A Brief Introduction on Shannon's Information Theory (old version)

Ricky X. F. Chen
University of Virginia

A Brief Introduction on Shannon's Information Theory

Ricky Xiaofeng Chen*

Abstract

This is an introduction of Shannon's information theory. Basically, it is more like a long note so that it is by no means a complete survey or completely mathematically rigorous. It covers two main topics: entropy and channel capacity. All these concepts will be developed in a totally combinatorial favor. Hopefully, it will be interesting and helpful to those in the process of learning information theory.

Keywords: information, entropy, channel capacity, encoding, decoding

1 Preface

Claud Shannon's paper "A mathematical theory of communication" [1] published in July and October of 1948 is the Magna Carta of the information age. Shannon's discovery of the fundamental laws of data compression and transmission marks the birth of Information Theory.

In this note, we will first introduce two main fundamental results in information theory: entropy and channel capacity, following Shannon's logic (hopefully). For more aspects, we refer the readers to the papers [1, 2, 3] and the references therein. At the end, we have some open discussion comments.

2 Information and Entropy

What is information? or what does it mean when somebody says he has gotten some information regarding something? Well, it means before someone else "communicate" some stuff about this "something", he is not sure of what this "something" is about. Note, anything can be described by several sentences in a language, for instance, English. A sentence or sentences in English can be viewed as a sequences of letters ('a','b','c',...) and symbols (',','.','␣',...). So, we can just think of sentences conveying different meaning

*Biocomplexity Institute and Dept. of Mathematics, Virginia Tech, 1015 Life Sciences Circle, Blacksburg, VA 24061, USA. *Email*: chen.ricky1982@gmail.com, chenshu731@sina.com

as different sequences. Thus, "he is not sure of what this "something" is about" can be understood as "he is not sure which sequence this "something" corresponds". Of course, we can assume he is aware of all possible sequences, only which one of them remains uncertain. He can get some information when someone else pick one sequence and "communicate" it to him. In this sense, we can say this sequence, even each letter there, contains certain amount of information.

Another aspect of these sequences is that not all sequences, words, or letters appear equally. They appear following some probability distribution. For example, the sequence "how are you" is more likely to appear than "ahaojiaping mei"; the letter 'e' is more likely to appear than the letter 'z' (the reader may have noticed that the letter 'z' has not appeared in the text so far). The rough ideas above are the underlying motivation of the following more formal discussion on what information is, how to measure information, and so on.

2.1 How many sequences are there

To formalize the ideas we have just discussed, we assume there is an alphabet $\mathbb{A}$ with n letters, i.e., $\mathbb{A} = \{x_1, x_2, \dots x_n\}$. For example, $\mathbb{A} = \{a, b, \dots, z, `,', `.', `\textvisiblespace', \dots\}$, or just as simple as $\mathbb{A} = \{0, 1\}$. We will next be interested in sequences with entries from the alphabet. We assume each letter x_i appears in all sequences interested with probability $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. To make it simple, we further assume that for any such sequence $s = s_1 s_2 s_3 \cdots s_T$, where $s_i = x_j$ for some j, the exact letter taking by s_i and s_j are independent (but subject to the probability distribution p_i) for all $i \neq j$. Now come to the fundamental question: with these assumptions, how many sequences are there?

It should be noted that a short sequence consisting of these letters will not properly reflect the statistic properties we assumed. Thus, the length T of these sequences we are interested should be quite large, and we will consider the situation as T goes to infinity, denoted by $T \to \infty$. Now, from the viewpoint of statistics, each sequence of length T can be viewed as a series of T independent experiments and the possible outcomes of each experiment are these events (i.e., letters) in $\mathbb{A}$, where the event x_i happens with probability p_i. By the Law of Large Numbers, for T large enough, in each series of T independent experiments, the event x_i will (almost surely) appear $T \times p_i$ (Tp_i for short) times. Assume we label these experiments by $1, 2, \dots T$. Now, the only thing we do not know is in which experiments the event x_i happens.

Therefore, the number of sequences we are interested is the number of different ways of placing Tp_1 number of x_1, Tp_2 number of x_2, and so on, into T positions. Equivalently, it is the number of different ways of placing T different balls into n different boxes such that there are Tp_1 balls in the first box, Tp_2 balls in the second box, and so on and so forth.

