THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Predictive Analytics using Machine Learning

Topic

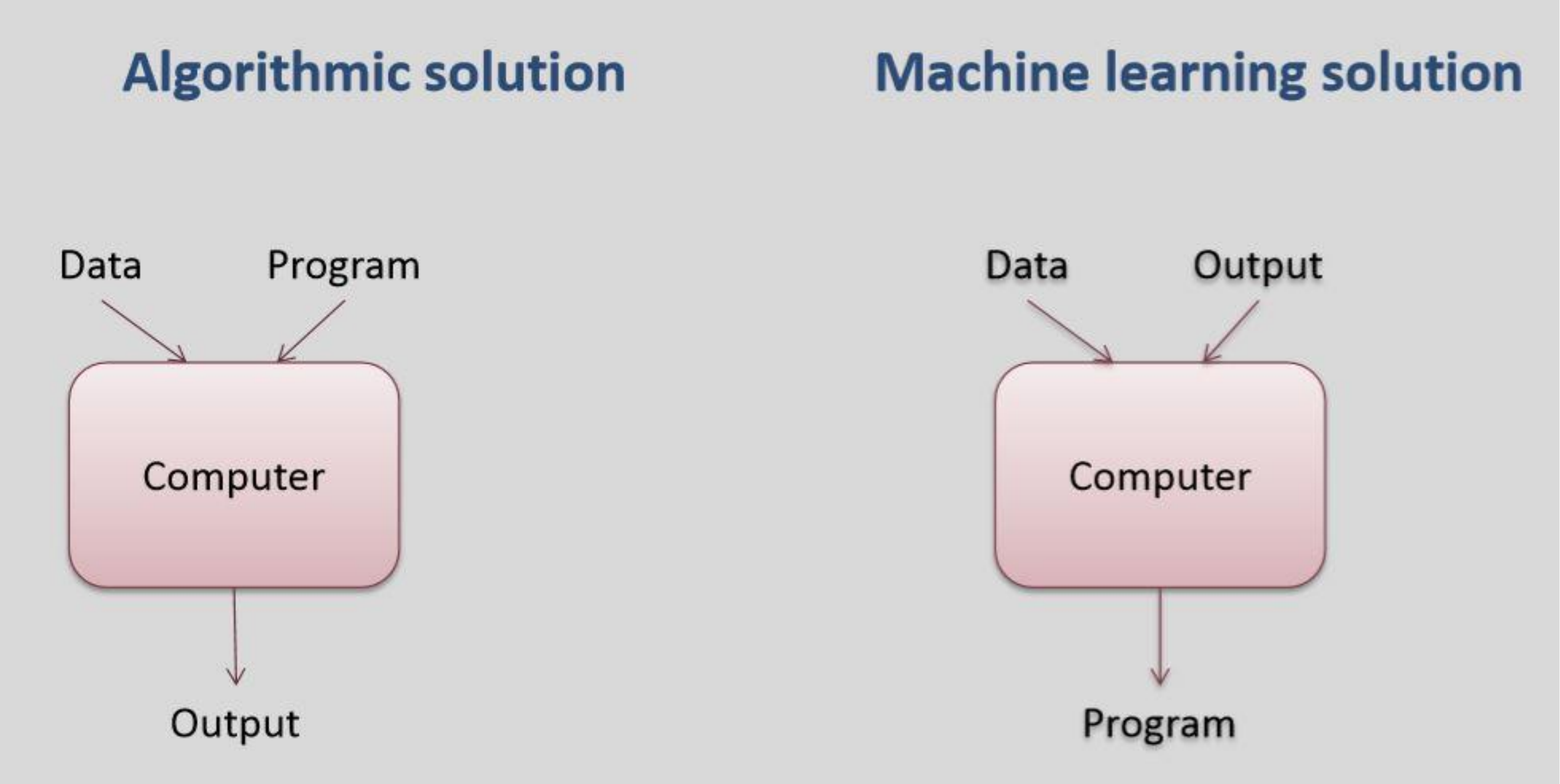
Machine Learning: A Perspective of Statistics

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Machine Learning: A Definition

A computer program is said to ***learn*** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T as measured by P, and improves with experience E.

Program Vs. Machine Learning



When to use Machine Learning?

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Human expertise does not exist
  - (navigating on Mars)
- Humans are unable to explain their expertise
  - (speech recognition)
- Solution changes in time
  - (routing on a computer network)
- Solution needs to be adapted to particular cases
  - (user biometrics)

Components of a Learning System

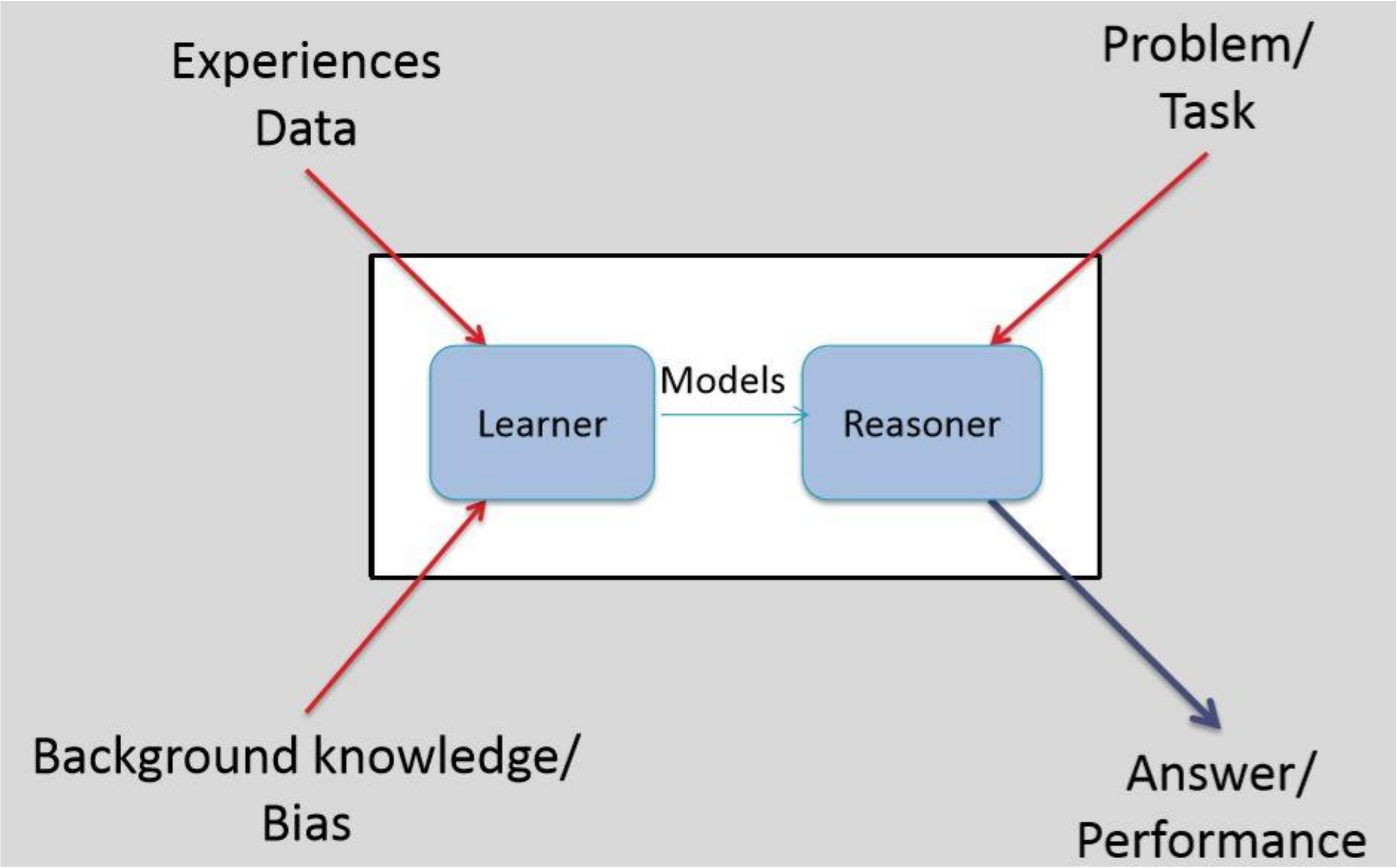
THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

