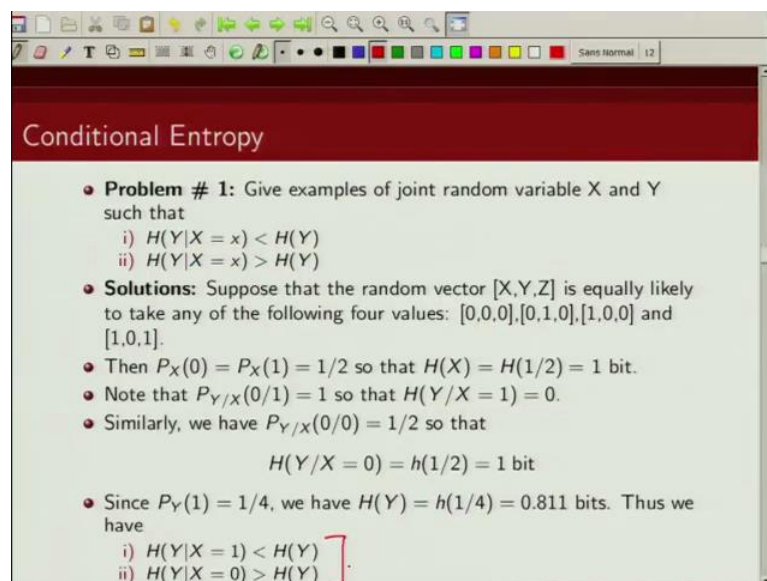


An Introduction to Information Theory
Prof. Adrish Banerjee
Department of Electronics and Communication Engineering
Indian Institute of Technology, Kanpur

Lecture – 2B
Problem Solving Session-I

Welcome to the course on an Introduction to Information Theory. In this lecture, we will try to solve some problems. So, this is a session devoted to problem solving.

(Refer Slide Time: 00:26)



The screenshot shows a presentation slide with a red header and a light green background. The title is "Conditional Entropy". The content includes a problem statement, a solution, and a list of properties. The solution part is highlighted with a red box.

Conditional Entropy

- **Problem # 1:** Give examples of joint random variable X and Y such that
 - $H(Y|X = x) < H(Y)$
 - $H(Y|X = x) > H(Y)$
- **Solutions:** Suppose that the random vector $[X, Y, Z]$ is equally likely to take any of the following four values: $[0, 0, 0]$, $[0, 1, 0]$, $[1, 0, 0]$ and $[1, 0, 1]$.
- Then $P_X(0) = P_X(1) = 1/2$ so that $H(X) = H(1/2) = 1$ bit.
- Note that $P_{Y|X}(0/1) = 1$ so that $H(Y|X = 1) = 0$.
- Similarly, we have $P_{Y|X}(0/0) = 1/2$ so that
$$H(Y|X = 0) = h(1/2) = 1 \text{ bit}$$
- Since $P_Y(1) = 1/4$, we have $H(Y) = h(1/4) = 0.811$ bits. Thus we have
 - $H(Y|X = 1) < H(Y)$
 - $H(Y|X = 0) > H(Y)$

So, let us start with the first problem. So, give examples of, joint random variable X and Y , such that the conditional entropy of Y , given a particular instant of X , is less than equal to uncertainty in Y . And uncertainty in Y , given a particular instant of an event X that has happened, that entropy is greater than entropy of Y . So, let us take an example, to illustrate this. So, we have seen in the lecture, before that, conditioning cannot increase, and conditioning under random variable, cannot increase entropy; However, as you can see from this example, conditioning on a particular event, can result in increase in uncertainty.

So, let us take an example. So, let us take, we have a, let us consider, we have a random vector $X Y$ and Z , and it takes 4 possible values. And what are those 4 possible values? $0 0 0, 0 1 0, 1 0 0$ and $1 0 1$, then, what is X , X is $0, 0, 1, 1$. So, then, probability of X being 0 is, so, it is 2 by 4 , similarly, probability of X being 1 is, 2 by 4 . So, this is half.

