

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

Lecture - 12
Variational formulae for Kullback-Leibler and Bregman Divergence

(Refer Slide Time: 00:21)

Lecture 11: Variational Formulae for KL Divergence

A Gibbs variational formula and a variational formula for divergence

We already saw the following formula for the log-moment generating function

$$\Psi_Z(\lambda) = \lambda \int_0^{\lambda} \frac{D(Q^{(tz)} \| P)}{t^2} dt \quad \left(\text{leads to the entropy method} \right)$$

where $\frac{dQ^{(tz)}}{dP} = \frac{e^{tz}}{E_P[e^{tz}]}$

→ We now present another variational formula for $\Psi_Z(\lambda)$...

Hi welcome to lecture 11. In this lecture, we will continue with our development of entropy method and we will take a slight detour because there are some very interesting formulae that we need to proceed, to apply entropy method to general random variables. In the last week, we saw how entropy method applies to Binary random variables as well as Gaussian random variables and now, we will proceed and extend the same bounds to general random variables.

Towards that we need some variational formula for KL divergence and it is not immediately clear what this formulae do. But in this class, we will just collect these formulae and then, in the next lecture, we will move forward to apply this formulae to obtain a desired bounds. So, we will start with a classic formula known as Gibb's variational principle, which is a variational formula for the log moment generating function.

Last week, we saw that, we saw this 2 weeks back, I think. We saw this beautiful formula for log moment generating functions, which relates log moment generating function to the divergence of the tilted measure from P ; where, the tilting this tilted measure is this exponentially tilted measure defined this way ok.

So, this measure has a density given by e^{tz} by its expected value with respect to P . This quantity can be related to the this quantity here, this ratio can be related this divergence can be related to the entropy of e^{tz} and that therefore, when we whenever we can bound entropy of e^{tz} that can be yield a bound on the log moment generating function.

This is what we called entropy method, what we have been calling entropy method. And in this lecture, we will present another variational formula for this log moment generating function. In fact, what we present in this lecture is perhaps out of 100 years or 200 years older than this formula. So, let me present that formula.

(Refer Slide Time: 02:35)

→ We now present another variational formula for $\Psi_Z(\lambda) \dots$

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

Proof: Let $Q^{(z)}$ be given by $\frac{dQ^{(z)}}{dP} = \frac{e^z}{\mathbb{E}_P[e^z]}$

$$0 \leq D(Q \| Q^{(z)})$$

$$= \mathbb{E}_{Q^{(z)}} \left[\frac{dQ}{dQ^{(z)}} \log \frac{dQ}{dQ^{(z)}} \right]$$

$$= \mathbb{E}_P \left[\frac{dQ^{(z)}}{dP} \cdot \frac{dQ}{dQ^{(z)}} \log \frac{dQ}{dQ^{(z)}} \right]$$

Diagram: A red curve labeled $Q^{(z)}$ starts at point P and ends at point $Q^{(z)}$. A red arrow points from P to $Q^{(z)}$ with the label $Q^{(z)} = P$.

So, we will show that the log moment generating function for a zero mean random variable; for simplicity, we will have a zero mean random variable at λ , maybe I should just write it this way so that the log moment generating function $e^{\lambda Z}$ - expected value of Z .

And every expectation here with respect to P , can be expressed as sup over all distributions Q that are absolutely continuous with respect to P expected value is λ here of Z under Q -

expected value of Z under P . This is because you have done the subtraction - divergence between Q and P ok.

So, if you take supremum of this quantity over all Q that are that have density with respect to P , that are absolutely continuous with respect to P that is = the log moment generating function. So, this is the formula we will derive. So, this formula here is there are various ways to see it.

One interesting way is that one thing to see here is that we were taking expectation of sort of a complicated function this is e to the power sorry about this. This is a complicated function of the random variable. But what we have here is just linear expectations. So, somehow this linearizes this whole quantity and it is a very famous formula which see which gives an alternative expression for the log moment generating function.

