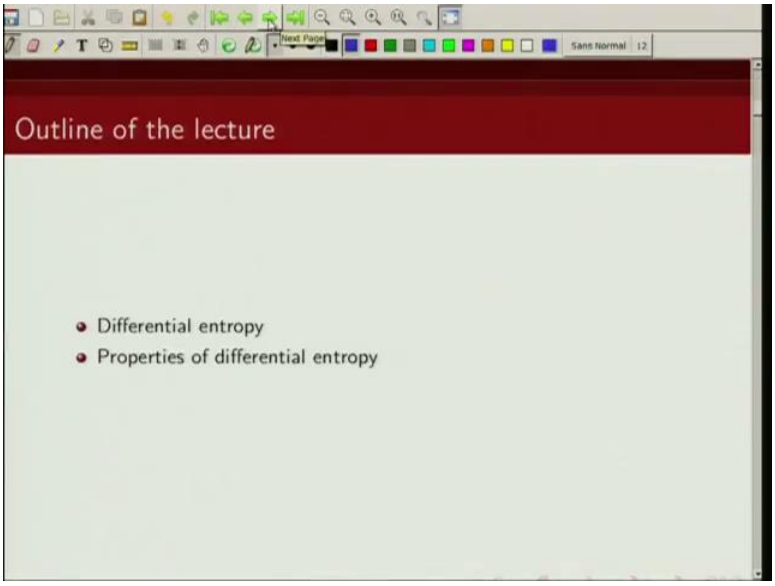
An Introduction to Information Theory

Prof. Adrish Banerjee
Department of Electronics and Communication Engineering
Indian Institute of Technology, Kanpur

Lecture – 11
Differential Entropy

Welcome to the course on An Introduction to Information Theory. So far we have been talking about discrete random variable and entropy associated with discrete random variable. Now, today we are going to talk about entropy for continuous random variable and that is known as differential entropy. So, it’s the entropy of continuous random variable.

(Refer Slide Time: 00:41)

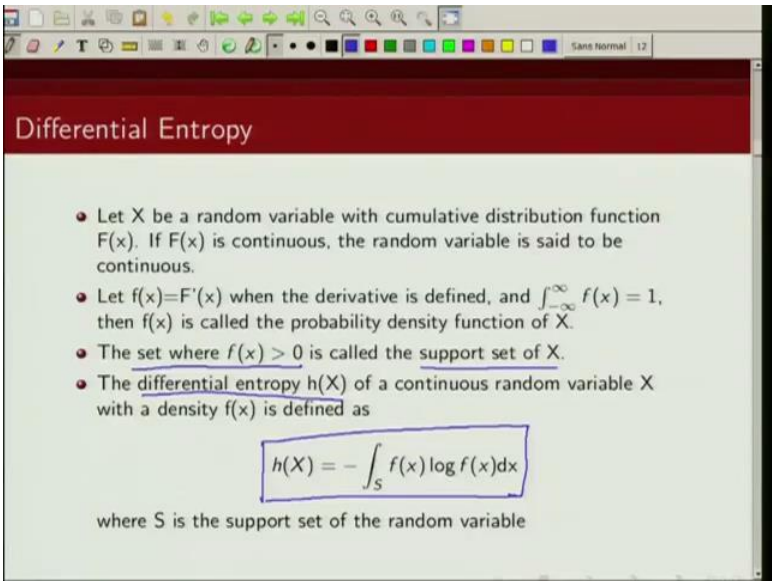


Outline of the lecture

- Differential entropy
- Properties of differential entropy

In this class, we will define differential entropy and we will prove some properties of differential entropy.

(Refer Slide Time: 00:52)



Differential Entropy

- Let X be a random variable with cumulative distribution function F(x). If F(x) is continuous, the random variable is said to be continuous.
- Let f(x)=F'(x) when the derivative is defined, and $\int_{-\infty}^{\infty} f(x) = 1$, then f(x) is called the probability density function of X.
- The set where $f(x) > 0$ is called the support set of X.
- The differential entropy h(X) of a continuous random variable X with a density f(x) is defined as

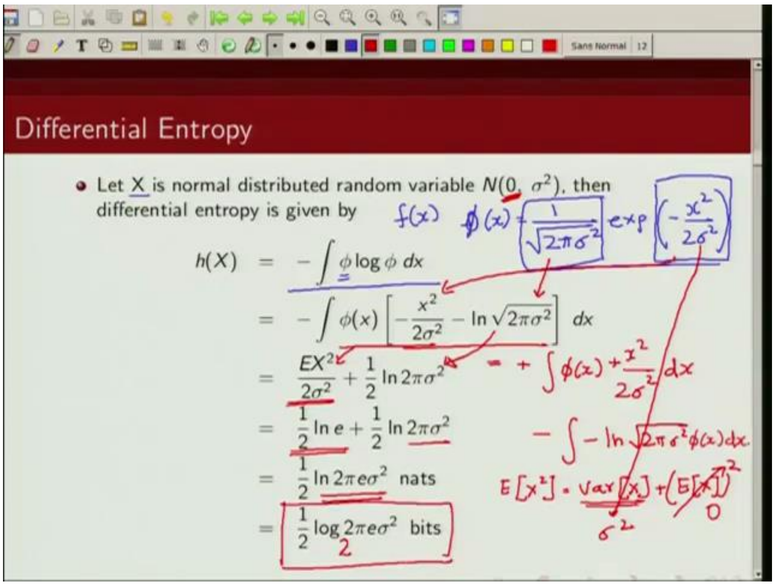
$$\boxed{h(X) = -\int_S f(x)\log f(x)dx}$$

where S is the support set of the random variable

Let x be a random variable whose cumulative distribution function is given by F of x. If F of x is continuous, then we say that random variable is continuous. So, if derivative of F of x exists, n is defined which has the following property that integration from minus infinity to infinity is 1, then F of x is known as the probability density function of x. The set over which F of x is greater than zero is known as support set of x. These definitions are very similar to the definitions we had for discrete case and if you recall for the discrete random variable, we define the support set as the set where probability p of x is greater than 0. Now, differential entropy of a continuous random variable is defined as minus of integration of this probability density function log of probability density function, integration over the support set of this random variable.

So, this is how we define differential entropy and if you recall for discrete random variable, we define entropy as minus of summation p of x log of p of x. This is very similar to that.

(Refer Slide Time: 02:34)



Differential Entropy

- Let X is normal distributed random variable $N(0, \sigma^2)$, then differential entropy is given by $f(x)$ $\phi(x) = \boxed{\frac{1}{\sqrt{2\pi\sigma^2}}} \exp \boxed{\left(-\frac{x^2}{2\sigma^2}\right)}$

$$
\begin{aligned}
h(X) &= -\int \phi \log \phi \, dx \\
&= -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] dx \\
&= \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2 \\
&= \frac{1}{2} \ln e + \frac{1}{2} \ln 2\pi\sigma^2 \\
&= \frac{1}{2} \ln 2\pi e\sigma^2 \text{ nats} \\
&= \boxed{\frac{1}{2} \log_2 2\pi e\sigma^2 \text{ bits}}
\end{aligned}
$$

$$= + \int \phi(x) + \frac{x^2}{2\sigma^2} dx$$

$$- \int - \ln \sqrt{2\pi\sigma^2} \phi(x) dx.$$

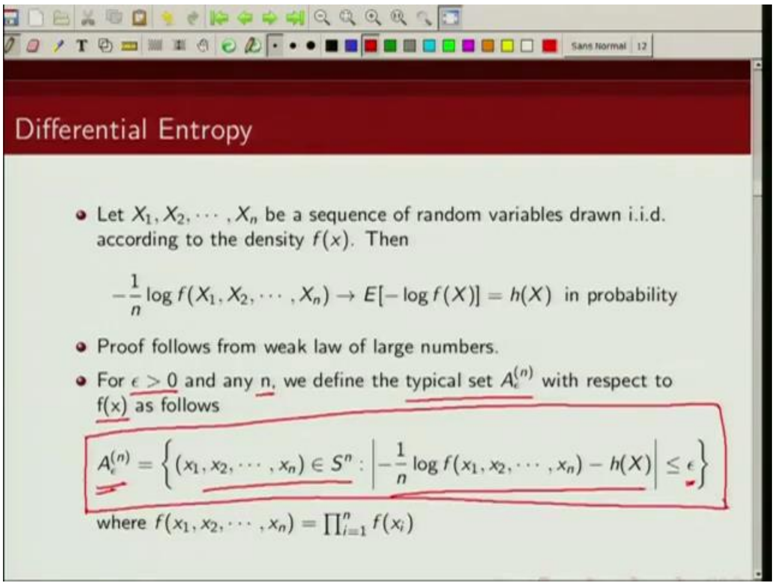
$$E[x^2] = \underbrace{\mathrm{Var}[x]}_{\sigma^2} + (\underbrace{\cancel{E[x]}}_{0})^2$$

