3F1 Information Theory, Lecture 1

Jossy Sayir

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Michaelmas 2013, 22 November 2013

Course Organisation

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

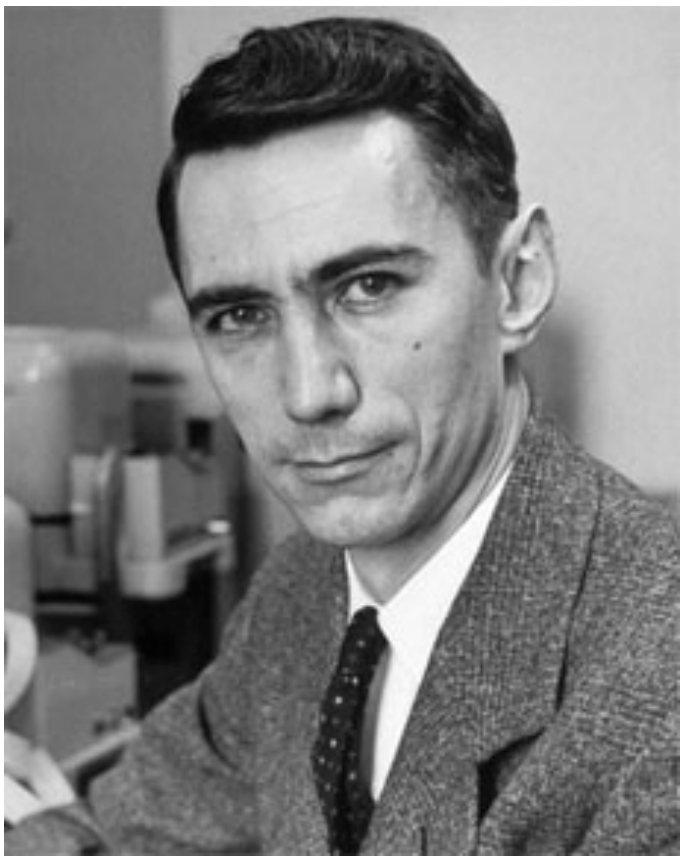
Course contents

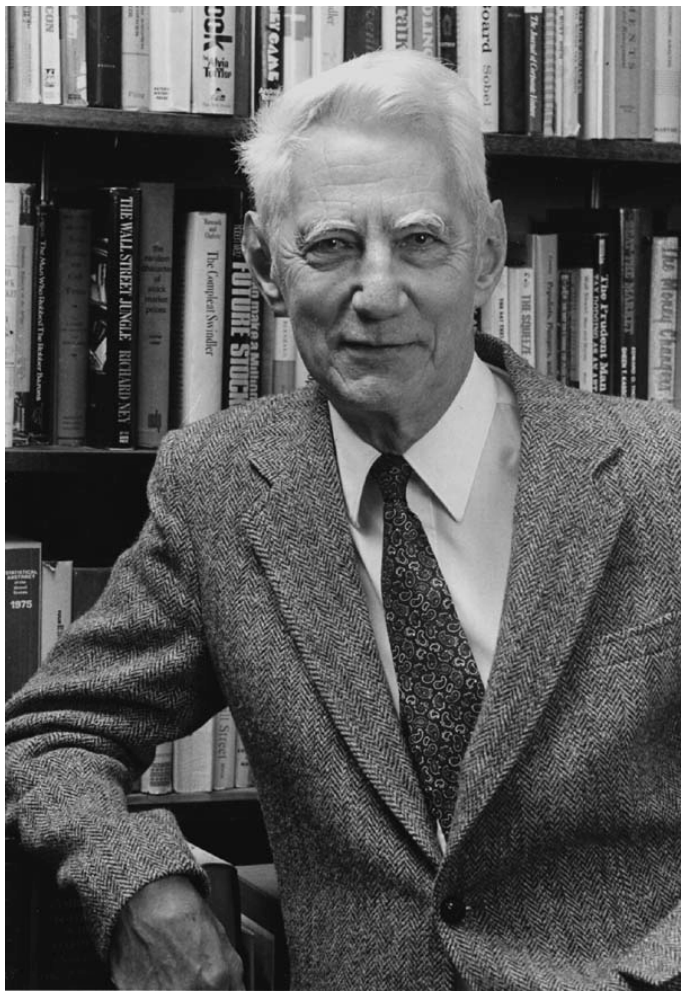
1. Introduction to Information Theory
2. Good Variable Length Codes
3. Higher Order Sources
4. Communication Channels
5. Continuous Random Variables

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

Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 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 Nyquist[1] and Hartley[2] 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