(i) Task (T)

(ii)Data (Experience, E)

(iii)Performance Measure (P)

Learning System



Mathematical Understanding

A dataset (D) comprise of two types of features as:

- **A set of features X = $\{x_1, x_2, x_3 \ldots\ldots x_n\}$**
- **A target feature Y = $f(x)$**

Task of Learner

To estimate the function $\hat{Y} = \hat{f}(x)$ from D, where, $\hat{Y} = f(x) + \varepsilon$

Task of Reasoner

To compute $\hat{Y} = \hat{f}(x)$ for a new value of x.

Mathematical Understanding

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Task of Learner

To estimate the function

$\hat{Y} = \hat{f}(x)$ **from D**

$$\hat{Y} = f(x) + \varepsilon$$

Task of Reasoner

To compute $\hat{Y} = \hat{f}(x)$ **for a new value of x.**

Types of features (X and Y)

(i) Categorical (such as blood group)
(ii) Ordinal (such as large, medium, or small)
(iii) Integer valued (such as no. of students)
(iv) Real valued (such as height, weight)

Categories of features (X and Y)

(i) Discrete
(ii) Continuous

Height (x1)	Age (x2)	Complexion (x3)	Weight (x4)
5.1	20	Fair	60.5
2.1	3	Dark	20.2
6.7	30	Dark	80.6
4	10	Fair	40.5

Types of Machine Learning

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

1. Supervised Learning

- ✓ Classification (When Y is discrete)
- ✓ Regression (When Y is continuous)

Classification

Training Data

Height (x1)	Age (x2)	Complexion (y)
5.1	20	Fair
2.1	3	Dark
6.7	30	Dark
4	10	Fair

Predict the value of Complexion for Height=2.5 and Age=5.

Regression

Training Data

Height (x1)	Age (x2)	Weight (y)
5.1	20	60.5
2.1	3	20.2
6.7	30	80.6
4	10	40.5

Predict the value of weight for Height=2.5 and Age=5.

Types of Machine Learning

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

2. Unsupervised Learning

It draws inferences from the values of X to obtain pattern of the data.

✓ Clustering

Type of Medicine	Weight	PH-Value
A	1	1
B	2	1
C	4	3
D	5	4

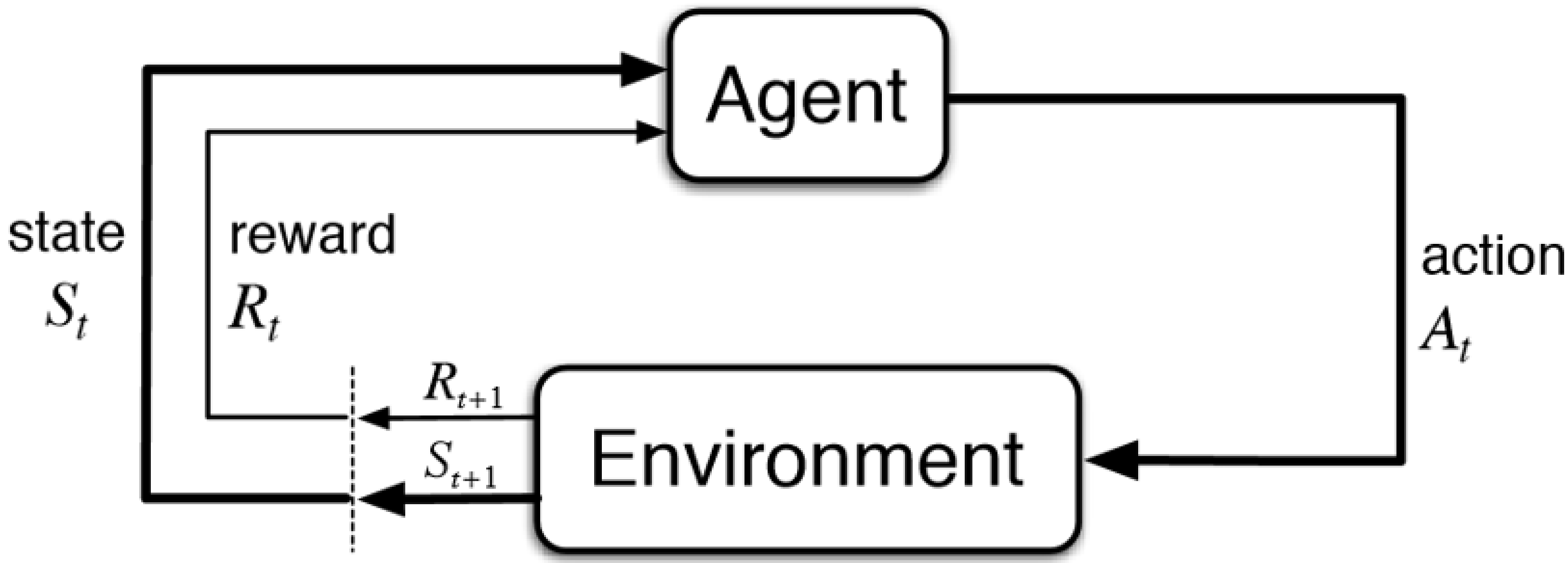
What is the type of a medicine with weight=2 and PH-value=2?

