

ESE 402-542

## Lecture 1:

5 modules:

- 1) Estimating parameters/distributions of data  
= (5-6 lecs.)
  - 2) Hypothesis testing (5-6 lecs)  
=
- 
- Midterm
- 3) supervised learning (4-5 lecs)
  - 4) unsupervised learning (4 lecs)
  - 5) Introduction to statistical learning theory.  
= (4 lecs)

---

Grading:

Exercises: 40%  $\rightarrow$  7-8 Hws

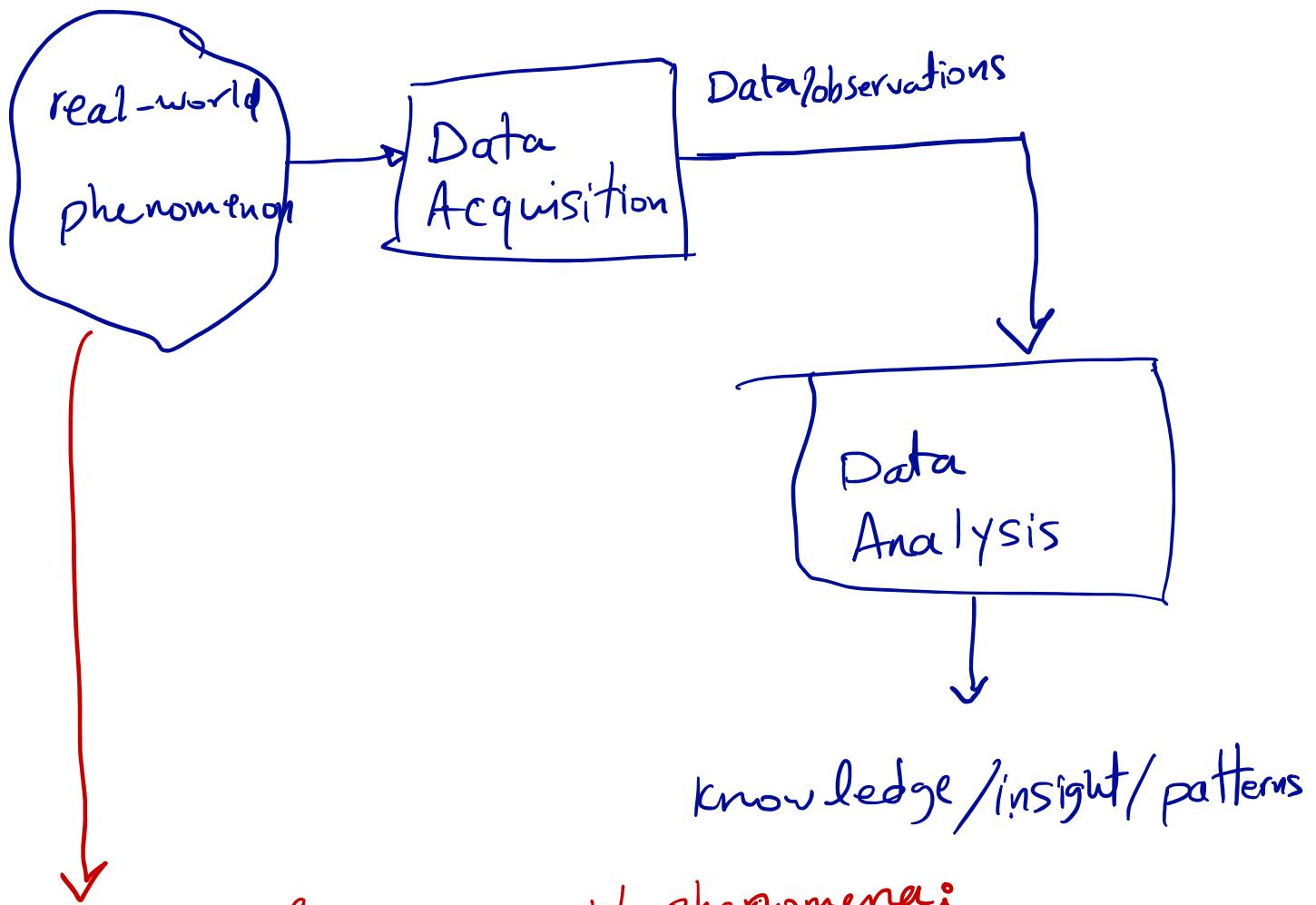
Exams (Midterm, Final): 60%

Text books  $\rightarrow$  See Canvas.

3 TAs. → office hours will be announced  
this week.

My office hours Fri 11am - 12pm  
EST

## Lecture 2 :



Examples of real-world phenomena:

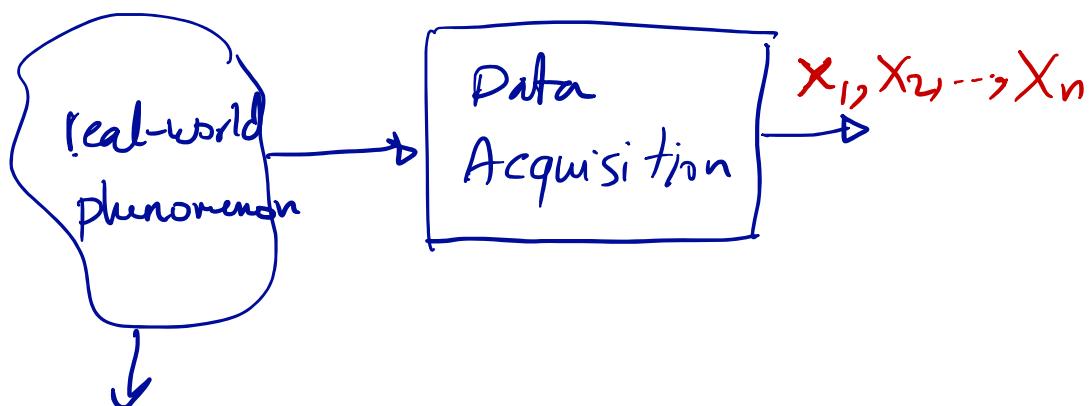
- finding the average height of women in the US
- percentage of people who'd vote for a certain candidate in the state of Pennsylvania

- likelihood of a person to develop breast cancer within the next 5 years
- average speed of molecules in a specific gas
- likelihood of a user to click on a certain ad in her face book page

Data Science involves developing methods of acquisition/storage/analysis of data to effectively extract useful knowledge/information/insight/patterns.

## Estimating Basic Statistics of Data:

Consider the task of estimation  
the average height of women  
across the US.



$$\{y_1, \dots, y_N\}$$

$N$  = population size

$$\left( \text{task} = \text{find } \mu = \frac{y_1 + y_2 + \dots + y_N}{N} \right)$$

Typically,  $N$  is a very large number (e.g.  $N = \text{tens of millions}$ )

and the naive method that goes over the whole population is often inefficient or infeasible.

So the basic idea of data acquisition is to sample from the real-world phenomenon.

---

Random Sampling: We will choose uniformly at random a subset of the population.

Let  $X_i$  be the data corresponding to an individual chosen uniformly at random in the population.

$$X_i = \left\{ \begin{array}{l} y_1 \\ y_2 \\ \vdots \\ y_N \end{array} \right\}$$

w.p.  $\frac{1}{N}$

w.p.  $\frac{1}{N}$

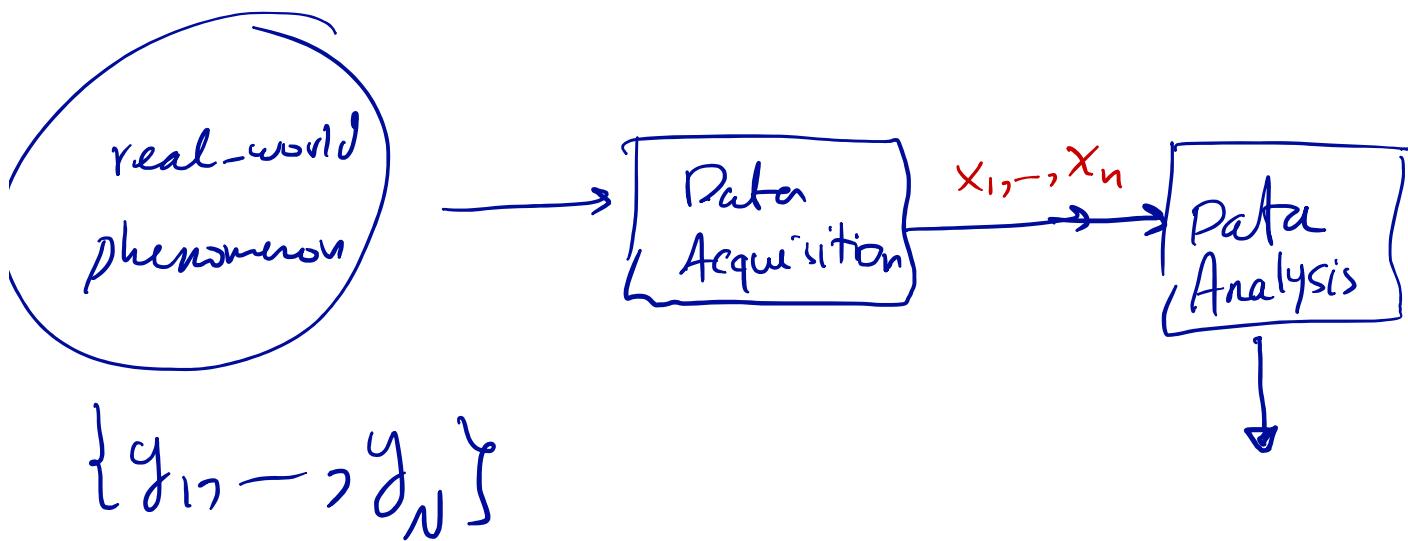
w.p.  $\frac{1}{N}$

furthermore  $X_i$ 's are independently chosen.

Data:  $X_1, X_2, \dots, X_n$

Typically  $n$  is much smaller

than  $N$ .

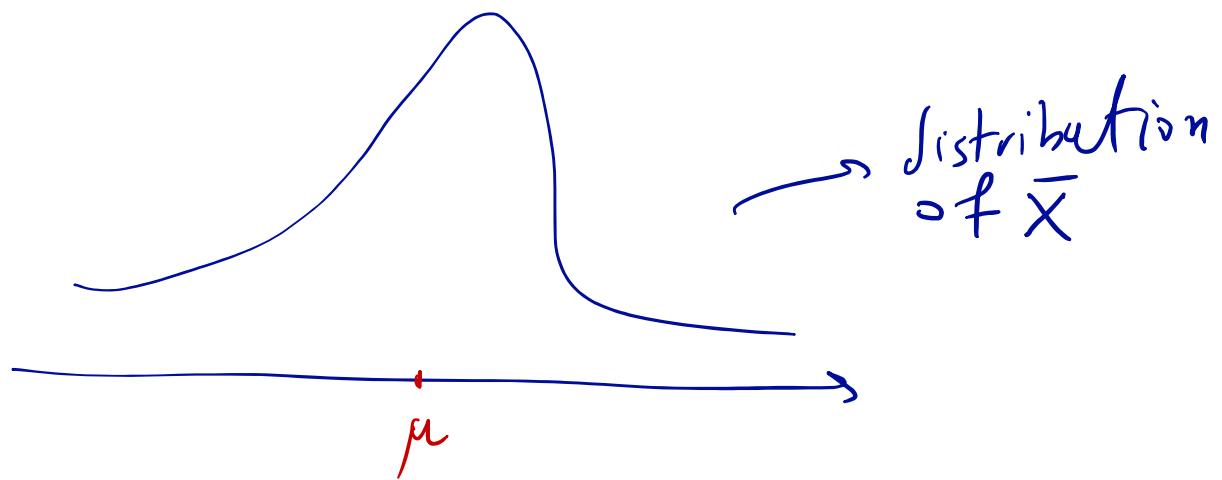


question : Estimate the average  
of  $y_i$ 's (i.e.  $\mu$ )  
by only using the  
data  $x_1, \dots, x_n$ .

Simplest answer:

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} \quad (\text{sample mean})$$

Note that  $\bar{X}$  is a random object/variable.



Let's first compute the expected value of  $\bar{X}$  and how it is related to  $\mu$ .

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{x_1 + x_2 + \dots + x_n}{n}\right] \\ &= \frac{E[x_1] + \dots + E[x_n]}{n} \end{aligned}$$

Since  $X_i$ 's have exactly the same distribution, then

$$E[X_1] = \dots = E[X_n]$$

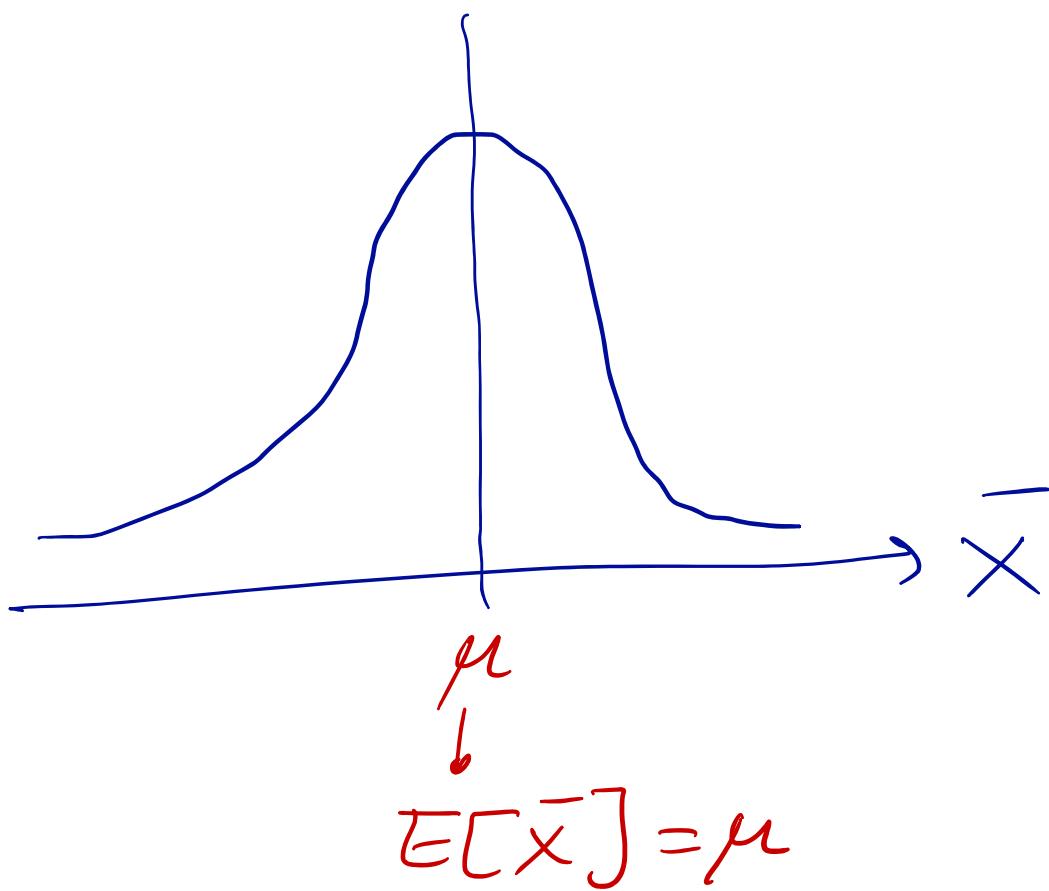
$$\bar{X}_i = \overbrace{\begin{cases} y_1 & \frac{1}{N} \\ y_2 & \frac{1}{N} \\ \vdots & \vdots \\ y_N & \frac{1}{N} \end{cases}}$$

$$\Rightarrow E[\bar{X}] = \frac{E[X_1] + E[X_2] + \dots + E[X_n]}{n}$$

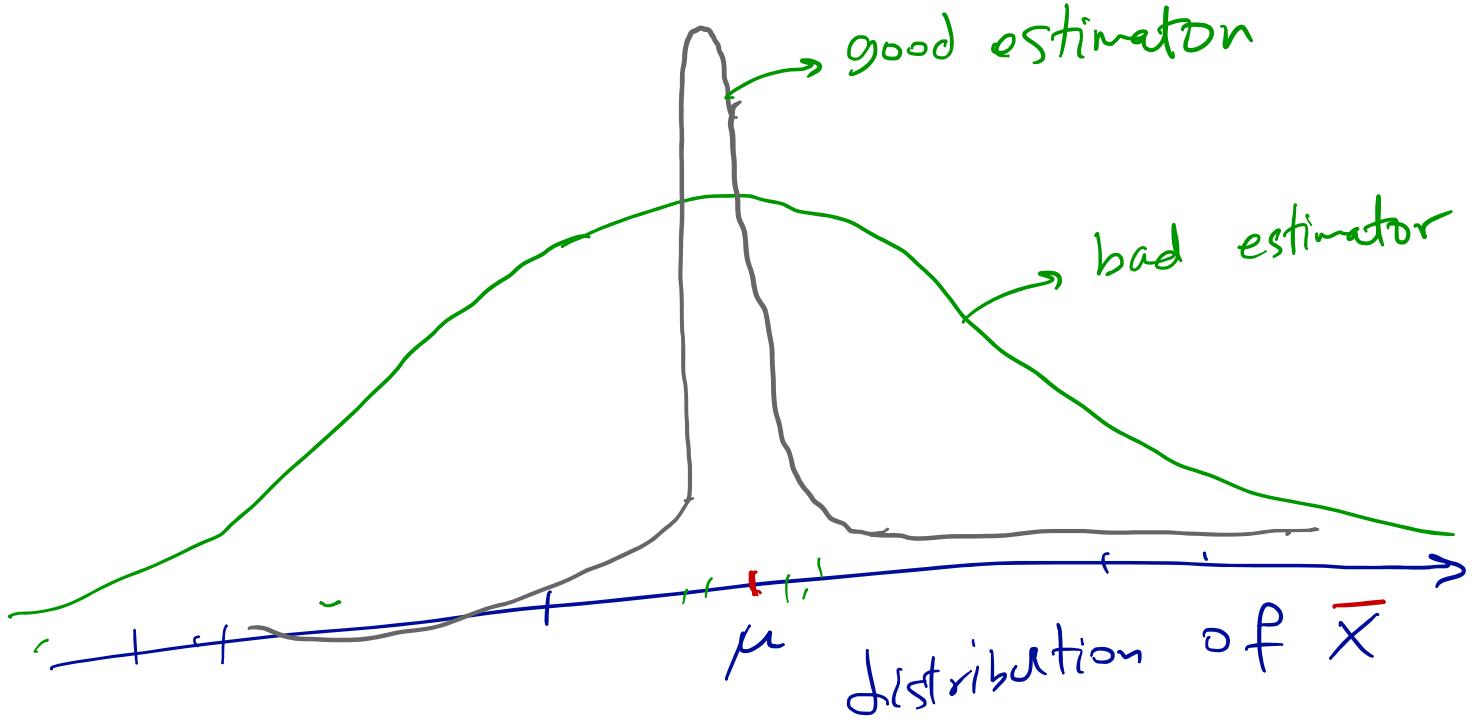
$$= E[X_1]$$

$$= \frac{1}{N} \cdot y_1 + \frac{1}{N} y_2 + \dots + \frac{1}{N} y_N$$

$$= \frac{y_1 + \dots + y_N}{N} = \mu$$



So,  $\bar{x}$  is on average a very good estimator because its expectation is  $\mu$ . The next question is how concentrated is  $\bar{x}$  around its average  $\mu$ .



In order to see how concentrated  $\bar{X}$  is around  $\mu$ , we need to compute ...

$$E[(\bar{X} - \mu)^2].$$

$$E[(\bar{X} - \mu)^2] \quad \left( \bar{X} = \frac{x_1 + \dots + x_n}{n} \right)$$

$$E E \left[ \left( \frac{x_1 + x_2 + \dots + x_n}{n} - \mu \right)^2 \right]$$

$$= E \left[ \left( \frac{x_1 + x_2 + \dots + x_n - n\mu}{n} \right)^2 \right]$$

$$= E \left[ \left( \frac{(x_1 - \mu) + (x_2 - \mu) + \dots + (x_n - \mu)}{n} \right)^2 \right]$$

$$= \frac{1}{n^2} E \left[ \left( (x_1 - \mu) + (x_2 - \mu) + \dots + (x_n - \mu) \right)^2 \right]$$

$\equiv -$

$$( \text{---} )^2 = \left( \sum_{i=1}^n (x_i - \mu) \right)^2$$

$$= \frac{1}{n^2} E \left[ \left( \sum_{i=1}^n (x_i - \mu) \right)^2 \right]$$

$$= \frac{1}{n^2} E \left[ \sum_{i=1}^n (x_i - \mu)^2 + 2 \sum_{\substack{i, j=1 \\ i \neq j}}^n (x_i - \mu)(x_j - \mu) \right]$$

$$= \frac{1}{n^2} \sum_{i=1}^n E[(x_i - \mu)^2] + \underbrace{\frac{2}{n^2} \sum_{\substack{i, j=1 \\ i \neq j}}^n E[(x_i - \mu)(x_j - \mu)]}_{\textcircled{2}}$$

---


$$\begin{aligned} (\alpha_1 + \alpha_2 + \dots + \alpha_n)^2 &= \alpha_1^2 + \alpha_2^2 + \dots + \alpha_n^2 \\ &\quad + 2\alpha_1\alpha_2 + 2\alpha_1\alpha_3 + \dots + 2\alpha_{n-1}\alpha_n \end{aligned}$$

②

$$E[(x_i - \mu)(x_j - \mu)] \quad (i \neq j)$$

when  $i \neq j$ ,  $x_i$  and  $x_j$  are independent. Hence

$$E[(x_i - \mu)(x_j - \mu)]$$

independence  $E[(x_i - \mu)] E[(x_j - \mu)]$

$$= (\underbrace{E[x_i] - \mu}_{=0})(\underbrace{E[x_j] - \mu}_{=0})$$

$$\Rightarrow ② = 0$$

if two random variables  $U, V$  are independent, then  $E[U \cdot V] = E[U]E[V]$   
 $E[g(U) \cdot h(V)] = E[g(U)] \cdot E[h(V)]$

$$\begin{aligned}
 ① &= \frac{1}{n^2} \sum_{i=1}^n E[(x_i - \mu)^2] \\
 &= \frac{1}{n^2} (E[(x_1 - \mu)^2] + E[(x_2 - \mu)^2] + \dots + E[(x_n - \mu)^2]) \\
 &= \frac{1}{n^2} \cdot n = E[(x_1 - \mu)^2] \\
 &\quad \uparrow x_i's \text{ are identically distributed (uniform)} \\
 &\equiv \frac{1}{n} E[(x_1 - \mu)^2]
 \end{aligned}$$

$$= \frac{1}{n} \left( \frac{1}{N} (y_1 - \mu)^2 + \frac{1}{N} (y_2 - \mu)^2 + \dots + \frac{1}{N} (y_N - \mu)^2 \right)$$

$$= \frac{1}{n} \cdot \sigma^2_{\text{population}}$$

$$\text{Where } \sigma^2_{\text{population}} = \frac{1}{N} \left( (y_1 - \mu)^2 + \dots + (y_N - \mu)^2 \right)$$

Variance of  $\bar{X}$ :

$$E[(\bar{X} - \mu)^2] = \textcircled{1} + \textcircled{2}$$

$\underset{\sigma^2}{=}$

$\frac{1}{n} \sigma_{\text{population}}^2 \xrightarrow{\text{fixed number}}$

---

So the variance of  $\bar{X}$

decays wrt the number of data points,  $n$ , like  $\frac{1}{n}$ .

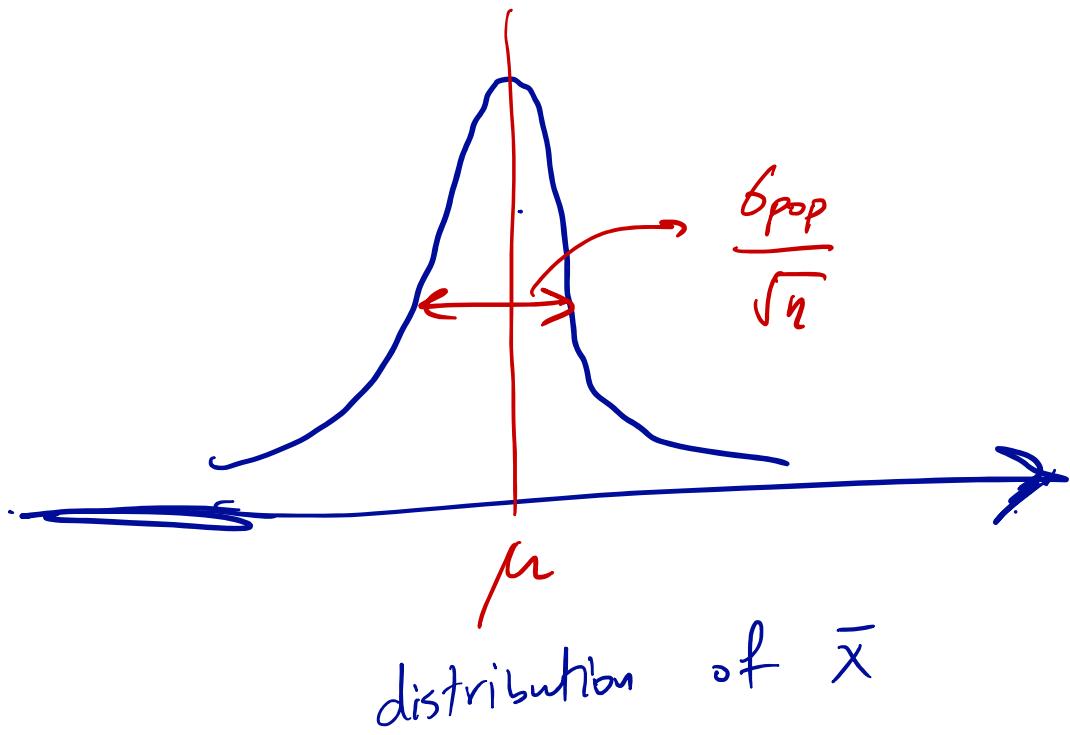
$\approx$

---

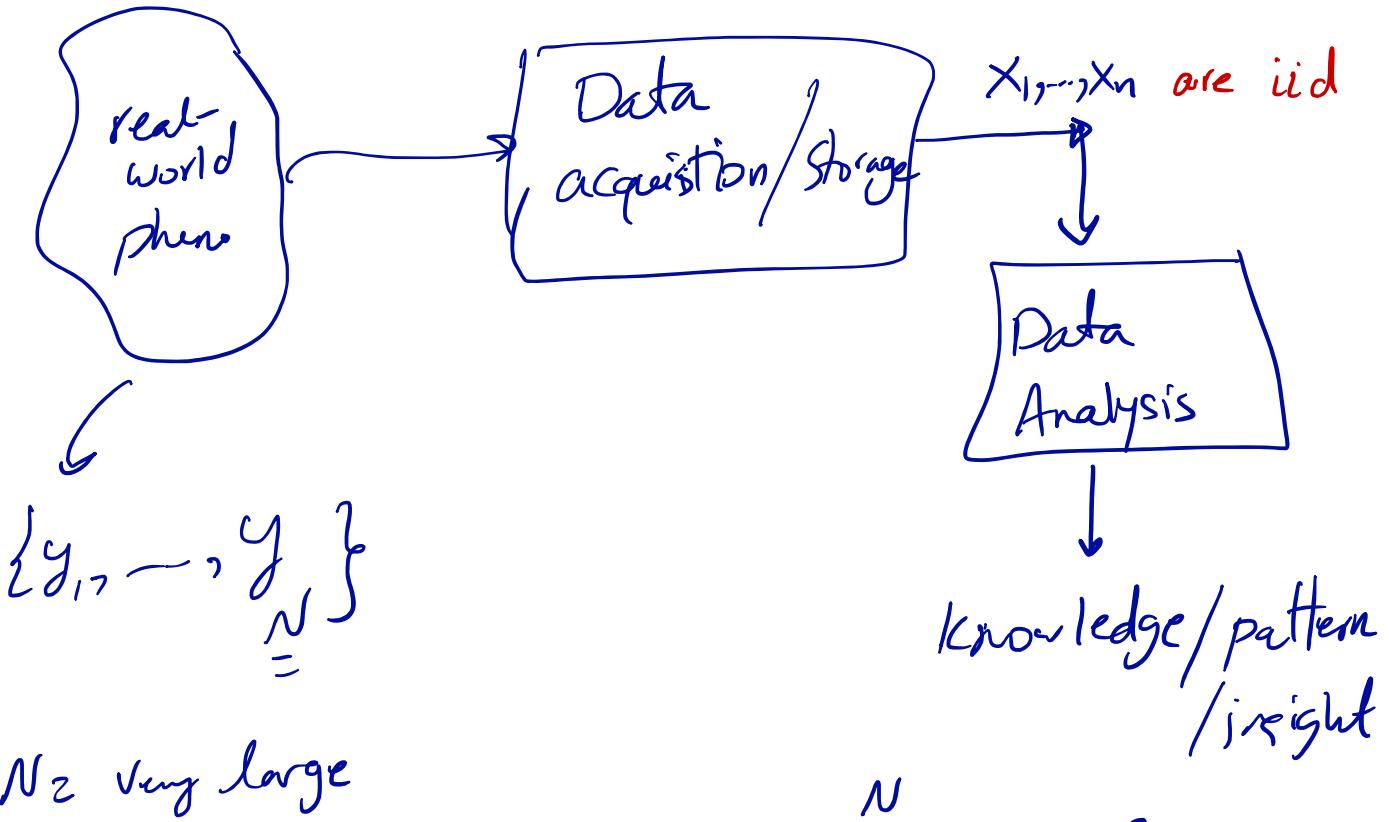
Standard deviation of  $\bar{X}$ :

$$\sqrt{E(\bar{X} - \mu)^2} = \frac{\sigma_{\text{pop}}}{\sqrt{n}}$$

Conclusion:



## Lecture 3 :

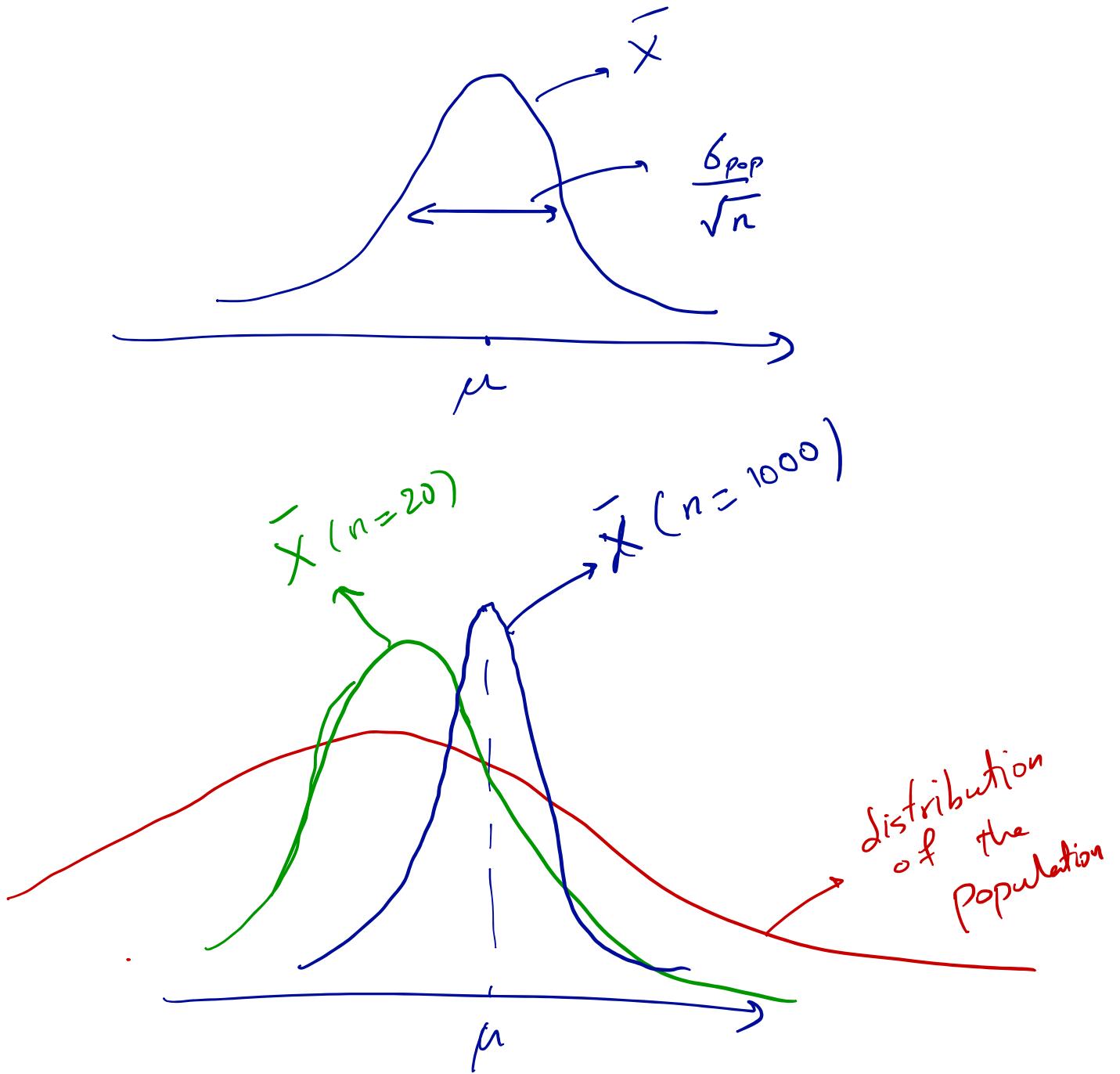


$$\mu = \frac{\sum_{i=1}^N y_i}{N}, \quad \sigma_{\text{pop}}^2 = \frac{\sum_{i=1}^N (y_i - \mu)^2}{N}$$

$$x_i = \begin{cases} y_1 & \frac{1}{N} \\ \vdots & \\ y_N & \frac{1}{N} \end{cases} \quad \bar{x} = \frac{x_1 + \dots + x_n}{n} \quad E[\bar{x}] = \mu$$

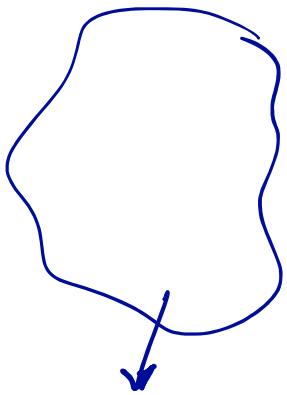
$\downarrow$

$$\text{Var}(\bar{x}) = \frac{\sigma_{\text{pop}}^2}{n}$$



Another example:

We'd like to find the percentage of people who would vote for a certain candidate in a state.



$$\{y_1, \dots, y_N\}$$

$$y_i = \begin{cases} 1 & \text{if } i \\ & \text{would vote} \\ & \text{for the} \\ & \text{candidate} \\ 0 & \text{o.w.} \end{cases}$$

$N$  = # of people  
in the state

$$\mu = \frac{y_1 + \dots + y_N}{N} \rightarrow \% \text{ of people}$$

who'd vote  
for the  
candidate

---

Let's summarize what we've learned so far. (and this is how we think of data acquisition/gathering procedure throughout the course).

In order to learn about a physical/real-world phenomenon, we will gather sample data.

Random Sample: statistically,

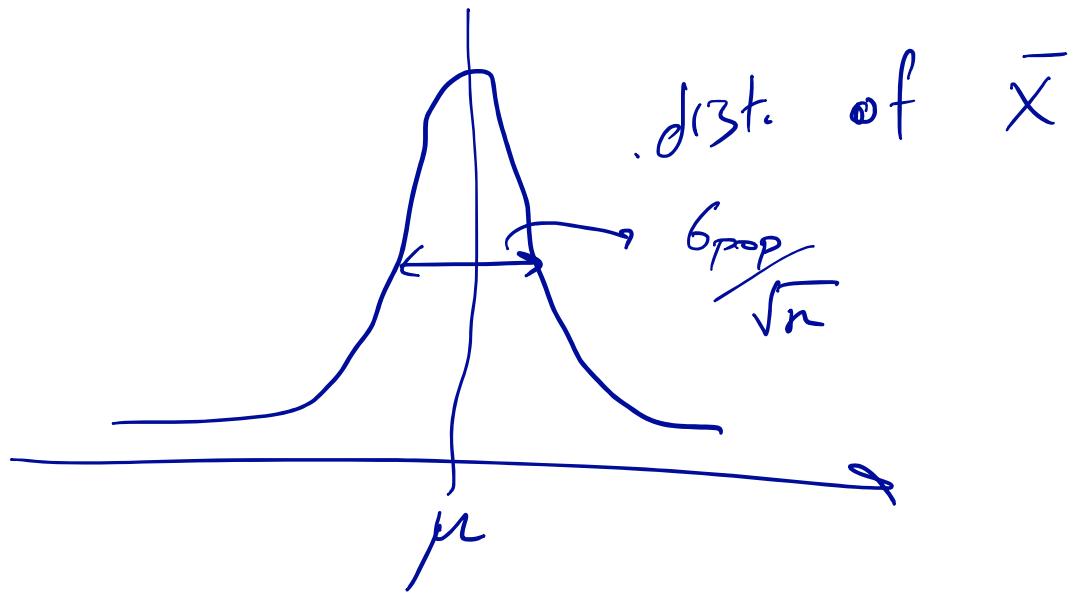
we model the sample data as

$x_1, x_2, \dots, x_n$  where  $x_i$ 's are assumed to be independent and identically distributed (i.i.d.).  
as the samples are generated from the same phenomenon.

↳  $x_i$ 's are random observations

- A statistic is any quantity whose value can be computed from sample data. Prior to obtaining data, there is uncertainty as to what value any particular statistic will result. Therefore, a statistic is typically a random variable.

e.g. Sample mean  $\bar{X}$  is a statistic ..



In order to understand more about the distribution of  $\bar{X}$ , we need to study the powerful framework of Central limit theorem.

### The Central Limit Theorem:

Take  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{dist}(\mu, \sigma^2)$

Then (informally):

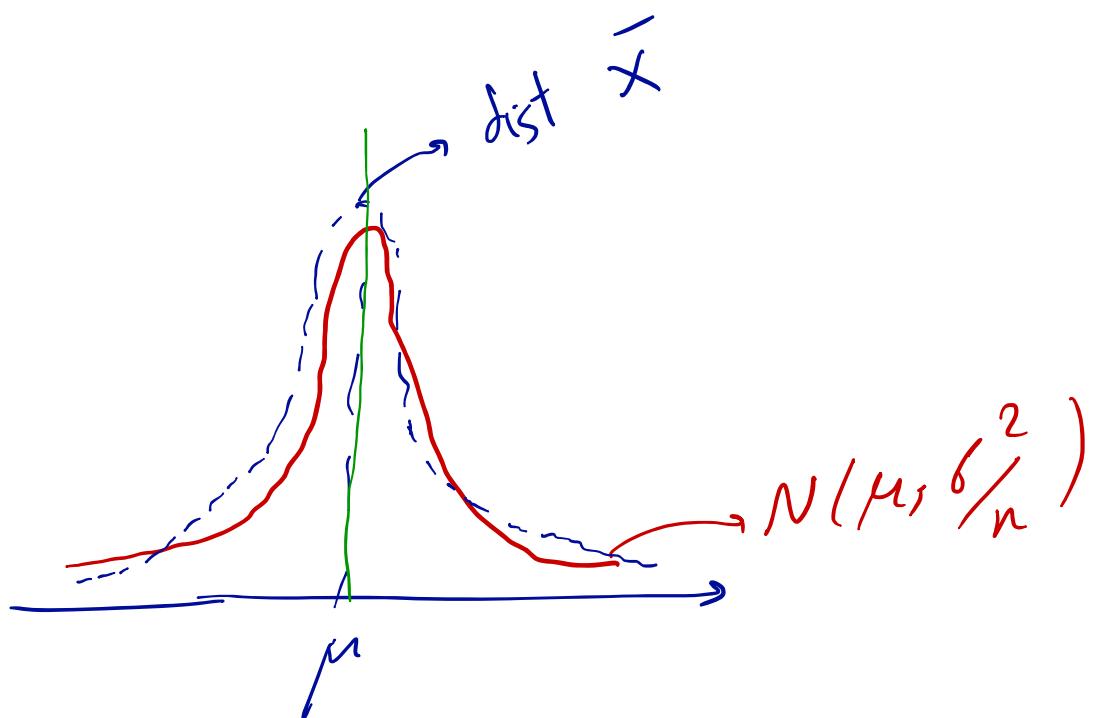
$\bar{X} \sim \text{gaussian} + \text{small error}$

we know that :

$$\begin{cases} E[\bar{X}] = \mu \\ \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \end{cases}$$

$$\bar{X} = \underbrace{\text{gaussian}}_{\downarrow} + \underbrace{\text{small error}}_{\frac{\sigma}{n}}$$

$$= N(\mu, \frac{\sigma^2}{n}) + \text{small error}$$



## Formal Statement ( CLT ):

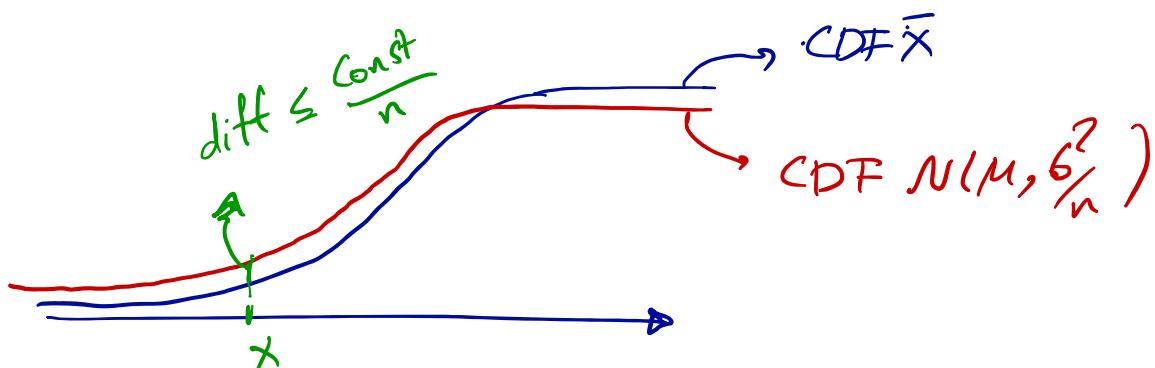
Let  $F_{\bar{X}}(x)$  be the CDF of  $\bar{X}$ :

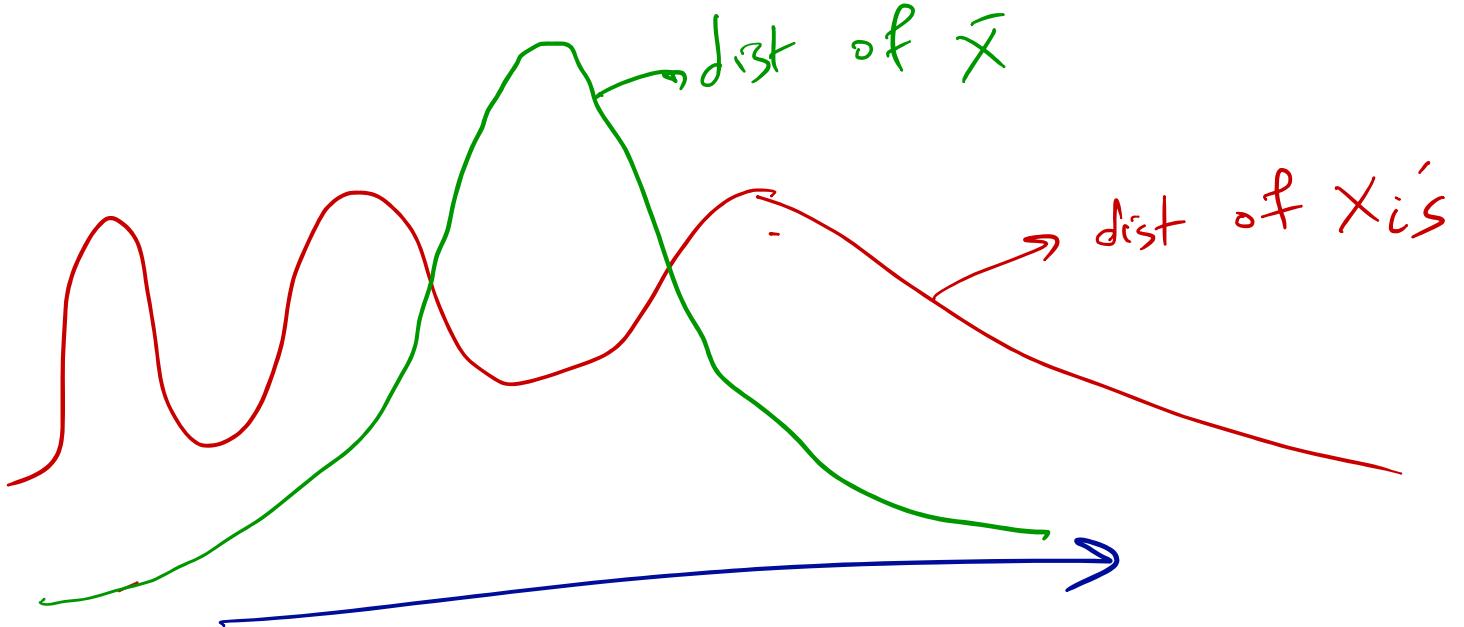
$$F_{\bar{X}}(x) = \Pr \{ \bar{X} \leq x \}$$

and Let  $\phi_{\mu, \sigma}(x)$  be the CDF of  $N(\mu, \sigma^2/n)$ . Then :

depends on  
the dis.  $X_i$ 's

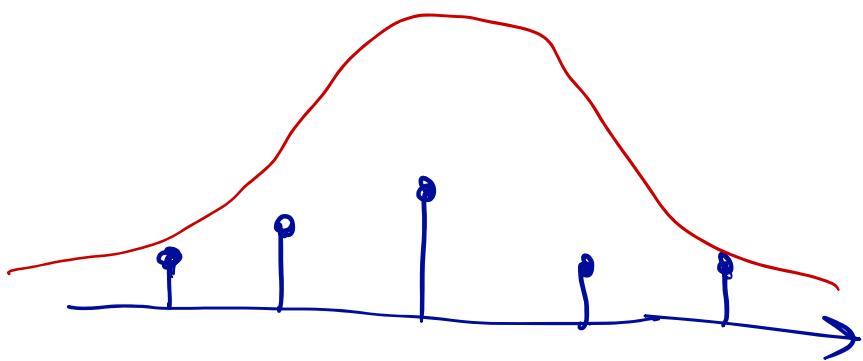
$$|F_{\bar{X}}(x) - \phi_{\mu, \sigma}(x)| \leq \frac{\text{Constant}}{n}$$






---

The theorem does not hold for  
Pdfs. (e.g. the dist of  $X_i$  could be discrete)

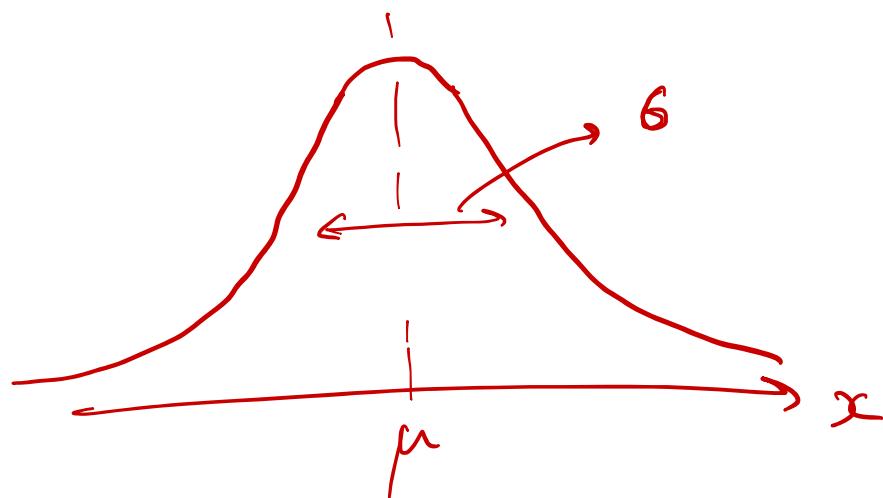


# Gaussian Distribution :

$$N(\mu, \sigma^2)$$

mean                      Variance

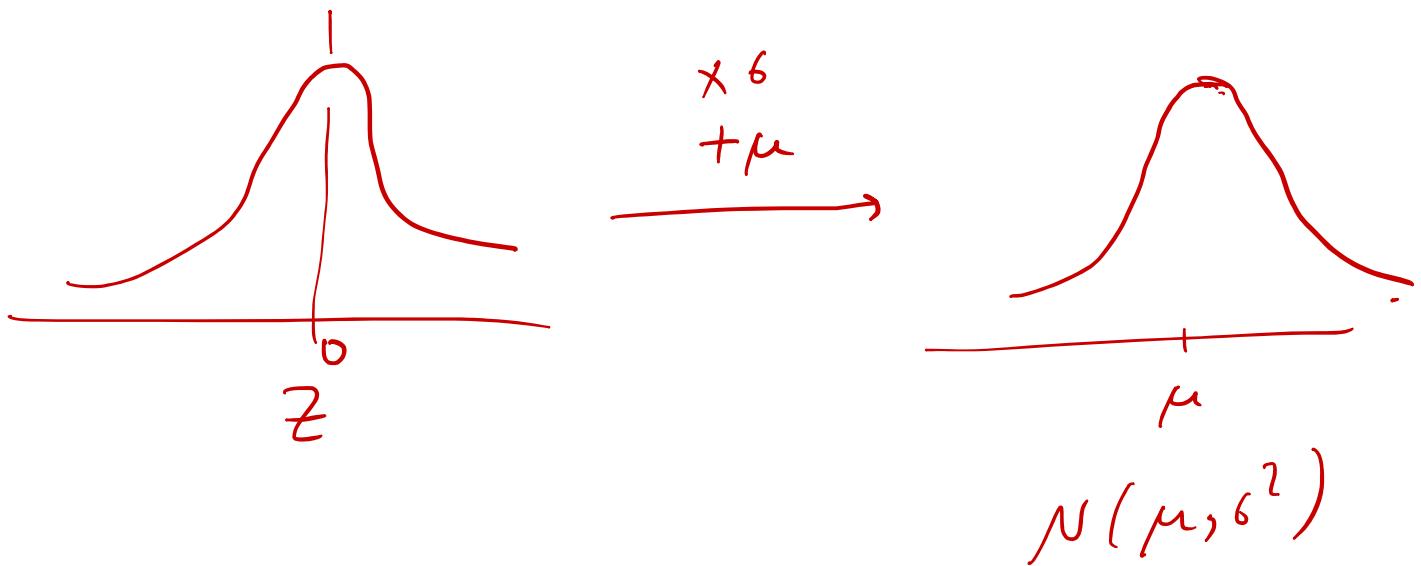
$$\text{Pf.} \quad \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



The standard gaussian or the normal distribution:

$$Z = N(0, 1)$$

$$N(\mu, \sigma^2) = \mu + \sigma \underbrace{Z}_{\sim}$$



Definitions:

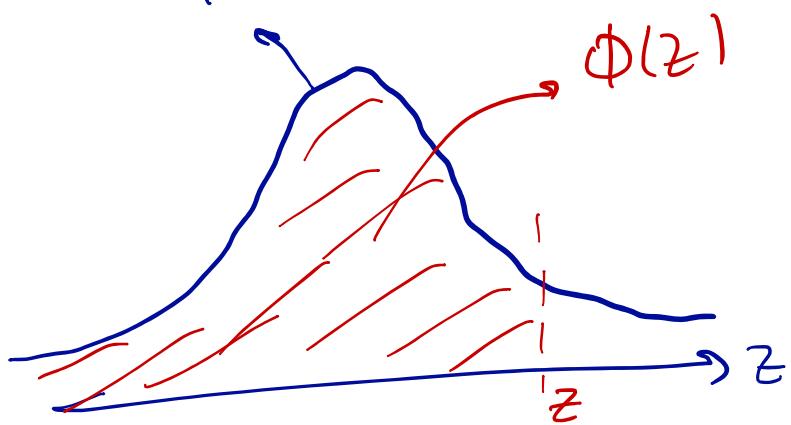
The Gaussian Cumulative Distribution Function:

Function:

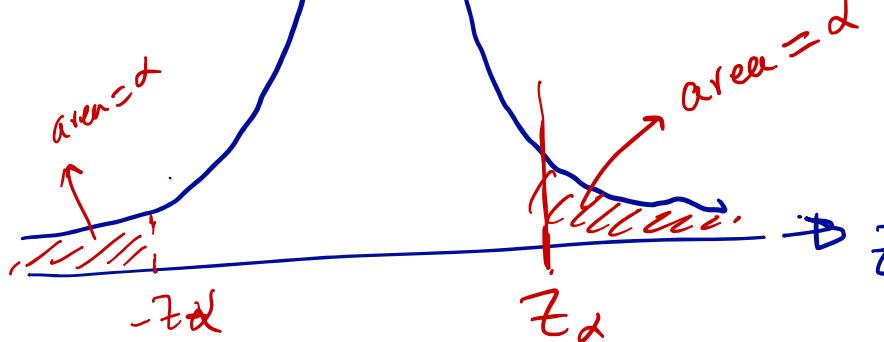
$$\Phi(z) = \Pr\{Z \leq z\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$

pdf of Normal dist



pdf of  $Z$



$z_\alpha$  is the ~~first~~ point on the  $Z$ -axis such that the area ~~is~~ under the Normal Pdf after  $z_\alpha$  is exactly  $\alpha$ .

## Confidence Intervals :

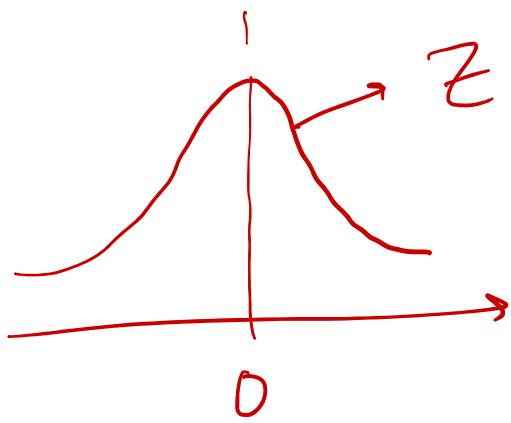
$\mu \rightarrow x_1, \dots, x_n \rightarrow \bar{x}$   
estimate for  $\mu$

we know that as we increase  
the number of data points,  $n$ ,

$\bar{x}$  will be a better estimate

for  $\mu$ . But we need guarantees.