Then what is the uncertainty in X? That is basically, given by, H of half, which is minus half log half, minus half log half. So, that is 1 bit. So, uncertainty in X is 1 bit. Now let us consider the uncertainty in Y, given X is 1. So, when is X 1? X is 1 in this case, and X is 1 in this case. Now what happens to Y when X is 1? We can see here, when X is 1, here Y is 0, and here also Y is 0. So, uncertainty in Y, given X, is 0. Because when X is 1, Y is always 0 in this example, right?

So, uncertainty in Y, given X equal to 1 is 0. Now similarly, let us try to calculate, uncertainty in Y, given X is 0. So, when X is 0, this is X is 0, this is X is 0. So, what happens to Y, when X is 0? In one instance it is 0, in other instance it is 1. So, with probability half it is 0, and probability half, it is 1. So, what is the uncertainty in Y, given X equal to 0? That is basically this quantity, and again this is minus half log of half, minus half log of half, which is 1 bit. So, what have we shown so far? We have shown that uncertainty in Y, given X equal to 1 is 0. Uncertainty in Y, given X equal to 0, is 1. Now let us compute uncertainty in Y. So, what is Y? Y is 0, Y is 1, Y is 0, Y is 0. So, probability of Y being 0 is half, so, probability of Y being 0 is $\frac{3}{4}$, and probability of Y being 1 is $\frac{1}{4}$. So, then what is the uncertainty in Y? That is given by this. So, this will be minus, $\frac{1}{4}$, log of $\frac{1}{4}$, and minus $\frac{3}{4}$, log of $\frac{3}{4}$. And that comes out to be 0.811 bits.

So, you can see in this particular example H of Y, is 0.811 bits. Now given, X equal to 1, uncertainty in Y, given X equal to 1 is 0. So, then H of Y, given X equal to 1, is less than H of Y, correct? And what is this? H of Y given X equal to 0 that is 1, which is greater than H of Y. So, hence we have shown that, condition on a particular event, it is possible that uncertainty in Y, may become more than uncertainty in Y without, conditioning on this event, or, that is this case, or the conditioning on a particular event, can reduce the uncertainty in Y.

Please note, however, this is always true. You can work it out from here also, we know that uncertainty in Y given X equal to 1, should multiply by probability of, X equal to 1 that is half. And what is the probability of X equal to 0 that is also half. So, if you compute uncertainty in Y given X, that is uncertainty in Y given X is given by, probability of X equal to 1, into H of Y given X equal to 1, plus probability of X equal to 0, multiplied by uncertainty in Y, given X equal to 0. This quantity is half, this quantity is 0, this quantity is half, and this quantity is 1.

So, this comes out to be 0.5, which is less than H of Y. H of Y is 0.811 bits. So, hence we have shown, with an example, with an example, that these both conditions are possible.

(Refer Slide Time: 07:27)

Mutual Information

- **Problem # 2:** Give examples of joint random variable X , Y and Z such that
 - $I(X; Y|Z) < I(X; Y)$
 - $I(X; Y|Z) > I(X; Y)$

Solutions: Let X , Y and Z form a Markov Chain.

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y)$$

- We note that $I(X; Z|Y) = 0$, by Markovity, and $I(X; Z) \geq 0$. Thus,

$$I(X; Y|Z) \leq I(X; Y) \quad (1)$$
- ii) Let X and Y be independent fair binary random variables, and let $Z = X + Y$.
 - Then $I(X; Y) = 0$, but $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = H(X|Z) = P(Z=1)H(X|Z=1) = \frac{1}{2}$ bit.

Handwritten notes in red and blue ink on the slide include:

 $H(X|Z) = H(X|Z=0)P(Z=0) + H(X|Z=1)P(Z=1)$

 $Z = X + Y$

 $H(X|Z) = 0$

Now, let us look at another problem. So, given example of Joint Random Variable X , Y and Z , such that, this condition holds, the Mutual Information between X and Y , condition on Z , is less than Mutual Information between X and Y , and Mutual Information between X and Y condition on Z , is greater than Mutual Information between X and Y . So, to prove the first case we will take an example, let us assume that, X , Y and Z , forms a Markov chain. So, X , Y and Z , forms a Markov chain, then probability of Z , given Y and X , is equal to probability of Z , given Y .

