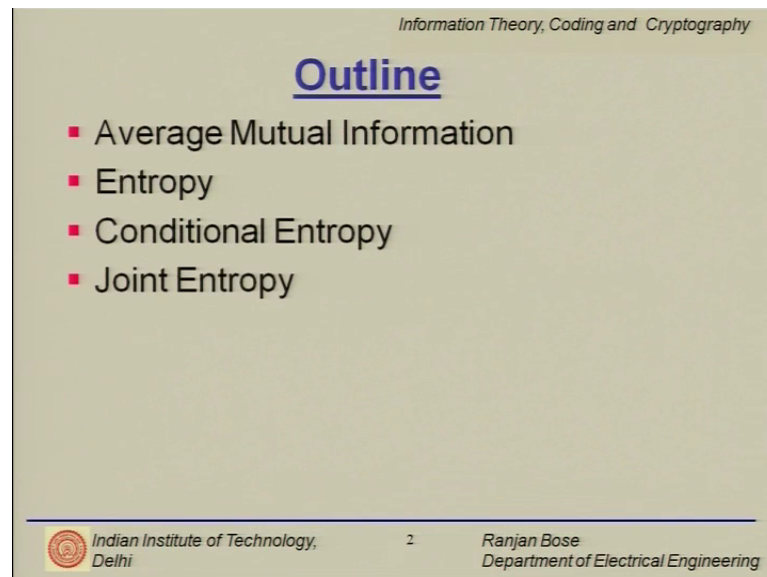


Information Theory, Coding and Cryptography
Dr. Ranjan Bose
Department of Electrical Engineering
Indian Institute of Technology, Delhi

Module - 02
Source Coding
Lecture - 02

Hello and welcome to module 2. Let us start with a brief outline for today's talk. We would visit average mutual information that we covered briefly in the last class.

(Refer Slide Time: 00:28)



Information Theory, Coding and Cryptography

Outline

- Average Mutual Information
- Entropy
- Conditional Entropy
- Joint Entropy

Indian Institute of Technology, Delhi 2 Ranjan Bose
Department of Electrical Engineering


Then we will introduce the concept of entropy, then we will move over to something very interesting conditional entropy followed by joint entropy, and then we will look at some examples.

(Refer Slide Time: 00:45)

Information Theory, Coding and Cryptography

Recap

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

 Indian Institute of Technology,
Delhi3Ranjan Bose
Department of Electrical Engineering

So, let us start with a brief recap what we did in the last class and we will see if there are any questions. We introduced the concept of uncertainty and related it to information, we found an inverse relationship between the probability of occurrence of an event and the information associated with it. Then we went on to define the notion of self information very important it has lots of practical implications. Then another very useful quantity that we looked at was mutual information; we will briefly go over it once again and then we graduated to average mutual information.

(Refer Slide Time: 01:30)


Information Theory, Coding and Cryptography

Self 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,
Delhi4Ranjan Bose
Department of Electrical Engineering

So, let us quickly recap what we learnt about self information. So, we always start with a discrete random variable; if there is no randomness, there is no information. Because we believe that if there is certainty this absolute knowledge about something that it really does not convey any information. Just like tomorrow the sun will rise in the east a good sentence no information for us. So, we would like to communicate whatever is there which is uncertain.

Now, we have defined already the self information as $1/P(x_i)$ log of that gives you x_i . So, it is minus log of $P(x_i)$ it is an inverse relation and in the last class we argued rather strongly why the logarithmic measure is the only useful measure. Please note the base of the log can be 2 and the units are bits if the base is e; the units are nats.

So, we made that observation in the last class, but it is the log which is important.


(Refer Slide Time: 02:41)

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
 5
Ranjan Bose
Department of Electrical Engineering

We also looked at mutual information; now we suddenly have 2 random variables x and y now x can have several possible outcomes x_1, x_2, x_3 up to x_i or even more; y on the other hand is another random variable y_1, y_2, y_3, y_j up to y_n and we can link x_i to y_j simply as x_i semicolon y_j equal to log $P(x_i$ given y_j over $P(x_i)$.

Now please note that here we have a conditional probability; we would like to know what is the probability of x_i given y_j , but at the same time if you do a little bit of mathematical jugglery; you will realize that $P(x_i$ given y_j can simply be written; if you

do these basic steps as P_{ij} given x_i divided by y_j . Therefore, we made this observation that $I(x_i; y_j)$. So, mutual information between x_i and y_j is the same as $I(y_j; x_i)$; so, it works both ways.

In the last class, we had observed that mutual information can indeed be negative. So, even though it is information; there is a notion of a negative quantity being attached to it. How does it help? Well, if we want to say that I received y_j and what is the chances that x_i was indeed say sent or in other words if you observe y_j and you would like to know how much information it conveys about x_i . Well, it is a relative quantity; so, it can be positive 0 or negative.

(Refer Slide Time: 04:47)

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**.

- **Average mutual information cannot be negative !**

 Indian Institute of Technology,
Delhi

6

Ranjan Bose
Department of Electrical Engineering

But when we average over all possible outcomes where we talk about average mutual information things are different. If you remember $I(X; Y)$; so, we have made a change now it is capital X which represents this random variable X . Of course, it can have several outcomes x_1, x_2, x_i, x_m .

And Y is again capital Y ; it is not y_j and it can be y_1, y_2, y_j, y_n and we are talking in general about x and y it is not a particular x_i related to a particular y_j . So, clearly we weighted with this joint probability; so, P_{x_i, y_j} weighting; so, multiplied. So, what do we multiply it; with $I(x_i; y_j)$ which is the mutual information and we do a double summation over all possible cases and we land up with this expression.

