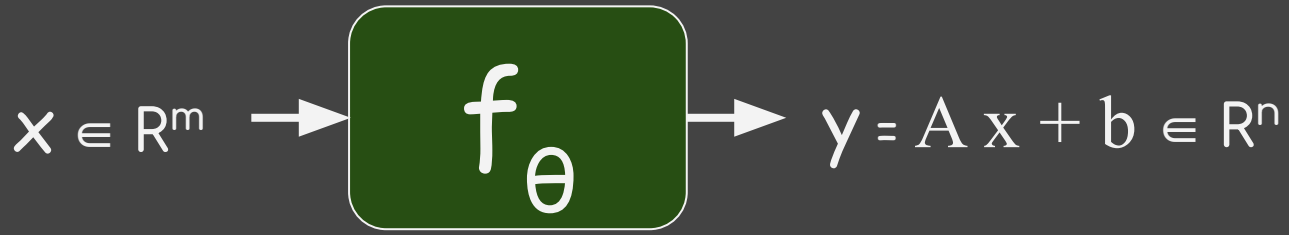
Linear Algebra for Computer Science

Lecture 31

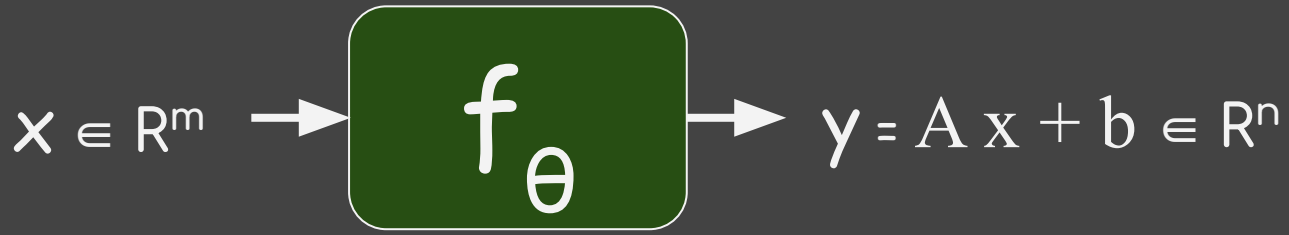
Parameter Learning, Cost Function, Overfitting, Cross-validation

Example: Linear Regression



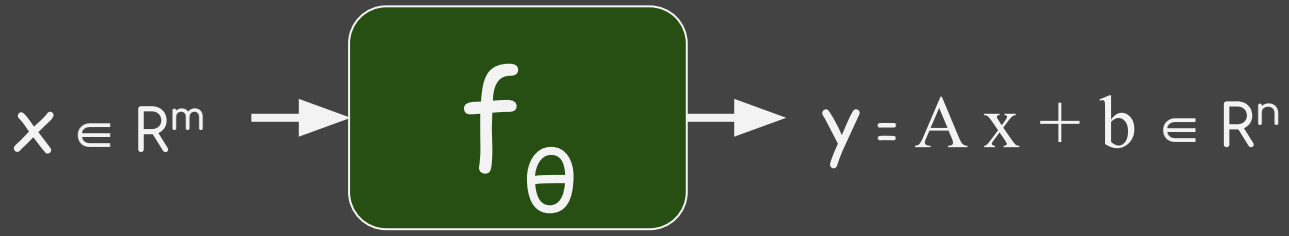
Example: Linear Regression

K. N. Toosi
University of Technology



A: ? by ? matrix
b: ?-D vector

Example: Linear Regression

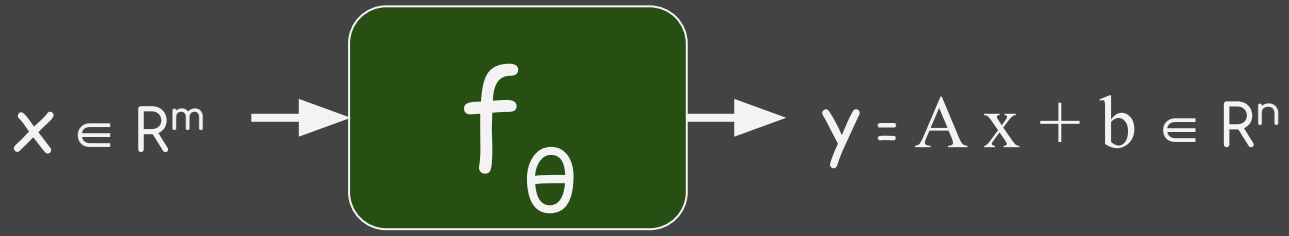


A: n by m matrix

b: n-D vector

Example: Linear Regression

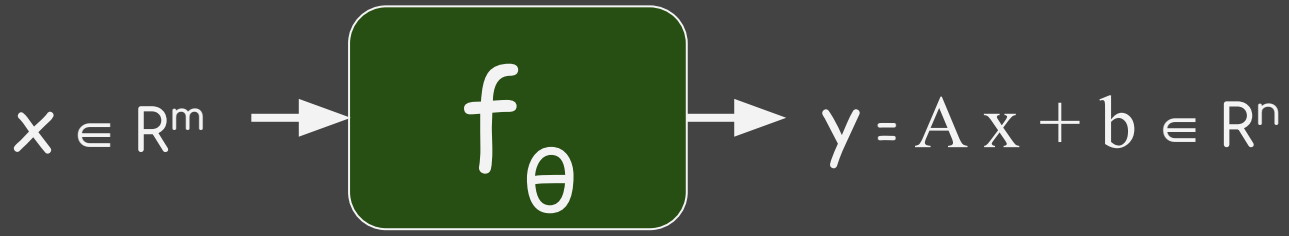
K. N. Toosi
University of Technology



$$y = f(\theta, x)$$
$$\theta = ?$$

Example: Linear Regression

K. N. Toosi
University of Technology



$$y = f(\theta, x)$$
$$\theta = (A, b)$$

Example: Linear Regression

K. N. Toosi
University of Technology

$f(x) = Ax + b$

$f: \mathbb{R}^m \longrightarrow \mathbb{R}^n$

is f linear?

$f(\alpha x) = A(\alpha x) + b = \alpha Ax + b$

$\alpha f(x) = \alpha Ax + \alpha b$

$\Rightarrow f$ is not linear in general $(b \neq 0)$

Affine maps

K. N. Toosi
University of Technology

$g(x) = f(x+x_0) - f(x_0) = A(x+x_0) + b - (Ax_0 + b) = Ax$

$g(x) = f(x) - f(0) = Ax$

$g(x) = f(x) - b = Ax$

$f(x) = Ax + b$

affine function

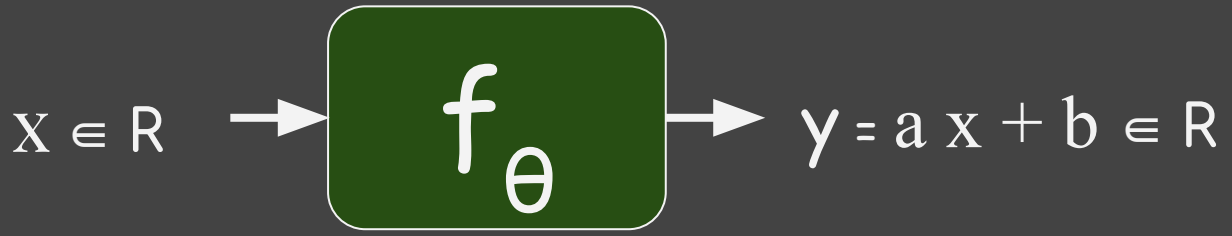
$f(x) : \mathbb{R}^m \to \mathbb{R}^n \quad \exists v \in \mathbb{R}^n \quad g(x) = f(x) - v$

linear

$g(x) = f(x - x_0) - f(x_0)$

Example: Linear Regression

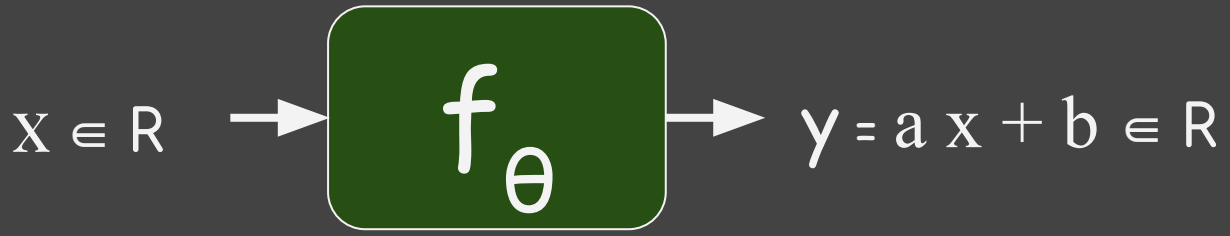
K. N. Toosi
University of Technology



$$y = f(\theta, x)$$
$$\theta = ?$$

Example: Linear Regression

K. N. Toosi
University of Technology



$$y = f(\theta, x)$$
$$\theta = (a,b)$$

Example: Linear Regression

K. N. Toosi
University of Technology

$f: \mathbb{R} \longrightarrow \mathbb{R}$

$f(x) = ax + b$ — f is not linear

$h(x, \theta) = h\left(x, \begin{bmatrix} a \\ b \end{bmatrix}\right) = ax + b$

- is h linear in x? NO
- is h linear in θ? YES

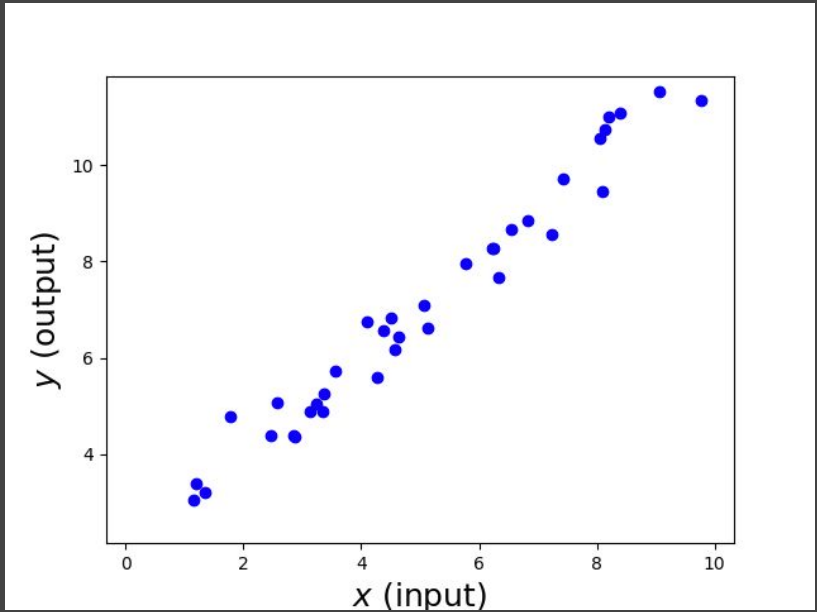
$\begin{bmatrix} a \\ b \end{bmatrix} = \theta \in \mathbb{R}^2$

$h(x, \theta) = ax + b = \begin{bmatrix} x & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = v^T \begin{bmatrix} a \\ b \end{bmatrix} = v^T \theta$

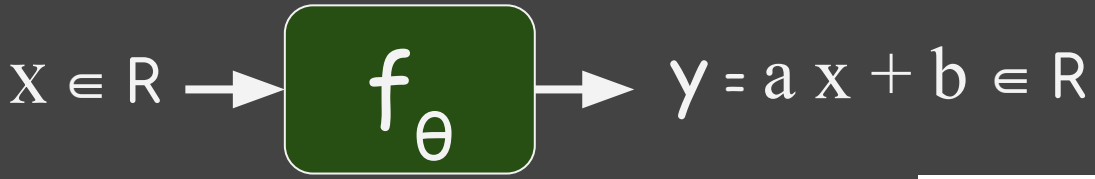
Linear Regression: Cost function

K. N. Toosi
University of Technology

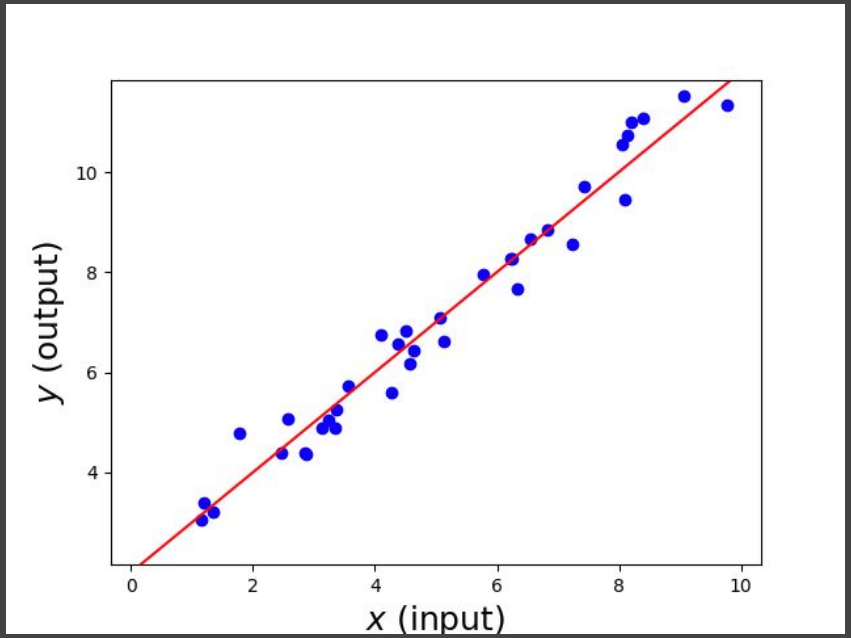
Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$



Linear Regression: Cost function



Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$



Linear Regression: Cost function

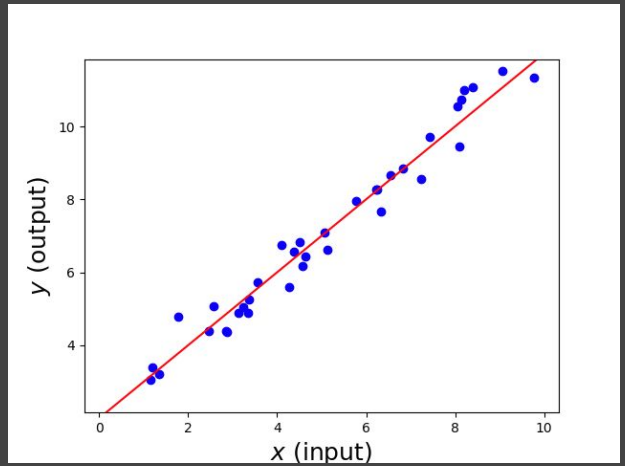
K. N. Toosi
University of Technology

$$x \in R \rightarrow f_\theta \rightarrow y = a\,x + b \in R$$

Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

cost function:

$$C(\theta) = \Sigma_{i=1..N}\, d(\, f(\theta, x_i), y_i\,)$$



Linear Regression: Cost function

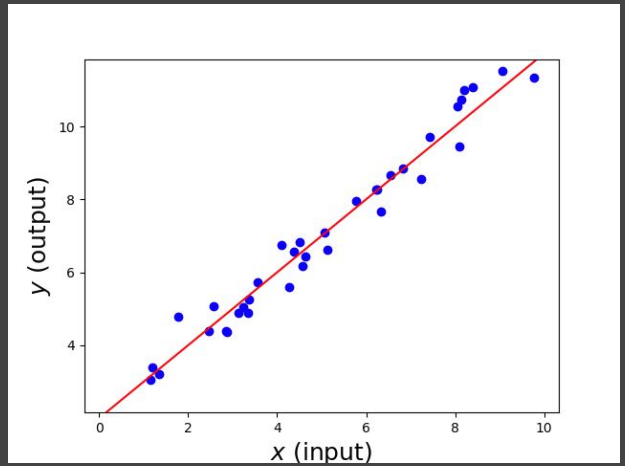
K. N. Toosi
University of Technology

$x \in R \rightarrow f_\theta \rightarrow y = a\,x + b \in R$

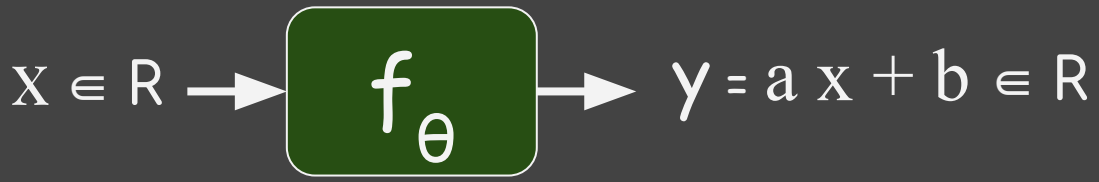
Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

cost function:

$$C(a,b) = \Sigma_{i=1..n}\, d(\, f(a,b, x_i), y_i\,)$$



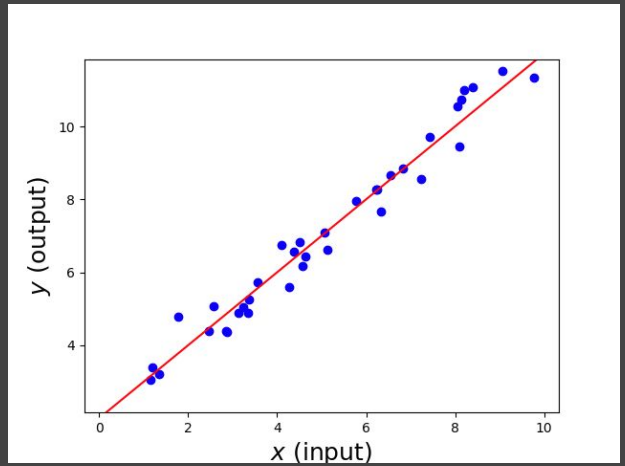
Linear Regression: Cost function



Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

cost function:

$$C(a,b) = \Sigma_{i=1..n}\, d(\, f(a,b, x_i), y_i\,)$$
$$= \Sigma_{i=1..n}\, d(\, a\, x_i + b, y_i\,)$$



Linear Regression: Cost function

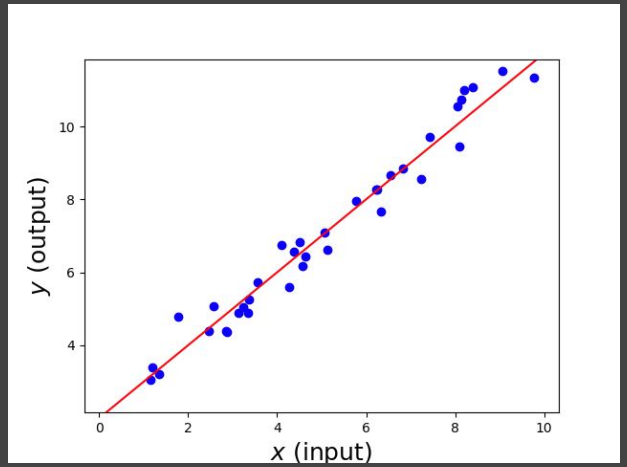
K. N. Toosi
University of Technology

$x \in R \rightarrow f_\theta \rightarrow y = a\,x + b \in R$

Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

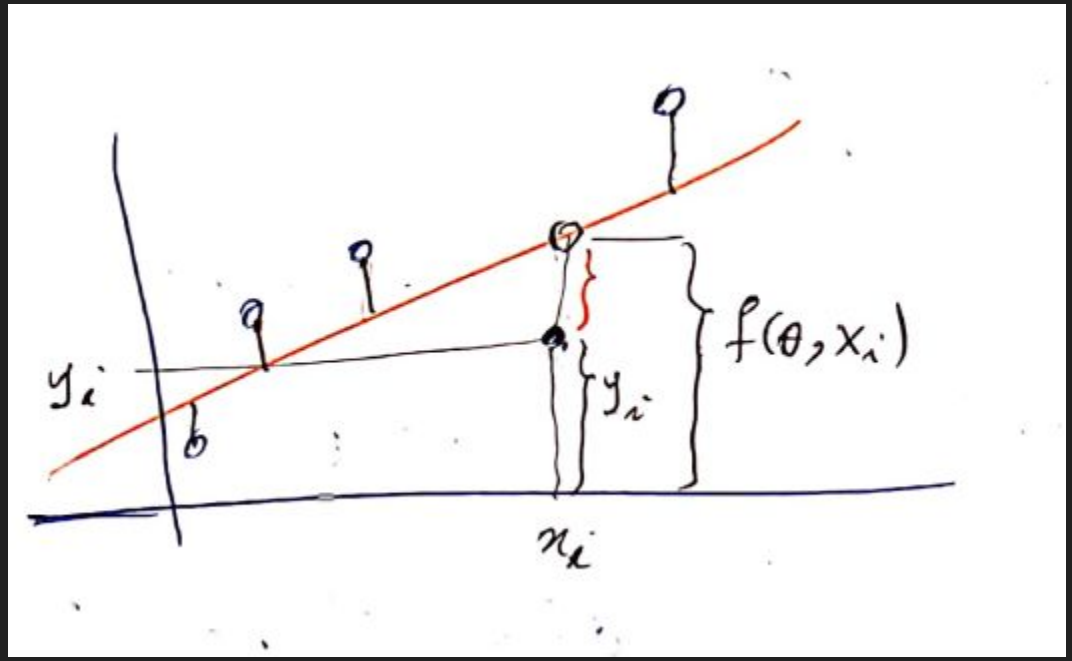
cost function (sum of squared errors):

$$C(a,b) = \Sigma_{i=1..n} \left(a\,x_i + b - y_i \right)^2$$

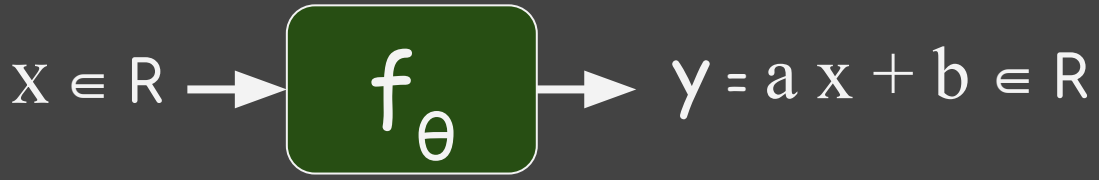


The cost function

K. N. Toosi
University of Technology



Linear Regression: Cost function

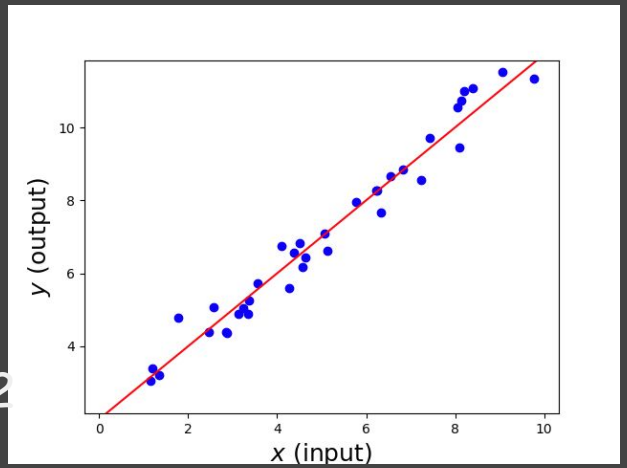


Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

cost function (sum of squared errors):

