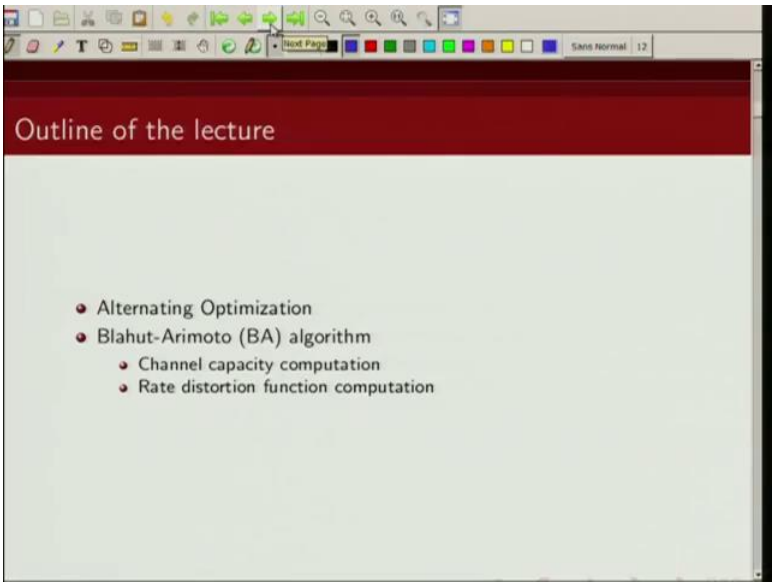
An Introduction to Information Theory
Prof. Adrish Banerjee
Department of Electronics and Communication Engineering
Indian Institute of Technology, Kanpur

Lecture – 14A
Blahut-Arimoto Algorithm

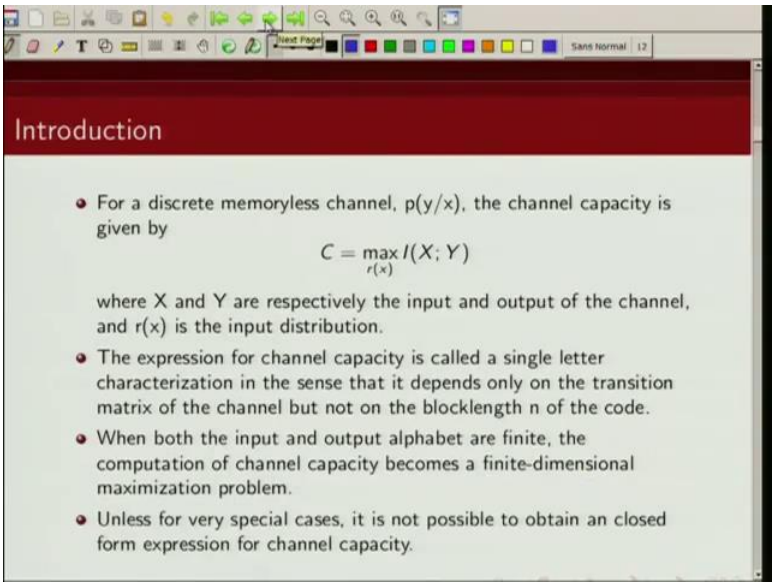
Welcome to the course on An Introduction to Information Theory. So, in this lecture, we are going to talk about an iterative algorithm to compute channel capacity and rate distortion function, and this is known as Blahut-Arimoto algorithm named after these two scientists who independently came up with this algorithm.

(Refer Slide Time: 00:54)



So, first we will talk about this algorithm in a general setting and then so will be talking about alternating optimization algorithm and then we will talk specific about this Blahut-Arimoto algorithm. In particular, we will talk about how we can use this algorithm to compute channel capacity, and how we can use this algorithm to compute rate distortion function. Now, for this lecture, we are going to use the book by Raymond Yeung on Information Theory and Network Coding, this is chapter 9 of that book.

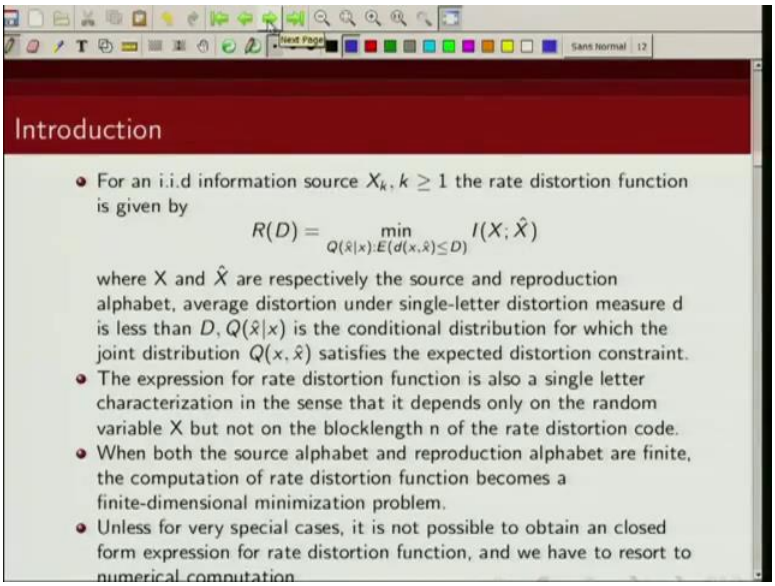
(Refer Slide Time: 01:27)



So, as we know for a discrete memoryless channel with transition matrix given by p of y given x, channel capacity is given as maximum mutual information between the input X and the output Y, and the maximization is taken over all input distribution. Here I am denoting the input distribution by r of x; X is input to the channel and Y is the output to the channel. Now, as we know this expression for channel capacity is what we call as single letter characterization. Now, we know this channel capacity depends on the transition matrix of the channel; it does not depend on the block length of the code used.

And when the input alphabet size and the output alphabet size that is finite in that case this computation of this channel capacity is nothing but a finite-dimensional maximization problem. Now, we know that except for some simple specific cases, we do not have a close form expression for channel capacity. So, to compute channel capacity in general, we will have to resort to let us say numerical computation of something like that. Now Blahut-Arimoto algorithm is an algorithm to iteratively compute the channel capacity and rate distortion function.

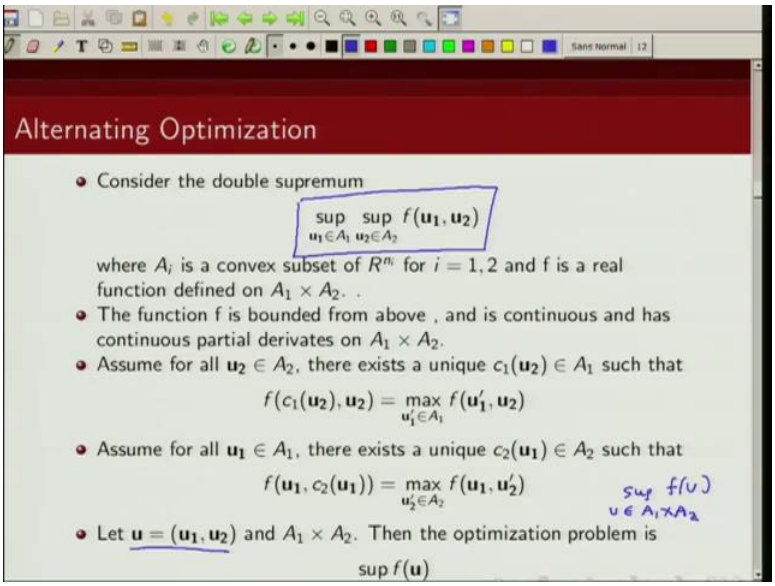
(Refer Slide Time: 03:10)



Similarly, if we look at the rate distortion function this is defined as minimum mutual information between the source alphabet and the reproduce alphabet which is denoted by X hat. And this minimization is taken over all conditional distribution of Q of x hat given x such that the expected single letter distortion is less than some quantity D. So, Q of x is the conditional distribution and of x hat given x and q of x, x hat denotes a joint distribution which satisfy the distortion constraint which is given by this here.

