

Fundamentals of MIMO Wireless Communication
Prof. Suvra Sekhar Das
Department of Electronics and Communication Engineering
Indian Institution of Technology, Kharagpur

Lecture - 33
Fundamentals of information Theory – 3

Welcome to the lectures in fundamental of MIMO Wireless Communications. We are currently looking at the basics of information theory or the definitions which will be very, very use full when analyzing the capacity of MIMO channel.

Till now we have covered discrete random variables and we have given the expression of entropy relations, we have also started the restriction of continuous random variables, because as we said it is important because we will finally dealing with wave form channels and we have started with the definition of differential entropy. So, we move forward in this lecture in defining some more description or some more relationships of continuous random variables which will finally, end up in helping us understanding or realizing the expression of capacity which is used in the MIMO channel conditions.

(Refer Slide Time: 01:15)

Joint Differential Entropy
 Let x_1, \dots, x_n densities $f(x_1, \dots, x_n)$

$$h(x_1, \dots, x_n) = - \int f(x_1, \dots, x_n) \log f(x_1, \dots, x_n) dx_1 \dots dx_n$$

Def Conditional Differential Entropy X, Y $f(x, y)$

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

$$f(x|y) = \frac{f(x, y)}{f(y)}$$

$$h(X|Y) = h(X, Y) - h(Y)$$

So, we continue from whatever we have done; that means, the differential entropy and we move forward to describe the joint entropy, the joint differential entropy. So, in this case it is similar to the discrete random variable and for the set. So, if we have the set of random variables x_1, x_2 up to x_n with densities given by $f x_1, x_2$ up to x_n ; that means,

this is joint density then we could define $h(x_1, x_2, \dots, x_n)$ which is the joint entropy as in the same manner minus $\int_{x_1, x_2, \dots, x_n} f(x_1, x_2, \dots, x_n) \log f(x_1, x_2, \dots, x_n) dx_1, dx_2, \dots, dx_n$. So, this density goes on top up, to x_n and of course, dx_1, dx_n .

And then we move on to the definition of we have this as definition of conditional differential entropy. So, the conditional differential entropy is also defined for the pair of random variables x and y with joint density of $f(x, y)$ as $h(x|y)$ or $h(x \text{ given } y)$ as $-\int f(x, y) \log f(x|y) dx dy$ and this is particular conditional density you could write it as $f(x|y)$ upon $f(y)$. So, using this you could expand this and the relationship you would get finally, this same as that of the earlier case as joint entropy of x and y minus of $h(y)$.

(Refer Slide Time: 03:45)

Entropy of a multivariate normal distribution

Let x_1, x_2, \dots, x_n . μ , covariance Matrix K

$N(\mu, K)$ $f(x_1, \dots, x_n)$

probability density x_1, x_2, \dots, x_n

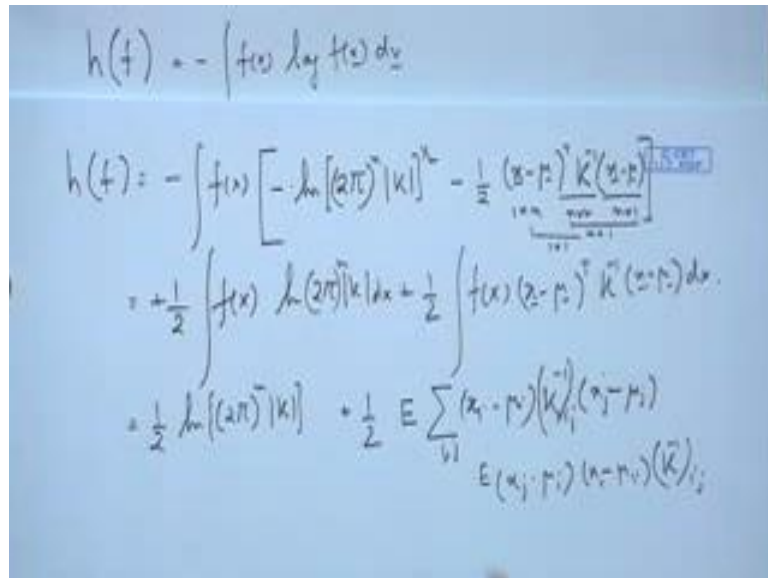
$$f(x) = \frac{1}{(\sqrt{2\pi})^n |K|^{0.5}} e^{-\frac{1}{2}(x-\mu)^T K^{-1} (x-\mu)}$$

So, these are some of the relationships which will be using as we move down further. So, moving on what we now need to look at is the multivariate the entropy. This is very, very important that is the entropy of a multivariate normal distribution. So, we are taking step by step forward.