Now, when X , Y and Z forms a Markov chain, let us first write the Mutual Information between X and Y, Z , using chain rule. So, this is an application of chain rule, for mutual information. So, using chain rule for mutual information, I can write this as Mutual Information between X and Z , plus Mutual Information between X and Y , given Z . Now I can apply chain rule in another fashion. So, if I apply chain rule again, I can write the same thing, as Mutual Information between X and Y , plus Mutual Information between X and Z , given Y . Now what is this quantity? Mutual information between X and Z , given Y ; now since X , Y and Z , forms a Markov chain. So, given, given, given Y , Mutual Information between X and Z is 0. Because if I know Y , I have complete knowledge, basically I do not need X . So, Mutual Information between X and Z , given Y is 0, and

this is because X, Y and Z , forms a Markov chain correct.

So, then what we can write here, here is, Mutual Information from this, lets call it A and B . So, equate A and B , what we get is Mutual Information between X , and is equal to Mutual Information between X and Z , plus Mutual Information between X and Y , given Z . Now, what is Mutual Information? Mutual Information is divergence between the joint probability distribution, and the marginals. So, Mutual Information between X and Z , would be divergence of, joint distribution of X and Z , and marginals between that, and the marginals P_X and P_Z . Now we know that divergence is always greater than equal to 0. So, this quantity is greater than equal to 0. So, then what have we proved? We have proved, that if this quantity is greater than equal to 0, then we have shown that Mutual Information between X and Y , is greater than equal to, Mutual Information between X and Y , given Z .

So, this we have proved. Now we are going to prove that, Mutual Information between X and Y , given Z , is greater than equal to, Mutual Information between X and Y . So, we will take, we will consider one example. So, let us consider that X and Y , are independent fair binary random variables. And Z is given by X plus Y . Now what is the Mutual Information between X and Y ? Because X and Y are independent, binary random variable, they do not convey any information about each other. So, the Mutual Information between X and Y is 0. Now what is the Mutual Information between X and Y , given Z ? Now following the definition of Mutual Information, we can write the expression for Mutual Information between X and Y , given Z as, uncertainty in X , given Z , minus uncertainty in X , given Y and Z . Now let us first look at, each of these terms. So, let us first look at this term. Uncertainty in Y , given uncertainty in X , given Y and Z , if Y and Z are given, clearly X is also known. If Y and Z are given, X is also known. So, there is no uncertainty in X , if Y and Z are known. So, then in this particular example, this term will be equal to 0.

So, then Mutual Information between X and Y , given Z , can be written as uncertainty in X , given Z . Now uncertainty in X , given Z , can be written as, uncertainty in X given Z , is equal to 0, multiplied by probability of Z being 0, plus, uncertainty in X , given Z is 1, probability of Z equal to 1. Now what happens when Z is 0? when Z is 0, since X and Y are binary random variables so, we know if Z is 0, Y will also be 0. So, there is no uncertainty in X , when Z is 0. So, this term again, goes to 0. So, then Mutual Information

which is this term, is equal to this we have proved, is now then equal to, this quantity. Now what is the uncertainty in X, given Z is 1? When Z is 1, X could be 0, and Y could be 1, or X could be 1, and Y could be 0.

So, this uncertainty in, and probability of Z equal to 1, is Z, could, could take values 0 and 1 with equal probability. So, then uncertainty in X, given Z equal to 1 is. So, this quantity, X given Z equal to 1 is, basically H of half, which is 1, multiplied by probability of Z equal to 1, this would give us 0.5. So, then Mutual Information between X and Y, given Z, we have shown that this is equal to, uncertainty in X given Z. We have shown that this is equal to uncertainty in X, given Z equal to 1, into probability of Z equal to 1, and this we have shown is equal to 0.5 bits. So, in this particular example, you can clearly see that, the Mutual Information between X and Y, given Z, which is given by 0.5 bits, is more than Mutual Information between X and Y, which is 0.

