

FUNDAMENTALS OF LEARNING AND INFORMATION PROCESSING

SESSION 09: PROBABILITY THEORY (V)

Dr Gang Li

Deakin University, Geelong, Australia

2018-11-06

Inequalities	3
<i>Cauchy-Schwarz</i> Inequality	4
<i>Jensen's</i> Inequality.	5
Concentration Inequalities	6
<i>Markov</i> Inequality	8
<i>Chebyshev's</i> Inequality	11
<i>Chernoff's</i> Bounds	13
Fundamental Inequality	14
<i>Hoeffding's</i> Inequality	15
Law of Large Numbers	17
The Weak Law of Large Numbers.	20
The Strong Law of Large Numbers.	21
Central Limit Theorem	25
Central Limit Theorem	26
Gaussian Distribution	32
Distribution of <i>Errors</i>	33
Univariate Gaussian	34
Multivariate Gaussian.	35
Affine Property.	37
Geometric Intuition	38

Table of Content

Inequalities

Cauchy-Schwarz Inequality

Jensen’s Inequality

Concentration Inequalities

Markov Inequality

Chebyshev’s Inequality

Chernoff’s Bounds

Fundamental Inequality

Hoeffding’s Inequality

Law of Large Numbers

The Weak Law of Large Numbers

The Strong Law of Large Numbers

Central Limit Theorem

Central Limit Theorem

Gaussian Distribution

Distribution of Errors

Univariate Gaussian

Multivariate Gaussian

Affine Property

Geometric Intuition

(None)-45903a7 (2018-11-06) – 2 / 41

Inequalities

3 / 41

Cauchy-Schwarz Inequality

Defn

The *Cauchy-Schwarz inequality* states that

■ When X and Y are vectors with the same dimension, we have

$|X \cdot Y| \leq |X| \times |Y|$

the two sides are equal if and only if x and y are linearly dependent.

■ When X and Y are random variables, then

$|E(X \cdot Y)| \leq \sqrt{E(X^2)E(Y^2)}$

Intuition.

Look at the correlation definition:

$-1 \leq \rho(X, Y)$

$= \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$

$= \frac{E(X \cdot Y)}{E(X)E(Y)}$

≤ 1

□

Random version.

Look at two expectations:

■ $E(aX + bY)^2 = a^2E(X^2) + b^2E(Y^2) + 2abE(XY) \geq 0$

■ $E(aX - bY)^2 = a^2E(X^2) + b^2E(Y^2) - 2abE(XY) \geq 0$

Now let $a^2 = E(Y^2)$ and $b^2 = E(X^2)$, we have

■ $2abE(X \cdot Y) \geq -2a^2b^2$

■ $2abE(X \cdot Y) \leq 2a^2b^2$

Dividing by $2ab$ results in $-\sqrt{E(X^2)E(Y^2)} \leq E(X \cdot Y) \leq \sqrt{E(X^2)E(Y^2)}$.

□

(None)-45903a7 (2018-11-06) – 4 / 41

Jensen’s Inequality

Defn If f is a convex function and X is a random variable, the *Jensen’s inequality* states that $E(f(X)) \geq f(E(X))$

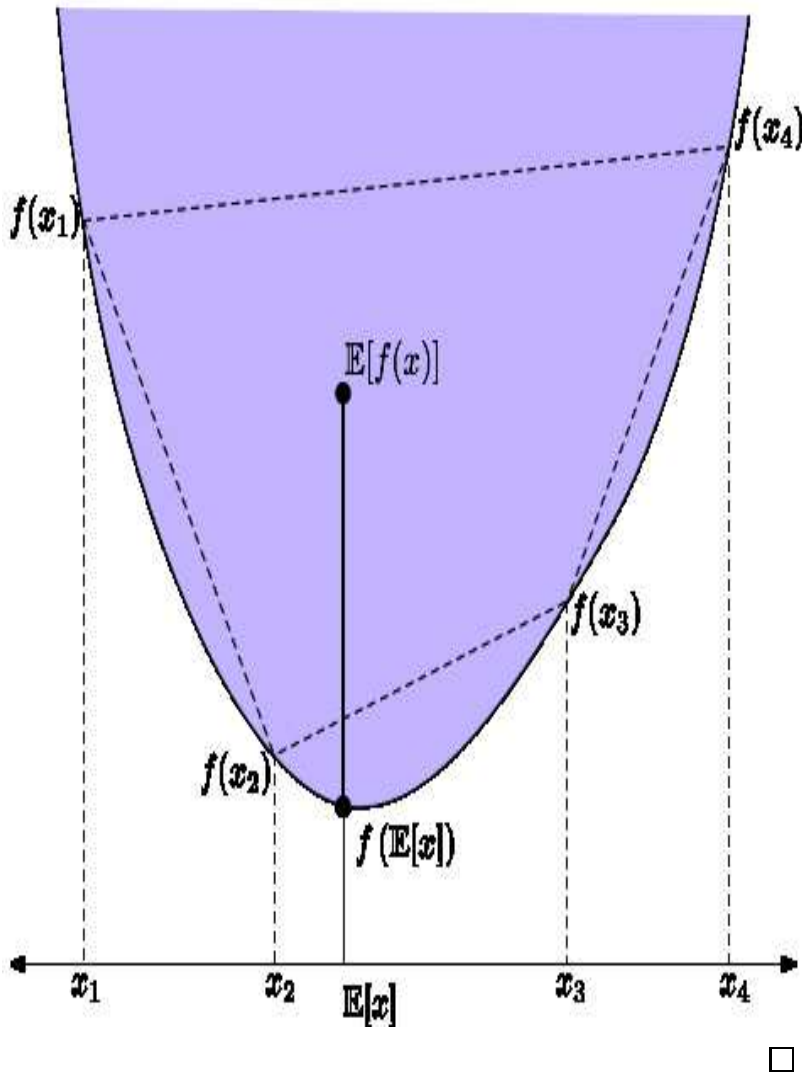
- If f is strictly convex, the equality implies that $X = E(X)$ with probability 1, namely X is a constant.
- We can define the *Jensen Gap*: $J_f(x) = E(f(X)) - f(E(X))$.

Intuition.

- The points $(x_i, f(x_i))$ form the vertices of a polygon which, must also be convex and lie within the epigraph (the blue shaded area above f).
- Since the $\sum p_i = 1$, the expected value of the random variable $(x, f(x))$ given by $E[(x, f(x))] = \sum_{i=1}^n p_i(x_i, f(x_i))$ is a convex combination and so must also lie within the dashed polygon. Since $E[(x, f(x))] = (E[x], E[f(x)])$, it must lie above $f(E[x])$.

Examples:

- For a convex function $f(X) = X^2$, we have $E(X^2) \geq (E(X))^2$.
- For a function $f(X) = \frac{1}{X}$ for positive X , we have $E(\frac{1}{X}) \geq \frac{1}{E(X)}$.
- For a function $f(X) = \ln(X)$, we have $E(\ln(x)) \leq \ln E(X)$.



Concentration Inequalities

Concentration Inequalities

Defn Let X_1, \dots, X_m be an i.i.d. sequence of random variables, and let μ be their mean. The SLLN states that when $m \rightarrow \infty$, $\frac{1}{m} \sum_{i=1}^m X_i$ converges to μ with probability 1. *Measure concentration inequalities* quantify the deviation of the empirical average from the expectation when m is finite.

Typical Concentration Inequalities. Concentration inequalities give probability bounds for a random variable to be concentrated around its mean, or for it to deviate from its mean or some other values.

- Markov’s Inequality
- Chebyshev’s Inequality
- Chernoff’s Bounds
- Hoeffding’s Inequality
- Bennet’s and Bernsein’s inequalities
- Slud’s inequality
- Concentration of χ^2 Variables

Markov Inequality

Defn

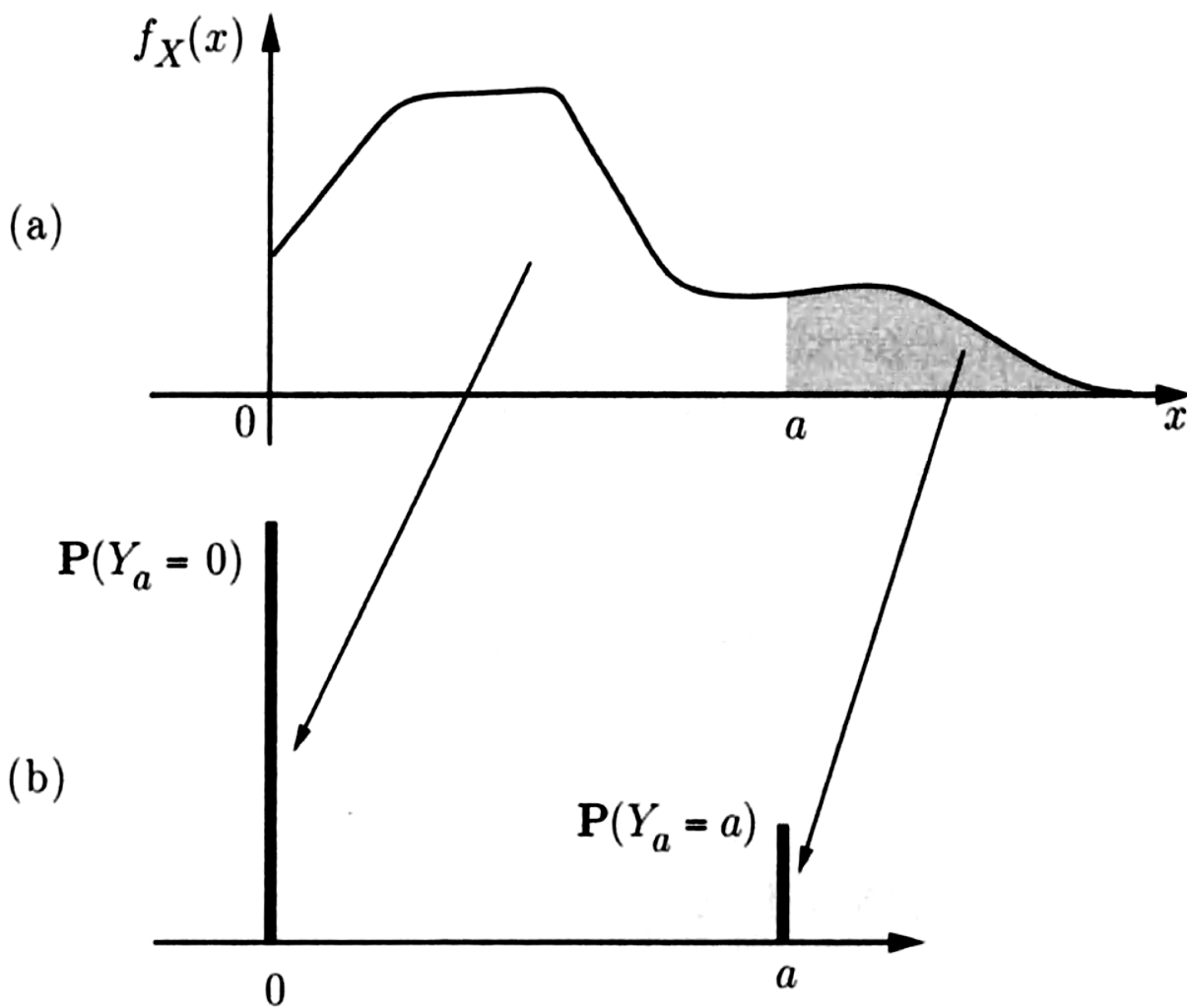
If a random variable X , the *Markov inequality* states that
$$P(|X|\geq a) \leq \frac{E(|X|)}{a}, \text{ for all } a > 0$$

Applications.

- Is it possible that at least 95% of people are younger than the average age?
- Is it possible that at least 50% of people are older than twice the average age?

□

Intuition.