So, that we finally, end up an expression which is very use full. So, for this we let x_1, x_2, \dots, x_n have a multivariate normal distribution, with a mean of μ I mean a vector μ and a covariance matrix, which is written K and we would use the notation $N(\mu, K)$ in this particular definition to indicate normal distribution with a mean and a variance. So, this you could change the variables, but they would look similar in the way you write it,

and we need to find h of $x_1 \dots x_n$ that is the multivariate distribution which is normal in this case. So, the probability distribution the probability density of the joint distribution, $x_1 \dots x_n$ is usually given by f of x . I write the underscore to indicate a vector you could find this is a bold index is given as 1 by 2π raised to the power of n this indicates the determinant of k to the half of square root; that means, e to the part of minus half x minus μ vector of course, transpose k inwards times x minus μ .

(Refer Slide Time: 05:56)



$$\begin{aligned}
 h(f) &= - \int f(x) \log f(x) dx \\
 h(f) &= - \int f(x) \left[-\ln[(2\pi)^n |k|] - \frac{1}{2} (x-\mu)^T k^{-1} (x-\mu) \right] dx \\
 &= \frac{1}{2} \int f(x) \ln[(2\pi)^n |k|] dx + \frac{1}{2} \int f(x) (x-\mu)^T k^{-1} (x-\mu) dx \\
 &= \frac{1}{2} \ln[(2\pi)^n |k|] + \frac{1}{2} E \sum_{i=1}^n (x_i - \mu_i) \left(\frac{1}{k_{ii}} \right) (x_i - \mu_i) + \dots
 \end{aligned}$$

So, this is the expression of the density and we have to find this following the first principle and we are interested in the expression which is valuable for us. So, as we move a head we have to write. So, far continuous random variables we said it is dependent only on the density.

So, we have this notation h of f which I have described in the previous lecture. So, this is written as minus f of x is the vector times $\log f$ of x $d x$ right this is what we know. So, using that we move forward and write what is h of f as minus $f x$. So, if we look at this expression this absolute expression, that we have here and take the log of it and what we are left with is two parts; one is this part the other is this part. So, we could write it as here we can take the natural logarithm. So, either log based to a log base n we will take $1/n$ for ease and then you can do for log based to similar result. So, when I take the $1/n$ of these we get a minus sign $1/n$ 2π to the power of n times determinant of k raised to the power of half.

So, this is what we have is the first expression and the second expression since it is $1/n$ of e to the power of something is the expression itself. So, we have $\frac{1}{2} x^T (x - \mu) K (x - \mu)$. So, this is the expression you have. So, for which this $\frac{1}{2}$ can come outside here this minus and minus adds together. So, we have plus half integration of $f(x)$ times this constant $(1/n^2 \pi)^{n/2} \det(K)^{-1/2}$ place to the 2π to the power of n and here again the minus and the minus cancels out and we have plus and that is $\int f(x) dx$.

Now, if we look at this term this is a row vector and this is the matrix this is the column vector. So, what you have over here is $1 \times n$ let us say n and this is a $n \times 1$ and this is a $n \times 1$. So, this would lead to a $n \times 1$ finally, a product over here would be one cross one. So, what we have is one single value end of it here. So, we still write it down $x^T (x - \mu)^T K^{-1} (x - \mu)$ of course.

So, we have this expression this particular integral this is a constant is this is a constant over here. So, the rest of it integrates to 1. So, what you have is half $(1/n^2 \pi)^{n/2} \det(K)^{-1/2}$ to the part of n times determinant of K this is one of the terms the second term which exists is also half and instead if we if we look at the this integral $\int f(x) dx$ it is basically the expectation of this terms. So, we could write the expectation of and if you would see this product you could write it as expectation over I and J 2 variables $x_I - \mu_I$ times K^{-1} times $x_J - \mu_J$ right this again I repeat this integral $\int f(x) dx$ I am replacing by this e this e can go inside and you will be having expectation of this times K^{-1} inverse; that means, the inverse of K times this is element times this and then we could bring this over here.

So, we will have an $e^{-(x - \mu)^T K^{-1} (x - \mu)}$ times $x_I - \mu_I$ times K^{-1} of I, J right. So, when we have this and the summation over I, J what this term would finally, lead to is an expression which is like half this this expression what you would see is kind of K^{-1} is the covariance matrix.

