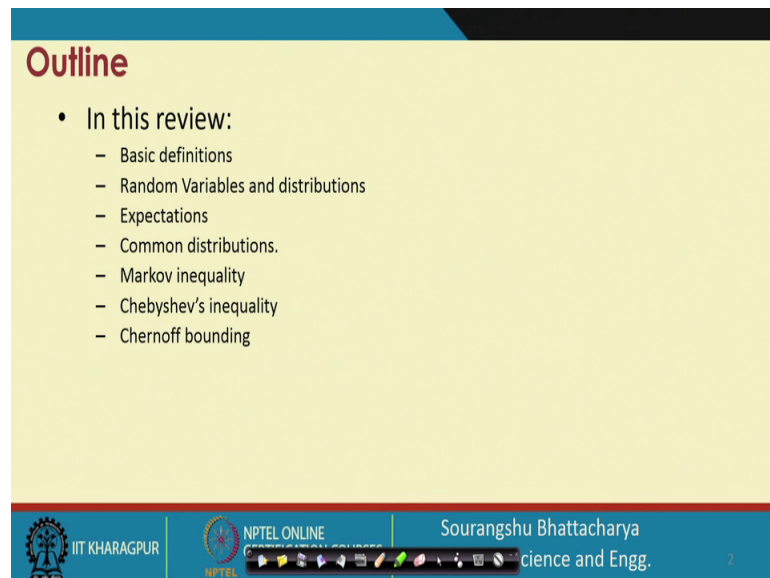


Scalable Data Science
Prof. Sourangshu Bhattacharya
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 02
Background Probability Theory

Hello. Welcome to the NPTEL course on Scalable Data Science, lecture number 2. Today we are going to discuss about the Background on Probability Theory which will be required for this course. I am Professor Sourangshu Bhattacharya of Department of Computer Science and Engineering, IIT, Kharagpur.

(Refer Slide Time: 00:36)



Outline

- In this review:
 - Basic definitions
 - Random Variables and distributions
 - Expectations
 - Common distributions.
 - Markov inequality
 - Chebyshev's inequality
 - Chernoff bounding

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
Department of Computer Science and Engg.

So, in this lecture we will discuss the basic definitions of probability, then we will discuss random variables and distributions. We will discuss expectations of random variables and we will describe some common distributions which are used in a data mining and machine learning. Then we will describe concentration 3 concentration inequalities. So, first we will describe the Markov inequality, then we will describe the Chebyshev's inequality and finally, we will describe the Chernoff bounding technique, ok.

(Refer Slide Time: 01:18)

Probability: Definition

- **Experiment:** toss a coin twice
- **Sample space:** possible outcomes of an experiment
 - $S = \{HH, HT, TH, TT\}$
- **Event:** a subset of possible outcomes
 - $A = \{HH\}, B = \{HT, TH\}$
- **Probability of an event:** an number assigned to an event $\Pr(A)$
 - Axiom 1: $\Pr(A) \geq 0$
 - Axiom 2: $\Pr(S) = 1$ →
 - Axiom 3: For every sequence of disjoint events
$$\Pr\left(\bigcup_i A_i\right) = \sum_i \Pr(A_i)$$
 - Example: $\Pr(A) = n(A)/N$ frequentist statistics

Handwritten notes: $n(A)$ and N

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya | Science and Eng

So, 4 terms come or are important a while defining probability. So, first is an experiment, an experiment is any real life phenomenon's which we want to study or we want discuss or we want to model, ok. So, for example, the tossing of a coin is an experiment tossing or. So, experiment could be also much more complex things like for example, tossing of two coins and or tossing of any number of coins.

Then we discuss, we use the term sample space which describes the set of all possible outcomes of an experiments. So, for example, if our experiment is toss of two coins or toss of a coin twice then the sample space is of size 4 and the and the sample space is basically a head followed by a head, a head followed by a tail, a tail followed by a head and a tail followed by a tail. So, these are the 4 possible outcomes.

Then the term event is used to describe a subset of possible outcomes. So, for example, eh as you can see event A may be just occurrence of one head followed by a head or another event B could be the occurrence of a head followed by a tail or a tail followed by a head. So, given this the probability of an event is a number which is assigned to an event, ok. So, let the event be A. So, the number is described as probability of A. And this number will should satisfy the following 3 axioms the first is it should be positive, second is the probability of entire sample space that is which is also an event should be one and lastly if you have a set of disjoint event say A_i , ok. Then the probability of

union of this disjoint events A_i is the sum of probabilities of this disjoint event, so this are the 3 axioms that it should satisfy.

As you can see if for example, simple way of calculating the probability is if you have N outcomes, so if you can describe your experiment in terms of let say capital N outcomes. And let say for a given event A N denotes the number of outcomes which favor the event a the probability of event A could simply be written as n of A divided by N you can see that this satisfy all the axioms of probability.

