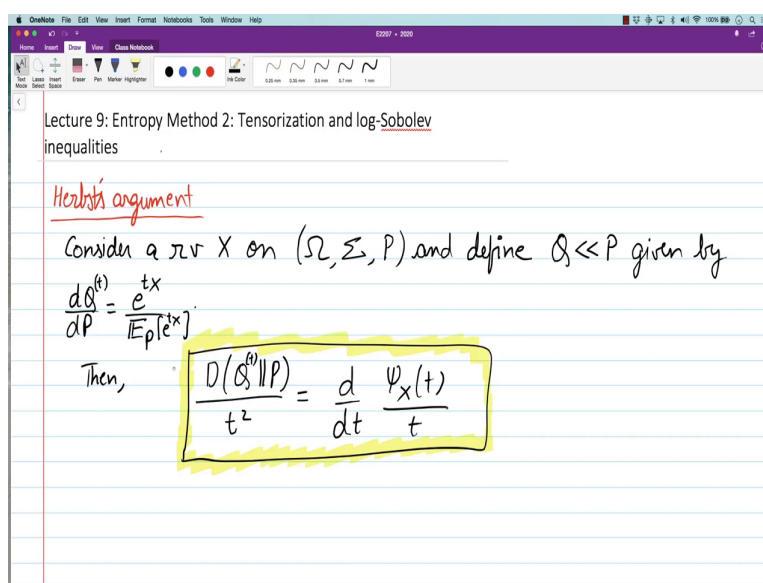


Concentration Inequalities
Prof. Aditya Gopalan
Prof. Himanshu Tyagi
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru

Lecture - 10
Log-Sobolev inequalities

(Refer Slide Time: 00:21)



Hello, in this lecture 9, we will continue with our discussion on Entropy Method and we will present the two main tools used in entropy method, namely that of Tensorization and log Sobolev inequalities. Just to refresh your memory, recall that entropy method builds on so called Herbst's argument and we presented what Herbst argument is last time.

But this time I will quickly review it, so that we can refresh your memory. So, consider a random variable X , I present a slightly different form then what we saw in last time and this form is slightly more abstract, but I think you can digest this generalization. So, we have a random variable X , defined on a probability space.

So, what all you, recall what all do you need to define a probability space; the input set ω , this is the universe and then the sigma is a + sigma and this probability measure P ok,

this is what a probability space is. This is set of events, this is the universe, and this is the probability measure.

So, consider a random variable defined on this. And define a new measure Q unless of; instead of P , we define a new probability measure Q , which has which is absolutely continuous with respect to P and is given by. So, it is a density with respect to P given by a density with respect to P ; we can just define a density with respect to P is given by e to the power x by expected value under p of e to the power x , ok. So, this is just a definition.

Then we showed this very nice formula for divergence between Q and P . Last time we were writing this whole thing in a slightly different way; we were using the measure Q we were looking at was e to the power t times f of x , ok. But it is exactly equivalent, ok.

So, the formula that we have is dQ by dP that the, that divergence between Q and P . Here let me put, let me put a t here; this divergence is $=$, this divergence by t^2 is $=$ the derivative of the log moment generating function at t divided by t , ok.

So, this is some function, its derivative is dQ the that diverges between Q t and P , ok. This is this formula which we can treat as, which leads to Herbst's argument, ok. So, how do we use this formula? Well, it can be, it gives us a nice formula for log moment generating function.

(Refer Slide Time: 03:46)

The image shows a digital notepad with handwritten mathematical notes. At the top, the expression $t^2 \frac{dQ}{dP}$ is written and underlined in yellow. Below this, the log moment generating function is defined as:

$$\Rightarrow \Psi_X(\lambda) = \lambda \int_0^\lambda \frac{D(Q^{(t)} \| P)}{t^2} dt$$

Below the integral, the text "Herbst's argument:" is written, followed by the condition:

$$\text{if } \frac{D(Q^{(t)} \| P)}{t^2} \leq \frac{\sigma^2}{2} \text{ for all } t \geq 0,$$

then the final inequality is derived:

$$\text{then } \Psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2} \text{ for all } \lambda \geq 0.$$

It gives us $\phi(\lambda) = \lambda \int_0^\lambda DQ_t P dt$, ok. This is the alternative form. So, if you want to derive an upper bound for the log moment generating function; then it suffices to derive an upper bound for this. So, that is what Herbst's argument is. So, Herbst's argument says that, if $DQ_t P \leq \frac{v}{2}$ for all $t > 0$; then $\phi(\lambda) \leq \lambda^2 s$.

Because integrate $p/2$, so that is $\lambda^2 v/2$ for all $\lambda > 0$, ok alright. So, this is Herbst's argument. So, we just have to derive a bound for this particular divergence and that will give us a bound for log moment generating function and that is what we want to do now.

(Refer Slide Time: 05:27)

then $\Psi_X(\lambda) \leq \frac{\lambda^2 v}{2}$ for all $\lambda \geq 0$.