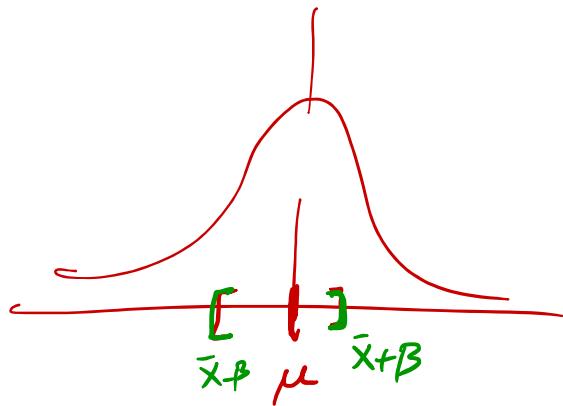
idea (i):  $\Pr\{\bar{X} = \mu\} = 0$



$$\begin{aligned}\Pr\{Z = 0\} \\ = 0\end{aligned}$$

$$(ii) \Pr \{ \mu \in [\bar{x} - \beta, \bar{x} + \beta] \}$$

typically  $\beta$   
 $\beta$  small.



$$\text{e.g. } \beta = 0.01$$

↳ if we prove that

$$\Pr \{ \mu \in [\cancel{\bar{x}} - 0.01, \cancel{\bar{x}} + 0.01] \} \geq 0.98$$

$\overbrace{[4.99, 5.01]}$

Problem:

Compute the value  $\beta$  such that

$$\Pr \left\{ \mu \in [\bar{x} - \underline{\beta}, \bar{x} + \underline{\beta}] \right\} = \overbrace{0.95}^{1-\alpha}$$

As an example, let's assume that

$X_i$ 's are generated iid from  
the gaussian  $N(\mu, \sigma^2)$  distribution.

Assumption:  $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$

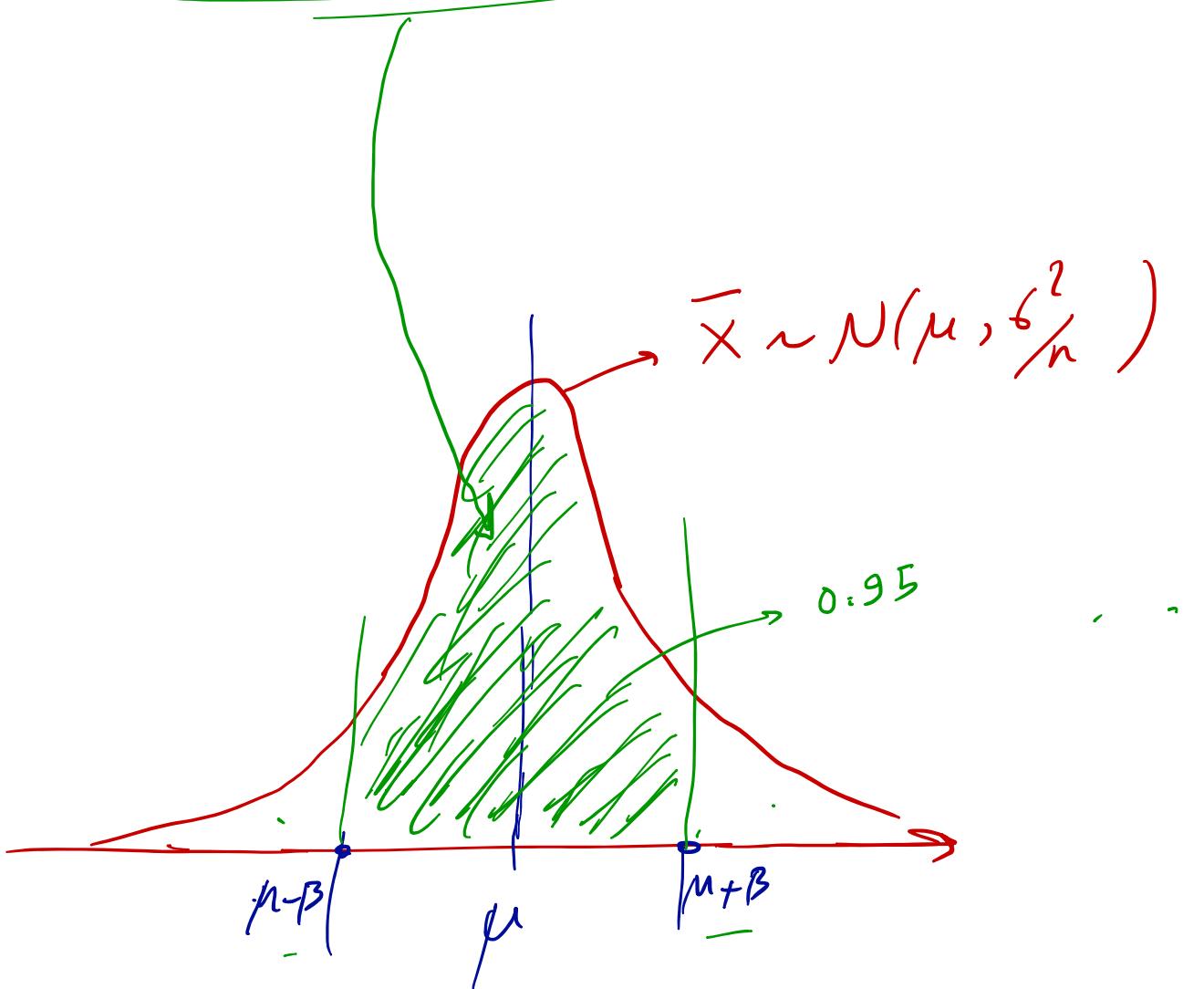
$$\bar{X} \sim N\left(\underline{\mu}, \frac{\sigma^2}{n}\right)$$

(sum of independent gaussians is a gaussian)

$$\Pr \left\{ \bar{x} - \beta \leq \mu \leq \bar{x} + \beta \right\} = 0.95$$



$$\Pr \left\{ \mu - \beta \leq \bar{x} \leq \mu + \beta \right\} = 0.95$$



$$\bar{x} - \beta \leq \mu \leq \bar{x} + \beta \quad (1)$$



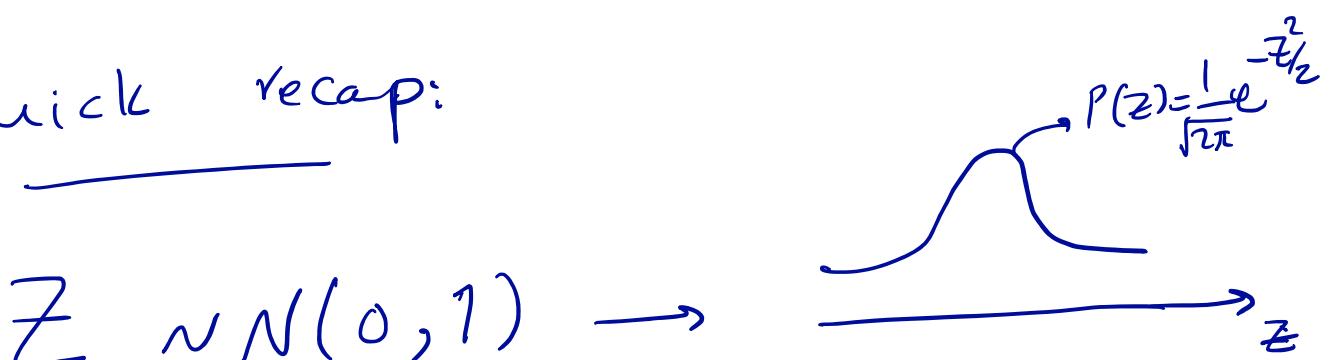
$$\boxed{\mu - \beta \leq \bar{x} \leq \mu + \beta}$$

$$(1) \rightarrow \mu \leq \bar{x} + \beta \Rightarrow \mu - \beta \leq \bar{x}$$

$$\mu \geq \bar{x} - \beta \Rightarrow \mu + \beta \geq \bar{x}$$

## Lecture 4:

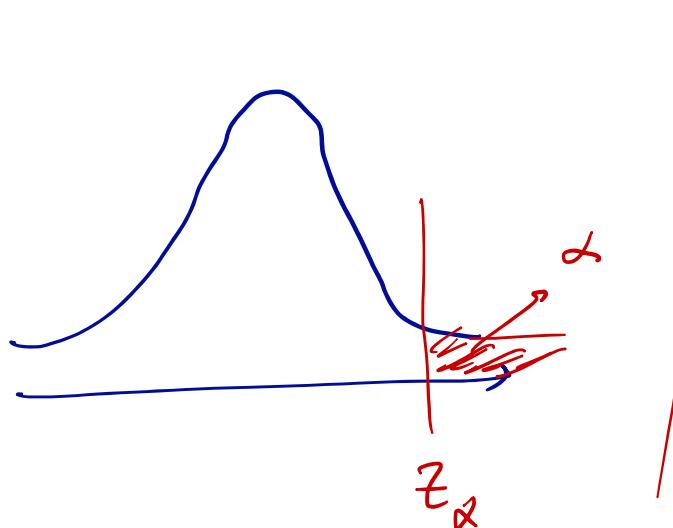
Quick recap:



$$\hookrightarrow N(\mu, \sigma^2) = \mu + \frac{\sigma}{\sqrt{2}} Z$$

In other words, if  $X \sim N(\mu, \sigma^2)$

then  $\frac{X - \mu}{\sigma} \sim Z$

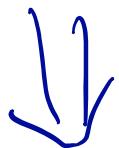


CDF of standard gaussian  
 $\Phi(z) = P(Z \leq z)$

from last lecture:

We'd like to compute the following  
when  $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ :  
 $\rightarrow (\sigma_{\text{pop}} = \sigma)$

$$\Pr \left\{ \mu \in [\bar{x} - \beta, \bar{x} + \beta] \right\} = 0.95$$



$$= \Pr \left\{ \mu - \beta \leq \bar{x} \leq \mu + \beta \right\}$$

$$= \Pr \left\{ -\beta \leq \bar{x} - \mu \leq \beta \right\}$$

$$= \Pr \left\{ -\frac{\beta}{\left(\frac{\sigma}{\sqrt{n}}\right)} \leq \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \leq \frac{\beta}{\left(\frac{\sigma}{\sqrt{n}}\right)} \right\}$$

What we've derived so far:

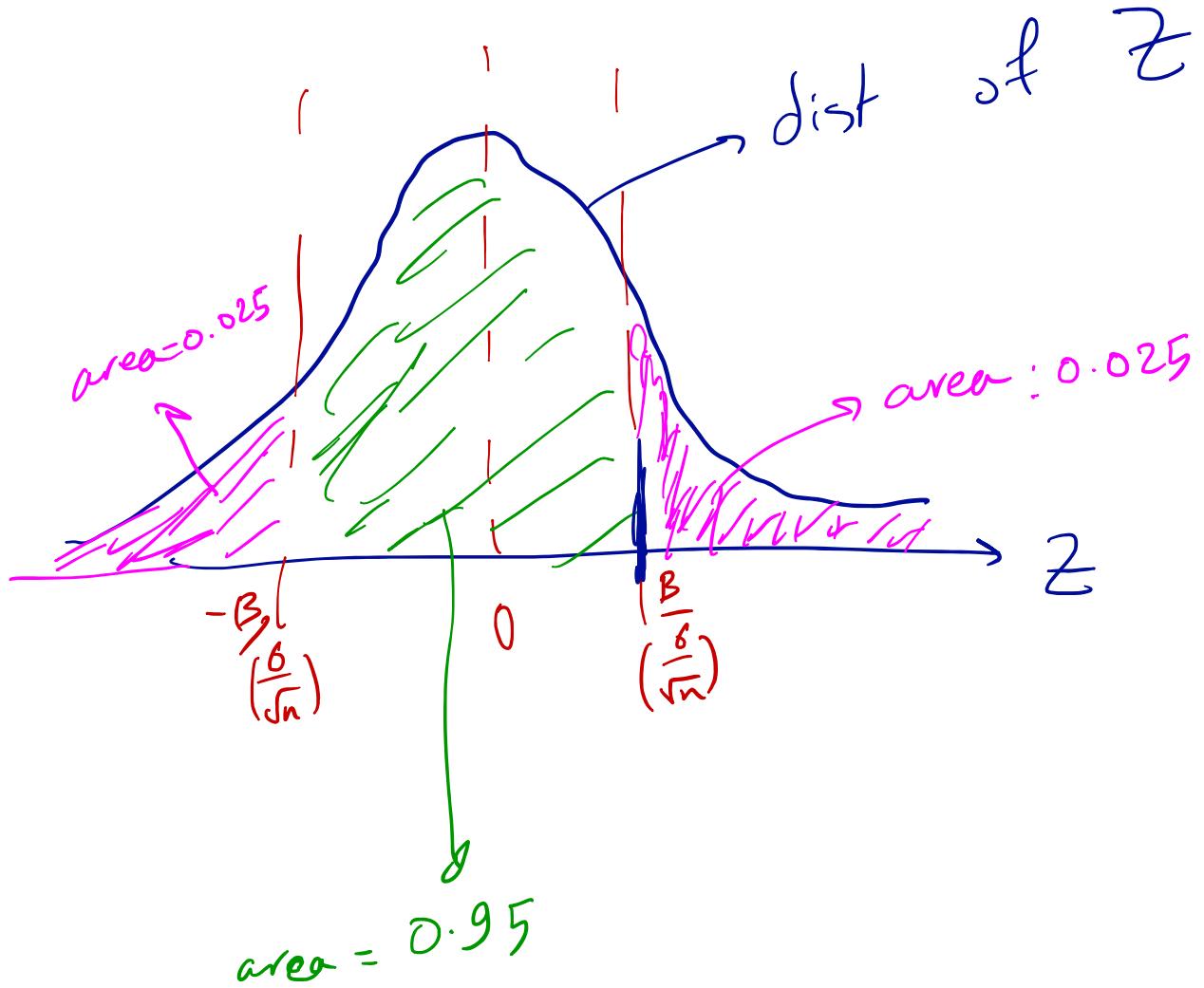
$$\Pr \{ \mu \in [\bar{x} - \beta, \bar{x} + \beta] \}$$
$$= \Pr \left\{ -\frac{\beta}{\left(\frac{6}{\sqrt{n}}\right)} \leq \frac{\bar{x} - \mu}{\left(\frac{6}{\sqrt{n}}\right)} \leq \frac{\beta}{\left(\frac{6}{\sqrt{n}}\right)} \right\}$$

What is the distribution of

$$\frac{\bar{x} - \mu}{\left(\frac{6}{\sqrt{n}}\right)} \sim Z = N(0, 1)$$

So our goal is to find  $\beta$   
such that

$$\Pr \left\{ -\frac{\beta}{\left(\frac{6}{\sqrt{n}}\right)} \leq Z \leq \frac{\beta}{\left(\frac{6}{\sqrt{n}}\right)} \right\} = 0.95$$



pre-computed or given

$$\Rightarrow \frac{\frac{B}{(\frac{6}{\sqrt{n}})}}{=} = \underbrace{Z_{0.025}}_{=} = \underbrace{1.96}_{=}$$

$$\Rightarrow B = Z_{0.025} \cdot \frac{6}{\sqrt{n}}$$

$$\Rightarrow \Pr \left\{ \mu \in \left[ \bar{x} - z_{0.025} \frac{6}{\sqrt{n}}, \bar{x} + z_{0.025} \frac{6}{\sqrt{n}} \right] \right\} = 0.95$$

Generally; assuming that

$x_i \sim N(\mu, \sigma^2)$ , then we

have:

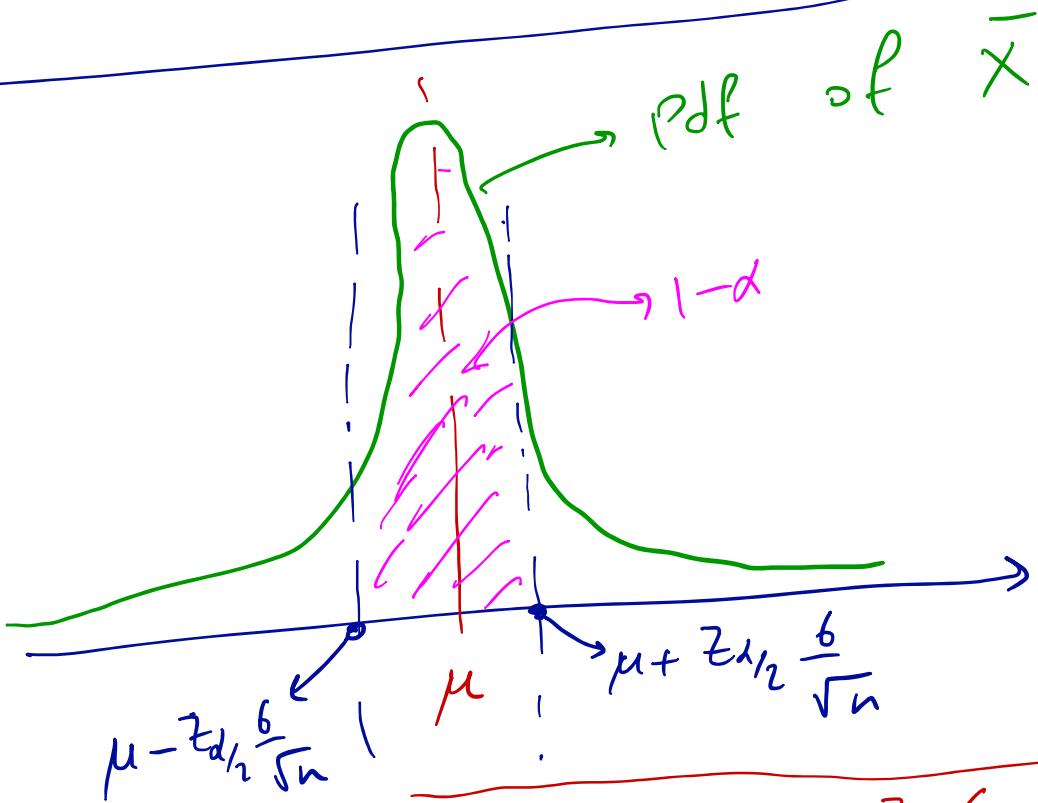
$$\Pr \left\{ \mu \in \left[ \bar{x} - \beta, \bar{x} + \beta \right] \right\} = 1 - \alpha$$

When  $\beta = z_{\alpha/2} \cdot \frac{6}{\sqrt{n}}$

$$\Pr \left\{ \mu \in \left[ \bar{x} - z_{\alpha/2} \frac{6}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{6}{\sqrt{n}} \right] \right\} = 1 - \alpha$$

$$X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \leftarrow$$



$$\Pr \left[ \mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = 1-\alpha$$

Next Question : What is the formula when  $X_i$ 's are not gaussian.

Let's assume  $X_i \stackrel{iid}{\sim} \text{dist}(\mu, \sigma^2)$

$$\Pr \left\{ \mu \in [\bar{x} - \beta, \bar{x} + \beta] \right\} = 1-\alpha$$

(when  $n$  is large, rule of thumb for  $n$  being large,  $n > 30$ ):

$\bar{X}$  is (approximately) gaussian.

$$\bar{X} \xrightarrow{\text{appr.}} N(\mu, \sigma^2/n)$$

---

Since  $N(\mu, \sigma^2/n)$  was also the distribution of  $\bar{X}$  in the previous example (when everything was gaussian), all the formulas are also (approximately) valid in this case.

$X_i \stackrel{iid}{\sim} \text{dist}(\mu, \sigma^2)$  then

$$\Pr \left\{ \mu \in \left[ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] \right\} \approx 1-\alpha$$

Definition:

A  $100(1-\alpha)\%$  confidence interval

for a population with mean  $\mu$

and (known) variance  $\sigma^2$  is

given by

$$\underbrace{\text{width of the confidence interval}}_{= 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}} = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\left[ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

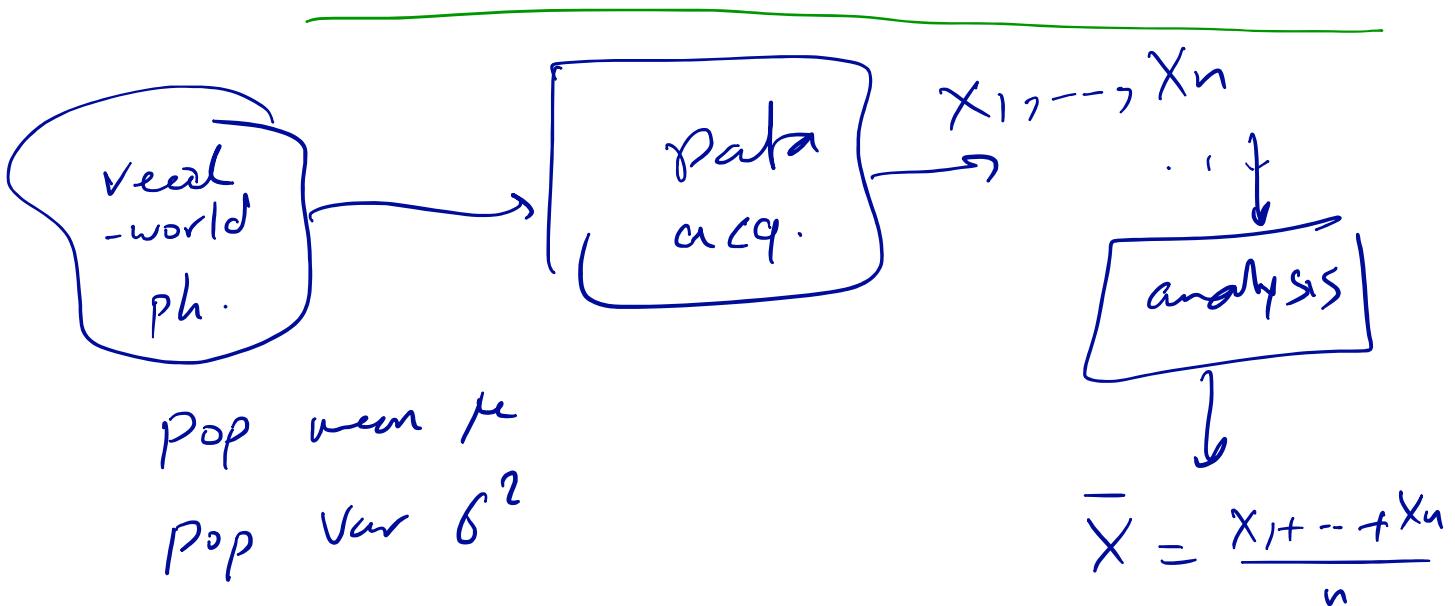
i.e.

$$\Pr \left\{ \mu \in [ \dots ] \right\} \stackrel{n \text{ is large}}{\approx} 1-\alpha$$

Exercise: Find the smallest value of  $n$  for which we can estimate the population mean  $\mu$ , using an interval of width  $w$ , with confidence at least  $1-\alpha$ ?

$$2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = w \implies \sqrt{n} = \frac{2 z_{\alpha/2} \sigma}{w}$$

$$\implies n = \left( \frac{2 z_{\alpha/2} \sigma}{w} \right)^2$$



There is still a PROBLEM  
with all the previous formulas  
when working with real data.

Confidence interval

problem is that we don't know  $\sigma$ .

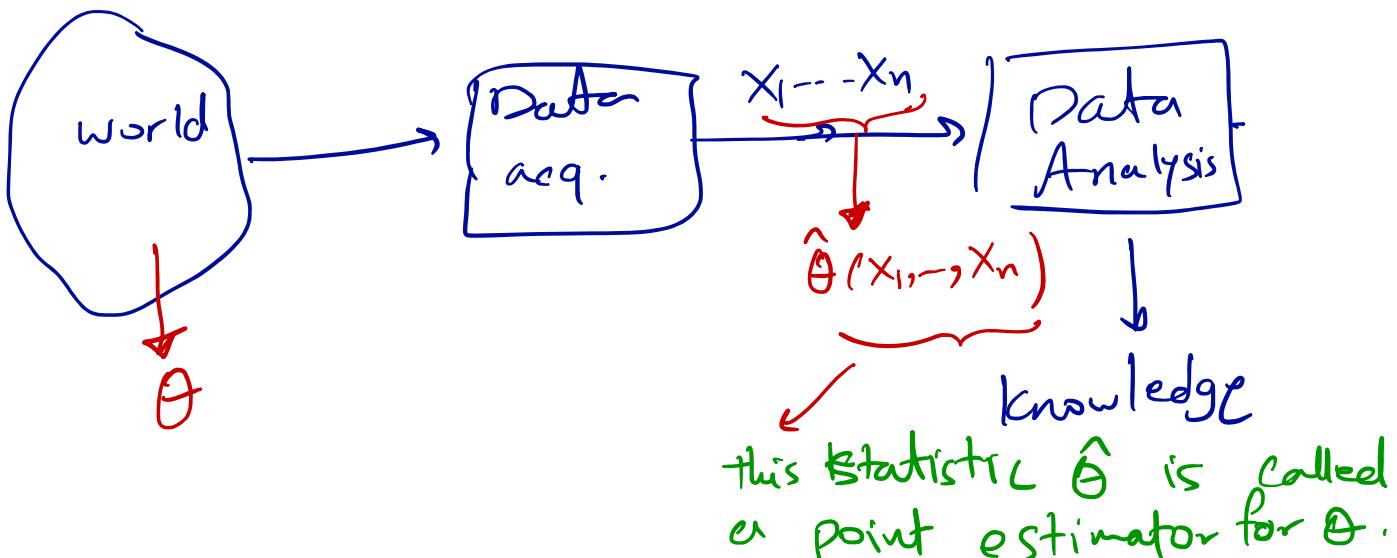
$$\left[ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Data:  $X_1, \dots, X_n$

In order to resolve the issue  
of not knowing  $\sigma$ , we will

estimate it using sample data

$X_1, \dots, X_n$



Statistically speaking, "knowledge" means that we estimate some statistic/parameter from data, or learn some pattern from data.

### Point estimation :-

A point estimate of a parameter  $\theta$  is a single number that can be regarded as a sensible value for  $\theta$ . A point estimate

is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the point estimator for  $\theta$ .

---

Example 1 (Population mean)

$\theta$  could be the population mean.

$$\hat{\theta}(x_1, x_n) = \frac{x_1 + \dots + x_n}{n}$$

Example 2 (Population Variance)

$\theta$  is the variance of the population, i.e  $\sigma^2$ .

So now, we are going to design an "estimator" for the Variance of the population.

$$\hat{\theta}(x_1, \dots, x_n)$$

Variance  $\rightarrow E[(x_i - \mu)^2]$

$E[(x_i - \bar{x})^2]$

↓ estimate

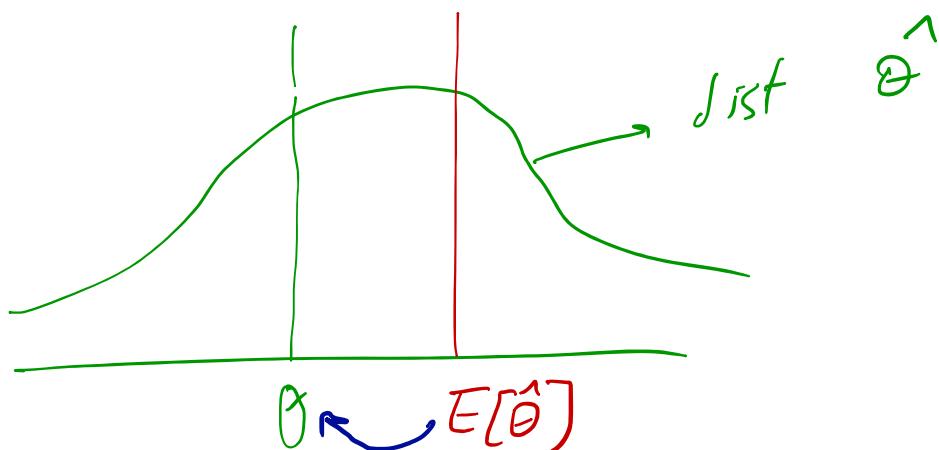
estimate of  $\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$

estimator  $\hat{\theta}(x_1, \dots, x_n)$

$$\underline{= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

The first step in evaluating the performance of a "designed" estimator is to see if it is biased / unbiased.

$$\theta \rightarrow E[\hat{\theta}(x_1, \dots, x_n)] = \theta$$



# Principle of Unbiased Estimation:

When choosing among several different estimators for a parameter  $\theta$ , select the one that is unbiased.

$$E[\hat{\theta}(x_1, \dots, x_n)] = \theta$$

$\uparrow$   
unbiased

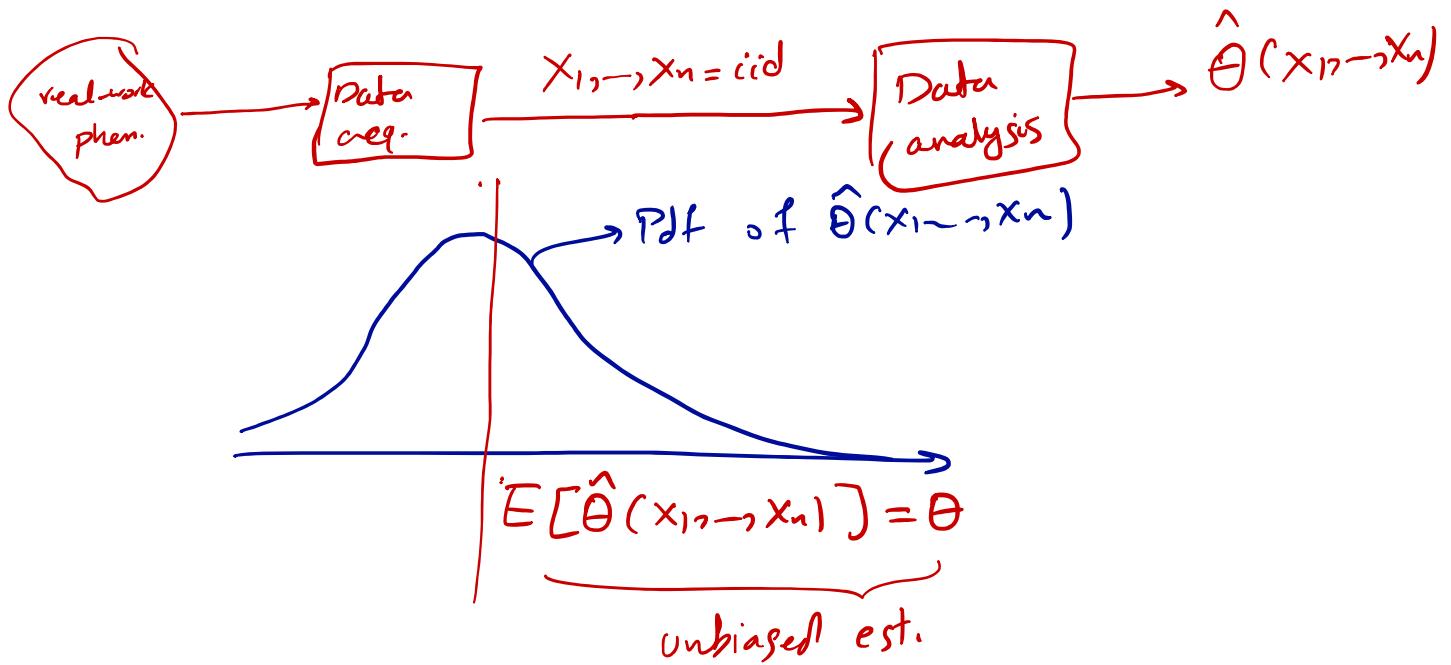
Definition: A point estimator  $\hat{\theta}$  is said to be an unbiased estimator of  $\theta$  if  $E[\hat{\theta}] = \theta$  for every possible value of  $\theta$ . If  $\hat{\theta}$  is not biased, the difference  $E[\hat{\theta}] - \theta$

is called the bias of  $\hat{\theta}$ .

Example: The average mean  $\hat{\theta}(x_1 - x_n) = \bar{x}$   
is an unbiased estimate of the  
population mean  $\mu$ .

## Lecture 5 :

$$\hat{\theta} \xleftarrow{\text{estimate}} \hat{\theta}(x_1, \dots, x_n)$$



$$\bar{x} \rightarrow \text{we've already studied} \rightarrow E[\bar{x}] = \mu$$

$$\text{Var}[\bar{x}] = \frac{\sigma^2}{n}$$

$$\Pr \left\{ \mu \in \left[ \bar{x} - Z_{d/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{d/2} \frac{\sigma}{\sqrt{n}} \right] \right\} \approx 1 - \alpha$$

$$\hat{\sigma}^2(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\hat{\sigma}^2 = \underline{E}[(x_i - \mu)^2]$$

In order to check if an estimator  $\hat{\theta}$  is unbiased, we should compute the mean of the estimator.

$$\underline{E[\hat{\theta}] \stackrel{?}{=} \theta}$$

$$\begin{aligned}
 & E[\hat{\theta}] = E\left[\sum_{i=1}^n (x_i - \bar{x})^2\right] \\
 & = E\left[\sum_{i=1}^n \left\{x_i^2 + \bar{x}^2 - 2x_i\bar{x}\right\}\right] \quad \cancel{\frac{1}{n} \cdot n E[\bar{x}^2]} \\
 & = \underbrace{E\left[\sum_{i=1}^n x_i^2\right]}_{(1)} + E[\bar{x}^2] - 2 \underbrace{E\left[\sum_{i=1}^n x_i \bar{x}\right]}_{(2)} - 2 \underbrace{E\left[\bar{x} \left(\sum_{i=1}^n x_i\right)\right]}_{(3)} \\
 & = E[x_i^2] + E[\bar{x}^2] - 2 E[\bar{x} \left(\sum_{i=1}^n x_i\right)] \\
 & = E[x_1^2] + E[\bar{x}^2] - 2 \overbrace{E[\bar{x}^2]}^{E[\bar{x}]^2} \\
 & = E[x_1^2] - E[\bar{x}^2]
 \end{aligned}$$

$$\textcircled{1} \quad E\left[\frac{1}{n} \sum_{i=1}^n x_i^2\right] = \frac{1}{n} \left( \underbrace{E[x_1^2]}_{=E[x^2]} + \underbrace{E[x_2^2]}_{=E[x^2]} + \dots + \underbrace{E[x_n^2]}_{=E[x^2]} \right)$$

$$= \frac{1}{n} (n \cdot E[x^2])$$

$$= E[x^2]$$

$$\textcircled{3} \quad E\left[\frac{1}{n} \sum_{i=1}^n \bar{x} x_i\right] = E\left[\bar{x} \sum_{i=1}^n x_i\right]$$

$$= E\left[\bar{x} \left(\frac{1}{n} \sum_{i=1}^n x_i\right)\right]$$

$$= E[\bar{x}^2]$$

---


$$E[\hat{\sigma}^2] = E[x_1^2] - E[\bar{x}^2]$$

$$x_i \sim \text{dist}(\mu, \sigma^2)$$

$$E[x_1^2] = \sigma^2 + \mu^2$$

$$\begin{cases} \text{Var}(x_1) = \sigma^2 \\ = E[x_1^2] - \mu^2 \end{cases}$$

---


$$\text{Formula for Variance} = \text{Var}(X) = E[X^2] - (E[X])^2$$

$$E[\bar{X}^2] = \frac{\sigma^2}{n} + \mu^2$$

$$\text{Var}(\bar{X}) = E[\bar{X}^2] - \overbrace{[E[\bar{X}]]^2}^{\mu^2}$$

$\mu$   
 $\sigma^2 / n$

---

$$\hat{E}[\hat{\sigma}^2(x_1, \dots, x_n)]$$

$$= E[x_1^2] - E[\bar{x}^2]$$

$$= \sigma^2 + \mu^2 - (\mu^2 + \frac{\sigma^2}{n})$$

$$= \underline{\sigma^2 \left(1 - \frac{1}{n}\right)} \neq \sigma^2$$

$\Rightarrow \hat{\sigma}^2(x_1, \dots, x_n)$  is a biased estimator.

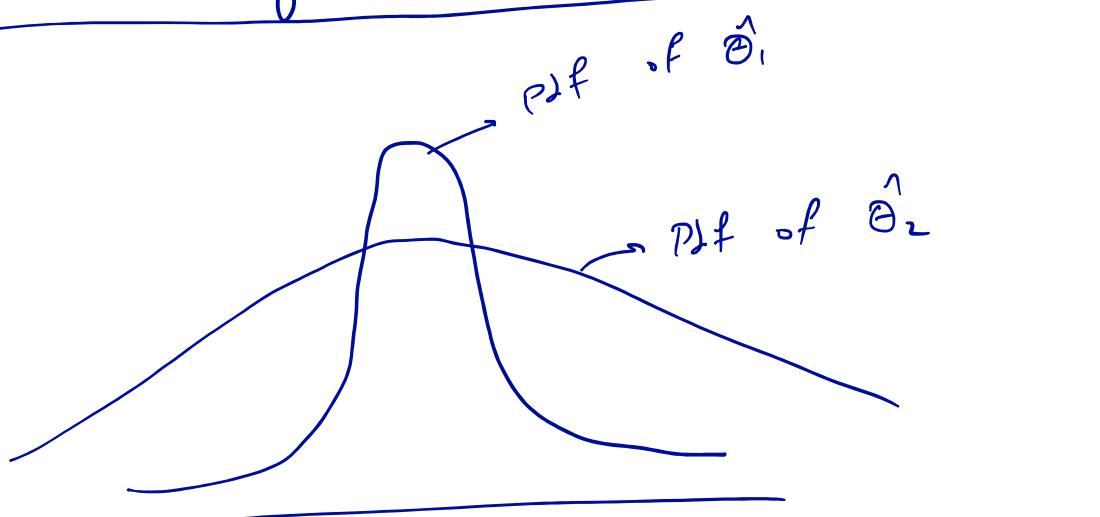
$\Rightarrow$  If we multiply  $\hat{\sigma}^2$  by  $\frac{1}{(1-\frac{1}{n})}$  then  
 the result will be an unbiased estimate.

Define:

$$\hat{\theta}(x_1, \dots, x_n) = \frac{1}{(1-\frac{1}{n})} \hat{\sigma}^2(x_1, \dots, x_n)$$
$$= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

then  $\hat{\theta}$  is an unbiased estimator for the variance  $\sigma^2$ .

Estimating with Minimum Variance:



Assume both estimator are unbiased (with mean  $\theta$ ).

Which one should we use?

**Principle of Minimum Variance Unbiased Estimation:**  
Among all estimators of  $\theta$  that are unbiased, choose the one that has minimum variance. The resulting  $\hat{\theta}$  is called the MVUE of  $\theta$ . (MVUE)



Systematic approaches for parameter estimation:

Basic assumption: Assume that we have some information about the pdf of the data distribution, i.e.

parameters that specify the distribution.

$$\Pr\{X_i = x\} = f(x | \theta_1, \dots, \theta_m)$$

e.g. if  $X_i \sim N(\mu, \sigma^2)$

$$\begin{aligned} P(X_i = x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= f(x | \mu, \sigma) \end{aligned}$$

e.g. If  $X_i \sim \text{exponential}(\lambda)$

$$\Pr\{X_i = x\} = e^{-\lambda x} \mathbb{I}\{x \geq 0\}$$

## The method of moments:

Definition: The  $k$ -th moment of a random variable  $X$  is defined as

$$\mu_k = \underline{E[X^k]}$$

e.g. first moment is just the mean:

$$\mu_1 = E[X].$$

Exercise: Given a sample  $x_1, x_n$  drawn independently from an identical distribution, design an unbiased estimator of  $\mu_k$  of the underlying distribution:

$$\hat{\mu}_k = \frac{x_1^k + x_2^k + \dots + x_n^k}{n}$$

The method of moments: Assume that  $x_1, \dots, x_n \stackrel{iid}{\sim} f(x | \theta_1, \dots, \theta_m)$ . We know  $f$  but we do not know the parameters  $\theta_1, \dots, \theta_m$ , and we'd like to estimate the parameters from data (using the moments of the distribution).

(1) Consider the first  $m$  moments of the distribution:

$$\mu_1 = E[X]$$

$$\mu_2 = E[X^2]$$

:

$$\mu_m = E[X^m].$$

assume that we can write

$$\theta_1 = g_1(\mu_1, \dots, \mu_m)$$

$$\theta_2 = g_2(\mu_1, \dots, \mu_m) \xrightarrow{\text{estimates be core}}$$

:

$$\theta_m = g_m(\mu_1, \dots, \mu_m)$$

$$\hat{\theta}_1 = \hat{g}_1(\hat{\mu}_1, \dots, \hat{\mu}_m)$$

$$\hat{\theta}_2 = \hat{g}_2(\hat{\mu}_1, \dots, \hat{\mu}_m)$$

:

$$\hat{\theta}_m = \hat{g}_m(\hat{\mu}_1, \dots, \hat{\mu}_m)$$

Example 1:  
 Let's assume  $x_i \sim \text{exponential}(\lambda)$ . We'd like to estimate  $\lambda$ .

$$\Pr\{X_i = x\} = f(x | \lambda) = \lambda e^{-\lambda x} \underbrace{\mathbb{1}_{\{x \geq 0\}}}_{\begin{array}{l} \mathbb{1}_{\{x > 0\}} \\ = \begin{cases} 1 & x \geq 0 \\ 0 & \text{o.w.} \end{cases} \end{array}}$$

$$\rightarrow \mu_1 = E[X] = \frac{1}{\lambda}$$

$$\hookrightarrow \mu_1 = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{\mu_1}$$

$$\hookrightarrow \left( \theta_1 = g_1(\mu_1) \right)$$

$$g_1(z) = \frac{1}{z}$$

$$\lambda = \frac{1}{\mu_1}$$

$$\Rightarrow \mu_1 \xrightarrow[\text{from data}]{\text{estimate}} \hat{\mu}_1 = \frac{x_1 + \dots + x_n}{n}$$

$$\Rightarrow \hat{\lambda} = \frac{1}{\hat{\mu}_1} = \frac{n}{x_1 + \dots + x_n}$$

final estimator :

$$\hat{\lambda}(x_1, \dots, x_n) = \frac{n}{x_1 + \dots + x_n} \left( = \frac{1}{\hat{\mu}_1} \right)$$

The estimator is biased because

$$E\left[\frac{1}{\bar{x}}\right] \neq \frac{1}{E[\bar{x}]}$$

$$\hat{\lambda} = \frac{1}{\hat{\mu}_1} \rightarrow E[\hat{\lambda}] = E\left[\frac{1}{\hat{\mu}}\right] \\ \cdot \neq \frac{1}{E[\hat{\mu}]} = \frac{1}{n}$$

Example 2:

$$x_1, \dots, x_n \sim N(\mu, \sigma^2) \rightarrow f(x | \theta_1, \theta_2)$$

$$\begin{cases} \mu_1 = E[X] = \theta_1 \\ \mu_2 = E[X^2] = \theta_1^2 + \theta_2^2 \end{cases}$$

$$\boxed{\begin{aligned} \text{Var}(X) &= E[X^2] - \underbrace{E[X]^2}_{\theta_1} \\ &= \theta_2^2 \end{aligned}}$$

$$\Rightarrow \begin{cases} \theta_1 = \mu_1 \\ \theta_2 = \sqrt{\mu_2 - \theta_1^2} = \sqrt{\mu_2 - \mu_1^2} \end{cases}$$

$$\text{recall: } \mu_1 \longrightarrow \hat{\mu}_1 = \frac{x_1 + \dots + x_n}{n}$$

$$\mu_2 \longrightarrow \hat{\mu}_2 = \frac{x_1^2 + \dots + x_n^2}{n}$$

$$\hat{\mu} = \hat{\theta}_1 = \hat{\mu}_1 = \frac{x_1 + \dots + x_n}{n}$$

$$\hat{\sigma} = \hat{\theta}_2 = \sqrt{\hat{\mu}_2 - \hat{\mu}_1^2}$$

$$= \sqrt{\frac{x_1^2 + \dots + x_n^2}{n} - \left( \frac{x_1 + \dots + x_n}{n} \right)^2}$$

simplify

$$= \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

## Lecture 6 :

Method of moments (cont'd) :

Example (Gamma Distribution) :

$$\text{Data: } x_1, \dots, x_n \stackrel{\text{iid}}{\sim} f(x|\alpha, \beta) = \frac{\alpha^{\alpha-1} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

$$\rightarrow (\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx)$$

for the Gamma dis. we know that:

$$\begin{aligned} \mu_1 &= E[X] = \frac{\alpha}{\beta} & \Rightarrow \begin{cases} \alpha = g_1(\mu_1, \mu_2) \\ \beta = g_2(\mu_1, \mu_2) \end{cases} \\ \mu_2 &= E[X^2] = \frac{\alpha(\alpha+1)}{\beta^2} \end{aligned}$$

$$\begin{aligned} \mu_1 &= \frac{\alpha}{\beta} \quad \Rightarrow \quad \mu_2 = \frac{\alpha^2}{\beta^2} + \frac{\alpha}{\beta^2} \\ \mu_2 &= \frac{\alpha(\alpha+1)}{\beta^2} \quad \Rightarrow \quad \mu_2 = \mu_1^2 + \frac{\alpha}{\beta^2} \end{aligned}$$

$$\Rightarrow \frac{\alpha}{\beta^2} = \mu_2 - \mu_1^2$$

$$\Rightarrow \frac{\alpha}{\beta} = \mu_1 \quad \Rightarrow \frac{1}{\beta} = \frac{\mu_2 - \mu_1^2}{\mu_1}$$

$$\beta = \frac{\mu_1}{\mu_2 - \mu_1^2}$$

$$\mu_1 = \frac{d}{\beta} \Rightarrow d = \mu_1 \cdot \beta$$

$$\Rightarrow d = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

$$\left. \begin{array}{l} d = g_1(\mu_1, \mu_2) = \frac{\mu_1^2}{\mu_2 - \mu_1^2} \\ \beta = g_2(\mu_1, \mu_2) = \frac{\mu_1}{\mu_2 - \mu_1^2} \end{array} \right\}$$

Now, we estimate  $\mu_1$  and  $\mu_2$  using data:

$$\left. \begin{array}{l} \hat{\mu}_1 = \frac{x_1 + \dots + x_n}{n} \\ \hat{\mu}_2 = \frac{x_1^2 + \dots + x_n^2}{n} \end{array} \right\} \Rightarrow \begin{aligned} \hat{d} &= \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} \\ \rightarrow \hat{\beta} &= \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \end{aligned}$$

# Confidence Interval Calculation using Data:

$$\Pr \left\{ \mu \in \left[ \bar{x} - z_{d/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{d/2} \frac{\sigma}{\sqrt{n}} \right] \right\} \approx 1-\alpha + \frac{1}{\sqrt{n}}$$

→ we also need to estimate  $\sigma$ :

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

the final confidence interval calculation becomes:

$$\Pr \left\{ \mu \in \left[ \bar{x} - z_{d/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{d/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] \right\} \approx 1-\alpha + \frac{1}{\sqrt{n}}$$

Intuitive/inexact reasoning of why the error in estimating  $\sigma$  does not affect the confidence intervals:

$$|\left( \bar{x} - z_{d/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right) - \left( \bar{x} - z_{d/2} \cdot \frac{\sigma}{\sqrt{n}} \right)| \approx \frac{| \hat{\sigma} - \sigma |}{\sqrt{n}} \approx \frac{\epsilon}{\sqrt{n}}$$

$$\left| \left( \bar{X} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right) - \left( \bar{X} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right) \right|$$

$$= z_{\alpha/2} \frac{|b - \hat{\sigma}|}{\sqrt{n}} \approx \frac{c}{n}$$

in the end

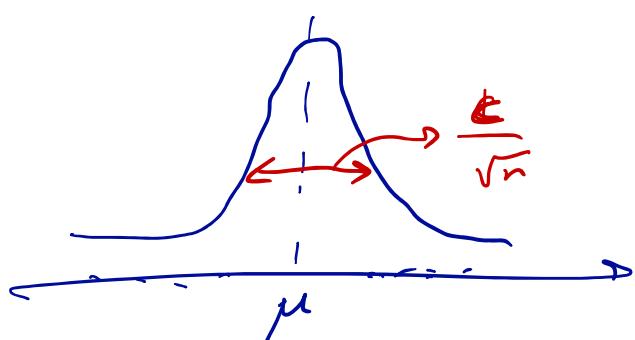
$$\Pr \left\{ \mu \in \left[ \bar{X} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] \right\}$$

$$= \Pr \left\{ \mu \in \left[ \bar{X} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] \right\}$$

$$\approx \frac{c}{n}$$

actual interval:  
 $\left[ \bar{X} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] =$   
 $\left[ \bar{X} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right] \pm \frac{c}{n}$

$$\bar{X} \rightarrow \mu \rightarrow E[\bar{X}] = \mu$$



$$|\bar{X} - \mu| \approx \frac{c}{\sqrt{n}}$$

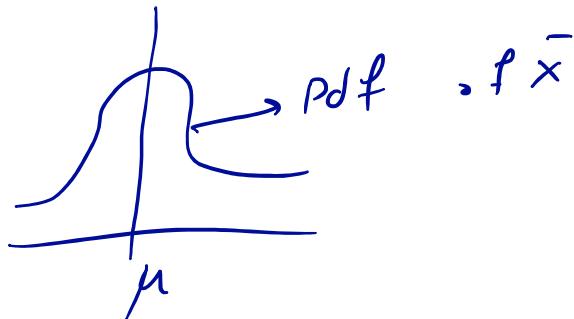
Same holds for  $|b - \hat{\sigma}|$

estimation error:

$$\mu \rightarrow \bar{X}$$

$$\begin{aligned} |\bar{X} - \mu| &\approx \frac{c}{\sqrt{n}} \\ \bar{X} &\stackrel{\text{CLT}}{\approx} \mu + \frac{6}{\sqrt{n}} N(0, 1) \\ \Rightarrow \bar{X} - \mu &\approx \frac{6}{\sqrt{n}} \underbrace{N(0, 1)}_{\frac{c}{\sqrt{n}}} \end{aligned}$$

$$E[\bar{X}] = \mu$$



Based on CLT, we derived confidence bounds:

$$\Pr \left\{ \mu \in \left[ \bar{X} - \frac{z_{d/2} \cdot 6}{\sqrt{n}}, \bar{X} + \frac{z_{d/2} \cdot 6}{\sqrt{n}} \right] \right\} \approx 1 - \alpha + \cancel{\frac{c}{\sqrt{n}}},$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\bar{x}_i - \bar{x})^2} \rightarrow |\hat{\sigma} - \sigma| = \frac{c}{\sqrt{n}}$$

# The Method of Maximum Likelihood:

Data:  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x|\theta_0)$   $\xrightarrow{(\theta_1, \dots, \theta_m)}$

Let's define the likelihood function  $lik(\theta)$  as follows:

$$lik(\theta) = f(x_1, \dots, x_n | \theta)$$

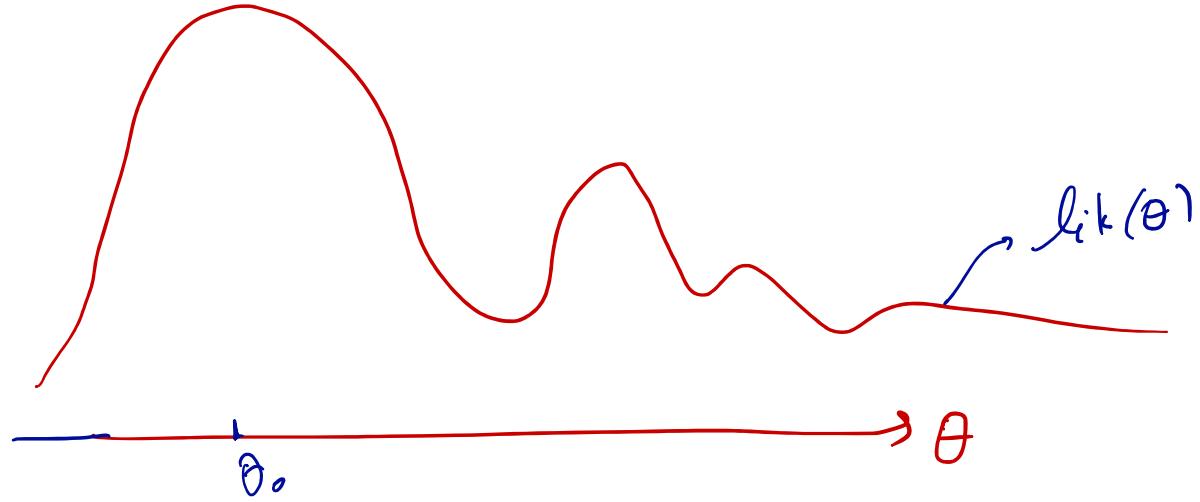
Note that the likelihood function is defined w.r.t. the current data

Sample  $(x_1, \dots, x_n)$

---

Problem: Data is generated with some parameter  $\theta_0$ . We <sup>do not</sup> know (ground truth) what  $\theta_0$  is and we'd like to estimate it from data.

$$\text{lik}(\theta) = \underline{f(x_1, \dots, x_n | \theta)}$$



we define the maximum likelihood estimator ( $\theta_{\text{mle}}$ ) as

$$\theta_{\text{mle}}(x_1, \dots, x_n) = \underset{\theta}{\operatorname{argmax}} \text{lik}(\theta)$$

remember  $\theta_{\text{mle}}$  is used to estimate  $\theta_0$ , which has generated the data.

$$\underbrace{x_1, \dots, x_n}_{\text{iid}} \sim f(x | \theta_0) \quad \substack{\text{ground truth}} \quad \theta_0$$

examples :

- if  $x_i \stackrel{\text{iid}}{\sim} N(\mu_0, \sigma_0^2)$

- if  $x_i \stackrel{\text{iid}}{\sim} \text{Gamma}(d_0, \beta_0)$

Let's assume the gaussian case :

$$x_i \sim N(\mu_0, \sigma_0^2)$$

$\Rightarrow$  Pdf of data is

$$\Pr\{x_i = x\} = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}$$

$$= f(x | \mu_0, \sigma_0^2)$$

$$f(x_1, \dots, x_n | \mu_0, \sigma_0^2) \stackrel{\text{iid}}{=} f(x_1 | \mu_0, \sigma_0^2) \times f(x_2 | \mu_0, \sigma_0^2) \times \dots \times f(x_n | \mu_0, \sigma_0^2)$$

$$f(x_1, \dots, x_n | \mu_0, \sigma_0)$$

$$= \prod_{i=1}^n f(x_i | \mu_0, \sigma_0)$$

$$= \frac{1}{(\sqrt{2\pi}\sigma_0)^n} \exp \left( -\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma_0^2} \right)$$

In reality (when we don't know what  $\sigma_0, \mu_0$  are), we know that

$$x_i \sim N(\mu, \sigma^2)$$

$$\hookrightarrow f(x_1, \dots, x_n | \tilde{\mu}, \tilde{\sigma}) = \prod_{i=1}^n f(x_i | \mu, \sigma)$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left( -\frac{\sum (x_i - \mu)^2}{2\sigma^2} \right)$$

$f(x_1, \dots, x_n | \mu, \sigma)$  is called the likelihood function.

$$\mu_{\text{mle}}, \theta_{\text{mle}} = \underset{\mu, \theta}{\operatorname{argmax}} \quad f(x_1, \dots, x_n | \mu, \theta)$$

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$$

$$\max_{\theta} f(x_1, \dots, x_n | \theta) = \max_{\theta} \log f(x_1, \dots, x_n | \theta)$$

↓  
because  $\log(\cdot)$  is  
an ~~not~~ strictly increasing  
function

$$\log(f(x_1, \dots, x_n | \theta))$$

$$= \log(f(x_1 | \theta) \times f(x_2 | \theta) \times \dots \times f(x_n | \theta))$$

$$= \log(f(x_1 | \theta)) + \log(f(x_2 | \theta)) + \dots + \log(f(x_n | \theta))$$

$$\operatorname{argmax}_{\theta} f(x_1, \dots, x_n | \theta)$$

$$= \operatorname{argmax}_{\theta} \log(f(x_1, \dots, x_n | \theta))$$

$$= \operatorname{argmax}_{\theta} \log \left( \prod_{i=1}^n f(x_i | \theta) \right)$$

$$= \operatorname{argmax}_{\theta} \sum_{i=1}^n \log f(x_i | \theta)$$

---

Why product?

if  $X \sim f(x)$  and  $Y \sim g(y)$   
and  $X$  and  $Y$  are independent; -  
then Pd f of  $(X, Y)$  is

$$\begin{aligned} \text{Pd f of } \underline{(X, Y)} &= \Pr \{ \underline{X=x}, \underline{Y=y} \} \\ &= \underline{f(x)} \cdot \underline{g(y)} \end{aligned}$$

$x_1, \dots, x_n \stackrel{iid}{\sim}$  Pdf of  $x_i$

$$\Pr \{ X_i = x_i \} = f(x_i | \theta)$$

$\xrightarrow{\text{Pdf}}$

$$f(x_1, \dots, x_n | \theta) \xrightarrow{\text{independence}} f(x_1 | \theta) \times \dots \times f(x_n | \theta)$$

MGF :

$$E[e^{tX}] \approx 1 + \sum_{i=1}^{\infty} E[X^i] \frac{t^i}{i!}$$

$E[X]$

$E[X^2]$

$$1 + E[X]t + E[X^2]\frac{t^2}{2} + \dots$$

$\uparrow \mu$        $\uparrow \mu_2$

## Lecture 7:

### Method of Maximum Likelihood Estimation (MLE)

Data:  $x_1, \dots, x_n \stackrel{iid}{\sim} f(x|\theta_0)$

Here,  $\theta_0$  denotes the <sup>true</sup> parameter(s) that generates the data.

