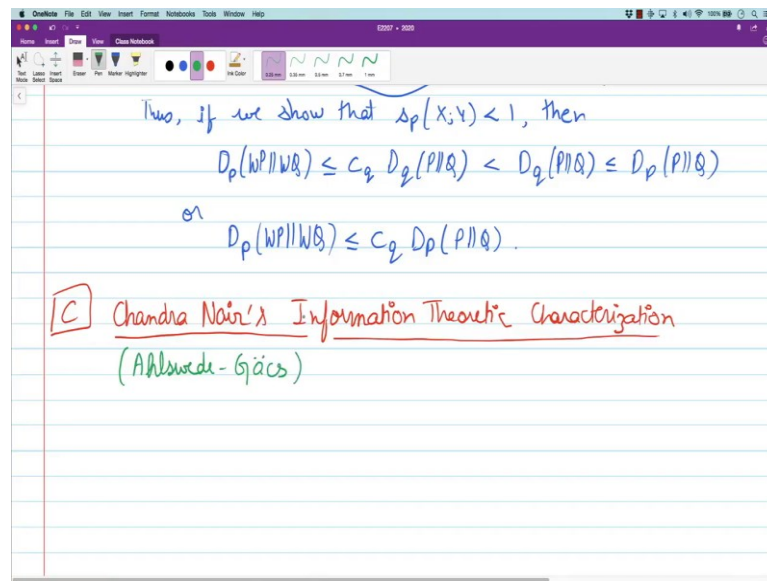


Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 23
An information theoretic characterization of hypercontractivity

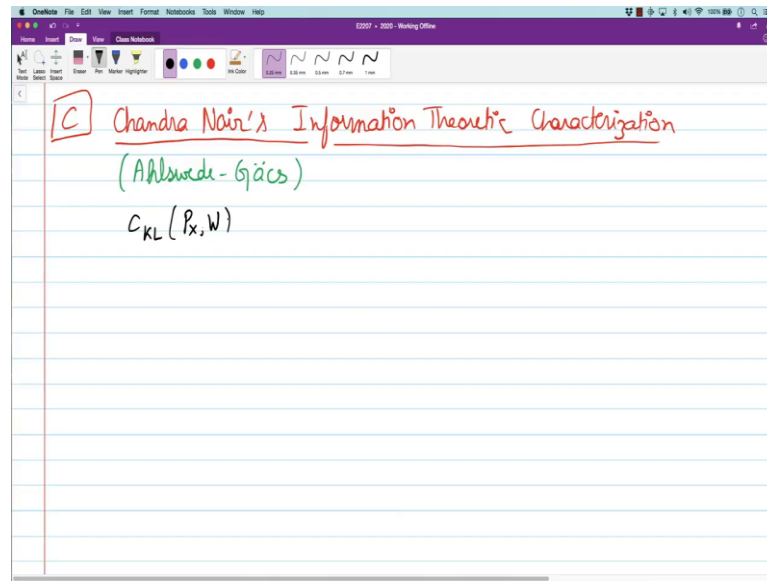
(Refer Slide Time: 00:20)



Finally, I will conclude with a slightly different characterization of hypercontractivity this constant $s_p(X; Y)$ and I will call this is an information theoretic characterization due to Chandra Nair. So, I will call this Chandra Nair's information theoretic characterization. I would like to remark that an earlier some of these results were already available in earlier work of Ahlswede and Gacs from the 70s and this is a recent paper from 2000 maybe 15 or so, I cannot remember the year. It was it is a conference paper.

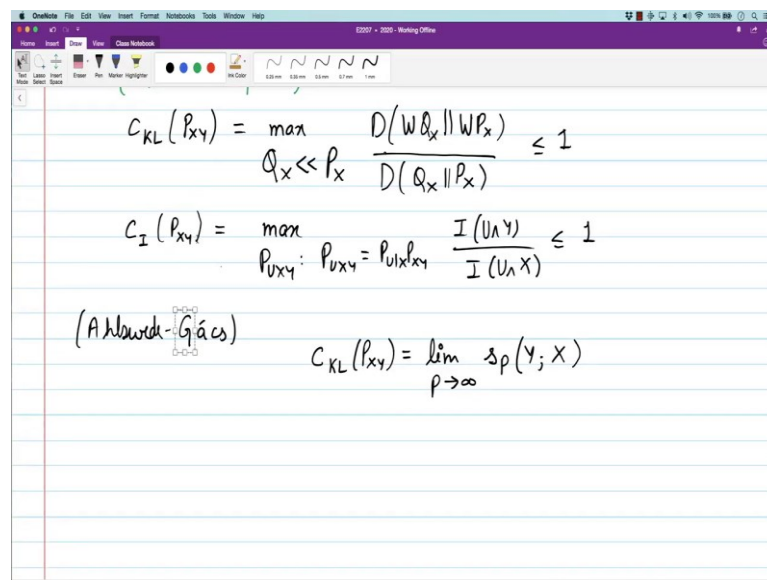
So, essentially what we will do is we will connect hyper contractive into so called strong data processing constants in information theory.

(Refer Slide Time: 01:33)



So, suppose you have this input distribution P_X and a channel w then the strong data processing constant associated with this P_X and W well.