Now, here also this is the single letter characterization; it does not depend on the block length of the rate distortion code, it already depends on the random variable x. So, when the source alphabet and reproduce alphabet size is finite again this becomes a finite dimensional minimization problem. And again here in most of the cases, we do not have a close form expression for the rate distortion function. So, to compute rate distortion function then we will have to resort to some sort of numerical method. So, this gives a motivation to study an iterative algorithm to compute this rate distortion function and channel capacity.

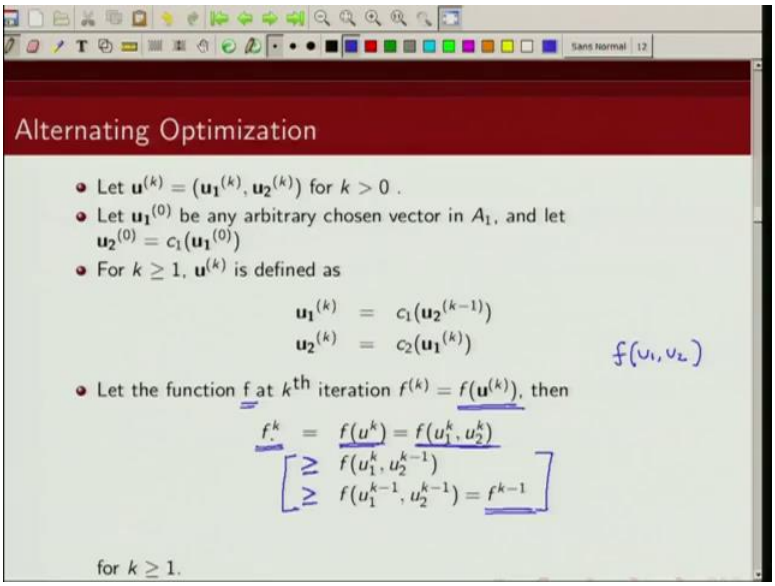
(Refer Slide Time: 04:58)



So, first let me talk about this algorithm in a general setting. So, I am going to talk about alternating optimization algorithm and then we will come to the specifics of Blahut-Arimoto algorithm. So, let us see you want to compute double supremum, so you have function f over u 1 and u 2; u 1 belongs to a convex set A 1, and u 2 belongs to convex set A 2. Now you want to compute double supremum of this function f, which is function of u 1 and u 2; and this function f is a real function defined over A 1 cross A 2. The function is also bounded from above. So, this is function is upper bounded by some quantity finite quantity, and it is continuous, also it is partial derivatives, it has continuous partial derivatives defined. So, this is the conditions on the function f.

Now, assume for all u 2 belonging to this convex set A 2, there exist mapping c 1 u 2 which belongs to this convex set A 1 such that f of c 1 u 2. And u 2 is equal to maximization of this function f of u 1 dash u 2 where u 1 dash belongs to this convex set A 1. Similarly, let us assume that for all u 1 belonging to this convex set A 1 there exist. A unique mapping c 2 u l which belongs to this convex set A 2 such that this condition is satisfied that f of u l and c 2 u 1 is equal to maximizing of this function f of u 1 and u 2 dash where u 2 dash belongs to A 2. So, then we can write this double supremum problem as we are computing a supremum of f of u where u is this and we are computing this over u which belongs to Cartesian to that of A 1 times A 2.

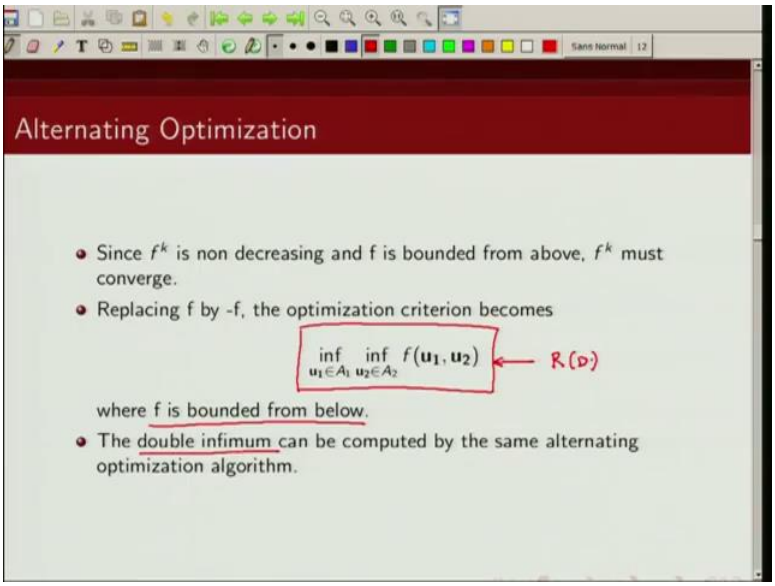
(Refer Slide Time: 07:55)



So, let us take the alternating optimization algorithm. So, let us see at some time k, u k is given by u 1 k and u 2 k. So, we start up with some initial time k is equal to 0, it start an initial value of u 1 which I am denoting by u 1 0. So, let this been arbitrary chosen vector, and it belongs to this convex set A 1. Then u 2 0 is given by c 1 u 1 0, and this belongs to this convex set A 2. In general, for k greater than equal to 1, we can define this u k by u 1 k and u 2 k where u 1 k is nothing but c 1 of u 2 k minus 1 and u 2 k is c 2 u 1 of a time k.

Now, let this function f, which is basically this at kth iteration, so we denoting this value by this is nothing but function f at kth iteration is f of u k, which is f of u 1 k and u 2 k. Now, this is greater than equal to f of u 1 k and u 2 k minus 1, which in turn is greater than f of u 1 k minus 1 and u 2 k minus 1. Now, these two relations follow from these two properties So, we have this property and this property. So, for all u 2 belonging to a 2 there exist a unique c 1 u 2 will belongs to this convex set A 1 such that f of c 1 u 2 u 2 is nothing but maximizing this function f of u 1 dash u 2 when s u 1 dash belongs to A 1. And similarly there exist for all u 1 belonging to A 1 there exist a unique c 2 u l such that this relation holds. So, because of these two properties we know that these two hold. Now so what we have shown then is f of k, f as a function of k as A is basically a non decreasing function. So, f is a non decreasing function.

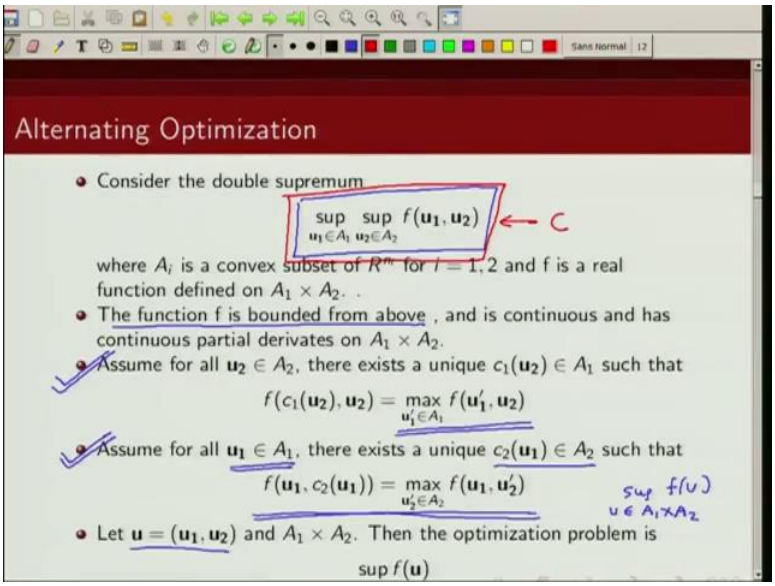
(Refer Slide Time: 11:05)



So, f of k is a non decreasing function, and remember it is bounded from above that was one of the property of f of k. If you go back, this function f of f is bounded from above. So, f of k is non-increasing, but it is bounded from above. So, what does it mean it means that f of k must converge must have a limit as k goes to very high. So, f of k converges to some finite quantity.