Later, we will develop the transportation method which exploits this formula to derive concentration bound. In this class, we will just give a proof for this formula. So, proof; so, we once again consider that exponential tilting of P . So, let Q_z be Q_z absolutely continuous with respect to P . So, Q_z which has a density with respect to P be given by.

So, we just have to define the density has before this guy has density e to the power z - well let us just prove it for e to the power z first; by expected value e to the power z ok. So, look at this guy and what we notice is that the distance the divergence between this guy and the divergence between Q and this guy must be ≥ 0 that is what we check that is what we notice.

And to talk about this divergence, we first must ensure that this Q_z also has divergence with respect to sorry Q also has density with respect to Q_z and that is something that can be verified ok. So, this guy here, what is this divergence? This divergence can be written as expected value over Q_z of the density of Q with respect to Q_z log density of Q with respect to Q_z .

(Refer Slide Time: 06:33)

$$\begin{aligned}
 &= \mathbb{E}_P \left[\frac{dQ^{(z)}}{dP} \cdot \frac{dQ}{dQ^{(z)}} \log \frac{dQ}{dQ^{(z)}} \right] \\
 &= \mathbb{E}_P \left[\frac{dQ}{dP} \log \frac{dQ}{dP} \cdot \frac{1}{\frac{dQ^{(z)}}{dP}} \right] \\
 &= \underbrace{\mathbb{E}_P \left[\frac{dQ}{dP} \log \frac{dQ}{dP} \right]}_{D(Q||P)} - \underbrace{\mathbb{E}_P \left[\frac{dQ}{dP} \log \frac{e^z}{\mathbb{E}_P[e^z]} \right]}_{\mathbb{E}_Q[\log e^z / \mathbb{E}_P[e^z]]} \\
 &= D(Q||P) - \mathbb{E}_Q[z] + \log \mathbb{E}_P[e^z].
 \end{aligned}$$

And this can be written as the expected value over P now. This is because this Q z has a density with respect to P density of Q z with respect to P times this log density of Q with respect to Q z. So, this guy here is expected value over P. This is how Radon-Nikodym derivative work. But you can verify this for discrete case easily. Again, I am leaving that as an exercise. This is just dQ by dP the density of Q with respect to P and log of dQ by dQ z.

So, I will write it again as dQ by dP times 1 by dQ z by dP that this is also true. This is almost surely these things are = this ok. So, if you are very confused about what is going on here, what is this voodoo math, you just write it as Σ for discrete case and then, you will see how this, this is a ratio of two PMFs and you can divide and multiply with you can bring this in and divide and multiply with these two densities ok.

So, that is how it is working or just the ratio of PMFs that is the easy way to see it or ratio of densities when everything has density with respect to the Lebesgue measure. So, everything is a contains a random variable. But here we are talking about densities with respect to some other measures, so I am using some technical result in the middle of all this. But they are all believable results.

So, that gives you expected value with respect to P log of this + expected value with respect to Q now log of because we have a dQ by d, there go slow here. So, this is P dQ by dP log.

So, log of $E_Q[z]$ by this, what is this density? We will by design this is e to the power z , my expected value with respect to P e to the power z alright. So, this first term here, you can recognize as the divergence between Q and P .

So, that is what this first term is and this now, this I will divide into two parts log of a constant expected value of some constant, this is just a could be this one here, the second term here you can recognize it as expected value with respect to Q because you have a Radon-Nikodym derivative here. This allows you to convert expectation with respect to P to expectation with respect to Q log of e to the power z by this guy.

So, this is - expected value with respect to Q of z and then, + log of expected value with respect to P e to the power z ok that is the inequality we get. We get that this whole thing here is non negative.