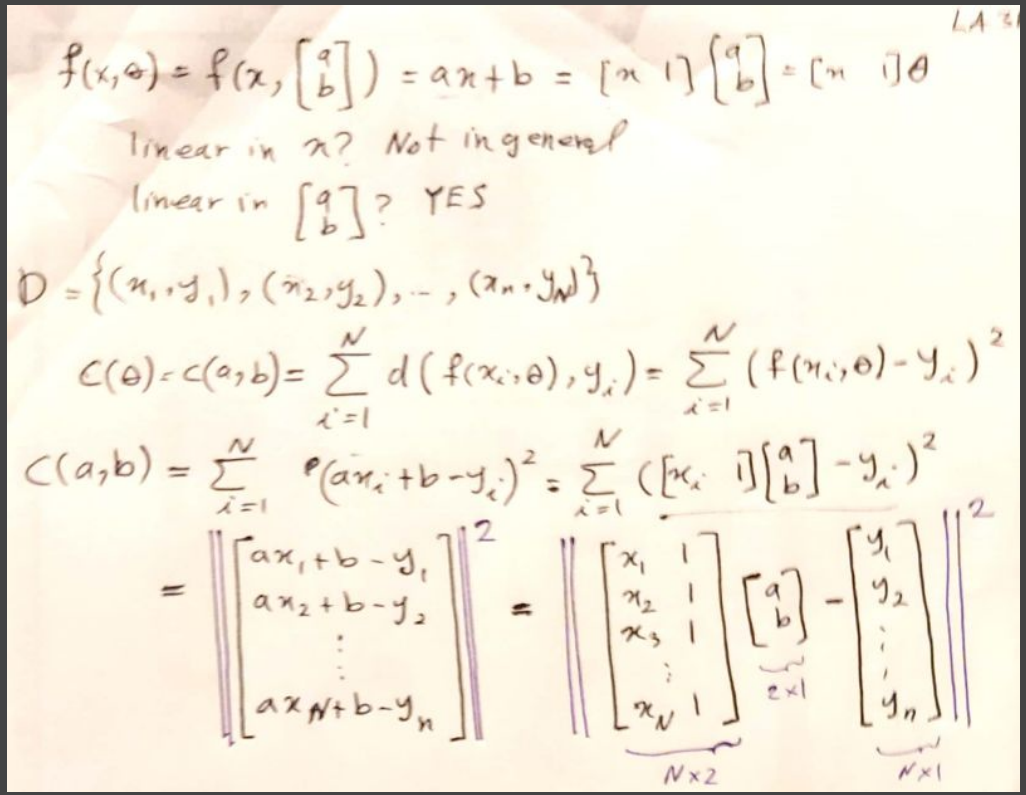
$$C(a,b) = \Sigma_{i=1..n} (a\,x_i + b - y_i)^2$$

$$a^*, b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a\,x_i + b - y_i)^2$$



Linear Regression: Cost function

K. N. Toosi
University of Technology



Example: Linear Regression

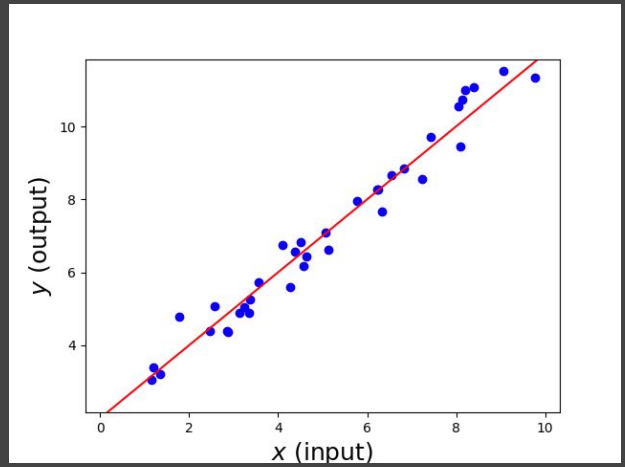
K. N. Toosi
University of Technology

cost function (sum of squared errors):

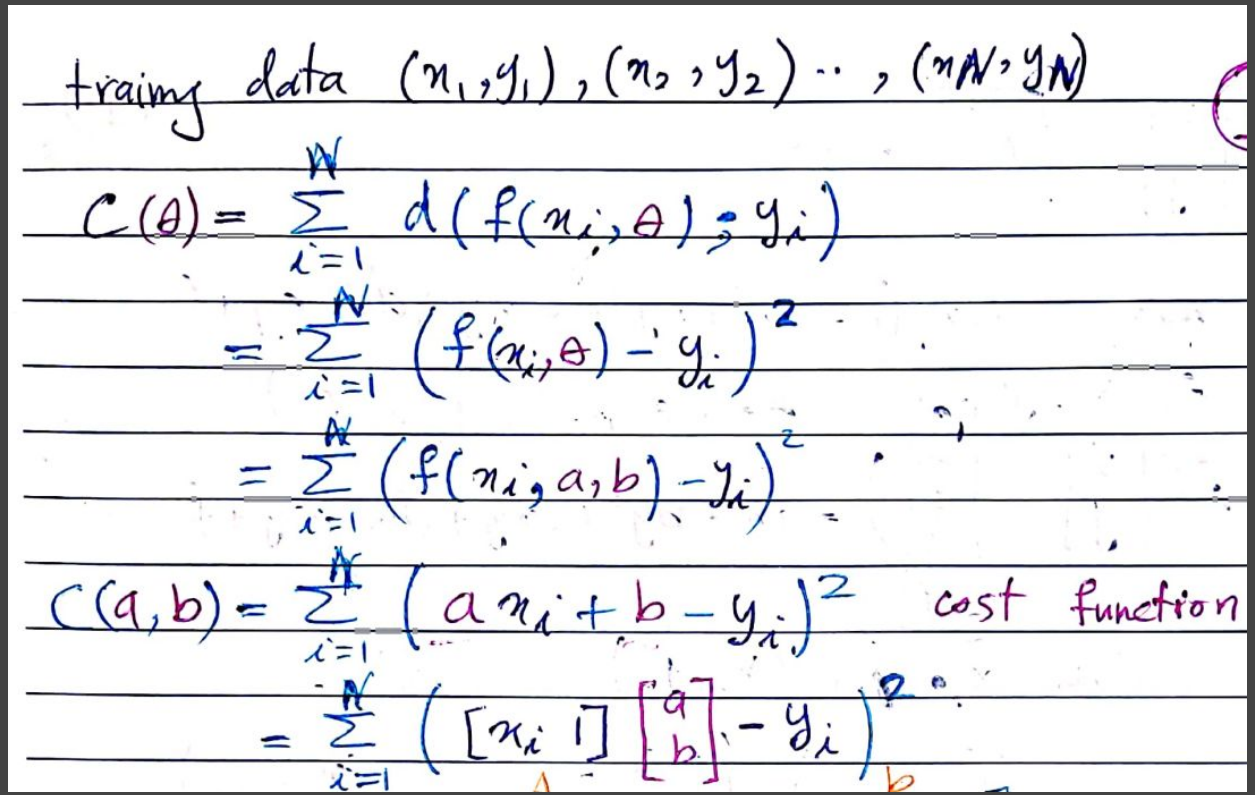
$$C(a,b) = \Sigma_{i=1..n} (a\, x_i + b - y_i)^2$$

$$a^*,b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a\, x_i + b - y_i)^2$$

How to find a^*,b^*?



Solution 1: Least squares



Solution 1: Least squares

K. N. Toosi
University of Technology

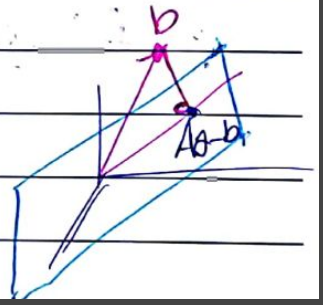
$$C(a,b) = \sum_{i=1}^{N} (a x_i + b - y_i)^2 \quad \text{cost function}$$

$$= \sum_{i=1}^{N} \left(\begin{bmatrix} x_i & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} - y_i \right)^2$$

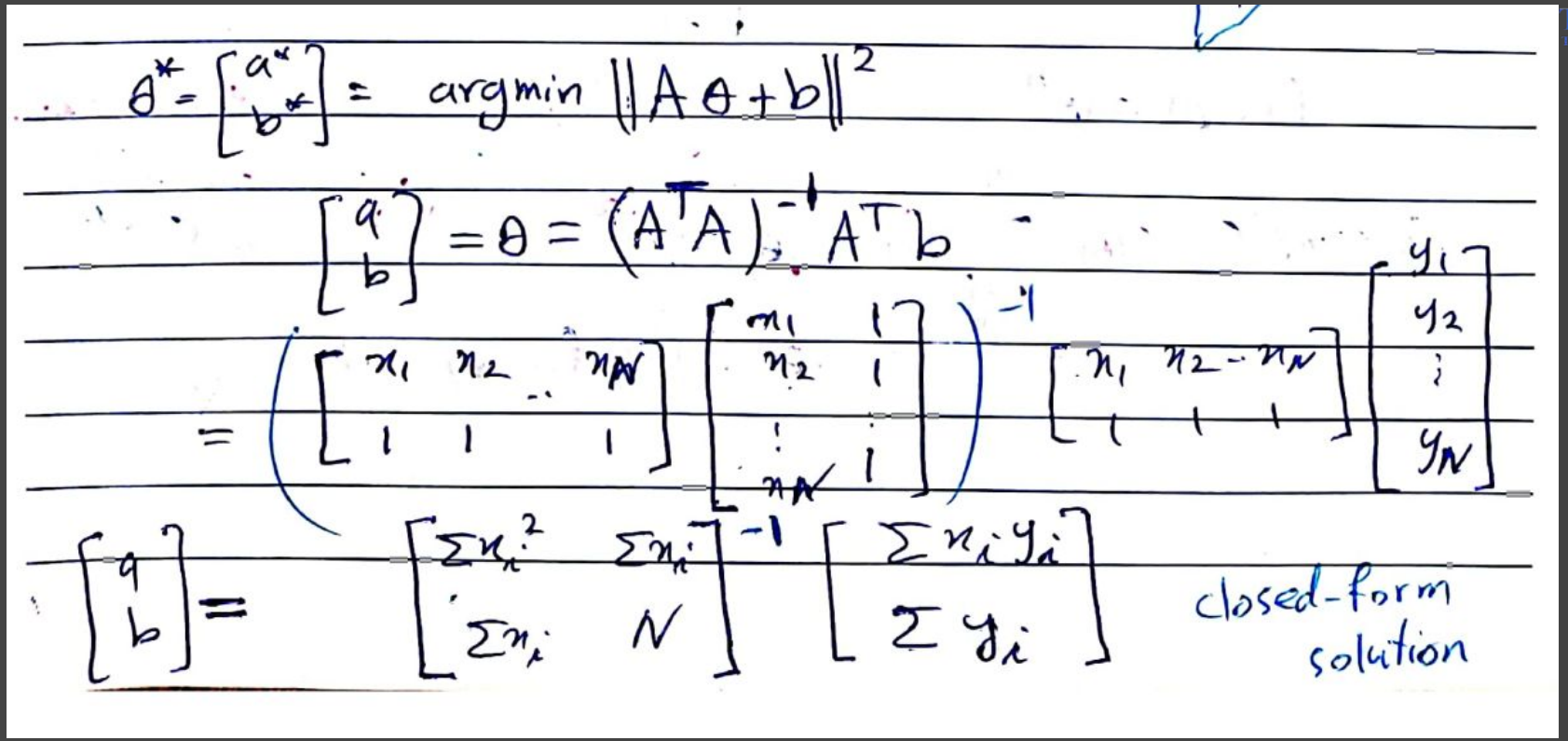
$$= \left\| \begin{bmatrix} a x_1 + b - y_1 \\ a x_2 + b - y_2 \\ a x_3 + b - y_3 \\ \vdots \\ a x_N + b - y_N \end{bmatrix} \right\|^2 = \left\| \underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\theta} - \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix}}_{b} \right\|^2 = \|A\theta - b\|^2$$

$N \times 2$ $\quad 2 \times 1$ $\quad N \times 1$

$N \times 1$



Solution 1: Least squares



Example: Linear Regression

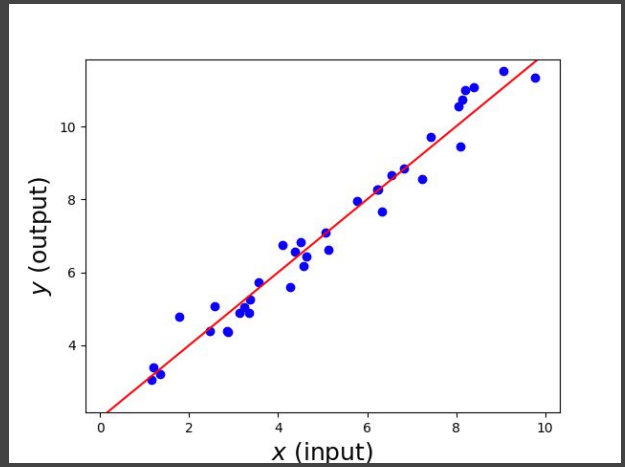
K. N. Toosi
University of Technology

cost function (sum of squared errors):

$$C(a,b) = \Sigma_{i=1..n} (a\, x_i + b - y_i)^2$$

$$a^*,b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a\, x_i + b - y_i)^2$$

How to find a^*,b^*?



Solution 2: partial derivatives

K. N. Toosi
University of Technology

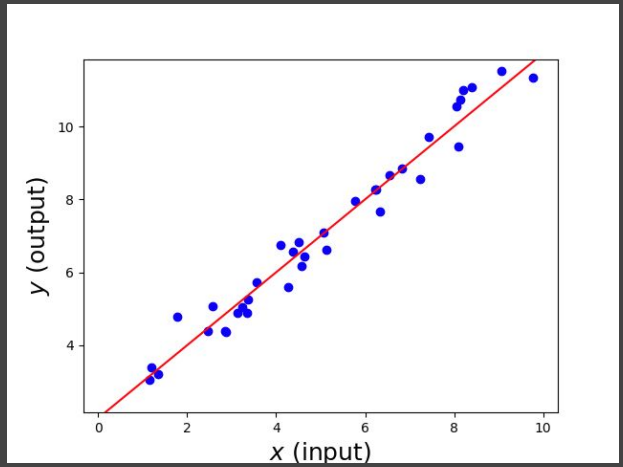
cost function (sum of squared errors):

$$C(a,b) = \Sigma_{i=1..n} (a\, x_i + b - y_i)^2$$

$$a^*, b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a\, x_i + b - y_i)^2$$

$$\partial\, C(a,b) / \partial\, a = 0$$

$$\partial\, C(a,b) / \partial\, b = 0$$



Solution 2: partial derivatives

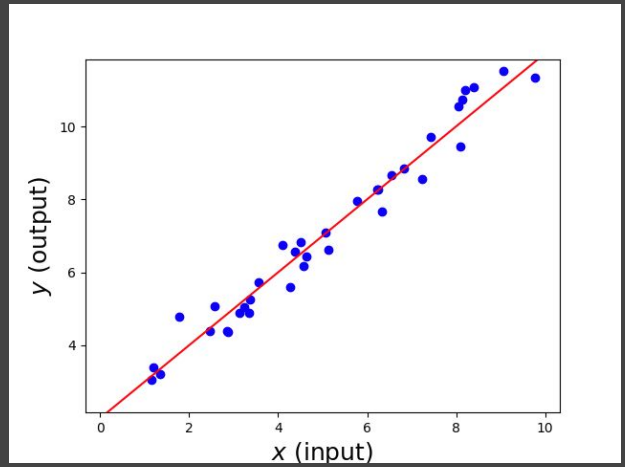
K. N. Toosi
University of Technology

cost function (sum of squared errors):

$$C(a,b) = \Sigma_{i=1..n} (a x_i + b - y_i)^2$$

$$a^*,b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a x_i + b - y_i)^2$$

$$\partial C(a,b) / \partial a = 2 \Sigma_{i=1..n} x_i (a x_i + b - y_i) = 0$$
$$\partial C(a,b) / \partial b = 2 \Sigma_{i=1..n} (a x_i + b - y_i) = 0$$



Solution 2: partial derivatives

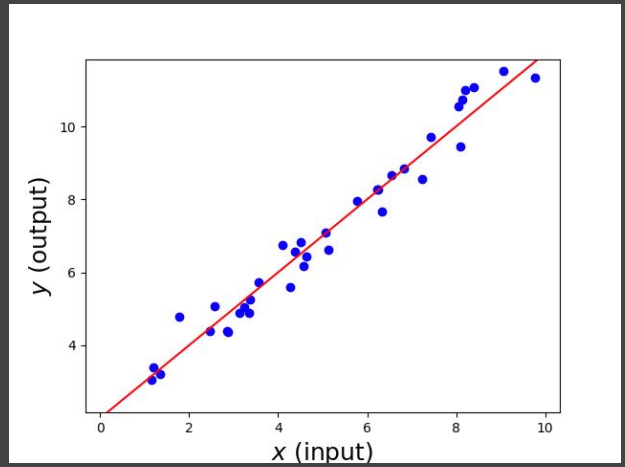
K. N. Toosi
University of Technology

cost function (sum of squared errors):

$$C(a,b) = \Sigma_{i=1..n} (a x_i + b - y_i)^2$$

$$a^*,b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a x_i + b - y_i)^2$$

$$\Sigma_{i=1..n} x_i (a x_i + b - y_i) = 0$$
$$\Sigma_{i=1..n} (a x_i + b - y_i) = 0$$



Solution 2: partial derivatives

$$a^*, b^* = \operatorname{argmin}_{a,b} \sum_{i=1..n} (a x_i + b - y_i)^2$$

$$\sum_{i=1..n} x_i (a x_i + b - y_i) = a \sum_{i=1..n} x_i^2 + b \sum_{i=1..n} x_i - \sum_{i=1..n} x_i y_i = 0$$

$$\sum_{i=1..n} (a x_i + b - y_i) = a \sum_{i=1..n} x_i + b \; n \; - \sum_{i=1..n} y_i = 0$$