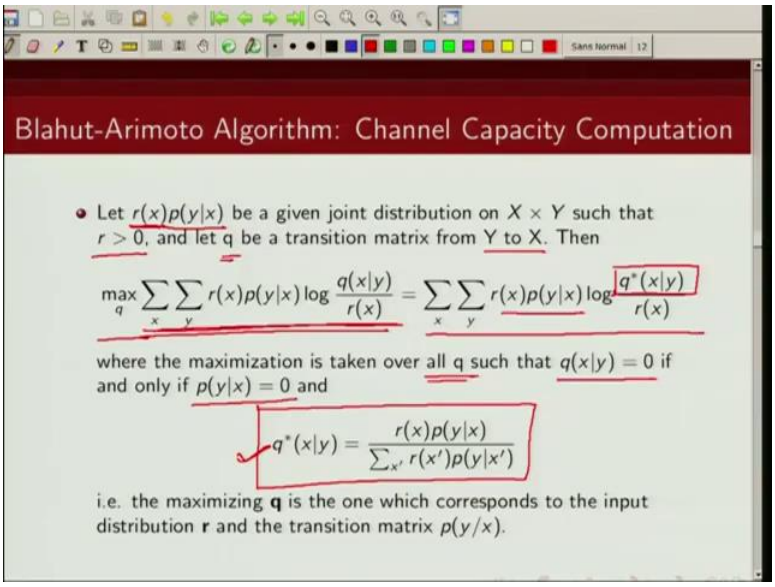
Now, if you replace this function f by minus f then in this problem if we replace this f by minus f what we get is a double infimum of this function f, where we have similar conditions that f has to be continuous partial derivatives has to be continuous. f is bounded from below then this double infimum can be computed also using this alternating optimization algorithm. Now, this form we are going to use for computing the rate distortion function

(Refer Slide Time: 12:34)



Whereas, this form we are going to use to compute channel capacity.

(Refer Slide Time: 12:42)

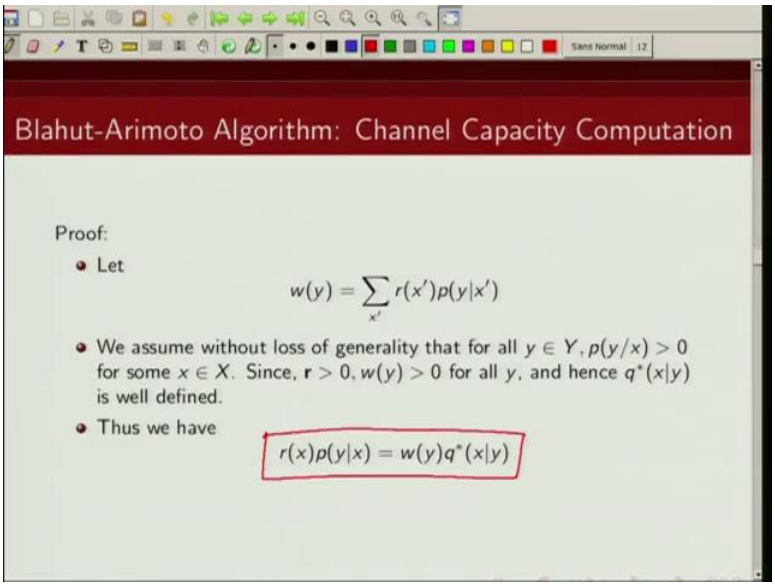


So, let us first prove a lemma that we are going to use and then we will state our Blahut-Arimoto algorithm for computation of channel capacity. So, r of x is my input distribution and p of y given x is my transition probability. Now, let r of x times p of y given x be the joint distribution on X and Y such that r of x is greater than intersects r of x is strictly positive. And let q be the transition matrix from Y to X. Then what this

lemma says is as follows. r of x multiplied by p of y given x log of q of x by y divide by r of x.

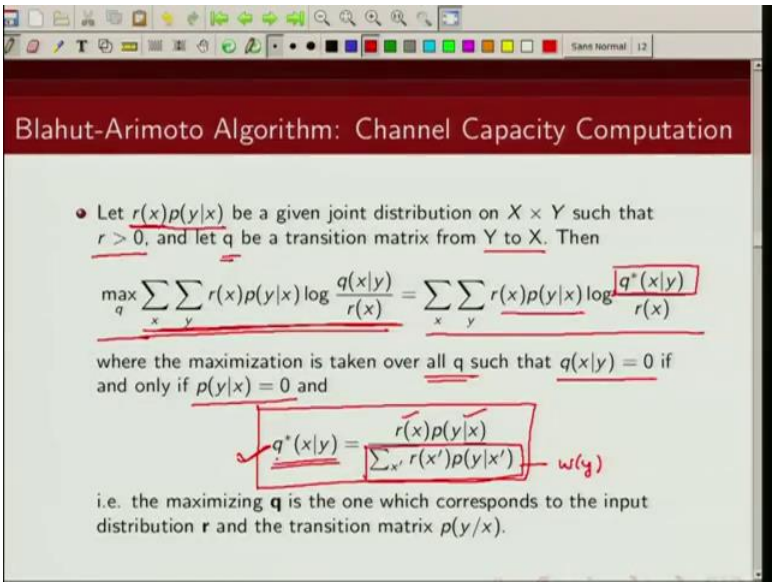
If you sum it over all x and y, and you maximize over all q then this is nothing but summation over all x and y of r x p y of x log of q star x given y divide by r x, where this q star of x given y is given by this expression. And this maximization is over all q such that q of x given y is 0, if and only if p of y given x is 0. So, what we have seen is if you try to maximize this over all q then this is basically given by this expression where expression of q star is given here. So, we are going to use the divergence inequality to prove this result.

(Refer Slide Time: 14:56)



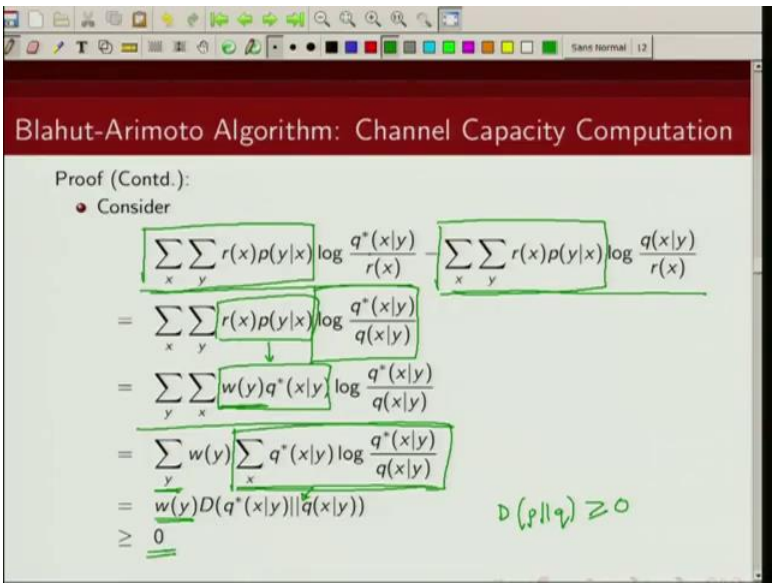
So, let us see how we are going to prove this. Let w y is given by this expression r of x prime p of y given x prime summation over all x prime. So, this is nothing but this one, this term let us denote this by w y. Now, without loss of generality, we will assume that for all Y p of y given x is greater than equal to 0 for some x and since we will consider a strictly positive r, so r greater than 0. So, then w of y will be greater than 0 for all y. And if w of y is greater than 0 you can go back and see if this is greater than 0 and these are greater than 0 then this will be also greater than 0.

