

The Significance of Entropy: Shannon's Lossless Source Coding Theorem

(Theorems 3 and 4, p. 54, in Shannon.)

Abstract: If we model an information source as a sequence X_1, X_2, \dots of independent, identically distributed random variables, one per second, at what rate is the source producing information? We shall prove that the answer is H bits per second, where H is the common entropy of the X_i 's.

1. The Memoryless Case.

Let $A = \{a_1, a_2, \dots, a_r\}$ be a finite alphabet, with individual letter probabilities $p(a_1), \dots, p(a_r)$. We denote the set of n -letter words over A by A^n , and define the probability of such a word $\mathbf{x} = (x_1, \dots, x_n) \in A^n$ as follows:

$$(1.1) \quad p(\mathbf{x}) = p(x_1)p(x_2) \cdots p(x_n).$$

We now regard A^n as a *discrete memoryless source*. A source word is modelled as a random vector $\mathbf{X} = (X_1, \dots, X_n)$, where the X_i 's are independent, identically distributed random variables, with common distribution given by the $p(a_i)$'s. The *entropy* of this source is defined to be

$$(1.2) \quad H = \sum_{i=1}^r p(a_i) \log \frac{1}{p(a_i)} \quad \text{bits.}$$

(Here and hereafter all logs are base 2.) This quantity, invented by Shannon, has a profound engineering significance. The entropy H represents the *minimum average number of bits needed to represent a source letter faithfully*. We shall give a proof of this fact, for the above memoryless source, in this section. Then in Section 2 we'll argue that provided H is defined properly, the same theorem holds for a much more extensive class of source models.

Suppose then that we wish to "compress" a source word of length n , in such a way that each individual source letter is represented by R bits, where $R > 0$ is an arbitrary real number. We do this by using a *source code* of rate R and length n , which is defined as follows. We choose a subset L_n of A^n containing at most 2^{Rn} words (the "lucky" words):

$$(1.3) \quad L_n \subseteq A^n$$

$$(1.4) \quad |L_n| \leq 2^{Rn}.$$

To each lucky word $\mathbf{x} \in L_n$ we can assign a unique k -bit codeword, where

$$(1.5) \quad k = \lceil Rn \rceil.$$

For each *unlucky* word, i.e., one not in L_n , we assign an *arbitrary* k -bit codeword, e.g., $00 \cdots 0$, *even though this codeword may already have been assigned to a lucky source word*. In this way, each n -letter word is mapped into a k -bit codeword. This is the *compression algorithm*. Each *individual symbol* from A is represented with k/n bits, which, at least if n is large, is very close to R , as desired.

Let us now agree on a *decompression* algorithm, as follows. Given a k -bit codeword \mathbf{y} , if there is a lucky sourceword \mathbf{x} for which \mathbf{y} is the codeword, then the decompression algorithm outputs \mathbf{x} . Otherwise, the decompression algorithm simply guesses about the sourceword. With such a scheme, we can achieve substantial compression, but only at the cost of losing the unlucky source words! (After all, if $R < \log r$, and n is large, 2^{Rn} is a tiny fraction of r^n , which is the total number of source words.) This may however be acceptable, if the *unlucky* source words are *unlikely* to occur. With this in mind, we define the *error probability* of the above-defined source code as the probability that the source produces an unlucky word:

$$(1.6) \quad P_e = \Pr\{\mathbf{x} \notin L_n\} = \sum_{\mathbf{x} \notin L_n} p(\mathbf{x}).$$

The “fundamental theorem of lossless data compression” says that if $R > H$, and n is large enough, P_e can be made exceedingly small, whereas if $R < H$, P_e must approach 1 for large n .

Example: Suppose that $A = \{0, 1\}$, with $p(1) = \alpha$, and $p(0) = 1 - \alpha$. Let us construct a source code of rate $R = 3/7$, and length $n = 7$, by choosing L_7 to be the set of source words whose binary weight is at most 1. Then the corresponding encoding and decoding tables could be as follows:

Encoding Table		Decoding Table	
sourceword	codeword	codeword	sourceword
0000000	000	000	0000000
0000001	001	001	0000001
0000010	010	010	0000010
0000100	011	011	0000100
0001000	100	100	0001000
0010000	101	101	0010000
0100000	110	110	0100000
1000000	111	111	1000000
(anything else)	000		

