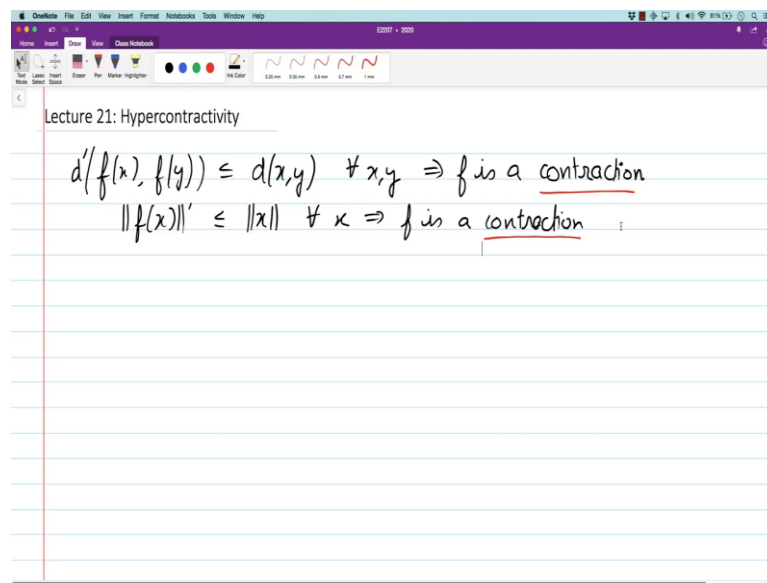


Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 22
Hypercontractivity and strong data processing inequality for Renyi divergence

(Refer Slide Time: 00:20)



In the last two lectures, we saw how log Sobolev inequality is connected to a classic information theoretic inequality called the Sturm's inequality. This was the connection between Gaussian log Sobolev inequality and Sturm's inequality. In the next two lectures, we will see the connection of Gaussian log Sobolev inequality with another very interesting notion namely that of Gaussian hyper contractivity.

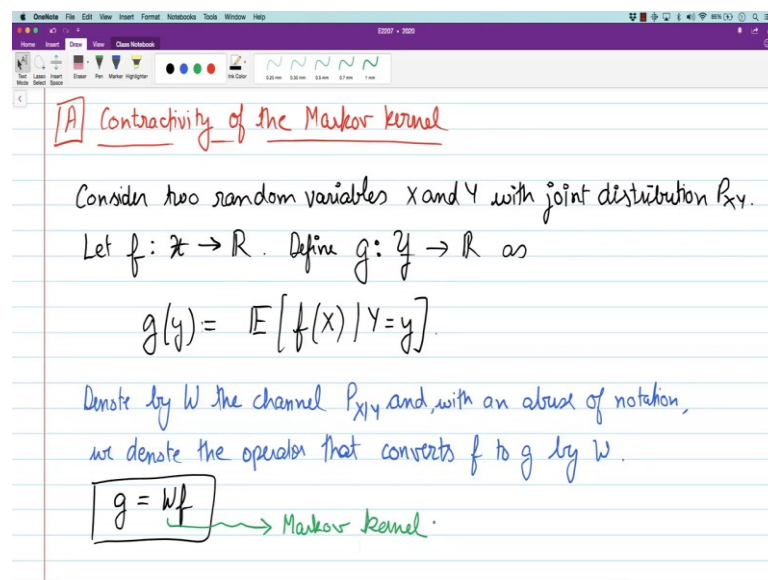
In order to talk about hypercontractivity, I should first describe contractivity in particular I should first talk about contraction. So, this is a quick review. I will not spend too much time on this and I will not be very formal. So, roughly speaking a mapping f on a metric space is a contraction if it shrinks distances. So, if the distance between $f(x)$ and $f(y)$ is less than equal to distance between x and y , then f will be a contraction ok.

So, this is what a contraction is because it shrinks distances, ok. And in fact, if this underlying metric is; if this underlying space is not only a metric space, but it has a norm then it suffices to talk about the norm of $f(x)$. By the way this distance

can be some other distance if you want we can have something d' instead of d , because f will take you to a different metric space; similarly this is a different norm, ok.

Suppose this is less than equal to norm x then for all x then implies f is a contraction. So, that is what a contraction is, ok. And in this lecture we will be interested in a very specific mapping namely the so called Markov Kernel.

(Refer Slide Time: 02:23)



And what we, let us begin by showing that the Markov kernel is a contraction. So, before we talk about hyper contractivity, let us talk about contractivity of the Markov Kernel, ok. So, what is the Markov Kernel?

So, consider two random variables X and Y with joint distribution $P_{X,Y}$. Let f be a mapping from x that is why this x takes values to \mathbb{R} , ok. So, $x \mapsto f(x)$ is a real value of mapping. I will put some rather than putting constraints on f up front I will put this constraints as we move forward when we require them. So, we are given this mapping f using this we can describe a mapping on Y as follows.

So, define g which is now a mapping from y to \mathbb{R} as $g(y) = \mathbb{E}[f(X) | Y=y]$ the conditional expectation of $f(X)$ given Y ; I am I will abuse the notation and write Y equal to y . So, this is basically this is a random variable and this is that, this is the realization of that random variable ok that is my $g(y)$ the conditional expectation, ok.

So, this operator we will use W to denote; denote by W the channel P_Y given X maybe X given Y ok, and with an abuse of notation we denote the operator that takes that denote the operator that converts f to g by W , ok. So this, so, what we have seen here is basically $g = W$ of f ok. So, that is that is the mapping.