(Refer Slide Time: 16:17)



Now, from this, this is my w y. So, I can write w y times q star of x given y to be equal to r of x p of y given x.

(Refer Slide Time: 16:39)

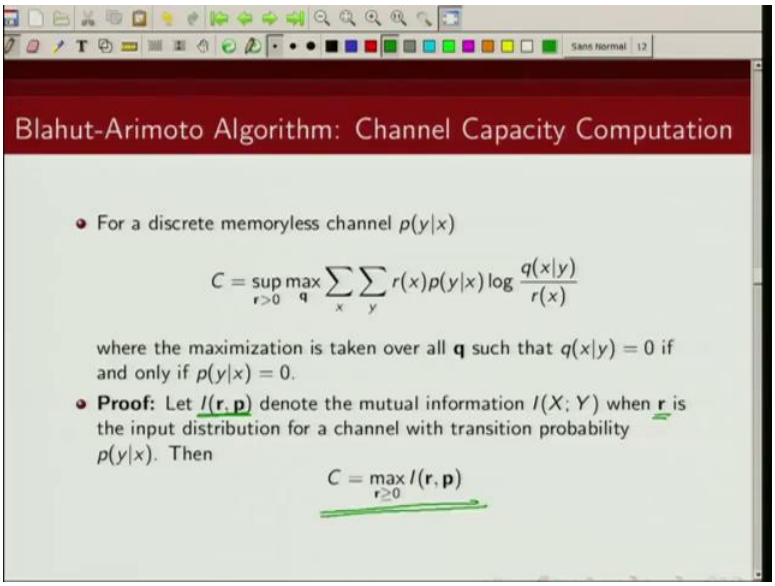


So, this is what I am writing here. Next, you just recall we want to show that if we maximize this over all q, this is given by this expression where q star is this. So, the way we are going to show that this is a maximum value of this function maximize over q. We are going to prove this by showing that this minus value of this function for some other q that is always greater than equal to 0. If we can show that then we have proved that q star

x given y is the 1 that maximizes this function. So, what we are going to show is this function evaluated at q star minus this function you have evaluated at q. Now, so if we take this is a common term, if we take this out this is common term here. If we take this out, we get log of q star x given y divided by r of x minus log of q of x given y divide by r of x. So, this can be written as log of q star x given y by q of x given y. Now we just now showed that this term is nothing but w y times q star. So, we plug in that value here, we get this expression, now this term does not depend on x. So, let us bring it out.

So, we have this term summation over x and this term summation over y. Now you can see this particular term can be written as divergence of q star and q. Now, we know that divergence between two distribution p and q is greater than is equal to 0, and since w y is also greater than 0. So, this whole thing will be greater than equal to 0. So, then what we have shown is this function evaluated at q star that has the maximum value. So, we have proved this result, this lemma which says if you try to maximize this over all q, and then this is a maximum value where q star is given by this expression.

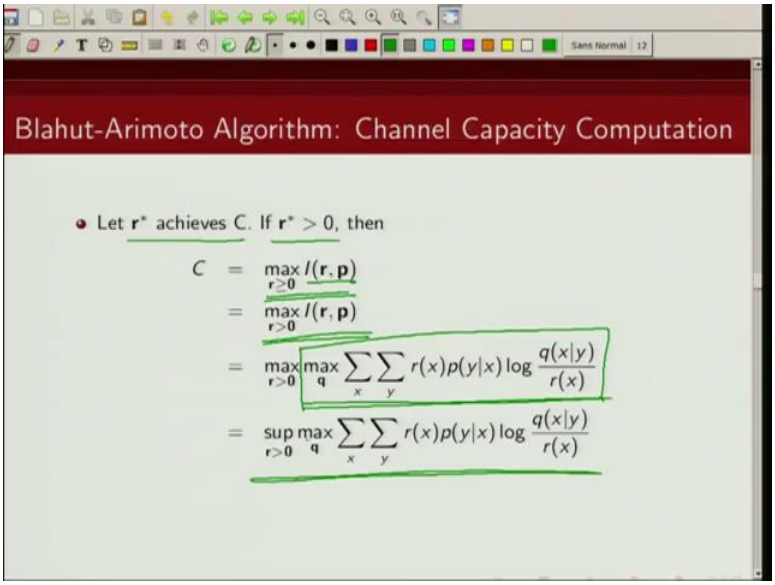
(Refer Slide Time: 20:00)



We state that for a discrete memoryless channel the capacity is given by this expression, where maximization is taken over all q such that q of x given y is 0 if and only if p of y given x is 0. So, let us prove this. So, we will first consider a strictly positive distribution of r let mutual information between x and y, let us denote by I of r p where r is input

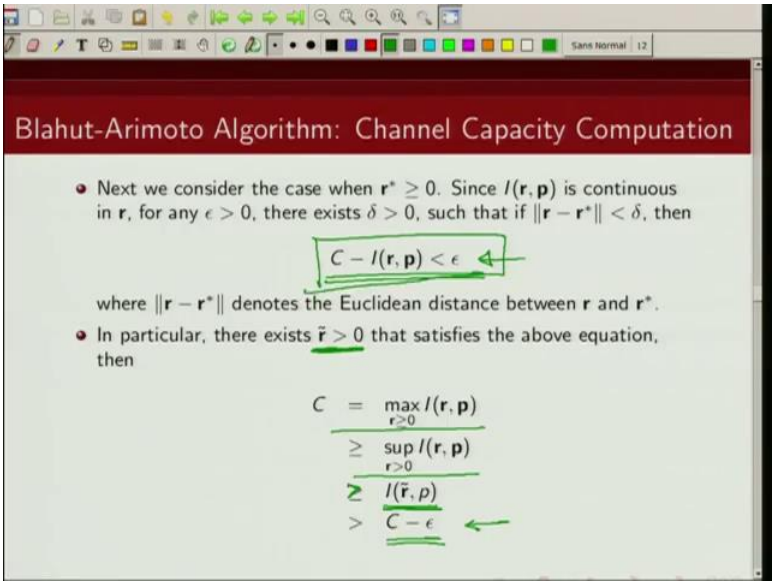
distribution and p is my channel transition probability. Now, capacity can be given as maximize mutual information and maximization over all input distribution r.

(Refer Slide Time: 20:54)



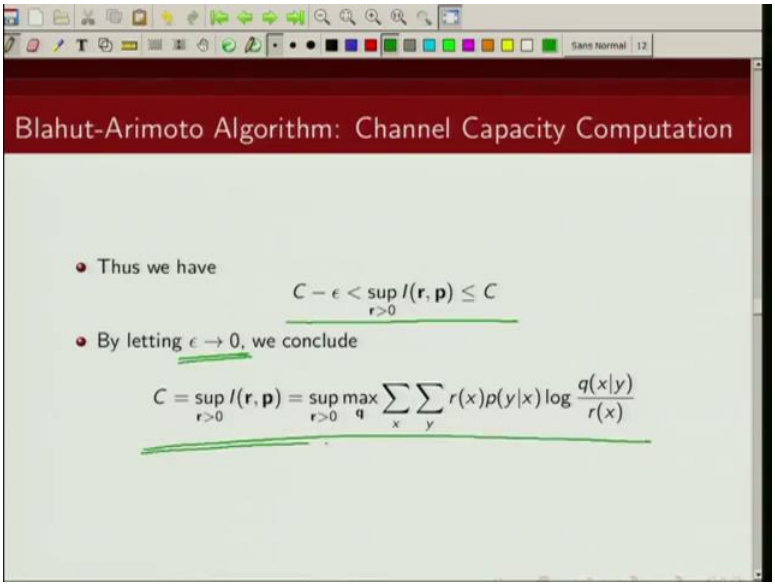
Now, let r star achieves capacity. And if r star is greater than 0 then we can write this as maximizing mutual information to r greater than equal to 0, this becomes maximizing mutual information when r is greater than 0. And from the definition of mutual information we can write this as in this particular way and this is nothing but supremum r greater than 0 maximizing over q of this function.