Here we have achieved substantial compression (2.33:1), but only at the cost of not being able to represent 120 of the 128 source words, viz., the words containing two or more ones! The error probability is (from (1.6))

$$\begin{aligned} P_e &= \sum_{k=2}^7 \binom{7}{k} \alpha^k (1 - \alpha)^{7-k} = 1 - (1 - \alpha)^7 - 7\alpha(1 - \alpha)^6 \\ &= 21\alpha^2 - 70\alpha^3 + \cdots \end{aligned}$$

For example, if $\alpha = .01$, we can achieve the compression of 2.33 : 1 with an error probability of .002. The following famous theorem of Shannon shows that in fact, the same source can be compressed by more than 12 : 1 with arbitrarily small error probability!

1.1 Theorem (Shannon, 1948). *If $R > H$, it is possible to choose a sequence of L_n 's satisfying (1.3) and (1.4), such that*

$$P_e \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Conversely, if $R < H$, for any sequence of L_n 's satisfying (1.3) and (1.4), it must be true that

$$P_e \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Proof: The proof uses the notion of *typical sequences*. For each $\mathbf{x} = (x_1, x_2, \dots, x_n) \in A^n$, we define the *self information* of \mathbf{x} by

$$(1.7) \quad I(\mathbf{x}) = \log \frac{1}{p(\mathbf{x})} = \sum_{k=1}^n \log \frac{1}{p(x_k)}.$$

If \mathbf{x} is regarded as a random variable, then $I(\mathbf{x})$ is the sum of n independent, identically distributed random variables, each with mean $E(\log \frac{1}{p(x)}) = H$ (from (1.2)), and so $I(\mathbf{x})$ will normally be “near” nH . To make the notion of “near” precise, for any fixed $\epsilon > 0$, we say that a source word \mathbf{x} is “ ϵ -typical,” and write $\mathbf{x} \in T_n(\epsilon)$, if

$$(1.8) \quad \left| \frac{1}{n} I(\mathbf{x}) - H \right| \leq \epsilon.$$

Since, as we observed above, $I(\mathbf{x})$ is the sum of n independent, identically distributed random variables, each with mean H , the weak law of large numbers implies that

$$(1.9) \quad \Pr\left\{ \left| \frac{1}{n} I(\mathbf{x}) - H \right| \leq \epsilon \right\} = \Pr\{\mathbf{x} \in T_n(\epsilon)\} \rightarrow 1 \quad \text{as } n \rightarrow \infty,$$

for any fixed $\epsilon > 0$. Using definition (1.7), we see that (1.8) is equivalent to

$$(1.10) \quad 2^{-n(H+\epsilon)} \leq p(\mathbf{x}) \leq 2^{-n(H-\epsilon)} \quad \text{for } \mathbf{x} \in T_n(\epsilon).$$

If we denote the number of ϵ -typical sequences by $|T_n(\epsilon)|$, then (1.10) implies that

$$|T_n(\epsilon)| 2^{-n(H+\epsilon)} \leq \Pr\{T_n(\epsilon)\} \leq |T_n(\epsilon)| \cdot 2^{-n(H-\epsilon)}.$$

Rearranging these inequalities, we obtain the following estimates of $|T_n(\epsilon)|$:

$$(1.11) \quad \Pr\{T_n(\epsilon)\} 2^{n(H-\epsilon)} \leq |T_n(\epsilon)| \leq \Pr\{T_n(\epsilon)\} 2^{n(H+\epsilon)}.$$

By (1.9), if we define $\delta_n = -\log \Pr\{T_n(\epsilon)\}$, then $\delta_n \geq 0$ and $\delta_n \rightarrow 0$, and so (using $\Pr\{T_n(\epsilon)\} \leq 1$) we can rearrange (1.11) to obtain

$$(1.12) \quad 2^{n(H-\epsilon)-\delta_n} \leq |T_n(\epsilon)| \leq 2^{n(H+\epsilon)}.$$

Taken together, (1.9), (1.10), and (1.12) say that for large n , the “typical sequences” account for nearly all of the “mass” in the probability space A^n ; that there are about 2^{nH} typical sequences; and that each typical sequence has probability about 2^{-nH} . In other words, asymptotically the probability space becomes a *uniform* probability space! This striking result is called the “asymptotic equipartition” property.

Now we can prove the theorem.

