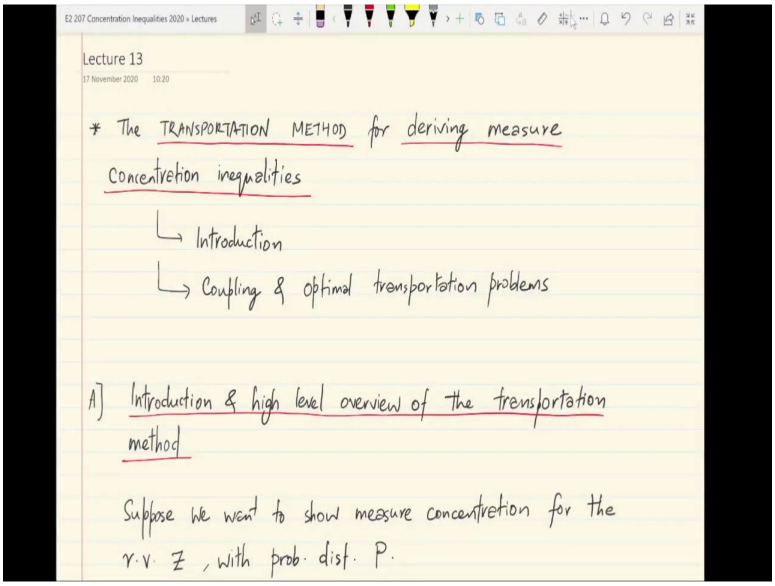
Concentration Inequalities

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

Lecture - 14
Introduction to the transportation method for showing concentration bounds

(Refer Slide Time: 00:21)



Lecture 13
17 November 2020 10:20

* The TRANSPORTATION METHOD for deriving measure concentration inequalities

↳ Introduction

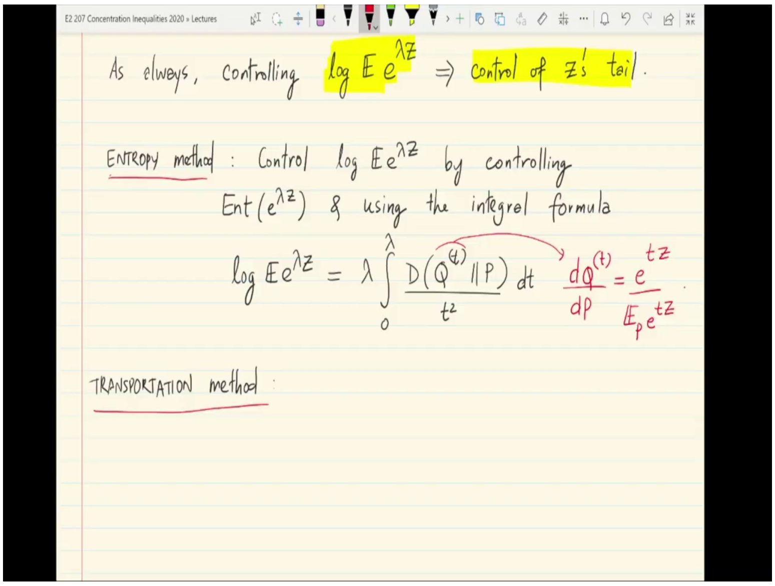
↳ Coupling & optimal transportation problems

A] Introduction & high level overview of the transportation method

Suppose we want to show measure concentration for the r.v. Z, with prob. dist. P.

Hi all. Today, we will study another very ingenious method called the transportation method for deriving concentration of measure inequalities. So, this class will be about an introduction or a high-level idea about what this method allows you to do and we will start digging into it by presenting the idea of coupling and optimal transportation which are important ideas and probability in optimization in their own right.

(Refer Slide Time: 00:55)



E2 207 Concentration Inequalities 2020 » Lectures

As elweys, controlling $\log \mathbb{E} e^{\lambda Z}$ $\Rightarrow$ control of Z's tail.

ENTROPY method : Control $\log \mathbb{E} e^{\lambda Z}$ by controlling $\mathrm{Ent}(e^{\lambda Z})$ & using the integrel formula

$$\log \mathbb{E} e^{\lambda Z} = \lambda \int_0^{\lambda} \frac{D(Q^{(t)} \| P)}{t^2} \, dt \qquad \frac{dQ^{(t)}}{dP} = \frac{e^{tZ}}{\mathbb{E}_P e^{tZ}}.$$

TRANSPORTATION method :

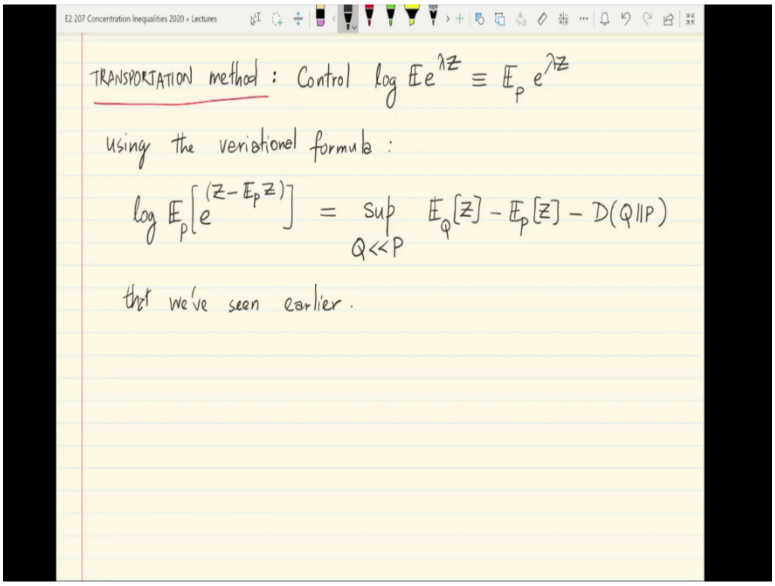
So, let us start with a high-level introduction of the transportation method for proving concentration inequalities. So, recall that suppose we want to show concentration of measure for some random variables Z having probability distribution P. As always, we know that if we control the log moment generating function of Z, then we can hope to control the tail of Z using the Chernoff bound. So, in particular, quadratic log moment generating functions give sub-Gaussian tail bounds and so on and so forth.

So, against this backdrop, what have we studied so far? We have looked at the well-known entropy method which essentially says one can control the moment generating function log expected value of e raise to λ Z for some real number λ by basically controlling the entropy of a certain random variable in particular Ent of e raise to λ Z

and then, by using or appealing to an integral formula for the log moment generating function which is of the form log expected value of e raise to λ Z expressed as an integral from 0 to λ of the relative entropy or KL divergence between two measures Q subscript t and the original measure P of Z divided by t ^{2} dt where if you recall Q of t was defined using the transformation or the Radon-Nikodym derivative dQ t by dP being defined as e raise to tz divided by the expected value under P of e raise to tZ.

So, the entropy method crucially relied on this kind of integral equality and further bounds on the entropy of e raise to λ Z which can be translated on the on a bound on this integrand to in turn control the log moment generating function in the sense of an upper bound.

(Refer Slide Time: 03:30)



TRANSPORTATION method : Control $\log \mathbb{E} e^{\lambda Z} \equiv \mathbb{E}_P e^{\lambda Z}$

using the veriational formula :

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

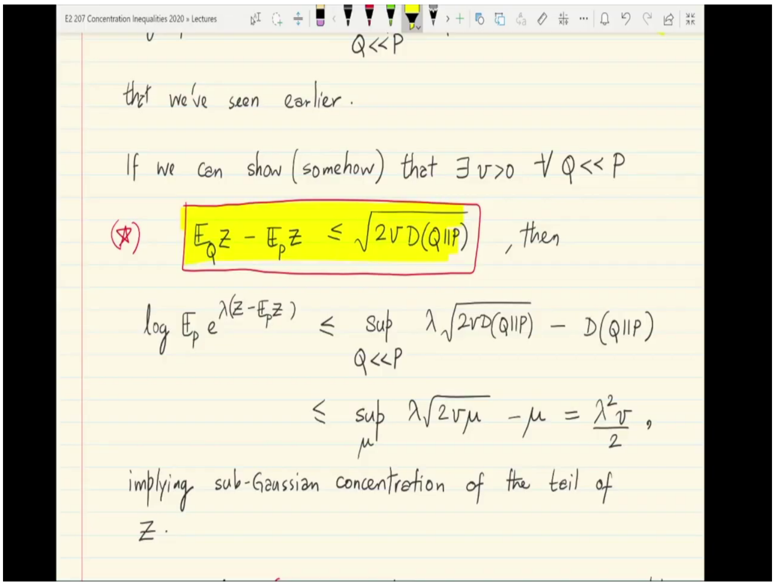
that we've seen earlier.

Now, what the transportation method does is somewhat related in spirit, but the exact tools are quite different and very elegant. It also seeks to control the log moment generating function ok which is to be more precise, it is the expectation according to Z distributed according to P of e raise to λ Z.

Using the variational formula for log moment generating function that we have seen earlier which is of the following form so, log expected value under P of e raise to let us say the centered version of Z, Z - its own expectation is a solution to an optimization problem over all measures or probability distribution skew absolutely continuous with respect to the original distribution P of Z of the expected value under Q of Z - the expected value of P under Z - the divergence between the measures Q and P.

So, we have seen this variational formula, we have derived this variational formula in a previous lecture, and this is exactly what the transportation method is going to appeal to ok. So, this is something that we saw earlier.

(Refer Slide Time: 05:18)



$Q \ll P$

that we've seen earlier.

If we can show (somehow) that $\exists\, v > 0 \;\forall\, Q \ll P$

$$(*) \quad \boxed{\mathbb{E}_Q Z - \mathbb{E}_P Z \le \sqrt{2v D(Q\|P)}} \;, \text{then}$$

$$\log \mathbb{E}_P e^{\lambda(Z - \mathbb{E}_P Z)} \le \sup_{Q \ll P} \lambda \sqrt{2v D(Q\|P)} - D(Q\|P)$$

$$\le \sup_{\mu} \lambda \sqrt{2v\mu} - \mu = \frac{\lambda^2 v}{2},$$

implying sub-Gaussian concentration of the teil of Z.