Solution 2: partial derivatives

$$a^*, b^* = \operatorname{argmin}_{a,b} \sum_{i=1..n} (a\, x_i + b - y_i)^2$$

$$\left(\sum_{i=1..n} x_i^2\right) a + \left(\sum_{i=1..n} x_i\right) b = \sum_{i=1..n} x_i\, y_i$$

$$\left(\sum_{i=1..n} x_i\right) a + n\, b = \sum_{i=1..n} y_i$$

Solution 2: partial derivatives

$$a^*, b^* = \operatorname{argmin}_{a,b} \sum_{i=1..n} (a\, x_i + b - y_i)^2$$

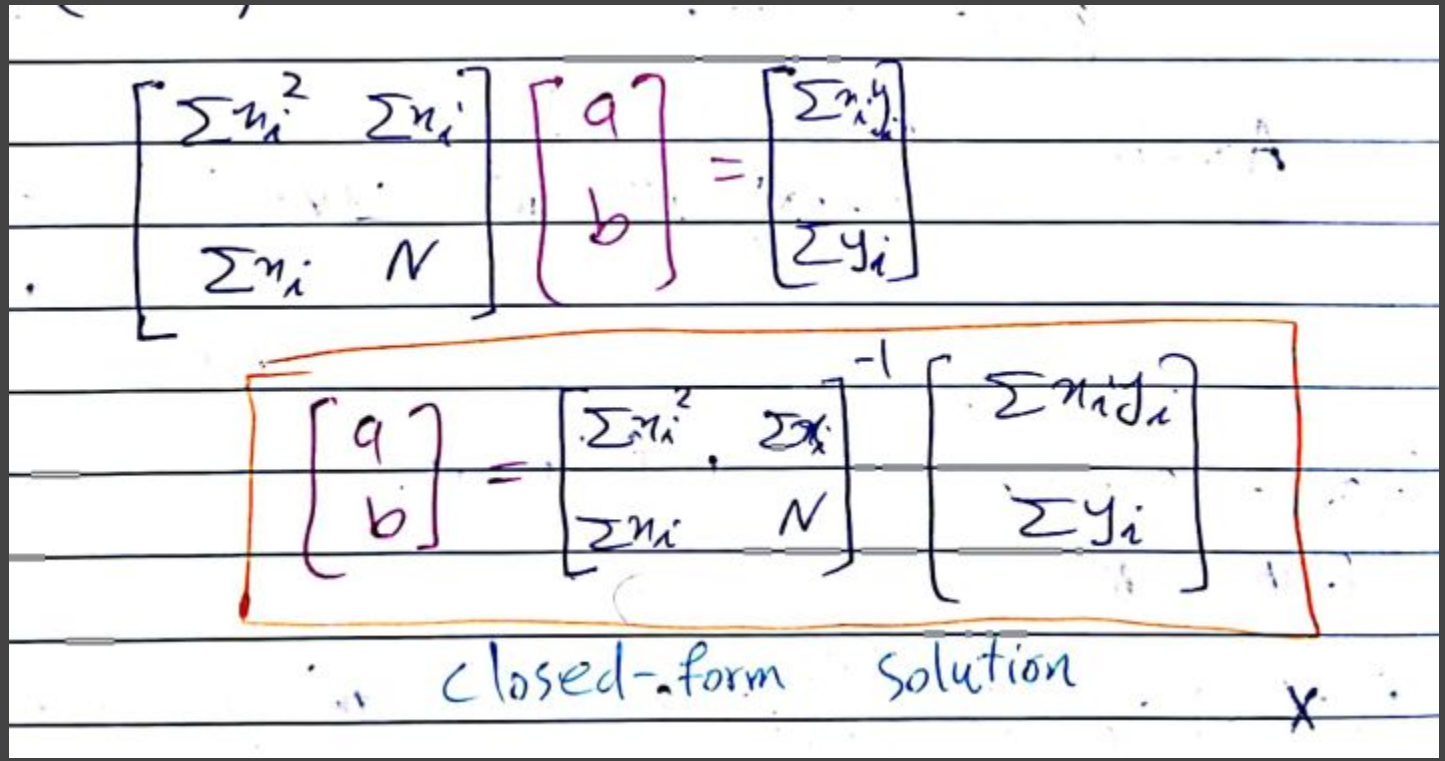
$$\left(\sum_{i=1..n} x_i^2\right) a + \left(\sum_{i=1..n} x_i\right) b = \sum_{i=1..n} x_i\, y_i$$

$$\left(\sum_{i=1..n} x_i\right) a + n\, b = \sum_{i=1..n} y_i$$

$a^*, b^* \Leftarrow$ solve system of linear equations

Solution 2: partial derivatives

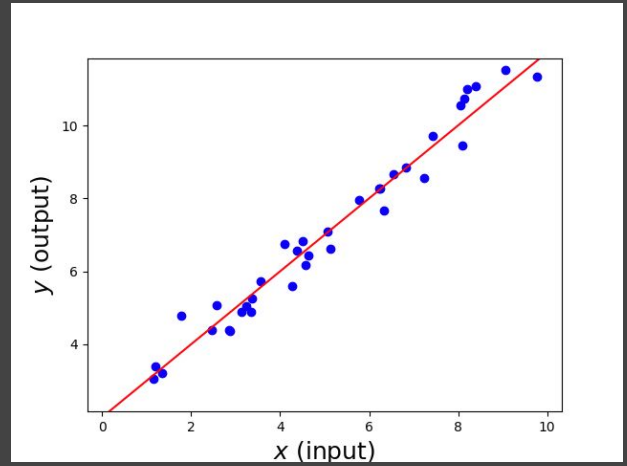
K. N. Toosi
University of Technology



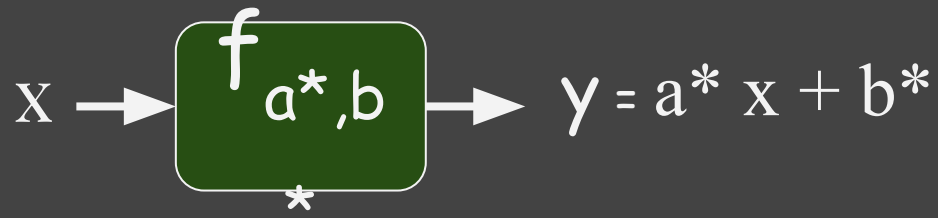
Example: Linear Regression

$a^*, b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a\, x^i + b - y^i)^2$

$a^*, b^* \Leftarrow$ solve system of linear equations



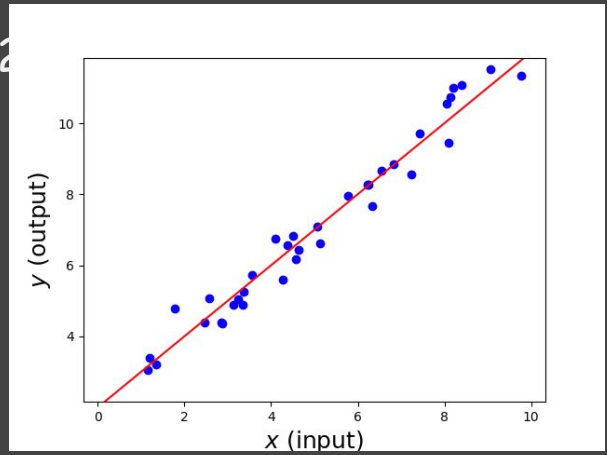
$y = a^* x + b^*$



Evaluation

K. N. Toosi
University of Technology

- Find good parameters θ
  - $\theta^* = \text{argmin}_\theta \sum_{i=1..n} (f(\theta, x_i) - y_i)^2$
  - another method
- How good θ^* is?
- How well the regressor works?

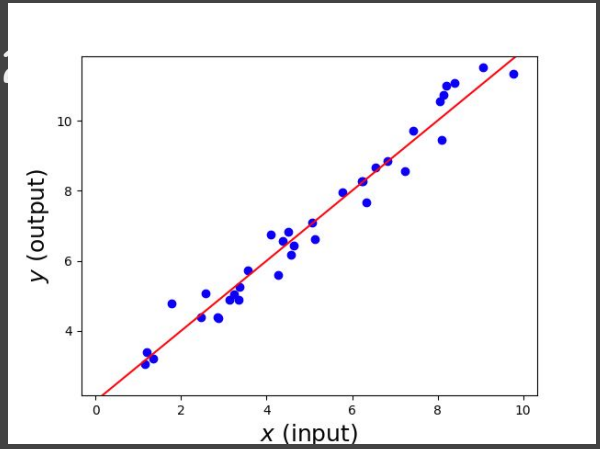


$y = a^* x + b^*$

Evaluation

K. N. Toosi
University of Technology

- Find good parameters θ
  - $\theta^* = \text{argmin}_\theta \sum_{i=1..N} (f(\theta, x_i) - y_i)^2$
  - another method
- How good θ^* is?
- How well the regressor works?
- Given Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$



$y = a^* x + b^*$

$$\text{Error} = C(\theta^*) = \sum_{i=1..N} (f(\theta^*, x_i) - y_i)^2$$

Learning from data

K. N. Toosi
University of Technology

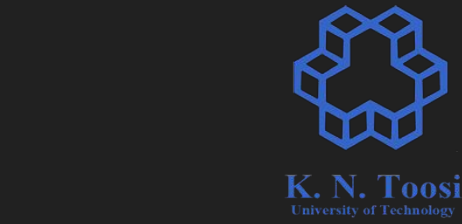
- Parameter Learning:
  - A collection of input-output paris $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$,
  - choose θ such that $y = f(\theta, x)$ is a reasonable output
    - for training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
    - for unseen data

Learning from data

K. N. Toosi
University of Technology

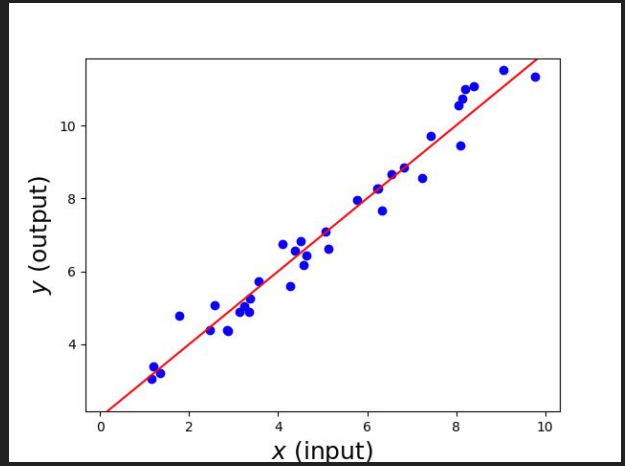
- Parameter Learning:
  - A collection of input-output paris $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$,
  - choose θ such that $y = f(\theta, x)$ is a reasonable output
    - for training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
    - for unseen data
  - Generalization: How well the model works on unseen data

Example: Linear Regression

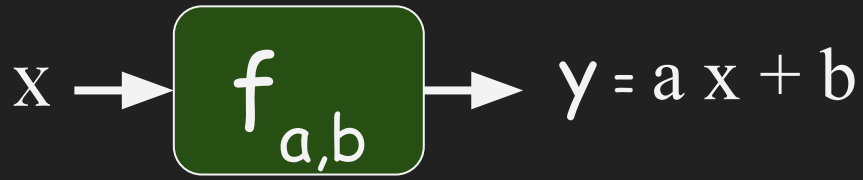


$a^*, b^* = \text{argmin}_{a,b} \Sigma_{i=1..n} (a\, x^i + b - y^i)^2$

$a^*, b^* \Leftarrow$ solve system of linear equations



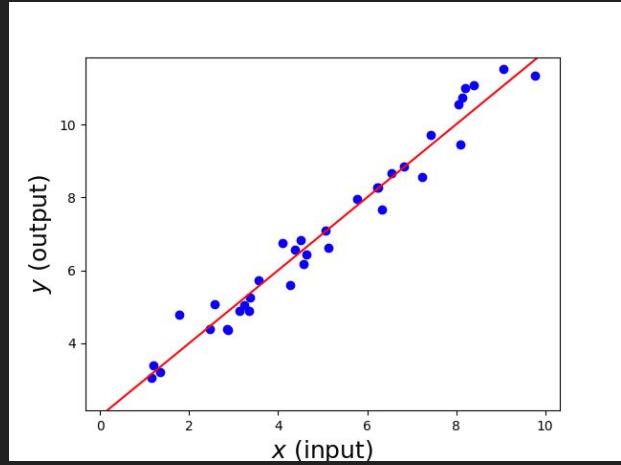
$y = a^* x + b^*$



Evaluation

K. N. Toosi
University of Technology

- Find good parameters θ
  - $\theta^* = \text{argmin}_\theta \Sigma_{i=1..n} (f(\theta, x_i) - y_i)^2$
  - another method
- How good θ^* is?
- How well the regressor works?

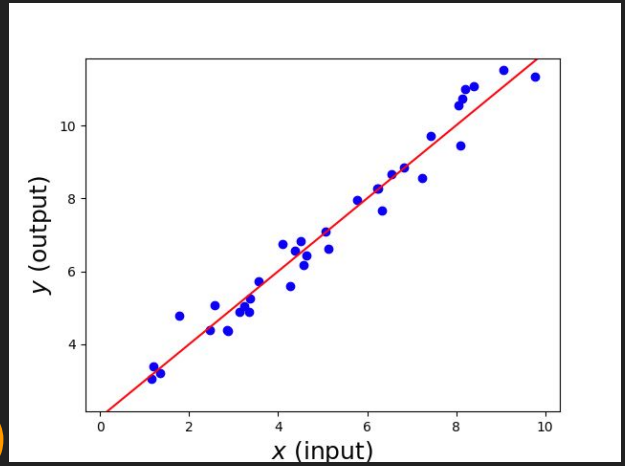


$y = a^* x + b^*$

Evaluation

K. N. Toosi
University of Technology

- Find good parameters θ
  - $\theta^* = \text{argmin}_\theta \Sigma_{i=1..n} (f(\theta, x_i) - y_i)^2$
  - another method
- How good θ^* is?
- How well the regressor works?
- Given Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$



$y = a^* x + b^*$

$$\text{Error} = C(\theta^*) = \Sigma_{i=1..N} (f(\theta^*, x_i) - y_i)^2$$

Learning from data

K. N. Toosi
University of Technology

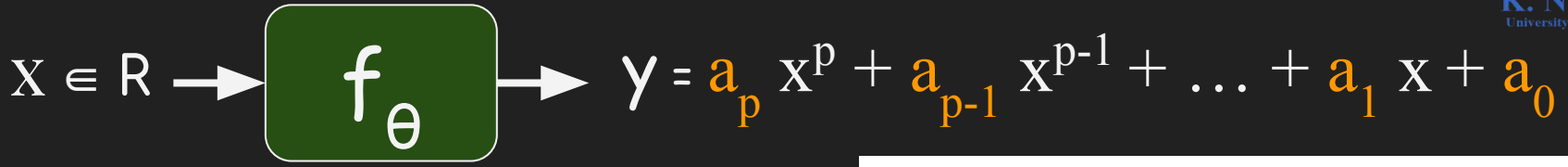
- Parameter Learning:
  - A collection of input-output paris $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$,
  - choose θ such that $y = f(\theta, x)$ is a reasonable output
    - for training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
    - for unseen data

Learning from data

K. N. Toosi
University of Technology

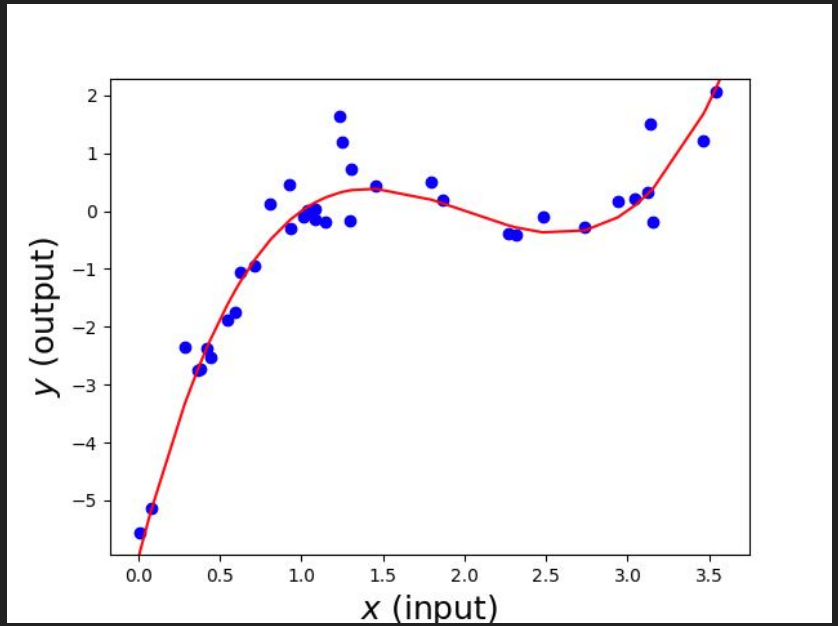
- Parameter Learning:
  - A collection of input-output paris $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$,
  - choose θ such that $y = f(\theta, x)$ is a reasonable output
    - for training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
    - for unseen data
  - Generalization: How well the model works on unseen data

Example: Polynomial Regression

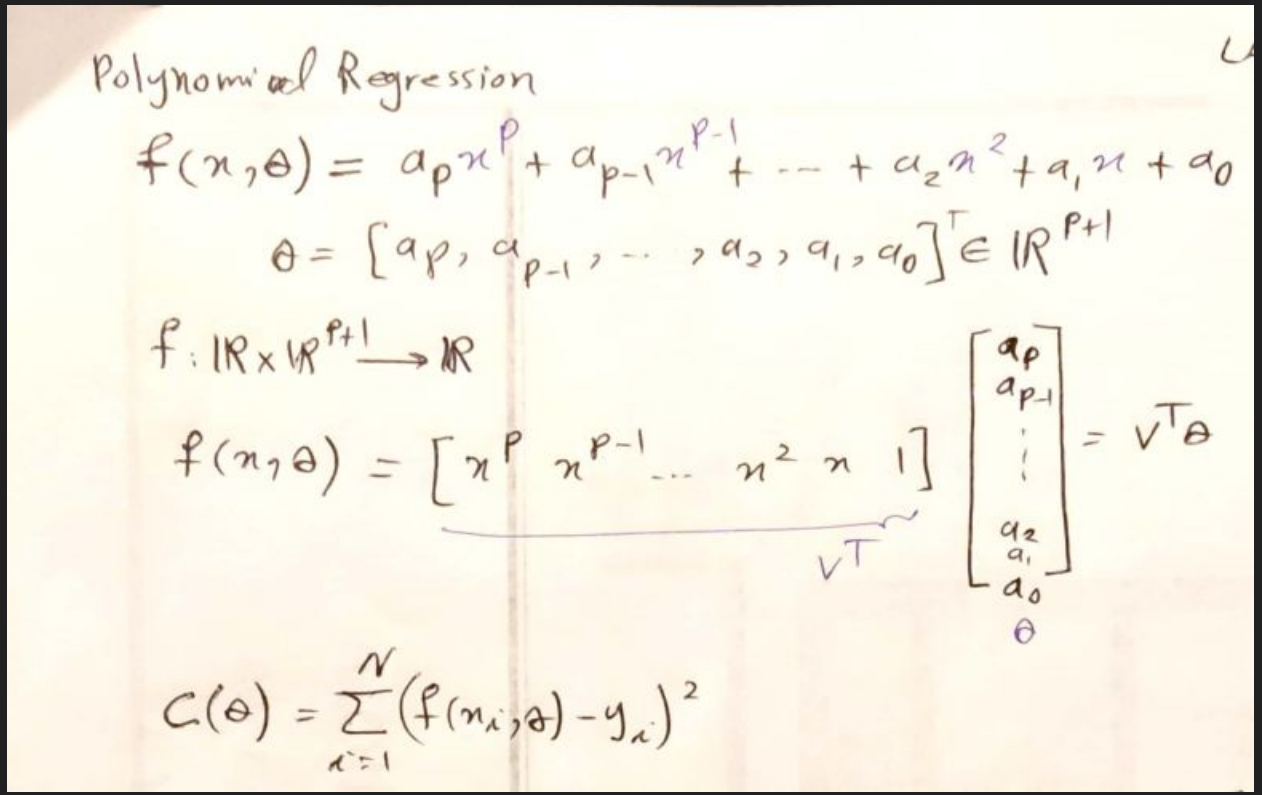


Training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

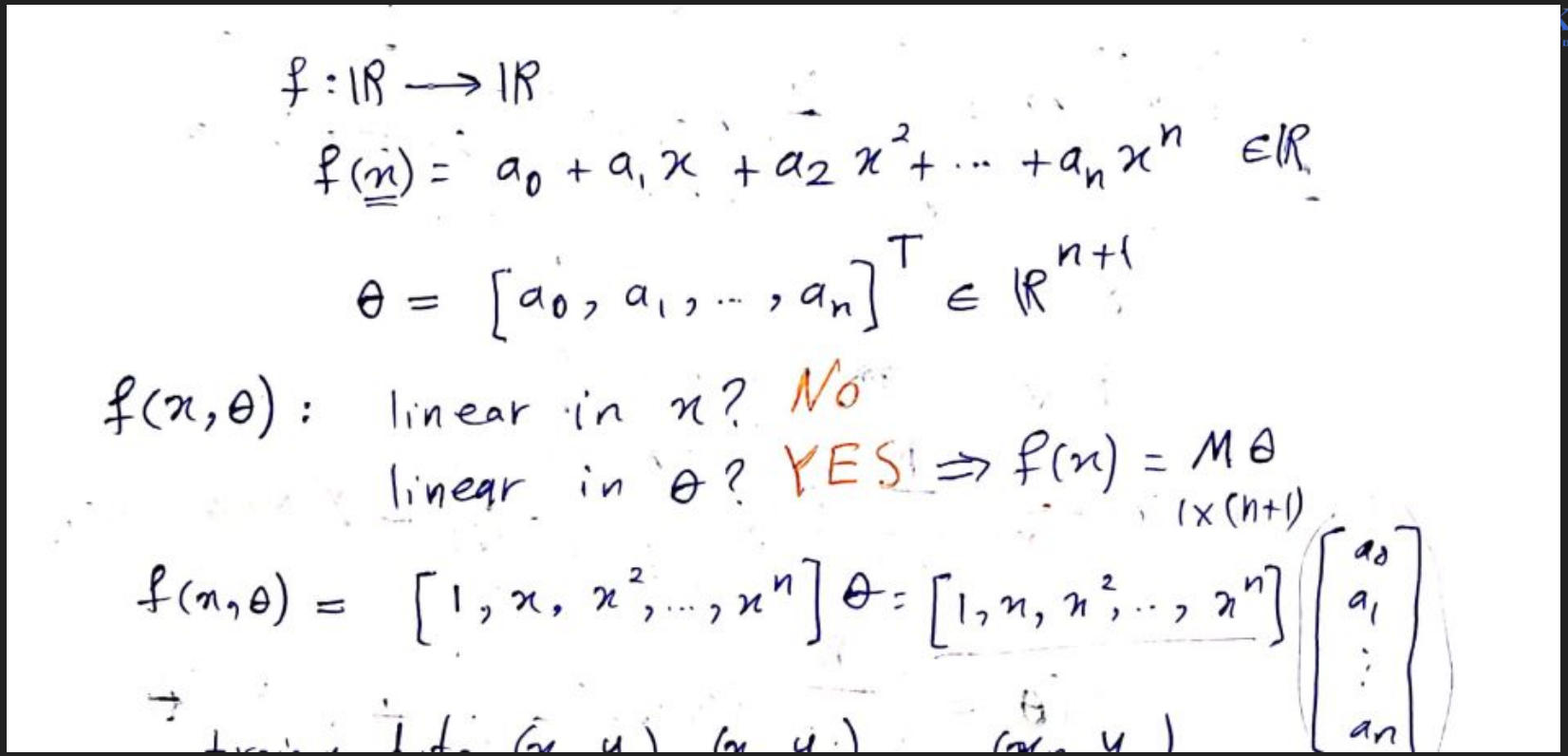
$\theta = (a_p, a_{p-1}, \ldots, a_1, a_0)$