- Since all the mass is shifted to the left, the expectation can only decrease from $E(X)$, hence $E(X) \geq E(Y_a) = aP(Y_a = a) = aP(X \geq a)$
- Or we can use $E(X) = \sum_x xP(x) \geq \sum_{x \geq a} xP(x) \geq \sum_{x \geq a} aP(x) = aP(X \geq a)$

□

Proof. Let us fix a positive number a and consider an indicator $I_{|X|\geq a}$.

- we have $aI_{|X|\geq a} \leq X$ always true, then $I_{|X|\geq a} \leq \frac{|X|}{a}$.
- we take expectation: $P(|X|\geq a) = E(I_{|X|\geq a}) \leq \frac{E(|X|)}{a}$

□

Applications. Let $a = kE(|X|)$. From *Markov inequality*, we have $P(|X|\geq kE(|X|)) \leq \frac{1}{k}$

- $P(|X|\geq 2E(|X|)) \leq 0.50$
- $P(|X|\geq 3E(|X|)) \leq 0.33$

□

Markov's Inequality

Lemma

Let a random variable X that takes values in $[0, 1]$ assume $E[X] = \mu$. Then for any $a \in (0, 1)$,
$$P[X > 1 - a] \geq \frac{\mu - (1 - a)}{a}$$

This also implies that

$$P[X > a] \geq \frac{\mu - a}{1 - a} \geq \mu - a$$

Proof.

- Let $Y = 1 - X$, then Y is a nonnegative random variable with $E[Y] = 1 - E[X] = 1 - \mu$.
- Applying Markov's inequality on Y we obtain

$$P[X \leq 1 - a] = P[1 - X \geq a] = P[Y \geq a] \leq \frac{1 - \mu}{a}$$

- Accordingly,

$$P[X \geq 1 - a] \geq 1 - \frac{1 - \mu}{a} = \frac{\mu - (1 - a)}{a}$$

□

Borel-Cantelli Lemma

LM Let X_i with $i \in N$ be a sequence of events such that $\sum_{i=0}^\infty P(X_i) < \infty$. Then, *almost surely*, only finitely many of these events occur.

Proof.

- From Markov’s inequality, $\forall t > 0$,

$$P[\sum_{i=0}^\infty 1_{X_i} \geq t] \leq \frac{1}{t} \sum_{i=0}^\infty P(X_i)$$

- Then,

$$P[\sum_{i=0}^\infty 1_{X_i} \leq \infty] = 1 - P[\sum_{i=0}^\infty 1_{X_i} \geq \infty] = 1 - \lim_{t \rightarrow \infty} P[\sum_{i=0}^\infty 1_{X_i} \geq t] = 1$$

□

Chebyshev’s Inequality

Defn If a random variable X with mean μ and variance σ^2 , the *Chebyshev inequality* states that

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}, \text{ for all } a > 0$$

Proof.

$$P(|X - \mu| \geq a) = P((X - \mu)^2 \geq a^2) \leq \frac{E(X - \mu)^2}{a^2} = \frac{\sigma^2}{a^2}$$

□

Proof. Let $a = k\sigma$ with $k > 0$. We have $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$. The probability that X takes a value between $\mu \pm k\sigma$ is at least $1 - \frac{1}{k^2}$.

- when $X \in [a, b]$, we can claim that $\sigma^2 \leq \frac{(b-a)^2}{4}$, and use $P(|X - \mu| \geq a) \leq \frac{(b-a)^2}{4a^2}$.

$$\begin{aligned} E(X - \gamma)^2 &= E(X^2) - 2E(X)\gamma + \gamma^2 \\ &\geq E(X - E(X))^2 \\ &= \sigma^2 \end{aligned}$$

If we take $\gamma = \frac{a+b}{2}$, we obtain:

$$\begin{aligned} \sigma^2 &\leq E(X - \frac{a+b}{2})^2 \\ &= E((X - a)(X - b) + \frac{(b-a)^2}{4}) \\ &\leq \frac{(b-a)^2}{4} \end{aligned}$$

□

Application. Roll a pair of fair dice n times; can we give a good estimate of the total value of the n rolls?

- We have the mean $7n$ and the variance $\frac{35}{6}n$.
- Let $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} = 0.01$. Then we can take $a = 10\sigma$.
- Accordingly the probability of $|X - \mu| \geq 10\sigma$ will be less than 1%, so X must be within $\mu - 10\sigma$ and $\mu + 10\sigma$, namely $[7n - 10\sqrt{\frac{35}{6}n}, 7n + 10\sqrt{\frac{35}{6}n}]$
- So the odds are better than 99 to 1 that the sum be roughly between $6.976M$ and $7.024M$ after $1M$ rolls.

□

Chebyshev’s Inequality

Lemma Let X_1, \dots, X_m be a sequence of i.i.d. random variables and assume $E[X] = \mu$ and $Var[X] \leq 1$. Then, for any $\sigma \in (0, 1)$, with probability at least $1 - \sigma$ we have

$$|\frac{1}{m} \sum_{i=1}^m X_i - \mu| \leq \sqrt{\frac{1}{\sigma m}}$$

Proof.

- Applying Chebyshev’s inequality we obtain

$$P[|\frac{1}{m} \sum_{i=1}^m X_i - \mu| > a] \leq \frac{Var[X]}{ma^2} \leq \frac{1}{ma^2} = \sigma$$

- The proof follows by denoting the right-hand side σ and solving for a .

□

Chernoff’s Bounds

Defn Let X_1, \dots, X_m be independent Bernoulli variables where $\forall i, P[X_i = 1] = p_i$ and $P[X_i = 0] = 1 - p_i$. Let $p = \sum_{i=1}^m p_i$ and $X = \sum_{i=1}^m X_i$. Then $\forall \sigma > 0$,

$$P[X > (1 + \sigma)p] \leq e^{-h(\sigma)p}$$

where $h(\sigma) = (1 + \sigma)\log(1 + \sigma) - \sigma$.

- Using $h(\sigma) \geq \frac{\sigma^2}{2+2\sigma/3}$, we have $P[X > (1 + \sigma)p] \leq e^{-p \frac{\sigma^2}{2+2\sigma/3}}$

Proof.

- From the monotonicity of the exponential function and Markov’s inequality,

$$P[X > (1 + \sigma)p] = P[e^{tX} > e^{t(1+\sigma)p}] \leq \frac{E[e^{tX}]}{e^{t(1+\sigma)p}}$$

- As $1 + x \leq e^x, E[e^{tX}] = E[e^{t\sum_i X_i}] = E[\prod_i e^{tX_i}] = \prod_i E[e^{tX_i}] = \prod_i (p_i e^t + (1 - p_i)e^0) = \prod_i (1 + p_i(e^t - 1)) \leq \prod_i e^{p_i(e^t - 1)} = e^{\sum_i p_i(e^t - 1)}$
- Choose $t = \log(1 + \sigma)$, we can prove the Bounds.

□

Fundamental Inequality

Defn

If a random variable X , a real number t and a strictly monotonically increasing nonnegative-value function $f(X)$, we have

$$P(X \geq t) = P(f(X) \geq f(t)) \leq \frac{E(f(X))}{f(t)}$$

Applications.

Markov Inequality Let $f(X) = X$, we have it.
Chebyshev Inequality Let $X = |Y - E(Y)|$, and $f(X) = X^2$, we have it.
Chernoff Bound Let $f(X) = e^{sX}$, we have

$$P(X \geq t) = P(e^{sX} \geq e^{st}) \leq \frac{E(e^{sX})}{e^{st}}$$

□

Hoeffding’s Inequality

Lemma

Let X be a random variable with $E[X] = 0$ and $a \leq X \leq b$ with $b > a$. Then $\forall t > 0$, the following inequality holds:

$$E[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{s}}$$

Proof.

Lemma See details of any textbook.

□

Hoeffding’s Inequality

Lemma Let X_1, \dots, X_m be independent random variables with X_i taking values in $[a_i, b_i]$ for all $i \in [1, m]$. Then $\forall \epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^m X_i$: $P[S_m - E[S_m] \geq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}}$ and $P[S_m - E[S_m] \leq -\epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}}$

Proof.

- Using the Chernoff bounds and Lemma, we have

$$\begin{aligned} P[S_m - E[S_m] \geq \epsilon] &\leq e^{-t\epsilon} E[e^{t(S_m - E[S_m])}] \\ &= \prod_{i=1}^m e^{-t\epsilon} E[e^{t(X_i - E[X_i])}] \\ &\leq \prod_{i=1}^m e^{-t\epsilon} e^{t^2(b_i - a_i)^2/8} \\ &= e^{-t\epsilon} e^{t^2 \sum_{i=1}^m (b_i - a_i)^2/8} \\ &\leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}} \end{aligned}$$

- The second statement can be proven in a similar way.

□

Two statements can be combined into one:

$$P[|S_m - E[S_m]| \geq \epsilon] \leq 2e^{\frac{-2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}}$$

Law of Large Numbers

Convergence of Random Variables

Defn Let Y_1, Y_2, \dots be a sequence of random variables, and a be a real number. The sequence Y_n *converges to a in probability* (**ALMOST ALL**), if $\forall \epsilon > 0$, we have

$$\lim_{n \rightarrow \infty} P(|Y_n - a| \geq \epsilon) = 0$$

Notes.

- The probability that this sequence deviates a is asymptotically decreasing and approaches 0 but never actually attains 0.
- $\forall \epsilon > 0$, and $\forall \delta > 0$, $\exists n_0$ such that $P(|Y_n - a| \geq \epsilon) \leq \delta$, $\forall n \geq n_0$.
 - ◆ ϵ is the *accuracy* level;
 - ◆ δ is the *confidence* level.

□

Defn Let Y_1, Y_2, \dots be a sequence of random variables, and a be a real number. The sequence Y_n *converges to a with probability 1* (**ALMOST SURELY**), if

$$P(\lim_{n \rightarrow \infty} Y_n = a) = 1$$

Notes.

- This sequence will equal a asymptotically but you cannot predict at what point it will happen.

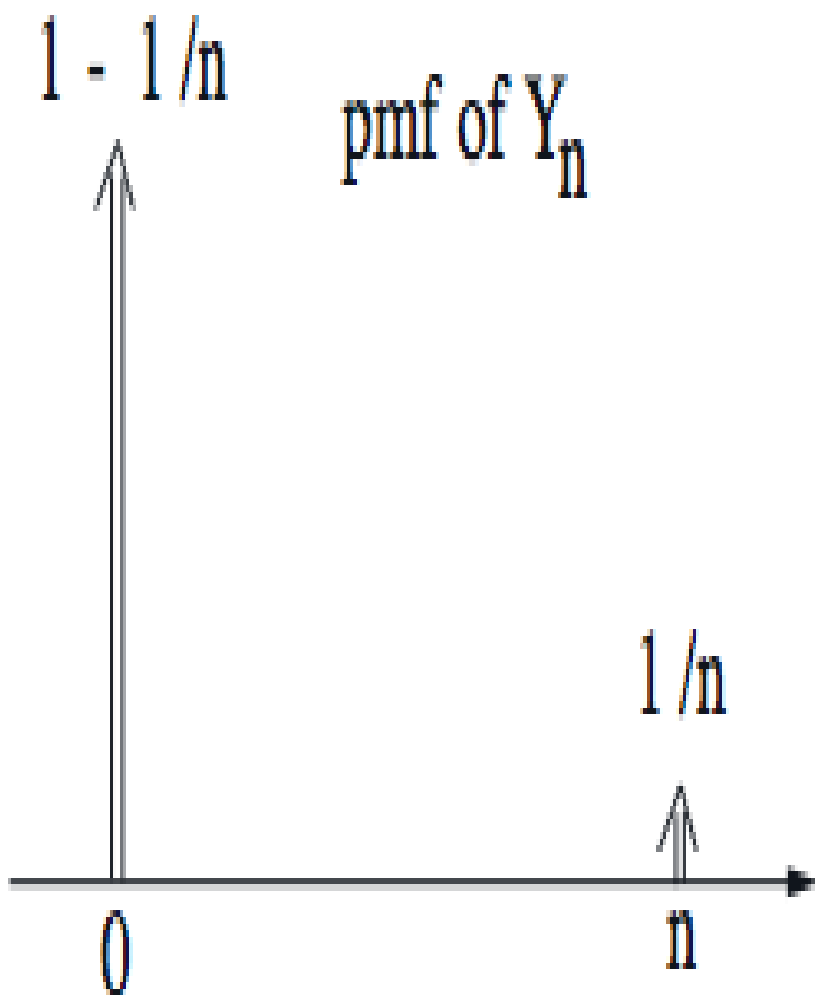
□

Convergence of Random Variables

Prob.

