



## On fuzzy information theory

A. Zarandi

*Department of Mathematics & Computer, Islamic Azad university Kerman branch, Kerman, Iran*  
afshin\_zarandi@yahoo.com

### Abstract

Information theory is generally considered to have been founded in 1948 by Claude Shannon. In this paper we introduce the concept of fuzzy information theory using the notion of fuzzy sets. We study it and give some example of it.

**Keywords:** Information theory, fuzzy set, fuzzy probability, entropy.

### Introduction

The main concepts of information theory can be grasped by considering the most widespread means of human communication: language. Two important aspects of a concise language are as follows: First, the most common words (e.g., "a", "the" & "I") should be shorter than less common words (e.g., "benefit", "generation" & "mediocre"), so that sentences will not be too long. Such a tradeoff in word length is analogous to data compression and is the essential aspect of source coding. Second if part of a sentence is unheard or misheard due to noise—e.g., a passing car—the listener should still be able to glean the meaning of the underlying message. Such robustness is as essential for an electronic communication system as it is for a language; properly building such robustness into communications is done by channel coding. Source coding and channel coding are the fundamental concerns of information theory (Rieke *et al.*, 1997, Burnham & Anderson, 2002).

Note that these concerns have nothing to do with the importance of messages. For e.g., a platitude such as "Thank you; come again" takes about as long to say or write as the urgent plea, "Call an ambulance!" while the latter may be more important and more meaningful in many contexts. Information theory, however, does not consider message importance or meaning, as these are matters of the quality of data rather than the quantity and readability of data, the latter of which is determined solely by probabilities (Anderson, 2003).

Information theory is generally considered to have been founded in 1948 by Claude Shannon in his seminal work, "A mathematical theory of communication." (Shannon, 1948; Shannon & Warren Weaver, 1949) The central paradigm of classical information theory is the engineering problem of the transmission of information over a noisy channel. The most fundamental results of this theory are Shannon's source coding theorem which establishes that on average the number of bits needed to represent the result of an uncertain event is given by its entropy; and Shannon's noisy-channel coding theorem which states that reliable communication is possible over noisy channels provided that the rate of communication is below a certain threshold called the channel capacity.

The channel capacity can be approached in practice by using appropriate encoding and decoding systems.

Information theory is closely associated with a collection of pure and applied disciplines that have been investigated and reduced to engineering practice under a variety of rubrics throughout the world over the past half century or more: adaptive systems, anticipatory systems, artificial intelligence, complex systems, complexity science, cybernetics, informatics, machine learning, along with systems sciences of many descriptions. Information theory is a broad and deep mathematical theory, with equally broad and deep applications, amongst which is the vital field of coding theory (Gibson, 1998).

A branch of communication theory devoted to problems in coding. A unique feature of information theory is its use of a numerical measure of the amount of information gained when the contents of a message are learned. Information theory relies heavily on the mathematical science of probability. For this reason the term information theory is often applied loosely to other probabilistic studies in communication theory, such as signal detection, random noise, and prediction. See also electrical communications; Probability (Gallager, 1968; Yeung, 2002).

In designing a one-way communication system from the standpoint of information theory, three parts are considered beyond the control of the system designer:

- (1) the source, which generates messages at the transmitting end of the system,
- (2) the destination, which ultimately receives the messages, and
- (3) the channel, consisting of a transmission medium or device for conveying signals from the source to the destination.

The source does not usually produce messages in a form acceptable as input by the channel. The transmitting end of the system contains another device, called an encoder, which prepares the source's messages for input to the channel. Similarly the receiving end of the system will contain a decoder to convert the output of the channel into a form that is recognizable by the destination. The encoder and the decoder are the parts to be designed. In

radio systems this design is essentially the choice of a modulator and a detector. See also Modulation.

A source is called discrete if its messages are sequences of elements (letters) taken from an enumerable set of possibilities (alphabet). Thus sources producing integer data or written English are discrete. Sources which are not discrete are called continuous, for example, speech and music sources. The treatment of continuous cases is sometimes simplified by noting that signal of finite bandwidth can be encoded into a discrete sequence of numbers.

The output of a channel need not agree with its input. For e.g., a channel might, for secrecy purposes, contain a cryptographic device to scramble the message. Still, if the output of the channel can be computed knowing just the input message, then the channel is called noiseless. If, however, random agents make the output unpredictable even when the input is known, then the channel is called noisy (Reza, 1994).