Example: Polynomial Regression



Example: Polynomial Regression



Example: Polynomial Regression

K. N. Toosi
University of Technology

$$f(x,\theta) = [1, x, x^2, \dots, x^n]\,\theta = [1, x, x^2, \dots, x^n] \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

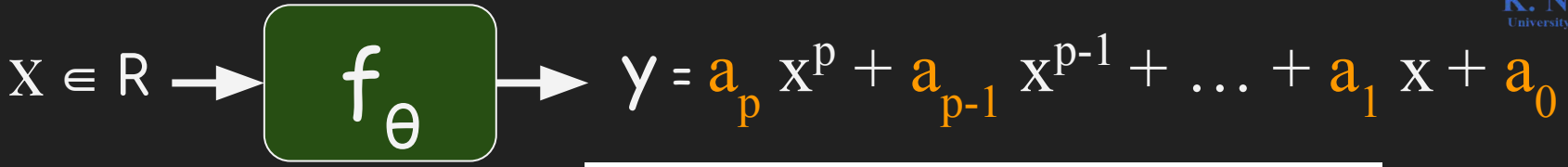
→ training data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

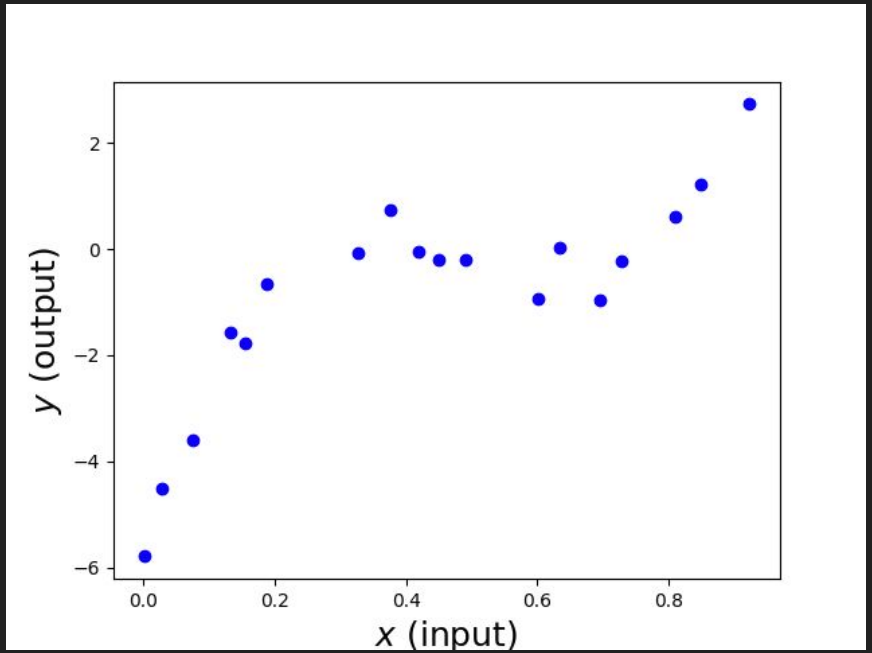
$$C(\theta) = \sum_{i=1}^{N} \left(f(\theta, x_i) - y_i \right)^2 = \left\| \begin{bmatrix} 1, x_1, x_1^2, \dots, x_1^n \\ 1, x_2, x_2^2, \dots, x_2^n \\ \vdots \\ 1, x_N, x_N^2, \dots, x_N^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \right\|^2$$

$$\| A\theta - Y \|^2$$

$$\theta^* = (A^T A)^{-1} A^T Y$$

Vary polynomial degree p





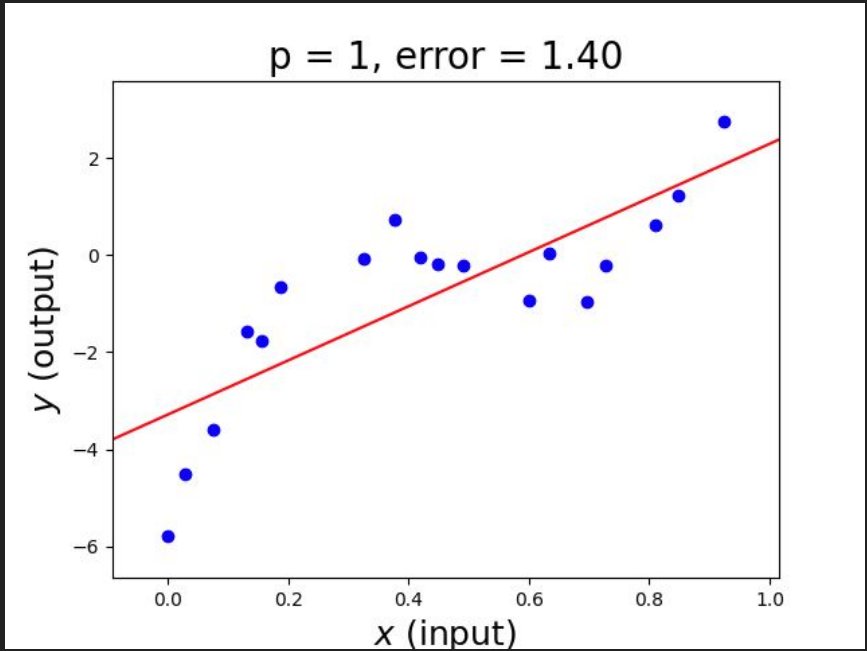
K. N. Toosi
University of Technology

polynomial degree p = 1

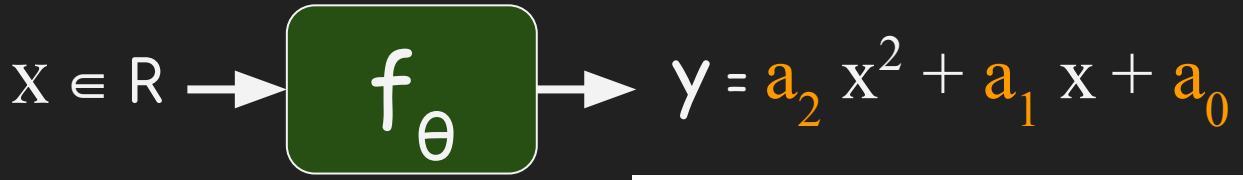
K. N. Toosi
University of Technology

$X \in R \rightarrow f_\theta \rightarrow y = a_1 x + a_0$

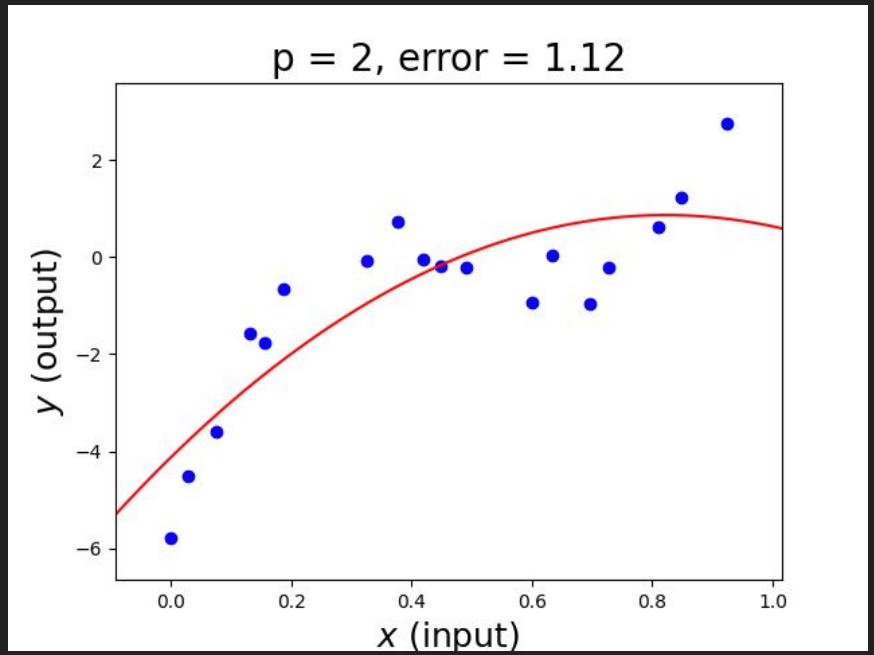
$$\text{Error} = \Sigma_{i=1..n} \left(f(\theta^*, x_i) - y_i \right)^2$$



polynomial degree p = 2



$$\text{Error} = \Sigma_{i=1..n} \left(f(\theta^*, x_i) - y_i \right)^2$$



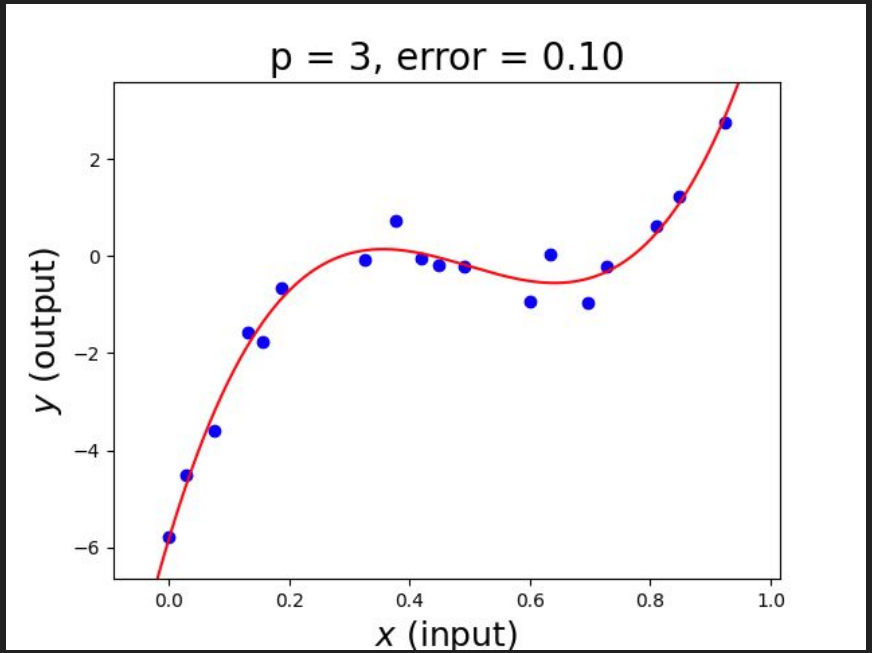
K. N. Toosi
University of Technology

polynomial degree p = 3

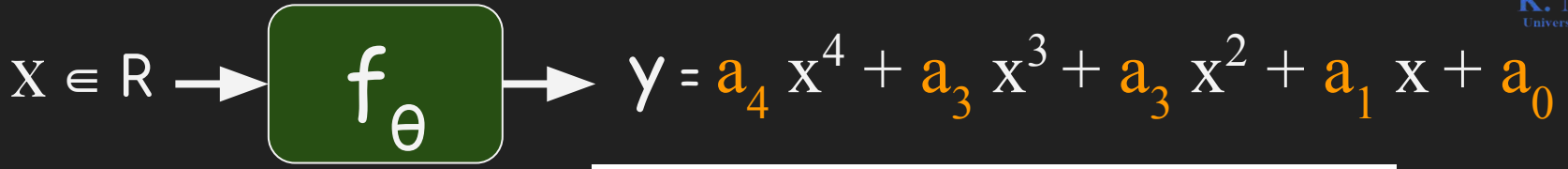
K. N. Toosi
University of Technology

$$X \in R \rightarrow \boxed{f_\theta} \rightarrow y = a_p x^3 + a_{p-1} x^3 + a_1 x + a_0$$

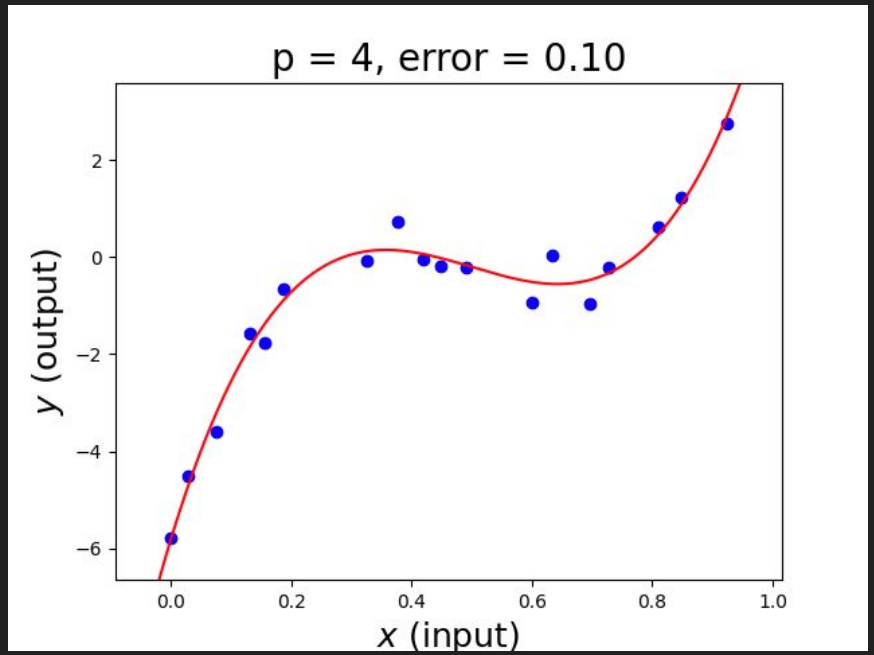
$$\text{Error} = \Sigma_{i=1..n} \left(f(\theta^*, x_i) - y_i \right)^2$$



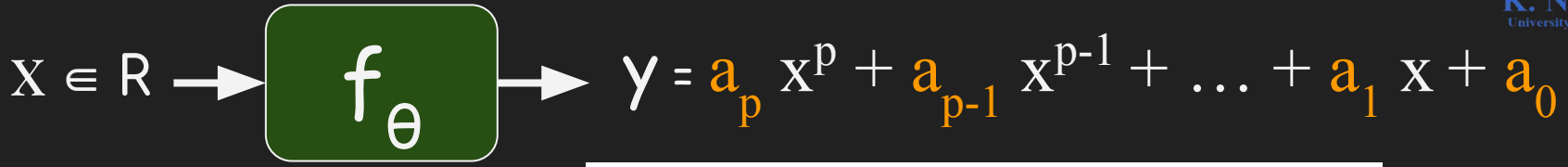
polynomial degree p = 4



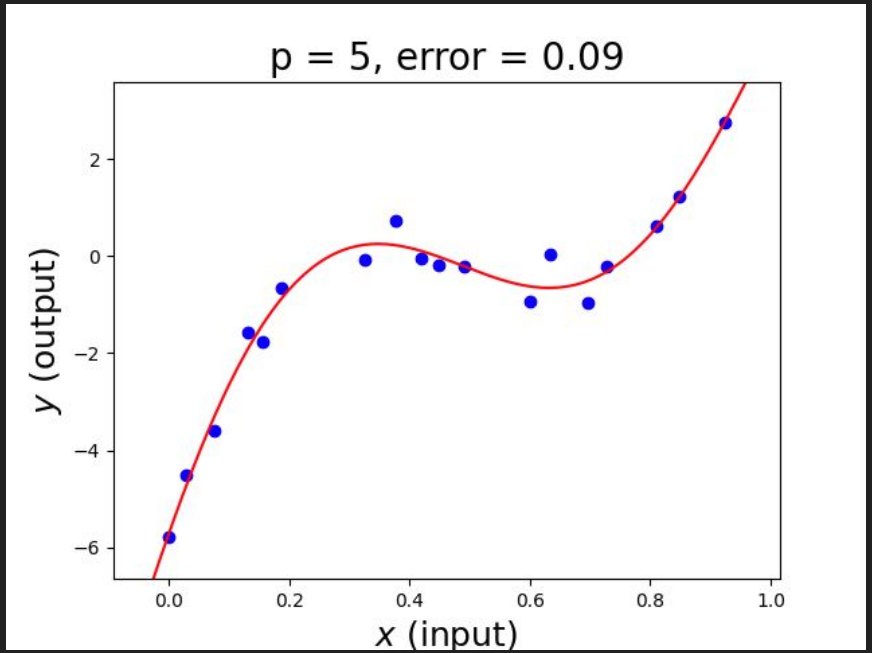
$$\text{Error} = \Sigma_{i=1..n} \left(f(\theta^*, x_i) - y_i \right)^2$$



