Information Theory

Informatioon theory is concerned with coding messages for transmission in the most economic way.

This "most frequent $\Leftrightarrow$ shortest code" is the basic principle that was developed mathematically in the 1940s. The pioneer of this new area of information theory was Claude Shannon (1916–2001) and his research greatly contributed to the mathematical notion of entropy.

Figure: Claude Shannon (1916–2001) and Robert Fano (1917–2016).

Information Theory

Shannon set out the basic principles of information theory and illustrated the notions of entropy and conditional entropy from this point of view. The question is here how to efficiently transmit messages through a channel and more complicated cluster of channels.

Information Theory

Shannon set out the basic principles of information theory and illustrated the notions of entropy and conditional entropy from this point of view. The question is here how to efficiently transmit messages through a channel and more complicated cluster of channels. Signals are here strings of symbols, each with potentially its own transmission time and conditions.

Definition Let $W(t)$ be the allowed number of different signals that can be transmitted in time t. The capacity of the channel is defined as

$$\mathrm{Cap} = \lim_{t \to \infty} \frac{1}{t} \log W(t). \tag{1}$$

Information Theory

If $X = \mathcal{A}^*$ is the collection of signals, and every symbol takes τ time units to be transmitted, then

$$W(t) = \#\mathcal{A}^{\lfloor t/\tau \rfloor} \text{ and } \mathrm{Cap} = \frac{1}{\tau} \log \#\mathcal{A}.$$

This $W(t)$ doesn't mean the number of signals can indeed be transmitted together in a time interval of length t, just the total number of signals each of which can be transmitted in a time interval of length t.

Information Theory

If $X = \mathcal{A}^*$ is the collection of signals, and every symbol takes τ time units to be transmitted, then

$$W(t) = \#\mathcal{A}^{\lfloor t/\tau \rfloor} \text{ and } \mathrm{Cap} = \frac{1}{\tau} \log \#\mathcal{A}.$$

This $W(t)$ doesn't mean the number of signals can indeed be transmitted together in a time interval of length t, just the total number of signals each of which can be transmitted in a time interval of length t.

Thus the capacity of a channel is the same as the entropy of the language of signals, but only if each symbol needs the same unit transmission time. If, on the other hand, the possible signals $s_1, \dots, s_n$ have transmission times $t_1, \dots, t_n$, then

$$W(t) = W(t - t_1) + \cdots + W(t - t_n),$$

where the j-th term on the right hand side indicates the possible transmissions after first transmitting s_j.

Information Theory

Using the ansatz $W(t) = ax^t$ for some $x \geq 1$, we get that the leading solution λ of the equation

$$1 = x^{-t_1} + \cdots + x^{-t_n},$$

solves the ansatz, and therefore $\mathrm{Cap} = \log \lambda$.

Information Theory

Theorem: Suppose the transmission is done by an automaton with d states, and from each state i any signal from a different group $S_{i,j}$ can be transmitted with transmission time $t^s_{i,j}$, after which the automaton reaches state j, see Figure 2. Then the capacity of the channel is $\mathrm{Cap} = \log \lambda$ where λ is the leading root of the equation

$$\det \left(\sum_{s \in S_{i,j}} x^{-t^s_{i,j}} - \delta_{i,j} \right) = 0,$$

where $\delta_{i,j}$ indicates the Kronecker delta.

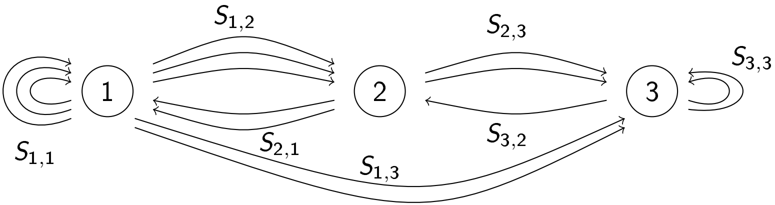


Figure: A transmission automaton.

