Epsilon®

26 OCT 2023

Introduction to Probability Statistics & Machine Learning

Rishi Chandra

© 2023 Epsilon Data Management, LLC. All rights reserved. Proprietary and confidential. V-3.20.23

Chapter-1 Contents

1. Summarizing Data Sets : Histograms
2. Summarizing Data Sets : Mean
3. Summarizing Data Sets : Variance
4. Percentile
5. Quartiles and Box Plots
6. Normal Data Sets Histograms / Distributions & Outliers
7. Correlation = strength of an assumed linear relationship

References:
1.Pattern-Recognition-and-Machine-Learning - Christopher M. Bishop, available at https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
2.Introduction to Probability and statistics for engineers and scientists - Sheldon M. Ross, available at https://minerva.it.manchester.ac.uk/~saralees/statbook3.pdf

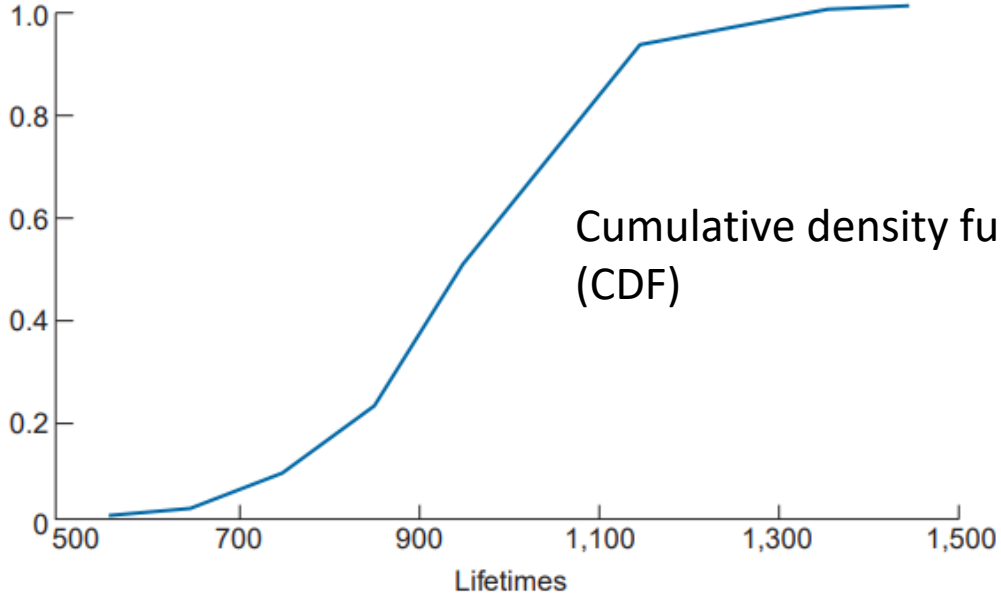
1.1 Summarizing Data Sets : Histograms

Relative frequency = Probability density (sums to 1)

For continuous values, bin into discrete buckets

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

Probability density function (PDF)

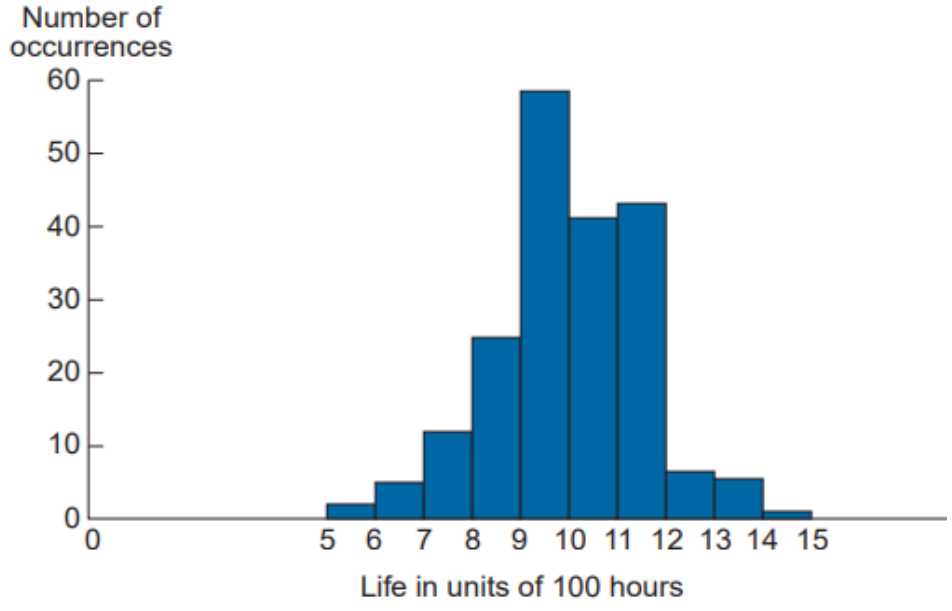


Cumulative density function (CDF)

A cumulative frequency plot.

TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1



A frequency histogram.

1.2 Summarizing Data Sets : Mean

- Sample Mean, Sample Median, Sample Mode

if Y = aX + b
then $\bar{Y} = a\bar{X} + b$

The *sample mean*, designated by $\bar{x}$, is defined by

Sample Mean $\bar{x} = \sum_{i=1}^{n} x_i/n$

if $y_i = ax_i + b, \quad i = 1, \ldots, n$

then the sample mean of the data set $y_1, \ldots, y_n$ is

$$\bar{y} = \sum_{i=1}^{n}(ax_i + b)/n = \sum_{i=1}^{n} ax_i/n + \sum_{i=1}^{n} b/n = a\bar{x} + b$$

➢ Sample mean is NOT population mean !

e.g., Voter exit polls are only a sample mean intended to representation of the overall population mean, but is not the population mean

Median

Median

8	15	50	200	1000

Median = (25+175)/2 = 100

1	3	25	175	880	9999

Order the values of a data set of size n from smallest to largest. If n is odd, the *sample median* is the value in position $(n+1)/2$; if n is even, it is the average of the values in positions $n/2$ and $n/2+1$.

Mode

- Most Frequent Value(s)

Another statistic that has been used to indicate the central tendency of a data set is the *sample mode*, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*.

1.3 Summarizing Data Sets : Variance

- Sample Variance and Standard Deviation, vs
- Population Variance and Std. Dev

The *sample variance*, call it s^2, of the data set $x_1, \ldots, x_n$ is defined by

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

$$\text{Var} = (\text{Std. Dev})^2$$

- Like sample mean is not population mean, **Sample Variance != Population Variance**
- Like sample mean is an estimate/approximation of the population mean, similarly sample variance is an estimate of the population variance.
- Sampling is done when it is not feasible to access/process the whole population.

Sample variance

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

S^2=sample variance

x_i =value of i th element

$\bar{x}$ = sample mean

n=sample size

Population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

σ^2=population variance

x_i =value of i th element

μ=population mean

N=population size

An Algebraic Identity

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

Variance of Translated variable

if $y_i = a + bx_i, \quad i = 1, \ldots, n$

if s_y^2 and s_x^2 are the respective sample variances, then $s_y^2 = b^2 s_x^2$

1.4 Percentile

If you score 99 percentile in an exam. It is not that you scored 99 % (percent).

It means that 99% of people scored <= you and 1% people score >= you

The *sample 100*p *percentile* is that data value such that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values.

To determine the sample $100p$ percentile of a data set of size n, we need to determine the data values such that

1. At least np of the values are less than or equal to it.
2. At least $n(1-p)$ of the values are greater than or equal to it.

Deciles

Arranging a set of data points in deciles means:

i. Sort the data points in increasing order,

ii. Distribute them in 10 bins in increasing order such that each bucket contains equal number of data points (10% in each bin). These bins are the deciles, e.g:

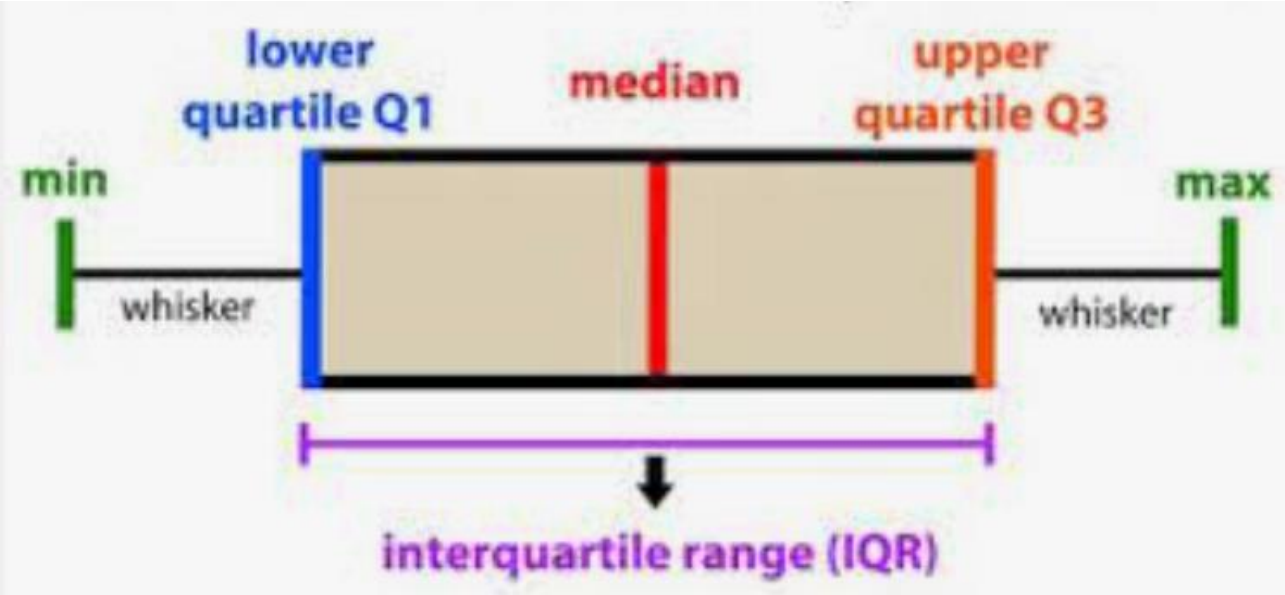
1	3	4	8	10	14	15	16	20	25	28	30	32	35	40	50	55	60	70	99
decile 1		decile 2		decile 3		decile 4		decile 5		decile 6		decile 7		decile 8		decile 9		decile 10	

1.5 Quartiles and Box Plots

- Box plot shows the min, 1st 2nd 3rd quartiles, and max values.
- 1st quartile = 25 percentile, 2nd quartile = 50 percentile = median, 3rd quartile = 75 percentile
- Box plot shows the skewness in data
- Also plot the mean in the box-plot. This gives a better view of the skew

The sample 25 percentile is called the *first quartile*; the sample 50 percentile is called the sample median or the *second quartile*; the sample 75 percentile is called the *third quartile*.

Box Plot

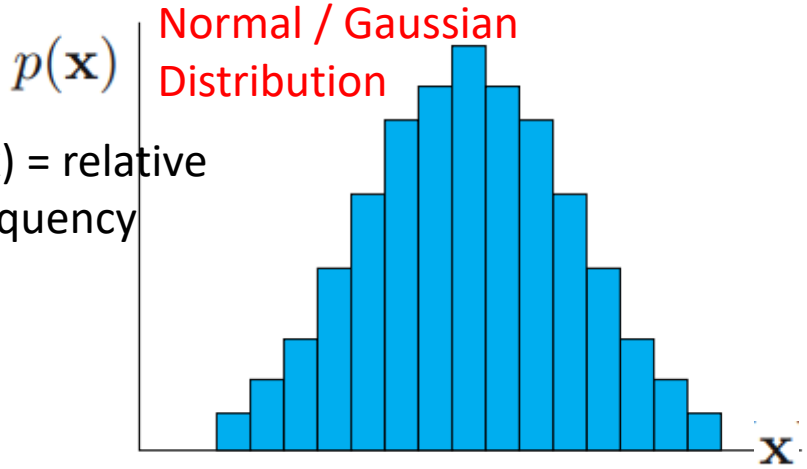


- The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the **range** of the data.
- the length of the box itself, equal to the third quartile minus the first quartile, is called the **interquartile range**.

1.6 Normal Data Sets Histograms / Distributions & Outliers

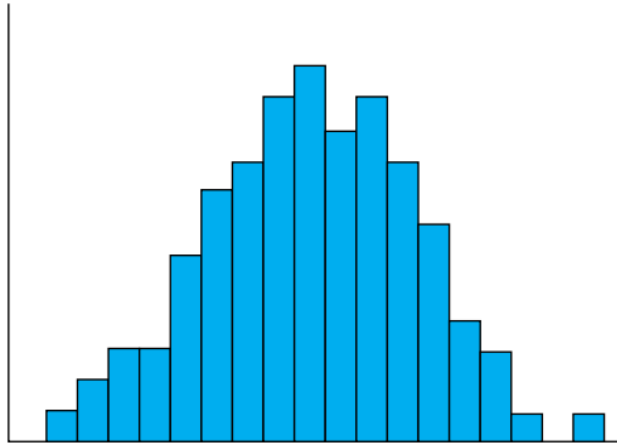
Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be **Normal** and their histograms are called **Normal histograms**

Normal / Gaussian Distribution



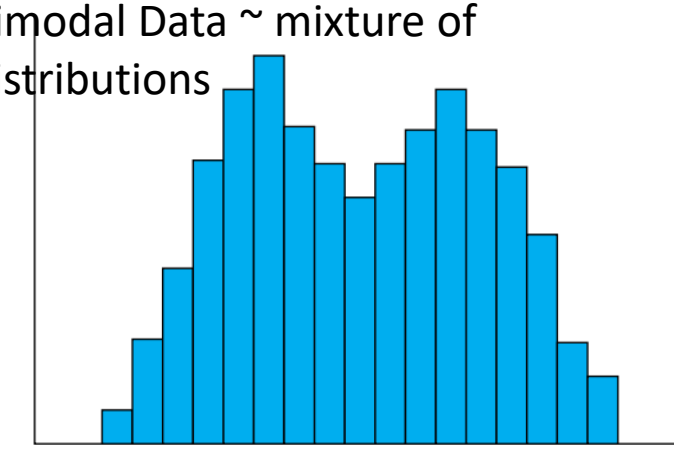
p(x) = relative frequency

Histogram of a normal data set.

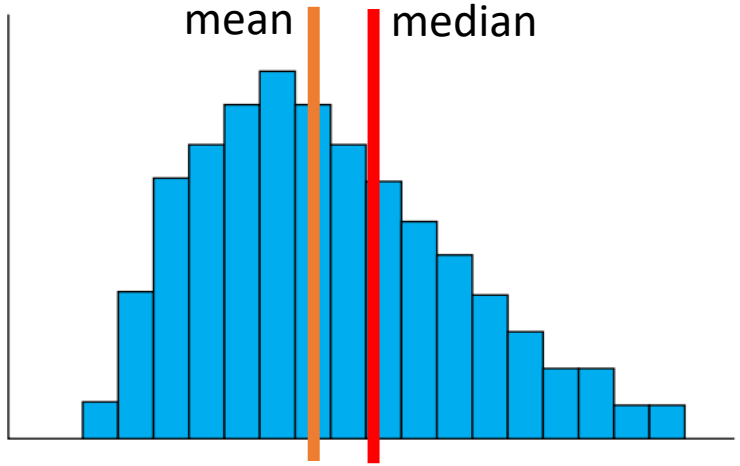


Histogram of an approximately normal data set.

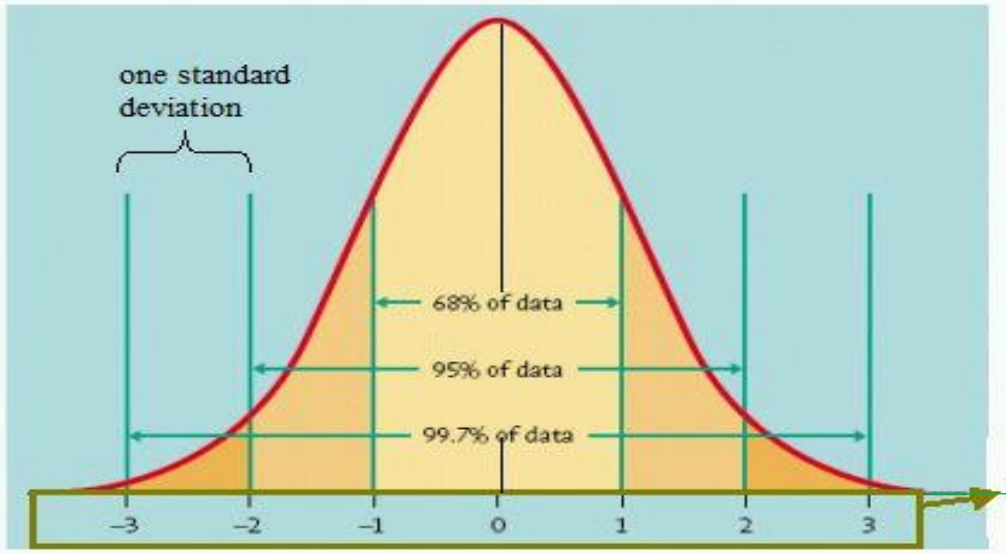
Bimodal Data ~ mixture of distributions



Histogram of a bimodal data set.



Histogram of a data set skewed to the right.



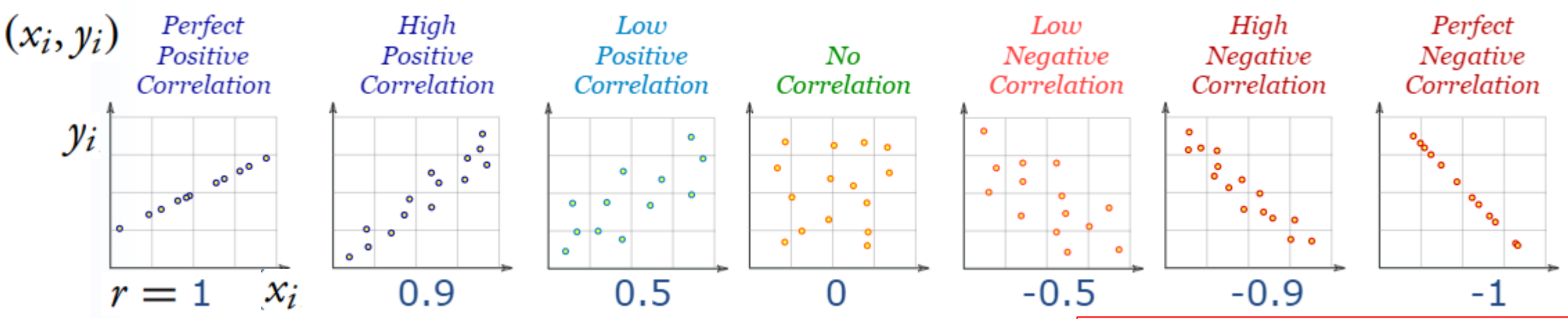
Z – score = how many StDev away is a point from its mean

If z > 3, it maybe an **outlier**

Z > 3: **outlier**

1.7 Correlation = strength of an assumed linear relationship

- Assumes and measures strength of linear relationship between two variables. Measures **Association, not Causation**
- An e.g., application: Features which are more correlated to the Target are more important. Fix all features at their mean except one, vary this by 1-sd and measure variation in target. The feature causing more variation in the Target is the more important one.



intuition: if $x_i - \bar{x}$ and $y_i - \bar{y}$ both have the same sign (either positive or negative), then their product $(x_i - \bar{x})(y_i - \bar{y})$ will be positive. Thus, it follows that when large x values tend to be associated with large y values and small x values are associated with small y values, then $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ will tend to be a large positive number.

$$\text{sample correlation} \quad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\text{population correlation} \quad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n \; s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Properties of *r*

1. $-1 \le r \le 1$
2. If for constants a and b, with $b > 0$,

   $$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

   then $r = 1$.
3. If for constants a and b, with $b < 0$,

   $$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

   then $r = -1$.
4. If r is the sample correlation coefficient for the data pairs $x_i, y_i, i = 1, \ldots, n$ then it is also the sample correlation coefficient for the data pairs

   $$a + bx_i, \quad c + dy_i, \quad i = 1, \ldots, n$$

   provided that b and d are both positive or both negative.

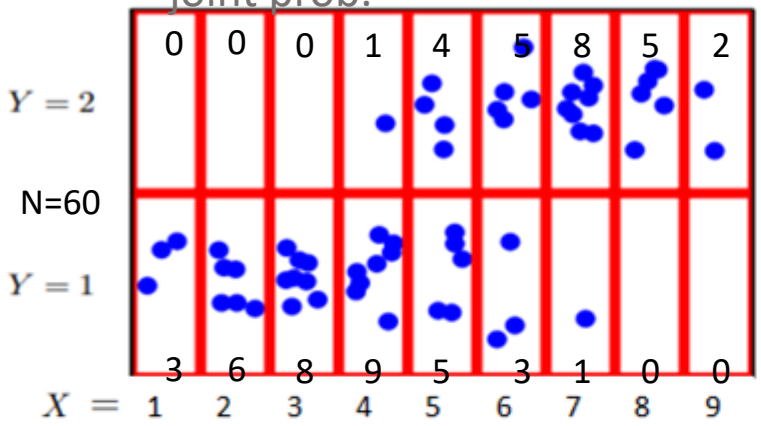
Chapter-2 Contents

1. Joint, marginal, conditional probability distribution
2. Three Laws of Probability
3. Expectation of f(x) where x has PDF p(x)
4. Point Estimates: Max Likelihood (MLE) vs Max Posterior (MAPE)
5. Covariance between bi-variates (x, y) as matrix operation
6. Covariance among multi-variates
7. Summary: Probability Densities, expectation, variance, covariance, correlation

2.1 Joint, marginal, conditional probability distribution

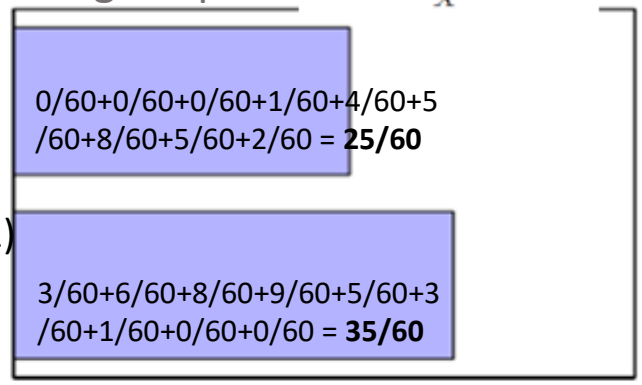
- Multivariate data – joint probability distribution e.g., p(x,y). (e.g., used for anomaly detection)
- Integrate the joint probability distribution to get marginal probability distribution e.g., p(x)

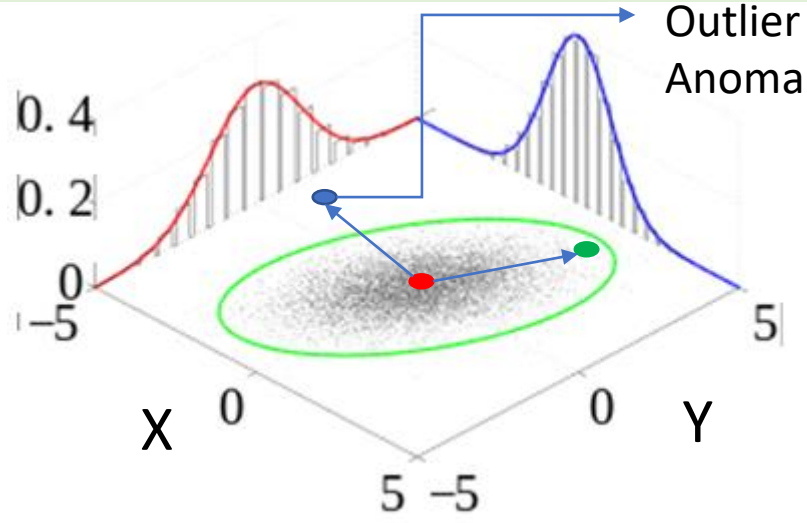
joint prob. $p(X,Y)$



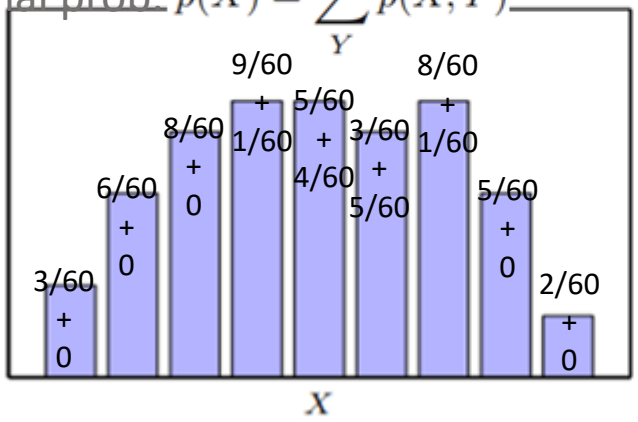
Note:
p(x=9) > 0
p(y=1) > 0
but
p(x=9, y=1) = 0 !

marginal prob. $p(Y) = \sum_X p(X,Y)$

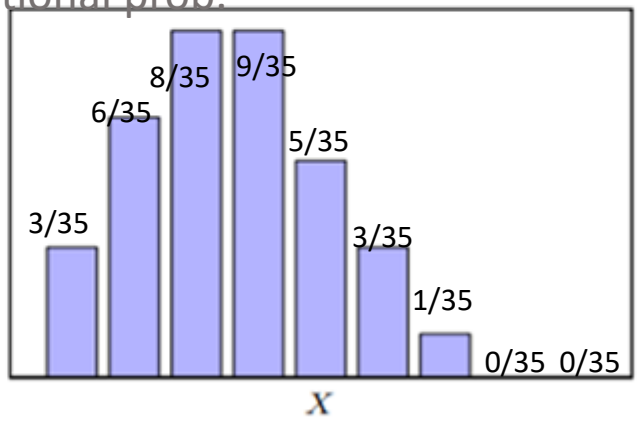




marginal prob. $p(X) = \sum_Y p(X,Y)$



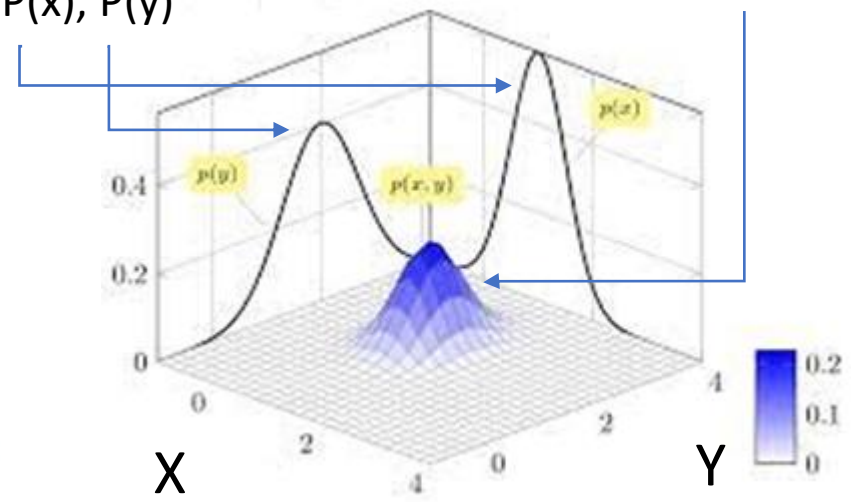
conditional prob. $p(X|Y=1)$



An illustration of a distribution over two variables, X, which takes 9 possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y=1)$ corresponding to the bottom row in the top left figure.