(Refer Slide Time: 10:15)

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

$$\Rightarrow \log E_P[e^z] \geq E_Q[z] - D(Q||P) \text{ for every } Q \ll P.$$

with equality iff $Q = Q^{(z)}$.

$$\log E_P[e^z] = \max_{Q \ll P} E_Q[z] - D(Q||P)$$

Replace z with $z - E_P[z]$

$$\log E_P[e^{z - E_P[z]}] = \max_{Q \ll P} E_Q[z] - E_P[z] - D(Q||P)$$

And so, what that implies is that the log of log moment generating function exceeds $E_Q[z] - D(Q||P)$ for every Q that has density with respect to P . But in fact, we also know the condition for equality. The condition for equality is when this Q is = this, this tilted measure here ok. So, with equality if and only if because divergence is non-negative unless the two distributions are equal; so, if and only if Q equals to $Q^{(z)}$ ok.

Therefore, not only this inequality holds, but there is a distribution for which there is equality and therefore, $\log \text{ expected value over } P \text{ of } e \text{ to the power } z = \max \text{ over all } Q \text{ that are absolutely continuous with respect to } P \text{ expected value with respect to } Q \text{ of } Z - D(Q|P)$ ok. So, you may wonder what happened to the centering part. Well, it comes for free. So, this centering part replace z by this guy z by the centered version and then, you will get; so, replace z with this guy.

So, you get $\text{expected value over } P \text{ of } e \text{ to the power } z - \text{this is} = \max \text{ over } Q \text{ expected value of } Z \text{ under } Q - \text{expected value of } Z \text{ under } P - D(Q|P)$ ok, that is the form we have seen in the claim ok. So, that is roughly the proof of this very classic formula and very useful formula.

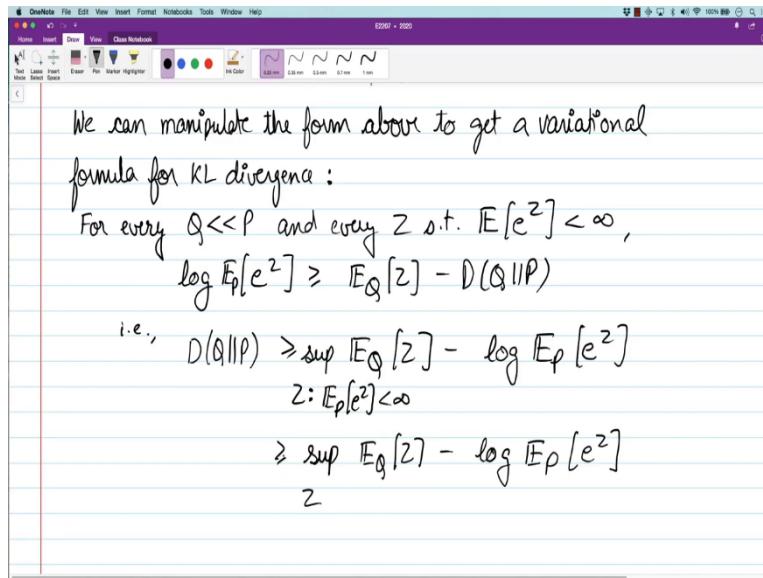
It has various names Laplace integration formula, Gibbs variational principle, Donsker-Varadhan variational principle, all these things are related to this formula itself and I wrote sup here. But we saw that we have a max actually, this is the measure which attains it with equality.

Just if you want to visualize this proof, all this proof is saying is you look at as you change these different random variables, this random variable z , you can define a measure which a measure which passes through P called a constant z for $z = 0$, this is $=$ the measure P .

So, you have a P and you for different z get something like this is your Q_z ok and Q of 0 equals to P and there is some there is some measure Q and we look at the distance of this from Q and we claim that for any Q , this distance is non-negative. Yeah of course and that is what this inequality is and it is equal if this Q lies on somewhere here. So, for that particular Q , these two things coincide and the measure is 0 and then, the distance is 0 that is what this inequality says ok.

So, that is the Gibbs's variational principle. Now, in fact, if you look at the title of this session I said I will prove the Gibbs's variational formula and its sometimes called the Gibbs's variational principle and a variational formula for KL divergence, the Kullback-Leibler divergence. So, as perhaps, now you can guess we can invert this formula by taking divergence to this side.

