INFORMATION THEORY

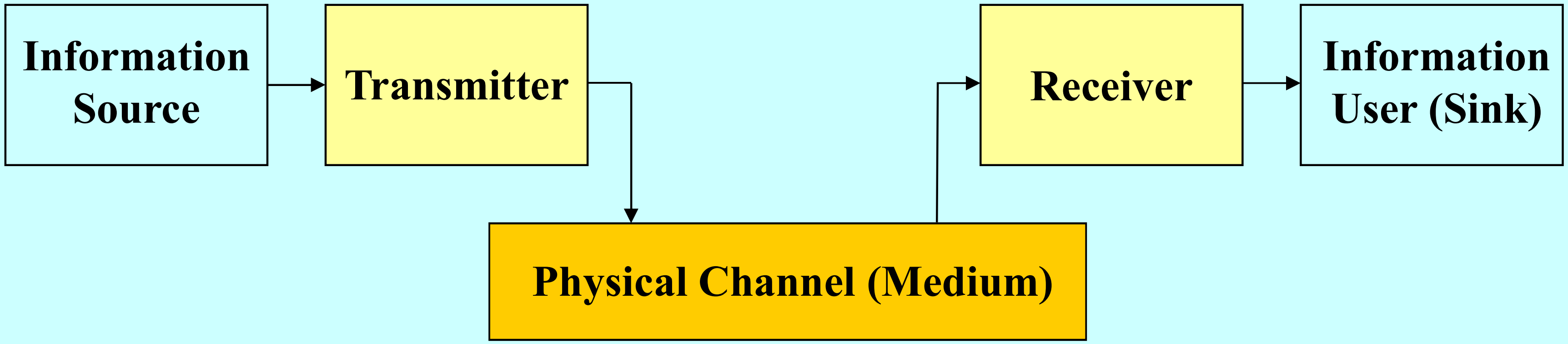
and

CODING

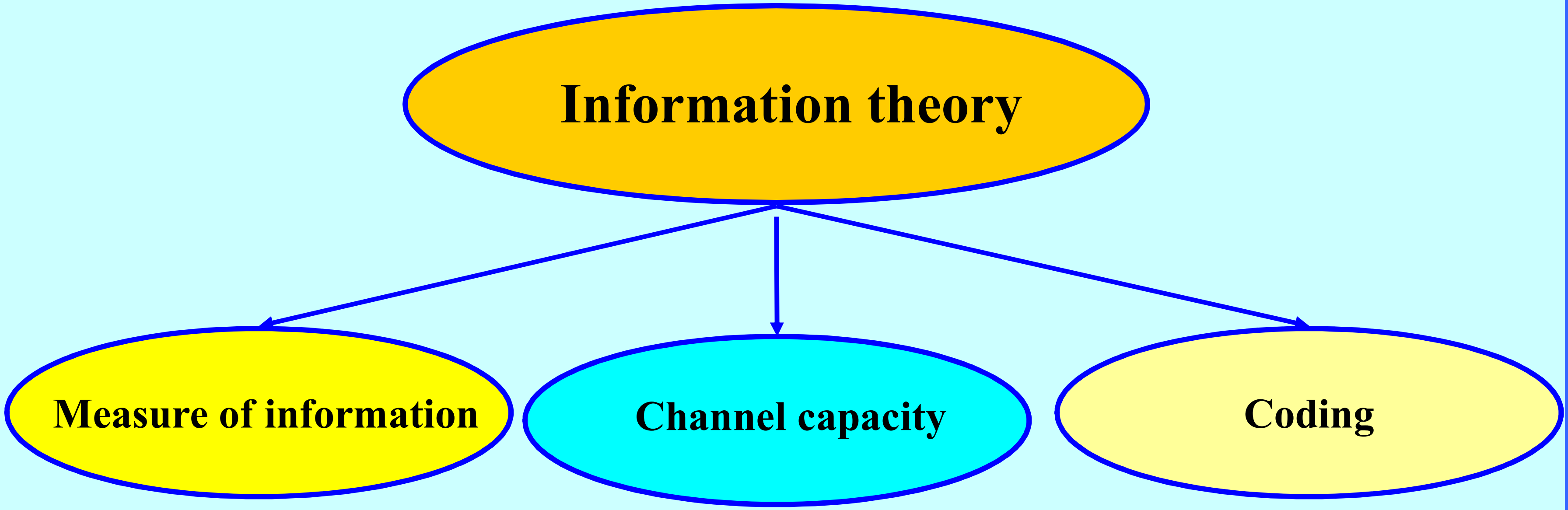
Husam Abduldaem Mohammed

Lecture 1

General Models of Information Transmission



INFORMATION THEORY



The purpose of a communication system is to carry **information**-bearing baseband signals from one place to another over a communication channel. with high efficiency and reliability.

Information theory provides a quantitative measure of the information contained in message signal and allows us to determine the capacity of a communication system to transfer this information from source to destination. Through the use of coding, a major topics of information theory, redundancy can be reduced from message signal so that channels can be used with improved efficiency. In addition systematic redundancy can be introduced to the transmitted signal so that channels can be used with improved reliability.

Examples:

Baseband signals: Audio signals and video signals

Channel: Optical fiber and free space

•"**Information**" in communication theory refers to the amount of uncertainty in a system that a message will get rid of.

Information Theory deals with mathematical modeling and analysis of a communication system rather than with physical sources and physical channels.

Information Sources. Amount of Information. Entropy

Mathematical Model of the Information Source. Discrete and Continuous Sources

Discrete memoryless source(DMS): is one for which each symbol produced is independent of the previous symbol and can be characterized by the list of symbols, the probability assignment to these symbols, and the specification of the rate of generating these symbols by the source.