Types of Machine Learning

3. Reinforcement Learning

- ✓ It enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

- ✓ It uses **rewards and punishments** as signals for positive and negative behavior. (The supervised learning consists positive signal only)



Performance Measures

(i) **Supervised Learning**

Regression – Squared Error or absolute error

Classification - Precision/Recall

(ii) **Unsupervised Learning**

Clustering – Scatter

(iii) **Reinforcement Learning** – Award/Punishment

Performance Measure for Regression

(i) Mean Absolute Error (MEA)

$$MEA = \frac{1}{N}\sum |Y - \hat{Y}|$$

(ii) Mean Square Error (MSE)

$$MSE = \frac{1}{N}\sum (Y - \hat{Y})^2$$

Performance Measures

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Performance Measure for Classification

Confusion Matrix

	Predicted		
Actual	Class	Cat	Dog
	Cat	1	3
	Dog	0	8

True-Positive

False-Negative

True-Negative

False-Positive

Performance Measures

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Performance Measure for Classification

<u>Confusion Matrix</u>

	Predicted		
Actual	Class	Cat	Dog
	Cat	0	3
	Dog	1	4

Preceision= $\frac{TP}{TP+FP}$

Recall= $\frac{TP}{TP+FN}$

Accuracy= $\frac{TP+TN}{TP+FP+TN+FN}$

Performance Measures

Performance Measure for Classification

Confusion Matrix

F-1 Score= $\frac{2*preceision*recall}{preceision+recall}$

You Explore

The AUC curve to measure performance of classification.

Bias and Variance

Consider the following training dataset.

X	Y
2	4
3	6
4	8
5	10
6	12

After applying linear regression algorithm over the training data, the following target function is estimated.

$$\hat{Y} = \hat{f}(x) = 2 * X + 1$$

X	Y	$\hat{Y}$
2	4	5
3	6	7
4	8	9
5	10	11
6	12	13

Mean of $\hat{Y}$= (5+7+9+11+13)/5=9

Bias and Variance

$E(\hat{Y})$=Mean of $\hat{Y}$= (5+7+9+11+13)/5=9

X	Y	$\hat{Y}$	$E(\hat{Y}) - Y$	$E(\hat{Y}) - \hat{Y}$
2	4	5	5	4
3	6	7	3	2
4	8	9	1	0
5	10	11	-1	-2
6	12	13	-3	-4

$$\text{Bias}^2 = E(\mathbf{E}(\hat{Y}) - Y)^2$$
$$= (25+9+1+1+9)/5$$
$$= 9$$

$$\text{Variance} = E(\mathbf{E}(\hat{Y}) - \hat{Y})^2$$

$$= (16+4+0+4+16)$$
$$= 8$$

Bias and Variance

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

$\text{Bias}^2 = E(\mathbf{E}(\widehat{\boldsymbol{Y}}) - Y)^2$
$= (25+9+1+1+9)/5$
$= 9$

Bias:

(i) Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

(ii) Model with high bias pays very little attention to the training data and oversimplifies the model.

(iii) It always leads to high error on training and test data.