(Refer Slide Time: 10:00)

$$\begin{aligned}
 &= +\frac{1}{2} \int f(x) \ln(2\pi) |K| dx + \frac{1}{2} \int f(x) (x-\mu)^T K^{-1} (x-\mu) dx \\
 &= \frac{1}{2} \ln((2\pi)^n |K|) + \frac{1}{2} E \sum_{i,j} (x_i - \mu_i) (K^{-1})_{ij} (x_j - \mu_j) \\
 &\quad \sum_{i,j} E(x_i - \mu_i)(x_j - \mu_j) (K^{-1})_{ij} \\
 &\quad \sum_{i,j} (K^{-1})_{ij} \\
 &= \frac{1}{2} \ln((2\pi)^n |K|) \quad \frac{1}{2} \sum_{i,j} \frac{1}{e} \\
 &= \frac{1}{2} \ln((2\pi e)^n |K|) \quad \text{and } \frac{1}{2} \sum_{i,j} (2\pi e)^n |K| \ln
 \end{aligned}$$

So, we left with the expression which is summing over k k inwards whole of these values. So, basically k k inverse would be an identity matrix and then you are basically adding up all of the terms. So, you are left with the half of n and here what you have is half $\ln(2\pi)^n$ times determinant of k . So, this half n that we have over here, for e s what we could do is we could have things in a nice way and here we could instead put this as $\ln e$. See if you have $\ln e$ and then this is equivalent to one and then again you could modify this term to say that it is half e to the power of n half $\ln e$ to the power of n .

So, we have this term and this term added together which would lead to half \ln of this plus \ln of this so; that means, \ln comes inside as a product $2\pi e$ to the part of n times determinant of k . So, this is the entropy if it is \ln it is a nats otherwise if it is in log base 2, it is half log base 2, $2\pi e$ are to the power of n times determinant of k is in bits right. So, this is the expression that you have for h of f and this is what we wanted to find and this will turn out to be very, very use full this particular expression is going to be turn out to very, very use full for all our expressions, that we finally, lead up to. So, at this point I would like to point out one more thing let is this what we have done is for real valued what we also need to do for complex valued because we will finally, dealing with complex valued. So, I will just give you the result what it appears for complex valued that is what we need for MIMO channel capacity.

(Refer Slide Time: 12:13)

Complex multivariate Gaussian Distribution
 $- (x-\mu)^T K^{-1} (x-\mu)$

$$p_f(x) = \frac{1}{\pi^n |k|} e^{- (x-\mu)^T K^{-1} (x-\mu)}$$

$$h(f) = \log((\pi e)^n |k|)$$

Relative Entropy: $D(f||g) = \int f \log \frac{f}{g}$

Mutual Information: $\int f(x,y) \log \frac{f(x,y)}{f(x)g(y)} dx dy$

$$I(x,y) = h(x) - h(x|y) = h(y) - h(y|x)$$

So, when we have the multivariate Gaussian distribution of complex multivariate Gaussian distribution in this case you could write distribution as p of or lets write in the same notation p of x f of x is equal to one upon pi to the power of n determinant of covariance matrix e to the power of minus x minus mu times k inverse is of course, the on top times x minus mu. So, this is the expression. So, if you use this expression in calculating h of f h of f would turn out to be log of pi e to the part of n times determinant of k. So, this is the other expression which would have valued to us. So, we have seen 2 important expressions which are fundamental in the expression of capacity which will finally, look at.

So, let us move on further from this point. So, this is one of the expressions that we have and the other expression which we had pointed out is here. So, these are the 2 expressions that we have which will be using in all our calculations after this further as we move on, we would also like to see the relative entropy as we have seen relative entropy for the discrete random variable case. So, here also it is the distance it is defined as d of f to g, where f and g are two distributions as integral f log f upon g which is similar to the expression that we had before. So, we will use this and we will define mutual information for this deferential for this continuous random variable as defined as integral f x y log f x y upon the f x times f y d x d y.

So, again using this end result that we get is $I(x, y)$ similar to that we have got in the discrete random variable case $h(x) - h(x|y)$ or we could also write it is $h(y) - h(y|x)$. So, the expression looks similar there is not much difference in the expression where in text books you could find all of these. So, these derivations or these expressions that, we have done now these are easily available in books and information theory one of the easy books to read very easy to read is the one on elements of information theory by Thomas m cover. There are many, many other books as I always said that you can choose a book according to your own preference and this is the particular 1 which gives these expressions in a straight forward manner which could be a it use to you in this particular course again I am iterating that we are not getting in to details of information theory to results we just reveal them, because they are. So, fundamental without which we will not be able to build upon the expression of capacity.

So, we are doing this because these are the elements which are going to support the expression of capacity finally. So, we are mainly interested in using the results of our information theory of for which we are undertaking this activity.