Many encoders first break the message into a sequence of elementary blocks; next they substitute for each block a representative code, or signal, suitable for input to the channel. Such encoders are called block encoders. For e.g., telegraph and teletype systems both use block encoders in which the blocks are individual letters. Entire words form the blocks of some commercial cablegram systems. It is generally impossible for a decoder to reconstruct with certainty a message received via a noisy channel. Suitable encoding, however, may make the noise tolerable (Csiszar & Janos, 1997).

Even when the channel is noiseless a variety of encoding schemes exists and there is a problem of picking a good one. Of all encodings of English letters into dots and dashes, the Continental Morse encoding is nearly the fastest possible one. It achieves its speed by associating short codes with the most common letters. A noiseless binary channel (capable of transmitting two kinds of pulse 0, 1, of the same duration) provides the following e.g. suppose one had to encode English text for this channel. A simple encoding might just use 27 different five-digit codes to represent word space (denoted by #), A, B, ..., Z; say # 00000, A 00001, B 00010, C 00011, ..., Z 11011. The word #CAB would then be encoded into 00000000110000100010. A similar encoding is used in teletype transmission; however, it places a third kind of pulse at the beginning of each code to help the decoder stay in synchronism with the encoder (Goldman, 2005).

Information theory is based on probability theory and statistics. The most important quantities of information are entropy, the information in a random variable and mutual information, the amount of information in common between two random variables. The former quantity indicates how easily message data can be compressed while the latter can be used to find the communication rate across a channel.

The choice of logarithmic base in the following formulae determines the unit of information entropy that is used. The most common unit of information is the bit, based on the binary logarithm. Other units include the nat, which is based on the natural logarithm, and the hartley, which is based on the common logarithm. In what follows, an expression of the form is considered by convention to be equal to zero whenever  $p = 0$ . This is justified because for any logarithmic base (Jaynes 1957, Mackay 2003).

*Entropy*

Entropy of a Bernoulli trial as a function of success probability often called the binary entropy function  $H_b(p)$ . The entropy is maximized at 1 bit per trial when the two possible outcomes are equally probable, as in an unbiased coin toss (Kolmogorov, 1968).

The entropy, H, of a discrete random variable X is a measure of the amount of uncertainty associated with the value of X. Suppose one transmits 1000 bits (0s & 1s). If these bits are known ahead of transmission (to be a certain value with absolute probability), logic dictates that no information has been transmitted. If, however, each is equally and independently likely to be 0 or 1, 1000 bits (in the information theoretic sense) have been transmitted. Between these two extremes, information can be quantified as follows. If is the set of all messages  $\{x_1, \dots, x_n\}$  that X could be, and  $p(x)$  is the probability of X given some, then the entropy of X is defined:

$$H(X) = - \sum_{x \in X} p(x) \log p(x)$$

Here,  $I(x)$  is the self-information, which is the entropy contribution of an individual message and is the expected value.

An important property of entropy is that it is maximized when all the messages in the message space are equiprobable  $p(x) = 1 / n$ , i.e., most unpredictable—in which case  $H(X) = \log n$ .

The special case of information entropy for a random variable with two outcomes is the binary entropy function, usually taken to the logarithmic base 2. To recapitulate, we assume the following four conditions as axioms:

- 1-  $H(1/M, 1/M, \dots, 1/M) = f(M)$  is a monotonically increasing function of M (M=1, 2, ...).
- 2-  $f(ML) = f(M) + f(L)$  (M, L=1, 2, ...)
- 3-

$$H(p_1, p_2, \dots, p_M) = H(p_1 + p_2 + \dots + p_r, p_{r+1} + \dots + p_M) + (p_1 + \dots + p_r) H\left(\frac{p_1}{\sum_{i=1}^r p_i}, \dots, \frac{p_r}{\sum_{i=1}^r p_i}\right) + (p_{r+1} + \dots + p_M) H\left(\frac{p_{r+1}}{\sum_{i=r+1}^M p_i}, \dots, \frac{p_M}{\sum_{i=r+1}^M p_i}\right)$$

4-  $H(p, 1-p)$  is a continuous function of  $p$ .

The four axioms essentially determine the uncertainty measure. More precisely, we have the following theorem.