(Refer Slide Time: 01:48)



We can just write it as the joint distribution P_{XY} is defined as max overall input distribution Q_X that are absolutely continuous with respect to P_X the divergence between $W P_X$ and $W Q_X$ divided by the divergence between Q_X and P_X ok; that is the strong data processing constant here.

And, similarly, we can define the strong data processing constant for mutual information as sorry, about this ok as the maximum over all distributions P_{UXY} such that U is such that P_{UXY} has a Markov relation. So, U can be formed only by looking at X is equal to $P_{U \text{ given } X} P_{X \text{ given } Y}$ of mutual information between U and Y divided by mutual information between U and X .

So, we know that all these coefficients by standard data processing inequalities that these coefficients are less than equal to 1 and when this coefficient is strictly less than 1, we call it a strong data crossing inequality, ok. So, it was shown by Ahlswede and Gacs. Sorry, I spelled it wrong. This is by this see the Gacs that this strong data processing constant for KL divergence is actually equal to limit P going to infinity $s P$ of not $X Y$, but $Y X$ ok.

Yeah, There is a there is an asymmetric treatment of X and Y here and this is the right order you have to take maybe, yeah. So, so that is what we defined this $s P$ in a particular order. So, that is the order in which this comes in here. In fact, here also it is not symmetric in $X Y$. This first coordinate is treated differently from the second one first is the input and second is the output ok. So, that is what it was shown that is quite interesting.

So, this strong data processing constant gets connected to hypercontractivity constant.

(Refer Slide Time: 05:31)

$$C_I(P_{XY}) = \lim_{p \rightarrow \infty} s_p(Y; X)$$

Define $K_P(X; Y) := \max_{Q_{XY} \ll P_{XY}} \frac{D(Q_Y \| P_Y)}{p D(Q_{XY} \| P_{XY}) - (p-1) D(Q_X \| P_X)}$

$$= \max_{Q_{XY} \ll P_{XY}} \frac{D(Q_Y \| P_Y)}{D(Q_X \| P_X) + p D(Q_{XY} \| Q_X P_{Y|X})}$$

And, later it was shown by I think Venkat Anantharam maybe I mean Gohari and Sudeep Kamath Chandra Nair is very recent about 2013 or so; that this guy itself is also equal to limit p going to infinity $s_p(Y, X)$ ok. So, these two strong data processing constants coincide, ok.

And what we will show now is a much more refined relation between strong data processing inequalities and hypercontractivity which will recover all of these results and this is the result due to Chandana. So, what is that result I will just call it a theorem. So, before I state theorem I need some more quantities to be defined now.

So, define κ for K κ for KL divergence κ_P let us say Y, X $\kappa_P(X, Y)$, it makes sense to think the way we define hypercontractivity, it makes sense to think of the first coordinate as input as the basic random variable here there and in this definition it makes sense to think of the other one as the input and that is why we have this sort of an switch in the order ok.

So, $\kappa(X, Y)$ is defined as \max over all distributions Q_{XY} that have density with respect to P_{XY} of divergence between P_Y and Q_Y divided by P_X * divergence between $P_{XY} Q_{XY} + - P_X - 1$ into divergence between sorry, P_X and Q_X and P_X , ok. The out the input output this is also equal to \max over Q_{XY} absolutely with respect to P_{XY} the divergence between Q_Y and P_Y just like the data processing inequality data processing strong data processing coefficient.

But, there is an additional term which comes and that additional term is p into this - this. So, that is the divergence between Q_{XY} and $Q_X P_Y$ given X ok that comes in.

(Refer Slide Time: 09:56)

$$Q_{XY} \ll P_{XY} \quad \frac{D(Q_{XY} \| P_X) + p D(Q_{XY} \| Q_X P_{Y|X})}{p}$$

$$\mu_p(X; Y) := \max_{P_{UXY}} \frac{I(U \wedge Y)}{p I(U \wedge X Y) - (p-1) I(U \wedge X)}$$

$$= \max_{P_{UXY}} \frac{I(U \wedge Y)}{I(U \wedge X) + p I(U \wedge Y | X)}$$

$$\lim_{p \rightarrow \infty} \kappa_p(X; Y) = C_{KL}(P_{XY}) \quad \text{and} \quad \lim_{p \rightarrow \infty} \mu_p(X; Y) = C_I(P_{XY})$$

Similarly, we can define μ ; μ for mutual information P_{XY} as max over all joint distribution P_{UXY} not necessarily the ones having the Markov relation of the output mutual information divided by $p \cdot \text{joint mutual information} - (p-1) \cdot I(U \wedge X)$ ok this is also equal to max over P_{UXY} by this $+ p$ into this - this which is just the conditional mutual information, ok.

So, we introduce these two new quantities κP_{XY} and μP_{XY} and if you take limit P going to infinity then look at this expression here; the only way this expression. So, if this guy is positive ok this is just a rough argument then as p goes to infinity, then this guy goes to 0. And, therefore, only way that you can have something more than 0 is when this is set to as P goes to infinity this also this thing also remains 0 and that means, that you want $P_{Y \text{ given } X}$ to be same as $Q_{Y \text{ given } X}$.