Now it should be easy to enumerate the number of these sequences. Let's first consider a toy example:

Example 2.1. Assume there are $T = 5$ balls and 2 boxes. How many different ways to place 2 balls in the first box and 3 balls in the second? The answer is that there are in total

$\binom{5}{2}\binom{5-2}{3} = 10$ different ways, where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ and $n! = n \times (n-1) \times (n-2) \times \cdots 1$.

The same as the example above, for our general setting here, the total number of sequences we are interested is

$$K = \binom{T}{Tp_1} \times \binom{T - Tp_1}{Tp_2} \times \binom{T - Tp_1 - Tp_2}{Tp_3} \times \cdots \binom{T - Tp_1 - \cdots Tp_{n-1}}{Tp_n}. \tag{1}$$

2.2 Average amount of restore resource

Next, if we want to index each sequence among these K sequences using binary digits, i.e., a sequence using only 0 and 1, what is the minimum length of the binary sequence? Still, let us look at an example first.

Example 2.2. If $K = 4$, all 4 sequence can be respectively indexed by 00, 01, 10 and 11. So, the binary sequence should have a length $\log_2 4 = 2$ to index each sequence.

Therefore, the binary sequence should have length $\log_2 K$ in order to index each sequence among all these K sequences. In terms of Computer Science, we need $\log_2 K$ bits to index (and restore) a sequence. Next, we will derive a more explicit expression of $\log_2 K$.

If m is large enough, $m!$ can be quite accurately approximated by the Stirling formula:

$$m! \approx \sqrt{2\pi m}\left(\frac{m}{e}\right)^m. \tag{2}$$

Thus, for fixed $a, b \geq 0$ and $T \to \infty$, we have the approximation:

$$\begin{aligned} \binom{Ta}{Tb} = \frac{(Ta)!}{(Tb)!(Ta-Tb)!} &\approx \frac{\sqrt{2\pi Ta}\left(\frac{Ta}{e}\right)^{Ta}}{\sqrt{2\pi Tb}\left(\frac{Tb}{e}\right)^{Tb}\sqrt{2\pi T(a-b)}\left(\frac{T(a-b)}{e}\right)^{T(a-b)}} \quad (3) \\ &= \frac{\sqrt{a}a^{Ta}}{\sqrt{2\pi T}\sqrt{b(a-b)}b^{Tb}(a-b)^{T(a-b)}}. \quad (4) \end{aligned}$$

Notice that for any fixed $p_i > 0$, $Tp_i \to \infty$ as $T \to \infty$, that means we can apply approximation eq. (4) to every term in eq. (1). Using it, we obtain

$$\begin{aligned} \log_2 K = -n\log_2\sqrt{2\pi T} - \log_2\sqrt{p_1} - \log_2\sqrt{p_2} - \cdots \log_2\sqrt{p_n} \\ -Tp_1\log_2 p_1 - Tp_2\log_2 p_2 - \cdots Tp_n\log_2 p_n. \end{aligned} \tag{5}$$

Now, if we consider the average number of bits a letter needs in indexing a sequence of length T, a minor miracle happens: as $T \to \infty$,

$$\frac{\log_2 K}{T} = -\sum_{i=1}^{n} p_i \log_2 p_i. \tag{6}$$

The expression on the right hand side of eq. (6) is the celebrated quantity associated with a probability distribution, called *Shannon entropy*.

Let's review a little bit what we have done: we have K sequences in total, and all sequences appear equally likely. Assume they encode different messages. Regardless specific messages they encode, we regard them as having the same amount of information. Let's just employ the number of bits needed to encode a sequence to count the amount of information a sequence encode (or can provide). Then, $\frac{\log_2 K}{T}$ can be viewed as the average amount of information a letter has. This suggests that we can actually define the amount of information of each letter. Here, we say "average" because we think the amount of information different letters have should be different as they may not "contribute equally" in a sequence, depending on the respective probabilities of the letters. Indeed, if we look into formula (6), it only depends on the probability distribution of these letters in $\mathbb{A}$. If we reformulate the formula as

$$\sum_{i=1}^{n} p_i \times \log_2 \frac{1}{p_i}.$$

It is clearly the expectation (i.e., average in the sense of probability) of the quantity $\log_2 \frac{1}{p_i}$ associated with the letter x_i, for $1 \le i \le n$. This matches the term "average" so that we can define *the amount of information* a letter x_i with probability p_i has to be $\log_2 \frac{1}{p_i}$ bits.