Bias and Variance

Variance= $E(\mathbf{E}(\widehat{\boldsymbol{Y}}) - \widehat{\boldsymbol{Y}})^2$

= (16+4+0+4+16)
= 8

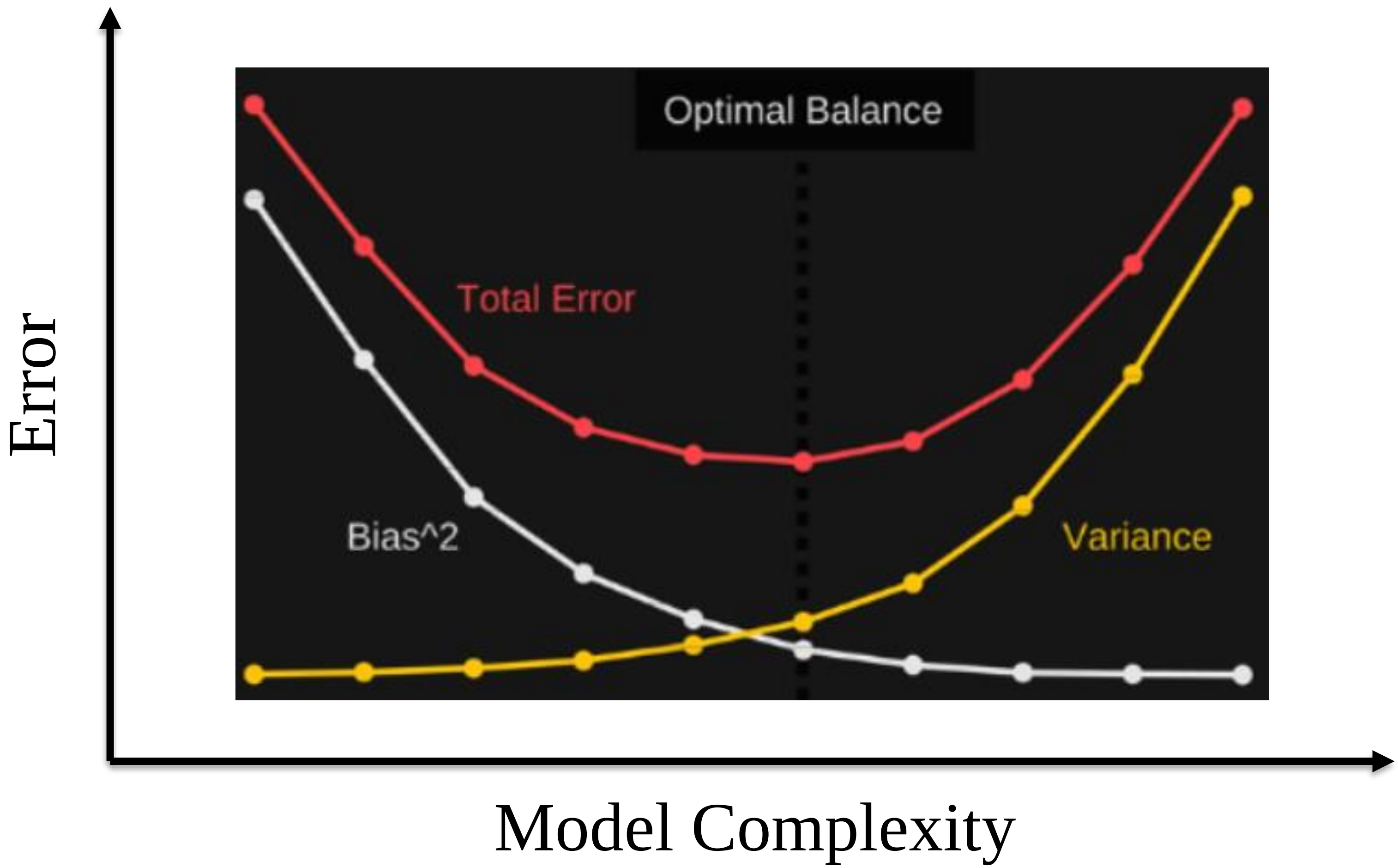
Variance

(i) Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.

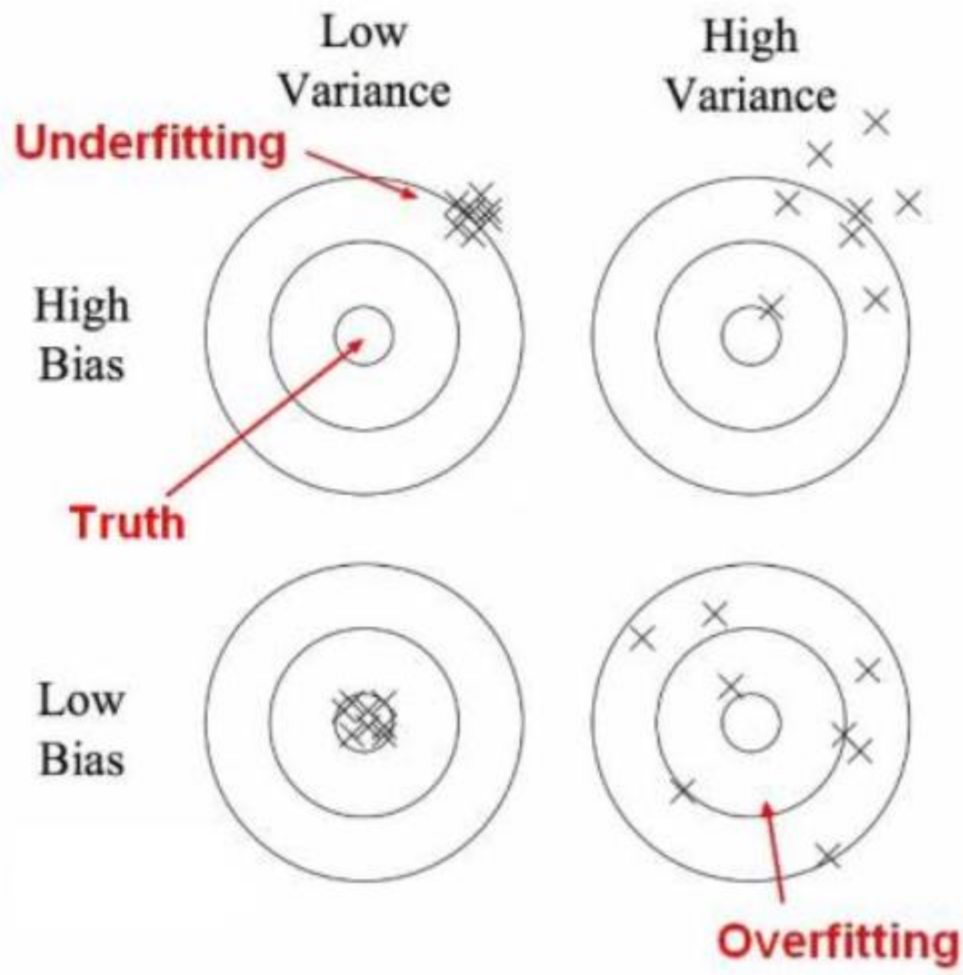
(ii) Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.

(iii) As a result, such models perform very well on training data but has high error rates on test data.