$\Omega = X_1 \times \dots \times X_n$

→ We are interested in product $P \equiv P_{X_1} \times P_{X_2} \times \dots \times P_{X_n}$

and $X \equiv f(X_1, \dots, X_n)$

General entropy method: (1) Tensorization: Suffices to establish a bound for 1 dim

(2) A basic inequality (log-Sobolev ineq.) for $n=1$

give conc. bounds when combined with Herbst's argument

Note that, so our interest we are interested in product measures in product P ok, which we associate with different coordinates of a random variable. So, $P \times 1$ (Refer Time: 05:53) $P \times 1$ times $P \times 2, P \times n$. So, this is our n dimensional product measure. And X we will replace with our function f of X_1 to X_n . Note that rank, so that the sample space Ω is this product set here, ok.

So, Ω is X_1 cross that is where we are interested in. And for this thing, we have to derive a bound on DQ_t by DQ_t on P and it will turn out that; with the first thing we will do is, we will show that although we want to derive a bound for n dimensional case, it suffices to derive a bound for one dimensional case and that is what is called the tensorization part, ok.

So, which this general recipe, general entropy method recipe as we saw last time has two parts. So, first step, has two steps; first step is tensorization, suffices to establish a bound for 1 dimension ok, $n = 1$. And the second step is some inequality, a basic inequality which is for 1 dimension.

Typically these are called log to see, typically you use what are called log Sobolev inequalities. It is not a single inequality, sort of a family of inequalities and we will not even go into the origin of this name for now; later in the course I will return to these inequalities and then perhaps I will have a more elaborate discussion on this, ok. But for now you can just imagine, this is some elementary inequality for $n = 1$, ok.

Once we have these two things; then combined with Herbst's argument, we get give concentration bound with Herbst's argument, when you combine them with Herbst argument ok that is the general plan here. So, let me show you both these steps now, what do we do in the tensorization step; how do we, how do we tensorize?

We want to bound this quantity; what we will show is that, it suffices to bound some similar 1 dimensional quantities ok, that is a tensorization step. And then we will show log Sobolev inequalities, that is the plan, ok.

(Refer Slide Time: 09:13)

A Tensorization of divergence

$$D(Q||P) = \begin{cases} \mathbb{E}_P \left[\log \frac{dQ}{dP} \right], & \text{if } Q \ll P, \\ \infty, & \text{o.w.} \end{cases} = \mathbb{E}_Q \left[\log \frac{dQ}{dP} \right]$$

In particular, for discrete distributions Q and P on \mathcal{X} ,

$$D(Q||P) = \begin{cases} \sum_x Q(x) \log \frac{Q(x)}{P(x)}, & \text{supp}(Q) \subseteq \text{supp}(P) \\ \infty, & \text{o.w.} \end{cases}$$

So, that is s, let me start with this was all review; let me start with the first part, tensorization of divergence, this Kullback Leibler divergence. So, before I do this, let me review some simple properties of this Kullback Leibler divergence. So, recall that $D(Q \parallel P)$ is defined as expected value of; if Q has density with respect to P , this density is noted by dQ/dP , then it is defined as $\int dQ/dP \log dQ/dP \, dP$, density with respect to P +, otherwise it is let us say infinite, that is how this guy is defined.

In particular for discrete distributions Q and P on the same set X , so discrete set X ; $D(Q \parallel P)$ is = summation over x .

Note that when I am, P is the bigger measure here; whenever I am not writing an expectation here, it is clear that the expectation is with respect to P . So, since you take expectation with respect to P , it is the same as taking expectation of this guy with respect to Q , it is the same thing ok; because that is how this expectation is defined, these are some small gymnastics.

So, once you get comfortable with this idea of densities with respect to another distribution, Radon Nikodym derivative; these are called Radon Nikodym derivatives, then you can do this manipulations, ok. So, that is the whole idea of density, when you want to compute expectation with respect to Q ; you can actually compute expectation with respect to P , but multiply it with the density.

So, that is why these two things are exactly the same, ok alright. So, for the discrete case, we can write this as summation over x $Q(x)$ any expectation with respect to Q of the log likelihood ratio of this density, that is $Q(x)/P(x)$. And this is only true when Q is absolutely continuous with respect to P , which in discrete case happens if the support of Q ; the set of point where Q is non zero is contained in the support of P and it is infinite otherwise, ok.

So, this is what Kullback Leibler divergence is, we saw this last time and this comes up in our in Herbst's argument, alright.

(Refer Slide Time: 12:36)

Properties of $D(Q||P)$

(1) $D(Q||P) \geq 0$

Proof. $D(Q||P) = \mathbb{E}_P \left[\frac{dQ}{dP} \log \frac{dQ}{dP} \right]$

$=: X$

$= \mathbb{E}_P [X \log X] = \mathbb{E} [f(X)]$ where $f(t) = t \log t$

$f'(t) = 1 + \log t$
 $f''(t) = \frac{1}{t} \geq 0$ for $t \geq 0$

$\Rightarrow f(t)$ is convex

So, now I will quickly recall quickly review two very basic properties of this guy. So, first property is that it is non negative. How do we show that? So, we will use Jensen's inequality for this. For discrete case you can perhaps show it easily, actually it can all will be shown easily $D(Q||P)$, the general definition is expected value with respect to P of; this expected value with respect to P of $dQ/dP \log dQ/dP$.

