

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

Lecture - 17
Establishing Marton's Conditional Transportation Cost Inequality

(Refer Slide Time: 00:21)

Lecture 16
 24 November 2020 15:51

Establishing
MARTON'S CONDITIONAL TRANSPORTATION COST INEQUALITY

$$\min_{(X,Y) \in \mathcal{P}(P^n, Q)} \mathbb{E} \left[\sum_{i=1}^n P[X_i \neq Y_i | X]^2 + P[X_i \neq Y_i | Y]^2 \right] \leq 2D(Q \| P^n) \quad (\star)$$

$\forall Q \ll P^n \equiv P_1 \times \dots \times P_n$

PROOF:

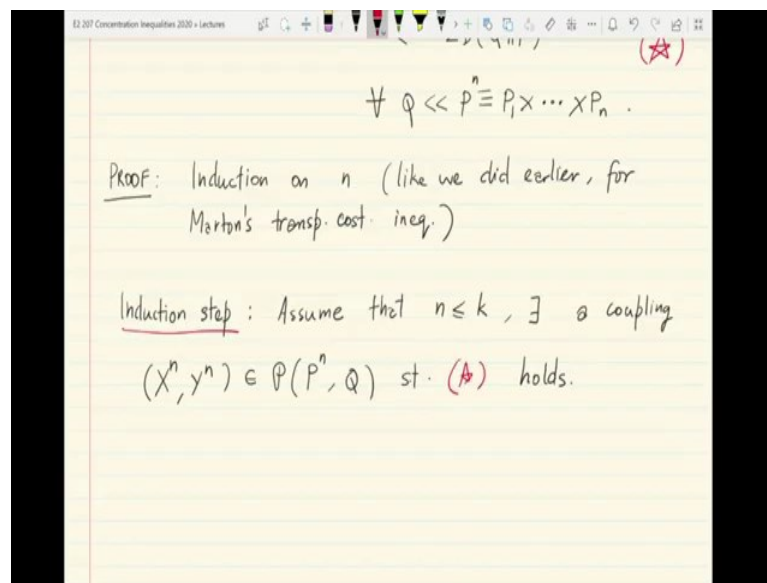
Hi, all. This lecture is devoted to Establishing Marton's Conditional Transportation Cost Inequality which is the following transportation cost inequality that we found last time was very useful and allowing us to prove concentration inequalities for functions which satisfy properties weaker than bounded differences or the MacDiarmid setting.

So, we did that in the last lecture and all that was remaining was to prove this kind of transportation cost inequality. So, what is Marton's condition transportation cost inequality? Just to recap it basically says that if you are given two probability measures one is P which is a product measure and one is Q which is some absolutely continuous measure with respect to P .

Then there exists a coupling between the P and the Q measures let us say for random variable sequences X and Y of length n such that the expected value of the sum of the conditional discrepancies squared is bounded by basically the divergence between these measures the $k l$ divergence between the Q and the P^n ok.

So, that was the content of Marton's conditional transportation cost inequality. We have already proven a similar related inequality which is Marton's transportation cost inequality basically where there was no conditioning on X or on Y as we were able to establish there that the inequality still holds that you can bound the left hand side by a quantity base in the divergence and we will follow a very similar proof template for doing the same exercise here.

(Refer Slide Time: 02:13)

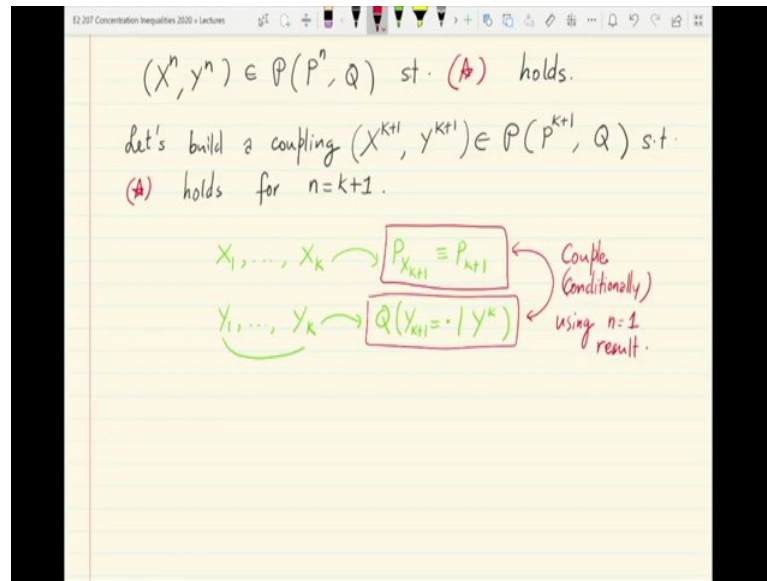


So, the proof is going to be as before induction by induction on the number of variables n . So, just as we did earlier ok. So, for Marton's transportation cost inequality. So, let us so, there are two arguments in an induction proof. So, one is basically the induction step and one is how you prove the base case. Let us look at the induction step before we go over to the base case.

So, what is the induction step here? So, the induction step consists of assuming that for all n less than or $=$ let us say some arbitrary integer k there exists a coupling between X^n and X_1 through n and Y_1 through n ok which have individual marginals P_1 through n and Q such that the inequality $(*)$ holds where $(*)$ is going to be denoted. So, this is $(*)$ here. So, let me make it red ok.

So, let us assume that this statement holds for all n at most some integer k what we will do is to show that the same inequality holds for $n = k + 1$ by building a new coupling.

(Refer Slide Time: 04:01)



Let us build a coupling. Let us now a couple using this one length so, the sequences of length one extra. So, X_{K+1} . Let us couple X_{K+1} to Y_{K+1} with desired marginal's P_{K+1} and Q such that. So, note that when we say P_{K+1} and Q this it is just sufficient to consider Q or that the marginal distribution of Q in the first $K+1$ symbols such that star holds for $n = K+1$ thereby extending the induction hypothesis by 1 unit.

So, it is that the idea is exactly the same as what we did to extend the induction hypothesis for Marton's transportation cost inequality. So, what is the idea here? So, in words what you have is that you have let us say you have a coupling in your hand which can couple X_1 through X_K to Y_1 through Y_K , ok. So, you basically have a coupling between these two and what we would like to do is to extend this to a coupling between $K+1$ length sequences on the top and the bottom.