Consider a sequence of discrete random variables Y_n with the following distribution:

$$P(Y_n = y) = \begin{cases} 1 - \frac{1}{n} & \text{for } y = 0 \\ \frac{1}{n} & \text{for } y = n^2 \\ 0 & \text{elsewhere} \end{cases}$$



Notes.

- For every $\epsilon > 0$, we have $\lim_{n \rightarrow \infty} P(|Y_n - 0| \geq \epsilon) = \lim_{n \rightarrow \infty} \frac{1}{n} = 0$. so Y_n *converges to zero in probability*.
- $E(Y_n) = 0 \times (1 - \frac{1}{n}) + n^2 \times \frac{1}{n} = n$.
- $E(Y_n^2) = 0 \times (1 - \frac{1}{n}) + n^4 \times \frac{1}{n} = n^3$, which goes infinity as n increases.

□

The Weak Law of Large Numbers

Defn

Let X_1, X_2, \dots be *independent identically distributed* (i.i.d.) random variables with mean μ and variance σ^2 . For every $\epsilon > 0$, we have *The Weak Law of Large Numbers*:

$$P(|\frac{X_1 + \dots + X_n}{n} - \mu| \geq \epsilon) \rightarrow 0, \text{ as } n \rightarrow \infty$$

This convergence is *Convergence in Probability*.

Proof. Let $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$. We have $E(\bar{X}_n) = \frac{n\mu}{n} = \mu$, using independence, we have $var(\bar{X}_n) = \frac{var(X_1) + \dots + var(X_n)}{n^2} = \frac{\sigma^2}{n}$. By *Chebyshev Inequality*, we have

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}, \text{ for any } \epsilon > 0$$

As $n \rightarrow \infty$, the right hand side goes to zero.

□

The Strong Law of Large Numbers

Defn Let X_1, X_2, \dots be a sequence of *independent identically distributed* (i.i.d.) random variables with mean μ . *The Strong Law of Large Numbers* states that: the sequence of sample mean $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$ *converges with probability 1* to μ .

$$P(\lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \mu) = 1$$

Notes. According to the *strong law*: with probability 1, \bar{X}_n converges to μ .

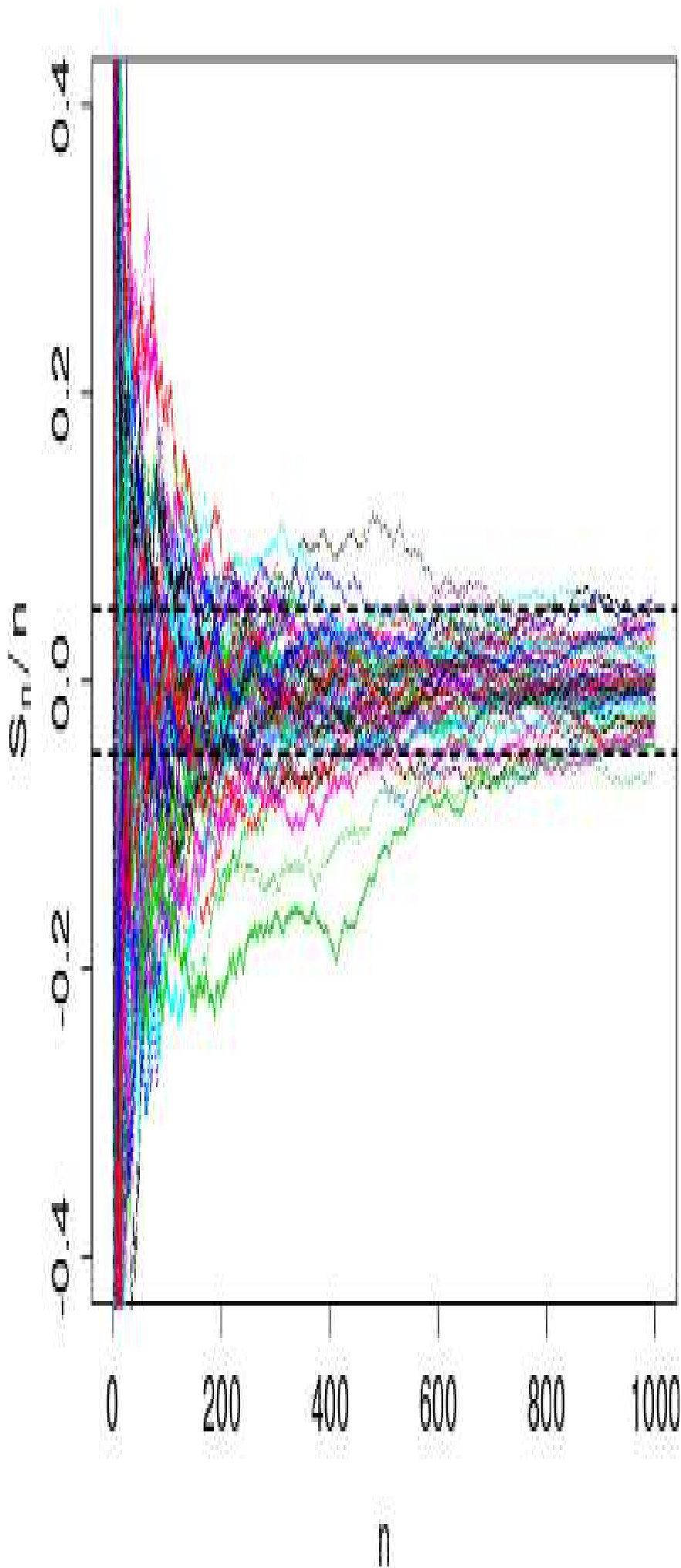
- This implies that for any given $\epsilon > 0$, the probability that the difference $|\bar{X}_n - \mu|$ will exceed ϵ an infinite number of times is equal to zero.
- While the *weak law* states that the probability $P(|\bar{X}_n - \mu| \geq \epsilon)$ of a significant deviation of \bar{X}_n from μ goes to zero as $n \rightarrow \infty$. Though, for any finite n , this probability can be positive and it is conceivable that once in a while, even if infrequently, \bar{X}_n deviates significantly from μ .

□

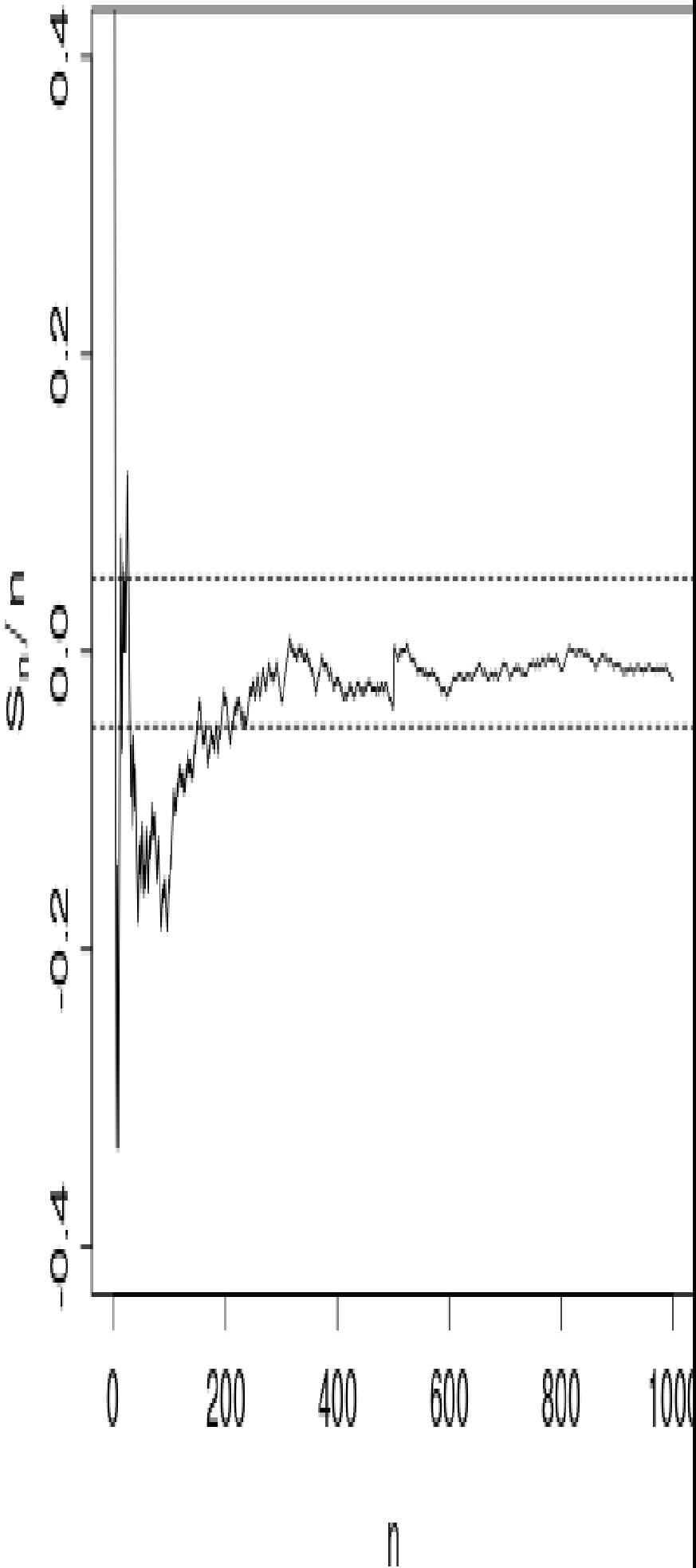
(None)-45903a7 (2018-11-06) – 21 / 41

WLLN vs SLLN

Weak Law of Large Numbers



Strong Law of Large Numbers



(None)-45903a7 (2018-11-06) – 22 / 41

Convergence of the Sample Mean

Defn Let X_1, X_2, \dots be *independent identically distributed* (i.i.d.) random variables with mean μ and variance σ^2 .

$$M_n = \frac{X_1 + \dots + X_n}{n}$$

- $E(M_n) = \frac{E(X_1) + \dots + E(X_n)}{n} = \mu$
- $Var(M_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$
- From *Chebyshev* Inequality, we have $P(|M_n - \mu| \geq \epsilon) \leq \frac{Var(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$
- This means $\forall \epsilon > 0$, we have $P(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$, which approximates to 0 as $n \rightarrow \infty$, namely M_n converges in probability to μ .

Different scalings of M_n . Look at three variants of their sum:

$S_n = X_1 + \dots + X_n$ variance $n\sigma^2$.
 $M_n = \frac{S_n}{n}$ variance $\frac{\sigma^2}{n}$, and converges in probability to μ .
 $\frac{S_n}{\sqrt{n}}$ constant variance. Asymptotic shape?