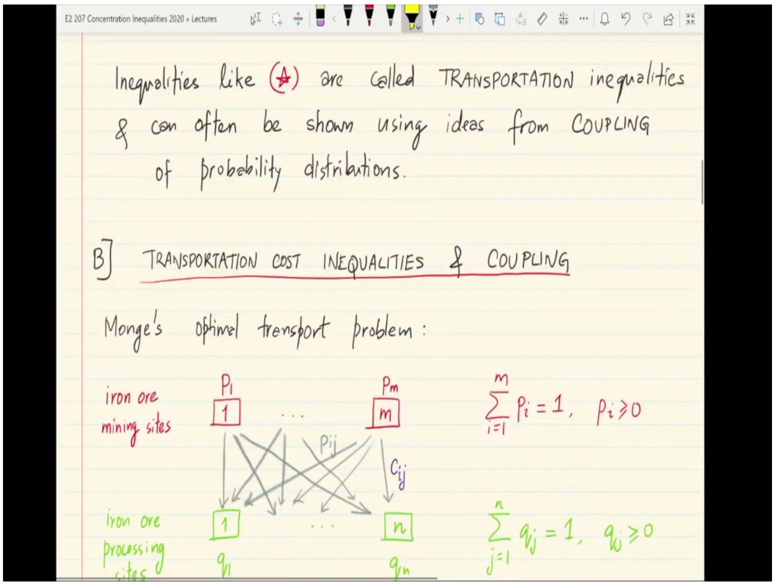
Now, with this formula in hand, suppose we can somehow show. So, if we can show somehow that there is some number v which is positive such that for all absolutely continuous measures Q with respect to P, this quantity here the E Q Z - E P Z can be bounded in terms of D Q, P itself as let us say E Q Z - E P Z is bounded by √2v D Q P ok all that is important here is the √dependence on D Q P, let us call this inequality, let us label this inequality as star.

Then, what we will have is an upper bound on this difference of expectations with respect to Q and P for Z and we can go ahead with upper bounding the right-hand side, the objective function in the variational formula as follows. So, log E P e raise to λ Z - E P Z ok is at most the supremum over all absolutely continuous Q of λ times so, recall that we have a λ × Z so, you can so; so, this E Q Z - E P Z now gets a λ multiplier.

And if star is true, then we can substitute the upper bound for E Q Z - E P Z and then finally, subtract D Q, P ok. Now, we can think of D Q, P as a variable in this optimization problem as any number, any non-negative number in fact, any real number and we can find a further upper bound here by just saying supremum over all numbers mu of λ √ 2v mu - mu ok because D Q, P is some number which is non-negative so, we are just allowing D Q, P, we are replacing D Q P by mu and allowing mu to take any real valued number.

And if you solve this explicitly, this you can just treat as a quadratic function of √ mu and the answer to this is in closed form λ^2 v by 2 ok. So, what we have shown is that if there are bounds of the form star where a difference of expectations of the same random variable with respect to two different measures is upper bounded by a function of the relative entropy or KL divergence, then you can look forward to actually showing bounds on the log moment generating function in particular of the sub-Gaussian type so, this means that so, this implies sub-Gaussian concentration as an example of the tail of Z.

(Refer Slide Time: 07:31)



E2 207 Concentration Inequalities 2020 » Lectures

Inequalities like (★) are called TRANSPORTATION inequalities & can often be shown using ideas from COUPLING of probability distributions.

B] TRANSPORTATION COST INEQUALITIES & COUPLING

Monge's optimal transport problem :

iron ore mining sites
p_1
1
. . .
p_m
m
p_{ij}
c_{ij}
iron ore processing sites
1
. . .
n
q_1
q_n

$\sum_{i=1}^{m} p_i = 1, \quad p_i \geqslant 0$

$\sum_{j=1}^{n} q_j = 1, \quad q_j \geqslant 0$

So, inequalities of the type star precisely what are called transportation cost inequalities. So, inequalities like star are called transportation inequalities for reasons that we will come to soon or transportation cost inequalities and can often be shown and can often be derived or shown using ideas from what is called coupling of probability distributions of probability measures or probability distributions ok.

So, for now, at a high-level, for a high-level overview of the transportation method, suffice it suffices to say that the transportation method is focused all the action in the transportation method is focused along proving inequalities of the following type where a difference of expectations with respect to the same of the same random variable with respect to different measures can somehow be suitably upper bounded by sort of in this case, functions of the KL divergence between these measures. And that is accomplished in turn using ideas from what is called optimal transportation theory and coupling theory from probability.

So, what are transportation cost inequalities and what is the relationship to coupling? So, for this, let us take a detour to what is called Monge’s optimal transport problem. So, this is a very old problem in actual transportation of materials and operations research that Monge’s actually formulated back in the 1700s many centuries ago.

So, to illustrate this optimal transport problem; let us consider the example of, the motivating example of transporting let us say iron ore that has been dug up the mines from various mines or sites and how to transport it to processing sites or factories. So, imagine that there are m sites one through m at which let us say some material is dug up or iron ore has been dug up and heaped up.

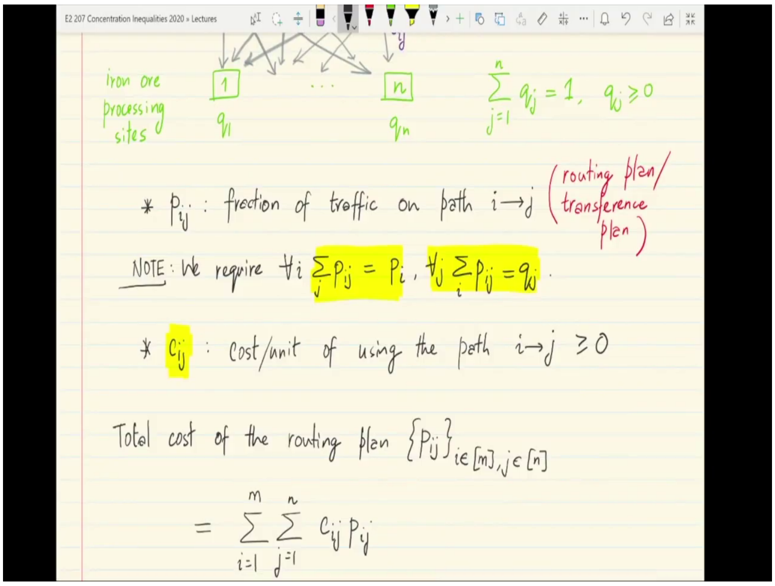
And essentially, the number of units of iron ore available at each site i is p i. So, p 1 through p m are the amounts of iron ore that are being generated or being dug up let us normalize things such that the summation of all these p i's is = 1 ok so, there is a total of one unit of ore that has been mined.

On the other hand, you have a bunch of n processing sites or factories each of which can process or mine or you know turn the iron ore into finished products at capacity q j ok. So, the capacity of the site j is q j so, in some sense, the demand from site q j for iron ore is q j is q j.

And since there is only a total of 1 unit totally being mined, let us also assume that the total amount of processing demand that is required for iron ore is = 1 ok. So, there are m numbers that sum to 1, m non-negative numbers summing to 1, one for each site on the top and there are n numbers, non-negative numbers that are also summing to 1 representing the demands ok.

So, the problem here is to essentially find a plan or a way to move iron ore from each site 1 through m to each site 1 through n if required by splitting the amounts at each site so, we have to come up with an allocation which is denoted by p ij. So, p ij is the fraction of traffic or iron ore that travels on the path i to j and reaches let us say from any i to any j or you can even think of the problem in reverse, there is no directionality here, it can even be going from j to i ok depends on your point of view.

(Refer Slide Time: 13:45)



iron ore processing sites $\boxed{1}$ . . . $\boxed{n}$

q_1 q_n

$\sum_{j=1}^{n} q_j = 1,\ q_j \geq 0$

* p_{ij} : fraction of traffic on path $i \to j$ (routing plan/ transference plan)

NOTE: We require $\forall i \sum_j p_{ij} = p_i$, $\forall j \sum_i p_{ij} = q_j$.

* c_{ij} : cost/unit of using the path $i \to j \geq 0$

Total cost of the routing plan $\{p_{ij}\}_{i \in [m], j \in [n]}$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} p_{ij}$$

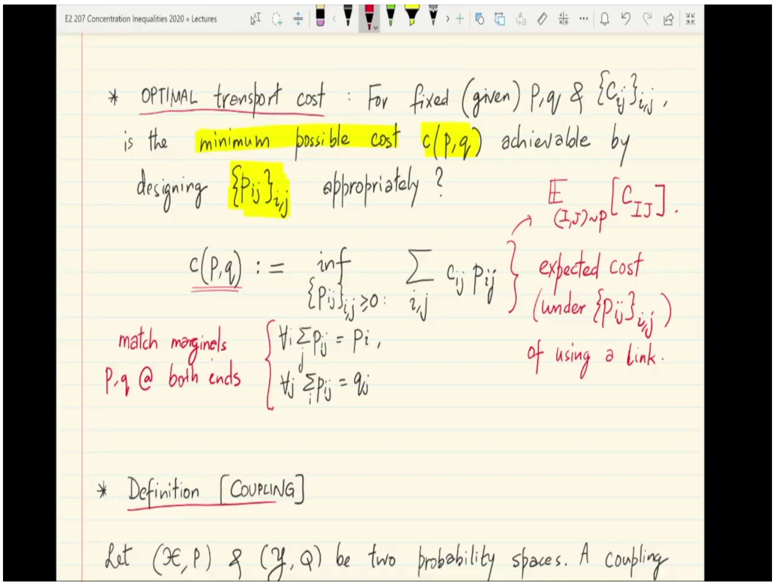
So, p ij is the amount of traffic out of the total p i that goes along the path i to j and naturally, in order to satisfy the exact demands at the receiving end which is the green end as well as the conservation of mass at the sending end, we require the following constraints to hold, equality constraint. So, for every source site i, the total amount of material leaving that site which is the sum of p ij overall j must equal p i and likewise the total amount of material entering each destination site which is the summation over all i of p ij must be = the total demand must satisfy the total demand q j at that; at that site.

So, essentially, this is a condition on p ij that says if you sum along, if p ij were a matrix of ij’s, the row sums and the column sums have been prescribed ok all that is left is for someone to design the exact numbers p ij for every i and j, the p ij is often called a routing plan or a transference plan ok and how can we design different transference plans?

Suppose we assume that there is a cost c ij to transporting material along the path ij. So, this could basically perhaps be determined by the distances of j and i if they are all aligned together in some space, it could depend on the route length and so on. But imagine that there is some c ij given that is a non-negative number which is a cost per unit of using the path i to j, ij.

So, every piece of material that travels along if a unit amount of material travels along the path i, j, then the cost of making a travel along the path I, j is going to be c ij. So, this implies that if we have a routing plan p ij which of which of course, satisfies the marginal constraints at both ends and the sum over all j in the sum over all i, then the total cost that this routing plan gives with respect to the given costs c and the given source and target distributions p and q is just obtained by summing over all i and all j or c ij, pij. So, this is the total cost of the routing plan ok.