So, this operator W can be thought of as mark is called Markov kernel ok. And, we will show that this Markov kernel is actually a contraction. So, to talk about contraction we should first think of a basic norm for random variables. So, let us use.

(Refer Slide Time: 06:11)

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. $\|Wf\|_p^p = \mathbb{E}[\mathbb{E}[f(X)^p | Y]] \leq \mathbb{E}[\mathbb{E}[f(X)^p | Y]] \quad (\text{by Jensen's inequality})$

$$= \mathbb{E}[f(X)^p] = \|f\|_p^p. \quad \square$$

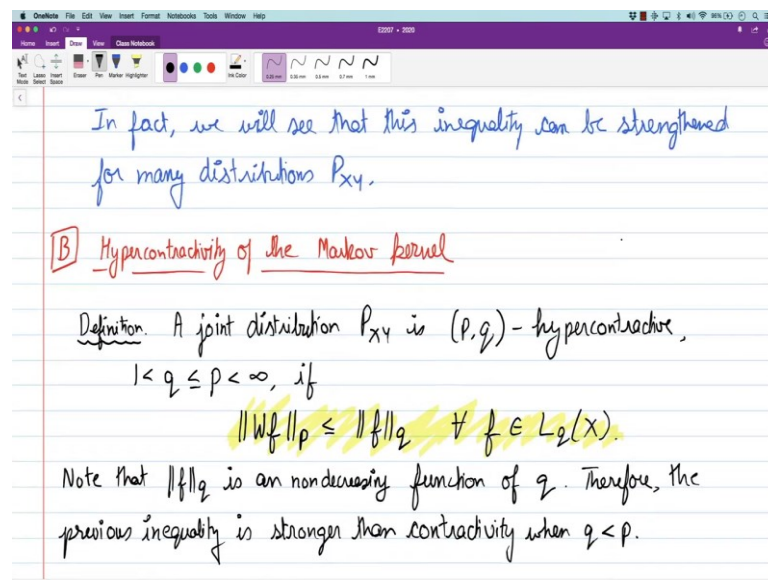
So, that is the claim; claim W is a contraction for we will use L_p norm ok; L_p norm. So, what is the L_p norm? So, how do we prove this? Firstly, what is the claim? So, claim is follows that is if you look at the random variable g of Y , look at its L_p norm. What is the L_p norm of g of Y ? This is the expected value of g of Y to the power p .

Claim is this guy is less than equal to; so, this is by the way this is norm of W of f is less than \mathbb{E} that is correct is less than equal to the L_p norm of f which is expected value of $f(x)$ to the power of p to the power 1 by p , ok that is the claim ok. So, how do we show this? Well, the proof is actually not so difficult. The proof is this Jensen's inequality ok. So, the let us see.

So, the L_p norm this L_p norm to the power p for this is simply expected value of expected value of $f(x)$ given Y to the power p . And, this is less than equal to if we take

this expectation inside right, this is by Jensen's inequality actually this is sometimes called conditional Jensen's ok. And now this is as we have seen several times this is exactly equal to expected value of X to the power p , right. So, that is the proof. So, indeed this is the contraction ok, and right.

(Refer Slide Time: 09:01)



And, so what we will show is that; in fact, we will see that this inequality can be improved can be strengthened for many distributions P_{XY} , ok. So, the fact that this inequality holds for all; this inequality here holds for all functions f ok, fact that Markov kernel is a contraction. That is a feature of the distribution P_{XY} it depends on the joint distribution P_{XY} and the claim here is that it can be strengthened ok. So, that is what and this strengthened inequality is called hypercontractivity.

So, you see that you are in a sense improving Jensen's whenever you can improve Jensen's you can have very very interesting deep consequences. So now, we will talk about the hypercontractivity of the Markov kernel, ok. So, let us bring in a definition. A, so this Markov kernel is associated with a joint distribution. So, we will just talk about a joint distribution.

A joint distribution P_{XY} is p, q hypercontractive ok; p, q hypercontractive where q is actually smaller can be smaller than or less than or q is less than or equal to p and so, we are ruling out p being infinity and 1 ok previous bound also holds for p equal to 1, but we

are ruling out those things. If this is just for convenience that we are ruling it out, it is not so useful if p is equal to 1.

If you look at the Markov kernel associated with P_{XY} and you take any function f like this p is less than equal to this q for all functions f which have finite q -th moment. So, this is basically functions f which lie in this space ok. These are the functions which have finite q -th moment. So, that is when we will call it p .

So, why is this stronger than; when can it be stronger than hypercontractivity when q is at least smaller than p . So, note that f_q is an increasing function of q . Therefore, the previous inequality is stronger than contractivity; stronger than contractivity when q is smaller than p . And in fact, we would like to find as smaller q as possible.

So this is a, this is the inequality that we want to prove. And, note that if this inequality holds for; if this inequality holds for a q it holds for all q prime greater than that q because of this monotonicity; is an increasing I should say non decreasing not increasing necessarily increasing ok. So, let us define the best q , ok.

(Refer Slide Time: 14:29)

previous inequality is stronger than contractivity when $q < p$.

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

We get an interesting inequality when $s_p(X; Y) < 1$.

Equivalent forms and simplifications

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

$$E[f(X)g(Y)] \leq \|f\|_{p'} \|g\|_q \quad \forall g \in L_2(Y)$$

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

So, let s_p for this joint distribution it is a function of joint distribution, but it is maybe I will just write it as $P_{X \times Y}$; P_{XY} is defined as; actually I think this notation is slightly better. So, I am viewing it as a function of the random variables, but surface function of the joint distribution, but this is for convenience.