Discrete ensemble is completely described by its set (ensemble) of possible messages $X=\{x_1, x_2, \ldots\}$ along with their probabilities $p(x_i)$, $i=1, 2, \ldots$. **Norming condition:**

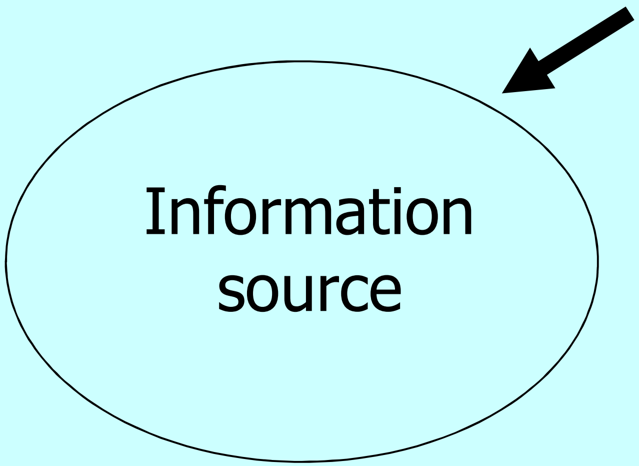
$$\sum_{i=1}^{M} p(x_i) = 1 \text{ or } \sum_{x \in X} p(x) = 1$$

Amount of Information

the amount of information received from a message is inversely related to the probability of its occurrence.

Source alphabet of M symbols

$\{x_1 , x_2 , x_3$3 ,iii $x_n\}$



Let the probability of producing x_i be for

$P(x_i) = P_i \qquad for \qquad P(x_i) >= 0$

Messages containing knowledge of high probability of occurrence convey relatively little information.

Those messages containing knowledge of a low probability of occurrence convey relatively large amount of information.

$P \uparrow \quad I \downarrow \qquad\qquad P \downarrow \quad I \uparrow$

Information axioms (requirements to an information measure):

1. Amount of information in a message x depends only on its probability:

$$I(x) = f(p(x)), \forall x \in X.$$

2. Information nonnegativity:

$$I(x) \geq 0, \ \forall x \in X,$$

with $I(x)$=0 only for the certain event.

3. Additivity for independent messages(Information from 2 Independent source will be the sum of the information from each separately).

$$I(x, y) = I(x) + I(y).$$

The only function of probability meeting these requirements is logarithm, thus

$$I(x) = -\log_b p(x) = \log_b \frac{1}{p(x)} (bit)$$

Standard convention of information theory is to take b=2

1 bit is the amount of information needed to choose between two equally likely alternatives.