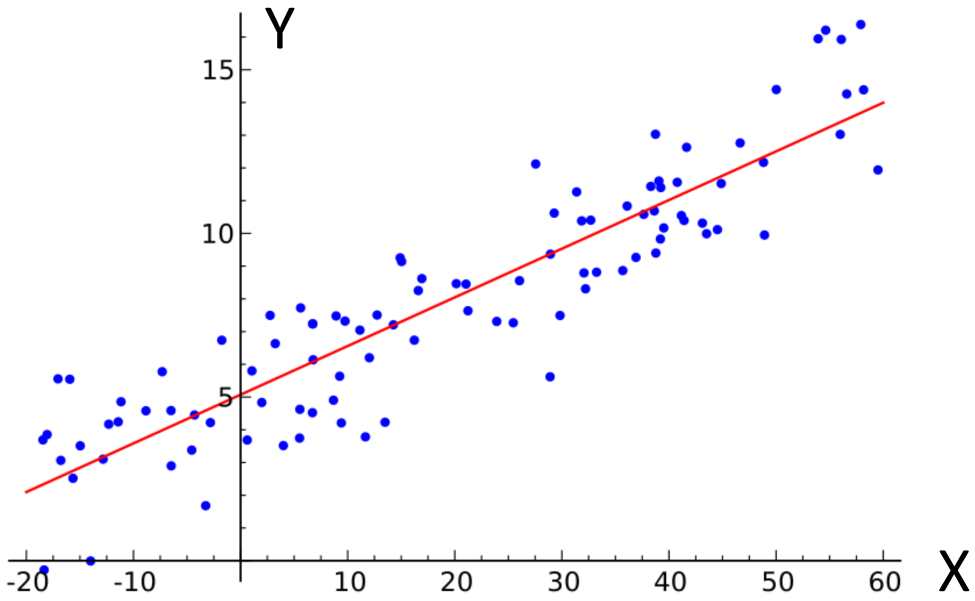
Bias and Variance Tradeoff



Bias and Variance Tradeoff



Simple Linear Regression



Response Variable

Covariate

Linear Model: $$Y = mX + b$$

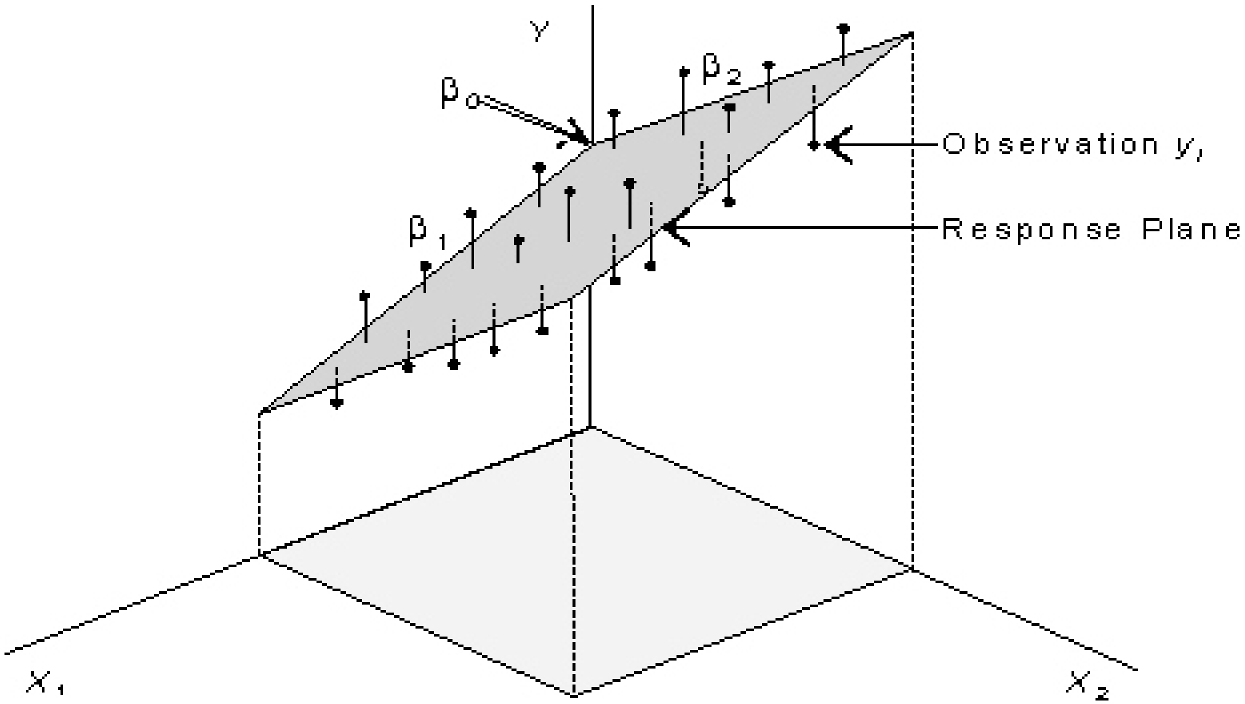
Slope

Intercept (bias)

Motivation

- One of the most widely used techniques
- Fundamental to many larger models
  - Generalized Linear Models
  - Collaborative filtering
- Easy to interpret
- Efficient to solve

Multiple Linear Regression



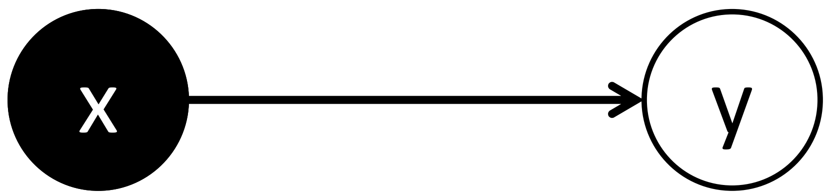
The Regression Model

- For a *single* data point *(x,y)*:

Independent Variable (Vector)

Response Variable (Scalar)

Observe: (Condition)



$x \in \mathbb{R}^p$

$y \in \mathbb{R}$

- Joint Probability:

$$p(x,y) = p(x)\boxed{p(y|x)}$$

Discriminative Model

The Linear Model

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY (Deemed to be University)

Scalar Response

$$y = \overbrace{\theta^T}^{\text{Vector of Parameters}} \overbrace{x}^{\text{Vector of Covariates}} + \epsilon$$

Real Value Noise

+ b

Noise Model:

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

Linear Combination of Covariates

$$\sum_{i=1}^{p} \theta_i x_i$$

What about bias/intercept term?

$$\text{Define: } x_{p+1} = 1$$

Then redefine p := p+1 for notational simplicity

Conditional Likelihood p(y|x)

- Conditioned on x:

$$y = \overbrace{\theta^T x}^{\text{Constant}} + \boxed{\epsilon \sim \overset{\text{Normal Distribution}}{N(\underset{\text{Mean}}{0}, \underset{\text{Variance}}{\sigma^2})}}$$

- Conditional distribution of Y:

$$Y \sim N(\theta^T x, \sigma^2)$$

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - \theta^T x)^2}{2\sigma^2}\right)$$

Parameters and Random Variables

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Parameters

$$y \sim N(\boxed{\theta^T} x, \boxed{\sigma^2})$$

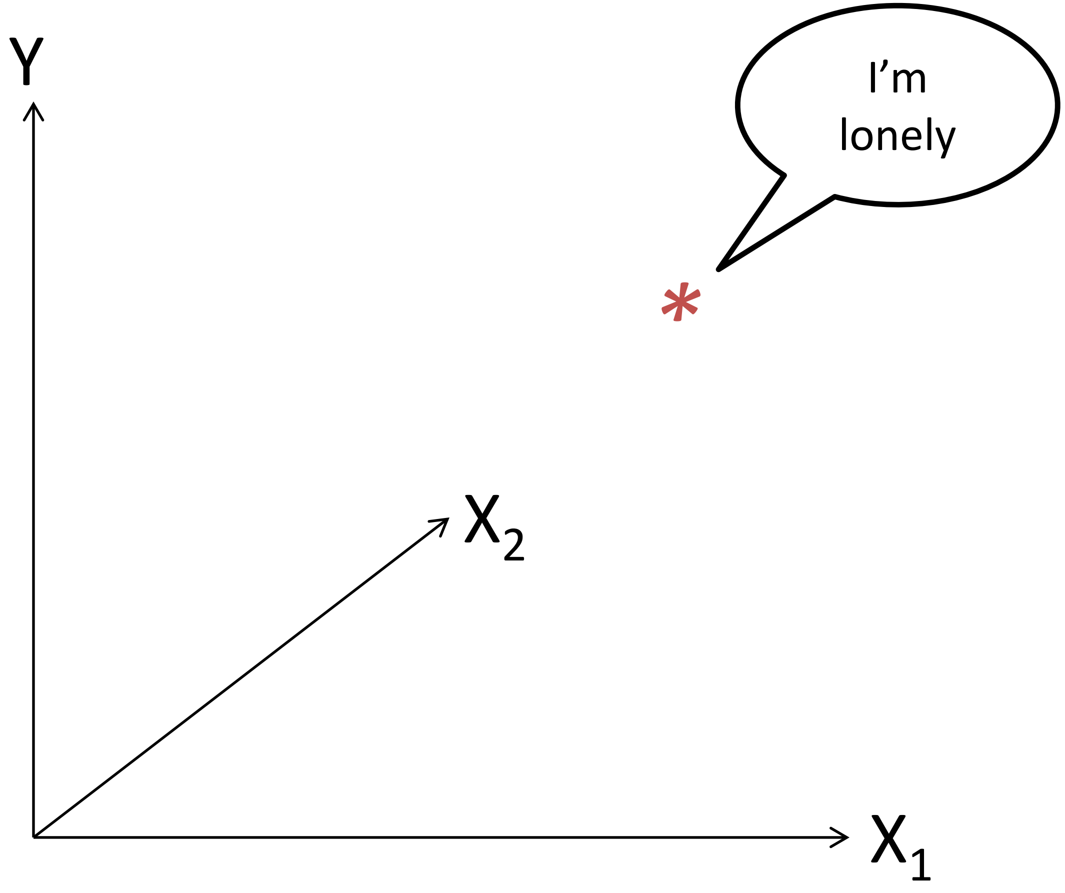
- Conditional distribution of y:
  - Bayesian: parameters as random variables