(Refer Slide Time: 16:06)



* OPTIMAL transport cost : For fixed (given) p, q & $\{c_{ij}\}_{i,j}$, is the minimum possible cost $c(p,q)$ achievable by designing $\{p_{ij}\}_{i,j}$ appropriately ?

$$\underline{\underline{c(p,q)}} := \inf_{\{p_{ij}\}_{i,j} \geq 0 :} \sum_{i,j} c_{ij} p_{ij}$$

expected cost (under $\{p_{ij}\}_{i,j}$) of using a link.

$\mathbb{E}_{(I,J)\sim p}[C_{IJ}]$.

match marginals p, q @ both ends $\left\{ \begin{array}{l} \forall i \sum_j p_{ij} = p_i , \\ \forall j \sum_i p_{ij} = q_j \end{array} \right.$

* Definition [COUPLING]

Let $(\mathcal{X}, P)$ & $(\mathcal{Y}, Q)$ be two probability spaces. A coupling

And one is essentially in the optimal transportation cost problem, interested in finding the allocation or transference plan p ij to minimize the total cost as best as possible. So, the optimal transportation cost problem says that for fixed p and q, fix distributions p and q on the i and j terminals and a given cost structure c ij for every i and j.

So, the question is; can how do you achieve the minimum possible cost c p, q denoted by c p, q achievable by careful design of the p ij ok. So, the optimal transportation cost problem is all about how one tries to design the p ij’s to actually minimize the total transport cost.

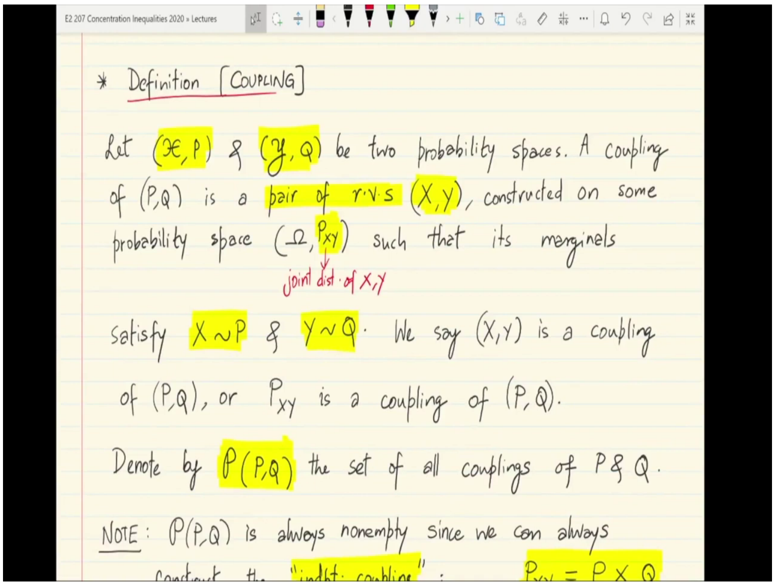
So, this leads one to formulate the following optimization problem which is defined by this number c p, q. So, c p, q is basically the least possible total transportation cost of moving all the material from the i terminals to the j terminals while satisfying on one hand, two types of constraints which are essentially kind of constraints about the marginal.

So, if you can imagine p ij as a distribution on you can imagine p ij as a distribution on edges i, j ok. So, p ij is a probability in some sense for every i, j and the row sums and column sums for p ij essentially are marginal conditions. So, this condition says you have to match the marginals or marginal distributions p and q at both ends while designing this joint distribution p ij

And the objective function itself now, can look looks like essentially a expected cost under p ij so, this is an expected, this has the interpretation of an expected cost under the distribution p ij ranging over all i and all j of using a link or using a link or a route ok. So, think of capital I , capital J being drawn from the distribution p ij and c of capital I , capital J would be the random cost of using that link. So, this essentially lays down an expected cost objective ok.

So, by this so, one can also write this objective as the expected value of I , J, capital I , capital J being drawn from p of c this notation its C capital I capital J ok. So, the aim is to minimize is to find a distribution p ij to try to minimize its expectation under a given cost structure right.

(Refer Slide Time: 20:02)



* Definition [COUPLING]

Let $(\mathcal{X}, P)$ & $(\mathcal{Y}, Q)$ be two probability spaces. A coupling of (P,Q) is a pair of r.v.s (X,Y), constructed on some probability space (Ω, P_{XY}) such that its marginals satisfy $X \sim P$ & $Y \sim Q$. We say (X,Y) is a coupling of (P,Q), or P_{XY} is a coupling of (P,Q).

$\underset{\text{joint dist. of } X,Y}{\uparrow}$

Denote by $\mathcal{P}(P,Q)$ the set of all couplings of P & Q.

NOTE: $\mathcal{P}(P,Q)$ is always nonempty since we can always construct the "indpt. coupling": $P_{XY} = P \times Q$

So, how does in some sense one go about solving such optimal transportation cost problems? One solution is given by the principle of coupling some probability theory. So, what is coupling here? Let us review this definition.

So, if you are given two probability spaces x let us say calligraphic x with the measure P and calligraphic y with measure Q. If they are two probability spaces, then a coupling of P , Q or a coupling of random variables to distributions P and Q is an appropriately constructed pair of random variables X and Y constructed on some probability space omega with some distribution, with some having its own measure let us say P XY such that its marginals satisfy X distributed according to P and Y distributed according to Q ok.

So, in short, if you are given two probability P and Q, any construction of joint random variables X and Y whose marginals are respectively X distributed according to P and the Y marginal is Q is called a coupling of the measures P and Q or the distributions P and Q.

So, notation for this is that we say X , Y is a coupling of P , Q or P XY is a coupling of these two distributions, the joint distribution P XY is a coupling of the individual, the tuple of marginals P and Q. We will also find it very convenient to denote by calligraphic P ok; calligraphic P of P , Q denotes the set of all possible couplings of P and Q.

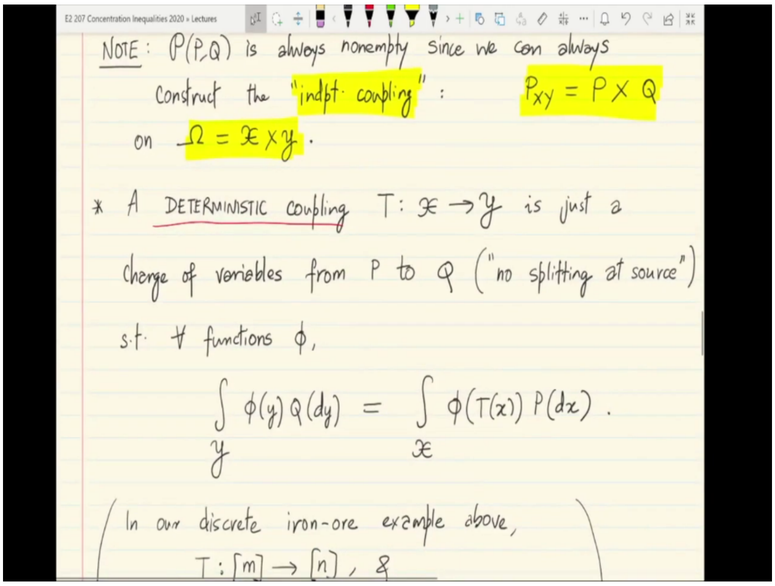
So, this is an abstractly defined set, but anytime that you can always, anytime that you can define a probability space with a joint distribution for two random variables with the individual marginal distributions being P and Q. Any such construction in some sense is assumed to belong to this calligraphic P P , Q.

So, by the way, it is easy to see that this calligraphic P of P , Q is always non-empty. So, one can in particular always construct a very trivial coupling between two measures P and Q which is the product distribution of P and Q so or the independent coupling.

So, you can just define on the probability space calligraphic x cross calligraphic y, the product measure P XY which is the product measure of P and Q and that trivially satisfies the marginal of X or the distribution of X being P and the distribution of Y being Q. So, this set of all couplings is certainly non-empty, you do not have to worry about the emptiness of it ok.

So, this is essentially the idea of coupling and it is often used to great effect in many areas of applied probability and stochastic processes and we will actually be using it to understand the connection to transportation cost problems and ultimately to concentration of measures inequalities ok.

(Refer Slide Time: 23:12)



NOTE: $\mathcal{P}(P,Q)$ is always nonempty since we can always construct the "indpt. coupling": $P_{XY} = P \times Q$ on $\Omega = \mathcal{X} \times \mathcal{Y}$.

* A DETERMINISTIC coupling $T : \mathcal{X} \to \mathcal{Y}$ is just a change of variables from P to Q ("no splitting at source") s.t. $\forall$ functions ϕ,

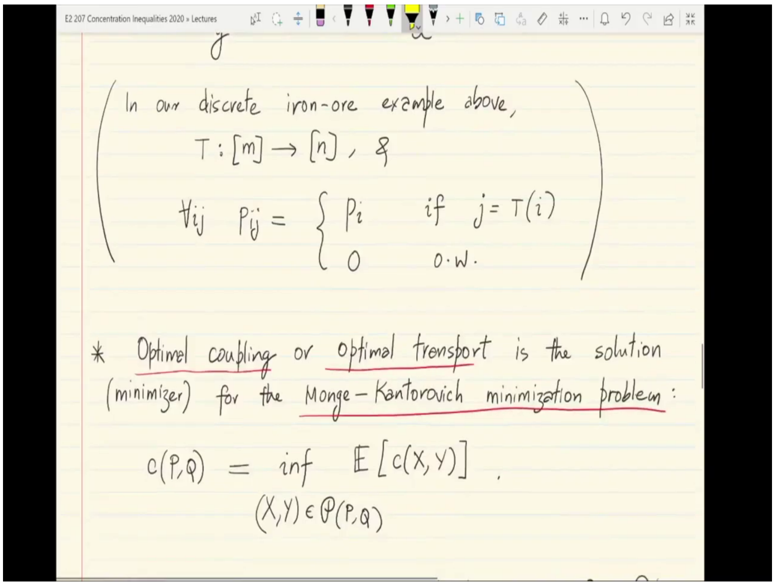
$$\int_{\mathcal{Y}} \phi(y)\, Q(dy) = \int_{\mathcal{X}} \phi(T(x))\, P(dx).$$

In our discrete iron-ore example above, $T : [m] \to [n]$, &

So, in particular, there are also objects called deterministic couplings which are just couplings with certain extra properties. So, in particular, a deterministic coupling T from x to y is you can think of it in different ways, it is just essentially a change of variables from P to Q ok. So, in our transportation cost problem, it essentially represents any transference plan that does not split material at source or destination. So, every the amount of iron ore at every site i is uniquely destined to go to a certain other site j and so on ok and vice versa.