Now, let us take an example. Let X is a normal distributed random variable. So, the density function phi of x is given by pi sigma square exponential minus x square by 2 sigma square. So, this is a zero mean variance sigma square normal distributed random variable whose density function can use the term f of x or I am using here phi of x is given by this expression. So, differential entropy is defined in this particular fashion. This can be written as phi of x, now if you take log of phi of x, we get log of this term and log exponential will be just this term. We get minus phi of x and log of phi is given by this expression. So, this and log of 1 by under root of 2 pi sigma square and now this can be written as minus phi of x minus x square 2 sigma square d of x and minus of minus actual log of 2 pi sigma square phi of x d x.

So, this minus minus becomes plus, this integration phi x x square 2 sigma square d x, it is nothing but expected value of x square by 2 sigma square. This 1st term that you see here is nothing but expected value of x square divided by 2 sigma square and there is no x here, so, this is a constant and if you integrate over phi of x we will get 1. So, this term is nothing but given by this. Now, the expected value of x square is given by variance of x plus expected value of x whole square and the expected value of x is 0. This term is 0 and variance is given by sigma square. This is variance sigma square and expected value of x square is given by sigma square.

So, sigma square sigma square cancels out, what we will get is half and I can also write it as half is half of natural log of e and this is of the form half of natural log of e plus half of natural log of 2 pi sigma square. This is log a plus log b and this can be written as log a times b and then this term be written like this. Now, note that we are talking about natural logs. So, this is in nats, if we take log to the base 2, then the units will be bits. The differential entropy of a 0 mean Gaussian random variable with variance sigma square is given by this expression half of log 2 pi e sigma square.

(Refer Slide Time: 06:28)



Differential Entropy

- Let $X_1, X_2, \cdots, X_n$ be a sequence of random variables drawn i.i.d. according to the density $f(x)$. Then

$$-\frac{1}{n}\log f(X_1, X_2, \cdots, X_n) \to E[-\log f(X)] = h(X) \text{ in probability}$$

- Proof follows from weak law of large numbers.
- For $\epsilon > 0$ and any n, we define the typical set $A_\epsilon^{(n)}$ with respect to f(x) as follows

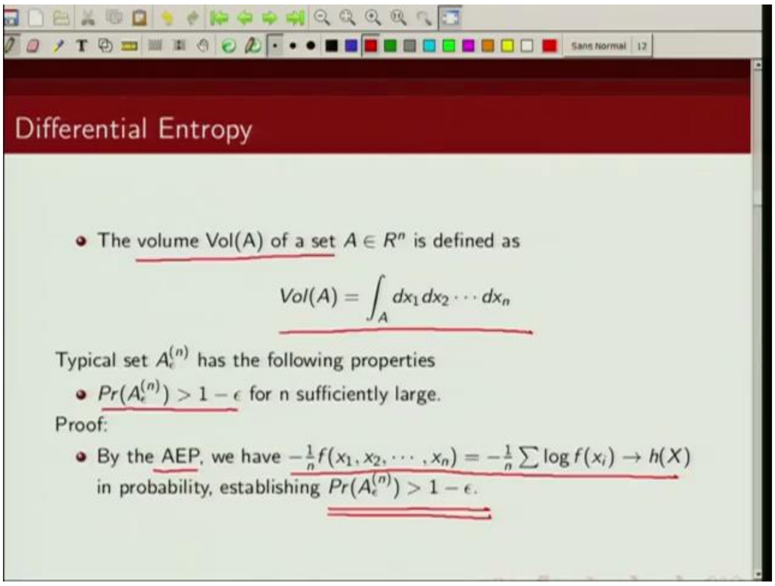
$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \cdots, x_n) \in S^n : \left| -\frac{1}{n}\log f(x_1, x_2, \cdots, x_n) - h(X) \right| \le \epsilon \right\}$$

where $f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$

Now, if x 1 x 2 x 3 are sequence of random variables drawn independent identically distributed according to distribution f of x then, 1 minus n log of f of X 1 X 2 X 3 X n, this converges to differential entropy in probability. The proof of this is very similar to the proof that we did in the discrete case, this follows from the weak law of large number. So, this can be written as product of f x i and when you take log of product term, you get summation and then you will get summation of minus 1 by n log of f x i and that by weak law of large numbers is basically expected value of this which is equal to differential entropy and this converges to differential entropy in probability.

So, we are skipping this proof because this is very similar to what we have done earlier now for an epsilon greater than 0 for any n we can define typical set with respect to this density function as follows. Typical set is defined as x 1 x 2 x 3 x n belonging to this set such that minus 1 by n log of f x 1 x 2 x 3, basically it is close to the true differential entropy and the difference is within epsilon.

(Refer Slide Time: 08:35)



Differential Entropy

- The volume Vol(A) of a set $A \in R^n$ is defined as

$$Vol(A) = \int_A dx_1 dx_2 \cdots dx_n$$

Typical set $A_\epsilon^{(n)}$ has the following properties

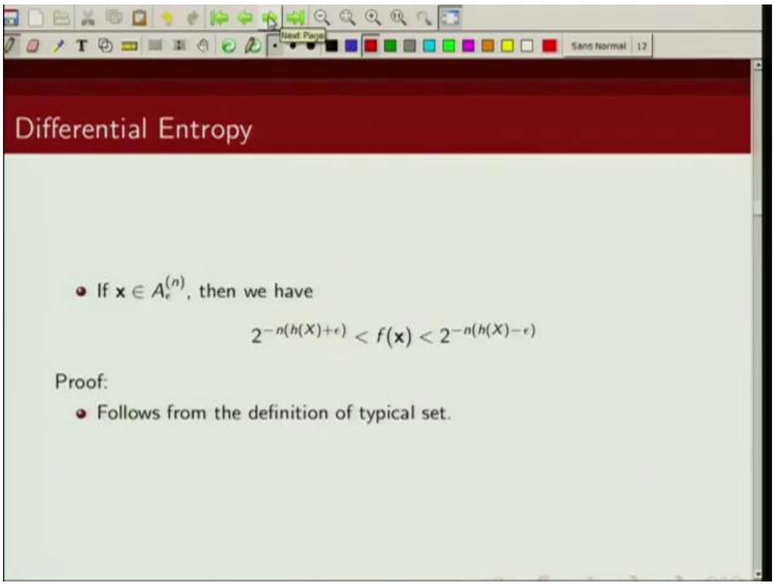
- $Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large.

Proof:

- By the AEP, we have $-\frac{1}{n} f(x_1, x_2, \cdots, x_n) = -\frac{1}{n} \sum \log f(x_i) \rightarrow h(X)$ in probability, establishing $Pr(A_\epsilon^{(n)}) > 1 - \epsilon$.

So, very similar to how we defined typical set, we for the discrete case and for the continuous random variable also we can define typical set. Now, in case of discrete random variable, we define how many such typical sequence is there, the similar analogous thing for continuous random variable is what we call volume of a set. So, volume of a set is defined as analogous to the number of typical sequence here we have the volume of the typical set. Now, typical set has following properties, the probability of typical sequence for very large n that is greater than one minus epsilon, now again this follows from AEP property that we know that this converges to differential entropy in probability and this establishes the proof that probability of this typical set is greater than 1 minus epsilon.