polynomial degree p = 5



$$\text{Error} = \Sigma_{i=1..n} \left(f(\theta^*, x_i) - y_i \right)^2$$



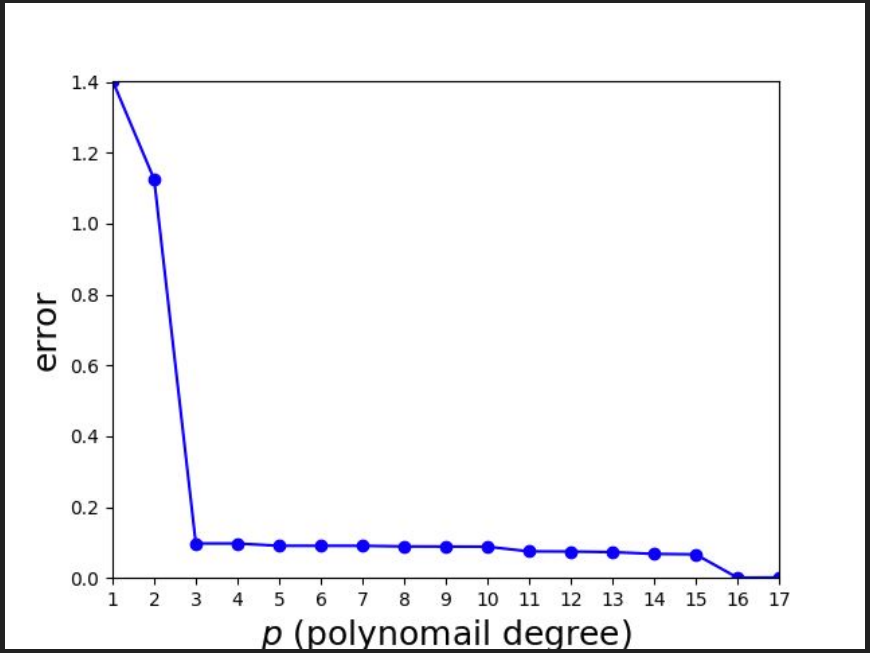
K. N. Toosi
University of Technology

polynomial degree p = 5

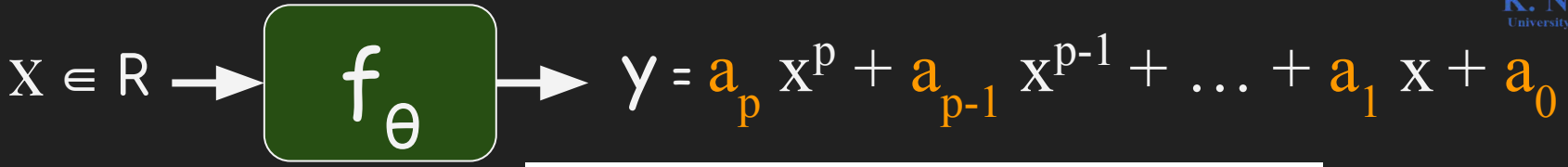
K. N. Toosi
University of Technology

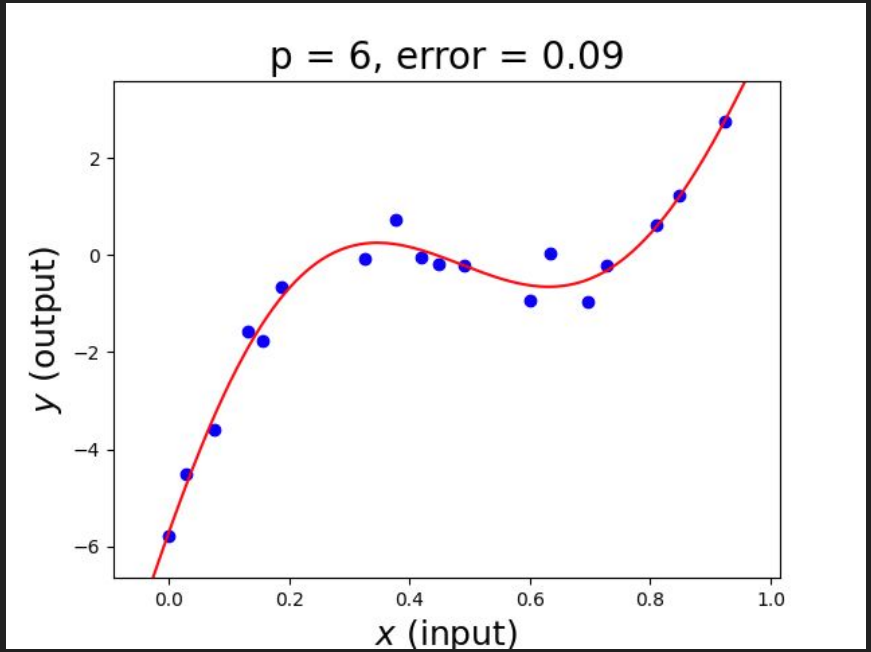
$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$

$$\text{Error} = \Sigma_{i=1..n} \left(f(\theta^*, x_i) - y_i \right)^2$$



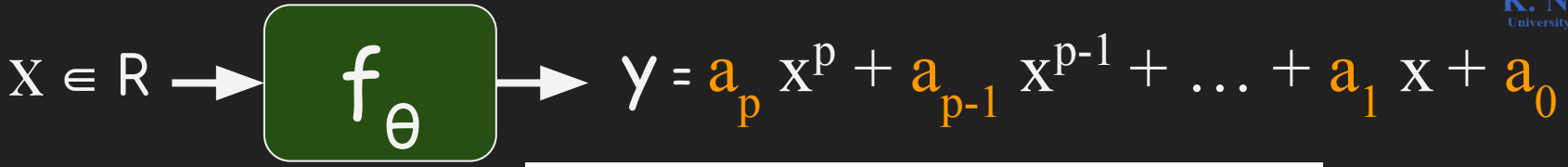
polynomial degree p = 6

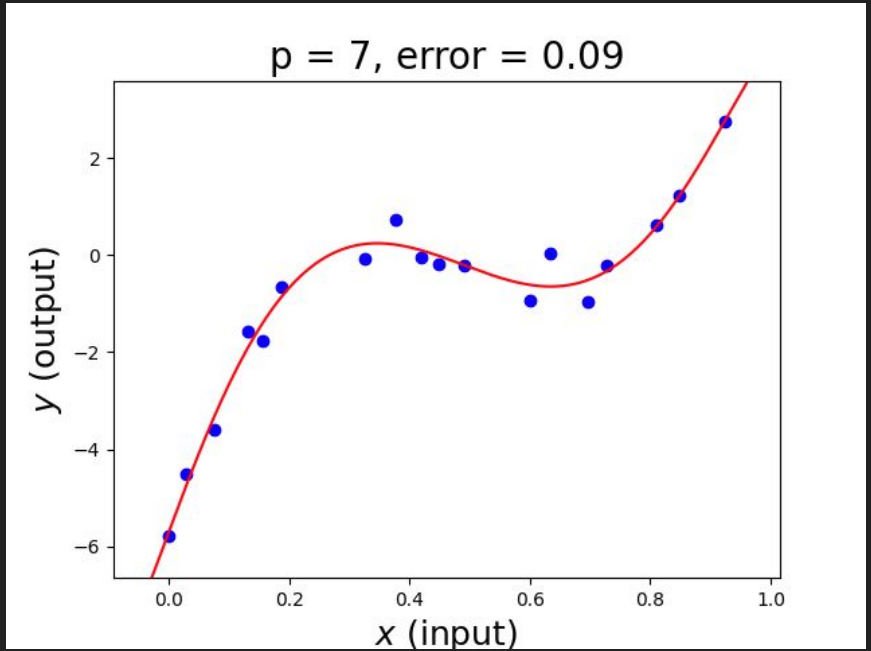




K. N. Toosi
University of Technology

polynomial degree p = 7



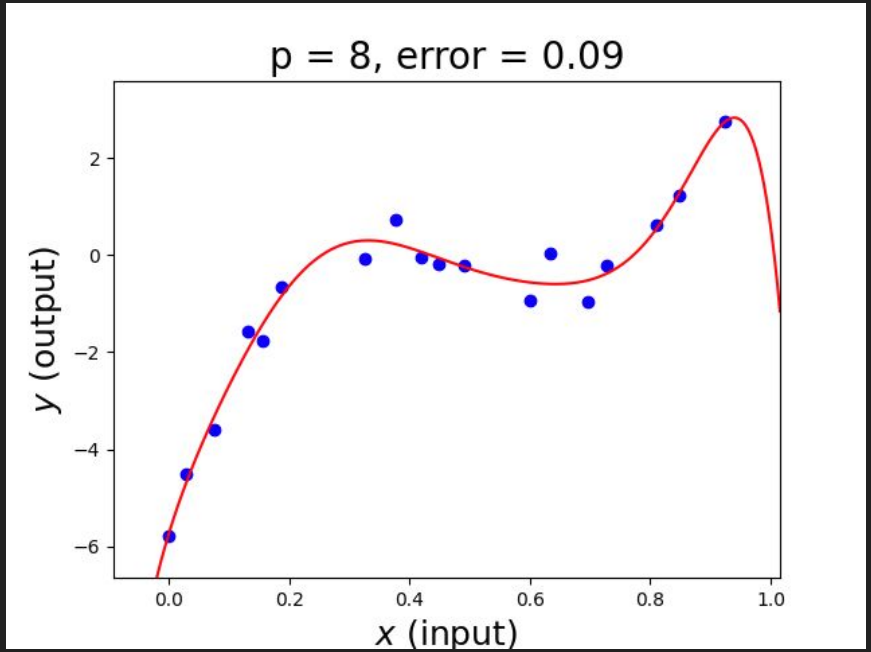


K. N. Toosi
University of Technology

polynomial degree p = 8

K. N. Toosi
University of Technology

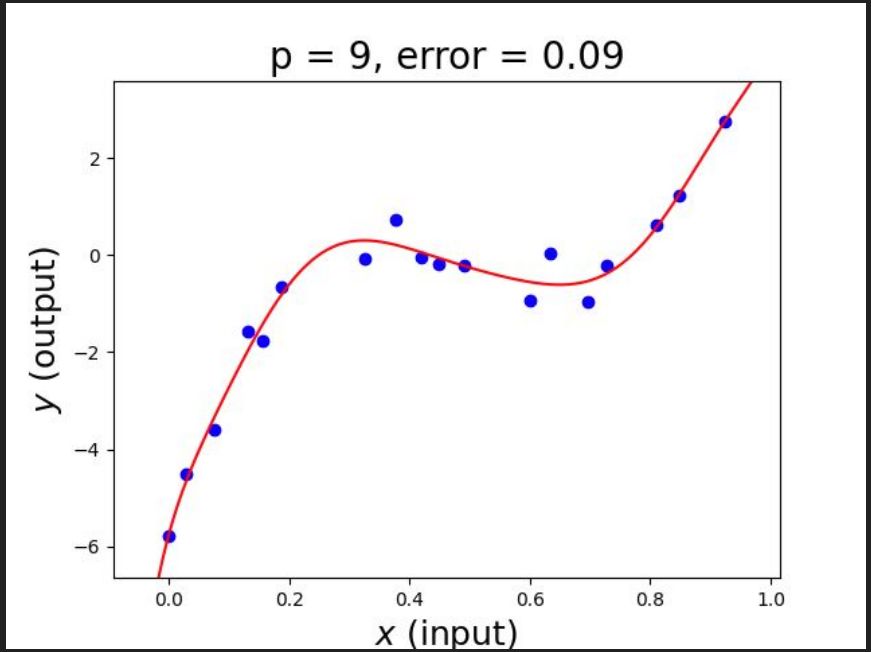
$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



polynomial degree p = 9

K. N. Toosi
University of Technology

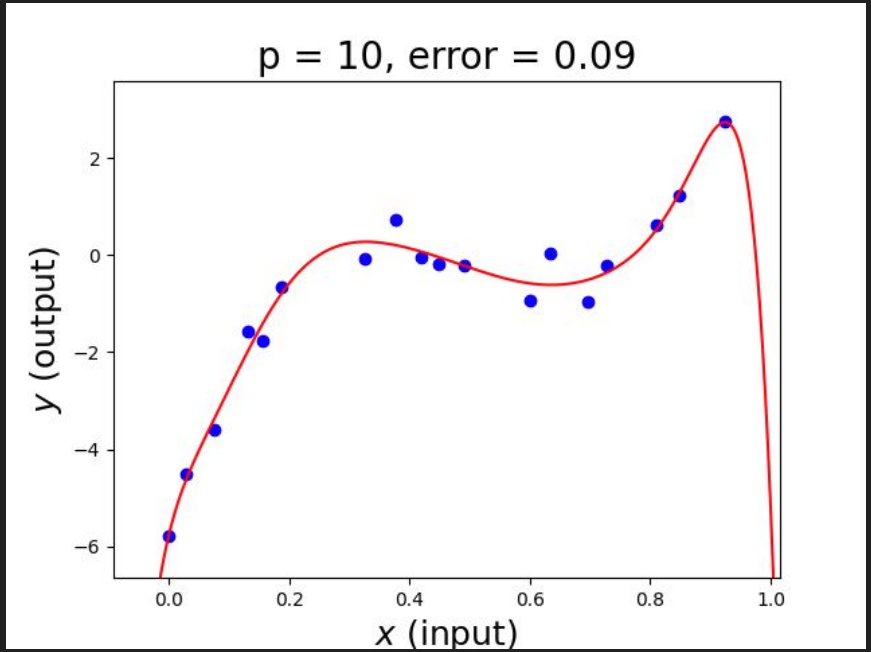
$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



