

Information Theory, Coding and Cryptography
Dr. Ranjan Bose
Department of Electrical Engineering
Indian Institute of Technology, Delhi

Module – 01
Source Coding
Lecture – 01

(Refer Slide Time: 00:19)

Information Theory, Coding and Cryptography

Outline

- Uncertainty and Information
- Self Information
- Mutual Information
- Average Mutual Information

Indian Institute of Technology, Delhi

2

Ranjan Bose
Department of Electrical Engineering

Welcome to this module on Source Coding. Let us start with a brief outline of today's talk. We will start with Uncertainty and Information, then we look at the notion of Self Information. We will then graduate to Mutual Information and finally, we will talk about Average Mutual Information. So, this is the general outline for today's talk.

(Refer Slide Time: 00:45)


Information Theory, Coding and Cryptography

Uncertainty and Information

Consider the following sentences:

- Tomorrow, the sun will rise from the East.
- The phone will ring in the next one hour.
- It will snow in Delhi this winter.

- The three sentences carry different amounts of **information**.
- In fact, the first sentence hardly carries any information.
- Everybody knows that the sun rises from the East and the probability of this happening again is *almost* unity
- *"Making predictions is risky, especially when it involves the future."* - N. Bohr

 Indian Institute of Technology, Delhi

3

Ranjan Bose
Department of Electrical Engineering

So, let us start with this notion of information. Let us start with these three sentences. Sentence one, tomorrow the sun will rise from the east or the phone will ring in the next 1 hour and the third sentence, it will snow in Delhi this winter.

Now, if you look, if you count the number of letters in each of the sentences, they are different. In fact, the first sentence appears to have the maximum number of letters and if you map one letter of the English alphabet to 8 bits, you will find the maximum number of bits are really used to represent the first sentence. But does it really warrant so many bits? Does it really need so many bits to be used to express whatever is being conveyed; tomorrow the sun will rise from the east?

So, let us have an intuitive feel first and then we will go to the mathematical rigor. The first sentence intuitively probably does not convey anything right. It hardly contains any information, why? Everybody knows, the sun will rise from the east regardless of how many bits I put in to express that sun will rise from the East; however, we should be a little careful making predictions as risky, when it involves the future. But coming back on the serious note, we will realize that a statement which has a very very high probability of occurrence is likely to contain very little information as the first sentence.

Now, if you go to the next one, which says looks like the phone will ring in the next 1 hour. Well it may ring or it may not. So, we look at the probability of that happening and

somehow you will see that the information contained in that sentence, the phone will ring in the next 1 hour is somehow related to the probability of occurrence of that event.

Now we come to the last sentence, it will snow in Delhi this winter. It has really never snowed in Delhi and if I say, it will it is likely to snow in Delhi this winter, it is likely to raise eyebrows. It will catch your attention, simply because it is a rare event. It is never happened and the probability of it happening is very very low.

So, it is possible that the information contained in the third sentence is really the highest even though if you measure the number of letters it has it is the least. So, somewhere there appears to be a mismatch. The number of bits we are using to represent an English sentence is completely different from the amount of information it is conveying.


(Refer Slide Time: 03:53)

Information Theory, Coding and Cryptography

Uncertainty and Information

- Tomorrow, the sun will rise from the East (most probable, least information).
- The phone will ring in the next one hour.
- It will snow in Delhi this winter (least probable, most information).

- **Intuitive Feel** : Occurrence of a *less probable* event conveys more information
- Since a lower probability implies a higher degree of uncertainty (and *vice versa*), a random variable with a higher degree of uncertainty contains more information.
- We will use this correlation between uncertainty and information for physical interpretations

 Indian Institute of Technology,
Delhi4Ranjan Bose
Department of Electrical Engineering

So, let us look at the intuitive feel. We observed from the first three sentences that the occurrence of a less probable event conveys more information. So, if you have a higher degree of uncertainty which is tantamount into a lower probability of occurrence, then it appears to be inversely connected to the quantity of information, it is conveying. In this lecture, we will use this correlation between uncertainty and information for all physical interpretation. As we move along, we will see that there is a strong intuitive feel to this information conveyed by a random variable.

(Refer Slide Time: 04:50)


Information Theory, Coding and Cryptography

Uncertainty and Information

- **Intuitive Feel** : Occurrence of a less probable event conveys more information
- Consider a discrete random variable X with possible outcomes $x_i, i = 1, 2, \dots, n$. The **self information** of the event $X = x_i$ is defined as

$$I(x_i) = \log\left(\frac{1}{P(x_i)}\right) = -\log P(x_i).$$

- When the base of the logarithm is 2 the units of $I(x)$ are in **bits**
- When the base is e , the units are in **nats** (natural units).

 Indian Institute of Technology,
Delhi5Ranjan Bose
Department of Electrical Engineering

So, let us start with our first point, our first observation that occurrence of a less probable event conveys more information. So, let us take it a one step further. Consider a discrete random variable X with possible outcomes x_i 's i equal to 1 to up to n . Now here we are saying up to n , but it could be up to infinity. So, the self-information of the event X is equal to x_i , we mathematically defined as $I(x_i) = \log$ over \log of 1 over $P(x_i)$ where $P(x_i)$ is the probability of occurrence of the event x_i . Well you can write it as minus $\log P(x_i)$.


