A CONVERGENT GAMBLING ESTIMATE OF THE ENTROPY OF ENGLISH

BY

THOMAS M. COVER
ROGER C. KING

TECHNICAL REPORT NO. 22
NOVEMBER 1, 1976

PREPARED UNDER THE AUSPICES
OF
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
CONTRACT AFOSR #F44620-74-C-0068

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A CONVERGENT GAMBLING ESTIMATE OF THE ENTROPY OF ENGLISH

by

Thomas M. Cover
Roger C. King

TECHNICAL REPORT #22
November 1, 1976

PREPARED UNDER THE AUSPICES
OF
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
CONTRACT AFOSR #F44620-74-C-0068

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A Convergent Gambling Estimate of the Entropy of English

Thomas M. Cover[†]
Roger C. King[††]

Abstract

In his original paper on the subject, Shannon found upper and lower bounds for the entropy of printed English based on the number of trials required for a subject to guess subsequent symbols in a given text. The guessing approach precludes asymptotic consistency of either the upper or lower bounds except for degenerate ergodic processes. In this paper, Shannon's technique of guessing the next symbol is altered by having the subject place sequential bets on the next symbol of text. If we denote by S_n the subject's capital after n bets at 27 for 1 odds, and if we assume that the subject knows the underlying probability distribution, then the entropy estimate is $\hat{H}_n(\underset{\sim}{X}) = \left[1 - \frac{1}{n} \log_{27} S_n\right] \log_2 27$ bits/symbol. If the subject does not know the true probability distribution for the stochastic process, then $\hat{H}_n(\underset{\sim}{X})$ is an asymptotic upper bound for the true entropy. If $\underset{\sim}{X}$ is stationary $E\,\hat{H}_n(\underset{\sim}{X}) \to H(\underset{\sim}{X})$, $H(\underset{\sim}{X})$ being the true entropy of the process. Moreover, if $\underset{\sim}{X}$ is ergodic, then by the Shannon-McMillan-Breiman theorem, $\hat{H}_n(\underset{\sim}{X}) \to H(\underset{\sim}{X})$ with probability one. Preliminary indications are that English text has an entropy of approximately 1.3 bits per symbol.

[†] Stanford University. The work of this author was supported under Contract AFOSR #F44620-74-C-0068.

[††] Stanford University. The work of this author was supported under Contract AFOSR #F44620-74-C-0068.

1. Introduction

The goal of this paper is to develop an accurate estimate of the entropy of printed English. For a discrete random variable Y, the entropy associated with Y is $H(Y) = -\sum_i p(y_i)\log_2 p(y_i)$ where Y takes the value y_i with probability $p(y_i)$. Let printed English be represented by the symbol $\underset{\sim}{X}$ and consist of strings of the form $(\ldots,x_{-1},x_0,x_1,\ldots)$. If we assume English to be a stationary random process, then we define the entropy $H(\underset{\sim}{X})$ of the process $\underset{\sim}{X}$ to be

$$\begin{aligned} H(\underset{\sim}{X}) &= \lim_{n\to\infty} H\left(X_n \middle| X_{n-1},\ldots,X_1\right) \\ &= \lim_{n\to\infty} \frac{1}{n} H\left(X_1,\ldots,X_n\right) \end{aligned} \qquad (1)$$

In addition, if English is an ergodic process, then the Shannon-McMillan-Breiman theorem states

$$-\frac{1}{n}\log p\left(X_1,\ldots,X_n\right) \to H(\underset{\sim}{X}) \quad \text{a.e.} \qquad (2)$$

If printed English is indeed an ergodic process, then for sufficiently large n, a good estimate of $H(\underset{\sim}{X})$ can be obtained from knowledge of $p(\cdot)$ on a randomly drawn string $(X_1,\ldots,X_n)$.

Additional comment is in order concerning what the "entropy of English" actually means. It should be realized that English is generated by many sources, and each source has its own characteristic entropy. The operational meaning of entropy is clear. It is the minimum expected number of bits per symbol necessary for the characterization of the text. A gambling approach will yield an estimate of the entropy which is consistent with the above operational meaning whether or not the assumption of ergodicity for the stochastic process of English text is satisfied.

Just as there are different entropies associated with various authors of English, there are different entropy estimates associated with different gamblers. Here, the difference in the entropy estimates is associated with the amount of money that each of the gamblers can make on the sequence and is profoundly effected by the gambler's ability to accurately quantify his previous empirical experience with English. Thus an intelligent well educated gambler will do better than a gambler untrained in quantitative thinking who is relatively unfamiliar with the language. Nonetheless, it will be true that there is an upper bound on how well a gambler can do. If there were no such bound, then the true entropy of the creative process of the writer would be zero and his writing totally predictable. This upper bound yields the entropy estimate we seek.

An extensive bibliography of papers relating directly to Shannon's paper [1] on the entropy of English is included. A brief discussion of these papers follows.

Several papers provide important theoretical material. Maixner [2] helps to clarify the details behind the derivation of Shannon's lower bound to N^{th} order entropy approximations. Background on entropy estimate limitations can be found in [3], [4], [5], and [6]. Important factors involved in eliciting probability assessments from experimental subjects can be found in Savage [7]. Consistent objective estimates of the entropy of finite alphabet ergodic processes with unknown distribution will appear in Bailey [8].

A completely different estimation technique can be found in Newman and Gerstman [9]. This paper has been quoted extensively in the psychology literature, but the theory involved does not include a proof of the consistency of its entropy estimate.

Several papers extend or comment on Shannon's empirical results for English text. Grignetti [10] recalculates Shannon's estimate of the average entropy of words in English text. Burton and Licklider [11] use longer passages of text for Shannon's estimate. Paisley [12] studies entropy variations due to authorship, topic, structure, and time of composition. Treisman [13] comments on contextual constraints in language; and Miller and Coleman [14] provide more data on the entropy of English using Newman and Gerstman's technique.

Shannon's results give emphasis to a paper on the encoding of English by White [15]. White attempts to compress English text using a dictionary encoding method.

Many other papers [16, 17, 18, 19, 20, 21] apply Shannon's technique to estimating the entropy of different languages. Tzannes, et. al., [22], and Parks [23] both measure the entropy of digitized images.

Many papers [24-27] use Shannon's estimate in related applications. The psychology literature is particularly rich with entropy estimates.