Input and Goal:

{  
  →  $(x_1, \dots, x_n)$   
  → we also know  $f$   
  → estimate  $\theta_0$

} Known  
} Goal

Design  $\hat{\theta}_{\text{MLE}}(x_1, \dots, x_n)$  to estimate  $\theta_0$ .

Example:  $x_1, \dots, x_n \stackrel{iid}{\sim} f(x | \lambda_0)$

$$= \lambda_0 e^{-\lambda_0 x} \mathbb{1}_{\{x \geq 0\}}$$

$\Pr \{x_i = x | \lambda_0\} = \lambda_0 e^{-\lambda_0 x} \mathbb{1}_{\{x \geq 0\}}$

In this case  $f(x | \lambda) = \lambda e^{-\lambda x} \mathbb{1}_{\{x \geq 0\}}$

the PDF of the  
data generating distribution

$\hat{\theta}_{mle}(x_1, \dots, x_n)$ :

- First define the  $lik(\theta)$  as:

$$lik(\theta) = f(\underbrace{x_1, \dots, x_n}_{\text{the data sample}} | \theta)$$

$$\hat{\theta}_{mle} \stackrel{\Delta}{=} \underset{\theta}{\operatorname{argmax}} \ lik(\theta)$$

$\stackrel{\Delta}{=}$  is equivalent to := (meaning that  $\hat{\theta}_{mle}$  is defined as)

when  $x_1, \dots, x_n$  is assumed to be iid:

$$f(\underline{x}_1, \dots, \underline{x}_n | \theta) \stackrel{iid}{=} f(x_1 | \theta) \times \dots \times f(x_n | \theta)$$

$$\text{lik}(\theta) = f(x_1 | \theta) \times \dots \times f(x_n | \theta)$$

log-likelihood  
function  $\underbrace{l(\theta)}_{\log(\text{lik}(\theta))}$

$$\underset{\theta}{\operatorname{argmax}} \text{lik}(\theta) = \underset{\theta}{\operatorname{argmax}} \log(l(\theta))$$

$$= \underset{\theta}{\operatorname{argmax}} \log \left( \prod_{i=1}^n f(x_i | \theta) \right)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n \log(f(x_i | \theta))$$

---

General rule:

$$\underset{x}{\operatorname{argmax}} f(x) = \underset{x}{\operatorname{argmax}} \log(f(x))$$

Example (cont'd) :

$$f(x_i | \lambda) = \lambda e^{-\lambda x_i} \quad \{x_i \geq 0\}$$

$$\begin{aligned} \text{lik}(\lambda) &= \prod_{i=1}^n f(x_i | \lambda) \\ &= \prod_{i=1}^n (\lambda e^{-\lambda x_i}) \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \end{aligned}$$

take the log - -

$$\begin{aligned} l(\lambda) &= \log(\text{lik}(\lambda)) \\ &= \log(\lambda^n e^{-\lambda \sum_{i=1}^n x_i}) \\ &= n \log \lambda - \lambda \sum_{i=1}^n x_i \end{aligned}$$

$$\hat{\lambda}_{\text{mle}} = \underset{\lambda \geq 0}{\arg \max} \left\{ n \log \lambda - \lambda \sum_{i=1}^n x_i \right\}$$

In order to find the maximizer we'll calculate the derivative  $\ell'(\lambda)$  and set it to zero:

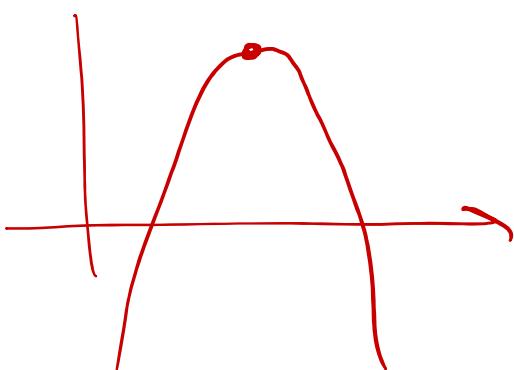
$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i$$

$$\underline{\ell'(\lambda)} = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

$$\underline{\ell'(\hat{\lambda}_{mle})} = 0 \Rightarrow \frac{n}{\hat{\lambda}_{mle}} - \sum_{i=1}^n x_i = 0$$

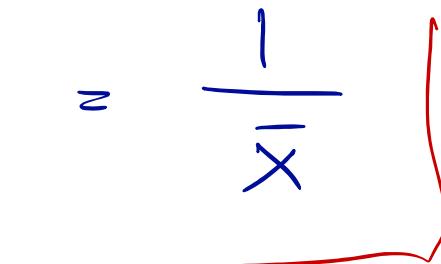
{

$\ell(\lambda)$



$$\Rightarrow \hat{\lambda}_{mle} = \frac{n}{\sum_{i=1}^n x_i}$$

$$= \frac{1}{\bar{x}}$$



So for the exponential distribution

we found that :

$$\hat{\lambda}_{mle} = \frac{1}{\bar{x}}$$

Is  $\hat{\lambda}_{mle}$  biased? Yes!

$$E[\hat{\lambda}_{mle}] = E\left[\frac{1}{\bar{x}}\right]$$

$$\neq \frac{1}{E[\bar{x}]} = \frac{1}{\frac{1}{\lambda_0}} = \lambda_0$$

for most distributions Y

$$E\left[\frac{1}{y}\right] \neq \frac{1}{E[y]}$$



## Large Sample Theory for mle:

In the section, we'd like to find out how good is the mle estimator as we increase the number of data points  $n$ ?

The first thing that we should expect is that  $\hat{\theta}_{\text{mle}}$  will become close to the true value  $\theta_0$  as the number of samples  $n$  increases.

$$\lim_{n \rightarrow \infty} \hat{\theta}_{\text{mle}} = \underline{\theta_0} \quad \text{law (1)}$$

$$\begin{aligned} \bar{X} &\xrightarrow{\mu} \\ \bar{X} &\approx \mu + \frac{\sigma}{\sqrt{n}} N(0, 1) \end{aligned}$$

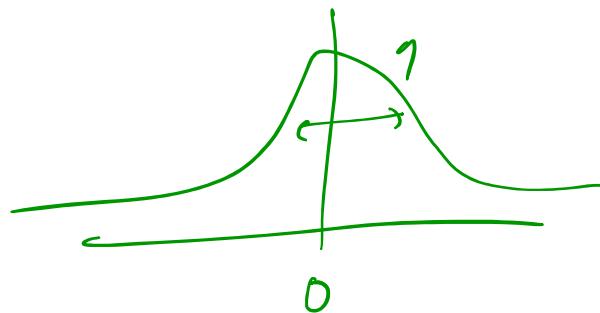
$$\hat{\theta}_{\text{mle}} - \underline{\theta_0} = \frac{1}{\sqrt{n I(\theta_0)}} N(0, 1) \quad \text{law (2)}$$

error  $\rightarrow \frac{\sigma}{\sqrt{n}}$

$$I(\theta_0) = E \left[ \left( \frac{d}{d\theta} \log f(x|\theta_0) \right)^2 \right] \rightarrow \text{The fisher information}$$

$$\bar{X} - \mu = \frac{\textcircled{6}}{\sqrt{n}} N(0, 1)$$

$$\hat{\theta}_{\text{mle}} - \theta_0 = \frac{1}{\sqrt{n} I(\theta_0)} N(0, 1)$$



Let's show law (1).

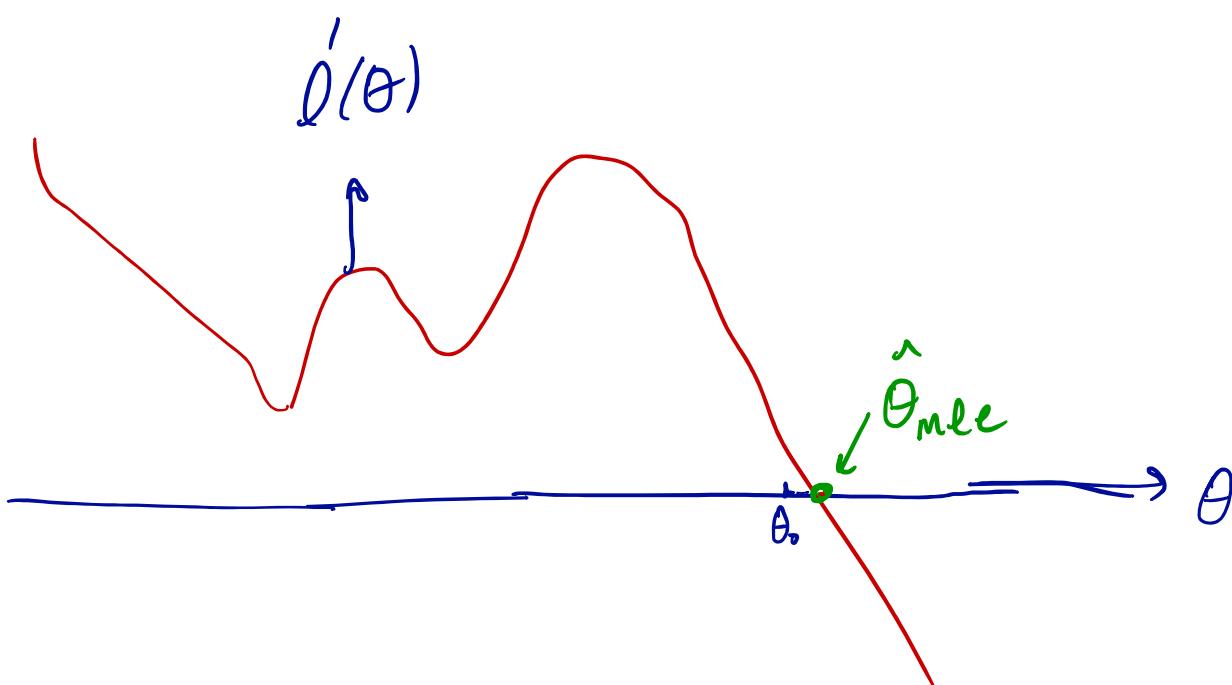
Lemma:  $\lim_{n \rightarrow \infty} \hat{\theta}_{\text{mle}} = \theta_0$

The full proof of this lemma is beyond the scope of our class. Here, I'll provide an intuitive picture.

Recall that in order to find  $\hat{\theta}_{mle}$ , we need to find the solution to the following equation:

$$l'(\hat{\theta}_{mle}) = \left. \frac{d}{d\theta} l(\theta) \right|_{\theta=\hat{\theta}_{mle}} = 0$$

$$\begin{cases} l'(\hat{\theta}_{mle}) = 0 \\ l'(\underline{\theta}_0) \xrightarrow{n \rightarrow \infty} 0 \end{cases}$$



as  $n$  increases, we show that  
 $\ell'(\theta_0) \rightarrow 0$ .

Instead of  $\ell(\theta)$ , let us look  
 at a normalized version of it,

$\frac{1}{n} \ell(\theta)$ , and see how  $\frac{1}{n} \ell'(\theta_0)$  behaves.

$$\ell(\theta) \rightarrow \frac{1}{n} \ell(\theta).$$

$$\frac{1}{n} \ell(\theta) = \frac{1}{n} \log \left( \prod_{i=1}^n f(x_i | \theta) \right)$$

$$= \frac{1}{n} \sum_{i=1}^n \log f(x_i | \theta)$$

$$\begin{aligned} \Rightarrow \frac{1}{n} \ell'(\theta) &= \frac{1}{n} \sum_{i=1}^n \left( \log f(x_i | \theta) \right)' \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\frac{\partial}{\partial \theta} f(x_i | \theta)}{f(x_i | \theta)} \end{aligned}$$

$$\frac{1}{n} \ell'(\theta) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\frac{d}{d\theta} f(x_i|\theta)}{f(x_i|\theta)} \right\}$$

$y_i \uparrow$

$x_i$ 's  $\equiv$  iid

$$y_i = \frac{\frac{d}{d\theta} f(x_i|\theta)}{f(x_i|\theta)} = \text{some function of } x_i$$

$y_i$ 's  $\stackrel{?}{\equiv}$  iid

$x_1, \dots, x_n \sim \text{iid}$

$\downarrow$

$\underbrace{g(x_1), \dots, g(x_n)}_{\text{iid}} \sim \text{iid}$

if  $x, y$  are independent  
 then any function of  
 $x, y$ , namely  $f(x)$  and  $g(y)$   
 are independent too.

$$\frac{1}{n} \ell'(\theta) = \frac{y_1 + y_2 + \dots + y_n}{n} \xrightarrow{n \rightarrow \infty} E[y]$$

$=$

law of large numbers

$$\frac{1}{n} \ell'(\theta) \xrightarrow{n \rightarrow \infty} E \left[ \frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)} \right]$$

Let's compute  
this term when  
 $\theta = \theta_0$ .

$$\frac{1}{n} \ell'(\theta_0) = E \left[ \frac{\frac{d}{d\theta} f(x|\theta_0)}{f(x|\theta_0)} \right]$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ell'(\theta_0) = E \left[ \frac{\frac{d}{d\theta} f(x|\theta_0)}{f(x|\theta_0)} \right]$$

$$E_x[-] = \int \dots f(x|\theta) \frac{dx}{d\theta}$$

$$= \int_x \frac{\frac{d}{d\theta} f(x|\theta) \Big|_{\theta=\theta_0}}{f(x|\theta_0)} f(x|\theta_0) dx$$

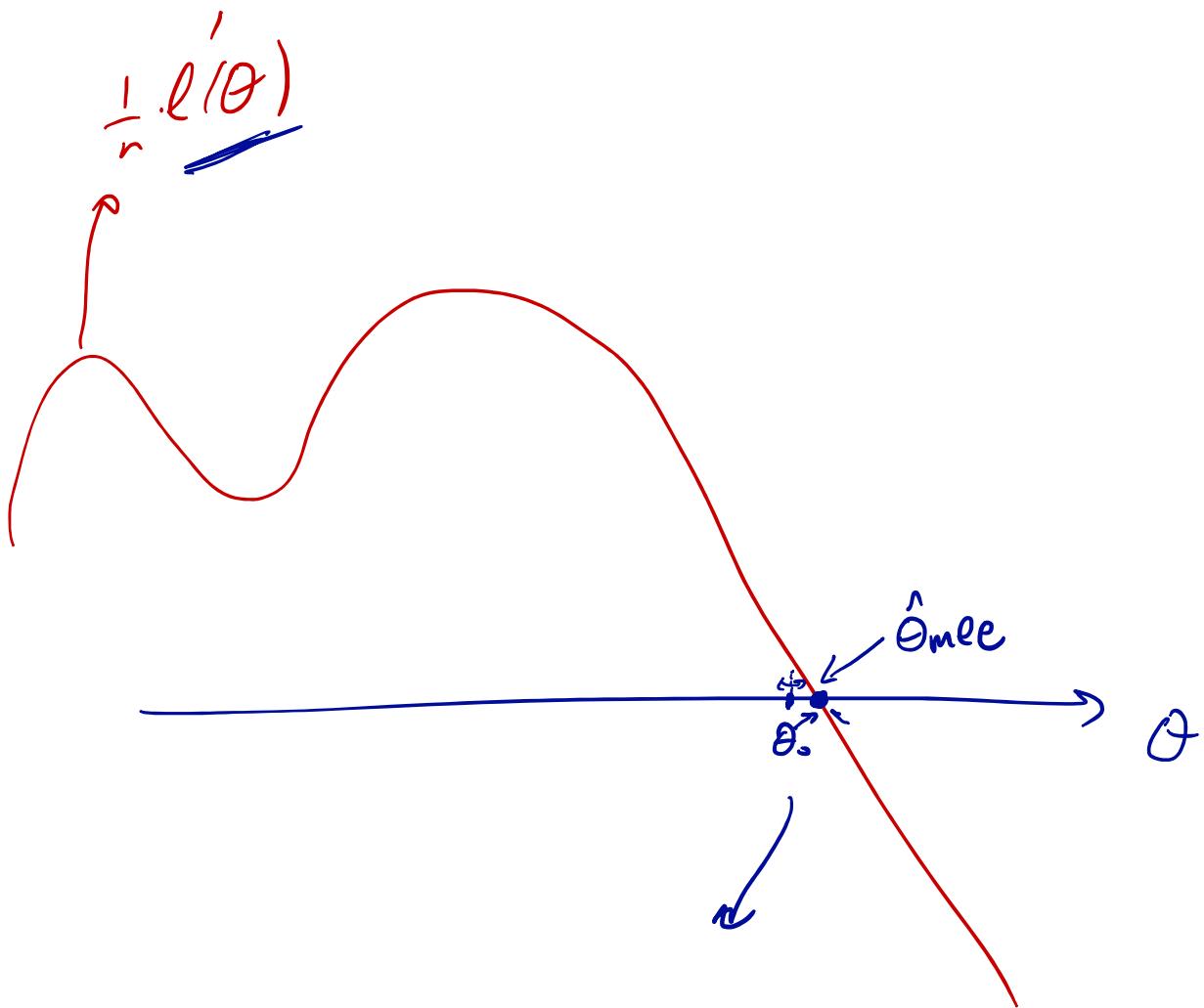
$$= \int_x \frac{d}{d\theta} f(x|\theta) \Big|_{\theta=\theta_0} dx$$

$$= \frac{d}{d\theta} \left( \int_x f(x|\theta) \right) \Big|_{\theta=\theta_0}$$

$$= \frac{d}{d\theta} (1) =$$

$$\frac{d}{d\theta} \overbrace{-}^{\text{---}} =$$

$$= 0$$



$$\lim_{n \rightarrow \infty} \hat{\theta}_{mle} = \theta_0$$

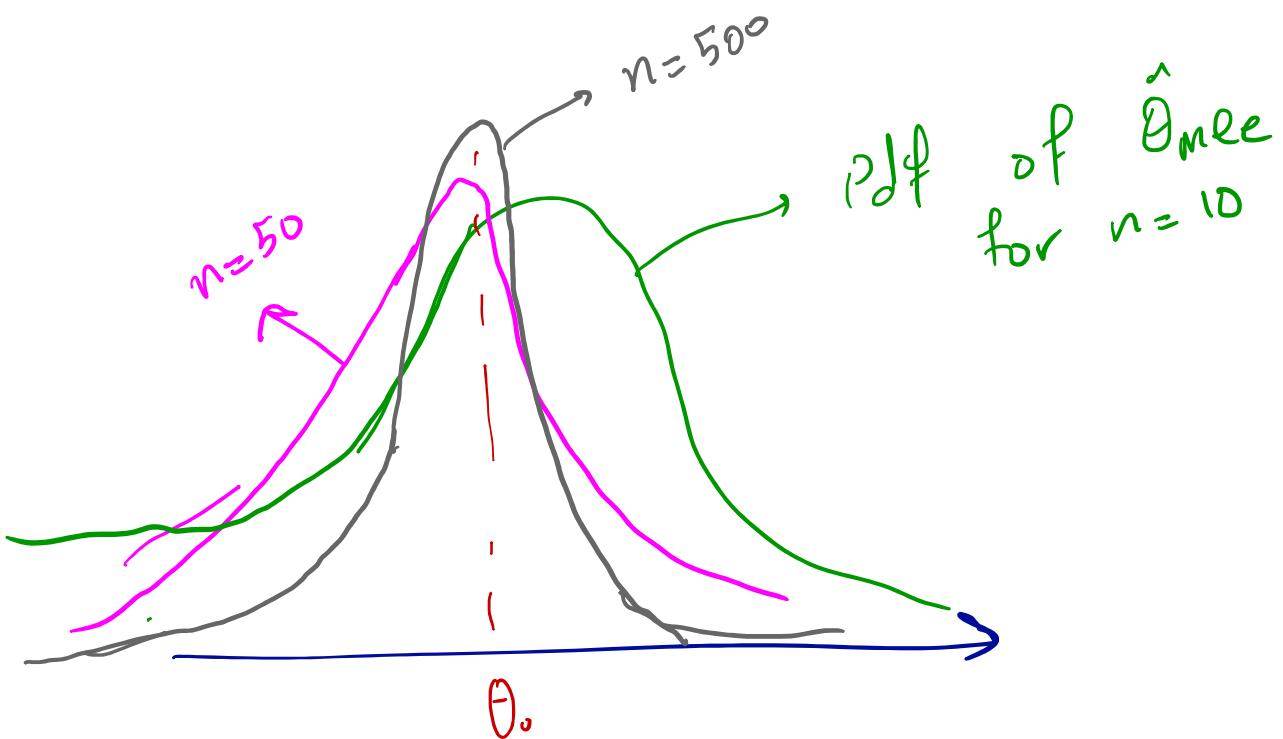
$$\left\{ \begin{array}{l} l'(\hat{\theta}_{mle}) = 0 \Rightarrow \frac{1}{n} l'(\hat{\theta}_{mle}) = 0 \\ \text{and} \\ \frac{1}{n} l'(\theta_0) \xrightarrow{n \rightarrow \infty} 0 \end{array} \right.$$

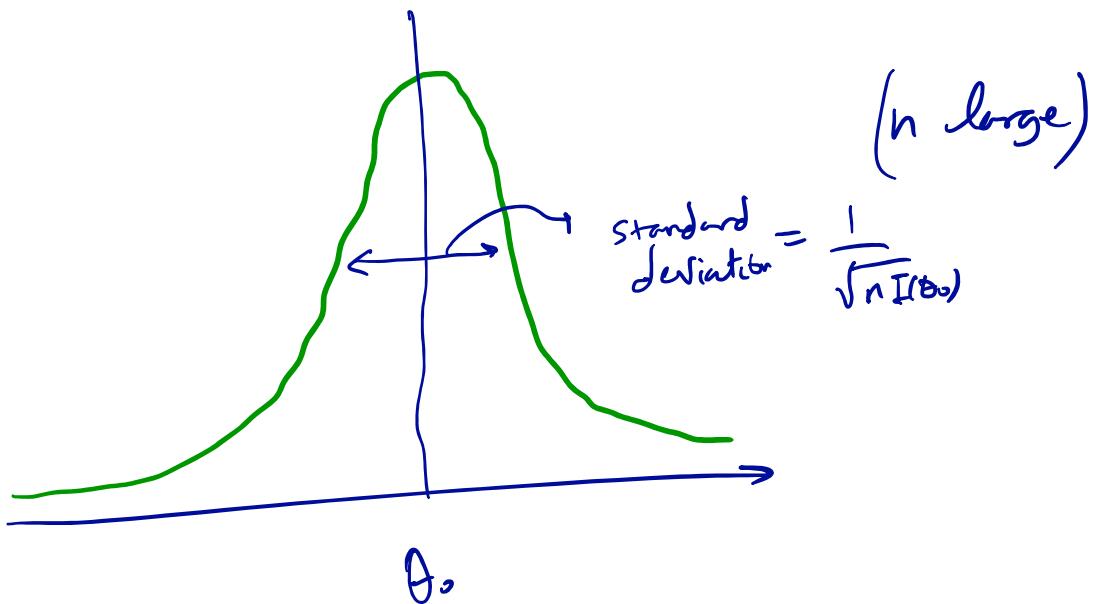
## Lecture 8:

$$\hat{\theta}_{\text{mle}} = \underset{\theta}{\operatorname{argmax}} \log \left( f(x_1, \dots, x_n | \theta) \right)$$
$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n f(x_i | \theta)$$
$$\qquad \qquad \qquad \ell(\theta)$$

(1)  $\lim_{n \rightarrow \infty} \hat{\theta}_{\text{mle}} = \theta_0$   $\sim N(\mu, \sigma^2)$

(2)  $\hat{\theta}_{\text{mle}} \approx \theta_0 + N(0, \frac{1}{n I(\theta_0)})$





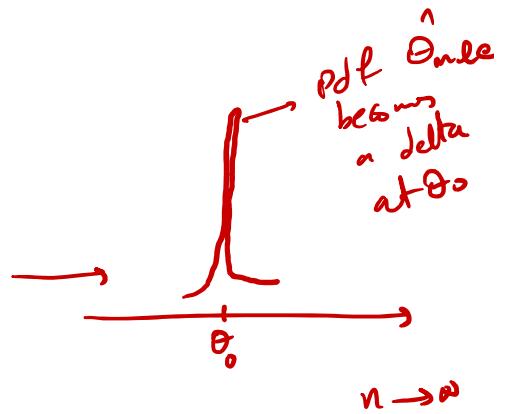
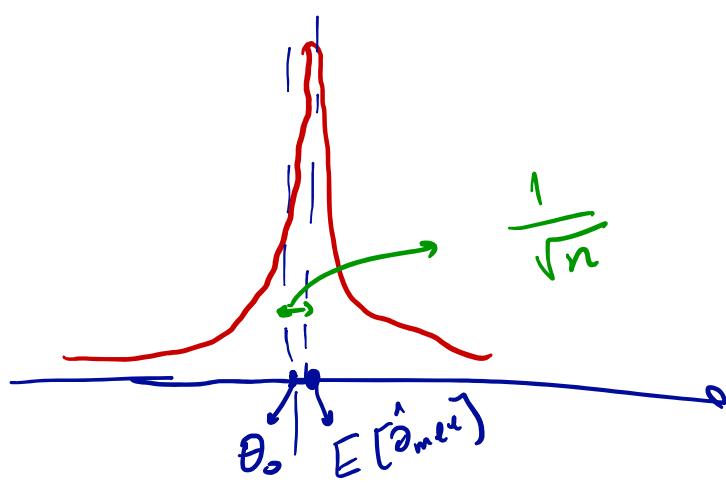
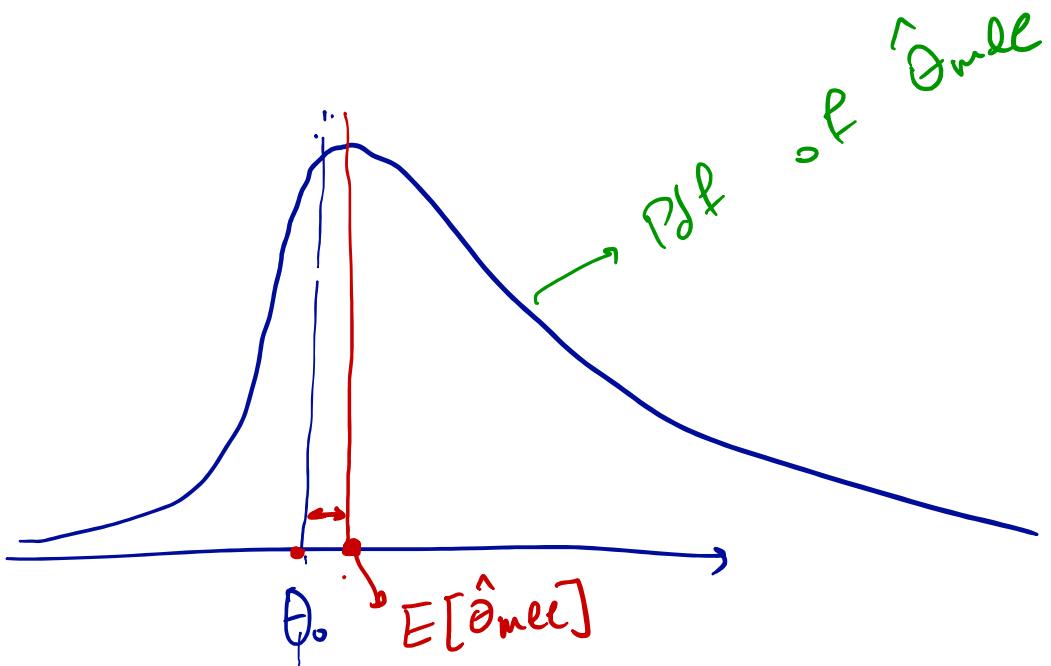
$E[\hat{\theta}_{mle}] \neq \theta_0 \rightarrow \hat{\theta}_{mle}$  is typically biased estimator for  $\theta_0$

$$\hat{\theta}_{mle} \xrightarrow{n \rightarrow \infty} \theta_0$$

e.g.  $\hat{\theta}_{mle} = \theta_0 + \frac{10}{\sqrt{n}}$

$$E[\hat{\theta}_{mle}] = \theta_0 + \underbrace{\left( \frac{10}{\sqrt{n}} \right)}_{\text{biased}}$$

when  $n \rightarrow \infty \rightarrow \hat{\theta}_{mle} = \theta_0 + \text{error}^o$



$$E[\hat{\theta}_{\text{meas}}] = \theta_0 + \underbrace{\frac{\sigma}{\sqrt{n}}}_{\xrightarrow{n \rightarrow \infty} 0}$$

$$\frac{\delta(x - x_0)}{\delta(\theta - \theta_0)} \rightarrow \begin{cases} x = x_0 & \text{w.p. 1} \\ & \end{cases}$$

Theorem 2 (Central limit type result for  $\hat{\theta}_{MLE} - \theta_0$ )

For  $n$  sufficiently large, we have

$$\hat{\theta}_{MLE} - \theta_0 \sim N(0, \underbrace{\frac{1}{n I(\theta_0)}}_{\text{variance is of order } \frac{1}{n}})$$

zero mean

variance is of order  $\frac{1}{n}$

$$= \frac{1}{\sqrt{n I(\theta_0)}} N(0, 1)$$

where  $I(\theta_0)$  is the so-called Fisher Information, given as follows:

(assuming  $x \sim f(x|\theta_0)$ ) :

$$I(\theta_0) = E \left[ \left( \frac{d}{d\theta} (\log f(x|\theta_0)) \right)^2 \right]$$

$$= \int_x \left( \frac{\frac{d}{d\theta} f(x|\theta_0)}{f(x|\theta_0)} \right)^2 f(x|\theta_0) dx$$

proof:

$$\hat{\theta}_{\text{mle}} - \theta_0 \sim \dots$$

Recall that  $\ell'(\hat{\theta}_{\text{mle}}) = 0$ . Using

Taylor Series, we have:

$$\ell'(\hat{\theta}_{\text{mle}}) = 0$$

$$\ell(\hat{\theta}_{\text{mle}}) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta}_{\text{mle}} - \theta_0)$$

$\stackrel{=0}{\cancel{\ell'(\hat{\theta}_{\text{mle}})}}$

$\stackrel{\frac{1}{n}}{\cancel{\ell''(\theta_0)}} \leftarrow$

$f = \ell'$   
 $x = \hat{\theta}_{\text{mle}}$   
 $x_0 = \theta_0$

+ neglecting the other terms  
↓  
 $O((\hat{\theta}_{\text{mle}} - \theta_0)^2)$

Taylor Series:  
 $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$

$$\hat{\theta}_{\text{mle}} - \theta_0 = - \frac{\ell'(\theta_0)}{\ell''(\theta_0)}$$

$$o = \ell'(\theta_0) + \ell''(\theta_0) (\hat{\theta}_{\text{mee}} - \theta_0)$$

$$\Rightarrow -\ell'(\theta_0) = \ell''(\theta_0) (\hat{\theta}_{\text{mee}} - \theta_0)$$

$$= \frac{-\ell'(\theta_0)}{\ell''(\theta_0)} = \hat{\theta}_{\text{mee}} - \theta_0$$

equivalent

$$-\frac{\ell'(\theta_0)}{\ell''(\theta_0)} \rightarrow N(0, \frac{I(\theta_0)}{n})$$

$$\frac{i}{n} \ell'(\theta_0) \rightarrow N(0, \frac{1}{n I(\theta_0)})$$

$$\frac{i}{n} \ell''(\theta_0) \rightarrow I(\theta_0)^2$$

$$\frac{1}{n} \ell''(\theta_0) = I(\theta_0)^2 + \frac{1}{\sqrt{n}} N(0, 1)$$

$$\Rightarrow N(0, \frac{1}{n I(\theta_0)})$$

$$\frac{1}{n} \ell'(\theta_0)$$

$$\frac{1}{n} \ell(\theta) = \frac{1}{n} \sum_{i=1}^n \log(f(x_i | \theta_0)) \Rightarrow$$

$$\frac{1}{n} \ell'(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{d}{d\theta} (\log f(x_i | \theta))$$

$$\frac{1}{n} \ell(\theta_0) = \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{\frac{\partial}{\partial \theta} f(x_i | \theta_0)}{f(x_i | \theta_0)}}_{Y_i}$$

$$\rightarrow \frac{1}{n} \ell(\theta) = \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{CLT} \mathcal{N}\left(\underset{=0}{\text{mean}}, \underset{\sqrt{n}}{\text{var}}\right)$$

$$E[Y_i] \xrightarrow{\substack{\text{computed} \\ \text{previous}}} \underset{x \sim f(x|\theta_0)}{E}\left[\frac{\frac{\partial}{\partial \theta} f(x_i | \theta_0)}{f(x_i | \theta_0)}\right]$$

$$Var[Y_i] = \int_x \frac{\frac{\partial}{\partial \theta} f(x_i | \theta_0)}{f(x_i | \theta_0)} f(x | \theta_0) dx$$

$$= \int_x \frac{\frac{\partial}{\partial \theta} f(x_i | \theta_0)}{f(x_i | \theta_0)} dx$$

$$= \frac{\partial}{\partial \theta} \left( \int f(x_i | \theta_0) dx \right)$$

$$= 0$$

$$\text{var}(Y_i) = E \left[ \left( \frac{d}{d\theta} (\log f(x_i|\theta)) \right)^2 \right] = I(\theta_0)$$

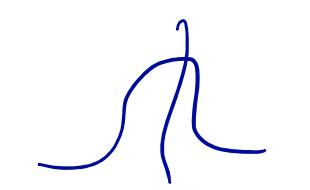
$\downarrow$                            $\downarrow$   
 $= E[Y_i^2] - E[Y_i]^2$   
 $= E[Y_i^2]$

$$Y_i = \frac{\frac{d}{d\theta} (\log f(x_i|\theta))}{f(x_i|\theta)}$$

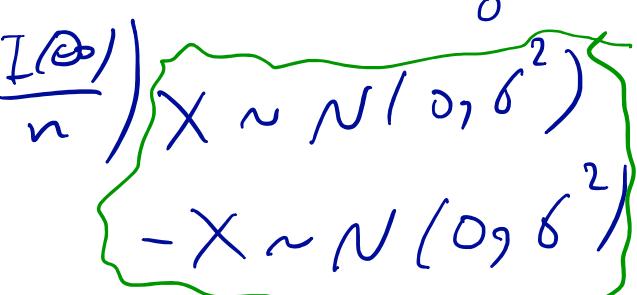
$$= \frac{\frac{d}{d\theta} f(x_i|\theta)}{f(x_i|\theta)}$$

---


$$\frac{1}{n} \ell'(\theta_0) \xrightarrow{CLT} N(0, \frac{I(\theta_0)}{n})$$



$$-\frac{1}{n} \ell'(\theta_0) \longrightarrow N(0, \frac{I(\theta_0)}{n})$$


  
 $X \sim N(0, \sigma^2)$   
 $-X \sim N(0, \sigma^2)$

$$\frac{1}{n} \ell''(\theta_0)$$

$$\frac{1}{n} \ell(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(x_i | \theta)$$

↓

$$\frac{1}{n} \ell''(\theta) = \frac{1}{n} \sum_{i=1}^n \frac{d}{d\theta^2} (\log f(x_i | \theta))$$

$\underbrace{\hspace{10em}}_{z_i}$

$$= \frac{1}{n} \sum_{i=1}^n z_i$$

$\xrightarrow{LLN}$

$$\mathbb{E}[z_i] + \cancel{\frac{1}{\sqrt{n}} N(0, \sigma^2)}$$

$\underbrace{\hspace{4em}}$

$\overbrace{\hspace{10em}}^{\mathcal{I}(\theta_0)}$

We will show that

$$\frac{1}{n} \ell''(\theta) = \mathcal{I}(\theta_0)$$

$$\rightarrow -\frac{1}{n} \ell'(\theta_0) \rightarrow N(0, \frac{I(\theta_0)}{n}) = \sqrt{\frac{I(\theta_0)}{n}} N(0, 1)$$

✓

$\left\{ \begin{array}{l} -\frac{1}{n} \ell''(\theta_0) \longrightarrow I(\theta_0) \\ \end{array} \right.$  → exercise

$$\Rightarrow \frac{-\frac{1}{n} \ell'(\theta_0)}{\sqrt{\frac{1}{n} \ell''(\theta_0)}} = \frac{\sqrt{\frac{I(\theta_0)}{n}}}{I(\theta_0)} N(0, 1)$$

$$= \sqrt{\frac{1}{n I(\theta_0)}} N(0, 1)$$

$$= N\left(0, \frac{1}{n I(\theta_0)}\right)$$

$\overbrace{\phantom{0.1234567890}}^{\text{Variance}}$



$\checkmark$   
 theorem  
 will be  
 proven.

Notation:

$$N(\mu, \sigma^2)$$

↑  
mean  
↑  
variance

in our case

$$\rightarrow N(0, \frac{1}{n I(\theta)})$$

↑  
variance

=  $\frac{1}{\sqrt{n I(\theta_0)}}$   $N(0, 1)$

↑  
Standard dev.

~~$N(\mu, \sigma)$~~  →  ~~$N(0, \frac{1}{\sqrt{n I(\theta_0)}})$~~

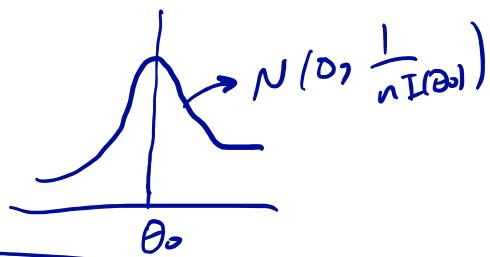
wrong notation

Proof : Exercise ( we will give it as an exercise - )

$$\frac{1}{n} L''(\theta_0) \longrightarrow I(\theta_0)$$

$$I(\theta_0) = \int_x \left( \frac{\frac{d\theta f(x|\theta_0)}{f(x|\theta_0)}}{\right)^2 f(x|\theta_0) d\theta$$

$$\hat{\theta}_{mle} - \theta_0 = N(0, \frac{1}{n I(\theta_0)})$$



$$N(0, \frac{1}{n I(\theta_0)})$$

asymptotic variance

of the mle estimator

$$\hat{\theta}_{mle} - \theta_0 = \frac{1}{\sqrt{n I(\theta_0)}} N(0, 1)$$

$$\Rightarrow \frac{\hat{\theta}_{mle} - \theta_0}{\sqrt{n I(\theta_0)}} = N(0, 1)$$

# Confidence Intervals for mle :

$$\hat{\theta}_{\text{mle}} - \theta_0 = N(0, \frac{1}{n I(\theta_0)})$$

$$\Pr \left\{ \theta_0 \in [\hat{\theta}_{\text{mle}} - \beta, \hat{\theta}_{\text{mle}} + \beta] \right\} \approx 1 - \alpha$$

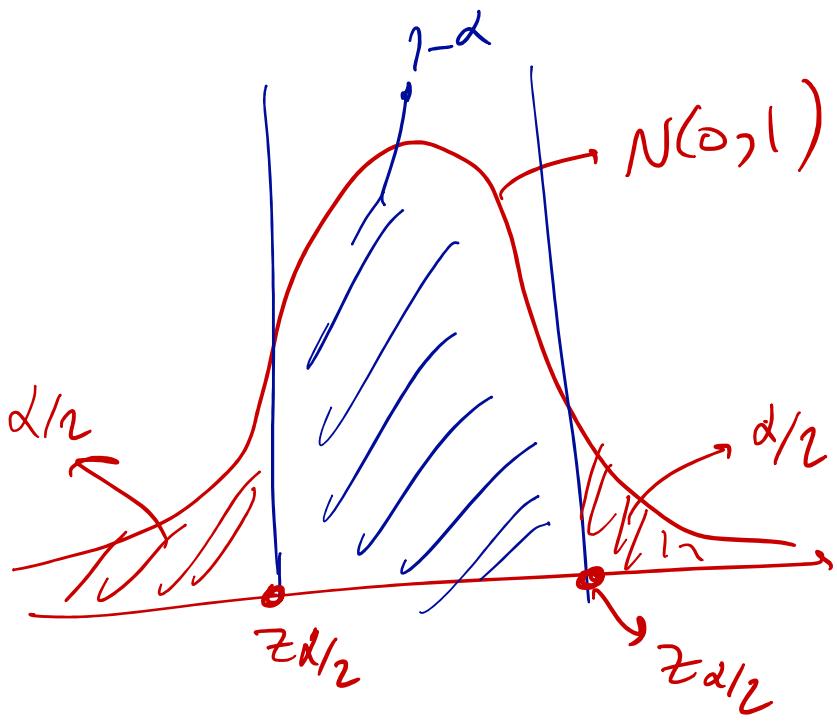


$$\Pr \left\{ \hat{\theta}_{\text{mle}} - \beta \leq \theta_0 \leq \hat{\theta}_{\text{mle}} + \beta \right\}$$

$$= \Pr \left\{ -\beta \leq \hat{\theta}_{\text{mle}} - \theta_0 \leq \beta \right\}$$

$$= \Pr \left\{ -\frac{\beta}{\sqrt{n I(\theta_0)}} \leq \frac{\hat{\theta}_{\text{mle}} - \theta_0}{\sqrt{\frac{1}{n I(\theta_0)}}} \leq \frac{\beta}{\sqrt{\frac{1}{n I(\theta_0)}}} \right\}$$

$$\Pr \left\{ -\frac{\beta}{\sqrt{n I(\theta_0)}} \leq N(0, 1) \leq \frac{\beta}{\sqrt{\frac{1}{n I(\theta_0)}}} \right\} = 1 - \alpha$$



$$\Rightarrow \beta = \frac{z_{\alpha/2}}{\sqrt{n I(\theta_0)}}$$

$$\hat{\theta}_{\text{mle}} - \theta_0 \sim \frac{C}{\sqrt{n}}$$

$$\Rightarrow \Pr \left\{ \theta_0 \in \left[ \hat{\theta}_{\text{mle}} - \frac{z_{\alpha/2}}{\sqrt{n I(\theta_0)}}, \hat{\theta}_{\text{mle}} + \frac{z_{\alpha/2}}{\sqrt{n I(\theta_0)}} \right] \right\}$$

The problem with this formula is that  $\theta_0$  is not given.  $= 1 - \alpha$

$$X_1 - X_n \longrightarrow \hat{\theta}_{\text{mle}}$$

$\Rightarrow$  Instead of  $\theta_0$ , we use  $\hat{\theta}_{\text{mle}}$ .

The final (computable) confidence  
Interval be comes:

$(1-\alpha)$  confidence interval  
for estimating  $\theta_0$

$$\Pr \left\{ \theta_0 \in \left[ \hat{\theta}_{mle} - \frac{z_{\alpha/2}}{\sqrt{n I(\hat{\theta}_{mle})}}, \hat{\theta}_{mle} + \frac{z_{\alpha/2}}{\sqrt{n I(\hat{\theta}_{mle})}} \right] \right\} = 1 - \alpha$$

f

## Lecture 9:

Point from last lecture:

$$-\frac{1}{n} \ell''(\theta_0) \rightarrow -\overline{E} \left[ \frac{\partial^2}{\partial \theta^2} (\log f(x_i | \theta)) \right] \leftarrow \\ = I(\theta_0)$$

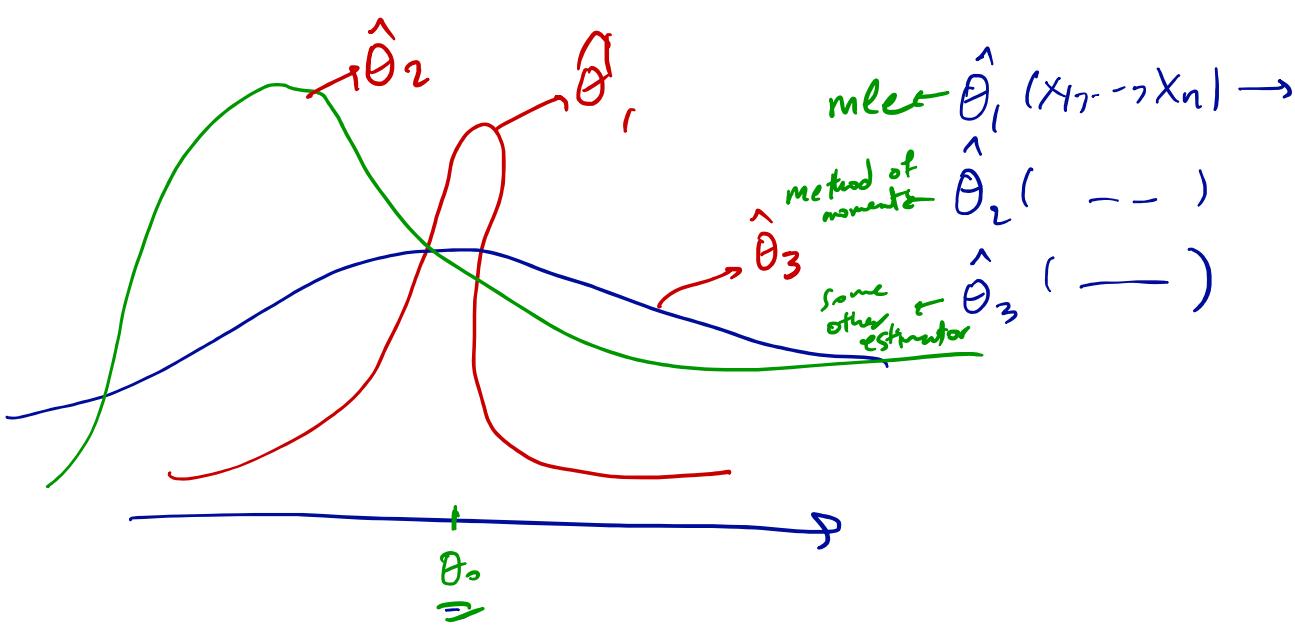
To compute  $I(\theta_0)$ , the main formula  
is:

$$I(\theta_0) = E \left[ \left( \frac{\partial}{\partial \theta} (\log f(x_i | \theta)) \right)^2 \right]$$

Cramer-Rao Bound:

Data:  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x | \bar{\theta}_0)$

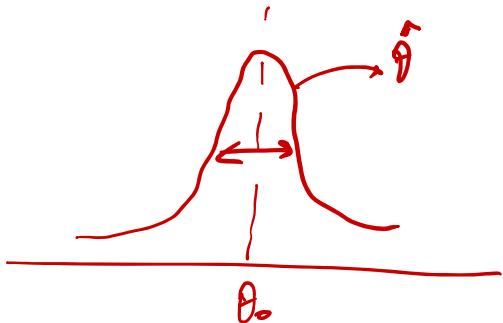
$\theta(X_1, \dots, X_n) \xrightarrow{\text{estimate}} \theta_0$



Recall that for our estimators, we typically prefer the one which has (assuming unbiasedness). a smaller variance. In this regard a fundamental question is "what is the minimum possible variance in terms of the underlying distribution  $f(x|\theta)$  and the number of samples  $n$ ?"

Data:  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x|\theta_0)$

<sup>suggest</sup>  $E[\|\hat{\theta} - \theta_0\|^2]$  ?



## The Cramer-Rao Bound:

For any unbiased estimator of  $\theta_0$ ,  
 Call it  $\hat{\theta}(x_1, \dots, x_n)$ , the variance  
 is bounded by:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n I(\theta_0)}$$

Recall that for the mle estimator  
 we had (for  $n$  large):

$$\hat{\theta}_{\text{mle}} - \theta_0 \sim N\left(0, \frac{1}{n I(\theta_0)}\right)$$



Asymptotically, the variance of the  
 mle estimator is

$$\frac{1}{n I(\theta_0)}$$

So, when  $n$  is large, we have

$$\text{Var}(\hat{\theta}_{\text{MLE}}) \approx \frac{1}{n I(\theta_0)}, \text{ which matches}$$

the Cramer - Rao lower bound.

In this sense,  $\hat{\theta}_{\text{MLE}}$  has the minimum possible variance when  $n$  is large.

$$\hat{\theta}_{\text{MLE}} - \theta_0 \sim N\left(0, \frac{1}{n I(\theta_0)}\right)$$

$$\hat{\theta}_{\text{MLE}} \sim N\left(\theta_0, \frac{1}{n I(\theta_0)}\right)$$

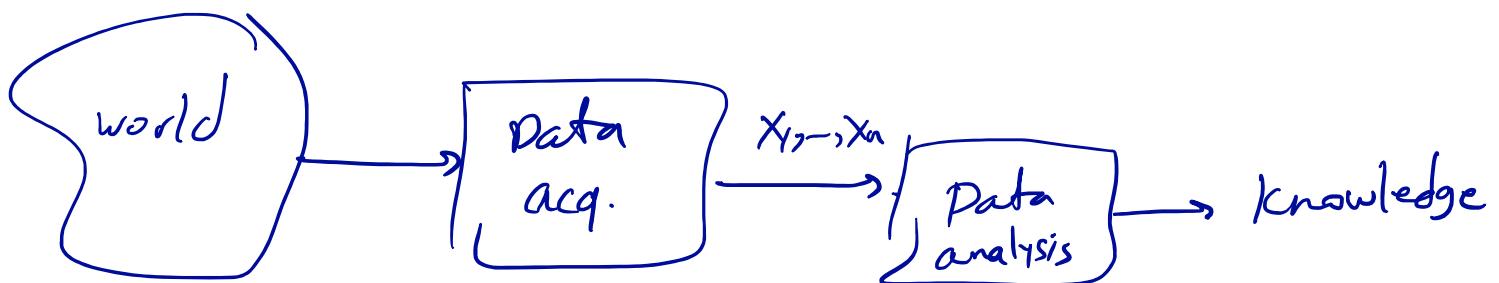
$$E[\hat{\theta}_{\text{MLE}}] = \theta_0$$

$\downarrow$   
 $n \text{ large}$

$$n = 50, 100$$

$$V_n(\hat{\theta}) \geq \frac{1}{n I(\theta)}$$

## Hypothesis testing:



In many occasions, we can "guess" some pattern of data (from experience, other related phenomena, physical laws, the data itself), and in order to verify this "guess", we need to "test" our "hypothesis" (i.e. the guessed pattern) using data.

Example: The rule of thumb in most engineering models is that the effect of the noise can be modeled

by adding a gaussian zero-mean random variable.

$$\text{output signal}(t) = \text{true-signal}(t) + N(t)$$

observation at  
time t

↓  
gaussian  
zero mean

Properties:

(1) Noise is additive  $\rightarrow$  Hyp 1

(2) Noise is gaussian distributed  
 $\hookrightarrow$  Hyp 2

(3) Noise is zero-mean.  
 $\hookrightarrow$  Hyp 3

(statistical) Hypothesis testing: Formal

Definition: Hyp. testing is a formal means of distinguishing between probability distributions on the basis of random variables generated from one of the distributions.

The null hypothesis,  $H_0$ : the claim that is initially assumed to be true (the "prior belief" claim).

The alternative hypothesis,  $H_1, H_a$ : is the assertion that is contradictory to  $H_0$ .

The null hypothesis will be rejected in favor of the alternative hypothesis only if the sample evidence strongly

suggests that  $H_0$  is false. If the sample evidence does not strongly contradict  $H_0$ , we will continue to believe in the possibility of the null hypothesis being true. The two possible conclusions of a hypothesis testing analysis are then "reject  $H_0$ " or "fail to reject  $H_0$ ".

↳ You can't reject  $H_0$  unless you have "strong evidence" against it.

---

The simplest form of hypothesis testing: Assume we have some prior belief on a parameter  $\theta$ .

$$x_1, \dots, x_n \sim f(x|\theta) .$$

$$H_0: \theta = \theta_0 \rightarrow \text{null value}$$

the alternative to the null hypothesis  
can have the following three  
forms:

$$1. H_a = \theta > \theta_0$$

$$2. H_a = \theta < \theta_0$$

$$3. H_a = \theta \neq \theta_0$$

---

### The Neyman - Pearson Paradigm:

---

We need to analyse the hypotheses  
using data  $x_1, \dots, x_n$ . A decision  
to whether or not to reject  $H_0$   
in favor of  $H_a$  is made based  
on a "statistic"  $T(x_1, \dots, x_n)$ , which

will be designed.