So:

$$I(x) = -\log_2 p(x) = \log_2 \frac{1}{p(x)} (bit)$$

$$\log_2(x) = \frac{\ln(x)}{\ln(2)} = \frac{\log_{10}(x)}{\log_{10}(2)}$$

Example: Consider a random experiment with 16 equally likely outcomes. The information associated with each outcome is

$$I(x_j) = -\log_2(1/16) = \log_2(16) = 4 bits$$

Where j ranges from 1 to 16. The information is greater than one bit since the probability of each outcome is less than half.

Example: Shannon information content of the outcomes a-z.

For:

- x = z has a Shannon information content of 10.4 bits.
- x = e has an information content of 3.5 bits.

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
	$\sum_i p_i \log_2 \frac{1}{p_i}$		4.1

The total information contained in the message set is the sum of all individual messages:

$$I_t = -\sum_{k=1}^{m} N p_k \log \frac{1}{p_k}.$$

Where N is the total number of messages sequence.

Entropy and information rate

Self information is defined in terms of the individual messages or symbols a source may produce. A communication system is not designed around a particular message but rather all possible messages.

In practice, a source may not produce symbols with equal probabilities. Then each symbol will carry different amount of information.

Need to know on average how much information is transmitted per symbol in a *M* - symbols message ? Or, the average information content of symbols in a long message.

Entropy

Consider a source emits one of *M* possible symbols in a statistically independent manner with probabilities , respectively. (i.e. occurrence of a symbol at any given time does not influence the symbol emitted any other time).
Assume that a total of *M* symbols were emitted, self information of the i^{th} possible symbol is

$$I(x_i) = -\log_2(p(x_i))bits$$

Entropy of a discrete source is average amount of information in its messages (i.e bit per symbol):

$$H(X) = \overline{I(X)} = -\sum_{x \in X} p(x)\log_2 p(x) = \sum_{x \in X} p(x)\log_2 \frac{1}{p(x)}.$$

Note that *entropy* is a function of the *distribution* of X.

Entropy properties:

1. Nonnegativity:

$$H(X) \geq 0,$$

with zero value only for the deterministic (noninformative) source.

2. Entropy is limited from above as

$$H(X) \le \log_2 M$$

where equality takes place for only source of *M* equiprobable messages $p_i = 1/M$.

3. Additivity for independent sources:

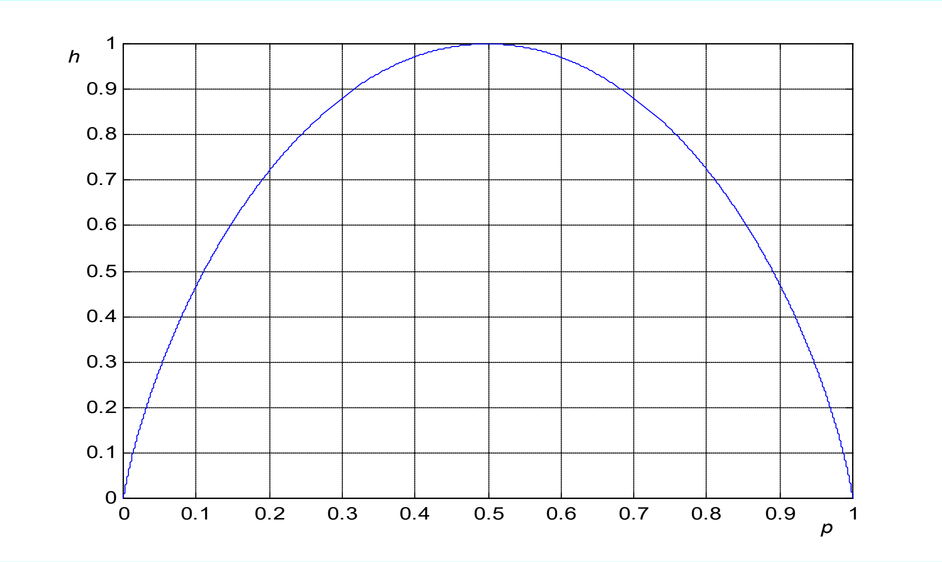
$$H(XY) = H(X) + H(Y).$$

- By analogy, if information is “disorderly”, there is less knowledge, or disorder is equivalent to unpredictability [in physics, a lack of knowledge about the positions and velocities of particles]
- Thus, entropy was also a measure of “order”: increase in entropy = decrease of order