In this definition of information, we observe that if a letter has a higher probability it has less information, and vice versa. In other words, more uncertainty, more information. Just like lottery, winning the first prize is less likely but more shocking when it happens, while you may think winning a prize of 10 bucks is not a big deal since it is very likely. Hence, this definition agrees with our intuition as well.

In the subsequent of the paper, we will omit the base in the logarithm function. Theoretically, the base could be any number and is 2 by default. Now we summarize information and Shannon entropy in the following definition:

Definition 2.3. Let X be a random variable, taking value x_i with probability p_i, for $1 \le i \le n$. Then, the quantity $I(p_i) = \log \frac{1}{p_i}$ is the amount of *information* encoded in x_i (or p_i), while the average amount of information $\sum_{i=1}^{n} p_i \times \log \frac{1}{p_i}$ is called the *Shannon entropy* of the random variable X (or the distribution P), and denoted by $H(X)$.

Question: among all possible probability distributions, which distribution gives the largest Shannon entropy? For finite case, the answer is given in the following proposition:

Proposition 2.4. *For finite n, when $p_i = \frac{1}{n}$ for $1 \le i \le n$, the Shannon entropy attains the maximum*

$$\sum_{i=1}^{n} \frac{1}{n} \times \log n = \log n.$$

Note by definition of Shannon entropy, if a distribution X has Shannon entropy $H(X)$, then there are $2^{TH(X)}$ sequences satisfying the distribution as T being large enough. Thus, for the distribution attaining the maximum entropy above, there are approximately

$$2^{T \log_2 n} = n^T$$

sequences satisfying that distribution.

On the other hand, for an alphabet with finite n of letters, there are in total n^T different sequences of length T. This appears to be little strange! Because it is clear that the total number of sequences is larger than a specific subclass of sequences.

Let's look at a more specific example: suppose $n = 2$, e.g., sequences of 0 and 1. The distribution attaining the maximum entropy is $P(0) = P(1) = \frac{1}{2}$. The number of sequences of length T for T large enough satisfying this distribution is $\binom{T}{T/2}$. The total number of 0, 1 sequences of length T is

$$2^T = \sum_{i=0}^{T} \binom{T}{i}.$$

The above strange fact implies

$$\lim_{T\to\infty} \frac{\log \binom{T}{T/2}}{T} = \lim_{T\to\infty} \frac{\log\{\binom{T}{T/2} + \sum_{i\neq T/2} \binom{T}{i}\}}{T},$$

which is not that clear. We stop this line of discussion here and remark that the discussion is related to the open discussion at the end.

2.3 Further definitions and properties

The definition of information and entropy can be extended to continuous random variables. Let X be a random variable taking real (i.e., real numbers) values and let $f(x)$ be its probability density function. Then, the probability $P(X = x) = f(x)\Delta x$. Mimic the discrete finite case, the entropy of X can be defined by

$$\begin{aligned} H(X) = \sum_x -P(x) \log P(x) &= \lim_{\Delta x\to 0} \sum_x -[f(x)\Delta x] \log[f(x)\Delta x] && (7)\\ &= \lim_{\Delta x\to 0} \sum_x -[f(x)\Delta x](\log f(x) + \log \Delta x]) && (8)\\ &= -\int f(x) \log f(x)\mathrm{d}x - \log \mathrm{d}x, && (9) \end{aligned}$$

where we used the definition of (Riemann) integral and the fact $\int f(x)\mathrm{d}x = 1$. The last formula here is called the absolute entropy for the random variable X. Note, regardless the probability distribution, there is always a positive infinity term $-\log \mathrm{d}x$. So, we can drop this term and define the (relative) entropy of X to be

$$-\int f(x) \log f(x)\mathrm{d}x.$$

Proposition 2.5. *Among all real random variables with expectation μ and variance σ^2, the Gauss distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ attains the maximum entropy*

$$H(X) = \log \sqrt{2\pi e\sigma^2}.$$

Note joint distribution and conditional distribution are still probability distributions. Then, we can define entropy there correspondingly.