Information Theory

It makes sense to expand this idea of transmission automaton to a Markov chain, where each transmission $s \in S_{i,j}$ happens with a certain probability $p_{i,j}^{s}$ such that $\sum_{j=1}^{R} \sum_{s \in S_{i,j}} p_{i,j}^{s} = 1$ for every $1 \leq i \leq d$.

Information Theory

It makes sense to expand this idea of transmission automaton to a Markov chain, where each transmission $s \in S_{i,j}$ happens with a certain probability $p_{i,j}^s$ such that $\sum_{j=1}^R \sum_{s \in S_{i,j}} p_{i,j}^s = 1$ for every $1 \leq i \leq d$. For example, if the states $i \in \mathcal{A}$ are the letters in the English alphabet, the transmissions are single letters $j \in \mathcal{A}$ and the probabilities $p_{i,j}^j$ are the diagram frequencies of ij, conditioned to the first letter i.

Information Theory

It makes sense to expand this idea of transmission automaton to a Markov chain, where each transmission $s \in S_{i,j}$ happens with a certain probability $p_{i,j}^s$ such that $\sum_{j=1}^R \sum_{s \in S_{i,j}} p_{i,j}^s = 1$ for every $1 \le i \le d$. For example, if the states $i \in \mathcal{A}$ are the letters in the English alphabet, the transmissions are single letters $j \in \mathcal{A}$ and the probabilities $p_{i,j}^j$ are the diagram frequencies of ij, conditioned to the first letter i. Ergodicity is guaranteed if the graph of this automaton is strongly connected. Also, if π_j is the stationary probability of being in state $j \in \{1, \dots, d\}$, then

$$\pi_j = \sum_{i=1}^{d} \pi_i \sum_{s \in S_{i,j}} p_{i,j}^s \qquad \text{for all } j \in \{1, \dots, d\},$$

see the Perron-Frobenius Theorem.

The Uncertainty Function

Shannon introduce an uncertainty function $H = H(p_1, \dots, p_d)$ as a measure of the amount of uncertainty of the state we are in, if only the probabilities $p_1, \dots, p_d$ of the events leading to this state are known. This function should satisfy the following rules:

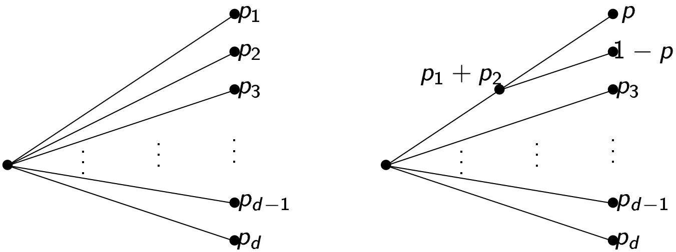
(1) H is continuous in all of its arguments;

(2) If $p_i = \frac{1}{d}$ for all $d \in \mathbb{N}$ and $i \in \{1, \dots, d\}$, then $d \mapsto E(d) := H(\frac{1}{d}, \dots, \frac{1}{d})$ is increasing;

The Uncertainty Function

(3) If the tree of events leading to the present state is broken up into subtrees, the uncertainty H is the weighted average of the uncertainties of the subtrees:

$$H(p_1, \ldots, p_d) = H(p_1 + p_2, p_3, \ldots, p_d) + (p_1 + p_2)H(p, 1 - p).$$



The Uncertainty Function

Theorem: Every uncertainty function satisfying rules (1)-(3) there is $c \geq 0$ such that

$$H(p_1, \ldots, p_d) = -c \sum_{i=1}^{d} p_i \log p_i$$

In particular, $E(d) = c \log d$ and $H(p_1, \ldots, p_d) = 0$ if $p_i \in \{0, 1\}$ for each i. If the total number of transmission words is d, then it is a natural to normalize, i.e., take $c = 1/\log d$.