$$p(y|x, \theta, \sigma^2)$$

  - Frequentist: parameters as (unknown) constants

$$p_{\theta,\sigma^2}(y|x)$$

So far …

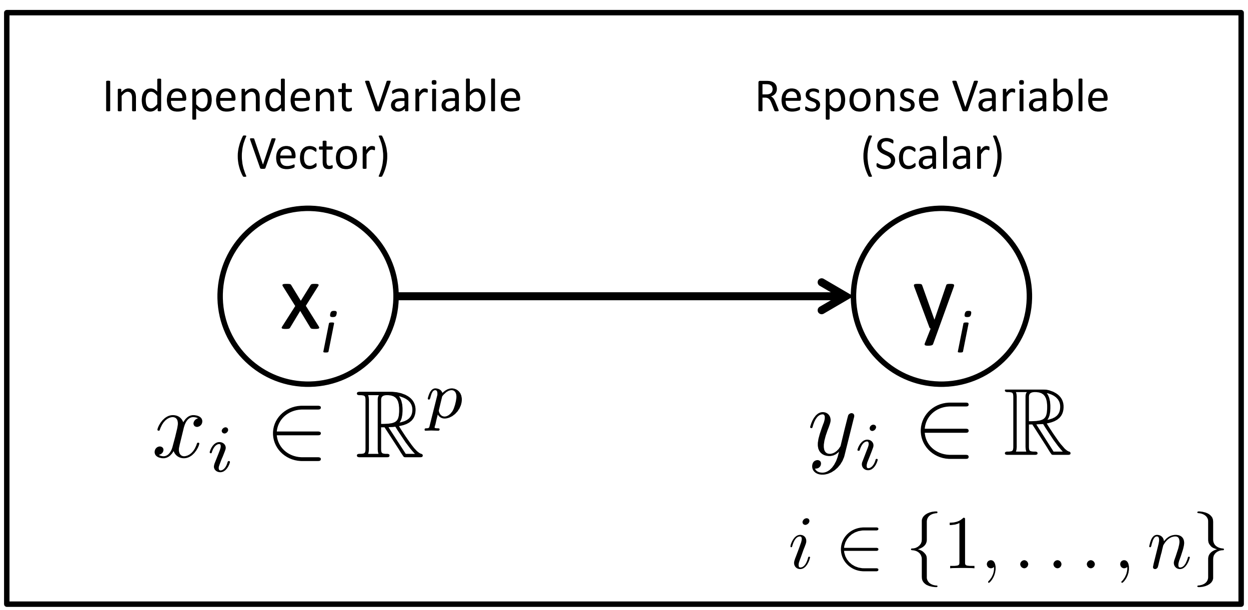


Independent and Identically Distributed (iid) Data

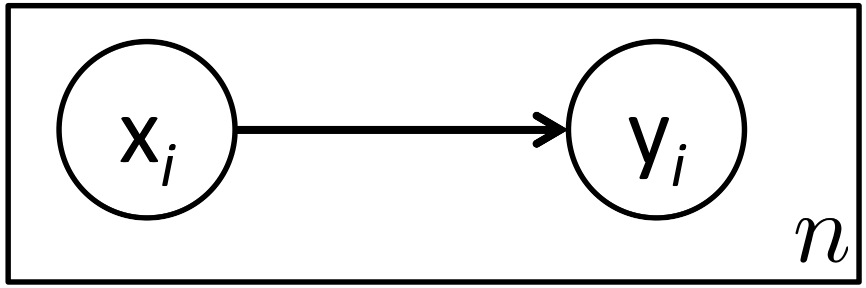
- For *n* data points:

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$
$$= \{(x_i, y_i)\}_{i=1}^{n}$$

Plate Diagram



Joint Probability



- For *n* data points **independent and identically distributed (iid)**:

$$p(\mathcal{D}) = \prod_{i=1}^{n} p(x_i, y_i)$$

$$= \prod_{i=1}^{n} p(x_i) p(y_i | x_i)$$

Rewriting with Matrix Notation

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

- Represent data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ as:

Covariate (Design) Matrix

$$X = n\left\{\underbrace{\begin{pmatrix} — & x_1 & — \\ — & x_2 & — \\ & \ldots & \\ — & x_n & — \end{pmatrix}}_{p}\right. \in \mathbb{R}^{np}$$

Assume *X* has rank p (not degenerate)