Case 1: $R > H$. Choose ϵ so that $R > H + \epsilon$ too. Then $2^{Rn} > 2^{n(H+\epsilon)}$, and so by (1.11) we can choose the subset L_n to contain $T_n(\epsilon)$. Then by (1.9), we have

$$1 \geq \Pr\{\mathbf{x} \in L_n\} \geq \Pr\{\mathbf{x} \in T_n(\epsilon)\} \rightarrow 1,$$

and so $P_e = 1 - \Pr\{\mathbf{x} \in L_n\} \rightarrow 0$, as promised. In other words, if we can provide codewords for all the typical sequences, then if n is large enough, our error probability will be negligible.

Case 2: $R < H$. Choose ϵ so that $R < H - \epsilon$ too. Then we have

$$\begin{aligned} 1 - P_e &= \Pr\{\mathbf{x} \in L_n\} = \Pr\{\mathbf{x} \in L_n \cap T_n(\epsilon)\} + \Pr\{\mathbf{x} \in L_n \cap T_n(\epsilon)'\} \\ &\leq \Pr\{\mathbf{x} \in L_n \cap T_n(\epsilon)\} + \Pr\{\mathbf{x} \in T_n(\epsilon)'\} \end{aligned}$$

By (1.9), the term $\Pr\{\mathbf{x} \in T_n(\epsilon)'\}$ approaches 0. To estimate $\Pr\{\mathbf{x} \in L_n \cap T_n(\epsilon)\}$, we reason as follows:

$$\begin{aligned} \Pr\{\mathbf{x} \in L_n \cap T_n(\epsilon)\} &= \sum_{\mathbf{x} \in L_n \cap T_n(\epsilon)} p(\mathbf{x}) \\ &\leq |L_n| 2^{-n(H-\epsilon)} \quad \text{by (1.10)} \\ &\leq 2^{Rn} 2^{-n(H-\epsilon)} \quad \text{by (1.4)} \\ &= 2^{-n(H-\epsilon-R)} \\ &\rightarrow 0 \quad \text{since } R < H - \epsilon. \end{aligned}$$

Thus $\Pr\{\mathbf{x} \in L_n\} \rightarrow 0$, and so $P_e \rightarrow 1$, again as promised. In other words, if we don't have enough bits to provide codewords for all the typical sequences, we're sunk. No matter what we do, as $n \rightarrow \infty$, our error probability will approach 1. ■

An alternative formulation of the theorem is as follows.

1.2 Corollary. For each real number $R > 0$, let $P(n, R)$ denote the total probability of the most likely 2^{Rn} source sequences of length n . Then

$$\lim_{n \rightarrow \infty} P(n, R) = \begin{cases} 0 & \text{if } R < H \\ 1 & \text{if } R > H. \end{cases}$$

We illustrate the corollary in Figure 1. There we see the case of a binary source, i.e., $A = \{0, 1\}$, with $p(0) = .9$ and $p(1) = .1$. For $n = 16, 64, 256, 1024, 4096$, we have plotted $P(n, r)$ vs. R . The “jump” at $R = H = .46899$ bits becomes increasingly pronounced, as n increases, just as predicted by the Corollary 1.2.

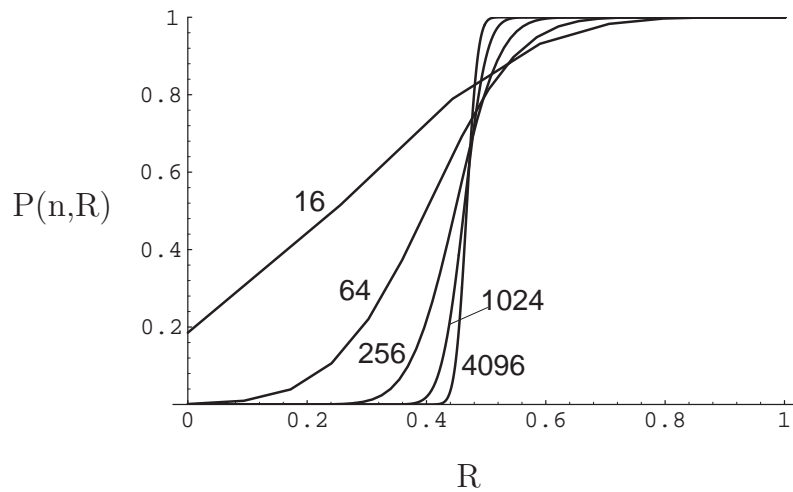


Figure 1. $P(n, r)$ (the total probability of the most probable 2^{nR} sequences of length n) vs. R for a binary source with $p(0) = .9, p(1) = .1$, for $n = 16, 64, 256, 1024, 4096$.