Definition 2.6. Let X, Y be two random variables with joint distribution $P(X = x, Y = y)$ ($P(x, y)$ for short). Then the *joint entropy* $H(X, Y)$ is defined by

$$H(X, Y) = -\sum_{x,y} P(x, y) \log P(x, y). \tag{10}$$

Definition 2.7. Let X, Y be two random variables with joint distribution $P(x, y)$ and conditional distribution $P(y \mid x)$. Then the *conditional entropy* $H(X \mid Y)$ is defined by

$$H(X \mid Y) = -\sum_{x,y} P(x, y) \log P(x \mid y). \tag{11}$$

Remark 2.8. Fixing $X = x$, $P(Y \mid x)$ is also a probability distribution. It's entropy equals

$$H(Y \mid x) = -\sum_{y} P(y \mid x) \log P(y \mid x)$$

which can be viewed as a function over X (or a random variable depending on X). It can be checked that $H(Y \mid X)$ is actually the expectation of $H(Y \mid x)$, i.e.,

$$H(Y \mid X) = \sum_{x} P(x) H(Y \mid x),$$

using the fact that $P(x, y) = P(x)P(y \mid x)$.

Example 2.9. If $Y = X$, we have

$$\begin{aligned} H(X \mid Y) = H(X \mid X) &= -\sum_{x,y} P(x, y) \log P(x \mid y) \\ &= -\sum_{x} P(x, x) \log P(x \mid x) = 0, \end{aligned}$$

where we used the fact that

$$P(x \mid y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{if } x \neq y. \end{cases}$$

Example 2.10. If Y and X are independent, we have

$$\begin{aligned} H(X \mid Y) &= -\sum_{x,y} P(x, y) \log P(x \mid y) \\ &= -\sum_{y} \sum_{x} P(x)P(y) \log P(x) = H(X), \end{aligned}$$

where we used the fact that $P(x, y) = P(x)P(y)$ and $P(x \mid y) = P(x)$ for independent X and Y.

3 Channel Capacity

In a communication system, we have three basic ingredients: the source, the destination and the media between them. We call the media *the (communication) channel*. A channel could be in any form. It could be physical wires, cables, open environment in wireless communication, antennas and certain combination of them.

3.1 Channel without error

Given a channel and a set $\mathbb{A}$ of letters (or symbols) which can be transmitted via the channel. Now suppose an information source generates letters in $\mathbb{A}$ following a probability distribution P (so we have a random variable X taking values in $\mathbb{A}$), and send the generated letters to the destination through the channel.

Assume the channel will carry the exact letters generated by the source to the destination. Then, what is the amount of information received by the destination? Certainly, the destination will receive exactly the same amount of information generated or provided by the source, which is $TH(X)$ in a time period of length of T symbols (with T large enough). Namely, in a time period of symbol-length T, the source will generate a sequence of length T, the destination will receive the same sequence, no matter what the sequence generated at the source is. Hence, the amount of information received at the destination is on average $H(X)$ per symbol.

The *channel capacity* of a channel is the maximum amount of information on average can be obtained at the destination in a fixed time duration, e.g., per second, or per symbol (time). Put it differently, the channel capacity can be characterized by the maximum number of sequences on $\mathbb{A}$ we can select and transmit on the channel such that the destination can in principle determine, without error, the corresponding sequences feed into the channel based on the received sequences.

If the channel is errorless, what is the capacity of the channel? Well, as discussed above, the maximum amount of information can be received at the destination equals the maximum amount of information can be generated at the source. Therefore, the channel capacity C for this case is

$$C = \max_{X} H(X), \text{ per symbol}, \tag{12}$$

where X ranges over all possible distributions on $\mathbb{A}$.

For example, if $\mathbb{A}$ contains n letters for n being finite, then we know from Proposition 2.4 that the uniform distribution achieves the channel capacity $C = \log n$ bits per symbol.