Now, the log we will come to it; why there is a log, but assuming that the base of log is 2; then the units of information here, self-information is expressed in bits. At the same time, if you make the log the natural logs, then the units are nats which is the natural units.

(Refer Slide Time: 06:00)

Information Theory, Coding and Cryptography

Example

- Consider a binary source which tosses a fair coin
- It produces an output equal to 1 if a head appears and a 0 if a tail appears.
- For this source, $P(1) = P(0) = 0.5$. The information content of each output from the source is
$$I(x_i) = -\log_2 P(x_i)$$
$$= -\log_2(0.5) = 1 \text{ bit}$$
- Indeed, we have to use only one bit to represent the output from this binary source
- We use a 1 to represent H and a 0 to represent T.

 Indian Institute of Technology, Delhi 6 Ranjan Bose
Department of Electrical Engineering

Let us start with an example. Consider a binary source which tosses a fair coin. So, how come a source is tossing a coin? Let us imagine a person sitting on a chair, tossing a fair coin and each time a head comes up, he shouts a 1 and if a tail appears, he shouts a 0. So, indirectly we assume that a source is generating a sequence of 1's and 0's. Of course, it depends on the heads or tails it is generating.

Now since it is a fair coin, let us assume that the probability of 1 and probability of 0 is equal to half. The question, then we ask is how much information is contained in each of the outputs? For example, how much information is in the event that heads come head comes up or the tail comes up. So, if you plug in the values, you find that I of x_i if you put in the formula is minus log to the base 2 $P \times i$, but $P \times i$ for head or for tail is half and so, it is 1 bit.

So, it intuitively fits in that every time either a head comes up or a tail comes up, you can represent that information in 1 bit. Luckily it falls into place because well, we had already designated 1 to be the head and 0 to be the tail right. So, it fits our intuition. But the comfort level stops right here because suppose I tell you that it is not a fair coin. It is indeed a biased coin with say probability of 1, probability of head appearing is 0.8 and probability of tail is 0.2. Then if you plug into the value then $I \times i$ is not going to be 1 bit, it will be something less.

So, that brings us to a very interesting question. Even though, I toss a coin and I get a full 1 or a full 0, the mathematical rendering tells us that we require less than 1 bit to represent the outcome. Here is where the interesting stuff starts. How can we represent the outcome of a head or a tail by fewer than 1 bits? That is the question we will address in the subsequent slides.

But for now since we are using a fair coin, we are happy that we use 1 bit to represent either a head or a tail from an information theoretic point of view. It is not just for convenience that a head is a 1 and tail is a 0, information theory tells me that if a fair coin is tossed, then the outcome can comfortably be represented using 1 bit, but not so for an unfair coin.

(Refer Slide Time: 09:13)


Information Theory, Coding and Cryptography

Example (cont'd)

- Suppose the successive outputs from this binary source are statistically independent, i.e., the source is memoryless.
- Consider a block of m binary digits.
- There are 2^m possible m -bit blocks, each of which is equally probable with probability 2^{-m} .
- The self information of an m -bit block is

$$I(x_i) = -\log_2 P(x_i)$$

$$= -\log_2 2^{-m} = m \text{ bits}$$
- Again, we observe that we indeed need m bits to represent the possible m -bit blocks

 Indian Institute of Technology,
Delhi

7

Ranjan Bose
Department of Electrical Engineering

Now, we do not stop at 1 toss. We toss the same coin several times. In fact, m times and suppose this source is memoryless that is my output of the second toss does not belong to or does not depend on the first toss, outcome of the first toss and subsequently any of the later tosses. So, if this is really an independent statistically independent series of binary bits, then if you look at it there are 2^m possible m bit blocks which are equally probable with a probability 2^{-m} .

So, now we ask for this series of m tosses, what is the self-information? If you plug in the value $I(x_i) = -\log_2 2^{-m}$, you get the answer m bits. We are relieved once more because here there were m tosses each time a


head or a tail came up and then to represent each one of them, I can represent it using m bit blocks. So, if suppose a head, head, tail, tail, head came; then I will write 1 1 0 0 1 because 1 represents a head and 0 represents a tail. So, far so good, we are representing the outcomes using certain number of bits.

(Refer Slide Time: 10:45)

Information Theory, Coding and Cryptography

Example

- Consider a discrete, memoryless source (source C) that generates *two* bits at a time.
- This source comprises of **two** binary sources (sources A and B), each source contributing one bit.
- The two binary sources within the source C are independent.
- Intuitively, the information content of the aggregate source (source C) should be the *sum* of the information contained in the outputs of the two independent sources that constitute this source C .
- $P(C) = P(A)P(B) = (0.5)(0.5) = 0.25$
- $I(C) = -\log_2 P(x_i) = -\log_2(0.25) = 2$ bits

 Indian Institute of Technology, Delhi 8 Ranjan Bose
Department of Electrical Engineering

Now, we extend this analogy a little further. Let us talk about a source C that generates 2 bits at a time. So, far we were concerned with only 1 bit at a time. Now consider 2 binary sources A and B which form a bigger source C combination ok. So, now, what we can do is if we assume A and B are independent; so independently A generates a bit, it could be a toss of a coin and B has its own coin and it tosses it and generates a 1 or a 0.