Now, what is interesting is that this average mutual information $I(X; Y)$ is non-negative; it is greater than or equal to 0. And this equality is achieved only if and only if the X and Y random variables are statistically independent. That is to say that X does not communicate any information about Y and vice versa; if $I(X; Y)$ is 0, it also implies that $I(Y; X)$ is also 0.

So, it is to say that suppose somebody gives me a channel and the channel on one side has an X and on the other side has a Y ; then if it has 0 mutual information, average mutual information then it does not communicate any information from the other side as well. So, this is an important observation average mutual information cannot be negative. So, channel can at most give you some information about what was sent, but it cannot be negative; this is the take home message from average mutual information.

(Refer Slide Time: 07:23)

Information Theory, Coding and Cryptography


Average Self Information

- Consider a discrete random variable X with possible outcomes $x_i, i = 1, 2, \dots, n$.
- The **average self information** of the event $X = x_i$ is defined as

$$H(X) = \sum_{i=1}^n P(x_i) I(x_i) = - \sum_{i=1}^n P(x_i) \log P(x_i)$$

- When the base of the logarithm is 2 the units of $I(x)$ are in **bits**
- The entropy of X can be interpreted as the expected value of

$$\log \left(\frac{1}{P(X)} \right)$$
- $H(X)$ is called the **Entropy**

 Indian Institute of Technology,
Delhi

7

Ranjan Bose
Department of Electrical Engineering

Now, we come to this notion of average self information please record we have already talked about self information, but now we move on to average self information. It has a lot of practical applications consider a discrete random variable X with possible outcomes $x_i, i = 1, 2, 3, \dots, n$ and the average self information of this event X is equal to x_i can be defined as H of X .

So, now, we take H of X as the definition for average self information. We take $I(x_i)$ and again weight it with the probability of x_i and add it; over standard way to take the

averaging and we come up with this quantity. This log measure as before is if the base is 2 we say the units are bits.

Now, what do you interpret X as? Where first important information is that H of X which is the average self information is also called entropy; we will decide why is this name good or bad, what is the physical correlation to that, but assuming for the time being that H of X can be termed also as entropy; it is nothing but the expected value of \log of 1 over $P(x_i)$ of $P(X)$. So, $\log 1$ over $P(X)$ is basically the self information and since we are talking about entropy being the average self information. It is nothing the average of that. H of X as I mentioned is called the entropy because it has bearings from statistical mechanics.

(Refer Slide Time: 09:15)

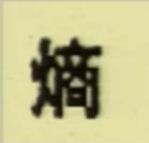
Information Theory, Coding and Cryptography


More on Entropy

- The term entropy has been borrowed from statistical mechanics, where it is used to denote the level of disorder in a system.
- We observe that since $0 \leq P(x_i) \leq 1$

$$\log\left(\frac{1}{P(x_i)}\right) \geq 0$$

- $H(X) \geq 0$
- It is interesting to see that the Chinese character for entropy looks like



 Indian Institute of Technology, Delhi8Ranjan Bose
Department of Electrical Engineering

So, just let us spend a couple of minutes on this quantity called entropy and we will lead it to information. So, the term entropy has been borrowed from statistical mechanics right; in statistical mechanics entropy is used to show or depict the level of disorder in a system ok. And people normally say the entropy of this universe is increasing, stars are moving far away from each other, the orderliness moves away. And it is a fact of life if I leave my room unattended in a few hours it becomes completely disorganized with papers and everything, but coming back to our subject at hand.

In information theory, if there is lot of uncertainty, there is lot of randomness then we say that there is a higher level of disorder and hence a higher value of entropy ok. Please

note: that probabilities lie between 0 and 1; so, every one of the quantities $\log \frac{1}{P_i}$ is positive and their weighted sum is also positive and hence H of X is necessarily greater than or equal to 0 ok. Now, just for the sake of information the Chinese character for entropy looks pretty complicated. If you remember they have a picture for everything and this is the picture they could think of for disorder just a side comment.

(Refer Slide Time: 11:00)

Information Theory, Coding and Cryptography

Example

- Consider a **discrete binary source** that emits a sequence of statistically independent symbols.
- The output is either a 0 with probability p and a 1 with a probability $1 - p$.
- The entropy of this binary source is

$$H(X) = -p \log_2(p) - (1 - p) \log_2(1 - p).$$

Indian Institute of Technology, Delhi

9

Ranjan Bose
Department of Electrical Engineering

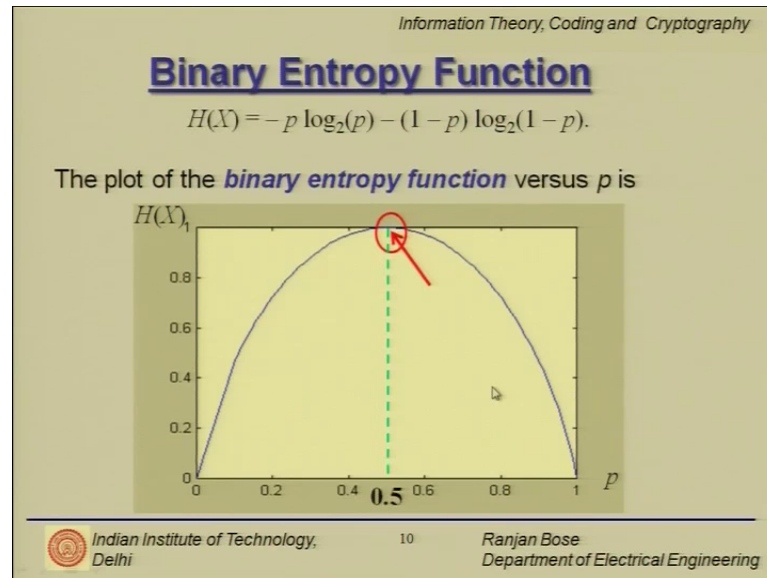
Now, let us understand the concept of entropy from a simple example. So, let us consider a discrete binary source ok; now this emits a sequence of statistically independent symbols. What could this source be? I can imagine it to be a person sitting on a chair, tossing a coin and every time the coin gives a head or a tail the person shouts 0, 1, 1, 0, 0 depending upon whatever is the outcome of the coin tossed.