Using this what we can also see I just roughly outline the proof that this guy in the limit as P goes to infinity is exactly the same as strong data processing in equal constant for KL divergence for P_{XY} . And similarly, you take limit p going to infinity and look at this quantity once again the only way to get positive constant is to have this as 0 and this will happen.

If this is 0 then the then you can only look at those channels for which this Markov condition holds and then it coincides with the strong data processing constant for mutual

information ok. So, this is true. These in the limit as p goes to infinity, these quantities coincide with the strong data processing coefficient.

(Refer Slide Time: 12:59)

$\lim_{p \rightarrow \infty} K_p(X; Y) = C_{KL}(P_{XY})$ and $\lim_{p \rightarrow \infty} \mu_p(X; Y) = C_I(P_{XY})$
Theorem For all $p \geq 1$,
 $\Delta_p(Y; X) = K_p(X; Y) = \mu_p(X; Y)$
Proof idea $\Delta_p(Y; X) \leq K_p(X; Y) \leq \mu_p(X; Y) \leq \Delta_p(Y; X)$
 ■ Let $\theta < \Delta_p(Y; X)$. Then, there exist functions f and g s.t.
 $E[f(Y)g(X)] > \|f\|_p \|g\|_{p'}$

So, the main result that we will present is for each $p \leq p$ of $Y; X$ is equal to K_p of $X; Y$ which is also equal to μ_p of $X; Y$ ok that is the that is the main result that we have. So, proof is a little bit involved I will just give you the main idea. So, what we will show is that Δ_p of $Y; X$ is less than equal to K_p of $X; Y$ is less than equal to μ_p of $X; Y$ ok.

So, let us show this part first I am sorry and this itself is less than equal to Δ_p of $Y; X$. So, for this part this green part this p inequality here. Let θ equal to Δ_p of sorry, but θ be smaller than Δ_p of $Y; X$ ok then there exists functions f and g such that expected value of $f(Y)g(X)$ this guy exceeds θ and θ is smaller than the smallest possible. So, this must exceed.

Note that without loss of generality we can assume that both these norms are one because we can divide by these norms and define new function and this inequality must hold for that function.

(Refer Slide Time: 16:22)

Let $\theta < s_p(Y; X)$. Then, there exist functions f and g s.t.

$$\mathbb{E}[f(Y)g(X)] > \|f\|_{p_p} \|g\|_{p'}.$$
 WLOG assume that $\|f\|_{p_p} = \|g\|_{p'} = 1$ and

$$C := \mathbb{E}[f(Y)g(X)] > 1$$
 Consider functions $\bar{f}(y) = f(y)^{\theta p}$ and $\bar{g}(x) = g(x)^{\theta p'}$. We have

$$\mathbb{E}[\bar{f}(Y)] = \mathbb{E}[\bar{g}(X)] = 1$$
 and
$$\mathbb{E}[\bar{f}^{1/\theta p}(Y) \bar{g}^{1/\theta p'}(X)] = C > 1$$
 Define $Q_{XY} \ll P_{XY}$ as
$$\frac{dQ_{XY}}{dP_{XY}} = \frac{\bar{f}^{1/\theta p}(Y) \bar{g}^{1/\theta p'}(X)}{C};$$

So, without loss of generality assume that this norm = this norm = 1 and in this case when we assume that we must have $f(Y)g(X)$ the expected value must exceed 1, ok. We will this is this will show up as a constant in the calculation we will define it as C , ok.

So, now we define our measure change. So, we will use this functions to construct this measure $Q_{X; Y}$ that is used in kappa. So, define maybe before that let us prepare a little bit for this definition. So, consider functions \bar{f} given by f to the power θp and \bar{g} given by g to the power $\theta p'$, ok.

And, what we know is that expected value over Y of \bar{f} = expected value over X of \bar{g} = 1. So, this is by the assumptions that the norms are 1 and that expected value of \bar{f} to the power $1/\theta p$ * \bar{g} to the power $1/\theta p'$ over Y this is X is equal to C which exceeds 1 ok.

So, we use these functions to define our measure change argument. So, define we would define a new measure Q_{XY} as so, we will define it is density. The density that Q has with respect to P is given by $\bar{f}^{1/\theta p}(Y) \bar{g}^{1/\theta p'}(X)$ and we need to divide by C ok. So, that it normalizes to 1, ok. So, this is a definition of Q now.

(Refer Slide Time: 20:08)

$$\begin{aligned}
 & \rightarrow p D(Q_{XY} \| P_{XY}) - (p-1) D(Q_X \| P_X) \\
 &= \frac{p}{\theta} E_{Q_{XY}} [\log \bar{f}(Y)] + \frac{p}{p'} E_{Q_{XY}} [\log \bar{g}(X)] \\
 &\quad - p \log C - (p-1) E_{Q_X} \left[\log \frac{dQ_X}{dP_X}(X) \right] \\
 &= \frac{1}{\theta} E_{Q_{XY}} [\log \bar{f}(Y)] + (p-1) E_{Q_X} \left[\log \frac{\bar{g}(X)}{\frac{dQ_X}{dP_X}(X)} \right] - p \log C
 \end{aligned}$$

Now, let us compute all the quantities involved for this Q. So, let us say D let us compute the denominator first of denominator of kappa for this Q first. So, we have $p * D Q XY P XY + - p - 1 * D Q X P X$ this is equal to $P * \text{expected value log of this guy}$. So, $\log f Y f$ bar Y to the by theta p that is the first term here + beta P by P prime expected value log g bar X and then $p \log C$, ok.