So, what you do is that having the Y sequence condition on the first K Y 's this basically gives you a conditional distribution for the $K+1$ st place. So, $Q(\cdot | y^K)$ given so, let us see Q of the $K+1$ st entry being something given the first K entries and you also have a target marginal distribution for the top side the X_{K+1} which is just the distribution of X_{K+1} which we also called P_{K+1} .

So, having conditioned on the realization of y_1 through K what you do is you take these this measure the conditional distribution of Y_{K+1} given the first K Y 's. And couple it to the X target distribution which is P_{K+1} . So, couple these couple

conditionally because we are conditioning on the Y_K here Y_1 through K here using the $n = 1$ result.

So, recall that we have assumed induction hypothesis to hold for all n less than $= K$. So, that means, it also holds for $n = 1$ and basically that allows us to find a coupling or a use a coupling between P_{K+1} and this conditional distribution of Y_{K+1} given Y_K to generate this new pair $X_{K+1} Y_{K+1}$ which we just append to the end ok.

And, we will be able to argue that the entire joint thing between X_1 to X_{K+1} and Y_1 through Y_{K+1} is the desired coupling on $K + 1$ length sequences with the designed to have the appropriate marginal's that is P on the top part on the X -side and Q on the Y -side ok.

(Refer Slide Time: 07:51)

Begin by using the $n=1$ construction conditionally:

$$\forall y^K \exists (X_{K+1}, Y_{K+1}) \in \mathcal{P}(P_{K+1}, Q_{Y_{K+1}|Y^K=y^K}) \text{ s.t.}$$

$$\mathbb{E} \left[\underbrace{\mathbb{P}[X_{K+1} \neq Y_{K+1} | Y_{K+1}, Y^K=y^K]}_{(3)}^2 + \underbrace{\mathbb{P}[X_{K+1} \neq Y_{K+1} | X_{K+1}, Y^K=y^K]}_{(2)} \right] \leq 2D(\underbrace{Q_{Y_{K+1}|Y^K=y^K}}_{(1)} \| P_{K+1}).$$

Let us set

$$P_{K+1} := P$$

So, formally what we do is we begin by using the $n = 1$ construction conditionally on so, conditionally on basically the previous K realization of the y 's. So, this just means for all realizations y_1 through y_K .

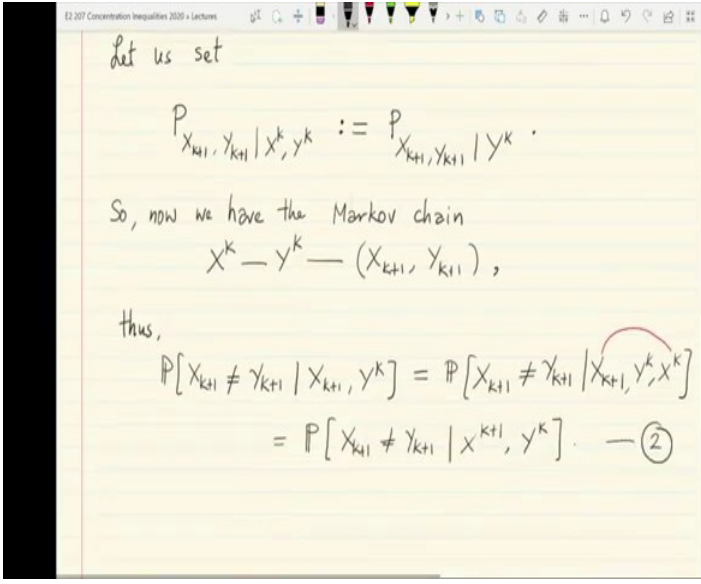
So, what is this construction that we are guaranteed to that is guaranteed to exist? There exists a coupling between X_{K+1} and Y_{K+1} distributed that couples the distributions P_{K+1} or $P_{X_{K+1}}$ with the conditional distribution of Y_{K+1} given the first K Y 's or y_1 through y_K ok.

So, also we know that this $n = 1$ conditional coupling also satisfies the inequality for $n = 1$. So, that the inequality basically reads expected value of the probability. So, this is the $n = 1$ form. So, the variables being measured for discrepancy are just X_{K+1} and Y_{K+1} given. So, this is all conditional on conditioned on Y_1 through K . So, we should always condition on Y_1 through K .

Firstly, there is the conditioning on one of these variables itself followed by the background conditioning Y_1 through $K = \text{small } y_1 \text{ through } K \text{ whole square} + \text{the same thing where the conditioning is on the } X \text{ variable}$. So, X_{K+1} and we will add to it the background conditioning $Y_K = y_K$ the whole square is upper bounded by twice this is the induction hypothesis for $n = 1$ twice the divergence between $Q_{Y_{K+1}} \text{ given } Y_K = \text{small } y_1 \text{ through } K$ with P_{K+1} , ok.

This is what the induction hypothesis guarantees us. This let us call this statement a statement 1. And, now how do we build this coupling? So, now, we are going to define the formal extension or a new coupling between X_1 through $K+1$ and Y_1 through $K+1$.

(Refer Slide Time: 10:37)



Let us set

$$P_{X_{k+1}, Y_{k+1} | X^k, Y^k} := P_{X_{k+1}, Y_{k+1} | Y^k}.$$

So, now we have the Markov chain

$$X^k - Y^k - (X_{k+1}, Y_{k+1}),$$

thus,

$$\begin{aligned} \mathbb{P}[X_{k+1} \neq Y_{k+1} | X_{k+1}, Y^k] &= \mathbb{P}[X_{k+1} \neq Y_{k+1} | X_{k+1}, Y^k, X^k] \\ &= \mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}, Y^k]. \quad \text{--- (2)} \end{aligned}$$

So, let us say $P_{X_{K+1}}$. So, the joint distribution between the $K+1$ first X and the $K+1$ first y condition on $X_K Y_K$ the previous scale and segment to be $=$; let us define this to be $=$ generating using this guaranteed coupling for the last two variables of the X and the Y sequences $X_{K+1} Y_{K+1}$ conditioned on Y_K alone, ok.

So, what this means is basically that it is exactly executes this diagram you take the first K y 's from then that gives you a conditional distribution on the $K + 1$ first Y and just use a coupling that can couple this conditional distribution for y_{K+1} with the standard distribution P_{K+1} for the X_{K+1} ok.