This is how we indicate even sometimes in which information and information theory as a function of random variables, but it is a function of the joint distribution. So, we will define this guy as the infimum over q by p such that q is smaller than p , q is greater than 1 and smaller than p and P_{XY} is p, q hypercontractive ok. And the point is that we get an interesting more than interesting actually inequality when $s p X Y$ is less than 1 can you make it strictly smaller than 1.

Of course, this is it is already clear that it is less than equal to 1 because any P_{XY} is hype is $P P$ hyper contractive because $P P$ hypercontractivity is just contractivity. And, we want to make we want to check if it can be made less than 1, ok, that is hypercontractivity. There are many equivalent forms of the same requirement this inequality and we will present that, ok. So, let us do that equivalent forms simplification, ok.

So, first we note that $W f p$ is less than equal to $f q$ if and only if expected value of $f X g Y$ is less than equal to f , yeah. So, $f p$ prime $g q$ ok for all functions g which have a finite q -th moment this function is of y where. So, where this p prime by the way is the holder conjugate this. So, you know 1 by p prime $+ 1$ by $p = 1$, ok. So, that is what p prime is ok. So, how do we show this? Proof is easy.

(Refer Slide Time: 18:36)

$$E[f(X)g(Y)] \leq \|f\|_q \|g\|_{p'} \quad \# g \in L_{p'}(Y)$$

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

Proof: $E[f(X)g(Y)] = E[Wf(Y)g(Y)]$

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

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

showing \Rightarrow direction.

This is just Holder's inequality. Suppose this holds then expected value of $f X g Y =$ expected value of $W f$ of $Y g Y$, because this is the conditional expectation and g when

you take conditional expectation given by g Y comes out because it is a constant given Y , ok.

And, now we apply Holder's inequality. So, that gives us $W f$; ok maybe a small change here. So, this will be p prime and this would be q , yeah. So, $W f$ of p and g of p prime, right. And now, this is by Holder's inequality. And now, this is less than equal to $f q g p$ prime, ok.

So, this is an improvement of Holder's inequality because you have replaced this p -th norm by p prime here. You could have applied Holder's and you would have got p prime directly you could have applied it here, but this is smaller than that. So, this is a smaller value, so you have improved Holder's inequality ok. So, this direction is done.

How about the opposite implication. So, this shows that hypercontractivity implies this new form ok; by hypercontractivity of W .

(Refer Slide Time: 21:06)

For the other direction, consider $f \in L_p(X)$.

$$\begin{aligned} E[Wf(Y)^p] &= E[Wf(Y) \cdot Wf(Y)^{p-1}] \\ &= E[E[f(X)|Y] Wf(Y)^{p-1}] \\ &= E[f(X) Wf(Y)^{p-1}] \end{aligned}$$

Note that $E[Wf(Y)^{p-1} \cdot \frac{f}{E[f(X)|Y]}] = E[Wf(Y)^p] \leq E[f(X)^p] < \infty$.

$$\Rightarrow Wf \in L_{p'}(Y).$$

And, now for the other direction suppose this inequality holds for all g , then consider f in $L_q X$ ok; that is what we need to consider here. So, if you look at such an f and we look at expected value of $W f$ to the power $W f y p$ the p -th moment this is equal to expected value of $W f Y$ times $W f Y$ to the $p - 1$. So, this is exactly equal to expected value of; expected value of $f X$ given $Y W f Y p - 1$. So, which can be also written as expected value of $f X$ times $W f Y p - 1$.

And, now we use the other form of hypercontractivity with this being g . And note that this is indeed we can view this as g because; note that expected value of $W f Y$ of; so, this is to the power $p - 1$ suppose we did it is it to the power p prime. So, that is p by $p - 1$, $ok =$ expected value of $W f Y$ to the power p and that guy here is less than this is the contractivity property we saw earlier; $f X$ to the power p and that is less than equal to X to the power p . So, we must maybe we should assume this f is in p ok and this is finite. This is the assumption here ok .

Therefore, $W f$ belongs to $L^{p'} Y$ this was a requirement here by the way. Therefore, we can apply Holder's inequality.

(Refer Slide Time: 24:15)

Handwritten mathematical derivation on a OneNote slide:

$$\begin{aligned} \text{Note that } \mathbb{E} |Wf(Y)|^{p'} &= \mathbb{E} |Wf(Y)|^p = \mathbb{E} |f(X)|^p < \infty. \\ \Rightarrow Wf &\in L^{p'}(Y). \\ \Rightarrow \mathbb{E} [f(X) Wf(Y)^{p-1}] &\leq \|f\|_q \|Wf\|_{p'}^{p-1} \\ \Leftrightarrow \|Wf\|_p^p &\leq \|f\|_q \|Wf\|_{p'}^{p-1} \\ \Leftrightarrow \|Wf\|_p^{p(1-\frac{1}{p'})} &\leq \|f\|_q \\ \Leftrightarrow \|Wf\|_p &\leq \|f\|_q. \quad \square \end{aligned}$$

So, expected value of we can apply that improvement that we are assuming here $W f Y$ if your loss what we are trying to show is this inequality implies this inequality; suppose this inequality holds for all function g then this must hold. So, this guy here is less than equal to $f q$ and $W f p$ prime, but that is the same as this is $f q$ and this is now $W f p$ prime, but what is this? This is exactly equal to this. So, this guy here is $W f p$ to the power p ok ; looks like $W f$ to the power $p - 1$ p prime norm of that.