2. Shannon's Estimate

Shannon [1] found an upper bound to the entropy of printed English by eliciting knowledge of $p(\cdot)$ from a subject through the use of a guessing scheme. A subject is shown $N-1$ consecutive symbols of unfamiliar text. He is then instructed to guess the next letter in the passage. Guesses are made in decreasing order of conditional probability until a correct guess occurs. Defining $\hat{q}_i^N$ to be the relative frequency of times the subject required i guesses to discover the correct letter given the $N-1$ previous letters, we can express Shannon's upper bound as

$$H(\underset{\sim}{X}) \le -\sum_{i=1}^{27} \hat{q}_i^N \log_2 \hat{q}_i^N \qquad (3)$$

We note that the upper bound is loose for 3 reasons: 1) N is finite, 2) $\hat{q}_i^N$ is the mixture of $\hat{q}_i^N$'s conditioned on the past, 3) the sample size n is finite and thus $\hat{q}_i^N$ is a random variable that has not yet converged to its mean. The first 2 reasons cause the upper bound to be strictly greater than $H(\underset{\sim}{X})$, and the 3rd reason implies that the expectation of the upper bound will be strictly greater than the upper bound of the expectation. Shannon's bounds are derived for a subject who knows the true conditional probabilities $p(X_n|X_n,\ldots,X_{n-N+1})$. For such a subject Shannon defines q_i^N to be equal to the probability that the subject requires i guesses to discover the correct letter following a sequence of $N-1$ symbols. The basis for Shannon's experimental bounds are the following bounds:

$$\text{i)} \quad \sum_{i=1}^{27} i\left(q_i^N - q_{i+1}^N\right) \log i \le H\left(X_n|X_{n-1},\ldots,X_{n-N+1}\right) \le -\sum_{i=1}^{27} q_i^N \log q_i^N$$

$$\text{ii)} \quad H(\underset{\sim}{X}) \le H\left(X_n|X_{n-1},\ldots,X_{n-N+1}\right) \le -\sum_{i=1}^{27} q_i^N \log q_i^N \qquad (4)$$

Define a map $\phi_N : \underset{\sim}{X} \rightarrow \underset{\sim}{S}_N$,

where $\underset{\sim}{S}_N$ is a new process taking values in $\{1,2,\ldots,27\}$. The map is determined by

$$\phi_N\left(X_n,X_{n-1},\ldots,X_{n-N+1}\right) = j \;, \quad \text{if } X_n \text{ is the } j^{th} \text{ most likely symbol given } X_{n-1},\ldots,X_{n-N+1}$$

Assuming $\underset{\sim}{X}$ is an ergodic process, it is shown in Shannon [1] and Maixner [2] that

$$H\left(X_n|X_{n-1},\ldots,X_{n-N+1}\right) = H\left(S_n|S_{n-1},\ldots,S_{n-N+1}\right) \qquad (5)$$

The second bound above follows immediately, since

$$H(\underset{\sim}{X}) \leq H\left(X_n | X_{n-1}, \ldots, X_{n-N+1}\right) = H\left(S_n | S_{n-1}, \ldots, S_{n-N+1}\right)$$
$$\leq H\left(S_n\right) = - \sum_{i=1}^{27} q_i^N \log q_i^N \qquad (6)$$

The distribution over which the entropy is calculated to find the upper bound is a very rough approximation to the distribution including past information. The point is that no guessing game of this type can in general estimate $H(\underset{\sim}{X})$ accurately if $H\left(S_n | S_{n-1}, \ldots, S_{n-N+1}\right) < H\left(S_n\right)$. A derivation of Shannon's lower bound, the first bound above, can be found in Shannon [1] and Maixner [2]. The upper and lower bounds are generally not equal, and the true entropy $H(\underset{\sim}{X})$ generally falls strictly below the upper bound.

3. Gambling Approach

The essence of the gambling estimate lies in an optimal gambling scheme. Instead of guessing symbols and counting the number of guesses until correct as in Shannon's technique, the subject wagers a percentage of his current capital in proportion to the conditional probability of the next symbol in the alphabet conditioned on the past. This process is repeated on subsequent symbols of text with the subject accumulating S_n dollars after n wagers. If we have an ideal subject and he divides his capital on each bet according to the true probability distribution on the next symbol, we shall show in this section that

$$\left[1 - \frac{1}{n} \log_{27} S_n\right] \log_2 27 \to H(\underset{\sim}{X}) \quad \text{a.e.} \qquad (7)$$

This is an extension of the work of Kelly [48] and Breiman [49] on gambling on favorable independent trials to gambling on ergodic processes [50]. If the subject does not bet according to the true distribution of the process, then

$$\left[1-\overline{\lim_{n\to\infty}}\frac{1}{n}\log_{27}S_n\right]\log_2 27 \geq H(\underset{\sim}{X}) \qquad \text{a.e.} \tag{8}$$

Let the English alphabet, augmented by blanks, be represented by X and denote the set of all finite strings of symbols from X by X^*.

<u>Definition</u>: Let $b(\cdot|\cdot) : X \times X^* \to \mathbb{R}$ be called a <u>sequential gambling system</u> if the following conditions are satisfied:

$$\begin{aligned}&\text{(i)}\quad b(\cdot|\cdot) \geq 0\\ &\text{(ii)}\quad \sum_{x_{k+1}} b\left(x_{k+1}|x_k,\ldots,x_1\right) = 1\end{aligned} \tag{9}$$

<u>Definition</u>: Associated with every gambling system is a <u>capital function</u> defined recursively by:

$$\begin{aligned}&S_0(\Lambda) = 1 \qquad \text{(where } \Lambda \text{ is the null string)}\\ &S_{n+1}(x_1,\ldots,x_{n+1}) = 27\; b(x_{n+1}|x_n,\ldots,x_1)S_n(x_1,\ldots,x_n),\; n=1,2,\ldots\end{aligned} \tag{10}$$

Thus if sequential bets are placed on a sequence $x \in X^\infty$ and at time k a proportion $b\left(x_{k+1}|x_k,\ldots,x_1\right)$ of the current capital is bet on the outcome x_{k+1}, with fair odds being paid, the resultant capital is $S_{k+1}(x_1,..,x_{k+1})$

<u>Definition</u>: $S : X^* \to \mathbb{R}$ is <u>achievable</u> if there exists a sequential gambling scheme with initial capital $S(\Lambda) = 1$ achieving $S(x)$, for all $x \in X^*$.