WLL Application (Sample Size)

Ex. Suppose f is the probability of population that with a characteristics. For the i -th randomly selected person polled

$$X_i = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if no} \end{cases}$$

We take n randomly selected persons, with $M_n = \frac{X_1 + \dots + X_n}{n}$ fraction of “yes”. How large the n needs to be, in order to have 95% confidence of $\leq 1\%$ error:

$$P(|M_n - f| \geq 0.01) \leq 0.05$$

Answer. Use Chebyshev’s inequality:

- $$P(|M_n - f| \geq 0.01) \leq \frac{\sigma_{M_n}^2}{(0.01)^2} = \frac{\sigma_x^2}{n(0.01)^2} = \frac{p_x(1 - p_x)}{n(0.01)^2} \leq \frac{1}{4n(0.01)^2}$$
- If $n = 50,000$, then we have $P(|M_n - f| \geq 0.01) \leq 0.05$.
-

CLT

Central Limit Theorem

Let X_1, X_2, \dots be a sequence of *independent identically distributed* (i.i.d.) random variables with common mean μ and variance σ^2 , define
$$Z_n = \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}},$$
which with $E(Z_n) = 0$ and $var(Z_n) = 1$

Then the CDF of Z_n converges to the standard normal CDF
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx$$

In the sense that $\lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z)$, for every z .

Intuition.

Look at three variants of the sum of random variables

- $S_n = X_1 + \dots + X_n$, with variance $n\sigma^2$
- $\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$, with variance σ^2/n . Converges in probability to $E(X)$ (WLLN)
- $\frac{S_n}{\sqrt{n}}$ with constant variance σ^2 .

Notes.

It is universal, only *means*, *variances* matter

- CDF of Z_n converges to normal CDF. Not a statement about convergence of *PDF*s or *PMF*s
- Treat Z_n as if normal, also treat S_n as if normal

Lognormal.

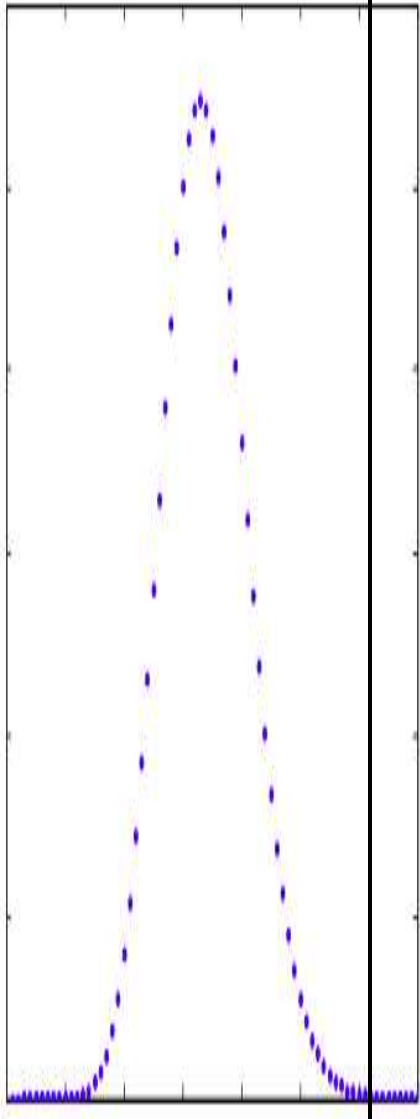
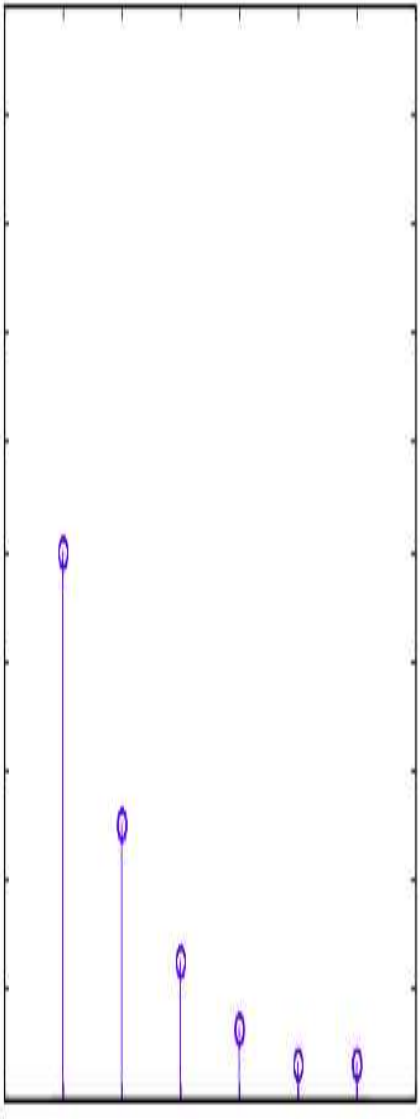
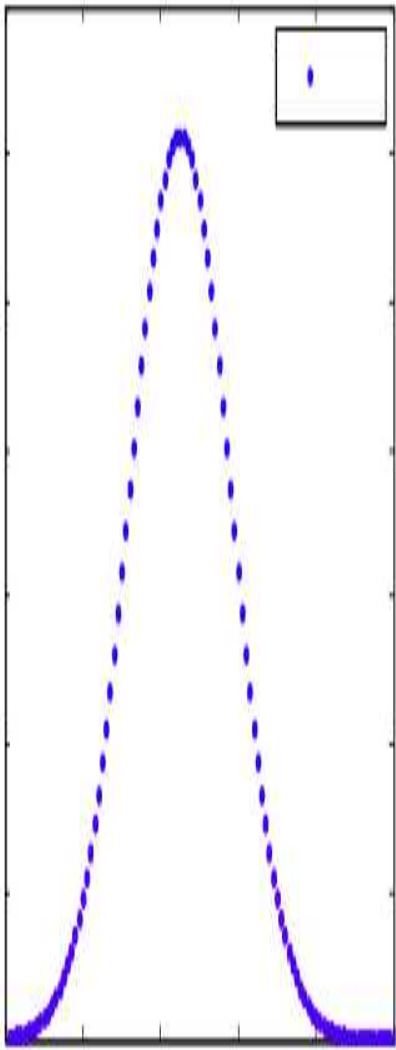
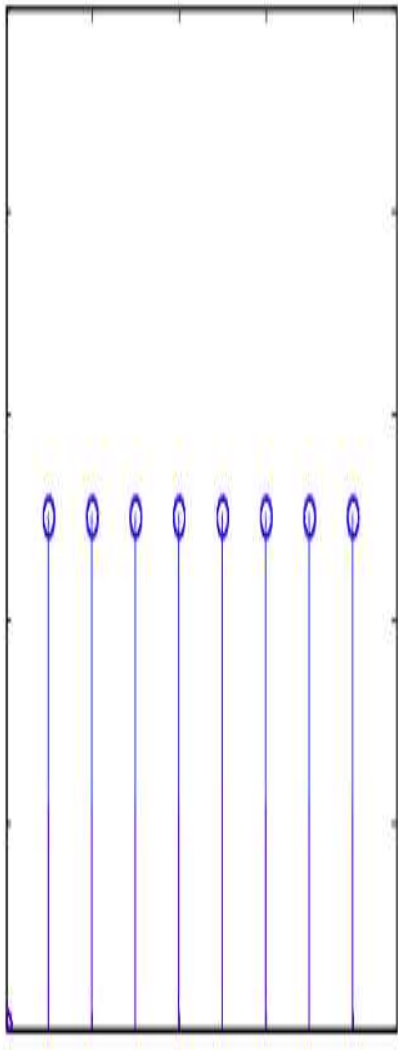
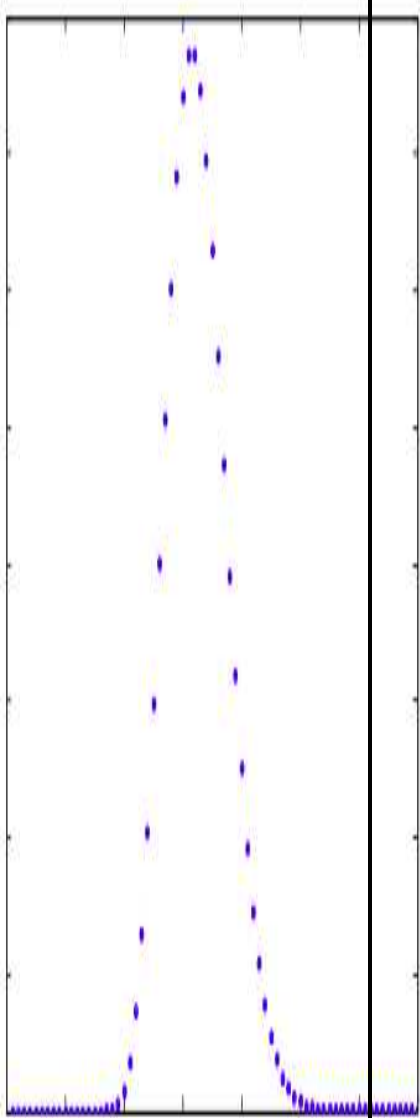
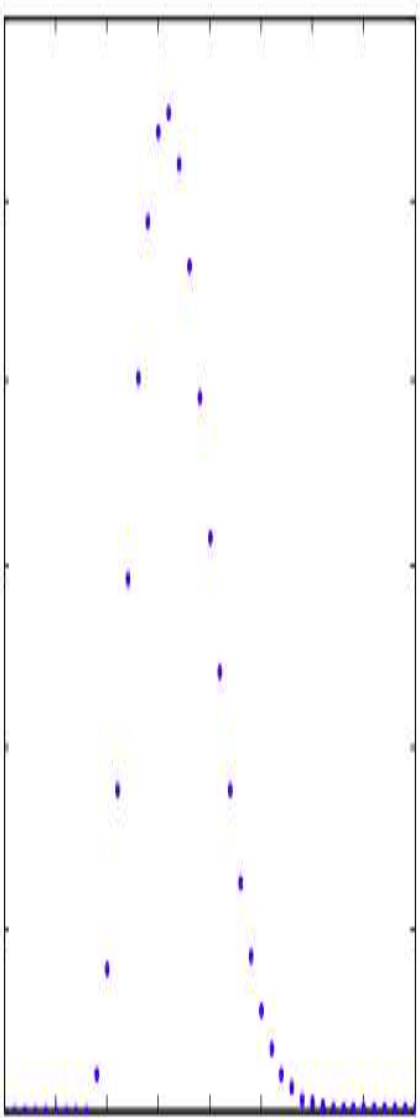
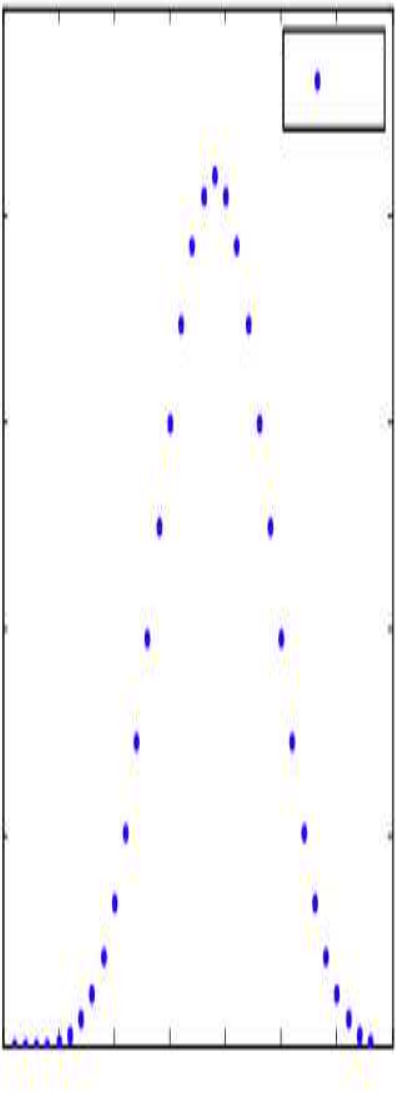
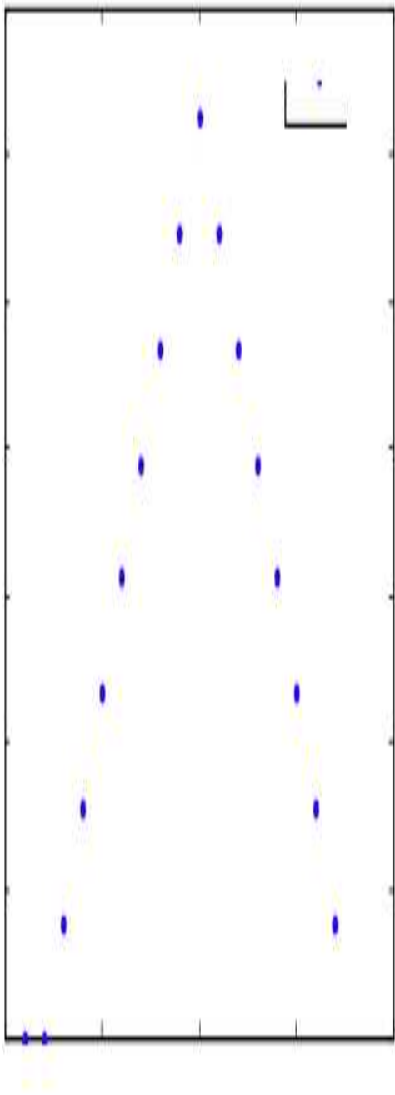
CLT says that the sum of many *independent*, *additive* effects is approximately normally distributed.

