UNIT-5

UNIT-5/LECTURE-1

INTRODUCTION TO INFORMATION THEORY

The main role of information theory was to provide the engineering and scientific communities with a mathematical framework for the theory of communication by establishing the fundamental limits on the performance of various communication systems. Its birth was initiated with the publication of the works of Claude E. Shannon who stated that it is possible to send in-formation-bearing signals at axed rate through a noisy communication channel with an arbitrarily small probability of error as long as the communication rate is below a certain quantity that depends on the channel characteristics; he baptized. this quantity with the name of channel capacity. He further pro-claimed that random sources {such as speech, music or image signals } possess an irreducible complexity beyond which they cannot be compressed distortion-free. He called this complexity the source entropy. He went on asserting that if a source has an entropy that is less than the capacity of a communication channel, then asymptotically error free transmission of the source over the channel can be achieved.

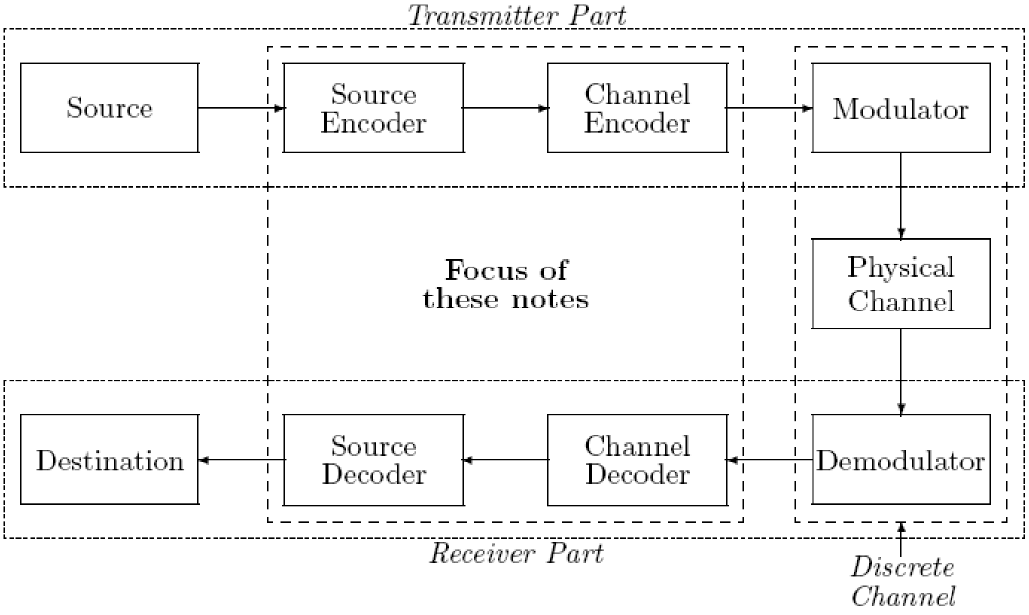


Figure 1.1: General model of a communication system.

A simple model of a general communication system is depicted in Figure 1.1.

Source: The source is usually modeled as a random process (the necessary background regarding random processes is introduced in Appendix B). It can be discrete (Finite or countable alphabet) or continuous (uncountable alphabet) in value and in time.

Source Encoder: Its role is to represent the source in a compact fashion by removing its unnecessary or redundant content (i.e., compression). Channel Encoder: Its role is to enable the reliable reproduction of the source encoder output after its transmission through a noisy communication channel. This is achieved by adding redundancy to the source encoder output.

Modulator: It transforms the channel encoder output into a waveform suitable for transmission over the physical channel. This is usually accomplished by varying the parameters of a sinusoidal signal in proportion with the data provided by the channel encoder output.

Physical Channel: It consists of the noisy (or unreliable) medium that the transmitted waveform traverses. It is usually modeled via a conditional (or transition) probability distribution of receiving an output given that a specific input was sent.

Receiver Part: It consists of the demodulator, the channel decoder and the source decoder where the reverse operations are performed. The destination represents the sink where the source estimate provided by the source decoder is reproduced.

UNIT-5/LECTURE-2

UNCERTAINTY AND INFORMATION

Let E be an event with probability Pr(E), and let I(E) represent the amount of information you gain when you learn that E has occurred (or equivalently, the amount of uncertainty you lose after learning that E has happened). Then a natural question to ask is \what properties should I(E) have?" The answer to the question may vary person by person. Here are some common properties that I(E), which is called the self-information, is reasonably expected to have.

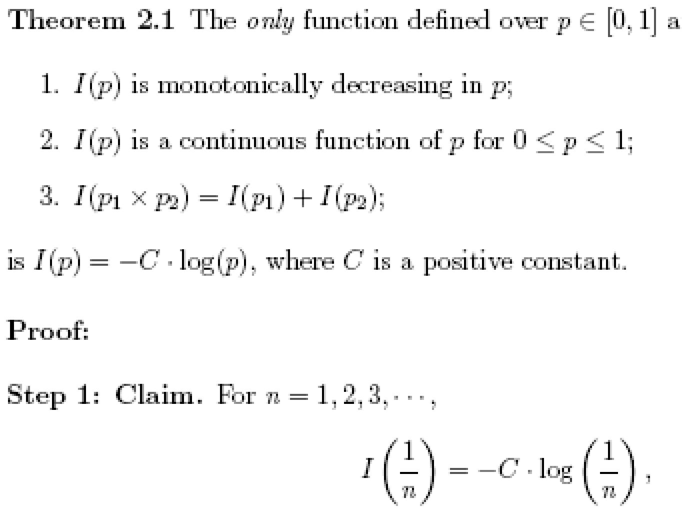
1. I(E) should be a function of Pr(E).

   In other words, this property says that I(E) = I(Pr(E)), where I(¢) is a function defined over an event space, and I(¢) is a function defined over [0; 1]. In general, people expect that the less likely an event is, the more information you have gained when you learn it has happened. In other words, I(Pr(E)) is a decreasing function of Pr(E).

2. I(Pr(E)) should be continuous in Pr(E).

   Intuitively, we should expect that a small change in Pr(E) corresponds to a small change in the uncertainty of E.

3. If E_1 and E_2 are independent events, then I(E_1 \ E_2) = I(E_1) + I(E_2), or equivalently, I(Pr(E_1) £ Pr(E_2)) = I(Pr(E_1)) + I(Pr(E_2)).
   This property declares that the amount of uncertainty we lose by learning that both E_1 and E_2 have occurred should be equal to the sum of individual uncertainty losses for independent E_1 and E_2.



Theorem 2.1 The *only* function defined over $p \in [0,1]$ and satisfying

1. $I(p)$ is monotonically decreasing in p;
2. $I(p)$ is a continuous function of p for $0 \leq p \leq 1$;
3. $I(p_1 \times p_2) = I(p_1) + I(p_2)$;

is $I(p) = -C \cdot \log(p)$, where C is a positive constant.

Proof:

Step 1: Claim. For $n = 1, 2, 3, \cdots$,

$$I\left(\frac{1}{n}\right) = -C \cdot \log\left(\frac{1}{n}\right),$$

where $C > 0$ is a constant.