So, if we think of this guy as a random variable; this guys are random variable X , let us just call it X , just defining it as X . So, this guy is expected value over P of $X \log X$, ok. And so, which is the same as expected value, just dropping this p part here of f of X , where f of t is defined as $t \log t$. So, what do we know about this function $t \log t$? So, at this function $t \log t$, how does it look? So, as t goes to infinity, this goes to infinity and as t goes to 0, what happens to this function?

So, in the limit as t goes to 0, this function also goes to 0, ok. So, this function f prime t is, it looks like what is f prime t here; f prime t is $1 + \log t$, f prime time t is $1/t$ and which is ≥ 0 for $t \geq 0$, implies f t is convex, right.

And therefore, the average so it is convex, maybe we can just try to plot it is convex and it is negative for t between 0 and 1 and then it is for and at $t = 1$, it is 0 and 0 here and then it goes

up all the way to infinity. So, it is something like this ok; note that it has only one point, where the derivative is 0. So, that also comes out here.

And yeah that is what this function looks like; therefore the since this is a convex function, so the value at its average at average is \leq average of values.

(Refer Slide Time: 16:37)

The image shows a digital notepad with the following handwritten text:

$$\begin{aligned} \Rightarrow \text{By Jensen's ing.,} \\ D(Q||P) &= \mathbb{E}[f(x)] \geq f(\mathbb{E}[x]) \\ &= f\left(\mathbb{E}_P\left[\frac{dQ}{dP}\right]\right) \xrightarrow{\quad} = \mathbb{E}_Q[1] \\ &= f(1) = -\log 1 = 0. \end{aligned}$$

So, this implies by Jensen's inequality expected value of $f(X)$ exceeds. So, this is what our $D(Q||P)$ was expected value of X , this exceeds f of expected value of X . So, what is this f of expected value of X for x ? Our choice of X is f of expected value under p of dQ/dP , ok.

I am deliberately writing it in this language of Radon Nikodym derivative, so that you become comfortable with this. So, now, note that this the way this density is defined; if you want to take expected value of a function with respect to Q ok, you can do that.

And you can switch to expectation with respect to P by multiplying it dQ/dP , right. So, these two are equal. So, this is $=$ this and this is 1, so this guy is just 1. So, that is just $= f$ of 1. If you want to understand this proof further, you try to rewrite it for the discrete case; you will understand each of the steps better ok, yeah I think that is a very important exercise you should do, alright.

So, now, what is f of 1? So, this is $1 \log 1$, which is 0, alright. So, we got our form. So, this is just by Jensen's inequality, $D(Q \parallel P)$ is non negative; that is the first property you want to cover that, $D(Q \parallel P)$ is non negative. So, first property $D(Q \parallel P)$ is non negative.

(Refer Slide Time: 18:42)

(2) Suppose P and Q prob. dist. on $\mathcal{X} \times \mathcal{Y}$. (We will use the notation P_{XY} and Q_{XY} for the joint dist.)

(chain-rule)

$$D(Q_{XY} \parallel P_{XY}) = D(Q_X \parallel P_X) + E_{Q_X} \left[D(Q_{Y|X} \parallel P_{Y|X}) \right]$$

↓

$$= D(Q_{XY} \parallel Q_X P_{Y|X})$$

$Q_X P_{Y|X}(x, y) = Q_X(x) P_{Y|X}(y|x)$

$$=: D(Q_{Y|X} \parallel P_{Y|X} \mid Q_X)$$

Second property that we have is, about is what is called chain rule sometimes. So, suppose you have P and Q are probability distributions on say X cross Y , ok. So, we will use the notation P_{XY} and Q_{XY} ok for the joint distribution. So, we will imagine this random variable X and Y and P and Q are distributions for this x and y .

Then we can relate this let us just. So, then we can relate this Q_{XY} to P_{XY} , this is $= \Sigma$ actually. So, the divergence of the first coordinate, this is called the chain rule for diversion. So, divergence of the first coordinate is $D(Q_X \parallel P_X) + D(Q_{Y|X} \parallel P_{Y|X} \mid Q_X)$.

Note that these guys are both random variables, because these are conditional distributions, ok. So, first guy and then the conditional distribution of next two conditional distributions. And you have to you have to average this guy; because these are conditional distribution, so this becomes a random variable.

So, you average it over some distribution of x and that distribution is the first one here, so Q_X , ok. So, this is the formula; yeah yes this is called chain rule for divergence and this will be this will be very useful for establishing the tensorization property we are looking for. So, if

you have divergence between $Q \times Y$ and $P \times Y$; it can be decomposed into divergence between the first part and the conditional divergence.

This second quantity here you can verify; you can verify that this is by just the form for discrete case you can easily verify, this is also = divergence between $Q \times Y$ and $Q \times P \times Y$ given X . That is the distribution between X and Y , when x is distributed with this with X and Y is distributed as $P \times Y$ given X .

So, this guy here is $Q \times P \times Y \times x$ of x comma y , the joint let us say $p_{m f}$ is $Q \times x \times P \times y$ given $X \times y$ given x , alright ok. So, this is = this and it has other, this is another notation; it is denoted by conditional divergence $D(Y \text{ given } X, P \text{ given } X)$, condition on this averaging major cubics. These are all the same things, this guy this is equal and this notation denotes these guys that is what we will show.