Now I am not saying the coin is a fair coin, it could be a biased coin with a probability p for head and $1 - p$ for tail. So, this is a source this is a discrete binary source that gives you bits. Now, if I ask this gentleman sitting on this chair tossing this coin to toss this coin once every second. So, he shouts 1, 0, 0, 1, 0 every second.

So, the rate at which this source is generating bits mind you not information is just 1 bit per second right, but this is this entropy which will tell me what is actually the source rate. There is a distinction between the bit rate and the source rate; we need to make that difference. But let us start with finding the entropy of this binary source; how much

uncertainty is there? So, we plug in the value H of X is minus p log to the base 2 minus $(1 - p)$ log to the base 2 minus p .

(Refer Slide Time: 12:54)



So, this quantity we just read out is called the binary entropy function and we will find his application at several places. So, for instance if you plot on the x axis the values of p and on the y axis the corresponding H of X ; we see an interesting bell curve ok; which reaches a maximum at 0.5. So, if you focus you will see that H of X is achieved for p equal to 0.5; remember the coin tossing experiment where p represents the probability of occurrence of head or tail if you want.

So, let us spend a minute figuring out this curve first thing is up to 0.5, the H of X increases. It rises pretty sharp in the beginning and then it starts flattening out there is a physical interpretation to that. So, when p is equal to 0; p is equal to 0 is the probability of head is 0; it is always a tail it is the Sholay coin right; it is tail on both sides.

Well, there is no uncertainty regardless of whether you toss it 1 time or 10 times, you are always going to get a tail and consequently there is no information contained in it. This gentleman who is my binary source is sitting on the chair tossing this tail only coin and he is screaming 0 0 0 0 0; he may shout as much as he want, but he is not communicating any information. There is absolutely no information in that sequence of bits it's all 0's it is a big chain of 0's.

Now, we change it a little bit more and we now have some probability of occurrence of head p is equal to 0.1 or 0.2. And suddenly the uncertainty jumps up the uncertainty jumps up and yes H of X is nonzero. So, it starts from 0.2 bits, goes up to point 4 bits and by the time I am close to probability of 0.2; I am nearing above 0.6 of the entropy which at most will be 1.

So, it is a fast rise and then it gradually flattens and reaches a maximum at 0.5; it tells me that if the guy is tossing a perfectly fair coin; half the time head comes up, half the time tail comes up there is a maximum uncertainty. There is no way I can guess what was going to come up next he just has to toss the coin. And indeed only in this case the entropy H of X is 1 bit and only in this case do we need 1 bit to represent that information.


Now flip the probabilities on the other side make head more probable this curve stops increasing, takes a downturn and starts going down right up to P equal to 1 when we have the H of X again going back to 0 right. So, this is our famous binary entropy function.

(Refer Slide Time: 16:36)

Information Theory, Coding and Cryptography

Entropy of English

- Consider the English language with alphabet $\{A, B, \dots, Z\}$.
- If every letter occurred with the same probability and was independent from the other letters, then the entropy per letter would be $\log_2 26 = 4.70$.
- This is the absolute upperbound.
- However, we know that all letters do not appear with equal probability.
- S, T, A, E are more frequent
- Q, J, Z, J are less frequent

 Indian Institute of Technology,
Delhi11Ranjan Bose
Department of Electrical Engineering

Now, let us see; what are the practical uses of this definition we just put together. So, I would like to know what is the entropy of English or entropy any language for that matter. The question is it a fair question to ask; entropy of a language well for me entropy is uncertainty is there an uncertainty in English? Of course, there is that is why

you buy a book and read it because you do not know what is there ok. If you know already there is no uncertainty you will never buy a book to read it because you know already.

So, consider this English language with alphabet with 26 alphabet A to Z; now as a first step let us assume that all of these alphabets are equiprobable ok; it is a bad assumption, but let us start somewhere. So, each one is equally likely probability 1 over 26 and I plug in to this value of H of X as summation of P x i log 1 over P x i over all i's and you will get this quantity log to the base 2 26 which is 4.70 bits basis 2; so, the units will be bits.

So, it tells me that if all these alphabets were equiprobable; I would need on an average 4.7 bits to represent them because that is the information they command; that is the resource I need to put in to represent each alphabet ok. So, please note there is a strong physical meaning to this information in terms of bits, but this is really the upper bound because we really know that all the alphabets are not equiprobable; A E S T are much much more frequent, if you just scan the dictionary as opposed to Q J Z etcetera.


So, the next logical step would be to plug in real probabilities; now how do you get real probabilities? You pick up your dictionary, do a frequency count and divide by the total number of letters you counted, you get an estimate of the number of times it occurs and hence the relative probability of occurrence.

(Refer Slide Time: 19:06)

Information Theory, Coding and Cryptography

Entropy of English

- Consider the English language with alphabet {A, B, ..., Z}.
- If we take into consideration the probabilities of occurrences of different alphabets (normalized letter frequency), the entropy per letter, H_L , would be
$$H_L \leq H(X) \approx 4.14 \text{ bits.}$$
- If X^2 denotes the random variable of bigrams in the English language, the upperbound on H_L can be refined as
$$H_L \leq \frac{H(X^2)}{2} \approx 3.56 \text{ bits}$$
- Here we consider the probabilities of all pairs.

 Indian Institute of Technology, Delhi 12 Ranjan Bose
Department of Electrical Engineering

So, now let us take the same English language with alphabet A up to Z and this time in our definition of H of X. We plug in the probabilities of occurrences of the different alphabets which I just now mentioned is nothing but the normalized letter frequency.