So, this is how the inductively defined so, this basically defines a probability distribution on X_1 through $K + 1$ and Y_1 through $K + 1$, ok. So, now we have basically by doing this we have ensured a Markov chain we have defined basically a Markov chain that say that if you have X_K from it you can generate Y_K just because you have a joint distribution between X_K and Y_K . This is using the existing coupling.

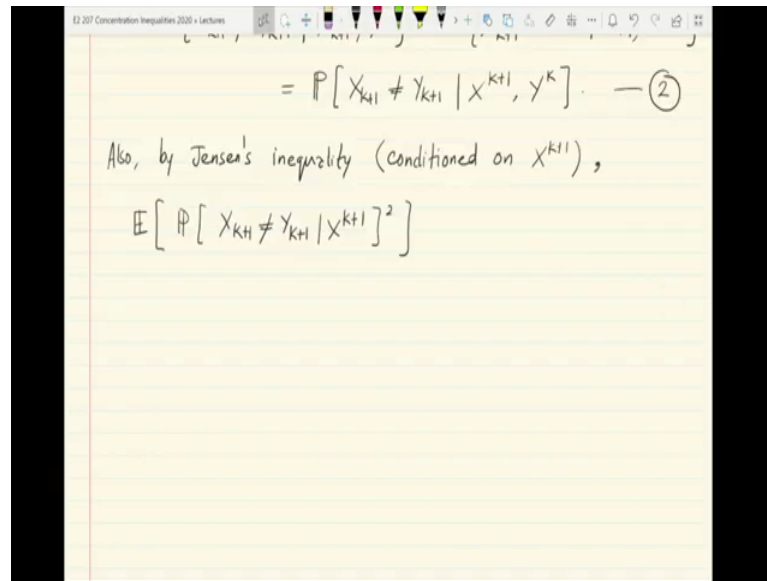
The induction hypothesis coupling for n equals K and from there Y_K alone is enough to generate the next pair $X_{K+1} Y_{K+1}$ ok. And, thus as a consequence of this we can write that the probability $X_{K+1} \neq Y_{K+1}$ given X_{K+1} and Y_K . So, recall a that this is just this term here ok this term is being analyzed conditioning is on X_{K+1} and the background conditioning on Y_K .

This is just = so recall what happens here if you map it to the Markov chain diagram here it is the probability of some event involving the rightmost Markov chain element conditioned on a part of it itself which is X_{K+1} and the Y_K . So, we can add by the Markov property X_K for free X_1 through K for free ok.

So, what the Markov chain picture means is that for free because of the Markov property we can add $X_{K+1} Y_K$ and we can add for free X_1 through K . And, now if we just combine these two elements it is just the entire vector X_1 through K given X_1 through $K + 1$ with Y_1 through K . So, let us record this as statement 2.

So, we are just analyzing the terms in this $n = 1$ inequality ok. Now, let us come to trying to prove, so we can manipulate this inequality further if we can find a lower bound for the first time your first conditional probability here and so, that is done by using Jensen's inequality.

(Refer Slide Time: 14:39)

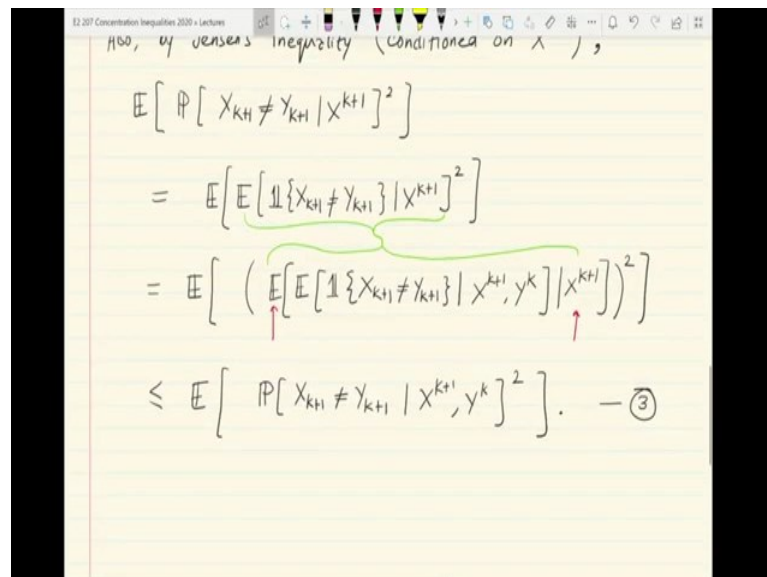


$$= \mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}, Y^k] \quad \text{--- (2)}$$
 Also, by Jensen's inequality (conditioned on X^{k+1}),

$$\mathbb{E} \left[\mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}]^2 \right]$$

So, notice that so, let us say also by Jensen's inequality so, in fact, expected value of the probability of $X_{k+1} \neq Y_{k+1}$ given X^{k+1} the whole square. So, let us analyze this term an expression of this form. So, in fact, let us write the Jensen inequality condition on in its conditional form X^{k+1} .

(Refer Slide Time: 15:27)



Also, by Jensen's inequality (conditioned on X^{k+1}),

$$\mathbb{E} \left[\mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}]^2 \right]$$

$$= \mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{X_{k+1} \neq Y_{k+1}\}} | X^{k+1} \right]^2 \right]$$

$$= \mathbb{E} \left[\left(\mathbb{E} \left[\mathbb{E} \left[\mathbb{1}_{\{X_{k+1} \neq Y_{k+1}\}} | X^{k+1}, Y^k \right] | X^{k+1} \right] \right)^2 \right]$$

$$\leq \mathbb{E} \left[\mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}, Y^k]^2 \right] \quad \text{--- (3)}$$

So, this is just the same as the expected value; probability it is just expectation of an indicator and one can write. So, let us start let us write the step by step. The inner probability which is being squared is just the expected value of an indicator event

conditioned on X_{K+1} the whole square ok whose expectation is being taken outside ok.

Now, what we can do is we can write the inner expectation which is the inner conditional expectation as a further conditional expectation where in the first level you take the expected value of $X_{K+1} \neq Y_{K+1}$ conditioned on two things: one is X_{K+1} as usual to it we add an additional Y_K ok and imagine taking the conditional expectation of this with respect to X_{K+1} .