(Refer Slide Time: 04:45)

Probability

- **Joint Probability:** For events A and B , joint probability $Pr(AB)$ stands for the probability that both events happen.
- **Independence:** Two events A and B are independent in case $Pr(AB) = Pr(A) Pr(B)$
- A set of events $\{A_i\}$ are independent in case $Pr\left(\bigcap_i A_i\right) = \prod_i Pr(A_i)$
- **Conditional Probability:** If A and B are events with $Pr(A) > 0$, the **conditional probability of B given A** is $Pr(B|A) = \frac{Pr(AB)}{Pr(A)}$

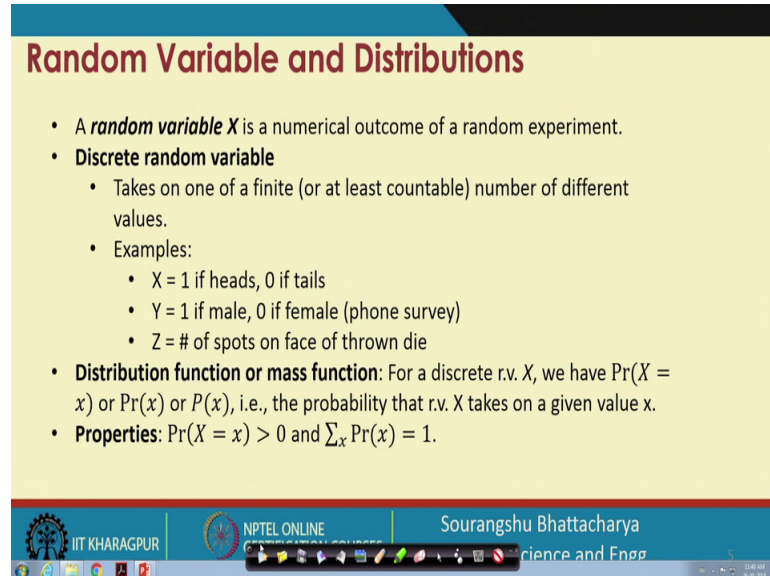
IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
science and Engg.

Next we just describe the term joint probability. So, the joint probability of two events A and B also described as probability of $A B$ in this the probability that both events occur, ok. So, this is also sometimes described as the probability of intersection of the events A and B , ok. So, then we describe the notion of independence. So, two events A and B are said to be independent if the probability of $A B$ or the joint probability of A and B is equal to the probability of occurrence of A times the probability of occurrence of B .

Another so if we have more than two events let say you have events A_i then the joint probability of occurrence of all the events A_i which is the probability of intersection of A_i is equal to product of probabilities of A_i if all the events are independent. Finally, we can describe the conditional probability of two events A and B , given that the probability of event a is strictly positive that is there is a probability that the event a will occur

and the probability of conditional probability of the event B given A is a joint probability of event A B divided by the probability of event A.

(Refer Slide Time: 06:44)



Random Variable and Distributions

- A **random variable** X is a numerical outcome of a random experiment.
- **Discrete random variable**
 - Takes on one of a finite (or at least countable) number of different values.
 - Examples:
 - $X = 1$ if heads, 0 if tails
 - $Y = 1$ if male, 0 if female (phone survey)
 - $Z = \#$ of spots on face of thrown die
- **Distribution function or mass function:** For a discrete r.v. X , we have $\Pr(X = x)$ or $\Pr(x)$ or $P(x)$, i.e., the probability that r.v. X takes on a given value x .
- **Properties:** $\Pr(X = x) > 0$ and $\sum_x \Pr(x) = 1$.

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
Science and Engg

Now, sometimes it is easier to describe the outcomes as values of random variables instead of sets of entities. So, for example, a random variable is a variable which takes a values which takes values in the sample space and they describe a numerical outcome of once such a random experiment that we have discussed earlier. So, there can be two types of random variables, one is the discrete random variable. So, a discrete random variable is a random variable which takes one of either a finite number of a different values or countable number of different values.

So, for example, if we consider if we consider our experiment of coin toss then the random variable X equal, so we can define a random variable X , as X is equal to 1 if the experiment turns out to be head and 0 if the experiment turns out to be tails. Similarly for example, in case of a phone survey the random variable Y could be 1 if the person you call is male and 0 if there the person you call is female or in case of throwing of a dies which is another random experiment the outcome could be the number of spots on the face of the thrown die. So, the random variable Z in that case will take on 6 values is the 1 to 6 because there are 6 faces.

The distribution function for a discrete random variable which is defined as probability of which is rather written as probability of X is equal to x or \Pr of x sub probability of x

is the probability that the random variable takes the value X . Now, from the previous or from the previous axioms it is easy to see that the. So, the random variable the probability of a random variable takes on values or the probability distribution function of a random variables takes on values which is greater than 0. And the sum over random variables or sum over the probability of sum over all values the all values say x the probability of x is equal to 1, where x basically ranges over the sample space or all possible outcomes.

(Refer Slide Time: 10:03)

Random Variable and Distributions

- **Continuous random variable**
 - Takes on one in an infinite range of different values
 - Examples:
 - $W = \% \text{ GDP grows (shrinks?) this year}$
 - $V = \text{hours until light bulb fails}$
- **Distribution function:**
 - What is the probability that a continuous r.v. takes on a specific value?
e.g. $\text{Prob}(V = 3.14159265 \text{ hrs}) = 0$
 - However, ranges of values can have non-zero probability.
e.g. $\text{Prob}(3 \text{ hrs} \leq V \leq 4 \text{ hrs}) = 0.1$
 - For a continuous r.v. X , we have $\Pr(x)$ or $P(x) = \Pr(x \leq X \leq x + dx)$.
- **Properties:** $P(x) > 0$ and $\int_x P(x) dx = 1$.

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
science and Engg.

Now, unlike discrete random variable continuous random variables are the random variables which take infinite number of values. So, examples include for example, the rate of GDP growth of a particular country per year or the number of hours till a light bulb fails or the number of hour a light bulb glows till it stops growing glowing or fails.

As you can see the total number of values that this particular random variable can take is infinite, hence what you can check or what you can see is that the probability of a continues random variable taking a particular value. For example, the light bulb glowing for particular number of hours let say 3.1415 hours is actually equal to 0. However, the probability of a random a continuous random variables taking values in a given range is a finite number so, for example, the light bulb glowing for 3 to 4 hours a before failure the probability of that may be 10 percent or 0.1, ok.

