Course Organisation

- 4 lectures
- Course material: lecture notes (4 Sections) and slides
- 1 examples paper
- Exam material: notes and example paper, past exams
Course contents

1. Introduction to Information Theory
2. Good Variable Length Codes
3. Higher Order Sources
4. Communication Channels
5. Continuous Random Variables
This lecture

- A bit of history...
- Hartley’s measure of information
- Shannon’s uncertainty / entropy
- Properties of Shannon’s entropy
Claude Elwood Shannon (1916-2001)
Shannon’s paper, 1948


A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

The recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.


Shannon’s paper, 1948

Fig. 1 — Schematic diagram of a general communication system.

A decimal digit is about $3\frac{1}{3}$ bits. A digit wheel on a desk computing machine has ten stable positions and therefore has a storage capacity of one decimal digit. In analytical work where integration and differentiation are involved the base $e$ is sometimes useful. The resulting units of information will be called natural units.

Change from the base $a$ to base $b$ merely requires multiplication by $\log_b a$.

By a communication system we will mean a system of the type indicated schematically in Fig. 1. It consists of essentially five parts:

1. An information source which produces a message or sequence of messages to be communicated to the receiving terminal. The message may be of various types: (a) A sequence of letters as in a telegraph of teletype system; (b) A single function of time $f(t)$ as in radio or telephony; (c) A function of time and other variables as in black and white television; (d) Two or more functions of time, say $f(t)$, $g(t)$, $h(t)$ — this is the case in "three-dimensional" sound transmission or if the system is intended to service several individual channels in multiplex; (e) Several functions of several variables — in color television the message consists of three functions $f(x; y; t)$, $g(x; y; t)$, $h(x; y; t)$ defined in a three-dimensional continuum — we may also think of these three functions as components of a vector field defined in the region — similarly, several black and white television sources would produce "messages" consisting of a number of functions of three variables; (f) Various combinations also occur, for example in television with an associated audio channel.
R.V.L. Hartley (1888-1970)
Uncertainty (Entropy)

Shannon Entropy

For a discrete random variable $X$,

$$H(X) \overset{\text{def}}{=} - \sum_{x \in \text{supp } P_X} P_X(x) \log_b P_X(x) = \mathbb{E}[- \log_b P_X(X)]$$

- $b = 2$, entropy in bits
- $b = e$, entropy in nats
- $b = 10$, entropy in Hartleys
Properties of the Entropy

Extremes

If $X$ is defined over an alphabet of size $N$,

$$0 \leq H(X) \leq \log_b N$$

with equality on the left if and only if $|\text{supp } P_X| = 1$ and on the right if and only if $P_X(x) = 1/N$ for all $x$.

**Proof:**

$$H(X) - \log_b N = -\sum_x P_X(x) \log_b P_X(x) - \sum_x P_X(x) \log_b N$$

$$= \frac{1}{\log_e b} \sum_x P_X(x) \log_e \frac{1}{NP_X(x)}$$

$$\leq \frac{1}{\log_e b} \sum_x P_X(x) \left( \frac{1}{NP_X(x)} - 1 \right) \quad \text{(IT inequality)}$$

$$= \frac{1}{\log_e b} \left( \sum_x \frac{1}{N} - \sum_x P_X(x) \right) = 0$$
Lemma (IT-Inequality)

\[ \log_e(x) \leq x - 1 \]

with equality if and only if \( x = 1 \).
Binary Entropy Function $h(p)$

Good to remember:

$$h(0.11) \approx \frac{1}{2}$$
Block and Conditional Entropy

- Entropy naturally extends to “vectors” or “blocks”:

\[
H(X) = -\sum_x P_X(x) \log P_X(x)
\]

\[
H(XY) = -\sum_{xy} P_{XY}(x, y) \log P_{XY}(x, y)
\]

- Entropy conditioned on an event:

\[
H(X|Y = y) = -\sum_x P_{X|Y}(x|y) \log P_{X|Y}(x|y)
\]

**Equivocation or conditional entropy**

\[
H(X|Y) = \sum_y P_Y(y) H(X|Y = y) = E[- \log P_{X|Y}(X|Y)]
\]
Chain rule of entropies

<table>
<thead>
<tr>
<th>Two random variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H(XY) = H(X) + H(Y</td>
</tr>
</tbody>
</table>

Follows directly from our definition of $H(Y|X)$

<table>
<thead>
<tr>
<th>Any number of random variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H(X_1X_2\ldots X_N) = H(X_1) + H(X_2</td>
</tr>
</tbody>
</table>

Follows from recursive application of the two variable chain rule
Does conditioning reduce uncertainty?

- **X** color of randomly picked ball
- **Y** row of picked ball (1-10)
- \( H(X) = h(1/4) = 2 - \frac{3}{4} \log_2 3 = 0.811 \) bits
- \( H(X|Y = 1) = 1 \) bit \( > H(X) \)
- \( H(X|Y = 2) = 0 \) bits \( < H(X) \)
- \( H(X|Y) = \frac{5 \times 1 + 5 \times 0}{10} = \frac{1}{2} < H(X) \)

**Conditioning Theorem**

\[ 0 \leq H(X|Y) \leq H(X) \]

Conditioning on a random variable only ever reduces uncertainty / entropy (but conditioning on an event can increase it).
Proof of conditioning theorem

\[ H(X | Y) - H(X) = H(XY) - H(X) - H(Y) \quad \text{(chain rule)} \]

\[ = \sum_{x, y} P(x, y) \log \frac{P(x)P(y)}{P(x, y)} \]

\[ \leq \sum_{x, y} P(x, y) \left[ \frac{P(x)P(y)}{P(x, y)} - 1 \right] \quad \text{(IT-inequality)} \]

\[ = \sum_{x, y} P(x)P(y) - \sum_{x, y} P(x, y) = 0 \]
Mutual Information

Definition

\[ I(X; Y) = H(X) - H(X|Y) \]

Mutual information is \textit{mutual}:

\[ I(X; Y) = H(X) + H(Y) - H(XY) = H(Y) - H(Y|X) \]
Positivity of Mutual Information

Theorem

\[ I(X; Y) \geq 0 \]

with equality if and only if \( X \) and \( Y \) are independent

Equivalent to \( H(X|Y) \leq H(X) \).