This is the expected value of log of this guy that is what divergence and then this expectation is with respect to p of course, when we do not write anything the expectation with respect to P that is the bigger probability measure. And, now, $- p - 1 * \text{expected value of log d Q X by d P X}$, ok.

So, in fact, this guy here this is equal to 1 by theta expected value of log f bar Y + p - 1 into expected value of log this g bar X by d Q X by P X, So, the these expectations have to be carefully taken all these expectations are with respect to Q XY QXY and this one is again Q X because we did not include the extra factor here. So, again Q XY Q XY ok all the expectations are with respect to Q XY - p log C that is what we have and what we.

(Refer Slide Time: 24:29)

Handwritten derivation on a OneNote page:

$$-p \log C$$

$$\leq \frac{1}{\theta} \mathbb{E}_{Q_{XY}} [\log \bar{f}(Y)] + (p-1) \mathbb{E}_{Q_X} \left[\log \frac{\bar{g}(X)}{\frac{dQ_X}{dP_X}(X)} \right]$$

$$\text{(Jensen's inequality)} \leq \frac{1}{\theta} \mathbb{E}_{Q_{XY}} [\log \bar{f}(Y)] + (p-1) \log \mathbb{E}_{Q_X} \left[\frac{\bar{g}(X)}{\frac{dQ_X}{dP_X}(X)} \right]$$

$$= \mathbb{E}_{P_X} \left[\frac{dQ_X}{dP_X} \cdot \frac{\bar{g}}{\frac{dQ_X}{dP_X}} \right]$$

$$= \mathbb{E}_{P_X} [\bar{g}(X)]$$

So, this C is greater than 1 therefore, this guy is less than equal to 0. So, since C is greater than 1. So, we can just drop this the great this is something positive we are subtracting that part is this one $\log \bar{f}(Y) + p - 1$. Now, let us look at this guy here. So, this one is expected value over Q_{XY} log of $\bar{g}(X)$ by the Radon – Nikodym Q derivative of Q_X with respect to P_{XX} .

If we take this expectation inside the log by Jensen's inequality it will only increase the thing, ok. We get $\frac{1}{\theta} \log \bar{f}(Y) + p - 1$ log of expected value with respect to Q_{XY} or just Q_X actually it depends only on the marginal. So, let us just say Q_X of $\bar{g}(X)$ by $\frac{dQ_X}{dP_{XX}}$.

But, this expectation is can be written as. So, now, let us look at just this part here just this part here. This can be written as expected value over P_X of $\frac{dQ_X}{dP_X}$ into \bar{g} bar divided by $\frac{dQ_X}{dP_X}$ ok if this you can do a $\frac{dP_X Q_X}{P_X}$ is a positive orbital. So, this cancels. So, this is only expected value over P_X of $\bar{g}(X)$ and what do we know about that? Well, what we know about that part is that the way we have normalized that part this is 1. So, this thing is 1 and therefore, this is log of 1 which is 0.

(Refer Slide Time: 27:31)

Handwritten derivation of Jensen's inequality for a convex function g :

$$\begin{aligned}
 \text{(Jensen's ineq)} &\leq \frac{1}{\theta} \mathbb{E}_{Q_{XY}} \left[\log \bar{f}(Y) \right] + (p-1) \log \mathbb{E}_{Q_X} \left[\frac{\bar{g}(X)}{\frac{dQ_X}{dP_X}(X)} \right] \\
 &= \mathbb{E}_{P_X} \left[\frac{dQ_X}{dP_X} \cdot \frac{\bar{g}}{\frac{dQ_X}{dP_X}} \right] \\
 &= \mathbb{E}_{P_X} [\bar{g}(X)] \\
 &\leq \cdot
 \end{aligned}$$

(Refer Slide Time: 27:36)

Handwritten derivation showing the relationship between KL divergence and the likelihood ratio:

$$\begin{aligned}
 &= \frac{1}{\theta} \mathbb{E}_{Q_{XY}} \left[\log \frac{\bar{f}(Y)}{\frac{dQ_Y}{dP_Y}(Y)} \right] + \frac{1}{\theta} \mathbb{E}_{Q_Y} \left[\log \frac{dQ_Y}{dP_Y} \right] \\
 &\leq 0 + \frac{1}{\theta} D(Q_Y \| P_Y) \\
 &\Rightarrow \frac{D(Q_Y \| P_Y)}{p D(Q_{XY} \| P_{XY}) - (p-1) D(Q_X \| P_X)} \geq \theta \Rightarrow \boxed{K_P(X; Y) \geq \theta} \\
 &\text{Since this holds for every } \theta < \underbrace{S_P(Y; X)}_{K_P(X; Y)}, \\
 &K_P(X; Y) \geq S_P(Y; X).
 \end{aligned}$$

So, this is less than equal to or this is exactly equal to 1 by theta expected value of $Q_{XY} \log f \bar{Y}$ and just like this calculation we can again multiply and divide by dQ_Y by P_Y the log likelihood ratio. Sorry, the Radon-Nikodym derivative for Q_Y with respect to P_Y the likelihood ratio we can do that.