So, clearly the probability of occurrence of e will be more than that of Z; if you plug that in you will get a number close to 4.14 bits.

(Refer Slide Time: 19:40)

Information Theory, Coding and Cryptography

Entropy of English

- Consider the English language with alphabet {A, B, ..., Z}.
- If every letter occurred with the same probability and was independent from the other letters, then the entropy per letter would be $\log_2 26 = 4.70$.
- This is the absolute upperbound.
- However, we know that all letters do not appear with equal probability.
- S, T, A, E are more frequent
- Q, J, Z, J are less frequent

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

So, clearly it is less than 4.70 bits.

(Refer Slide Time: 19:44)


Information Theory, Coding and Cryptography

Entropy of English

- Consider the English language with alphabet {A, B, ..., Z}.
- If we take into consideration the probabilities of occurrences of different alphabets (normalized letter frequency), the entropy per letter, H_L , would be

$$H_L \leq H(X) \quad 4.14 \text{ bits.}$$
- If X^2 denotes the random variable of bigrams in the English language, the upperbound on H_L can be refined as

$$H_L \leq \frac{H(X^2)}{2} \approx 3.56 \text{ bits}$$
- Here we consider the probabilities of all pairs.

 Indian Institute of Technology, Delhi
12
Ranjan Bose
Department of Electrical Engineering

So, this brings us to a very interesting observation; just our observation of the fact that some letters are more frequent than others. We are making a bold statement; on an average we now need 4.14 bits to represent each letter.

Now, you will hear me talk about this on an average because this entropy by definition is the average self information whatever gain we make is over the averaging part. So, right now we have seen that on an average an alphabet can be represented with just 4.14 bits. Now, this is way below the ASCII representation which uses 7 or 8 bits per letter, but let us see; what more can information theory deliver.

Now, we make some more observation Q is always followed by U; T H E comes more frequently T H comes more frequently as a pair right ING comes together; so, if we start talking about pairs and also look at the probability of occurrence of pairs which we now called bigrams in English language. Then we again calculate this entropy and what we get is $H(X)$ the square represents for the bigram divided by 2, because now we have double the number of pairs. And the entropy now is calculated to 3.56 bits per letter the division by 2 represents per letter because bigram transits of 2 letters at a time.


So, it is coming down it is also getting us more excited to see how it goes it tells me that if I further look at the distribution of frequencies 2 at a time; I need fewer bits. This brings to me a very important fact that if I really have to efficiently represent the English language, then I should not only consider one letter at a time, maybe 2 letters at a time and then why should I stop at 2 letters at a time maybe I can do better.

(Refer Slide Time: 22:16)

Information Theory, Coding and Cryptography

Entropy of English

- The logic can be extended to **n-grams**. Thus the entropy of the language can be defined as
$$H_L = \lim_{n \rightarrow \infty} \frac{H(X^n)}{n}$$
- Even though the exact value of H_L may be difficult to determine, statistical investigations show that for the English language $1 \leq H_L \leq 1.5$ bits.
- So each letter in the English text gives at most 1.5 bits of information.
- Let us assume the value of H_L 1.25 bits. Thus the redundancy of the English language is
$$R_{Eng} = 1 - \frac{H_L}{\log_2 26} = 1 - \frac{1.25}{4.70} \approx 0.75$$

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

So, how about going to n grams? So, why do not we define the entropy of a language; it could be English as limit n tending to infinity $H(X^n)$ standing for n gram and since I wanted per letter it is divided by n. Now clearly it is very difficult to find this quantity and some statistical investigations over English language have led us to believe that this quantity lies somewhere between 1 and 1.5 bits; this is very interesting. So, each letter in English text really conveys only 1.5 bits of information right. If a word is 6 letter long its only 6 into 1.5 bits is the amount of information it conveys bits is a unit of information. So, since it lies between 1 and 1.5 for the sake of discussion, let us assume that this H_L is 1.25 bits.

Then really what is the redundancy of English language? Redundancy well we will spend a minute on that, but let us understand assume that all letters were equiprobable, they were all independent right and they have 26 letters. So, we found out that the upper bound was log to the base 2 of 26 which was 4.7, but after this statistical investigation; we found it to be 1.25. So, the redundancy is 1 minus 1.25 divided by 4.7 is approximately 0.75 yes there is a question.

Student: (Refer Time: 24:26).

Yes.

Student: N grams sir then how can we find that call let us the entropy is 1.5 or it is an nth associate.

Right the question being asked is about this n grams. So, it is an n tuple we are taking n letters at a time right.

Now, you look at all the probabilities suppose for the sake of discussion n is 4. So, you start with a a a a a a b a a a c going up to z z z y z z z z and 4 letters at a time right. Now each one will have a probability associated with it; so, you have a long table of probabilities you plug in that value into summation of $P \times i \log 1 \text{ over } P \times I$, you compute that value will get a value of the entropy, but that is for the n tuple right, but that is the amount of bits conveyed by n letters at a time.

So, I need to divide it by n to get one entropy for one letter because I have to compare apples with apples right. So, we have covered all possible cases when we consider this n tuple and mind you n tends to infinity here and I have really gone down to the actual entropy of the language fine ok.

So, the observation is a little a nerving it says that the redundancy of English language is close to 75 percent. That is if your book that you have bought is hundred pages fact, 75 page pages of those book is redundant. Only the 25 percent is worth it if we really look at look at it from the information theory point of view. I am talking to you in English; so, out of every 4 sentences; 3 of my sentences are redundant there is a question.