(Refer Slide Time: 16:18)

$D(f||g) \geq 0$
 with equality iff $f=g$
 let S be the support set of f .
 $D(f||g) = \int_S f \log \frac{f}{g}$
 $\leq \int_S f \log f$
 $= h(f)$
 $= h(g)$
 $= 0$

$I(x,y) \geq 0$
 $h(x) \geq h(x|y)$
 $h(y) \geq h(y|x)$

So, moving on further what we have is that d ; that means, the relative differential relative entropy d of f to g is greater than or equal to 0, we have seen this in the case of discrete random variables and this is with equality of course, with equality if and only if f is equal

to g almost everywhere in that set. So, let s be the support set we have defined the support set support set of f , we have defined support set that have it is the set where the probability density is greater than 0. So, that is the support set of f and then we have minus d again we do it in the same way minus d of f g we begin with a minus we could do it other way also within the support set $f \log$ of g of f . Now you would remember from n since in equality these are concave function. So, you could write this as less than \log of f g of f . So, what we have done is this is the function and integral over f d f is basically the expectation operation.

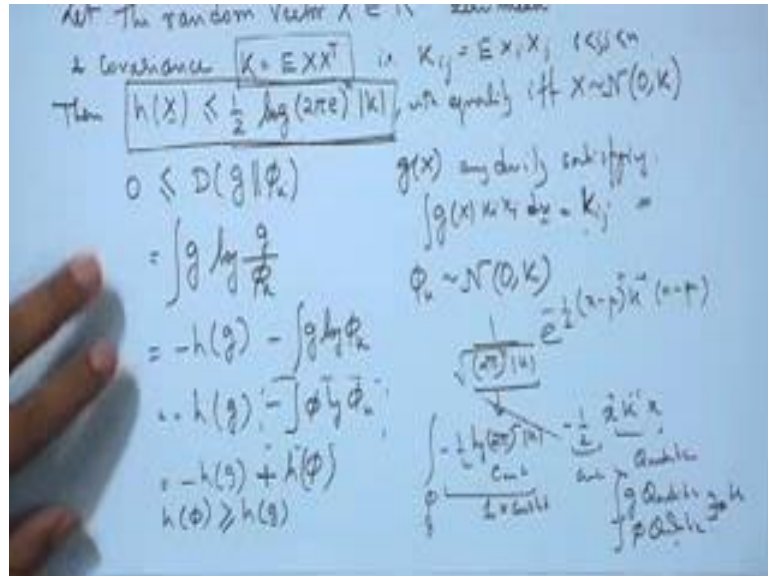
So, expectation of a function is greater than or equal to the function of the expectation of the variable for convex or concave it is the reverse. So, that is what we have over here. So, this is equal to \log of this and this cancels g and this is basically \log of this basically integrated over s which supports set. So, basically the probability density functions were integrated over the support set leads to one. So, this will be \log of one which is equal to 0. So, that critically means that minus g is less than or equal to 0 or in other words what we have is d of f from g is greater than or equal to 0. So, this also holds true for the continuous random variable case as we have seen for the different for the discrete random variable. Now this gives some important results because we have defined mutual information in terms of relative entropy you could also start with this definition and you could prove many things as well. So, d is greater than or equal to 0. Therefore, we have this $I(x, y)$ is also greater than or equal to 0.

So, basically if $I(x, y)$ bring the definition of d it is greater than or equal to 0. So, which in other words mean that h of x is greater than or equal to h of x conditional y or again h of y is greater than or equal h of y conditional x . So that means, conditioning reduces entropy that is what we had seen earlier same thing holds in this particular case also, we do not have any different result now with this we would like to move on to a very, very important result at this stage.

Where what we have seen is expression of entropy and we have seen also expression of the entropy for Gaussian distribution. We have also seen an expression of entropy for multivariate Gaussian distribution for real as well as for complex case. So, what we would like to see is which distribution maximizes entropy this could be done in many, many ways one of the way is doing straight forward, where like finding the distribution is maximizes the other one that we could do is to propose a distribution and see that this

distribution whether it is the maximum or not. So, that is the second approach is the one that what we are going to follow in this particular case.

(Refer Slide Time: 20:00)



So, we say that let the random vector x which is an element of \mathbb{R}^n ; that means, a n dimensional vector and 0 mean have 0 mean and covariance given by k is equal to e of x times x transpose that is how you define the covariance matrix and that is k_{ij} is equal to e of $x_i x_j$ this is the expectation operator for i line i and j both line between one and n then if you say this then we can say that h of x h of x is be differential entropy of x is less than or equal to half log of two pi e raise to the power of n times determinant of k with equality if and only if x is distributed as a normal distribution with 0 mean and covariance matrix given by k .