So, the distribution function of a continuous random variable which is also written as probability of x is actually the probability that the continuous random variable capital X takes on a value between x and x plus dx which is a small increment from x . We can see that again the distribution function of a continuous random variable takes is a positive quantity and if we integrate over all possible values of x the so it integrates to 1, because the probability of the sample space has to be 1.

(Refer Slide Time: 12:36)

Expectation

- A discrete random variable $X \sim P(X = x)$. Then, its expectation is:

$$E[X] = \sum_x x P(X = x)$$
- For an empirical sample, x_1, x_2, \dots, x_N , expectation can be estimated as:

$$E[X] = \frac{1}{N} \sum_{i=1}^N x_i$$
- Continuous random variable: $E[X] = \int_x x P(x) dx$
- Expectation of sum of random variables: $E[X_1 + X_2] = E[X_1] + E[X_2]$.
- A measure of central tendency. Other measures: median, mode, etc.

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
 science and Eng

Next we define the interesting concept of expectation. So, expectation of discrete random variables is the sum over all possible values of the random variable X , X times the probability of x , ok. So, and similarly you can see that for a continuous random variable this summation is replaced by the integral. So, it is the integral over all possible values again the infinite number of values the continuous random variable can take and x times P of x dx .

Now, if we have finite number of samples say x_1 till x_N of the random variables then we can estimate the expectation of x using this formula $\frac{1}{N}$ summation over i is equal to 1 to N x_i . As you can see this is the formula for the average value of the random variable, hence the expectation of the random variable also denotes or it describes the average value of the distribution of the expectation.

Now, there are other ways of denoting. So, expectations is a measure of central tendency, ok. So, there are many other ways of measuring such as mean mid, so mean is also the

expectation is also called the mean but median and more are also other ways of measuring the central tendency of a random variable.

(Refer Slide Time: 14:29)

Variance

- The variance of a random variable X is the expectation of $(X - E[X])^2$:

$$\begin{aligned} \text{Var}(X) &= E((X - E[X])^2) \\ &= E(X^2 + E[X]^2 - 2XE[X]) \\ &= E(X^2 - E[X]^2) \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Handwritten notes:

$$E[(X - E[X])^2]$$

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

Footer: IIT KHARAGPUR, NPTEL ONLINE, Sourangshu Bhattacharya

Another very very important quantity is the variance of the random variables is the variance of the random variable is defined as the expectation of this quantity X minus expectation of x whole square. So, for example, if random variable as an expectation or mean of E of x and if we take any value x in the sample space what we are measuring the square of the difference and taking the expectation of this square ok. So this is; what is the variance of the random variable X .

Now, as you can see this is a measure of a spread of the random variable X . More over this small derivation here shows that you can actually calculate the variance of X and expectation of the random variable X square minus the expectation of X whole square. And this is because of the linearity property of the expectation. So, what we can show is that expectation of X plus Y which is another random variable is nothing but expectation of X plus expectation of Y , ok. And using this formula we can show that we can expand we can expand the square and then we can percolate the expectation in side and see that it turns out to be expectation of x square minus expectation of X whole square.

(Refer Slide Time: 16:53)

Bernoulli Distribution

- The outcome of an experiment can either be success (i.e., 1) and failure (i.e., 0).
- $\Pr(X = 1) = p, \Pr(X = 0) = 1 - p$, or

$$p_{\theta}(x) = p^x(1-p)^{1-x}$$

- $E[X] = p, \text{Var}(X) = p(1-p)$

The slide includes a footer with the IIT KHARAGPUR logo, NPTEL ONLINE text, and the name Sourangshu Bhattacharya. A Windows taskbar is visible at the bottom of the slide image.

Now, we describe some compound distributions which are used in practice. So, the first common distribution that is used in practice is the Bernoulli distribution. So, this is the distribution over binary outcomes, ok. So, if the outcome of an experiment is either success which is denoted by 1 or failure which is denoted by 0, then we can describe the outcome. So, then we can describe the distribution over these outcomes using probability of success parameter which is the p parameter, ok. So, note that probability of failure is automatically becomes 1 minus p .

Hence we can see that the Bernoulli distribution can be written as p to the power x times one minus p to the power $1-x$. So, the distribution function of the Bernoulli distribution can be written as p to the power x times 1 minus p to the power $1-x$ and the expectations of the Bernoulli distribution is just p and the variance of the Bernoulli distribution will be turn out to be p times 1 minus p .

(Refer Slide Time: 18:24)

Binomial Distribution

- n draws of a Bernoulli distribution
 $X_i \sim \text{Bernoulli}(p), X = \sum_{i=1}^n X_i, X \sim \text{Bin}(p, n)$
- Random variable X stands for the number of times that experiments are successful.