Proof: Conditions 1 and 3 respectively imply

$$n < m \ \Rightarrow\ I\left(\frac{1}{n}\right) < I\left(\frac{1}{m}\right). \qquad (2.1.1)$$

and

$$I\left(\frac{1}{mn}\right) = I\left(\frac{1}{m}\right) + I\left(\frac{1}{n}\right) \qquad (2.1.2)$$

where $n, m = 1, 2, 3, \cdots$. Now using (2.1.2), we can show by induction that

$$I\left(\frac{1}{n^k}\right) = k \cdot I\left(\frac{1}{n}\right) \qquad (2.1.3)$$

for all positive integer n and non-negative integer k Note that (2.1.3) already proves the claim for the case of $n = 1$.

Now let n be a fixed positive integer greater than 1. Then for any positive integer r, there exists non-negative integer k such that

$$n^k \le 2^r < n^{k+1}.$$

By (2.1.1), we obtain

$$I\left(\frac{1}{n^k}\right) \le I\left(\frac{1}{2^r}\right) < I\left(\frac{1}{n^{k+1}}\right),$$

which together with (2.1.3), yields

$$k \cdot I\left(\frac{1}{n}\right) \le r \cdot I\left(\frac{1}{2}\right) < (k+1) \cdot I\left(\frac{1}{n}\right).$$

Hence, by $I(1/n) > I(1) = 0$,

$$\frac{k}{r} \le \frac{I(1/2)}{I(1/n)} \le \frac{k+1}{r}.$$

On the other hand, by the monotonity of logarithm, we obtain

$$\log n^k \le \log 2^r \le \log n^{k+1} \ \Leftrightarrow\ \frac{k}{r} \le \frac{\log(2)}{\log(n)} \le \frac{k+1}{r}.$$

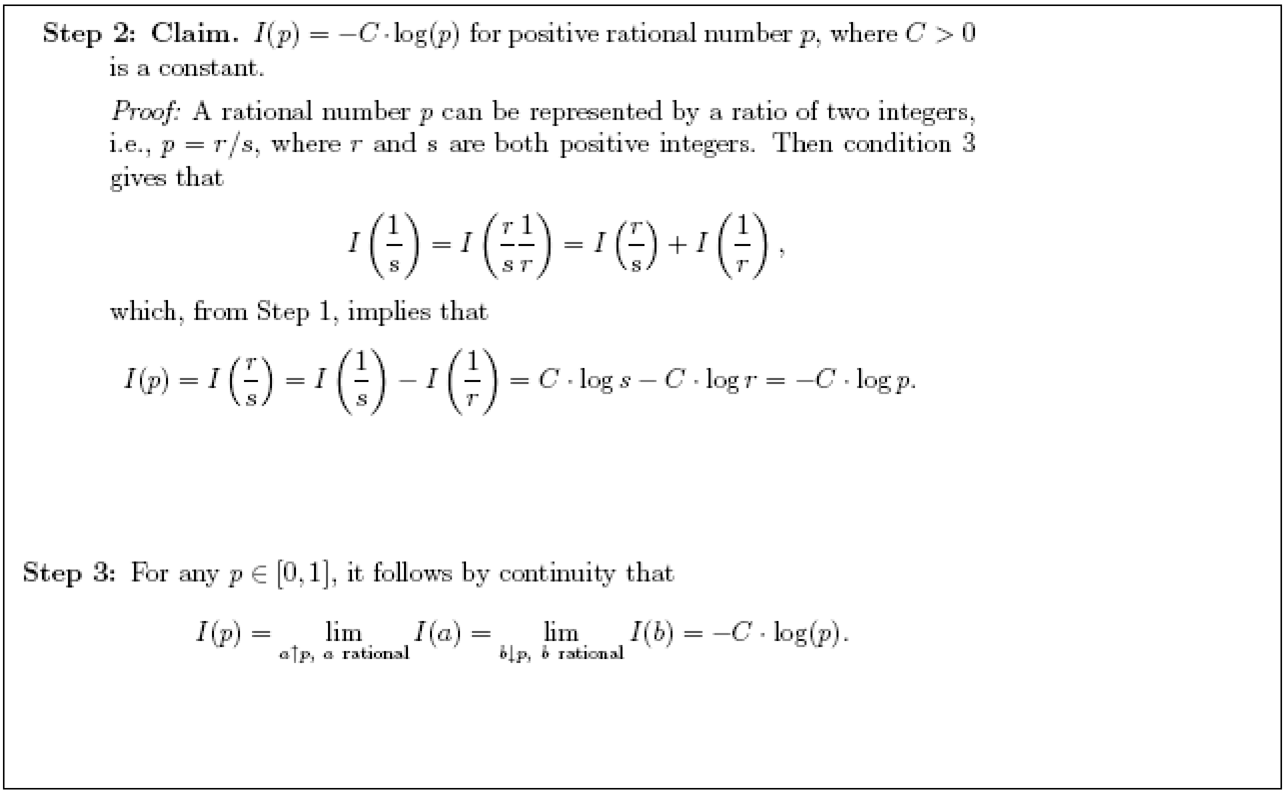
Therefore,

$$\left| \frac{\log(2)}{\log(n)} - \frac{I(1/2)}{I(1/n)} \right| < \frac{1}{r}.$$

Since n is fixed, and r can be made arbitrarily large, we can let $r \to \infty$ to get:

$$I\left(\frac{1}{n}\right) = C \cdot \log(n).$$

where $C = I(1/2)/\log(2) > 0$. This completes the proof of the claim.



Step 2: Claim. $I(p) = -C \cdot \log(p)$ for positive rational number p, where $C > 0$ is a constant.

Proof: A rational number p can be represented by a ratio of two integers, i.e., $p = r/s$, where r and s are both positive integers. Then condition 3 gives that

$$I\left(\frac{1}{s}\right) = I\left(\frac{r}{s}\frac{1}{r}\right) = I\left(\frac{r}{s}\right) + I\left(\frac{1}{r}\right),$$

which, from Step 1, implies that

$$I(p) = I\left(\frac{r}{s}\right) = I\left(\frac{1}{s}\right) - I\left(\frac{1}{r}\right) = C \cdot \log s - C \cdot \log r = -C \cdot \log p.$$

Step 3: For any $p \in [0, 1]$, it follows by continuity that

$$I(p) = \lim_{a \uparrow p,\ a \text{ rational}} I(a) = \lim_{b \downarrow p,\ b \text{ rational}} I(b) = -C \cdot \log(p).$$

UNIT-5/LECTURE-3

ENTROPY

Entropy is a measure of the amount of information (or uncertainty) contained in the source. The source can be modeled as a random process, which is a collection of random variables indexed through an index set (cf. Appendix B). For simplicity, we first assume that the index set associated with the random process corresponding to the source consists of only one index. It is also assumed that the source alphabet X is finite. Then as indicated in the previous subsection, the self-information can be probabilistically defined as:

$$I(x) = - \log P_X (x);$$

Where $P_X (x)$ is the probability distribution of the source X.

This definition fits the intuition that a less likely outcome will bring more information. By extending the concept, entropy is defined as follows.

Entropy

For a source X, the entropy H(X) is defined by

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} P_X(x) \cdot \log P_X(x) = E[-\log P_X(X)] = E[\mathcal{I}(X)].$$

By the above definition, entropy can be interpreted as the expected or average amount of (self-) information you gain when you learn that one of the lXl out-comes has occurred, where lXl is the cardinality of X. Another interpretation is that H(X) is a measure of uncertainty of random variable X. Sometimes, H(X) is also written as $H(P_X)$ for notation convenience.