Student: Sir, (Refer Time: 26:40) H of X raised to power n is the entropy of like n letters consecutive n letters h let is the average information associated with those n letters and then it is already a average, then you divided it by n like a n letters. So, for per letter, but in combination like abcd like 4 letter combination abcd like it is like a inside there will be there will be having individual probability whether like if I am searching for a combination abcd; it might not be possible that I will find a combination abcd. So, sir it will vary.

No. So, let me repeat the question the question is that we looking at n grams and they could be various combination. In fact, there will be all possible combinations for n the first position.

Student: Sir.

Can we have.

Student: (Refer Time: 27:34) only (Refer Time: 27:36) dictionary words.

No we are not considering only the dictionary words we are considering all possible worlds which can be constructed with 26 letters.

Student: (Refer Time: 27:47) in case of 4 4 (Refer Time: 27:49) equals to 4 we evaluate by 4 like for abcd like I calculate to for all divide by 4, but like it is abcd, but if I take acbd then also answer comes out to be same because like b is coming before or after, but I am dividing by 4 in the end.

No, but the probability of abcd and probability of acbd may not be the same for example, probability of e g g is much higher than gee or geg because g e g will never occur, but e g g will occur. And e is more frequent than z; so, all those things will be taken into considerations. Therefore, we talk about a statistical investigation what occurs in a big fat dictionary, but you look at all possible words like q qq qq you it will tend to 0 right never occur.

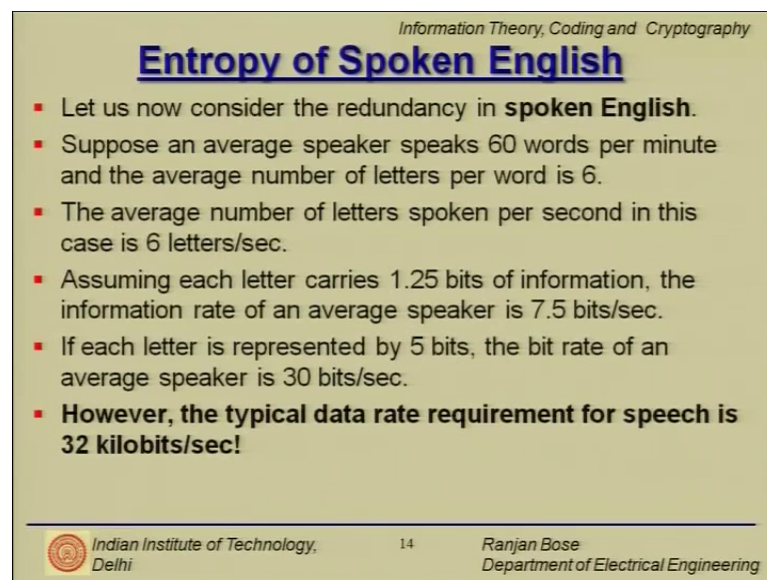
So, if you do that analysis because we are talking about English language as we use it will show that you really need less than 1.5 bits per letter; this is remarkable. This first thing this says is that I can save humongous amounts of bandwidth, bandwidth is expensive. Half of the time I text my messages maybe I am not using the bandwidth optimally; it tells me that when I save my data on cd maybe I should compress and save it because there is so, much of redundancy.

The more fundamental question is why does the language have redundancy; the answer to that is that languages all over the world have evolved and this has its own built in error correction technique. That even if I say hello and I miss out one of the else and you read h e l o; we can almost guess that it is a hello, but the redundancy is built in it helps us communicate more reliably. What is very very uncanny is languages developed across the world in different parts at different times. But, if you calculate the entropy of a particular language say Hindi or German or English their entropy is very very close to the same number of between 1 and 1.5 except Chinese which is a pictorial language and

so they have very smart ways. So, one tree is a tree 2 trees is a forest right; so it is a picture based things. So, they do not have that same concept; so, they can compress it much much more.

So, if you look at any normal language across the world the entropy is very very close to between lying between 1 and 1.5.

(Refer Slide Time: 21:23)



Information Theory, Coding and Cryptography

Entropy of Spoken English

- Let us now consider the redundancy in **spoken English**.
- Suppose an average speaker speaks 60 words per minute and the average number of letters per word is 6.
- The average number of letters spoken per second in this case is 6 letters/sec.
- Assuming each letter carries 1.25 bits of information, the information rate of an average speaker is 7.5 bits/sec.
- If each letter is represented by 5 bits, the bit rate of an average speaker is 30 bits/sec.
- **However, the typical data rate requirement for speech is 32 kilobits/sec!**

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

But, let us do look at a very interesting thing. If you go and buy a modem which should be able to give you real time speech communication. Or if you want bandwidth which will permit you to send speech over say wireless; you always ask for a certain data rate and it is typically 32 kilobits per second or 64 kilobits per second for toll quality speech. Why do not we use this tool that we just learnt the entropy and find out the entropy of spoken English, alright.

So, consider me as a speaker an average speaker and if you listen carefully; I speak roughly 60 words per minute. Typically a speaker in English speaks 60 words per minute, and if you assume that every word has about 6 letters; then you can calculate that the average number of letters spoken per second is roughly 6 letters per second, but now we just calculated that each letter is roughly 1.25 bits of information do you agree?