Response Vector

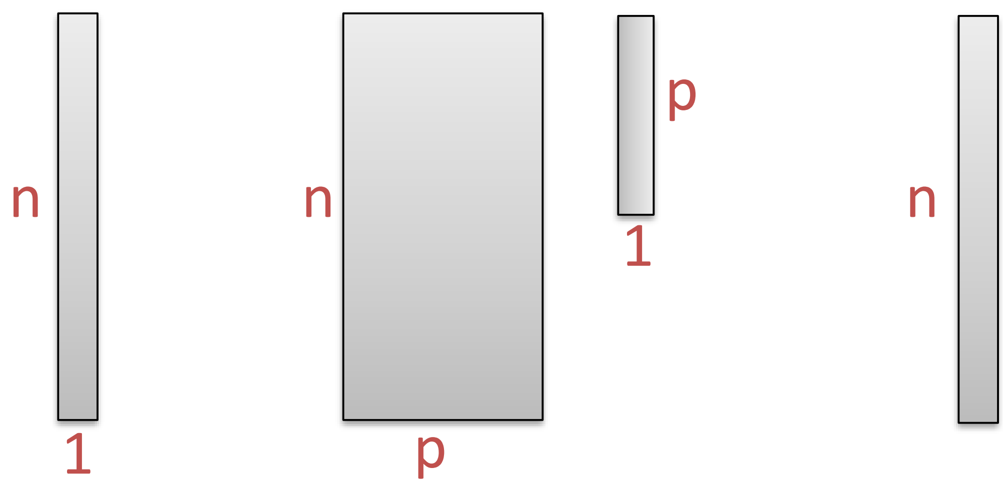
$$Y = n\left\{\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{1}\right. \in \mathbb{R}^{n}$$

Rewriting with Matrix Notation

- Rewriting the model using matrix operations:

$$Y = X\theta + \epsilon$$

$$Y \quad = \quad X \quad \theta \quad + \quad \epsilon$$



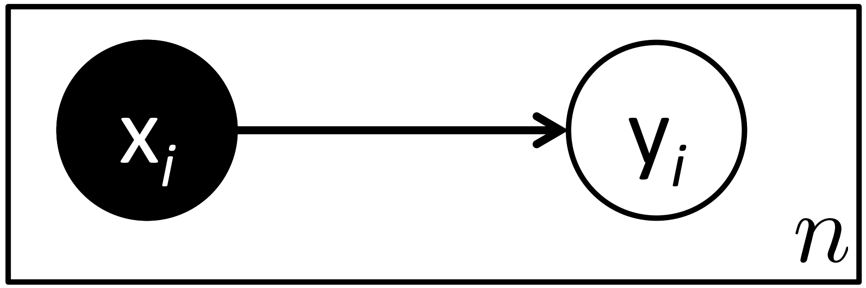
Estimating the Model

- Given data how can we estimate θ?

$$Y = X\theta + \epsilon$$

- Construct maximum likelihood estimator (MLE):
  - Derive the log-likelihood
  - Find θ_{MLE} that maximizes log-likelihood
    - Analytically: Take derivative and set = 0
    - Iteratively: (Stochastic) gradient descent

Joint Probability



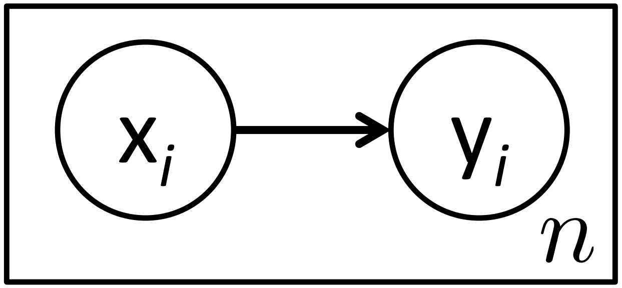
- For n data points:

$$p(\mathcal{D}) = \prod_{i=1}^{n} p(x_i, y_i)$$

$$= \prod_{i=1}^{n} \overset{\text{“1”}}{\cancel{p(x_i)}}\, \boxed{p(y_i | x_i) \quad \text{Discriminative Model}}$$

Defining the Likelihood

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)



$$p_\theta(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\theta^T x)^2}{2\sigma^2}\right)$$

$$\begin{aligned}
\mathcal{L}(\theta|\mathcal{D}) &= \prod_{i=1}^{n} p_\theta(y_i|x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i-\theta^T x_i)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta^T x_i)^2\right)
\end{aligned}$$

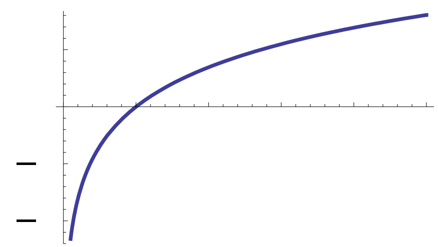
Maximizing the Likelihood

- Want to compute:

$$\hat{\theta}_{\mathrm{MLE}} = \operatorname*{arg\,max}_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta | \mathcal{D})$$

- To simplify the calculations we take the log:

$$\hat{\theta}_{\mathrm{MLE}} = \operatorname*{arg\,max}_{\theta \in \mathbb{R}^p} \boxed{\log} \mathcal{L}(\theta | \mathcal{D})$$



which does not affect the maximization because log is a monotone function.

$$\mathcal{L}(\theta|\mathcal{D}) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2\right)$$

- Take the log:

$$\log \mathcal{L}(\theta|\mathcal{D}) = -\log(\sigma^n (2\pi)^{\frac{n}{2}}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$

- Removing constant terms with respect to θ:

$$\log \mathcal{L}(\theta) = -\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$

Monotone Function
(Easy to maximize)

$$\log \mathcal{L}(\theta) = -\sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

- Want to compute:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \mathbb{R}^p} \log \mathcal{L}(\theta | \mathcal{D})$$

- Plugging in log-likelihood:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \mathbb{R}^p} - \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