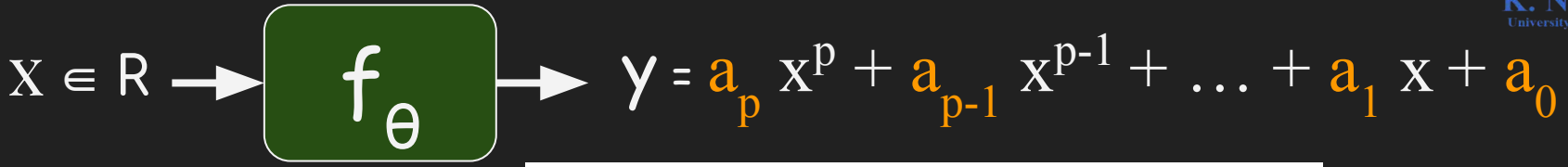
polynomial degree p = 10

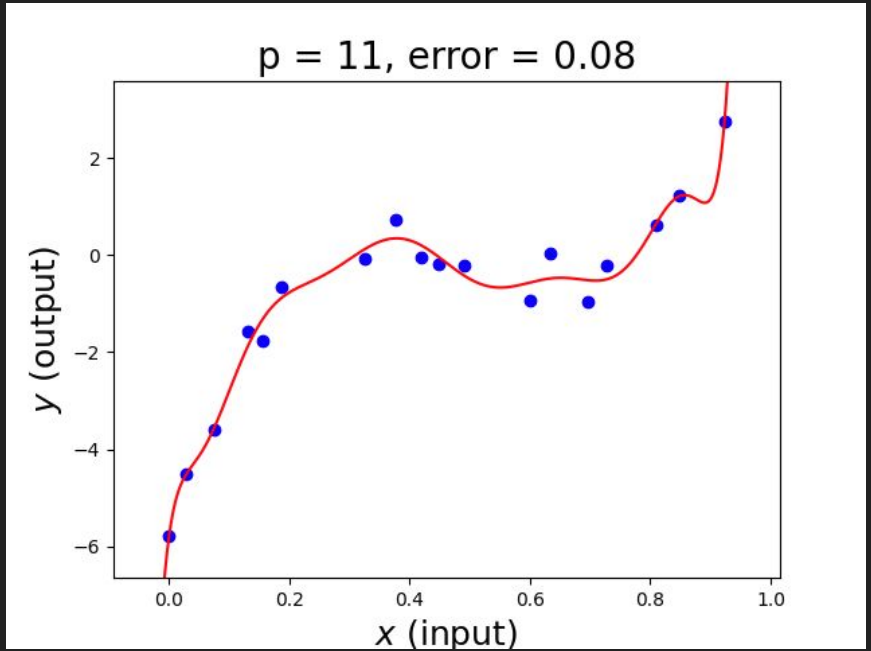
K. N. Toosi
University of Technology

$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



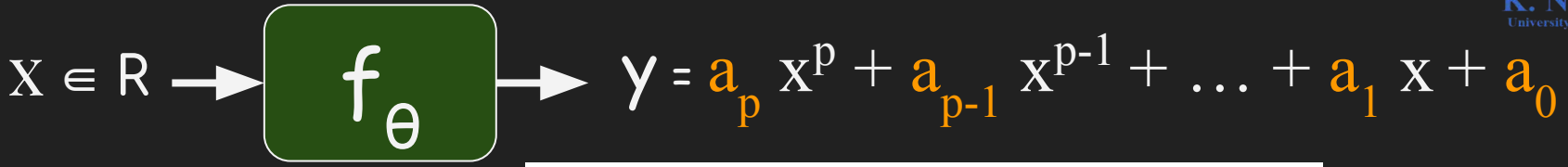
polynomial degree p = 11

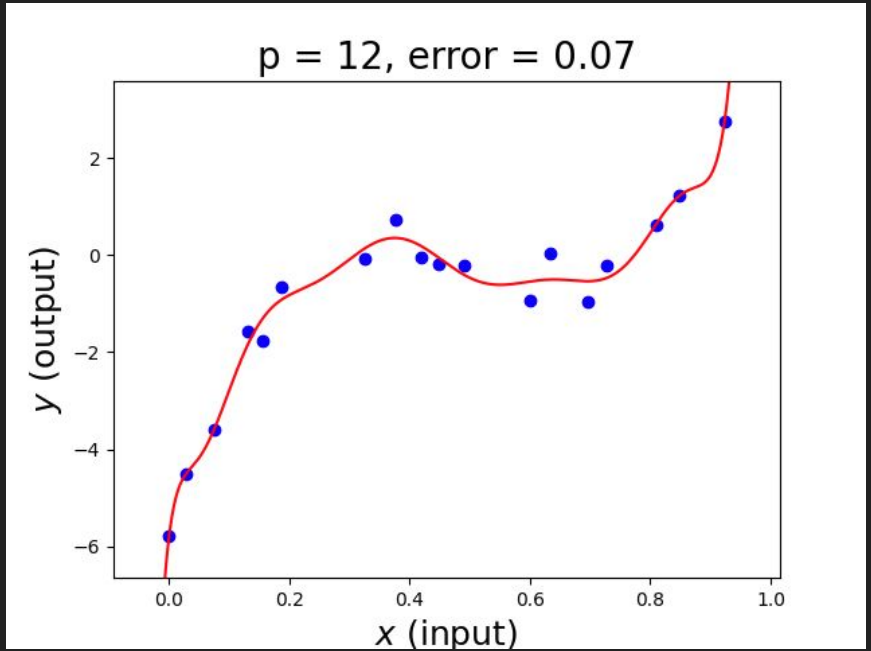




K. N. Toosi
University of Technology

polynomial degree p = 12

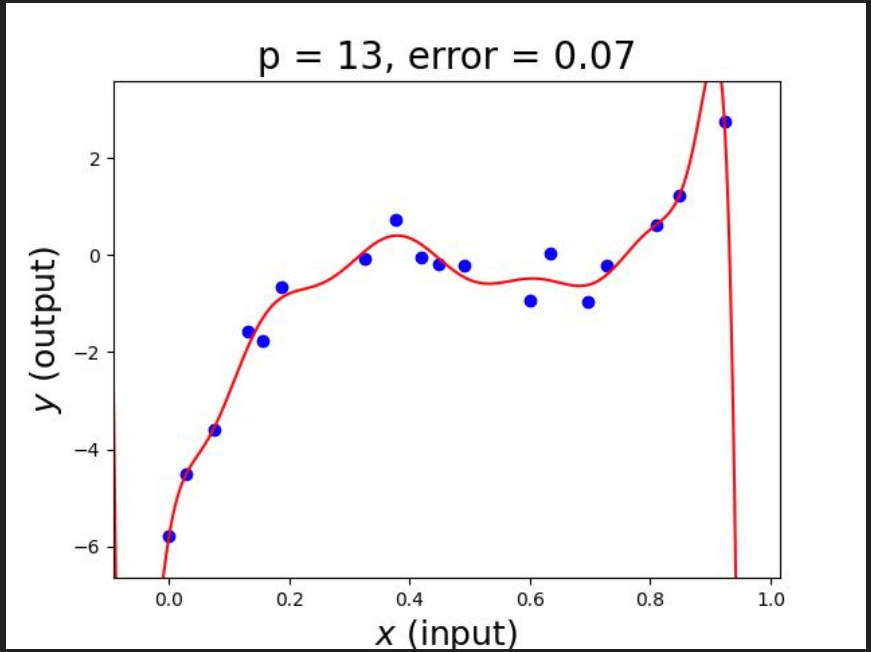




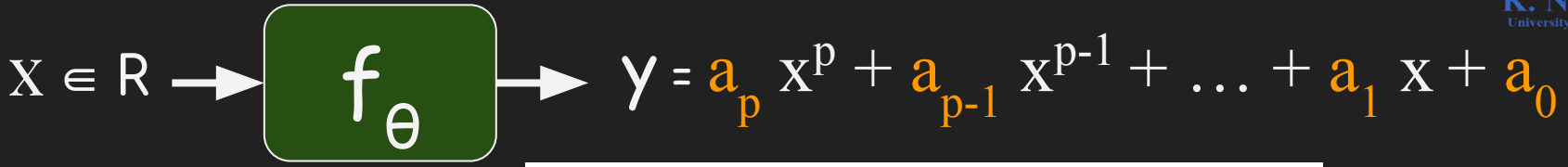
polynomial degree p = 13

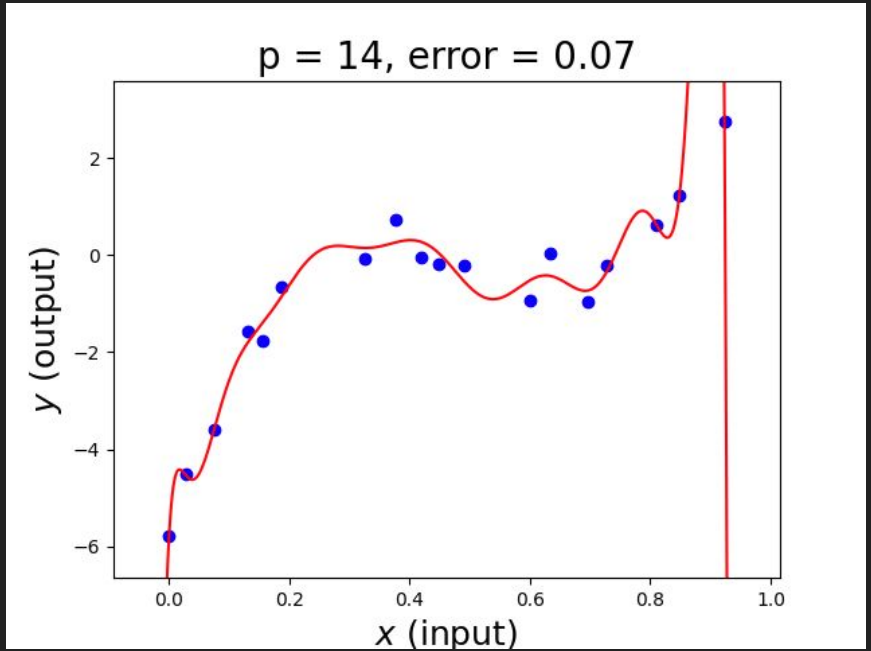
K. N. Toosi
University of Technology

$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



polynomial degree p = 14



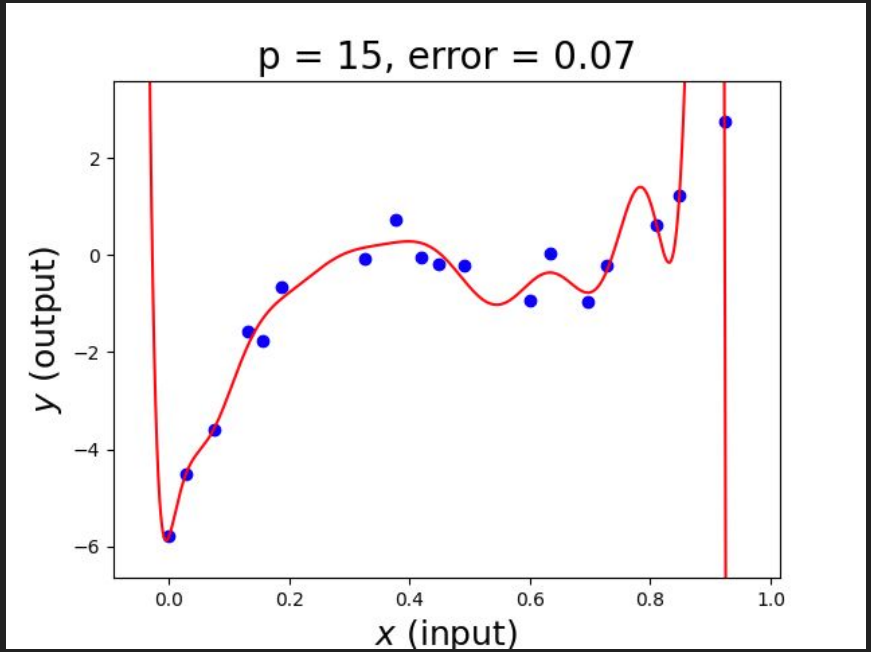


K. N. Toosi
University of Technology

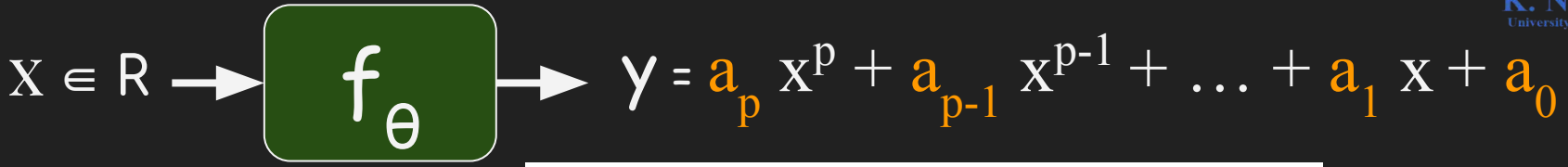
polynomial degree p = 15

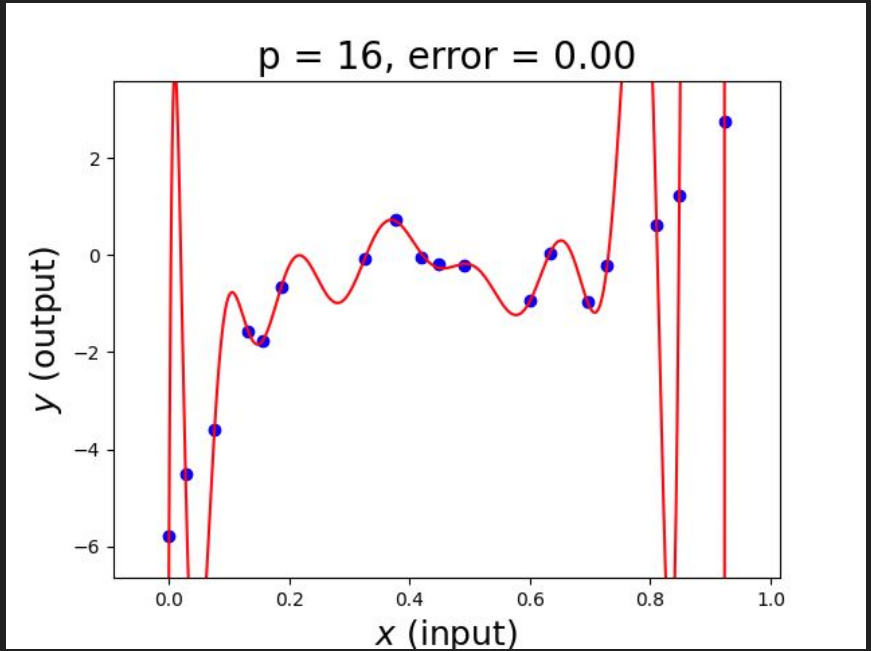
K. N. Toosi
University of Technology

$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



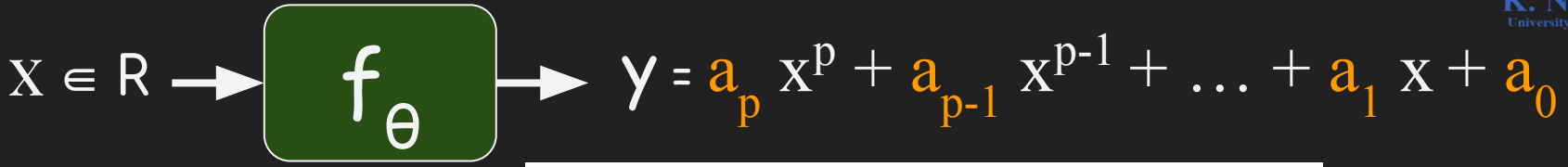
polynomial degree p = 16

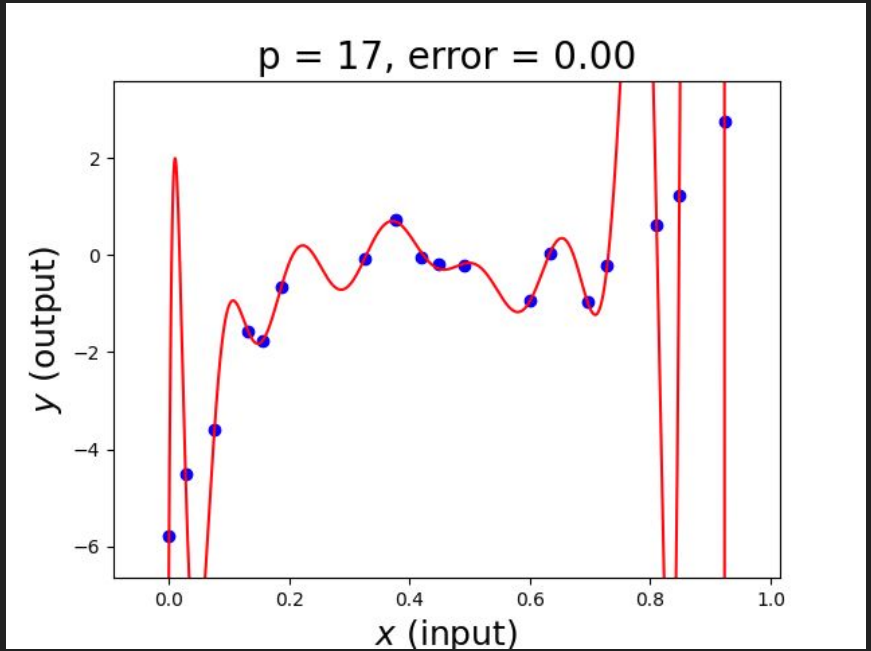




K. N. Toosi
University of Technology

polynomial degree p = 17



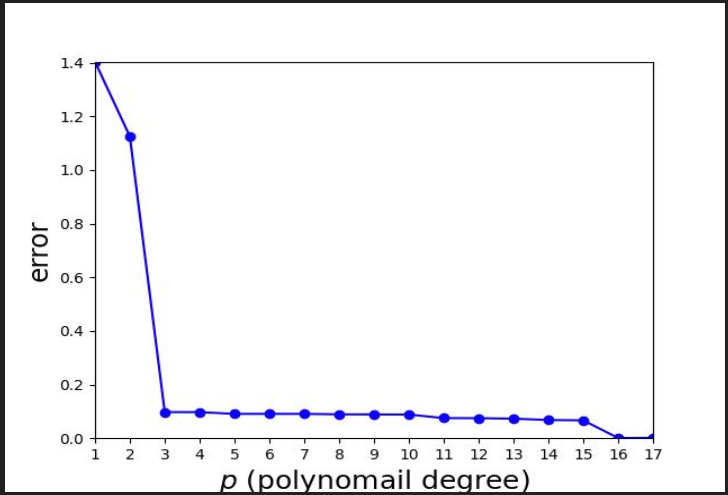


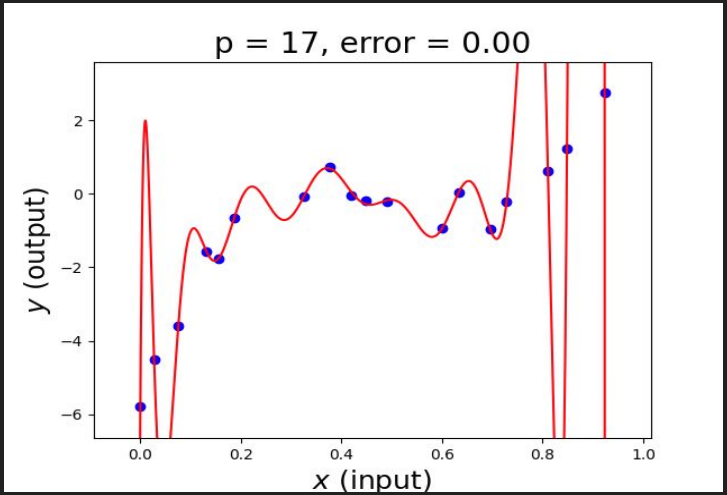
K. N. Toosi
University of Technology

polynomial degree p = 17

K. N. Toosi
University of Technology

$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



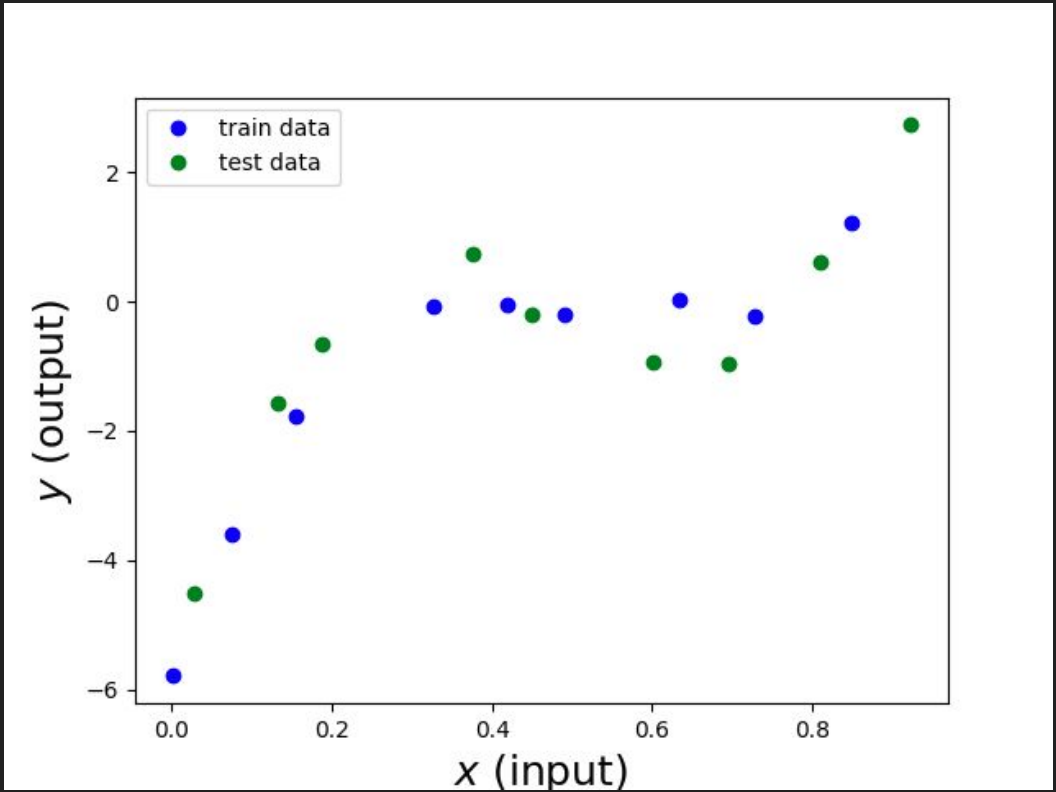


Learning from data

K. N. Toosi
University of Technology

- Parameter Learning:
  - A collection of input-output paris $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$,
  - choose θ such that $y = f(\theta, x)$ is a reasonable output
    - for training data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$
    - for unseen data
  - Generalization: How well the model work on unseen data
  - How to evaluate error on unseen data?

Train-Test split



Train-Test split

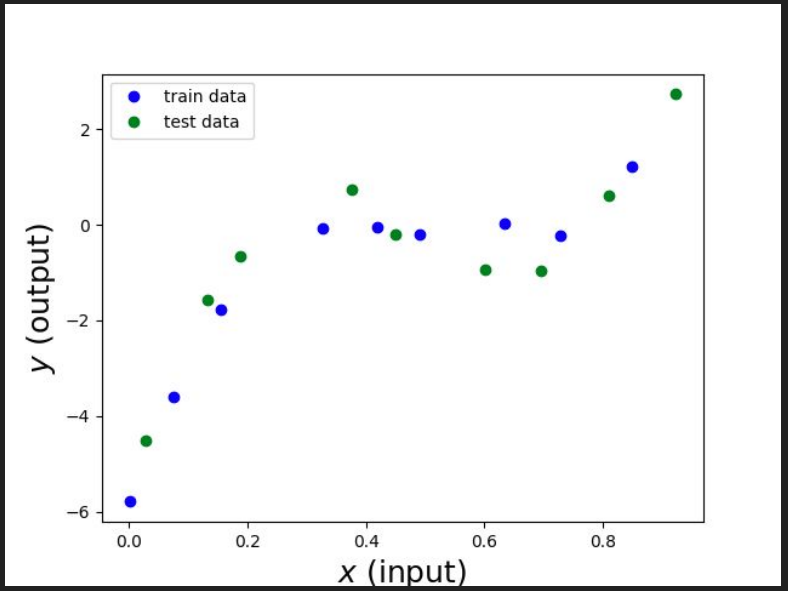
K. N. Toosi
University of Technology

- Divide data into train and test sets
- Train the model using train data
- Evaluate on test data

$$D = D_{test} \cup D_{train}$$

$$\text{Train Error} = \sum_{(x_i, y_i) \in D_{train}} \left(f(x_i, \theta) - y_i\right)^2$$

$$\text{Test Error} = \sum_{(x_i, y_i) \in D_{test}} \left(f(x_i, \theta) - y_i\right)^2$$



Train-Test split

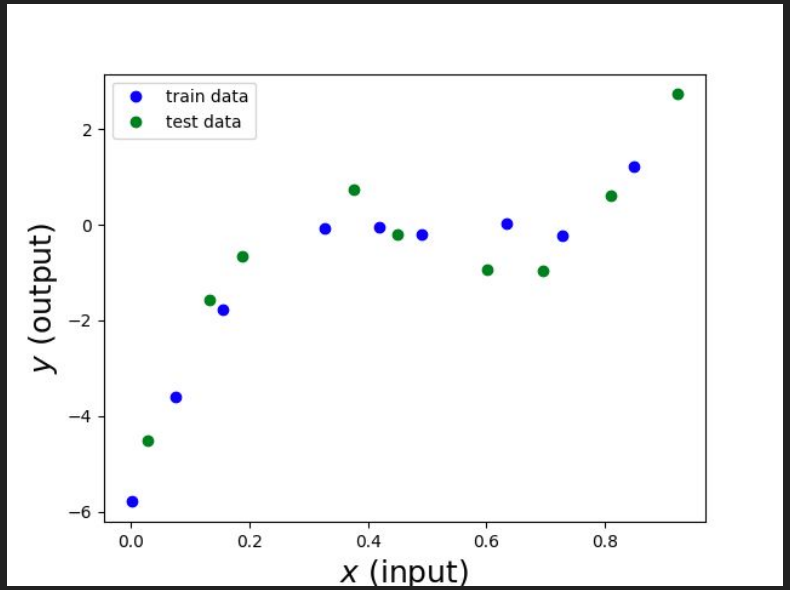
K. N. Toosi
University of Technology

- Divide data into train and test sets
- Train the model using train data
- Evaluate on test data

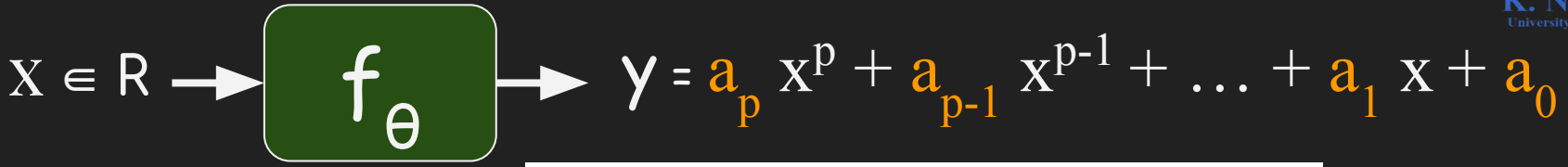
$$\theta^* = \text{argmin}_\theta \sum_{i \in \text{train}} (f(\theta, x_i) - y_i)^2$$

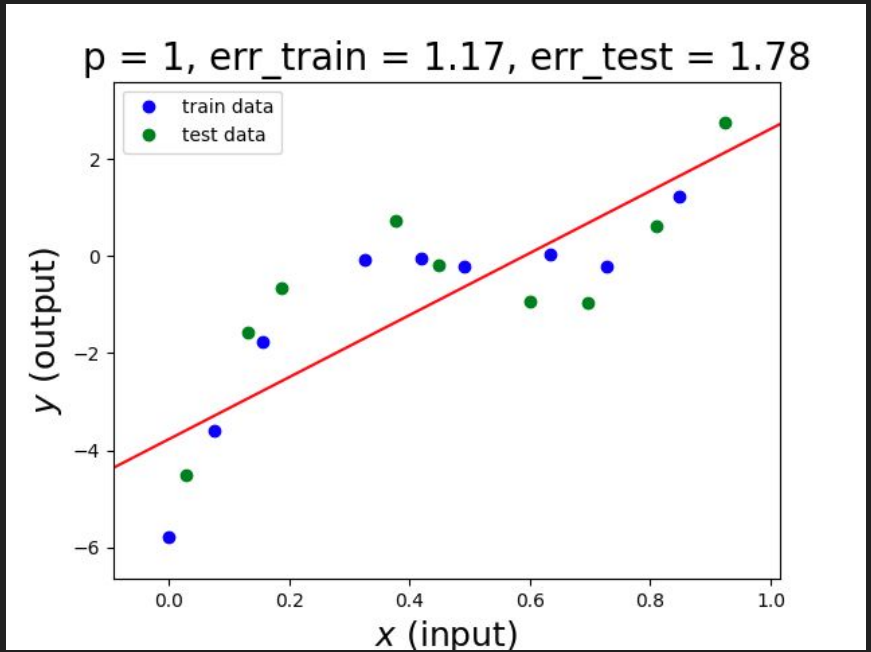
$$\text{Err}_{\text{train}} = \sum_{i \in \text{train}} (f(\theta^*, x_i) - y_i)^2$$

$$\text{Err}_{\text{test}} = \sum_{i \in \text{test}} (f(\theta^*, x_i) - y_i)^2$$