(Refer Slide Time: 09:43)



Differential Entropy

- If $\mathbf{x} \in A_\epsilon^{(n)}$, then we have

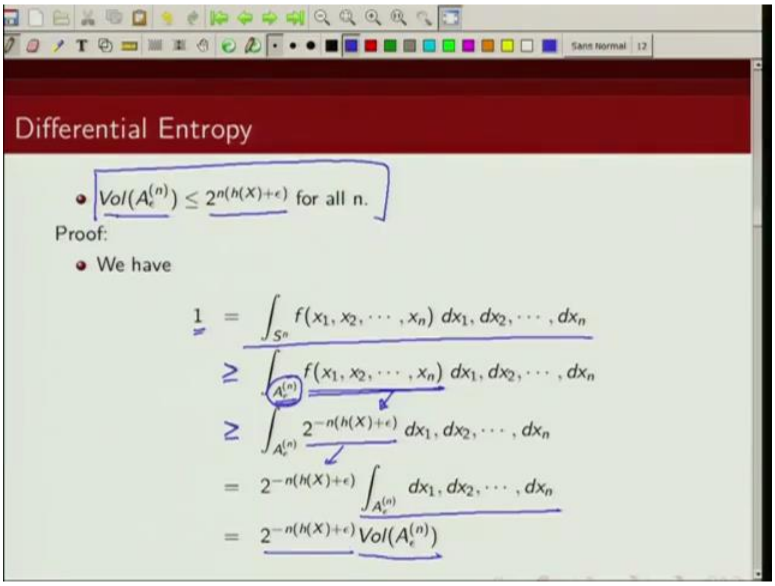
$$2^{-n(h(X)+\epsilon)} < f(\mathbf{x}) < 2^{-n(h(X)-\epsilon)}$$

Proof:

- Follows from the definition of typical set.

The next property which says the f of x lies between this and this. Now, this can be proved from the definition of typical set. So, if you go back and look at the definition of typical set, this empirical differential entropy, the difference between that and the two differential entropy. The absolute difference should be less than or equal to epsilon.

(Refer Slide Time: 10:48)



Differential Entropy

- $\boxed{\underline{Vol(A_\epsilon^{(n)})} \leq \underline{2^{n(h(X)+\epsilon)}} \text{ for all n.}}$

Proof:

- We have

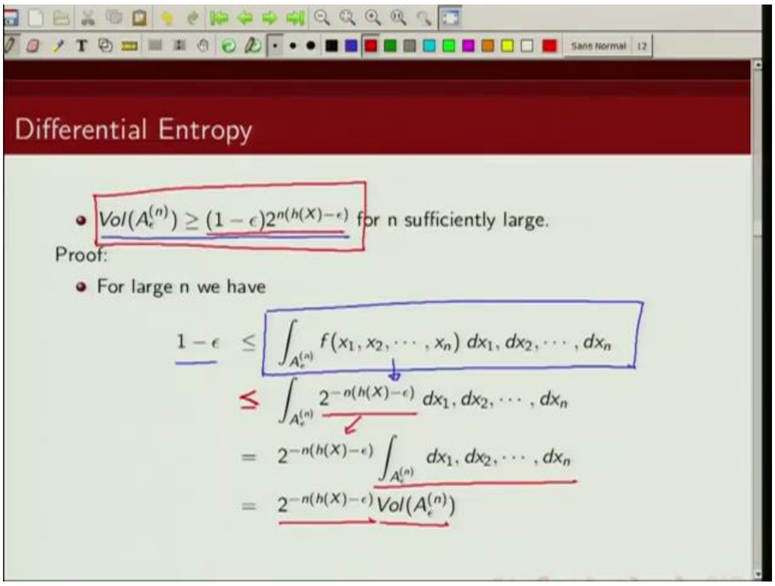
$$\begin{aligned}
\underline{1} &= \underline{\int_{S^n} f(x_1, x_2, \cdots, x_n)\, dx_1, dx_2, \cdots, dx_n} \\
&\geq \int_{A_\epsilon^{(n)}} \underline{f(x_1, x_2, \cdots, x_n)}\, dx_1, dx_2, \cdots, dx_n \\
&\geq \int_{A_\epsilon^{(n)}} \underline{2^{-n(h(X)+\epsilon)}}\, dx_1, dx_2, \cdots, dx_n \\
&= 2^{-n(h(X)+\epsilon)} \underline{\int_{A_\epsilon^{(n)}} dx_1, dx_2, \cdots, dx_n} \\
&= \underline{2^{-n(h(X)+\epsilon)}}\, \underline{Vol(A_\epsilon^{(n)})}
\end{aligned}$$

So, depending on whether this term is greater or this term is greater because we are taking absolute difference and we will get either this inequality or we will get this inequality. Again the proof is similar to the proofs we have done many times before so, I am just skipping the details of the proof. The next property that we are going to show is the volume of the typical set is upper bounded by this.

Now, we know that if you integrate this density function over the support set, you will get probability 1 and if we integrate over typical set then since the typical set is subset of this S of n. So, we get here this greater than equal to because we are integrating over a smaller set and that is why this greater than equal to sign comes. We know that this density function is lower bounded by this. So, if we plug in this lower bound on f of x 1 x 2 x n here, what we get is this and that is why we are writing it as greater than equal to. We can take this out and if we integrate it over the typical set, what we get is a volume of the typical set and volume of the typical set multiplied by this is less than equal to 1.

So, from here we get volume of typical set is upper bounded by this 2 raise power n times differential entropy plus epsilon. Similarly, we can prove that volume of a typical set is lower bounded by this or just follows from we know probability of this typical set is greater than equal to 1 minus epsilon. So, 1 minus epsilon is less than equal to this probability of typical set which is given by this expression and now we know the upper bound on this f of x 1 x 2 x 3 x n, this follows from this property. So, we know the upper bound on f of x.

(Refer Slide Time: 13:24)



Differential Entropy

- $\boxed{\underline{Vol(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}}}$ for n sufficiently large.

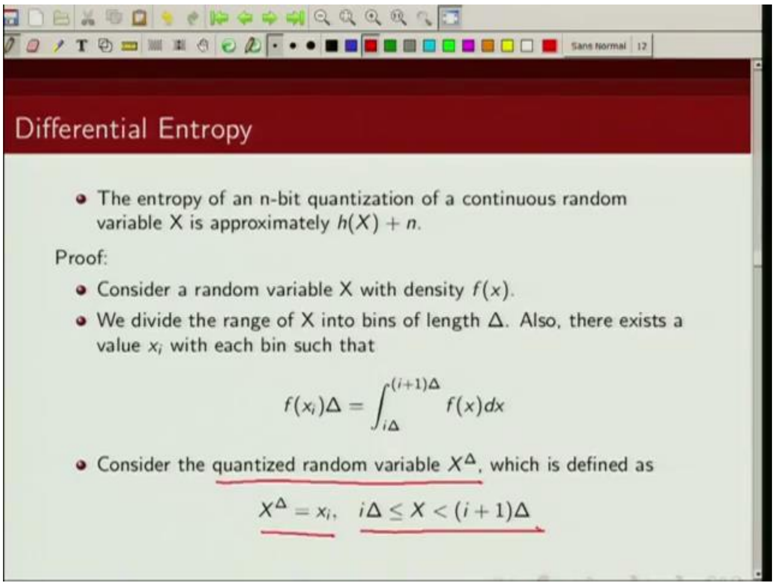
Proof:

- For large n we have

$$
\begin{aligned}
\underline{1-\epsilon} \quad &\leq \quad \boxed{\int_{A_\epsilon^{(n)}} f(x_1, x_2, \cdots, x_n)\, dx_1, dx_2, \cdots, dx_n} \\
&\leq \quad \int_{A_\epsilon^{(n)}} \underline{2^{-n(h(X)-\epsilon)}}\, dx_1, dx_2, \cdots, dx_n \\
&= \quad 2^{-n(h(X)-\epsilon)} \underline{\int_{A_\epsilon^{(n)}} dx_1, dx_2, \cdots, dx_n} \\
&= \quad \underline{2^{-n(h(X)-\epsilon)}}\, \underline{Vol(A_\epsilon^{(n)})}
\end{aligned}
$$

Now, if we plug in the upper bound of f of x 1 x 2 x n here, we get here less than equal to and subsequently you take this out and integrate it over the typical set what we get is volume of typical set multiplied by this and this is greater than equal to 1 minus epsilon. So, from here we get volume of typical set is at least given by this.