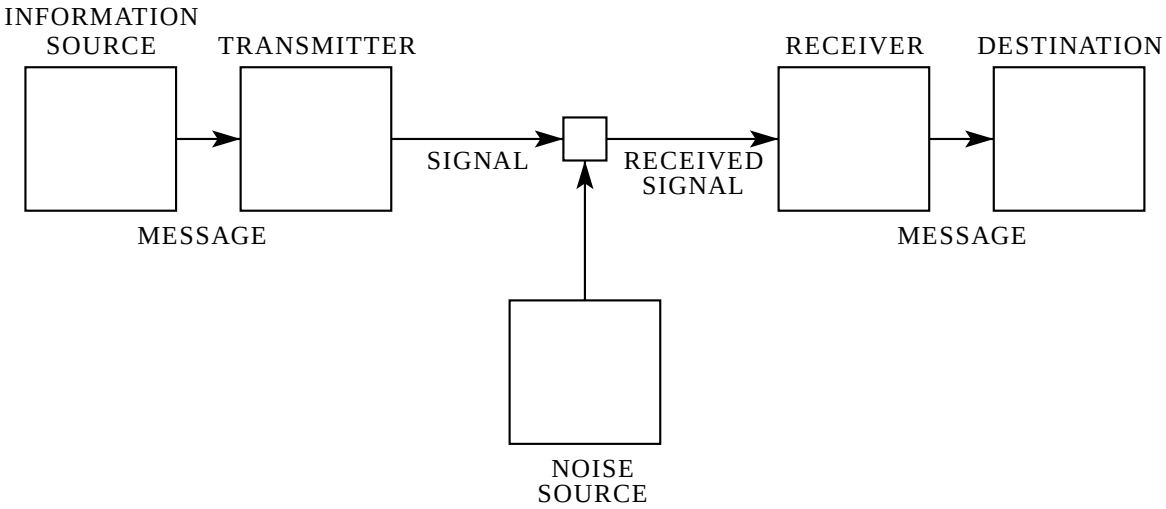


Fig. 1—Schematic diagram of a general communication system.

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—here the message may be thought of as a

R.V.L. Hartley (1888-1970)

Uncertainty (Entropy)

Shannon Entropy

For a discrete random variable X,

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

- $b = 2$, 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,

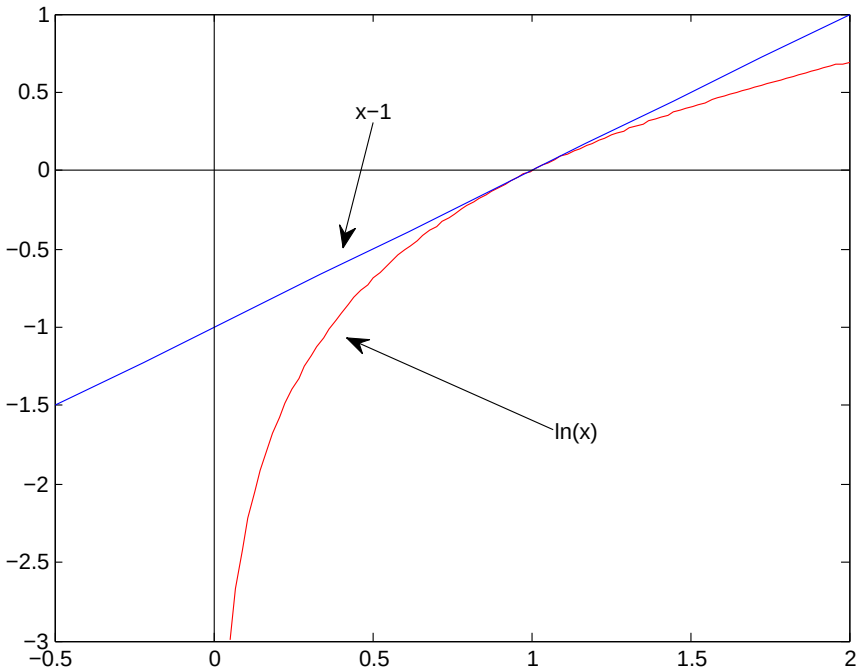
$$0 \leq H(X) \leq \log_b N$$

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

Proof:

$$\begin{aligned} H(X) - \log_b N &= -\sum_x P_X(x) \log_b P_X(x) - \sum_x P_X(x) \log_b N \\ &= \frac{1}{\log_e b} \sum_x P_X(x) \log_e \frac{1}{NP_X(x)} \\ &\leq \frac{1}{\log_e b} \sum_x P_X(x) \left(\frac{1}{NP_X(x)} - 1 \right) \quad \text{(IT inequality)} \\ &= \frac{1}{\log_e b} \left(\sum_x \frac{1}{N} - \sum_x P_X(x) \right) = 0 \end{aligned}$$

