FACTA UNIVERSITATIS (NIŠ)
Series: Electronics and Energetics vol. 11, No.2 (1998), 129-141

INFORMATION THEORY AFTER 50 YEARS

Dušan Drajić and Dragana Bajić

Abstract. In this paper an attempt is made to give a very short surway of the development of Information Theory. Also, some thoughts concerning the future of Information Theory are given.

1. Introduction

"Scientific theories deal with concepts, not with reality. All theoretical results are derived from certain axioms by deductive logic. In physical sciences the theories are so formulated as to correspond in some useful sense to the real world, whatever that may mean. However, this correspondence is approximate, and the physical justification of all theoretical conclusions is based on some form of inductive reasoning."

A. Papoulis: Probability, Random Variables
and Stochastic Processes (Preface)

In Communications it is especially important to have the above citation in mind. The "real world" in Communications (messages, electrical signals etc.) is not always easy to see and models must be made for the "invisible" things.

Some parallels can be drawn.

In 19th century human race started to use more and more energy. At the end of the century the corresponding theory was developed - Statistical Mechanics with the well known Second Law of Thermodynamics (where the notion of **entropy** was introduced, the entropy being regarded as a quantitative measure of order against disorder).

In 20th century the "information era" started. Now, we are living in the **Information Age**. In the middle of the century the corresponding

Manuscript received July 10, 1998.

The authors are with Faculty of Electrical Engineering, Bulevar revolucije 73, 11000 Beograd, Yugoslavia, E-mail: edrajic(ebajic)@ubbg.etf.bg.ac.yu.

mathematical theory of communication was brought by Claude Shannon (including also the quantity named entropy).

One more parallel: at the beginning of the century Einstein gave the relationship between mass, light velocity and energy. Shannon gave the relationship between the attainable information rate, frequency band and the signal power (energy).

It should be also noted that Shannon's theory appeared practically at the beginning of the information era. So, some solutions had to wait for the corresponding technology to be used in practice.

In fact, the theory in engineering sciences is usually a little "behind" the practice, confirming the practical experience. However, in Information Theory the theory was partially far ahead and waited to be confirmed by practice.

In this paper a short surway of the development of Information Theory is presented. At the end some thoughts concerning the future of Information Theory will be given.

2. Communication processes modelling

Although the human race started to communicate from the very beginning, the paper will start only from the beginning of the 20th century. The electrical signals were used to transmit alphanumerical characters and voice (a little later pictures as well). The model used was **deterministic** by its nature. The Fourier analysis, invented a century ago to solve some other problems, was used. The signals were regarded as a sine waves or as a their sum (finite yielding the Fourier series, an infinite yielding the Fourier integral). The obtained results were used in designing classical analogue systems (needed bandwidth, power etc.). In fact, these systems worked quite well in "normal" conditions. But they operated badly in "severe" conditions, or they could not operate at all (FM is unusable for negative – in dB – SNR). The severe conditions are encountered also during the war, where security is needed, too.

So, a better model had to be found. It was based on the **probabilistic approach**. Indeed, two such models appeared.

Norbert Wiener, borrowing the notion of "ensemble" from Statistical Physics and generalising harmonic analysis gave the basic principles of the so–called Statistical Communication Theory.

Claude Shannon, on the other hand, started "from inside", i.e. from communication problems themselves and provided a brilliant and elegant solutions. He created original mathematical concepts. His fundamental paper "A Mathematical Theory of Communication" [1] (transcribed as "**The Mathematical Theory of Communication**" by some scientists) is the basis of Information Theory.

Both models are based on the probabilistic approach. Both use the same mathematical apparatus. But, it is the only common thing.

The main problem in Statistical Communication Theory can be formulated as follows: The transmitted signals are corrupted by noise (the signals, as well as the noise are modelled as stochastic processes). How to extract signals from noise (i.e., how to improve SNR)? The goal is achieved by the so–called optimum filtering and by correlation methods. For digital signal transmission the matched filter can be also regarded as a result of this theory.

In fact, the early beginnings of Information Theory are the works of Nyquist [2] and Hartley [3]. Nyquist found the minimum frequency band to transmit independent discrete signals at a given rate. Hartley proposed to use the logarithmic measure for information (in fact, he said that the information transmitted is proportional to the logarithm of the number of different signals we use – to the alphabet size).