So, what we get is one by theta expected value of $Q_Y \log dQ_Y$ by dP_Y ok, but this quantity again by Jensen's inequality this is just the same calculations here and the

because of the fact that expected value of \bar{f} is 1 under P . This is this is $0 + 1$ by θ and we can recognize the term here as a divergence between Q_Y and P_Y .

So, if you look at this is the numerator of the expression of κ , this is less than equal to 1 by $\theta * \frac{1}{\theta} * \frac{1}{\theta} * \frac{1}{\theta}$ that is the denominator this is the numerator. So, we get that $D(Q_Y \| P_Y)$ by P into $D(Q_{XY} \| P_{XY}) - p - 1 D(Q_X \| P_X)$ is greater than equal to θ which implies that $\kappa_{P_X Y}$ is greater than equal to θ . And since this holds for every θ less than $S_{p_Y; X}$ we must have $\kappa_{P_X Y}$ greater than equal to $S_{p_Y; X}$, ok.

This is almost complete proof some infinite we have treated some sup as max and so on and so forth. In fact, we used max in a definition of $\kappa_{P_X Y}$ should have used sup if you wanted to be more careful, ok. So, we have shown this inequality. Now, let us show the other inequality which is this one here. So, the fact that $\kappa_{p_X Y}$ is less than equal to $\mu_{p_X Y}$.

(Refer Slide Time: 30:58)

Handwritten notes on a OneNote page showing mathematical derivations for the inequality $\kappa_P(X; Y) \leq \mu_P(X; Y)$.

The notes include the following content:

- At the top, the inequality $\kappa_P(X; Y) \geq S_P(Y; X)$ is written.
- Below it, the inequality $\kappa_P(X; Y) \leq \mu_P(X; Y)$ is boxed.
- Then, the expression for $\kappa_P(X; Y)$ is given as:

$$\kappa_P(X; Y) = \frac{D(Q_Y \| P_Y)}{p D(Q_{XY} \| P_{XY}) - (p-1) D(Q_X \| P_X)}$$
- Following this, it is noted: "(WLOG, we may assume that $\frac{dQ_{XY}}{dP_{XY}} \leq M$ a.s.)"
- Finally, it says: "Define $P_{U|XY}$ as follows: $U \in \{0, 1\}$ with $P_U(0) = \varepsilon$, $P_U(1) = 1 - \varepsilon$."

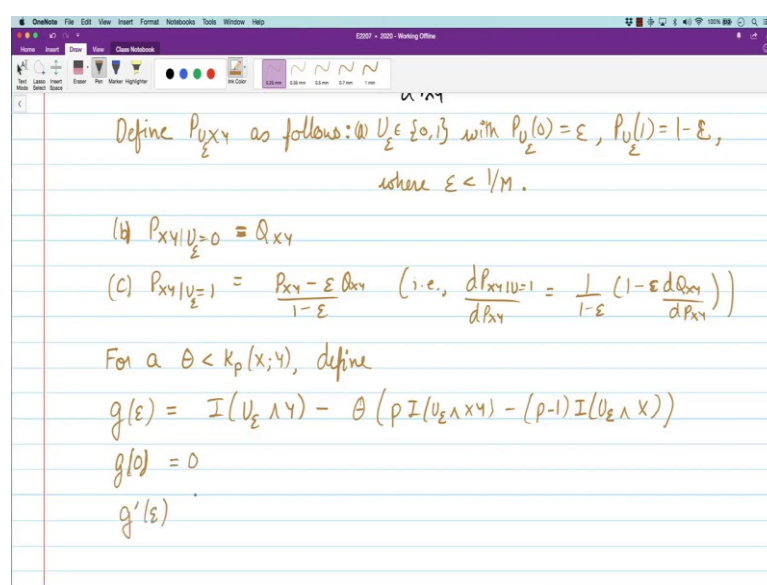
Again I am showing this because these are very interesting counts tell you about how divergence behave and how divergence how divergences can be related to mutual information. This is the next thing we will show, ok and this is a very interesting proof, it is a nice perturbation argument, ok.

So, suppose because you were writing kappa as a max although it is a sup. So, this is a little bit informal, but that is fine suppose kappa p X; Y is attained by a particular distribution Q, suppose this is equal to this we will use this distribution Q X Y to define our random variable u which is used in the definition of mu X Y mu p X Y. So, how do we define that?

So, consider the distribution before that it is a technical statement I am putting it here without loss of generality, we may assume that this guy here is less than equal to some large constant M almost surely with probability 1, you can assume that it is bounded ok this boundedness will be used in our proof, ok. So, let us assume that.

Now, with this assumption we can define this define this distribution P UXY which needs to have the marginal of X; Y corresponding to P X; Y as follows. So, P U so, U takes the value 0 and 1 with P U of 0 is epsilon and P U of 1 is 1 - epsilon, ok.

(Refer Slide Time: 33:44)



Define $P_{U_{\epsilon}XY}$ as follows: $U_{\epsilon} \in [0,1]$ with $P_{U_{\epsilon}}(0) = \epsilon$, $P_{U_{\epsilon}}(1) = 1 - \epsilon$, where $\epsilon < 1/M$.