Marginal Prob. Distbn. P(x), P(y)

Joint Prob. Distbn. P(x,y)



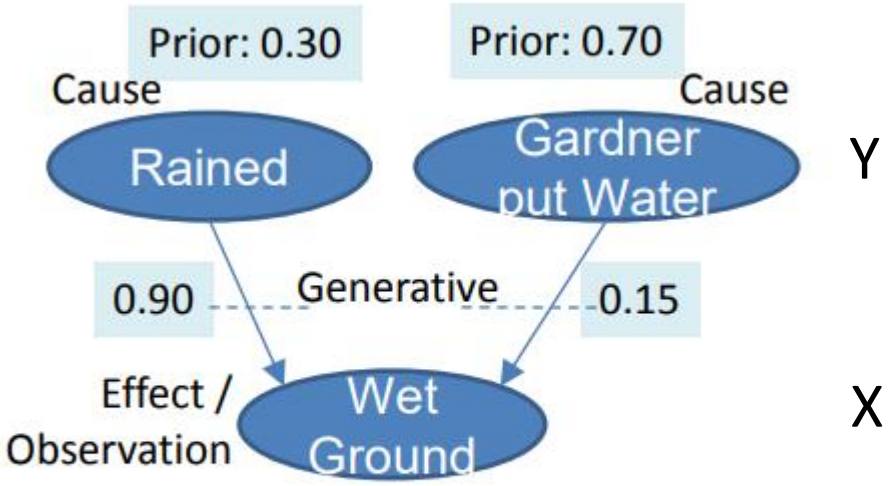
2.2 Three Laws of Probability

- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y)\, p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad p(X) = \sum_Y p(X|Y)p(Y)$$



'Wet Ground' is a weak entity. Its probability can be accounted only upon conditioning over the independent entities.

$$P(rain) = 0.3,\ P(gardener) = 0.7$$
$$P(wet \mid rain) = 0.9,\ P(wet \mid gardener) = 0.15$$
$$P(rain \mid wet) = ?$$

Baye's Theorem

Prior : $P(Y)$

Generative : $P(X|Y)$ – *this is given.*

Predictive : $P(Y|X)$ – *need to estimate this*

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(Y)\, P(X|Y)}{P(X)} = \frac{P(Y)\, P(X|Y)}{\sum_Y P(X, Y)} = \frac{P(Y)\, P(X|Y)}{\sum_Y P(Y)\, P(X|Y)}$$

$$Hence\ P(Y|X) = \frac{P(Y)\, P(X|Y)}{\sum_Y P(Y)\, P(X|Y)}$$

Beauty *of Baye's Theorem is that :*
future *(Prediction) is modeled through* ***past*** *(prior, generative)*

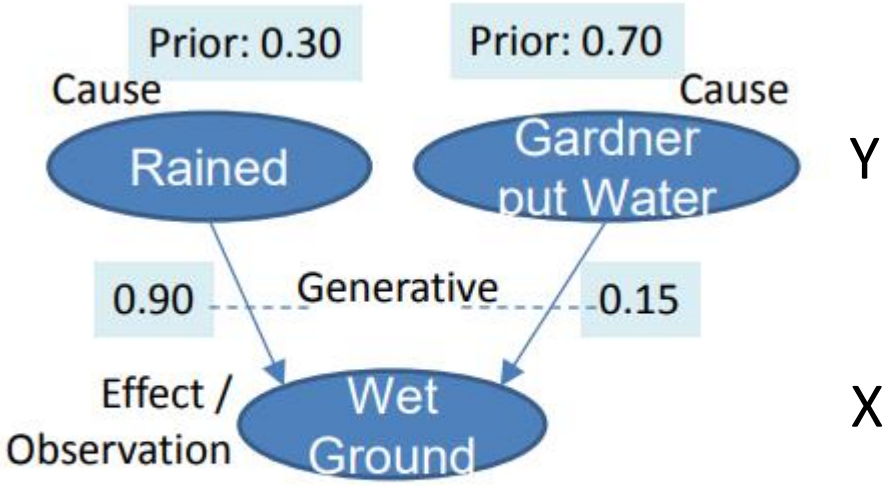
2.2 Three Laws of Probability: question-1

- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y)\, p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad p(X) = \sum_Y p(X|Y)p(Y)$$



'Wet Ground' is a weak entity. Its probability can be accounted only upon conditioning over the independent entities.

$$P(rain) = 0.3,\ P(gardener) = 0.7$$
$$P(wet \mid rain) = 0.9,\ P(wet|\ gardener) = 0.15$$
$$P(rain|\ wet) = ?$$

$$P(rain|\ wet) = \frac{P(rain)\ P(wet|rain)}{P(rain)\ P(wet|rain) + P(gardener)\ P(wet|gardener)}$$
$$= \frac{0.3 * 0.9}{0.3 * 0.9 + 0.7 * 0.15} = 0.72$$

Baye's Theorem

Prior : $P(Y)$ – *this is given.*
Generative : $P(X|Y)$ – *this is given.*
Predictive : $P(Y|X)$ – *need to estimate this*

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(Y)\ P(X|Y)}{P(X)} = \frac{P(Y)\ P(X|Y)}{\sum_Y P(X, Y)} = \frac{P(Y)\ P(X|Y)}{\sum_Y P(Y)\ P(X|Y)}$$

$$Hence\ P(Y|X) = \frac{P(Y)\ P(X|Y)}{\sum_Y P(Y)\ P(X|Y)}$$

Beauty *of Baye's Theorem is that :*
future *(Prediction) is modeled through* ***past*** *(prior, generative)*

For weak/dependent entities we can only derive marginal probabilities (sum rule)

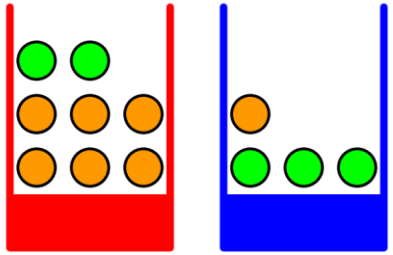
2.2 Three Laws of Probability : question-2

- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y)\, p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad p(X) = \sum_Y p(X|Y)p(Y)$$



Total #apple = 5
Total #Fruits = 12
Hence p(apple) = 5/12 ?
– No, this is **Wrong**

Given a red and blue box. Red box has 2 apples & 6 oranges. Blue box has 3 apples and 1 orange. Your eyes are closed and you have to pick any fruit. Blue box is bigger in size hance probability of selecting blue-box is 0.6. **Problem**:

i. What is probability to pick an apple ?
ii. If an orange is picked, what is prob. if the box is red ?

Baye's Theorem

Prior : $P(Y)$ – *this is given.*
Generative : $P(X|Y)$ – *this is given.*
Predictive : $P(Y|X)$ – *need to estimate this*

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(Y)\, P(X|Y)}{P(X)} = \frac{P(Y)\, P(X|Y)}{\sum_Y P(X, Y)} = \frac{P(Y)\, P(X|Y)}{\sum_Y P(Y)\, P(X|Y)}$$

$$Hence\ P(Y|X) = \frac{P(Y)\, P(X|Y)}{\sum_Y P(Y)\, P(X|Y)}$$

Beauty *of Baye's Theorem is that :*
future *(Prediction) is modeled through* ***past*** *(prior, generative)*

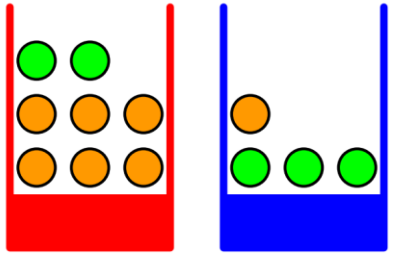
2.2 Three Laws of Probability : question-2

- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y)\, p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad p(X) = \sum_Y p(X|Y)p(Y)$$



Total #apple = 5
Total #Fruits = 12
Hence p(apple) = 5/12 ?
– No, this is **Wrong**

Given a red and blue box. Red box has 2 apples & 6 oranges. Blue box has 3 apples and 1 orange. Your eyes are closed and you have to pick any fruit. Blue box is bigger in size hance probability of selecting blue-box is 0.6. **Problem**:

i. What is probability to pick an apple ?
ii. If an orange is picked, what is prob. if the box is red ?

Random Vars

B (Box):
r = red
b = blue

F (Fruit):
a = apple (green color)
o = orange (orange color)

Given

$$p(B = r) = 4/10$$
$$p(B = b) = 6/10$$
$$p(F = a|B = r) = 1/4$$
$$p(F = o|B = r) = 3/4$$
$$p(F = a|B = b) = 3/4$$
$$p(F = o|B = b) = 1/4.$$

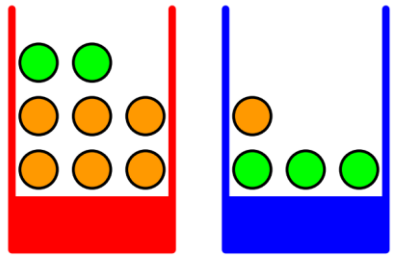
2.2 Three Laws of Probability : question-2

- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y)\, p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad p(X) = \sum_Y p(X|Y)p(Y)$$



Total #apple = 5
Total #Fruits = 12
Hence p(apple) = 5/12 ?
– No, this is **Wrong**

Given a red and blue box. Red box has 2 apples & 6 oranges. Blue box has 3 apples and 1 orange. Your eyes are closed and you have to pick any fruit. Blue box is bigger in size hance probability of selecting blue-box is 0.6. **Problem**:

i. What is probability to pick an apple ?
ii. If an orange is picked, what is prob. if the box is red ?

Random Vars

B (Box):
r = red
b = blue

F (Fruit):
a = apple (green color)
o = orange (orange color)

Given

$$p(B = r) = 4/10$$
$$p(B = b) = 6/10$$
$$p(F = a|B = r) = 1/4$$
$$p(F = o|B = r) = 3/4$$
$$p(F = a|B = b) = 3/4$$
$$p(F = o|B = b) = 1/4.$$

P (F = a) = ?

$$p(F = a) = p(F = a, B = r) + p(F = a, B = b) \quad \text{: Sum Rule}$$

$$p(F = a) = p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b) \quad \text{: Product Rule}$$

$$= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} = 0.55$$

P (F = o) = 1 – P (F = a) = 9/20 = 0.45

$$p(x) = \int p(x, y)\, \mathrm{d}y \quad \text{If continuous vars}$$

P (B = r | F = o) = ?

: Baye's Theorem

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$

$$= \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}.$$

If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $p(B)$. We call this the *prior probability* because it is the probability available *before* we observe the identity of the fruit. Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $p(B|F)$, which we shall call the *posterior probability* because it is the probability obtained *after* we have observed F.

Posterior P(B=r|F=o) is higher even when prior P(B=b) is higher!

2.3 Expectation of f(x) where x has PDF p(x)

$$E[f(x)] = \sum_x f(x)\,p(x) \quad or \quad \int_x f(x)\,p(x)\,dx$$

$$if\ f(x) = x \rightarrow E[x] = \int_x x\,p(x)\,dx$$

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n). \quad x\ sampled\ from\ p(x)$$

A person can walk max. up to 10 km. Probability that a person walks a distance of x km. is proportional to x. What is the expected distance a person walks ?

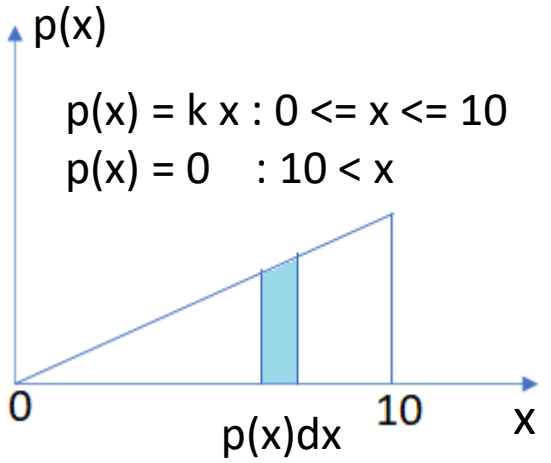
2.3 Expectation of f(x) where x has PDF p(x): question-1

$$E[f(x)] = \sum_x f(x)\,p(x)\,dx \quad or \quad \int_x f(x)\,p(x)\,dx$$

$$if\ f(x) = x \rightarrow E[x] = \int_x x\,p(x)\,dx$$

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n). \quad x\ sampled\ from\ p(x)$$

A person can walk max. up to 10 km. Probability that a person walks a distance of x km is proportional to x.
What is the expected distance a person walks ?



$$p(x) = k\,x$$

$$\int_0^{10} p(x)\,dx = 1$$

$$i.e. \quad \int_0^{10} k\,x\,dx = 1$$

$$i.e. \quad \left[\frac{k\,x^2}{2}\right]_0^{10} = 1$$

$$i.e. \quad k\,[100 - 0] = 2$$

$$hence \quad k = 0.02$$

$$E(x) = \int_0^{10} x\,p(x)\,dx$$

$$= \int_0^{10} x\,(k\,x)\,dx$$

$$= 0.02 \int_0^{10} x^2\,dx$$

$$= 0.02 \left[\frac{x^3}{3}\right]_0^{10}$$

$$= \frac{0.02 * 1000}{3}$$

$$= 6.67\ km\ !$$

2.4 Point Estimates: Max Likelihood (MLE) vs Max Posterior (MAPE) Estimate

- MLE: An estimate which maximizes the likelihood of the observation.
- MAPE: Models the posterior probability distribution.

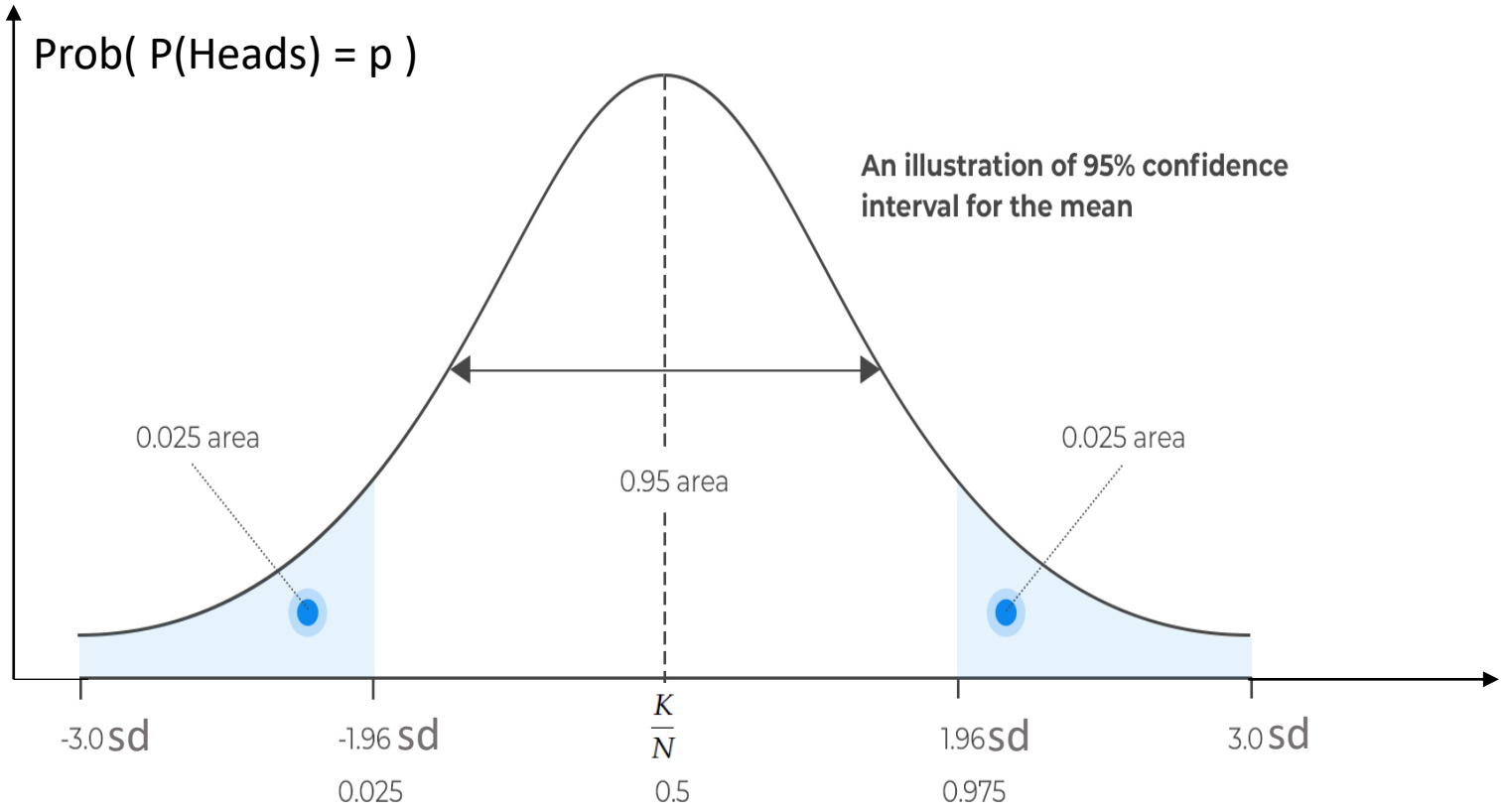
$$E[f(x)] = \sum_x f(x)\, p(x)\, dx \quad or \quad \int_x f(x)\, p(x)\, dx$$

Prior info : *A coin can show up Heads with probability anywhere between* $[0, 1]$.

Event E : *out of N trials K heads occur.*

Question : *what is P(Head) now ?*

$$Max.\ Likelihood: P(Head) = \frac{K}{N}$$



But can you predict a probability distribution for P(Heads) and suggest confidence intervals ?

2.4 Point Estimates: Max Likelihood (MLE) vs Max Posterior (MAPE) Estimate: question

- MLE: An estimate which maximizes the likelihood of the observation.
- MAPE: Models the posterior probability distribution.

$$E[f(x)] = \sum_x f(x)\, p(x)\, dx \quad or \quad \int_x f(x)\, p(x)\, dx$$

Prior info : *A coin can show up Heads with probability anywhere between* $[0, 1]$.

Event E : *out of N trials K heads occur.*

Question : *what is P(Head) now ?*

$$Max.\ Likelihood: P(Head) = \frac{K}{N}$$

Let $\mathbf{p}$ *be a randon var. that is P(Head). Let* $f()$ *represent a prob. density func.*

$$P(E|\mathbf{p} = p) = p^K (1-p)^{(n-K)}$$

$$P(E) = \int_0^1 P(E|\mathbf{p} = p)\, f(p)\, dp$$

$$P(E) = \int_0^1 p^K (1-p)^{(n-K)}\, dp = \frac{(n-k)!\, k!}{(n+1)!}$$

By Baye's Theorem :

$$f(p|E) = \frac{f(p)\, P(E|\mathbf{p} = p)}{P(E)}$$

MAPE

$$f(p|E) = \frac{(n+1)!\, p^K (1-p)^{(n-K)}}{(n-k)!\, k!}$$

– *MAPE is considered more robust because it incorporates prior information*
– *Considering a probability distribution is more robust than a point estimate*

MAPE: the point p at which this is maximized. In this case MAPE estimate is same as MLE: K/N

2.5 Covariance between bi-variates (x, y) as matrix operation

Covariance between variables (x, y): $COV(x,y) = \frac{[x-\bar{x}]^T[y-\bar{y}]}{n}$ and variance : $VAR(x) = \frac{[x-\bar{x}]^T[x-\bar{x}]}{n}$

X	Y
X1	Y1
X2	Y2
X3	Y3
…	…
Xn	Yn

$$COV(x,y) = \frac{[x-\bar{x}]^T[y-\bar{y}]}{n} = \frac{1}{n}[x_1-\bar{x},\ x_2-\bar{x},\ \ldots,\ x_n-\bar{x}]\begin{bmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \ldots \\ y_n-\bar{y} \end{bmatrix}$$

$$= \frac{(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \ldots + (x_n-\bar{x})(y_n-\bar{y})}{n}$$

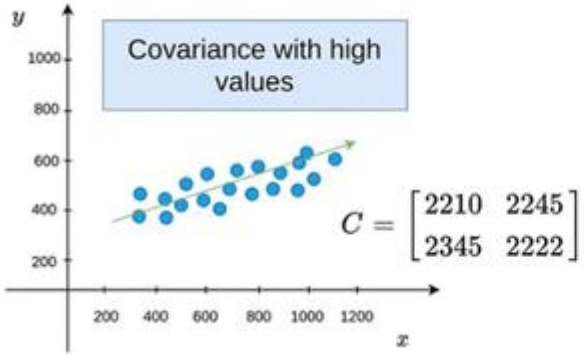
similarly

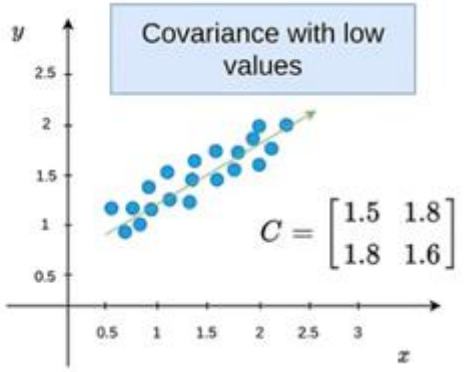
$$VAR(x) = \frac{[x-\bar{x}]^T[x-\bar{x}]}{n} = \frac{1}{n}[x_1-\bar{x},\ x_2-\bar{x},\ \ldots,\ x_n-\bar{x}]\begin{bmatrix} x_1-\bar{x} \\ x_2-\bar{x} \\ \ldots \\ x_n-\bar{x} \end{bmatrix}$$

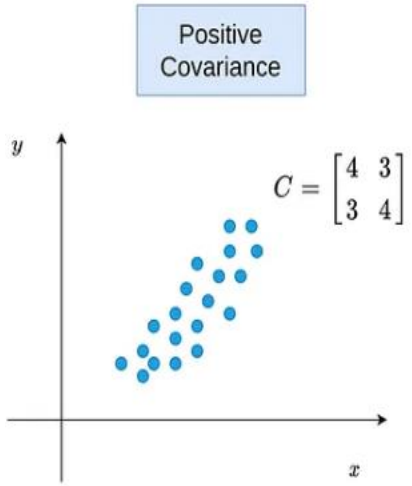
$$= \frac{(x_1-\bar{x})(x_1-\bar{x}) + (x_2-\bar{x})(x_2-\bar{x}) + \ldots + (x_n-\bar{x})(x_n-\bar{x})}{n}$$