Example:Binary source (ensemble) entropy (*p* is probability of one of two messages):

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

Binary Ensemble Entropy Plot



Homework: When considering an M message set it can be shown that the maximum entropy is possible when these messages are likely to occur equally. PROVE**.**

- Suppose we have a source that emits one of m possible symbols, in statistically independent sequence. Now, in a long message containing N symbols, the symbol x_i will occur on the average $p_i N$ times.
- Treading individual symbols as messages of length one, we have define the information content of the ith symbol as $\log_2(1/p_i)$ bits.

Hence, the $p_i N$ occurrences of s_i contributes an information of

$\mathbf{p_i N \log_2(1/p_i)}$ bits.

- The total information content of the message is then the sum of the contribution due to each of the m symbols of the source alphabet and is given by

$$I_{total} = \sum_{i=1}^{m} N p_i \log_2(1/p_i) \quad bits$$

- The average information per symbol is

$$H = \frac{I_{total}}{N} \sum_{i=1}^{m} p_i \log_2(1/p_i) \quad bits\ /\ symbol$$

Entropy

Entropy	**Number of possible messages**	**Amount of Uncertainty**	**Amount of Information in message**
Low Entropy	One message out of 10	Small uncertainty	Small amount of information
High Entropy	One message out of 1,000,000	Large uncertainty	Large amount of information

Entropy rate (Information rate) R:

$$R = H r_s (bit/s)$$

r_s = symbol rate in symbol / sec.

Example: a discrete source emits one of five symbols once every millisecond. The symbol probabilities are 0.5, 0.25, 0.125, 0.0625, and 0.0625, respectively. Find the source entropy and information rate.

Solution:

$$H(X) = \sum_{i=1}^{5} p(x_i) \log_2 \frac{1}{p(x_i)}.$$

$H = 0.5 \log_2(2) + 0.25\log_2(4) + 0.125\log_2(8) + 0.125 \log_2(16)$

$= 0.5 + 0.5 + 0.375 + 0.5$

$= 1.875$ bit per symbol

Information rate $R = r_s H = (1000)(1.875) = 1875$ bps

Example: calculate the entropy and entropy rate of a telegraph source having $p_{dot}=2/3$, $T_{dot}= 0.2$ s, $p_{dash}=1/3$, and $T_{dash}=0.4$s?

Solution:

$$H = \sum_{i=1}^{2} p(x_i)\log_2 \frac{1}{p(x_i)}$$

$$H = \frac{2}{3}\log_2 \frac{3}{2} + \frac{1}{3}\log_2 3$$

H=0.923 bit/symbol

$$\bar{\tau} = \sum_{i=1}^{2} p(x_i)\tau_i = \frac{2}{3}\otimes 0.2 \oplus \frac{1}{3}\otimes 0.4 = 0.266s$$

$$R = \frac{H}{\overline{\tau}} = \frac{0.923}{0.266} = 3.47 \; {bit}/{s}$$

Homework:

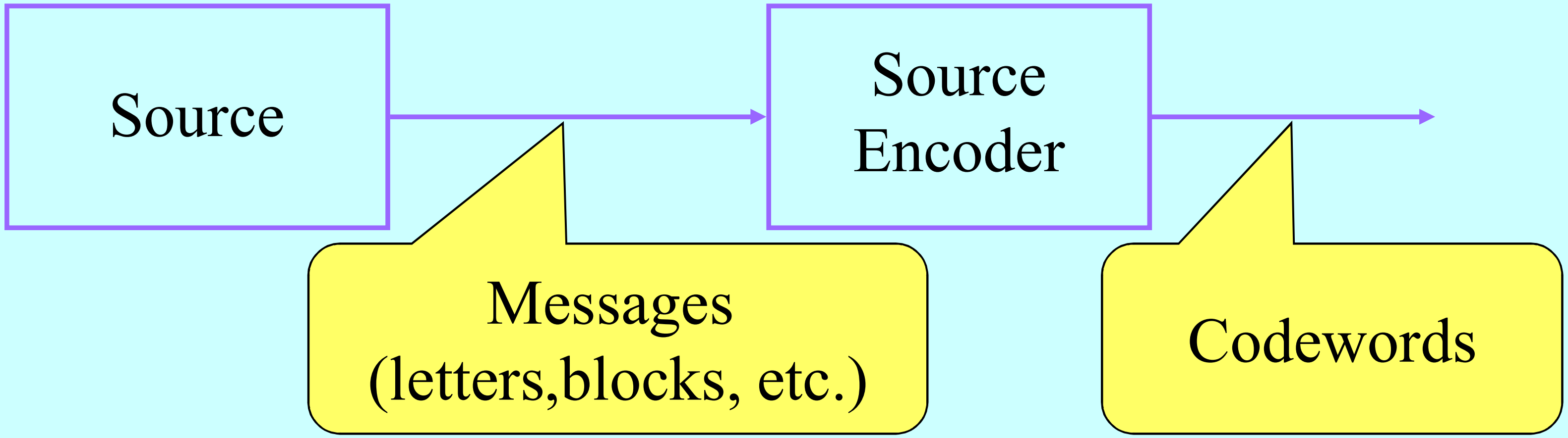
1) Consider a PCM system whose input is the continuous signal x(t) with bandwidth of w=50 Hz. Suppose x(t) is sampled the Nyquist rate $f_s=2w$, and let there be four quantum levels such the quantized value s have equally likely probabilities. Identifying each possible quantized value as a symbol, the output of the quantizer source with m=4. Find the entropy and entropy rate of the system.

2) consider the PCM source in example (1). Now, let the quantized values have probabilities such that 1/2, 1/4, 1/8 and 1/8.

Source Coding

Some Definitions

Under source coding M messages of a source ensemble are mapped onto M **codewords** or **code vectors** whose elements are taken from some **alphabet**. **Code** itself is just a set of codewords. If all codewords have the same number of symbols (length) the code is called **uniform** or **fixed-length**. The simplest uniform binary code is just message number written in a binary number system (example: ASCII code). Code whose words have different lengths is called **nonuniform** (**variable-length**, **run-length**).



The target of source coding is the most economical representation of messages, i.e. codewords are desirable to be as short as possible.

Let M is the number of symbols.then

$N = [\ \log_2 L] + 1.$ L is not a power of 2.

$N = \log_2 L.$ L is not a power of 2

The efficiency of a fixed length code = H / N

Example: A DMS has output alphabet of five letters x_i where I = 1,2,3,4,5 each with probabilirty of 0.2. Evaluate the efficiency of a fixed length code in which:

i) each letter is encoded separately into a binary sequence.

ii) two letters at a time are encoded into a binary sequence.

Solution:

i) $N = [\log_2 L] + 1$

$= [2.32] + 1 = 3$ bit

$H = 2.32$

Efficiency = H/ N = 2.32 / 3 = 0.77

ii) L is not a power of two

$L = 5^2 = 25$

$N = [\log_2 L] + 1 = [\log_2 25] + 1$

$= [4.6] + 1 = 5$ bit

Efficiency = H/ N = 2.32 / 2.5 = 0.93

Shannon-Fano Code

At the first step of this algorithm the whole source ensemble is divided into two subsets whose total probabilities are as close to each other (and, hence, to 1/2) as possible. All messages of one of the subsets are assigned symbol 0 as the first codeword element while all messages of the other are assigned symbol 1. At the second step the same procedure is repeated with each of the subsets obtained at the first step and the second code symbol is obtained. As soon as some message proves the only in a subset it is encoded completely. We proceed this way until all messages are encoded.