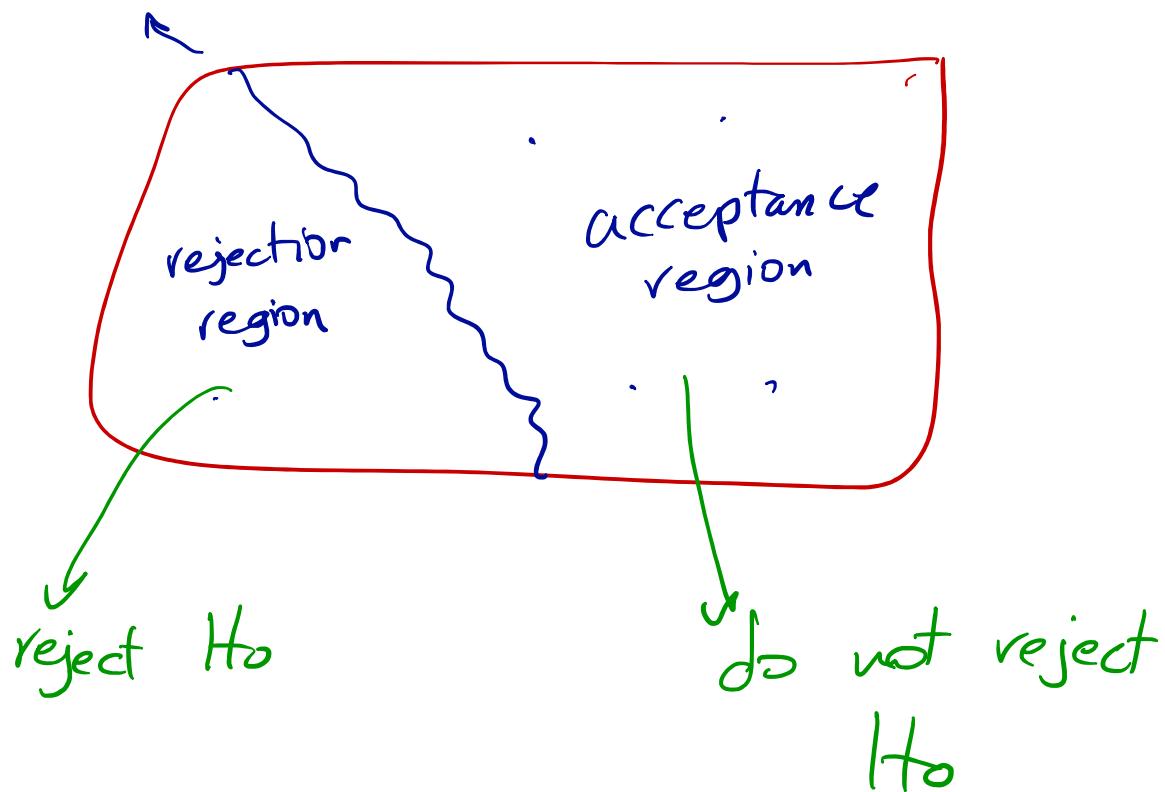
The set of values of  $T$  for which  $H_0$  is "accepted" is called the "acceptance region". And the set of values of  $T$  for which  $H_0$  is "rejected" is called the "rejection region".

$H_0$

$H_a$

$$T(x_1, \dots, x_n) = t$$

all the possible values that  $t$  could take



## Lecture 10 :

Quick hint for problem 2e in HW2:

$$X_1, \dots, X_n \sim \text{Poisson}(\lambda)$$

$$\theta = e^{-\lambda}$$

if  $X \sim \text{Poisson}(\lambda)$  then:

$$P(X=j) = \frac{e^{-\lambda} \lambda^j}{j!}$$

in particular  $\Pr(X=0) = \underbrace{e^{-\lambda}}_{\text{design an unbiased estimator for } \Pr(X=0)} = \theta$

## Errors in hypothesis testing:

Recall that Data is generated according to either  $H_0$  or  $H_a$ .

A Type I error consists of rejecting the null hypothesis  $H_0$  when it is true.

A type II error involves not rejecting the null hypothesis when  $H_0$  is false.

### Notation:

- The probability of type I error is denoted by  $\alpha$ .

$$\alpha = \Pr \{ \text{type I error} \}$$
$$= \Pr \{ H_0 \text{ is rejected} \mid H_0 \text{ is true} \}$$

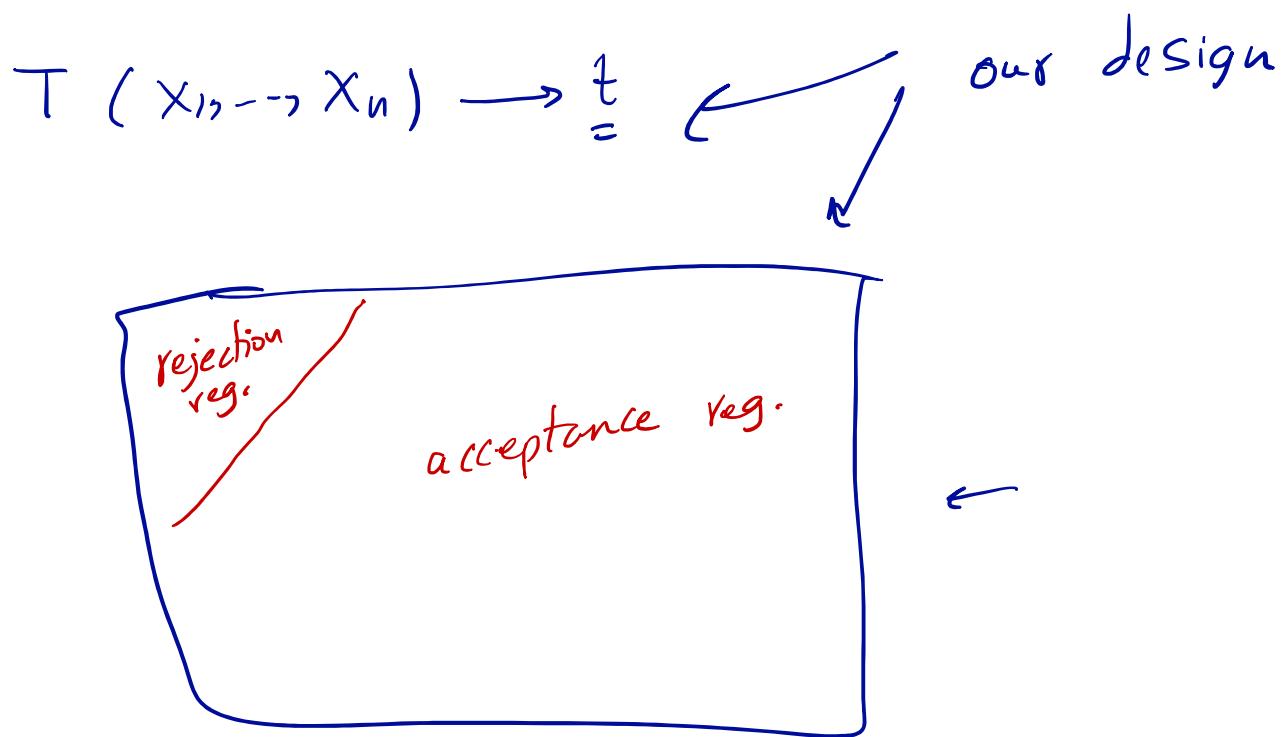
- The probability of Type II error is denoted by  $\beta$ .

$$\beta = \Pr \{ \text{type II error} \}$$
$$= \Pr \{ H_0 \text{ is not rejected} \mid H_0 \text{ is false} \}$$

Ideally, we'd like to make both errors  $\alpha, \beta$  as small as possible

(e.g.  $\alpha \approx 0, \beta \approx 0$ )

But is it possible to have  $\alpha, \beta \approx 0$  at the same time?



$\alpha \downarrow \Rightarrow$  acceptance region  $\uparrow \Rightarrow$  rejection reg.

$\beta \downarrow \Rightarrow$  acceptance region  $\downarrow \Rightarrow \alpha \uparrow \Rightarrow \beta \uparrow$

In hypothesis testing, we typically fix  $\alpha$  to a tolerance level (e.g.  $\alpha = 0.05$ ) and within this constraint we will try to minimize  $\beta$  (by optimizing  $T$ ).

(Type I error is more serious than Type II error)

Significance level: The largest value of  $\alpha$  that can be tolerated.

---

As statisticians we have two goals:

$H_0 \rightsquigarrow X_1, \dots, X_n \rightsquigarrow$  design a suitable  
 $H_a$  test statistic  
 $T(X_1, \dots, X_n)$

} (1) given a significance level  $\alpha$ ,  
the probability of type I  
error is less than  $\alpha$ .  
(specify the acceptance/rejection  
regions accordingly)

} (2) within the constraint above we would  
like to find the best  $T$  that minimizes  $\beta$ .

## Tests about the population mean:

Data:  $X_1, \dots, X_n \sim \text{dist}(\text{mean } \mu, \text{variance } \sigma^2)$   
 $N(\mu, \sigma^2)$

assume that we know what  $\sigma$  is  
but we are not sure about  $\mu$ .

$$H_0: \mu = \mu_0 \quad \}$$

$$H_a: \mu \underline{\> \>} \mu_0 \quad \}$$

(there are two other ways for  $H_a$ , i.e.

$$H_a: \mu < \mu_0$$

$$H_a: \mu \neq \mu_0$$

---

$$T(\underbrace{X_1, \dots, X_n}_{\text{---}}) \rightarrow 0$$

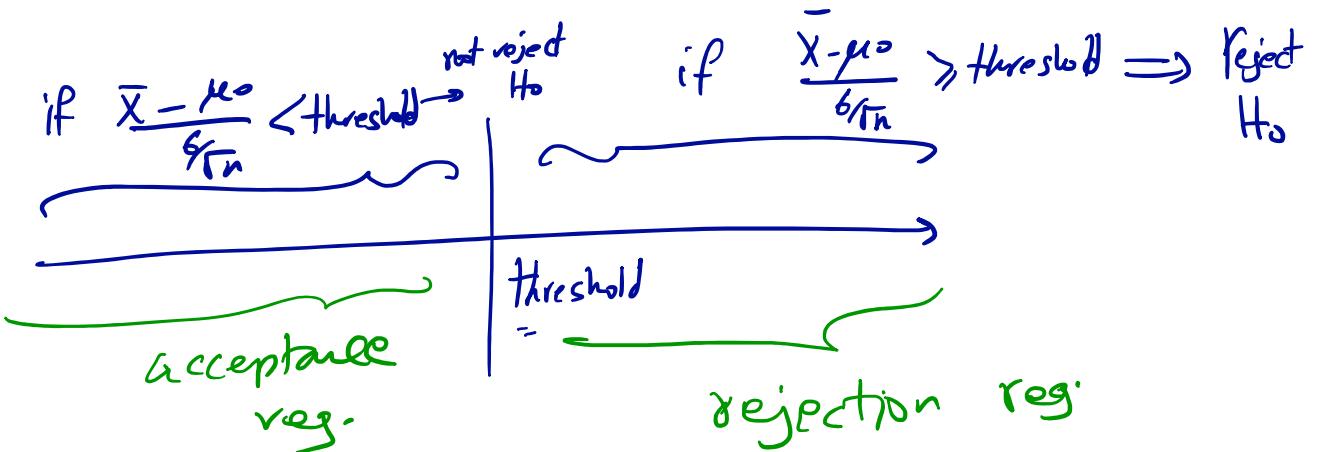
Test statistic  $T(x_1, \dots, x_n)$

$$= \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

If  $H_0$  is true ( $H_0 : \mu = \mu_0$ )  
 $x_1, \dots, x_n \sim N(\mu_0, \sigma^2)$

↙

$$T(x_1, \dots, x_n) = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \underline{N(0, 1)}$$



$$H_a : \mu > \mu_0$$

recall that we always require that the probability of type I error is equal (or less than) a significance level  $\alpha$ .  
 e.g. 0.05

$$\Pr \{ \text{type I error} \mid H_0 \text{ is true} \} = \alpha \quad \text{e.g. } 0.05$$

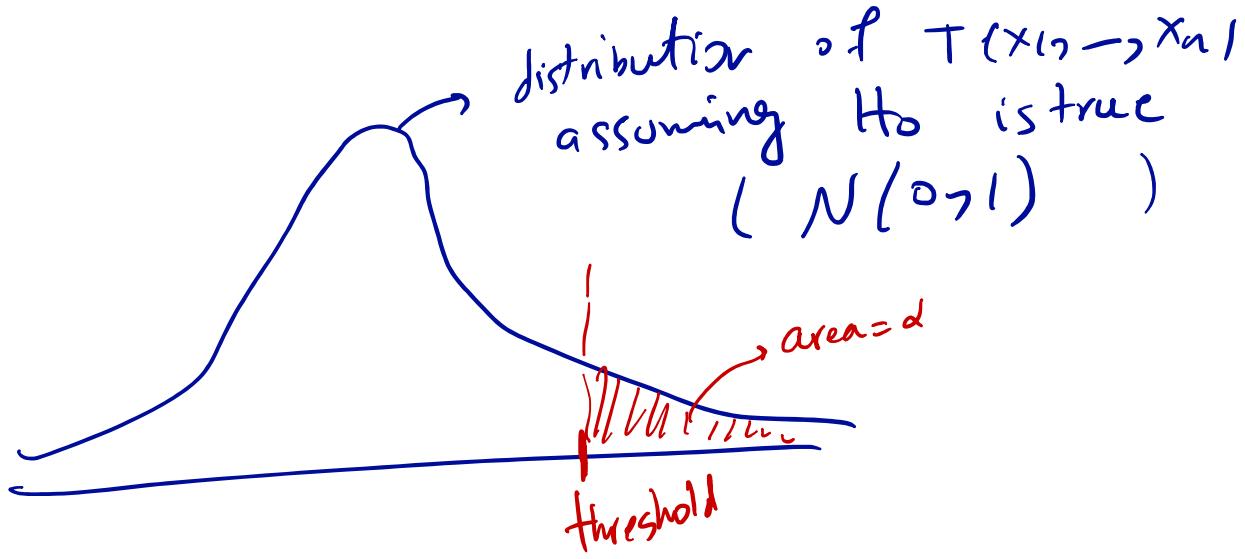
$$\Pr \{ \text{type I error} \mid H_0 \text{ is true} \} = \alpha$$

from this equation we'll find  
 the value of the threshold.

$$\alpha = \Pr \{ \text{type I error} \mid H_0 \text{ is true} \}$$

$$= \Pr \left\{ T(x_1, \dots, x_n) \geq \text{threshold} \mid \underbrace{H_0 \text{ is true}}_{T(x_1, \dots, x_n) \sim N(0, 1)} \right\}$$

$$= \Pr \left\{ N(0, 1) \geq \underline{\text{threshold}} \right\} = \alpha$$



$$\Rightarrow \boxed{\text{threshold} = z_\alpha}$$

## Z-test:

Hypothesis testing procedure (given significance level  $\alpha$ )

Data :  $x_1, \dots, x_n$

$\xrightarrow{\text{z-test}}$

Compute  $T(x_1, \dots, x_n) = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$

} reject  $H_0$  if  $T(x_1, \dots, x_n) > z_\alpha$

} not reject  $H_0$  if  $T(x_1, \dots, x_n) < z_\alpha$ .

Let's find  $\beta$  for this test:  
 (assuming significance level =  $\alpha$ )

$$\begin{aligned}\beta &= \Pr \left\{ \underbrace{\text{not reject } H_0}_{\text{H}_0 \text{ is wrong}} \right\} \\ &= \Pr \left\{ T(x_1, \dots, x_n) \leq z_\alpha \mid H_0 \text{ is wrong} \right\}\end{aligned}$$

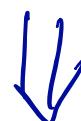
if  $H_0$  is wrong  $\Rightarrow x_1, \dots, x_n \sim N(\mu, \sigma^2)$

where  $\mu > \mu_0$



$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$= \mu + \frac{\sigma}{\sqrt{n}} N(0, 1)$$



$$T(x_1, \dots, x_n) = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

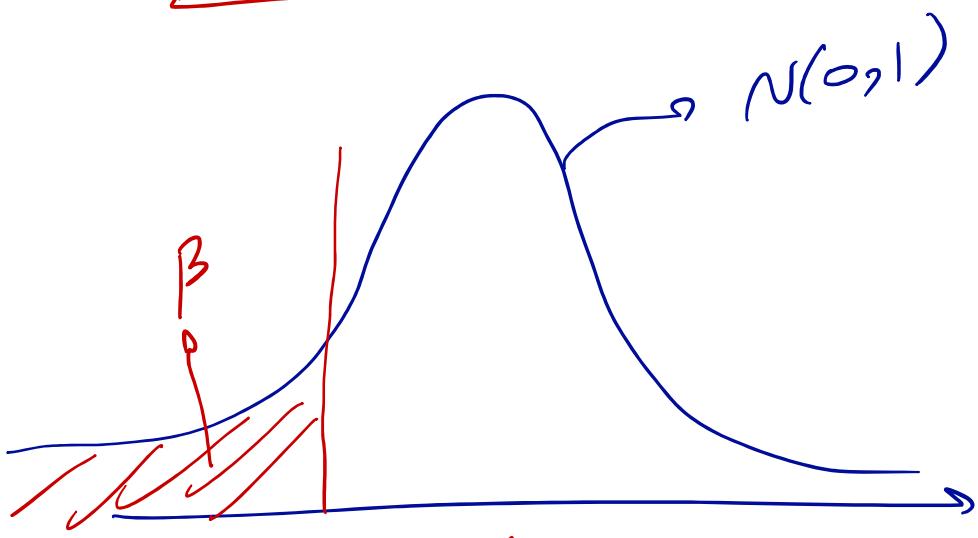
$$\sim \frac{\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}}{\frac{\sigma}{\sqrt{n}}} + N(0, 1)$$

$$\beta = \Pr \left\{ T(x_1, \dots, x_n) \leq z_\alpha \mid H_0 \text{ is wrong} \right\}$$

$$\stackrel{ID}{=} \Pr \left\{ \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} + N(0, 1) \leq z_\alpha \right\}$$

$$\stackrel{ID}{=} \Pr \left\{ N(0, 1) \leq z_\alpha - \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\}$$

$$= \Phi \left( z_\alpha - \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right)$$



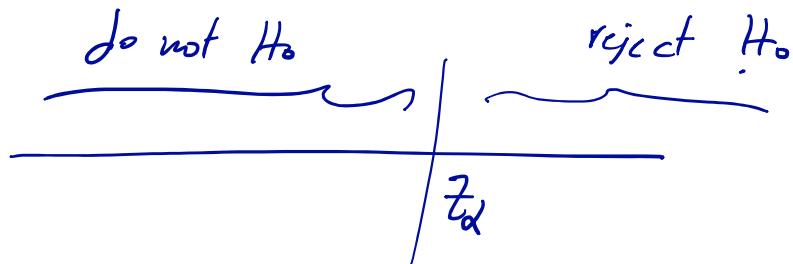
$$z_\alpha - \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

## Lecture 11:

Data:  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

$$\left. \begin{array}{l} H_0 : \mu = \mu_0 \\ H_a : \mu > \mu_0 \end{array} \right\} \quad \left( \begin{array}{l} \text{Significance} \\ \text{level} = \alpha \end{array} \right)$$

$$T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$



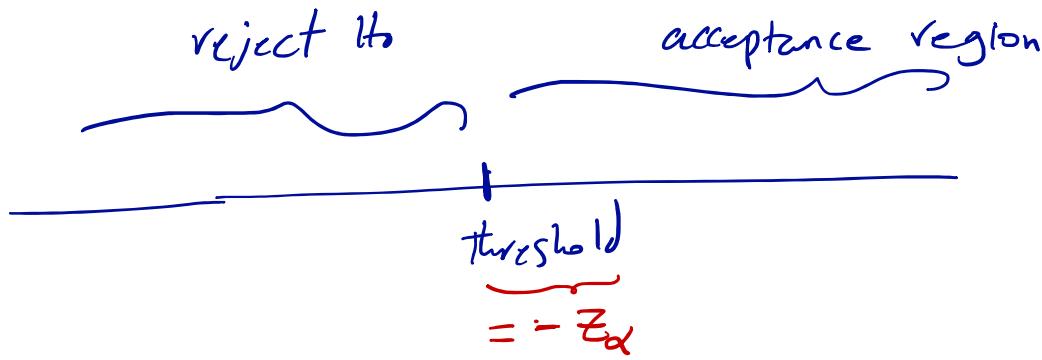
$$T(X_1, \dots, X_n) \begin{cases} \leq z_\alpha & \text{do not reject } H_0 \\ \geq z_\alpha & \text{reject } H_0 \end{cases}$$

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

Case II :

$$\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_a : \mu < \mu_0 \end{array} \right.$$

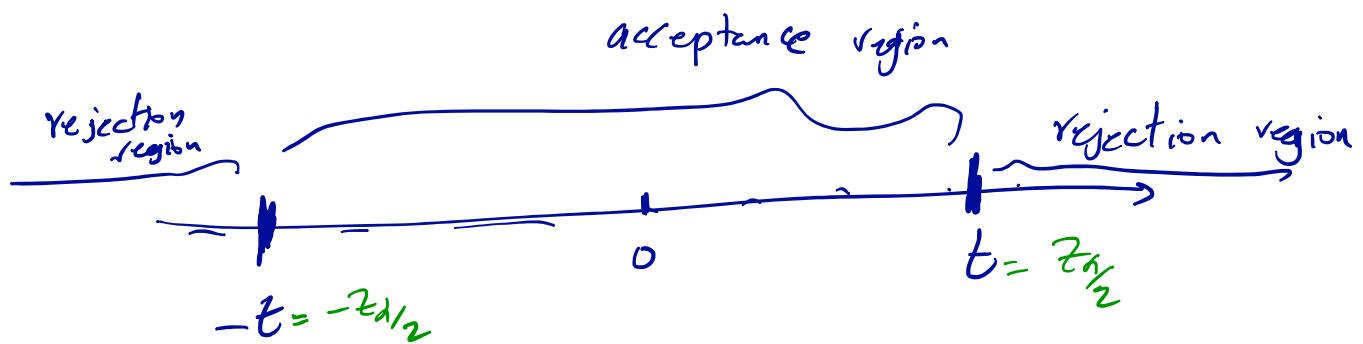
$$T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$



$$T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \begin{cases} > -Z_\alpha & \text{accept } H_0 \\ \leq -Z_\alpha & \text{reject } H_0. \end{cases}$$

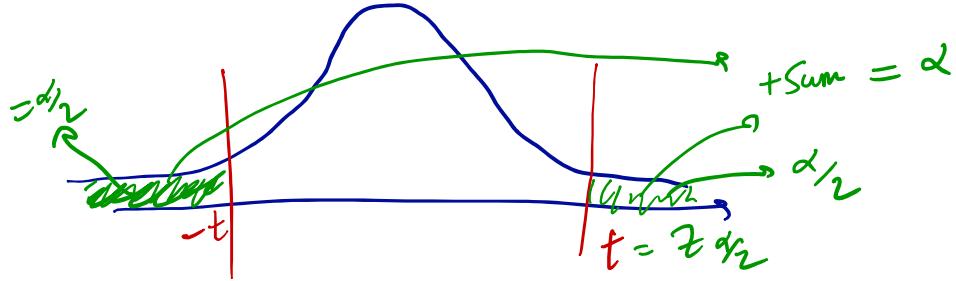
Case III :  $\left\{ \begin{array}{l} H_0 : \mu = \mu_0 \\ H_a : \mu \neq \mu_0 \end{array} \right.$

$$T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$



$$\Pr \{ H_0 \text{ is rejected} \mid H_0 \text{ is true} \} = \alpha$$

$$= \Pr \{ N(0,1) \notin [-t, t] \} = \alpha$$



} if  $\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \in [-z_{1/2}, z_{1/2}]$  then accept  $H_0$   
 . reject  $H_0$   
 O. w.

Testing the mean assuming that we have  
 a large number of samples available:

Data:  $x_1, \dots, x_n = \text{dist}(\underbrace{\text{mean} = \mu}_{\text{unknown}}, \underbrace{\text{variance} = \sigma^2}_{\text{known}})$

$H_0: \mu = \mu_0$   
 $H_a: \mu > \mu_0$

$$T(x_1, \dots, x_n) = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x} \xrightarrow{\text{CLT}} \mu + \frac{\sigma}{\sqrt{n}} N(0, 1)$   
 ↓ assuming  $H_0$   
 $\bar{x} \approx \mu_0 + \frac{\sigma}{\sqrt{n}} N(0, 1)$

⇒ Test is the same as  
 Case I and the acceptance  
 and rejection regions are  
 also the same.

What happens if we do not know  
the value of  $\sigma$ ?

$$T(x_1, \dots, x_n) = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

If we don't know  $\sigma$ , then  
we should estimate it from data

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$T(x_1, \dots, x_n) = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}.$$

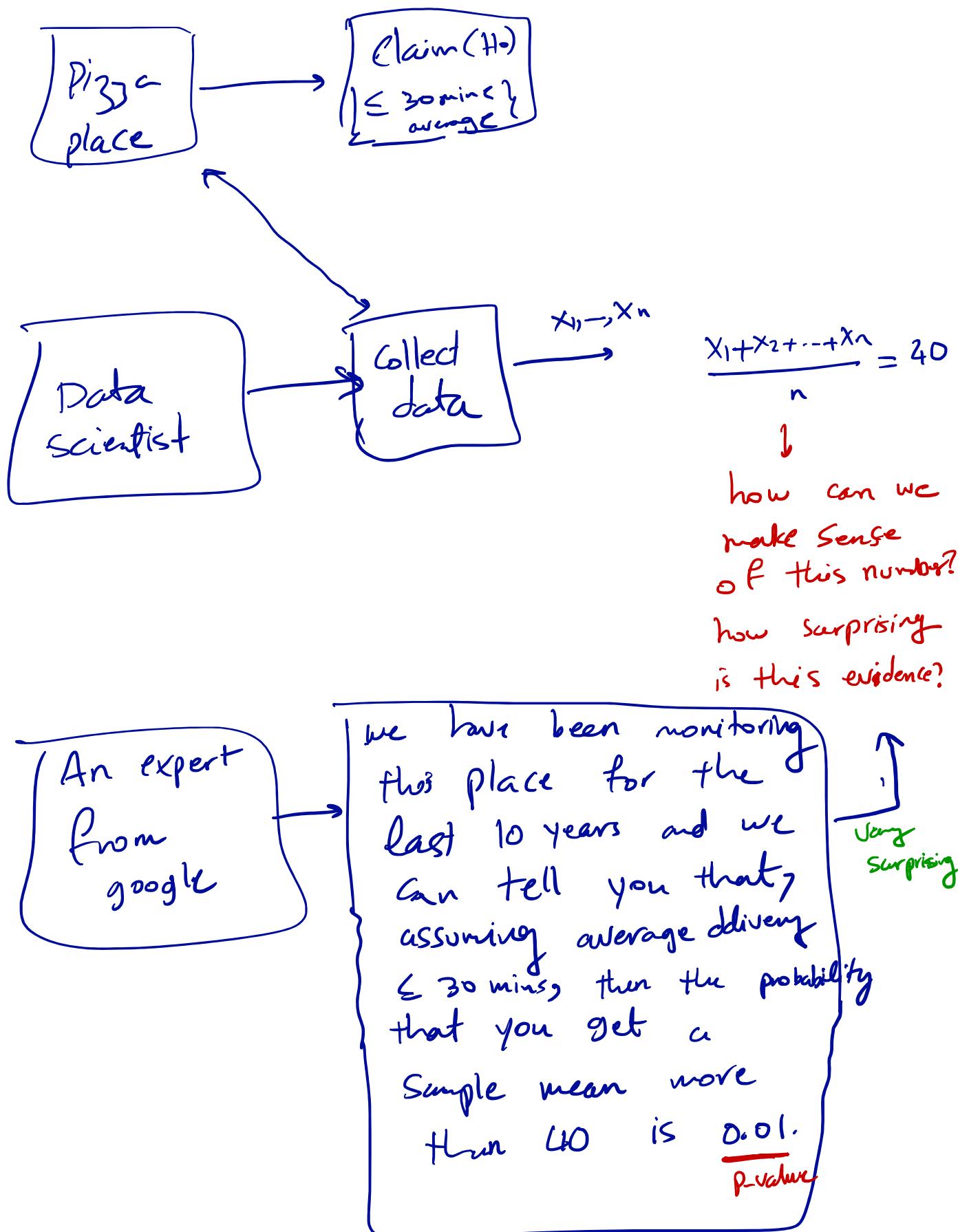
P-Value:

Definition of P-Value: The P-value is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory as the value calculated from the available sample.

---

Pizza Delivery Example:

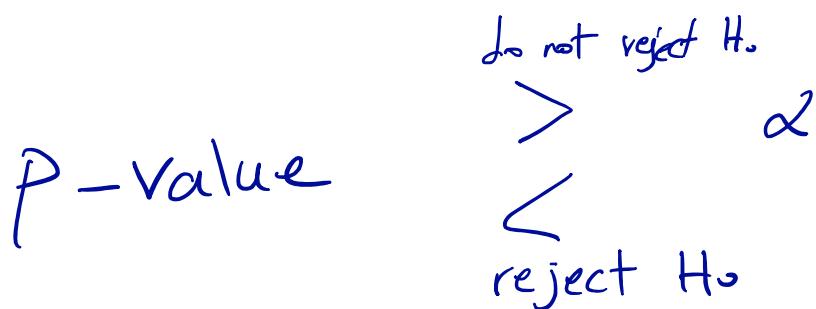
A pizza place claims that their delivery time is 30 mins or less on average. We would like to check this. So we order pizza from a bunch of different locations and find that the sample mean delivery is 40 mins.



- \* The lower the P-value  $\Leftrightarrow$  the more surprising the evidence is  $\Leftrightarrow$  the more unrealistic our null hypothesis looks.
- \* (more formal) the lower the P-Value, the more evidence there is in sample data against the null hypothesis.
- \* A p-value does not "prove" anything. It is simply a way to use surprise as a basis for making a reasonable decision.

Decision rule based on the P-value:

Given significance level  $\alpha$ :



\* The two procedures, namely the rejection/acceptance region rule and the p-value rule, are in fact identical. ~

---

p-value  $\rightarrow$  T

$p\text{-value} = \Pr \left\{ \begin{array}{l} \text{the outcome of the test} \\ \text{is more contradictory to} \\ \text{to the outcome the} \\ \text{we have currently from} \\ \text{data} \end{array} \mid H_0 \text{ is true} \right\}$

---

Our simplest hypothesis testing problem:

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

$$\begin{aligned} H_0: \mu &= \mu_0 & \text{test:} \\ H_a: \mu &> \mu_0 & T(X_1, \dots, X_n) = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \end{aligned}$$

given the current sample, > define:

$$t_{\text{data}} = T(x_1, \dots, x_n) = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$P\text{-Value} = \Pr \left\{ \underbrace{T(x_1, \dots, x_n)}_{= \bar{x}} \geq t_{\text{data}} \mid H_0 \text{ is true} \right\}$$

↑

$$\left. \begin{array}{l} H_0: \mu = \mu_0 \\ H_a: \mu > \mu_0 \end{array} \right\}$$

$$T(x_1, \dots, x_n)$$

$x_1, \dots, x_n$

$\downarrow \quad \downarrow$

$x_1 \quad x_n$

$t_{\text{data}}$

$$t_{\text{data}} = 40$$

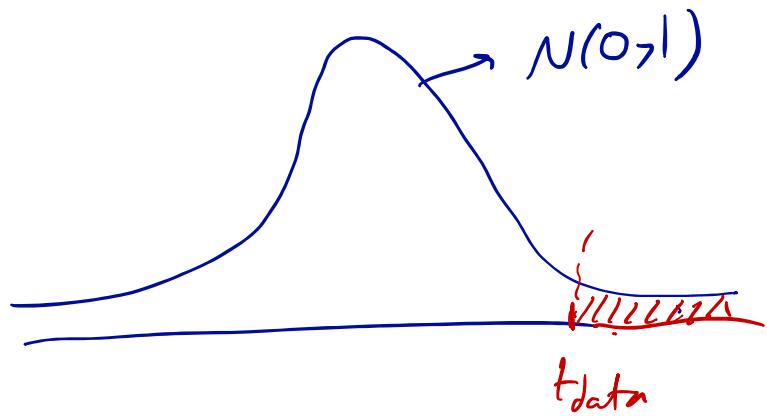
$$\Pr \left\{ T(x_1, \dots, x_n) \geq 40 \right\} = 0.01$$

$$\Pr \left\{ T(x_1, \dots, x_n) \geq t_{\text{data}} \mid \underbrace{H_0 \text{ is true}} \right\}$$

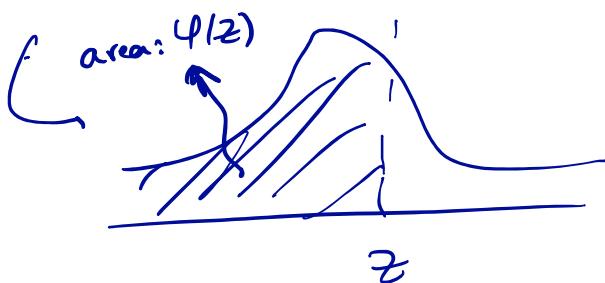


$$T(x_1, \dots, x_n) \sim N(0, 1)$$

$$\Pr \left\{ N(0, 1) \geq t_{\text{data}} \right\} = \underline{1 - \Phi(t_{\text{data}})}$$



$$\Phi(z) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$



P-value test:

$$\begin{array}{c} \text{P-value} \geq \alpha \\ \text{do not reject } H_0 \\ \text{reject } H_0 \end{array}$$



$$\begin{array}{c} 1 - \Phi(t_{\text{data}}) \geq \alpha \\ \text{acc } H_0 \\ \text{rej } H_0 \end{array}$$

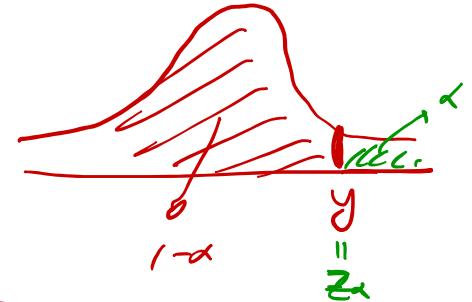


$$\begin{array}{c} \Phi(t_{\text{data}}) \geq 1 - \alpha \\ \text{reject } H_0 \\ \text{accept } H_0 \end{array}$$



$$t_{\text{data}} \geq \varphi^{-1}(1 - \alpha)$$

$$\begin{aligned} \varphi^{-1}(1 - \alpha) &= y \\ 1 - \alpha &= \varphi(y) \end{aligned}$$



Claim:

$$\varphi^{-1}(1 - \alpha) = z_\alpha$$

①

$$1 - \alpha = \underbrace{\Phi(z_\alpha)}$$

$\downarrow$   
the area  
to the  
left of  
 $z_\alpha$



## Likelihood Tests:

Data:  $X_1, \dots, X_n \sim f(x | \underline{\theta})$  <sup>known</sup> is being tested

Consider the following hypothesis testing problem:

$$\begin{cases} H_0 : \underline{\theta} = \underline{\theta}_0 & \underline{\theta}_0 \neq \underline{\theta}_a \\ H_a : \underline{\theta} = \underline{\theta}_a & \end{cases}$$

So far our tests/problems have been mostly about the mean (in which we used  $\bar{x}$  and z-test). Now, we are testing a general parameter  $\underline{\theta}$ .

Test  $T(X_1, \dots, X_n) \rightarrow$  number

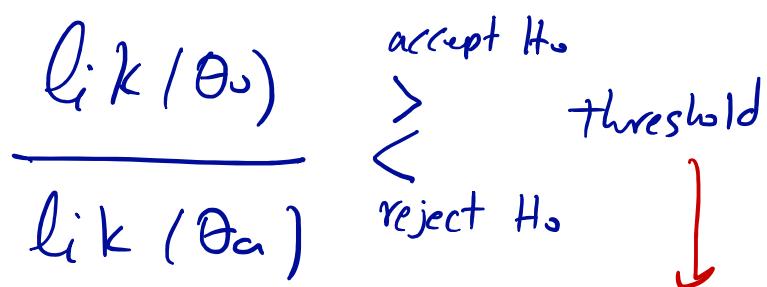
How can we design a test for the hypothesis testing problem mentioned above?

what kinds of other methods did we have to estimate  $\theta$ ?

$$\text{Likelihood test statistic} = \frac{\text{lik}(\theta_0)}{\text{lik}(\theta_a)}$$

$T(x_1 \rightarrow x_n)$

---

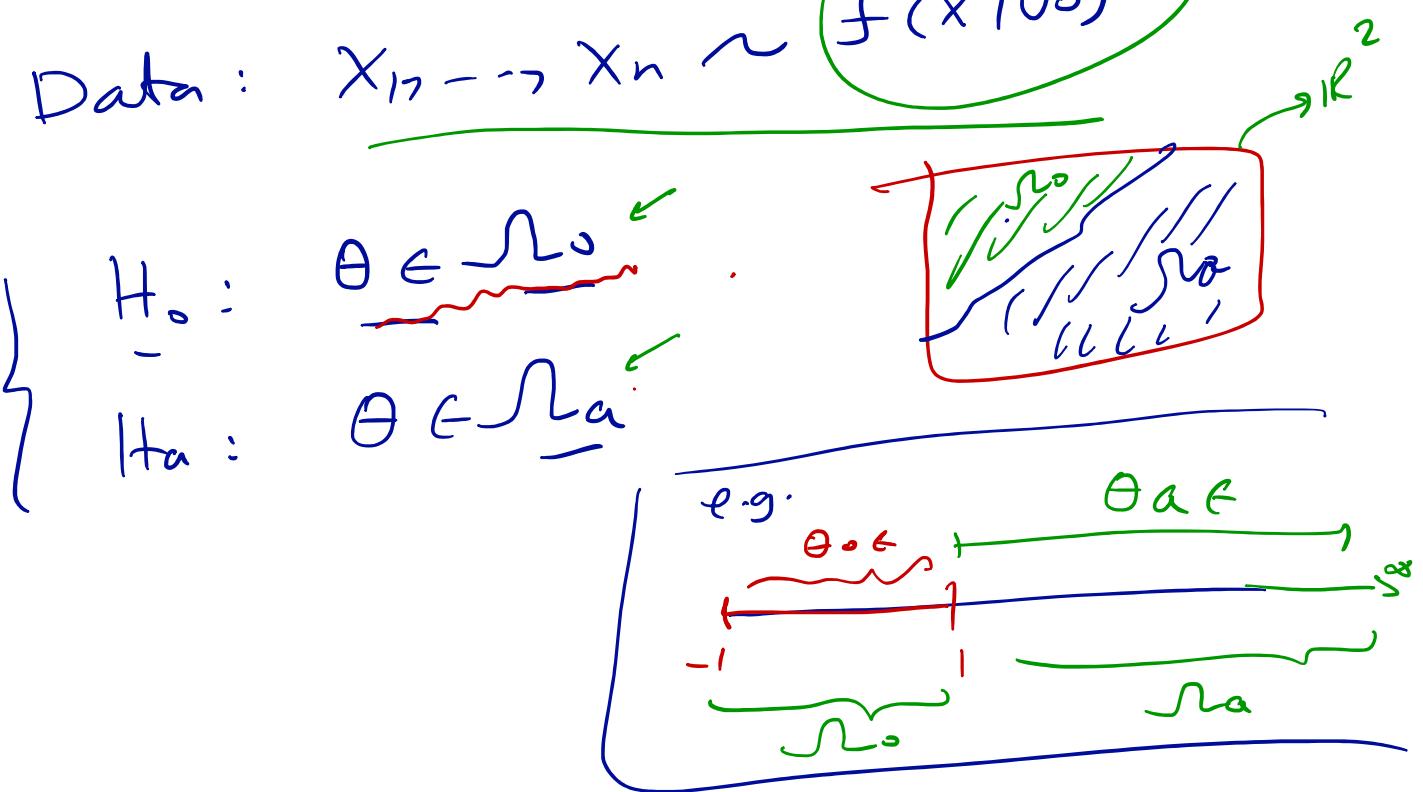


threshold will be found using the significance level  $\alpha$ .

## General form of the likelihood test:

Consider the following (general form)

hypothesis testing problem:



Generalized likelihood test:

$$T(X_1, \dots, X_n) = \frac{\max_{\theta \in \Omega_0} \text{lik}(\theta)}{\max_{\theta \in \Omega_a} \text{lik}(\theta)}$$

intuition: if I tell you that  $\theta$  is inside  $\Omega_0$ , then to estimate  $\theta$ , we can use the rule  $\rightarrow \arg \max_{\theta \in \Omega_a} \text{lik}(\theta)$

(intuition : )

→ if  $\theta_{true} \in \mathcal{R}_0$ , then what is  
a good estimate of the likelihood of  
 $\theta_{true}$ ? (what's a good estimate of  
 $lik(\theta_{true})$ ?)

↳ if we know  $\theta_{true} \in \mathcal{R}_0$ , then

$$\hat{\theta}_{true} = \underset{\theta \in \mathcal{R}_0}{\operatorname{argmax}} lik(\theta)$$

$\uparrow$

$$lik(\hat{\theta}_{true}) = \underset{\theta \in \mathcal{R}_0}{\max} lik(\theta)$$

---

question:

if  $\theta \in \mathcal{R}_0$  then we'd expect

that :

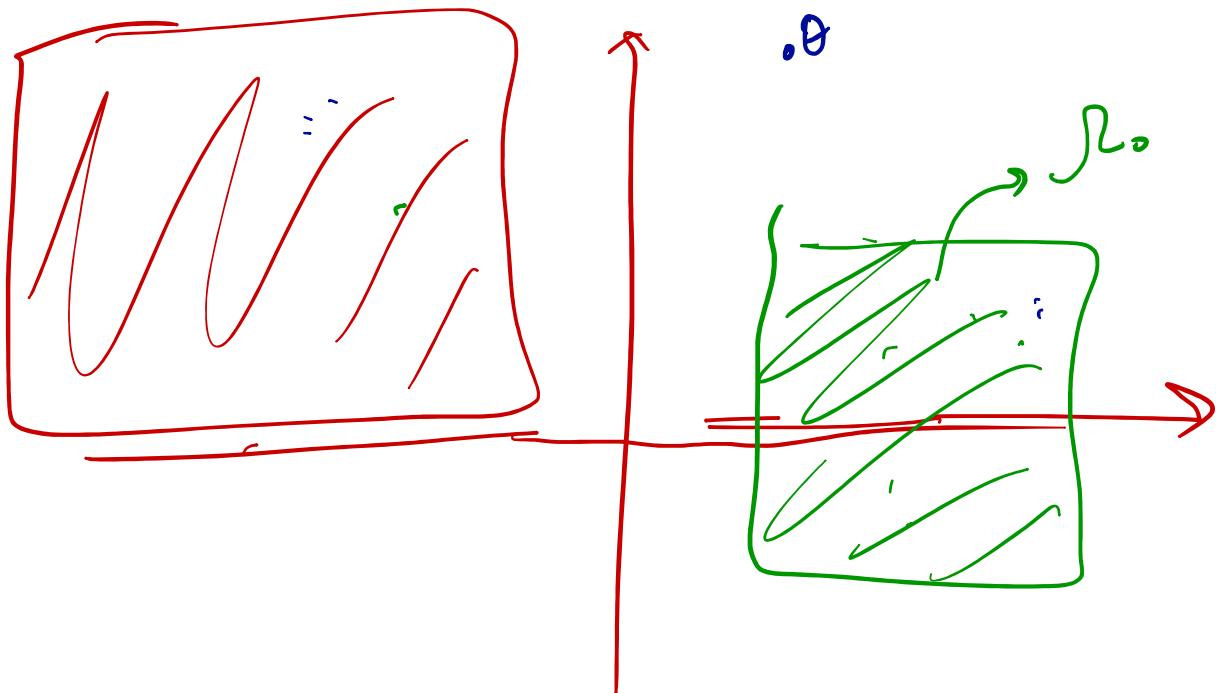
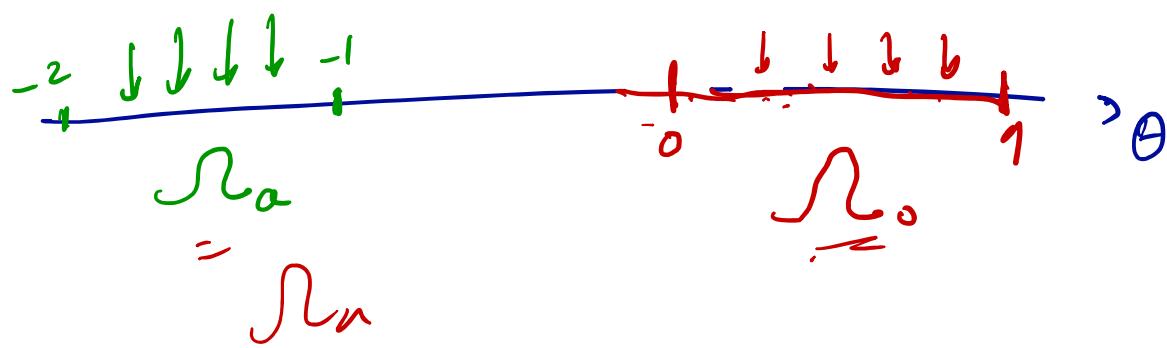
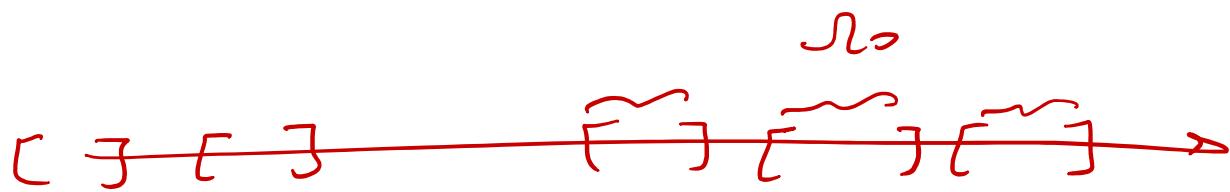
$$\max_{\theta \in \mathcal{R}_0} lik(\theta)$$



$$\max_{\theta \in \mathcal{R}_0} lik(\theta) \uparrow$$

(intuition:)  
 $x_1, \dots, x_n \sim f(x|\theta) \rightarrow \theta \in \mathbb{R}$

$$\left\{ \begin{array}{l} H_0: \theta \in \mathcal{N}_0 \\ H_a: \theta \in \mathcal{N}_a \end{array} \right.$$



Generalized likelihood test:

$$T(X_1, \dots, X_n) = \frac{\max_{\theta \in \mathcal{L}_0} \text{lik}(\theta)}{\max_{\theta \in \mathcal{L}_{\alpha}} \text{lik}(\theta)}$$

accept.  $H_0$   
 $\geq k$   
reject  $H_0$ .

The value of  $k$  would be specified wrt the significance level  $\alpha$ .

Example:  $X_1, \dots, X_n \sim \text{Exponential}(\lambda)$

$$\left\{ \begin{array}{l} H_0: \lambda = \lambda_0 \rightarrow \mathcal{L}_0 = \{\lambda_0\} \\ H_a: \lambda = \lambda_a \rightarrow \mathcal{L}_a = \{\lambda_a\} \end{array} \right.$$

(assume  $\lambda_a > \lambda_0$ )

likelihood test:

$$T(X_1, \dots, X_n) = \frac{\text{lik}(\lambda_0)}{\text{lik}(\lambda_a)}$$

$$f(x|\lambda) = \lambda e^{-\lambda x} \mathbb{1}\{x \geq 0\} -$$

given data:  $x_1, \dots, x_n$

$$\text{lik}(\lambda) = \prod_{i=1}^n f(x_i | \lambda)$$

$$= \lambda^n e^{-\lambda \sum x_i}$$

$$T(x_1, \dots, x_n) = \frac{\text{lik}(\lambda_0)}{\text{lik}(\lambda_a)}$$

$$\frac{e^{-\lambda_0}}{e^{-\lambda_a}} = e^{\frac{-\lambda_0 - (-\lambda_a)}{x - y}}$$

$$= \frac{\lambda_0^n e^{-\lambda_0 \sum_{i=1}^n x_i}}{\lambda_a^n e^{-\lambda_a \sum x_i}}$$

$$e^{-(\lambda_0 - \lambda_a) \sum_{i=1}^n x_i} \stackrel{H_0}{>} \stackrel{H_a}{<} K \cdot \left(\frac{\lambda_a}{\lambda_0}\right)^n$$

$$\stackrel{\text{II}}{\iff} (\lambda_a - \lambda_0) \sum x_i \stackrel{H_0}{>} \stackrel{H_a}{<} \ln k + \frac{n \ln \frac{\lambda_a}{\lambda_0}}{\lambda_0}$$

$\Leftrightarrow$

$$\sum_{i=1}^n x_i$$

$$\begin{matrix} > \\ < \\ H_a \end{matrix}$$

$$\frac{\ln k + n \ln \frac{\lambda_a}{\lambda_0}}{\lambda_a - \lambda_0}$$

$$\frac{1}{n} \sum_{i=1}^n x_i$$

$$\begin{matrix} > \\ < \\ H_a \end{matrix}$$

$$\frac{\ln k + n \ln \frac{\lambda_a}{\lambda_0}}{n(\lambda_a - \lambda_0)}$$

$$k'$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\begin{matrix} > \\ < \\ H_a \end{matrix}$$

$$k'$$

$$R_a = \{\lambda_a\}$$

$$\hookrightarrow \text{lik}(\lambda_a) = \underbrace{\lambda_a e^{-\lambda_a \sum x_i}}_{\neq 0}$$

next step will be to find  $k'$ ?

from now on we assume that

$n$  is large (hence CLT is a valid approximation)

$$\alpha = \Pr \{ \text{type I error} \}$$

$$\underline{\alpha} = \Pr \{ H_0 \text{ is rejected} \mid H_0 \text{ is true} \}$$

$$= \Pr \left\{ \bar{X} < \underline{k}' \mid H_0 \text{ is true} \right\}$$

assuming  $H_0$  is true,

what's the distribution of  $\frac{\sum_{i=1}^n x_i}{n}$ ?

$$x_i \stackrel{iid}{\sim} \exp(\lambda_0)$$

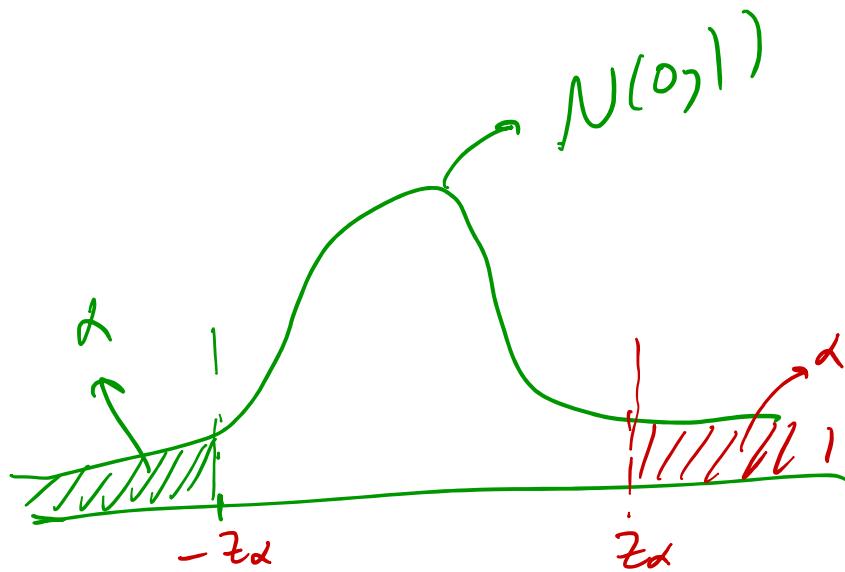
$$\begin{cases} E[x_i] = \frac{1}{\lambda_0} \\ \text{var}[x_i] = \frac{1}{\lambda_0^2} \end{cases} \rightarrow \frac{\sum_{i=1}^n x_i}{n} \sim \mathcal{N}\left(\frac{1}{\lambda_0}, \frac{1}{n\lambda_0^2}\right)$$

$$\alpha = \Pr \left\{ \mathcal{N}\left(\frac{1}{\lambda_0}, \frac{1}{n\lambda_0^2}\right) < \underline{k}' \right\}$$

$$\alpha = \Pr \left\{ \frac{1}{\lambda_0} + \frac{1}{\sqrt{n}\lambda_0} N(0, 1) < \underline{k}' \right\}$$

$$\lambda = \Pr \left\{ N(0, 1) < \frac{k' - \frac{1}{\lambda_0}}{\frac{1}{\lambda_0 \sqrt{n}}} \right\}$$

$-z_\alpha$



$$\frac{k' - \frac{1}{\lambda_0}}{\frac{1}{\lambda_0 \sqrt{n}}} = -z_\alpha$$

$$\Rightarrow k' = -\frac{z_\alpha}{\lambda_0 \sqrt{n}} + \frac{1}{\lambda_0}$$

So far our framework has been to fix the 'significance level  $\alpha$  (and try to minimize  $\beta$ ). Now, the question that needs to be answered is what is the best (smallest)  $\beta$  that we could possibly have? And which test would lead to this  $\beta$ ? The likelihood test is a very powerful test which gives us the most powerful test (least  $\beta$ ) in several cases.

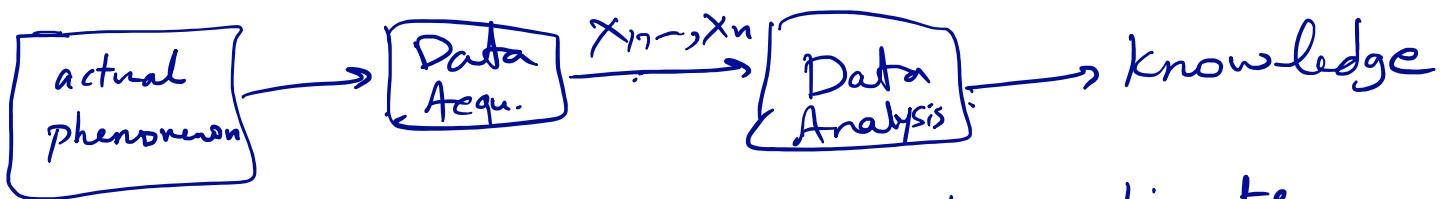
Neyman - Pearson Lemma:

For the problem where  $\mathcal{R}_0 = \{\Theta_0\}$  and  $\mathcal{R}_a = \{\Theta_a\}$ , the likelihood test is the most powerful test among all the tests with significance level  $\alpha$ .

(most powerful = least  $\beta$ ).

## Lecture 13:

### Module 3: Supervised learning



Module 1,2: {

- studied how to estimate different parameters connected to the underlying phenomenon
- how different hypotheses can be evaluated from data

Module 3: "learn" some "predictive relations"  
regarding the phenomenon.

#### Supervised Learning:

We assume that each data point  $x_i$  is associated with a "label" (which is a part of the data). That is, each data point is of the form

$(x_i, y_i)$ , where  $y_i$  is the associated label.

Examples:

$\left\{ \begin{array}{l} x_i \text{ could be the text of an email} \\ y_i \text{ could be spam/harm} \end{array} \right.$

$\left\{ \begin{array}{l} x_i \text{ could be an image (street view) image} \\ y_i \text{ could whether/not there is a car in that image} \end{array} \right.$

$\left\{ \begin{array}{l} x_i \text{ could be a vector representing a person's profile} \\ \text{e.g. } x_i = (\text{age}, \text{years of education}, \dots) \end{array} \right.$

$y_i = \text{person's income}$

---

we're interested in understanding the relation between  $x_i$  and  $y_i$ , i.e.

$$y = f(x) + \epsilon$$

the relation (function)  $f$  is to  
be learnt from sample data  
 $(x_1, y_1), \dots, (x_n, y_n)$ .