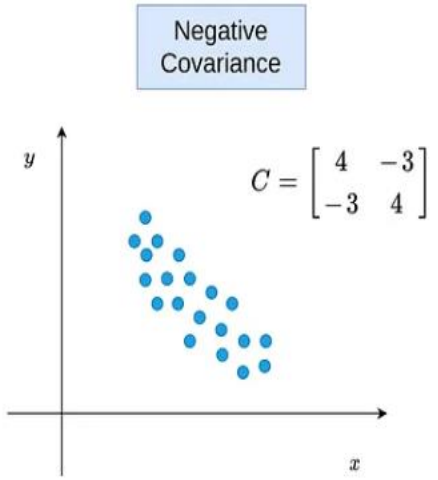
Covariance Matrix =

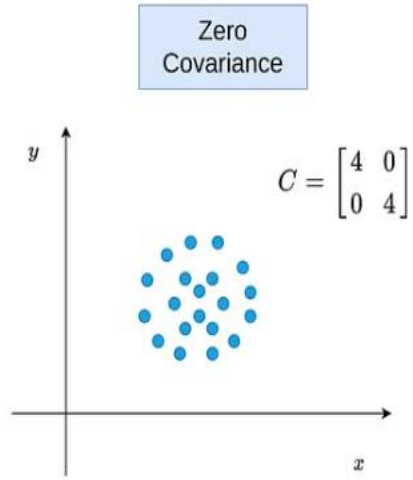
$$\begin{matrix} x \\ y \\ z \end{matrix}\begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix}$$

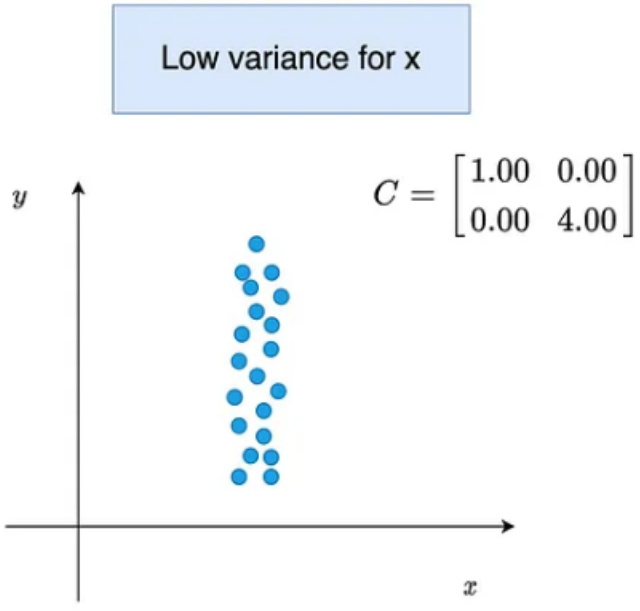


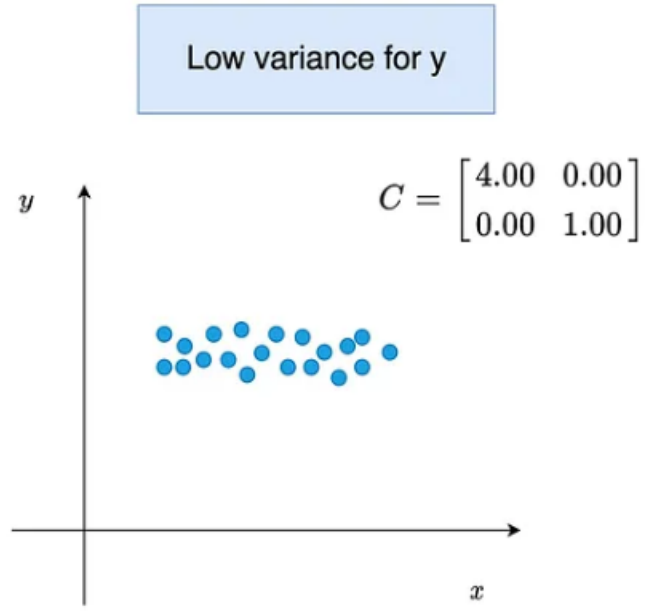












2.6 Covariance among multi-variates say W = [x, y, z] is $Cov\,(W) = [W-\overline{W}]^T N^{-1} [W-\overline{W}]$

$$W = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_1 & z_2 \\ .. & .. & .. \\ x_n & y_n & z_n \end{bmatrix} \quad W-\overline{W} = \begin{bmatrix} x_1-\bar{x} & y_1-\bar{y} & z_1-\bar{z} \\ x_2-\bar{x} & y_2-\bar{y} & z_2-\bar{z} \\ .. & .. & .. \\ x_n-\bar{x} & y_n-\bar{y} & z_n-\bar{z} \end{bmatrix} \quad N^{-1}_{\substack{(n\times n)\\ diagonal\\ matrix}} = \begin{bmatrix} 1/n & 0 & 0.. & 0 \\ 0 & 1/n & 0.. & 0 \\ 0.. & 0.. & 1/n.. & 0.. \\ 0 & 0 & 0.. & 1/n \end{bmatrix}$$

for population, but if for a sample then

$$N^{-1}_{\substack{(n\times n)\\ diagonal\\ matrix}} = \begin{bmatrix} 1/(n-1) & 0 & 0.. & 0 \\ 0 & 1/(n-1) & 0.. & 0 \\ 0.. & 0.. & 1/(n-1).. & 0.. \\ 0 & 0 & 0.. & 1/(n-1) \end{bmatrix}$$

then $Cov\,(W) = [W-\overline{W}]^T N^{-1} [W-\overline{W}]$

$$= \begin{bmatrix} x_1-\bar{x} & x_2-\bar{x} & .. & x_n-\bar{x} \\ y_1-\bar{y} & y_2-\bar{y} & .. & y_n-\bar{y} \\ z_1-\bar{z} & z_2-\bar{z} & .. & z_n-\bar{z} \end{bmatrix} \begin{bmatrix} 1/n & 0 & 0.. & 0 \\ 0 & 1/n & 0.. & 0 \\ 0.. & 0.. & 1/n.. & 0.. \\ 0 & 0 & 0.. & 1/n \end{bmatrix} \begin{bmatrix} x_1-\bar{x} & y_1-\bar{y} & z_1-\bar{z} \\ x_2-\bar{x} & y_2-\bar{y} & z_2-\bar{z} \\ .. & .. & .. \\ x_n-\bar{x} & y_n-\bar{y} & z_n-\bar{z} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{(x_1-\bar{x})}{n} & \frac{(x_2-\bar{x})}{n} & .. & \frac{(x_n-\bar{x})}{n} \\ \frac{(y_1-\bar{y})}{n} & \frac{(y_2-\bar{y})}{n} & .. & \frac{(y_n-\bar{y})}{n} \\ \frac{(z_1-\bar{z})}{n} & \frac{(z_2-\bar{z})}{n} & .. & \frac{(z_n-\bar{z})}{n} \end{bmatrix} \begin{bmatrix} x_1-\bar{x} & y_1-\bar{y} & z_1-\bar{z} \\ x_2-\bar{x} & y_2-\bar{y} & z_2-\bar{z} \\ .. & .. & .. \\ x_n-\bar{x} & y_n-\bar{y} & z_n-\bar{z} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{[x-\bar{x}]^T[x-\bar{x}]}{n} & \frac{[x-\bar{x}]^T[y-\bar{y}]}{n} & \frac{[x-\bar{x}]^T[z-\bar{z}]}{n} \\ \frac{[y-\bar{y}]^T[x-\bar{x}]}{n} & \frac{[y-\bar{y}]^T[y-\bar{y}]}{n} & \frac{[y-\bar{y}]^T[z-\bar{z}]}{n} \\ \frac{[z-\bar{z}]^T[x-\bar{x}]}{n} & \frac{[z-\bar{z}]^T[y-\bar{y}]}{n} & \frac{[z-\bar{z}]^T[z-\bar{z}]}{n} \end{bmatrix} = \begin{bmatrix} var\,(x) & cov\,(x,y) & cov\,(x,z) \\ cov\,(y,x) & var\,(y) & cov\,(y,z) \\ cov\,(z,x) & cov\,(z,y) & var\,(z) \end{bmatrix} = \text{Cov (W)}$$

if $W = [x, y, z, \ldots]$ *be a multivariate*

Let $W_{centered} = [(x-\bar{x})\ (y-\bar{y})\ (z-\bar{z}) \ldots]$

then $COV\,(W) = W^T_{centered}\, N^{-1}\, W_{centered}$

Let $W_{norm} = \left[\frac{(x-\bar{x})}{S_x}\ \frac{(y-\bar{y})}{S_y}\ \frac{(z-\bar{z})}{S_z}\ \ldots\right]$

where $S_x\ S_y\ S_z \ldots$ *are the Std. dev of vars x y z ...*

then $CORR\,(W) = W^T_{norm}\, N^{-1}\, W_{norm}$

2.7 Summary: Probability Densities, expectation, variance, covariance, correlation

Important to get a hang of Expectation, Variance, Covariance, Correlation

Probability distribution/density properties

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x. \quad \text{or} \quad \sum_{a<x\leqslant b} p(x)$$

if x is continuous if x is discrete

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1. \text{ or } \sum_x p(x) = 1$$

$$p(x) = \int p(x,y)\,\mathrm{d}y \quad \text{or} \quad p(x) = \sum_y p(x,y)$$
$$p(x,y) = p(y|x)p(x).$$

Expected value of f(x)

$$E[f(x)] = \sum_x f(x)\,p(x) \quad or \quad \int_x f(x)\,p(x)\,dx$$

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n). \quad x \text{ sampled from } p(x)$$

$$E[X] = \sum_i x_i P\{X = x_i\}$$

Variance of f(x)

If x independent y: $E[XY] = E[X]E[Y]$.

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2.$$
$$\mathrm{var}[x] = E\left[(x - E[x])^2\right] = \mathbb{E}[x^2] - \mathbb{E}[x]^2.$$

$$E[aX + b] = aE[X] + b$$

$$E[X + Y] = E[X] + E[Y]$$

$$avg(x - \bar{x}) = 0 \text{ or } E(x - E(x)) = 0$$
$$\text{or } \sum_x (x - E[x])p(x) = 0$$

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

for independent $X_1, \ldots, X_n$, $\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i)$

Covariance

$$\mathrm{Cov}(X,Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y]$$

$$cov(aX,\ bY) = ab\ cov(X,\ Y)$$

If X and Y are independent random variables, then

$$\mathrm{Cov}(X,Y) = 0$$

Population Covariance Formula

$$\mathbf{Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}}$$

Sample Covariance

$$\mathbf{Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - y)}{N - 1}}$$

Correlation

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

sample correlation

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

population correlation

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n\ s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Chapter-3 Contents

1. Special Random Variables: Uniform and Normal RVs
2. Uniform Random Variable: Max. Likelihood Estimate (MLE) example
3. Special Random Variables: Chi-Square, t-Distribution, F-Distribution
4. Distributions of Sampling Statistics

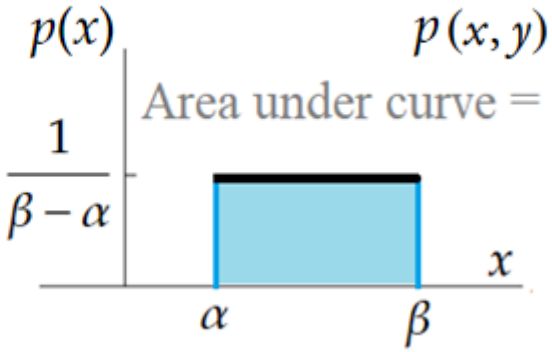
3.1 Special Random Variables: Uniform and Normal RVs

- Define Randomness: randomness is a probability distribution (think histogram) e.g.:
- Uniform randomness, Gaussian/Normal Randomness, etc..

THE UNIFORM RANDOM VARIABLE

A random variable X is said to be uniformly distributed over the interval $[\alpha, \beta]$ if its probability density function is given by

$$p(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases} \qquad \textit{Note} \quad \sum_{\alpha}^{\beta} p(x) \text{ or } \int_{\alpha}^{\beta} p(x)\,dx = \frac{1}{\beta - \alpha}\int_{\alpha}^{\beta} dx = 1$$



$$p(x,y) = \begin{cases} c & \text{if } (x,y) \in R \\ 0 & \text{if otherwise} \end{cases}$$

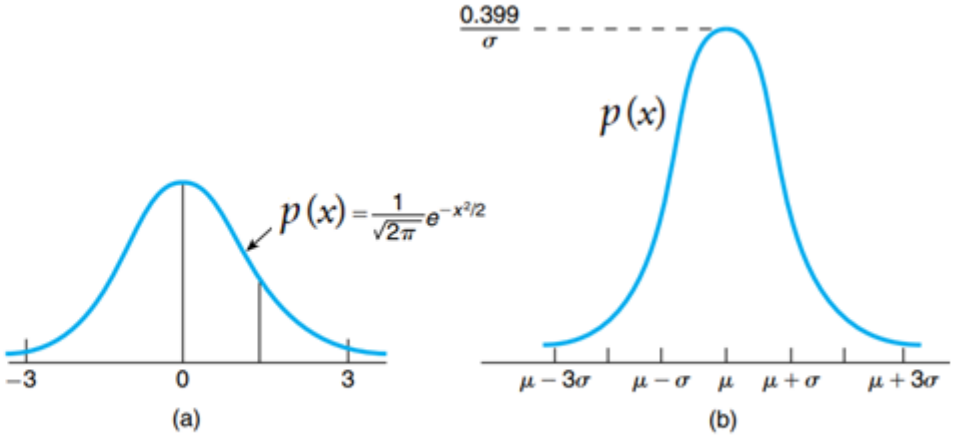
Joint uniform distribution

NORMAL RANDOM VARIABLES

A random variable is said to be normally distributed with parameters μ and σ^2, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty^*$$

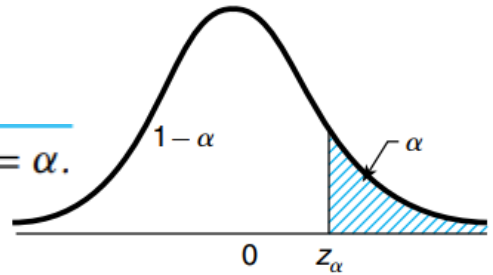
The normal density $p(x)$ is a bell-shaped curve that is symmetric about μ and that attains its maximum value of $\frac{1}{\sqrt{2\pi}\,\sigma} \approx 0.399/\sigma$ at $x = \mu$



The normal density function (a) with $\mu = 0, \sigma = 1$ and (b) with arbitrary μ and σ^2.

if $X \sim N(\mu, \sigma^2)$ *then* $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$ *is* **Std. Normal**

$P\{Z > z_\alpha\} = \alpha.$



3.2 Uniform Random Variable: Max. Likelihood Estimate (MLE)

What is some param (e.g. mean) θ of a given population distribution when some samples from that population are observed ?

Maximum Likelihood Estimate of θ = that value which maximizes the likelihood of the observations.

$$Given: \ Samples \ X_1, \ X_2, \ ....., \ X_n \ from \ a \ uniformly \ distributed \ population \ over \ (0, \ \theta)$$

$$Question: \ What \ is \ the \ population \ mean \ ?$$

3.2 Uniform Random Variable: Max. Likelihood Estimate (MLE): question-1

What is some param (e.g. mean) θ of a given population distribution when some samples from that population are observed ?

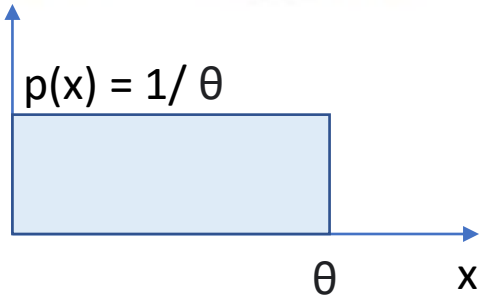
Maximum Likelihood Estimate of θ = that value which maximizes the likelihood of the observations.

Given : Samples X_1, X_2, ..., X_n from a uniformly distributed population over $(0, \theta)$

Question : What is the population mean ?

Estimating the Mean of a Uniform Distribution Suppose $X_1, \ldots, X_n$ constitute a sample from a uniform distribution on $(0, \theta)$, where θ is unknown. Their joint density is thus

$$f(x_1, x_2, \ldots, x_n|\theta) = \begin{cases} \frac{1}{\theta^n} & 0 < x_i < \theta, \quad i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$



This density is maximized by choosing θ as small as possible. Since θ must be at least as large as all of the observed values x_i, it follows that the smallest possible choice of θ is equal to $\max(x_1, x_2, \ldots, x_n)$. Hence, the maximum likelihood estimator of θ is

$$\hat{\theta} = \max(X_1, X_2, \ldots, X_n) \qquad \text{mean} = \frac{\hat{\theta}}{2} = \frac{\max(X_1, X_2, \ldots, X_n)}{2}$$

3.3 Special Random Variables: Chi-Square, t-Distribution, F-Distribution

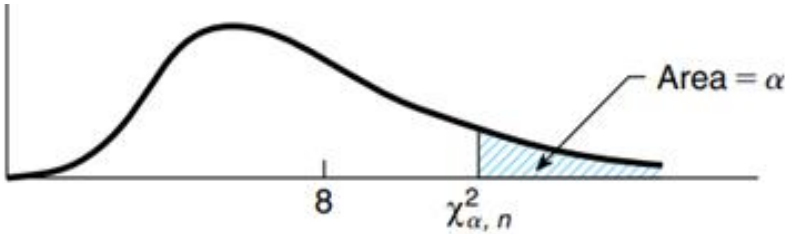
Some distributions are 1-sided (x>=0), some are 2-sided

The Chi-Square Distribution

If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables,

then X, defined by $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ $\quad X \sim \chi_n^2$

is said to have a *chi-square distribution with n degrees of freedom.*



$P\{X \geq \chi^2_{\alpha,n}\} = \alpha$

The chi-square density function with 8 degrees of freedom.

The t-Distribution

If Z and χ_n^2 are independent random variables, with Z having a standard normal distribution and χ_n^2 having a chi-square distribution with n degrees of freedom, then the random variable T_n defined by

$$T_n = \frac{Z}{\sqrt{\chi_n^2/n}} \qquad \text{where} \qquad \frac{\chi_n^2}{n} = \frac{Z_1^2 + \cdots + Z_n^2}{n}$$

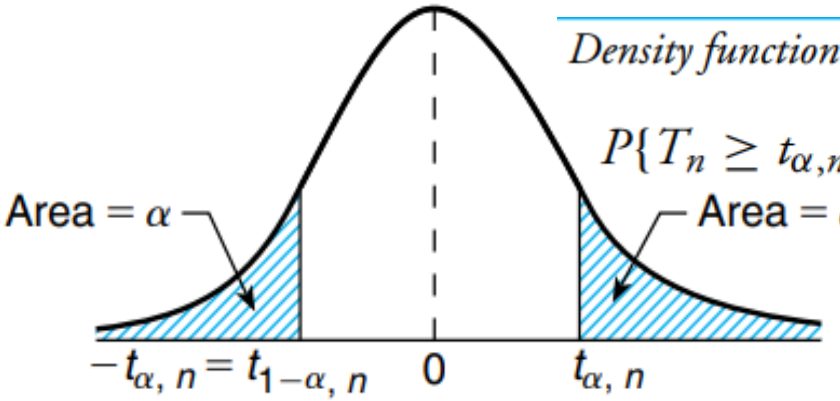
$$E[T_n] = 0, \qquad n > 1$$

$$\text{Var}(T_n) = \frac{n}{n-2}, \qquad n > 2$$

is said to have a *t-distribution with n degrees of freedom.*

Density function of T_n.

$P\{T_n \geq t_{\alpha,n}\} = \alpha$



The F-Distribution

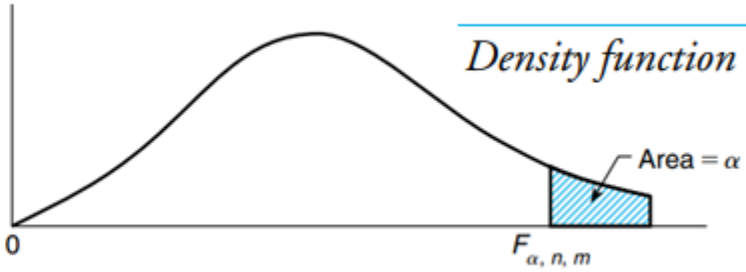
If χ_n^2 and χ_m^2 are independent chi-square random variables with n and m degrees of freedom, respectively, then the random variable $F_{n,m}$ defined by

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$$

is said to have an *F-distribution with n and m degrees of freedom.*

For any $\alpha \in (0, 1)$, let $F_{\alpha,n,m}$ be such that $P\{F_{n,m} > F_{\alpha,n,m}\} = \alpha$

Density function of $F_{n,m}$.



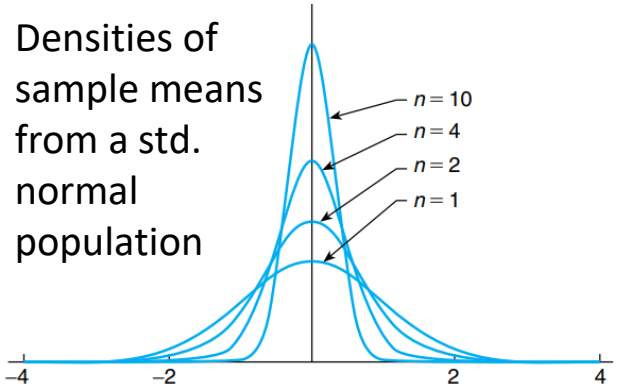
3.4 Distributions of Sampling Statistics

Sample mean is a point estimate of the population mean. Sample mean is a random variable

Properties of Mean

The quantities μ and σ^2 are called the *population mean* and the *population variance*, respectively. Let $X_1, X_2, \ldots, X_n$ be a sample of values from this population. The sample mean is defined by $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ $\quad E(\bar{x}) = \mu \quad Var(\bar{x}) = \dfrac{\sigma^2}{n}$

Densities of sample means from a std. normal population



The Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of $X_1 + \cdots + X_n$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

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

A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least **30**.

Properties of Variance

If $X_1, \ldots, X_n$ is a sample from a normal population having mean μ and variance σ^2, then $\overline{X}$ and S^2 are independent random variables, with $\overline{X}$ being normal with mean μ and variance σ^2/n and $(n-1)S^2/\sigma^2$ being chi-square with $n-1$ degrees of freedom.

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \qquad \sqrt{n}\frac{(\overline{X} - \mu)}{S} \sim t_{n-1}$$

Binomial Distribution

The probability mass function of a binomial random variable with parameters n and p

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n \qquad \text{where } \binom{n}{i} = n!/[i!(n-i)!]$$

$$E[X] = np \quad \text{and} \quad SD(X) = \sqrt{np(1-p)}$$

$$E(\bar{x}) = np$$

$$\text{Var}(\overline{X}) = \text{Var}(X)/n^2 = p(1-p)/n$$

$$\bar{x} \sim N\left(np, \frac{p(1-p)}{n}\right)$$

Chapter-4 Contents

1. Interval Estimate (Confidence Interval) for population mean
2. Point Estimate (MLE) vs Interval Estimate (Confidence Interval)
3. Two-sided and one-sided Confidence Interval
4. Distributions of Sampling Statistics
5. Parameter Estimation
6. Hypothesis Testing

4.1 Interval Estimate (Confidence Interval) for population mean

If we do sampling multiple times, each time the sample mean maybe different. This tells that the sample mean is a random var and 95% times actual population mean will lie within 1.96*Stdev of sample mean.

X_1	X_2	X_2	.....	X_n

.....

X_1	X_2	X_2	.....	X_n

.....