So, this is just a change of variables from P to Q or in other words, for all functions, for all nice enough functions phi, you have that the integral; the integral over y of phi y Q dy is the same as the integral over x of phi of T x, T of x is essentially where x gets sent to in the y domain integrated against the distribution P.

(Refer Slide Time: 24:16)



In our discrete iron-ore example above,

$T : [m] \to [n]$, &

$$\forall ij \quad p_{ij} = \begin{cases} p_i & \text{if } j = T(i) \\ 0 & \text{o.w.} \end{cases}$$

* Optimal coupling or optimal transport is the solution (minimizer) for the Monge – Kantorovich minimization problem :

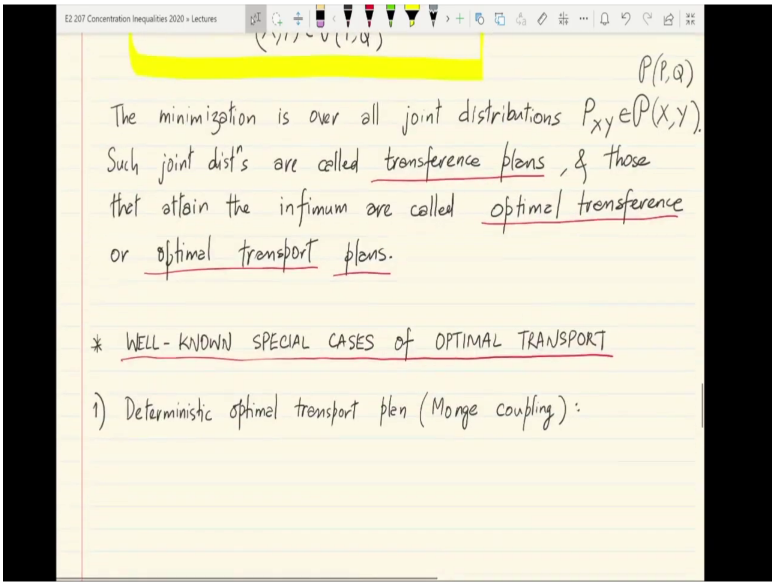
$$c(P,Q) = \inf_{(X,Y) \in \mathcal{P}(P,Q)} \mathbb{E}\left[c(X,Y)\right] .$$

So, in our discrete iron ore example as I said deterministic coupling T would be any function that maps the set of integers one through m to the set of integers one through n such that for all i , j P ij = exactly P i if j is = T of i ok. So, that is what is called a deterministic coupling. It is a deterministic way of mapping any x to any y ok. So, you do not split, you do not have multiple destinations for any of the sites or domain elements of x, each domain element of x is assumed to send material or transport material only to a uniquely identified site in the y domain.

So, optimal coupling or optimal transport is the solution or minimizer for the what is called the Monge-Kantorovich minimization problem which is a very general problem defined in general probability spaces. So, if one is given two measures P and Q, c P, Q is defined to be the minimum or more generally the infimum of X of the expected value of c of X , Y.

So, if c is a given function on X cross Y which is the transportation cost function, then the Monge-Kantorovich minimization problem asks you to find the minimizer of expected value of c X, Y over all possible joint distributions of X and Y that respect the given marginals P and Q ok. So, this is the optimal transportation cost problem and there is an entire branch called optimal transportation theory that studies the existence and properties of solutions of the Monge-Kantorovich minimization problem.

(Refer Slide Time: 26:21)



$\mathcal{P}(P,Q)$

The minimization is over all joint distributions $P_{XY} \in \mathcal{P}(X,Y)$.
Such joint distⁿ's are called <u>transference plans</u>, & those that attain the infimum are called <u>optimal transference</u> or <u>optimal transport plans</u>.

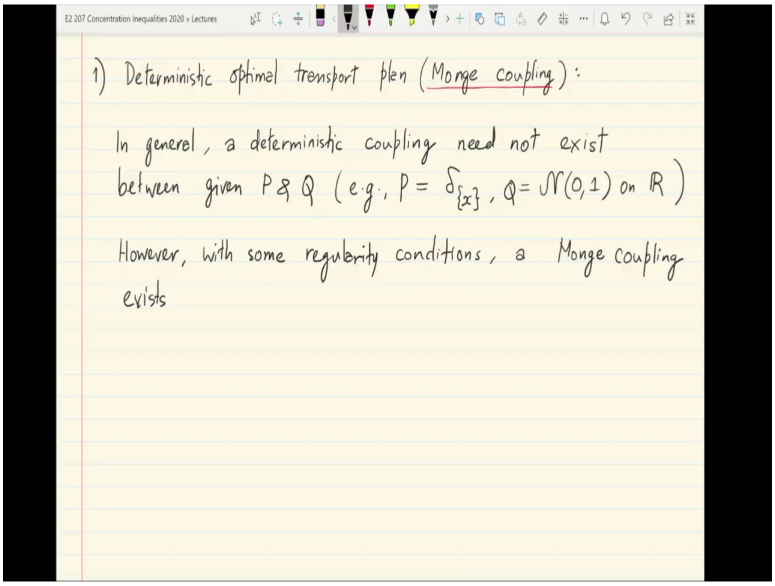
* WELL-KNOWN SPECIAL CASES of OPTIMAL TRANSPORT

1) Deterministic optimal transport plan (Monge coupling):

The minimization is over all joint distributions P XY lying in the set of couplings of X of X and Y in fact, sorry this is not Q, but this is Y or P, Q depending on the notation that you choose to adopt and such joint distribute, any such joint distribution is or a coupling is called a transference plan and those that attain the infimum are called optimum transference or optimal transport plans.

So, this is a every general and abstract problem that has been studied for a lot of time in analysis and probability, but let us look at some well-known special cases of the optimal transportation problem in terms of the cost structures c ok.

(Refer Slide Time: 27:28)



1) Deterministic optimal transport plan (Monge coupling):

In general, a deterministic coupling need not exist between given P & Q (e.g., $P = \delta_{\{x\}}$, $Q = \mathcal{N}(0,1)$ on $\mathbb{R}$)

However, with some regularity conditions, a Monge coupling exists

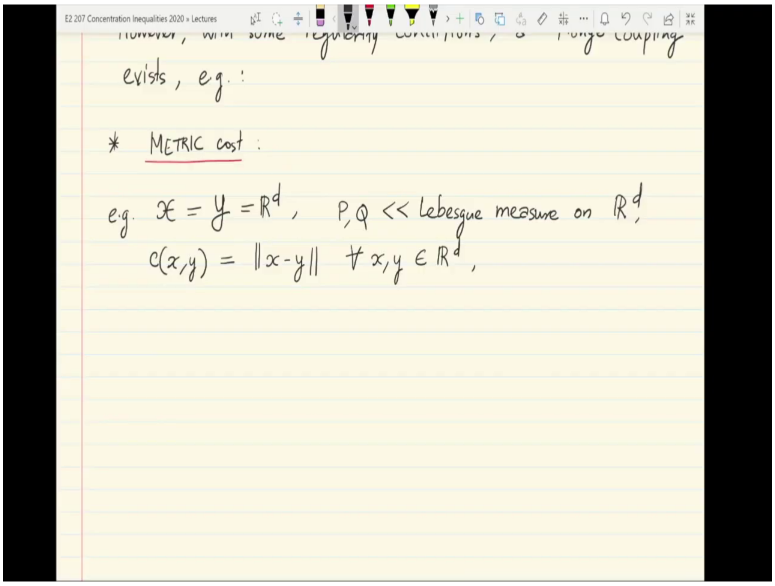
So, the first example that I would like to present here is what is called the existence of a deterministic optimal transport plan or a Monge coupling. So, recall deterministic optimal transport plan is just a plan that sends material from every location i uniquely to a to a unique destination j ok.

So, in general so, we begin with a remark saying that in general, one cannot hope to find a deterministic coupling between two given marginals P and Q. So, in general, a deterministic coupling, or an optimal transport plan need not exist; need not even exist between given marginals, marginal distributions P and Q.

So, as a simple example, you can just take P so, let us just take the space as the set of space of all real numbers and if you have P as a direct delta probability distribution supported on some point x and Q as being a very nice absolutely continuous with respect to Lebesgue measure distribution let us say a Gaussian measure on R, then there is clearly no optimal transport plan; there is no clear there is no deterministic coupling that sends that can basically send probability mass all of which lies at x in the P side to a unique location on the Q side because Q insists that the probability mass must be spread all over the place across real numbers.

So, there is no way of mapping in a deterministic way sending material from P to Q ok. So; so, that is what the statement says. So, however, with some appropriate regularity conditions, a Monge’s coupling actually can seem to be exist; to seem to exist. So, however, with some regularity conditions or smoothness conditions on the probability measures P and Q, a Monge coupling which is essentially a deterministic transport plan, plan from x to y exists.

(Refer Slide Time: 29:55)



However, with some regularity conditions, a Monge coupling exists, e.g.:

* METRIC cost :

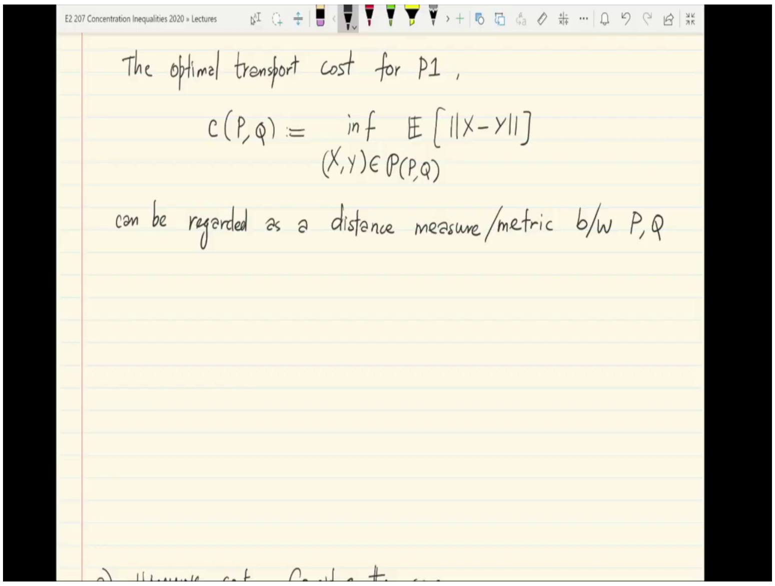
e.g. $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $P, Q <<$ Lebesgue measure on $\mathbb{R}^d$;

$c(x,y) = \|x-y\| \ \forall x,y \in \mathbb{R}^d$,

So, here is a set of regularity conditions as an example that makes this possible and this is again a famous setting where we have a metric cost. So, what do we mean by this? Let us suppose that the x and y domains are both let us say R d so, if both of these are R d and P and Q are both given to be the marginals P and Q and R d are both given to be absolutely continuous measures with respect to Lebesgue measures so, there are two probability densities ok.