The Uncertainty Function

Theorem: Every uncertainty function satisfying rules (1)-(3) there is $c \geq 0$ such that

$$H(p_1, \ldots, p_d) = -c \sum_{i=1}^{d} p_i \log p_i$$

In particular, $E(d) = c \log d$ and $H(p_1, \ldots, p_d) = 0$ if $p_i \in \{0, 1\}$ for each i. If the total number of transmission words is d, then it is a natural to normalize, i.e., take $c = 1/\log d$.

Proof: If we break up an equal choice of d^2 possibilities into first d equal possibilities followed by d equal possibilities, we obtain

$$\begin{aligned} E(d^2) &:= H(\frac{1}{d^2}, \ldots, \frac{1}{d^2}) \\ &= H(\frac{1}{d}, \ldots, \frac{1}{d}) + \sum_{i=1}^{d} \frac{1}{n} H(\frac{1}{d}, \ldots, \frac{1}{d}) = 2E(d). \end{aligned}$$

Induction gives $E(d^r) = rE(d)$.

The Uncertainty Function

Now choose $2 \leq a, b \in \mathbb{N}$ and $r, s \in \mathbb{N}$ such that $a^r \leq b^s < a^{r+1}$. Taking logarithms gives $\frac{r}{s} \leq \frac{\log b}{\log a} \leq r + 1s$. The monotonicity of rule (2) also gives

$$rE(a) = E(a^r) \leq E(b^s) = sE(b), \qquad (*)$$

Taking logarithms again: $\frac{r}{s} \leq \frac{E(b)}{E(a)} \leq r + 1s$. Combining the two, we obtain

$$\left| \frac{E(b)}{E(a)} - \frac{\log b}{\log a} \right| \leq \frac{2}{s}.$$

Since $s \in \mathbb{N}$ can be taken arbitrarily large, it follows that

$$E(b) = c \log b \qquad \text{for } c = \frac{E(a)}{\log a}.$$

The monotonicity of rule (2) implies that $c \geq 0$.

The Uncertainty Function

Now assume that $p_i = n_i/N$ for integers n_i and $N = \sum_{i=1}^d n_i$. By splitting the choice into N equal possibilities into d possibilities with probability p_i, each of which is split into n_i equal possibilities, by (3), we get

$$E(N) = H(p_1, \ldots, p_d) + \sum_{i=1}^{d} p_i E(n_i).$$

Inserting (*), we obtain

$$\begin{aligned} H(p_1, \ldots, p_d) &= -c \sum_{i=1}^{d} p_i(\log n_i - \log N) \\ &= -c \sum_{i=1}^{d} p_i \log \frac{n_i}{N} = -c \sum_{i=1}^{d} p_i \log p_i. \end{aligned}$$

This proves the theorem for all rational choices of $(p_1, \ldots, p_d)$. The continuity of rule (1) implies the result for all real probability vectors. This concludes the proof.

Information Theory

Remark: Suppose we compose messages of n symbols in $\{0,1\}$, and each symbol has probability p_0 of being a 0 and $p_1 = 1 - p_0$ of being a 1, independently of everything else. Then the bulk of such messages has np_0 zeros and np_1 ones. The exponential growth rate of the number of such words is, by Stirling's formula

$$
\begin{aligned}
& \lim_{n\to\infty} \frac{1}{n} \log \binom{n}{np_0} \\
= \quad & \lim_{n\to\infty} \frac{1}{n} \log \frac{n^n e^{-n} \sqrt{2\pi n}}{(np_0)^{np_0} e^{-np_0} \sqrt{2\pi n p_0}\ (np_0)^{np_0} e^{-np_0} \sqrt{2\pi n p_0}} \\
= \quad & -p_0 \log p_0 - p_1 \log p_1 = H(p_0, p_1).
\end{aligned}
$$

Information Theory

Recall the convenience of using logarithms base d if the alphabet $\mathcal{A} = \{1, 2, \ldots, d\}$ has d letters. In this base, the exponential growth rate is $H(p_1, \ldots, p_d) \leq 1$ with equality if and only if all $p_a = 1/d$. Thus the number of the most common words (in the sense of the frequencies of $a \in \mathcal{A}$ deviating very little from p_a) is roughly $d^{nH(p_1,\ldots,p_d)}$. This suggests that one could recode the bulk of the possible message with words of length $nH(p_1, \ldots, p_d)$ rather than n.