- For example, the effects of dominant and recessive genes act more like *max* and *min* than *addition*.
- if effects are independent but *multiplicative* rather than additive, the result may be approximately *log-normal* rather than normal
 - ◆ If X follows $\text{lognormal}(\mu, \sigma)$ distribution, then $\log(X) \sim N(\mu, \sigma)$.

(None)-45903a7 (2018-11-06) – 26 / 41

12

Central Limit Theorem



CLT Applications (I: Sample Size)

Ex.

Suppose f is the probability of population that with a characteristics. For the i -th randomly selected person polled

$$X_i = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if no} \end{cases}$$

We take n randomly selected persons, with $M_n = \frac{X_1 + \dots + X_n}{n}$ fraction of “yes”. How large the n needs to be, in order to have 95% confidence of $\leq 1\%$ error:

$$P(|M_n - f| \geq 0.01) \leq 0.05$$

Answer. Event of Interest:

$$P(|M_n - f| \geq 0.01) \approx P\left(\left|\frac{X_1 + \dots + X_n - nf}{\sqrt{n}\sigma}\right| \geq \frac{0.01\sqrt{n}}{\sigma}\right) \leq P(|Z| \geq 0.02\sqrt{n})$$

■ Find a $n = 9604$ that makes $P(|Z| \geq 0.02\sqrt{n})$.

□

CLT Applications (II)

Ex.

Suppose you are running a *Cloud Storage* service, such as *Baidu Pan*, *Dropbox*, etc. If you sold this service to 100 customers, whose storage size are independent random variables uniformly distributed between 5G to 50G Bytes. What is the probability that the total storage size will exceed 3T Bytes?

CLT. It is not easy to calculate the CDF of the total storage size and the desired probability, but we can use CLT to approximate it.

■ We want to calculate $P(S_{100} > 3000)$, where S_{100} is the sum of the storage size for 100 customers. The mean and the variance of the wight of a single customer is: $\mu = \frac{5+50}{2} = 27.5$, and $\sigma^2 = \frac{(50-5)^2}{12} = 168.75$.

■ Based on CLT, we calculate the normalized value

$$z = \frac{3000 - 100 \times 27.5}{\sqrt{168.75 \times 100}} = 1.92$$

■ Use the standard normal table, we have

$$P(S_{100} \leq 3000) \approx \Phi(1.92) = 0.9726$$

□

CLT Applications (III)

Ex.

Bernoulli process consists of a sequence X_1, X_2, \dots of independent *Bernoulli* random variable X_i : at each trial for each i , we have,

- $P(X_i = 1) = P(\text{success at the } i\text{-th trial}) = p$
- $P(X_i = 0) = P(\text{failure at the } i\text{-th trial}) = 1 - p$

Let $S_n = \sum_{i=1}^n X_i$, what is the probability that S_n is between a and b ?

- CLT.* We can use CLT to approximate the *Binomial*.
- Recall that for *Bernoulli distribution*, the mean and the variance is: $\mu = p$, and $\sigma^2 = p(1 - p)$.
 - Based on CLT, we calculate the normalized value $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$
 - From CLT, we have

$$\begin{aligned} P(a \leq S_n \leq b) &= P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq Z_n \leq \frac{b - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

□

Normal Approximation vs Poisson Approximation

Poisson distribution also approximates the sum: sum of n independent *Poisson* arrivals during n intervals of length $\frac{1}{n}$.

Normal when p fixed, $n \rightarrow \infty$

Poisson when np fixed, $n \rightarrow \infty$, and $p \rightarrow 0$

CLT Applications (IV: Stirling’s Approximation)

Ex.

Stirling’s approximation is an approximation for factorials. It leads to accurate results even for small values of n :

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

where \approx means that two quantities are (*asymptotic*): their ratio tends to 1 (*asymptotic*) as n tends to ∞ . Typically it is used as $\ln n! = n \ln n - n + O(\ln n)$, or the bounded form $\sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n} \leq n! \leq e n^{n+\frac{1}{2}} e^{-n}$.

- CLT.* We can use CLT to approximate the *Stirling’s Approximation*.
- Assume X_1, X_2, \dots, X_n are iid with Poisson probability $\lambda = 1$. Take $S_n = \sum_{i=1}^n X_i$, from the additivity, $S_n \sim \text{Poisson}(n)$ with mean and variance n .
 - From CLT, we have $Z_n = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n}{\sqrt{n}} \rightarrow Z$, and $Z \sim N(0, 1)$ with pdf $f(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}$.
 - So we have

$$\begin{aligned} P(S_n = n) &= P(n - 1 < S_n \leq n) = P\left(-\frac{1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leq 0\right) \approx P\left(-\frac{1}{\sqrt{n}} < Z \leq 0\right) = \int_{-\frac{1}{\sqrt{n}}}^0 f(z) dz \\ &\approx f(0)\left[0 - \left(-\frac{1}{\sqrt{n}}\right)\right] = \frac{1}{\sqrt{2\pi n}} \end{aligned}$$

- From the Poisson distribution, we have $P(S_n = n) = \frac{e^{-n} n^n}{n!}$
- Hence we have $\frac{e^{-n} n^n}{n!} \approx \frac{1}{\sqrt{2\pi n}}$, that gives us $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$

□

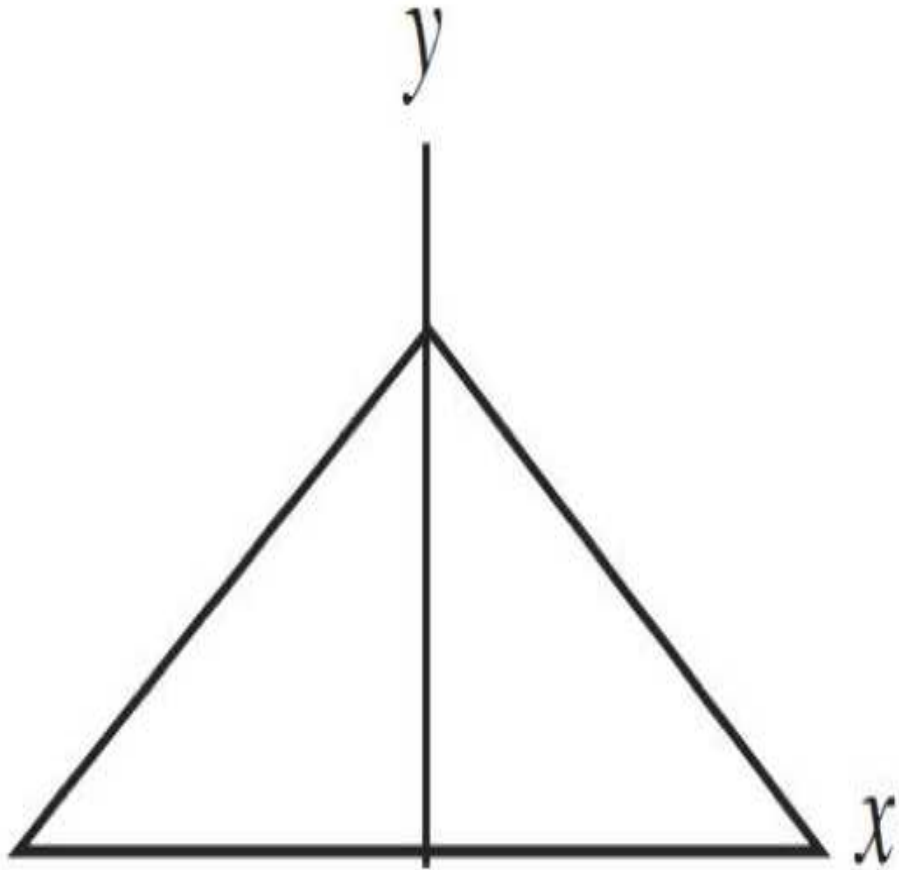
Distribution of *Errors*

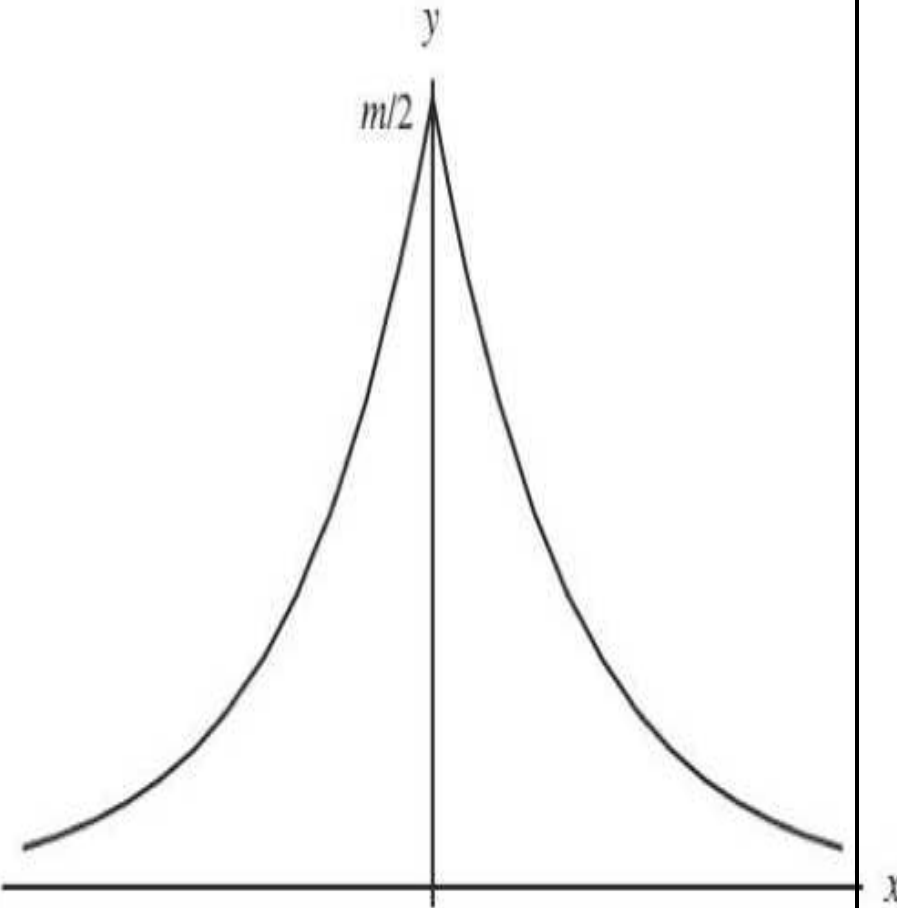
Ex.