So, let us say P and Q are two; two given probability densities with the cost structure c x , y between two points in R d just being the standard norm difference between x and y, this can be any norm by the way defined on R d.

(Refer Slide Time: 31:08)



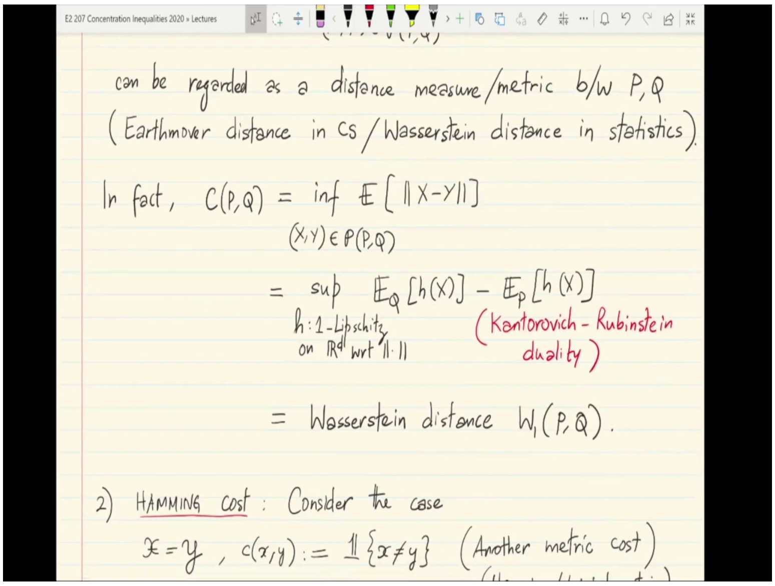
The optimal transport cost for P1 ,

$$c(P,Q) := \inf_{(X,Y)\in \mathbb{P}(P,Q)} \mathbb{E}\left[\|X-Y\|\right]$$

can be regarded as a distance measure/metric b/w P, Q

Then, the optimal cost, the optimal transportation cost so, let us call this setting as P1. So, the optimal transport cost for P1 which is defined, which we defined as c P, Q being the inf over X , Y in the set of all couplings of P, Q of expected value of non-difference between; non-difference between X and Y. So, in fact, this turns out to be actually a measure of distance or a metric on the set of all probability distributions P and Q. So, this can actually can be regarded as the distance measure or a metric between probability distributions P and Q. So, the smaller this is you can think of P and Q being more similar.

(Refer Slide Time: 32:25)



can be regarded as a distance measure/metric b/w P, Q
(Earthmover distance in CS / Wasserstein distance in statistics)

In fact, $C(P,Q) = \inf_{(X,Y)\in \mathcal{P}(P,Q)} \mathbb{E}[\|X-Y\|]$

$= \sup_{\substack{h: 1\text{-Lipschitz} \\ \text{on } \mathbb{R}^d \text{ wrt } \|\cdot\|}} \mathbb{E}_Q[h(X)] - \mathbb{E}_P[h(X)]$ (Kantorovich - Rubinstein duality)

$=$ Wasserstein distance $W_1(P,Q)$.

2) HAMMING cost : Consider the case

$\mathcal{X} = \mathcal{Y}$, $c(x,y) := \mathbb{1}\{x \neq y\}$ (Another metric cost)

And in fact, this is also called by other names such as the earthmover distance which is commonly used in computer science or the alternative name is the Wasserstein distance in statistics ok. In fact, one gets this nice duality principle here where you can express C P, Q as both a minimization problem and an optimizer and a maximization problem.

So, recall that we defined the C P, Q as the inf of the expected value of X - Y, the norm difference between X and X - Y where X , Y are coupled according to P , Q and it turns out that you can also write this in a dual way, it is the supremum over all, it the supremum over E Q h X - E. P h X over all functions h that are 1 Lipschitz continuous on R d with respect to the norm, whatever norm you take.

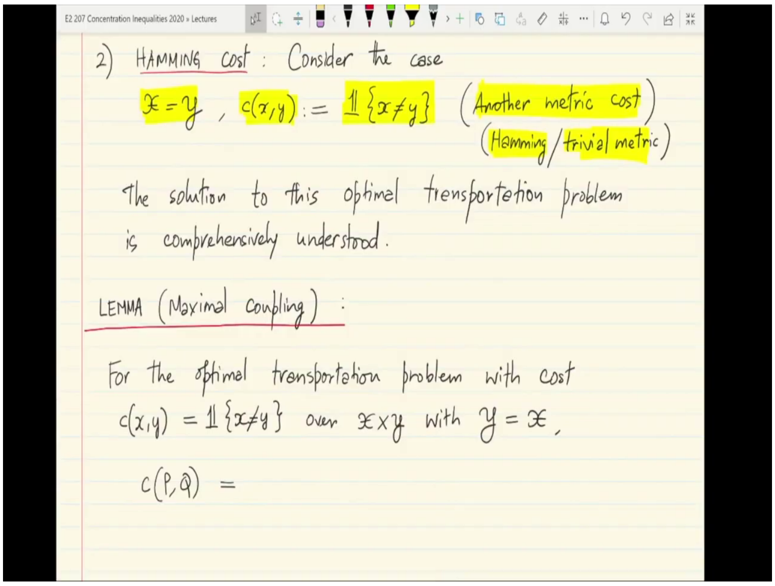
So, in fact, this is what is called Kantorovich duality or Kantorovich-Rubinstein duality ok, this has the following really nice form which in some sense is connected to what we want to do with deriving concentration inequality. So, this is the first time you are seeing the appearance of functions like the expected value of a certain random variable with respect to two different probability measures P and Q and this is also called the as we said this is called the Wasserstein 1 distance between probability measures P and Q ok.

So, just a side note, the Kantorovich-Rubinstein duality at the simplest level essentially says that the solution of a certain infimization problem can also be represented as a solution of certain supremization or maximization problem and one of the simplest ways of seeing this is in the discrete setting. So, in the discrete setting, if we recall what our problem was, it was minimizing the total because a minimization of the following cost which is linear over the variables P ij ok so, the objective function is linear over P ij's and de-constraints are also linear over the P ij's ok.

So, this is actually a nice linear program, the solution for c p, q or the optimal transportation costs in the discrete case and as we know, you can always take so, every linear program has a dual linear program and in fact, interpreting the dual linear program appropriately is what gives you the Kantorovich Monge duality result.

So, it works to much greater generality than just the discrete case and so, we have results of this form that if you define a stand if you use a standard metric as the cost function in your two spaces, then the optimal transport cost turns out to be related to another dual problem which essentially for who's the, for whom the set the domain is the set of all Lipschitz functions between those two spaces.

(Refer Slide Time: 36:30)



2) HAMMING cost : Consider the case

$\mathcal{X} = \mathcal{Y}$, $c(x,y) := \mathbb{1}\{x \neq y\}$ (Another metric cost)

(Hamming / trivial metric)

The solution to this optimal transportation problem is comprehensively understood.

LEMMA (Maximal coupling) :

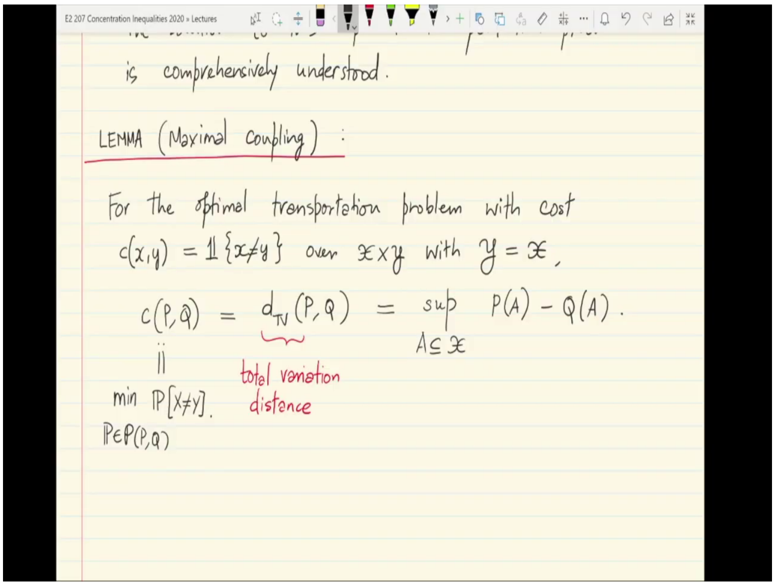
For the optimal transportation problem with cost $c(x,y) = \mathbb{1}\{x \neq y\}$ over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \mathcal{X}$,

$c(P,Q) =$

The second example which is very well studied of transportation cost problems is the one of the hamming cost. So, let us consider the case where you want to transport from two given measures on the same space that is script x = script y and the cost function between x and y is just 1 if x is not = y and 0 otherwise. So, each time that you transport material, any material, any amount of material from a site x to a site y which is not x, you end up paying me a cost you end up paying a unit one transportation cost of one per unit material.

So, this is in fact, another metric cost indicator x naught = y is what is called the trivial metric over the space x or the hamming metric ok. So, imagine that you have the spaces being the same and you have two given distributions p and q on x and you want to try to find the coupling between x between P and Q that minimizes the expected value of the indicator that x is not = y.

(Refer Slide Time: 38:26)



is comprehensively understood.