What is the p prime norm of $W f p - 1$? We just saw this. This is exactly equal to $W f p$ to the power p by p prime ok that is this norm here. So, what we get is this is the same as saying $W f p$ to the power p into $1 - \frac{1}{p}$ prime less than $f q$ and since p prime is the

holder conjugate of p , so, $1 - 1$ by p prime is 1 by p . So, this is just 1 . So, which is the same as saying $W f p$ is less than equal to $f q$, ok.

So, this is a standard way of expressing the these norm inequalities in terms of this inner products here, ok. So, you may have seen it in function analysis where you show some bound holes for all test functions in this class, then the same bound holds for all function; if this holds for this then this also holds. Yeah, these kind of tricks are used. And, we have chosen g carefully here, but it works out ok.

So, in other words we see that this particular form of hypercontractivity is form is equivalent to this strengthening of this strengthening of Holder's inequality, ok.

(Refer Slide Time: 27:17)

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

The next simplification that we introduce is the familiar tensorization property. So, if you want to establish hypercontractivity for product measure it suffices to show hypercontractivity for individual coordinates. In particular, remember this quantity $s_p X; Y$ it was defined as the least value of the ratio q by p such that this joint distribution P_{XY} is this is the least of all q greater than exceeding 1 , but less than equal to P . And such that P_{XY} is $p q$ hypercontractive, ok.

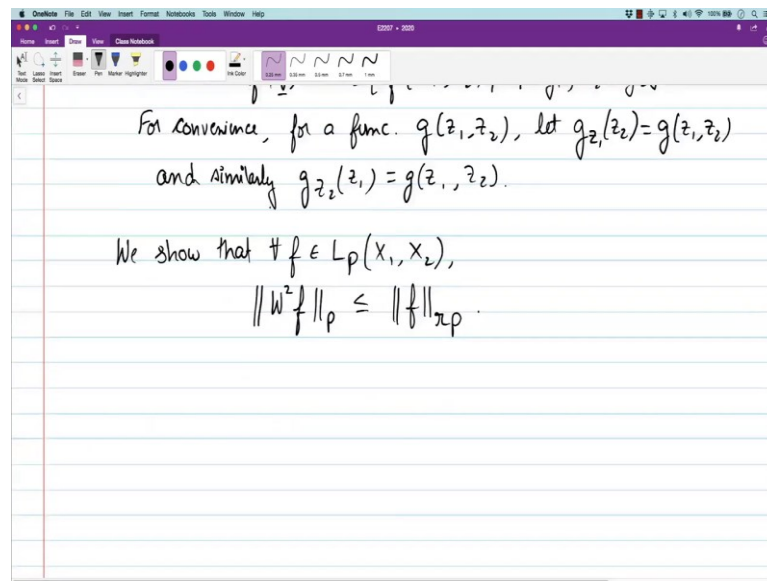
We have defined this quantity. Now, suppose you want to evaluate this quantity for product measures so, X_n and Y_n denote two independent distributed random variables. So, let us say $X_i Y_i$ for i equal to 1 to n are iid P_{XY} . Then this the quantity s_p of X_n

Y_n which must be less than 1 is equal to $\max_{1 \leq i \leq n} \mathbb{E} \sum_{j=1}^n X_j Y_j$, ok. This is the so called tensorization property. In particular, if they are identically distributed then they are then $\mathbb{E} \sum_{j=1}^n X_j Y_j = \mathbb{E} \sum_{j=1}^n X_j^2$. So, how do we show this?

I will present the proof sort of an old proof of this result from a paper by Ahlswede and Gacs, a little less known paper, but it is one of the initial papers looking at hypercontractivity.

So, this is how they prove it. Note that it suffices to show this for n equal to 2 and that is what we will do we will fix n to 2. So, we need to show that $\mathbb{E} \sum_{j=1}^2 X_j Y_j = \mathbb{E} \sum_{j=1}^2 X_j^2$ is what we need to show; $\max_{1 \leq i \leq 2} \mathbb{E} X_i, \mathbb{E} Y_i$ let us call this guy something let us call it r , the max we are denoting this max with r ok. So, how do we show this?

(Refer Slide Time: 30:25)



So, consider a function f from $X_1 \times X_2$ to \mathbb{R} and denote $W^2 = W_1 \otimes W_2$, the two channels in applied independently. That is what this W^2 is to be applied to the function f . The condition the; so, we will use this W^2 operator as before W^2 of f will be the conditional expectation of f given Y_1, Y_2 . So, this is again Y_1, Y_2 W^2 this is expected value of $f(X_1, X_2)$ given $Y_1 = y_1$ and $Y_2 = y_2$ that is the W^2 operator. And maybe one more convenient notation.

If you have a function say g , for a function g which has two arguments Z_1 and Z_2 we define g subscript Z_1 as a function of Z_2 as g of $Z_1 Z_2$. And, similarly g Z_2 of Z_1 is g of Z_1 comma Z_2 , ok. This is just for convenience ok. So, all the notation is set. So, what do we want to show?

Let us look at. So, we show that expect that for all f that are I think an L^p expected to an W^2 of f p is less than equal to f of r p ok that will complete the proof. Yeah, I am being a little bit informal here there is an \inf in this definition and I am treating it like a \min in the sense that I am thinking there is indeed a number for which there is an indeed a q for which this is q p hypercontractive.