(Refer Slide Time: 14:02)



Differential Entropy

- The entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

Proof:

- Consider a random variable X with density $f(x)$.
- We divide the range of X into bins of length Δ. Also, there exists a value x_i with each bin such that

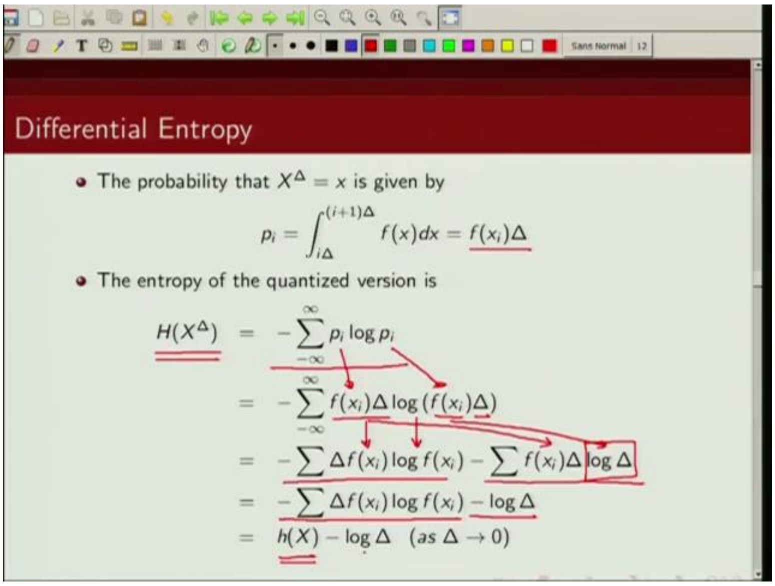
$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$$

- Consider the $\underline{\text{quantized random variable } X^\Delta}$, which is defined as

$$\underline{X^\Delta = x_i}, \quad \underline{i\Delta \leq X < (i+1)\Delta}$$

Now, what is the relation between differential entropy and discrete? So, we are comparing the differential entropy of a continuous random variable with the entropy of a discrete random variable that we get by discretizing that continuous random variable. This result says that entropy of n bit quantization of a continuous random variable is approximately given by differential entropy of X plus n. Let us consider a random variable X whose density function is given by f of x. Now, as we said we want to compare the entropy of the continuous random variable with the discretized version of this continuous random variable. So, we divide the range of X into bins of length delta and from mean value theorem, we know that there exist an x i such that f of x i times delta is equal to this.

(Refer Slide Time: 15:37)



Differential Entropy

- The probability that $X^{\Delta} = x$ is given by

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = \underline{f(x_i)\Delta}$$

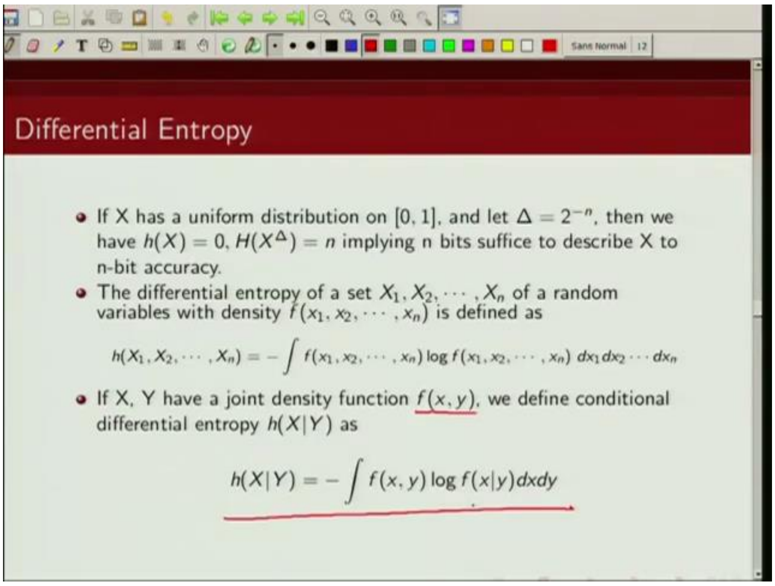
- The entropy of the quantized version is

$$\begin{aligned}
\underline{\underline{H(X^{\Delta})}} &= \underline{-\sum_{-\infty}^{\infty} p_i \log p_i} \\
&= -\sum_{-\infty}^{\infty} \underline{f(x_i)\Delta} \log (\underline{f(x_i)}\underline{\Delta}) \\
&= \underline{-\sum \Delta f(x_i) \log f(x_i)} - \underline{\sum f(x_i)\Delta \boxed{\log \Delta}} \\
&= \underline{-\sum \Delta f(x_i) \log f(x_i)} \underline{- \log \Delta} \\
&= \underline{\underline{h(X)}} - \log \Delta \quad (as\ \Delta \to 0)
\end{aligned}$$

Then we are considering a quantized random variable and calling it x delta which is equal to X i if X lies between i delta and i plus 1 delta and the probability that X delta is equal to x is then given by this f of x d x integrate over this interval i delta to i plus 1 delta which by mean value theorem is given by f x i into delta. So, let us compute the entropy of this quantized random variable X delta. This is given by this expression which it follows from the definition of entropy, now p i is given by f x i delta. So, we plug in the value of p i here, we plug in the value of p i here and here we have log of a times b. This can be written as log of a plus log of b. So, then log of f x i delta can be written as log of f x i and log of delta and this comes here, this comes here.

So, what we have here is this term plus this term, now this term does not depend on x i s and if you sum f x delta I, this will basically be 1 and we get minus log of delta and this term. Thus, the small delta approximates to the differential entropy.

(Refer Slide Time: 17:37)



Differential Entropy

- If X has a uniform distribution on $[0, 1]$, and let $\Delta = 2^{-n}$, then we have $h(X) = 0$, $H(X^{\Delta}) = n$ implying n bits suffice to describe X to n-bit accuracy.
- The differential entropy of a set $X_1, X_2, \cdots, X_n$ of a random variables with density $f(x_1, x_2, \cdots, x_n)$ is defined as

$$h(X_1, X_2, \cdots, X_n) = -\int f(x_1, x_2, \cdots, x_n) \log f(x_1, x_2, \cdots, x_n)\, dx_1 dx_2 \cdots dx_n$$