Example 2.3.1. M=8 and probabilities of messages are given in table.

X	$p(x)$	1	2	3	4	5
x_1	0.50	0				
x_2	0.20	1	0	0		
x_3	0.05	1	0	1		
x_4	0.15	1	1	0		
x_5	0.04	1	1	1	0	0
x_6	0.03	1	1	1	1	0
x_7	0.02	1	1	1	1	1
x_8	0.01	1	1	1	0	1

Source entropy: $H(X) \approx 2.11$ bit, $\log M = 3$, **average length:**

$$\overline{n} = 0.5 \cdot 1 + (0.2 + 0.05 + 015) \cdot 3 + (0.04 + 0.03 + 0.02 + 0.01) \cdot 5 = 2.2.$$

Huffman Code

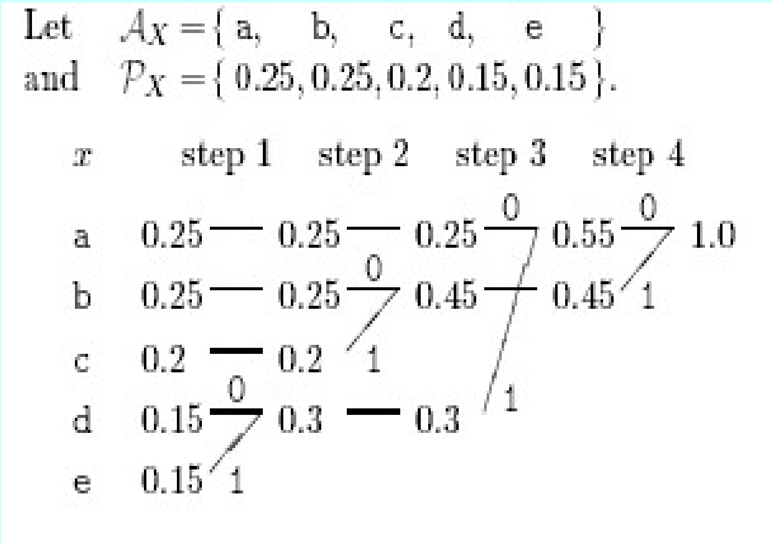
This code is an optimal one, i.e. no other nonuniform prefix code can have smaller average codeword length. At the first step of Huffman coding two least probable messages are combined into one compound message whose probability equals sum of probabilities of initial messages. One of these two initial messages is assigned code symbol 0 while the other obtains 1. At the second step new ensemble consists of M-1 messages and again two least probable messages are combined into one with assignment of symbol 0 to one of them and 1 to the other. The procedure goes on until only two messages are left whereupon one of them is assigned symbol 0 and the other symbol 1. As a result a **code tree** is built up which is read out from right to left to obtain codewords for all initial messages.

It is advisable but not necessary to list all messages in a nonincreasing order of their probabilities before starting to run algorithm. It is noteworthy also that contrary to Shannon-Fano algorithm, in a progress of Huffman coding any codeword appears in a reverse order.

We now present a beautifully simple algorithm for finding an optimal prefix code. The trick is to construct the code backwards starting from the tails of the codewords; we build the binary tree from its leaves.