When the base of the logarithm operation is 2, entropy is expressed in bits; when the natural logarithm is employed, entropy is measured in nats. For ex-ample, the entropy of a fair coin source is 1 bit or log(2) nat.

Example:-

Let X be a random variable with $P_X (1) = p$ and $P_X (0) = 1 - p$.

Then $H(X) = -p \log p - (1-p) \log(1-p)$.

This is called the binary entropy function.

Joint entropy and conditional entropy

We now consider the case where the index set associated with the random source consists of two indexes. Then the self-information of such a source is probabilistically defined as:

I(x; y) = - log P $_{(X;Y)}$ (x; y);

Joint entropy

$$\begin{aligned} H(X,Y) &\triangleq -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\cdot\log P_{X,Y}(x,y) \\ &= E[-\log P_{X,Y}(X,Y)]. \end{aligned}$$

Conditional entropy

$$H(Y|X) \triangleq \sum_{x\in\mathcal{X}} P_X(x)\left(-\sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\cdot\log P_{Y|X}(y|x)\right)$$

Relationship Between different types of Entropy

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

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

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

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

$$H(X|Y) \leq H(X)$$

	RGPV QUESTIONS	Year	Marks
Q.1	What is entropy? show that the entropy is maximum when all the symbols are equiprobable.Assume M=2.	DEC-2012,	10
Q.2	What is entropy? show that the entropy is maximum when all the symbols are equiprobable.Assume M=3.	DEC-2013,DEC-2014	7,7

UNIT-5/LECTURE-4

MUTUAL INFORMATION

MUTUAL INFORMATION

For two random variables X and Y , the mutual information between X and Y is the reduction in the uncertainty of Y due to the knowledge of X (or vice versa).

For example the mutual information of the channel is the first argument X_1. A dual definition of mutual information states that it is the average amount of information that Y has (or contains) about X or X has (or contains) about Y . Under this definition, we can say that the shared (or mutual) uncertainty (or information) between channel sender and channel receiver is Uncertainty X_1.

We can think of the mutual information between X and Y in terms of a channel whose input is X and whose output is Y . Thereby the reduction of the uncertainty is by definition the total uncertainty of X (i.e. H(X)) minus the uncertainty of X after For a source X, the entropy H(X) is defind by observing Y (i.e. H(XjY) Mathematically, it is

$$\text{mutual information} = I(X;Y) \triangleq H(X) - H(X|Y).$$

Properties of mutual information

1. $I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$.
2. $I(X;Y) = I(Y;X)$.
3. $I(X;Y) = H(X) + H(Y) - H(X,Y)$.
4. $I(X;Y) \leq H(X)$ with equality holds if, and only, if X is a function of Y (i.e., $X = f(Y)$ for some function $f(\cdot)$).
5. $I(X;Y) \geq 0$ with equality holds if, and only if, X and Y are independent.

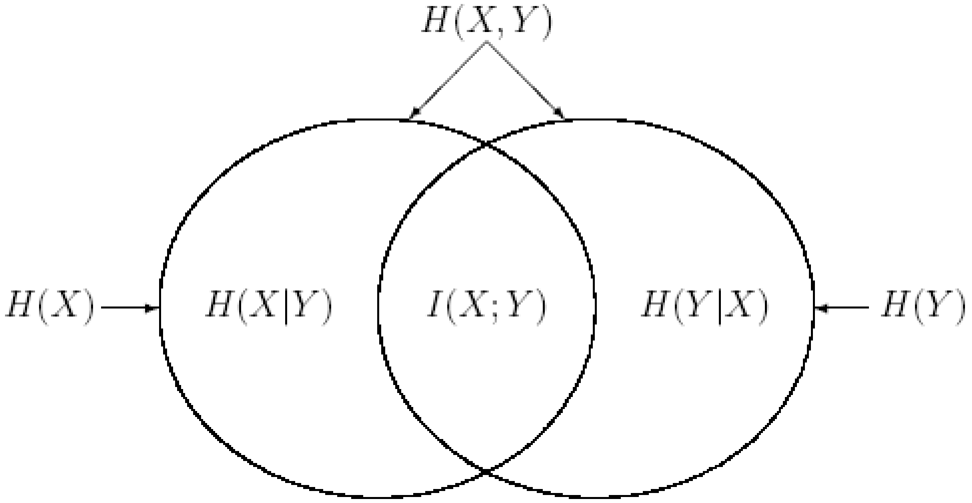


Figure : Relation between entropy and mutual information

	RGPV QUESTIONS	Year	Marks
Q.1	Find the mutual information and channel capacity of the channel shown in figure below. Given P(X1)= 0.6 and P(X2) = 0.4	DEC 2013	7

UNIT-5/LECTURE-5

SOURCE CODING THEOREM,INFORMATION CAPACITY THEOREM

Shannon–Hartley theorem

In information theory, the Shannon–Hartley theorem tells the maximum rate at which information can be transmitted over a communications channel of a specified bandwidth in the presence of noise. It is an application of the noisy channel coding theorem to the archetypal case of a continuous-time analog communications channel subject to Gaussian noise.

The theorem establishes Shannon's channel capacity for such a communication link, a bound on the maximum amount of error-free digital data (that is, information) that can be transmitted with a specified bandwidth in the presence of the noise interference, assuming that the signal power is bounded, and that the Gaussian noise process is characterized by a known power or power spectral density. The law is named after Claude Shannon and Ralph Hartley.

Statement of the theorem

Considering all possible multi-level and multi-phase encoding techniques, the Shannon–Hartley theorem states the channel capacity C, meaning the theoretical tightest upper bound on the information rate (excluding error correcting codes) of clean (or arbitrarily low bit error rate) data that can be sent with a given average signal power S through an analog communication channel subject to additive white Gaussian noise of power N, is:

$$C = B \log_2 \left(1 + \frac{S}{N} \right)$$

where

C is the channel capacity in bits per second;

B is the bandwidth of the channel in hertz (passband bandwidth in case of a modulated signal);

S is the average received signal power over the bandwidth (in case of a modulated signal, often denoted C, i.e. modulated carrier), measured in watts (or volts squared);

N is the average noise or interference power over the bandwidth, measured in watts (or volts squared); and

S/N is the signal-to-noise ratio (SNR) or the carrier-to-noise ratio (CNR) of the communication signal to the Gaussian noise interference expressed as a linear power ratio (not as logarithmic decibels).

Shannon's source coding theorem

This article is about the theory of source coding in data compression. For the term in computer programming, see Source code.

In information theory, Shannon's source coding theorem (or noiseless coding theorem) establishes the limits to possible data compression, and the operational meaning of the Shannon entropy.

The source coding theorem shows that (in the limit, as the length of a stream of independent and identically-distributed random variable (i.d.) data tends to infinity) it is impossible to compress the data such that the code rate (average number of bits per symbol) is less than the Shannon entropy of the source, without it being virtually certain that information will be lost. However it is possible to get the code rate arbitrarily close to the Shannon entropy, with negligible probability of loss.

The source coding theorem for symbol codes places an upper and a lower bound on the minimal possible expected length of code words as a function of the entropy of the input word (which is viewed as a random variable) and of the size of the target alphabet.