$$\Pr(X = x) = p_\theta(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

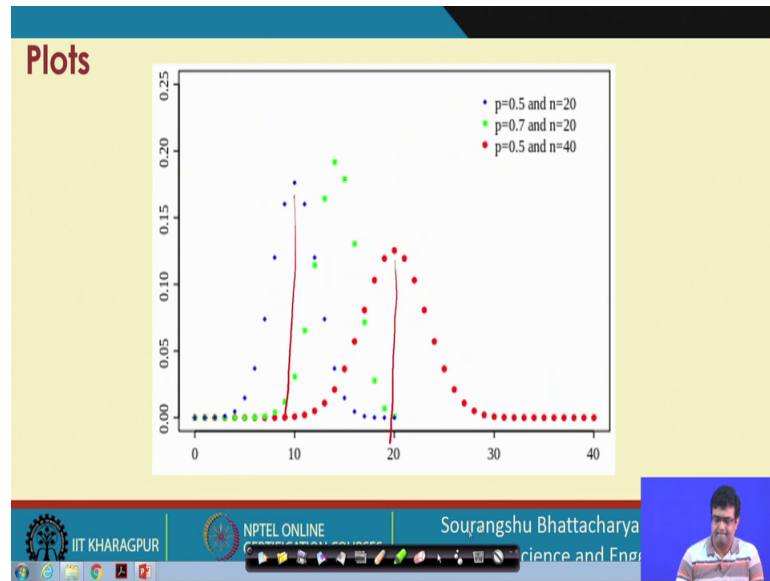
- $E[X] = np, \text{Var}(X) = np(1-p)$

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
Science and Engg

Next, we describe another very interesting distribution which is the binomial distribution the binomial distribution models instead of one draw of it models N draws of a Bernoulli random variables, ok. So, instead of measuring success or failure of just one random variables something like a heads or a tail, it measures the number of successes that one can get if we toss if we toss the coin or if we perform the experiment N number of times, ok. So, the binomial distribution has two parameters p and n the first parameter p denotes the probability of success and the second parameter n denotes the number of times you have tossed the coin.

And we can see that the probability of or the probability distribution function of the binomial distribution is n choose x , p to the power x times 1 minus p to the power 1 minus x and the expectation of a the binomial distribution is n times p and the variance of the binomial distribution is n times p into 1 minus p .

(Refer Slide Time: 20:12)



And this is a plot which shows the distribution function for the binomial distribution for different values of p and n , as you can see as the n increases or the number of coin tosses increases for the same success probability the mean increases and the distribution shifts to the right where as this blue distribution is for 20 coin tosses, so for mean for this is 0.5 times 20 which is 10.

(Refer Slide Time: 20:53)

Poisson Distribution

- Distribution of number of arrivals, given the average rate of arrival, λ .
- Coming from Binomial distribution
 - Fix the expectation $\lambda=np$
 - Let the number of trials $n \rightarrow \infty$A Binomial distribution will become a Poisson distribution

$$\Pr(X = x) = p_{\theta}(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

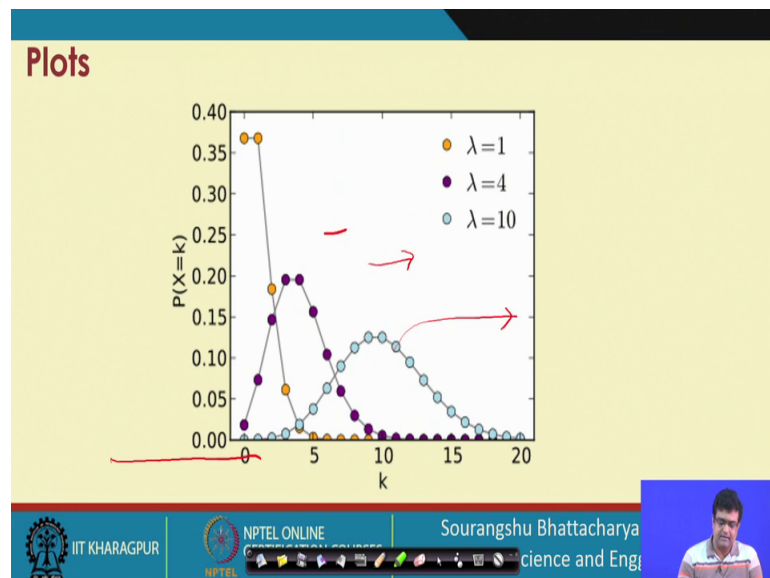
- $E[X] = \lambda, \text{Var}(X) = \lambda$

The slide includes a navigation bar at the bottom with 'IIT KHARAGPUR', 'NPTEL ONLINE', and 'Sourangshu Bhattacharya'.

Next, we describe another important distribution which is the Poisson distribution which is a distribution over the number of arrivals of let us say objects or customers in a queue

or a packets in a networks given that the average rate of arrival is lambda, ok. So, if we come from a binomial distribution then the we know that the average of a binomial distribution is np and hence if we set np equals to lambda and then if we let the number of trials n tends to infinity because there can there can be infinite number of customers who arrive. Then we arrive at the distribution function of the Poisson distribution and the distribution function of the Poisson distribution is given by this number where probability of x number of arrivals is lambda to the power x by x factorial into E to the power minus lambda. And we can see that both the mean and the variance of the Poisson distribution is lambda.

(Refer Slide Time: 22:23)



And here we show the distribution function of Poisson distribution for various values of the parameter or of the mean parameter lambda.