The Shannon approach was totally different from the Wiener one. One should say it was at a higher level. He did not consider the signals, but the information. The information is represented (encoded) by signals, which are carriers of information. That means also that it is possible that transmitted signals do not carry any information at all (from the Information Theory point of view). Of course, these signals may be needed for the proper functioning of the communication system itself (synchronisation etc.).

Shannon defined the quantity of information emitted by information source and tried to find how to represent (encode) the information by the signals so that the information remains undistorted even if the transmitted signals are corrupted or distorted by noise. He investigated the limits of such a system having in mind the source information rate and the channel characteristics (parameters) – bandwidth and SNR.

A communication system from an Information Theory point of view is presented in Fig. 1.

The first thing when one wish to describe mathematically some process is to define a corresponding quantity as well as a unit to measure it. So Shannon had to define the quantity of information in a message. For a discrete source with finite number of messages he defined the quantity of information as a logarithm of the inverse of the message (symbol) probability. The average information rate per symbol (from the source) is obtained by averaging it over all symbols. If 2 is taken as a base for logarithm, the quantity of information is measured in *shannons* and information rate (suitably named entropy, the expression being the same as for the entropy of ideal gas) is measured in *shannons per symbol*. The entropy – H – can be thought as a measure of our uncertainty which message (symbol) will be chosen and emitted by the source. For a source with higher entropy this uncertainty is higher.

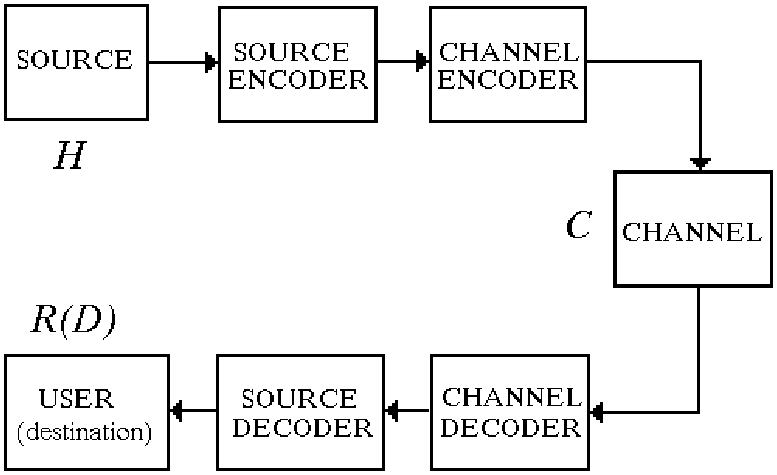

Figure 1. Communication system as "seen" by Information Theory.

The next block is "source encoder". Its task is to represent (encode) the information (messages – symbols) by signals in an efficient way. Shannon showed that the number of signals needed depends on the source entropy (Source Coding Theorem).

The next one is "channel encoder" having the task to encode (represent) the information by the channel signals (symbols) is such a way that no information is lost if the signals are distorted and even if some finite error probability exists. Shannon showed that the error probability can be made as small as possible if the information flow is smaller then the channel capacity – C – depending on bandwidth and SNR (Channel Coding Theorem).

After the corresponding decoders the last block is the "user". It was not taken into account in the beginning. About ten years later [4] Shannon brought the basis of the so–called **Rate Distortion Theory** where $R(D)$ is the minimum amount of information (*shannons per symbol*) needed by the user allowing some distortion of information, quantitatively described by D.

3. Basic results of information theory

3.1 Source coding

The Source Coding Theorem simply states that (for binary signals – bits) the average number of bits needed to encode the source symbol can be made as small as entropy, but not smaller.

The source encoding is accomplished by giving the shorter code words to the symbols with higher probabilities (the same thing did Morse without knowing Information Theory). The algorithms for source encoding (Shannon–Fano and Huffman) were published just a few years after the basic Shannon's paper. Generally, the Huffman algorithm is the optimum one. Shannon also mentioned "arithmetic" encoding based on the cumulative probability.