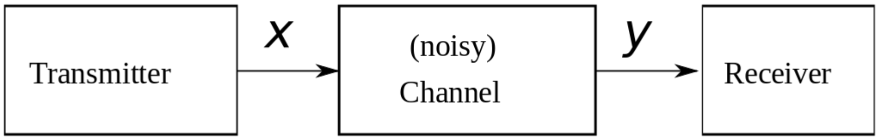
	RGPV QUESTIONS	Year	Marks
Q.1	State and prove Shannon Hartley theorem.	JUNE-2014	7

CHANNEL CAPACITY

Channel capacity

In information theory, channel capacity is the tightest upper bound on the rate of information that can be reliably transmitted over a communications channel. By the noisy-channel coding theorem, the channel capacity of a given channel is the limiting information rate (in units of information per unit time) that can be achieved with arbitrarily small error probability.[

Information theory, developed by Claude E. Shannon during World War II, defines the notion of channel capacity and provides a mathematical model by which one can compute it. The key result states that the capacity of the channel, as defined above, is given by the maximum of the mutual information between the input and output of the channel, where the maximization is with respect to the input distribution.



Let $X$and Ybe the random variables representing the input and output of the channel, respectively. Let $p_{Y|X}(y|x)$be the conditional distribution function of Ygiven X, which is an inherent fixed property of the communications channel. Then the choice of the marginal distribution $p_X(x)$completely determines the joint distribution $p_{X,Y}(x,y)$due to the identity

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)\,p_X(x)$$

which, in turn, induces a mutual information $I(X;Y)$. The channel capacity is defined as

$$C = \sup_{p_X(x)} I(X;Y)$$

Where the sup is taken over all possible choices of $p_X(x)$.

UNIT-5/LECTURE-7

CHANNEL MODELS

Identity channel

An identity channel has equal-size in input and output alphabets (lXl = lYl), and channel transition probability satisfying

$$Q_{Y|X}(y|x) = \text{either } 1 \text{ or } 0.$$

In such channel, H(Y l X) = 0. since no extra information provides by Y when X is given. As a consequence,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y), \end{aligned}$$

and the channel capacity is

$$\max_X I(X;Y) = \max_X H(Y) = \log|\mathcal{Y}| \text{ nats/channel usage.}$$

Binary symmetric channels

A binary symmetric channel (BSC) is a channel with binary input and output alphabet, and the probability for one input symbol to be complemented at the output is equal to that for another input symbol as shown in Figure.

This is the simplest model of a channel with errors; yet it captures most of the complexity of the general problems. To compute the channel capacity of it

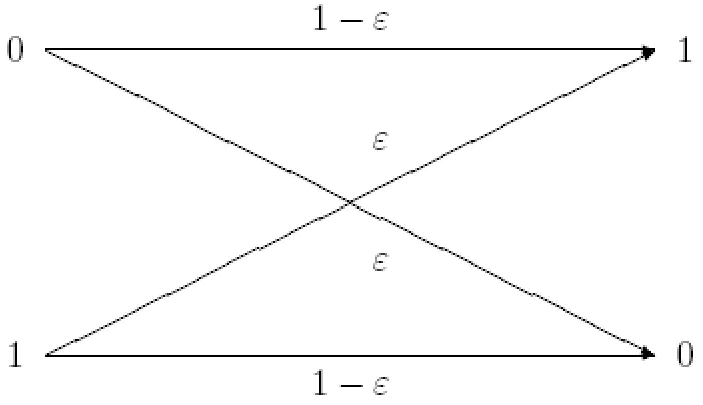
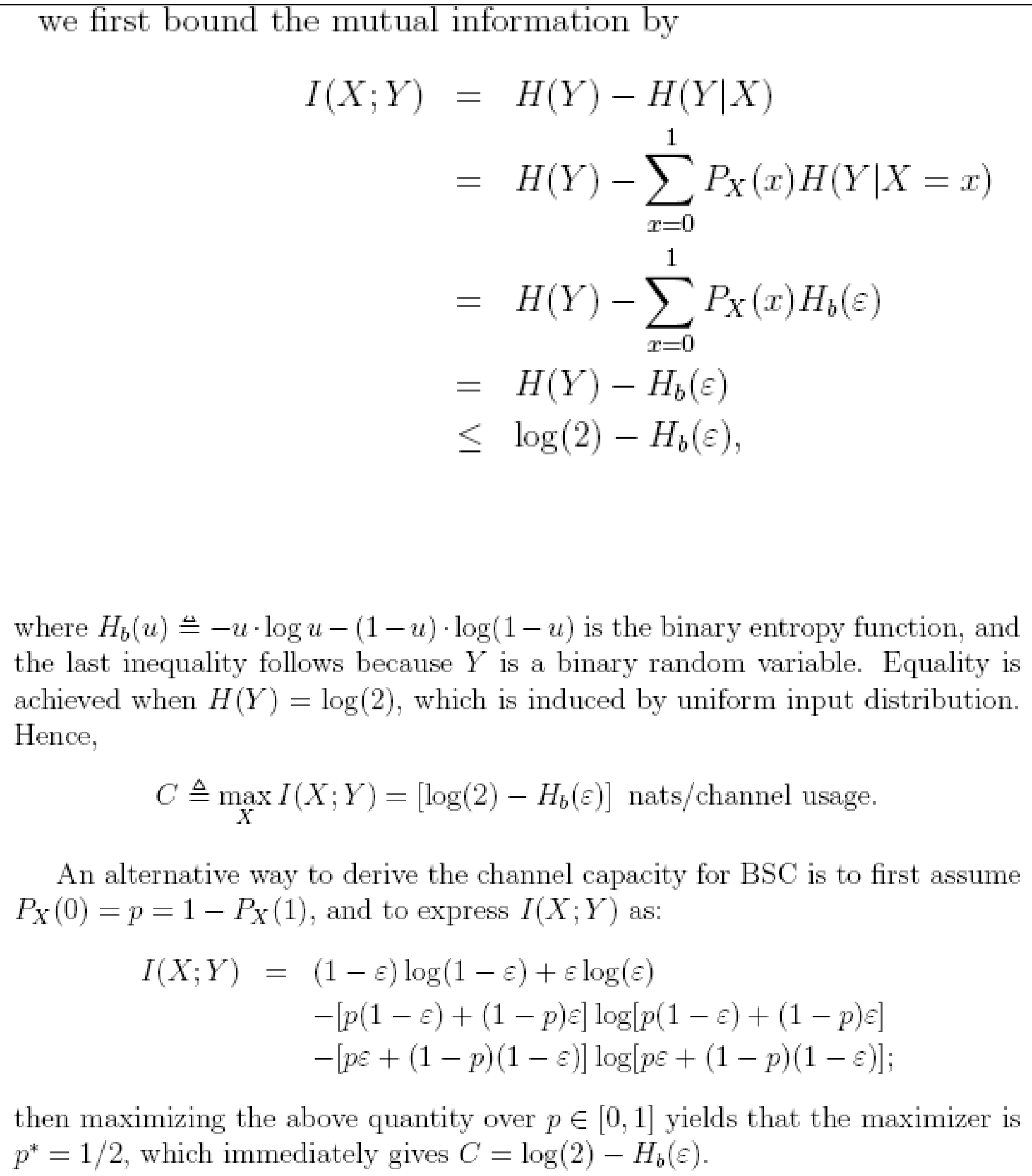


Figure: Binary symmetric channel.