Information Theory

Recall the convenience of using logarithms base d if the alphabet $\mathcal{A} = \{1, 2, \ldots, d\}$ has d letters. In this base, the exponential growth rate is $H(p_1, \ldots, p_d) \leq 1$ with equality if and only if all $p_a = 1/d$. Thus the number of the most common words (in the sense of the frequencies of $a \in \mathcal{A}$ deviating very little from p_a) is roughly $d^{nH(p_1,\ldots,p_d)}$. This suggests that one could recode the bulk of the possible message with words of length $nH(p_1, \ldots, p_d)$ rather than n. Said differently, the bulk of the words $x_1 \ldots x_n$ have measure

$$p(x_1, \ldots x_n) = \prod_{i=1}^{n} p_{x_i} \approx e^{-nH(p_1,\ldots,p_d)}.$$

Information Theory

Recall the convenience of using logarithms base d if the alphabet $\mathcal{A} = \{1, 2, \ldots, d\}$ has d letters. In this base, the exponential growth rate is $H(p_1, \ldots, p_d) \leq 1$ with equality if and only if all $p_a = 1/d$. Thus the number of the most common words (in the sense of the frequencies of $a \in \mathcal{A}$ deviating very little from p_a) is roughly $d^{nH(p_1,\ldots,p_d)}$. This suggests that one could recode the bulk of the possible message with words of length $nH(p_1, \ldots, p_d)$ rather than n. Said differently, the bulk of the words $x_1 \ldots x_n$ have measure

$$p(x_1, \ldots x_n) = \prod_{i=1}^{n} p_{x_i} \approx e^{-nH(p_1,\ldots,p_d)}.$$

By the Strong Law of Large Numbers, for all $\varepsilon, \delta > 0$ there is $N \in \mathbb{N}$ such that for all $n \geq N$, up to a set of measure ε, all words $x_1 \ldots x_n$ satisfy

$$\left| -\frac{1}{n} \log_d p(x_1 \ldots x_n) - H(p_1, \ldots, p_d) \right| < \delta.$$

Shannon's Source Coding Theorem

Thus, such δ-typical words can be recoded using at most $n(H(p_1, \dots, p_d) + o(1))$ letters for large n, and the compression rate is $H(p_1, \dots, p_d) + o(1)$ as $n \to \infty$. Stronger compression is impossible. This is

Shannon's Source Coding Theorem: For a source code of entropy H and a channel with capacity Cap, it is possible, for any $\varepsilon > 0$, to design an encoding such that the transmission rate satisfies

$$\frac{\text{Cap}}{H} - \varepsilon \leq \mathbb{E}(R) \leq \frac{\text{Cap}}{H}. \tag{2}$$

No encoding achieves $\mathbb{E}(R) > \frac{\text{Cap}}{H}$.

Shannon's Source Coding Theorem

Thus, such δ-typical words can be recoded using at most $n(H(p_1,\ldots,p_d)+o(1))$ letters for large n, and the compression rate is $H(p_1,\ldots,p_d)+o(1)$ as $n\to\infty$. Stronger compression is impossible. This is

Shannon's Source Coding Theorem: For a source code of entropy H and a channel with capacity Cap, it is possible, for any $\varepsilon>0$, to design an encoding such that the transmission rate satisfies

$$\frac{\mathrm{Cap}}{H}-\varepsilon \le \mathbb{E}(R) \le \frac{\mathrm{Cap}}{H}. \tag{2}$$

No encoding achieves $\mathbb{E}(R) > \frac{\mathrm{Cap}}{H}$.

That is, for every $\varepsilon>0$ there is N_0 such that for very $N\ge N_0$, we can compress a message of N letter with negligible loss of information into a message of $N(H+\varepsilon)$ bits, but compressing it in fewer bit is impossible without loss of information.

