

Paper review**A Mathematical Theory of Communication***Jin Woo Shin, Sang Joon Kim***1 Introduction**

This paper opened the new area - the information theory. Before this paper, most people believed that the only way to make the error probability of transmission as small as desired is to reduce the data rate (such as a long repetition scheme). However, surprisingly this paper revealed that it does not need to reduce the data rate for achieving that much of small errors. It proved that we can get some positive data rate that has the same small error probability and also there is an upper bound of the data rate, which means we cannot achieve the data rate with any encoding scheme that has small enough error probability over the upper bound.

For reviewing this noble paper, first, we summarize the paper at our point of view and then discuss several topics concerned with what the paper includes.

2 Summary of the paper

The paper was categorized as five parts - PART I: Discrete Noiseless systems, PART II: The Discrete Channel With Noise, PART III: Mathematical Preliminaries, PART IV: The Continuous Channel, PART V: The Rate For A Continuous Source.

We reconstruct this structure as four parts. We combine PART I and PART III to 2.1 Preliminary. PART I contained the important concepts for demonstrating the following parts such as entropy and ergodic source. PART III extended the concept of entropy to the continuous case for the last parts of paper. PART II,IV,V dealt with 'discrete source and discrete channel' case, 'discrete source and continuous channel' case and 'continuous source and continuous channel' case respectively. We renamed each PARTs and summarize them with this point of view.

2.1 Preliminary

There are three essential concepts which are used for developing the idea of the paper. One is entropy to measure the uncertainty of the source and another is ergodic property to apply the AEP theorem. The last is capacity to measure how much the channel allow to transfer the information.

2.1.1 Entropy

To measure the quantity of information of a discrete source, the paper defined "entropy". To define the measure $H(p_1, p_2, \dots, p_n)$ reasonably, it should satisfy the following three requirements:

1. H should be continuous in p_i .

2. If all $p_i = \frac{1}{n}$ then, H is monotonically increasing function of n .
3. H is the weighted sum of two H s divided from the original H .

And the paper proved that the only function that satisfies the above three properties is:

$$H = -K \sum_{i=1}^n p_i \log p_i$$

The entropy easily can be extended in the continuous case as following:

$$h = - \int_{-\infty}^{\infty} p(x) \log p(x) dx$$

**we use h for representing the entropy in continuous case to avoid confusing that with the entropy H in discrete case.

2.1.2 Ergodic source

In the Markov process, we call the graph is ergodic if it satisfies the following properties:

1. IRREDUCIBLE: the graph does not contain the case of two isolated nodes A and B, i.e. every node can reach to every node.
2. APERIODIC: the GCD of the lengths of all cycles in the graph is 1.

If the source is ergodic then, with any initial conditions, $P_j(N)$ approach the equilibrium probability when $N \rightarrow \infty$ and for $\forall j$. Equilibrium probability is as following:

$$P_j = \sum_i P_i p_i(j)$$

2.1.3 Capacity

The paper defined capacity as two different ways and in Theorem 12 at page 25, it proved these two definitions are identical. Two different definitions are as following:

1. $C = \lim_{T \rightarrow \infty} \frac{\log N(T)}{T}$
where $N(T)$ is the number of possible signals and T is time duration.
2. $C = \text{Max}(H(x) - H(x|y))$

2.2 Discrete Source and Discrete Channel

Discrete source and Discrete Channel case means that the source information is discrete value and the received data that are sent through the channel are also discrete values. In this case the paper proved the famous theorem (Theorem 11 at the page 22) that there is a rate upper bound that we can achieve the error-free transmission. The detail theorem is as following:

Theorem 1. *If the discrete source entropy H is less than or equal to the channel capacity C then there exists a code that can be transmitted over the channel with arbitrarily small amount of errors. If $H > C$ then there is no method of encoding which gives equivocation less than $H - C$.*

For proving this theorem, the paper used the assumption that the source is ergodic. With this assumption, we can use the AEP theorem and the result is the above theorem.

There are several examples that make sure the result of the theorem and the last one is interesting for us. The last example is a simple encoding function whose rate is equal to the capacity of the given channel, i.e. the most efficient encoding scheme in that channel. The reason why this code is interesting is that it is [7,4,3] Hamming code - the basic linear code.