Let's denote the function that we  
learn by  $\hat{f}$ : Once we learn  $\hat{f}$ , we  
can use it for prediction:

If a new email,  $x_{\text{new}}$  arrives, we  
can compute  $\hat{y} = \hat{f}(x_{\text{new}})$  and predict  
using  $\hat{y}$  if it is spam or ham.  
As another example:  $\hat{y} = \hat{f}(x_{\text{new}})$   
 $x_{\text{new}}$ : person's profile  
 $\hat{y}$ : income give  
the person's  
later.

---

Main Question: How do we find  $\hat{f}$ ?

Setting: Data :  $(x_1, y_1), \dots, (x_n, y_n)$

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$$

- An image with pixels
- profile of the  $i$ -th person

$y_i$  = label

supervised learning

Regression =  $y_i \in \mathbb{R}$   
(label could be any real number)

classification =  
 $y_i \in A$  ( $A$  is a discrete set)  
e.g.  $A = \{0, 1\}$

Let's start with the simplest setting (for regression):

1-dimensional linear regression:

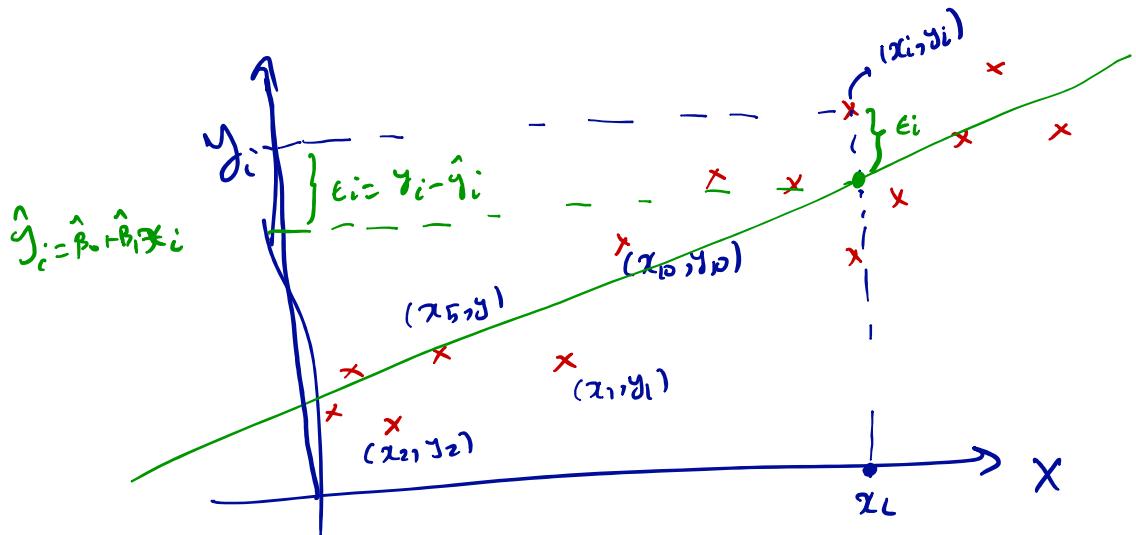
$$(x_1, y_1), \dots, (x_n, y_n)$$

$$\begin{cases} x_i \in \mathbb{R} \\ y_i \in \mathbb{R} \end{cases} \quad (p=1)$$

Goal: learn a relation  $\hat{y} = \tilde{f}(x) + \epsilon$

(Example,  $x$  = blood sugar level at the age of 20  
 $y$  = probability of having Diabetes at 40.)

or  $x = \text{years of education}$  ).  
 $y = \text{income}$



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

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

the error term models what we miss with this simple linear model: the true relation is probably not linear as there may be other variables

Goal: "learn" or "estimate" the parameters  
 $\hat{\beta}_0, \hat{\beta}_1$

Question: What should the basic principle be in learning  $\hat{\beta}_0$  and  $\hat{\beta}_1$ ?

best fit, lowest error, --

Option 1: Minimize  $\hat{\beta}_0, \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n |\epsilon_i|$  (hard to minimize)

$$= \min_{\hat{\beta}_0, \hat{\beta}_1} \frac{1}{n} \sum_{i=1}^n |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|$$

$$y = mx$$

option 2:

$$\text{minimize}_{\hat{\beta}_0, \hat{\beta}_1} \frac{1}{n} \sum_{i=1}^n e_i^2$$

easy to minimize

$$= \text{minimize}_{\hat{\beta}_0, \hat{\beta}_1}$$

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$g(\hat{\beta}_0, \hat{\beta}_1)$

to find  $\hat{\beta}_0, \hat{\beta}_1$  → derivate = 0

$$(1) \frac{\partial g}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$(2) \frac{\partial g}{\partial \hat{\beta}_1} = -\frac{2}{n} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$(1) -\frac{1}{n} \sum_{i=1}^n y_i + \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{\beta}_0}_{\hat{\beta}_0} + \underbrace{\frac{\hat{\beta}_1}{n} \sum_{i=1}^n x_i}_{\hat{\beta}_1 \bar{x}} = 0$$

$-\bar{y}$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}^{(*)}$$

$$(2) \quad \frac{1}{n} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \bar{x} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i^2}{n} = 0$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^n x_i y_i = \underbrace{\hat{\beta}_0 \bar{x}} + \hat{\beta}_1 \frac{\sum_{i=1}^n x_i^2}{n}$$

$$\stackrel{(*)}{\Rightarrow} \frac{1}{n} \sum_{i=1}^n x_i y_i = (\bar{y} - \hat{\beta}_1 \bar{x}) \bar{x} + \hat{\beta}_1 \frac{\sum_{i=1}^n x_i^2}{n}$$

$$\Rightarrow \frac{1}{n} \sum x_i y_i = \bar{x} \bar{y} + \hat{\beta}_1 \left( \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 \right)$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \hat{\beta}_1 \left( \underbrace{\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2}_{\text{brace}} \right)$$

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2}$$

$$\begin{cases} \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

$$\hat{B}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

using (1)

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x} = \bar{y} - \bar{x} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Data  $(x_i, y_i) \rightarrow \hat{B}_0, \hat{B}_1$

Question: How good are these estimates?

To evaluate the performance of our estimates of  $\hat{B}_0, \hat{B}_1$ , let's assume that the data is generated according to the following relation:

$$y = B_0^{\text{true}} + B_1^{\text{true}} x + \epsilon,$$

where  $\epsilon$  is an independent zero-mean noise with  $\text{Var}(\epsilon) = \sigma_{\text{noise}}^2$ .

Given  $n$  data points  $(x_i, y_i)$ ,  $i=1, \dots, n$ ,  
 that are generated iid from this model,  
 we find estimates  $\hat{\beta}_0, \hat{\beta}_1$  according to the  
 formulas derived above..

Question: How close is  $\hat{\beta}_0$  to  $\beta_0^{\text{true}}$  and  
 $\hat{\beta}_1$  to  $\beta_1^{\text{true}}$ ?

Note  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are estimated  
 using data and are random quantities

$\left\{ \begin{array}{l} \hat{\beta}_0 \text{ is the estimate of } \beta_0^{\text{true}} \\ \hat{\beta}_1 \text{ is " " " } \beta_1^{\text{true}} \end{array} \right.$

(1) Are these unbiased estimates?

Let's consider  $\hat{B}_1$ .

$$\hat{B}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$y_i = B_0^{\text{true}} + B_1^{\text{true}} x_i + \hat{\epsilon}_i$$

$$E_{\text{data}} \left[ \hat{B}_1 \right] \stackrel{?}{=} B_1^{\text{true}}$$

## Lecture 15:

A General frame that we've been developing:

Data:  $(x_1, y_1), \dots, (x_n, y_n)$

Goal:  $\hat{y} = \hat{f}(x)$

Step (1)

define a parametric model that relates the label to the input

Step (2)

define a metric that measures how the model fits wrt training data

$$x_i \in \mathbb{R}$$

$$y_i \in \mathbb{R}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \leftarrow \text{Step (1)}$$

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &\downarrow \text{residual sum of squares} \end{aligned} \quad \leftarrow \text{Step (2)}$$

$$\text{For linear Reg. : } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\text{minimize RSS} \quad \left\{ \begin{array}{l} \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{array} \right.$$

For the analysis:

$$y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} x_i + \epsilon_i$$

<sup>zero-mean</sup>  
independent  
 $\text{var}(\epsilon_i) = \sigma^2_{\text{noise}}$

Question:

$$E[\hat{\beta}_0] \stackrel{?}{=} \beta_0^{\text{true}}, \quad E[\hat{\beta}_1] \stackrel{?}{=} \beta_1^{\text{true}}$$

We assume fixed  $x_1, \dots, x_n \in \mathbb{R}$  and  $y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} x_i + \epsilon_i$

↗ random

$$E[\hat{\beta}_1] = E \left[ \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} x_i + \epsilon_i$$

$$\bar{y} = \beta_0^{\text{true}} + \beta_1^{\text{true}} \bar{x} + \bar{\epsilon} \quad | \quad \bar{\epsilon} = \frac{\epsilon_1 + \dots + \epsilon_n}{n}$$

$$E[\hat{\beta}_1] = E\left[ \frac{\sum_{i=1}^n (x_i - \bar{x}) (B_1^{\text{true}}(x_i - \bar{x}) + \epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$= E\left[ \frac{\sum_{i=1}^n B_1^{\text{true}} (x_i - \bar{x})^2 + \sum_{i=1}^n (\epsilon_i - \bar{\epsilon})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$= \beta_1^{\text{true}} E\left[ \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] + E\left[ \frac{\sum_{i=1}^n (\epsilon_i - \bar{\epsilon})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

1      2

$E[\epsilon_i] - E[\bar{\epsilon}]$

$$= \beta_1^{\text{true}} + \frac{\sum_{i=1}^n (x_i - \bar{x}) (E[\epsilon_i - \bar{\epsilon}])}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

0

$$= \beta_1^{\text{true}}$$

In general :

$$E[\hat{B}_1] = \beta_1^{\text{true}}$$

$$E[\hat{B}_0] = \beta_0^{\text{true}}$$

$$\begin{aligned} \text{Var}[\hat{B}_0] &= E[(\hat{B}_0 - \beta_0^{\text{true}})^2] \\ &= \sigma_{\text{noise}}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \end{aligned}$$

$$\text{Var}[\hat{B}_1] = \frac{\sigma_{\text{noise}}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

---

### Multi-dimensional Linear Regression:

---

Data:

$$(x_1, y_1), \dots, (x_n, y_n)$$

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$$

$$y_i \in \mathbb{R}$$

linear :

$$x_i = (x_{i1}, \dots, x_{ip})$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}$$

(Step<sup>(1)</sup>)

$\leftarrow$  p+1 parameters  $\hat{\beta}_0, \dots, \hat{\beta}_p$

(Step<sup>(2)</sup>)

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RSS(\hat{\beta}_0, \dots, \hat{\beta}_p) = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}))^2$$

. we'd like to solve  $\min_{\hat{\beta}_0, \dots, \hat{\beta}_p} RSS(\hat{\beta}_0, \dots, \hat{\beta}_p)$

The Derivations to find the best choice for the parameters:

$$x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p \rightarrow \tilde{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^{p+1}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

$$RSS = \sum_{i=1}^n (y_i - \tilde{x}_i \hat{\beta})^2$$

$$\begin{pmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip} \\ = \tilde{x}_i \hat{\beta} \end{pmatrix}$$

Define the data matrix as:

$$X = \begin{bmatrix} \frac{\tilde{x}_1}{\tilde{x}_2} \\ \vdots \\ \frac{\tilde{x}_n}{\tilde{x}_n} \end{bmatrix}_{n \times (p+1)}$$

label vector

$$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$RSS = \| y - X \hat{\beta} \|^2$$

Where  $\| (z_1, \dots, z_n) \|^2 = \sum_{i=1}^n z_i^2$   $(\|\cdot\| \text{ is the Euclidean norm})$

Goal:  
minimize  $RSS(\hat{\beta})$

$$\Rightarrow \nabla RSS(\hat{\beta}) = 0$$
$$\nabla (\| y - X \hat{\beta} \|^2) = -2 X^T (y - X \hat{\beta})$$

$$\nabla \text{RSS} = 0 \implies -2x^T(y - x\hat{\beta}) = 0$$

$$\Rightarrow x^T x \hat{\beta} = x^T y$$

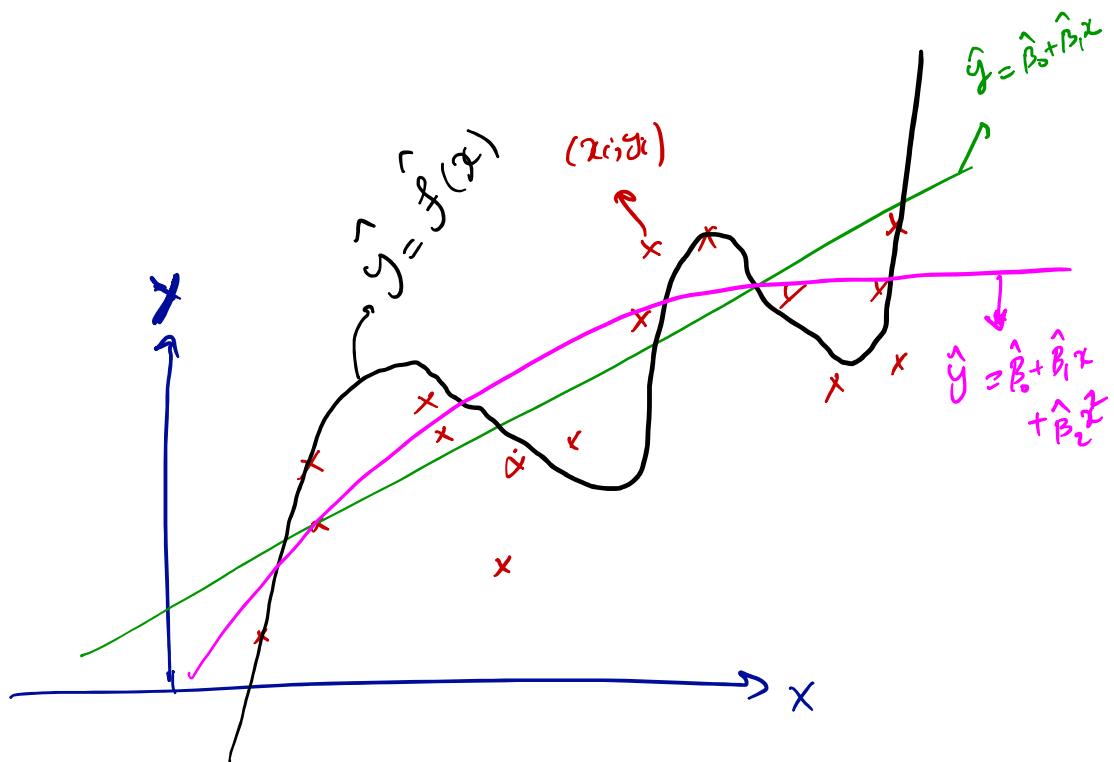
$$\Rightarrow \hat{\beta} = (x^T x)^{-1} x^T y$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$x = \left( \begin{array}{c} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{array} \right) = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & & x_{np} \end{pmatrix}$$

Let's go back to the 1-D setting

Data:  $\{(x_i, y_i)\}_{i=1}^n$   $x_i, y_i \in \mathbb{R}$



linear:  $y = \hat{\beta}_0 + \hat{\beta}_1 x$

polynomial:  $y = p(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 + \dots + \hat{\beta}_d x^d$

One way to define non-linear parametric models  
is by using polynomial models:

{ degree of the model = d  
parameters  $\hat{\beta}_0, \dots, \hat{\beta}_d$  }  $\rightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \dots + \hat{\beta}_d x^d$

E.g.

$$d=2$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$$

Step (2): "train" the parametric model

(let's assume)  
d=2 for simplicity

$$\begin{aligned} RSS &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_i + \hat{\beta}_2 \tilde{x}_i^2))^2 \end{aligned}$$

Multi-dim linear regression

$$y_i = (x_{i1}, \dots, x_{ip})$$

$$RSS = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}))^2$$

$$\hat{\beta} = (\vec{X}^T \vec{X})^{-1} \vec{X}^T \vec{y}$$

$$\vec{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})$$

$$\Rightarrow RSS = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 \vec{x}_{i1} + \hat{\beta}_2 \vec{x}_{i2} + \dots + \hat{\beta}_d \vec{x}_{id}))^2$$

The by drawing this analogy between

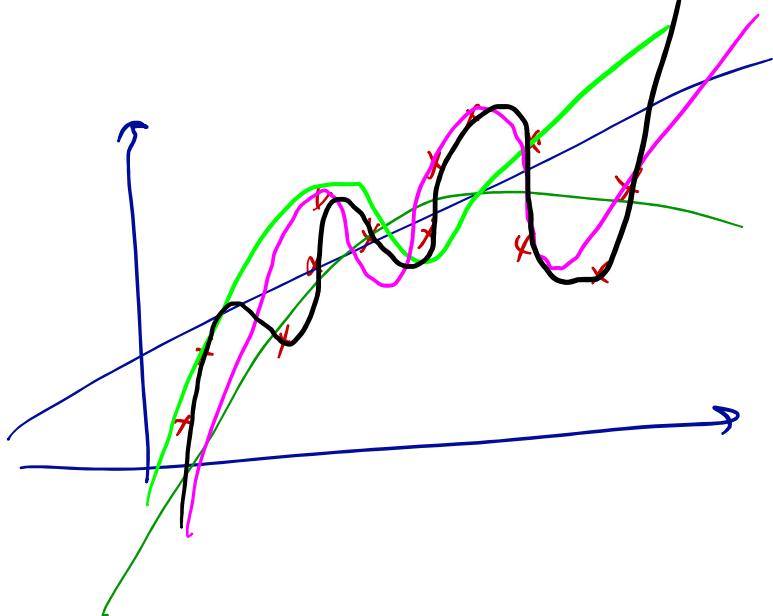
Non-linear (polynomial) regression and multi-dim linear regression we obtain:

a closed form solution for polynomial regression:

$$\hat{B} = \begin{pmatrix} \hat{B}_0 \\ \vdots \\ \hat{B}_d \end{pmatrix}$$

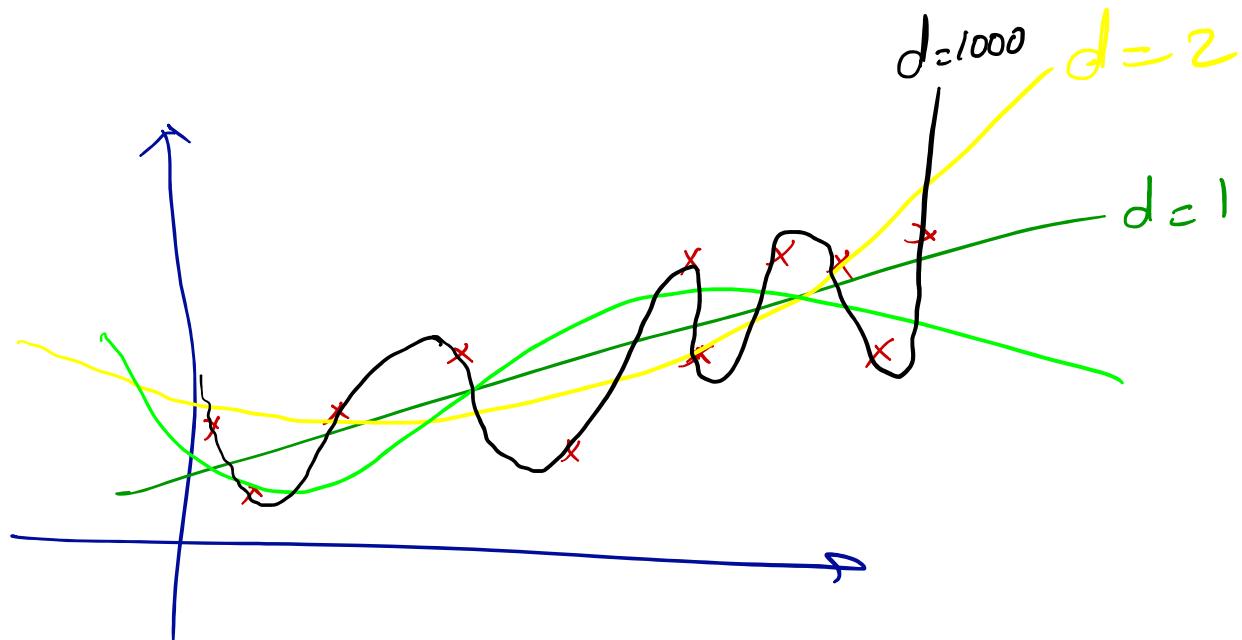
$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & & x_1^d \\ 1 & x_2 & x_2^2 & & x_2^d \\ 1 & x_3 & . & & x_3^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & & x_n^d \end{pmatrix}_{n \times (d+1)}$$

$$\hat{B} = (X^T X)^{-1} X^T y.$$



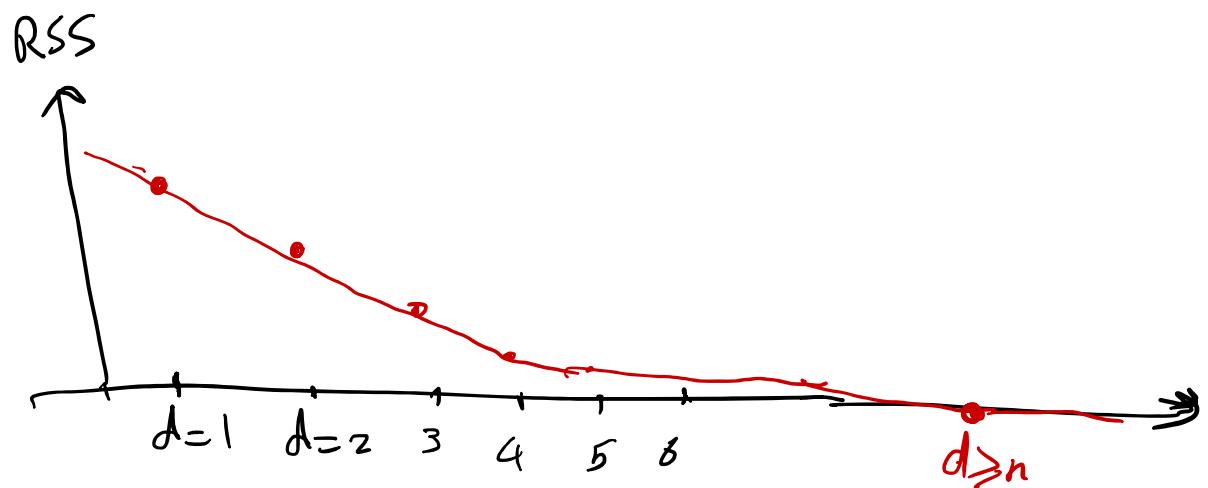
## Lecture 16 :

Data :  $(x_1^{t^R}, y_1^{t^R}), \dots, (x_n, y_n)$



$$\hat{y} = p(x)$$

degree = d



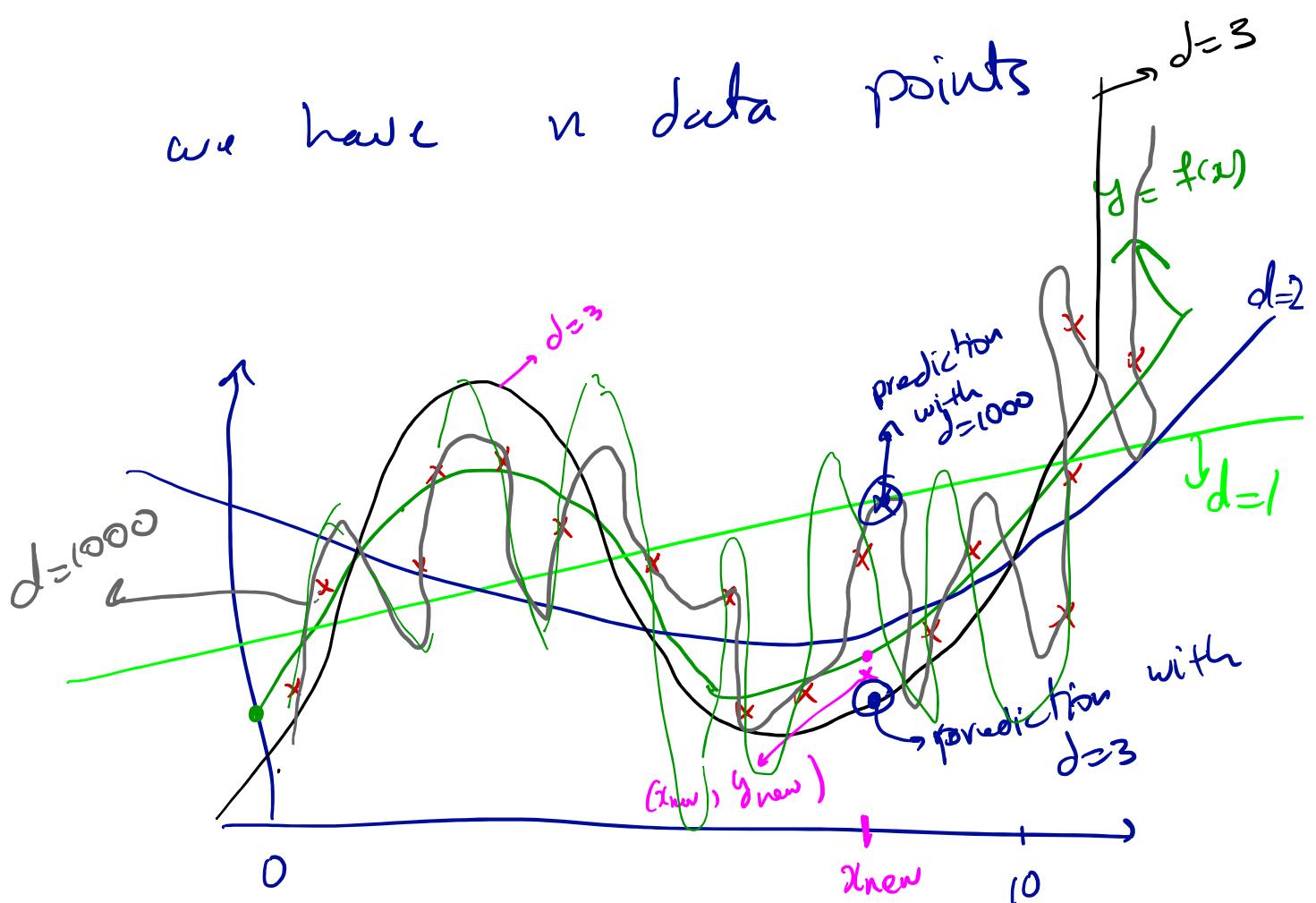
Let's assume that the data is generated by the following model:

$$y_i = \underbrace{f(x_i)}_{\text{true function}} + \epsilon_i^{\text{noise}}$$

$$f(x) = \underline{x^3 + 2x^2 - 3x + 1}$$

$$x_i \in \text{Uniform } [0, 10]$$

we have  $n$  data points



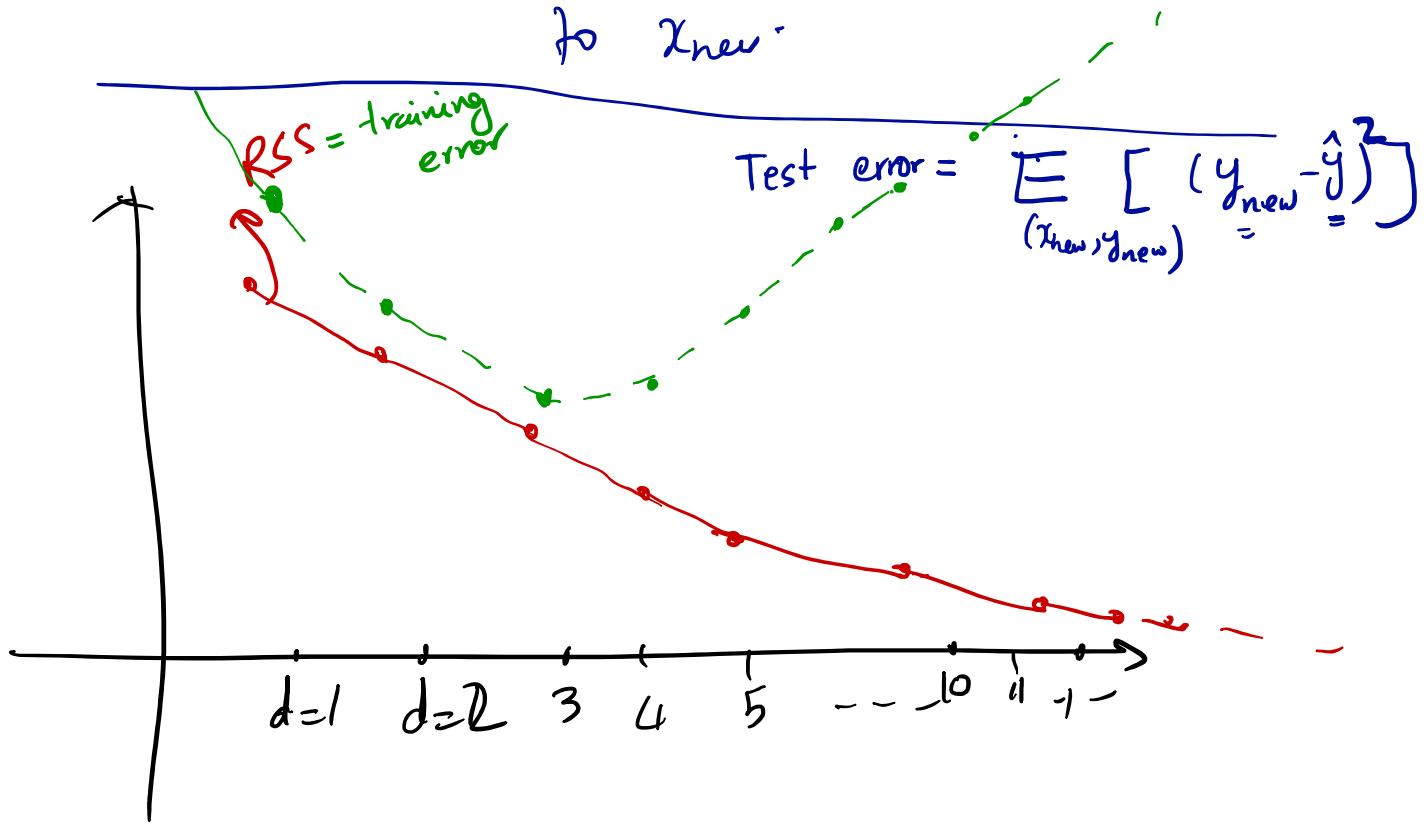
Let's now assume that we'd like find the best polynomial fit to the training data.

\* recall that the main goal of supervised learning is to predict as accurately as possible on new, unseen inputs.

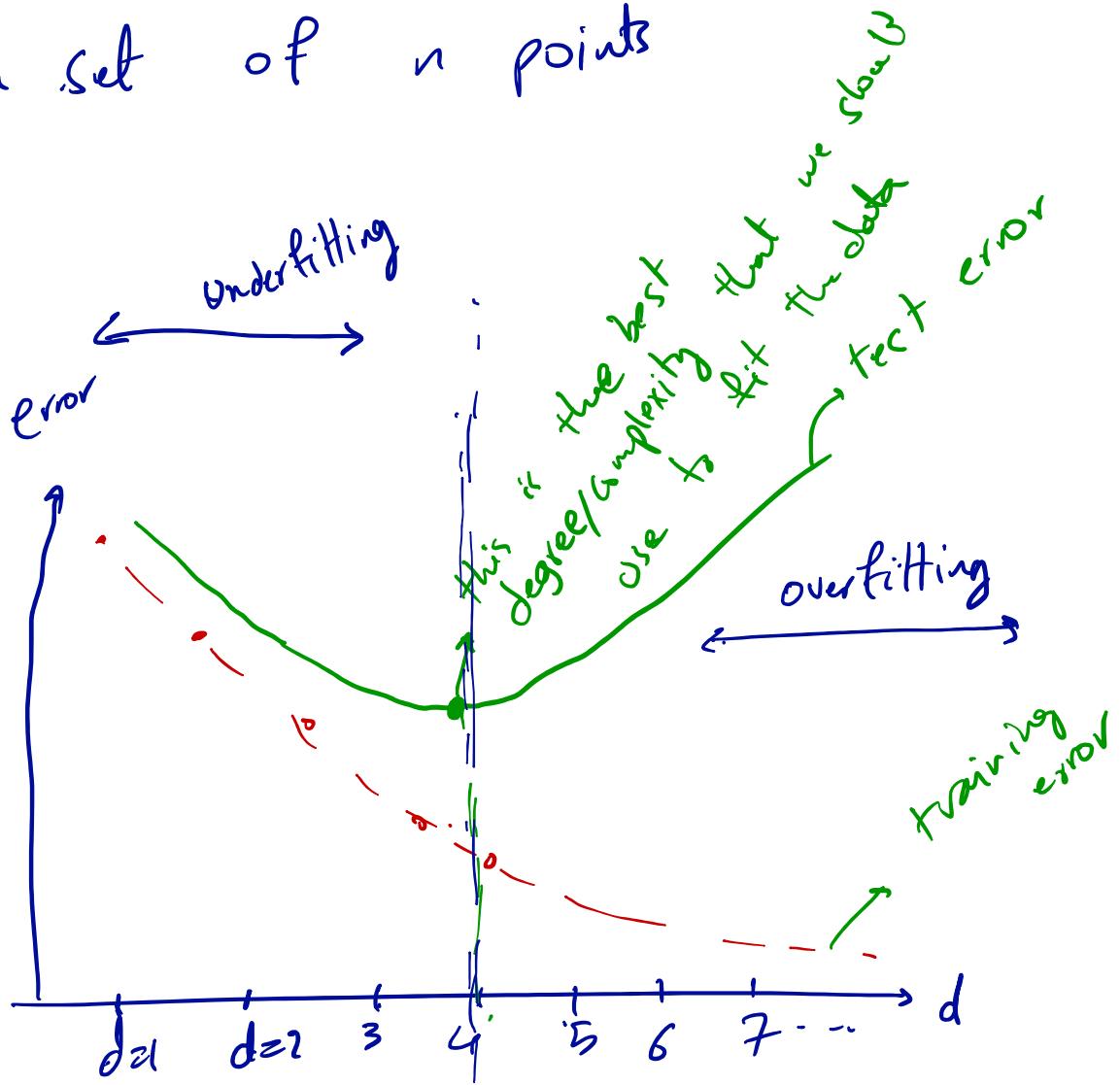
⇒ On a new input

$$x_{\text{new}} \in [0, 10] \rightarrow \underline{\hat{y}_{\text{new}}} = f(x_{\text{new}}) + \epsilon$$

} predict the label associated to  $x_{\text{new}}$



In practice, we only have access to a data set of  $n$  points



We have two regions' depending on the complexity (degree) of the parametric model that we're fitting to the data:

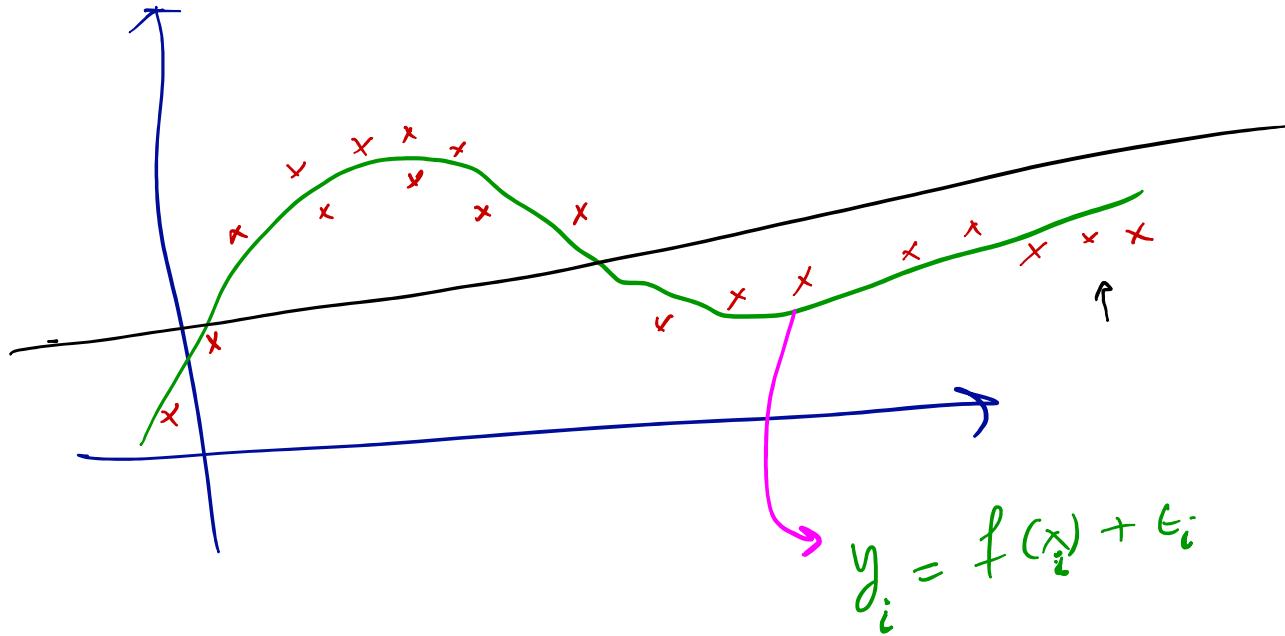
Underfitting : - model is not sufficiently complex to explain the data.  
( high bias )

- the trained model does not change much if we add new training data points  
( low variance )

Overfitting :

- model is too complex for training the data  
( low bias )
- the trained model will change significantly if we add new data points.  
( high variance )

Underfitting :

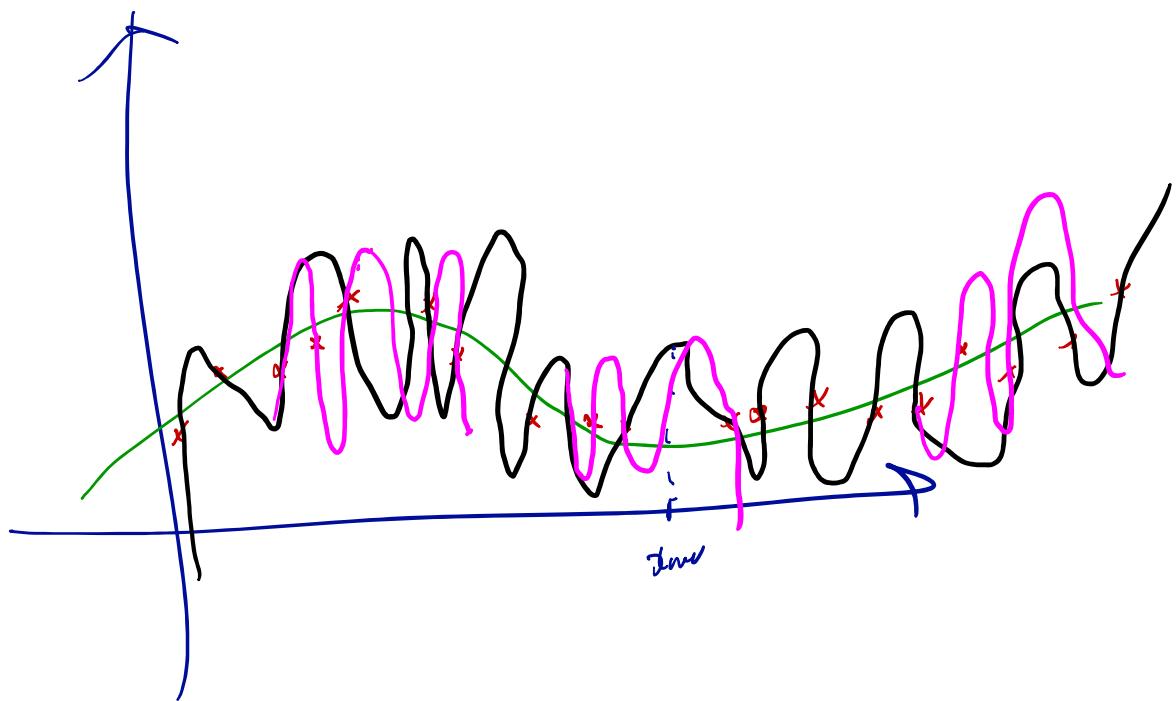


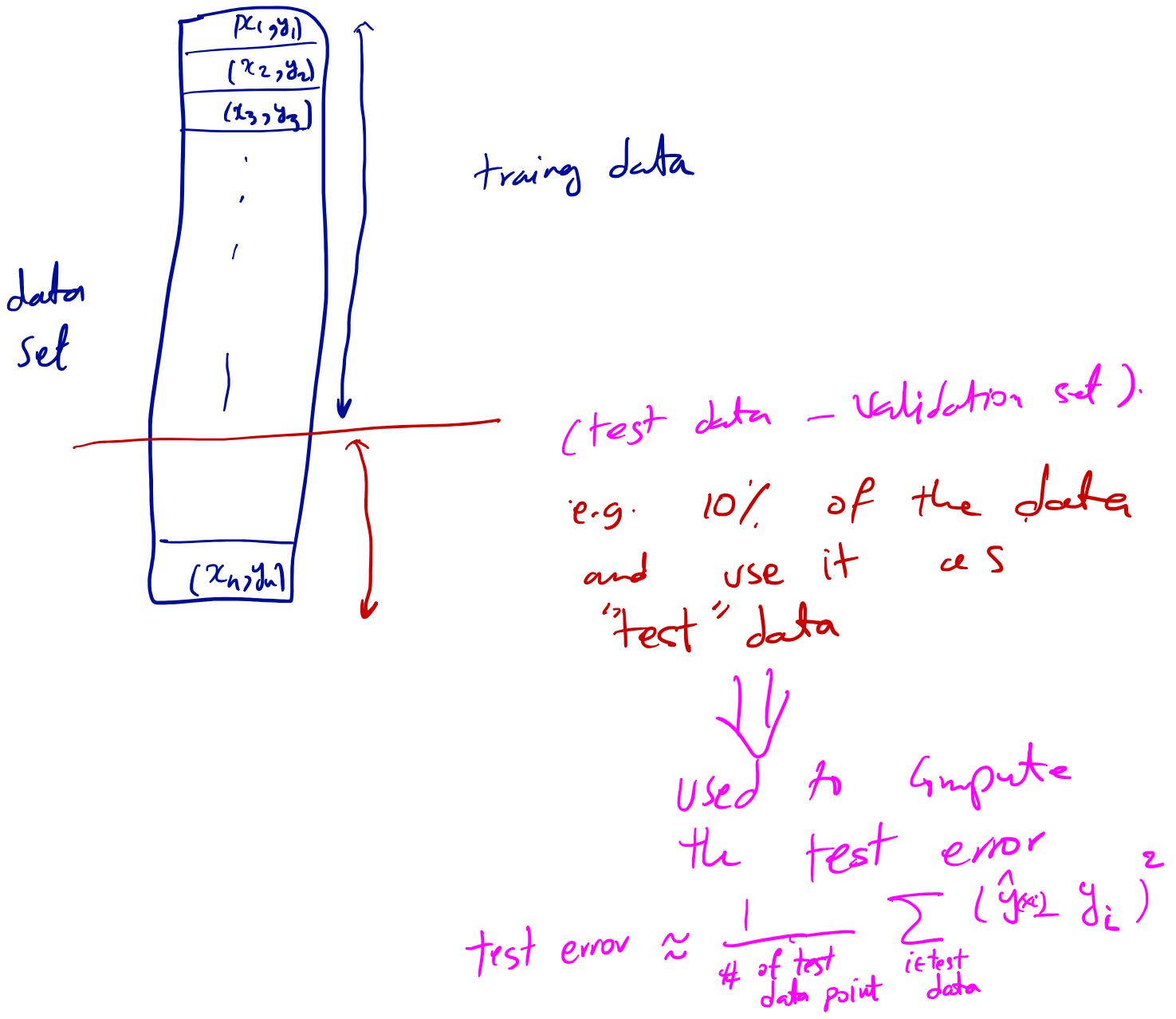
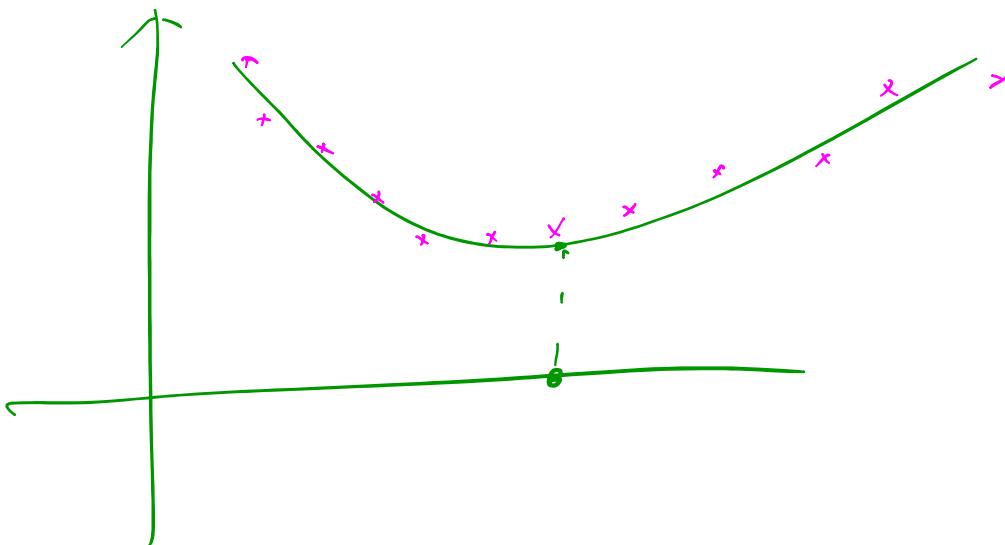
true model: degree 3

{ class of linear models :  
high bias

overfitting .

class of models: degree-100





## Validation/hold-out data

Randomly choose a given portion of data and use it as test data  
(useful when size of data is sufficiently large).

issue: Not using all the available data for training.

## k-fold cross-validation:

Divide the data randomly into  $k$  parts:

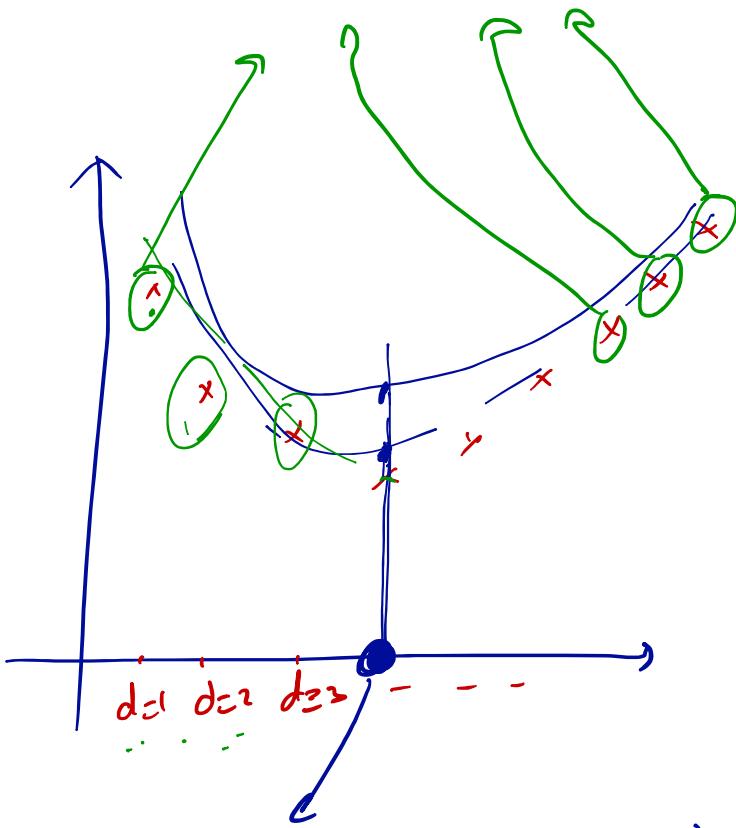
$$D = D_1 \cup D_2 \cup D_3 \dots \cup D_k.$$

for each  $i = 1, \dots, k$

train on  $D \setminus D_i$

test on  $D_i \rightarrow \text{test}_i = \begin{matrix} \text{test error of} \\ \text{the trained model} \\ \text{on } D_i \end{matrix}$

$$\text{test}_{\text{error}} = \frac{1}{K} (\text{test}_1 + \text{test}_2 + \dots + \text{test}_K)$$



Once we find the best complexity  
we train the parameters on the  
whole data set with that complexity

$$k^{20}$$

$$d = 1, 2, 3, 4, \dots, 1000$$

## Lecture 17.

Regression

label  $y_i \in \mathbb{R}$

classification

$y_i \in$  discrete set  
e.g.  $y_i \in \{0, 1\}$

---

Data:  $(x_i, y_i)$ ,  $i=1, \dots, n$

$$\hat{y}_i = f(x_i)$$

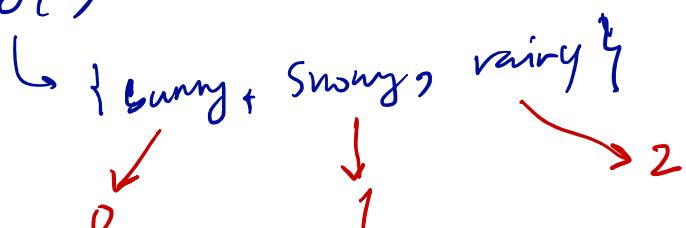
---

Example:

$x_i$  = weather data for today

$y_i = (\text{rainy}, \text{snowy}, \text{sunny})$

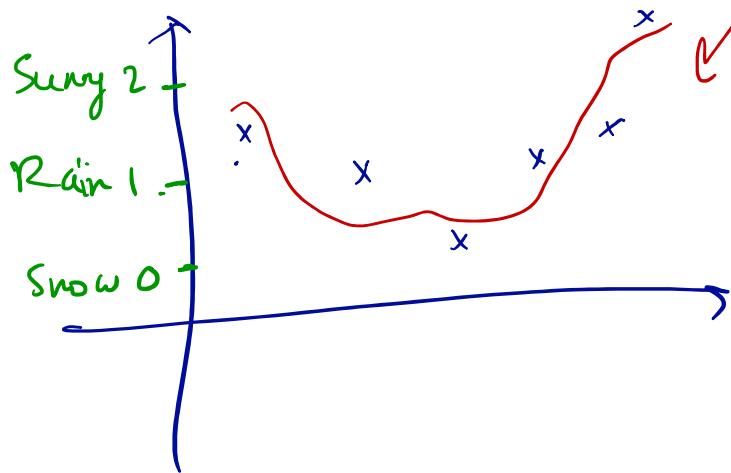
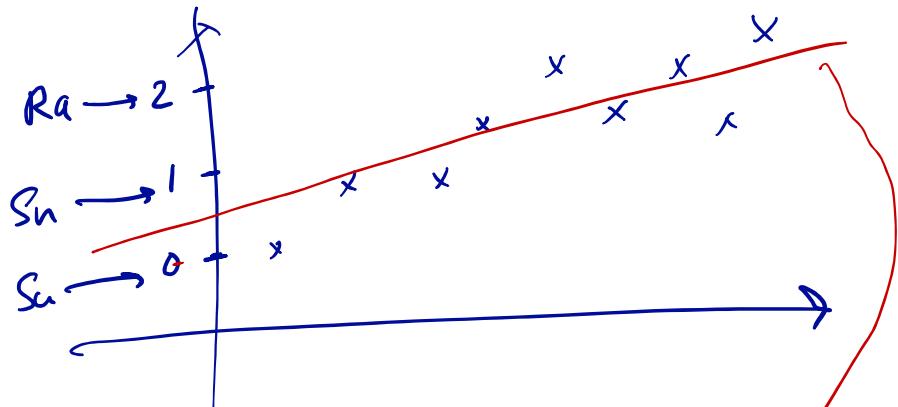
$(x_i, y_i)$ ,  $i=1, \dots, n$



$(x_i, y_i)$

As a naive idea, we can fit a regression model to the above setting.

$$Y_i = \left\{ \begin{array}{l} S_0 \\ S_n \\ R_a \end{array} \right\} \quad \left\{ \begin{array}{l} 0 \\ 1 \\ 2 \end{array} \right\}$$



does not work X

Recall:

(1)

for Parametric model

(2)

find the best fit to the data

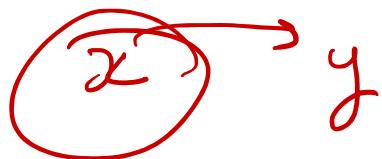
$$y_i = f(x_i) + \epsilon_i$$

---

$S_u$	$\rightarrow$	x
		$n_1$
$S_n$	$\rightarrow$	x
		$n_2$
$R_a$	$\rightarrow$	x
		$n_3$

---

0.2	$\Pr \{ \text{sunny}   x \}$	e $\{0, 1\}$
0.5	$\Pr \{ \text{snowy}   x \}$	e $[0, 1]$
0.3	$\Pr \{ \text{rainy}   x \}$	e $[0, 1]$


 $x \rightarrow y$

---

What we really need to learn/estimate  
is the conditional probabilities:

$$\Pr \{ \text{class}_k | x \} \quad \text{for } k=1, \dots, K$$

For now assume that we have two classes (binary classification problem)

class 0  $\rightarrow y_i = 0$   $x \in \mathbb{R}$

class 1  $\rightarrow y_i = 1$

Goal: given  $x \rightarrow \Pr\{0|x\}$   
 $\Pr\{1|x\}$ .

specifically, we're going to devise  
a parametric model for estimating

$\Pr\{0|x\}$ ,  $\Pr\{1|x\}$ .  
=

Let's begin with linear models.

What is a linear model in the binary classification setting?

naive approach:

$$P\{0|x\} = \beta_0 + \beta_1 x$$

$$P\{1|x\} = 1 - P\{0|x\}$$

of course, we should ensure that probabilities are between 0 and 1.  
So the above parametric model does not work.

$$\left. \begin{array}{l} P\{0|x\} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \in [0,1] \\ P\{1|x\} = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \in [0,1] \end{array} \right\}$$

So the parametric model is given as above. (Step (1) ✓)

Step (2) : fit the parameters to training data.