So, hence we have shown 2 different examples that, Mutual Information between X and Y, condition on Z, can be greater than Mutual Information between X and Y, or it can be less as well.

(Refer Slide Time: 16:06)

Divergence

- **Problem # 3:** Let $P_X(X = 0) = P_X(X = 1) = 0.5$, $Q_X(X = 0) = 0.25$, $Q_X(X = 1) = 0.75$ and $R_X(X = 0) = 0.2$, $R_X(X = 1) = 0.8$. Show that triangle inequality does hold for divergence, i.e. $D(P_X||R_X) > D(P_X||Q_X) + D(Q_X||R_X)$
- **Solution:**

$$D(P_X||Q_X) = 0.5 \log \frac{0.5}{0.25} + 0.5 \log \frac{0.5}{0.75} = 0.208$$

$$D(Q_X||R_X) = 0.25 \log \frac{0.25}{0.2} + 0.75 \log \frac{0.75}{0.8} = 0.011$$

$$D(P_X||R_X) = 0.5 \log \frac{0.5}{0.2} + 0.5 \log \frac{0.5}{0.8} = 0.322$$
- Since, 0.322 > 0.208 + 0.011 = 0.219, triangular inequality is not satisfied.

Now let us look at another problem. We want to demonstrate, with a help of a simple example, that divergence, the relative entropy, does not satisfy triangle inequality. So, I have given you, 3 probable distribution, P of X, X equal to 0, and X equal to 1 is same as 0.5. Q of X, X equal to 0 is given by 0.25, and X equal to 1 is 0.75. So, these are all

(Refer Time: 16:41) binary random variables 0 and 1, and R of X , X equal to 0 is 0.2 and R of X , X equal to 1 is 0.8. We want to show, that triangular inequality, in other words divergence between P and Q , plus divergence between Q and R , is not greater than divergence between P and R .

So, with this simple example, we want to illustrate that, triangular inequality does not hold for divergence. Soon, note the definition of divergence between two distribution, P and Q , is expected value of \log of, P by Q , and expectation is taken respect to P . So, this would be then, P of X equal to 0 is 0.5, \log of 0.5 when probability of Q X equal to 0 is 0.25. Similarly P of X equal to 1 is 0.5, this is P of X equal to 1, and Q of X equal to 1 is 0.75. So, this is the divergence between P and Q . Now to find the divergence between Q and R , this is probability of Q , X , X being 0, probability of. So, this is basically Q X , X being 0, and this is R being 0. Similarly, this is Q of X , X equal to 1. This is this quantity and this quantity, and this is R X , X equal to 1. That is this quantity. This is follows exactly from the definition of divergence, and similarly we can find out the divergence between P and R that is basically this quantity is P of X , X equal to 0, same quantity, this is R of X , X equal to 0. And this is P X , X equal to 1, same thing, and this is R of X , X equal to 1.

Now, note here, this plus this, is not greater than this. So, we have shown with a simple example that, triangle inequality, does not hold for divergence. So, you can see, divergence between P and R is more than divergence between P and Q , plus, divergence between Q and R . So, the triangular inequality does not hold.

(Refer Slide Time: 19:35)

Mutual Information

- **Problem # 4:** Consider a discrete memoryless channel with inputs X and outputs Y . The input X takes values from a ternary set with equal probability and it is known that the probability of error for the system is p . Using Fano's lemma, find a lower bound to the mutual information $I(X; Y)$ as a function of p .
- **Solutions:** Mutual information can be written as
$$I(X; Y) = H(X) - H(X|Y)$$
- By Fano's inequality, we get
$$H(X|Y) \leq H(P_e) + P_e \log(3 - 1) = H(p) + p$$
- Thus
$$I(X; Y) \geq H(X) - H(p) - p = \log 3 - H(p) - p$$

Now, let us consider a discrete memoryless channel. So, we have a channel, discrete memoryless channel, that has input X , and output Y . The input X , takes values from ternary set, with equal probability. So, let us say at 0, 1 and 2, with equal probability. It is known that the probability of error, of this system is P . So, the probability of error for this system, where we are sending X through this channel, this probability of error is given by P . Now using Fano's lemma, you have been asked to find a lower bound on Mutual Information, between X and Y , as a function of P .