Let θ be the true value, and x_1, \dots, x_n be n independent measures of θ , and every measure with the error $e_i = x_i - \theta$. What is the PDF of errors?

Typical Assumptions.

- Errors are symmetric;
- The probability is low for large errors, and is high for small errors;
- Laplace assumed the error distribution function $f(x)$ satisfies $-f'(x) = mf(x)$.





Simpson Error $P(|\bar{e}| < x) \geq P(|e_i| < x)$

Laplace Error $f(x) = \frac{m}{2}e^{-m|x|}$

Gauss's method. Assume the PDF of error e_i as $f(x)$, then the jointed probability for n measure errors on θ is

$$L(\theta) = f(e_1)f(e_2)\cdots f(e_n) = f(x_1 - \theta)\cdots f(x_n - \theta)$$

- let $\frac{d \log L(\theta)}{d\theta} = 0$, then $\sum_{i=1}^n \frac{f'(x_i - \theta)}{f(x_i - \theta)} = 0$
- let $g(x) = \frac{f'(x)}{f(x)}$, we have $\sum_{i=1}^n g(x_i - \theta) = 0$.
- From LSA, the solution to MLE is the mean μ , then we have $\sum_{i=1}^n g(x_i - \mu) = 0$.
 1. take $n = 2$, then $g(x_1 - \mu) + g(x_2 - \mu) = 0$. As $x_1 - \mu = -(x_2 - \mu)$, we have $g(-x) = -g(x)$.
 2. take $n = m + 1$, and $x_1 = \dots = x_m = -x$, $x_{m+1} = mx$, then $\mu = 0$ and $\sum_{i=1}^n g(x_i - \mu) = mg(-x) + g(mx)$. We then have $g(mx) = mg(x)$.
- The only continuous function satisfies both is $g(x) = cx$, hence $f(x) = Me^{cx^2}$.
- After normalizing, $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-x^2/2}$

Univariate Gaussian

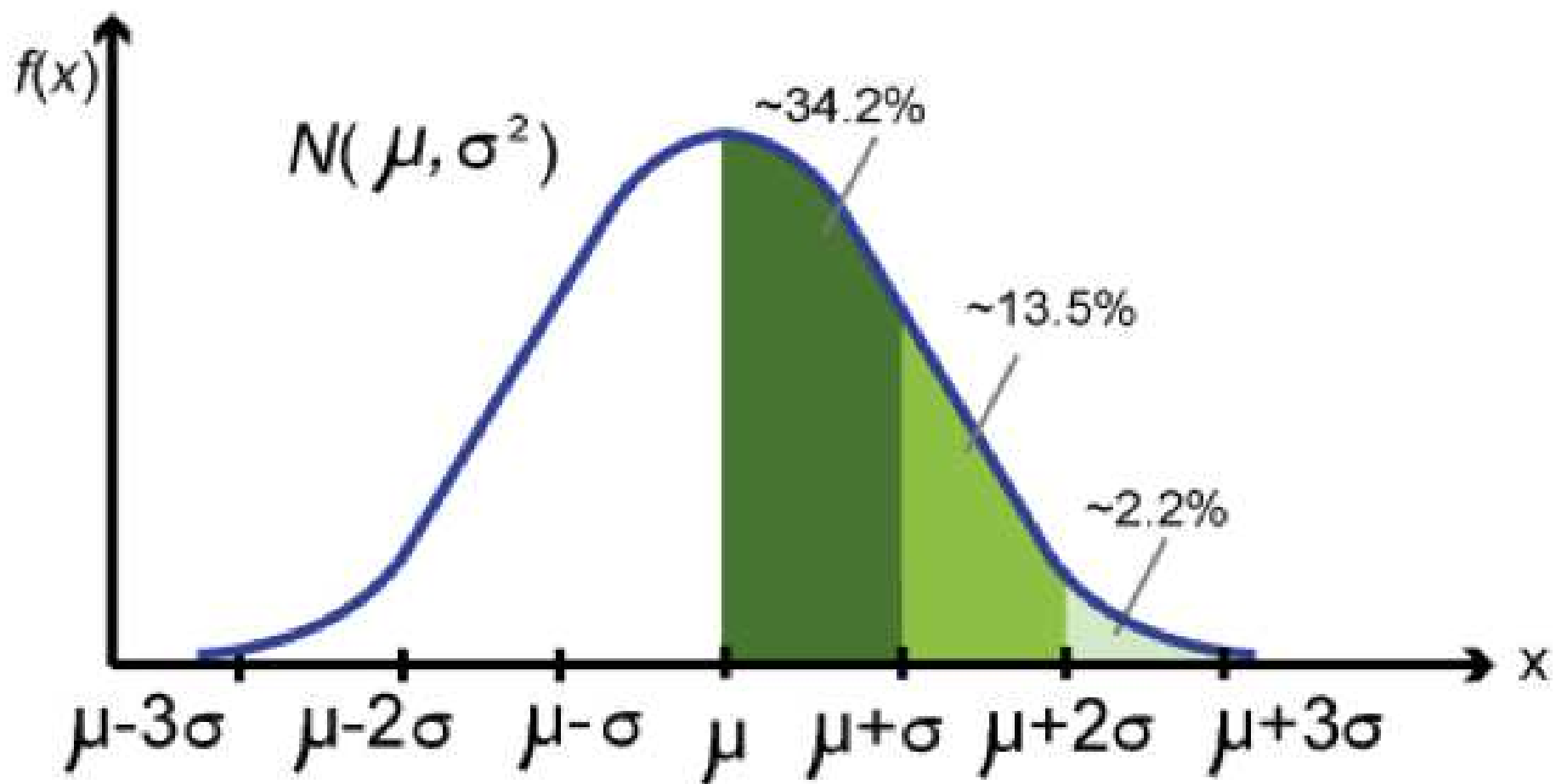
Let $\mu \in \mathcal{R}$ and $\sigma^2 > 0$, a *Gaussian distribution* in a random variable X with mean μ and variance σ^2 , is a distribution with probability density function:

Defn
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for all } x \in \mathcal{R}$$

After normalization, the pdf is:

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \text{ for all } x \in \mathcal{R}$$

We write it as $X \sim N(\mu, \sigma^2)$ or $X \sim N(0, 1)$ respectively.



Degenerated univariate Gaussian. When $\sigma = 0$, we have $X \sim N(\mu, 0)$, which means:

- $X \equiv \mu$,
- $X(\omega) \equiv \mu, \forall \omega \in \Omega$.

□

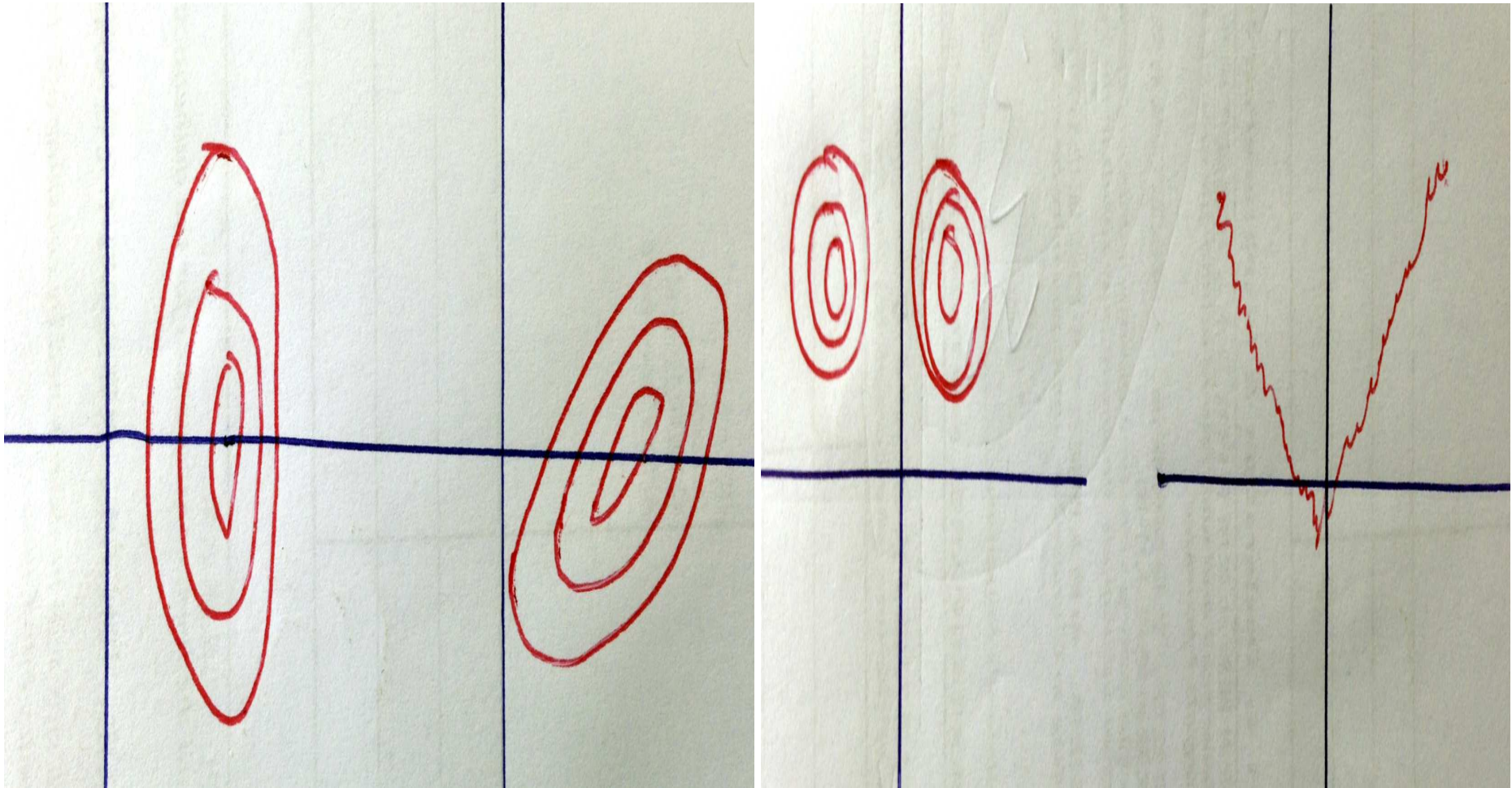
Multivariate Gaussian

Defn A random variable $X \in \mathbb{R}^n$ is *multivariate Gaussian* if any linear combination is *univariate Gaussian*: $a^T x = \sum_{i=1}^n a_i x_i$ is Gaussian distributed for all $a \in \mathbb{R}^n$.

Remarks. $X \sim N(\mu, C)$, with $\mu \in \mathbb{R}^n$ and $C \in \mathbb{R}^{n \times n}$ is *positive semi-definite*, means X is Gaussian with $E(X_i) = \mu_i$, $Cov(X_i, X_j) = C_{ij}$.

- μ and C uniquely determines the distribution $N(\mu, C)$;
- $X \sim N(\mu, C)$ is degenerated if $det(C) = 0$.