- If X, Y have a joint density function $f(x, y)$, we define conditional differential entropy $h(X|Y)$ as

$$h(X|Y) = -\int f(x, y) \log f(x|y) dxdy$$

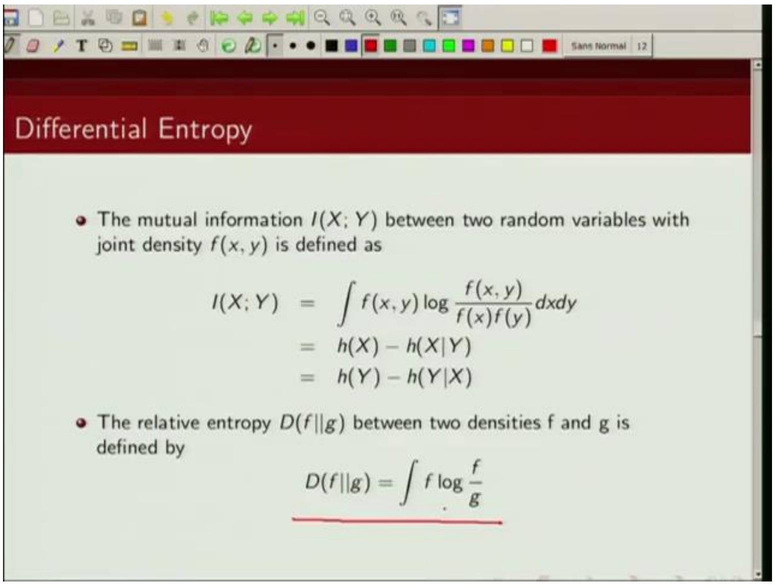
We can write the entropy of the quantized random variable to be equal to differential entropy minus log of delta when delta is very small and if we assume X to have uniform distribution between 0 and 1 and we take our delta to be 2 raise power minus n, then in this case differential entropy is given by log of a which is a log of 1 which is 0 and log of delta minus log of delta will be n. So, H of delta will be h of X plus h of X minus log of delta minus log of delta is n, h of X is zero and then H of X delta will be n which implies that basically n bits are sufficient to describe X within n-bit accuracy.

Now, here we have made use of the fact that what is the differential entropy of a uniformly distributed random variable. So, we have a random variable which is uniformly distributed between zero to a, its differential entropy from the definition h of X is given by minus our support group f of x log of f of x d of x, now if it is uniformly distributed between 0 to a, it looks like this. So, 0 to a is uniformly distributed, this will be 1 by a, this is my f of x, this will be minus 1 by a log of 1 by a d x and this integration will be found 0 to a. So, this can be written as 1 by a integration log of a d x 0 to a, this is 1 by a into a into log of a.

So, this will be log of a, in this case a is 1, then log of 1 is 0 and that is why I wrote differential entropy for X which is uniformly distributed is between 0 to 1, its differential entropy is 0. One point I just wanted to make which is different for differential entropy compared to the entropy for discrete random variable is, for discrete random variable the entropy is greater than equal to zero, here the entropy can be less than zero. For example, if a is a fraction, let us say a is half then entropy in this case will be log of half which is minus log of 2 which is a negative quantity. So, differential entropy can be negative. However, volume of a typical set is positive.

Next, we are going to define differential entropy for random variable X 1 X 2 X 3 X n, so, differential entropy for a set x 1 x 2 x n whose joint density is given by this and can be similarly defined like this. So, it is a very straight forward extension of definition of differential entropy with single random variable. Similarly, we can also define conditional differential entropy. So, if X and Y have joint density function given by this, then we can define conditional differential entropy in this fashion. Again this is, if you look at the form of this expression, it is very similar to the form that we had for the discrete case.

(Refer Slide Time: 21:55)



Differential Entropy

- The mutual information $I(X;Y)$ between two random variables with joint density $f(x,y)$ is defined as

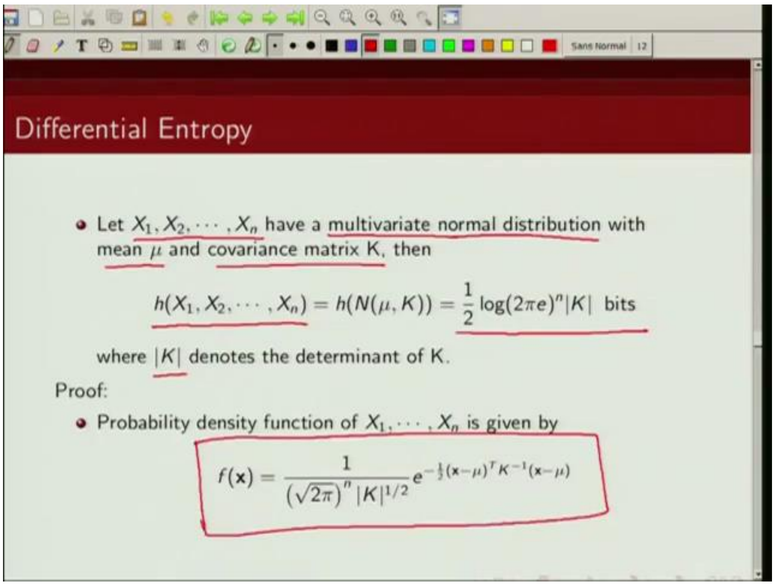
$$\begin{aligned} I(X;Y) &= \int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy \\ &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \end{aligned}$$

- The relative entropy $D(f||g)$ between two densities f and g is defined by

$$\underline{D(f||g) = \int f\log\frac{f}{g}}$$

Similarly, we can define mutual information. So, mutual information between two random variable with joint density function given by this is nothing but divergence between f x y and the marginal's f x and f y and this can be written as differential entropy of X minus conditional differential entropy of X given Y or differential entropy of Y minus differential entropy of Y given X.

(Refer Slide Time: 22:56)



Differential Entropy

- Let $X_1, X_2, \cdots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix K, then

$$\underline{h(X_1, X_2, \cdots, X_n)} = h(N(\mu, K)) = \underline{\frac{1}{2}\log(2\pi e)^n|K| \text{ bits}}$$

where $|K|$ denotes the determinant of K.

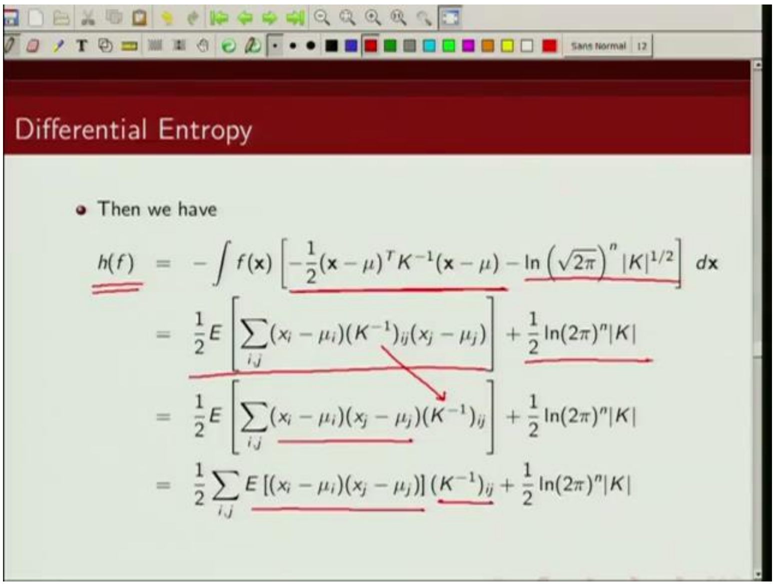
Proof:

- Probability density function of $X_1, \cdots, X_n$ is given by

$$\boxed{f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n|K|^{1/2}}e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}}$$

Similarly, we could define the divergence between two densities f and g. This is defined as expected value f of log of f by g. This is the divergence between f and g. Now, let us consider an example, we are considering a multivariate normal distribution. So, X 1 X 2 X 3 X n have a multivariate normal distribution with mean given by mu and covariance matrix given by X show that the joint differential entropy is given by this expression where this is nothing but determinate of K, the covariance matrix. So, since X 1 X 2 X 3 X n follows a multivariate normal distribution such that density function is given by this expression.