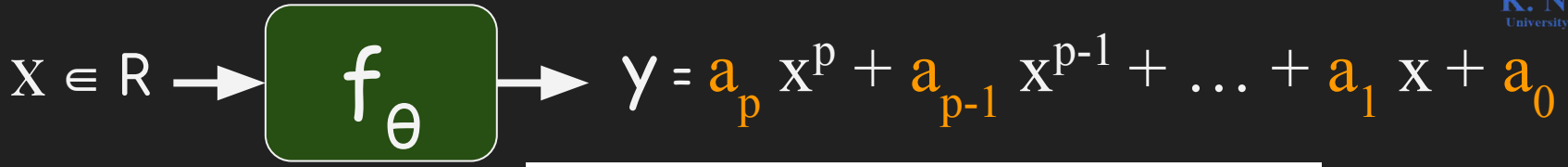
polynomial degree p = 1

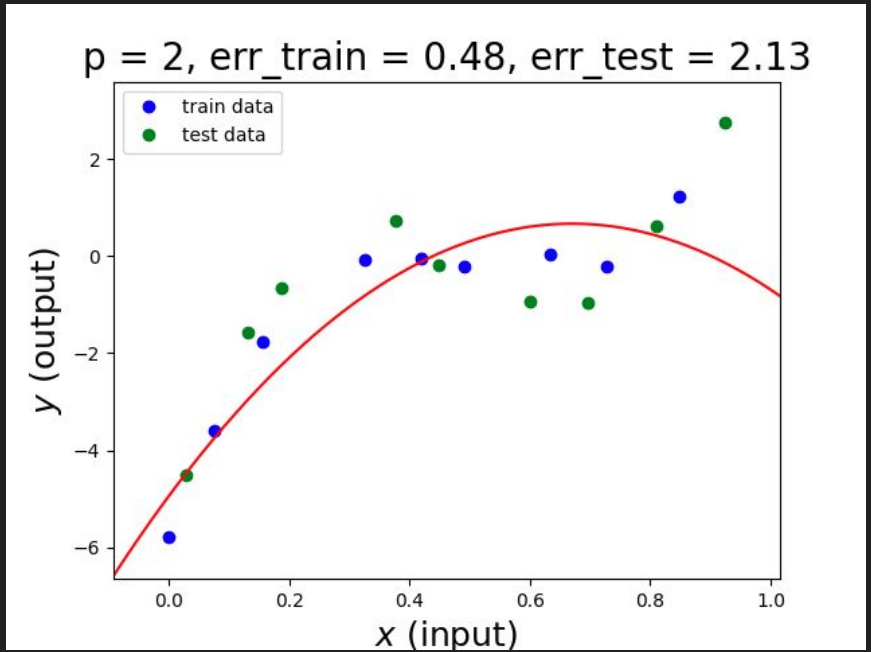




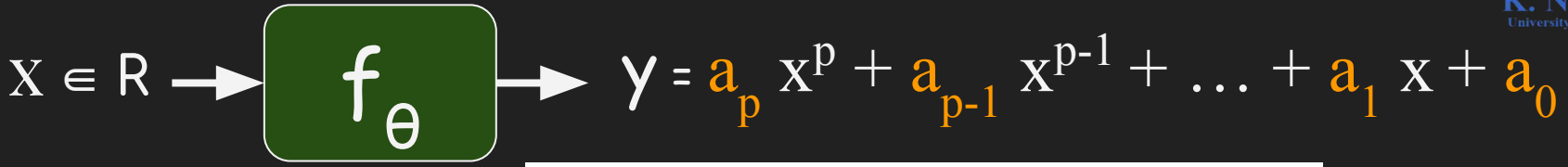
K. N. Toosi
University of Technology

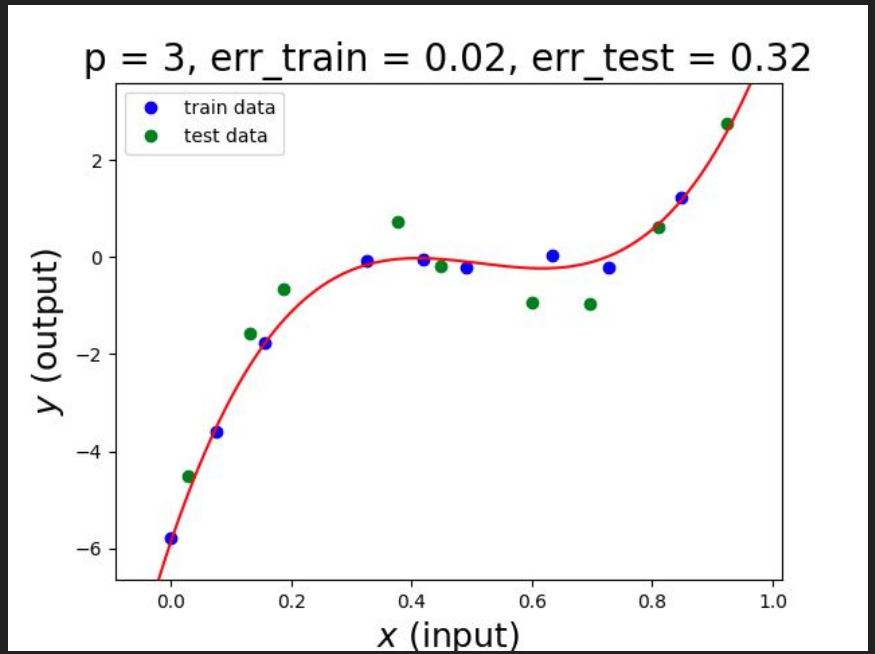
polynomial degree p = 2





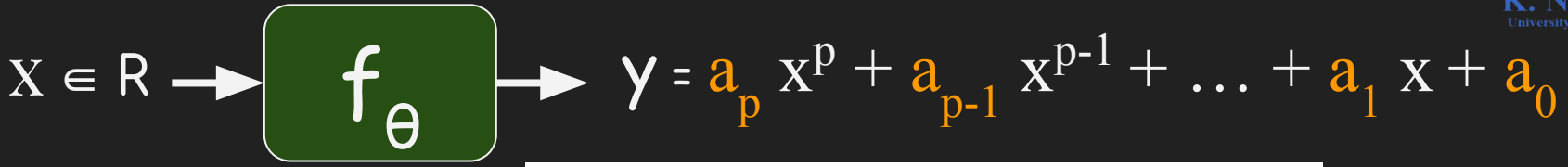
polynomial degree p = 3

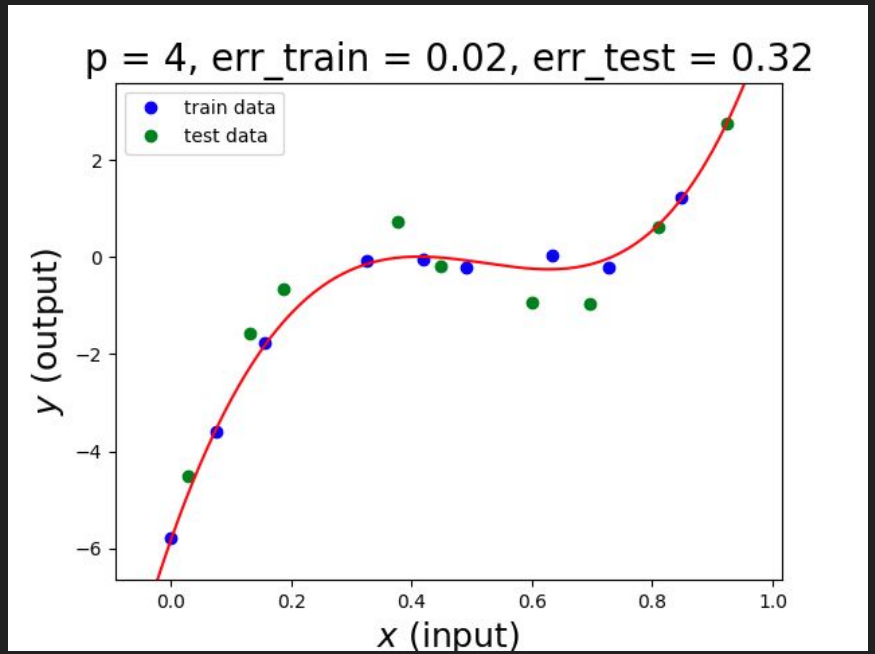




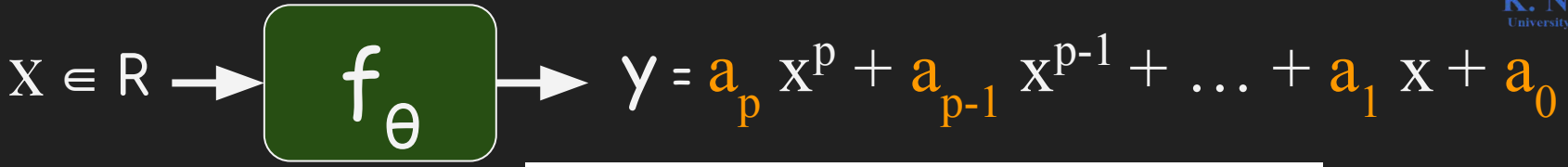
K. N. Toosi
University of Technology

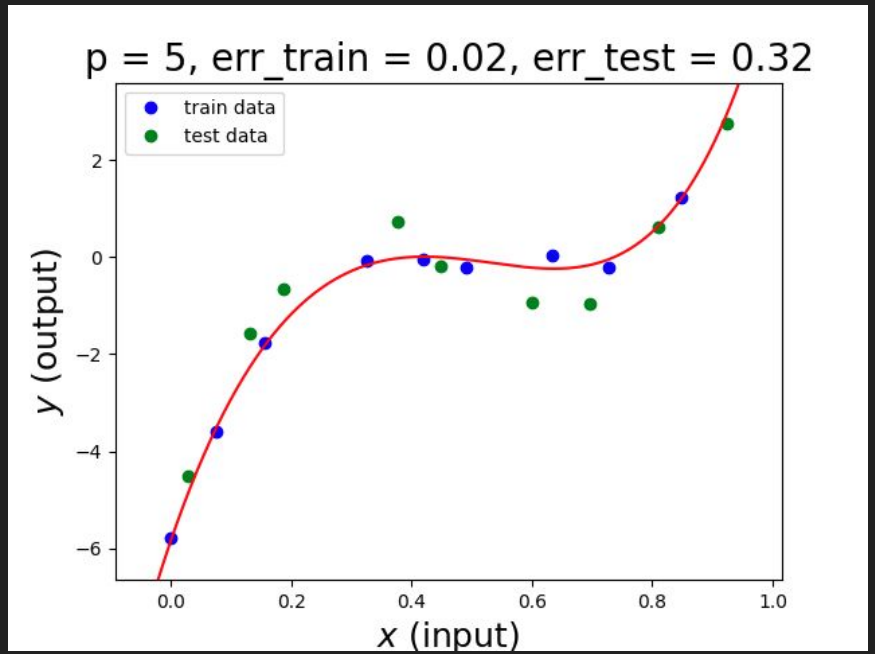
polynomial degree p = 4





polynomial degree p = 5



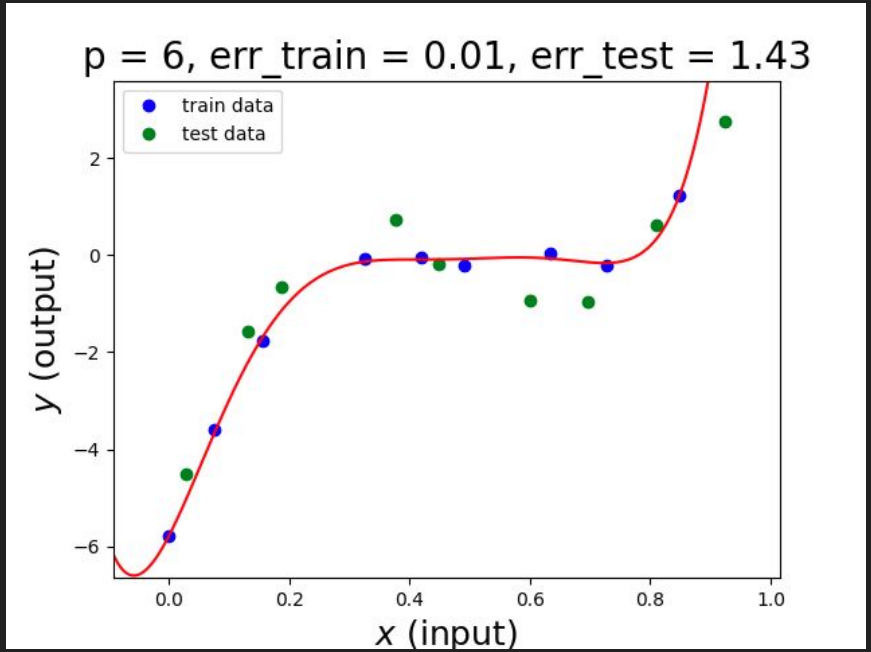


K. N. Toosi
University of Technology

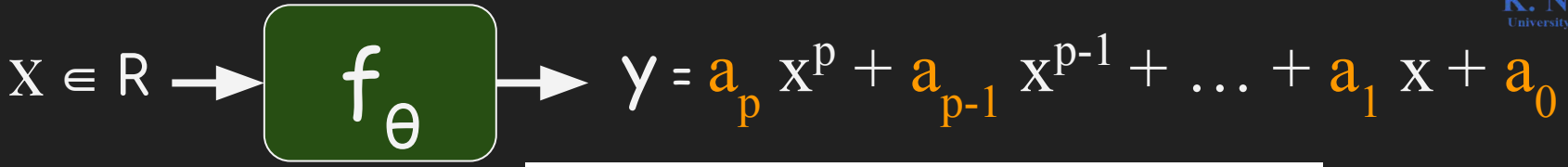
polynomial degree p = 6

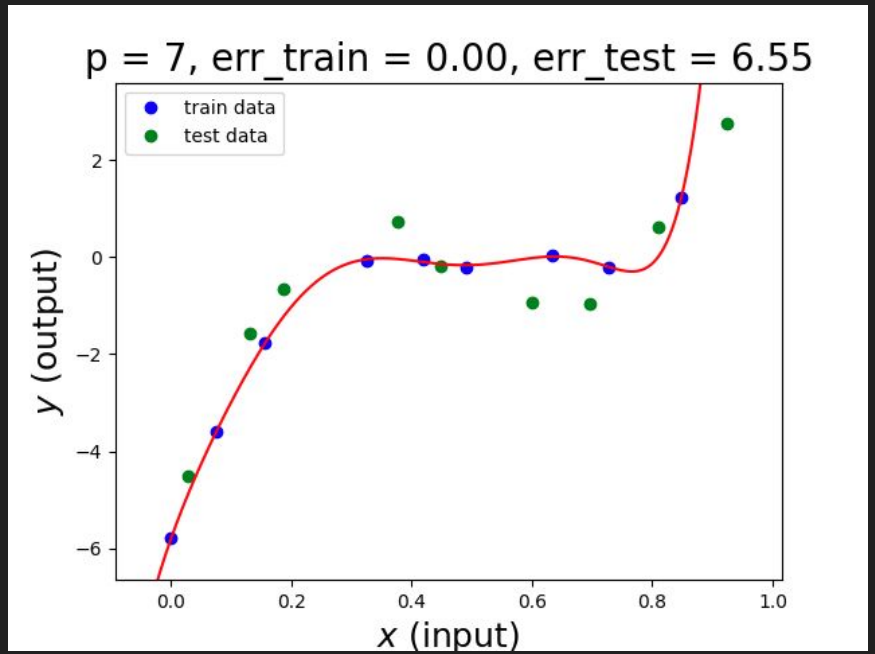
K. N. Toosi
University of Technology

$$x \in R \rightarrow \boxed{f_\theta} \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



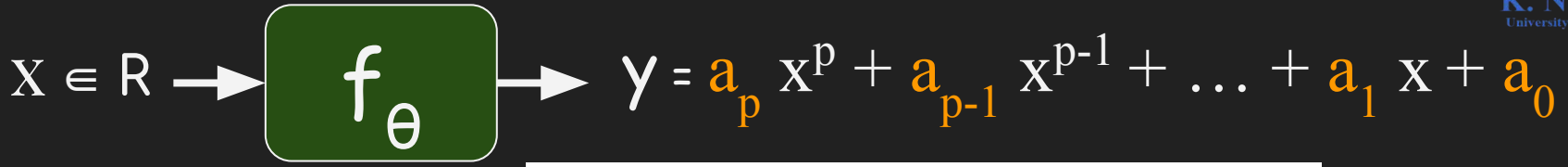
polynomial degree p = 7

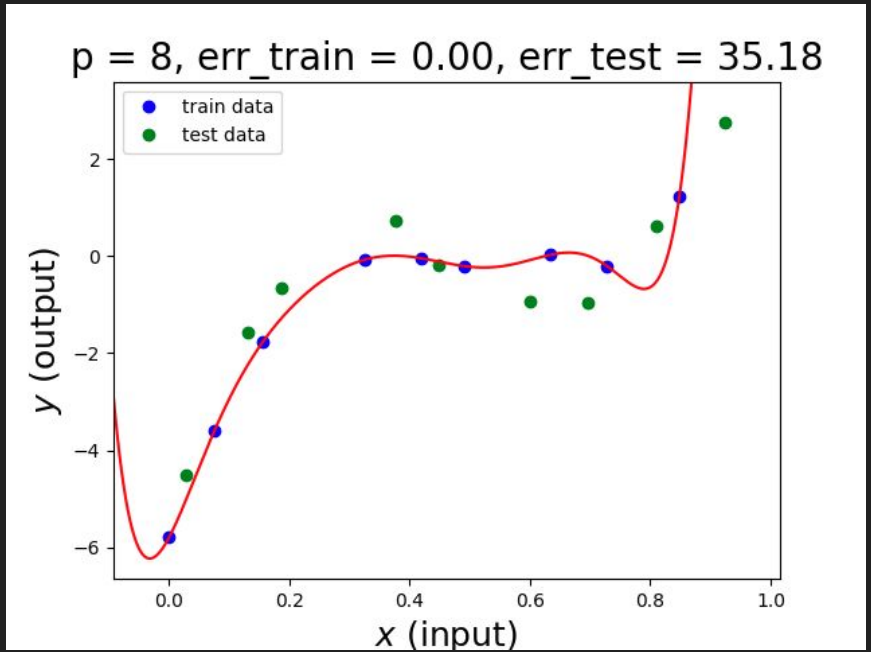




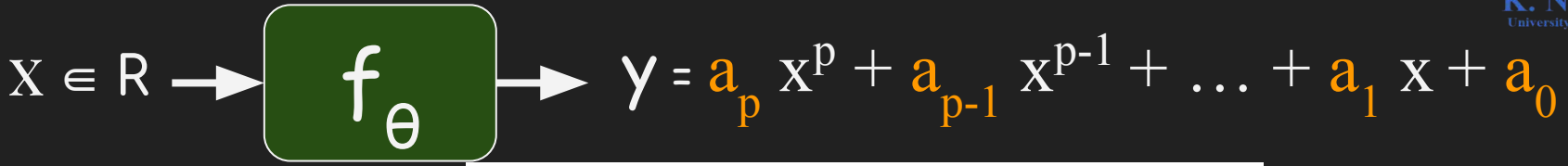
K. N. Toosi
University of Technology

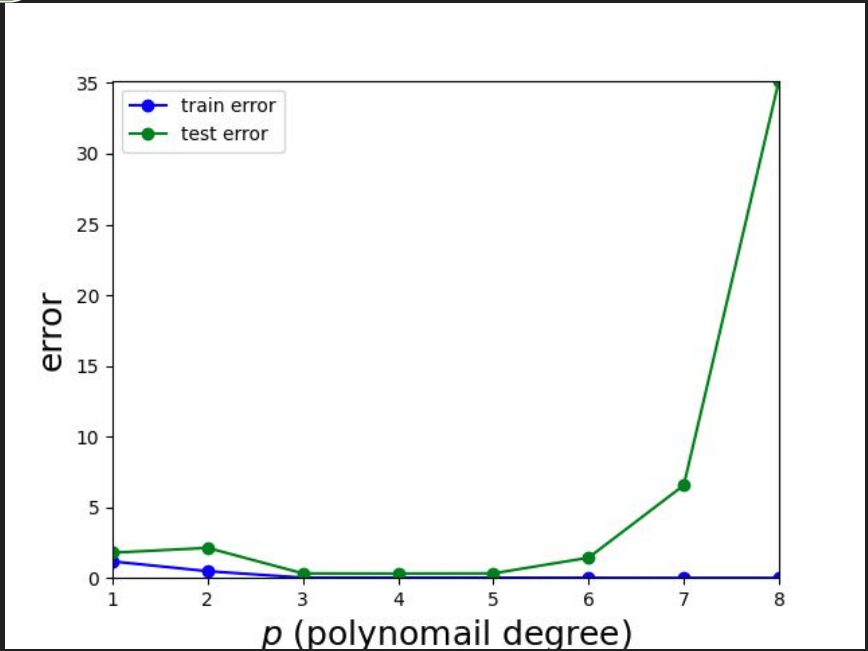
polynomial degree p = 8





Train and test errors

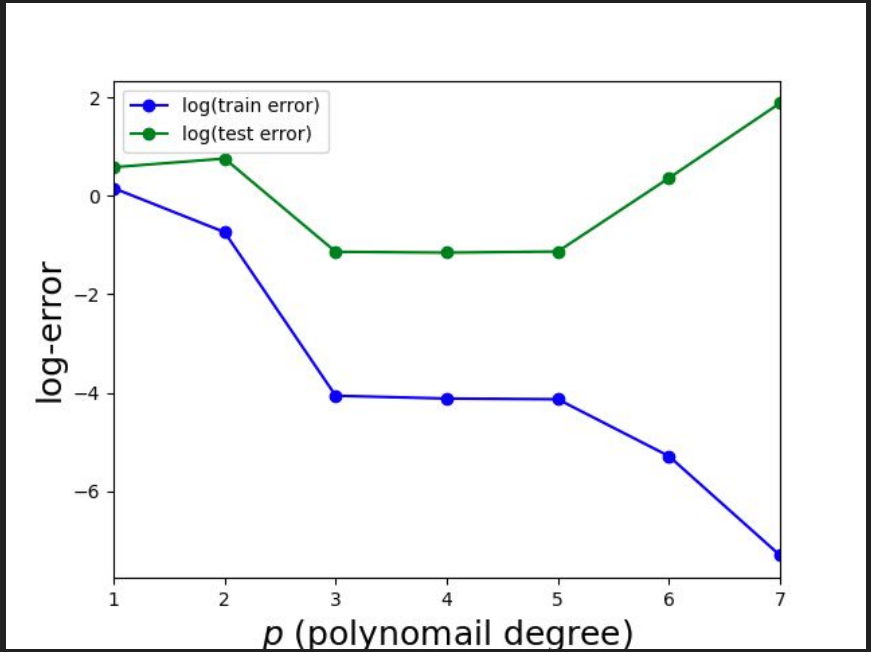




Train and test errors

K. N. Toosi
University of Technology

$$x \in R \rightarrow f_\theta \rightarrow y = a_p x^p + a_{p-1} x^{p-1} + \ldots + a_1 x + a_0$$