3.2 Channel with error

What is the channel capacity of a channel with error? A channel with error means that the source generated a letter $x_i \in \mathbb{A}$ and transmitted it to the destination via the channel, with some unpredictable error, the received letter at the destination may be x_j. Assume statistically, x_j is received with probability $p(x_j \mid x_i)$ when x_i is transmitted. These probabilities are called transit probabilities of the channel. We assume, once the channel is given, the transit probabilities are determined and will not change.

We start with some examples.

Example 3.1. Assume $\mathbb{A} = \{0, 1\}$. If the transit probabilities of the channel are

$$p(1 \mid 0) = 0.5, \quad p(0 \mid 0) = 0.5,$$
$$p(1 \mid 1) = 0.5, \quad p(0 \mid 1) = 0.5,$$

what is the channel capacity? The answer should be 0, i.e., the destination cannot obtain any information at all.

Because no matter what is being sent to the destination, the received sequence at the destination could be any $0 - 1$ sequence, with equal probability. From the received sequence, we can neither determine which sequence is the one generated at the source, nor can we determine which sequences are not the one generated at the source.

In other words, the received sequence has no binding relation with the transmitted sequence on the channel at all, we can actually flip a fair coin to generate a sequence ourself instead of looking into the one actually received at the destination.

Example 3.2. Assume $\mathbb{A} = \{0, 1\}$. If the transit probabilities of the channel are

$$p(1 \mid 0) = 0.1, \quad p(0 \mid 0) = 0.9,$$
$$p(1 \mid 1) = 0.9, \quad p(0 \mid 1) = 0.1,$$

what is the channel capacity? The answer should not be 0, i.e., the destination can determine something in regard to the transmitted sequence.

Further suppose the source generates 0 and 1 with equal probability, i.e.,$\frac{1}{2}$. Observe the outcome at the destination for a long enough time, that is a sequence long enough, for computation purpose, say a 10000-letter long sequence is long enough (to guarantee the Law of Large Numbers to be effective). With these assumptions, there are 5000 1's and 0's, respectively, in the generated sequence at the source. After the channel, $5000 \times 0.1 = 500$ 1's will be changed to 0's and vice versa. Thus, the received sequence has also 5000 1's and 5000 0's. Assume the sequence received at the destination has 5000 1's for the first half entries and 5000 0's for the second half entries.

With these probabilities and received sequence known, what can we say about the generated sequence at the source? Well, it is not possible immediately to know what is the generated sequence based on these intelligences, because there are more than one sequence which can lead to the received sequence after going through the channel. But, the sequence generated at the source can certainly not be the sequence that contains 5000 0's for the first half and 5000 1's for the second half, or any sequence with most of 0's concentrating in the first half subsequence. Since if that one is the generated one, the received sequence should contain about 4500 0's in the first half of the received sequence, which is not the case observed in the received sequence.

This is unlike the example right above, for which we can neither determine which is generated nor those not generated. Thus, the information obtained by the destination should not be 0.

Let us come back to determine the capacity of the channel. Recall the capacity is the maximum number of sequences on $\mathbb{A}$ we can select and transmit on the channel such that the destination can in principle determine without error the corresponding sequences feed into the channel based on the received sequences. Since there is error in the transmission on the channel, we can not select two sequences which potentially lead to the same sequence after going through the channel at the same time, otherwise we can never determine which one of the two is the transmitted one on the channel based on the same (received) sequence at the destination.

Basically, the possible outputs at the destination are also sequences on $\mathbb{A}$, where element x_i, for $1 \leq i \leq n$, appears in these sequences with probability

$$p_Y(x_i) = \sum_{x_j \in \mathbb{A}} p(x_j)p(x_i \mid x_j).$$

Note this probability distribution will depend only on the distribution X since the transit probabilities are fixed. Denote the random variable associating to this probability distribution at the destination Y (note Y will change as X change). Shannon [1] proved that for a given distribution X, we can choose at most $2^{T[H(X)-H(X|Y)]}$ sequences (satisfying the given distribution) to be the sequences to transmit on the channel such that the destination can determine, in principle, without error, the transmitted sequence based on the received sequences. That is, the destination can obtain $H(X) - H(X \mid Y)$ bits information per symbol. Therefore, the channel capacity for this case is

$$C = \max_X [H(X) - H(X \mid Y)], \text{ per symbol,} \tag{13}$$

where X ranges over all probability distributions on $\mathbb{A}$.