(Refer Slide Time: 22:41)

Normal (Gaussian) Distribution

- Continuous valued distribution $[-\infty, \infty]$
- $X \sim N(\mu, \sigma)$ \longrightarrow $[-\infty, \infty]$

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$
$$\Pr(a \leq X \leq b) = \int_a^b p_{\theta}(x) dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx$$

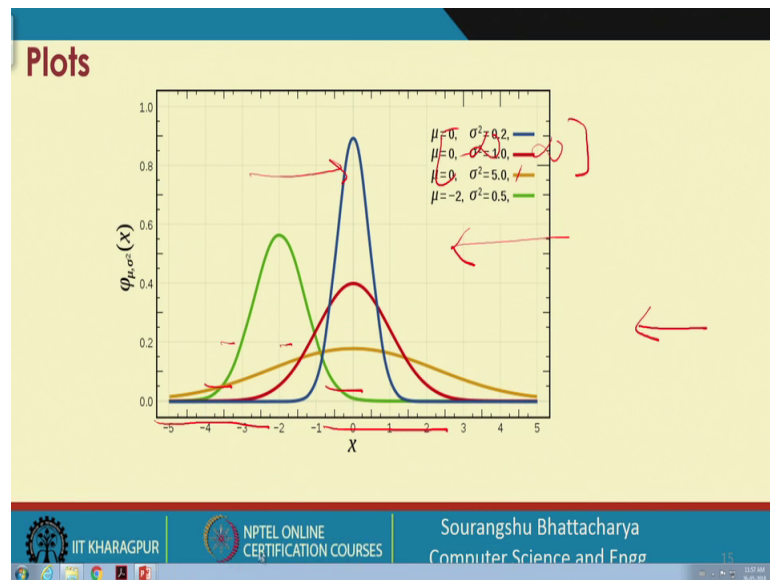
- $E[X] = \mu, \text{Var}(X) = \sigma^2$
- If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, $X = X_1 + X_2$, then $X \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
Science and Engg

Another very important distribution is the normal distribution or the Gaussian distribution which is a continuous distribution. So, the Gaussian distribution random variables takes values. So, the Gaussian random variable is X it takes values in the intervals minus infinity to plus infinity, ok. And the probability distribution function of the Gaussian random variable is given by this quantity and if we want to calculate the probability of x taking values between two numbers A and B it is given by this integral as we have discussed.

Now, the important thing is that the Gaussian distribution is actually parameterized by its mean and its variance. So, the Gaussian distribution is parameterized by its mean μ and variance σ^2 more over it has the nice property that is you have two Gaussian random variables x_1 and x_2 with means μ_1 and μ_2 and standard deviation σ_1^2 and σ_2^2 . Then the sum x_1 plus x_2 is also a Gaussian distribution with mean μ_1 plus μ_2 and standard deviation σ_1^2 plus σ_2^2 .

(Refer Slide Time: 24:15)



And this is the plot or this is the curve of the distribution function of the standard normal or of the Gaussian random variable for various parameter values.

(Refer Slide Time: 24:39)

Motivation

Many times we do not need to calculate probabilities **exactly**.

Sometimes it is enough to know that a probability is very small (or very large)

E.g. $P(\text{earthquake tomorrow}) = ?$

This is often a lot **easier**

I toss a coin 1000 times. The probability that I get **14 consecutive heads** is

A	B	C
< 10%	$\approx 50\%$	> 90%

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | Sourangshu Bhattacharya | Computer Science and Engg. | 17

Now, we go into concentration inequalities. So, many a times we do not need to calculate the probabilities exactly but it is enough to have a bound on the probability. So, for example, we may want to know whether the probability of an earthquake occurring tomorrow is let say less than 10 percent or equal to 50 percent or greater than 90 percent. So, we do not nearly need to know the exact value of the probability but we

need to know the value of we need a bound on the probability. So, concentration inequalities allow us to do that in case of very very complex events for which the exact calculation of probability may not be possible.

(Refer Slide Time: 25:41)

Consecutive heads

Let N be the number of occurrences of 14 consecutive heads in 1000 coin flips.

$$N = I_1 + \dots + I_{987}$$

where I_i is an indicator r.v. for the event "14 consecutive heads starting at position i "

$$E[I_i] = P(I_i = 1) = 1/2^{14} \rightarrow$$
$$E[N] = 987 \cdot 1/2^{14} = 987/16384 \approx 0.0602$$

The slide includes a footer with logos for IIT KHARAGPUR, NPTEL ONLINE, and Sourangshu Bhattacharya, along with a small video inset of the speaker.

So, for example, consider the situation where we want to let say calculate the probability or calculate the expectation of a random variable N which counts the number of occurrences of 14 heads in a coin toss of thousand coins, ok. So, how many times would a contiguous sequence of 14 heads occur if I toss a coin for 1000 times, ok. So, how will we calculate the expectation of this?

So, first we define an indicator random variable I_i which denotes the event that 14 consecutive heads occur starting at position i . The probability of this happening is 1 by 2 to the power 14 because probability of one head occurring is half and hence probability of 14 consecutive heads occurring is 1 by 2 to the power 14. Now, then N is actually sum of I_1 till I_{987} which is in all the ways in which the 14 heads can occur and hence by linearity of expectation we can see that expectation of N the number of times 14 heads occur in thousand coin tosses is roughly 6 percent.

(Refer Slide Time: 27:25)

Markov's inequality

For every **non-negative** random variable X and every value a :

$$P(X \geq a) \leq E[X] / a.$$

$E[N] \approx 0.0602$

$P[N \geq 1] \leq E[N] / 1 \leq 6\%.$

The slide includes a red arrow pointing from the title to the boxed text, and red underlines under the terms $P(X \geq a)$, $E[X]$, and a in the inequality. Below the inequality, the terms $E[N]$ and 1 are also underlined. The slide footer contains logos for IIT KHARAGPUR, NPTEL ONLINE, and Sourangshu Bhattacharya, along with a small video inset of the speaker.

Now, we discuss the Markov's inequality which will give us our first concentration inequality. So, the Markov inequality says that the probability of a random variable taking value greater than a specific value a is less than expectation of that random variable X by a note that the random variable X has to be a non-negative random variable. So, if we use this if we use this inequality Markov's inequality then we can try to calculate the probability that the number of times 14 head occur heads occur in a sequence is greater than exactly 1, ok.