X_1	X_2	X_2	.....	X_n

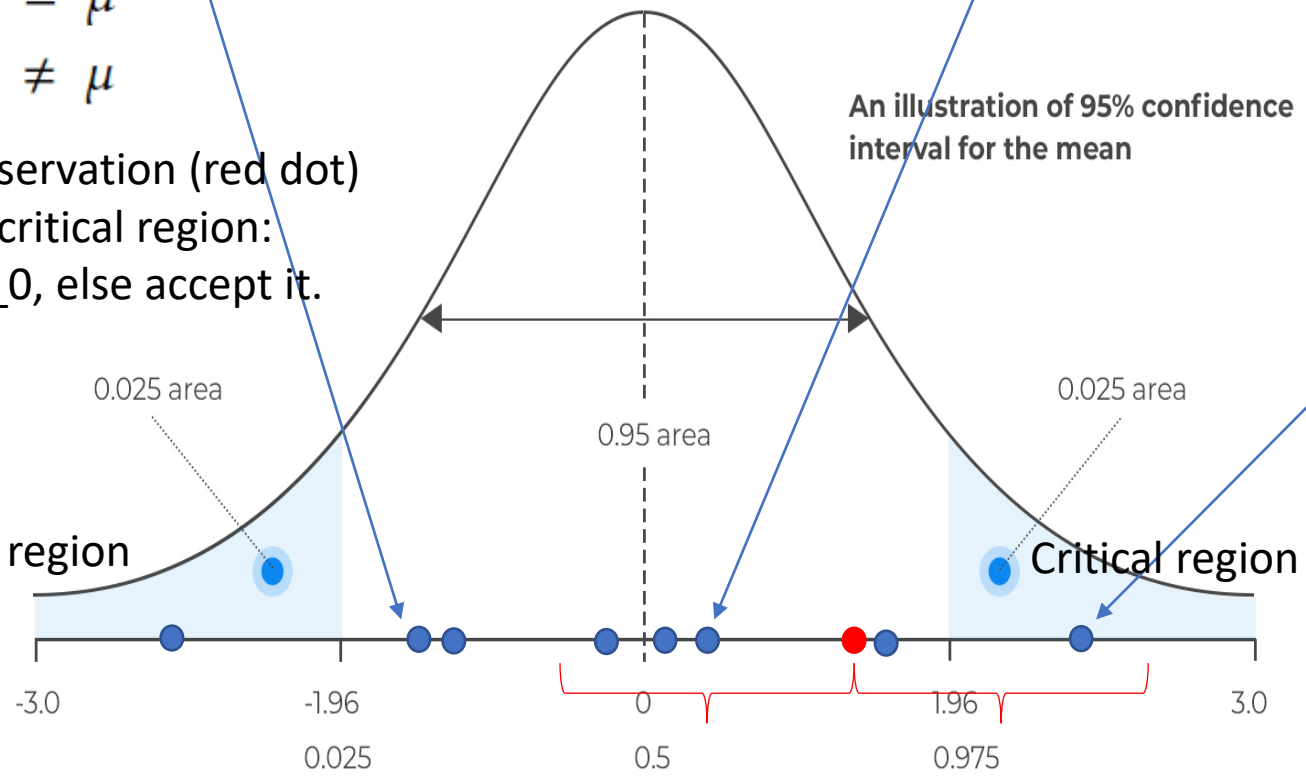
$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim Normal\ (mean = 0,\ stdev = 1)$$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$ Z-score

$$H_0: \bar{x} = \mu$$
$$H_a: \bar{x} \neq \mu$$

If the observation (red dot) is in the critical region: reject H_0, else accept it.



- Suppose the red dot is the only observed value.
- Then the red range is the 95% confidence interval estimate:

$$-1.96 < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1.96$$

or, equivalently,

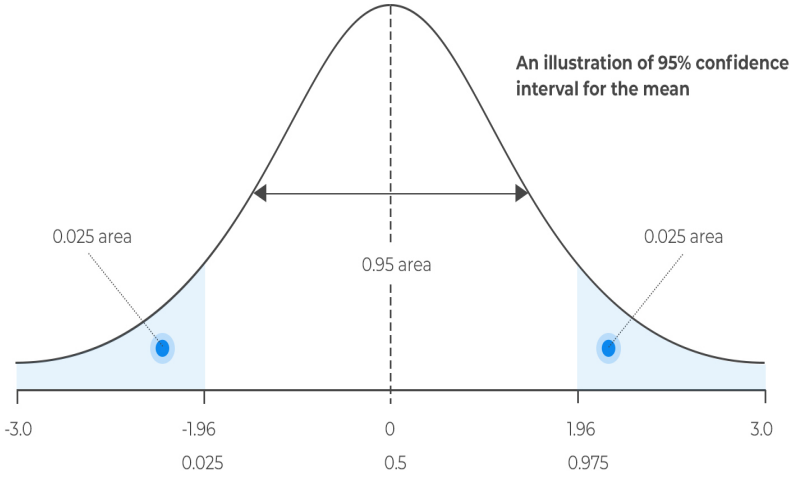
$$P\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

Confidence interval

4.2 Point Estimate (MLE) vs Interval Estimate (Confidence Interval)

Two-sided confidence interval assuming stdev is known.

Suppose that $X_1, \ldots, X_n$ is a sample from a normal population having unknown mean μ and known variance σ^2. It has been shown that $\overline{X} = \sum_{i=1}^{n} X_i/n$ is the maximum likelihood estimator for μ. However, we don't expect that the sample mean $\overline{X}$ will exactly equal μ, but rather that it will "be close." Hence, rather than a point estimate, it is sometimes more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within.



$\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\dfrac{(\overline{X} - \mu)}{\sigma}$ has a standard normal distribution. Therefore,

$$P\left\{-1.96 < \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma} < 1.96\right\} = .95$$

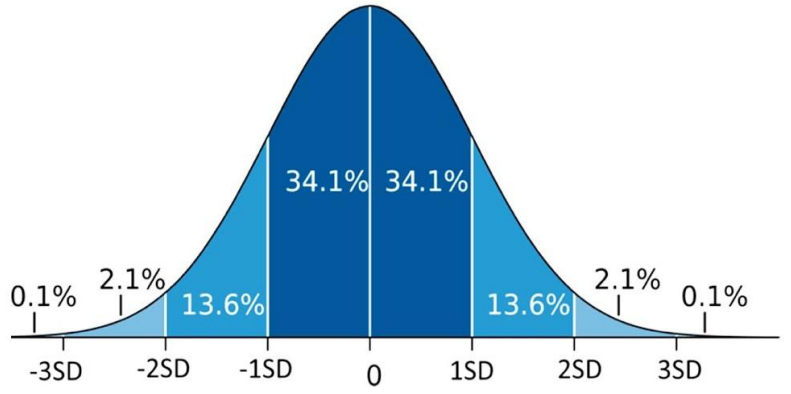
or, equivalently,

$$P\left\{\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = .95$$

That is, 95 percent of the time the value of the sample average $\bar{X}$ will be such that the distance between it and the mean μ will be less than $1.96\sigma/\sqrt{n}$. If we now observe the sample and it turns out that $\overline{X} = \bar{x}$, then we say that "with 95 percent confidence"

That is, "with 95 percent confidence" we assert that the true mean lies within $1.96\sigma/\sqrt{n}$ of the observed sample mean. The interval

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$ is called a *95 percent confidence interval estimate* of μ.



4.3 Two-sided and one-sided Confidence Interval

Two-sided vs. one-sided confidence interval assuming stdev is known.

We can also obtain confidence intervals of any specified level of confidence. To do so, recall that z_α is such that

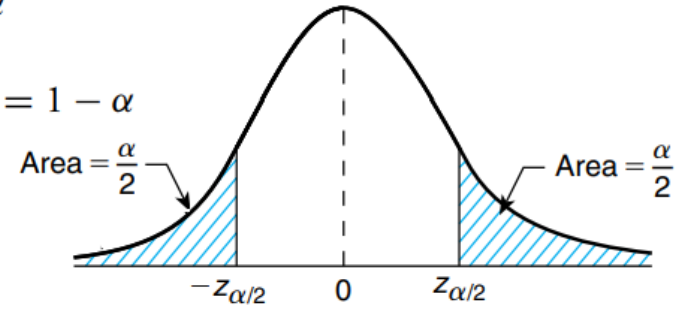
$$P\{Z > z_\alpha\} = \alpha$$

when Z is a standard normal random variable. But this implies (see Figure 7.1) that for any α

$$P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha$$

hence

$$P\left\{-z_{\alpha/2} < \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma} < z_{\alpha/2}\right\} = 1 - \alpha$$



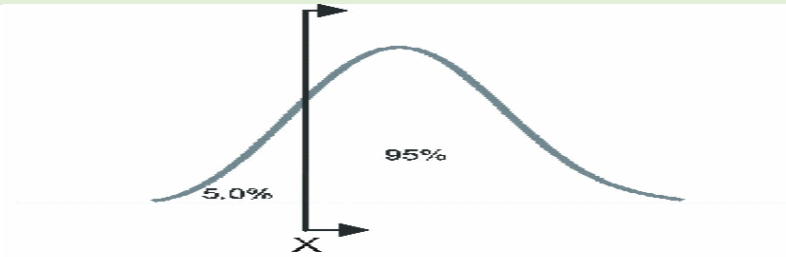
$P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha.$

that is

$$P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

Hence, a $100(1 - \alpha)$ percent two-sided confidence interval for μ is

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$



Similarly, knowing that $Z = \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma}$ is a standard normal random variable, along with the identities

$$P\{Z > z_\alpha\} = \alpha$$

and

$$P\{Z < -z_\alpha\} = \alpha$$

results in one-sided confidence intervals of any desired level of confidence. Specifically, we obtain that

$$\left(\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}, \ \infty\right)$$

and

$$\left(-\infty, \ \bar{x} + z_\alpha\frac{\sigma}{\sqrt{n}}\right)$$

are, respectively, $100(1 - \alpha)$ percent one-sided upper and $100(1 - \alpha)$ percent one-sided lower confidence intervals for μ.

Chapter-5 Contents

1. Hypothesis Testing
2. Known variance: Normal Distbn., Unknown variance: T-Distribution
3. Testing equality of means of two Normal populations
4. Tests comparing the Variance of populations

5.1 Hypothesis Testing

When null hypothesis H_0 is True then observations X_is are very unlikely to be in the critical region C_0

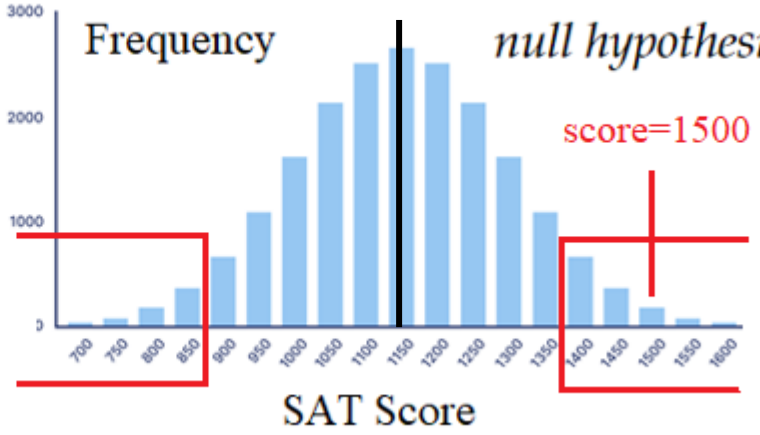
Hence, if X_is in C_0 then H_0 is very likely False ! Usually C_0 is the region of least 0.05 or 5 % probability

null hypothesis

(a) $H_0 : \theta = 1$ **Simple hypothesis**: completely specifies the population distribution

(b) $H_0 : \theta \leq 1$ **Composite hypothesis**: incomplete specification of the distribution

in the figure, 1500 is too far away from the sample mean, and lie in the Critical region. Hence we reject the null hypothesis !



Histogram of SAT scores of a sample of students. Can we estimate the population mean through sample mean ?

Normal distribution with known variance

null hypothesis $H_0 : \mu = \mu_0$ against alternative hypothesis $H_1 : \mu \neq \mu_0$

$$P_{\mu 0}\{|\bar{X} - \mu_0| > c\} = \alpha$$

$$\text{or } P_{\mu 0}\left\{\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} > \frac{c}{\sigma/\sqrt{n}}\right\} = \alpha$$

$$\text{or } P_{\mu 0}\left\{|Z| > \frac{c}{\sigma/\sqrt{n}}\right\} = \alpha$$

$$\text{or } 2\,P_{\mu 0}\left\{Z > \frac{c}{\sigma/\sqrt{n}}\right\} = \alpha$$

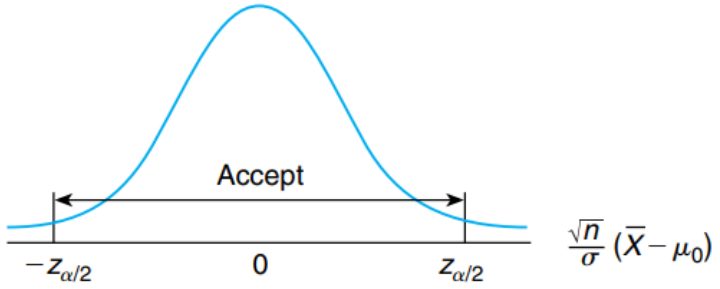
$$\text{or } P_{\mu 0}\left\{Z > \frac{c}{\sigma/\sqrt{n}}\right\} = \alpha/2$$

$$\text{and } P_{\mu 0}\{Z > z_{\alpha/2}\} = \alpha/2$$

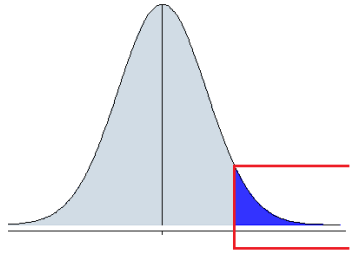
$$\Longrightarrow \frac{c}{\sigma/\sqrt{n}} = z_{\alpha/2} \quad \text{or} \quad c = \frac{z_{\alpha/2}\,\sigma}{\sqrt{n}}$$

reject H_0 if $\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| > z_{\alpha/2}$

accept H_0 if $\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \leq z_{\alpha/2}$



One sided test



$H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$

$$P_{\mu_0}\{\bar{X} - \mu_0 > c\} = \alpha$$

accept H_0 if $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) \leq z_\alpha$

reject H_0 if $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) > z_\alpha$

5.2 Known variance: Normal Distbn., Unknown variance: T-Distribution

- Reject null hypothesis if p-value < α (level of significance)
- Smaller the p-value worse the fit

$X_1, \ldots, X_n$ *Is a Sample from a* $\mathcal{N}(\mu, \sigma^2)$ *Population* $\quad \sigma^2$ *Is Known,* $\overline{X} = \sum_{i=1}^{n} X_i/n$

H_0	H_1	Test Statistic TS	Significance Level α Test	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/\sigma$	Reject if $\lvert TS\rvert > z_{\alpha/2}$	$2P\{Z \geq \lvert t\rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/\sigma$	Reject if $TS > z_\alpha$	$P\{Z \geq t\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/\sigma$	Reject if $TS < -z_\alpha$	$P\{Z \leq t\}$

Z is a standard normal random variable.

$X_1, \ldots, X_n$ *Is a Sample from a* $\mathcal{N}(\mu, \sigma^2)$ *Population*

σ^2 *Is Unknown,* $\overline{X} = \sum_{i=1}^{n} X_i/n \qquad S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$

H_0	H_1	Test Statistic TS	Significance Level α Test	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/S$	Reject if $\lvert TS\rvert > t_{\alpha/2, n-1}$	$2P\{T_{n-1} \geq \lvert t\rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/S$	Reject if $TS > t_{\alpha, n-1}$	$P\{T_{n-1} \geq t\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}(\overline{X} - \mu_0)/S$	Reject if $TS < -t_{\alpha, n-1}$	$P\{T_{n-1} \leq t\}$

T_{n-1} *is a t-random variable with* $n-1$ *degrees of freedom:* $P\{T_{n-1} > t_{\alpha, n-1}\} = \alpha$.

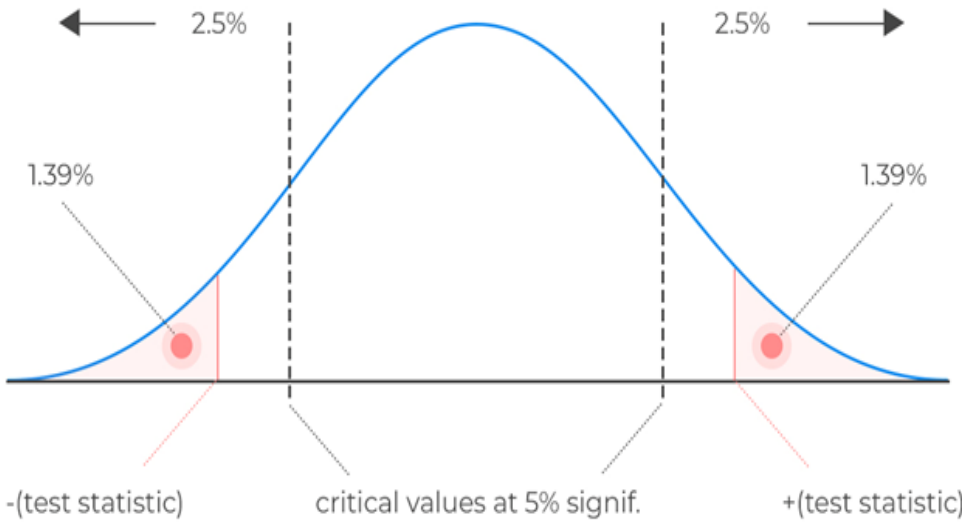
if σ *is unknown, estimate it by* S

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

hence

$$T_{n-1} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

is t-distribution with n-1 degrees of freedom



5.3 Testing equality of means of two Normal populations

Paired t-test can work even if the two populations are not independent and have different variances.

Separate tests for i) known variance, ii) unknown but equal variance, iii) unknown, unequal variance.

The Paired *t*-Test

Before	After	$A - B$
30.5	23	−7.5
18.5	21	2.5
24.5	22	−2.5
32	28.5	−3.5
16	14.5	−1.5
15	15.5	.5

$$W_i = X_i - Y_i, i = 1, \ldots, n.$$

$$H_0 : \mu_w = 0$$
versus
$$H_1 : \mu_w \neq 0$$

normal population having unknown mean and unknown variance

$X_1, \ldots, X_n$ Is a Sample from a $\mathcal{N}(\mu_1, \sigma_1^2)$ Population; $Y_1, \ldots, Y_m$ Is a Sample from a $\mathcal{N}(\mu_2, \sigma_2^2)$ Population

The Two Population Samples Are Independent

to Test

$H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 \neq \mu_2$

Assumption	Test Statistic *TS*	Significance Level α Test	*p*-Value if $TS = t$
σ_1, σ_2 known	$\frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$	Reject if $\lvert TS\rvert > z_{\alpha/2}$	$2P\{Z \geq \lvert t\rvert\}$
$\sigma_1 = \sigma_2$	$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}\sqrt{1/n + 1/m}}$	Reject if $\lvert TS\rvert > t_{\alpha/2, n+m-2}$	$2P\{T_{n+m-2} \geq \lvert t\rvert\}$
n, m large	$\frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n + S_2^2/m}}$	Reject if $\lvert TS\rvert > z_{\alpha/2}$	$2P\{Z \geq \lvert t\rvert\}$

accepting H_0 if $-t_{\alpha/2, n-1} < \sqrt{n}\frac{\overline{W}}{S_w} < t_{\alpha/2, n-1}$

rejecting H_0 otherwise

5.4 Tests comparing the Variance of populations

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \text{vs.} \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

$$P_{H_0}\left\{\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha$$

Therefore, a significance level α test is to

$$\text{accept} \quad H_0 \quad \text{if} \quad \chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{\alpha/2,n-1}$$

$$\text{reject} \quad H_0 \quad \text{otherwise}$$

$$H_0 : \sigma_x^2 = \sigma_y^2 \qquad \text{versus} \qquad H_1 : \sigma_x^2 \neq \sigma_y^2$$

$$S_x^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \qquad S_y^2 = \frac{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}{m-1}$$

denote the sample variances, then as shown in Section 6.5, $(n-1)S_x^2/\sigma_x^2$ and $(m-1)S_y^2/\sigma_y^2$ are independent chi-square random variables with $n-1$ and $m-1$ degrees of freedom, respectively. Therefore, $(S_x^2/\sigma_x^2)/(S_y^2/\sigma_y^2)$ has an F-distribution with parameters $n-1$ and $m-1$. Hence, when H_0 is true

$$S_x^2/S_y^2 \sim F_{n-1,m-1}$$

$$S_x^2/S_y^2 \sim F_{n-1,m-1} \qquad \text{hence}$$

$$P_{H_0}\{F_{1-\alpha/2,n-1,m-1} \leq S_x^2/S_y^2 \leq F_{\alpha/2,n-1,m-1}\} = 1-\alpha$$

hence

$$\text{accept} \quad H_0 \quad \text{if} \quad F_{1-\alpha/2,n-1,m-1} < S_x^2/S_y^2 < F_{\alpha/2,n-1,m-1}$$

$$\text{reject} \quad H_0 \quad \text{otherwise}$$

Chapter-6 Contents

1. One way - Analysis of Variance (ANOVA)
2. Chi-squared test – test independence of two categorical columns

6.1 One way - Analysis of Variance (ANOVA)

Get two estimates of variance: i. When assuming all group means are same, ii. irrespective of the group means being same or not. If they differ much then the group means are not same.

Consider m independent samples, each of size n, where the members of the ith sample — $X_{i1}, X_{i2}, \ldots, X_{in}$ — are normal random variables with unknown mean μ_i and unknown variance σ^2. That is,

$$X_{ij} \sim N(\mu_i, \sigma^2), \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, n$$

We will be interested in testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ versus H_1 : not all the means are equal

Let $\mu_i = \mu$ for all i.

expected value of a chi-square random variable is equal to its number of degrees of freedom

$$E\left(X_n^2\right) = n$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - E[X_{ij}])^2/\sigma^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \mu_i)^2/\sigma^2 \sim \chi^2_{nm}$$

Sum of squares of n*m std norm vars is chi^2 with n*m degrees of freedom

Let $X_{i.} = \sum_{j=1}^{n} X_{ij}/n$ then $\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i.})^2/\sigma^2 \sim \chi^2_{nm-m}$

1 degree of freedom is lost for each parameter that is estimated

Let $SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_{i.})^2$ then $E[SS_W]/\sigma^2 = nm - m$ or $E[SS_W/(nm-m)] = \sigma^2$

$SS_W/(nm-m)$ is an estimator of σ^2. irrespective of H0 being true or not

SS_W is called the *within samples sum of squares*

if H_0 is true then $\dfrac{\sqrt{n}\,(X_{i.} - \mu)}{\sigma}$ *is Std. Normal. Hence, sum of squares of m such vars is chi – squared random variable with m degrees of freedom:* $n\sum_{i=1}^{m}(X_{i.} - \mu)^2/\sigma^2 \sim \chi^2_m$

the estimator of μ is $X_{..} = \dfrac{\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}}{nm} = \dfrac{\sum_{i=1}^{m} X_{i.}}{m}$

hence $n\sum_{i=1}^{m}(X_{i.} - X_{..})^2/\sigma^2 \sim \chi^2_{m-1}$

Let $SS_b = n\sum_{i=1}^{m}(X_{i.} - X_{..})^2$ when H_0 is true, $E[SS_b]/\sigma^2 = m - 1$ or $E[SS_b/(m-1)] = \sigma^2$

SS_b is called the *between samples sum of squares.*

6.1 One way - Analysis of Variance (ANOVA)

H0: all group means are same. i. Calculate the test statistic and p-value.

ii. Reject H0 if p-value < 0.05 (95% confidence interval)

Thus we have shown that

$SS_W/(nm-m)$ always estimates σ^2

$SS_b/(m-1)$ estimates σ^2 when H_0 is true

Because it can be shown that $SS_b/(m-1)$ will tend to exceed σ^2 when H_0 is not true, let the test statistic be given by

$$TS = \frac{SS_b/(m-1)}{SS_W/(nm-m)} \sim F_{m-1,nm-m}$$

reject H_0 when TS is sufficiently large.

The significance level α test of H_0 is as follows:

$$\text{reject } H_0 \text{ if } \frac{SS_b/(m-1)}{SS_W/(nm-m)} > F_{m-1,nm-m,\alpha}$$

do not reject H_0 otherwise

If the value of the test statistic is $TS = v$,

then the p-value will be given by $p\text{-value} = P\{F_{m-1,nm-m} \geq v\}$

6.2 Chi-squared test – test independence of two categorical columns

H_0: X and Y are independent. If H_0 is rejected, we conclude that the two columns are dependent/correlated. Hence columns with minimum p-values are the most important ones!

X	Y
1	9
1	8
1	9
2	6
3	3
3	6
4	8
...	...

Suppose that there are r possible values for the X – characteristic, and s for the Y – characteristic, and let $P_{ij} = P\{X = i,\ Y = j\}$ *for* $i = 1, \ldots, r, \quad j = 1, \ldots, s.$

Let $p_i = P\{X = i\} = \sum_j P_{ij}$ *and* $q_j = P\{Y = j\} = \sum_i P_{ij}$

We are interested in testing :

$H_0 : P_{ij} = p_i q_j$ *for all* i, j *vs* $H_a : P_{ij} \neq p_i q_j$ *for some* i, j

Solution : Let $N_i = \sum_j N_{ij}$ then $\widehat{p_i} = \frac{N_i}{n} \quad i = 1, \ldots, r$

Let $M_j = \sum_i N_{ij}$ then $\widehat{q_j} = \frac{M_j}{n} \quad j = 1, \ldots, s$

$$T = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{(N_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{N_{ij}^2}{n\hat{p}_i\hat{q}_j} - n$$

$$T = \sum_i \sum_j \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}$$

reject H_0 if $T \geq \chi^2_{\alpha,(r-1)(s-1)}$

not reject H_0 otherwise

Let $T = t$, *then*

$p - value = P\left\{X^2_{\alpha,\,(r-1)(s-1)} \geqslant t\right\}$

at 95% confidence interval i.e at 5% level of significance

$p - value = P\left\{X^2_{0.05,\,(r-1)(s-1)} \geqslant t\right\}$

6.2 Chi-squared test – test independence of two categorical columns: question-1

Question: is gender independent of political affiliation ?