(Refer Slide Time: 22:25)

(Chain-rule)

$$D(Q_{XY} \parallel P_{XY}) = D(Q_X \parallel P_X) + E_Q \left[D(Q_{Y|X} \parallel P_{Y|X}) \right]$$

$$= D(Q_{XY} \parallel Q_X P_{Y|X})$$

$$=: D(Q_{Y|X} \parallel P_{Y|X} \mid Q_X)$$

(In particular, if $Q_{XY} = Q_X Q_Y$ and $P_{XY} = P_X P_Y$, then

$$D(Q_{XY} \parallel P_{XY}) = D(Q_X \parallel P_X) + D(Q_Y \parallel P_Y) .)$$

Proof. We can assume that $Q_{XY} \ll P_{XY}$:

And an important consequence in particular if $Q \times Y$ equals to $Q \times X \times Y$, so it is a product measure, they are independent of Q and $P \times Y$ equals to $P \times X \times Y$; then it is then divergence is just additive $D(Q \times Y \times P \times X \times Y) = D(Q \times X \times P \times X) + D(Q \times Y \times P \times Y)$, ok. So, that is the claim. So, how do we show this?

Proof is simple, you can just expand things, I will give an informal proof; first you notice that we can assume that $Q \times Y$ is absolutely continuous with respect to $P \times X \times Y$, ok. This is just

some, maybe I should not prove this for the general case, it will look very technical; I will give the proof for discrete case, for discrete distributions.

(Refer Slide Time: 24:08)

The image shows a handwritten proof in a OneNote application. The text is as follows:

$$D(Q_{XY} \parallel P_{XY}) = D(Q_X \parallel P_X) + D(Q_{Y|X} \parallel P_{Y|X}).$$

Proof. (for discrete distributions) WLOG assume $\text{supp}(Q_{XY}) \subseteq \text{supp}(P_{XY})$

$$D(Q_{XY} \parallel P_{XY}) = \mathbb{E}_{Q_{XY}} \left[\log \frac{Q_{XY}(x, y)}{P_{XY}(x, y)} \right]$$

$$= \mathbb{E}_{Q_{XY}} \left[\log \frac{Q_X(x)}{P_X(x)} + \log \frac{Q_{Y|X}(y|x)}{P_{Y|X}(y|x)} \right]$$

$$= \mathbb{E}_{Q_X} \left[\log \frac{Q_X(x)}{P_X(x)} \right] + \mathbb{E}_{Q_X} \left[\mathbb{E}_{Q_{Y|X}} \left[\log \frac{Q_{Y|X}(y|x)}{P_{Y|X}(y|x)} \right] \right]$$

$$= D(Q \parallel P) + \mathbb{E}_{Q_X} [D(Q_{Y|X} \parallel P_{Y|X})]. \quad \square$$

So, discrete distribution the proof is very simple; this divergence here $Q_{XY} \parallel P_{XY}$ is = expected value with respect to Q_{XY} , when the support is contained, otherwise they are both infinite. So, I can assume without loss of generality, assume support of Q_{XY} is contained in the support of P_{XY} ; because if not if this is not true, then both sides here will become infinity, you can check that.

So, then you have Q_{XY} of X comma Y . So, this expectation is for this X comma Y by P_{XY} of X comma Y , ok. So, this these are the random variables and the expectation with respect to these random variables. So, this is = expected value with respect to Q_{XY} of log of Q_{XY} by P_{XY} + log $Q_{Y|X}$ given X / $P_{Y|X}$ given X , ok.

And this guy here is exactly =. So, what is the expectation of; this is a function only of X . So, this part of Y over Y disappears. So, what you get is just this guy here, + now expected value over Q_X and expected value over Q of Y given X , because that is how expectation works.

This is the conditional expectation given X under Q of log of Q of Y given X P of Y given X Y given X . So, this is, this notation is a bit ugly; but what I am saying is the expectation of this guy can be written as expected value of the conditional expectation and

the conditional expectation will happen to be with respect to Q of Y given X and the outer expectation is Q X .

So, when you do this what you get here is, this is just the first term is just $D Q P$ and second term is expected value over Q X of what you see inside is $D Q Y$ given $X P Y$ given X more or less elementary ok, it is by definition, ok. So, that is what we get. So, this was the second property that we wanted to cover that, this chain rule here, ok.

(Refer Slide Time: 27:33)

The image shows a handwritten derivation in a notebook. At the top, there is a small diagram showing a sequence of variables X_1, X_2, \dots, X_n with arrows indicating dependencies. The main derivation is as follows:

$$\begin{aligned} \rightarrow D(Q_{X^n} \| P_{X^n}) &= \sum_{i=1}^n E_{Q_{X^{i-1}}} [D(Q_{X_i | X^{i-1}} \| P_{X_i | X^{i-1}})] \\ &= \sum_{i=1}^n D(Q_{X_i | X^{i-1}} \| P_{X_i | X^{i-1}} | Q_{X^{i-1}}) \end{aligned}$$

Below the equations, there is a note in green ink: "Notation. Recall P_{ij} \rightarrow conditional dist. given (X^{i-1}, X_{i+1}^n) . Let $X^{(i)} := (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$."