we first bound the mutual information by

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_{x=0}^{1} P_X(x) H(Y|X=x) \\ &= H(Y) - \sum_{x=0}^{1} P_X(x) H_b(\varepsilon) \\ &= H(Y) - H_b(\varepsilon) \\ &\leq \log(2) - H_b(\varepsilon), \end{aligned}$$

where $H_b(u) \triangleq -u \cdot \log u - (1-u) \cdot \log(1-u)$ is the binary entropy function, and the last inequality follows because Y is a binary random variable. Equality is achieved when $H(Y) = \log(2)$, which is induced by uniform input distribution. Hence,

$$C \triangleq \max_X I(X;Y) = [\log(2) - H_b(\varepsilon)] \text{ nats/channel usage.}$$

An alternative way to derive the channel capacity for BSC is to first assume $P_X(0) = p = 1 - P_X(1)$, and to express $I(X;Y)$ as:

$$\begin{aligned} I(X;Y) = \;& (1-\varepsilon)\log(1-\varepsilon) + \varepsilon \log(\varepsilon) \\ & -[p(1-\varepsilon) + (1-p)\varepsilon] \log[p(1-\varepsilon) + (1-p)\varepsilon] \\ & -[p\varepsilon + (1-p)(1-\varepsilon)] \log[p\varepsilon + (1-p)(1-\varepsilon)]; \end{aligned}$$

then maximizing the above quantity over $p \in [0,1]$ yields that the maximizer is $p^* = 1/2$, which immediately gives $C = \log(2) - H_b(\varepsilon)$.

Binary erasure channels

The binary erasure channel (BEC) has a form similar to BSC, except that bits are erased with some probability. It is shown in Figure.

We calculate the capacity of the binary erasure channel as follows:

$$\begin{aligned} C &= \max_X I(X;Y) \\ &= \max_X [H(Y) - H(Y|X)] \\ &= \max_X [H(Y)] - H_b(\varepsilon). \end{aligned}$$

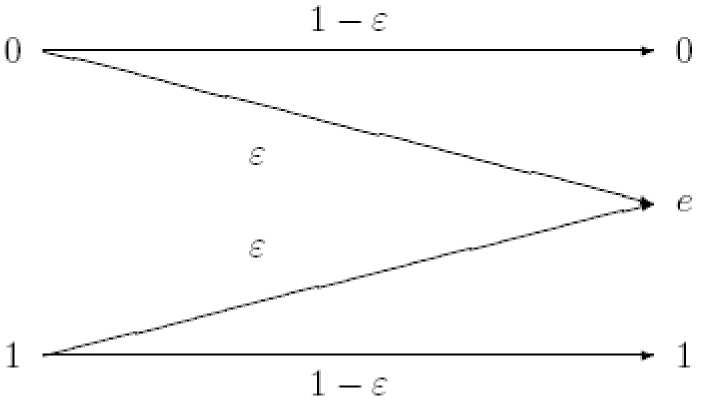
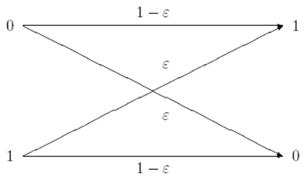


Figure : Binary erasure channel.

Now we note that $H(Y) \leq \log(3)$ because the size of the output alphabet $\{0, 1, e\}$ is 3, which is achieved by uniform channel output. But since there is no input distribution yielding uniform channel output, we cannot take $\log(3)$ as an achievable maximum value. Some specific approach needs to be created for the calculation of its channel capacity.

	RGPV QUESTIONS	Year	Marks
Q.1	For BSC shown in fig, find the channel capacity for E (Probability)=0.9 	JUNE-2014	7

UNIT-5/LECTURE-8

CHANNEL CODING

Channel coding

The purpose of channel coding theory is to find codes which transmit quickly, contain many valid code words and can correct or at least detect many errors. While not mutually exclusive, performance in these areas is a trade off. So, different codes are optimal for different applications. The needed properties of this code mainly depend on the probability of errors happening during transmission. In a typical CD, the impairment is mainly dust or scratches. Thus codes are used in an interleaved manner. The data is spread out over the disk.

Although not a very good code, a simple repeat code can serve as an understandable example. Suppose we take a block of data bits (representing sound) and send it three times. At the receiver we will examine the three repetitions bit by bit and take a majority vote. The twist on this is that we don't merely send the bits in order. We interleave them. The block of data bits is first divided into 4 smaller blocks. Then we cycle through the block and send one bit from the first, then the second, etc. This is done three times to spread the data out over the surface of the disk. In the context of the simple repeat code, this may not appear effective. However, there are more powerful codes known which are very effective at correcting the "burst" error of a scratch or a dust spot when this interleaving technique is used.

Other codes are more appropriate for different applications. Deep space communications are limited by the thermal noise of the receiver which is more of a continuous nature than a bursty nature. Likewise, narrowband modems are limited by the noise, present in the telephone network and also modeled better as a continuous disturbance. Cell phones are subject to rapid fading. The high frequencies used can cause rapid fading of the signal even if the receiver is moved a few inches. Again there are classes of channel codes that are designed to combat fading.

Linear codes

The term algebraic coding theory denotes the sub-field of coding theory where the properties of codes are expressed in algebraic terms and then further researched.

Algebraic coding theory is basically divided into two major types of codes:

1. Linear block codes
2. Convolution codes.

It analyzes the following three properties of a code – mainly:code word length

- total number of valid code words
- the minimum distance between two valid code words, using mainly the Hamming distance, sometimes also other distances like the Lee distance.

Linear block codes

Linear block codes have the property of linearity, i.e. the sum of any two codeword is also a code word, and they are applied to the source bits in blocks, hence the name linear block codes. There are block codes that are not linear, but it is difficult to prove that a code is a good one without this property.

Linear block codes are summarized by their symbol alphabets (e.g., binary or ternary) and parameters (n,m,d_{min})[where

1. n is the length of the codeword, in symbols,
2. m is the number of source symbols that will be used for encoding at once,
3. d_{min} is the minimum hamming distance for the code.

There are many types of linear block codes, such as

1. Cyclic codes (e.g., Hamming codes)
2. Repetition codes
3. Parity codes
4. Polynomial codes (e.g., BCH codes)
5. Reed–Solomon codes
6. Algebraic geometric codes
7. Reed–Muller codes
8. Perfect codes.

Block codes are tied to the sphere packing problem, which has received some attention over the years. In two dimensions, it is easy to visualize. Take a bunch of pennies flat on the table and push them together. The result is a hexagon pattern like a bee's nest. But block codes rely on more dimensions which cannot easily be visualized. The powerful (24,12) Golay code used in deep space communications uses 24 dimensions. If used as a binary code (which it usually is) the dimensions refer to the length of the codeword as defined above.

The theory of coding uses the *N*-dimensional sphere model. For example, how many pennies can be packed into a circle on a tabletop, or in 3 dimensions, how many marbles can be packed into a globe. Other considerations enter the choice of a code. For example, hexagon packing into the constraint of a rectangular box will leave empty space at the corners. As the dimensions get larger, the percentage of empty space grows smaller. But at certain dimensions, the packing uses all the space and these codes are the so-called "perfect" codes. The only nontrivial and useful perfect codes are the distance-3 Hamming codes with parameters satisfying (2^r – 1, 2^r – 1 – r, 3), and the [23,12,7] binary and [11,6,5] ternary Golay codes.[

Another code property is the number of neighbors that a single codeword may have.[Again, consider pennies as an example. First we pack the pennies in a rectangular grid. Each penny will have 4 near neighbors (and 4 at the corners which are farther away). In a hexagon, each penny will have 6 near neighbors. When we increase the dimensions, the number of near neighbors increases very rapidly. The result is the number of ways for noise to make the receiver choose a neighbor (hence an error) grows as well. This is a fundamental limitation of block codes, and indeed all codes. It may be harder to cause an error to a single neighbor, but the number of neighbors can be large enough so the total error probability actually suffers.