(b) $P_{XY|U_{\epsilon}=0} = Q_{XY}$

(c) $P_{XY|U_{\epsilon}=1} = \frac{P_{XY} - \epsilon Q_{XY}}{1 - \epsilon}$ (i.e., $\frac{dP_{XY|U_{\epsilon}=1}}{dP_{XY}} = \frac{1}{1 - \epsilon} (1 - \epsilon \frac{dQ_{XY}}{dP_{XY}})$)

For a $\theta < \kappa_p(x,y)$, define

$g(\epsilon) = I(U_{\epsilon} \leq Y) - \theta (\rho I(U_{\epsilon} \leq X,Y) - (\rho-1) I(U_{\epsilon} \leq X))$

$g(0) = 0$

$g'(\epsilon)$

And, where this epsilon is less than 1 by M. So, this is the first part of the definition uh. Second part of the definition is P XY given the fact that u equal to 0 is the same as Q XY same as Q XY, then P XY given U = 1 = P XY - epsilon Q XY by 1 - epsilon. So, what does it mean? That is this guy has a density with respect to P XY and that density is given by 1 - 1 by 1 - epsilon into 1 - epsilon * density of Q XY with B XY. And, since this density is less than equal to M, this thing is always less than 1 and therefore, this is a

valid density and you can also show that it integrates to one easily ok. So, that is the second distribution ok.

Now, so, this is a distribution that we have defined. Now, let us define this function for a lambda for let us say theta again theta less than kappa p XY define g of epsilon as mutual information. So, this U depends on epsilon. So, I will call it U epsilon, ok. So, it is the mutual information between U epsilon and Y - theta * the denominator that comes up in the definition of mu. So, I mu epsilon XY - p - 1 * I mu epsilon U epsilon X, ok.

So, one thing you can check is that g of 0, so, what happens when you have 0? Then under 0 this coincides this distribution just has one value which is P XY and therefore, the mutual information is just U is a constant actually. So, this is just 0 0 0. So, everything is 0 here. So, this is 0. What about now the derivative of this guy around epsilon?

(Refer Slide Time: 37:25)

$$\lim_{\epsilon \rightarrow 0} g'(\epsilon) = D(Q_Y \| P_Y) - \theta (p D(Q_{XY} \| P_{XY}) - (p-1) D(Q_X \| P_X))$$

$$> 0$$

Therefore, $\exists \epsilon' > 0$ s.t. $g(\epsilon') > 0$

$$\Rightarrow \frac{I(U_{\epsilon'} \wedge Y)}{p I(U_{\epsilon'} \wedge XY) - (p-1) I(U_{\epsilon'} \wedge X)} > \theta$$

$$\Rightarrow \mu_p(X; Y) > \theta$$

Thus, $\mu_p(X; Y) > \kappa_p(X; Y)$.

In fact, we can comment on its derivative around epsilon in the limit as epsilon goes to 0. And it turns out that this is an observation I think it is I it is Chandra Nair paper, but I think Chandra Nair it is to attributes this to Guhari and maybe it goes back further. So, if you take at this take this limit this guy here the all these all this mutual information they go back to divergences of the two perturbations used in constructing this U E ok.

So, these divergences are actually the limiting values for this mutual information and this you can actually validate very easily by differentiating and taking the limit. So, all the mutual information becomes the divergence between corresponding distributions used in defining this ok.

And since we have assumed that so, since we have assumed that θ is less than κ and this ratio by the way is κ therefore, if you take this part out which is the non-negative part then you get some non-negative part into $\kappa - \theta$ which must be strictly greater than 0, ok.

So, here is a function which starts at 0 and its derivative is strictly greater than 0 and 0 therefore, there exist some ϵ prime greater than 0 such that g of ϵ prime is strictly greater than 0 which in turn implies that. So, this is greater than 0 what does it mean to have this greater than 0 which means this by $P_{XY} - p_X - p_Y + 1$ $I(U, \epsilon)$ Y is greater than θ which in turn implies that $\mu_{X; Y}$ is greater than θ .

Now, as before we have shown that for any θ that is less than, so, we choose some arbitrary θ here for any θ where θ is my θ chosen for any θ less than κ $\mu_{X; Y} > \theta$ $\mu_{X; Y}$ must exceed θ thus $\mu_{X; Y}$ must exceed κ $\mu_{X; Y}$ ok. So, that this completes this completes the second part ok this is this is the second part and both these parts are sort of self contained analytics, so, I could present them

However, this last part requires a lot of ideas from information theory. In fact, so, this that is why I am calling this as sketch.

(Refer Slide Time: 41:24)

Theorem For all $p \geq 1$,

$$\Delta_p(Y; X) = K_p(X; Y) = \mu_p(X; Y)$$

Proof idea $\Delta_p(Y; X) \leq K_p(X; Y) \leq \mu_p(X; Y) \leq \Delta_p(Y; X)$

Let $0 < \Delta_p(Y; X)$. Then, there exist functions f and g s.t. $\mathbb{E}[f(Y)g(X)] > \|f\|_p \|g\|_{p'}$. (uses ideas from multi-terminal info theory)

WLOG assume that $\|f\|_p = \|g\|_{p'} = 1$ and

$$C := \mathbb{E}[f(Y)g(X)] > 1$$

Consider functions $\tilde{f}(Y) = f(Y)^{1/p}$ and $\tilde{g}(X) = g(X)^{1/p'}$. We have

$$\mathbb{E}[\tilde{f}(Y)] = \mathbb{E}[\tilde{g}(X)] = 1$$

So, for this last part here uses ideas from multi terminal information theory actually. So, I will just tell you how we do it. We use the fact that this Δ_p tends to arise. First we use that $\mu_p(X; Y) \leq \mu_p(X; Y \cup Z)$ that is what that is that is true because X and X and Y has more options. And then we show that $\mu_p(X; Y \cup Z)$ actually is less than equal to $\Delta_p(Y; X)$ for n sufficiently large or in the limit as n goes to infinity and that completes the proof.