And we can calculate this by we can we can say that this probability is less than about 6 percent because, we have already calculated that expectation of N is 0.6. And hence, we by a Markov inequality we know the this is expectation of x by 1 which is less than 0.6.

(Refer Slide Time: 28:56)

Markov's inequality

For every non-negative random variable X :
and every value a :

$$P(X \geq a) \leq E[X] / a.$$

$$E[X] = E[X | X \geq a] P(X \geq a) + E[X | X < a] P(X < a)$$

$$E[X] \geq a P(X \geq a) + 0.$$

The slide includes a graph of a probability density function with a vertical line at a . Handwritten red annotations include arrows pointing to the terms in the expectation formula, a circled a on the graph, and the text $= 0$ under the second term of the expectation formula.

IIT KHARAGPUR NPTEL ONLINE Sourangshu Bhattacharya
Science and Engg

Now, in order to see this it is easy to see this. So, we write the expectation of x into two parts. So, let say this is the random variable X and this is the value a . So, we can take the expectation of X condition on the fact that X is less than a . So, all the expectation over all values of X which is less than a and so this is the second term. Now, we note that since X is a positive random variable the probability of X less than a is greater than 0 and also the expected value of X such that X less than a is greater than or equal to 0 because for all values of a the random variable X takes only positive values, ok.

Similarly, we know that expectation of X such that X greater than a is always going to be greater than a simply because all values of X in this part of the term are actually greater than a . So, all values of a that I take are greater than a . From this we can see by setting this term to 0 we get the Markov inequality.

(Refer Slide Time: 30:50)

Patterns

A coin is tossed 1000 times. Give an **upper bound** on the probability that the pattern **HH** occurs:

(a) at least 500 times

(b) at most 100 times

IIT KHARAGPUR NPTEL ONLINE Sourangshu Bhattacharya
science and Eng

Now, suppose a coin is we can use the Markov inequality to calculate the probabilities of events or other bound the probabilities of events in many cases. So, for example, suppose a coin is tossed 1000 times we and we want to give bounds on probability that the pattern each occurs probability that the pattern each occurs at least 500 times or at most 100 times, ok.

(Refer Slide Time: 31:34)

Patterns

(a) Let N be the number of occurrences of **HH**.
Last time we calculated $E[N] = 999/4 = 249.75$.

$$P[N \geq 500] \leq E[N] / 500 = 249.75 / 500 \approx 49.88\%$$

so 500+ **HHs** occur with probability $\leq 49.88\%$.

(b) $P[N \leq 100] \leq ?$

$$\begin{aligned} P[N \leq 100] &= P[999 - N \geq 899] \leq E[999 - N] / 899 \\ &= (999 - 249.75) / 899 \\ &\leq 83.34\% \end{aligned}$$

IIT KHARAGPUR NPTEL ONLINE Sourangshu Bhattacharya
science and Eng

And this can be given using. So, the first bound can be given is a straight forward application of Markov inequality where N is greater than 500 and we can see that

expectation of N is actually just 999 by 4 which is 249.75 . Now, for the second calculation we can define a new random variables which is 999 minus N and calculate the probability that this is greater than 899 and we can come up with the probability of 83 percent and this one is probability of 50 percent.

(Refer Slide Time: 32:20)

Chebyshev's inequality

For every random variable X and every t :

$$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$$

where $\mu = E[X]$, $\sigma = \sqrt{\text{Var}[X]}$.

$P(|X - \mu| \geq t\sigma) = P((X - \mu)^2 \geq t^2\sigma^2) \leq E[(X - \mu)^2] / t^2\sigma^2 = 1 / t^2.$

The slide includes a normal distribution curve with the area under the curve outside the interval $[\mu - t\sigma, \mu + t\sigma]$ shaded in red. Handwritten red notes include $\frac{1}{t^2}$ and $x \pm t\sigma$.

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya | Science and Engg.

Now, the Markov inequality have some problems the first problem is that it assumes that the random variables is positive, ok. The second problem is that it does not utilize the information about the variance of the random variable ok; it only utilizes the information about the expectation or the means of the random variable. So, Chebyshev's inequality solves that problem by taking care of both the mean and the variance of the random variable and giving the bound ok.

So, as we can see here the Chebyshev's inequality says that the probability of a random variables deviating from its mean by more than t times the standard deviation is less than 1 minus 1 by σ 1 by t square, ok. So, if we draw this random variable the pds of this random variable and let say this is the mean. So, it is giving us the probability that the random variable deviates by more than t times σ . So, it deviates on either side by more than t times σ or t times standard deviation and this quantity is less than 1 by t square.

So, how do we derive this? So, we can derive this by using Markov inequality on the random variables mod of X minus μ . So, if we if we use if we or we can also use the

random variable X minus μ square, if we use the random variable x minus μ square the probability that of this event occurring that X minus μ mod of this is greater than t times σ is nothing, but the probability of X minus μ square being greater than t square times σ square and by Markov inequality this is less than or equal to expectation of X minus μ square by t square σ square.

(Refer Slide Time: 35:09)

Chebyshev's inequality



For every random variable X and every t :

$$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$$

where $\mu = E[X]$, $\sigma = \sqrt{Var[X]}$.

$E[(x-\mu)^2] = \sigma^2$

$$P(|X - \mu| \geq t\sigma) = P((X - \mu)^2 \geq t^2\sigma^2) \leq E[(X - \mu)^2] / t^2\sigma^2 = 1 / t^2.$$

Sourangshu Bhattacharya

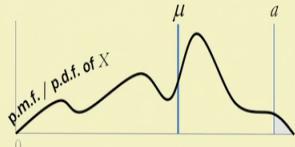
Science and Engg.