<u>Theorem 1</u>: The capital function $S : X^* \to \mathbb{R}$ is achievable by a sequential gambling scheme <u>iff</u> for all n and for all $x \in X^*$, $S_n(x_1,\ldots,x_n)2^{-n} = p(x_1,\ldots,x_n)$ are marginal distributions for some stochastic process $\{X_i\}_{i=1}^\infty$.

The proof is given in Cover [50].

Theorem 2: For any sequential gambling scheme b', $\left(n - E \log_{27} S_n(x_1,\ldots,x_n)\right)\log_2 27 \geq H(x_1,\ldots,x_n)$ for all n, with equality iff $b' = b^*$, where

$$b^*\left(x_{k+1}|x_k,\ldots,x_1\right) = p\left(x_{k+1}|x_k,\ldots,x_1\right) \quad , \quad k=1,2,3,\ldots \quad (11)$$

The proof is given in Cover [50]. See Kelley [48] for the same result for i.i.d. processes.

Thus we have the intuitively satisfying result that to gamble optimally we simply place bets according to the conditional probability of possible outcomes given the past.

Theorem 3: Let $\{X_n\}_{n=1}^{\infty}$ be an ergodic process with distribution p.

i) If the b^* scheme is used then $\left[1 - \frac{1}{n}\log_{27} S_n\right]\log_2 27 \to H(\underset{\sim}{X})$ a.e.;

ii) For any other scheme b, $\Pr\left\{\left[1 - \overline{\lim}\,\frac{1}{n}\log_{27} S_n\right]\log_2 27 \geq H(\underset{\sim}{X})\right\} = 1$.

Proof of (i)

$$\left[1-\frac{1}{n}\log_{27} S_n\right]\log_2 27 = \left[1 - \frac{1}{n}\log 27^n p\left(X_1,\ldots,X_n\right)\right]\log_2 27 \to H(\underset{\sim}{X}) \quad \text{a.e.} \quad (12)$$

by the SMB theorem.

Proof of (ii): See [50].

We now wish to extend the b^* scheme to conditioning on the infinite past. In order to do this, the following propositions are needed. Proofs are given in Appendix A.

Proposition 1: If $\underset{\sim}{X}$ is an ergodic process

$$H(\underset{\sim}{X}) = H\left(X_n|X_{n-1},X_{n-2},X_{n-3},\ldots\right) \quad (13)$$

Proposition 2: If $\underset{\sim}{X}$ is an ergodic process

$$-\frac{1}{n}\log_2 p\left(X_1,X_2,\ldots,X_n|X_0,X_{-1},X_{-2},\ldots\right) \to H(\underset{\sim}{X}) \quad \text{a.e.} \quad (14)$$

The above propositions indicate gambling schemes conditioned on the past, finite or infinite, will achieve the same end result as the b^* scheme.

Theorem 4: If $\underset{\sim}{X}$ is an ergodic binary process and gambling schemes b^* and $\bar{b}$ are defined by

$$\begin{aligned} b^*\left(X_n|X_{n-1},\ldots,X_1\right) &= p\left(X_n|X_{n-1},\ldots,X_1\right) \\ \bar{b}\left(X_n|X_{n-1},X_{n-2},\ldots\right) &= p\left(X_n|X_{n-1},X_{n-2},\ldots\right) \end{aligned} \qquad (15)$$

are used, inducing capital functions S_n^* and S_n respectively, then

$$\left[1 - \frac{1}{n}\log_{27}S_n^*\right]\log_2 27 \rightarrow H(\underset{\sim}{X}) \quad \text{a.e.}$$

$$\text{and} \quad \left[1 - \frac{1}{n}\log_{27}S_n\right]\log_2 27 \rightarrow H(\underset{\sim}{X}) \quad \text{a.e.} \qquad (16)$$

Proof: The proof is analogous to the proof of Proposition 2 in the Appendix.

The $\bar{b}$ gambling scheme provides the tool with which to find an asymptotically correct estimate of the entropy of printed English.

The subject inspects the text thoroughly up to a point X_0. Starting with $S(\Lambda) = 1$ unit at time zero, the subject places bets according to the $\bar{b}$ scheme on the next outcome X_1. Fair odds are paid (27 for 1) and the process continues to symbol X_n of the text at which time the subject has S_n dollars where

$$\begin{aligned} S_n = S_n\left(x_1,\ldots,x_n\right) &= \bar{b}\left(x_1,\ldots,x_n\right)27^n \\ \bar{b}\left(x_1,\ldots,x_n\right) &= \prod_{k=1}^{n} b\left(x_k|x_{k-1},x_{k-2},\ldots\right) \end{aligned} \qquad (17)$$

By Theorem 4,

$$\left[1 - \frac{1}{n} \log_{27} S_n \right] \log_2 27 \to H(\underset{\sim}{X}) \quad \text{bits/symbol} \quad \text{a.e.} \qquad (18)$$

We use the $\bar{b}$ scheme since allowing the subject to inspect as much past text as he wishes allows him to formulate the best subjective opinion he can of the true statistics of the given text. Thus, roughly speaking, convergence to the entropy of the process should take place faster than if the past were limited.

4. Education of the Gambler

How is it that asking a subject to gamble will elicit an accurate entropy estimate? We have already argued that there is no way to gamble in such a manner that the expected log capital $E \log S_n$ exceeds $n - H\left(X_1, \ldots, X_n\right)$, but how can we be assured that ordinary human gamblers will choose to achieve this limit?

First let us observe that each gambler has the vague motivation to increase his capital to a large amount with high probability. We present the gambler with 3 arguments for the proportional gambling scheme:

1) Maximizing the expected log of the return is achieved by proportional gambling. Thus if the gambler's utility function is logarithmic in money, betting in proportion to the probabilities is optimal. Of course, we do not believe that a given gambler's utility function is precisely logarithmic in money, so this point is not emphasized.