How do we show this last part? The goal is to use this distribution is auxiliary U which will be defined for remember that there will be this U we will have a joint distribution $P_{U \times Y \times Z}$ and we want to use that distribution to define these functions which satisfy to define these functions so that for those functions this inequality that the hypercontractivity inequality is violated. This holds in the opposite direction ok.

How do we find these functions? Well we actually construct functions as indicator functions of sets, ok. And those sets so, we have to actually construct sets and constructing such sets is a typical activity is a standard activity, a standard problem one needs to solve a multi terminal information theory when we derive some coding theorem. So, we use some typicality arguments so called typicality arguments to define those sets.

One more observation here is that while for a given X, Y, Z $\Delta_p(Y; X)$ can be attained by must be attained by arbitrary functions, but $\Delta_p(X; Y)$ and $\Delta_p(Y; Z)$ for sufficiently large n can almost be attained by almost be attained by sets, ok. So, there is actually no loss in generality in

using these sets. In fact, we show that this upper bound holds therefore, we show that for large enough n it suffices to consider it suffices to achieve $s p Y$ and $X n$, it suffices to consider indicator functions of sets subsets of $X n$ and $Y n$ rather than this arbitrary functions ok.

Yeah, I went over this part quite quickly, but this one requires a lot more effort if you do not know these ideas from information theory, but on the other hand if you know these ideas this is a very elegant proof which uses some standard construction from information theory to some typicality construction to show this inequality and exploits the tensorization property of $s p Y X$, ok. So, this is the proof.

(Refer Slide Time: 44:33)

Consider two random variables X and Y with joint distribution P_{XY} .

Let $f: \mathcal{X} \rightarrow \mathbb{R}$. Define $g: \mathcal{Y} \rightarrow \mathbb{R}$ as

$$g(y) = \mathbb{E}[f(X) | Y=y].$$

Denote by W the channel $P_{X|Y}$ and, with an abuse of notation, we denote the operator that converts f to g by W .

$g = Wf$ → Markov kernel

Claim: W is a contraction for L_p norm, $p \geq 1$, i.e.,

$$\|W\| = \|W\|_p = \mathbb{E}[P(Y)]^{1/p}$$

So, in summary what we have seen in this lecture is, we have defined this we have started studying this Markov Kernel the conditional expectation.

(Refer Slide Time: 44:36)

we denote the operator that converts f to g by W .

$$g = Wf \rightarrow \text{Markov kernel}$$

Claim: W is a contraction for L_p norm, $p \geq 1$, i.e.,

$$\|Wf\|_p = \|g(y)\|_p = \mathbb{E} [g(y)^p]^{1/p} \leq \|f\|_p = \mathbb{E} [f(x)^p]^{1/p}.$$

Proof. $\|Wf\|_p^p = \mathbb{E} [\mathbb{E} [f(x)^p | y]^p] \leq \mathbb{E} [\mathbb{E} [f(x)^p | y]]$ (by Jensen's inequality)

And, we have checked that this Markov Kernel is actually a contraction for any LP norm.

(Refer Slide Time: 44:44)

B Hypercontractivity of the Markov kernel

Definition. A joint distribution $P_{X,Y}$ is (p, q) -hypercontractive, $1 < q \leq p < \infty$, if

$$\|Wf\|_p \leq \|f\|_q \quad \forall f \in L_2(X).$$

Note that $\|f\|_q$ is a nondecreasing function of q . Therefore, the previous inequality is stronger than contractivity when $q < p$.

$$\Delta p(X; Y) := \inf \left\{ \frac{q}{p} : 1 < q \leq p \text{ and } P_{X,Y} \text{ is } (p, q)\text{-hypercontractive} \right\}.$$

We get an interesting inequality when $\Delta p(X; Y) < 1$.

And, then we showed that when we define this notion of hypercontractivity when can Markov Kernel be more than a contraction when can this norm on the right side be less than p .

(Refer Slide Time: 44:54)

Equivalent forms and simplifications

(1) $\|Wf\|_p \leq \|f\|_q$ if and only if

$$\mathbb{E}[f(x)g(y)] \leq \|f\|_q \cdot \|g\|_{p'} \quad \forall g \in L_{p'}(Y)$$

where

$$p' = \frac{p}{p-1}$$

Proof: $\mathbb{E}[f(x)g(y)] = \mathbb{E}[Wf(y)g(y)]$

(by Hölder's inequality) $\leq \|Wf\|_p \|g\|_{p'}$

(by hypercontractivity of W) $\leq \|f\|_q \|g\|_{p'}$

And, so, we define this quantity $s_p(X, Y)$ rather than finding when or what is this $s_p(X, Y)$ for a particular distribution. What we have been doing is trying to understand this question itself by expressing various equivalent forms we saw that there is an inner product form which strengthens holder inequality.