These two bits are now thrown out by the source C and this is the output of the source C . So, my source C is generating 2 bits at a time. What do we expect? Well from information theoretic perspective, the information contained in C , the big source C , the aggregate source C should be the sum of the information generated by A or B for example, if B switches off and only A keeps generating its bits. So, the total information coming out of the source C is the information of A .

Similarly, if A shuts down and only B keeps generating information, then it is the output of B . So, it intuitively, we are dealing with the sum of the information right. And here if you look at it, if A and B are indeed independent, then the probability of the occurrence

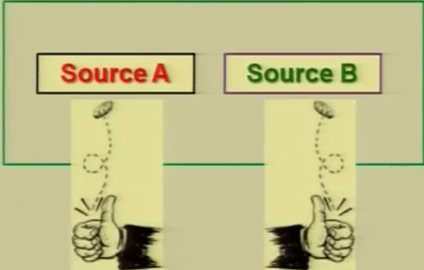
of 2 bits, it is related by the product because independent in events their probabilities multiply and if you look at the self information for C, it comes out to be 2 bits.

(Refer Slide Time: 12:45)


Information Theory, Coding and Cryptography

Why the logarithm?

- Consider two independent sources
- Independent events \Rightarrow Probabilities multiply
- Independent sources \Rightarrow Information must add up
- Logarithm seems to do the job !



Source A Source B

 Indian Institute of Technology, Delhi 9 Ranjan Bose
Department of Electrical Engineering

So, far we are consistent. This gives us an answer to why the proposers of this logarithmic measure of information chose the log? See they needed a mathematical operation that could change the product into a sum and it should be such that independent events whose probabilities multiply, the information contained in that should add up. Logarithm seems to do the job. So, if you do a visual representation, this green box is my aggregate source and I can pack in as many sources as I want and I am interested in the amount of information being generated by this aggregate source green box. So, I have packed in source A and I have got source B, they can be individually present or they can be independently present.

So, why do not we put them together source A and source B and we would like to know what is the total information contained? Only when A and B are independent would their probabilities multiply and there an we will have the output as the sum of the 2 sources. So, the independent sources should be such that the information should add up. So, for all you know source is toss in its own coin A source B toss in its own coin B and the total information generated is the sum of A and B. Please note coin A could be a fair coin, coin B could be a biased coin.

So, the rate of information generation from source A may not be the same at the rate of information generation by source B. What do you mean by rate of information generation? Well if I toss a coin, a fair coin once every second and I shout 1 0 0 0 1 every second, then clearly the rate of information that I am generating is 1 bit per second.


On the other hand, if I am tossing a biased coin, say source B is a biased coin again 1 toss per second, I am generating 1 0 0 1 0 whatever be the output of source B, but please note that the rate of information is less than 1 bit per second simply because the self information for the biased coin is less than 1. We will delve into it in further detail as we move along, but it was important to understand why a logarithm has been used as a measure for information, why the logarithm measure is important.


(Refer Slide Time: 15:50)

Information Theory, Coding and Cryptography

Mutual Information

- Consider *two* discrete random variables X and Y with possible outcomes $x_i, i = 1, 2, \dots, n$, and $y_j, j = 1, 2, \dots, m$ respectively.
- Suppose we observe some outcome $Y = y_j$ and we want to determine the **amount of information** this event provides about the event $X = x_i, i = 1, 2, \dots, n$
- That is, we want to mathematically represent the **mutual** information.



 Indian Institute of Technology, Delhi 10 Ranjan Bose
Department of Electrical Engineering

Now, since we are setting the stage up for a larger task, let us look at the notion of mutual information. This is to say that we have two discrete random variables X and Y maybe they are related somehow; maybe they are connected through a channel. What we would like to do is having observed Y , how much can we say about X ? Would not it be great to have a measure of information which is conveyed mutually? So, having observed Y , what can I say about the occurrence of X ? Having observed X , what can we say about the occurrence of Y ?

So, this notion is captured by mutual information. Let us look at it further, but please note, it is not just a channel which links X and Y . It could be in the area of DNA

sequencing, it could be relating diseases to certain genetic defects. So, you make certain observations Y could be a profile of the disease and X could be the genetic code and you can derive a mutual information relationship between X and Y . X could be the Bombay stock exchange, Y could be the London stock exchange and there could be a mutual information between these two.

So, having observed the performance of the London stock exchange, how much does you how much do you say or conclude about the performance of X ? Can we predict the outcome? Can we do something more with it? Those are the questions that mutual information will help us answer. But right now, this green arrow that you see could be as well be the channel connecting X and Y .

(Refer Slide Time: 17:55)

Information Theory, Coding and Cryptography

Mutual Information

- We note the following:
- If X and Y are independent, in which case the occurrence of $Y = y_j$ provides **no information** about $X = x_i$.