2) The results of Kelly [48] and Breiman [49] indicate, for independent gambles, that maximizing the expected logarithm of the return on each gamble (which is achieved by proportional gambling) will cause one's money to grow to infinity at the highest possible rate on the condition that one does not go broke. We then argue (as shown in Cover [50]) that

if the stochastic process is ergodic then conditional proportional gambling will cause S_n to grow to infinity at the highest possible rate, with probability one. Moreover, we show that even if one is willing to go broke with probability $\lambda > 0$, that conditional proportional gambling is still optimal and the growth rate of capital is unchanged, i.e., independent of λ, $0 \le \lambda < 1$. The proof is similar to the strengthening of Shannon's weak converse to Wolfowitz's strong converse.

3) We have been able to show that proportional gambling is also competitively best. This is exciting because it is consistent with the motivation of many gamblers approached for this project, in the sense that they were interested in achieving more money on the given sequence than any of the other participants. In fact, we have the following results. Let $b(x)$ be any gambling scheme on the random variable X, $P(X=x) = p(x)$, $x \varepsilon X$. Thus $\Sigma\, b(x) = 1$, $b(x) \ge 0$, $\forall\, x \varepsilon X$. Let $O(x)$ be the odds offered given that alternative x is the outcome of the drawing of the random variable X. Thus the gambling scheme b induces the capital $S(x) = O(x)b(x)$, with probability $p(x)$. Consider the proportional gambling scheme $b^*(x) = p(x)$, $\forall x$, with induced capital $S^*(x) = O(x)p(x)$ with probability $p(x)$. Then we have the result [50]:

Theorem 5: $P\{S(X) \ge t\, S^*(X)\} \le 1/t$, for $t \ge 0$. (19)

Corollary: Let $\{X_i\}_{i=1}^{\infty}$ be a stochastic process. Let $b^*\left(x_{k+1}|x_1,\ldots,x_k\right) = p\left(x_{k+1}|x_1,\ldots,x_k\right)$ and let $b(\cdot)$ be any other sequential gambling scheme. Then

$$p\left(\frac{1}{n}\log S\left(X_1,\ldots,X_n\right) \ge \frac{1}{n}\log_2 S^*\left(X_1,\ldots,X_n\right)+\frac{t}{n}\right) \le 2^{-t} \quad \text{for all } t. \quad (20)$$

Summarizing, we see that proportional gambling is best for log utility functions, and is competitively best as well as causing one's capital to grow at the highest possible interest rate. Thus it behooves any gambler motivated by any one of these 3 considerations to gamble in a proportional manner, alloting his next bet, independently of the odds, according to the conditional probability distribution of the next symbol given the available past.

5. Operational Meaning of Gambling Estimate: Compression and Decompression Using Identical Twins

It has been asserted that the gambling approach elicits both an estimate of the true probability of the text sequence as well as an estimate of the entropy of the ensemble of English from which the sequence was drawn. In this section, we investigate the operational significance of gambling in terms of data compression. Specifically, we shall argue that if the text in question results in capital S_n , then $\log_2 S_n$ bits can be saved in a naturally associated deterministic data compression scheme. We shall further assert that if the gambling is optimal, then the data compression achieves the Shannon limit.

We shall make the assumption that there is an identical twin to the gambler who will be receiving some encoding of the text. This identical twin is assumed to have precisely the same thought processes as the encoder. (See also the Shannon twin [1].)

The scheme we shall describe is essentially the Elias coding scheme for stochastic processes with respect to the distribution $p(x(n)) \overset{\Delta}{=} 2^{-n}S(x(n))$, where we have set $x(n) = (x_1,x_2,\ldots,x_n)$. (See Elias's unpublished manuscript and Jelinek's discussion of Elias's scheme [51].)

Consider the following data compression algorithm which maps the text $x_1, x_2, \ldots, x_n$ into a code sequence $c_1, c_2, \ldots, c_k$, where $c_i \varepsilon \{0,1\}$, $x_i \varepsilon \{0,1\}$, $i = 1,2,\ldots$. (We have assumed the text to be binary, without loss of generality, to obviate certain notational problems concerning bases of logarithms, etc.) Both the compressor and the decompressor know n. Let the 2^n text sequences be arranged in lexicographical order. Thus for example, $0101101 < 0100101$. The encoder observes the sequence $x(n) = (x_1, x_2, \ldots, x_n)$. He then inspects his mental processes to calculate what his capital $S_n(x'(n))$ would have been on all sequences $x'(n) \leq x(n)$ and calculates $F(x(n)) = \sum_{x'(n) \leq x(n)} 2^{-n} S_n(x'(n))$. Clearly $F(x(n)) \varepsilon [0,1]$. Let $k = \lceil n - \log S_n(x(n)) \rceil \overset{\Delta}{=} \lceil -\log p(x(n)) \rceil$. Now express $F(x(n))$ as a binary decimal to k place accuracy: $\lfloor F(x(n)) \rfloor = .c_1 c_2 \ldots c_k$. The sequence $c(k) = (c_1, c_2, \ldots, c_k)$ is transmitted to the decoder.

The decoder twin can calculate the precise $S(x'(n))$ associated with each of the 2^n sequences $x(n)$. He thus knows the cumulative sum of $2^{-n} S(x'(n))$ up through any sequence $x(n)$. He tediously calculates this sum until it first exceeds $.c(k)$. The first sequence $x(n)$ such that the cumulative sum falls in the interval $[.c_1 \ldots c_k, .c_1 \ldots c_k + (1/2)^k)$ is uniquely defined, and the size of $S(x(n))/2^n$ guarantees that this sequence will be precisely the encoded $x(n)$. Thus the twin has uniquely recovered $x(n)$. The number of bits required is $k = \lceil n - \log S(x(n)) \rceil$. The number of bits saved is $n - k = \lfloor \log S(x(n)) \rfloor$. For proportional gambling, $S(x(n)) = 2^n p(x(n))$; thus $Ek = \sum p(x(n)) \lceil -\log p(x(n)) \rceil \leq H(X_1, \ldots, X_n) + 1$. (An encoding-decoding algorithm for optimal data compression using only two operations per bit has been developed by Pasco [52].)

Thus we see that if the betting operation is deterministic and is known both to the encoder and the decoder that the number of bits necessary to encode $x_1,\ldots,x_n$ is approximately $n - \log S_n$ and that the expected value of this quantity is $H(X_1,\ldots,X_n)$. Thus for the text used in this experiment we argue that the gambling results correspond precisely to the data compression that would have been achieved by the given human encoder-decoder indentical twin pair.

