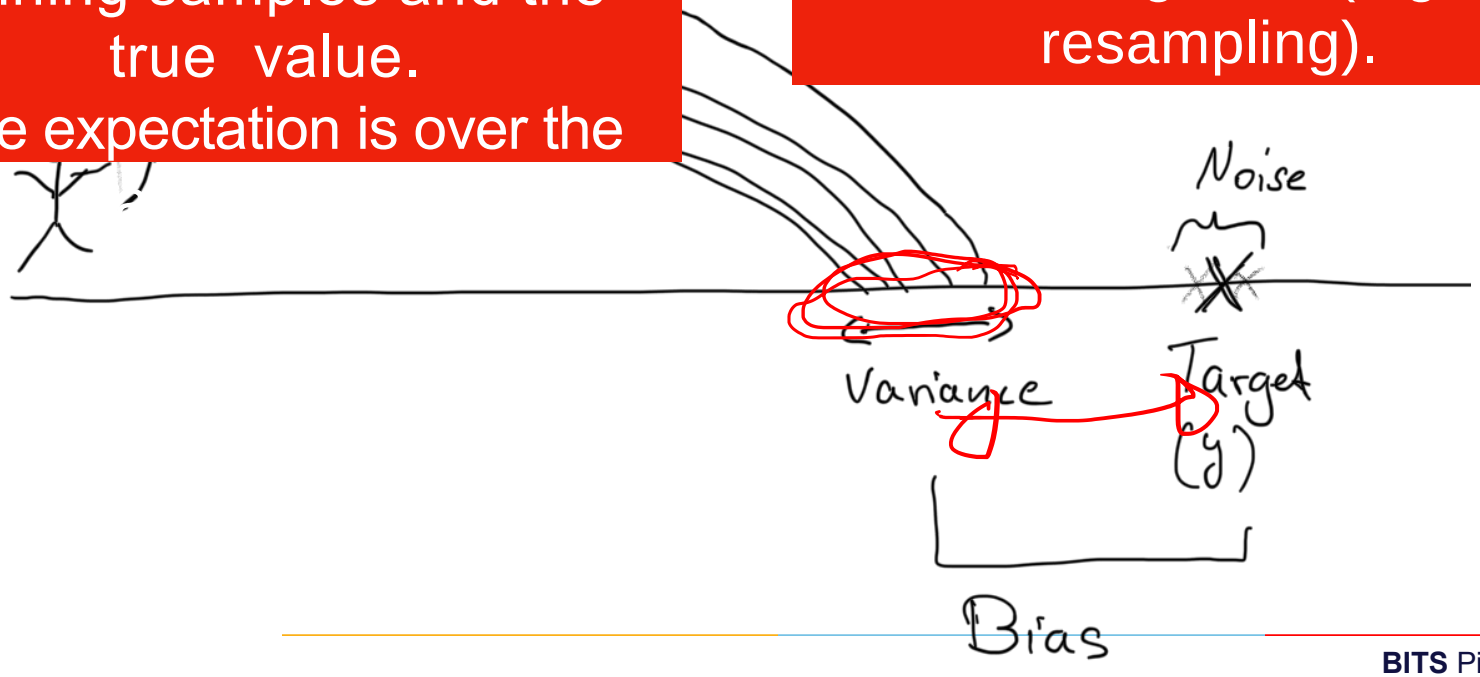


$$\text{Bias}[\hat{\vartheta}] = E[\hat{\vartheta}] - \vartheta$$

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

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

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



# Bias-Variance Decomposition

innovate achieve lead

$$\text{Loss} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

# Bias-Variance of the Squared Error

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

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

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

**"ML Notation" for Squared Error Loss**

$y = f(x)$ target

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

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

for simplicity, we ignore the noise term

(Next slides: the expectation is over the training data, i.e, the average estimator from different training samples)

# Bias-Variance of the Squared Error

**"ML Notation" for**

$y = f(x)$ target

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

**Squared Error Loss**

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

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

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

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

# Bias-Variance of the Squared Error

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

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

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

???