Proof of Shannon's Source Coding Theorem

Proof: Assume that the source messages are in alphabet $\{1, \ldots, d\}$ and letters s_i appear independently with probability p_i, so the entropy of the source is $H = -\sum_i p_i \log p_i$. For the uppoer bound, assume that the ith letter from the source alphabet require t_i bits to be transmitted.

Proof of Shannon's Source Coding Theorem

Proof: Assume that the source messages are in alphabet $\{1, \ldots, d\}$ and letters s_i appear independently with probability p_i, so the entropy of the source is $H = -\sum_i p_i \log p_i$. For the uppoer bound, assume that the ith letter from the source alphabet require t_i bits to be transmitted.

The expected rate $\mathbb{E}(R)$ should be interpreted as the average number of bits that a bit of a "typical" source message requires to be transmitted. Let $\mathcal{L}_N$ be the collection of N-letter words in the source, and μ_N be the N-fold Bernoulli product measures with probability vector $p = (p_1, \ldots, p_d\}$.

Proof of Shannon's Source Coding Theorem

Let

$$A_{N,p,\varepsilon} = \{s \in \mathcal{L}_N : |\frac{|s|_i}{N} - p_i| < \varepsilon \text{ for } i = 1, \dots, d\}.$$

Proof of Shannon's Source Coding Theorem

Let

$$A_{N,p,\varepsilon} = \{s \in \mathcal{L}_N : |\frac{|s|_i}{N} - p_i| < \varepsilon \text{ for } i = 1, \dots, d\}.$$

By the Law of Large Numbers, for any $\delta, \varepsilon > 0$ there is N_0 such that $\mu_N(A_{N,p,\varepsilon}) > 1 - \delta$ for all $N \geq N_0$. This suggests that a source message s being "typical" means $s \in A_{N,p,\varepsilon}$, and the transmission length of s is therefore approximately $\sum_i p_i t_i N$. Thus typical words $s \in \mathcal{L}_N$ require approximately $t = \sum_i p_i t_i N$ bits transmission time, and the expected rate is $\mathbb{E}(R) = \sum_i p_i t_i)^{-1}$.

Proof of Shannon's Source Coding Theorem

For the capacity, the number of possible transmissions of t bits is at least the cardinality of $A_{N,p,\varepsilon}$, which is the multinomial coefficient $\binom{N}{p_1 N, \ldots, p_d N}$. Therefore, by Stirling's Formula,

$$\begin{aligned}
\text{Cap} &\geq \frac{1}{t} \log \binom{N}{p_1 N, \ldots, p_d N} \\
&\geq \frac{1}{\sum_i p_i t_i N} \log \left((\sqrt{2\pi N})^{1-d} \prod_{i=1}^{d} p_i^{-(p_i N + \frac{1}{2})} \right) \\
&= \frac{-\sum_i p_i \log p_i}{\sum_i p_i t_i} - \frac{\sum_i \log p_i}{2 \sum_i p_i t_i N} - \frac{\frac{d-1}{2} \log 2\pi N}{\sum_i p_i t_i N} \geq RH,
\end{aligned}$$

proving the upper bound.

Proof of Shannon's Source Coding Theorem

The coding achieving the lower bound in (2) that was used in Shannon's proof resembled one designed by Fano. It is now known as the Shannon-Fano code and works as follows:
For the lower bound, let again $\mathcal{L}_N$ be the collection of words B of length N in the source, occurring with probability p_B. The Shannon-McMillan-Breiman Theorem implies that for every $\varepsilon > 0$ there is N_0 such that for all $N \geq N_0$,

$$|-\frac{1}{N}\log p_B - H| < \varepsilon \text{ for all } B \in \mathcal{L}_N \text{ except for a set of measure } < \varepsilon.$$

Thus the average

$$G_N := -\frac{1}{N}\sum_{B\in\mathcal{L}_N} p_B \log p_B \to H \qquad \text{as } N \to \infty.$$