One of the drawbacks of source encoding is that when one symbol is in error, the synchronisation between encoder and decoder will be lost for some time resulting in a series of erroneously decoded symbols. The efforts have been made to find the suitable code words so as that resynchronisation is obtained as fast as possible.

One may ask oneself: Now we have disks with memory measured in gigabytes, also more gigabits per second (1.1 terabit per second is the last result) can be transmitted through the fibre. Is there any need to compress? The answer is: *yes*!

Every engineer knows that any system should be used efficiently. So, why not put twice (or even three times) more data on the same disk without any change in hardware? Why not transmit twice more data per second through the same channel?

It should be noted that the mentioned algorithms for source encoding are based on the complete knowledge of source statistics (i.e. symbol probabilities). But sometimes we do not know the source statistics. We have no time to store the whole incoming sequence, analyse it, and choose the corresponding code. Then, the adaptive methods are used.

The statistics is performed on the incoming part of a sequence and a corresponding adaptive encoding is performed. The adaptiveness is based on the fact that we know more about incoming sequence, as we have a larger part of it.

Almost all codes for data compression are made in such a way. The best known procedure is Ziv–Lempel encoding (LZ codes) with many versions. If a processed sequence is relatively long, then the adaptive method will attain the limit prescribed by Source Coding Theorem (attainable with Huffman encoder, but with theoretically infinite delay – waiting the end of a sequence to start).

It should be noted also that in the case of LZ codes and their versions there is no need for encoder and decoder to communicate before the start (by sending a list of code words as in the Huffman coding), because they begin to work on the same sequence. So decoder can draw the same conclusions as the encoder at the beginning and then to apply them for further decoding.

It is also assumed that the statistics (known or not known) of the sequence does not change with time. For such a case there is a mathematical argument that the coding will be efficient. It is Asymptotic Equipartition Property the consequence of which is that almost all (i.e. with probability approaching to 1) sequences emitted by the source will be from the "typical set", i.e., all will have the entropy near to the source entropy as the length of a sequence approaches to infinity. For example, the number of possible binary sequences of n bits is 2^n, but if the source entropy is H, then the number of the sequences in a typical set is 2^{nH}.

For those wishing to have a better insight into LZ encoding, just a hint. Kholmogorov defined the complexity of a sequence as a length of a computer program to generate it. LZ algorithms can be thought out as an attempt to write such a program during observing the incoming sequence.

So, the Source Coding Theorem is the basis of **non–destructive (lossless) text compression**, i.e. the original text can be reconstructed in the whole.

One of the interesting examples is the compression of the English text (Shannon wrote the paper "Prediction and Entropy of Printed English" [5]). He considered a text to be a Markov chain and tried to predict the "true" entropy of English finding it to be near to 1 *shannon per letter*, instead between 3 and 4, the result obtained when considering the text as being without memory.

It is interesting to note that Markov took also the text (the Russian one!) as an example, when formulated his theory.

Here is the very place for a little discussion about modelling. Almost all models of English (as well as of other languages), suppose that it can be modelled as a Markov chain with constant memory, considering space sign as a part of the alphabet. In fact, the sounds (phonemes, as linguists call them) are written using letters. The sounds are generated by human being using vocal tract. So, the sequence of letters depends on the possibility of successive sounds pronunciation in a continuous speech. If there are pauses between words, as they should be, then the pauses can not be treated in the same way as other sounds. They just ease the pronunciation. In fact, in a better model the text should be considered as a "pulsed" Markov chain where statistical dependence between letters is disrupted at the end of a word (at least at the end of a sentence). With a new word, i.e. after a space sign, a dependence starts anew from the first letter.

It is also possible to model a text (speech) using words as units (at a syntactic level!) instead of letters. It is interesting for automatic translation.

3.2 Channel coding

The Channel Coding Theorem states that the information can be transmitted with the probability of error being as small as possible (but not zero!) until the information flow is less than the channel capacity. The theorem was proved on the basis of random coding argument. So, it did not show how to find a channel code. It only gave the limit that can be approached by long coding sequences (theoretically when length approaches to infinity). This theorem is a basis for error control coding. These codes are used for transmission as well as for the information storage (magnetic disks, CD etc.). The error control coding theory flourished for many years based sometimes on very specific mathematical apparatus (e.g. Galois Fields!) or giving sometimes the results having to wait to be put in practice with a new technology.