EXAMPLE 11.4a A sample of 300 people was randomly chosen, and the sampled individuals were classified as to their gender and political affiliation, Democrat, Republican, or Independent. The following table, called a *contingency table*, displays the resulting data.

	j			
i	**Democrat**	**Republican**	**Independent**	**Total**
Women	68	56	32	156
Men	52	72	20	144
Total	120	128	52	300

Thus, for instance, the contingency table indicates that the sample of size 300 contained 68 women who classified themselves as Democrats, 56 women who classified themselves as Republicans, and 32 women who classified themselves as Independents; that is, $N_{11} = 68, N_{12} = 56$, and $N_{13} = 32$. Similarly, $N_{21} = 52, N_{22} = 72$, and $N_{23} = 20$.

Use these data to test the hypothesis that a randomly chosen individual's gender and political affiliation are independent.

6.2 Chi-squared test – test independence of two categorical columns : question-1

Question: is gender independent of political affiliation ?

Answer: No, political affiliation is dependent on gender.

$n\hat{p}_i\hat{q}_j = N_iM_j/n$ are as follows:

$$\frac{N_1M_1}{n} = \frac{156 \times 120}{300} = 62.40$$
$$\frac{N_1M_2}{n} = \frac{156 \times 128}{300} = 66.56$$
$$\frac{N_1M_3}{n} = \frac{156 \times 52}{300} = 27.04$$
$$\frac{N_2M_1}{n} = \frac{144 \times 120}{300} = 57.60$$
$$\frac{N_2M_2}{n} = \frac{144 \times 128}{300} = 61.44$$
$$\frac{N_2M_3}{n} = \frac{144 \times 52}{300} = 24.96$$

The value of the test statistic is thus

$$TS = \frac{(68 - 62.40)^2}{62.40} + \frac{(56 - 66.56)^2}{66.56} + \frac{(32 - 27.04)^2}{27.04} + \frac{(52 - 57.60)^2}{57.60} + \frac{(72 - 61.44)^2}{61.44} + \frac{(20 - 24.96)^2}{24.96}$$
$$= 6.433$$

Since $(r - 1)(s - 1) = 2$, we must compare the value of TS with the critical value $\chi^2_{.05,2}$. From Table A2

$$\chi^2_{.05,2} = 5.991$$

Since $TS \geq 5.991$, the null hypothesis is rejected at the 5 percent level of significance. That is, the hypothesis that gender and political affiliation of members of the population are independent is rejected at the 5 percent level of significance. ■

Chapter-7 Contents

1. Linear regression (linear in weights)
2. MLE solution: Moore Penrose pseudo inverse – Least square approximation
3. Evaluation: R^2 - higher the better
4. Gradient Descent : 1st order approximation of Taylor's theorem
5. Sequential Gradient Descent Optimization
6. Regularized Least Squares Error (weight decay)
7. Underfitting and Overfitting
8. Bootstrap Aggregation (Bagging) – reduces both Bias and Variance
9. Bagging demo
10. Linear Regression: Loss = Modeling loss + Noise
11. Bias – Variance tradeoff !
12. Regularization – balance between Bias and Variance
13. Kernel Regression

7.1 Linear regression (linear in weights)

Target (t) = Linear (w.x) + GaussianNoise

E(t|x) = y = w.x (Linear in w, not in x)

$$\textit{Assumption}: \; t = y(\boldsymbol{w}, \boldsymbol{x}) + \epsilon \;\; \textit{where} \;\; \epsilon \sim N(0, \sigma^2)$$
$$\textit{or} \quad t = N(y(\boldsymbol{w}, \boldsymbol{x}), \sigma^2) \;\; \textit{hence} \;\; E[t] = y(\boldsymbol{w}, \boldsymbol{x})$$

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

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

$\phi_j(\mathbf{x})$ are known as *basis functions*.

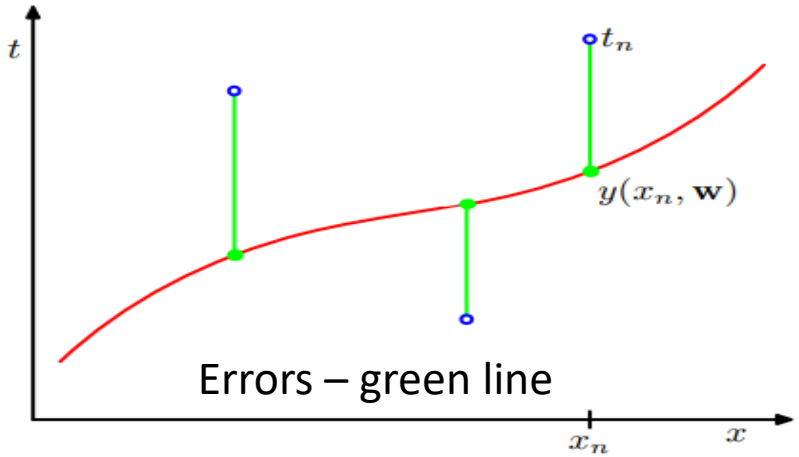
define an additional dummy 'basis function' $\phi_0(\mathbf{x}) = 1$

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

$$\textit{common basis functions} \sim x^n, \; e^x, \; e^{x^2}, \; log(x), \; tanh(x) \quad e.g. \; \textit{if} \; \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\textit{let} \; y(\boldsymbol{w}, \boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 e^{x_1} + w_6 e^{x_2} = \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_6 \end{bmatrix}^T \begin{bmatrix} 1 \\ x_1 \\ \dots \\ e^{x_2} \end{bmatrix} = \boldsymbol{w}^T \phi(\boldsymbol{x})$$

$$\textit{Assumption}: \; t = y(\boldsymbol{w}, \boldsymbol{x}) + \epsilon \;\; \textit{where} \;\; \epsilon \sim N(0, \sigma^2)$$
$$\textit{or} \quad t = N(y(\boldsymbol{w}, \boldsymbol{x}), \sigma^2) \;\; \textit{hence} \;\; E[t|x] = y(\boldsymbol{w}, \boldsymbol{x})$$



Errors – green line

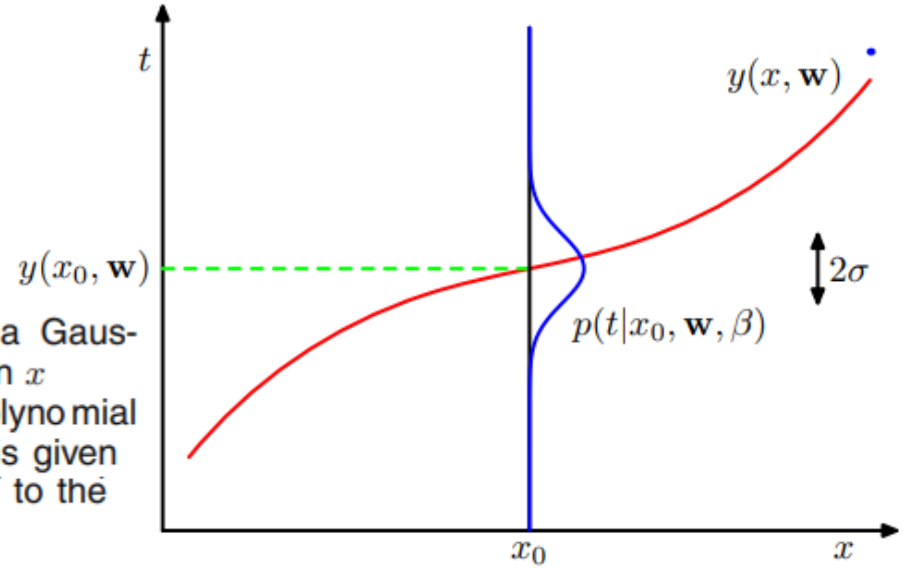


Figure Schematic illustration of a Gaussian conditional distribution for t given x in which the mean is given by the polynomial function $y(x, \mathbf{w})$, and the precision is given by the parameter β, which is related to the variance by $\beta^{-1} = \sigma^2$.

7.2 MLE solution: Moore Penrose pseudo inverse – Least square approximation

- Moore Penrose pseudo inverse can go out of memory. - It represents a least square projection of targets T (n-dim) onto the vector-subspace spawned by columns of X (k-dim): k<n

$$Let \quad \mathbf{X} = \begin{matrix} \\ \phi(\boldsymbol{x}_n) \rightarrow \end{matrix} \begin{bmatrix} x_{10} & x_{11} & x_{12} & .. & x_{2k} \\ x_{20} & x_{21} & x_{22} & .. & x_{2k} \\ : & : & : & :: & : \\ x_{n0} & x_{n1} & x_{n2} & .. & x_{nk} \end{bmatrix} \quad \boldsymbol{T} = \begin{bmatrix} t_1 \\ t_2 \\ : \\ t_n \end{bmatrix}$$

sum-of-squares error function is defined by

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2.$$

This is also the Max. Likelihood Estimate (MLE) of W.

$$then\ MLE\ estimate\ for\ \mathbf{W}\ s.t. \quad \boldsymbol{T} = \mathbf{W}^T\mathbf{X} + \boldsymbol{\epsilon}$$

$$is \quad \mathbf{W}_{MLE} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{T} \quad where$$

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = Moore\ Penrose\ pseudo\ inverse$$

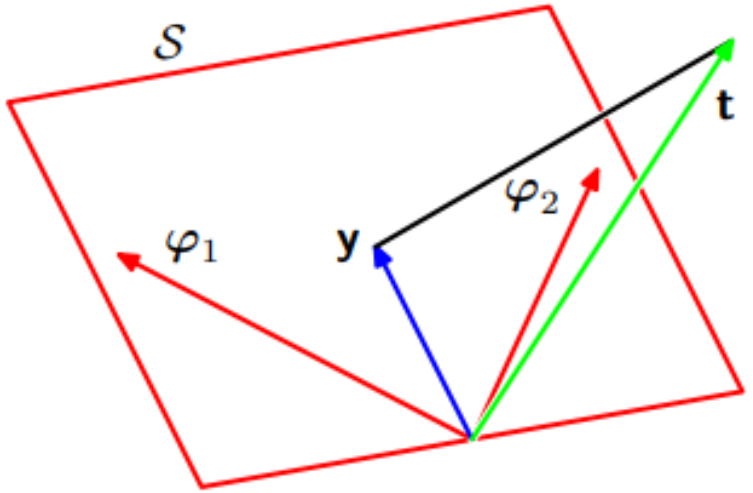
if features = $\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$

Then:

X: large matrix can go out of memory. Hence, use sequential optimization approaches as Seq. Gradient Descent.

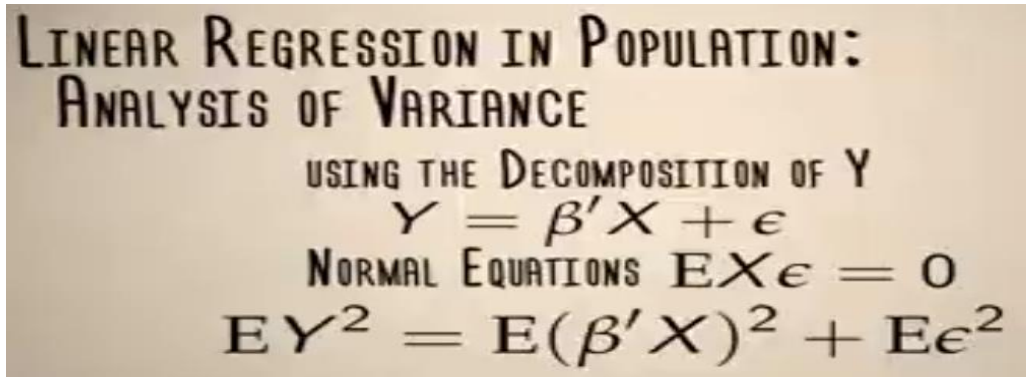
Online Learning: Sequential learning is also appropriate for real-time applications in which the data observations are arriving in a continuous stream, and predictions must be made before all of the data points are seen.

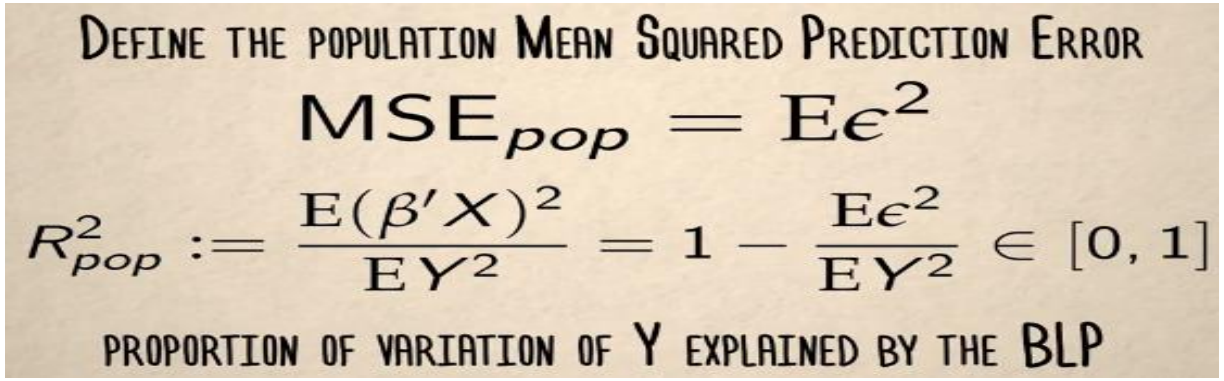
Geometrical interpretation of the least-squares solution, in an N-dimensional space whose axes are the values of $t_1, \ldots, t_N$. The least-squares regression function is obtained by finding the orthogonal projection of the data vector **t** onto the subspace spanned by the basis functions $\phi_j(\mathbf{x})$ in which each basis function is viewed as a vector φ_j of length N with elements $\phi_j(\mathbf{x}_n)$.

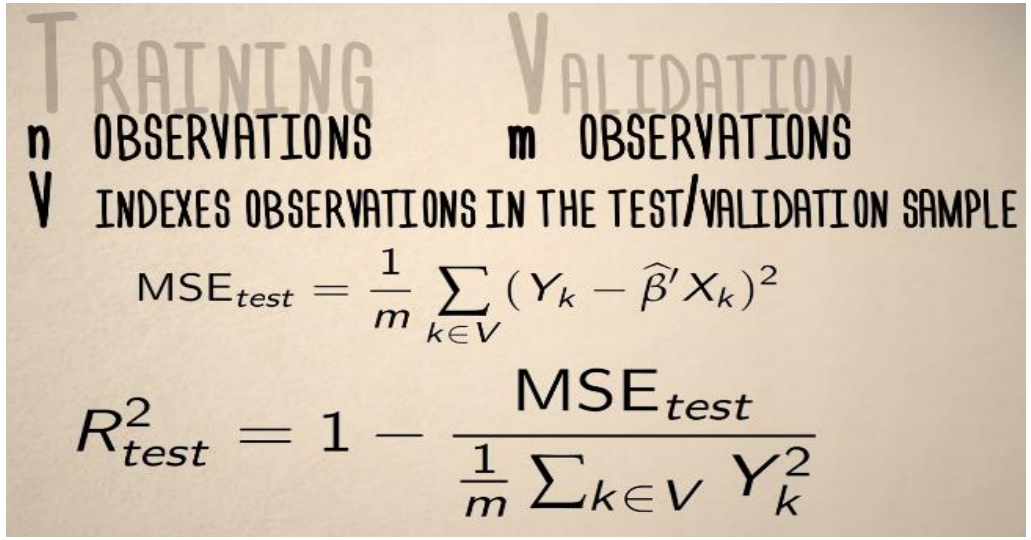


7.3 Evaluation: R^2 - higher the better

R^2 = Proportion of variance explained by the linear model. It is a statistical measure used to evaluate the goodness of fit of a regression model. max R^2 = 1 is the best it can get.







7.4 Gradient Descent : 1st order approximation of Taylor's theorem

$1^{st} order\ approx.: f(x) = f(a) + f'(a)(x-a)\ \ where\ \ x\ is\ close\ to\ a$

$x = x - \mu f'(x)\ \ decreases\ f(x)\ :\ Gradient\ Descent\ !$

If a real-valued function $f(x)$ is differentiable at the point $x = a$, then it has a linear approximation near this point. This means that there exists a function $h_1(x)$ such that

$$f(x) = f(a) + f'(a)(x-a) + h_1(x)(x-a), \quad \lim_{x \to a} h_1(x) = 0.$$

Here *Limiting Taylor's theorem to 1st derivative gives the 1st order approximation*

$$P_1(x) = f(a) + f'(a)(x-a) \qquad \approxeq\ f(x)\ \ if\ \ x\ is\ near\ a$$

is the linear approximation of $f(x)$ for *x* near the point *a*, whose graph $y = P_1(x)$ is the tangent line to the graph $y = f(x)$ at *x* = *a*. The error in the approximation is:

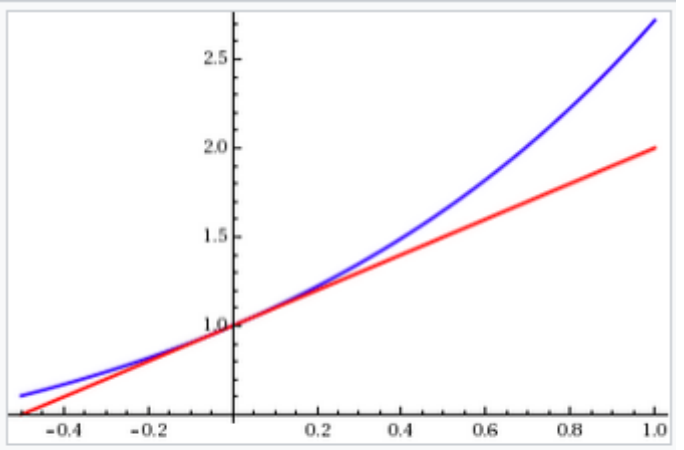
$$R_1(x) = f(x) - P_1(x) = h_1(x)(x-a).$$

As *x* tends to *a*, this error goes to zero much faster than $f'(a)(x-a)$, making $f(x) \approx P_1(x)$ a useful approximation.

For a better approximation to $f(x)$, we can fit a quadratic polynomial instead of a linear function:

Limiting Taylor's theorem to 2nd derivative gives the 2nd order approximation

$$P_2(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2. \qquad \approxeq\ f(x)\ \ if\ \ x\ is\ near\ a$$



Graph of $f(x) = e^x$ (blue) with its linear approximation $P_1(x) = 1 + x$ (red) at $a = 0$.

$1^{st} order\ approx.: f(x) = f(a) + f'(a)(x-a)\ \ when\ \ x\ is\ close\ to\ a$

$Let\ x_{new} = a - \mu f'(a)\ \ here\ \mu \ll 1\ \ s.t\ \ \mu f'(a) \equiv 0.01,\ then:$

$$f\big(a - \mu f'(a)\big) = f(a) + f'(a)\big(a - \mu f'(a) - a\big)$$
$$or\ f\big(a - \mu f'(a)\big) = f(a) + f'(a)\left(-\mu f'(a)\right)$$
$$or\ f\big(a - \mu f'(a)\big) = f(a) - \mu f'^{2}(a)$$

$now\ f'^{2}(a) \geqslant 0\ \ because\ it\ is\ squared,\ hence\ f\big(a - \mu f'(a)\big) \leqslant f(a)$

$thus\ x = x - \mu f'(x)\ \ decreases\ f(x)\ :\ Gradient\ Descent\ !$

7.4 Gradient Descent : gradually reduce the learning rate

$1^{st} order\ approx.: f(x) = f(a) + f'(a)(x-a)\ \ where\ \ x\ is\ close\ to\ a$

$x = x - \mu f'(x)\ \ decreases\ f(x)\ :\ Gradient\ Descent\ !$

Limiting Taylor's theorem to 1st derivative gives the 1st order approximation

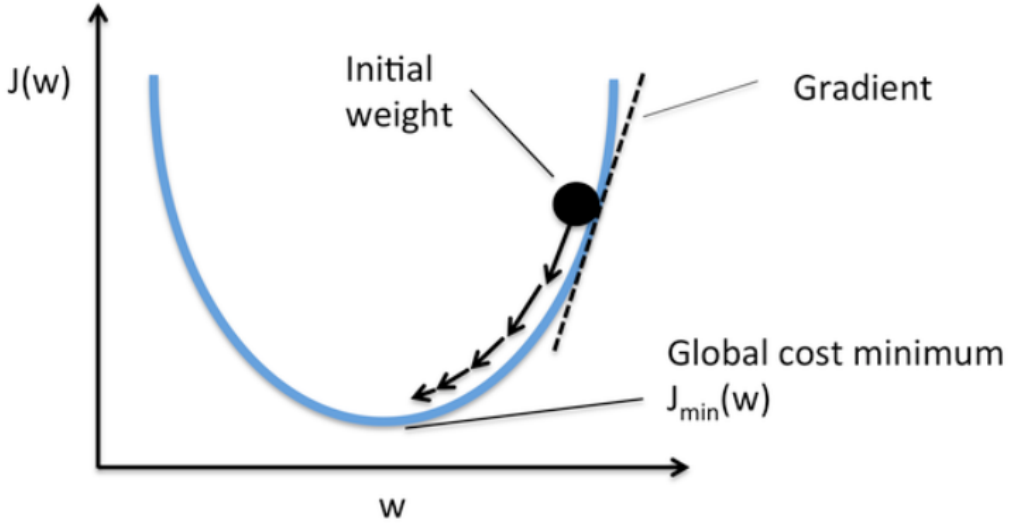
Limiting Taylor's theorem to 2nd derivative gives the 2nd order approximation

If learning rate (LR) is high, GD may never converge.

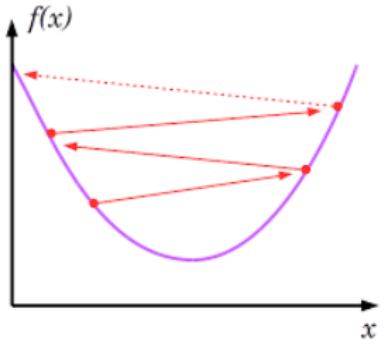
Hence, gradually reduce the LR.

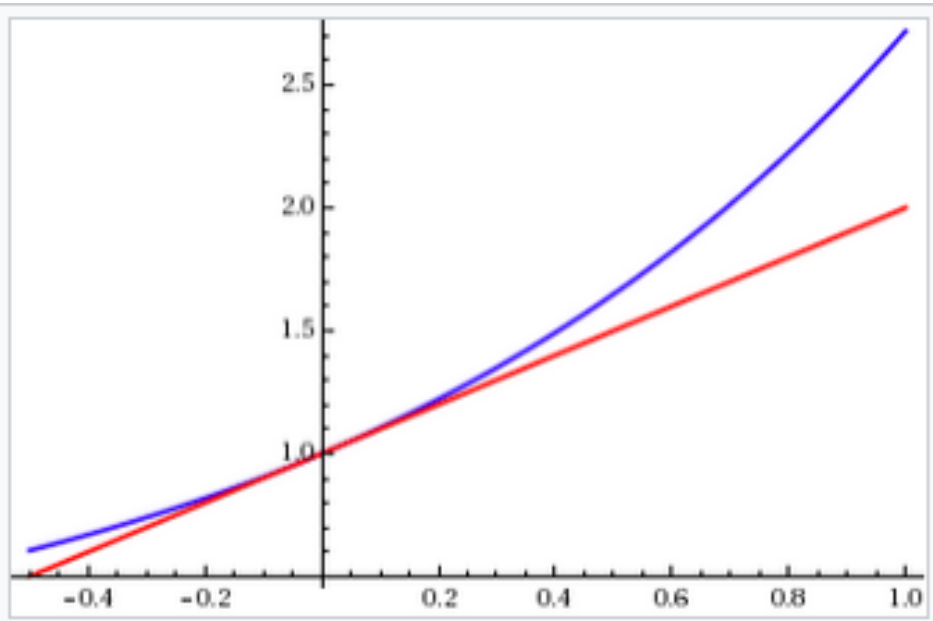
The gradients matrix is called as **Jacobian**. Matrix operations can be accelerated on GPUs.

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix}$$



f the alpha is high the step oscillates and global minimum is not guaranteed.





Graph of $f(x) = e^x$ (blue) with its linear approximation $P_1(x) = 1 + x$ (red) at $a = 0$.

There are libraries such as JAX which use N-th order approximation for GD optimization.
https://jax.readthedocs.io/en/latest/index.html

7.5 Sequential Gradient Descent Optimization

Sequential mini-batch gradient descent can help in online-learning also

$$w_{\tau+1} = w_\tau + \eta \sum_{n=1}^{B} \{t_n - w^T \phi_n\} \phi_n$$

$B =$ *no. of samples in a mini batch*

error function comprises a sum over data points $E = \sum_n E_n$.