LEMMA (Maximal Coupling) :

For the optimal transportation problem with cost $c(x,y) = \mathbb{1}\{x \neq y\}$ over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \mathcal{X}$,

$$c(P,Q) = \underbrace{d_{TV}}_{\text{total variation distance}}(P,Q) = \sup_{A \subseteq \mathcal{X}} P(A) - Q(A).$$

$$\parallel$$

$$\min_{\mathbb{P} \in \mathcal{P}(P,Q)} \mathbb{P}[X \neq Y].$$

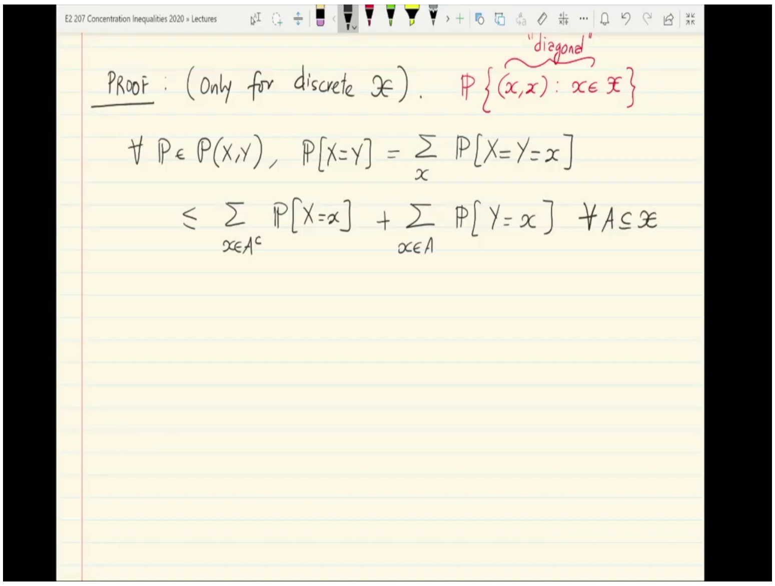
So, the solution to this optimal transportation problem has been well studied and understood and in fact, can be it can be seen by the following result that we will prove, this is called by various names one of them being the maximal coupling lemma which we will also use in the subsequent lecture.

So, for the optimal transportation problem with a cost structure given by the hamming distance or the hamming metric over x cross y with y = x, we have that the solution, the optimal transportation cost is nothing, but the total variation distance between P and Q ok which we will call d TV. So, d TV is the classic total variation distance measure, which is defined, which can be defined again in various ways, one of the most common ways to define it is the supremum over all possible subsets of the space or between the so, of the difference between the probability A as measured by P and the probability of A as measured by Q ok.

So, thanks to this lemma, we also know that c P, Q, the total variation distance as an has the flavor of a minimization problem as well because c P, Q was defined to be the minimum over all P which as a coupling of P and Q of the probability that. So, expected value of the indicator that x is not = y so, it is the probability that x is not = y over all joint distributions P with given marginals P and Q.

So, the content of the lemma essentially says that the total variation distance is actually the solution of a transportation cost problem where the cost structure is given by indicator x naught = y and in fact, the name maximum maximal coupling arises as we will see in the proof of the lemma by actually explicitly constructing in its a coupling between P and Q that achieves the d TV, the total variation distance ok.

(Refer Slide Time: 40:23)



PROOF : (Only for discrete $\mathcal{X}$). $\mathbb{P}\{\overbrace{(x,x)}^{\text{"diagonal"}} : x \in \mathcal{X}\}$

$$\forall\, \mathbb{P} \in \mathcal{P}(X,Y),\ \mathbb{P}[X=Y] = \sum_{x} \mathbb{P}[X=Y=x]$$

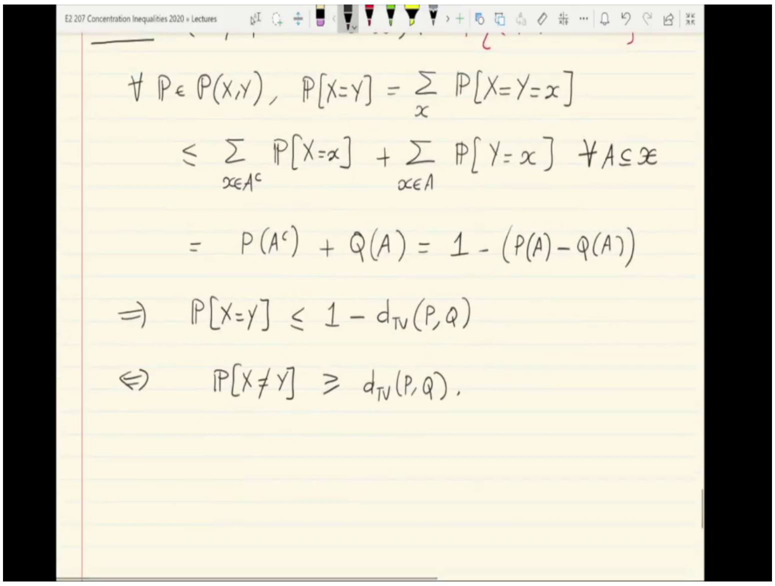
$$\leq \sum_{x \in A^c} \mathbb{P}[X=x] + \sum_{x \in A} \mathbb{P}[Y=x] \quad \forall A \subseteq \mathcal{X}$$

So, let us move on to the proof which is a rather elegant proof. So, for this for the reason for reasons of simplicity, we will only show the proof for discrete spaces x. The general case can be shown with some more technical effort. In fact, we can assume that x is finite just for the sake of simplicity.

So, the first step is to say that for every so, for every possible coupling P of X and Y or P and Q be probability that X = Y is exactly the sum over all x in script x or calligraphic x of the probability that capital X = capital Y = small x. So, it is basically the sum of probabilities of pairs that lie on the diagonal ok. So, the set x , x so, it is the probability of the set x , x where x ranges in x and this is called the diagonal, is often called the diagonal in the joint for the joint probability distribution.

And one can split the sum over x in all x all script x over x in any set A and the complement so, doing that gives raise to the inequality. So, if you just take the take x running over A complement of this expression and drop the condition that Y has to be = x just get probability that X = x and likewise, the remainder of the sum is x lying in a of the probability that you insist that only Y be = x and this holds for any subset A of x ok.

(Refer Slide Time: 42:33)



$$\forall\, \mathbb{P} \in \mathcal{P}(X,Y),\quad \mathbb{P}[X=Y] = \sum_{x} \mathbb{P}[X=Y=x]$$

$$\leq \sum_{x\in A^c} \mathbb{P}[X=x] + \sum_{x\in A} \mathbb{P}[Y=x] \quad \forall A \subseteq \mathcal{X}$$

$$= P(A^c) + Q(A) = 1 - (P(A) - Q(A))$$

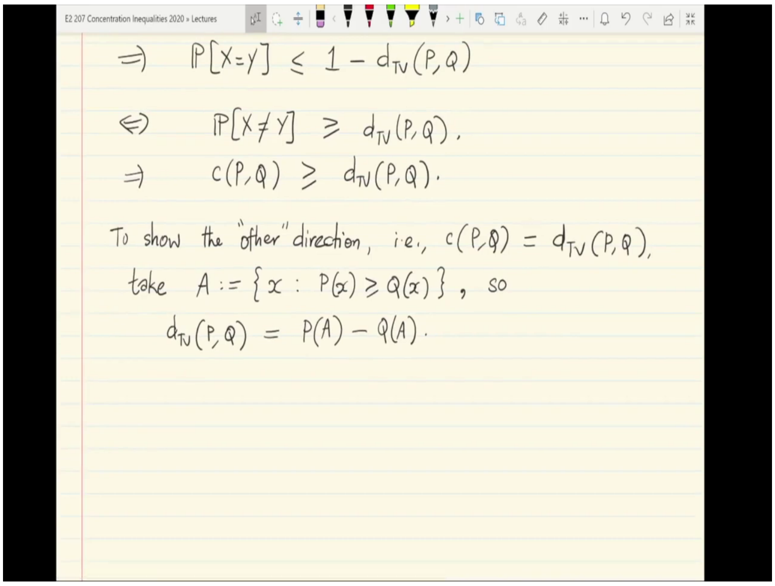
$$\Rightarrow \quad \mathbb{P}[X=Y] \leq 1 - d_{TV}(P,Q)$$

$$\Leftrightarrow \quad \mathbb{P}[X\neq Y] \geq d_{TV}(P,Q).$$

And now, by the definition of coupling the P, the bold P of X = x is simply the regular P of a A compliment, P of A compliment, the second term is by definition of the coupling it is Q of A and this is 1 - P A - Q A ok. So, P of x naught x being = y is upper bounded by any such expression for all A in x.

And in particular, one can basically take the A, the subset A that gives you the tightest upper bound or the smallest upper bound which is to say take the set A that essentially reaches the supremum in a definition of total variation distance and so, this implies that the probability that X = Y is at most 1 - the total variation distance between P and Q, this is just the same as saying that probability that X not = Y is lower bounded by d TV of P, Q. So, this gives us one side of the result that we wanted, we essentially want to show that this holds with equality so, we obtain the greater than or = part ok.

(Refer Slide Time: 43:51)



$$\Rightarrow \quad \mathbb{P}[X=Y] \le 1 - d_{TV}(P,Q)$$

$$\Leftrightarrow \quad \mathbb{P}[X\neq Y] \geq d_{TV}(P,Q).$$

$$\Rightarrow \quad c(P,Q) \geq d_{TV}(P,Q).$$

To show the "other" direction, i.e., $c(P,Q) = d_{TV}(P,Q)$,

take $A := \{x : P(x) \geq Q(x)\}$, so

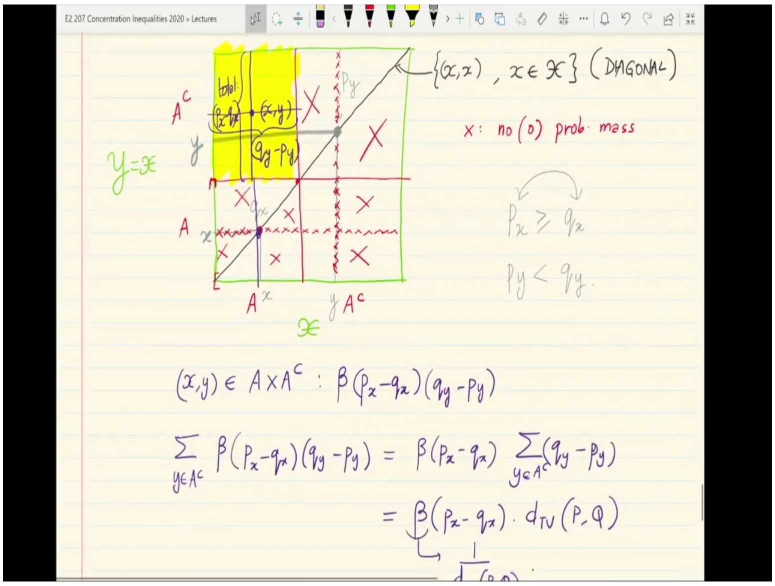
$$d_{TV}(P,Q) = P(A) - Q(A).$$

So, this is just the same as saying c P, Q. So, for any such coupling since there is a lower bound of d TV P, Q it means that c P, Q is lower bounded by d TV P, Q, the d hamming cost. So, we want to essentially show the other direction as well that c P, Q is ≤d TV P, Q so, for that, we will actually construct an explicit coupling.