Now, after Ungerboeck's invention of trellis-coded modulation (TCM) [6], we have the first case where we approached to the limit given by Information Theory. In fact, even with commercial modems (with data rates 28800 b$/s$ and even 33600 b$/s$) we are very near to the capacity of the telephone channel now! The telephone channel is regarded as band–limited Gaussian noise channel.

As a newest things in this field we will mention only turbo–codes [7] being very efficient at low SNR as well as that some non–linear codes (Preparata) can be regarded as linear in some higher mathematical structures [8] easing the corresponding coding and decoding procedures.

It should be noted that capacity was defined for point–to–point transmission (channel). Now, the Network Information Theory is developing. It is a system with many senders and receivers (multi–user) and many new elements as well (interference, co–operation and feedback). The general problem is: Given many senders and receivers and the channel transition matrix (describing the effects of interference and noise), decide whether or not the sources can be transmitted over the channel (for example TDMA, CDMA). In fact we do not talk about channel capacity but about the whole medium capacity. The general problem has not yet been solved, but only for some special cases.

3.3 Rate distortion theory

While the Source Coding Theorem concerns to so–called **non–destructive data compression** where the original information must not be distorted, here the "wishes" of the user, concerning the "distortion" of the received information are taken into account. In fact, the user dictates the "fidelity criterion" and the average allowable distortion per symbol (D). So Rate Distortion Function $R(D)$ is defined as giving the minimum mutual information between the source and the user needed for the average distortion being smaller than allowed (D).

So, it is, in fact, **destructive** (in good will, of course) **(lossy) data compression**. This part of Information Theory is the basis for finding the limits when quantising the analogue signals (speech, picture (image) etc.). The theory can be applied to the discrete sources as well.

3.4 General impact of information theory on communications

Firstly, it should be noted that all important quantities are obtained by statistical averaging. So, all results should be taken "on the average" – over long symbol sequences. But our aim is the same – our system should work efficiently on the long run.

Further, the ultimate limits in Communications are given by Information Theory, sometimes without the clear algorithm how to approach them. So, we know what we can do and what we cannot do. But we have often ourselves to find how to do that what we can do according to theory.

Last, but not least, let us discuss the capacity of the "human channel", .i.e. what are our limits when we communicate (consciously) with the outside world. The information rate (flow) of the language is smaller than 50 shannons per second (according to the Information Theory it is near to 10 shannons per second). So, we need only 50 bits (or less) per second to transmit speech. In practice, there are commercial vocoders using 1200 b/s or less. In laboratories we are under 300 b/s. For understanding the received speech, the capacity of our ear (hearing system) should not be greater than 50 Sh/s. In fact it is shown that the human being cannot consciously communicate (taking into account all five senses) with the surrounding faster then at the rate of 300 Sh/s. It is the human channel capacity. So, there is still much room, especially to compress the pictures. Many standards were created or are created now for efficient speech and picture transmission (JPEG, MPEG1, MPEG2 etc.).

We are now in picture (image) compression under a hundred kb/s and try to transmit a picture by a modem over the telephone channel (having videoconferencing and a multimedia in view). From the Information Theory point of view it is possible to approach the rate of 300 b/s for picture transmission, but we still do not know how. Of course, this limit can be approached only asymptotically.

3.5 Information theory and other fields

Information Theory was born inside the Communications and primarily for Communications. Still, it sometimes pays to apply the fresh ideas from one field into another field.

Shannon, himself, formulated a theory of cryptography (secrecy systems) in terms of the concepts of Information Theory [9]. He showed that the entropy of the language (in fact its complement – the redundancy) is related to the possibility for solving cryptograms in this language. For simple substitution cipher he calculated the minimum length of the cryptogram (the number of letters in it) needed to break the cipher The result was confirmed in practice. He also defined a "perfect secrecy" using entropy concept. The same approach can be used to obtain the average number of frames needed to obtain the frame synchronisation.

There were also many other attempts to apply the concepts of. Information Theory in other fields (biology, genetics, psychology, linguistics, economics etc.). Some of them gave results, some did not. Why?