So, what is Mutual Information? So, from the definition of Mutual Information, we can write Mutual Information between X and Y as, uncertainty in X , minus uncertainty in X , given Y . Now from Fano's lemma, we know that this uncertainty in X , given Y , note this is in the Fano's lemma that we have proved, this is my U , and this is estimate of U , we had. So, uncertainty in X , given Y , from Fano's lemma, can be written as, binary entropy function of the error probability, plus error probability log of, L minus 1, where L is the number of possible values, this random variable is taking. Now in this particular example, my random variable takes 3 different values, with equal probability. So, then, L in this case is 3. So, log of 3 minus 1, that is basically 1, and we know probability of error is P . So, then using Fano's lemma, I am able to get an upper bound on, uncertainty in X , given Y , and this is given by this expression.

Now, if I plug-in this upper bound on, $H(X|Y)$ in this expression I will get a lower

bound on, Mutual Information between X and Y. So, plugging in this value of, uncertainty in X, given Y, plugging in this upper bound on this, I get a lower bound on Mutual Information. So, Mutual Information is H X minus, H of X, given Y, which is upper bounded by this. So, then the lower bound on, Mutual Information is given by this expression.

(Refer Slide Time: 22:24)

Concave Function

- **Problem # 5:** Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. the mutual information $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$
- **Solutions:** To prove, we expand the mutual information

$$I(X; Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x)$$

concave $f(x)$
concave $f_p(x)$
linear

- If $p(y|x)$ is fixed, then $p(y)$ is a linear function of $p(x)$.
- Hence $H(Y)$, which is a concave function of $p(y)$, is a concave function of $p(x)$.
- The second term is a linear function of $p(x)$. Hence, the difference is a concave function of $p(x)$.

Next, we are going to show, that, Mutual Information, is a concave function of P of X, for a fixed P of Y, given X. So, Mutual Information is a concave function of P of X. for a given fixed P of Y, given X. So, how do we prove this? So, we write down from the definition of Mutual Information between X and Y, we can write it as, uncertainty in Y, minus uncertainty in Y, given X.

Now, the uncertainty in Y, given X, I can write it in this particular way. These uncertainties in Y, given X equal to a particular event X, multiplied by the probability of that event, right? Now lets look at, so, now, let us look at each of these terms separately. Let us look at H of Y. Now we have already proved that, entropy is a concave function, of P of Y. And what is P of Y? P of Y if we fixed, P of Y given X, then P of Y is a linear function of P of X. And we know that H of Y, entropy function, is a concave function of P of Y. So, if it is a concave function of P of Y, it will also be a concave function of P of X. Because P of Y, for a fixed P of Y, given X, is a linear function of P of X. So, then we can say that, since H of Y, is a concave function of P of Y, and P of Y is a linear function

of P of X , so then H of Y , must also be, a concave function of P of X . So, what we have proved so far is, this term H of Y , is a concave function of P of X .

Now, let us look at this. This is a linear function of P of X , right? You can look, look at this expression; it is a linear function of P of X . So, then we have Mutual Information which is, consist of 2 term, one, which is a concave function of P of X , another, which is a linear function of P of X . So, the sum of that also, will also be a concave function of P of X . So, as I said, since the second term, is a linear function of P of X the difference or the sum, will also be a concave function. So, we have now proved that, Mutual Information between X and Y is a concave function of P of X . Because we have shown that, Mutual Information is nothing, but subtraction of these two functions. This function is a concave function of P of X . This function is a linear function of P of X , this is a concave function. So, concave function minus linear function is a concave function. So, this will be a concave function of P of X . And that is our proof.

So, I will stop here. In the next class, we will talk about; we will start discussion on, Source Compression Algorithm. So, we will start with Block Two Variable Length Coding.

Thank you.