So, if we look at this what we are saying is we are proposing that the maximum entropy of this random vector x is determined by this expression. Now if you would remember this expression it is the entropy of multivariate Gaussian distribution, and where k is the covariance matrix. So, what this is saying is that suppose your random vector x whose covariance matrix is given by k all this is saying is that the entropy of such a random vector whose covariance matrix is given against the maximum value when the distribution is normal and the entropy is given by that of the normal.

So, anything any other would any other random vector would have an entropy which is less than that of the normal vector whose covariance is the same as that of the vector which

is given. So, to do this again will take advantage of d . So, we will begin with that d is greater than or equal to 0. So, this is well known. So, this is where we begin with g and ϕ^k . So, where we assume that this ϕ^k in our expression is normal distribution and we assume that g of x is any density any density satisfying $\int g(x) x_i x_j dx$ is equal to k_{ij} this is the element of the covariance matrix. So, all this is saying is that this is the definition of the element of the covariance matrix for all i, j and ϕ^k is normally distributed with 0 mean and k is the covariance matrix of equated the covariance matrix and we have said the μ is equal to 0.

So, when we take. So, so basically we will be using the distance of g from ϕ^k and will be showing the important result. So, when we move forward we look at this expression this is basically $g \log g$ upon ϕ^k right. So, if you look at the numerator it is minus $g \log$ one upon g . So, or $g \log g$ is. So, basically it is minus h of g right. So, this is from the numerator and from the denominator term you have $\int g \log \phi^k$ of k . So, when we look at $g \log \phi^k$ of ϕ^k we said it is $\frac{1}{2} \pi$ to the power of n times determinant of k square root of that e to be part of minus half x minus μ . So, in this case μ is 0 transpose k inverse x minus μ . This is what we had. So, if you take \log of ϕ^k . We have seen this expression there is a constant term which goes there. So, basically \log of $\frac{1}{2} \pi$ to the power of n times determinant of k half this is one of the constant and if you take a look at the second term you again have there is a minus half depending upon the base you will have 1 or \log of $2 \log$ of e and then you have x times k inverse x transpose times x . So, basically there is a constant there is another constant and there is a quadratic term.

So, what we have said is basically this term, if you look at this term the g times the quadratic term is what you are going to encounter over here and g times the quadratic term is equal to k and we could we could also say that this could be replaced by $\int \phi^k \log \phi^k$. Because what we have described in this is that the covariance of this random vector x and that of the normal distribution are basically same. So, we are taken a normal distribution which as this. So, if I would do $\int \phi^k \log k$ what we are going to encounter is this term. So, in this term when we do the integral we do the integral. So, this would turn out to be one whether, we have a ϕ or a g . So, whether we have a $\int \phi$ or we are $\int g$ along with this term this integral terms in one times the constant this is the same constant which is due to $\log \phi$. So, we have \int

g times the quadratic term in one case here and in the other case you are going to have integral ϕ times the quadratic term.

So, these are the two things that we have. So, as we have said that g times the quadratic terms leads to k leads to the expression of k and again ϕ times the quadratic term is also equal to k because these both lead to the expression of the covariance matrix. So, since they are equal and this integral turns out to be the constant because integral of ϕ times of constant is equal to one integral $\log g$ times is constant is equal to 1. So, one we are left with these terms again which are equal by what of this covariance. So, then we could replace integral $\phi \log g \log \phi$ with a $\phi \log \phi$. So, what we have over here is minus h of ϕ and we have minus over here.

So, that would lead to minus h of g plus h of ϕ if you look at this expression. So, all we have done is replaced integral $g \log \phi$ with in integral $\phi \log \phi$ because $\log \phi$ leads to expression where there is a constant and constant times of quadratic term. So, integral of constant times g is equal to the constant and integral of the constant time the quadratic term is equal to the same constant time the quadratic term as for the normal case and they are both equal to the k component. So, I could replace g with ϕ . So, this is equal to minus h of g and this whole expression is basically h of ϕ .

So, since this is a greater less greater than or equal to 0 all, we can say is that h of ϕ is greater than or equal to h of g and h of ϕ is the expression which we already have is given by this. So, basically h of ϕ is given by this expression. So, over all we could say thus this holds true for any continuous random vector whose covariance is given by k .

So, we conclude this particular lecture at this point other very important things to remember is we have come to the point, where we have shown that Gaussian distribution or multi variant Gaussian distribution gives the maximum entropy for any random vector which has the same covariance.

Thank you.