Intuitive Examples:



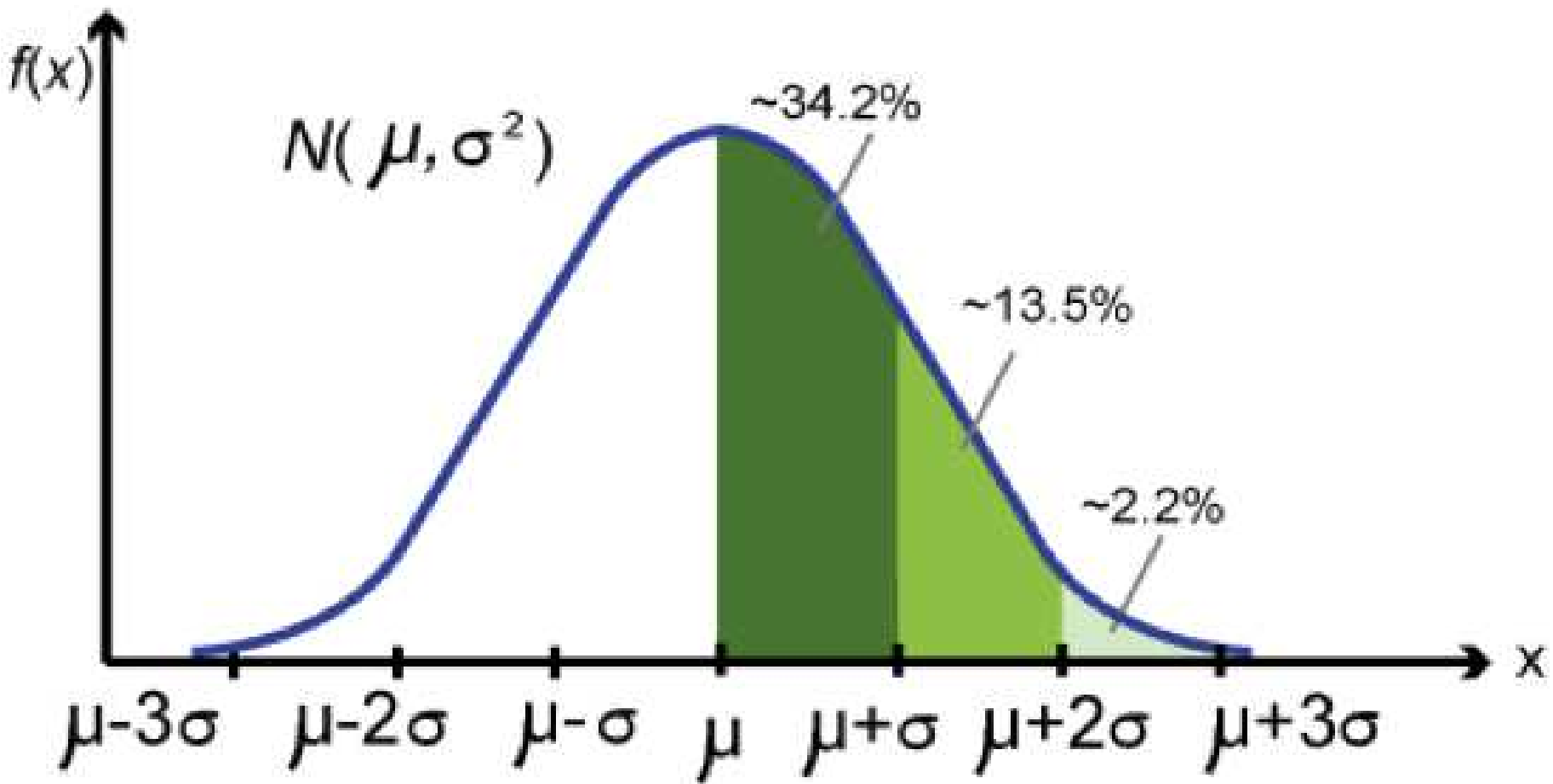
□

Facts (Independent Components). X_1, \dots, X_n are independent with $X_i \sim N(\mu_i, \sigma_i^2)$ if and only if $X = (X_1, \dots, X_n) \sim N(\mu, C)$,

where $\mu = (\mu_1, \dots, \mu_n)$ and $C = diag(\sigma_1^2, \dots, \sigma_n^2) = \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_1^2 & \\ & & \ddots \\ & & & \sigma_n^2 \end{bmatrix}$.

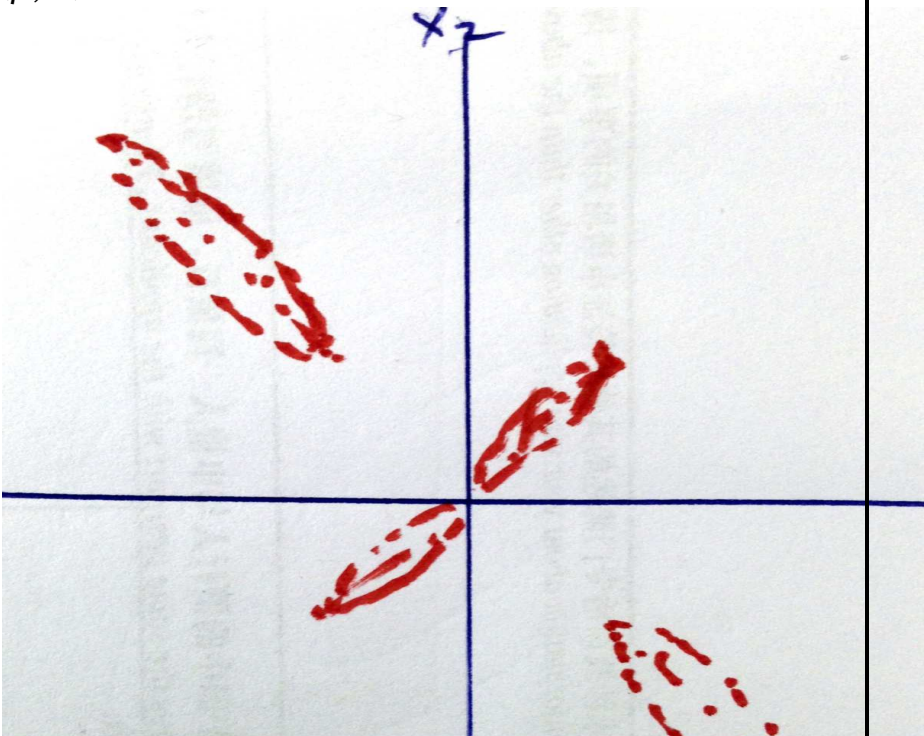
□

Intuitive Examples:



WARNING! X_1, \dots, X_n each univariate Gaussian $\nRightarrow X = (X_1, \dots, X_n) \sim N(\mu, C)$

- Let $X_1 \sim N(0, 1)$,
$$X_2 = \begin{cases} X_1 & \text{if } |X_1| \leq 1 \\ -X_1 & \text{if } |X_1| \geq 1 \end{cases}$$
- (X_1, X_2) is not Gaussian.



Multivariate Gaussian PDF

Defn A multivariate Gaussian random variable $X \sim N(\mu, C)$ with $\mu \in \mathcal{R}^n$ and $C \in \mathcal{R}^{n \times n}$, has a density if and only if it is non-degenerated, namely $\det(C) \neq 0$:

$$p(x) = \frac{1}{\sqrt{|2\pi C|}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)}, \text{ for all } x \in \mathcal{R}^n$$

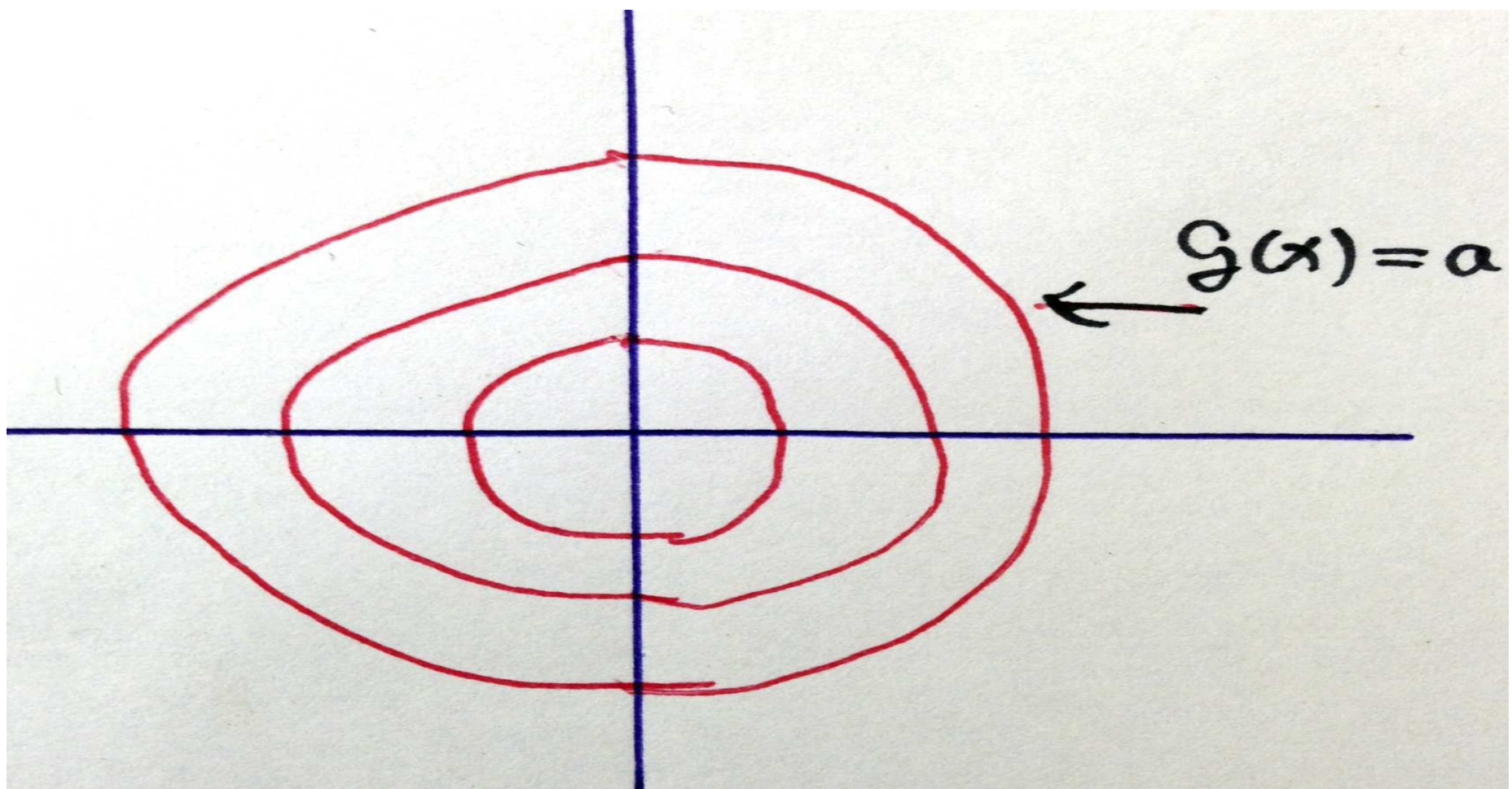
Notes. This representation is similar to univariate form:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for all } x \in \mathcal{R}$$

- Here $|A| = \det(A)$, so $\det(2\pi C) = (2\pi)^n \det(C)$.
- C^{-1} is the inverse matrix of C , that is why C needs to be *positive definite*.
- when $n = 1$, $C = \sigma^2$, we have $C^{-1} = \frac{1}{\sigma^2}$, and $\text{Cov}(X, X) = \sigma^2$.

□

Quadratic function. when C is symmetric positive semi-definite, $X^T C X = g(x)$ is a quadratic form.