In the section on evaluation of experimental results, we see that the possibility of the identical twin encoding-decoding scheme applies in Section 7 to 1) the average capital scheme, where now we need an identical twin committee on the other end; to 2) the best subject estimate scheme, where now we need an extra log n bits of information to specify which of the gamblers is used for decoding, and to 3) the committee gambling scheme, where we have an identical committee on the other end. Thus the computed entropy estimates correspond to the actual compressions which are achieved.

6. Experimental Results

The above gambling procedure was carried out using twelve subjects and a sample of text from the same source Shannon used, Jefferson the Virginian, by Dumas Malone*. The sample of text used is given in Appendix B. Table I shows the resultant entropy estimates.

* Awarded Pulitzer Prize in 1974 for his five volume series, Jefferson and His Time.

Table I Experimental Results on Estimating the Entropy of English Using a Sequence of 75 Symbols From Jefferson the Virginian

Subject	Capital Achieved	Resultant Entropy Estimate
1	1.50×10^{78}	1.29 bits/sym.
2	1.46×10^{76}	1.38
3	3.36×10^{75}	1.41
4	2.37×10^{73}	1.51
5	6.45×10^{71}	1.57
6	3.22×10^{71}	1.59
7	2.30×10^{70}	1.64
8	4.00×10^{70}	1.67
9	2.21×10^{69}	1.68
10	9.63×10^{68}	1.70
11	3.88×10^{67}	1.76
12	3.60×10^{64}	1.90
Average Capital Achieved:	1.28×10^{77}	1.34

Under the assumption that a more current piece of literature relating more directly to the subjects involved in the experiment might give a better estimate, Contact: The First Four Minutes, by Leonard and Natalie Zunin, was chosen as a second text source. The passage used appears in Appendix B. The experiment is still proceeding and thus the actual test segment of text does not appear. The results from the first two subjects are given in Table II.

Table II Experimental Results on Estimating the Entropy of English Using a Sequence of 220 Symbols from Contact.

Subject	Capital Achieved	Resultant Entropy Estimate
1	5.62×10^{231}	1.26 bits/sym.
2	6.01×10^{228}	1.30

7. Evaluation of Experimental Results

If only one experimental subject is available, then $\hat{H} = \left[1 - \frac{1}{n}\log_{27} S_n\right]\log_2 27$ is the natural estimate of the entropy, as argued previously. Now we consider natural methods of combining the performance of several experimental subjects in order to obtain a better estimate of $H(\underset{\sim}{X})$. There are two sources of error:

1) Bias. A subject may use an "incorrect" $p(x(n))$.

2) Statistical error. The sequence $x(n)$ may not be typical of the process, i.e., $-\frac{1}{n}\log_2 p(x(n))$ may differ significantly from $H(\underset{\sim}{X})$. The first source of error is handled by convexity, and the second by the asymptotic equipartition property.

Let subject i, $i = 1,2,\ldots,m$, use gambling scheme $b_i(x(n))$, thus accumulating capital $S_n^{(i)} = b_i(x(n))27^n$. Consider the following four natural estimates for S_n and $H(\underset{\sim}{X})$. In the first 3 of these, let $\hat{H} = (1 - \frac{1}{n}\log_{27} S_n)\log_2 27$.

a) Average Capital: $\overline{S}_n = \frac{1}{m}\sum_{i=1}^{m} S_n^{(i)}$. This is equivalent to a gambling scheme

$$b_{avg}(x(n)) = \sum_{i=1}^{m} \frac{1}{m} b_i(x(n)) , \qquad (21)$$

i.e., each gambler begins with $(\frac{1}{m})^{\underline{th}}$ of 1 unit.

b) Best Subject Estimate: $S_n = \max\limits_{i \in \{1,2,\ldots,m\}} S_n^{(i)}$ (22)

c) Committee Gambling:

$$b(x_k|x(k-1)) = \sum_{i=1}^{m} \alpha_k^{(i)} b_i(x_k|x(k-1)) \tag{23}$$

where $\sum\limits_{i=1}^{m} \alpha_n^{(i)} = 1$, $\alpha_1 \varepsilon [0,1]$ $i = 1,\ldots,m$ and $S_n = 27^n \prod\limits_{k=1}^{n} b(x_k|x(k-1)$.

d) Average Entropy Estimate:

$$\hat{H} = \frac{1}{m} \sum_{i=1}^{m} \left(1 - \frac{1}{n} \log_{27} S_n^{(i)}\right) \log_2 27 = \frac{1}{m} \sum_{i=1}^{m} \hat{H}^{(i)} \tag{24}$$

We reject (d) immediately because it is too sensitive to poor gambling schemes on the part of one or more of the subjects. Suppose, for example, that b_1 bets all of his capital on one symbol at time 1 and loses. Then $S_n^{(1)} \equiv 0$, $n = 1,2,\ldots,$ and $\frac{1}{n} \log_{27} S_n^{(1)} = -\infty$, for all n , thus yielding $\hat{H} = +\infty$. This is an absurd use of the data.

Suppose that subject i achieves a limit $H^{(i)}$, i.e.,

$$\left[1 - \frac{1}{n} \log_{27} S_n^{(i)}\right] \log_2 27 \to H^{(i)} \tag{25}$$

Thus $H^{(i)}$ is his asymptotic estimate of the entropy.

We now show that (a) and (b) both yield $\hat{H} = \min\limits_i H^{(i)}$ as the asymptotic estimate of H . Without loss of generality, let

$$H^{(1)} < H^{(i)} , \quad \text{for all } i \tag{26}$$

Note that in (a),

$$\left[1-\frac{1}{n}\log_{27}\bar{S}_n\right]\log_2 27$$

$$= \left[1-\frac{1}{n}\log_{27}\frac{1}{m}\sum_{i=1}^{m} S_n^{(i)}\right]\log_2 27$$

$$= \left[1-\frac{1}{n}\log_{27}\left(S_n^{(1)}\right)-\frac{1}{n}\log_{27}\frac{1}{m}\sum_{i=1}^{m}\left(S_n^{(i)}\Big/S_n^{(1)}\right)\right]\log_2 27$$

$$\to H^{(1)} \tag{27}$$

since the last term → 0 , by (25) and (26). Thus (a) and (b) have the same limit for the number of subjects m fixed and $n \to \infty$.