(Refer Slide Time: 23:46)



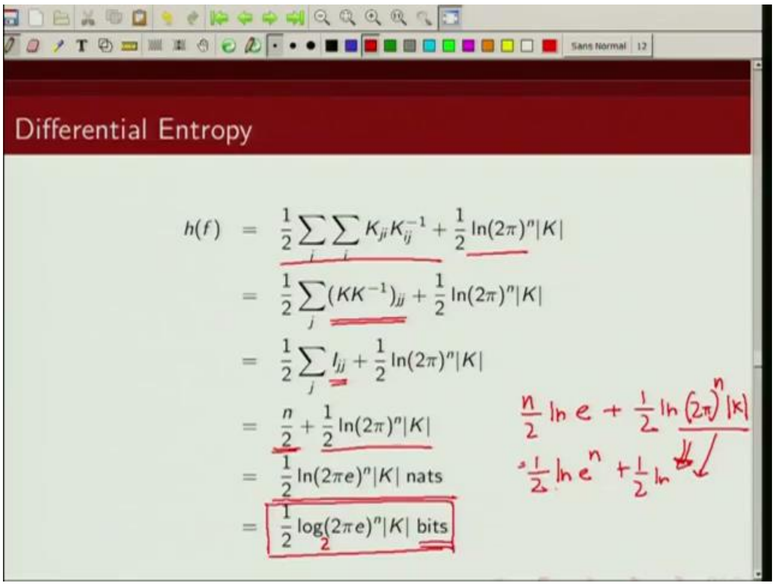
Differential Entropy

- Then we have

$$
\begin{aligned}
\underline{\underline{h(f)}} &= -\int f(\mathbf{x}) \left[\underline{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)} - \underline{\ln \left(\sqrt{2\pi}\right)^n |K|^{1/2}} \right] d\mathbf{x} \\
&= \underline{\frac{1}{2} E \left[\sum_{i,j} (x_i - \mu_i)(K^{-1})_{ij}(x_j - \mu_j) \right]} + \underline{\frac{1}{2} \ln (2\pi)^n |K|} \\
&= \frac{1}{2} E \left[\sum_{i,j} \underline{(x_i - \mu_i)(x_j - \mu_j)}(K^{-1})_{ij} \right] + \frac{1}{2} \ln (2\pi)^n |K| \\
&= \frac{1}{2} \sum_{i,j} E\, \underline{[(x_i - \mu_i)(x_j - \mu_j)]}\, \underline{(K^{-1})_{ij}} + \frac{1}{2} \ln (2\pi)^n |K|
\end{aligned}
$$

Next, we apply the definition of joint differential entropy. So, this will be minus f of x log of f of x d x. When we take log of this function, we get log of this plus this term minus half x minus mu transpose K inverse x minus mu. So, this is what I am writing here and I have minus half x minus mu transpose K inverse x minus mu that is one term and then we had this minus of 1 by under root 2 pi n determinant of K. So, we have these two terms. Now, second term if I integrate over f of x, I will just get this and let us look at the first term which is nothing but expected value of this. So, in matrix form I can write this in this particular form, the summation over i and j, x i minus mu i K inverse i j x j minus mu j. So, after I can keep it here and I can combine these two terms, this can be written as expected value of this multiplied by K inverse of this.

(Refer Slide Time: 25:25).



Differential Entropy

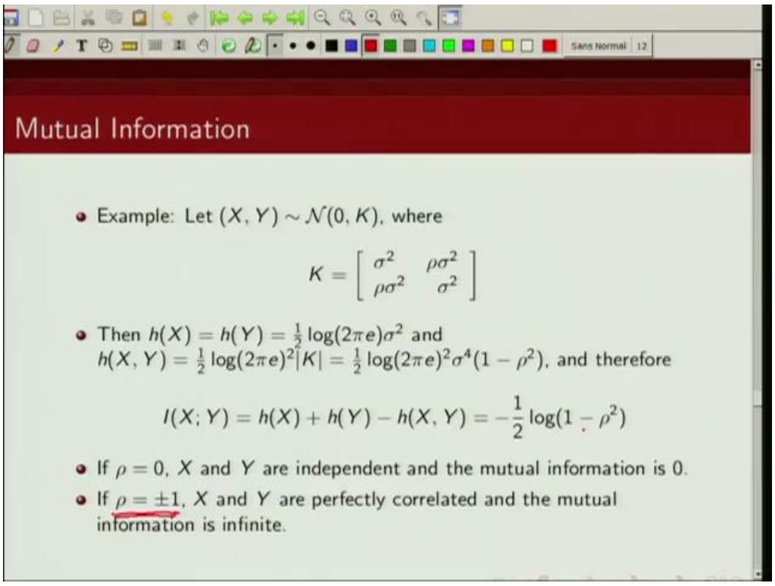
$$
\begin{aligned}
h(f) &= \underline{\frac{1}{2}\sum_i\sum_j K_{ji}K_{ij}^{-1}} + \frac{1}{2}\underline{\ln(2\pi)^n}|K| \\
&= \frac{1}{2}\sum_j \underline{(KK^{-1})_{jj}} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \frac{1}{2}\sum_j \underline{I_{jj}} + \frac{1}{2}\ln(2\pi)^n|K| \\
&= \underline{\frac{n}{2}} + \underline{\frac{1}{2}\ln(2\pi)^n|K|} \\
&= \underline{\frac{1}{2}\ln(2\pi e)^n|K| \text{ nats}} \\
&= \boxed{\frac{1}{2}\log_2(2\pi e)^n|K| \text{ bits}}
\end{aligned}
$$

$$\frac{n}{2}\ln e + \frac{1}{2}\ln \underline{(2\pi)^n|K|}$$

$$= \frac{1}{2}\ln e^n + \frac{1}{2}\ln$$

Next, I can write this as K j i and K i j inverse. So, this will be non-zero only for those terms where i is equal to j and that is basically this and this is going to be i. So, summation over all j, this will give me n and this one I am getting n by 2 and then this term is coming from earlier, so we get this. Now, n by 2 I can write as n by 2 natural log of e and then I have this term which is half of natural log of 2 pi n and determinant of K. This I can also write as half of natural log of e raise power n plus this term. Now, I have half log of this particular term here. So, log of a plus log of b kind of form and this can be written as log of a b. So, this can be combined into this, now if I write unit in terms of log to the base 2, I get my differential entropy to be this. For a multivariate normal distribution with mean mu and co variance matrix K, the differential entropy is given by this.

(Refer Slide Time: 27:08)



Mutual Information

- Example: Let $(X, Y) \sim \mathcal{N}(0, K)$, where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

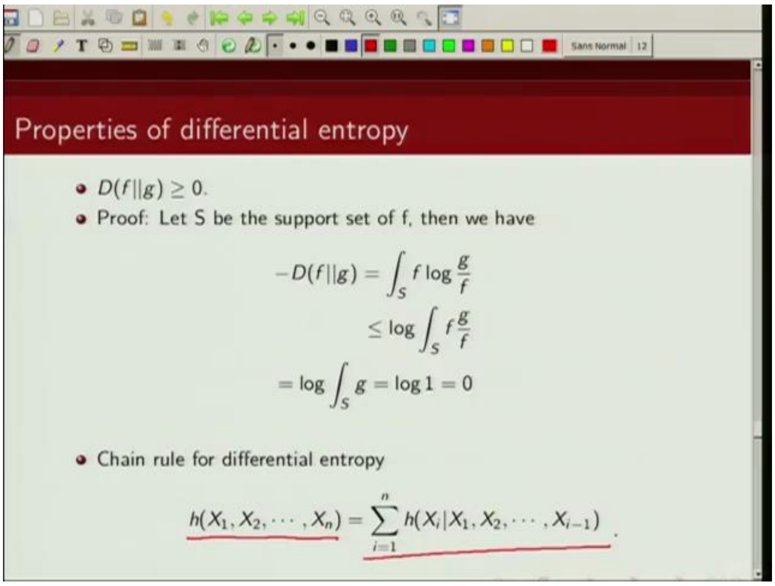
- Then $h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$ and $h(X, Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}\log(2\pi e)^2\sigma^4(1 - \rho^2)$, and therefore

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2}\log(1 - \rho^2)$$

- If $\rho = 0$, X and Y are independent and the mutual information is 0.
- If $\underline{\rho = \pm 1}$, X and Y are perfectly correlated and the mutual information is infinite.