So, by the law of iterated expectation this double expectation is the same as this expectation ok. And so, all we are doing is we are writing the inner conditional expectation as a further expectation on an extra quantity which is Y_K and then we are squaring it ok.

Now, let us apply Jensen's inequality to this conditional expectation with respect to X_{K+1} which is being square ok. So, when we do that so, it is the square of an expectation because the square the square is a convex function we can basically bring the conditional expectation outside the square. Sorry, we can bring the conditional expectation outside the square, yes. So, less than = the expected value the outer E remains the same and so, imagine bringing the conditional expectation outside the square.

So, what you have is the thing that comes out combines with the E outside ok and then what you what is left inside is the probability that $X_{K+1} \neq Y_{K+1}$ given X_{K+1} comma Y_K the innermost probability which gets squared ok. So, this is basically Jensen's inequality for the square function that is in it is conditional form that we apply. So, let us put 2 and 3. So, if you take 2 2 gave you an equivalent form let us put 2 and 3 into 1.

(Refer Slide Time: 18:25)

$\forall (X_{k+1}, Y_{k+1}) \in \mathcal{P}(P_{k+1}, Q_{Y_{k+1}} | Y^k) \text{ s.t.}$

$$\mathbb{E} \left[\mathbb{P}[X_{k+1} \neq Y_{k+1} | X_{k+1}, Y^k = y^k]^2 + \mathbb{P}[X_{k+1} \neq Y_{k+1} | X_{k+1}, Y^k = y^k]^2 \right] \leq 2D(Q_{X_{k+1}|Y^k=y^k} \| P_{k+1})$$
②, ③ — ①
 Let us set

$$P_{X_{k+1}, Y_{k+1}}^{X^k, Y^k} := P_{X_{k+1}, Y_{k+1}} | Y^k$$

 So, now we have the Markov chain
 $X^k - Y^k - (X_{k+1}, Y_{k+1})$
 Thus,

So, let us go back to equation 1 here. If you look at equation 1 above what equations 2 and 3 have helped us do is to basically give a lower bound for the second term on the left hand side. So, we can now go back and using 2 and 3 into 1 we can write the following.

(Refer Slide Time: 18:45)

$$\leq \mathbb{E} \left[\mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}, Y^k]^2 \right]. \text{ --- ③}$$

 Putting ② & ③ into ①, we get

$$\mathbb{E} \left[\mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}]^2 + \mathbb{P}[X_{k+1} \neq Y_{k+1} | Y^{k+1}]^2 \right] \leq 2D(Q_{X_{k+1}|Y^k} \| P_{X_{k+1}} | Q_{Y^k}). \text{ --- ④}$$

So, let us say putting 2 and 3 into 1 get. So, we get this following simple form expected value of P the conditional probability of discrepancy at the K + first position given the entire expected + P X K + 1 not = Y K + 1 the same conditional probability given the Y vector up to K + 1 is just at most the divergence between the conditional measure Q K +

1 given Y_1 through K with respect to $P_{X_{K+1}}$ or P_{K+1} it is the same thing conditioned on Q_{Y_K} , ok.

So, this notation is just saying that right so, what we have is basically that. So, what we do to equation 1 to inequality 1 is just that we put in these expressions and then since the left hand side now and then we take expectation over Y_K . You take just take a expectation over Y_K to get this ok. Now, right, so we get this equation for and so, now, what we want. So, we can now go ahead to finish the proofs of the induction step the extending the induction hypothesis.

(Refer Slide Time: 20:29)

At the top level, we have

$$\sum_{i=1}^{K+1} \mathbb{E} \left[\underbrace{\mathbb{P}[X_i \neq Y_i | X^{K+1}]^2}_{(\forall i \leq k)} + \mathbb{P}[X_i \neq Y_i | Y^{K+1}]^2 \right]$$

$$= \mathbb{P}[X_i \neq Y_i | X^k, X_{K+1}]$$

At the top level what we wanted to do to extend the induction hypothesis was to prove a bound on the following expression sum over $i = 1$ to $K+1$ expected value of $\mathbb{P}[X_i \neq Y_i | X^{K+1}]^2 + \mathbb{P}[X_i \neq Y_i | Y^{K+1}]^2$ whole square ok.

So, we wanted to basically bound this using the KL divergence, but we can already see the following from. So, let us now look at it from the point of view of the coupling we have constructed between X_1 through $K+1$ and Y_1 through $K+1$ we have that for all i . So, let us look at i at most K ok.

So, let us look at the first K terms of the sum any any term in the first K terms of the sum. So, for any such i that is at most K we have. So, we have that this expression is just

= probability $X_i \neq Y_i$ given X_K and X_{K+1} this is by definition X this is everything up to K and then the $K+1$ (Refer Time: 22:00).

But, by our coupling construction basically it does not matter whether $K+1$ is there or not ok because X_K, X_{K+1} is really independently generated of X_K, Y_K, X_{K+1} just follows its own distribution P_{K+1} . So, once X_K is there the conditional distribution of X_i, Y_i just depends on X_K does not depend on X_{K+1} ; just because of the nature of our particular coupling here ok.

(Refer Slide Time: 22:35)

$$\begin{aligned}
 &= \sum_{i=1}^k \mathbb{E} [\mathbb{P}[X_i \neq Y_i | X^k]^2 + \mathbb{P}[X_i \neq Y_i | Y^k]^2] \\
 &\quad + \mathbb{E} [\mathbb{P}[X_{k+1} \neq Y_{k+1} | X^{k+1}]^2 + \mathbb{P}[X_{k+1} \neq Y_{k+1} | Y^{k+1}]^2] \\
 &\quad \text{(induction hyp. for } n=k+1 \text{)} \\
 &\leq
 \end{aligned}$$

So, since X_{K+1} is independent of X_K, Y_K ok that is the reason. The other term here for any the conditioning or Y_{K+1} so, this is just = so, for any i at most K this is just = the probability that $X_i \neq Y_i$ ok given. So, recall that Y_{K+1} consists of Y_K and the Y all the way up to K and then the last Y_{K+1} .

So, again we can drop Y the last Y_K, Y_{K+1} first element of Y since we have the Markov chain Y_K . So, Y_{K+1} just depends on Y_K and it does not depend on X_K ok. So, Y_{K+1} is irrelevant to find the conditional probability of $X_i \neq Y_i$ given all the first $K+1$ elements ok because of this Markov chain structure here.