So, to show the other direction that is the achievability, we will actually show c P, Q = d TV P, Q. Let us start by taking the subset A as the set of all elements x where P is larger than Q, P is at least Q ok. So, it is easily seen that in this situation, the total variation distance between P and Q is simply the size of A as measured by P - the size of A as measured by Q. So, this gives us an expression for d TV P, Q in terms of this distinguished set A.

And now, we will set about constructing a coupling that a joint distribution between P and Q or between X and Y whose x marginal is P and whose Y marginal is Q and which essentially minimizes the probability that X and Y are different in the joint sense.

(Refer Slide Time: 45:57)



$$(x,y) \in A \times A^c : \beta(p_x - q_x)(q_y - p_y)$$

$$\sum_{y \in A^c} \beta(p_x - q_x)(q_y - p_y) = \beta(p_x - q_x) \sum_{y \in A^c} (q_y - p_y)$$

$$= \beta(p_x - q_x) \cdot d_{TV}(P, Q)$$

So, for doing that, its the operation is quite naturally is illustrated by drawing a figure so, let me draw the following figure where so, I will; I will draw a two-dimensional figure. So, let the square on the xy plane represent the cartesian product of so, this is x. So, the x axis is the space x and the y axis is the space y which is assumed to be = x, but think of it as y.

And we can so, let us assume that we have arranged the elements of x and in the same order arranged it along y as well such that the first block of elements represents the set A ok. So, the set A where P is always larger than Q and the remainder is A complement. So, imagine that the elements of x are arranged in order, in some order such that the elements in A come first and then, the elements in A complement come later and you follow the same ordering along the y; y axis as well. So here lies A and here lies A compliment ok.

And imagine lining up all these elements and then, the square represents the cartesian product of all pairs x , y where x belongs to script x and y belongs to script y. So, we have this diagonal here ok. So, this diagonal is essentially the set of all x , x where x ranges over big script x so, that is a diagonal in the space.

Now, what we want to do? So, really there are imagine there are only finitely many points in x on the x and y axis so, this is a finite cartesian product so, it is a grid of finitely many points. So, what we want to intuitively do is to define for every grid point on the square, a probability number P of x , y ok such that two properties must hold, one is that the sum along all vertical lines must be = the vertical the horizontal distribution P x and the sum over all horizontal lines must be = the given number Q y ok.

So, let us in so, let me demarcate four important quadrants here ok. So, these are four corners defined by the cartesian products of A with A, A with A compliment, A compliment with A and A compliment with A compliment. So, all we have to do is design a distribution on this entire plane. So, distribute probability mass across this entire square such that the projection of that probability mass on the x axis gives you the marginals p x and, on the y-axis, gives you the marginals p y.

Now, what do we have to intuitively do to minimize the transportation cost, we have to try to keep x and y as equal as possible ok so that intuitively translates to trying to put as much probability on the diagonals as we can ok. So, think of it as if you put, if you are able to put as much probability on the diagonal as you can, then you effectively minimize the chance that x and y are different or not lying on the diagonal.

So, let us consider a point. So, let us consider a point x here on the diagonal or x , x on the diagonal ok. So, let us consider a point x , x on the diagonal. So, suppose we want to put the largest possible probability on this diagonal point x , x such that the marginal constraints are met, what is the largest probability we can put? So, on one hand, we know that if we sum along the vertical line which passes through x , x, we must get p x and if we sum on the horizontal line that passes through x , x, we must get q x finally, but we know that p x is ≥ q x so, the theoretical maximum probability that one can put at the point x , x is limited by the smaller one out of p x and q x which in this case happens to q x.

So, what we can do is we can put a probability mass of q x on each such diagonal point x , x where x belongs to A ok. The moment we put such a probability q x on this diagonal point x , x, the horizontal sum already has q x so, there can be no more probability on the vertical strip passing through x , x at any other point apart from x ,x.

So, what it says is that it rules out so, by x is I mean that there can be no probability mass sitting at any of these points along this horizontal line ok. So, x means that there is no or 0 probability mass because you have exhausted all the q x by placing it exactly at the diagonal point. So, you cannot put any more probability mass along this vertical line passing through x , x ok.

And this essentially holds for you can do this construction for every x in the set A and x , x on the diagonal. So, that effectively says that you know apart from this diagonal, there can be essentially no probability in this entire part, you cannot put probability mass anywhere here.

By the same token, one can also consider points x , x with x lying in A complement ok. So, in fact, let me call this point y, y , y take a typical y , y on the diagonal where y lies in A compliment. So, here, p y is <is strictly <q y. So, it follows that the largest amount of probability mass you can put here is p y and it also follows that by putting p y, you have already exhausted everything on this vertical strip so, you cannot put anything on this vertical anymore probability mass on all points on this vertical line apart from the diagonal ok.

And so, that essentially rules out by generalizing this y along the entire, by sweeping y along A compliment that you cannot put probability mass anywhere here and anywhere here ok. So, we managed to place probability mass on the diagonals by using the principle of largest possible probability on the diagonal ok. So, ok and so, the only remaining mass so, there is of course, remaining probability mass that we have to place on this figure and it follows that we can only place it that the only space to place it is in this northwest quadrant of this figure which is essentially over x , y such that x lies in A and y lies in A compliment ok.

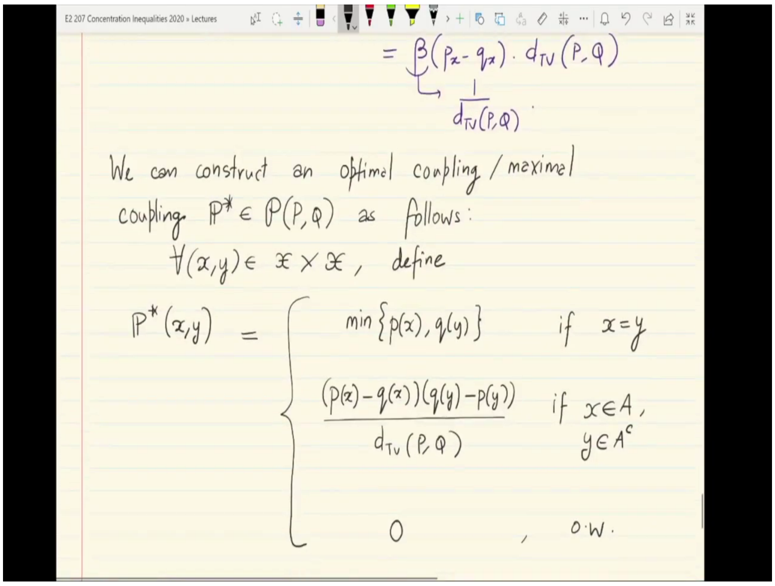
So, what do you have to satisfy here? What let us look at what are the operational constraints here? So, the operational constraints here say that take a typical point here which we will call x , y so, the x coordinate is small x, and the y coordinate is small y. What we must ensure is that if you imagine placing masses at every point in this, every grid point in this highlighted yellow square, then what it means is that along this vertical strip here, in this northwest quadrant, the total mass, probability mass must be the difference between what was left over, what you put on the diagonal and what you totally wanted to put so, that is in fact, the difference between p x - q x.

So, you had a total of p x to put on the vertical diagonal out of which you put q x on the diagonal and so, on the part, apart from the diagonal, there is a remainder of p x - q x that you have to put and similarly, along this word this horizontal part, the total mass in the horizontal part must be q y - p y. So, for every such horizontal and vertical line passing through x and y in the northwest quadrant, the distribution must be such that you should find p x - q x on the horizontal part and q p; p y - q y as a total on the vertical line segment ok.

So, this is in fact, now just it is like a product constraint. So, the simplest way of achieving this is just by putting let us say at each point x , y so, for x and y belonging to A cross A complement that is the northwest quadrant perhaps we can put a mass which is proportional to what we want there ok, proportional to the product of what we want there, we want p x - q x on the vertical segment passing through it and q y - p y on the horizontal segment passing through it and we have to multiply by some constant here ok beta, that we will find soon.

So, if you put this kind of probability distribution so, take the total mass that you have remaining for the northwest quadrant and put masses at each x, y following this product rule with an appropriate constant beta such that the total mass is exactly takes care of the remaining mass from earlier from after putting on the diagonal points. So, what that means is if we sum over all y in A complement of a beta so, let sum be total vertical let us try to find the total vertical mass here so, that is the summing over y in A compliment beta into p x - q x and q y - p y. What this gives you is beta into p x - q x comes out and you have y in A compliment q y - p y and by definition, this is exactly beta into p x - q x into the total variation distance between p and q ok.

(Refer Slide Time: 57:44)



E2 207 Concentration Inequalities 2020 » Lectures

$$= \underbrace{\beta}_{\frac{1}{d_{TV}(P,Q)}}(p_x - q_x) \cdot d_{TV}(P,Q)$$

We can construct an optimal coupling / maximal coupling $\mathbb{P}^* \in \mathcal{P}(P,Q)$ as follows:

$\forall (x,y) \in \mathcal{X} \times \mathcal{X}$, define

$$\mathbb{P}^*(x,y) = \begin{cases} \min\{p(x), q(y)\} & \text{if } x = y \\ \dfrac{(p(x)-q(x))(q(y)-p(y))}{d_{TV}(P,Q)} & \text{if } x \in A, \\ & \quad y \in A^c \\ 0 & , \text{ o.w.} \end{cases}$$

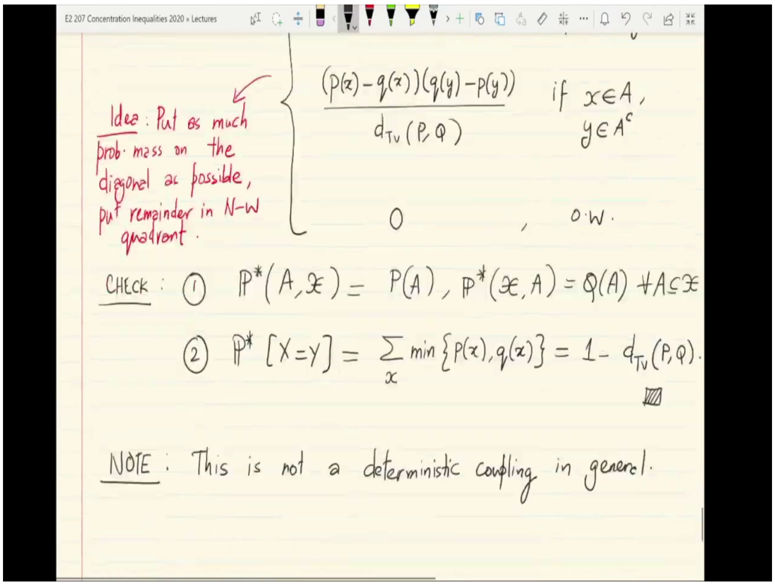
So, in order that this should equal p x - q x so, you can just take beta as 1 over d total variation distance between P and Q and that essentially fixes everything beautifully. So, this was just motivation, this was sort of the natural way of trying to distribute mass to intuitively minimize the amount of probability mass, joint probability mass away off the diagonal and in fact, we can make this formal by saying that we can construct what all of this has resulted in is the statement that we can construct an optimal coupling or a maximal coupling. Notice that there can be several different optimal coupling, this is just one example that we have argued intuitively that can be constructed.