(Refer Slide Time: 14:37)



We can manipulate the form above to get a variational formula for KL divergence:

For every $Q \ll P$ and every Z s.t. $E[e^Z] < \infty$,

$$\log E_P[e^Z] \geq E_Q[Z] - D(Q \| P)$$

i.e.,

$$D(Q \| P) \geq \sup_{Z: E_P[e^Z] < \infty} E_Q[Z] - \log E_P[e^Z]$$
$$\geq \sup_Z E_Q[Z] - \log E_P[e^Z]$$

So, the formula we get is we can manipulate the form above to get a variational formula for KL divergence. So, indeed, we have shown that for every Q , for every Q that has density with respect to P , the log moment generating function, this is for every Q and this is for every random variable z and every z such that expected value e to the power z is finite.

A log moment generating function is = not just = it exceeds expected value with respect to Q of $z - D(Q \| P)$. So, that is this divergence between Q and P exceeds the expected value under Q of this random variable $z - \log$ expected value over P of z and this can this holds for every z . So, therefore, this holds for sup over all z , such that expected value under P e to the power z is finite.

In fact, if this thing is infinite, then you have infinity here and $-\log$ infinity is $-\infty$. So, this holds automatically. But for sanity check, we just put to be finite ok. So, maybe I do not know, just this maybe you like this one. So, because when this is infinite, the bound holds together ok. So, this is \geq this. So, we have this bound. Now, this bound holds, this bound holds; but can we find a Z for which this for equality holds?

(Refer Slide Time: 17:25)

Handwritten derivation in OneNote:

$$\geq \sup_Z \mathbb{E}_Q[Z] - \log \mathbb{E}_P[e^Z]$$

On the other hand, for $Z = \log \frac{dQ}{dP}$ we have

$$\mathbb{E}_Q[Z] - \log \mathbb{E}_P[e^Z]$$

$$= \mathbb{E}_Q\left[\log \frac{dQ}{dP}\right] - \log \mathbb{E}_P\left[\frac{dQ}{dP}\right]$$

Annotations: A green arrow points from the second term to 1, and a blue bracket under the first term points to $\mathbb{E}_P\left[\frac{dQ}{dP} \log \frac{dQ}{dP}\right]$.

$$= \mathbb{E}_P\left[\frac{dQ}{dP} \log \frac{dQ}{dP}\right] - \log 1$$

$$= D(Q||P) - 0$$

And that z exists on the other hand for $Z =$ the Radon-Nikodym derivative. Note that this Radon-Nikodym derivative is the density, but that is also a random variable because it is a function of the underlying probability space. So, this is a random variable. For this random variable, let us compute this right side.

Expected value over Q of $Z - \log$ sorry let us take a log here log of the guy expected value e to the power z . This guy equals to expected value under Q of $\log dQ$ by dP ok, make space. This guy is to expected value under Q of \log of dQ by dP - log of expected value under P of e to the power z which is just dQ by dP ; dQ by dP .

Now, if you look at the second term here, this guy is 1, so this second term here is 0 because this expectation is 1. This just changes the measure. It is this expected value of 1 with respect to Q and this guy here is another term disguising our Kullback-Leibler divergence, we have seen this equivalent expression before.

Yeah, so once again, remember that if you want to switch from measure Q to P only, you have to multiply with this density. So, you have this is $= dQ/P - 0$ ok. So, there is indeed a random variable for which we get $D(Q||P)$ on for the right side. So, in fact, this inequality which holds it actually holds with equality because there is something here for which this is =

this. So, to conclude what we have obtained is this another alternative form of the Gibb's variational principle.

(Refer Slide Time: 19:51)

$$D(Q||P) = \max_{Z: E[e^Z] < \infty} [E_Q[Z] - \log E_P[e^Z]]$$

[B] Another variational formula for KL divergence
 Uses Bregman divergence ...
 Given a convex and differentiable function $f: I \rightarrow \mathbb{R}$, where $I \subseteq \mathbb{R}$ is an interval,

I think it is just an alternative form, but maybe you do not want to call it that. So, this is max over all random variable Z . Note that this random variable Z actually has e to the power expected value of Z under P as 1. So, this is finite for sure. So, if you want you can write it as max over all random variables Z such that expected value of this is finite, but as I said this can be drawn.