(Refer Slide Time: 21:42)



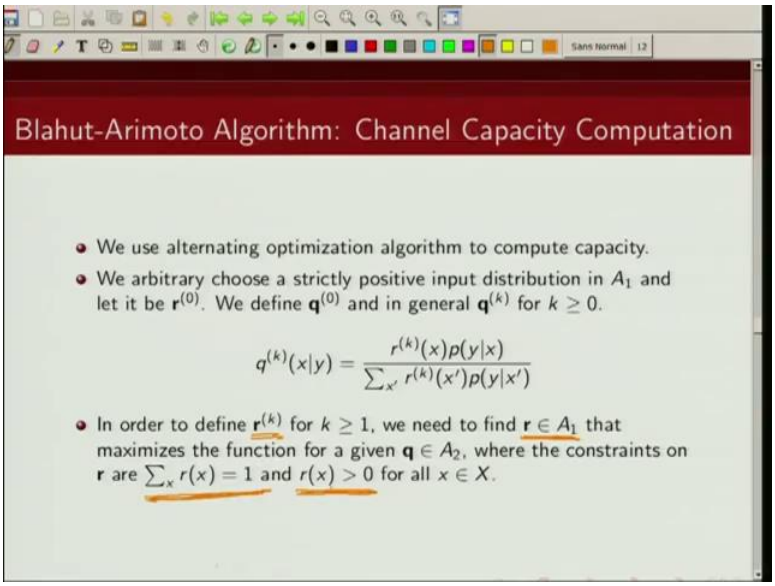
Now, in case r star is not strictly positive then since mutual information is continuous in r, so for any epsilon greater than 0 there exist a delta such that the Euclidean distance between r and r star is less than delta then this relation over the c capacity minus mutual information is less than epsilon. So, in particular, there exist an input distribution r tilde greater than 0 such that this relation holds. So, capacity is given by maximizing mutual information over all input distribution r this can be written as this is greater than equal to 2 supremum of r p a supreme over all r greater than equal to 0. And since r tilde r tilde this, this, this can be written as greater than is equal to mutual information between some distribution r tilde p and some distribution r tilde p and since r tilde satisfies this relation a satisfy this relation satisfy this condition. So, we can write that this mutual information between r tilde p will be C minus epsilon this last things follows from the fact that there exist in r tilde which satisfies this equation.

(Refer Slide Time: 23:24)



So, we have this, now if we let go epsilon to 0, we get the expression of capacity as this.

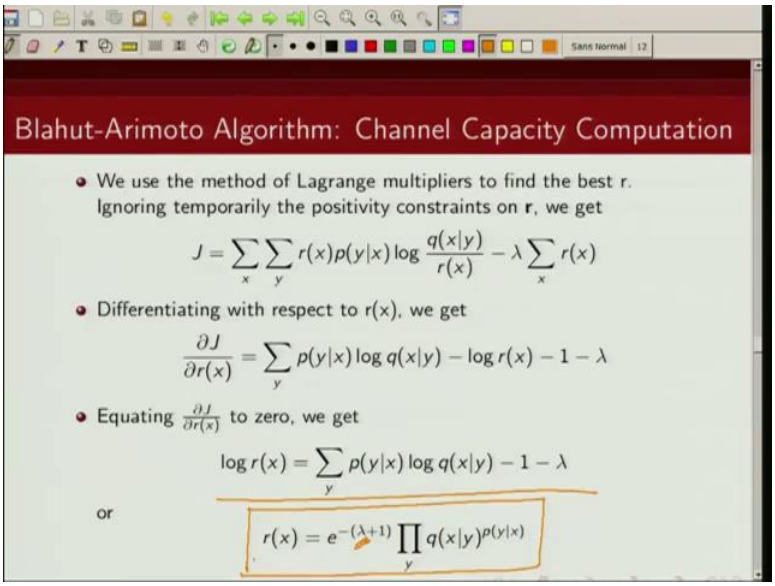
(Refer Slide Time: 23:36)



Now, let us see how we are going to use alternating optimization algorithm to compute the channel capacity. So, first we are going to choose a strictly positive input distribution which I am denoting by r 0. Of course, this belongs to this convex set a 1. And we define q of 0 or in general q of k by this relation, why, if you recall we have proved the lemma earlier and this lemma is this, what does this lemma says the maximum value of this function maximizing over all q is given by this where q is given by this expression. So, given an r, we should find q in this particular fashion because this will maximize my - this function over all q. So, this is how r and a, they are related, so that is why we start off with some initial arbitrary input distribution which is strictly positive r 0. And then we can we have to find q 0 for that we are going to use this.

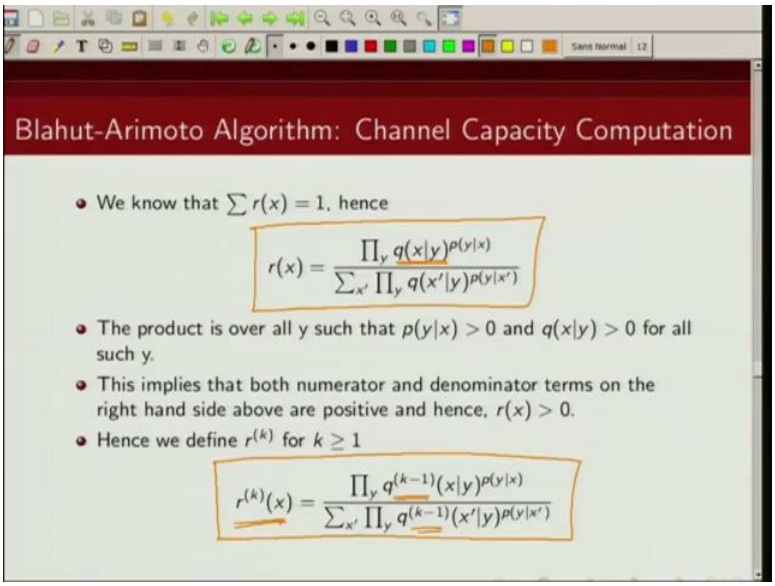
Now once you know q 0 you need to find r 1 and how do we find r 1 or in general for any k greater than equal to 1 we need to find r at time k. So, we need to find an r which belongs to this convex set a 1 that maximizes the functions for a given q. Now, remember in addition to maximizing this function we also have some constraints on r and what are these constraints first one is sum of probabilities should be 1, and the probability is are basically greater than greater than 0. So, these are the two constraints we have. So, we want to find the optimal value of r given these two constraints. So, how do we proceed?

(Refer Slide Time: 26:20)



So, we will take help of this method of Lagrange multiplier. So, we will define this, this is my objective function, this is my constraint related to the fact the summation of this r of x is 1. Now, at the moment, I am ignoring these positivity constraints that because later on we will see when we compute the value of r of x that are already taken care of. So, this is a Lagrangian. Now, the next step is to find up to value I differentiate it with respective to r of x. So, when I differentiate with respect to r of x I get this summation over y p of y given x log of q given y minus log of r of x minus 1 minus lambda. Now, when we equate this to 0 a unique collect terms what we get is log of r x is equal to summation over y q of y given x log of q x given y minus 1 minus lambda or in other words I can write r of x. In this particular way, now we need to find lambda right now we know that summation of r x of x over all x that is equal to 1.

