Shannon's Source Coding Theorem

Shannon's source coding theorem provides both a bound on lossless encodings and an assurance that an optimal coding is achievable for large data streams.

Recall the definition of entropy of a random variable X:

$$H(X) = \sum_{x \in \text{range}(X)} p_X(x) \log\left(\frac{1}{p_X(x)}\right),$$

where $p_X(x) = \Pr[X = x]$.

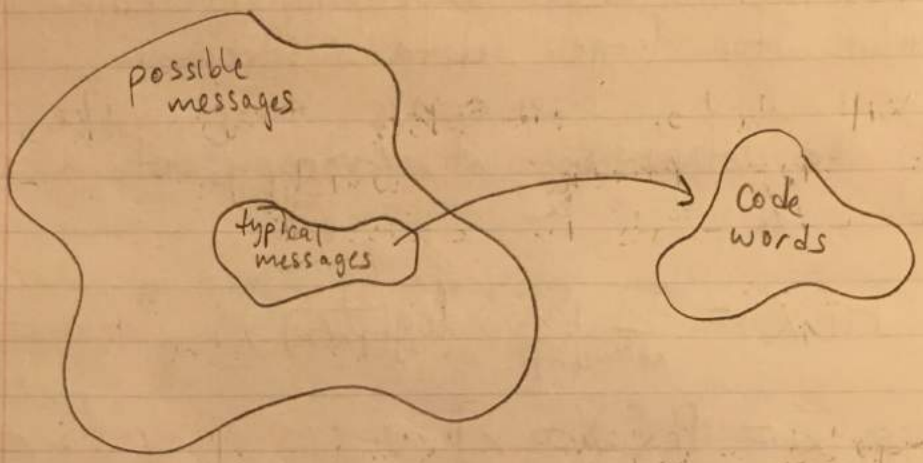
Shannon's source coding theorem relates the size of an encoding of n random variables to the entropy of the variables:

Theorem (Shannon's Source Coding Theorem):

A collection of n i.i.d. random variables, each with entropy $H(X)$, can be compressed into $nH(X)$ bits on average with negligible loss as $n \to \infty$. Conversely, no uniquely decodable code can compress them to fewer than $nH(X)$ bits without loss of information.

The main idea behind Shannon's source coding theorem is to encode only "typical" messages. In the limit, the probability of the occurence of non-typical strings tends to zero, while

considering only typical strings makes our encoding more efficient.



We define a random message as a string of letters $x = x_1 \dots x_n$ drawn from an alphabet $A = \text{range}(X) = \{a_1, \dots, a_K\}$ with probabilities

$$p_X(a_k) = p_k \in (0,1], \quad k = 1, \dots, K.$$

Each string is a draw from the random variable

$$Y = \underbrace{X X \dots X}_{n \text{ copies}}$$

obtained by concatenating n copies of X. Any particular message $x = x_1 \dots x_n$ occurs with probability

$$p_Y(x_1 \dots x_n) = p_X(x_1) \cdots p_X(x_n),$$

because each letter in a string is independent from every other letter in the string.

Consider a long message x. Typically, a letter a_k will appear $N_k \approx n p_k$ times. The probability of a typical message, in which each letter appears approximately N_k times for $k = 1, \ldots, K$, is

$$P_Y(x) \approx P_{typ} = p_1^{N_1} \cdots p_K^{N_K} = \prod_{k=1}^{K} p_k^{np_k}.$$

If typical messages comprise all messages that occur with any significant probability, then the set T of typical messages has size

$$|T| \approx \frac{1}{P_{typ}}$$

since typical messages are uniformly distributed by P_{typ}. If we enumerate and encode each member of T using a binary string, we need

$$I_n = \log_2 |T| = -n \sum_{k=1}^{K} p_k \log_2 p_k = n H(X)$$

bits. The average number of bits per letter is

$$I = \frac{I_n}{n} = H(X).$$

This is the idea behind the proof of Shannon's source coding theorem. To prove it rigorously, we must define what we mean by a typical

message and show that T includes most messages (weighted by probability). First we require some machinery; the law of large numbers.

Given a random variable X, the function $f: \mathcal{A} \to \mathbb{R}$ defines a discrete, real random variable. The realizations of f are the real numbers $f(x)$, $x \in \mathcal{A}$. The <u>average</u> of f is defined as

$$\langle f(X) \rangle = \sum_{x \in \mathcal{A}} p_X(x) f(x) = \sum_{k=1}^{K} p_k f(a_k).$$

The <u>variance</u> is defined as

$$\Delta^2 f(X) = \langle f^2(X) \rangle - \langle f(X) \rangle^2$$

For the sequence $f(Y) = f(X_1), \ldots, f(X_n)$, we define the <u>arithmetic average</u> as

$$A = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

where $X_i = X$ for all i. Note that A is also a random variable and that its average is the same as that of f:

$$\langle A \rangle = \frac{1}{n} \sum_{i=1}^{n} \langle f(X_i) \rangle = \frac{1}{n} n \langle f(X) \rangle = \langle f(X) \rangle.$$