Cross-validation

K. N. Toosi
University of Technology

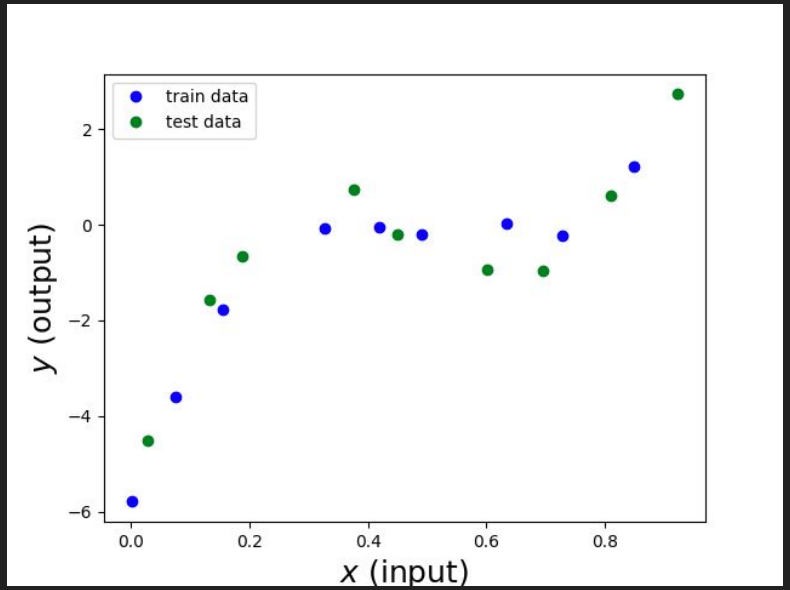
How good model work on new data?

- Divide data into train and test sets
- Train the model using train data
- Evaluate on test data

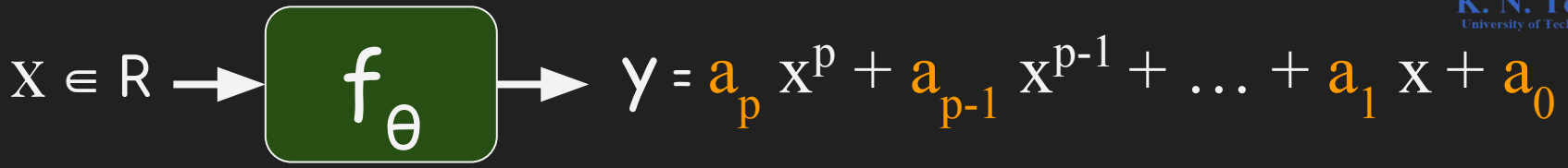
$$\theta^* = \text{argmin}_\theta \sum_{i \in \text{train}} (f(\theta, x_i) - y_i)^2$$

$$Err_{test} = \sum_{i \in \text{test}} (f(\theta^*, x_i) - y_i)^2$$

- divide the data in different ways and

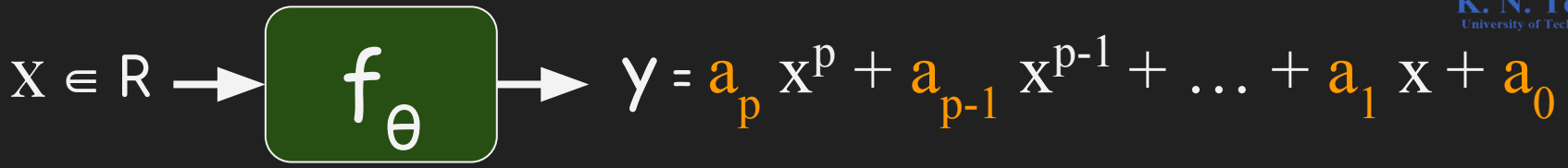


How to choose hyper-parameters?



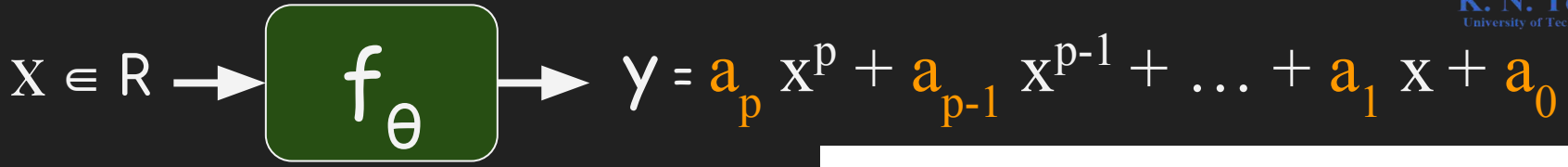
- Hyperparameters
  - Are not learned during model fitting
  - Used for
    - Learning the structure
    - Choosing the right model
    - …
- In polynomial regression the degree parameter (p) is a hyperparameter.

How to choose hyper-parameters?

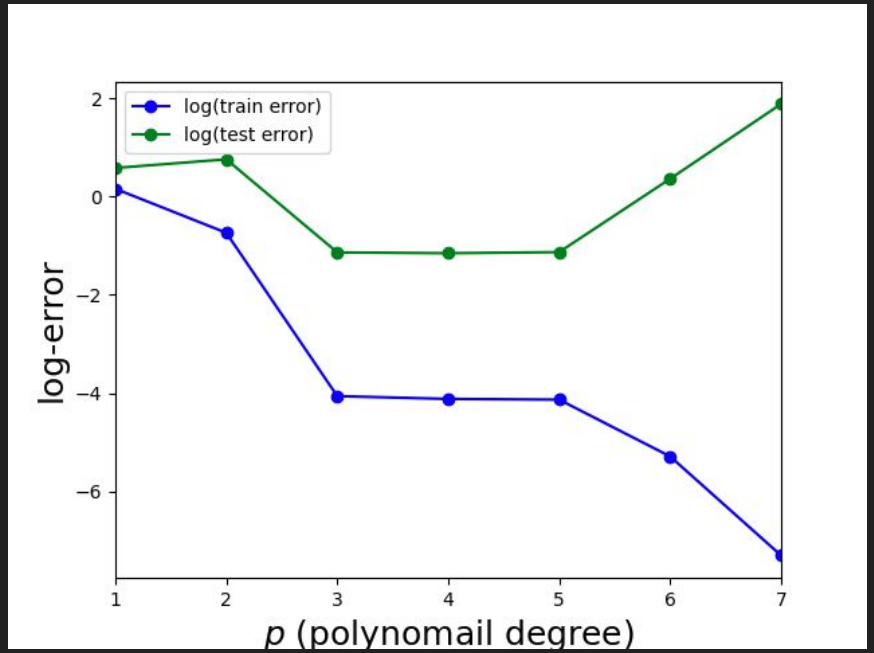


- Hyperparameters
  - Are not learned during model fitting
  - Used for
    - Learning the structure
    - Choosing the right model
- In polynomial regression the degree parameter (p) is a hyperparameter.
- How to choose hyper parameters?

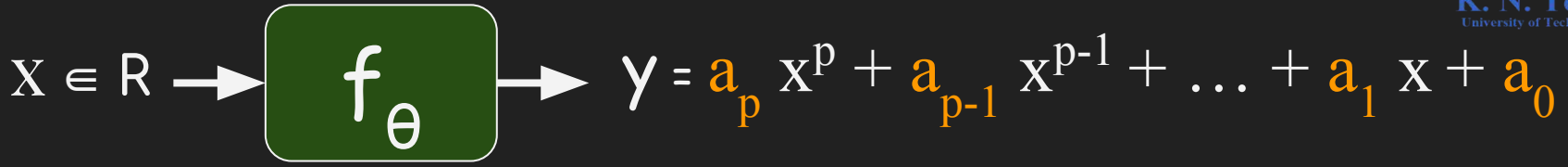
How to choose hyper-parameters?



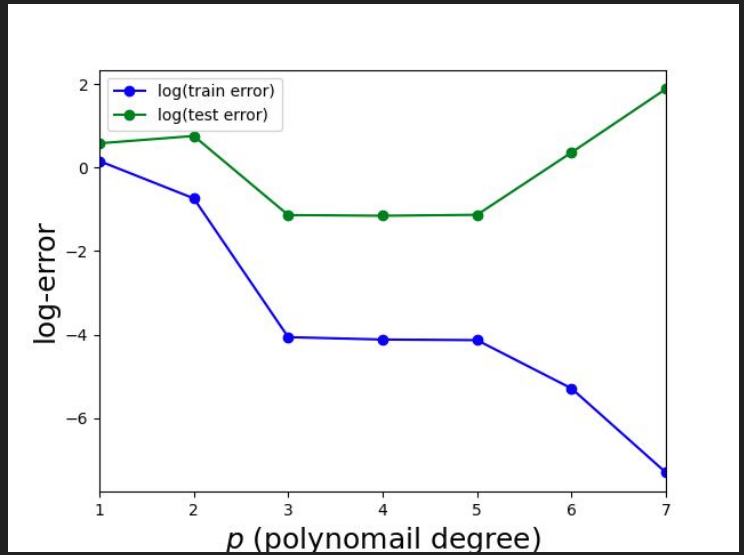
- How to choose p?



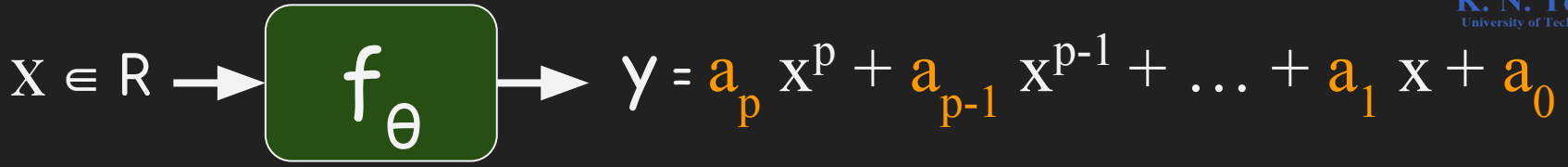
How to choose hyper-parameters?



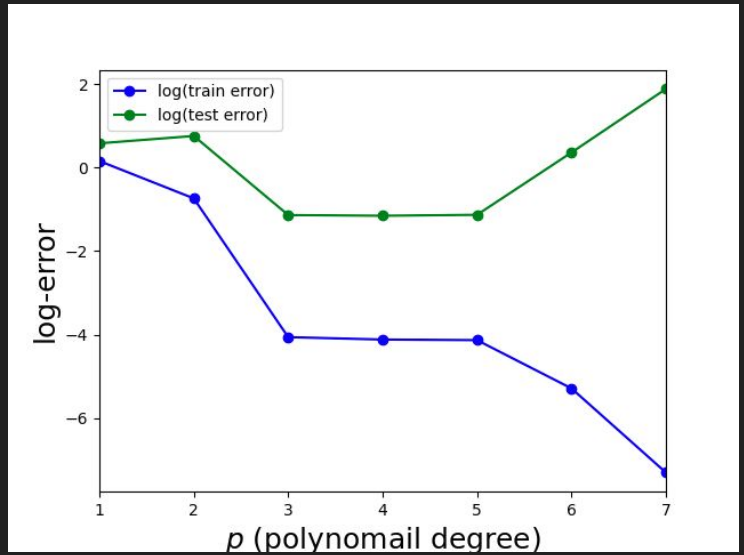
- How to choose p?
  - Split data to train and test sets
  - Train regular parameters using training data
  - Select the p that minimizes test error



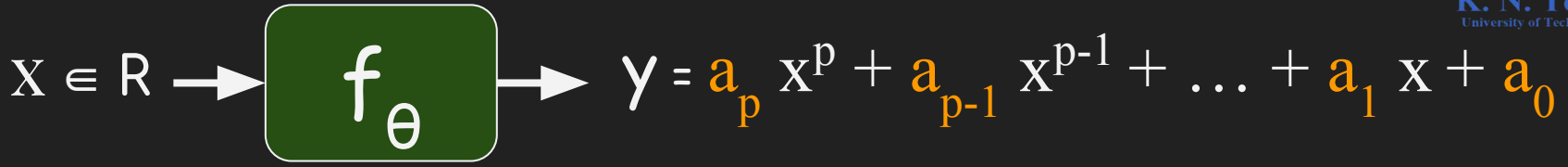
How to choose hyper-parameters?



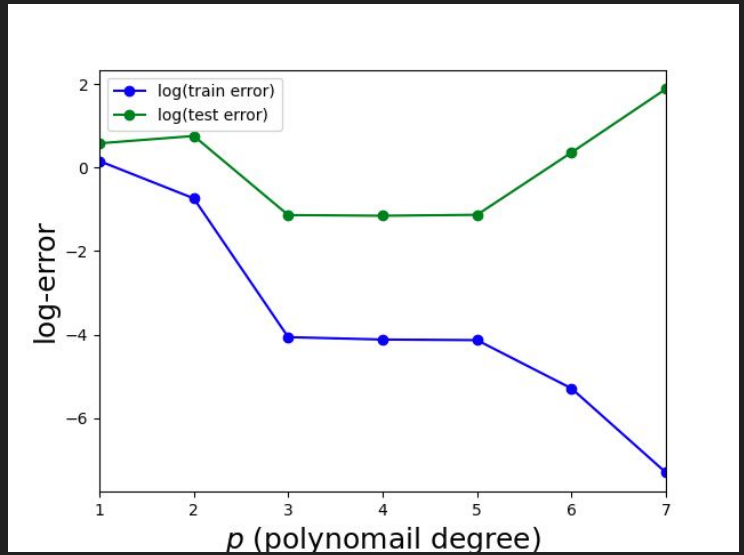
- How to choose p? (cross-validation)
  - Split data to train and test sets
  - Train regular parameters using training data
  - Select the p that minimizes test error
- [Divide in different ways and use the average error.]



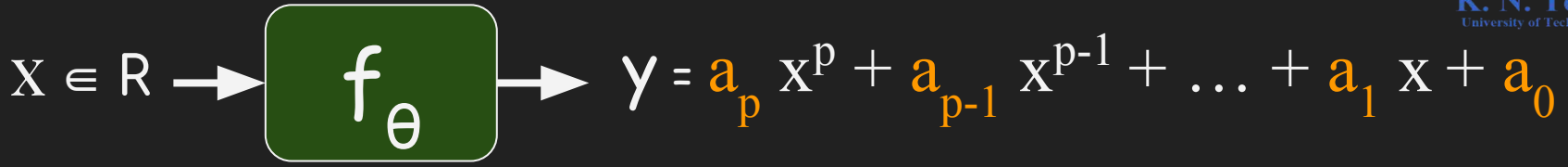
How to choose hyper-parameters?



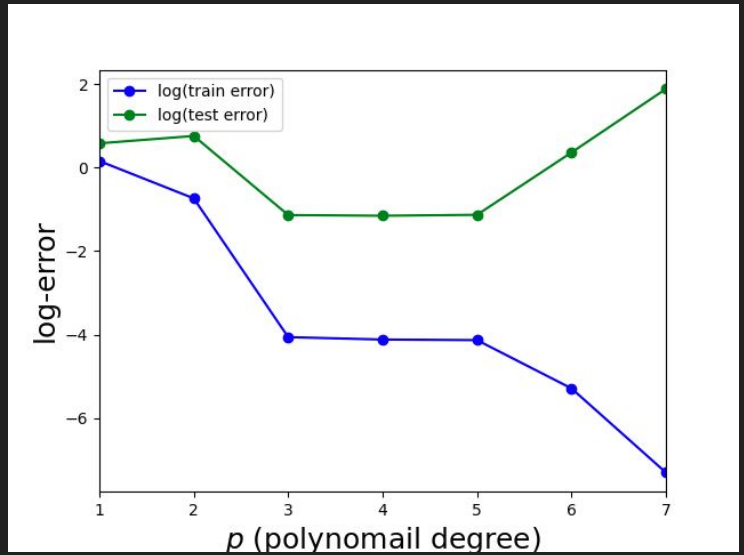
- How to choose p?
  - Split data to train and test sets
  - Train regular parameters using training data
  - Select the p that minimizes test error
- How to evaluate model on unseen data?



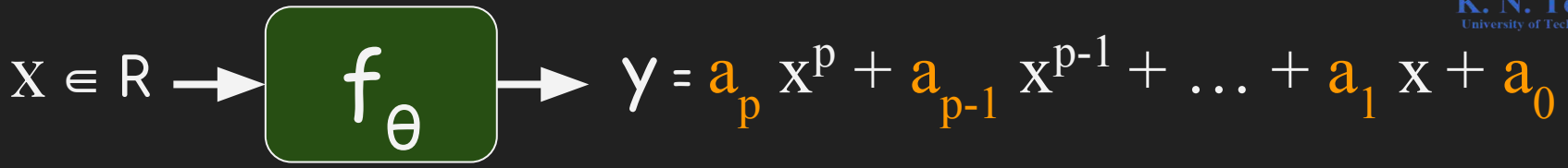
How to choose hyper-parameters?



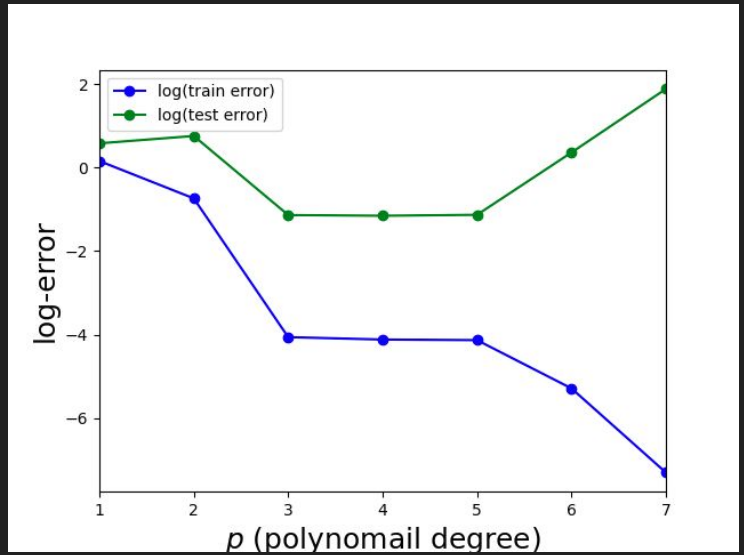
- How to choose p?
  - Split data to train and test sets
  - Train regular parameters using training data
  - Select the p that minimizes test error
- How to evaluate model on unseen data?
  - Error on train set?
    - train data is used to learn parameters
  - Error on test set?
    - test data is used to choose hyperparameters



How to choose hyper-parameters?



- How to evaluate model on unseen data?
  - Split data into 3 subsets: train, validation, and test
  - Train regular parameters using train set
  - Choose hyperparameters using validation set
  - Report system error on test set
- [Divide in different ways and report the average error.]



How to avoid overfitting

- Limit model complexity (e.g. degree of polynomial)
- **Regularization**