So, what we have finally, is that this is = if you split the sum into two parts – one containing all terms up to K and one containing the $K+1$ first term. So, in the first term up to $i = 1$ through K we will use these equivalences and we can essentially this allows us to

replace K X all the way up to $K + 1$ with X all the way only up to $K + P$ X i not $= Y$ i given Y all the way up to K only.

And, then we have the last term which is the term initialize that $K + 1$. So, that is P X $K + 1$ not $= Y$ $K + 1$ given X all the way up to $K + 1$ the whole square + the same thing same event conditioned on Y $K + 1$ whole square.

And, now we are in a position to finally, bond this we will use the induction hypothesis on the first sum ok. Note that we have massage the first sum into exactly the form that the induction hypothesis for $n = K$ guarantees us and we will use inequality for only $K + 1$ part ok on the remaining part.

(Refer Slide Time: 25:33)

$$\begin{aligned}
 & \leq 2 \left\{ D(Q_{Y^K} \| P_{X^K}) + D(Q_{Y^{K+1} | Y^K} \| P_{X^{K+1} | X^K} | Q_{Y^K}) \right\} \\
 & \stackrel{\text{(KL chain rule)}}{=} 2 D(Q_{Y^{K+1}} \| P_{X^{K+1}})
 \end{aligned}$$

So, let me say induction hypothesis for $n = K +$ inequality 4. This is what gives us the authority to write this as upper bounded by D Q Y K with P X $K +$ by 4 it is a right hand side of 4 is basically Q Y $K + 1$ given Y all the way up to K with P X $K + 1$ which is itself the conditional distribution of X $K + 1$ with respect to X K given Q Y K and by the chain rule of KL divergence this is exactly what we need which is $2 D$ Q by $K + 1$ with P X $K + 1$ this is by the KL divergence chain rule ok. So, that basically completes the induction argument.

(Refer Slide Time: 26:39)

Base case : $n = 1$

Define $d_2(Q, P) := \sqrt{\sum_x \frac{(P(x) - Q(x))_+^2}{P(x)}}$ ($x_+ = \max(0, x)$)

It only remains to now show the base case of Marton's conditional transportation cost inequality which is $n = 1$. So, we just have to show the inequality holding for the $n = 1$ case. So, for this let us introduce some new notation. Let us introduce this new notion of a distance between two probability measures which we will call $d_2(Q, P)$ distance or divergence between probability measures as the sum over X of $P(x) - Q(x)$.

So, we are doing this for discrete distributions only, but it can easily be extended by the Radon-Nikodym derivative mechanism for arbitrary distributions ok all under the square root ok where the $+$ notation just means that. So, x_+ , for instance, is just taking max of x with 0 which just gives you the positive part of x .

(Refer Slide Time: 27:39)

$$= \left\| \left(1 - \frac{dQ}{dP}\right)_+ \right\|_2.$$

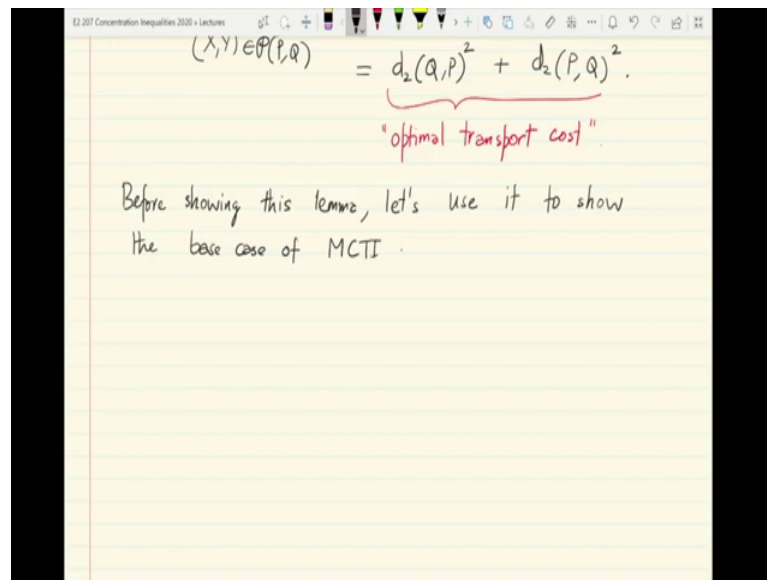
LEMMA: $\min_{(X,Y) \in \mathcal{P}(P,Q)} \mathbb{E} [P[X \neq Y | X]^2 + P[X \neq Y | Y]^2]$

$$= d_2(Q, P)^2 + d_2(P, Q)^2.$$

You can also write it in terms of the two norm over of random variables as the is the L 2 norm of the Radon – Nikodym and derivative of 1 - dQ by dP. So, the + sign here and the two norm as defined in the standard. So, it is the two norm of any random variable is the square root of the expected value of its square, this is ok. Square root of its second moment.

So, here is a lemma which is very similar to the sort of a transportation cost lemma for the total variation distance. So, we have that given any two random variables any two distributions P and Q there exists a coupling X, Y of P and Q such that the least value of $\mathbb{E} P[X \neq Y | X]^2 + P[X \neq Y | Y]^2$ equals $d_2(Q, P)^2 + d_2(P, Q)^2$ ok, this is actually an optima.

(Refer Slide Time: 29:05)

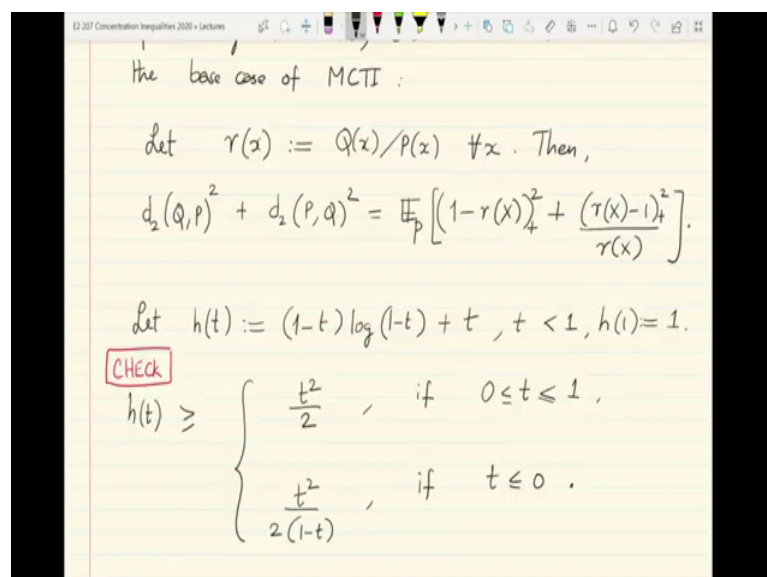


$(X, Y) \in \mathcal{P}(P, Q) \quad = \quad d_2(Q, P)^2 + d_2(P, Q)^2.$
"optimal transport cost"

Before showing this lemma, let's use it to show the base case of MCTI.