Expected value under Q of the z , so then, you measure - the correction due to log moment generating function, ok. So, this is we converted Gibb's variational principle into a sort of a variational formula for divergence ok. So, this is the first variational formula for divergence that we saw right. This is what I wanted to say in the first part of this lecture.

Now, moving to the next part of this lecture, we will present another slightly different variational formula again for KL divergence and this one uses something called a Bregman divergence. So, another variational formula for KL divergence ok and this uses thing called Bregman divergence, alright. So, what is this Bregman divergence based formula?

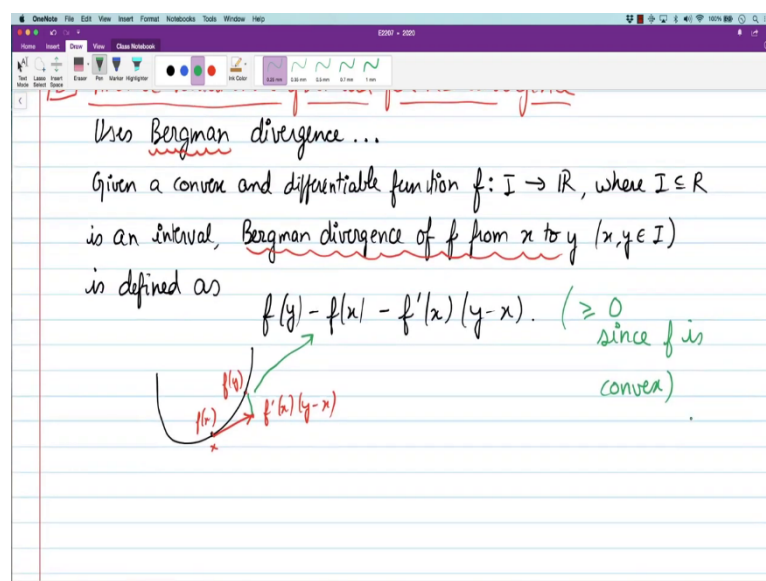
In fact, to ask me what is this Bregman divergence, so, let me start by defining Bregman divergence. So, given a convex function; given a convex and differentiable function f , so on

\mathbb{R} , but let us say just an interval of \mathbb{R} , where this is an interval yeah. So, just I just switched context very quickly.

So, this divergence this variational formulae may I, firstly, you may wonder why are they called variational formulae. I never described that and that is a very big story. But one way to think of this formula is it is a series of bounds and any of these bounds is a lower bound for divergence, but these this connection of bound is tight because there is one which meets it with equality.

And the first formula that we presented is sort of a sort of a family of lower bounds for KL divergence. What I am presenting now is a family of upper bounds for KL divergence ok that is the second variational formula, we are presenting and it uses a different idea. It uses this idea of Bregman divergence.

(Refer Slide Time: 23:45)



So, you are given this convex function f ; convex and differentiable function f . For this such a function, we can define the Bregman divergence of f and it requires two points from x to y ok; here, x comma y are in this interval. So, Bregman divergence from x to y . So, this is the, is defined as yeah.

We can use the notation, but since we never go back to this one, I will not give a notation. It is $f(y) - f(x) - f'(x)(y-x)$. So, you go from x to y , $f(y) - f(x) - f'(x)(y-x)$. So, can someone

recognize what this quantity is? And you have to keep in mind that f is a convex function right. So, since f is a convex function, looks something like this.

We took two points x and y and so, this is let us say x and this is y and we looked at the tangent at x f' prime x , this is here in this direction, so that is the direction f' prime x and we took a step $= x - y$ in this direction that will get you somewhere here. So, that is my y . So, this point here when you move is what is the value of this point?