□

$p(x, y)$. *Herschel* (1850) and *Maxwell* (1860) derived the *error distribution* function:

- Assume that x and y are independent on orthogonal directions;
- Assume that *error distribution* is symmetric upon rotation, namely the distribution is independent of angles.
- from assumption 1, we have $f(x, y) = f(x) \times f(y)$
- under polar system, $f(x, y) = f(r \cos \theta, r \sin \theta) = g(r, \theta)$, which equals to $g(r)$.
- so we have $f(x)f(y) = g(r) = \sqrt{x^2 + y^2}$.
- set $y = 0$, we have $g(x) = f(x)f(0)$, hence $\log f(x) + \log f(y) = \log f(\sqrt{x^2 + y^2}) + \log f(x)$, and then $\log \frac{f(x)}{f(0)} + \log \frac{f(y)}{f(0)} = \log \frac{f(\sqrt{x^2 + y^2})}{f(0)}$.
- let $h(x) = \log \frac{f(x)}{f(0)}$, we have $h(x) + h(y) = h(\sqrt{x^2 + y^2})$.
- Solution is $h(x) = \alpha x^2$, and then $f(x) = \sqrt{\frac{\alpha}{\pi}} e^{-\alpha x^2}$, which is the $N(0, \frac{1}{\sqrt{2\alpha}})$.
- $f(x, y) = \frac{\alpha}{\pi} e^{-\alpha(x^2 + y^2)}$

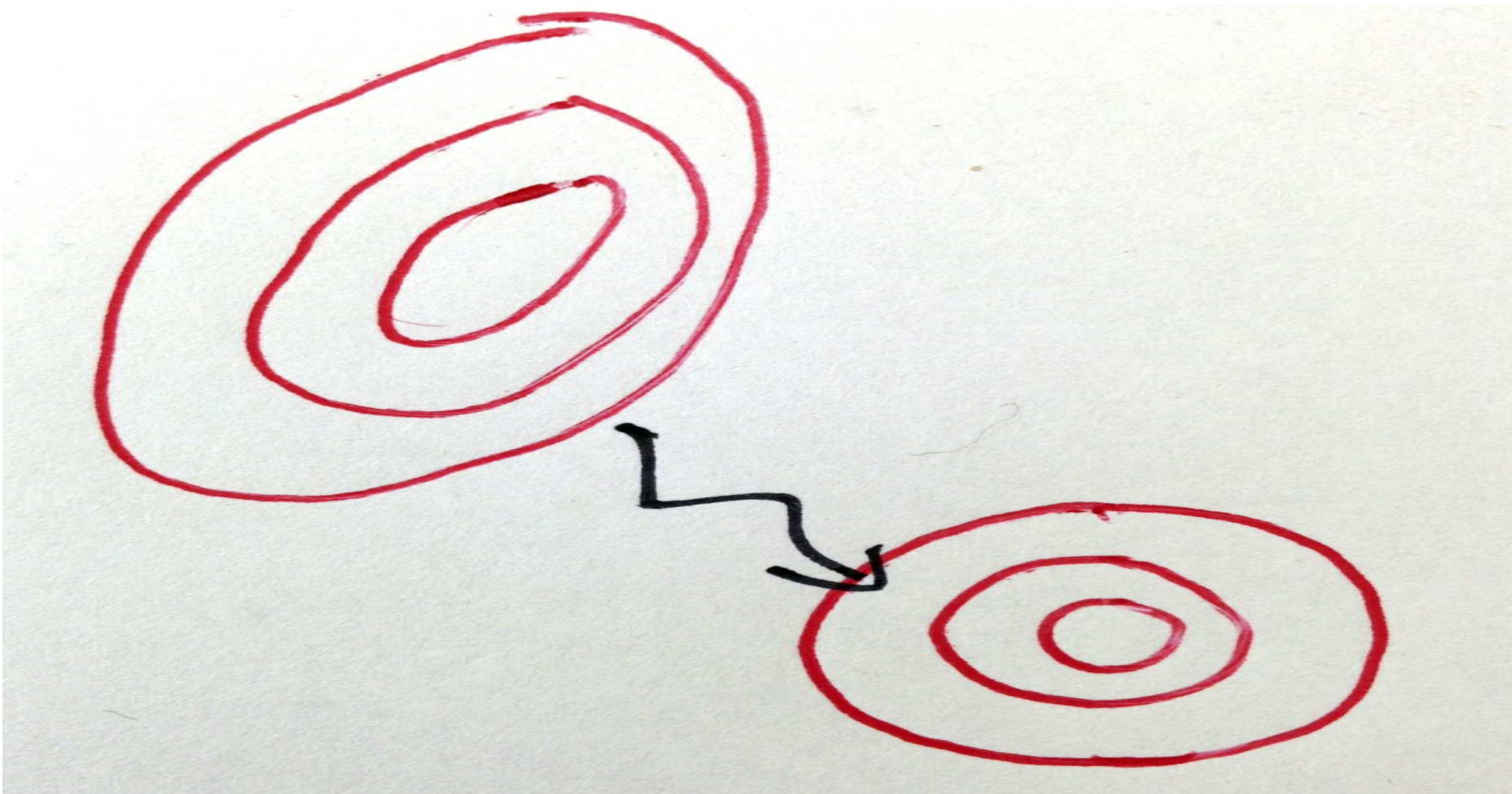
□

Affine Property

Defn Any affine transformation $f(X)=AX+b$ of a Gaussian variable is still a Gaussian variable. That is, if $X \sim N(\mu,C)$, then $AX+b \sim N(A\mu+b,ACA^T)$ for any $\mu \in \mathbb{R}^n$, any positive semi-definite $C \in \mathbb{R}^{n \times n}$, and any $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

Facts.

Constructing n independent Gaussian random variables $X_1, \dots, X_n \sim N(0,1)$, namely $X=(X_1, \dots, X_n) \sim N(0,I)$, then we have $AX+\mu \sim N(\mu,C)$, where $C=AA^T$, for any μ and A .
Sphering If C is positive definite, and $Y \sim N(\mu,C)$, we can have $A^{-1}(Y-\mu) \sim N(0,I)$, where $C=AA^T$

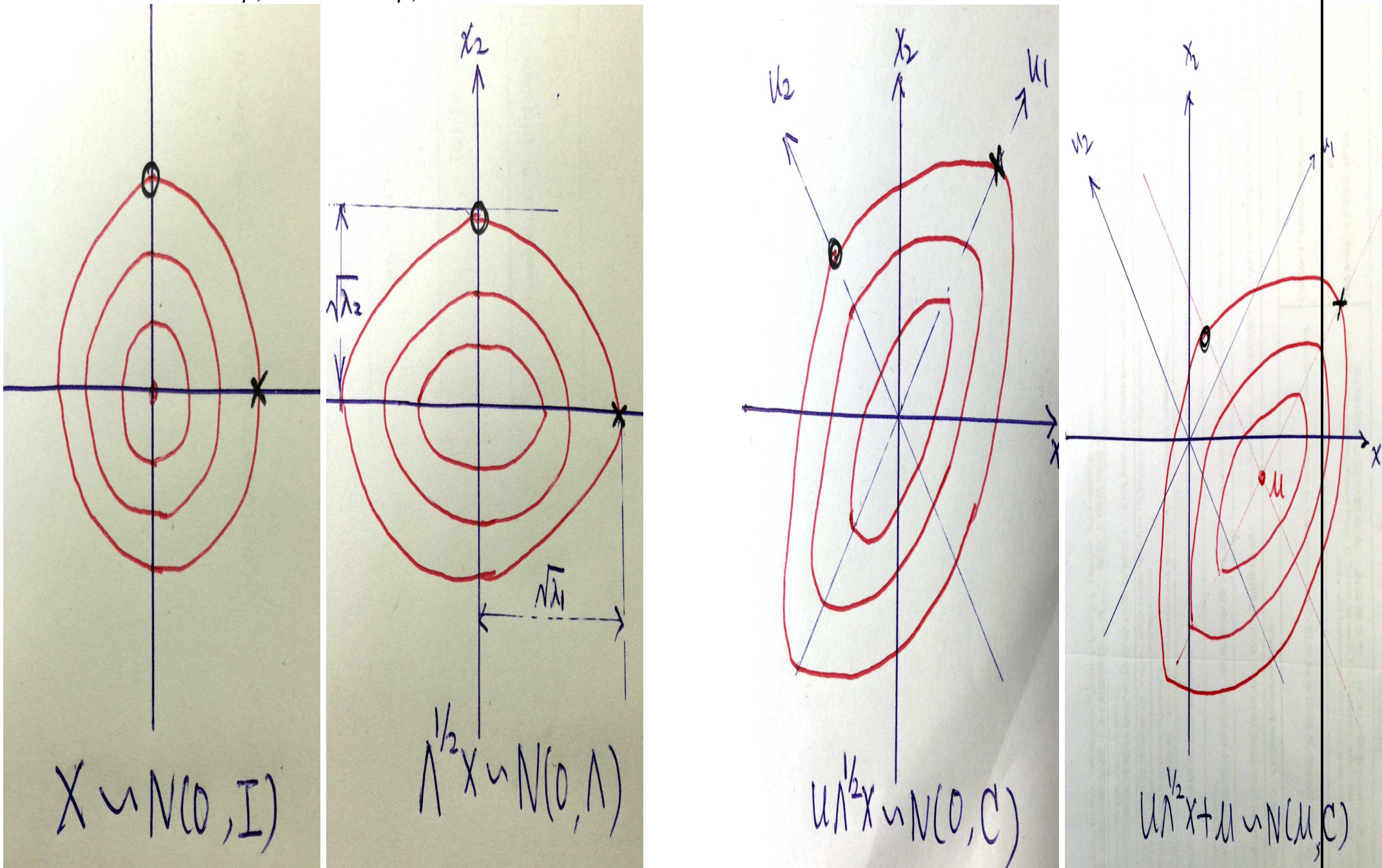


□

Geometric Intuition

Defn Let $X \sim N(0,I)$, C be a covariance matrix and $\mu \in \mathbb{R}^n$. Any symmetric matrix C can be diagonalized: $C=U\Lambda U^T=U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T=(U\Lambda^{\frac{1}{2}})(U\Lambda^{\frac{1}{2}})^T$, where $\Lambda=\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$, and $U=(u_1, \dots, u_n)$ is orthogonal space formed by eigenvectors u_i .

Intuition. Let $Y=AX+\mu$, then $Y \sim N(\mu,C)$.



□

Gaussian Marginals and Conditionals

Defn For a *multivariate Gaussian* distribution, all of its *marginals* and *conditionals* are *Gaussian*.

Marginals Specific Case. Let $X = (X_1, X_2) \in \mathcal{R}^2$ is Gaussian, then X_1 and X_2 are both Gaussian.

- Let $A = (1, 0)$ and $b = (0, 0)$,
- assume $X \sim N(\mu, C)$, then from *Affine Property*, we have

$$AX + b \sim N(A\mu + b, ACA^T)$$

namely $X_1 \sim N(A\mu + b, ACA^T)$.

- Similarly, the X_2 can also be proved to be *Gaussian*.

□

Marginals General Case. Let $X \sim N(\mu, C)$, $a = (1, \dots, k)$ and $b = (k + 1, \dots, n)$, where $1 \leq k \leq n$.

- We can rewrite $X = \begin{bmatrix} X_a \\ X_b \end{bmatrix}$, where $X_a = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}$ and $X_b = \begin{bmatrix} X_{k+1} \\ \vdots \\ X_n \end{bmatrix}$, and $\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}$, $C = \begin{bmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{bmatrix}$, where $C_{aa} = \begin{bmatrix} C_{11} & \cdots & C_{1k} \\ \vdots & \vdots & \vdots \\ C_{k1} & \vdots & C_{kk} \end{bmatrix}$.

- Let $k \times n$ matrix $A = \begin{bmatrix} 1 & & & \\ & 1 & & \\ \vdots & & & \\ & & 1 & 0 & 0 \end{bmatrix}$

- $AX = X_a \sim N(\mu_a, C_{aa})$, where $A\mu = \mu_a$ and $ACA^T = C_{aa}$.

□


Conditionals Specific Case. Let $X = (X_1, X_2) \in \mathcal{R}^2$ be Gaussian, then the conditional distribution $P(X_1|X_2 = x_2)$ is also Gaussian.

□

Questions?

Contact Information

Associate Professor **Gang Li**
School of Information Technology
Deakin University, Australia

 GANGLI@TULIP.ORG.AU
 TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING