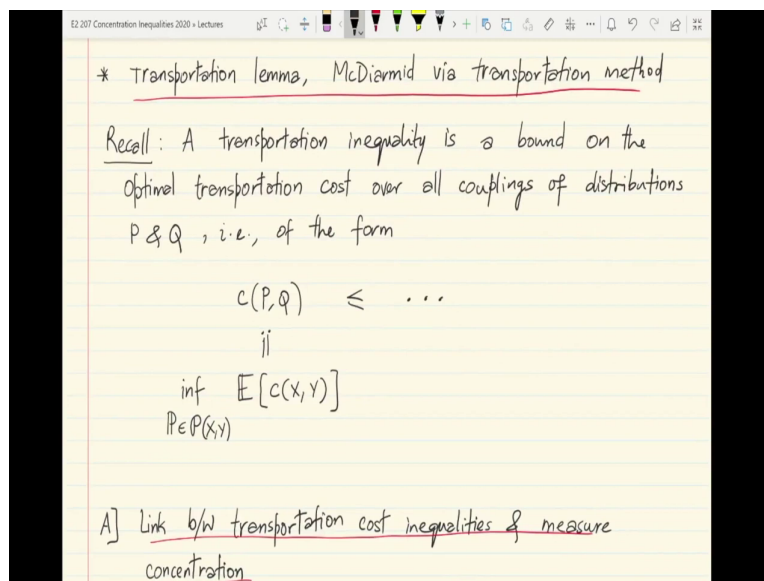


Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 15
Transportation lemma and a proof of McDiarmid's inequality using the transportation method

(Refer Slide Time: 00:21)



Welcome to this lecture. This lecture will cover what is called the transportation lemma and it will also expose us to the first concrete application of the transportation method to prove familiar concentration inequality which we have you shown using other methods which is the term inequality for functions with boundary differences.

So, let us recall from the previous lecture that, a transportation inequality or a transportation cost inequality is basically useful bound on the optimal transportation cost or the optimal cost for a transportation problem which is defined over all couplings or transportation plans between marginal distributions P and Q and we had this notation $c(P, Q)$ for it.

So, $c(P, Q)$ is just the minimum possible expected transportation cost over all join distributions with given marginals and any upper bound on it is essentially a transportation inequality.

(Refer Slide Time: 01:27)

LEMMA (Transportation Lemma)

$$\psi_{Z - \mathbb{E}_P Z}(\lambda) \leq \frac{v \lambda^2}{2} \quad \forall \lambda \geq 0$$

$$(\because \log \mathbb{E}_P e^{\lambda(Z - \mathbb{E}_P Z)})$$

if & only if

$$\forall Q \ll P \quad \mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{2v D(Q \| P)}.$$

So, how do transportation inequalities help lead to results on measure concentration? Ok. So, the core part of this is because of the following result which is called which one can call a transportation lemma. So, this result essentially says that bound on the log moment generating function of a random variable Z ok.

So, any bound of the following form which is a quadratic bound exists for a random variable Z if and only if corresponding bound exists for the differences of expectations of Z with respect to two measures Q and P being bounded by a the $\sqrt{\cdot}$ the relative entropy or $k l$ divergence between Q and P .

So, the statement that there is a sub Gaussian log moment generating function for Z for centered Z is equivalent to the fact that for every possible probability distribution Q absolutely continuous with respect to the original distribution P of Z . We have that the expected value of Z according to Q - the expected value of Z according to P can be bounded by some constant times the $\sqrt{k l}$ divergence between Q and P . So, this is the transportation lemma.

In fact, we have already seen one side of it in the introduction to the transportation method, we argued essentially that if we had bounds of the latter form where $\mathbb{E}_Q Z - \mathbb{E}_P Z$ is bounded as a square as a function of $\sqrt{\cdot}$ of the $k l$ divergence. Then you could get bounds on

the log moment generating function and that was basically by appealing to the variational formula for KL divergence.

(Refer Slide Time: 03:18)

if & only if

$$\forall Q \ll P \quad \mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{2V D(Q||P)},$$

Proof: Recall the (Gibbs') variational formula

$$\log \mathbb{E}_P \left[e^{\lambda(Z - \mathbb{E}_P Z)} \right] = \sup_{Q \ll P} \lambda(\mathbb{E}_Q Z - \mathbb{E}_P Z) - D(Q||P).$$

So, that is what we will exhibit again in this proof. But this lemma essentially formalizes the fact that bounds for moment generating functions and transportation cost type bounds are essentially two sides of the same coin. So, we start by recalling the standard Gibbs variation formula for KL divergence which we have seen and derived in earlier lectures.

Which states that the log moment generating function of a centered random variable is in fact, = the solution of a optimization problem over all absolutely continuous distributions Q with respect to P of $\lambda \mathbb{E}_Q Z - \mathbb{E}_P Z - D(Q||P)$ with respect to P .

(Refer Slide Time: 04:36)

Handwritten notes on a slide:

$$\log \mathbb{E}_P \left[e^{\lambda(Z - \mathbb{E}_P Z)} \right] = \sup_{Q \ll P} \lambda (\mathbb{E}_Q Z - \mathbb{E}_P Z) - D(Q \| P).$$

The term $\lambda(Z - \mathbb{E}_P Z)$ is underlined and labeled $\psi_{Z - \mathbb{E}_P Z}(\lambda)$.

"Only if" part: If $\forall \lambda \quad \psi_{Z - \mathbb{E}_P Z}(\lambda) \leq \frac{v\lambda^2}{2}$

So, for showing the forward part which is the only if part or the necessity part. So, if we assume that for all λ the centered log moment generating function right. So, by the way we will refer to the log moment generating function of $Z - \mathbb{E} Z$ under P as this C function.

So, if for alright. So, if for all λ we have this quantity being bounded by $v\lambda^2/2$ which is the first statement here right. In fact, we can just say that for all λ non negative.

(Refer Slide Time: 05:41)

Handwritten notes on a slide:

"Only if" part: If $\psi_{Z - \mathbb{E}_P Z}(\lambda) \leq \frac{v\lambda^2}{2} \quad \forall \lambda \geq 0$,

then for $Q \ll P$,

$$D(Q \| P) \geq \lambda (\mathbb{E}_Q Z - \mathbb{E}_P Z) - \frac{v\lambda^2}{2}.$$

Maximizing the RHS over λ gives

$$D(Q \| P)$$

Then for Q absolutely continuous with respect to P we can use the Gibbs variational formula to give us that $D(Q||P) \geq$. So, we just. So, since the log moment generating function is the supremum of this, we can always apply this for any $Q <$ absolutely continuous with respect to P and we will only get a quantity that is lower than log moment generating function and we can rearrange terms algebraically to get that $D(Q||P)$ is at least λ times $E_Q Z - E_P Z - \frac{\lambda^2}{2\sigma^2}$ ok.

Which is an upper bound on the log moment generating function and we can just treat this right hand side as a function of λ . So, if we just maximize the right hand side of this expression. In fact, we can just maximize this over λ gives to give $D(Q||P) \geq (E_Q Z - E_P Z)^2 / 2\sigma^2$ ok.

(Refer Slide Time: 07:22)

Maximizing the RHS over λ gives

$$D(Q||P) \geq \frac{(E_Q Z - E_P Z)^2}{2\sigma^2}.$$

"If" part : If $\forall Q \ll P \quad E_Q Z - E_P Z \leq \sqrt{2\sigma^2 D(Q||P)},$

Now, for the other side for the if part or the sufficiency part of the lemma. So, if for all absolutely continuous Q the hypothesis is that the expected value of Z with $Q - E_P Z$ is at most $\sqrt{2\sigma^2 D(Q||P)}$.

(Refer Slide Time: 08:11)

"If" part: If $\forall Q \ll P \quad \mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{2v D(Q||P)}$,
then $\forall \lambda \geq 0$:

$$\psi_{Z - \mathbb{E} Z}(\lambda) \leq \sup_{Q \ll P} \lambda \sqrt{2v D(Q||P)} - D(Q||P)$$

$$\leq \sup_x \lambda \sqrt{2v} \cdot x - x^2 = \frac{\lambda^2 v}{2}$$

NOTE: The lemma extends to any "nice" function g
as follows:

$$\forall \lambda \geq 0 \quad \psi_{Z - \mathbb{E} Z}(\lambda) \leq g(\lambda) \Leftrightarrow \forall Q \ll P \quad \mathbb{E}_P Z - \mathbb{E}_Q Z \leq g^{*-1}(D(Q||P)).$$

B] McDiarmid's inequality via Transportation Lemma

Then we have that for every λ non negative, the log moment generating function of Z with respect to the measure P is at most. So, we just substitute the upper bound for $\mathbb{E}_Q Z - \mathbb{E}_P Z$ in Gibbs variational formula.

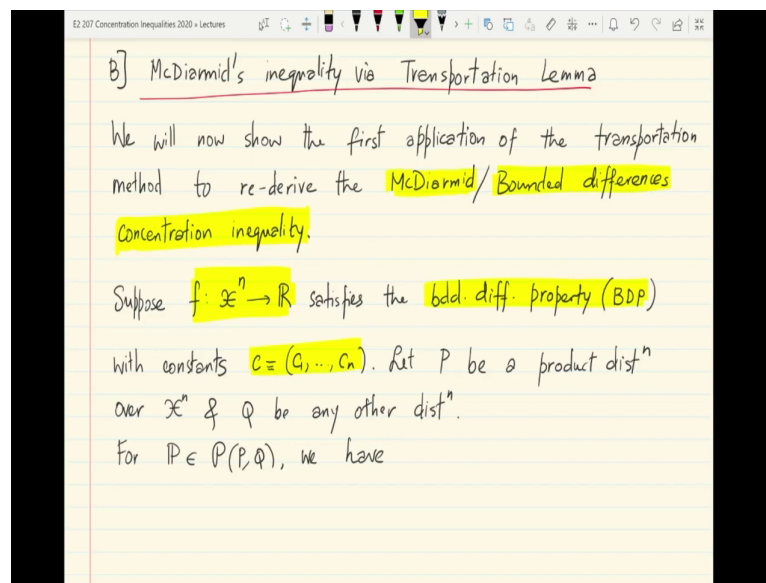
So, we get $\lambda \sqrt{2v D(Q||P)} - D(Q||P)$ and we can just treat this as a maximization problem over the variable $D(Q||P)$ or $\sqrt{D(Q||P)}$. So, the supremum over x and constraint $\lambda \sqrt{2v}$ into $x - x^2$ and this by exact analytical minimization gives you the closed form answer $\lambda^2 v / 2$.

So, this is the content of this transportation lemma. In fact, this lemma is little more general than what we wrote as then what we just wrote. So, we will just make a remark here that the lemma extends to any nice function nice enough function g . So, essentially we need g to be convex and with the smooth derivatives as follows. You can check this in the textbook for the more general version. So, for all λ non negative if we have the log moment generating function of $Z - \mathbb{E} Z$ being bounded by g of λ .

Some function g of λ this is equivalent to the statement that for all absolutely continuous $Q \ll P$ $\mathbb{E}_P Z - \mathbb{E}_Q Z$ is bounded above by the dual function g^* the Legendre dual of g its inverse applied to $D(Q||P)$. And so, this be specific lemma we wrote above is just a special case of the more general result when the g function is the quadratic function ok. I

So, with this transportation lemma in hand this transportation lemma essentially gives you the bridge between you know transportation type inequalities like these and log moment generating functions of random variables from which we already know there is a route to go to showing tail concentration using the standard Cherenkov technique.

(Refer Slide Time: 11:15)



So, we now come to what is probably the most important initial application of the transportation method. To derive something that we already seen before which is MacDiarmid's inequality or the bounded differences concentration inequality for stable functions of independent random variables.

So, we have derived MacDiarmid's inequality using other techniques like the entropy method and even much more you know more basic sense by using the Asuma Hobting method. But here we will derive it through the lengths of the transportation lemma. So, we will assume the same hypothesis as MacDiarmid's inequality which is that there is a function f defined on n variables which satisfies the bounded difference property with a set of constants one per dimensions C_1 through C_n .

So, it means that if we keep all, but one coordinates x_i the same and wiggle x_i the function value you can change by at most an amount C_i . Let P be a product distribution over x^n was to say that f acts on a bunch of n independent random variables and let Q be any other

distribution. So, we can start by saying that for any coupling or join distribution yielding marginals P and Q.

(Refer Slide Time: 12:45)

Over X & Y be any other dist.

For $P \in \mathcal{P}(P, Q)$, we have

$$\mathbb{E}_Q f(X) - \mathbb{E}_P f(X) = \mathbb{E}_{Y \sim Q} f(Y) - \mathbb{E}_{X \sim P} f(X)$$

$\because P \in \mathcal{P}(P, Q) \Rightarrow \mathbb{E}[f(Y) - f(X)] \stackrel{\text{BDP}}{\leq} \mathbb{E}\left[\sum_{i=1}^n c_i \mathbb{1}_{\{X_i \neq Y_i\}}\right]$

We can always write that $\mathbb{E}_Q f(X) - \mathbb{E}_P f(X)$. So, recall that we basically get want to use the transportation lemma to get a bound on the log moment generating function and the central object on which we need an upper bound for using the transportation lemma is expressions like these.

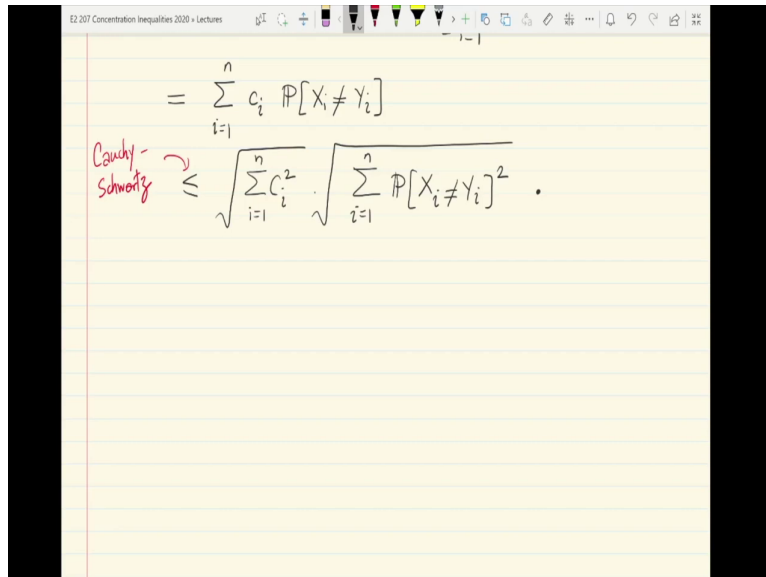
$\mathbb{E}_Q f(X) - \mathbb{E}_P f(X)$. So, think of $f(X)$ as z . So, you can interpret this so this is the power of the coupling method. So, you can interpret this as the expected value of y distributed according to Q of $f(Y) -$ the expected value of X distributed according to P of $f(X)$ ok.

The same random variables expectation being taken under to different measures and we can actually now use the coupling which is bold face P to actually write things under a common expectation as the E without any subscripts which is sort of attached to this P of $f(Y) - f(X)$.

So, its a single expectation that is equivalent to the difference of two different expectations. So, this is just because P is a coupling between P and Q ok. So, the individual marginals are the right marginals and now that we have $f(Y) - f(X)$ inside the expectation.

We can actually use almost sure properties or point wise properties of f which in this case is the bounded difference property to write that this expectation. In fact, the inner term is bounded almost surely. So, applying the expectation does not change things $i = 1$ to n . We know that this almost sure bound holds this is by the bounded differences property we assumed and this now sets the states for this to look like a transportation cost inequality ok.

(Refer Slide Time: 14:57)



$$= \sum_{i=1}^n c_i \mathbb{P}[X_i \neq Y_i]$$

Cauchy-Schwarz \Rightarrow

$$\leq \sqrt{\sum_{i=1}^n c_i^2} \cdot \sqrt{\sum_{i=1}^n \mathbb{P}[X_i \neq Y_i]^2}.$$

So, this is exactly = if you interchange summations and expectations the sum of C_i times the probability that X_i and Y_i are different ok and we can further upper bound this by using something like Cauchy Schwartz sum over all i of C_i^2 under the $\sqrt{\quad}$ into the sum over all i of the squares of these probabilities under the joint distribution P .

This is let us say by using Cauchy Schwartz ok and so, what we have achieved using this exercise is an upper bound on $E Q f - E P f$ in terms of some constant which is the $\sqrt{\quad}$ of the this is the l_2 norm of the C vector the bounded difference vector C into some other term.

So, we will only benefit if we can upper bound this remaining the last term here on the right in terms of the $k l$ divergence between P and Q distributions P and Q ok. So, that will help us to directly use the transportation lemma and push things through.

(Refer Slide Time: 16:25)

So, by the Transportation lemma, it is enough to show:

$$\exists \text{ a coupling } P \in \mathcal{P}(P, Q) \text{ s.t.}$$

$$\sum_{i=1}^n P[X_i \neq Y_i]^2 \leq \frac{1}{2} D(Q || P)$$

(#)
(A transportation cost inequality)

to get: $\psi_{Z - E_P Z}(\lambda) \leq \left(\sum_{i=1}^n C_i^2\right) \lambda^2 / 8 \quad \forall \lambda \geq 0$,

which in turn implies: (by Chernoff)

$$P[f(X) \geq E_P f(X) + t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n C_i^2}\right).$$

So, by the transportation lemma it stands to a reason that you know it is enough to show that for instance there exists a couplings. So, if we show there exists a coupling a nice enough coupling that we can find a P between P and Q such that the sum over all i of P probability that $X_i \neq Y_i$ use let us say at most some multiple of the k l divergence $D(Q || P) / 2$ ok.

So, it is enough to show this kind of property to get. So, using the transportation lemma we can just push this through to get the log moment generating function of Z at λ being bounded $\sum_{i=1}^n C_i^2$ into $\lambda^2 / 8$ for all λ non negative.

Which in turn will imply a right side tail bound for the deviation of f of X about its mean ok by Chertoff or sub Gaussian random variables. The original P probability of f of x the random variable f of X exceeding its own expectation by a number t is at most it has a sub Gaussian type tail modulated by the constants C_i ok.

So, the main thing that the transportation lemma assures us is enough to show is a statement of the following form that one can couple P to Q while keeping this sort of transportation cost like expression on the left not more than a multiple of its k l divergence.

So, this statement or property that we require which we will let us we will call hash is essentially a unique quality involving transportation costs because for each i probability we have already seen that for each specific i the probability that $X_i \neq Y_i$ is actually an

expected transportation cost where the cost structure is given by the indicator that $X_i \neq Y_i$ for every individual tuple X_i and Y_i ok.

So, it's the hamming coster hamming transportation cost and so, this is essentially an inequality that involves transportation cost and a bound on them ok. So, the strong result that Marton actually established is basically this exact result this transportation cost inequality that allows us to prove successfully the bounded differences inequality.

(Refer Slide Time: 20:13)

(#) was shown to hold by Marton in 1986.

* Marton's Transportation cost inequality

Theorem: For $X = (X_1, \dots, X_n) \sim P = P_1 \times \dots \times P_n$, & $Q \ll P$, let $Y = (Y_1, \dots, Y_n) \sim Q$. Then, there exists a coupling P of P & Q s.t.

$$\sum_{i=1}^n \mathbb{P}[X_i \neq Y_i]^2 \leq \frac{1}{2} D(P \parallel Q). \quad (*)$$

So, in fact, let us remark that this property the existence of such a coupling star was shown affirmatively it was answered. So, was shown to hold by first time by Catlin Marton in 1986 and there have been several development after this in the use of transportation inequalities to bridge to concentration of major inequalities.

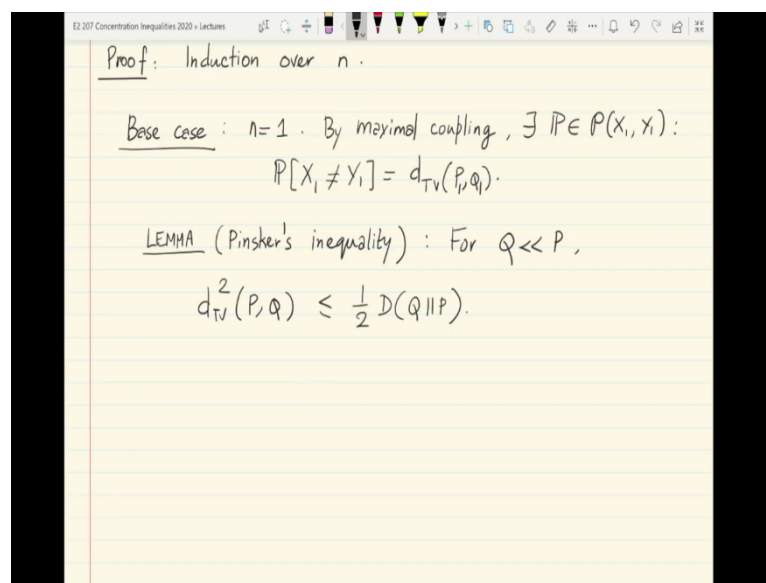
So, let us go ahead and study Marton's transportation cost inequality which is the last Ps remaining in this exercise to successfully derive MacDiarmid's bounded differences inequality. So, here is Marton's transportation cost inequality. It says that you have a bunch of random variables X_1 through X_n distributed independently according to the joint measure P . So, P with the i th marginal of P being denoted as P_i P_1 through P_n .

And Q is any absolutely continuous distribution with respect to P for the set of random variable X_1 through X_n in fact, let us assume that Y is a bunch of random variables n

random variables distributed jointly according to Q . Then there actually exists a coupling between P and Q for these n random variables X_1 through n and Y_1 through n satisfying the property that the summation was the disagreement probability squared over all i between X_i and Y_i is at most $\frac{1}{2}$ times the k_l divergence between Q and P .

So this is exactly Marton's transportation cost inequality which completes the derivation of MacDiarmid inequality via the transportation lemma and finally, using Chernoff ok. So, the last part of this lecture is going to be the proof of Marton's transportation cost inequality we will again we will actually build a build or find an explicit coupling which satisfies this property, just as we did with the maximal coupling for instance.

(Refer Slide Time: 22:28)



So, let us go ahead and do the proof of Marton's transportation inequality. The proof is going to be by induction over the number of elements n over the number of random variables to be coupled. So, for the base case this is itself an interesting problem see $n = 1$ case and we already know by the maximal coupling lemma that there exists one nice maximal coupling which can couple X_1 through Y_1 .

$P[X_1 \neq Y_1] = d_{TV} \leq d_{TV}(P, Q)$ ok. So, there exists a P which can couple X_1 to Y_1 such that this holds and so, all that is required to show is that the square of this quantity which is

the total variation distance P yeah. So, $P \perp$ and $Q \perp$ I would say. So, all we need to show by the theorem for the base case is that the square of the total variation distance is at most $1/2$ $k \perp$ divergence that is exactly what is Pinsker's inequality which we will now prove.

So, the base case for Marton's transportation cost inequality is Pinsker's inequality. So, it is a generalization of Pinsker's inequality. So, which simply says that for any Q absolutely continuous with respect to P . The total variation between P and Q 2 is at most $1/2$ of $D Q P$ ok and the proof is by appealing to the transportation lemma.

(Refer Slide Time: 24:48)

PROOF: Put $A := \left\{ \frac{dQ}{dP} \geq 1 \right\}$, so that

$$d_{TV}(P, Q) = Q(A) - P(A) = \mathbb{E}_Q Z - \mathbb{E}_P Z, \text{ for}$$

$$Z := \mathbb{1}_A. \text{ By Hoeffding's lemma,}$$

$$\psi_{Z - \mathbb{E}_P Z}(\lambda) \leq \frac{\lambda^2}{8}, \text{ so using the}$$

transportation lemma

So, let us put A as the let us define A as the set where $d Q / d P$ the radon Nikodym derivatives of Q with respect to P or in the discrete case. Ah. So, this is larger than $= 1$ discrete case A is just the set of all elements whose Q probability is at least the P probability. So, that $d T V$ the total variation distance between P and Q is $Q A - P A$ which one can write equivalently as $\mathbb{E} Q Z - \mathbb{E} P Z$ for the random variable Z being defined as the indicator random variable for the set A .

Now, by Hoeffding's lemma for the bounded random variable Z which takes values only 0 between 0 and 1 we have that the log moment generating function of Z if we center it is at most its a sub Gaussian log moment generating function ok $\lambda^2 / 8$ there is a bound for it.

(Refer Slide Time: 26:32)

$\forall \lambda: \psi_{Z-EZ}(\lambda) \leq \frac{\lambda^2}{8}$, so using the
 transportation lemma,
 $d_{TV}(P,Q) = \mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{\frac{1}{2} D(Q||P)}$.

And so, using the transportation lemma we can convert this into a bound on the $\mathbb{E}_Q Z - \mathbb{E}_P Z$. So, this by the way holds for all λ ok follows that the $\mathbb{E}_Q Z - \mathbb{E}_P Z$ must be at least must be at most $\sqrt{\frac{1}{2} D(Q||P)}$ ok. And that finishes the lemma because $\mathbb{E}_Q Z - \mathbb{E}_P Z$ is exactly the total variation distance $d_{TV}(P, Q)$ ok. So, the base case is done which is the familiar Pinsker inequality that relates that helps us control total variation distance in terms of the KL divergence.

(Refer Slide Time: 27:18)

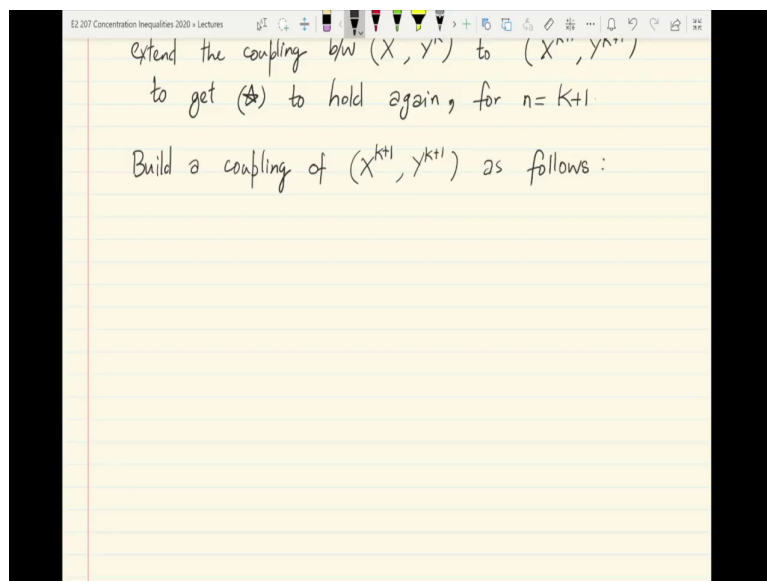
$d_{TV}(P,Q) = \mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{\frac{1}{2} D(Q||P)}$.

Induction step: Suppose $(*)$ holds for $n=K$. We'll
 extend the coupling b/w (X^K, Y^K) to (X^{K+1}, Y^{K+1})
 to get $(*)$ to hold again.

Now, for the induction step let us assume that the claimed inequality in the theorem holds for $n = k$. So, suppose the inequality in the theorem let us call it. So, we have already called it star. So, suppose that the statement star holds the conclusion of the theorem holds for collection of K random variables X_1 through K and Y_1 through K .

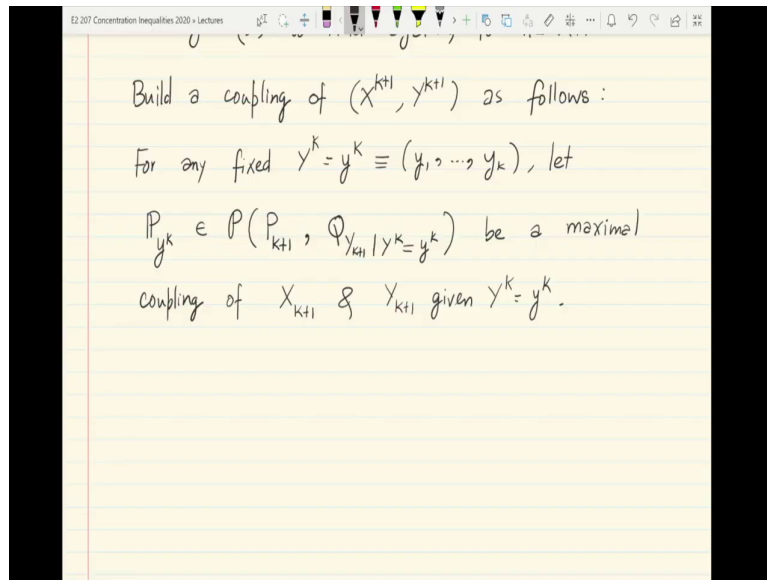
That means there is coupling that gives you the desired property. So, we will just extend that coupling we will extend the coupling between X_k . So, $X^{(k)}$ will be used to denote X_1 through X_k . The collection of variables random variables x_1 through X_k and Y_1 through Y_k to get. So, we will extend this coupling between these random variables to a larger collection X up to $k + 1$ and Y up to $k + 1$ to get star to hold again for the larger collection.

(Refer Slide Time: 28:50)



So for $n =$ precisely for $n = K + 1$. So, the key to doing this extension of the existing coupling is as follows. So, let us first couple the last part so build a coupling. So, let us build a grand coupling of the $k + 1$ collection of random variables. X from 1 to $k + 1$ and Y from 1 to $k + 1$ as follows. So, first for any fixed Y . So, first for any fixed configuration of the first K Y 's.

(Refer Slide Time: 29:26)



So, the first K y 's let them let me let us denote them as y_1 up to y_k or small y superscript k . Let us define P_{y^k} to P subscript y^k you should read this as bold P given the configurations small y^k small y superscript k as a coupling. So, let this P_{y^k} denote a coupling that couples the $k + 1$ first marginal of x which is P_{k+1} with the $k + 1$ first joint distribution of $K + 1$ was conditional distribution to y_{k+1} given the previous y 's.

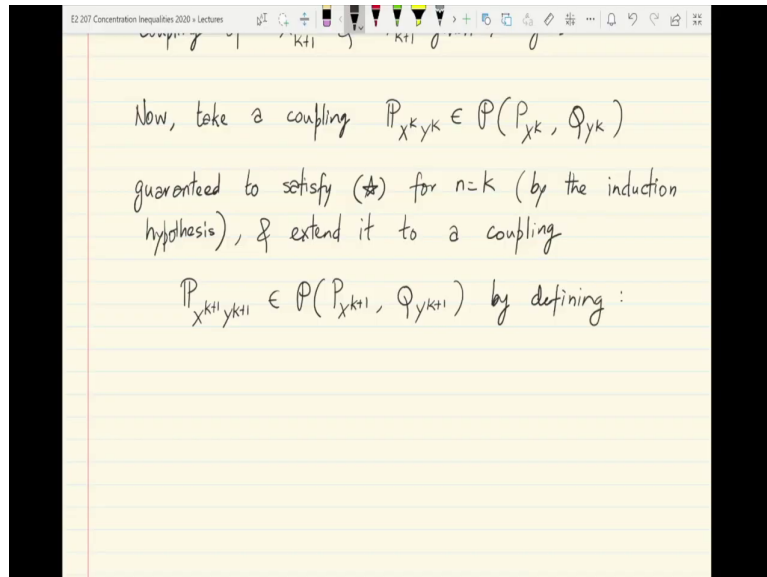
So, denote this as Q the distribution the Q distribution of the $k + 1$ first y given that the previous k y 's have been observed to have configuration small y^k ok. So, let this be a maximal coupling of these two distributions of. So, as random variables we can say its a maximal coupling of X_{k+1} the single random variable X_{k+1} and Y_{k+1} given $Y^k =$ small y^k .

We all we always know by the maximal coupling lemma we know that such a maximal coupling always exists ok such that the discrepancy probability between these two random variables that it couples is exactly = the $k + 1$ divergence between the marginals that it couples ok.

So, let. So, just get hold of. So, for any given conditioning or fixed past configuration small y^k of the y 's let P_{y^k} denote maximal coupling that jointly delivers a pair x_{k+1} and y_{k+1} given that small y^k has happened ok.

And which also achieves the total variation constraint with equality in the transportation cost problem. So, we will write it down explicitly later. So, given the past y 's this basically gives you a joint pair a fresh joint pair x_{k+1} and y_{k+1} ok.

(Refer Slide Time: 32:15)



Now, having done this take a coupling P let us call it that couples X^k the first K X 's and the first k Y 's. So, this couples P_{X^k} and Q_{Y^k} that is guaranteed to satisfy the inequality star for $n = k$. We know that this is guaranteed to exist by the induction hypothesis. So, grab hold of any coupling between X^k and Y^k that satisfies the inequality star and we will now use this conditional extra coupling to extend the original coupling P_{X^k, Y^k} ok.

And extend it to a coupling let us call the new coupling $P_{X^{k+1}, Y^{k+1}}$ coupling these two sequences with one extra element each which we will insist couples $P_{X^{k+1}}$ and $Q_{Y^{k+1}}$ by. So, its defined in the usual manner.

(Refer Slide Time: 33:54)

$$P_{X^{k+1}, Y^{k+1}} \in \mathcal{P}(P_{X^{k+1}}, Q_{Y^{k+1}}) \text{ by defining:}$$

$$P_{X^{k+1}, Y^{k+1}} = P_{Y^k} \times P_{X^k | Y^k}.$$

By the maximal coupling property of P_{Y^k} for each y^k , we have

So, symbolically what one can say is that if we want the joint distribution on X^{k+1}, Y^{k+1} how you can generate a joint sequence X^{k+1}, Y^{k+1} with the respective marginals is that you first generate X^k, Y^k ok. So, you first generate X^k, Y^k using the existing coupling that is guaranteed by the induction hypothesis. Now having the Y^k sequence in your hand the previous Y all the way from 1 to k in your hand you use this new extra coupling P_{Y^k} .

So, you take small y^k as the actual y^k that has happened in the past and using that you can basically generate a new X^{k+1} and a new Y^{k+1} conditioned on the exact Y^k in the past. There is no condition required in required on X^k in the past because the new X is going to be independent from the past X s. So, what you do is you take. Firstly, you generate X and Y up to length k and then you generate the P_{Y^k} .

So, you generate X^{k+1} and Y^{k+1} conditioned on y all the way up to k using the extension coupling the conditional coupling to basically give you a sample from the entire sequence X up to $k+1$ and Y up to $k+1$. So, this part gives you X^k and Y^k and this part by definition of P_{Y^k} gives you a new tuple X^{k+1} and Y^{k+1} that you can append to the existing sequence X^k, Y^k .

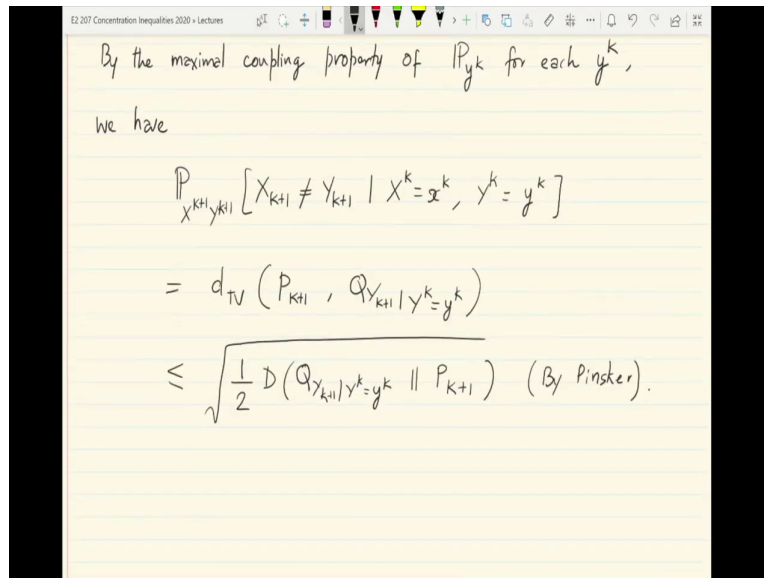
So, now, what we can write is that by the maximal coupling property of the tailed the tail coupling P_{Y^k} for each possible value of the configuration in the past y^k .

(Refer Slide Time: 36:34)

The image shows a digital notepad with handwritten text. At the top, the equation $\pi_{X^{k+1}Y^{k+1}} = \pi_{Y^k} \times \pi_{X^kY^k}$ is written, with red underlines under π_{Y^k} and $\pi_{X^kY^k}$. Below this, the text reads: "By the maximal coupling property of P_{Y^k} for each y^k , we have $P_y \int X_{k+1} \neq Y_{k+1} \mid X^k = x^k, Y$ ".

We have that. So, this is the maximum couple maximal coupling property which ensures that the probability under this y^k of X_{k+1} been unequal from Y_{k+1} given $X^k = x^k$ and Y sorry. So, we do not need the subscript P_{y^k} this is under the extension coupling.

(Refer Slide Time: 37:04)



By the maximal coupling property of P_{y^k} for each y^k ,
we have

$$\begin{aligned} & \mathbb{P}_{X^{k+1}|Y^{k+1}} [X_{k+1} \neq Y_{k+1} | X^k = x^k, Y^k = y^k] \\ &= d_{TV}(P_{k+1}, Q_{Y_{k+1}|Y^k=y^k}) \\ &\leq \sqrt{\frac{1}{2} D(Q_{Y_{k+1}|Y^k=y^k} || P_{k+1})} \quad (\text{By Pinsker}). \end{aligned}$$

So in fact, let me just be more precise and say that P under $X_{k+1} Y_{k+1}$ let me just constructed of X_{k+1} not being = in the $k+1$ first symbol Y given the past of X is small x^k and the past of Y is small y^k must be = the total variation distance between the distributions that the tail and coupled which are P_{k+1} and conditional distribution of Y_{k+1} given $Y^k = Y^k$ small y^k .

And by Pinsker's inequality we know that this is at most $\sqrt{1/2}$ the $k+1$ divergence between the conditional distribution $Q_{Y_{k+1}|Y^k=y^k}$ with respect to the P_{k+1} . Distribution of the $k+1$ first X_{k+1} by Pinsker which is just a one dimensional inequality and so, we can just take expectation over the conditioning here to make it a single probability. So, this implies that we when we take expectation over the past of x and the past of y up to time k .

(Refer Slide Time: 38:31)

$$\Rightarrow P[X_{k+1} \neq Y_{k+1}] \leq E_{Q_{Y^k}} \sqrt{\frac{1}{2} D(Q_{Y_{k+1}|Y^k} \| P_{k+1})}$$

Jensen \rightarrow

$$\leq \sqrt{\frac{1}{2} E_{Q_{Y^k}} D(Q_{Y_{k+1}|Y^k} \| P_{k+1})}$$

We get that under. So, of course, I am omitting the subscript $X_{k+1} Y_{k+1}$, but all of this is under the new coupling for X up to $k+1$ and Y up to $k+1$. So, this satisfies the property that probability X_k unconditionally X_{k+1} not being $= Y_{k+1}$ is at most the expected value under Q_{Y^k} of the $\sqrt{\frac{1}{2}}$ of $\frac{1}{2}$ the $k+1$ divergence $Q_{Y_{k+1}|Y^k}$ the conditional distribution with respect to P_{k+1} .

And this by Jensen's inequality for the concave $\sqrt{\cdot}$ function is bounded by the $\sqrt{\frac{1}{2}}$ the expected value of the divergence of the conditional divergence ok right. So, this is by Jensen's inequality.

(Refer Slide Time: 40:03)

Thus,

$$\sum_{i=1}^{k+1} \mathbb{P}[X_i \neq Y_i]^2 \leq \frac{1}{2} D(Q_{Y^k} \| P_{X^k}) + \mathbb{P}[X_{k+1} \neq Y_{k+1}]^2$$

induction hyp.

$$\leq \frac{1}{2} D(Q_{Y^k} \| P_{X^k}) + \frac{1}{2} D(Q_{X_{k+1}} \| P_{Y_{k+1}} | Q_{Y^k})$$

So, finally, what we have is that the statement star for $n = k + 1$ its left hand side is simply the sum over all i going from i to 1 to $k + 1$ of the probability square that $X_i \neq Y_i$.

So, you can spilt this into the sum from 1 to k and then the $k + 1$ first term the induction hypothesis gives us a bound for the first k terms sum of the first k terms as $D(Q_{Y^k} \| P_{X^k})$ this is by the induction hypothesis + the additional term the last square term at position $k + 1$ which we just bounded above as $\frac{1}{2} \frac{1}{2}$ the $k + 1$ divergence between $Q_{Y^{k+1}}$ with respect to $P_{Y^{k+1}}$ conditioned on Q_{Y^k} its the expectation expected $k + 1$ divergence which we just written under this notation.

So, its the conditional $k + 1$ divergence of $Q_{Y^{k+1}}$ given the entire past of Y with respect to $P_{Y^{k+1}}$ and then average doubt across the past Q_{Y^k} .

(Refer Slide Time: 41:39)

$$\begin{aligned}
 & \leq \frac{1}{2} D(Q_{Y^k} || P_{X^k}) + \frac{1}{2} D(Q_{X^{k+1}} || P_{Y^k}) \\
 & \stackrel{\text{chain rule for KL divergence}}{=} \frac{1}{2} D(Q_{Y^{k+1}} || P_{X^{k+1}}). \quad \square
 \end{aligned}$$

And this precisely by the chain rule of KL divergence equals what we want which is D between Q_{Y^k} up to $k+1$ and P_{X^k} up to $k+1$. So, this is by the chain rule for KL divergence which we seen earlier and that completes the proof. So, this was MacDiarmid's inequality or the bounded difference inequality derived using a very different method which is via the transportation lemma.

Thank you.