It is f' prime x into $y - x$ that is the value of this point ok. So, and what about this difference here? So, this is $f y$, this is $f x$ and what we are looking for is how much larger is $f x$ in comparison to $f y$. So, this quantity here is this distance and its always non negative because the function is convex ok.

So, now, we state a result which is a generalization of a well-known result that I will tell I will tell you about later.

(Refer Slide Time: 26:43)

Lemma (Jensen's gap variational formula). $f: I \rightarrow \mathbb{R}$ is convex, diff.

$$E[f(X)] - f(E[X]) = \min_{a \in I} E[f(X) - f(a) - f'(a)(X-a)]$$

Remark. Consider $f(x) = x^2$.

$$\begin{aligned} E[X^2] - E[X]^2 &= \text{Var}(X) = E[(X - E(X))^2] \\ &= \min_a E[X^2 - a^2 - 2a(X-a)] \\ &= \min_a E[(X-a)^2] \end{aligned}$$

So, here is a another variational formula. So, we will just call it a lemma. So, you know this is the Jensen's gap variational formula. So, for a random variable X , consider X f expected value of f of X . So, once again by the way f is a convex differentiable function, maybe f from

I to R as above is convex differentiable. So, now, we consider the Jensen gap, what I am calling Jensen gap.

So, remember that the expected value of $f(X) - f(\text{expected value of } X)$. By Jensen's inequality, this quantity is non-negative and what we are claiming is actually this guy, we will write a variational formula for this guy. It is \inf infimum over all a in the domain of f . So, all the best constant a , that you can choose expected value of $f(X) - f(a) - f'(a)(X - a)$.

In fact, I have written it as infimum, but you see that, but you see that when you substitute a equals to expected value of X . So, this becomes $f(X) - f(\text{expected value of } X)$ and this guy is $-f'(a)$ and something here and so, when you take in linearity of expectation, this linear this expected value becomes 0.

So, in fact, this can be attained by a equals to expected value of X . So, in fact, this is a min ok; this is a min. If expected value of X belongs to this interval, but that is true because X is a random variable on this I , only when only then this can make sense and therefore, expected value of X is also inside this I , right. So, this is this nice formula that that we were after this is the second yeah.

So, we have not still talked about how this is a formula of the rational formula for divergence but before we do that some, one remark. So, let us choose a very simple convex function, ok. Consider $f(x)$ equals to x^2 perhaps the first convex function one can think of ok.

In that case, what is this formula saying? Saying that expected value of x^2 - the mean; so, expected value of X^2 - expected value of X^2 ok. So, this is basically we know we recognize this as variance of X which is expected value of X^2 - expected value of X^2 ok.

This guy is = minimum over all a expected value of $X^2 - a^2 - 2a$ that is the derivative into $X - a$ and if you look at this formula carefully, so you get $+2a^2 - 2aX$. This is nothing but $X - a^2$. So, this brings out the familiar formula ok; yeah, the familiar formula.

(Refer Slide Time: 31:09)

Expected value minimizes the MSE.

Corollary. By applying the variational formula for Jensen gap to $f(x) = x \log x$, we get

$$\mathbb{E}[X \log X] - \mathbb{E}[X] \log \mathbb{E}[X]$$

$$= \min_a \mathbb{E} \left[X \log X - a \log a - (1 + \log a)(X - a) \right]$$

$$= \min_a \mathbb{E} \left[X \log \frac{X}{a} - (X - a) \right]$$

$\rightarrow \mathbb{E}[X] = a$

It says that expected value minimizes the Mean Squared Error, MSE. This is the mean squared error; mean squared error and the constant which minimizes this is expected value that is something you have seen, perhaps you know about it; otherwise, you can here is a proof ok by applying this formula, which we have not proved yet. So, that is what we have seen.

Now, an interesting so what happens when you look at the other convex function that we have been using in this course? So, here is the corollary that we were after by applying the variational formula for Jensen gap to. So, now, we applied to the other convex function $x \log x$. Remember just like we have extended (Refer Time: 32:27) time to entropy method by replacing this x^2 with $x \log x$, we do the same thing here.