So, note that we did it for two random variables, but you can do it for multiple random variables. So, if you have X^n and P_{X^n} , that is $= \sum_{t=1}^n$ or maybe rather $i=1$ to n divergence between Q of X_i given the past under Q , P of X_i given the past under P . And now this is a random variable depending on X - X the past and so, you take expectation with respect to the past, but under the measure Q .

So, most essentially this is represented as conditional divergence. Somehow this conditional divergence notation is very handy; but it is not so popular outside the information theory literature, ok. So, that is what this divergence is, alright. So, we have these two properties, chain rule and the non-negativity of divergence.

Now, using these two properties, we will establish our main result. So, some notation, recall that we had this measure we were showing P_i and P_i was measured when you; P_i was

basically conditional distribution, given everything, but the i th, right. So, given $X_{1:i-1}$ and $X_{i+1:n}$, ok.

So, we will introduce some, we will introduce a notation for this guy; let $X^{(i)}$ be defined as everything, but the i th one. This is a random variable which appears in, which appeared in Efron-Stein inequality, ok. So, with this is the notation, $X^{(i)}$ superscript i ; superscript because this is a vector and it is everything, but the i th guy, just like these notations this, $X_{1:i-1}$ was everything up to $i-1$.

This one is everything starting from $i-1+1$ to n and this is everything, but the i th.

(Refer Slide Time: 30:27)

Let $X^{(i)} := (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$.

Lemma ("Entropy" tensorization lemma) Consider independent random variables X_1, \dots, X_n (under P). Let Q_{X_1, \dots, X_n} be any other distribution. Then,

$$D(Q || P) = D(Q_{X^n} || P_{X^n})$$

$$\leq \sum_{i=1}^n \mathbb{E}_{Q_{X^{(i)}}} \left[D(Q_{X_i | X^{(i)}} || P_{X_i | X^{(i)}}) \right]$$

Definition: $\text{Ent}(f) = \mathbb{E}[\log f(X)] = -\mathbb{E}[\log \frac{1}{f(X)}]$
 $D(Q || P) = \frac{\text{Ent}(f)}{\mathbb{E}[e^f]} - \mathbb{E}[\log \frac{1}{f(X)}]$

With this notation, here is a lemma and I will call this the entropy tensorization lemma. So, consider independent random variables X_1 to X_n under P ; let Q_{X_1, \dots, X_n} be any other distribution. Then divergence between Q and P is the overall thing and P this is basically (Refer Time: 32:09) divergence between Q_{X_1, \dots, X_n} and P_{X_1, \dots, X_n} .

This divergence is less than or = mention $i = 1$ to n , expected value with respect to this Q of X everything, but the i th guy ok; divergence between $Q_{X_i | X^{(i)}}$ of the i th guy given everything, P the i th guy given everything, that is the key, sort of a very interesting claim. Looks something like chain rule, but we are not conditioning on the past.

We are conditioning on everything, but the i th entry. So, all these divergences are just changing one part of the function ok, it of the measure. So, it fixes everything, but just changes one point. And I am calling it entropy tensorization lemma, but we do not see entropy anywhere, we will see entropy later. By the way I never reviewed in this class what entropy is.

Recall that in the last class, we introduced this entropy of f and we showed that this entropy of f is exactly $=$; I mean we defined it as yeah, this is some side note, it is a little bit of a regression. So, entropy of any function we had defined as expected value of $f \times \log f \times$ - expected value $f \times \log$ of expected value $f \times$, that is what we defined entropy to be and divergence between this Q of f .

So, Q of f we had defined as a tilting with respect to f . So, this was $e^f \times$ that is a tilting; from the from this P is exactly $=$ entropy of f by expected value of, I am sorry this divergence is exactly $=$ entropy of e^f by expected value of e to the power, ok.

So, this formula from last time, this divergence can also be related to this ratio, this is just by definition, ok. So, since we are since we are establishing some tensorization property for this divergence that will tend to amount to tensorization property for entropy, ok.

Of course we are assuming here this expected value of e^f is 1, that is why divergence looks like entropy and that we can do for most of our proofs. So, we are just calling it tensorization of entropy, but actually we are showing some tensorization for divergence, ok alright ok.

(Refer Slide Time: 35:47)

$$D(Q||P) = D(Q_{X^n} || P_{X^n})$$

$$\leq \sum_{i=1}^n \mathbb{E}_{Q_{X^{(i)}}} \left[D(Q_{X_i|X^{(i)}} || P_{X_i|X^{(i)}}) \right]$$