- If X and Y are fully dependent events, in which case the occurrence of $Y = y_j$ **determines** the occurrence of the event $X = x_i$.

Indian Institute of Technology, Delhi
11
Ranjan Bose
Department of Electrical Engineering

So, let us do some more intuitive thinking on it. Suppose X and Y are really independent that is the connection is really broken. Whatever X does has no relation to whatever, you observe for Y . So, in this case mutual information should be 0. So, when we build up a mathematical framework, we must make sure that if the X and Y events are really independent, then the mutual information conveyed should be 0.

On the other hand, if Y and X are completely dependent on each other right. So, Y actually can determine, what X did or vice versa, then there should be a high level of mutual information being communicated. So, we will keep these two observations in mind. With that we come to the mathematical definition of mutual information.

(Refer Slide Time: 19:00)


Information Theory, Coding and Cryptography

Mutual Information

- **Definition** : The **mutual information** $I(x_i; y_j)$ between x_i and y_j is defined as

$$I(x_i; y_j) = \log \left(\frac{P(x_i | y_j)}{P(x_i)} \right)$$

- As before, the units of $I(x)$ are determined by the base of the logarithm, which is usually selected as 2 or e.
- When the base is 2 the units are in **bits**.

 Indian Institute of Technology,
Delhi12Ranjan Bose
Department of Electrical Engineering

Now please note, we are talking about the mutual information of the occurrence of x_i and y_j . Please note capital X in as we showed in the last time is a random variable. It has many possible outcomes. Y is also random variable with Y_1, Y_2 up to Y_n possible output and y_j happens to be a particular output.

So, we can relate x_i to y_j and how do we do that? The mutual information x_i semicolon y_j ; so please note, how we denote it is x_i semicolon y_j is given by log in the numerator, this conditional probability $P(x_i | y_j)$ by $P(x_i)$. Again log has been used without specifying the base of the log. So, if the base is 2, we will have bits. If the base is e, we will have nats. So, this is the mathematical definition of mutual information.


(Refer Slide Time: 20:11)

Information Theory, Coding and Cryptography

Mutual Information

- **Mutual information**

$$I(x_i; y_j) = \log \left(\frac{P(x_i | y_j)}{P(x_i)} \right)$$
- **Observe that**
- $$\frac{P(x_i | y_j)}{P(x_i)} = \frac{P(x_i | y_j)P(y_j)}{P(x_i)P(y_j)} = \frac{P(x_i, y_j)}{P(x_i)P(y_j)} = \frac{P(y_j | x_i)}{P(y_j)}$$
- **Therefore**
- $$I(x_i; y_j) = \log \left(\frac{P(x_i | y_j)}{P(x_i)} \right) = \log \left(\frac{P(y_j | x_i)}{P(y_j)} \right) = I(y_j; x_i)$$

 Indian Institute of Technology, Delhi
13
Ranjan Bose
Department of Electrical Engineering

Now, let us make an interesting observation. So, we have just established x_i ; y_j , the mutual information between these two is $\log \frac{P(x_i | y_j)}{P(x_i)}$, but if you write $P(x_i | y_j)$ over $P(x_i)$ as $P(x_i | y_j)$ into $P(y_j)$ and same with the denominator, then the numerator can now be represented as $P(x_i, y_j)$ and in the denominator we have $P(x_i)P(y_j)$ which can again be written as $P(y_j | x_i)P(y_j)$. So, what do we see? Well we see that x_i ; y_j is nothing, but $I(y_j ; x_i)$. So, there is a two way relationship. It is symmetric. This will come handy as we move along.

(Refer Slide Time: 21:19)


Information Theory, Coding and Cryptography

Physical Interpretation

- When the random variables X and Y are **statistically independent**, $P(x_i | y_j) = P(x_i)$, which leads to $I(x_i; y_j) = 0$.
- When the occurrence of $Y = y_j$ **uniquely determines** the occurrence of the event $X = x_i$, $P(x_i | y_j) = 1$, the mutual information becomes

$$I(x_i; y_j) = \log \left(\frac{1}{P(x_i)} \right) = -\log P(x_i).$$

This is the self information of the event $X = x_i$.

 Indian Institute of Technology, Delhi
14
Ranjan Bose
Department of Electrical Engineering

So, let us look at the physical interpretation. Clearly we look at two extremes. Suppose X and Y are statistically independent, what does it mean? It means $P(x_i | y_j)$ is just $P(x_i)$, it does not matter what y_j is. They are independent. And if this happens, then if you plug into the formula for mutual information, you get 0. This fits very well with our intuition because the mutual information is indeed 0 for 2 statistically independent random variables.

Now, we look at the other extreme where the occurrence of Y uniquely determines the occurrence of the event x_i . This means given y_j , the probability of x_i is 1. It uniquely determines. In that case, the numerator becomes 1 and the mutual information reduces to $\log_2 \frac{1}{P(x_i)}$ which is nothing, but minus $\log_2 P(x_i)$ which is a self-information of x_i . Whatever uncertainty was there in $P(x_i)$ remains so, because having observed y_j your transformed right up to x_i and now you are dealing with x_i directly fine. So, if you see there is a very strong physical feel for this mathematical definition of mutual information. Now why do not we put this to use in terms of a channel?