So, we apply this here, we get expected value of $X \log X$ - expected value of $X \log$ expected value of X . This guy here is = minimum over all a constants a, expected value of let us see $f(X) - f(a)$, $f'(a)(X - a)$. So, what is $f'(a)$? That is $1 + \log a$ into $X - a$.

(Refer Slide Time: 33:29)

The image shows a digital notepad with handwritten mathematical derivations. The first line is $= \min_a \mathbb{E} \left[X \log \frac{X}{a} - (X - a) \right]$. The second line, enclosed in a box, is $\Rightarrow \text{Ent}(X) = \min_{a>0} \mathbb{E} \left[X \log \frac{X}{a} - (X - a) \right]$. The third line is $\text{When } X = \frac{dQ}{dP}, \text{ Ent}(X) = D(Q||P) = \min_{a>0} \mathbb{E}_P \left[\frac{dQ}{dP} \left(\log \frac{dQ}{dP} - \log a \right) - \left(\frac{dQ}{dP} - a \right) \right]$. The final line is $= \min_{a>0} \mathbb{E}_P \left[\dots \right]$.

So, it is the same as minimum over a expected value of $X \log$. So, this $a \log a$ cancels with this $a \log a$ here and so, you are left with $X - a + X \log a$; $- X \log a$. So, you get this times this $- X - a$. So, what we have? If you see this guy there, then you recognize it perhaps as the entropy.

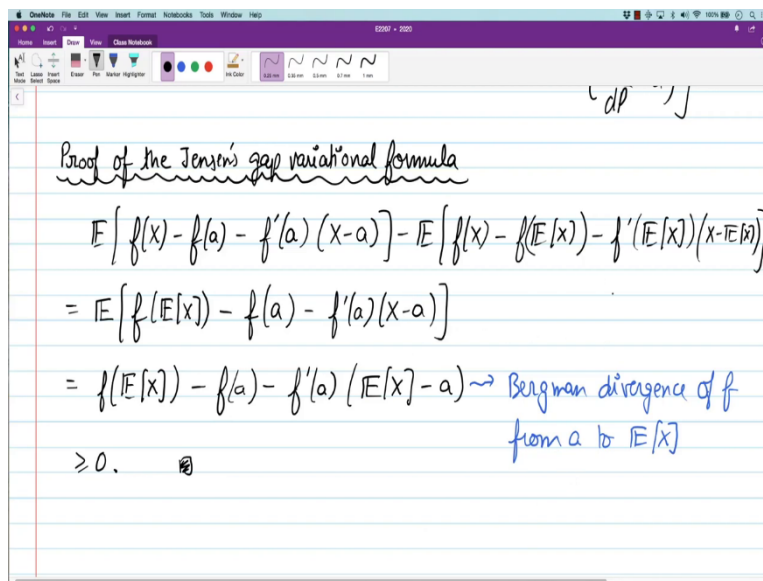
So, we have shown that entropy of X is = minimum over a expected value of $X \log X$ by a . Let us just because we are talking about entropy, we take non-negative random variables here and this just to be shown, I clarify that we choose I to be 0 to infinity here. In fact, we should rule out 0 also, no problem; $- X - a$ ok.

So, this is the this is the formula variational formula for entropy that will be very useful for us and in fact, as we know that when $X > 0$ is given by. So, we so, let us say when X is $= \frac{dQ}{dP}$ for some Q that is absolutely continuous with respect to P entropy of X is = divergence between Q and P .

And then, this guy we are saying is $= \min$ over $a > 0$, expected value of this expectation is with respect to P here. So, $\frac{dQ}{dP} \log \frac{dQ}{dP} - \log a - \frac{dQ}{dP} + a$. So, this guy here is $=$ this guy here, \min over $a > 0$ expected value over P , I think it is ok, yeah. So, this is another formula for divergence and now this time, we have a \min here instead of \max ok.

