Shannon's Source Coding Theorem

Kim Boström
Institut für Physik, Universität Potsdam, 14469 Potsdam, Germany *

The idea of Shannon's famous source coding theorem [1] is to encode only *typical* messages. Since the typical messages form a *tiny* subset of all possible messages, we need less resources to encode them. We will show that the probability for the occurence of non-typical strings tends to zero in the limit of large message lengths. Thus we have the paradoxical situation that although we "forget" to encode most messages, we loose no information in the limit of very long strings. In fact, we make use of *redundancy*, i.e. we do not encode "unnecessary" information represented by strings which almost never occur. Recall that a *random message* of length N is a string

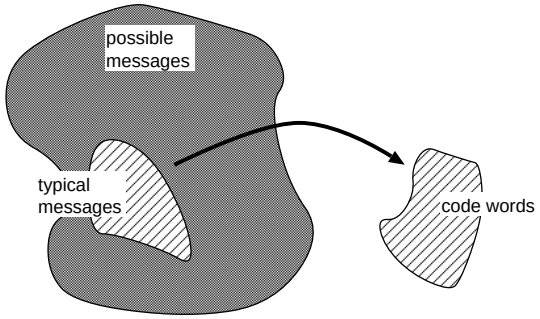

FIG. 1: Lossy coding.

$\boldsymbol{x} \equiv x_1 \cdots x_N$ of letters, which are independently drawn from an alphabet $\mathcal{A} = \{a_1, \ldots, a_K\}$ with *a priori* probabilities

$$p(a_k) = p_k \in (0, 1], \quad k = 1, \ldots, K \tag{1}$$

where $\sum_k p_k = 1$. Each given string $\boldsymbol{x}$ of a random message is an *instance* or *realization* of the message ensemble $\boldsymbol{X} \equiv X_1 \cdots X_N$, where each random letter X_n is identical to a fixed *letter ensemble* X,

$$X_n = X, \quad n = 1, \ldots, N. \tag{2}$$

A particular message $\boldsymbol{x} = x_1 \cdots x_N$ appears with the probability

$$p(x_1 \cdots x_n) = p(x_1) \cdots p(x_n), \tag{3}$$

which expresses the fact that the letters are *statistically independent* from each other.

*Electronic address: bostroem@qipc.org

Now consider a very long message $\boldsymbol{x}$. Typically, the letter a_k will appear with the frequency $N_k \approx N p_k$. Hence, the probability of such *typical message* is roughly

$$p(\boldsymbol{x}) \approx p_{typ} \equiv p_1^{N_1} \cdots p_K^{N_k} = \prod_{k=1}^{K} p_k^{N p_k}. \tag{4}$$

We see that typical messages are *uniformly distributed* by p_{typ}. This indicates that the set T of typical messages has the size

$$|T| \approx \frac{1}{p_{typ}}. \tag{5}$$

If we encode each member of T by a binary string we need

$$I_N \;=\; \log |T| = -N \sum_{k=1}^{K} p_k \log p_k \equiv N H(X), \tag{6}$$

bits, where $H(X)$ is the Shannon entropy of the letter ensemble. Thus for very long messages the average number of bits per letter reads

$$I \equiv \frac{1}{N} I_N = H(X). \tag{7}$$

This is Shannon's source coding theorem in a nutshell. Now let us get a bit more into detail. In order to rigorously prove the theorem we need the concept of a random variable and the law of large numbers. Given the letter ensemble X, the function $f : \mathcal{A} \to \mathbb{R}$ defines a discrete, real *random variable*. The realizations of $f(X)$ are the real numbers $f(x), x \in \mathcal{A}$. The *average* of $f(X)$ is defined as

$$\langle f(X) \rangle := \sum_{x \in \mathcal{A}} p(x)\, f(x) = \sum_{k=1}^{K} p_k\, f(a_k), \tag{8}$$

and the *variance* is given by

$$\Delta^2 f(X) \;:=\; \langle f^2(X) \rangle - \langle f(X) \rangle^2. \tag{9}$$

For the sequence $f(\boldsymbol{X}) \equiv f(X_1), \ldots, f(X_N)$ we define its *arithmetic average* as

$$A := \frac{1}{N} \sum_{n=1}^{N} f(X_n), \tag{10}$$

which is also a random variable. Since the X_n are identical copies of the letter ensemble X, the average of A is equal to the average of $f(X)$,

$$\langle A \rangle \;=\; \frac{1}{N} \sum_{i=1}^{N} \langle f(X_n) \rangle = \langle f(X) \rangle, \tag{11}$$

and the variance of A reads

$$\begin{aligned}\Delta^2 A &= \langle A^2\rangle - \langle A\rangle^2 && (12)\\ &= \frac{1}{N^2}\sum_{n,m}\langle f(X_n)f(X_m)\rangle \\ &\quad -\frac{1}{N^2}\sum_{n,m}\langle f(X_n)\rangle\langle f(X_m)\rangle && (13)\\ &= \frac{1}{N^2}\sum_{n}\left\{\langle f^2(X_n)\rangle - \langle f(X_n)\rangle^2\right\} && (14)\\ &= \frac{1}{N}\Delta^2 f(X). && (15)\end{aligned}$$

The relative standard deviation of A yields

$$\frac{\Delta A}{\langle A\rangle} = \frac{1}{\sqrt{N}}\left(\frac{\Delta f(X)}{\langle f(X)\rangle}\right). \qquad (16)$$

Concluding, in the limit of large N the arithmetic average of the sequence $f(\boldsymbol{X})$ and the ensemble average of $f(X)$ coincide. This is the law of large numbers. It is responsible for the validity of statistical experiments. Without this law, we could never verify statistical properties of a system by performing many experiments. In particular, quantum mechanics would be free of any physical meaning.