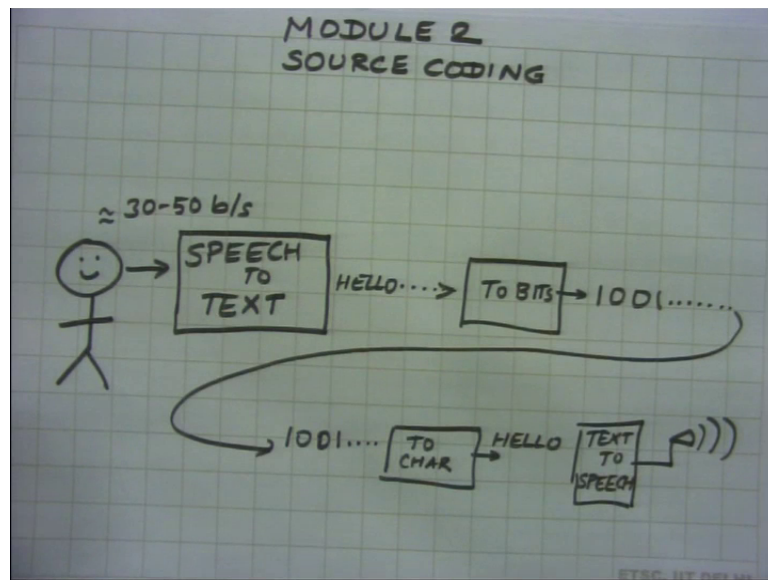
Then my information rate as a speaker of English language who is speaking about 60 words per minute; I am actually speaking information worth only 7.5 bits per second

right. So, if you look at it those that is for the letters and if you talk about the bitrate of an average speaker right; it comes out to be hardly 30 bits per second.

It is just multiplying the numbers right each letter is represented by 5 bits right even if the letter is represented by 5 bits or 7 bits we will just take it to a little bit higher. But I am talking about in the range of 30 bits per second, maybe 45 bits per second max 50 bits per second. Regardless of how fast I speak how big words I speak, I am only communicating to you at 30 bits per second.

But if you look at the typical data rate requirement for speech it is 32 kilobits per second. So, where is the catch? So, let me explain with a diagram.

(Refer Slide Time: 34:19)



So, if you see there is a speaker and he has this wonderful device called speech to text. So, whatever he is speaking is going in terms of something like hello how are you, and so and so forth. But this is converted into bits. So I convert to bits and I get a sequence of 1001, but this speaker is only speaking at max 30 to 50 bits per second; these bits go over a channel and are received.

And then I convert them to characters and then I get this hello out and then I have another device which converts text to speech right. And here I have my invention which let us me go from 30 bits per second to give you a good quality audio; where is the catch? I do not need 32 kilobits per second, I just need 30 bits per second there is 3

orders of magnitude difference. The catch is that I will not be able to hear the same guy speak and the quality of speech will not be the same; here I have a prerecorded way of converting text to speech.

So, it is a mechanical output here, but indeed those 32 kilobits per second that we reserved to communicate speech data is only to ensure that the listener and the other end gets to hear my sound including all the quality of my sound and all the (Refer Time: 36:44) interest is attached with it that is the price we pay in terms of transmission of speech, but entropy of spoken English is really very very low.

(Refer Slide Time: 37:00)

Information Theory, Coding and Cryptography


Conditional Entropy

- The **Average Conditional Self information** called the **Conditional Entropy** is defined as

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

- The physical interpretation of this definition is as follows. $H(X|Y)$ is the information (or uncertainty) in X after Y is observed.
- Based on the definitions of $H(X|Y)$ and $H(X, Y)$ we can write

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

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

So now, we move to the next concept called conditional entropy; the average conditional self information; so, also called the conditional entropy right. So, it this is not a big leap; average self information is entropy, average conditional self information is called conditional entropy and it is simply defined as H of X given Y as double summation joint probability $P(x_i, y_j) \log$ of 1 over $P(x_i | y_j)$. So, this is the place where you have the conditionality.

So, what is the physical interpretation of this definition? The physical interpretation is as follows H of X given Y the conditional entropy is the information for me it is also the uncertainty in X after observing Y . Let me repeat X is a random variable, Y is also a random variable maybe they are connected ok.

So, if you observe Y you can probably say something more about X ; how much more ?
 So, H of X given y is the uncertainty in X remaining after you observe Y . So, we already know this definition of H of X given Y right and what we can write is that. We already have a definition of H of X . So, you can write average mutual information which is $I(X; Y)$ is nothing but $H(X) - H(X|Y)$. But since $I(X; Y)$ the average mutual information is also equal to $H(Y) - H(Y|X)$; you can always write is equal to $H(Y) - H(Y|X)$.


Later, in this lecture we will use a Venn diagram to explain this is what is important to note in this equation is this average mutual information. What is the information X communicates about Y and what Y communicates about X is linked to the self information and the conditional self information this is a very useful formula. And we will be using it time and again.

(Refer Slide Time: 40:00)

Average Mutual Information and Conditional Entropy Information Theory, Coding and Cryptography

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

- Since $I(X; Y) \geq 0$, it implies that $H(X) \geq H(X|Y)$.
- The case $I(X; Y) = 0$ implies that $H(X) = H(X|Y)$, which is possible if and only if X and Y are statistically independent.
- Since $H(X|Y)$ is the average amount of uncertainty (information) in X after we observe Y and $H(X)$ is the average amount of uncertainty (self information) of X , $I(X; Y)$ is the average amount of uncertainty (mutual information) about X having observed Y .
- Since $H(X) \geq H(X|Y)$, the observation of Y does not increase the entropy (uncertainty). It can only decrease the entropy. That is, observing Y cannot reduce the information about X , it can only add to the information.

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

So, let us look at the average mutual information and link it to conditional entropy. We have already established that average mutual information $I(X; Y)$ is nothing but $H(X) - H(X|Y)$, which is equal to $H(Y) - H(Y|X)$.

We already know that average mutual information is non negative; it clearly says that $H(X)$ has to be greater than or equal to $H(X|Y)$. And I am very happy to note that for a very simple reason it is very intuitive that $H(X|y)$ the uncertainty in X having observed Y is at most lower or equal to $H(X)$.

If I observe Y it cannot increase the uncertainty either to tell me something about X or it will not, but there will be no way that observing Y; another random variable Y will increase the uncertainty in X, there is a very physical interpretation available to this equation.