(Refer Slide Time: 23:00)

Information Theory, Coding and Cryptography

Mutual Information - Example

CHANNEL

A binary symmetric channel (BSC)

$$P(Y=0) = P(X=0)P(Y=0|X=0) + P(X=1)P(Y=0|X=1)$$

$$= 0.5(1-p) + 0.5(p) = 0.5$$

$$P(Y=1) = P(X=0)P(Y=1|X=0) + P(X=1)P(Y=1|X=1)$$

$$= 0.5(p) + 0.5(1-p) = 0.5$$

$$I(x_0, y_0) = I(0, 0) = \log_2 \left(\frac{P(Y=0, X=0)}{P(Y=0)} \right) = \log_2 \left(\frac{1-p}{0.5} \right) = \log_2 2(1-p)$$

$$I(x_1, y_0) = I(1, 0) = \log_2 \left(\frac{P(Y=0, X=1)}{P(Y=0)} \right) = \log_2 \left(\frac{p}{0.5} \right) = \log_2 2p$$

Indian Institute of Technology, Delhi
15
Ranjan Bose
Department of Electrical Engineering

So, if you look at a binary symmetric channel, it transmits a 0 or a 1. So, most of the time when I send a 0, it goes as a 0 and when I send a 1, it goes as a 1 and I am happy about it, but once in a while it makes an error. So, even though you sent a 0 with probability small p , it appeared as a 1. There is noise in the channel; if it is a wireless channel. There is

fading, there is distortion; there is so, many extraneous factors which can cause an error in the channel. .

So, what we do is now we find out, what is the probability of Y being 0 and Y being 1 given the input probabilities of 0 being 0.5 and 1 being 0.5; So, if you work out the math, then based on that you can calculate the mutual information between x 0 and y 0. What does it mean physically? What is the information being conveyed having observed 0 at the output about a 0 being sent? Let me repeat, we are asking a very fundamental question. I receive as 0 at the far end of the channel.

Now I would like to know, whether a 0 was sent or a 1 was sent? So, the information that is been conveyed having observed y 0 what is the chance that x 0 was sent, the mutual information between them is characterized by $I(0; 0)$ and if you plug into that formula, you get it is $\log_2 2(1 - p)$.

So, we are relieved to see that small p enter the picture. Because if p were indeed 0 that the channel makes no mistake, we look at this case; then we should get 0 always as a 0 and 1 always as a 1. On the other hand suppose y 0 was received that is you receive a 0 at the receiver, but you are curious to know that 1 was indeed sent. What is the mutual information between x 1 and y 0? So, if you plug in you get an expression $\log_2 2p$.

(Refer Slide Time: 25:47)

Information Theory, Coding and Cryptography

Mutual Information - Example

Suppose, $p = 0$, i.e., it is an ideal channel (noiseless).
In that case

$$I(x_0; y_0) = I(0; 0) = \log_2 2(1 - p) = 1 \text{ bit.}$$

Hence having observed **with certainty** the output we can determine what was transmitted.

Recall that the self information about the event $X = x_0$ was 1 bit.

Indian Institute of Technology,
Delhi
16
Ranjan Bose
Department of Electrical Engineering

So, we will look at the interesting applications now. Suppose it is indeed an ideal channel right. So, 0 goes always as a 0, 1 goes always as a 1. It never makes a mistake. In that case, if you compute the mutual information $I(X; Y)$, you get the answer 1 bit. What does it mean? It means simply that 1 bit of information is conveyed right through this channel, every time you use this channel right. So, having observed with certainty the output, we can determine what was indeed transmitted.

So, where does it come to? Well the self-information about the event X is equal to x was 1 bit. So, whatever uncertainty, please remember information deals with uncertainty is the uncertainty of the input. The channel is not introducing any further uncertainty. So, the take home message from this is, whatever uncertainty you observe at the receiver side is primarily the uncertainty at the transmitter side.

There is uncertainty; sometimes 1 comes, sometimes 0 comes. If there was no uncertainty, there would be no need to transmit the information ok. So, this channel does not introduce any uncertainty on its own. It merely communicate without error what was set.