(Refer Slide Time: 45:13)

(2) Tensorization of hypercontractivity

$$s_p(X, Y) = \inf \left\{ \frac{q}{p} : 1 < q \leq p \text{ and } P_{X,Y} \text{ is } (p, q)\text{-hypercontractive} \right\}$$

$$s_p(X^n, Y^n) = \max_{1 \leq i \leq n} s_p(X_i, Y_i) \quad (\text{Tensorization})$$

$(X_i, Y_i)_{i=1}^n \sim \text{iid } P_{X,Y}$

Proof Suffices to show this for $n=2$.

To show: $s_p(X^2, Y^2) = \max \{ s_p(X_1, Y_1), s_p(X_2, Y_2) \} = s_p(X, Y)$

Consider $f: X_1 \times X_2 \rightarrow \mathbb{R}$ and denote $W^2 := W_1 \otimes W_2$

$$W^2 f(y) = \mathbb{E}[f(X_1, X_2) | Y_1 = y_1, Y_2 = y_2]$$

For convenience, for a func. $g(z_1, z_2)$, let $g_2(z_2) = g(z_1, z_2)$

Then we saw that this $s_p(X, Y)$ tensorizes. So, this is equal to $s_p(X)$ and $s_p(Y)$ and for independent and identically distributed $X_1, \dots, X_n, Y_1, \dots, Y_n$.

(Refer Slide Time: 45:29)

(3) Relate hypercontractivity to strong data processing inequality

Recall that

$$D(WP \| WQ) \leq D(P \| Q) \quad (\text{data processing inequality})$$

→ The same inequality holds for Rényi divergence:

Rényi divergence of order α between $P \ll Q$ is given by

$$D_\alpha(P \| Q) = \frac{1}{\alpha-1} \ln \mathbb{E} \left[\left(\frac{dP}{dQ} \right)^\alpha \right]$$

$$(D_\alpha(P \| Q) = \frac{1}{\alpha-1} \ln \sum_x P(x)^\alpha Q(x)^{1-\alpha})$$

$$D_\alpha(WP \| WQ) \leq D_\alpha(P \| Q)$$

And, we also saw the connection of this hypercontractivity in fact, equivalence of this with the strong data processing inequality for any divergences.

(Refer Slide Time: 45:34)

$$D_\alpha(WP \| WQ) \leq D_\alpha(P \| Q)$$

→ Suppose $P_{X,Y}$ is (p, q) -hypercontractive, $1 < q \leq p < \infty$.

Then, consider $f(x) = \frac{dP}{dQ}$ and assume that $\|f(x)\|_p < \infty$.

$$\|Wf\|_p \leq \|f\|_q$$

$$\Leftrightarrow \frac{p-1}{p-1} \ln \|Wf\|_p \leq \frac{q-1}{q-1} \ln \|f\|_q$$

$$\Leftrightarrow \frac{p-1}{p-1} \ln \|Wf\|_p = \frac{(q-1)}{q} D_q(P \| Q)$$

Note that $Wf(y) = \mathbb{E}[f(x) | Y=y]$

And, finally, we presented this is that equivalence showing this for all Q that that have density with respect to P is equivalent to $P \ll Q$ hypercontractivity.

(Refer Slide Time: 45:40)

Therefore, P_{XY} is (p,q) -hypercontractive \Leftrightarrow
for all $Q \leq P$ s.t. $D_p(P||Q) < \infty$,

$$D_p(WP||WQ) \leq \left(\frac{1 - 1/q}{1 - 1/p} \right) D_q(P||Q) \rightarrow < 1 \text{ if } q < p.$$

Thus, if we show that $s_p(X;Y) < 1$, then

$$D_p(WP||WQ) \leq c_q D_q(P||Q) < D_q(P||Q) \leq D_p(P||Q)$$

or

$$D_p(WP||WQ) \leq c_q D_p(P||Q).$$

And, finally, we saw we saw this alternative characterization in terms of divergence and mutual information $s_p XY$ is equal to this $\kappa_p XY$ is equal to $\mu_p XY$. By the way once we establish this equality this identity here, then by taking limit p going to infinity both these results follow, ok. Clearly the strong data processing constant is equal to is equal to $s_p p$ going to infinity and again the both the strong data processing inequality constants are equal to this and therefore, both of them must also be equal, ok.

So, this concludes this lecture. As I said we have not proved anything any particular bound on $s_p X Y$ or $s_p Y X$ for any distribution. We have just shown equivalent form of this hypercontractivity requirement. In the next lecture, we will try we will study hypercontractivity for Gaussian distribution.

So, called Gaussian hypercontractivity which is first shown by Gacs and remarkably what we will show is that Gaussian hypercontractivity is exactly equivalent to Gaussian log Sobolev inequality. So, this amazing phenomenon of hypercontractivity is also equivalent to for the Gaussian cases is also equivalent to this super inequality called the log Sobolev Gaussian log Sobolev inequality which we earlier saw was equivalent to stams in.