So, it is better to remember this formula in terms of entropy. The important point is you have a min here instead of max. This is the second formula for variational formula for divergence that we wanted alright. So, this is all corollary of this Jensen's gap variational formula, which is in terms of this Bregman divergence of f from a to x . So, how do we show this formula? This is the proof of the Jensen's gap variational formula. I am making this name up, it is not a standard name; but you know why I am calling it ok, alright.

(Refer Slide Time: 37:11)



The image shows a handwritten proof in a OneNote application. The title is "Proof of the Jensen's gap variational formula". The proof consists of the following steps:

$$\begin{aligned} & \mathbb{E} [f(X) - f(a) - f'(a)(X-a)] - \mathbb{E} [f(X) - f(\mathbb{E}[X]) - f'(\mathbb{E}[X])(X-\mathbb{E}[X])] \\ &= \mathbb{E} [f(\mathbb{E}[X]) - f(a) - f'(a)(X-a)] \\ &= f(\mathbb{E}[X]) - f(a) - f'(a)(\mathbb{E}[X] - a) \sim \text{Bregman divergence of } f \\ & \quad \text{from } a \text{ to } \mathbb{E}[X] \\ &\geq 0. \end{aligned}$$

So, for Jensen's gap variational formula, how do we show this? So, since this is a generalization of this proof of the fact that expected value minimizes the mean square error we will follow a similar proof. We will take the difference between the Jensen gap, so in this right side here that is how you prove this for this case.

So, this right side here, this expression here this one and for $a = \text{expected value of } X$. So, $f(X) - f(a) - f'(a)(X-a)$ that is something we would like to claim exceeds the other thing $f(X) - f(\mathbb{E}[X]) - f'(\mathbb{E}[X])(X - \mathbb{E}[X])$, I would like to claim that this whole thing is non-negative.

So, let us just take expectation outside and we see that $f(X)$ cancels. So, you get $f(\mathbb{E}[X]) - f(a) - f'(a)(\mathbb{E}[X] - a)$ and now, if you see this expected value here is 0.

So, that is all ok, because this is a constant, this these two match. So, this is this ok, but this is exactly $= f$ of expected value of $X - f$ of $a - f$ prime a expected value of $X - a$, what is that?

That is the Bregman divergence of f from a to expected value of X and this guy is non-negative ok and of course, you can have equality here by setting $a =$ expected value of X and that is the proof. This is always non-negative and there is equality when a equals z expected value of X alright. So, this is the proof of Jensen's gap variation formula and we can apply different functions here and get all these different variational formula.

So, two takeaways from this lecture. Actually there are three takeaways, but I am I need not overplay the first one because we will spend a lot of time on that. And the first one, the main one perhaps in the long term is this Gibb's variation formula here. It allows you it gives you a new formula for log moment generating function ok. We will come back to this later. But for our immediate use, we want to recall two variational formulae.

One for this divergence, it is the max of expected value of max over all random variable z that expected value of z under $Q -$ the log moment generating function with respect to P . So, log of expected value with respect to P of e^z ok, that is a divergence and second is this Gibb's variational formula; a second is this variational formula for this Jensen's gap thing that I derived.

But instead of that particular formula, what we can remember is its consequence which gives a nice formula variational formula for entropy which is similar to divergence that entropy of X is $=$ min of all non-negative a expected value of $X \log X - \log a - X - a$ ok. So, in fact, we our motivation for covering this variational formula or both this variational formula at this point was so that we could derive this particular formula for entropy.

And in the next lecture, we will apply this formula to obtain a sort of a log Sobolev inequality that will different log Sobolev inequality. Till now, we have seen the Binary log Sobolev inequality and the Gaussian log Sobolev inequality as its consequence and now, we will use this variational formula for entropy to derive the so called modified log Sobolev inequality which will allow us to extend the bounds that we saw in the previous, the concentration bound that we saw in the previous lecture for uniform random variables over the Boolean

hypercube and for Gaussian random variables to arbitrary random variables ok. That is what we will do in the next class. Just remember this formula.

See you in the next lecture.