In fact, This is the question of the model. The basic results of Information Theory are aimed at a very specific direction – a direction that may not be necessarily relevant to all fields. So, everyone trying to apply concepts of Information Theory in some other (his!) field should know also the mathematical foundations of Information Theory as well as its communication application. That would help him to evaluate the applicability of Information Theory concepts in his own field.

If the model is adequate (and well understood) then some results can be obtained intuitively (and later confirmed mathematically). Two examples will be given.

Example 1:

The following inequality

$$\ln x \leq x - 1, \qquad x \geq 0$$

where equality holds only at the point $x = 1$, can be easily verified. Consider any two probability distributions $\{p_0, p_1, \ldots, p_{K-1}\}$ and $\{q_0, q_1, \ldots, q_{K-1}\}$ of the alphabet $S = \{s_0, s_1, \ldots, s_{K-1}\}$ of a discrete memoryless source. We may then write

$$\sum_{k=0}^{K-1} p_k \log\left(\frac{q_k}{p_k}\right) \leq \frac{1}{\log_2 e}\sum_{k=0}^{K-1} p_k \ln\left(\frac{q_k}{p_k}\right)$$

using the above inequality we get

$$\begin{aligned}\sum_{k=0}^{K-1} p_k \log_2\left(\frac{q_k}{p_k}\right) &\leq \frac{1}{\log_2 e}\sum_{k=0}^{K-1} p_k\left(\frac{q_k}{p_k}-1\right)\\ &\leq \frac{1}{\log_2 e}\sum_{k=0}^{K-1} p_k(q_k-p_k)\\ &\leq \frac{1}{\log_2 e}\left(\sum_{k=0}^{K-1} q_k - \sum_{k=0}^{K-1} p_k\right) = 0.\end{aligned}$$

Thus, we have the fundamental inequality

$$\sum_{k=0}^{K-1} p_k \log_2\left(\frac{q_k}{p_k}\right) \leq 0$$

where the equality holds only if $q_k = p_k$ for all k. Suppose we next put

$$q_k = \frac{1}{K}, \qquad k = 0, 1, \dots, K-1$$

which corresponds to an alphabet S with equiprobable symbols. The entropy in this case equals

$$\sum_{k=0}^{K-1} q_k \log_2\left(\frac{1}{q_k}\right) = \log_2 K$$

also, putting $q_k = 1/K$ in the fundamental inequality yields

$$\sum_{k=0}^{K-1} p_k \log_2\left(\frac{1}{p_k}\right) \leq \log_2 K$$

So, the entropy of a discrete memoryless source with an arbitrary probability distribution for the symbols of its alphabet is bounded with $\log_2 K$. The equality holds only if the symbols are equiprobable.

It is a proof that maximum entropy for a case of discrete source without memory will be obtained when all symbols are equiprobable. The intuitive proof is: The entropy measures our uncertainty about the source (about the symbols – which one will be emitted). The uncertainty will be maximum when all the symbols are equally likely.

Now, another example, a little more sophisticated.

Example 2:

The entropy of a continuous random variable X, often called differential entropy, is defined by

$$H(X) = -\int\limits_{-\infty}^{\infty} w_X(x) \log_2 w_X(x) dx$$

How to find the probability density function $w_X(x)$ for which the entropy is maximum, subject to the following constraints:

$$\int\limits_{-\infty}^{\infty} w_X(x) dx = 0$$

and

$$\int\limits_{-\infty}^{\infty} (x-\mu)^2 w)X(x) dx = \sigma^2 = \text{const}$$

where μ is the mean of X and σ^2 is its variance.