You can make it formal very easily by replacing this r with $r - r + \epsilon$ and then taking limit ϵ going to 0 that is the standard way of handling this \inf and \sup . So, I am not bothered about that. So, we will show this. How do we show this?

(Refer Slide Time: 33:54)

Handwritten mathematical derivation on a digital notepad:

$$\begin{aligned} &\text{We show that } \forall f \in L_p(X_1, X_2), \\ &\quad \|W^2 f\|_p \leq \|f\|_{L_p} \\ \rightarrow &\quad \mathbb{E} \left[\mathbb{E} [f(X_1, X_2) | Y_1, Y_2]^p \right] \\ &= \mathbb{E} \left[\|W^2 f_{Y_1, Y_2}\|_p^p \right] = \mathbb{E} \left[\|W^2 f_{Y_1, Y_2}\|_p^p \right] \\ &\|W^2 f_{Y_1, Y_2}\|_p = \mathbb{E} \left[\mathbb{E} [f(X_1, X_2) | Y_1 = y_1, Y_2 = y_2]^p \right]^{1/p} \end{aligned}$$

So, let us start with this conditional expectation; conditional expectation of f X . So, X here denotes the vector X_1, X_2 maybe it is better to write the whole thing Y_1, Y_2 to the power p whole to the power; whole to the power p let us look at this guy. This can be written as expected value of. So, here the expectation is over X_1 . So, expected value of f of W^2 of f of Y_1, Y_2 to the power p I will write it as $Y_1; Y_1$ of Y_2 to the power p , ok. This is f of Y_1, Y_2 I am W^2 f of Y_1, Y_2 I am using a notation here to simplify this, ok.

And what is this quant this quantity here? So, for each Y_1 this guy here is simply expected value of the p -th norm of this guy this expectation is over Y_1 outside this point. So, the functions are different, but for each of these random variables which is just dependent Y_2 now, this is this expectation ok.

So now, let us focus on this quantity here. So, for this quantity we note that this W_2 of f of Y_1 it is p -th norm this is equal to expected value of. So, Y_1 is fixed in this parts maybe I will use a small Y_1 here. So, then this is X equal to expected value this expectation will be over Y_2 W_2 . The conditional expectation of $f(x)_1, X_2$ given $Y_1 = y_1, Y_2$. This is my random variable whose expectation I am taking over Y_2 . So, this random variable to the power p whole to the power 1 by p , ok.

That is what, that is what this quantity is there is another power we will come back to that later.

(Refer Slide Time: 36:54)

$$\begin{aligned}
 \|W_2 f_{Y_1}\|_p &= \mathbb{E} \left[\mathbb{E} \left[f(X_1, X_2) | Y_1 = y_1, Y_2 \right]^p \right]^{1/p} \\
 &= \mathbb{E} \left[\left(\int \mathbb{E} [f(X_1, X_2) | Y_2] d\mu(x_1, y_2) \right)^p \right]^{1/p} \\
 &= \left\| \int W_2 f_{X_1}(Y_2) d\mu(x_1, y_2) \right\|_p \\
 &\leq \int \|W_2 f_{X_1}(Y_2)\|_p d\mu(x_1, y_2) \\
 &\leq \int \|f_{X_1}(X_2)\|_p d\mu(x_1, y_2)
 \end{aligned}$$

And so, this part here if you look at this inner part here it can be viewed as expected value of. So, let us make some space 1 by p and let us look at this inner expectation here there is something to the power p , ok. So, what is this inner expectation here? So, this inner expectation is over X_1 is over X_1 and X_2 , but let us look at the expectation over X_1 part.

So, that is look like, this is the conditional expectation. So, we will do d of this is notation for convenience, so, that things become explicit. This is the conditional distribution given Y_1, X_1 and it does not depend on Y_2 because of independence. So, we can just do this part and we have expected value of $f(x)_1$ and so, this part is $f(x)_2$. Now, if you look at this particular conditional expectation you have to average over both Y_1, Y_2 , but there is no dependence. So, given Y_2 , X_2 is independent of Y_1 and X_1 and therefore, we can only write it as this.

So, once again the point here is that this inner part can be written as W_2 just the W_2 operator not the superscript to the second channel of $f(x)_1$ that is this conditional expectation evaluated at Y_2 , ok evaluated at Y_2 that is the random variable we have d of $p \times 1$ given y_1 . And, what we are looking at is the, so, this is a function of Y_2 and we are looking at the L_p norm of this random variable of this random variable.

Now, by Minkowski inequality this is less than equal to the L_p norms of individual guys weighed by this thing ok this is Minkowski inequality ok. Because this is just the fact that L_p norm is a norm and this is roughly the triangular inequality for L_p norm which is called Minkowski inequality.

Now, here once we have this part, now we can see that this part here only entails the W_2 function for the second channel. And therefore, by using the by using definition of r and the fact that the second channel is $p \times r$ become hypercontractive this is less than equal to $f(x)_1, X_2$ ok that is the function here given q -th norm of this sorry, r p -th norm of this ok that is what this green expression here can be bounded and we have used now the hypercontractivity for the second coordinate. So, we will substitute this guy in this form here.