A similar argument can be formulated in the case of committee gambling for the special case of

$$\alpha_k^{(i)} = \alpha_k^{(i)}(x(k-1)) = \frac{S_k^{(i)}(x(k-1))}{\sum_{i=1}^{m} S_k^{(j)}(x(k-1))}$$

This corresponds to a weighted average of several betting schemes where the weighting factor is the proportion of money won by the i$^{\underline{th}}$ scheme at time n . From (c) we see:

$$b(x_{k+1}|x(k)) = \sum_{i=1}^{m}\frac{S_k^{(i)}}{\sum_{j=1}^{m} S_k^{(i)}}\; b_i(x_{k+1}|x(k))$$

$$= \sum_{i=1}^{m}\frac{b_i(x(k))}{\sum_{j=1}^{m} b_j(x(k))}\; b_i(x_{k+1}|x(k))$$

$$= \frac{\frac{1}{m}\sum_{i=1}^{m} b_i(x(k+1))}{\frac{1}{m}\sum_{j=1}^{m} b_j(x(k))}$$

$$= \frac{b_{avg}(x(k+1))}{b_{avg}(x(k))} = b_{avg}(x_{k+1}|x(k)) \qquad (28)$$

where b_{avg} is the gambling scheme resulting in an average capital estimate. Thus (a), (b), and (c) are all equivalent in the special case when the weighting factor $\alpha_n^{(i)}$ is proportional to the current capital earned by the $i^{\underline{th}}$ gambler.

In general, any other linear combination is possible and may do better than (a), (b), or (c). As an example consider $\alpha_n^{(i)} = \frac{1}{m}$. Using the data in Table I and the probability distributions guessed at each point by each of the 12 participants we arrive at $\hat{H} = 1.25$, a lower estimate than the best subject achieved. However, any choice of $\alpha_n^{(i)}$ yields an estimate, the expectation of which is an upper bound to $H(\underset{\sim}{X})$.

A summary of all of the above schemes as applied to the data in Table I is given in Table III.

Table III Data Evaluation Using Different Entropy Estimates on Data in Table I.

a) Average Capital Estimate:

$$\bar{S}_{75} = 1.28 \times 10^{77} \; ; \quad \hat{H} = 1.34$$

b) Best Subject Estimate:

$$S_{75} = 1.50 \times 10^{78} \; ; \quad \hat{H} = 1.29$$

c) Committee Gambling Estimate:

$$\alpha_n^{(i)} = \frac{1}{12} \; ; \quad S_{75} = 1.24 \times 10^{79} \; ; \quad \hat{H} = 1.25$$

d) Average Entropy Estimate (a rejected method):

$$\hat{H} = 1.59$$

8. Conclusions

Using the committee decision estimate as the estimate of the entropy of printed English, we discover a redundancy of at least 64%. The gambling winnings leading to this estimate have a direct data compression interpretation (Section 5). Thus the ability of the experimental subjects to quantify their predictions would enable them to describe the given text in 36% of the original length. The subjects used during the experiment might be improved. An English professor who spends half his time in the Las Vegas casinos would be a good choice. The results of this paper also apply to the complexity of images and to the extensive psychological research literature developed around Shannon's technique.

Appendix A

For purpose of clarity, the following propositions are proved for a binary random process.

Proposition 1: If $\underset{\sim}{X}$ is an ergodic binary process

$$H(\underset{\sim}{X}) = H(X_n|X_{n-1},X_{n-2},X_{n-3},\dots) \tag{29}$$

Proof: Let $h(p) = -p \log_2 p-(1-p) \log_2 (1-p)$ (30)

We have by definition

$$H(X_n|X_{n-1},\dots,X_{n-k}) = E\ h(p(X_n|X_{n-1},\dots,X_{n-k})) \tag{31}$$

where the expectation is with respect to $(X_n,X_{n-1},\dots,X_{n-k})$. But $\{h(p(X_n|X_{n-1},\dots,X_{n-k}))\}_{n=1}^{\infty}$ is a sequence of real random variables satisfying

$$0 \leq h(p(X_n|X_{n-1},\dots,X_{n-k})) \leq 1 \tag{32}$$

and by Martingale convergence

$$P(X_n|X_{n-1},X_{n-2},\dots) = \lim_{k\to\infty} p(X_n|X_{n-1},\dots,X_{n-k}) \quad \text{a.e.} \tag{33}$$

Thus by the continuity and boundedness of h ,

$$h(p(X_n|X_{n-1},\dots,X_{n-k})) \to h(p(X_n|X_{n-1},X_{n-2},\dots)) \quad \text{a.e.} \tag{34}$$

Finally, by dominated convergence

$$\begin{aligned} H(X_n|X_{n-1},X_{n-2},\dots) &= E\ h(p(X_n|X_{n-1},X_{n-2},\dots)) \\ &= E \lim_{k\to\infty} h(p(X_n|X_{n-1},\dots,X_{n-k})) \\ &= \lim_{k\to\infty} E\ h(p(X_n|X_{n-1},\dots,X_{n-k})) \\ &= H(\underset{\sim}{X}) \end{aligned} \tag{35}$$

Proposition 2: If $\underset{\sim}{X}$ is an ergodic binary process

$$- \frac{1}{n} \log_2 p(X_n, \ldots, X_1 | X_0, X_{-1}, \ldots) \rightarrow H(\underset{\sim}{X}) \quad \text{a.e.} \tag{36}$$

