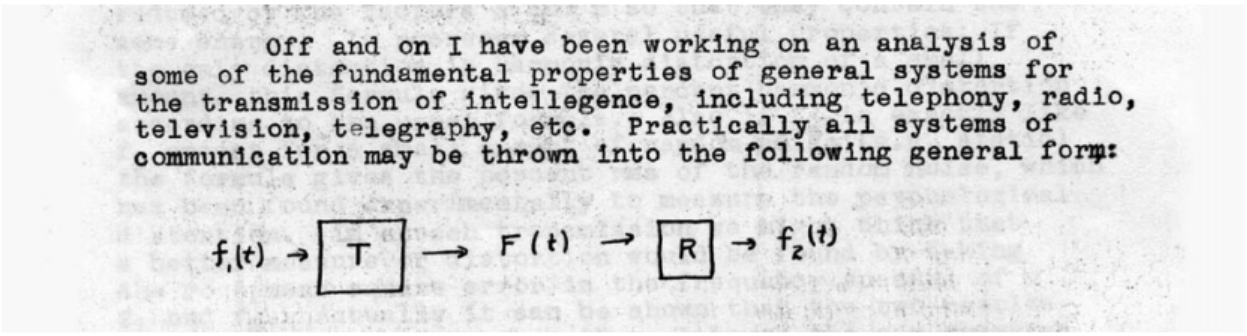
EE 376A: Information theory **Winter 2017**

Lecture 1 — January 10

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

1.1 Historical Context

The 18^{th} & 19^{th} century witnessed birth of devices such as telegraph, telephone, radio, television. Even though these devices were used for communicating information, they were viewed as separate entities and were being developed separately. However, that changed in 1937 when Claude Shannon 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:

$f_1(t) \rightarrow$ [T] $\rightarrow F(t) \rightarrow$ [R] $\rightarrow f_2(t)$

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, he successfully captures few essential parts of the problem, but leaves out a couple important points:

- No channel effects - The channel was assumed to have a deterministic model.
- Analog signals - The signals were thought in the analog framework.

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 the transmission pipeline:

Shannon had made two major modifications that would have huge impacts over the next 70 years:

- The source and channel were now modeled probabilistically
- Bits became the currency of communication

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

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 $H < I$.

Achievements of Information theory - After 70 years **(i)** All communication systems are designed based on the principals of information theory and **(ii)** Various bench marks for comparing different schemes and channels are formulated from information theory.

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

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

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.

1.2.1 Example: Coin flip

Let's 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 $H(X) = 0.999 \log \frac{1}{0.999} + 0.001 \log \frac{1}{0.001} \approx 0.011$.

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 $p = 0.5$.

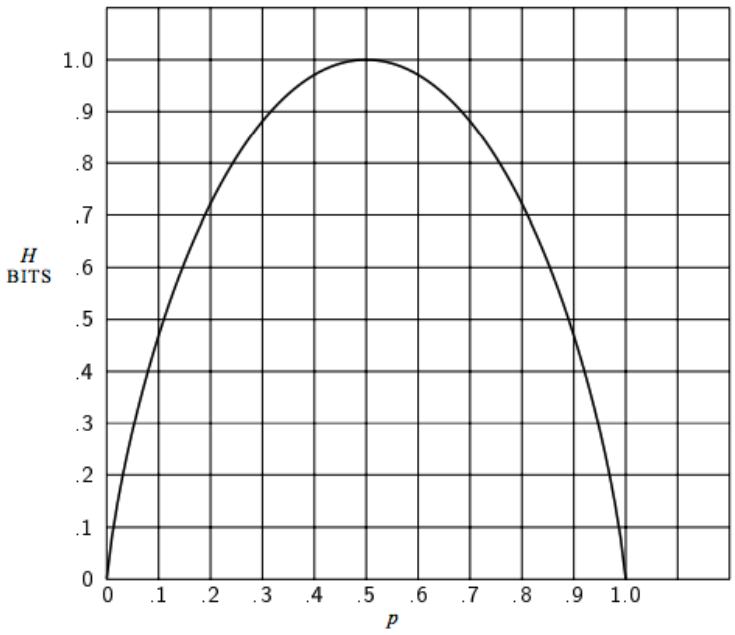


Figure 1.2: Reproduced from Shannon's 1948 paper.

1.2.2 Joint entropy

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

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

.

If X_1 and X_2 are independent, we can show that

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

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.