---

Data:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Model :  $\Pr\{y_i=0 | x_i\} = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$

(also called  
Logistic Regression)

$$\Pr\{y_i=1 | x_i\} = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

---

How do we learn  $\beta_0$  and  $\beta_1$ ? MLG

$$\begin{aligned} \text{lik}(\beta_0, \beta_1) &= \prod_{i=1}^n \Pr\{x_i, y_i | \beta_0, \beta_1\} \\ &= \prod_{i=1}^n \underbrace{\Pr\{x_i\}}_? \times \underbrace{\Pr\{y_i | x_i, \beta_0, \beta_1\}} \\ &= \prod_{i=1}^n \Pr\{x_i\} \times \begin{cases} \text{if } y_i=0 \quad \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(-)} \\ \text{if } y_i=1 \quad \frac{1}{1 + \exp(-)} \end{cases} \end{aligned}$$

$$\arg \max_{B_0, B_1} \text{lik}(B_0, B_1)$$

does not depend on  $B_0, B_1$

$$= \arg \max_{B_0, B_1} \prod_{i=1}^n \Pr\{x_i\} \times \prod_{i:y_i=0} \frac{\exp(B_0 + B_1 x_i)}{1 + \exp(B_0 + B_1 x_i)} \\ \times \prod_{i:y_i=1} \frac{1}{1 + \exp(B_0 + B_1 x_i)}$$

$$= \arg \max_{B_0, B_1} \prod_{i:y_i=0} \frac{\exp(B_0 + B_1 x_i)}{1 + \exp(-)} \quad \prod_{i:y_i=1} \frac{1}{1 + \exp(-)}$$

$$= \arg \max_{B_0, B_1} \frac{\prod_{i:y_i=0} \exp(B_0 + B_1 x_i) \prod_{i:y_i=1} 1}{\prod_{i:y_i=0} (1 + \exp(B_0 + B_1 x_i)) \prod_{i:y_i=1} (1 + \exp(B_0 + B_1 x_i))}$$

$$= \arg \max_{B_0, B_1} \frac{\prod_{i:y_i=0} \exp(B_0 + B_1 x_i)}{\prod_{i=1}^n (1 + \exp(B_0 + B_1 x_i))}$$

We need to solve the argmax problem.

$$\frac{\partial \log l(B_0, B_1)}{\partial B_0} = 0$$

$$\frac{\partial \log l(B_0, B_1)}{\partial B_1} = 0$$

$\Rightarrow$  there is no closed form <sup>solution</sup> to these equations.

$\Rightarrow$  from now on,  $\hat{B}_0, \hat{B}_1$  will be the solutions of the above equations.

---

$$x \in \mathbb{R} \rightarrow x \in \mathbb{R}^P \rightarrow x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$$

$$\Pr\{0|x_i\} = \frac{\exp(B_0 + B_1 x_{i1} + B_2 x_{i2} + \dots + B_p x_{ip})}{1 + \exp(B_0 + B_1 x_{i1} + \dots + B_p x_{ip})}$$

$$\Pr\{1|x_i\} = 1 - \Pr\{0|x_i\}$$

$\Rightarrow$  formulate the likelihood function

$\Rightarrow$  MLE to find  $\hat{B}_0, \hat{B}_1, \dots, \hat{B}_p$ .

$$P(0|x) = \frac{\exp(B_0 + B_1 x)}{1 + \dots}$$

$$P(1|x) = \frac{1}{1 + \exp(\dots)}$$

$$\log\left(\frac{P(0|x)}{P(1|x)}\right) = B_0 + B_1 x$$

*wgit*

So once we find  $\hat{\beta}_0, \hat{\beta}_1$ , in order to predict the class on a new data point  $x$ , we should do as follows:

$$x \rightarrow \begin{cases} \Pr\{0|x\} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\dots)} & \text{prediction:} \\ \Pr\{1|x\} = 1 - \Pr\{0|x\} & \text{Pick the class with highest probability} \end{cases}$$

to summarize:

Classification  $\xrightarrow{\text{1) estimate } \Pr\{\text{class}_k | x\}}$   
(2) choose the class with  
the highest (estimated)  
probability.

---

Bayes Optimal Classifier:

Assume for the moment that all  
the conditional probabilities are known.

$\Pr\{\text{class}_k | x\}$  is known.

How should we (optimally) predict  
the class associated to input  $x$ ?

Bayes optimal classifier:

$\hat{y}_{\text{Bayes}}(x) = \text{predicted class}(x) = \underset{k}{\operatorname{argmax}} \Pr\{\text{class}_k | x\}$

Theorem: The Bayes optimal classifier has  
the smallest classification error  
among all the possible classifiers.

That means: for every other classifier  
 $\hat{y}(x)$ , we have:

$$\Pr_{(x,y)} \{ \hat{y}(x) \neq y \}$$

$$\geq \Pr_{(x,y)} \{ \hat{y}_{\text{Bayes}}(x) \neq y \}$$



## Lecture 18:

$$y \in \{0, 1\} \longrightarrow x \quad \begin{cases} P\{y=0|x\} \\ P\{y=1|x\} \end{cases}$$

---

logistic Reg.:  $P\{y=0|x\}$

$$= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

---

multi-dimensional?  $\rightarrow$   
multiple classes?  $\rightarrow$

# Multi-dimensional Logistic Regression;

Assume the binary classification setting  $y \in \{0, 1\}$ .  $x \in \mathbb{R}^P$ .  
 $x = (x_1, x_2, \dots, x_p)$

$$\Pr\{y=0|x\} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p)}{1 + \exp(\hat{\beta}_0 + \dots + \hat{\beta}_p x_p)}$$

$$\Pr\{y=1|x\} = 1 - \Pr\{y=0|x\}.$$

We'd then find the parameters  $\hat{\beta}_0, \dots, \hat{\beta}_p$  using the mle (from training data).

Multi-class Setting:

(logistic regression is typically used  
for the binary classification problem)  
and often times it does not  
perform well for the multi-class  
setting

$$y \in \{1, 2, 3\} . \quad x \in \mathbb{R}^P \\ x = (x_1, \dots, x_p)$$

$$\Pr \{ y=1 \mid x \} = \frac{\exp(\hat{\beta}_0^1 + \hat{\beta}_1^1 x_1 + \dots + \hat{\beta}_p^1 x_p)}{1 + \exp(\hat{\beta}_0^1 + \dots + \hat{\beta}_p^1 x_p) + \Pr \{ y=2 \}}$$

$$\Pr \{ y=2 \mid x \} = \frac{\exp(\hat{\beta}_0^2 + \hat{\beta}_1^2 x_1 + \dots + \hat{\beta}_p^2 x_p)}{1 + \exp(\hat{\beta}_0^1 + \dots + \hat{\beta}_p^1 x_p) + \Pr \{ y=2 \}}$$

$$\Pr \{ y=3 \mid x \} = 1 - \Pr \{ y=1 \mid x \} - \Pr \{ y=2 \mid x \}$$

$$y=1 \rightarrow \hat{\beta}_0^1, \dots, \hat{\beta}_p^1$$

$$y=2 \rightarrow \hat{\beta}_0^2, \dots, \hat{\beta}_p^2$$



estimated using rule.

---

## Linear Discriminant Analysis (LDA):

logistic Regression:

$$\hookrightarrow \boxed{Pr\{y=k \mid X=x\}}$$

↑  
estimate all these conditional  
probabilities (directly)

LDA: estimate  $Pr\{y=k \mid X=x\}$   
indirectly using the Bayes rule

Assume  $K$  classes:  $y \in \{1, \dots, K\}$

$$\Pr\{y = k \mid x = x\}$$

$$= \Pr\{y = k, x = x\}$$

$$= \frac{\Pr\{x = x\}}{\Pr\{y = k\} \Pr\{x = x \mid y = k\}}$$

$$\Pr\{x = x\}$$

$$= \sum_{k=1}^K \Pr\{x = x, y = k\}$$

$$= \sum_{k=1}^K \Pr\{x = x \mid y = k\} \Pr\{y = k\}$$

$$= \sum_{k=1}^K f_{ik}(x) \pi_k$$

$$\pi_k = \Pr\{y=k\} = \frac{\# \text{ training data points with label } y=k}{\# \text{ total number of data points}}$$

$$= \frac{n_k}{n} \rightsquigarrow \text{unbiased estimator}$$

assume  $x \in \mathbb{R}$

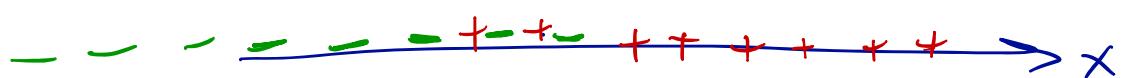
$$f_k(x) = \Pr\{x=x | y=k\} = N(\mu_k, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$$

$(\mu_k, \sigma)$  are parameters to be estimated from data.

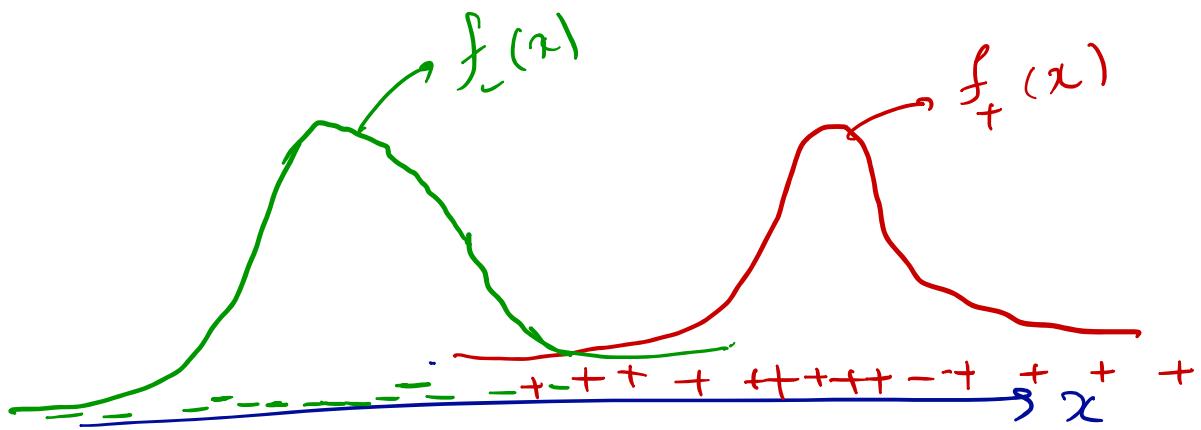
training data in classification:

$$(x_i, y_i) \quad y_i \in \{1, \dots, K\}$$

e.g. when data is 1-d,  $K=2, y \in \{+, -\}$



E.g if data is 1-d and  $y = \{-, +\}$

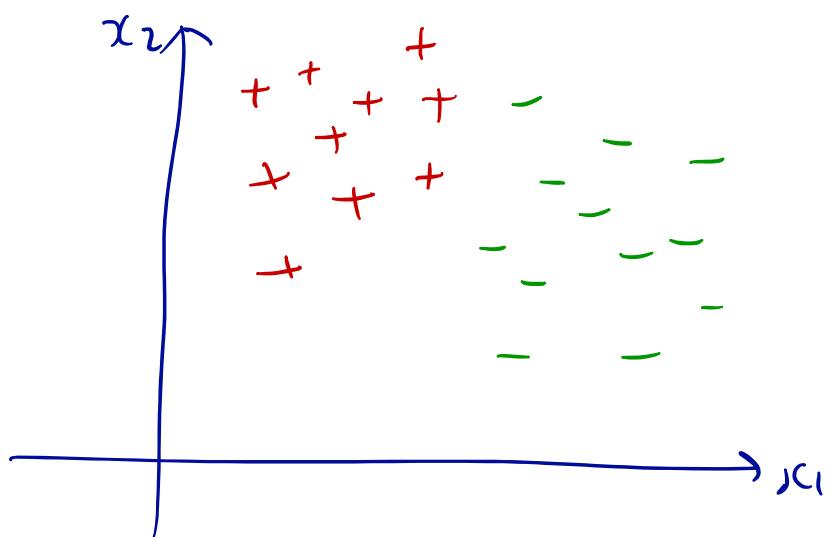


Let's estimate the parameters  
 $\theta$  and  $\mu_k$  using training data-

---

$$x \in \mathbb{R}^2, y = \{-, +\}$$

$$x = (x_1, x_2)$$

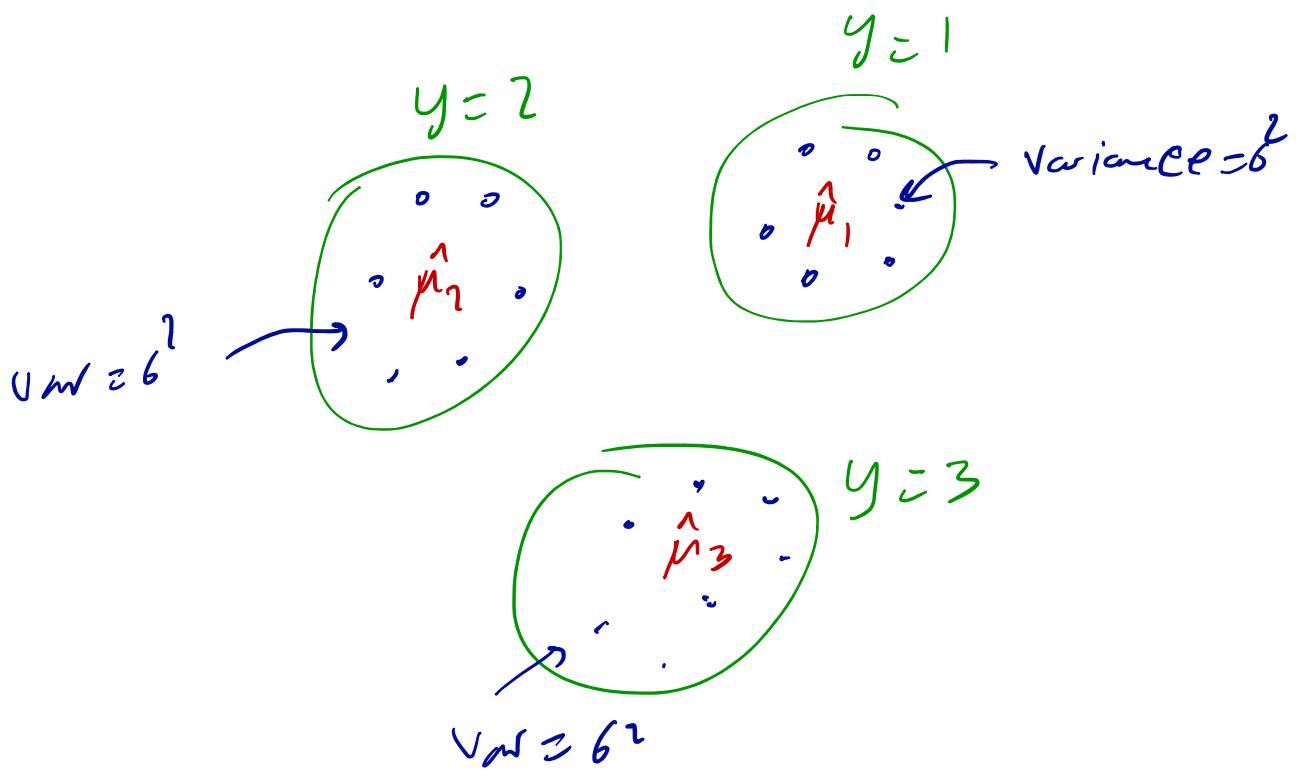


$$\hat{\mu}_k \rightarrow f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$$

$$\hat{\mu}_k = \frac{\sum_{i:y_i=k} x_i}{n_k}$$

estimate of variance inside class  $k$

$$\hat{\sigma}^2 = \frac{\sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2}{n - K}$$



$$f_{ik}(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}^2} \exp\left(-\frac{(x - \hat{\mu}_{ik})^2}{2\hat{\sigma}^2}\right)$$

$$\bar{\pi}_k \approx \hat{\pi}_k = \frac{n_k}{n}$$

Let's predict the class using  
the above (trained) model:

Given a new input  $x$ , our  
prediction will be as follows;

$$\hat{y} = \arg \max_k \Pr\{y=k \mid x=x\}$$

$$= \arg \max_k \frac{\Pr\{x=x \mid y=k\} \Pr\{y=k\}}{\Pr\{x=x\}}$$

independent  
of  $k$ .

$$f_k(x) \quad \pi_k$$

$$= \arg \max_k \hat{\pi}_k e^{-\frac{(x - \hat{\mu}_k)^2}{2\hat{\sigma}^2}}$$

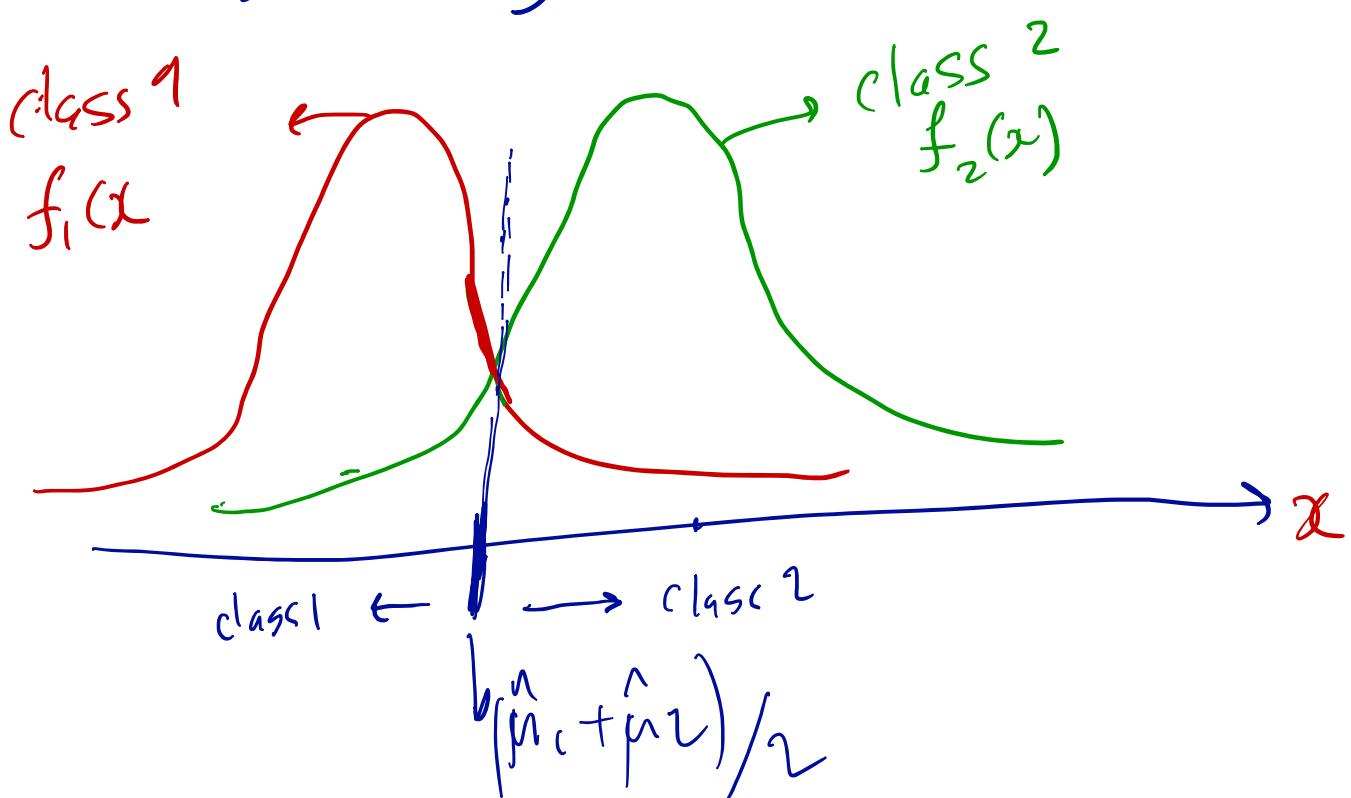
$$= \arg \max_k \left\{ \log \hat{\pi}_k - \frac{(x - \hat{\mu}_k)^2}{2\hat{\sigma}^2} \right\}$$

$$= \arg \max_k \left\{ \log \hat{\pi}_k - \cancel{\frac{x^2}{2\hat{\sigma}^2}} + \frac{x\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} \right\}$$

$$\hat{y}_{(x)} = \arg \max_k \left\{ \log \hat{\pi}_k + \frac{x \hat{\mu}_k}{\gamma^2} - \frac{\hat{\mu}_k^2}{2\gamma^2} \right\}$$

Example: Let's assume that we have two classes and also we have  $\pi_1 = \pi_2$ .

$$\pi_1 = \pi_2 \Rightarrow \log \pi_1 = \log \pi_2$$



$$\hat{y}(x) = \underset{k=1,2}{\operatorname{argmax}} \left\{ \log \hat{\pi}_k + \frac{x \hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2 \hat{\sigma}^2} \right\}$$

$$\frac{x \hat{\mu}_1}{\hat{\sigma}^2} - \frac{\hat{\mu}_1^2}{2 \hat{\sigma}^2} \begin{matrix} > \\[-1ex] < \end{matrix} \frac{x \hat{\mu}_2}{\hat{\sigma}^2} - \frac{\hat{\mu}_2^2}{2 \hat{\sigma}^2}$$

$$\frac{(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 + \hat{\mu}_1)}{2 \hat{\sigma}^2} \begin{matrix} > \\[-1ex] < \end{matrix} \frac{x(\hat{\mu}_2 - \hat{\mu}_1)}{\hat{\sigma}^2}$$

$$\frac{\hat{\mu}_2 + \hat{\mu}_1}{2} \begin{matrix} > \\[-1ex] < \end{matrix} x$$

# Multi-dimensional LDA:

$x \in (x_1, x_2, \dots, x_p) \in \mathbb{R}^n$

$$\pi_k = \frac{n_k}{n}$$

$$f_k(x) = N(\mu_k, \Sigma)$$

Multi-dimensional Gaussian PDF:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^p, \quad \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \in \mathbb{R}^p$$

$$f(x | \mu, \Sigma_{p \times p}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

$$\Sigma_{r,s} = E[(x_r - \mu_r)(x_s - \mu_s)]$$

For class  $k$ ,  $\Pr\{X=x \mid Y=k\}$  is modeled by  $f_k(x) = N(\mu_k, \Sigma)$

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \underbrace{|\Sigma|^{\frac{1}{2}}}_{\text{ }} \text{ }} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k)\right)$$

Parameters  $\mu_k$  and  $\Sigma$  will be learned from data.

$$\hat{\mu}_k = \frac{\sum_{i: y_i=k} x_i}{n_k}$$

$$\hat{\Sigma}_{ij} = \frac{1}{n-K} \sum_{k=1}^K \sum_{i: y_i=k} (x_{i,r} - \hat{\mu}_r)(x_{i,s} - \hat{\mu}_s)$$

$$\Sigma = \begin{pmatrix} & \vdots \\ \vdots & \Sigma_{rs} \end{pmatrix}_{p \times p}$$

$$\Sigma_{rs} = E[(x_r - \mu_r)(x_s - \mu_s)]$$

## Lecture 19:

$$Pr(y=k | x=x) \rightarrow$$

$$\frac{\overbrace{f_k(x)}^{\text{Pr}(x|k)} \cdot \overbrace{\pi_k}^{\text{Pr}(y=k)}}{\overbrace{\Pr(x=x)}^2}$$

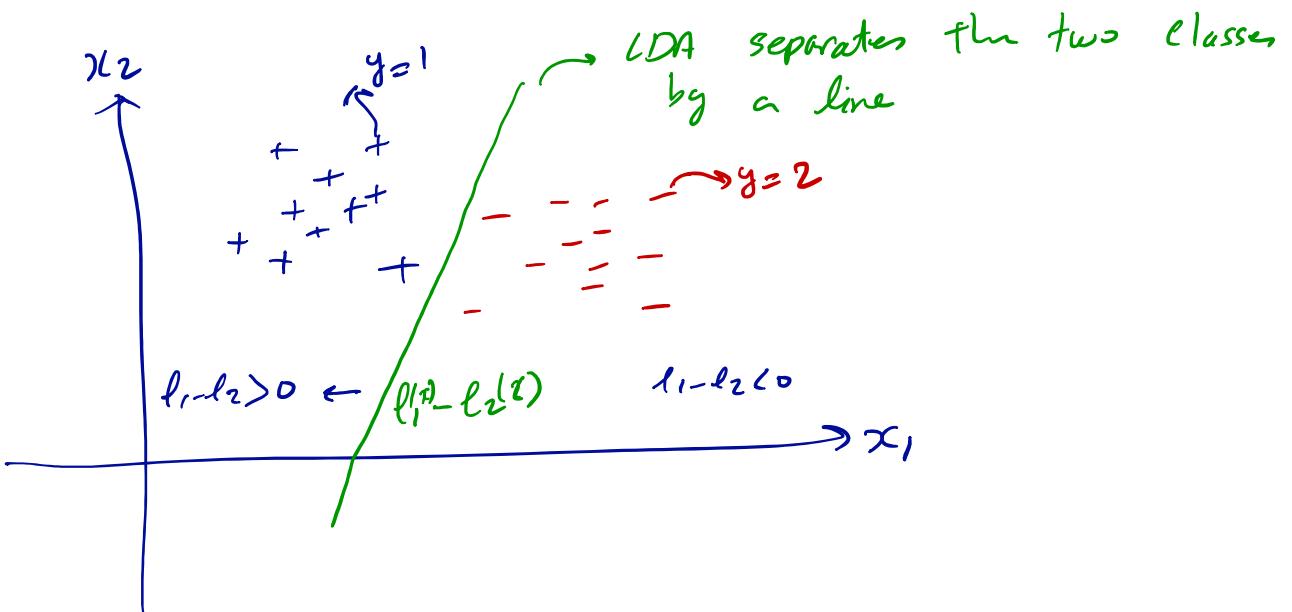
$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$$

## Multi-dim LDA

$\Sigma \rightarrow$  estimated from data

$\hat{\mu}_k \rightarrow$  estimated from data

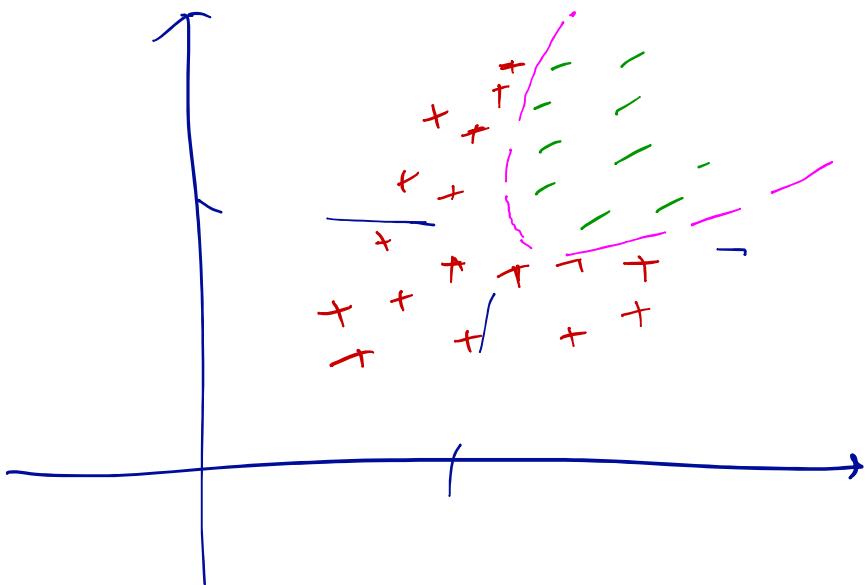
$$\begin{aligned} \hat{y}(x) &= \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} f_k(x) \underbrace{\pi_k}_{\text{gaussian PDF}} \\ &= \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} \left\{ x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k \right\} \end{aligned}$$



$$x = (x_1, x_2)$$

$\underset{k \in \{1, 2\}}{\operatorname{argmax}}$   $\left\{ x^T \underbrace{\sum_{k=1}^{n+1} \mu_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k}_{\ell_k(x)} \right\}$

$$\begin{cases} \ell_1(x) > \ell_2(x) \\ \ell_1(x) < \ell_2(x) \end{cases} \Rightarrow \ell_1(x) - \ell_2(x) \begin{cases} > 0 \\ < 0 \end{cases}$$



LDA could be restrictive because the classification boundary in LDA becomes a hyperplane.

Quadratic Discriminant Analysis(QDA):

QDA is similar to LDA except that the covariance matrix (the variance in the one-dimensional case) will depend on the class.

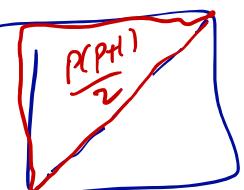
$$1-d: f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^T}{2\sigma_k^2}}$$

$$2-d: f_k(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

$\hat{\Sigma}_k$  and  $\hat{\mu}_k$  would be estimated from the data associated with class  $k$ .

$$\hat{y}_{\text{QDA}}(x) = \underset{k}{\operatorname{argmax}} \left\{ -\frac{1}{2} (x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k) - \frac{1}{2} \log |\hat{\Sigma}_k| + \log \hat{\pi}_k \right\}$$

↑  
Quadratic

LDA	vs	QDA
# of parameters: $K \cdot P + \frac{P(P+1)}{2}$	"	# parameters: $K \cdot P + \frac{K P(P+1)}{2}$
for each class: $\hat{\mu}_k \rightarrow P$ parameters	"	a lot more parameters
we also need to learn	"	to learn compared to LDA
$\Sigma \rightarrow \frac{P(P+1)}{2}$ parameters	"	↓ more complexity
$\Sigma =$ 	"	

LDA



less variance

less flexibility / complexity



If the common covariance assumption is off, then LDA can suffer from high bias (underfitting)

QDA

might overfit

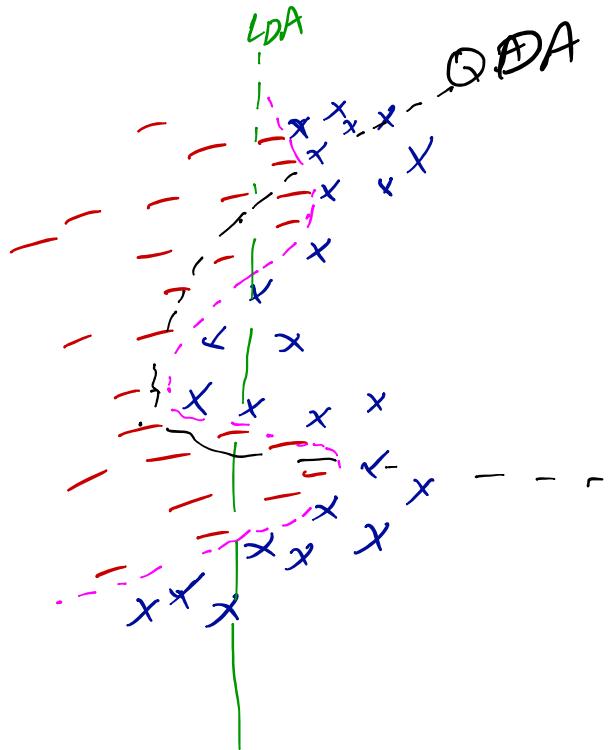
Roughly speaking: LDA is a better bet

than QDA if there are relatively few training data points.

In contrast, QDA is a better choice if the number of training data points is large or if the assumption of a common covariance matrix for the  $k$  classes is off.

$\cdot$   $T$ -nearest neighbor classifier:  
(a.k.a.  $kNN$  method)

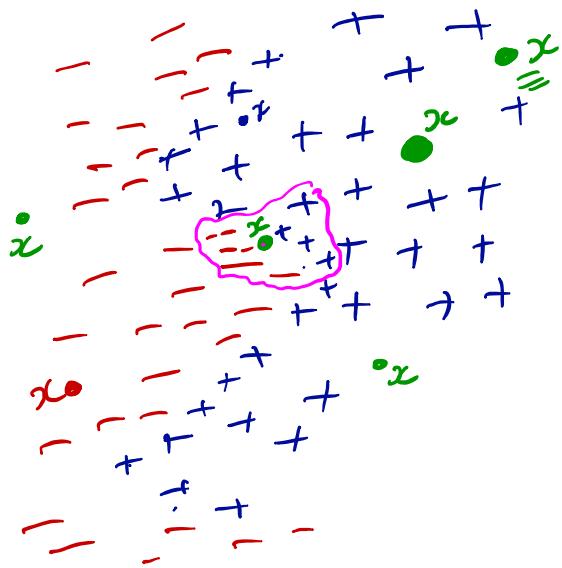
Non-parametric models:



estimate  $\Pr\{y=k \mid x=x\}$

Intuition: Estimate the probability of  
each class using labeled (training)  
data points which are close.

Let's consider the  $T$  closest training  
data points to the input  $x$ .



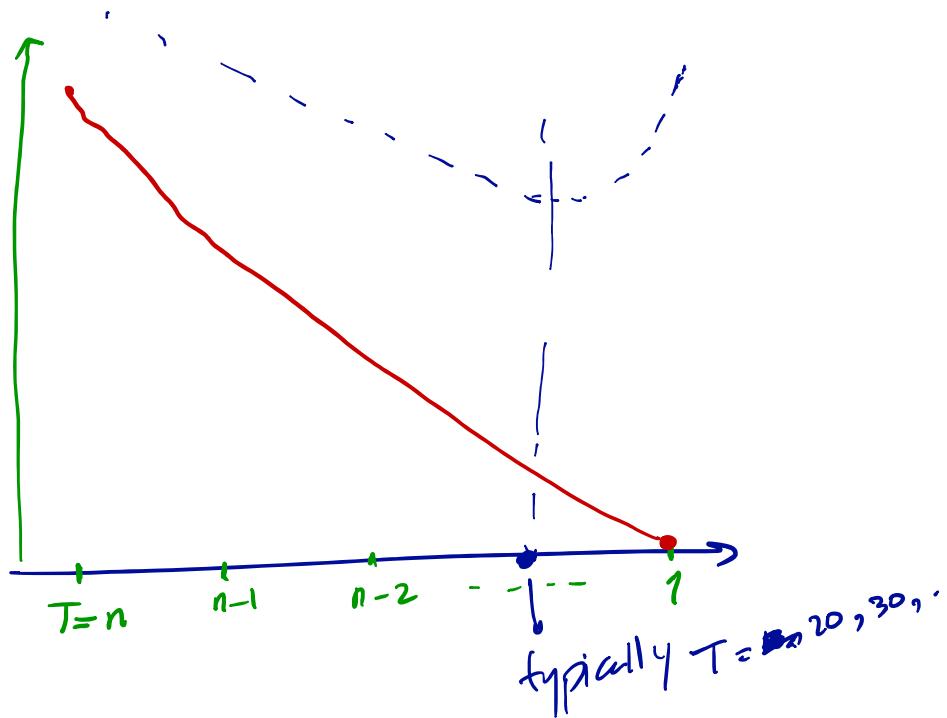
Let  $N_x$  be the set of  $T$  closest points wrt  $x$ .

$$\hat{P}(y_i=k \mid x=x) = \frac{\sum_{x_i \in N_x} \mathbb{1}\{y_i=k\}}{T}$$

$$\hat{y}_{TNN}(x) = \underset{k}{\operatorname{argmax}} \hat{P}(y_i=k \mid x=x)$$

↓  
majority

How can find the best  $T$ ?



---

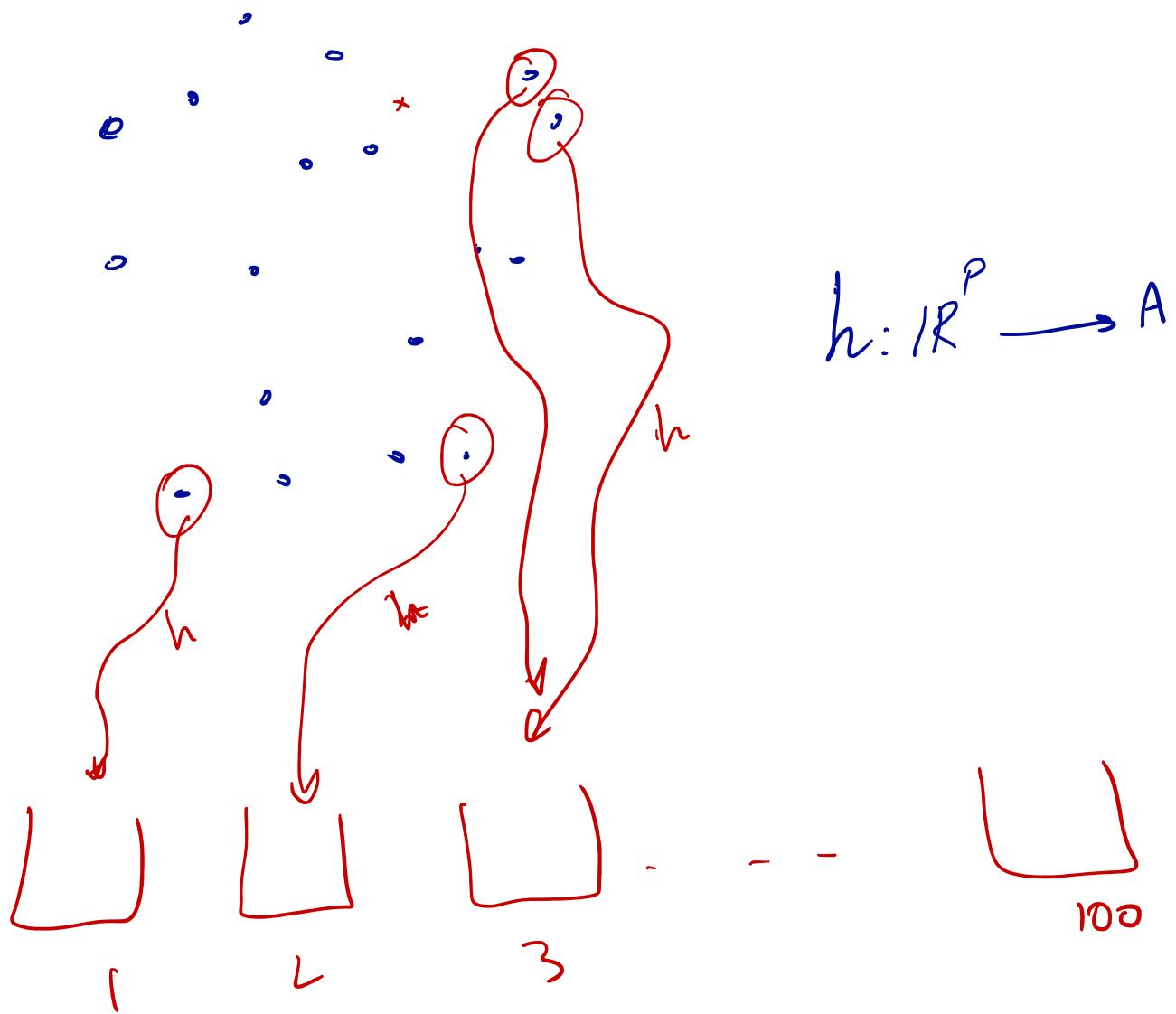
$T = 1 ? \rightarrow \text{train error} = 0$

$T = n ? \rightarrow \begin{aligned} \text{training error} &= \text{high} \\ \text{the prediction} &\text{ is the} \\ \text{same for any } x. \end{aligned}$

---

In high-dimensional problems we have "the curse of dimensionality" meaning that the (Euclidean) distance  $\rightarrow$  is not so informative.

# Locality sensitive Hashing:



$$x \rightarrow h(x) = 3$$

## Lecture 2D:

### Module 4 : Unsupervised learning.

Supervised learning:

$$(x_1, y_1), \dots, (x_n, y_n)$$

↳ Goal: learn a predictive relation  $f: X \rightarrow Y$  and use it for prediction.

Unsupervised learning;

There is no label.

Data:  $x_1, x_2, \dots, x_n$

Goal: Discover informative patterns  
structures, subgroups, etc within data

In this course, we'll consider two specific instances of unsupervised learning: clustering and dimensionality reduction.

---

Clustering: partition data into distinct sub-groups such that data points within each group are similar to each other and points from different groups are "dissimilar" to each other.

similar? dissimilar?

Similarity will be quantified  
by a distance function:

$$d(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$$

$d(x, y)$

$x = (x_1, \dots, x_p) \in \mathbb{R}^p$        $y = (y_1, \dots, y_p) \in \mathbb{R}^p$

e.g.  $d(x, y) = \left( \sum_{i=1}^p (x_i - y_i)^2 \right)^{1/2}$

- Euclidean distance.      Also denoted by  $\|x - y\|_2$

- L<sub>2</sub> distance.

$$d(x, y) = \sum_{i=1}^P |x_i - y_i|$$

↳ - Manhattan Distance

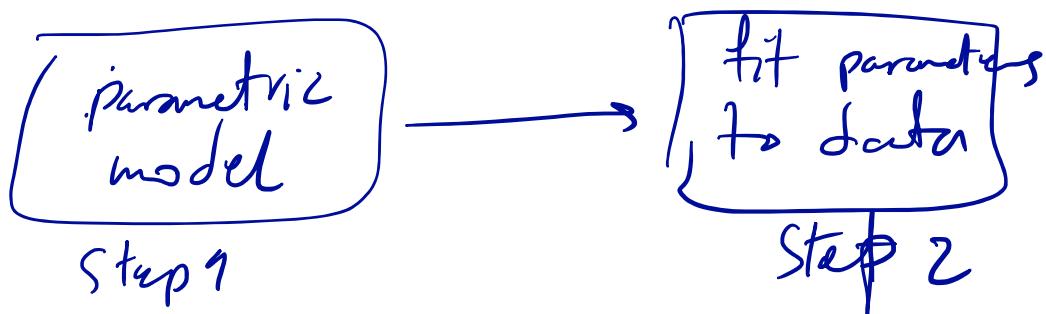
-  $L_1$  - distance

⋮

many other choices

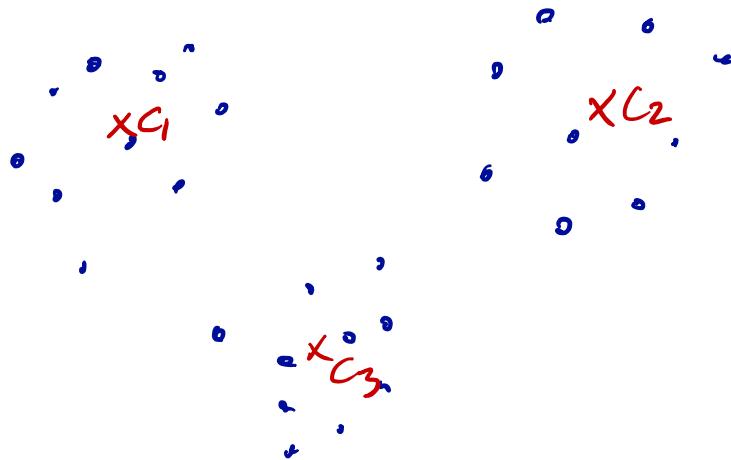


Clustering  $\rightarrow$  Goal: partition data into  $k$  groups ( $k$ : given) such that points in each group are similar and points in different groups are dissimilar.



Step 1:

$$K=3$$



step 1 (parametric modeling):

How do we represent a group?

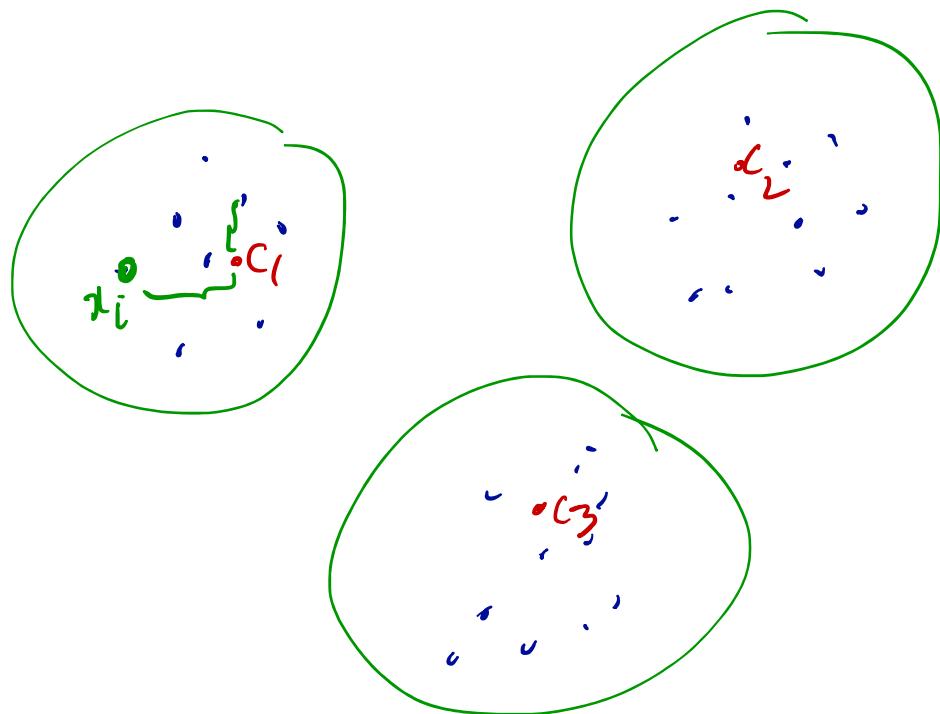
- let's assume we'll have  $k$  groups ( $k$  given).
- assign to each group  $i$  a "center",  $c_i \in \mathbb{R}^P$ , such that points in group  $i$  are close to  $c_i$  and points outside of group  $i$  are not close to  $c_i$ .

Parameters of the clustering problem:  $c_1, c_2, \dots, c_k \in \mathbb{R}^p$

Step 2 : Design an objective/loss function wrt  $c_1, \dots, c_k \in \mathbb{R}^p$ . (\*)

$$\min_{c_1, \dots, c_k} \sum_{j=1}^k \sum_{i \in \text{Group } j} \|x_i - c_j\|_2^2$$

Internal distance of group j



## $K$ -means clustering:

Goal: Solve  $(*)$  and find  $c_1, \dots, c_K$ .

Difficulty: Finding the exact solution of  $(*)$  is very difficult (most probably it is intractable meaning that it needs exponential complexity in terms of  $p$ ). Instead, we'll come up with algorithms that work well in practice.

Our algorithm (i.e. the K-Means algorithm) is based on two main principles:

principle 1:

$$\sum_{j=1}^k \sum_{x_i \in \text{Group}_j} \|x_i - c_j\|^2$$
$$= \sum_{i=1}^n \|x_i - c(x_i)\|_2^2$$

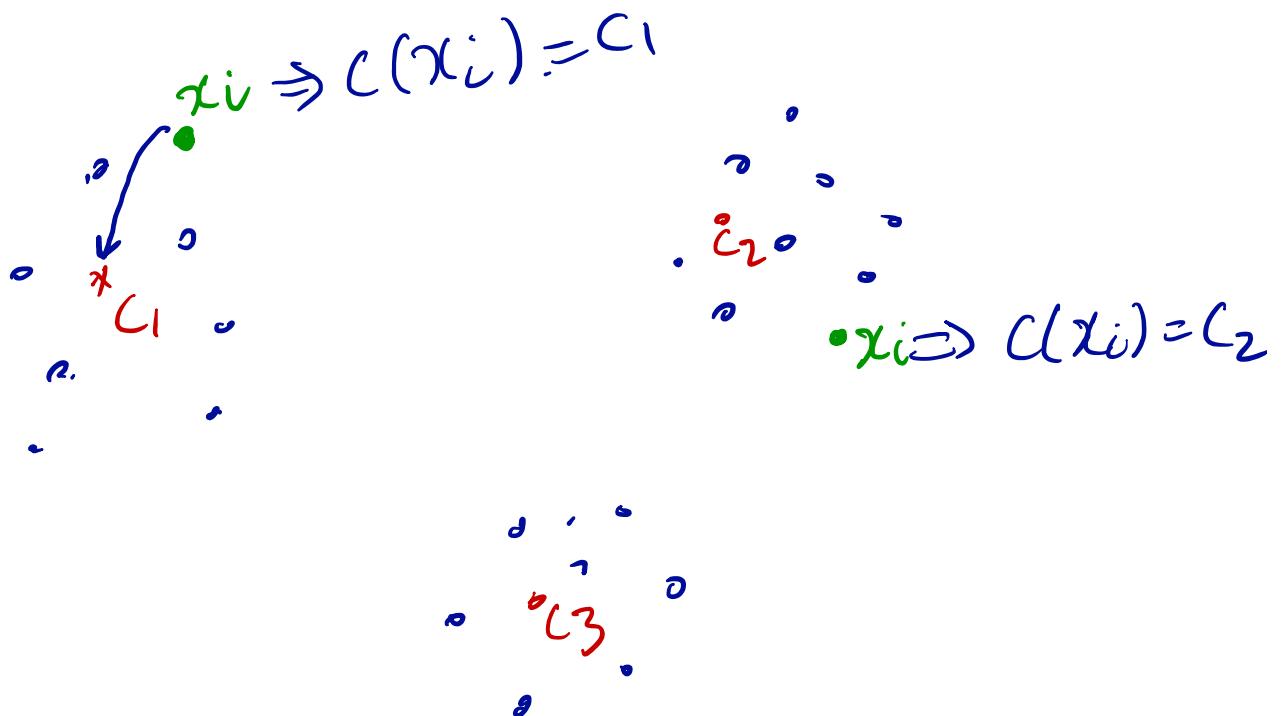
( $c(x_i)$  is the center of the group that  $x_i$  belongs to).

principle 9: Given centers  $c_1 \rightarrow c_k$

the center (group) that we assign to data point  $x_i$  is the one that is closest to

$x_i$ :

$$c(x_i) = \underset{c_j \in \{c_1, \dots, c_k\}}{\operatorname{argmin}} \|x_i - c_j\|.$$



Principle 2:

Let's assume that we fix the groups:

Group<sup>1</sup>, --> Group<sub>K</sub>. Then

$$c_j = \frac{\sum_{x_i \in \text{Group } j} x_i}{N_j \rightarrow \# \text{ points in Group } j}$$

---

Let's try to solve the 1-Means problem.

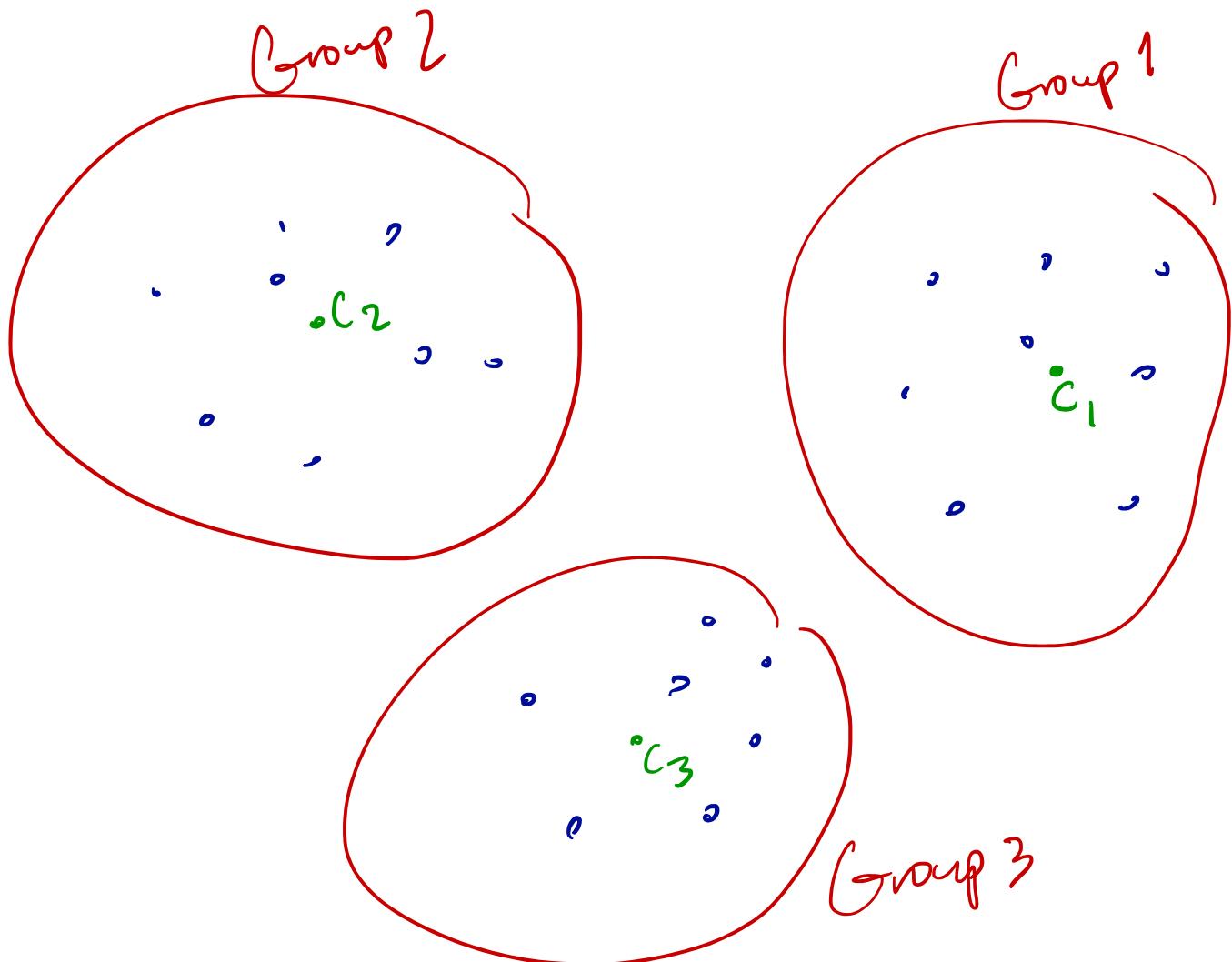
$$\min_{c_1 \in \mathbb{R}^P} \left[ \sum_{i=1}^n \|x_i - c_1\|^2 \right]$$

$$c_1 = \frac{\sum_{i=1}^n x_i}{n}$$

If Group j is given  
then:

$$c_j = \frac{\sum_{x_i \in \text{Group}_j} x_i}{N_j}$$

$$\min_{c_1, \dots, c_K} \sum_{j=1}^K \sum_{\substack{x_i \in \text{Group}_j}} \|x_i - c_j\|^2$$



## $k$ -means algorithm :

- choose  $\{c_1^0, \dots, c_k^0\}$  according to an arbitrary choice
- For  $t=1, 2, \dots$  :
  - (1) Define  $\text{Group}_j$  to be the set of all the data points whose closest center is  $c_j^{t-1}$ .

(2) update  $c_j^t = \frac{\sum_{x_i \in \text{Group}_j} x_i}{N_j}$

(3) If  $c_j^t = c_j^{t-1}$  for all  $j \in \{1, \dots, k\}$ , then stop!