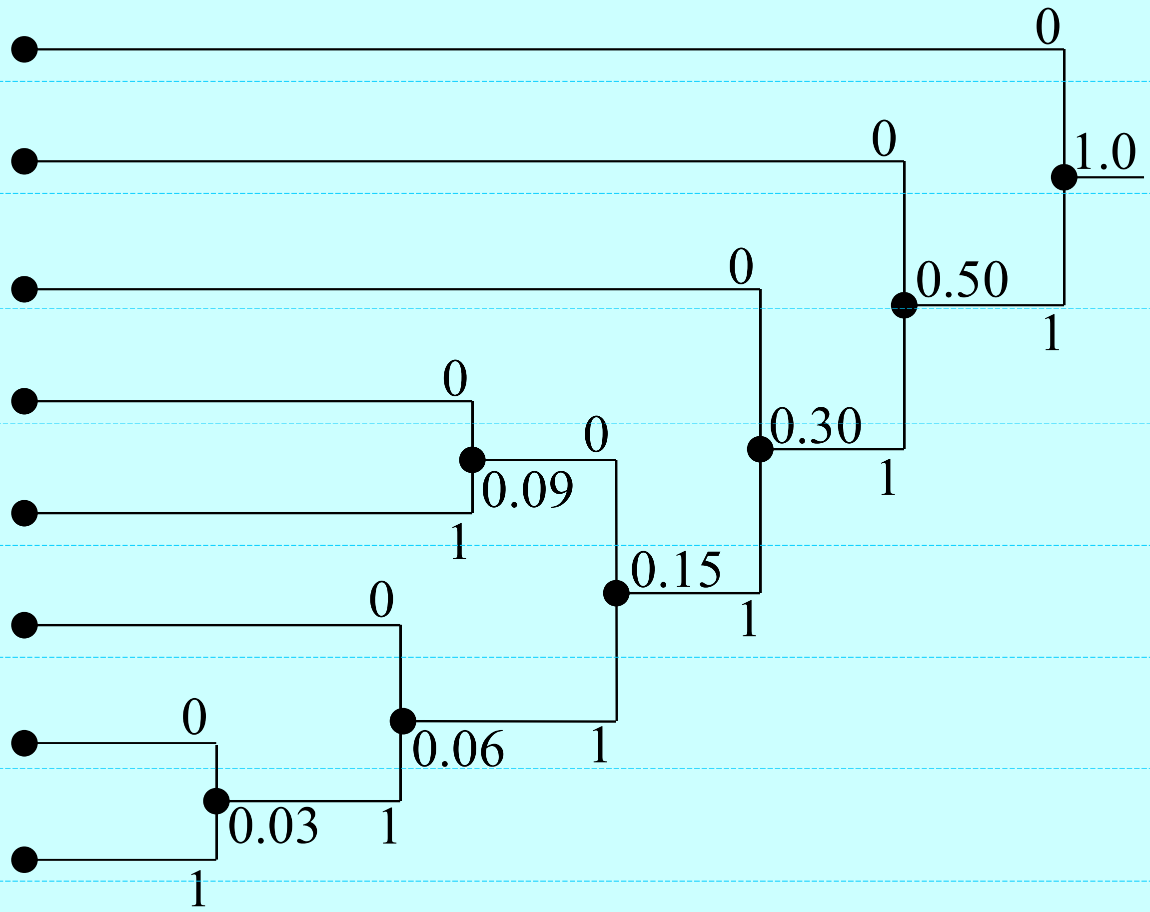
1. Take the two least probable symbols in the alphabet. These two symbols will be given the longest codewords, which will have equal length, and differ only in the last digit.
2. Combine these two symbols into a single symbol, and repeat.

Example:



Example 2.4.1. Encode the same ensemble as in Example:

X	$p(x)$	**Codewords**
x_1	0.50	0
x_2	0.20	10
x_3	0.15	110
x_4	0.05	11100
x_5	0.04	11101
x_6	0.03	11110
x_7	0.02	111110
x_8	0.01	111111



Average length: $\overline{n} = 1 \cdot 0.5 + 2 \cdot 0.2 + 3 \cdot 0.15$

$$+ 5 \cdot (0.05 + 0.04 + 0.03) + 6 \cdot (0.02 + 0.01) = 2.13$$

Homwork:

•Prove that there is no better symbol code for a source than the Huffman code.

• Find the optimal binary symbol code for the ensemble:

AX = { a ; b ; c ; d ; e ; f ; g}

PX = { 0:01; 0:24; 0:05; 0:20; 0:47; 0:01; 0:02 }

then find the efficiency of the code.

• Write a program to implement the Huffman and Shannon Fano algorithms

Disadvantages of the Huffman code

The Huffman algorithm produces an optimal symbol code for an ensemble, but this is not the end of the story. Both the word `ensemble' and the phrase symbol code need careful attention.

•**Changing ensemble**

•**The extra bit**

•**Beyond symbol codes**