2.3 Discrete Source and Continuous Channel

2.3.1 The capacity of a continuous channel

In a continuous channel, the input or transmitted signals will be continuous functions of time $f(t)$ belonging to a certain set, and the output or received signals will be perturbed versions of these. (Although the title of this section includes 'discrete source', a continuous channel assumes a continuous source in general.) The main difference from the discrete case is the domain size of the input and output of a channel becomes infinite. The capacity of a continuous channel is defined in a way analogous to that for a discrete channel, namely

$$C = \max_{P(x)} (h(x) - h(x|y)) = \max_{P(x)} \int \int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy$$

Of course, h in the above definition is a differential entropy.

2.3.2 The transmission rate of a discrete source using a continuous channel

If the logarithmic base used in computing $h(x)$ and $h(x|y)$ is two then C is the maximum number of binary digits that can be sent per second over the channel with arbitrarily small error, just as in the discrete case. We can check this fact in two aspects. At first, this can be seen physically by dividing the space of signals into a large number of small cells sufficiently small so that the probability density $P(y|x)$ of signal x being perturbed to point y is substantially constant over a cell. If the cells are considered as distinct points the situation is essentially the same as a discrete channel. Secondly, in the mathematical side, it can be shown that if u is the message, x is the signal, y is the received signal and v is the recovered message, then

$$h(x) - h(x|y) \geq H(u) - H(u|v)$$

regardless of what operations are performed on u to obtain x or on y to obtain v . This means no matter how we encode the binary digits to obtain the signal, no matter how we decode the received signal to recover the message, the transmission rate for the binary digits does not exceed the channel capacity we defined above.

2.4 Continuous Source and Continuous Channel

2.4.1 Fidelity

In the case of a continuous source, an infinite number of binary digits needs for exact specification of a source. This means that transmitting the output of a continuous source with exact recovery at the receiving point requires is impossible using a channel with a certain amount of noise and a finite capacity. Therefore, practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a certain tolerance. The question is, can we assign a transmission rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way. A criterion of fidelity can be represented as follows.

$$\nu(P(x, y)) = \int \int P(x, y) \rho(x, y) dx dy$$

$P(x, y)$ is a probability distribution over the input and output signals, and $\rho(x, y)$ can be interpreted as a "distance" between x and y , in other words, it measures how undesirable it is to receive y when x is transmitted. We can choose a suitable $\rho(x, y)$, for example,

$$\rho(x, y) = \frac{1}{T} \int_0^T [x(t) - y(t)]^2 dt$$

,where T is the duration of the messages taken sufficiently large. In discrete case, $\rho(x,y)$ can be defined as the number of symbols in the sequence y differing from the corresponding symbols in x divided by the total number of symbols in x .

2.4.2 The transmission rate for a continuous source to a fidelity evaluation

For given a continuous source(i.e $P(x)$ fixed) and a fidelity valuation ν (i.e $\int \int P(x,y)\rho(x,y) dx dy$ fixed), we define the rate of generating information as follows.

$$R = \min_{P(y|x)} (h(x) - h(x|y)) = \min_{P(y|x)} \int \int P(x,y) \log \frac{P(x,y)}{P(x)P(y)} dx dy$$

This means that the minimum is taken over all possible channel. The justification of this definition lies in the following result.

Theorem 2. *If a source has a rate R_1 for a fidelity valuation ν_1 , it is possible to encode the output of the source and transmit it over a channel of capacity C with fidelity near ν_1 as desired provided $R_1 \leq C$. This is not possible if $R_1 > C$.*

3 Discussion

3.1 Practical approach to Shannon limit

This paper amazingly provided the upper limit of data rate that can be achieved by any encoding schemes. However, it only proved the "existence" of the encoding scheme that has the data rate under the capacity. Of course, there are several kinds of provided encoding scheme in the paper but they are not practical. Therefore, finding a "good" practical encoding scheme is another issue. There are many attempts to find an encoding scheme to get closed to the capacity. Currently, there are two coding schemes that are most efficient. One is Turbo code, and the other is LDPC(Low Density Parity Check) code.

The encoding schemes of these two codes are quite different. Turbo code is a kind of convolutional code, LDPC is a linear code. However, the decoding algorithms are very similar. They use the iterative decoding algorithm, which reduces the decoding computational complexity.