Properties of linear block codes are used in many applications. For example, the syndrome-coset uniqueness property of linear block codes is used in trellis shaping, one of the best known shaping codes. This same property is used in sensor networks for distributed source coding

Convolutional codes

The idea behind a convolution code is to make every codeword symbol be the weighted sum of the various input message symbols. This is like convolution used in LTI systems to find the output of a system, when you know the input and impulse response.

So we generally find the output of the system convolution encoder, which is the convolution of the input bit, against the states of the convolution encoder, registers.

Fundamentally, convolution codes do not offer more protection against noise than an equivalent block code. In many cases, they generally offer greater simplicity of implementation over a block code of equal power. The encoder is usually a simple circuit which has state memory and some feedback logic, normally XOR gates. The decoder can be implemented in software or firmware.

The Viterbi algorithm is the optimum algorithm used to decode convolution codes. There are simplifications to reduce the computational load. They rely on searching only the most likely paths. Although not optimum, they have generally been found to give good results in the lower noise environments.

Convolution codes are used in voice band modems (V.32, V.17, and V.34) and in GSM mobile phones, as well as satellite and military communication devices.

	RGPV QUESTIONS	Year	Marks
Q.1	What is coding efficiency .Show that the coding efficiency is maximum when p(0)=p(1)	DEC-2012	10

UNIT-5/LECTURE-9

HUFFMAN CODING,SHANNON FANO ELIAS CODING

Huffman code : a variable-length optimal code

In this subsection, we will introduce a simple optimal variable-length code, named Huffman code. Here optimality means that it yields the minimum average codeword length among all codes on the same source. We now begin our examination of Huffman coding with a simple observation.

Give a source with source alphabet f1; : : : ; Kg and probability fp$_1$; : : : ; p$_K$ g. Let `$_i$ be the binary codeword length of symbol i. Then there exists an optimal uniquely-decodable variable-length code satisfying:

1. $p_i > p_j$ implies $\ell_i \leq \ell_j$.

2.The two longest codeword have the same length.

3.The two longest codeword differ only in the last bit and correspond to the two least-frequent symbols.

Proof: First, we note that any optimal code that is uniquely decodable must satisfy the Kraft inequality. In addition, for any set of codeword lengths that satisfy the Kraft inequality, there exists a prefix code who takes the same set as its set of codeword lengths. Therefore, it succes to show that there exists an optimal prefix code satisfying the above three properties.

1. Suppose there is an optimal prfix code violating the observation. Then we can interchange the codeword for symbol i with that for symbol j, and yield a better code.
2. Without loss of generality, let the probabilities of the source symbols satisfy

$$p_1 \leq p_2 \leq p_3 \leq \cdots \leq p_K.$$

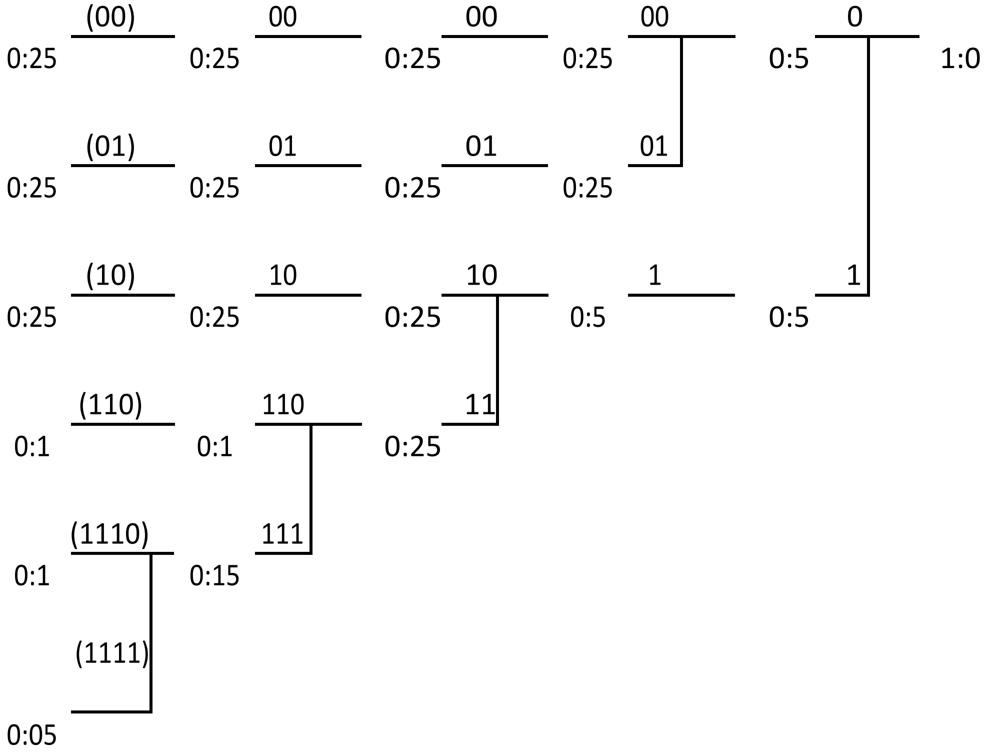
3. Since all the code words of a prefix code reside in the leaves, we can interchange the siblings of two branches without changing the average codeword length. Property 2 implies that the two least-frequent code words has the same codeword length. Hence, by repeatedly interchanging the siblings of a tree, we can result in a prefix code which meets the requirement.

The above observation proves the existence of an optimal prefix code that statistics the stated properties. As it turns out, Huffman code is one of such codes. In what follows, we will introduce the construction algorithm of Huffman code.

We now give an example of Huffman encoding.

Example : Consider a source with alphabet f1; 2; 3; 4; 5; 6g with probability 0:25; 0:25; 0:25; 0:1; 0:1 and 0:05, respectively. By following the Huffman encoding procedure as shown in Figure 3.6, we obtain the Huffman code as

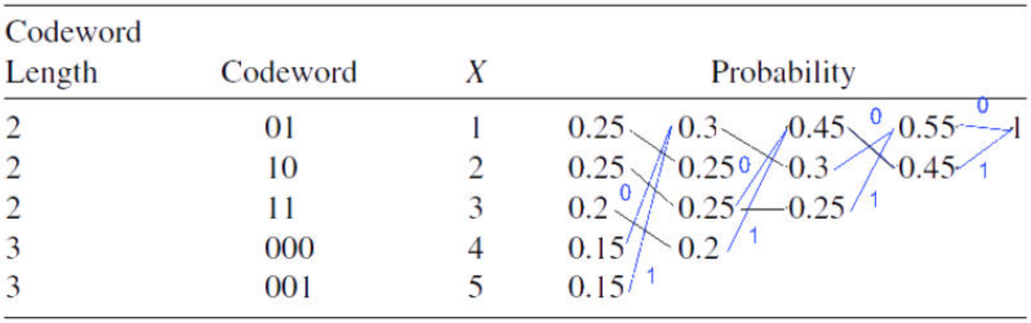
00; 01; 10; 110; 1110; 1111:



Example:

	1	2	3	4	5
X	0.25	0.25	0.2	0.15	0.15

We can combine the symbols 4 and 5 into a single source symbol, with a probability assignment 0.30. Proceeding this way, combining the two least likely symbols into one symbol until we are nally left with only one symbol, and then assigning codeword to the symbols, we obtain the following table:



0
1
0
1
0
1
0
1
0
1

Codeword Length	Codeword	X	Probability				
2	01	1	0.25	0.3	0.45	0.55	1
2	10	2	0.25	0.25	0.3	0.45	
2	11	3	0.2	0.25	0.25		
3	000	4	0.15	0.2			
3	001	5	0.15				

This code has average length L = 2.3 bits and H(X) = 2.286.