*Theorem 1.* (Yeung 2008)

The only function satisfying the four given axioms is

$$H(p_1, \dots, p_M) = -C \sum_{i=1}^M p_i \log p_i$$

Where,  $C$  is an arbitrary positive number and the logarithm base is any number greater than 1. Unless otherwise specified, we shall assume  $C=1$  and take logarithms to the base 2. The units of  $H$  are sometimes called bits (a contraction of binary digits). Thus the units are chosen so that there is one bit of uncertainty associated with the toss of an unbiased coin. Biasing the coin tends to decrease the uncertainty. We remark in passing that the average uncertainty of a random variable  $X$  does not depend on the values the random variable assumes or on anything else except the probabilities associated with those values. The average uncertainty associated with the toss of an unbiased coin is not changed by adding the condition that the experimenter will be shot if the coin comes up tails.

*Joint entropy*

The joint entropy of two discrete random variables  $X$  and  $Y$  is merely the entropy of their pairing:  $(X, Y)$ . This implies that if  $X$  and  $Y$  are independent, then their joint entropy is the sum of their individual entropies. For e.g, if  $(X, Y)$  represents the position of a chess piece –  $X$  the row and  $Y$  the column, then the joint entropy of the row of the piece and the column of the piece will be the entropy of the position of the piece.

$$H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y)$$

Despite similar notation, joint entropy should not be confused with cross entropy.

*Conditional entropy (equivocation)*

The conditional entropy or conditional uncertainty of  $X$  given random variable  $Y$  (also called the equivocation of  $X$  about  $Y$ ) is the average conditional entropy over  $Y$ :

$$H(X|Y) = - \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(y)}$$

Because entropy can be conditioned on a random variable or on that random variable being a certain value, care should be taken not to confuse these two definitions of conditional entropy, the former of which is in more common use. A basic property of this form of conditional entropy is that (Gallager 1968):

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

*Mutual information (transinformation)*

Mutual information measures the amount of information that can be obtained about one random variable by observing another. It is important in communication where it can be used to maximize the amount of information shared between sent and received signals. The mutual information of  $X$  relative to  $Y$  is given by:

$$I(X, Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(y)p(x)}$$

Where, SI (Specific mutual Information) is the point wise mutual information.

A basic property of the mutual information is that

$$I(X, Y) = H(X) - H(X|Y)$$

That is, knowing  $Y$ , we can save an average of  $I(X; Y)$  bits in encoding  $X$  compared to not knowing  $Y$ .

Mutual information is symmetric:

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

Mutual information can be expressed as the average Kullback-Leibler divergence (information gain) of the posterior probability distribution of  $X$  given the value of  $Y$  to the prior distribution on  $X$ .

$$I(X; Y) = E_{p(y)} [D_{KL}(p(X|Y=y) || p(X))]$$

In other words, this is a measure of how much, on the average, the probability distribution on  $X$  will change if we are given the value of  $Y$ . This is often recalculated as the divergence from the product of the marginal distributions to the actual joint distribution:

$$I(X; Y) = D_{KL}(p(X, Y) || p(X)p(Y))$$

Mutual information is closely related to the log-likelihood ratio test in the context of contingency tables and the multinomial distribution and to Pearson's  $\chi^2$  test: mutual information can be considered a statistic for assessing independence between a pair of variables, and has a well-specified asymptotic distribution.

*Kullback-Leibler divergence (information gain)*

The Kullback-Leibler divergence (or information divergence, information gain, or relative entropy) is a way of comparing two distributions: a "true" probability distribution  $p(X)$ , and an arbitrary probability distribution  $q(X)$ . If we compress data in a manner that assumes  $q(X)$  is the distribution underlying some data, when, in reality,  $p(X)$  is the correct distribution, the Kullback-Leibler divergence is the number of average additional bits per datum necessary for compression. It is thus defined

$$D_{KL}(p(X) || q(Y)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Although it is sometimes used as a 'distance metric', it is not a true metric since it is not symmetric and does not

satisfy the triangle inequality (making it a semi-quasimetric) (Mansuripur, 1987).

*Other quantities*

Other important information theoretic quantities include Rényi entropy, (a generalization of entropy,) differential entropy, (a generalization of quantities of information to continuous distributions,) and the conditional mutual information.

*Coding theory*

Coding theory is one of the most important and direct applications of information theory. It can be subdivided into source coding theory and channel coding theory. Using a statistical description for data, information theory quantifies the number of bits needed to describe the data, which is the information entropy of the source.