Let

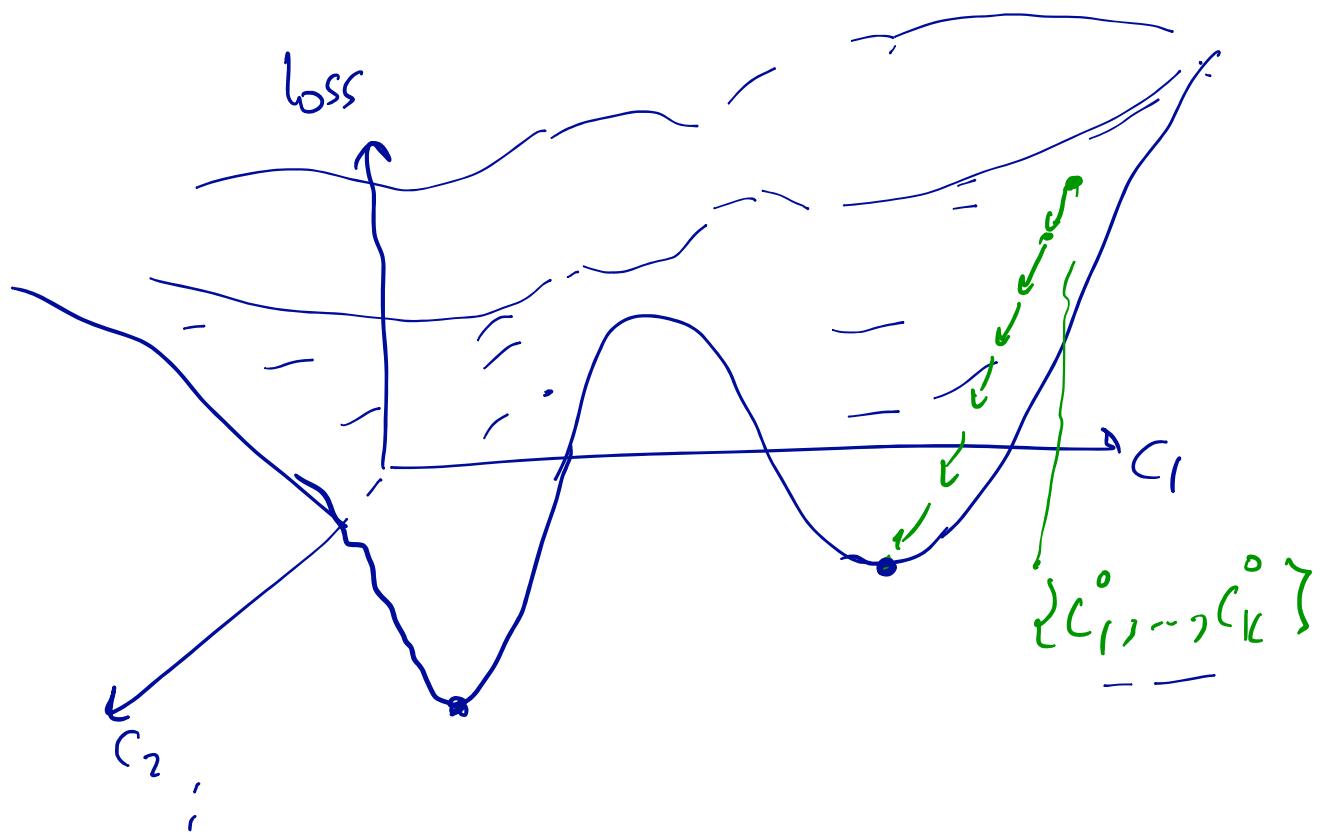
$$L(\{c_1, \dots, c_k\}) = \sum_{i=1}^n \|x_i - c(x_i)\|_2^2$$

-Then it can be proven that

$$L(\{c_1^t, \dots, c_k^t\}) \leq L(\{c_1^{t-1}, \dots, c_k^{t-1}\})$$

- Also, it can be shown that after a "finite" number of iterations the K-Means algorithm will stop (converge).

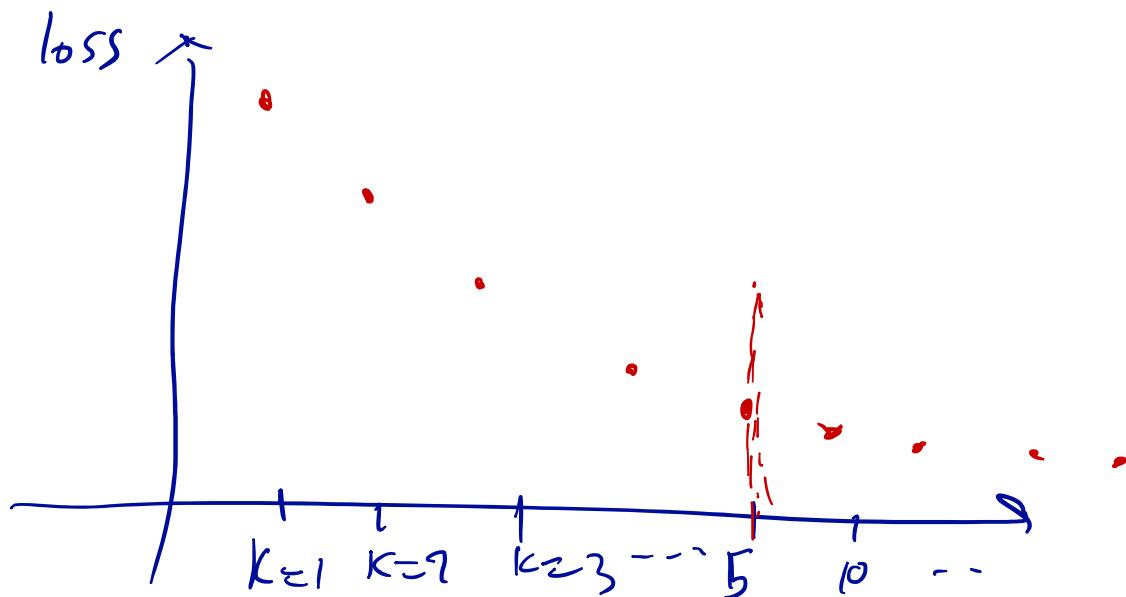
# Some important practical points about K-means:



The quality of the final set of centers that we obtain from the K-means alg. depends on the initial centers  $\{c_1^o, \dots, c_K^o\}$ .

In practice, we often choose ~ 10-20-50 different initializations and choose the set of centers that has the best loss.

- So far we've assumed that  $K$  is given.



## Lecture 21:

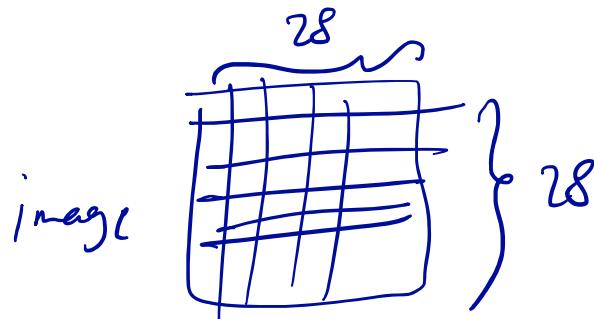
Unsupervised learning

clustering  
Dimensionality Reduction  
(this lecture)

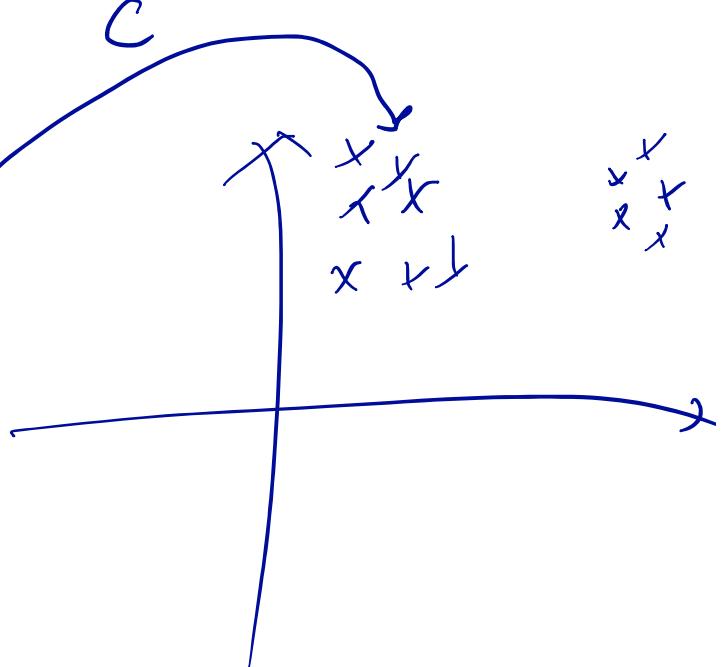
→ data visualization in  $\mathbb{R}^2$

Example: Consider a data set of images: e.g. 10000 data points

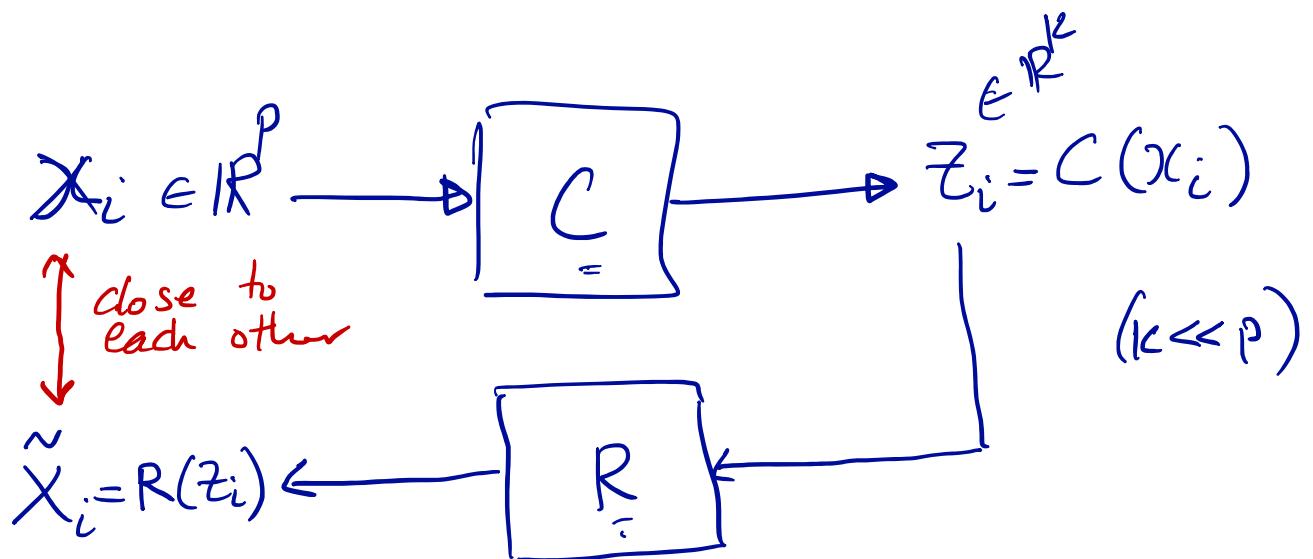
$$x_1, \dots, x_n \in \mathbb{R}^{724}$$



$$x_1, \dots, x_n$$



$x_1, \dots, x_n \in \mathbb{R}^P$ ; want to represent the data points in some lower-dimensional space  $\mathbb{R}^K$  (e.g.  $K=2$ )



$$\text{minimize}_{C, R} \sum_{i=1}^n \|x_i - \tilde{x}_i\|_2^2$$

$$\tilde{x}_i = R(C(x_i))$$

$$\min_{R, C} \sum_{i=1}^n \|x_i - R(C(x_i))\|_2^2$$

↓ Simplify to the simplest class of mappings

Linear mappings:  $C = \text{matrix} \in \mathbb{R}^{K \times P} \rightarrow \begin{bmatrix} \cdot & \cdot & \cdot \\ \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdot \end{bmatrix}_{K \times P}$

$R = \text{matrix} \in \mathbb{R}^{P \times K}$

$$z_i = C(x_i) = \underset{\substack{\downarrow \\ \text{matrix}}}{C_{K \times P}} \cdot x_i$$

$$\tilde{x}_i = R(z_i) = \underset{\substack{\uparrow \\ \text{matrix}}}{R_{P \times K}} \cdot z_i$$

The final version (after considering linear mappings):

$$\min_{\substack{C \in \mathbb{R}^{K \times P} \\ R \in \mathbb{R}^{P \times K}}} \sum \|x_i - R \cdot C \cdot x_i\|_2^2$$

This problem is called  
the Principal Component  
Analysis (PCA)

Goal: Given Data set  $x_1, \dots, x_n \in \mathbb{R}^p$   
and the target dimension  $k$   
(e.g.  $k=2$ ) we'd like to minimize  
the objective of PCA to  
find the best linear maps  
 $C, R$ .

$$C = [ \quad ]_{2 \times p}$$

$$R = [ \quad ]_{p \times 2}$$

To solve this optimization problem,  
we need to review an important  
tool in linear algebra

(which is also very important for data science) called the Singular Value Decomposition (SVD).

Singular Value Decomposition (SVD)  
of Matrices:

Theorem (SVD): Let  $A$  be a  $P \times n$  matrix ( $P \leq n$ ). Then we can always write:

$$A = U_{P \times P} \cdot S_{P \times n} \cdot V_{n \times n}^T$$

where  $U$  and  $V$  are unitary matrices

$$U V^T = I_{P \times P}, \quad V V^T = I_{n \times n}$$

and  $S$  is a diagonal matrix:

$$S = \begin{bmatrix} s_{11} & & & \\ & s_{22} & & \\ & & \ddots & \\ & 0 & & s_{pp} \end{bmatrix}_{p \times n}$$

$$\left\{ \begin{array}{l} s_{ii} = s_i \geq 0 \\ s_{ij} = 0 \end{array} \right.$$

We further have  $s_1 \geq s_2 \geq \dots \geq s_p$

$$p=2, n=4$$

$$S = \begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \end{bmatrix}_{2 \times 4}$$

$$A = \begin{bmatrix} & \\ & \\ & \end{bmatrix}_{pxn}$$

$$A = U_{pxp} \cdot S_{pxn} \cdot V_{n \times n}^T$$

$$\begin{bmatrix} & \\ & \\ & \end{bmatrix}_{pxn}$$

$$= \left[ \begin{array}{c} U \\ \downarrow \\ UU^T = I \end{array} \right]_P \times \left[ \begin{array}{ccc} b_1 & \geq & 0 \\ b_2 & \geq & \ddots \\ 0 & \ddots & b_p \geq 0 \end{array} \right]_S \times \left[ \begin{array}{c} & \\ & \\ & \end{array} \right]_{pxn}$$

$$x \begin{bmatrix} & \\ & \\ & \end{bmatrix}_x \quad V^T \downarrow \quad \begin{array}{c} V \\ \downarrow \\ VV^T = I \end{array} \quad \begin{bmatrix} & \\ & \\ & \end{bmatrix}_{n \times n}$$

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}_{2 \times 2}$$

$$P=n=2$$

$$A = U \cdot S \cdot V^T$$

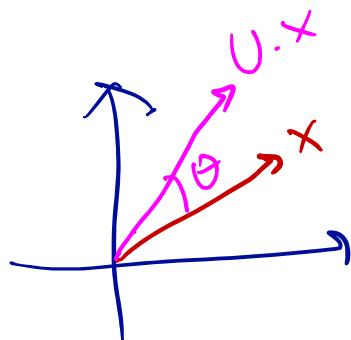
$$= \begin{bmatrix} & \\ & U \end{bmatrix} \cdot \begin{bmatrix} & 0 \\ 0 & \sigma \end{bmatrix} \cdot \begin{bmatrix} & \\ & V^T \end{bmatrix}$$

$$U U^T = I_{2 \times 2} \rightarrow U = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$$

where  $\theta \in [0, 2\pi]$

rotation matrix

$$x \rightarrow U \cdot x$$



$$V V^T = I \implies V = \begin{pmatrix} \cos \theta' & \sin \theta' \\ -\sin \theta' & \cos \theta' \end{pmatrix}$$

↓  
also a rotation

$$S = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \rightarrow \text{Scaling}$$


---

any matrix A

$$= \text{rotation} \times \text{scaling} \times \text{rotation}$$


---

every linear mapping = Rotations scaling o rotation

Let's try to perform singular value decomposition on the data:

Let's the "Data Matrix" which is defined as;

$$X = \begin{pmatrix} x_1 & | & x_2 & | & \dots & | & x_n \\ \in \mathbb{R}^P & & \in \mathbb{R}^P & & & & \in \mathbb{R}^P \end{pmatrix}_{P \times n}$$

$$X \xrightarrow{\text{SVD}} U_{P \times P} \cdot S_{P \times n} \cdot V_{n \times n}^T$$

$$- U = U_{P \times P} = \left( \begin{array}{c|ccc|c} U_1 & U_2 & \cdots & & U_P \\ \hline U_1^T & U_2^T & \cdots & & U_P^T \end{array} \right)$$

$$- UU^T = I_p \Rightarrow \|U_i\|_2^2 = 1 \quad \text{and}$$

$$\langle U_i, U_j \rangle = 0 \quad i \neq j$$

$\{U_1, \dots, U_p\}$  is an orthonormal set.

- Since  $\{U_1, \dots, U_p\}$  is an orthonormal set, we can always write:

$$x_i = \sum_{j=1}^p \underbrace{\langle x_i, U_j \rangle}_{\text{inner product}} U_j$$

$$\|x_i\|^2 = \sum_{j=1}^p \langle x_i, U_j \rangle^2$$

Let's define the energy of the data as

Energy of data =

$$\sum_{i=1}^n \|x_i\|_2^2$$

Theorem:

$$\sum_{i=1}^n \|x_i\|_2^2 = \sum_{j=1}^p b_j^2$$

proof (for completeness):

$$\begin{aligned}\sum_{i=1}^n \|x_i\|_2^2 &= \text{trace}(X^T X) \\ &= \text{trace}(V S^T U \underbrace{U^T S V^T}_I) \\ &= \text{trace}(V S^T S V^T)\end{aligned}$$

$$= \text{trace}(S^T S) = \sum_{j=1}^p b_j^2.$$

$$S^T S = \begin{bmatrix} b_1^2 & & \\ & b_2^2 & \\ & & \ddots & \\ & & & b_p^2 \end{bmatrix}_{p \times p}$$

$$\boxed{\sum_{i=1}^n \|x_i\|^2 = \sum_{j=1}^p b_j^2}$$

Recall that :

$$x_i = \sum_{j=1}^p \langle x_i, u_j \rangle u_j$$

$$\sum_{i=1}^n \|x_i\|_2^2 = \sum_{j=1}^p \underbrace{\sum_{i=1}^n \langle x_i, u_j \rangle^2}_{b_j^2}$$

$$\forall j \rightarrow b_j^2 = \sum_{i=1}^n \langle x_i, u_j \rangle^2$$

Two important results:

Energy of Data set

$$\sum_{i=1}^n \|x_i\|^2 = \sum_{j=1}^P \delta_j^2$$

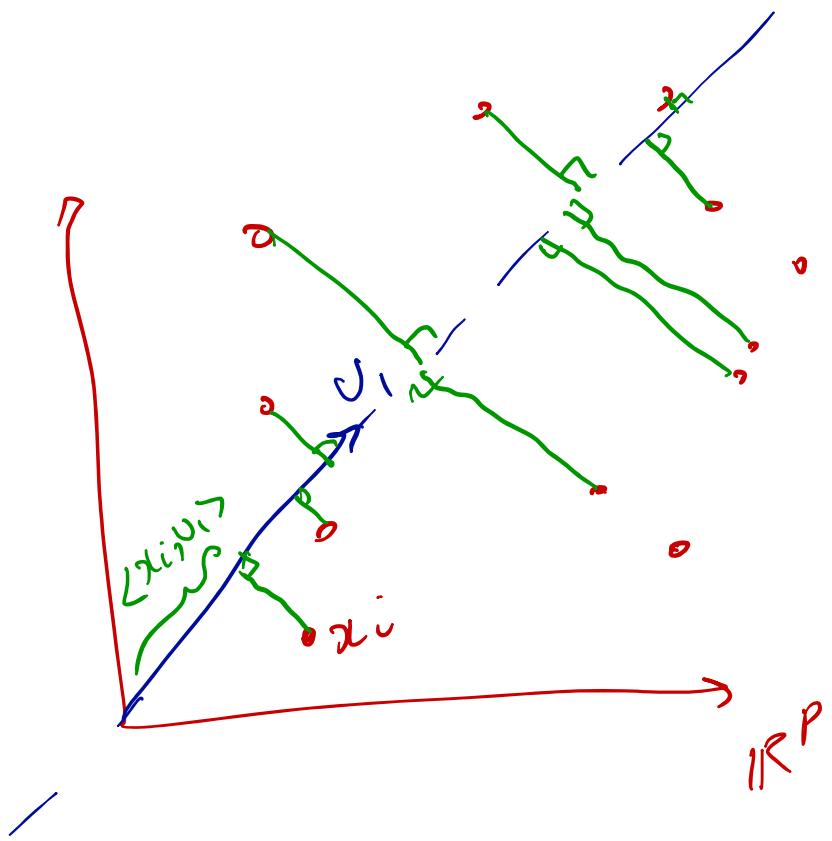
$$\delta_j^2 = \sum_{i=1}^n \langle x_i, U_j \rangle^2$$

the energy of data along  
the direction  $U_j$

---

$$U = (U_1 | U_2 | \dots | U_P)$$

$\longrightarrow U_1$



$$\sum_{i=1}^n \langle x_i, U_1 \rangle^2 = \sigma_1^2$$

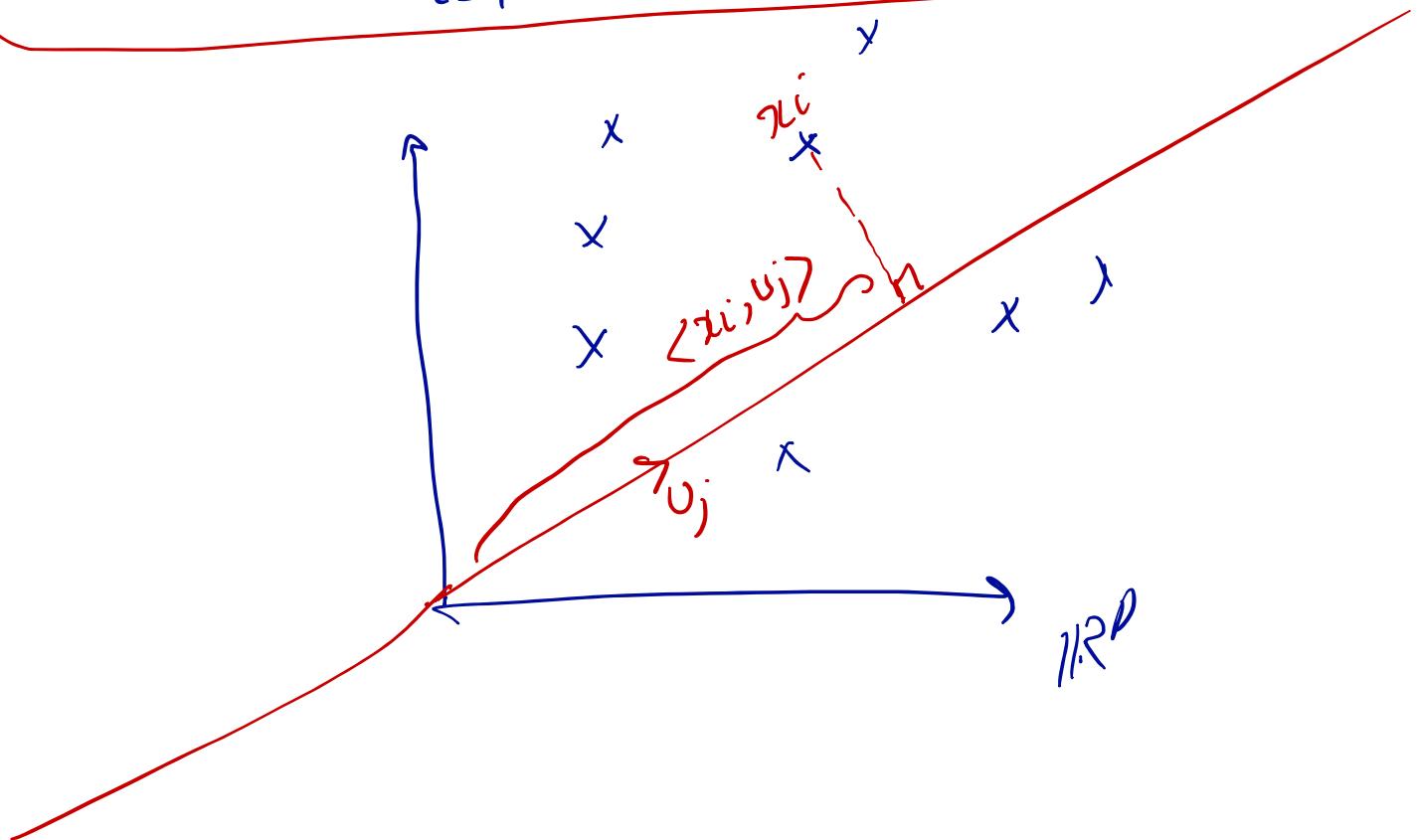
$$\sum_{i=1}^n \langle x_i, U_j \rangle^2 = \sigma_j^2$$

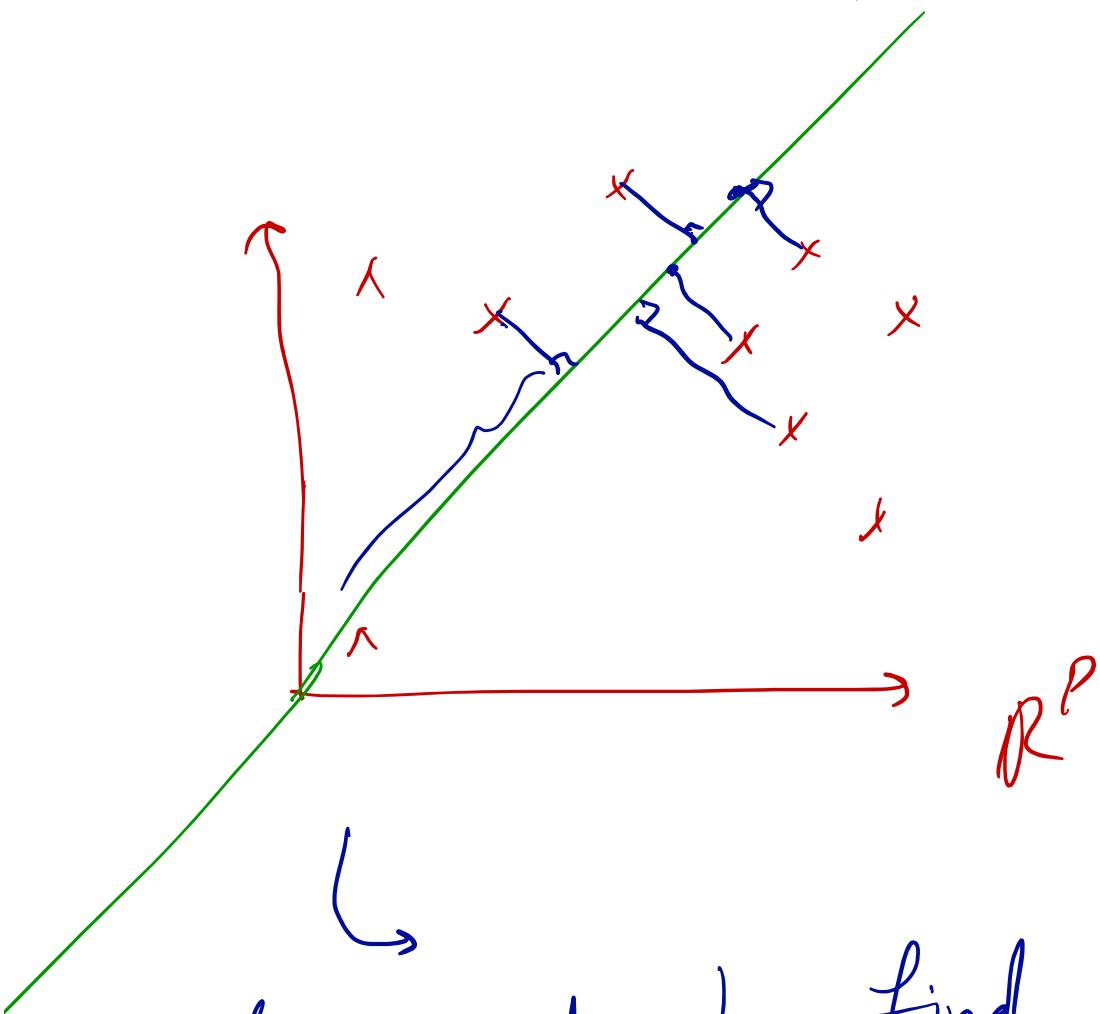
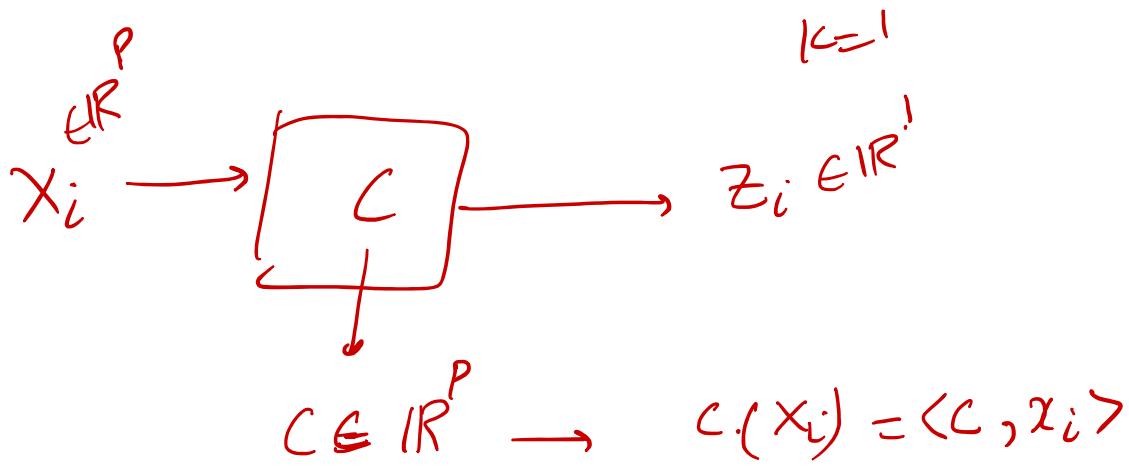
Energy of data:

$$\sum_{i=1}^n \|x_i\|^2$$

$$= \sum_{j=1}^p b_j^2$$

$$b_j^2 = \sum_{i=1}^n \langle x_i, u_j \rangle^2$$





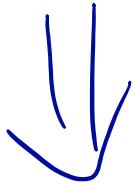
if we want to find a direction  
 such that the energy of  
 the data along that direction  
 is maximized then that

direction is  $U_1$ .

$U_1$  is called the principal singular vector.

$$C : \mathbb{R}^P \longrightarrow \mathbb{R}^I$$

$$\min_{C, R} \sum_{i=1}^n \|x_i - R \cdot C \cdot x_i\|^2$$



the minimizer is  $C = U_1$   
and  $R$  would be  $U_1^T$ .

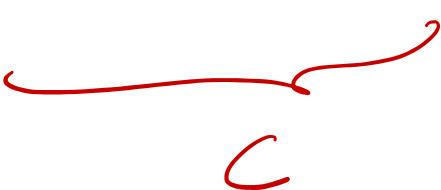
1-dim representation of data:  
1-dim representation of data:

$$z_i = \langle U_1, x_i \rangle$$

$$\tilde{x}_i = \langle U_1, x_i \rangle \cdot U_1^T$$

When  $K=2$  :

$$C(x_i) = [\langle U_1, x_i \rangle, \langle U_2, x_i \rangle]$$
$$= \begin{bmatrix} -\frac{U_1^T}{U_2^T} \end{bmatrix} \cdot x_i$$



The reconstruction matrix

$$R = [U_1, U_2]$$

$$\tilde{x}_i = R \cdot C \cdot x_i = \langle U_1, x_i \rangle U_1 + \langle U_2, x_i \rangle U_2$$

for general  $k$ :

$$C(x_i) = (\langle U_1, x_i \rangle, \langle U_2, x_i \rangle, \dots, \langle U_k, x_i \rangle)$$

$$\tilde{x}_i = R \cdot C(x_i)$$

$$= \sum_{j=1}^k \langle U_j, x_i \rangle U_j .$$

---

$$U = (U_1 | \dots | U_p) \rightarrow \text{orthonormal}$$

the best 2-d representation of the data

$$x_i = \underbrace{\langle x_i, U_1 \rangle}_{\text{best 1-d representation of the data}} U_1 + \langle x_i, U_2 \rangle U_2 + \dots + \langle x_i, U_p \rangle U_p$$

$$c: \mathbb{R}^P \rightarrow \mathbb{R}^K$$

$$\begin{aligned} c(x_i) &= \begin{pmatrix} \langle u_1, x_i \rangle \\ \langle u_2, x_i \rangle \\ \vdots \\ \langle u_K, x_i \rangle \end{pmatrix} \\ &= \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_K^T \end{pmatrix} \times x_i \end{aligned}$$

$c$

# Module 5: An Introduction to the Statistical Learning Theory

We will consider the supervised learning problem, but the framework can also be extended to unsupervised learning.

A formal framework for learning theory:

(1) Data is assumed to be generated according to a distribution  $(x, y) \sim D(x, y)$ . The input domain is denoted by  $X$  ( $x \in X$ ), the label domain is denoted  $\mathcal{Y}$  (i.e.  $y \in \mathcal{Y}$ ), and  $D$  is a distribution over  $X \times \mathcal{Y}$ .

(2) The learner's output: we are

looking for predictive relations

$h: X \rightarrow Y$  with low prediction

error. Formally speaking, for each function  $h: X \rightarrow Y$  we

define its error as:

$$\begin{aligned} L(h) &= \Pr_{(x,y) \sim D} \{ h(x) \neq y \} \\ &= E_{(x,y) \sim D} [ \mathbb{I}_{\{h(x) \neq y\}} ] \end{aligned}$$

prediction error  
generalization error  
true error  
Distribution of data

Ideally, we'd like to find a predictor  $h$  that has the

Smallest true error, i.e. our gold standard is to solve

$$\underset{h}{\text{minimize}} \ L_D(h) \quad (1)$$

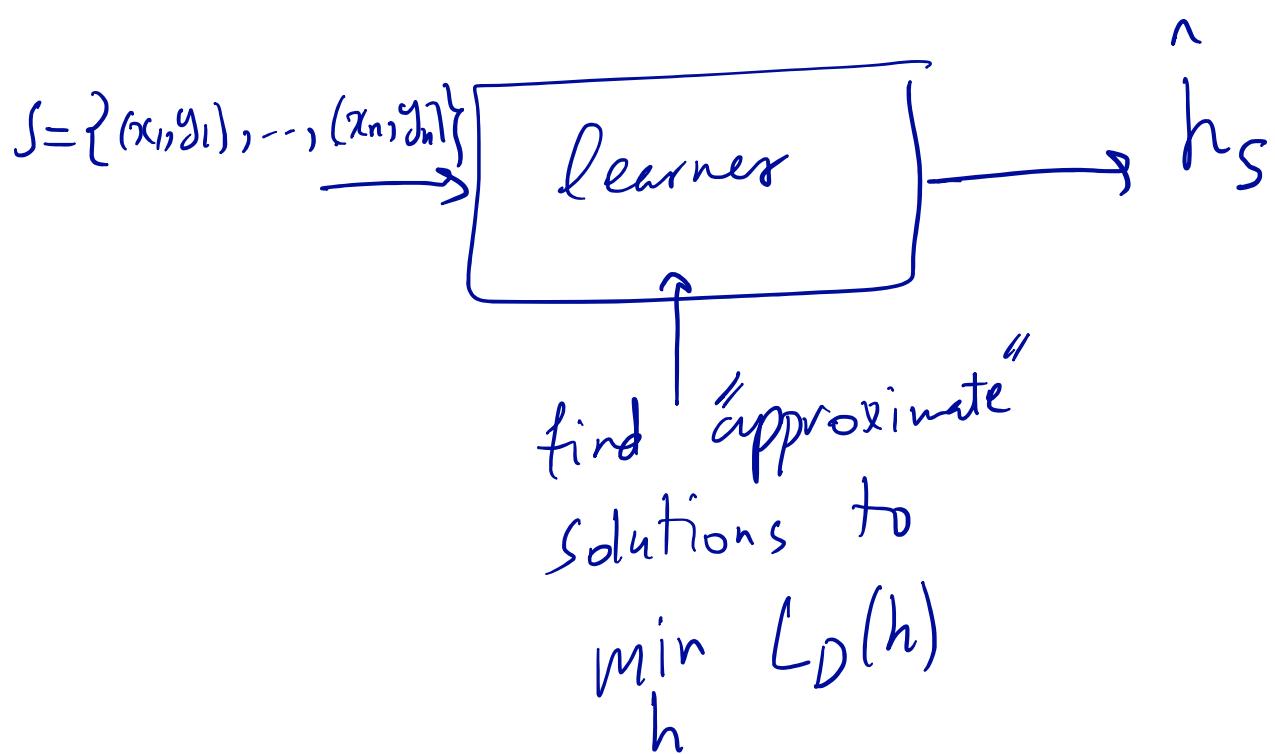
$$L_D(h) = E_{(x,y) \sim D} [\# \{ h(x) \neq y \}]$$

However, this task is impossible

since  $D$  is unknown.

(3) the learner's input: Training data! The only information that the learner has about the data distribution is a set of training samples (data points),  $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$  where each  $(x_i, y_i)$  is

Generated i.i.d. according to the distribution  $D$ . Hence, the learner has to find "approximate" solutions to problem (1) using the training data set.



$$(x_i, y_i) \stackrel{iid}{\sim} D$$

#### 4) Empirical Risk Minimization (ERM):

$$\underline{L_D(h)} \longrightarrow \begin{matrix} \text{find an} \\ \text{unbiased} \\ \text{estimator} \end{matrix}$$

↑  
Can't compute  
exactly

$$L_D(h) = E_{(x,y) \sim D} \left[ \mathbb{1}_{\{h(x) \neq y\}} \right]$$

unbiased estimate

$$L_S(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{h(x_i) \neq y_i\}}$$

---

$$\text{Note: } E_S [L_S(h)] = L_D(h)$$

Empirical risk minimization (ERM) is the task of finding a predictor that minimizes the training error:

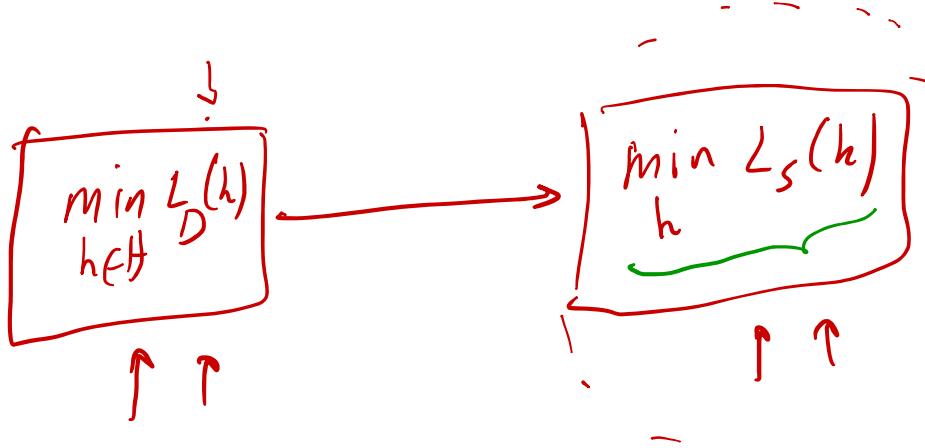
$$\min_h L_S(h) \quad (2)$$

$\min_h L_D(h) \xrightarrow{\text{instead}} \min_h L_S(h)$ 

$L_D(h)$  we can't do

$$\begin{aligned}
 L_D(h) &\xrightarrow{\text{Instead}} L_S(h) \\
 &= \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{h(x_i) \neq y_i\}
 \end{aligned}$$

## Lecture 23:



Claim:

$$\min_h L_S(h) = 0$$

$$L_S(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{h(x_i) \neq y_i\}}$$

Define  $\tilde{h} : X \rightarrow Y$  as follows:

$$\tilde{h}(x) : \begin{cases} y_i & \text{if } x = x_i \\ 0 & \text{if } x \in \{x_1, \dots, x_n\} \end{cases}$$

It's easy to see that

$$L_S(\tilde{h}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\tilde{h}(x_i) \neq y_i\} = 0$$

---

$$L_S(\tilde{h}) = 0 \rightarrow$$

$$\min_h L_S(h) = 0$$

5) Although ERM is natural,  
it can fail miserably if  
we are not careful: It  
can "overfit" easily.

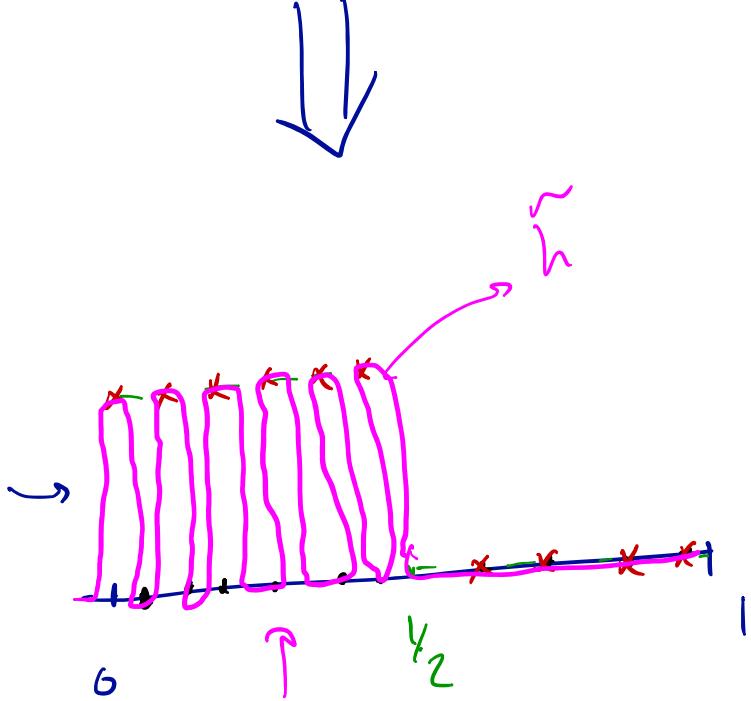
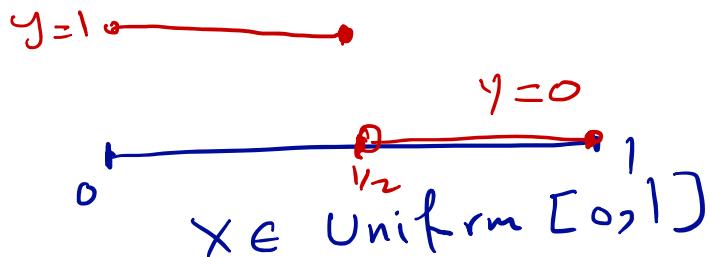
The minimizer of

$$\min_h L_S(h) \quad (2)$$

is always zero.

Example:

Distribution of data:



Prediction  
on unseen  
data here  
is very bad

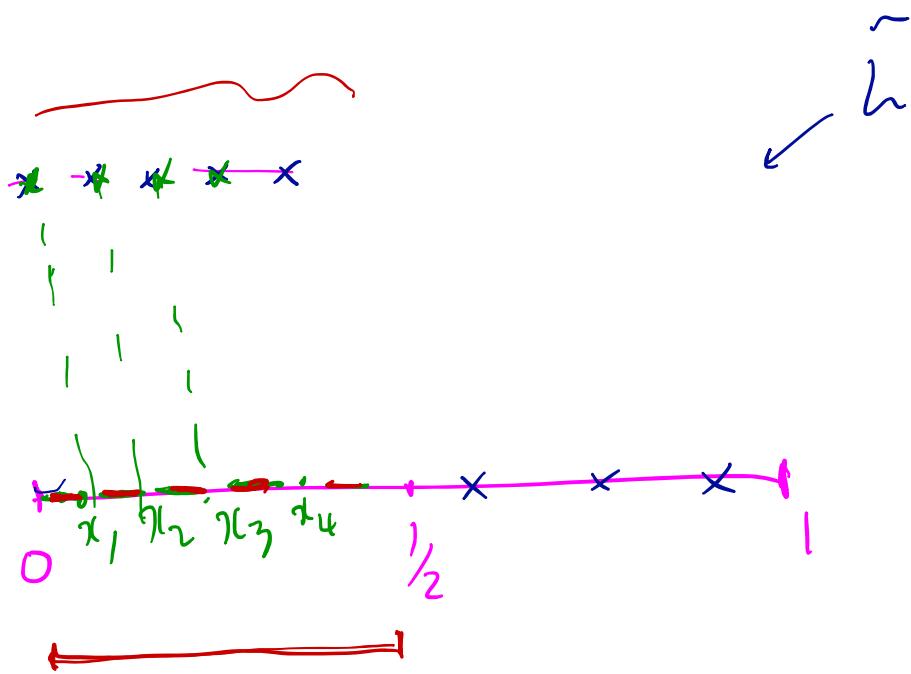
$$\mathcal{L}_S(\tilde{h}) = 0$$

$$\left\{ \mathcal{L}_D(\tilde{h}) \approx 0.5 \right.$$

---

$$\min_h \mathcal{L}_D(h)$$

$$\min_h \mathcal{L}_S(h) \hat{f} \tilde{h}$$

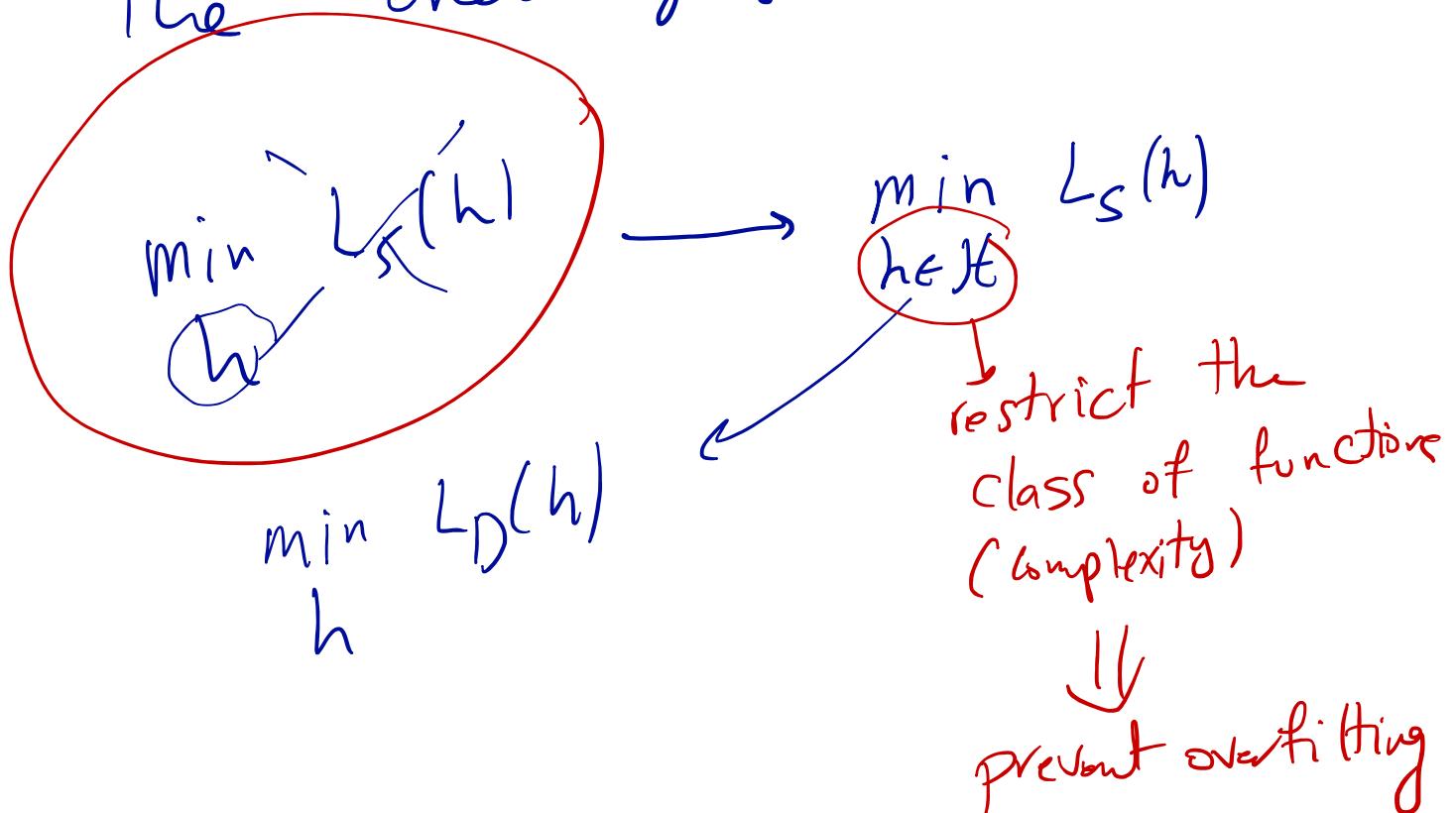


$$\tilde{h} = \begin{cases} h(x_i) = y_i & x \in \{x_1, \dots, x_n\} \\ h(x) = 0 & \text{otherwise} \end{cases}$$

$$\left\{ \begin{array}{l} L_S(\tilde{h}) = 0 \\ \neg D^{(n)} = \bigcup_{(x,y) \in D} [h(x) \neq y] \end{array} \right.$$

$$= \frac{1}{2}$$

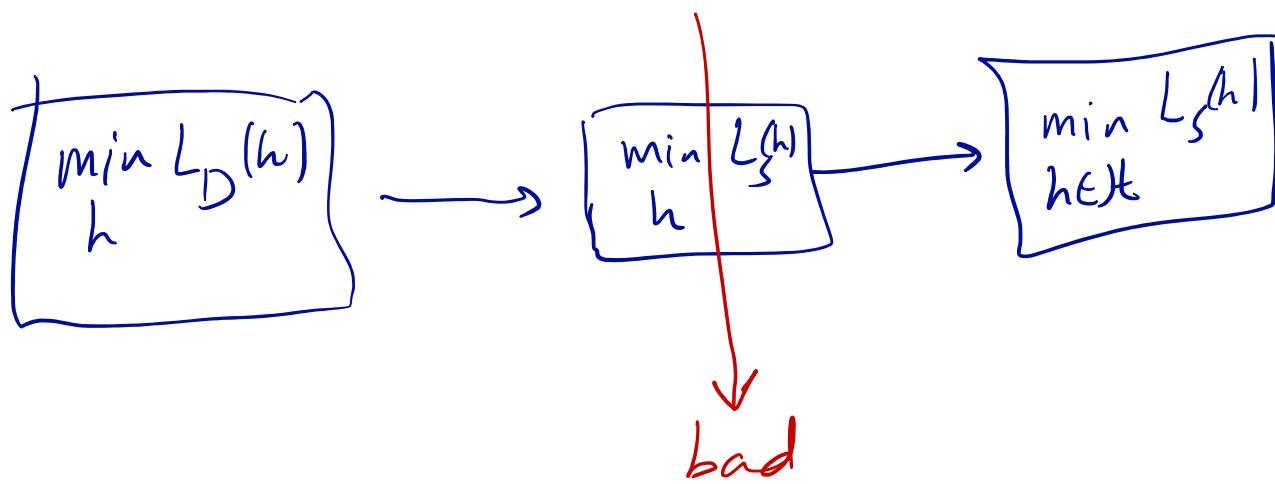
f) we need to search for conditions under which there is a guarantee that ERM does not overfit, namely conditions under which when the ERM predictor has a good performance with respect to the training data, it is also likely to perform well over the underlying (true) distribution



A common solution is to apply ERM over a restricted set of functions (classifiers). Formally the learner should choose in advance (before seeing the data) a set of functions (classifiers). This set is called "the predictor class" or "the function class" or "the hypothesis class" and is denoted by  $H$ . Each  $h \in H$  is a function from  $X$  to  $Y$ .

And the restricted ERM problem ( $ERM_{\mathcal{H}}$ ) is :

$$\text{minimize}_{h \in \mathcal{H}} L_S(h) -$$



So we are biasing the learner towards a particular set of predictors.