2. More Realistic Source Models.

Very few (if any) real data sources are accurately modelled as memoryless. Therefore as practical engineers we naturally want to know if the results of Section 1 can be extended to more general models. The answer is a resounding “yes,” but only at the cost of a considerable increase in the mathematical sophistication needed. In this section I will sketch what is known about lossless data compression for general random processes.

We begin again with a finite source alphabet $A = \{a_1, \dots, a_r\}$. Now, however, we model the source output as a *stationary random process* X_n , for $n = 0, \pm 1, \pm 2, \dots$, taking values in A . The idea is that X_n represents the output of the source at time n . Our first goal is to define the *entropy* of this source. We shall give two equivalent definitions.

The first definition of the process entropy is in terms of the n th *marginal entropy* H_n of the process, which is defined as

$$(2.1) \quad H_n = H(X_n | X_{n-1}, \dots, X_1).$$

(This is Shannon’s “ F_N ” defined on p. 55) The quantity H_n represents the uncertainty about the current source symbol, given the previous $n - 1$ source symbols. In particular, H_1 is the unconditional entropy of a single source symbol. Plainly $0 \leq H_n \leq \log r$, for all n . Also, it is easy to see that the sequence (H_n) is *decreasing*, i.e., $H_{n+1} \leq H_n$ for $n = 1, 2, \dots$. Here is a proof:

$$\begin{aligned} H_{n+1} &= H(X_{n+1} | X_n, \dots, X_2, X_1) \quad (\text{definition of } H_{n+1}) \\ &\leq H(X_{n+1} | X_n, \dots, X_2) \quad (\text{conditioning decreases } H) \\ &= H(X_n | X_{n-1}, \dots, X_1) \quad (\text{by stationarity}) \\ &= H_n \quad (\text{definition of } H_n). \end{aligned}$$

By a well-known theorem from calculus, a decreasing sequence which is bounded from below must approach a limit. Thus we define the *entropy* of the source as

$$(2.2) \quad H = \lim_{n \rightarrow \infty} H_n.$$

Intuitively, this definition says that H is the conditional uncertainty of the next source symbol, given that we have seen a very large number of previous source symbols.

The second definition is in terms of the n th *average entropy* H'_n , which is defined as

$$(2.3) \quad H'_n = \frac{H(X_1, X_2, \dots, X_n)}{n}.$$

(This is Shannon’s “ G_N ” also defined on p. 55.) An easy exercise shows that $(n + m)H'_{n+m} \leq nH'_n + mH'_m$. Mathematicians call a sequence a_n with the property $a_{n+m} \leq$

$a_n + a_m$ a *subadditive* sequence, and it can be shown that for any subadditive sequence, the limit of a_n/n , as $n \rightarrow \infty$, exists. Thus we can also define the source entropy as

$$(2.4) \quad H' = \lim_{n \rightarrow \infty} H'_n.$$

Fortunately we don't have to choose between H and H' as the definition of the entropy of a stationary source, since we have a Theorem.

2.1 Theorem. *The two entropies defined in (2.2) and (2.4) are equal, i.e., $H = H'$.*

Proof: The proof can be found in Cover and Thomas, Theorem 4.2.1. ■

As an example, consider a discrete ergodic Markov chain of the kind discussed by Shannon in Section 2.5. For simplicity, we will assume that the initial random variable X_0 is already in “equilibrium,” so that X_0, X_1, \dots is a stationary process. Then we have

$$\begin{aligned} H_n &= H(X_n | X_{n-1}, \dots, X_1) && \text{(definition of } H_n) \\ &= H(X_n | X_{n-1}) && \text{(since } (X_i) \text{ is a Markov chain)} \\ &= H(X_1 | X_0) && \text{(by stationarity)} \end{aligned}$$

Thus using the definitions (2.2) and (2.1), we see that the entropy of the Markov chain is given by

$$H = H(X_1 | X_0) = \sum_i p_i \sum_j p_{i,j} \log \frac{1}{p_{i,j}},$$

where (p_i) is the equilibrium distribution and $(p_{i,j})$ are the transition probabilities for the chain. (This is stated by Shannon in the second displayed equation in Section 2.7, p. 53.)