Proof. by chain rule for KL divergence,

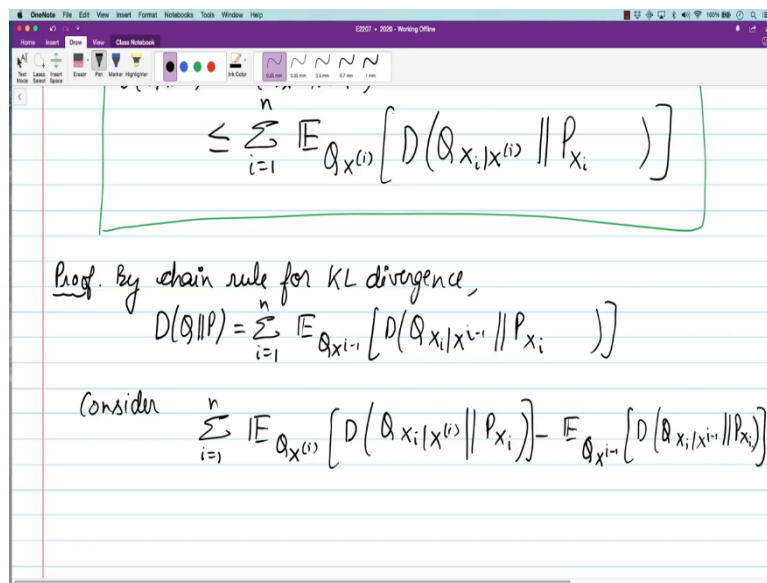
$$D(Q||P) = \sum_{i=1}^n \mathbb{E}_{Q_{X^{i-1}}} \left[D(Q_{X_i|X^{i-1}} || P_{X_i|X^{i-1}}) \right]$$

Consider $\sum_{i=1}^n \mathbb{E}_{Q_{X^{(i)}}} \left[D(Q_{X_i|X^{(i)}} \right]$

So, now we have this claim, how do we show this? So, proof just by chain rule for divergence, by chain rule for Kullback Leibler divergence $D(Q||P)$, $D(Q||P)$ is $= \sum_{i=1}^n D(Q||P)$ let me write in a different way expected value over the past divergence of Q_{X_i} given the past from P_{X_i} given the past ok, that is equal. So, and we want to show this less than this.

So, we can just take the difference of these guys and show that the difference is non negative. So, consider $\sum_{i=1}^n$ expected value over Q of X_i of divergence between Q_{X_i} given X_i , ok so right. So, this, claim I was; the most important thing in this claim is that these guys are independent under P .

(Refer Slide Time: 37:32)



$$\leq \sum_{i=1}^n \mathbb{E}_{Q_{X^{(i)}}} [D(Q_{X_i|X^{(i)}} \| P_{X_i})]$$

Proof. by chain rule for KL divergence,

$$D(Q \| P) = \sum_{i=1}^n \mathbb{E}_{Q_{X^{i-1}}} [D(Q_{X_i|X^{i-1}} \| P_{X_i})]$$

Consider

$$\sum_{i=1}^n \mathbb{E}_{Q_{X^{(i)}}} [D(Q_{X_i|X^{(i)}} \| P_{X_i})] - \mathbb{E}_{Q_{X^{i-1}}} [D(Q_{X_i|X^{i-1}} \| P_{X_i})]$$

And so under P , there is no past ok, just P_{X_i} , that is the most; otherwise actually this bound does not hold [Laughter], you can find the counter example. So, this is only for independent case, so yeah this is P_{X_i} . And similarly this guy, because they are independent; this is only P_{X_i} , does not depend on what the past is, ok.

So, the difference of these two terms would be this from P_{X_i} - the expected value under Q of the past of divergence between $Q_{X_i|X^{i-1}}$ given X^{i-1} from P_{X_i} , ok. And so, let us now look at each term here.

(Refer Slide Time: 38:32)

Consider
$$\sum_{i=1}^n \mathbb{E}_{Q_{X^{(i)}}} \left[D(Q_{X_i|X^{(i)}} \| P_{X_i}) \right] - \mathbb{E}_{Q_{X^{i-1}}} \left[D(Q_{X_i|X^{i-1}} \| P_{X_i}) \right]$$

$$\mathbb{E}_{Q_{X^n}} \left[\log \frac{Q_{X_i|X^{(i)}}(X_i|X^{(i)})}{P_{X_i}(X_i)} \right] - \mathbb{E}_{Q_{X^n}} \left[\log \frac{Q_{X_i|X^{i-1}}(X_i|X^{i-1})}{P_{X_i}(X_i)} \right]$$

$$= \mathbb{E}_{Q_{X^n}} \left[\log \frac{Q_{X_i|X^{(i)}}(X_i|X^{(i)})}{Q_{X_i|X^{i-1}}(X_i|X^{i-1})} \right]$$

So, we just considered this guy here. So, this guy is, it can be written as expected value of this and let us just look at a discrete key. So, there will be expectation with respect to this conditional distribution of log of probability here and probability here. It can all compactly be written as first term expected value over Q , the entire Q . Q of X^n of log of Q of X_i given the past / P of X_i , ok.

This is what the first term will turn out to be. So, there are two steps here, I had an expectation of a Q_{X_i} outside and then there is an expected value which comes in the definition of divergence; you could have done this here also, maybe it is something good to put it down, I have to put down here.