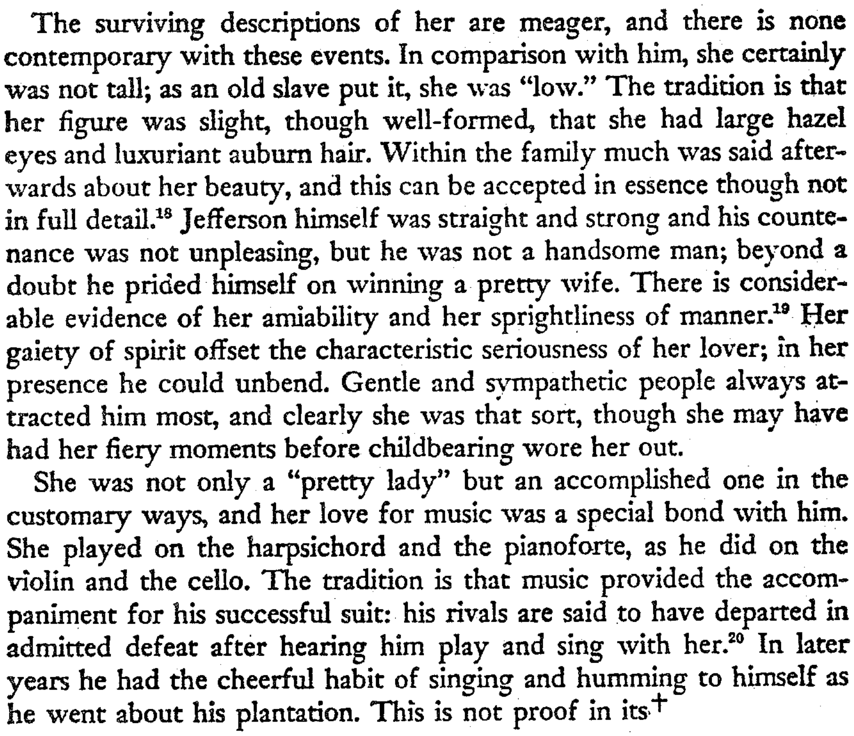
Proof:

$$\begin{aligned} & - \frac{1}{n} \log p(X_1, \ldots, X_n | X_0, X_{-1}, \ldots) \\ = & - \frac{1}{n} \sum_{i=1}^{n} \log p(X_i | X_{i-1}, X_{i-2}, \ldots) \rightarrow H(\underset{\sim}{X}) \quad \text{a.e.} \end{aligned} \tag{37}$$

by the ergodic theorem in combination with Proposition 1 and the Shannon-McMillan-Breiman theorem.

APPENDIX B

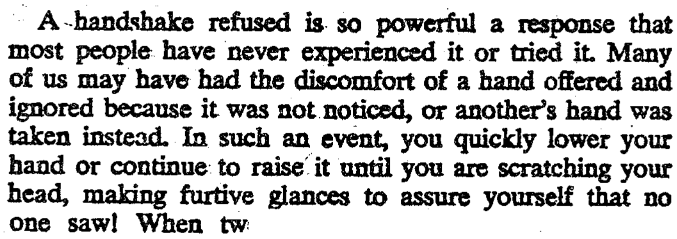
Excerpt from Jefferson the Virginian, by Dumas Malone (test section given in footnote).



The surviving descriptions of her are meager, and there is none contemporary with these events. In comparison with him, she certainly was not tall; as an old slave put it, she was "low." The tradition is that her figure was slight, though well-formed, that she had large hazel eyes and luxuriant auburn hair. Within the family much was said afterwards about her beauty, and this can be accepted in essence though not in full detail.[18] Jefferson himself was straight and strong and his countenance was not unpleasing, but he was not a handsome man; beyond a doubt he prided himself on winning a pretty wife. There is considerable evidence of her amiability and her sprightliness of manner.[19] Her gaiety of spirit offset the characteristic seriousness of her lover; in her presence he could unbend. Gentle and sympathetic people always attracted him most, and clearly she was that sort, though she may have had her fiery moments before childbearing wore her out.

She was not only a "pretty lady" but an accomplished one in the customary ways, and her love for music was a special bond with him. She played on the harpsichord and the pianoforte, as he did on the violin and the cello. The tradition is that music provided the accompaniment for his successful suit: his rivals are said to have departed in admitted defeat after hearing him play and sing with her.[20] In later years he had the cheerful habit of singing and humming to himself as he went about his plantation. This is not proof in its[†]

Excerpt from Contact, by Leonard and Natalie Zunin (test section omitted).



A handshake refused is so powerful a response that most people have never experienced it or tried it. Many of us may have had the discomfort of a hand offered and ignored because it was not noticed, or another's hand was taken instead. In such an event, you quickly lower your hand or continue to raise it until you are scratching your head, making furtive glances to assure yourself that no one saw! When tw

[†] Test sequence: elf that he was a pleasing vocal performer but with Martha in the parlor it

References

[1] Shannon, C.E., "Prediction and Entropy of Printed English," The Bell System Technical Journal, pp. 50-64, January 1951.

[2] Maixner, V., "Some Remarks on Entropy Prediction of Natural Language Texts," Information Stor. Retr., Vol. 7, pp. 293-295, 1971.

[3] Blyth, C.R., "Note on Estimating Information," Tech. Report #17, Dept. of Statistics, Stanford University, 1958.

[4] Nemetz, T., "On the Experimental Determination of the Entropy," Kybernetik 10, pp. 137-139, 1972.

[5] Basharin, G.P., "On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables," Theory of Probability Applications, Vol. 4, No. 3, pp. 333-336.

[6] Pfaffelhuber, E., "Error Estimation for the Determination of Entropy and Information Rate from Relative Frequencies," Kybernetik 8, pp. 50-51, 1971.

[7] Savage, Leonard J., "Elicitation of Personal Probabilities," Journal of the American Statistical Association, Vol. 66, No. 336, pp. 783-801, December 1971.

[8] Bailey, David H., "Sequential Schemes for Classifying and Predicting Ergodic Processes," Ph.D. thesis, Stanford University, 1976.

[9] Newman, E.B. and Gerstman, L.J., "A New Method for Analyzing Printed English," Journal Exp. Psych. 44, pp. 114-125, 1952.

[10] Grignetti, M., "A Note on the Entropy of Words in Printed English," Information and Control 7, pp. 304-306, 1964.

[11] Burton, N.G. and Licklider, J.C.R., "Long-Range Constraints in the Statistical Structure of Printed English," American Journal of Psychology, No. 68, pp. 650-653, 1955.

[12] Paisley, W.J., "The Effects of Authorship, Topic, Structure, and Time of Composition on Letter Redundancy in English Texts," Journal of Verbal Learning and Verbal Behavior 5, pp. 28-34, 1966.

[13] Treisman, A., "Verbal Responses and Contextual Constraints in Language," Journal of Verbal Learning and Verbal Behavior 4, pp. 118-128, 1965.

[14] Miller, G.R. and Coleman, E.B., "A Set of Thirty-six Prose Passages Calibrated for Complexity," Journal of Verbal Learning and Verbal Behavior 6, pp. 851-854, 1967.