(Refer Slide Time: 41:22)

Therefore,

$$\begin{aligned} \mathbb{E} \left[W^2 f(Y_1, Y_2)^p \right] &\leq \mathbb{E} \left[\mathbb{E} \left[\|f_{X_1}(X_2)\|_{\mathcal{H}_p}^p \mid Y_1 \right] \right] \\ &= \mathbb{E} \left[W_1 g(Y_1)^p \right] \\ &\leq \|g(X_1)\|_{\mathcal{H}_p}^p \\ \Rightarrow \|W^2 f\|_p &\leq \|g(X_1)\|_{\mathcal{H}_p} \\ &= \mathbb{E} \left[g(X_1)^{1/p} \right]^{1/p} = \mathbb{E} \left[\|f_{X_1}(X_2)\|_{\mathcal{H}_p}^{1/p} \right]^{1/p} \\ &= \mathbb{E} \left[\mathbb{E} \left[\|f_{X_1}(X_2)\|_{\mathcal{H}_p}^{1/p} \mid X_1 \right] \right]^{1/p} \\ &= \mathbb{E} \left[\|f(X_1, X_2)\|_{\mathcal{H}_p}^{1/p} \right]^{1/p} = \|f\|_{\mathcal{H}_p}. \end{aligned}$$

So, on substituting what we get is; yeah let us maybe zoom out to see both of them at the same time, right. So, we will substitute this we will substitute this in this guy here. So, therefore, expected value of $W^2 f$ of Y_1, Y_2 to the power p is less than equal to expected value.

Now, I will substitute this guy. This is just the conditional expectation given of X_1 given Y_1 and then there is an expectation of Y_1 . So, that is just expectation of Y_1 . And, this whole part of the whole thing to the power p , so that is conditional expectation of $\|f_{X_1}(X_2)\|_{\mathcal{H}_p}$ the p -th norm of this random variable this expectation given Y_1 whole to the power p , ok that is what we have, that is where we have reached ok.

So, we have already used hyper contractivity for the second coordinate and now, it is time to use hypercontractivity for the first coordinate. So, towards that end this is less than equal to. So now, we apply hypercontractivity for the first coordinate you see this is exactly that form. This is exactly equal to expected value of we will think of this function as some we will name this function something.

Let us call this function here as this norm already takes expectation over X_2 and this is only some function of X_1 . So, we will call this h of X_1 ok or maybe g of X_1 . And so, this is just W_1 of g evaluated at Y_1 to the power p that is what this is. And, now using the hypercontractive; using hypercontractivity for the first coordinate this is less than equal to g of X_1 p -th norm to the power p ok. Which implies that W^2 of f p -th norm is

less than equal to $g(X_1)$ r -th norm this is where we use hypercontractivity for the first coordinate.

So, let us expand what this g looks like. So, what is its r -th norm that is expected value over X_1 g of X_1 to the power r and whole to the power $1/r$ and then this is equal what is g to the power r ? Well, this is g of X_1 is this norm this norm to the power r . So, that is $f(x)$ X_2 to the power r .

So, what is this inner norm? This inner norm is nothing but expectation of $f(x)$ this expectation is over X_2 and this one is over X_1 , ok to the power r given X_1 just to explicit here and this is nothing but $f(x)$, X_2 this is just $f(x)$, X_2 to the power r ok. So, essentially we have obtained ok, alright. So, that is what we get, ok.

(Refer Slide Time: 45:54)

$$= \mathbb{E} \left[f(x_1, x_2)^{rp} \right]^{\frac{1}{rp}} = \|f\|_{rp}.$$

Thus, $s_p(X^2; Y^2) \leq \pi.$

→ From here on, we focus only on $n=1$.

(3) Relate hypercontractivity to strong data processing inequality

And, this is exactly this exactly shows that thus s_p of $X_2 Y_2$ is also less than equal to r ok and indeed s_p of X_2, Y_2 must be greater than equal to r , because we can consider functions which are just depending on one coordinates and for them the hypercontractivity coefficient the best one here will depend only on the hypercontract; will equal the hypercontractivity for that coordinate and, so, $s_P X_2, Y_2$ must be greater than equal to max of $s_P X_1, Y_1$ and $s_P X_2 X Y_2$, ok. So, this must be equal to r that is the proof.

Here the proof was a bit tedious because of notation, but I think this is a very delicate proof. For example, here instead of taking the p-th norm if you take p-th norm to the power p and used Jensen's inequality here instead of Minkowski inequality, then you will not get this bound you will get a weaker bound which will not give you this hypercontractivity, ok.

So, long story short. This s p tensorizes just max of each coordinate ok. So, great. So, if you want to show hypercontractivity for product distribution its suffices to show hypercontractivity for individual coordinates. This is something we have been doing for other inequalities as well and it is good to see that it works for hypercontractivity too ok, alright. So, this was the second simplification.

Now, 3rd simplification so, from now from; so, from here on before we go to the third simplification. So, where are we currently? Let me take a big eraser. So, from here on we focus only on the one dimensional case because by tensorization the result for general dimensions follows naturally for product distributions, ok.

Now, 3rd simplification, this is very interesting. This one will relate hypercontractivity to so called strong data processing inequality. So, we have seen tensorization, we have seen this alternative form and now we are seeing that in fact, you can relate hypercontractivity to strong data processing inequality.