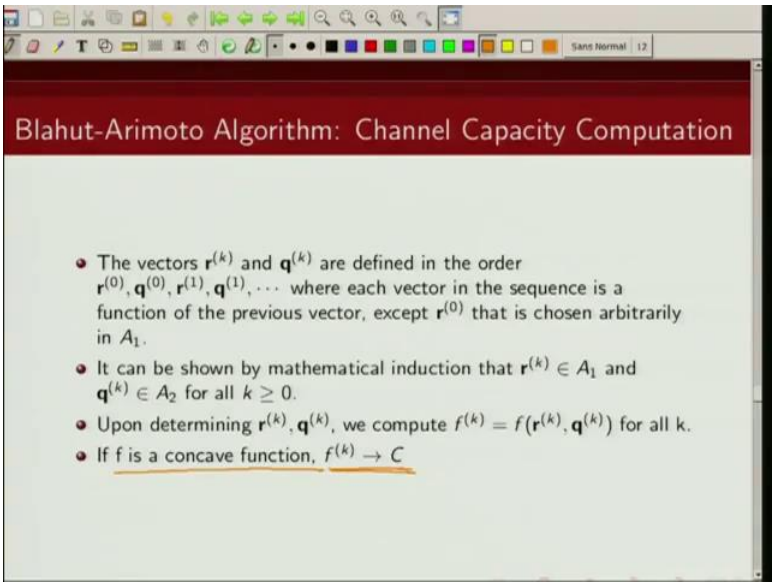
(Refer Slide Time: 28:10)



So, if we put in that constraint, we get the value of r of x to be this. Now, note that here we are taking product over all y here we are taking product over all y we have this term which is greater than 0 we have this term which is greater than zero. So, r of x is going to be greater than 0. So, earlier we have not taken this positivity constraints, but you can see that that r of x which comes out is constant is greater than 0, so that is implicitly taken care of. Now, So, this is the optimal value of r of x in terms of q of x.

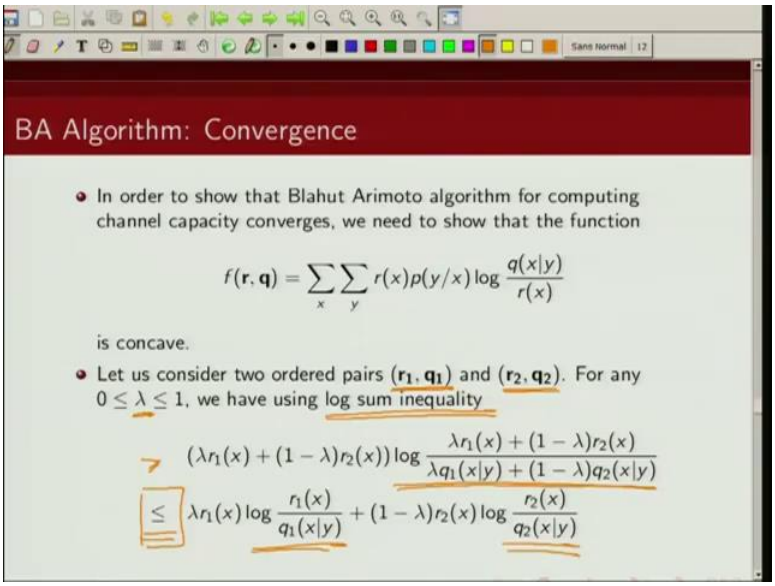
Earlier, we have computed the optimal value of q of x in terms of r of x. So, at kth iteration basically I can write r of k x as product over all y q for k minus 1 time x given y raise to power p of y given x to address summation of this. So, note now we started with an arbitrary strictly positive distribution of r, r 0 we plug that in to get q of 0. Once, we get q of 0, it plugs the value of q of 0 here to that r of 1. Now, once we get r of 1 we are going to make use of once we know r of 1 we are going to make use of this expression to get q of 1. Now, once we get q of 1 we will make use of this to get r of 2, so that is how we are that is how we are proceeding.

(Refer Slide Time: 30:27)



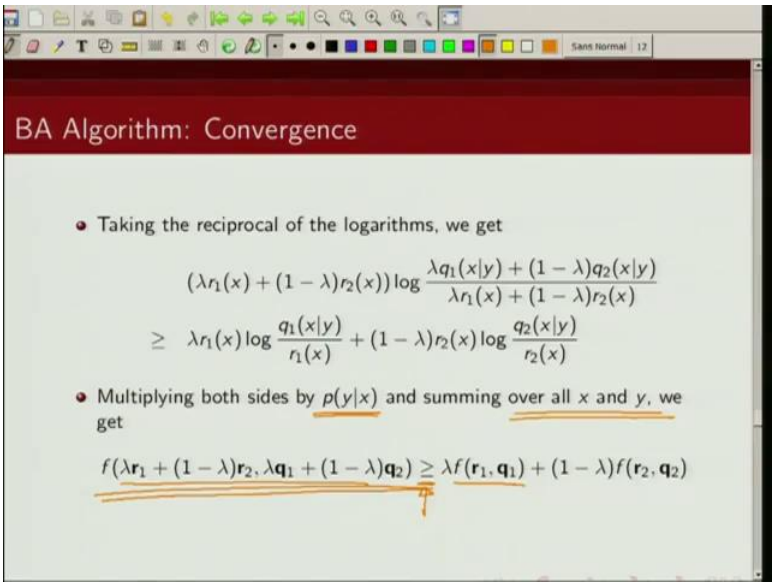
So, this you can see this is an iterative way we are, so these vectors r k and q k are defined in order we first start in arbitrary r of 0 which belongs to this convex set a 1 then we use r 0 to compute q 0 then we use q 0 compute r 1 and so on. So, you can see that each vector in this sequence is a function of previous vector except r of 0 initial values. Now, I am not showing you the proof, but it can be shown that r of k belongs to this convex set a 1 and q of k belongs to this convex set a 2 this can be proved using mathematical induction. Now, once we define and once we determine r of k and q of k we evaluate the function f at kth iteration, which is given by this. Now, the next obvious question is this function guaranteed to converge and when we will converge. So, when f is the concave function this function will converge to expression for capacity.

(Refer Slide Time: 32:00)



So, next we are try we will show that this expression for channel capacity this function is a concave function. So, to show that this algorithm converges to channel capacity we are going to show that this function that we evaluated this is the concave function of r and q. So, let us consider two ordered pair r 1, q 1 and r 2, q 2. And let lambda be and there will be between 0 and 1. Now, using log sum inequality, we can write lambda times r 1 x plus 1 minus lambda times r 2 x log of lambda times r 1 x plus 1 minus lambda times r 2 x divided by lambda times q 1 x given y plus 1 minus lambda times q 2 x given y. This is less that equal to lambda times r 1 x log of r 1 x by q 1 x given y plus 1 minus lambda times r 2 x log of r 2 x divided by q 2 x given y. So, this follows from log sum inequality. Now we are going to take the reciprocal, so reciprocal of this log. So, what you will notice is, so what I did was I just took a reciprocal of these log terms. If you take reciprocal of this log time is less than equal to term will become greater than equal to and that is what happened here.

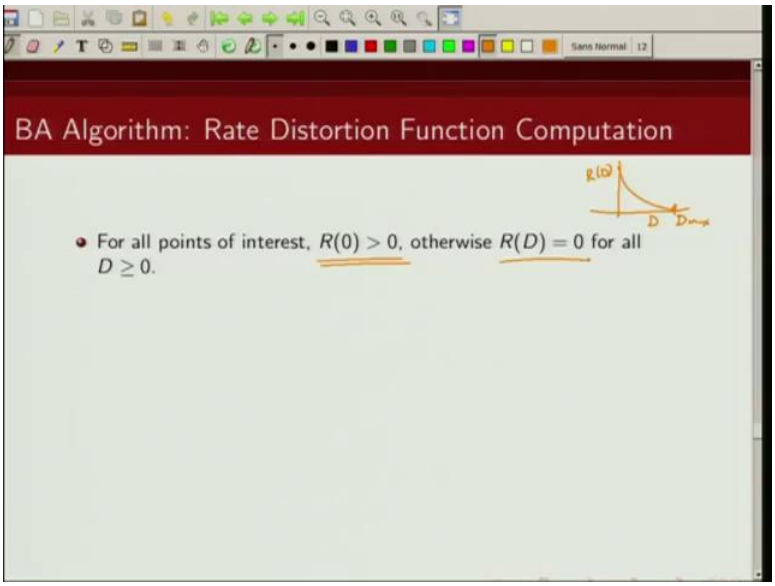
(Refer Slide Time: 34:04)



So, I took the reciprocals of the logarithm. So, then this log of this by this is now greater than equal to lambda times r 1 x log of q 1 divided by r r 1 plus 1 minus lambda r 2 log of q 2 r 2. So, I essentially took reciprocals of this log terms and this term which was earlier less than equal to is now greater than equal to. Next, I multiply both sides by p of y given x and sum over all x and y. So, if I do that what I get on the left hand side is function evaluated at lambda times r 1 plus 1 minus lambda times r 2 and lambda times q 1 plus 1 minus lambda time q 2 is greater than equal to lambda times function evaluated at r 1 q 1 plus 1 minus lambda times function evaluated at evaluated at r 2 q 2. And we know from the definition of concavity that if function evaluated at this is greater than equal to expected value of the function basically this is the condition for concavity.