These codes have a very high performance(near Shannon limit) but the block size should be large enough to get the performance. As this paper revealed, the message size should go to infinity to get the rate that has the negligible error probability. Practically, for get $P_e = 10^{-7}$ in Low SNR ($< 2.5dB$), the block size should be larger than several tens of thousands. However, practically, the above kinds of large message block cannot be applied because of the limitation of decoding and encoding processing power. Therefore it is necessary to find a code that has high performance with reasonable block size.

3.2 Assumption of ergodic process

The fundamental assumption in the paper is that the source information is ergodic. With this assumption, the paper proved the AEP property and capacity theorems. Therefore, one curiosity is arisen that "what happens if the source is not ergodic?". If the information is not ergodic, it is reducible or periodic. If AEP property holds with this source(not ergodic), shannon's capacity theorem also satisfies in this case because capacity theorem is not based on ergodic source but on AEP property. Therefore, to find a source that is not ergodic and holds AEP property is one of meaningful works. Following example is one of these sources.

If a source has following transition probability, then:

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$$

It is not ergodic because of reducibility. However it has a unique stationary probability, $P = [01]$, therefore it holds AEP property.

3.3 Discrete Source with fidelity criterion

For dealing with a continuous source, Shannon introduced the 'fidelity' concept, which is also called 'Rate Distortion'. This concept can also apply to the case of a discrete source as follows.

Definition 3. The information rate R of a discrete source for a given distortion(fidelity) ν is defined as follows.

$$R^{(I)}(\nu) = \min_{P(y|x)} (H(x) - H(x|y))$$

,where $P(x)$ and $\nu = \sum_{x,y} P(x,y)\rho(x,y)$ fixed.

Consider the case when $\nu = 0$. In this case, the channel between X and Y becomes a noiseless channel. Therefore, $H(X|Y) = 0$, and $R = H(X)$. Thus, the rate of a continuous source that Shannon defined in this paper turns out to be the generalized concept of entropy(=the rate of a discrete source) of a discrete source. In other words, entropy of a discrete source is the rate with 0 distortion.

Shannon returned to it and dealt with it exhaustively in his 1959 paper('Coding theorems for a discrete source with a fidelity criterion'), which proved the first rate distortion theorem. To state it, define some notions.

Definition 4. A $(2^{nR}, n)$ rate distortion code consists of an encoding function,

$$f_n : \mathcal{X}^n \rightarrow \{1, 2, \dots, 2^{nR}\}$$

and a decoding function,

$$g_n : \{1, 2, \dots, 2^{nR}\} \rightarrow \mathcal{X}^n$$

, and the distortion ν associated with the code is defined as follows.

$$\nu = \nu(f_n, g_n) = \sum_{x^n} p(x^n)\rho(x^n, g_n(f_n(x^n)))$$

Informally, ν means the expected distance between original messages and recovered messages.

Definition 5. For a given ν , R is said to be a achievable rate if there exists a sequence of $(2^{nR}, n)$ rate distortion codes (f_n, g_n) with $\lim_{n \rightarrow \infty} \nu(f_n, g_n) \leq \nu$

Theorem 6. For an i.i.d source X with distribution $P(x)$ and bounded distortion function $\rho(x, y)$, $R^{(I)}(\nu)$ is the minimum achievable rate at distortion ν .

This main theorem implies that we can compress a source X up to $R^{(I)}(\nu)$ ratio when allowing ν 'distortion'.

4 Conclusion

The reason why we selected this paper for our term project thought the most contents which are contained in this paper already have been studied in the class is that we want to grab the original idea of Shannon when he opened the information theory. The organization of the paper is intuitively clear and there are a lot of useful examples that help understand the notions and concepts. The last part about continuous source is the most difficult and confusing to understand and it may be because it has not been dealt in the class.

We estimate that the notion of entropy and the assumption of ergodic process are the key of this paper because all fruits are from this root. After this paper came out, many scientists are involved in this field and there may be a great improvement from the start point. However, even if the achievement in the information theory has become plentiful it is still under the fundamental notion "entropy" and "ergodic".

It is a great opportunity to read this paper rigorously. We can get the fundamental idea of information theory and we hope to know more about what we discussed one chapter above.