(Refer Slide Time: 39:41)

(Chain-rule)

$$D(Q_{XY} || P_{XY}) = D(Q_X || P_X) + E_Q \left[D(Q_{Y|X} || P_{Y|X}) \right]$$

$$= D(Q_X || P_X) + E_Q \left[\log \frac{Q_{Y|X}(Y|X)}{P_{Y|X}(Y|X)} \right]$$

$$= D(Q_X || P_X) + D(Q_{Y|X} || P_{Y|X} | Q_X)$$

(In particular, if $Q_{XY} = Q_X Q_Y$ and $P_{XY} = P_X P_Y$, then

$$D(Q_{XY} || P_{XY}) = D(Q_X || P_X) + D(Q_Y || P_Y) .)$$

Proof. (for discrete distributions) WLOG assume $\text{supp}(Q_{XY}) \subseteq \text{supp}(P)$

$$D(Q_{XY} || P_{XY}) = E_{Q_{XY}} \left[\log \frac{Q_{XY}(X,Y)}{P_{XY}(X,Y)} \right]$$

So, I had all these guys here, right. So, one more way to define conditional divergence is you can take expected value with respect to Q_X of \log of $Q_{Y|X}$ given X by $P_{Y|X}$ given X ; I should put some derivative, but let us just look at the discrete case. So, that is just y of, because this Y and X are generated from Q , that is what this divergence is, it is equivalent.

So, we will have this conditional expectation inside that makes the conditional divergence and then the outside expectation, which you can combine to get the overall expectation, ok. So, this is first term and the second term again is expected value with respect to Q of $X \log$ of $Q_{Y|X}$ given X by $P_{Y|X}$ given X . So, this sort of harmless trick of writing both in terms of common distribution is something which will come up again and again in this course.

And this is sometimes called couplings; they in principle these two, these two quantities did not did not require a common joint distribution, they were two different quantities related to two different distributions. But we realize that they can both be expressed as coming out from the same joint distribution Q_{XY} .

And the advantage of doing that is that, I can take this expectation, the common expectation outside and take both these quantities inside. We have \log of this - \log of this; is P_X part cancels and this is where we have used independence, because that is because, otherwise you

would have conditioning on X_i here and X past X_{i-1} here and X past here. So, now, you just have this one.

(Refer Slide Time: 41:34)

$$\begin{aligned}
 & E_{Q_{X^n}} \left[\log \frac{Q_{X_i|X^{(i)}}(X_i|X^{(i)})}{P_{X_i}(X_i)} \right] - E_{Q_{X^n}} \left[\log \frac{Q_{X_i|X^{i-1}}(X_i|X^{i-1})}{P_{X_i}(X_i)} \right] \\
 &= E_{Q_{X^n}} \left[\log \frac{Q_{X_i|X^{(i)}}(X_i|X^{(i)})}{Q_{X_i|X^{i-1}}(X_i|X^{i-1})} \right] \\
 &= D(Q_{X_i|X^{(i)}} \| Q_{X_i|X^{i-1}} | Q_{X^{(i)}}) \geq 0.
 \end{aligned}$$

So, now what is this guy X_i given X_{i-1} ; this again can be written in terms of this is your P , this is your Q and you are looking at this conditional divergence. So, this is conditional divergence (Refer Time: 41:49) remove, may be the, we have better notation as that of conditional, this is the conditional divergence between X_i given X past.

From another measure where we just take the marginal given only the first $i-1$, not $i-1$ and averaged over X of i ok and this guy must be non-negative. So, this guy is non-negative, alright. So, this difference each term in this difference is non negative; therefore the overall difference is non negative and therefore, this guy exceeds this guy, that is the proof.

Nothing very difficult, actually that is it looks all very simple; because that is how profound this method is. So, now, what we do is, we apply this sort of tensorization property of divergence to get a tensorization property of entropy; because we have called this entropy tensorization lemma. Let us see how we get tensorization of entropy from this.

(Refer Slide Time: 42:55)

$$\text{Ent}(f) = E[f \log f] - E[f] \log E[f], \quad f \geq 0$$

If $E[f] = 1$

$$\frac{dQ}{dP} = f$$

$$D(Q||P) = E\left[\frac{dQ}{dP} \log \frac{dQ}{dP}\right] = \text{Ent}(f)$$

$$\leq \sum_{i=1}^n E_i$$

And recall that the definition of entropy for us was expected value of $f \log f$ ok - expected, this is the definition of entropy that we have. And for the particular case where, so the entropy of f by expected value of f ; it will be better way to say is, if expected value of f equals to 1, which is something we will assume for now. By the way this entropy is defined only for non-negative functions f , ok.

Other otherwise they have a log of a negative function, it does not make sense. So, this is the definition; for now let us assume just this part that, expected value of f is 1 and we will see that this can be removed later. So, under this assumption, consider a Q which is defined by dQ/dP equals to f ; just like we had seen earlier that we had $e^{\lambda f}$, now we are defining dQ/dP equals to f .

For this Q we saw that the divergence between Q and P , which could be bound, could be bounded above using that tensorization property; but this divergence is exactly equals to expected value of $dQ/dP \log dQ/dP$ and this is exactly = entropy of f , ok. So, for this Q , this divergence is exactly the entropy of f , ok.

And we saw that we saw that, this entropy of this divergence is less than $= \sum_{i=1}^n$ to n expected, divergence expectation over expected value over this $Q \times i$. So, Q of all, but the i th guys of divergence between $Q \times i$ given I will put the i th guy and $P \times i$.