The variance of A is

$$\Delta^2 A = \langle A^2 \rangle - \langle A \rangle^2$$

$$= \frac{1}{n^2} \sum_{i,j} \langle f(X_i) f(X_j) \rangle - \frac{1}{n^2} \sum_{i,j} \langle f(X_i) \rangle \langle f(X_j) \rangle$$

$$= \frac{1}{n^2} \sum_{i} \left[\langle f^2(X_i) \rangle - \langle f(X_i) \rangle^2 \right]$$

$$= \frac{1}{n} \Delta^2 f(X).$$

Thus,

$$\frac{\Delta A}{\langle A \rangle} = \frac{1}{\sqrt{n}} \left(\frac{\Delta f(X)}{\langle f(X) \rangle} \right).$$

This is the law of large numbers. As the number of trials of a random process (i.e., n) increases, the arithmetic average tends to coincide with the average for each trial, $\langle f(X) \rangle$.

We can reformulate the law of large numbers using an ε, δ definition. For $\delta > 0$, define the typical set T_δ of a random sequence y as the set of realizations $x = x_1 \ldots x_n$ such that

$$\langle f(X) \rangle - \delta \leq \frac{1}{n} \sum_{i=1}^{n} f(x_i) \leq \langle f(X) \rangle + \delta.$$

The law of large numbers says that for every $\varepsilon, \delta > 0$ there exists a natural number n_0 such that for all $n > n_0$ the total

probability of all typical messages fulfills

$$P_T = \sum_{x \in T} p_y(x) \geq 1 - \varepsilon .$$

We can see this by applying Chebyshev's inequality.

$$1 - P_T = P_r\left[\left|A - \langle f(X) \rangle\right| > \delta\right] \leq \frac{\Delta^2 A}{\delta^2}$$

$$= \frac{\Delta^2 f(X)}{n \delta^2} .$$

By choosing large enough n, we can make this probability arbitrarily small (e.g. by setting $n_0 = \Delta^2 f(X) / \delta^2 \varepsilon$).

Now, consider the random variable

$$f(X) = -\log_2 p_X(X) .$$

The average of $f(X)$ is the entropy of X:

$$\langle f(X) \rangle = \sum_{x \in \mathcal{A}} p_X(x) f(x)$$

$$= \sum_{x \in \mathcal{A}} p_X(x) \frac{1}{\log_2 p_X(x)}$$

$$= H(X) .$$

Given this choice of f, the typical set contains messages x satisfying

$$H(X) - \delta \leq -\frac{1}{n} \sum_{i=1}^{n} \log_2 p_X(x_i) \leq H(X) + \delta.$$

Equivalently,

$$2^{-n(H(X)+\delta)} \leq p_y(x) \leq 2^{-n(H(X)-\delta)}.$$

By the law of large numbers, if we encode only typical messages, the probability of error is

$$P_{err} = 1 - P_T \leq \varepsilon$$

and can be made arbitrarily small by making n large enough.

Now, let us determine how many typreal messages there are. We have for typical messages $x \in T$,

$$p_y(x) \geq 2^{-n(H(X)+\delta)}$$

$$\sum_{x \in T} p_y(x) \geq |T| 2^{-n(H(X)+\delta)}$$

and

$$p_y(x) \leq 2^{-n(H(X)-\delta)}$$

$$\sum_{x \in T} p_y(x) \leq |T| 2^{-n(H(X)-\delta)}$$

Thus,

$$|T|2^{-n(H(X)-\delta)} \geq P_T$$
$$\geq 1-\varepsilon$$
$$|T| \geq (1-\varepsilon)2^{n(H(X)-\delta)}$$

We have

$$(1-\varepsilon)2^{n(H(X)-\delta)} \leq |T| \leq 2^{n(H(X)+\delta)}.$$

As $n \to \infty$, $\varepsilon, \delta \to 0$, and

$$|T| \to 2^{nH(X)}.$$

We need $I_n \to nH(X)$ bits to encode any message. This proves the first statement in Shannon's source coding theorem.

Next, we prove the converse, by investigating whether or not we can further improve the compression. Let us compress by fixing some $\varepsilon' > 0$ and encoding only sequences that lie in a "subtypical set" $T' \subset T$ with size

$$|T'| \leq (1-\varepsilon)2^{n(H(X)-\delta-\varepsilon')} < 2^{n(H(X)-\delta-\varepsilon')}.$$

This will enable us to encode using fewer than $H(X)$ bits per message on average.

The probability that a message is in T' is

$$P_{T'} = \sum_{x \in T'} p_y(x)$$
$$\leq |T'| \, 2^{-n(H(X)-\delta)},$$

since $p_y(x) \leq 2^{-n(H(X)-\delta)}$ for all $x \in T$, and therefore also for $x \in T'$.

$$P_{T'} \leq 2^{-n(H(X)-\delta-\varepsilon')} \, 2^{-n(H(X)-\delta)}$$
$$= 2^{-n\varepsilon'}.$$

Since $\varepsilon' > 0$, the probability of a successful encoding goes to 0 as $n \to \infty$:

$$P_{T'} \to 0.$$

Thus, if we compress the message below $nH(X)$ bits, we are not likely to encode all typical messages. As $n \to \infty$, we lose all information. This concludes the proof of Shannon's source coding theorem.