The method of Lagrange multipliers will be used. The entropy will attain its maximum value only when the integral

$$\int\limits_{-\infty}^{\infty} [-w_X(x) \log_2 w_X(x) + \lambda_1 w_X(x) + \lambda_2 (x-\mu)^2 w_X(x)] dx$$

is stationary (the parameters λ_1 and λ_2 are Lagrange multipliers. So, the entropy is maximum only when the derivative of the integrand with respect to $w_X(x)$ is zero, i.e.

$$-w_X(x) \log_2 w_X(x) + \lambda_1 w_X(x) + \lambda_2 (x-\mu)^2 w_X(x) = 0.$$

This yields the result

$$\begin{aligned} -\log_2 e + \lambda_1 + \lambda_2 (x-\mu)^2 &= \log_2 w_X(x) \\ &= (\log_2 e) \ln w_X(x) \end{aligned}$$

Solving for $w_X(x)$, we get

$$w_X(x) = \exp[-1 + \frac{\lambda_1}{\log_2 e} + \frac{\lambda_2}{\log_2 e}(x-\mu)^2]$$

Taking into account the constraints we get

$$\lambda_1 = \frac{1}{2}\log_2\left(\frac{e}{2\pi\sigma^2}\right)$$

and

$$\lambda_2 = -\frac{log_2 e}{2\sigma^2}$$

yielding finally

$$w_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

i.e. the well known Gaussian probability density.

It is a proof that a "continuous" information source (in fact, a continuous random process) for specified variance will have a maximum differential entropy for Gaussian probability density. The intuitive proof is as follows: According to the Central Limit Theorem, the resultant probability density for the sum of independent random variables is Gaussian. So, a Gaussian probability density is a result of "independent causes" and our uncertainty (and the entropy as well) for such a random variable must be maximum.

4. Conclusion

At the end, a few thoughts about the future of Information Theory It is better not to be a prophet, but some trends can be seen now.

Firstly it is a development of multi–user Information Theory.

Further, it is the so–called "universal coding" i.e. coding without knowing the exact source statistics (or where such a statistics is not easy to model – for example images).

Also, there will be always some new error control codes, as well as some new decoding algorithms for a known codes – the algorithms more suitable for technology to come.

Turbo codes were firstly discovered and later properly understood. they were obtained by concatenation of two or more convolutional codes and decoded by iterative decoding. They are very efficient at a very low SNR. In fact, these codes approach the channel capacity. There is still more room to try various codes as well as to implement more efficient decoding algorithms.

It should also be mentioned the possibility of "soft–decision" decoding for some well known and frequently used codes (e.g. Reed–Solomon codes [10]). The similar ideas can be tried also for some other codes.

An interesting unsolved theoretical problem is the "zero-error channel capacity". i.e. the capacity of channel without errors.

At the end, we should remind ourselves that on the cover of Transactions on Information Theory it is written that "the boundaries of these transactions are deliberately not sharply defined".

REFERENCES

1. C.E. Shannon: *A Mathematical Theory of Communication.* BSTJ, Vol. 27, pp. 379–423 (July 1948), 623–656 (October 1948).
2. H. Niquist: *Certain Topics in Telegraph Transmission Theory.* Trans. of the AIEE, Vol. 47. pp. 617–644, April, 1928.
3. R.V.L. Hartley: *Transmission of Information.* BSTJ, Vol. VII, pp. 535–563, July, 1928.
4. C.E. Shannon: *Coding Theorems for a Discrete Source with a Fidelity Criterion.* IRE Nat. Convention Record, Part 4, pp. 142–163, 1959.
5. C.E. Shannon: *Prediction and Entropy of Printed English.* BSTJ, Vol. 30, pp. 50–64, January 1951.
6. G. Ungerboek: *Channel coding with multilevel/phase signals.* IEEE Trans. on Inform. Theory, Vol. IT–28, pp. 55–67, Jan. 1982.
7. C. Berrou, A. Clavieux, P. Thitimajshima: *Near Shannon Limit Error–Correcting Coding and Decoding Turbo-Codes.* Proc of ICC'93, Geneve, pp. 1064–1070, May 1993.
8. A.R. Hammons, P.V. Kumar, A.R. Calderbank, N.J.A. Sloane, P. Sole: *The Z_4–linearity of Kerdock, Preparata, Goethals, and related Codes.* IEEE Trans. Inform. Theory, Vol. 40, pp. 301–319, March, 1994.
9. C.E. Shannon: *Communication Theory of Secrecy Systems.* BSTJ, Vol. 28, pp.656–715, October, 1949.
10. J.S. Vučković, B.S. Vučetić: *Maximum–Likelihood Decoding of Reed–Solomon Codes.* in 1997 IEEE Intern. Symposion on Information Theory, Ulm, June 29–July 4, 1997, p. 400.