Data compression (source coding): There are two formulations for the compression problem: lossless data compression: the data must be reconstructed exactly; lossy data compression: allocates bits needed to reconstruct the data, within a specified fidelity level measured by a distortion function. This subset of Information theory is called rate-distortion theory. Error-correcting codes (channel coding): While data compression removes as much redundancy as possible, an error correcting code adds just the right kind of redundancy (i.e., error correction) needed to transmit the data efficiently and faithfully across a noisy channel.

This division of coding theory into compression and transmission is justified by the information transmission theorems, or source-channel separation theorems that justify the use of bits as the universal currency for information in many contexts. However, these theorems only hold in the situation where one transmitting user wishes to communicate to one receiving user. In scenarios with more than one transmitter (the multiple-access channel), more than one receiver (the broadcast channel) or intermediary "helpers" (the relay channel), or more general networks, compression followed by transmission may no longer be optimal. Network information theory refers to these multi-agent communication models.

*Source theory*

Any process that generates successive messages can be considered a source of information. A memory-less source is one in which each message is an independent identically-distributed random variable, whereas the properties of ergodicity and stationarity impose more general constraints. All such sources are stochastic. These terms are well studied in their own right outside information theory.

*Rate*

Information rate is the average entropy per symbol. For memory-less sources, this is merely the entropy of

each symbol, while, in the case of a stationary stochastic process, it is

$$r = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, X_{n-2}, X_{n-3}, \dots)$$

that is, the conditional entropy of a symbol given all the previous symbols generated. For the more general case of a process that is not necessarily stationary, the average rate is

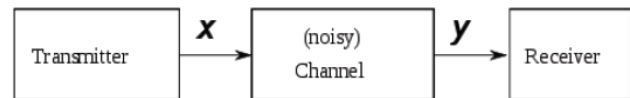
$$r = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, X_3, X_4, \dots)$$

that is, the limit of the joint entropy per symbol. For stationary sources, these two expressions give the same result. It is common in information theory to speak of the "rate" or "entropy" of a language. This is appropriate, for example, when the source of information is English prose. The rate of a source of information is related to its redundancy and how well it can be compressed, the subject of source coding (Arndt, 2004).

*Channel capacity*

Communications over a channel—such as an ethernet wire is the primary motivation of information theory. As anyone who's ever used a telephone (mobile or landline) knows, however, such channels often fail to produce exact reconstruction of a signal; noise, periods of silence, and other forms of signal corruption often degrade quality. How much information can one hope to communicate over a noisy (or otherwise imperfect) channel?

Consider the communications process over a discrete channel. A simple model of the process is shown below:



Here X represents the space of messages transmitted and Y the space of messages received during a unit time over our channel. Let  $p(y | x)$  be the conditional probability distribution function of Y given X. We will consider  $p(y | x)$  to be an inherent fixed property of our communications channel (representing the nature of the noise of our channel). Then the joint distribution of X and Y is completely determined by our channel and by our choice of  $f(x)$ , the marginal distribution of messages we choose to send over the channel. Under these constraints, we would like to maximize the rate of information, or the signal, we can communicate over the channel. The appropriate measure for this is the mutual information, and this maximum mutual information is called the channel capacity and is given by:

$$C = \max_f I(X;Y)$$

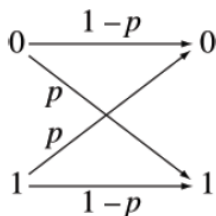
This capacity has the following property related to communicating at information rate R (where R is usually bits per symbol). For any information rate  $R < C$  and coding error  $\epsilon > 0$ , for large enough N, there exists a code of length N and rate  $\geq R$  and a decoding algorithm, such that the maximal probability of block error is  $\leq \epsilon$ ; that is, it

is always possible to transmit with arbitrarily small block error. In addition, for any rate  $R > C$ , it is impossible to transmit with arbitrarily small block error.

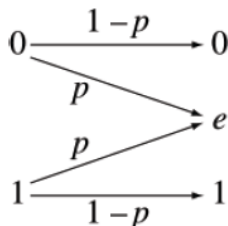
Channel coding is concerned with finding such nearly optimal codes that can be used to transmit data over a noisy channel with a small coding error at a rate near the channel capacity.