And we know that expectation of X minus μ square is nothing but σ square, hence this probability is bounded by 1 by t square.

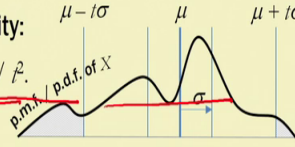
(Refer Slide Time: 35:24)

An illustration



Markov's inequality:

$$P(X \geq a) \leq \mu / a.$$


Chebyshev's inequality:

$$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$$


$E[(x-\mu)^2] = \sigma^2$

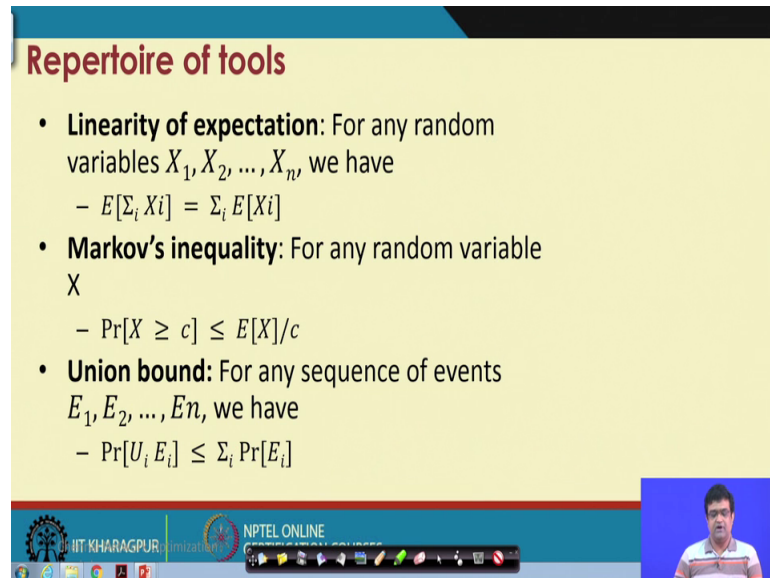



Sourangshu Bhattacharya

Science and Engg.

So, this is an illustration of the kinds of bounds that the Markov inequality or the Chebyshev's inequality gives, ok.

(Refer Slide Time: 25:40)



Repertoire of tools

- **Linearity of expectation:** For any random variables X_1, X_2, \dots, X_n , we have
 - $E[\sum_i X_i] = \sum_i E[X_i]$
- **Markov's inequality:** For any random variable X
 - $\Pr[X \geq c] \leq E[X]/c$
- **Union bound:** For any sequence of events E_1, E_2, \dots, E_n , we have
 - $\Pr[\cup_i E_i] \leq \sum_i \Pr[E_i]$

The slide also features logos for IIT KHARAGPUR and NPTEL ONLINE, along with a small video inset of a presenter in the bottom right corner.

So, we have already seen linearity of expectation and we have also seen the Markov inequality. Another important bound is the union bound which is the which is the bound on probability of union of events which is which will also be heavily used in this course ok, and the bound says that the probability of union of a certain number of a events either they can be independent or they may not be independent is less than the probability of the sum over the probability of each of this events. So, this is the union bound.

(Refer Slide Time: 36:32)

Chernoff bounding

The Chernoff bound for a random variable X is obtained as follows: for any $t > 0$,

$$\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}] \leq E[e^{tX}] / e^{ta}$$

Similarly, for any $t < 0$,

$$\Pr[X \leq a] = \Pr[e^{tX} \geq e^{ta}] \leq E[e^{tX}] / e^{ta}$$

The value of t that **minimizes** $E[e^{tX}] / e^{ta}$ gives the best possible bounds.

When $X = X_1 + \dots + X_n$:

$$\Pr[X \leq a] \leq \min_{t > 0} e^{-ta} \prod_i E[e^{tX_i}]$$

Handwritten notes: $e^{tx} \geq e^{ta}$, $t > 0$, $t < 0$

Video inset: Sourangshu Bhattacharya, IIT KHARAGPUR, NPTEL ONLINE, Science and Eng

Now, we describe our final technique which is the Chernoff bounding technique. The Chernoff bounding technique tries to bound give bound on the probability of a random variable X taking a value greater than a or it can also be derived for random variable X taking a value less than a . And here note that unlike Markov bound there is no the Markov inequality there is no restrictions on the value of the random variable X . So, the random variable x need not be positive.

Now, we can do this simply by raising E to the power t of X and the event that X is greater than a is same as the event that E to the power t of X is t times X rather is greater than the event E to the power t times is greater than E to the power t times a for t greater than 0 . And this using Markov inequality is expectation of E to the power t times X by E to the power t times a .

Similarly, for the event X is less than a we can use a t less than 0 derive the same bound. Now, the interesting thing is that we can actually optimize this bound over all possible values of t to derive a much tighter bound than what could normally be given by Markov inequality. So, for example, if this random variable X is sum over X_1 till X_N then you can see that the Chernoff bounding technique can be used to derive a bound on some of random variables being less than a particular value a as the minimum over t greater than or greater than 0 E to the power minus t a times product over all i is equal to 1 to N

expectation of E to the power t is X . So, here the random variables x_i has to be independent, ok.

(Refer Slide Time: 39:30)

Chernoff bounding

- **Def:** The **moment generating function** of a random variable X is $M_X(t) = E[e^{tX}]$.
- $E[X^n] = M_X^n(0)$, which is the n th derivative of $M_X(t)$ evaluated at $t = 0$.
- Fact: If $M_X(t) = M_Y(t)$ for all t in $(-c, c)$ for some $c > 0$, then X and Y have the same distribution.
- If X and Y are independent r.v., then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
79

So, the function E to the power t is actually called the moment generating function because the derivative of so.