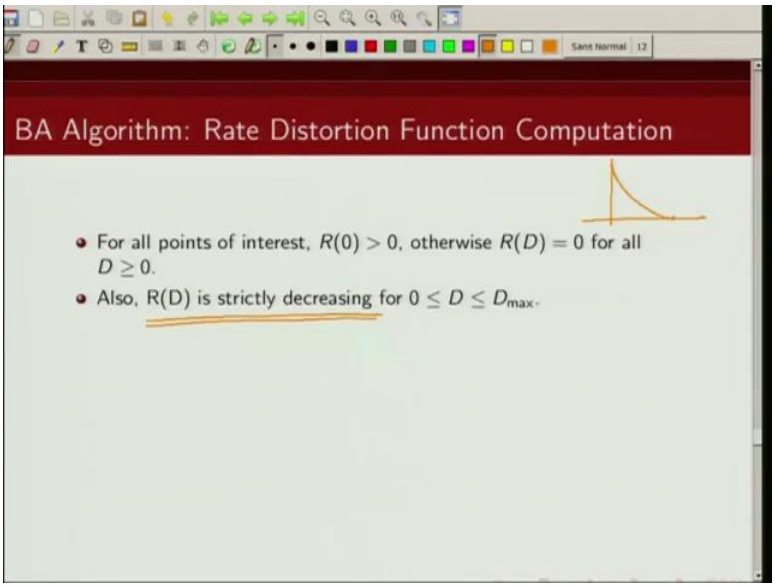
So, this shows that the function f is a concave function. So, hence we have shown that when we iteratively compute R and Q as k increases basically f evaluated at kth equation will tends towards the expression for channel capacity and this is Blahut-Arimoto algorithm for computation of channel capacity. Now Blahut-Arimoto algorithm for rate distortion function is very, very is a similar we will this quickly browse over the results.

(Refer Slide Time: 36:31)



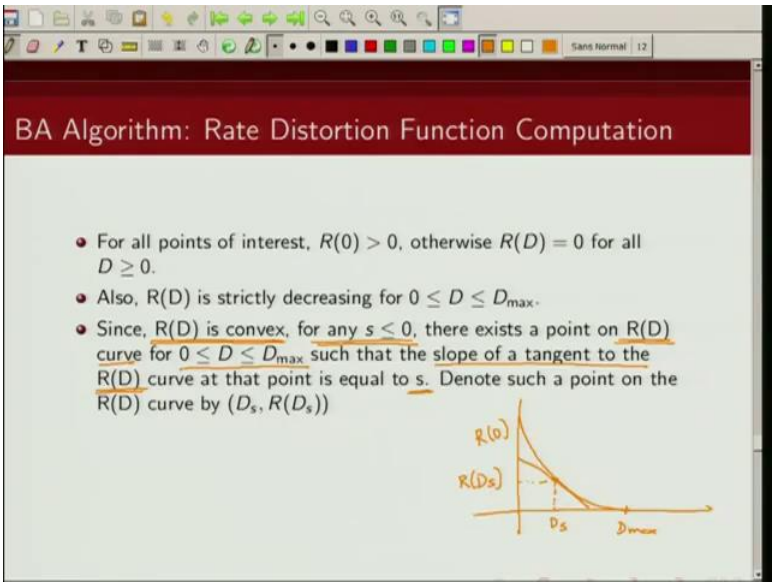
So, we know the rate distortion function looks like this typically something like this you have this is a rate this distortion. So, this is maximum distortion D max, R 0 typically greater than is equal to 0 or otherwise R 0, R D 0 for D greater than 0.

(Refer Slide Time: 37:03)



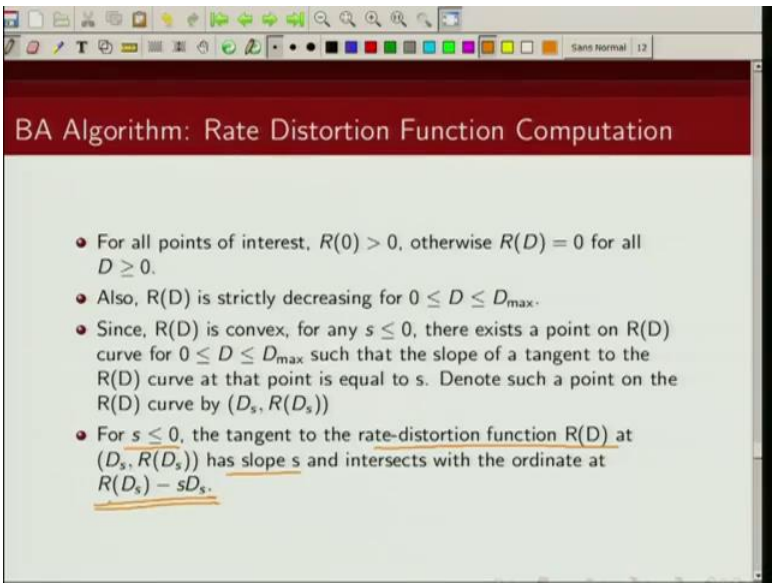
It is a typically a strictly decreasing function as we saw basically R of D typically is like this something like that.

(Refer Slide Time: 37:16)



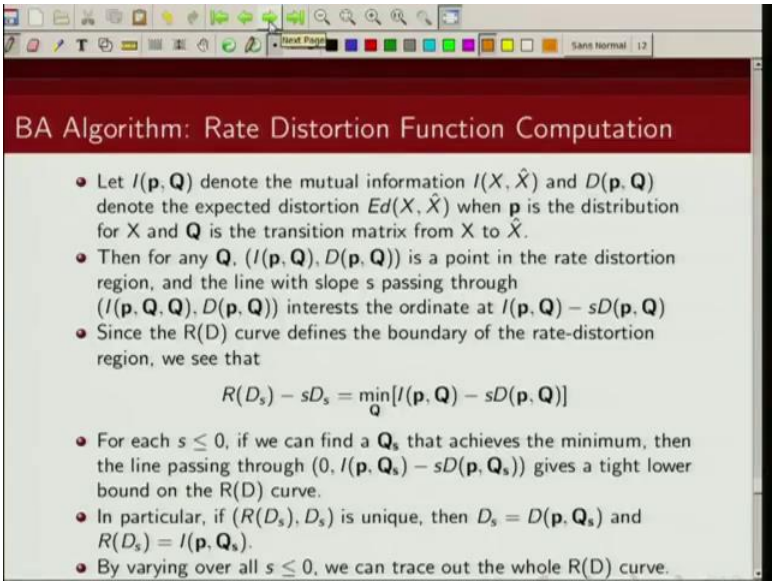
Now, we also know that this rate distortion function is a convex function. So, then for any s which is negative there exist a point on this rate distortion curve for D between 0 to D max such that the slope of the tangent to the rate distortion curve at that point is equal to this slope s. So, what I am saying is if you have a rate distortion function let us say something like this, this like D max. So, there exist a point and this rate distortion curve such that the slope of the tangent to this curve at that point is equal to s something s here. and let us denote this point s this is my D of s and this is my R of D of s.