(Refer Slide Time: 48:55)

(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

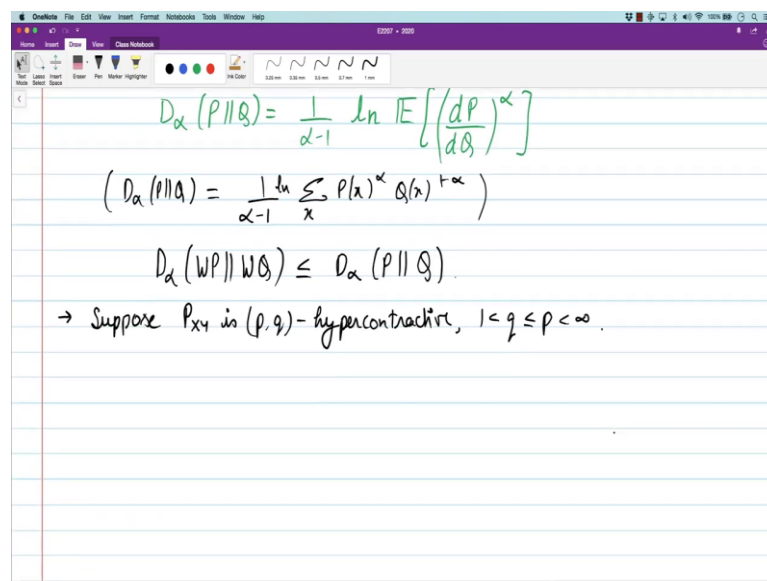
$$D_\alpha(P \| Q) = \frac{1}{\alpha - 1} \ln \mathbb{E} \left[\left(\frac{dP}{dQ} \right)^\alpha \right]$$

So, recall the data processing inequality. If you look at divergence between W_P and W_Q these are the output distributions when the input distributions are P and Q pass to the same channel. This must be less than equal to $D(P||Q)$. This is the Kullback Leibler divergence. This inequality is called the data processing inequality. In fact, the same inequality holds for a larger family of divergences.

So, let us introduce that larger family. The same inequality holds for so called Renyi divergence. So, what is Renyi divergence? So, what is this Renyi divergence? So, Renyi divergence of order α between two distributions P and Q , let us say P that is absolutely continuous with respect to Q .

So, it has a density with respect to Q is given by D_α . So, this α here is something that is greater than 0, but not equal to 1. $D_\alpha(P||Q) = \frac{1}{\alpha-1} \log \int \left(\frac{dP}{dQ} \right)^\alpha dQ$ it is a natural log expected value of. So, P has a density with respect to Q take that density and take its α -th moment.

(Refer Slide Time: 51:44)



Handwritten notes on a digital notepad showing the definition of Renyi divergence and its data processing inequality.

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

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

$$D_\alpha(W_P||W_Q) \leq D_\alpha(P||Q)$$

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

So, this looks yeah. So, this is what Renyi divergence is. And for discrete case you can simplify this. $D_\alpha(P||Q)$ for discrete case can be written as $\frac{1}{\alpha-1} \log \sum_x P(x)^\alpha Q(x)^{1-\alpha}$ let us say distributions are both on X log summation $\sum_x P(x)^\alpha Q(x)^{1-\alpha}$. So, $P(x)^\alpha$ by $Q(x)^{1-\alpha}$ into this expectation with respect to Q , that is, the bigger measure $Q(x)$ into $Q(x)$. So, you get $P(x)^\alpha Q(x)^{1-\alpha}$ ok. You may have seen this

form, but yeah this is Renyi divergence expected value of this $d P Q$ to the power α ok.

So, just like Kullback Leibler divergence you can show this is not a difficult group this Renyi divergence also satisfies data processing inequality. So, $D_\alpha W P W Q$ is less than equal to $D_\alpha P Q$, this is also true, ok. So, that is true. Now, suppose $P \rightarrow Y$ is p , q hypercontractive, where p is where q is less than equal to p .

(Refer Slide Time: 53:30)

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

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

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

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

Note that $Wf(y) = E[f(x) | Y=y]$

Suppose, this is true, then consider this function consider $f X = d P$ by $d Q$ ok then D . So, we have to make an assumption; we have to make consider this such that and assume that the p -th norm of the X is finite, ok. So, what do we have? By hypercontractivities into this $p \rightarrow q$ hypercontractive W of f p -th norm is less than equal to f q -th norm, ok. So, what is this guy here? Well, let us take log on both sides. So, that we can related to divergences, this is less than equal to log of f of q .

And, let us make this Renyi divergence. So, which is the same as saying that so, this right side here this guy here is exactly equal to $q - 1$ into the q -th Renyi divergence between P and Q . But what is this left side? We will see what this left side is, ok.

So, note that this W of f of y is equal to conditional expectation of $f X$ given $Y = y$.

(Refer Slide Time: 56:40)

Note that $W(y) = E[f(x) | Y=y]$
 $= E\left[\frac{dP}{dQ} | Y=y\right]$

Claim: $E\left[\frac{dP}{dQ} | Y=y\right] = \frac{dWP}{dWQ}(y)$

Proof: (When X, Y are discrete)

$$E_Q\left[\frac{dP}{dQ}(x) | Y=y\right] = \sum_x \frac{P(x)}{Q(x)} \frac{Q(x) W(y|x)}{E_Q[W(y)|X]}$$

So, what is this guy? This is the conditional expectation of dP by dQ given Y equal to y . So, what is this quantity on the right? We claim that this quantity on the right expected value of dP by dQ which is the function of X actually given $Y = y$ is exactly equal to it is almost surely equal to the density of the measure WP with respect to WQ .

Since P has a density with respect to Q the output distribution when you pass through channel W also has a density with input distribution P also has a density with respect to this WQ , if that density is exactly this conditional distribution ok. Now, this proof is a bit technical, but we will just show it for the discrete case where it is easy to see.