*Capacity of particular channel models*

A continuous-time analog communications channel subject to Gaussian noise—see Shannon-Hartley theorem. A binary symmetric channel (BSC) with crossover probability  $p$  is a binary input, binary output channel that flips the input bit with probability  $p$ . The BSC has a capacity of  $1 - H_b(p)$  bits per channel use, where  $H_b$  is the binary entropy function to the base 2 logarithm:



A binary erasure channel (BEC) with erasure probability  $p$  is a binary input, ternary output channel. The possible channel outputs are 0, 1, and a third symbol 'e' called an erasure. The erasure represents complete loss of information about an input bit. The capacity of the BEC is  $1 - p$  bits per channel use.



*Applications to other fields*

*Intelligence uses and secrecy applications:* Information theoretic concepts apply to cryptography and cryptanalysis. Turing's information unit, the ban, was used in the Ultra project, breaking the German Enigma machine code and hastening the end of WWII in Europe. Shannon himself defined an important concept now called the unicity distance. Based on the redundancy of the plaintext, it attempts to give a minimum amount of ciphertext necessary to ensure unique decipherability. Information theory leads us to believe it is much more difficult to keep secrets than it might first appear. A brute force attack can break systems based on asymmetric key algorithms or on most commonly used methods of symmetric key algorithms (sometimes called secret key algorithms), such as block ciphers. The security of all such methods currently comes from the assumption that

no known attack can break them in a practical amount of time.

Information theoretic security refers to methods such as the one-time pad that are not vulnerable to such brute force attacks. In such cases, the positive conditional mutual information between the plaintext and ciphertext (conditioned on the key) can ensure proper transmission, while the unconditional mutual information between the plaintext and ciphertext remains zero, resulting in absolutely secure communications. In other words, an eavesdropper would not be able to improve his or her guess of the plaintext by gaining knowledge of the ciphertext but not of the key. However, as in any other cryptographic system, care must be used to correctly apply even information-theoretically secure methods; the Venona project was able to crack the one-time pads of the Soviet Union due to their improper reuse of key material (Ash, 1990, Cover & Joy, 2006).

*Entropy by fuzzy probability*

In this section we use the fuzzy probability instead of probability in entropy.

*Definition 1.* Let  $\Omega$  be a sample space and  $P$  be a probability measure on  $\Omega$ . If  $\tilde{A}$  is a fuzzy event of  $\Omega$ , then probability of  $\tilde{A}$  is defined as

$$P(\tilde{A}) = \begin{cases} \int \tilde{A}(w) dP(w) & \text{if } \Omega \text{ is not discrete} \\ \sum_{w \in \Omega} \tilde{A}(w) P(w) & \text{if } \Omega \text{ is discrete} \end{cases}$$

*Example 2.* Let tossing a coin and  $\tilde{A}$  be a small number occurs and  $\tilde{B}$  be approximately number 5. Then the events are:

$$\tilde{A} = \left\{ \frac{1}{1}, \frac{0.8}{2}, \frac{0.5}{3}, \frac{0.2}{4}, \frac{0.1}{5} \right\}$$

$$\tilde{B} = \left\{ \frac{0.1}{1}, \frac{0.2}{3}, \frac{0.6}{4}, \frac{1}{5}, \frac{0.6}{6} \right\}.$$

In this case the probability of event  $\tilde{A}$  is

$$\begin{aligned} p(\tilde{A}) &= \sum_{t \in \Omega} \tilde{A}(t) P(t) \\ &= \tilde{A}(1)P(\{1\}) + \tilde{A}(2)P(\{2\}) + \dots + \tilde{A}(6)P(\{6\}) \\ &= 1 \times \frac{1}{6} + 0.8 \times \frac{1}{6} + 0.5 \times \frac{1}{6} + 0.2 \times \frac{1}{6} + 0.1 \times \frac{1}{6} \\ &= \frac{2.6}{6} = 0.433. \end{aligned}$$

and

$$p(\tilde{B}) = \sum_{t \in \Omega} \tilde{B}(t) P(t) = 0.416.$$

Therefore we can in above example the probability of a fuzzy event may be differ with the (crisp) probability of an

event since the fuzzy event was differed with (crisp) event.