Shannon-Fano-Elias code

Assume $\mathcal{X} = \{0, 1, \ldots, L-1\}$ and $P_X(x) > 0$ for all $x \in \mathcal{X}$. Define

$$F(x) \triangleq \sum_{a \le x} P_X(a),$$

and

$$\bar{F}(x) \triangleq \sum_{a < x} P_X(a) + \frac{1}{2} P_X(x).$$

Encoder: For any $x \in \mathcal{X}$, express $\bar{F}(x)$ in binary decimal, say

$$\bar{F}(x) = .c_1 c_2 \ldots c_k \ldots,$$

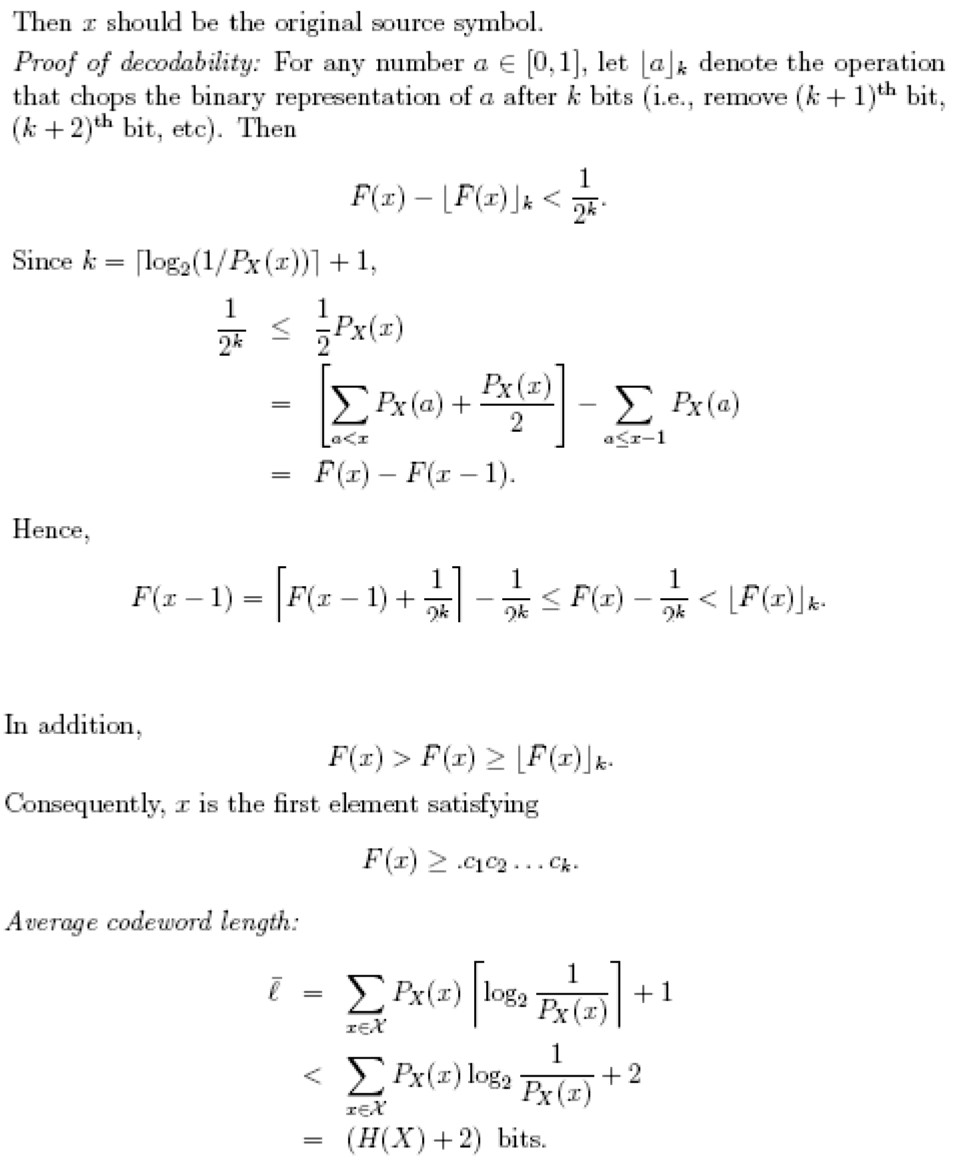
and take the first k bits as the codeword of source symbol x, i.e.,

$$(c_1, c_2, \ldots, c_k),$$

where $k \triangleq \lceil \log_2(1/P_X(x)) \rceil + 1$.

Decoder: Given codeword $(c_1, \ldots, c_k)$, compute the cumulative sum of $F(\cdot)$ starting from the smallest element in $\{0, 1, \ldots, L-1\}$ until the first x satisfying

$$F(x) > .c_1 \ldots c_k.$$



Then x should be the original source symbol.

Proof of decodability: For any number $a \in [0,1]$, let $\lfloor a \rfloor_k$ denote the operation that chops the binary representation of a after k bits (i.e., remove $(k+1)^{\text{th}}$ bit, $(k+2)^{\text{th}}$ bit, etc). Then

$$\bar{F}(x) - \lfloor \bar{F}(x) \rfloor_k < \frac{1}{2^k}.$$

Since $k = \lceil \log_2(1/P_X(x)) \rceil + 1$,

$$\begin{aligned} \frac{1}{2^k} &\le \frac{1}{2} P_X(x) \\ &= \left[\sum_{a<x} P_X(a) + \frac{P_X(x)}{2} \right] - \sum_{a \le x-1} P_X(a) \\ &= \bar{F}(x) - F(x-1). \end{aligned}$$

Hence,

$$F(x-1) = \left[F(x-1) + \frac{1}{2^k} \right] - \frac{1}{2^k} \le \bar{F}(x) - \frac{1}{2^k} < \lfloor \bar{F}(x) \rfloor_k.$$

In addition,

$$F(x) > \bar{F}(x) \ge \lfloor \bar{F}(x) \rfloor_k.$$

Consequently, x is the first element satisfying

$$F(x) \ge .c_1 c_2 \ldots c_k.$$

Average codeword length:

$$\begin{aligned} \bar{\ell} &= \sum_{x \in \mathcal{X}} P_X(x) \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil + 1 \\ &< \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{1}{P_X(x)} + 2 \\ &= (H(X) + 2) \text{ bits.} \end{aligned}$$