Let us take an example. So, X and y are multivariate Gaussian distributed mean 0, co variance matrix given by K which is this. We can write h of X and h of Y which is nothing but half of log 2 pi sigma square and similarly you can write joint differential entropy which is given by half of log 2 pi square and determinant of k which is given by this and now we can write mutual information also as h of X plus h of Y minus joint entropy and this comes out to be this. Now, if rho 0 which is basically these terms of 0, you can see X and Y are independent and in that case the mutual information between X and Y is going to be 0 because X and Y are independent random variables and if rho is plus minus 1 that means they are perfectly correlated then you can see mutual information is going to be infinite.

(Refer Slide Time: 28:39)



Properties of differential entropy

- $D(f||g) \geq 0$.
- Proof: Let S be the support set of f, then we have

$$
\begin{aligned}
-D(f||g) &= \int_S f \log \frac{g}{f} \\
&\leq \log \int_S f \frac{g}{f} \\
&= \log \int_S g = \log 1 = 0
\end{aligned}
$$

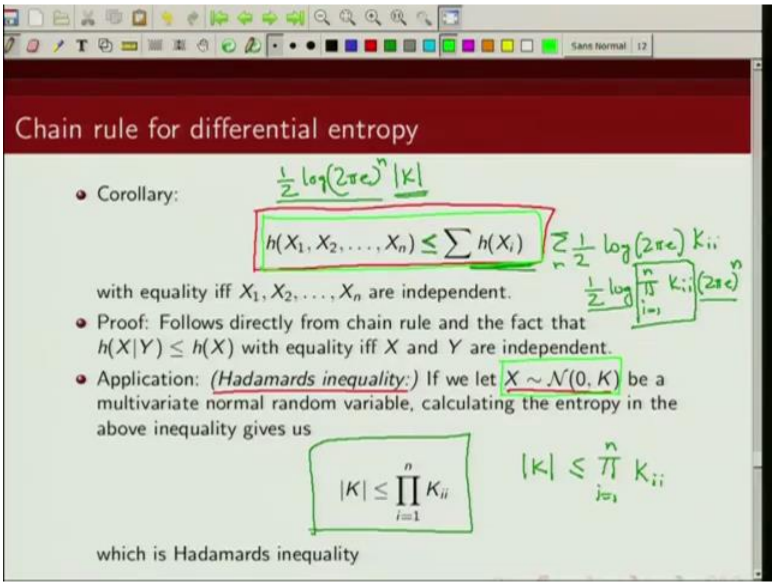
- Chain rule for differential entropy

$$\underline{h(X_1, X_2, \cdots, X_n)} = \underline{\sum_{i=1}^{n} h(X_i|X_1, X_2, \cdots, X_{i-1})}.$$

Now, similar to the discrete random variable case, here also the divergence is greater than equal to 0. So, divergence between two density function f and g is greater than equal to 0, again we can prove it in the similar fashion, it is minus of divergence between f and g which is given by this expression, now log is a concave function if you recall log is a concave function, log is like this and from Jensen’s inequality, the expected value of a function is less than equal to function evaluated at expected value and if it is a concave function then Jensen’s inequality says this. So, this is like expected value of this log function. Since, log is a concave function then by Jensen’s inequality, the log of expected value of the X should be more than expected value of the function.

From Jensen’s inequality, we know this relation holds if f of x is concave, if this is a concave function then this expected value of log will be less than equal to log expected value and this basically is then integration of g over the support set is 1. So, log of 1 will be 0 and what we have proved is minus of divergence is less than equal to 0 and if multiplied by minus 1 both sides, we get the divergence between two densities f and g is greater than equal to 0. Now, similarly we could define chain rule also for continuous random variable. So, differential entropy between X 1 X 2 X 3 X n can be written using chain rule in this particular fashion and since we know conditioning cannot increase entropy, again this can be very easily proved. We have just now shown divergence is greater than equal to 0 and we know mutual information is divergence between the joint densities and the marginal's.

(Refer Slide Time: 31:09)



Chain rule for differential entropy

- Corollary:

$$\underline{\tfrac{1}{2}\log(2\pi e)^n}\,\underline{|K|}$$

$$\boxed{h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i)} \quad \sum_n \frac{1}{2}\log(2\pi e) K_{ii}$$

$$\frac{1}{2}\log \boxed{\prod_{i=1}^{n} K_{ii}}\,\underline{(2\pi e)^n}$$

  with equality iff $X_1, X_2, \ldots, X_n$ are independent.
- Proof: Follows directly from chain rule and the fact that $h(X|Y) \leq h(X)$ with equality iff X and Y are independent.
- Application: *(Hadamards inequality:)* If we let $\boxed{X \sim \mathcal{N}(0, K)}$ be a multivariate normal random variable, calculating the entropy in the above inequality gives us

$$\boxed{|K| \leq \prod_{i=1}^{n} K_{ii}} \qquad |K| \leq \prod_{i=1}^{n} K_{ii}$$

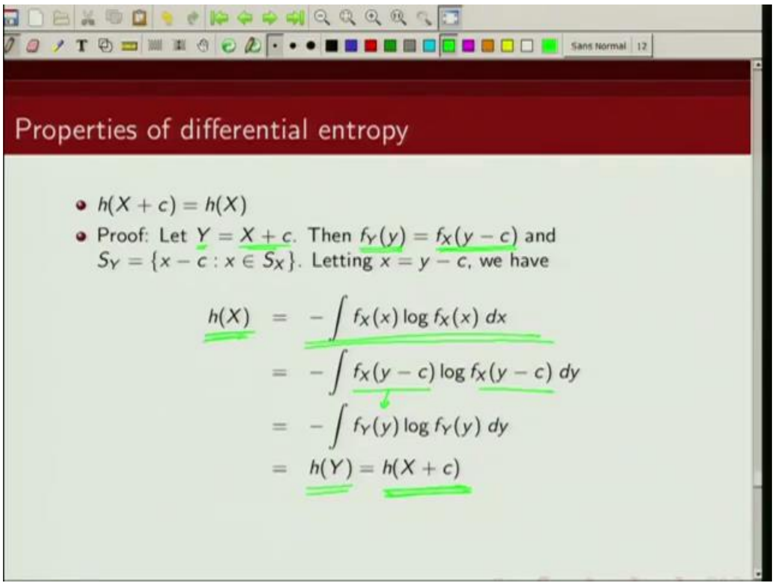
  which is Hadamards inequality

So, mutual information should be greater than equal to 0 and from there we can prove that conditioning cannot increase entropy and from the chain rule which is given here if we apply this condition that conditioning cannot increase entropy, we get this condition and of course, this joint differential entropy is equal to this integral differential entropy when these X i s are independent. So, I may have given the proof, this follows from the chain rule and the fact that conditioning cannot increase entropy only when X and Y are independent that this is equal.

Now, we can make use of this result to prove what is known as Hadamards inequality. So, if we consider a multivariate random variable with 0 mean and covariance matrix given by K, now if we compute its differential entropy, this will be given by this expression half of log 2 pi e raise power n determinant of x and this h of X I, this half of log of 2 pi e K i i. So, if we compute this differential entropy here, we will get terms of the form, this is half log 2 pi e, this K i i form summation over all n, we will get something of this form whereas, this differential entropy is the form half log of 2 pi e raise to power n determinant of this..

So, summation of log, this can be written as half log of product from i equal to 1 to n K i i and you have this 2 pi e raise to power n term here and if you compare these two forms you have half log of 2 pi e n and you have half log of two pi e n and since this is less than equal to this, we will get the condition that determinant of K is less than equal to this particular term. So, this follows from two results, one is this one and second is the differential entropy of this multivariate Gaussian random variable.