In entropy formula  $H(p_1, \dots, p_M) = -C \sum_{i=1}^M p_i \log p_i$  we

consider the fuzzy probability instead of (crisp) probability for generalization of this notion and we introduce the fuzzy entropy. Therefore we can generalize all of the above notions and then we can introduce the fuzzy information theory.

*Definition 3.* Let  $p_i$  be fuzzy probability. Then we can use

the entropy formula  $H(p_1, \dots, p_M) = -C \sum_{i=1}^M p_i \log p_i$

and compute the entropy for the fuzzy event with fuzzy probability instead of (crisp) probability.

For convenience we put  $C=1$  and then we

have  $H(p_1, \dots, p_M) = -\sum_{i=1}^M p_i \log p_i$ .

*Example 4.* Consider the fuzzy events of Example 2; we can compute the entropy for these fuzzy events.

$$H(\tilde{A}, \tilde{B}) = -\sum_{i=1}^2 p_i \log p_i = -(0.433 \log(0.433) + 0.416 \log(0.416)) = 2.473$$

## Conclusion

In this paper we study the information theory, entropy, coding and applications. We introduced the entropy with fuzzy probability as a generalization of entropy with probability which is studied before. Since we live in fuzzy world and always we work with vague notions therefore for transferring the fuzzy events we must use another way since studied method was about crisp events. We hope that this notion is used for compactification the data for transformation and decrease the speed of networks. As a next work we can study this notion and develop this method for practical problem.

## Acknowledgement

The author would like to express his sincere thanks to the referees for their valuable suggestions and comments.

## References

1. Anderson DR (2003) Some background on why people in the empirical sciences may want to better understand the information-theoretic methods. <http://www.jyu.fi/science/laitokset/bioenv/en/coevolution/events/itms/why>. Retrieved on Dec. 30th, 2007.
2. Arndt Christoph (2004) Information measures, information and its description in science and engineering. *Signals and Communication Technology*, Springer Series.

3. Ash B. Robert (1990) Information theory. Interscience, NY; Dover, NY.
4. Burnham KP and Anderson DR (2002) Model selection and multimodel inference: A practical information-theoretic approach, 2nd Edition, Springer Science, NY.
5. Cover M. Thomas and Joy A. Thomas (2006) Elements of information theory. 2nd Edition, Wiley-Inter-science, NY.
6. Csiszar Imre and Janos Korner (1997) Information theory: coding theorems for discrete memoryless systems. Akademiai Kiado, 2nd edition.
7. David J and C. MacKay (2003) Information theory, inference, and learning algorithms. University Press, Cambridge.
8. Gallager Robert (1968) Information theory and reliable communication. John Wiley & Sons, NY.
9. Gibson Jerry D (1998) Digital Compression for Multimedia: Principles and Standards. Morgan Kaufmann.  
[http://books.google.com/books?id=aqQ2Ry6spu0C&pg=PA56&dq=entropy+ate+conditional&as\\_brr=3&ei=YGDsRtzGGKjupQKa2L2xDw&sig=o0UCtf0xZOof11IPlexPrjOKPgNc#PPA57](http://books.google.com/books?id=aqQ2Ry6spu0C&pg=PA56&dq=entropy+ate+conditional&as_brr=3&ei=YGDsRtzGGKjupQKa2L2xDw&sig=o0UCtf0xZOof11IPlexPrjOKPgNc#PPA57), M1.
10. Goldman Stanford (2005) Information theory. Dover, NY.
11. Jaynes ET (1957) Information theory and statistical mechanics. *Phys. Rev.* 106, 620.
12. Kolmogorov Andrey (1968) Three approaches to the quantitative definition of information. *Intl.J. Computer Mathematics. Problems of Information Transmission.* No. 1, 3-11.
13. Mansuripur Masud (1987) Introduction to information theory. Prentice Hall, NY.
14. Raymond W. Yeung (2008) Information theory and network coding. Springer.
15. Reza Fazlollah (1994) An introduction to information theory. Dover, NY.
16. Rieke F, Warland D, Ruyter van Steveninck R and Bialek Spikes W (1997) Exploring the neural code. The MIT press.
17. Shannon CE (1948) A Mathematical Theory of Communication. *Bell System Technical J.* 27, 379-423 & 623-656.
18. Shannon CE and Warren Weaver (1949) The Mathematical theory of communication. Univ of Illinois Press.
19. Yeung Raymond W (2002) A first course in information theory. Kluwer Academic/Plenum Publ.