According to the definitions (2.3) and (2.4), the the average value of $\log p(\mathbf{x})^{-1}$ is near nH , if n is large. We are thus led to define a “typical sequence” for a general stationary source as a sequence $\mathbf{x} = (x_1, \dots, x_n)$ for which $\log p(\mathbf{x})^{-1}$ is near nH . More formally, we say an n -symbol source word \mathbf{x} is ϵ -*typical*, and write $\mathbf{x} \in T_n(\epsilon)$, if

$$\left| \frac{1}{n} \log \frac{1}{p(\mathbf{x})} - H \right| \leq \epsilon.$$

We now say that the process satisfies the (weak) *Asymptotic equipartition Property* (AEP) if*

$$(2.5) \quad \lim_{n \rightarrow \infty} \Pr\{T_n(\epsilon)\} \rightarrow 1,$$

* I use the term “weak” to denote convergence in probability, as in the weak law of large numbers. Thus the weak AEP is that $(1/n) \log p(\mathbf{x})^{-1} \rightarrow H$ in probability. The “strong” AEP is that $(1/n) \log p(\mathbf{x})^{-1} \rightarrow H$ with probability 1. The fact that every ergodic process satisfies the weak AEP is called the Shannon-McMillan theorem, and the fact that every ergodic process satisfies the strong AEP is called the Shannon-McMillan-Breiman theorem.

for all $\epsilon > 0$.

Not all stationary processes satisfy the AEP. However, an important class of them, the *ergodic* processes, do. Roughly speaking, an ergodic process is one for which almost every sample sequence exhibits statistical properties that are characteristic of the whole process. For example, in an ergodic process, almost all sample sequences exhibit the right proportion of each symbol from A , and the right proportion of all pairs, triples, etc. An example of a stationary process which is *not* ergodic is the process consisting of just two sample sequences, viz., $(\cdots 00000 \cdots)$, and $(\cdots 11111 \cdots)$, each occurring with probability $1/2$. Then for the process as a whole, $\Pr\{X = 0\} = \Pr\{X = 1\} = 1/2$, but neither of the sample sequences exhibits this distribution. But this is the exception, rather than the rule. Most naturally occurring stationary processes are ergodic. For example, any i.i.d. process is ergodic, and any first-order Markov chain for which it is possible to get from any state to any other, is also ergodic.

So from now on let's assume that our process satisfies the AEP. Then by using exactly the same reasoning we used in the proof of Theorem 1.1, we can prove that we have

$$(2.6) \quad 2^{-n(H+\epsilon)} \leq p(\mathbf{x}) \leq 2^{-n(H-\epsilon)} \quad \text{for } \mathbf{x} \in T_n(\epsilon),$$

and that if $|T_n(\epsilon)|$ denotes the number of ϵ -typical sequences, that

$$(2.7) \quad 2^{n(H-\epsilon)-\delta_n} \leq |T_n(\epsilon)| \leq 2^{n(H+\epsilon)},$$

where $\delta_n \rightarrow 0$. Thus, as in the memoryless case, we see that for processes satisfying the AEP, for large n :

1. Almost every sequence is typical.
2. Every typical sequence has probability $\approx 2^{-nH}$.
3. There are $\approx 2^{nH}$ typical sequences.

In fact, it is these three properties, rather than (2.5), that suggests the name “asymptotic equipartition property.” So here is the big generalization of Theorem 1.1.

2.2 Theorem (The Noiseless Source Coding Theorem). *Suppose the stationary source X_n , $n = 0, \pm 1, \pm 2, \dots$, has entropy H and satisfies the AEP. If $R > H$, it is possible to choose a sequence of L_n 's satisfying (1.3) and (1.4), such that*

$$P_e \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Conversely, if $R < H$, for any sequence of L_n 's satisfying (1.3) and (1.4), it must be true that

$$P_e \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

Proof: The proof is exactly the same as the proof of Theorem 1.1. The only thing to notice is that the proof of Theorem 1.1 was based on the properties (1.9), (1.10), and (1.12), which, as we have seen, hold for any AEP source, not just for i.i.d. sources. ■