So, let us call this P star which we will show is a coupling of P , Q as follows. So, we are now just going to mathematically and precisely define what we just the procedure that we outlined for putting mass in this 2d plane, on this 2d^2 So, for all x , y which are pairs in this square, a cartesian product of x with x, define P star, the mass sitting at x , y as follows.

So, if x is = y, if it is a diagonal point, then you just put probably the largest possible probability mass there which has been of p x , q y is x = y by our intuition. If you are not on the diagonal, but you are in the northwest quadrant, then you put a probability mass of p x - q x into q y - p y divided by so, into beta which is 1 over the total variation distance between P and Q, if x is in A and y is in A complement and everywhere else you do not put any probability, 0 otherwise ok.

So, this essentially defines a complete probability distribution where no probability mass is wasted over the entire 2D plane; over the entire 2D domain which is the cartesian product of x and x as we just outlined or argued intuitively.

(Refer Slide Time: 60:09)



Idea: Put as much prob. mass on the diagonal as possible, put remainder in N-W quadrant.

$$\begin{cases} \frac{(p(x)-q(x))(q(y)-p(y))}{d_{TV}(P,Q)} & \text{if } x\in A, \\ & \quad y\in A^c \\ 0 \quad , & \text{o.w.} \end{cases}$$

CHECK: ① $\mathbb{P}^*(A,\mathcal{X}) = P(A)$, $\mathbb{P}^*(\mathcal{X},A) = Q(A)$ $\forall A \subseteq \mathcal{X}$

② $\mathbb{P}^*[X=Y] = \sum_x \min\{p(x),q(x)\} = 1 - d_{TV}(P,Q)$.

▨

NOTE: This is not a deterministic coupling in general.

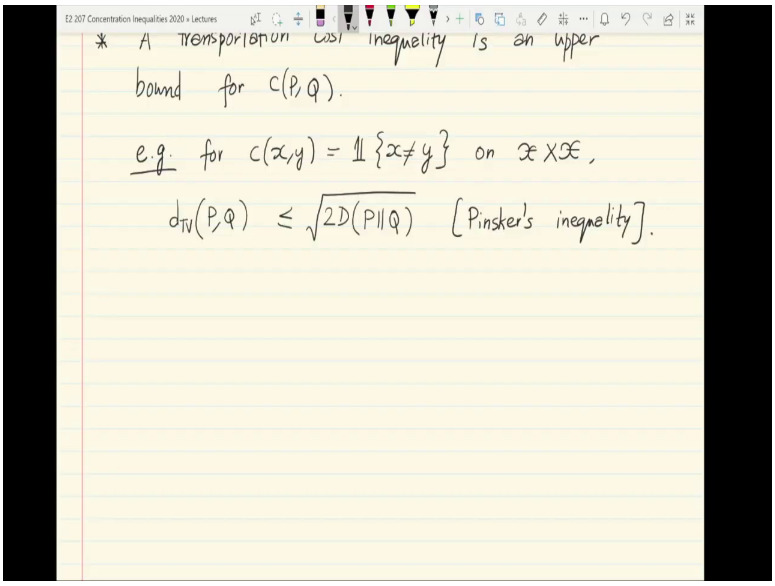
And with this formal definition, I mean it is this is almost true by how we constructed it, but you can; you can mathematically verify the following claims that the right marginal properties are made. So, P star of the set A, the entire of any subset A with the entire x just gives you the marginal over A and P star measured over the rectangle given by the entire space x in the first coordinate and any subset A in the second coordinate gives you a key ok.

This is following our construction, it is almost immediate by what we argued in the geometric figure and more importantly, you can also check that P star X = Y using this definition about of this coupling is exactly so, this is exactly the sum of all masses on the diagonal which we argued that every x , x, you put the minimum of p x and q x and this is easily seen to be 1 - d TV of P, Q.

So, equality holds here that the total variation distance and so, this is exactly one such coupling that achieves equality and proves this lemma ok. So, just so, that you do not get confused, the idea the underlying idea here which is what we drew on the figure is to put as much probability mass on the diagonal as possible.

And this naturally forces you to redistribute the remaining probability mass in the northwest corner, redistribute or distribute or put the remaining mass or the remainder of the probability mass in the northwest quadrant; quadrant ok. So, that is the construction of a maximal couple and you can also see that in general, this is not this maximal coupling that we constructed is not a deterministic coupling in general namely because you know for a point x in A, the p x probability actually was being spread out over this entire vertical strip and that is why it is not in general A deterministic coupling right.

(Refer Slide Time: 63:11)



* A Transportation cost inequality is an upper bound for $c(P,Q)$.

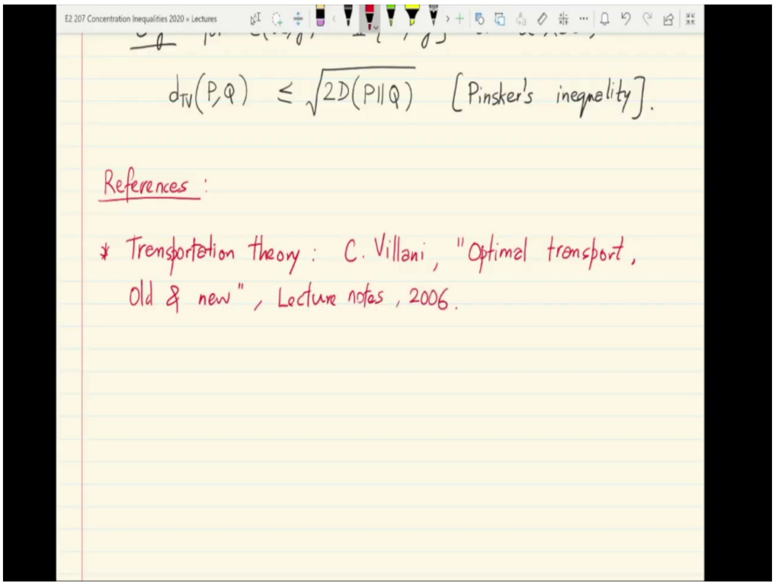
e.g. for $c(x,y) = \mathbb{1}\{x \neq y\}$ on $\mathcal{X} \times \mathcal{X}$,

$$d_{TV}(P,Q) \leq \sqrt{2D(P\|Q)} \quad \text{[Pinsker's inequality]}.$$

So, we now return to; so, we now return to the relationship between transportation cost inequalities and we will move towards the relationship with the concentration of measure bounds. So, we will just conclude this lecture right now by saying that A transportation cost inequality. So, in general what is called a transportation cost inequality is simply an upper bound for the optimal transportation costs c P, Q is just an upper bound for c P, Q ok.

So, a famous example of such an upper bound for c P, Q is again for this hamming distance. So, for c x, y, the cost structure given by the hamming metric or the trivial metric between x and y on x cross x, we already know that the optimal transportation cost with this cost structure is the total variation distance between P and Q and one upper bound on d TV is what is given by Pinsker's inequality which is 2 times the KL divergence between P and Q. So, this is called in information theory as Pinsker's inequality which we will see and prove in the next lecture.

(Refer Slide Time: 64:44)



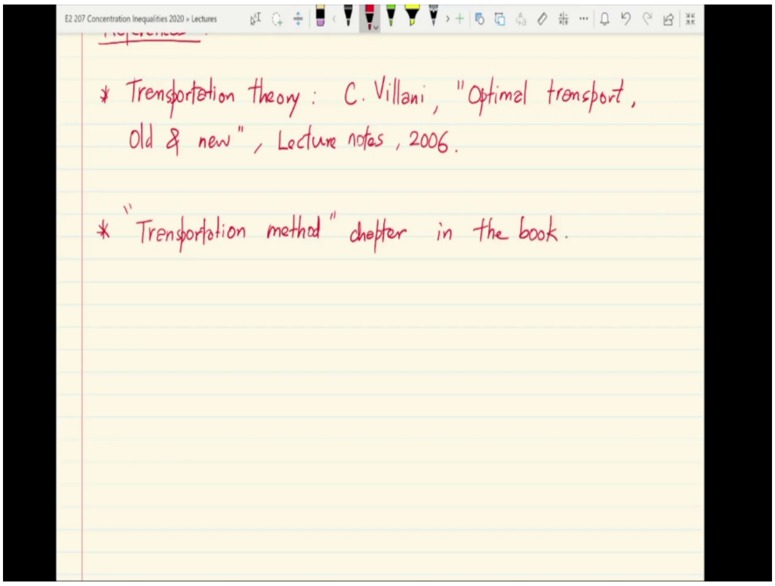
$$d_{TV}(P,Q) \leq \sqrt{2D(P\|Q)} \quad [\text{Pinsker's inequality}].$$

References :

- Transportation theory : C. Villani, "Optimal transport, old & new", Lecture notes, 2006.

So, this concludes this lecture. A very nice reference if you are interested for reading up about optimal transportation is the following. So, for Transportation theory so, you can read the lecture notes by Cedric Villani on Optimal Transport: old and new, these are lecture notes, these are series of lecture notes from 2006.

(Refer Slide Time: 65:28)



* Transportation theory : C. Villani, "Optimal transport, Old & new", Lecture notes, 2006.

* "Transportation method" chapter in the book.

And of course, you can always refer to the chapter on transportation the transportation method chapter in the book that we are following which is the Concentration Inequalities book by Boucheron.

Thank you.