So, what happens when $H(X; Y)$ is 0; it means that $H(X)$ interpreted as uncertainty of X becomes equal to $H(X|Y)$ that is uncertainty of X given Y; whether you observe Y, you do not observe Y $H(X|Y)$ is the same as $H(X)$ and clearly this is possible if X and Y are statistically independent.

Since, $H(X|Y)$ is the average amount of uncertainty in X after we observe Y. So, this is how to physically interpret it and $H(X)$ is the average amount of uncertainty of X anyway what does it mean for $H(X; Y)$? Well, $H(X; Y)$ is nothing but the uncertainty in X remaining after you remove the uncertainty in X having observed Y right.

So, we put it on record that since $H(X)$ is greater than $H(X|Y)$ the observation of Y in no case increases the entropy of X. We can at best decrease the entropy observing Y cannot reduce the information about a X it can only add to the information. So, this is the crux of this slide.

(Refer Slide Time: 43:02)

Information Theory, Coding and Cryptography

Joint Entropy


- The **Joint Entropy** of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log p(x_i, y_j)$$
- By using the mathematical definitions of $H(X)$, $H(X, Y)$ and $H(X|Y)$ we obtain the following **chain rule**

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

and consequently,

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

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

We now go to the notion of joint entropy; the joint entropy of a pair of discrete random variables X and Y which have a joint distribution $P_{X,Y}$ is defined as follows $H(X,Y)$ is double summation this joint probability the joint distribution \log of $p_{X,Y}$.

Now, we can use mathematical definitions of $H(X)$, $H(X,Y)$ and $H(X|Y)$ to obtain the following chain rule. What is the chain rule? $H(X,Y)$ joint entropy what is the joint the uncertainty present both in X and Y the uncertainty of both X and Y taken together is nothing but the uncertainty of X right plus uncertainty of Y given X . So, it is intuitive if you follow it or you can also write it as uncertainty of Y plus $H(X|Y)$; that is having observed Y what is uncertainty of X ?

In reality suppose I give you a case where X and Y are statistically independent in that case $H(X|Y)$ is nothing but $H(X)$. So, the joint entropy the total uncertainty in X and Y taken together is nothing but uncertainty of X and uncertainty of Y added together and same is the case with this equation.

Now, we put together all that we know and we write this average mutual information $I(X;Y)$ as $H(X) + H(Y) - H(X,Y)$. This is a very very useful relation which you will use throughout this module. So, this slide gives you not only the chain rule, but also links the average mutual information to the uncertainties of X , Y and the joint entropy of X and Y . Whenever you want a physical interpretation I should be interpreted as the uncertainty of. So, $H(X)$ is uncertainty of x , $H(Y)$ is uncertainty of Y , $H(X,Y)$ is uncertainty of X and Y taken together.

(Refer Slide Time: 46:14)

Information Theory, Coding and Cryptography


Self Entropy revisited

- How much does X convey about itself?

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

$$I(X; X) = H(X) - H(X|X) = H(X).$$

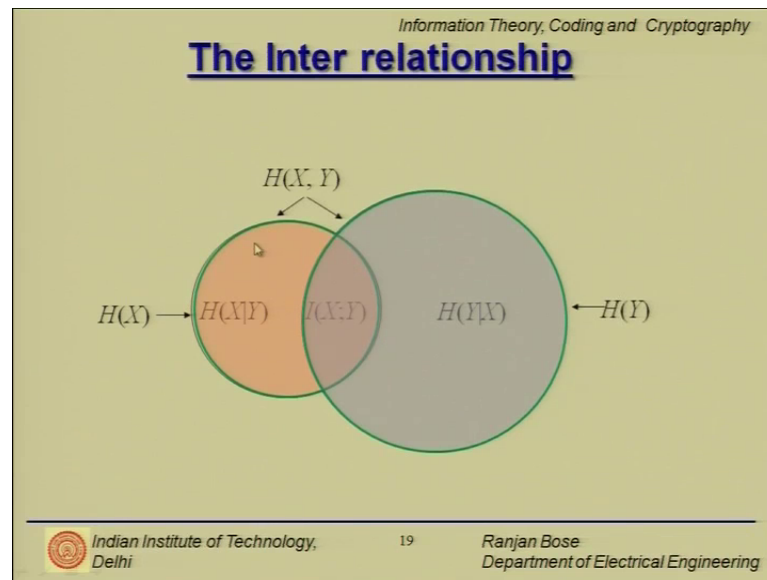
- The average self information is also called the entropy

 Indian Institute of Technology,
Delhi18Ranjan Bose
Department of Electrical Engineering

So, let us see whether everything fits together can we close the loop? Can we close the chain? Question that I am asking is how much does X convey about itself X is a random variable and X is a random variable. So, if I ask this basic question having observed X how much information does it convey about itself? So, I am tempted to put I of X semicolon Y as I of X semicolon X . What is the mutual information of X and X if my math is right I should get an answer?

So, if you put it there this is nothing but H of X minus H of X given Y , but Y is nothing but here X . So, what is the uncertainty of X given X 0? So, this quantity goes and you are left with H of X and therefore, it is called the self information this is the entropy that we have already figured out. So, it all fits in together it makes perfect sense all the definitions that we have given so far are consistent with each other.

(Refer Slide Time: 47:40)



So, we now look at a very interesting interrelationship between H of X , H of Y , and H of Y given X . So, consider this Venn diagram here the pink represents the uncertainty in X , size can be taken as the amount of uncertainty. So, X could be a coin toss experiment with a biased coin. So, there is uncertainty, but there is some amount of uncertainty.