(Refer Slide Time: 39:47)

Chernoff bounding

- Let X_1, X_2, \dots, X_n be n independent random variables in $\{0,1\}$, with $X = X_1 + X_2 + \dots + X_n$.
- For any nonnegative δ $\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$
- For any δ in $[0,1]$ $\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu\delta^2/3}$

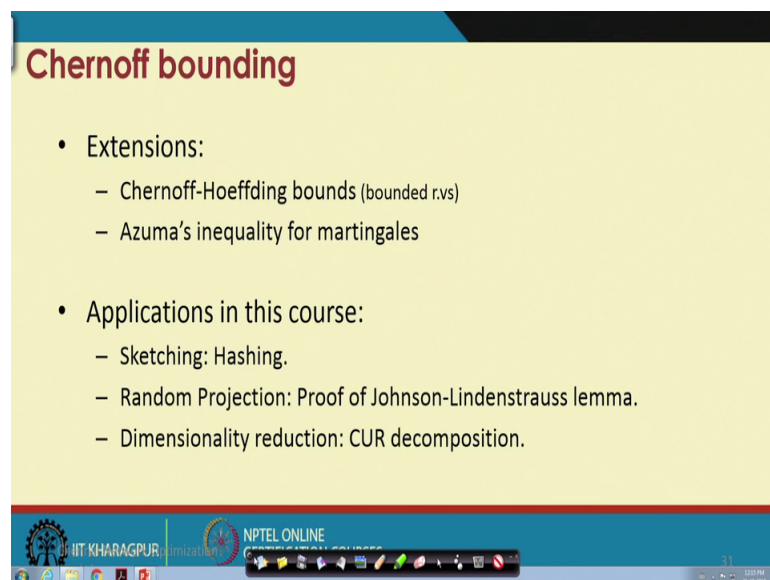
IIT KHARAGPUR | NPTEL ONLINE | 30

So, if we call this function as M of x then the derivative of M of the rather the N X derivative of M of X actually gives the n th moment of the distribution function of E to the power t of X . So, these are some of the results that you can derive using Chernoff

bounding technique. The first one is that if you have X as sum of X_1 till X_N and all these are independent binary variables.

Then with mean μ which is also the probability of these being 1, then the probability of that sum being greater than $1 + \delta$ times μ is always less than or equal to E to the power δ by $1 + \delta$ to the power δ whole to the lower μ . And this is another simpler form where we can show that the probability of X minus μ being greater than δ times μ is less than or equal to 2 to the power 2 times E power minus μ delta square by 3 .

(Refer Slide Time: 41:26)



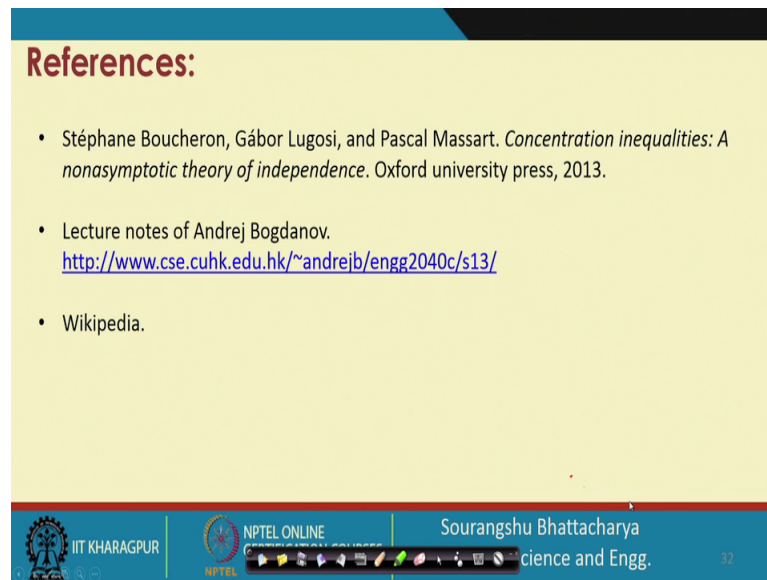
Chernoff bounding

- Extensions:
 - Chernoff-Hoeffding bounds (bounded r.v.s)
 - Azuma's inequality for martingales
- Applications in this course:
 - Sketching: Hashing.
 - Random Projection: Proof of Johnson-Lindenstrauss lemma.
 - Dimensionality reduction: CUR decomposition.

The slide is part of an NPTEL ONLINE presentation from IIT KHARAGPUR. The footer includes the IIT Khargapur logo, the text 'NPTEL ONLINE', and a taskbar with various application icons.

So, this brings us to the end of the Chernoff bounding techniques. The reason we have discuss this Chernoff bounding technique is because these techniques will be used in various portion or various topics in this course especially in the topic of sketching in the topic of random projection and in the topic of dimensionality reduction.

(Refer Slide Time: 42:05)



References:

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Lecture notes of Andrej Bogdanov.
<http://www.cse.cuhk.edu.hk/~andrejb/engg2040c/s13/>
- Wikipedia.

IIT KHARAGPUR | NPTEL ONLINE | Sourangshu Bhattacharya
science and Engg. 32

And these are the references for this particular lecture. So, some of the material is borrowed from the lecture notes of sorry, some of the material are borrowed from the lecture notes of Andrej Bogdanov and the book by Boucheron Lugosi and Massart is very nice book on the topic.

Thank you.