IT-Inequality

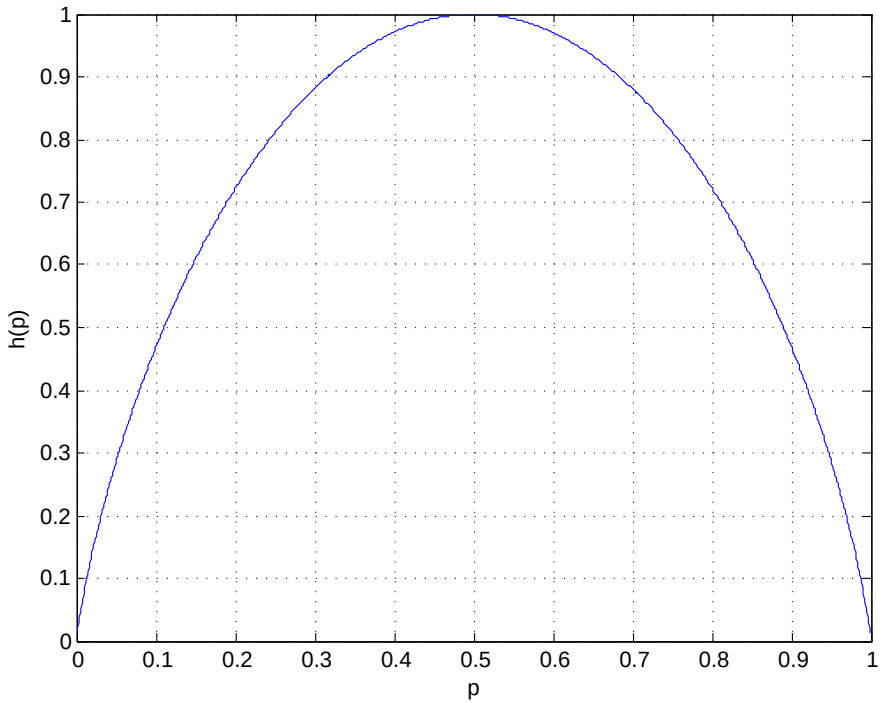


Lemma (IT-Inequality)

$$\log_e(x) \leq x - 1$$

with equality if and only if $x = 1$.

Binary Entropy Function $h(p)$



Good to remember:

$$h(.11) \approx \frac{1}{2}$$

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

Two random variables

$$H(XY) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

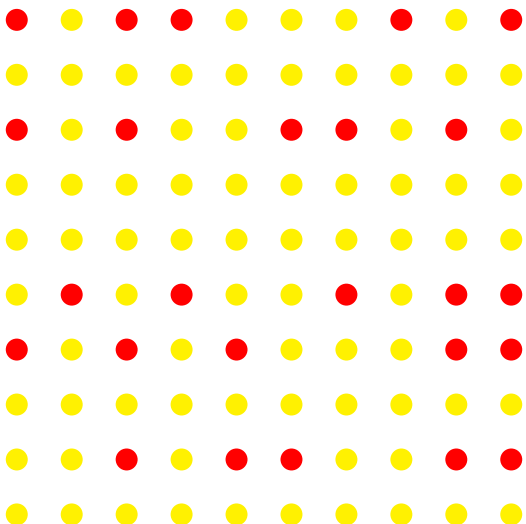
Follows directly from our definition of $H(Y|X)$

Any number of random variables

$$H(X_1 X_2 \dots X_N) = H(X_1) + H(X_2|X_1) + \dots + H(X_N|X_1 \dots X_N)$$

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 increaase it).

Proof of conditioning theorem

$$
\begin{aligned}
H(X|Y) - H(X) &= H(XY) - H(X) - H(Y) && \text{(chain rule)} \\
&= \sum_{x,y} P(x,y) \log \frac{P(x)P(y)}{P(x,y)} \\
&\leq \sum_{x,y} P(x,y) \left[\frac{P(x)P(y)}{P(x,y)} - 1 \right] && \text{(IT-inequality)} \\
&= \sum_{x,y} P(x)P(y) - \sum_{x,y} P(x,y) = 0
\end{aligned}
$$

Mutual Information

Definition

$$I(X;Y) = H(X) - H(X|Y)$$

Mutual information is 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 $H(X|Y) \leq H(X)$.