$$E[2(y - E[\hat{y}])(E[\hat{y}] - \hat{y})] = 2E[(y - E[\hat{y}])(E[\hat{y}] - \hat{y})]$$

$$= 2(y - E[\hat{y}])E[(E[\hat{y}] - \hat{y})]$$

$$= 2(y - E[\hat{y}])(E[E[\hat{y}]] - E[\hat{y}])$$

$$= 2(y - E[\hat{y}])(E[\hat{y}] - E[\hat{y}])$$

$$= 0$$

# Bias-Variance of the Squared Error

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

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

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

Bias Variance

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

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

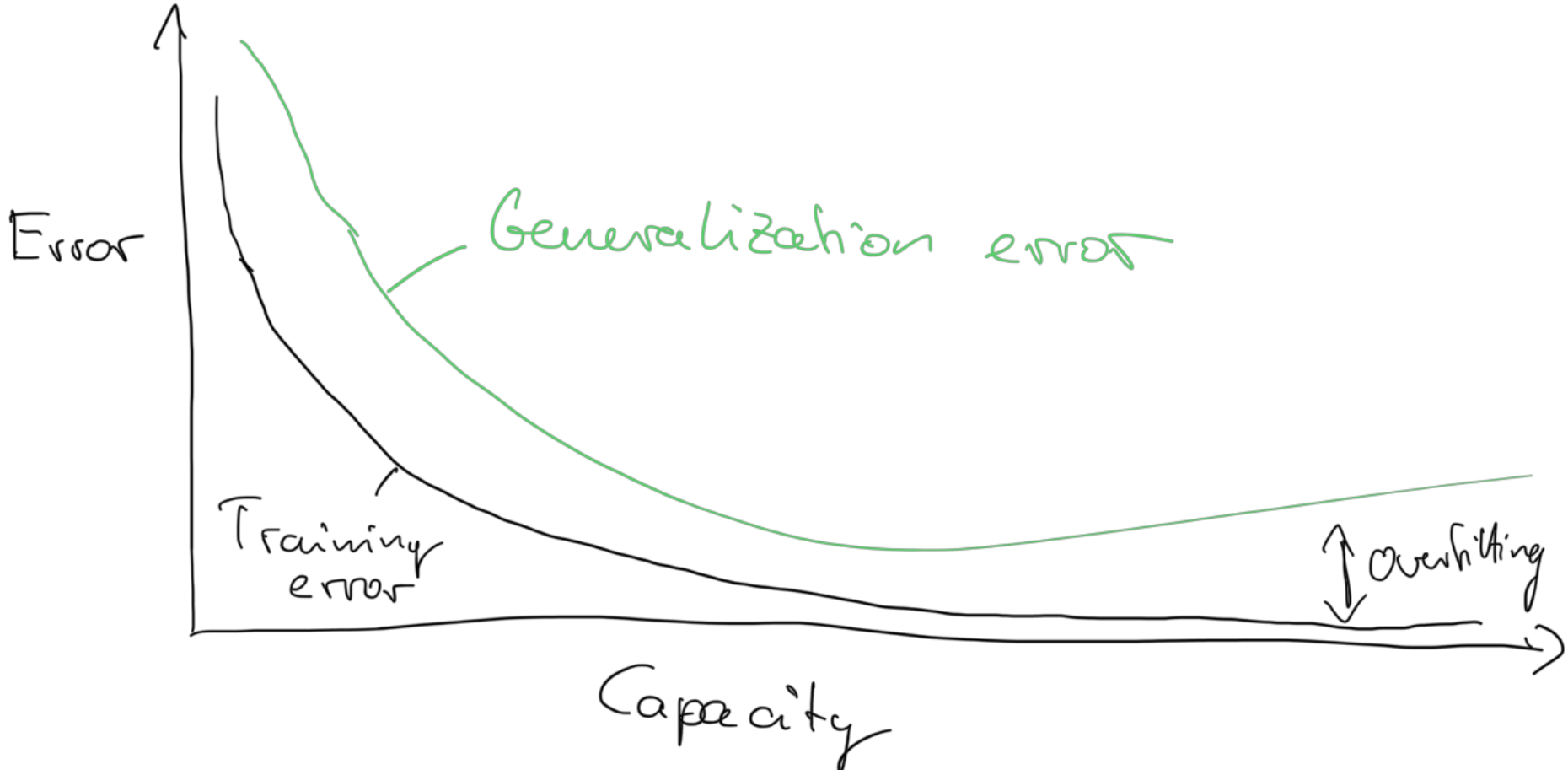
$$= \text{Bias}^2 + \text{Var}$$

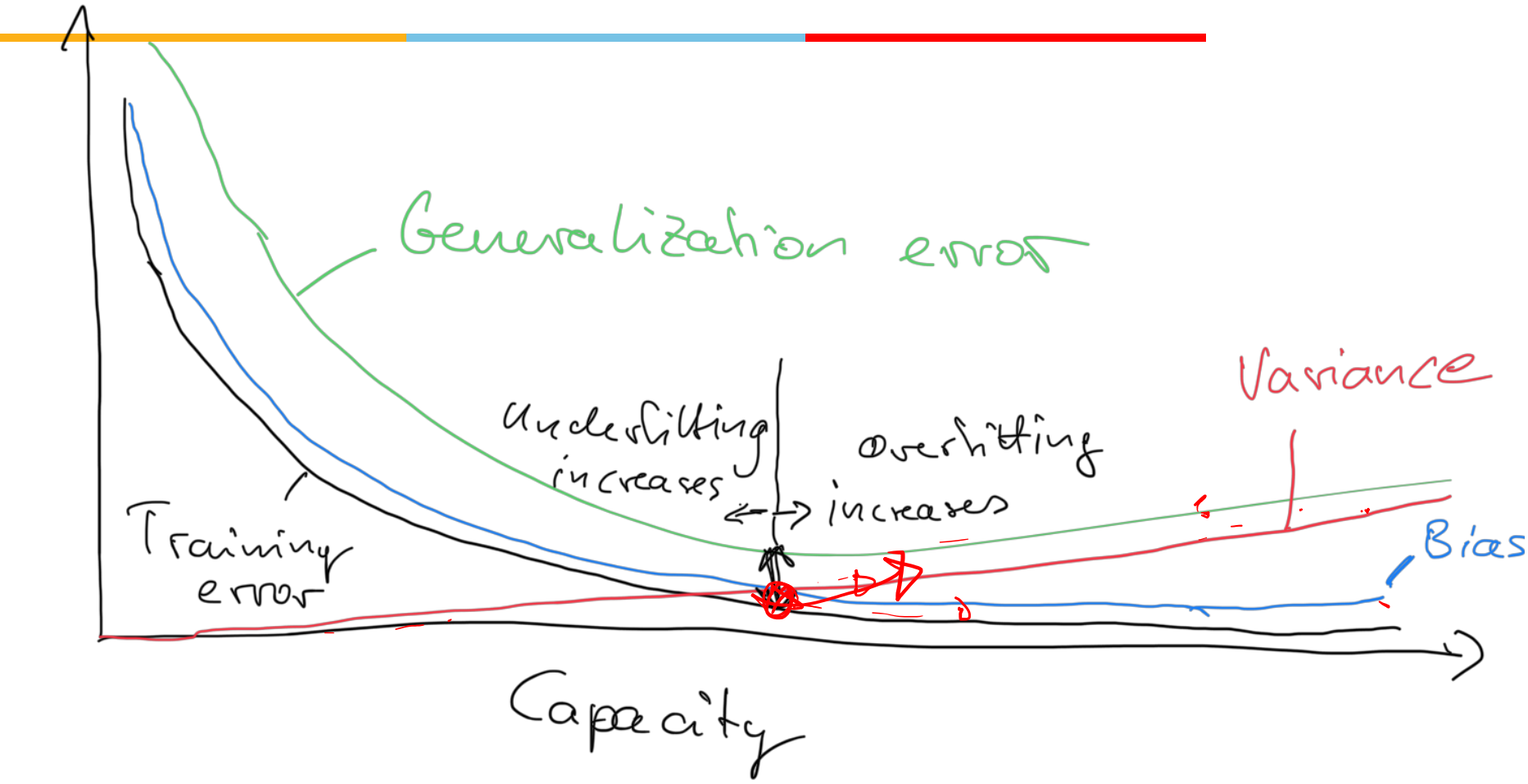
$$\text{Bias}[\hat{\vartheta}] = E[\hat{\vartheta}] - \vartheta$$

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

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

# Now, how is this related to overfitting and underfitting?


Error
Generalization error
Training error
overfitting
Capacity


Generalization error
Variance
Underfitting increases
overfitting increases
Training error
Bias
Capacity

# Thank you !