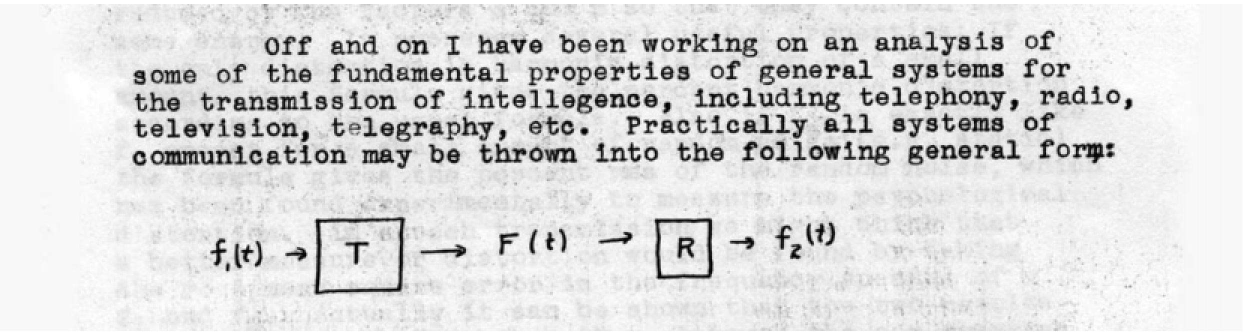
EE 376A: Information theory **Winter 2017**

Lecture 1 — January 10

Lecturer: David Tse *Scribe: Neal Jean, Martin Zhang, Chayakorn Pongsiri*

1.1 Historical Context

The 19^{th} & 20^{th} century witnessed the birth of electronic communication: the telegraph, the telephone, the radio and the television. Even though these devices were all used for communicating information, they were viewed as separate entities and were being developed separately using methodologies which are tied to specific sources and specific channels. In the 1930's, Claude E. Shannon started thinking about a unified theory of communication. In 1939 when he wrote the following letter to his advisor Vannevar Bush:



Off and on I have been working on an analysis of some of the fundamental properties of general systems for the transmission of intellegence, including telephony, radio, television, telegraphy, etc. Practically all systems of communication may be thrown into the following general form:

$$f_1(t) \to \boxed{T} \longrightarrow F(t) \longrightarrow \boxed{R} \to f_2(t)$$

Excerpt of a letter from Shannon to Bush. Feb. 16, 1939. From Library of Congress

From this letter, we see that Shannon had identified the common feature among these devices and was on the route of mathematically formulating the general problem of transmitting information. In his early formulation of the problem, two particular points are of note:

- No channel effects - The channel was assumed to be deterministic.
- Analog signals - The prevalent thought of the day.

1.1.1 *A Mathematical Theory of Communication*

In 1948, Shannon published *A Mathematical Theory of Communication*[1], giving birth to the field of Information Theory. It contained an updated view of communication:

Shannon had made two major modifications that would have huge impact to communication design:

- The source and channel are modeled probabilistically;

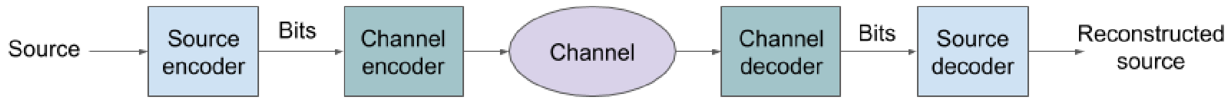


Figure 1.1: View of communication in Shannon's 1948 paper.

- Bits became the common currency of communication

In this paper Shannon proved the following three theorems:

Theorem 1. Minimum compression rate of the source is its entropy rate H

Theorem 2. Maximum reliable rate over the channel is its mutual information I.

Theorem 3. End-to-end reliable communication happens if and only if $H < I$, i.e. there is no loss in performance by using a digital interface between source and channel coding.

After almost 70 years, all communication systems are designed based on the principles of information theory. The fundamental limits delineated by information theory puts a limit on the flow of information through a communication system, just like physical laws impose a limit on what physical systems can be engineered. The limits not only serve as benchmarks for evaluating communication schemes, but also provide insights on designing good ones. In fact, the basic information theoretic limits in Shannon's theorems have now been successfully achieved using efficient algorithms and codes.Beyond communications, information theory has impact on other fields via the measures of information such as entropy and mutual information, as well as its philosophy of understanding the fundamental limits of information processing systems. In this course, we will discuss information theory as both a mathematical theory of communication as well as its broader impact to statistics and machine learning.