(Refer Slide Time: 27:33)

Information Theory, Coding and Cryptography

Mutual Information - Example

However, if $p = 0.5$, we obtain

$$I(x_0; y_0) = I(0; 0) = \log_2 2(1 - p) = \log_2 2(0.5) = 0.$$

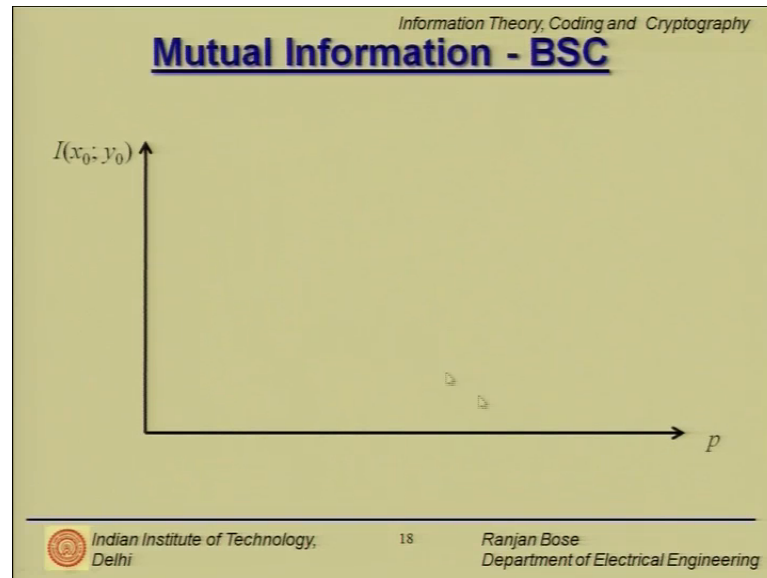
- This implies that having observed the output, we have *no information* about what was transmitted.
- Thus, it is a useless channel.
 - For such a channel, there is no point in observing the received symbol and trying to make a guess as to what was sent.
 - Instead we can as well toss a fair coin at the receiver in order to estimate what was sent!

Indian Institute of Technology, Delhi
17
Ranjan Bose
Department of Electrical Engineering

So, now we look at another case, the worst part. Suppose this channel makes a mistake with probability 0.5, the half the time the bit is flipped; I send a 0, half the time it is received as a 0, half the time it is received as a 1. Same is the story with 1. It has a probability 0.5 being 0 and probability 0.5, it being received as a 1. If you plug in the

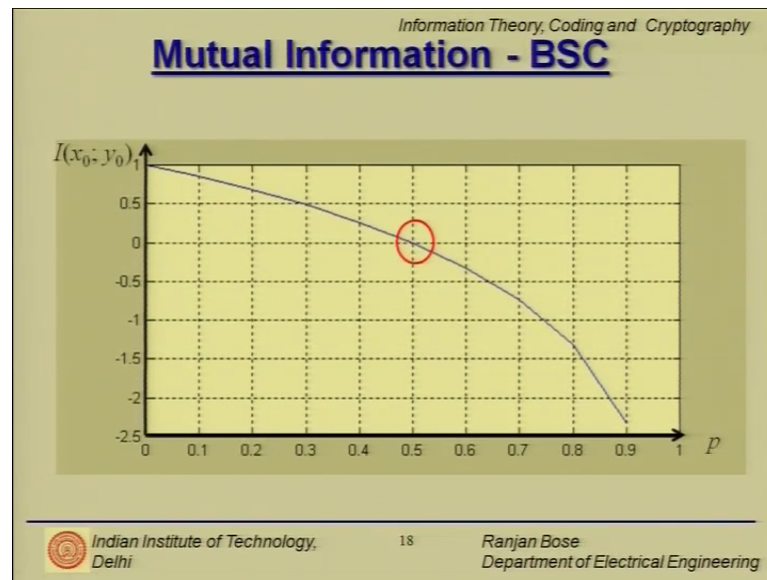
values, now it says that $I(x_0; y_0)$ is 0. This channel basically conveys no information. Having observed output 0 or 1, really you cannot say anything about with a 0 or 1 was sent. It is a useless channel. I mean might as well throw away the channel and toss a fair coin at your other end because a fair coin was being tossed at the transmitter side. There is no need for this channel. The mutual information is 0.

(Refer Slide Time: 28:48)



So, let us now see the variation of $I(x_0; y_0)$ with p . Remember we looked at two cases; p equal to 0, ideal channel and p is equal to 0.5 were in half the time, we were making errors, but what are the other values.

(Refer Slide Time: 29:14)



So, if you look at it, you see an interesting variation. We have plotted from p is equal to 0 up to p is equal to 0.9 and you see a monotonically decreasing value of $I(x_0; y_0)$, but this is part of the story. If you look carefully at 0.5, the value of $I(x_0; y_0)$ that is 0, that is what we found. But if you increase p beyond 0.5, $I(x_0; y_0)$ becomes negative. So, can information be negative? Well here is the first example that mutual information has touched the negative boundary. It has gone below 0 and continues to go do so.

So, we will talk about the physical interpretation of this negative information shortly, but it is important to observe that this mutual information can be negative right. The way to look at negative mutual information is that if you go back to your original binary symmetric channel, if I have a high value of x_0 , the mutual information between these two; that means, having observed 0 you high chance that 0 was indeed sent right. If the mutual information $I(x_0; y_0)$ is 0; that means, you cannot say anything, this channel is not telling you anything.

On the other hand if you have a negative answer; that means that having observed 0, you should more likely to guess 1 was being sent and the channel made a mistake rather than a 0 being sent. So, it is indicating to you that given a choice go up for 1 rather than 0 because there is a negative mutual information between $I(x_0; y_0)$; that is the physical

way to look at this negative in mutual information. So, we already went to this figure and please note the take how message I_{x0} semicolon $y0$ is negative for p greater than 0.5.