So, it basically says that $d_2(Q, P)^2 + d_2(P, Q)^2$ is in some sense an optimal transportation cost. So, it is the solution to a transportation cost problem ok because the left hand side is a variational problem over all couplings of P and Q . Now, before proving this lemma so, let us say before showing this lemma let us use it to. So, let us use it to show the base case of Marton's conditional transport conditional transportation inequality ok.

(Refer Slide Time: 30:05)



the base case of MCTI :

let $r(x) := Q(x)/P(x) \quad \forall x$. Then,

$$d_2(Q, P)^2 + d_2(P, Q)^2 = \mathbb{E}_P \left[\frac{(1-r(x))_+^2}{r(x)} + \frac{(r(x)-1)_+^2}{r(x)} \right].$$

let $h(t) := (1-t) \log(1-t) + t, \quad t < 1, \quad h(1) = 1.$

CHECK

$$h(t) \geq \begin{cases} \frac{t^2}{2}, & \text{if } 0 \leq t \leq 1, \\ \frac{t^2}{2(1-t)}, & \text{if } t \leq 0. \end{cases}$$

So, let us define again the Radon-Nikodym derivative between Q and P as r so let r of x for every discrete outcome x we defined as the ratio Q of x by P of x for all x . Then, we can write this equality that says that $d_2(Q, P)^2 + d_2(P, Q)^2$ is nothing, but the expected value under the P measure of this quantity $1 - rX + \text{square} + rX - 1 + \text{square}$ divided by rX . This is easy to check.

And, let h of t denote the function $1 - t \log 1 - t + t$ for t less than 1 and h of 1 is defined to be 1. So, this is the familiar h function that we have seen before while analyzing the tail of the Poisson distribution it is connected intimately to moment generating functions of the Poisson random variable.

So, one can prove this easy inequality that says that h of t is lower bounded by t^2 if t lies between 0 and 1 and t^2 by $2(1 - t)$ if t is negative. So, this is something that you can easily check and these (Refer Time: 32:04) will come in very handy.

By the way h is also the same function that was used to in the manipulations to derive the Bennett and Bernstein inequalities. So, moving on let us, so with these lower bounds for h we can actually use these lower bounds for h to bound each of these terms in the right hand side of the $tQ d_2(Q, P)^2 + d_2(P, Q)^2$ expression.

(Refer Slide Time: 32:31)

$$\left(\frac{t}{2(1-t)} \right)^2$$

Thus, $(1-r(x))_+^2 \leq 2h((1-r(x))_+)$, and

$$\frac{(\tau(x)-1)_+^2}{\tau(x)} \leq 2h(-(\tau(x)-1)_+), \text{ which yields}$$

$$d_2(Q, P)^2 + d_2(P, Q)^2 \leq 2 \mathbb{E}_P \left[h((1-r(x))_+) + h(-(\tau(x)-1)_+) \right]$$

$$= 2 \sum_x p(x) \left[\tau(x) \log r(x) + 1 - r(x) \right]$$

$$= 2D(Q \parallel P).$$

So, thus with this we can write that $1 - r$ of x for any x $1 - r x + \text{square}$ just at most $2 * h$ of $1 - r x + \text{ok}$.

Note that $1 - r x +$ is always bounded between 0 and 1 and on the other hand, we have $r x - 1 + \text{square}$ by $r x$ is at most twice h of the negative of $r x - 1 + \text{ok}$ which in turn yields that if you go back to $d^2(Q, P) + d^2(P, Q)$ square and plugging in the upper bounds on the right hand side we have $d^2(Q, P) + d^2(P, Q)$ is at most twice the expectation under P of h of $1 - r X + + h$ of the negative of $r X - 1 +$.

And, this if you write out explicitly turns out to be exactly the sum over all x of $t x$ the expectation under P of $r x \log r x + 1 - r x$, ok . So, the second term just becomes 0 because it is the sum of $P x$ which is 1 - the sum of $Q x$ which is again 1. So, $1 - 1$ cancels out and what remains is just twice the KL divergence between Q and P ok .

So, what we have shown is that $d^2(Q, P) + d^2(P, Q)$ assuming this lemma. So, this lemma allows you to relate it is a transportation cost lemma that allows you to basically say that there exists a coupling for which this equals $d^2(Q, P) + d^2(P, Q)$ square and what we have shown is the right hand side is upper bounded by twice $D(Q, P)$. So, that completes the proof Marton's transportation conditional transportation cost inequality modulo the proof of the lemma.

(Refer Slide Time: 35:09)

PROOF OF LEMMA:

Given $(X, Y) \in \mathcal{P}(P, Q)$,

$$P[X=Y | X=x] = \frac{P[X=x, Y=x]}{P[X=x]} \leq \min \left\{ 1, \frac{Q(x)}{P(x)} \right\}.$$

$$\therefore E \left[P[X \neq Y | X]^2 \right] \geq E_P \left[\left(1 - \frac{Q(x)}{P(x)} \right)_+^2 \right] = d_2(Q, P)^2.$$

Similarly,

$$E \left[P[X \neq Y | Y]^2 \right] \geq d_2(P, Q)^2.$$

So, let us now do the proof of the lemma to wrap up this entire discussion. So, what does the lemma ask us to show? The lemma basically says find a coupling between X and Y . So, with that in some sense the probability the conditional probabilities of discrepancy between X and Y are kept to the least possible amount ok .