Instead of the total E at once, apply E_n sequentially for all x_n:

sum-of-squares error function is defined by

$$E_D(\mathbf{w}) = \frac{1}{2} \sum^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n)\}^2.$$

$$E'_D(w) = -\sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\} \phi(x_n)$$

$$\textit{Gradient Descent}: w_{n+1} = w_n + \eta \sum_{n=1}^{N} \{t_n - w^T \phi_n\} \phi_n$$

$\eta =$ *learning rate parameter e.g.* $\mu = 0.001$

$\phi_n = \phi(\mathbf{x}_n)$

The value of $\mathbf{w}$ is initialized to some starting vector $\mathbf{w}^{(0)}$.

for epoch = 1 to K:

  for n = 1 to N:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n \quad \text{or} \quad \mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta (t_n - \mathbf{w}^{(\tau)\mathrm{T}} \boldsymbol{\phi}_n) \boldsymbol{\phi}_n$$

where τ denotes the iteration number, and η is a learning rate parameter. $\phi_n = \phi(\mathbf{x}_n)$.

Sequential Gradient Descent

- Faster, less expensive
- Less overfitting

The value of $\mathbf{w}$ is initialized to some starting vector $\mathbf{w}^{(0)}$.

for epoch = 1 to k:

  for batch = 1 to M:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n \quad \text{or} \quad w_{\tau+1} = w_\tau + \eta \sum_{n=1}^{B} \{t_n - w^T \phi_n\} \phi_n$$

$B =$ *no. of samples in a mini batch* $\quad B \times M = N$: *Total no of training samples*

Stochastic (mini batch) Gradient Descent (fastest)

- In between Sequential and full Grad. descent

7.6 Regularized Least Squares Error (weight decay)

Regularization constrains W from growing as it penalizes W.

Thus, it prevents overfitting.

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

Err = $\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$

If q = 1 we have L1 regularization (Lasso)

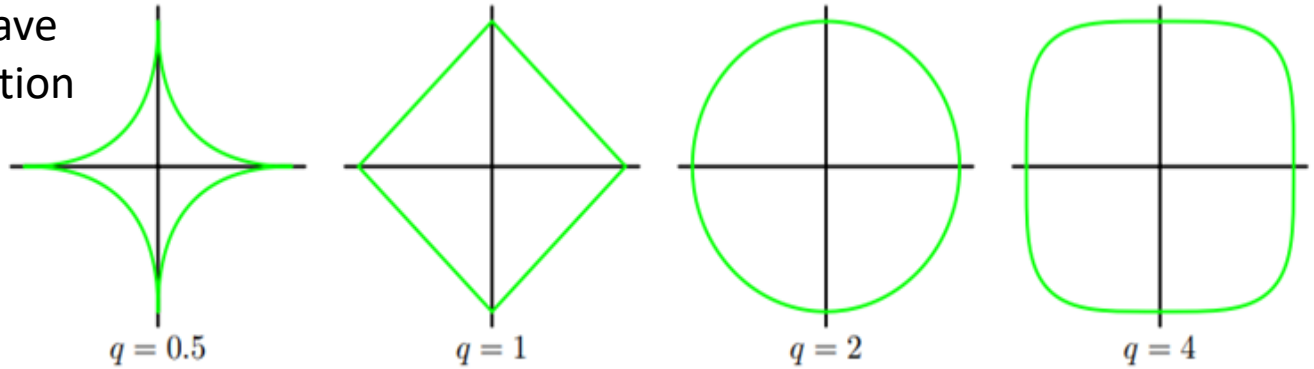


Figure Contours of the regularization term for various values of the parameter q.

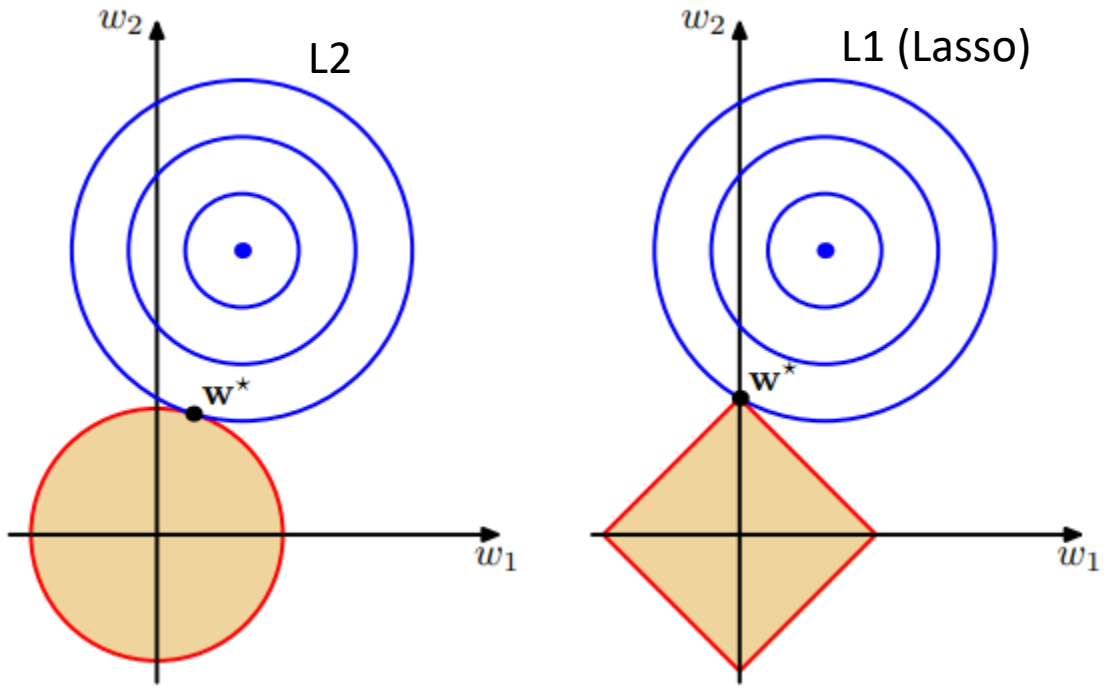
If q =2 we have L2 regularization (Ridge regression):

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}.$$

Now $\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$

L1 regularizer (Lasso) yields sparse W as it sets most of the not so important dimensions to 0 or very small values close to 0.
(i.e. features which are not so important their weights will be near 0.)

Figure Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$. The lasso gives a sparse solution in which $w_1^\star = 0$.



7.7 Underfitting and Overfitting

Underfitting ~ high Bias : Scope to add more complex features.

Overfitting ~ high Variance : Starts memorizing training data and give near 100% accuracy on it.

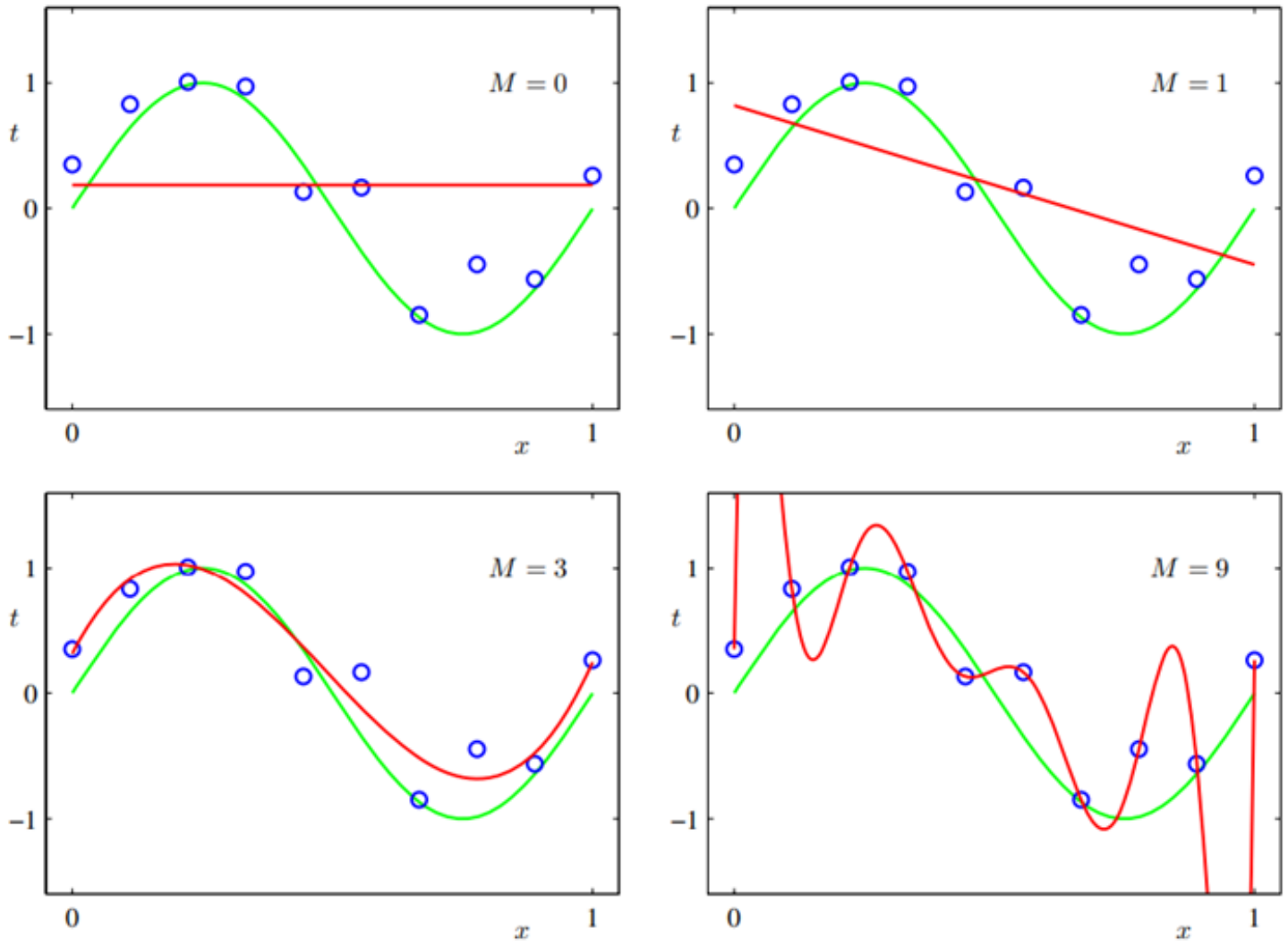
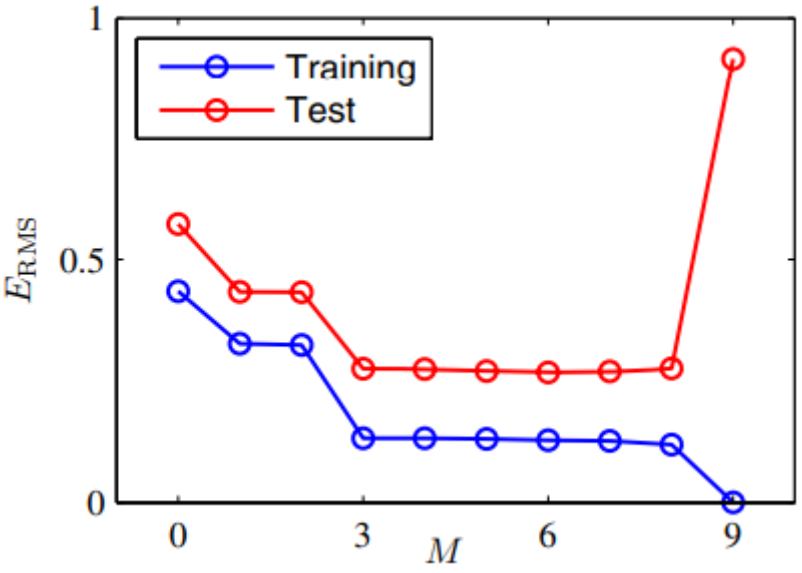


Figure Plots of polynomials having various orders M, shown as red curves

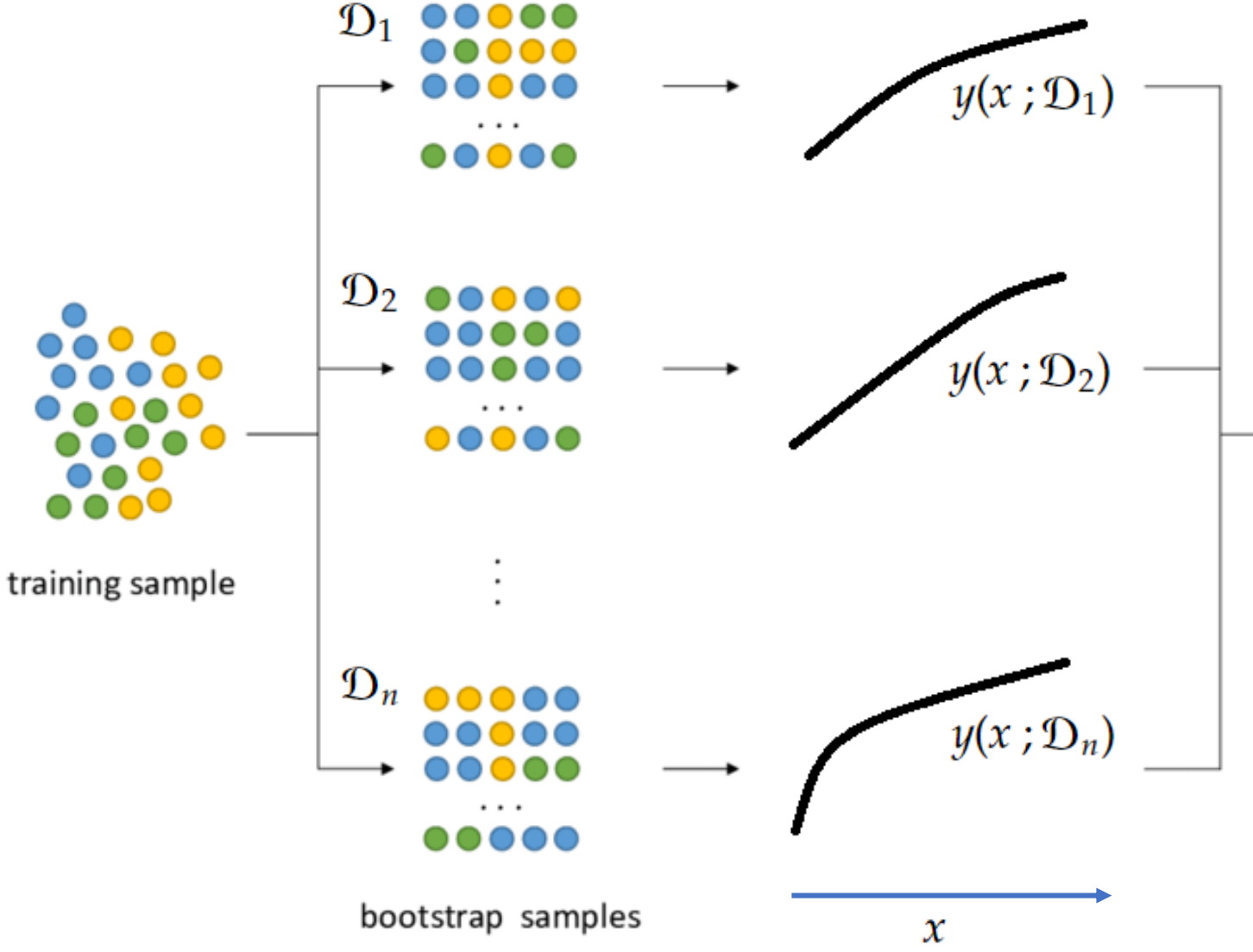
Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M.



7.8 Bootstrap Aggregation (Bagging) – reduces both Bias and Variance

create models $y(x;\mathcal{D}_1)\ y(x;\mathcal{D}_2) \ldots y(x;\mathcal{D}_n)$ *on samples* $\mathcal{D}_1\ \mathcal{D}_2\ \ldots\ \mathcal{D}_n$

Final model $y(x) =$ *Avg. or Expectation over* $\mathcal{D}$ *of* $y(x;\mathcal{D}) = E_{\mathcal{D}}[y(x;\mathcal{D})] = \dfrac{y(x;\mathcal{D}_1) + y(x;\mathcal{D}_2) + \ldots + y(x;\mathcal{D}_n)}{n}$



Let variable $\mathcal{D} = \{\mathcal{D}_1\ \mathcal{D}_2\ \ldots\ \mathcal{D}_n\}$

Final model $y(x) =$ *Avg. or Expectation over* $\mathcal{D}$ *of* $y(x;\mathcal{D}) = E_{\mathcal{D}}[y(x;\mathcal{D})]$

$$y(x) = E_{\mathcal{D}}[y(x;\mathcal{D})] = \frac{y(x;\mathcal{D}_1) + y(x;\mathcal{D}_2) + \ldots + y(x;\mathcal{D}_n)}{n}$$

predictions

- Averaging minimizes variance in predictions.
- Averaging is costly

Bagging is an ensemble algorithm that fits multiple models on different subsets of a training dataset, then combines the predictions from all models.
Random forest is an extension of bagging that also randomly selects subsets of features used in each data sample.

7.9 Bagging demo

Ensembel tends not to overfit (less variance)

Ensemble tends to be more accurate (less bias)

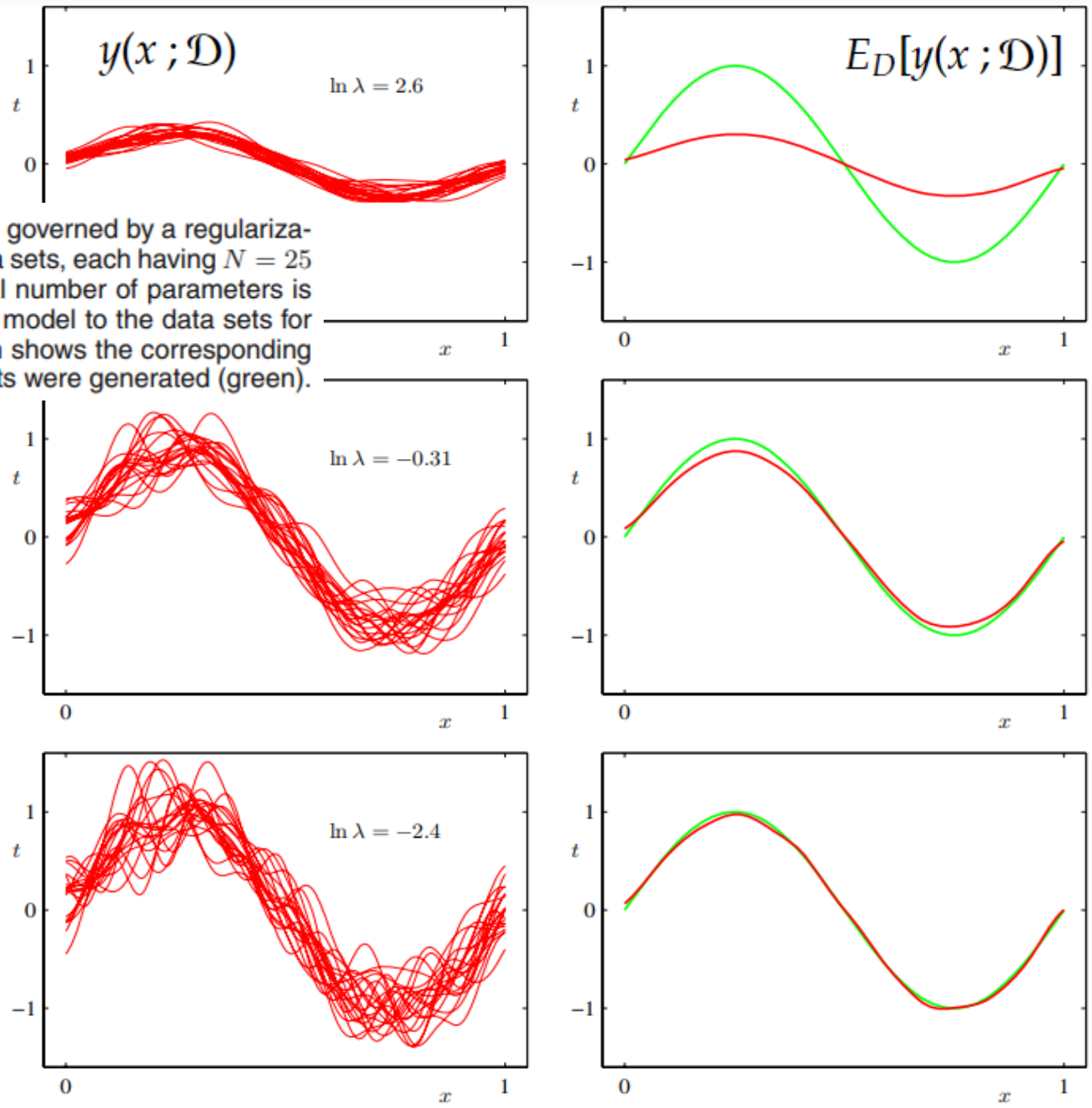


Figure Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ, using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln\lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

$$E_D[y(x;\mathcal{D})] = \frac{y(x;\mathcal{D}_1) + y(x;\mathcal{D}_2) + \ldots + y(x;\mathcal{D}_n)}{n}$$

7.10 Linear Regression: Loss = Modeling loss + Noise

Least square loss decomposed into :

i) Modeling loss, and ii) Noise. Independent noise signals cannot be predicted.

$$Let\ E[t|x] = \int t\, p(t|x)\, dt = h(x)$$

$$Loss = E_{xt}\left[(y(x) - t)^2\right] = E_{xt}\left[(y(x) - E[t|x] + E[t|x] - t)^2\right]$$

$$= E_x\left[(y(x) - E[t|x])^2\right] + E_{xt}\left[(E[t|x] - t)^2\right] + 2E_{xt}[(y(x) - E[t|x])(E[t|x] - t)]$$

$$now\ 2E_{xt}[(y(x) - E[t|x])(E[t|x] - t)] = 0 \quad hence:$$

$$Loss = E_x\left[(y(x) - E[t|x])^2\right] + E_{xt}\left[(E[t|x] - t)^2\right]$$

$$Loss = \underbrace{E_x\left[(y(x) - h(x))^2\right]}_{modeling\ loss} + \underbrace{E_{xt}\left[(h(x) - t)^2\right]}_{Noise}$$

$$Noise = E_{xt}\left[(h(x) - t)^2\right] = \int_x \int_t (h(x) - t)^2\, p(t|x)\, p(x)\, dt\, dx$$

Noise cannot be predicted, hence this is the minimum expected loss !

The loss component which can be minimized is the :

$$modeling\ loss = E_x\left[(y(x) - h(x))^2\right] = \int_x (y(x) - h(x))^2\ p(x)\, dx$$

$$min.\ modeling\ loss = 0\ when\ prediction\ y(x) == h(x) = E[t|x]$$

7.11 Bias – Variance tradeoff !

Modeling loss = bias^2 + variance

If bias decreases, variance can increase and vice-versa.

$$\text{Hence}$$
$$\text{Loss} = \text{bias}^2 + \text{variance} + \text{noise}$$

If we have multiple trained models across different datasets $\mathcal{D}_1\mathcal{D}_2..\mathcal{D}_n$
and instead of avg. the predictions, we avg. the loss, then :

$$\text{modeling loss} = \text{Avg. loss across datasets } \mathcal{D}_1\mathcal{D}_2..\mathcal{D}_n = E_{xD}\left[(y(x;\mathcal{D}) - h(x))^2\right]$$

$$\text{Hence}$$
$$\text{Loss} = \text{modeling loss} + \text{noise} = \text{bias}^2 + \text{variance} + \text{noise}$$

$$\text{remember } E_D[y(x;\mathcal{D})] = \frac{y(x;\mathcal{D}_1) + y(x;\mathcal{D}_2) + \ldots + y(x;\mathcal{D}_n)}{n}$$

$$\text{then modeling loss} = E_{xD}\left[(y(x;\mathcal{D}) - E_D[y(x;\mathcal{D})] + E_D[y(x;\mathcal{D})] - h(x))^2\right]$$
$$= E_{xD}\left[\{y(x;\mathcal{D}) - E_D[y(x;\mathcal{D})]\}^2\right] + E_x\left[\{E_D[y(x;\mathcal{D})] - h(x)\}^2\right]$$
$$+ 2E_{xD}[\{y(x;\mathcal{D}) - E_D[y(x;\mathcal{D})]\}\{E_D[y(x;\mathcal{D})] - h(x)\}]$$
$$\text{the last term} = 0 \quad \text{hence}$$

$$\text{modeling loss} = \underbrace{E_{xD}\left[\{y(x;\mathcal{D}) - E_D[y(x;\mathcal{D})]\}^2\right]}_{\text{variance}} + \underbrace{E_x\left[\{E_D[y(x;\mathcal{D})] - h(x)\}^2\right]}_{\text{bias}^2}$$

$$= \underbrace{\int_x E_D\left[\{y(x;\mathcal{D}) - E_D[y(x;\mathcal{D})]\}^2\right] p(x)dx}_{\text{variance}} + \underbrace{\int_x \{E_D[y(x;\mathcal{D})] - h(x)\}^2 p(x)dx}_{\text{bias}^2}$$

There is a trade-off between bias and variance, with very flexible models having low bias and high variance, and relatively rigid models having high bias and low variance. The model with the optimal predictive capability is the one that leads to the best balance between bias and variance.