So, for when x, y when random variables X, Y are discrete. In that case expected value of dP by dQ as a function of X given $Y = y = \text{summation over } x$; by the way this expectation is with respect to Q x this ratio is $P(x)$ by $Q(x)$ the density is just $P(x)$ by $Q(x)$. And you want to look at this probability of; you want to look at probability of x given y that is probability of x, y it is $Q(x)$ into $W(y|x)$ given x by probability of y .

What is probability of y ? Probability of y is simply summation over x $Q(x) W(y|x)$ given x . Let us write slightly better this is equal to $P(x)$ by $Q(x)$ that is this function times $Q(x) W(y|x)$ given x divided by the marginal which is this guy ok that is what the density is.

(Refer Slide Time: 59:20)

Handwritten notes on a digital notepad:

$$= \frac{E_P[W(y|x)]}{E_Q[W(y|x)]} = \frac{dW_P(y)}{dW_Q(y)}$$

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

$$D_P(W_P \parallel W_Q) \leq \left(\frac{1 - 1/2}{1 - 1/p} \right) D_Q(P \parallel Q) \rightarrow < 1 \text{ if } q < p.$$

Thus, if we show that $\Delta_P(X; Y) < 1$, then.

But, if you look at this guy here, this is the probability of Y under W and if you look at this guy here this Q x canceled. So, that is probability Y under P by probability of Y under Q this is essentially this is exactly equal to W P, ok.

So, indeed this conditional expectation is equal to this density and this can be proved in general under some conditions on the channel some technical conditions of channel. We will not get into all that. In the way I need; the way I know how to show this proof is by defining the channel having what is called regular conditional densities, but maybe there are more general proofs available ok.

So, very nice so, even this W f is actually the density of the output measures $P W_P$ with W_Q . And therefore, just as we could write this guy as a divergence we can write this guy also as a divergence. So, therefore, P_{XY} is (p, q) hypercontractive implies $D_P(W_P \parallel W_Q)$ is less than $1 - 1/p$. So, we will take this guy this side. Actually there is this $1 - 1/q$ power here which must be brought down. So, it is here ok because we do not need this $1 - 1/q$ power and $D_Q(P \parallel Q)$ and so, this is less than equal to so, this is $1 - 1/p$ by $1 - 1/q$ $D_Q(P \parallel Q)$.

So, the only difference between this and the hypercontractivity statement is that hypercontractivity it holds for all function and this one only holds for functions which can be expressed as these densities, which means they must be normalized to 1. But note that hypercontractivity inequality is homogeneous. If you divide both sides; if you divide

if you replace a function f with c times f , the inequality continues to hold. Therefore, we can always normalize.

So, this is actually if and only if it is p q hypercontractive; I will just replace it with if and only if condition. If and only if for all Q that have density with respect to P such that $D_p(P||Q)$ is finite $D_W(P||Q)$ is less than equal to $1 - 1/q$ by $1 - 1/p$ into $D_q(P||Q)$ ok. This is the same as; this is exactly equivalent to hypercontractivity that is what we are claiming, ok. This is not so clean, ok. So, this is hypercontractivity.

So, what is important here is that if you look at this constant here this part here this part here, this is less than 1 if $1/q$ is greater than $1/p$; if q is less than p ok. So, if you can show that because if we show that $s_p(X; Y)$ is less than 1, then the data processing inequality holds with this constant less than 1. By the way if q is less than p this $D_q(P||Q)$ is actually an increasing function of non-decreasing function of q or order α .

(Refer Slide Time: 64:20)

Handwritten derivation on a digital notepad:

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

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 therefore, thus $D_p(WP||WQ)$ is less than equal to C_q times D_q this, and which is less than because see this is less than 1, so this guy is also less than 1. This is less than D_q which is less than equal to $D_p(P||Q)$ ok; or alternatively D_p ok.

And, therefore, that this is the data processing inequality we saw earlier, except that this constant can be made less than 1. And when is this constant the smallest? This constant is the smallest when this guy is the smallest, that is the best we can do, ok.

Note that s_P is not exactly this guy s_p is related to q by p ; this is $q - 1$ this is slightly different, but related quantity ok. And, in fact, there is a lot of work on connection between the best data processing constant which is obtained by taking sup over this C_q here and which is obtained by minimizing over the C_q here and this $s_{P_X; Y}$. Mostly this guy is attained when p goes to infinity ok, and that quantity coincides with the best data processing constant that is usually the case. But here is an exact equivalence.

So, at $P_X Y$ is P_Q hypercontractive if this Renyi divergence of order p satisfies strong this particular strong data processing inequality. The Renyi divergence of order p between W_P and W_Q is less than equal to $1 - 1/q$ by $1 - 1/p$ $D_q(P_Q)$ ok. So, this is another alternative form of hypercontractivity.

So, in conclusion. We have seen that hypercontractivity is actually equivalent to strong data processing inequality for Renyi divergence. We have also seen two different forms of hypercontractivity. We have seen tensorization which says that it suffices to show hypercontractivity for one dimension and then you will automatically get it for any arbitrary dimension.

In improving this equivalent of equivalence of hypercontractivity and strong data processing inequality for Renyi divergence. Along the way we remark that it suffices to show hypercontractivity for functions which are normalized in any norm. In fact, if that inequality holds for f it must also hold for $C f$. So, this is all I wanted to show say about hypercontractually different form.