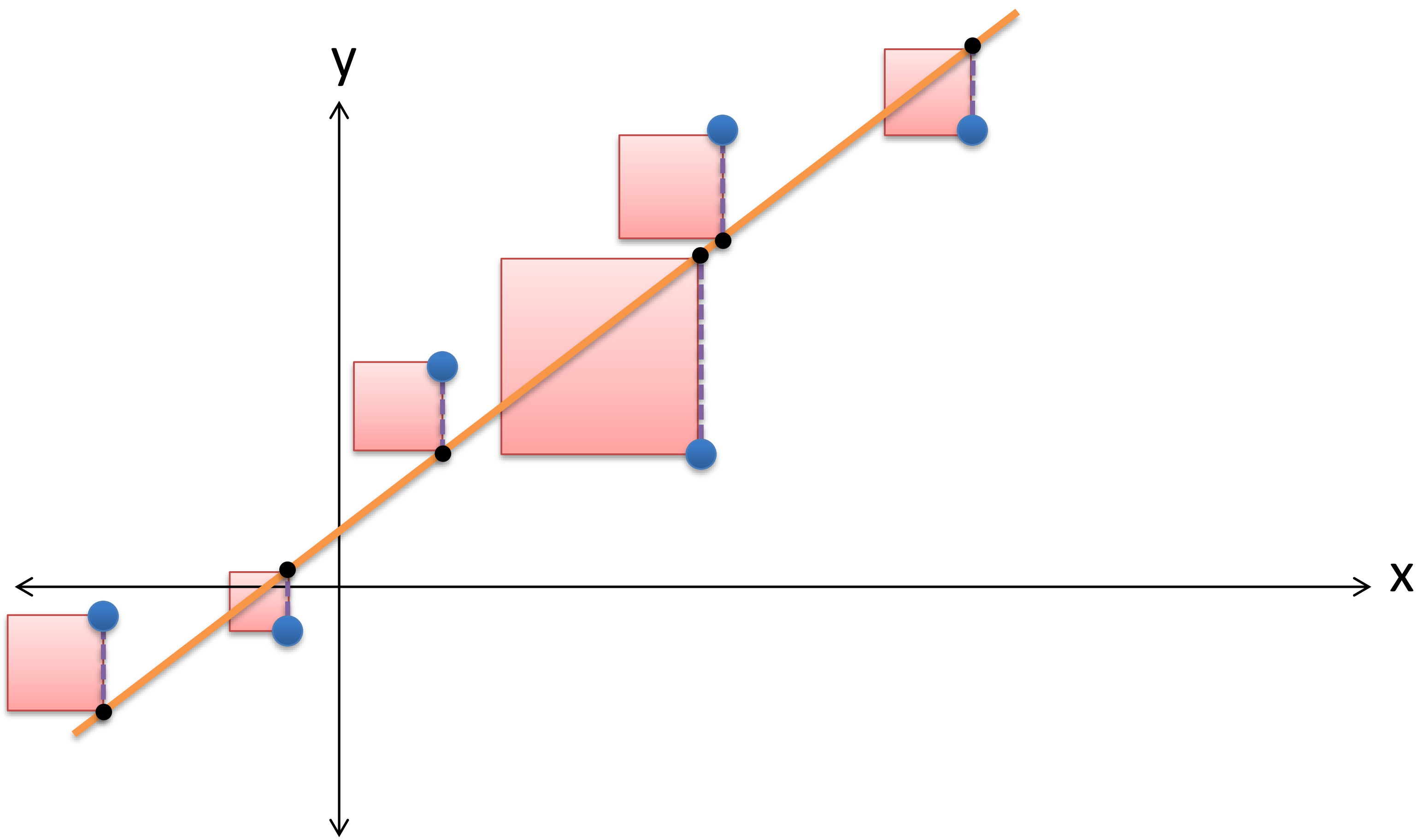
$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \mathbb{R}^p} - \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

- Dropping the sign and flipping from maximization to minimization:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \underbrace{(y_i - \theta^T x_i)^2}_{\text{Minimize Sum (Error)}^2}$$

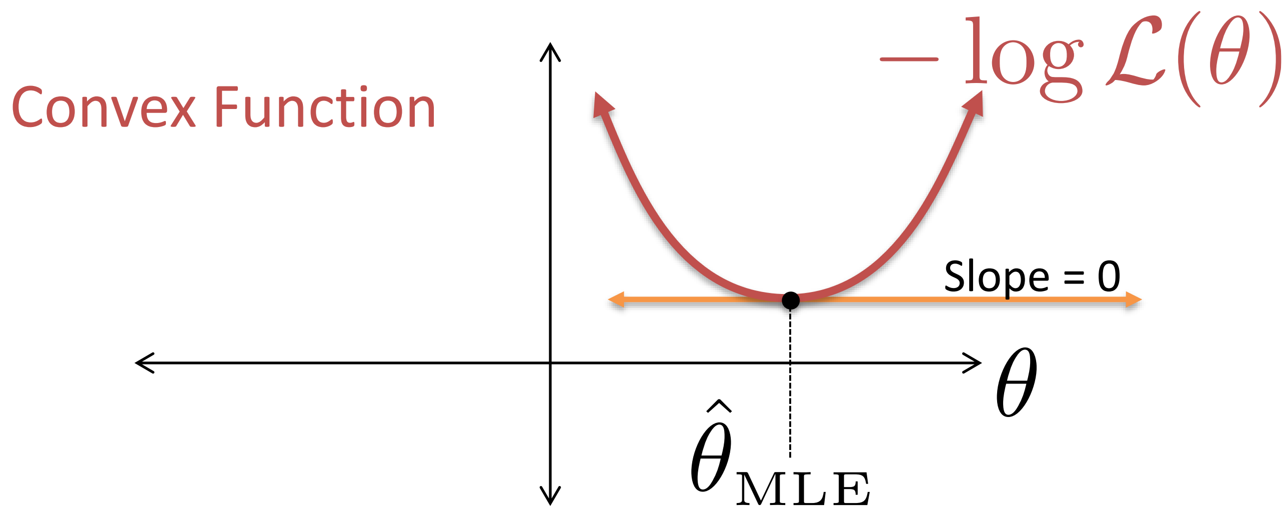
- Gaussian Noise Model → Squared Loss
  – Least Squares Regression

Pictorial Interpretation of Squared Error



Maximizing the Likelihood (Minimizing the Squared Error)

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$



- Take the gradient and set it equal to zero

Minimizing the Squared Error

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

- Taking the gradient

$$-\nabla_\theta \log \mathcal{L}(\theta) = \nabla_\theta \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

Chain Rule →

$$= -2 \sum_{i=1}^{n} (y_i - \theta^T x_i) x_i$$

$$= -2 \sum_{i=1}^{n} y_i x_i + 2 \sum_{i=1}^{n} (\theta^T x_i) x_i$$

- Rewriting the gradient in matrix form:

$$-\nabla_\theta \log \mathcal{L}(\theta) = -2\sum_{i=1}^{n} y_i x_i + 2\sum_{i=1}^{n}(\theta^T x_i)x_i$$

$$= -2X^TY + 2X^TX\theta$$

- To make sure the log-likelihood is convex compute the second derivative (Hessian)

$$-\nabla^2 \log \mathcal{L}(\theta) = 2X^TX$$

- If *X* is full rank then X^TX is positive definite and therefore θ_{MLE} is the minimum
  - Address the degenerate cases with regularization

$$-\nabla_\theta \log \mathcal{L}(\theta) = -2X^T y + 2X^T X\theta = 0$$

- Setting gradient equal to 0 and solve for θ_{MLE}:

$$(X^T X)\hat{\theta}_{\mathrm{MLE}} = X^T Y$$

$$\hat{\theta}_{\mathrm{MLE}} = (X^T X)^{-1} X^T Y$$

Normal Equations
(Write on board)