Let us reformulate the law of large numbers in the ϵ,δ-language. For $\delta > 0$ we define the *typical set* T of a random sequence $\boldsymbol{X}$ as the set of realizations $\boldsymbol{x} \equiv x_1 \cdots x_N$ such that

$$\langle f(X)\rangle - \delta \leq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \leq \langle f(X)\rangle + \delta. \qquad (17)$$

The law of large numbers implies that for every $\epsilon,\delta > 0$ there is a natural number N_0, such that for all $N > N_0$ the total probability of all typical sequences fulfills

$$P_T \equiv \sum_{\boldsymbol{x}\in T} p(\boldsymbol{x}) \geq 1-\epsilon. \qquad (18)$$

The total probability P_T represents the probability for a randomly chosen sequence $\boldsymbol{x}$ to lie in the typical set T. Now consider the special random variable

$$f(X) := -\log p(X). \qquad (19)$$

The average of $f(X)$ equals the Shannon entropy of the ensemble X,

$$\langle f(X)\rangle = -\sum_{x\in\mathcal{A}} p(x)\log p(x) = H(X). \qquad (20)$$

The typical set now contains all messages $\boldsymbol{x}$ whose probability fulfills

$$H-\delta \leq -\frac{1}{N}\sum_{n=1}^{N}\log p(x_n) \leq H+\delta, \qquad (21)$$

or equivalently

$$2^{-N(H+\delta)} \leq p(\boldsymbol{x}) \leq 2^{-N(H-\delta)}, \qquad (22)$$

where $H \equiv H(X)$. By the law of large numbers, the probability for a randomly drawn message $\boldsymbol{x}$ to be a member of T reads

$$P_T \equiv \sum_{\boldsymbol{x}\in T} p(\boldsymbol{x}) \geq 1-\epsilon. \qquad (23)$$

If we encode only typical sequences, the probability of error

$$P_{err} := 1 - P_T \leq \epsilon \qquad (24)$$

can be made arbitrarily small by choosing N large enough. Now let us determine how many typical sequences there are. The lefthand side of (22) gives

$$\begin{aligned} p(\boldsymbol{x}) &\geq 2^{-N(H+\delta)} && (25)\\ \Leftrightarrow \sum_{\boldsymbol{x}\in T} p(\boldsymbol{x}) &\geq |T|\,2^{-N(H+\delta)}. && (26)\end{aligned}$$

The righthand side of (22) gives

$$\begin{aligned} p(\boldsymbol{x}) &\leq 2^{-N(H-\delta)} && (27)\\ \Leftrightarrow \sum_{\boldsymbol{x}\in T} p(\boldsymbol{x}) &\leq |T|\,2^{-N(H-\delta)}, && (28)\end{aligned}$$

which yields together with (23)

$$\begin{aligned} |T|\,2^{-N(H-\delta)} &\geq 1-\epsilon && (29)\\ \Leftrightarrow |T| &\geq (1-\epsilon)\,2^{N(H-\delta)}. && (30)\end{aligned}$$

Relations (28) and (30) can be combined into the crucial relation

$$(1-\epsilon)\,2^{N(H-\delta)} \leq |T| \leq 2^{N(H+\delta)}. \qquad (31)$$

For $N\to\infty$ we can choose $\epsilon,\delta = 0$ and obtain the desired expression

$$|T| \to 2^{NH(X)}, \qquad (32)$$

thus we need $I_N \to NH(X)$ bits to encode the message. Equivalently, the information content per letter reads $I = H(X)$ bits. Finally, let us investigate if we can further improve the compression. Relation (30) gives a lower bound for the size of the typical set. Let us compress below H bits per letter by fixing some $\epsilon' > 0$ and encode only sequences that lie in a "subtypical set" $T' \subset T$ whose size reads

$$|T'| \leq (1-\epsilon)2^{N(H-\delta-\epsilon')} < 2^{N(H-\delta-\epsilon')}. \qquad (33)$$

The righthand side of (22) states that the probability of a typical sequence is bounded from above by

$$p(\boldsymbol{x}) \leq p_{max} \equiv 2^{-N(H-\delta)}. \qquad (34)$$

If we encode only the typical sequences in the subtypical set T', the probability that a sequence is in T' fulfills

$$\begin{aligned} P_{T'} &= \sum_{\boldsymbol{x}\in T'} p(\boldsymbol{x}) && (35)\\ &\leq |T'| \cdot p_{max} = 2^{N(H-\delta-\epsilon')}\, 2^{-N(H-\delta)} && (36)\\ &= 2^{-N\epsilon'}. && (37) \end{aligned}$$

Because $\epsilon' > 0$, the probability of a successful encoding goes to 0 for $N \to \infty$,

$$P_{T'} \to 0. \qquad (38)$$

Concluding, if we compress the messages below $NH(X)$ bits, we are not able to encode all typical messages and for $N \to \infty$ we will loose all information. A good review on the issue can also be found in [2, 3].

[1] C. E. Shannon and W. Weaver. A mathematical Theory of communication, *The Bell System Technical Journal*, **27**, 379–423,623–656, (1948).

[2] D.J.C. MacKay. Information theory, inference, and learning algorithms, http://wol.ra.phy.cam.ac.uk/mackay/itprnn/book.html, (1995-2000).

[3] J. Preskill. Lecture notes. http://www.theory.caltech.edu/people/preskill/ph219/, (1997-1999).