Proof of Shannon's Source Coding Theorem

If we define the condition entropy of symbol a in the source alphabet following a word in $\mathcal{L}_N$ as

$$F_{N+1} = H(Ba|B) = -\sum_{B \in \mathcal{L}_N} \sum_{a \in \mathcal{S}} p_{Ba} \log_2 \frac{p_{Ba}}{p_B},$$

then after rewriting the logarithms, we get $F_{N+1} = (N+1)G_{N+1} - NG_N$, so $G_N = \sum_{n=0}^{N-1} F_{n+1}$. Because the conditional entropy is decreasing as the words B get longer. Thus F_N is decreases in N and G_N is a decreasing sequence as well.

Proof of Shannon's Source Coding Theorem

Assume that the words $B_1, B_2, \ldots, B_n \in \mathcal{L}_N$ are arranged such that $p_{B_1} \geq p_{B_2} \geq \cdots \geq p_{B_n}$. Shannon encodes the words B_i in binary as follows. Let $P_s = \sum_{i<s} p_{B_i}$, and choose $m_s = \lceil -\log p_{B_s} \rceil$, encode m_s as the first m_s digit of the binary expansion of P_s, see Table 1.

p_{B_s}	P_s	m_s	Shannon	Fano
$\frac{8}{36}$	$\frac{28}{36}$	3	110	11
$\frac{7}{36}$	$\frac{21}{36}$	3	101	101
$\frac{6}{36}$	$\frac{21}{36}$	3	011	100
$\frac{5}{36}$	$\frac{15}{36}$	3	010	011
$\frac{4}{36}$	$\frac{6}{36}$	4	0010	010
$\frac{3}{36}$	$\frac{3}{36}$	4	0001	001
$\frac{2}{36}$	$\frac{1}{36}$	5	00001	0001
$\frac{1}{36}$	$\frac{0}{36}$	6	00000(0)	0000

Table: An example of encoding using Shannon code and Fano code.

Proof of Shannon's Source Coding Theorem

Because $P_{s+1} \geq P_s + 2^{-m_s}$, the encoding of B_{s+1} differs by at least one in the digits of the encoding of B_s. Therefore all codes are different.

The average number of bits **per symbol** is $H' = \frac{1}{N}\sum_s m_s p_{B_s}$, so

$$
\begin{aligned}
G_N &= -\frac{1}{N}\sum_s p_{B_s} \log p_{B_s} \\
&\leq H' < -\frac{1}{N}\sum_s p_{B_s}(\log p_{B_s} - 1) = G_N + \frac{1}{N}.
\end{aligned}
$$

Proof of Shannon's Source Coding Theorem

Because $P_{s+1} \geq P_s + 2^{-m_s}$, the encoding of B_{s+1} differs by at least one in the digits of the encoding of B_s. Therefore all codes are different.

The average number of bits **per symbol** is $H' = \frac{1}{N} \sum_s m_s p_{B_s}$, so

$$\begin{aligned} G_N &= -\frac{1}{N} \sum_s p_{B_s} \log p_{B_s} \\ &\leq H' < -\frac{1}{N} \sum_s p_{B_s} (\log p_{B_s} - 1) = G_N + \frac{1}{N}. \end{aligned}$$

Therefore the average rate of transmission is

$$\frac{\text{Cap}}{H'} \in \left[\frac{\text{Cap}}{G_N + \frac{1}{N}}, \frac{\text{Cap}}{G_N} \right].$$

Since G_N decreases to the entropy H, the above tends to Cap $/H$ as required.

Proof of Shannon's Source Coding Theorem

Fano used a different and slightly more efficient encoding, but with the same effect (the difference negligible for large values of N). He divides $\mathcal{L}_N$ into two groups of mass as equal to $1/2$ as possible. The first group gets first symbol 1 in its code, the other group 0. Next divide each group into two subgroups of mass as equal to $1/4$ as possible. The first subgroups get second symbol 1, the other subgroup 0, etc. See Table 1.