7.12 Regularization – balance between Bias and Variance

Bias: deviation between h(x) and averaged curve

Variance: deviation between y(x) and averaged curve

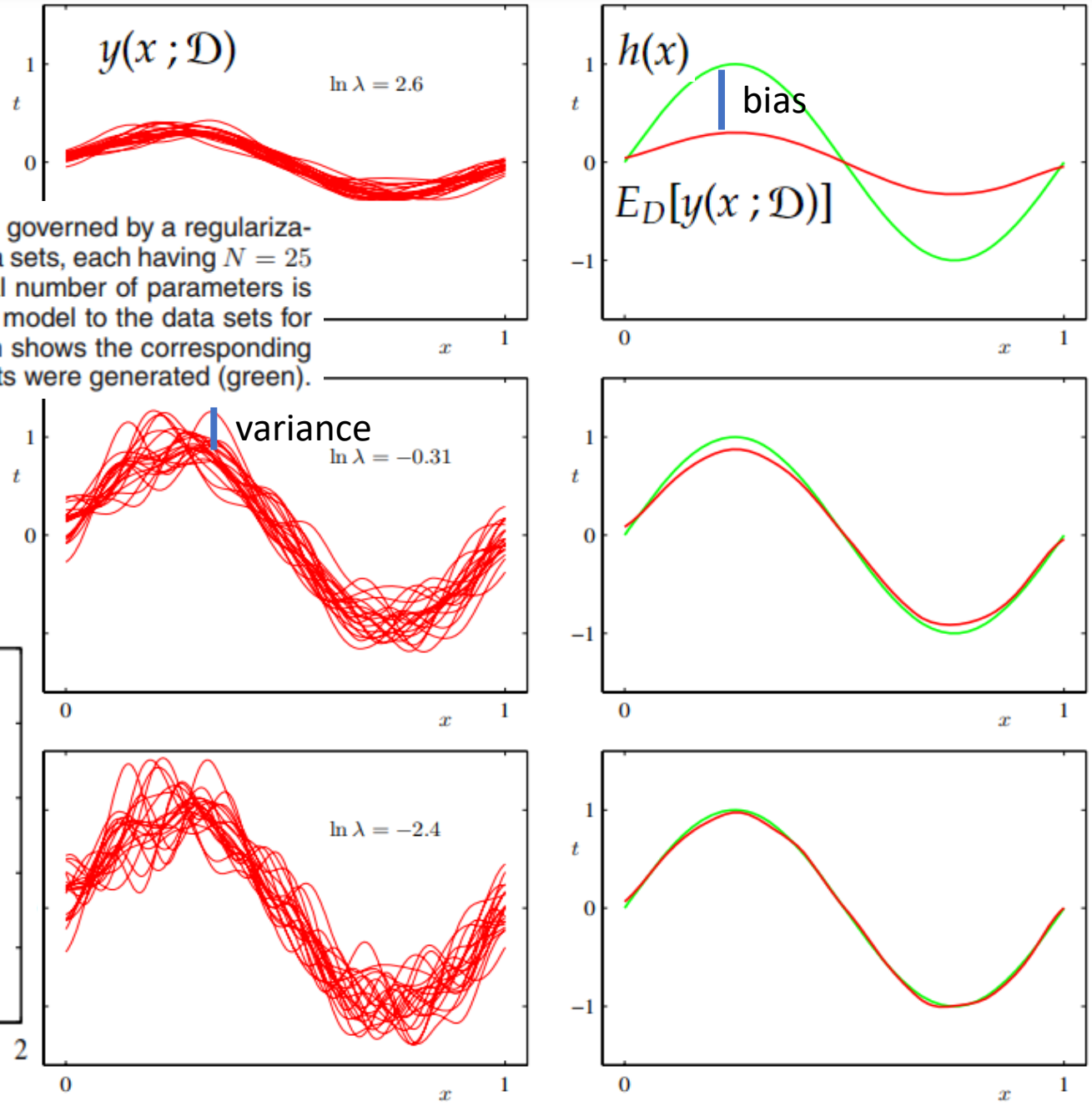
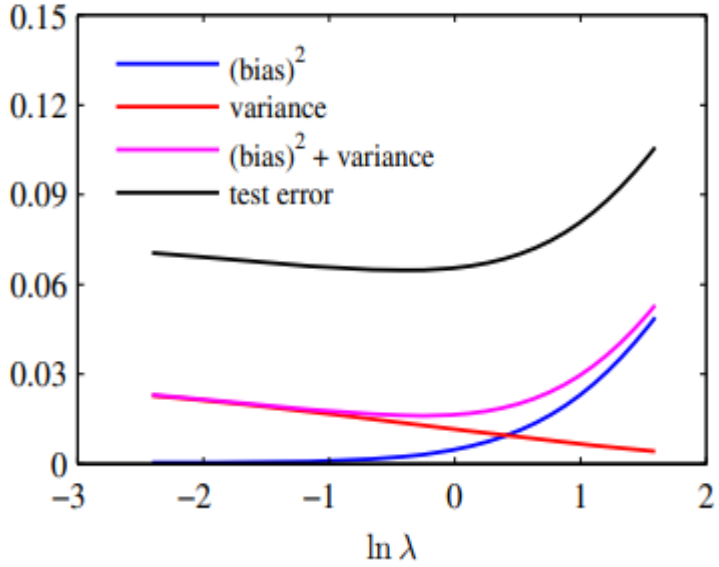


Figure Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ, using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln\lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

$$bias^2 = E_x\left[\{E_D[y(x;\mathcal{D})] - h(x)\}^2\right]$$

$$variance = E_{xD}\left[\{y(x;\mathcal{D}) - E_D[y(x;\mathcal{D})]\}^2\right]$$

Plot of squared bias and variance, together with their sum, corresponding to the results shown in Figure 3.5. Also shown is the average test set error for a test data set size of 1000 points. The minimum value of $(\text{bias})^2 + \text{variance}$ occurs around $\ln\lambda = -0.31$, which is close to the value that gives the minimum error on the test data.



Stronger regularization yields higher bias & lower variance

7.13 Kernel Regression

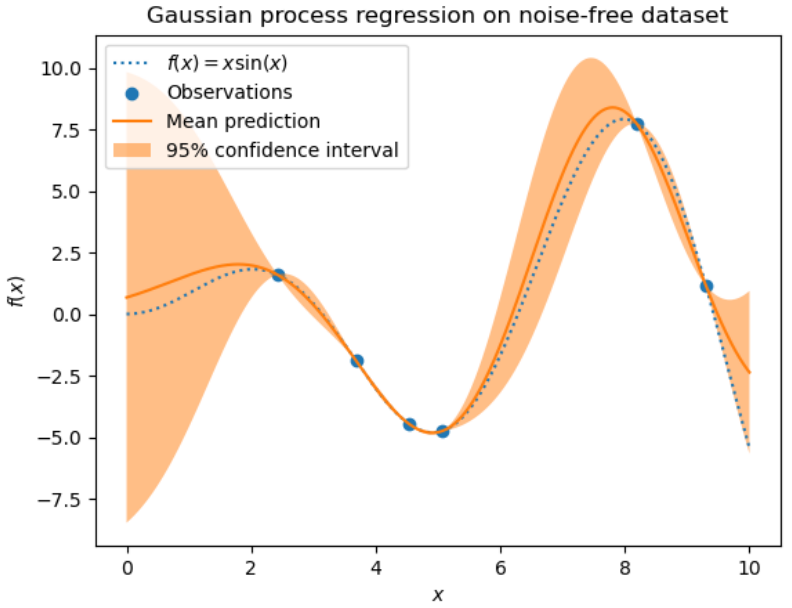
Purely data driven: interpolate smoothly between the observations

Not assuming any kind of linear relationships

Gaussian Process Regression

https://scikit-learn.org/stable/modules/gaussian_process.html#

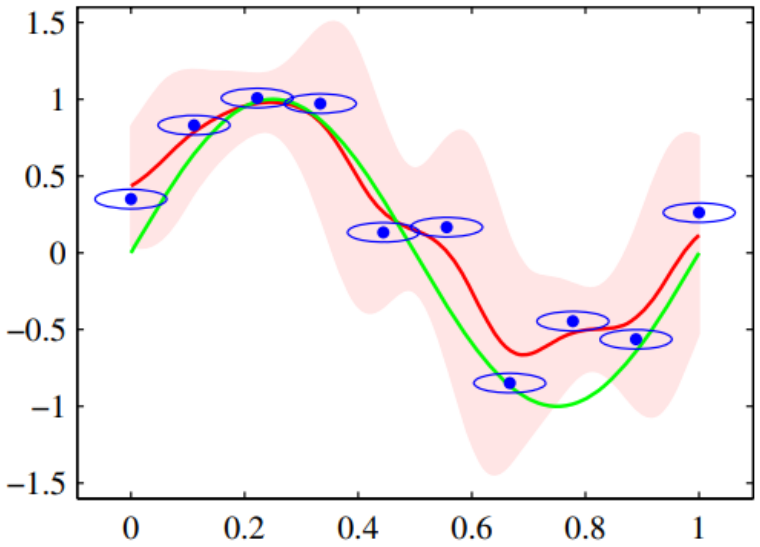
https://github.com/jwangjie/Gaussian-Processes-Regression-Tutorial



Kernel Regression (hands-on)

https://fda.readthedocs.io/en/latest/auto_examples/plot_kernel_regression.htm

Figure Illustration of the Nadaraya-Watson kernel regression model using isotropic Gaussian kernels, for the sinusoidal data set. The original sine function is shown by the green curve, the data points are shown in blue, and each is the centre of an isotropic Gaussian kernel. The resulting regression function, given by the conditional mean, is shown by the red line, along with the two-standard-deviation region for the conditional distribution $p(t|x)$ shown by the red shading. The blue ellipse around each data point shows one standard deviation contour for the corresponding kernel. These appear noncircular due to the different scales on the horizontal and vertical axes.



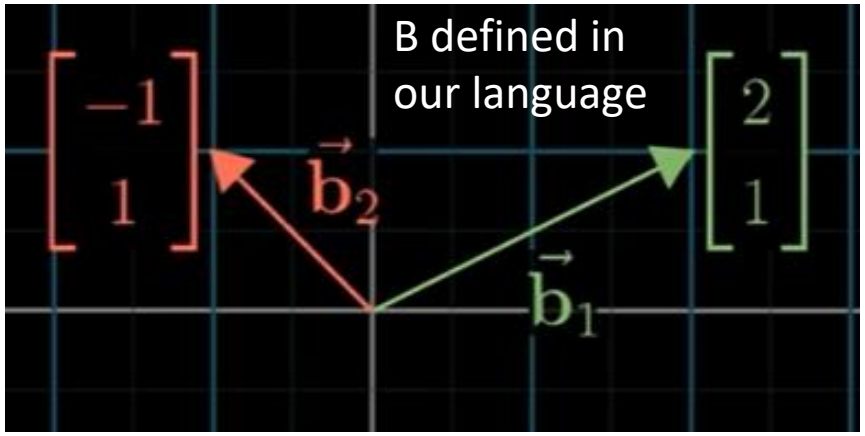
8 Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)

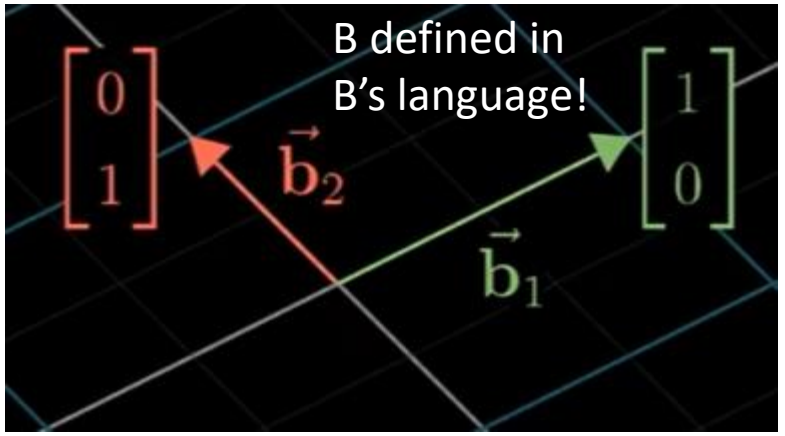
Contents

Contents

8.1.1 Change of Axis/Bases (but origin {0, 0} is always same)

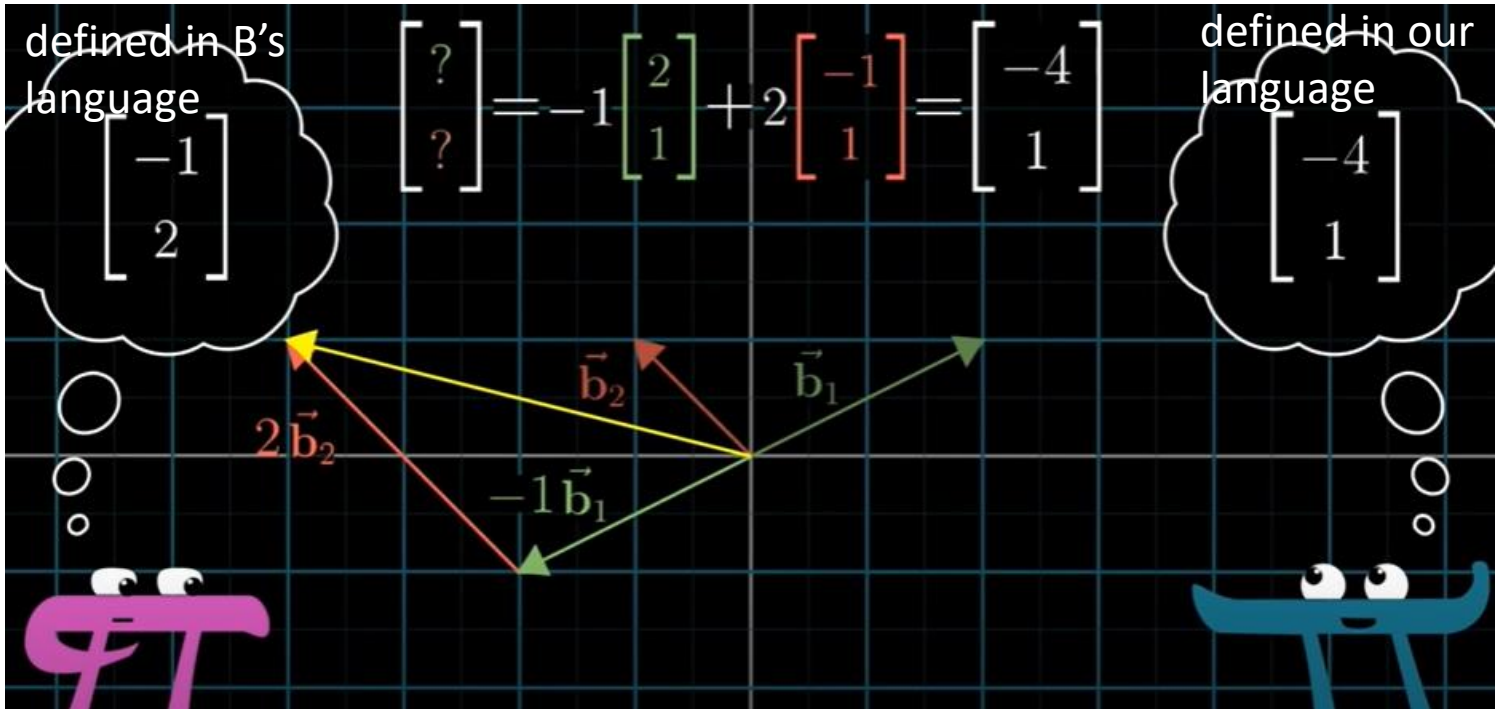
Let b1 b2 are the new basis. Basis need not be orthogonal always. A point say (-1, 2) in the new coordinate system is represented in our coordinate system as = -1b1 + 2b2 = $[b_1, b_2]\begin{bmatrix}-1\\2\end{bmatrix}$





Let this new coordinate system be represented by the new bases B:

$$B = [b_1\ b_2] = \begin{bmatrix}2 & -1\\1 & 1\end{bmatrix}$$



a point $\begin{bmatrix}-1\\2\end{bmatrix}$ *in B will be represented in our coordinate system as* $[b_1\ b_2]\begin{bmatrix}-1\\2\end{bmatrix} = \begin{bmatrix}2 & -1\\1 & 1\end{bmatrix}\begin{bmatrix}-1\\2\end{bmatrix} = \begin{bmatrix}-4\\1\end{bmatrix}$

$$B = [b_1\ b_2] = \begin{bmatrix}2 & -1\\1 & 1\end{bmatrix} = \textit{Change of Basis Matrix}$$

and translates vectors in B's language to our language.

Note : B defines the new basis in our language!

8.1.2 Represent a vector defined in our language/axis to B's language/axis

$B = [b_1\ b_2] = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}$ = *Change of Basis Matrix*

and translates vectors in B's language to our language.

B^{-1} = *inverse Change of Basis Matrix*

it translates vectors in our lang to B's lang

$$B^{-1}\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} x_j \\ y_j \end{bmatrix}$$

Vector in B's language — Vector in our language

$$B\begin{bmatrix} x_j \\ y_j \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

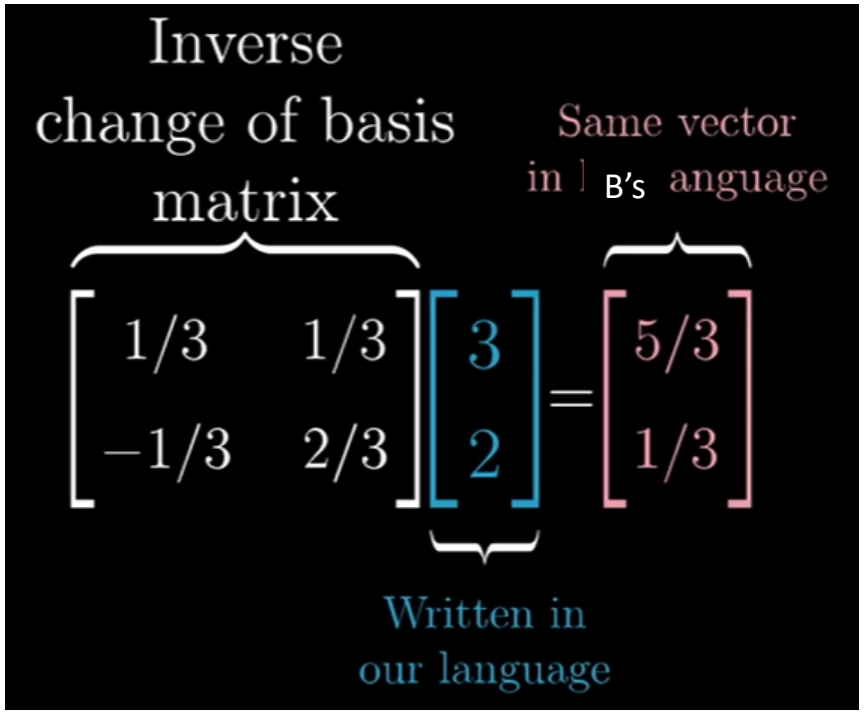
so

$$\begin{bmatrix} x_j \\ y_j \end{bmatrix} = B^{-1}\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

Vector in our language — Vector in B's language

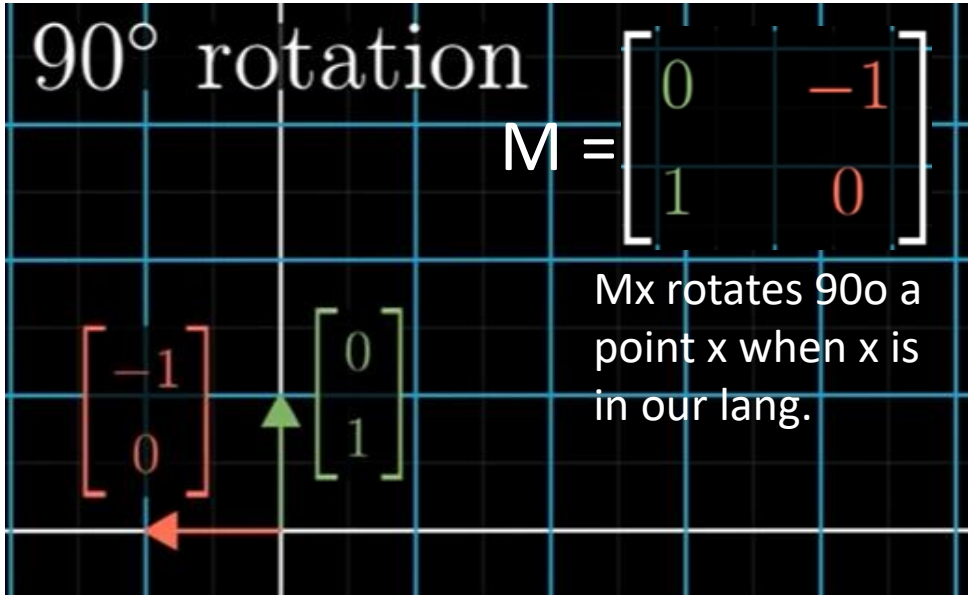
$$B^{-1}\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} x_j \\ y_j \end{bmatrix}$$

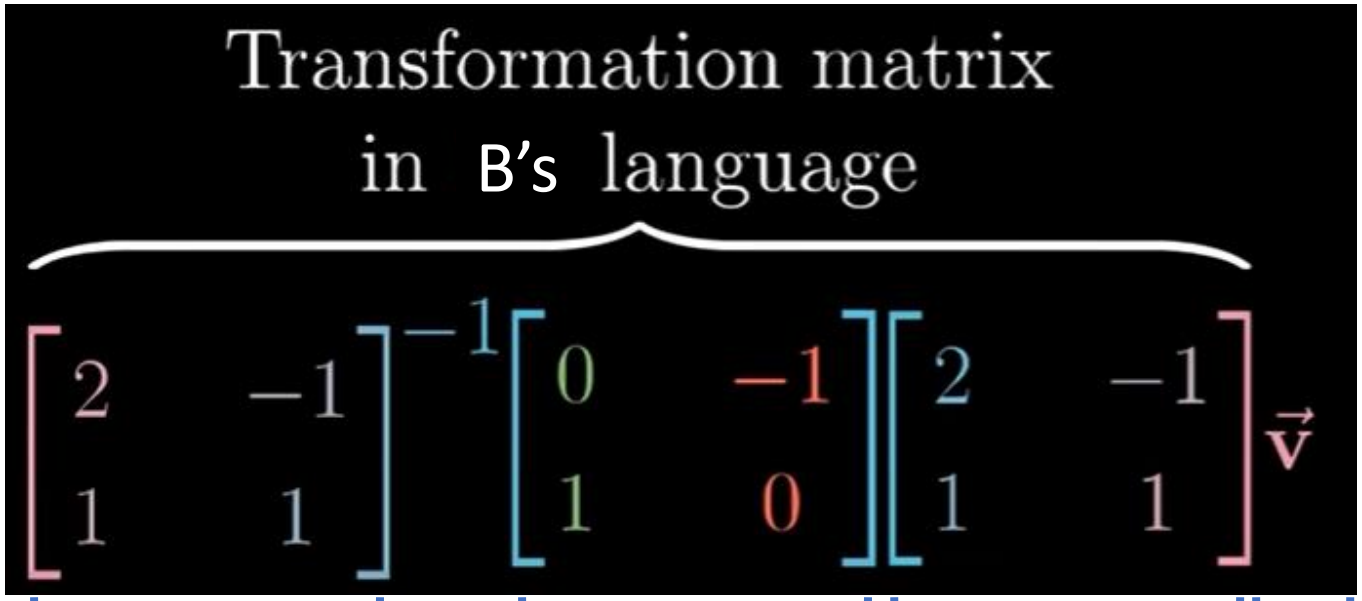
$$B^{-1} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{-1}{3} & \frac{2}{3} \end{bmatrix} = \textit{inverse Change of Basis Matrix}$$



8.1.3 Applying transformation in our coordinate system to B's system

Given a transformation M (e.g. 90o rotation) in our coordinate system. How to apply the same transformation in B's coordinate system ? $M \rightarrow B^{-1}MB$





Translate the transformed (rotated) vector back to B's lang.

Transform (90o rotation) in our coordinate system

Change of Basis matrix (B): translate V from B's lang to our lang.