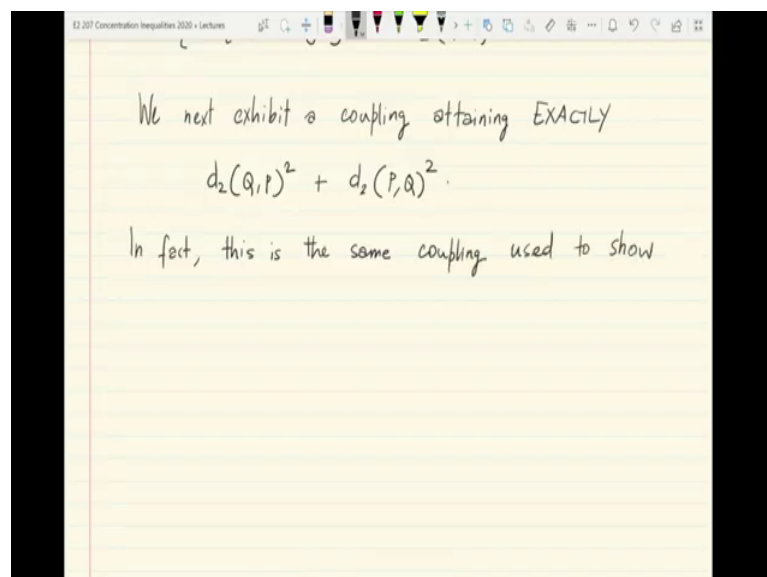
So, given so, let us take any coupling of X and Y following the marginals P and Q given marginals P and Q we have that the probability of $X = Y$ given so, not discrepancy, but equality. So, probability that the joint pair lines on the diagonal given $X = \text{small } x$ is by definition probability $X = x, Y = x$ divided by probability $X = x$.

And, you can from the top numerator you can factor our probability $Y = x$ into the conditional probability of X given $Y = x$ to show that this is bounded above. So, because this is a conditional probability it has to be bounded by 1 above and it is also bounded by $Q(x)$ by $P(x)$ which is what you get when you factor in factor out the numerator as $P(Y)$ into $P(X)$ given by ok.

So, therefore, the expected value of $E(X \neq Y)$ given X the whole square which is one part of the lemma is left hand side is lower bounded by the expected value $1 - Q(x)$ by $P(x)$ + the whole square just by algebra and this we know is $= d^2 \text{ square } Q, P$ ok or $d^2 Q, P$ square ok.

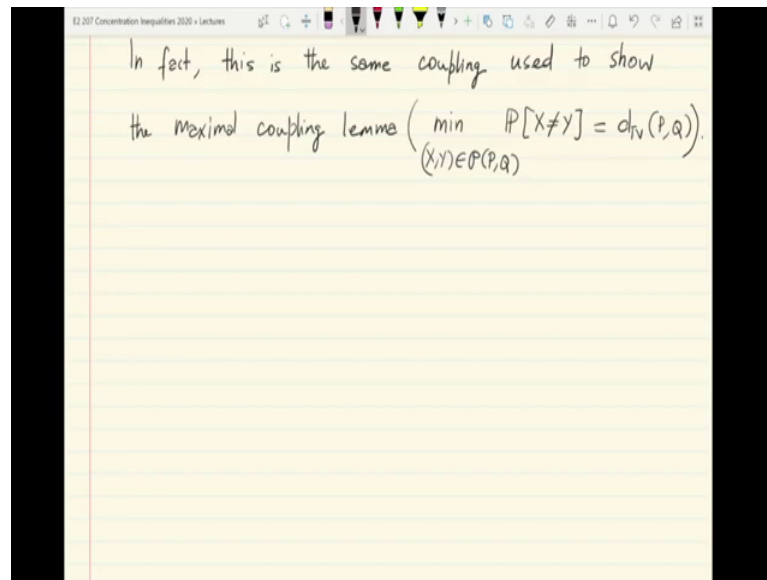
Similarly, if you condition on Y we have expected value in that of a $P(X \neq Y)$ given Y whole square is at least $d^2 P, Q$ whole square ok. So, what we have shown is basically that the left hand side in the lemma is lower bounded by the right hand side. So, this left hand side. So, the minimum value is at least the right hand side. Now, what we have to complete in order to show the to finish the proof is to show that there actually exists a coupling that attains equality.

(Refer Slide Time: 38:09)



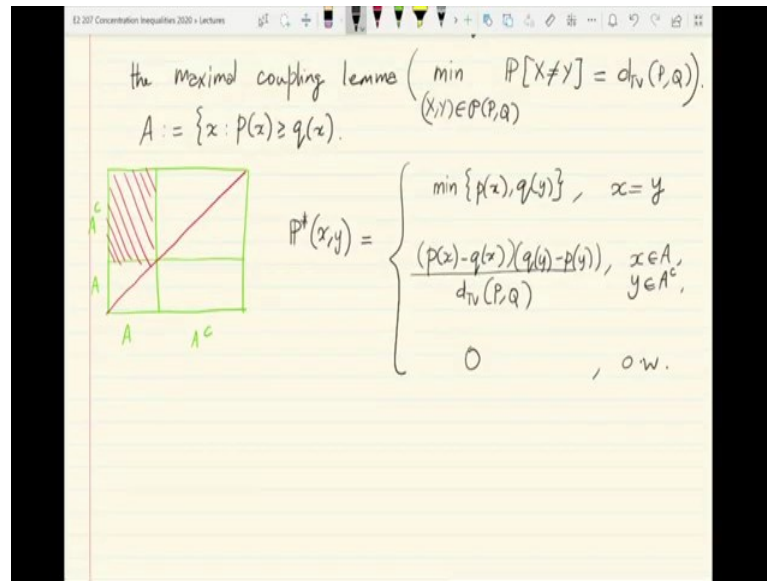
So, we next exhibit a coupling attaining exactly $d_2(Q, P^2) + d_2(P, Q^2)$ and what is this coupling? In fact, we can reuse the same coupling that we used to show the maximal coupling lemma ok.

(Refer Slide Time: 38:55)



In fact, this is the same coupling your use used to show the maximum coupling lemma, ok. Recall what was the maximal coupling lemma we showed? Basically we showed that there exists a coupling. So, rather the minimum over all couplings between P and Q of the probability of $X \neq Y$ and $=$ the total variation distance between P and Q . So, we will use the same coupling and just to recall what that coupling was so, we drew a picture let me just recall that picture.