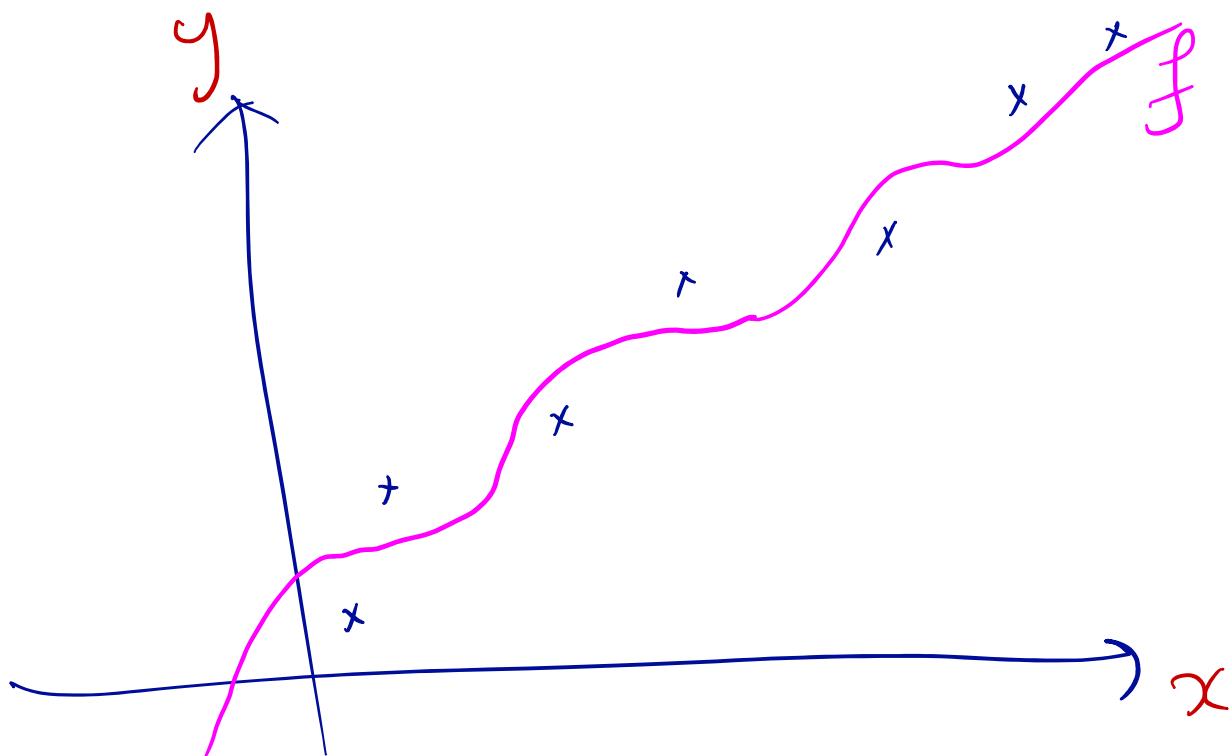
To explain things further, let's look at an example about regression (even though it is not classification, it will help us understand the concepts better).

$$y = f(x) + \epsilon$$

independent gaussian  
 $N(0, \sigma^2)$

We don't know  $f$  and we'd like to estimate the predictive relation between  $x$  and  $y$  using training data.

$$S = \{(x_i, y_i)\}_{i=1}^n$$



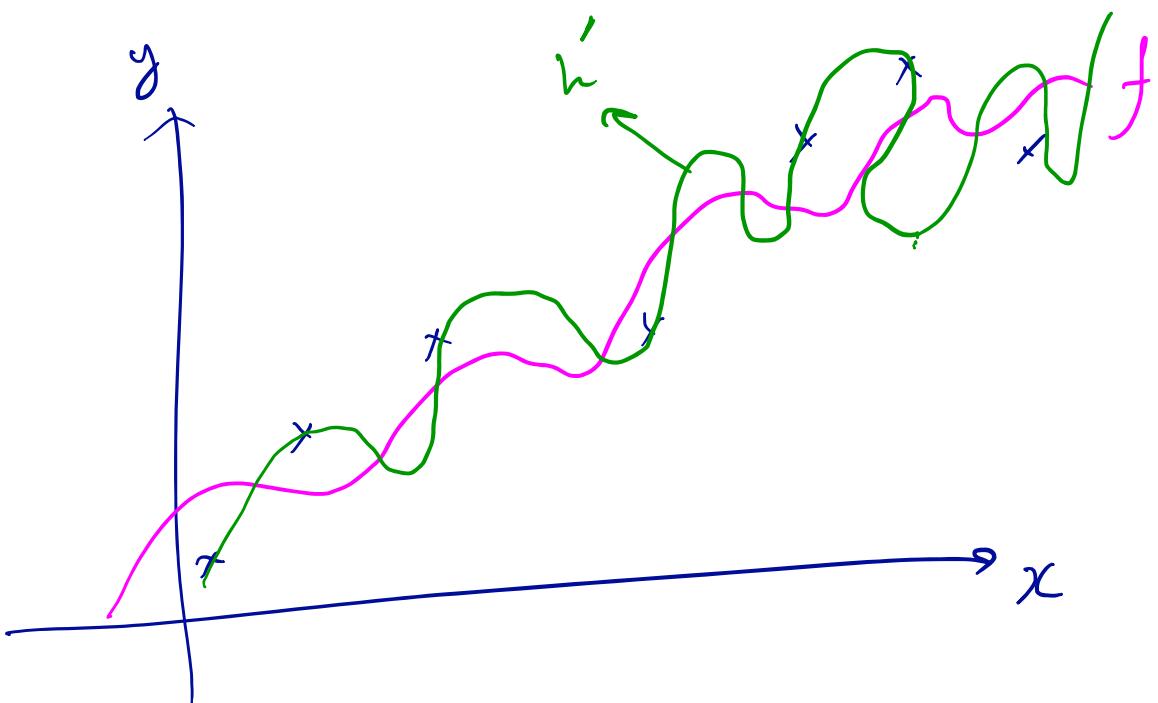
For any  $h: X \rightarrow Y$

$$L_D(h) = E_{(x,y) \sim D} [(h(x) - y)^2]$$

$$\min_h L_D(h) = 6^2$$

$\hookrightarrow$  minimizer  $\rightarrow f$

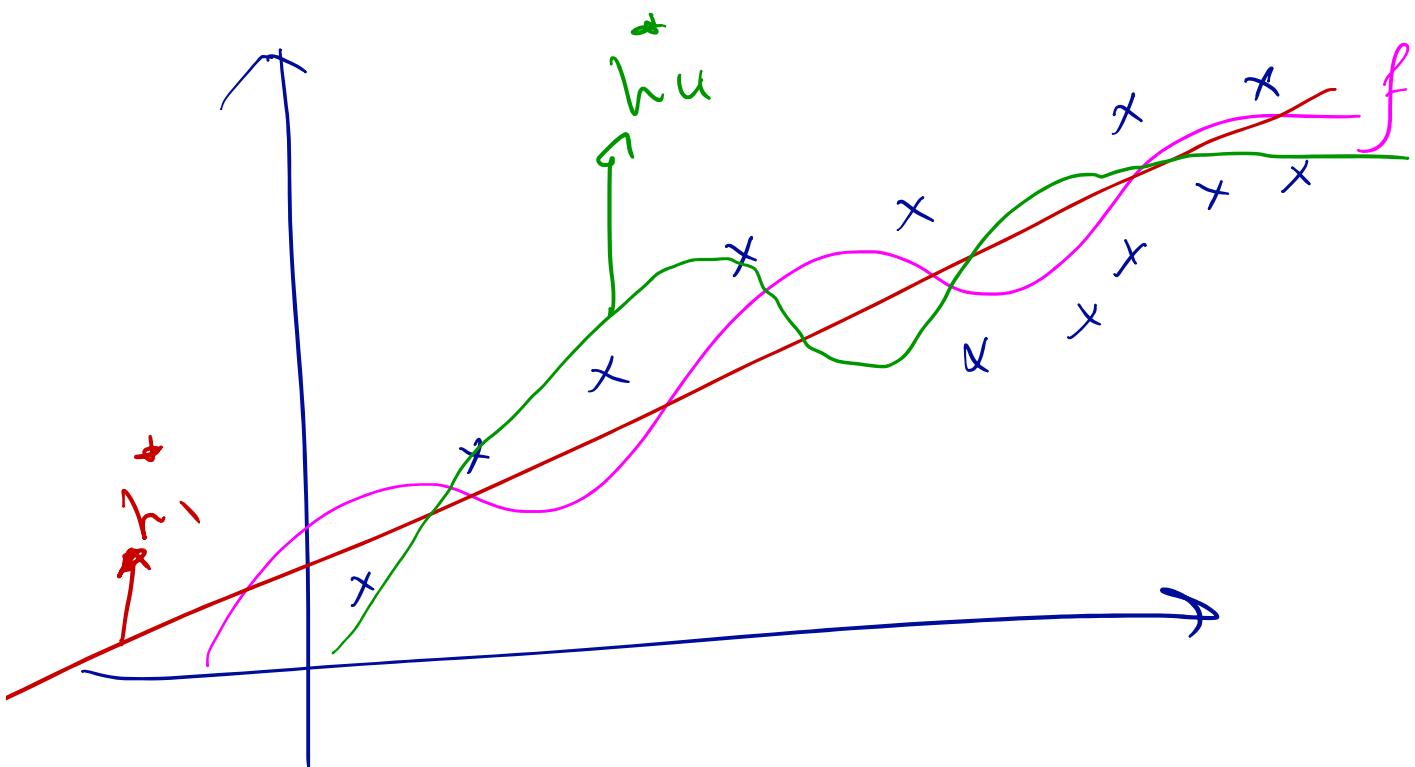
$$L_D(f) = E[(y - f(x))^2] = 6^2$$



$$\left\{ \begin{array}{l} \min_h L_D(h) = 6^2 \\ \min_h L_S(h) = 0 \end{array} \right. \rightarrow L_S(h') = 0$$

Let's now consider restricted  
ERM objectives:

$$\min_{h \in \mathcal{H}} L_S(h)$$



$$ERM_{\mathcal{H}} \rightarrow \min_{h \in \mathcal{H}} L_S(h)$$

e.g. when  $\mathcal{H}_1 = \left\{ \begin{array}{l} \text{polynomials of} \\ \text{degree 1} \end{array} \right\}$

$$\hookrightarrow \min_{h \in \mathcal{H}_1} L_S(h) \rightarrow h_1^*$$

$\mathcal{H}_4 = \left\{ \begin{array}{l} \text{polynomials of} \\ \text{degree 4} \end{array} \right\}$

$$\min_{h \in \mathcal{H}_4} L_S(h) \rightarrow h_4^*$$

These are two important points:

(1) since we are restricting our function class, our best bet would be to achieve

$$\min_{h \in \mathcal{H}} L_D(h)$$

which is larger than

$$\min_{h \in \mathcal{H}} L_D(h) \geq \min_h L_D(h)$$

$$\min L_D(h) \neq 0$$

Hence, if the class  $\mathcal{H}$  is not rich enough we may have (underfitting)

$$\min_{h \in \mathcal{H}} L_D(h) > \min_h L_D(h)$$

$\Rightarrow \mathcal{H}$  should be sufficiently complex

(2) The minimizer of

$$\min_{h \in \mathcal{H}} L_S(h)$$

is not the same as  
the minimizer of

$$\min_{h \in \mathcal{H}} L_D(h)$$

In other words if we denote

$$\rightarrow h_s^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$$

↳ this is computable

then it is possible that

$$L_D(h_s^*) \gg \min_{h \in \mathcal{H}} L_D(h)$$

↑  
overfitting

To avoid over fitting in  $\mathcal{H}$

We should make sure  
that we have sufficiently  
many training data points.

$h_S^*$  → minimizer of  $L_S(h)$  within the class  $\mathcal{H}$ .

since  $h^* \in \mathcal{H} \Rightarrow L_D(h^*) \geq \min_{h \in \mathcal{H}} L_D(h)$

$$L_D(h^*) \gg \min_{h \in \mathcal{H}} L_D(h)$$

↑  
Overfitting

two potential Potential problems:

(1)  $\min_{h \in \mathcal{H}} L_D(h) \gg \min_{h \in \mathcal{H}} L_D(h)$

↑  
Underfitting

we need  $\downarrow$  (  $h \in \mathcal{H}$  )  
to increase  
the complexity of  $\mathcal{H}$

Solution of ERM  $\mathcal{H}$

(2)  $L_D(h_S^*) \gg \min_{h \in \mathcal{H}} L_D(h)$

we need to  
increase the # of training data points

↓  
Overfitting

## Lecture 24:

$$\min_h L_D(h) \rightarrow \min_h \underbrace{L_S(h)}_D$$

$$\tilde{h} = \begin{cases} y_i & x = x_i \\ 0 & x \notin \{x_1, \dots, x_n\} \end{cases}$$

$$L_S(\tilde{h}) \geq 0$$

$$L_D(\tilde{h}) \text{ very large}$$

— — — —

$$\min_{h \in \mathcal{H}} L_S(h)$$

$$h \in \mathcal{H}$$

1. Since we're restricting our search to  $\mathcal{H}$

$$\min_{h \in \mathcal{H}} L_D(h) \geq \min_h L_D(h)$$

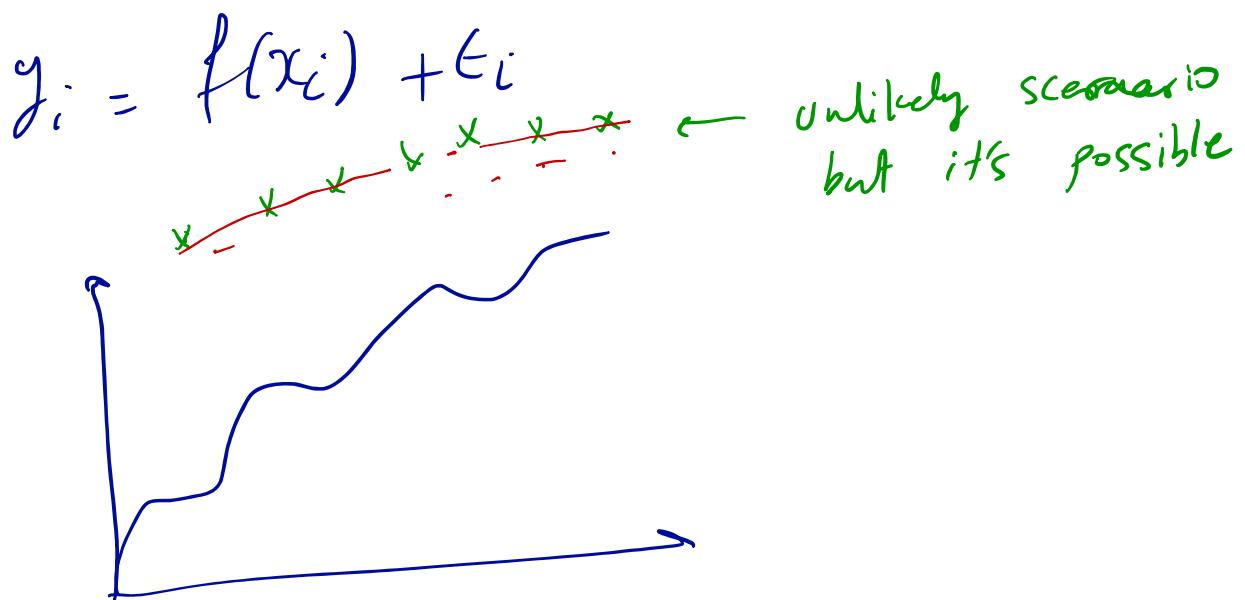
2. what we're actually optimizing is

$$\min_{h \in \mathcal{H}} L_S(h) \implies \text{Let's denote the minimizer by } h^*$$

$$L_D(h^*) \geq \min_{h \in H} L_D(h)$$

↳ if  $L_D(h^*)$  is much larger than  $\min_{h \in H} L_D(h)$ , then this is called overfitting.

3. All the guarantees and bounds are probabilistic (they'll hold with high probability excluding the unlikely scenarios).



$$L_D \rightarrow E_D$$

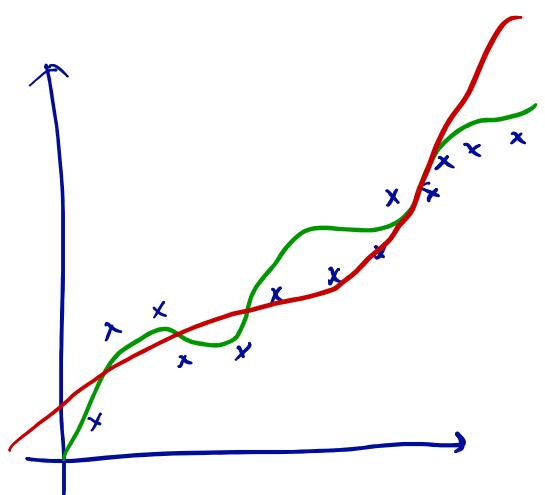
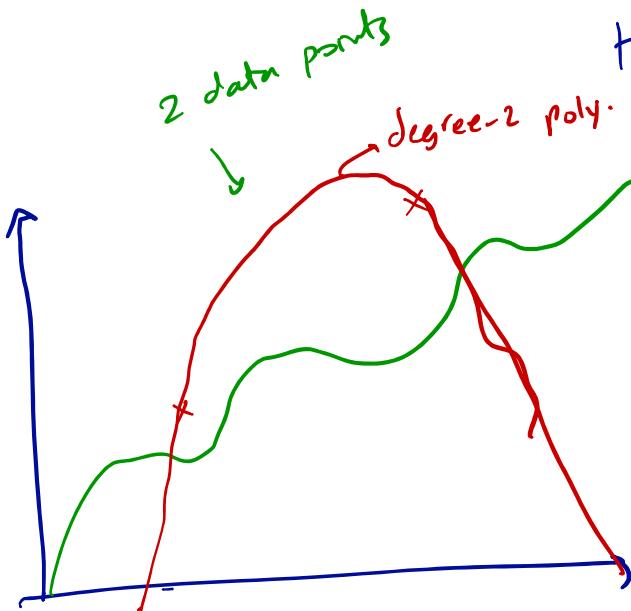
$$L_S \rightarrow \frac{1}{n} \sum_i$$

$$L_S \rightarrow L_D$$

$$\min_{h \in H} L_S(h)$$

$$\rightarrow$$

in order to ensure that overfitting does not take place, we need to have a sufficient number of training data points.



$$y_i = f(x_i) + \epsilon_i$$

$$H = \{ \text{polynomials of degree } 3 \}$$

$\Rightarrow$  overfitting can be avoided provided that we have sufficiently many training data points.

Hence, the main question is:

How many training data points do we need to make sure that overfitting does not happen?

↪ # of data points that we need will depends on:

$$H, \epsilon, \delta$$

What does "avoiding overfitting" mean?

$$h^*_{\text{S}} = \underset{h \in H}{\operatorname{argmin}} L_S(h)$$

what we can  
do computationally

$$\min_{h \in H} L_D(h)$$

the best error  
that we can  
hope for.

overfitting is avoided when:  $L_D(h^*)$  is close to  $\min_{h \in H} L_D(h)$ .

overfitting is avoided when

$$\min_{h \in \mathcal{H}} L_D(h) \leq L_D(h^*) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon$$

↓  
small value

With probability  $1 - \delta$  we have:

$$L_D(h^*) - \min_{h \in \mathcal{H}} L_D(h) \leq \epsilon$$

↓

provided that  
sufficiently  
many data  
points are  
available.

if we have more than no data  
points, then with high probability  
the following is true: with probability  
at least  $1 - \delta$

$$L_D(h^*) - \min_{h \in \mathcal{H}} L_D(h) \leq \epsilon$$

$\epsilon$  depends on  $\epsilon, \delta, \mathcal{H}$ .

PAC Learning:

PAC = Probably Approximately Correct

Definition (PAC): A function class  $H$  is called PAC learnable if for every  $\epsilon, \delta \in (0, 1)$ , there exists a number  $n_0(\epsilon, \delta)$  such that the following holds:

There exist a learning algorithm that for any distribution  $D$  over  $X \times Y$ , by using a set  $S$  of  $n_0(\epsilon, \delta)$  data points, we obtain a predictor (function)  $h_S^*$  such that with probability  $1 - \delta$  we have

$$\min_{h \in H} L_D(h) \leq L_D(h_S^*) \leq \min_{h \in H} L_D(h) + \epsilon.$$

$$\mathcal{H} \xrightarrow{\quad} \left\{ \begin{array}{l} \text{Algorithm} \rightarrow h^*_S \\ n_0(\epsilon, \delta) \end{array} \right. \Rightarrow \mathcal{H} \text{ is PAC learnable}$$

Let's start with the simplest possible hypothesis class and see what PAC-learning means with that class.

$$\mathcal{H} = \{h_1, h_2, \dots, h_m\}$$

$$\min_{h \in \mathcal{H}} L_D(h) \longrightarrow \min_{h \in \mathcal{H}} L_S(h)$$

$h^*_S$  is one of the functions  $h \in \mathcal{H}$  with smallest empirical error.

$$\min_{h \in \mathcal{H}} L_S(h) = \min \{L_S(h_1), \dots, L_S(h_m)\}$$

What we'd like to find in this case is a number  $n_0(\epsilon, \delta)$  such that for my distribution  $D$  over data we have:

with probability  $1-\delta$ :

$$L_D(h_S^*) \leq \min_{h \in H} L_D(h) + \epsilon$$


---

In other words we're asking how large the number of training data points should be s.t. we have w.p.  $1-\delta$  that  $L_D(h_S^*) \leq \min_{h \in H} L_D(h) + \epsilon$ .

---

In order to find  $n_0(\epsilon, \delta)$  we need to answer two main questions.

(1) Given a fixed function  $h: x \rightarrow y$ ,  
 how many training data points should  
 we have such that:  
 with probability  $1-\delta$ :

$$\left| L_S(h) - L_D(h) \right| < \epsilon$$

$$\left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{h(x_i) \neq y_i\} - E_D[\mathbb{1}\{h(x) \neq y\}] \right| < \epsilon$$

What should be  $|S| = n_*(\epsilon, \delta)$ ?

To answer this question we'll use  
 a very important probabilistic tool -  
 which is called the Hoeffding's  
 inequality. This inequality has very  
 useful in a variety of applications  
 in data science (both in terms of theory

and algorithm design).

## Hoeffding's Inequality:

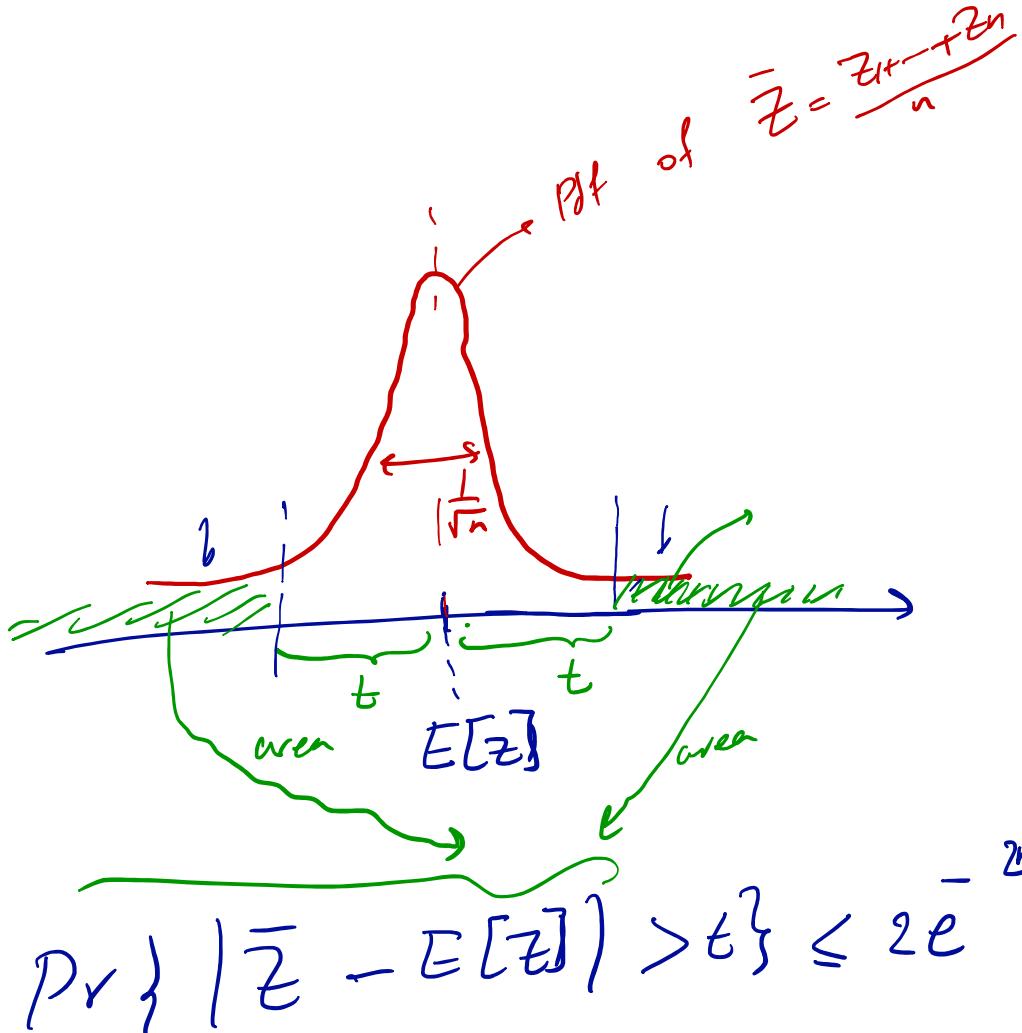
Let  $Z_1, Z_2, \dots, Z_n$  be  $n$  iid random variables that take value in the unit interval (i.e.  $Z_i \in [0, 1]$ ).

Then:

$$\Pr \left\{ \left| \frac{\sum_{i=1}^n Z_i}{n} - E[Z_i] \right| > t \right\} \leq 2e^{-2nt^2}$$

$Z_1, Z_2, \dots, Z_n$  iid  $Z \rightarrow \delta^2 = \text{var}(Z)$

$$\frac{Z_1 + Z_2 + \dots + Z_n}{n} \sim E[Z] + \frac{1}{\sqrt{n}} N(0, \delta^2)$$



e.g.  $n=1000$ ,  $t=0.5$

$$e^{-2 \cdot 1000 \cdot (0.5)^2} = e^{-500} \rightarrow 0$$

# Lecture 25:

$H \rightarrow$  PAC learnable:

(1)  $n_0(\epsilon, \delta)$

(2) learning algorithm that takes as input a set  $f$  of training data points  $S$ , and outputs  $h^*$ .

for any data distribution  $D$ , as

long as  $|S| \geq n_0(\epsilon, \delta)$  then

with probability  $1 - \delta$  we

have:

$$L_D(h_s^*) \leq \min_{h \in H} L_D(h) + \epsilon$$

Example:

$$\mathcal{H} = \{ h_1, \dots, h_m \}$$

(1) Specify what  $n_{\mathcal{E}, \delta}$  is.

(2) Algorithm: ERM  $\mathcal{H}$ :

$$h_S^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h)$$

What is  $n_{\mathcal{E}, \delta}$ , s.t.

w.p.  $1 - \delta$

$$L_D(h_S^*) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon.$$

Two questions:

(1) Given a fixed function  $h$ ,  
how many data points,  $S$ , do  
we need such that

W.P. 1-8

$$\left| L_S(h) - L_D(h) \right| < \varepsilon$$

$$\frac{1}{|S|} \sum_{i=1}^{|S|} \mathbb{1}\{h(x_i) \neq y_i\}$$

↑ equivalent

$$\Pr \left\{ \left| L_S(h) - L_D(h) \right| \geq \varepsilon \right\} \leq \delta$$

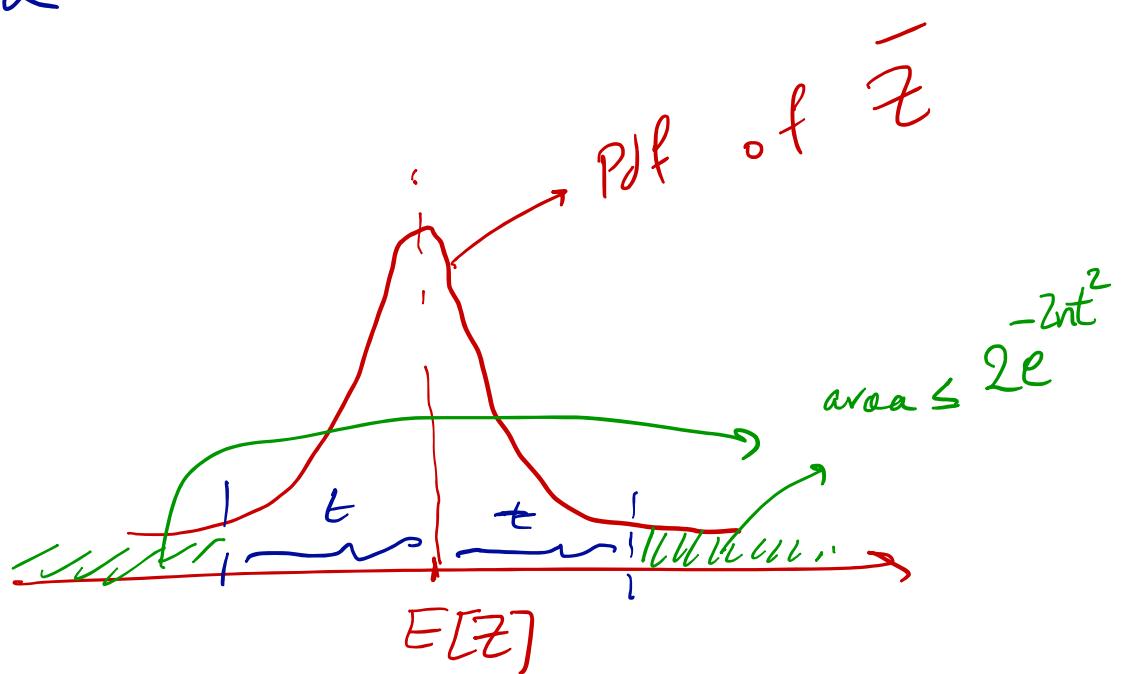
Recall that Hoeffding's inequality was given as follows:

For any iid random variables

$Z_1, Z_2, \dots, Z_n$ , s.t.  $Z_i \in [0, 1]$ ,

we have

$$\Pr \left\{ \left| \overline{\frac{1}{n} \sum_{i=1}^n Z_i} - E[\bar{Z}] \right| > t \right\} \leq 2e^{-2nt^2}$$



$$L_S(h) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{1}\{h(x_i) \neq y_i\}}_{z_i}$$

( \$|S|=n\$ )

$$L_D(h) = E_{(x,y) \sim D} [ \mathbb{1}\{h(x) \neq y\} ]$$

$E[z]$

$$L_S(h) = \frac{1}{n} \sum_{i=1}^n z_i = \bar{z}$$

$$z_i = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & \text{o.w.} \end{cases} \rightarrow z_i \in [0,1]$$

$$E[z_i] = E_{(x_i, y_i) \sim D} [ \mathbb{1}\{h(x_i) \neq y_i\} ]$$

$$= L_D(h)$$

Hoeffding:

$$\Pr \left\{ \left| \underbrace{\frac{1}{n} \sum z_i}_{L_S(h)} - \underbrace{E[z]}_{L_D(h)} \right| > \epsilon \right\} \stackrel{(1)}{\leq} 2e^{-2n\epsilon^2}$$

for any  $\epsilon > 0$

Recall that we're looking for the smallest  $n$  s.t.

$$\Pr \left\{ |L_S(h) - L_D(h)| > \epsilon \right\} \leq \delta \quad (2)$$

using (1), and in order to guarantee

(2), we can let:

$$2e^{-2n\epsilon^2} \leq \delta \quad \left( \log \frac{1}{\delta} = -\log \frac{1}{2} \right)$$

$$\Rightarrow n \geq \frac{1}{2\epsilon^2} \log \frac{1}{\delta}$$

So, the final statement is:

Given any  $\epsilon, \delta$ , as long as the number of training data points is larger than

$$n_1 = \frac{1}{2\epsilon^2} \log \frac{2}{\delta}, \text{ we have}$$

$$\Pr_{(X, g) \in D} \left\{ |L_S(h) - L_D(h)| > \epsilon \right\} \leq \delta.$$

| for a fixed function  $h$  )

(2) Let's now assume that we have  $m$  functions  $h_1, h_2, \dots, h_m$ . What is the smallest value  $n_0(\epsilon, \delta)$  such that

With probability  $1 - \delta$ : (3)

$$\forall i \in \{1, \dots, m\} : |L_S(h_i) - L_D(h_i)| < \epsilon.$$

To answer this question, we're going to write an equivalent formulation of relation (3):

Let  $A_i$  be the event that

$$|L_S(h_i) - L_D(h_i)| > \epsilon .$$

Then (3) is equivalent to:

$$\Pr \left\{ A_1 \cup A_2 \cup A_3 \cup \dots \cup A_m \right\} \leq \delta . \quad (4)$$

Remember that we are searching for the smallest value of  $n$  such that (4) holds.

$$\Pr \{ A_1 \cup A_2 \cup \dots \cup A_m \} \leq \delta$$

To answer this, we're going to use the Union bound:

$$\Pr \{ A \cup B \} \leq \Pr \{ A \} + \Pr \{ B \}$$

$$\Pr \{ A_1 \cup A_2 \cup \dots \cup A_m \} \leq \Pr \{ A_1 \} + \Pr \{ A_2 \} + \dots + \Pr \{ A_m \}$$

bad event # $i$  :  $A_i$  :

$$|L_S(h_i) - L_D(h_i)| > \epsilon$$

We'd like to make sure that none of the bad events,  $A_i$ , would take place.

So in order to guarantee

$$\Pr \{ A_1 \cup A_2 \cup \dots \cup A_m \} \leq \delta \quad (5)$$

it is sufficient to guarantee that

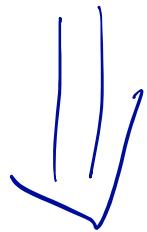
$$\Pr \{ A_1 \} + \Pr \{ A_2 \} + \dots + \Pr \{ A_m \} \leq \delta \quad (6)$$

Note that if (6) holds, then (5) would also hold as a result of the union bound.

Now, to guarantee (6), it is sufficient to choose  $n$  large enough s.t. for every  $i$

We have

$$\Pr \{ A_i \} \leq \frac{\delta}{m} \quad (7)$$



$$\Pr \{ A_1 \} + \Pr \{ A_2 \} + \dots + \Pr \{ A_m \} \leq \delta$$



$$\Pr \{ A_1 \cup A_2 \cup \dots \cup A_m \} \geq \delta$$

---

Now given our answer to question 1,

in order to guarantee that

$$\Pr \{ A_i \} \leq \frac{\delta}{m} \text{ we need to}$$

choose:

$$\text{if } n \geq n_1(\epsilon, \frac{\delta}{m}) = \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}$$

$\underbrace{\phantom{\frac{1}{2\epsilon^2} \log \frac{2m}{\delta}}}_{n_1(\epsilon, \delta)}$

then

$$\Pr \{ A_i \}$$

$$= \Pr \{ |L_S(h_i) - L_D(h_i)| > \epsilon \} \leq \frac{\delta}{m}$$

Hence,

Statement: If the number of training data points,  $|S|$ , is larger

$$\text{then } n_0(\epsilon, \delta) = \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}, \text{ then}$$

with probability  $1 - \delta$  we have

$$\forall i \in \{1, \dots, m\} : |L_S(h_i) - L_D(h_i)| < \epsilon.$$

What we've shown is that we need

$\frac{1}{2\epsilon^2} \lg \frac{2m}{\delta}$  data points to

~~guarantee that for all the~~

$m$  functions  $h \in \mathcal{H}$  the value of  $L_S(h)$  is close to the

Value of  $L_D(h)$ :

w.p.  $1-\delta$ :

$\forall h \in \mathcal{H} : |L_S(h) - L_D(h)| < \epsilon.$

Now, let  $h_s^*$  be the minimizer of ERM over the class  $\mathcal{H}$ :

$$h_s^* : \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) \quad \left( \begin{array}{l} L_S(h_s^*) \\ \text{is the} \\ \text{smallest} \\ \text{among } \mathcal{H} \end{array} \right)$$

Let  $h_{\text{true}}^* : \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_D(h)$

We will show that if the number of training data points is at least  $N_0(\epsilon, \delta)$  then with  $1 - \delta$  we have

$$L_D(h_s^*) - \min_{h \in \mathcal{H}} L_D(h) \leq 2\epsilon. \quad (8)$$

Hence,  $\mathcal{H}$  is PAC-<sup>enable</sup>  
with  $n_{\epsilon}(\epsilon, \delta)$  data points.

Let's see why (8) holds.

$$\begin{aligned}
 & L_D(h_S^*) - L_D(h_{\text{true}}^*) \\
 &= \underbrace{L_D(h_S^*) - L_S(h_S^*)}_{\leq 0} + \underbrace{L_S(h_S^*) - L_S(h_{\text{true}})}_{\substack{\text{$h_S^*$ is the} \\ \text{minimizer of} \\ L_S(\cdot)}} \\
 &\quad + \underbrace{L_S(h_{\text{true}}^*) - L_D(h_{\text{true}}^*)}_{\leq \epsilon} \\
 &\leq \epsilon + 0 + \epsilon \leq 2\epsilon.
 \end{aligned}$$

$$\leq \epsilon + 0 + \epsilon \leq 2\epsilon.$$

Hence, if the number of training data points is

at least  $n_0(\epsilon, \delta) = \frac{1}{2\epsilon^2} \log \frac{2m}{\delta}$

then w.p.  $1-\delta$  we have

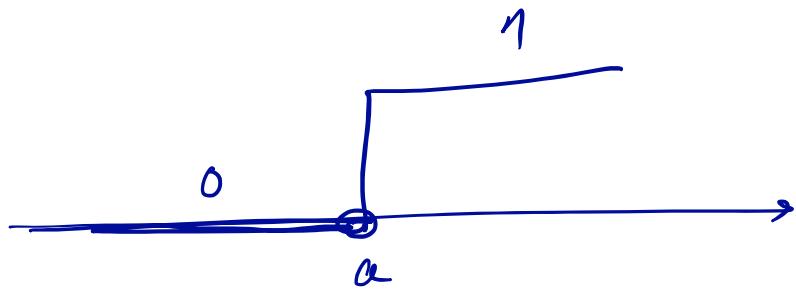
$$0 \leq L_D(h^*) - \min_{h \in \mathcal{H}} L_D(h) \leq 2\epsilon.$$

minimizer of

ERM over  $\mathcal{H}$ .

When  $\mathcal{H} = \{h_1, \dots, h_m\}$ .

$$h_a(x) =$$



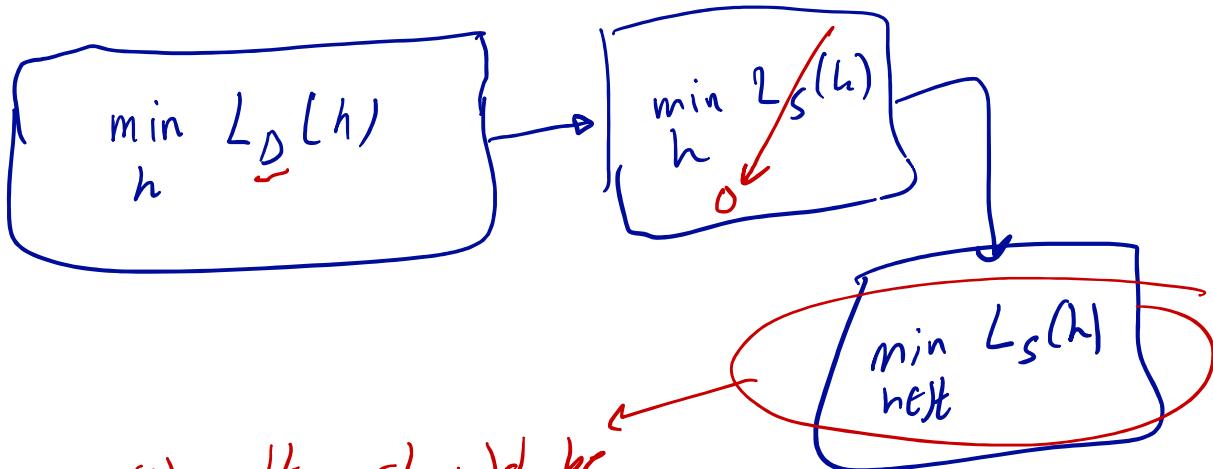
$$\mathcal{H} = \left\{ h_a(x) , a \in [-\infty, \infty) \right\}$$

infinitely many functions

One important consequence of PAC-learnability is that it's sufficient to work with a ~~finite~~ finite data set and we don't lose anything in terms of generalization

## Lecture 26:

-  $(x, y) \sim D$   $\rightarrow$



(1) It should be sufficiently complex

(2) Avoid overfitting = have sufficiently many data points

PAC:  $H$  is PAC learnable:

(1)  $n_0(\epsilon, \delta) \in \mathbb{N}$  training data

(2) learning alg.  $S \rightarrow h_S^* \in H$

If  $|S| \geq n_0(\epsilon, \delta)$ :

w.p.  $1 - \delta$ :

$$L_D(h_S^*) \leq \min_{h \in H} L_D(h) + \epsilon$$

$$\mathcal{H} = \{h_1, \dots, h_m\}$$

up to constants

$$\frac{\log \frac{1}{\delta} + \overbrace{\log m}^{\text{up to constants}}}{\epsilon^2}$$

- $n_0(\epsilon, \delta) = \frac{1}{\epsilon^2} \log \frac{2m}{\delta} \approx \frac{\log \frac{1}{\delta} + \log m}{\epsilon^2}$
- learning algorithm:  $h_s^* = \underset{i=1, \dots, m}{\operatorname{argmin}} L_s(h_i)$

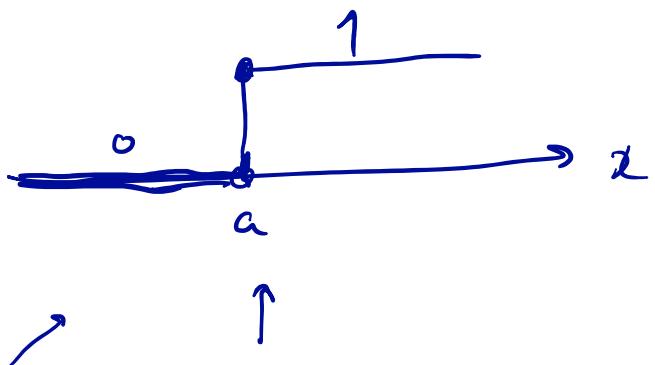
$\mathcal{H}$  is PAC-learnable with  
 $n_0(\epsilon, \delta)$  samples and ERM-like as  
the learning algorithm.

---

What about infinite  $\mathcal{H}$ ?

$$\mathcal{H} = \left\{ h_a(x) : \mathbb{R} \rightarrow \{0, 1\}, a \in \mathbb{R} \right\}$$

$$h_a(x) \rightarrow$$



The fundamental theorem of learning theory:

For any function class  $H$ , if the number of training data point is larger than

$$n_0(\epsilon, \delta) = \frac{\log \frac{1}{\delta} + \overbrace{\text{VC-dim}(H)}^{\text{Complexity of the class}}}{\epsilon^2}, \text{ then}$$

w.p.  $1 - \delta$ :

$$L_D(h_s^*) \leq \min_{h \in H} L_D(h) + \epsilon.$$

where  $h_s^* = \underset{h \in H}{\operatorname{argmin}} L_S(h)$ .

---

→  $\text{VC-dim}(H) = \text{Complexity of the class}$

→ if  $\text{VC-dim}(H)$  is finite, then

It is PAC-learnable with

$$n_0(\epsilon, \delta) = \frac{\log \frac{1}{\delta} + \text{VC-dim}(H)}{\epsilon^2} \text{ and}$$

Term as the learning algorithm.

Important implication:

We can learn from finite data

sets.

the best that we could

$$L_D(h_s^*) \leq \underbrace{\min_{h \in H} L_D(h)}_{\text{intractable}} + \epsilon$$

$L_D(h_s^*)$   
↓  
tractable

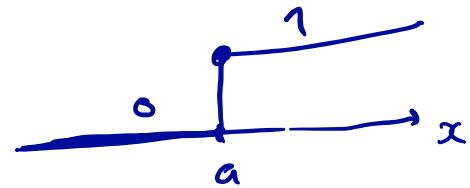
As long as we have sufficiently many data points we'll be fine with solving the "tractable" ERM-like problem and we lose at most a small value  $\epsilon$  with respect to the optimal accuracy.

VC dimension:

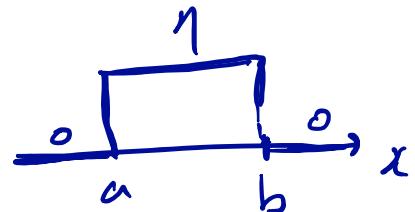
Vapnik-Chernovskii

Consider the following function classes:

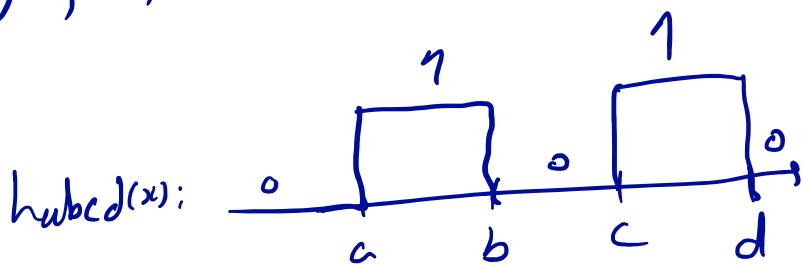
$$H_1 = \{ h_a(x), a \in \mathbb{R} \} \rightarrow$$



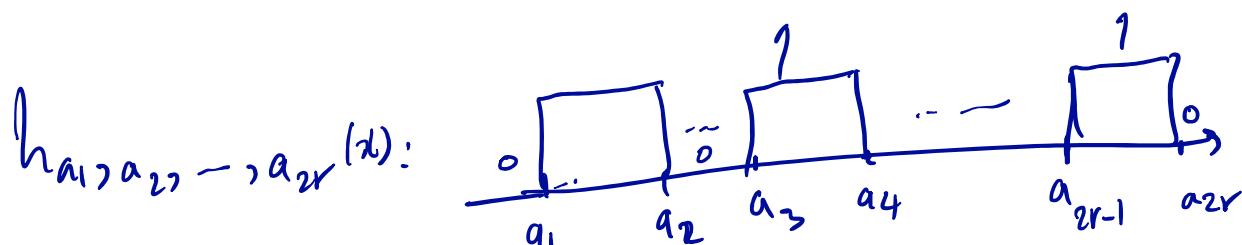
$$H_2 = \{ h_{ab}(x), a, b \in \mathbb{R} \} \rightarrow$$



$$H_3 = \{ h_{abcd}(x), a, b, c, d \in \mathbb{R} \}$$



$$H_4 = \{ h_{a_1, a_2, \dots, a_{2r}}(x), a_1, \dots, a_{2r} \in \mathbb{R} \}$$



$H_1 < H_2 < H_3 < H_4 \rightarrow$  complexity

Definition: (restriction):

Let  $\mathcal{H}$  be a class of functions from  $X$  to  $\{0,1\}$  ( $\forall h \in \mathcal{H}, h: X \rightarrow \{0,1\}$ ).

Let  $C = \{x_1, x_2, \dots, x_k\} \subseteq X$ .

The restriction of  $\mathcal{H}$  to  $C$  is the set of all the  $k$ -tuples

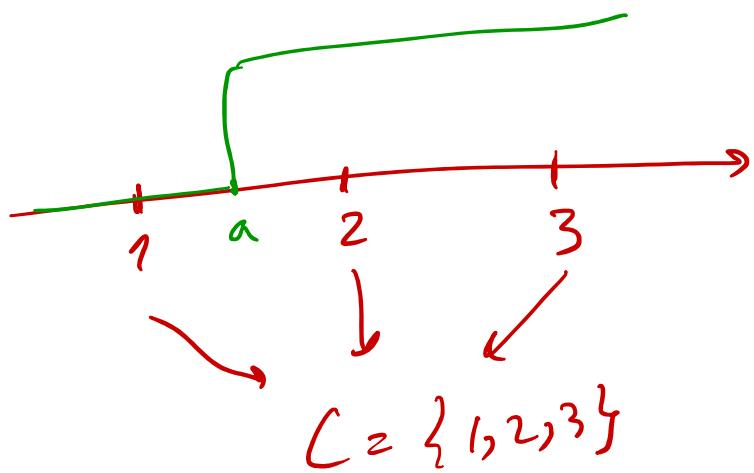
$$\mathcal{H}_C = \{ (h(x_1), h(x_2), \dots, h(x_k)) , h \in \mathcal{H} \}.$$

Example:  
e.g. if  $\mathcal{H} = \{ h_a(x) , a \in \mathbb{R} \}$

$$h_a(x) = \begin{cases} 1 & x > a \\ 0 & x \leq a \end{cases}$$

Let  $C = \{1, 2, 3\}$ .

$$\mathcal{H}_C = \{ (h_a(1), h_a(2), h_a(3)) , \begin{matrix} h_a \in \mathcal{H} \\ a \in \mathbb{R} \end{matrix} \}$$



$$a < 1 : (h_a(1), h_a(2), h_a(3)) = (1, 1, 1)$$

$$a \in [1, 2) : \quad " \quad = (0, 1, 1)$$

$$a \in [2, 3) \quad " \quad = (0, 0, 1)$$

$$a > 3 \quad " \quad = (0, 0, 0)$$

$$\mathcal{H}_C = \{(1, 1, 1), (0, 1, 1), (0, 0, 1), (0, 0, 0)\}$$

$$|\mathcal{H}_C| = 4 < 2^{|C|} = 2^3 = 8$$

$\hookrightarrow \mathcal{H}$  is not sufficiently complex to produce all the possible binary 3-tuples on  $C$ .

Note :

If  $C = \{x_1, \dots, x_k\}$

$$|\mathcal{H}_C| \leq 2^{|C|}$$



$$\mathcal{H}_C = \{(0, 0, \dots, 0), (1, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (1, 1, \dots, 1)\}$$

$$2^k$$

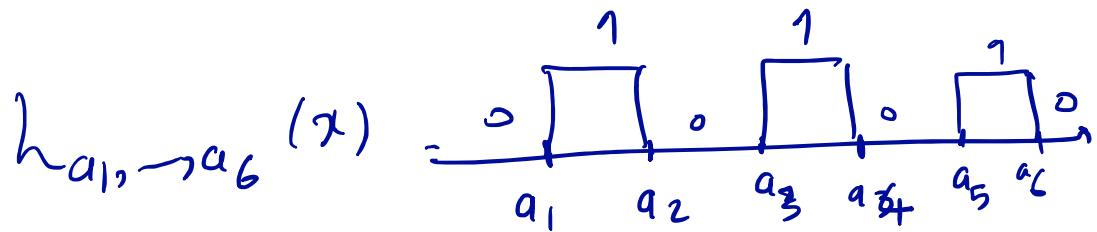
Definition (shattering). We say that

a function class  $\mathcal{H}$  shatters a

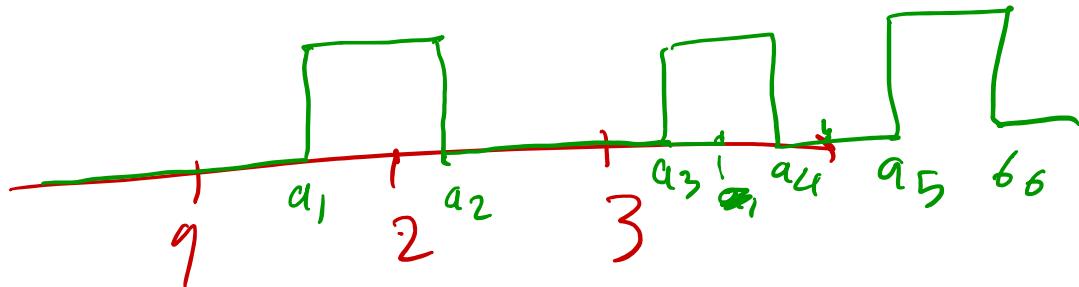
Set  $C$  if  $|\mathcal{H}_C| = 2^{|C|}$

Example:

$$\mathcal{H} = \left\{ h_{a_1 a_2 a_3 a_4 a_5 a_6}(x) , a_1, \dots, a_6 \in \mathbb{R} \right\}$$



$$C = \{1, 2, 3\}$$



$$\mathcal{H}_C = \{(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1), -1^{(1, 1, 1)}\}$$

$\downarrow$   
 $\mathcal{H}_C$  is all the 8 possible  
binary 3-tuples on  $C$ .  
 $\hookrightarrow$  It shutters  $C$ .

Definition (VC-dimension):

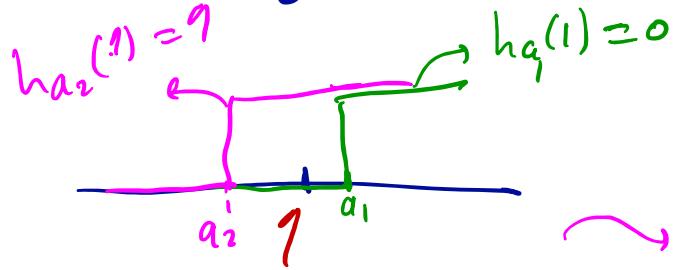
The VC-dimension of a function class  $\mathcal{H}$  is the largest number  $k$  such that exists a set  $C$  of size  $k$  which is shattered by  $\mathcal{H}$ .

Examples:

$$\mathcal{H} = \{ h_a(x) , a \in \mathbb{R} \} \rightarrow \begin{array}{c} \nearrow \\ \text{---} \\ \nearrow \end{array} \quad a$$

VC-dim( $\mathcal{H}$ ) =

Is there a set of size 1 that's shattered by  $\mathcal{H}$ ? YES.

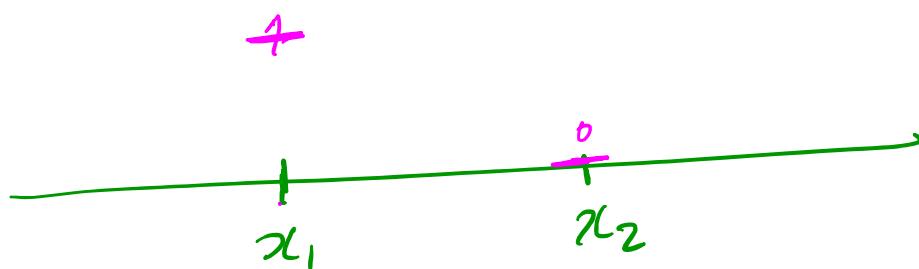


$$\mathcal{H}_C = \{0, 1\}$$

$$C = \{1\}$$

$\hookrightarrow \mathcal{H}$  shatters  $C$ .

Is there a set of size 2 which  
is shattered by  $\mathcal{H}$ ? No



$$C = \{x_1, x_2\} \quad (\text{assume } x_1 < x_2)$$

↳  $\mathcal{H}_C = \{(\circ, \emptyset), (\underline{1}, 0), (0, 1), (1, 1)\}$

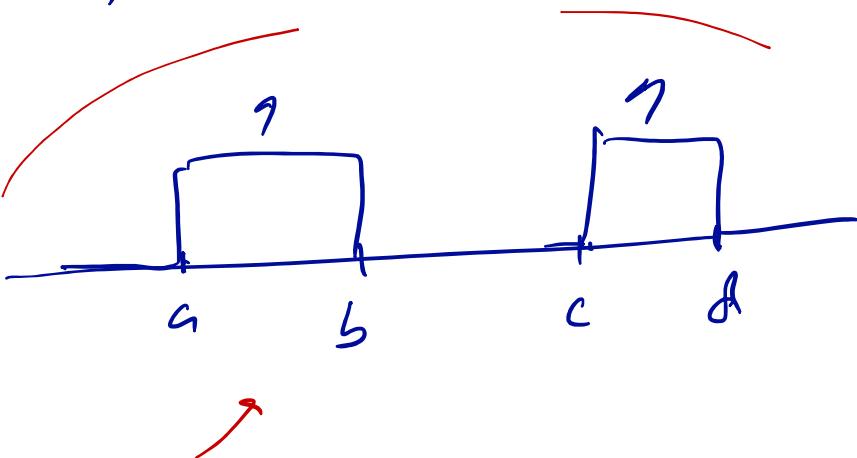
↳ there is no  $h_a(x)$   
that produces this  
binary 2-tuple

$$VC\text{-dim}(\mathcal{H}) =$$

there exists no set  $C$   
of size 2 which is shattered  
by  $\mathcal{H}$

Example:

$$\mathcal{H} = \left\{ h_{a,b,c,d}(x) : a, b, c, d \in \mathbb{R} \right\}$$



is there a set  $C$  of size  $k$   
which is shattered by  $\mathcal{H}$ ?

$k=1 \rightarrow \text{YES}$

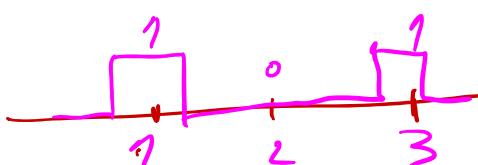
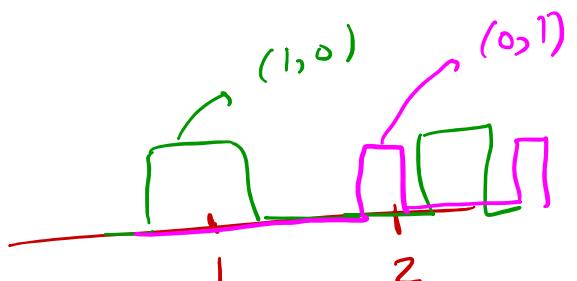
$k=2 \rightarrow \text{YES}$

$k=3 \rightarrow \text{YES}$

$k=4 \rightarrow \text{YES}$

$k=5 \rightarrow \text{NO}$

$$\mathcal{H}_C = \left\{ (0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,0,1), (1,1,0), (0,1,1), (1,1,1) \right\}$$



$k=5$



$\rightarrow h$  should have 3 jumps from 0 to 9.