(Refer Slide Time: 32:00)

Information Theory, Coding and Cryptography

Binary Channel

CHANNEL

From the channel transition probabilities we have

$$P(Y = 0) = P(X = 0)P(Y = 0 | X = 0) + P(X = 1)P(Y = 0 | X = 1)$$

$$= 0.5(1 - p_0) + 0.5(p_1) = 0.5(1 - p_0 + p_1).$$

$$P(Y = 1) = P(X = 0)P(Y = 1 | X = 0) + P(X = 1)P(Y = 1 | X = 1)$$

$$= 0.5(p_0) + 0.5(1 - p_1) = 0.5(1 - p_1 + p_0).$$

Indian Institute of Technology,
Delhi

19

Ranjan Bose
Department of Electrical Engineering

So, we have been talking about this binary symmetric channel, but now why do not we kind of remove this symmetric part and make it a binary channel ok. So, please note that the p_0 and p_1 's may not be the same. That is a 1 being an error, may not have the same probability of being an error as a 0 right. The probability of 1 flipping is not the same as 0 flipping over the channel. Are the real world examples? Of course, there are right. I just put different power associated with transmission of 1 or a 0 and I will get different probabilities of error and the receiver. So, there are many practical channels where it does not have to be symmetric. So, let us look at this binary channel.

So, if you do the basic math, you can first calculate what is the probability of occurrence of a 0 and probability of occurrence of a 1 at the receiver end fine; in terms of p_0 and p_1 ; Of course, here we have assumed that the input probabilities are equal half and half. Question, does it always happen that the input probabilities are half and half? Answer is no. In many real life applications, the occurrence of 0's and 1's may not be the same.

Over the internet, if you just count the 1's and 0's, chances are they may not be equal or near equal. But here for most of the applications, we can assume that occurrence of 0's and 1 are the same. They are equiprobable and therefore, we have the expressions for y_0 and y_1 right.

(Refer Slide Time: 34:00)

Information Theory, Coding and Cryptography

Binary Channel

CHANNEL

The mutual information about the occurrence of the event $X = 0$ given that $Y = 0$ is

- $I(x_0; y_0) = I(0; 0) = \log_2 \left(\frac{P(Y=0|X=0)}{P(Y=0)} \right) = \log_2 \left(\frac{2(1-p_0)}{1-p_0+p_1} \right)$
- $I(x_1; y_0) = I(1; 0) = \log_2 \left(\frac{P(Y=0|X=1)}{P(Y=0)} \right) = \log_2 \left(\frac{2p_1}{1-p_0+p_1} \right)$

Indian Institute of Technology,
Delhi
20
Ranjan Bose
Department of Electrical Engineering

So now, we look at the binary channel in further detail and we were talking about. So, we are talking about binary channel and we are now talking about the application of this mutual information to this binary channel. As i mentioned before, the probabilities of 0's and 1's flipping over being in error is not the same.

So, we do the same exercise that we did for binary symmetric channel, but this time we do only for binary channel. So, $I(x_0; y_0)$ that is $I(0, 0)$; If you work out the basic math is given by log to the base 2 in the numerator $2(1-p_0)$ over $1-p_0+p_1$. So, my expressions are in terms of p_0 and p_1 where p_0 is a probability of 0 being in error and p_1 is a probability of 1 being in error.

Similarly, if you look at the mutual information between $I(x_1; y_0)$, you get the expression log to the base 2 in the numerator $2p_1$ over $1-p_0+p_1$. How does it help us? Well it helps us analyze this channel, in terms of p_0 and p_1 . After all mutual information will yield 1 more very very interesting thing. It will tell me the goodness of the channel. If you ask an information theoretic person compared to channels, here she will look at the mutual information and therein it will do some extra work and derive the goodness of the channel based on mutual information.

So, physically mutual information is a measure of goodness of a channel. After all what is a channel for? It is supposed to communicate information. How much? How much

information can a channel convey? That is exactly, what mutual information tells us right. So, I would like all of you to understand this basic idea about mutual information.

(Refer Slide Time: 36:45)


Information Theory, Coding and Cryptography

Average Mutual Information

- **Definition** The **average mutual information** between two random variables X and Y is given by

$$I(X; Y) = \sum_{i=1}^n \sum_{j=1}^m P(x_i, y_j) I(x_i; y_j) = \sum_{i=1}^n \sum_{j=1}^m P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$

$I(X; Y) \geq 0$, with equality if and only if X and Y are **statistically independent**.

 *Indian Institute of Technology, Delhi*
21
Ranjan Bose
Department of Electrical Engineering

Now, sometimes we would like to know, what is the average mutual information because in all of the previous exercises, we found out the information between the mutual information between x_0 and y_0 ; that is having observed y_0 , how much information is conveyed about the occurrence of x_0 . I can observe 0 at the receiver and argue how much information is communicated about the occurrence of 1. Similarly having observed 1 at the receiver, what information was communicated about 0 and 1? So, there are four combinations.