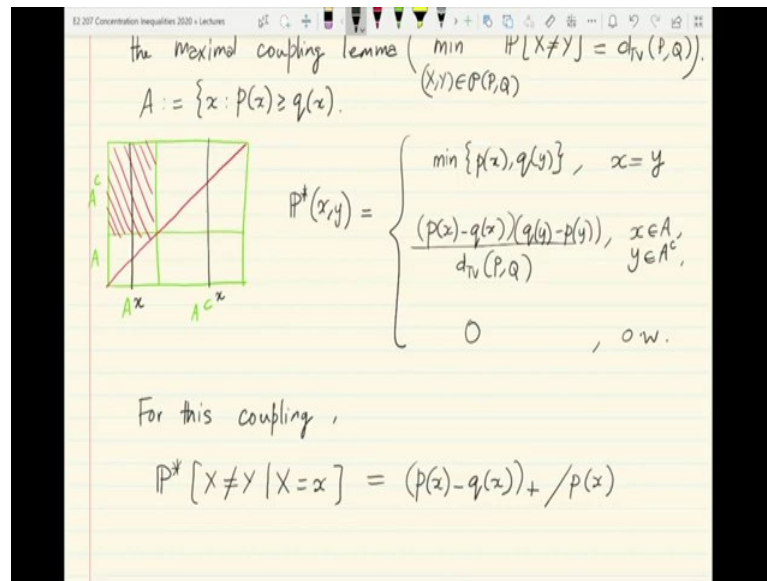
(Refer Slide Time: 39:39)



So, if you define A as the set where. So, let us say A is the set of all x where P dominates Q. So, we basically split this 2D square into a and a complement this is a on the X part a complement on the X part a and a complement on the X and Y parts and this was the diagonal where we put as much probability mass as possible. And the remaining probability mass was all put on the northwest quadrant here, ok.

So, now, let us just use the same coupling. So, just to put it down formally this coupling we called P^* x, y which is this define to be $\min\{p(x), q(y)\}$ if $x = y$, it was defined to be $\frac{p(x) - q(x)}{d_{TV}(P,Q)} \cdot \frac{q(y) - p(y)}{d_{TV}(P,Q)}$ the northwest quadrant divided by d_{TV} the total variation distance between P and Q on x belonging to A and y not belonging to A and 0 otherwise ok. So, that was exactly this coupling which put mass exactly on the diagonal and the remaining mass on the northwest corner.

(Refer Slide Time: 41:37)

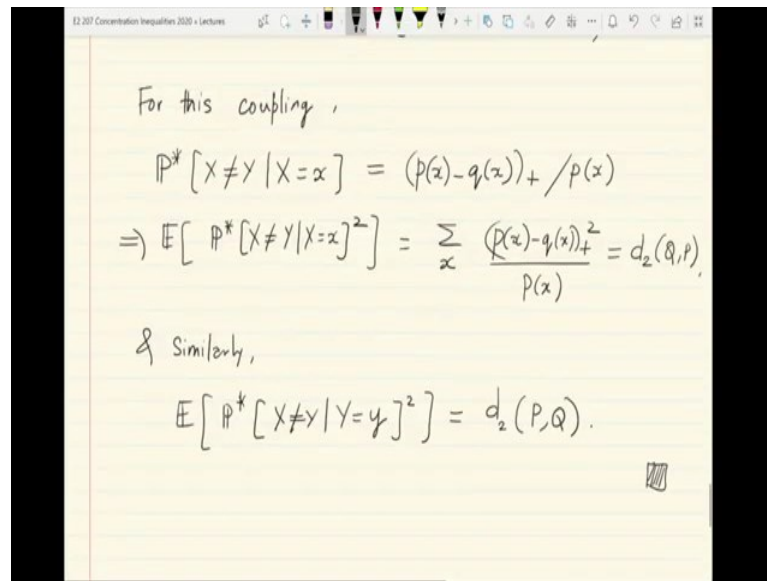


So, we have in this case so, for this coupling we can easily so, we can evaluate $P^* X \neq Y$ given $X = \text{small } x$; now, what is probability $X \neq Y$ given $X = \text{small } x$? Well, if $X = \text{small } x$ and $\text{small } x$ is in A , then there is some problem some there is potentially some nonzero probability of Y not equal in X .

But, on the other hand if X is in a complement then in fact, Y can Y always has to be $= X$ by our construction ok. So, this is only non this is this can only be nonzero if $\text{small } x$ lies in A , ok. So, this in fact, is just by our construction it is $p(x) - q(x) +$ divided by $p(x)$, ok. So, the conditional probability for any.

So, if you look at the first vertical black line here the probability that Y is not $= X$ given X is at this vertical X is at $= \text{small } x$ is simply the probability of the segment in the northwest the probability on the segment in the northwest corner divided by the total probability of this vertical line which is $p(x)$ and that is exactly what we have written.

(Refer Slide Time: 43:09)



The image shows a digital notepad with a yellow background and a toolbar at the top. The text is handwritten in black ink. It starts with 'For this coupling,' followed by the equation $P^*[X \neq Y | X=x] = (p(x)-q(x))_+ / p(x)$. Then, it shows the expectation of the square of this expression: $\Rightarrow E[P^*[X \neq Y | X=x]^2] = \sum_x \frac{(p(x)-q(x))_+^2}{p(x)} = d_2(Q, P)$. Next, it says 'Similarly,' and gives the symmetric equation: $E[P^*[X \neq Y | Y=y]^2] = d_2(P, Q)$. A small square symbol is at the bottom right of the text.

For this coupling ,

$$P^*[X \neq Y | X=x] = (p(x)-q(x))_+ / p(x)$$
$$\Rightarrow E[P^*[X \neq Y | X=x]^2] = \sum_x \frac{(p(x)-q(x))_+^2}{p(x)} = d_2(Q, P)$$

Similarly,

$$E[P^*[X \neq Y | Y=y]^2] = d_2(P, Q)$$

And, this implies that the expected value of this probability conditional probability square is just the sum over all x of. So, imagine squaring the right hand side expression. So, divided by $p(x)$ square, but you also multiply by a $p(x)$ because you are taking expectation with respect to small x . We just get $p(x)$ in the denominator sorry, as opposed to a $p(x)$ square this is exactly $d_2(Q, P)$ ok.

And, exactly along the same lines we also has expected value of $P^*[X \neq Y | Y=y]$ given $Y = \text{small } y$ whole square just by symmetry in fact, we have this = $d_2(P, Q)$. And, that completes the proof of this lemma and hence of Marton's conditional transportation cost inequality.

Thank you.