Now, Y is another big circle represented by light blue which represents the uncertainty in Y . So, what could Y be? Y could be another coin toss experiment, but this type is a fair coin. So, this circle is much bigger, but they overlap. So, probably my friend 1 who is tossing an unfair coin and giving me the red circle and my friend 2 who is tossing a fair coin and giving me the blue circle are talking to each other; there is some dependency ok; so, there is an overlap.

Now, the joint entropy H of X comma Y is the outside perimeter of this structure ok. This together represents the H of X comma Y and if you look at just the pure pink area is H of X given Y . So, if I remove it this white portion this part of the circle which was H of X is the uncertainty of X given Y because that much has been removed, because Y is right here.

On the other hand, if I want for Y this part of this circle is H of Y given X and the intersection is the average mutual information. So, if the average mutual information is high; I will start making the overlap larger and this intersection will increase. So, much

so, that I can make this H of X completely end up being inside H of Y ok. So, the average mutual information can at best be H of X.

(Refer Slide Time: 50:33)

Information Theory, Coding and Cryptography

Example

- The entropy of this binary source is

$$H(X) = -q \log_2(q) - (1-q) \log_2(1-q).$$
- The conditional entropy is given by

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

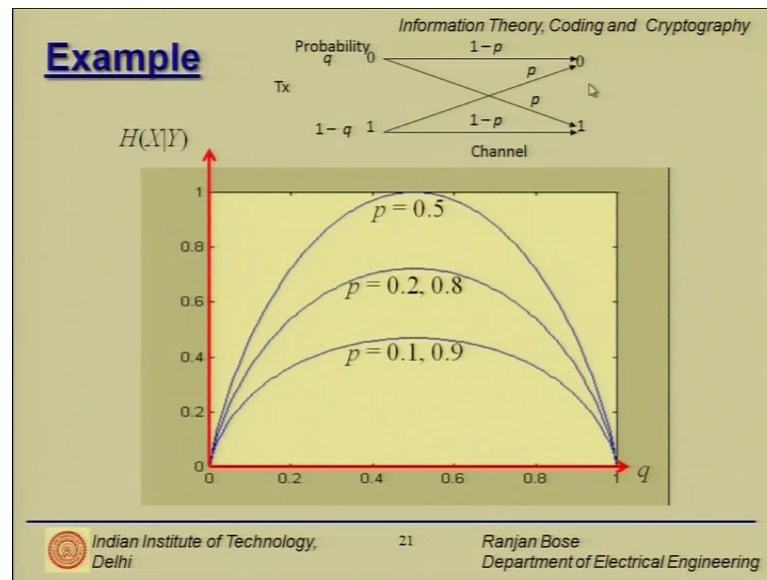
Indian Institute of Technology, Delhi
20
Ranjan Bose
Department of Electrical Engineering

Now, let us quickly take a look at an example this is our friend the binary symmetric channel, but unlike the last time this time the input probabilities are not half and half, we argued in the last class that it is not always necessary that the input probabilities should be 0.5 and 0.5. If I tap my internet line and I measure the 1s and 0's chances are that the number of 1s will not equal to number of 0's over a long enough period of time. So, for instance let us say probability of 0 is q and probability of one is 1 minus q.

Now, my binary symmetric channel has the probability of flipping of bits equal to P. So, 1 goes as 1 most of the time, but once in a while it becomes a 0 with probability small p. Likewise 0 sometimes flips and becomes 1 with probability small p; so, if you look at the entropy of the binary source at the input it is nothing but minus q log to the base q minus 1 minus q log to the base 1 minus q I am lonely looking at the input probabilities.

So, there is a certain amount of uncertainty in the input itself and if you look at the formula for conditional entropy H X given Y it is given by this double summation.

(Refer Slide Time: 52:06)



So, I would like to plot by H of X given Y for different values of q. So, let me pose a problem to you we have seen that this binary symmetric channel makes error once in a while, but I am more excited about finding out how does H of X given Y; the conditional entropy of X given the observation of Y at the further end of the channel change as I change my input probabilities. So, x axis is Q I have got 3 figures for each one for different values of p. So, first thing to note is as we have seen in this previous formulation H of X given Y that it depends on the probabilities of P_{x_i, y_j} together and they depend in turn on p and q.

So, H of X given y the uncertainty of x given the observation of Y increases as Q increases go through a maximum and then goes down again, but if I make my channel worse. So, p is equal to 0.2 the overall uncertainty of X increases remember physical interpretation H of X given Y; I am trying to figure out the uncertainty on X having made an observation of Y; it is a regular communication problem, you have a handle on y you observe y and you guess what is X.

Now if I am excited about I X semicolon Y which is nothing but the average mutual information over this channel I get the following curves again the x axis is q and the y axis is I X semicolon Y. For p is equal to say 0.3 or 0.7 you will realize that whether you flip it to p or 1 minus p; you get the similar results the I X semicolon Y increases with q reaches a maximum and goes down.

But then if; so we are talking about average mutual information how much is the channel communicating $I(X; Y)$ is the measure of the goodness of the channel. So, when probability of error is high points is pretty bad it communicates, but it is not good I make the channel better, I reduce the probability of error p is equal to 0.1, look it has been able to communicate much more. And when it is an ideal channel right probability of error is 0; then indeed it can go right up to 1. So, it can actually effectively communicate one bit per use fine.

(Refer Slide Time: 55:36)

Information Theory, Coding and Cryptography

Summary

- Average Mutual Information
- Entropy and Self Information
- Conditional Entropy
- Joint Entropy

Indian Institute of Technology, Delhi

23

Ranjan Bose
Department of Electrical Engineering

This brings us to the end of this lecture; let us summarize what we have learned today. We started off by revisiting average mutual information, we will be looked at entropy and self information followed by the definition of conditional entropy joint entropy and we looked at the chain rule. And finally, we looked at an example of a binary symmetric channel and how the input probabilities affect $I(X; Y)$.

So with that, we come to the end of module 2.