(Refer Slide Time: 38:38)



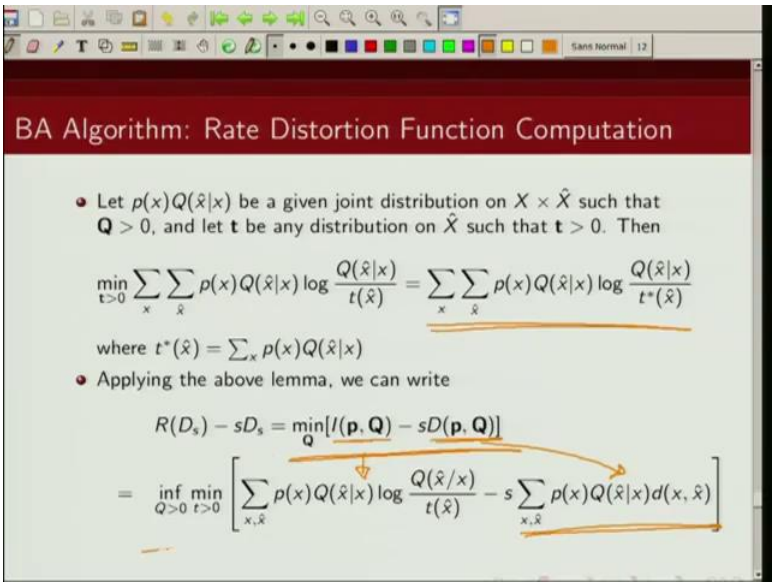
So, s less than 0, the tangent to the rate distortion function have slope s, and it is y intercept is given by R of D x minus s times D s.

(Refer Slide Time: 38:59)



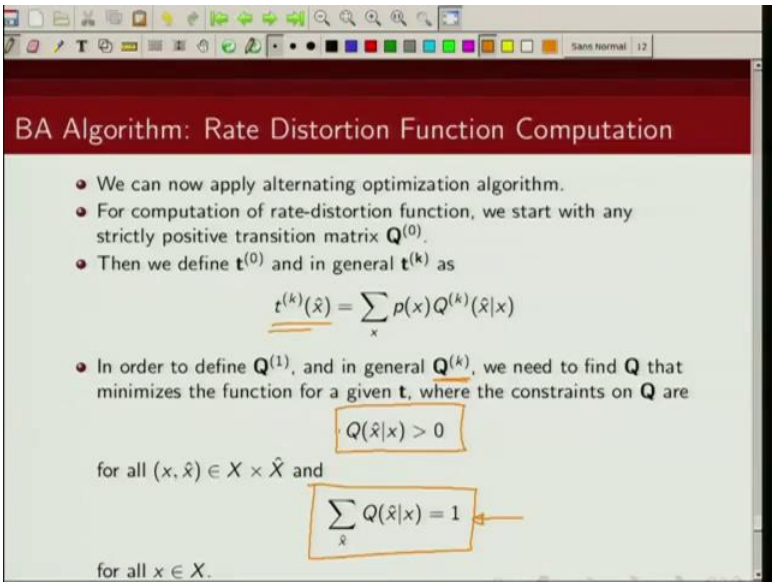
Now, let I p, Q denotes the mutual information between X and X hat and D p, Q denotes the expected distortion, where the p is the distribution for X and Q is the transition matrix from X to X hat. For any Q, then this is a point in this rate distortion region, and the line with slope s is going to pass through this point. So, this will give a y intercept of this. Now, since this rate distortion curves defines the boundary of the rate distortion region, we can write this as minimizing mutual information minus s time this distortion minimizing over Q. So, for each s less than equal to 0, we can find a Q of x such that this is minimize. So, we can find the Q s that will achieve this minimum value. So, then a line passing through this point is going to give me a tight lower bound on the r D curve and if I do this for all S's, I essentially can trace out this rate distortion curve.

(Refer Slide Time: 40:53)



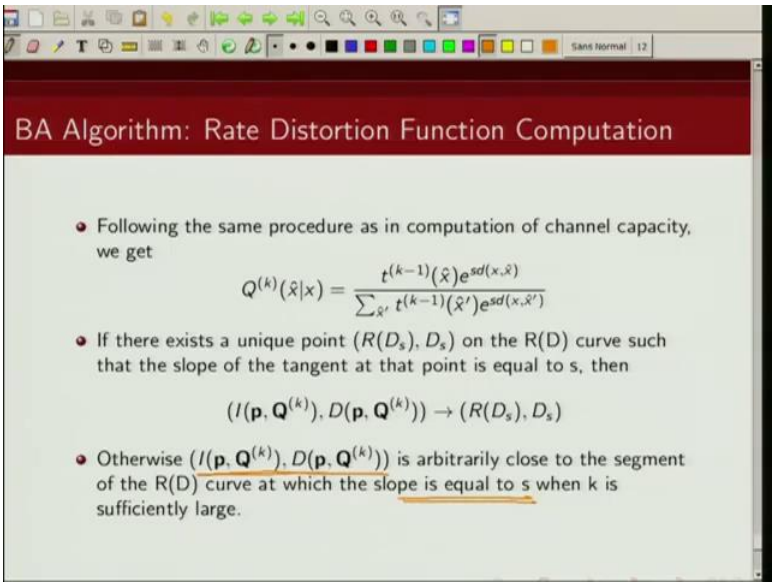
Now, let us talk about an iterative way to compute this thing. So, without proving, I am just stating the lemma that I am going to use the proof of this lemma is all on the similar lines as the proof that we did for lemma for computation of channel capacity. So, let p of x Q of x hat given x be the joint distribution on X and X hat such that Q is strictly positive and t is the distribution on x hat such that t is also strictly positive. Then minimizing this over all t is given by this expression where t star is this. Again this proof is very similar to we can use divergence inequality to proof this result, I am not stating it here again; again very similar to prove that we did for channel capacity, if now this is mutual information and this is my distortion, basically, R of D of minus s of D which is nothing, but minimizing this over all Q. Now, this mutual information is given by this and this is given by this channel. So, then I can write R of D s minus s of D s is basically this term.

(Refer Slide Time: 42:46)



Now, we are going to use alternative minimization algorithm to compute this. So, we start with any strictly positive matrix Q of 0. The next step is we need to find t at time at iteration index k and this is given by this. Now, this follows from the lemma that minimum value of this minimum over all t is given by this, where the value of t star is given by this expression. So, given a Q, you know what is the next t. Now, once you find t, you need then next find Q of 1 at Q at iteration index 1. Now how do I find Q in general for k, again this process is very, very similar to what we did you want to maximize this with some constraints. And what are this constraints we have those positivity constraints Q of x hat given x is greater than 0 and we have this sum of probability adding up to 1. So, we will again follow the method of Lagrange. So, we will form the Lagrangian of objective function plus lambda times this constraint, we will initially ignore the positivity constraints and later on we will show that Q comes out to be positive, so that condition is expressively taken care.

(Refer Slide Time: 44:46)



So, without showing you the details of this expression I am just directly writing the expression for Q for iteration index k and again this is exactly same procedure that we followed for when we compute computed R of at some iteration index x for computation of channel capacity. So, then if there exist a point R of D s D s on this rate distortion curve, such that the slope of the tangent at that point is equal to s then what we will see is as k increases, this converges to R of D s and D s, and if otherwise you will see that this term will be arbitrary close to a segment on this rate distortion curve, where slope is equal to s when k is sufficiently large. And the condition for convergence is so this function has to be convex function. So, I am not proving this, again very similar proof to what we did for the computation of the channel capacity.

So, to summarize, we have given to example one for the case of computation of channel capacity. We have shown how we can iteratively compute it. And similarly for the rate distortion function, as we said initially start off with some strictly positive value of Q of 0 and then we used that to compute t of 0 and then we use t of 0 to compute Q of 1, and we process that in a iterative fashion to compute the rate distortion function. So, with this, I will conclude this lecture.

Thank you.