[15] White, H.E., "Printed English Compression by Dictionary Encoding," Proceedings of the IEEE, Vol. 55, No. 3, pp. 390-396, March 1967.

[16] Jamison, D. and Jamison, K., "A Note on the Entropy of Partially-Known Languages," Information and Control 12, pp. 164-167, 1968.

[17] Rajagopalan, K.R., "A Note on Entropy of Kannada Prose," Information and Control 8, pp. 640-644, 1965.

[18] Newman, E., and Waugh, N., "The Redundancy of Texts in Three Languages," Information and Control 3, pp. 141-153, 1960.

[19] Siromoney, G., "Entropy of Tamil Prose," Information and Control 6, pp. 297-300, 1963.

[20] Balasubrahmanyam, P. and Siromoney, G., "A Note on Entropy of Telugu Prose," Information and Control 13, pp. 281-285, 1968.

[21] Wanas, M.A., Zayed, A.I., Shaker, M.M. and Taha, E.H., "First- Second- and Third-Order Entropies of Arabic Text," Information Theory, Vol. IT-22, No. 1, pp. 123, January 1976.

[22] Tzannes, N., Spencer, V., and Kaplan, A., "On Estimating the Entropy of Random Fields," Information and Control 16, pp. 1-6, 1970.

[23] Parks, J.R., "Prediction and Entropy of Half-Tone Pictures," Behavioral Science, Vol. 10, pp. 436-445, 1965.

[24] Miller, G.A. and Frick, F.C., "Statistical Behavioristics and Sequences of Responses," Psychology Review 56, pp. 311-324, 1949.

[25] Miller, G.A. and Selfridge, J.A., "Verbal Context and the Recall of Meaningful Material," American Journal of Psychology, Vol. 63, pp. 176-185, 1950.

[26] Newman, E.B., "Computational Methods Useful in Analyzing Series of Binary Data," American Journal of Psychology, Vol. 64, pp. 252-262, 1951.

[27] Newman, E.B., "The Pattern of Vowels and Consonants in Various Languages," American Journal of Psychology, Vol. 64, pp. 369-379, 1951.

[28] Frick, F.C. and Miller, G.A., "A Statistical Description of Operant Conditioning," American Journal of Psychology, Vol. 64, pp. 20-36, 1951.

[29] Chapanis, A., "The Reconstruction of Abbreviated Printed Messages," Journal of Experimental Psychology, Vol. 48, No. 6, pp. 496-510, 1954.

[30] Bennett, W.F., Fitts, P.M. and Noble, M., "The Learning of Sequential Dependencies," Journal of Experimental Psychology, Vol. 48, No. 4, pp. 303-312, 1954.

[31] Miller, G.A., Newman, E.B. and Friedman, E.A., "Length-Frequency Statistics for Written English," Information and Control 1, pp. 370-389, 1958.

[32] Carson, D.H., "Letter Constraints within Words in Printed English," Kybernetik, Vol. 1, pp. 46-54, 1961.

[33] Bourne, C.P. and Ford, D.F., "A Study of the Statistics of Letters in English Words," Information and Control 4, pp. 48-67, 1961.

[34] Hogan, J.A., "Copying Redundant Messages," Journal of Experimental Psychology, Vol. 62, No. 2, pp. 153-157, 1961.

[35] Blachman, N.M., "Prevarication vs. Redundancy," Proc. of the IRE, pp. 1711-1712, 1962.

[36] Tulving, E., "Familiarity of Letter-Sequences and Tachistoscopic Identification," American Journal of Psychology, Vol. 76, pp. 143-146, 1963.

[37] Shepard, R.N., "Production of Constrained Associates and the Informational Uncertainty of the Constraint," American Journal of Psychology, Vol. 76, pp. 218-228, 1963.

[38] Schwartz, E.S., "A Dictionary for Minumum Redundancy Encoding," Journal of the Association of Computing Machinery, Vol. 10, pp. 413-439, 1963.

[39] Bluhme, H., "Three-Dimensional Crossword Puzzles in Hebrew," Information and Control 6, pp. 306-309, 1963.

[40] Tannenbaum, P.H., Williams, F. and Hillier, C.S., "Word Predictability in the Environments of Hesitations," Journal of Verbal Learning and Verbal Behavior 4, pp. 134-140, 1965.

[41] Raviv, J., "Decision Making in Markov Chains Applied to the Problem of Pattern Recognition," IEEE Transactions on Information Theory, Vol. IT-3, No. 4, October 1967.

[42] Schwartz, E.S. and Kleiboemer, A.J., "A Language Element for Compression Coding," Information and Control 10, pp. 315-333, 1967.

[43] Thomas, R.B. and Kassler, M., "Character Recognition in Context," Information and Control 10, pp. 43-64, 1967.

[44] Cornew, R.W., "A Statistical Method of Spelling Correction," Information and Control 12, pp. 79-93, 1968.

[45] McNicol, D., "The Confusion of Order in Short-term Memory," Australian Journal of Psychology, Vol. 23, No. 1, pp. 77-84, 1971.

[46] Tuinman, J.J. and Gray, G., "The Effect of Reducing the Redundancy of Written Messages by Deletion of Function Words," The Journal of Psychology, Vol. 82, pp. 299-306, 1972.

[47] Mandelbrot, B., "Simple Games of Strategy Occurring in Communication through Natural Languages," IRE Transactions on Information Theory, Vol, PGIT-3, pp. 124-137, 1954.

[48] Kelly, J. Jr., "A New Interpretation of Information Rate," The Bell System Technical Journal, pp. 917-926, July 1956.

[49] Breiman, L., "Optimal Gambling Systems for Favorable Games," Proc. Fourth Berkeley Symposium, Vol. 1, pp. 65-78, 1961.

[50] Cover, T., "Universal Gambling Schemes and the Complexity Measures of Kolmogorov and Chaitin," Tech. Report No. 12, Statistics Dept., Stanford University, October 1974, to appear Ann. Stat.

[51] Jelinek, F., Probabilistic Information Theory, McGraw-Hill, 1968.

[52] Pasco, R.C., "Source Coding Algorithms for Fast Data Compression," Ph.D. thesis, Department of Electrical Engineering, Stanford University, May 1976.