(Refer Slide Time: 45:13)

$$\leq \sum_{i=1}^n \mathbb{E}_{Q^{(i)}} \left[D(Q_{x_i | x^{(i)}} \| P_{x_i}) \right]$$

$$= \sum_{i=1}^n \mathbb{E}_Q \left[\log \frac{Q_{x_i | x^{(i)}}}{P_{x_i}} \right]$$

$$\frac{Q_{x_i | x^{(i)}}}{P_{x_i}} = \frac{f(x_i | x^{(i)})}{\mathbb{E}_P[f(x_i | x^{(i)})]}$$

Now, if you look at this guy, this can be expressed as summation $i = 1$ to n expected value over Q , the whole Q now X^n of \log of $Q_{X_i | X^{(i)}}$ given X_i / P_{X_i} $d Q_{X_i}$. So, now we need to convince ourselves that, in fact this measure is absolutely contents with respect to P_{X_i} and their log likelihood ratio can be expressed in some nice form.

So, let us look at that. So, d of $Q_{X_i | X^{(i)}}$ given $X_i / d P_{X_i}$ and this random variable we will look at it as a function of x_i ; but of course, we have conditioned on all the paths, so I will use this notation of conditioning here, ok. This guy here is = ok, you can verify this part, may be by using a discrete case; this is = the f of x_i given x_i ok, f of x_i given x_i / the conditional expectation under P of f of x_i given X_i , ok.

I am sorry. So, this I will this conditional expectation over p of X_i , this is the random variable given the past; but P this does not matter, because P for independent measure this is just leveraging only over X^n . So, this is something you can verify, you can verify for discrete case; because the product form of P , you can show that this one also has this form.

(Refer Slide Time: 48:12)

The image shows a digital notepad with the following handwritten equations:

$$\frac{d \log_{X_i|X^{(i)}}(x_i | x^{(i)})}{d p_{X_i}} = \frac{f(x_i | x^{(i)}) \rightarrow f(x^{(i)}; x_i, x^{(i)})}{\mathbb{E}_P[f(x_i | x^{(i)})]}$$

$$= \sum_{i=1}^n \mathbb{E}_Q \left[\log \frac{f(x_i | x^{(i)})}{\mathbb{E}_P[f(x_i | x^{(i)})]} \right]$$

$$= \sum_{i=1}^n \mathbb{E}_P \left[f(x_i | x^{(i)}) \log f(x_i | x^{(i)}) - \mathbb{E}_P[f(x_i | x^{(i)})] \log \mathbb{E}_P[f(x_i | x^{(i)})] \right]$$

And therefore, by observing this guy here, if you continue from this point; this expression here is exactly $\sum_{i=1}^n$ expected value over Q of \log of, I did not define this notation, this is just my notation for $X_{i-1} \times x_i, x_{i+1}$ to n , right. So, you have all the remaining arguments in the i th argument, is just for convenience. So, this becomes \log of f of X .

So, then the Q by f of X divided by expected value the P expected to write it as f of X_i given X_i and under P f of X_i given X_i . So, if you look at this quantity here, this would look like $\sum_{i=1}^n$ expected value under P of $f(X)$ and we have assumed that expected value of X is 1, so there is no division required.

\log and this $f(X)$ we can write as $f(X_i | x^{(i)}) \log f(X_i | x^{(i)})$; this is just f of X differently for a specific reason, - expected value over P of $f(X_i | x^{(i)}) \log$ of expected value of P $f(X_i | x^{(i)})$ given X_i , ok.

(Refer Slide Time: 50:36)

$$\begin{aligned}
 &= \sum_{i=1}^n \mathbb{E}_p \left[f(x_i | x^{(i)}) \log \frac{f(x_i | x^{(i)})}{\mathbb{E}_p[f(x_i | x^{(i)})]} \right] \\
 &= \sum_{i=1}^n \mathbb{E} \left[\text{Ent}_{(i)}(f) \right] \\
 \Rightarrow & \boxed{\text{Ent}(f) \leq \sum_{i=1}^n \mathbb{E} \left[\text{Ent}_{(i)}(f) \right]} \quad \text{if } \mathbb{E}[f] = 1
 \end{aligned}$$

Condition $\mathbb{E}[f]=1$ can be removed by dividing both sides with $\mathbb{E}[f]$.

So, essentially what this thing we have written here is, $i = 1$ to n expected value of entropy; but given all, but the i th coordinate of f ok, that is that this is similar to the Efron-Stein form that we saw earlier, ok. So, we have the Efron-Stein for entropy we show that, entropy of f ; we have shown that entropy of f is less than $= \sum_{i=1}^n$, sorry even n expected value entropy given everything, but the i th coordinate of f , ok.

And we have shown it only for the case when expected value of f is 1; but now you note that if you divide both these sides with the constant ok, this is true for expected value of f , this is if expected value of $f = 1$. But, if this inequality holds and then you divide both these sides with a constant, then still the inequality holds.

And therefore, this can be removed. So, this condition can be removed by dividing both sides with expected value of f ; it is a non-negative function, so we can do that, ok. So, this is something that we will retain from this lecture; this is the Efron-Stein counter part of entropy and we will call this a tensorization property of entropy. And it just follows from the tensorization property of this divergence ok and that is something we have retained from this lecture, ok.

So, I think this is becoming a long lecture already. So, we will stop here and in the next class, I will show the show the binary logs Sobolev inequality.