1.2 Entropy

Entropy is a fundamental concept in information theory, as it is the measure of the information content contained in any "message", or flow of information. For a discrete random variable X with probability mass function $p(x) \triangleq Pr[X = x]$, we define entropy as

$$H(X) = \mathbb{E}\left[\log_2 \frac{1}{p(X)}\right] = \sum_x p(x) \log_2 \frac{1}{p(x)}.$$

In this course, we will always use logarithm to the base 2, in which case entropy has the unit of bits.

Label-invariance Entropy is *label-invariant*, meaning that it depends only on the probability distribution and not on the actual values that the random variable X takes on. In contrast, quantities like $\mathbb{E}[X^2]$ are label-variant.

1.2.1 Example: Coin flip

Let us work through the simple example of flipping a coin that can take on two values, Heads or Tails. In this scenario, $X \in 0, 1$ and $Pr[X = 0] = p$, so we can compute the entropy of the distribution (dropping the base 2) as

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}.$$

In the case of a fair coin, $p = 0.5$, we find that $H(X) = 0.5(1) + 0.5(1) = 1$. What about for a coin that almost always lands Tails ($X = 0$), with $p = 0.999$? With this heavily-biased coin, we get $H(X) = 0.999 \log \frac{1}{0.999} + 0.001 \log \frac{1}{0.001} \approx 0.011$. Note that entropy can be interpreted as an average of the information one gets when one sees a Tails ($\log(1/0.999)$, very small) and the information when one sees a Heads ($\log(1/0.001)$, fairly big). On the average, however, the entropy of the biased coin is still quite small.

From this example, we can see that we gain more information from more surprising events (i.e., $\log \frac{1}{p(x)} \uparrow$ as $p(x) \downarrow$), but they also happen less often. If we plot the entropy of a Bernoulli distribution, we get the curve in Figure 1.2 which reaches a maximum of 1 when $p = 0.5$. In particular, one sees that the entropy of a binary random variable cannot exceed 1 bit. We will see this as a special case of a more general result later on.

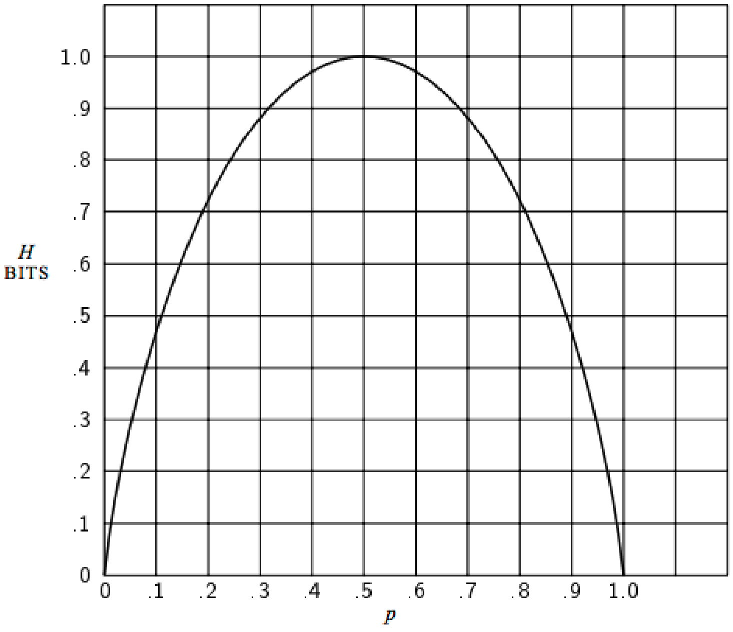


Figure 1.2: Reproduced from Shannon's 1948 paper.

1.2.2 Joint entropy

If we have two random variables X_1 and X_2, then we can compute the **joint entropy** as

$$H(X_1, X_2) = \mathbb{E}\left[\log \frac{1}{p(X_1, X_2)}\right].$$

Note that this is not really a new definition, but is just an application of the definition of entropy to the random vector (X_1, X_2).

If X_1 and X_2 are independent, then

$$\begin{aligned} H(X_1, X_2) =& \mathbb{E}\left[\log \frac{1}{p(X_1)p(X_2)}\right] \\ =& \mathbb{E}\left[\log \frac{1}{p(X_1)}\right] + \mathbb{E}\left[\log \frac{1}{p(X_2)}\right] \\ =& H(X_1) + H(X_2). \end{aligned}$$

Why does log make sense in the definition of entropy? Because of log in the definition, the entropy of independent random variables is a sum of the entropy of individual random variables, which intuitively makes sense.

Bibliography

[1] Shannon, Claude Elwood. "A mathematical theory of communication." ACM SIGMOBILE Mobile Computing and Communications Review 5.1 (2001): 3-55.