Note, the definition of capacity in eq. (13) applies to channels without error as well. Just noticing that for a channel without error, $Y = X$ so that $H(X \mid Y) = 0$ as discussed in Example 2.9.

4 Open Discussion

There are many articles and news claiming that the Shannon capacity limit defined above was broken. In fact, these are just kind of advertisements on new technologies with more advanced settings than Shannon's original theory, e.g., multiple-antenna transmitting/receiving technologies (MIMO). Essentially, these tech. are still based on Shannon capacity. They have not broken Shannon capacity limit at all.

Can Shannon capacity limit be broken?

In my personal opinion, all "limit" problems are optimization problems under certain conditions. So, it is really about settings. If you believe Shannon's settings are the most suitable one in modelling a communication system, then I suspect you cannot break Shannon capacity limit. Anyway, we end this note with some open discussions:

1. There is no problem to model information sources as random processes. However, given a channel and the set $\mathbb{A}$ of letters transmittable on the channel. To discuss the capacity of the channel, why are we only allowed to select sequences obeying the same probability distribution? Given two probability distributions on $\mathbb{A}$, P_1 and P_2. If there exists x_i for some $1 \le i \le n$ (for discrete finite case) such that

$$\sum_{x_j \in \mathbb{A}} P_1(x_j) p(x_i \mid x_j) \neq \sum_{x_j \in \mathbb{A}} P_2(x_j) p(x_i \mid x_j), \tag{14}$$

we call X_1 and X_2 compatible. Note, in this case, if we transmit a sequence satisfying distribution X_1 and another sequence satisfying distribution X_2, the destination should know, based on inspecting the number of x_i in the received sequence, that the transmitted sequence is from the X_1-class or the X_2-class. We call a set F of probability distributions on $\mathbb{A}$, such that any two probability distributions in F are compatible, an admissible set. Then, in theory, the maximal number of sequences of length T, $T \to \infty$, distinguishable at the destination is

$$\max_F \sum_{X \in F} 2^{T(H(X) - H(X|Y))}.$$

Should the capacity of the channel be defined as

$$\tilde{C} = \lim_{T \to \infty} \frac{\log\{\max_F \sum_{X \in F} 2^{T(H(X) - H(X|Y))}\}}{T}, \tag{15}$$

where F ranges over all admissible sets on $\mathbb{A}$?

Another question: is the number $\tilde{C}$ defined in eq. (15) *really larger than Shannon capacity C?*

To answer the question, let's consider channel without error and channel with error, respectively:

without error. For this case, any two distributions are compatible so that any set of distributions is admissible. Thus, the number defined in eq. (15) reduces to

$$\tilde{C} = \lim_{T \to \infty} \frac{\log\{\sum_X 2^{TH(X)}\}}{T},$$

where the summation is over all possible distributions. The number $\sum_X 2^{TH(X)}$ here should in theory be equal to the total number of sequences of length T. That is, for a finite alphabet with n letters, we have

$$\sum_X 2^{TH(X)} = n^T.$$

Thus, we have

$$\tilde{C} = C = \log n.$$

So, for channels without error, the new definition of channel capacity does not gain much, at least for finite discrete case.

with error. *Is it still true that $C = \tilde{C}$ for a channel with error?* It seems not easy to see the answer[1].

2. Set aside theoretical discussion. There are lots of real systems or simulations. For channels with error, if $\tilde{C} \neq C$, why is there no report of breaking Shannon capacity limit? Possibly, the reason is that these channel encoding methods used in these real systems or simulations just generate sequences (of modulation symbols, e.g., QPSK,16QAM, etc.) satisfying almost the same probability, i.e., agree with Shannon's settings on obtaining channel capacity. *Is that the case?*

References

[1] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J., vol. 27, pp. 379-423, 623-656, July-Oct. 1948.

[2] C.E. Shannon, Communication in the presence of noise, Proc. IRE, 37 (1949), 10-21.

[3] S. Verdú, Fifty years of Shannon theory, IEEE Transactions On Information Theory, 44 (1998), 2057-2078.

[1]The author did not read through all literature on information theory, so the author is not sure if this problem has been addressed and solved.