(Refer Slide Time: 35:29)



Properties of differential entropy

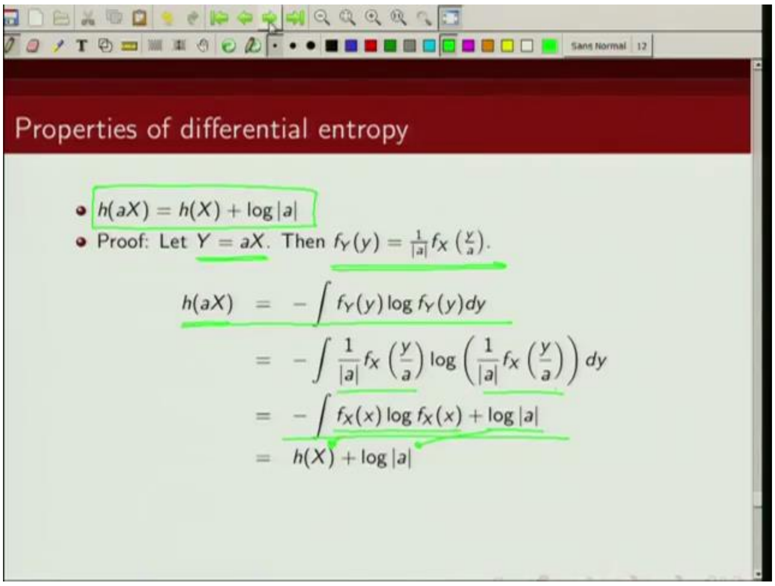
- $h(X+c) = h(X)$
- Proof: Let $Y = X + c$. Then $f_Y(y) = f_X(y-c)$ and $S_Y = \{x - c : x \in S_X\}$. Letting $x = y - c$, we have

$$\begin{aligned} h(X) &= -\int f_X(x) \log f_X(x)\, dx \\ &= -\int f_X(y-c) \log f_X(y-c)\, dy \\ &= -\int f_Y(y) \log f_Y(y)\, dy \\ &= h(Y) = h(X+c) \end{aligned}$$

We will prove some more properties of differential entropy. So, translation does not change differential entropy. So, differential entropy of X plus c is same as h of X. This is straight forward to prove and we have a new random variable Y which is X plus c and we can write the density function of y in terms of density function of x.

Now, we can write density differential entropy of X in this particular fashion and X is nothing but y minus c. So, this is given by this and this is nothing of f of y and this is equal to differential entropy of Y which is nothing but differential entropy of y is X plus c. and this is differential entropy of X plus c. So, what we have proved is differential entropy of X is same as differential entropy of X plus c. So, translation does not change differential entropy.

(Refer Slide Time: 36:47)



Properties of differential entropy

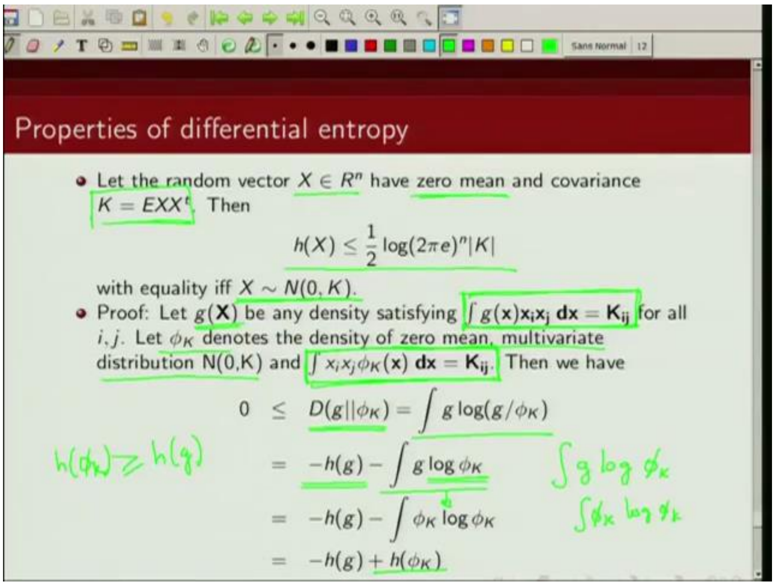
- $\boxed{h(aX) = h(X) + \log|a|}$
- Proof: Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$.

$$\begin{aligned} h(aX) &= -\int f_Y(y)\log f_Y(y)dy \\ &= -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \\ &= -\int f_X(x)\log f_X(x) + \log|a| \\ &= h(X) + \log|a| \end{aligned}$$

This is the effect of scaling. So, differential entropy of a of a times X is given by differential entropy of X plus log of absolute value of a. Here, y is given by a of X, we can write the density function of y in terms of density function of x and from the definition, differential entropy of y is given by this expression, now we plug in the densities of y in terms of densities of x which is this and this we simplify log of 1 by absolute value of a plus log of f of x of this and y by a is nothing but my x and this can be simplified into two terms, one is this particular term and second term is log of a.

So, we can see basically this will be log of 1 by a, log of 1 by a is minus log of a and minus minus that becomes plus and when you integrate it over the density function, this density function will become one. So, you got log of a integration of this density function will give you ,once it is log of a and the next term that you will get is this log of f of x y a which is nothing but f of x log of f of x. This is nothing but differential entropy of X, this is log of absolute value of a. So, the effect of scaling is as follows, h of differential entropy of a of X is differential entropy of X plus log of absolute value of a.

(Refer Slide Time: 38:57)



Properties of differential entropy

- Let the random vector $X \in R^n$ have zero mean and covariance $\boxed{K = EXX^t}$. Then

$$h(X) \leq \frac{1}{2}\log(2\pi e)^n|K|$$

with equality iff $X \sim N(0, K)$.

- Proof: Let $g(\mathbf{X})$ be any density satisfying $\boxed{\int g(\mathbf{x})\mathbf{x_i}\mathbf{x_j}\,\mathbf{dx} = \mathbf{K_{ij}}}$ for all i, j. Let ϕ_K denotes the density of zero mean, multivariate distribution N(0,K) and $\boxed{\int x_i x_j \phi_K(\mathbf{x})\,\mathbf{dx} = \mathbf{K_{ij}}}$. Then we have

$$\begin{aligned} 0 \leq\ & D(g\|\phi_K) = \int g\log(g/\phi_K) \\ =\ & -h(g) - \int g\log\phi_K \qquad \int g \log \phi_K \\ =\ & -h(g) - \int \phi_K \log\phi_K \qquad \int \phi_K \log \phi_K \\ =\ & -h(g) + h(\phi_K) \end{aligned}$$

$h(\phi_K) \geq h(g)$

Now, the next result that we are going to show is as follows. So, if you have a random vector X with zero mean and variance given by this, then its differential entropy is upper bounded by half log of 2 pi e raise power n determinant of this covariance matrix and equality happens when X is multivariate normal distributed random variable.

So, if X is the random variable with zero mean and covariance matrix K then, we are saying it is the multivariate Gaussian distributed random variable which will have the maximum differential entropy and if you recall its counterpart in case of discrete random variable, it is the uniform distributed random variable which will have the maximum entropy. Now, how do we prove it? We have been given that mean is 0 and covariance is fixed K. So, let g of X is a density satisfying this condition, this is basically the covariance. Let phi of x is zero mean multivariate distribution which has the same second order moment which is given by K. Now, let us compute the divergence between the density g and this zero mean multivariate normal distributed phi K. So, this divergence between g and phi k can be written from the definition as g log of g by phi K.

This can be written as minus differential entropy of g minus integration of g log phi K, now log phi K is of quadratic form, log phi K is of quadratic form and we have been given that x and phi K has the same covariance that is they have the same. So, integration of x i x j phi K d x is same as integration of x i x j g x d x and then this term g of log phi K will be same as phi K log phi K because this has a quadratic form log of phi x will be a quadratic form x i x j form and we know that integration of g x x i x j, this is same as integration of x i x j phi K. So, this is equal to this and that is why you are able to write it like this.

Now, this is nothing but plus of differential entropy of this zero mean multivariate distribution and this is greater than equal to 0 that means, h of phi k is greater than equal to h of g. So, in other words it is the multivariate normal distribution random variable which will have the maximum differential entropy. So, with this we will conclude our discussion on differential entropy. In the next class, we will talk about Gaussian channel.

Thank you