Vector in B's lang

$$M \rightarrow B^{-1}MB$$

Transformation M in our coordinate system $\rightarrow$ $B^{-1}MB$ *transformation in B's coordinate system*

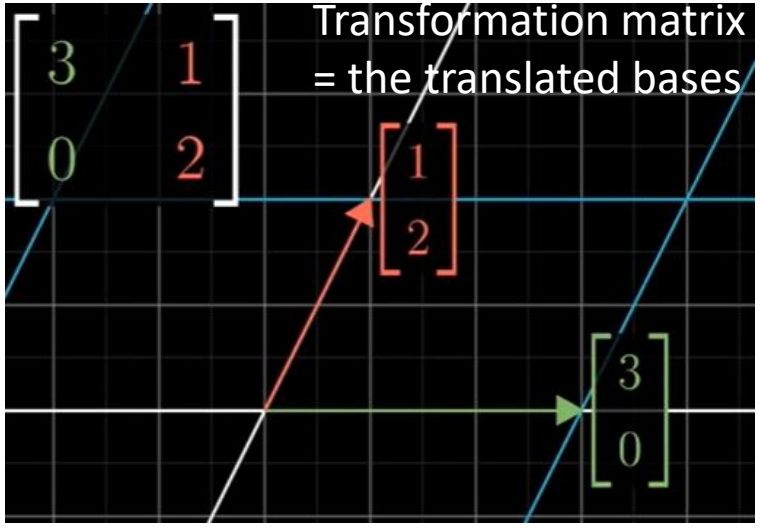
Contents

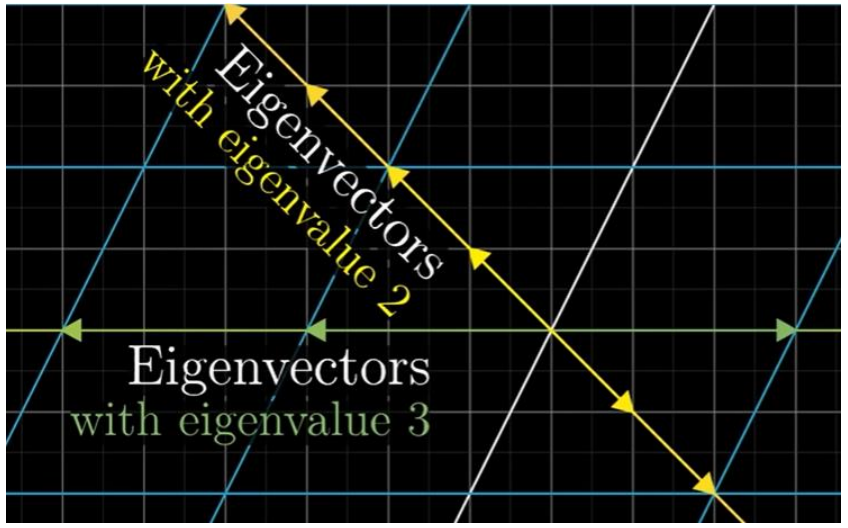
8.2.1 Transformation matrix A's eigenvector v is only scaled when v is multiplied by A

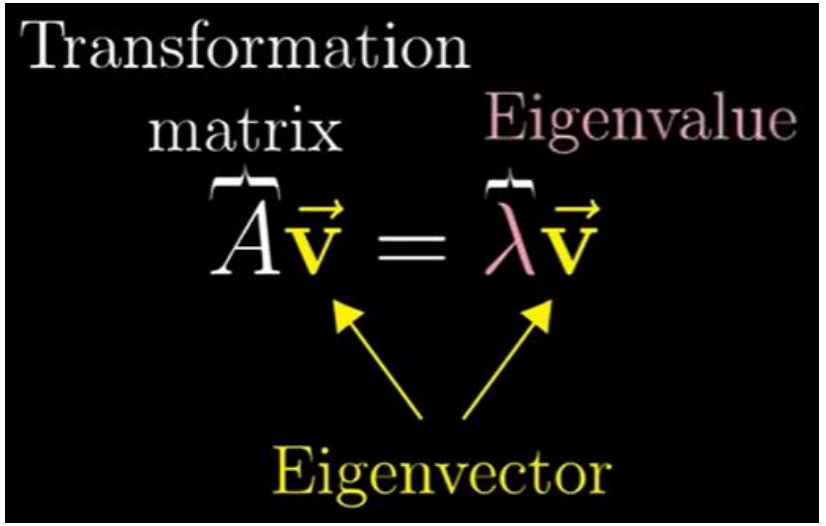
$$(A-\lambda I)\vec{v}=\vec{0}$$

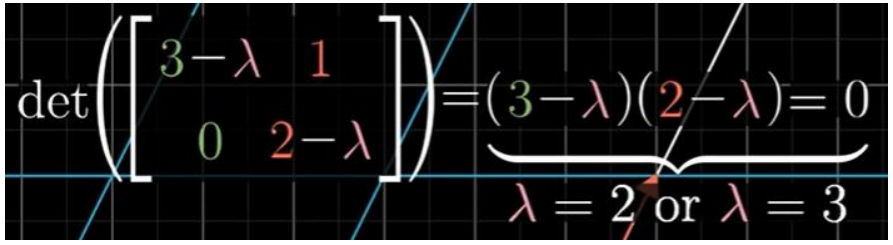
$(A-\lambda I)$ *is linearly dependent*

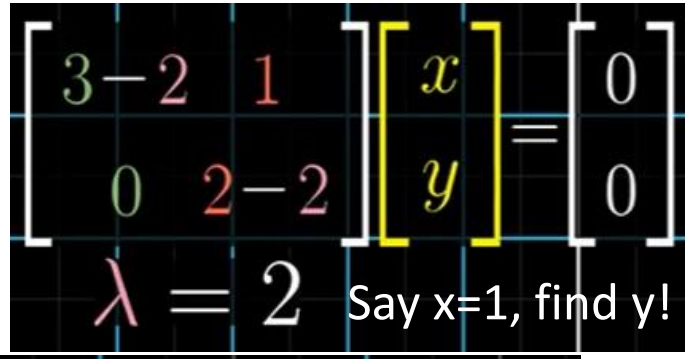
$$\det(A-\lambda I)=0$$



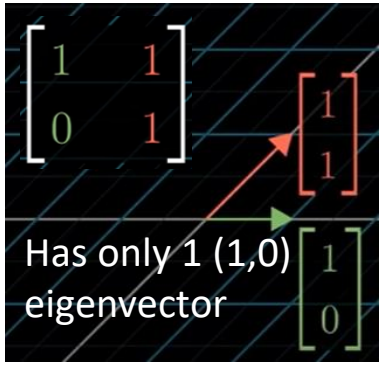


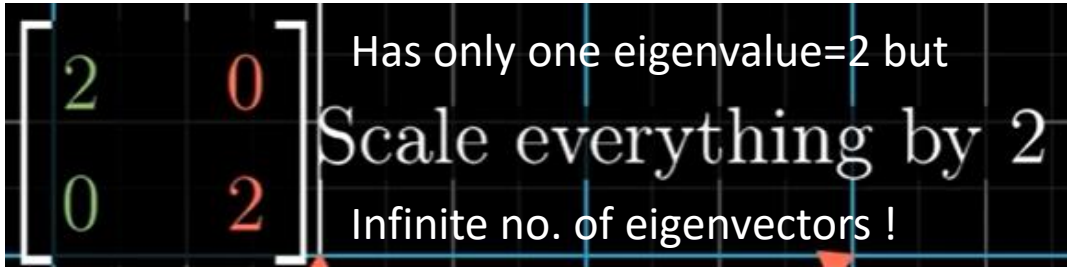






- Some transformations e.g rotation do not have any eigenvector!
- They have complex eigenvalues.
- Some NxN matrices have N-k eigenvectors
- Some matrices have infinite many eigenvectors.





8.2.2 Eigen-decomposition

$$Mx = BDB^{-1}x \text{ where } D = B^{-1}MB = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \text{diagonal matrix}$$

B = eigenvectors of M and are used as the bases i.e Eigenbases
Sometimes it is easier to apply transformation M in B's language

if $b_1 b_2$ are eigenvectors of M: $Mb_1 = \lambda_1 b_1$, $Mb_2 = \lambda_2 b_2$

$$\text{then } MB = M[b_1\ b_2] = [Mb_1\ Mb_2] = [\lambda_1 b_1\ \lambda_2 b_2] = [b_1\ b_2]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$\text{thus } B^{-1}MB = B^{-1}(MB) = B^{-1}[b_1\ b_2]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$= B^{-1}B\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

hence if bases $b_1\ b_2$ ... are in decreasing order of importance (variance) then eign values $\lambda_1\ \lambda_2$ ... are also in decreasing order.

$$\text{Hence } B^{-1}MB = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

where

M = a transformation in our coordinate system
B = Eigenbases of M
$\lambda_1\ \lambda_2$ = eigen values of the eigenvectors of B
$B^{-1}MB$ = transformation M in the B's coordinate system

$$\text{note } MB = B\lambda \text{ where } \lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$\text{Let } B^{-1}MB = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = D \text{ (diagonal matrix)}$$

D is directly not applicable to vectores in our coordinate system hence we have to first convert a given vector to B's language, apply the transformation D and convert back to our language.

thus

$$Mx = BDB^{-1}x$$

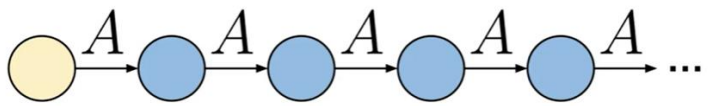
suppose we need to compute $M^{100}\ x$. Its difficult to compute M^{100}

$$\text{but } M^{100}x = (BDB^{-1})(BDB^{-1}) \ldots_{100\ times}\ x$$

$$\text{or } M^{100}x = BDB^{-1}BDB^{-1}BDB^{-1} \ldots_{100\ times}\ x$$

$$\text{or } M^{100}x = BDDD \ldots_{100\ times} B^{-1}x = BD^{100}B^{-1}x$$

$$\text{now } D^{100} = \begin{bmatrix} \lambda_1^{100} & 0 \\ 0 & \lambda_2^{100} \end{bmatrix}$$



if M is a symmetric then its eigenvectors B are orthogonal and $B^{-1} = B^T$
hence if M is symmetric $M = BDB^T$ else $M = BDB^{-1}$

Contents

8.3.1 PCA: Principal components = Eigenvectors of the covariance matrix

Variance is maximized in the direction of the covariance matrix eigenvector having the largest eigenvalue

Variance along this direction = the eigenvalue. These eigenvectors = principal components

To begin with, consider the projection onto a one-dimensional space ($M = 1$). We can define the direction of this space using a D-dimensional vector $\mathbf{u}_1$, which for convenience (and without loss of generality) we shall choose to be a unit vector so that $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$ (note that we are only interested in the direction defined by $\mathbf{u}_1$, not in the magnitude of $\mathbf{u}_1$ itself). Each data point $\mathbf{x}_n$ is then projected onto a scalar value $\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n$. The mean of the projected data is $\mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}$ where $\overline{\mathbf{x}}$ is the sample set mean given by

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{12.1}$$

and the variance of the projected data is given by

$$\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\right\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 \tag{12.2}$$

where $\mathbf{S}$ is the data covariance matrix defined by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}. \tag{12.3}$$

Since co – variance matrix M is symmetric $M = BDB^T$

$$\text{or } M = B\begin{bmatrix}\lambda_1 & 0 & 0 & 0\\ 0 & \lambda_2 & 0 & 0\\ 0 & 0 & \lambda_3 & 0\\ 0 & 0 & 0 & \lambda_4\end{bmatrix}B^T \text{ where } \lambda_1 > \lambda_2 > \lambda_3 > \ldots$$

We now maximize the projected variance $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$ with respect to $\mathbf{u}_1$. Clearly, this has to be a constrained maximization to prevent $\|\mathbf{u}_1\| \to \infty$. The appropriate constraint comes from the normalization condition $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$. To enforce this constraint, we introduce a Lagrange multiplier that we shall denote by λ_1, and then make an unconstrained maximization of

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1\left(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1\right). \tag{12.4}$$

By setting the derivative with respect to $\mathbf{u}_1$ equal to zero, we see that this quantity will have a stationary point when

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \tag{12.5}$$

which says that $\mathbf{u}_1$ must be an eigenvector of $\mathbf{S}$. If we left-multiply by $\mathbf{u}_1^{\mathrm{T}}$ and make use of $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$, we see that the variance is given by

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1 \tag{12.6}$$

and so the variance will be a maximum when we set $\mathbf{u}_1$ equal to the eigenvector having the largest eigenvalue λ_1. This eigenvector is known as the first principal component.

To reduce dims set lower values e.g $\lambda_4 = \lambda_3 = 0$ *then we are left with only* λ_1 *and* λ_2 *and corresponding two eigenbases / principal components*

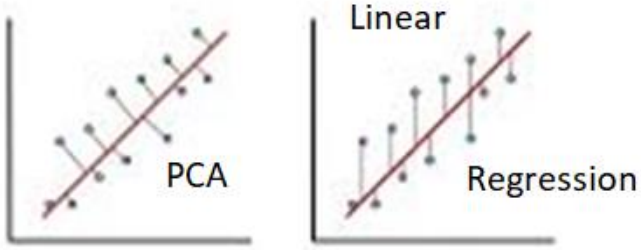
$$\text{thus } M' = B\begin{bmatrix}\lambda_1 & 0 & 0 & 0\\ 0 & \lambda_2 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}B^T = \begin{bmatrix}B_1 & B_2\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}B_1^T\\ B_2^T\end{bmatrix}$$

8.3.2 PCA – Change of axis : new axis = Eigenbases of the covariance matrix

- PCA: moves the center of the coordinate system to the center of data and finds new axes by rotating the coordinate system
- Whitening: Standardize the PCA features! The 1st principal component has more variance, hence standardize them to have unit var.

$$Since\ co-variance\ matrix\ M\ is\ symmetric\ \ M = BDB^T\ \ \left(B^{-1} = B^T\right)$$

$$or\ \ M = B\begin{bmatrix}\lambda_1 & 0 & 0 & 0\\ 0 & \lambda_2 & 0 & 0\\ 0 & 0 & \lambda_3 & 0\\ 0 & 0 & 0 & \lambda_4\end{bmatrix}B^T\ \ where\ \lambda_1 > \lambda_2 > \lambda_3 > \ldots$$



To map X to PCA dims 1st center on origin and rotate the axis :

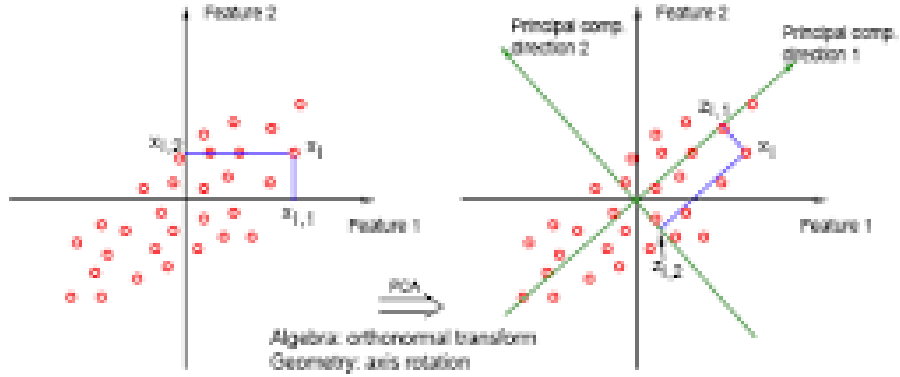
1. *X has to be centered on 0 i.e. let* $X_c = X - \overline{X}$

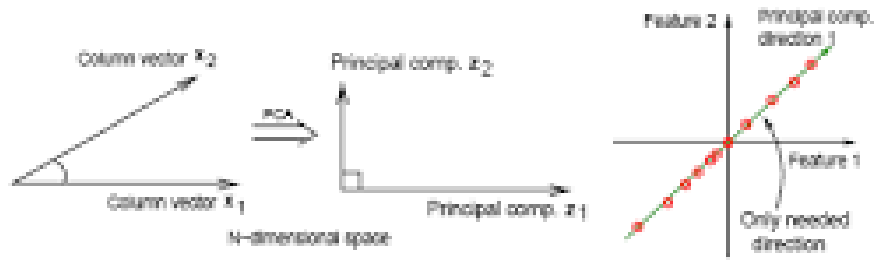
2. *Map* X_c *from our coordinate system to B's coordinate system as* X_{pca} :

$$we\ know\ \ BX_{pca} = X_c$$ ---- To recover Xc from Xpca use this

$$hence\ X_{pca} = B^{-1}X_c = B^TX_c$$ ----- To map Xc to Xpca use this

- Eigenvectors (B) of symmetric matrices are orthogonal. Normalize them to get orthonormal bases.
- Orthonormal matrices imply rotation.
- B-inverse = B-transpose.





8 Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)

Contents

8.4.1 SVD diagonalizes any kind of rectangular matrix !

$SVD: \quad A = U\sum V^T$

SVD = sequence of transformations: 1. rotate/map the vector from our lang to the lang of V, 2. remove/add dimension, stretch along the axis, 3. rotate/map back to our language.

$$if\ A_{m\,n}\ then\ AA^T = S_{m\,m} = S_L = Left\ symmetric\ matrix$$
$$if\ A_{m\,n}\ then\ A^TA = S_{n\,n} = S_R = Right\ symmetric\ matrix$$
$$The\ m\ Eigenvectors\ of\ S_L = Left\ singular\ vectors\ of\ A$$
$$The\ n\ Eigenvectors\ of\ S_R = Right\ singular\ vectors\ of\ A$$
$$S_L and\ S_R\ are\ +ve\ Semi\ Definite\ (PSD)\ i.e\ their\ eigenvalues\ \geqslant 0$$
$$S_L and\ S_R\ have\ same\ non-zero\ eigenvalues.\ Let\ m < n\ then$$
$$1st\ m\ eigenvalues\ of\ both\ S_L\ and\ S_R\ are\ same\ (\lambda_1 \ldots \lambda_m),\ and$$
$$remaining\ n-m\ eigenvalues\ of\ S_R = 0,\ and\ vice-versa.$$
$$SqRoot(eigenvalues) = \sqrt{\lambda_1} \ldots \sqrt{\lambda_m} = Singular\ Values\ of\ A$$

$$SVD: \quad A = U\sum V^T$$
$$U = eigenbases\ of\ S_L\ i.e\ AA^T$$
$$V = eigenbases\ of\ S_R\ i.e\ A^TA,\ \ V^{-1} = V^T$$
$$\sum = diagonal\ matrix\ of\ singular\ values\ of\ A$$

No. of singular values of A (on the diagonal) = Rank of A

$$m \times n\ matrix\ transforms\ a\ vector\ in\ n-dims\ to\ m-dims: M_{m\,n} \times R_{n\,1} \rightarrow R_{m\,1}$$

$$Ax = U\sum V^T x = U\left(\sum (V^T x)\right)$$

V^Tx maps x to the axis of eigenbases

$\sum$ is the diagonal transformation in that axis and is also a dimension adder / remover

U maps it back to our axis

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 4 \\ 8 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \\ 0 \end{bmatrix}$$

dim-adder

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 7 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

dim-eraser !

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$\sum$ Combined

B scale x-axis by 4, scale y-axis by 2

A Dimension Eraser

8.4.2 SVD for low rank approximation: set the lowest diagonal values to 0

$$A_{mn} = U_{mr} \sum_{rr} V^T_{rn}$$

The diagonal matrix can be reduced to a RxR square matrix if rank(A) = R. Now set the lowest diagonal values = 0 to get a low rank K<R approx. of A. Save only the K columns from each of U and V – image compression!

$$A = U \Sigma V^T$$

$eigenvectors\ of\ S_L = AA^T$ | $\sigma_1 = \sqrt{\lambda_1} \ldots \sigma_m = \sqrt{\lambda_m}$ = singular values of A | $eigenvectors\ of\ S_R = A^TA$

$$\begin{bmatrix} & A & \\ & 2\times 3 & \end{bmatrix} = \begin{bmatrix} | & | \\ \vec{u_1} & \vec{u_2} \\ | & | \end{bmatrix}_{2\times 2} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \end{bmatrix}_{2\times 3} \begin{bmatrix} — \vec{v_1}^T — \\ — \vec{v_2}^T — \\ — \vec{v_3}^T — \end{bmatrix}_{3\times 3}$$

$$A = U \Sigma V^T$$

m x m

m x n

diagonal-rectangle: same shape as A

n x n

m x n

$$\begin{bmatrix} & A & \\ & 4\times 2 & \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ \vec{u_1} & \vec{u_2} & \vec{u_3} & \vec{u_4} \\ | & | & | & | \end{bmatrix}_{4\times 4} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \end{bmatrix}_{4\times 2} \begin{bmatrix} — \vec{v_1}^T — \\ — \vec{v_2}^T — \end{bmatrix}_{2\times 2}$$

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i^T}$$

$$\begin{bmatrix} & A & \\ & & \end{bmatrix} = \begin{bmatrix} | & | & | & | & | & | \\ \vec{u_1} & \vec{u_2} & \vec{u_3} & \vec{u_4} & \vec{u_5} & \vec{u_6} \\ | & | & | & | & | & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{bmatrix} \begin{bmatrix} — \vec{v_1}^T — \\ — \vec{v_2}^T — \\ — \vec{v_3}^T — \end{bmatrix}$$

Matrix of rank R: A = sum of R rank 1 matrices :

$$\begin{bmatrix} & A & \\ & & \end{bmatrix} = \sigma_1 \begin{bmatrix} | \\ \vec{u_1} \\ | \end{bmatrix} \begin{bmatrix} — \vec{v_1}^T — \end{bmatrix} + \sigma_2 \begin{bmatrix} | \\ \vec{u_2} \\ | \end{bmatrix} \begin{bmatrix} — \vec{v_2}^T — \end{bmatrix} + \sigma_3 \begin{bmatrix} | \\ \vec{u_3} \\ | \end{bmatrix} \begin{bmatrix} — \vec{v_3}^T — \end{bmatrix}$$

Optimal low rank approximation of A = Nearest matrix to A in lower rank = set the lowest diagonal elements = 0 above. -> image compression!

$$Let\ A\ has\ rank\ 3 \quad then\ approx\ A_{rank\,2} = U \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 = 0 \\ 0 & 0 & 0 \end{bmatrix} V^T$$

We can safely remove those rows / cols which are all 0s from $\sum$

Thus $\sum$ *is an* $R \times R$ *square matrix if* $rank(A) = R$

Thus we can also retain the first R columns from U, hence U is $M \times R$ *and retain the first R rows from* V^T*, hence* V^T *is* $R \times N$

$$Thus\ A_{mn} = U_{mr} \sum_{rr} V^T_{rn}$$

8.4.3 SVD: dimensionality reduction

- Set lowest singular values on the diagonal = 0 to reduce the dimensions.
- PCA / SVD embeddings bring out meaningful patterns/clusters in data. Clustering – use PCA/SVD

$$X_{m\,n} - \overline{X}(column\ mean) = X_c = U_{m\,r} \sum_{r\,r} V^T_{r\,n}$$

	item1	item2	item3
Cust1			
Cust2		Xc	
Cust3			
Cust4			
cust5			M x N

=

	dim1	dim2	dim3
Cust1	Loadings on the principal axis		
Cust2		U	
Cust3			
Cust4			
cust5			M x R

Σ

s1		
	s2	
		s3

R x R

	item1	item2	item3
dim1	Principal axis/component for Cust		
dim2		V^T	
dim3			R x N

Dimensionality reduction: Set lowest singular values on the diagonal = 0

	item1	item2	item3
Cust1			
Cust2		Xc	
Cust3			
Cust4			
cust5			M x N

=

	dim1	dim2	dim3
Cust1			
Cust2		U	
Cust3			
Cust4			
cust5		M x R-1	

Σ

s1		
	s2	
		0

R-1 x R-1

	item1	item2	item3
dim1		V^T	
dim2			
dim3			

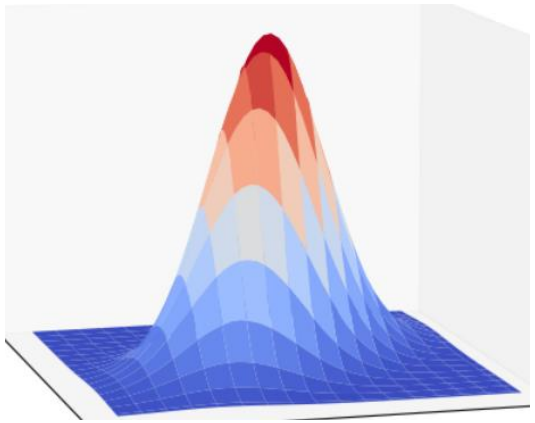
R-1 x N

My two favorite books from where I re-used most of the material

References

- Pattern-Recognition-and-Machine-Learning - Christopher M. Bishop, available at https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

- Introduction to Probability and statistics for engineers and scientists - Sheldon M. Ross, available at https://minerva.it.manchester.ac.uk/~saralees/statbook3.pdf



Joint probability
density Pr(x1, x2)