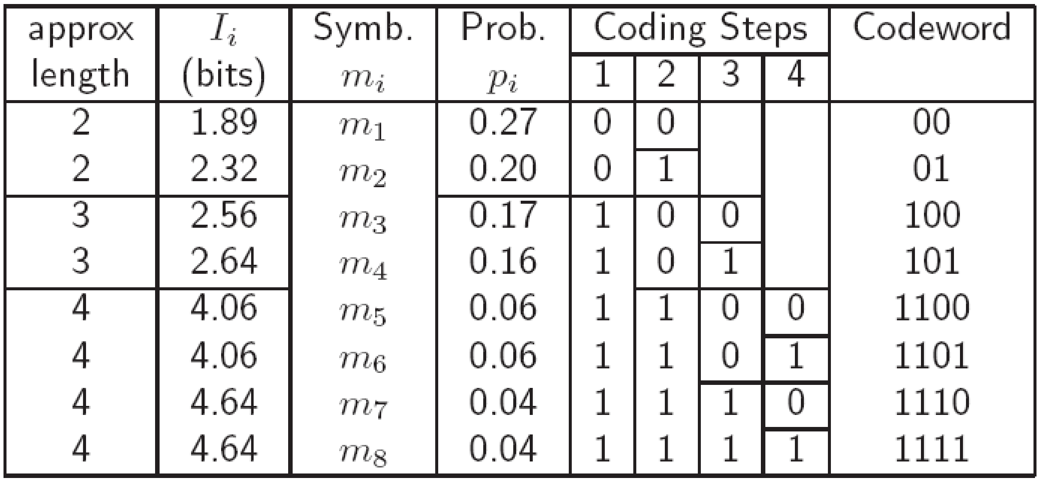
Shannon-Fano Coding

Shannon-Fano source encoding follows the steps

1. Order symbols m_i in descending order of probability
2. Divide symbols into subgroups such that the subgroup's probabilities (i.e. information contests) are as close as possible can be two symbols as a subgroup if there are two close probabilities (i.e. information contests), can also be only one symbol as a subgroup if none of the probabilities are close

3. Allocating codeword: assign bit 0 to top subgroup and bit 1 to bottom subgroup.
4. Iterate steps 2 and 3 as long as there is more than one symbol in any subgroup
5. Extract variable-length codewords from the resulting tree (top-down)

Example



approx	I_i	Symb.	Prob.	Coding Steps				Codeword
length	(bits)	m_i	p_i	1	2	3	4	
2	1.89	m_1	0.27	0	0			00
2	2.32	m_2	0.20	0	1			01
3	2.56	m_3	0.17	1	0	0		100
3	2.64	m_4	0.16	1	0	1		101
4	4.06	m_5	0.06	1	1	0	0	1100
4	4.06	m_6	0.06	1	1	0	1	1101
4	4.64	m_7	0.04	1	1	1	0	1110
4	4.64	m_8	0.04	1	1	1	1	1111

- Less probable symbols are coded by longer code words, while higher probable symbols are assigned short codes

- Entropy for the given set of symbols: H = 2.6906 (bits/symbol)

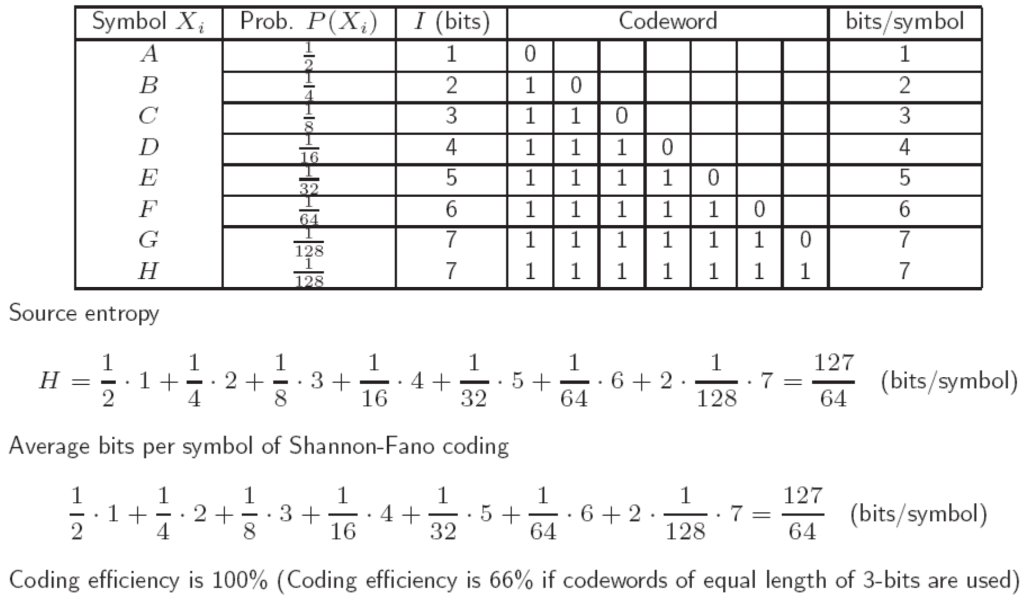
- Average code word length with Shannon-Fano coding:

$$0.47 \cdot 2 + 0.33 \cdot 3 + 0.2 \cdot 4 = 2.73 \quad \text{(bits/symbol)}$$

Coding efficiency:

$$\frac{\text{source information rate}}{\text{average source output rate}} = \frac{R_s \cdot H}{R_s \cdot 2.73} = \frac{2.6906}{2.73} = 98.56\%$$

Example



Symbol X_i	Prob. $P(X_i)$	I (bits)	Codeword							bits/symbol
A	$\frac{1}{2}$	1	0							1
B	$\frac{1}{4}$	2	1	0						2
C	$\frac{1}{8}$	3	1	1	0					3
D	$\frac{1}{16}$	4	1	1	1	0				4
E	$\frac{1}{32}$	5	1	1	1	1	0			5
F	$\frac{1}{64}$	6	1	1	1	1	1	0		6
G	$\frac{1}{128}$	7	1	1	1	1	1	1	0	7
H	$\frac{1}{128}$	7	1	1	1	1	1	1	1	7

Source entropy

$$H = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{32}\cdot 5 + \frac{1}{64}\cdot 6 + 2\cdot\frac{1}{128}\cdot 7 = \frac{127}{64} \quad \text{(bits/symbol)}$$

Average bits per symbol of Shannon-Fano coding

$$\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{32}\cdot 5 + \frac{1}{64}\cdot 6 + 2\cdot\frac{1}{128}\cdot 7 = \frac{127}{64} \quad \text{(bits/symbol)}$$

Coding efficiency is 100% (Coding efficiency is 66% if codewords of equal length of 3-bits are used)

	RGPV QUESTIONS	Year	Marks
Q.1	Apply the Shannon-fano coding procedure to find coding efficiency for the following message. [X]=[X1,X2,X3,X4,X5,X6 ,X7,X8] [P(x)]=[0.2,0.2,0.15,0.15,0.1,0.1,0.05.0.05]	DEC-2012	10
Q.2	Apply the Shannon fano coding procedure for the following essemble and find the coding efficiency (take M=2). [x] = [x1 x2 x3 x4 x5 x6 x7 x8] [p] = [1/4 1/8 1/16 1/16 1/16 ¼ 1/16 1/8]	JUNE-2014	7

REFERENCCE

BOOK	AUTHOR	PRIORITY
Principle of communication systems	Taub and schilling	1
Modern analog and Digital Communication	Lathi	2