Now, sometimes more information is conveyed about 0 versus 0 0 versus 1 and so and so forth. We had seen in the previous case, in the case of binary symmetric channel that the mutual information can even be negative. So, if you go back to your binary symmetric channel, it is possible that because of the choice of p right sometimes positive, sometimes negative information is being communicated; if you observe 0 about 0, if you observe 0 about 1.

So, we did that exercise earlier and now we have moved to binary channel and in this binary channel we observe that depending upon this specific values of p_0 and p_1 , I can make this fraction greater than 0 or less than 0. I can make this right hand side, positive or negative. I can playing with just the numbers p_0 and p_1 , I can make this I_{x_0}

semicolon y 0 positive negative more less. But there are specific cases. Given 0, what can I say about 0? Given 1, what can I say about 1 and so and so forth?

People observed that the channel is conveying not only 0 or not only 1 is communicating both either 0 or 1 sometimes 0, sometimes one. So, it makes sense to talk about the average mutual information. Hey it is possible because it is a binary channel, it is not symmetric that it is it creates 0 very well. Almost always when I send a 0, I get a 0. So, I am very confident about receiving a 0 right, I get excited whenever I receive a 0, but this channel is partial to 1 as you can see from the different probabilities of error.

So, whenever I get a 1, I am worried, I am scared whether 1 was sent, whether 0 was sent and it became a 1; I do not know. So, what do I tell about the channel? Is this a good channel? Is this a bad channel? Well a mathematician would tell you, hey average it right. It would be great for 0, but not so, good for 1. So, maybe if you want a 1 line answer, you average it. So, that is what I will do. I would define the average mutual information between two random variables x and y and if you look at it, it is nothing, but $I(x; y) = \sum_i \sum_j P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$ which is the mutual information multiplied by the probabilities $P(x_i, y_j)$. So, the joint probability x_i, y_j and then double summation which can be expressed as $\sum_i \sum_j P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$ right.

Now, here is the interesting part. Even though mutual information as you saw could be negative and we also saw the physical intuitive feel, for what does a negative mutual information say? The average mutual information can never be negative. It is greater than or equal to 0 right. It is fairly easy to show this, but right now we are only stating the fact that the average mutual information can never be negative. It is a non negative quantity and 0 is achieved if and only if X and Y are statistically independent.

So, if my channel is such that having observed Y , the output nothing is conveyed about the occurrence of X . In that case, the mutual information is 0, but a channel cannot convey negative information and we are very relieved to hear that because it was a hard task explaining on an average, how can a channel convey negative information? Yes it makes sense that if I have a connect between 0 being sent and a 0 being received, I can have a negative mutual information. But the average mutual information which tells you about the overall capability of the channel, how much information it conveys is non-negative ok. This is a very important observation and we have established earlier that $I(X; Y) \geq 0$.

semicolon Y is the same as $Y X$ semicolon Y is the same as $X Y$ semicolon X . So, we can change the order.

So, the amount of information being conveyed from X to Y is the same as the amount of information being communicated from Y to X and again we are happy that this mathematical result exists for a very simple reason that if it were not so, then I would be worried which side of the wire to plug in to my laptop and the other side to plug in on to the internet jack ok. I can just flip over the channel the wire and I am not worried about whether I get more information from one side to other and vice versa.

So, please note what we are stating is that we are able to communicate, the same amount of information from X to Y as from Y to X . Make sure this is different from different bandwidth allocation in the uplink and downlink of certain wireless channels, where you can indeed communicate different amounts of information one way or the other way, but those are two different channels; one for the uplink and one for the downlink. So, we have come to that.

(Refer Slide Time: 44:00)

Information Theory, Coding and Cryptography

Summary

- Uncertainty and Information
- Self Information
- Mutual Information
- Average Mutual Information

Indian Institute of Technology, Delhi 22 Ranjan Bose
Department of Electrical Engineering

So, this brings us to the end of today's lecture and I would just summarize what we have discussed today. We started off with the notion of uncertainty and connected it to information, what we observed was that uncertainty which is linked to probability of occurrence has an inverse relationship. The less likely an event is the less likely it is to

occur the higher amount of information, it contains ok; this is intuitive also, but what is not intuitive is the logarithmic measure.

So, we went over and discussed, why it is important to have a logarithmic measure of information and the basic idea behind it was that if you have independent sources, then the information should add up whereas, the probabilities multiply and log is the only function that does the job. And of course, it is a monotonically increasing function. So, we have the probabilities, if we have a log of that it does not change the inverse relation that we wanted.

So, we talked about self information and then we quickly graduated to the notion of mutual information wherein we brought in two random variables X and Y . Now these could be the input of the channel, the output of the channel or we looked at some other examples where it could be the stock markets in two different countries and we can relate them through mutual information. Lot of investment banking people look at this kind of stuff.

They use information theory, mutual information to predict the behavior of certain stock markets based on the observation of other stock markets. And then we finally, concluded this module with the notion of an average mutual information. We made an observation that even though mutual information in parts can be negative, the average mutual information is never negative. So, this brings us to the end of this module. If there are any questions we can address those questions all right.

Thank you.